Application of Knowledge Discovery in Databases Process in the Production Systems

Dominika Jurovatá, Michal Kebísek, Júlia Kurnátová
Faculty of Materials Science and Technology
Slovak University of Technology
Paulinska 16, 917 24 Trnava, Slovakia
{dominika.jurovata, michal.kebisek, julia.kurnatova}@stuba.sk

Abstract. *This article deals with knowledge discovery in databases (abbr. KDD) and their application in the industrial area. The article is focused on methodology of process KDD and the overview of theoretical knowledge in the field of knowledge discovery in databases and data mining. Knowledge discovery in the production databases is minimally used for the process of planning and control. In this article, authors present the objectives and steps of the project.*

Keywords. data mining, knowledge discovery in databases, simulation, production system

1 Introduction

The most important business process is the production system which leads to the transformation of inputs into outputs. Ironically, it is the most underrated. As often as not, production is controlled and planned in Excel, while for the other business processes (such as accounting, finance, etc.) are used advanced tool. Planning and control are very difficult. In order to successfully operate business, its managers must have sufficient and accurate information. Appropriate sources of these information seem the simplification of data analysis and flexible creation of different reports, but also the process of data mining, too. Better understanding of system control and new interesting knowledge predictive of future behavior of the production system can be achieved by introduction the process of knowledge discovery into the control of production systems. New discovered knowledge will help managers in their decision-making process.

The process of KDD can be applied to solve many problems in the control and planning of production processes [10]. For example here belong:
- Production process optimization,
- Identification of production parameters influence on a production process,
- Identification of breakdowns in production process,
- Detection of deviations from production plan,
- Failure states detection of production equipments,
- Identification of time series of final production,
- Workstations layout optimization,
- Optimization of storage subsystem,
- Failures prediction in production process etc.

There are many problems that occur in the production process. It is important to choose the problem correctly and appropriate way of solving. The process of knowledge discovery in databases gives a lot of tools for solution of these problems.

2 Knowledge discovery in databases

The term „Knowledge Discovery in Databases“ is generally used to refer to the overall process of discovering useful knowledge from data, where data mining is a particular step in this process [5], [7].

There are many views on the classification and description of the KDD process. Therefore, there is an effort to standardize the process of KDD. The result of this initiative is very promising step towards the definition of standard methodology CRISP-DM for implementation of projects knowledge discovery.

View of CRISP-DM (Cross Industry Standard Process for Data Mining) on knowledge discovery in databases and data mining consists of six main phases (see Fig. 1). There are close relationships and always requires passage in both directions between different phases. Steps of project are based on this methodology.

Figure 1. CRISP-DM [7]

There are the following 6 phases within Cross Industry Standard Process for Data Mining [1], [7]:

1. Business understanding
2. Data understanding
3. Data preparation
4. Modeling
5. Evaluation
6. Deployment

3 Data mining

The term “data mining” is primarily used by statisticians, database researchers and business communities. Fayyad defined the data mining as a particular step in process KDD, is the application of specific algorithms for extracting patterns from data. The distinction between the KDD process and the data-mining step (within the process) is in the additional steps in the KDD process, such as data preparation, data selection, data cleaning, incorporation of appropriate prior knowledge and proper interpretation of the results of mining. They are essential to ensure that useful knowledge is derived from the data. [2]

Program STATISTICA from Statsoft were used to implement project. Data mining model was created from the following methods and techniques [6]:

- Neural Network Regression,
- Generalized K-Means Cluster Analysis,
- MARSplines (Multivariate Adaptive Regression),
- SVM (Support Vector Machines).

3.1 Neural Network for Regression

This technique uses a series of weights and hidden neurons to detect complex relationships. It can perform well in the presence of complicated, noisy and imprecise data. Data is pre-processed by scaling to (0,1) using a linear transformation.

In regression problems, the objective is to estimate the value of a continuous output variable, given the known input variables. Regression problems are represented by data sets with non-nominal (standard numeric) output(s).

A particularly important issue in regression is output scaling, and extrapolation effects. The most common neural network architectures have outputs in a limited range (e.g., (0,1) for the logistic activation function). This presents no difficulty for classification problems, where the desired output is in such a range. However, for regression problems there clearly is an issue to be resolved, and some of the consequences are quite subtle [3], [6].

3.2 Generalized K-Means Cluster Analysis

In general, the k-means is a method which aims to produce exactly k different clusters of greatest possible distinction. It should be mentioned that the best number of clusters k leading to the greatest separation (distance) is not known as a priori and must be computed from the data. The type of research question that can be addressed by the k- means clustering algorithm is for example: Suppose that you already have hypotheses concerning the number of clusters in your cases or variables. You may want to "tell" the computer to form exactly 3 clusters that are to be as distinct as possible.

