machine learning, classifier competence
multiple classifier system,
dynamic competence threshold

Maciej KRYSMANN[1], Marek KURZYŃSKI[1]

ON NEW METHODS OF DYNAMIC ENSEMBLE SELECTION BASED ON RANDOMIZED REFERENCE CLASSIFIER

In the paper two dynamic ensemble selection (DES) systems are proposed. Both systems are based on a probabilistic model and utilize the concept of Randomized Reference Classifier (RRC) to determine the competence function of base classifiers. In the first system in the selection procedure of base classifiers the dynamic threshold of competence is applied. In the second DES system, selected classifiers are combined using weighted majority voting rule with continuous-valued outputs, where the weights are equal to the class-dependent competences. The performance of proposed MCSs were tested and compared against DES system with better-than-random selection rule using eleven databases taken from the UCI Machine Learning Repository. The experimental results clearly show the effectiveness of the proposed methods.

1. INTRODUCTION

Nowadays, dynamic ensemble selection (DES) methods are strongly developed as an effective approach to the construction of multiple classifier systems ([3, 6, 12]). In the DES scheme, first an ensemble of base classifiers is dynamically selected from the entire set (pool) of base classifiers and then the members of ensemble are combined by a fusion method (usually weighted majority voting). The most DES methods use the concept of classifier competence such as the local accuracy estimation [2], Bayes confidence measure [5] or multiple classifier behavior [4], to name only a few.

In [10, 11] and [12] the new competence measure of classifiers based on the probabilistic model has been proposed. In the method, first a randomized reference classifier (RRC) whose class supports are realizations of the random variables with beta probability distributions is constructed. The parameters of the distributions are chosen in such a way that, for each feature vector in a validation set, the expected values of the class supports produced by the RRC and the class supports produced by a modeled classifier are equal. This allows for using the probability of correct classification of the RRC as the competence of the modeled classifier. The competences calculated for a validation set are then generalized to an entire feature space by constructing a competence function (measure) based on a potential function model. Next, the DES-competence based system (DES-C) was constructed which classifies an object x in the following manner. First, the competences are determined for each base classifier in the pool. Then a subset of the classifiers with the competences greater than the probability of random classification is selected from the pool for an object x. The selected classifiers are combined using the weighted majority voting rule with continuous-valued outputs, where the weights are equal to the competences. Finally, the DES system classifies x using the maximum rule.

In this paper two new DES systems are proposed which significantly develop the presented DES-C system:

1. DES system with dynamic threshold of competence (DES-DT): This system is the same as the DES-C system except that now a subset of the classifiers with the competences greater than dynamically determined threshold is selected from the pool.
2. DES system with dynamic threshold of competence and class-dependent weights in majority voting procedure (DES-CD): This system is the same as the DES-DT system except that in the majority voting procedure weights are equal to the class-dependent competences.

[1] Department of Systems and Computer Networks, Wroclaw University of Technology,
Wybrzeze Wyspianskiego 27, 50-370 Wroclaw, Poland, email: {maciej.krysmann, marek.kurzynski}@pwr.wroc.pl

The paper is organized as follows. In section 2 the randomized reference classifier (RRC) is presented and competence measure of base classifier is developed. Sections 3 describe DES-DT and DES-CD systems. The experiments conducted and results with discussion are presented in section 4. Section 5 concludes the paper.

2. THEORETICAL FRAMEWORK

2.1. PRELIMINARIES

In the multi-classifier (MC) system we assume that a set of trained classifiers $\Psi = \{\psi_1, \psi_2, \dots, \psi_L\}$ called base classifiers is given. A classifier ψ_l $(l = 1,2,\dots,L)$ is a function $\psi_l : X \to \mathcal{M}$ from a feature space to a set of class labels $\mathcal{M} = \{1,2,\dots,M\}$**.** Classification is made according to the maximum rule:

$$\boldsymbol{\psi_l(x) = i \Leftrightarrow d_{li}(x) = \max_{j \in \mathcal{M}} d_{lj}(x)}, \tag{1}$$

where $[d_{l1}(x), d_{l2}(x), \dots, d_{lM}(x)]$ is a vector of class supports produced by ψ_l. Without loss of generality we assume, that $d_{lj}(x) \geq 0$ and $\sum_j d_{lj}(x) = 1$.

