Management and Production Engineering Review
Volume 7 • Number 4 • December 2016 • pp. 97–108
DOI: 10.1515/mper-2016-0040

DE G DE GRUYTER OPEN

mper

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

Michał Rogalewicz[1], Robert Sika[2]

[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

methods, requires application of large resources (maintenance and updating of databases, hiring specialists for knowledge extraction etc.); not every company is capable of doing it.

- Expectations from the users' side that the Data Mining methods will replace them in drawing conclusions and making decisions – are unfounded.

Despite these problems, threats and limitations, it seems that the Data Mining methods will achieve higher and higher popularity, not only in the areas where they have already been used (customer relationship management, fraud detection, banking etc. [32]), but also in the industry, which so far managed to virtually "neglect" them. Methodologies of conduct during the Data Mining implementation, developed along the Data Mining methods themselves, may surely become useful here. More or less universal, they help a user to consider all the necessary steps to properly realize a process of knowledge extraction from the data. The next section describes development and characteristics of Data Mining methodologies used in industry.

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.

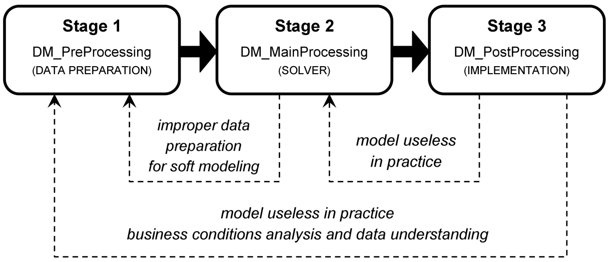
Stage 1
DM_PreProcessing
(DATA PREPARATION)
Stage 2
DM_MainProcessing
(SOLVER)
Stage 3
DM_PostProcessing
(IMPLEMENTATION)
improper data preparation for soft modeling
model useless in practice
model useless in practice
business conditions analysis and data understanding

Fig. 1. Diagram of approach to chronological knowledge discovery process [own work].

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 exploration scenario 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).

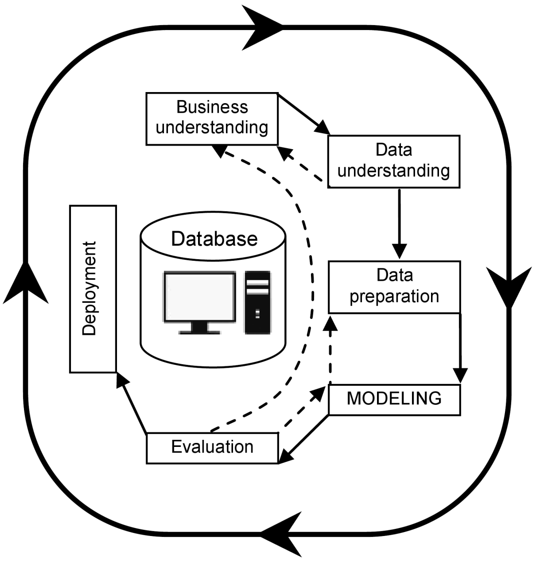

Fig. 2. CRISP-DM methodology diagram (own study based on [33]).

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:

- Sampling (Input Data Source, Sampling, Data Partition).
- 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

Table 1
Comparison of selected Data Mining methodologies in three main aspects of knowledge discovery process [own work].

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

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 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.

