1999

A model to integrate Data Mining and On-line Analytical Processing: with application to Real Time Process Control

Rahul Singh

Virginia Commonwealth University
School of Business

This is to certify that the dissertation prepared by Rahul Singh entitled "A model to integrate Data Mining and On-Line Analytical Processing: with application to Real Time Process Control" has been approved by her committee as satisfactory completion of the dissertation requirement for the degree of Doctor of Philosophy.

[redacted]

Richard T. Redmond, Ph.D., Chair of Dissertation Committee, School of Business

[redacted]

Youngohc Yoon, Ph.D., Co-Chair of Dissertation Committee, School of Business

[redacted]

Richard J. Coppins, Ph.D., School of Business

[redacted]

George M. Kasper, Ph.D., School of Business

[redacted]

Elliott D. Minor, Ph.D., School of Business

[redacted]

Lorraine M. Parker, Ph.D., College of Humanities and Sciences

[redacted]

E.G. Miller, Ph.D., Dean, School of Business

[redacted]

Jack L. Haar, Ph.D., Dean, School of Graduate Studies

Nov. 3, 1999

Date

A model to integrate Data Mining and On-line Analytical Processing: with application to Real Time Process Control

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy at Virginia Commonwealth University

By

Rahul Singh
Bachelor of Engineering
Birla Institute of Technology

Director(s): **Dr. Richard T. Redmond**
Associate Professor, Information Systems

Dr. Youngohc Yoon
Associate Professor, Information Systems

Virginia Commonwealth University
Richmond, Virginia
June 1999

ACKNOWLEDGEMENTS

I wish to express special thanks to Dr. Redmond and Dr. Yoon, my research directors for their patience, guidance and support through the entire duration of my doctoral work. Their constant encouragement along with providing an independent work environment has helped me to grow personally and professionally

I would like to thank all my committee members for their guidance, patience and understanding in the development of this work. Special thanks to Dr. Coppins for his time and patience in helping me improve my work. I want to thank Dr. Kasper for his guidance and advice, which have been an invaluable in my research and academic career. I would like to thank Dr Minor and Dr. Parker for their guidance, support and encouragement during my doctoral work. I want to thank the faculty and staff for their constant support and friendship. I would also like to thank the students in the Information Systems department for their friendship.

My heartfelt thanks to my parents for their love and constant support throughout my education. I dedicate my thesis to them. I am very grateful to my wife, Kamlakshi, for her constant encouragement, assistance, patience and everlasting friendship. Without her I would not have achieved my academic endeavors.

TABLE OF CONTENTS

Page

List of Figures

List of Tables

ABSTRACT

A model to integrate Data Mining and On-line Analytical Processing: with application to Real Time Process Control

By Rahul Singh

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy at Virginia Commonwealth University. Virginia Commonwealth University, 1999.

Director(s): Dr. Richard T. Redmond, Associate Professor, Information Systems
Dr. Youngohc Yoon, Associate Professor, Information Systems

Since the widespread use of computers in business and industry, a lot of research has been done on the design of computer systems to support the decision making task. Decision support systems support decision makers in solving unstructured decision problems by providing tools to help understand and analyze decision problems to help make better decisions. Artificial intelligence is concerned with creating computer systems that perform tasks that would require intelligence if performed by humans. Much research has focused on using artificial intelligence to develop decision support systems to provide intelligent decision support.

Knowledge discovery from databases, centers around data mining algorithms to discover novel and potentially useful information contained in the large volumes of data that is ubiquitous in contemporary business organizations. Data mining deals with large

volumes of data and tries to develop multiple views that the decision maker can use to study this multi-dimensional data. On-line analytical processing (OLAP) provides a mechanism that supports multiple views of multi-dimensional data to facilitate efficient analysis. These two techniques together can provide a powerful mechanism for the analysis of large quantities of data to aid the task of making decisions.

This research develops a model for the real time process control of a large manufacturing process using an integrated approach of data mining and on-line analytical processing. Data mining is used to develop models of the process based on the large volumes of the process data. The purpose is to provide prediction and explanatory capability based on the models of the data and to allow for efficient generation of multiple views of the data so as to support analysis on multiple levels. Artificial neural networks provide a mechanism for predicting the behavior of non-linear systems, while decision trees provide a mechanism for the explanation of states of systems given a set of inputs and outputs. OLAP is used to generate multidimensional views of the data and support analysis based on models developed by data mining. The architecture and implementation of the model for real-time process control based on the integration of data mining and OLAP is presented in detail. The model is validated by comparing results obtained from the integrated system, OLAP-only and expert opinion. The system is validated using actual process data and the results of this verification are presented. A discussion of the results of the validation of the integrated system and some limitations of this research with discussion on possible future research directions is provided.

Chapter 1: Introduction

1.1 Introduction

Every modern organization uses business processes to provide goods or services to its customers. Depending on the nature of the market that the organization competes in, some processes have greater relevance to the final mix of goods and services that the organization delivers. Manufacturing processes are one of the most important processes for any firm involved in the manufacture of tangible goods and have a significant bearing on its competitive advantage. Quality of a product may be defined as "the total composite product and service characteristics of marketing, engineering, manufacture and maintenance through which the product and service in use will meet the expectation by the customer" (Feigenbaum, 1991). Considerable human and financial resources are involved in the control of the manufacturing process to ensure that its products are of "good" quality. Decision-making for quality control is a central activity in manufacturing organizations.

Numerous data gathering devices collect and from all parts of the manufacturing process and store the data in centralized or distributed databases. Though manufacturing

processes are usually continuous processes, data is typically collected at discrete time intervals depending on the sampling frequency of the data collection instruments. In the case of modern manufacturing and highly automated organizations, data may be collected every minute for hundreds of variables throughout the production process. This data is available to all levels of personnel in the organization that need information about the manufacturing process. Production data may be used for multiple purposes within the organization such as scheduling, sales, and purchasing. Quality control is a central issue in the context of manufacturing organizations. The production data serves as vital input to decisions made about product quality.

1.2 Decision-making and Decision Support

Nature of Decision Problems

Early applications of computers in business solved problems that were repetitive, requiring few simple arithmetic operations on large volumes of data and involved little algorithmic complexity. The entire process could be programmed for the computer to generate reports with no supervision or intervention required by humans. Most business decision problems are "ill-structured" which have no simple algorithmic solutions. These problems are interchangeably referred to in the literature as unprogrammed, semi-structured or ill-structured problems. Solution of ill-structured problems often requires the judgment of the decision maker as vital input into the decision-making process. The design of computer-based systems that support the decision maker in solving ill-structured problems is a challenging task.

Decision Support Systems and Expert Systems.

Decision support systems and expert systems are commonly used to provide computer support for decision-making in the business environment. Decision support systems provide support in investigating alternatives and their relationships with the corresponding outcomes. Knowledge-based expert systems provide support in solving unstructured problems where the knowledge of a domain expert is beneficial to the solution. A primary difference between expert systems and decision support systems is in the role of the user. Expert systems are designed to make the decision for the decision maker while decision support systems support the decision-maker by guiding the analysis of the relationships between alternatives and their associated outcomes. A goal of artificial intelligence research is to design systems that can improve their performance with experience. Expert systems formalize the current knowledge of domain experts and make it available for non-expert decision makers. There is no learning involved in expert systems.

Data Mining and On-line Analytical Processing

An alternative to using the knowledge of a domain expert is to use the large volumes of data collected by organizations as the source of knowledge about the problem domain. Knowledge discovery from databases uses data mining to find hidden relationships in data that can provide useful information about decision problems. Data mining is a collection of algorithms derived from artificial intelligence, mathematics, pattern recognition and statistics. Data mining techniques can be applied to develop models to

help decision-makers' understanding of the problem domain and help inform the decision process. These algorithms provide a means for classification and categorization of the data to extract the nascent relationships within the data and build descriptive and prescriptive models of the processes from the data. Knowledge extracted by using data mining reflects the experiences of the organization and represents previously -unknown information. The size of the data repositories of any modern organization is constantly growing and the active mining of process data can provide a means of evolution to the knowledge extracted at earlier times. Data mining can be performed on virtually any kind of data storage format, from simple flat files to the most complex relational databases and data warehouses. Hence the process is quite versatile.

Data stored in the repositories of any modern organization has multiple dimensions and any combination of these dimensions may be important to a decision maker faced with a given problem. In recent years, there has been rapid growth in the services required of organizational databases to support decision-making in the organization. Analyses of different problems require different views of the data and different levels of analytical support. The amount of data used to make decisions, the number of people responsible for making the decision, the extent of distribution of the data and the types of information that are available to make the decision are constantly increasing. The process of analysis can benefit from multi-dimensional databases that are organized in a manner that supports the analytical demands of the problem. On-line analytical processing (OLAP) provides fast and flexible access to large amounts of derived data whose inputs may be constantly changing (Thomsen, 1997). OLAP requires multidimensional database

technology to support the analysis of large amounts of data with a view to making business decisions.

Integration of Data Mining and On-line analytical processing for Intelligent Decision-making

Intelligent decision-making requires that the analysis of data be driven by knowledge of the business processes. The goal is a decision-making environment that provides accurate models of the decision problem and flexible mechanisms to examine the dimensions of the data. Data mining techniques provide accurate and sophisticated models of the process involved in the decision problem. These models are based on actual data from the business processes and reflect the nuances of the business process from which they are derived. Actively mining the data allows for dynamic models that capture emerging relationships in the data. Analysis based on these models is current in its depiction of the problem environment. OLAP methods allow the structuring of relevant data to facilitate analysis. It stands to reason that research in decision-making should investigate the integration of the two techniques to provide analytical views of the data based on intelligent models of the problem environment. Data mining algorithms can identify the data items that bear relevance to the goal of the decision problem and their relationships to characteristics of the problem domain. In developing models of the problem environment, these algorithms define a structure for the analysis of decision problems. OLAP techniques can take these structures and provide access to the data to facilitate analysis and decision-making in the domain. This research develops an integrated model

of data mining and OLAP to support intelligent decision-making in the context of a real-time process control application.

1.3 Process Control.

Nature of Process Control Problems

Information systems help gather and store raw production data and allow access to this data in multiple formats. Quality control problems occur when the quality of the final product is not within the established acceptable parameters defined for normal operation. In these situations, decisions need to be made regarding the identification of the problem, identification of its causes and selection of a requisite course of action to solve the problem. Decision-making for process control involves the following activities:

i) Accurate detection of errors in the production process
ii) Identification of the possible causes of errors
iii) Support for selecting a course of action to correct errors
iv) Working within the temporal bounds of the problem context.

If errors are identified in the manufacturing processes, they need to be corrected as soon as possible to avoid waste and consequent financial losses. There is a practical temporal bound to decision-making regarding course of action to take when errors occur in a production line. An ideal system would incorporate early warning mechanisms to warn

operators of imminent failures in the system so that action could be taken to pre-empt such situations.

Requirements of systems support to Process Control

Data from modern manufacturing environments has many complex relationships due to the many processes that raw materials are subjected to in creating the final product. This complexity is compounded by the fact that modern information systems allow multiple data points to be collected and stored at frequent intervals in manufacturing data repositories. To be effective, any set of models that improves the understanding of the decision problem must be sophisticated and realistic enough to take into account the complex relationships within the data. The models must provide accurate descriptions of the many aspects of the process that the decision maker is interested in. Manufacturing processes are dynamic processes affected by changes in environment, machinery components, raw materials and product characteristics. Models that attempt to explain the relationships in the process must be dynamic and adaptive. Information systems that incorporate these models must provide the user with analytical support to explore the alternatives and their respective outcomes. Systems that support decision-making for process control should have the following characteristics:

i) Detect errors in the process.

ii) Work within the temporal bounds of the problem.

iii) Provide early warning of imminent failures.

iv) Use sophisticated adaptive and dynamic models that can adjust the model parameters based on changes in the problem environment.

v) Provide analytical support for decision makers.

vi) Provide understandable presentation of results and outcomes.

Many advances have been made in this direction spanning the range from statistical process control to the application of artificial intelligence techniques to manufacturing process control. The following section introduces approaches to process control problems. They are discussed in more detail in subsequent chapters.

Process Control Techniques

Statistical process control is one of the most commonly used approaches to process control. Statistical process control examines pre-established measures of quality in the product and their association with critical measures of performance of the manufacturing process. Statistical methods inform the user of the extent of conformity of the process with the established measures of stability by examining measures of central tendency and deviations. Quality control personnel decide how to make changes to the process so that the product can conform to quality requirements. Statistical process control techniques are not capable of explaining the cause of non-adherence of process measures to established parameters. Statistical process control offers no analytical support to help decision makers understand the process, examine alternatives and choose corrective actions.

A commonly used improvement of statistical process control is multivariate statistical process control that takes into account the multidimensionality of the data. Multivariate approaches identify the major contributors to variations in the process. Using techniques such as factor analysis and principal component analysis, multivariate statistical process control allows for the reduction in the dimensionality of the process, and makes it easier to understand the variations in the data. Multivariate techniques do not offer any analytical support for decision-making and it is very difficult for users who are not trained in multivariate methods to understand the output of multivariate statistical process control.

Object-oriented methods provide an effective approach to modeling the manufacturing process and incorporating the relationships between the entities of the system. Simulation methods recognize that process control systems are event-triggered systems that model and explain the relationships in the process and use these models to predict future behavior. Simulation models are typically theory-based and may not reflect real operating conditions. Thus, many critical nuances of the implementation of the manufacturing process may not be incorporated in the model (Bennett, 1995). Both object oriented and simulation methods have considerable limitations in analyzing the massive volume of complex data inherent in manufacturing process data (Grega, 1996; Ham et. al., 1996). More effective methods are needed to analyze the large amounts of data from complex and continuous processes in order to determine the steps required to keep a process stable and to bring it back to stability when errors occur.

Artificial Intelligence Techniques in Process Control

Expert systems and neural networks are two techniques from the artificial intelligence arena that have been applied to provide support for process control. Expert systems can be used to build models of the system and provide excellent analytical support for the decision makers. Their strength lies in their ability to explain the alternatives and the decision choices to the user. Such models, however, are usually rule-based and do not capture all nuances of the system. Expert systems formalize the knowledge of domain experts and make this available to non-experts (Dhar, 1987). Expert Systems are not adaptive and changes in the problem environment render the system inaccurate. Expert systems, by themselves, do not make effective process control systems (Alexander, 1987).

Neural networks are very effective in developing models for non-linear systems that require the ability to handle noisy data. They are useful for manufacturing process data since they typically contains noisy and missing data due to intermittent failures of data collection devices. Neural networks can be used to provide effective process control with on-line, real-time data. The prediction capabilities of neural networks can be used to provide early warning of failures in the outputs of the system. Neural networks can be trained to build accurate, sophisticated, and dynamic models of the system. They are commonly used as embedded intelligent components for control loops of individual pieces of machinery and are rarely used for modeling the entire manufacturing process (Calabrese, 1991). They can provide little support to help the user understand the process

and fare poorly in providing analytical support and understandable representation of the system (Dagli, 1994).

1.4 Summary of effectiveness of Process Control Techniques

Table 1.1 summarizes the effectiveness of the process control techniques discussed earlier with respect to the dimensions and requirements of the process control problem. None of these techniques address all aspects of the process control problem and more research is needed to provide a method to address the issues of the process control problem more effectively.

	Effective Process Control	On-Line System	Early Warning	Accurate models	Adaptive models	Analytical Support	Simple User Interface
Statistical Process Control	No	Yes	No	No	No	None	Yes
Multivariate Statistical Process Control	Yes	Possible	No	Yes	No	No	No
Object Oriented Methods	No	No	Yes	No	No	No	Yes
Simulation Methods	No	Possible	No	No	No	Yes	No
Expert Systems	No	No	No	Possible	No	Yes	Yes
Artificial Neural Networks	Yes	Yes	Yes	Yes	Yes	No	No

Table 1.1 Comparison of Process Control Methods
Columns represent dimensions of the process control problem and rows show current techniques.

1.5 Purpose of Research

This research develops a model that integrates data mining and OLAP technologies to support intelligent decision-making for real-time process control. Data mining is used to discover knowledge from the large volumes of data, which can be used as information in making intelligent decisions about the environment. An evolutionary approach is suggested in which the models are constantly reviewed as new data is gathered. This data is organized and presented for decision-making using OLAP to allow multidimensional views of the data. This integrated approach can be used to analyze incoming real-time data to locate and explain possible error conditions. As an improvement on existing approaches, the integrated approach offers explanatory and predictive capabilities based on accurate and adaptive models of the process and provides early warning of imminent failures.

The proposed solution relies on the integration of data mining and OLAP to build accurate and dynamic models of the process and provide analytical views of the data that support decision-making in this environment. In manufacturing environments, data mining can unearth novel patterns useful to predict future trends and behaviors of systems and enable proactive and knowledge-driven decision-making. Data from production processes is multi-dimensional. This data is collected from the multiple processes of the system and has information about multiple aspects of the production system. The data has a temporal component since it is collected at regular time intervals.

A large continuous manufacturing process, which is typical of many chemical process industries and other heavily automated manufacturing environments, is considered as the problem context. Such environments usually have enormous operational data logs that contain data collected from various parts of the manufacturing process. This data is usually collected at regular and frequent time intervals and stored in the production data repository. Data about the everyday operations contains a wealth of information about the numerous processes of the production system. The raw data itself, however, does not generate any direct benefits. The data needs to be analyzed to develop descriptive models to understand, explain and predict imminent errors in the manufacturing process. Such models can provide insight and direction to decision-making activity in the problem context.

1.6 Organization of Dissertation

This chapter introduces the concepts from the literature that are relevant to this research. The real-time process control problem is introduced and a description of the problem environment and its requirements, including the methodologies currently used for this problem, are presented. The next two chapters serve as a review of the literature pertaining to decision-making in organizations and process control. Chapter two presents an evolutionary view of decision-making in organizations and describes the nature of the decision-making problem and the different ways in which computerized support has facilitated the decision-making task in business organizations. Chapter three addresses the decision-making requirements in the process control environment.

Chapters four and five present the integrated model and discuss the implementation of its prototype. Chapter four presents the design goals for the proposed system and introduces the model for real-time process control based on the integration of data mining and OLAP. Chapter five presents the components of the integrated system and discusses their implementation in the prototype of the integrated system. Chapter six describes the model validation approach. Chapter seven presents the results of the model validation. The system is validated using actual process data and the results of this verification are presented in chapter seven. Chapter eight concludes the dissertation with a discussion of the results of the verification of the integrated system and presents some limitations of this research with discussion on possible future research directions.

Chapter 2: Decision-making in Organizations

2.1 Decision-making

Administration of an organization involves the determination of appropriate courses of action to help the organization achieve its objectives (Simon, 1976). Decision-making is the act of selecting a requisite course of action among a number of alternatives so as to achieve certain objectives. Theories of organization provide multiple perspectives on what the task of decision-making in organizations entails. March and Simon (1958) compare the rationality of decision makers in organizations as embodied by the classical and statistical decision theories with the concept of "administrative man" or "rational man". They note that the traditional theories of organization, such as those postulated by Fredrick Taylor (Theory of Scientific Management) make certain assumptions about the problem domain, which may not hold true in the context of organizational decision-making. This scientific-management view of decision-making assumes that all possible alternatives are given, i.e., all possible courses of action are completely known at the time that the decision is made. The decision maker has a predefined utility function, or a system of preferences, that can be used to order the outcomes. Such problems typically conform to an algorithmic solution and have known models for their analysis.

Operational control functions, suggested by Anthony (1965), are often concerned with this category of problems.

There are several limitations to this algorithmic model of solving business decision problems, as pointed out by Simon, Cyert and Trow (1956):

i) "It is questionable whether the problem is a given for the decision maker, whether all alternatives and their exact relationships with associated outcomes are known.
ii) It is arguable that relationships between objectives and alternatives may change, and that all alternatives may not be available over time.
iii) It is certainly debatable that there may be more than one objective to a decision and that the decision to satisfy one of the goals may adversely affect the outcome with respect to another objective."

March and Simon (1958) postulate three sets of theories regarding alternatives available to a decision maker and their associated outcome:

i) Certainty of outcomes,
ii) Risk involved with obtaining outcomes, and
iii) Uncertainty regarding outcomes.

Certainty of outcomes implies that the decision maker has "complete and accurate" (March and Simon, 1958) information on the consequence of every alternative. Risk

implies that a probability distribution of the consequences associated with each outcome exists. In such circumstances, the decision maker would choose the alternative that minimizes expected risk and yet achieves the objectives of the decision-making process. Uncertainty theories assume that the consequences of each alternative belong to a subset of all the known consequences and that no information on how the consequences are associated with the outcomes is available. Critical to this concept is the lack of knowledge of the association between alternatives and outcomes. The latter two categories involve stochastic models or uncertainty and are referred to as unprogrammed decisions (Simon, 1957). There are no algorithmic solutions for these problems and their solution often involves judgement on the part of the decision maker (Scott Morton, 1971). Scott Morton also categorizes business decision problems into unstructured, ill-structured and structured problems based on the extent to which these problems fit into known models of decision problems.

2.2 Characteristics of Business Decision Problems

The American Heritage College Dictionary defines a model as:

"A schematic description of a system, theory, or phenomenon that accounts for its properties and may be used for further study of its characteristics. Such a work or construction used in testing or perfecting a final product".

In other words, a model is a concise representation of the problem domain and presents an organized view of the aspects of the problem that the problem solver is interested in. It

represents a simplification of the problem to the extent that it facilitates the understanding of the problem environment and its solution. The extent to which a problem can be reduced into known models is the extent to which their solution can be simplified. Structured problems are those for which there exist known, well-defined, models or algorithmic solutions. Ill-structured problems present a challenge in that there may not be a complete fit of the problem characteristics into known models of solution. Decision problems typically faced by managers are ill-structured since there are no known models for them and their solution often requires judgment on the part of the decision maker.

2.3 Decision Support Systems.

Decision-making is the primary function of administrators and managers in an organization. Decision support systems have been developed for this set of users to aid the decision-making process. Keen and Scott Morton (1978) define a decision support system as "a coherent system of computer based technology, including hardware, software, and supporting documentation, used by managers as an aid to their decision-making in semi-structured tasks". These systems help the decision maker examine more alternatives, evaluate the complex relationships between the alternatives and their associated outcomes while satisfying the objectives of the decision-making process. It is the goal of these systems to serve as a support system that helps the effectiveness of the decision-making process, and not to make the decisions for the decision maker. These systems are geared towards problems where there is a sufficient amount of structure for analytical aids to be helpful, but the nature of the problems makes the judgment of managers critical to the decision-making process (Keen and Scott Morton, 1978).

Typically decision support systems consist of a data management system, a model management system and a user interface (Olson and Courtney, 1992; Turban and Aronson, 1998). The data management system consists of the data that is required for analyzing the decision problem and some form of database management system that allows for easy storage and retrieval of the data. The content of the database varies based on the type of problem and may range from the entire corporate database to very domain-specific data. The model management system is responsible for analyzing data and the representation of the data in the context of the problem so that the problem can be easily understood and analyzed by the decision maker. Decision support systems are typically constructed with apriori knowledge of the decision problem. Hence, the models used in the model base are chosen based on the kind of analysis suitable for the decision problem for which the system is designed. For example, a multiple criteria decision support system may use the analytical hierarchy process as the model base and retrieve and store data to support this form of analysis. Such a system would not be suitable to perform any kind of analysis that does not conform to the analysis methodology of the analytical hierarchy process. Decision support systems provide interactive support to the decision maker through the user interface.