Construction of proposed DES systems is based on competence function $c(\psi_l|x)$ of base classifiers ψ_l $(l = 1,2,\dots,L)$, which can be considered as a measure of capability to correct classification of ψ_l at a point $x \in X$. Competences of base classifiers at a point x – on the one hand – are a basis of selection procedure to create an ensemble of competent classifiers and – on the other hand – are used for calculation of weights in the majority voting method of fusion. Since selection and fusion depends on feature vector *x*, both procedures are realized in dynamic fashion.

In this paper trainable competence function is proposed what leads to the assumption that a validation set containing pairs of feature vectors and their corresponding class labels is available, viz:

$$\boldsymbol{V = \{(x_1, j_1), (x_2, j_2), \dots, (x_N, j_N)\}; x_k \in X, j_k \in \mathcal{M}} \text{ .} \tag{2}$$

In the next subsection the original concept of randomized reference classifier (RRC) will be presented [10, 11, 12], which is the convenient tool for determining competences $c(\psi_l|x)$ of base classifiers.

2.2. RANDOMIZED REFERENCE CLASSIFIER – RRC

The RRC is a stochastic classifier modeling the activity of a classifier ψ from the pool Ψ (throughout this description, the index *l* of the classifier ψ_l and its class supports is dropped for clarity). RRC is for each $x \in X$ a probability distribution over the set of class labels $\mathcal{M}$ or – assuming the canonical model of classification – over the product of class supports $[0,1]^{\mathcal{M}}$. In other words, the RRC produces a vector of class supports $[\delta_1(x), \delta_2(x), \dots, \delta_M(x)]$ for the classification of the feature vector x, where the *j*-th support is a realization of a random variable (rv) $\Delta_j(x)$. Final decision is made according to (1).

The probability distributions of the rvs are chosen in such a way that the following conditions are satisfied:

1. $\Delta_j(x) \in [0,1]$;
2. $\boldsymbol{E}[\Delta_j(x)] = d_j(x),\ j = 1,2,\dots,M$;
3. $\sum_{j=1,2,\dots,M} \Delta_j(x) = 1$;

where $\boldsymbol{E}$ is the expected value operator. From the above definition it results that the RRC can be considered as equivalent to the classifier ψ for the feature vector x since it produces, on average, the same vector of class supports as the modeled classifier.

Since the RRC performs classification in a stochastic manner, it is possible to calculate the probability of classification an object x to the $\boldsymbol{i}$-th class:

$$\boldsymbol{P^{(RRC)}(i|x)} = \mathbf{Pr}\left[\forall_{\boldsymbol{k=1,\dots,M,k\neq i}}\Delta_{\boldsymbol{i}}(\boldsymbol{x}) > \Delta_{\boldsymbol{k}}(\boldsymbol{x})\right] \tag{3}$$

In particular, if the object x belongs to the i-th class, from (3) we simply get the conditional probability of correct classification $Pc^{(RRC)}(x)$.

The key element in the modeling presented above is the choice of probability distributions for the rvs $\Delta_j(x), j \in \mathcal{M}$ so that the conditions 1 – 3 are satisfied. In this paper beta probability distributions are used with the parameters $\alpha_j(x)$ and $\beta_j(x)$ $(j \in \mathcal{M})$. The justification of the choice of the beta distribution, resulting from the theory of order statistics can be found in [11].