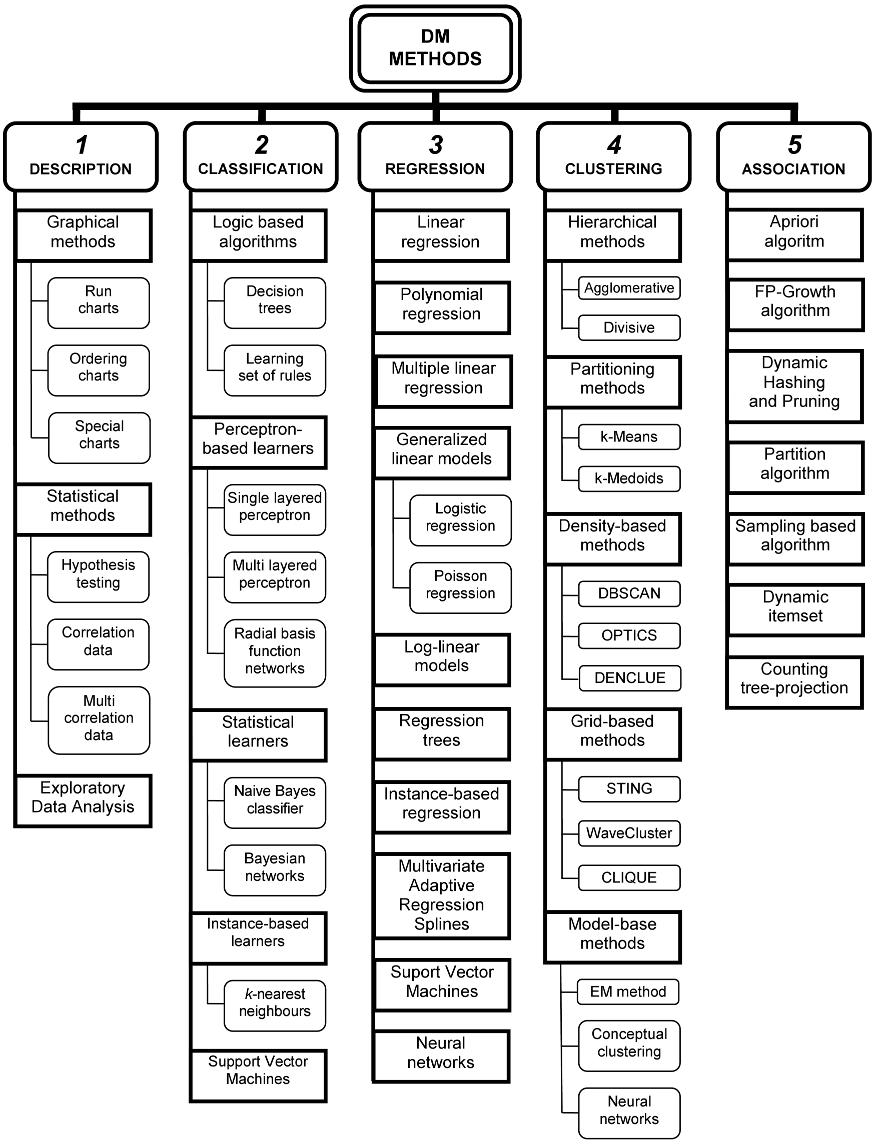

Fig. 3. Classification of Data Mining methods (own study based on [48–51]).

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 2
Review of applications of selected Data Mining methods used in mechanical engineering.

Area	Authors	Short description
Looking for associations		
Total Preventive Maintenance	Djatna T. et al. 2015 [57]	Formulation of Association Rule Mining finds rules that show the well computed relationship between measurable indicators of OEE with the response of action required to take in certain condition of machine utilization
Fault diagnosis	Hu Y. et al. 2015 [58]	Association rules proved feasible to build agricultural machinery maintenance fault knowledge base
	Jia Z. et al. 2013 [59]	Associated fault diagnosis rules in warship's power plants help to judge the type of fault appeared and predict future faults
Failures in the manufacturing process diagnosis	Martinez-de-P. F.J. et al. 2012 [60]	Use of association rules from multiple time series captured from industrial process. The main goal is to seek useful knowledge for explaining failures in the process of galvanizing steel coils
Product design and development	Zhang L. et al. 2011 [61]	Approach based on association rule mining to identify mapping relationships between functions and technologies. A case study of microlithography machines is reported to demonstrate how the proposed approach identifies mapping relationships from given function and technology data and how these relationships help select technologies and determine details
	Yang X. et al. 2008 [62]	This paper applies association rule mining to reveal the mapping relations between customer affective needs and configuration of design elements. A goodness criterion was introduced to refine association rules. A case study of Volvo truck cab design was conducted based on the proposed methods
	Shahbaz M. 2006 [63]	The authors show the methodology helping to apply association rules to extract knowledge from product and processes databases aiming to improve the design of the product and discover some constraints in manufacturing system. They applied their methodology to fan blade production process.
Clustering		
Materials properties clustering	Sobh A.S. et al. 2015 [64]	The authors adopt hierarchical clustering to create clusters of engineering materials having the same properties
Product quality prediction	Hayajneh M. 2005 [65]	The author used fuzzy subtractive clustering based system identification and Sugeno type fuzzy inference system to develop a model of process parameters influence on surface quality after fine turning process
Product design	Jing H.L. et al. 2011 [66]	This paper shows a retrieval efficiency of 3D mechanical models improvement with the use of k-means clustering method
Process anomaly detection	Zhou X. et al. 2006 [67]	The authors used an improved k-means algorithm to cluster historically disqualified products data. It is used as a prediction model to detect and prevent from quality problems. They implemented this approach in gear-plant.
Fault detection	Yiakopoulos C.T. et al. 2010 [68]	K-means clustering partitional method used to identify and classify bearing defects

[Table 2. Cont.]

