*pattern recognition, compound methods, classifier ensembles,
machine learning, unbalanced classification, two-stage recognition*

Bartosz KRAWCZYK[1], Michał WOŹNIAK[1]

HYPERTENSION DIAGNOSIS USING COMPOUND PATTERN RECOGNITION METHODS

The paper presents a hypertension type classification task where the decisions should be made only on the basis of blood pressure, general information and basis biochemical data. This problem has a great importance to the medical decision support systems, yet results achieved so far are not satisfactory. When the canonical approaches tend to fail we should look for the compound pattern recognition systems, such as multiple classifiers systems. This article presents the results of an experimental investigation of the pool of compound classifiers which have their origin in classifiers ensembles, random forest, and random subspace. Presented methods returned good, satisfactory results, outperforming canonical approaches for this problem.

1. INTRODUCTION

The aim of the recognition task [10] is to classify a given object of interest by assigning it to some predefined category, on the basis of observing the features of the object. Depending of the practical application, these objects (so-called patterns) can be images, signal waveforms or any type of measurements that need to be classified [24].

Since the publication of Frank Rosenblatt's work devoted to the idea of a perceptron [21] that time, the progress of computers technology has increased the demand for practical applications of pattern recognition and caused the development of new efficient theoretical methods of recognition required by more and more sophisticated decision problems.

There is much current research into developing even more efficient and accurate recognition algorithms, like neural networks, statistical and symbolic learning, fuzzy methods to name only a few. Such methods are implemented in the form of computer software and applied in many practical areas, like character and speech recognition, machine vision, computer aided medical diagnosis, prediction of customer behavior, fraud detection etc.

Medical diagnosis is a very important and attractive area of implementation decision support systems. About 11% of expert systems are dedicated to the medical aided diagnosis and ca 21% of papers connected with application of aforementioned methods are illustrated by the medical cases [15].

In the paper we present four types of compound pattern recognition algorithms: two-stage classifier, random forest, random subspace and feature driven space division. Additionally we discus if balancing the uneven class representation can improve the quality of object recognition for the real medical problem.

The content of the work is as follows. Section 2 consists of short descriptions of proposed algorithms. In Section 3 we describe mathematical model of the hypertension's type. Next section presents results of the experimental investigations of the algorithms. Section 5 concludes the paper.

[1] Department of Systems and Computer Networks, Wroclaw University of Technology,
Wybrzeze Wyspianskiego 27, 50-370 Wroclaw, Poland.
email: {bartosz.krawczyk , michal.wozniak}@pwr.wroc.pl.

2. ALGORITHMS

2.1. DECISION TREE INDUCTION

The basic idea involved in a multistage approach is to break up a complex decision into a set of simpler classifications [22]. A decision tree classifier is a one of the possible approaches to the multistage pattern recognition. Hierarchical classifiers are a special type of the multistage classifiers which allow rejection of class labels at intermediate stages. The synthesis of hierarchical classifier is a complex problem. It involves specification of the following components [12]:

- design of a decision tree structure,
- selection of features used at each noterminal node of decision tree,
- choice of decision rules for performing the classification.

The decision tree induction algorithms have been developed for several years [2]. From the mathematical point of view they propose how to estimate discrete functions which could be adapted to classification tasks. From the practical point of view decision trees achieve pretty good results in many real decision tasks. Among different methods of tree training top down decision tree induction concept is very popular. Algorithms based on aforementioned idea train a tree from a root node to leaf ones using a splitting attribute's choosing measure. The most famous representative of algorithm family using aforementioned concept is ID3 developed by Quinlan [18]. ID3 uses information gain measure to decide which attribute should be tested in a given node. Proposed measure evaluates how homogenous are subsets of training set (according to the given class labels) obtained on the basis of original set split using the chosen attribute values.

Descendants of IDs improve its main features. The main disadvantage of information gain, is that it prefers features with high number of values. C4.5 [19] uses another attribute measure information ratio. Both measures based on information theory which uses Shannon's entropy as a measure, but many other measures are proposed like Gini metric used e.g., by CART or χ^2 statistic [4] to enumerate only a few. Aforementioned algorithms propose also methods which protect tree classifiers against overtraining like reduce-error pruning or rule post pruning, show how to deal with continuous attributes, how to handle with missing attribute values and attributes with weights, how to reduce computational complexity, and how to use algorithm in distributed computing environments.

