cost-sensitive classification; multiple classifier systems,
pattern recognition genetic algorithm

Michał WOŹNIAK[1], Marcin ZMYŚLONY[1]

COST-SENSITIVE CLASSIFIER ENSEMBLE FOR MEDICAL DECISION SUPPORT SYSTEM

Multiple classifier systems are currently the focus of intense research. In this conceptual approach, the main effort focuses on establishing decision on the basis of a set of individual classifiers' outputs. This approach is well known but usually most of propositions do not take exploitation cost of such a classifier under consideration. The paper deals with the problem how to take a test acquisition cost during classification task under the framework of combined approach on board. The problem is known as cost-sensitive classification and it has been usually considered for the decision tree induction. In this work we adapt mentioned above idea into choosing members of classifier ensemble and propose a method of choosing a pool of individual classifiers which take into consideration on the one hand quality of ensemble on the other hand cost of classification. Properties of mentioned concept are established during computer experiments conducted on chosen medical benchmark databases from UCI Machine Learning Repository.

1. INTRODUCTION

The aim of a pattern recognition task is to classify a given object by assigning it to the one of the predefined categories [5]. There are many propositions how to automate this process, but according to „no free lunch theorem" there are not a single solution which could solve all tasks, because classifiers have different domains of competence [22].

It is worth noting that a chosen classifier could make mistakes because of:

- each classifier has its preferences which are usually are cold inductive bias, therefore its model could not fit to the real considered target concept or model is simplified,
- learning set is limited,
- learning set is unrepresentative or includes errors (wrong labels or measurement errors).

Fortunately we are not doomed to failure because usually for each classification task we could obtain a pool of different classifiers and for such a case we can choose the best one on the basis of evaluation process or use all available classifiers to exploit the individual classifiers' strengths.

In several review articles multiple classifier systems (MCSs) have been mentioned as one of the most promising pattern recognition approach [9]. In this conceptual method, the main effort concentrates on combining knowledge of the set of elementary classifiers [3]. The main motivations of using MCSs are as follows:

- for small sample MCSs could avoid selection of the worst classifier [12],
- there are many evidences that classifiers combination can improve the performance of the best individual ones and it can exploit unique classifier strengths [20],
- additionally combined classifier could be used in distributed environment, especially in the case that database is partitioned from privacy reason and in each node of computer network only final decision could be available.

During designing computer decision support systems the cost of their designing and exploitation plays the key role. The cost of exploitation can be considered as the expenses of incorrect diagnosis or expenses of feature value acquisition. The first problem is the typical problem for decision theory where one wants to find the classifier with the lowest cost of misclassification [5]. This work focuses on the problem where the cost depends on real expenses of feature values acquisition for decision making [7,18].

[1] Department of Systems and Computer Networks, Wroclaw University of Technology, Wyb.Wyspianskiego 27, 50-370 Wroclaw, Poland, email: {Michal.Wozniak, Marcin.Zmyslony}@pwr.wroc.pl.

The typical example of cost sensitive classification is medical diagnosis. Let’s note that nowadays for the many decision tasks we can to make the high-quality medical decision on the basis of the expensive medical tests. Unfortunately, in real cases physicians have to balance the costs of various tests with the expected benefits or doctors have to make the diagnosis fast on the basis of the fast measured (low cost) features because therapeutical action has to be taken without delay, but it is worth mentioning that the problem of cost-sensitive decision making arises frequently in many fields of human activities except medicine [14] as industrial production process [21], robotics [16], technological diagnosis [11] to enumerate only a few.

Our paper addresses with the popular method of combined classifiers based on the majority voting rule. We propose the cost sensitive algorithm of classifier ensemble design which respects on one hand the cost of feature values acquisition and on the other hand the quality of compound classifier based on the mentioned above ensemble.

The content of the work is as follows. Section 2 provides idea of combined classifiers and the related works which take into account cost during learning process. Section 3 describes our modification of cost-sensitive approach to ensemble design. In the next section results of the experimental investigations are presented. The last part concludes the paper.

