DE GRUYTER
OPEN

DOI 10.1515/ama-2016-0036

acta mechanica et automatica, vol.10 no.3 (2016)

COMPARATIVE EVALUATION OF THE DIFFERENT DATA MINING TECHNIQUES USED FOR THE MEDICAL DATABASE

Anna KASPERCZUK[*], Agnieszka DARDZIŃSKA[**]

[*]Department of Mechanics and Computer Science, Bialystok University of Technology, ul. Wiejska 45c, 15-351 Bialystok, Poland

a.kasperczuk@doktoranci.pb.edu.pl, a.dardzinska@pb.edu.pl

received 2 February 2016, revised 22 July 2016, accepted 25 July 2016

Abstract: Data mining is the upcoming research area to solve various problems. Classification and finding association are two main steps in the field of data mining. In this paper, we use three classification algorithms: J48 (an open source Java implementation of C4.5 algorithm), Multilayer Perceptron - MLP (a modification of the standard linear perceptron) and Naïve Bayes (based on Bayes rule and a set of conditional independence assumptions) of the Weka interface. These classifiers have been used to choose the best algorithm based on the conditions of the voice disorders database. To find association rules over transactional medical database first we use apriori algorithm for frequent item set mining. These two initial steps of analysis will help to create the medical knowledgebase. The ultimate goal is to build a model, which can improve the way to read and interpret the existing data in medical database and future data as well.

Key words: Data Mining, Classification, WEKA, J48, MLP, Apriori, Association Rules

1. INTRODUCTION

The past 20 years show dynamic growth in the amount of information in electronic formats. The accumulation of this data has taken place at an explosive rate and it has been estimated that the amount of information in the world doubles every two years (Dardzinska, 2013; Dardzinska and Romaniuk, 2015a). Collected data often hold valuable and interesting information.

Intensive rise of the field of knowledge discovery in databases (KDD) and data mining (DM) is a response to a sharp increase in the amount of information collected in databases and data warehouses. Data mining techniques allow us to find new, previously unknown relationships and patterns in databases that can be used later to build support decision-making information system (Dardzinska and Romaniuk, 2015b). This phenomenon is largely reflected in medicine, where the progress of information technology has contributed to the sudden increase in the amount of data. Using these technologies, we are able to bring unprecedented knowledge that can be useful in the treatment of various diseases (Yoo et al., 2012).

In this paper we present how to choose the best classifier, verify it, and then extract interesting association rules in medical database. For the purpose of this paper we use voice disorders database, which data was collected among academic staff. The occupational voice diseases are chronic diseases that are directly related to the profession and working conditions. In the case of vocal organ teacher, the diseases are the results of continuous voice strain. Increasingly, it takes into account also the psychophysical load occurring in their professional teacher as a risk factor increasing the likelihood of disease burden and vocal organ (Sliwinska-Kowalska et al., 2006). Therefore, it becomes extremely important to find such traits among patients which have the greatest impact on their recovery.

2. MATERIAL AND METHODS

Based on the survey of 240 people we built a database consisting of 240 objects and 68 attributes. Data refers to issues related to fonoaudiology, speech therapy and voice diseases. This database has been prepared in the extension .arff, which is accepted by Weka 3.6.11. There are 68 original classification attributes including age, place of residence, workplace, frequency of voice work, frequency of clinical control, surgical treatments, smoking and other features important from the point of view of fonoaudiology.

In this work (for testing and verification) we use Weka interface (Wakaito Environment for Knowledge Analysis), developed at University of Wakaito, New Zealand. It is a collection of machine learning algorithms for data mining tasks. Weka supports several standard data mining tasks, more specifically, data preprocessing, clustering, classification, regression, visualization, and feature selection. All techniques of Weka software are predicated on the assumption that the data is available as a single flat file or relation, where each data point is described by a fixed number of attributes (numeric, normally, or nominal attributes, but some other types of attributes are also supported by this software).

This software has many important advantages, so that we use it in our work:

- it is fully implemented in the Java programming language, therefore runs on almost any architecture;
- it is easy to use due to its graphical user interface;
- it is a huge collection of data preprocessing and modeling techniques.

