J Intell Manuf (2009) 20:501–521
DOI 10.1007/s10845-008-0145-x

Data mining in manufacturing: a review based on the kind of knowledge

A. K. Choudhary · J. A. Harding · M. K. Tiwari

Received: 14 December 2007 / Accepted: 2 June 2008 / Published online: 15 July 2008

Abstract In modern manufacturing environments, vast amounts of data are collected in database management systems and data warehouses from all involved areas, including product and process design, assembly, materials planning, quality control, scheduling, maintenance, fault detection etc. Data mining has emerged as an important tool for knowledge acquisition from the manufacturing databases. This paper reviews the literature dealing with knowledge discovery and data mining applications in the broad domain of manufacturing with a special emphasis on the type of functions to be performed on the data. The major data mining functions to be performed include characterization and description, association, classification, prediction, clustering and evolution analysis. The papers reviewed have therefore been categorized in these five categories. It has been shown that there is a rapid growth in the application of data mining in the context of manufacturing processes and enterprises in the last 3 years. This review reveals the progressive applications and existing gaps identified in the context of data mining in manufacturing. A novel text mining approach has also been used on the abstracts and keywords of 150 papers to identify the research gaps and find the linkages between knowledge area, knowledge type and the applied data mining tools and techniques.

A. K. Choudhary · J. A. Harding (✉)
Wolfson School of Mechanical and Manufacturing Engineering, Loughborough University, Loughborough, Leicestershire LE113TU, UK
e-mail: J.A.Harding@lboro.ac.uk

A. K. Choudhary
e-mail: a.k.choudhary@lboro.ac.uk

M. K. Tiwari
Department of Industrial Engineering and Management, Indian Institute of Technology, Kharagpur 721302, India
e-mail: mkt09@hotmail.com

Keywords Knowledge discovery · Data mining · Manufacturing · Text mining · Literature review

Introduction

Knowledge provides power in many manufacturing contexts enabling and facilitating the preservation of valuable heritage, new learning, solving intricate problems, creating core competencies and initiating new situations for both individuals and organizations now and in the future (Choudhary et al. 2007). In most sectors, manufacturing is extremely competitive and the financial margins that differentiate between success and failure are very tight, with most established industries needing to compete, produce and sell at a global level. To master these trans-continental challenges, a company must achieve low cost production yet still maintain highly skilled, flexible and efficient workforces who are able to consistently design and produce high quality and low cost products. In higher-wage economies, this can generally only be done through very efficient exploitation of knowledge (Harding and Popplewell 2006; Choudhary et al. 2006). However knowledge can take many forms and it is necessary to identify the kind of knowledge to be mined when examining the huge amount of data generated during manufacturing.

In modern manufacturing, the volume of data grows at an unprecedented rate in digital manufacturing environments, using barcodes, sensors, vision systems etc. These data may be related to design, products, machines, processes, materials, inventories, maintenance, planning and control, assembly, logistics, performances etc., and may include patterns, trends, associations and dependencies. However, the use of accumulated data has been limited, which has led to the "rich data but poor information" problem (Wang and McGreavy

1998). The collected manufacturing data contains valuable information and knowledge that could be integrated within the manufacturing system to improve decision making and enhance productivity (Elovici and Braha 2003).

The huge amounts of data in manufacturing databases, which contain large numbers of records, with many attributes that need to be simultaneously explored to discover useful information and knowledge, make manual analysis impractical. All these factors indicate the need for intelligent and automated data analysis methodologies, which might discover useful knowledge from data. Knowledge discovery in databases (KDD) and data mining (DM) have therefore become extremely important tools in realizing the objective of intelligent and automated data analysis. Data mining is a particular step in the process of KDD, involving the application of specific algorithms for extracting patterns (models) from data. The additional steps in the KDD process, such as data preparation, data cleaning, data selection, incorporation of appropriate prior knowledge and proper interpretation of the results of mining, ensure that useful knowledge is derived from the data (Mitra et al. 2002).

KDD incorporates theories, algorithms and methods from the intersection of several research fields including database technology, machine learning, statistics, artificial intelligence, knowledge based systems and data visualization. A natural question that arises is how KDD is different from other fields such as machine learning or artificial intelligence or related fields? The answer is that these fields provide specific data mining tools that can be used in various steps of a KDD process. Recently, with the growth of data mining technology, researchers and practitioners in various aspects of manufacturing and logistics have started applying this technology to search for hidden relationships or patterns which might be used to equip their systems with new knowledge. Early applications of data mining were mostly applied to financial applications, for example Zhang and Zhou (2004) described data mining in the context of financial applications from both technical and application perspectives. In this area, the competitive advantage gained through data mining included increased revenue, reduced cost, much improved market place responsiveness and awareness. A recent survey carried out by Harding et al. (2006) and a special issue published on "data mining and applications in engineering design, manufacturing and logistics" (Feng and Kusiak 2006) clearly indicated the potential scope of data mining in these areas to achieve competitive advantages. A major advantage of data mining over other experimental techniques is that the required data for analysis can be collected during the normal operation of the manufacturing process being studied. Therefore, it is generally not necessary to specially dedicate machines or processes for data collection.

The diversity of data mining tools, techniques and functionalities provides great opportunities, but the profusion of options can cause confusion. Han and Kamber (2001) classified data mining systems based on various criteria such as *kind of database mined, kind of knowledge mined, kind of technique utilized and application areas adopted*. Pham and Afify (2005) reviewed machine learning techniques in the manufacturing domain. They evaluated several machine learning techniques discussing their advantages and disadvantages and examined application areas in which they have been successfully deployed. Harding et al. (2006) surveyed data mining systems in different application areas of manufacturing, including some less considered areas such as manufacturing planning and shop floor control. Over the decades, data mining has been applied in various but limited aspects of manufacturing and logistics domains. However, in the last few years, data mining research in manufacturing has increased at an exponential rate. In the last 3 years, approximately 75 papers have been published, which is 50% of the relevant literature identified and reviewed in this paper.

Han and Kamber (2001) mentioned that the kind of knowledge to be mined determines the data mining functions to be performed. Possible kinds of knowledge include *concept description (characterization and discrimination), association, classification, clustering, and prediction*. The aim of this paper is therefore to consolidate the existing state-of-the art research efforts concerning the current practices in data mining applications in manufacturing based on the *kind of knowledge mined* and *the kind of technique* utilized, thereby identifying promising areas for study. This paper also identifies the most utilized techniques for mining certain types of knowledge, by using a novel text mining (TM) approach which has been applied on the abstracts and keywords of 150 papers. The results identify the research gaps, and show the linkages between knowledge area, knowledge mined and kind of technique used to mine data in the manufacturing domain.

The remainder of the paper is organized as follows. The next section briefly discusses about KDD, data mining, and the kinds of knowledge that particularly occur in manufacturing contexts. Section "Concept description (characterization and discrimination) in manufacturing" will discuss concept descriptions which include characterization and discrimination in manufacturing. Classification in manufacturing is discussed in section "Classification in manufacturing", followed by clustering in manufacturing in section "Clustering in manufacturing". Section "Prediction in manufacturing" discusses prediction in manufacturing, and association in manufacturing is discussed in section "Association in manufacturing". Details of our novel text mining approach are given in section "Detailed analysis and discussion: a text mining perspective on reviewed literature" and this is followed by conclusions in section "Conclusion".

KDD, data mining and knowledge types

Basic definitions

KDD is the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data (Fayyad et al. 1996a). Data mining(DM) is a particular step in this process, involving the application of specific algorithms for extracting patterns (models) from data. The additional steps in the KDD process, such as data preparation, data selection, data cleaning, incorporation of prior knowledge, and proper interpretation of the results of mining ensure that useful knowledge is derived from the data. An important notion of "interestingness" is usually taken as an overall measure of pattern value, combining validity, novelty, usefulness, simplicity and understandability. As a matter of fact, knowledge in this definition is purely user oriented and domain specific and is determined by whatever function and threshold the user chooses. The role of interestingness is to threshold the huge number of discovered patterns and report only those which may be of some use (Macgarry 2005).

KDD process in manufacturing

The overall KDD process applied in manufacturing is delineated in Fig. 1. The KDD process is interactive and iterative involving more or less the following steps (Fayyad et al. 1996b; Mitra et al. 2002).

- I Understanding the manufacturing domain: This includes the relevant prior knowledge related to manufacturing application and targeted goal.
- II Collecting the targeted data: This includes the collection of raw data, selecting the data set and focussing on the set of variables affecting the manufacturing problem.
- III Data cleaning, pre-processing and transformation: This includes the pre-processing of data such as noise removal, missing value replacement and data cleaning. Data are consolidated into forms appropriate for mining.
- IV Data integration: This includes integrating the multiple manufacturing heterogeneous data sources.
- V Choosing the functions of data mining: Based on the kind of knowledge required, various data mining functions (clustering, classification, prediction, association, regression, summarization etc.) need to be performed to derive the model.
- VI Choosing the appropriate data mining algorithm: This includes the selection of techniques to perform the desired function to find the patterns in the data.
- VII Data mining: This includes searching for patterns of interest in a particular representational form or a set of such representations.
- VIII Interpretation and visualization: This includes the interpretation and visualization of patterns to derive novel knowledge.
- IX Implementation of discovered knowledge: The discovered knowledge is incorporated into the manufacturing domain performance system. Feed back is received and the knowledge can be modified further based on feedback.
- X Knowledge storage, reuse and integration into the manufacturing system: This includes the storage of discovered knowledge for future reuse and possible integration in to the manufacturing system.

Data mining in manufacturing

Data mining is an interdisciplinary field with the general goal of predicting outcomes and uncovering relationships in data. It makes use of automated tools and techniques, employing sophisticated algorithms to discover hidden patterns,

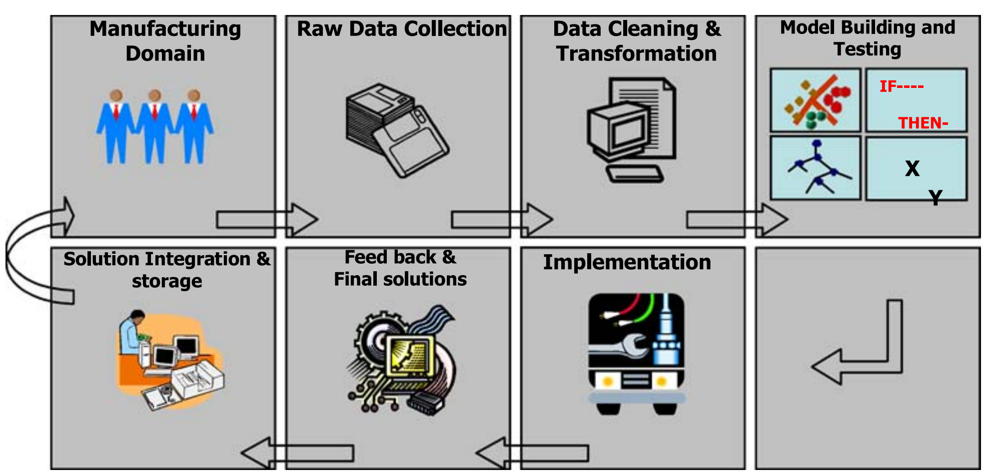

Fig. 1 Knowledge discovery and data mining in manufacturing systems environment

associations, anomalies and/or structure from large amounts of data stored in a data warehouse or other information repositories. In the context of manufacturing, two high level primary goals of data mining are prediction and description. Descriptive data mining focuses on discovering interesting patterns to describe the data. Predictive data mining focuses on predicting the behaviour of a model and determining future values of key variables based on existing information from available databases. The boundaries between, descriptive and predictive data mining are not sharp, e.g. some aspects of the predictive model can be descriptive, to the degree that they are understandable and vice versa. The goals of prediction and description can be achieved by using a variety of data mining tools and techniques.

Kind of knowledge to be mined

As explained in the introduction, before performing a data mining task in manufacturing, it is necessary to identify the kind of knowledge to be mined. Kind of knowledge mined determines the data mining functions. The next section therefore describes a range of functions and reviews their applicability in manufacturing domains.

Concept description (characterization and discrimination) in manufacturing

Descriptive data mining describes the data set in a concise and summarative manner and presents interesting general properties of data. In most manufacturing problems, it is necessary to view the summarized data in concise, descriptive terms to provide an overall picture of the manufacturing domain's data or distinguish it from a set of comparative classes. This type of data mining is called concept description and includes characterization and discrimination. For example in printed circuit board (PCB) manufacturing quality control problems, it is necessary to identify the various features that cause defects. Characterization can be used to identify the features that significantly impact the quality. Characterization provides a concise and succinct summarization of the given collection of data, while concept or class discrimination or comparison provides descriptions that compare two or more collections of data. In manufacturing contexts, these functions are basically used to understand the process.