2.2. BOOSTING

Boosting [23] is general method of producing an accurate classifier on base of weak and unstable one. It is often called meta-classifier. The idea of boosting has its root in PAC (Probably Approximately Correct) theory. The underlying idea of boosting is to combine simple classifiers to form an ensemble such that the performance of the single member of ensemble is improved. As we see the main problem of the boosting is how to construct ensemble. The main advantage of boosting is that it often does not suffer from overfitting.

The one of the most popular algorithm AdaBoost produces at every stage, a classifier which is trained with the modified learning set. The output of the classifier is then added to the output of classifier ensemble, with the strength proportional to how accurate obtained classifier is. Then, the elements of learning set are reweighted: examples that the current learned function gets wrong are "boosted" in importance, so that the classifier obtained at the next stage will attempt to fix the errors. We adapted the AdaBoost for classifiers which cannot use the weights in decision making process. In this case we have to generate learning sequence according to the weights (distributions) of elements in each iteration. Generated learning set is the base of WeakLearn algorithm. This concept is similar, but not the same, as presented in [5] called boosting by subsampling.

2.3. RANDOM FOREST

Random forest was introduced by Breiman [6]. It is to some extent an extension of the boostrap aggregation [5] algorithm. The classifier itself consist of a number of decision trees, each of which uses only a randomly selected subspace of features for training. Every tree is fully grown and no pruning is used. Apart from that, on the bases of the given dataset, like in bagging, a number of subsets were generated by uniformly sampling the examples with replacement from the standard training set. Random forest is used to increase the predictive performance of weak tree classifiers and to introduce the diversity into the ensemble. It can be used additionally as a feature selection method, as we can see the performance of base trees, created with different features.

2.4. RANDOM SUBSPACE

Random subspace method (or attribute bagging) [7] is an ensemble classifier that consists of several classifiers and outputs the class based on the outputs of these individual classifiers. Random subspace method is a generalization of the random forest algorithm. Whereas random forests are composed of fully-grown decision trees, a random subspace classifier can be composed from any underlying classifiers. Random subspace method has been used for various combinations of decision trees, linear classifiers, support vector machines and other types of classifiers. It consists of three main steps: adjusting the number of base classifiers, adjusting the number of subspaces and randomly selecting features for each of those subspaces, on which the classifiers are trained. This allow go increase the diversity of the ensemble. The Random subspace method does not impose itself what fusion method should be chosen.

2.5. FEATURE DRIVEN SPACE DIVISION

Feature driven space division introduced by Krawczyk [13], is a novel classifier ensemble method designed for complex data. A feature space is partitioned into the much smaller, disjoint subspaces. Each of them is created by the usage of a feature selection algorithm and then it is used to train the classifier. One proposes to use the random subspaces method, which deliver good results. By the usage of feature selection algorithms for this task it is possible to ensure that created subspaces consist of relevant features. Therefore this algorithm divides N-dimensional feature space into L subspaces, where $N = N_1 \cup N_2 \cup ... \cup N_L$. The example of such division, for nine features and three target subspaces, is presented in Fig. 1. Then the most relevant classifiers are selected and their outputs are fused to deliver the final decision. They are ranked according to their individual accuracy.

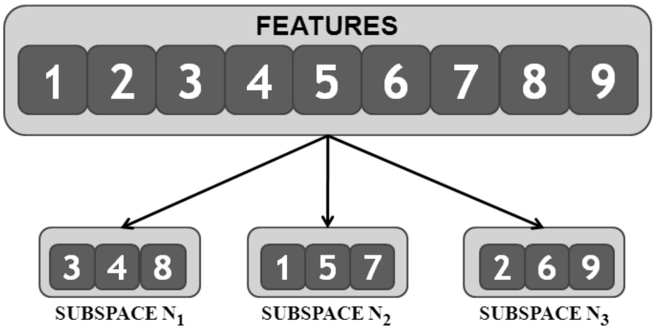

Fig. 1. Idea of splitting feature space between 3 different subspaces/classifiers.

2.6. SMOTE

A data set is imbalanced if the classification categories are not approximately equally represented. Often real-world data sets are predominately composed of normal examples with only a small percentage of abnormal or interesting examples. The performance of classification algorithms is typically evaluated using predictive accuracy. However, this is not appropriate when the data is imbalanced. Therefore this is a crucial problem in the process of pattern recognition.