2.4 Classification of Decision Support Systems

Alter (1980) developed a taxonomy to classify decision support systems based on the degree to which the system's outputs can directly determine the appropriate decision for the problem under consideration. Decision support systems are used to retrieve

information, provide a mechanism for data analysis, estimate the consequences of proposed decisions (what-if or sensitivity analysis), propose solutions and even make decisions for the user. Based on the type of operation that they perform, Alter (1980) classifies decision support systems into seven "reasonably distinct" types:

i) **File Drawer Systems:** Systems that provide data and information retrieval.

ii) **Data Analysis Systems:** Systems that allow the manipulation of data in a manner that is either specific to the nature of the problem or based on general operators.

iii) **Analysis Information Systems:** Systems that provide access to multiple databases and models, thereby facilitating analysis of information.

iv) **Accounting Models:** Systems that provide what-if analyses by revealing the consequences of planned actions based on accounting definitions (cost-benefit, what-if analysis).

v) **Representation Models:** Systems that estimate the consequences based on models that are partially definitional. The difference between this type of model and the accounting models is that these systems may evaluate the consequences of actions based on associations other than those defined by accounting models or by cost-benefit ratio analysis.

vi) **Optimization Models:** Systems that provide guidelines for action by generating the optimal solution consistent with a series of constraints. The nature of the problems that these systems solve requires that an optimization model can explain the problems and some form of associated cost is known for each alternative. These models are largely derived from management science and operations research literature.

vii) **Suggestion Models:** Systems that provide the mechanical work that leads to a specific suggested decision for a fairly structured task. Again, the nature of the problem and its ability to be solved using a fairly structured task is critical for the success of this form of decision support systems.

An analysis of the classification of decision support systems proposed by Alter reveals that the classifications differ along the dimension of model support required for the decision-making activity. For example, file drawer systems are data oriented systems that require very little analytical or model based support while representation model systems are model-oriented systems that rely heavily on the analytical support provided by the underlying analytical models in the system. Others have classified decision support systems based on different aspects. Scott Morton (1971) classifies decision support systems as structured, ill-structured and unstructured based on the problem type that they are intended to solve. Decision support systems have also been classified based on the level of the management control function that they are intended to support, such as strategic planning, management control, and operational or task control (Anthony, 1956;

Anthony, Dearden, Bedford, 1984). In summary, there are many different dimensions along which decision support systems have been classified. For example:

i) The degree of structure in the problems that the decision support system tries to solve (Scott Morton – structured, unstructured and ill-structured problems).

ii) The extent to which the problems can be algorithmically reduced to simpler problems (Simon's programmable and non-programmable decisions).

iii) The level of organizational function that the decision problem is intended to support (Anthony's strategic planning, management control, and operational control).

iv) The extent to which the decision-making task requires the support of data or models (Alter's data oriented and model oriented decision support systems).

An additional dimension upon which decision support systems can be classified is the extent to which the decision-maker makes the decision. A distinguishing feature of suggestive systems (Alter, 1980), from the other types of systems discussed in Alter's classification scheme, is that here the decision is made by the system instead of by the decision maker. Suggestive systems represent a class of systems that can recognize the problem, obtain data required to analyze the problem, search for viable alternatives, analyze the appropriateness of the alternatives and select the most suitable alternative to

meet the objective of the decision problem. Alter's work is based on case studies of existing systems and the suggestive systems that he describes are used for very structured tasks that can be completely automated, a notion similar to that of Simon's "programmable decisions". Scott Morton (1971) speculates on the idea of using artificial intelligence to support human decision-making by the system taking an active role in the decision-making process as the expert, and offering meaningful suggestions about the alternatives and their associated outcomes to aid the decision maker. Computer systems can record and store the entire decision sequence taken by an expert. Over time, and with the experience gained over a number of cases, these records of the expert decision sequence can provide useful precedence and direction to the decision maker in the decision-making process. Expert systems are developed using this philosophy of capturing and formalizing the expertise of a domain expert and making it available to non-experts to solve specific problems that require the expertise of a domain expert. Such problems are typical and recurring in their nature so as to justify the expense of developing a system centered on the solution of particular decision problems.

2.5 Artificial Intelligence

Artificial intelligence (AI) is concerned with the study of machines that can perform tasks that are thought to require human intelligence. Earlier definitions of AI focused on its being a field concerned with creating systems that think like humans. Bellman (1978) defines AI as the automation of "activities that we associate with human thinking, activities such as decision-making, problem solving, learning...". Others focus on the aspect of AI concerned with creating systems that can think and act rationally. Winston

(1992) defines AI as "the study of the computations that make it possible to perceive, reason and act". Schlakoff (1990) gives another definition of AI: "a field of study that seeks to explain and emulate intelligent behavior in terms of computational processes". Albus (Albus, 91) defines intelligence as "the ability of a system to act appropriately in an uncertain environment, where appropriate action is that which increases the probability of success, and success is the achievement of behavioral sub-goals that support the system's ultimate goal." The goals and success that are key to this definition are defined outside of the system by the designers of the systems.

Artificial intelligence inherits from a number of fields of study including philosophy, mathematics, psychology, computer engineering and linguistics. The first work in artificial intelligence was by McCulloch and Pitts in 1943, on modeling any computational activity as a network of neurons. They proposed a model of artificial neurons that could be either on or off, with a switch to on occurring as a response to stimulation by a sufficient number of neighboring neurons (Russell and Norvig, 1995). AI is concerned with symbolic representation and manipulation and theorem proving as demonstrated by Newell and Simon's research on the Logic Theorist and the General Problem Solver. Artificial intelligence has been used in a number of application areas such as robotics, intelligent manufacturing, marketing, banking and finance. Feigenbaum et. al. (Feigenbaum, McCorduck and Nii, 1988), list several applications of expert systems in such diverse areas as agriculture, communications, computers, construction, geology and medicine.

2.6 Using Artificial Intelligence for Decision Support

2.6.1 Expert Systems

Knowledge-based expert systems, or expert systems as they are commonly known, are computer systems that can perform the role of a domain expert in the area of a problem that is being investigated by the decision maker. One definition of an expert system is computer software that performs a highly specialized task that would normally require human expertise. The expertise of the human in terms of either knowledge of the task at hand or knowledge of the problem area is incorporated into the system (Murray and Tanniru, 1987). Expert systems provide a means of formalizing a lot of mostly experiential and subjective knowledge that may have been heretofore unexpressed and unrecorded (Dhar, 1987). This creates a formal body of knowledge that the organization can draw upon in solving problems.

The concept of knowledge-assisted decision-making is not new. Scott Morton (Scott Morton, 1971) speculated about the use of artificial intelligence in designing computer systems that could become active participants in the decision-making process by making useful suggestions to the decision maker. Expert systems offer "the possibility of a major contribution" (Scott Morton, 1971) to the decision maker. Such systems could record the process that domain experts follow as they are solving problems in their area of expertise. This knowledge could grow, over time and over multiple instances of problems, into a systematic body of knowledge about the key decisions that an organization makes. A key feature of such systems is the institutionalization of this domain specific expertise so that

movements of people in the organization do not affect the knowledge. Newell and Simon's General-Purpose Problem Solver (Newell and Simon, 1973) is another example of an early system that bears resemblance to the expert systems of today. A major contribution to the development of expert systems is the development of theories describing how to represent the knowledge of domain experts in a form that can be used by these systems. The fundamental problem that Artificial Intelligence tries to solve is how to represent large volumes of knowledge so that decision makers can efficiently use the knowledge in their decision-making tasks (Goldstein and Papert, 1977).

Generically, an expert system contains a knowledge acquisition system. This system incorporates the methods used to acquire knowledge from the domain expert and, with the help of the knowledge engineer, formalize it into a knowledge base that can be used to store the knowledge. Knowledge typically consists of domain specific data and rules used to solve specific problem in the problem area. This domain specific knowledge is stored in the knowledge base in a format that can be drawn upon by the inference engine to help the manager solve business problems. The inference engine incorporates analytical models that are used to provide the decision maker with analytical support. It interacts with the knowledge base to provide the user with the information contained in the knowledge base. The inference engine can also utilize the explanation subsystem to call upon the knowledge base to explain the rationality behind the choices made by the system. The components of a model for a generic expert system are shown in Figure 2.1.

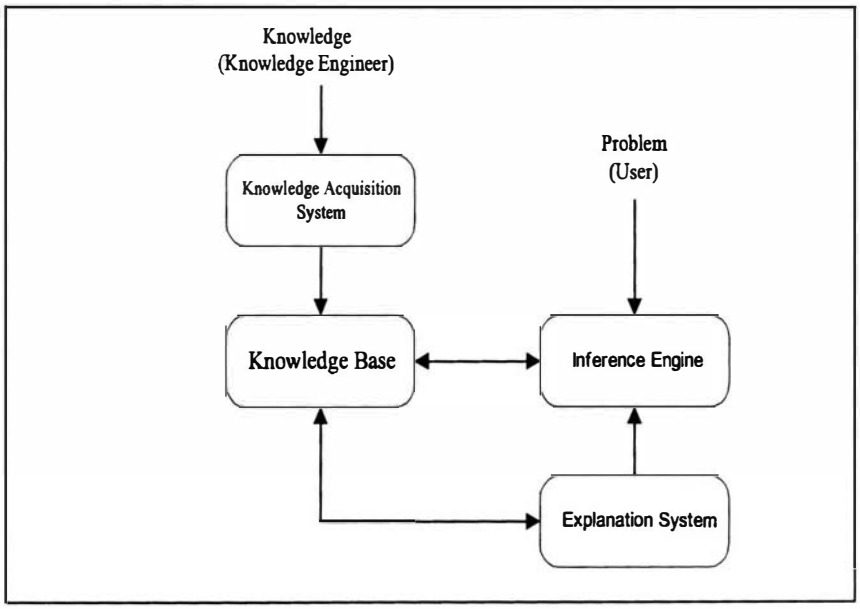

Figure 2.1 A General Expert System
(Murray and Tanniru, 1987)

The goal of an expert system is to capture the knowledge and expertise of a domain expert in solving highly domain specific problems. This expertise is incorporated into the expert system and can be used by non-experts to make decisions in similar problem domains within the organization. Expert Systems have limitations on the range of problems that they can solve. Problems usually involve ambiguous and incomplete data and an expert must be able to judge the reliability of the facts to clarify the problem and evaluate competing conceptualizations of a problem (Dhar, 1987). The scope of application of expert systems is limited to the specific problem domain for which they are developed. The methodology used by experts in solving the problem may be extended to similar problems with which organization is routinely faced. As the problem domain changes, attempts to add new knowledge to the system may affect the system in unforeseen ways. Traditionally, expert systems are developed to solve a very narrow range of well-structured problems that require the expertise of a domain expert. There is no learning involved in expert systems; if there are changes in the problem domain, then a new expert system needs to be built. This is a prominent drawback of expert systems.

Researchers have speculated on the idea of using AI-based components in decision support systems and it is noted that many of the advancements in DSS design has come from the field of artificial intelligence (Goul, Henderson and Tonge, 1992). There is debate in the AI community on whether its purpose is to create machines that act rationally thereby replacing human actions, or to create machines that think rationally and can support human cognitive activity. These opposing points of view are reflected in the

debate on the role of AI in the DSS area. Expert systems represent one end of the spectrum of decision support systems that aim to replace the human decision maker with an artificial one - the system. Another role for AI is in contributions to the design of decision support systems that have intelligent components to support the decision maker, as in traditional decision support systems. Goul, Henderson and Tonge (1992) propose "Artificial Intelligence can broaden DSS research beyond its original focus on supporting rather than replacing human decision-making by selectively incorporating machine based expertise in order to deliver the potential of DSS in the knowledge era." This proposition calls for synergy between research in both fields to find aspects of artificial intelligence that can help decision support and conversely find areas of decision support systems that can benefit from machine intelligence.

Earlier sections of this chapter have established the agreement in the decision support systems literature that decision support systems design can benefit from interaction with artificial intelligence research to provide knowledge based components for decision support systems. Traditionally, expert systems have provided the means of providing knowledge based support to the decision-making task. These systems serve the purpose of formalizing, documenting and institutionalizing the domain specific knowledge of an expert so that non-domain expert decision makers in the organization can utilize the domain specific knowledge in the decision-making task. It has also been pointed out that expert systems are very domain specific and it is not easy to expand their application scope. Such an endeavor invariably entails the development of a new system. This shortcoming of expert systems has received some attention in the literature and research

has been done on designing systems that expand their application domain. The ideas of evolutionary systems and self-evolving systems, and research in these directions, bear evidence of this fact (Liang and Jones, 1987; Alavi and Henderson, 1981; Hurst, 1983).

2.6.2 Data Mining and Knowledge Discovery in Databases

Another viable source of knowledge about the organization and its problem domain is the large volume of data that are typically stored in data repositories of organizations. Knowledge about the organization is contained in the many relationships hidden in organizational data, and that these patterns can provide useful insight into the functionality of the business. This approach is known as knowledge discovery from databases. The process of knowledge discovery from databases is defined as (Fayyad, 1996):

> *"the non-trivial process of identifying valid, novel, potentially useful and ultimately understandable patterns in data"*

Fayyad et. al. (1996), define the process of data mining in the context of the knowledge discovery process as:

> *"the process by which patterns are extracted and enumerated from the data."*

An analysis of these definitions shows that data mining is the central theme to the process of knowledge discovery in databases. These terms are often used interchangeably in the

literature. The key aspects of this definition are that the process discovers knowledge, in the form of patterns, from the existing data. These patterns should be understandable and potentially useful to the organization so that decision makers are able to understand the knowledge and use it for decision-making. Knowledge discovery from databases involves techniques for acquiring knowledge from organizational data that can eventually be used to support the decision-making activity in organizations. This is a departure from the technique used by expert systems that use the domain expert as the source of knowledge and attempt to capture and institutionalize this knowledge. Knowledge discovery from databases presents an alternative technique in that it uses corporate data as the source of knowledge and attempts to extract rules, patterns and associations that can be useful in making decisions in organizations. Contemporary organizations continuously collect data about their operations and various aspects of their business environment. This suggests that the nature of the knowledge gathered from data can be evolutionary, in that it can grow continuously to incorporate changes in the business and its environment. New knowledge can emerge and old knowledge can become redundant, or even irrelevant, as an organization progresses and reacts to changing situations.

2.7 Data Mining

Data mining is the technique by which new and meaningful patterns are discovered from data. Data mining uses models and algorithms from a number of related fields of study, such as statistics, machine learning and pattern recognition to:

a) Achieve a better understanding of large volumes of data,

b) Develop means of classifying the data, and

c) Discover patterns and associations in the data.

Fayyad et. al., (1996) take a reductionist point of view in defining three major components which comprise most data mining algorithms:

a) **Model:** contains parameters that are to be determined from the data. The function of models and their representational forms are factors that need to be decided based on the application domain and the type of data, before the data is mined.

b) **Preference criteria:** form the basis for preference of one model over another.

c) **Search algorithm:** algorithm for finding particular models and parameters given the data, a family of models and preference criteria.

The data mining algorithm is an instantiation of the model and the preference search algorithm components (Fayyad, et. al., 1996). The literature classifies data mining into two major forms (Simoudis, 1996; Fayyad et. al., 1996):

i) **Verification-driven data mining:** This form of data mining presumes that users already have postulated a hypothesis that they want to verify from the data. The

data mining algorithms are selected to verify, or nullify, the a priori hypothesis. Statistical techniques are often used for this form of data mining.

ii) **Discovery-driven data mining:** New rules and associations are extracted from the data in this form of data mining in which hypotheses are not postulated a priori. Discovery-driven data mining relies heavily on artificial intelligence techniques such as symbolic processing, association discovery and supervised induction.

There are many attempts in the literature to classify the operations performed by data mining algorithms to understand the functionality of the data mining task. Some researchers have developed classifications based on the most common data mining techniques, such as artificial neural networks, decision trees, genetic algorithms and rule induction methods.

Some common data mining operation are listed below:

1. **Classification:** Classification refers to the rapid discovery of categories that observational data can be divided into. Classification looks at the difference between dimensions of an observation and categorizes the observation based on that dimension. Some common techniques used for classification of data include clustering analysis, nearest neighbor analysis and factorial analysis.

2. **Link Analysis:** Discovery of associations in data that may lead to the explanation of causality and reveal previously unknown associations between variables in a data set. It is common to find variables that co-vary, particularly in large data sets. Such associations may provide useful insight and provide a powerful tool for the reduction of the dimensionality in data.

3. **Regression Techniques:** Regression attempts to establish relationships between variables in the data set so that associations between the rules can be inferred. This is a relatively simple, yet powerful, technique for the induction of rules and associations in the data set.

4. **Visualization Techniques:** These techniques contend that people who are familiar with the application area can identify patterns in the data if they have powerful tools available that allow them to view the data in multiple ways. This class of techniques provides a way to form hypotheses about relationships in the data that can be further verified.

It seems intuitive that some techniques would be better suited for different objectives of the data mining task, and for data with different characteristics, than others. Evidence of such a mapping between data mining techniques and the objective of the data mining task is not found in the literature. Case studies in the literature report on the success or failure of the data mining undertaken at certain companies. Perhaps, as the field matures, and more case studies become available, this type of research could provide valuable

guidelines for practitioners and researchers who want to design the data mining process. This type of research could be in the flavor of Alter's study of companies that use decision support systems to aid the decision-making task in their organizations.

The Institute of Electrical and Electronics Engineers (IEEE) devoted an entire issue of its journal that focuses on expert systems development, *IEEE Expert*, to data mining in October, 1996, as did the Association for Computing Machinery in its journal, *Communications of the ACM*, in November, 1996. This attests to the fact that data mining is an emerging research area that brings research in artificial intelligence and machine learning to business applications.

2.8 On-line Analytical Processing

On-line Analytical Processing (OLAP) is a class of technologies that provides multidimensional views of data and is supported by multidimensional database technology. These multidimensional views provide the technical basis for the calculations and analysis required by intelligent applications for providing fast, responsive analysis of data (Han 1997). OLAP provides powerful analytical processing for applications and is optimized for analysis of information. This technology is especially suitable for multidimensional data that includes a temporal component, as is the case in manufacturing process data (Simoudis, 1996). OLAP provides multi-dimensional structures and summarization techniques that enable fast and intuitive access for complex analytical queries (Thomsen, 1997).

OLAP technologies are commonly used in the increasingly popular organizational data warehouses that many large organizations are investing in (Elkins, 1998). Organizations store data about their various business processes in data warehouses. Due to the large variety of the business processes, typical organizational databases contain vast amounts of data. This data is stored for the purpose of managerial decision-making. OLAP presents a multi-dimensional logical view of the data to facilitate analysis. OLAP operations are provided to allow users to interact with the data for multi-dimensional data analysis to facilitate decision-making. OLAP alone does not generate any models about the nature of the data and these are normally developed at the time of the development of the OLAP system. Hence, these models are developed prior to the analysis that the data is used for. OLAP systems interact with statistical analysis algorithms, such as trend analysis or linear modeling to allow for statistical analysis of the data. The models that the analysis is based on are a task for the designers of the OLAP system or for the analyst. Given models of the environment, OLAP technologies can provide an efficient method for easy and flexible access to data to facilitate analysis for decision-making purposes.

2.9 Summary

This chapter discussed different models of decision-making in organizations, from simple, structured models to complex, unstructured models. The ways in which computerized support can be provided to aid the decision-making process by using simple data processing systems, management information systems and decision support systems were examined. Expert systems were presented as a class of systems that

formalize the domain knowledge of an expert to allow non-experts to make decision for decision problems that require domain expertise. The primary limitation of expert systems is the lack of learning involved. If the problem environment changes, then the expert system may no longer be useful to the problem domain. It become apparent that such a solution is not suitable for the dynamic environments of typical business decisions problems.

A suitable decision support solution needs to be adaptive to changes in the business environment in which it operates. The concepts of evolutionary systems and adaptive systems were introduced in this context. It is clear that the ability of machines to perform complex analysis and provide support in terms of the search for, and analysis of, the alternatives that are available to the decision maker is a desirable feature in any system designed to support decision-making in organizations. Another desirable quality is that these analyses and searches are knowledge-driven and that this knowledge is dynamically acquired to keep it synchronous with the changes in the business environment. In other words, the systems that provide support for decision-making should be able to learn from the environment and the decision-making task in order to improve.

Chapter 3: Decision-making for Process Control.

3.1 Processes and Systems in Organizations

This chapter presents the techniques commonly used for process control and the nature of decisions that have to be made in this context. Literature regarding statistical process control and other methods of controlling manufacturing processes and contributions of artificial intelligence to control manufacturing processes are reviewed. The shortcomings of statistical process control and some alternatives to overcome these limitations are presented. The concept of real-time process control as an improvement over process control, since it serves the problem environment better if the solution has a temporal requirement, is explored.

The application of artificial intelligence to process control is investigated. In particular, machine learning, a field of study concerned with the use of artificial intelligence algorithms in computation so that machines can learn and exhibit "intelligent" behavior, is examined. Many machine learning approaches have been applied towards the design of embedded systems to intelligently control the behavior of sub-components of the manufacturing process. Most artificial intelligence-based techniques for process control approach intelligent manufacturing by developing machines that can exhibit intelligent

behavior and can be controlled based on conditions of the environment. This form of solution is usually applied to individual control loops or small pieces of machinery. A shortcoming of these current solutions is that they do not provide a mechanism to control the complete process. They are usually implemented with the intention of controlling the process with no human intervention and have no explanatory component as to the cause of the errors and its consequences to the entire production line. These approaches do not support intelligent managerial and engineering decision-making for the entire process.

Every organization makes use of many processes to provide goods and services to its customers. Modern organizations have a variety of computerized systems that help administer and control the various business processes in which the organizations are involved. It is not uncommon for modern business organizations to have accounting systems and production systems as well as systems to support the marketing function. These systems gather a variety of organizational data to perform their function. Such data is available for other internal and external functions of the organization such as financial reporting and auditing. The data stored by these systems is also used by various decision support systems to support decision-making activities in the organization. Based on the nature of the market in which the organization competes, some of these processes have greater impact on the final mix of goods and services that it delivers than other processes. Consequently, the decision support systems and information systems that support these critical processes have a greater impact on the competitive advantage of the organizations. For example one of the most vital systems for a financial services firm is the one responsible for the gathering and analysis of financial data. The manufacturing

process is one of the most important processes for an organization primarily involved with the production of tangible goods.

Significant resources are involved in the control of manufacturing processes to produce high quality products. Ensuring quality in the final product is an organization-wide effort that involves all levels of management of the organization. Quality control involves operational personnel, supervisors, engineers, multidisciplinary quality assurance teams, research and technology scientists and multiple levels of management. Clearly, systems support for this activity in manufacturing firms is very important. Modern manufacturing organizations spend a considerable amount of money on automatic data gathering devices to capture and process data from all parts of the manufacturing process. This data is typically stored in some form of data repository that may use database technology or log files to be used for analysis of historical trends and report on the overall performance of the production process. Production data is collected at discrete time intervals. The size of these time intervals depends on the response time and sampling frequency of data gathering equipment. The manufacturing process itself is a continuous process that performs a series of sequential transformations to the inputs to produce the final output. Outputs from one part of the production process become inputs to the next and at any intermediate stage of the process. The properties of the intermediate product are the net result of all the transformations that have been applied to it. This suggests that in such systems there will be a high level of interrelationships in the data as raw inputs are transformed into a final product. These relationships play an integral part in any decision-making activity to determine causes of errors and possible alternatives for the correction

of these errors. Statistical process control is one of the most commonly used approaches to quality control for manufacturing and industrial processes.

3.2 Statistical Process Control.

The application of statistical techniques for quality control and improvement of manufacturing processes can be traced to the work of Shewhart in the 1920s, "The application of statistical methods in mass production makes possible the most efficient use of raw materials and manufacturing processes, effects economies in production, and makes possible the highest economic standards of quality for the manufactured goods" (Shewhart and Deming, 1986). Statistical process control is an approach to ensure the control of quality in a product so that it meets the needs of the customer. Quality can be defined as "the total composite product and service characteristics of marketing, engineering, manufacture and maintenance through which the product and service in use will meet the expectation by the customer" (Feigenbaum, 1991). Much of the early research in statistical process control examined the use of statistical methods in implementing the philosophy of total quality control that stems from Deming's work on total quality management (Deming, 1986). Total quality management is a management philosophy whose purpose is to deliver quality to the customer by creating value in a product and making continuous incremental improvements to the product. The review of the literature on quality control presented here is concerned with the implementation of quality control in process industries by using statistical process control.

