METHODOLOGIES OF KNOWLEDGE DISCOVERY FROM DATA AND DATA MINING METHODS IN MECHANICAL ENGINEERING

Michał Rogalewicz1, Robert Sika2

1 Poznan University of Technology, Chair of Management and Production Engineering, Poland
2 Poznan University of Technology, Institute of Materials Technology, Poland

Corresponding author:
Michał Rogalewicz
Poznan University of Technology
Chair of Management and Production Engineering
Piotrowo 3, 61-138 Poznań, Poland
phone: (+48) 61 665-27-98
e-mail: michal.rogalewicz@put.poznan.pl

Received: 9 September 2016 Accepted: 1 December 2016

Abstract
The paper contains a review of methodologies of a process of knowledge discovery from data and methods of data exploration (Data Mining), which are the most frequently used in mechanical engineering. The methodologies contain various scenarios of data exploring, while DM methods are used in their scope. The paper shows premises for use of DM methods in industry, as well as their advantages and disadvantages. Development of methodologies of knowledge discovery from data is also presented, along with a classification of the most widespread Data Mining methods, divided by type of realized tasks. The paper is summarized by presentation of selected Data Mining applications in mechanical engineering.

Keywords
knowledge discovery, Data Mining methods, Data Mining methodology.

Introduction

Nowadays, competitiveness of companies is determined mostly by the possibility of fulfilling client’s needs in the best way possible. It means supplying the client with products or services which will meet his/her requirements in terms of proper timing, quality and price. Free market economy and the dominating role of the client make these requirements more and more diversified and individualized.

Because of this, companies aim at flexibility of realized processes, with simultaneous minimization of financial, material or energetic resources spent. It is related, on the one hand, to improvement of the main processes (limitation of variability [1], reduction of time, introduction of better supervision [2] and improvement techniques [3, 4]) and elimination of waste in the supplying processes [5] on the other hand. The proper approach to management in a company plays a significant role here, in support of projects for improvement, innovation, as well as analysis and implementation of solutions increasing the company’s competitiveness.

In modern times, progress in science, computing and technology, with observable results in almost every branch of human activity, lifted competing for clients and ways of fulfilling their needs to an entirely new level. In the production companies, modern solutions, e.g. Virtual Reality [6, 7] and Rapid Prototyping and Manufacturing [8], as well as integrated systems for communication with clients and suppliers, make it possible to significantly shorten the time needed for design [9] and manufacturing of a product and putting it in the market. Information technologies allow better management of processes realized in an organization (for example, production planning and organizing, manufacturing process control [10, 11]), while automation and modern manufacturing technologies allow obtaining products of very high quality.
Using modern technologies is related to gathering large amounts of data in all areas of company operation. Proper use of the data, often stored in large databases, to build process knowledge and realize constant improvement, as well as determination of the development strategy, is one of the challenges facing modern companies. This is particularly visible in the Industry 4.0 concept [12, 13]. This concept assumes the use of various modern information technologies, such as the Cyber-Physical Systems (CBS) or Internet of Things (IoT) – processing the Big Data [14]. The main idea here is the preparation of a computerized manufacturing environment, which will smartly allow us to increase flexibility and efficiency of production through integration of various activities and effective communication between a client and a producer (Customer to Business, C2B), as well as between a producer and a supplier (Business to Business, B2B) [15, 16].

In view of the above, methods of acquisition, gathering, processing and, most of all, exploration and analysis of data become particularly important. It seems that predictions of experts from an Internet magazine ZDNET News have come true – at the beginning of this millennium they stated that data exploration would be a revolutionary achievement of the following or coming decade [17, 18].

Discovery (extraction) of knowledge in large data sets is made possible thanks to the so-called Data Mining methods. The range of techniques used in DM is very wide. These are methods based on mathematical statistics and/or artificial intelligence. As a rule, they are used for soft modeling, as opposed to hard modeling (where models are based on differential equations from mathematical physics). These methods are used to model unknown phenomena with high level of complexity [19, 20]. In case of soft modeling, measurement results are used to build the models.