2. COMBINED CLASSIFIERS

Designing a combined classifier is similar to the design of a classical pattern recognition system [6]. When designing a typical classifier, the aim is to select the most valuable features and choose the best classification method from the set of available ones. The design of a classifier ensemble is similar – it is aimed to create a set of complementary and diverse classifiers. The design of a fuser is aimed to create a mechanism that can exploit the strength of classifiers from an ensemble and combine them optimally.

There are a number of important issues when building such a system, which can be grouped into two main problems:

- how to design the classifier ensemble,
- how to design the fuser.

Apart from increasing the computational complexity, combining similar classifiers should not contribute to the system becoming constructed. Therefore, selecting members of the committee with different components seems interesting. An ideal ensemble consists of classifiers with high accuracy and high diversity, i.e. mutually complementary. Several papers introduce different types of diversity measures that allow for the possibility of a coincidental failure to be minimized [1]. A strategy for generating the committee of individual classifiers must seek to improve the ensemble’s diversity. To enforce classifier diversity we could use varying components:

- different input data e.g., we could use different partitions of a data set or generate various data sets by data splitting, cross-validated committee, bagging, boosting [10], because we hope that classifiers trained on different inputs are complementary,
- classifiers with different outputs i.e., each individual classifier could be trained to recognize a subset of only predefined classes (e.g., binary classifier - one class against rest ones strategy) and the fusion method should recover the whole set of classes. A well known technique is Error-Correcting Output Codes [4],
- classifiers with the same input and output, but trained on the basis of a different model or model’s versions.

The problem of assuring high diversity of classifier ensemble is crucial for quality of above mentioned compound classifiers.

The second problem is how to negotiate a common decision by an ensemble of selected individual classifiers. The first group of methods includes algorithms using class numbers of individual classifiers only [17] when the second one focuses on problem of combined the support function. In this work we concentrate on the first group which includes intuitive and flexible propositions because they can combine outputs of classifiers using different pattern recognition models. Initially only majority-voting schemes were implemented, but in later work more advanced methods were proposed.

Many known conclusions regarding the classification quality of combined classifiers have been derived analytically, but are typically valid only under strong restrictions, such as particular cases of the majority vote [8] or make convenient assumptions, such as the assumption that the classifier committee is formed from independent classifiers. Unfortunately, such assumptions and restrictions are of a theoretical character and not useful in practice. From aforementioned research we can make the following conclusion that it is worthy to combine classifiers only if the difference among their qualities is relatively small. It has to be noted that the higher the probability of misclassification of the best classifier, the smaller quality difference should be in order to get an effective committee that outperforms their components. Some additional information about voting classifier quality can be found in [1,10].

3. ALGORITHM

Let us assume that we have *n* classifiers $\Psi^{(1)}, \Psi^{(2)}, ..., \Psi^{(n)}$ and each of them decides if a given object belongs to class $i \in \boldsymbol{M} = \{1, ..., M\}$. The majority voting decision rule of combining classifier $\overline{\Psi}$ is as follows :

$$\overline{\Psi}(x) = \arg\max_{j \in \boldsymbol{M}} \sum_{l=1}^{n} \delta\left(j, \Psi^{(l)}(x)\right) \Psi^{(l)}(x) \tag{1}$$

and

$$\delta(j,i) = \begin{cases} 0 & \text{if} \quad i \neq j \\ 1 & \text{if} \quad i = j \end{cases} \tag{2}$$

In our case we would like to choose a pool of individual classifiers which take into consideration on the one hand quality of ensemble on the other hand cost of classification. Therefore we have to formulate the optimization task with the following optimization criterion *C* which assesses the quality of the combined classifier $\overline{\Psi}$

$$C(\Psi) = \frac{2^{Error(\overline{\Psi})} - 1}{\left(Cost(\overline{\Psi}) + 1\right)^{\omega}}, \tag{3}$$

where:

