Trendy v podnikání - Business Trends (2018), 8(3), 14-23.
https://doi.org/10.24132/jbt.2018.8.3.14_23

A USE OF DATA MINING METHODS IN THE CZECH REPUBLIC AND IN THE WORLD

Jan Brčák[1], Mikuláš Gangur[2]

[1] Ing. Jan Brčák, Západočeská univerzita v Plzni, Fakulta ekonomická, jbrcak@kem.zcu.cz
[2] doc. RNDr. Mikuláš Gangur, Ph.D., Západočeská univerzita v Plzni, Fakulta ekonomická, gangur@kem.zcu.cz

Abstract: The data mining or a knowledge discovery from data becomes more significant these days. Our world faces an enormous amount of data which are produced every day. It is important to use clever softwares to help companies sort the information and use them in a right way. Regarding the world areas it describes which methods are used and briefly describes the most important of them. Overall it contains the overview of 33 methods. Regarding the Czech Republic it is based on revisions of 42 articles which are focused on the application of data mining methods. The result from the world revision of data mining methods are decision tree, including C4.5 decision tree and classification and regression tree, genetic algorithm, k-nearest neighbor, multilayer perceptron neural network, Naïve Bayes, support vector machine, association rule, expectation maximization and k-means. For the Czech Republic mix of methods were found and they went through the whole spectrum of all areas from business, environment to healhtcare. Between main methods which are used most often are decision tree and its variations, classification and regression trees, genetic algorithm, neural network or logistic regression. The result of the comparison of all the data mining methods used in the Czech Republic and in the world is satisfactory – the same main methods are used in the Czech Republic as well as in the rest of the world. The purpose of this article is to provide the overview of commonly used methods in data mining in different areas in the Czech Republic and in the world. Another goal of this article is to provide an overview of the various areas where data mining is used with regard to the amount of resources dealing with the use of data mining methods in the Czech Republic and in the world. The conducted research made by authors shows a trend in a growth of new published articles in the Czech Republic. The research also shows that the most popular spheres where data mining methods could be used are business, environment and educational sectors.

Keywords: Data mining, methods, data collection

JEL Classification: C10, C80, C83

INTRODUCTION

Every company needs data if they want to succeed in business. Before starting with a classification of methods, it must be clarified what the data mining is itself. Data mining can also be called a knowledge discovery from data.

According to (Larose, 2014) data mining is the process of discovering useful patterns and trends in a large data set. According to (Clancy, 2016) data mining is an interdisciplinary area focusing upon methodologies for extracting useful knowledge from common data. According to (Romero, 2011) data mining is concerned with developing, researching and applying computerized methods to detect patterns in large collections of data that would otherwise be hard or impossible to analyze due to the enormous volume of data within they exist.

A mutual denominator for each definition are patterns and large data. It is essential to clarify these factors. Main seven steps for choosing right methods are mentioned in (Han, 2011) where the first step is data clean, the second step is data integration, the third step is data selection, next is data transformation, then data mining, then pattern evaluation and the seventh step is knowledge presentation. Data mining methods are widely used in companies around the world across the sectors.

It is important to realize that in today world when storing and getting new data is very cheap that we need some kind of system how to work with them and be able to recognize which of the methods could be used.
The main purpose of this article is to introduce an overview of data mining methods which are commonly used in companies, enterprises and other businesses in the Czech Republic and in the world. Another goal of this article is to provide an overview of the various areas where data mining is used with regard to the amount of resources dealing with the use of data mining methods in the Czech Republic and in the world.

1. THE DATA FLOW

Nowadays it is possible to access data from many different resources. This can be done thanks to the accessibility of data which increases rapidly due to cheaper data storages and better connectivity. The main goal is always optimization. Every company aims to optimize business performance as much as possible. That is why this development opens many new questions within companies how to arrange interconnections among application domain, data storages, technology available and analytical methodologies and of course to do this the most effectively (Horakova, 2013). Today there exist 2,7 Zettabytes of data in the digital universe as shown in the IBM Big Data Flood Infographics (Ularu, 2012). This study also shows that around 100 Terabytes are daily updated only through one social network - Facebook. There is a lot of other activity on different social networks which leads to an estimate of 35 Zettabytes of data annually generated by 2020 (Ularu, 2012). To have idea 1 Zettabyte is equal 10^{15} MB.
Another issue is to find data which are relevant and which are aplicable in a business environment. This is the reason why we need to determine methods which can help us to find data which we really need.
Nowadays, in the time of the digitalization all companies are starting to use more and more sophisticated softwares which are able to collect and evaluate extensive data. Today softwares can collect data for example about a temperature of oil in the machine, about rpm, they can remember all fails which happen in the machine and they can also collect data about what is the weather outside of the building, how long the sun shined and much more information which would be separately absolutely not usable.
A very important target is to realize that when we are able to find patterns belong data, we can create methods which can help us evaluate data correctly and the data mining methods could lead to a better setting of a production line. It is crucial to know methods which can help users to classify and evaluate all kind of data.

