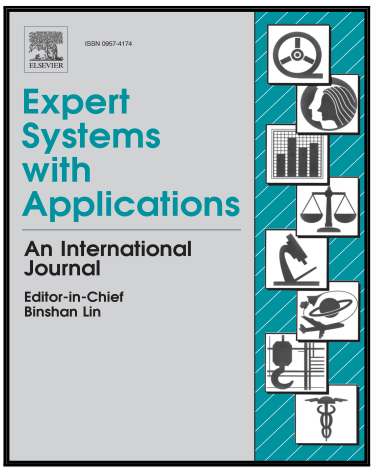


# Aggregation of classifiers ensemble using local discriminatory power and quantiles

Bartosz Swiderski , Stanislaw Osowski , Michal Kruk , Walid Barhoumi

PII: S0957-4174(15)00742-3
DOI: 10.1016/j.eswa.2015.10.038
Reference: ESWA 10368

To appear in: *Expert Systems With Applications*

Received date: 1 December 2014
Revised date: 25 August 2015
Accepted date: 28 October 2015

Please cite this article as: Bartosz Swiderski , Stanislaw Osowski , Michal Kruk , Walid Barhoumi , Aggregation of classifiers ensemble using local discriminatory power and quantiles, *Expert Systems With Applications* (2015), doi: 10.1016/j.eswa.2015.10.038

## Highlights

- We developed new method of integrating classifiers in an ensemble based on quantiles.
- We have shown superiority of our solution on the benchmark problems.
- We have applied this solution to recognition of melanoma and proved its superiority.

**Aggregation of Classifiers Ensemble Using Local Discriminatory Power and Quantiles**

**Corresponding author:**

**Stanislaw OSOWSKI**

Warsaw University of Technology, Faculty of Electrical Engineering, Koszykowa 75, Warsaw and Military University of Technology, Faculty of Electronics, Kaliskiego 2, 00-908 Warsaw, POLAND

email: sto@iem.pw.edu.pl

tel: +48-22-234-7235

fax: +48-22-234-5642

**Other authors:**

**Bartosz SWIDERSKI**

Warsaw University of Life Sciences, The Faculty of Applied Informatics and Mathematics, Nowoursynowska 159, 02-796 Warsaw, POLAND

Email: jbswiderski@wp.pl

**Michal KRUK**

Warsaw University of Life Sciences, The Faculty of Applied Informatics and Mathematics, Nowoursynowska 159, 02-796 Warsaw, POLAND

Email: michal_kruk@sggw.pl

**Walid BARHOUMI**

University of Manouba, Campus Universitaire de *Manouba*, TUNISIA

Email: walid.barhoumi@laposte.net.

**Bartosz SWIDERSKI[1], Stanislaw OSOWSKI[2,3], Michal KRUK[1], Walid BARHOUMI[4]**

[1]University of Life Sciences, swidersb@iem.pw.edu.pl

[2]Warsaw University of Technology, [3]Military University of Technology, sto@iem.pw.edu.pl

[4] University of Manouba, TUNISIA, walid.barhoumi@laposte.net

***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.*

**Keywords:** Ensemble of classifiers; dynamic classifier selection; quantiles; data mining; machine learning; melanoma.

## 1. Introduction

The combination of many classifiers in an ensemble is a well-known method of increasing the quality of recognition and classification tasks (Xu at al., 1992; Kuncheva, 2004; Osowski et al., 2008; Omari and Figueiras-Vidal, 2015;Parvin et al. 2015). 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 (Xu at al., 1992; Kuncheva, 2004; Osowski et al., 2008; Haghighi et al. 2011; Kim et al. 2006; Omari and Figueiras-Vidal, 2015). Boosting, bagging, random subspace methods play a major part of such solutions (Efron and Tibshirani, 1993; Friedman et al., 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) (Didaci et al., 2005; Britto et al., 2014; Parvin et al., 2015; Ko et al., 2008). 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. 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 (Didaci et al., 2005; Britto et al., 2014; Parvin et al., 2015; Ko et al., 2008). 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 and 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 $x_1$, $x_2$, …, $x_n$. The empirical cumulative distribution function is defined by the formula

$$F(x) = \frac{\#\{x_i : x_i \le x\}}{n} \qquad (2)$$

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

$$q_p = \min\{x : F(x) \ge p\} \qquad (3)$$

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**tiedrank*(x)-1)/(2**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 **X** containing *K* observations, each characterized by *N* variables (input attributes). The observations are associated with the proper destination vector **d** representing classes to which the observations belong.

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

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 and Tibshirani, 1993; Friedman et al., 2000) to the data set (**X**, **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 4 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 calculations are performed using the quantile representation of the samples (Chu and Nakayama, 2010; Matlab, 2012). The quantiles provide approximately uniform distribution of [0-1] support for each feature. Thanks to this we get higher resistance to the outliers and varying distributions or scales of the input attributes.

As a result, we get the distance of all validation samples of $\mathbf{X}_v$ to the testing sample $\mathbf{x}_t$. These distances will be associated with the weights. The closer is the distance the higher weight is associated with the particular validation sample. In this way the higher weights correspond to higher similarity of the testing sample $\boldsymbol{x}_t$ to the actual sample from the set $\mathbf{X}_v$ .

