Knowledge Extraction in Manufacturing using Data Mining Techniques

Catarina Dudas[1], Amos Ng[1], Henrik Boström[2]
[1]Centre for Intelligent Automation, University of Skövde, Skövde, Sweden
[2]School of Humanities and Informatics, University of Skövde, Skövde, Sweden
catarina.dudas@his.se

ABSTRACT

Nowadays many production companies collect and store production and process data in large databases. Unfortunately the data is rarely used in the most value generating way, i.e., finding patterns of inconsistencies and relationships between process settings and quality outcome. This paper addresses the benefits of using data mining techniques in manufacturing applications. Two different applications are being laid out but the used technique and software is the same in both cases. The first case deals with how data mining can be used to discover the affect of process timing and settings on the quality outcome in the casting industry. The result of a multi objective optimization of a camshaft process is being used as the second case. This study focuses on finding the most appropriate dispatching rule settings in the buffers on the line.

The use of data mining techniques in these two cases generated previously unknown knowledge. For example, in order to maximize throughput in the camshaft production, let the dispatching rule for the most severe bottleneck be of type Shortest Processing Time (SPT) and for the second bottleneck use any but Most Work Remaining (MWKR).

Keywords: Data mining, Quality engineering, Knowledge extraction.

1. INTRODUCTION

Knowledge is the key aspect to become a successful and well organized business of today. Knowledge can be on different levels such as knowing the number of operators working on a certain day or a complex chemical formula describing the relationship between different materials in a liquefied compound. It can also be generated from the result of optimizing a discrete simulation model, i.e., simulation based optimization. Although extensive amounts of knowledge are known and widely used within a business, there is most likely a lot of unknown knowledge stored in in-house databases which can be further exploited.

In this paper, knowledge extraction by data mining in two different applications is considered. In the first experiment, important process variables for quality improvement are found from a process data base combined with data from quality control in the casting industry. The second experiment is performed with data generated by a simulation-based optimization model, where different dispatching rules in a production line are used to predict both throughput and total tardiness.

2. INTRODUCTION TO DATA MINING

Computer aided techniques are widely used within companies in many fields and data regarding many different aspects of a business is regularly collected and stored, such as process data, break down data and quality outcome of the finished product. Often these types of data have been stored in different databases for one or the other reason, in some cases even for no specific reason, over many years. One of the key aspects of data mining is that it can be used for analyzing data that has been collected during the normal operations of the manufacturing process, i.e., data does not have to be specifically collected for this purpose [1]. However, a drawback of having collected data without analysis in mind is that the data may not be optimal for the intended purpose.

One predicament is the huge amount of data stored in addition to the lack of knowledge. The future goal would be to work continuously with information technology in a pro-active manner to achieve process knowledge and control. To accomplish this goal, the in-house business systems must be investigated in a systematic way finding the additional data required or inconsistencies and to extract important knowledge. The next step is to apply data mining techniques on the essential and interesting data.

Knowledge Discovery in Databases (KDD) is defined by Fayyad et al [2] as "the non-trivial process of identifying valid, novel, potential useful and ultimately understandable patterns in data". According to Fayyad et al, data mining is only one step in the KDD process (Fig. 1), whereas others often use data mining as a synonym for KDD, as done in this paper.

2.1 The data mining process

Data mining is an automated or semi-automated technique used to discover and interpret hidden relationships, patterns or trends in large data sources.