2. DATA MINING METHODS IN THE WORLD

The author in (Lin, 2017) classifies 33 data mining methods in 7 different applications area in his survey. There is a classification of mining methods divided by supervised and unsupervised learning methods in (Tab. 1). He also chooses top 10 data mining methods used in a business application.
According to this study the most frequent methods for data mining are: decision tree including C4.5 decision tree [C4.5] and classification and regression tree [CART], genetic algorithm [GA], k-nearest neighbor [K-NN], multilayer perceptron neural network [MP], Naïve Bayes [NB], support vector machine [SVM], association rule [AR], expectation maximization [EM] and k-means algorithm.
These methods are related especially to areas like bankruptcy prediction, customer relationship management, fraud detection, intrusion detection, recommender systems, software development, effort estimation and stock prediction/investment.

Tab. 1: A classification of mining methods

Supervised Learning Statistical Methods	Supervised Learning Intelligent Methods	Unsupervised Learning Statistical Methods	Unsupervised Learning Intelligent Methods
ARIMA	ACO	EM	AR
BN	C4.5	LA	DEA
Chi-square	CART	MCMC	K-means
LDA	CBR	PCA	SOM
LR	EA	SA	
MCMC	FL		
NB	GA		
	GRNN		
	K-NN		
	MP		
	MOEA		
	PSO		
	RBFNN		
	RS		
	SVM		
	SVR		

