Expert Systems With Applications 46 (2016) 316–323

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Aggregation of classifiers ensemble using local discriminatory power and quantiles

Bartosz Swiderski [a], Stanislaw Osowski [b,c,*], Michal Kruk [a], Walid Barhoumi [d]

[a] Warsaw University of Life Sciences, The Faculty of Applied Informatics and Mathematics, Nowoursynowska 159, 02-796 Warsaw, Poland
[b] Warsaw University of Technology, 00-661 Warsaw, Koszykowa 75, Poland
[c] Military University of Technology, 01-476 Warsaw, Kaliskiego 2, Poland
[d] Intelligent Systems in Imaging and Artificial Vision, RIADI Laboratory, ISI 2 Street, Abou Rayhane Bayrouni, 2080 Ariana, Tunisia

ARTICLE INFO

Keywords:
Ensemble of classifiers
Dynamic classifier selection
Quantiles
Data mining
Machine learning
Melanoma

ABSTRACT

The paper presents a new approach to the dynamic classifier selection in an ensemble by applying the best suited classifier for the particular testing sample. It is based on the area under curve (AUC) of the receiver operating characteristic (ROC) of each classifier. To allow application of different types of classifiers in an ensemble and to reduce the influence of outliers, the quantile representation of the signals is used. The quantiles divide the ordered data into essentially equal-sized data subsets providing approximately uniform distribution of [0–1] support for each data point. In this way the recognition problem is less *sensitive to* the outliers, scales and noise contained in the input attributes. The numerical results presented for the chosen benchmark data-mining sets and for the data-set of images representing melanoma and non-melanoma skin lesions have shown high efficiency of the proposed approach and superiority to the existing methods.

1. Introduction

The combination of many classifiers in an ensemble is a well-known method of increasing the quality of recognition and classification tasks (Kuncheva, 2004; Omari & Figueiras-Vidal, 2015; Osowski, Markiewicz, & Tran Hoai, 2008; Parvin, Babouli, & Alinejad-Rokny, 2015; Xu, Krzyżak, & Suen, 1992). Each classifier, which relies its operation on different principle, may attain different degree of success in a specific application problem. Maybe none of them is perfect or as good as expected. Thus, there is a need to combine different solutions of classifiers, so that a better result could be obtained. Combining many trained networks together helps to integrate the knowledge acquired by the component classifiers and in this way to improve the accuracy of the results of final classification.

There are many different solutions to the integration problem. The usual approach relies on applying all classifiers from the ensemble to classify the testing patterns and on the basis of their results the final response is formed. Different static fusion strategies are applied in practice. Among the most often used is the voting principle organized in different ways, application of naive Bayes rule, Dempster–Shafer methods, Kullback–Leibler rule, meta-evolutionary ensemble, principal component analysis or application of additional integrating classifier (Haghighi, Vahedian, & Hadi, 2011; Kim, Stree, & Mencher, 2006; Kuncheva, 2004; Omari & Figueiras-Vidal, 2015; Osowski et al., 2008; Xu et al., 1992). Boosting, bagging, random subspace methods play a major part of such solutions (Efron & Tibshirani, 1993; Friedman, Hastie, & Tibshirani, 2000). These rules take into account all classifiers of an ensemble to perform the classification task and then exploit the statistics of their results to elaborate the final classification decision.

This paper applies different strategy, called in general dynamic classifier selection (DCS) (Britto, Sabourin, & Oliveira, 2014; Didaci, Giacinto, Roli, & Marcialis, 2005; Ko, Sabourin, & Britto, 2008; Parvin et al. 2015; Woods, Kegelmeyer, & Bowyer, 1997). The final classification of each testing sample is done by only one classifier from an ensemble, which is the best suited to the particular analyzed task. The best classifier is selected on the basis of its local discriminatory power in the neighborhood of the testing sample. Closely, we examine the generalization ability of all classifiers in the neighborhood of the testing sample. In computation of the discriminatory power of the classifier we assign higher weights to the analyzed observations which are closer to the actual testing sample. Selection of the best suited classifier is dependent on the distance of the testing sample to the samples used in learning. The selection is done by estimating the competence of the classifiers available in the pool on local regions of the feature space. In this way the classifier of the highest classification accuracy in the region is chosen.

* Corresponding author at: Warsaw University of Technology, Faculty of Electrical Engineering, Koszykowa 75, Warsaw, Poland. Tel.: +48 22 234 7235; fax: +48 22 234 5642.
E-mail addresses: swidersb@iem.pw.edu.pl, jbswiderski@wp.pl (B. Swiderski), sto@iem.pw.edu.pl (S. Osowski), michal_kruk@sggw.pl (M. Kruk), walid.barhoumi@laposte.net (W. Barhoumi).

http://dx.doi.org/10.1016/j.eswa.2015.10.038
0957-4174/

Thanks to this we are able to achieve the highest yield, since each classification task is performed by the classifier best suited to this particular task.

Comprehensive review of DCS is done in recent publications (Britto et al., 2014; Didaci et al., 2005; Ko et al., 2008; Parvin et al., 2015). The most important point in DCS is to select the most accurate classifier in the neighborhood of the analyzed sample. Different approaches are used in this task: the overall local accuracy, local class accuracy, a priori selection, a posteriori selection or k-nearest oracle (Didaci et al., 2005). All of them are done on the original data points. Another approach combines static ensemble with DCS (Parvin et al., 2015), by selecting classifiers based on clustering principle. The DCS has been also extended to selection of an ensemble for every test data point (Ko et al., 2008). These classical approaches to selection of the most accurate classifier suffer from such problems, as different ranges of output signals of used classifiers, influence of outliers and noise contaminating data or difficult choice of number of learning samples taken into account in the process of the best classifier selection.

