Data Mining – Novel Statistical Method

[1]A. Horníková, [2] N.M. Durakbasa, [2]E. Güclü, [2]G. Bas

[1]University of Economics, Faculty of Economic Informatics, Dept. of Statistics, Dolnozemska cesta 1/b, 852 35 Bratislava, Slovakia,
[2]Vienna University of Technology, Institute for Production Engineering and Laser Technology, Dept. of Interchangeable Manufacturing and Industrial Metrology, Karlsplatz 13/3113, 1040 Vienna, Austria
Email: adriana.hornikova@euba.sk

Abstract. *Data Mining (DM) is a new and very dynamic discipline that is oriented on finding new knowledge in databases. DM is predicted to be „one of the most revolutionary developments of the next decade", according to the online technology magazine ZDNET News. [7] DM is a set of methods originating from practically attained knowledge. They are based though on a thorough statistical basis and machine learning. DM is usually used in following steps: starting with problem definition, tasks allocation, processes description, variables assignment, up to learning and evaluating the ensemble models, all phases should be explained in detail. This paper presents the general principles of data mining methodology applicable to re-calibration intervals to minimize the costs of re-calibration of measurement devices.*

Keywords: Data Mining, Applied Statistics, Clustering, Classification, Association Rules

1. Introduction

It was the beginning of the nineties of last century when scientists started to call this scientific discipline **knowledge discovery in databases or knowledge discovery process**. Knowledge Discovery in Databases (KDD) is the name coined by Gregory Piatetsky-Shapiro in 1989 to describe the process of finding interesting, interpretable, useful and novel data. Later the term of DM has become a synonym for the whole process of knowledge discovery in databases. The major distinguishing characteristic of DM is that it is data driven as opposed to other methods that are often model driven. [1] The evolution of KDD has undergone three distinct phases:

1. first generation systems that have been providing only one DM technique with very weak support for the overall process framework,
2. second generation systems (suites) provided multiple types of integrated data analysis,
3. third generation systems introduced the vertical approach enabling to address specific problems.

In today´s understanding is DM an interactive and iterative process of finding knowledge in experimental data sets. It comprises from following steps: problem specification, hypothesis building, data collection, pre-processing of collected data, model building or estimation, results interpretation and summarizing. [10]

2. Philosophy of Data Mining

DM methodology nowadays uses one of these two approaches for obtaining results: predictive or descriptive approach. Predictive approach uses the known variables for prediction of unknown values of other variables. Different approaches are used for different tasks. Descriptive methods use the pattern recognition approach that uses the description

process of experimental data and that can be interpreted according to statistics. Output from the knowledge discovery in databases is the generated new knowledge, usually described in terms of rules, patterns, classification models, associations, trends, statistical analysis, etc.

But what is the actual purpose of DM ? It is the process of making decisions. Decisions in organizations should be based on extensive Data Mining and analytics to model what-if scenarios, forecast the future, and minimizes risks. Nowadays we can apply DM in energy consumption predictions, prediction of exchange rates on markets, classification of bank customers or insurance companies customers, analysis of service providers change, reliability analysis for different kinds of machines or their parts, analysis of patients in hospitals, analysis of the consumers´ baskets and similar with large data sets. [2, 4]

3. Methodology of Data Mining

To formalize the knowledge discovery process within a common framework introduced was the process model concept or the standardized process model. Roughly DM steps are: to pre-processing raw data, mine the data, and interpret the results. In general there are several standard methodologies currently enabling to use DM.
Once the objective for DM is known, a target data set must be assembled. A data source is a datamart or data warehouse. Pre-processing of raw data involves "cleaning" of data which is the dismissal of noise or missing data. The cleaned data is reduced into feature vectors, usually one vector per observation. Feature vectors are divided into two sets, the "training set" and the "test set" (and the validation set). The training set is used to "train" DM algorithm(s), while the test set is used to verify the accuracy of any patterns found. For mining the data are commonly used four classes of tasks: classification, clustering, regression and association rules. The final step of KDD is to verify the patterns produced by DM algorithms.
Not all patterns found by DM algorithms are necessarily valid (overfitting). Further a number of statistical methods may be used to evaluate algorithms by the lift chart, the ROC (receiver operating characteristic) chart, the profit chart, the Lorentz curve, the K-S assessment chart (to evaluate a combination of two or more of the models one can use to evaluate the ensemble model). If learned patterns do not meet the desired output requirement then it is necessary to re-evaluate, update pre-processing and DM phases.