Applying the RRC to a validation point x_k and putting in (3) $i = j_k$, we get the probability of correct classification of RRC at a point $x_k \in V$:

$$\boldsymbol{Pc^{(RRC)}(x_k) = P^{(RRC)}(j_k|x_k),\ \ x_k \in V\ .} \tag{4}$$

Since the RRC can be considered equivalent to the modeled base classifier $\psi_l \in \Psi$, it is justified to use the probability (4) as the competence of the classifier ψ_l at the validation point $x_k \in V$, i.e.

$$\boldsymbol{C(\psi_l|x_k) = Pc^{(RRC)}(x_k)}. \tag{5}$$

Using the normalized Gaussian potential function method [10], [11] for extending competence values (5) to the entire feature space X, we get competence function for base classifier ψ_l $(l = 1,2,\dots,L)$:

$$\boldsymbol{c(\psi_l|x)} = \frac{\sum_{x_k\in V}\boldsymbol{C(\psi_l|x_k)}\mathbf{exp}(-\boldsymbol{dist(x,x_k)^2})}{\boldsymbol{max}_{x\in X}\sum_{x_k\in V}\boldsymbol{C(\psi_l|x_k)}\mathbf{exp}(-\boldsymbol{dist(x,x_k)^2})}, \tag{6}$$

and class-dependent competence functions $(i \in \mathcal{M})$:

$$\boldsymbol{c_i(\psi_l|x)} = \frac{\sum_{x_k\in V:\, j_k=i}\boldsymbol{C(\psi_l|x_k)}\mathbf{exp}(-\boldsymbol{dist(x,x_k)^2})}{\boldsymbol{max}_{x\in X}\sum_{x_k\in V:\, j_k=i}\boldsymbol{C(\psi_l|x_k)}\mathbf{exp}(-\boldsymbol{dist(x,x_k)^2})}, \tag{7}$$

where $dist(x,y)$ is the Euclidean distance between the objects x and y.

3. DYNAMIC ENSEMBLE SELECTION SYSTEMS

In [11] DES system based on competences (6) and with continuous-valued outputs (DES-C) was developed. In this system first the competences (6) are determined for each base classifier and a subset Ψ_x^* of the classifiers better-than-random (with competences $c(\psi_l|x) > 1/M$) is selected from the pool for a given object x. Next the selected classifiers are combined using the weighted majority voting rule with weights equal to the competences. The weighted vector of class supports of DES-C system is given by

$$\boldsymbol{d_j^{DES-C}(x)} = \sum_{\psi_l\in\Psi_x^*}\boldsymbol{c(\psi_l|x)\, d_{lj}(x)} \tag{8}$$

Finally the maximum rule (1) is used for the classification x.

In the next subsections two novel MC systems based on DES scheme are proposed using measures of competence (5) and (6).

3.1. DYNAMIC THRESHOLD DES SYSTEM (DES-DT)

In this system the competence threshold in the selection procedure is not constant but changes dynamically depending on the number of selected classifiers and values of their competences.

The correct choice of the threshold of competence is not an easy task. As the value of competence threshold increases, it also increases the risk that Ψ_x^* is an empty set. And vice versa – when the threshold is decreased then inaccurate (incompetent) classifiers are not properly eliminated from the ensemble Ψ_x^*.

In the proposed mechanism, the competence threshold is determined iteratively so as to ensure the minimum number of member classifiers with the highest competences.

The pseudo-code of the algorithm of DES-DT system is as follows:

Input data: V – validation set; Ψ – the pool of base classifiers; $x \in X$ – the testing point, $\propto$ - initial value of threshold; r – the grain of threshold changes, n_{min} – the minimal size of ensemble Ψ_x^* (in the further experiments $\propto = 0.99$, $r = 0.1$, $n_{min} = 3$)