3.3 Multivariate Adaptive Regression Splines

Multivariate Adaptive Regression Splines (abbr. MARSplines) is an implementation of techniques introduced by Friedman in 1991. This method is presented for flexible regression modeling of high dimensional data, where the curse of dimensionality would likely create problems for other techniques [4]. It is for solving regression-type problems, with the main purpose to predict the values of a continuous dependent or outcome variable from a set of independent or predictor variables.

MARSplines is a nonparametric regression procedure that makes no assumption about the underlying functional relationship between the dependent and independent variables. Instead, MARSplines constructs this relation from a set of coefficients and basis functions that are entirely "driven" from the regression data. In a sense, the method is based on the "divide and conquer" strategy, which partitions the input space into regions, each with its own regression equation [6].

3.4 Support Vector Machines

Support Vector Machines (abbr. SVM) are a group of supervised learning methods that can be applied to classification or regression. This method performs regression and classification tasks by constructing nonlinear decision boundaries. Because of the nature

of the feature space in which these boundaries are found, SVM can exhibit a large degree of flexibility in handling classification and regression tasks of varied complexities.

Support Vector Machines are based on the concept of decision planes that define decision boundaries. A decision plane is one that separates between a set of objects having different class memberships. A set of features that describes one case (i.e., a row of predictor values) is called a vector. So the goal of SVM modeling is to find the optimal hyperplane that separates clusters of vector in such a way that cases with one category of the target variable are on one side of the plane and cases with the other category are on the other size of the plane. The vectors near the hyperplane are the support vectors [6], [9].

4 Steps of project

Steps of project realization are shown in Fig. 2. It is difficult to obtain data of the real production system, so we used the simulation method on it. Simulation models of typical representatives of selected production systems are created in Witness program. These are systems of unit, serial and mass production. Created simulation models were used to generate production data about systems control. These data are stored in a data warehouse. These models serve for verify discovered knowledge. The discovered knowledge was implemented in the systems management strategies. This process represents a validation of discovered knowledge.

First, according to real production systems were developed simulation models of these systems. The simulation model was proposed control (management) strategy to achieve specific production objectives. The data were generated through simulation experiments. These were stored in a relational database. This procedure was repeated for multiple models, typical representatives of the production processes. The data warehouse was created and it represents a long-term data on the production of more types of production systems.

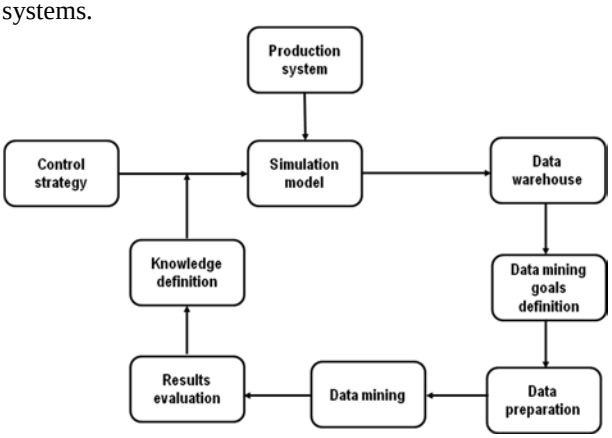

Figure 2. Steps of project realization

Data warehouse was built to better pre-processing and editing data. It serves as mine of data for data-mining.

The process based on the standard methodology CRISP-DM consists of the following key steps:

- Define the objectives of data mining, focusing on improving the manufacturing process,
- Data preparation: extraction of data sets from the data warehouse with respect to defined objectives, grouping and cleaning of data with respect to the chosen method of data mining,
- Data Mining: creating a model and application of data mining methods on the transformed data set,
- Evaluation of the results: an analysis with respect to the objectives,
- Definition of knowledge: the obtained results are formulated in the knowledge production process and its control and planning,
- According to the knowledge, the initial control strategy can be modified.

This process is iterative, you can return to any step in the process.

5 Project objectives

Production objectives are often conflicting and their achievement depends on many factors. Control and planning of production try to achieve the different production objectives in a given timeframe. Many dependencies are so far very little explored, for example relationship of capacity utilization and value of flow time, depending on the size of the production batch.

The problem of minimizing variable costs depending on the necessary operating supplies possession, alternatively with the possibility of increasing the value added percentage parameter (metric according to Lean Production). There are also few issues that are very little investigated, like the impact of priority rules for allocating the operations on the production objectives. These problems and dependencies would be possible to solve by using data mining methods [8].