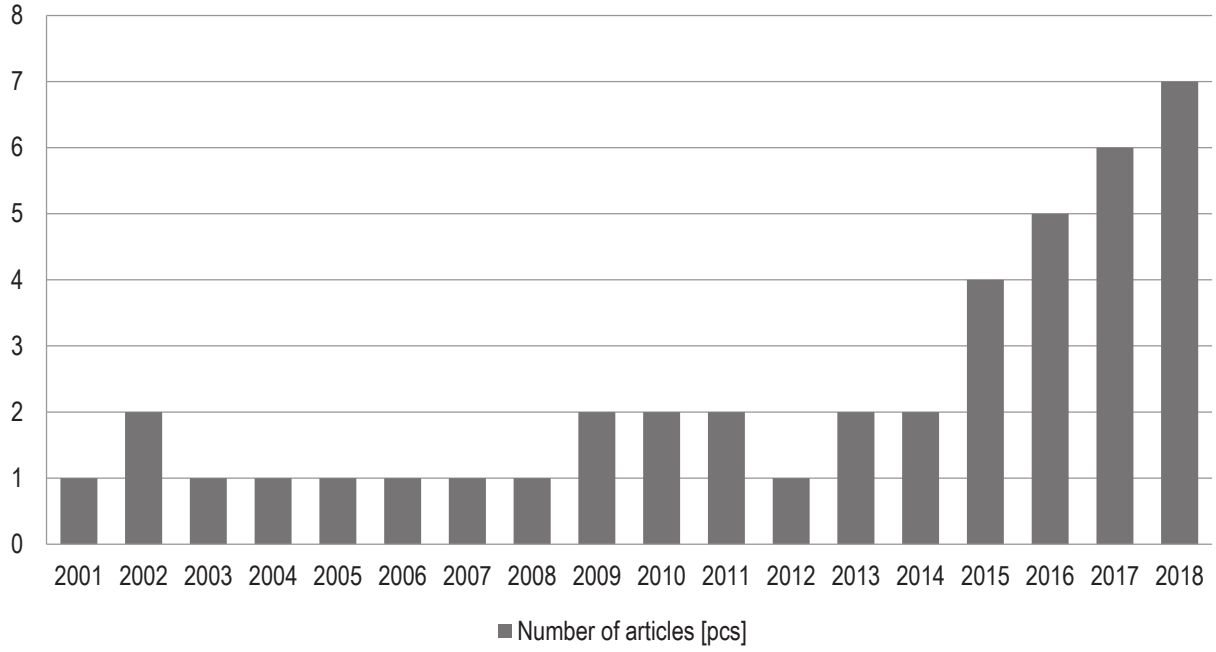
Source: Lin, 2017

According to (Han, 2011) supervised learning methods are the synonym for a classification. The result is known and it is essential to train algorithm to find the right class. For example it is possible to train the algorithm to recognize a car brand from the set of brand pictures using some specific characters.
Unsupervised learning methods are the synonym for clustering. We don‘t know the result, we want to train an algorithm to find its own solution via some specific characters. For example we can let it solve some problem which can happen and the algorithm, due to its own intelligence, solves this problem by its own.
Above in the Tab. 1: A classification of mining methods, there are used abrreviations of mining methods which are described subsequently. Bayesian network [BN], linear discriminant analysis [LDA], Markov chain Monte Carlo [MCMC], ant colony optimization [ACO], radial basis function neural network [RBFNN], case-based reasoning [CBR], data envelopment analysis [DEA], rough set [RS], fuzzy logic [FL], support vector regression [SVR], generalized regression neural network [GRNN], evolutionary algorithm [EA], multiobjective evolutionary algorithm [MOEA], particle swarm optimization [PSO], autoregressive integrated moving average model [ARIMA], link analysis [LA], principal component analysis [PCA], survival analysis [SA], self-organizing map [SOM]

3. DATA MINING METHODS IN THE CZECH REPUBLIC

This section is focused on comparison and revision methods for data mining used in the Czech Republic. 42 different scientific articles were reviewed and selected thoughts from them were explained here. They cover all the spectrum of various application from business environment like bankruptcy prediction or fraud detection up to the healthcare application.
In (Fig. 1) made by authors there is noticeable a number of published articles during period from 2001 to 2018. During last four years there is visible a significant trend of publishing higher number of articles.

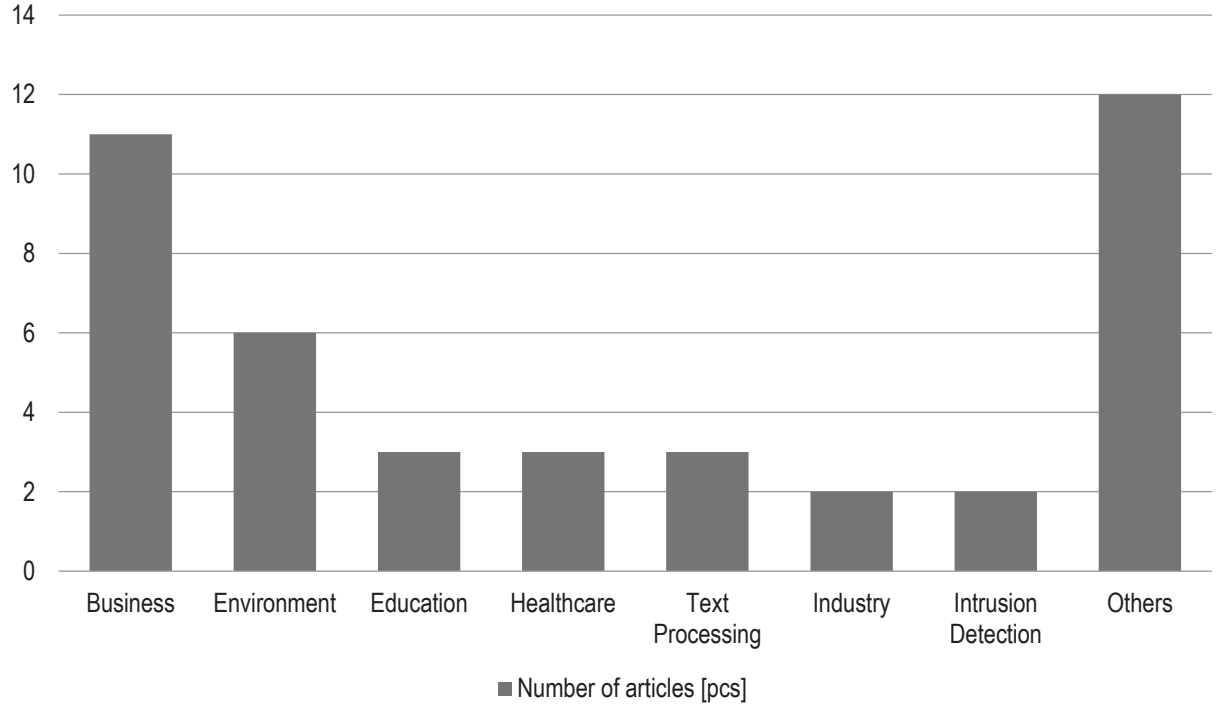
Fig. 1: The number of articles published in the Czech Republic during 2001 – 2018 according to years



Source: Own processing, 2018

In an overview in (Fig. 2) made by authors there is a perspicuous number of published articles divided by sectors. Most articles are found from the business sector, as the next the articles from the environment sector are used and these are followed by the education, healthcare and text processing sectors. The least used areas are the industry and intrusion detection. It ought to be mentioned that a high number of articles belongs to the group „others“. These articles and studies are selected mainly from the technical sector.