1. Initial values: $\Psi_x^* = \emptyset$; $n = 0$
2. For each $\psi_l \in \Psi$ calculate competence $c(\psi_l|x)$ at the point x
3. For each classifier $\psi_l \in \Psi$ do
   If $c(\psi_l|x) > \propto$ then do
      $\Psi_x^* = \Psi_x^* \cup \psi_l$
      $\Psi = \Psi - \psi_l$
      $n = n + 1$
   endfor
4. If $n < n_{min}$ then do
      $\propto = \propto - r$
      go to 3
5. Calculate supports of DES-DT system according to (8)
6. Classify the object x according to the maximum rule

3.2. CLASS-DEPENDENT DES SYSTEM (DES-CD)

This system is the same as DES-DT system except that supports are calculated as follows:

$$\boldsymbol{d_j^{DES-CD}(x) = \sum_{\psi_l \in \Psi_x^*} c_j(\psi_l|x)\, d_{lj}(x).} \tag{9}$$

where $c_j(\psi_l|x)$ denotes class-dependent competence of classifier ψ_l at a point *x* given by formula (7).

4. EXPERIMENTS

4.1. DATABASES AND EXPERIMENTAL SETUP

In order to evaluate the performance of DES systems developed, several computer experiments were made. All experiments were conducted in MATLAB with own procedures and PRTools toolbox [8] for base classifiers implementations. Benchmark databases used in experiments were obtained from UCI Machine Learning repository [9]. A brief description of each database is given in Table 1. The training and testing dataset were extracted from each dataset using two-fold cross-validation method.

Table 1. The databases used in experiments

Database	Objects	Features	Classes	Database	Objects	Features	Classes
Dermatology	366	34	6	**Pima Indians**	768	8	2
EColi	336	7	8	**Sonar**	208	60	2
Glass	214	9	6	**Spam**	4601	57	2
Haberman	306	3	2	**Wine**	178	13	3
Ionosphere	351	34	2	**Yeast**	1484	8	10
Iris	150	4	3				

The experiments were made using the pool consisted of the following nine base classifiers:

- k-nearest neighbors classifiers with k=1, 5, 15
- nearest mean classifier
- Parzen density based classifiers with the Gaussian kernel and smoothing coefficient h_{opt} and $\frac{h_{opt}}{2}$,
- decision tree with Gini splitting criterion
- neural network based classifiers with two hidden layers with 5 neurons and one hidden layer with 10 neurons, both with maximum 80 training epochs.

The performances of DES-DT and DES-CD systems were tested against original DES-C system in order to answer the question if proposed modifications of combining procedures are effective and lead to the better results.

4.2. RESULTS AND DISCUSSION

Results of experiments are presented in Tables 2, 3 and 4. Table 2 gives classification accuracies (the percentage of correct classification) of DES-C, DES-DT and DES-CD systems. The accuracies are averaged values obtained over 5 replications of two-fold cross-validation. The last row contains average ranks of tested methods (lower rank denotes better classifier). Results of all experiments were tested statistically. Firstly, a Friedman test with Iman-Davenport correction was used [1]. The test demonstrates that there are differences between DES systems tested. In turn, a post-hoc Holm test [1] showed that DES-DT is statistically significant better than DES-C and DES-CD (the level of $p < 0.11$ was considered statistically significant). However test have showed that on this significance level it cannot determine if DES-C and DES-CD systems are different, though it can be seen in Table 2 that the DES-CD system gives much worse classification accuracies.

Table 2. Classification accuracies of DES-C, DES-DT and DES-CD systems

Database	DES-C	DES-DT	DES-CD	Database	DES-C	DES-DT	DES-CD
EColi	84,98	83,66	70,54	**Haberman**	95,52	89,70	90,87
Ionosphere	80,42	83,79	73,86	**Glass**	75,12	73,15	74,01
Iris	95,70	96,38	93,29	**Yeast**	53,53	56,24	49,92
Wine	74,68	89,12	88,67	**Sonar**	69,62	72,40	65,63
Pima	70,03	69,09	69,05	**Spam**	81,69	83,78	82,30
Dermathology	82,67	94,63	89,92	**Average rank**	1,909	1,546	2,546

Table 3 and 4 presents results of testing influence of parameter r on DES-DT system, which determines the grain of change of dynamic threshold value α. Similarly as previously, the last rows contain average ranks of each part of test. Table 3 presents mean time of classification of a single object. As expected, this time decreases with increasing r value. But statistical test (as for results from Table 2) showed that – in general – differences are not statistically significant (on the same significance level $p < 0.11$).

