

Feature Selection for Classification Incorporating Less Meaningful Attributes in Medical Diagnostics

Agnieszka Wosiak

Lodz University of Technology
Institute of Information Technology
ul. Wólczajska 215, 90-924 Łódź, Poland
Email: agnieszka.wosiak@p.lodz.pl

Danuta Zakrzewska

Lodz University of Technology
Institute of Information Technology
ul. Wólczajska 215, 90-924 Łódź, Poland
Email: danuta.zakrzewska@p.lodz.pl

Abstract—In medical diagnostics there is a constant need of searching for new methods of attribute acquiring, but it is difficult to assess if these new features can support the existing ones and can be useful in medical inference. In the paper the methodology of discovering features which are less informative while considering independently, however meaningful for diagnosis making, is investigated. The proposed methodology can contribute to better use of attributes, which have not been considered in the diagnostics process so far. The experimental study, which concerns arterial hypertension as one of the civilization diseases demanding early detection and improved treatment is presented. The experiments confirmed that additional attributes enable obtaining the diagnostic results comparable to the ones received by using the most obvious features.

I. INTRODUCTION

IN MEDICAL research the process of diagnosis is usually provided by experts with the necessary knowledge. Towards facilitating this task some automatic actions may be performed, such as feature selection for choosing the set of attributes appropriate for particular diagnostics problem. In most instances the results of the automated selection cover or even improve the expert judgment [1], [2], [3].

The paper deals with cases where attributes regarded as the most informative should be excluded to discover new dependencies and, as a consequence, new therapies. We consider the methodology, which aims at indicating these features from among less meaningful for medical classification, that can be used in automated diagnosis of the disease. The proposed method includes result evaluation by using clustering taking into account independently features indicated by the method and the recognized ones. The obtained clusters are compared to check if the presented methodology can contribute to better use of attributes acquired from new diagnostics process. Presented approach will be investigated for arterial hypertension, which is acknowledged as one of the civilization diseases demanding early detection and improved treatment [4]. The proposed methodology is evaluated by experiments carried out on real data.

The rest of the paper is organized as follows. In Section II relevant work is presented. Next we describe the proposed methodology and discuss classification and clustering techniques, which are expected to be the most appropriate for the considered case. Section IV is dedicated to the experiments

conducted on real data. Finally, in Section V, the results and some concluding remarks are discussed.

II. RELATED WORK

The paper addresses an issue of feature selection methods for medical diagnosis supporting. The new contribution of this work is the exclusion of the most informative features to find out additional dependencies among the attributes derived from a modern process of data acquiring. Such an approach was not considered in the literature so far, however the feature selection analysis was the subject of interests of many researchers.

A survey on feature selection methods was presented in [5]. The main objective was to provide a generic introduction to variable elimination which can be applied to a wide range of machine learning problems. The authors described filter, wrapper and embedded methods. Moreover they applied some of the feature selection techniques on standard data sets to demonstrate the applicability of the proposed methods. As the conclusion they pointed out that comparison between feature selection algorithms can only be done on the same data set since each underlying algorithm behaves differently depending on data characteristic.

The problem of factors that are considered as less important for disease diagnosis but still, according to medical literature, deserve to be included in diagnostics process, was mentioned in [6]. However the main goal of that paper was to compare classifiers for the detection of heart disease. The paper presented effects of using automated feature selection and a medical knowledge based on Motivated Feature Selection (MFS) process. MFS combined with the Computerized Feature Selection (CFS) process was analyzed and good performance was observed for Naive Bayes, k-nearest-neighbors and SMO classifiers.

In [7] the research was based on data of clinical diagnosis, symptoms and medical intervention classification for the patients after surgical intervention with recurrent pelvic cyst. The decision tree was used to find the meaningful characteristic as well as classification rules. The experiment results were to help the clinical faculty doctors in effective diagnosing and providing treatment reference for future patients.

The authors of [8] presented a study of a diverse set of machine learning algorithms on a large number of biomedical

datasets. They concluded that the nature of a given dataset plays an important role on the classification accuracy of algorithms. Therefore it is necessary to choose an appropriate algorithm for a particular data set. However they identified some general rules for machine learning technique selection: using resampling based classifier enhancement techniques (bagging and boosting) over individual classifiers, using boosting on stable algorithms like SMO, JRip, and J48 and recommended using bagging MLP for classification if the nature of a biomedical data set is unknown.