4. Usage of Data Mining

DM should be a white-box approach with understanding of the algorithm and should be implemented in conjunction with utilized software. There are currently in use several wide-spread methods for discovering knowledge in databases. The dominating methods are: SEMMA methodology (Sample, Explore, Modify, Model and Assess), 5A methodology (Assess, Access, Analyze, Act and Automate) and CRISP-DM methodology (the CRoss Industry Standard Process for Data Mining). [3-4]
SEMMA model´s abbreviation stands for Sample (identify input datasets), Explore (explore datasets statistically and graphically), Modify (prepare data for analysis), Model (fit a predictive model) and Assess (compare predictive models). Specialized licensed module of the SAS Company package dedicated to DM is the SAS Enterprise Miner®. Other software for DM tasks is WEKA (Waikato Environment for Knowledge Analysis), Minitab statistical software, Traceis software and many more.
As a result of an European research project was created CRISP-DM methodology. The aim was to create a new framework (standard) on higher level of generalization that would be useful for any DM application. CRISP-DM is an iterative and adaptive process of six basic steps that can be used in differing order when analyzing a DM related problem:

1. research understanding phase or business understanding phase (with several sub-steps: determination of business objectives, assessment of the situation, determination of DM goals and generation of a project plan),
2. data understanding phase (with several sub-steps: collection of initial data, description of data, exploration of data and verification of data quality),
3. data preparation phase (with several sub-steps: selection of data, cleansing of data, construction of data, integration of data and formatting of data subsets),
4. modelling phase (with several sub-steps: selection of modelling techniques, generation of test design, creation of models and assessment of generated models),
5. evaluation phase (with several sub-steps: evaluation of results, process review and determination of the next step) and
6. deployment phase (with several sub-steps: plan deployment, plan monitoring and maintenance, generation of final report and review of process sub-steps).

This methodology can be represented by a circle of DM project having six phases. The order of steps implementation is not fixed, just outputs of one step influence the selection of approaches within the next step. Sometimes it is needed to re-start and re-evaluate the analysis. A circle becomes then a suitable symbol for representing cycles of CRISP-DM methodologies´ steps. CRISP-DM methodology is supported by Clementine® DM software suite by SPSS.

5. Re-Calibration Interval Application

The supervision of measuring equipment is an essential quality requirement for modern production especially at the higher demands of micro and nanotechnology. The efficiency of the confirmation can be increased and expenses can be reduced substantially through computer assistance with flexible checking intervals. A special method developed at the Institute for Metrology for this purpose allows increasing of the flexibility level and efficiency of a system for the intelligent management and supervision of measuring devices.

On the basis of preceded calibrations the intervals are to be shortened if necessary, to secure the precision continually. They can be also enlarged, if it from the calibration results clearly emerges, that this measure the trust in the precision does not hurt the measuring and test devices. The system must ensure that the measuring and test equipment will be calibrated according to the determined timetable.

"Optimal Interval" is defined as the one having the total costs at minimum. If the interval is chosen too small, the checking costs goes up, because there are more checking in the equal period than necessarily. If the interval is chosen too large it raises the probability to find the measuring and test equipment as "inadmissible" at the next checking. Poor quality is joined always with a rise of the costs [6]. But how to model and set the optimal intervals ?