Quality control

Tseng et al. (2004a) developed an extended rough set theory (RST) based approach to solve the quality control (QC) problems of PCB manufacturing. The algorithm is able to derive the rules and identify the most significant features to solve the QC problems. The rules derived from the dataset provide an indication of how to best investigate the problem further. Unnatural control chart patterns (CCP) are associated with a particular set of assignable causes for process variation. Guh (2005a) presented a hybrid neural network and decision tree based model to detect and discriminate typical unnatural CCPs, while identifying the major parameters and starting point of the CCP detected.

Job shop scheduling

Huyet (2006) proposed an evolutionary optimization and data mining based approach to produce the knowledge of systems behaviour in a simulated job shop based production process. Assigning proper dispatching rules is an important issue in enhancing the performance measures for a flexible manufacturing system (FMS). Liu et al. (2005) presented SAMA (supervised attribute mining algorithm) which is based on the fuzzy set theoretic approach and genetic algorithm (GA) for the attribute selection problem of dispatching rules. Li et al. (2006) improved the accuracy of flexible manufacturing systems (FMS) scheduling for small datasets, as at the early stage of manufacturing only a few data points can be obtained. Their study develops a data trend estimation technique and combines it with mega-fuzzyfication and an adaptive network based fuzzy inference system (ANFIS). Li and Olafsson (2005) introduced a novel methodology for generating scheduling rules using data mining. This approach includes preprocessing of historic scheduling data into an appropriate data file, discovery of key scheduling concepts and then represents the data mining results so that they can be used for job scheduling. Koonce and Tsai (2000) applied data mining methodologies to explore the patterns in data generated by a GA performing scheduling operations and to develop a rule set scheduler which approximates the GA's scheduler. The Attribute Oriented Induction approach was used to characterize the relationship between the operations' sequences and their attributes. Knowledge generated from data mining can be used to analyze the effect of decisions made at any stage. Belz and Mertens (1996) used SIMULEX coupled with a knowledge based system to model the plant and evaluate the results of various rescheduling measures. They used multivariate analysis of variance (MANOVA) for statistical analysis and the collected data was analysed to identify normal and abnormal patterns.

Fault diagnostics

Lee and Ng (2006) presented a hybrid case based reasoning (Hy-Case) system for online technical support of PC fault diagnosis. Hy-Case consists of a natural language (keyword) as input and a graph theoretic constraint-net. Fountain et al. (2003) described a decision theoretic approach to die level functional test (DLFT) in which historical test data from integrated circuit (IC) products manufactured by Hewlett Packard were mined to create a probabilistic model of patterns for die failure. This model is combined with greedy value-of-information computation to decide in real time, which die to test next and when to stop testing. The KDD process is basically a user driven task but the user driven approach is not efficient enough for use in manufacturing. Maki and Teranishi (2001) therefore developed an intelligent system in Hitachi for online data analysis using a data mining approach. Their approach consists of three step feature extraction, combinatorial search and presentation. They applied this method to fault analysis and found that data mining is useful for indicating to engineers where to focus their attention when looking for faults. Subsequently a similar methodology was applied on QC of production processes (Maki et al. 2000). A rough set theory based approach was used by Shen et al. (2000) to get the final reducts and extract the rules for fault diagnosis of diesel engines. These rules were used to distinguish the fault type and to inspect the dynamic characteristics of the machinery.

Manufacturing process

Jung et al. (2006) proposed a vertical group wise threshold (VGWT) procedure for the multiple high dimensional functional data. Here the "Functional Data" refers to the analogue or digital signal measured during each operational cycle of a manufacturing process. The proposed methodology increases the class separability with a reasonably small loss in data reduction efficiency. Dofasco, a fully integrated Canadian Steel manufacturer applied multivariate statistical analysis to several steel production processes to improve the productivity and reduce maintenance costs (Zhang and Dudzic 2006). Caskey (2001) developed a general environment for providing the right knowledge to achieve a specific factory condition at the right time. In addition to discrete event simulation, he used GAs and neural networks (NN) in identifying the structure of the data. The knowledge extracted was in the form of "actual control applied → performance obtained" and the knowledge generated could be used to increase the accuracy of the system or validate the performance model. Menon et al. (2005) analyzed the product development process based on a textual database using data mining to deliver the right information at the right time to satisfy particular needs. The data mining tools were applied on product development process data to provide quick and reliable feedback that would facilitate faster product development.

Manufacturing system

Neaga and Harding (2005) presented a framework for the integration of complex enterprise applications including data mining systems. Their approaches provide the definition and development of a common knowledge enterprise model, which represents a combination of previous projects on manufacturing enterprise architectures and Object Management Group (OMG) models and standards related to data mining. The Decision tree technique has proved to be the valuable tool for description, generalization and classification. Chen et al. (2000) applied hyperspace data mining in materials manufacturing. Master Miner is provided with several algorithms to provide better understanding of materials design and manufacturing. Holden and Serearuno (2005) developed *iGem*, an artificial intelligence tool that integrates rule based knowledge representation, fuzzy logic and genetic algorithms to produce a system for automating and introducing consistency into the grading of diamonds and gemstones. It derives knowledge from repeated examples of previously correctly graded stones and improves its performance by learning from experience. A french truck manufacturer applied KDD techniques on the datasets of measures recorded during the test of diesels manufactured at their production lines. By analyzing the data, they significantly (about 25%) reduced the processing time (Gertosio and Dussauchoy 2004). This provided production engineers and operators with useful and easy to understand results.

Maintenance

Romanowski and Nagi (1999) applied a decision tree based data mining approach on a scheduled maintenance dataset and a vibration signal dataset. Subsystems which are most responsible for low equipment availability are recognized in the scheduled maintenance data and a recommendation for preventive maintenance interval is made. The vibration data is analyzed to find the sensors and frequency responses which give the most information about the type of fault present in the equipment. Batanov et al. (1993) researched knowledge-based maintenance systems and developed a prototype system called EXPERT-MM, which works on historical failure data and provides suggestions for an appropriate preventive maintenance schedule.

Defect analysis

Kusiak and Kurasek (2001) applied a RST based data mining approach to identify the cause of solder defects in a PCB. The rules derived from the dataset provide a robust indication of where the solder is too narrow, and thereby make more effective further investigations into the cause of solder ball defects. For ceramics manufacturing, Dengiz et al. (2006) presented a two stage data-mining approach to accurately estimate flaw distribution. The first stage of data mining is image processing for automatic identification and recovery of large flaws from noisy microscopic images. The second stage is to use the flaw information to fit the extreme value distribution function. These provide the manufacturer and designer with a better understanding of the processing method.

Yield improvement

Data mining has proved to be important for yield improvement which is a key element to ensure the profitability of semiconductor manufacturing. Bergeret and Gall (2003) applied a Bayesian network to solve yield issues, where the root cause comes from a failure at a single process stage. Three different methods Statistical Process Control (SPC) technique, Design Of Experiment (DOE) and Bayesian Network have been applied to resolve two critical yield issues (a) faulty batch of raw material and (b) process change. The advantage of this approach is that the only data required are the process dates of the lots at each process stage and the probe results. Dabbas and Chen (2001) presented an integrated relational database approach for modelling and collecting semiconductor manufacturing data from a multi-database system and transforming the data into useful reports. These reports are utilized to monitor the performance of one of Motorola's wafer fabs by tracking different key metrics and in turn improving the factory performance.

In addition, Murthy (1998) surveyed the applicability of data mining in various domains and identified the important issues involved. He concluded that the hierarchical tree construction methodology is very powerful and has repeatedly been shown to be useful for diverse real world problems. Hou and Yang (2006) presented a data model and technology mining model to investigate the technology categories with more return and client demand in a Technology and Service Provider (TSP) environment. The proposed model can be used to derive suggestions including the critical issues such as technology development, return of technology, R&D investment and outreach sequence of technology.

From this review, the main areas where data mining is used for concept description (characterization and discrimination) include fault diagnostics, scheduling, dispatching, maintenance and manufacturing system and processes. The next section will describe classification, which is another functionality performed on manufacturing data.

Classification in manufacturing

Classification is a useful functionality in many areas of manufacturing, for example, in the semiconductor industry, defects are classified to find patterns and derive the rules for yield improvement. Online control chart pattern recognition (CCPR) is another example of classification for SPC, because unnatural patterns displayed by a control chart can be associated with specific causes that adversely impact the manufacturing process.

Classification is a learning function that maps (classifies) a data item into one of several predefined categorical classes. Generally, classification is performed in two steps. In the first step, a model is built to describe a predetermined set of data classes or concepts, and this is done by analyzing the database tuples described by attributes, which collectively form the training dataset. This step is also known as supervised learning as the class label of each training sample is provided. The learned model is represented in the form of classification rules, decision trees, or mathematical formulae. In the second step, based on classifier accuracy, the model is used for classification of the future data or test data. General techniques used for classification are decision tree induction, Bayesian classification, Bayesian belief network and neural networks (NN). Other techniques such as K Nearest Neighbour (KNN), Case Based Reasoning (CBR), GA, RST, Fuzzy Logic (FL) and various hybrid methods are also used for classification purposes (Han and Kamber 2001).

Quality control

False recognition or incorrect classification is a frequently encountered issue in the CCPR. In this regard, Guh (2005b) presented a hybrid learning based system that integrates neural network and decision tree learning to overcome the CCP classification problem in real time. This hybrid system consists of three sequential modules namely feature extraction, coarse classification and fine classification. It outperformed the conventional approaches in terms of recognition speed and accuracy. Rokach and Maimon (2006) applied a feature set decomposition methodology for quality improvement. They developed the Breadth Oblivious Wrapper (BOW) algorithm and showed its superiority over existing tools on datasets from IC fabrication and food processing. The idea is to find the classifier that is capable of predicting the quality measure of product or batch based on its manufacturing parameters.

Job shop scheduling

Kwak and Yih (2004) presented a data-mining based production-control system for testing and rework in dynamic CIM. Their system analyses the present situation and suggests dispatching rules to be followed and also how data mining can be used to evaluate the effect of those decisions. It uses a decision tree based module to generate classification rules on the partitioned data that are suitable for interpretation and verification by users and stores the rule in the competitive decision selector (CDS) knowledge bases.

Fault diagnosis

Knowledge acquisition is a well known problem in building expert systems. In one of the earliest applications, Irani et al. (1993) developed an expert system for diagnosis and process modelling of semiconductor manufacturing that uses a generalized ID3 (GID3) based machine learning technique for knowledge acquisition. The discovered pattern is consistent with the data and conformed with the engineers' expectations. GID3 has been used for classification purposes in areas targeted for automation by the Semiconductor Research Corporation, a consortium of US semiconductors manufacturers. Skormin et al. (2002) presented a classification model for the database containing the information downloaded from a dedicated monitoring system of flight critical hardware. A decision tree based data mining model has been used for the accurate assessment of the probability of failure of any avionics unit by utilizing historical data relating to environment and operation condition. As data sets increase in size, exploration, manipulation and analysis become more complicated and resource consuming. To deal with these situations, Jeong et al. (2006) proposed a data compression and data denoising method, which was used with tree based classification for identifying classes of process faults. Rojas and Nandi (2006) applied the support vector method (SVM) for fault classification of rolling element bearings. The classification procedure was fast and effective (95% successful).

Condition based monitoring

Extensive research has been carried out in developing knowledge based condition monitoring systems. To deal with noise, uncertainty and decrease the rate of misclassification, Peng (2004) proposed a hybrid Fuzzy-inductive learning based approach. This method has been successfully deployed in diagnosing the conditions of a tapping process. The results showed that the accuracy of classification is better than the traditional inductive learning method. Pasek (2006) explored a rough set theory based classifier for cutting tool wear monitoring applications, but, this method is only applicable to offline situations. For online monitoring, Hou et al. (2003) integrated neural network and rough sets technique for manufacturing process monitoring and fault diagnosis. A neural network was used to classify the quality faults such as wrinkles and uneven thickness. Apart from providing operating guidelines, rough set is used to determine the causal relationship between parameters like process temperature and output quality measures. A fuzzy set theory with a fuzzy variable rough set approach has been used in an extension of this work (Hou and Huang 2004). The results showed that the proposed approach was robust and able to deal with noisy data and produced better rules.

Manufacturing process

Braha and Shmilovici (2002) presented three classification based data mining methods (decision tree induction, neural network and composite classifier) for a new laser based wafer cleaning process called advanced wafer cleaning. The purpose of the data mining based classifier is to enhance understanding of the cleaning process by categorizing the given data into a given predefined number of categorical classes and determine to which the new data belongs. A data mining based system called RMINE was developed to deal with the incomplete and noisy data in condition based monitoring of a centrifugal pump system. The major advantage of this system is its self learning ability to cope with the changing condition based data (Li et al. 2006). Liao et al. (2001) discussed a multi-layer perceptron neural network to model radiographic welding data.

