DATA MINING IN MANUFACURING: THE NATURE AND IMPLICATIONS

Kesheng Wang
Department of Production and Quality Engineering, Norwegian University of Science and Technology, N-7491 Trondheim, Norway. Email:kesheng.wang@ntnu.no.

Abstract: Recent advances in computers and manufacturing techniques have made it easy to collect and store all kinds of data in manufacturing enterprises. Traditional data analysis methods are no longer the best alternative to be used. The problem is how to enable engineers and managers to understand large amount of data. Data mining approaches have created new intelligent tools for automated extracting, useful information and knowledge. All these changes will have a profound impact on current practices in manufacturing. In this paper the nature and trends of these changes and their implications on product design and manufacturing are discusses.

Key words: Knowledge discovery, data mining, design, manufacturing.

1. INTRODUCTION

The productivity and efficiency of manufacturing enterprises can be affected by product design, manufacturing decisions and production management. With recent advances in computers and manufacturing techniques, powered data acquisition systems are actively in use in manufacturing enterprises. This supports huge datasets which are related to bill of materials, product design, process planning and scheduling, process and systems, monitoring and diagnosis, and market forecasting, which are collected and stored in databases at various stages for easy manufacturing. The knowledge which is valuable for enterprises including rules, regulars, clusters, patterns, associations and dependencies, etc. is hidden in the

datasets. How to enable engineers and managers to extract, understand and use the knowledge from large amount of datasets is still a problem for most of companies. For example, Charts and tables are places on walls, but not read by all, reports and manuals generated, only to be stored on bookshelves, databases created, only to be archived away. Critical manufacturing-related knowledge may be hidden in the data. Example of such knowledge may include rules or regulations for finding defects of the quality of the products. Human operators may never find the rules by manually investigation. That means he may never discover such hidden knowledge from the data. Traditional data analysis methods are no longer the best alternative to be used. Data Mining (DM) approaches based on Computational Intelligence (e.g. Artificial Neural Networks, Fuzzy Logic Systems and Genetic Algorithms), Machines Learning (Decision Tress and Association Rules) and statistics have created new intelligent tools for automated extracting useful information and knowledge. These changes will have a profound impact on current practices in manufacturing.

The applications of DM primarily are related to the fields of business and health care. Applications of DM to manufacturing are still underutilized and infrequently used. In this paper the nature and trends of these changes and definition, functions, techniques and applications of Data Mining on manufacturing are discusses. A case study is used to demonstrate the overall DM process.

2.   DATA MING

2.1   Definition

Data mining is an integration of analysis and modeling technologies developed over the last twenty years. Data mining is often defined as the process of extracting valid, previous unknown, comprehensible information from large data bases in order to improve and optimize business decision (Fayyad, et al, 1996; Last and Kandel, 2004).

DM techniques are at the core of DM process, and can have different functions (tasks) depending on the intended results of the process. In general, data mining functions can be divided into two broad categories: discovery data mining and predictive data mining.

(1). Discovery data mining

Discovery data mining is applied to a range of techniques, which find patterns inside your data without any prior knowledge of what patterns exist.

The following examples of discovery data mining: (1). Clustering; (2). Link analysis; and (3). Frequency analysis; etc.

(2). *Predictive data mining*

Predictive data mining is applied to a range of techniques that find relationships between a specific variable (called the target variable) and the other variables in your data. The following are examples of predictive data mining techniques: (1). Classification; (2). Value predication; and (3). Association rules; etc.

2.2   Techniques

A variety of techniques are available to enable the above functions. The most commonly used techniques can be categorized in the following groups (Kantardzic, 2003; Wang, 2005):

1. Classical statistical methods (e.g., linear, quadratic, and logistic discriminate analyses),
2. Modern statistical techniques (e.g., projection pursuit classification, density estimation, k-nearest neighbor, Bayes belief networks),
3. Artificial Neural Networks (ANNs),
4. Support Vector Machines (SVM),
5. Decision Trees (DT) and Rule Induction (RI) algorithms
6. Association Rules (AR),
7. Case Based Reasoning (CBS),
8. Fuzzy Logic Systems (FLS) and Rough Sets (RS),
9. Sequential Patterns (SP),
10. Genetic Algorithms (GAs),
11. Evolutionary Programming (EP), and
12. Visualization Methods (VM), etc.

Each class containing numerous algorithms, for example, there are more than 100 implementations of the decision tree algorithms. A hybrid DM system in which several techniques with different functions can be integrated to achieve a desired result are often more effective and efficient than a single one. For example, in order to identify the attributes that are significant in a manufacturing process, clustering can be used first to segment the process database into a given predefined number of categorical classes, then classification can be used to determine to which group a new data is belongs.

