Data Mining Integrated Architecture For Shop Floor Control System

By

Srinivas

A Doctoral Thesis

Submitted in Partial fulfilment of the requirements for the award of

Doctor of Philosophy of Loughborough University

2007

Abstract

Organizations are becoming increasingly complex with emphasis on decentralized decision making Recent advances in the field of information systems and networking have greatly changed the characteristics of demands from shop floor of an enterprise It is not only viewed as a production centre but is also considered as a nucleus of information and knowledge. This knowledge may consist of system's behaviour, limitation, capability, etc Therefore, the manufacturing system must have an information system that facilitates generation, sharing and integration of knowledge for effective and efficient decision making. In the competitive environment, organisational knowledge is not perpetual, but has a lifecycle Organisational knowledge value deteriorates with time due to changes in the competitive environment Enterprises generate an avalanche of data and information that may be critical and valuable in nature but hard to manage and leverage properly Effective decision making in a data intensive environment is likely to determine future business activities and differentiate a company from its competitors. Knowledge generation, accumulation and maintaining knowledge bases are time consuming processes but essential to development and successful application of a knowledge base system.

The two prime sources of manufacturing knowledge in an organization are its employees and its operational data, which are supported by secondary sources such as technical specifications, plant practices and operating theory. Modern manufacturing businesses store most of their data and knowledge electronically (although some tacit knowledge will only exist within their employees) The current research is built on the belief that these data stores can be valuable assets and potentially important sources to explore for new information and knowledge Data mining tools and techniques have been explored during this research as they provide a methodology for analysing real time data and can generate useful information and knowledge in many areas important for shop floor control

In this research, a proposal has been made for agent based shop floor control which incorporates data mining capabilities The architecture has been devised to provide useful information and knowledge for shop floor control The proposed model also provides a mechanism for continuous learning in the system. It also provides an approach for integration

of data mining processes for the generation of required knowledge and information by analysing operational databases for different activities on the shop floor into a decision support framework by means applying intelligent agent technology. An approach for linking different data mining agents situated at different process sites has also been discussed. The application of the data mining agent has also been examined on three different data sources (relating to three different types of functions) from the shop floor to demonstrate how knowledge can be identified and generated for reuse within the proposed architecture This research also address the challenges of management of data mining tasks and selection of appropriate tools and algorithms for it. Data cleaning and transformation in different contexts has also been discussed The research also suggests ways to improve the quality of results and knowledge generated through various data mining algorithms

Keywords:

Knowledge Discovery in Database, Agent, Agent Based System, Shop Floor Control, Decision Support System, Data Mining Algorithms, Data Mining Process, Control Signature

ACKNOWLEDGEMENT

It is the constant effort of numerous people who have helped me not to get lost during the development phases of this thesis. Dr. J.A.Harding, my supervisor, provided a motivating, enthusiastic and critical atmosphere during the plethora of discussions we had Her constructive criticisms and stimulating suggestions added great value to this thesis. It was a great pleasure for me to pursue my research under her supervision.

Whatever I be, wherever I am, I cannot forget my days when I was working under Prof. M K.Tiwari. The research which I carried out under his supervision helped me to develop a drive for research, acquire some of the useful analytical tools and unwind up my mind for further research. I would ever be grateful to him for the exposure which he provided to me in this area

I would also like to thank all the colleagues of my department especially Shahbaz, Rahul, Shilpa, Alok and my flatmate Vasil for the cooperation they rendered me.

There are some wonderful people on this earth, my Father, Mother, Brother, Sister without their moral and emotional support this work could not have been carried out.

At last, I thank God for giving me a good supervisor, parents, relatives and friends, without which it would not have been possible to reach up to this level.

Glossary

SFC	Shop floor Control
MAS	Multi Agent System
DMA	Data Mining Agent
DBMS	DataBase Management System
KDD	Knowledge Discovery in Databases. KDD is the process of identifying valid, novel, potentially useful and ultimately understandable patterns and/or models in data
Data Mining	Data mining is a step in the knowledge discovery process consisting of selecting and applying particular data mining algorithms that, under some acceptable computational efficiency limitations, find patterns or models in data
Control signature	a control signature contains a preferred set of features' values leading to selected performance measures for various operations of a manufacturing enterprise
AMRF	Automated Manufacturing Research Facility
MSI	Manufacturing System Integration
PAC	Production Activity Control
MPCS	Manufacturing Planning and Control System
FACT	Factory Activity ConTrol
CRISP-DM	Cross-Industry Standard Process for Data Mining
SEMMA	Sample, Explore, Modify, Model, Assess
SPC	Statistical Process Control
DOE	Design Of Experiment
KQML	Kernel Query and Manipulation Language
CMM	Co-Ordinate Measuring Machine

Contents

Chapter 1

Introduction

The greatest challenges facing operating and production enterprises are maintaining and often increasing, operational effectiveness, revenues and customer satisfaction, whilst simultaneously reducing capital, operating and support costs. If manufacturing firms are to compete in the ever-changing global market, it is crucial that they exploit and develop their competitive advantages and achieve low cost production This is especially difficult for companies whose operating costs rely on high investment in assets since the only way to reduce the cost of operation may be to reduce the complexity of workflow and achieve better utilization of the assets and resources within the company

The shop floor represents a fundamental part of a manufacturing enterprise, where value adding processes take place, transforming raw materials into final products Shop floor control manages the operations responsible for transforming planned orders into a set of outputs through activities such as reactive scheduling, order release, resource allocation, online process planning, data collection, monitoring, etc. Shop floor manufacturing environments consist of a number of interconnected systems with operations that are interdependent from a decision making perspective The large number of components found on the shop floor require protocols for interaction and integration of control components and reference architectures to specify their dependencies, interfaces and interaction mechanisms, constraints and rules for different activities However, assessment of past and current research efforts in manufacturing control reveals that the key assumptions of any particular control paradigm cause the inherent drawbacks on the general applicability of that paradigm (see chapter 3)

Competition for products' prices, delivery performance, customer preferences, etc , drives major changes in the production style and configuration of manufacturing organisations Traditional centralised and sequential configurations are too inflexible to respond to these

competitive drivers and traditional centralised approaches to shop floor control problems are being challenged by Multi Agent System (MAS) approaches (see chapter 4) A MAS is a distributed problem-solving paradigm, which breaks complex problems into small and manageable sub-problems which can be solved by individual agents working co-operatively. MASs provide rapid responses and dynamically reconfigurable structures which facilitate rapidly changing environments ([1, 2]). Increasing market competition and unstable demands from customers have forced manufacturers to become adaptive and responsive and therefore a shop floor controller in a dynamic environment must contain particular mechanisms and characteristics [3-5] which are explained and discussed in chapter 7. In particular better knowledge management is required to improve shop floor control, particularly as, in a competitive environment, the value of organisational knowledge is not perpetual, but has a lifecycle [6] Enterprises must therefore have the capability to learn new knowledge, propagate them in the system and then retire them at the end of their shelf life.

Enterprises often rely heavily on the expertise and experience of their employees in decision making processes, yet do not fully exploit other less clearly defined experiences and information that may also exist in their recorded product life cycle data. The relationships between operational variables and performance are usually not thoroughly analyzed and interpreted by enterprises Effective knowledge discovery approaches could enable organisations to identify, maintain and exploit their dispersed knowledge sources and experience for the long term benefit of the corporation, through improved shop floor systems The need for improved knowledge management to support decision making processes is also justified by the inability of decision makers to diagnose efficiently many of the malfunctions that arise at machine, cell and entire system levels during manufacturing [7].

Knowledge based systems embed human know how into a computer using a knowledge base, which can then be used to reason through a problem, and solve different problems, within the domain of an existing knowledge base, without reprogramming Knowledge based systems can help shop floor managers to consider various alternatives quickly, and useful knowledge to support this can come from many business functions, including purchasing, marketing, design, production, maintenance and distribution, etc The two prime sources of manufacturing knowledge in an organization are its employees and its operational data, which are supported by secondary sources such as technical specifications, plant practices and operating theory The capture of human knowledge is generally achieved through knowledge-elicitation using many established techniques But, useful knowledge may not all be well-known by human experts, or be immediately obvious or visible and instead, it may lie 'hidden' as patterns within the plant data. Data mining techniques have been recently used to extract

such knowledge from plant data and in consequence help in validating and providing alternative sources for experts' knowledge. Knowledge discovery, knowledge management and knowledge engineering are therefore currently topics of importance to manufacturing researchers and managers intent on identifying and exploiting current knowledge assets (see chapter 5).

Although potentially very valuable for improved exploitation of knowledge assets, knowledge discovery and data mining can be complex, time-consuming and expensive processes. Work published in this area of data analysis has generally been one shot experiments, and consequently has been poorly exploited, since the analysis has only been done to address particular problems Hence, efforts have not been put into reusing the lessons learnt in other contexts and exploiting the experiences gained by encoding them into the production system or actively providing ongoing feedback to other users. The primary motivation of this research is therefore to design a shop floor knowledge-based system that has an ability to identify, accumulate and exploit the (possibly hidden) knowledge that can be generated from a variety of enterprise databases, using a range of tools (see chapter 6)

An objective of this research has therefore been

To design a data supported manufacturing shop floor control system, which can benefit from its historical or legacy systems, as well as from its current databases.

This has been achieved through the satisfaction of the following aims:

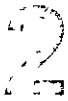
1. To design a multi agent based manufacturing system which uses a knowledge based decision support system (with a knowledge pool updated from operational databases using data mining techniques) to provide enhanced information, knowledge and alternatives to decision makers so that they can make better decisions to meet the enterprise's objectives.
2. To show how systematic enterprise knowledge discovery could be provided by exploring the operational databases of an enterprise to generate knowledge
3. To investigate knowledge discovery technology in order to determine the requirements and constraints of incorporating data mining functionality within a shop floor control system.
4 To ensure that the distributed control system of the proposed design satisfies the requirement for ease of the process of capturing knowledge associated with different groups in the manufacturing system located at different places and where available provides a correct solution or range of possible solutions in response to user queries

5 To generate control signatures (a control signature contains a preferred set of features' values leading to selected performance measures for various operations of a manufacturing enterprise) These enable feedback to be provided for disturbances in order to react quickly in different scenarios

The approach taken to meet the above aims and objective has been to initially investigate knowledge discovery through the application of data mining techniques in several scenarios and from this to determine how data mining functionality could be incorporated within a distributed, intelligent shop floor decision support tool A variety of data mining tools and techniques have been explored during this research as an understanding of their capabilities is fundamental to providing a methodology for analysing real time data that can generate useful information and knowledge in several areas important for shop floor control The results of this research work are proposals for an intelligent decision support tool that incorporates data mining and intelligent agent technology. The architecture adopted is similar to that presented by Wang [8], taking advantage of the intelligent, autonomous and active aspects of agent technology. However, in addition, it makes an original contribution to knowledge by also integrating data mining processes within the architecture for the identification of required knowledge and information for different activities on the shop floor through the application of intelligent agent technology, using a data mining agent (DMA) The knowledge generated by the DMA is integrated in the decision support system for future use

The purpose of the DMA is to extract useful knowledge from large datasets obtained from product life cycle data and store them in a knowledge pool which can be further reused. This research is based on the belief that knowledge generated from mining enterprise wide data can result in a better understanding of the consequences of decisions made at all levels of the company The multi-agent based shop floor control system has been designed with a decentralised control mechanism which uses knowledge, information and alternatives provided by the decision support tool. The knowledge pool can be generated and updated by a data mining agent, which enhances flexibilities of functioning and tackles the problem of dynamic management in real time by exploiting the various flexibilities offered by this system and provides various alternatives in different scenarios. The proposed shop floor system could then utilize the performance information provided by the knowledge pool and the data mining agent to select the optimum alternatives. Feedback can be provided to different levels in a formalised way so that discovered knowledge can be exploited and reused in various ways in the future. An important aspect of this type of knowledge based systems is that they have the potential for their knowledge to be continuously updated

Chapter 2



Research Scope

2.1 Introduction

A manufacturing enterprise is a group of interacting subsystems. Conflicts arise from the allocation of scarce resources among the competing subsystems, so the interests of each subsystem have to be arbitrated in the light of the overall effectiveness of the entire manufacturing plant. Traditional decision making has an orientation towards centralised approaches that emphasize a deterministic view of the technological and logistic production process With the increase in automation, the availability of instantaneous market information, the growing need to have much faster product cycles and the globalisation of manufacturing activities, a radical change is occurring in manufacturing strategies which are shifting from the conventional "make-and-sell" model to a new "sense-and-respond" paradigm Fast responses are needed to rapidly changing customer requirements and changing market opportunities. Greater and continuously updated process knowledge is expected to be critical to ensure effective and efficient decision making, since better quality decisions are possible if they are based on accurate and reliably updated relevant information

In recent years, information growth has proceeded at an explosive rate. Database management systems (DBMS) provide the basic tools for efficient storage and look-up of large data sets, but the capabilities for collecting and storing data have far outpaced our abilities to analyse, summarize and extract knowledge from this data. Traditional methods of data analysis were based mainly on humans dealing directly with data. Large volumes of data overwhelm the traditional manual methods of data analysis and make the task of analysis more difficult and less efficient Also, traditional methods of analysis tend to create informative reports from data, rather than analysing the contents of those reports by focusing on important knowledge

In consequence, there is likely to be additional knowledge hidden in operational databases, which, if discovered, could improve manufacturing processes and/or support more accurate modelling of the system's behaviour. This research is therefore based on the assumption that improved exploitation of such hidden knowledge through integrating data analysis processes into everyday decision making could help decision makers by providing them with enhanced information about the various alternative solutions or opportunities that might be available.

2.2 Scope of Research – Issues

To address the challenges of the research aims stated in chapter 1, it has been important to thoroughly review the current state of the art of manufacturing control systems and explore the technologies which are most likely to underpin advances in this research area. As this is a very wide-ranging topic, it has been necessary to focus research effort primarily into the areas where clear benefits and contributions have been identified Therefore priority has been given to the system design and partial implementations for demonstration of concepts, rather than to trying to implement a fully functioning prototype control system which would not be feasible in the time available. The main areas of research and associated issues for each topic are now discussed to clarify the scope of work undertaken in each of the research topics and relevant support technologies that have been considered

2.2.1 Shop Floor Control

Shop floor control manages the operations responsible for the transformation of planned orders into a set of outputs It has offered challenging problems for researchers and practitioners for many years Chapter 3 presents a detailed review of the different architectures and approaches used in resolving the issues of shop floor control and show that the challenges resulting from ever changing customer demands require dynamic, adaptive and robust architectures. During this research project, effort has primarily been focused on studying and analysing control architectures and strategy used in a shop floor control system These, and recommendations identified in the literature, have been used as the basis for the proposed architecture which has been designed to specifically address shortcomings that have been reported in published literature.

2.2.2 Agent Technology

Agent technology is derived from distributed artificial intelligence MASs are considered to be the most promising architecture for next generation manufacturing [9] since they provide rapid responses and dynamic, reconfigurable structures to facilitate flexible and efficient use of

manufacturing resources in rapidly changing environments. A detailed review of agent technology has been provided in chapter 4. During this research project, effort has primarily been focused on understanding and exploiting the flexibility and other benefits of existing agent technology, whilst determining how to incorporate enhanced knowledge management and knowledge discovery functionality in MASs.

2.2.3 Knowledge Discovery and Data Mining

Knowledge discovery in databases (KDD) and data mining is a rapidly growing interdisciplinary field which merges together database management, statistics, machine learning and related areas and is aimed at extracting useful knowledge from large collections of data The review provided in chapter 5 shows that data mining techniques have produced useful results in many manufacturing contexts. However nearly all of the reported research relates to the application of data mining techniques to solve single or "one off" problems. It is likely therefore that valuable (discovered) operational knowledge is being under-exploited. A key aspect of the approach to KDD taken in this research project is the requirement that discovered knowledge should be reused and regularly maintained for the ongoing improvement of shop floor control. In order to design an effective method for incorporating KDD within an agent-based shop floor control system it is necessary to also understand and experiment with the application of data mining techniques on different types of operational data The scope of effort on this topic has therefore been to determine when and how particular algorithms should be selected and also how to assess the quality of the information/knowledge discovered. Chapter 6 therefore explains the range of data mining approaches available and criteria for their selection whilst the architecture for reusing their results has been proposed in chapter 7 The practical examples in chapters 8, 9 and 10 demonstrate examples of data mining applications on a range of operational data, and show how the results from data mining may be evaluated

2.3 Structure of thesis

The following figure shows the structure of this thesis and how the three main background topics of research contribute to the proposed data supported manufacturing shop floor control system, that is introduced in chapter 7

The whole thesis can be divided into four sections as shown in figure 2.1.

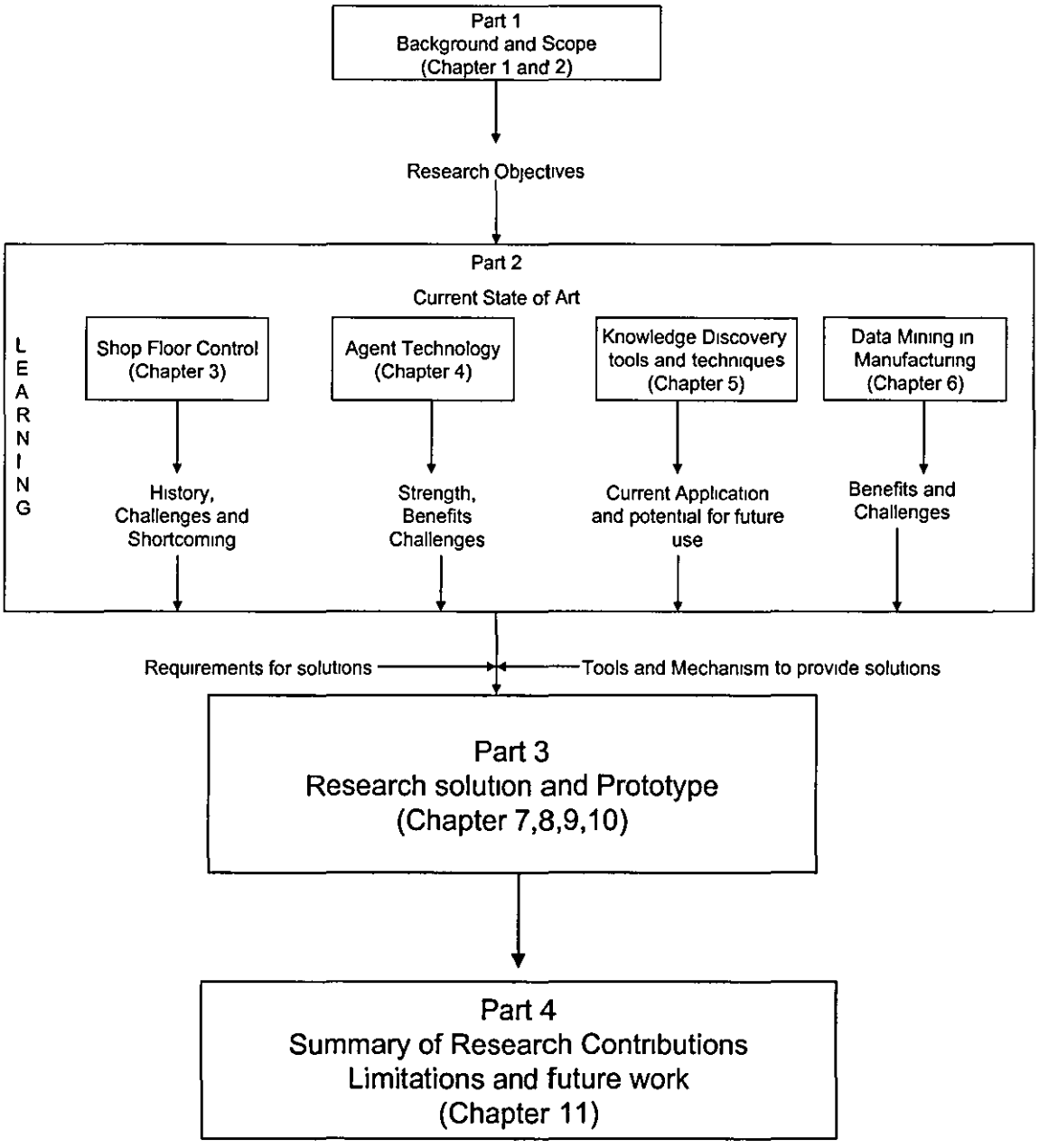


Figure 2 1: Structure of Thesis

1. **Background and Scope:** This section consists of two chapters i e. Introduction and Research Scope and it establishes the research objectives and research domain The first chapter presents the general overview of the challenges in the shop floor of manufacturing enterprises and identifies the need for better knowledge discovery and management in its decision making. It also summarises the aims and objective of the

research. The second chapter presents the scope of the research and area of learning and contributions.

2 **State of the Art:** This section presents the literature review on four different topics i.e. manufacturing control architectures, agent technology in manufacturing, application of data mining in manufacturing and different techniques in data mining This section presents wide ranging views on different areas and sets the context for the contributions reported in this thesis. Each of the chapters, 3, 4, 5 and 6, contributes to the thesis in a different way by providing the necessary understanding and learning that enabled the requirements for the proposed original research solution to be identified The review on manufacturing control architectures indicates that these control systems should be versatile and adaptive and that new knowledge (trends) must be integrated into their decision making processes. This therefore identifies the requirements for the proposed shop floor control system. The literature survey on agent technology reveals its advantages for the manufacturing domain in terms of modularity, reconfigurability, adaptability, scalability, upgradeability and robustness. These characteristics have therefore been exploited in the proposed shop floor control system The review of data mining tools and their application in the manufacturing domain identifies that operational databases in manufacturing enterprises can be used as a source of new knowledge (trends) Chapters 5 and 6 therefore support the validity of the proposed system and also provide background understanding of the range data mining tools that exist and understanding of how they can best be applied

3 **Research Solution and Prototype:** This section consists of four chapters describing the architecture that has been designed for integrating and reusing the knowledge generated by the data mining process Chapter 7 describes the functioning of different agents in the system and also presents a partial prototype to demonstrate its concepts of utilising the knowledge stored in knowledge pools These knowledge pools are created and updated through data mining processes being applied on the operational databases of a manufacturing enterprise The next three chapters show examples of the application of different data mining techniques on operational data from different sources in manufacturing contexts to present a methodology to generate knowledge Chapter 8 discusses the application data mining process on product data to generate knowledge about manufacturing capabilities and design limitations. Chapter 9 shows how the control signature for the product can be

generated and also provides feed back for SPC processes Chapter 10 generates control signature for a maintenance operation

4. **Conclusion:** This section consists of chapter 11 which presents a summary of the work pursed in this research and also discusses its benefits and the future extensions of this work.

Chapter 3

Manufacturing Control: State-of-the-Art

3.1 Introduction

The shop floor represents a fundamental part of the enterprise, where value adding processes occur and transform raw materials into final products. The shop floor environment in manufacturing is comprised of a number of systems that are not only interconnected, but whose operations are interdependent from a decision making perspective. The large number of components found on the shop floor requires an architecture that specifies the protocols for the interaction and integration of the control components. This chapter presents an overview of published research in the area of manufacturing control architectures The approach taken in this thesis is to define shop floor control and manufacturing control to include resource allocation (scheduling) aspects as well as process specific aspects (process planning, etc) This is an addition to most of the reported architectures, which only consider resource allocation aspects Assessment of past and current research efforts in manufacturing control reveals that the key assumptions of each control paradigm cause inherent drawbacks on the general applicability of the paradigm.

3.2 Existing Manufacturing Control Architecture

Shop floor control manages the operations responsible for the transformation of planned orders into a set of outputs It consists of several activities reactive scheduling, order release, resource allocation, online process planning, data collection, monitoring, etc. The reference architectures specify the dependencies, interfaces and interaction mechanisms, constraints and rules of the system for different activities.

Dilts et al [10] give a detailed overview of the evolution from centralised control, over hierarchical and modified hierarchical, to heterarchical control. Their paper presents a detailed discussion of the characteristics, advantages, and drawbacks of each of these concepts, however, since its publication, control paradigms have evolved to reduce the identified drawbacks, and increase their applicability to a wider range of manufacturing systems.

The following section presents an overview of architectures, control concepts and typical assumptions made about manufacturing systems and their environments.

3.2.1 Hierarchical Control System

Hierarchical control architectures have a strong association with the software structures of traditional control architectures and are characterised by a fixed structure and large volume of global information [11]. The motivation for such architectures is derived from the natural hierarchy existing in all complex organisational systems. In these architectures, a master module co-ordinates the operations of all the facilities The communications between the master and its sub-ordinate module is carried out by passing commands and receiving feedback from the sub-ordinate modules through well organised and defined communication protocols. An advantage of hierarchical control is that it allows the control problem to be divided to limit the complexity of the entire structure [12] The following overview of these architectures concentrates on three tracks which depict the evolution from hierarchical to modified hierarchical control

3.2.1.1 Track 1: NIST Originated Architectures

3.2.1.1.1 AMRF(Automated Manufacturing Research Facility)

The National Bureau of Standards(NBS) now known as the National Institute of Standards and Technology(NIST) established a Hierarchical control model AMRF (Automated Manufacturing research facility) [13]. The architecture utilizes the hierarchical structures (tree shaped) that exist in many complex organizational models and recognizes five levels: Facility, Shop, Cell, Workstation and Equipment The Facility level consists of three major sub-systems: manufacturing engineering, information management and production management The Shop level is comprised of two modules· task and resource management and carries out the real time management of jobs and resources on the shop floor The Cell level performs sequencing of batches and material handling facilities The Workstation level directs and co-ordinates a set of equipment on the shop floor Instruction/feedback control flow is limited only to sub ordinate/supervisor level. However several levels share some data. The tree structure at the cell level varies depending on the part to be manufactured. The concept of

virtual manufacturing cells is exploited to avoid the limitations of Group technology (GT) cells [14], as it provides greater flexibility. At a higher level, the shop floor control system schedules the virtual GT cell depending upon the part to be produced and the priority of different batches. Although this virtual cell concept allows some dynamic configuration at the cell level, mixed model production still causes problems as a workstation can belong to only one virtual cell at any given time and thereby only one virtual cell can be active at any given time The purpose of the AMRF is to provide a test bed for research directed towards the development of standards for use in manufacturing systems ([12, 13]).

3.2.1.1.2 MSI (Manufacturing System Integration)

Senehi[15] elaborated the Manufacturing System Integration Architecture (MSI), which is a product of the MSI project conducted from 1990-1993 at NIST. The shop controller is at the

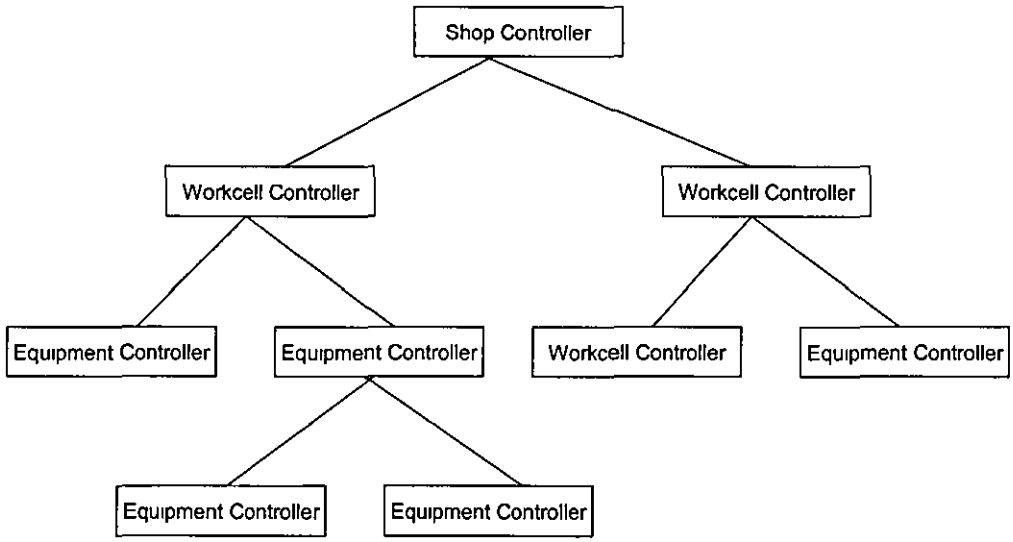


Figure 3.1: The branches of MSI control Architecture (15)

top and the equipment controller is at the bottom of the hierarchy The number of hierarchies between these levels depends on the branches in the tree structure, as shown in figure 3.1. Every subordinate controller has exactly one supervisor controller and the control hierarchy remains unchanged during the operation of the shop floor. However, the architecture can be reconfigured by adding or removing a controller, but at any fixed time the MSI architecture has a single control hierarchy Most entities interact by passive sharing of information in a database Peer controllers on the same level may share information. However, there exists a strong master/slave relationship between two adjacent layers in the hierarchies The shared information model contains process plans, resources, orders, product models, etc. MSI exploits the natural hierarchy present in the organisation MSI uses dynamic scheduling, reallocation of

resources and dynamic process planning as error recovery means The primary concern in the MSI project has been to develop standards for systems (i e controller) interconnection

The original NIST implementation lacked sophisticated planning and scheduling at lower levels in the hierarchy[16]. As a result there exist only two distinguishable layers. decision making and control Taking this fact into consideration, Jones and Saleh[16] present a "multi-layer/multi-level" control architecture based on both spatial and temporal decomposition of the system The chief feature of this architecture is the decomposition of the control task at each level into adaptation, optimisation and regulation functions. Joshi et al [17] developed a three level hierarchical control model based on the lower three levels of the NIST hierarchy in which the tasks for each controller are separated into planning, scheduling and control.

3.2.1.2 TRACK 2: PAC Based Architecture

3.2.1.2.1 PAC(Production Activity Control)

The Production Activity Control (PAC) architecture evolved from the ESPRIT project 477, COSIMA (COntrol Systems for Integrated Manufacturing) ([18, 19]). In this architecture,

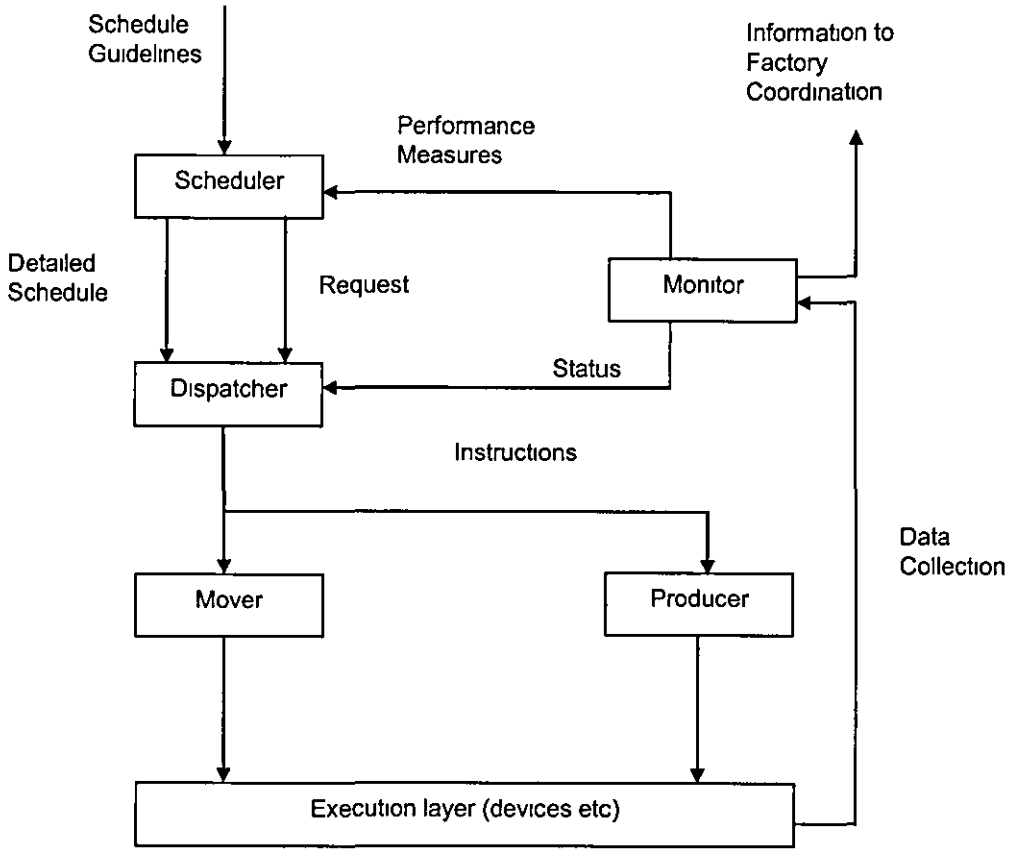


Figure 3 2· Production Activity Control ([18])

production planning and control is separated into three hierarchical activities strategic, tactical, and operational Shop floor control is one of the concerns of the operational level. The shop is assumed to be composed of a series of product based manufacturing cells controlled by PAC controllers These PAC controllers are supervised by a Factory Co-ordination (FC) level The architecture is shown in figure 3.2 and it consists of a scheduler, a dispatcher, a monitor, movers and producers. The scheduler accepts the production requirements from the factory coordination level and generates a detailed plan for the use of the manufacturing facilities in the given time horizon, which is implemented by the remaining four modules The Dispatcher controls the flow of work within the cell on a real time basis. The Monitor observes and passes information back to the higher level The Mover organizes the movement of materials and the Producer controls the sequence of operations at each work station.

The Factory Co-ordination executes the two functions of controlling and designing the production environment The production environment design task consists of process planning, layout maintenance and analysis of the production system to ensure continuous improvement and reorganization to support product based layout. The control task is similar to the PAC in which the movers and producers are replaced by the PAC cells The scheduler, dispatcher, and monitor co-ordinate the instructions for the PAC cells

3.2.1.3 TRACK 3: From MPCS to FACT

3.2.1.3.1 MPCS (Manufacturing Planning and Control System)

Biemans [20] proposed a Manufacturing Planning and Control System Reference model (MPCS) consisting of two layers, MPCS execution and MPCS management, see figure 3 3. MPCS execution performs the control task and has a six layer hierarchy consisting of factory controller, cell/line controller, work station controller, automation module controller, device controller and sensors/actuators. The architecture depicts horizontally layered controllers, each with different tasks Each controller enhances the service of the aggregate of all its subordinate controllers. MPCS executer receives commands from the MPCS manager for transforming the raw materials into final products in a certain time frame. The MPCS executor realizes those commands. The constraints of the manufacturing system cannot be changed during execution. The MPCS management consists of a master planner, product and process developer, a supervisor, etc. It interacts with the factory controller to receive commands for dispatch of products at due dates to customers. The MPCS manager controls the various other activities of the shop floor.

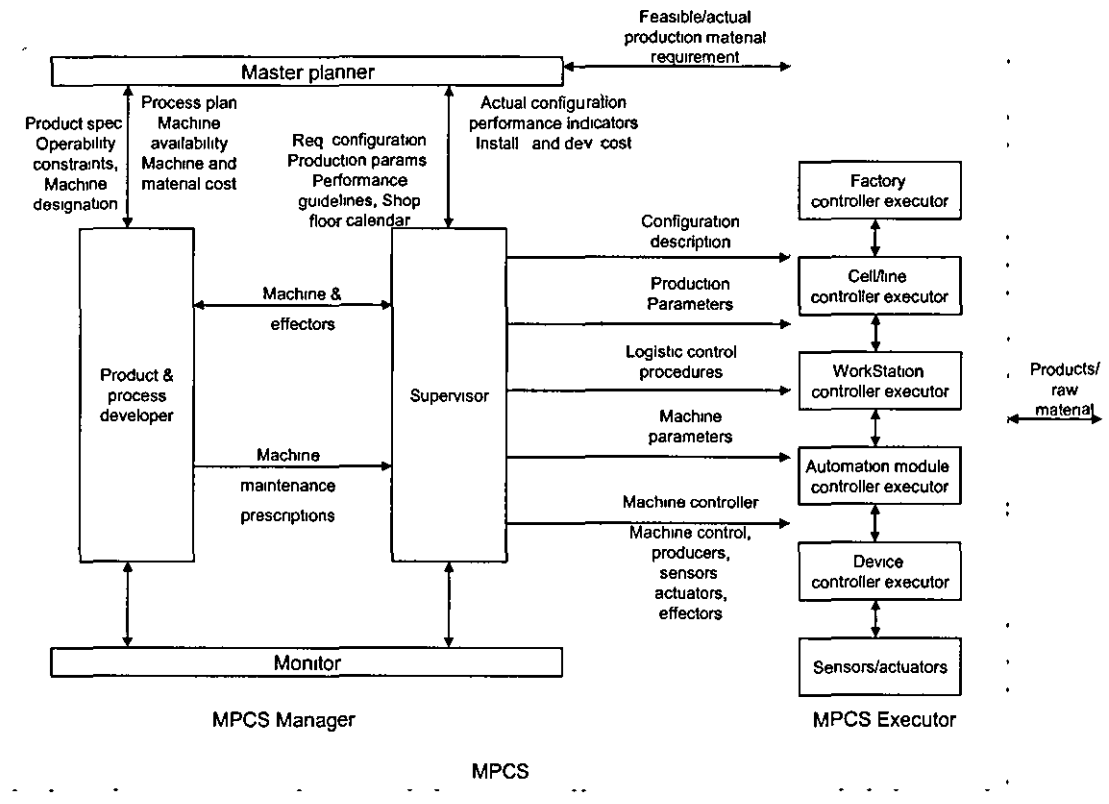


Figure 3.3: The MPCS reference model([20])

3.2.1.3.2 FACT (Factory Activity ConTrol)

Aretsen [21] proposed the Factory Activity Control Model (FACT) for make-to-order and engineer-to- order manufacturing. In this architecture, technical information eg. Process plans are treated as online activity This generates open flexibilities in specifying process plans based on actual availability of workstations. FACT allows a "direct-request" between workstations, as shown in Figure 3 4 and these are considered as high priority jobs for help in unexpected situations. However, such direct requests may disturb the nominal planning for the auxiliary stations

3.2.1.4 Summary

The hierarchical architecture for manufacturing control systems is suitable for the computer integrated manufacturing systems in the interior of the factory. These are usually composed of three main layers, i.e. the master planning scheduling, short term production scheduling and the shop floor control All hierarchical control systems have a strong master/slave relationship between different modules while the system is running and assume a deterministic behaviour of the components Each module must send responses and receive orders from its superior control model, but the subordinate equipment cannot directly pass messages to each other In

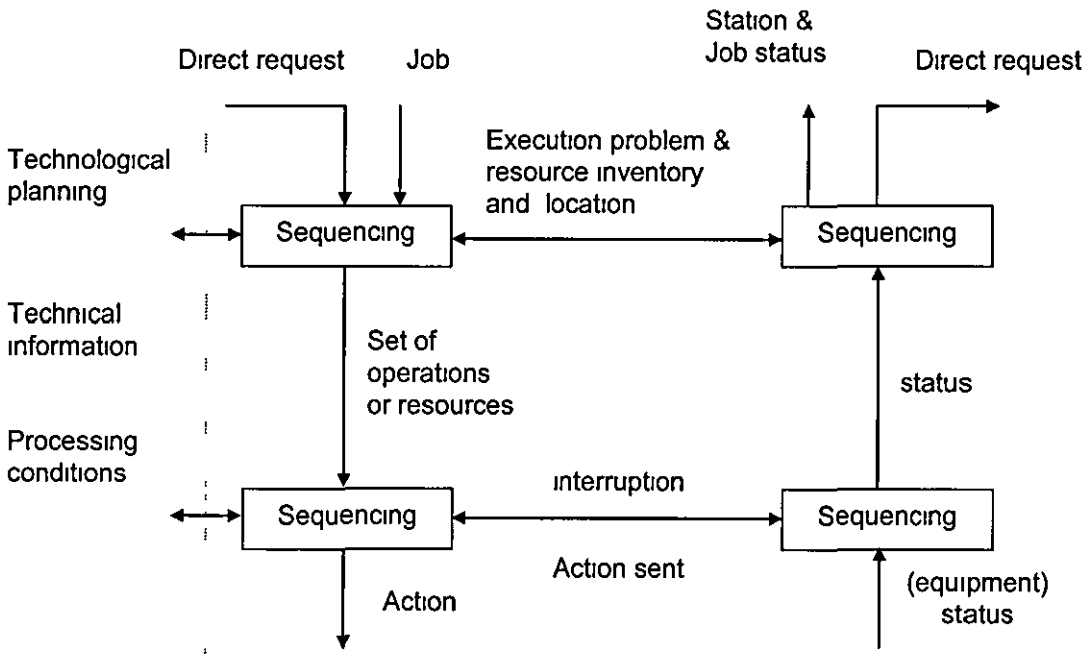


Figure 3.4: FACT station level control ([21])

manufacturing systems, material processing machines and handling equipment need to exchange many messages, so it becomes inconvenient if all these messages cannot be passed directly [22]. The advantages of hierarchical architectures include strictness, correctness and effectiveness The following implicit assumptions cause the main drawbacks of hierarchical architectures.

- Long lead times, high inventory, delays and high tardiness are considered to be the main indicators of the weaknesses of hierarchical shop floor management [23].
- Modification of the structure is a tedious task, as operations must be shutdown to enable updates to be made to higher level data structure for accurate prediction of lower levels of behaviour
- New technology and the incorporation of unforeseen modifications are almost impossible [24].
- Run time disturbances such as machine breakdowns cannot be handled easily, in some cases, the schedule becomes invalid before it is completely generated [25]
- Failure of a single part of the system may cause the global disorder
- The system is unable to respond to a chaotic environment quickly and agilely.

3.2.2 Heterarchical Control System

In an address to the 21st Conference on Decision and Control, Vamos [26] outlined the inadequacy of traditional control approaches and defined a cooperative system as :

1. A system free of any coalition
2. A system with incomplete knowledge of the overall system
3. A system operating on exchange of information

The control is distributed (with no centralized controller) and is based on negotiation between system components to handle unplanned events Hatvany [27] suggested the need for a manufacturing control model, which allows total system analysis from incomplete knowledge, having an automatic recognition and diagnosis of fault situations. The model should also incorporate automatic remedial action against all disturbances and maintain optimal operating conditions. Therefore, heterarchical architectures have been suggested, since these reject a master/slave relationship and provide full autonomy, decision making capabilities for fulfilment of one's goals, localised information and negotiation ability to each entity which represents a physical resource in the manufacturing environment Since every author applies the heterarchical control paradigm differently, typical properties of this approach have been examined here rather than attempting to examine an evolutionary trend.

The heterarchical control architecture is based on full local autonomy (distributed control) resulting in a control environment in which autonomous components co-operate in order to reach global objectives through local decision making These autonomous components are often referred to as agents, and co-operation is structured via a negotiation protocol These agents typically represent resources and/or tasks. Task allocation to resources is performed by exploiting a dynamic market mechanism, e g bid announcement, bid evaluation, etc and this results in a simple fault tolerant system, since the complete information about the system is not required. This helps in absorbing many disturbances on the shop floor and changes can be easily accommodated

Duffie and Piper [28, 29] compared the three control architectures centralized, hierarchical and heterarchical and illustrated the advantages of the heterarchical system as being reduced complexity, reduced software development costs, high modularity, high flexibility and improved fault tolerance (Hierarchical and heterarchical systems have been discussed in this chapter in detail and the centralized control system can be defined as an architecture which employs a centralized computer or controller to manage and maintain the records of all planning and information processing functions. Machines employed on a shop floor execute

the commands released from the centralized controller, and then feed back the results to the centralized controller. Traditionally, shop floor control has been performed on a centralized computer This architecture approach is most suited for completely deterministic environments) However, they also identified and listed several problems including deadlock detection and resolution and contradiction between local and global objectives Duffie and Prabhu [25] presented a look ahead cooperative scheduling algorithm to enhance global system performance and address the problems previously identified by Duffie and Piper [28, 29]

3.2.2.1 Independent Entities

Hatvany [27] proposed co-operative heterarchies as a substitute to rigid hierarchies. These systems are characterised by a strong local decision making mechanism which can easily absorb disturbances such as reconfiguration of resources, rescheduling, etc The entities in these control architectures are marked by equal opportunities for resources, mutual access and accessibility to each other, their ability to follow the protocol of the system and to function independently of each other This leads to the difficult task of establishing and codifying rules which not only permit maximum autonomy and flexibility but also optimise overall performance information and commands are exchanged via a negotiation protocol.

Heterarchical systems can easily be implemented by using agent technology [30]. An agent is considered to be an active software object, able to act on its initiative and able to interact with other agents in the system. Baker [31] enumerated a set of criteria which the multi-agent heterarchy must satisfy:

- Balanced distribution: the computational load must be equally distributed over agents.
- Physical correspondence: agents must correspond to the physical entities of the manufacturing system and as the system grows, the number of agents grows proportionally
- Scalable growth the amount of computation required for the addition of a new agent must not exceed the computational capability of the new agent.

3.2.2.2 Co-Ordination

Workloads are co-ordinated by agents using a market mechanism There exist numerous approaches (such as auction based, pricing-based, bulletin-based and game-theory-based approaches) for carrying out distributed decision making in a manufacturing system. The "Contract-net protocol" is one of the most widely acknowledged examples of such a

mechanism introduced by [32, 33] It consists of five steps: task announcement, task evaluation by contractors, contactors bidding for the tasks, winner determination and communication between agents to execute the task. Protocols establish the rules of performing the above tasks Typically bids are evaluated as a function of job or machine data ([34-36]). Tilley [37] discussed a number of protocol areas that are poorly defined in the steps of contract nets. Their inability to predict the performance of a system and the lack of global optimality were pointed out as major drawbacks and presented bidding based heterarchical system's behaviour from the computational and communication point of view and the length of time taken by contractors to interpret the task announced by the task manager was considered to be a major problem for system operations

Ozaki et.al [38] described a system of heterogeneous robotic agents which arrange themselves in collaborating teams An agent searches for collaborators whenever it cannot perform the task alone and the first agent acts as co-ordinator and commands the co-operator to carryout the task

Baker et.al [31] mentioned that not all heterarchical systems necessarily involve a dynamic market based co-operation mechanism. Kanban systems, or the distributed pull-mechanism applied by Timmermans [39] for batch production are examples of static heterarchical systems However, these systems are termed as simplified degraded versions of heterarchical systems since they lack fault tolerant characteristics Liu and Sycara [40] proposed a coordination mechanism called Constraint Partition and Coordinated Reaction(CP&CR) for job shop constraint satisfaction This system assigns each resource to a resource agent responsible for enforcing capacity constraints on the resource, and each job to a job agent responsible for enforcing temporal precedence and release-date constraints within each job. Moreover, a coordination mechanism called Anchor & Ascend is proposed for distributed constraint optimization Anchor & Ascend employs an anchor agent to conduct local optimization of its sub-solution and interacts with other agents that perform constraint satisfaction through CP&CR to achieve global optimization [41].

Nawana et al [42] described four co-ordination techniques· organisational structuring, contracting, planning, and negotiation. Faratin [43] suggested a reasoning based model for service-oriented negotiation between autonomous agents Several tactics and strategies are used to generate bids and the model is used to evaluate bids and suggest counter proposals Olivera [44] presented several co-operating policies to minimize the occurrence of conflict for an assembly robotic cell Finin et al [45] propose knowledge query and manipulation language (KQML) as a communication language between software agents It has predefined

performatives to clarify message intentions Miyasshita [46] exploited constraints for problem decomposition and co-ordination and the negotiation process iteratively builds a feasible and mutually acceptable schedule for agents. Maione and Naso[47] used a genetic adaptation technique for integration and co-ordination. Gao et al [48] proposed a stigmergic co-operation mechanism for shop floor control In their system, each work piece had a corresponding piece agent, which carries al information of the work piece and makes all decisions about the work piece The piece agent selects the manufacturing resource randomly in the light of pheromones stored in information environment. The piece agent also modifies these pheromones by awarding or penalizing to guide subsequent agents' routing

3.2.2.3 Fault- Tolerance

Fault tolerance indicates the ability of the system to continue to function, perhaps in a degraded state, despite the occurrence of system failures [49] Duffie and Piper[28] and Duffie et.al[24] established that the independent functioning of agents yields simple fault-tolerant control systems Minimization of global information in the system not only requires that agents maintain their own local data and manage local decisions but also make no assumptions about other agents. These fault-tolerant design approaches help in absorbing unplanned disturbances and have lower development costs The above architecture can absorb unplanned disturbances as only the resources that are capable of performing the particular task at that time take part in the bidding process

3.2.2.4 Global Optimization

Optimisation of global performance (throughput) or prediction for individual jobs (flow time and make span) becomes difficult due to the absence of global information

Lin and Solberg [36, 50] proposed and applied a heterarchical intelligent agent framework which treats each part and resource unit as an agent All part agents have alternative process plans and negotiate with resource agents in real time via market-like bidding mechanisms to optimise a weighted set of objectives. Each agent has the power to accept or deny a bid placed by another agent. Macchiaroli and Riemma [51] extended the above negotiation process which is iterative in nature and forces a convergence between demands and offers

Duffie and Prabhu [25] developed a local feedback algorithm for a real-time distributed scheduling system All the plans developed by the agents are evaluated in the accelerated simulation of the system and global performance measures are fedback by a central evaluator agent Each agent selects the plan which produces the best result in the simulator However

this system is unable to outperform a schedule generated by a trivial reactive centralised scheduler as the agents are not allowed to evaluate the impact of their decision on other agents

Baker [31] discussed various potential dispatching, scheduling and pull algorithms for planning in heterarchical systems and how they can be implemented Lagrange relaxation is a type of scheduling algorithm which can be distributed among agents [52] Capacity constraints can be easily handled but other scheduling constraints are tedious to satisfy

Shaw ([35, 53]) presented a distributed scheme, with several one level cells, for dynamic scheduling in a cellular manufacturing system The job assignment takes place dynamically by negotiation between cell controllers and there is no global controller. An augmented Petri-Net is used to model the bidding scheme and the performance of the proposed scheme is compared with centralised short processing time dispatching scheme via a simulation model. Usher [54] presented an experimental approach for performance analysis of a multi-agent system for job routing in job-shop settings i) under various information levels for constructing and evaluating bids, and ii) under actual real time process data for the negotiation process. Some simple but practical mechanisms are proposed and implemented

Dewan and Joshi [55], Veermani and Wang([56, 57]), Kutanoglu and Wu [58] and Veermani et al [59] presented an auction based mechanism that can be used for scheduling within a distributed decision making environment The Auction mechanism can be implemented in either a single serial processor or a distributed processor environment They have presented a viable distributed scheduling alternative to dispatching heuristics that is not only mathematically structured and leads to a predictable system performance but also outperforms the dispatching heuristics within a reasonable computation time Other examples of similar distributed scheduling mechanisms are given by Lewis et al [60], Kaihara and Fujii [61] and Tharumarajah and Wells [62] Lewis et al [60] utilized data flow model for a manufacturing control system Kaihara and Fujii [61] presented a multi-agent scheduling method and Tharumarajah and Fujii [62] used a behaviour based approach.

Ramaswamy and Joshi [63] combine a centralised off-line scheduling algorithm based upon Lagrange relaxation together with a distributed on-line control system based on a market mechanism Bongaerts et al [64] use a centralised reactive scheduler which sends the generated schedule activate order agents and resource agents They introduce schedule perturbation analyses and the use of "partial derivative" of the schedule to a local decision variable

3.2.2.5 Tuning to Market Rules

Lin and Solberg [50] stated that different price systems and price setting methods in a dynamic market mechanism result in different control strategies and system performance This tuning problem is intrinsic to price-based market mechanisms as there is loss of information The multi-dimensional decision problem needs to be scaled into a single price rate Therefore, the prediction of system performance becomes difficult for a certain price system. Simulation of the system may give an indication of the system performance under a certain price system. However, there is no guarantee that this behaviour will be replicated under a different pricing system

3.2.2.6 Typical Application

The ease of development and the reaction to disturbances are the main advantages of heterarchical systems, but they have only be successfully applied to certain types of manufacturing systems They are most successful in applications with homogenous agents, some over capacity and preferably containing capacity constraints. The reduction in time required for development of the system is also obtained because in many applications agents have an identical base code with different data However, this is only obtained when all task agent and resource agents perform identically Automatic handling of disturbances functions properly when alternative identical workstations exist and hence is more successful in homogenous systems.

Peng et al [65] presented a consortium for intelligent manufacturing planning execution (CIMPLEX), which is an agent system architecture for integrating manufacturing system planning and execution They built several specialized agents such as a parameter agent, a process rate agent, a monitoring agent and a scenario co-ordination agent to provide the necessary functionality for exception handling Two service agents, i e a server and a facilitator, were built for facilitating collaboration between agents The agents can communicate with each other directly or through a facilitator agent. The gateway agent is used for communication with the existing systems such as ERP etc.

Parunak [66] considered agents to be the extended step to object-oriented programming in software evolution and mentioned that agent technology is best suited to problems which can be characterised as modular, decentralised, changeable, ill structured and complex. Baker et al [31] and Parnuk et. al [67] described an agent architecture for shop floor control and scheduling. They attempt to provide a mechanism for direct dialog between customers and distributed manufacturing systems for mass customisation using intelligent agent technology.

Khool et al [68] presented an agent based manufacturing scheduler for multiple shop floor manufacturing systems. Two software agents, i e a manufacturing scheduling server (MSS) and a shop scheduling client system (SCSS), generate an optimal schedule for the complex manufacturing system. The MSS consists of a knowledge base with production rules for decision making and conflict resolution Wang et al [69] constructed a hybrid Petri nets model in which the continuous part describes the dynamics of the production process within manufacturing and the discrete part describes the dynamics of the ordering and delivering process between every two manufacturing systems.

3.2.2.7 Summary

The heterarchical control architecture is based on full local autonomy (distributed control) resulting in a control environment in which autonomous components co-operate in order to reach global objectives through local decision making These autonomous components are often referred to as agents, and co-operation is structured via a negotiation protocol These agents typically represent resources and/or tasks. Task allocation to resources is performed by exploiting a dynamic market mechanism, e g. bid announcement, bid evaluation, etc. this results in a simple fault tolerant system, since the complete information about the system is not required This helps in absorbing many disturbances on the shop floor and changes can be easily accommodated However, the basic assumption of independence of agents, which prohibits the use of global information can be a drawback and obstructs their widespread application in some industrial application areas The independence of agents makes central scheduling and resource planning a difficult task. The control system cannot guarantee a minimum performance level for cases outside the scope for which the rules are tuned Prediction of system performance indices for individual orders is difficult. The flow time of an order depends on the nature and status of the other orders in the system. The global system performance is sensitive to the market rules. The advantage of automatically handling machine breakdowns is only appropriate when alternative machines are available However, since most publications covering this control strategy consider process planning as giving static information, the use of alternative machines actually means the use of identical machines. Therefore this strategy has been most successful in conditions which have homogenous sets of resources eg distributed computing applications, or mobile robot applications [70] A knowledge pool populated with knowledge about various scenarios and alternatives can provide better information about the system and help the decision makers(agents) in making optimal decisions

3.2.3 Recent Control Paradigms

Hierarchical and heterarchical control systems have some desirable characteristics but also have limitations. Several attempts have been made to capture the positive aspects of both These have mostly been focused on the simple goal of enabling the manufacturing system to efficiently survive and adapt to the ever changing manufacturing environment Parunak [71] developed an agent based manufacturing system called 'yet another manufacturing systems' (YAMS). The architecture consisted of a static hierarchical structure of the global scheduler, workstations, and workcells. The Global scheduler generates a course schedule for the entire factory and the detailed schedule is generated by the real time negotiation between the successive layers of the manufacturing system An agent is allowed to negotiate with its sibling agents as well as with parent and children agents even though a hierarchical structure is used

Butler and Ohtsubo [72] described an architecture for distributed dynamic manufacturing scheduling (ADDYMS) having several levels of work cells A work cell has a site agent and may also have sub-work cells and act as a physical division of resources Tasks are allocated to a sub-work cell by negotiation and by using a database to store information about its capabilities, resource list, knowledge of assigned operations and their state and the addresses of other agents for communication based on a heuristic

Tawegoum et al [73] presented a hybrid control architecture for a flexible manufacturing system The upper level scheduler readjusts the schedule if the sub-level cannot meet the schedule due to unplanned events. Brennan et al [74] presented a hybrid control architecture suitable for constantly changing manufacturing environments and they also discussed the concepts and reference architecture of a partial dynamic hierarchy.

Ou-Yang and Lin [75] presented a hybrid control model using a bidding method for job dispatching. Each cell controller submits a bid depending upon the processing cost. Inventory and shortage cost and the shop controller selects a bid based on the submitted price and utilization level Each cell has a fixed routing and there is no concept of a part agent in the model. Overmars and Tonich [76] suggested a hybrid flexible numerical control (FNC) architecture where an intelligent controller is attached to each servo axis. These controllers respond to the commands from a hierarchical host scheduler and co-operate with each other in a heterarchical manner

Ottaway and Burns ([77, 78]) described an adaptive production control system (APCS), where the transition between a heterarchical and hierarchical control system occurs dynamically depending on the system work load It consists of job, resource, supervisory agent having coordination, production and interface knowledge along with an inference engine When the utilisation of a resource goes below a certain level, the corresponding agent requests a supervisor agent which has the jurisdiction over the concern resource and hence a hierarchy is introduced dynamically. The bidding price calculation mechanism of the simulated non-hierarchical control system in their model, does not lead to a uniform resource utilisation. One machine has a utilisation level of 97 17% whilst the other has a utilization of only 17 71%.

Mathurana [79] and Mathurana et.al [80] proposed MetaMorph, a multi-agent architecture for a distributed manufacturing system They suggested two type of agents a resource agent for physical resources and a mediator agent for co-ordination. The mediator agents use brokering and recruiting mechanisms for co-operation, and act as system co-ordinator by enabling co-operation between the intelligent agents and by helping them to find other agents. Individual resource agents are registered with the mediator agent. Virtual clusters or organisations of intelligent agents are created and disbanded dynamically. Selective communication and agent cloning mechanisms are used to reduce communication overload. A learning mechanism is proposed for co-ordination and presented in the context of capacity planning They use learning from the future by simulation as well as from history. Kernel Query and Manipulation Language (KQML) protocol is used as communication standard in MetaMorph They implemented a prototype and tested it on two shop floors and for three products. Shen and Noorie [81] have extended the MetaMorph architecture to integrate enterprise activities with those of suppliers, partners and customers in their MetaMorph II project. They use a hierarchical mediator and bidding mechanism for the cooperative negotiation among resource agents.

Brussel et al [70], Valckenaers et al [82], Sousa and Ramos [83] and Wyns [84] described a product-resource-order-staff architecture (PROSA) for a holonic manufacturing system (HMS). The architecture tends to achieve stability with disturbances, adaptability, flexibility with change and efficient use of resources. The architecture consists of three basic holons: order, product and resource holons. Each of these basic holons is responsible for logistics, technological planning and determination of resource capabilities, respectively A staff holon assists the other basic holons and is provided with a centralised algorithm and provides a hierarchical control behaviour to the system which helps in enhancing the global performance of the system. The system hierarchies can be formed by holons, and the aggregate holons can be created dynamically by the self-organisation of interacting holons, or by initial system

design. A resource holon may belong to several holarchies at several different levels Wyns [84] also discussed the PROSA application framework and a deadlock handling mechanism Valckenaers et al [82] presented an overall system design and basic principle of software development of a holonic manufacturing system. Bongaaerts et al [85] described a reactive scheduler for HMS, which tends to determine the effect of local decisions on global performance by partial derivative of the global performance to the local decision parameter Brusel et al [70] presents a method of identifying holons and holarchies Sousa and Ramos [83] presented a contract net protocol for a dynamic scheduling holon in manufacturing orders.

Bussmann [86] compares the concept of a holonic manufacturing system with an agent based manufacturing system He mentioned that both approaches recognize that a manufacturing system consists of cooperating autonomous manufacturing units and he concluded that HMS concerns the overall structure of the manufacturing process, whilst agent based systems concentrate on the design of information processing in a control system He also suggested that agent technology could be utilized to design and implement the information processing of a holon. Busssmann and McFarlane [87] described the control system for a future manufacturing system as a decentralised/resource based control architecture, providing generalised and flexible control interaction, reactive, proactive and self-organising control and suggested that HMS can support the requirements since it has properties of autonomy, cooperation, self-organisation, and reconfigurability. McFarlane and Bussmann [88] review the use of a holonic manufacturing control system in production planning and control.

Gou et al [89] and Luh and Hoitomt [90] presented a lagrangian relaxation method for holonic scheduling. The model consists of two levels factory and cell level. The factory level has product, part, cell, factory coordinator and factory holons. The cell level has part, machine-type, cell coordinator and cell holons The part precedence constraints of the factory level and machine capacity constraints of the cell level are relaxed with Lagrangian multipliers The scheduling problems are decomposed to cell and part level sub-problems and the later is solved by dynamic programming. The factory level and cell level problems are solved by the conjugate subgradient method based on the sub-problem's objective function values The factory and cell coordinator holons generate appropriate coordination information to guide the schedule quality. The part precedence and machine capacity constraints are relaxed and the generated solution may not be feasible.

Gibels [91] proposed EtoPlan (Engineer-to-order Planning), to integrate the design, as well as, process and production planning tasks in manufacture-to-order environments Three generic information structures for products, resources and orders are presented for integration of the

three planning tasks and an evolution based control model is proposed. The temporary planning hierarchies of applicability groups (AGs) can be dynamically created or deleted in their model The AG controller has four functions planning, dispatching, monitoring and diagnostics. Each AG controller can interact with its parent, children and siblings within the same order while directly interacting with other agents via resources. The lower level AGs make the detailed operational plan autonomously within the boundary constraints set by scheduling group An aggregate order planning method is also presented for higher level integration of macro process planning and resource loading.

Okino ([92, 93] incomplete reference) introduced Bionic manufacturing systems inspired by a biological metaphor The aim of this project was to design a system capable of self-organising and automatically adapting to changing needs, similar to biological systems. Ueda ([94, 95]) and Ueda and Ohkura [96] described the concept of Genetic Manufacturing systems which mimic the functioning of the DNA found in genes. The main emphasis was laid on the diversification of products and the adaptability of the dynamic manufacturing environment. Iwata and Onosato [97] presented a Random Manufacturing system having a heterarchical control algorithm based on four concepts.