The next part of the paper presents reasons for application, advantages and limitations of the Data Mining methods. Then, development of methodologies of knowledge extraction from data is presented, as well as classification of Data Mining methods with respect to realized tasks. The final section indicates application areas of selected DM methods in the mechanical engineering industry.

**Reasons for application, advantages and limitations of Data Mining**

Computerization and development of computing techniques along with artificial intelligence methods brought about a rapid progress in development of methods of automated knowledge extraction from large sets of data. The uni- and multivariate statistical methods, known beforehand, ceased to be sufficient in view of the size of databases, but frequently were a starting point to prepare more methodologically complex, sophisticated tools, often making use of achievements in the field of artificial intelligence. In the beginning of the 21st century, the Data Mining methods are increasingly used because of the following reasons:

- The number of data sets and their size rapidly increases, along with the size of memory to store the databases and computing power needed to process the data. The only problem lies in discovering useful knowledge in these sets [21].
- Humans are unable to process large amounts of information in a “manual” way, so automated ways of extracting knowledge from data are necessary [22].
- DM methods allow rapidly obtaining the (often hidden) knowledge about analyzed products, processes and phenomena [23].
- They assist in decision making, for example in preparing prognoses and detection of frauds [24].
- They do not require performing very expensive experiments – they are based on already gathered data [25].
- They allow obtaining knowledge from data sets which are noisy, contain missing values or correlated variables – i.e., sets which are not dealt well by the traditional data analysis methods [26].
- They are universal and can be applied to a wide spectrum of problems [27].
- Dedicated methods were created to realize the knowledge extraction process, presenting step-by-step ways of conduct [28].
- The increase in literature in this is notable – the literature presents successful implementations of Data Mining methods. They may become an inspiration for new applications [29–31].

Unfortunately, application of Data Mining methods involves certain problems, limitations and threats. Some of them are listed below [21, 25]:

- There is a problem to ensure the safety of data gathered in databases, as well as of extracted knowledge.
- There is a threat to use Data Mining for a wrong purpose (unethically, against safety of a company, a country or its citizens).
- Improper use of Data Mining methods may lead to incorrect results, which means improper conclusions and decisions made on their basis.
- Implementation of a well-functioning system, which systematically utilizes the Data Mining
Data Mining methodologies applied in industry

From the point of view of the user, application of specific Data Mining methods for extraction of knowledge hidden in data is the least labor-consuming stage, in comparison with the often cumbersome and technically complicated preceding stages, related to understanding of a problem, proper data preparation, filtering and converting data with regard to a given task. In the knowledge extraction process the data exploration results can be obtained automatically. However, the preceding and the final stages (the latter focused on the analysis of obtained results) require the user to be familiar with problems of mathematics, statistics, as well as to have specialized knowledge regarding the studied branch (bank services, medicine, logistics and transport, production, etc.).

The authors propose a general scheme of approaching data analysis with regard to the knowledge discovery process using the DM methods – see Fig. 1.

Stage 1 concerns proper preparation of data for modeling, especially regarding elimination or minimization of gross errors through dealing with missing values, and removing outliers and unreal values. At this stage filtering should also be applied, i.e. selection of the appropriate range and type of data for analysis. For example it consists in choosing a specific product assortment or narrow down a range of variability of data, often from millions of data records. Data conversion should be applied if a specific DM methods use requires it (it most often consists in normalization or standardization of data). It is necessary, for example, in case of using neural networks, where it is recommended to use a MIN-MAX normalization during the model building stage (SOLVER) [21, 25, 33, 34]. Therefore, the danger of excess influence of the data order of magnitude on the model result is decreased.