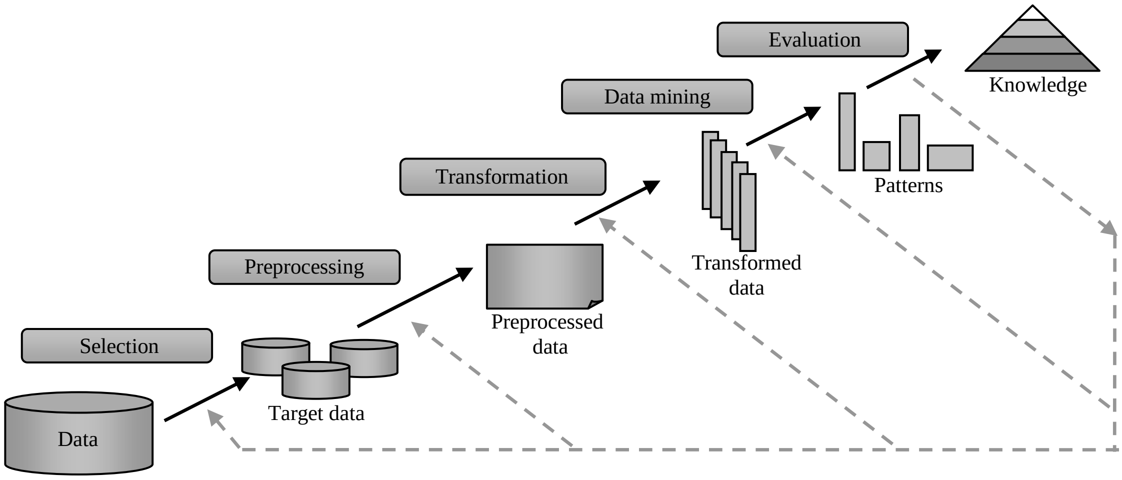

Figure 1: Knowledge Discovery in Databases according to Fayyad et al [2]

Figure 1 shows the data mining process as an iterative procedure, going from data to knowledge. A blend of concepts and algorithms from machine learning, statistics, artificial intelligence, and data management are borrowed to the field of data mining.

The process can be divided into three parts: selection/preprocessing, mining and presentation. The first step involves gathering, organizing and cleaning data before it can be used. The mining process concerns choosing appropriate method(s) to be used for searching patterns in data. The final step concerns how to present the results of the prior processes in a suitable way. After evaluation of the presented results, the entire process, or parts of it, may be re-iterated.

Data mining is a rapidly expanding field with growing interests and importance. Manufacturing is indeed an application area where the use of this technology can provide a significant competitive advantage [1].

3. LITERATURE REVIEW

Data mining is a technique which has been used in both private and public sectors and clearly with different objectives. Companies within banking, insurance and retailing use data mining to reduce cost, detect frauds and to advertise in more effective ways. Homeland security is yet another application area of growing interest, in which data mining also has been used.

The first use of artificial intelligence in manufacturing applications appeared in the 1980’s according to [1]. In the beginning of the 1990’s, the use of data mining techniques was introduced for production, something which has been growing since then. A comprehensive review of papers considering data mining applications within manufacturing is presented by Kusiak in [1]. Manufacturing operations, fault detection, design engineering and decision support systems have been in focus as research topics, but there is still an enormous potential for further research in other application areas, such as maintenance, layout design, resource planning and shop floor control. Below is a brief listing of the main current application areas of data mining within manufacturing.

3.1 Preventive maintenance

Preventive maintenance normally integrates with quality control, and preventive maintenance plans can often be designed by accessing quality control databases. Different data mining techniques, including decision trees, regression, neural networks, have been used to predict component failure based on the data collected from manufacturing process allowing maintenance actions to be undertaken whenever such failures can be expected [3]. The preventive maintenance area was one of the first areas within manufacturing to take advantage of data mining.

3.2 Decision support system

The main reason for using data mining in decision support systems is to discover relevant system knowledge before the decision making process. The knowledge extracted from databases can be integrated with existing expert systems in order to modify or finding patterns in job shop scheduling sequences [4].

3.3 Fault detection

One procedure for fault detection is to examine historical data for better understanding of the process, and to use this knowledge to predict and improve the process performance. Data mining techniques can not only be used for classifying, e.g., the products not fulfilling the quality requirements, but also for determining the most influencing risk factors for failures.

Shi et al. [5] developed an artificial neural network model for a chemical manufacturing process using historical data for validation. The model was used to predict the outputs for well-designed process settings. The predicted result was then used to perform statistical tests and identify the significant factors and interactions that affect the quality-related output variables. The data mining approach showed potential to achieve a better understanding of process behavior and to improve the process quality efficiently.

In the paper by Karlsson et al. [6], the fusion of different sources for fault detection is investigated. To determine whether an industrial motor is worn out or not, a technique combining data from different vision systems for pattern recognition and signal processing is used for classification of the motor status. In that study, nine different fusion methods were used for classification of data extracted from these vision images.

3.4 Operational control

Data mining for analyzing the effect of local dynamic behavior for operational control can be used for extracting knowledge to generate control policies, e.g., for intelligent scheduling systems. These operational systems are often inherently adaptive, and since data is accumulated in real-time, baseline policies generated by the data mining algorithms can be updated on the fly [7].

