Probabilistic Approach to the Dynamic Ensemble Selection Using Measures of Competence and Diversity of Base Classifiers

Rafal Lysiak, Marek Kurzynski, and Tomasz Woloszynski

Wroclaw University of Technology, Dept. of Systems and Computer Networks, Wyb. Wyspianskiego 27, 50-370 Wroclaw, Poland
{rafal.lysiak,marek.kurzynski,tomasz.woloszynski}@pwr.wroc.pl

Abstract. In the paper measures of classifier competence and diversity using a probabilistic model are proposed. The multiple classifier system (MCS) based on dynamic ensemble selection scheme was constructed using both measures developed. The performance of proposed MCS was compared against three multiple classifier systems using six databases taken from the UCI Machine Learning Repository and the StatLib statistical dataset. The experimental results clearly show the effectiveness of the proposed dynamic selection methods regardless of the ensemble type used (homogeneous or heterogeneous).

Keywords: Dynamic ensemble selection, Classifier competence, Diversity measure.

1 Introduction

Dynamic ensemble selection (DES) methods are recently intensively developed as an effective approach to the construction of multiple classifier systems ([10], [14], [17]). In these methods, first an ensemble of base classifiers is dynamically selected and then the selected classifiers are combined by majority voting. The most DES schemes use the concept of classifier competence on a defined neighbourhood or region [15], such as the local accuracy estimation [19], [6], [16], Bayes confidence measure [12], multiple classifier behaviour [11] or probabilistic model [17], among others.

The performance of multiclassifier system based on DES approach can be significantly improved through the use of diverse ensemble of classifiers ([2], [4], [5]). Then, it is intuitively clear that classifiers to be selected should be competent (accurate) as well as diverse (different from one another).

In this paper a novel procedure for ensemble selection is developed, in which both a competence and a diversity of member classifiers are simultaneously taken into consideration in the dynamic fashion. Methods for calculating classifier competence and diversity using a probabilistic model are based on the original concept of a randomized reference classifier (RRC) [17], which – on average – acts like classifier evaluated. The competence of a classifier is calculated as the probability of correct classification of the respective RRC and the class-dependent

E. Corchado, M. Kurzyński, M. Woźniak (Eds.): HAIS 2011, Part II, LNAI 6679, pp. 229–236, 2011.

error probabilities of RRC are used for determining the diversity measure, which evaluates the difference of incorrect outputs of classifiers [1]. Next, the procedure for dynamic ensemble selection using both measures is proposed, in which incompetent classifiers are eliminated and the ensemble is kept maximally diverse. The subset of classifiers selected in the DES procedure is combined using continuous-valued outputs and weighted majority voting where the weights are equal to the competence values.

The paper is organized as follows. In section 2 the randomized reference classifier (RRC) is presented and measures of base classifier competence and ensemble diversity are developed. Section 3 describes the multiple classifier system that was constructed using both measures. The experiments conducted and results with discussion are presented in section 4. Section 5 concludes the paper.

2 Theoretical Framework

2.1 Preliminaries

Consider a classification problem with a set $\mathcal{M} = \{1, 2, \ldots, M\}$ of class labels and a feature space $\mathcal{X} \subseteq \mathcal{R}^n$. Let a pool of classifiers, i.e. a set of trained classifiers $\Psi = \{\psi_1, \psi_2, \ldots, \psi_L\}$ be given. Let

$$\psi_l : \mathcal{X} \rightarrow \mathcal{M} \tag{1}$$

be a classifier, that produces a vector of discriminant functions $[d_{l1}(x), d_{l2}(x), \ldots, d_{lM}(x)]$ for an object described by a feature vector $x \in \mathcal{X}$. The value of $d_{lj}(x)$, $j \in \mathcal{M}$ represents a support given by the classifier ψ_l for the fact that the object x belongs to the j-th class. Assume without loss of generality that $d_{lj}(x) \geq 0$ and $\sum_j d_{lj}(x) = 1$. Classification is made according to the maximum rule

$$\psi_l(x) = i \Leftrightarrow d_{li}(x) = \max_{j \in \mathcal{M}} \; d_{lj}(x). \tag{2}$$