In the experiments we have limited the highest weight to three and the lowest to one. All other weights have been distributed linearly between these two extreme values.

In the following step we create the weighted receiver operating characteristic (ROC) curve representing the relation between the true positive rate and false positive rate for the samples from the validation data set (Tan et al., 2006). In forming ROC curve we take into account the weights associated with the samples, duplicating them proportionally to their weight value. Each sample of weight *w* is duplicated *w* times in the process of ROC forming. In this way we include the information of proximity of the tested sample to the samples of the validation set.

The ability of the classifier to proper classifying the sample $\mathbf{x}_t$ is measured by the area under the weighted ROC curve (AUC) created for the validation set. The AUC reflects the probability of correct recognition of classes. It was shown for the balanced classes (Ali and Deserno, 2012) that when the AUC value is in the range (0.9 – 1) the accuracy of class recognition is excellent. When this value is within the range (0.8 – 0.9) the accuracy is good. The value of AUC below 0.8 corresponds to fair or poor accuracy. Therefore, the closer this area to the value of 1, the better is the classifier.

We perform the classification steps, i.e. boosting and sampling of the validation set and learning and testing procedures *m* times (in this research, *m*=100) for each classifier from the ensemble. Taking the mean value of the weighted AUC at many bootstrap resampling, we obtain an objective assessment of the discriminatory power of the particular classifier for the sample $\boldsymbol{x_t}$ under recognition. In each case the classifier chosen to do the classification task of the testing sample is the one, which has the highest value of the average weighted AUC obtained in all trials. This procedure can be summarized by the following pseudo code.

for $j$=1:$m$
begin
-draw bootstrap samples: $X_L$ and $X_V$ for $i^{th}$ classifier, $i$ = 1,2, …, $M$;
-learn $i$th classifier on $X_L$ and test on $X_V$ and $x_t$;
-transform the continuous outputs of the classifiers for testing data into ordered quantile representation;
-calculate the distance between the quantile representation of $X_V$ and $x_t$;
-associate higher weights to the $X_V$ samples which are closer to $x_t$;
-create the weighted ROC for the samples of $X_V$ and -calculate the weighted AUC for each classifier;
end
calculate the average weighted AUC for each classifier and choose the classifier of the highest value of AUC for the analyzed sample $x_t$. .

Let us consider now many testing samples $\mathbf{x}_t$. While repeating the presented above procedure on this set, we select the best classifier for each testing sample. However, the chosen classifiers may produce the continuous output signals in different ranges of values. For example, the logistic regression classifier produces the output signal in the range [0, 1], while the SVM classifier signal $y(\mathbf{x})$ may take any real value. The important point is to find the common platform for all classifiers, in which all output signals are in the unified range [0, 1].

We solve this problem using again the quantile representation of the output signals. The chosen (best) classifier is learnt and tested on the whole available data set X and then tested on the actual testing sample $\mathbf{x}_t$. In each case, we get the continuous output signals from the classifier (before applying the sign operation). The set of the output signals on the data set X and the output of the testing sample are converted to the common quantile representation in the normalized range [0, 1]. The output signal of the testing sample is associated with the proper order of quantiles in the combined set. Thanks to this we avoid the problem of diversity of types of the output signals of the applied classifiers, since the quantile representation is always in the range [0, 1], irrespective of the applied classifier.

Each sample $\mathbf{x}_t$ from the testing set is classified by the locally best classifier according to its quantile order. For example, the output signal of the best classifier for the $i^{\text{th}}$ sample is equal to 0.8674 and for the $j^{\text{th}}$ one is equal to 0.2347. The first result is closer to one, so it will be associated with the class one. The second is closer to zero, hence it will be associated with the alternative class. The point discriminating two classes may be set on the level chosen by the user, however, the typical threshold is 0.5 in the case of balanced classes.

## 4. Application of quantiles in classification

In the following subsections we introduce the details of our approach, starting from application of quantiles in selecting the best classifiers up to formation of weighted AUC of the ROC characteristics.

### 4.1 Weighted AUC formation using quantile representation

The quantile representation is used by us to create the weighted AUC for the testing data. To explain the details of this approach let us take the exemplary validation data set $\mathbf{X}_V$ and testing vector $\mathbf{x}_t$ in the form shown in Table 2, where each row represents observation and column the variable. The quantile representation is built for each column separately.

On the basis of the quantile representation we calculate the Euclidean distance between $\mathbf{x_t}$ and each row of $\mathbf{X}_V$. Next, each row of $\mathbf{X}_V$ is weighted in a reverse order. The vector closest

to $\mathbf{x}_t$ is associated with the highest weight (the value 3 was assumed in experiments) and the farthest one with the smallest weight (the value of 1). For intermediate points, the linear weighting between 3 and 1 was applied.