2.1. Classification

First we focus on finding the best classification algorithm for given database. The classifier is an algorithm that implements

DOI 10.1515/ama-2016-0036

classification, especially in a concrete implementation. We use for this classification – model finding process that is used for portitioning data into different classes according to some initial assumptions. In other words, we can say that classification is the process of generalizing the data according to different instances. There are many different classifiers and many different types of dataset resulting in difficulty in knowing which will perform most effectively in any given case. It is already widely known that some classifiers perform better than others on different datasets. It is always possible that another classifier may work better. To decide which classifier will work the best for a given dataset there are two options. First is to put all the trust in an expert's opinion based on knowledge and experience. Second is to run through every possible classifier that could work on the dataset, identifying rationally the one which presents the best results (Cheng, Greiner, 2001; Dardzinska, 2013).

Classification is a data mining algorithm that creates a step-by-step guide for how to determine the output of a new data instance. It is the process of finding a set of models that differentiate data classes and concepts. We used it to predict group memberships for data instances. In first step we describe a set of predetermined classes. Each tuple is assumed to belong to a predefined class as determined by class label attribute, the set of tuples are used for model construction, called training sets. The model is represented as classification rules, decision trees or mathematical formulas. Model usage that is used for classifying future data trends and unknown objects. It estimates the accuracy of the constructed model by using certain test cases. Test sets are always independent of the training sets (Dardzinska, 2013; Frawley et al., 1991).

In Weka we have three basic steps for classification:

- preparing the data;
- choose classify and apply algorithm;
- analyze the result or output.

Tab.1. Classification attributes

Attribute	Value
Smoke	{0-never, 1-no, but I used, 2-yes}
Allergy	{no, yes}
Thyroid_disease	{no, yes}
Reflux	{no, yes}
Horm_disorders (hormonal dosorders)	{no, yes}
Reflux_treat (reflux tratment)	{no, yes}
Cons_therapy (conservative therapy)	{no, yes}
Voice_rehabilitation	{0 -never, 1-once, 2-few times}
Laring_surgery (larynx surgery)	{no, yes}
Infection_resp (upper respiratory tract infections)	{no, yes}

In the following subsections we discussed various classification algorithms, which we used in our work (Thair, 2009).

J48 is a popular machine learning algorithm based upon J.R. Quilan C4.5 algorithm. All data are of the categorical type and therefore continuous data will not be examined at this stage. The algorithm will however leave room for adaption to include this capability. The algorithm was tested against C4.5 for verification purposes (Freund, 1999; Ras and Dardzinska, 2011).

Multilayer Perception (MLP) is a network, which can be built step by step by user, created by an algorithm or both. The network can also be monitored and modified during the whole training time. The nodes in this network are all sigmoid (except for when the class is numeric, when the output nodes become threshold linear units).

Naive Bayes is a numeric estimator, where precision values are chosen based on analysis of the training set. This classifier will use a default precision of 0.1 for numeric attributes when built classifier is called with zero training instances (Bouckaert, 2004).

Based on the knowledge of voice hygiene and factors affecting the occurrence of voice diseases, we chose the class attributes of classification (Tab.1).

2.2. Association rules

Let us assume that $S = (X, A, V)$ is an information system, where (Agrawal and Srikant, 1993; Dardzinska, 2013):

- X is a nonempty, finite set of objects;
- A is a nonempty, finite set of attributes;
- V is a set of all attributes values.

Then, $a : X \rightarrow V_a$ is a function for any $a \in A$, that returns the value of the attribute of a given object. The attributes are divided into three different categories: set of stable attributes A_1 (the values of such attributes cannot be changed in time), set of flexible attributes A_2 and set od decision attributes D (in both of them the values of attributes can change), such that $A = A_1 \cup A_2 \cup D$ (Han et al., 2000; Pauk and Dardzinska, 2012).