Fig. 2: The number of articles published in the Czech Republic during 2001 – 2018 according to areas



Source: Own processing, 2018

The data mining methods which are the most commonly mentioned in different articles follow below. They are divided by sectors.

3.1 Business

In the world and also in the Czech Republic a business segment is one of the most favourite topics and is very important. The importance of predicting bankruptcy has increased in recent years after the crisis in 2008 and after long term ongoing bull market in the world economy.
In the article from (Misankova, 2016), there is a comparison of the Visegrad four with world data mining methods for the bankruptcy sector. There were 528 methods analyzed worldwide with 30 methods of the Visegrad four (in this article the Slovak Republic was taken as a reference country). The comparison showed us that main methods used in a bankruptcy prediction are Linear discriminant analysis [LDA], multivariete discriminant analysis [MDA], logistic regression [LR], neural network [NN], data envelopment analysis [DEA] and decision tree [DT]. In Karas's (2017) article, there a bankruptcy prediction and its accuracy based on Classification and regression trees [CART] is analyzed. In the next article (Klepáč, 2017), there is predicted distress of 250 EU agriculture business companies based on support vector machine [SVM] and decision tree [DT] algorithms. In (Čamská, 2015), there is examined technical assumptions of models predicting corporate financial distress. It uses linear discriminant analysis [LDA] and logistic regression [LR]. In (Kubišta, 2018), there the effect of weather conditions on the sales of the FMCG market is described. There are more than 130 variables. For this study the methods classification and regression tree [CART] and random forrest [RF] are choosen. Authors in (Poláček, 2016) describe reconciliation as a way how to help a company deal with a debt situation. For this purpose the author uses decision tree [DT] method. The article (Mittigová, 2018) is focused on a financial sector and predicting core default predictors which cause consumer credit risk. In this article, there are examined the borrowers age, monthly income, region or how many children the borrower has. It uses seven methods: logistic regression [LR], linear discrimination analysis [LDA], quadratic discrimination analysis [QDA], classification tree [CT], random forest [RF], k-nearest neighbors [k-NN] and support vector machine [SVM]. In another article (Klepáč, 2018), there are used methods such as decision tree [DT], support vector machines [SVM], neural networks [NN] and genetic algorithms [GA] for describing the accuracy of these methods for EU companies. In (Kočenda, 2014), there is examined a prediction of default in retail credit scoring. It considers financial and socioeconomic variables. For this purpose logistic regression [LR] and classification and regression tree [CART] are used. In (Hájek, 2010), there are used methods such as support vector machine [SVM], neural networks [NN] and linear regression [LR] for the prediction budget of revenue for a municipality. Many authors try to use different methods for their investigation. Next article (Kuchař, 2017) describes methods which are based on decision trees, these methods are the following: frequent pattern outlier factor [FPOF], longer frequent pattern outlier factor [LFPOF], maximal frequent pattern outlier factor [MFPOP] and weighted closed frequent pattern outlier factor [WCFPOF]. They are used for a detection of using frequent data mining patterns in a financial sector and in a healthcare sector. As we can see there are many methods which are used in a business sector.

3.2 Environment

In the next study (Kasparova, 2008), which is focused on an observation of an air pollution in different regions, decision tree [DT] is used together with classification and regression trees [CART] and Chi-squared automatic interaction detection [CHAID]. The environment area is described in an article (Miháliková, 2014) where they try to use k-nearest method for the estimation of the missing wilting points. In the article about spread and future spread of Ambrosia artemisiifolia (Pyšek, 2012) there are used neural network [NN] and decision tree [DT]. In (Mikulecký, 2007), there is used the method decision tree [DT] for a river basin management. In (Divíšek, 2014), there is examined a landscape classification based on the distribution of different types of natural habits. Classification and regression tree [CART] is used. In (Saberioon, 2018), there is described an issue with new diets for fish where is

researched which type of a diet is better for fish. This is recorded on a consumer grade digital camera. For this purpose was chosen random forrest [RF], support vector machine [SVM], logistic regression [LR] and K-Nearest neighbours [k-NN]. In (Pořízka, 2018), there is described how to verify a characteristic feature of a grapevine cultivation with classification and regression tree [CART].