At Stage 2, software is mostly used; specific Data Mining methods are implemented in order to perform the exploration tasks. Data sets prepared at Stage 1 are used at this point. The speed of work with the database and of building the model depends on the complexity of the problem itself, as well as on the type, amount and character of data (qualitative or quantitative data, the model obtained with a teacher or without a teacher, the number of data records and dependencies between input and output variables). This stage of use of the Data Mining methods is most frequently conducted automatically, while the time of its realization is related to the above mentioned problem complexity, but also performance of the computing equipment itself. Data Mining solvers do not require potent graphics cards or capacious hard disks (more and more often cloud data distribution and grid computing methods are used), but it is important to have a good CPU and a large amount of RAM [35]. However, a powerful computer unit with a properly configured graphics card can significantly accelerate the computing [36].

Stage 3 concerns interpretation of results obtained at Stage 2. It is important to involve in the analysis an expert in statistical methods and Data Mining. Insofar as Stage 2 may seem relatively simple, especially using capabilities of the modern computers, the other stages require expert knowledge.

Figure 1 presents the idea of the knowledge discovery process for a specific case, which is usually realized in science. In companies, complex of data exploration projects are realized, which require coordinated efforts of experts from many branches and
company divisions. In literature, various Data Mining methodologies are proposed, in form of scenarios of gathering and preparing data for further analysis, as well as dissemination of results for implementation of certain solutions. Below, selected concepts of data explorationscenario are presented, focusing only on description of the most frequently used ones (CRISP-DM, KDD, SEMMA, VC-DM).

The methodology which is most frequently used in practice and cited in literature is known as the CRISP-DM (CRoss-Industry Standard Process for Data Mining) [21, 37]. This name was proposed in 1996 by a consortium formed by 3 companies: Daimler Chrysler AG (Germany), SPSS Inc. (USA) and NCR Systems Engineering Copenhagen (USA and Denmark) with support from a bank – OHRA Verzekeringen en Bank Groep B.V (Netherlands). Initially, version 1.0 was developed; since 2006 version 2.0 has been in use. According to the CRISP-DM concept, the lifecycle of a data exploration project consists of 6 stages (Fig. 2).

In the CRISP-DM methodology, particular attention is paid to the understanding of business conditions of data. It is visible by synergic connections between the three first stages, which can be treated as the pre-processing stage (see Fig. 1). The next stage is the modeling (main-processing), while the two last ones consist in assessment of obtained results and implementation of results acquired on the basis of modeling.

The term Knowledge Discovery in Database (KDD) was used in 1991 for the first time [38]. Subsequent work [39, 40] led to establishing the KDD process as a methodology, described in [41, 42]. It contained results of cooperation of many researchers and business analysts. Fayyad et al. do not focus on description of the DM methods themselves, but they claim that their use is a part of the KDD process, which is used to discover new knowledge. According to Fayyad et al. [42], the KDD process comprises 5 stages:

- Selection (of a data set).
- Pre-Processing (data cleaning and preparation to modeling).
- Transformation (converting data for application of a specific method).
- Data Mining (use of DM tools to search for hidden patterns).
- Interpretation/Evaluation (interpretation and assessment of “unearthed” knowledge).

Despite Fayyad using the “Pre-Processing” notion only for Stage 2, the authors of this paper propose to use this term to label the first three stages (compare with Fig. 1). The fourth stage is the Main-Processing, while the last one is the Post-Processing.

Another methodology, used just as frequently, assumed to be competitive to CRISP-DM, is known as SEMMA. The name was proposed by Bulkley in 1991, but it was not commercially implemented until 2008 [43]. It is used in the Enterprise Miner (EM) software and obviously is most effective with this software. There are 5 distinct stages of knowledge discovery, for which there are tools available in the EM:

- Exploration (Distribution Explorer, Multiplot, Insight, Association, Variable Selection, Link Analysis).
- Modification (Data Set Attributes, Transform Variables, Filter Outliers, Replacement, Clustering, SOM/Kohonen, Time Series).
- Model (Regression, Tree, Neural Network, Princomp/Dmneural, User Defined Model, Ensemble, Memory-Based Reasoning, Two Stage Model).
- Verification (Assessment and Reporter).