The most common approach to this problem is the SMOTE algorithm [9]. For a subset $S_{\min} \in S$, where S stands for training set and $S_{\min}$ for minority class, the k-nearest neighbors (k-NN) are considered for each one of the examples $x_i \in S_{\min}$. To create a synthetic data, one of the k-NN is randomly selected, its corresponding feature vector difference is multiplied with a random number between [0 - 1] and added to x_i:

$$x_{new} = x_i + (\hat{x}_i - x_i) \times \delta, \tag{1}$$

where x_i is the minority example under consideration, $\hat{x}_i$ is one of the k-NN of x_i (randomly chosen) and δ is a random number between 0 and 1. The resulting synthetic example is some point along the segment joining x_i under consideration and the randomly selected $\hat{x}_i$.

Example of SMOTE mechanism in a two dimensional space is shown in Fig. 2. In the presented example the number of neighbours is set to three. "Pluses" stands for positive, minority class, "minuses" stands for negative, majority class and "star" presents the introduced new synthetic object.

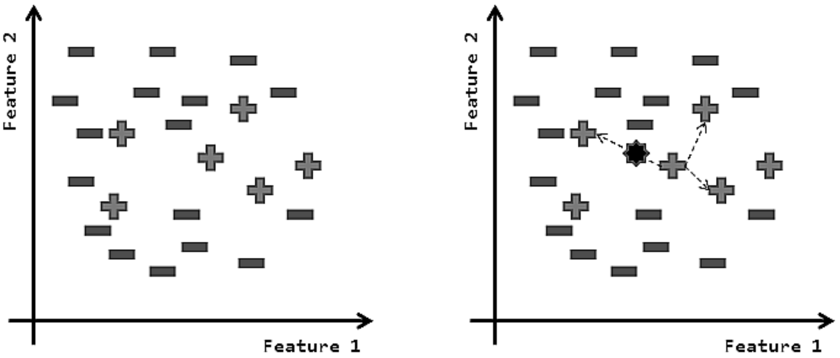

Fig. 2. Idea of the SMOTE. On the left an imbalanced data set and on the right a synthetic object created.

3. MODEL OF TYPE OF HYPERTENSION DIAGNOSIS

During the hypertension’s therapy is very important to recognize the state of patient and the correct treatment. The physician is responsible for deciding if the hypertension is of an essential or a secondary type (so called the first level diagnosis). The senior physicians from the Broussais Hospital of Hypertension Clinic and Wroclaw Medical Academy suggest 30% as an acceptable error rate for the first level diagnosis. The presented project was developed together with Service d'Informatique Médicale from the University Paris VI. All data was gathered from the medical database *ARTEMIS*, which contains the data of the patients with hypertension, whose have been treated in Hôpital Broussais in Paris.

The mathematical model was simplified. However our experts from the Broussais Hospital, Wroclaw Medical Academy, regarded that the stated problem of diagnosis is very useful. It leads to the following classification of type of hypertension:

1. essential hypertension (abbreviation: essential),
2. fibroplastic renal artery stenosis (abbreviation: fibro),
3. atheromatous renal artery stenosis (abbreviation: athero),
4. Conn’s syndrome (abbreviation: conn),
5. renal cystic disease (abbreviation: poly),
6. pheochromocytoma (abbreviation: pheo).

Although the set of symptoms necessary to correctly assess the existing HT is pretty wide, in practice for the diagnosis results of 18 examinations (which came from general information about patient, blood pressure measurements and basis biochemical data) are used. They are presented in Table 1.

4. EXPERIMENTAL INVESTIGATION

4.1. EXPERIMENTS SETUP

All experiments were carried out in the *R* environment, with classification algorithms taken from the dedicated packages, thus ensuring that the results achieved the best possible efficiency and that the performance was not decreased by a bad implementation. All tests were done by a 10-fold cross validation.

Table 1. Clinical features considered.