The process of knowledge discovery from management of the production systems can be applied for solution of many problems. There were defined objectives that we try to obtain by using process of knowledge discovery in databases in the first stage of project. We analyzed the influence of some production process parameters as flow time of production batches, production equipment utilization and number of finished products.

Consequently we picked the relevant data. We selected the suitable data set from data warehouse with reference to defined objectives and used methods and techniques of data mining.

6 Application of knowledge

Based on the resulting reports from each data-mining methods and techniques, we proceeded to the evaluation. Evaluation was performed individually according to each goal. It was identified a parameter, which has a big influence on a production process – production batch. The new and more suitable values of lot size for production batch were gained by analyzing of reached results.

The lot size value is very important input parameter for management strategies of the production system. The lot size is defined as a number of pieces which is processed at the same time at one workplace with one-off (time) and at the same costs investment for its set up [11]. The lot size is one of the production directions which influence the production costs markedly. Therefor the lot sizes of batches require very accurate setting for the individual production system.

We can formulate general knowledge that greater or smaller lot sizes lead to improve parameters of production system. The gained knowledge for each goal is following:

- The lot size to ensure increased capacity utilization for the first product is 6 pieces, for the second product is 4 pieces, for the third product is 5 pieces.
- The lot size to ensure minimized flow time for the first product is 1 piece, for the second product is 2 pieces, for the third product is 2 pieces.
- The lot size to ensure increased the number of finished parts for the first product is 4 pieces, for the second product is 5 pieces, for the third product is 3 pieces.

The new discovered knowledge was applied to the designed simulation models. The process of knowledge discovery from the production system could by verified by this way. We compared the results of production system according to original management strategy with the results of production system that have been gained according to modified strategy.

Table 1. Partial results with priority on one production aim - capacity utilization

	Avg. flow time [s]	No. of product [pc]	Capacity utilization [%]
initial value	1329	8746	73,31
new value	1674	7358	90,37
difference	-25,96%	-15,87%	23,27 %

Table 1 presents a quantitative expression of the objectives that were achieved according to knowledge for influence of parameters analyses of production process, such as average flow time, number of products and capacity utilization. The minus and plus before the difference values indicate negative or positive influence to production system. When the value of average flow time is on the increase, it is negative for system, because the whole production process takes time. For number of finished products is negative influence its decrease, because it is a trend to produce maximum.

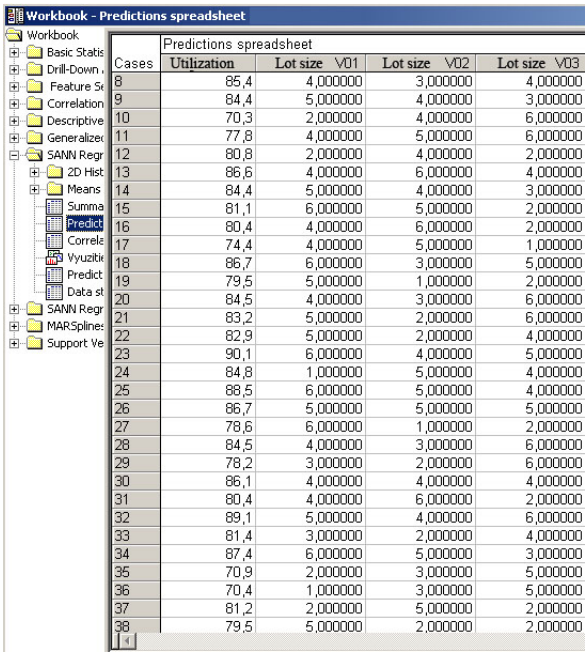

Figure 3. Obtained results of capacity utilization considering the lot size

The partial results are given in this article. The change of required parameter (see Table 1) i.e. capacity utilization was reached. Capacity utilization was increased from the original value 73,31% to the new value 90,37%. This change was caused by modification of lot size parameter as you can see in Fig. 3.

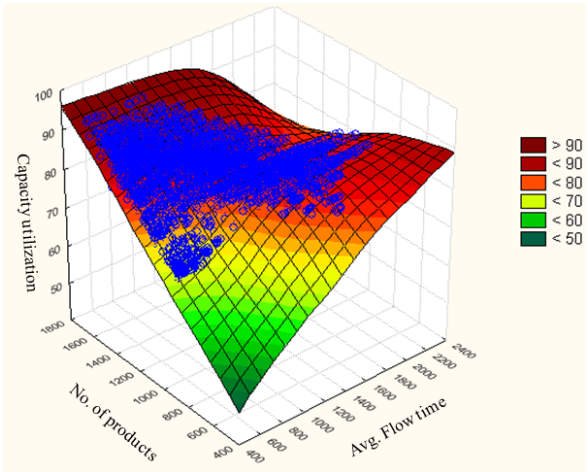

Figure 4. Plot data in 3D

Outcomes document the utilization of discovered knowledge to fulfill the defined objectives. Data are arranged in one display to allow for comparisons

between the subsets of data (categories) in Fig. 4.