Table 3. Mean classification times of a single object (in seconds)

Database	r=0,025	r=0,05	r=0,1	r=0,2	Database	r=0,025	r=0,05	r=0,1	r=0,2
EColi	0,311	0,311	0,308	0,301	**Haberman**	0,171	0,172	0,172	0,175
Ionosphere	0,175	0,174	0,174	0,174	**Glass**	0,184	0,182	0,184	0,189
Iris	0,216	0,216	0,214	0,213	**Yeast**	0,618	0,587	0,569	0,565
Wine	0,117	0,119	0,119	0,121	**Sonar**	0,237	0,237	0,233	0,232
Pima	0,165	0,168	0,171	0,169	**Spam**	0,548	0,548	0,561	0,562
Dermathology	0,176	0,175	0,173	0,173	**Avg. rank**	2,682	2,546	2,364	2,409

Table 4 presents percentage of correct classifications of DES-DT system depending on value of parameter r. As previously, the last row contains average ranks of DES-DT system for different values of r. Statistical test showed that there are no differences between all average rank values.

Table 4. Classification accuracies of DES-DT system depending on parameter *r*.

Database	r=0,025	r=0,05	r=0,1	r=0,2	Database	r=0,025	r=0,05	r=0,1	r=0,2
EColi	84,83	85,38	85,63	85,29	**Haberman**	91,30	90,76	90,64	91,48
Ionosphere	84,23	84,86	84,61	84,4	**Glass**	72,56	72,46	72,66	72,85
Iris	96,65	96,45	96,45	96,31	**Yeast**	55,51	55,57	55,00	55,74
Wine	90,08	90,08	89,51	89,8	**Sonar**	73,12	73,51	73,08	72,69
Pima	68,58	68,42	68,62	68,66	**Spam**	83,51	84,13	83,49	83,54
Dermathology	93,65	93,87	93,54	93,92	**Avg. rank**	2,727	2,364	2,546	2,364

Results of tests have proven the superiority of DES-DT modification over an original DES-C method. The DES-DT system achieved the highest overall classification accuracy averaged over all

datasets. On the other hand, the tests demonstrate, that modification of fusion procedure in DES-C system and using class-dependent weights in majority voting scheme leads to the worse classification results. It means that deeper exploring the competence space is not always effective and justified.

Analysis of Fig 1, which presents classification accuracy and classification time of DES-DT system for different values of parameter r, leads to the conclusion that there is no a simple relation between r and performance of DES-DT system. Time for most databases is decreasing with increasing of r value, but the quality is breaking this dependence. The best quality is achieved for most databases for $r = 0.05$, and 0.2. The worst result is achieved for $r = 0.025$. Because of lack of statistical differences between classification accuracies there is no way to determine which result is statistically the best one. For that reason value or r should be chosen experimentally for specific application.

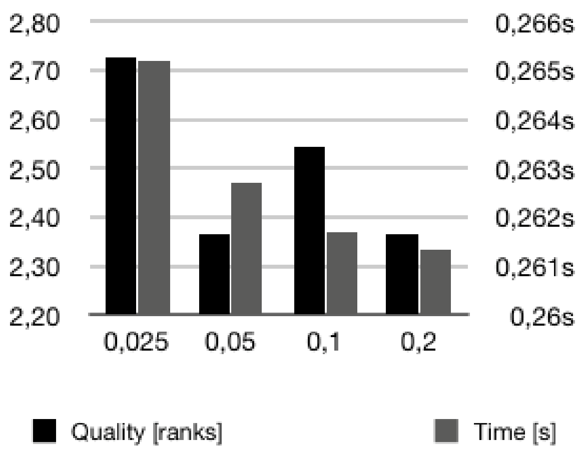

Fig. 1. Comparison of classification time and classification accuracy of DES-DT system for different values of parameter *r*.