Tab.2. Information System

Object	Stable attributes A_1	Flexible attributes A_2	
	Attribute a	Attribute b	Attribute c
x_1	$a1$	$b1$	L
x_2	$a1$	$b2$	L
x_3	$a1$	$b3$	H
x_4	$a2$	$b3$	H
x_5	$a2$	$b2$	L
x_6	$a2$	$b3$	L

Example of the information system $S = (X, A, V)$ is presented in Tab.2. The set of objects consists of six elements $X = \{x_1, x_2, x_3, x_4, x_5, x_6\}$. The set of attributes consists of two subsets A_1, A_2, where A_1 includes stable attributes $\{a\}$, and A_2 is a set with only flexible attributes $\{b, c\}$. The domain of attribute a consists of two values $\{a1, a2\}$, attribute b can reach three values $\{b1, b2, b3\}$, while the attribute c has two different values $\{L, H\}$ (Agrawal and Srikant, 1993; Dardzinska and Ras, 2003).

Information systems can be also seen as decision tables. In Tab. 3 we have decision System $S = (X, A, V \cup \{d\})$, with one stable attribute a, two flexible attributes b and c and the decision attribute d. "Place of birth" is an example of a stable attribute. "Blood pressure" or "Glucose level" of diagnosed patient is an example of a flexible attribute. "Operation", "Hospitalization", "Medical Treatement" are the examples of decision attributes values.

DOI 10.1515/ama-2016-0036

Tab.3. Decision System

Object	Atribute a	Attribute b	Attribute c	Decision d
x_1	$a1$	$b1$	L	+
x_2	$a1$	$b2$	L	+
x_3	$a1$	$b3$	H	-
x_4	$a2$	$b3$	H	+
x_5	$a2$	$b2$	L	-
x_6	$a2$	$b3$	L	+

Extracting association is one of the most important data mining tasks, which works on the principle of association rules between items that are significant in the database. Obtained results form the basis for decision-making and forecasting, which is undoubtedly a great advantage of the described method (Dardzinska and Romaniuk, 2015a; Ras et al., 2008). First, each set of items is called an itemset, if the support for the set is higher than a minimum threshold of support (Bouckaert, 2004; Ras and Joshi, 1997). Next we generate rules. To confirm the rule, for example $X \rightarrow Y$, where X and Y are itemsets, the support and the confidence of the rule are calculated in a standard way, i.e. by the support of the rule we mean the number of objects in information system S satisfying $X \cap Y$ (number of transactions that contain both X and Y) $\sup(r) = \mathrm{card}(X \cap Y)$, while the confidence is the ratio between the number of objects satisfying $X \cap Y$ and the number of objects saisfying X: $\mathrm{conf}(r) = \frac{\mathrm{card}(X \cap Y)}{\mathrm{card}X)}$ (Dardzinska and Romaniuk, 2015b; Deogun, et al., 1994; Han and Kamber, 2006). The rule with support and confidence above the minimum thresholds (given at the begining by the user) is the rule which should be added to the knowledge base (Dardzinska and Romaniuk, 2015b).

3. RESULTS AND DISCUSSION

We adapted the data prepared in the form of surveys and prepared them in the form of a database. Further, the data is saved with extension ARFF (Attribute Relation File Format) format to process in WEKA.

Then we start with the Weka tool use the explorer application and select the preprocess button followed by this open the result analysis data set. After that we can choose filter, which can be used to transform the data from one format to other e.g. numeric attributes into discrete ones. It is also possible to delete instances and attributes according to specific criteria on the preprocess screen.

3.1. Mining classification rules

To find the best classifier we should pay attention to the following parameters we receive in output (Bouckaert, 2004; Han et al., 2000):

- TP Rate - rate of true positives (instances correctly classified as a given class);
- FP Rate - rate of false positives (instances falsely classified as a given class);
- Precision - proportion of instances that are truly of a class divided by the total instances classified as that clases;
- Recall - proportion of instances classified as a given class divided by the actual total in that class (equivalent to TP rate);
- F-Measure - general indicator of quality of the model;
- ROC Curve (ROC Area) - a graphical plot that illustrates the performance of a binary classifier system as its discrimination threshold is varied. The accuracy of the test depends on how well the test separates the group being tested into those with and without the disease in question. Accuracy is measured by the area under the ROC curve;
- Kappa Statistic - it is a measure of conformity between the proposed allocation instance of the class and the actual, which is about the overall accuracy of the model;
- Number of correctly classified instances.

As part of the development of data we compared the individual parameters for each classifier. The results are presented in Tab. 4, Tab. 3, Fig. 1 and Fig. 2.