In [9] the efficiency of the classification methods including SVM, RBF Neural Nets, MLP Neural Nets, Bayesian, Decision Tree and Random Forrest methods were compared. Some of the common clustering techniques including K-means, DBC, and EM algorithms were applied to the datasets and the efficiency of these methods has been analyzed. In each case these methods were applied to eight different binary (two class) microarray datasets. As a conclusion the authors stated that the classification success depends on the choice of feature selection methods, the number of attributes and the number of cases (samples). Results revealed the importance of attribute selection in accurately classifying new samples and the importance of integration of the feature selection and classification algorithms.

The advantages of using and perspectives of applications for AdaBoost algorithm were discussed in [10]. The authors stated that the main significance of AdaBoost concerns providing new ideas to theoretical study as well as practical problems. Moreover while most of the machine learning algorithms tend to seek complicated classifiers to improve the accuracy, AdaBoost takes the approach to obtain an accurate classifier by combining simple and weak classifiers whose accuracies are slightly better than random guessing. Besides AdaBoost does not need any parameters except the number of iterations and therefore the authors suggested that it can be used in many practical applications.

The overview of recent research done to analyze the different machine learning schemes on various medical domains leads to the conclusion that experiments are usually carried out by using the limited choice of algorithms from the machine learning repository. Then the technique which gives relatively the best results for the considered domain is selected. There is no guidelines which indicates the best classifier for a particular type of data.

III. MATERIALS AND METHODS

The proposed methodology of indicating less meaningful features to use in diagnostics process consists of four steps:

- data preparation, which results in the initial dataset,
- classification process, which enables the selection of the set of attributes crucial for the automated diagnosis,
- clustering based on the attributes derived from the previous step,
- verification process by using expert feature selection.

The research will focus on arterial hypertension case study.

A. Data Description

Arterial hypertension is a significant problem in pediatric practice. Therefore, finding effective methods which support early diagnosis of hypertension is a crucial issue for the researchers to solve. The assessment of arterial hypertension include physical examination, manual and twenty-four hour blood pressure measurements and medical imaging derived from two- and three-dimensional echocardiography.

The tissue Doppler echocardiography (TDE) allows exact evaluation of a number of additional parameters that indicate myocardial functions. Many studies confirmed that regional analysis comes from the method of tissue Doppler imaging as a sensitive way to detect clinically silent changes when standard echocardiographic parameters are still within the normal range [11]. This method is mainly used for research carried out for adult population, however Zamojska et al. considered assigning this approach also for children [12].

The initial cardiac data can be characterized by over 50 attributes. All patients undergo physical examination, manual arterial blood pressure measurements (RRmanSBP, RRmanDBP), ambulatory blood pressure monitoring (ABPM-S, ABPM-D), echocardiographic examination to evaluate cardiac function using standard parameters (ejection fraction - EF, shortening fraction - SF and myocardial performance index - MPI) and tissue Doppler examination (systolic mitral annular velocity profile and regional function parameters: velocity, strain, strain rate).

The aim of medical analysis is to evaluate the characteristics of the variables in the data sets of healthy and diagnosed children and to discover the relationships between all the parameters. The process of diagnosis performed by medical expert is mainly based on the blood pressure measurements (either manual or ambulatory monitored). The rest of the attributes are usually supportive for medical staff as each of them separately cannot indicate the disease and multivariate analysis is difficult to perform without any computer support.

B. Classification Task

For the classification purpose we will consider two approaches: decision trees and adaptive boosting. Decision trees represent one of the main techniques for discriminant analysis in data mining and knowledge discovery [13], [14]. They predict the class membership (dependent variable) of an instance using its measurements of predictor variables. They provide higher classification accuracy and offer an easy way to understand graphic representation of gathered knowledge [15]. Moreover decision trees are easy to understand and analyze, as they reflect a hierarchical way of human decision making. Therefore they are the opposite of the 'black-box' approaches where model parameters are not understandable [16] and can be easily understood by human experts [17].

In the paper we have chosen C4.5 for a decision tree algorithm as one of the most popular. Namely J48 algorithm, which is the open source Java implementation of the C4.5 in the Waikato Environment for Knowledge Analysis (WEKA) data mining tool [18] has been chosen.