Defect analysis

McDonald (1999) discussed the applicability of data mining in online inspection and control, and automatic defect classification. He made a proposal to extend data mining tools for improvement of reliability of products during wafer fabrication, packaging and final testing steps. Wang et al. (2006) proposed a data mining based approach for automatic defect detection and classification (ADD/C), using hierarchical clustering with a k-means partitioning approach followed by a Gaussian EM algorithm to classify the defects patterns. A fractal dimension based classifier was proposed by Purintrapiban and Kachitvichyanukul (2003) for detection of unnatural patterns in process data.

Manufacturing system

Kusiak (2002a), Kusiak (2002b) applied data mining to support decision making processes by using different data-mining algorithms to generate rules for a manufacturing system. A subset of these rules was then selected to produce a control signature for the manufacturing process where the control signature is a set of feature values or ranges that lead towards an expected output. These control signatures are updated using a learning classifier based framework. Busse et al. (2005) compared two data mining approaches for the classification of imbalanced data. In the context of a flexible manufacturing system, Horng and Lin (2004) developed a hybrid classification tree (HCT) to classify the products of complicated machines. They used a separation matrix and fuzzy rules to split the data set into terminal clusters and then further classification and a regression tree (CART) were used for classification of the terminal clusters. This approach achieved a 40% reduction in training time. Hsu and Wang (2005) used a decision tree based approach to identify and classify significant patterns in the body shape of soldiers. This sizing system provided the garment manufacturers with size specifications, design development, pattern grading and market analysis. Zhang and Jiao (2007) proposed an associative classification based recommendation system for personalization in B2C e-Commerce applications to anticipate customers' heterogeneous requirements. They also discussed its implementation, design and system analysis in a web based environment.

From this review, the major application areas where data mining tools and techniques are used for classification include fault diagnosis, quality control and condition monitoring. In order to perform the classification task, decision tree, rough set theory, hybrid neural network and other hybrid approaches have been successfully used. In hybrid approaches, Fuzzy logic is used often in combination with other techniques to deal with noise and uncertainty in the data. The next section will deal with clustering and its performance on manufacturing databases.

Clustering in manufacturing

Clustering is an important data mining function that can be performed on specified manufacturing data such as order picking in logistics and supply chain. For example order picking is routine in distribution centres and before picking a large set of orders, orders are clustered into batches to accelerate the product movement within the storage zone. Clustering is also useful in the formation of cells in cellular manufacturing where it is used for the simultaneous design of the part families and machine cells.

Clustering is also known as unsupervised learning. Unlike classification (supervised learning), in clustering the class object of each data object is not known. Clustering maps a data item into one of several clusters, where clusters are natural groupings of data items based on similarity metrics or probability density models (Mitra et al. 2002; Xu and Wunsch 2005). Within the same cluster data objects are similar to one other but they are dissimilar to the objects in other clusters. Generally clustering techniques are classified in to the following categories: partitioning methods, hierarchical methods, density based methods, grid based methods, and model based methods. Liao and Wen (2007) reviewed the application of artificial neural networks (ANN) for clustering and classification. A detailed review and study of clustering techniques' application areas are mentioned in Xu and Wunsch (2005).

Yield improvement

A combination of self organizing map (SOM) neural networks and rule induction was used by Gardner and Bieker (2000) to identify the critical poor yield factors from normally collected wafer manufacturing data and thus increase the yield. The system is implemented with CorDex, which creates a two dimensional relational topology, called a "cluster map" that maintains the dimensional data interrelationship. To increase the yield of semiconductor manufacturing, Chien et al. (2007) developed a framework consisting of kruskal-wallis test, k-means clustering, and variance reduction splitting criteria.

Manufacturing process

Sebzalli and Wang (2001) applied principal component analysis and fuzzy c means clustering to a refinery catalytic process to identify operational spaces and develop operational strategies for the manufacture of desired products and to minimize the loss of product during system changeover. Four operational zones were discovered, with three for product grade and the fourth region giving high probability of producing off-specification product. Liao et al. (1999) presented fuzzy clustering based techniques for the detection of welding flaws. A comparative study between fuzzy k-nearest neighbours clustering and fuzzy c means clustering has been performed.

Design

Engineering design is knowledge intensive. Jin and Ishino (2006) proposed a design activity knowledge acquisition (DAKA) framework to discover useful design activity knowledge from CAD event data. The discovered knowledge reflects a trajectory of designers approach to reach a final design product model and identifies key design operations for specific design tasks. Kim and Ding (2005) proposed a data mining aided optimal design method for fixture layout in a four station SUV side panel assembly process. Clustering and classifications are carried out to generate a design library and design selection rules, respectively. They also carried out a comparative study of fuzzy c means clustering and crisp methods on cell design problems. Torkul et al. (2006) showed the outperformance of fuzzy c means clustering over crisp methods on a selected data set. Romanowski and Nagi (2001) proposed a design system which supports the feedback of data mined knowledge from life cycle data to the initial stages of the design process. Romanowski and Nagi (2005) and Romanowski and Nagi (2004) also applied a data-mining approach for forming generic bills of materials (GBOMS) entities that represent the different variants in a product family and facilitate the search for similar designs and the configurations of new variants. They resolved the technical difficulties associated with GBOM by combining data-mining approaches such as text and tree mining in a new tree union procedure that embodies the GBOM and design constraints in constrained XML. In these works, BOM are clustered into product families and subfamilies. With the increase of automated data generation and gathering in manufacturing, mining interesting patterns is of utmost concern. Kusiak (2000a) and Kusiak (2000b) presented various algorithms and models for design data analysis including cluster analysis, precedence analysis and other data mining methods.

Defect detection

Lee et al. (2001) proposed an intelligent in-line measurement sampling method for process excursion monitoring and control in semiconductor manufacturing. Chip locations were clustered within the wafer through SOM classifications. It has been shown that the generated sampling method can be very effective for all defect detection despite small sampling size. Crespo and Webere (2005) presented a fuzzy c means clustering based methodology for dynamic data mining and applied it on real life data.

Fault diagnosis

Huang et al. (2005) presented an integrated diagnostic support system which uses hybrid rough set theory and a genetic algorithm. The proposed approach has been applied at a mother board manufacturing company to discover decision rules for electro magnetic induction (EMI) faults. The average diagnostic accuracy of 80% showed that this hybrid model is promising for an EMI diagnostic support system. Hui and Jha (2000) investigated the application of data mining techniques to extract knowledge from the customer service database for decision support and fault diagnosis. Both structured and unstructured data are mined over the world wide web and their proposed data mining technique integrates neural networks and case based reasoning, rule based reasoning for classification and clustering purposes.

Supply chain

Symeonidis et al. (2003) applied data mining to make ERP systems more versatile and adaptive by integrating them with knowledge extracted from companies' selling policies in order to determine the customer trends. Qian et al. (2006) applied clustering algorithms for churn detection via customer profile modelling. Chen et al. (2005) proposed an association rule mining based algorithm, which automatically consolidates orders into batches for minimizing the cost in terms of distance and time. In another work, they used association rule mining and 0–1 integer programming model for the same objective (Chen and Wu 2005).

Caramia and Felici (2006) applied a clique based approach to mine relevant information on the web in order to enhance the capability of a search engine by, for example identifying structural information in a set of pages, or by providing semantic extensions to searches for key words. Liao et al. (2006) proposed an adaptive genetic clustering method for exploratory mining of feature vectors and time series data. The proposed approach produced comparable or better clustering accuracies than the k-means approach when tested on the bench mark datasets. Inada and Teraano (2005) presented a useful and effective method "QC Chart Mining" to extract the systematic patterns from quality control charts in order to manage the clinical test data.

From this review, the major areas where clustering is used for deriving useful knowledge include design, yield improvement and fault diagnostics etc. The major tools and techniques used for clustering purposes include neural network, fuzzy logic with k means or c means clustering. However, several hybrid tools with fuzzy techniques have also been developed for the problem specific domains. The next section will deal with prediction in manufacturing.

Prediction in manufacturing

Predictability of manufacturing processes, quality, maintenance, defects, or even within manufacturing systems is of vital importance. For example in the context of maintenance, predictions can be made about what condition maintenance will be required or how equipment will deteriorate based on the analysis of past data. Prediction is a learning function that maps a data item to a real valued prediction variable. Prediction can be viewed as the construction and use of a model to assess the class of an unlabelled sample, or to assess the value or value range of an attribute that a given sample is likely to have.

Manufacturing process

Model selection and cross-validation (CV) are critical issues in data collection and data mining for predicting the performance of manufacturing processes. Feng and Kusiak (2006), Feng et al. (2006) showed that there is no significant statistical advantage of using fivefold CV over threefold CV and or of using a two hidden layer neural network over a one hidden layer neural network for turning surface roughness data. Pasek (2006) used the rough set theory based classifier for the prediction of cutting tool wear. For tool condition monitoring Sun et al. (2005) applied a neural network for recognition of tool condition in a monitoring system. Then, the SVM approach with two regularization parameters was employed to adjust the recognition ability for each tool wear condition separately. The experimental result showed that the proposed model can reliably identify tool flank wear and reduce the over due prediction of worn tool conditions and its relative loss. Tseng et al. (2005a) and Tseng et al. (2004b) proposed an RST based approach to solve the quality assurance problem in predicting the acceptance of CNC machined parts, rather than focussing on the prediction of precise surface roughness values. The rule composition algorithm and rule validation procedure provide a higher accuracy prediction tool for investigating features. In the context of an e-manufacturing environment, (Tseng et al. 2005b) applied data mining and type II fuzzy system for predicting the system output of a CNC turning operation. Feng and Wang (2004) investigated the application of two competing data mining techniques, regression analysis and artificial neural network to develop a predictive model of a knurling process. Fractional factorial design and design of experiment (DOE) were used to plan the experiments and hypothesis testing was conducted to test each model. Their research concluded that the neural network performed better when compared to regression modelling. A comparative study of implicit and explicit methods to predict the non-linear behaviour of the manufacturing process using statistical procedure, neural network and case based reasoning (CBR) was discussed by Kim and Lee (1997). Hsieh (2004) developed a neural network and statistical method based thermal profile model for stress level prediction under voltage stress. The neural network performance was found to perform better when compared to regression analysis. Novel research has been done by Mere et al. (2004) in developing the predictive model for optimal mechanical properties of galvanised steel by using a combination of clustering and neural networks. Clustering was used in the first instance and then neural networks were applied to the clusters to predict the mechanical properties of the steel. Another interesting piece of work is reported by Yuan et al. (2000) in determining the toxicity (Microtox) of process effluents from a chemical plant by using neural networks and principal component analysis. Their software analyser predicts the toxicity level and helps in developing strategies in process operations for toxicity reduction in the effluents. A data mining approach for analyzing significance of nonlinearity effects in a multi station process has been applied by Ren et al. (2006). Regression tree based predictive model was used in identifying the critical factors, and responses. They also provided guidelines for users to decide when a nonlinear model should be used instead of a linear model.

Customer relations management

Customer relations management is an area where data mining is extensively used for predicting the behaviours of customers. Morita et al. (2000) developed a data mining server program that uses rule induction and memory based reasoning to effectively predict customer behaviour in telecommunication industry. Tseng et al. (2006) present a data mining based hybrid approach consisting of a new rough set algorithm for feature selection and a multi-class SVM method for more accurate prediction. This technique has been applied on a supplier selection case study to predict the preferred supplier of a video game system.

Maintenance

Sylvain et al. (1999) used different data mining techniques including decision trees, rough sets, regression and neural networks to predict component failure based on the data collected from the sensors of an aircraft. Their results also led to the design of preventive maintenance policies before the failure of any component. Lin and Tseng (2005) introduced a cerebellar model articulation controller (CMAC) neural network based machine performance estimation model. CMAC-PEM was used to fuse sensory data to predict machine reliabilities, and for condition based predictive maintenance. Yam et al. (2001) presented an intelligent predictive decision support system (IPDSS) for condition based maintenance. The proposed IPDSS model uses the recurrent neural network model in predicting the faults of critical equipment in a power plant. Earlier indication of failure provides more time for proper maintenance planning and scheduling.