Area	Authors	Short description
Classification		
Products defects classification	Ma H.W. et al. 2011 [69]	The authors proposed an algorithm connecting rough set approach based on entropy with multi-class v-SVM based on binary tree to defects classification of steel cord conveyor belt
Fault classification	Muralidharan V. et al. 2013 [70]	In authors' approach vibration signals are used for fault diagnosis of centrifugal pumps using wavelet analysis. They used rough set theory to generate the rules from the vibration signals
	Muralidharan V. et al. 2012 [71]	A vibration based condition monitoring system was introduced. Wavelet analysis for feature extraction was used and Naive Bayes algorithm and Bayes nets were compared for fault diagnosis of monoblock centrifugal pump
	Jegadeeshwaran R. et al. 2015 [72]	In this study the authors introduce a Clonal Selection Classification Algorithm (CSCA) for condition monitoring of a hydraulic brake system. through vibration analysis. The algorithm is used for a brake fault diagnosis
Fault diagnosis	Moosavian A. et al., 2013 [73]	The paper introduces an idea of fault diagnosis on a main engine journal-bearing. The authors used power spectral density (PSD) technique and based their scheme on K-nearest neighbor and artificial neural network
Pattern recognition in control charts	Lesany S.A. et al. 2014 [74]	The authors suggest a model for the recognition of patterns on control charts with the use of LVQ and MLP networks along with a fitted line analysis. Model is also usefeul when common paterns appear simultaneously
Tool condition monitoring	Brezak D. et al. 2012 [75]	The authors suggest a hybrid tool wear estimator consisting of two modules: for classification and regression. The former utilizes analytic fuzzy logic concept and the latter support vector machine algorithm
Regression		
Fault diagnosis	Yasa R. et al. 2014 [76]	The authors apply decision trees and nonlinear regression approaches to develop engineering design formulae for estimation of the current induced scour depth in both live bed and clear water conditions
	Perzyk M. et al. 2014 [77]	The authors used an evaluation of various methodologies to determine relative significances of input variables in data-driven models. Significance analysis applied to manufacturing process parameters can be a useful tool in fault diagnosis for various types of manufacturing processes.
Quality prediction	Lu Z.J. et al., 2015 [78]	The authors suggest using a prediction model based on support vector regression (SVR) to solve the yarn quality prediction problem in textile production management
Manufacturing process control	Jin R. et al., 2012 [79]	In the article a methodology for feedworward control based on a reconfigured piecewise linear regression tree was presented. The authors verify an effectiveness of this methodology in a multistage wafer manufacturing process
Process optimization	Pashazadeh H. et al., 2016 [80]	In the paper some results of Resistance Spot Welding process optimization were introduced. The authors used full factorial experiment and hybrid combination of the artificial neural networks and multi-objective genetic algorithm to specify optimal parameters of this process
Product design	Verbert J. et al., 2011 [81]	The authors describe and validate the use of MARS (Multivariate Adaptive Regression Splines) as a tool to predict deviations in uncompensated tests and improve the accuracy of parts produced by FSPIF (Feature Assisted Single Point Incremental Forming)
Materials properties study	Mareci D. et al., 2013 [82]	The authors studied the corrosion resistance of new ZrTi alloys to various working conditions using adaptive instance-based regression model designed for experiments with insufficient or missing data

[Table 2. Cont.]

Area	Authors	Short description
Quality control in foundry	Perzyk M. et al., 2014 [83]	Results of the paper authors' research concerning two important issues of RST (Rough Set Theory) and DTs (Decission Trees) applications in foundry production are presented. They include assessment of correctness of relative significances of process parameters of arbitrary nature (e.g. physical, human, organizational etc.) and evaluation of reliability of engineering knowledge in the form of logic rules.
Description		
Production improvement (optimization)	Jansen F.E. et al., 1996 [84]	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 waterflood in order to identify and rank areas of potential improvement
	Abonyi J., 2007 [85]	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

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

[1] Hamrol A., *Intelligent components for quality control in manufacturing*, Proc. of 3rd IFAC Symp. on Intel. Comp. and Instr. for Con., pp. 613–618, 1997.