Tab.4. WEKA results for Recall, F-Measure, Precison

	Recall			F-Measure			Precision		
	J48	NaiveBayes	MLP	J48	NaiveBayes	MLP	J48	NaiveBayes	MLP
Smoke	0.987	0.992	1	0.987	0.991	1	0.988	0.992	1
Allergy	0.983	0.936	1	0.983	0.937	1	0.983	0.939	1
Thyroid_disease	0.992	0.949	1	0.992	0.97	1	0.992	0.97	1
Reflux	0.966	0.97	1	0.963	0.951	1	0.967	0.954	1
HORM_DISORDES	0.987	1	1	0.987	1	1	0.987	1	1
REFLUX_TREAT	1	1	1	1	1	1	1	1	1
CONS_THERAPY	0.966	0.905	1	0.965	0.906	1	0.965	0.908	1
INFECTION_RESP	0.97	0.894	1	0.97	0.896	1	0.971	0.901	1
VOICE_REHABILITATION	0.992	0.907	1	0.991	0.912	1	0.992	0.935	1
THYROID_SURGERY	0.979	0.967	1	0.977	0.97	1	0.98	0.978	1
LARING_SURGERY	0.996	0.983	1	0.996	0.984	1	0.996	0.986	1

DOI 10.1515/ama-2016-0036

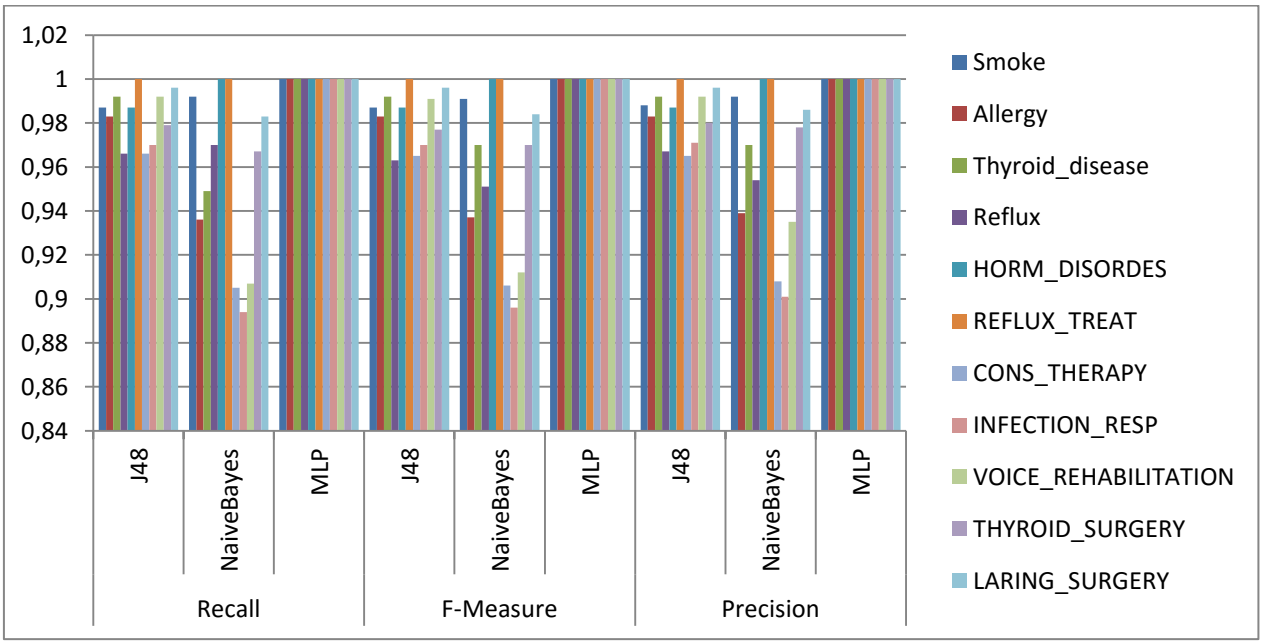

Fig.1. WEKA results for Recall, F-Measure, Precision

If we compare the Precision parameter, it can be noticed that in most cases the values are also the highest for the algorithm MLP (Fig. 1). The highest value (close to 1) indicate a good classifier. However, it should be noted that for some attributes, such us: Smoke, Allergy, Reflux, we have similar value for all examined classifiers (Tab. 4).