Statistical process control tries to achieve quality in the product by examining characteristics of the product and the association of these measures to the characteristics of the process. The intent is to identify and minimize the sources of variations in the process so that the final product conforms to established standards of quality. Emphasis is placed on constant monitoring and interpretation of process variables to identify cause of variations in the quality of the final product. Process data is monitored continuously for abnormal variations. If variations occur, adjustments are made to process characteristics to remove these abnormal variations and return the process to a state of normal operation. Statistical process control monitors process data through the use of control charts. These charts plot the progress of process variables as the process is running and allow operators and process engineers to visualize the process. Shewhart originally devised control charts as a means of viewing the state of a process in 1924. The original Shewhart chart measures the mean of a process variable and sets up an upper and lower control limit for the variable at mean ± three standard deviations. The process is said to be in statistical control if the value of the mean is within the control limits. The process is out of statistical control when the mean leaves the mean ± three standard limits. Other methods for developing process control charts, such as the cumulative sum chart and the exponentially weighed moving average chart, have been devised. A sample control chart based on the process mean and ± three standard deviations is shown in Figure 3.1.

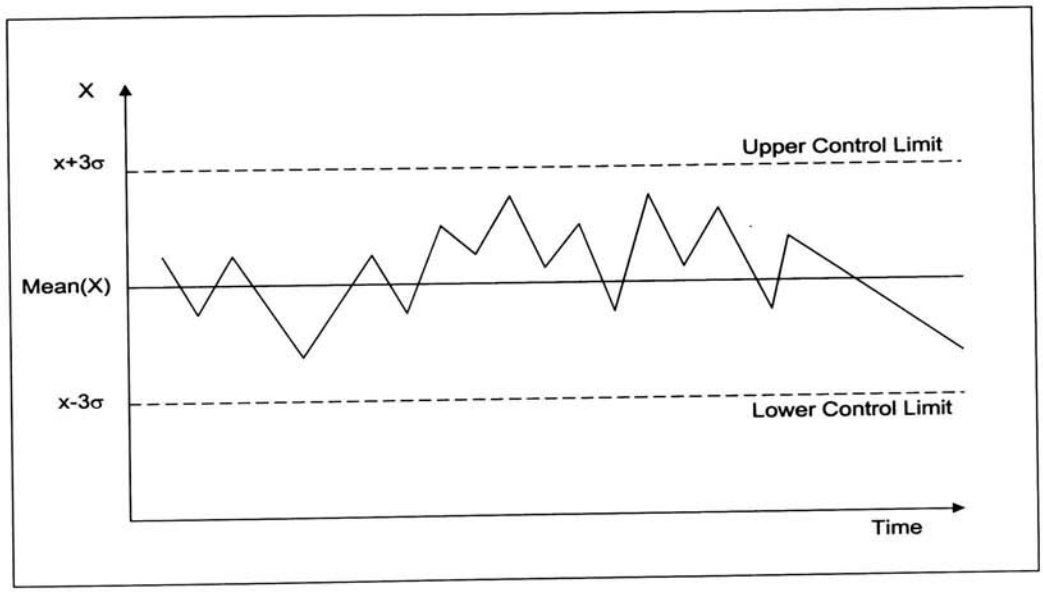

Figure 3.1 A Typical Control Chart for Statistical Process Control

Typically, modern manufacturing environments have a large number of process variables that are collected frequently by automated data gathering equipment. These variables are usually sampled from the entire production process and stored in a data repository. The large number of variables and the high frequency of data collection make it impossible to examine all process variables simultaneously. A common solution to this problem is that a smaller and more manageable subset of the process variables is chosen as the set of variables to be monitored. Operating characteristics that are known, or believed, to have a significant impact on the quality of the product are chosen from various stages of the manufacturing process. Statistical process control is employed to identify unusual variations in any of these critical characteristics. This allows for a manageable subset of the process data to be viewed in the form of control charts. This subset is selected as part of the design of the quality control policy using expert opinion of the process engineers and the research and technology groups in the organization (Oakland, 1996). Much of the current research in quality control and statistical process control is in the area of statistical methods for developing process control charts and their applicability to the nature of the process being controlled.

There is widespread use of information technology in quality control primarily to deliver on-line measurements of numerous variables and provide the ability to store and retrieve these measurements for analysis. Information technology provides the means to perform large scale analysis of data in a short period of time, which is frequently required for analysis of the problem environment. The constant monitoring of critical measures of

product quality provides the detection of out-of control states of the manufacturing process. In addition, information technology provides access to control charts that are updated in real-time as the data is gathered. The interface presented to the user is in the form of control charts; some form of audio/visual alarm is sent if the process is out of statistical control. These interfaces provide the ability to retrieve historical information and allow various forms of analysis. The emphasis of information technology is on supporting the accuracy of control charts and providing constant updates in real-time. These tools provide the control engineer with the ability to add and remove variables that make up the control charts. Information technology can simultaneously monitor a large number of variables and notify the users of out-of-control conditions.

As noted previously, there are some serious shortcomings in the statistical process control approach. One major limitation is the lack of support for analysis in these tools. For example, once an out-of-control process variable is discovered, the statistical process control tools do not provide methods to analyze the data to search for causes for this condition. The lack of ability to rapidly identify the source of variation in a product characteristic is a major drawback of statistical process control and the tools used to support it (Palm, Rodriguez, Spring and Wheeler, 1997). There is no support for providing information on requisite courses of action that can be taken upon the detection of out-of-control conditions. Hence, the focus of these tools is on providing information to the operators and engineers in a variety of formats, not on the analysis of the data or for the support of decision-making based on these data. Thus, statistical process control tools typically provide the means to monitor a process and its state. They do not provide a

means for analysis of the relationships in the process and do not support decision-making for the choice of requisite corrective actions.

An underlying theme of statistical process control is the determination of causality in the relationship between the process variables and the product characteristics. It is assumed that abnormal variations in product quality are caused by variations in process characteristics. Adjustments made to the process will allow the product to conform to quality specifications. Desirable output from a complex manufacturing process is often the result of a combination of multiple simultaneous and sequential treatments on raw materials so that there is a great deal of interdependence throughout the process. In modern manufacturing environments, for example, data for hundreds of variables are collected and examined from all parts of the manufacturing process at frequent intervals. The use of statistical process control to examine the results of single measurements of process characteristics, while simple and easy to interpret, fails to capture the multivariate nature of complex processes. A linear relationship between process characteristics and product quality measures is assumed by statistical process control. Based on this relationship, any changes made to one process variable will cause a corresponding change in the product characteristics. Linearity may be an overly simplistic assumption for the relationships in the process and product quality measures.

Another basic assumption of statistical process control is that the individual data points are independent of one another and that the data are distributed normally. However, for many technological reasons, there is a natural tendency for data that are collected from

physically close sources to be related to one another. This phenomenon is known as autocorrelation. Autocorrelation may exist in data that are collected from the same machine, from the same production shift or from the same batch of the product. The existence of autocorrelation in data violates the assumptions of linearity and independence of individual data made by statistical process control. If autocorrelation exists, then data on standard control charts may appear to be out of statistical control for the mean at times when the process may be running as a stable product and producing good quality product. This is due to the possibility of confounding effects of certain variables on others that cause the resultant process to be within statistical control. If any corrective action is taken on the process in response to these situations, the operator runs the risk of causing a stable process to become out-of-control. It is imperative to consider these sequential relationships within the process characteristics when considering their overall effect on product quality.

In situations where there is significant autocorrelation between process variables the following situations are possible:

i) Individual process variables are in statistical control, i.e.; they lie within the mean ± 3 standard deviations range while the overall process is out of statistical control.

ii) Individual variables appear to be out of statistical control on process control charts while the process is producing good quality product.

Such cases may occur if multiple variables are collected from the same part of the production process so that there is significant autocorrelation among the variables. This is a common situation in modern manufacturing environments where computerized data gathering instruments automatically collect data from multiple machines. In these cases it becomes clear that statistical process control may not be a suitable method for quality control. Hotelling (1947) first discussed a multivariate approach to statistical process control in such circumstances in 1947 applying multivariate methods to bombsight data during World War II. Hotelling developed the T^2 statistic and proposed the use of T^2 charts for multivariate quality control. Jackson (1956) and Jackson and Morris (1959) extended Hotelling's T^2 procedure by using principal components.

Multivariate statistical process control uses multivariate statistical analysis techniques to account for the existing relationships in the data. They provide a more suitable method to detect errors in the production process. Manufacturing environments that produce a lot of correlated data usually have some variables that display a trend, while others follow this trend due to the existing correlation. This data typically has a small number of dimensions and a lot of variables that co-vary with these dimensions. Multivariate approaches make use of techniques such as principal component analysis and contribution plots (Kourti and McGregor, 1996) to identify the direction in the process data while taking into account the existing autocorrelation. Contribution plots can be used to identify the variable(s) that contribute the most to an out-of-control process. This approach is particularly useful for large and ill-conditioned data sets due to the use of multivariate methods that take into account the various relationships in the data and

provide a more accurate technique for the identification of problems in the manufacturing process (Kourti and McGregor, 1996). One drawback of multivariate control charts is that they do not directly provide the information an operator needs, such as, the location of problems in the process and an explanation of its causes.

An important contribution of information technology to statistical process control is the rapid delivery of on-line measurements of process data. This is achieved through a combination of multiple data collecting instruments and computer networks that collect and deliver information to sites that can assimilate all the data and update the control charts. Such an approach, combining the techniques of statistical process control and engineering process control, can provide an important tool for quality improvement (Montgomery, Keats, Runger and Messina, 1994). Information technology can help achieve the rapid identification of the sources of an out-of-control condition through database access and query techniques. Such techniques, however, often involve a time delay in which relevant data is retrieved from the database and analyzed and processed so it can be displayed. Depending on the size of the database and methods of access, this time delay may be too large to serve the purpose of real-time analysis and display. This delay is further compounded when multiple sets of control limits must be maintained and multiple charts must be updated simultaneously. The inability to manage large amounts of data is often an obstacle to the maintenance and real-time update of multiple control charts. More informative and intuitive graphical interfaces are needed to make the information presented by the control charts understandable for all operators and engineers.

Recently, object-oriented techniques (Ham, Jeong, and Kim 1996) and simulation methods (Grega, 1996) have been proposed as possible solutions for process control. Object-oriented methods provide an effective way to model the manufacturing process and incorporating the relationships between the entities of the system but they have no predictive or explanatory capability (Ham, Jeong, and Kim 1996). Simulation methods recognize that process control systems are event-triggered systems and attempt to explain the relationships of the system and predict its future behavior. They provide a powerful method for the analysis of the process data and provide a mechanism to support decision-making for process control (Grega, 1996). The goal of simulation-based methods is to provide techniques for the analyst to understand the current environment and predict future states. Model-based approaches provide a deep understanding of the process and facilitate decision-making within the process control environment. Depending on the level of accuracy and sophistication of the simulation model, simulation driven techniques can assist decision-making for linear and non-linear models of the process. However, simulation models are typically theory-based and may not reflect real operating conditions (Grega, 1996) and the nuances of the actual implementation may not be incorporated in the model. More effective methods are needed to analyze a large amount of complex process control data in order to unearth knowledge useful for controlling large, data intensive, manufacturing process in real-time.

3.3 Real-time Requirement of the Process Control Problem

An important and practical element of the design of systems to support decision-making for manufacturing processes is the temporal constraint of the problem domain. With the advance of manufacturing technology, more sophisticated methods produce more product in smaller amounts of time. When the process is producing a product of inferior quality, it will continue to do so, leading to more waste and consequently larger losses for the organization. This makes the task of detection and correction of errors more demanding in modern manufacturing environments. It seems reasonable for any system that is designed for the monitoring of such processes to function within a temporal bound to provide critical information on the status of the product in real-time. This is an important consideration in the design of such systems in order to minimize the waste and provide warnings to operators as early as possible so that measures can be taken to try to correct the problem.

Real-time processing is an interesting area of research that has not received much attention in the literature. A real-time system is defined as a system that can perform state transitions bounded in the temporal dimensions of the problem domain (Kratzer, 1992). All systems are required to enact changes of state in the current environment; hence the additional requirement placed on a real-time system is the temporal bound. A real-time system can be thought of as a conventional system with temporal bounds. The violation of such bounds may invalidate the operational consistency requirements of the problem domain. Real-time algorithms can then be defined as those that can be guaranteed to execute within a specified response-time window. Real-time systems have greater

requirement in terms of speed, interrupt scheduling and prioritization as compared to conventional process (Kratzer, 1992). This research uses a loose definition of the temporal bound to define the real-time requirement of the system. A stricter definition would require virtually instantaneous response times that may not be a requirement of the system as defined by the problem environment. This definition of the temporal bound is used to define the real-time requirement of the process control problem. This requirement is a pragmatic one in the context of continuous manufacturing environments where a faster response time can directly translate in to a decrease in the number of out-of-specification products produced and a consequent decrease in waste of resources.

Modern manufacturing environments can be characterized as continuous, data-intensive and very dynamic problem environments. Suitable modeling techniques for these environments must be dynamic in that it must be able to detect changes of state and make the required adjustments to incorporate these changes. Computations performed must be completed within the temporal bounds defined for the system in order to ensure effectiveness. Many current real-time systems in manufacturing process control use a real-time implementation of statistical process control (Badavas, 1993). Data from all parts of the manufacturing process is collected in a data repository. This is achieved through the use of data communication software that updates this repository of production and other data. The software responsible for the updating control charts accesses the real-time database for the most current information. Users may be allowed to add or remove the charted elements based on variables that are of particular interest at the time. A schematic for a typical real-time statistical process control system is provided in

Figure 3.2. Such systems are on-line versions of statistical process control systems and other than satisfying the temporal requirements, do not provide any additional analytical or modeling benefit over traditional statistical process control systems.

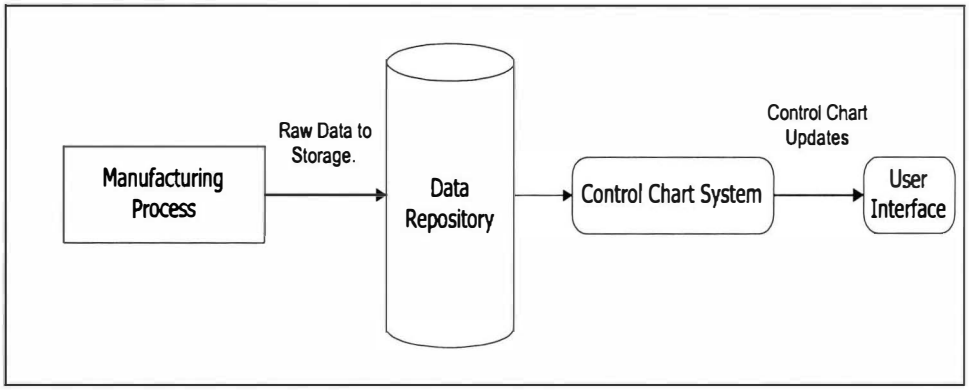

Figure 3.2 A Typical Real-Time Statistical Process Control System

3.4 Artificial Intelligence in Process Control

In recent years, there has been much interest in the application of artificial intelligence to provide techniques for process control in the engineering and manufacturing literature. The ability of artificial intelligence techniques to learn the state of the system from process data, to explain and model the system and be able to handle imprecise or fuzzy and complex information is seen to have potential for the highly demanding process control problem domain (Pham and Oztemel, 1996). Neural networks and expert systems are prominent technologies developed by the artificial intelligence community that have been applied to develop intelligent applications for control of manufacturing processes. Machine learning algorithms deal with the development of machines that can improve with experience and increase their efficiency in the domain of application. Many of the algorithms that are used for machine learning are similar to the algorithms in data mining. The following sections discuss the use of expert systems in process control, machine learning and neural networks.

3.4.1 Expert systems in Process Control

As stated earlier, expert systems have had demonstrable benefits in aiding non-expert decision makers in making decisions in problem domains that require the expertise of domain expert. Expert systems have been applied to the problem of process control primarily to provide off-line analysis of error situations in the manufacturing process. Typically, expert systems have been applied elsewhere in manufacturing industries to difficult and unstructured problems such as planning and scheduling (Alexander, 1987). The primary difference between statistical process control and expert system applications

is the level of support for analysis of states of the process. Historically the analysis of error conditions has been left to research and technology scientists, quality experts, or very experienced operators who are perceived to be experts in the process (Oakland, 1996). Expert systems have typically been applied to formalize the domain-specific knowledge of such experts so that it can be used by the system to offer analytical support to explain error conditions (Affisco and Chandra, 1990). These systems have typically been implemented as off-line systems that analyze the data and provide analysis.

Many systems have been designed to support quality control using expert systems technology and have focused on the methodologies for knowledge engineering and the design of explanation systems. Few of these systems, however, are reported to be performing satisfactorily (Pham and Oztemel, 1996). Much of the focus of the application of expert systems in manufacturing control is in the selection of the appropriate control charts to be used for the analysis (Dagli and Stacey, 1988). The application of expert systems to process control suffers from the same problems that expert systems applications in decision-making for other organizational activities suffer from. Their lack of learning and hence the lack of growth in their applicable problem domain is a significant. This drawback is especially significant for production systems, which are very dynamic in nature due to rapid changes in production and information technology. Little support is offered by expert systems based approaches to provide effective means for analyzing the underlying causality in the relationships of process variables or for the automatic interpretation of out-of-control conditions (Pham and Oztemel, 1996). There is some evidence in the literature of attempts to combine machine learning algorithms,

including neural networks and decision trees, and expert systems (Calabrese, et. al., 1991; Smith and Yazici, 1992). Pham and Oztemel (1996) observe that more research is needed in this integration. The goal of these attempts is to design a system with the learning capabilities of neural networks and the explanatory strengths of expert systems.

3.4.2 Machine Learning

A simple definition of machine learning is given by Mitchell (Mitchell, 1997) as an area concerned with the construction of computer programs that automatically improve with experience. This embodies a definition of learning as the ability to improve with experience. It simply states that the field of machine learning is the construction of machines, computers or computer programs that can learn. This definition of machine learning defines learning as the ability to improve, with experience, in the task that they are assigned. Simon (Simon in Michalski, 1983) provides a more formal definition of learning:

> *"Learning denotes changes in the system that are adaptive in the sense that they enable the system to do the same task or tasks drawn from the same population more efficiently and more effectively the next time."*

Machine learning draws from a number of parent fields including artificial intelligence, statistics, control theory, information theory and cognitive science. Machine learning involves the search of very large spaces of possible hypotheses for one that best fits the

observed data (Mitchell, 1997). This perspective is consistent with many problems that are dealt with by statistical or by artificial intelligence algorithms.

This discussion of machine learning is limited to machine learning algorithms since they have a strong bearing on some very important data mining algorithms. Carbonell, et. al., present a taxonomy of machine learning research to classify artificial intelligence based machine learning research (Carbonell et. al., in Michalski et. al., 1983). They classify machine learning systems along three dimensions:

i) **Underlying learning strategies used:** This dimension considers the amount of inductive inference that the algorithm is able to develop. Learning can come from a number of different methods such as rote learning, learning from instruction, learning from examples, learning by analogy, and learning from observation and discovery. Learning from examples can be classified based on whether the learning comes from looking at positive examples or from negative examples, or a combination of both.

ii) **Representation of knowledge:** The knowledge that is learned may be represented in the multiple forms depending on the functional model of the task used. Some forms include:

a) **Parameters to algebraic expressions:** as in regression analysis.

b) **Decision trees:** Knowledge is represented as the various branches that a decision process may take.

c) **Production rules:** These represent a mapping of the conditions under which certain actions are taken.

d) **Graphs and networks:** The learning from the application of methods such as neural networks usually generate knowledge representations as graphs and networks.

iii) **Application Domain:** The domains in which machine learning systems have been applied offers a dimension along which machine learning algorithms can be classified. There are a number of such problem domains including voice and image recognition, medical diagnosis, chemistry, natural language processing and robotics.

Some major classes of machine learning algorithms that are pertinent to this research are neural networks, decision trees, and statistical algorithms that implement methods of clustering, regression analysis and principal component analysis.

3.4.3 Neural Networks

The study of artificial neural networks has been partly inspired by the branch of artificial intelligence that seeks to develop machines that can act like human beings. Attempts to model computing activity on knowledge of the working of the brain have been the major inspiration of the development of artificial neural networks. Warren McCulloch and Walter Pitts developed the concept of neural networks in their work entitled "A Logical Calculus of the Ideas Imminent in Nervous Activity" (McCulloch and Pitts, 1943). This work combined ideas about finite state machines, linear thresholds and decision elements and logical representations of various forms of behavior and memory (Minsky and Papert, 1988). Limitations in single-layered networks were discovered when Minsky and Papert proved that these networks could not represent simple functions such as the Boolean XOR function (Minsky and Papert, 1969). Work on neural networks saw resurgence in the 1980s with the advent of the back propagation algorithm (Rumelhart and McClelland, 1986) and work done on parallel processing and multi-layer networks (Mitchell, 1997).

Mitchell (Mitchell, 1997) observes that there have been two directions of research in artificial neural networks: the first direction looks at attempts to model the working of the human brain, while the other has been motivated by the attempts to obtain highly effective machine learning algorithms. This discussion of neural networks is limited to machine learning algorithms that can be applied to decision-making situations.

The back propagation algorithm is very commonly used for training neural networks. The model of a neural network used by the back propagation algorithm is a multi-layered network that consists of a number of nodes that are connected to other nodes by edges as shown in Figure 3.3. The neural network consists of input layers which represent the input to the system; output layers, which represent outputs of the system; and one or more hidden layer that are between the input and output layers and are responsible for intermediate processing. The number of input layer nodes is the number of inputs to the system and the number of output layer nodes is the number of outputs of the system. The number of hidden layers is decided either by some heuristic based on apriori knowledge of the relationship between the input and output nodes. A trial and error approach that minimizes error rates, number of iterations and prediction errors is often used to determine the number of hidden layers. The literature also does not offer any rules on the number of nodes in the hidden layer(s). Each layer in a neural network consists of a number of nodes. Each node in a layer is connected to every node in the layer above and below by an edge.

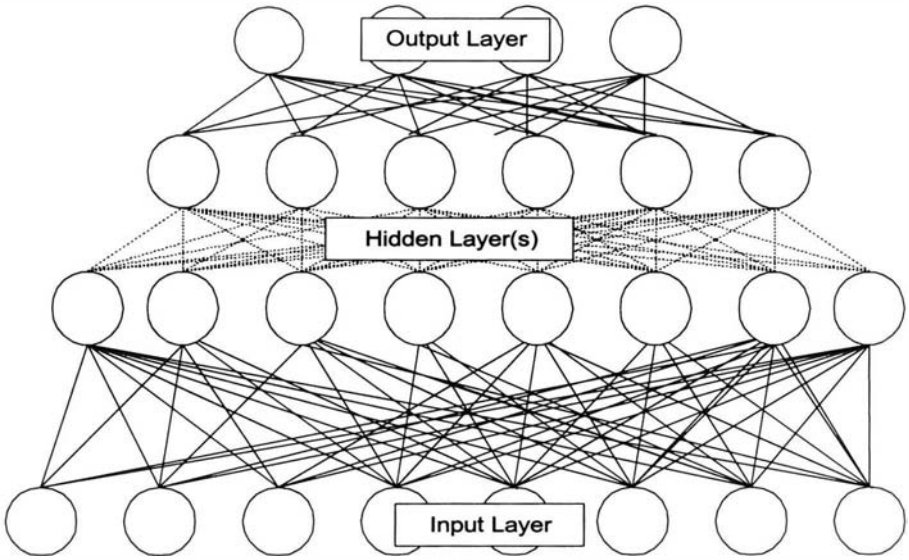

Figure 3.3 A Multi-Layered Neural Network

The back propagation algorithm initially assigns random weights to each edge and searches the space of possible hypotheses by an iterative process, reducing the overall error in the model to fit the training data. The algorithm learns the weights for a multi-layer network by minimizing the squared error between the network output values and the target values for these outputs. Many attribute-value pairs represent instances of the target function that is to be learned. The target function output can be singular or a vector of several real or discrete values. The training examples may contain noise, errors, or a combination of these. This feature adds robustness to the algorithm. Mitchell (Mitchell, 1997) offers some characteristics of problems for which this technology is well suited. The terminating condition for learning in the back propagation algorithm is usually implemented as the point when the overall training error, measured by the difference between the outputs and the targets, falls below an acceptable level.