Table 2
Transformation of observations into their quantile representations.

| | Original data | | | | | Quantile representation | | | |
|---|---|---|---|---|---|---|---|---|---|
| Observations | $\mathbf{X}_V$ | 1.1 | -100 | 200 | 324 | 0.3333 | 0.1666 | 0.8333 | 0.8333 |
| | | -2.3 | 200 | 4.5 | 52 | 0.1666 | 0.5000 | 0.2777 | 0.5000 |
| | | 1.1 | -400 | -3 | 5 | 0.3333 | 0.0555 | 0.0555 | 0.1666 |
| | | 4.5 | 200 | 24 | 5 | 0.7222 | 0.5000 | 0.5000 | 0.1666 |
| | | 5.7 | 300 | 354 | 2345 | 0.8333 | 0.7777 | 0.9444 | 0.9444 |
| | | 8.9 | 200 | 31 | 55 | 0.9444 | 0.5000 | 0.6111 | 0.6111 |
| | | -4.3 | 300 | 53 | 35 | 0.0555 | 0.7777 | 0.7222 | 0.3888 |
| | | 3.1 | 150 | 15 | 5 | 0.6111 | 0.2777 | 0.3888 | 0.1666 |
| | $\mathbf{x}_t$ | 2.5 | 500 | 2 | 100 | 0.5000 | 0.9444 | 0.1666 | 0.7222 |

Table 3 presents the original Euclidean distances between the vectors (rows) of $\mathbf{X}_V$ and vector $\mathbf{x}_t$ in the example (all in quantile representations). The column denoted by "Weights" represents the weights associated with the succeeding samples (rows) of observations.

Table 3
The original Euclidean distances between $\mathbf{x}_t$ and observations from $\mathbf{X}_v$ and the corresponding weights associated with the observations.

| Observations | Original distance | Weight |
|---|---|---|
| 1 | 1.04379 | 1.1020 |
| 2 | 0.60858 | 3.0000 |
| 3 | 1.06719 | 1.0000 |
| 4 | 0.81650 | 2.0933 |
| 5 | 0.89062 | 1.7700 |
| 6 | 0.77778 | 2.2621 |
| 7 | 0.80316 | 2.1514 |
| 8 | 0.90267 | 1.7175 |

Then, we form the weighted ROC before computing the weighted AUC. In the process of calculating the true positive and false positive cases in ROC creation, we take into account not the real number of samples satisfying the appropriate conditions, but the sum of weights associated with these samples.

### 4.2 Selecting the best classifier for testing samples using quantiles

Applying the procedure described above, we can select the classifier, which is the most suitable for the classification of the particular tested sample $\mathbf{x}_t$ . This is the classifier of the highest value of the mean of the weighted AUC, built on the basis of the validation set. Different type classifiers may be found as the best for the particular testing samples. Therefore, we should take into account, that their output signals may be placed in different ranges of values. To create common platform, we have to transform their outputs to the unified range of [0, 1]. To solve the problem, we have used once again the quantile approach.

In this approach the best classifier is learnt on the whole data set $\mathbf{X}$ and then tested on both: learning data (output set $\mathbf{Y}$) and the testing sample $\mathbf{x}_t$ (output $\mathbf{y}(\mathbf{x}_t)$). As a result we get the output signals of the classifier in the form of common set [$\mathbf{Y}$; $\mathbf{y}(\mathbf{x}_t)$]. They are represented as the continuous signals (before application of sign function).

This common set of signals is converted to the quantile representation, including the tested signal $y(\mathbf{x}_t)$. In this way we get automatically the quantile value associated with $y(\mathbf{x}_t)$. The final class recognition of the sample $\mathbf{x}_t$ depends on this value. If it is above the threshold we recognize sample as the class one. In the opposite case $\mathbf{x}_t$ is associated with the alternative class.

## 5. Empirical results of experiments on benchmark data

The proposed approach to classification problem was first tested on the 2-class benchmark problems taken from (UCI, 2014). The following benchmark problems have been considered: breast cancer, flare, soybean, credit cards, glass, heart, diabetes and horse. The data values of all problems were prepared in the same way. The nominal and discrete variables were coded using weights of evidence (WOE) technique (Bonham-Carter et al., 1989; Zhixiao, 2013). The basic characteristics of data sets are presented in Table 4.

The set of classifiers taking part in experiments was composed of the following units:

1. Support Vector Machine of Gaussian kernel (SVMG) of sigma=1 and regularization factor $C$=1 (Scholkopf and Smola, 2002).
2. SVM of linear kernel (SVML) and $C$=1 (Scholkopf and Smola, 2002).

3. Decision tree (DT) (Rokach, 2008).
4. Multilayer perceptron (MLP) of 3 neurons in hidden layer and *tansig* as an activation function (Haykin, 2000).
5. Fuzzy K-nearest neighbor classifier (FKNN) of 5 neighbors (Keller et al., 1985).
6. Logistic regression based on generalized linear model (LR) with link logit and binomial distribution (Tomassi et al., 2006).
7. The proposed ensemble of the above classifiers, applying local discriminatory power and quantiles (denoted further by LDPQ).
8. The comparative approach to DCS based on local accuracy estimation (DCS-LA) (Didaci et al., 2005)

Table 4
The characterization of the 2-class benchmark data.

| Name of problem | Dimension of input vector | Population of one class | Population of second class |
|---|---|---|---|
| Breast Cancer | 9 | 458 | 241 |
| Flare | 10 | 884 | 112 |
| Soybean | 35 | 40 | 40 |
| Credit cards | 15 | 383 | 307 |
| Glass | 9 | 70 | 76 |
| Heart | 13 | 441 | 265 |
| Diabetes | 8 | 500 | 268 |
| Horse | 20 | 224 | 88 |

The internal parameters of the individual classifiers were not optimized. Our aim in this part of experiments is to show that application of the proposed approach allows improvs the final results of class recognition. The results of classification in the form of AUC values for each benchmark problem are presented in Table 5.