[2] Hamrol A., Kowalik D., Kujawińska A., *Impact of selected work condition factors on quality of manual assembly process*, Hum. Fact. and Erg. in Man. and Serv. Ind., 21, 2, 156–163, 2011.

[3] Starzyńska B., Hamrol A., *Excellence toolbox: Decision support system for quality tools and techniques selection and application*, Tot. Qual. Man. and Bus. Exc., 24, 5–6, 577–595, 2013.

[4] Diering M., Dyczkowski K., Hamrol A., *New method for assessment of raters agreement based on fuzzy similarity*, Adv. in Intell. Sys. and Comp., 368, 415–425, 2015.

[5] Trojanowska J., Żywicki K., Varela M.L.R. et al., *Shortening changover time – an industrial study*, Proc. of the 2015 10th Iberian Conf. on Inf. Sys. and Tech., 2015.

[6] Górski F., Buń P., Wichniarek R., Zawadzki P., Hamrol A., *Immersive City Bus Configuration System for Marketing and Sales Education*, Proc. Comp. Sc., 75, 137–146, 2015.

[7] Pandilov Z., Milecki A., Nowak A., Górski F., Grajewski D., Ciglar D., Mulc T., Klaić M., *Virtual Modelling And Simulation Of A CNC Machine Feed Drive System*, Trans. of FAMENA, 39, 4, 37–54, 2016.

[8] Grajewski D., Diakun J., Wichniarek R. et al., *Improving the skills and knowledge of future designers in the field of ecodesign using virtual reality technologies*, Proc. Comp. Sc., 75, 348–358, 2015.

[9] Zawadzki P., Górski F., Hamrol A., Kowalski M., Paszkiewicz R., *An Automatic System for 3D Models and Technology Process Design*, Trans. of FAMENA, 35, 2, 69–78, 2011.

[10] Hamrol A., *Intelligent system for quality control in manufacturing*, Proc. of 7th Int. Conf. on Hum. Comp. Int., 21, 321–324, 1997.

[11] Hamrol A., *Process diagnostics as a means of improving the efficiency of quality control*, Prod. Plan. and Con., 11, 8, 797–805, 2000.

[12] Lee J., Kao H., Yang S., *Service innovation and smart analytics for Industry 4.0 and big data environment*, Prod. Serv. Sys. and Val. Creat., CIRP Procedia, 16, 3–8, 2014.

[13] Schuh G., Potente T., Varandani R., Schmitz T., *Global Footprint Design based on genetic algorithms – An "Industry 4.0" perspective*, CRIP Ann. Man. Tech., CIRP Procedia, 63, 1, 433–436, 2014.

[14] Chen F., Deng P., Wan J., Zhang D., *Data Mining review for the Internet of Things: Literature Review and Challenges*, Int. J. of Dis. Sen. Net., vol. 2015, Art. ID 431014, 14 pages, 2015.

[15] Gorecky D., Schmitt M., Loskyll M., Zuhlke D., *Human – Machine – Interaction in the Industry 4.0 ERA*, Ind. Inf. (INDIN), in 12th IEEE International Conference, pp. 289–294, 2014.

[16] Zawadzki P., Żywicki K., *Smart product design and production control for effective mass customization in the Industry 4.0 concept*, Manag. and Prod. Eng. Rev., 7, 3, 105–112, 2016.

[17] Rachel K., *Scientists want more computing power*, ZDNET Magazine, 2001, http://www.zdnet.com/article/scientists-want-more-computing-power/.

[18] Canepa M., *Virtual Data Storage, new uses*, ZDNET Magazine, 2002, http://www.zdnet.com/article/virtual-data-storage-old-concept-new-uses/.

[19] Ignaszak Z., Hajkowski J., Popielarski P., *Example of New Models Applied in Selected Simulation System with Respect to Database*, Arch. of Found. Eng., 13, 1, 45–50, 2013.

[20] Ignaszak Z., Hajkowski J., Popielarski P., *Sensivity of Models Applied in Selected Simulation System with Respect to Database Quality for Resolving of Casting Problems*, Def. and Diff. Forum., 334–335, 314–321, 2013.