The back propagation algorithm iteratively reduces the difference between the observed value and the computed result of inputs by adjusting the weights of its edges between each pair of nodes in the neural network. This difference between the observed output values and the output values generated by the neural network is known as the training error. For a neural network using the back propagation algorithm, the training error typically has an exponentially decaying pattern that assumes an asymptotic form after achieving a certain error rate. The neural network is trained to the region just before it achieves this asymptotic value. Stopping the training process well before this value is

achieved decreases the accuracy of the neural network, while training beyond this region reduces the generalizability of the neural network.

The learning rate represents the aggressiveness with which the neural network achieves the trained weights for the nodes. A higher learning rate in a neural network will achieve lower error rates in a fewer number of iterations but may fail to capture the nuances in the variations of the data. Lower learning rates may over-involve the neural network in capturing random variations in the data without significant gain in the training, and thereby significantly increase training times without corresponding gains in accuracy. Learning rates typically range between five percent and ten percent depending on the nature of the data and the extent of random variations within them. There is a trade-off between over-fitting the model and the ability to generalize the model that must be considered here. As the model iteratively reduces the error rate, it has a tendency to over-fit the training data. This tendency may result in a loss of the ability to generalize the network to other problems. There is little in the literature in terms of offering guidelines on how to deal with this trade-off and a combination of judgment, heuristics and standard values are commonly used. Neural networks have been very successfully applied in a large number of application such as control systems, robotics, automation of manufacturing, control of self-driving vehicles and aircraft, voice recognition, image recognition, economic prediction modeling and many more engineering and business applications (Mitchell, 1997).

Neural networks have been successfully applied to solve many manufacturing problems. They have been used for their ability to capture complex, non-linear relationships in scheduling, computer integrated intelligent manufacturing and process control (Dagli, 1994). The design of hybrid intelligent systems that use a combination of neural networks and rule-based expert systems have also been suggested in the literature to utilize the strengths of each technique for providing explanatory and predictive capabilities to the process control system (Madey, Weinroth, and Shah, 1994). Intelligent systems need to learn autonomously and adapt in uncertain or partially known situations in order to progress to full engineering implementation (Stacey, 1994). They need to be able to predict future states of the system and be able to offer plausible explanations to users as to why the states were predicted. Process control systems ultimately will be used as decision support systems to help users make decisions about the manufacturing process. Such a use of process control systems would benefit from modeling support to help users better understand the manufacturing systems.

Chapter 4: An Integrated Model

4.1 Design of the integrated model

This research proposes a model for the integration of data mining and on-line analytical processing to provide intelligent decision-making capabilities. The problem context of monitoring and controlling a large automated continuous manufacturing process is used as the basis to develop this design. The environment is sufficiently mature in terms of the volume of data available, which makes it an ideal candidate for data mining. Typically, a large number of variables are involved and there are many complex relationships in the data that have bearing on decisions to be made in this environment.

Design is the use of scientific principles, technical information and imagination in the definition of a system to perform pre-specified functions with the maximum efficiency (Fielden, 1975). The design of information systems is a goal-oriented activity. Some design goals for this system are:

i) The system must be based on accurate models of the process and use these models to support analysis of the process.

ii) The system must be able to take a proactive role in the identification of imminent errors in the process and detect the possibility of their occurrence.

iii) The system should use the knowledge from the process models to answer questions about the process, provide information about normal operations and probable causes of error.

iv) The system should be able to react to anomalies in the process in a responsive manner and suggest possible causes.

v) Models of the process must constantly adapt to changes in the process.

These characteristics imply that the system can quickly and intelligently process huge amounts of data and react to subtle changes in process characteristics, evaluate their threat and offer adaptations to deal with these threats.

Based on the above goals of the system, the following components of the proposed system are developed.

i) A set of **Accurate Models** derived from the data, which can be used to explain the relationships that exist in the data.

ii) A **Model Updating** component to allow for evaluation and re-generation of the process models so that they reflect the current states of the process.

iii) A **Proactive Analysis Component** that can analyze current data as the system acquires them and check for the extent of conformity of the current data with known models of the system. If the current data falls into any known patterns of failure, the system can serve as an early warning system so that potential failure can be avoided.

iv) A **Query Response Component** that can answer questions from users about the current state of the process based on the models of the system and current and historical data.

A number of methodologies can be used to create sophisticated models depending on the data mining technique employed and the nature of the data used to create these models. For example, the use of artificial neural networks will create a complex model that is very accurate in terms of predictions and learning the nature of the data sets. Models created using neural networks are not very easy for humans to understand and effort effort is required to explain the results generated by neural network models. On the other hand, decision trees offer a mechanism of creating models of the data that is easy to understand. However, the level of accuracy and extent of conformity with actual data using this approach is not as high as that obtained by using neural networks.

The data used to develop the models may be known as "good" process data so that the system learns descriptions of a good process and recognizes the normal state of operation of the process. An equally valid approach is to train the system with "bad" process data so that the system is focused towards recognizing out-of-control states of the process and can quickly recognize such states from the current process data. There is also the possibility of creating a system that combines multiple data mining models and the types of data used for training. This approach seems to hold more promise for creating a set of sophisticated models; however, it has high processing and analytical requirements.

Data from dynamic processes is inherently dynamic. This implies that the relationships in the data are subject to change. Therefore, any system that supports decision-making based on these models should dynamically update the models to reflect current states of the process in light of changes in the operating environment. Otherwise, users run the risk of making decisions on information that does not hold true in the current environment. In the proposed system, the data mining component responsible for the maintenance of the explanatory models of the system must constantly evaluate these models based on new data. This process would keep the models up-to-date with the current data from the production process. This must be done in parallel to, and separate from, the active, on-line components of the system.

OLAP allows for the analysis of large quantities of multi-dimensional data by giving the user multiple views of the data. OLAP stores these multiple views of data and stores aggregates with the data so that these views can be made available to the users in a much

more responsive manner. These views of the system data are defined based on the models of the process created by the data mining component, so that the system can explain and provide suggestions on queries regarding the states of the process. For example, suppose that for a set of variables, the values for of means and standard deviations are seen to be critical to the identification of out-of control states. Data cubes can be constructed with these dimensions and made available to the OLAP components so that it can be available for on-line analysis of the system.

The above sections have described the design goals of the system in general to provide an explanation of the rationale behind these choices. The following sections describe the model for the system to integrate the technologies of data mining and OLAP to provide real-time process monitoring and control that will satisfy the above design goals. Descriptions of each component's implementation to achieve the design goals are provided.

4.2 Components of the Model

The proposed model of the integrated system consists of the following five components, as shown in Figure 4.1:

1. Manufacturing Process.
2. Data Repository.
3. OLAP Component.
4. Data Mining Component.
5. User Interface.

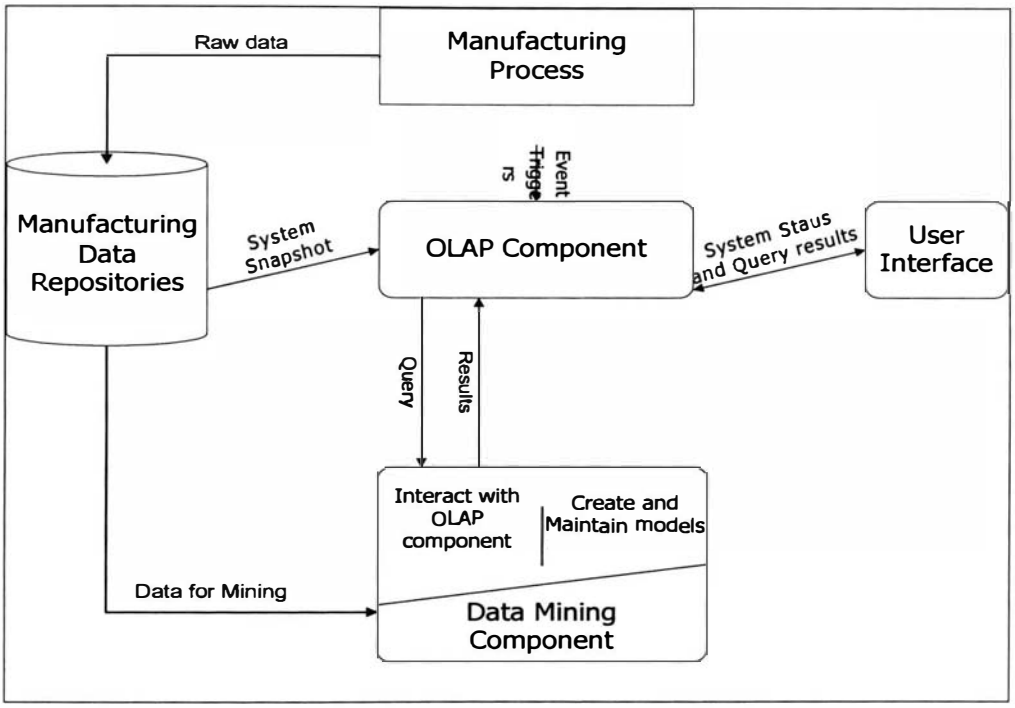

Figure 4.1 Model of Integrated System

4.2.1 Manufacturing Process

The application domain is a large, fully automated, continuous manufacturing unit, such as, an assembly line environment, where data collecting equipment periodically collects various established performance measures. These data points are typically established before the design of the process control system. Data collection instruments, with the help of data storage systems, deposit this data in a data repository. Production monitoring systems use this data to monitor the state of the system and establish whether the system is running within established parameters. An important feature of the production process data is that it is multi-dimensional, i.e., it is derived from the many processes of the production system and has information about multiple aspects of it.

Process control systems are event driven systems. In current process control systems, if critical variables are out of their established ranges they are "flagged" and manual intervention is required to investigate and solve the problem. To achieve this, certain data items are identified as being critical measures of process performance. The process is assumed to be running in a normal state of operation when these variables are within the established parameters. When these variables are outside of the established range, a "triggering" event is said to occur. Such events identify critical error conditions that signal a disruption of the steady state of the process. In current systems, the characteristics of such events are typically identified by expert opinion.

In the proposed model, the event triggers are identified by a combination of expert opinion and applying data mining techniques to the manufacturing data repository for a set of variables that can be used as predictors for the state of the system, with acceptable levels of accuracy. A possible criterion would be to identify variables that have a high correlation with the likelihood of system failure. This set of variables would be constantly under review by the data-mining component of the system, which identifies the best performance predictors and their relationships with quality measures. These variables would be made available directly to the OLAP sub-component from the data collection instruments of the manufacturing process. This allows the system to quickly identify and react to emerging trends in the current process and keep the response time for the system to a minimum. The raw system performance data is also passed to the data repository for permanent storage and future processing purposes. Logic can easily be incorporated into the OLAP sub-component to sieve out these variables based on information obtained from the data mining sub component. It is recognized that this introduces a level of data redundancy in the system; however, this choice is made in the interest of response time.

4.2.2 Manufacturing Data Repository.

The manufacturing data repository stores multitudes of data from all parts of the process. The different data collection units from the entire process automatically store this data in the data repository. For most firms, this data is critical process data that is collected to monitor and control the manufacturing process. Scientific data collection instruments can collect millions of bytes of information in very short periods of time. The data is typically time-indexed to facilitate easy retrieval and processing. As discussed earlier, most of this

data is usually not used by process control systems that use statistical process control (SPC), or similar technologies. The ailment that these systems suffer from is not lack of data but rather too much data and not enough information, which creates a high rate of under-utilization of data.

Techniques such as data mining, machine learning and neural networks rely heavily on the availability of sufficient volumes of "good" data to develop models of processes in sufficient detail to explain and predict the performance of the system. In environments where multiple systems are collecting data and feeding it into a common data repository, it is quite common that some data points will be lost. This could happen for a variety of reasons such as a malfunctioning data collection device or data loss in transmission. Such phenomena are causes of missing data. It is critical to have a data-cleaning step to examine and process such conditions and guarantees an acceptable "quality" of the data. These data points may be replaced by nominal data, or some other technique may be applied to account for such conditions. It is very likely that raw data would be collected in a number of different scales. An additional pre-processing step is to normalize all raw data to a common scale through mean centering and variance scaling, to allow for further processing. The manufacturing data repository provides the data input interface for two critical components of the system, the OLAP component and the data mining component.

4.2.3 Data Mining Component

In the context of a real-time process control system, data mining can be used to extract meaningful relationships between the various data items in the production data. The data-mining component of this system is divided into two parts:

a) **Creation of the Model Base** - responsible for creation and maintenance of the models associated with the process.

b) **Interaction with the OLAP component** - responsible for interaction with the OLAP component and passing the correct model parameters of the system to use for analysis.

a) Creation of the Model Base

The result of data mining is a set of models that describe the operation of the process in normal operation and model the conditions under which failure may occur. Also the models should be able to predict future states of the system, given information on the current state. The objective of the data mining process is to create a "model base" which describes the correct and incorrect operation of the production process in terms of process variables. Upon creation and validation, these models can be used by the OLAP component to evaluate the process at any instant in time with respect to its stability and likelihood of failure.

For this research, the data mining task involves searching for patterns or trends in data elements that frequently occur preceding the occurrence of an error condition. These may involve establishment of acceptable parameters for data elements, composite or otherwise, that lead to an error condition. Initially, the models are created from the complete data repository of the system in an off-line mode separated from the on-line component of the process control system with the help of expert opinion. They are tested and validated on historical data before being deployed in the system. The model base is updated on a regular basis, especially if known changes are made to the process parameters. The two primary functions that the models developed using data mining must serve are:

i. Predict failures of the manufacturing process.

ii. Provide models for the analysis of the process.

i. Models to predict failure of process.

A primary requirement of models needed to support process control is that they must be able to identify imminent failures of the system and provide early warning. The system needs to learn the historical patterns that have historically led to failure by using data mining techniques. Before a model can be used to predict a state of any system, the characteristics of the state to be predicted must be operationalized. In other words, there must be a set of defined inputs and outputs that can be used to describe the states of the system. A set of variables that are perceived to be critical to the steady operation of the system was gathered from expert opinion. When anomalous deviations in these variables

occur, the process is out-of-control. A snapshot of the entire set of process variables will be used as the input to this model to develop predicted values of the critical variables.

An artificial neural network based on the back propagation algorithm is developed and trained on the actual data from the manufacturing process. This neural network is used to examine data from the system and predict whether the critical variables are within their established ranges. Information regarding the predicted values of these variables, obtained from the neural network, will be used to predict failure in the manufacturing process. This is different from other methods of error detection in that the prediction is obtained from non-linear models of the system that consider the inter-relationships among process variables.

ii. Models for analysis of the process

Data mining can construct models from the process data but does not provide any guarantees on the effectiveness of the models. The criteria for which the search was carried out and the types of models, patterns and associations that were being mined for determine the effectiveness of these models. The models that are generated must be interpreted, evaluated, validated and tested on the real system. These models may be examined by a variety of techniques including expert opinion and testing. The predictions of the developed models can be tested on historical data sets where there is a known occurrence of failure to see if, and when, the system predicts the failure. The output of these steps is a set of descriptive and predictive models that explain when and why error conditions occur in terms of the target data set.

b) Interaction with the OLAP component

The data mining component identifies the views that the OLAP component needs to support to the efficient analysis of the data. For effective analysis, the OLAP component requires views of the process based on the model that is currently being used to analyze the data. The neural network is trained for steady state operation of the process and made available to the OLAP component to analyze the current snapshots of the process. A snapshot of the process is defined as a vector of all the inputs to the neural network.

The OLAP component must also provide means for the analysis of the variables responsible for error conditions as they are detected. The data mining component will react to requests by the OLAP component and provide models required for analysis based on the current error condition being analyzed. The application of the data mining techniques gives the system a "model base" to describe causes of possible error conditions in the data. A model of the system would identify a set of trends in the process that should continuously be checked. A model would define a "time-window" that corresponds to the number of data points required to make valid and accurate predictions of imminent failures from the incoming, real-time process data. It also would define a set of parameters that identify and describe a set of key variables that can be used as event triggers to identify possible error conditions in the process. The interaction of the models of the systems as defined by the data mining components is passed to the OLAP component for constant analysis of the system. On identification of non-conformity to the process model, the OLAP component queries the data-mining component for either a

changed model or an error condition. This information is passed back to the OLAP component. If there is no conformance to any model present in the data mining component's model base, then an error condition is said to occur and this information is passed to the OLAP component. Otherwise, a different set of trends and parameters are passed to the OLAP component and a process model changeover occurs. New data may be obtained from the data repository and the process starts to be monitored.

4.2.4 OLAP Sub-Component

OLAP is a class of technologies that provides multidimensional views of data supported by multidimensional database technology. This technology is suitable for multidimensional data that includes a temporal component, as is the case in manufacturing process data. The OLAP component accepts event trigger data from the multiple data collecting devices of the process and analyzes them to see if sufficient evidence can be found for the likelihood of an error condition. This process follows the simplistic view of process control where certain variables or clusters are checked for conformity to known models of a normal state of operations. These models are derived from data mining algorithms that consider the normal operation of the process. For this research, this would be the trained neural network that analyzes process data to see if the process is in or out control as defined by abnormal variations in these critical variables. When the critical variables are within appropriate ranges, the OLAP component does not have to do anything. The actual results as obtained by the process would serve as a confirmation of the fact that the process is not going out-of-control. When an event trigger detected by the OLAP component identifies an imminent problem in the process,

action is required. The first course of action is to flag the process to be leaving the normal operating range and hence inform the user of an imminent problem in the process. This can be done by a simple comparison of the predictions from the neural net and normal operating range means of the critical variables of the process. This is further reinforced by comparison with the actual values from the process.

When the process is leaving normal operating range, two questions need to be answered by analysis of the variable(s) that identify the process to be out-of-control. The process control system needs to identify causes of these abnormal variations in the critical variable from abnormal variations in variables that occur before the critical variable, and it needs to identify what the normal values of these variables are so as to provide some indication of corrective action required. The data mining component develops decision tree models for relationships between the critical variables and all variables that occur downstream. Once an unsteady variation is identified, the model(s) pertaining to the out-of-control variables is requested by the OLAP component from the data mining component. In such situations, the OLAP component requests a snapshot of the current state of the system from the production data repository. This snapshot would consist of a time window of process data. For example, the OLAP module may require all process data from the last hour of operation, or it may require data from a defined cluster of variables for a certain period of time. The content of this data cube would be defined by the model of the effect of the critical variable that is out-of-control and the set of process characteristics that it is known to affect. The fact that these models are predictive and

descriptive in nature allows for the forecasting of results from the out-of-control condition.

If sufficient evidence of process error is not discernible from the event triggers, then it must be the case that models for the current condition do not exist. The OLAP component passes the variables under consideration to the data mining component and new associations must be derived for those data items as relationships develop between the critical variables and the process variables under consideration. Depending on the observed data and the extent of system information, the OLAP component may query the data mining component for its data for error conditions that may possibly be developing. If this query returns a positive result, then the results will be passed to the user interface along with explanations of the possible errors that may be developing, a prognosis and possible remedies. In either case, a set of output and input variables are passed to the data mining component for it to search for associations.

4.2.5 User Interface.

The user interface interacts with the OLAP component to display information on the status of the process. If an error condition occurs the user interface uses audio alarms and graphical displays to alert the user. By interacting with the data-mining component, the OLAP component obtains information on possible causes and remedies for the occurring errors. This information can be passed to the user using drill-down features. Depending on the complexity and sophistication of the model base, it may be possible to provide decision support features such as queries on deciding possible remedies and their efficacy

in the given situation. Under normal operating conditions, the OLAP component has a complete set of information on the operating conditions of the process and can provide information on the various parameters and components of the production system. This research is concerned with the implementation of the model to study the interaction of the data mining and OLAP components to provide a real-time process control system that supports intelligent decision-making. The interface component is not of primary concern in this research.

4.3 Summary

Production environments collect large amounts of data from all parts of the production process at frequent intervals. Over time, this results in enormous repositories of data that are not being used. This multidimensional data contains vital information about the production system and the relationships of the components of the system. Current techniques in process control do not offer explanatory and predictive capabilities required to make sense of the complex relationships and interactions in the process data. Systems must be able to use these large volumes of complex data effectively. This research proposes a model for a real-time process control system that integrates the data mining and OLAP technologies to take advantage of the wealth of information contained in these data repositories. The resulting system can predict and explain the occurrence of error conditions in the system and adapt itself to changes in the operating environment. The model in this research was presented with reference to its application in a large automated manufacturing environment, a problem domain that featured continuously generated data and required fast responsive and intelligent processing.

Chapter 5: Prototype of the Integrated Model

5.1 Introduction

This chapter describes the details of a prototype of the integrated system that was developed to test the model. The inputs and outputs of the system and their association with the manufacturing process that the system models are described. The implementation details of applying an OLAP-only apprach and the implementation of a solution using an integration of OLAP and data mining components are also described in detail.

5.2 System Inputs and Outputs

The inputs to the system are a set of forty-one process variables from a critical sub-part of the complete manufacturing process called the extrusion process. This sub-process is responsible for the final processing of the melted raw materials for conversion to the final product and represents a critical stage in the manufacturing process. This sub-process is identified by the process experts to be an error-prone component of the manufacturing process and is a suitable target application area. A set of observations that represent the normal operation of the manufacturing process with nominal errors in the final output was used to create a training set for the data mining components of the integrated system.

Another set of observations that produced known errors in the output is used as the verification set to test the effectiveness of the system. For the prototype, the training set contained 10,000 observations of all input and output variables, while the test data set contains 1500 observations, including a known number of errors in the final output. The following sections describe the implementation of the integrated system and the OLAP subsystem.

5.3 Integrated System Implementation

The integrated system was implemented as an object-oriented system using the C++ programming language. It consists of three major classes that represent the major components that are incorporated into the system:

i) An artificial neural network based on the back propagation algorithm,

ii) Decision trees for each output based on the ID3 algorithm (Quinlan), and

iii) A data class that represents the multi-dimensional data component and contains methods for creating multiple views of the data, and stores the actual data with statistical information about the data.

The artificial neural network and decision tree are parts of the data mining component of the integrated system, while the data objects comprise the OLAP component. These components work together to develop knowledge about the manufacturing process and provide access to the data to provide a medium for intelligent decision-making. Figure

5.1 shows an expanded version of the model presented in chapter four and forms the schematic for implementation and serves as a reference for the subsequent discussion in this chapter.