1) *J48 Algorithm*: The J48 algorithm is the WEKA implementation of the C4.5 top-down decision tree learner proposed by Quinlan [19]. The algorithm uses the greedy technique. It deals with numeric attributes by determining where thresholds for decision splits should be placed. J48 algorithm employs an automatic procedure capable to select relevant features from the training data. It is able to cut the poor or non-meaningful branches into an efficient pruning process as well as able to handle both continuous and discrete attributes. In handling continuous attributes, J48 creates a threshold and then splits the list into those attributes, which values are above the threshold and the ones, which are less than or at least equal to the threshold value. It enables handling training data with missing attribute values by employing gain and entropy calculations. Therefore the J48 algorithm may cut the poor and non-meaningful branches into an efficient pruning process [20].

2) *AdaBoost Algorithm*: The possibility of boosting the prediction quality of a weak learner was firstly introduced by Freund and Schapire [21]. The adaptive boosting algorithm (AdaBoost) solved many practical shortcomings of earlier algorithms [21]. The AdaBoost is a machine learning algorithm which feeds an input training set to a weak learner algorithm repeatedly. During these repeated calls, the algorithm maintains and updates a set of weights, which indicate how difficult it is for the weak learner to identify a particular element of the training data set. Initially, all weights are equal. However, after each call, the weights are updated, in the way, which guarantee that the weights of misclassified training set elements grow. This forces the weak learner to concentrate on the difficult elements of the training set. In the study, we use a decision stump as a weak learner algorithm for the AdaBoost classifier. This model is composed of a single-level decision tree (DT), which uses one of the input parameters [22].

C. Clustering

Cluster analysis algorithms group objects taking into account a certain similarity metric. They divide the objects into a predetermined number of groups in a manner that maximizes a similarity function. During investigations of the proposed methodology, two different approaches, commonly used in medical studies ([9]) will be considered: the Expectation Maximization (EM) probabilistic approach and deterministic k-means algorithm.

1) *k-means Algorithm*: The k-means algorithm divides a data set into k clusters, where k is a user-defined value. The algorithm starts with k random clusters, and then move objects between those clusters to minimize variability within clusters and maximize variability between clusters. In other words, the similarity rules apply maximally to the members of one cluster and minimally to members belonging to the rest of the clusters. Usually, the means for each cluster on each dimension are calculated for assigning objects into the closest ones [23]. In most of the cases Euclidean metric is considered as the distance function for k-means algorithm [24], [25].

2) *EM Algorithm*: An expectation-maximization (EM) algorithm finds maximum likelihood estimates of parameters in probabilistic models. EM performs repeatedly between an expectation (E) and maximization (M) steps. Within the E step an expectation of the likelihood of the observed variables is computed and then the M step computes the maximum expected likelihood found on the E step. EM assigns a probability distribution to each instance which indicates the probability of its belonging to each of the clusters [25]. By cross validation, EM can decide how many clusters to create. The goal of EM clustering is to estimate the means and standard deviations for each cluster so as to maximize the likelihood of the observed data. K-means assigns observations to clusters to maximize the distances between clusters. The EM algorithm computes classification probabilities, not actual assignments of observations to clusters.

D. Verification of results

In order to confirm the correctness of the obtained results, clusters based on the most meaningful attributes selected by classification algorithm are built. They are compared with groups created by clustering using attributes indicated by experts. If the groups, which are built taking into account two different sets of attributes, are of similar characteristics, then the attributes indicated by classification can be effectively used in diagnostics process.

Methodology verification consists of the following steps:

- classification using all the available features including most informative ones, which results in the feature subset selection,
- clustering based on the attributes derived from the previous step,
- comparison of clusters obtained after exclusion of most informative features with the clusters from the previous step.

IV. EXPERIMENTAL ANALYSIS AND RESULTS

The main objectives of the experiments were to prove, that by performing clustering based on particular set of less meaningful features acquired in automated classification, we can obtain the output results close to data sets acquired by using most important attributes derived from the process of feature selection and pointed out by medical experts. The presented methodology was evaluated on the real data, which were gathered for early diagnosis of arterial hypertension in children. The data set was described earlier in the section III-A.

During experiments 2 initial data sets were considered: the first one (A - Study group), consisted of data of 30 children diagnosed with primary arterial hypertension, without being overweight or obese, hospitalized in the University Hospital No 4, Department of Cardiology and Rheumatology, Medical University of Lodz. The second set (B - Control group) consisted of 30 data of children with normal blood pressure. The decision process of this initial judgment (the value of

the dependent variable for our experiments) was performed by medical experts.