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

[22] Morzy M., *Data Mining – Review of available methods and fields of application* [in Polish], Retrieved 15.10.2016, http://www.cs.put.poznan.pl/mmorzy/papers/cpi06.pdf.

[23] Fronczak E., Michalczewicz M., *Application of Data Mining tools to create models and knowledge management* [in Polish], Pol. Comp. of Know. Manag., Series: Studies and Materials, Vol. 27, 2010.

[24] Ngai E., Hu Y., Wong Y., Chen Y., Sun X., *The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature*, Dec. Supp. Sys., 50, 559–569, 2011.

[25] Tadeusiewicz R., *Data Mining as a chance for relatively cheap scientific perform of scientific discoveries digging seemingly unexploited empirical data* [in Polish], Statsoft Inc. Web Site, 2006, Retrieved 03.11.2016, www.statsoft.pl/czytelnia/.

[26] Zhang P., Zhu X., Shi Y., Guo L., Wu X., *Robust ensemble learning for mining noisy data streams*, Dec. Supp. Sys., 50, 469–479, 2011.

[27] Liao S., Chu P., Hsiao P., *Data mining techniques and application – A decade review from 2000 to 2011*, Exp. Sys. with App., 39, 11303–11311, 2012.

[28] Azevedo A., Santos M., *KDD, SEMMA and CRISP-DM: a parallel overview*, IADIS European Conf. Data Mining, pp. 182–185, 2008.

[29] Witten I.H., Frank E., Hall M.A., *Data Mining: Practical machine learning tools and techniques, 3rd edition*, Morgan Kaufmann Series in Data Management Systems, Elsevier, 2011.

[30] Han J., Kamber M., Pei J., *Data Mining: Concepts and Techniques, 3rd Edition*, Morgan Kaufmann Series in Data Management Systems, Elsevier, 2012.

[31] Wu X. and others, *Top 10 algorithms in data mining*, Know. Inf. Sys., 14, 1–37, 2008.

[32] Berry M.J.A., Linoff G., *Data mining techniques: for marketing, sales, and customer support*, Wiley & Sons, 1997.

[33] Manikandan G., Sairam N., Sharmili S., Venkatakrishnan S., *Achieving Privacy in Data Mining Using Normalization*, Ind. J. of Sc. And Tech., 6, 4, 4268–4272, 2013.

[34] Perzyk M., *Statistical and Visualization Data Mining Tools for Foundry Production*, Arch. of Foun. Eng., 7, 3, 111–116, 2007.

[35] KDnuggets, *Computing resources for analytics, data mining, data science work or research Pool*, 2015, Retrieved 05.11.2016, http://www.kdnuggets.com/polls/2015/computing-platform-hardware-analytics-data-mining.html.

[36] Dean J., *Big Data, Data Mining and Machine Learning. Value Creation for Business Leaders and Practitioners*, Wiley, 2014.

[37] KDnuggets, *CRISP-DM, still the top methodology for analytics, data mining, or data science projects*, 2014, Retrieved 25.10.2016, http://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html.

[38] Piatetsky-Shapiro G., Frawley W., *Knowledge Discovery in Databases*, MIT Press Cambridge, 1991.

[39] Frawley W., Piatetsky-Shapiro G., Matheus C.J., *Knowledge Discovery in Databases: An Overview*, Art. Int. Mag., 13, 3, 57–70, 1992.

[40] Piatetsky-Shapiro G., Matheus C.J., Smyth P., Uthurusamy R., *KDD-93: Progress and Chellenges in Knowledge Discovery in Databases*, Art. In. Mag., 15, 3, 77–82, 1994.

[41] Fayyad U.M., Piatetsky-Shapiro G., Smyth P., *From Data Mining to Knowledge Discovery in Databases*, Art. Int. Mag., 17, 3, 37–53, 1996.

[42] Faayad U.M., Piatetsky-Shapiro G., Smyth P., Uthurusamy R., *Advances in knowledge discovering and data mining*, American Association for Artificial Intelligence, 1996.

[43] Marban O., Mariscal G., Segovia J., *A Data Mining & Knowledge Discovery Process Model*, Dat. Min. and Know. Disc. Proc., INTECH Open Science, 2009.