Decision support system

Zhou et al. (2005) developed an agent based framework for intelligent prediction and monitoring of equipment failure and thereby support equipment prognostics and diagnostics. Data mining was used for the intelligent prediction engine, which is the key component of the system. Kusiak and Shah (2006) proposed a data mining based robust alarm system architecture for predicting incoming faults of water chemistry faults. The decision tree (DT) algorithm is used for learning and prediction purposes. However, the modular architecture of the developed system allows alternative knowledge generation modules to be incorporated using other data mining approaches such as simulation, analytical models and domain knowledge. Tsai et al. (2006) presented a case based reasoning (CBR) system using intelligent indexing and reasoning approaches for PCB defect prediction. Knowledge elicitation is a technique that is generally used for producing rules based on human expertise. A method was developed by Browne et al. (2006) to fuse knowledge elicitation and data mining using an expert system.

Job shop

Lead time and due date estimation are important issues as they critically affect customer relations and shop floor management practices. A regression tree based DM approach has been used by Ozturk et al. (2006), to estimate manufacturing lead time. Sha and Liu (2005) incorporated a data mining tool with a due date assignment method called total work content.

The decision tree is used for mining the knowledge of job scheduling to support due date assignment in a dynamic job shop environment, which is represented by IF-THEN rules. Song et al. (2005) applied a rough set theory based approach to predict the feasibility of a plan in a complex remanufacturing system. Their approach includes selecting the training dataset, attribute discrimination, attribute reduction, rule generalization and decision making. This research provided the basis for developing a data mining based simulation method with complicated constraints. A hybrid knowledge discovery model using decision tree and back propagation neural network has been developed by Wang et al. (2005) to determine the appropriate dispatching rule based on production data with noise information and predicting rule's performance. Chang et al. (2005) developed an evolving fuzzy rule (EFR) method based on Wang and Mendel's method (Wang and Mendel 1992) for due date assignment problems in manufacturing. The efficacy of their approach is demonstrated by comparing its performance with CBR and multi-layer perceptron. Chen et al. (2005) developed a methodology for predicting cycle time based on data mining and domain knowledge, given production status, WIP and throughput.

Manufacturing system

Kusiak et al. (2000) and Kusiak (2001a, 2000c, 2005, 2001b) extensively used data mining for medical and engineering applications. He introduced a novel data transformation method called feature bundling to improve the classification accuracy of decision rules (Kusiak 2001a). A list of decomposition methods has been presented and described in Kusiak (2000c) to enhance the quality of knowledge mined from vast industrial databases. A data mining based approach was also applied for identifying in-variant objects (a set of feature or parameter values) in semiconductor applications (Kusiak 2005). Rough set theory has been proved to be a powerful tool for discovering new knowledge and autonomous decision making. Kusiak (2001b) discussed the basic concept of RST as a prediction model. All these concepts follow the evolutionary computational approach for extending the longevity of knowledge. For cases where data are highly noisy, Last and Kandel (2004) presented a novel perception based method called the automated perception network (APN) for automated construction of compact and interpretable models. Their proposed method was applied on a real world dataset from the semiconductor industry, demonstrating the improved prediction capability of the constructed model. A comparative study of regression tree, KNN and clustering has been presented by Backus et al. (2006) to predict the intermediate cycle time of products. Giess et al. (2002), Giess and Culley (2003) developed a predictive model for manufacturing and an assembly database of gas turbine rotors to determine and quantify relationships between the various balance and vibration tests to highlight critical areas. This knowledge could then be fed back to the designers to improve tolerance decisions in the future design of components. They used a decision tree at the initial stage to determine appropriate areas of investigation and to identify problems with the data. At the next stage, a neural network was used to model the data. Ho et al. (2006) proposed an intelligent production workflow mining system with integrated online analytical process and data mining technology, combining ANN and fuzzy rule sets to realize knowledge discovery and decision support in high quality manufacturing. The result showed that their model is capable of producing accurate predictions over different manufacturing processes within production workflow. A knowledge based approximate life cycle assessment system (KALCS) was developed by Park and Seo (2006) to assess the environmental impacts of product design alternatives.

Yield improvement

Li et al. (2006) presented a genetic programming based DM approach for a yield prediction system and performed automated discovery of the significant factors that might cause low yield. The robustness and effectiveness of their approach has been shown on well known DRAM fab's real data. An integrated yield management system, called the Advanced Statistical Bin Limit (A-SBL) system has been proposed by Kang et al. (1998). It uses inductive decision trees and a neural network to manage yields over major manufacturing processes and provides functions to identify causal relationships between them such as yield prediction, process capability monitoring, feature selection, and wafer map stability monitoring. Zhou et al. (2001) applied the C4.5 algorithm for drop test analysis of electronic goods. They focussed on predicting the integrity of the solder joints for large components on the PCBs but their approach can also be extended to other parts.

Quality control

A supervised learner for eradicating the difficulties encountered in MEWMA has been developed by Li et al. (2006). They show that the tree based supervised learner can potentially detect change points in the control chart as it produces predictions that are piecewise constant. Rajagopal and Castillo (2006) proposed a Bayesian method that is able to set a tolerance limit on one or more responses to provide a given desired probability of conformance, and to determine at the same time the optimal setting of the control factors that the responses depend upon.

From this review, major areas where prediction has been applied as a data mining function include maintenance, manufacturing system, manufacturing processes, job shop type problems and as decision support systems. The data mining tools used include regression tree, ANN, rough set theory, decision induction trees and hybrid algorithms.

Association in manufacturing

Association rules mining was first introduced in 1993, and is used to identify relationships between a set of items in a database (Agrawal et al. 1993). These relationships are not based on inherent properties of the data themselves (as with functional dependencies), but rather are based on co-occurrence of the data items. Generally it is used by management for market basket type analysis to discover association rules to increase the effectiveness (and reduce the cost) associated with advertising, marketing, inventory and stock location on the floor. In manufacturing, Association rules can be used for prediction of failure in telecommunications networks by identifying what events occur before a failure.

In design contexts, the associations between requirements may provide additional information useful for the design. For example, technical specifications might state that a car that has two doors and a diesel engine requires a specific speed transmission. In such cases, knowing the number of cars with two doors and the number of cars with a diesel engine is not relevant whilst the number of cars with two doors and a diesel engine is useful, for example to determine the capacity of a manufacturing process. The nature of these association can be extracted by applying data mining algorithms on the database. More over, these associations may have pre-specified strength and confidence. A detailed survey of association rule has been carried out by Hipp et al. (2000) and Zhao and Bhowmick (2003).

Product design

Kusiak (2002b) applied rough set theory to derive associations between the process control parameters and the product families in the form of decision rules. Using an example in the metal forming industry, he showed that a control signature can be found to produce good quality products. Using data from the aerospace industry, Shahbaz et al. (2006) applied the association rule mining on historical product (Fan blade) data to extract information about the process limitations and knowledge of the associations between particular product dimensions. Generated information can then be feedback to establish design change requirement and product quality improvement. Agard and Kusiak (2004b) applied data mining to customer response data for its utilization in the design of product families. They used clustering for customer segmentation, i.e. to group the customers. The requirements from the product were then analyzed using association rules for the design of the product. Woon et al. (2003) proposed a method termed as Product development miner (PDMiner) to mine web logs efficiently and effectively using association rule and sequential pattern mining. Tsai and Chang (2003) proposed an intelligent design retrieval system that applies feature relation association and object form association. A Fuzzy ART neural network was used to search for relevant designs based on object form associations.

Manufacturing process

Wang et al. (2005) proposed the integration of variable precision rough set and fuzzy clustering to generate effective association rules for manufacturing process planning. The algorithm they used achieved the superiority performance over fuzzy decision technique and entropy based analysis method. Agard and Kusiak (2004a) applied an association rule mining base algorithm to generate knowledge for the selection of subassemblies based on the analysis of prior orders received from the customers. The generated knowledge can be applied to construct a model for selection of subassemblies for timely delivery from the suppliers to the contractors. Cunha et al. (2006) applied the association rule mining for improving the quality of assembly operations. The computational result showed that the source of assembly faults can be detected using association rule, even in the presence of noise. The associations extracted can be used to improve production quality by avoiding “risky” sequences. Chen (2003) used association rules for cell-formation problems. Associations among the machines are found from the process database, which leads to the identification of the occurrences of other machines with the occurrence of a machine in the cell. This approach also clusters the parts and machines into families and cells simultaneously and hence requires minimal manual judgement. Chen et al. (2005) generated association rules for defect detection in semiconductor manufacturing. They determined the association between different machines and their combination with defects to determine the defective machine.

CRM

Buddhakulsomsiri et al. (2006) developed a new association rule generation algorithm to extract the knowledge (in the form of rules) which could then be used to identify the root cause of a particular warranty problem and to develop useful conclusions from automotive warranty data. This knowledge was presented in the form of IF-THEN association rules, where the IF portions of the rule contain the product feature attributes and the THEN portion includes a problem related labour code. Chen and Wu (2005) applied association rule mining to derive rules, which imply the customer demand pattern directly from the order databases. By applying the association function, to an order database certain sets of customer orders that frequently consists of certain product items can be inducted.

Decision support

Jiao and Zhang (2005) developed explicit decision support to improve the product portfolio identification issue by using association rule mining from past sales and product records. Their research discovers associations among customers needs, marketing folk and designers. In a further extension of their work, they applied an association rule mining technique to deal with product and process variety mapping (Jiao et al. 2008). The mapping relationships are embodied in association rules, which can be deployed to support production planning of product families within existing production processes. A case study of mass customization of vibration motors has been illustrated to demonstrate how the association rule mining mechanism helps maintain the coherence between product and process variety. Shao et al. (2006) proposed a data mining based architecture to discover customer group based configuration rules in configuration design. The association rule mining based on a priori algorithm was used to get the association rules between of clusters of products specifications and configuration alternatives. Chao et al. (1997) presented an intelligent system to generate associative data for input in layout generation tools. They used an expert system, object oriented database and cluster analysis, which ensures data consistency and determines the strength of relationship between the two items under consideration.

This review shows that the major areas where association as a data mining function has been applied include product design, process control, mass customization, cellular design etc. Association rule mining has been applied as a dominating tool to identify the associations among variables.

Detailed analysis and discussion: a text mining perspective on reviewed literature

The reviewed literature shows that there is a rapid growth in the application of data mining in manufacturing, particularly in the semiconductor industry. However, there is still slow adoption of this technology in some manufacturing industries for several reasons including:

- In any particular knowledge area it is very difficult to determine what type of data mining function should be performed.
- In many cases a number of data mining techniques are possible, but which technique should be used, or which one is most appropriate?

To address these challenges, a novel text mining based approach has been applied on the reviewed literature's keywords and abstracts to identify patterns in the applications areas, techniques and data mining functions.

Benefits of applying text mining on keyword and abstracts of reviewed literature

The objectives of this text mining are as follows:

- to automate the process of finding a research gap among the reviewed literature.
- to identify any overlooked and under examined areas.
- to extract common patterns of good practices for data mining applications in manufacturing.
- to find a linkage between knowledge area, data mining function, and data mining technique used.
- to find key features which are novel for data mining users in manufacturing.

The next section discusses the text mining experiments undertaken using the abstract and keywords of the 150 published works reviewed in this paper.

Knowledge discovery in text and text mining applications on the literature review

Following the definition of KDD by Fayyad et al. (1996a), Karanikas and Theodoulidis (2002) defined KDT as "the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in unstructured data". Text Mining (TM) is also a step in the KDT process consisting of particular data mining and natural language processing algorithms that under certain computational efficiency and limitations produce a particular enumeration of patterns over a set of unstructured textual data. KDT in reviewed literature mainly consists of three steps as follows:

1. Abstract and keyword collection: In our experiments, the abstracts and key words of the literature reviewed in this paper have been collected. Where necessary, additional key words have also been identified from the papers and added to the abstract for text mining. This is important as the published abstracts often did not include full details of the type of data mining function (s) and areas of application discussed in the paper.
2. Retrieving and pre-processing documents: Abstracts have only been taken from papers which deploy data mining methodology to solve problems of manufacturing. The additional key words have been identified based on knowledge area, function performed and technique used. The major knowledge areas examined include manufacturing system, quality control, fault diagnosis,

maintenance, job shop, yield improvement, manufacturing process, fault diagnosis, product design, production control, and supply chain management. Similarly, the functions considered include concept description, classification, clustering, prediction and association. Major techniques used include rough set theory, decision tree, statistics, neural network, association rule, fuzzy c means clustering, regression analysis and hybrid algorithms. In this context, the term “hybrid algorithm” indicates that either a group of algorithms have been used in combination to solve a particular problem, or a group of algorithms have been used at different stages of data mining.

3. Text mining: For the current purpose, text analysis and link analysis were used to extract patterns, trends, useful knowledge and meet the listed benefits. The text mining was performed as an automatic process with manual interventions during the pre-processing stage. Polyanalyst, which is one of the leading data/text mining software package in the market was used for this purpose. All the results shown and interpretations made were automatically generated using this software. The following subsections describe how the abovementioned objectives were achieved.