A. Data Preprocessing

As the first step all the cases were put together to form one data set consisted of 60 children. We decided to exclude from the process of automatic classification these attributes that are in the straight relation to the expert judgment: manual arterial blood pressure measurements (RRmanSBP, RRmanDBP) and ambulatory blood pressure monitoring (ABPM-S, ABPM-D). Moreover we removed fundus_oculi as the feature that is usually correlated to arterial hypertension but can not determine this disease. As a result we took into consideration 42 attributes listed in table I, where the first column contains names of all selected parameters, the second one describes these parameters and the third column gives the domain definitions.

B. Classification

According to the methodology described in section III we used two classification methods: decision trees and adaptive boosting.

1) *J48 Results.*: The J48 algorithm has chosen for classification 9 attributes listed in table II out of all the attributes (table I).

As a result we obtained 58 correctly classified instances (96.67%) and 2 incorrectly (3.33%) which made the precision and recall equal to 0.967, the same for both classes.

2) *AdaBoost - Results.*: The AdaBoost algorithm choose for classification 6 attributes listed in table III out of all the parameters (table I).

Despite the fact that this method has chosen the set of attributes different from the J48 algorithm, the results were satisfactory enough. We obtained 49 instances correctly classified (81.67%) and 11 incorrectly (18.33%). The weighted average of precision was equal to 0.831 (0.757 for the 1st class and 0.913 for the 2nd class) and the weighted average of recall was equal to 0.817 (0.933 for the 1st class and 0.7 for the 2nd class).

C. Clustering

We performed clustering taking into account the sets of attributes selected by classification algorithms in the previous step of analysis (section IV-B).

1) *EM Algorithm with J48 Subset of Attributes*: Performing EM algorithm we firstly used the same subset of attributes as it was chosen by J48 algorithm. We obtained 2 clusters automatically by using cross-validation [25]. The first cluster consisted of 21 instances: 19 instances from the set A and 2 cases from the set B. The second cluster included 39 instances: 11 from the set A and 28 from the set B.

2) *K-means Algorithm with J48 Subset of Attributes*: While testing k-means technique with the same subset of attributes as it was chosen by J48 algorithm the number of 2 clusters was indicated. As a result the first cluster consisted of 35 instances. It contains 8 instances from the set A and 27 ones from the

TABLE I
THE LIST OF PARAMETERS TAKEN FOR ARTERIAL HYPERTENSION CLASSIFICATION

Parameter name	Parameter description	Domain
Group type	Dependent attribute	Integer
Body mass	Body mass	Real
BMI	Body mass index	Real
BSA	Body surface area	Real
Phys act	Physical activity	Integer
Family hist	Family history risk factor	Integer
EF	Ejection fraction	Integer
SF	Shortening fraction	Integer
IVSs	Interventricular septum-systole	Real
IVSd	Interventricular septum-diastole	Real
PWDs	Posterior wall thickness in systole	Real
PWDd	Posterior wall thickness in diastole	Real
LVDs	Left ventricular systolic diameter	Real
LVDd	Left ventricular diastolic diameter	Real
S long	Longitudinal strain	Real
MPI	Myocardial performance index	Real
LVMPI	Left ventricular myocardial performance index	Real
Sm [cm/s]	Systolic mitral annular velocity at the intraventricular septum level	Real
Sml [cm/s]	Systolic mitral annular velocity profile at the lateral level	Real
LVM Sim	Left ventricular mass by de Simone	Real
LVM Dev	Left ventricular mass by Devereux	Real
V long	Systolic longitudinal regional velocity	Integer
V circ	Systolic circumferential regional velocity	Integer
V rad	Systolic radial regional velocity	Integer
S long	Longitudinal strain	Integer
Time to peek 1	Time to peek for longitudinal strain	Integer
S circ	Circumferential strain	Integer
Time to peek 2	Time to peek for circumferential strain	Integer
S rad	Radial strain	Integer
Time to peek 3	Time to peek for radial strain	Integer
SRI long	Longitudinal strain rate	Integer
SRI rad	Radial strain rate	Integer
SRI circ	Circumferential strain rate	Integer
V long basal	Longitudinal regional systolic velocity - basal segments	Integer
V long mid	Longitudinal regional systolic velocity - middle segments	Integer
V long apex	Longitudinal regional systolic velocity - apical segments	Integer
S long basal	Longitudinal strain - basal segments	Integer
S long mid	Longitudinal strain - middle segments	Integer
S long apex	Longitudinal strain - apical segments	Integer
SRI long basal	Longitudinal strain rate - basal segments	Integer
SRI long mid	Longitudinal strain rate - middle segments	Integer
SRI long apex	Longitudinal strain rate - apical segments	Integer

TABLE II
THE LIST OF PARAMETERS CHOSEN BY J48 ALGORITHM.