4. PREDICTIVE AND DESCRIPTIVE DATA MINING

Data mining techniques have become standard tools to develop predictive and descriptive models in situations where one wants to exploit data collected from earlier observations in order optimize future decision making [8]. In the case of predictive modeling, one typically tries to estimate the expected value of a particular variable (called the dependent variable), given the values of a set of other (independent) variables. In the case of a nominal dependent variable (i.e., the possible values are not given any particular order), the prediction task is usually referred to as classification, while the corresponding task when having a numerical dependent variable is referred to as regression. One usually wants the model to be as correct as possible when evaluated on independent test data, and several suggestions for how to measure this have been proposed. For classification, such measures include accuracy, i.e., the percentage of correctly classified test examples, and the area under the ROC curve (AUC), i.e., the probability that a test example belonging to a class is ranked as being more likely belonging to the class than a test example not belonging to the class [9]. Besides the ability to make correct predictions, one is also often interested in obtaining a comprehensible (descriptive) model, so that the reasons behind a particular classification can be understood, and also that one may gain insights into what factors are important for the classification in general. Examples of such comprehensible models are decision trees and rules, e.g. [10], while examples of models not belonging to this group, often called black-box, or opaque, models, include artificial neural networks and support vector machines (see e.g. [11]).

4.2 Decision trees and ensembles

Techniques for generating decision trees are perhaps among the most well-known methods for predictive data mining. Early systems for generating decision trees include CART [12] and ID3 [13], the latter being followed by the later versions C4.5 [10] and C5.0 [14]. The basic strategy that is employed when generating decision trees is called recursive partitioning, or divide-and-conquer. It works by partitioning the examples by choosing a set of conditions on an independent variable (e.g., the variable has a value less than a particular threshold, or a value greater or equal to this threshold), and the choice is usually made such that the error on the dependent variable is minimized within each group. The process continues recursively with each subgroup until certain conditions are met, such as that the error cannot be further reduced (e.g., all examples in a group belong to the same class). The resulting decision tree is a graph that contains one node for each subgroup considered, where the node corresponding to the initial set of examples is called the root, and for all nodes there is an edge to each subgroup generated from it, labeled with the chosen condition for that subgroup.

Decision trees have many attractive features, such as allowing for human interpretation and hence making it possible for a decision maker to gain insights into what factors are important for particular classifications. However, recent research has shown that significant improvements in predictive performance can be achieved by generating large sets of models, or ensembles, which are used to form a collective vote on the value for the dependent variable [15]. It can be shown that as long as each single model performs better than random, and the models make independent errors, the resulting error can in theory be made arbitrarily small by increasing the size of the ensemble. However, in practice it is not possible to completely fulfill these conditions, but several methods have been proposed that try to approximate independence, and still maintain sufficient accuracy of each model, by introducing randomness in the process of selecting examples and conditions when building each individual model. One popular method of introducing randomness in the selection of training examples is bootstrap aggregating, or bagging, as introduced by Breiman [16]. It works by randomly selecting n examples with replacement from the initial set of n examples, leading to that some examples are duplicated while others are excluded. Typically a large number (at least 25-50) of such sets are sampled from which each individual model is generated. Yet another popular method of introducing randomness when generating decision trees is to consider only a small subset of all available independent variables at each node when forming the tree. When combined with bagging, the resulting models are referred to as random forests [17], and these are widely considered to be among the most competitive and robust of current methods for predictive data mining. The drawback of ensemble models are however that they can no longer be easily interpreted and hence provide less guidance into how classifications are made.

The Rule Discovery System (RDS) [18] addresses this problem by providing some insight into what factors are of importance in an ensemble of decision trees by presenting the variable importance of each independent variable, i.e., how much the variable, relative to all other variables, contributes to reducing the squared error of the dependent variable.

5. TWO DATA MINING APPLICATIONS

The use of data mining in manufacturing applications can have different aims and purposes. In this paper two applications are presented: data mining for identifying variables affecting quality and for identifying dispatching rules setting in a production line.

5.1 Process data in casting industry

This study is done in cooperation with Volvo Powertrain in Skövde, Sweden. Volvo Powertrain supplies power train parts, cylinder blocks, gear boxes and drive shafts, to the business areas within the Volvo Group.