The succeeding columns from 1 to 6 correspond to the results of application of the particular classifier, learned in a classical way. Column 7 depicts the results of application of our ensemble strategy of classification (LDPQ). Last column presents the comparative results of application of DCS with the local accuracy estimation "a posteriori" (DCS-LA), the alternative solution of dynamic classifier selection presented in the paper (Didaci et. al.,

2005). All results correspond to the application of leave-one-out procedure (Haykin, 2000). By the bold we denote the best result of classification.

Note, that the final solution selected on the basis of set of six predefined classifiers for each observation has been chosen by the dynamic selection approach (two last columns). With the fixed parameters of the classifiers we were able to improve the classification results almost in all cases by using our approach (LDPQ). The only exceptions were the credit card problem, where the linear SVM classifier was better and soybean problem, where DCS-LA method was slightly better.

Table 5
The values of AUC in the benchmark class recognition problems

| Benchmark problem | Classification system | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | SVMG | SVML | DT | MLP | FKNN | LR | LDPQ | DCS-LA |
| Breast Cancer | 0.985 | 0.991 | 0.969 | 0.985 | 0.985 | 0.991 | **0.992** | 0.984 |
| Flare | 0.529 | 0.494 | 0.535 | 0.657 | 0.521 | 0.673 | **0.693** | 0.585 |
| Soybean | 0.970 | 0.999 | 0.986 | 0.967 | 0.961 | 0.986 | 0.999 | **1.000** |
| Credit cards | 0.906 | **0.938** | 0.904 | 0.924 | 0.904 | 0.931 | 0.934 | 0.912 |
| Glass | 0.924 | 0.730 | 0.777 | 0.738 | 0.921 | 0.716 | **0.950** | 0.835 |
| Heart | 0.788 | 0.862 | 0.767 | 0.848 | 0.805 | 0.867 | **0.871** | 0.843 |
| Diabetes | 0.779 | 0.833 | 0.726 | 0.813 | 0.764 | 0.835 | **0.838** | 0.803 |
| Horse | 0.634 | 0.680 | 0.740 | 0.721 | 0.762 | 0.811 | **0.816** | 0.751 |

Table 6 presents the percentage of observations for which the particular classifiers were chosen as the best in LDPQ. In three cases (breast cancer, credit cards and heart) our system has chosen the solution of only one type of classifier. In the remaining cases, different classifiers have been selected for doing the classifications task. Observe that applying different system of interpreting the output signals (ordinary numerical values of individual classifiers and quantile interpretation of results in the proposed ensemble) has resulted in slight differences even in the case when only one classifier was selected for recognition of all samples.

Table 6

The percentage of observations for which the particular classifiers were chosen as the best.

| Benchmark problem | Classifiers | | | | | |
|---|---|---|---|---|---|---|
| | SVMG | SVML | DT | MLP | FKNN | LR |
| Breast Cancer | 0,00% | 0,00% | 100,00% | 0,00% | 0,00% | 0,00% |
| Flare | 0,60% | 97,69% | 1,71% | 0,00% | 0,00% | 0,00% |
| Soybean | 0,00% | 0,00% | 0,00% | 7,50% | 92,50% | 0,00% |
| Credit cards | 0,00% | 100,00% | 0,00% | 0,00% | 0,00% | 0,00% |
| Glass | 0,00% | 69,87% | 0,00% | 4,79% | 0,00% | 25,34% |
| Heart | 0,00% | 100,00% | 0,00% | 0,00% | 0,00% | 0,00% |
| Diabetes | 0,00% | 96,87% | 3,13% | 0,00% | 0,00% | 0,00% |
| Horse | 23,08% | 44,23% | 18,27% | 14,42% | 0,00% | 0,00% |

The presented dynamic classifier selection approach to the benchmark problems has been also compared based on the accuracy of classification. Once again the local accuracy (DCS-LA) approach applying "a posteriori" estimation has been used (Didaci et. al., 2005) for comparative analysis. Table 7 presents the numerical results of such comparison.

Table 7

The comparison of accuracy of class recognition in benchmark problems

| | SVMG | SVML | DT | MLP | FKNN | LR | LDPQ | DCS-LA |
|---|---|---|---|---|---|---|---|---|
| Breast Cancer | 0,96 | 0,96 | 0,95 | 0,96 | 0,97 | 0,97 | **0,971** | 0,961 |
| Flare | 0,77 | 0,18 | 0,65 | 0,60 | 0,52 | 0,67 | 0,784 | **0,883** |
| Soybean | 0,95 | 0,98 | 0,98 | 0,92 | 0,92 | 0,97 | **0,992** | 0,951 |
| Credit cards | 0,86 | 0,88 | 0,86 | 0,86 | 0,86 | 0,87 | **0,881** | 0,852 |
| Glass | 0,86 | 0,73 | 0,77 | 0,72 | 0,84 | 0,68 | **0,923** | 0,810 |
| Heart | 0,74 | 0,77 | 0,76 | 0,78 | 0,71 | 0,77 | **0,784** | 0,763 |
| Diabetes | 0,72 | 0,74 | 0,72 | 0,73 | 0,71 | 0,74 | **0,772** | 0,731 |
| Horse | 0,74 | 0,66 | 0,72 | 0,66 | 0,69 | 0,73 | **0,754** | 0,740 |

The results presented in columns LDPQ are superior to the DCS-LA fusion method. Only in one case (Flare) the DCS-LA approach was better.