2.3   Procedures

The generic data mining procedures seen from IBM's viewpoint involves the following seven steps:

1. Defining the business issue in a precise statement,
2. Defining the data model and data requirements,
3. Sourcing data from all available repositories and preparing the data (the data could be relational or in flat files, stored in a data warehouse, computed, created on-site or bought from another party). They should be selected and filtered from redundant information,
4. Evaluating the data quality,
5. Choosing the mining function and techniques,
6. Interpreting the results and detecting new information, and
7. Deploying the results and the new knowledge into your business.

To understand how DM can overcome a variety of problems in manufacturing, we consider activities in a manufacturing company.

3.   DM IN MANUFACTURING

3.1   Challenges

Application of data mining to several domain areas such as business, telecommunication and medicine has been investigated. Unfortunately, there has not been similar research interest and activity in the manufacturing domain, despite of the potential benefits. Different reasons can be explained:

1. The majority of researchers in the manufacturing domain area are not familiar with data mining algorithms and tools.
2. The majority of theoretical data mining researchers are not familiar with the manufacturing domain area.
3. The few researchers who are skilled in both data mining algorithms and manufacturing domain area are not able to access to often proprietary and sensitive manufacturing enterprise data.

The effort to explore the use of data mining in manufacturing enterprises started only a few years ago, mostly by the researchers in the manufacturing domain. The amount of interest in data mining has definitely grown in the manufacturing research community. However, most of manufacturing researchers are not familiar with DM algorithms and which domain in manufacturing is suitable for DM. The challenge will be where DM should be and what DM technique should be used.

3.2   Applications

Today's manufacturing systems and processes are very complex and complicated. There are many stages of operations and many variables related

to each operation in each stage. Even most experienced engineers, charged with controlling variables to make quality consistent, time short and cost low, face problems that have unknown causes to defects and failures in the system and processes. These problems lead to product variability, reproduce, and rejects. System and process engineers always try to understand the relationship between variables using system and process models which are based on mathematics with much of assumptions. DM gives this burden to computers to quickly and exhaustively find those relationships which are useful for manufacturing systems and processes. With the knowledge, the system and process engineers can focus on improvement of the system and the process. (Braha, D., 2001)

Companies such as IBM, ORACLE, SPSS and SAS have developed DM software to large organizations for their data warehouses or data mart. Implementation of the DM systems requires the employment of people with sophisticated mathematical and computer skills. More recently, better users interface with advanced visualization have made DM accessible to a broader range of people. DM now can be applied to engineering and manufacturing easily, but experience and knowledge of how the tools work is still important. Users can encounter difficulty in getting their data loaded into the software, or in knowing which technique would work best for their particular problem. Once learned, however, DM can give the manufacturing engineer a powerful tool to uncover new truths and tackle the hardest problems.

The specific DM applications in advanced manufacturing include: (1). Manufacturing system modeling; (2). Manufacturing process control; (3) Quality control; (4) Monitoring and diagnosis; (5) Safety evaluation; and (6) Process planning and scheduling; (Chen, et al., 2005; Huang, et al., 2005; Hsu, et al., 2005; Tsironis, et al., 2005)

As we discussed above, collecting and cleaning of datasets are time consuming (80% of DM process) and critical to the success of data mining. Examples of variables in manufacturing datasets include:

1. Manufacturing process variables: machining, casting, forgings, extrusions, stampings, assembling, cleaning, etc.
2. Manufacturing process variables: Cutting speed, pressure, lubricants, coolants, voltage, current, forces, toques, vibration, etc.
3. Other variables: product materials, numbers, machines, fixtures, etc.
4. Environnement conditions variables: humidity, temperature, toxic, etc.
5. Working condition variables: duration, shift, injuries and accidents, etc.
6. Target variables: quality, yield, productivity, performance index, etc.

Successful implementations of data mining are not easy tasks. Some of the success factors are listed as the following: (Kasravi, 1997)

1. Datasets available – Availability of appropriate datasets is essential to the success of DM project.
2. Data preprocess – Cleanness of datasets is the major effort in DM.
3. Information content – not all data are suitable for data mining. Care must be taken to ensure that relevant variables and records are collected such that they could serve as the basis of any knowledge to be solved.
4. Correlation vs. causality – DM is a computer-based technique and it can only discover correlations, not causal relationships among data sets.
5. Privacy – Data sets are normally related to privacy of both personnel and companies which sometimes make it difficult to collect data and interpret the results.
6. Corporations – One of the important steps in data mining is interpretation. The patterns and rules obtained from data mining are domain-specific and the interpretation is a difficult task. Data mining experts should work with domain expert in order to make success of data mining.
7. Tools and software selection – Many data mining algorithms are standard, therefore a number of commercial data mining tools are currently available on the market. It will help manufacturing engineers to launch DM projects easily and quickly.