3.3 Education

Another article (Kuzilek, 2015) is focused on a prediction of a student risk based on their demographic origin. For this examination was used Naïve Bayes [NB], k-nearest neighbors [k-NN] and classification and regression tree [CART]. In an educational area (Bydžovská, 2013), there are used methods Naïve Bayesian [NB], support vector machine [SVM], logistic regression [LR], linear regression [LiR], neural networks [NN], decision tree [DT] and rule-based classifier [RBC]. The previously mentioned study describes a dedication, course recommendation and requirements for university students based on a use of data mining methods. Another example from the educational area where data mining methods are used is decision tree [DT]. It is used to predict the likelihood of a high school student enrolling for university studies (Dobesova, 2018). Data mining methods can be used almost everywhere in case of a lot of data.

3.4 Healthcare

The next big area where data mining methods can be used is a healthcare. For evaluating and recording electrical activity of the heart can be used decision tree [DT] (Lhotska, 2004). In the next article (Vévoda, 2016), there is considered a factor why Czech nurses leave their job. For this study decision tree [DT] and Chi-squared automatic interaction detection [CHAID] are used. In (Chudáček, 2006), there is used decision tree [DT] for a body surface potential mapping for the better analysing of a heart cycle.

3.5 Text processing

In (Kravalová, 2009), there is examined a recognition of named entities in Czech texts. It uses support vector machine [SVM]. The article (Lenc, 2017) is focused on an automatic multi-label document classification of Czech text documents. It uses convolutional networks [CN] and multilayer perceptron neural network [MP]. In (Bojar, 2005), there is examined a software on evaluating sentences and it uses decision tree [DT] method.

3.6 Industry

In (Drchal, 2003), there is used genetic algorithm [GA] and neural networks [NN] for testing materials in a civil engineering. Another article (Švec, 2015) is focused on a steel production and how to achieve all tasks in just in time base. For this examination k-means method is used.

3.7 Intrusion detection

In (Scherer, 2011), there is described an intrusion detection system for monitoring a network traffic. The methods k-means, support vector machine [SVM], Farthest First Traversal [FFT] and COBWEB/CLASSIT are used. In a detection of network anomalies (Nevlud, 2013), there are used the methods decision tree [DT] and Naïve Bayesian [NB] related to attacks on a network and prevention.

3.8 Others

In the telecommunication sphere, there can be used neural network [NN] and genetic algorithm [GA] as it is in the article (Kejik, 2010) about the capacity of CDMA. Next article (Šimek, 2001) describes the usage of the genetic algorithm [GA] for an application to a multilayer structure determination. In the article (Šlápek, 2017), which describes decision making strategies for auction, there are used two methods: decision tree [DT] and genetic algorithm [GA]. For a highway traffic control (Kuklová, 2017) is also used decision tree [DT]. In (Hrabec, 2015), there is described a transportation network design

problem where the object is to build and modificate a transportation network. For solving this problem was choosen genetic algorithm [GA]. The article (Franc, 2002) is about transformation support vector machine [SVM] into multi-class support vector machine [M-SVM] and it is compared with sequential minimal optimizer [SMO]. In (Matoušek, 2002), there is examined an optimization problem known as the 0/1 Knapsack problem. Methods Genetic algorithm [GA] and hill climbing algorithm [HCA] are used for this examination. In (Belohlavek, 2009), there is examined a using of a new method of decision tree based on a formal concept analysis. For this purpose are used C4.5 decision tree [C4.5], ID3 and formal concept analysis [FCA]. In (Straková, 2016), there is examined a language agnostic named entity recognition system based on neural networks [NN]. The next area is Customer Relationship Management [CRM] where decision tree [DT] and classification and regression trees [CART] are used for a modeling customer lifetime value in online retail (Jasek, 2018). For recommender systems, for a small-medium sized e-commerce portal and absent feedbacks (Peska, 2016) are used linear regression [LinReg], Lasso Regression [Lasso], AdaBoost regression [Ada LinReg], decision tree [DT] and AdaBoost classification [Ada Tree] methods.

In a summary shown in (Tab. 2), which was compiled by authors of this study, there is an overview showing how often the data mining methods are mentioned in reviewed articles. This is because the business sector is the most popular for analyzing in general. In (Fig. 2) is visible that 11 articles from this area were analyzed. Next sectors with a plenty of methods are environment and education sectors.