No	Feature	No	Feature
1	sex	10	effusion
2	body weight	11	artery stenosis
3	high	12	heart failure
4	cigarette smoker	13	palpitation
5	limb ache	14	carotid or lumbar murmur
6	alcohol	15	serum creatinine
7	systolic blood pressure	16	serum potassium
8	diastolic blood pressure	17	serum sodium
9	maximal systolic blood pressure	18	uric acid

4.2. CANONICAL PATTERN RECOGNITION METHODS

Most of the users turn to the canonical pattern recognition algorithms as the first algorithms of choice. They are well-established in the pattern recognition field, delivering good results in many areas of usage. Therefore they were used as the first base-line algorithms for further comparison. We have tested six algorithms: C4.5, Alternative Decision Tree (ADTree), Support Vector Machine (SVM), Neural Network (NN), *k*-Nearest Neighbours (*k*-NN) with *k* set to 3 and Quadratic Classifier (QDA) [11,12]. First two of them were used by Woźniak [22], four remaining were tested for the purpose of this study. Their results are showed in Table 2. Unfortunately we had to reject the classifiers because their quality did not satisfy expert.

Table 2. Canonical classifiers performances.

C4.5	ADTree	SVM	NN	k-NN	QDA
67,79%	58,48%	**68,21%**	64,35%	52,54%	57,97%

4.3. COMPOUND PATTERN RECOGNITION METHODS

As seen in previous subsection, the canonical methods delivered the unsatisfactory results and were discarded by our experts. To deal with such complex problem we propose to use the compound pattern recognition methods. Idea behind them can be explained as follow: when one classifier tend to fail, introduce a combination of the classifiers and somehow fuse their outputs to deliver final result. It was shown that such approach tends to behave better for many problems. Now this area is very broad, consisting of such approaches as multistage recognition [8], multiple classifier systems [14] and problem binarization [17]. For the examined problem we focused on boosting, two-stage classifier, random forest and random subspace methods.

Boosting for hypertension recognition was introduced by Woźniak [25]. He used this algorithm to increase the quality of C4.5 and ADTree classifiers.

In the same paper Woźniak [25] proposed a two-stage classifier approach to this problem. He constructed a classifier ensemble based on the two-stage classifier concept. Its idea is depicted in Fig. 3.

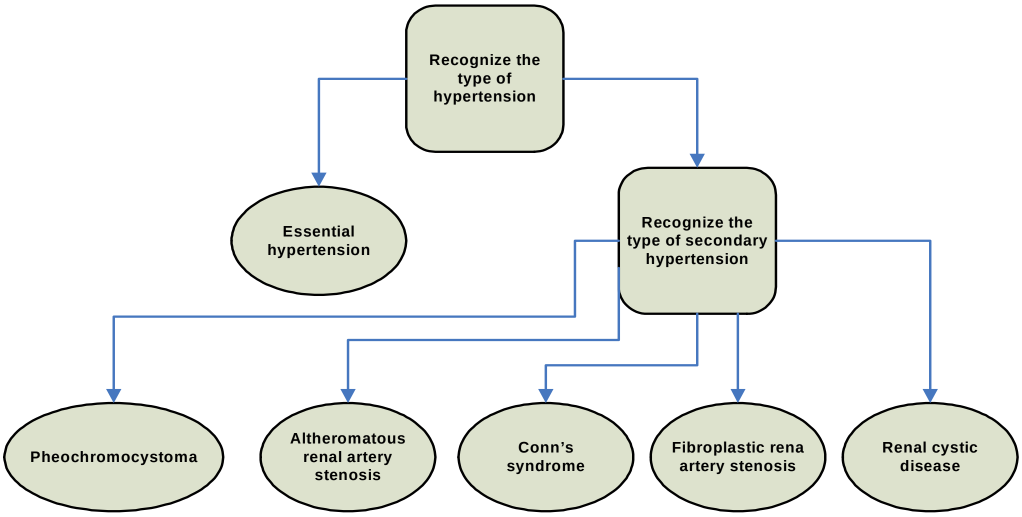
Recognize the type of hypertension
Essential hypertension
Recognize the type of secondary hypertension
Pheochromocystoma
Altheromatous renal artery stenosis
Conn's syndrome
Fibroplastic rena artery stenosis
Renal cystic disease

Fig. 3. Two-stage classifier of hypertension's type.

Boosted ADTree classifier was used for the first stage of recognition. For the second stage recognition a rule-based classifier was proposed. Human experts were queried for the set of rules for this problem and 17 of them were obtained.

Random Forest algorithm were tested for this problem with a different number of trees in ensemble (20 and 80), a different number of features for each of the trees (2,3,5 and 10) and also with the Principal Component Analysis (PCA) data pre-processing method (2,10 and 18 first components were used).