Complexity in the casting process: The casting process is complex with multivariate interactions of known but also unidentified factors which makes it impossible for humans to overview. Humans are normally not able to simultaneously analyze situations involving more than three variables very effectively and this becomes even more difficult when the data are corrupted by noise and uncertainty [19]. The current way of analyzing the casting process is done in a one-variable-at-a-time manner. Due to the complex relationships, there is a requirement for a more sophisticated and accurate way of analyzing data from such processes and it is believed that data mining can achieve this in a useful way.

Data used in the experiment: Process data and quality data has been collected from the pre-processing line before casting of products. It is believed that some of these variables or process settings have a more crucial impact on the resulting product, i.e., the cylinder head. In order to determine these variables, we create a model consisting of a number of input variables and using the product quality, i.e., if the cylinder head is rejected or not, as the response variable. This is done to get an understanding of how the processes interact and will be helpful to keep these variables under a more severe control.

Rejection codes: The quality data contains a number of different rejection codes. These codes are used to identify what kind of defect in quality the cast but rejected cylinder head has.

Since the coding of rejection is done manually it is believed that similar codes can be used in various ways by different operators. In order to be able to use the quality data in a more appropriate way, the rejection codes have been divided into fewer subgroups.

One expectation is that the division into subgroups will generate more accurate result, since the rejections are affected by different process variables, see Fig. 3. A global model, where the rejection code is binary, is also used for comparison.

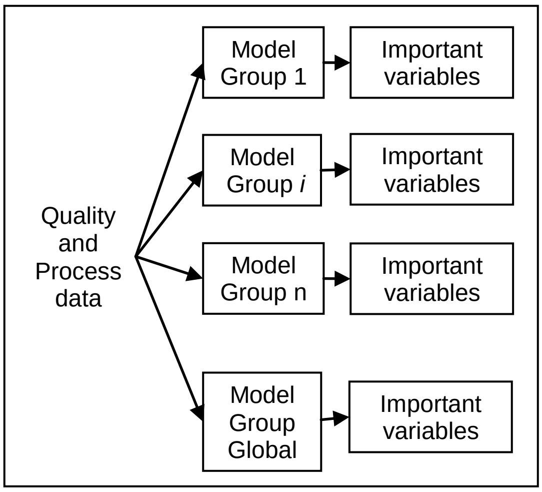

Figure 3: The set up of different models to generate more accurate results.

Input and output variables: In the data mining tool, the process data is used as input variables and the quality data is used as the output variable which is being classified.

5.2 Results in the casting experiment

RDS [18] was used as data mining tool for this study. One run is done for each rejection subgroup and one for the global data. The modeling methods used are a single tree structure and an ensemble of 50 trees. 10-fold cross-validation is used as validation method for both modeling methods.

The rejection codes present in this data set can be divided into the following six subgroups:

- Group 1: Sand influence
- Group 2: Core defects
- Group 3: Manual process
- Group 4: Blisters
- Group 5: Damage after casting
- Group 6: Not completely cast

These groupings are believed to be affected by different variables. For instance, it is probably difficult or even impossible to use process data for prediction of rejections of Group 3, which is a manual process step.

Performance measurement: The models generated obtain high AUC for all sub groups except number 2 and 3, as shown in table 1. Due to results close to random for group number 3, these are not further explored in this section.

Table 1: The AUC value for the models built for each subgroup of rejection codes.

Ensemble Model	AUC
Global	0.72
Group 1	0.80
Group 2	0.67
Group 3	0.47
Group 4	0.73
Group 5	0.73
Group 6	0.78

Affecting process variables: The subgroups with specific variables are listed in Table 2.

Table 2: Specific variables for some subgroups.

Ensemble Model	Specific variable
Global	No specific
Group 1	Weight after casting
Group 2	Time_3
Group 4	No specific
Group 5	Time_2
Group 6	Time_6

In Table 2, the following times are declared:

- Time_2 – time between start of production and the assembly process
- Time_3 – time between when the upper and lower forms are put together and when it enters the chill
- Time_6 – time between a control process and when the upper and lower forms are put together.

It must also be noted that a few variables appear more frequently in all models, and these are:

- Time_1 – time between the drying process and a control station
- Time_8 – time between the coating process and a control station
- Time_9 – time between start of production and the coating process
- Chill and cast bin – 12 chills and 2 cast bins for each chill.

A checkup is done to see how the process data is set according to these significant variables, see Table 3. This implies that the coating process is important as well as the timing before and after for all types of rejection.