The chart for Recall (Fig. 1) shows that the highest values, close to 1, we calculated for Multilayer Perception algorithm. For every choosen attribute this value is equal to 1 (Tab. 4). In this case we calculated, that the worse algorithm is Naïve Bayes, because of low values of the proportion of instances classified.

The same situation we have for value of F-Measure. For every classification attribute the value is close to 1 (Tab. 4). It proves the high quality of the generated model.

When we describe the quality of the generated model of a classification, it is important to turn attention into two parameters: Kappa Statistic and number of correctly classified instances. For J48 algorithm we received the average score 0.92, which is a satisfying result. The Multilayer Perception algorithm give us the value of Kappa Statistic equal to 1. It indicates very high quality. For Naïve Bayes we got the lowest values among the considered classifiers. For properly classified instances distribution is similar. The value of Kappa Statistic for MLP is equal to 1 and the number of correctly classified instances was 100%.

We got high dispersion of results for the ROC Area (Fig. 2). We note that values for J48 and MLP are similar and ranges from 0.883 to 1 (Tab. 5). The largest value equal to 1 has occurred for Multilayer Perception, which is why it can be expected as the best classifier.

Tab.5. WEKA results for TP Rate, FP Rate, ROC Area

	TP Rate			FP Rate			ROC Area		
	J48	**NaiveBayes**	**MLP**	**J48**	**NaiveBayes**	**MLP**	**J48**	**NaiveBayes**	**MLP**
Smoke	0.987	0.992	1	0.031	0.021	0	0.996	0.994	1
Allergy	0.983	0.936	1	0.046	0.073	0	0.987	0.97	0.974
Thyroid_disease	0.992	0.97	1	0.047	0.119	0	0.952	0.925	0.965
Reflux	0.966	0.949	1	0.299	0.154	0	0.863	0.993	0.989
HORM_DISORDES	0.987	1	1	0.137	0	0	0.882	0.953	0.916
REFLUX_TREAT	1	1	1	0	0	0	0.96	0.981	0.876
CONS_THERAPY	0.966	0.905	1	0.054	0.104	0	0.948	0.953	1
INFECTION_RESP	0.97	0.894	1	0.105	0.154	0	0.984	0.94	0.995
VOICE_REHABILITATION	0.992	0.907	1	0.027	0.015	0	0.996	0.983	1
THYROID_SURGERY	0.979	0.967	1	0.292	0.002	0	0.932	0.996	1
LARING_SURGERY	0.996	0.983	1	0.046	0.002	0	1	1	1