Random Subspace algorithm was tested for C4.5 and RandomTree algorithms, with subspace size fixed to 0.5 of the whole feature space. PCA with 18 components was also used for this approach. The results of each of the base classifiers were combined by the majority voting procedure [14].

Feature Driven Space Division was used with SVM algorithm and ReliefF feature selection algorithm [20]. Subspace size was fixed to four and three most relevant classifiers were used in the final decision making step (with 12 features in total).

The results are showed in Table 3.

Table 3. Compound methods performances.

BoostC4.5	**BoostADT**	**Two-stage**	**RF20,10**	**RF20,5**	**RF20,3**
69,42%	68,90%	***73,07%***	64,23%	68,45%	70,02%
RF20,2	**RF80,10**	**RF80,5**	**RF80,3**	**RF80,2**	**RF80,2 + PCA2**
70,15%	65,12%	68,28%	70,03%	70,53%	68,43%
RF80,2 + PCA10	**RF80,2 + PCA18**	**RS,C4.5**	**RS,RT**	**RS,RT + PCA18**	**FDSD**
68,89%	70,89%	71,24%	71,65%	70,73%	70,83%

As we can see the difference in the best received results are oscillating about 2-3%. That is why we decided to use a statistical significance test, to compare the results and see if their differences are statistically significant. For this purpose we used a Combined 5 × 2 cv F Test [1]. As this test is done by comparison all versus all, we decided to test only the best method (two-stage classifier) with four others, giving closest results. As a test score we used the probability of rejecting the null hypothesis – that both classifiers have the same error rates. A small difference in error rate implies that the different algorithms construct two similar classifiers with similar error rates; thus the hypothesis should not be rejected. For a large difference, the classifiers have different error rates, and the hypothesis should be rejected. Results of statistical test are shown in Table 4.

Table 4. Probabilities of Rejecting the Null Hypothesis.

Two-stage vs. RS,RT	**Two-stage vs. RS,C4.5**	**Two-stage vs. RF80,2 + PCA18**	**Two-stage vs. FDSD**
0.095	**0.129**	0.440	0.545

As we can see in the Table 4 two-stage classifier is statistically similar to the random subspace with random tree method. Also low probability of rejecting the null hypothesis was received for the random subspace with C4.5. It means, that despite the differences in their accuracies, we cannot say from statistical point of view, that one is better than the other.

To our experts the threshold reached by the two-stage classifier was satisfactory, yet what is worth noticing all of the proposed compound pattern recognitions methods achieved similar results, using only one-stage classification. These results encouraged us to find an additional way to improve their accuracy.

4.4. IMBALANCED CLASS DISTRIBUTION PROBLEM

We noticed that the weaker results may occur from the imbalanced class distribution. Let us note that essential hypertension is represented by 912 objects. Each of the other classes consists of 80-140 objects. This is major disproportion, which for sure has an effect on overall classification performance. To deal with this problem we used SMOTE algorithm. We tested it on the four most promising one-stage compound approaches – random subspace with random tree / C4.5, random forest with 80 trees, each consisting of two features with PCA used and feature driven space division.

With the usage of SMOTE comes the main problem – how many of the artificial samples we should generate. Intuition points that best results should be achieved when classes have equal number of objects. Yet with so big disproportion (approximately 9:1) after some repetitions of this algorithm the new objects will be created only on the basis of previously artificially created ones. Therefore it is hard to conclude if so many artificial objects will be representative for the problem. To analyze this we propose a following approach. SMOTE was tested for increasing the number of objects from 100% to 900%. The result of classification was tested by 10CV and additionally by a validation set, consisting of 300 original objects, randomly removed from the training set. Therefore we can see how the number of artificial objects influence the ability of recognizing new, unseen real-life objects. To get the best possible comparison with ten-fold cross validation and avoid unfortunate selection of the objects, the testing set was randomly chosen 10 times. Results for random subspace with random tree method are shown in Fig. 4. All other methods behave similarly, achieving the best results for the same SMOTE parameters - so there is no point with presenting their individual graphs. Results achieved for the four tested methods with SMOTE set to 200% of new objects, are presented in the Table 5.