Table 3: Comparison of rejected and accepted products for important variables.

	Time_1	Time_8	Time_9
Aver. time for rejection	9971	11102	7610
Aver. time for non-rejection	5179	7728	8564

Choice of cast bin and chill: Two variables which have great impact on all subgroups is the choice of casting bin and chill, which can be seen in Fig. 4 and 5. Casting in the second bin results in greater number of rejections which also is the case for choice of chill number 6.

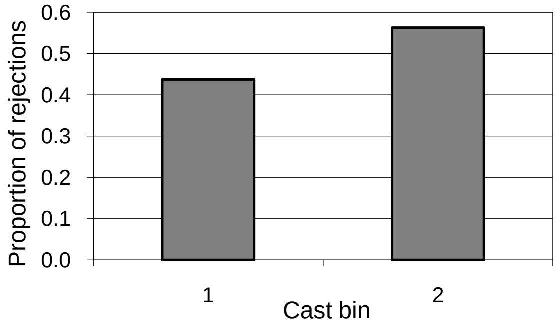

Figure 4: The number of rejections differs with choice of cast bin.

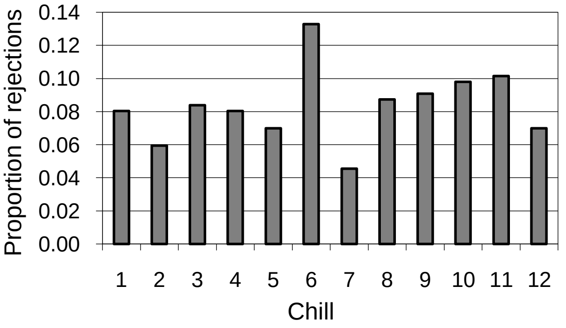

Figure 5: The choice of casting chill affects the quality of the cast product.

5.3 *Dispatching rules in a production line*

The aim of this experiment is to understand how dispatching rules affect the outcome of a production line. The result of a multi-objective optimization study is used to discover patterns in dispatching rule settings in order to maximize throughput or minimize total tardiness.

Simulation model at Volvo cars: In the experiment at Volvo Cars a Discrete Event Simulation (DES) model is used for bottleneck detection in a production line. The saying; A chain is as strong as it weakest link, is applicable in the theory of bottleneck detection. The goal of the study is to discover the bottlenecks in the production line. One should be aware of that when one bottleneck has been eliminated, another station or buffer will then be the current bottleneck.

Input to the data mining experiment is output from a DES model, which represents the H-factory at Volvo Cars in Skövde. The H-factory is committed to camshaft processing and 15 variants are handled on the production line. The H-factory consists of thirteen different groups of operations with one to seven machines in each group. All machines within a group of operations have the same capability and in front of every group of machines is a buffer.

H-factory based on dispatching rules: When a product enters to a buffer it is checked if a machine is free, if so the product is directly moved there. But, if there is no machine available then the product is placed on a free spot in the buffer. The buffer is checked every time a machine has finished a product. If there is only one product there; it is moved to the machine. The dispatching rules are considered each time there is more than one product in the buffer. The product to pick is dependent on the current dispatching rule assigned to that specific buffer.

There are eight different dispatching rules: shortest processing time (SPT), longest processing time (LPT), earliest due date (EDD), total work remaining (TWR), least work remaining (LWKR), most work remaining (MWKR), minimum slack time (MST) and operation due date (OPNDD).

Simulation Based Optimization (SBO): There are 13 groups of operations and 8 dispatching rules. Due to the great number of different dispatching rule settings, simulation based optimization is used to generate an optimal configuration of the production line. The optimization goal parameters are throughput, total tardiness and number of late products. The output of the simulation based optimization is the dispatching rules used for each operation with its resulting throughput and total tardiness. The number of different settings is 8^{13}, approximately $5.5 \cdot 10^{11}$, and about 400 of these are given by the SBO model and used for further exploration.

5.4 Results in the bottleneck detection experiment

The data set with 13 input variables (used dispatching rule for each buffer) and throughput and total tardiness was used in RDS to generate decision trees. Decision trees were used due to that they allow for interpretation in contrast to ensembles of trees.

It can easily be seen in Fig. 6 that the bottlenecks, i.e., the most important variables, are op20 and op90. In order to generate a small and more interpretable model all other input variables were excluded and a new model was built.