In the above presented approach, data analysis is started by identification of the research problem, while further exploration is conducted on a data sample obtained from a larger data set. Then, relations between data are mostly looked for using data visualization tools (Explore) and the data set is prepared for modeling (Modification). These first three stages can be included in the pre-processing phase (compare with Fig. 1). Subsequently, the DM techniques are used to discover hidden knowledge (Model – the main processing phase). The last stage (Assess) is the
evaluation of obtained results and attempt at their translation into real conditions of company functioning (post-processing phase).

Independent studies [37] indicate that the three methodologies described above were the most frequently used ones between 2007 and 2014. It is worth noting that in 2008 Azevedo et al. [28] proposed guidelines aimed at establishing a connection between the CRISP-DM, KDD and SEMMA methodologies.

The last methodology considered significant by the present authors, despite being rarely used, is known as the Virtuous Cycle of Data Mining (VC-DM), proposed by Berry and Linoff as early as in 1997 [32, 44]. It consists of four basic stages:

• Identify the business problem.
• Transform data into actionable result.
• Act on the information.
• Measure the results.

In the first stage, which can be included in the pre-processing (compare with Fig. 1), the problem or group of problems which will possibly be solved by data exploration must be identified. Specialists from selected divisions of a company should cooperate with analysts and clearly define the problem(s), along with presenting the frequency or special circumstances of its/their occurrence. The second stage can be partially included in the pre-processing, as it is related with appropriate preparation of data for modeling (cleaning, transformation, selection of learning sample). On this stage, selection of a DM method and data exploration with the selected methods (main-processing) also take place. The third and the fourth stage, which can be both included in the post-processing, are directly related to dissemination of results in a company, i.e., the use of obtained models e.g. for process improvement and evaluation of results, which allows verification of their effectiveness and applicability. A good practice is isolation of a validation set, to which the Data Mining methods were not applied, from the analyzed data, and applying this set to DM model.

As mentioned earlier, only the most frequently used and worth noting methodologies of knowledge discovery from data were presented. They are summarized in the context of the three main aspects of knowledge discovery from data (compare with Fig. 1) in Table 1. Description of other methodologies can be found in the literature [43, 45, 46].

### Classification of Data Mining methods

Development of artificial intelligence, statistical methods, machine learning and computational intelligence makes for more and more Data Mining methods. They can be classified according to various criteria: simplicity of operation and implementation, speed of operation, scalability or way of data processing. However, the most widespread way of classifying the Data Mining methods is by realized tasks. The most frequently distinguished tasks are [21, 30, 47]:

- Description.
- Classification.
- Regression.
- Clustering.
- Looking for associations.

**Description** consists in concise summarizing of analyzed data. During realization of this task, graphs are frequently used alongside one-dimensional or multidimensional tables or rules for data description.

**Classification** belongs to the so-called supervised (“learning with teacher”) group and is aimed at creating a dependency model between independent variables describing given objects or phenomenon and a dependent variable in an attribute form. It is done on the basis of the so-called teaching set, containing a set of objects with known values of independent and dependent variables. The purpose is to

### Table 1

Comparison of selected Data Mining methodologies in three main aspects of knowledge discovery process [own work].

<table>
<thead>
<tr>
<th>DM Methodology</th>
<th>Pre-Processing</th>
<th>Main-Processing</th>
<th>Post-Processing</th>
</tr>
</thead>
<tbody>
<tr>
<td>CRISP-DM</td>
<td>Business Understanding</td>
<td>Model</td>
<td>Evaluation, Deployment</td>
</tr>
<tr>
<td>CRoss-Industry Standard Process for Data Mining</td>
<td>Data Understanding</td>
<td></td>
<td></td>
</tr>
<tr>
<td>KDD</td>
<td>Selection, Pre-Processing</td>
<td>Data Mining</td>
<td>Interpretation and Evaluation</td>
</tr>
<tr>
<td>Knowledge Discovery in Database</td>
<td>Transformation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SEMMA</td>
<td>Sample, Explore, Modify</td>
<td>Model</td>
<td>Assess</td>
</tr>
<tr>
<td>Sampling, Exploration, Modification, Model, Verification</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VC-DM</td>
<td>Identify</td>
<td>Transform</td>
<td>Act, Measure</td>
</tr>
<tr>
<td>Virtuous Cycle of Data Mining</td>
<td>(Pre-Processing and Main-Processing)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

...
apply this model for assigning new cases to a selected class of the dependent variable. The most frequently used methods in this area are classification trees, neural networks, support vector machines, the naive Bayesian classifier or Bayesian networks.