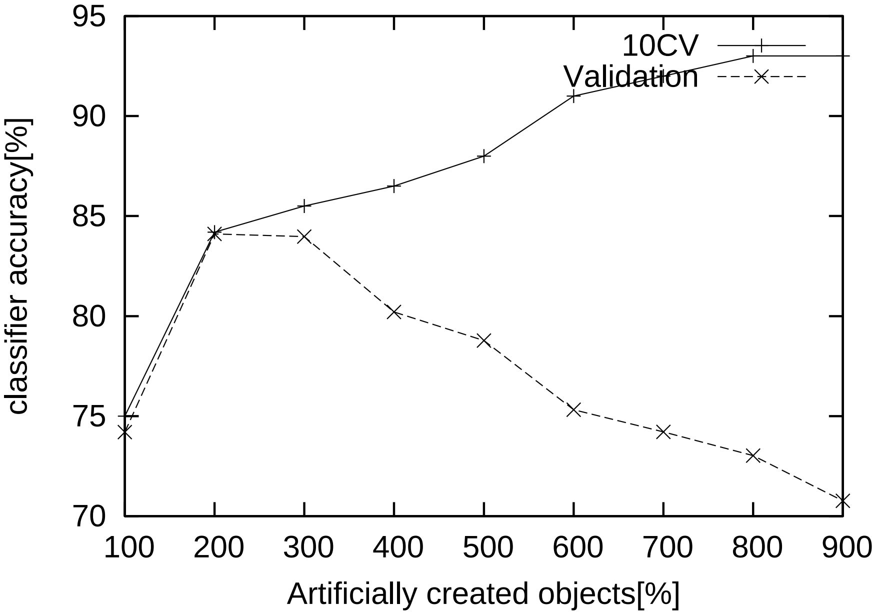

Fig. 3. SMOTE effect on the classification process.

Table 5. Performances of the compound methods after usage of the SMOTE algorithm.

RS,RT	RS,C4.5	RF80,2 + PCA18	FDSD
84,20%	83,47%	81,75%	82,55%

Best results were obtained for the random subspace method using the random tree algorithm. As previously we compared the best received method with others using the statistical significance test. Results are presented in the Table 6.

Table 6. Performances of the compound methods after usage of the SMOTE algorithm.

RS,RT vs. RS,C4.5	RS,RT vs. RF80,2 + PCA18	RS,RT vs. FDSD
0.252	0.700	0.490

As we can see in the Table 6 the best method is statistically quite similar to the random subspace with C4.5 algorithm.

5. DISCUSSION AND CONCLUSION

The methods of inductive learning were presented. The classifiers generated by these algorithms were applied to the medical decision problem (recognition of the type of hypertension). For the real decision problem we compared the canonical classifiers and compound methods. Generally the compound methods outperformed significantly the simple classifiers. Additionally by usage of the SMOTE algorithm we managed to outperform the two-stage classifier, which returned the best results up to date.

Several interesting conclusions can be drawn from those experiments. Feature Space Division algorithm was created for small sample, high dimensionality problems and in a normal tasks did not outperform random subspaces. Random Forests for this task returned better results, when consisting of alarge number of small trees. Increasing the tree size decreased their accuracy. Using PCA additionally improved their performance. On the other hand PCA did not cope well with Random Subspace method. Class re-balancing had a great impact on the classification process, yet it should be stopped at the threshold of 200% of artificial examples. Further increasing their numbers lead to the drop of overall accuracy. Statistical test of significance showed that despite differences in received accuracy some of the methods give similar classifiers.

The similar problem of the computer-aided diagnosis of hypertension's type was described in [3] but authors used another mathematical model and implement Bayes decision rule. They obtained slightly better classifier than two-stage recognition, yet usage of the other compound patter recognition method combined with introducing artificial objects outperformed that approach. Additional advantage of our proposition is simplified and cheaper model than presented in [3] (we use 18 features, authors of mentioned work use 28 ones).

Advantages of the proposed methods make it attractive for a wide range of applications in medicine, which might significantly improve the quality of the care that the clinicians can give to their patients.

6. ACKNOWLEDGEMENT

This work is supported by the Polish National Science Centre under a grant for the period 2011-2014.

BIBLIOGRAPHY

[1] ALPAYDIN E., Combined 5 x 2 cv F Test for Comparing Supervised Classification Learning Algorithms, Neural Computation, Vol. 11, 1998, pp. 1885-1892.