(i) The machine takes autonomous decisions
(ii) Machine grouping is dynamic
(iii) Orders are communicated via a blackboard
(iv) Shop floor control is exerted by rewards and penalties

Roy et al [98] proposed "SYROCO", a system based on a two fold hybrid multi-agent platform Control is hierarchically distributed and the decision making is centralised. Schedules are partially modified during the process. Peeters et.al [99] proposed an approach for better handling of reconfiguration in production environments and better response in case of disturbances. They exploited the behaviour of ant colonies for co-ordination and suggested a pheromone based control scheme. The chief feature of this architecture is a layered approach for decision making The advantage of this approach is its capability to handle dynamic situations and automatic guidance towards an optimized solution.

Chen et.al. [100] presented an integrated information system for use in shopfloor controlling systems for knitting parts, producing goods and distributing them, to enhance the performance of the build-to-order/configuration-to-order production system Several information technology devices were adopted to support the relevant logistics Wullink et.al [101] focused on developing planning methods that operate at the tactical level of manufacturing planning

and control. They used a Mixed Integer Linear Programming model to minimise the expected cost on resource loading under uncertainty

RapidCIM is an intelligent approach to the fast generation of shop floor control code ([102-108]). It attempts to build data driven shop floor control systems and proposes the following development methodology for CIM controllers:

- A model of system is developed
- The model is then run in a simulation system to optimise decision logic
- Decision logic is automatically translated to generate the code for the shop floor control system.

Shen et al [109] presented an intelligent shop floor control system based on the internet, web and agent technology It focussed on the implementation of distributed intelligence in the manufacturing shop floor to work together as a whole rather than as a disjoint set. Lima et al [110] proposed an agent based production planning and control system that can be dynamically adaptable to local and distributed utilization of production resources and materials Shin and Cho [4] proposed a rapid development methodology that fulfils given characteristics of a shop floor through a formal model-based control software specification The formal models were represented in XML format to make them neutral to any model-building or development tools.

3.2.3.1 Summary

Researchers have attempted to overcome some of the problems of hierarchical and heterarchical by combining features from both types of frameworks. They have developed hierarchical structures to enhance global performance with coordination between agents. While ([70, 79-84, 91, 97]) adopt dynamic hierarchical structures, other researchers use static hierarchical framework in their modelling frameworks Some researchers, for example, ([73, 77, 78]) chose dynamic introduction of a supervisor or high level controllers when a low level controller is unable to meet the forecast Others used static introduction of supervisor controllers Most of the researchers have used negotiation mechanism for real time task allocation, and the decision in these framework is made by the lower level agent by a negotiation mechanism which may result in a globally inferior decision and deadlocks

3.3 Manufacturing Control Model

Shop floor control (SFC) is a complex process which lies at the heart of operations for manufacturing companies, and primarily involves job scheduling, progress monitoring, status reporting, and corrective actions [19, 111] SFC has to rapidly reflect the current system status to allow job processing to be controlled in a real time mode. However, the manufacturing system behaviour, which is an accumulation of status in time, is highly dependent on the control system architecture, control function classification, and allocation of controls to different control levels (i e control allocation). All this has had a great effect on the way production facilities are designed and managed with large changes in manufacturing systems, in levels of automation and in the application of computer based technology at all levels of the company The earlier sections of this chapter have presented a literature survey of different architectures for SFC. This section reviews modelling techniques for selecting control strategy.

Performance of the manufacturing system relies heavily on the control strategies adopted. The modelling of manufacturing control can be classified into three different groups

1. **Mathematical approaches:** Multi-dimensional spaces are searched for the 'best' solutions to provide optimal results for given decision criteria As long as the decision criteria are weighted correctly the results will be valid. Unfortunately the assigning of weightings, such as cost or set-up time is often arbitrary and relies on value judgements, leading to results that are often less than optimal. Another problem with this approach to scheduling is that the computational efforts can be vast due to reasons of combinatorial explosion Duffie and Prabhu[25] and Miyasshita [46] exploited constraint to generate the solution Shaw [53], Park et al[112] and Wang et.al [69] used Petri-nets for problem solving. Gou et al [52, 89], Ramaswamy and Joshi [63], Bongaerts et al [64] and Luh and Hoitomt[90] used lagrangian relaxation method for modelling SFC. Bongaerts et al [85] used partial derivatives for modelling SFC. Willnk et al [101] presented mixed integer programming model for SFC. Simulation has also been widely used for decision making in SFC (Roy et al [98], Joshi et al [102], Smith et al [104], Smith and Joshi [105], Manuel et al [106], Son and Wysk [113] , Son et al [107], Qiu et al [108], Monch et al [114], Rogers and Gordon [115] and Lastra and Colombo [116]). Garbot et al [117] used fuzzy logic and theory of possibility for SFC modelling

2 **Heuristic approach** For most manufacturing systems with large numbers of machines and many jobs with various routings competing for the various resources an algorithmic solution to the manufacturing control problem is not possible. In these instances heuristics or 'rules of thumb' are often used, and these evolve over time through trial and error and are based on past experience of successful decisions For large problems 'the best' solution generally cannot be found within real-world time constraints, and a heuristic approach is better than a random solution. Baker [30], Lin and Solberg [36, 50], Tilley [37], Macchiaroli and Riemma [51] and Shen et al [109] used market based pricing heuristic for modelling SFC. Smith [32], Smith and Davis [33], Shaw [35] and Lima et al [110] applied contract-net heuristic methodology for SFC modelling Ou-Yang and Lin [75], Ottaway and Burns [77, 78], Dewan and Joshi [55], Veeramani and Wang [56, 57], Veermani et al [59], Chan and Zhang [118] and Heragu et al [119] used auction mechanisms for decision making in SFC modelling. Evolutionary techniques have also been exploited for decision making in SFC (Maione and Naso [47],Giebels [91], Ueda [94, 95] Gao et.al [48]). Lewis et al [60] used FIFO for decision making Morton and Pentico [120] used SPT for decision making Heuristic solutions can be generated in less time but are myopic in nature (Gao et al [48]) These methods cannot independently consider real time information and are unable to consider parallel and alternative process plans

3 **Knowledge based system:** The use of expert systems within control can be seen as an extension of the use of heuristics, where the selection of the rules to apply are suggested by the Expert System based on the encoding of an expert's domain specific knowledge This allows the non-expert to apply the heuristics as the expert scheduler would By eliciting the expertise of certain key individuals, such as for example a section foreman, and encoding this expertise within a set of rules, the expertise can be called upon repeatedly and reliably. The expert systems may often act as expert control decision making assistant rather than a stand alone 'expert' decision maker. The systems may suggest search heuristics to the scheduler / operator in certain conditions, then the system would carry out the heuristic search with possibly an algorithmic base to start off Metaxiotis et.al [121] presented an survey of expert systems. Iwamura et al [122], Sun et al [123], Shuie and Guh [3], Shen and Noorie [124], Patriotta [125], Ozbayrak and Bell [7], Eberts and Nof [126], Khool et al [68] used a knowledge based approach for SFC modelling The knowledge base consisted of knowledge about the different rules and actions that are to be taken in different scenarios. These system have been successful to a certain extent, as they help shop floor managers to consider various alternatives quickly The major disadvantage of

knowledge based systems is that they do not scale up well as they become disordered They have poor conflict resolution capabilities and inherit the disadvantages of heuristic approaches and updating the knowledge base is a tedious task

The shop floor must have a module that is in charge of data collection, classification, management, analysis and message passing on the shop floor It must cooperate with the control system to execute production activities, deliver high-level decisions and monitor the production status to accomplish the required job processing on time.

3.4 Conclusion

A manufacturing control reference architecture specifies a generic solution for manufacturing control applications, using models to represent system structure, system components and their responsibilities, dependencies, interfaces, design rules etc The assessment of control shows that the basic assumptions of an architectural paradigm lead to constraints being built into the control system, affecting the structure, control algorithms and genericity of the architecture. Hence, when the initial assumptions do not hold, the architecture does not deliver the expected result.

Existing modelling frameworks for manufacturing system control can be classified into hierarchical, heterarchical and recent control frameworks The hierarchical framework assumes a master/slave relationship between higher and lower levels of control. They ignore uncertainty and complexity of the real world system and do not provide accurate models or react properly in modern manufacturing environments. The heterarchical framework focuses on interactions between unit controllers to allow system flexibility, and lack predictability and global perspective Recent frameworks include features from both hierarchical and heterarchical frameworks, to allow direct interactions amongst the lower level controllers as well as between higher and lower controllers. Analysis of existing reference architectures have revealed some important shortcomings, and to avoid these, the control architecture should address the following issues

Dynamic structure: The manufacturing systems are forced to work in an ever changing environment due to instantaneous market information and decreasing product life cycle. Heterarchical systems have the property of modifying, adding or removing resources when the system is running. However, dynamic structural changes can also be introduced in hierarchical systems by releasing relationships so that the dependencies change between the layers.

Decoupling of structural aspects from the control algorithm: Both hierarchical and heterarchical control architectures directly resemble the structure of control algorithms used

for planning and resource allocation Decoupling can be made possible by defining a dynamic hierarchical structure in which prediction based control as well as dynamic mechanisms can be applied.

Reactive scheduling and process planning: All evolutionary trends towards hierarchical architectures stress the importance of reactive scheduling as a means of reducing the impact of disturbances and use reactive process planning in decision making. This allows an online process plan to be generated to cope with disturbances based on the current status of the system.

Generic applicability The new architecture should be as generic as possible by assuming as little as possible about the underlying manufacturing process

Adaptability. The manufacturing system must react quickly to changes in the environment and the controller must learn and adapt to its environment in real time The system must be able to monitor and identify changes to its environment in real time, present relevant information to managers to make timely informed decisions and have an inbuilt learning mechanism.

Efficient knowledge management and discovery: Organizations are becoming increasingly complex with emphasis on decentralized decision making Knowledge is a organizational asset that enables sustainable competitive advantage in hypercompetitive environments The manufacturing system must have an information system that facilitates generation, sharing and integration of knowledge for effective and efficient decision making.

The main area addressed in this work has been to provide a methodology for knowledge discovery from data recorded during different operations in the product life cycle. Data mining techniques have been explored to utilise their ability to generate knowledge in different contexts at the shop floor and also a methodology has been provided for incorporating data mining techniques into the shop floor control system Data mining techniques also provide a methodology for a continuous learning process and help in the discovery and generation of new knowledge The resulting knowledge can then be used to update the knowledge base of the system. Agent technology has been used in the proposed architecture as it is considered to be the most promising architecture for the next generation of shop floor control systems to meet the requirements of adaptability, robustness, etc.

Chapter 4

Agent Technology in Manufacturing: A Review

4.1 Introduction

For many years, manufacturing researchers have looked to computing and information technology to overcome the current challenges in manufacturing and solutions have been sought by using knowledge-based systems and artificial intelligence. Agent technology is derived from distributed artificial intelligence (DAI) and originated from Carl Hewitt's DAI Actor Model [127], which states that "An actor is a computational agent which has a mail address and a behaviour. Actors communicate by message passing and carry out their actions concurrently". However, it is only during the last 10 to 15 years that agent technology has established a role in manufacturing research This chapter shows the application of agent technology in manufacturing control and analyses the different approaches taken

4.2 Definitions of Agents

Before the research into agent technology within manufacturing can be thoroughly reviewed and evaluated, it is important to understand what is meant by the term Agent. There is still little agreement on a single definition of this term and in fact, Franklin and Graesser [128] argue whether agents actually exist or whether they are simply programs Among the wide number of definitions for agents, the following provide a consensus of views

Agents have individual internal states and goals and they act in such a manner as to meet their goals on behalf of their user. A key element of their autonomy is their proactiveness, i e their ability to 'take the initiative' rather than acting simply in response to their environment [129]

An entity that resides in environments where it interprets data that reflect events in the environment and executes commands that produce effects in the environment An agent can be purely software or hardware. In the latter case a considerable amount of software is needed to make the hardware agent [130]

An Agent is a computer system situated in some environment and that is capable of autonomous action in this environment in order to meet its design objectives[131]

An autonomous and interactive unit in complex systems with the aim of process optimisation and stabilisation, intelligence and the ability of co-operation and co-ordination [132].

The term agent can be used to denote a software-based computer system that enjoys the following properties([129, 133])

1. Autonomy· Agents should be proactive, goal directed and act on their own (i e. exhibit self starting behaviour) or perform tasks on some user's behalf. Effectiveness of achieving goals is an important property of an agent.
2. Co-operative Agents should co-operate with other agents to achieve a common goal
3. Trustful the agents should be reliable when exerting their autonomy in performing the task designated by humans
4. Reactive Agents should perceive their environment and have the ability to learn and improve their functionality with experience in order to respond in a timely fashion
5. Flexible agents should be flexible in terms of system configuration and task delegation. They should be able to join and participate in the community at any time.
6 Interactive· Agents should be able to communicate and interoperate efficiently with humans, other systems and information sources.
7. Mobility: agent should be able to travel through computer networks.

Agents are seldom (if ever) found alone and a multi-agent system (MAS) is a finite set of bounded-rational, individually operating agents that can co-ordinate their actions through cooperation and competition in an environment determined by rules Agents are generally provided with sensors, limited memory, computational capabilities and effectors MASs provide rapid responses and dynamic reconfigurable structures to facilitate flexible and efficient use of manufacturing resources in rapidly changing environments An agent controls one or many resources, including, for example, humans, computers, machines, robots, tools etc., which are integrated by an internal communications network called an intranet. An inter-

networking communication bus enables agents to interact with each other using a predefined protocol The protocol aims to distribute tasks quickly and effectively between the different agents so that the tasks are completed smoothly and efficiently. The agent is equipped with a master monitor unit that keeps a record of the resources and posts their status in a database. This sharpens the instinct and impulsiveness of the agent. The instinct of an agent makes it proactive and this enables it to anticipate necessary steps and avoid functional failure The impulsiveness of an agent determines how it responds to an external stimulus from the environment in which it is functioning

4.3 Agents in Manufacturing Control

Manufacturing systems commonly evolve by modifying old systems and adding on new systems. This can result in a large network of independent systems forced to work together with make shift bridges The importance and challenges of shop floor control were discussed in chapter 3 and the review of recent literature on this topic showed that many researchers are looking to MASs for more effective solutions. MASs are distributed and work autonomously, they support reactivity, and are more robust than centralised systems against both local and global failures. Modular hardware can be developed, and their software is highly reactive to scheduling policies. Agents can also react faster to local changes than a centralised system can, and they have the ability to cooperate to define a globally feasible schedule. The application of multi agents leads to dynamic scheduling systems that are emergent rather than planned, and concurrent rather than sequential. The core issue in multi-agent organisational design problems is the definition of the agent roles in the organisation, as this is dependent on the agent encapsulation. Many different encapsulation approaches are possible, but most fall into two categories (1) a function oriented approach, where agents are used to encapsulate some function such as task decomposition, activity coordination, conflict detection and resolution; (2) a physical entity oriented approach, where agents represent physical entities such as managers, workers, machines and components [81] The second approach is more appropriate for modelling a manufacturing environment [134], where more physical entities are involved than in transactional oriented information system domains Most MAS research projects adopt schemes of interaction between agents based on the metaphor of negotiations in micro-economic environments In general terms, negotiation algorithms can be regarded as task dispatching strategies using fictitious currency and heuristic pricing policies based on the current conditions of the tasks and the servers involved in the decision.

Several works have been identified in this area Lin and Solberg [135] modelled the manufacturing floor shop as a market place Tasks and resources were represented by agents.

Each task agent enters the market carrying certain currency and it bargains with each resource agent on which it can be processed Similarly, each resource agent competes with other agents to get a more valuable task Autonomous Agents for Rock Island Arsenal [136] (AARIA), was a MAS for manufacturing scheduling developed for an army manufacturing facility. Manufacturing resources were encapsulated as autonomous agents and cooperation between the agents took place through the manager agent. Ouelhadj et al. [137, 138] described a multi agent architecture for dynamic scheduling in flexible manufacturing systems where resources were represented by agents. The resource agents were responsible for scheduling the resources, and they cooperated using the contract net protocol (CNP) A multi agent scheduler, also based on the Contract-Net approach, has been suggested in Saad et al [139] in this scheduler, agents model the articles to manufacture and the machines, with manufacturing objectives to minimise resource use and cycle time and to meet the due dates. Maturana and Noorie [140] described the mediator architecture for an intelligent manufacturing system which provided a virtual organisation through virtual clustering and distributed decision making through the CNP. A single high level mediator agent dynamically created clusters of heterogeneous agents as needed, and then co-ordinated their activities A similar architecture was used in Metaphor II [141], where the cooperative negotiation among resource agents was realised by combining the mediation mechanism based on hierarchical mediators and the bidding mechanism based on CNP for generating and dynamically maintaining schedules. Vancza and Markus [142] presented an agent model based on economic concepts using markets rules and incentive mechanism and reconciled autonomy and cooperation. They integrated order processing, advance scheduling and dynamic dispatching to solve distributed production scheduling. Macchiaroli and Riemma [51] presented a negotiation process between agents in a heterarchical model. Each agent pursued its objective and the iterative negotiation process forced a convergence between demand and offer. The convergence can lead to a globally optimal schedule Usher [54] employed an agent based system to dynamically route job orders through production using a single step production reservation approach involving a modified contract net based negotiation mechanism

Sousa and Ramos [83] and Ramos and Sousa [143] proposed a holonic architecture for scheduling in manufacturing systems in which tasks and resources are represented by holons and used the CNP for scheduling/rescheduling of tasks Recently levelled commitment contracts were proposed as an extension of the CNP for increasing the economic efficiency of contracts between self interested agents in the presence of incomplete information about future events Tiwari and Mondal [144] specified the communication protocols and subsequently synthesised and clustered the individual parties into autonomous agents in accordance with the basic constraints of a holonic manufacturing system. They used a fuzzy c-means clustering

algorithm to club the parties and effectively capture the uncertainty and imprecision associated with them

Sousa et al [145] presented a prototype system (named Fabricare) for the scheduling of manufacturing orders, based on holonic manufacturing systems using extended logic programming The holons cooperated among themselves using an extension to CNP named contract net with a constraint propagation protocol. Cowling [146] presented a multi agent architecture for integrated dynamic scheduling of a hot strip mill and continuous caster Each process was assigned to an agent which independently determined an optimal dynamic schedule by considering its objective and information about other agents and the environment. Arboleda and Das [147] presented a multi agent based methodology for dynamic control of a stochastic lot scheduling problem They developed a simulation optimisation methodology using reinforcement learning and implemented it for a single server multiple product lot scheduling problem. Tripathi et al. [148] presented a multi agent architecture of intelligent agents that controls the part flow within highly automated manufacturing systems Chan et al [149] discussed a conceptual infrastructure for an information based control architecture with special emphasis on multi level coordination The architecture provides a global perspective to each of the constituent agents, so that both the hierarchical and heterchical control mechanisms can be adopted.

In other studies, the MAS approach has been used to repair a schedule (e g. [150]) or for real time control of a flexible cell [151] In this last case, the interaction between static agents modelling the product system environment (robots, machines, etc) and dynamic agents (modelling parts, tools or Numeric Control programs) allowed a production scenario to be generated. Baker [30] described a market driven contract net control architecture for advanced factory scheduling in a heterarchical architecture. Each agent controls one or more manufacturing resources and is connected with others in a network The agents negotiate autonomously with others using estimated costs and market based negotiation mechanisms Peng et al. [152] presented the consortium for intelligent integrated manufacturing planning execution (CIMPLEX) agent system architecture, which was designed for integrating planning and execution in a manufacturing system They focus on exception handling and their resolution Brun and Portioli [153] presented an agent based approach for shop floor scheduling to cope with assembly line coordination Khoo et al [154] presented an agent based manufacturing scheduler for a multiple shop floor manufacturing system Two software agents, i e. a manufacturing scheduling server (MSS) and a shop scheduling client system (SCSS) generated an optimal schedule for a complex manufacturing system The MSS consists of a knowledge base with production rules for decision making and conflict

resolution A novel algorithm termed as Distributed Probabilistic Scheduling (DPS) was proposed by Bochmann et al [155] for dynamic production scheduling DPS defines a probabilistic time window in which Resource Holons can schedule tasks received dynamically from Order Holons A part dispatching approach that utilizes intelligent agents and Evolutionary Algorithm (EA) was proposed by Maione and Naso [47, 156]. The decision was made based on multiple fuzzy criteria Peeters et al [99] proposed an agent-based manufacturing system scheduling and control approach that utilizes a coordination mechanism found in insect society Roy and Anicaux [157] presented a MAS based shop floor control system to solve dynamic production problems and adapt to changing environments The control was performed by quick reaction of the lower part agent and an efficient communication system, which allowed the upper part to react and adapt its decision immediately. Unver and Analgan [158] presented a framework for adapting agent based shop floor control system Windows DNA(Windows Distributed interNet application Architecture) platform based on contract net and sub-contraction model. Chan and Zhang [118] presented a CORBA based multi agent framework for an agile manufacturing system which integrated different activities of the shop floor in a distributed intelligent open environment. Odrey and Mejia [159] presented an agent based architecture to control a flexible manufacturing system which provided responsive and adaptive capabilities for error recovery in the control of a large scale discrete event production system Heragu et al [119] presented an intelligent agent based framework for manufacturing system control In their framework, entities and resources were modelled as an holonic structure that uses intelligent agents to function in a cooperative manner so as to accomplish both individual and system objectives Wang [160] presented a model for production control based on distributed shortest path algorithm using a synchronous protocol Frey et al. [161] identified the environmental constraints for the successful application of multi-agent systems. Monch et al [114] presented an architecture of agent based systems for semi conductor manufacturing [110].

Liu and Sycara [40] suggested a coordination mechanism requiring the definition of scheduling problems as a constraints satisfaction problem Resource-agents express the capacity constraints, whereas job agents express the precedence constraints between tasks, so that the constraints for earliest beginning and delivery dates are satisfied. A similar method is suggested by Miyashita [46] with the difference that each agent addresses its own constraint satisfaction problem separately from the others. A negotiation process then solves the conflicts Krothpalli and Deshmukh [1] proposed new inter-agent and intra-agent negotiation mechanisms for improving the performance of multi-agent or decentralised manufacturing systems. The system considers parts and machines as agents with communication capabilities The primary objective of a part is to be finished before the due date, whilst the objective of a

machine is to maximise its utilisation rate A manufacturing system modelling approach has been developed by Huang and Nof [162] The manufacturing system was designed as a society of autonomous agent networks (AAN). The system tasks were accomplished through the communication and information exchange definitions and protocols. Arbib and Rossi [163] presented a multi-agent environment for optimal resource allocation in a manufacturing system. The mathematical properties of the model were used to guarantee or approximate an optimal behaviour of the agents with respect to local and global objectives Ghiassi [164] identified that integration should create knowledge driven systems, which would result in more agile, interpretable, efficient and responsive manufacturing and production systems. Barber et al [165] described a MAS for conflict detection and resolution in plan integration based on extensions to PERT diagrams Wang et.al [166] presented a multi agent and ruler based distributed approach for scheduling in agile manufacturing system. They used simulation to generate a feasible schedule. Gorodetski et al. [167] used a multi agent system for resource allocation and scheduling within shipping logistics problems. Karageorgos et al. [168] presented an agent based approach for supporting logistics and production planning, taking into account not only production schedules but also availability and cost of logistics providers achieved through negotiation based on an extended contracting protocol. Naso and Turchiano [169] presented a MAS approach for dynamic part routing in automated manufacturing systems, where the part agent takes decisions about both imminent and subsequent operations Wu and Weng [170] presented a multi agent scheduling method to integrate job routing and sequencing in a flexible job shop environment with the objective to minimise earliness/tardiness. They utilised a heuristic algorithm for decision making. Wong et al [171] presented a multi agent manufacturing system for integration of planning and scheduling They presented an hybrid contract net protocol for negotiation in two MAS architectures (one with a supervisor agent and the other without).

Mobile agents are ideal for the development of distributed scheduling systems, which are becoming more of a necessity as supply chains grow on a global scale made possible by distributed networks such as the Internet. Trentesaux et al. [172] used agents in a decision support system for dynamic task allocation in a distributed structure for a flexible manufacturing system The agents act as entities of a manufacturing system which co-operate with other agents to achieve a global production program Ming et al [173] also used agents to determine the behaviour of a distributed manufacturing system Each agent has independent interests, values and modes of operation Xue et al. [174] presented an agent based intelligent scheduling mechanism that identified an optimal schedule which satisfied both product and process constraints Yen and Wu [175] presented a new paradigm to solve scheduling problems in a distributed and collaborative manner using internet scheduling agents. These

agents shared communication resources to resolve the problems using market based control Lim and Zhang [176] developed an agent based system that integrated dynamic process planning and production scheduling to increase the responsiveness of an adaptive manufacturing system Cavalieri et al. [177] presented a comparison of multi-agent heterarchical architectures and pointed out their limitations, peculiarities and applicability. Shen et al [178] reviews the literature on agent based approaches for manufacturing process planning and scheduling Lastra and Colombo [116] reviewed the state of art in implementation of agent based manufacturing system and identifies the lack of engineering tools as the technological gap for the widespread industrial adoption of the paradigm.

4.4 Comparisons of the different approaches

Two distinct approaches for agent encapsulation were found in the literature: *physical decomposition* and *functional decomposition* Both approaches have distributed implementations In the physical decomposition approach, agents are used to represent physical entities such as operators, machines, tools, products, parts and operations with explicit relationship, while in the functional decomposition approach, there are no explicit relationships between agents and the physical entities. The latter approach uses agents to encapsulate modules assigned to functions such as order acquisition, planning, scheduling, material handling, transport management, and product distribution [179]

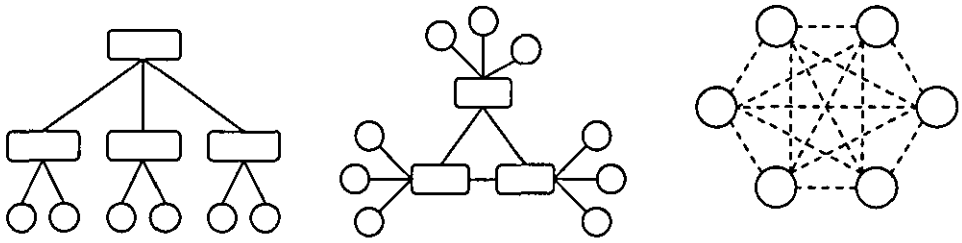


Figure 4 1: System architecture for agent-based system [179]

Agent system architectures provide the framework within which agents are designed and constructed. Three types of architectures exist, i e. hierarchical, federal and autonomous, as shown in figure 4.1 [179]

An architecture is said to be hierarchical if there are a number of physically distributed, semi–automated units, each with some control over local resources and with different information requirements. As the hierarchical architecture suffers problems due to its centralised nature,

the federal multi–agent architecture is increasingly considered to be a compromise solution for industrial agent based applications In the federal architecture, communication between agents take place through facilitators. A federated agent–based system stores all data in local databases and handles updates and changes through message passing An autonomous agent, as the name says, is not managed by other software agents or human It can communicate and interact directly with other agents and systems and has knowledge about other agents and their environments These types of agents have their own goals and motivations.

4.5 Conclusions

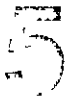
This Chapter presents a review of the state–of–the–art literature on agent technology, which has recently gained popularity in manufacturing control due to its distributed and autonomous nature Agent-based approaches offer many advantages for manufacturing control modularity, reconfigurability, adaptability, scalability, upgradeability, and robustness (including fault recovery). The results achieved so far in the agent research community provide excellent motivation for further development of solutions in this area. Moreover, at present, there are no other ways to solve these complex problems However, whether the potential advantages of agent-based approaches can actually be realized in industrial systems will depend on the selection of a suitable system architecture for agent organization and an appropriate approach for agent encapsulation, on the design and implementation of effective mechanisms and protocols for communication, cooperation, coordination, and negotiation; and on the design and implementation of advanced internal architectures and efficient decision schemes of individual agents Learning from experience and predicting the future by analysis of database information is a promising area for providing efficient decision schemes in agent systems.

The paradigm of agent-based computation has revolutionized the building of intelligent and decentralized systems. This helps in meeting the technological requirements in manufacturing domains where problems of uncertainty and temporal dynamics, information sharing and distributed operation, or coordination and cooperation of autonomous entities had to be tackled Intelligent agents help in building large and complex systems by leveraging the strengths of object-oriented, peer-to-peer and service oriented architectures while providing a process-centric design paradigm. The key benefits of agent technology are:

- Dynamic Planning – The ability to develop distributed workflows using rules and domain knowledge that is appropriate to the current situation. This benefit allows enterprises to create more accurate and appropriate plans and to react more quickly and appropriately when conditions change.

- Adaptation and Evolution – The ability to allow significant business changes to be implemented quickly and dynamically by actual users who can easily manage adjustments to the business rules or policies - without engaging consultants to significantly alter their systems. This benefit allows enterprises to be agile and adaptive as conditions change,
- Collaborative Execution – The ability to easily share information and coordinate changes with your partners, suppliers and customers

Chapter 5



Data Mining: Processes and Algorithms

5.1 Introduction

The efficiency and effectiveness of decision making in shop floor control is dependent on the accuracy and relevancy of knowledge that underpins the rules and heuristics upon which the decisions are based Several writers ([3, 4]) have recommended that shop floor control systems require the capability of ongoing learning The reviews in Chapter 5 and Appendix I have shown that data mining is valuable in analysing, classifying and understanding a variety of complex manufacturing processes. This research therefore proposes that existing MAS for shop floor control should also be linked with data mining technology to provide ongoing learning mechanisms for shop floor control However this is not a straightforward undertaking due to the wide-ranging variety and complexity of data mining technology This chapter will therefore introduce several necessary processes and algorithms for data mining.

A particular challenge in data mining in manufacturing is how to determine the best (most effective and efficient) type of algorithm in any particular context. This is important as manufacturing industry is unlikely to whole-heartedly adopt data mining principles until data mining can be shown to frequently produce cost effective, high quality solutions. This chapter therefore also addresses the challenge of appropriate algorithm selection Data mining can be defined as the process of exploration and analysis by automatic or semi-automatic methods of databases to extract meaningful information. Data mining tools and techniques are used to try to find patterns in the data and infer rules from them to guide decision making and forecast the effects of decisions

This chapter provides a review of data mining process models, stages and algorithms. It also addresses the challenges of management of data mining projects and the selection of appropriate tools and algorithms, and discusses some of the most widely used data mining tools The tools discussed are not designed specifically for any particular problems and can therefore be tried for any kind of knowledge exploration process on existing (rather than purposely collected) data. Comments and observations are also made about data mining methods and these are based on the author's personal experiences of applying data mining approaches to many manufacturing based case studies

5.2 Knowledge Discovery in Databases (KDD) and Data Mining

KDD is the process of identifying valid, novel, potentially useful and ultimately understandable patterns and/or models in data ([180]) Data mining is a step in the knowledge discovery process consisting of selecting and applying particular data mining algorithms that, under some acceptable computational efficiency limitations, find patterns or models in data [181]. Methods and algorithms used in data mining come from many computing and mathematical fields, including statistics, database management, machine learning and artificial intelligence. During this research it has been considered important to study and understand the range of techniques that form data mining technology because any of the different approaches or algorithms may provide useful functionality to enable learning within different implementations of the proposed shop floor control system.

It is often possible to identify patterns of some form within a set of data, however, not all patterns are understandable or interesting. Patterns here refers to the relationships that might exist between different fields (variables and output) within the database. Many such relationships exist in manufacturing environments because of the physical relationships between the process variables e g temperature, pressure, speed, etc , and the product quality measurements like length, thickness, height, etc. The "degree of interest" of the discovered knowledge can be characterized by several criteria Evidence indicates the significance of a finding can also be measured by a statistical criterion. Redundancy amounts to there being similarity of a finding with respect to other findings and measures to what degree any particular finding follows from another one. Usefulness relates a finding to the goal of the users Novelty includes the deviation from prior knowledge of the user or system Simplicity refers to the syntactical complexity of the presentation of a finding, and how generality is determined An important notion, called interestingness, is usually taken as an overall measure

of pattern value, combining validity, novelty, usefulness and simplicity. Interestingness functions can be explicitly defined or can be manifested implicitly through an ordering placed by the KDD system on the discovered patterns or models

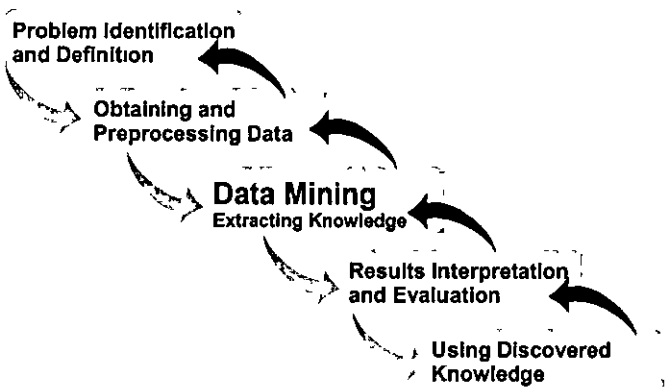


Figure 5.1: The KDD process

The process of knowledge discovery inherently consists of several steps, as shown in figure 5 1 ([182-186]) The first step is to understanding the application domain and formulating the problem This step is an important step for extracting knowledge and for selecting data mining algorithms in step 3. The second step is to collect and pre-processes the data This step includes the selection of data sources, removal of noises and outliers, treatment of missing value, transformation and reduction of data, etc. The third step is to apply the selected data mining algorithm(s) (common types of data mining algorithms are discussed in section 5.5) to extract the knowledge from the database The fourth step is to interpret the discovered knowledge. In this step the results obtained in step 3 are translated into a form that is appropriate for the application domain. The final step is to utilise the discovered knowledge within the application domain

KDD consists of the following tasks as shown is figure 5 2

- Develop an understanding of the application domain: relevant prior knowledge, goals of end users, etc
- Create or select a target data set select a data set, or focus on a subset of variables for the data mining task that is to be performed
- Supply missing value replace any missing values with the a suitable value, e g. the mean, median or one with maximum frequency. The decision of which values are suitable for substitution and how decisions should be taken for handling missing values will vary in different contexts

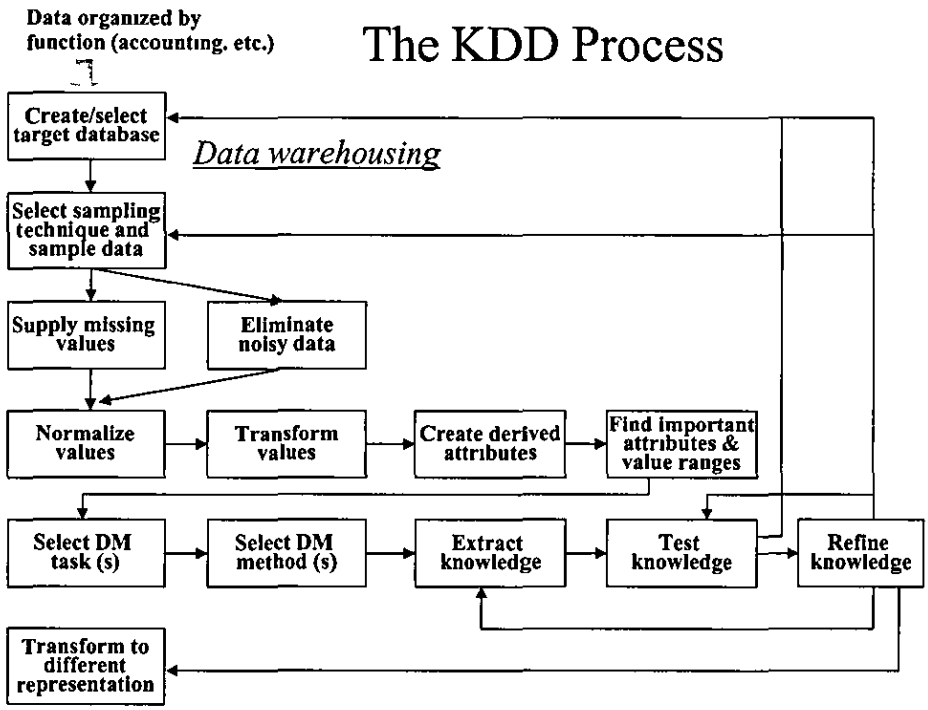


Figure 5 2. Tasks in the KDD process

- Eliminate noisy data The data values which are considered to be noise should be eliminated to improve the efficiency of the result obtained from the application of the data mining algorithms
- Normalize value· databases are normalised to reduce data duplication and possibly eliminate various kinds of logical inconsistencies that could lead to loss of integrity of the database
- Transform value data bases are transformed to different values or forms depending on the problem domain and data mining algorithms used.
- Create derived attributed· the transformed and/or normalised data values should be stored as separate variables to facilitate the application of the data mining algorithms.
- Find important attributes. different approaches are explored to determine effect of each variable and eliminate the ones with least effect
- Select Data mining Task decide the goals of data mining task e g classification, prediction, etc.

- Select data mining method: selecting method(s) to be used for searching for patterns in the data This includes deciding which models and parameters may be appropriate (e g , models for categorical data are different to models on vectors over the real numbers) and matching a particular data mining method with the overall criteria of the KDD process
- Extract knowledge: The algorithm is applied to obtained and extract the patterns from the data base
- Test knowledge. the result obtained above is tested statistically to determine its validity over the entire universal set
- Refine knowledge. the obtained knowledge should be examined and tested by domain experts to determine the validity and novelty of the obtained results.
- Transform to different representations: various graphical representation techniques are used to explain the knowledge thus generated.

It should be emphasised that not all of the above tasks are performed in all data mining applications. The tasks performed in any particular application depend on the problem domain, data type, data source and algorithms used These tasks tend to be iterative in nature and their results are progressively improved through repeated iterations that can occur in each step, or between any step and any preceding one. There are many challenges for KDD, which can limit the quality and accuracy of results, these include

Size of dataset: very large databases may be too unwieldy or time consuming to process and alternatively some algorithms cannot be applied to very small datasets. Number of fields in the dataset also increases the dimensionality of the problem A high dimensional data increases the search space for model induction in a combinatorial explosion manner.

Over-fitting: there is a possibility that when an algorithm searches for best parameters for one particular model using limited data set, it may over fit the data This leads to degradation of performance on test data

Missing and noisy data: data bases are not generally designed with data mining tasks as one of its criteria and generally have high error rates

Understandability of patterns: In many applications, it is important to make the discoveries more understandable by humans

User interaction and prior knowledge: many KDD methods and tools are not truly user interactive, has poor interface design and it is difficult to incorporate relevant information into the models in simple way.

Integration with other systems: most of KDD methods are stand alone application and may not be useful.

5.3 Data Mining Process Models

KDD works in a systematic way by applying data mining algorithms to discover any hidden information in the data Data mining relies on a set of processes rather than on a set of tools In data mining literature, different general frameworks have been proposed to serve as guidelines for how to collect and analyse the data, prepare and present the results, how to exploit the results and how to monitor the improvement The well established and developed data mining process models are

- CRISP-DM [187]
- SEMMA [188]

5.3.1 CRISP-DM

CRISP-DM (Cross-Industry Standard Process for data mining) was proposed in late 1996 by a European consortium of companies to serve as a non-proprietary standard process model for data mining CRISP-DM was a project to develop an industry-neutral and tool-neutral data mining process model that can be summarized into six basic stages,

1. Problem/Business Understanding: this step focuses on understanding the project objectives and requirement from business perspective and then converting this knowledge into data mining problems definition.
2. Data Understanding phase starts with initial data collection and proceeds with activities to get familiarise with data
3. Data Preparation: covers all the activities that are carried out to transform the raw data that can be fed into different algorithms.
4. Modelling: In this phase various algorithms are applied to determine the trends, knowledge and information
5. Evaluation: the models generated in the above steps are thoroughly evaluated to determine the accuracy of results and to check if all the steps were performed in accordance with the business objectives.
6 Deployment In this phase, the obtained model is implemented within application domain

5.3.2 SEMMA

The SEMMA (Sample, Explore, Modify, Model, Assess) data mining methodology was developed by SAS Institute Inc. as a systematic and structured guide to mining the data. This methodology is easy to implement but could require many iterations before getting the required results. In this methodology, the sample data is chosen in the beginning and therefore there is always a probability that during the assessment phase the model will fail. The main phases of SEMMA are

- Sample. in this step, appropriate data is selected for data mining task
- Explore: in this step, data is explored to find noise and outliers.
- Modify: in this step various cleaning, selection and transformation are performed on data set
- Model in this step, various data mining algorithms are applied to generate information and knowledge
- Assess in this step, the generated knowledge is evaluated for their accuracy and strength

5.3.3 Comparison of Process Models

These process models are widely used by the data mining community. CRISP-DM and SEMMA provide a step by step guide for data mining implementation. CRISP-DM is easier to use than SEMMA in a sense that it provides a detailed neutral guideline that can be used by any novice in the data mining field SEMMA has been developed as a set of functional tools for SAS's Enterprise Miner software. Therefore those who use this specific software for their tasks are more likely to adopt this methodology Secondly if the discovered relationships do not appear consistently throughout the whole database then new samples must be examined which means repeating the whole data mining process. Kurgan and Musilek [189] presented a review of data mining process models.

5.4 Data Mining Stages

Data Mining can be defined as the exploration and analysis by automatic or semi automatic means of large or small quantities of data in order to discover meaningful trends, patterns or rules [182, 183] Data Mining is a small step in the overall process of KDD In the real world of data-mining applications, more effort is generally expended preparing data than applying a data mining (or prediction) program to the data Data mining methods are quite capable of

finding valuable patterns in data and the usual straightforward approach is to apply a method to the data and then judge the value of its results based on the estimated predictive performance. However, this should not under-estimate or diminish the importance of careful data preparation. While the prediction methods may have very strong theoretical capabilities, in practice all these methods may be limited by a shortage of data relative to the unlimited space of possibilities that may be searched for. The whole process can be divided into four iterative sub-stages as shown in figure 5 3 below.

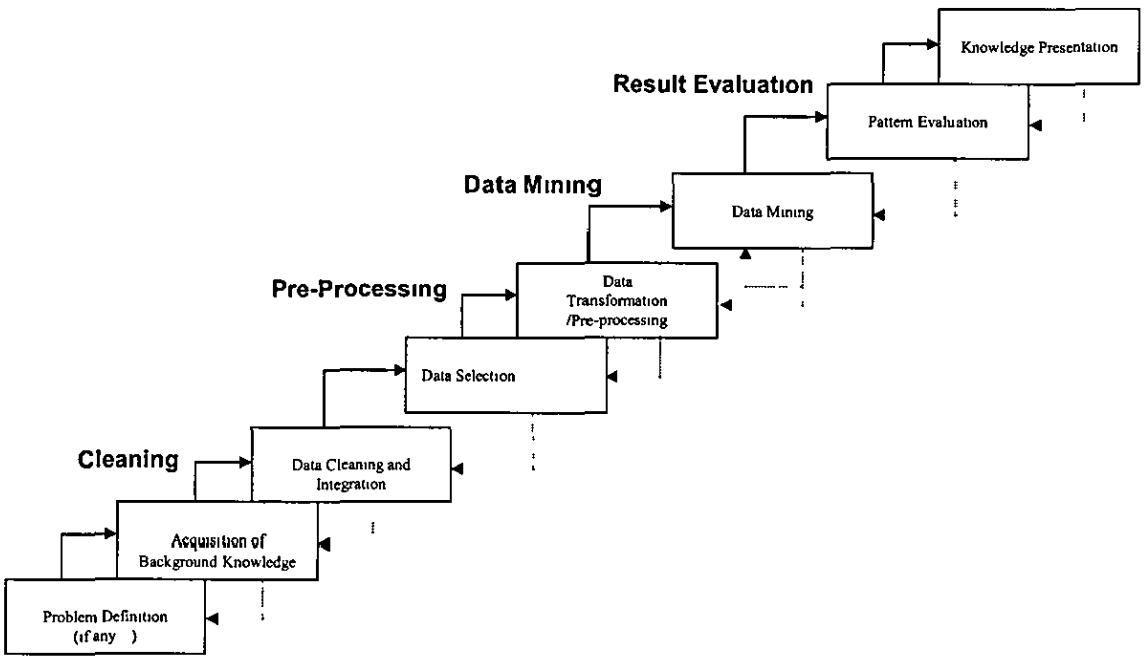


Figure 5.3: The Main Steps in Data Mining

5.4.1 Data Cleaning

Real world data tend to be incomplete, noisy and inconsistent. The data cleaning step attempts to fill in missing values, smooth out noise while identifying outliers and correct inconsistencies in the data Data Cleaning is a complex and time consuming task and requires substantial resources to be reserved for the whole data mining process. Cleaning can only be done when process and data details are understood properly. There are no defined rules for this step. It varies from project to project and the processes where cleaning is performed The sources of error can arise from the data entry operators and / or through the malfunctioning of machines Some different methods that can be used as guide lines for data cleaning are as follow·

5.4.1.1 Missing Records

If the missing values can be isolated to only a few features, the prediction program could be used to find several alternative solutions, i e. one solution using all features and other solutions that do not use the features with many expected missing values. A substitute feature might be found that approximately mimics the performance of the missing feature Some of the measures that can be taken are:

- Ignore the record
- Fill in the missing value manually
- Use a global constant to fill in the missing value
- Use the mean of the attribute values to fill the missing value
- Use the most probable value to fill the in missing value

5.4.1.2 Noise

Noise is a random error or variance in measure values. Smoothing techniques and outliers analysis can be used for noise reduction. Changes should only be made in consultation with experts in those areas or processes that generated the original data and all changes must be documented

5.4.2 Pre-Processing

Most of the time raw data (possibly from different file sources) may need to be consolidated and then converted into different formats depending on the type of the problem that generated the data and data mining algorithm selected for use. The pre-processing stage may require some intelligent thinking of alternative data transformations and definitely also requires the intended data mining algorithm to be well understood. Data transformation can involve the following tasks:

- Smoothing
- Aggregation
- Generalisation
- Normalisation
- Data Transformations
- Attribute Constructions

- Pre-processing also can also involve data reduction, which involves removing some records or features or limiting the value of a feature.

5.4.3 Data Mining

This is the stage where the data mining algorithm is implemented on the clean and transformed data [190]. This step may require the use of several algorithms in order to obtain the desired results. Different data mining algorithms are detailed in section 5.5.

5.4.4 Result Evaluation

As data mining is the discovery of understandable and useful patterns or results, all results need to be validated on the system from where the data was collected The results can be presented in the forms of graphs, tables or rules Therefore results are validated on the system and if some errors are produced then the results should be fine tuned to generate more exact and reliable results. These results can be used on the current system and stored for future reference or consultation

5.5 Data Mining Approaches

5.5.1 Statistical Methods

Statistics is the science of collecting and organising data and drawing conclusions from data sets. Descriptive statistics deals with organisation and description of the general characteristics of datasets and statistical inference describes the methodology of drawing conclusions from data In this section, emphasis is placed on the basic principle of statistical inference. Statistical data analysis is the most well established set of methodologies for data mining Historically, the first computer based analyses of data were developed with the support of statisticians Statistical methods of data analysis vary from one dimensional to multivariate data analysis and provide a variety of data mining techniques including different types of regression [191]. In some types of problem, statistics can be used to determine explicit relationships within the data However, in more complex problems this may not be possible, but statistics may still be useful for an investigative stage to better understand the data and the problem

5.5.2 Statistical Inference

The relation between data sets and the systems that they describe can be used for inductive reasoning: from observed data to improve knowledge of a (partially) unknown system Statistical inference is the main form of reasoning relevant to data analysis. The theory of statistical inference consists of those methods by which one can make inference or generalisations about a population It consists of arriving at conclusions concerning a population when it is impossible or impractical to observe the entire set of observations that make up the population These methods can be categorised into two major areas, ***estimation*** and ***test of hypothesis.***

Population *refers to the totality of observation which are of statistical interest, whether it is a group of people, events, or objects The number of observations in the population is defined as the* ***size of population*** *A subset of population is called* ***sample*** *and it describes a finite data set of n-dimensional vectors. In this report, the term* ***data sets*** *will be used as a synonym for* ***sample.***

In *estimation* a possible value or range of values is estimated for the unknown parameters in the system. The goal is to gain information from a data set in order to estimate one or more parameters belonging to the model of a real world system Statistical testing is used to decide whether a *hypothesis* concerning the value of a population characteristic should be accepted or rejected in the light of an analysis of the data set. A *statistical hypothesis* is an assertion concerning one or more populations. The truth and falsity of a statistical hypothesis can never be known with absolute certainty unless a observation on complete data is performed. This is impractical for large databases and therefore data sets are tested instead. If inconsistency is observed, then the hypothesis is rejected, whereas if consistency is observed then the hypothesis is accepted or more precisely, it is said that the data sets do not provide sufficient evidence to refute the hypothesis [191].

5.5.3 Summary Statistics

In data mining analysis, it quite informative to know about the central tendencies and data dispersions of the data sets. These describe the differences between the data sets Typical measures of central tendency include the metrics, *mean, median* and *mode*, while measures of data dispersion include *range, variance* and *standard deviation*

The most common and effective measure of the centre of data set is the mean value For a set of n numeric values $x_1, x_2, \ldots x_n$, for a given feature, the *mean* is

$$mean = \frac{1}{n} \sum_{i=1}^{n} X_i$$

These values can be associated with weight w_i (reflecting importance, frequency, significance) In those cases the weighted *mean* value is

$$mean = \frac{\sum_{i=1}^{n} w_i \quad X_i}{\sum_{i=1}^{n} w_i}$$

For skewed data sets, a better measure of the centre of data is the *median.* It is the middle value of the ordered set of feature values if the set consists of an odd number of elements and it is the average of the middle two values if the number of elements in the set is even.

$$median = \begin{cases} X_{(n+1)/2} & if \quad n \quad is \quad odd \\ (X_{n/2} + X_{(n/2)+1} & if \quad n \quad is \quad even \end{cases}$$

Another measure of the central tendency of a data set of *mode.* The mode for the set of data is the value that occurs most frequently in the set

Mean and median are characteristics of primarily numeric data sets, the mode can also be applied to categorical data, but it has to be interpreted carefully as the data sets are not ordered It is possible for the greatest frequency to correspond to several different values in the dataset and this results in more than one mode existing in the data set.

The degree to which numeric data tend to spread is called the dispersion of the data, and the most common measures of dispersion are *the standard deviation* σ and the variance σ^2 the variance of n numeric value $x_1, x_2, \ldots, x_n$ is

$$\sigma^2 = (1/(n-1)) \sum_{i=1}^{n} (x_i - mean)^2$$

The *standard deviation* σ is the square root of the *variance* σ^2. The basic properties of the *standard deviation* σ as a measure of spread are

σ measures spread about the *mean* and should be used only when the *mean* is used as the measure of the centre.

$\sigma = 0$ only when there is no spread in the data

The summary statistics help to provide a quick overview of the data set and comparing the characteristics and distributions of two or more data sets, compare different frequencies of

attributes and identify any anomalies in data Summary statistics also provide an idea of the shape of a dataset [191]

5.5.4 Predictive Regression

The prediction of continuous value can be modelled by statistical techniques called regression. The objective of regression analysis is to determine the best model that can relate the output variable to various input variables [192]. Common reasons for performing regression analysis include

1 The output is expensive to measure but the inputs are not and so a cheap prediction of output is sought

2 The values of inputs are known before the output is known and a working prediction of output is required

3. Controlling the input value, the corresponding output behaviour can be predicted

4 There exists a link between some of the inputs and the output, and the regression technique is used to identify it.

Generally linear models are currently the most frequently applied statistical techniques They are used to describe the relationship between the trend of one variable and the values taken by several other variables The relationship that fits a set of data is characterized by a prediction model called the regression equation. The most widely used form of the regression model is the general linear model formally written as

$$Y = \alpha + \beta_1.X_1 + \beta_2 X_2 + \beta_3 X_3 + ... + \beta_n X_n$$

Where α , β_1, β_2, ... β_n are regression coefficients determined by the method of least squares of errors about the regression line

The values of β's are relatively easy to find in problems with several hundred training samples The number of samples in real data mining problems may be up to several millions In these situations because of the extreme dimensions of the matrixes and the exponentially increased complexity of the algorithm, it is necessary to find modifications and/or approximations in the algorithm, or to use totally different regression methods.

There is a large class of regression problems, initially nonlinear, that can be converted into the form of a general linear model For example, a polynomial relation such as

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_3 + \beta_4 X_2 X_3$$

can be converted into linear form by setting new variables $X_4 = X_1 . X_3$ and $X_5 = X_2 . X_3$ Also, polynomial regression can be modelled by adding polynomial terms to the basic linear model. For example, the cubic polynomial curve

$$Y = \alpha + \beta_1 . X + \beta_2 X^2 + \beta_3 . X^3$$

can be linearized by applying a transformation to the predictor variable ($X_1 = X$, $X_2 = X^2$, $X_3 = X^3$) so that the problem is transformed into a multiple regression problem, which can be solved by the method of least squares.

The major effort for a user, in applying multiple–regression technology lies in identifying the relevant independent variables from an initial set and in selecting the regression model using only relevant variables Two general approaches are common for this task

1. **Sequential search approach:** which consists primarily of building a regression model with an initial set of variables and then selectively adding or deleting variables until some overall criterion is satisfied or optimized
2. **Combinatorial approach:** which is, in essence, a brute force approach, where the search is performed across all possible combinations of independent variables to determine the best regression model

Irrespective of whether the sequential or combinatorial approach is used, the maximum benefit to model building occurs from a proper understanding of the application domain

Correlation analysis attempts to measure the strength of a relationship between two variables. The correlation coefficient r, denotes this strength, and is calculated as

$$r = \beta . \sqrt{(S_{xx} / S_{yy})} = S_{xy} / \sqrt{(S_{xx} S_{yy}}$$

Where,

$$S_{xx} = \sum_{i=1}^{n} (x_i - mean_x)^2$$

$$S_{yy} = \sum_{i=1}^{n} (y_i - mean_y)^2$$

$$S_{xy} = \sum_{i=1}^{n} (X_i - mean_x)(Y_i - mean_y)$$

The value of r lies between -1 and 1, where negative values for r correspond to regression lines with negative slopes and a positive r shows a positive slope.

5.5.5 Association Rule

Association Rule is one of the tools for KDD to find the relationships or affinity between different attributes of the provided data. An Association Rule discovers how one attribute is related with another attribute or items by finding the items that appear frequently in the same records or transactions. In other words, the relationship is discovered by finding the frequent itemsets appearing in the database with more than a defined threshold value of support and confidence.

An Association Rule is made up of two parts called the antecedent and the consequent. The rules are typically shown with an arrow from the antecedent towards the consequent, since it indicates or measures the affinity of the antecedent towards the consequent Discovering Association Rules is a two-stage process. In the first stage all the frequent item sets are found using the Association Rule algorithm and in the second stage the rules are extracted from those frequent item sets which have a defined confidence limit.

This algorithm can be used whenever one seeks to find which events occur together and have data in the proper form The data must be categorical and arranged in lexicographical order (in general). It works quickly with large numbers of attributes (none of the other exploration engines can handle large number of attributes simultaneously in a reasonable amount of time). It is also an ideal technique for early usage in exploration as it does not require any attribute reduction or target selection

5.5.5.1 Definitions

It is useful to understand some of the basic terminology and definitions of Association Rules

Association Rule: Let $I = \{I_1, I_2, I_3, \ldots I_m\}$ be the set of items and $T = \{t_1, t_2, t_3, \quad t_n\}$ represent the transactional database, where $t_i = \{I_{i1}, I_{i2}, I_{i3}, \ldots I_{ik}\}$ and $I_{ij} \in I$. If X, Y are the subsets of I, called the itemsets then Association Rule will be an implementation of the form $X \Rightarrow Y$ such that $X \cap Y = \phi$.

Itemset Any set of items in a transaction is called an itemset

Support: The support indicates the percentage of the data, which contains both the antecedent and consequent of the Association Rule

For the Association Rules $X \Rightarrow Y$, support can be defined as the number of transactions or percentage of transactions in T, which contain both sides of the rules i e $X \cup Y$ or this, can be expressed as;

$\text{support}(X \Rightarrow Y) = P(X \cup Y)$

Confidence: Confidence is the percentage of the support of both antecedent and consequent together to the support of antecedent or the left side of the rule

Confidence can be defined as the ratio of number of transactions that contain X ∪ Y to the number of transactions that contain X for the Association Rule X ⇒ Y. It can be written as,

$$\text{confidence}(X \Rightarrow Y) = \frac{Support(X \cup Y)}{Support(X)} = P(Y \mid X))$$

Lift: Lift is the ratio of support of the rule to the ratio of the supports of antecedent and consequent For an Association Rule X ⇒ Y lift can be defined as;

$$\text{Lift}(X \Rightarrow Y) = \frac{Support(X \cup Y)}{Support(X)Support(Y)} = \frac{Confidence(X \Rightarrow Y)}{Support(Y)}$$

If the value of lift is more than 1 then it shows that Y is more frequent in those transactions where X is also present than those that do not have X

Leverage: Leverage is the difference between the observed support of a rule and the support that would be expected if the two were independent. It can be written as,

Leverage(X ⇒ Y)= Support(X ⇒ Y) – Support(X) * Support(Y)

Chi Square: Chi Square is a non-parametric test of statistical significance for bivariate tabular analysis It can be used to check the quality of the discovered Association Rules The chi squared significance test takes into account both the presence and the absence of items in sets which makes this test a much better measure than support or confidence measurements for the discovered Association Rules

5.5.5.2 Association Mining

Association mining is a two step process:

1. Find all frequent itemsets
2. Generate strong association rules from the frequent itemsets

Additional interestingness measures can also be applied, if desired. The second step is the easiest and the overall performance of the rule is defined by the first Efficient counting of large items sets is thus the focus of most of the algorithms The Apriori algorithm provides one early solution to association rule mining and most subsequent algorithms have been built upon it.

5.5.5.3 Apriori Algorithm

The Apriori algorithm takes the transactional data and returns all the frequent itemsets that are present in the data with at least the minimum defined support through several iterations The subroutine Apriori_gen takes frequent (k-1) itemsets and minimum support and makes all the possible combinations from those by joining them It returns the candidate itemsets for the next iteration [193]

Once the frequent itemsets are found then extraction of Association Rules is very simple According to the Apriori property all the subsets of the frequent itemsets are also frequent therefore Association Rules are made of all of the subsets of the frequent itemsets Association Rule for all the subsets 's' of a frequent itemset 'l' will be of the form $s \Rightarrow l\text{-}s$.

It is always likely that a large number of rules will be discovered and that these will include many useless rules In order to get good quality rules some kind of filtration is required. A measure of lift and leverage help in determining the interesting rules, and using Chi Square to determine which rules are poor and should be rejected ensures reliable results

There are two main categories of development of Association Rule mining algorithms One branch can be categorized as sequential algorithms and the other as parallel algorithms

Sequential Algorithms include -

- AIS Algorithm
- Apriori Algorithm
- Apriori Tid
- Apriori Hybrid
- Hash Based Techniques
- Transaction Reduction Technique
- Partitioning Technique
- Sampling Technique

Parallel Algorithms include -

- Data Parallelism- CDA (Count Distribution Algorithm)
- Task Parallelism- DD (Data Distribution Algorithm)

5.5.6 Clustering

Clustering is an unsupervised data classification technique which groups data objects based on information found in the data that describes the objects and their relationships The goal of clustering is that the objects in a group be similar (or related) to one another and different from (or unrelated to) the objects in other groups. The resulting cluster tends to capture the `natural' structure of the data [194].

The input for a system of cluster analysis is a set of samples/data and a measure of similarity (or dissimilarity) between two samples. The output from cluster analysis is a number of groups (clusters) that form a partition, or a structure of partitions of the data set One additional result of cluster analysis is a generalized description of every cluster and this is especially important for a deeper analysis of the data set's characteristics.

The "Clustering" process can be described as below:

An input to a clustering analysis can be described as an ordered pair (X,s) or (X,d), where X is a set of descriptions of samples and s and d are measures for similarity or dissimilarity between samples, respectively Output from a clustering system is a partition $\prod = \{G_1, G_2, \ , G_n\}$, where G_k, k= 1,2,,3, .., n is a crisp subset of X such that

$$G_1 \cup G_2 \cup G_3 \quad \cup G_n = X \text{ and}$$

$$G_i \cap G_j = \varphi, \ i \neq j$$

The members $G_1, G_2, \ , G_n$ of $\prod$ are called clusters. There are several methods for a formal description of discovered clusters

1. Centroid based
2 Clustering Tree based
3. Logical Expression based

The availability of a vast collection of clustering algorithms in literature and also in different software environments can easily confound a user attempting to select an approach suitable for problems at hand. It is important to mention that there is no clustering techniques that is universally acceptable in covering the variety of structure present in multidimensional data sets The user's understanding of the problem and the corresponding data type will be the best criteria to select the appropriate method

Clustering algorithms can be divided into five main categories as listed below

Hierarchical Methods

- Agglomerative Algorithms use bottom-up strategy and start by placing each object in its own cluster and then merge these atomic clusters to form larger clusters, until all the objects are in a one cluster or the termination criteria has been met
- Divisive Algorithms use top down strategy and act in reverse of agglomerative approach by starting with all objects in one cluster and then sub dividing them into smaller pieces until each object forms a cluster of its own or a termination criteria is fulfilled.