In fact the increase of capacity utilization was ensured by changing of the lot size. However the next production objectives were changed at the same time. The number of products was decreased and production flow time was increased simultaneously. Because of the both parameters got worse, the changes of these parameters can be evaluated in a negative way from the point of view of the production system management. This fact means that the production objectives are contradictory.

The discovered knowledge was resulted in the improvement of production process and the defined objectives were fulfilled simultaneously. It is important to mention that some parameters of production process got worse in some cases at the same time. Therefore it is necessary to determine the target priority. Consequently the application of process KDD to the management strategies will be realized on the basis of priority system. Understandably the priority system will influence the parameters selection that will be monitored, collected and processed.

It is necessary to solve each specific case individually because it is not possible to define generally validated priority system for the global objectives of production systems analyses. For example if it is important to increase the number of finished products during the short time period then it will be needed to specify higher priority just to this parameter during the data mining process. The management strategy can be modified according to the new knowledge. The number of products increases although some parameters of production systems can get worse simultaneously. The very important part of this process is the determination of priority in the analysis of the production system. The priorities influence the stages of knowledge discovery process and also specifically gained knowledge.

Table 2. Results for strategy of all production goals together

	Avg. flow time [s]	No. of product [pc]	Capacity utilization [%]
initial value	1329	8746	73,31
new value	1234	9478	83,48
difference	7,15%	8,37%	13,87 %

All three parameters of the manufacturing process were improved. Improvement of parameter values is not as significant as the results of the analysis of separate parameters. Any of the parameters did not worsen unlike previous analyzes. As in previous cases, this analysis confirmed that the knowledge gained from the model contributed to the improvement of the analyzed manufacturing process parameters.

7 Conclusion

The aim of this article was to describe process KDD for planning and control manufacturing processes in industry. In the conclusion, we can observe that the bigger lot size ensures the increasing capacity utilization and the number of finished parts. On the other hand, the smaller lot size ensures minimizing of flow time. It is needed specific setting for the individual production system.

New discovered knowledge from production systems will help to identify the impact of manufacturing parameters on the system and the subsequent optimization of production system. And this will help managers in their decision making.

8 Acknowledgments

This contribution was written with a financial support VEGA agency in the frame of the project 1/0214/11 „The data mining usage in manufacturing systems control“.

References

[1] Chapman, P. et al. CRISP-DM 1.0 – Step-by-step data mining guide, http://www.spss.ch/upload/1107356429_CrispDM1.0.pdf, downloaded: March 12th 2012.

[2] Fayyad, U. M. et al. *From data mining to knowledge discovery: an overview*, In Advances in knowledge discovery and data mining. AAAI Press / The MIT Press, 1996.

[3] Fred, Y. W; Kang, K. Y. Applications of neural network in regression analysis, *Computers & Industrial Engineering*, Volume 23, Issues 1–4, 1992, pages 93-95, ISSN 0360-8352, http://www.sciencedirect.com/science/article/pii/036083529290071Q, downloaded: March 5th 2012.

[4] Friedman, H. J. Multivariate Adaptive Regression Splines, *In Annals of Statistics*, Volume 19, Number 1, pages. 1-67, 1991.

[5] Han, J; Kamber, M. *Data Mining: Concepts and Techniques*, Morgan-Kaufmann, Academic Press, San Francisco, 2001.

[6] Hill, T; Lewicki, P. *STATISTICS: Methods and Applications.* StatSoft, Tulsa, OK, 2007.

[7] Jackson, J. Data mining: A conceptual overview, *In Communications of the Association for Information Systems*, Volume 8, pages 267-296, 2002.

[8] Larose, D. *Data Mining Methods and Models*, John Wiley & Sons Ltd, New Jersey, 2006.

[9] Sherrod, H. P. Introduction to Support Vector Machine (SVM) Models, http://www.dtreg.com/ svm.htm, downloaded: March 7th 2012.

[10] Trnka, A. Results of application data mining algorithms to (lean) six sigma methodology. *International Journal of Engineering* 2012/1 ISNN 1584-2665

[11] Važan, P; Moravčík, O. The alternative procedure of lot size determination in flexible manufacturing systems, *Annals of DAAAM for 2007 & Proceedings of the 18th International DAAAM*, 24th -27th October, Zadar, Croatia, 2007, ISBN 3-901509-58-5.