**Regression** also belongs to the supervised group and plays a similar role to classification, but in the dependency model created on the basis of the teaching set, the dependent variable is in a numerical form. Examples of methods used to realize the regression task are SVM, neural networks, simple and multiple regression, and regression trees.

**Clustering** does not use a teacher (there is no dependent variable here) and consists in creating clusters (groups) of objects in a way to ensure the highest possible similarity between objects in one cluster, as regards values of the considered independent variables, with simultaneously maintained maximal possible differences between particular clusters. In this area, two groups of methods are applied: hierarchical ones, building the so-called dendrograms, and non-hierarchical ones, creating entirely separate clusters.

**Looking for associations** consists in finding dependencies in an analyzed data set. These dependencies do not have a functional character; rather, they are based on coexistence of values of particular variables.

Figure 3 presents the classification of methods and techniques used to perform the above mentioned tasks. Only the most frequently used methods are considered.

---

Fig. 3. Classification of Data Mining methods (own study based on [48–51]).
Review of Data Mining methods used in mechanical engineering

Data Mining is successfully applied in various areas of human activity: telecommunication, banking, transport, aeronautics, and marketing [21, 27, 29, 30, 32, 44, 52]. Regarding mechanical engineering, it seems that they do not play a major role just yet. The following factors may influence this [49, 53]:

• Most researchers operating in the industry are not familiar with the DM methods, at the same time, the DM specialists do not know complex production processes well.
• Industrial data is often confidential or sensitive – it makes it harder to perform analysis.
• It is difficult to unequivocally estimate benefits and efficiency of DM methods implemented in the industry.

However, the literature indicates that the appreciation of data mining in industrial applications is gradually growing [54–56]. In Table 2, the present authors review the successful applications of selected Data Mining methods in mechanical engineering. Information about the area of use and short description of contents is included.