Parameter name	Parameter description
Body mass	Body mass
BMI	Body mass index
EF	Ejection fraction
IVSs	Interventricular septum-systole
PWDs	Posterior wall thickness in systole
PWDd	Posterior wall thickness in diastole
Sml	Systolic mitral annular velocity profile at the lateral level
Sm	Systolic mitral annular velocity at the intraventricular septum level
S long mid	Longitudinal strain - middle segments

TABLE III
THE LIST OF PARAMETERS CHOSEN BY ADABOOST ALGORITHM.

Parameter name	Parameter description
EF	Ejection fraction
SF	Shortening fraction
IVSs	Interventricular septum-systole
PWDd	Posterior wall thickness in diastole
Family hist	Family history risk factor
Time to peek 2	Time to peek for circumferential strain

set B. The second cluster included 25 instances: 22 from the set A and 3 from the set B.

3) *EM Algorithm with AdaBoost Subset of Attributes*: In the third test we executed EM algorithm with the same subset of attributes as it was chosen by AdaBoost. We also obtained 2 clusters automatically by using cross-validation [25]. The first cluster consisted of 21 instances. It was built up of 18 instances from the set A and 3 cases from the set B. The second cluster included 39 instances: 12 from the set A and 27 from the set B.

4) *K-means Algorithm with AdaBoost Subset of Attributes*: The last run was performed by using k-means technique with the same subset of attributes as it was chosen by AdaBoost algorithm and 2 clusters indicated. Consequently we obtained the first cluster consisted of 24 instances: 18 instances from the set A and 6 cases from the set B. The second cluster included 36 instances: 12 from the set A and 24 from the set B.

The results of all the combinations of methods introduced in section IV-C are presented in table IV, where the first column describes the methods and the last two columns contain the numbers of cases obtained for particular cluster with the reference to the initial data sets A (healthy children) and B (diagnosed children).

It can be easily noticed that in more than 70% of cases group contents were consistent with groups created by the initial expert diagnosis being the result of the most informative attribute analysis.

TABLE IV
THE RESULTS OF CLUSTERING PERFORMED USING PROPOSED METHODOLOGY.

Method	cluster "0"	cluster "1"	% of cases of initial groups
J48-EM	19A / 2B	11A / 28B	78%
J48-k-means	22A / 3B	8A / 27B	82%
AdaBoost-EM	18A / 3B	12A / 27B	75%
AdaBoost-k-means	18A / 6B	12A / 24B	70%

TABLE V
THE LIST OF PARAMETERS CHOSEN BY J48 ALGORITHM OUT FROM ALL THE ATTRIBUTES.

Parameter name	Parameter description
ABPM-S	ambulatory blood pressure monitoring - systolic
ABPM-D	ambulatory blood pressure monitoring - diastolic

D. Verification of results

As the first step in the process of verification we performed classification enabling all the features - also the most informative derived from the standard echocardiography examination. As a result the J48 algorithm pointed to 2 attributes (table V) for the classification task and the AdaBoost algorithm chose 6 attributes (table VI).

After the clustering process we obtained 2 clusters for each combination of methods: EM after J48 classification, k-means after J48, EM after AdaBoost and k-means after AdaBoost. The detailed results are presented in table VII. The first column of the table describes the combination of methods for the classification and clustering. The second and the third columns contain the number of cases in reference to the initial data set of healthy children (A) and data set of diagnosed children (B).

The comparison of results gathered in tables IV and VII allows to conclude that the proposed methodology incorporating less meaningful features produces the cluster structure similar to the clustering based on most informative attributes derived from standard echocardiography. For all the applied algorithms, we obtained more than 60% of cases assigned to the corresponding clusters. The best results were obtained using the combination of J48 classification and EM clustering (80%), and the worst for AdaBoost classification with k-means clustering (65%).

V. CONCLUSIONS AND FUTURE WORK

In medicine, as well as in other fields of science, which include diagnostics techniques, there is a constant need of searching for new methods of attribute acquiring. However it may be difficult to assess if these new features can replace the existing ones and can be useful in medical inference.