Partitioning Methods

- Probabilistic Clustering In the probabilistic approach, data is considered to be a sample independently drawn from a mixture model of several probability distributions The main assumption is that data points are generated by, first, randomly picking a model j with probability τj j=1 K, and, second, by drawing a point x from a corresponding distribution The area around the mean of each distribution constitutes a natural cluster.
- K-medoids Methods: In k-medoids methods a cluster is represented by one of its points and has embedded resistance against outliers since peripheral cluster points do not affect them. Clusters are defined as subsets of points close to respective medoids, and the objective function is defined as the averaged distance or another dissimilarity measure between a point and its medoid.
- K-means Methods The k-means algorithm is by far the most popular clustering tool used in scientific and industrial applications The name comes from representing each of k clusters C by the mean (or weighted average) c of its points, the so-called centroid
- Density-Based Algorithms Density-based approaches apply a local cluster criterion Clusters are regarded as regions in the data space in which the objects are dense, and which are separated by regions of low object density (noise) These regions may have an arbitrary shape and the points inside a region may be arbitrarily distributed. The two major types of this approach are

    I - Density-Based Connectivity Clustering

    II - Density Functions Clustering

Grid-Based Methods

Grid-based methods quantize the object space into a finite number of cells that form a grid structure All of the clustering is performed on the grid structure.

Clustering Algorithms Used in Machine Learning

- Gradient Descent and Artificial Neural Networks· k-mean objective functions are slightly modified to incorporate fuzzy errors, i.e. it accounts for distances not only to the closest, but also to the less fit centroids
- Evolutionary Methods: evolutionary algorithms like genetic algorithms (Gas) are also used in clustering Populations are generally "k-means" systems represented by grid segments instead of centroids The population is improved through mutation and crossover

Algorithms For High Dimensional Data

- Subspace Clustering Subspace clustering aims at computing all clusters in all subspaces of the feature space The information of objects clustered differently in varying subspaces is conserved. Objects may be assigned to several clusters (in different subspaces)
- Co-Clustering Techniques: Co-clustering is a simultaneous clustering of both points and their attributes. This approach reverses the struggle: to improve clustering of points based on their attributes, it tries to cluster attributes based on the points

The most important of these clustering techniques are hierarchical and partitioning techniques Clustering methodology is particularly appropriate for the exploration of interrelationships among samples to make preliminary assessment of the sample structure. Clustering is often used to help in the selection of a target attribute for one of the other exploration engines, or to begin an analysis, or identifying outliers in the data Clustering works autonomously and is a good choice whenever one lacks the information necessary to choose a target attribute for one of the more directed exploration engines. However, interpretation of clustering results is a very difficult problem Data may reveal clusters with different shapes and sizes in an n-dimensional data space that is difficult to visualize The number of clusters depend upon the resolution with which data is assessed and deciding the number of clusters in a data is a major problem in clustering

The following list provides the typical requirements of clustering algorithms in data mining

1. Scalability
2. Ability to deal with different types of attribute
3. Discovery of attributes with arbitrary shapes
4. Minimal requirement for domain knowledge to determine input parameters
5 Ability to deal with noisy data
6 Insensitivity to the order of input records
7 High dimensionality
8. Constraint based clustering
9. Interpretability and usability

5.5.6.1 Some Examples of Similarity and Distance Measures

If the goal of clustering is to put similar objects in the same cluster, then the measure of similarity is crucial Informally, a similarity is a numerical measure of the degree to which two objects are alike. The only absolute requirement on similarities is that they are higher when pairs of objects are more alike However, similarities are usually non-negative and are often between 0 (no similarity) and 1 (complete similarity).

Similarity and Dissimilarity between Simple Attributes

Attribute Type	Dissimilarity	Similarity
Nominal	d = 0 if p = q 1 if p ≠ q	s = 1 if p = q 0 if p ≠ q
Ordinal	d = (|p-q|)/(n-1)	s = 1- (|p-q|)/(n-1)
Interval or Ratio	d = |p-q|	s = -d, s = (1/(1+d)) s = 1– ((d- min,d)/(max,d – min,d))

Here p and q are two objects that have one attribute of a given type

5.5.6.2 Distance

In some clustering techniques distance between the two records or their distance from some other point is used as a measure for similarity.

The distance between two points (data objects), p and q, in two, three, or higher dimensional space is given by the following familiar formula for Euclidean distance:

$$dis\tan ce(p,q) = \sqrt{\sum_{k=1}^{n}(p_k - q_k)^2}$$

where n is the number of dimensions and p_k and q_k are, respectively, the kth attributes (components) of p and q.

The Euclidean distance measure above is generalized by the Minkowski distance as:

$$dis\tan ce(p,q) = \left(\sum_{k=1}^{n}| p_k - q_k |\right)^{\frac{1}{r}}$$

Where r is a parameter and not to be confused with the dimension n

r = 1 City block (L1 norm) distance

r = 2. Euclidean distance (L2 norm)

r = ∞. `supremum' (Lmax norm, L∞ norm) distance This is the maximum difference between any attribute of the objects.

5.5.6.3 Similarity between Objects with Binary Attributes

SMC = (Number of matching attribute values) / (number of attribute values)

J = (Number of matching attributes) / (Number of attribute values excluding 00 matches)

Cosine Similarity

$$\text{Cosine}(p,q) = (p \bullet q)/ ((||p||) (||q||))$$

Where • denotes the dot products and ||p|| is the length of vector p, i e. the square root of p • p

5.5.7 Decision Trees

Decision trees are a popular and commonly used classification type of algorithm where classification is done by generating tree like structures that have different test criteria for a variable at each of the nodes and new leaves are generated based on the results of the tests at the nodes. Decision tree is a non parametric approach for building a classification model as it does not require any prior assumptions to be made about the probability distributions of classes and attributes [195] It assumes that once a specific class has been created, it is then expected to work for all future instances or data with the same dimensions Decision tree induction is a useful classification technique for data mining as it provides the following features

1. It is highly expressive for representing functions of discrete variables

2. It is relatively inexpensive to construct and extremely fast at classifying new instances.
3. For small sized trees, it is relatively easy to interpret.
4. It can effectively handle both missing values and noisy data.
5. It can achieve good accuracy that is comparable to other classification techniques in many domains

To apply decision tree methods, several key requirements have to be satisfied

1. **Attribute-value description.** the data to be analysed must be in a flat file form All information about one object must be expressible in terms of a fixed collection of properties or attributes Each attribute may be either discrete or numeric values, but must not vary from one case to another.
2. **Predefined classes** the categories to which samples are to be assigned must have been established beforehand
3. **Discrete class**· the classes must be sharply delineated. It is expected that there will be far more samples than classes.
4 **Sufficient data** the differentiation usually depends upon on statistical tests; there must be a sufficient number of data samples for each class to allow these tests to be effective

Decision tree is generally used for classifying attributes or predicting outcomes These can also be used to discover the numerical dependencies One major drawback of decision tree induction is the data fragmentation problem. At the leaf nodes, the number of instances could be too small to make any statistically significant decision about the class representation of the instances Redundant attributes have quite little adverse effect on the accuracy of the decision tree but they tend to produce a larger tree.

Decision Tree is a supervised learning system in which classification rules are constructed from the decision tree. It takes in a set of objects, the training data set, and builds the decision tree by partitioning the training set Attributes are chosen based on the "information content" and "gain" to split the set, and a tree is built for each subset, until all members of the subsets belong to the same class. While various forms of decision tree induction algorithm have been developed in recent years, the greedy top-down recursive partitioning approach is still the most popular strategy. In general growing a decision tree involves the following task

1. **Determine how to split the instances:** decision tree algorithms often use the greedy heuristic to make a series of locally optimal decisions about which attribute to use for portioning the data.
2. **Determine when to stop splitting:** a stopping condition is needed to terminate the tree growing process Two of the most widely used conditions are: (1) stop extending a node if all the instances belong to the same class. (2) stop extending when instances have similar attribute values

The ID3 and C4 5 algorithms for Decision Trees were introduced by Quinlan Most other algorithms (except for CHAID which is older than C4 5) are enhancements to C4 5 ID3, stands for "Inductive Dichotomizer" and is a greedy algorithm, which constructs a decision tree in the top down recursive manner, and never checks back on its previous decisions In the ID3 algorithm, the attribute selection step is done using information content and information gain calculations.

The ID3 algorithms work out the best attribute in the training instances to separate the given example. If the selected attribute exactly classifies the training sets then ID3 stops otherwise it recursively operates to determine the best possible attribute from the remaining possibilities until it exactly classifies all of them

5.5.7.1 Attribute Selection

Attribute selection is the main decisive factor during the building of the decision tree. The selection of the attribute is aimed at minimizing the depth of the tree so each attribute is tested against a metric or value called the information gain. This method helps in taking closest, on average, to the decisive attribute to split by indicating the highest value for that attribute compared with the others.

Information Content is the quantity of weighted information gained for an attribute corresponding to the decision class variables present in the training data It can be calculated on the basis of the probability of the possible outcomes as follows:

If there are m different answers (classes),v_i, each one having a probability p_i then the information content, I, will be,

$$I(p_1, p_2, ..., p_m) = \sum_{i=1}^{m} - p_i \log_2 p_i$$

If the decision tree node associated with a set of s examples, S, where the class label has m values defining classes C_i, with s_i examples of each class (i = 1,2,3, ,m), the information content can be estimated as:

$$I(s_1, s_2, s_3, ..., s_m) = \sum_{i=1}^{m} - p_i \log_2 p_i = \sum_{i=1}^{m} \frac{s_i}{s} \log_2 \frac{s_i}{s}$$

Now let the attribute X have v different values [x_1, x_2, x_3, ..., x_v] and suppose the X attribute is used to partition the set of examples, S, into subsets [S_1, S_2, S_3, .. S_v], where S_j contains the examples from S which have the value a_j for X. Now let s_{ij} be the number of examples of class C_i in subset S_j. Then the expected information based on the partition using X will be the sum of the multiples of the probability of choosing a branch and the information content of that branch summed over all v available branches or:

$$E(X) = \sum_{j=1}^{v} \frac{s_{1j} + ... + s_{mj}}{s} I(s_{1j}, ..., s_{mj})$$

In the above equation the information content of the branch is

$I(s_{1j}, s_{2j}, ..., s_{mj}) = \sum_{i=1}^{m} - p_{ij} \log_2 p_{ij}$ where $P_{ij} = \frac{S_{ij}}{|S_j|}$ i e. the probability that an example in S_j belongs to class C_i

The "information gain" by branching on X can be calculated as,

Gain(X) = I(s1, s2, ..., sm) – E(X)

Problems with ID3

There are certain problems associated with the ID3 algorithm like the unnecessary leaf extensions or bushy tree structure development due to noisy data This problem can be handled either by getting rid of the noisy data at the data cleaning stage or by tree pruning techniques

There are two common techniques used for tree pruning.

1- Pre-pruning Approach

5.5.7.2 Further Developments in Decision Tree Algorithms

C4 5 was the starting point of a new and advanced era of decision tree algorithms Many developments have been suggested in C4 5 dealing with different issues including incremental issues, scalability issues and parallel processing issues. CHAID, CART, FACT, CRUISE, QUEST (Quick, Unbiased and Efficient Statistical Tree) are a few of the many examples in decision tree classification research work. Several incremental versions of decision tree algorithms have been proposed which include ID4 and ID5. These algorithms work in cases when the new training data is given for training or classifying the data These incremental algorithms restructure the old tree rather than building the new tree based on the new training data Another important area of research in the decision trees was scalability of the decision trees. Important algorithms concerning this issue include, SLIQ (Supervised Learning In Quest), SPRINT and BOAT.

5.5.8 Neural Network

An artificial neural network is an abstract computational model of the human brain A neural network is a network structure consisting of a number of nodes connected through directional links. Each node represents a processing unit and the links between nodes specify the causal relationship between connected nodes All nodes are adaptive, which means that the outputs of these nodes depend on modifiable parameters pertaining to these nodes Artificial neural networks are popular because they have a proven track record in many data mining and decision-support applications The appeal of neural networks is that they bridge this gap by modelling, on a digital computer, the neural connections in human brains When used in well-defined domains, their ability to generalize and learn from data mimics our own ability to learn from experience Each neural processing element acts as a simple pattern recognition machine. It checks the input signals against its memory traces (connection weights) and produces an output signal that corresponds to the degree of match between those patterns In typical neural networks, there are hundreds of neural processing elements whose pattern recognition and decision making abilities are harnessed together to solve problems [196]

5.5.8.1 Neural Network Topologies

The arrangement of neural processing units and their interconnections can have a profound impact on the processing capabilities of the neural networks In general, all neural networks

have some set of processing units that receive inputs from the outside world, which we refer to appropriately as the "input units" Many neural networks also have one or more layers of "hidden" processing units that receive inputs only from other processing units. A layer or "slab" of processing units receives a vector of data or the outputs of a previous layer of units and processes them in parallel The set of processing units that represents the final result of the neural network computation is designated as the "output units". There are three major connection topologies that define how data flows between the input, hidden, and output processing units These main categories are feed forward, limited recurrent and fully recurrent networks.

5.5.8.2 Neural Network Models

The combination of topology, learning paradigm (supervised or non-supervised learning), and learning algorithm define a neural network model. There is a wide selection of popular neural network models For data mining, perhaps the back propagation network and the Kohonen feature map are the most popular However, there are many different types of neural networks in use Some are optimized for fast training, others for fast recall of stored memories, others for computing the best possible answer regardless of training or recall time But the best model for a given application or data mining function depends on the data and the function required The following table 5 1 gives a summary of different models and their functions.

Model	Training paradigm	Topology	Primary functions
Adaptive Resonance Theory	Unsupervised	Recurrent	Clustering
ARTMAP	Supervised	Recurrent	Classification
Back propagation	Supervised	Feed-forward	Classification, modelling, time series
Radial basis function networks	Supervised	Feed-forward	Classification, modelling, time series
Probabilistic neural networks	Supervised	Feed-forward	Classification
Kohonen feature map	Unsupervised	Feed-forward	Clustering
Learning vector quantization	Supervised	Feed-forward	Classification
Recurrent back propagation	Supervised	Limited recurrent	Modelling, time-series
Temporal difference learning	Reinforcement	Feed-forward	Time-series

Table 5 1: Summary of Different Neural Network Models

The selection of model and architecture for neural network depends on data type and quantity, training and functional requirements Neural networks are versatile and provide good results in

complicated domains. It can handle categorical and continuous data types However, input and output must be constrained between some ranges. One of the advantages of neural networks is that they can handle data on which they have not been trained so they are very flexible in their applications and data classification compared with the decision trees, which are limited to their training data's set patterns and always search for the same patterns in the data. It has certain advantage in those cases when an explored dataset contains a large number of records and relatively few attributes. However, it cannot explain the result and also may converge to an inferior solution

5.5.9 Rough Set Theory

Pawlak [197] introduced Rough Set Theory in the early 1980's. Rough set theory is a classification tool which works with discrete variables. Therefore all continuous variables must be transformed into discrete variables.

Rough set is based on the set approximation methods or establishment of equivalent classes with the given training data The data sets within the training data that forms equivalent classes are indiscernible i e the data entries or samples are the same in all the attributes defining that class. But in the real world this does not always happen, as there is always a possibility of undistinguished classes existing that are based on the attributes Rough set can be used to approximately or roughly define such classes

Another important use of Rough set theory is to reduce the dimensions of the data by indicating those attributes, which do not effect the indiscernibility relation present in the data The rejected attributes are redundant since their removal cannot have any adverse effects on the classification of the data There are usually several sets of such attributes and those with minimum number of attributes are called "reducts". Finding such reducts is very important for analyzing very large databases using any of the available data mining algorithms as the time required for analysis reduces considerable when working only with reducts.

Rough set theory's classification algorithm called MD-heuristic, developed by Komorowski, can be used on smaller sized databases and can handle Boolean class output to exactly classify the data or to find the lower approximation sets. This classification can be used to extract governing rules for the output class variables [198]

5.5.10 Genetic Algorithm

Genetic algorithms are stochastic adaptive search techniques They commonly maintain a constant-sized population of individuals that represent samples of the space to be searched. Each individual is evaluated based on its overall fitness with respect to the given application domain. New individuals are constructed by selecting individuals to produce the next generation that will preserve many of the characteristics of their parents The process results in an evolving population that has improved fitness. The two main genetic operators often used to create the next generation are crossover and mutation The main function of these operators is to exchange information between individual parents without any loss of information. GAs avoid exhaustive search by relying on a random mechanism to generate new generations of features, the fittest individuals will be selected for the next round of generation and selection. Knowledge of the domain is utilised to construct new individuals as it helps in avoiding impossible combinations and determines effective measures to evaluate new compound features The process of evolution stops when there is no significant change in the fitness function from one generation to the next or after a certain number of generations GAs are more robust than exiting directed search methods They are quite popular as they do not depend on the functional derivatives They provide a parallel search procedure, applicable to both continuous and discrete optimisation problems, are less likely to get trapped in local optima and can facilitate both structure and parameter identification in complex models. However, practical applications do not always follow the theory The main reasons being that the coding of the problem often moves the GA to operate in a different space than the problem itself. Limited or noisy training data may result in inconsistent, meaningless output [199].

5.5.11 Expert Systems

An expert system as the name indicates consists of a knowledge base of rules (extracted from experts), facts (or data), and a logic based inference engine (or control) which creates new rules and facts based on previously accumulated knowledge and facts An expert system can mimic, to some extent, the reasoning of experts whose knowledge of a narrow domain is deep, thus permitting human experts and expert systems to arrive at similar conclusions

They are highly supervised i e the original configuration or training of the system is not automatic but must have significant user supervision. The knowledge engineer elicits expert knowledge from either human experts through interviews or from textbook procedures This procedure is the most difficult and time-consuming aspect of developing good expert systems

Expert systems are not well suited for use on broad domains of knowledge and are not very robust since they are brittle and cannot easily support illogical complexities, poor clarity (in the facts and/or the rules) or internal inconsistencies of the data set [200] They are not easy to scale, i e. if more rules are added to the knowledge base or perhaps even if the rules are simply rearranged, unforeseen results may occur. Expert systems do, however (unlike the other data mining tools), possess a high degree of explanatory power.

5.5.12 Fuzzy Expert Systems

Fuzzy expert systems are a modified form of expert systems and they help solve the brittleness problem inherent in expert systems. Fuzzy logic has been applied very successfully in many areas where conventional model based approaches are difficult or not cost effective to implement. However, as system complexity increases, reliable fuzzy rules and membership functions used to describe the system behaviour are difficult to determine, furthermore, due to the dynamic nature of economic and financial applications, rules and membership functions must be adaptive to the changing environment in order to continue to be useful The truth and the falsity of a fact can be measured in a fuzzy way using values from the real number interval zero to one inclusive (i e (0,1)) In expert systems information is crisp, in that is it is either totally false or true but in fuzzy expert systems true values can lie anywhere in the interval [0 0 to 1 0] of real numbers Some facts may be close to true and some may be close to false. Fuzzy expert systems can perform as well as, or sometimes better than, human experts can on problem domains consisting of cognitive based, very specific expertise Ideal knowledge domains are very narrow in scope and allow experts to resolve problems in a relatively short period of time. The knowledge should be easy to capture, simple to explain, straightforward to represent (and/or code, typically as fuzzy if/then rules), and should avoid too much dependency on common sense Fuzzy Expert Systems afford the knowledge user most flexibility in generating solutions since consistency and exactness (i e. crispness) restrictions are loosened

Like Expert systems, the most difficult and time-consuming aspect of developing good fuzzy expert systems is attaining knowledge from the experts However, Fuzzy Expert Systems do make knowledge elicitation simpler than the conventional expert systems Fuzzy expert systems are robust (i e. not as brittle as expert systems) and can easily support illogical complexities, internal inconsistencies, and poor clarity or contradiction in the facts and/or rules. In addition Fuzzy Expert Systems are easy to scale and they have a high degree of explanatory power.

5.6 Data Mining Tools and Algorithm Selection Procedures

Categorizing data mining techniques will guide the user, prior the start of the KDD process or during the data-mining phase, in the selection of the best subset of techniques to resolve a particular problem or data mining task

The current KDD process presents several problems. Apart from the basic problem of pre-processing i e cleaning and transformation which are considered to be the most time and resource consuming steps, the selection of a specific type of data mining algorithm is also a problem The data-mining step involves typically the use of one or more inductive learning algorithms, often requiring the user to iterate this step several times, especially when the initial results are not good enough, either in terms of performance or accuracy or understanding of the rules generated for the model. To carryout data mining tasks successfully, it is vital to have knowledge of·

1. which data mining approaches or techniques are appropriate (or best) for which type of task and
2 under what conditions will the identified relationships remain valid

Although data mining has been used for a wide variety of tasks, data mining activities can be divided into two main types: predictive data mining and descriptive data mining Prediction involves using some attributes of the data base to predict the unknown future values of another variable, whereas description, in turn, focuses on finding human-interpretable patterns, which describe the data in order to get insight of the data before trying to predict anything Both of these goals, prediction and description, are really complementary and they use some of the following primary data mining tasks classification, regression, clustering, summarization, dependency, modelling, link analysis and sequence analysis. Appropriate data mining approaches can be determined according to the high level data mining goal, the specific data mining tasks that the user wants to perform and the characteristics of the data set being mined A decision-making process can therefore be designed to determine the specific data mining method to be used which matches the goal of the data mining task as illustrated in figure 5 4

5.6.1 Algorithms Versus Types of Problem

A fundamental issue in the application of data mining algorithms is how to determine beforehand the usefulness and applicability of the algorithm for the class of problems being considered. In other words, before starting the KDD process using a specific data mining

algorithm A_i, it is desirable to know how well it may perform in solving a specific problem P, which, given its features belongs to the type C_i of problems or tasks.

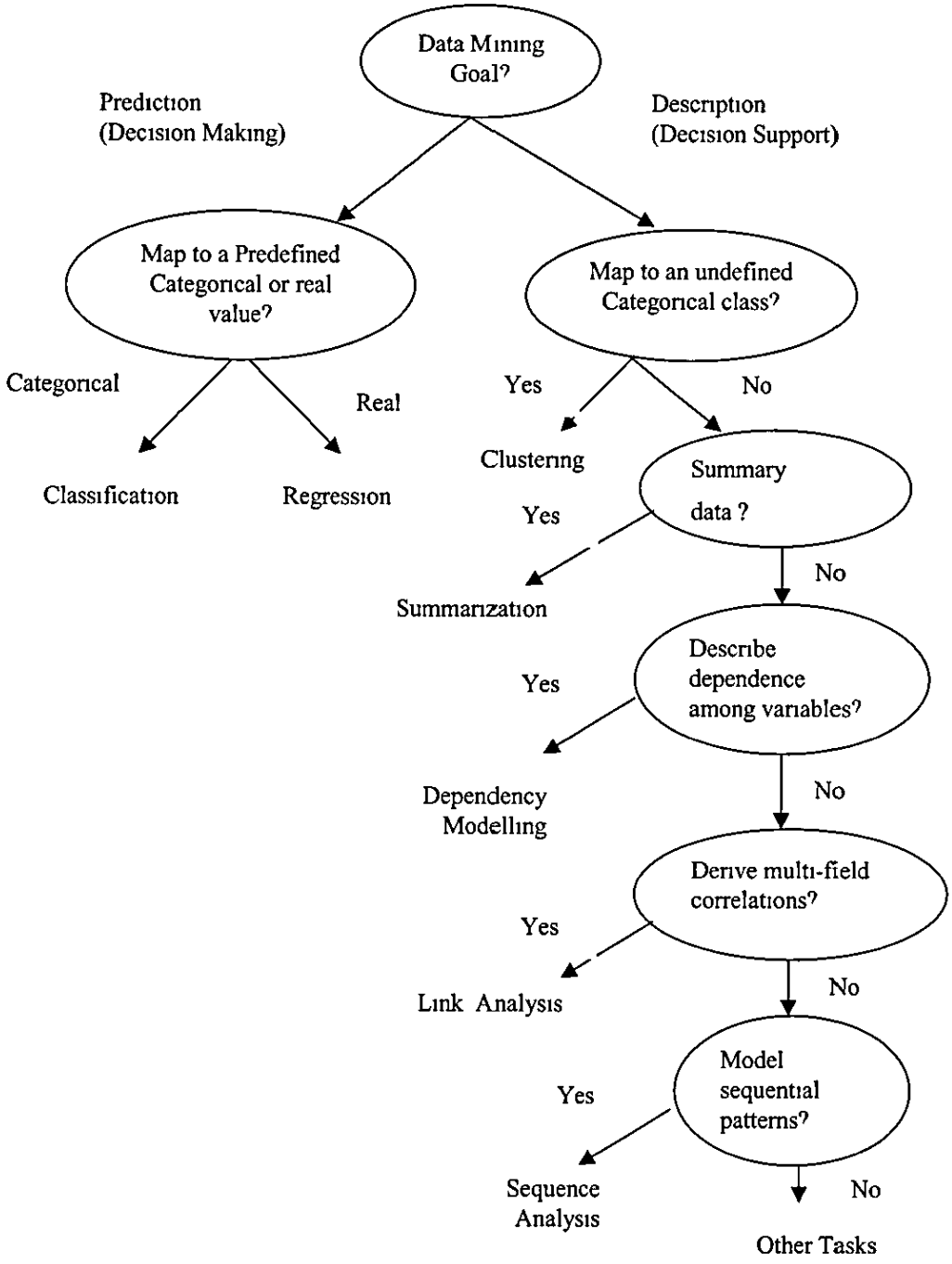


Figure 5 4· Determining the target data-mining task [source [201]]

Moustakis et al [202] performed a survey among the machine learning community, about the usefulness of certain machine learning techniques to solve different types of problems. They considered the set A= $\{A_1, A_2, A_3, A_4, A_5, A_6\}$ of machine learning techniques, and the set C= $\{C_1, C_2, C_3\}$ of types of tasks, where:

A_1. k-nearest neighbour

A_2· Decision Trees

A_3: Association Rules

A_4: Neural Networks

A_5 Genetic Algorithms

A_6 Inductive Logic Programming

And

C_1· Classification

C_2· Problem Solving

C_3: Knowledge Engineering

Most of the above machine learning approaches have been discussed in earlier sections of this report except Inductive Logic programming and this is discussed now. Inductive Logic Programming (A_6) is referred to as the approach that uses First Order Logic (FOL) to represent the learned knowledge. The aim is to construct an FOL program that together with the domain knowledge has the training set as its logical consequence Inductive Logic Programming I(ILP) algorithms learn a set of rules containing variables, called first-order Horn clauses Two well-know approaches for ILP are the Sequential Covering algorithms and FOIL programme

The results of the survey are shown in figure 5 5 which, shows grade of usefulness of each technique in the set A in performing a type of task included in the set C.

Figure 5 5 indicates that neural network algorithms (A_4) perform better for classification tasks than genetic algorithms (A_5), which in turn seem to be more appropriate for problem solving tasks. Also, inductive logic programming (A_6) clearly shows advantages in performing knowledge engineering tasks over the other algorithms considered In addition, decision trees (A_2), k-nearest neighbour (A_1), and association rules (A_3) algorithms seem, in general, to perform better in classification problems than in problem solving and knowledge engineering.

It is evident from the above discussion that no single technique performs best for all types of tasks that are usually involved in solving a real life problem. In some specific problems some algorithms say A_i may perform better than the other algorithms A_j but in other specific problems A_j may give better analysis that Ai even for the same data set [201].

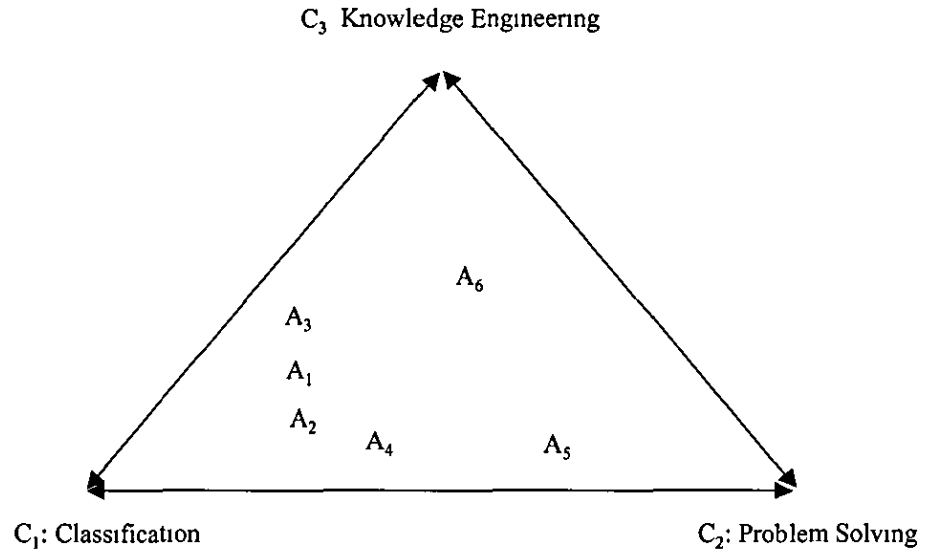


Figure 5 5 Machine Learning algorithms versus types of task [202]

5.6.2 The Quality of the Inductive Learning Algorithm

The type of available data and the nature of a data mining problem typically determine which data mining methodologies are appropriate. In order to select the appropriate algorithm for a particular type of problem, the best way is to assess the different key features of a data-mining algorithm for that particular kind of problem Some data mining algorithms are able to handle larger input data sets than others, some of them build models which are easier to understand and derive rules from them where as some algorithms demand less computational resources (CPU time, memory space) than others etc.

Pieter and Dolf [203] define a set of features F = $\{f_1, f_2, \ldots f_{11}\}$, to evaluate the quality of a data mining algorithm They logically ordered these features in four groups: D_1, D_2, D_3, and D_4, where

$D_1 = \{f_1, f_2, f_3, f_4\}$: Characteristics of the input

$D_2 = \{f_5, f_6, f_7\}$: Characteristics of the output

$D_3 = \{f_8, f_9\}$ Efficiency (performance) for learning

$D_4 = \{f_{10}, f_{11}\}$: Efficiency for applying the model

Where

f_1: Ability to handle large number of records

f_2: Ability to handle large number of attributes

f_3: Ability to handle numeric attributes

f_4 Ability to handle strings

f_5: Ability to learn transparent rules

f_6: Ability to learn incrementally

f_7. Ability to estimate statistical significance

f_8 Disk load in the learning phase

f_9: CPU load in the learning phase

f_{10}. Disk load in the application phase

f_{11}: CPU load in the application phase

The quality of each feature defined in the set F is assessed for each machine learning algorithm in the set A= {A_1, A_2, A_3, A_4, A_5, A_6} where,

A_1: k-nearest neighbour

A_2 Decision Trees

A_3. Association Rules

A_4 Neural Networks

A_5: Genetic Algorithms

A_6 Inductive Logic Programming

In this case, the quality of each feature f_i can assume one categorical value in the set R= {r_1, r_2, r_3} where

r_1 Poor quality,

r_2· Average quality

r_3 Good quality

The assessments based on the study is shown in figured 5 6, 5 7 and 5.8 Each machine learning algorithm in A is evaluated by its quality in each feature f_i As illustrated in figures 5 6, 5 7 and 5 8, each feature f_i, is located in a corner of rectangular area. The quality categories (good, average, poor) are translated to geometric distance to the corresponding feature f_i being considered, such that

If an algorithm A_i performs good in the feature fi, then it is located close to the corner of f_i.

If an algorithm Ai performs average in the feature f_i, then it is located on the diagonal (dashed line) of the rectangular area for f_i, and

If an algorithm A_i performs poor in the feature f_i, then it is located far to the corner of f_i.

Figure 5 6 shows the assessment of the quality of the different algorithms in the set A with different features associated to the input. Thus, decision trees (A_2) and association rules (A_3) perform better in handling large numbers of records than the other algorithms (A_1, A_4, A_5), which have an average performance In terms of ability to handle large numbers of attributes (f_2), decision trees (A_2) perform well, but neural networks (A_4) and genetic algorithms (A_5) perform poorly, because their efficiency deteriorates considerably as the number of attributes

becomes large in the input data set Based on the types of attributes as criteria to select an algorithm, it can be observed that k-nearest neighbour (A_1), decision trees (A_2) and neural networks (A_4) perform well in handling numeric attributes, and association rules (A_3) and genetic algorithms (A_5) perform poorly in this aspect, however when the attributes are strings, a better selection may be genetic algorithms and neural networks, which perform better than the other algorithms considered.

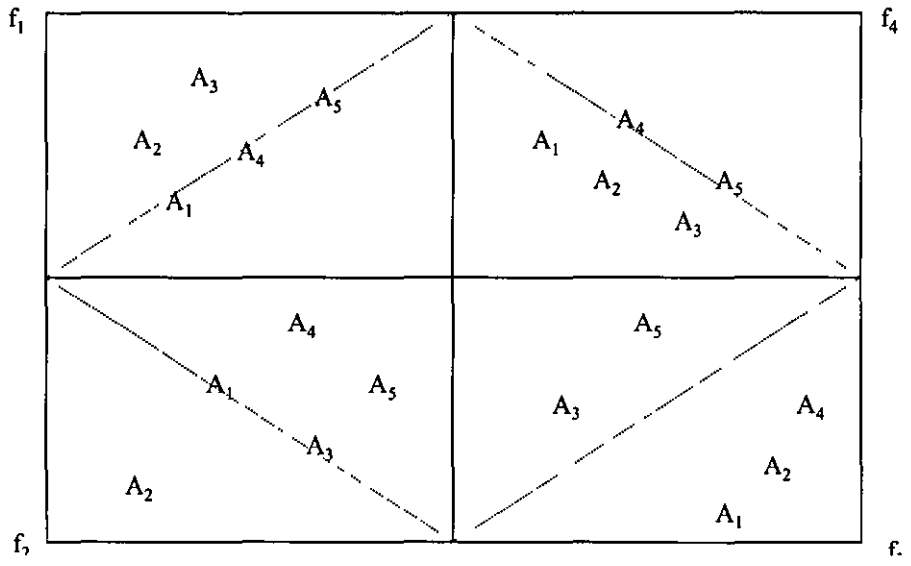


Figure 5 6 Quality of DM Algorithms based on the characteristics of the input

Figure 5.7 shows the performance of different algorithms based on the characteristics of the output produced by the algorithm, K-nearest neighbour (A_1) and neural network (A_4) perform poorly in learning transparent rules. Although they can provide a yes/no answer, no explanations are provided about how the response is reached In terms of the ability to learn incrementally, which is very important with large data sets, as the inductive process does not need to re-start again when new examples are added, association rules (A_3) perform well, but k-nearest neighbour (A_1) and decision trees (A_2) are inappropriate when new cases need to be incorporated to the model.

Neural networks (A_4) and genetic algorithms (A_5) perform poorly if they are judged based on the ability of the algorithm to estimate the statistical significance of the results This is because it is difficult to evaluate their results from a statistical point of view, better choices in this aspect are k-nearest neighbour, decision trees, or association rules algorithm

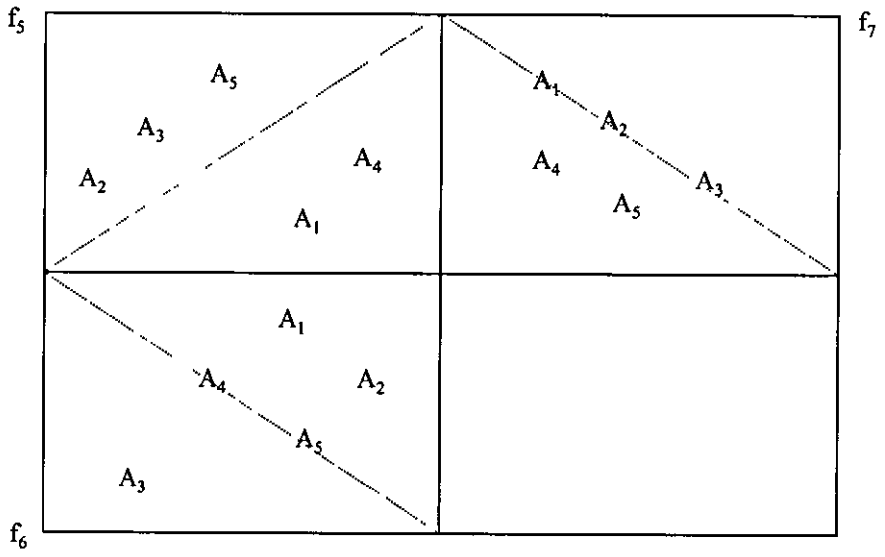


Figure 5.7 Quality of DM Algorithms based on the characteristics of the output

Finally, figure 5 8 shows the quality of the algorithms based on their efficiency in the learning phase and application phase Although, k-nearest neighbour has a good disk/CPU load performance in the learning phase, it performs poorly in the application phase. In contrast, decision trees (A_2) and association rules (A_3) have a good disk/CPU load performance in the application phase and an average rate in the learning phase. When these factors are put together with their good scores (on average) to handle input and output there are clear reasons why they are widely used in data mining applications

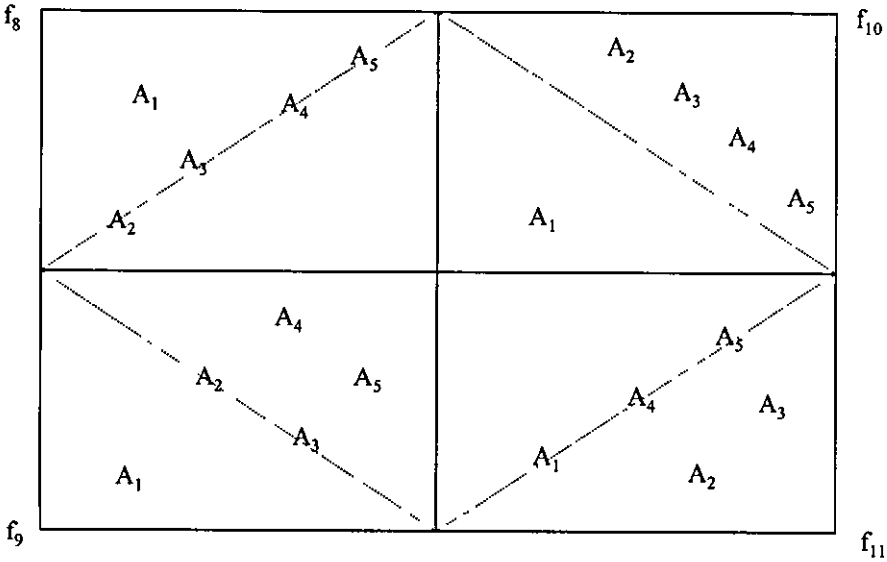


Figure 5 8 Quality of Data Mining Algorithms based on performance

5.7 Critical Comments

Data mining applications typically rely on observational data. Interpreting observed associations in such data is challenging; sensible inferences require careful analysis and detailed consideration of the underlying factors. In general, analysis of observational data demands care and comes with no guarantees, since:

- Associations in the database may be due in whole or part to unrecorded common causes
- The population under study may contain a mixture of distinct causal systems, resulting in statistical associations that are due to the mixing rather than to any direct influence of variables as one another or any substantive common cause
- Missing values of variables for some units may result in misleading associations among the recorded values
- Membership in the database may be influenced by two or more factors under study, which will create spurious statistical association between those variables
- Many models with quite distinct causal implications may fit the data equally or almost equally well.
- The frequency distributions in samples may not be well approximated by the most familiar families of probability distributions.
- The recorded values of variables may be the result of feedback mechanisms which are not well represented by simple non-recursive statistical models

In this research, data mining tools and techniques have been used to generate knowledge in different contexts. It is important to emphasise that no single approach will give the best solutions in all contexts and hence different algorithms need to be applied in different cases to obtain reliable results. The main algorithms that will be pursued in this research are regression analysis, association rules, decision tree and clustering Regression is learning a function that maps a data item to a real-valued prediction variable. It helps in determining whether any correlation exists between variables in the data sets The distribution and variation in data are not known and hence to obtain useful information machine learning needs to be employed Therefore clustering, decision tree and association rule have also been pursued in different contexts. However, other algorithms can also be utilised to generate knowledge

Clustering is a common descriptive task where one seeks to identify a finite set of categories or clusters to describe the data The categories may be mutually exclusive and exhaustive, or consist of a richer representation such as hierarchical or overlapping categories Decision trees are powerful and popular tools for classification and prediction. The attractiveness of tree-based methods is due in large part to the fact that, in contrast to neural networks, decision trees represent rules Rules can readily be expressed so that we humans can understand them or in a database access language like SQL so that records falling into a particular category may be retrieved Association Rule is one of the tools of a knowledge discovery process to find the relationships or affinity between different attributes of the provided data. It helps in determining the relationship between different fields of the database by counting their co-occurrence. An Association Rule is made up of two parts called the antecedent and the consequent The rules are typically shown with an arrow from the antecedent towards the consequent, since it indicates or measures the affinity of the antecedent towards the consequent Discovering Association Rules is a two-stage process. In the first stage all the frequent item sets are found using the Association Rule algorithm and in the second stage the rules are extracted from those frequent item sets which have a defined confidence limit

Chapter 6

Data Mining in Manufacturing: A Review

6.1 Summary

A very thorough review of Data Mining applications in manufacturing contexts has been undertaken by the author in partnership with other members of Product Realisation Technologies Research Group at Loughborough University This has been published as Harding et al [204] and is included in Appendix 1 A summary of the findings of this review, updated by further recently published papers is presented in this chapter

The key asset of any manufacturing enterprise is its knowledge and to properly exploit this resource it needs to be maintained, revised and at times replaced by new knowledge Modern manufacturing businesses store most of their data and knowledge electronically (although some tacit knowledge will only exist within their employees) These data can be a source of valuable assets that are implicitly coded within it. In manufacturing, these data captures performance and optimisation opportunities, as well as the keys to improving processes. Data mining research in manufacturing contexts is primarily focused on attempts to identify, extract and make explicit knowledge that may lie hidden within electronic files or databases, by using various statistical or artificial intelligence techniques and algorithms.

The use of databases and statistical techniques are well established in engineering ([182]) The first applications of artificial intelligence in engineering in general and in manufacturing in particular were developed in the late 1980s ([205, 206]). The scope of these activities, however, has recently changed Current technological progress in information technology (IT), data acquisition systems and storage technology permits the storage and access of large amount of data at virtually no cost. The main problem in an information-centric world remains

how to properly put the collected raw data to use The true value is not gained from the knowledge that is only stored as data, but rather from the ability to extract useful reports and to find interesting patterns and correlations. The extracted knowledge can be used to model, classify, and make predictions for numerous applications.

Data mining has made a significant impact on numerous industries, but its application and benefits in manufacturing have only received moderate attention to date. The strengths of data mining lie where it is difficult or impossible to capture all aspects of a system a priori in a model, either because of its complexity or because of incomplete existing knowledge in situations where large volumes of data are generated by the system. Both situations commonly exist in production environments, which are often too complex for simple mathematical models to adequately capture all essential elements of the system, and much of the knowledge of the operation of the system may be implicit and would thus not be captured by the mathematical models.

The review of approximately 90 papers relating to data mining research in which has been published in Harding et al [204] shows that data mining methods have been successfully introduced in many fields It is still a research topic, but industry is also increasingly showing interest in data mining techniques in order to solve their real-world problems Consequently, the research in data mining is not only driven by theoretical aspects Perhaps more than any other field, data mining is being influenced by currently existing practical problems and researchers are consequently also trying to address the special needs of the industry by incorporating them into their research plans and activities The use of data mining techniques in manufacturing began in the 1990s ([207-209]) and it has gradually progressed by receiving attention from the production community. Data mining is now used in many different areas in manufacturing engineering to extract knowledge for use in predictive maintenance, fault detection, design, production, quality assurance, scheduling and decision support systems. Data can be analyzed to identify hidden patterns in the parameters that control manufacturing processes or to determine and improve the quality of products A major advantage of data mining is that the required data for analysis can be collected during the normal operations of the manufacturing process being studied and it is therefore generally not necessary to introduce dedicated processes for data collection

6.2 Data Mining in Manufacturing

Data mining technology provides many tools and algorithms for the identification of new knowledge and would therefore appear to be an obvious choice for manufacturing

organisations wishing to fully exploit their data resources. The temporal stacked area chart in figure 6 1, which is an updated version of figure 1 from Harding et al [204], (see appendix 1) shows the data mining research reported in different application areas of manufacturing It clearly indicates the current trends of industry towards applications of data mining and shows that particularly since the beginning of the new century people have started to focus on solving their problems using historical databases. Areas such as manufacturing operations, fault detection, design engineering and decision support systems have gained the attention of the research community, although there is still enormous potential for research in these areas. Other areas like maintenance, layout design, resource planning and shop floor control require even greater attention and further exploration

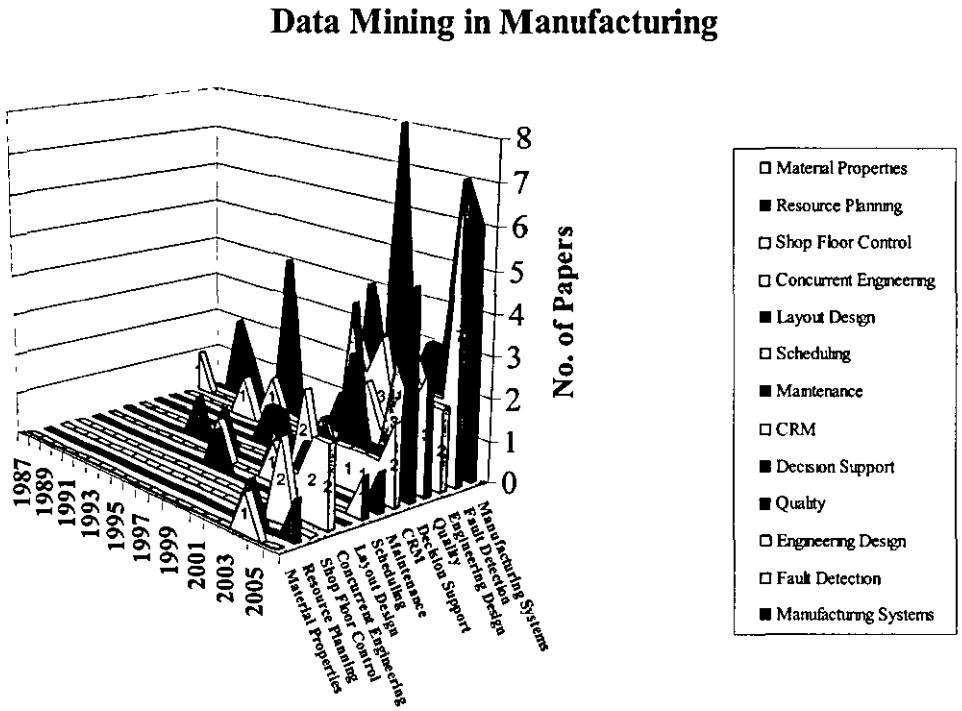


Figure 6.1 History of application of data mining in manufacturing

6.2.1 Engineering Design

Engineering design is a multidisciplinary, multidimensional and non-linear decision-making process where parameters, actions and components are selected This selection is often based on historical data, information and knowledge. It is therefore a prime area for data mining applications, with many published pieces of research Most recently, Shao et.al [210] proposed a methodology and system architecture for engineering and requirement configuration in product design and management Jin and Ishino [211] presented a data mining approach to generate design activity knowledge by analysing CAD operation event data

6.2.2 Manufacturing System

Data collection in manufacturing is common but its use tends to be limited to rather few applications Machine learning, computational intelligence and data mining tools provide excellent potential for better control of manufacturing systems, especially in complex manufacturing environments where detection of the causes of problems is difficult Browne et al [212] utilised data mining to determine mill-set points and generated knowledge for supervisory control of aluminium hot strip mill. Hsu and Wang [213] used decision tree on anthropometric database to extract important sizing variables for soldiers garments Vullers et.al [214] analysed system logs to generate a process model for the system.

Semiconductor manufacturing is complex, facing many challenges, and many data mining approaches have been suggested to overcome these problems Li et.al [215] used data mining based on genetic programming to generate a yield prediction system and perform automatic discovery of significant factors that might cause low yield.

Performance and quality issues have also been considered while applying data mining techniques in manufacturing process related areas. Holden and Serearuno [216] proposed an genetic and fuzzy logic based approach for improving yield in precious stone manufacturing Sadoyan et al. [217] presented a data mining algorithm based on rough set theory for manufacturing process control and illustrated it with rapid tool making process. They also presented a method for controlling output parameters with the help of data mining results.

Efforts have also been made to develop models to study the entire factory or enterprise data altogether to discover the problem areas instantly affecting any subsequent processes Chen and Tsai presented an integration of ERP system and data mining Ren et al [218] presented an data mining approach to analyse the significance of non linearity in assembly processes. Rokach and Maimon [219] presented a data mining approach to generate useful patterns in complicated manufacturing processes.

6.2.3 Decision Support Systems

Decisions are commonly made based on a combination of judgement and knowledge from various domains. The knowledge extracted from databases (prescriptive data mining) can be integrated with existing expert systems to provide different alternatives to decision makers. Lau et.al [220] presented an OLAP based neural network approach for decision support in

resource allocation. Kaya and Alhaji [221] presented a fuzzy OLAP association rule mining approach to effectively process the information generated in the system. Hamilton-Wright and Stashuk [222] presented a statistical reasoning and fuzzy inference method for decision support. Kusiak [223] presented an data mining based framework for organizing and applying knowledge for decision making in manufacturing and services and offered a new data driven paradigm for manufacturing and service organization

6.2.4 Shop Floor Control and Layout

The shop floor control and layout problems are concerned with the efficient and effective utilisation of resources, at the lowest level of control in manufacturing A vast amount of data is recorded during the operation of a shop floor, often to ensure that parts and production steps can be traced This data can also be used to optimise the process itself, since the knowledge generated from mining historical work-in-process data helps in characterising process uncertainty and parameter estimation of the system concerned Knowledge generated from data mining can be used to analyze the effect of decisions made at any stage Sha and Liu [224] used data mining for assigning due dates in a dynamic job shop environment. Backus et al [225] used data mining for determining cycle time for a product in semi conductor manufacturing. Li and Olafsson [226] presented a novel methodology for generating schedule using data driven approach Shiue and Guh [3] presented a hybrid genetic algorithm and decision tree approach to select an optimal subset of system attributes based on production requirements to generate a knowledge base for a production control system. Browne et al [212] used data mining for generating knowledge that helped supervisory control of an aluminium hot strip mill by determining its mill set points

6.2.5 Fault Detection and Quality Improvement

Fault diagnosis is an area that has seen some of the earliest applications of data mining, e g Malkoff [205]. Data mining can help in identifying the patterns that lead towards potential failure of manufacturing equipment. This methodology helps by identifying the defective products and can also simultaneously determine the significant factors that influence the success or failure of the process. The knowledge thus generated by searching large databases can be integrated with the existing knowledge-based systems to enhance process performance and product improvement. Cunha et al [227] presented a data mining approach that used production data to determine the sequence of assemblies that minimizes the risk of faulty production, whilst Li et.al. [228] studied condition based fault diagnosis from incomplete data Buddhakulsomsiri et al. [229] used association rule mining on automotive warranty data to develop useful relationships between product attributes and their causes of failure. Dengiz

et al [230] used data mining for flaw detection in ceramics manufacturing. Rokach and Maimon [219] used data mining on manufacturing data to generate patterns that can then be used for improving its quality. Hou et al [231] developed a data mining based approach to detect and diagnose sensors faults by analysing past data of an air conditioning system using rough set theory and artificial neural network Feng et al [232] used neural networks to predict surface roughness in a machining process

6.2.6 Maintenance

Maintenance is of key importance in process and manufacturing engineering. Databases containing the events of failure of the machines and the behaviour of the relevant equipment at the time of the failure can be used in the design of the maintenance management systems Raheja et al [233] presented a data fusion/data mining based architecture for condition based maintenance

6.2.7 Customer Relationship Management

The marketing model has shifted from being product-focused to being customer-focused, and customer relationship management (CRM) is concerned with increasing the value of interactions with customers and maximizing the profit. Data mining helps in understanding customer demand data and the information obtained is used to determine product design features to meet customer requirements Symeonidis et al [234] used data mining to generate knowledge from an ERP system and then incorporated it into the company selling policies. Tseng et al [235] presented a data mining approach based on rough set theory and support vector mechanics to extract decision rules for accurate prediction and illustrated it with supplier selection in a video game system. Qian et al [236] presented a clustering based functional mixture approach to model customer profile. Crespo and Weber [237] used fuzzy clustering for customer segmentation

6.2.8 Conclusion

Numerous applications of data mining in manufacturing have been surveyed in this research In recent years there has been a significant growth in the number of publications in some areas of manufacturing, such as fault detection, quality improvement, manufacturing systems and engineering design. In contrast, other areas such as customer relationship management and shop floor control have received comparatively less attention from the data mining community. An exponential growth of data mining applications in the semiconductor industry has also been observed The reasons for this may be that large volumes of data are generated

during manufacture and that small improvements can have a significant impact in this industry

Many reported applications are related to the causes of malfunctioning of different types of manufacturing systems or processes and hence the discovered knowledge should lead towards the better functioning of the manufacturing enterprise. Most reported applications in data mining have been "one-off" single shot experiments Whilst these are useful and can often solve particular problems, the knowledge obtained in these projects may not be fully exploited unless the lessons learned in one project are also utilised in other projects and the knowledge generated should also be integrated in the manufacturing system Future work in this area will be directed towards integration of data mining approaches at different level in the manufacturing enterprise and utilising the results obtained at those levels These approaches should be able to use diverse data located at different places and integrated in the database management system Future work should also be directed more towards optimizing the process and suggesting pre-emptive measures rather than just predictions

The research reviewed in Harding et al [204] and this thesis has mainly concentrated on applications of the algorithms. The quality of the data and data preparation issues, particularly relating to manufacturing databases have not been discussed Major effort is needed in the data preparation process, as this is often simply based on practitioner's instinct and experience. A more generic process for data cleaning is essential to enable the growth of data mining in manufacturing industry Also manufacturing data-mining research often does not consider the quality of the rules or knowledge discovered. The knowledge generated is sometimes cumbersome and the relationships obtained are too complex to understand. Future research effort is therefore also needed to enhance the expressiveness of the knowledge and also develop new algorithms that can learn and use data from varying sources.

Chapter 7

7

Data mining Integrated Shop Floor Control

7.1 Introduction

A shop floor manufacturing environment is comprised of different systems that are interconnected and whose operations are interdependent from a decision making perspective Shop floor control plays an important role in coordinating the operations necessary to process production orders across manufacturing resources [238] The main objective of this control is effective and efficient usage of resources [110] The literature review of shop floor control (chapter 3) revealed the ineffectiveness of deterministic and long range solutions in real time manufacturing problems involving various uncertainties such as machine failures, breakdown etc. The complex challenges posed by uncertainties on the shop floor, have motivated researchers to look at intelligent and adaptive techniques to carry out real time manufacturing.

The shop floor is an important unit of any manufacturing system and has offered challenging problems for researchers and practitioners in the last few decades ([5, 13]). Most research has been aimed towards the development of mathematical models [4, 112, 113, 239], heuristics [110, 119, 120] and knowledge based systems [3, 100, 126], each of which has met with varying success. Mathematical models, can provide optimal solutions, but cannot be implemented on large scale problems because of their inability to provide solutions in a reasonable amount of time [240] Heuristic approaches can deliver very fast solutions but tend to be myopic in nature [48] Some dispatching rules such as the shortest processing time and earliest due date, are popular and provide good results [120, 241]. However, these rules independently cannot consider real time information and are unable to consider parallel and

alternative process plans. Knowledge based systems have been successful to some extent as they can help shop floor managers to consider various alternatives quickly. The major disadvantage of knowledge based systems is that they do not scale up well as they become disordered and search intensive as they increase in size They also have poor conflict resolution capabilities and inherit some of the problems of heuristic approaches Therefore an intelligent shop floor controller for a dynamic environment must (according to [3-5])

- incorporate mechanisms, which monitor the environment in real-time,
- present relevant information to mangers so that they can make timely informed decisions and
- support learning mechanisms that are robust in the face of various changes in the environment

These requirements provide a context for the current research, as they indicate that better knowledge discovery and knowledge management are required to improve shop floor control. Data mining tools and techniques have been explored during this research as they provide a methodology for analysing real time data and can generate useful information and knowledge in many areas important for shop floor control.

To contribute towards the satisfaction of the stated objectives of this research, this chapter presents a proposal for an intelligent decision support tool which incorporates data mining and intelligent agent technology to provide useful information and knowledge for a shop floor control system. The architecture presented is similar to the architecture presented by Wang [8] that takes advantage of the intelligent, autonomous and active aspects of agent technology It also provides additional functionality through the integration of data mining processes for the generation of required knowledge and information for different activities on the shop floor into a decision support framework by applying intelligent agent technology.

Manufacturing enterprises routinely generate large amounts of data during their normal operation and the current research is built on the belief that these data stores can be valuable assets and potentially important sources to explore for new information and knowledge By exploiting these assets and becoming more data-aware, manufacturing systems can respond more quickly to market changes and challenges in their business and production activities This may be achieved by intensive and intelligent analysis of existing databases with the objectives of identifying new trends, to predict results and improve performance. Hence the proposed shop floor control system should be able to apply data mining technology to extract and exploit valuable knowledge from its existing and historical operational databases.

It is recognised that information and knowledge are at the core of manufacturing operations and that managers need to make use of stored data to gain valuable insights into the system behaviour [125] The advances in manufacturing system operations have led to an increase in the quantities of information which can be effectively analysed for knowledge extraction. In recent years, information growth has proceeded at an explosive rate. While database management systems (DBMS) provide basic tools for the efficient storage and examination of large data sets, the capabilities for collecting and storing data have far outpaced our abilities to analyse, summarize and extract knowledge from this data Traditional methods of data analysis were based mainly on humans dealing directly with data Large volumes of data overwhelm the traditional manual methods of data analysis and make the task of analysis more difficult and less efficient. Traditional methods of analysis can create informative reports from data, but cannot always analyse the contents of those reports by focusing on important knowledge. There is therefore potentially far more information and knowledge hidden in such databases that needs to be discovered for improvement of the whole of manufacturing and to accurately model the system's behaviour [125]

Knowledge is the most valuable asset of a manufacturing enterprise, as it enables the business to differentiate itself from competitors and compete efficiently and effectively to the best of its ability. The major activity of manufacturing firms is no longer confined to production but lies in the systematic management of knowledge to rapidly meet customer demand [242] Another key reason for deployment of knowledge management systems is the recent advances in information technologies that enable firms to build systems that integrate and consolidate experts' experiences, thereby enabling the companies to provide better services to their customers Therefore, knowledge management becomes a crucial tool for corporations to survive in the volatile marketplace and to achieve a competitive edge [243] Development of a knowledge based system to support the decision making process is also justified by the inability of decision makers to diagnose efficiently many of the malfunctions that arise at machine, cell and entire system levels during manufacturing [7].

Many knowledge based systems have been employed to automate different operations in manufacturing, such as expert systems for decision support, intelligent-scheduling systems for concurrent production, and fuzzy controllers [244]. Knowledge based systems embed human know how into a computer using a knowledge base, which can then be used to reason through a problem. Thus different problems, within the domain of an existing knowledge base, can be solved using the same program without reprogramming The ability of these systems to explain the reasoning process through back traces and to handle levels of confidence and uncertainty provides an additional feature that conventional tools do not have [7] Encoding

the required expert knowledge in a knowledge based system is a tough task [245]. Automation of this process would increase the speed and reduce the cost of development by decreasing the amount of time needed to acquire knowledge from experts and knowledge engineers Every piece of knowledge has a life and it is therefore essential to continuously review and update any expert system or knowledge based system. Effective decision making in a data intensive environment is likely to differentiate future business activities In order to add value and enhance decision making, knowledge based systems need capabilities to categorize and sort tremendous amounts of data while generating information to support decision making with intelligence features Data mining tools and techniques provide an automatic way of carrying out this activity through a series of analytical and computational techniques. Data mining also has the advantage that it is "exploratory" in nature and it can be carried out on existing rather than specially collected data. It is therefore not necessary to do costly experimentation for the special collection of data and moreover the datasets to be used do not have to be complete.

The review of manufacturing applications of data mining in chapter 5 showed that these are mostly "one-off" applications or experiments and no research has yet been reported into a generic data mining enabled architecture that would be relevant across different domains This research therefore attempts to fill this research gap by proposing a new data mining enabled architecture for manufacturing, shop floor control A data mining agent (DMA) has been integrated into a multi agent architecture of shop floor control The purpose of the DMA is to extract useful knowledge from large datasets obtained from product life cycle data and store them in a knowledge pool which can be further reused Knowledge generated from mining enterprise wide data can result in a better understanding of the consequences of decisions made at all levels of the company The combination of a reusable knowledge pool and the automatic discovery and update of knowledge that is generated through data mining analysis should provide managers with the information that they require for decision making on the shop floor The classification model based on data mining algorithms [224, 246, 247] can support various functionalities of shop floor control In this chapter the architecture for decision support in shop floor control has been described first and then each of its different components is described in turn In chapters 8, 9 and 10 the application of the DMA is illustrated with industrial examples demonstrating how knowledge can be identified and generated into re-useable forms.

The proposed data mining enabled decision support tool provides feedback to different levels in a formalised way so that discovered knowledge can be exploited and reused in various ways in the future. One of the important attributes of knowledge based systems is that they have the potential for their knowledge to be continuously updated. However, one of the challenges for

such systems has always been how to efficiently achieve ongoing knowledge updates in a cost efficient manner that helps in learning new knowledge conforming to recent changes in the environment Data mining tools and techniques potentially provide an automatic way of carrying out this activity. The objective of the proposed learning based controller is therefore to provide continuous learning, thus enabling potential automatic updating of its knowledge pool from time to time This learning ability should potentially overcome some of the imperfections in previously reported knowledge-based research. It is also important to note that the blind application of data mining to generate knowledge can be dangerous, leading to the use of meaningless patterns. It is therefore always desirable to incorporate expertise and existing prior knowledge and to properly interpret and validate mined patterns. Additional verification of new knowledge by domain experts is therefore also recommended. Hence a form of semi-automatic update of the knowledge pool is considered to be a more realistic and reliable option than a fully automatic system. Data mining can provide people with useful support for the proper interpretation of mined patterns and strategic decisions, by presenting a set of symbolic rules The proposed data mining enabled decision support system would therefore assist the decision makers by providing them with more alternative options and potentially provide comments about the implications of choosing each of the different possible options It should also be noted that this research tries to provide a generic solution and therefore does not provide a detailed specification of the hardware or software that may be required for the implementation of the proposed system, instead it provides a conceptual description and explanation of a set of modules that can be combined to develop a data mining enabled system for decision support in shop floor control.