Our approach avoids most of these problems by applying the quantile representation of data. The quantiles divide the ordered data into essentially equal-sized data subsets providing approximately uniform distribution of [0-1] support for each data point. Thanks to this the recognition problem is less sensitive to the outliers, scales and noise contained in the input attributes. Additionally, they form an ideal platform for cooperation of different types of classifiers arranged in an ensemble. Moreover, we propose novel way of choosing the best suited classifier for the particular testing sample. The choice is done on the basis of the area under curve (AUC) of the receiver operating characteristic (ROC) of each classifier.

The experiments performed on the benchmark problems and on the real task of recognition of melanoma from the non-melanoma lesions have shown very high efficiency of the proposed approach. In all cases the results of our method were better in the classification accuracy than the stand alone individual solutions.

The outline of the paper is as follows. Section 2 introduces the quantile representation of data. Section 3 presents the general description of the presented approach. Section 4 is devoted to the application of quantiles in classification of the data. The results of numerical experiments performed on the benchmark data are presented in Section 5. Section 6 is devoted to the real problem of melanoma recognition. The quality of solution is measured on the basis of area under ROC curve in all these experiments and the accuracy of class recognition. The last section presents the conclusions.

2. Quantile representation of data

In our approach the important role is fulfilled by the quantile representation of the data (Chu & Nakayama, 2010; Matlab, 2012). Quantiles are the points taken at the regular intervals from the cumulative distribution function (CDF) of a random variable. They mark the boundaries between consecutive subsets. Let us assume there is a given feature (variable) x of the particular values $x_1, x_2,\ldots, x_n$. The empirical cumulative distribution function is defined by the formula

$$F(x) = \frac{\#\{x_i : x_i \le x\}}{n} \tag{1}$$

for all $x \in R$. Formally, the quantile of order p is defined by:

$$q_p = \min\left\{x : F(x) \ge p\right\} \tag{2}$$

Roughly speaking, it means the quantile of the order p divides the ordered series of the random variable into two subsets in the proportions: p and $1 - p$.

For estimating a quantile representation we have used the Matlab function *tiedrank* (Matlab, 2012) applied in the Matlab notation as $(2 \times tiedrank(x) - 1)/(2 \times length(x))$. For example, let us consider the data in the ordered series of random variable as shown in the first column of the Table 1.

Table 1
The exemplary series of data (column 1) and the corresponding quantiles (column 2).

q_p	p
−3	0.0555
4	0.2222
4	0.2222
5	0.3889
100	0.5000
1001	0.72220
1001	0.7222
1001	0.7222
2000	0.9444

We get their quantile representation of the form expressed in column 2 of the table (variable p). Observe that irrespective of the distribution of the original series, the quantile representation is always uniform and is in the range [0, 1]. The observations, which are far from each other in original space (for example 1001 and 2000), may be very close in the quantile representation (0.7222 and 0.9444, respectively). It depends only on their positions in the ordered series.

The quantiles are useful measures because they are less sensitive to the fat-tailed distributions and outliers. At the same time they are well supported by the functions *quantile* and *tiedrank* of Matlab.

3. The proposed classification method – general description

Let us assume the data set $\mathbf{X}$ containing K observations, each characterized by N variables (input attributes). The observations are associated with the proper destination vector $\mathbf{d}$ representing classes to which the observations belong.

$$\mathbf{X} = \begin{bmatrix} x_{11} & \cdots & x_{1N} \\ \vdots & \ddots & \vdots \\ x_{K1} & \cdots & x_{KN} \end{bmatrix}, \quad \mathbf{d} = \begin{bmatrix} d_1 \\ \vdots \\ d_K \end{bmatrix} \tag{3}$$

Consider one testing observation denoted by $\mathbf{x_t}$ and its proper class represented by $d_t \in \{0, 1\}$. In further considerations we assume the binary classifiers. Assume M classifiers employed to solve the classification problem. Our task is to choose the classifier of the best generalization ability to recognize and classify the testing sample. The proposed procedure is as following.

First, apply the bootstrap strategy (Efron & Tibshirani, 1993; Friedman et al., 2000) to the data set ($\mathbf{X}$, $\mathbf{d}$) of K observations. A bootstrap set is created by sampling K instances uniformly from the original data (with replacement). This bootstrap set is split into the learning samples ($\mathbf{X}_L$, $\mathbf{d}_L$) containing 75% of data and validation set ($\mathbf{X}_V$, $\mathbf{d}_V$) of the remaining 25% samples.

To provide the proportional representation of classes in the sets, we first separate the samples of both classes. For each class, 75% of observations form the potential learning set and the remaining 25% the validation set. Then, we apply the bootstrap strategy for each of these four groups of data. The bootstrap selection is repeated as many times as is the number of observations in each subgroup. In the last step, we fuse the learning subsets of both classes, forming the final learning set and in the same way we fuse two validation subsets to form the final validation set. The learning set selected in this way is used as the learning base for all classifiers included in an ensemble. The learned classifiers are tested on the validation data set.

In the next step, we check the generalization ability of each member of an ensemble, paying the greatest attention to the samples placed in the neighborhood of the testing sample $\mathbf{x_t}$. This process is done using the validation set ($\mathbf{X}_V$, $\mathbf{d}_V$). We calculate the Euclidean distance of $\mathbf{x_t}$ to each sample of the validation set. The