Tab. 2: A classification of mining methods by a sector

Sector	**Used Method**
Business	SVM, NN, LR, CART, LDA, QDA, CT, RF, K-NN, MDA, LR, DEA, DT, GA, FPOF, LFPOF, MFPOP, WCFPOF
Environment	CART, SVM, RF, LR and k-NN, DT, CART, CHAID, k-nearest, NN
Education	CART, NB, k-NN, DT, NB, SVM, LG, LR, NN, DT, RBC
Healthcare	DT, CHAID
Text Processing	CN, MP, SVM, DT
Industry	K-means, GA, NN
Intrusion Detection	SVM, K-means, FarhesFirst, COBWEB, DT, NB
Others	GA, DT, CART, NN, C4.5, FCA, ID3, HCA, SMO, SVM, M-SVM, LR, Lasso, Ada LinReg, Ada Tree

Source: Own processing, 2018

In the Tab. 2 above, mining methods divided by a sector are introduced. The method used in a business sector most often is LR, in an environment sector it is DT, in an education sector also DT as well as in a healthcare sector and the method used at the highest frequency in a sector called others is GA. Regarding sectors text processing, industry and intrusion detection, only one most commonly used method cannot be determined.

THE CONCLUSION AND FURTHER RECOMMENDATION

The task of this study is to review methods used in data mining in the Czech Republic and in the world. In many areas of business applications, data mining techniques are used but a common employee in a company does not know which of them are used. In most cases the employees know that systems work and that is all. This article should bring more light on a methods usage in the Czech Republic and in the world.

In the second part of this study, a review of world methods is made and main use of them is described. The research shows that the main methods such as decision tree, including C4.5 decision tree

and classification and regression tree, genetic algorithm, k-nearest neighbor, multilayer perceptron neural network, Naïve Bayes, support vector machine, association rule, expectation maximization and k-means algorithm belong to frequently used methods choosen for applying in a business.
In comparison to the conducted research realized by authors where 42 Czech articles were reviewed there is a possibility to recognize a tendency or a trend across the business. There are described a lot of methods which are usually used in different fields. But main methods such as decision tree and its variation, classification and regression trees, genetic algorithm, neural network and logistic regression are most common methods for data mining in the Czech Republic.
This article is not about a classification of the only possible method which should be used or which is the best one. The main contribution of this article is to clarify which methods are often used in different businesses in the Czech Republic and in the world.
In the conducted research it is shown that more methods such as k-means, k-NN, MP, etc. are also used in the Czech Republic however these methods are not utilized so often.
Each of the mentioned methods is very specific and could be suitable for some kind of a specific application. It is important to have in mind that it is not possible to generalize this whole topic and say that only one single method would be appropriate for some application because it will probably always be a mix of methods which should be used in a certain situation to achieve the best results.
The article can help students or scientists orient in the main methods and make a brief overview about applicable methods.
For the next research it is possible to continue in this analysis, go deeper and focus on a connection between data mining and business intelligence. It is recommended to find common features and classify them in more detail.
The conducted research made by authors gathers information from a relevant science source such as Google Scholarship and other relevant sources. In (Fig. 1), there is shown the trend of new published articles dealing with data mining methods and especially the trend in recent years where there is visible a rising number of published articles in the Czech Republic. In (Fig. 2), there is shown a number of articles divided by sectors. It is mentioned that a business sector is very popular and has the highest quantity of articles. Next areas are the environment, education and healthcare sector. In (Tab. 2), there are shown data mining methods used in the Czech Republic. Again it is obvious that most methods are used in the business, environment and education sector according to the biggest number of articles.

REFERENCES

Belohlavek, R., De Baets, B., Outrata, J., & Vychodil, V. (2009). Inducing decision trees via concept lattices. *International Journal of General Systems*. 38(4), 455-467.
Bojar, O., Semický, J., & Bnešová, V. (2005). VALEVAL: Testing VALLEX Consistency and Experimenting with Word-Frame Disambiguation. *The Prague Bulletin of Mathematical Linguistics.* 5-17.
Bydžovská, H. (2013).*Toward Prediction and Recommendation in Higher Education.Course Enrolment Recommender System*. Brno. Rigorózní práce. Masarykova Univerzita.
Čamská, D. (2015). Technical Assumptions of Models Predicting Corporate Financial Distress. *The 9th International Days of Statistics and Economics.* Prague, 2015, 1-9.
Chudáček, V., Huptych, M., & Lhotská, L. (2006). Feature Selection in Body Surface Potential Mapping. *IEEE: ITAB International Special Topics Conference on Information Technology in Biomedicine.* Pistacaway.
Clancy, T.R., & Gelinas, L. (2016). Knowledge discovery and data mining implications for nurse leaders. *Journal of Nursing Administration*. 46(9), 422-424.
Divíšek, J., Chytrý, M., Grulich, V., & Poláková, L. (2014). Landscape classification of the Czech Republic based on the distribution of natural habitats. *Preslia.* 86(3), 209-231.