Automated process and identification of the research gaps

Our experiments tried to show if text mining could be used to automate the process of searching for relevant literature in a particular topic area and subsequently be used to identify research gaps. Figure 2 schematically represents the extracted keywords in the form of rules. These show that very little work has been carried out in the area of supply chain management and logistics with only five identified examples being found within the 150 papers. Barcode and Radio frequency identification (RFID) are currently the key to supply chain infrastructure. Since these techniques generate a huge amount of data they potentially provide sources of useful data for data mining based intelligent decision support systems. However no examples of these tools being used in data mining contexts were found hence, it could therefore be argued that the text mining has identified a potential research gap. In addition, data mining could be used as an integral part of supply chain planning and optimization software to analyze data and transactions to give managers a range of decision choices. Other areas with low application rates of data mining and potential research gaps identified include integration of data mining with ERP systems, production planning and control. In the area of distributed manufacturing, data mining can be integrated with agent technology to create a productive data mining system.

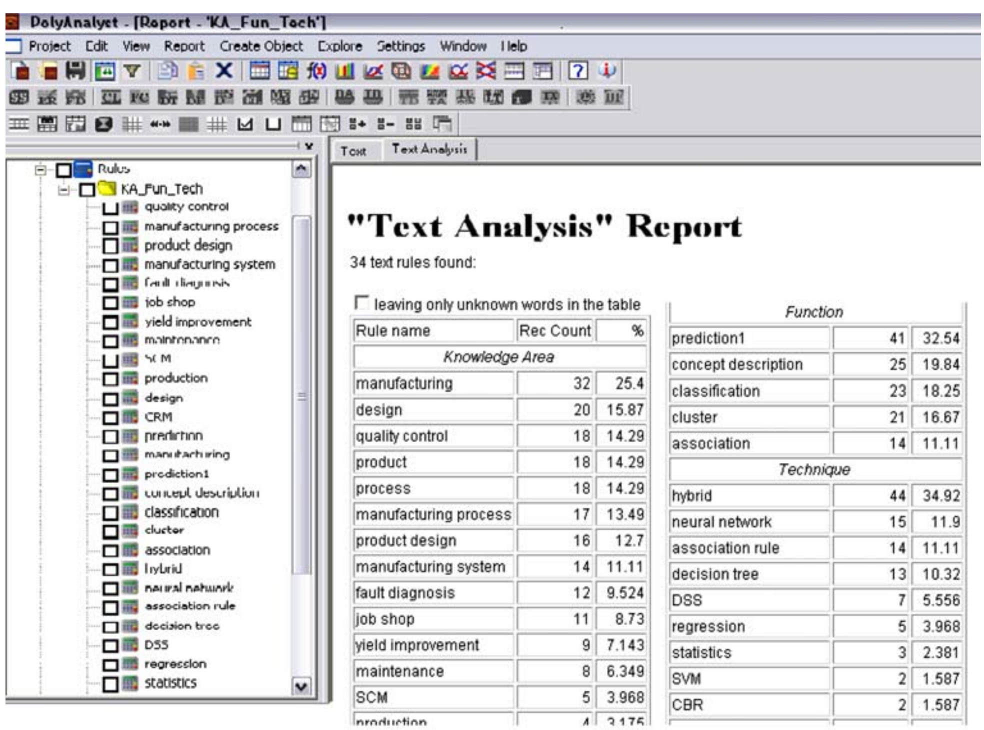

Fig. 2 Text analysis of reviewed literature and generated rules

Identifying the overlooked and under examined areas

Figure 2 shows the number of papers published in various areas of manufacturing. It shows that several pieces of research work have been carried out in the areas of data mining applications in quality control, manufacturing processes, product design, fault diagnosis, maintenance, manufacturing systems, job shop type problems, and yield improvement of semiconductor industries. The reasons for this may be that large volumes of data are generated during manufacture and these are therefore easily available for use. Another important aspect is that most of the work has been carried out in the semiconductor industries where small improvements can have a significant impact on cost. The figure also clearly shows areas where little research has been carried out including supply chain management, production planning and distributed manufacturing (agent technology). This may mainly be due to problems encountered during integration issues of the existing system with a data mining system. Less work has been carried out in performing the "association" function than in other data mining functions in manufacturing as shown in Fig. 2.

Extracting common patterns of good practices

The TM experiments also showed that techniques like neural network, regression analysis, rough set theory and decision tree are mostly used for prediction purposes in manufacturing. For clustering purposes, hybrid algorithms and fuzzy c means clustering have mostly been used. Most of the hybrid algorithms include neural network as a primary tool. Similarly for classification purposes hybrids of neural network, or rough set theory have been used. However, association rule mining has also proved to be an effective tool for identifying the associations in manufacturing. These results are represented in the link analysis as shown in Fig. 3. This pictorially shows the linkage between patterns found, where the bolder lines show greater support in the links.

Finding a linkage between manufacturing knowledge area, data mining function, and data mining technique used

One of the major benefits of the application of text mining on the reviewed literature is to establish a linkage between various knowledge areas, data mining techniques and functions performed using the link term analysis tool of the text mining software. This visually represents the relations between key words in the textual abstract. Results of the link terms analysis are presented as a graph displaying a cluster of linked objects supporting various object manipulations and drill down operations. This mechanism provides the quickest way to understand the most prominent semantic characteristics of the explored data. Figure 4 shows the linkages between various manufacturing areas and data mining techniques used to solve the problem. Again, the bolder lines show stronger links (or greater use).

It can be seen from Fig. 4 that some of the techniques have been more frequently used in a particular type of manufacturing area. For example, neural network have been most commonly applied to analyse the manufacturing process related data. Similarly, association rule mining has been applied more frequently to analyse product design related data.

Figure 5 shows the linkages between various manufacturing areas and data mining functions performed. It is evident from the figure that certain types of function are more frequently applied on a particular manufacturing area. These include, prediction in manufacturing processes, maintenance and job shop, association in product design, concept description for job shop type problems, classification in quality control. Figure 6 represents the linkages between manufacturing area, function and techniques used. These results can be used to prioritize the data mining functions and techniques for a particular manufacturing area. For example, Fig. 6 shows that that there is deep link between manufacturing process, prediction and neural network. Therefore, neural networks can be applied as one of the several useful tools for prediction purposes in manufacturing processes. Similarly, to identify association type relationships in product design related problems, priority should be given to association rule mining.

Finding novel features for data mining users in manufacturing

The text mining experiments also enabled some of the novel features which were commonly used by data mining practitioners in manufacturing to be identified. One of them is

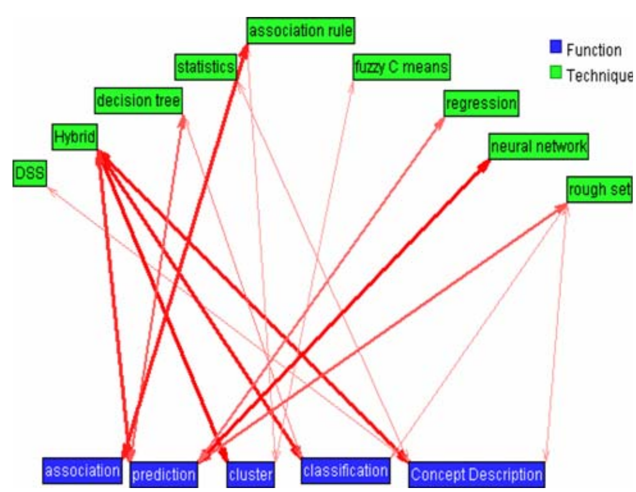

Fig. 3 Linkage and correlation between data mining functions and techniques

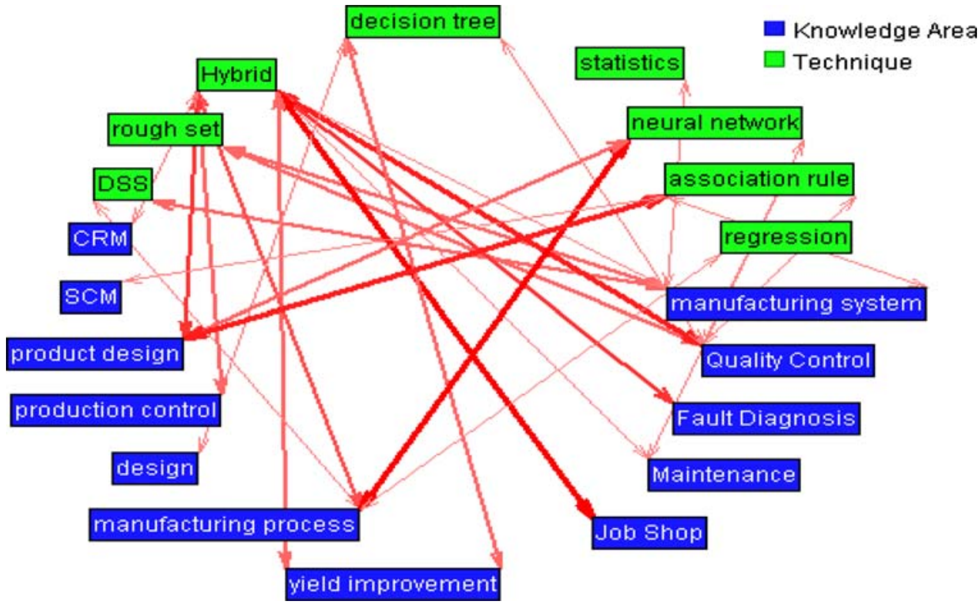

Fig. 4 Linkage and correlation between manufacturing area and data mining techniques

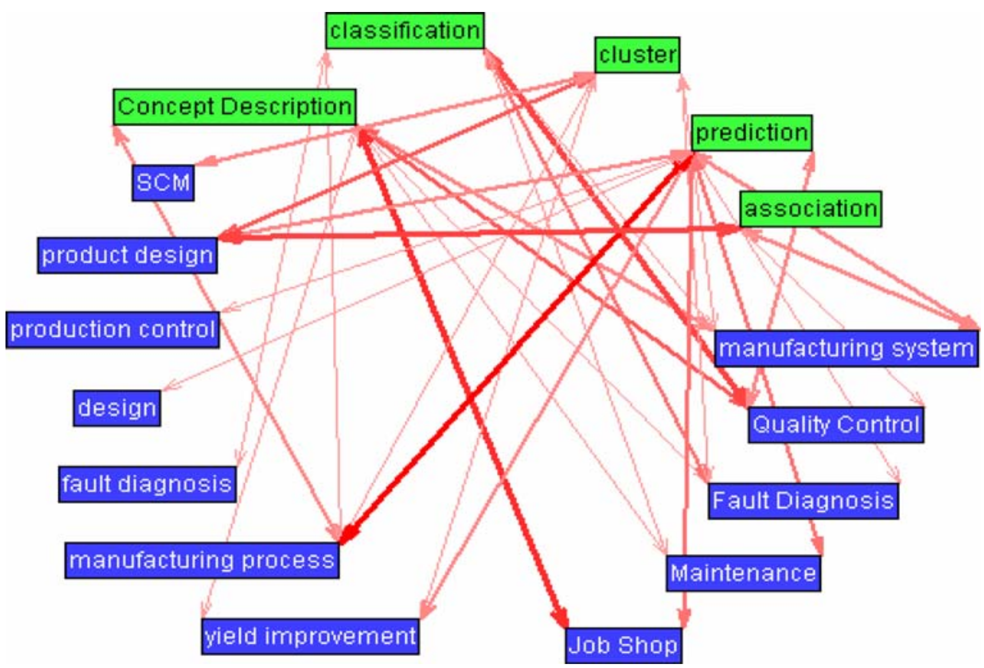

Fig. 5 Correlation and linkage between data mining functions and manufacturing area

the increased use of hybrid algorithms in manufacturing. There is no universally best data mining method for all manufacturing contexts. Therefore, there is an increased used of combinations of traditional data mining algorithms to capture advantages from each technique used. It has been found that in order to perform several functions in a problem area several algorithms may be used at different stages of data mining. Another noticeable factor is the use of fuzzy sets with various data mining techniques. One of the reason for this may be, that fuzzy sets are inherently suitable for coping with linguistic domain knowledge and producing more interpretable solutions. In addition, when fuzzy sets are used in association with a different data mining technology, they are capable of handling the issues related to understadability of patterns, incomplete and noisy data, and human interaction, and can provide approximate solutions faster. It has been found that hybridization of neural network and fuzzy logic is of increased use in manufacturing. In addition,