7.2 Proposed Architecture

The application of multi-agent systems based on the concept of distributed artificial intelligence is considered to be the most promising architecture for next generation manufacturing [9] (see chapter 4). These consist of distributed heterogeneous agents and make use of flexible control mechanisms for creating and co-ordinating the resulting society of agents. This society of agents provides the foundation for the creation of an architecture that possesses the capability to benefit manufacturing by enhancing a system's reliability, maintainability, flexibility, fault recovery and stability, as well as providing a means for real time decision making on the shop floor [140]

Figure 7 1 represents the schema of any control system. Figure 7 1(a) represents the control module in real time. A network of autonomous agents residing in different locations on the shop floor and consisting of physical objects or logical decision making units (such as job

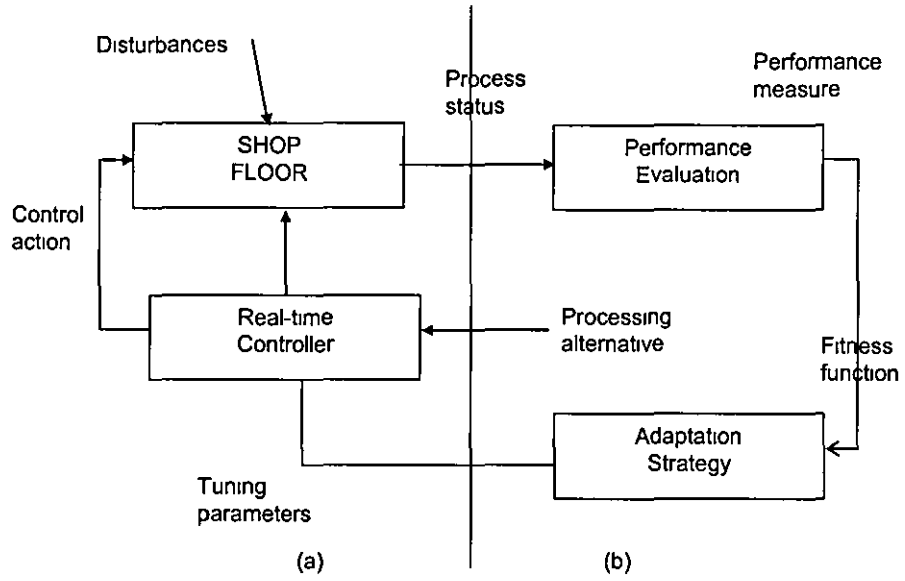


Figure 7 1· Schema of the control loop (a) real time module (b) Decision support module (Adapted from [23,47])

agent, scheduler agent) makes the control loop for the shop floor Each agent measures the actual values of the variables characterising the current processing status and consequently, makes a decision among a set of possible alternatives An actuating agent's decision influences the dynamics of the whole shop floor. Many Multi Agent System (MAS) for shop floor control work in a similar way and can be described by the figure 7.1 a The performance of such MAS is generally evaluated, by means of a set of performance indices (PI) The figure 7.1 b describes the decision support module that uses the real time values of a set of PI to suggest various alternatives and provide the performance information of those alternatives

Manufacturing control is concerned with efficient and effective utilization of resources at the lowest level of control in a manufacturing facility. It involves the co-ordination of the flow of both physical items, as well as information Therefore, the agents within a MAS in this context are found to represent either the physical or informational entities that are required by the system In order for an agent to perform the necessary tasks to achieve its objectives, the agent must be able to participate in a society governed by some set of basic protocols In order to enhance the decision making process, the proposed system makes use of quasi-heterarchical structures adding essentially one or more layers to the single layer heterarchical architectures These additional layers are for the purpose of creating agents with more global awareness allowing them to take the role of manager or mediator of subsets of the agent society [89, 140]

Decision support systems are computer-mediated tools that assist decision making by presenting information and interpretations for various alternatives, enabling decision makers to make more effective and efficient decisions. Incorporating data mining techniques in a decision support system can aid the decision making process through a set of recommendations reflecting domain expertise [248]. It provides information and knowledge which can be crucial for the decision making process. It has been proved that the best way to capture and share knowledge is to embed it in the jobs of workers and to ensure that learning new knowledge is not a separate task that requires additional time and effort to encode what they have learned and what to learn from others [242] In order to build a knowledge repository that captures and embeds the value added knowledge and enhances decision making, it is essential to build a knowledge based system with data analysis capability It is indeed difficult to capture and embed informal knowledge that resides within minds of people in the system McDonnell Douglas, which is now part of Boeing, tried to develop an expert system that contained the expert knowledge necessary to determine whether an aircraft is positioned properly for landing. They gathered the human knowledge by interview and observation. The system took two years and a tremendous amount of resources to capture the human expertise and demonstrated how difficult it is to capture and embed tacit knowledge in a system [245] However, it should be less time consuming and require less resources to develop a system that captures and embeds structured knowledge [242] In the proposed system product life cycle data (product, process and machine data) can be used to generate structured knowledge, with the primary aim that the generated output should be as good as the decisions made by domain experts. The proposed DMA analyses data and generates the knowledge by capturing knowledge embedded in the existing databases

The multi-agent based shop floor control system that has been designed in this research has a decentralised control mechanism which uses knowledge, information and alternatives provided by the decision support tool The knowledge pool can be generated and updated by the DMA, which reduces complexities, enhances flexibilities of functioning and addresses the problem of dynamic management in real time by exploiting the various flexibilities offered by this system and provides various possible alternative measures to be taken in different scenarios The proposed system utilizes the performance information provided by the knowledge pool and the DMA to select the optimum alternatives. Figure 7 2 shows the proposed architecture for shop floor control consisting of multiple autonomous agents. All constituent agents are interconnected via a communication bus with each other together with the shop floor controller. The shop floor controller is also interfaced with this communication bus

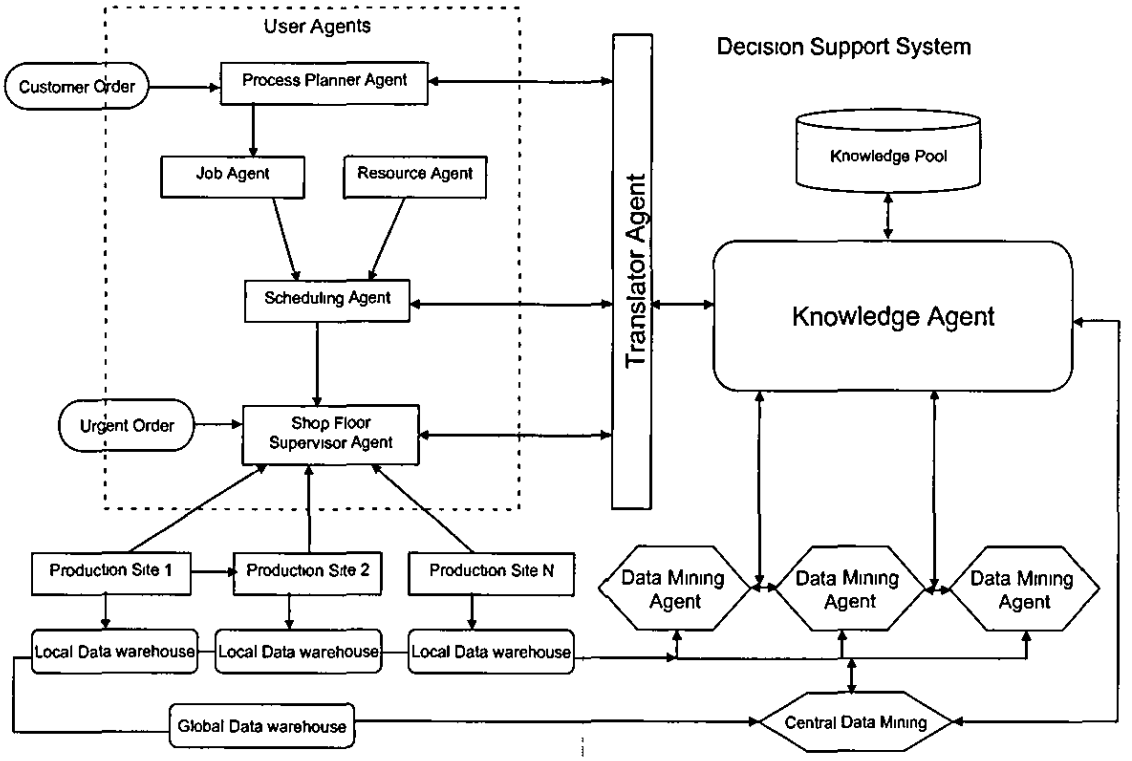


Figure 7 2 Proposed Agent Based System

The functional characteristics of the constituent agents in the proposed control architecture are as follow:

1. **Process planner agent**: The process planner agent dissects the customer orders into a set of alternative manufacturing operations and these alternative operations are represented with the help of a logical AND/OR graph The proposed approach aims to minimise early commitment by constructing a process plan based on the most recent state of the shop and by using knowledge representations which are open for alterations.
2. **Scheduling Agent:** In this context scheduling explicitly refers to the optimisation process of looking for the best schedule It traditionally generates a Gantt-chart which represents the planned operations on every machine/workstation This information may be used by the other agents to improve their logistical decision making process. The scheduling agent performs the above task and sets a guideline for the other agents to follow as exactly as possible. The scheduling agent may have many resource allocation algorithms, each generating a schedule, but it selects the 'best' depending on the goal of the organisation.
3 **Shop floor supervisor agent:** The basic function of the shop floor agent is to follow the schedule generated by the Scheduling agent as closely as possible and to

determine feasible actions in case of breakdowns and urgent orders. The local schedules are compared pair wise with a check for potential conflicts, which are resolved using the rules generated by an operational data warehouse.

4 **Resource agent:** These represent the basic components of the manufacturing system, and bridge between the abstract world (control world) and real world (physical world) The degree of complexity depends on the physical resources represented Information is given about the state of any of the associated resources and estimates of the time during which the resource will remain in a given state

5. **Job agent:** These are created by the process planning agent according to the production objectives and have information about the resources they will use and also the sequence of operations they will perform. These agents also have information about the standard times for manufacturing operations, and thus, they can send an alarm message if these times are exceeded

6 **Data Mining Agent:** These are responsible for accessing data and extracting higher level useful information from the data These agents will perform the various steps of the data mining process and will then select suitable algorithms from the library to provide the required information online and generate knowledge for updating the knowledge pool The information and rules that they generate will therefore be utilised, shared and reused by other agents in decision making

7. **Translator Agent:** This acts as the intelligent interface agent between the decision makers (Process Planning agent, Scheduling agent, Supervisor agent) and the decision support system It provides the interaction between any particular decision maker and the knowledge agent

8 **Knowledge Agent:** The knowledge agent provides system co-ordination, facilitates knowledge communication and evaluates the results obtained by the DMA before updating the knowledge pool. When a decision maker requests support through the translator agent, the knowledge agent interrogates and evaluates the available knowledge in the knowledge pool to identify what relevant information and rules may be known about the current problem and what advice it might provide to help prioritise options for selection. It then provides the required knowledge and information back to the decision makers via the translator agent.

9. **Knowledge Pool:** consists of domain knowledge rules and expertise generated through data mining but verified and evaluated by human experts and then structured in ways that can be utilised by the decision support system

The proposed system organises the agents functionality into three layers (figure 7.3). Like most multilayer architectures, as one moves up the architectures the planning horizon is

extended and the decisions are more global in scope. The top two layers of the architecture support the deliberative behaviour of the agents with the bottom layer providing their reactive behaviour (A deliberative agent is able to reason about its environment and beliefs, to create plans of actions, and to execute those plans. A reactive agent reacts to changes in its environment or to messages from other agents.)

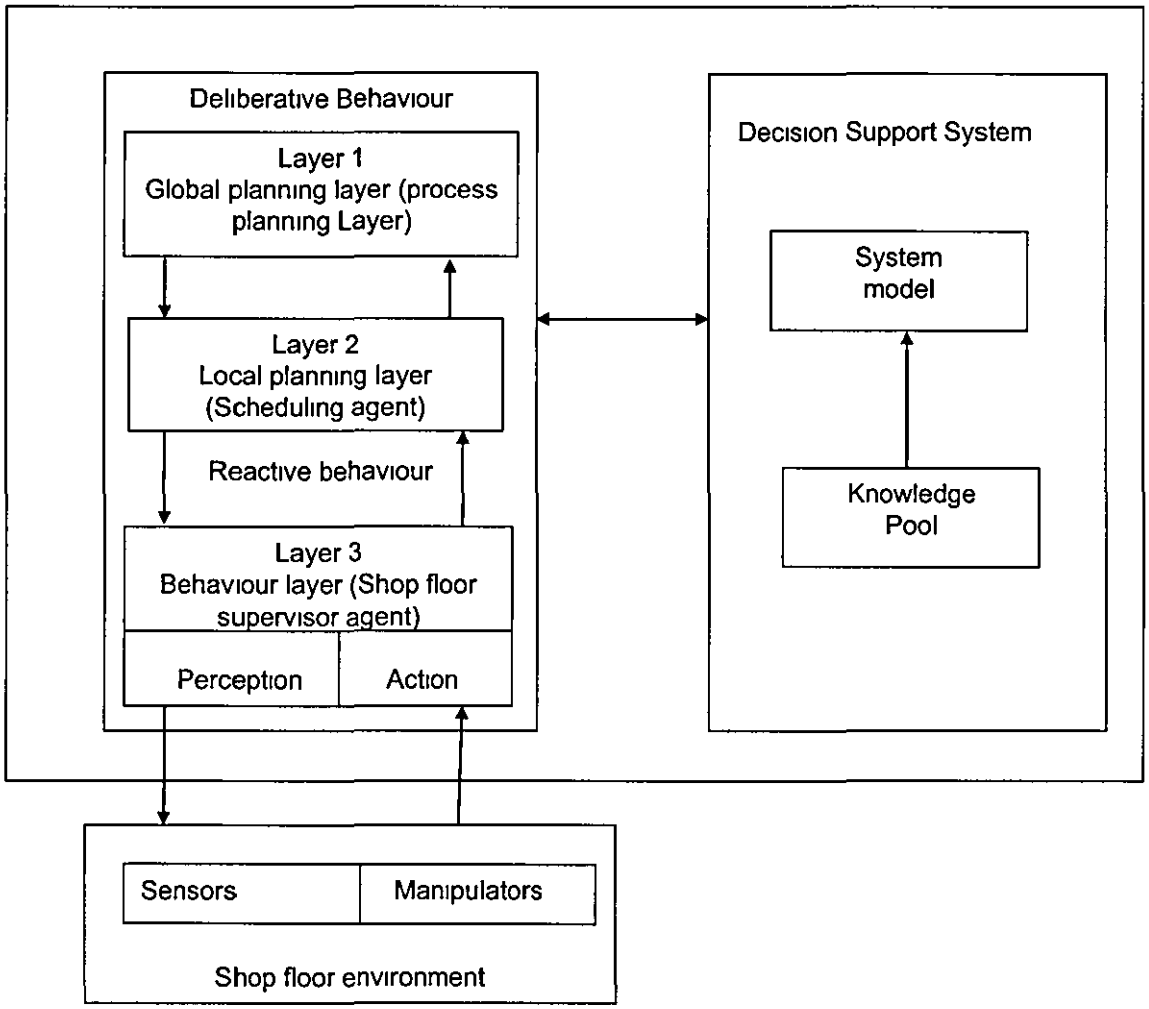


Figure 7 3 · Hierarchy of Shop floor control Environment (Adapted from Bauer et al [19])

The global planning layer (process planning agent) is responsible for performing the higher level planning functions necessary to ensure that the agent is contributing to the goals of the overall manufacturing system. This job involves formulating local objectives that account for the needs and current actions of the environment within which the agent operates The performance of this job requires that the global planning layer maintains an awareness of the overall manufacturing system through the use of a system model and communicates with system agents The system model is used to estimate the effect of the impact of local decisions

on the overall system The global layer within an agent is also involved in initiating and managing interactions with other agents to effect global changes and satisfy system objectives

The local objectives created by the global layer are passed down to the second layer (Scheduling agent), the local planning layer, where the planning functions determine the actions necessary to achieve these objectives. These two layers have the information about the shop floor environment and create plans of actions to meet the demands of it Thus providing the deliberate behaviour of this agent system These actions are passed to the behavioural layer (Shop floor supervisor agent) in the form of commands that the agent will carry out to effect the necessary system changes In addition to executing the required actions, the behavioural layer is also programmed to monitor its environment and respond to events directly affecting it. The behavioural layer would employ schemas to provide the capability to connect perceptual signals with actions providing for the reactive behaviour of this agent system.

In the proposed model, the process planning agent, job agent, resource agent, scheduling agent and shop floor supervisor agent have all been considered as user agents DMA is a part of the decision support system which provides information, knowledge and alternatives to the queries provided by these agents. The decision support system communicates with these agents organised in different layers by providing the different information required by its domain. The DMA for each layer uses different data sources to extract the required information It provides a set of recommendations reflecting domain expertise. It provides useful features for the application of domain knowledge in decision-making It should be able to provide information for re-routing, process parameter variables and feedback for statistical process control, etc, to the decision makers. The next three chapters demonstrate the application of data mining techniques on manufacturing data to create useful knowledge and information The decision support system should also be able to provide performance information generated in those methods to assist decision makers in the system when there are alternative/possible choices. The knowledge agent provides the performance information with different alternatives either by sifting through the knowledge pool or by invoking the DMA The application of data mining can potentially help in generating knowledge autonomously. Thus the knowledge generated in one data mining application can potentially be further used

The main contribution of this research is the incorporation of data mining techniques for decision support system in shop floor control. Therefore the data mining agent, translator agent, knowledge agent and knowledge pool agent are discussed further in later sections The descriptions of other agents are widely available in literature and hence have not been pursued further in this thesis.

7.2.1 Components of a Data Mining Agent

As explained in chapter 5 and 6, data mining is a process of extracting useful knowledge from data automatically. It combines the tools and techniques from machine learning, statistics, artificial intelligence and data management. It tends to simulate human know how and intelligence into a knowledge base, which can then be used to reason through and determine possible solutions to a problem Modern distributed control and data logging systems collect large volumes of data in real time by using bar codes, sensors and integrated vision systems in computer integrated manufacturing environments As shown in chapter 6 the application of data mining to manufacturing problems is not new and successful results have been reported in a variety of problem areas The inclusion of a data mining agent (techniques) within the manufacturing system environment should therefore offer many potential benefits since the available manufacturing data may contain valuable, but hidden, information for operational and control strategies as well as information about normal and abnormal operational patterns which may provide useful information that may be crucial for decision making The data mining agent therefore should provide a set of recommendations reflecting domain expertise

The primary purpose of the proposed data mining agent is to extract as much useful knowledge as possible from existing operational data to build, populate and update the knowledge pool so that it consists of knowledge and experience from the manufacturing environment that can be regularly and repeatedly shared and reused to assist ongoing decision making. At present, successful data mining tends to rely on the experience and expertise of both data mining practitioners and domain experts (i e. it is only a semi-automatic process, relying on some human interaction). However, it can be used to generate useful information about the system performance based on the current status data and past records The performance target for a successful DMA should therefore be that it is able to generate outputs that are as good as or even better than the decisions made solely by human experts in the same situation with the same set of input data.

A data mining agent must be able to perform the following functions in order to achieve its main objectives in knowledge discovery. It must be able to (1) collect appropriate data, (2) prepare the data so that it is free from errors and structured appropriately for the application of particular data mining techniques, (3) determine and apply prior knowledge (if available) to simplify or facilitate the main data mining tasks, (4) apply a range of data mining algorithms (5) analyse results and determine next appropriate course of action (which might be to stop as suitable results have been obtained or to continue with further data mining). These steps may

be iterative i e the previous function may be invoked again depending on the requirement of the problem and algorithm selected for data mining.

To achieve the above functionalities, it is proposed that the data mining agent should consist of the following modules (see figure 7 4)·

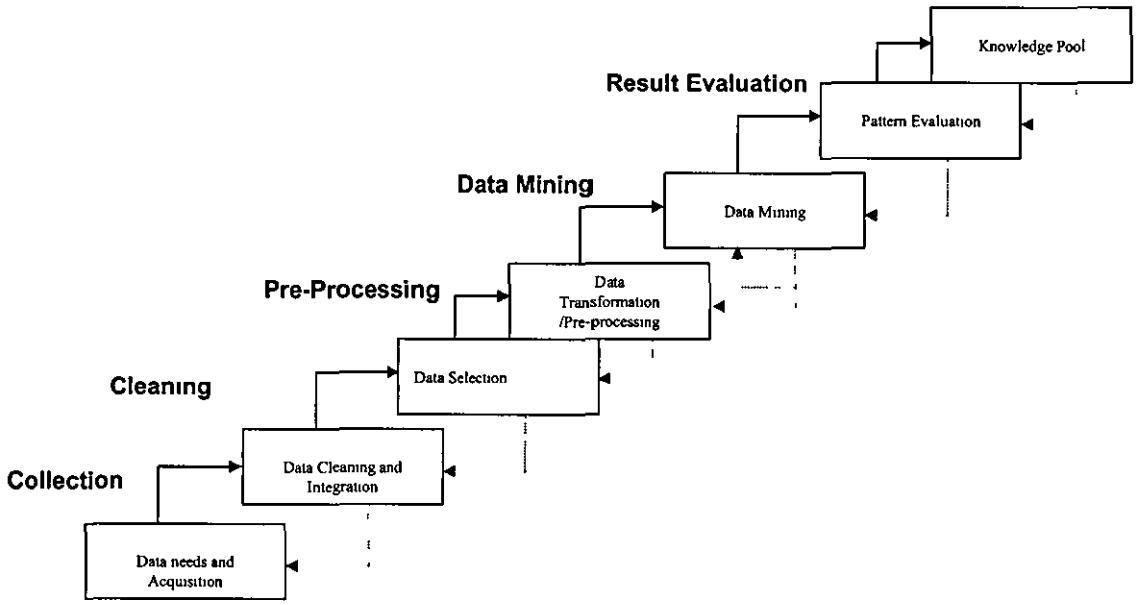


Figure 7 4 Main steps in Knowledge Discovery

a) **Data Collection Module:** There are many different automated methodologies that companies can use to store their product life cycle data in files and databases In the proposed system, there is a local data warehouse for each station and this is linked to the global data warehouse. The Data Collection Module performs the following two tasks.

   1 ***Define data needs:*** The data required depends on the type of information sought. The relevance of attributes differs from problem to problem For example the measurements of a component might be indispensable information in solving one data mining problem e g to provide knowledge of accuracy achievable from a particular manufacturing resources and setup, but might not be essential for another such as to provide knowledge of typical availability of a type of resource

   2. ***Data Acquisition:*** The initial step in any data mining approach is the selection of a historical dataset for analysis. There are three main considerations within this step: data accessibility, population of required data attribute and data heterogeneity. The accessibility of data can be revoked for several reasons

Data might not be stored electronically or is not accessible physically. Population of relevant attributes is crucial to the quality of discovered knowledge. Null values can have some semantic meaning, but they degrade the quality of result. Data heterogeneity is a major concern when there are different data sources. Semantic inconsistencies have to be considered in case of data heterogeneity. Data exporting methods have to incorporate these inconsistencies. Domain expert knowledge needs to be employed in this step

b) **Data Cleaning and Pre-Processing Module:** Data cleaning is a time consuming but essential operation in any data mining task. The process of data cleaning depends upon the particular data mining task and objective The raw data collected from actual manufacturing processes may contain impurities, missing values, and outliers The data cleaning procedure attempts to remove noise, outliers and duplicate records. It takes values from bills of materials and process specifications to determine the specified or required operating value (or range of values) and values lying outside these ranges are treated as noise for process variables Any irrelevant and redundant values for the process variables should be removed as they can potentially deteriorate the modelling results E g chapter 8 section 8.3.1, describes the strategies used for cleaning product dimension related data obtained from a metrological process applied to the product. The section describes the steps that can be followed to eliminate different types of error found in database which records the data from the measurements. However, the steps required to "clean" the data will vary with problem domain and knowledge sought. After the initial cleaning, the data will need to be transformed into different formats depending upon the algorithm selected for data mining Data pre-processing is a significant and important task in data mining as it may affect the data mining algorithm's efficiency and accuracy. Data pre-processing may consists of data clustering, predicting and filling in missing values, coding and heterogeneity resolution, etc. Data pre-processing also includes any steps needed to transform data into a different format if this is required by the chosen algorithm Data cleaning and pre-processing is an iterative task. Data cleaning and transformation is also carried out after identifying critical variables and algorithm selection. The common data transformation tasks are data smoothing, normalisation, categorisation, etc. These tasks may also be carried out after the selection of particular data mining algorithms as individual algorithms may be dependent on different forms of data transformation E g. Chapter 8, section 8 3 2 describes a process of categorising the measurements of product dimensions into different categories. The product dimension data are grouped in bands to generate a few discrete values from the continuous values obtained during the metrological process.

c) **Determination of critical variables module:** dealing with a large number of process variables has several disadvantages First, optimising all the variables takes too much computing time Second, adjusting all the optimised process variable at the manufacturing facility takes too long. One type of information which is usually irrelevant are primary keys, since they are unique by nature and thus do not contain any patterns To determine the critical variables, the process variables can be ranked according to their importance index values using different multivariate statistical techniques, principal component analysis, OLAP, or discriminant analysis [249] In addition to the mathematical indexes the domain knowledge is also employed This domain knowledge can be embeded with the help of humans or can be based on the knowledge gained in similar data mining tasks The identification of critical process variables makes it possible to achieve the desired levels of quality by optimizing only a small number of controllable process variables E g Chapter10, section 10 3 1 1, in cleaning the data of process variable for liquid chromatograph data, the process variable data whose value remains the same are eliminated

d) **Mining module:** Data mining algorithms are then applied on the dataset to characterise and identify the different interrelationships among variables The rules thus generated are checked for their validity and quality. The comprehensible information is then released to assist the decision makers The mining algorithm predicts the various outcomes of the decision and also estimates the limitations of different parameters that existed in the decision. The mining module selects the algorithm to be applied on the data set depending on the data type, information required and other constraints The rules thus obtained are statistically validated and after consultation with the subject matter expert (human expert) they are added to the knowledge pool The human expert also needs to analyse the rules generated and validate them before storing them in the knowledge pool, e g in chapter 8, section 8 3.3, different data mining algorithms have been applied on the product data (which has previously been collected, cleaned and stored in a database) to obtain the relationships between different fields (variables) in it and hence new knowledge is generated and this needs to be validated (by domain experts) before it can be exploited and used

e) **Result Evaluation Module:** The model (results) generated with high quality from the above module are evaluated before being stored in knowledge pool for future use In this step, various assumptions are reviewed to check if the business objectives have been sufficiently considered while performing different data mining steps E g in chapter 8, section 8 3 3 3, a chi-square test has been performed on the rules obtained

from the application of association rules on the product dimension data The test result helps in statistically validating the rules generated

7.2.2 Data Mining Agent

Manufacturing databases are dynamic as new records are added regularly. The data thus obtained must be monitored without interruption so that faults can be diagnosed and eradicated immediately. Therefore, the DMA (KDD process) must run concurrently with the production process [250] The DMA must provide multi-dimensional views on scattered manufacturing data which might exist in several different types of files or databases and generate aggregated data from these to gather and provide the information for further assessment to whichever agent has requested information and support within the shop floor control system The DMA therefore needs to analyse data from the system repositories which store product life cycle data

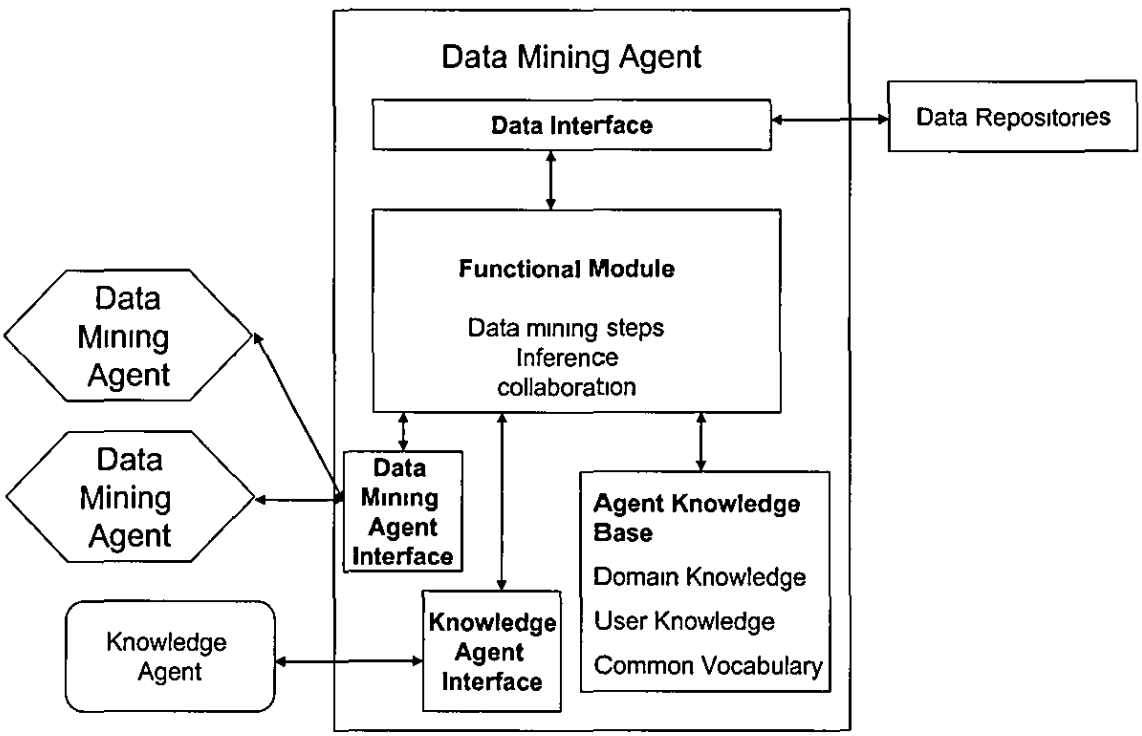


Figure 7 5 Data Mining Agent Architecture

The operational facility and its data interface are shown in figure 7.5. There are 5 modules in DMA. The knowledge agent interface manages the communication between the DMA and the knowledge agent. The communication is message based on shared understanding of domain called an ontology. Ontologies are conceptualisations of a domain into a form which can be understood by all the agents involved They remove ambiguity from communication language

though careful design. The detail of this ontology is beyond the scope this research However, detail about it can be found in young et al [251] The interface agent translates the messages received from the knowledge agent from the common format into the local format based on the ontology and also converts the message from the DMA into common format before transmitting them to knowledge agent. The required knowledge to perform the mining task, common vocabulary, information about the different users and domain knowledge are stored in the agent knowledge base. This knowledge drives the different functionality of this agent. The data interface provides a means for communication with data bases. The databases can be accessed through the structured query language The data mining agent interface allows the DMA to collaborate with other DMAs The knowledge generated about the data mining process/stages (e g strategies for data cleaning, transformation, etc can therefore be communicated to other DMAs and thus this helps in reuse of knowledge generated in applying data mining tools. The functional module provides the basic functionality of this agent It

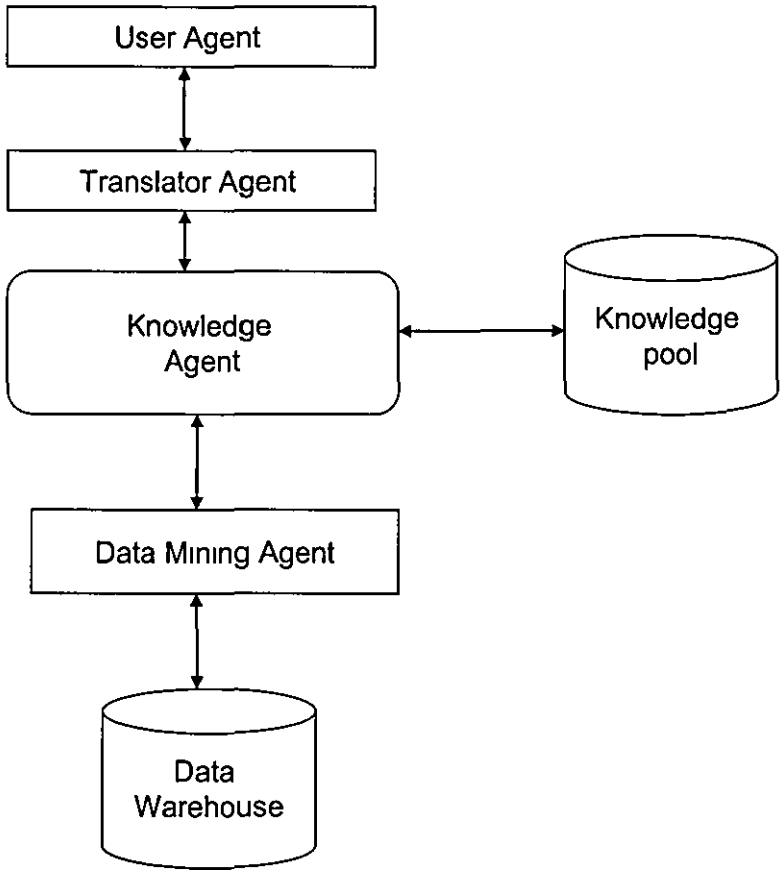


Figure 7 6 Knowledge Generation Process

carries out the task of data analysis and knowledge generation This module consists of various sub-functions that are required to carry out the data mining task and result evaluation like data cleaning, data transformation etc.. The tasks performed by this module depend on the

information received from the knowledge agent interface and it uses the knowledge stored in the agent knowledge base module for its functionality. After completing the task of data mining, the functional module communicates its result through the knowledge interface and the functional module is then terminated

The DMA needs to be able to utilise data mining tools, techniques and methodologies (as discussed in chapter 6) to perform a wide range of data mining applications (similar to those presented in chapter 5) To demonstrate this, three application areas have been examined and are provided as examples of the application of the DMA in the next three chapters (8, 9 and 10) with the corresponding databases to generate the knowledge and update the knowledge pool (Figure 7.6). When the DMA has completed its mining activity on a particular dataset, it transmits the resulting knowledge and information to the knowledge agent to be stored in knowledge pool for future use

7.2.3 Integration of Data Mining

The myriad of organisational informational sources require a team of data miners to distributively discover data relationships or patterns that are contained in the information sources. The individual data miners will need to exchange the captured knowledge between themselves and make such discovered information available to decision makers (via the knowledge pool). The DMAs therefore need to be able to operate in at least two contexts, firstly to provide the results that are of importance to any particular decision issue at hand and provide assistance for an individual instance of the organisational decision making process and secondly to capture data relationships with respect to some general knowledge objectives, to regularly maintain and update the knowledge pool with accurate and valid knowledge, which requires that the data mining processes are constantly executed whenever the source data has been updated.

In the previous section different modules of the DMA have been discussed. This agent can be applied at different micro levels i e within different sections of the manufacturing operation and also at a macro level i e. to examine the overall manufacturing processes A central DMA can also be created which co-ordinates, shares, and exchanges data and knowledge with other similar DMAs Thus a network of DMAs can be build by connecting different DMAs with the central DMA (figure 7.7), which helps in mining the data from entire manufacturing processes. The data relating to each process and its adjacent processes are mined independently. The central DMA mines data related to different steps of the manufacturing process Each process will have its own local DMA and data warehouse where the data will be

stored and analyzed when required. The data can also be transferred to the main data warehouse (if required), which has a direct link with a central DMA for the analysis of the whole system's data. The whole process is thus supported and can be explored by the remainder of the network The activities and results are consistently communicated to the knowledge pool through the central DMA In an integrated manufacturing environment where the product is developed at separate locations or in discrete steps data mining can best be used to control any individual process or step through identification of hidden information in its associated data The manufacturing knowledge and limitations thus discovered can be utilized for better quality control of future productions.

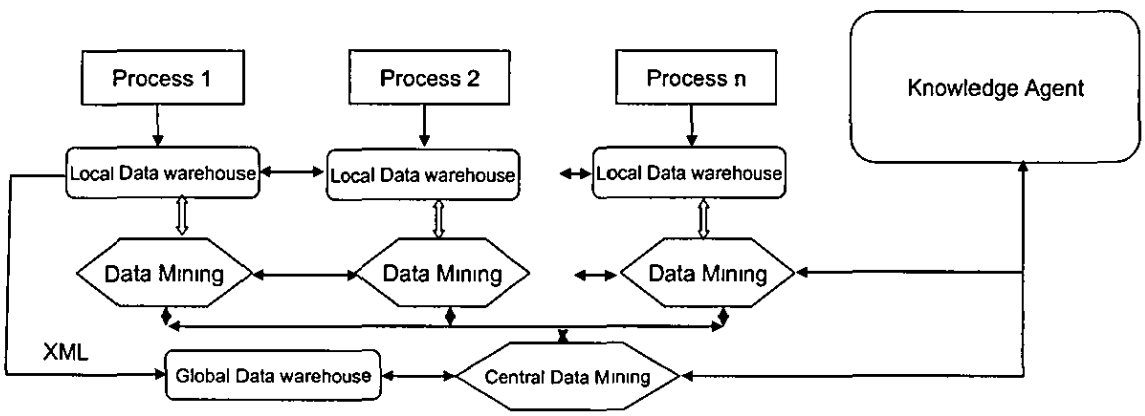


Figure 7.7. Integration of Data Mining Agent

The basic idea for integration of DMAs is that rules, principles and concepts applicable to one manufacturing stage may also be utilized (tested and applied) for other similar stages, (by the exchange of activity and rules information via the Main Data Mining Engine) This integrated DMA can be applied at a factory level where a product goes into different stages and data for each and every step is collected and stored in a pre-designed data warehouse or in the pattern warehouse as knowledge is much more compact than data Data mining activities and the main data warehouse will work in parallel during the whole activity with the production process. A standard format is used to build a central data warehouse for an integrated system where production is carried out at different locations. The data from the individual sites will be transferred to the main data warehouse using XML format where the data will be mined for the whole process and rules/knowledge extracted will be returned back in the same format.

The integrated data mining will be productive in the sense that if different rules are identified for two individual, but similar small manufacturing/production steps The rules could then be tested to see if they are applicable to both steps and if so, the rules can be shared, and each data mining engine can use its knowledge to refine the "best" one for its particular application In this way, knowledge can be fed into the main data warehouse, so results can be reused in

the future Future applications can then make use of the stored patterns and rules instead of always having to return to the original manufacturing process databases.

7.2.4 Translator Agent:

The Translator agent manages the communication between decision makers (User agents) and the Knowledge agent Figure 7 8 shows the different components of this agent The user interface manages the interaction with the decision maker and the knowledge agent interface manages the knowledge communications with the knowledge agent The required knowledge to perform these tasks are stored in the agent's knowledge base. These three components are controlled by the functional module. The translator agent enables the decision makers to view the state of available knowledge, information and data mining processes. It also interprets the data mining results. The conversion of format (Use of ontology [251])is beyond the scope of the present work. However, these issues must be considered in future extension of this work.

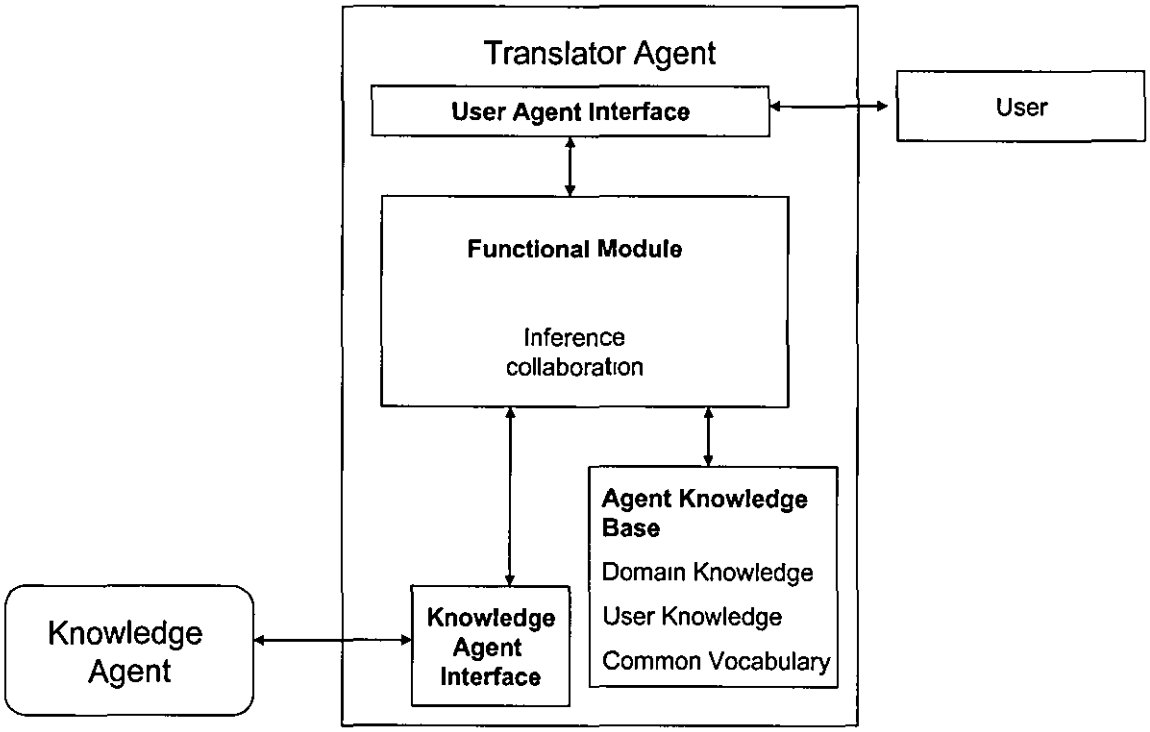


Figure 7 8 Translator Agent Architecture

7.2.5 Knowledge Agent

The Knowledge agent provides the required knowledge and information of different alternatives and its performance information to the decision makers (User agents) Figure 7.9

represents the architecture of this agent. The knowledge agent mediates requests from the Translator agent interface, analyses these requests with its knowledge base and inference module. It then sifts through the knowledge pool to extract the required information or initiate a DMA to perform the required task. It also mediates the knowledge extracted by the DMA and stores them in the knowledge pool for future use. It also transmits the relevant knowledge through the translator agent interface to the decision makers

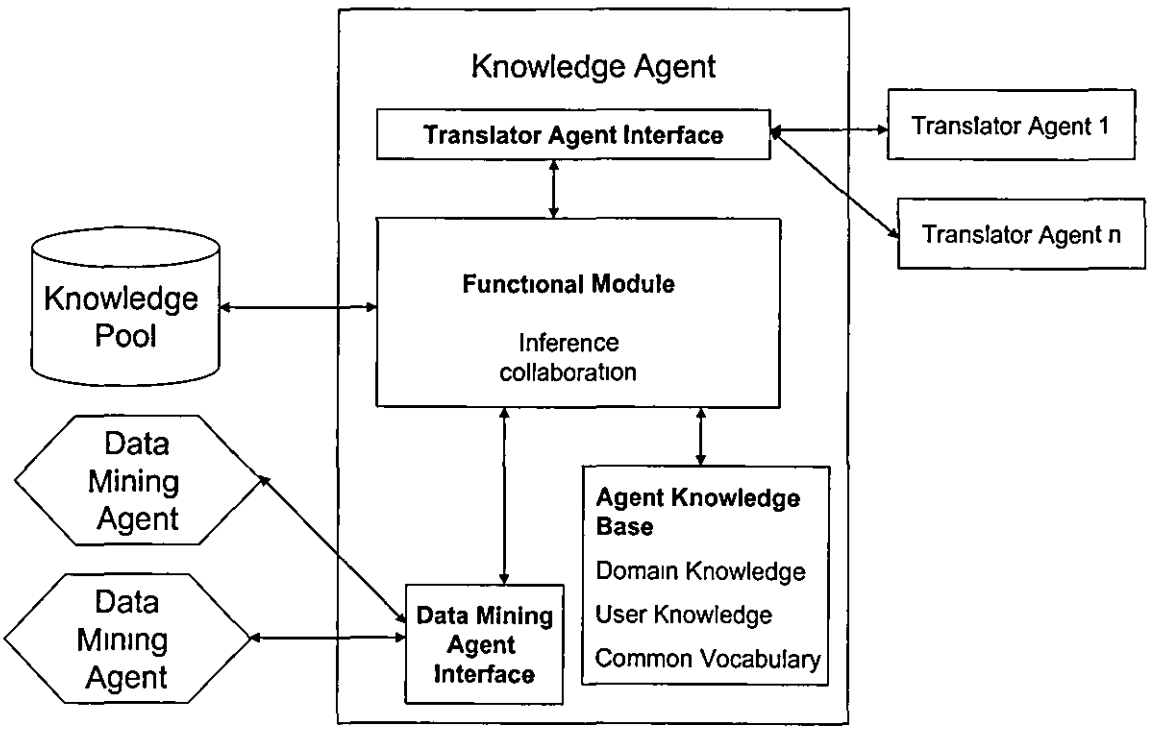


Figure 7 9 Knowledge Agent Architecture

7.2.6 Knowledge Pool

The Knowledge pool consists of domain knowledge rules and expertise required by the system. Different types of knowledge are integrated in it such as knowledge about the mining process, knowledge about the different operational processes and the extracted knowledge, such as rules specifying efficient combinations of input parameters for processes or rules for identifying that maintenance interventions are likely soon, etc., etc. The knowledge pool is updated whenever the DMA finds new or revised solutions The exact structure of the knowledge pool and incorporating knowledge management ([252, 253]) and best practise is beyond the scope of this work However, these things must be considered in future. In this work, the knowledge pool has been considered to consist simply as a set of rules. Knowledge pool content can include all types of digitised knowledge without format restriction. The

knowledge pool consist of Procedural Knowledge (Steps To Solve Problems), Declarative Knowledge (Descriptions of Problems), Meta Knowledge (Knowledge About Knowledge), Common Knowledge (Rules Of Thumb) and Structural Knowledge (Knowledge Structures) i e. all types of knowledge that an organisation may use to describe its functioning and its environment End users would interact with the knowledge pool through client tools, e g Java and Web applications, to insert documents and their associated metadata into the knowledge pool, search for relevant documents, and download them from the knowledge pool. Special care should be taken during development of these tools to ensure that they would be adaptable and that changes to the underlying structure of the knowledge pool should be relatively easy to accommodate

7.3 Prototype Of The System

The goal of Shop Floor control is to achieve the best possible correspondences between external demands and internal possibilities of the enterprise. The necessary functions can be summed up as "logistic process chain" and "technological process chains" forming the basis of production [23] (figure 7.10). Control represents the implementation of planning instructions, in the form of value adding processes.

The planning and scheduling process begins when a customer order is released This order will be in the form of similar structure or parts and then the Process Planner agent generates a linked list of feature based operations In addition to the process plan alternative, the system will also have knowledge of job's due date When the job is released to the Scheduler agent, a Job agent is created to represent the physical order and given the process planning and due date information for storage in its own database A Resource agent exists for each system resource and possesses knowledge of its capability and cost in terms of the feature it can produce, the part shape and size it can hold (fixtures), as well the transporting devices it needs or possesses The Scheduler agent then generates an optimal plan based on information provided by the Job and Resource agent. The shop floor supervisor agent receives the information about the optimal sequence along with different process plans and other constraints on the task. Its main function is to closely follow the sequence generated by the scheduling agent. Whenever an urgent order enters the system, it notifies its advent to the shop floor supervisor The Shop floor supervisor then re-sequences the jobs depending upon the scenario.

In this section, the prototype of the proposed decision support system is presented The objective is to enable the user agents to get alternatives for decision making in different

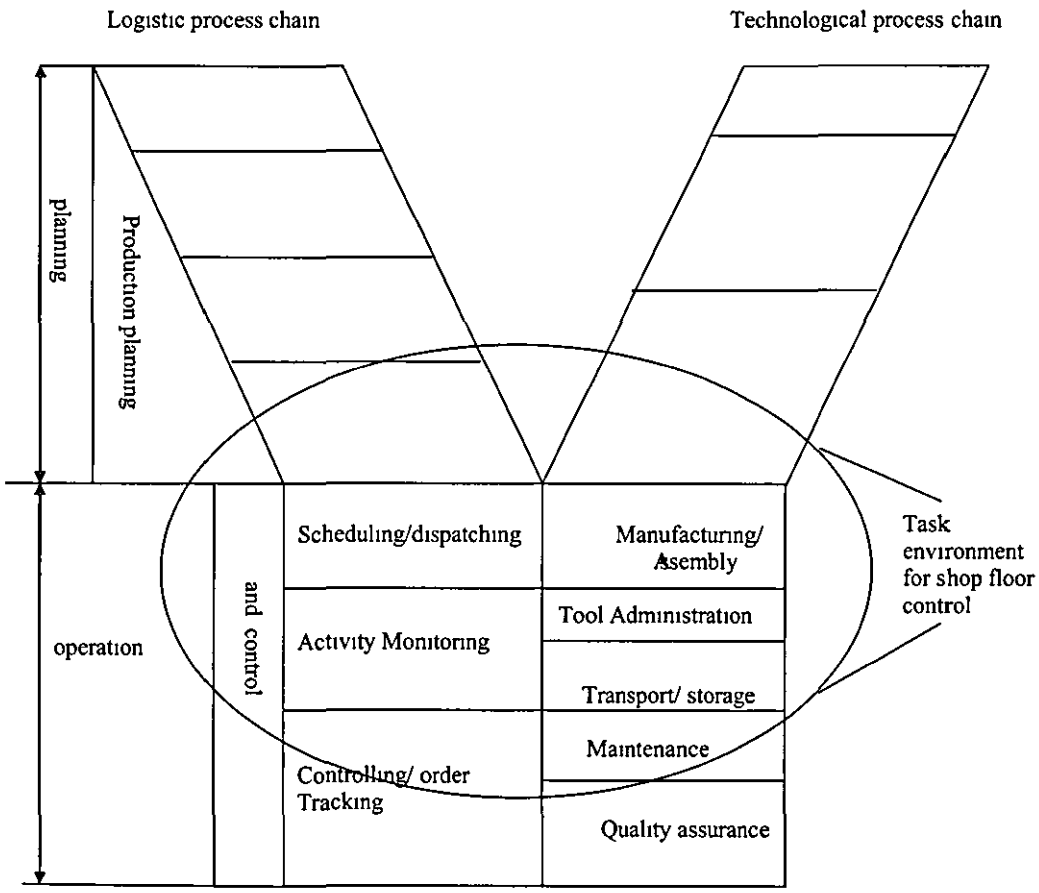


Figure 7 10 Scherer's[23] Y-Model applied to shop floor control

scenarios through agent interactions (using the knowledge agent) For simplicity, in the present prototype, it is assumed that the desired knowledge is already available in the knowledge pool and the knowledge agent does not require the data mining agent to be invoked to extract them However, the functioning of the DMA has been discussed further in next three chapters for different subsystems of a manufacturing enterprise, to more fully demonstrate how knowledge can be acquired by the DMA to add to the knowledge pool.

Each agent is provided with a set of objectives to meet and has a constraint to fulfil These agents would need information and knowledge to meet those goals in the ever changing environment. These agents will communicate with the decision support system for the required information and knowledge. The proposed system functioning has been represented in the figure 7 11

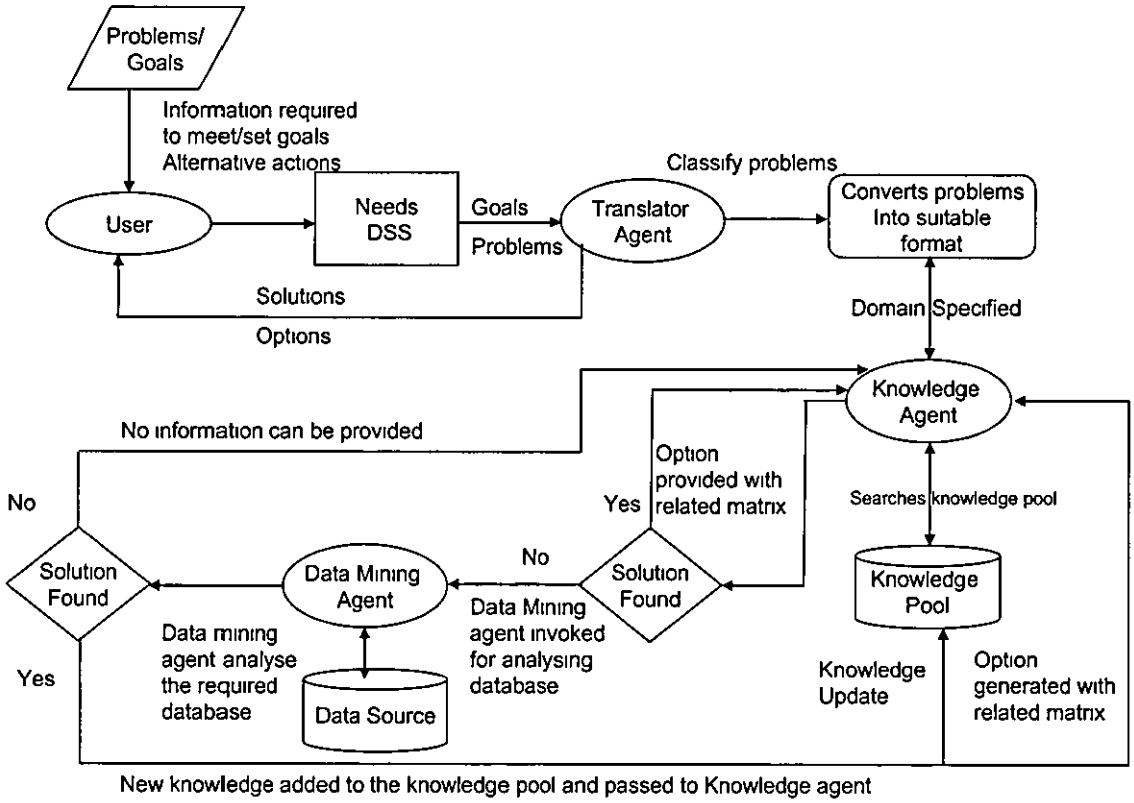


Figure 7 11 Proposed System Functioning

Each user agent (e g Shop floor supervisor agent) is provided with a set of goals/ targets (e g. product dimension) to achieve by the higher level of planning in the system The agent aims to achieve those goals while fulfilling its constraints The user agent needs to cope with changes in the dynamics of the shop floor and take optimal decisions. The decision support system helps the user agent by providing relevant information and knowledge. The user agent logs into the translator agent by specifying their goals/problem (e g. output dimensions are out of tolerance) The translator agent then converts the problem into suitable format for analysis by knowledge agent (e g control signature for product range of process variable that will lead to product within tolerance) The knowledge agent then searches the knowledge pool to find whether such information and knowledge are available or not. The information and knowledge if found is passed to the user agent through the translator agent (e g. a set of rules that include process variables and associated ranges of values for those variables which should produce products with the specified output dimensions in tolerance The quality of the knowledge provided should also be considered and therefore details of the support and confidence for the provided rules could also be passed to the user agent) Otherwise the data mining agents are invoked to analyse the historical data for generating such information and knowledge. The solution if found is passed to the user agent through a translator agent and is

also added to the knowledge pool. Otherwise the message that no support can be provided is passed to the user agent

Typically, the decision support system would be invoked when the user agents cannot meet the set of goals or target assigned to them e g. the shop floor supervisor agent is provided with the task to meet the product dimension It controls the process input variable to meet this requirement. It will invoke decision support, whenever the product dimensions are going out of tolerance. This agent will then logs into the translator agent with information about the dimension (e g length at section "a1") for which information is sought. The translator agent will then seek for a control signature from the knowledge agent The knowledge agent will search the knowledge pool to see if relevant information exists or if not, will invoke the DMA (as shown in chapter 9) to generate the result and will pass the information back to shop floor supervisor agent.

The following section describes the prototype implementation and its application to the proposed decision support model as presented in the previous sections

7.4 Prototype Application

To simplify the demonstration of the prototype, it has been assumed that there are only 3 agents in the system i e. user agent, translator agent and knowledge agent.

1. **User Agent:** in this application process planner agent, job agent, resource agent, scheduler agent, supervisor agent have been considered as user agents. These agents carry out certain functions in the shop floor and seek information from the decision support system on different scenarios. Simple version of these agents can be represented using the following code as

```
CLASS UserAgent
{
  IS-A (Agent)
  ATTRIBUTES // attributes for its domain functioning
  VECTOR wishlist // vector of scenarios in which the system is functioning      //accuracy
required and information required
  Negotiation // basic negation for the knowledge inquired.
  HASHTABLE Negotiation // current negotiation in operation
  Initialise() // register the user agent in the environment and initialise its wishlist
 AgentBody() // domain specific code of agent
 Process() // starts the agent functioning
```

```
If Decision support needed
{
  Register() // register the environment in decision support system
  Processdsspop() // starts negotiation with decision support if there is any item in //wishlist
and no negotiation is in process
  Msg() // communicates message with translator
  Notifydss() // message received from translator
  Processdssagentevent() // starts the negotiation with decision support system
  Processdssmsg() // process the information obtain from translator
Negotiation() // perform any negotiation
  }
}
```

The user agent class as shown above contains the base functionality of all user agents The major data elements from the decision application point of view is *wishlist*, which is a Vector of data elements representing scenarios on which the user agent is seeking assistance from decision support system The *negotiations*, are recoded as part of the Hashtable of negotiations that are in progress. The *AgentBody* method contains the domain specific tasks for the agent and the *process* method initialises the thread enabling the agent to carryout its functions This method registers the agent in the decision support system environment if it detects changes in its operations that it cannot handle by itself. The registration in the decision support method is performed with help of *Register* method and it enables the user agent to add scenarios from the wish list The *Processdsspop* method kicks off the negotiation with the decision support environment by sending a message to the translator agent by invoking *Processdssagent* method It processes the message received from the decision support environment with *Processdssmsg* method

Figure 7.12 represents a GUI for this agent when it is initialised to seek assistance. In this application the user agent logs into the translator agent by using three pieces of information i e user (identifier), information (type of information required) and side (dimension of interest) to describe the scenarios in the environment. These are just a set of simple elements used in the prototype application as they enable information and mined knowledge from the example in chapter 8 to be used in this experimental application. However, for a real time application more elements needs to be added and an adaptive GUI which can accommodate multiple parameters depending on the particular current user requirements would be useful and could be incorporated The lower text box is the space reserved for the information that

would be obtained from the decision support environment (this example will be discussed in chapter 8)

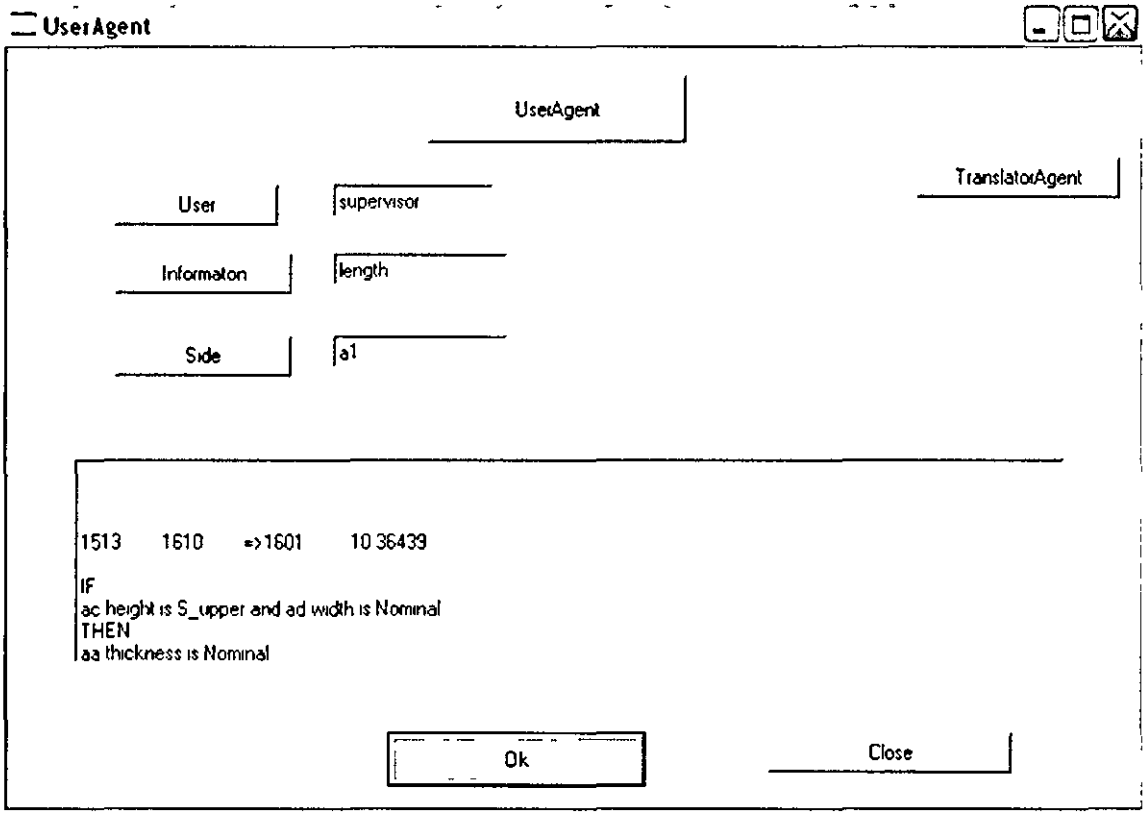


Figure 7.12 User Agent

2 **Translator Agent:** It manages the communication between user and knowledge agent. This agent can be represented as

CLASS Translator

{

IS-A(Agent) // single instance of this agent is created

Hashtable allagent // used for selecting agent

Msgcur // current message that is processed

Reset() // clears all the agents and communities in hashtable

Register() // register the agent in the environment

Initialise() // initialise the agent

StartAgentProcess() // starts the thread

GetTaskDescription() // retrieves the scenarios information

Process() // starts the thread with task description as argument

Processtime() // gives he trace message

ProcessAgentEvent() // process the event/ task for which it is invoked

Msg() // handles the different message

Routemsg () // routes the message to proper agent

Delagent // removes the agent from the Hashtable once its request has been served

}

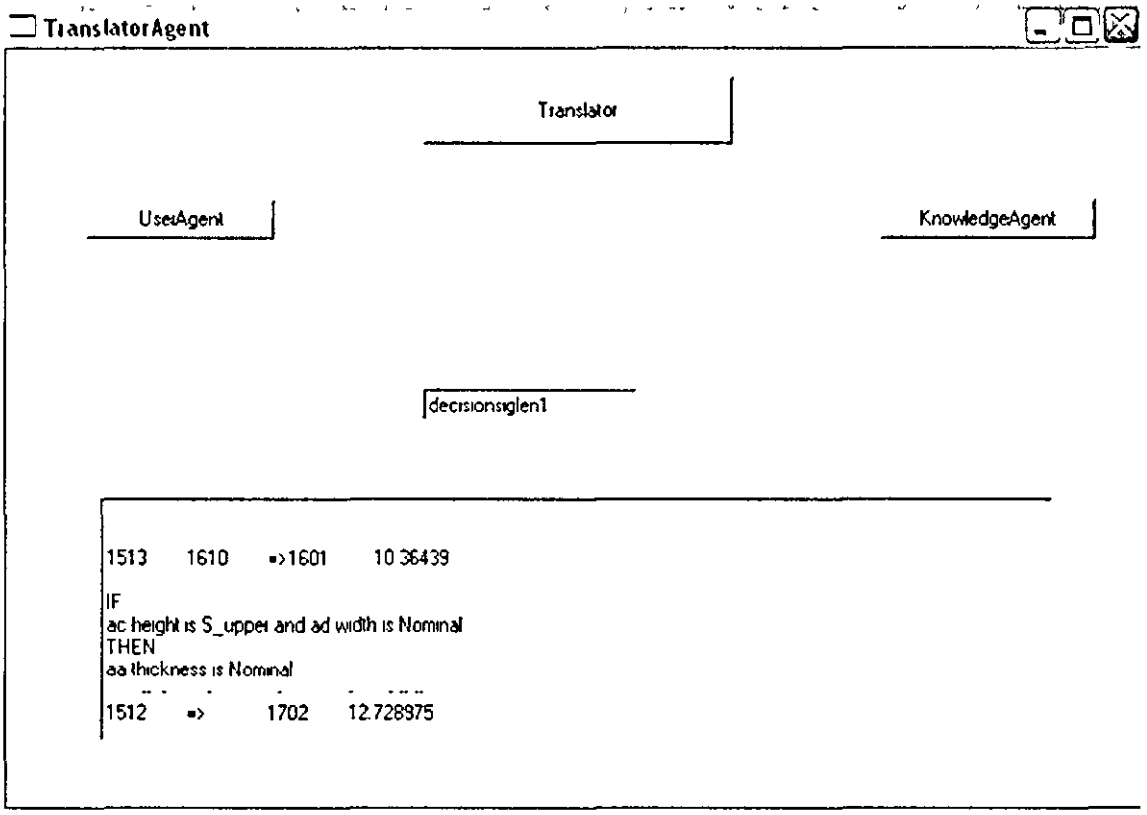


Figure 7 13: Translator Agent

The code shown above contains the base functionality of a translator agent and in the prototype application only a single instance of translator agent has been implemented for experimentation. This feature ensures that all users find the decision support environment through the global translator agent. The translator agent contains two Hashtables- one that is a registry for all agents in the environment and one that contains communities that are of interest in the environment. The *Msgcur* attribute handles the current message that is being handled. The *Reset* method clears all the agents and communities from the Hashtables The *initialise* method initialises the translator agent and registers it with the decision support environment The *StartAgentProcess* method is used to start its thread and make it become runnable when an user agents logs into it The *GetTaskDescription* method retrieves the scenarios from the user agent and *process* method initiates the thread for this inquiry The ProcessEventAgent method implements the *EventListner* interface and implements the

domain code for the previously obtained task description and generates the message to be routed. The *Routemsg* method directs the message to the required agent (user or knowledge agent). It takes *msg* as parameter and uses the reference from Hashtable to direct the message The *Delagent* method removes the agent from the hashtable for which the information has been provided.