[44] Berry M.J.A., Linoff G., *Mastering data mining*, Wiley & Sons, 2000.

[45] Alnoukari M., Sheikh A., *Knowledge Discovery Process Models: From Traditional to Agile Modeling*, IGI Glob., pp. 72–100, 2012.

[46] Rohanizadeh S.S., Moghadam M.B., *A proposed Data Mining Methodology and its Application to Industrial Procedures*, J. of Ind. Eng., 4, 37–50, 2009.

[47] Hastie T., Tibshirani R., Friedman J., *The Elements of Statistical Learning. Data Mining, Inference, and Prediction*, Springer, 2009.

[48] Jain K., Murty M., Flynn P., *Data clustering: a review*, ACM Comp. Surv. (CSUR), 31, 3, 264–323, 1999.

[49] Kotsiantis S.B., Zaharakis I.D., Pintelas P.E., *Machine learning: A review of classification and combining techinques*, Art. Int. Rev., 26, 159–190, 2006.

[50] Murtagh F., Contreras P., *Algorithms for hierarchical clustering: an overview*, Wiley Int. Rev.: Data Mining and Know. Disc., 2, 1, 86–97, 2012.

[51] Popat S., Emmanuel M., *Review and comparative study of clustering techniques*, Int. J. of Comp. Sc. and Inf. Tech., 5, 1, 805–812, 2014.

[52] Więcek-Janka E., Mierzwiak R., Kijewska J., *Competencies' model in the succession process of family firms with the use of grey clustering analysis*, J. of Grey Sys., 28, 2, 121–131, 2016.

[53] Wang K., *Applying data mining to manufacturing: the nature and implications*, Springer, 2007.

[54] Choudhary A.K., Harding J.A., Tiwari M.K., *Data mining in manufacturing: a review based on the kind of knowledge*, J. of Int. Man., pp. 500–521, 2009.

[55] Harding J.A., Shahbaz M., Srinivas S., Kusiak A., *Data Mining in Manufacturing: A Review*, J. of Man. Sc. and Eng., 2006.

[56] Wang K., Tong S., Eynard B. et al., *Review on Application of Data Mining in Product Design and Manufacturing*, Fourth Int. Conf. on Fuz. Sys. And Know. Dis., 2007.

[57] Djatna T., Muharram, A.I., *An application of association rule mining in total productive maintenance strategy: an analysis and modelling in wooden door manufacturing industry*, Proc. of Inter. Conf. on Ind. Eng. and Serv. Sc., pp. 336–343, 2015.

[58] Hu Y., Guo Z., Wen J., *Research on knowledge mining for agricultural machinery maintenance based on association rules*, Proc of Inter. Conf. on Ind. Elect. and App., pp. 901–906, 2015.

[59] Jia Z., Gou Y., Han X., *The Fault Diagnosis for Warship's Power Plant Based on Association Rules*, Adv. in Mech. and Cont. Eng. II, 433–435, 960–963, 2013.

[60] Martinez-de-Pison F.J., Sanz A., Martinez-de-Pison E. et al., *Mining association rules from time series to explain failures in a hot-dip galvanizing steel line*, Comp. and Ind. Eng., 63, 1, 22–36, 2012.

[61] Zhang L., Jiao R., *Identifying Mapping Relationships between Functions and Technologies: an Approach based on Association Rule Mining*, Proc. of Int. Conf. on Ind. Eng. and Eng. Manag., pp. 1596–1601, 2011.

[62] Yang X., Wu D., Zhou F., *Association rule mining for affective product design*, Proc. of Int. Conf. on Ind. Eng. and Eng. Manag., pp. 748–752, 2008.

[63] Shahbaz M., Srinivas V., Harding J.A. et al., *Product design and manufacturing process improvement using association rules*, Proc. of the Int. of Mech. Eng. Part B – J of Eng. Manuf., 220, 2, 243–254, 2006.

[64] Sobh A.S., Salem A.S., Darwish R. et al., *Unsupervised clustering of materials properties using hierarchical techniques*, Int. J of Coll. Ent., 5, 1–2, 74–88, 2015.

[65] Hayajneh M.T., *Fuzzy clustering modelling for surface finish prediction in fine turning process*, Mach. Sc. and Tech., 9, 3, 437–451, 2005.

[66] Jing H.L., Li C., Huang M., *A Fast Retrieval Method Based on K-means Clustering for Mechanical Product Design*, Adv. Manuf. Tech., 156–157, 98–101, 2011.