$Error(\overline{\Psi})$ is classification error of $\overline{\Psi}$ using decision rule (1),

$Cost(\overline{\Psi})$ is a sum of exploitation costs of classifiers from the committee, where exploitation cost of an individual classifier is the cost of tests used by it,

ω is the strength of the bias toward the lowest cost attributes. In the case of $\omega = 0$ the feature acquisition cost is ignored and $C(\overline{\Psi})$ has the same features as the $Error(\overline{\Psi})$, if $\omega = 1$ the mentioned cost plays the most important role.

The mentioned above proposition of the criterion is similar to the split criterion used by Nunez in his cost-sensitive decision tree induction algorithm EG2 [15].

To solve the optimization problem we propose learning algorithm which uses the genetic approach with the traditional binary representation [13]. Our objective is to find such an classifier ensemble which maximizes the criterion (3). Let us assume that we have a pool of k individual classifiers and we have to choose the best *n* classifiers (*n*<*k*) according criterion (3). As the representation we propose *k* bits word. If *l*th bit is equal 1 then it means that *l*th classifier from a pool is nominated to ensemble, otherwise it is not member of the committee. Of course *n* bits are equal 1 exactly.

Let us present some important elements of the implementation of the algorithm for ensemble learning problem. Information concerning the recognition task that being the basis for proposed algorithm are included in the following sets of input data:

- learning set,
- validation set,
- set of elementary classifiers.

As the setting parameters we should establish

- upper limit of number of algorithm’s cycles,
- population quantity,
- elite fraction size,
- probability of crossover and mutation,
- crossover model.

Initially we have to generate a set of members preserving all the constraints of the model. For each member of the population, a value of the fitness function is calculated according to (3). The learning set is exploited for that purpose. A certain number of members that are characterized by the highest fitness are taken from the population. The elite is put into descendant population, not being treated by mutation and crossover processes as well as selection procedure.

We use the traditional one-point crossover rule. The mutation operator assures that number of positive bits is equal n, therefore it chooses and it involves reversing value of two randomly chosen bits (one equals 1 and the second one 0 respectively).

Selection of individuals from population formed by merging descendant population and a set of individuals created by mutation and crossover. The probability of selection of a particular individual is proportional to the value of its fitness, according to the roulette wheel selection rule. A number of drawings is calculated so that a number of members of new population will be the same as previous population, including the elite that has been previously promoted. The procedure breaks the optimization process if deterioration of the result obtained by the best individual is observed in the course of given number of subsequent learning cycles.

4. EXPERIMENTAL INVESTIGATION

4.1. SET UP OF EXPERIMENT

The experiment was carried out in Matlab environment using PRTools toolbox [19] and our own software. The fusion block was realized according to majority voting rules and optimization task (3). In each iteration of the experiment different ω (from 0 to 1) and different number of individual classifiers (3, 5, 7, 9) from available classifiers' pool were used.

Table 1. Misclassification error and the cost value for each classifiers on different databases.

		Database			
		Hepatitis	**Liver disorders**	**Pima Indians diabetes**	**Heart disease**
Classifier1	Cost	3,00	21,81	19,61	280,50
	Error	0,37984	0,42258	0,34877	0,49022
Classifier2	Cost	10,27	29,08	20,61	4,00
	Error	0,43411	0,57097	0,56281	0,33088
Classifier3	Cost	17,54	46,21	43,39	302,37
	Error	0,43411	0,44516	0,34009	0,25735
Classifier4	Cost	24,81	53,48	43,39	206,67
	Error	0,40310	0,50645	0,40955	0,54044
Classifier5	Cost	9,27	21,81	3,00	208,70
	Error	0,41085	0,44839	0,35311	0,33824
Classifier6	Cost	16,54	31,67	25,78	197,47
	Error	0,34109	0,45806	0,38640	0,48021
Classifier7	Cost	16,54	24,40	42,39	204,67