<table>
<thead>
<tr>
<th>Table 2</th>
<th>Review of applications of selected Data Mining methods used in mechanical engineering.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Area</td>
<td>Authors</td>
</tr>
<tr>
<td>Total Preventive Maintenance</td>
<td>Djatna T. et al. 2015 [57]</td>
</tr>
<tr>
<td>Fault diagnosis</td>
<td>Hu Y. et al. 2015 [58]</td>
</tr>
<tr>
<td></td>
<td>Jia Z. et al. 2013 [59]</td>
</tr>
<tr>
<td>Failures in the manufacturing process diagnosis</td>
<td>Martinez-de-P. F.J. et al. 2012 [60]</td>
</tr>
<tr>
<td>Product design and development</td>
<td>Zhang L. et al. 2011 [61]</td>
</tr>
<tr>
<td></td>
<td>Yang X. et al. 2008 [62]</td>
</tr>
<tr>
<td></td>
<td>Shahbaz M. 2006 [63]</td>
</tr>
<tr>
<td>Clustering</td>
<td>Sobh A.S. et al. 2015 [64]</td>
</tr>
<tr>
<td></td>
<td>Hayajneh M. 2005 [65]</td>
</tr>
<tr>
<td></td>
<td>Jing H.L. et al. 2011 [66]</td>
</tr>
<tr>
<td></td>
<td>Zhou X. et al. 2006 [67]</td>
</tr>
<tr>
<td>Area</td>
<td>Authors</td>
</tr>
<tr>
<td>-------------------------------------------</td>
<td>----------------------------------</td>
</tr>
<tr>
<td><strong>Classification</strong></td>
<td></td>
</tr>
<tr>
<td>Products defects classification</td>
<td>Ma H.W. et al. 2011 [69]</td>
</tr>
<tr>
<td>Fault classification</td>
<td>Muralidharan V. et al. 2013 [70]</td>
</tr>
<tr>
<td>Fault classification</td>
<td>Muralidharan V. et al. 2012 [71]</td>
</tr>
<tr>
<td>Fault diagnosis</td>
<td>Jegadeeshwaran R. et al. 2015 [72]</td>
</tr>
<tr>
<td>Pattern recognition in control charts</td>
<td>Moosavian A. et al., 2013 [73]</td>
</tr>
<tr>
<td>Tool condition monitoring</td>
<td>Lesany S.A. et al. 2014 [74]</td>
</tr>
<tr>
<td><strong>Regression</strong></td>
<td></td>
</tr>
<tr>
<td>Fault diagnosis</td>
<td>Yasa R. et al. 2014 [76]</td>
</tr>
<tr>
<td></td>
<td>Perzyk M. et al. 2014 [77]</td>
</tr>
<tr>
<td>Quality prediction</td>
<td>Lu Z.J. et al., 2015 [78]</td>
</tr>
<tr>
<td>Manufacturing process control</td>
<td>Jin R. et al., 2012 [79]</td>
</tr>
<tr>
<td>Process optimization</td>
<td>Pashazadeh H. et al., 2016 [80]</td>
</tr>
<tr>
<td>Product design</td>
<td>Verbert J. et al., 2011 [81]</td>
</tr>
<tr>
<td>Materials properties study</td>
<td>Mareci D. et al., 2013 [82]</td>
</tr>
<tr>
<td>Area</td>
<td>Authors</td>
</tr>
<tr>
<td>------</td>
<td>---------</td>
</tr>
<tr>
<td>Quality control in foundry</td>
<td>Perzyk M. et al., 2014 [83]</td>
</tr>
</tbody>
</table>

**Description**

<table>
<thead>
<tr>
<th>Production improvement (optimization)</th>
<th>Jansen F.E. et al., 1996 [84]</th>
<th>In the article some Exploratory Data Analysis tools are suggested to be useful in description of dynamic flow process in the reservoir. The paper presents a simple approach to examine the interwell communication and interference of a mature water flood in order to identify and rank areas of potential improvement.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Abonyi J., 2007 [85]</td>
<td>The article presents some Exploratory Data Analysis tools to use in the description of polyethylene production. Box plots and quantile-quantile plots help to analyse the relationships between different operating and product quality variables.</td>
</tr>
</tbody>
</table>

**Conclusions**

The paper presents Data Mining methodologies and methods the most frequently used in industry, focusing on the mechanical engineering field. Dynamical development of systems and production processes, along with ongoing automation in connection with the mass customization of products are a reason behind new requirements towards available methods of data analysis. In large databases, there is more and more data gathered, but automated discovery of useful knowledge from this data is still a problem in many companies. The Data Mining methods can be used to meet these requirements. As indicated in the paper, they are more and more often used in industry in areas of fault diagnosis and product design and development. They can also play a significant role in supervision and control of processes, as well as in detection of irregularities.

The authors predict increasing interest in the Data Mining methods used in the mechanical engineering industry. Successful applications indicated in the paper may inspire to search for new application areas.

**References**


Chen F., Deng P., Wan J., Zhang D., Data Mining

Lee J., Kao H., Yang S., Scientists want more computing power

Schuh G., Potente T., Varandani R., Schmitz T., Discovering Knowledge in Data: An Introduction to Data Mining


Gorecky D., Schmitt M., Loskyll M., Zuhlke D., Data Mining as a Chance for Reliability of New Models Applied in Selected Simulation Systems


Larose T., Discovering Knowledge in Data: An Introduction to Data Mining, Wiley & Sons, 2005.


[68] Yiakopoulos C.T., Gryllias K.C., Antoniadis I.A., Rolling element bearing fault detection in indus-