The size of it depends on a number of different factors: frequency of utilization, mode of using, behavior of abrasion, consequences at lapses, permissible tolerance range, number of users, status in the calibration chain, etc [5]. Since these elements are not temporally constant, the optimal interval cannot be a constant size either.

For this purpose has been a special method developed, which is based on the artificial intelligence [7]. This method is based on the demand, to consider on the environment conditions of the past as well as of the expected future. This happens exclusively through the application of fuzzy-logic in conjunction with clustering approach (DM based method), which operates also with linguistic variables.

Three input variables shall be enough for defining an output variable, which means that there is a three-dimensional model. The mathematical method namely, fuzzy clustering, fuzzy k-

means clustering, could be suitable. These methods compute the membership function that represents the presence of each observation to a cluster, and a membership score.
Other clustering methods those are more easily utilizable as they are a part of several packages for DM. We have compared several clustering approaches (hierarchical as well as non-hierarchical clustering methodologies) for creating clusters of measurement devices.

6. Conclutions

DM is defined as the process of extracting patterns from data. The authors have applied selected DM methods of a database of re-calibration intervals of measurement devices with the purpose to minimize the re-calibration costs. Adjustable quantity is the length of the re-calibration interval which is seen by the authors as not constant. This method aims a financial efficiency of the re-calibration process.
DM has been in recent years widely used in area of science and engineering, such as bioinformatics, genetics, medicine, education and electrical power engineering. But there is also an ethical horizon of DM. It requires data preparation which can uncover information or patterns which may defeat confidentiality and privacy obligations. A common way to resolve this is through data aggregation. [3]

Acknowledgements

Authors would like to thank the Slovak Research and Development Agency and the OeAD-GmbH/ICM Centre for International Cooperation and Mobility for the financial support of the Slovak-Austrian Bilateral Cooperation project entitled Statistical Analysis in Support of Technical Development and Evaluation of Measurement (SASoTDEM) with number SK-AT-0013-10 within the framework of the Slovak-Austrian Research and Development Cooperation. This contribution was elaborated with the support of the grant agency VEGA in the framework of the projects number 1/0543/10 and 1/0437/08.

References

[1] Fayyad, Usama; Gregory Piatetsky-Shapiro, and Padhraic Smyth (1996). "From Data Mining to Knowledge Discovery in Databases".
[2] 17.12.2008 at: http://www.kdnuggets.com/gpspubs/aimag-kdd-overview-1996-Fayyad.pdf.
[3] Cios, K. J., Pedrycz, W., Swiniarski, R. W., Kurgan, L. A.: Data Mining. A Knowledge Discovery Approach. Springer 2007. ISBN 978-0-387-33333-5.
[4] Terek, M., Labudová, V., Horníková, A.: HÍbková analýza údajov. Iura Edition, 2010.
[5] Osanna, P.H., Durakbasa, N.M.,: Prüfmittelüberwachung und –verwaltung im Qualitätsmanagement – leistungsfähig und flexibel durch Rechnereinsatz, Elektrotechnik und Informationstechnik, 4, pp. 207/212, 1998.
[6] Durakbasa, M.N., Pfeiffermann, G.G.: Computer Aided Confirmation and Management of Inspection, Measuring and Test Equipment in the Quality Management Systems, Proceedings „Measurement ‚97“, International Conference on Measurement, pp. 63/66, Smolenice, Slovak Republic, May, 1997
[7] Zadeh, L.A.: Fuzzy Sets, Information and Control; Vol.8 (1965), S 338/353.
[8] Myatt, G. J., Johnson, W P.: *Making Sense of Data II. A Practical Guide to Data Visualization, Advanced Data Mining Methods, and Applications.* USA: J. Wiley and Sons, 2009. ISBN 978-0-470-22280-5.
[9] February 8, 2001 at www.wikipedia.com/datamining
[10] Lyman, Peter; Hal R. Varian (2003). "How Much Information".Retrieved on 2008-12-17 at http://www.sims.berkeley.edu/how-much-info-2003.