DOI 10.1515/ama-2016-0036

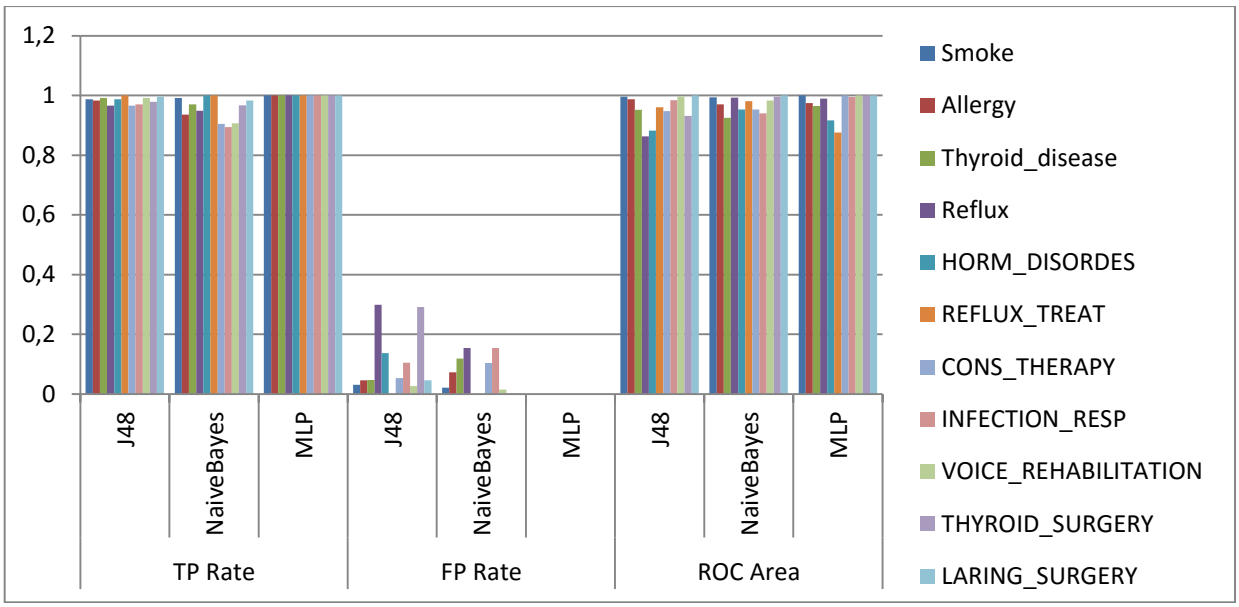

Fig.2. WEKA results for TP Rate, FP Rate, ROC Area

It can be seen that the value of TP Rate is the best for Multiplayer Perception (Fig. 2). For every attribute this value is close to 1, which is a very good result. Not so impressive results we received for Naïve Bayes classifier (Tab. 5).

For the best classification algorithm value of FP Rate is close to 0. In this case the best results we calculated also for Multilayer Perception (Fig. 2). For the majority of selected attributes classifying this value amounted to 0, which indicates a very good operating algorithm.

The results indicate, that the MLP is the best performing classifier for considered voice disorders database.

3.2. Association rules

Based on the survey of 60 people we built a database consisting of 60 objects and 68 attributes. The data refer to issues related to fonoaudiology. speech therapy and voice diseases. This database has been prepared in the extension .arff. which is accepted by the installation program. Weka. We use the Apriori Algorithm and we want to find the association rules in our database.

For the purposes of analysis the following values:

- minimum support: 0.9 (54 instances);
- minimum metric <confidence>: 0.9.

are taking into consideration.

During the analysis several interesting association rules were obtained. Some ot them are given below:

1. REFLUX_TREAT=no, 57==>ASTHMA=no, 57 conf:(1)
2. VOICE_PER=2,REFLUX_TREAT=no,55==>ASTHMA=no,55 conf:(1)
3. HORM_DISORDES=no,54==>ASTHMA=no,54 conf:(1)
4. EDU=2,REFLUX_TREAT=no,54==>ASTHMA=no,54 conf:(1)
5. EDU=2,57==>ASTHMA=no,56 conf:(0.98)
6. VOICE_PER=2,57==>ASTHMA=no,56 conf:(0.98)
7. THYROID_SURGERY=no,56==>ASTHMA=no,55 conf:(0.98)
8. VOICE_PER=2,ASTHMA=no,56 ==>REFLUX_TREAT=no,55 conf:(0.98)
9. GENDER=K,55==>ASTHMA=no,54 conf:(0.98)
10. LARYNG_SURGERY=no,55==>ASTHMA=no,54 conf:(0.98)

These are the best rules extracted from given data. All of them have the support and the confidence above given minimal thresholds. The attributes forming these rules are described below. Others, with the support below the minimal value (given by the user and consulted with the expert) are passed over.

@attribute	'REFLUX_TREAT'	{'no', 'yes'}
@attribute	'ASTHMA'	{'no', 'yes'}
@attribute	'VOICE_PER'	{'0','1','2'}
	0-(0-2y), 1-(2-10 y), 2-(> 10 y)	
@attribute	'HORM_DISORDES'	{'no', 'yes'}
@attribute	'EDU'	{'0','1','2'}
	0-primary, 1-secondary, 2-higher	
@attribute	'THYROID_SURGERY'	{'no', 'yes'}
@attribute	'GENDER'	{'K', 'M'}
@attribute	'LARYNG_SURGERY'	{'no', 'yes'}
@attribute	'EDU'	{'0','1','2'}
	0-primary, 1-secondary, 2-higher	
@attribute	'THYROID_SURGERY'	{'no', 'yes'}
@attribute	'GENDER'	{'K', 'M'}
@attribute	'LARYNG_SURGERY'	{'no', 'yes'}

4. CONCLUSION

In this paper we propose method to find the best classifier and association rules in voice disorders database using WEKA methods. The voice disorders database was collected among academic professionals. The occupational diseases are chronic diseases that are directly related to the profession and working conditions. In the case of vocal organ teacher, these diseases are the result of continuous voice strain. It becomes important to find such traits among patients have the greatest impact on their recovery. The obtained results are interesting, however we will wish on finding new algorithm which will be more useful in people with voice disorders treatment.