The LDPQ approach presented by us for benchmark problems was found superior also to the static fusion of an ensemble. For example the same breast cancer problem solution using

the ensemble of classifiers was considered in (Haghighi et al. 2011). The authors have reported the following accuracy rates: 96.5% in majority voting, 96.25% in Dempster-Shafer method, 96.5% in decision template and 96.85% in extended decision template approach. The accuracy of class recognition in our method was 97.1% and AUC measure of 0.992.

The same problems of credit cards and diabetes have been considered in (Kim et al. 2006) by using the meta-evolutionary approach to fusion. In the case of credit cards the reported accuracy was 86.4%. Our result for the same data set was 88.1% and it corresponds to the AUC equal 9.935. The two classes in diabetes problem have been recognized in the same paper with the accuracy of 76.8%, while our result is 77.2% and AUC=0.838.

Our results are also superior to these obtained by using the specialized individual methods (SVM, naïve Bayes, decision tree) applied to the same benchmark problems. For example, the best breast cancer and credit cards recognition results presented in (Huang and Ling, 2005) show AUC=0.973 and accuracy 96.5% for breast cancer and AUC=0.904, and accuracy=86.5% for credit cards.

## 6. Application of proposed approach to melanoma recognition

The proposed solution was applied for the real-world problem of recognition of melanoma and non-melanoma lesions of the skin. Melanoma belongs to the most dangerous human skin disease (Ganster et al., 2001;Zagrouba and Barhoumi, 2004). Early recognized melanoma changes allow the patient to recover completely. Therefore, the early diagnosis of lesion image of the skin is a crucial issue for dermatologists. Lesions can vary in color, saturation, shape and size. Fig. 1 shows examples of two images of melanoma and two of non-melanoma lesions.

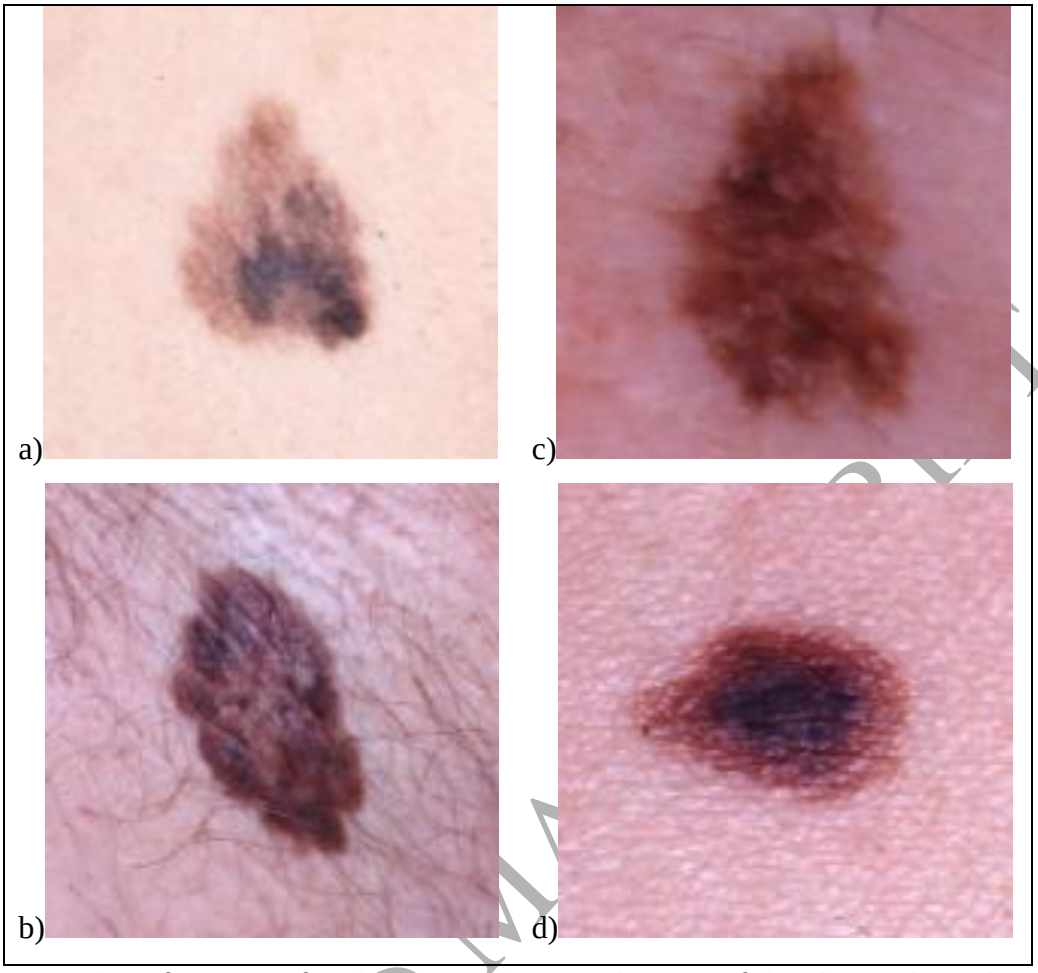


Fig. 1. Examples of images of melanoma and non-melanoma of the skin: a, b) melanoma, c, d) non-melanona lesions.