[2] ALPAYDIN E., Introduction to Machine Learning. Second edition, The MIT Press, Cambridge, MA, USA, London, UK, 2010.

[3] BLINOWSKA A., et al., Bayesian Statistics as Applied to Hypertension Diagnosis, IEEE Trans. on Biomed. Eng., Vol. 38, No. 7, 1991, pp. 699-706.

[4] BREIMAN L., FRIEDMAN J.H., OLSHEN R.A., STONE C.J., Classification and regression trees, Wadsworth and Brooks, Monterey, CA, 1984.

[5] BREIMAN L., Bagging predictors, Technical Report 421, Department of Statistics, University of California, Berkeley, 1994.

[6] BREIMAN L., Random forests, Machine Learning, Vol. 45, No. 5, 2001, pp. 32.

[7] BRYLL R., Attribute bagging: improving accuracy of classifier ensembles by using random feature subsets, Pattern Recognition, Vol. 20, No. 6, 2003, pp. 1291–1302.

[8] BURDUK R., Case of Fuzzy Loss Function in Multistage Recognition Algorithm, Journal of Medical Informatics & Technologies, Vol. 5, 2003, pp. 107-112.

[9] CHAWLA N.V., BOWYER K.W., HALL L.O., KEGELMEYER W.P., SMOTE: Synthetic Minority Over-sampling Technique, Journal of Artificial Intelligence Research, Vol. 16, 2002, pp. 321-357.

[10] DUDA R.O., HART P.E., STORK D.G., Pattern Classification, Wiley-Interscience, 2001.

[11] HASTIE T., TIBSHIRANI R., FRIEDMAN J., The Elements of Statistical Learning. Data Mining, Inference, and Prediction, Springer Series in Statistics, Springer Verlag, New York 2001.

[12] JAIN A.K., DUIN P.W., MAO J., Statistical Pattern Recognition: A Review, IEEE Transaction on Pattern Analysis and Machine Intelligence, Vol. 22., No. 1, 2000, pp. 4-37.

[13] KRAWCZYK B., Classifier committee based on feature selection method for obstructive nephropathy diagnosis, Semantic Methods for Knowledge Management and Communication, KATARZYNIAK R. et al. (Eds.), Springer, Studies in Computational Intelligence, Vol. 381, 2011, pp. 115-125.

[14] KUNCHEVA L.I., Combining Pattern Classifiers: Methods and Algorithms, Willey, 2004.

[15] LIEBOWITZ J. (ED), The Handbook of Applied Expert Systems, CRC Press, 1998.

[16] MUI J., FU K.S., Automated classification of nucleated blood cells using a binary tree classifier, IEEE Trans. Pattern Anal. Mach. Intell., Vol. 2, 1980, pp. 429-443.

[17] OPITZ D., MACLIN R., Popular Ensemble Methods: An Empirical Study, Journal of Artificial Intelligence Research, Vol. 11, 1999, pp. 169-198.

[18] QUINLAN J.R., Induction of decision trees, Machine Learning, Vol. 1, No. 1, 1986, pp. 81-106.

[19] QUINLAN J.R., *C4*.5: Programs for Machine Learning, Morgan Kaufmann, 1993.

[20] ROBNIK-SIKONJA M., KONONENKO I., Theoretical and empirical analysis of relieff and rrelieff. Machine Learning, Vol. 53, No. 23, 2003, pp. 69.

[21] ROSENBLATT F., The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain, Psychological Review, Vol. 65, No. 6, 1958, pp. 386-408.

[22] SAFAVIAN S.R., LANDGREBE D., A survey of decision tree classifier methodology, IEEE Trans. Systems, Man Cyber., Vol. 21, No. 3, 1991, pp. 660-674.

[23] SCHAPIRE R.E., The boosting approach to machine learning: An overview. Proc. Of MSRI Workshop on Nonlinear Estimation and Classification, Berkeley, CA, 2001.

[24] THEODORIDIS S., KOUTROUMBAS K., Pattern Recognition, Academic Press, Amsterdam, 2003.

[25] WOZNIAK M., Two-Stage Classifier for Diagnosis of Hypertension Type, Lecture Notes in Bioinformatics, Springer-Verlag, Berlin Heidelberg New York, Vol. 4345, 2006, pp. 433-440.