	Error	0,30233	0,51290	0,34732	0,45956
Classifier8	Cost	17,54	29,08	21,61	506,00
	Error	0,40310	0,55161	0,35890	0,27941
Classifier9	Cost	22,81	24,40	25,78	392,87
	Error	0,51163	0,45806	0,55123	0,52041

Simple classifiers were realized as neural networks. To provide diversity of simple classifiers that allows their local competences to be exploited, only slightly undertrained networks has been used. The details of used neural nets are as follow:

- 5 neurons in hidden layer,
- sigmoidal transfer function,
- back propagation learning algorithm,
- number of neurons in last layer equals number of classes of given experiment.

It is important that the cost of using each simple classifier was different. To get such situation, each simple classifier was trained and was using in classification only the subset of the available vector values. The exact cost value and misclassification rate of each simple classifier was presented in Table 1.

To evaluate the experiment we used four databases from UCI Machine Learning Repository [2], which are described in Table 2.

Table 2. Databases' description.

	database	**number of**		
		examples	**attributes**	**classes**
1	Pima Indians diabetes	768	8	2
2	Heart disease	303	13	2
3	Hepatitis	155	19	2
4	Liver disorders	345	5	2

For each database the experiment was repeated 10 times. Results of all experiments are presented below in figures 1-4.

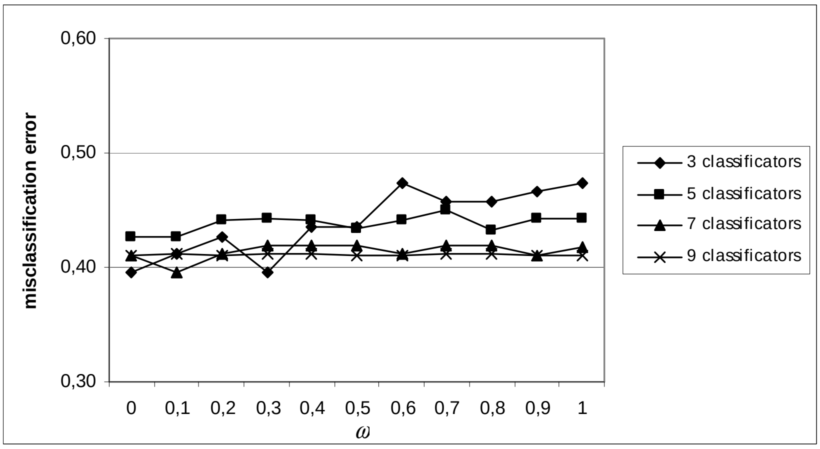

Fig. 1. Misclassification rate for “Hepatitis” problem.

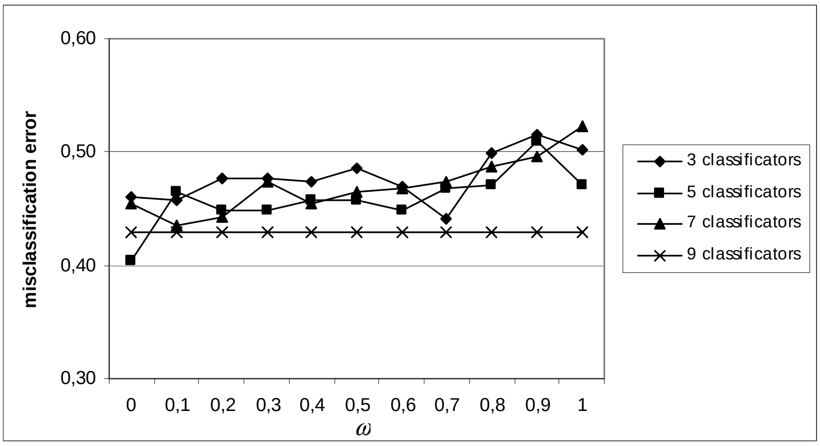

Fig. 2. Misclassification rate for "Liver disorders" problem.