There are many approaches to the automatic melanoma recognition. Different techniques of image preprocessing and classification methods have been proposed up to date (Ganster et al., 2001; Zagrouba and Barhoumi, 2004; Abbas et al., 2013; Aswin et al., 2013, Sheha et a., 2012). However, the obtained accuracy is still not satisfactory and can be improved by applying more sophisticated approaches to the classification.

The preprocessing of the acquired images to get their characterization in a numerical form (diagnostic features) includes such steps as segmentation of the lesions from the surrounding skin, filtering for removing the noise, and finally their transformation into diagnostic features. In this paper, we have applied the procedure proposed in the papers (Zagrouba and Barhoumi, 2004; Abbas et al., 2013). As a result, we got the dermatoscopic features applying the ABCD descriptions (A- asymmetry, B – border irregularity, C – color variation, D – diameter) of the skin lesion. These features form the input attributes to the classifiers.

Nine parameters based on ABCD description have been defined and applied in experiments. They include: the asymmetry index, lengthening index of the lesion, compactness index, border irregularity index, mean and variance of the gradient magnitude of edge abruptness, color homogeneity, correlation between geometry and photometry and fractal dimension (Zagrouba and Barhoumi, 2004).

The numerical experiments have been performed on a set of 200 RGB images, where 120 images represented the melanoma and the rest the non-melanoma lesions. All images were of 150×150 pixels size with the spatial resolution of 0.234mm×0.234mm per pixel. The images used BMP format coded in 24 bits representing three colors: red, green and blue. Each image has one lesion region located near the center. The lesion changes are surrounded by the normal skin of variable hue.

The images have been transformed to nine features as indicated above. These features form the input attributes applied to the ensemble of 6 classifiers, defined earlier in this paper. The parameter values of the classifiers were the same as in the benchmark data. The individual classifiers (from 1 to 6) have been trained and tested in a classical mode without using the concept of quantiles. The experiments have been performed in the cross-validation leave-one-out mode (Haykin, 2000). We compare the results of application of individual classifiers working in a classical mode to the results of their ensemble integrated by using our approach.

The numerical results for each individual classifier and the applied ensemble, as well as DCS-LA approach are presented in Table 8. The second row (denoted as AUC) represents the area under ROC curve. The third row depicts the percentage of cases (*pc*) for which the particular classifier was selected as the best in an ensemble.

It is evident that the quality of melanoma recognition performed by the individual classifiers arranged in a classical way (one classifier recognizes all samples) is poor. However, organizing them in an ensemble integrated by our approach has resulted in a significant improvement. The AUC for the best individual classifier was 0.665. After LDPD integration this value was increased to 0.924. The improvement was possible thanks to the fact that for each testing sample the best performing classifier was used. As a result of such organization of classification almost all classifiers took part in the recognition process. The third row of the table confirms this fact. Once again our approach to the dynamic classifier selection performed better than the comparative DCS-LA.

Table 8

The averaged results of AUC in recognition of melanoma and non-melanoma lesions.

| Classifiers | SVMG | SVML | DT | MLP | FKNN | LR | LDPQ | DCS-LA |
|---|---|---|---|---|---|---|---|---|
| AUC | 0,665 | 0,303 | 0,567 | 0,653 | 0,632 | 0,640 | 0,924 | 0.671 |
| *pc* | 13% | 0 | 30% | 41% | 1% | 15% | | |

Fig. 2 presents the exemplary ROC curve in one of the cross-validation experiments. The area under curve is the assumed measure of the quality of the applied classification approach. It is evident, that the best result corresponds to LDPQ. The average sensitivity of this method calculated over 100 trials was above 93%.