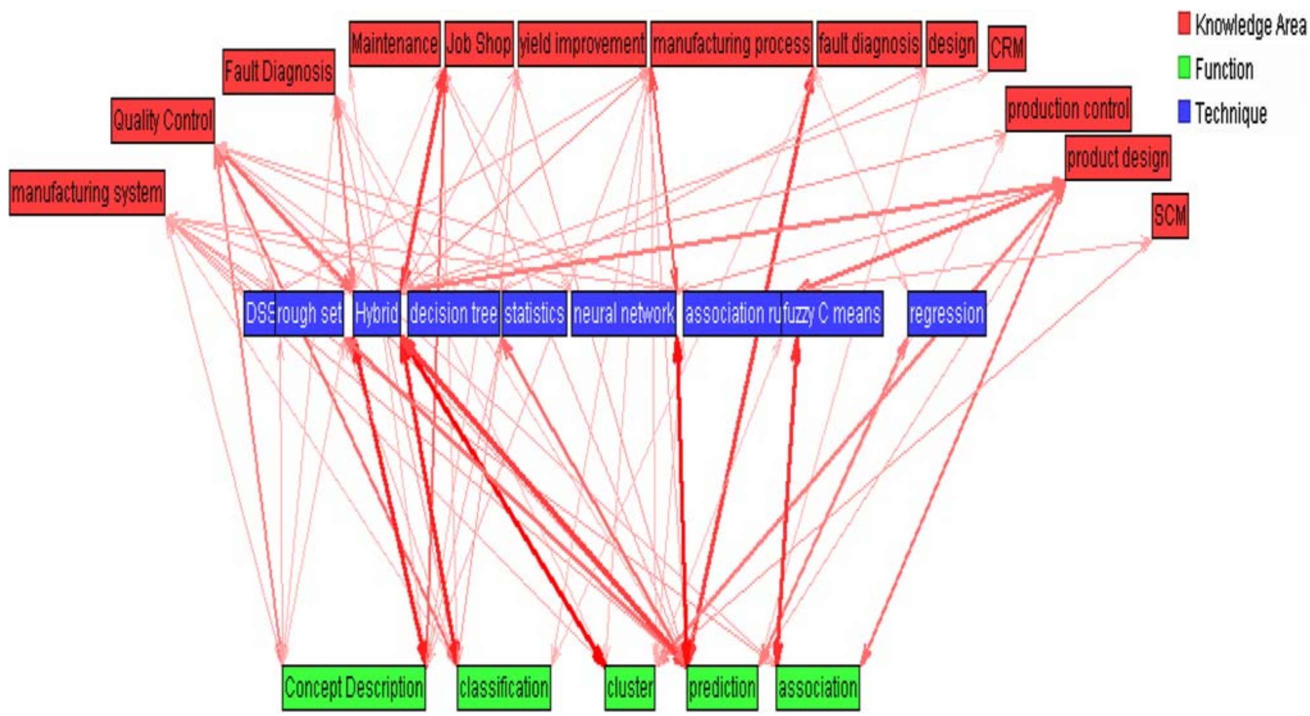

Fig. 6 Linkages between knowledge area, knowledge mined and techniques used

researchers and practitioners have also modelled, or advanced the existing algorithms according to the problem environment. Thus, it can be said that hybridization of data mining algorithms may lead to a better solution quality than existing and traditional algorithms.

Conclusion

Knowledge discovery and data mining have created new intelligent tools for extracting useful information and knowledge automatically from manufacturing databases. The present article provides a survey of the available literature on data mining applications in manufacturing with a special emphasis on the kind of knowledge mined. The types of knowledge identified indicates the major data mining functions to be performed include characterization and description, association, classification, prediction, clustering in data. It has also been found that there is an exponential growth in the number of papers in last 3 years.

A novel text mining approach has been applied on the reviewed literature to identify the popular and successful research tools and existing research gaps, examine the under looked and overlooked areas, identify good practices in data mining in manufacturing and some key features unknown to data mining practitioners. A correlation and linkage has been established between knowledge area, knowledge mined and techniques used based on the past practices to provide the user with appropriate choices of technique. It has been found that some areas such as supply chain management, production planning and control and integration of data mining systems with manufacturing systems need attention from the research community.

References

Agard, B., & Kusiak, A. (2004a). Data mining for subassembly selection. *Journal of Manufacturing Science and Engineering, 126*, 627–631. doi:10.1115/1.1763182.

Agard, B., & Kusiak, A. (2004b). Data-mining based methodology for the design of product families. *International Journal of Production Research, 42*, 15, 2955–2969.

Agrawal, R., Imielinski, T., & Swami, A. N. (1993). Mining association rules between sets of items in large databases. In *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data* (pp. 207–216). Washington, D.C., May 1993.

Backus, P., Janakiram, M., Mowzoon, S., Runger, G. C., & Bhargava, A. (2006). Factory cycle time prediction with a data-mining approach. *IEEE Transactions on Semiconductor Manufacturing, 19*(2). doi:10.1109/TSM.2006.873400.

Batanov, D., Nagarur, N., & Nitikhumkasem, P. (1993). Expert—MM: A knowledge based system for maintenance management. *Artificial Intelligence in Engineering, 8*, 283–291. doi:10.1016/0954-1810(93)90012-5.

Belz, R., & Mertens, P. (1996). Combining knowledge based systems and simulation to solve rescheduling problems. *Decision Support Systems, 17*, 141–157. doi:10.1016/0167-9236(95)00029-1.

Bergeret, F., & Gall, C. L. (2003). Yield improvement using statistical analysis of process dates. *IEEE Transactions on Semiconductor Manufacturing, 16*(3), 535–541. doi:10.1109/TSM.2003.815204.

Braha, D., & Shmilovici, A. (2002). Data mining for improving a cleaning process in the semiconductor Industry. *IEEE Transactions on Semiconductor Manufacturing, 15*(1), 91–101. doi:10.1109/66.983448.

Browne, W., Yao, L., Postlethwaite, I., Lowes, S., & Mar, M. (2006). Knowledge elicitation and data mining: Fusing human and industrial plant information. *Engineering Applications of Artificial Intelligence, 19*, 345–359. doi:10.1016/j.engappai.2005.09.005.

Buddhakulsomsiri, J., Siradeghyan, Y., Zakarian, A., & Li, X. (2006). Association rule generation algorithm for mining automotive warranty data. *International Journal of Production Research, 44*(14), 2749–2770. doi:10.1080/00207540600564633.

Busse, J. W. G., Stefanowski, J., & Wilk, S. (2005). A comparison of two approaches to data mining from imbalanced data. *Journal of Intelligent Manufacturing, 16*, 565–573. doi:10.1007/s10845-005-4362-2.

Caramia, M., & Felici, G. (2006). Mining relevant information on the web: A clique based approach. *International Journal of Production Research, 44*(14), 2771–2787. doi:10.1080/00207540600693713.

Caskey, K. R. (2001). A manufacturing problem solving environment combining evaluation, search and generalisation methods. *Computers in Industry, 44*, 175–187. doi:10.1016/S0166-3615(00)00072-5.

Chang, P. C., Hieh, J. C., & Liao, T. W. (2005). Evolving fuzzy rules for due date assignment problem in semiconductor manufacturing factory. *Journal of Intelligent Manufacturing, 16*, 49–557. doi:10.1007/s10845-005-1663-4.

Chao, K. M., Guenov, M., Hills, B., Smith, P., Buxton, I., & Tsai, C. F. (1997). An expert system to generate associativity data for layout design. *Artificial Intelligence in Engineering, 11*, 191–196. doi:10.1016/S0954-1810(96)00037-4.

Chen, M. C. (2003). Configuration of cellular manufacturing systems using association rule induction. *International Journal of Production Research, 41*(2), 381–395. doi:10.1080/0020754021000024184.

Chen, M. C., & Wu, H. P. (2005). An association based clustering approach to order batching considering customer demand pattern. *The International journal of Management Science, 33*, 333–343.

Chen, N., Zhu, D. D., & Wang, W. (2000). Intelligent material processing by hyper space data mining. *Engineering Applications of Artificial Intelligence, 13*, 527–532. doi:10.1016/S0952-1976(00)00032-4.

Chen, W. C., Tseng, S. S., & Wang, C. Y. (2005). A novel manufacturing defect detection method using association rule mining techniques. *Expert Systems with Applications, 29*, 807–815. doi:10.1016/j.eswa.2005.06.004.

Chen, M. C., Huang, C. L., Chen, K. Y., & Wu, H. P. (2005). Aggregation of orders in distribution centres using data mining. *Expert system with application, 28*, 453–460.

Chen, F. C., Chih, W. H., Cheng, M., Kuo, T. H., & Szu, T. W. (2005). Cycle time prediction and control based on production line status and manufacturing data mining. In *IEEE International Symposium on Semiconductor Manufacturing, ISSM* (pp. 327–330), 13–15 September 2005.

Chien, C. F., Wang, W. C., & Chang, J. C. (2007). Data mining for yield enhancement in semiconductor manufacturing and an empirical study. *Expert System With Applications, 33*(1), 192–198.

Choudhary, A. K., Harding, J. A., & Popplewell, K. (2006). Knowledge discovery for moderating collaborative projects. In *Proceedings of the 4th IEEE International Conference on Industrial Informatics 2006 (INDIN '06)* (pp. 519–524). Singapore, August 2006.

Choudhary, A. K., Harding, J. A., & Lin, H. K. (2007). Engineering moderator to universal knowledge moderator for moderating collaborative projects. In *Proceedings of GLOGIFT 07* (pp. 529–537), 15–17 November 2007, UP Technical University, Noida.

Crespo, F., & Webere, R. (2005). A methodology for dynamic data mining based on fuzzy clustering. *Fuzzy Sets and Systems, 150*, 267–284. doi:10.1016/j.fss.2004.03.028.

Cunha, D., Agard, B., & Kusiak, A. (2006). Data mining for improvement of product quality. *International Journal of Production Research, 44*(18–19), 4027–4041. doi:10.1080/00207540600678904.

Dabbas, R. M., & Chen, H. N. (2001). Mining semiconductor manufacturing data for productivity improvement—An integrated relational database approach. *Computers in Industry, 45*, 29–44. doi:10.1016/S0166-3615(01)00079-3.

Dengiz, O., Smith, A. E., & Nettleship, I. (2006). Two stage data-mining for flaw identification in ceramics manufacturing. *International Journal of Production Research, 44*(14), 2839–2851. doi:10.1080/00207540500534454.

Elovici, Y., & Braha, D. (2003). A decision theoretic approach to data mining. *IEEE Transactions on Systems, Man, and Cybernetics. Part A, Systems and Humans, 33*(1), 42–51. doi:10.1109/TSMCA.2003.812596.

Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P., & Uthuruswamy, R. (Eds.). (1996a). *Advances in knowledge discovery and data mining*. Menlo Park, CA: AAAI/MIT Press.

Fayyad, U., Shapiro, G. P., & Smyth, P. (1996b). The KDD process for extracting useful knowledge from volume of data. *Communications of the ACM, 39*, 27–34. doi:10.1145/240455.240464.

Feng, C.-X. J., & Wang, X. F. (2004). Data mining technique applied to predictive modelling of the knurling process. *IIE Transactions, 36*, 253–263. doi:10.1080/07408170490274214.

Feng, J. C. X., & Kusiak, A., (Eds.). (2006). Special issue on data mining and applications in engineering design, manufacturing and logistics. *International Journal of Production Research, 44*(14), 2689–2694.

Feng, C. X. J., Yu, Z. G., & Kusiak, A. (2006). Selection and validation of predictive regression and neural network models based on designed experiments. *IIE Transactions, 38*, 13–23. doi:10.1080/07408170500346378.

Fountain, T., Dietterich, T., & Sudykya, B. (2003). Data mining for manufacturing control: An application in optimizing IC test. In G. Lakemeyer & B. Nebel (Eds.), *Exploring artificial intelligence in the new millennium* (pp. 381–400). San Francisco, CA: Morgan Kauffmann Publisher Inc.

Gardner, M., & Bieker, J. (2000). Data mining solves tough semi conductor problems (pp. 376–383). Boston: KDD 2000.

Gertosio, C., & Dussauchoy, A. (2004). Knowledge discovery from industrial databases. *Journal of Intelligent Manufacturing, 15*, 29–37. doi:10.1023/B:JIMS.0000010073.54241.e7.

Giess, M. D., & Culley, S. J. (2003). Investigating manufacturing data for use within design. *ICED 03*, Stockholm, Sweden, p. 14.

Giess, M. D., Culley, S. J., & Shepherd, A. (2002). *Informing design using data mining methods* (pp. 98–106). Montreal, Canada: ASME DETC.

Guh, R. S. (2005a). A hybrid learning based model for online detection and analysis of control chart patterns. *Computers & Industrial Engineering, 49*, 35–62. doi:10.1016/j.cie.2005.03.002.