DOI 10.1515/ama-2016-0036

REFERENCES

1. **Agrawal R., Srikant R.** (1993), Fast algorithm for mining assocation rules, *International Conference on Very Large Databases*, 487-499.
2. **Bouckaert R** (2004), Naive Bayes Classifiers That Perform Well with Continuous Variables*, Lecture Notes in Computer Science Volume*, 3339, 1089-1094.
3. **Cheng J, Greiner R** (2001), Learning Bayesian Belief Network Classifiers, *Algorithms and System In Stroulia & Matwin LNAI 2056*, 141-151.
4. **Dardzinska A, Romaniuk A** (2015a), Incomplete distributed information systems optimization based on queries*, Advances in Swarm and Computational Intelligence*, Volume 9142 of LNCS Springer, 265-274.
5. **Dardzinska A.** (2013), *Action Rules Mining*. Springer, pp.90.
6. **Dardzinska A., Ras Z.** (2003), On Rules Discovery from Incomplete Information Systems, *Proceedings of ICDM'03 Workshop on Foundations and New Directions of Data Mining*, Melbourne, Florida, IEEE Computer Society.
7. **Dardzinska A., Romaniuk A.** (2015b) Queries for detailed information system selection, *Position Papers of the 2015 Federated Conference on Computer Science and Information Systems, Annals of Computer Science and Information Systems vol. 6*, Computer Science and Information Systems: FedCSIS, 11-15.
8. **Deogun J., Raghavan V., Sever H.** (1994), Rough set based classification methods and extended decision tables*, International Workshop on Rough Sets and Soft Computing*, 302-309.
9. **Frawley W., Piatetsky-Shapiro G., Matheus C.** (1991), Knowledge discovery in databases, *An overview. Knowledge Discovery in Databases*, 1–27.
10. **Freund Y, Mason L.** (1999), The alternating decision tree algorithm, *In Proceedings of the 16th International Conference on Machine Learning,* 124-133.
11. **Han J., Kamber M.** (2006), Data Mining: Concepts and Techniques, *Morgan Kaufmann Publishers*, Second Edition, 21-27.
12. **Han J., Pei J., Yin Y.** (2000), Mining frequent patterns without candidate generation, *ACM SIGMOD International Conference on Management of Data*, 1–12.
13. **Pauk J., Dardzinska A.** (2012), New method for finding rules in incomplete information systems controlled by reducts in flat feet treatment, *Image Proc. and Communications Challenges. Advances in Intelligent and Soft Computing*, 184, 209-214.
14. **Ras Z., Dardzinska A.** (2011), From Data to Classification Rules and Action*,. International Journal of Intelligent Systems*, Wiley, 26(6), 572-590.
15. **Ras Z., Dardzinska. A.,Tsay. L., Wasyluk H.** (2008), Association Action Rules, *IEEE Interna-tional Conference on Data Mining Workshops*, 283-290.
16. **Ras Z., Joshi S.** (1997), Query approximate answering system for an incomplete DKBS*, Fundamenta Informaticae Journal*, 20(3/4), 313-324.
17. **Sliwinska-Kowalska M., Niebudek-Bogusz E., Fiszer M., et al**. (2006), The prevaence and risk factors for occupational voice disorders in teachers, *Folia Phoniatr. Logop*, 58(2), 85-101.
18. **Thair Nu Phyu** (2009), Survey of Classification Techniques in Data Mining, *Proceedings of the International MultiConference of Engineers and Computer Scientists*, Vol I IMECS.
19. **Yoo I, Alafaireet P, Marinov M, et al.** (2012), Data mining in healthcare and biomedicine, *A survey of the literature. Journal of medical systems*, 36(4), 2431-2448.

Research was performed as a part of projects MB/WM/8/2016 and financed with use of funds for science of MNiSW