Dobesova, Z., & Pinos, J. (2018). Using Decision Trees to Predict the Likelihood of High School Students Enrolling for University Studies. *Computational and Statistical Methods in Intelligent Systems*. 111-119.

Drchal, J., Kučerová, A., & Němeček, J. (2002). *ICECT'03 Proceedings of the third international conference on Engineering computational technology*. Scotland, 2002, 211-212.

Franc, V., & Hlaváč, V. (2002). Multi-class support vector machine. *Pattern Recognition*. 2.

Han, J., Kamber, M., & Jian J. (2011). *Data mining: concepts and techniques*. Burlington, MA: Elsevier, c2011. ISBN 978-012-3814-791.

Hájek, P., & Olej, V. (2010). Municipal revenue prediction by support vector machine ensembles. *WSEAS Transactions on Computers*. 9(11), 1255-1264.

Horakova, M., & Skalska, H. (2013). Business Intelligence and Implementation in a Small Enterprise. *Journal of Systems Integration*.2, 12.

Hrabec, D., Popela, P., Roupec, J., & Novotný, J. (2015). Hybrid Algorithm for Wait-and-See Transportation Network Design Problem with Linear Pricing. *21st International Conference on Soft Computing MENDEL*. 2015, 183-188.

Jasek, P., Vrana, P., Sperkova, L., Smutny, Z., & Kobulsky, M. (2018). Modeling and Application of Customer Lifetime Value in Online Retail. *Informatics.* 2018, 5(1).

Karas, M., & Rezankova, M. (2017). Predicting the Bankruptcy of Construction Companies: A CART-Based Model. *Engineering Economics.* 28(2), 145-154.

Kasparová, M., Krupka, J., & Jirava, P. (2008). Application of decision trees in problem of air quality modelling in the Czech Republic locality. *WSEAS Transactions on Systems*. 7(10), 1166-1175.

Kejik, P., Hanus, S., & Blumenstein, J. (2010). Comparison of Fuzzy Logic and Genetic Algorithm Based Admission Control Strategies for UMTS System. *Radioengineering*. 19(4).

Klepáč, V., & Hampel, D. (2018). PREDICTING BANKRUPTCY OF MANUFACTURING COMPANIES IN EU. *E&M Economics and Management*. 21(1).

Klepáč, V., & Hampel, D. (2017). Predicting financial distress of agriculture companies in EU. *Agricultural Economics.* 63(8), 347-355.

Kocenda, E., & Vojtek, M. (2011). Default Predictors in Retail Credit Scoring: Evidence from Czech Banking Data. *Emerging Markets Finance and Trade*. 47(6), 80-98.

Kravalová, J., & Žaokrtský, Z. (2009). Czech Named Entity Corpus and SVM-based Recognizer - SVM. *NEWS '09 Proceedings of the 2009 Named Entities Workshop: Shared Task on Transliteration*. 194-201.

Kubišta, M. (2018). Analysis of Weather Effect on Sales in the Czech FMCG Market. Praha, Bachelor Thesis. Faculty of Social Sciences.

Kuchař, J., & Svátek, V. (2017). Spotlighting Anomalies using Frequent Patterns. *Proceedings of the KDD 2017 Workshop on Anomaly Detection in Finance*. Canada.

Kuklová, J., & Přibyl, O. (2017). Changeover from decision tree approach to fuzzy logic approach within highway management. *Neural Network Work.* 1(1), 1-16.

Kuzilek, J., Hlosta, M., Herrmannova, D., Zdrhal, Z., & Wolf, A. (2015). OU Analyse: analysing at-risk students at the Open University. *Learning Analytics Review*. 15(1), 1-16.

Larose, D., & D Larose, Ch. (2014). *Discovering knowledge in data: an introduction to data mining*. Hoboken: Wiley.

Lenc, L., & Král, P. (2017). Deep Neural Networks for Czech Multi-label Document Classification. *CICLing.*

Lhotksa, L., Macek, J., & Peri, D. (2004). Evaluation of ECG: Comparison of decision tree and fuzzy rules induction.