Guh, R. S. (2005b). Real time pattern recognition in statistical process control: A hybrid neural network/decision tree- based approach. *Proceedings of the Institution of Mechanical Engineers. Journal of Engineering Manufacture: Part B, 219*, 283–298. doi:10.1243/095440505X28963.

Han, J., & Kamber, M. (2001). *Data mining concepts and techniques*. Morgan Kaufmann Publisher.

Harding, J. A., & Popplewell, K. (2006). Knowledge reuse and sharing through data mining manufacturing data. In *Proceedings of IERC 2006, Industrial Engineering Research Conference, The Institute of Industrial Engineers*, Orlando, May 2006, 6 pp, [CD-ROM and WWW], Available from: http://www.iieannual.org,paperIIE06/Research/1038.pdf.

Harding, J. A., Shahbaz, M., Srinivas, & Kusiak, A. (2006). Data mining in manufacturing: A review. *American Society of Mechanical Engineers (ASME). Journal of Manufacturing Science and Engineering, 128*(4), 969–976. doi:10.1115/1.2194554.

Hipp, J., Guntzer, U., & Nakhaeizadeh, G. (2000). Algorithms for association rule mining—A general survey and comparison. *ACM SIG KDD Explorations, 2*(1), 58–64.

Ho, G. T. S., Lau, H. C. W., Lee, C. K. M., Ip, A. W. H., & Pun, K. F. (2006). An intelligent production workflow mining system for continual quality enhancement. *International Journal of Advanced Manufacturing Technology, 28*, 792–809. doi:10.1007/s00170-004-2416-9.

Holden, T., & Serearuno, M. (2005). A hybrid artificial intelligence approach for improving yield in precious stone manufacturing. *Journal of Intelligent Manufacturing, 16*, 21–38. doi:10.1007/s10845-005-4822-8.

Horng, S. C., & Lin, S. Y. (2004). A hybrid classification tree for products of complicated machines in flexible manufacturing system. *IEEE International Conference on System Man and Cybernetics, 4*, 3775–3780.

Hou, T. H., & Huang, C. C. (2004). Application of fuzzy logic and variable precision rough set approach in a remote monitoring manufacturing process for diagnosis rule induction. *Journal of Intelligent Manufacturing, 15*, 395–408. doi:10.1023/B:JIMS.0000026576.00445.d8.

Hou, J. L., & Yang, S. T. (2006). Technology mining model concerning operation characteristics of technology and service providers. *International Journal of Production Research, 44*(14), 3345–3365. doi:10.1080/00207540500490996.

Hou, T. S., Liu, W. L., & Lin, L. (2003). Intelligent remote monitoring and diagnosis of manufacturing process using an integrate approach of neural networks and rough sets. *Journal of Intelligent Manufacturing, 14*, 239–253. doi:10.1023/A:1022911715996.

Hsieh, S.-J. (2004). Artificial neural networks and statistical modelling for electronic stress prediction using thermal profiling. *IEEE Transactions on Electronics Packaging Manufacturing, 27*(1), 49–58. doi:10.1109/TEPM.2004.830517.

Hsu, H. C., & Wang, M. J. J. (2005). Using decision tree based data mining to establish sizing system for the manufacture of garments. *International Journal of Advanced Manufacturing Technology, 26*, 669–674. doi:10.1007/s00170-003-2032-0.

Huang, C. L., Li, T. S., & Peng, T. K. (2005). A hybrid approach of rough set theory and genetic algorithm for fault diagnosis. *International Journal of Advanced Manufacturing Technology, 27*, 119–127. doi:10.1007/s00170-004-2142-3.

Hui, S. C., & Jha, G. (2000). Data mining for customer service support. *Information & Management, 38*, 1–13. doi:10.1016/S0378-7206(00)00051-3.

Huyet, A. L. (2006). Optimization and analysis aid via data-mining for simulated production system. *European Journal of Operational Research, 173*, 827–838. doi:10.1016/j.ejor.2005.07.026.

Inada, M., & Teraano, T. (2005). QC chart mining: Extracting systematic error patterns from quality control charts. *IEEE International Conference on System Man and cybernetics, 4*, 3781–3787.

Irani, K. B., Cheng, J., Fayyad, U. M., & Qian, Z. (1993). Applying machine learning to semiconductor manufacturing. *IEEE Expert, 8*(1), 41–47. doi:10.1109/64.193054.

Jeong, M. K., Lu, J.-C., Huo, X., Vidakovic, B., & Chen, D. (2006). Wavelet-Based data reduction techniques for process fault detection. *Technometrics, 48*(1), 26–40. doi:10.1198/004017005000000553.

Jiao, J., & Zhang, Y. (2005). Product portfolio identification based on association rule mining. *Computer Aided Design, 37*, 149–172. doi:10.1016/j.cad.2004.05.006.

Jiao, J., Zhang, L., Zhang, Y., & Pokharel, S. (2008). Association rule mining for product and process variety mapping. *International Journal of Computer Integrated Manufacturing, 21*(1), 111–124. doi:10.1080/09511920601182209.

Jin, Y., & Ishino, Y. (2006). DAKA: Design activity knowledge acquisition through data mining. *International Journal of Production Research, 44*(15), 2813–2837. doi:10.1080/00207540600654533.

Jung, U. K., Jeong, M. K., & Lu, J. C. (2006). Data reduction for multiple functional data with class information. *International Journal of Production Research, 44*(14), 2695–2710. doi:10.1080/00207540500539784.

Kang, B.-S., Lee, J.-H., Shin, C.-K., Yu, S.-J., & Park, S.-C. (1998). Hybrid machine learning system for integrated yield management in semiconductor manufacturing. *Expert System With Application, 15*, 123–132. doi:10.1016/S0957-4174(98)00017-7.

Karanikas, H., & Theodoulidis, B. (2002). Knowledge discovery in text and text mining software. *Technical Report, UMIST-CRIM*, Manchester.

Kim, P., & Ding, Y. (2005). Optimal engineering system design guided by data-mining methods. *Technometrics, 47*(3), 336–348. doi:10.1198/004017005000000157.

Kim, S. H., & Lee, C. M. (1997). Non linear prediction of Manufacturing systems through explicit and implicit data mining. *Computers and Industrial Engineering, 33*(3–4), 461–464.

Koonce, D. A., & Tsai, S. C. (2000). Using data mining to find patterns in genetic algorithm solutions to a job shop schedule. *Computers & Industrial Engineering, 38*, 361–374. doi:10.1016/S0360-8352(00)00050-4.

Kusiak, A. (2000a). Data analysis: Models and algorithms. In P. E. Orban & G. K. Knopf (Eds.), *Proceeding of the SPIE conference on intelligent systems and advanced manufacturing* (Vol. 4191, pp. 1–9). Boston, MA: SPIE.

Kusiak, A. (2000b). *Computational intelligence in design and manufacturing*. New York: John Wiley.

Kusiak, A. (2000c). Decomposition in data mining: An industrial case study. *IEEE Transactions on Electronics Packaging Manufacturing, 23*(4), 345–353. doi:10.1109/6104.895081.

Kusiak, A. (2001a). Feature transformation methods in data mining. *IEEE Transactions on Electronics packaging manufacturing, 24*, (3), 214–221.

Kusiak, A. (2001b). Rough set theory: A data mining tool for semiconductor manufacturing. *IEEE Transactions on Electronics Packaging Manufacturing, 24*(1), 44–50. doi:10.1109/6104.924792.

Kusiak, A. (2002a). Data mining and decision making. In *SPIE Conference on Data Mining and Knowledge Discovery: Theory, Tools and Technology IV* (pp. 155–165). Orlando, FL, 2002.

Kusiak, A. (2002b). A data mining approach for generation of control signatures. *Journal of Manufacturing Science and Engineering, 124*, 923–926. doi:10.1115/1.1511524.

Kusiak, A. (2005). Selection of invariant Objects With a data-mining approach. *IEEE Transactions on Electronics Packaging Manufacturing, 28*(2), 187–196. doi:10.1109/TEPM.2005.846832.

Kusiak, A., & Kurasek, C. (2001). Data mining of printed circuit board defects. *IEEE Transactions on Robotics and Automation, 17*(2), 191–196. doi:10.1109/70.928564.

Kusiak, A., & Shah, S. (2006). Data-Mining based system for prediction of water chemistry faults. *IEEE Transactions on Industrial Electronics, 15*(2), 593–603. doi:10.1109/TIE.2006.870706.

Kusiak, A., Kernstine, K. H., Kern, J. A., Mclaughlin, K. A., & Tseng, T. L. (2000). Data mining: Medical and engineering case studies. In *Industrial Engineering Research Conference* (pp. 1–7). Ohio, Cleveland.

Kwak, C., & Yih, Y. (2004). Data mining approach to production control in the computer integrated testing cell. *IEEE Transactions on Robotics and Automation, 20*(1), 107–116. doi:10.1109/TRA.2003.819595.

Last, M., & Kandel, A. (2004). Discovering useful and understandable pattern in manufacturing data. *Robotics and Autonomous Systems, 49*, 137–152. doi:10.1016/j.robot.2004.09.002.

Lee, S. G., & Ng, Y. C. (2006). Hybrid case-based reasoning for on-line product fault diagnosis. *International Journal of Advanced Manufacturing Technology, 27*, 823–840.

Lee, J. H., Yu, S. J., & Park, S. C. (2001). Design of intelligent sampling methodology based on data mining. *IEEE Transactions on Robotics and Automation, 17*(5), 637–648. doi:10.1109/70.964664.

Li, X., & Olafsson, S. (2005). Discovering dispatching rules using data mining. *Journal of Scheduling, 8*, 515–527. doi:10.1007/s10951-005-4781-0.

Li, D. C., Wu, C. S., Tsai, T. I., & Chang, F. M. (2006). Using mega-fuzzification and data trend estimation in small set learning for early scheduling knowledge. *Computers & Operations Research, 33*, 1857–1869. doi:10.1016/j.cor.2004.11.022.

Li, F., Runger, G. C., & Eugene, T. (2006). Supervised learning for change point detection. *International Journal of Production Research, 44*(14), 2853–2868. doi:10.1080/00207540600669846.

Li, J. R., Khoo, L. P., & Tor, S. B. (2006). RMINE: A rough set based data mining prototype for the reasoning of incomplete data in condition based fault diagnosis. *Journal of Intelligent Manufacturing, 17*, 163–176. doi:10.1007/s10845-005-5519-8.

Li, T. S., Huang, C. L., & Wu, Z. Y. (2006). Data mining using genetic programming for construction of a semiconductor manufacturing yield rate prediction system. *Journal of Intelligent Manufacturing, 17*, 355–361. doi:10.1007/s10845-005-0008-7.

Liao, S. H., & Wen, C. H. (2007). Artificial neural networks classification and clustering of methodologies and applications- literature analysis from 1995–2005. *Experts Systems With Applications, 32*(1), 1–11.

Liao, T. W., Li, D. M., & Li, Y. M. (1999). Detection of welding flaws from radiographic images with fuzzy clustering methods. *Fuzzy Sets and Systems, 108*, 145–158. doi:10.1016/S0165-0114(97)00307-2.

Liao, T. W., Ting, C. F., & Chang, P. C. (2006). An adaptive genetic clustering method for exploratory mining of feature vector and time series data. *International Journal of Production Research, 44*(15), 2731–2748. doi:10.1080/00207540600600130.

Liao, T. W., Wang, G., Triantaphyllou, E., & Chang, P. C. (2001). A data mining study of weld quality models constructed with MLP neural networks from stratified sample data. In *Industial Engineering Research Conference* (p. 6). Dallas, Tx.

Lin, C. C., & Tseng, Y. H. (2005). A neural network application for reliability modelling and condition-based predictive maintenance. *International Journal of Advance Manufacturing Technology, 25*, 174–179.

Liu, Y.-H., Huang, H. P., & Lin, Y. S. (2005). Attribute selection for the scheduling of flexible manufacturing systems based on fuzzy set theoretic approach and genetic algorithm. *Journal of the Chinese Institute of Industrial Engineers, 22*(1), 46–55.

Macgarry, K. (2005). A survey of interestingness measures for knowledge discovery. *The Knowledge Engineering Review, 20*(1), 39–61. doi:10.1017/S0269888905000408.

Maki, H., & Teranishi, Y. (2001). Development of automated data mining system for quality control in manufacturing. *Lecture Notes in Computer Science*, Springer-Verlag: Berline, Vol. 2114, pp. 93–100.

Maki, H., Maeda, A., Morita, T., & Akimori, H. (2000). Applying data mining to data analysis in manufacturing. *International Conference on Advances in Production Management Systems* (pp. 324–331). Berlin, Germany: Kluwer Academic Publishers; MA, USA: Norwell.