Figure 7 13 presents the GUI for this agent In this the text field box represent the field of knowledge that is to be searched and the lower text box will be filled with the knowledge that will be transferred by knowledge agent. The field of knowledge in this example is the decision signature obtained in the example shown in chapter 8.

3. **Knowledge Agent:** The Knowledge agent provides the required information in response to the message received from the translator agent It searches the knowledge stored in the knowledge pool in correspondence to the information sought by the user agent and if suitable knowledge exists it returns this to user agent via the Translator Agent. This agent can be represented as

```
CLASS KnowledgeAgent
{
  Seed // current message being processed
  HashTable KnowledgePool // knowledge pool organisation
  HashTable negotiation // negotiation that are in progress
  GetTaskDescription() // retrieves the task description for its performance
  Register() // register the agent in the environment
  Initialise() // initialise the agent when translator logs into it.
 Taskdescription() // // retrieves the task
 ProcessAgentEvent() // runs the domain code of it
 Message () // used for passing message
 Msg() // messages generated
 ProcessMessage() // processing the message
 Checkiteminhashtable() // Checks if any item is left in hashtable
  Deliteminhashtable () // removes items in hashtable for which information has been
//served
 }
```

The knowledge agent class, as shown above contains the base functionality necessary for knowledge agent to provide the user agents with answers to their queries. The major data

members of this class include *KnowledgePool*, which is a hashtable of elements that help in searching knowledge pool and *Negotiation* which are hashtable of negotiation that are in progress. The *Initialise* method initialises the different data elements in this agent and registers it in the environment The *register* method gets invoked when the translator agent logs into it and starts the agents threading running. The *TaskDescription* method enables it to retrieve the task which it is required to perform. *ProcessAgentEvent* method searches the knowledge pool for task description It passes the information obtained from knowledge pool using *Message* method and processes the messages received from other agent through *ProcessMessage* method The thread goes to sleep when there are no items in the hashtable and deletes the items from the hashtable for which information has been provided Figure 7.14 represents the GUI for this agent In this, the text box contains the field for which knowledge pool has been searched and the lower text box contains the information that has been obtained from it

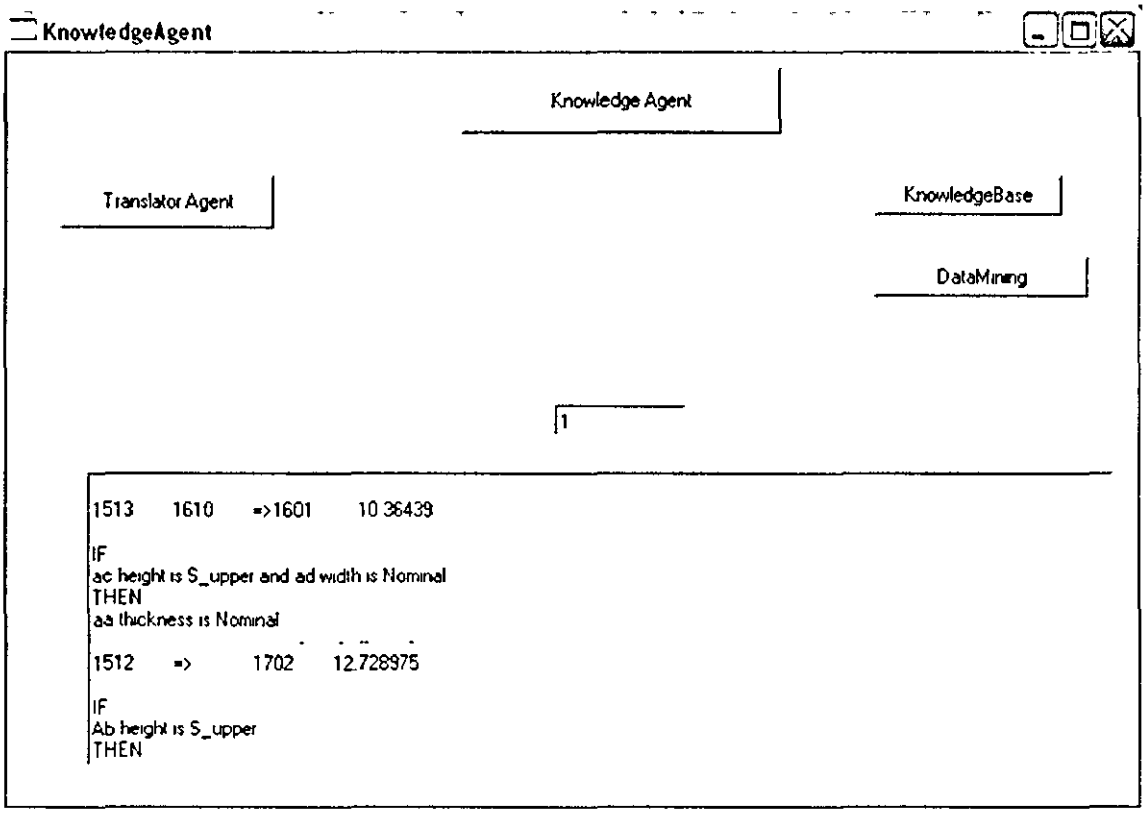


Figure 7.14. Knowledge Agent

7.5 Implementation

This section presents the prototype implementation of the proposed architecture for the decision support system in the shop floor control A java application was developed to present

the behaviour of the proposed system where user, translator and knowledge agent interact. The decision support application class contains the main() method and its user interface as shown in figure above. The applications were constructed using Borland Jbuilder 3 0 interactive environment The visual builder enabled the complete GUI to be created by using swing component in a drag-and-drop style and automatically generate codes for creating GUI controls and event handlers The logic for other agents was then added to set up the decision support infrastructure

Figure 7.15 shows the main panel of the decision support system application User Agent, Translator agent and knowledge agent are three data members for this application class The two text areas are used to display messages from translator and other agents in the system The translator agent manages the implementation environment and includes other agents to interact within the application domain. The translator agent acts a medium for communication between the other agents in the system.

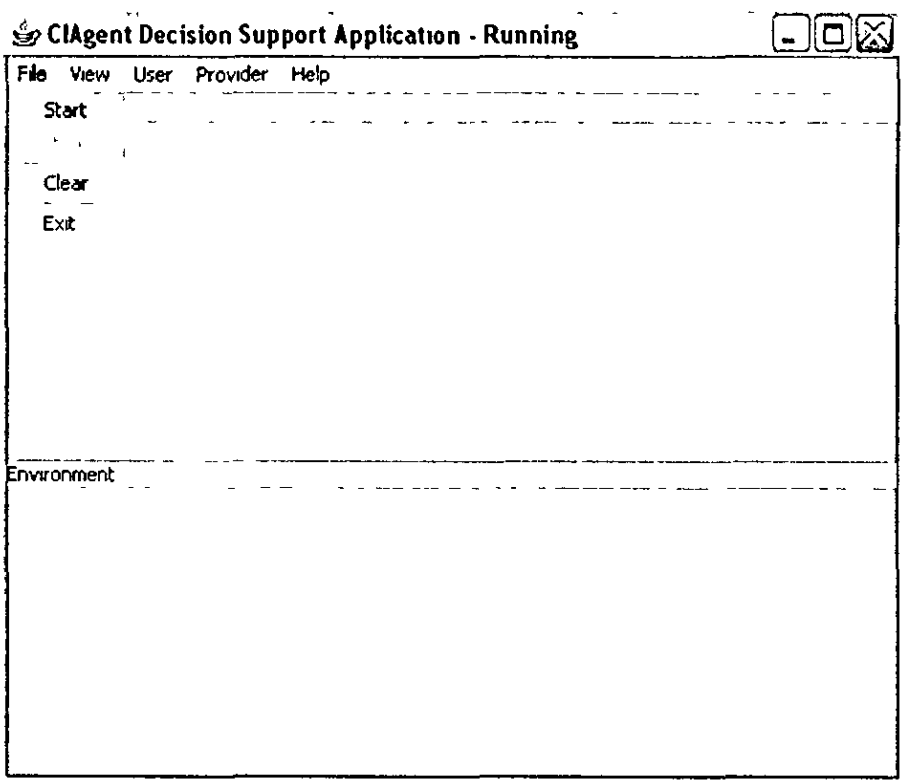


Figure 7.15 : Application running

When the user agent selects the start option (figure7.15) from the file menu, a single translator agent (facilitator) is created along with other agents. The user agent then supplies the input to seek information from the knowledge agent (provider) figure 7.16 The user agent provides the pieces of information as input to describe the scenarios to seek information from the

knowledge agent (provider) via translator agent(facilitator) (figure 7 16). In this example, the user agent (supervisor agent) provides three inputs to describe the scenarios i e user name (user identifier), information required (type of information) and side (dimension of interest) to the translator agent along with the accuracy level The accuracy level defines the minimum accuracy (confidence in result) in the knowledge which the user agent is ready to accept.

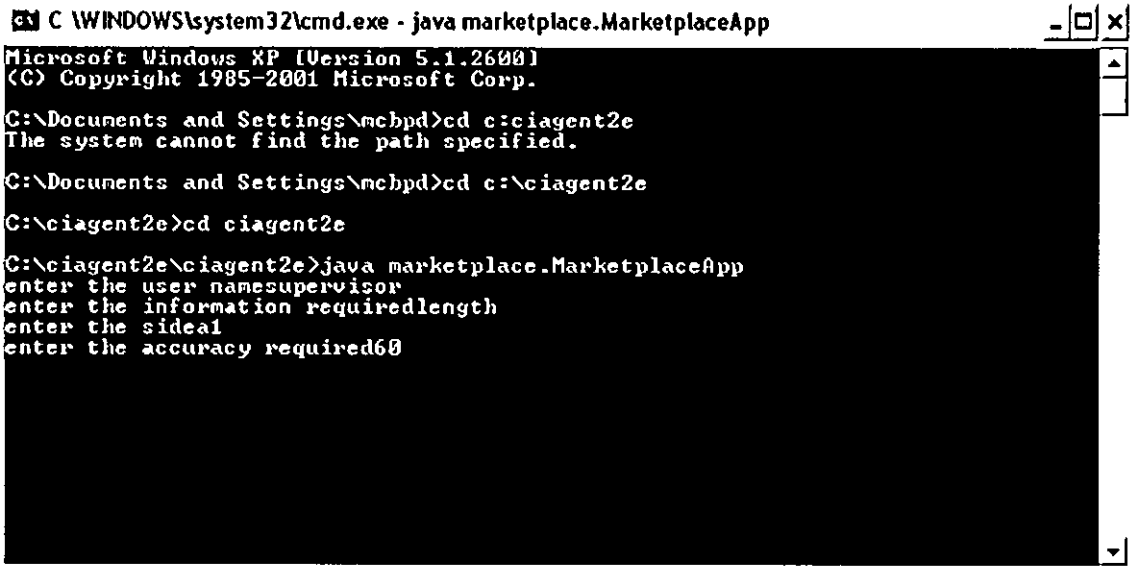


Figure 7 16 User Input

The user agent is then initialised in the system The translator agent (facilitator) uses the first three inputs and translates the user input to the type of information it is seeking and passes it to knowledge agent (provider) by adding it in the wishlist. It acts as a match making between user and knowledge agent (figure 7 17) The translator agent (facilitator) initialise the knowledge agent (provider) and figure 7 17 show the main window after knowledge agent (provider) and user agent have been initialised and before any knowledge has been passed. The translator agent (facilitator) adds the knowledge pool to the knowledge agent (user community) as shown in the upper text box in figure 7.17. The two agents use the accuracy level for negotiation to obtain the different information that is above it. The translator agent (facilitator) recommends the knowledge community and the knowledge agent (provider) makes offer to user agent along with its accuracy, as shown in upper text box of figure 7.18. The user agent accept the knowledge community which have accuracy higher than the desired The knowledge agent transfer the stored knowledge to the user agent through translator agent. The lower text box figure 7 18 represents the content of the knowledge transferred

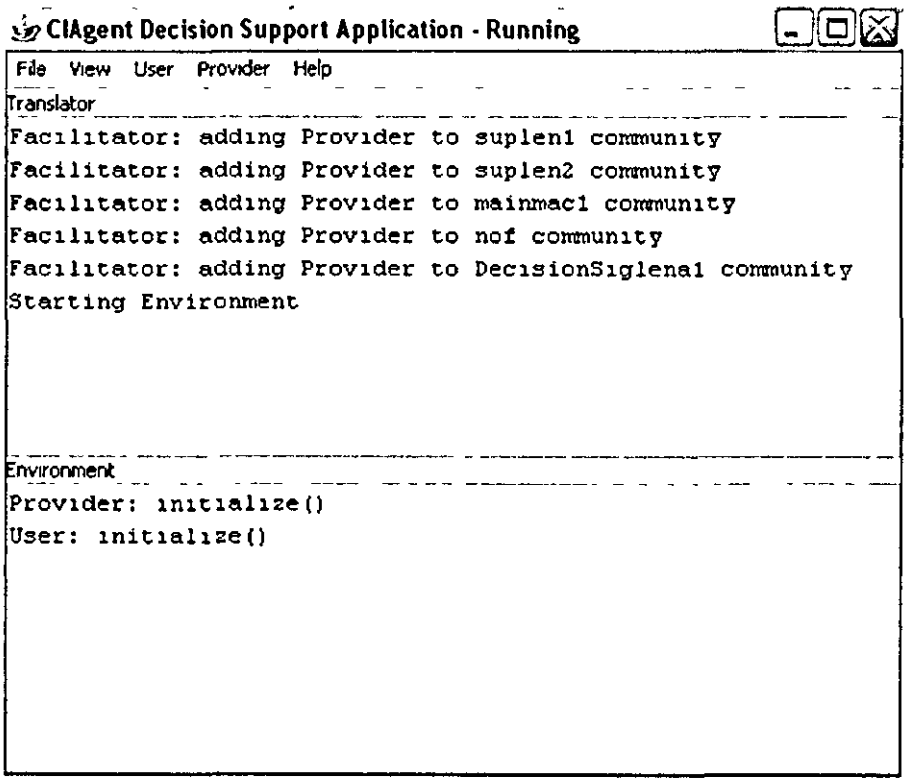


Figure 7.17. Initialize

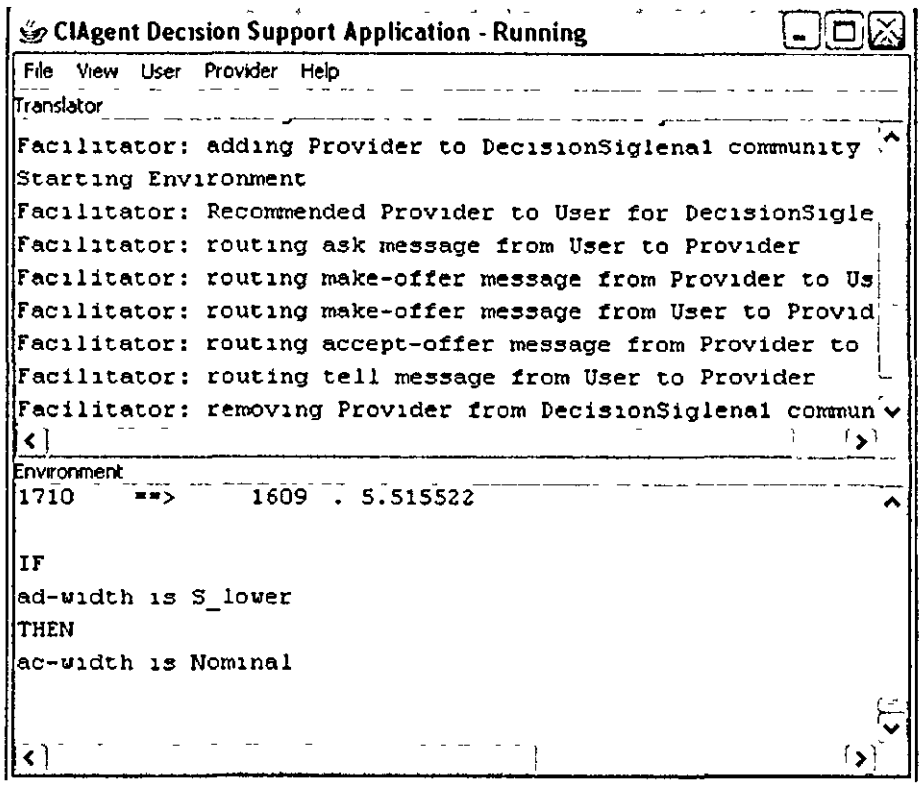


Figure 7.18. Information passed

7.6 Summary

The dynamics of the market are forcing shop control systems to enhance their intelligence and reactivity to changes. The control system must also have mechanisms to process the vast amount of data, which has been generated with the application of cheap data logging system. The shop floor control system must also incorporate a mechanism for continuous learning. Data mining provides a mechanism for analysing these data and generating new knowledge In this chapter, an intelligent knowledge based decision support system for shop floor control has been provided which incorporates data mining techniques and intelligent agent technology to provide useful information and knowledge. The functioning of the system has also been discussed The architecture of different agents that constitute the system has been discussed and the different modules required have been described A process to integrate the different data mining agent associated with different process has also been described. A prototype of the system has also been presented to demonstrate the application of this agent system.

Chapter 8

Data Mining on Product Dimension Data

8.1 Introduction

A shop floor control system helps in coordinating and managing the activities required for processing the production orders across the manufacturing resources It is responsible for selecting process routing, allocating resources, scheduling the work piece, downloading the process instructions, detecting and recovering from errors [238] The controller must be able to analyse real time data and present required information about the changes in the environment These information can be about changes in the parameter setting, manufacturing limitations, etc of the resources as the process routing is changed. Data mining tools and techniques can analyse the data and provide the required information for decision making.

The data mining enabled architecture has been presented in detail in chapter 7 and three examples will now be given of the types of data mining activity which could be undertaken by the data mining agent within this architecture The first example describes how to obtain knowledge through data mining actual product measurements during manufacture, or how other product related data, can be used to identify rules or trends which may be usefully employed or reused in the redesign of the product or similar products. The analysis of product data can also reveal the manufacturing constraints and design limitations of the product This knowledge can be stored in the knowledge pool and used in decision making in scenarios when process routing is changed. The details of various data mining algorithms have already been presented in chapter 6, and the application of any of these algorithms can contribute

towards achieving the goals of this research, since various reported examples exist in the literature to demonstrate the application and value of particular data mining algorithms on product related data, for example [254] and other examples given in chapter (5) and appendix . This chapter therefore presents the first example and provides guidelines for the application of data mining agent on product data to generate product related knowledge which can be reused.

Manufacturing industry relies on complex systems and machines. Operational efficiency, reliability and cost have all been improved by the skill of designers and the use of a variety of analytical, computational and manufacturing techniques Managers and designers may receive feedback from manufacturing operators as to what is effective and what is not so effective, but this tends to be based on anecdotal experience. The overall objective of the work presented here is to use emerging data mining techniques to provide this feedback in a more formalized manner which can then be used in a semi-automatic way after evaluation by human operators and show that this can then be linked to the data mining enabled architecture

8.2 Problem Overview

A typical product goes through many different manufacturing processes before being shipped out. A manufacturing system with different processes named as process 1 to process n has been represented as a block diagram (figure 8 1). The data (manufacturing parameters) may be recorded at each individual station or machine and the dimensions or quality of the product may also be measured after every important step

The example in figure 8.1 shows data being extracted from just two example processes. The earlier process generates data related to different parameters of the machines and the second data extraction point may be a metrology process, which measures different dimensions of the product, for quality control purposes Each manufacturing process is important and contributes to the required quality of the product but some manufacturing processes and their parameters may be more influential than others In a real, complex manufacturing system, it can be very difficult to determine the overall effect of particular parameters of a certain manufacturing process on the final quality of the product However, this is essential if the process is to be effectively controlled and the desired product quality and throughput achieved Further complexities exist due to any unknown (or unconfirmed) inter-relationships between processes and the inter-relationships between dimensions and parameters of the product.

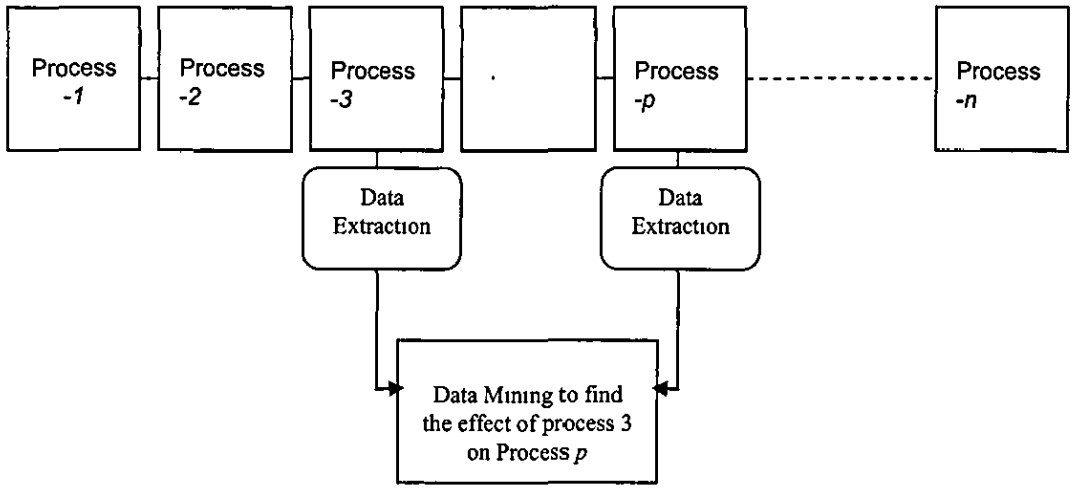


Figure 8 1. A sample block diagram of a manufacturing process show the flow of a product and data extraction

In this chapter, the data mining agent has been implemented on manufacturing data as part of a knowledge acquisition process This is done to determine how explicit knowledge can be extracted from existing databases, so that it can be used (1) to improve the design and production of the product and (2) to discover any constraints that might exist in the manufacturing system The example included here illustrates how a typical data mining algorithm could be applied on product related data using the data mining agent and how the knowledge outputs can subsequently be captured within the data mining enabled architecture for future reuse Although the example included here has been simplified in the interests of clarifying the processes involved, it is based on industrial case studies of a real, highly complex product The results obtained are valuable and have been confirmed by experts in this field. The full case study experiments, using several different data mining approaches were reported in [255] A simplified example from the case studies, demonstrating the application of Association Rules to manufacturing data, has recently also been published in [256]. This has been chosen as the basis for the first example discussed here, since the author and Shahbaz were jointly responsible for the novel contribution of applying association rules to manufacturing data

The ability to identify explicit rules which relate different characteristics (and therefore implicitly provide metrics of quality) from production data of manufactured products, has therefore previously been demonstrated and reported, Indeed other examples of successful data mining of product related manufacturing data can be found in appendix Hence, this research is based on the understanding that data mining can provide potentially useful and reusable discovered knowledge to improve the production processes involved However, all

such reported examples of successful data mining in manufacturing have been based on finding solutions of individual or "one-off" problems. A distinguishing aspect of this research is that it through the specification and design of the DMA as part of the data mining enabled architecture, a mechanism for ongoing knowledge discovery and exploitation is provided The discovered knowledge in this case study also provides an improved understanding of the inter-relationships of characteristics of the product, which can be used in the production of the product and also feedback into the design of similar products or the redesign or improvement of the current product. The improved understanding, based on experience, may also influence concept and embodiment stages of new product designs

8.3 Data Mining Agent

Product's output dimensions are a good measure of the quality of the production cycle and can help in suggesting any dependency or relation between different dimensions resulting from the manufacturing process or any alteration in the design. It is very important for designers and production engineers to understand the inter-relationships between different dimensions of the product, particularly when components have complex geometry and need to be manufactured to a high level of precision. If one dimension is positively related to another, i e improvements to the first also improve the second, then it may be beneficial to put resources into improving manufacturing accuracy to consistently achieve outputs within even smaller tolerance bands In contrast, knowledge that particular dimensions are negatively

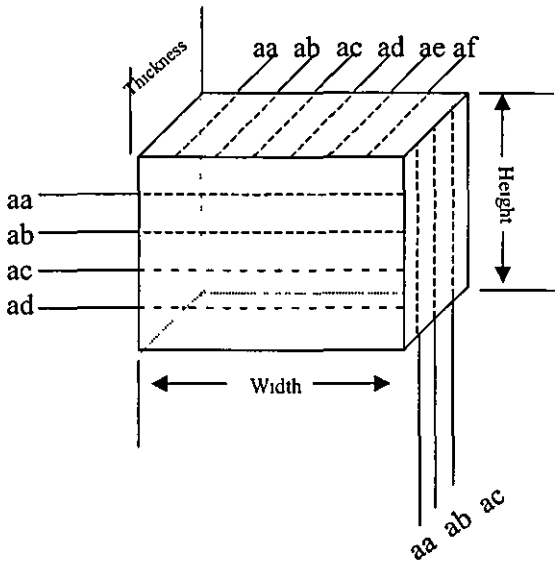


Figure 8 2 A sample product showing different dimensions at different sections

related can reduce waste. Since it would be detrimental to spend time improving accuracy on one dimension if this is also likely to make another dimension less accurate This type of lifecycle knowledge would also be valuable to designers on future projects and would provide vital information to shop floor managers in cases of re-routing. In this chapter, the data mining agent has been described for identifying the associations between different measures or dimensions of the products

In this section, the data mining agent is primarily discussed in the general context of how it can be applied to manufacturing product data. The whole process of knowledge discovery works in several stages including understanding the problem and process, data cleaning, data selection and transformation, data mining, pattern evaluation and knowledge representation. However, as previously shown in chapters 5 and 7, these steps can occur and reoccur in many different iterative cycles The different modules within the DMA have therefore been provided with particular functionalities to perform each of the individual stages of knowledge discovery as and when it is required The type of data cleaning and transformation required depends upon the data source and type. Product data as measured by metrological processes are continuous It is extremely difficult to manufacture two products of exactly the same dimensions Therefore, tolerances are assigned to every dimension. In the manufacturing process, it can only be predicted that if particular conditions are fulfilled then the output will lie in a certain range The data used in this process are therefore transformed into categorical form The data cleaning and data transformation stages are primarily discussed in the general context of how these stages should be applied on manufacturing data The data mining stage can include the application of one or more type of data mining algorithm, such as regression, clustering, etc. as previously explained, the data mining stage is presented in the context of the simplified example of a cuboid as shown in figure 8 2 The product data of this cuboid consists of dimensions of width, height and thickness at different sections. The product under consideration is not a perfect cuboid and its measurements vary at different sections across different dimensions The edges of the product are not parallel and also do not have exactly the same dimensions across length, width and height The dimensions of the cuboid are considered in this case study across and different sections of each dimension have been shown in table 8.2. An example of the types of data used for analysis in chapter 8 and 9 is shown in appendix II. It should be emphasised however that these techniques are included here purely as examples of the application of the DMA, since no single technique can perform best on all types of data and the data mining stage module of the data mining agent should be able to apply a variety of different data mining algorithms depending on their appropriateness for the data that is being examined Finally, pattern evaluation is discussed by examining ways of assessing the quality of the generated rules

	Thickness						Width				Height		
	aa	ab	ac	ad	ae	af	aa	ab	ac	ad	aa	ab	ac
D	7 55	7 5	7 45	7.45	7 5	7 55	7.55	7.45	7 45	7.55	7.55	7 5	7 55
+	02	02	02	.02	02	02	.02	.02	.02	02	.02	02	02
-	.02	.02	.02	02	02	.02	02	02	02	.02	02	.02	02

Table 8.1: Dimensions of different sections with tolerances. --- D: Dimension

8.3.1 Data Cleaning

The requirement of the data cleaning module is that it takes raw data, which may be in the form of one or more types of file, possibly containing errors (e g duplications, gaps, etc), and combines it to output well structured, consolidated, error-free tables in a data base The functionality of the Data Cleaning module is shown in the figure 8.3

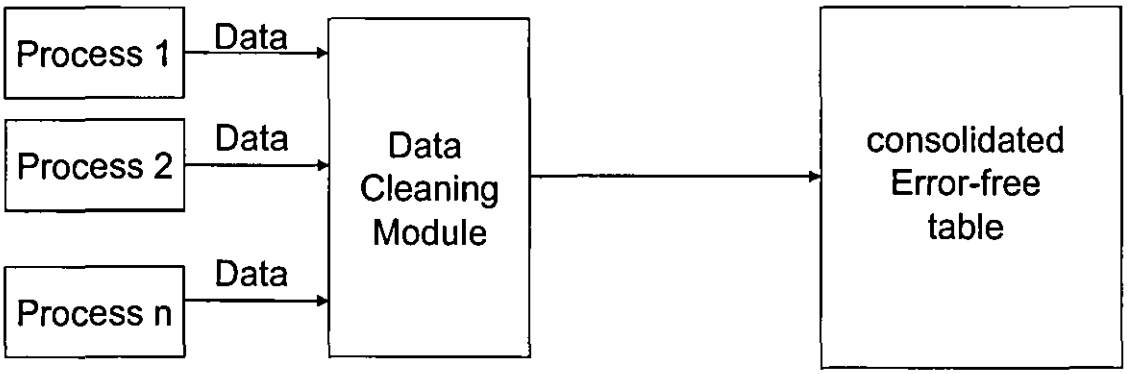


Figure 8 3 Data cleaning module

It has been reported that data cleaning is an important stage in the knowledge discovery process and consumes most of the resources [257] However, the amount of data cleaning that is needed largely depends on the type of raw data that is being considered Manufacturing data are recorded in several ways as manufacturing systems may have some manual data entry systems, analog and/or digital data acquisition systems and/or automatic data loading from Computer Aided Manufacturing (CAM) systems. If data is manually entered, or is collected at multiple points of different processes and needs to be combined and consolidated cleaning can be a complex operation for manufacturing data, since this data is likely to contain more noise and inaccuracies than other kinds of data, such as telecommunication data, market basket data from retail purchases, or data from web logs etc. It is therefore crucial that manufacturing data be cleaned carefully and thoroughly to make it ready for the different types of transformations that may be necessary before particular data

mining techniques can be applied However, if the data mining is to be carried out on data from a single source, such as an Enterprise Resource Planning system or from a previously consolidated data warehouse, the data cleaning is likely to be much simpler.

In general terms data cleaning involves the identification and if possible, the correction of irregular, duplicate or identical records. If the data cannot be cleaned it should be removed from the datasets that will be used for either training or testing in the data mining stage of the knowledge discovery, since poor quality records will adversely affect the quality and reliability of the data mined results. All changes and deletions should be carefully documented, as it may be necessary to refer back to them if any anomalies are found or problems occur during the data mining stage.

The first step in cleaning, if multiple files are being examined, is to join the data into a single record (or tuple) for each product This can only be achieved if each file contains key field(s) (or unique identifiers) for each product This can be a problem, particularly when data is collected from several different processes and stored as separate data sets or files. In such cases, key values may need to be matched across several different data sources and then consolidated in the form of a relational database In practise, this type of matching can reduce the number of data sets considerably. There are several reasons for this, including missing data (entries not fully keyed in by the operators), lost data, data not in electronic format (for example data stored in the form of ultrasonic or x-ray images or drawings), data without the main or primary ID, partial transformation of non electronic data into electronic information and confusing data with some kind of duplication or other error

When this consolidation has been completed, several other types of cleaning can be done, as summarised in the following list - ·

1 **Identical Records:** It is easy to identify and clean identical records Identical records have tuples that are exactly the same in all respects So it is simple to get rid of one of the duplicate tuples and it is not necessary to keep any kind of longer-term record of these errors having been corrected.
2 **Duplicate Records but with Missing Values:** Such records are also easy to identify but need a little bit of consideration in processing. The tuple with all fields filled with different values are kept as the original records but such records are noted down in a separate table for future reference so that they can be easily identified if any kind of unusual relationships that include them are found
3 **Confusing Records:** Examples of confusing records includes those with the same primary key but different values in the different fields Some times most of the fields

may be identical except for a few. Some records may have entirely different values in many different fields These types of records need to be checked manually with the data sources.

4. **Missing Records:** Missing records are those, which have no records in some fields of the table In such cases a decision has to be taken about whether or not to include the record in further analysis. For example if more than 25% of the values are missing from any individual record then that record could be excluded from analysis. But if less than 25% values were missing then those missing values might be replaced with the average value (or some other agreed typical value) of that particular dimension. These types of decisions need to be made in the context of the particular study and in consultation with experts who know the particular processes involved
5. **Noise Reduction:** After dealing with all the missing, duplicated, multiple and confusing records each attribute is checked for noise by plotting their scatter plots Scatter plots help to quickly identify the abnormal values that result due to any abnormality or keystroke errors. In most contexts such data values need to be removed from the data and recorded separately with any reasons, if any for the noise

8.3.2 Data Transformation

The data transformation module takes the data tuples from the consolidated data base and where necessary transforms it into a form suitable for input into the chosen data mining algorithm Data transformation is not always necessary in data mining generally. However as manufacturing data is often continuous, (e g. product dimensional data) and many data mining techniques (including Association Rules) work better on discrete data, data transformation often becomes a necessary stage It is also more appropriate to run some kinds of data mining algorithms with only a few divisions of the data since too many different values will result in very indistinct results It should also be remembered that knowledge discovery is best applied to "non-trivial" complex problems,[180] so if explicit relationships can be calculated between the relevant variables in other ways, it is generally not worth using data mining algorithms Therefore, transformation is necessary because it may be virtually impossible to discover relationships for exact, continuous, measured values The data should therefore be transformed into some appropriate, representative bands or divisions, before being used as input to the chosen data mining techniques Transformation is equally important when results have been determined and output from the data mining stage, as then, the results need to be translated back into the appropriate range of variable values, using a reverse transformation process In the case of manufacturing data the transformation stage requires detailed understanding of the manufacturing process, its constraints and operations, the

importance of particular dimensions of the products and the range of manufacturing variations that are likely It is therefore essential for the data mining expert to work with the manufacturing engineers or process experts during this stage Different data will need different kinds of transformation, however, experience gained on data mining case studies has shown that for continuous product measurement data, techniques such as drawing the simple distribution curve and finding the standard deviation can give a good guide for transforming the data into some suitably identified ranges.

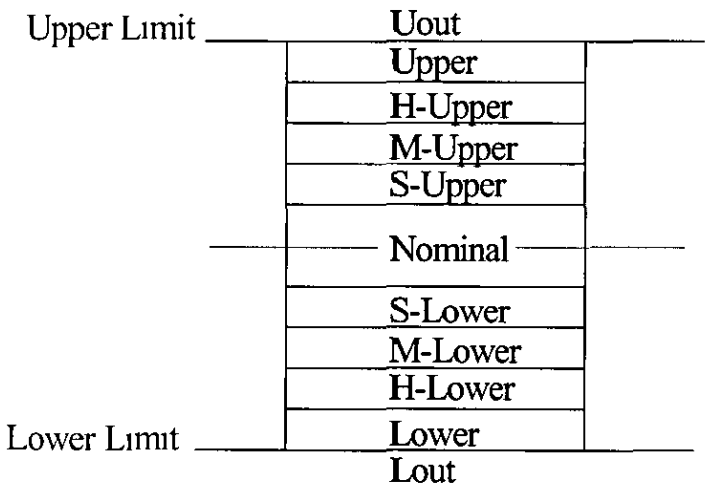


Figure 8.4. Difference sections/divisions for the output dimension of a product

In the current example, the approach adopted was to initially consider the current tolerance bands. The output dimensions were then divided into different sections from the upper engineering tolerance to the lower engineering tolerance as shown in figure 8.4. The reason for transforming the data into these ranges is that the manufacturing process will achieve these divisions for large batches of products, which is not the case for individual precise output dimensions

8.3.3 Data Mining

The requirement of the data mining module is that it applies an appropriate data mining algorithm on the previously cleaned and transformed data, which should now be in a single consolidated table. There are many possible algorithms which can be applied at this stage, for example, decision trees, clustering, rough set theory, or Association Rules algorithms. As previously explained, regression analysis, Association Rules technique and k-median clustering will be used to illustrate this example further

8.3.3.1 Regression Analysis

Different types of regression analysis have already been discussed in chapter 6 In this research regression analysis was used as one of the first data mining techniques to find the relationships between pairs of dimensional variables and help in predicting their trends The information generated through this analysis could then be exploited in other data mining techniques and also to control the different dimensions of the product by controlling one dimension In this implementation, both linear and non-linear regressions were used to determine any significant relationship between the variables. The existence of relationships between the variables could have generated a set of governing equation for controlling output dimensions Table 8.2 shows some of the regression results and shows the accuracy of the prediction result. A strong relationship would be shown by a value near 1. It was concluded from the table that regression analysis was not sufficient in this case to generate governing equations with high confidence

		Thickness						Height			Width			
		aa	ab	ac	ad	ae	af	aa	ab	Ac	aa	ab	ac	ad
Thickness	aa	*	.23	25	.37	41	11	.12	14	29	.51	25	36	.42
	ab		*	.28	.14	15	11	.17	.19	22	.27	29	31	.33
	ac			*	42	.34	32	.11	.12	09	.13	.17	.16	.18
	ad				*	.30	20	.21	16	.14	.11	19	.33	.41
	ae					*	11	.19	.16	.15	26	.37	44	.50
	af						*	21	.35	33	.26	.29	38	.40
Height	aa							*	.12	16	15	.25	36	.29
	ab								*	.14	13	.17	.19	.20
	ac									*	14	.17	.19	.28
Width	aa										*	28	27	.26
	ab											*	25	.24
	ac												*	.34
	ad													*

Table 8.2 Regression results (value of r^2)

8.3.3.2 Association Rule on Product data:

Association Rule algorithms were discussed in chapter 6. In this implementation of Association Rules the Apriori Algorithm was used, which is based on the property that states, "any subset of a large itemset must be large" [258] Here large means a defined support or occurrence level of a single or multiple items in the transactions. The Apriori principle, states that any subset of the frequent itemsets discovered having minimum support level would have the same or higher support level The Association Rules can therefore be found using all those subsets of the frequent itemsets. This section demonstrates the application of Apriori algorithm on product data using different dimensions on the cuboid, as shown in figure 8.2.

The width, height and thickness dimension of the cuboid need to be transformed into appropriate values The dimension at each section has an nominal value and tolerance. This tolerance band is then divided into 11 (or any other appropriate division as convenient) sections as shown in figure 8 4 Now the measured dimensions of all the sections are translated into the appropriate bands. The data from the simplified product section are shown in table 8 2, with appropriate dimensions

The transformation table for section "ab" of "Thickness" is shown in table 8 3 If a measured value of ab-Thickness equals to 7.509 then according to the transformation it will be translated as M-Upper. This transformation is necessary since if valid association rules are discovered, the manufacturing setup must be able to operate between limits instead of requiring exact measured values

If	Then replace the value with
Dimension>7.52	Uout
7 52>Dimension>=7 516	Upper
7.516>Dimension>=7.512	H-Upper
7 512>Dimension>=7.508	M-Upper
7.508>Dimension>=7.504	S-Upper
7.504>Dimension>=7.496	Nominal
7.496>Dimension>=7.492	S-Lower
7 492>Dimension>=7 488	M-Lower
7 488>Dimension>=7 484	H-Lower
7.484>Dimension>=7 480	Lower
Dimension<7 480	Lout

Table 8.3: Data Categorization

Apriori algorithm works well on numerical data. The product's data which has already been transformed into different categories is further transformed from strings into integer identifiers A simple nomenclature is defined to identify different output dimensions into integer identifiers.

8.3.3.2.1 Second Transformation:

The apriori algorithm used in this example works best on data which has integer identifiers The data was therefore transformed second time and the algorithm was applied to find the

frequent itemsets and then the association rules The integer identifier transformation matrix is shown in the table 8 4. Each integer identifier is a combination of two, two or three digit numbers (these numbers can be chosen according to the requirements). The first two digits show the dimensional band and the last two digits show the section of the product. For example, if the "ac" section of Width has a measured value in the S_Lower band, then that value will be translated as 1709.

		U-out	Upper	H-Upper	M-Upper	S-Upper	Nominal	S-Lower	M-Lower	H-Lower	Lower	L-Out
		11	12	13	14	15	16	17	18	19	20	21
aa_Thickness	01	1101	1201	1301	1401	1501	1601	1701	1801	1901	2001	2101
ab_Thickness	02	1102	1202	1302	1402	1502	1602	1702	1802	1902	2002	2102
ac_Thickness	03	1103	1203	1303	1403	1503	1603	1703	1803	1903	2003	2103
ad_Thickness	04	1104	1204	1304	1404	1504	1604	1704	1804	1904	2004	2104
ae_Thickness	05	1105	1205	1305	1405	1505	1605	1705	1805	1905	2005	2105
af_Thickness	06	1106	1206	1306	1406	1506	1606	1706	1806	1906	2006	2106
aa_Width	07	1107	1207	1307	1407	1507	1607	1707	1807	1907	2007	2107
ab_Width	08	1108	1208	1308	1408	1508	1608	1708	1808	1908	2008	2108
ac_Width	09	1109	1209	1309	1409	1509	1609	1709	1809	1909	2009	2109
ad_Width	10	1110	1210	1310	1410	1510	1610	1710	1810	1910	2010	2110
aa_Height	11	1111	1211	1311	1411	1511	1611	1711	1811	1911	2011	2111
ab_Height	12	1112	1212	1312	1412	1512	1612	1712	1812	1912	2012	2112
ac_Height	13	1113	1213	1313	1413	1513	1613	1713	1813	1913	2013	2113

Table 8.4: Integer Identifier Transformation Matrix for Cuboid Example

The complete set of dimensional data for the manufactured product is therefore transformed in this manner into integer identifiers, and is then treated as one transaction (i e forming one record, with multiple fields, in a database table) The whole training dataset is then fed into the association rule algorithm programme This requires the minimum acceptable support level to be selected so that the frequent itemsets can be identified in the data. Support is defined as the minimum number of occurrences of the itemsets in each stage of the iteration of the algorithm. If the support level is set too high, there is a chance that some important items in the frequent itemset data could be missed, which could lead to some valuable relationships being missed and remaining hidden By decreasing the required level of support too far, too many "low quality" frequent itemsets may be found resulting in numerous association rules, which may mostly be of little value.

It is important that mining decisions are made in the context of the physical realities of the data that is being examined When data is the result of very highly controlled manufacturing processes, unusual combinations of results are particularly interesting as they may represent occasional "problem" situations Hence, in this research, it was considered important not to risk missing possibly valuable hidden knowledge, simply because a particular combination of

manufacturing outputs only rarely occur Therefore the decision was taken to keep the support level very low, and the best possible solution found was to calculate the support level by counting the occurrence of each integer identifier in the whole data and then setting the support level to equal the minimum of these counted occurrences.

Association rules were then generated from each of the frequent itemsets. Each of the frequent itemsets was split into all the possible subsets, and rules generated using these subsets, in the form subset x -> subset y. A confidence level was then calculated for each rule based on the number of times the set (subset x U subset y) occurred in the original data The minimum acceptable confidence level was selected for the analysis. When the association rules were generated, the certainty of each of the rules was tested against the minimum acceptable confidence level. If the confidence of any rule was less than the defined level the rule was discarded otherwise it was kept in the final output

In initial tests of the method, the association rule algorithm generated frequent itemsets and thousands of rules. Each rule had its confidence level, which was equal or higher than the defined threshold value When the rules were checked against the manufacturing process output some interesting facts were identified and these results are discussed in the next section There is an important issue about the quality of the generated rules, and it is therefore very important to find the validity of each of the generated rules before time is spent analysing all the output, as this can be a very long, tedious and time consuming task. It is therefore advantageous to quickly dispose of any rules that are actually misleading or simply not valid

8.3.3.3 Rule Quality

Rules should be eliminated which are not statistically valid. This module takes the different rules that have been generated, tests them using some statistical analysis and removes the rules which cannot be validated. Rules generated with very high support and confidence level are less likely to be misleading than rules generated with lower support levels But these two measures can only partially help in detecting useless rules The degree of association needs to be measured for both side of the rules to determine its usefulness. This can be illustrated with the example data shown in table 8.5. The data shows fifteen transactions or products, which have been carefully chosen for illustration purposes only.

Consider the following valid rules

1 1612 → 1507

(IF ab section of 'height' is nominal THEN aa section of 'width' is s_upper.)

2: 1612 → 1303

(IF ab section of 'height' is nominal THEN ac section of 'thickness' is h_upper)

Where,

Nominal: 7 496 –7.504

s_upper. 7.504 – 7 508

h_upper. 7.508 – 7 512

When the above rules are checked against the data, both seem to be valid, the first rule having 40% support and 100% confidence and the second rule also having support of 40% and 83 3% confidence

ID	Data
1	1612, 1303, 1507, ...
2	1703, 1303, 1703, .
3	1611, 1303, 1602,
4	1612, 1303, 1507, ...
5	1408, 1303, 1404, ...
6	2106, 1303, 1703,
7	1408, 1703, 1504, ...
8	1603, 1303, 1609, ...
9	1612, 1303, 1507, .
10	1312, 1303, 1803, ...
11	1713, 1303, 1601, ...
12	1612, 1404, 1507,
13	1612, 1303, 1507, ..
14	1401, 2006, 1507, ...
15	1612, 1303, 1507, .

Table 8 5 Example Data

The first rule is a valid rule because 1507 and 1612 complement each other in the data, whilst the second rule is misleading as 1612 is complementing 1303 but 1303 does not complement 1612 Examination of the data shows that 1303 appears several times even when 1612 is not present so 1303 is independent of 1612 This situation clearly leads to the deduction that there is an association present in the data but its strength is uncertain. An association is strong only

when both sides of the rule come together and neither element appears elsewhere in the data independently. The absence of one variable in the rule or too many appearances of one side of the rule makes it less important.

A popular technique for finding the correlation or the significance of the rules is the chi square test Chi square is popular for finding the correlation of bi-variant data in the statistics community. A higher value of chi square for the bi-variant (both 'if' statement and 'then' statement) shows a strong relationship and a lower value indicates a weak relation. The strength of the relations can be found using the chi square table under the specific degree of freedom (1 in this case). In the above quoted example the chi square values of first rule was 11 42 and for the second rule it was only 0.5113 showing a confidence (strength) of more than 99% in the first case and in the second case of about 50% These results show that even though the confidence of the second rule was very high, when the quality of the rule is checked it resulted in a very poor rule, which is not significant.

8.3.3.4 Result and Discussion

The importance of "support" and "confidence" levels was discussed in section 8 4 3 2 1 and the effects of varying these levels can be seen for example in the following table produced from tests run on a sample batch of 2200 test records (after cleaning)

Set Number	Support	Confidence	Total Rules
1	20	80	35319
2	40	70	2077
3	40	80	1762
4	50	70	805
5	50	80	619
6	60	70	436
7	60	80	325
8	70	70	220
9	70	80	175

Table 8 5: Statistics of different analysis runs.

However, experience on case studies and discussions with manufacturing and production engineers has lead to the conclusion that in data mining studies using manufacturing data, a very low support level (e g 20%) should be chosen This is so that possibly important rules

and relationships that only occur infrequently in the manufacturing data are not missed. However, as can be seen from table 8.5, this is likely to result in large quantities of rules being generated In this simplified cuboid example, there are 13 different dimensions considered and each can occur in 11 different categories (since there are 11 bands between the upper and lower tolerance levels shown in figure 8.4). In the results that are possibly generated by the application of the data mining association rule algorithms there are therefore 13 x 11 unique elements that can occur in different groups or combinations in the various different rules of types similar to those discussed in section 8 4.3.3.

It is also important to consider that this is in fact a small quantity of unique elements (and consequently lower number of potential rules) than is likely to occur in case studies with real products For example the fan blade product shown and discussed in [256] was reported as containing 25 attributes and sections in 11 different categories, making 25 x 11 unique items which were reported to appear in different groups or combinations within over 20,000 types of rule. It is clearly therefore a major task to reduce this huge set of rules into a core set of useful rules which can benefit the manufacturing operation. It is also important that methods are found to perform the necessary reduction quickly and efficiently.

In this research, the method used to extract useful, valid rules from the total sets of rules generated is -
Apply chi square testing with a high confidence requirement (e g 95% or 99% depending on the context of the study) In [256] it is reported that this reduced the set of useful rules to nearly half of those originally identified.

Extract rules that display particularly interesting or useful forms of relationships, for example "one-to-one" relationships It may then be possible to logically reason whether the resulting individual types of rule are useful and worth keeping or whether they are unlikely to add value to the current knowledge pool

One to one relationships are considered to be a particularly useful form of relationship in manufacturing contexts as they are very important in identifying any design constraints on the product or any kind of manufacturing process limitation. One to one relationships which indicate that changes in one element result in improvements in another element are encouraging to identify (since this is likely to be the desired or anticipated behaviour of the product or process), but they are unlikely to therefore add any additional useful knowledge for the knowledge pool. However, one to one relationships which indicate that changes in one

element result in adverse effect on another element are likely to be much more interesting and important, potentially adding very useful additional information for the knowledge pool.

For example, such rules could indicate that one particular 'Nominal' dimension corresponds to another dimensional band being non-Nominal This type of result requires more attention, and as well as provide useful knowledge for the shop floor control of the manufacturing of the current product It provides valuable feedback for the design process, as such rules can help in redefining design constraints.

The extracted rules should always be tested against new production data to see if the identified relationships remain valid in the new test data. The test results indicated that strong rules with high chi square values still hold in the new data with similar confidence of chi square index.

It is important to note that relationships identified in data mining applications of the type discussed in this chapter can indicate two important aspects of manufacturing The first is where a design error may exist as the relationship shows that naturally the two correspondent dimensions do not have a 'nominal-to-nominal' relationship Clearly particular care must be taken in the precision to which these dimensions are manufactured on the shop floor in order to prevent waste (and detrimental) work being done. The other important aspect could be identification of errors, faults or limitations in the manufacturing process. Identification of these types of problem might need more careful reconsideration to improve manufacturing practice and strategies in order to remove any related faults The obtained information about manufacturing and design limitation provide information about the different constraints that can be encountered in case of re-routing and the shop floor managers must be informed to take action to offset them.

8.3.3.5 Clustering on Product Data

Clustering algorithms are another popular type of data mining algorithms and were discussed in chapter 6 Clustering techniques were also applied on the data discussed in this chapter. In this implementation of clustering, k-median algorithm and Minkowski distance was used for finding cluster in the data The reasons for applying clustering was to find any natural grouping in the data The data obtained after second transformation was fed to the clustering algorithm program. This requires the number of clusters to be selected so that clusters can be obtained The selection of number of clusters depends on the resolution with which the data is accessed However, the selection of this number is a tough task In this research, different

numbers of clusters were used The application of the clustering algorithm resulted in a mixture of clusters having different class (band) variables spread everywhere in those clusters Uniform distribution of class (band) variables indicated that no natural grouping of data existed in it. The high number of unique elements discussed in the previous section meant that clusters that were obtained were difficult to visualise as there are 13 dimensions in the cuboid This illustrates that clustering was not suitable to generate information from this type of data The major cause for its failure were high dimensionality and interpretability

8.4 Novelty and Contribution

The work and experiments described in this chapter contribute substantially to the aim and objectives of this thesis. A methodology for the application of data mining techniques has been presented. The application of the proposed approach on product data helped in generating valuable knowledge about design constraints, manufacturing limitations and capabilities. It helps in providing the content of formalised feedback from the manufacturing to design The results also indicate that the design knowledge which are gained through experiences can be generated from data This indicates that experts are not the only source of information The novel aspects of this work can be summarised as follow

- The proposed methodology provides techniques for systematic enterprise knowledge discovery particularly focusing on data cleaning and transformation requirements
- A methodology of providing information to shop floor managers about some constraints to be considered in case of re-routing of raw material
- A process of generating content of formalised feedback to design stages from manufacturing
- A process to update the knowledge pool. The updated knowledge can be reused by different sections of the enterprise.

A novel methodology for the application of Association Rules on manufacturing data, was jointly identified and reported by the author (Srinivas) and M Shahbaz, and this has been discussed in chapter 8 and has been recently published in [256] The example used to explain the methodology in chapter 8 concentrates totally on product data. It is very important to understand the relationships of different product parameters since, for example (as shown in this chapter), attempts to improve one dimension or feature of a product during manufacture could result in either improvement or alternatively have detrimental effects on one or more of the other dimensions or features of the product. However, this is only one source of potential variation and other important sources of variation in product quality are the manufacturing

processes and sub-processes themselves. Chapter 9 will therefore consider how the DMA can be used on a combination of product and process data.

Chapter 9

Data Mining Agent on Process Control Data

9.1 Introduction

This chapter will examine the use of a DMA on operational data coming from more than one different sources, in particular from the product (through metrology processes) and from the manufacturing processes (through input or set up variables).

This type of knowledge is very important in the context of shop floor control as

1 Decisions made on the input or set up variables for individual processes will affect the overall quality of the product
2. Knowledge of the current state of a particular product and understanding of the influence that particular variations in process input parameters can have on that state, are likely to influence routing or re-routing decisions when work is required to bring a particular feature or dimension of a product into specification range (see section 9 2)
3 Existing Process control software should be able to interpret the knowledge obtained to generate the process input automatically.
4. Knowledge about the relationship between product quality and process variables enhances the process monitoring capability This knowledge should enable compensation to be provided (if possible) for any degradation in a proactive manner and it would reduce re-work, scrap, lead time and increase productivity.

It is important therefore to consider the application of the DMA on combined product and process data in order to improve decision making for manufacturing, shop floor control as this

is concerned with effective and efficient utilisation of the resources at the lowest level of control in manufacturing facilities The extraction of knowledge through the DMA can be more beneficial in cases where a clear relationship of how variables affect the quality of the product is difficult to achieve Shop floor control involves the co-ordination of physical items as well as information. In order to keep the manufacturing system at its peak efficiency and to take decisions on process levels based on the changes in the system dynamics, the shop floor control system must be able to continuously process recent operational data and provide enhanced information and knowledge about it to update the contents of the knowledge pool

Drives to increase quality are important in the context of this research because this generally implies increases in achieved precision, and therefore the cost of improved quality increases as engineering tolerances are reduced and become more difficult to achieve Manufacturing engineers work hard to address this issue by using their engineering skills and statistical tools. There are many possible reasons for poor quality, including possible faults in the design of the products or limitations of the manufacturing machinery or errors by the work force. Hence, quality problems can be the result of any one or a combination of these or other more obscure causes, and it may be impossible to detect most design or manufacturing limitations with traditional statistical tools

In the past, manufacturing enterprises have used several quality assurance tools successfully to improve the quality of their products and production rates. Statistical process control (SPC) is extremely useful in maintaining process stability by providing a methodology for measuring process capability and performance. Control charting is the key part of SPC implementation The power of a control chart lies in its ability to separate out special disturbances from inherent variability in the process. However, SPC can only detect process abnormality but cannot provide remedies and feedback to adjust the input variables to compensate for the difference between target and measured value Data mining techniques can explore the process data and provide the required information

This chapter presents a generic methodology to implement a DMA for identifying process and product relationships in a manufacturing environment The methodology generates control signatures from the partitioned process data that results in a particular product output class. This methodology helps in improving the manufacturing process and the product quality by determining the process controlling variables that result in a particular output class where the class can be any section between the engineering tolerances (as explained in the previous chapter) and providing information for the compensation. This methodology also contributes substantially to current approaches to quality assurance as it supports improvement in

manufacturing processes and product quality by determining the process controlling variables that result in particular output classes and providing the feedback information to SPC process for compensating the difference in measured and target value An additional benefit is that it provides a useful alternative to the expensive and time-consuming classical and full factorial experimental design approaches in this context. It also offers alternative yet complementary techniques to modern Design of Experiments (DOE) approaches where it provides particular benefits in the early stages of screening experiments

9.2 Problem Overview

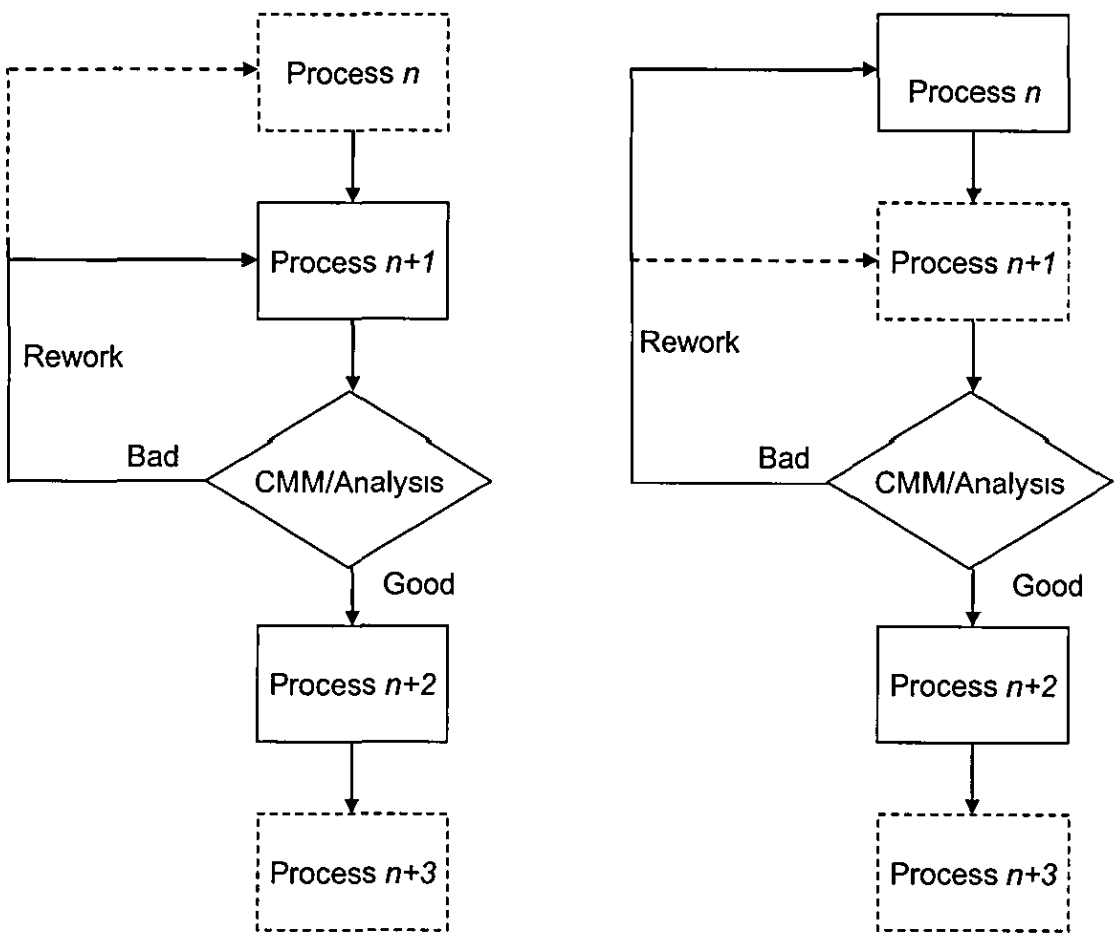


Figure 9 1a(left), 9.11b(right): Manufacturing Process flow diagram

The aim of all manufacturing processes is to produce quality products according to the design and within the acceptable engineering tolerances High quality manufacturing processes try to produce perfect products by constantly reducing the number of errors or faulty products manufactured This generally means that they aim to output products with dimensions and other characteristics as close to the specified nominal value as possible, so that the range of

the distribution of manufactured parts all fall close to the specified nominal value and well within the specified tolerances Thus the engineering tolerances are often squeezed to produce products as precisely as possible. Manufacturing process parameters may therefore be adjusted or changed regularly to accommodate variations or deterioration in the manufacturing process equipment, and to consistently obtain the desired product quality as output

A block diagram for a typical manufacturing process flow diagram is illustrated in figure 9 1 During manufacture, the product is likely to pass through many different processes. The manufacturing input parameters may be recorded at each individual station or machine It is also likely that the dimensions or quality of the product may be measured at several stages during manufacture (as implied by the CMM / Analysis box in figure 9.1) however it is unlikely that the product will be measured after every step of manufacture. The effect of each different manufacturing process is important to the quality of the product, but some of the processes will be more influential than others, and indeed particular dimensions or other characteristics of the product are likely to be affected more by some processes than by others. This is illustrated by figure 9 1, since figure 9 1(a) shows a case where if an error is found in a particular product characteristic (dimensionA) then in most cases the product should be rerouted to return to process n+1. However, if an error is found in a different product characteristic (dimensionB), then figure 9 1(b) shows that in most cases the product should be rerouted to return to process n. These types of relationships between product characteristics and particular sub-processes may only exist as tacit knowledge in the experienced operators of the processes but this chapter will show that data mining may be used to identify them. The bold line in figure 9.1 indicate the path through which the product may be re-routed to return for rework on it. E g in figure 9.1 (a), the product will return to process n+1 while in figure 9.1(b) it would return to process n

Several different statistical techniques are commonly used to check the quality of the current products going through the production resources at that time (e g Statistical Process Control (SPC) etc) However, traditional SPC systems only detect process abnormality (Baliga [259]), they do not have a feedback controller to recommend how the input process parameters should be adjusted to offset the variations in output. The control signature generated with historical data (identified through data mining) could help in providing such information.