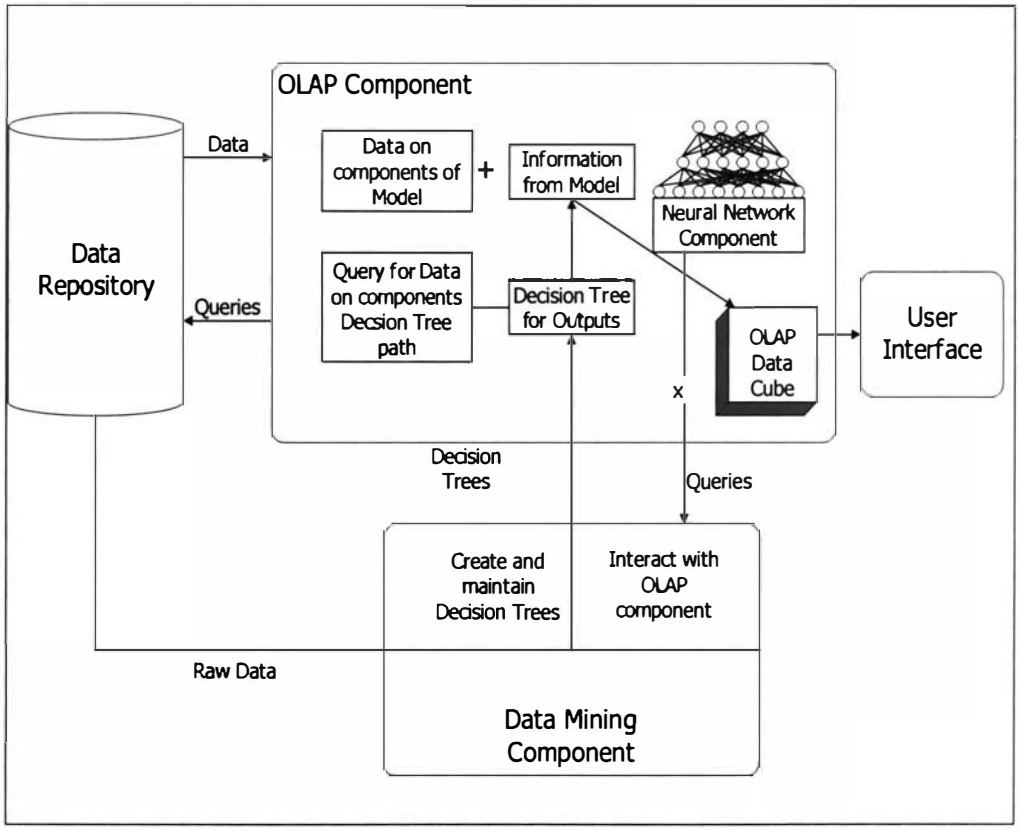

Figure 5.1 Expanded Model for implementation of integrated system

Using the training data set, the neural network was trained with the set of inputs of the process using the back propagation algorithm. The training data was also given to the ID3 algorithm to develop decision trees for each of the output variables. The trained neural network and the decision trees were developed, tested and stored. This process represents the training of the integrated system. The result of the training process was an artificial neural network that was trained for the detection of whether an output variable is within acceptable limits of operation. Another output of the training process is one decision tree for each output variable. A set of rules can be developed from each decision tree that identify the different variables that are causes for the errors in the output, as identified by the decision tree algorithm.

After the rules for each of the outputs of the system have been developed and the neural network has been trained using the data mining components of the integrated system, these components are loaded into memory. The OLAP component of the system initially loads the test data as a matrix of the variables of the system across the number of observations that are available. Each observation in the test data represents samples taken at one-minute intervals. This data is normalized with means and standard deviations that are created from the training data that represents the normal operation of the manufacturing process. The normalized data is then passed to the trained neural network and output values from the neural network are calculated. These outputs are then compared with the acceptable limits for the process outputs and are also verified against the actual outputs at the time so that the level of accuracy of the neural network can be constantly evaluated.

If an actual output does not conform to the acceptable limits the system recognizes that there is an error in the process. The output detected as out-of-specification is then provided as an input to the decision tree component of the integrated system. The integrated system stores the rules that were learned as part of the training procedure as an array of decision tree objects. These objects take the normalized value of the inputs and outputs at that time and trace a path on the decision tree to generate a rule that represents the decision tree's estimate of a cause for the error. Rules are represented by a set of value for a set of input variables that explain the reason why the output is out of specification. In the training and verification stage, the values for the inputs that are generated by the rule sets are compared to the actual values for the inputs to verify the rules created by the decision trees.

The result of following a path along the decision tree is a set of variables that lie along this path. This set of variables is the integrated system's best estimate of causes for the error under consideration. Once the variables that are believed to be the causes of the error are known, the set of variables is then passed to the OLAP system as the dimensions along which the trends in the process are to be viewed for the error under consideration. The OLAP component of the integrated system maintains additional information for each variable in the system, including both inputs and outputs for the system. This information includes means and standard deviations for each of the variables. This information is incorporated into the views of the system developed for the user that is analyzing the process. The user can analyze trends in the key variables that are causes of error across

time. For instance the user may view the standard deviations that are outside of acceptable control limits, note deviations of the key variables from their means and create additional customized views on other dimensions. These dimensions may include variables organized by physical proximity in the manufacturing process, historically error-prone parts of the process and additional dimensions. Many of these dimensions are available in the prototype that was created as an illustration of the model. The prototype does not include a graphical interface; however, data for these dimensions are available in memory as data objects and procedures exist for the creation of views based on these dimensions.

The following sections describe each of the components of the integrated system: the neural network component, the decision tree component and the OLAP component, in more detail.

5.4 Neural Network Component

The Neural Network class is implemented as an array of layers of nodes, with an input layer, a hidden layer and an output layer. The number of nodes in the input layer is equal to the total number of input variables with which the system is initialized. The number of outputs for the neural network is the number of outputs for the system. The number of nodes in the hidden layer is calculated by the integer result of the division of the number of input and output layer nodes. In the case of the prototype system, the number of hidden layer nodes is approximately fifteen. The two outputs represent the variables that have been identified by process experts to be critical measures of process stability. The neural

network was trained on data that has been normalized using the model means and standard deviations for the production process. A standard learning rate of 5 percent was used to train the network. The neural network is trained to within acceptable range of error equal to 10% by using the training data that represents the normal operation of the manufacturing process.

After one instance of the neural network is created, the training behavior of the neural network object requires references to data objects that represent the process inputs and outputs. These data objects contain the data in normalized form and the model means and standard deviations for the process as data members. The neural network object slices row vectors of these data members and feeds them into the input layer and moves them through the neural network layers according to the back propagation algorithm. Errors in the output layer are calculated according to the standard back propagation algorithm. The entire training set is treated as the input and output matrix for the training of the neural network and the neural network is trained on this training set until the network achieves an error rate of 10 percent. The edge weights of the neural network and all other parameters of the neural network are stored in file so that the network can be reloaded from file without having to re-train the network.

5.5 Decision Tree Component.

The decision tree component of the integrated system is implemented as an array of pointers to decision tree objects with one tree for each output. The decision trees are trained with the same data as the neural network. The purpose of the decision tree

creation is to create decision paths for each output of manufacturing process in terms of the inputs of the process.

The inputs are categorized based on the standard deviations of the variables. For example, input variables that have a normalized value of one are placed in a different category than variables that have a normalized value of two. The entire training set is used to create the decision trees. The root node of each of the decision trees is the input variable that has the largest discriminating power to discriminate between the states of the process. Each subsequent node of the decision tree is an input variable of the manufacturing process. The number of children for each node of the decision tree is based on the number of categories that the variable exhibits in the training data set.

The output variables are categorized based on the acceptable control limits as set by the manufacturing process experts, which is ± 3 standard deviations. The classification of the output variable is a Boolean classification, where a 0 represents an output variable that is within acceptable limits while a 1 represents an output variable that lies outside of the control limits. Once the training set is adapted into categorized values the decision tree algorithm uses these categorized values to create the decision tree that creates branches at each node based on these values. Every leaf node of the decision tree contains a 0 or 1 that represents whether the path from the root to this particular leaf node results in an acceptable or unacceptable value of the output.

Each possible path of the decision tree contains a set of input variables, their respective value ranges and a state of the system represented by the value of the output. This can easily be formatted into an if-then rule that lays out a condition for the output of the system to be within, or outside of acceptable limits. For example, assume that Figure 5.2 represents a decision tree with three input variables. Each of these input variables has two categories and the output variables have two categories, 0 and 1, that represent the output being within or outside control limits.

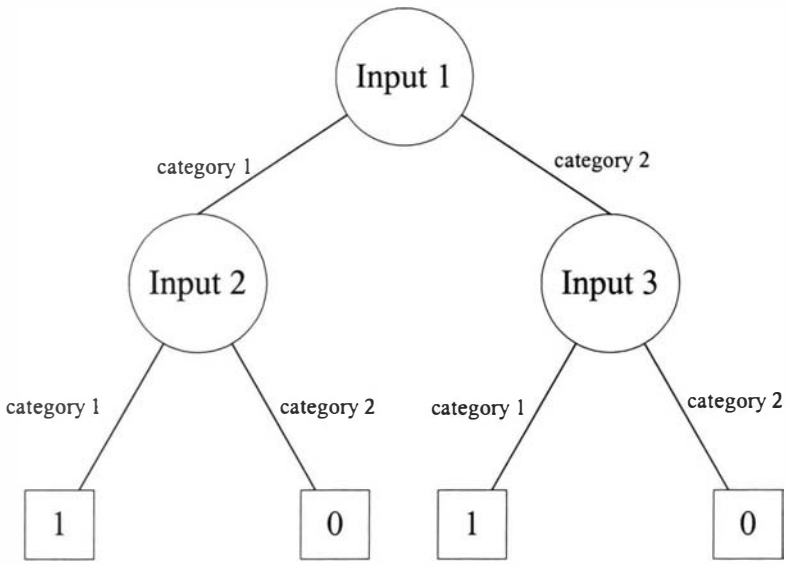

Figure 5.2 Example Decision Tree

The decision tree of figure 5.2 can be interpreted into rules based on the values shown. For example, the left-most path of the decision tree of figure 5.2 generates the following rule:

If Input 1 is in category 1 and Input 2 is in its category 1,

Then the output is out of specifications.

Rules are created for each possible path of the decision tree, from the root to every leaf node. All such rules can easily be created and stored in the system by the above path traversal procedure once the decision tree is loaded into memory. They can be retrieved to support the user by analyzing the process for error conditions that have occurred in certain output variables. Once the decision tree has been created, it is saved by the system so that it can be easily reloaded.

These data mining components model the existing relationships in the training data set and store it in formats that can be later used for predictive and interpretive functions. It is important to note that the knowledge gathered by these components is from only the information that is contained in the training data set. The trained system represents the environment to the extent that the training data set is representative of the environmental conditions. Also, this modeling is accurate with respect to the conditions that were placed on the data mining algorithms and the assumptions made about the process characteristics. These include parameters for acceptable control limits for output

variables and the relative magnitude of each category of the input variables. For this reason, if any of these environmental conditions or the system parameters were to change, the system must be retrained on data that represents the changed environmental conditions. Since training the data mining components is a computing-intensive process, it must be done off-line. During the time that the data mining components are being re-trained, the on-line system works with the models that are available until the system can be re-trained and the new models made available to the system.

5.6 Data processing component

The OLAP component of the integrated system is responsible for managing the data that is used by the data mining component. The available data consists of raw data from the manufacturing process collected at a continuous interval of time from the same part of the process. The data represents regions of relatively stable operation of the manufacturing process as well as data that has a comparatively greater number of errors in the manufacturing process. This data was collected over a period of time during which three different types of product were being produced. The process means and standard deviations for the stable manufacturing process that are used to create the normalized values for the raw data are also available.

The data component of the integrated system is implemented as a set of data objects that contain a collection of multi-dimensional arrays that include the raw process data, the normalized data and the names of all the variables. The data objects are instantiated from multiple files that contain the raw data collected from the manufacturing process and files

that contain the names of all variables. The files are loaded as part of the instantiation of the data objects and normalized data matrices are created from the raw data. Once the data objects are instantiated, they can interact with the data mining components. They provide normalized data, raw data and variable names used by the data mining components to develop and train models of the relationships in the manufacturing process. The data component of the integrated system contains behaviors that extract information from its components to provide multiple views of the data to the user.

5.7 On-line Analytical Processing

The OLAP-only approach can easily be implemented by extracting the set of variables that are outside of their specified control limits at the same time that any of the output variables are out of their specified control limits. This set of variables is extracted from comparing the normalized values of all inputs to the control limits. If a variable has a normalized value magnitude greater than its control limit then it is flagged as being out-of-control. This is achieved with a single pass through all input variables and all variables flagged as out-of-control can be presented to the user for analysis in the same way as described above for the integrated system. The presentation of summary information for analysis done by the integrated system is similar to that for the OLAP-only approach with the exception that the variables that are analyzed are different. We will describe the presentation of summary information by the OLAP-only system in the description of the OLAP component of the integrated system. Therefore, the primary difference between the two approaches is in the set of variables that are exposed for analysis.

This difference leads to the development of the validation procedure for the model. The model is validated by comparing the set of variables identified as causes of errors by each approach and their extent of conformity with what are believed to be the "true" causes of the errors, as identified by process experts. The next chapter explains the approaches to validation of the model in detail.

5.8 Summary

The components of the integrated system are implemented with interaction between the artificial intelligence components that comprise the data mining part of the integrated system and the data components which make up the OLAP components. The system is implemented using object-oriented methods with objects for the data mining and OLAP components. The purpose of the system is to model the relationships that exist in the data and extract. These relationships are then used by the OLAP component of the integrated system to provide support for analysis used for decision-making regarding the manufacturing process.

Chapter 6: Model Validation Approach

6.1 Introduction and Model Validation Approach

This chapter presents the validation of the integrated model by comparing the integrated approach, using a combination of data mining and OLAP, with an approach that uses OLAP-only to solve a process control problem. The results obtained by these approaches are compared with the results provided by experts of the manufacturing process under consideration. The model is validated by comparing the causes of errors identified by the integrated approach with those identified by using OLAP-only and those identified by experts in the manufacturing process.

The approach to validating the model is a three-step process:

i) Comparing the integrated approach and the OLAP-only approach,

ii) Presenting arguments as to why the integrated approach is more suitable to the problem, and

iii) A formal presentation of results that support the findings.

First, key dimensions of the process control problem that are important to gauge the effectiveness of any solution to the problem are extracted. The integrated system

approach and the OLAP-only approach are then compared based on their effectiveness to solve the process control problem. Their respective strengths and weaknesses on some key dimensions of the process control problem and on the overall system effectiveness are evaluated.

Secondly, arguments discussing why the integrated system is expected to perform better than the OLAP-only approach are presented. These arguments are grounded in the benefits of the two approaches as they apply to the process control problem under consideration. OLAP is not a standard approach to the solution of process control problems. It involves the automatic generation of queries to display data when certain variables are outside established parameters. The results of such queries on critical variables provide trends of the behavior of these variables to the users to support decision-making. Hence, the OLAP-only approach captures the essential functionality of the statistical process control approach. There is an underlying assumption that these critical variables are effective measures of process stability and that any unexplained variance in any one of these variables that is outside the pre-specified control limits is an indication of instability in the manufacturing process.

As a third step, some hypotheses that compare the results obtained by the integrated system, the OLAP-only approach and the view of the experts on the specific errors, are presented. These steps evaluate effectiveness of the integrated system and the OLAP-only approach in identifying the causes of errors in the process and providing explanations for

the causes of these errors. Figure 6.1 illustrates and summarizes the model validation approach.

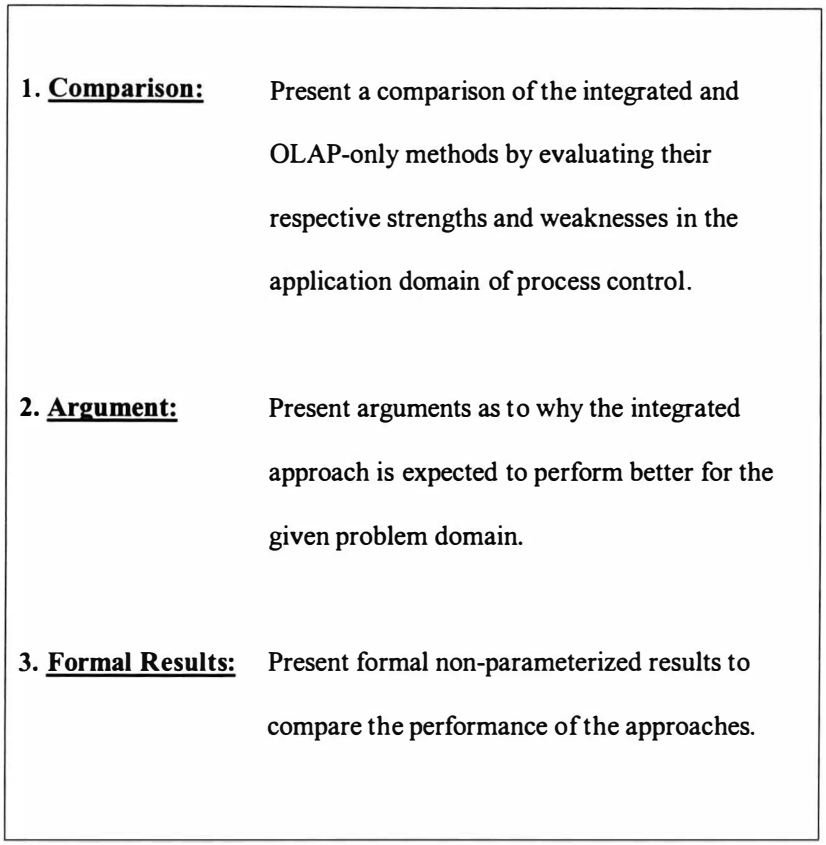

Figure 6.1 Model Validation Approach

6.2 Bases for Comparison of Integrated Approach vs. OLAP-only

The comparison between the two approaches is made on dimensions that are relevant in the assessment of the extent to which these approaches provide a viable and useful solution for the process control problem. These dimensions are:

i) The ability to detect and explain errors in the process,

ii) The flexibility and adaptivity of the approach with respect to changes in the product type and changes in standards for individual products and environmental conditions,

iii) Access to summary information about the product and process characteristics,

iv) The ability to predict errors in the outputs from examination of system inputs.

The following sections describe each of the above dimensions and their applicability to this comparison.

6.2.1 Detection and explanation of errors

A primary functional requirement of an effective approach to process control is that it should be able to detect errors in the process by examining the process data. This is operationalized by identifying critical process variables that are key indicators of process stability. These are based on the criterion that if these variables are outside of normal parameters, then the stability of the process is questionable, at best. These process experts identified two variables that satisfy this criterion for the data under consideration. For this

discussion, a process error is defined as a condition where one or both of these variables are outside their parameters of normal operation.

As previously stated, the purpose of the system is to identify error conditions as well as existing approaches and to provide causes for this error to provide better input to the decision-making process. The proposed system offers explanations of the causes of the error so the decision maker can decide on a requisite course of action to correct the error. These explanations are the primary contribution of the integrated approach to the existing state of the art in process control.

6.2.2 Flexibility and Adaptivity

The process control approach should be flexible enough to incorporate changes in the manufacturing environment. Changes in product types and their corresponding changes in process specifications are frequent occurrences in modern manufacturing processes given the prevalence of high levels of automation and flexible manufacturing environments. Industrial processes typically produce many types of products that are similar. Production of these products place different standards for production on the manufacturing process. The design of process control systems must take into consideration the changes in standards as production shifts from one product type to another. Process control systems should be flexible enough to accommodate changes in the values of the control limits as required by production changes from one type of product to another.

The stability of a process is measured by the degree of conformance of process characteristics to established standards. These standards are routinely revised due to changes in manufacturing technology or product characteristics. Changes in standards for process cause a change in the acceptable control limits of the process control system. A process control system must be able to adapt to changes in the operating environment.

Artificial intelligence-based systems adapt to their operating environment through training. This time-consuming and critical task needs to be done carefully so that the system parameters reflect the conditions of the environment. Training a system requires the selection of a set of model training data that contains both good and bad examples so that the system can learn the intricacies of each. A trained system can identify the different states of the environment that it models. Care needs to be taken not to over-train the system so that the generated model is extendable to other data sets. However, the training should be complete enough that a comprehensive set of relationships in the data is incorporated in the system, making it effective in identifying and explaining good and bad cases.

The effective modeling of environments whose characteristics change is a very challenging task. For an artificial intelligence-based system to be adaptive to changes in conditions of the environment, the system needs to be retrained. This includes changes due to change in the product type and change due to revisions of the established standards for production. Retraining of the system is done off-line whenever external changes to the

operating conditions of the environment occur. New models, which reflect the changes in the environment, are generated and the system adapts.

6.2.3 Access to summary information

At any point in time the system should be able to provide the user with summary information regarding the various process and product characteristics to support decision-making regarding the process. The system should be able to respond to queries for summary information as well as provide some process critical information on a regular basis as an indication of process stability at any given point in time. The presentation and content of this information should be done to facilitate making decisions regarding the process.

6.2.4 Prediction Capability

The ability of a system to accurately predict the conformance of the quality of the product to standards by examination of the process characteristics is a desirable feature of a process control system. Most current process control methods do not have the capability of predicting future product quality by examining current process characteristics. This prediction capability is different from looking at current process characteristics to indicate current product quality, which is the principle of statistical process control methods. Hence, this capability adds to the functionality of existing methods.

6.3 Comparison of Integrated Approach vs. OLAP-only

6.3.1 Detection and Explanation of errors:

OLAP-only:

Using an OLAP-only approach, the system will always offer an explanation for errors in the output. As stated earlier, an error is a set of observations in which at least one of the critical process variables is outside its acceptable range of operation. The OLAP-only approach can easily identify all the variables that are out of range while the output is out of range. Hence, the explanation that the OLAP-only approach offers is a set of input variables that are outside their range of specification while the output is out of its specified limits. At best, this approach offers information on out-of-range co-variance of the input variables with the output variables. This output is limited to the pre-specified limits on the input and output variables of the system. It is clear that the results provided by the OLAP-only approach suffer from the same limitations as those offered by traditional statistical process control. As with traditional statistical process control, there is no information on the causes of these errors or the relationships between the erring input and output variables.

Integrated Approach:

The ability of the integrated approach to detect and explain errors comes from the combination of the data mining and OLAP components in the integrated system. Recall that the OLAP component detects a process error by detecting a

condition where an output variable is outside its established range. The out-of-control output variable is used to traverse a decision tree to determine the variables that may explain the causes for the error. These variables are then used to retrieve summary information via the OLAP component to inform the user of the causes and their behaviors leading up to the detected error in the process. Hence, the integrated system detects the errors in the process, offers causes for these errors, and provides information related to these causes to allow the user to make an informed decision regarding the cause and subsequent correction of the error.

This approach relies on the availability of a path on the decision tree, which was created by sufficient training examples so that the tree is trained in this type of error. For this reason, the integrated approach may not be able to explain all instances of process errors without retraining the data mining based components of the integrated system. Specifically, the integrated approach will not be able to offer explanations for errors that are novel because they were not part of the training set and are new to the learning-based components of the integrated approach.

6.3.2 Flexibility and Adaptivity

OLAP-only:

Process control systems should be flexible enough to handle changes in the product type and be able to adapt to changes in the standards for any given

product. Flexibility and adaptivity in process control address the ease with which the underlying assumptions of the system can be modified so the system works with a new set of descriptions about the environment. In the OLAP-only approach these assumptions represent the control limits of both the input and output parameters. In order to change these assumptions, the system has to update these parameters. This process can be as easy as loading these parameters from a file. Therefore, using the OLAP-only approach, it is relatively easy to modify the underlying assumptions of the system.

For example, assume that ± one standard deviation in the output variable(s) is taken to denote a process that is in control. Upon revaluation of the process conditions, suppose that it is felt this should be changed to two standard deviations. When using OLAP-only, such a change would merely require a change in one of the parameters to a query. The same concept can be easily extrapolated to incorporate changes in the acceptable control limits for each variable in the system, including inputs and outputs of the system. Similar changes would be required for changes due to a change in product type in which the standards files for each product type are loaded by the system based on the product type being manufactured. Hence, changing the operating parameters of a system using OLAP-only is relatively easy, showing that the system is rather flexible.

Integrated Approach:

Changes in the environment imply that the inherent relationships in the environment have also changed. The data mining component of the integrated approach models these relationships in the environment. Therefore, if there are any changes in the environmental conditions, the data mining components need to be trained to incorporate the new relationships in the environment. If any of the underlying assumptions are altered, the system needs to be retrained. Changes in these assumptions may be due to a change in the product type or revisions in the standards for the product. In each case the models that are used by the data mining component of the integrated system will have to change.