Lin, W-Ch., Ke, S-W., & Tsai, Ch-F. (2017). Top 10 data mining techniques in business applications: a brief survey. *Browse Journals & Books.* 46(7), 1158-1170.

Matoušek, R., (2002). Hybrid Genetic Algorithms and Knapsack Problem in MATLAB Environment. *Matlab*. 315.

Miháliková, M., Matula, S., & Doležal, F. (2014). Application of k-Nearest code to the improvement of class pedotransfer functions and countrywide Field Capacity and Wilting Point maps. *Soil and Water Research.* 9(1), 1-8.
Mikulecky, P., Štekerová, K., & Ponce, D. (2007). Knowledge-based approaches for river basin management. *Hydrology and Earth System Sciences Discussions*. 4(3).
Misankova, M., & Barosova, V. (2016). Comparison of selected statistical methods for the prediction of Bankruptcy. *The 10th International Days of Statistics and Economics*. Prague.10.
Mittigová, P. (2018). *Consumer Credit Risk Analysis: Evidence from the Czech Republic*. Praha, 2018. Master Thesis. UK Fakulta sociálních věd. Vedoucí práce Evžen Kočenda.
Nevlud, P., Bureš, M.., Kapičák, L., & Zdrálek, J. (2013). Anomaly-based Network Intrusion Detection Methods. *Advances in Electrical and Electronic Engineering*. 11(6), 468 - 474.
Peska, L. (2016). Using the Context of User Feedback in Recommender Systems. *MEMICS 2016*. 1-12.
Poláček, T., Doubravský, K., & Dohnal, M. (2016). Reconciliation as a tool for decision making within decision tree related to insolvency problems. *Trends Economoc and Management*. 10(25).
Požízka, J., Diviš, P., & Dvořák, M. (2018). Elemental analysis as a tool for classification of Czech white wines with respect to grape varieties. *JOURNAL OF ELEMENTOLOGY.* 23(2), 709-727.
Pyšek, P., Chytry, M., Pergl, J., & Wild, J. (2012). Plant invasions in the Czech Republic: current state, introduction dynamics, invasive species and invaded habitats. *Preslia.* 84(3), 576-630.
Romero, C., Ventura, S., Pechenizkiy, M., & Baker, R.S.J.D. (2011). *Handbook of educational data mining. Boca Raton*.
Sarioon, M., Císař, P., Labbé, L, Souček, P, Pelissier, P., & Kerneis, T. (2018). Comparative Performance Analysis of Support Vector Machine, Random Forest, Logistic Regression and k-Nearest Neighbours in Rainbow Trout (Oncorhynchus Mykiss) Classification Using Image-Based Features. *Sensors. Basel.* 18(4).
Scherer, P., Vicher, M., Drázdilová, P., & Snasel, V. (2011). Using SVM and clustering algorithms in IDS systems. *DATESO 2011: databases, texts, specifications, and objects: proceedings of the Dateso 2011 Workshop.* Ostrava, 108-119.
Šimek, D., Rafaja, D., & Kub, J. (2001). Genetic algorithm applied to multilayer structure determination. *Materials Structures*. 8(1).
Šlápek, M., & Neruda, R. (2017). Tree based decision strategies and auctions in computational multi-agent systems. *Revista Investigación Operacional*. 38(4), 335-342.
Strakova, j., Straka, M., & Hajič, J. (2016). Neural Networks for Featureless Named Entity Recognition in Czech. *Text, Speech, and Dialogue: 19th International Conference, TSD 2016*, Brno, Czech Republic. 12-16, 173-181.
Švec, p., Frischerová, L., & David, J. (2016). Usage of clustering methods for sequence plan optimization in steel production. *Metalurgija.* 55(3), 485-488.
Ularu, E.G., Puican, F.C., Apostu, A., & Velicanu, M. (2012). Perspective on Big Data and Big Data Analytics. *Database Systems Journal*. 2(4), 12.
Vévoda, J., Vévodová, Š., Bubeníková, Š., Kisvetrová, H., & Ivanová, K. (2016). Datamining techniques – Decision tree: New view on nurses' intention to leave. *Cent Eur J Nurs Midw.* 7(4), 518-526.
Wen, S., Li, H., & Wang, Ch. (2018). Naïve Bayes regression model and its application in collaborative filtering recommendation algorithm. *International Journal of Internet Manufacturing and Services*. 5(1).