Another common way of generating process knowledge is through the design of experiment process [260, 261]. However, as Schmidt and Launsby [261] state, unless experimenters are well informed and their experiments are well designed, results can be misleading and large

amounts of resources can be consumed unnecessarily In DOE different manufacturing variables are chosen with certain levels to produce sample products to determine the optimal manufacturing conditions Conventional, full factorial experiments in particular are an expensive approach to obtain the desired results when multiple variables are involved There is also always a possibility that the results may fail due to improper selection of the manufacturing process parameter and/or unidentified sources of noise from the machines. Hence, thorough understanding of the manufacturing process and careful selection of appropriate quality variables are critical to good experimentation [261]. The methodology introduced here could greatly enhance this understanding and support the variable selection process

9.3 Data Mining Agent

The importance of data oriented knowledge discovery techniques cannot be denied in any industry especially those recording their process and product life cycle information. Data collected from manufacturing and metrology processes could be a source of hidden knowledge about the process limitations, process improvements, design constraints, product quality and the deterioration of the manufacturing process machinery etc. The idea of finding patterns in manufacturing is not new, as shown by the many examples in chapter 5 In this section, the DMA is primarily discussed in the general context of how it can be used on manufacturing product and process data to determine relationships between them. Each of the different modules of the DMA, i.e. data cleaning, data selection, data transformation, data mining and pattern evaluation are now discussed in turn, although as previously shown in chapter 6 and 7, these steps can occur and reoccur in many different iterative cycles. The type of data cleaning and transformation depends upon the data source and type, algorithm selected and information required Product and process data is generally continuous and it is extremely difficult to control manufacturing parameters and dimensions to an exact value Therefore, tolerances are assigned to every dimension and manufacturing variable In the manufacturing process, it can only be predicted that if particular conditions are fulfilled then the output will lie in a certain range. The data used in this process are therefore transformed into categorical form. Though in some algorithm applications untransformed data have been used The data cleaning and data transformation stages are primarily discussed in the general context of how these stages should be applied on manufacturing data It should be emphasised however that no single technique can perform best on all types of data and the data mining module of the DMA should be able to apply a variety of different data mining algorithms depending on their appropriateness for the data that is being examined. Therefore the

selection and application of different data mining algorithms has been in the context of manufacturing product and process data Finally, pattern evaluation is discussed by examining ways of assessing the quality of the generated rules

9.3.1 Data Cleaning

The requirement of the data cleaning module is that it takes raw data, which may be in the form of one or more types of file, possibly containing errors (e g duplications, gaps, etc), and combines it to output well structured, consolidated, error-free tables in a data base Data cleaning is important in order to remove all the records that can produce errors or problems during the data mining algorithm application stage. The importance and functioning of data cleaning has been discussed in section 8 4.1. Manufacturing engineering data commonly contains duplications and confusing records There are also many cases where one or more data values are missing within a record which requires special attention Duplication in records may be the result of reprocessing or rework when a process does not fully complete properly in the first instance and therefore needs restarting, resulting in duplications of the records. Duplications can also result when the operator enters information twice either by unintentionally pressing an information feed or enter button twice or more often due to the malfunctioning of any of the manufacturing machinery

In the reported research two different methods were adopted to clean confusing records

1. If confusing records are found in the manufacturing process data then the earlier entry of the confusing record is deleted and the later entry in the time domain is accepted as the genuine record. The reason for this is that an earlier record is often due to an incomplete process or the process being stopped during the manufacture of that particular product. In such cases the later or final entry indicates the successful completion of the process (unless otherwise stated in the data) and therefore is the correct representative of the process.
2 If the confusing records are found in the metrology data where the dimensions of the product or the quality of the product is recorded then the earlier records are kept and the later records are discarded. The earlier records are actually the true representations of the previous step of the manufacturing process whereas the later records is likely to be the result of rework on the product. Such reworks may not be documented or may be documented in a different way to the normal production data and therefore in data mining or information hunting process these records do not have any kind of relationship with the previous work done on that particular product.

The data cleaning module of the DMA should be provided with the above knowledge and knowledge from section 8 4 1 and should also be provided with training or experiences that are gained through the application of data mining in the particular contexts that are learnt to be useful for the particular manufacturing company

9.3.2 Data Transformation

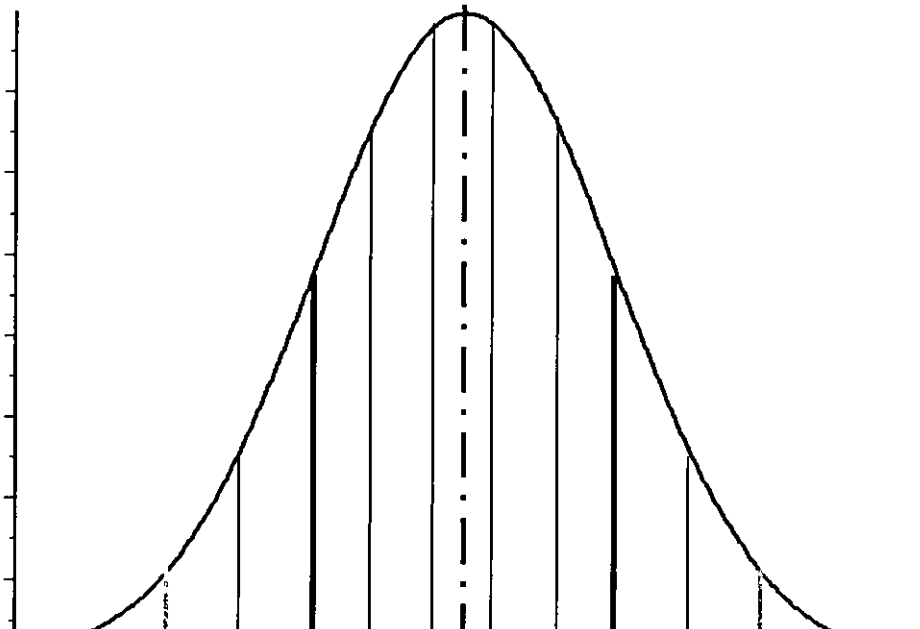


Figure 9 2 Normal Distribution Curve for Process data

The importance and functioning of the data transformation module has already been discussed in section 8 4 2. This module takes the data tuples from the data base and transforms them into a suitable form as required by the data mining algorithm This may include for example converting continuous data into suitable discrete "bands" or "ranges".

9.3.2.1 Process Data Transformation

Since transformation of product data was considered in Chapter 8, this section will only examine the different data variables collected from the different manufacturing processes through which a product passes (figure 9 1). There are many different possible types of data variables, including for example temperatures, average pressures, machining speeds or time

etc These values can be continuous or discrete and there are millions of possible combination sets of these values Therefore finding a relationship for these process parameters' values with output dimensions can be complex These manufacturing data will generally have normal distributions. In cases where most of the values lie near the average, the data transformation band or ranges should be made finer near the centre and can be slightly coarser nearer to the edges of the normal distribution curve as shown in figure 9 2 hence the ranges can be split into a number of bands or different sizes to suit the particular manufacturing context, for example in this research experiments were made with the data values from the centre up to the 1σ range on either side being divided into five ranges named as 04, 05, 06, 07, 08. The next set of data values, which are between 1σ and 2σ from normal value have been divided into four ranges named as 02, 03, 09, 10 The data beyond 2σ on either side has been consolidated into two ranges 01 and 11 as shown in the figure It must be emphasised that the division of data can be varied into any desired number of partitions depending on the context, requirements and distribution of data.

The categories or bands of the transformed data each form a discrete value, so if six process variables were chosen and named as 51, 52, 53, 54, 55, 56 A data matrix in table 9.1 shows how the particular values of these process variables could be translated into integer identifiers which are suitable for several types of data mining algorithms

Band	PVar1	PVar2	PVar3	PVar4	PVar5	PVar6
	51	52	53	54	55	56
01	5101	5201	5301	5401	5501	5601
02	5102	5202	5302	5402	5502	5602
03	5103	5203	5303	5403	5503	5603
04	5104	5204	5304	5404	5504	5604
05	5105	5205	5305	5405	5505	5605
06	5106	5206	5306	5406	5506	5606
07	5107	5207	5307	5407	5507	5607
08	5108	5208	5308	5408	5508	5608
09	5109	5209	5309	5409	5509	5609
10	5110	5210	5310	5410	5510	5610
11	5111	5211	5311	5411	5511	5611

Table 9 1. Data Transformation Matrix for Process data

The data transformation matrix shows the transformed values in the form of integer identifiers for each of the data bands for all six attributes of the process. In this transformation, the data has been divided into 11 ranges (as described above), making a total number of 66 discrete data groups. This type of transformation results in a much smaller number of combinations than would have occurred using the original recorded data values.

The data transformation technique described above can be varied into any number of possible sub-ranges If the normal distribution curve is very steep then fine slicing, using a greater number of bands near the centre (or average value) is required whereas in case of a smooth distribution curve equally spaced limits are adequate for the transformation In some case these values may also be normalised.

9.3.3 Data Mining Algorithm

The requirement of the data mining module is that it applies an appropriate data mining algorithm on the previously cleaned and transformed data, which should now be in a single consolidated table There are many possible algorithms which can be applied at this stage, for example: regression, decision trees, clustering, rough set theory, or Association Rules algorithms. A major challenge in data mining is choosing which algorithm (or type of algorithm) should be applied in a particular context as only a limited amount of guidance is available in the published literature, as shown in chapter 6 Machine learning techniques are generally used in data mining if the data does not follow a set pattern Moutkais et al [202] surveyed different machine learning techniques which are used to solve different type of problems. The most practical (machine learning) techniques used in data mining is classification ([219]). The aim of this task is to build a classification model from the stored data and the model can then be used to classify unclassified data There are many approaches for the classification method The survey of Moutkais et al [202] suggest that decision tree, k-nearest neighbourhood and association rules may perform better in classification problems (figure 6 5). However, it should be noted that no single algorithm performs best for all problems The performance of the algorithm will vary on the task performed and data used.

During this research, several algorithms have been tried on real manufacturing data to better understand their use and similarity or differences of application. In the following sections some personal observations are made about the application of certain algorithms in the context of manufacturing product and process data

9.3.3.1 Regression Analysis

Different types of regression analysis have already been discussed in chapters 6 and 8 of this thesis As regression is quick and easy to apply, it is worthwhile to test whether this type of relationship can be found between pairs of variables, to relate individual process input parameters with particular product characteristics or dimensions. If such relationships can be found it is useful as it may mean that explicit mathematical rules can be used or that the

number of variables to be examined can be reduced, thereby simplifying some of the future data mining experiments However such simple relationships have seldom been found during this research.

9.3.3.2 Decision tree

Decision tree algorithms were discussed in chapter 6 These algorithms are popular and have widely been used in literature for data classification ([219]) The algorithms classify the predicted class based on the different attributes, so in the context of manufacturing process and product data, they could be applied to determine the range of one or more process variables that result in different classes of precision in manufacturing the product. Rules are generated from the classification and can be used as a set of governing rules for controlling the process If the rules are going to be used to improve the performance of a manufacturing process, and control production, it is very important to validate the rules It is therefore recommended that part of the manufacturing data be used as training data and other (or subsequent data) be used as test data with the product data being used as the classifying variable.

The nodes in the decision tree represent the attributes (process variables) that are being tested for classification in the data (product quality) The outgoing branches of a node correspond to all the possible outcomes of the test (different range of process variables) at the node. The nodes and branches present a linear combination of attributes that can be used in decision making and classifying the other unclassified data. Decision tree can be also be re-expressed as an ordered list of IF-THEN rules, which are easy to comprehend. These rules can be used to update the knowledge pool and employed by the process controller to make correct decisions The rules generated can also be translated as a simple SQL query and help in understanding complex domains. These can be easily interpreted by a process monitor and also incorporated in it for decision making.

The problem found when applying decision trees was that when complex manufacturing data with several process variables, from well controlled manufacturing processes was used, the rules generated from these algorithms were complicated and the depth of tree was up to 20 (if all the examples are taken in the training dataset) levels with hundreds of nodes. The main reason for this could be complexity of data or the complex interactions between variables or the precision of the current manufacturing process making it difficult to classify Error rates in

the classification were also too high to generate a definite range in process data leading to a specific class in product data

The algorithms were initially tried on transformed product data and the original continuous process data, and then on transformed process data However, the application of the decision tree algorithms on transformed process data did not improve the results

9.3.3.3 Clustering

Another way of grouping the manufacturing data would be by applying clustering algorithms which were discussed in chapter 6 Clustering is used to detect any natural grouping of data which may result in specific output The algorithm can be run with a minimal requirement of domain knowledge and can be performed on diverse data type. In this research, a clustering algorithm was used to determine if any natural grouping of input parameters existed, which leads towards a specific output class. The objects in the same clusters have very high similarity and are dissimilar from the ones in other clusters based on certain measures. These similarity measures and the algorithm can be used to determine any grouping in the process variable and product quality. The cluster thus generated can be used to classify new data instances The new data member will belong to its nearest cluster based on the similarity measure These clusters and similarity measure can be incorporated in the process monitoring system to make an effective decision

In this application, k-mean clustering was used on clean process data and transformed products data (based on figure 8.4 of chapter 8) The datasets were dissected with respect to number of classes present in them Each portion of data corresponds to a particular class and hence a cluster if found would lead towards the specific class. Each data is then clustered in n dimensional space (n here corresponds to number of process variables) The shape of the cluster was determined by the distance of each member from the centre of the cluster. In the application of k-mean clustering, determination of k is a tough task and in this research different values of k were tried to group the data. Each point in the data set was assumed to belong to its nearest cluster but because of the complexity of the data and the interaction of the variables, the clusters were very complex multi-dimensional shapes, and different choices of the value k resulted in quite different clusters This was not promising and when the results were tested on a subsequent batch of manufacturing data a high level of inaccuracy was seen It is believed that the main reasons for the failure of these algorithms in the tests may be the effect of the complex interaction of manufacturing variables (which is likely to exist in complex manufacturing processes or combinations of manufacturing processes). It

must also be noted that it is difficult to visualise the space of n dimensions so these techniques may not be easy to apply on manufacturing data with multiple variables.

9.3.3.4 Association Rule Mining on Product and Process Data

The Association rule algorithm was discussed in chapters 6 and 8 and a methodology for its application on product data was shown in section 8 4 The apriori algorithm has been utilised to determine product and process relationships.

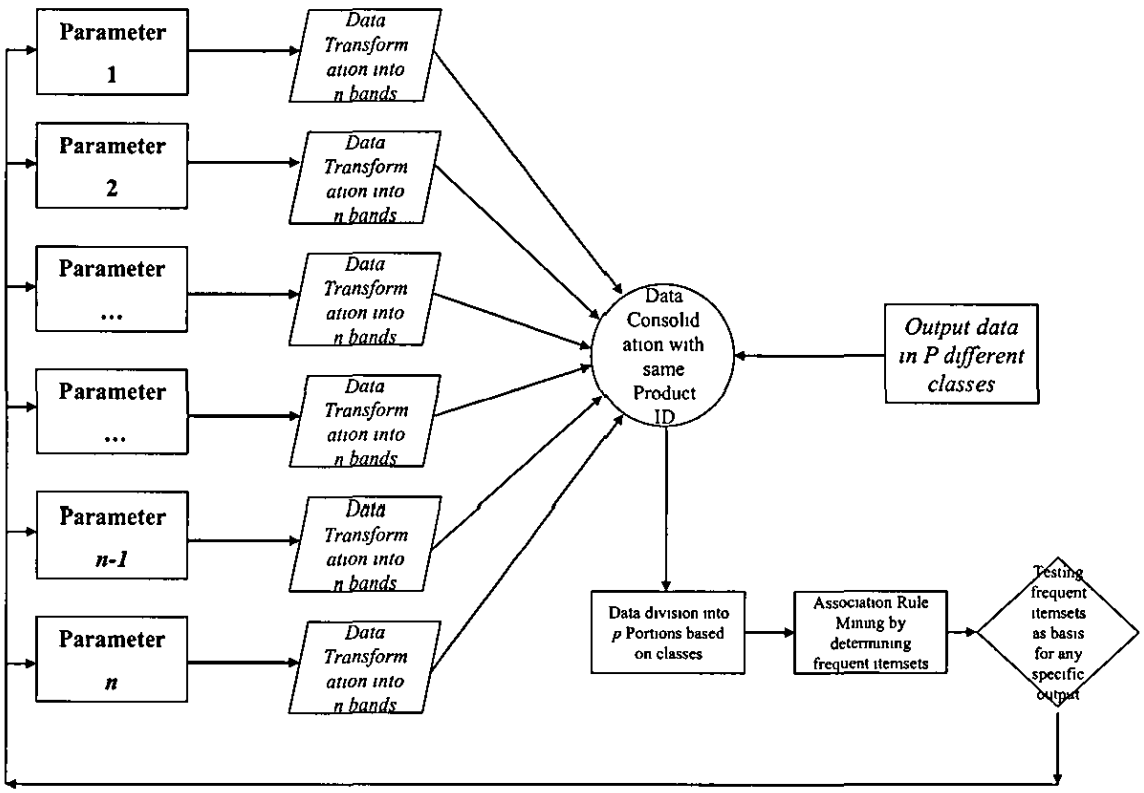


Figure 9 3 Association Rule Mining to Determine Product and Process Relationships

To discover the relationships between product and process data, the data of the manufacturing process related to one product and its output dimensions should be treated as one complete transaction. In the previous chapter, product data were divided into n different classes figure 8.2 and the process data should be treated in a similar manner and also divided into n different classes. Hence, the product and process database is dissected into n different parts and each portion of it corresponds to a particular class (figure 9.3) Association among process variables discovered in these portions will lead to a conclusion that those variables will lead to the specific class to which the data corresponds. Hence a range of possible values for the manufacturing variables will be mapped to a range of possible values for a product dimension or characteristic, i e an appropriate set of manufacturing process input variables (or set up

variables) could be identified to produce a product to a particular level of accuracy Table 9 2 shows a sample of data with transformed process variables and product dimension as discussed in the previous section. The sample data has 4 different classes and hence the datasets can be divided into 4 different portions as shown in table 9.3. These four smaller datasets are then used to determine the frequent itemsets within each group The frequent itemsets of these small datasets will be the classified sets of process values for the specific output class to which it belongs.

Var1	Var2	Var3	Var4	Var5	Var6	Class
5107	5204	5310	5410	5504	5611	F
5107	5208	5307	5409	5504	5607	G
5102	5204	5306	5405	5503	5605	F
5103	5205	5306	5406	5504	5605	G
5103	5206	5306	5406	5503	5605	G
5103	5205	5306	5405	5503	5605	G
5103	5206	5305	5405	5504	5606	F
5111	5208	5306	5406	5503	5605	F
5103	5204	5306	5405	5503	5604	G
5109	5205	5305	5404	5505	5610	I
5109	5205	5306	5405	5506	5611	G
5109	5209	5302	5402	5503	5606	H
5111	5209	5308	5406	5503	5604	F
5103	5205	5306	5406	5503	5604	F
5109	5207	5301	5401	5505	5611	H

Table 9 2: Example of Manufacturing process and Dimensional transformed Data

Var1	Var2	Var3	Var4	Var5	Var6	Class
5109	5205	5305	5404	5505	5610	I
5109	5209	5302	5402	5503	5606	H
5109	5207	5301	5401	5505	5611	H

5107	5208	5307	5409	5504	5607	G
5103	5205	5306	5406	5504	5605	G
5103	5206	5306	5406	5503	5605	G
5103	5205	5306	5405	5503	5605	G
5103	5204	5306	5405	5503	5604	G
5109	5205	5306	5405	5506	5611	G
5107	5204	5310	5410	5504	5611	F
5102	5204	5306	5405	5503	5605	F
5103	5206	5305	5405	5504	5606	F
5111	5208	5306	5406	5503	5605	F
5111	5209	5308	5406	5503	5604	F
5103	5205	5306	5406	5503	5604	F

Table 9.3: Data split into 4 portions based on the 4 classes present in the output product's dimension.

9.3.3.4.1 Application

The comments made in section 8.3.3 4 about the choice of level of support for product data are equally true for process data. In manufacturing contexts, there are some important error conditions or faults that may occur very rarely, but it is important to know about them Hence it is very important not to miss some infrequent but useful rules However, as in the case of product data, when the support level is set to a low level (so that useful rules are not missed), a large number of rules will be generated It is therefore important to also check the validity of the identified rules with manufacturing experts

9.3.3.4.2 Testing

It is very important to check the validity of the discovered frequent itemsets so subsequent sets of manufacturing data should be collected and used to check the results. The Chi square test can also be used to confirm the quality of the rules.

In the case of manufacturing process data, confusing results can occur when the process variable with particular range occurs in more than one different class Hence, the discovered itemsets also need to be checked for the complete dataset to find out which of them are producing confusing records This type of result can also be discussed with manufacturing experts to help to validate the rule quality.

When the best quality results have been identified, these frequent item sets can then be translated to determine the exact process parameters and then the rules can be generated from them, enabling values for manufacturing processes to be determined in order to manufacture product dimensions or other characteristics to required levels of accuracy. The above methods of data transformation and application of association rules were found to work well on the test data. One example of such a frequent itemset for a particular variable aa-thickness Uout might be ('5111', '5406', '5503') can be translated into a rules as, If the processvar1 is in band 11, processvar4 in band 6 and processvar5 is in band 3 then the output will be with in 'Uout' range having more than 95% confidence limit. This rule proves an linear association among process variables that leads to a product in specific range and can be incorporated in the system for use The process monitoring can detect the output and adjust the parameters for other process variables accordingly to adjust the offset.

The Association rules algorithm searches for the relationship among records in fields in a given data The association in the process variable data can be determined that leads to product quality in some specific ranges These associations are easy to understand and can readily be expressed in a query languages such as SQL The associations in the data can be easily expressed as "IF-THEN" statements which can be used as governing rules to be incorporated in process monitoring system. The rules can also be updated in the knowledge pool and used by other decision makers in the system This algorithm can work on variable length data However, it requires exponentially more computational effort as problem sizes grows and has limited capability to generate rare rules.

9.4 Discussion

This chapter presents the application of the DMA on product and process data to generate relationships between them that can be used in the production system as corrective measures or providing information to the shopfloor managers. The chapter presents the application four different data mining techniques with varying success for this It should be noted that these algorithms will have different success rates when the data changes The discovered frequent itemsets or sets of process variable values generated in association rule algorithm application

give the relationship between the process parameter with a specific product dimensional class in the form of rules Therefore the proposed technique can be applied to determine the process parameter values that are required to produce a product within a certain range of specified parameter values. Since the proposed method helps find the process variable values that best produce a product within the required tolerance band, this technique could also, in some cases be used as an alternative or be complementary to the design of experiment approach, to generate control signatures or in system design for real time feedback control These are discussed below

1. **Design of Experiment:** Design of Experiment is a disciplined approach to identify the causative factors for manufacturing problems In DOE the controllable input factors are systematically varied and their effects on output parameters are observed and analysed [191, 262] The overall objective is to predict the levels of controllable factors to minimize the effects of uncontrollable variables and noise The experimentation methodology ensures a statistically significant result The overall steps for the implementation of DOE are shown in Figure 9 4. The most crucial stage in this methodology is Step 3 In this step, the engineers exploit their collective knowledge and experience to determine the most likely causative factors for the manufacturing problem(s) in question [263] Each of these proposed factors are a speculation that will be subsequently tested for validity as input parameters in the DOE The number of DOE experiments increases exponentially with the number of variables and their levels and hence, restrict the number of independent causative variables The whole process of DOE would have to be repeated if the engineer fails to select all the causative factors in their experiment A typical manufacturing process generally has many variables and they interact at a wide range of operational (levels) Hence, selecting them correctly can be a difficult task In steps 4-10, the engineer designs, runs and validates the experiments based on the experimental conditions and manufactures a number of products. The steps of fabricating, testing and validating the experimental product are not only time-consuming but also expensive. The success of the experiment will lead towards corrective measures to be taken while its failure would force to start the experiment from beginning It is not unusual for a complex manufacturing problem to require dozens of iterations It is therefore a costly cycle time problem resulting in scrapped products, wasted materials and man hours on unsuccessful experiments, product shipping delays, lost market windows and undermined customer relationships So there is a requirement to greatly increase the odds of identifying the actual causative factors quickly and correctly so that the lengthy DOE process needs to only be done once The previous discussions show that Association Rules can be used to determine the best possible process variables and their values to manufacture a well controlled product Using the proposed method does not require any planned experimentation All that is required is quantities of historical operational data that can be used to determine the process variables that

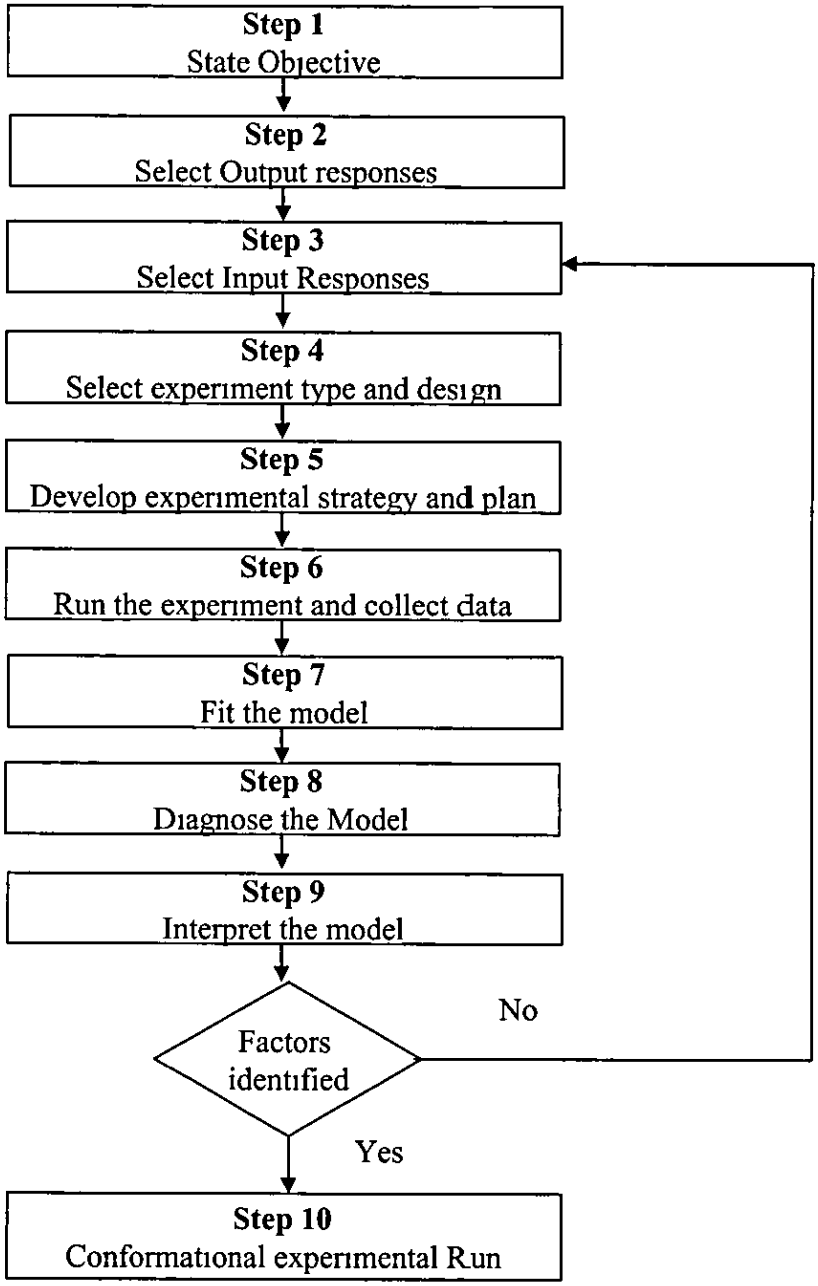


Figure 9 4 Steps for Design of Experiment

are best suited to manufacture a controlled product In contrast the DOE exercise requires specially collected very controlled process data during the experimentation, therefore, in situations where product and process information is routinely collected, there are clear benefits to using the proposed method of Association Rules as an alternative or pre-process approach to DOE.

2. **Control Signature:** A decision (control) signature is a set of feature (attribute) value necessary for making a decision. It simplifies and improves process control by integrating

parameters that might be otherwise be independently controlled In this research, an association algorithm based approach has been used to derive associations among control parameters and product quality in the form of decision rules. The methodology used in this research produces control signatures leading to good quality (nominal) products. As discussed in the testing section (9 3 3 4 2), each rule can be translated back in the form of some rule A complete set of such rules can be automatically generated and used after appropriate validation. The computational results obtained in this research opens new avenues for decision making process in manufacturing industries.

3 **System Design for Real time feedback control:** A Data mining enabled manufacturing system can help enhance the analysis and prediction capability of the current enterprise SPC (most widely used) system can only detect process abnormality. They do not automatically adjust the variables through feedback mechanism to lower the difference between target and measured value The results obtained in this research might be used to feedback in the manufacturing system to compensate for the change in the environment The feedback mechanism works in following way.

1 The real time data of process and product data is recorded in the data warehouse
2. The data is then analyzed to check if it meets the required condition
3 If not, then the measures are sent to data mining module to generate the changes that are required in the process parameters to meet the condition (This involves finding in which band the product data lies and finding out the change in value of parameters required for compensation)
4. The compensation value is then feedback for the next lot

The result obtained in this research is stored in a database The feedback loop when initiated, finds the output class from which it has to move to where. It then selects from the available feature sets which requires least changes in the parameter value

Novelty in the work:

- The proposed methodology provides techniques for systematic enterprise knowledge discovery particularly focusing on data cleaning and transformation requirements
- The proposed approach when used with SPC can provide the much needed feedback in the enterprise
- The results obtained provide information for manufacturing decision making in the enterprise
- The proposed methodology generates control signature leading to good quality of product and hence it helps in waste reduction.

- The proposed methodology of using Association Rules is novel
- This technique is a very useful supportive or alternative/complementary of the expensive design of experiments technique for quality improvement.

Chapter 10

10

Data Mining Agent on Maintenance

10.1 Introduction

Maintenance is an important function in many industries and directly affects shop floor control through the availability (or lack of availability) of resources and consequent needs for rescheduling or rerouting The costs due to failures and repairs can contribute a significant portion of the operating cost of the manufacturing facilities. System safety and reliability are important goals in industries and can be hard to achieve due to different regulations Therefore, it is essential that the maintenance system has the knowledge of the system degradation. It should be able to incorporate expert knowledge, feedback observations and know how of degradation to quantify the system performance, identify the major variables affecting the performance and help in decision making. It should be able to detect and predict faults and suggest corrective measures for them

A fault can be defined as an abnormal state of a machine or a system The cost and technical expertise required for systematic maintenance for a system increase substantially with the increase in functionality and complexity of the machine Fast and precise identification of faults and problems in equipment makes a crucial contribution to the enhancement of reliability in manufacturing and efficiency in product testing The knowledge of system behaviour i.e. identification of patterns leading to acceptable system behaviour and fault diagnosis is increasingly becoming important for improving the quality of manufacturing, decreasing cost of production, reducing the cost for product testing and increasing equipment improvement. Scheduled periodic maintenance is also expensive and may not always be necessary. It would be more efficient and cost effective if times when essential maintenance is

required could be accurately predicted, as this would ensure that no unnecessary time was wasted due to resources being unavailable since maintenance would only be carried out when it is really needed and no time was wasted through unexpected breakdowns and failures.
The maintenance system should be able to diagnose large amounts of data efficiently and effectively gathered from various sensors placed on-board in equipment. These sensors recording can be regarded as evidence of origin for recognising the working condition of a machine (e g normal operation, electric failure). The experts of this field can make proper judgement of failures by inspection of the measured signals in many circumstances but may fail without supporting tools in cases with high noise measurement. The major difficulties for such systems arise from contaminated sensor readings caused by heavy background noise as well as the unavailability of experienced technicians for support. This chapter considers and explores the use of the DMA for diagnosis of faults and understanding the system behaviour to predict maintenance requirements by using data mining techniques Ways of identifying and predicting the performance, likely behaviours and need for maintenance in machinery are important in the context of shop floor control as these factors will affect the availability and utilisation of particular resources which in turn will affect scheduling and routing decisions. This chapter will therefore discuss each module of the DMA in turn in the context of machinery whose performance can be tested through use on a known standard or a test medium. (This is therefore a different type of example to the one considered in chapter 9 where manufacturing process data is collected as different products are manufactured) In this way an approach will be proposed for the application of the DMA for fault diagnosis and system behaviour prediction for data obtained from measurement instruments and machinery The results from the DMA could then be used to build a predictive model for system performance. The model could also identify key variables, sensors and frequency responses that affect the performance and give information about fault type The mined result could then be used to create a decision signature for the system and predict the schedule for calibration of the system

10.2Preventive Maintenance Overview

Maintenance can be defined as the combination of all technical and associated administrative actions intended to retain an item or system in, or restore it to, a state in which it can perform its required function. The maintenance objectives can be summarized under four headings--ensuring system function (availability, efficiency and product quality); ensuring system life (asset management), ensuring safety, ensuring human well-being For production equipment, ensuring the system function should be the prime maintenance objective Here, maintenance

has to provide the right (but not the maximum) reliability, availability, efficiency and capability (i e , producing at the right quality) of production systems, in accordance with the need for these characteristics. Costs of maintenance have to be minimized while keeping the risks within strict limits and meeting statutory requirements.

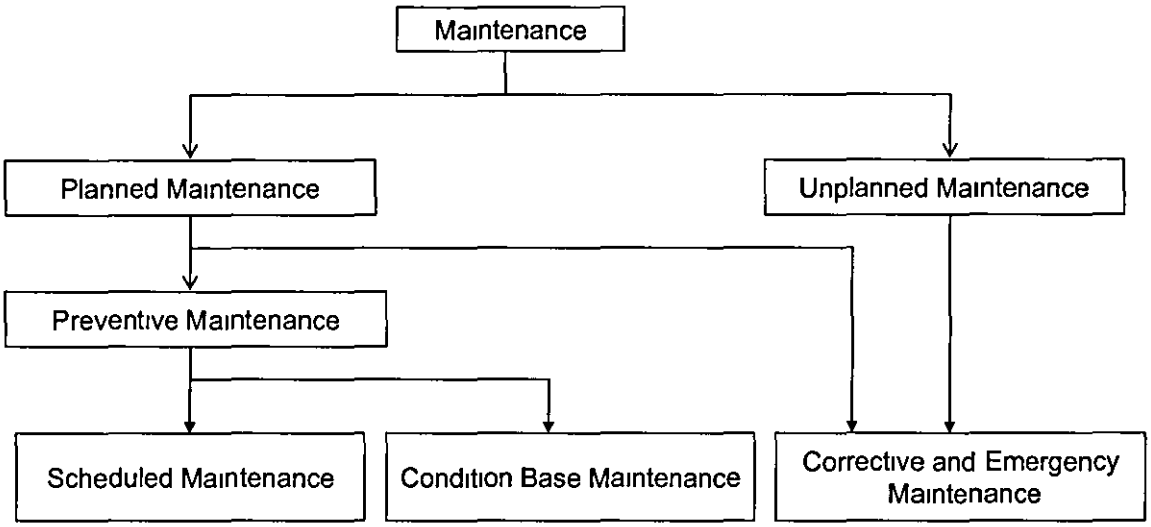


Figure 10 1 maintenance Policies (Adapted from Williams [264])

There are three main types of maintenance policies predominate in both theory and practice unplanned maintenance, planned maintenance and preventive maintenance [264]. Each sub type within those types are intermediate steps towards the next level of complexity. The relationships between the different maintenance policies are shown in figure 10.1. Unplanned maintenance is a reactive program that performs only emergency repairs The inventory level for this activity may be quite high as there is not much knowledge about failures and it is also a longer and more costly process than planned downtimes [265] Schedule maintenance programs carry out maintenance at predetermined intervals. The entire system is overhauled in the process independent of its requirement, and although this generally costs less than the corrective maintenance, production schedules may still be disrupted and unscheduled breakdowns may still occur frequently[266]. Parameters for scheduling activities such as time from the last maintenance, amount of usage, etc need to be determined Condition based maintenance (CBM) is the most comprehensive of the policies and generates large amounts of data. CBM is a maintenance philosophy wherein equipment repair or replacement decisions are based on the current and projected future health of the equipment [267] This approach measures the state of the machine either at intervals or continuously through sensors and/or monitoring of process parameters. Monitoring data is analysed to classify the varying states of the machine's condition and extracted knowledge is used to forecast when breakdowns are

likely to occur. The constituents and sub processes within CBM include sensors and signal processing techniques that provide the mechanism for condition monitoring

In general, maintenance optimization models cover four aspects: (i) a description of a technical system, its function and its importance, (ii) a modelling of the deterioration of the system in time and possible consequences for the system, (iii) a description of the available information about the system and the actions open to management and (iv) an objective function and an optimization technique which helps in finding the best balance.

The DMA could add significant functionality to a CBM system, as it would enable many levels of increasingly sophisticated assessments to be made, from raw data analysis to situation and impact analysis The objective of the DMA in this context would be to find rules that can be used as a set of governing rules for CBM of the machine or system, thus enabling more accurate and efficient shop floor control.

10.3 Problem overview

Manufacturing firms strive hard to reduce their operating cost, though they rarely mention reducing their maintenance costs. Machine maintenance contributes a significant amount (15-40%) to the cost of production and around one third of it is spent on unnecessary or improper maintenance activities ([268]) Traditional time-based machinery maintenance is being replaced by maintenance based on the condition of the machinery

The key to the successful implementation of CBM strategies is the accurate diagnosis of existing component faults, and the ability to predict when components are going to fail [264] The latter is the real key to successful CBM, since the maintenance system needs to know that a part is going to fail during the next task before that task is assigned The maintenance system needs to be able to reliably predict when components are going to fail, and furthermore, develop analysis techniques that can be implemented on embedded processing systems to automatically identify the remaining useful life of components, with little or no intervention from a human expert Application of data mining techniques helps in generating information about many queries inherent in CBM Ideally the DMA modules can sift through the historical data and find previously unknown patterns to help in determining how and when the machine should be monitored Potentially it can therefore help in detecting the type of failure that occurs on the machine and also provide information about warning signs

Due to the importance of maintenance in shop floor control, some experimental data mining work has been carried out in order to explore the potential use of the DMA in a maintenance context. The proposed DMA has therefore been explored in the context of instruments or other machinery whose performance can be tested through use on a known standard or a test medium. The performance of many types of measurement instruments or systems can be checked by running them on a known standard or test medium for which the result is known e g use of artifact in calibrating Co-ordinate Measuring Machine [269, 270] and use of different samples in mass spectrometer[271] Therefore, if the instrument produces the correct (known) value, it is assumed to be working accurately but if it produces an incorrect (unanticipated) value then it is assumed that some fault exists in the instrument and therefore some maintenance may be required. In this context, the DMA is explored for performance monitoring of a liquid chromatogram system. Liquid chromatograms are widely used in research and industry to detect the quality (chemical composition) of product and process waste ([272]) They are generally used to confirm the presence of specific substances or measure how much of it is present This process is highly useful in environmental pollutant work and in pharmacokinetic studies where the goal is quantisation at very low concentration in complex mixtures The aim of this work has been to identify if useful information can be generated about relationships among variables in the output results from the instrument and their relationship to the status or functioning of the machine by analysing the sensors data of the machine.

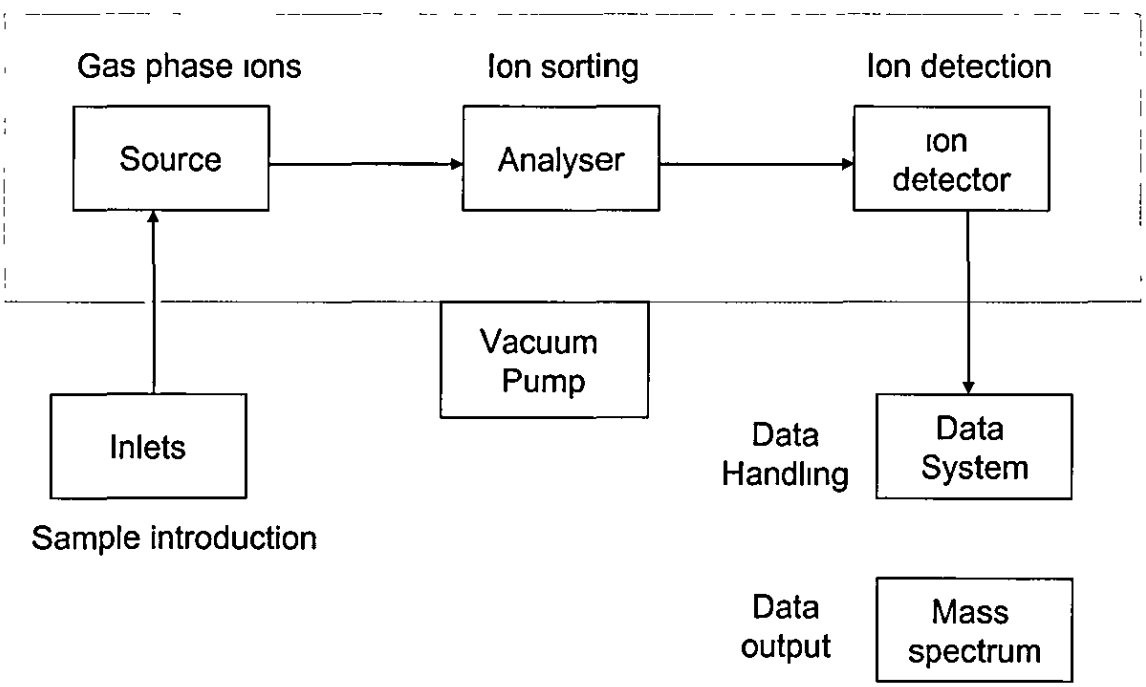


Figure 10 2 The different components of a mass spectrometer[274, 275]

The functionalities of a spectrometer are described below

A mass spectrometer is an instrument that measures the masses (mass-to-charge ratio) of individual molecules that have been converted into ions ([273]) The different functional units of a mass spectrometer are represented conceptually in the following block diagram [274, 275](figure 10 2) The sample is introduced through the inlet and source and converts the molecules from the samples into ions and accelerates them towards the mass analyser The mass analyser separates them in accordance with the charge and mass with use of electric and magnetic fields These ions are then allowed to pass onto an ion detector producing an electric current which amplified and detected ([276]).

Mass spectrometer cannot detect neutral molecules as its path cannot be guided by electrical and magnetic fields. Therefore, ions are created from injected samples for mass sorting and detection process that occurs in a mass spectrometer. There are various ionisation processes available e g. chemical ionisation, electron ionisation, fast atom bombardment, etc Early mass spectrometers required a sample to be a gas, but with development the applicability of mass spectrometry has been extended to include samples in liquid solutions or embedded in a solid matrix The gas (mobile in solid matrix) phase ions are sorted in the mass analyzer according to their mass-to-charge (m/z) ratios and then collected by a detector. In the detector, the ions flux is converted (magnified) to a proportional electrical current The data system records the magnitude of these electrical signals as a function of m/z and converts this information into a mass spectrum (a graph of ion intensity as a function of mass-to-charge ratio) This record of ions and their intensities serve to establish the molecular weight and structure of the compound being mass analysed Chromatographs are also generated by the data output system([272, 274-277]) A chromatograph is a temporal graph for the intensities of the molecules falling on the detector at any instance of time The data processing system calculates different features of the graph ([278]). The performance of the spectrometer can be checked at any time by running the system on a sample of a known compound that generates only one kind of ions. The resulting chromatograph will have a single peak and other result quantifying parameters can be compared with the standard and deviation can be easily measured and analysed for its occurrence The aim of this experimental work was to see if the DMA could be used to find patterns or clusters in this type of test data as the existence of these may indicate common faults in the condition or behaviour of the equipment In the next section the application of the data mining agent and its individual modules has been discussed for these systems

10.3.1 Data Mining Agent

The analysis of historical data can potentially provide valuable information about the systems status and performance Data collected from sensors on the machines should be a good source to determine its key states during operations and help in detecting and measuring values of machines states and their profile and unearth hidden symptoms ([231, 233]). A maintenance system based on these concepts should be able to reason about the different types of faults associated with the machine or make prediction about the potential problems that may occur in future In this section, data mining agent is primarily discussed in the general context how it can be used on machine sensors data to determine the key variables affecting its performance and also relating them to performance. The different modules of this agent including data cleaning, data selection, data transformation, data mining and pattern evaluation have therefore been discussed with respect to machines sensors data However, as previously shown in chapter 6,7,8 and 9, these steps can occur and reoccur in many different iterative cycles. The type of data cleaning and transformation depends upon the data source and type, algorithm selected and information required. The types of data considered in this chapter is quite different from the process data considered in chapter 9. For example it is generally all collected within one or more files on the instrument for each reading that is taken. How easy it is to access and read these files depends on the "openness" of the individual system as the files are normally created solely for the purposes of processing results by the measurement system itself It therefore may be necessary to obtain or create pieces of code to read the existing file format and convert the file's contents into a more useable form. This therefore can be a difficult process, however once the files can be accessed systematically the data can often be processed in a consistent manner as the data has generally been created automatically by the machine (rather than by human data entry) so there tends to be less anomalies in it

Signals data are different formats and a signal processing module is therefore also required (as part of the data cleaning module for data collection) to convert all the sensors data into a common format. These data may be highly contaminated by the noises and these must be removed before applying any data mining algorithms. It may also be possible to classify the data obtained from sensors into different classes to group or cluster the data. Different data mining algorithms have been applied on the data to try to extract useful information It should be emphasised however that these have been initial, exploratory investigations with algorithms as no single technique can perform best on all type of data and the data mining stage module of the DMA should therefore be able to apply a variety of different data mining algorithms depending on their appropriateness for the data that is being examined. The knowledge extracted from the application of data mining algorithms should always be validated and

evaluated (as discussed in previous chapters) to assess its usefulness This may require input from human experts If additional information is also available about the operation of the instrument, this may add value to the knowledge that can be extracted from it. For example in this context, if additional information can be obtained about set up changes, or maintenance carried out over a period of time or of the behaviour of operators, it may be possible that the knowledge extracted could help in predicting future behaviours and trends, allowing companies to make proactive decisions based on past experiences and knowledge obtained from the databases

10.3.1.1 Data Cleaning

In liquid chromatograph/ mass spectrum (LC/MS), the data consists of a series of mass spectra (mass-to-charge ratio, peak intensities) obtained by scanning a particular mass range (e g , m/z 30-500) of every scan separately during the experimentation [275]. If a scan is taken every second and the run is 30 min long, 1800 spectra are recorded. This information may be

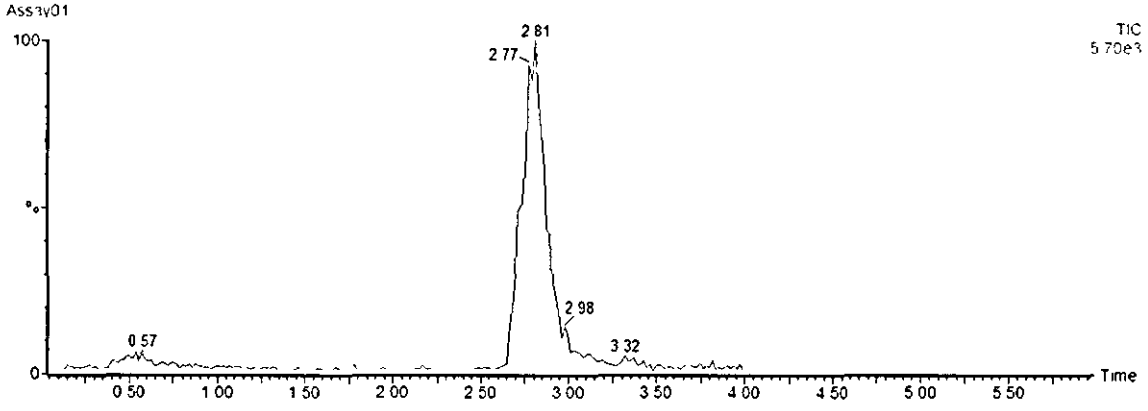


Figure 10 3 Total Ion Chromatogram

displayed in several ways as shown in figure 10 3, 10.4. First the intensities of all the ions in each spectrum can be summed (irrespective of mass-to-charge ratio), and this sum plotted as a function of chromatographic retention time to give a total ion chromatogram (TIC). Each peak in the TIC represents a compound that can be identified by interpretation of the mass spectrum recorded for the peak. The intensities at a single m/z ratio over time can be displayed to yield a selected ion current profile or mass spectrum. This technique can be used to find components of interest in a complex mixture without having to examine each individual mass spectrum The setup data for each experiment that is carried out is also recorded [275, 279]

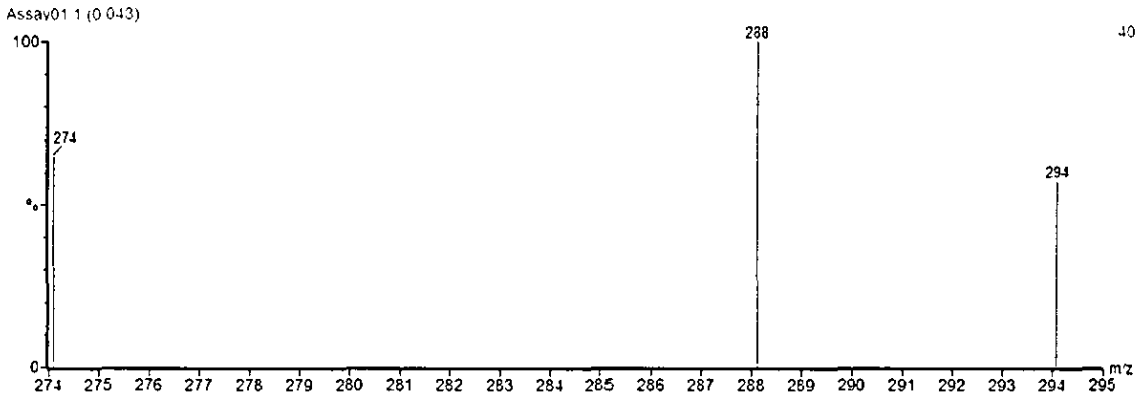


Figure 10 4 Spectrum

The requirement of the data cleaning module is that it takes raw data, which may be in the form of one or more types of file, possibly containing errors (e g duplications, gaps, etc), and combines it to output well structured, consolidated, error-free tables in a data base. Data cleaning is important in order to remove all the records that can produce errors or problems during the data mining algorithm application stage. The importance and functioning of data cleaning has been previously discussed in chapters 7, 8 and 9. However, the type of instrument data examined in this example has some different requirements to the product and process data that has previously been considered and therefore imposes addition functionality requirements for the data cleaning module of the DMA

Data consolidation and cleaning is an important and time consuming phase in the application of the data mining agent The data considered in this chapter was obtained from readings made with a known compound that is commonly used for performance evaluation No data was available about maintenance or operator behaviour over the period of the readings A significant challenge to data cleaning in this case was that the data recorded by LC/MS instruments were in different files and format. These data therefore needed to be converted into a single format and consolidated using a primary key. The accessibility and complexity of transforming the different data types are major challenges to applying the proposed DMA approaches on data from a variety of different sources (machines and instruments) The selection of data cleaning tools required must be guided by the data conversion mechanism available in the software system of the sensors and requirements of and ease of application of individual data mining algorithms on these data. The data collected on the machines to measure its performance can be divided into three different groups i e spectrum/chromatogram data, set up data and performance quantifying measures LC/MS spectrum data are converted into ASCII format and recorded with product id being the primary

key. The data in this file consist of series of mass spectra that is obtained during experimentation Peaks generated in the chromatograph are integrated and the results obtained are recorded in a file with product id being the primary key This way in which data is stored (i e its format and file structure) is dependent on the type and make of instrument used The data thus obtained therefore needs to be processed to obtain data that can be used for analysis i.e. to extract data that characterises or summarises the key feature of the reading obtained so that these can be compared with the anticipated key features for the standard The spectrum data file is processed therefore to obtain different results like retention time and scan number that produce spectrum has highest value recorded, highest m/z value in those scans, the signal value, scan number and retention time when the tested compounded was found to be maximum etc. The results obtained from peak integration file can give characteristics such as the value peak height, peak width, peak area, etc. these values are recorded in a file with product id as primary key

The data obtained from LC/MS usually has two kinds of noise in them One is measurement (electrical) noise generated due to intrinsic imprecision of the electric and magnetic system used in the apparatus It is very difficult to identify and reason with this noise and it therefore needs to be removed (cleaned) before applying any data mining algorithms. The decision as what is "noise" is rather subjective as logically any signal at a low% can be noise. Therefore a decision was made that collectively all the signals which were at a low %, for example below a level somewhere in the region of 2-7 % of the intensity of the signal (compound used in testing) should be treated as electrical noise The value of the electrical noise can therefore be calculated by processing the data in the spectrum file and recording the sum of these low % signals in the data as an "electrical noise" attribute value The other source of noise is external (chemical noise) It is caused by the contamination in the sample by solute or solvent. Its value is obtained from the peak integration of chromatographs The sum of the areas of all the peaks except signal peak is classified as chemical noise. The value of the "chemical noise" can therefore also be calculated and its value also recorded as an attribute

Extraction of experimental setup data may also be a tough task, as it is likely that this is also stored somewhere in the output files in a particular format, therefore in most cases purpose written piece of code will be required to retrieve it from the files. Different instrument data variables that should be recorded from the setup file of the process, relate to instrument set up parameters such as voltage, magnetic field strength, etc However, setup values that are always the same, i.e. the value of variables that are the same in all records (in all the data files) can removed from the database as they would not add any value to any knowledge that might be discovered in the data.

The data obtained directly from instruments or machinery in this way generally do not have confusing records as it is collected automatically without any human interaction However, duplicate and missing value records may be obtained and they should be deleted from the data base. There is high probability that some tuples in the database obtained above has constant value e g the strength of magnetic field applied will remain same in different runs for it These therefore will not have any effect on the performance of the instrument and hence are removed from the database

The process of data cleaning in this context can therefore be summarised as follows:-

1. Extract and consolidate the data in one record for each reading. This is likely to need purpose written program code to deal with the format(s) of the particular instrument's data files. The record for each reading should have a primary key.
2 In cases where there are many data points (as in this case where one reading which lasted 30 minutes would record 1800 spectra values), use background understanding of the instrument, process or context to reduce this to a few meaningful attributes which characterises the data For example, in this case, useful attributes would record details of the main peak (for the known compound), the chemical noise, the electrical noise, etc. This is likely to require purpose written program code to do this
3. Assess the contents of these reduced records further, particularly considering set up information etc. If any values are the same for all records, these fields can be removed from the record as they will not add value to the discovered knowledge
4. Create a clean consolidated table (or file) of all the readings (from the reduced records created in steps 1, 2 and 3), taking care to remove any duplicated or incomplete records

10.3.1.2 Data transformation

The importance and functioning of the data transformation module has already been discussed in previous chapters 8 and 9 This module takes the data tuples from the data base and transforms them into a suitable form as required by the particular data mining algorithm which is going to be applied Data transformation is not always necessary in data mining generally. However as LC/MS data is often continuous, (e g retention time, peak area) and many data mining techniques work on discrete data, data transformation often becomes a necessary stage. Some data mining algorithms also require data to be normalised so that the large variations in one of the variables does not overpower any variables with only small variations It is more appropriate to run some kinds of data mining algorithms with only a few divisions of the data

since too many different values will result in very indistinct results. Data transformation in general for data value that lies within limits and for those having normal distributions has already been discussed in previous chapters However, these types of data transformation will not work on data obtained by quantifying chromatograph and spectrum and set up file data as these variables generally do not have normal distributions

As mentioned in section 10 3 1 1, there were several instrument set-up variables routinely recorded in the files for each reading Variables from the input experimental setup data which are simply recorded as a discrete variable value, which is just a fixed number set for any particular reading, generally do not require any transformation. However to avoid any confusion that may arise due to the same values appearing for different variables, each variable was made distinct from each other (but with similar magnitude) Clearly, suitable transformation rules, based on the range and magnitude of the values for each individual variable, need to be found for each individual case, but the approach can be illustrated with the following example set of three transformation rules which, in each case will transform each value of the particular variable into a 4 digit number where the values for each variable begin with a particular integer. E g

Case 1: If the value of variable1 varies from 1 to 30 and 1000 is added to it to distinguish it from other variables, the values of variable1 will all lie in the range [1001 to 1030]

Case 2: If the value of variable2 was of the order of 10^{-6} and three decimal places stored, it could first be multiplied by 10^{6} and then 2000 added to it This will produce a range of values for variable 2 between 2000 and 2999

Case 3. If variable3 lies between 1 to 10 and it was first multiplied with 10 and then 3000 was added to it, this would produce a range of values for variable3 between 3010 and 3100.

These transformations help to distinguish the variables and produce integers which are better for processing in many data mining algorithms Any variables that have normal distributions can be divided in 11 different zones as in section 9 3 2 and demonstrated in table 10.1.

	Set_UpVariable4	*Set_UpVariable 5*	*Set_UpVariable6*	*Set_upVariable7*
	51	**52**	**53**	**54**
Band 01	5101	5201	5301	5401
Band 02	5102	5202	5302	5402
Band 03	5103	5203	5303	5403
Band 04	5104	5204	5304	5404
Band 05	5105	5205	5305	5405
Band 06	5106	5206	5306	5406
Band 07	5107	5207	5307	5407
Band 08	5108	5208	5308	5408

Band 09	5109	5209	5309	5409
Band 10	5110	5210	5310	5410
Band 11	5111	5211	5311	5411

Table 10 1 Transformation to integer identifiers for normally distributed variables

As explained in section 10 3 1.1, it is important to use background understanding of the instrument, process or context to identify a limited number of meaningful attributes in the data. These should provide characteristics to enable the output from the instrument (for readings of the known compound) to be classified as ***Good*** (i e. the machine is performing as expected) or ***Bad*** (i e. the machine is performing badly – it is faulty).

Clearly the chosen characteristics will vary in different contexts, but for example, in this study, the previously discussed variables such as chemical noise and electrical noise, position and size of the main peak (which should relate to the known compound being measured) etc., were considered.

The process of data transformation in this context can therefore be summarised as follows -

1. Consider the requirements of the particular data mining algorithm(s) which are to be applied Do they work better on integer data? Do they work better on a few categories of data, or can it be continuous? Etc. In this way identify whether any of the data needs to be transformed If it does not, go on step 4 if it does, continue with step 2
2. Is the data continuous and does it have a normal distribution? If so apply transformation techniques described in chapters 8 and 9.
3. Is data continuous but does not have a normal distribution? If so create a transformation rule of type discussed in this section
4. Consider the range and magnitude of all the variables – are there big differences in these between different variables? If so, normalise variables as discussed in this section
5 Use background knowledge of the instrument, process or content to classify each reading (record) as good or bad performance This will generally be done based on the values of each variable.