In this paper we proposed the methodology of searching for features which are less informative while considering independently, but still meaningful in the process of diagnosis. This approach is mainly useful when new attributes derived from new diagnostics techniques are introduced. These features may

TABLE VI

THE LIST OF PARAMETERS CHOSEN BY ADABOOST ALGORITHM OUT FROM ALL THE ATTRIBUTES.

Parameter name	Parameter description
ABPM-D	ambulatory blood pressure monitoring - diastolic
SF	Shortening fraction
IVSs	Interventricular septum-systole
PWDd	Posterior wall thickness in diastole
SrRRmanSBP	manual arterial blood pressure measurements - systolic
SrRRmanDBP	manual arterial blood pressure measurements - diastolic

TABLE VII

THE RESULTS OF CLUSTERING PERFORMED USING ALL THE FEATURES.

Method	cluster "0"	cluster "1"	% of cases of initial groups
J48-EM	26A / 0B	4A / 30B	93%
J48-k-means	30A / 8B	0A / 22B	87%
AdaBoost-EM	29A / 2B	1A / 28B	95%
AdaBoost-k-means	29A / 7B	1A / 23B	87%

seem to be less meaningful at first and hard to be assessed by medical staff due to multivariate analysis, but the experimental studies confirmed that they enable obtaining the diagnostic results comparable to the ones received by using features recognized as the most informative.

In the first step feature set classification is applied, then taking into account the selected set of attributes clustering is performed. Two different algorithms of classification with two methods of clustering were combined: J48 + k-means, J48 + EM, AdaBoost + k-means, and AdaBoost + EM. During experiments, conducted on real data, we obtain satisfactory results in comparison to the corresponding ones received by cluster analysis carried out by using all the features. Moreover the results did not differ significantly while comparing with the initial groups created by features indicated by experts.

Despite the fact, that the mining methods chosen for the research were widely recommended in the literature as appropriate for medical data, in the future we intend to verify other approaches and build different hybrid solutions to find out methods, which enable discovering new features assuring more precise disease diagnosing.