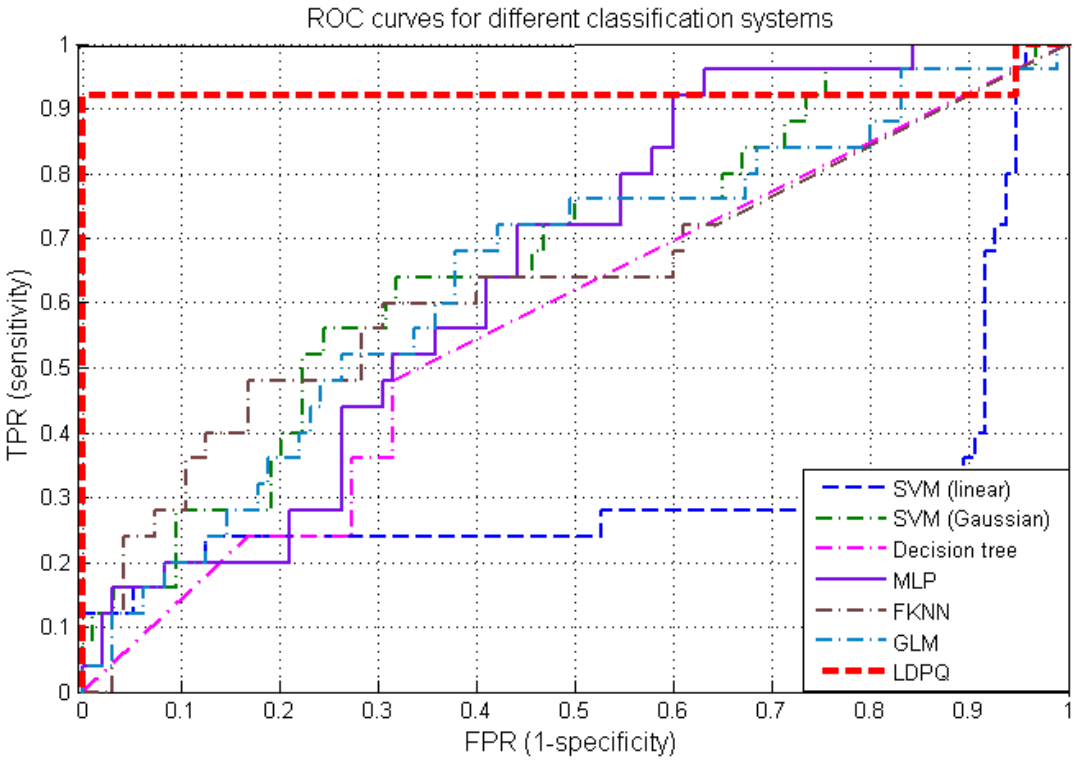


Fig. 2 The exemplary ROC curves obtained in recognition of melanoma cases by using different classification systems.

Table 9 presents the accuracy of melanoma recognition by the individual classifiers and their ensemble aggregated using our approach and the local accuracy method presented in (Didaci et al., 2005). In this particular application the advantage of our method is evident.

Table 9

The comparison of accuracy of class recognition for melanoma

| SVMG | SVML | DT | MLP | FKNN | LR | LDPQ | DCS-LA |
|---|---|---|---|---|---|---|---|

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 0,16 | 0,67 | 0,64 | 0,49 | 0,75 | 0,63 | 0,98 | 0,80 |

The obtained results using the proposed dynamic ensemble are also well compared to the results reported in recent publications. For example the sensitivity declared in the paper (Ganster et al., 2001) was 87%, 75.1% in (Zagrouba and Barhoumi, 2004), 88.2%, in (Abbas et al., 2013), 84%, in (Aswin et al., 2013) and from 70.5% to 92.3% in (Sheha et al., 2012). Our sensitivity result of 93% is among the best. Moreover, it should be noted that we applied the classifiers of the standard (not optimized) parameters. Our main task in the paper was to show that lower quality classifiers combined in an ensemble integrated by using our approach are able to achieve very good accuracy of performance.

## 7. Conclusions

The paper has presented the novel approach to the dynamic classifier selection in the integration of classifiers in an ensemble, based on the local discriminatory power and quantile representation of data. Instead of taking into account the results of all classifiers, only one, the best suited to the particular task, is chosen. Thanks to this the final result is never decreased by the least efficient classifier and the overall statistical accuracy and sensitivity of the class recognition is enhanced.

The most important difference to the existing methods of DCS (Didaci et al. 2005; Britto et al., 2014; Parvin et al., 2015; Ko et al., 2008) is the way of choosing the best suited classifier for the testing sample. This choice is done on the basis of the area under curve of the receiver operating characteristic of each classifier. AUC is recently regarded as the most objective way of comparing the quality of classifiers (Tan et al., 2006). In this way the choice is always optimal and at the same time insensitive to the number of samples taken into account in the neighborhood of the testing sample.

Application of the quantile representation of data has allowed avoiding many common problems in pattern recognition, such as sensitivity to the outliers and the noise or difference of output ranges of individual classifiers in an ensemble. The quantile method allows reducing the influence of the noise contaminating the input data. Additionally, it forms an ideal platform for cooperation of different types of classifiers arranged in an ensemble.

The statistical results performed on the benchmark problems have shown the superiority of our approach to the dynamic classifier selection over the already existing techniques of DCS. Moreover, our dynamic ensemble scheme performed better than other investigated static fusion methods. Very good results have been obtained for demanding task of melanoma

recognition. The most objective measure of the classification quality (AUC) achieved by us in this case has assumed the value of 0.924, much larger than the values of AUC corresponding to individual classifiers.

The benefits of our method may be limited when small amount of training data is available, or when the classification accuracy of individual classifiers is sufficiently high. In such cases the traditional approach to fusion of ensemble may be competitive. Some questions and problems regarding future practical applications in pattern recognition problems may arrive also for large number of classes. Higher number of recognized classes leads to more complex pattern recognition task, which results in increasing the computational cost. There will be the need for optimizing the computation algorithm, leading to acceleration of the problem solution. However, in these problems DCS approach to classification represents great potential, since such tasks are usually too complex for most individual classifiers.

It is important also to continue study of the influence of the choice of classifiers forming an ensemble on the performance quality of DCS, especially in the high dimensionality complex classification problems. Another problem worth of study is extension of our dynamic classifier selection to dynamic ensemble selection, in which best ensemble of classifier will be chosen for each testing sample. The additional direction of study is the extension of the method to the regression problems, through the choice of the best individual regressors minimizing the objective function in the neighborhood of the testing sample.