To handle changes in the product type, the integrated system stores the trained model parameters for each type of product and loads the correct model based on knowledge of the current product type. For revisions of the standards for any given product, the data mining components of the integrated system must be retrained. For example, if the control limits are altered for a given product, then the system will need to be retrained so that these changes in the operating environment can be incorporated into the models maintained by the system.

The integrated system is flexible in that it can easily incorporate changes in the product type by loading the correct model from storage. The integrated system is comparatively less adaptive since the data mining components must be retrained in order for it to incorporate the changes in the environmental conditions. Training data must be identified and the data mining components of

the system must be trained off-line before the system can adapt to changes in the operating environment.

6.3.3 Access to Summary Information

OLAP-only:

Flexible and efficient access to summary information about multiple aspects of the environment is a major strength of the OLAP-only approach. This information is based on dimensions that must be supplied to the OLAP-only approach by the user. Given the dimensions along which the data is to be analyzed, the OLAP-only approach provides efficient access to historical and summary information.

Integrated Approach:

The Integrated approach incorporates all the benefits of the OLAP-only approach through the OLAP component of the integrated system. The primary difference between the two approaches is that the integrated approach generates the summary information based on the dimensions that are identified by the artificial intelligence-based components of the integrated approach. Hence, the integrated approach can provide efficient access to summary information on the dimensions that are identified by the data mining components of the integrated system, in addition to those identified by users' queries.

6.3.4 Prediction Capability:

OLAP-only:

The OLAP-only approach does not offer any predictive capabilities. There is no information with respect to what may happen to the outputs in the future based on the current values of the inputs of the system. Poor quality of the product is detected by the errors in the critical variables that are constantly monitored. If there is an error in these critical variables, then there is an error in the process, and hence, the product at that time.

Integrated Approach:

The artificial neural network component of the integrated approach has predictive capabilities to determine future values of outputs based on the current values of the inputs. The neural network can be trained on inputs in the present to predict outputs at a later point in time.

Summary

Table 6.1 summarizes the comparisons between the OLAP-only and the integrated approach.

	OLAP-only	Integrated System
Detection and Explanation of Errors.	Detects and provides information on causes of errors based on SPC model	Can detect errors in the system. Decision tree component provides explanation capability
Flexibility and Adaptivity of Approach	Flexible and adaptive approach to process control problem	Training is critical. Retraining is required when environmental conditions change
Access to Summary Information	Provides quick and efficient summary information on multiple dimensions of the environment.	Provides summary information on multiple dimensions based on rules in the decision tree.
Prediction of errors	No prediction capability.	Neural network component provides prediction capability.

Table 6.1 Comparison of features of OLAP-only and the Integrated System

6.4 Arguments for Integrated Approach vs. OLAP-only

Both the OLAP-only and the integrated approaches offer desirable capabilities for process control problems. OLAP offers efficient and flexible access to summary information about the system. It also offers flexibility in making modifications to the underlying assumption of the system. Some typical queries for an OLAP-based system include:

i) The retrieval and display of data for a variable over a particular time period,

ii) The retrieval and display of data for any set of variables over any time period,

iii) The retrieval and display of data over time periods where certain variables are outside of their control specifications,

iv) The retrieval and display of data for variables that are out of specifications at any given instance of time, and

v) Re-evaluating the overall state of the system when the underlying assumptions that decide when a variable is in or out-of-control are modified.

A primary strength of the OLAP-only approach is the efficiency with which it provides the user with effective access to summary information. Since there is no knowledge-based component in an OLAP-only approach, the content of this summary information has to be decided upon by the users. In other words, the user must decide on the queries to make to the OLAP system. Hence, the nature of the analysis has to be a predefined input to the analysis process.

A major drawback of using OLAP-only is that any explanation offered by the system is based on pre-specified assumptions of simultaneous independent variations in process variables. The explanations that such an approach offers are based on the whether a certain variable was within its pre-specified acceptable range of variation. Hence, at best, OLAP-only provides a list of variables that are outside their acceptable range at the same time that the output is out of range. Since there is no process knowledge involved in this approach, OLAP-only approach does not offer any indication of causality of the variations in the output due to the set of input variables that are outside specifications. It can only identify the co-variants of the output from the set of inputs at any instant of time. In highly auto-correlated processes, as the most continuous manufacturing processes are, it is common for numerous variables to be outside specifications simultaneously. Hence, the outputs of the OLAP query provide a large number of variables that are simultaneously outside normal parameters, as are the outputs. This provides little utility to the user searching for the cause of variations in processes' numerous variables.

OLAP is very efficient in extracting known information from the large volumes of data and providing summary information with multiple views of the information. Once the relationships in the data are known, using OLAP can significantly enhance analysis of the data based on these known relationships. The goal of the integrated approach is to augment this capability of OLAP with the knowledge that can be gained from mining existing data. Thus, this approach integrates the capabilities of OLAP with the knowledge

gained from mining the data to enable the knowledge-driven analysis of multidimensional data.

6.5 Formal Comparison of Results

Errors identified by process experts to be representative of common errors that occur frequently in the manufacturing process were used. The experts described the measures that were taken to correct these errors. Through these successful corrective measures, it is possible to identify the variations in the input variables that would explain why these errors took place, according to the experts. The data from the manufacturing process containing these errors was analyzed with both the OLAP-only approach and the integrated approach. The set of variables identified by each of these systems to be the cause of the errors was recorded for each error under consideration. The sets of variables identified by using OLAP-only as the causes of error will be compared with the set of variables identified by the integrated system. In both cases, the sets of variables identified by process experts form the basis of comparison. The schematic of figure 6.2 represents the comparisons to be made.

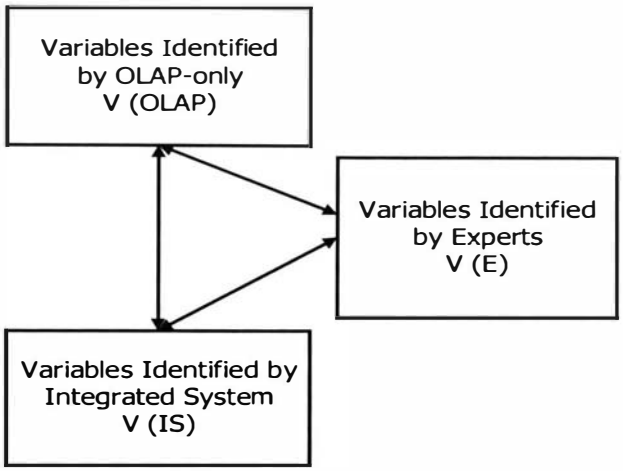

Figure 6.2 Comparisons of Results from the Three Possible Approaches

When an error occurs, three sets of variables can be identified:

V (E): The set of variables identified by the experts to be the cause of error;

V (OLAP): The set of variables identified by the OLAP-only approach to be the cause of error; and

V (IS): The set of variables identified by the integrated approach to be the cause of error.

For these comparisons, the set difference operation is used to compare the difference between the set of variables identified by the approaches. The set difference between two sets, x and y, is a set of variables that exist in x but not in y. For the comparisons, the set difference operator is defined as:

Let $x = \{A, B, C, D, E\}$ and $y = \{C, E, F, H\}$;

then $x - y = \{A, B, D\}$ and $y - x = \{F, H\}$.

Given the sets of variables defined above, the following comparisons can be made:

V (E) – V (OLAP): Denotes the variables that **experts have identified** as contributing to the errors, **but the OLAP-only approach does not**. This result represents **missing information** since these are variables that are

actual causes of error, as identified by the experts, but were missed by the OLAP-only system.

V (E) – V (IS): Denotes the variables that **experts have identified** as contributing to the errors, but the **Integrated System does not**. This result represents **missing information** since these are variables that are actual causes of error, as identified by the experts, but were missed by the integrated approach.

V (IS) – V (E): Denotes the variables that the **Integrated System has identified** as contributing to the errors, but the **experts do not**. This result represents **misleading information** provided by the integrated system.

V (OLAP) – V (E): Denotes the variables that the **OLAP-only approach has identified** as contributing to the errors, but the **experts do not**. This result represents **misleading information** identified by the OLAP-only system.

These comparisons show that two types of misinformation may occur as a result of the differences between the three sources of information about the causes of error in the

output. The result may contain missing information or contain misleading information.

Table 6.2 summarizes these comparisons.

	OLAP-only vs. Expert Opinion	**Integrated System vs. Expert Opinion**
Misleading Information	**V (OLAP) – V (E)** If this is non-null, it implies that **OLAP-only provides misleading information.**	**V (IS) – V (E)** If this is non-null, it implies that **IS provides misleading information**
Missing Information	**V (E) – V (OLAP)** If this is non-null, it implies that **Experts provide information that is missed by the OLAP-only approach**	**V (E) – V (IS)** If this is non-null, it implies that **Experts provide information that is missed by the Integrated approach**

Table 6.2 Comparisons of the Three Approaches and their Implications

The following hypotheses can be made to assert that the integrated system is more effective compared to the OLAP-only approach and is closer to expert opinion than the OLAP-only approach. This assertion is made on the assumption that expert opinion represents the true causes of errors in the outputs and provides the baseline.

Hypotheses:

H1: $\{V(OLAP) - V(E)\} = \Phi$

This hypothesis states that the set difference between the set of variables identified by the OLAP-only approach and the set of variables identified by the manufacturing process experts is the null set. This implies that OLAP-only also identifies the variables that are identified by manufacturing process experts to be the cause of errors. The set difference between the set of variables identified by the OLAP-only approach and the set of variables identified by the manufacturing process experts represents misleading information provided by the OLAP-only approach. If this hypothesis is true, then the OLAP-only approach does not offer any misleading information about the errors that occur in the manufacturing process.

H2: $\{V(IS) - V(E)\} = \Phi$

This hypothesis states that the set difference between the set of variables identified by the integrated system and the set of variables identified by the manufacturing process experts is the null set. This implies that the integrated system also identifies the variables that are identified by manufacturing process experts to be the cause of errors. The set difference between the set of variables identified by the integrated system and the set of variables identified by the manufacturing process experts represents misleading information provided by the integrated system. If this hypothesis is true, then this difference must be a null set, which implies that the integrated system does not offer any misleading information about the errors that occur in the manufacturing process.

H3: $\{V(E) - V(OLAP)\} = \Phi$

This hypothesis states that the set difference between the set of variables identified by the manufacturing process experts and the set of variables identified by OLAP-only is the null set. This implies that the manufacturing process experts also identify the variables that are identified by the OLAP-only approach to be the cause of errors in the manufacturing process. The set difference between the set of variables identified by the manufacturing process experts and the set of variables

identified by the OLAP-only approach represents missing information not identified by the OLAP-only approach. If this hypothesis is true, then this difference must be a null set, which implies that the manufacturing process experts do not offer any information that is missing from the explanations offered by the OLAP-only approach about the errors that occur in the manufacturing process.

H4: $\{V(E) - V(IS)\} = \Phi$

This hypothesis states that the set difference between the set of variables identified by the manufacturing process experts and the set of variables identified by integrated system approaches a null set. This implies that the manufacturing process experts also identify the variables that are identified by the integrated system approach to be the cause of errors in the manufacturing process. The set difference between the set of variables identified by the manufacturing process experts and the set of variables identified by the integrated system represents information about causes of error missing from the explanations offered by the integrated system. If this hypothesis is true, then this set must be a null set, which implies that manufacturing process experts do not offer any information that is missing from the explanations offered by the integrated system approach about the errors that occur in the manufacturing process.

6.6 Summary

A set of comparisons between the OLAP-only and the integrated system approach to the process control problem were presented. A set of arguments why we believe that the integrated approach of combining data mining and OLAP would be a more effective approach to the problem than the application of an OLAP-only approach are discussed. The procedure for a formal presentation of results from the comparison of the results obtained by the integrated system, the OLAP-only approach and by experts of the manufacturing process was outlined. Some hypotheses to test the effectiveness of the integrated approach to the real time process control problem were presented. The above hypotheses are tested on known cases of output errors for which expert opinion has been obtained. The next chapter presents these results and the tests of these hypotheses.

Chapter 7: Model Validation Results

7.1 Introduction

The data sets used for the validation of the model are described, and the characteristics of the data used for the training and verification of the models of the manufacturing process as part of the model validation procedure are discussed in this chapter. The training data sets are used by the data mining components of the integrated system to develop the models of the process that can predict and explain errors in the process. The verification data sets are used to verify the models that were developed. The results obtained for training the models on manufacturing process data are presented. The implications of these training results with respect to their usability and effectiveness are presented. Trained models of the manufacturing process are then exposed to the verification data sets to predict and explain the errors in the verification data. The results obtained from the exposing of the models to the verification data sets are presented and discussed. These results are then summarized and used to test the hypotheses developed in the last chapter. The results of testing the hypotheses are then summarized.

7.2 Description of Data Sets

The complete data set consists of ten thousand observations. Each observation is comprised of forty-two input variables and two output variables that represent the complete set of data collected from one part of a continuous manufacturing process. Output variables are identified by process experts to be critical measures of the stability of the part of the manufacturing process under consideration. The input variables are all variables that are collected from the part of the manufacturing process under consideration.

On examination of the output values, it is observed that the entire data set is divided into three distinct regions that significantly differ in their characteristics. The manufacturing process experts confirmed this fact. They explained that three different products were being manufactured in the time period. This fact is also reflected in distinct classifications in the input variables that also change at the same periods of time. Therefore, the complete data set is divided into three sets of observations. Figure 7.1 shows one output variable and its distinct beak-down into the three output regions. This output is divided into 3 separate regions which is shown in figures 7.2 (a), (b) and (c).

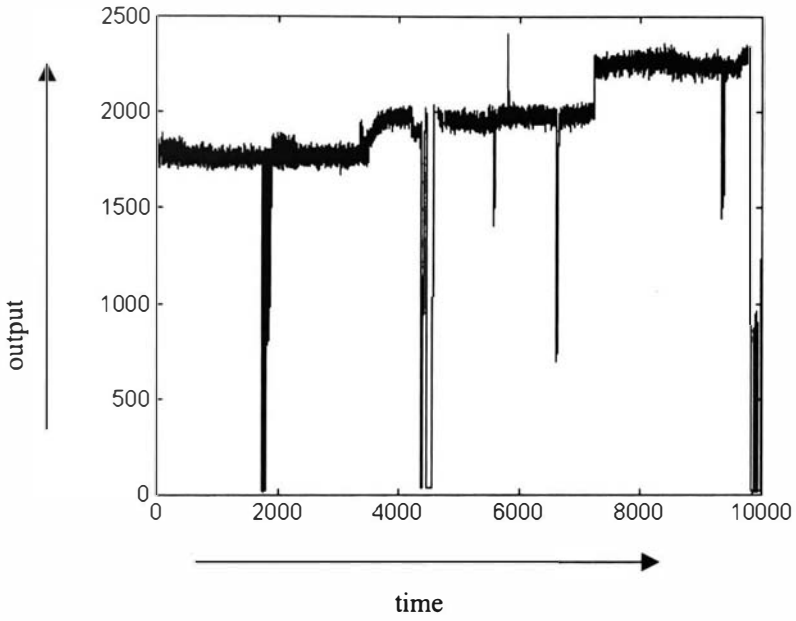

Figure 7.1 Plot of Output Showing the Distinct Breakdown into three Regions

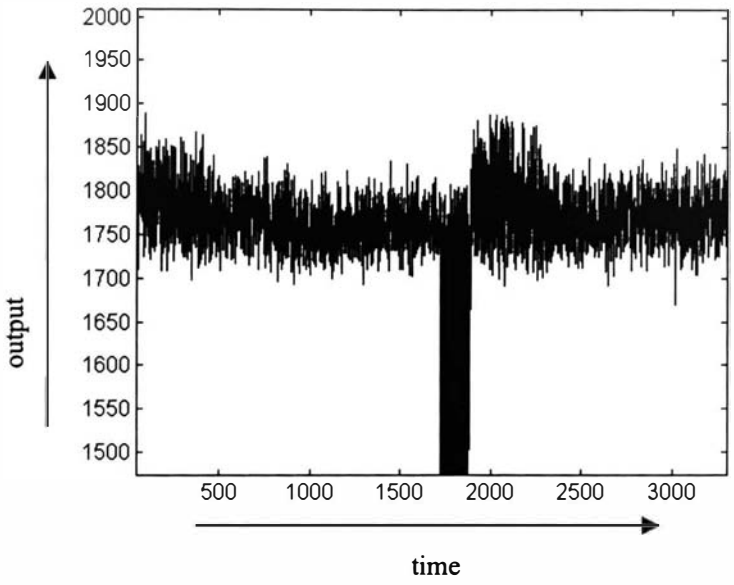

Figure 7.2(a) Output Region One

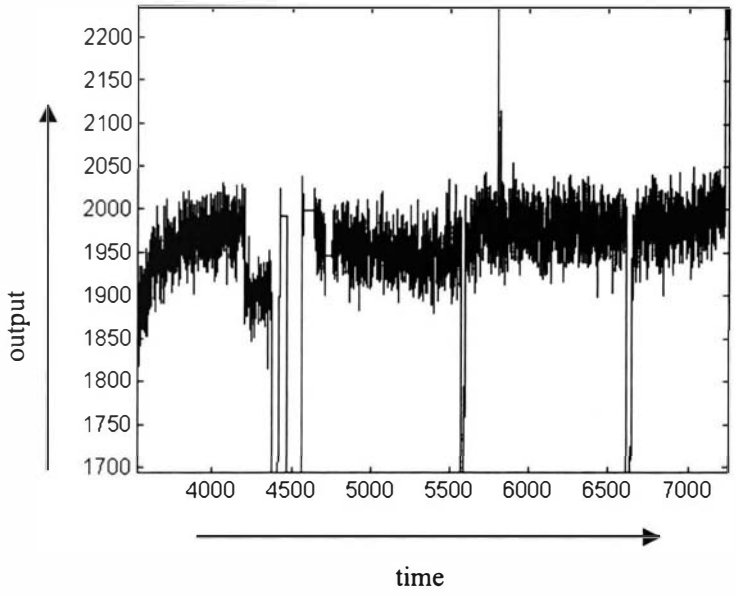

Figure 7.2(b) Output Region Two

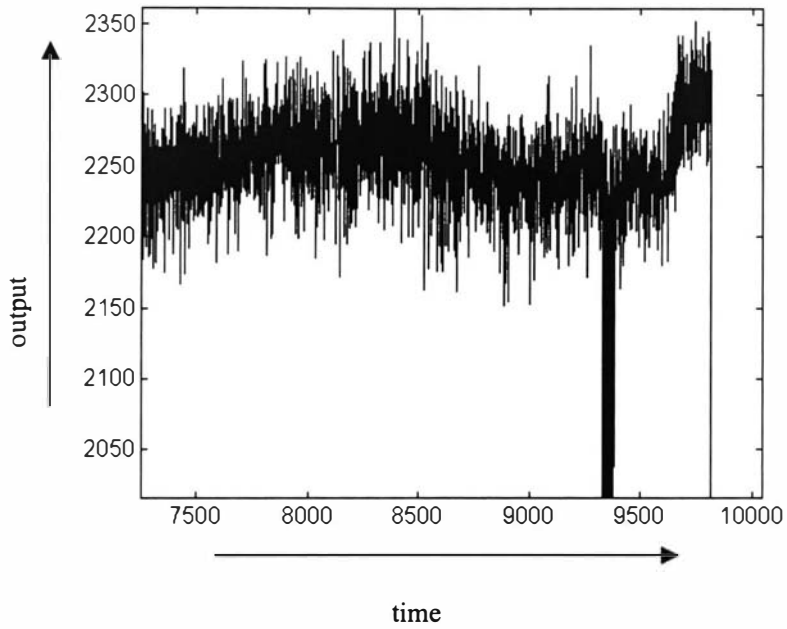

Figure 7.2(c) Output Region Three

This division of the data into three regions provides three sets of data for analysis. Each output region contains three thousand observations of the forty-two input variables and the two output variables. Some values were removed since they did not contribute in terms of normal operation or error conditions as noted by the process experts. The manufacturing process under consideration has automatic data collection equipment that takes measurements of various characteristics of the process at predefined regular intervals. These observations are taken once every minute and time stamped. The data is ordered sequentially by time stamp so that observation number 199 was taken exactly one minute before observation number two hundred. Each of the output regions has data that are collected in a continuous period of time. Each of these regions is further subdivided sequentially into a training data set that contains 2500 observations and a verification data set of five hundred observations.

For each output region, the data mining components of the integrated system are trained on the training set data. The result of this training procedure is a set of models that explain the behavior of the system for the type of data that it is supplied with. The trained models are then subjected to the verification data for that region, and results are generated for predictions and explanations generated by the integrated system. This analysis procedure is repeated for each of the three output regions. The ability to perform these repetitions of analysis provides a level of generalizability to the validation procedure across different product characteristics.

7.3 Procedure

For each of the three output regions described above, the data mining components of the integrated system are trained on the training set for that output region and the models for that region are obtained in the form of a trained neural network and decision trees for each output variable. The trained models are then applied to the verification data for that output region and the variables that are outside pre-specified control limits are identified. This set of variables is compared with the set of outputs obtained by application of the OLAP-only approach and the set of output variables identified by the manufacturing process experts. The entire set of results is then summarized. Figure 7.3 summarizes the steps for training and verification of the integrated system.

1. For each output region:

    1.1. Train the data mining components of the integrated system on the training data for that region;

    1.2. Verify by applying data mining models to the verification data to identify erroneous observations;

    1.3. Generate explanations from the data mining and OLAP-only components;

    1.4. Compare explanations generated from each approach and with those provided by manufacturing process experts;

2. Repeat for each output region.

Figure 7.3 Steps Involved in Procedure for Obtaining Results

7.4 Training

The system is trained on the training data set for each output region to capture the nuances of the manufacturing process data and develop models that capture the relationships in the data. The data mining component of the integrated system consists of a neural network and a decision tree component. Training involves the training of the neural network component of the integrated system using the back propagation algorithm until an adequate training error rate is achieved. The neural network is trained for each of the output regions identified above with the training data set for that region. For the decision tree component of the integrated system, training involves the development of the decision trees using the ID3 algorithm. Decision trees are also developed for each of the output regions with the training data set for that region. Each of these components is later verified using the trained data mining components for that region.

The neural network component of the integrated system is trained using the standard back propagation algorithm. The neural network is trained to the region just before it achieves this asymptotic value of training error. Stopping the training process well before this value is achieved decreases the accuracy of the neural network, while training beyond this region reduces the generalizability of the neural network. Learning rates typically range between five percent and ten percent depending on the nature of the data and the extent of random variations within them. The implementations of the neural network in the integrated system are trained using the back propagation algorithm with a learning rate of five percent until an acceptable level of error is achieved. The total sum of errors

that the neural network is trained to is ten percent. In other words, as soon as the error rate drops below ten percent the training is stopped, and the parameters of the neural network is saved for use in the verification phase of the experiment. The neural network typically achieves a satisfactory error rate in approximately 35,000 to 50,000 iterations after which the error rate takes an asymptotic value. Table 7.1 lists the parameters that were used for the training of the neural network component of the integrated system for each of the three output regions.

Neural Network Parameters	Values
Number of Levels	3
Number of Output Nodes	2
Number of Hidden Layer Nodes	15
Number of Input Nodes	42
Acceptable Error rate	10 %
Learning Rate	5 %

Table 7.1 Training Parameters for Neural Network

The decision tree component of the integrated system is trained for each output region based on the standard deviations of the training data. Training a decision tree involves the generation of "if-then" rules to classify each output value of the system based on the observed ranges of input values. The modified ID3 algorithm develops categories for each of the input and output variables and classifies output observations based on the observed categories of the input variables. Standard deviations are used to generate categories for the input variables. Using standard deviations to generate categories makes implementation efficient since the algorithm is applied to normalized data whose magnitude represents their standard deviations from the mean for that variable. A difference of three standard deviation units are used to classify the input variables, which is consistent with standard practice followed in manufacturing process control. Using this method, a variable with a value of three standard deviations or below will be classified into one category, while another with a value between four and six will be classified into another category. A stricter bound is used for the output variables where any magnitude less than two standard deviations is classified to be within specifications, while anything higher is considered an error condition.