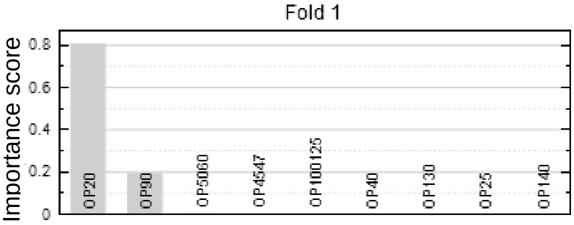

Figure 6: Important variables in the H-factory.

A decision tree is normally depicted with the root at the top having the ancestor nodes below it, see Fig. 7. An example is classified by the tree by following a path from the root to a leaf node, such that all conditions along the path are fulfilled by the example, where the conditions are formed from the variable names directly below each node and from the edge labels (e.g., the condition op20 = SPT follows from the root and the leftmost edge from the root).

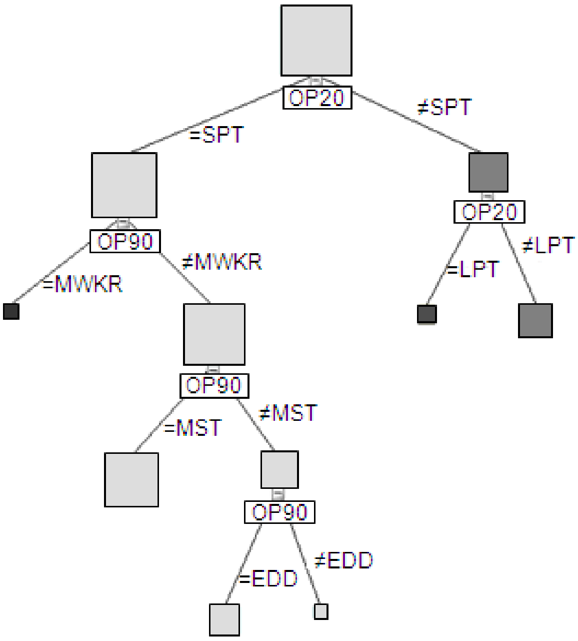

Figure 7: The tree structure for dispatching rules maximizing throughout.

The estimated regression value at a reached leaf node is used to assign the value of the prediction variable, i.e., throughput or total tardiness. The relative sizes of the area of each square chart in Fig. 7 correspond to number of observations used to estimate the regression value, where a dark sector corresponds to low average throughput, and a light sector corresponds to high average throughput.

Throughput as output variable: New information that may be extracted from the decision tree in Fig. 7 is that for a higher average throughput use dispatching rule SPT in op20 and use any of the dispatching rules but MWKR for the other buffers.

Total tardiness as output variable: The experiment for the total tardiness is performed in a similar manner. As an initial study, all input variables are used to identify the variables with most importance. As in the throughput experiment, op20 and op90 are the bottlenecks. The corresponding decision tree is shown in Fig. 8.

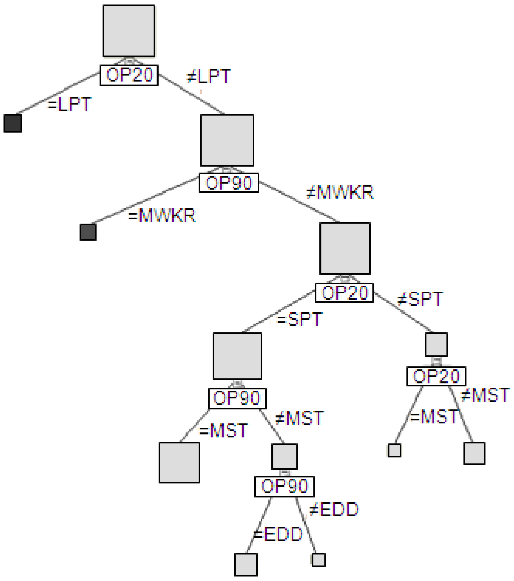

Figure 8: The tree structure for dispatching rules minimizing total tardiness.

This structure is not as easy to interpret as in the previous case, but still there is some valuable information that may be extracted. In contrast to the throughput experiment, the total tardiness case focuses on finding the settings to obtain low prediction values. One finding is that letting the most severe bottleneck have a dispatching rule that is not LPT and the second one having any but MWKR will result in a low total tardiness.