10.3.2 Data Mining

The requirement of the data mining module is that it applies an appropriate data mining algorithm on the previously cleaned and transformed data, which should now be in a single consolidated table. There are many possible algorithms which can be applied at this stage, for example regression, decision trees, or Association Rules algorithms This section shows the

methodology of application of these algorithms on the data obtained from spectrum and experimental set up. In this application of algorithm, the main objective is relate the variables obtained from setup files with the parameters defining the quality of the curve.

10.3.2.1 Regression Analysis

Different types of regression analysis have already been discussed in chapters 6, 8 and 9 In this research regression analysis was used as one of the first data mining technique to find the relationship between experimental setup variables and output curve quantifying variable to help in predicting their trends. The information generated through this analysis can be exploited in other data mining techniques and also to relate the setup files will quality of curve. In this implementation, both linear and non-linear regressions were applied on clean and untransformed data were used to determine any significant relationship between the variables The existence of relationships between the variables could have generated a set of governing equation for controlling output dimensions However, in this example, no strong correlation values were identified

10.3.2.2 Decision Tree

Decision tree algorithms were discussed in chapter. These algorithms have widely been used in literature for data classification ([219]) The algorithms classify the predicted class based on the different attributes and rules are generated from those classification These rules can be used as a set of governing rules for controlling the process In this application, ID3 and C 4 5 algorithms were applied on setup file data and curve quantifying data. The aim of this application was to determine the range of process variable that results in different quality of the curve. In this application, half of the data were used as training data and other half as test data with product data as classifying variable. Decision Trees results were not very promising as they included very few attributes in the rules. They identified some obvious relationships, since most of the times they gave only one attribute to classify the output curve, such as IF Set_upVariable4 is in band 1, THEN Electric Noise is high. However, such results do not provide any useful knowledge as they are too obvious result to quote as such, and are the kind of infinite rules are always present in the data. Also the classification errors were all greater more than 10%

10.3.2.3 Association Rule

The Association Rule algorithms, as discussed in the previous chapters, was also considered in this context Any associations found could then be translated back to generate an operating

range for good performance The methodology used in this context works on the same principle, but in this case, there are only two classes (accepted or rejected) for each reading (record) defining the output. A new tuple is generated in this application which is based on a Boolean function. This tuple has two classes, accepted, if all the variable defining the output are in accepted range and rejected, if any of them is out of acceptable limit.

The input parameters i e. data from setup files, sensors, sensors processing unit and output classifying variables makes up for one complete transaction. The data is divided into two classes (accepted or rejected) If any kind of association among variables is generated within these sections then this will lead to a conclusion that this association results in that specific class to which data section corresponds.

As the available data set was small, the support level had to be kept very low, so as not to miss any potentially useful rules. In consequence a large number of rules were identified (see discussion in section 8 3 3.4). The rules were checked using the chi square method discussed in chapter 8 and many of them were found to have high confidence level. The association rule method therefore again appears to show good potential but a much larger study and further work working with domain experts with further background information would be needed to properly evaluate this type of data mining in this context. However the current explorative work has demonstrated that the DMA shows potential for ongoing knowledge discovery in all three context areas that were considered in chapters 8, 9 and 10

10.4Discussion and Novelty

This chapter has discussed the application of the DMA in the context of machinery whose performance can be tested through use on a known standard or a test medium The tasks which need to be carried out by each of the DMA's modules have been considered in turn and the essential steps summarised In this context, it is important to identify sets of process variables recorded at the instrument and use these to quantify the instrument's performance. The objective of the proposed approach is that it can be used to determine the range of values of machine variables for acceptable performance level, and help in generating a decision signature and providing information for condition based maintenance system

1. **Control Signature:** A decision signature is a set of feature (attribute) values necessary for making a decision. It simplifies and improves instrument performance integrating parameters that might be otherwise be independently controlled In this research, an association algorithm

based approach has been used to derive associations among parameters recorded at instrument and its performance in the form of decision rules The methodology used in this research shows potential in identifying decision signatures leading to its acceptable performance level However further work is required in this area, working with domain experts to properly evaluate the quality of rules identified

2 **Condition Based Maintenance:** The DMA shows potential for providing information to the shop floor control system about performance monitoring The rules generated in this research provides information about system performance and the variables affecting them The association rule generated in this research are also based individual variable defining performance. These rules provide information about fault detection and gives warning sings when they are going out of acceptable range

Novelty in the work

- The proposed methodology provides a systematic knowledge discovery about instrument functioning and its performance monitoring particularly data cleaning and transformation requirements
- The result obtained might be used to provide compensation for the degradation in performance
- The rules generated show potential in providing decision signatures leading to acceptable performance level or warning about unacceptable performance levels

.

Chapter 11

Conclusion

Recent advances in the field of information systems and networking have greatly changed the characteristics of demands from shop floor of an enterprise It is not only viewed as a production centre but is also considered as a nucleus of information and knowledge. This knowledge may consist of system's behaviour, limitation, capability, etc and its utilisation could enable the enterprise to differentiate itself from competitors These information and knowledge can be extracted and integrated in the system which could enhance its performance Enterprise must therefore have the capability to learn new knowledge, propagate them in system and discard after its shelf life i e. must have a systematic knowledge management and maintenance system

The increasing use of computers at different levels in manufacturing enterprises coupled with advances in information technology and decreasing cost of data analysis have enabled enterprises to easily store and maintain large volumes of data These data sources can be valuable assets and potentially important source to explore and generate knowledge and information about the system behaviour. The enterprises can become data-aware and respond quickly by exploring these databases. This can be achieved by intensive and intelligent analysis of existing databases with the aim to identify new trends and knowledge. The main objective of this research was stated in chapter 1 as being to design a data supported manufacturing shop floor control system, which can benefit from its historical or legacy systems, as well as from its current databases

A thorough review of shop floor control systems is presented has been chapter 3. The review classified the manufacturing system control into three different types of framework and its assessment showed that the basic assumptions of an architecture paradigm leads to constraints

being induced in the control system, affecting the structure and genericity of the architecture It was identified that the control architecture must address issues of adaptability, generic applicability, efficient and effective knowledge management and discovery, reactive scheduling and process planning and has dynamic structure The intelligent shop floor controller for a dynamic environment must also incorporate mechanisms to monitor the environment in real time, support learning mechanism and present relevant information and alternatives to decision makers and assist in making better decisions. These features indicated that improved knowledge management and ongoing learning (through knowledge discovery and reuse) were needed to improve the effectiveness and efficiency of shop floor control systems

This review also identified that existing architectures and structures for shop floor control systems cannot support these requirements and that many authors are proposing MAS as the best approach to address current weaknesses and improve the performance of the next generation of shop floor control systems. This approach offers many advantages for manufacturing control: modularity, adaptability, reconfigurability, etc. Agent systems were therefore studied in chapter 4 and after a thorough examination of their scope and capabilities; MAS were accepted as the most appropriate architecture for this research

Data Mining has been studied as a way of addressing the needs for regularly updated knowledge and ongoing learning The reviews presented in chapter 5 and appendix I show that increasing numbers of manufacturing problems have been successfully addressed using a wide range of data mining tools and techniques However, it is clear from the published literature on this topic that most applications of data mining in manufacturing have been done to address specific "one off" problems, rather than giving any consideration to how the discovered knowledge can be fully exploited, reused and integrated in the system, or to how the many varied approaches to data mining can be best applied. Therefore if these approaches are to be adopted in the context of this research, to tackle the knowledge management and learning challenges of shop floor control, additional capabilities need to be incorporated in the shop floor control system to also support knowledge reuse and knowledge maintenance. To identify the necessary capabilities, a thorough review of data mining tools, techniques, methodologies and performance has been presented in chapter 6. It also addresses the challenges of management of data mining projects and the selection of appropriate tools and algorithms for it.

In chapter 7, a proposal was made for a data integrated shop floor control system which is agent based (using background knowledge gained from chapter 4) and includes the data

mining capabilities that were identified in chapters 5 and 6 in order to address the weaknesses of current shop floor control architectures that were identified in chapter 3. The proposed system makes use of quasi-heterarchical structure to enhance the decision making process The proposed system has a decentralised control mechanism which uses knowledge, information and alternatives provided by the decision support. The intelligent decision support tool presented in this work incorporates data mining and intelligent agent technology for providing useful information and knowledge. These techniques provide a mechanism for learning in the system. It also provides an approach for integration of data mining processes for the generation of required knowledge and information for different activities on the shop floor into a decision support framework by means applying intelligent agent technology. An approach for linking different data mining agents situated at different process sites has also been discussed. This linkage helps in transferring the knowledge generated at one site to be reused at other The proposed data mining enabled decision support tool provides feedback to different levels in a formalised way so that discovered knowledge can be exploited and reused in various ways in the future. The proposed data mining enabled decision support system would thus assist the decision makers in making decisions by providing them with more alternatives with their implications

Two different approaches have been made to test and demonstrate the proposed system. In chapter 7, sections 4 and 5, a partial implementation has been coded to demonstrate the application of a MAS in this context This demonstrates the interactions of three typical agents within the proposed system to presents its functionality This research presents the conceptual description and explanation of the set of modules that would be required to build the system and in order to maintain the generic applicability of the proposed approach The partial implementation has therefore also been kept flexible by using JAVA code rather than using particular agent environments which constraint the required detail specification of hardware and software required for it

A key determining factor in the potential value of the proposed system is how well the data mining agent would be able to supply the SFC system with useful, well-maintained knowledge There are many different data-mining tools and existing commercial software tools and systems which include many of them. The purpose of this research is therefore not to simply show the application of different data mining techniques, but rather to consider the different types of data which might exist in different manufacturing operational contexts and show how the data mining agent would operate in each of these contexts. It is important to emphasise that no single data mining approach will give the best solutions in all contexts, it is therefore important to establish a general approach or methodology for the operation of the

DMA in a variety of contexts Chapters 8, 9 and 10 therefore demonstrate the differences and challenges to be faced by the DMA and the potential types of knowledge that it might discover when it is working with (1) product data – as measured on CMM or other measurement systems (2) combined product and process data – when process input variables are known as well as measured product data and (3) instrument or machine data to possibly provide knowledge about the performance or maintenance needs of the resource. The data mining studies in chapters 8, 9 and 10 are therefore exploratory in nature rather than trying to solve any particular problems and primarily have been included to explain the tasks and activities of the various modules within the DMA in the three different contexts. It should also be emphasised that due to the complexity and wide variety of different types of data and contexts that the DMA must deal with it is very unlikely that the DMA could operate in a fully automatic manner in the foreseeable future It is likely that human interaction and domain expertise will be needed to guide the selection of data and also to critically evaluate the data mining results

The proposed data mining enabled decision support tool provides feedback to different levels in a formalised way so that discovered knowledge can be exploited and reused in various ways in the future One of the important attributes of this approach that they have the potential for their knowledge to be continuously updated and provides a way for generating knowledge from operational databases This methodology contributes substantially to current approaches to quality assurance as it supports improvement in manufacturing processes and product quality by determining the process controlling variables that result in particular output classes and providing the feedback information to SPC process for compensating the difference in measured and target value An additional benefit is that it provides a useful alternative to the expensive and time-consuming classical and full factorial experimental design approaches in this context It also offers alternative yet complementary techniques to modern Design of Experiments (DOE) approaches where it provides particular benefits in the early stages of screening experiments. Data mining enabled manufacturing system can help enhance the analysis and prediction capability of the current enterprise

The main achievements of this work are·

- Presents a design and prototype of an intelligent decision support system for shop floor control which incorporates data mining techniques The data mining techniques help it to analyse large volumes of data The system is capable of learning new knowledge
- Provides a methodology for systematic enterprise knowledge discovery which particularly focuses on data cleaning, transformation and rule quality requirements

- Provides a methodology for knowledge generation in manufacturing shop floor contexts
- Provides an approach to generate control signatures which can be utilised in decision making

Future Work:

It is also important to note that the blind application of data mining to generate knowledge can be dangerous, leading to the use of meaningless patterns. It is therefore always desirable to incorporate prior knowledge and to properly interpret mined patterns. The successful methodology of data mining application should only be incorporated in the decision support system Additional verification of new knowledge by domain experts is therefore also recommended. The application of data mining tools and techniques has been carried out in three different contexts in this study. It must be applied to many other to determine the different tasks that are needed for its successful application. It should also be noted that this research does not provide a detailed specification of the hardware or software that may be required for the implementation of such a system, but it provides a conceptual description and explanation of a set of modules that can be combined to develop an data mining enabled system for decision support in shop floor control Knowledge organisation will be an important issue in the functioning of this decision support system A functional specification for knowledge management should be incorporated which generates a synergy between knowledge generation, operationalisation and support quicker access to it

Reference

1. Krothapallı, N.K.C. and A.V. Deshmukh, *Design of Negotiation Protocols for Multi-Agent Manufacturing System* International Journal of Production Research, 1999. **37**: p 1601-1624.
2. Sıkora, R. and M J. Shaw, *Coordination Mechanism for Multi-Agent Manufacturing Systems. Application to Integrated Manufacturing Scheduling".* IEEE Transactions on Engineering Management, 1997 **44**(2): p. 175-187.
3. Shiue, Y.-R. and R.-S. Guh, *The Optimization of Attribute Selection in Decision Tree-Based Production Control* International Journal of Advance Manufacturing Technology, 2006. **28**: p. 737-746.
4. Shin, J. and H. Cho, *Rapid Development of a Distributed Shop Floor Control System From an XML Model-Based Control Software Specification* International Journal of Production Research, 2006. **44**(2): p 329-350.
5. Yoon, H.J. and W. Shen, *Simulation Based Real Time Decision Making for Manufacturing Automation System. A Review* International Journal of Manufacturing Technology And Management, 2006. **8**(1-3) p. 188-202.
6. Siemieniuch, C.E and M.A. Sinclair, *Organizational Aspects of Knowledge LifeCycle Management in Manufacturing* International Journal of Human-Computer Studies, 1999. **51**(517-547).
7. Ozbayrak, M. and R. Bell, *A Knowledge Based Decision Support System for The Management of Parts and Tools in FMS* Decision Support Systems, 2003. **35**: p. 487-515.
8. Wang, H., *Intelligent Agent-Assisted Decision Support Systems Integration of Knowledge Discovery* Expert Systems with Applications, 1997. **12**(3): p. 323–336.
9. Babiceanu, R.F and F.F. Chen, *Development and Application of Holonic Manufacturing Systems. A Survey* Journal of Intelligent Manufacturing, 2006. **17**. p. 111-131.
10 Dilts, D M , N.P. Boyd, and H.H. Whorms, *The Evolution of Control Architectures for Automated Manufacturing System* Journal of Manufacturing Systems, 1991. **10**(1): p. 79-93.
11. Jones, A.T. and A Saleh, *A Proposed Hierarchical Control Architecture for Automated Manufacturing System* Journal of Manufacturing Systems, 1986. **5**(1): p. 15-25.
12. Simpson, J.A., H. R J., and J.S. Albus, *The Automated Manufacturing Research Facility of the National Bureau of Standards.* Journal of Manufacturing Systems, 1982. **1**(1): p. 17-31.
13. Jones, A.T. and R C. McLean, *A Proposed Hierarchical Control Model for Automated Manufacturing Systems* Journal of Manufacturing Systems, 1986. **5**(1) p 15-25.
14. McLean, C., H. Bloom, and T. Hopp. *The Virtual Manufacturing Cell.* in *Proceeding of the IFAC/IFIP Conference on Information Control Problems in Manufacturing Technology.* 1982. Gaithersburg.
15. Senehi, M.K., T.R. Kramer, S.R Ray, R. Quintero, and J.S. Albus, *Hierarchical Control Architectures from Shop Level to End Effectors*, in

Computer Control of Flexible Manufacturing Systems, S.B Joshi and J.S. Smith, Editors 1994, Chapman and Hall.

16. Jones, A T. and A. Saleh, *A Multi-Level/ Multi-Layer Architecture for Intelligent Shop Floor Control.* International Journal of Computer Integrated Manufacturing, 1990. **3**(1): p. 60-70.
17. Joshi, S.B , R.A. Wysk, and A. Jones. *A Scaleable Architecture for CIM Shop Floor Control.* in *Proceedings of Cimcon '90, National Institute of standards and technology.* 1990.
18. Duggan, J. and J. Brown, *Production Activity Control: A Practical Approach to Scheduling* International Journal of Flexible Manufacturing Systems, 1991. **4**· p. 79-103.
19. Bauer, A., R. Browen, J. Browne, J. Duggan, and G. Lyons, *Shop Floor Control Systems - From design to implementation.* 1991: Chapman and Hall, UK. 344.
20. Biemans, F P.M., *A Reference Model for Manufacturing Planning and Control.* 1989, University of Twente: Enschede.
21. Arentsen, A.L., *A Generic Architecture for Factory Activity Control.* 1995, University of Twente: Enschede
22 Zhang, J., F.T S. Chan, and P. Li, *A Generic Architecture of Manufacturing Cell Control System.* International Journal of Computer Integrated Manufacturing, 2002. **15**(6): p. 484-498.
23. Scherer, E., *Shop Floor Control- A System Perspective From Deterministic Models Towards Agile Operations Management.* 1998: Springer Verlog.
24. Duffie, N.A., R. Chitturi, and J. Mou, *Fault-Tolerant Heterarchical Control of Heterogeneous Manufacturing System Entities* Journal of Manufacturing Systems, 1988. **7**(4): p. 175-179.
25. Duffie, N.A. and V.V. Prabhu, *Real-Time Distributed Scheduling of Heterarchical Manufacturing Systems.* Journal of Manufacturing Systems, 1994. **13**(2): p 94-107.
26 Vamos, T , *Cooperative Systems-An Evolutionary Perspective*, in *IEEE Control System Magazine.* 1983. p. 9-14.
27. Hatvany, J. *Intelligence and Cooperation in Heterarchical Manufacturing Systems* in *Proceedings of the CIRP Seminars.* 1985.
28. Duffie, N.A. and R S. Piper, *Non Hierarchical Control of Manufacturing Systems.* Journal of Manufacturing Systems, 1986. **5**(2)· p 137-139
29. Duffie, N.A and R.S. Piper, *Non-Hierarchical Control of a Flexible Manufacturing Cell* Robotics and Computer Integrated Manufacturing, 1987. **3**(2) p 175-179.
30. Baker, A D , *Metaphor or Reality: A Case Study Where Agents Bid with Actual Costs to Schedule A Factory*, in *Market-Based Control. A Paradigm for Distributed Resource Allocations*, S H. Clearwater, Editor. 1996, World Scientific: River Edge, NJ. p. 184-223.
31. Baker, A.D., *A Survey of Factory Control Algorithms which Can be Implemented in a Multi-Agent Heterarchy Dispatching, Scheduling, and Pull.* Journal of Manufacturing Systems, 1998 **17**(4): p. 297-320.
32. Smith, R G , *The Contract Net Protocol. High-Level Communication and Control in a Distributed Problem Solver* IEEE Transactions on Computers, 1980. **C-29**(12). p. 1104-1113.

33. Smith, R.G. and R. Davis, *Frameworks for Cooperation in Distributed Problem Solving* IEEE Transactions on Systems, Man, and Cybernetics, 1981. **11**(1): p. 61-70.
34. Upton, D., M. Brash, and M. Metheson, *Architectures and Auctions in Manufacturing* International Journal of Computer Integrated Manufacturing, 1991. **4**(1): p. 23-33.
35. Shaw, M., *A Distributed Knowledge-Based Approach to Flexible Automation: The Contract Net Framework* International Journal of Flexible Manufacturing Systems, 1988. **1**: p. 85-104.
36. Lin, G.Y. and J J. Solberg, *Integrated Shop-Floor Using Autonomous Agents.* IIE Transactions, 1992. **24**(3): p. 57-71.
37. Titley, K J , *Machining Task Allocation in Discrete Manufacturing Systems*, in *Market-Based Control A Paradigm for Distributed Resource Allocations*, S H. Clearwater, Editor 1996, World Scientific: River Edge, NJ. p. 225-252.
38. Ozaki, K., H., Y.I. Asama, K. Yokota, A. Matsumoto, H. Kaetsu, and I. Endo, *Negotiation Method for Collaborating Team Organisation among Multiple Robots*, in *Distributed Autonomous Robotic Systems*, H. Asama, et al , Editors. 1994, Springer: Tokyo.
39. Timmermans, P.J.M., *Modular Design of Information Systems for Shop Floor Control.* 1993, TUEindhoven Eindhoven.
40. Liu, J.S and K. Sycara. *Distributed Problem Solving Through Coordination in a Society of Agents.* in *13th International Workshop on Distributed Artificial Intelligence.* 1994. Seattle.
41. Liu, J. and K.P. Sycara. *Exploiting Problem Structure for Distributed Constraint Optimization.* in *First International Conference on Multi-Agent Systems* 1995. San Francisco, California: AAAI Press/MIT Press.
42. Nwana, H.S., L Lee, and N.R Jennings, *Co-ordination in Multi-agent systems*, in *Software Agents and Soft Computing*, H.S. Nwana and N. Azarmi, Editors. 1997, Springer. New York. p. 42-58.
43. Faratin, P., C. Sierra, and N.R. Jennings, *Negotiation Decision Functions for Autonomous Agents* International Journal of Robotics and Automation System, 1998. **24**(3-4) p. 159-182.
44. Oliveira, E., *Cooperative Multi-Agent System for an Assembly Robotics Cell* Robotics and Computer Integrated Manufacturing, 1994. **11**(4): p 311-317.
45. Finin, T , Y. Labrou, and J. Mayfield, *KQML as an Agent Communication Language*, in *Software Agent*, J.M. Bradshaw, Editor. 1997, MIT Press. Cambridge, MA.
46. Miyasshita, K., *CAMPS: A Constraint Based Architecture for Multi Agent Planning and Scheduling* Journal of Intelligent Manufacturing, 1998. **9**(2): p. 147-155.
47. Maione, B. and D. Naso, *Evolutionary Adaptation of Dispatching Agents in Heterarchical Manufacturing Systems.* International Journal of Production Research, 2001. **37**(7): p. 1481-1504.
48. Gao, Q., X. Luo, and S. Yang, *Stigmergic Cooperation Mechanism For Shop Floor System.* International Journal of Advance Manufacturing Technology, 2005. **25**: p. 743-753.
49. Booch, G., *Object Oriented Design with Application.* 1991: Benjamin/Cummings Publishing

50. Lin, G.Y. and J.J. Solberg, *Autonomous Control for Open Manufacturing Systems*, in *Computer Control of Flexible Manufacturing Systems*, S.B. Joshi and J S. Smith, Editors 1994, Chapman & Hall. p. 169- 205.
51. Macchiiaroli, R. and S. Riemma, *A Negotiation Scheme for Autonomous Agents in Job Shop Scheduling* International Journal of Computer Integrated Manufacturing, 2002. **15**(3). p. 222-232.
52. Gou, L , P B. Luh, and Y. Kyoya. *Holonic Manufacturing Scheduling· Architecture, Cooperation Mechanism, and Implementation.* in *IEEE/ASME International Conference on Advanced Intelligent Mechatronics.* 1997. Tokyo Japan
53. Shaw, M J , *A Distributed Scheduling Method for Computer Integrated Manufacturing· The Use of Local Area Networks in Cellular Systems* International Journal of Production Research, 1987 **25**(9): p. 1285-1303
54. Usher, J.M., *Negotiation Based Routing in Job Shops Via Collaborative Agents* Journal of Intelligent Manufacturing, 2003. **14**: p. 485-499.
55. Dewan, P. and S.B. Joshi, *Implementation of an Auction Based Distributed Scheduling Model for a Dynamic Job Shop Environment* International Journal of Computer Integrated Manufacturing, 2001. **14**(5): p. 446-456.
56. Veeramani, D. and K.J. Wang *Comparison of Distributed Control Schemes From the Prospective of Communication System.* in *Proceedings of the Third IIE Research Conference* 1994. Norcross, GA.
57 Veeramani, D. and K.J. Wang, *Performance Analysis of Auction Based Distributed Shop Floor Control Schemes from the Prospective of Communication System.* International Journal of Flexible Manufacturing Systems, 1997. **9**. p. 121-143
58. Kutanoglu, E. and S D. Wu, *On Combinatorial Auction and Lagrangean Relaxation for Distributed Resource Scheduling* IIE Transactions, 1999. **31**(9) p. 813-826.
59. Veeramani, D., K.J Wang, and J Rojas, *Modelling and Simulation of Auction Based Shop Floor Control Using Parallel Computing* IIE transactions, 1998. **30** p. 773-783.
60. Lewis, W., M. Barash, M , and J.J. Solberg, *Computer Integrated Manufacturing System Control. A Data Flow Approach* Journal of Manufacturing System, 1987. **6**(3) p. 177-191.
61. Kaihara, T. and S. Fujii, *A Self-Organising Scheduling Paradigm Using Co-Ordinated Autonomous Agents*, in *Rapid Product Development*, N. Ikawa, T. Kishinami, and F. Kimura, Editors. 1997, Chapman & Hall.
62 Tharumarajah, A. and A.J. Wells, *A Behaviour-Based Approach to Scheduling in Distributed Manufacturing Systems* Journal of Computer Aided Engineering, Special issue on Intelligent Manufacturing Systems, 1996.
63. Ramaswamy, S.E. and S.B. Joshi. *Distributed Control of Automated Manufacturing Systems* in *Proceeding of the 27th CIRP Seminar on Manufacturing Systems.* 1995 Michigan
64. Bongaerts, L , B.H. Brussel, and P. Valckenaers *Schedule Execution Using Perturbation Analysis.* in *International Symposium on Non-Linear Dynamics in Production Processes and Systems.* 1997 Hannover.
65. Peng, Y , T. Finin, Y. Labrou, R S. Cost, B. Chu, J. Long, W.J. Tolone, and A. Boughannam, *Agent-Based Approach for Manufacturing Integration: The CIIMPLEX Experience* Application of Artificial Intelligence, 1999. **13**(1): p. 39-63.

66. Parunak, H.V.D. *Agents in Overalls. Experiences and Issues in the Development and Deployment of Industrial Agent-Based Systems.* in *PAMM'99.* 1999.
67. Parunak, H V.D , A.D. Baker, and K. Krol, *Manufacturing Over the Internet and into Your Living Room: Perspectives from the AARIA Project.* 1997, ECECS Dept., University of Cincinnati, Cincinnati, OH: Tech Rep TR208-08-27.
68. Khool, L P., S.G. Lee, and X F. Yin, *Agent Bases Multiple Shop Floor Manufacturing Scheduler* International Journal of Production Research, 2001 **39**(14): p. 3023-3040.
69 Wang, Z., J. Zhang, and F.T.S. Chan, *A Hybrid Petri Nets Model of Networked Manufacturing Systems and its Control System Architecture.* Journal of Manufacturing Technology Management, 2005. **16**(1): p. 36-52.
70. Brussel, H.V., J. Wyns, P. Valckenaers, and L. Bongaerts, *Reference Architecture for Holonic Manufacturing Systems: PROSA.* Computers in Industries, 1998. **37**: p 255-274.
71. Parunak, H.V.D., *Distributed Artificial Intelligence Systems*, in *Artificial Intelligence Systems. Implications for CIM*, A. Kusiak, Editor 1988, IFS: Bedford, U K p. 225-251.
72. Butler, J. and H. Ohtsubo, *ADDYMS Architecture for Distributed Dynamic Manufacturing Scheduling*, in *Artificial Intelligence Applications in Manufacturing*, A. Famili, D.S. Nau, and S.H Kim, Editors. 1992, AAAI Press/The MIT Press. Menlo Park, CA. p. 199-213.
73. Tawegoum, R , E. Castelain, and J.C. Gentina, *Hierarchical and Dynamic Production Control in Flexible Manufacturing Systems* Robotics and Computer Integrated Manufacturing, 1994 **11**(4): p 327-334.
74. Brennan, R.W., S Balasurbramanian, and D.H. Norrie. *A Dynamic Control Architecture for Metamorphic Control of Advanced Manufacturing Systems.* in *Proceedings of International Conference on Intelligent Systems for Advanced Manufacturing.* 1997. Pittsburgh, PA.
75. Ou-Yang, C. and J.S Lin, *The Development of a Hybrid Hierarchical/Heterarchical Shop Floor Control Applying Bidding Method in Job Dispatching* Robotics and Computer Integrated Manufacturing, 1998. **14**: p. 199-217
76. Overmars, H. and D.J. Tonich, *Hybrid FMS Control Architectures Based on Holonic Principles.* International Journal of Flexible Manufacturing Systems, 1996. **8**: p. 263-278.
77. Ottaway, T.A. and J.R. Burns, *Adaptive, Agile Approaches to Organizational Architecture Utilizing Agent Technology* Decision Science, 1997. **28**(3): p. 483-511.
78. Ottaway, T.A and J.R. Burns, *An Adaptive Production Control System Utilizing Agent Technology* International Journal of Production Research, 2000. **38**(4): p. 721-737.
79. Maturana, F., *MetaMorph: An Adaptive Multi-Agent Architecture for Advanced Manufacturing Systems.* 1997, University of Calgary. Calgary,Canada.
80. Maturana, F., W. Shen, and D.H. Norrie, *Metamorph An Adaptive Agent-Based Architecture for Intelligent Manufacturing.* International Journal of Production Research, 1999. **37**(10): p. 2159-2173.

81. Shen, W. and D.H. Norrie, *Agent-Based Systems for Intelligent Manufacturing. A state-of-the-Art Survey.* Knowledge and Information System, an International Journal, 1999. **1**(2): p. 129-156.
82. Valckenaers, P , H.V. Brussel, J. Wyns, L. Bongaerts, and P. Peeters, *Designing Holonic Manufacturing Systems.* Robotics and Computer Integrated Manufacturing, 1998. **14**: p. 455-464.
83. Sousa, P. and C. Ramos, *A Dynamic Scheduling Holon for Manufacturing Orders* Journal of Intelligent Manufacturing, 1998. **9**(2): p. 107-113.
84. Wyns, J., *Reference Architecture for Holonic Manufacturing Systems. The Key to Support Evolution and Reconfiguration.* 1999, Katholieke University. Leuven, Belgium.
85. Bongaerts, L , H.V. Brussel, and P. Valckenaers. *Scheduling Execution Using Perturbation Analysis.* in *Proceeding of 1998 IEEE International Conferenceon Robotics and Automation.* 1998 Leuven, Belgium.
86 Bussmann, S. *An Agent-Oriented Architecture for Holonic Manufacturing Control.* in *Proceeding of 1st International Workshop on IMS.* 1998. Lausanne, Switzerland.
87. Bussmann, S. and D C. McFarlane. *Rationales for Holonic Manufacturing Control.* in *Proceeding of 2nd International Workshop on IMS* 1999. Leuven, Belgium.
88. McFarlane, D.G. and S. Bussmann, *Developments in Holonic Planning and Control.* Production Planning and Control, 2000. **11**(6): p. 522-536.
89. Gou, L , P.B. Luh, and Y. Kyoya, *Holonic Manufacturing Scheduling: Architecture, Cooperation Mechanism and Implementation* Computers in Industries, 1998 **37**: p. 213-231.
90. Luh, P. and D J. Hoitomt, *Scheduling of Manufacturing Systems Using the Lagrangian Relaxation Technique* IEEE Transactions on Automatic Control, 1993. **38**(7): p 1066-1079.
91. Giebels, M.M.T., *Etoplan: A Concept for Concurrent Manufacturing Planning and Control-Building Holarchies for Manufacture-to-Order Environments.* 2000, University of Twente, Netherlands.
92 Okino, N. *A Prototyping of Bionic Manufacturing System.* in *Proceeding of ICOOMS'92.* 1992
93. Okino, N. *Bionic Manufacturing Systems, Flexible Manufacturing Systems, Past, Present, Future.* in *CIRP, Faculty of Mechanical Engineering.* 1993. Ljubljana.
94 Ueda, K. *An Approach to Bionic Manufacturing Systems Based on DNA-Type Information.* in *Proceeding of ICOOMS '92.* 1992.
95. Ueda, K. *A Genetic Approach Toward Future Manufacturing Systems, Flexible Manufacturing Systems, Past, Present, Future.* in *CIRP.* 1993 Faculty of Mechanical Engineering, Ljubljana.
96. Ueda, K. and K. Ohkura. *A Biological Approach to Complexity in Manufacturing systems.* in *Proceeding of the 27th CIRP Seminar on Manufacturing Systems.* 1995 Michigan, Iwata.
97. Iwata, K. and K. Onosato, M., *Random Manufacturing System: A New Concept of Manufacturing Systems for Production to Order.* Annals of the CIRP, 1994. **43**(1): p. 379-384.
98. Roy, D., D. Anciaux, and F. Vernadat, *SYROCO: A Novel Multi-Agent Shop Floor Control System.* Journal of Intelligent Manufacturing, 2001. **12**(3): p. 295-308

99. Peeters, P., H.V. Brussel, P. Valckenaers, J. Wyns, L. Bongaerts, M. Kollingbaum, and T. Heikkila, *Pheromone Based Emergent Shop Floor Control System for Flexible Flow Shop.* Artificial Intelligence in Engineering, 2001. **15**: p. 343-352.
100. Chen, R S., K.Y. Lu, S C. Yu, H.W. Tzeng, and C.C. Chang, *A Case Study in Design of BTO/CTO Shop Floor Control System* Information & Management, 2003. **41**: p. 23-37.
101. Wullink, G., A.J.R.M. Gademann, E W. Hans, and A.V. Harten, *Scenario-Based Approach for Flexible Resource Loading under Uncertainty.* International Journal of Production Research, 2004. **42**(24): p. 5079-5098.
102. Joshi, S.B., J.S Smith, R.A Wysk, and C.D Pegden. *Software for Control of FMS - The RAPIDCIM Concept and Approach. Intelligent automation and soft computing: trends in research, development, and applications.* in *Proceeding. of the WAC94, World Automation Congress.* 1994
103. Joshi, S B., J.S. Smith, and C.D. Pegden. *RapidCIM: An Approach to Rapid Development of Control Software for FMS Control.* in *Proceeding of the 27th CIRP Seminar on Manufacturing Systems.* 1995. Michigan
104 Smith, J.S., W C. Hoberecht, and S.B. Joshi, *A Shop Floor Control Architecture for Computer Integrated Manufacturing.* Texas A&M University and Pennsylvania State University, 1993.
105. Smith, J.S. and S.B Joshi, *Message-Based Part State Graph (MPSG): A Formal Model for Shop Floor Control* Department of Industrial Engineering, Texas A&M University, 1994
106 Manuel, J M l., R.A. Wysk, J. Hong, and V.V. Prabhu, *A Hybrid Shop-Floor Control System for Food Manufacturing.* IIE Transactions, 2001. **33**: p. 193-202.
107. Son, Y.J., R A Wysk, and A.T. Jones, *Simulation-Based Shop Floor Control. Formal Model, Model Generation and Control Interface.* IIE Transactions, 2003. **35**(1): p. 29-48.
108. Qiu, R., R. Wysk, and Q. Xu, *An Extended Structured Adaptive Supervisory Control Model of Shop Floor Controls for an e-Manufacturing System* International Journal of Production Research, 2003. **41**(8) p. 1605-1620.
109. Shen, W., Y T. Lang, and L. Wang, *iShopFloor: An Internet-Enabled Agent-Based Intelligent Shop Floor.* IEEE Transactions on Systems, Man, and Cybernetics—Part C: Applications and Reviews, 2005. **35**(3): p. 371-381.
110. Lima, R.M., R.M. Sousa, and P.J. Martins, *Distribute Production Planning and Control Agent Based System.* International Journal of Production Research, 2006. **44**(18-19): p. 3693-3709.
111. Melnyk, S.A. and P L. Carter, *Production Activity Control*, ed. R.D. Irwin. 1987 Homewood, IL
112. Park, S., P.Y. Jang, and H. Cho, *Petri-Net Based Rapid Development of a Task Execution Module of Equipment Controller for Distributed Shop Floor Control* Computers in Industries, 2001. **45**: p. 155-175.
113. Son, Y.J. and R.A. Wysk, *Automatic Simulation Model Generation for Simulation-Based, Real-Time Shop Floor Control* Computers in Industries, 2001. **45**: p 291-308.
114 Monch, L , M. Stehli, and J. Zimmermann, *FABMAS An Agent Based System for Production Control of Semiconductor Manufacturing Processes* Lecture Notes in AI, 2003. **2744**. p. 258-267.

115. Rogers, P. and R.J. Gordon *Simulation for Real-Time Decision Making in Manufacturing Systems.* in *25th Conference on Winter simulation* 1993 Los Angeles, California: ACM Press.
116. Lastra, J.L.M. and A.W. Colombo, *Engineering Framework for Agent Based Manufacturing Control.* Engineering Applications of Artificial Intelligence, 2006. **19**: p. 625-640.
117 Garbot, B., J.C. Blame, and C. Binda, *A Decision Support System for Production Activity Control.* Decision Support Systems, 1996. **16**: p. 87-101.
118. Chan, F.T.S. and J. Zhang, *A Multi Agent Based Agile Shop Floor Control System.* International Journal of Advanced Manufacturing Technology, 2002. **19**(764-774).
119. Heragu, S.S., R J. Graves, B.I. Kim, and A.S. Onge, *Intelligent Agent Based Framework for Manufacturing Systems Control* IEEE Transactions on Systems, Man and Cybernetics- Part A: Systems and Humans, 2002 **32**(5): p. 560-573.
120. Morton, T.E. and D. Pentico, *Heuristic Scheduling System.* 1993, Ny· Wiley.
121. Metaxiotis, K.S., D. Askounis, and J. Psarras, *Expert Systems in Production Planning And Scheduling: A State-of-the-Art Survey* Journal of Intelligent Manufacturing, 2002 **13**(4). p. 253-260.
122. Iwamura, K , N. Okubo, Y. Tanimizu, and N. Sugimura, *Real-Time Scheduling for Holonic Manufacturing System Based on Estimation of Failure Status* International Journal of Production Research, 2006. **44**(18-19): p. 3657-3675.
123. Sun, Y.L., T.M. Chang, and Y. Yih, *Learning Based Adaptive Controller for Dynamic Manufacturing Cells.* International Journal of Production Research, 2005. **43**(14,15) p. 3011-3025
124 Shen, W. and D.H. Norrie, *An Agent-Based Approach for Information and Knowledge Sharing in Manufacturing Enterprise Networks* International Journal of Networking and Virtual Organisations, 2004. **2**(2): p. 173-190.
125. Patriotta, G., *Sense-Making on the Shop Floor· Narratives of Knowledge in Organizations* Journal of Management Studies, 2003. **40**(2): p 349-375.
126. Eberts, R.E. and S Y. Nof, *Distributed Planning of Collaborative Production* International Journal of Advance Manufacturing Technology, 1993. **8**: p 57-71.
127. Hewitt, C , *Viewing Control Structures as Patterns of Passing Messages* Artificial Intelligence, 1977. **8**(3) p. 323-364.
128. Franklin, S and A. Graesser. *Is it an Agent, or just a Program?. A Taxonomy for Autonomous Agents.* in *Proceedings of the Third International Workshop on Agent Theories, Architectures, and Languages* 1996: Springer-Verlag.
129. Wooldridge, M. and N. Jennings, *Intelligent Agents: Theory and Practice.* The Knowledge Engineering Review, 1995. **10**(2): p. 115-152.
130 Chiariglione, L., *Foundation For Intelligent Physical Agents (FIPA) - Rationale.* 1996. p. http·//www cselt.it/fipa_rationale htm.
131. Jennings, N R. and M.J. Wooldridge, *Applications of Intelligent Agents.*, in *Agent Technology· Foundations, Applications, and Markets*, N.R. Jennings and M.J. Wooldridge, Editors. 1998, Springer. p. 3-28
132. Westkamper, E., A. Ritter, and C. Schaeffer. *Asimov-Holonic Multi-agent system for AGV's.* in *Proceedings of the 2nd International Workshop on Intelligent Manufacturing Systems.* 1999.

133. Zhu, Q., *Topologies of Agents Interactions in Knowledge Intensive Multi-Agent Systems for Networked Information Services.* Advanced Engineering Informatics 2006. **20**(1): p. 31-45.
134. Balakrishnan, A., S R.T. Kumara, and S. Sundaresan, *Manufacturing in the Digital Age: Exploring Information Technologies for Product Realization* Information Systems Frontiers, 1999. **1**: p. 25-50.
135. Lin, Y.J.G. and J J. Solberg, *Integrated Shop Floor Control Using Autonomous Agents* IIE Transactions, 1990. **24**(3): p. 57-71.
136. Parunak, H.V., A.D. Baker, and S.J. Clark *The AARIA Agent Architecture: An Example of Requirements-Driven Agent Based System Design.* in *1st International Conference on Autonomous Agents.* 1997. California.
137. Ouelhadj, D., C. Hanachi, and B. Bouzouia. *A Multi Contract Net Protocol for Dynamic Scheduling in Flexible Manufacturing Systems.* in *ICRA* 1999. Detriot
138. Ouelhadj, D., C. Hanachi, and B. Bouzouia *Multi Agent Architecture for Distributed Monitoring in Flexible Manufacturing Systems.* in *ICRA.* 2000. San Francisco.
139. Saad, A., K. Kawamura, and G. Biswas, *Performance Evaluation for Contract Net Based Heterarchical Scheduling for Flexible manufacturing Systems* Intelligent Automation and Soft Computing, 1997. **3**(3): p. 229-248.
140. Maturana, F. and D H. Norrie, *Multi-Agent Mediator Architecture for Distributed Manufacturing* Journal of Intelligent Manufacturing, 1996 **7**: p. 257-270.
141. Shen, W., F. Maturana, and D.H. Norrie, *Metaphor II: An Agent Based Architecture for Distributed Intelligent Design and Manufacturing.* Journal of Intelligent Manufacturing, 2000. **11**(3): p. 237-251.
142. Vancza, J. and A Markus, *An Agent Model for Incentive-Based Production Scheduling* Computers in Industry, 2000 **43**: p. 173-187
143. Ramos, C and P. Sousa, *A Distributed Architecture and Negotiation Protocol for Scheduling in Manufacturing Systems.* Computers in Industry, 1999. **38**(2): p. 103-113.
144 Tiwari, M K. and S. Mondal, *Application of an Autonomous Agent Network to Support the Architecture of a Holonic Manufacturing System* International Journal of Advanced Manufacturing Technology, 2002. **20** p. 931-942.
145. Sousa, P., C. Ramos, and J. Neves, *The Fabricare Scheduling Prototype Suite: Agent Interaction and Knowledge Base* Journal of Intelligent Manufacturing, 2003. **14**: p 441-455.
146. Cowling, P.I., D. Ouelhadj, and S. Petrovic, *A Multi-Agent Architecture for Dynamic Scheduling of Steel Hot Rolling* Journal of Intelligent Manufacturing, 2003. **14**: p. 457-470.
147. Arboleda, C.D.P. and T K. Das, *A Multi-Agent Reinforcement Learning Approach to Obtaining Dynamic Control Policies for Stochastic Lot Scheduling Problem* Simulation Modelling Practice and Theory, 2005. **(in Press)**.
148. Tripathi, A.K , M K Tiwari, and F.T.S. Chan, *Multi-Agent Based Approach to Solve Part Selection and Task Allocation Problem in Flexible Manufacturing Systems* International Journal of Production Research, 2005 **43**(7): p. 1313-13335.

149. Chan, F.T S , R. Swarnkar, and M.K. Tiwari, *Infrastructure for Coordination of Multi–Agent in a Network–Based Manufacturing System* International Journal of Advanced Manufacturing Technology, 2005
150. Tranvouez, E., B. Espinasse, and J P. Chirac. *A Multi Agent Based Scheduling System: A Cooperative and Reactive Approach* in *9th Symposium on Information Control in Manufacturing*. 1998. Nancy-Metz, France.
151. Sohier, C., B. Denis, and J.J. Lesage. *Applying Eco Problem Solving to the Control of an Adaptive Manufacturing Cell.* in *Multi Conference on Computational Engineering in Systems Applications*. 1998. Lille, France.
152 Peng, Y., T. Finin, Y. Labrou, R S. Cost, B. Chu, J. Long, W.J Tolone, and A. Boughannam, *Agent Based Approach for Manufacturing Integration· The Cimplex Experience* Applied Artifical Intelligence, 1999. **13**(1): p. 39-63.
153 Burn, A. and A. Portioli, *Agent Based Shop Floor Schedulingof Multi Stage System* Computers and Industrial Engineering, 1999 **37**: p. 457-460
154. Khoo, L.P , S G. Lee, and X.F. Yin, *Agent Based Multiple Shop Floor Manufacturing Scheduler.* International Journal of Production Research, 2001. **39**(14) p. 3023-3040.
155. Bochmann, O., P. valckenaers, and H.V. Brussel. *Negotiation Based Manufacturing Control in Holonic Manufacturing Systems.* in *ASI Conference* 2000.
156 Maione, B and D. Naso *Multi-Agent Routing Control in Heterarchical Manufacturing Systems.* in *IIA'99 ICSC Synposium on Intelligent Industrial Automation*. 1999.
157. Roy, D. and D Anicaux, *Shop-Floor Control: A Multi-Agent Approach.* International journal of Computer Integrated Manufacturing, 2001. **14**(6): p. 535-544.
158. Unver, H O. and O. Anlagan, *Design and Implementation of an Agent-Based Shop Floor Control System Using Windows-DNA* International Journal of Computer Integrated Manufacturing, 2002. **15**(5) p. 427-439
159 Odrey, N G. and G. Mejia, *A Re-Configurable Multi-Agent for Error Recovery in Production System* Robotics and Computer Integrated Manufacturing, 2003. **19**· p. 35-43.
160. Wang, K J., *Negotiation-Based Multi-Stage Production Control Using Distributed Shortest Path Algorithm* International Journal of Computer Integrated Manufacturing, 2003. **16**(1). p 38-47.
161. Frey, D., J. Nimis, H. Worn, and P. Lockemann, *Benchmarking and Robust Multi-Agent Based Production Planning and Control* Engineering Applications of Artificial Intelligence, 2003. **16**: p. 307-320.
162. Huang, C.Y. and S.Y. Nof, *Formation of Autonomous Agent Network for Manufacturing Systems.* International Journal of Production Research, 2000. **38**(3): p. 607-624.
163 Arbib, C and F. Rossi, *Optimal Resource Assignment Through Negotiation in a Multi-Agent Manufacturing System* IIE Transactions, 2000. **32**· p 963-974.
164. Ghiassi, M *Internet-Based Manufacturing Integration* in *10th International Conference on Flexible Automation and Intelligent Manufacturing*. 2000.
165 Barber, K S., T.H. Liu, and S. Ramaswamy, *Conflict Detection During Plan Integration for Multi-Agent Systems* IEEE Transactions on Systems, Man and Cybernetics- Part B: Cybernetics, 2001. **31**(4): p. 616-628.
166. Wang, Y H., C.W. Yin, and Y. Zhang, *A Multi-Agent and Distributed Ruler Based Approach to Production Scheduling of Agile Manufacturing Systems*

International Journal of Computer Integrated Manufacturing, 2003. **16**(2): p. 81-92.
167. Gorodetski, V., O. Karsaev, and V. Konushy, *Multi-Agent System for Resource Allocation and Scheduling* Lecture Notes in AI, 2002. **2691**· p 236-246.
168. Karageorgos, A., N. Mehandjiev, G. Weichhart, and A Hammerle, *Agent Based Optimisation of Logistics and Production Planning* Engineering Applications of Artificial Intelligence, 2003. **16**: p. 335-348.
169. Naso, D. and B. Turchiano, *A Coordination Strategy for Distributed Multi-Agent Manufacturing Systems.* International Journal of Production Research, 2004. **42**(12): p. 2497-2520.
170 Wu, Z. and M.X. Weng, *Multiagent Scheduling Method With Earliness and Tardiness Objectives in Flexible Job Shops.* IEEE Transactions on Systems, Man and Cybernetics- Part B: Cybernetics, 2005. **35**(2): p. 293-301.
171. Wong, T N , C.W. Leunga, K L. Maka, and R.Y.K. Fung, *Dynamic Shopfloor Scheduling in Multi-Agent Manufacturing Systems* Expert Systems with Applications, 2006 **31**(3): p. 486-494
172 Trentesaux, D , R. Dindeleux, and C. Tohan, *A Multi-Criteria Decision Support System for Dynamic Task Allocation in a Distributed Production Activity Control System.* International journal of Computer Integrated Manufacturing, 1998. **11**(1)· p. 3-17.
173. Ming, L., Y. Xiaohong, M.M. Tseng, and Y. Shuzi, *A Corba-Based Agent-Driven Design for Distributed Intelligent Manufacturing System* Journal of Intelligent Manufacturing, 1998 **9** p. 457-465.
174 Xue, D., J. Sun, and D.H. Norrie, *An Intelligent Optimal Production Scheduling Approach Using Constraint-Based Search and Agent Based Collaboration* Computers in Industry, 2001. **46**: p. 209-231.
175. Yen, P.C.B. and O Q. Wu, *Internet Scheduling Environment with Market Driven Agents* IEEE TRANsactions on Systems, Man and Cybernetics- Part A: Systems and Humans, 2004. **34**(2): p. 281-289.
176. Lim, M.K. and D Z. Zhang, *An Integrated Agent Based Approach for Responsive Control of Manufacturing Resources.* Computers and Industrial Engineering, 2004 **46**(2): p. 221-232
177. Cavalieri, S , M. Garetti, M. Macchi, and M. Taisch, *An Experimental Bechmarking of Two Multi-Agent Architectures for Production Scheduling and Control* Computers in Industry, 2000. **43**· p. 139-152.
178. Shen, W , L. wang, and A. Hao, *Agent Based Distributed Manufacturing Process Planning and Scheduling. A State-of-Art Survey* IEEE Transactions on Systems, Man and Cybernetics- Part C: Applications and Reviews, 2006 **36**(4)· p. 563-577.
179. Shen, W., *Distributed Manufacturing Scheduling Using Intelligent Agents* IEEE Intelligent Systems, 2002. **17**: p. 88-94.
180. Usama, M F., G. Piatetsky-Shapiro, P.Smyth, and R.Uthurusamy,, *Advances in Knowledge Discovery and Data Mining.* 1996: AAAI/MIT Press.
181. Brierley, P. and B. Batty, *Knowledge Discovery and Data Mining*, ed M.A. Bramer. Vol. 1. 1998: IEE Professional Applications of Computer Series. 240-303.
182. Han, J. and M. Kamber, *Data Mining· Concepts and Techniques.* 2001: Morgan Kaufmann Publishers. 550.

183. Klosgen, W. and J.M. Zytkow., *Handbook of Data Mining and Knowledge Discovery*. 2002: Oxford University press.
184. Adriaans, P. and D. Zantinge, *Data Mining*. 1996, Harlow, England: Addison-Wesley.
185. Fayyad, U. and E Simoudis. *Data Mining and KDD: An Overview*. in *Tutorial in the Third International conference on Knowledge Discovery and Data Mining*. 1997 Newport Beach, California.
186. Gertosio, C and A. Dussauchoy, *Knowledge Discovery from Industrial databases* Journal of Intelligent Manufacturing, 2004. **15**: p 29-37.
187. SPSS, C.-D.S.-b -S.D M.G., http://www.crisp-dm org/.
188. EnterpriseMiner, S., http //www sas.com/technologies/analytics/datamining/miner/semma html.
189. Kurgan, L.A. and P. Musilek, *A survey of Knowledge Discovery and Data Mining Process Models*. Knowledge Engineering Review, 2006 **21**(1). p. 1-24.
190. Witten, I.H. and E. Frank, *Tools for Data Mining, Practical Machine Learning Tools and Techniques*. 1999: Morgan Kaufmann. 371.
191. Montgomery, D C., *Introduction to Statistical Quality Control*. Vol. 4. 2001, NY: John Wiley & Sons Inc.
192. Breiman, L., J. Friedman, R. Olshen, and C. Stone, *Classification and Regression Trees* 1984, Monterey, CA: CRC Press.
193. Agrawal, R , T. Imielinski, and A Swami. *Mining Associations Between Sets of Items in Massive Databases*. in *ACM SIGMOD International Conference on Management of Data*. 1993. Washington, D.C.
194. Kaufman, L. and P.J. Rousseeuw, *Finding Groups in Data. An Introduction to Cluster Analysis*. 1990, New York: John Wiley & Sons.
195. Breslow, L.A. and D.W. Aha, *Simplifying Decision Trees: A survey* Knowledge Engineering Review, 1997. **12**. p. 1-40.
196. Caudill, M. and C. Butler, *Understanding Neural Networks: Computer Explorations*. 1992, Massachusetts: The MIT Press
197 Pawlak, Z , *Rough Sets* International Journal of Computer and Information Sciences, 1982. **11**: p. 341-356
198. Komorowski, J., L. Polkowski, and A. Showron, *Rough Sets. A Tutorial* Rough Fuzzy Hybridization, A New Trend in Decision Making
1999, Singapore: Springer Verlag. 3-98.
199. Pal, S.K., S. De, and A. Ghosh, *Genotypic and Phenotypic Assertive Mating in Genetic Algorithm* Journal of Information Sciences, 1998. **105**: p. 209-226.
200. Michael, L.G. and R.G. Bell, *Data Mining – A Powerful Information Creating Tool. OCLC Systems and Services*. Vol. 15 1999. MCB University Press. 81-90.
201. Claudio, M. and G. Grinstein. *Data Mining*. 1998: WIT Press/ Computational Mechanics Publications.
202. Moustakis, V.S , M. Letho, and G. Salvendy, *Survey of Expert Opinion: Which Machine Learning Method may be Used for Which Task?* Machine Learning of International Journal of HCI, 1996. **8**(3): p. 221-236.
203. Pieter, A. and Z. Dolf, *Data Mining*. 1996: Addison-Wesley.
204. Harding, J.A , M. Shahbaz, Srinivas, and A. Kusiak, *Data Mining in Manufacturing: A Review* Transactions of the American Society of Mechanical Engineers ASME: Journal of Manufacturing Science and Engineering, 2006. **128**: p. 969 - 976.

205 Malkoff, D.B., *A Framwork for Real-Time Fault Detection and Diagnosis Using Temporal Data* Artificial Intelligence in Engineering, 1987 **2**(2): p. 97-111.
206. Ramamoorthy, C.V. and B.W. Wah, *Knowledge and Data Engineering.* IEEE Transactions on Knowledge and Data Engineering, 1989. **1**(1): p. 9-16.
207. Lee, M.H., *Knowledge Based Factory.* Artificial Intelligence in Engineering, 1993. **8**: p. 109-125.
208. Irani, K.B., J. Cheng, U.M. Fayyad, and Z. Qian, *Applying Machine Learning to Semiconductor Manufacturing.* IEEE Expert, 1993(Feb): p. 41-47.
209. Piatetsky-Shapiro, G., *The Data Mining Industry Coming of Age* IEEE Intelligent Systems, 1999(Nov-Dec): p. 32-34.
210 Shao, X.Y., Z.H Wang, P.G. Li, and C.X.J. Feng, *Integrating Data Mining and Rough Set for Customer Group Based Discovery of Product Configuration Rules* International Journal of Production Research, 2006. **44**(15): p. 2789-2811.
211. Jin, Y. and Y. Ishino, *DAKA. Design Activity Knowledge Acquisition Through Data Mining* International Journal of Production Research, 2006. **44**(15): p. 2813-2837.
212. Browne, W., L. Yao, I. Postlethwaite, S. Lowes, and M. Mar, *Knowledge Elicitation and Data Mining: Fusing Human and Industrial Plant Information* Engineering Applications of Artificial Intelligence, 2006 **19**: p. 345-359.
213. Hsu, C.H. and M.J. Wang, *Using Decision Tree Based Data Mining to Establish a Sizing System for The Manufacture of Garments.* International Journal of Advance Manufacturing Technology, 2005. **26**: p. 669-674.
214 Vullers, M.H J , W.M.P. Van-der-Aalst, and M. Rosemann, *Mining Configurable Enterprise Information Systems.* Data and Knowledge Engineering, 2006. **56**. p. 195-244.
215. Li, T.S , C L. Huang, and Z.Y. Wu, *Data Mining Using Genetic Programming for Construction of a Semiconductor Manufacturing Yield Rate Prediction System* Journal of Intelligent Manufacturing, 2006. **17**: p. 335-361.
216. Holden, T. and M Serearuno, *A Hybrid Artificial Intelligence Approach for Improving Yield in Precious Stone Manufacturing* Journal of Intelligent Manufacturing, 2005. **16**: p. 21-38.
217. Sadoyan, H., A. Zakarian, and P. Mohanty, *Data Mining Algorithm for Manufacturing Process Control* International Journal of Advance Manufacturing Technology, 2006. **28**· p. 342-350.
218. Ren, Y , Y. Ding, and S. Zhou, *A Data Mining Approach to Study the Significance of Nonlinearity in Multistation Assembly Processes* IIE Transactions, 2006. **38**: p. 1069-1083.
219. Rokach, L. and O. Maimon, *Data Mining for Improving the Quality of Manufacturing: A Featured Set Decomposition Approach.* Journal of Intelligent Manufacturing, 2006. **17**: p. 285-299.
220. Lau, H.C.W., A. Ning, W.H. Ip, and K L. Choy, *A Decision Support System to Facilitate Resource Allocation. An OLAP-Based Neural Network Approach* Journal of Manufacturing Technology Management, 2004. **15**(8): p. 771-778.
221. Kaya, M. and R. Alhajı, *Fuzzy OLAP Association Rules Mining-Based Modular Reinforcement Learning Approach for Multiagent System.* IEEE Transactions on Systems, Man and Cybernetics- Part B: Cybernetics, 2005. **35**(2): p. 326-338.

222. Hamilton-Wright, A. and D.W. Stashuk, *Transparent Decision Support Using Statistical Reasoning and Fuzzy Inference* IEEE Transactions on Knowledge and Data Engineering, 2006. **18**(8): p. 1125-137.
223. Kusiak, A., *Data Mining: Manufacturing and Service Applications* International Journal of Production Research, 2006. **44**(18-19): p. 4175-4191.
224. Sha, D Y. and C.H. Liu, *Using Data Mining for Due Date Assignment in a Dynamic Job Shop Environmnet.* International Journal of Advance Manufacturing Technology, 2005. **25**: p. 1164-1174.
225. Bakus, P., M. Janakiran, S. Mowzoon, G C. Runger, and A. Bhargava, *Factory Cycle Time Prediction With a Data Mining Approach* IEEE Transactions on Semiconductor Manufacturing, 2006. **19**(2): p. 252-258.
226. Li, X. and S. Olafsson, *Discovering Dispatching Rules Using Data Mining.* Journal of Scheduling, 2005. **8**: p. 515-527.
227. Cunha, C.D., B. Agard, and A. Kusiak, *Data Mining for Improvement of Product Quality* International Journal of Production Research, 2006 **44**(18): p 4027-4041.
228. Li, J R., L.P. Khoo, and B. Tor, *RMIE· A Rough Set Based Data Mining Prototype for the Reasoning of Incomplete Data in Condition Based Fault Diagnosis.* Journal of Intelligent Manufacturing, 2006 **17**: p. 163-176
229. Buddhakulsomsiri, J., Y. Siradeghyan, A. Zakarian, and X. Li, *Association Rule Generation Algorithm for Mining Automative Warranty Data.* International journal of Production Research, 2006. **44**(14-15): p. 2749-2770.
230. Dengiz, O , A.E. Smith, and I. Nettleship, *Two-Stage Data Mining for Flaw Identification in Ceramics Manufacture* International Journal of Production Research, 2006. **44**(14-15): p. 2839-2851.
231. Hou, Z., Z Lian, Y. Yao, and X. Yuan, *Data Mining Based Sensor Fault Diagnosis and Validation for Building Air Conditioning System* Energy Conservation and Management, 2006. **47**(15-16): p. 2479-2490.
232. Feng, C.X J., Z.G. Yu, U. King, and M P Baig, *Threefold vs Fivefold Cross Validation in One-Hidden Layer and Two Hidden Layer Predictive Neural Network Modelling of Machine Surface Roughness Data* Journal of Manufacturing Systems, 2005. **24**(2)· p 1-15.
233. Raheja, D., J. Llinas, R. Nagi, and C. Romanowski, *Data Fusion/Data Mining Based Architecture for Condition Based Maintenance.* International Journal of Production Research, 2006. **44**(14-15)· p 1869-2887.
234. Symeonidis, A.L., D.D. Kehagias, and P.A Mitkas, *Intelligent Policy Recommendation on Enterprise Resource Planning by the use of Agent Technology and Data Mining Techniques* Expert Systems with Applications, 2003. **25**: p. 589-602
235. Tseng, T.L., C.C. Huang, F. Jiang, and J.C. Ho, *Applying a Hybrid Data Mining Approach to Prediction Problems: A Case of Preferred Suppliers Prediction.* International journal of Production Research, 2006. **44**(14-15). p. 2935-2954.
236. Qian, Z., W. Jiang, and K.L. Tsui, *Churn Detection via Customer Profile Modelling* International journal of Production Research, 2006. **44**(14-15): p. 2913-2933.
237. Crespo, F. and R. Weber, *A Methodology for Dynamic Data Mining Based on Fuzzy Clustering* Fuzzy Sets and Systems, 2005. **150**(2): p. 267-284.