5. FINAL REMARKS

In the paper two new methods of dynamic ensemble selection based on RRC idea are developed. The DES-DT system which uses dynamic threshold of competence in the selection procedure has proven its effectiveness and better classification accuracy than base DES-C system, verified by appropriate statistical tests. Dynamic threshold of competence in the selection procedure on the one hand fully eliminates inaccurate classifiers and on the other ensures that ensemble is not an empty set.

The DES-CD system in which class-dependent weights (competences) were applied in the weighted majority voting method gave the worse classification accuracy compared with the original DES-C system.

The DES systems based on RRC concept were successfully applied in many practical decision within biomedical engineering area including diagnosis of knee osteoarthritis based on radiographic images [13, 14] and recognition of hand grasping movements based on EMG signals [15].

As it seems, the RRC concept has great potential to built new competence–based multiple classifier systems, which till now are not fully utilized. This justified further research in this area.

BIBLIOGRAPHY

[1] DEMSAR J., Statistical comparison of classifiers over multiple data sets, Journal of Machine Learning Research 7, 2006, pp. 1-30.
[2] DIDACI L., GIACINTO G., ROLI F., MARCIALIS G., A study of the performances of dynamic classifier selection based on local accuracy estimation, Pattern Recognition 38, 2005, pp. 2188-2191.
[3] EULANDA M., SANTOS D, SABOURIN R., A dynamic overproduce and choose strategy for the selection of classifier ensembles, Pattern Recognition 41, 2008, pp. 2993-3009.
[4] GIACINTO G., ROLI F., Dynamic classifier selection based on multiple classifier behavior, Pattern Recognition 34, 2001, pp. 1879-1881.
[5] HUENUPAN F., YOMA N., Confidence based multiple classifier fusion in speaker verification, Pattern Recognition Letters 29, 2008, pp. 957-966.
[6] KO A., SABOURIN R., BRITTO A., From dynamic classifier selection to dynamic ensemble selection, Pattern Recognition 41, 2008, pp. 1718-1733.
[7] KUNCHEVA L.I., Combining Pattern Classifiers: Methods and Algorithms, John Wiley & Sons, (2004).
[8] PRTools, http://www.prtools.org/.
[9] UCI Machine Learning repository, http://archive.ics.uci.edu/ml/.
[10] WOLOSZYNSKI T., KURZYNSKI M., A measure of competence based on randomized reference classifier for dynamic ensemble selection, In: 20th Int. Conf. on Pattern Recognition, 2010, pp. 4194-4197.
[11] WOLOSZYNSKI T., KURZYNSKI M., A probabilistic model of classifier for dynamic ensemble selection, Pattern Recognition 44, 2011, pp. 2656-2668.
[12] WOLOSZYNSKI T., KURZYNSKI M., A measure of competence based on random classification for dynamic ensemble selection, Information Fusion 13, 2012, pp. 207-213.
[13] WOLOSZYNSKI T, KURZYNSKI M., et al., Prediction of progression of radiographic knee osteoarthritis using tibial trabecular bone texture, Arthritis and Rheumatism 64, 2012, pp. 688-695.
[14] WOLOSZYNSKI T, KURZYNSKI M., et al., A signature dissimilarity measure for trabecular bone texture in knee radiographics, Medical Physics 37, 2011, pp. 2030-2042.
[15] KURZYŃSKI M., , WOLCZOWSKI A., Dynamic selection of classifiers ensemble applied to the recognition of EMG signal for the control of bioprosthetic hand, Proc.11th Int. Conf. on Control, Automation and Systems, Seoul 26-29 Oct. 2011, pp. 382-386.