6. SUMMARY AND FUTURE WORK

This paper has focused on the use of data mining techniques in the manufacturing area. This is still a rather unexplored but exciting area for making the Swedish production industry more efficient. As a first step, the two cases described here have only applied a subset of available data mining techniques for knowledge extraction. A second step would be to work in a pro-active manner using data mining further as a prediction model.

One interesting result in this paper is the contradiction in the dispatching rule experiment where two different dispatching rule settings were discovered for op20. If throughput is to be maximized then SPT should be used as dispatching rule and LPT if total tardiness should be minimized. In RDS, only one dependent variable can be set for each model, and using a different technique would be an interesting future step. For instance, an artificial neural network can be used for this purpose, but is not as easy to interpret as a decision tree.

7. ACKNOWLEDGEMENT

This work was supported by the Information Fusion Research Program (www.infofusion.se) at the University of Skövde, Sweden, in partnership with the Swedish Knowledge Foundation under grant 2003/0104.

8. REFERENCES

[1] Kusiak, A. (2006). Data Mining in Manufacturing: A Review, *Journal of Manufacturing Science and Engineering*, **Vol. 128, Issue 4**, 969-976.

[2] Fayyad, U. M.; Piatetsky-Shapiro, G. and Smyth, P., (1996). From Data Mining to Knowledge Discovery in Databases, *AI Magazine*, **Vol. 17, Issue 3**, 37-54.

[3] Sylvain, L., Fazel, F., and Stan, M., (1999), Data Mining to Predict Aircraft Component Replacement, *IEEE Intelligent Systems*, **Vol. 14, Issue 6**, 59–65.

[4] Koonce, D.A. and Tsai, S. C., (2000), Using Data Mining to Find Patterns in Genetic Algorithm Solutions to a Job Shop Schedule, Computers and Industrial Engineering, Vol. 38, 361-374

[5] Shi, X., and Boyd, D. and Schillings P., (2004), Applying Artificial Neural Network and Virtual Experimental Design to Quality Improvement of Two Industrial Processes, *International Journal of Production Research*, **Vol. 42, Issue 1**, 101–118.

[6] Karlsson B., Järrhed J.O., and Wide P., (2002), A fusion toolbox sensor data fusion in industrial recycling, *IEEE Transactions on Instrumentation and Measurement*, **Vol. 51, Issue 1**, 144-149.

[7] Braha, D., (2001). *Data mining for Design and Manufacturing*, Preface pages, Kluwer Academic Publishers, Dordrecht

[8] Witten I. and Frank E., (2005) Data Mining: Practical Machine Learning Tools and Techniques (Second Edition), Morgan Kaufmann Publisher, San Francisco.

[9] Provost F., Fawcett T. and Kohavi R., (1998) *The case against accuracy estimation for comparing induction algorithms*, Proc. Fifteenth Intl. Conf. Machine Learning, 445-553

[10] Quinlan J.R., (1993). *C4.5: Programs for Machine Learning*, Morgan Kauffman, San Francisco

[11] Hastie T., Tibshirani R. and Friedman J., (2001) The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Springer-Verlag, London.

[12] Breiman L., Friedman J.H., Olshen R.A. and Stone C.J., (1984). *Classification and Regression Trees*, Wadsworth, Belmont

[13] Quinlan J.R., (1986). Induction of decision trees, *Machine Learning*, **Vol. 1, Issue 1,** 81-106

[14] Quinlan J.R., Data Mining Tools See5 and C5.0, http://www.rulequest.com/see5-info.html (accessed June. 2, 2008)

[15] Bauer E. and Kohavi R., (1999). An Empirical Comparison of Voting Classification Algorithms: Bagging, Boosting, and Variants, *Machine Learning*, **Vol. 36, Issue 1-2**, 105-139

[16] Breiman L., (1996). Bagging Predictors, *Machine Learning*, **Vol. 24, Issue 2**, 123-140

[17] Breiman L., (2001). Random Forests, *Machine Learning*, **Vol. 45, Issue 1**, 5-32

[18] Rule Discovery System, v. 2.6.0, Compumine AB, http://www.compumine.com/web/public/rds (accessed May 29, 2008)

[19] Wang, X. Z. (1999). Data Mining and Knowledge Discovery for Process Monitoring and Control, Springer-Verlag, London.