238. Cho, H and R.A. Wysk, *Intelligent Workstation Controller for Computer Integrated Manufacturing· Problems and Models.* Journal of Manufacturing System, 1995. **14**(4): p. 252-263.
239. Pinedo, M., *Scheduling-Theory, Algorithm and Systems.* 1995, NJ: Prentice Hall.
240. Hong, J., V. Prabhu, and R.A. Wysk, *Real-Time Batch Sequencing Using Arrival Time Control Algorithm.* International Journal of Production Research, 2001. **39**(17): p. 3863 - 3880.
241 Srinivas, M.K. Tiwari, and V. Allada, *Solving The Machine-Loading Problem in a Flexible Manufacturing System Using A Combinatorial Auction-Based Approach* International Journal of Production Research 2003. **42**(9): p. 1879-1893.
242. Lau, H.C.W., A. Ning, K.F. Pun, K.S. Chin, and W.H. Ip, *A Knowledge Based System To Support Procurement Decision.* Journal of Knowledge Mangement, 2005. **9**(1): p. 87-100
243. Tiwana, A., *The Knowledge Management ToolKit.* 2000, Prentice-Hall: EngleWood Cliffs, NJ
244. Pham, D.T. and A.A. Afify, *Machine-Learning Techniques and Their Applications in Manufacturing* Proceeding of IMechE, Part B: Journal of Engineering Manufacture, 2005. **219**: p. 395-412.
245. Davenport, T.H. and l Prusak, *Working Knowledge.* 2000, Boston, MA: Harvard Business School Press.
246. Shiue, Y.R. and C.T. Su, *An Enhanced Knowledge Representation for Decision Tree Based Learning Adaptive Scheduling* International Journal of Computer Integrated Manufacturing, 2003. **16**(1). p. 48-60.
247. Koonce, D.A and S.C. Tsai, *Using Data Mining to Find Patterns in Genetic Algorithm Solutions to a Job Shop Schedule* Computers and Industrial Engineering, 2000. **38** p. 361-374.
248 Lee, J.H. and S C Park, *Agent and Data Mining Based Decision Support System and its Adaptation to a New Customer-Centric Electronic Commerce.* Expert Systems with Applications, 2003. **25**: p. 619-635.
249. Guo, H., W.-C. Hou, F. Yan, and Q. Zhu. *A Monte Carlo Sampling Method for Drawing Representative Samples from Large Databases.* in *Scientific and Statistical Database Management.* 2004. Greece.
250. Maki, H. and Y. Teranishi, *Development of Automated Data Mining System for Quality Control in Manufacturing* Lecture Notes in Computer Science, Springer-Verlag Berline, 2001. **2114**: p 93-100.
251. Young, R I., A G. Gunendran, and A.F. Cutting-Decelle. *Sharing manufacturing knowledge in PLM: are manufacturing ontologies the answer? Advances in Manufacturing Technology and Management.* in *Proceedings of the 3rd International Conference on Manufacturing Research.* 2005. Cranfield University, UK.
252. Young, R.I., *Informing Decision Makers in Product Design and Manufacture.* International Journal of Computer Integrated Manufacture, 2003. **16**(6): p 428-438.
253. Gunendran, A.G. and R I. Young, *An Information and Knowledge Framework for Multi Perspective Design and Manufacture.* International Journal of Computer Integrated Manufacture, 2006. **19**(4): p. 326-338.
254. Kusiak, A., *A Data Mining Approach for Generation of Control Signatures.* Journal of Manufacturing Science and Engineering, 2002. **124**(4): p. 923-926.

255. Shahbaz, M., *Product and Manufacturing Process Improvement Using data Mining*, in *Mechanical Nad Manufacturing*. 2005, Loughborough University: Loughborough. p. 199.
256. Shahbaz, M., Srinivas, J.A. Harding, and M Turner, *Product Design and Manufacturing Process Improvement Using Association Rules*. Proceedings of the Institution of Mechanical Engineers, Part B: Journal of Engineering Manufacture, 2006
257. Meneses, C.J. and G.G. Grinstein, *Categorization and Evaluation of Data Mining Techniques* Data Mining, ed. N.F.F. Ebecken. 1998. WIT Press/Computational Mechanics Publications. 53-79.
258. Dunham, M H., *Data Mining, Introductory and Advanced Topics*. 2003, New Jersey: Pearson Education Inc. 315.
259. Baliga, J , *Advanced Process Control: Soon to be a Must*, in *Semiconductor International*. 1999. p. 76-88.
260. Taguchi, G., E Elsayed A, and T. Hsiang, *Quality Engineering and Production Systems*. 1989: McGraw-Hill Publishing Company.
261. Schmidt, S.R. and R.G. Launsby, *Understanding Industrial Designed Experiments* 4 ed 1998, Colarado Springs, USA: Air Academy Press and Asssociates.
262. Phadke, M S , *Quality Engineering Using Robust Design*. 1989, NJ: Prentice Hall, Englewood Cliffs.
263. Kear, F.W., *Statistical Process Control in Manufacturing Practice*. 1998, New Maxico: Marcel Dekker Inc. NY.
264. Williams, J.H., A. Davies, and P R. Drake, *Condition-Based Maintenance and Machine Diagnostics*. 1994, London. Chapman and Hall.
265. Tsang, A.H C., *Condition Based Maintenenace Tools and Decision Making* Journal of Quality in Mainteneace Engineering, 1995. **1**(3): p. 3-17.
266. Mobley, R E , *An Introduction to Predictive Maintenance*. 1990, New York· Van Nostrand Reinhold. 189.
267. Yam, R.C.M., P.W. Tse, L. Li, and P. Tu, *Intelligent Predictive Decision Support System for Condition Based Maintenance* International Journal of Advance Manufacturing Technology, 2001 **17** p. 383-391.
268 Shenoy, D. and B. Bhadury, *Maintenance Resource Management· Adapting MRP*. 1998, London: Taylor and Francis 144.
269 Lee, E.S. and M. Burdekin, *A Hole-Plate Artifact Design for the Volumetric Error Calibration of CMM* International Journal of Advance Manufacturing Technology, 2001. **17**: p. 508-515.
270. Trapet, E. and F. Waldele, *A Reference Object Based Method to Determine the Parametric Error Components of Coordinate Measuring Machine and Machine Tools*. Measurement, 1991. **9**(1): p 17-22.
271. Swadesh, J.K., *HPLC- Practical and Industrial Applications*. 2001, London: CRC Press 154-155.
272. Ardrey, B., *Liquid Chromatography- Mass Spectrometry· An Introduction*. 2003, Sussex: John Willey and Sons. 1-75.
273. Duckett, S. and B. Gilbert, *Foundations of Spectroscopy*. 2000, Oxford· Oxford Science Publications.
274. Barker, J., *Mass Spectrometry*. 1999, Sussex: Hohn Willey and Sons. 1-193.
275. Spectrometry, A.S f.M., *What is Mass Spectrometry*. 2001, http·//www asms org/whatisms/.

276. Downward, K., *Mass Spectrometry A Foundation Course*. 2004, Cambridge: The Royal Society OF Chemistry. 1-65.
277. Herbert, C.G. and R A.W. Johnstone, *Mass Spectrometry Basics*. 2003, Florida: CRC Press. 245-269.
278. Kaiser, R.E. and A.J. Rackstraw, *Computer Chromatography Volume 1*. 1983, Heidelberg: IFC.
279. Niessen, W M A , *Liquid Chromatograpgy- Mass Spectrometry*. 2001, New York· Marcel Dekker Inc.

APPENDIX I

J. A. Harding
e-mail j a harding@lboro ac uk

M. Shahbaz
e-mail m shahbaz@lboro ac uk

Srinivas
e-mail s srinivas@lboro ac uk

Wolfson School of Mechanical and Manufacturing Engineering,
Loughborough University,
Loughborough, Leicestershire LE2 4LA, UK

A. Kusiak
Department of Mechanical and Industrial Engineering,
The University of Iowa,
Iowa City, IA 52242-1527
e-mail andrew-kusiak@uiowa edu

Data Mining in Manufacturing: A Review

The paper reviews applications of data mining in manufacturing engineering, in particular production processes, operations, fault detection, maintenance, decision support, and product quality improvement Customer relationship management, information integration aspects, and standardization are also briefly discussed This review is focused on demonstrating the relevancy of data mining to manufacturing industry, rather than discussing the data mining domain in general The volume of general data mining literature makes it difficult to gain a precise view of a target area such as manufacturing engineering, which has its own particular needs and requirements for mining applications This review reveals progressive applications in addition to existing gaps and less considered areas such as manufacturing planning and shop floor control [DOI 10 1115/1 2194554]

1 Introduction

Knowledge is the most valuable asset of a manufacturing enterprise, as it enables a business to differentiate itself from competitors and to compete efficiently and effectively to the best of its ability Knowledge exists in all business functions, including purchasing, marketing, design, production, maintenance and distribution, but knowledge can be notoriously difficult to identify, capture, and manage Knowledge can be as simple as knowing who is best to contact if particular materials are running short, or can be as complex as mathematical formulas relating process variables to finished product dimensions Spiegler [1] reviewed two models of knowledge The first model follows a conventional hierarchy and transformation of data into information and knowledge with a spiral and recursive way of generating knowledge The second model presents a reverse hierarchy where knowledge may appear before data and information processing Knowledge discovery, knowledge management, and knowledge engineering are currently topics of importance to manufacturing researchers and managers intent on exploiting current assets Database technology is central to all these knowledge-based research topics

The use of databases and statistical techniques are well established in engineering [2] The first applications of artificial intelligence in engineering in general and in manufacturing in particular were developed in the late 1980s [3,4] The scope of these activities, however, has recently changed The advancements in information technology (IT), data acquisition systems, and storage technology as well as the developments in machine learning tools have enticed researchers to move forward toward discovering knowledge from databases (KDD) Data from almost all the processes of the organization such as product and process design, material planning and control, assembly, scheduling, maintenance, recycling, etc , are recorded These data stores therefore offer enormous potential as sources of new knowledge Making use of the collected data is becoming an issue and data mining is a natural solution for transforming the data into useful knowledge The extracted knowledge can be used to model, classify, and make predictions for numerous applications

The idea of finding patterns in manufacturing, design, business, or medical data is not new Databases have been processed to derive the underlying relationships within the data for many years as evidenced by the developments in statistics Traditionally, it was the responsibility of analysts, who generally used statistical techniques, but increasingly data mining, which is an emerging area of computational intelligence, is providing new systems, techniques, and theories for the discovery of hidden knowledge in large volumes of data Data mining is a blend of concepts and algorithms from machine learning, statistics, artificial intelligence, and data management With the emergence of data mining, researchers and practitioners began applying this technology on data from different areas such as banking, finance, retail, marketing, insurance, fraud detection, science, engineering, etc , to discover any hidden relationships or patterns Data mining is therefore a rapidly expanding field with growing interests and importance and manufacturing is an application area where it can provide significant competitive advantage

The use of data mining techniques in manufacturing began in the 1990s [5–7] and it has gradually progressed by receiving attention from the production community Data mining is now used in many different areas in manufacturing engineering to extract knowledge for use in predictive maintenance, fault detection, design, production, quality assurance, scheduling, and decision support systems Data can be analyzed to identify hidden patterns in the parameters that control manufacturing processes or to determine and improve the quality of products A major advantage of data mining is that the required data for analysis can be collected during the normal operations of the manufacturing process being studied and it is therefore generally not necessary to introduce dedicated processes for data collection Since the importance of data mining in manufacturing has clearly increased over the last 20 years, it is now appropriate to critically review its history and application

This paper presents a comprehensive overview of data mining applications in manufacturing, especially in the areas of production processes, control, maintenance, customer relationship management (CRM), decision support systems (DSS), quality improvement, fault detection, and engineering design The remainder of the paper briefly describes pertinent manufacturing enterprise applications where data mining is applied to extract knowledge for improvement The paper also approaches the challenging area

Contributed by the Manufacturing Engineering Division of ASME for publication in the JOURNAL OF MANUFACTURING SCIENCE AND ENGINEERING Manuscript received April 4 2005 final manuscript received December 9, 2005 Review conducted by C J Li

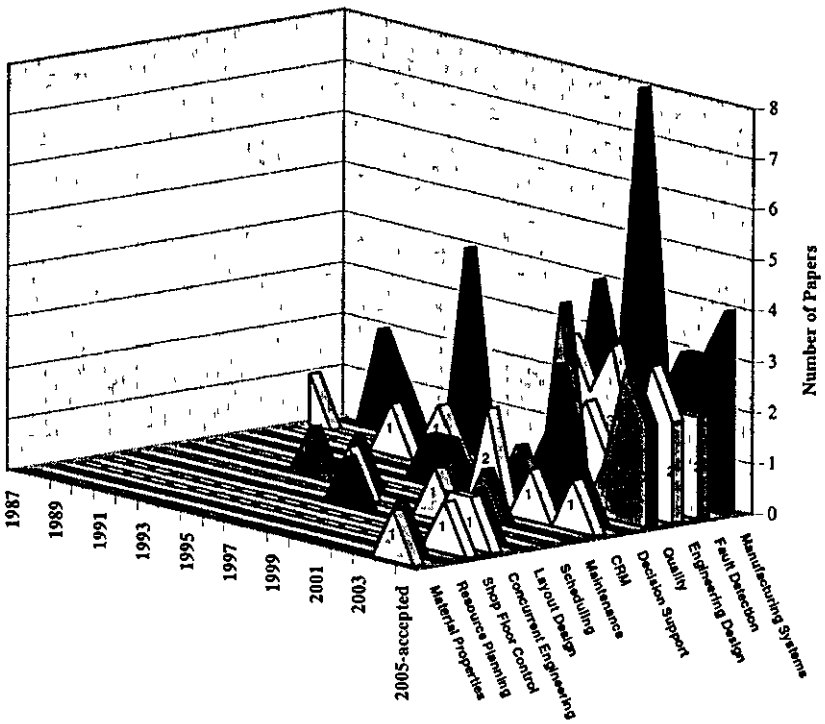


Fig 1 History of manufacturing applications of data mining

of data mining system integration Finally, the conclusions and future research directions outline the progress made by the ongoing research related to manufacturing control and quality improvement

2 Data Mining Models for Manufacturing Applications

CRISP-DM™ (Cross Industry Standard Process for Data Mining), *SEMMA*, *SolEuNet* (Data Mining and Decision Support for Business Competitiveness A European Virtual Enterprise), *Kensington Enterprise Data Mining* (Imperial College, Department of Computing, London, UK), and *Data Mining Group (DMG)* have established methodologies and developed languages and software tools for the standardization of industrial applications of data mining However, most products focus on the implementation of data mining algorithms and application development rather than on the ease-of-use, integration, scalability, and portability Most published research on data mining in manufacturing reports dedicated applications or systems, tackling specific problem areas, such as fault detection (see Fig 1) Only limited research has been done to address the integration of data mining with existing manufacturing-based enterprise reference architectures, frameworks, middleware, and standards such as Common Object Request Broker Architecture (CORBA), Model-Driven Architecture (MDA), or Common Warehouse Metamodel (CWM) Neaga [8] explained the neglect of these issues and highlighted their importance and the investments already committed to the existing efforts for enterprise integration Neaga and Harding [8,9] presented a holistic approach to a wide range of data mining applications suitable for manufacturing enterprises The areas of manufacturing enterprise design, engineering and re-engineering, information modeling and the suitability of applying data mining techniques to use previous knowledge and information about an enterprise are examined in Refs [8,9]

The CRISP-DM and SEMMA methodologies are most widely used by the data mining community CRISP-DM and SEMMA provide a step by step guide for data mining implementation CRISP-DM is easier to use than SEMMA as it provides detailed neutral guidelines that can be used by any novice in the data mining field SEMMA is developed as a set of functional tools for SAS's Enterprise Miner software Therefore those who use this specific software for their tasks are more likely to adopt this methodology Using SEMMA, results may be found quickly by mining samples of data from the whole database, but if the discovered relationships do not follow in the whole database then new samples must be examined which means repeating the whole data mining process

The details of each step of CRISP-DM [10] make it a reliable methodology that is easy to use and fast to implement The detailed sub-stages are optional guidelines and can be skipped as required The CRISP-DM methodology can therefore be fully or partially adopted depending on the problem and its requirements This was the main reason that the CRISP-DM was used by the authors as a standard guide to implement the data mining research that is reported in Ref [11]

Neaga and Harding also presented a framework for the integration of complex enterprise applications including data mining systems [12,13] The presented approaches provide the definition and development of a common knowledge enterprise model, which represents a combination of previous projects on manufacturing enterprise architectures and Object Management Group (OMG) models and standards related to data mining

3 Data Mining Applications Relevant to Manufacturing

This section details the contributions of researchers and practitioners in different areas of manufacturing from the late 1980s to date The literature was searched extensively in different journals,

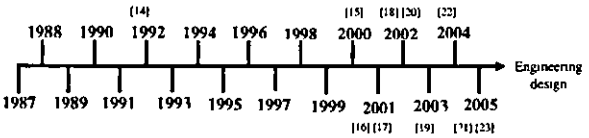


Fig 2 Engineering design literature over time

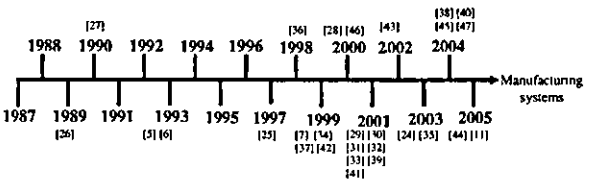


Fig 3 Manufacturing systems literature over time

personal web pages, internet and citeseers web sites [1] This review is particularly focused on data-mining applications and case studies in manufacturing and closely related fields

The temporal stacked area chart in Fig 1 shows the data mining research reported in different application areas of manufacturing It clearly indicates the current trends of industry toward applications of data mining and shows that particularly since the beginning of the new century people have started to focus on solving their problems using historical databases Areas such as manufacturing operations, fault detection, design engineering, and decision support systems have gained the attention of the research community, although there is still enormous potential for research in these areas Other areas like maintenance, layout design, resource planning, and shop floor control require even greater attention and further exploration

In each of the following subsections, a time series progress figure has been provided for quick reference to the history of data mining development and implementation in the particular area

3.1 Engineering Design. Engineering design is a multidisciplinary, multidimensional, and non-linear decision-making process where parameters, actions, and components are selected This selection is often based on historical data, information, and knowledge It is therefore a prime area for data mining applications and although as yet only a few papers have reported applications of data mining in engineering design (see Fig 2), this has been an area of increased research interests in recent years These recently published papers form an important part of this review paper due to the essential synergies between design and manufacturing The importance of considering how a product should be manufactured during the design stage and the constraints imposed on design by particular manufacturing processes and technologies have been accepted for many years There is indeed great potential for data mined knowledge to integrate manufacturing, product characteristics, and the engineering design processes

Sim and Chan [14] developed a knowledge-based system for the selection of rolling bearings They used heuristic knowledge supported by a manufacturer's catalogue to optimize design specifications by matching the temporal data of the new product against the knowledge base Kusiak et al [15] proposed a rough-set theory approach to predict product cost Ishino and Jin [16] used data mining for knowledge acquisition in design from the data obtained through observing design activities using a CAD system They developed a method called Extended Dynamic Programming to extract the knowledge Romanowski and Nagi [17] proposed a design system which supports the feedback of data mined knowledge from the life cycle data to the initial stages of the design process Giess et al [18,19] mined a manufacturing and assembly database of gas turbine rotors to determine and quantify relationships between the various balance and vibration tests and highlight critical areas This knowledge could then be fed back to the designers to improve tolerance decisions in the future design of components They used a decision tree at the initial stage to determine appropriate areas of investigation and to identify problems with the data At the next stage, a neural network was used to model the data Hamburg [20] applied data mining techniques to support product development by analyzing global environment aspects, market situation, strategy, philosophy, and culture of the manufacturing and customer behavior He utilized a decision-tree algorithm to mine and integrate the enterprise data in the product development Romanowski and Nagi [21,22] applied a data-mining approach for forming generic bills of materials (GBOMS), entities that represent the different variants in a product family and facilitate the search for similar designs and the configuration of new variants By combining data-mining approaches such as text and tree mining in a new tree union procedure that embodies the GBOM and design constraints in constrained XML, the technical difficulties associated with a GBOM are resolved

Kim and Ding [23] presented a data mining aided optimal design method capable of finding a competitive design solution with a relatively low computation cost They applied the method to facilitate the optimal design of fixture layout in a four-station SUV side panel assembly process The literature reviewed in this section is summarized in Fig 2

3.2 Manufacturing Systems. Data collection in manufacturing is common but its use tends to be limited to rather few applications Machine learning and computational intelligence tools provide excellent potential for better control of manufacturing systems (see Fig 3), especially in complex manufacturing environments where detection of the causes of problems is difficult Piatesky-Shapiro et al [7] argued that the data mining industry is coming of age However, this review of data mining in manufacturing shows that although there are several areas in manufacturing enterprises that have benefited from data-mining algorithms, there are still numerous areas that could benefit further [24] In manufacturing environments the need and importance of data collection is ever present for statistical process control purposes Lee [5] discussed and suggested several principles leading to a knowledge-based factory environment utilizing the data collected over several stages of the manufacturing-related processes A comparative study of implicit and explicit methods to predict the non-linear behavior of the manufacturing process, using statistical and artificial intelligence tools, was discussed by Kim and Lee [25]

Semiconductor manufacturing is complex and faces several challenges relating to product quality, scheduling, work in process, cost reduction, and fault diagnosis To overcome these problems several methods and systems have been developed, e g , Rule-Based Decision Support Systems (RBDSS) [26], CAQ [27], Knowledge Acquisition from Response Surface Methodology (KARSM), and GID3 [6] or generalized ID3, a decision-tree algorithm for fault diagnostics and decision making have been developed and used Gardener and Bieker [28] showed a substantial savings in the manufacture of semiconductors by applying decision-tree algorithms and neural networks to solve the yield problem in the wafer manufacture Sebzalli and Wang [29] applied principal component analysis and fuzzy c-means clustering to a refinery catalytic process to identify operational spaces and develop operational strategies for the manufacture of desired products and to minimize the loss of product during system changeover Four operational zones were discovered, with three for product grade and the fourth region giving high probability of producing off-specification product Lee and Park [30] used self-organizing maps and Last and Kandel [31] applied information fuzzy networks for quality checks and extracted useful rules from their model to check the quality of the products Kusiak [32]

[1] http //citeseer ist psu edu/cs

proposed a rule-structuring algorithm that can handle data from different sources to extract rules, which is very helpful in semiconductor manufacturing The algorithm formed relevant metastructures enhancing the utility of the extracted knowledge Dabbas and Chen [33] proposed the consolidation and integration of data from different semiconductor manufacturing sources into one database to generate different factory performance reports Their method can be further exploited to use data mining to extract information from these reports

Different data mining tools for improvement in integrated circuit manufacturing were presented in Ref [34] Another successful application of a sophisticated data mining algorithm was reported by Fountain et al [35] They used the Naive Bayes probabilistic model in their theoretic decision-making approach to optimize testing of dies (ICs in the wafer form) during a die-level functional test Their results showed substantial savings in testing costs and hence reduced overall costs compared with other testing policies, such as "exhaustive," "package all," and "Oracle"

An interesting area of research in manufacturing enterprises has been determination of optimal machining parameters to minimize machining errors such as tool wear, tool breakage, and tool deflection, which could result in slower production rates and increased costs Park and Kim [36] reviewed different techniques based on CAD systems, operational research, and computational intelligence to determine the optimal solutions to these errors and for online adaptive control using knowledge-based expert systems Other knowledge-based systems have also been proposed in Ref [37] for condition interpretation of tools and quality of the products

Performance and quality issues have also been considered while applying data mining techniques in manufacturing process related areas Gertosio and Dussauchoy [38] have used linear regression analysis to determine and establish the relationships between test parameters and the performance of truck engines Their simple methodology showed up to 25% saving in test process time A method to reduce the component testing time required before assembly was proposed by Yin et al [39] They applied genetic and rough-set algorithms on past test data to find the optimal test criteria to substantially reduce the overall testing time Another successful application of a regression model is presented in Ref [40] to predict the performance of the knurling process and the quality of the knurls Similar results were achieved by using both regression and neural networks

Efforts have also been made to develop models to study the entire factory or enterprise data altogether to discover the problem areas instantly affecting any subsequent processes Maki et al [41,42] developed an intelligent system in Hitachi for online data analysis using a data mining approach Their system used a rule induction algorithm to extract rules using an automated data mining engine and delivered the results using an intranet for easy access Adams [43] analyzed different software that can be used to mine a factory's data and compare the features of information sharing Shahbaz's [24] integrated data mining model is the next level of knowledge sharing, as once the data are mined the relevant data and data mining results can be shared within the factory and beyond at other sites, by using a neutral data format Shahbaz et al [44,45] used association rules for product design improvement and applied supervised association rules for controlling the product dimensions by controlling the process variables using supervised association rules [11,45] Their methodology can be used as an alternative and/or a support to the design of experiments methodology

Finally, the last two papers to be included and reported in this section are in the area of material properties Chen et al [46] applied data mining in hyperspace to identify material properties They used MasterMiner to build a hyperspace data mining model, which uses n principle factors or most relevant variables and built a mathematical model to find the solution equation in n dimensional space for a specific material property This technique is

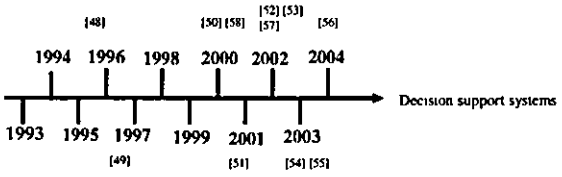


Fig 4 Decision support systems literature over time

useful in chemical and material industry where different variables affect one or more properties of the material or chemical reaction Interesting research has also been done by Mere et al [47] to determine the optimal mechanical properties of galvanized steel by using a combination of clustering and neural networks Clustering was used in the first instance and then neural networks were applied to the clusters to predict the mechanical properties of the steel

3.3 Decision Support Systems. Knowledge is the most valuable asset of an organization Decisions are made based on a combination of judgement and knowledge from various domains Decision support, knowledge management, and processing are interdependent activities in many organizations Data mining applications related to Decision Support Systems are shown in Fig 4 Ideally, all relevant knowledge should be available before making a decision The knowledge extracted from databases (prescriptive data mining) can be integrated with existing expert systems Grabot [48] used fuzzy logic to compliment the decision-support system to modify schedules Koonce et al [49] applied data mining to assist engineers in understanding the behavior of industrial data They developed a software tool called DBMine using Bacons algorithm, decision trees, and DB learn They applied the tool to find patterns in job shop scheduling sequences generated by a genetic algorithm [50] Caskey [51] developed a general environment for providing the right knowledge at the right time He used GAs and neural networks in identifying the structure of the data The knowledge extracted was in the form of "actual control applied → performance obtained" and the knowledge generated could be used to increase the accuracy of the system or validate the performance model Kusiak [52] applied data mining to support decision-making processes Different data-mining algorithms were used to generate rules for a manufacturing system A subset of these rules was then selected to produce a control signature of the manufacturing process The control signature is a set of feature values or ranges that lead toward an expected output Kusiak [53] used rough-set theory to determine the association between control parameters and the product quality in the form of decision rules and generated the control signature from those rules Lee and Park [54] presented an agent-based customer centric electronic commerce model in a make-to-order semiconductor manufacturing environment They used data mining for a decision support system by providing a set of recommendations reflecting domain knowledge Knowledge-based systems can be used to enhance the application range of simulation They offer the necessary knowledge required to make decisions in scheduling and rescheduling of manufacturing operations Symeonidis et al [55] applied data mining to make the ERP system more versatile and adaptive by integrating the knowledge extracted in companies' selling policies Huang [56] presented an agent-based system for knowledge management focused on the decision support of modular and collaborative product design and manufacture He used a neural network for the development of a decision support system

Bolloju et al [57] suggested integrating decision support systems and knowledge management processes across organizations using OLAP (On Line Analytical Processing) Based on this approach, a general framework was proposed for an enterprise decision support system using model marts and model warehouses for structured repositories of knowledge obtained through various

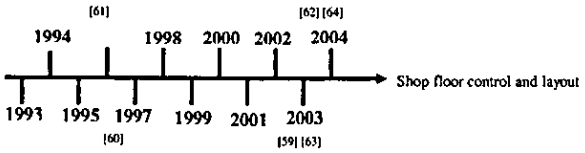


Fig 5 Shop floor control and layout literature over time

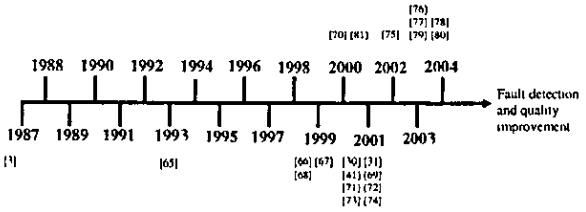


Fig 6 Fault detection and quality improvement literature over time

sources They assume that decision makers combine different types of data (e g , internal data and external data) and knowledge (both tacit and explicit knowledge) available in various forms in the organization

An interesting combination of neural networks and OLAP, called Neural On-Line Analytical Processing System (NOLAPS), was developed to enhance the decision support functionality of a network of enterprises NOLAPS used a neural network for extrapolating probable outcomes based on available patterns of events and OLAP for converting complex data into new information and knowledge The areas addressed are the selection of business partners, coordination in the distribution of production processes, and the prediction of production problems The adoption of NOLAPS in real industrial situations was also suggested [58]

3.4 Shop Floor Control and Layout. The shop floor control and layout problems are concerned with the efficient and effective utilization of resources, at the lowest level of control in manufacturing A vast amount of data is recorded during the operation of a shop floor, often to ensure that parts and production steps can be traced These data can also be used to optimize the process itself, since the knowledge generated from mining historical work-in-process data helps in characterizing process uncertainty and parameter estimation of the system concerned Data Mining literature related to these topics is shown in Fig 5

Chen [59] used association rules for cell-formation problems Associations among the machines are found from the process database, which leads to the identification of the occurrences of other machines with the occurrence of a machine in the cell This approach also clusters the parts and machines into families and cells simultaneously and hence requires minimal manual judgement Chao et al [60] presented an intelligent system to generate associative data for input in layout generation tools They used an expert system, object oriented database, and cluster analysis, which ensures data consistency and determines the strength of relationship between the two items under consideration

Knowledge generated from data mining can be used to analyze the effect of decisions made at any stage Belz and Mertens [61] used SIMULEX coupled with a knowledge based system to model the plant and evaluate the results of various rescheduling measures They used MANOVA for statistical analysis The collected data can be analyzed to identify the normal and abnormal patterns in it Kwak and Yin [62] presented a data-mining based production-control system for testing and rework in dynamic CIM Their system analyzes the present situation and suggests dispatching rules to be followed and also how data mining can be used to evaluate the effect of those decisions The knowledge generated can be used as the intelligence in a multi-agent system, Mitkas et al [63] presented a multi-agent system for concurrent engineering equipped with a data-driven inference engine The behavior and intelligence of each agent in the system is obtained by performing data mining on available application data and the respective knowledge domain Srinivas et al [64] presented a multi-agent based control architecture which uses data mining for decision support systems

3.5 Fault Detection and Quality Improvement. Fault diagnosis is an area that has seen some of the earliest applications of data mining, e g Ref [3], (see Fig 6) A common and intuitive approach to problem solving is to examine what has happened in the past to better understand the process, then predict and improve the future system performance Hence, the error rates in manufacturing are commonly used for knowledge acquisition to assist the quality control engineers Data mining can help in identifying the patterns that lead toward potential failure of manufacturing equipment This methodology helps in identifying not only the defective products but can also simultaneously determine the significant factors that influence the success or failure of the process The knowledge thus generated by searching large databases can be integrated with the existing knowledge-based systems to enhance process performance and product improvement

Data mining can be used to improve quality control, for example Apté et al [65] used computational techniques for quality control in manufacturing They deployed it in a disk-drive manufacturing line to reduce the number of expensive tests while meeting the performance criteria They applied rule induction, neural network, decision tree, and *k*-nearest neighbor in their experimentation Lee and Park [30] applied self-organizing maps to determine the optimal areas of inspection for a manufactured wafer This technique can save considerable time that is used in carrying out a 100% inspection of the semiconductor wafer

An important aspect of quality improvement is accurate fault diagnosis, and determining types of fault and failures Malkoff [3] introduced a methodology which uses temporal data in performing fault diagnosis in a subsystem of a Navy Ship propulsion system The methodology used patterns of the binary tree to generate the corrective actions Liao et al [66] presented fuzzy clustering based techniques for the detection of welding flaws They also presented a comparison between two fuzzy clustering methods, i e , fuzzy *k*-nearest neighbors and fuzzy c-means Liao et al [67,68] presented an integrated database and expert system for assisting the human analyst in identifying the failure mechanism of mechanical components Liao et al [69] discussed a multi-layer perceptron neural network to model radiographic welding data Last and Kandel [31] used information fuzzy networks to build a prediction model for quality checks and then used this model for the extraction of rules Shen et al [70] applied rough-set theory to diagnose more than one category of faults in a generic manner, since it was used to extract the rules leading to the failure These rules were used to distinguish the fault type or to inspect the dynamic characteristics of the machinery They demonstrated their approach for the identification of valve faults in a multi-cylinder diesel engine Lu [71] mined enterprise data to improve the quality of the product They decreased the dimension of the data and then applied different data mining tools for quality improvement

Maki and Teranishi [41] developed an automated data mining approach for data analysis in manufacturing and used it on an LCD production line Their system consisted of three main features First, it defined the data feeding and mining that were automated concurrently with the production process They used an induction method for mining and also determined its statistical significance Second, their system had the facility to store the generated rules in the intranet of the company and the third feature of their system was that it could also predict the temporal variance in the process Zhou et al [72] applied the C4 5 algorithm for drop test analysis of electronic goods Kusiak and

Kurasek [73] used data mining to solve the quality engineering problems (solder ball defects) in the manufacture of printed circuit boards (PCB) They applied rough set theory to determine the causes of defects which needed further investigation Oh et al [74] presented an intelligent control system using a data mining architecture for quality improvement in the process industry They used a neural network modeling method to establish the relationships between process and quality variables and identified the main causes of defects, which also provided optimized parameter adjustments Skormin et al [75] applied data mining for accurate assessment and forecasting of the probability of failure of hardware, such as avionics based on the historical data of environmental and operational conditions They developed a heuristic for the determination of informative subspace in low dimension and then used a decision tree to model the data Chen et al [76] generated association rules for defect detection in semiconductor manufacturing They determined the association between different machines and their combination with defects to determine the defective machine They used the Piatetsky-Shapiro formula to determine the statistical significance of the identified association Tseng et al [77,78] used rough set theory to resolve quality control problems in PCB manufacturing by identifying the features that produce solder ball defect and also determined the features that significantly affect the quality of the product Tseng et al [79] used rough-set theory on machining data to identify the relationships between the features of the machining process and surface roughness Shi et al [80] applied a neural network to model nonlinear cause and effect relationships and applied it in the chemical and PCB manufacturing process

Another interesting work is reported by Yuan et al [81] in determining the toxicity (Microtox) in the process effluents from a chemical plant using neural networks and principal component analysis Their software analyzer predicts the toxicity level and helps in developing strategies in process operations for toxicity reduction in the effluents

3.6 Data Mining in Maintenance. Preventive maintenance is of key importance in process and manufacturing engineering Databases containing the events of failure of the machines and the behavior of the relevant equipment at the time of the failure can be used in the design of the maintenance management systems Batanov et al [82] researched knowledge-based maintenance systems and developed a prototype system called EXPERT-MM, which works on historical failure data and provides suggestions for an appropriate preventive maintenance schedule A data-based design of optimal maintenance methods has also been proposed by Hsu and Kuo [83] They started working on this project at the beginning of the 1990s and they suggested that 100% of the inspection should start after the manufacture of a certain number (n) of parts and when the percentage of bad parts reaches a certain threshold value Preventive maintenance should then start to bring the process under control again When the process has been controlled and an additional n parts have been manufactured, the procedure can be repeated Maintenance operations and quality control are interrelated and quality control databases can therefore be used to design preventive maintenance plans Sylvain et al [84] used different data mining techniques, including decision trees, rough sets, regression, and neural networks to predict component failure based on the data collected from the sensors of an aircraft Their results also led to the design of preventive maintenance policies before the failure of any component Romanowski and Nagi [85] applied data mining in a maintenance domain to identify which subsystems were responsible for low equipment availability They recommended a preventive schedule and found that sensors and frequency response provide the most information about faults They used a decision tree to model the data

Although only a few reports of data mining have been identified in maintenance applications, this was one of the first areas of manufacturing to take advantage of data mining based solutions (see Fig 7)

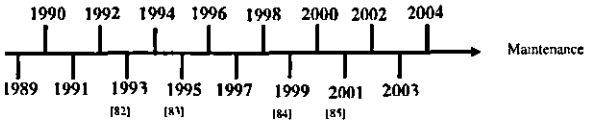


Fig 7 Maintenance literature over time

3.7 Customer Relationship Management. The marketing model has shifted from product-focused to customer-focused Customer Relationship Management (CRM) is concerned with increasing the value of interaction with customers and maximizing the profit In this competitive and global business environment, the application of data mining in CRM related to manufacturing industry has attracted research interest (see Fig 8)

CRM is as important as producing high quality and low cost products and is complementary to demand management which may be defined as a set of practices aimed at managing and coordinating a demand chain, starting from the end customer and working backwards to raw material and suppliers To collect appropriate information, customer demand data are collected and analyzed and then the product design features are changed to meet the customer's demands Similarly, in service industries, data from customers is the only source of knowledge about their satisfaction with the product Morita et al [86] applied data mining for customer segmentation to determine which customers are likely to shift from one cellular company to another They used a rule induction algorithm on the transformed data to build rules and then predict the potential moves of the customers Hui and Jha [87] used DBMiner to develop a decision-support system using a customer service database and they used neural networks and case based reasoning to mine the unstructured customer data to identify the machine faults Rygielski et al [88] discussed different data-mining techniques and provided an overview of customer relationship management They presented two case studies, one using neural networks and the other using Chi-Square Automation Interaction Detection (CHAID) to improve the business by targeting customer's data They reviewed both the models, comparing the simplicity of implementation of CHAID against the accuracy of neural networks Agard and Kusiak [89] applied data mining to customer response data for its utilization in the design of product families They used clustering for customer segmentation, i e to group the customers The requirements from the product were then analyzed using association rules for the design of the product Padmanabhan and Tuzhilin [90] described different ways in which optimization and data mining can help another for certain customer relationship management applications in e-commerce

4 Future Directions and Conclusions

This paper has surveyed numerous applications of data mining in manufacturing In recent years there has been a significant growth in the number of publications in some areas of manufacturing, such as fault detection, quality improvement, manufacturing systems, and engineering design In contrast, other areas such as customer relationship management and shop floor control have received comparatively less attention from the data mining community An exponential growth of data mining applications in the

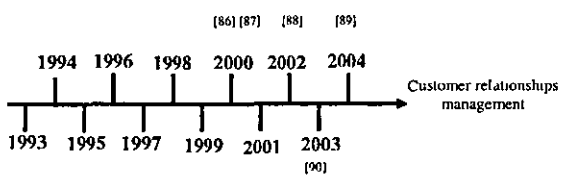


Fig 8 Customer relationships management literature over time

semiconductor industry has been observed The reasons for this may be that large volumes of data are generated during manufacture and that small improvements can have a significant impact in this industry No other sector of manufacturing industry reports such large increases of data mining applications This is rather surprising as other industries such as aerospace routinely collect huge quantities of data during product manufacture and hence are good potential environments for data mining studies

Many reported applications are related to the causes of malfunctioning of different types of manufacturing systems or processes and hence the discovered knowledge leads toward the better functioning of the manufacturing enterprise Developments in data mining are generally directed at the refinement of algorithms and their application in manufacturing, their integration with existing systems, standardization, the use of common methods and tools, and the definition of repeatable projects Recent trends indicate an increasing awareness as more and more people are using data mining for problem solving in manufacturing It is expected that future research will be directed at analyzing data related to design, shop floor control, scheduling, ERP, supply chain, and in developing a generic system where these can be integrated with existing knowledge based systems to enhance their capability

The research reviewed in this paper has mainly concentrated on applications of the algorithms The quality of the data and data preparation issues, particularly relating to manufacturing databases have not been discussed Major effort is needed in the data preparation process, as this is often simply based on practitioner's instinct and experience A more generic process for data cleaning is essential to enable the growth of data mining in manufacturing industry

The manufacturing data-mining research often does not consider the quality of the rules or knowledge discovered The knowledge generated is sometimes cumbersome and the relationships obtained are too complex to understand Future research effort is therefore also needed to enhance the expressiveness of the knowledge

The CRISP-DM methodology provides high level step-by-step instructions for applying data mining in engineering Further research is needed to develop generic guidelines for a variety of different data and types of problems, which are commonly faced by manufacturing engineering industry

References

[1] Spiegler I 2003 "Technology and Knowledge Bridging a 'Generating' Gap," Inf Manage , 40 pp 533–539

[2] Han, J and Kamber M , 2001, *Data Mining Concepts and Techniques*, Morgan Kaufmann New York, 550 pp

[3] Malkoff D B , 1987, "A Framwork for Real-Time Fault Detection and Diagnosis Using Temporal Data " Artif Intell Eng , 2(2), pp 97–111

[4] Ramamoorthy C V and Wah B W, 1989, Knowledge and Data Engineering ' IEEE Trans Knowl Data Eng , 1(1) pp 9–16

[5] Lee, M H , 1993 ' Knowledge Based Factory," Artif Intell Eng 8 pp 109–125

[6] Irani K B , Cheng J , Fayyad, U M , and Qian Z. 1993 "Applying Machine Learning to Semiconductor Manufacturing IEEE Expert, 8(1), pp 41–47

[7] Piatetsky-Shapiro, G , 1999 "The Data Mining Industry Coming of Age," IEEE Intell Syst , 14(6) pp 32–34

[8] Neaga, E I , and Harding J A , 2002, "A Review of Data Mining Techniques and Software Systems to Improve Business Performance in Extended Manufacturing Enterprises Int J Adv Manuf Syst (IJMAS), Spec Issue Decis Eng , 5(1) pp 3–19

[9] Neaga E I and Harding, J A , 2001, Data Mining Techniques for Supporting Manufacturing Enterprise Design," *International Conference on Industrial and Production Management* Quebec City Canada, pp 232–241

[10] SPSS, 1999, *CRISP DM 1 0 Step By Step Data Mining Guide*

[11] Shahbaz, M 2005 "Product and Manufacturing Process Improvement Using Data Mining " Ph D thesis, Wolfson School of Mechanical and Manufacturing Engineering Loughborough University Loughborough, Leicestershire 194 pp

[12] Neaga E. I , and Harding, J A , 2005, ' An Enterprise Modelling and Integration Framework Based on Knowledge Discovery and Data Mining," Int J Prod Res , 43(6), pp 1089–1108

[13] Neaga, E. I , 2003 ' Framework for Distributed Knowledge Discovery Systems Embedded in Extended Enterprise " Ph D thesis, Wolfson School of Mechanical and Manufacturing Engineering Loughborough University Loughborough UK. 192 pp

[14] Sim S K and Chan Y W, 1992, "A Knowledge Based Expert System for Rolling-Element Bearing Selection in Mechanical Engineering Design," Artif Intell Eng , 6(3), pp 125–135

[15] Kusiak, A , Kernstine K H Kern J A , McLaughlin, K A , and Tseng, T L , 2000, "Data Mining Medical and Engineering Case Studies " *Industrial Engineering Research Conference* Cleveland, OH, pp 1–7

[16] Ishino Y, and Jin, Y, 2001, ' Data Mining and Knowledge Acquisition in Engineering Design " *Data Mining for Design and Manufacturing Methods and Applications*, D Braha, ed Kluwer Academic, Dordrecht pp 145–160

[17] Romanowski C J , and Nagi, R , 2001, ' A Data Mining for Knowledge Acquisition in Engineering Design " *Data Mining for Design and Manufacture Methods and Applications* D Braha, ed , Kluwer Academic, Dordrecht pp 161–178

[18] Giess, M D Culley, S J , and Shepherd, A , 2002, Informing Design Using Data Mining Methods," *ASME DETC*, Montreal Canada, pp 98–106

[19] Giess, M D , and Culley S J , 2003, "Investigating Manufacturing Data for Use Within Design *ICED 03* Stockholm Sweden, pp 1408–1413

[20] Hamburg I , 2002, Improving Computer Supported Environment Friendly Product Development by Analysis of Data," *2nd European Conference on Intelligent Systems and Technologies*, Iasi, Romania, 6 pp

[21] Romanowski, C J , and Nagi R , 2005 "On Comparing Bills of Materials A Similarity/Distance Measure for Unordered Trees," IEEE Trans Syst. Man Cybern , Part A Syst Humans 35(2) pp 249–260

[22] Romanowski C J and Nagi R 2004 "A Data Mining Approach to Forming Generic Bills of Material in Support of Variant Design Activities," ASME J Comput Inf Sci Eng , 4(4), pp 316–328

[23] Kim, P, and Ding, Y, 2005, "Optimal Engineering System Design Guided by Data-Mining Methods " Technometrics 47(3) pp 336–348

[24] Shahbaz, M , and Harding J A 2003 ' An Integrated Data Mining Model for Manufacturing Enterprises " *International Conference on Manufacturing Research* Glasgow pp 539–545

[25] Kim, S H , and Lee C M , 1997 ' Non Linear Prediction of Manufacturing Systems Through Explicit and Implicit Data Mining," Comput Ind Eng , 33(3-4) pp 461–464

[26] Adachi, T , Talavage, J J , and Modie, C L , 1989 "A Rule Based Control Method for a Multi Loop Production System " Artif Intell Eng 4(3) pp 115–125

[27] Whitehall B L , Lu, S C Y, and Stepp, R E , 1990, "CAQ A Machine Learning Tool for Engineering," Artif Intell Eng , 5(4) pp 189–198

[28] Gardner M , and Bieker, J , 2000, Data Mining Solves Tough Semi Conductor Problems," *KDD 2000*, Boston, pp 376–383

[29] Sebzalli, Y M , and Wang, X Z 2001 "Knowledge Discovery From Process Operational Data Using PCA and Fuzzy Clustering,' Eng Applic Artif Intell , 14 pp 607–616

[30] Lee, J H , and Park, S C , 2001, "Data Mining for High Quality and Quick Response Manufacturing " *Data Mining for Design and Manufacturing Methods and Applications*, D Braha, ed , Kluwer Academic, pp 179–205

[31] Last, M , and Kandel, A , 2001, "Data Mining for Process and Quality Control in the Semiconductor Industry," *Data Mining for Design and Manufacturing Methods and Applications* D Braha, ed , Kluwer Academic, pp 207–234

[32] Kusiak, A , 2001, "Rough Set Theory A Data Mining Tool for Semiconductor Manufacturing," IEEE Trans Electron Packag Manuf , 24(1), pp 44–50

[33] Dabbas, R M , and Chen, H N , 2001, "Mining Semiconductor Manufacturing Data for Productivity Improvement—An Integrated Relational Database Approach," Comput Ind , 45 pp 29–44

[34] McDonald, C J , 1999, ' New Tools for Yield Improvement in Integrated Circuit Manufacturing Can They be Applied to Reliability?," Microelectron Reliab 39(6-7) pp 731–739

[35] Fountain T, Dietterich, T , and Sudyka B , 2003 "Data Mining for Manufacturing Control An Application in Optimizing IC Test," *Exploring Artificial Intelligence in the New Millennium* G Lakemeyer and B Nebel, eds , Morgan Kaufmann San Francisco, CA pp 381–400

[36] Park, K S , and Kim, S H , 1998 ' Artificial Intelligence Approaches to Determination of CNC Machining Parameters in Manufacturing A Review,' Artif Intell Eng 12 pp 127–134

[37] Liao, T W Chen J H , and Triantaphyllou, E , 1999, Data Mining Applications in Industrial Engineering A Perspective," *Proceedings of the 25th International Conference on Computers and Industrial Engineering* New Orleans LA, pp 265–276

[38] Gertosio, C , and Dussauchoy A 2004 "Knowledge Discovery From Industrial databases," J Intell Manuf , 15, pp 29–37

[39] Yin, Z L , Pheng, K L , and Cheong F S 2001 "Derivation of Decision Rules for the Evaluation of Product Performance Using Genetic Algorithms and Rough Set Theory " *Data Mining for Design and Manufacturing Methods and Applications*, D Braha, ed , Kluwer Academic, Dordrecht, pp 337–353

[40] Feng C X J , and Wang, X F, 2004, Data Mining Techniques Applied to Predictive Modelling of the Knurling Process," IIE Trans , 36 pp 253–263

[41] Maki, H , and Teranishi Y 2001 Development of Automated Data Mining System for Quality Control in Manufacturing " *Lecture Notes in Computer Science*, Springer-Verlag Berlin pp 93–100

[42] Maki, H , Maeda, A Morita T and Akimon H , 1999, Applying Data Mining to Data Analysis in Manufacturing," *International Conference on Advances in Production Management Systems* pp 324–331

[43] Adams, L , 2002, "Mining Factory Data " Quality Magazine, Software and Analysis, May, pp 5

[44] Shahbaz, M , Srinivas, Harding J A , and Turner M ' Product Design and

Manufacturing Process Improvement Using Association Rules" Proc Inst. Mech Eng Part B (in press)
[45] Shahbaz, M Srinivas and Harding, J A, 2004 "Knowledge Extraction from Manufacturing Process and Product Databases using Association Rules," *PDT Europe*, Stockholm Sweden pp 145–153
[46] Chen N Zhu D D and Wang, W, 2000, Intelligent Material Processing by Hyper Space Data Mining," Eng Applic Artif Intell, **13**, pp 527–532
[47] Mere, J B O, Marcos, A G Gonzalez, J A, and Rubio, V L, 2004 'Estimation of Mechanical Properties of Steel Strip in Hot Dip Galvanising Lines' Ironmaking Steelmaking **31**(1), pp 43–50
[48] Grabot, B, Blanc, J C and Binda, C, 1996, A Decision Support System for Production Activity Control," Decision Support Sys, **16**, pp 87–101
[49] Koonce D A, Fang C H, and Tsai, S C, 1997 "A Data Mining Tool for Learning From Manufacturing Systems" Comput Ind Eng, **33**(1-2), pp 27–30
[50] Koonce, D A, and Tsai S C, 2000, Using Data Mining to Find Patterns in Genetic Algorithm Solutions to a Job Shop Schedule Comput Ind Eng, **38**, pp 361–371
[51] Caskey, K R, 2001 "A Manufacturing Problem Solving Environment Combining Evaluation Search and Generalisation Methods, Comput Ind, **44**, pp 175–187
[52] Kusiak A, 2002, "Data Mining and Decision Making" *SPIE Conference on Data Mining and Knowledge Discovery Theory, Tools and Technology IV* Orlando FL, pp 155–165
[53] Kusiak A, 2002, 'A Data Mining Approach for Generation of Control Signatures," ASME J Manuf Sci Eng, **124**, pp 923–926
[54] Lee J H, and Park, S C, 2003 "Agent and Data Mining Based Decision Support System and its Adaptation to a New Customer Centric Electronics Commerce," Expert Sys Applic, **25**, pp 619–635
[55] Symeonidis, A L Kehagias, D D, and Mitkas P A, 2003, "Intelligent Policy Recommendations on Enterprise Resource Planning by the Use of Agent Technology and Data Mining Techniques Expert Sys Applic, **25**, pp 589–602
[56] Huang, C C, 2004 "A Multi-Agent Approach to Collaborative Design of Modular Products Concurr Eng Res Appl, **12**(1), pp 39–47
[57] Bolloju N, Khalifa, M, and Turban, E, 2002, Integrated Knowledge Management Into Enterprise Environment for the Next Generation Decision Support," Decision Support Sys **33**, pp 163–176
[58] Lau H C W, Chinb, K S, Punb K F, and Ninga, A, 2000, "Decision Supporting Functionality in a Virtual Enterprise Network," Expert Sys Applic, **19** pp 261–270
[59] Chen M C, 2003, "Configuration of Cellular Manufacturing Systems Using Association Rule Induction," Int J Prod Res, **41**(2) pp 381–395
[60] Chao K M Guenov, M, Hills, B Smith P Buxton I, and Tsai, C F, 1997, An Expert System to Generate Associativity Data for Layout Design" Artif Intell Eng, **11**, pp 191–196
[61] Belz, R, and Mertens P, 1996, Combining Knowledge Based Systems and Simulation to Solve Rescheduling Problems Decision Support Sys, **17**, pp 141–157
[62] Kwak, C, and Yih, Y, 2004, 'Data Mining Approach to Production Control in the Computer Integrated Testing Cell," IEEE Trans Rob Autom, **20**(1), pp 107–116
[63] Mitkas, P A, Symeonidis A L Kehagias, D, and Athanasiadis, I, 2003 'Application of Data Mining and Intelligent Agent Technologies to Concurrent Engineering" *International Conference on Concurrent Engineering Research and Applications* Madeira, Portugal, pp 11–18
[64] Srinivas, Harding J A and Shahbaz, M, 2004, 'Agent Oriented Planning Using Data Mined Knowledge," *10th International Conference on Concurrent Engineering Adaptive Engineering for Sustainable Value Creation*, Seville, Spain pp 301–307
[65] Apte, C, Weiss S and Grout, G, 1993, Predicting Defects in Disk Drive Manufacturing A Case Study in High Dimensional Classification,' *IEEE Annual Computer Science Conference on Artificial Intelligence in Application*, Los Alamitos, pp 212–218
[66] Liao T W Li D M, and Li, Y M, 1999, "Detection of Welding Flaws From Radiographic Images With Fuzzy Clustering Methods," Fuzzy Sets Syst, **108** pp 145–158
[67] Liao, T W, Khan, Z H and Mount, C R, 1999, "An Integrated Database and Expert System for Failure Mechanism Identification Part II—The System and Performance Testing" Eng Failure Anal **6** pp 407–421
[68] Liao, T W, Khan Z. H and Mount, C R, 1999, "An Integrated Database and Expert System for Failure Mechanism Identification Part I—Automated Knowledge Acquisition" Eng Failure Anal, **6** pp 387–406
[69] Liao T W, Wang, G, Triantaphyllou E. and Chang P C, 2001, 'A Data Mining Study of Weld Quality Models Constructed With MLP Neural Networks From Stratified Sample Data," *Industial Engineering Research Conference*, Dallas, TX, p 6
[70] Shen L Tay F E H, Qu, L. S, and Shen, Y, 2000 "Fault Diagnosis Using Rough Set Theory," Comput Ind **43** pp 61–72
[71] Lu, J C, 2001, Methodology of Mining Massive Data Sets for Improving Manufacturing Quality/Efficiency," *Data Mining for Design and Manufacturing Methods and Applications*, D Braha, ed, Kluwer Academic Dordrecht, pp 255–288
[72] Zhou, C, Nelson P C, Xiao, W, Tirpak, T M, and Lane, S A 2001 "An Intelligent Data Mining System for Drop Test Analysis of Electronic Products," IEEE Trans Electron Packag Manuf, **24**(3) pp 222–231
[73] Kusiak, A, and Kurasek, C, 2001, 'Data Mining of Printed Circuit Board" IEEE Trans Rob Autom, **17**(2), pp 191–196
[74] Oh, S, Han, J, and Cho, H, 2001 "Intelligent Process Control System for Quality Improvement by Data Mining in the Process Industry" *Data Mining for Design and Manufacturing Methods and Applications*, D Braha, ed, Kluwer Academic, Dordrecht, pp 289–310
[75] Skormin, V A, Gorodetski, V I, and PopYack, I J, 2002, "Data Mining Technology for Failure of Prognostic of Avionics," IEEE Trans Aerosp Electron Syst, **38**(2), pp 388–403
[76] Chen W C, Tseng S S, and Wang, C Y, 2004, "A Novel Manufacturing Defect Detection Method Using Data Mining Approach" Lecture Notes in Artificial Intelligence, **3029**, pp 77–86
[77] Tseng T L, JothiShankar, M C, and Wu, T, 2004 "Quality Control Problems in Printed Circuit Board Manufacturing—An Extended Rough Set Theory Approach, J Manuf Syst, **23**(1) pp 56–72
[78] Tseng, T L, JothiShankar, M C, Wu, T, Xing, G, and Jiang, F, 2004, "Applying Data Mining Approaches for Defect Diagnosis in Manufacturing Industry," *IERC 2004*, Houston TX p 7
[79] Tseng T L Leeper, T, Banda, C, Herren S M and Ford J 2004 "Quality Assurance in Machining Process Using Data Mining," *IERC 2004*, Houston, TX p 6
[80] Shi, X, and Boyd P S D 2004 "Applying Artificial Neural Network and Virtual Experimental Design to Quality Improvement of Two Industrial Processes" Int J Prod Res, **42**(1) pp 101–118
[81] Yuan B Wang X Z and Morris, T, 2000, "Software Analyser Design Using Data Mining Technology for Toxicity Prediction of Aqueous Effluents" Waste Manage **20** pp 677–686
[82] Batanov, D, Nagarur, N, and Nitikhumkasem P 1993, 'Expert—MM A Knowledge Based System for Maintenance Management," Artif Intell Eng, **8**, pp 283–291
[83] Hsu L F and Kuo, S, 1995, Design of Optimal Maintenance Policies Based on Online Sampling Plans Eur J Oper Res, **86**, pp 345–357
[84] Sylvain, L Fazel, F, and Stan, M, 1999, 'Data Mining to Predict Aircraft Component Replacement' IEEE Intell Syst, **14**(6), pp 59–65
[85] Romanowski C J and Nagi, R, 2001, Analysing Maintenance Data Using Data Mining Methods," *Data Mining for Design and Manufacture Methods and Applications* D Braha ed Kluwer Academic, Dordrecht, pp 235–254
[86] Morita, T, Sato, Y, Ayukawa E and Maeda, A, 2000, 'Customer Relationship Management Through Data Mining" *Informs Korms 2000* Seoul pp 1956–1963
[87] Hui S C, and Jha, G, 2000, Data Mining for Customer Service Support" Inf Manage, **38** pp 1–13
[88] Rygielski, C, Wang, J C, and Yen, D C, 2002 'Data Mining Techniques for Customer Relationship Management," Technol Soc, **24** pp 483–502
[89] Agard B and Kusiak A 2004 Data-Mining Based Methodology for the Design of Product Families," Int J Prod Res, **42**(15) pp 2955–2969
[90] Padmanabhan, B, and Tuzhilin A 2003, "On the Use of Optimization for Data Mining Theoretical Interactions and eCRM Opportunities Manage Sci, **49**(10), pp 1327–1343

APPENDIX II

Sample Data

ID	Var1	Var2	Var3	Var4	Var5	Var6	aat	abt	act	adt	aet	aft	aaw	abw	acw	adw	aah	abh	ach
1	4 05	16	387.84	505 41	360	468	7.56	7.50	7 48	7 46	7 52	7.57	7.58	7.45	7.49	7.56	7.56	7.53	7.54
2	4.22	13	397.73	504 31	264	619	7.57	7 51	7 43	7.45	7.51	7.54	7.59	7.43	7.45	7.58	7.58	7.54	7.56
3	2.54	19	394.22	509 5	185	286	7.55	7.52	7 45	7.78	7.48	7.55	7.55	7.44	7.46	7.54	7.53	7.51	7.57
4	4.01	14	404 59	512 92	238	581	7.53	7.48	7.44	7 49	7.49	7.56	7.56	7.48	7.49	7.55	7.52	7.50	7.55
5	4.06	5	404.94	513 7	180	257	7.54	7.44	7.45	7.43	7.49	7.52	7.54	7.45	7.46	7.59	7.54	7.52	7.52
6	4.04	17	405 56	513 29	258	605	7.54	7.54	7.42	7.44	7.48	7.58	7.56	7.46	7.48	7.53	7.55	7.48	7.59
7	3 89	5	401.77	512 09	236	561	7.57	7.52	7.43	7.45	7.47	7.59	7 57	7.43	7.44	7.54	7.56	7.49	7.56
8	4.13	12	407.5	514 37	205	248	7.58	7.51	7.47	7.43	7.53	7.55	7.55	7.45	7 46	7 51	7.58	7.52	7.55
9	3.96	18	405.34	514.23	181	255	7.51	7.50	7.45	7.44	7.52	7.56	7.53	7.47	7 48	7 53	7.59	7.51	7.54
10	4.04	13	406 05	514.12	253	603	7.53	7.51	7.46	7.45	7.49	7.54	7 54	7.46	7 47	7 58	7.54	7.48	7.49
11	4.1	13	407.94	514.8	189	239	7 57	7.53	7.45	7.46	7.47	7.56	7.57	7.48	7.49	7 54	7.51	7.50	7.56
12	4.08	15	406.22	514.35	182	239	7 54	7.54	7.43	7.43	7.48	7 57	7.58	7.43	7.44	7.56	7.52	7.51	7.58
13	4.12	5	409.31	515.69	208	514	7 55	7.51	7.44	7.48	7 49	7 55	7.51	7.45	7.46	7 57	7.55	7.52	7.54
14	2 99	5	397.22	504.09	137	98	7.56	7.50	7.48	7.47	7 52	7 53	7.53	7.44	7.45	7.55	7 54	7.48	7.55
15	3.11	8	398.45	505	141	119	7.52	7 52	7.45	7.45	7 50	7.54	7.57	7 45	7.46	7.52	7.54	7.44	7.59
16	3.15	10	396.53	504.44	132	110	7.58	7 48	7.46	7.43	7.51	7.54	7 54	7.42	7.43	7.59	7.52	7.54	7.53
17	3.2	9	396 55	504.1	110	100	7.59	7 49	7.43	7.47	7.53	7.57	7.55	7.43	7.44	7.56	7.53	7.52	7.54
18	3 2	11	396.13	504 01	139	149	7.55	7.52	7.45	7 49	7.54	7.58	7.56	7.47	7.43	7.55	7.54	7 51	7.51
19	6 21	14	397.14	504.34	117	131	7.56	7.51	7.47	7 45	7.51	7.51	7.52	7.45	7.41	7.54	7.58	7 50	7.53
20	3 3	6	397.33	503.99	120	87	7.54	7.48	7.46	7 42	7.48	7.53	7.58	7.46	7.40	7.49	7.59	7.51	7.58

- var1, var2, var3, var4, var5 and var6 are measures of process variables 1-6
- aat, abt, act, adt, aet, aft represents the value of dimensions measured at thickness section aa,ab,ac,ad,ae,af
- aaw, abw, acw, adw represents the value of dimensions measured at width section aa,ab,ac,ad
- aah, abh, ach represents the value of dimensions measured at height section aa,ab,ac