These categories are used to create the branches of the decision tree. Every path of the decision tree represents a sequence of categories of input variables that lead to a classification of the output as being within or outside of acceptable limits. In addition to path information, each node of the decision tree also stores the number of examples in the training data that follow each of the possible branches disseminating from a node. The

number of examples along a path provides a measure of the strength of the path by counting the number of examples in the training data set that conform to the path of the decision tree.

The result of training the decision tree component of the integrated system is a set of trained decision trees from which a set of "if-then" rules for each output variable can be extracted. Decision trees categorize input variables into ranges of values. Each branch of a decision node is created based on the ranges of values, which are incorporated in the explanations offered by the integrated system. For example, an explanation by the decision tree component of the integrated system would be in the form:

Output A is out of range because

Input X is in the range x1: x2, and Input Y is less than y1.

Hence, the explanations offered by the integrated system offer richer content towards the support of making decisions than those offered by using the OLAP-only approach. The result of training the neural network is a trained neural network that can be used for prediction of output values from input values of the manufacturing process. These trained components are stored for verification on verification data for each of the output regions. The trained neural network and the trained decision tree with satisfactory training results are the output of the training process.

7.5 Results

After the system was trained satisfactorily, the trained system was exposed to the verification data set for each output region. Recall that the verification data set for each region consists of five hundred observations. Each observation was given to the trained neural network, which predicts output values for the given set of inputs based on the given set of inputs. Each error in the output was identified by the neural network component of the integrated system and was verified by the actual outputs for that observation in the verification set of output variables. The difference in the predicted value and actual values was calculated to measure the accuracy of prediction obtained by the neural network components of the integrated system. Table 7.2 presents the average, maximum and minimum prediction errors by the neural network component for the verification data set in each output region, expressed as absolute percentage values.

	Number of Incorrectly Predicted Observations	**Hit Rate**
Region 1	17	96.6 %
Region 2	41	91.8 %
Region 3	29	94.2 %

Table 7.2 Verification Results for Neural Network Component Expressed as Percentage Difference Between Predicted and Observed Values.

Each identified error observation was given to the decision tree and the OLAP components of the integrated system. For each output variable, the decision tree component of the integrated system followed the appropriate decision path in the trained decision tree and returned the set of variables along the path as the cause of the error. In addition to the set of variables, the decision tree also generated a range of values for each variable on its path. This set of variables and their respective value ranges were then formatted to generate natural language explanations for the cause. These natural language explanations formed the explanations generated by the integrated system.

The OLAP component of the integrated system was also exposed to the same set of error observations. Each observation was examined for conformance to specifications. If the outputs were out of specifications, then that observation was an error and the set of inputs for that observation were examined for errors. This procedure generated explanation for the error by the OLAP component based on the knowledge available to it. This knowledge was represented by predefined rules that identify an error condition as the condition when variable values are outside their specified limits. The OLAP component examined the observation and identifies the set of input variables that were outside of their specified limits for that observation. The result of this procedure was a set of variables for each error condition that was outside the pre-specified control limits defined by the mean $\pm$ 3 standard deviations. These variables were the explanations generated by the OLAP-only approach.

In summary, the integrated system and the OLAP-only approach were independently given the same verification data set for each of the output regions. The results obtained from these procedures are in the form of identified observations that are believed to be errors and a corresponding set of variables that explain the error identified by each approach. The verification data was also made available to the process experts, and their explanations for the same sets of errors are obtained. Hence, there are three independent sets of explanations for the errors in the manufacturing process to make comparisons with. The set of explanations identified by the manufacturing process experts was used as the basis for comparisons made between the explanations offered by the OLAP-only approach and those offered by the integrated system.

Errors in a manufacturing process usually exist for a few minutes. During this time the process stabilizes and the process parameters return to the normal conditions of operation due to corrective action taken by operators. Occasionally, the process re-establishes without any corrective action. In the data that used for this research, this time period for error typically spans multiple observations. In fact, typical errors span multiple observations, while single observation errors are usually incorrect readings or "spikes" that are essentially outliers that have a negligible effect on the quality of the product. Hence groups of error observations that occur in continuous blocks of time are of greater concern than individual errors that are occur in single observations.

Region One

Figure 7.4 shows one output variable used for verification for the first output region. It is clear from the figure that this is a relatively stable output variable in the verification data since it varies relatively close to the mean parameter of 531.7 units. The errors identified by the integrated system and the OLAP-only approaches for this region are shown in Table 7.3.

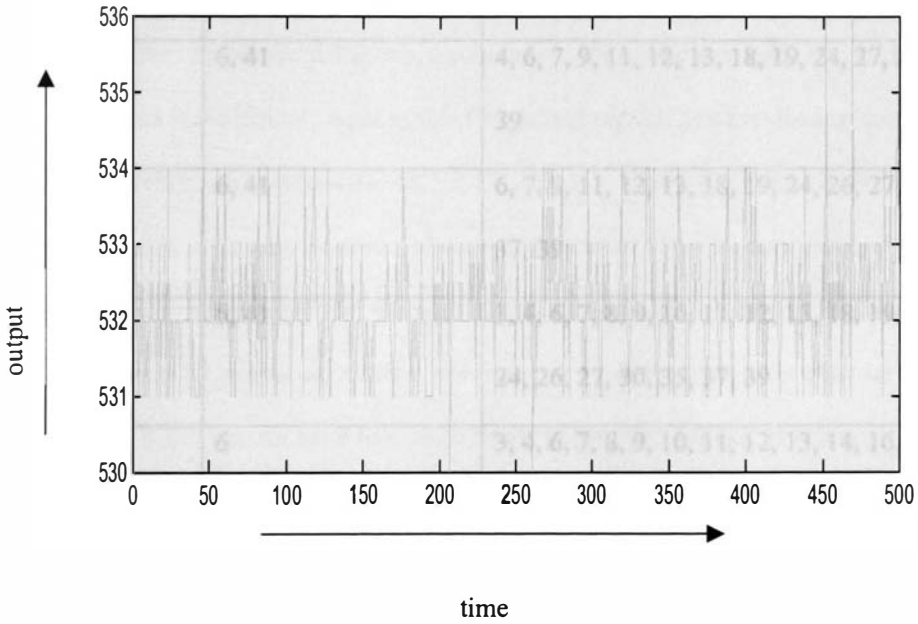

Figure 7.4 Verification Output for first output region

Observation Number	**Variables identified as causes by the Integrated approach**	**Variables identified as causes by OLAP-only**
23	6, 41	4, 6, 7, 9, 11, 12, 13, 18, 19, 24, 27, 30, 35, 37, 39
46	6, 41	6, 7, 9, 11, 12, 13, 18, 19, 24, 26, 27, 30, 35, 37, 39
55	6, 41	3, 4, 6, 7, 8, 9, 10, 11, 12, 13, 18, 19, 20, 22, 24, 26, 27, 30, 35, 37, 39
82	6	3, 4, 6, 7, 8, 9, 10, 11, 12, 13, 14, 16, 18, 19, 24, 26, 27, 30, 35, 37, 39
176	6, 41	1, 3, 4, 6, 7, 8, 11, 13, 14, 19, 24, 27, 35, 37
335 - 337	6	1, 4, 6, 7, 12, 13, 18, 19, 30, 35

Table 7.3 Verification Results for Integrated System and OLAP-only for First Region

The numbers listed in the first column of Table 7.3 are observation numbers, numbered from 1 to 499, that represent the time period at which the data was collected. The second and third columns represent the indices of the variables that are identified by the integrated system and by the OLAP-only as the causes of the errors in the observation numbers listed in the first column. Some observations that occur as errors in a group and have the same set of variables identified as causes by both the integrated system and the OLAP-only approach are grouped together as a set of errors as is true in the last row of the table above. As mentioned above, errors in manufacturing processes occur in a group; hence, this set is the primary error in this first output region. The previous errors that are identified earlier, observations 23, 46, 55, 82, 176, also have similar explanations offered by both systems with one exception, observation number 176. The error in observations 23 is referred to as error set one; the error in observation 46 is error set # two; the error in observations in 55 is error set # three; and the error in observation 82 is error set # four; observation # 176 is error set # five while errors in the observations 335 through 337 are grouped as error set # six. It is noteworthy that none of the error causes identified by the OLAP-only approach list variable 41. However, this variable is identified as a cause in multiple instances by the integrated approach.

Region Two

Figure 7.5 shows both the verification output values for region two. This region has the largest number of errors in all the verification data sets. It is clear from the figures below that that there are two distinct error regions in this set of output values. The errors in this region of verification outputs can easily be grouped together into the two sets that are the

two primary occurrences of errors in this region. Table 7.4 displays the set of variables that are identified by the integrated system and the OLAP-only approach.

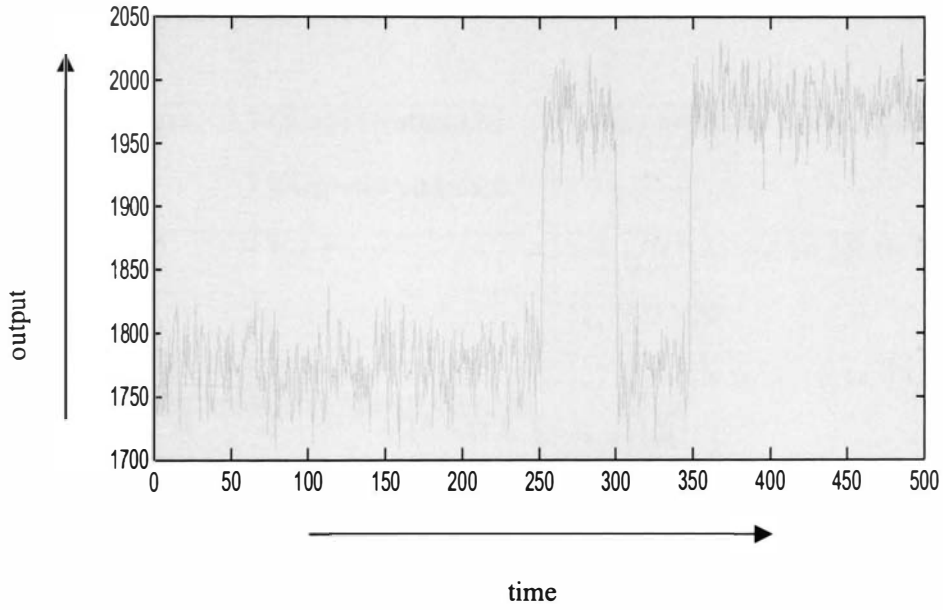

Figure 7.5(a) Verification Output Values for Second Region for Output one

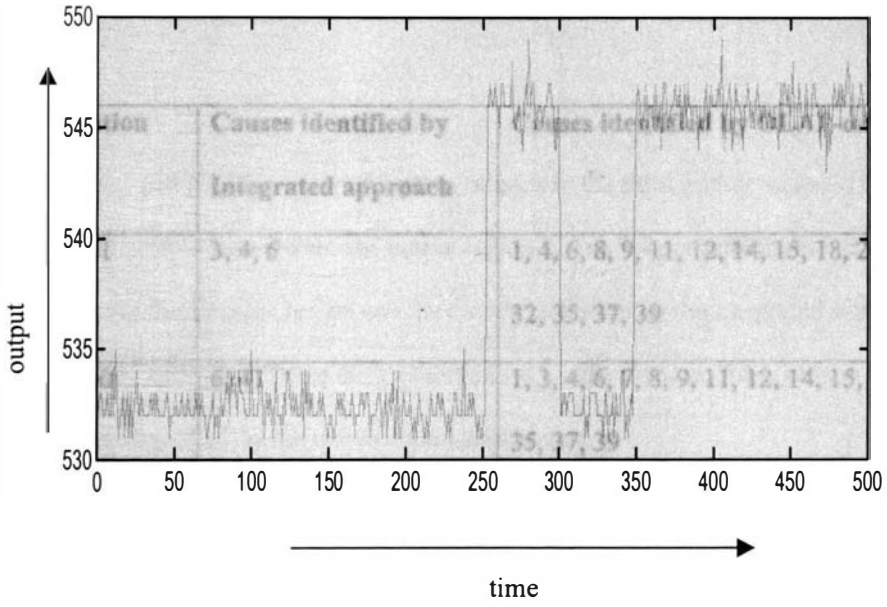

Figure 7.5(b) Verification Output Values for Second region for Output two

Observation Number	**Causes identified by Integrated approach**	**Causes identified by OLAP-only**
253 – 301	3, 4, 6	1, 4, 6, 8, 9, 11, 12, 14, 15, 18, 20, 30, 32, 35, 37, 39
351 – 500	6, 41	1, 3, 4, 6, 7, 8, 9, 11, 12, 14, 15, 18, 30, 35, 37, 39

Table 7.3 Verification Results for Integrated System and OLAP-only for Second Region

The integrated approach identifies input variables 3, 4 and 6 as the causes of the errors for the first group of errors and variables 6 and 41 as the set of causes for the second set of errors. Input variable 6 is common across all variables identified as a cause of the error. As in the first output region, the OLAP-only approach identifies a much larger set of variables as the causes of error in each of the set of errors. Input variable 41 is not indicated as a cause of error by the OLAP-only approach.

Region Three

Figures 7.6 (a) and (b) show the verification outputs in the third output region. This region is more stable than the second output region. Table 7.4 presents the errors that are observed in the third output region and the causes identified by the integrated system and the OLAP-only approach. These error observations are divided into three groups based on the causes identified by the two different approaches.

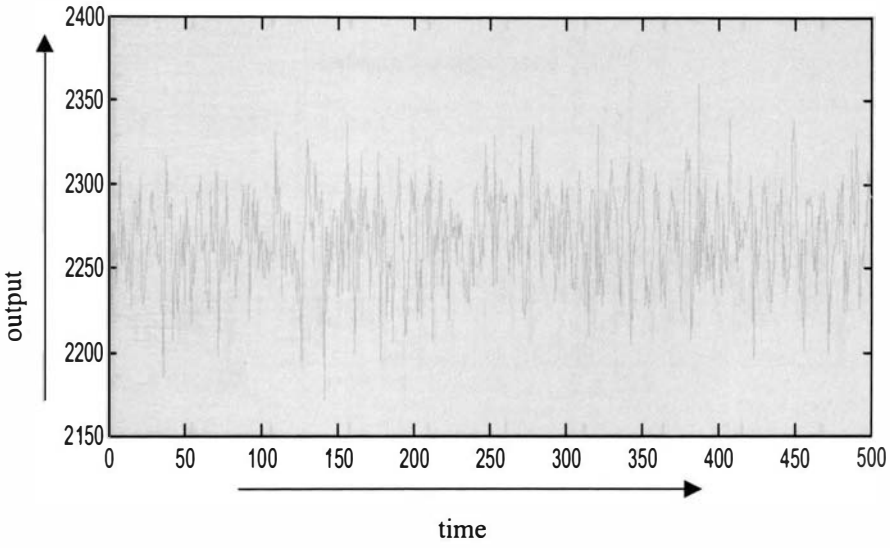

Figure 7.6(a) Verification Output Values for third region for Output 1

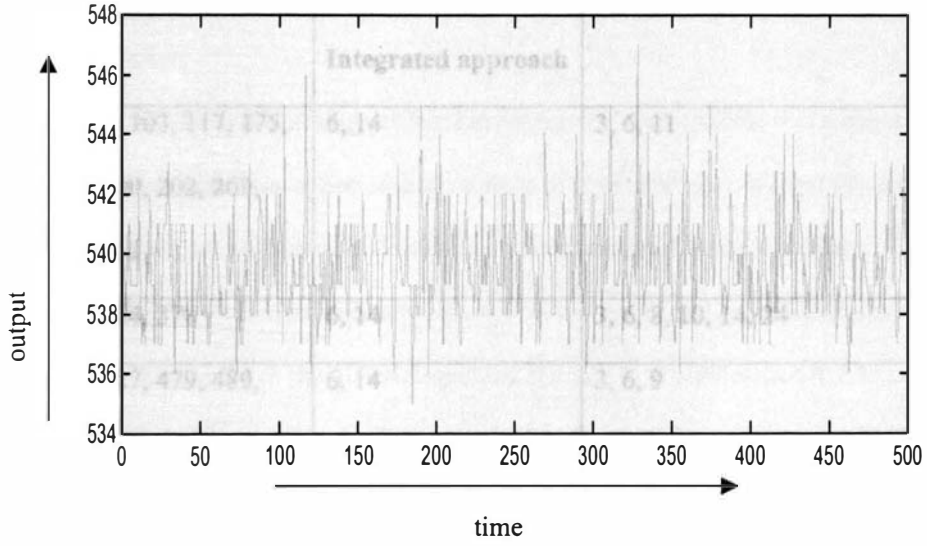

Figure 7.6(b) Verification Values for Third Region for Output 2

Observation Number	**Causes identified by Integrated approach**	**Causes identified by OLAP-only**
30, 81, 103, 117, 175, 190, 200, 202, 269, 289, 311,	6, 14	3, 6, 11
370, 374, 378	6, 14	3, 6, 8, 10, 14, 24
421, 427, 479, 489, 497	6, 14	3, 6, 9

Table 7.4 Verification Results for Integrated System and OLAP-only for Third Region

As shown in figure 7.11, the integrated approach offers variables 6 and 14 as the cause of the errors in this set of outputs. Again, the OLAP-only approach identified a larger number of variables as the causes for the errors. In addition, as in the above two output regions, variable 6 is a common cause of error that is identified by the two approaches. These groups of errors are labeled error sets 7, 8 and 9 in the summary presented later.

Expert Opinion

The experts in the manufacturing process have suggested that variable 6 is a major cause of error. This variable was collected from a piece of machinery that is regularly serviced as part of scheduled maintenance. As this machine component starts to perform inconsistently, there are some characteristic spikes that occur due to non-uniform temperature differences in the materials passing through it. It was learned that some of the machinery was serviced in the time period when the data from which variable 6 was collected. Variable 41 is the last piece of data collected for the part of the manufacturing process under consideration in this research. It is the experts' opinion that any disparities in the region of the manufacturing process from which variable 6 is collected, should also appear in one or more of the variables 35, 37, 39 or 41. Experts did not identify variable 14 as a cause of errors in the output in either output region.

7.6 Summary of Results

In the first output region, the integrated system identified variables 6 and 41 as the causes of error with variable 6 occurring as the cause of error in all cases of identified errors.

Input variable 6 is also identified as a cause of error for each error identified by the OLAP-only approach. It is clear from the above results that the set of variables identified by the OLAP-only approach for each error occurrence is much larger than the set that is identified by the integrated system. Based on the experts' analysis, this disparity in the size of the two sets implies misleading information was given by the OLAP-only approach.

Table 7.5 summarizes the errors identified by the integrated system and the explanations for these errors that were offered by the integrated system, the OLAP-only approach, and the manufacturing process experts. The errors are grouped by occurrence as explained in previous sections. This figure groups the errors together and enumerating the explanations that were offered by the process experts for the time periods represented by the verification data set.

Error Observation Sets	Variables Identified by Integrated System: V(IS)	Variables Identified by OLAP-only: V(OLAP)	Variables identified by Experts: V(E)
Error Set 1	6, 41	4, 6, 7, 9, 11, 12, 13, 18, 19, 24, 27, 30, 35, 37, 39	6, 35, 37, 39, 41
Error Set 2	6, 41	3, 4, 6, 7, 8, 9, 10, 11, 12, 13, 18, 19, 20, 22, 24, 26, 27, 30, 35, 37, 39	
Error Set 3	6, 41	1, 3, 4, 6, 7, 8, 11, 13, 14, 19, 24, 27, 35, 37	
Error Set 4	6	1, 4, 6, 7, 12, 13, 18, 19, 30, 35	
Error Set 5	3, 4, 6	1, 4, 6, 8, 9, 11, 12, 14, 15, 18, 20, 30, 32, 35, 37, 39	6, 35, 37, 39, 41
Error Set 6	6, 41	1, 3, 4, 6, 7, 8, 9, 11, 12, 14, 15, 18, 30, 35, 37, 39	
Error Set 7	6, 14	3, 6, 11	6, 35, 37, 39, 41
Error Set 8	6, 14	3, 6, 8, 10, 14, 24	
Error Set 9	6, 14	3, 6, 9	

Table 7.5 Summary of Verification Output Values

7.7 Hypotheses Testing

The last chapter presented hypotheses about the efficacy of the integrated system vis-à-vis the OLAP-only approach and the opinions of the process experts. The following sections describe the validity of these hypotheses to the data and discuss the implications of rejecting or failing to reject these hypotheses.

H1: $\{V(OLAP) - V(E)\} = \Phi$

Reject this hypothesis.

This hypothesis states that the set difference between the set of variables identified by the OLAP-only approach and the set of variables identified by the manufacturing process experts is the null set. Failure to reject this hypothesis would imply that the OLAP-only approach also identifies the set of variables identified by manufacturing process experts to be the cause of errors and does not offer any misleading information in the verification data about the errors that occur in the manufacturing process.

This hypothesis must be rejected based on the data presented above. For each output region and for all identified sets of error, the number of variables identified by the OLAP-only approach is greater than those identified by the manufacturing process experts. From figure 7.12, it is clear that this hypothesis was false across all sets of observations in the verification data sets. The OLAP-only approach

identifies more variables as causes of errors than the manufacturing process experts do. Rejecting this hypothesis implies that for the data under consideration, the OLAP-only approach offers misleading information compared to the manufacturing process experts.

H2: $\{V(IS) - V(E)\} = \Phi$

Fail to reject this hypothesis.

This hypothesis states that the set difference between the set of variables identified by the integrated system and the set of variables identified by the manufacturing process experts is the null set. Failure to reject this hypothesis would imply that the integrated system also identifies the set of variables identified by manufacturing process experts to be the cause of errors and does not offer misleading information in the verification data about the errors in the manufacturing process.

This hypothesis cannot be rejected based on the results in the table above. The set of variables identified by the integrated system to be the causes of error in the data sets considered include the variables that are identified by the manufacturing process experts. From the results presented in figure 7.12, it is clear that this hypothesis cannot be rejected for the results from the verification procedure. There is not sufficient evidence in the data to reject this hypothesis. The integrated system also identifies those variables as causes of errors that the

manufacturing process experts do. Failure to reject this hypothesis implies that for the data under consideration, the integrated system does not offer misleading explanations in the verification data about errors in the manufacturing process.

H3: $\{V(E) - V(OLAP)\} = \Phi$

Reject this hypothesis.

This hypothesis states that the set difference between the set of variables identified by the manufacturing process experts and the set of variables identified by OLAP-only is the null set. Failure to reject this hypothesis would imply that the manufacturing process experts also identify the variables that are identified by OLAP-only approach to be the cause of errors in the manufacturing process. Failure to reject this hypothesis would imply that the manufacturing process experts do not offer any information that is missing from the explanations offered by the OLAP-only approach about the errors in the verification data about the manufacturing process.

This hypothesis must be rejected based on the data presented above. In each output region for all identified sets of error, the number of variables identified by the OLAP-only approach is greater than those identified by the manufacturing process experts. Also, all the variables identified by the manufacturing process experts are not also identified by the OLAP-only approach. The OLAP-only approach provides explanations that are consistently missing some variables

identified by the manufacturing process experts to be the causes of error. From the results presented in Figure 7.12, it is clear that this hypothesis is false across all sets of observations in the verification data sets. There is sufficient evidence in the data to reject this hypothesis, therefore this hypothesis is rejected. The OLAP-only approach identifies more variables as causes of errors than the manufacturing process experts do. Rejecting this hypothesis implies that for the data under consideration, the OLAP-only approach offers information that is missing variables that have been identified by the manufacturing process experts as causes of errors in the verification data.

H4: $\{V(E) - V(IS)\} = \Phi$

Reject this hypothesis.