[67] Zhou X., Peng W., Shi H., *Improved K-means algorithm for manufacturing process anomaly detection and recognition*, Proc. of 1st Int. Symp. on Dig. Manuf., 1–3, 1036–1041, 2006.

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

trial environments based on a K-means clustering approach, Exp. Sys. with Appl., 38, 3, 2888–2911, 2011.

[69] Ma H.W., Mao Q.H., Zhang X.H. et al., *Defects Classification of Steel Cord Conveyor Belt Based on Rough Set and Multi-Class v-SVM*, Adv. Mat. Res., 328–330, 1814–1819, 2011.

[70] Muralidharan V., Sugumaran V., *Rough set based rule learning and fuzzy classification of wavelet features for fault diagnosis of monoblock centrifugal pump*, Measuement, 46, 9, 3057–3063, 2013.

[71] Muralidharan V., Sugumaran V., *A comparative study of Naive Bayes classifier and Bayes net classifier for fault diagnosis of monoblock centrifugal pump using wavelet analysis*, App. Soft Comp., 12, 8, 2023–2029, 2012.

[72] Jegadeeshwaran R., Sugumaran V., *Brake fault diagnosis using Clonal Selection Classification Algorithm (CSCA) – a statistical learning approach*, Eng. Sc. and Tech., 18, 1, 14–23, 2015.

[73] Moosavian A., Ahmadi H., Tabatabaeefar A. et al., *Comparison of two classifiers; K-nearest neighbor and artificial neural network, for fault diagnosis on a main engine journal-bearing*, Shock and Vib., 20, 63–272, 2013.

[74] Lesany S.A., Koochakzadeh A., Fatemi Ghomi S.M.T., *Recognition and classification of single and concurrent unnatural patterns in control charts via neural networks and fitted line of samples*, Int. J. of Prod. Res., 52, 6, 1771–1786, 2014.

[75] Brezak D., Majetić D., Udiljak T. et al., *Tool wear estimation using an analytic fuzzy classifier and support vector machines*, J. of Int. Man., 23, 3, 797–809, 2012.

[76] Yasa R., Etemad-Shahidi A., *Classification and Regression Trees Approach for Predicting Current-Induced Scour Depth Under Pipelines*, J. Off. Mech Arct. Eng., 136, 1, 2014.

[77] Perzyk M., Kochański A., Kozłowski J., Soroczyński A., Biernacki R., *Comparison of data mining tools for significance analysis of process parameters in applications to process fault diagnosis*, Inf. Sc., 259, 380–392, 2014.

[78] Lu Z.J., Xiang Q., Wu Y. et al., *Application of Support Vector Machine and Genetic Algorithm Optimization for Quality Prediction within Complex Industrial Process*, Proc. of IEEE 13th Int. Conf. on Ind. Inf., pp. 98–103, 2015.

[79] Jin R., Shi J., *Reconfigured piecewise linear* regression*tree for multistage manufacturing process control*, IIE Trans., 44, 4, 249–261, 2012.

[80] Pashazadeh H., Gheisari Y., Hamedi M., *Statistical modeling and optimization of resistance spot welding process parameters using neural networks and multi-objective genetic algorithm*, J. of Int. Man., 2014.

[81] Verbert J., Behera A.K., Lauwers B., Duflou J.R., *Multivariate Adaptive Regression Splines as a Tool to Improve the Accuracy of Parts Produced by FSPIF*, Key. Eng. Mat., 473, 841–846, 2011.

[82] Mareci D., Sutiman D., Chelariu R. et al., *Evaluation of the corrosion resistance of new ZrTi alloys by experiment and simulation with an adaptive instance-based regression model*, Corros. Sc., 73, 106–122, 2013.

[83] Perzyk M., Soroczyński A., Kozłowski J., *Application of rough sets theory in control of foundry processes*, Arch. of Metall. and Mat., 55, 3, 889–898, 2010.

[84] Jansen F.E., Kelkar M.G., *Exploratory Data Analysis of Production Data*, Proc. of Permian Basin Oil and Gas Rec. Conf., 1996.

[85] Abonyi J., *Application of Exploratory Data Analysis to Historical Process Data of Polyethylene Production*, Hung. J. of Ind. and Chem., 35, 1, 85–93, 2007.