**ACKNOWLEDGMENT**

This work was supported by The National Centre for Research and Development of Poland under the grant which is being realized in years 2015-2018.

REFERENCES

Abbas Q., Emre Celebi M., Fondo I., Ahmad W. (2013). Melanoma recognition framework based on expert definition of ABCD for dermoscopic images. Skin Research and Technology, 19, 93-102.

Ali A.R, Deserno T.M, (2012). A systematic review of automatic melanoma detection in dermatoscopic images and its ground truth data. Proc. of SPIE, 8318 83181I-1, 1-12.

Aswin R.B, Jaleel A., Salim S. (2013). Implementation of ANN classifier using Matlab for skin cancer detection. International Journal of Computer Science and Mobile Computing, ICMIC13, 87 – 94.

Bonham-Carter G.F., Agterberg F.P., Wright D.F. (1989). Weights of evidence modeling: a new approach to mapping mineral potential (in Agterberg F.P., Bonham-Carter G.F. Eds. Statistical applications in the earth sciences, Geological Survey of Canada, Montreal, pp. 171-183.

Britto A.S., Sabourin R., Oliveira L. (2014). Dynamic selection of classifiers—A comprehensive review. Pattern Recognition 47 3665-3680.

Chu F., Nakayama M.K. (2010). Confidence intervals for quantiles when applying variance-reduction techniques, 8th International Workshop on Rare Event Simulation, 1-10, (conference proceedings).

Didaci L. Giacinto G., Roli F., Marcialis G.L. (2005). A study on the performances of dynamic classifier selection based on local accuracy estimation. Pattern Recognition, 38(11) 2188-2191

Efron B. Tibshirani R. (1993). *An introduction to bootstrap*. London: Chapman and Hall.

Friedman J., Hastie T., Tibshirani R. (2000). Additive logistic regression: a statistical view of boosting. The Annals of Statistics, 28, 337-407.

Ganster H., Pinz A., Röhrer R., Wildling E., Binder M., Kittler H. (2001). Automated melanoma recognition. IEEE Transactions on Medical Imaging, 20, 233-239.

Haghighi M.S., Vahedian A.Y., Hadi S. (2011). Extended decision template presentation for combining classifiers. Expert Systems with Applications, 38(7), 8414-8418.

Haykin S. (2000). *Neural networks, a comprehensive foundation*, London: Macmillan.

Huang J., Ling C. (2005).Using AUC and accuracy in evaluating learning algorithms, IEEE Trans. on Knowledge and Data Engineering, 17(3), 299-310.

Keller J., Gray M., Givens J. (1985). A fuzzy K-nearest neighbour algorithm. IEEE Trans. Systems, Man and Cybernetics, 15, 580-585.

Kim Y.S, Stree W.N., Mencher F. (2006). Optimal ensemble construction via meta-evolutionary ensembles. Expert Systems with Applications, 30(4), 705-714.

Ko A. H., Sabourin R., Britto A.S. (2008). From dynamic classifier selection to dynamic ensemble selection. Patteren recognition, 41(5) 1718-1731.

Kuncheva L. (2004). *Combining pattern classifiers: methods and algorithms*. New York: Wiley.

Matlab user manual (2012). Natick: MathWorks.

Omari A., Figueiras A.R. (2015) Post-aggregation of classifier ensembles. Information Fusion 26 96-102.

Osowski S., Markiewicz T., Tran Hoai L. (2008). Recognition and classification system of arrhythmia using ensemble of neural networks. Measurement, 41, 610-617.

Parvin H., Babouli M.M, Alinejad-Rokny H. (2015). Proposing a classifier ensemble framework based on classifier selection and decision tree. Eng. Appl. of Artif. Intel. 37 34-42.

Rokach L. (2008). *Data mining with decision trees*, Singapore: World Scientific.

Schölkopf B., Smola A. (2002). *Learning with kernels*. Cambridge MA.: MIT Press.

Sheha M.A., Mabrouk M.S., Sharawy A. (2012). Automatic detection of melanoma skin cancer using texture analysis. International Journal of Computer Applications, 42, 22-26.

Tan P.N., Steinbach M., Kumar V. (2006). *Introduction to data mining*, Boston: Pearson Education Inc.

Tommasi T., La Torre E., Caputo B. (2006). Melanoma recognition using representative and discriminative kernel classifers. International Workshop on Computer Vision Applications for Medical Image Analysis, Graz, 1-12 (conference proceedings).

UCI Benchmarks of machine learning repository (2014).

Woods K., Kegelmeyer W.P., Bowyer K. (1997). Combination of multiple classifiers using local accuracy estimates. IEEE Trans. PAMI, 19(4) 405-410.

Xu L., Krzyżak A., Suen C.Y. (1992). Methods of combining multiple classifiers and their applications to handwriting recognition. IEEE Trans. Systems, Man and Cybernetics, 22, 418-434.

Zagrouba E., Barhoumi W. (2004). A preliminary approach for the automated recognition for malignant melanoma. Image Analysis and Stereology, 23(2) 121—135.

Zhixiao L.A. (2013). Variable reduction in SAS by using weights of evidence and information value. SAS Forum, http://support.sas.com/resources/papers/proceedings13/095-2013.pdf.