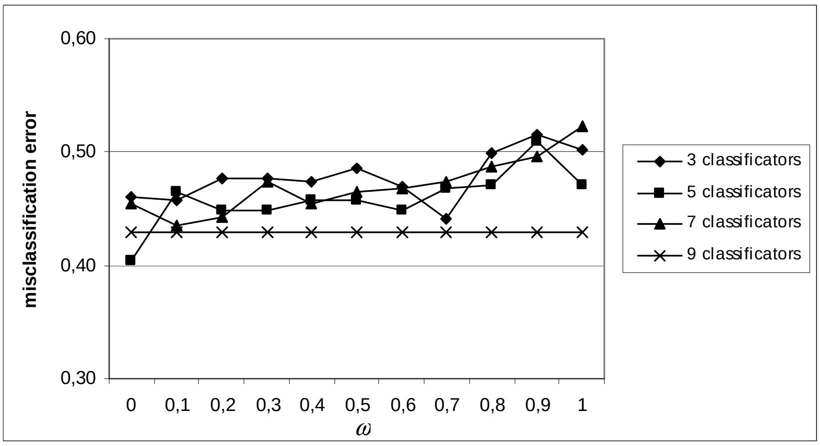

Fig. 3. Misclassification rate for "Heart disease" problem.

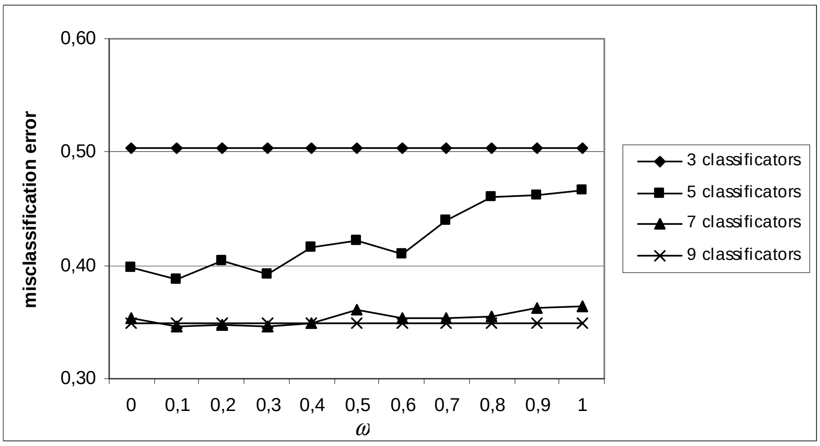

Fig. 4. Misclassification rate for "Pima Indians diabetes" problem.

4.2. EXPERIMENTAL RESULTS EVALUATION

The results of experiments show that in all repetition of experiments proposed algorithm got better results than the worst simple classifier from available pool. On the other hand obtained results were not as good as the results observed on the best classifier from the pool. In most cases when ω was growing, the misclassification error of proposed algorithm was increasing because cheaper individual classifiers were chosen. We can predict that cheaper classifiers did not have enough discriminated power and because of the pretty small set of the available features they were not differ each other (small diversity of a pool of individual classifiers). What is worth noting is that for big enough classifier committee the combined classifiers did not depend on ω.

5. CONCLUSION

The idea of a cost-sensitive combined classifier training was presented in this paper. The properties of the proposed concept were established during computer experiments carried out on four benchmark databases from the medical area. The results did not surprise us but we noted some interesting properties of the method under consideration. In our opinion presented idea of a cost-sensitive combined classifier is good direction in constructing real decision systems, especially for decision-aided systems where the cost of decision making plays the crucial role. Although we realize that the scope of computer experiments were limited and it is still a lot of work to do in this field.

6. ACKNOWLEDGEMENT

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

BIBLIOGRAPHY

[1] ALEXANDRE L.A., CAMPILHO A.C., KAMEL M., Combining Independent and Unbiased Classifiers Using Weighted Average, Proc. of the 15th Internat. Conf. on Pattern Recognition, Vol.2, 2000, pp. 495-498.