McDonald, C. J. (1999). New tools for yield improvement in integrated circuit manufacturing: Can they be applied to reliability? *Microelectronics Reliability, 39*(6–7), 731–739.

Menon, R., Tong, L. H., & Sathiyakeerthi, S. (2005). Analyzing textual databases using data mining to enable fast product development process. *Reliability Engineering & System Safety, 88*, 171–180. doi:10.1016/j.ress.2004.07.007.

Mere, J. B. O., Marcos, A. G., Gonzalez, J. A., & Rubio, V. L. (2004). Estimation of mechanical properties of steel strip in hot dip galvanising lines. *Iron making and Steel making, 31*(1), 43–50.

Mitra, S., Pal, S. K., & Mitra, P. (2002). Data mining in soft computing framework: A survey. *IEEE Transactions on Neural Networks, 13*(1), 3–14. doi:10.1109/72.977258.

Morita, T., Sato, Y., Ayukawa, E., & Maeda, A. (2000). Customer relationship management through data mining. Informs-Korms2000, Seoul, (pp. 1959–1963).

Murthy, S. K. (1998). Automatic construction of decision trees from data: A multidisciplinary survey. *Data Mining and Knowledge Discovery, 2*, 345–389. doi:10.1023/A:1009744630224.

Neaga, E. I., & Harding, J. A. (2005). An enterprise modelling and integration framework based on knowledge discovery and data mining. *International Journal of Production Research, 43*(6), 1089–1108. doi:10.1080/00207540412331322939.

Ozturk, A., Kayaligil, S., & Ozdemirel, N. E. (2006). Manufacturing Lead time estimation using data mining. *European Journal of Operational Research, 73*, 683–700. doi:10.1016/j.ejor.2005.03.015.

Park, J.-H., & Seo, K.-K. (2006). A knowledge based approximate life cycle assessment system for evaluating environmental impacts of product design alternatives in a collaborative design environment. *Advanced Engineering Informatics, 20*, 147–154.

Pasek, Z. J. (2006). Exploration of rough sets theory use for manufacturing process monitoring. *Proceedings of the Institution of Mechanical Engineers. Journal of Engineering Manufacture: Part B, 220*, 365–373. doi:10.1243/095440505X69337.

Peng, Y. (2004). Intelligent condition monitoring using fuzzy inductive learning. *Journal of Intelligent Manufacturing, 15*, 373–380. doi:10.1023/B:JIMS.0000026574.95637.36.

Pham, D. T., & Afify, A. A. (2005). Machine learning techniques and their applications in manufacturing. *Proceedings of the Institution of Mechanical Engineers. Journal of Engineering Manufacture: Part B, 219*, 395–412. doi:10.1243/095440505X32274.

Purintrapiban, U., & Kachitvichyanukul, V. (2003). Detecting patterns in process data with fractal dimension. *Computers & Industrial Engineering, 45*, 653–667. doi:10.1016/j.cie.2003.09.004.

Qian, Z., Jiang, W., & Tsui, K. L. (2006). Churn detection via customer profile modelling. *International Journal of Production Research, 44*(4), 2913–2933. doi:10.1080/00207540600632240.

Rajagopal, R., & Castillo, E. (2006). A bayesian method for Robust tolerance control. *IIE Transactions, 38*, 685–697. doi:10.1080/07408170600692283.

Ren, Y., Ding, Y., & Zhou, S. (2006). A data mining approach to study the significance of nonlinearity in multi-station assembly processes. *IIE Transactions, 38*, 1069–1083. doi:10.1080/07408170600735538.

Rojas, A., & Nandi, A. K. (2006). Practical scheme for fast detection and classification of rolling element bearing faults using support vector method. *Mechanical Systems and Signal Processing, 20*, 1523–1536. doi:10.1016/j.ymssp.2005.05.002.

Rokach, L., & Maimon, O. (2006). Data mining for improving the quality of manufacturing: A feature set decomposition approach. *Journal of Intelligent Manufacturing, 17*, 285–299. doi:10.1007/s10845-005-0005-x.

Romanowski, C. J., & Nagi, R. (1999). Improving preventive maintenance scheduling using data mining techniques. In *8th Industrial Engineering Research Conference*, Phoenix AZ, May 1999.

Romanowski, C. J., & Nagi, R. (2001). A data mining for knowledge acquisition in engineering design: A research agenda. In D. Braha (Ed.), *Data mining for design and manufacture: Methods and applications* (pp. 161–178). Dordrecht: Kluwer Academic Publisher.

Romanowski, C. J., & Nagi, R. (2004). A data mining approach to forming generic bills of material in support of variant design

activities. *ASME Journal of Computing and Information Science in Engineering, 4*(4), 316–328. doi:10.1115/1.1812556.
Romanowski, C. J., & Nagi, R. (2005). On comparing bills of materials: A similarity/distance measure for unordered trees. *IEEE Transactions on System Man and Cybernetics Part A, 35*(2), 249–260.
Sebzalli, Y. M., & Wang, X. Z. (2001). Knowledge discovery from process operational data using PCA and fuzzy clustering. *Engineering Applications of Artificial Intelligence, 14*, 607–616. doi:10.1016/S0952-1976(01)00032-X.
Sha, D. Y., & Liu, C. H. (2005) Using data mining for due date assignment in a dynamic job shop environment. *International Journal of Advance Manufacturing Technology, 25*, 1164–1174.
Shahbaz, M., Srinivas, Harding, J. A., & Turner, M. (2006). Product design and manufacturing process improvement using association rules. *Proceedings of the Institution of Mechanical Engineers, Part B: Journal of Engineering Manufacture, 220*, 243–254.
Shao, X.-Y., Wang, Z.-H., Li, P.-G., & Feng, C.-X. J. (2006). Integrating data mining and rough set for customer group-based discovery of product configuration rules. *International Journal of Production Research, 44*(14), 2789–2811. doi:10.1080/00207540600675777.
Shen, L., Tay, F. E. H., Qu, L. S., & Shen, Y. (2000). Fault diagnosis using rough set theory. *Computers in Industry, 43*, 61–72. doi:10.1016/S0166-3615(00)00050-6.
Skormin, V. A., Gorodetski, V. I., & PopYack, I. J. (2002). Data mining technology for failure of prognostic of avionics. *IEEE Transactions on Aerospace and Electronics Systems, 38*(2), 388–403. doi:10.1109/TAES.2002.1008974.
Song, C., Guan, X., Zhao, Q., & Ho, Y.-C. (2005). Machine learning approach for determining feasible plan of a remanufacturing system. *IEEE Transactions on Automation Science and Engineering, 2*(3), 262–275. doi:10.1109/TASE.2005.849090.
Sun, J., Hong, G. S., Rahman, M., & Wong, Y. S. (2005). Improved performance evaluation of tool condition identification by manufacturing loss consideration. *International Journal of Production Research, 43*(6), 1185–1204. doi:10.1080/00207540412331299701.
Sylvain, L., Fazel, F., & Stan, M. (1999). Data mining to predict aircraft component replacement. *IEEE Intelligent Systems, 14*(6), 59–65.
Symeonidis, A. L., Kehagias, D. D., & Mitkas, P. A. (2003). Intelligent policy recommendations on enterprise resource planning by the use of agent technology and data mining techniques. *Expert Systems with Applications, 25*, 589–602. doi:10.1016/S0957-4174(03)00099-X.
Torkul, O., Cedimoglu, I. H., & Geyik, A. K. (2006). An application of fuzzy clustering to manufacturing cell design. *Journal of Intelligent and fuzzy systems, 17*(2), 173–181.
Tsai, C. Y., & Chang, C. A. (2003). Fuzzy neural network for intelligent design retrieval using associative manufacturing features. *Journal of Intelligent Manufacturing, 14*, 183–195. doi:10.1023/A:1022951430109.
Tsai, C. Y., Chiu, C. C., & Chen, J. S. (2006). A case based reasoning system for PCB defect prediction. *Expert Systems with Applications, 28*, 813–822. doi:10.1016/j.eswa.2004.12.036.
Tseng, T. L., Jothishanker, M. C., & Wu, T. (2004a). Quality control problem in printed circuit board manufacturing—An extended rough set theory approach. *Journal of Manufacturing System, 23*(1), 56–72.
Tseng, T. L., Leeper, T., Banda, C., Herren, S. M., & Ford, J. (2004b). Quality assurance in machining process using data mining. In *Proceedings of Industrial Engineering Research Conference* (pp. l–6). Houston, Taxes, May 15–19, 2004.
Tseng, T. L., Kwon, Y., & Ertekin, Y. M. (2005a). Feature-based rule induction in machining operation using rough set theory for quality assurance. *Robotics and Computer Integrated Manufacturing, 21*, 559–567. doi:10.1016/j.rcim.2005.01.001.
Tseng, T. L., Kwon, Y., Ho, J., & Jiang, F. (2005b). Hybrid data mining and type II fuzzy system approach for surface finish from the perspective of E-manufacturing. *Proceedings of the Society for Photo-Instrumentation Engineers, 5999*, 1–12.
Tseng, T. L., Huang, C. C., Jiang, F., & Ho, J. C. (2006). Applying a hybrid data mining approach to prediction problems: A case of preferred supplier prediction. *International Journal of Production Research, 44*(14), 2935–2954. doi:10.1080/00207540600654525.
Wang, L. X., & Mendel, J. M. (1992). Generating fuzzy rules by learning from example. *IEEE Transactions on system, man and cybernetics, 22*(6), 1414–1427.
Wang, X. Z., & McGreavy, C. (1998). Automatic classification for mining process operational data. *Industrial & Engineering Chemistry Research, 37*(6), 2215–2222. doi:10.1021/ie970620h.
Wang, K. J., Chen, J. C., & Lin, Y. S. (2005). A hybrid knowledge discovery model using decision tree and neural network for selecting dispatching rules of a semiconductor final testing factory. *Production Planning and Control, 16*(6), 665–680. doi:10.1080/09537280500213757.
Wang, Z., Shao, X., Zhang, G., & Zhu, H. (2005). Integration of variable precision rough set and fuzzy clustering: An application to knowledge acquisition for manufacturing process planning. *Lectur Notes of Artificial Intelligence, 3642*, 585–593.
Wang, C. H., Kuo, W., & Bensmail, H. (2006). Detection and classification of defects patterns on semiconductor wafers. *IIE Transactions, 38*, 1059–1068. doi:10.1080/07408170600733236.
Woon, Y. K., Ng, W. K., Li, X., & Lu, W. F. (2003). Efficient web log mining for product development. In *Proceedings of the International Conference on Cyberworld*.
Xu, R., & Wunsch, D. (2005). Survey of clustering algorithms. *IEEE Transactions on Neural Networks, 16*(3), 645–678. doi:10.1109/TNN.2005.845141.
Yam, R. C. M., Tse, P. W., Li, L., & Tu, P. (2001). Intelligent predictive decision support system for condition based maintenance. *International Journal of advance Manufacturing Technology, 17*, 383–391.
Yuan, B., Wang, X. Z., & Morris, T. (2000). Software analyser design using data mining technology for toxicity prediction of aqueous effluents. *Waste Management (New York, N.Y.), 20*, 677–686. doi:10.1016/S0956-053X(00)00045-3.
Zhang, D., & Zhou, L. (2004). Discovering golden nuggets: Data mining in financial applications. *IEEE Transactions on System Man and Cyberntics-Part C Applications and Revies, 34*(4), 513–521.
Zhang, Y., & Dudzic, M. S. (2006). Online monitoring of steel casting process using multivariate statistical technologies: From continuous to transitional operations. *Journal of Process Control, 16*, 819–829. doi:10.1016/j.jprocont.2006.03.005.
Zhang, Y., & Jiao, J. (2007). An associative classification based recommendation system for personalization in B2C e-commerce application. *Expert Systems with Applications, 33*(2), 357–367.
Zhao, Q., & Bhowmick, S. S. (2003). Association rule mining: A survey. Technical Report, CAIS, Nanyang Technological University, Singapore, No. 2003116.
Zhou, C., Nelson, P. C., Xiao, W., Tirpak, T. M., & Lane, S. A. (2001). An intelligent data mining system for drop test analysis of electronic products. *IEEE Transactions on Electronics Packaging Manufacturing, 24*(3), 222–231. doi:10.1109/6104.956808.
Zhou, J., Li, X., Andernroomer, A. J. R., Zeng, H., Goh, K. M., Wong, Y. S., & Hong, G. S. (2005). Intelligent prediction monitoring system for predictive maintenance in manufacturing. In *32nd Annual Conference of IEEE Industrial Electronics Society* (pp. 6–12). 6–10 Nov. 2005.