4.   AN IMPLEMENTATION CASE

4.1   The scheduled maintenance data

Many examples of application of data mining in manufacturing can successfully include: Conceptual design, Design manufacturing systems, Fault diagnosis, Preventive machines maintenance, Manufacturing knowledge acquisition, Material requirements planning, Adaptive scheduling, Process planning, Manufacturing process control, Quality control, Optimization of yield, Assemble selection, and Learning in robotics. This section presents a case to show how to use data mining techniques to solve a manufacturing problem.

There are two goals for this data mining task: 1. Classification - to determine what subsystems or components are most responsible for downtime; 2. Predication - to predict when preventive maintenance would be most effective in reducing failures. This information can then be used in setting maintenance policy guidelines such as planed maintenance schedules.

Classification and predication can be done by the use of Decision Trees (DT). The advantages of DT are: 1. Robust performance, 2. Clear visualization of output, 3. Ease interpretation. Especially DT is a good choice when the number of classes are low.

The one of mining tasks for scheduled maintenance dataset is a classification of machine health. The target variable of DT is Equipment Availability (EA):

$$EA = \frac{ScheduledTime - DownTime}{ScheduledTime} \times 100$$

Because EA takes into account the amount of time needed to repair a machine, it is a good indicator of machine health. The target of the classification can be shown in the next section.

4.2 Machine health mining (Wang, et al., 2004)

The case study for machine health is run following the procedure and the data mining development is based on the commercial software, IBM Intelligent Miner. [Reinschmidt, et al, 1999]

4.2.1 The business issues

The machine health analysis is an important task for maintenance activities of a company because it permits it to find the most sensible machines, which are the most responsible to the low Equipment Availability of the whole company. The purpose of data mining is to use DT to detect a group of machines with low EA. It is useful in order to plan a maintenance action and we are able to focus on the group of machines in order to make the maintenance more effective and save time and cost.

4.2.2 The Data model

This datasheet has been created in order to obtain results in a particular way. Indeed, we would like here to classify the machine related to the importance of their downtime inside the company. Each generated node of the tree contains the different machines, which should be classified by the evaluation of the value of Equipment Availability. In the original datasheet, all downtime have been registered with respect to scheduling time. It is not suitable for the data mining process. Therefore we define each machine as a column of the datasheet with the values of the total downtime, which is the cumulative value (30+30 = 60). The modified datasheet is shown in Fig. 1.

4.2.3   Evaluating the data quality

The original dataset from the company is not quite suitable for our data mining purpose. It contains some errors, such as missing data, wrong data and other errors. A cleaning process is required. Afterwards a pure and modified datasheet can be ready for a data mining process.

As seen in Fig. 1, each record has the value of the total downtime for each machine. The last record contains the value of Equipment Availability of each machine
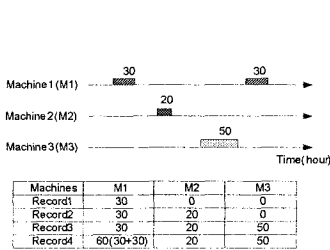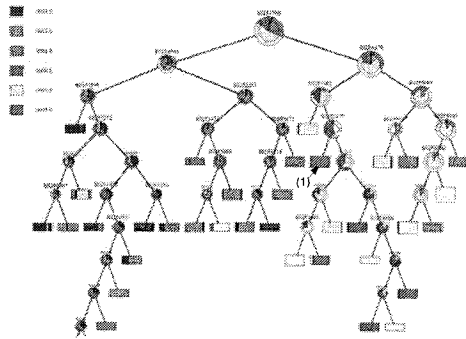

Figure 1. The datasheet for machine health.

Figure 2. The tree structure of the machine health.

4.2.4   Selecting DT as a data mining approach

DT has been selected as a data mining algorithm for the task. We input the modified datasheet to IBM Intelligent Miner. We define six targets of classification, depicted from 0001,0 to 0006,0. The decision tree generated is shown in Fig. 2.

4.2.5   Interpretation of the tree

On the top left corner of the tree (Fig. 2), we can see the 6 main classes related to the variable of EA. 0001,0 corresponds to the highest availability and 0006,0 to the lowest one. Then the data are classified in the leaf nodes where you can see a repartition of colors. We have a repartition, a number of records and purity for each leaf node.

On each node of the tree there is a condition on the number of hours of downtime for each machine. For example, the root node corresponds to the Machine 80254580. And the condition is: "*If 80254580 < 322 hours of downtime, **Then** it goes to the left branch*".