This hypothesis states that the set difference between the set of variables identified by the manufacturing process experts and the set of variables identified by integrated system approaches a null set. Failure to reject this hypothesis would imply that the manufacturing process experts also identify the variables that are identified by the integrated system approach to be the cause of errors in the manufacturing process. Accepting this hypothesis would imply that the manufacturing process experts do not offer any information that is missing from the explanations already offered by the integrated system about the errors in the manufacturing process that occur in the verification data.

This hypothesis must be rejected based on the results presented in figure 7.12. In each output region, for all identified sets of error, the set of variables identified by the integrated system approach is consistently different from those identified by the manufacturing process experts. The variables identified by the manufacturing process experts are not also identified by the integrated system. The integrated system approach provides explanations that are consistently missing some of the variables identified by the manufacturing process experts to be the causes of error. From the summary table presented in figure 7.12, it is clear that this hypothesis is false across all sets of observations in the verification data sets. There is sufficient evidence in the data to reject this hypothesis. The integrated system approach fails to identify all the variables that are identified by the manufacturing process experts as causes of errors in the verification data. Rejecting this hypothesis implies that for the data under consideration, the integrated system approach offers information that is missing variables that have been identified by the manufacturing process experts as causes of the errors in the in the verification data.

7.8 Summary

In summary, hypotheses H1, H3, and H4 are not supported by the data and must be rejected while hypothesis H2 cannot be rejected based on the results obtained. Table 7.6 summarizes the results of the hypothesis testing and its implications.

Hypotheses	Result	Implication
H1: {V (OLAP) – V (E)} = Φ	Reject	OLAP-only approach offers misleading information.
H2: {V (IS) – V (E)} = Φ	Fail to reject	Integrated system does not offer misleading information.
H3: {V (E) – V (OLAP)} = Φ	Reject	OLAP-only approach misses Information provided information provided by experts
H4: {V (E) – V (IS)} = Φ	Reject	Integrated System misses information provided by experts

Table 7.6 Summary of Hypotheses Testing Results

Rejecting hypothesis H1 implies that for the data under consideration, the OLAP-only approach offers misleading information as compared to the manufacturing process experts. Failure to reject hypothesis H2 implies that for the data under consideration, the integrated system does not offer any misleading explanations about the errors in the manufacturing process. Rejecting hypothesis H3 implies that for the data under consideration, the OLAP-only approach offers information that is missing variables that have been identified by the manufacturing process experts as causes of the errors in manufacturing process. Rejecting hypothesis H4 implies that for the data under consideration, the integrated system approach offers information that is missing variables that have been identified by the manufacturing process experts as causes of the errors in manufacturing process.

From these results, it can be inferred that the OLAP-only approach provides information that is misleading by identifying variables as causes of errors that are not verified by the process experts. The integrated system does not offer any misleading information as it identifies variables that manufacturing process experts believe to be the causes of errors in the manufacturing process for the verification data set. The OLAP-only approach and the integrated system offer information that is missing some of the variables that have been identified by the manufacturing process experts to be causes of errors in the verification data.

Chapter 8: Conclusions

8.1 Conclusions

An integrated system consisting of data mining and OLAP components to support intelligent decision-making was presented in the context of a real time process control problem. The proposed integrated system uses data mining to discover the complex relationships hidden in large volumes of manufacturing process data to classify error conditions in the manufacturing process. These relationships are discovered from real manufacturing process data using an artificial neural network component for prediction of future states of the environment and a decision tree component to offer explanations for states of the environment. The knowledge in these models can be used for making process control decisions for the manufacturing process. This data is organized and presented for the decision maker using OLAP to support multidimensional views of the data. These multidimensional views are created by the results from the models discovered by mining the data from the manufacturing process under consideration. An evolutionary approach is suggested in which these models can be constantly updated whenever there are any changes in the environment that may cause changes in the relationships modeled by the system.

The integrated approach can be used to analyze incoming real-time data to predict, identify, and explain possible error conditions in the process. As an improvement on existing approaches, this approach offers explanatory and predictive capabilities based on accurate and adaptive models of the process and offers early warning of imminent failures. Once an error occurs, the system identifies this by comparing the incoming data with the models of the process. If the error is confirmed, then the current parameters of the process are used to generate explanations for why the error has occurred. These explanations are provided to the user in the form of easy to understand if-then rules with information on current values of the system parameters. The system identifies the process variables and reports values that are causes of error. This information is intended to be a set of alternatives with which the user can investigate in the physical process in order to solve problems with the current manufacturing process.

The proposed solution relies on the integration of data mining and OLAP to build accurate and dynamic models of the process and to provide analytical views of the data that support decision-making in this environment. The solution is tested by comparing the results obtained by the proposed system with those obtained from using an OLAP-only approach. Both these results are validated using opinions given by manufacturing process experts as the basis of comparison. Results show that the integrated approach is able to identify and explain errors in the process data. It also offers explanations that provide information for decision-making about the environment. The integrated approach offers content rich explanations about the nature of the errors and their causes. The integrated approach also provides additional information about these causes of error using decision

tree models that supply information about the output variable in question and the input values associated with the output. These explanations take the form of natural language explanations of the output variables' states due to values of the inputs. These explanations can also take the form of queries used to materialize multi-dimensional views of the data from actual operation of the system. This information is knowledge based, multi-dimensional and concerns the operations of the system under consideration. Therefore the integrate system can provide valuable information to support the decision-making process.

Comparing the explanations of errors offered by the integrated system, the OLAP-only approach, and opinions of manufacturing process experts, validates the system. Each approach is exposed to the same set of data, and the explanations offered by each approach are compared. These explanations are offered in the form of variables that are identified by the integrated system as the causes for error. The data reveals that the OLAP-only approach offers explanations that are misleading since they contain variables that are not identified by the experts as causes of the errors. It is also found that the explanations offered by the OLAP-only approach misses some of the variables identified by the process experts as causes of error. The integrated system approach does not identify any misleading information about the errors in the manufacturing process data. However, the integrated system approach also misses some of the information that is identified by the process experts as causes of error in the process data. Though neither system identically matches the variables identified by the process experts to the causes of

the errors under consideration, the explanations offered by the integrated system are more concise and consistently match more closely with those offered by the process experts.

Results from the integrated system also differ from those of the OLAP-only approach since they provide information about the ranges of values for the variables that form the explanations for each identified error. Hence, the information provided by the integrated system is richer in content, as it does more than merely identify variables that are believed to be the cause of the error. Decision trees categorize input variables into ranges of values. Each branch of a decision node is created based on the values, of these ranges of values and these are incorporated in the explanations offered by the integrated system. The explanations offered by the integrated system offer richer content towards the support of making decisions than those offered by using the OLAP-only approach. These explanations can be automatically structured by the integrated system to generate analytical views of the data that can support analysis of the errors. The explanations from the integrated system are generated based on sophisticated models of the environment and can support intelligent decision-making about the environment for which they are trained.

8.2 Limitations

The information from the integrated system is a set of variables and their values that identify the cause of errors in the manufacturing process. The result of this system is not a set of actions that the user needs to perform in order to correct any problems that are identified by the system. The system provides causes for the errors and explains these

causes and the circumstances under which they occur. It does not suggest any remedial action to correct these errors. Hence, users need a level of sophistication to take the information supplied by this system and translate it into components of the physical system that need to be investigated or adjusted. With the recent advances in knowledge-engineering and multimedia databases, it is easily conceivable that modules can be added to translate variables into physical components of machinery and characteristics of the physical system. These modules can be presented to the user as creative graphical user interfaces to give the typical users of process control systems, such as production line operators that are not trained in process control techniques, the ability to visualize errors in the manufacturing process and use sophisticated techniques to troubleshoot problems. Such research would greatly enhance the usability of systems built on the principles developed in this research and advance the state of the art in real time process control systems.

Some variables that are part of the set of inputs to the system directly translate in to a piece of machinery that can be investigated for malfunction. In many cases however single pieces of machinery provide multiple input variables that often co-vary. The selection of the individual input and output variables existed prior to this research and were taken as a given for the creation of the models by data mining techniques. It is not possible to determine from this research whether the selection of alternative variables as inputs or outputs to the system would improve the efficacy of the models and, hence, the system.

This research treats the set of output variables identified by the process experts as a direct measure of process stability and, hence, measures of the quality of the overall production process. These variables are critical measures of the stability of the part of the manufacturing process under consideration; however, they are not measures of the quality of the final product. The variables treated as outputs in this research are provided by the experts as variables that are believed to be critical to the stability of the manufacturing process. Specifically, these variables were identified by process experts to be critical measures of the stability of the sub-process that is considered in this research. These associations were established in internal research done by the manufacturing process experts and are treated as valid associations for this research. Any changes in the validity of these associations will require retraining of the system developed in this research to accommodate for these changes. Future research may look at developing models for the overall process where measures of overall product quality are used as outputs. Research on the development of comprehensive process models can be done in addition to developing models for critical components of the process, as done in this research.

8.3 Future Research

To provide a real world problem context, this research considers a large continuous manufacturing process typical of many chemical process industries and other heavily automated manufacturing environments. Such environments typically have enormous operational data repositories that contain data collected from various parts of the manufacturing process at regular time intervals. Data collected from everyday operations of the production system contains a wealth of information about the processes of the

system. However, the raw data itself does not generate any direct benefits. The raw data needs to be analyzed to develop descriptive models that can be used to understand, explain, and predict imminent failures and errors in the manufacturing process. Models are required to provide answers when errors occur in the manufacturing process to provide insight into the causes of errors and provide direction and understanding to decision-making requirements for the problem context. Data mining extracts novel and ultimately comprehensible knowledge useful for making crucial business decisions and has been successfully applied in a large number of systems and in many diverse application areas. In manufacturing environments, data mining can unearth novel patterns useful to predict future trends and behaviors of systems and, in turn, enables proactive and knowledge-driven decision-making. This research provides an approach to model the dynamic relationships in the data so that they can be used to make decisions about correcting errors that have occurred or are about to occur in the process. More research into generating the appropriate type of models and their applicability to the problem domain will enhance the current state of the art in making decisions about process control and quality control problems in manufacturing processes.

The manufacturing environment can be characterized as a dynamic process that is rich in terms of the volume of data available and often requires making decisions for standard, repetitive, and novel, problems that arise in the environment. This research concentrates on one type of manufacturing process. More research in the applicability of these models to other types of manufacturing processes will enhance the generalizability of this model to industrial processes.

The model for integration of data mining and OLAP to support intelligent decision-making developed by this research can help the decision-making process by providing a set of models to explain the different states of the system. The proposed model provides a means for knowledge driven analysis of large volumes of data by combining methods to develop analytical models of data with means for analysis of large volumes of multi-dimensional data at multiple levels of abstraction as the decision problem requires. This approach needs to be tested on other environments and problem contexts in order to address the issue of generalizability of the approach. Sufficient data needs to exist for the data mining models to be developed to make this approach applicable to a problem domain. This requirement is necessary for the opportunity to discover complex and heretofore unknown relationships in the data that may be potentially useful for making decisions in the problem domain. Many business environments, such as financial markets, credit analysis, marketing analysis, and banking share these characteristics. Research has been done on the applicability of data mining and OLAP in these areas with reasonable amounts of success. This research focuses on off-line data mining of the environment to develop explanatory and predictive models of the environment to provide appropriate multidimensional views of the data. Little has been done to develop methods to integrate data mining and OLAP to provide a systematic method for decision-making that allows users to examine multiple views of the data that are generated using knowledge about the environment and the decision problem. Further research in the applicability of systems that focus on technologies to support intelligent decision-making can advance the state of the art in the areas of data mining, OLAP research and intelligent decision-making.

References

1. Affisco, J. F., and Chandra, M., Quality assurance and expert systems – A framework and conceptual model, *Expert Systems with Applications.* 1990.

2. Alavi, M., and Henderson, J. C., An Evolutionary Strategy for implementing a decision support system. *Management Science*, Vol. 27, No. 11, 1981.

3. Albus, J. S. Outline for a Theory of Intelligence. *IEEE Transactions on Systems, Man, and Cybernetics.* Vol. 21, No. 3, May/June 1991.

4. Alexander, S. M., The application of expert systems to manufacturing process control, *Computers and Industrial Engineering,* Vol. 10, No.3, 1987.

5. Alter, S., *Decision Support Systems: Current Practice and Continuing Challenges*, Reading, Mass., Addison-Wesley, 1980.

6. Anthony, Robert, *Planning and Control Systems: A Framework for Analysis*, Division of Research, Graduate School of Business Administration, Harvard University, Boston, 1965.

7. Aström, K. J., Wittenmark, B., *Computer Controlled Systems*, Prentice Hall, London 1989.

8. Badavas, P. C., *Real-Tme Statistical Process Control.* Prentice-Hall, NJ., 1993.

9. Bellman, R. E., *An Introduction to Artificial Intelligence: Can Computers Think?*, Boyd & Fraser Publishing Company, San Francisco., 1978.

10. Bennett, B. S., *Simulation Fundamentals*, Prentice-Hall, 1995.

11. Bennett, John, L., *Building Decision Support Systems*, Reading, Mass., Addison-Wesley, 1983.

12. Calabrese, C., Gnerre, E., and Fratesi, E., An expert system for quality assurance based on neural networks. *Parallel Architectures and Neural Networks, 4th Workshop, International Institute for Advanced Scientific Studies, 1991.*

13. Dagli, C. H., and Stacey, R., A prototype expert system for selecting control charts. *International Journal of Production Research*, Vol. 26, No. 5, 1988.

14. Dagli, C. H., Intelligent Manufacturing Systems, in Dagli, C. H., (editor) *Artificial Neural Networks for Intelligent Manufacturing,* Chapman & Hall, 1994.

15. Deming, W. E., *Statistical Adjustment of Data*, Massachusetts Institute of Technology, Center for Advanced Engineering Study, 1986.

16. Dhar, V. On the Plausibility and Scope of Expert Systems in Management. *Journal of Management Information Systems*, Vol. 4, No. 1, Summer 1987.

17. Elkins, S. B., Open – OLAP. *DBMS*, April 1998.

18. Fayyad, U. M., Haussler, D., Stolorz, P., Mining Scientific Data, Communications of the ACM, Vol. 39, No. 11.

19. Fayyad, U. M., Piatetsky-Shapiro, G. and Smyth, P., From Data Mining to Knowledge Discovery: An Overview. In Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P., and Uthurusamy, R. (editors) *Advances in Knowledge Discovery and Data Mining*, AAAI Press/ The MIT Press. Menlo Park California, 1996.

20. Fayyad, U. M., Piatetsky-Shapiro, G. and Smyth, P., The KDD Process for extracting Useful Knowledge from Volumes of Data. *Communications of the ACM*. Vol. 39, No. 11, November 1996.

21. Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P., The KDD Process for Extracting Useful Knowledge from Volumes of Data, Communications of the ACM, Vol. 39, No. 11.

22. Feigenbaum, A. V., *Total Quality Control*, 3rd Edition, McGraw-Hill, New York, USA., 1991.

23. Feigenbaum, E., McCorduck, P. and Nii, P. *The rise of the expert company.* Times Books, NY. 1988.

24. Fielden, G. D. R., *Engineering Design*, 1975.

25. Goldstein, I. and Papert, S. Artificial Intelligence, language, and the study of knowledge. *Cognitive Science*, 1 1977.

26. Gorry, Anthony G., and Scott Morton M. S., *A Framework for Management Information Systems*, Sloan Management Review, 1971.

27. Goul. M., Henderson, J. C., and Tonge F. M., The Emergence of Artificial Intelligence as a reference Discipline for Decision Support Systems Research., *Decision Sciences*, Vol. 23, No. 6. November/December 1992.

28. Grega, W. Integrated environment for real-time control and simulation. *Computers in Industry* Vol. 31. 1996.

29. J. Han, `` OLAP Mining: An Integration of OLAP with Data Mining", *Proc. 1997 IFIP Conference on Data Semantics (DS-7)*, Leysin, Switzerland, Oct. 1997, pp. 1-11

30. Ho-Sang Ham, Seok-Chan Jeong, Young-Hui Kim, Real-time shop floor control for a PCB Auto-Insertion line based on Object-Oriented approach: *Computers In Industrial Engineering*., Vol. 30, No. 3, pp. 543-555, 1996.

31. Hotelling, H. "Multivariate Quality Control." In C. Eisenhart, M. W. Hastay, and W. A. Wallis, eds. *Techniques of Statistical Analysis.* New York: McGraw-Hill Book Company, 1947, 111-184.

32. Hurst, E. G., Jr., et. al. Growing DSS: A flexible evolutionary approach. In Bennett, J. L., Ed. *Building Decision Support Systems.* Reading Mass. : Addison-Wesley, 1983, 133-172.

33. Jackson, J. E. "Quality Control Methods for Several Related Variables." *Technometrics*, 1, 359-377, 1959.

34. Jackson, J. E., and Morris, R. H. "An Application of Multivariate Quality Control to Photographic Processing." *Journal of the American Statistical Association*, 52(278), 186-189 June 1959.

35. Keen , P.G.W., and Scott Morton, M. S., *Decision Support Systems: An Organizational Perspective*, Reading, Mass., Addison-Wesley, 1978.

36. Kratzer, K. P., Artificial Intelligence Techniques in Real-Time Processing, in Schiebe, M., and Pferrer, S. (editors) *Real-Time Systems Engineering and Applications*, Kluwer Academic Publishers, 1992.

37. Kohavi, R., Sommerfield, D., Dougherty, J., Data Mining using MLC++: A Machine Learning Library in C++, Tools with Artificial Intelligence, 1996.

38. Kourti and McGregor, Mutlivariate SPC Methods for Process and Product Monitoring. *Journal of Quality technology,* 28, pp. 409-428, 1996

39. Liang, T. and Jones, C. V., Design of a Self-Evolving Decision Support System. *Journal of Management Information Systems*. Vol. 4., No. 1, Summer 1987.

40. Madey, G. R., Weinroth, J., and Shah, V., Hybrid intelligent systems: Tools for decision-making in intelligent manufacturing, in Dagli, C. H., (editor) *Artificial Neural Networks for Intelligent Manufacturing,* Chapman & Hall, 1994.

41. March, J. G. and Simon, H. A, *Organizations.* John Wiley and Sons, 1958.

42. McCulloch, W. S., and Pitts, W. A Logical Calculus of the Ideas Immanent in Neural Nets, *Bulletin on the Mathematical Biophysics,* Vol. 5, pp. 115 - 137. 1943.

43. Michalski, R. S., Carbonell, J. C. and Mitchell, T. M. *Machine Learning: An Artificial Intelligence approach* Vol. I and Vol. II. Morgan Kaufmann Publishers, Inc., Los Altos, CA, 1983.

44. Minsky, M. and Papert, S., *Perceptrons: An Introduction to Computational Geomentry.* MIT Press, Cambridge, MA 1969. 3rd edition 1988.

45. Mitchell, T. M., *Machine Learning.* McGraw-Hill, 1997.

46. Montgomery, D. C., and Wadsworth, H. M. "Some Techniques for Multivariate Control Chart Applications." *American Society for Quality Control Annual Technical Conference Transactions*, 427-435, 1972.

47. Montgomery, D. C., *Introduction to Statistical Proess Control*, 2nd Edition. John Wiley and Sons, 1991.

48. Montgomery, D.C., Keats, J. B., Runger G. C., and Messina, W. S., Integrating Statistical Process Control and Engineering Process Control. *Journal Of Quality Technology*, Vol. 26. No. 2, 1994.

49. Murray, T. J. and Tanniru, M. R., A Framework for Selecting between Knowledge-based and Traditional Systems Design. *Journal of Management Information Systems*, Summer 1987, Vol. 4, No. 1, pp. 42 – 58.

50. Narendra, K. S., Adaptive Control using Neural Networks. in *Neural Networks for Control.* Miller, W. T., Sutton, R. S., and Werbos, P. J. (editors). MIT Press, Cambridge, Mass. 1990.

51. Newell, A. and Simon, H. A., GPS, a program that simulates human thought. In Billing, H., editor, *Lernede Automaten,* pp. 109-124, R. Oldenburg, Munich, Germany. Reprinted in Feigenbaum and Feldman (1963).

52. Newell, A. and Simon, H. A., *Human Problem Solving*, Prentice Hall, Englewood Cliffs, NJ., 1973.

53. Newell, A., Shaw, J. C., and Simon, H. A., Programming the Logic Theory Machine. *Proceedings of the Western Joint Computer Conference*, 15: 218 – 239. Reprinted in Feigenbaum and Feldman (1963).

54. Oakland, J. S., *Statistical Process Control: A Really Practical Guide*, 3rd Edition, Butterworth-Heineman, Oxford. 1996.

55. Olson, D. L., and Courtney, J. F., Jr. *Decision Support Models and Expert Systems*, Macmillan Publishing, NY. 1992.

56. Palm, A. C., Rodriguez, R. N., Spring, F. A., and Wheeler, D. J., Some perspectives and challenges for Control Chart Methods. *Journal of Quality Technology.* Vol. 29, No. 2, April 1997.

57. Pham, D. T., and Oztemel, E. *Intelligent Quality Systems* Springer-Verlag, London, 1996.

58. Pilot Software OLAP White Paper. An Introduction to OLAP: Multi-Dimensional Terminology and Technology. Dun and Bradstreet Corporation. 1996.

59. Rietman, Edward A., Patel, S., Lory Earl R., Modeling and control of a Semiconductor Manufacturing process with an automata Network: An Example in Plasma Etch Processing., Computers in Operations Research, 1996, Vol. 23, No. 6, pp. 573 – 585.

60. Rosenblatt, F. *Principles of Neurodynamics.* Spartan Books, New York, 1962.

61. Rumelhart, D. E. and McClelland, J. L. *Parallel distributed processing: exploration in the microstructure of cognition.* (Vols. 1 and 2) MIT Press, Cambridge, MA. 1986.

62. Russell, S. J., and Norvig, P., *Artificial Intelligence: A Modern Approach*, Prentice Hall, Englewood Cliffs, NJ., 1995.

63. Schalkoff, R. J., *Artificial Intelligence: An Engineering Approach*, McGraw-Hill, Highstown, NJ., 1990.

64. Scott Morton M. S., *Management Decision Systems: Computer Based Support for Decision-making*, Graduate School of Business Administration, Harvard University, Boston, 1971.

65. Shewart, W. E., and Deming, W. E., (editor and foreword) *Statistical Method from the viewpoint of quality control.* Dobver Publications Inc., 1986.

66. Shortliffe, E. *MYCIN: Computer Based Medical Consultation.* American Elsevier, 1976.

67. Simon, H. A., *Administrative Behavior: A Study of Decision-Making Processes in Administrative Organization*, The Free Press, NY., 1976.

68. Simon, H. A., Cyert, R. M., Trow, D.B., "Observation of Business Decision," *The Journal of Business*, 1956, pp. 237-248.

69. Simon, H. A., *The New Science of Management Decision.* Englewood Cliffs, NJ: Prentice Hall. 1977.

70. Simon, H.A., Why should machines learn? in Michalski, R. S., Carbonell, J. C. and Mitchell, T. M. (editors) *Machine Learning: An Artificial Intelligence approach* Vol. I. Morgan Kaufmann Publishers, Inc., Los Altos, CA, 1983.

71. Simoudis, E., Reality Check for Data Mining. *IEEE Expert* October 1996.

72. Smith, A. E., and Yazici, H. An intelligent composite system for statistical process control, *Engineering Applications of Artificial Intelligence.*, Vol. 5, No.6., 1992.

73. Stacey, D., Intelligent systems architecture: Design techniques, in Dagli, C. H., (editor) *Artificial Neural Networks for Intelligent Manufacturing,* Chapman & Hall, 1994.

74. Thomsen, E. *OLAP Solutions: Building Multidimensional Information Systems.* John Wiley and Sons, Inc. 1997.

75. Turban, E., and Aronson, J. E., *Decision Support Systems and Intelligent Systems,* Prentice Hall, NJ. 1998.

76. Winston P. H., *Artificial Intelligence.,* Addison-Wesley, Reading, Massachusetts, 1990.