[2] ASUNCION A., NEWMAN D.J., UCI ML Repository, Irvine, CA: University of California, School of Information and Computer Science, 2007, http://www.ics.uci.edu/~mlearn/MLRepository.html.

[3] CHOW C.K., Statistical independence and threshold functions, IEEE Trans. on Electronic Computers, EC-16, 1965, pp. 66-68.

[4] DIETTERICH T.G., BAKIRI G., Solving multiclass learning problems via error-correcting output codes, Journal of Artificial Intelligence Research, 2, 1995, pp. 263-286.

[5] DUDA R.O., et al., Pattern Classification, Wiley-Interscience, 2001.

[6] GIACINTO G., Design Multiple Classifier Systems, PhD thesis, Universita Degli Studi di Salerno, 1998.

[7] GREINER R., GROVE A., ROTH D., Learning active classifiers, Proceedings of the 13th International Conference on Machine Learning, 1996, pp.207-215, 1996.

[8] HANSEN L.K., SALAMON P. , Neural Networks Ensembles, IEEE Trans. on PAMI, 12(10), 1990, pp. 993-1001.

[9] JAIN A.K., DUIN P.W., MAO J., Statistical Pattern Recognition: A Review, IEEE Trans. on PAMI, 22(1), 2000, pp. 4-37.

[10] KUNCHEVA L.I., Combining pattern classifiers: Methods and algorithms, Wiley, 2004.

[11] LIROV, Y., YUE, O.C., Automated network troubleshooting knowledge acquisition, Journal of Applied Intelligence, 1, 1991, pp. 121-132.

[12] MARCIALIS G.L., ROLI F., Fusion of Face Recognition Algorithms for Video-Based Surveillance Systems, in FORESTI G.L., REGAZZONI C., VARSHNEY P., (eds.), Multisensor Surveillance Systems: The Fusion Perspective, Kluwer Academic Pub., 2003.

[13] MICHALEWICZ Z., Genetics Algorithms + Data Structures = Evolutions Programs, Springer-Verlag, Berlin 1996.

[14] NUNEZ, M., Economic induction: A case study, Proceedings of the Third European Working Session on Learning EWSL-88, California: Morgan Kaufmann, 1998, pp. 139-145.

[15] NUNEZ, M., The use of background knowledge in decision tree induction, Machine Learning, 6, 1991, pp. 231-250.

[16] TAN M., SCHLIMMER J., Cost-sensitive concept learning of sensor use in approach and recognition, Proceedings of the Sixth International Workshop on Machine Learning ML-89, Ithaca, New York, 1989, pp. 392—395.

[17] TUMER, K., GHOSH, J., Analysis of Decision Boundaries in Linearly Combined Neural Classifiers, Pattern Recognition, 29, 1996, pp. 341–348.

[18] TURNEY P.D., Cost-sensitive classification: Empirical evaluation of a hybrid genetic decision tree induction algorithm, J. Artif. Intell. Res., 2, 1995, pp. 369-409.

[19] Van der HEIJDEN, F., DUIN, R.P.W., de RIDDER, D., TAX D.M.J., Classification, parameter estimation and state estimation - an engineering approach using Matlab, John Wiley and Sons, 2004.

[20] Van ERP M., VUURPIJL L.G., SCHOMAKER L.R.B., An overview and comparison of voting methods for pattern recognition, Proc. of IWFHR.8, Canada, 2002, pp. 195–200.

[21] VERDENIUS, F., A method for inductive cost optimization, Proceedings of the Fifth European Working Session on Learning EWSL-91, New York: Springer-Verlag, 1991, pp. 179-191.

[22] WOLPERT D.H., The supervised learning no-free-lunch theorems, Proceedings of the 6th Online World Conference on Soft Computing in Industrial Applications, 2001.