In order to generate some rules, the principle is the same. We look at a leaf node. If the number of records and the purity are important enough, we

can keep the rule and say that this rule is obviously relevant.

We can take an example of the leaf node (1) as shown in Fig. 3. We can see that label is 0005,0, Records is 145, Class Distribution is (1 0 0 1 2 141), Purity 97,2 %.
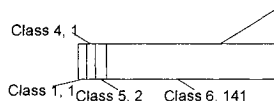

Figure 3. Zoom in the leaf node (1).

4.2.6   Generation of rules

Here we can say that this classification is meaningful because the number of records is important with a high purity so we can look at the rules:

   IF   (80254580 > 322)
   And   (80254270 < 22,5)
   And   (80253462 > 202,5)
   And   (80252477 < 36)
   THEN Class is 0005,0

So for example the maintenance staff will be able to look at the machines 80254580 and 80253462 which seem to be the most responsible for the low availability in this part of the classification.

4.2.7   Deploying the results and the new knowledge into your business

When we look at the induced tree in Fig. 2, we clearly see that the emphasis on preventive maintenance should be on machines 80254580, 80254593, and 80254270 since these machines are close to the root node and are associated with low EA.

So the tree structure is interesting for the maintenance management because it's pretty easy to detect the different part (high equipment availability, low equipment availability) with the colors on the charts and by the way that the machines are related to these results.

5.   CONCLUSION

In the quest for a sustainable competitive advantage, manufacturing companies have finally come to realize that technology alone is not enough. What makes it sustainable is the knowledge residing in people's head, and process and experience in the companies. Modern manufacturing enterprises

have collected and stored a huge amount of manufacturing datasets for easy access, such as product design data, manufacturing resource data, manufacturing system and process data, which are not fully used for increased competitive advantages in the global economy. Business Intelligence (BI), Data Mining (DM) and Knowledge Management (KM) will change the stakes. These techniques definitely are able to help the enterprises to collect, extract, create and delivery manufacturing knowledge in a competitive environment. In this paper we focus on DM techniques, which can be used to discover and extract previous unknown manufacturing knowledge from manufacturing data bases. The knowledge discovered can be used to improve the product manufacturing processes.

6.   REFERENCES

Braha, D., (2001), Ed: Data mining for Design and Manufacturing – Methods and applications, Kluwer Academic Publishers, The Nederlands.

Chen, W. C., Tseng, S. S., and Wang, C. Y., (2005), A novel manufacturing defect detection methods using association rule mining techniques, *Expert Systems with Applications*, V. 29, No. 4, pp. 807-815.

Fayyad, U. Piatetsky-Shapiro, G., Smyth, (1996), From data mining to Knowledge discovery: An overview, in: *Advantaces in Knowledge Discovery and Data Mining,* Fayyad, U. Piatetsky-Shapiro, G., Smyth, P. and Uthurusamy, Eds., Cambridge, MA: MIT Press, pp. 1-36.

Hsu, C. H. and Wang, M. J. J., (2005), Using decision tree-based data mining to establish a sizing system for manufacture of garments, *Int J Adv Manuf Technol*, v. 26, pp. 669-674.

Huang, H., and Wu, D., (2005), *Lecture Notes in Artificial Intelligence* (sub series of Lecture Notes in Computer Science) V. 3614, No. Part II, *Fuzzy Systems and Knowledge Discovery: Second International Conference, FSKD, Proceedings*, pp.577-580.

Kantardzic, M., (2003), *Data mining – concepts, Models, Methods, and Algorithms*, Wiley-Interscience.

Karavi, K. (1997), Data mining and knowledge discovery in manufacturing, *AUTOFACT'97, Information Technologies for the Manufacturing Enterprise*, Michigan.

Last, M. and Kandel, A., (2004), Discovering useful and understandable patterns in manufacturing data, Robotics and Autonomous Systems, v. 49, pp. 137-152.

Reinschmidt, J., Gottschalk, H., Kim, H., and Zwietering, D., 1999, *Intelligent Miner for data: Enhance your business intelligence*, International technical support organization, IBM (http://www.redbooks.ibm.com).

Tsironis, L., Bilalis, N. and Moustakis, V., (2005), Using machine learning to support quality managm,ent – framework and experimental investigation, The TQM magazine, V. 17, No. 3, pp. 237-248.

Wang, K., (2005), *Applied Computational Intelligence in Intelligent Manufacturing Systems*, Advanced Knowledge International.

Wang, K., Fabien, C., Wang, Y., Yuan, Q. and Fang, M., (2004), The 6th International Conference on Frontiers of Design and Manufacturing, June 26-28, Xi'an, China.