REFERENCES

- I. Guyon and A. Elisseeff, *An Introduction to Variable and Feature Selection*, Mach Learn Res, vol. 3, 2003, pp. 1157-1182
- Z. Xu, I. King and M. R.-T. Lyu, *Discriminative Semi-Supervised Feature Selection Via Manifold Regularization*, IEEE Transactions on Neural Networks, Vol. 21, No. 7, 2010, pp. 1033-1047, DOI: 10.1109/TNN.2010.2047114
- A. Hamdy and A. E. Hassanien, *The importance of handling multivariate attributes in the identification of heart valve diseases using heart signals*, In: M. Ganzha, L. Maciaszek, M. Paprzycki (eds.) Proceedings of the 2012 Federated Conference on Computer Science and Information Systems, IEEE, 2012, pp. 75-79
- L. Ostrowska -Nawarycz and T. Nawarycz, *Prevalence of excessive body weight and high blood pressure in children and adolescents in the city of Łódź*, Kardiol Pol. Vol. 65, 2007, pp. 1079-1087
- G. Chandrashekar and F. Sahin F, *A survey on feature selection methods*, Computers and Electrical Engineering, Vol. 40, 2014, pp. 16-28, DOI: dx.doi.org/10.1016/j.compeleceng.2013.11.024
- J. Nahar, T. Imama, K.S. Tickle and Y.-P.P. Chen, *Computational intelligence for heart disease diagnosis: A medical knowledge driven approach*, Expert Systems with Applications Vol. 40, 2013, pp. 96-104, DOI: 10.1016/j.eswa.2012.07.032
- Y.F. Wang, M.Y. Chang, R.D. Chiang, L.J. Hwang, C.M. Lee and Y.H. Wang, *Mining Medical Data: A Case Study of Endometriosis*, J Med Syst 37:9899, 2013, DOI: 10.1007/s10916-012-9899-y, DOI: 10.1007/s10916-012-9899-y
- A.K. Tanwani, M.J. Afridi, M.Z. Shafiq and M. Farooq, *Guidelines to Select Machine Learning Scheme for Classification of Biomedical Datasets*, In: C. Pizzuti, M.D., Ritchie, M., Giacobini (eds.), EvoBIO, Springer, 2009, pp. 128-139, DOI: 10.1007/978-3-642-01184-9_12
- M. Pirooznia, J. Yang, M.Q. Yang and Y. Deng, *A comparative study of different machine learning methods on microarray gene expression data*, BMC Genomics, Vol. 9, 2008, DOI:10.1186/1471-2164-9-s1-s13
- Y. Cao, Q.-G. Miao, J.-Ch. Liu and L. Gao, *Advance and Prospects of AdaBoost Algorithm*, Acta Automatica Sinica, Vol. 39, 2013, pp. 745-758, DOI: 10.1016/S1874-1029(13)60052-X
- S. Yuda, L. Short, R. Leano and T.H. Marwick, *Myocardial abnormalities in hypertensive patients with normal and abnormal left ventricular filling: a study of ultrasound tissue characterization and strain*, Clin Sci, Vol. 103(3), 2002, pp. 283-293
- J. Zamojska, K. Niewiadomska-Jarosik, A. Wosiak and J. Stańczyk, *Evaluation of left ventricular systolic function with the use of tissue Doppler echocardiography in children with primary arterial hypertension (Ocena funkcji skurczowej lewej komory z wykorzystaniem metody doplera tkankowego u dzieci z nadciśnieniem tętniczym pierwotnym)*, Pol J Cardiol Vol. 4(2), 2012, pp. 95-100
- S.K. Murthy, *Automatic construction of decision trees from data: a multi-disciplinary survey*, Data Mining and Knowledge Discovery, vol.2, 1998, pp. 345-389, DOI: 10.1023/A:1009744630224
- L. Rokach and O. Maimon, *Data mining with decision trees: theory and applications*, Machine perception and artificial intelligence, vol. 69, 2008, WorldScientific Publishing (Singapore)
- J. Cerquides, M. López-Sánchez, S., Ontañón, E. Puertas, A. Puig, O. Pujol and D. Tost, *Classification Algorithms for Biomedical Volume Datasets*, In: R. Marín, E. Onaindía, A. Bugarín and J. Santos, (eds.) Current Topics in Artificial Intelligence. LNCS, Springer Berlin Heidelberg, vol. 4177, 2006, pp. 143-152, DOI: 10.1007/11881216_16
- M. Czajkowski, M. Grześ and M. Kretowski, *Multi-test decision tree and its application to microarray data classification*, Artif Intell Med, 2014, DOI: http://dx.doi.org/10.1016/j.artmed.2014.01.005
- C.E. Brodley and P.E. Utgoff, *Multivariate decision trees*. Machine Learning, 1995, pp. 45-77
- M. Hall, E. Frank, G. Holmes, B. Pfahringer, R. Reutemann, I. H. Witten, *The WEKA data mining software: an update*, SIGKDD Explor. NewsL. vol. 11, 2009, pp. 10-18, DOI: 10.1145/1656274.1656278
- J.R. Quinlan, *Bagging, Boosting, and C4.5*, In: Thirteenth National Conference on Artificial Intelligence, AAAI Press, 2006, pp. 725-730
- S.R. Konda, *A Comparative Evaluation Of Symbolic Learning Methods and Neural Learning Methods*, https://www.cs.umd.edu/grad/scholarlypapers/papers/ShravyaKonda.pdf
- Y. Freund and R. E. Schapire, *A decision-theoretic generalization of on-line learning and an application to boosting*, In: Proceedings of the Second European Conference on Computational Learning Theory, Paul M. B. Vitányi (Ed.), Springer-Verlag, London, UK, 1995, pp. 23-37, DOI: 10.1006/jcss.1997.1504
- W. Iba and P. Langley, *Induction of one-level decision trees*, In: Ninth International Workshop on Machine Learning, Morgan Kaufmann Publishers Inc., San Francisco, USA, 1992, pp.233-240, URL: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.23.2878
- J. B. MacQueen, *Some Methods for Classification and Analysis of Multi-Variate Observations*, In: Fifth Berkeley Symposium on Mathematical Statistics and Probability, University of California Press, 1967, pp. 281-297,
- V. Ankita, R. V. Satyanarayana and K. Kamalakar, *An Experiment with Distance Measures for Clustering*, In: International Conference on Management of Data, Technical Report, 2008,
- I. H. Witten, E. Frank and M. A. Hall, *Data Mining: Practical Machine Learning Tools and Techniques (3rd ed.)*, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2011