Now, our purpose is to determine the following characteristics, which will be the basis for dynamic selection of classifiers from the pool:

1. a competence measure $C(\psi_l|x)$ of each base classifier ($l = 1, 2, \ldots, L$), which evaluates the competence of classifier ψ_l i.e. its capability to correct activity (correct classification) at a point $x \in \mathcal{X}$.
2. a diversity measure $D(\Psi_E|x)$ of any ensemble of base classifiers Ψ_E, considered as the independency of the errors made by the member classifiers at a point $x \in \mathcal{X}$.

In this paper trainable competence and diversity functions are proposed using a probabilistic model. It is assumed that a learning set

$$\mathcal{S} = \{(x_1, j_1), (x_2, j_2), \ldots, (x_N, j_N)\}; \;\; x_k \in \mathcal{X}, \; j_k \in \mathcal{M} \tag{3}$$

is available for the training of competence and diversity measures.

In the next section the original concept of a reference classifier will be presented, which – using probabilistic model – will state the convenient and effective tool for determining both competence and diversity measures.

2.2 Randomized Reference Classifier - RRC

A classifier[1] ψ from the pool Ψ is modeled by a randomized reference classifier (RRC) [17] which takes decisions in a random manner. A randomized decision rule (classifier) is, for each $x \in \mathcal{X}$, a probability distribution on a decision space [3] or – for the classification problem (2) – on the product $[0,1]^M$, i.e. the space of vectors of discriminant functions (supports).

The RRC classifies object $x \in \mathcal{X}$ according to the maximum rule (2) and it is constructed using a vector of class supports $[\delta_1(x), \delta_2(x), \ldots, \delta_M(x)]$ which are observed values of random variables (rvs) $[\Delta_1(x), \Delta_2(x), \ldots, \Delta_M(x)]$. Probability distributions of the random variables satisfy the following conditions:

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

where E is the expected value operator. In other words, class supports produced by the modeled classifier ψ are equal to the expected values of class supports produced by the RRC.

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

$$P^{(RRC)}(i|x) = Pr[\forall_{k=1,\ldots,M,\ k\neq i}\ \Delta_i(x) > \Delta_k(x)]. \tag{4}$$

In particular, if the object x belongs to the i-th class, from (4) 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 can be found in [17] and furthermore the MATLAB code for calculating probabilities (4) was developed and it is freely available for download [18].

Applying the RRC to a learning point x_k and putting in (4) $i = j_k$, we get the probability of correct classification of RRC at a point $x_k \in \mathcal{S}$, namely

$$Pc^{(RRC)}(x_k) = P^{(RRC)}(j_k|x_k), \quad x_k \in \mathcal{S}. \tag{5}$$

Similarly, putting in (4) a class $j \neq j_k$ we get the class-dependent error probability at a point $x_k \in \mathcal{S}$:

$$Pe^{(RRC)}(j|x_k) = P^{(RRC)}(j|x_k), \quad x_k \in \mathcal{S}, \quad j(\neq j_k) \in \mathcal{M}. \tag{6}$$

In next sections probabilities of correct classification (5) and conditional probabilities of error (6) for learning objects will be utilized for determining the competence and diversity functions of base classifiers.

[1] Throughout this subsection, the index l of classifier ψ_l and class suppports $d_{lj}(x)$ is omitted for clarity.

2.3 Measure of Classifier Competence

Since the RRC can be considered equivalent to the modeled base classifier $\psi_l \in \Psi$, it is justified to use the probability (5) as the competence of the classifier ψ_l at the learning point $x_k \in \mathcal{S}$, i.e.

$$C(\psi_l|x_k) = Pc^{(RRC)}(x_k). \tag{7}$$

The competence values for the validation objects $x_k \in \mathcal{S}$ can be then extended to the entire feature space $\mathcal{X}$. To this purpose the following normalized Gaussian potential function model was used ([17]):

$$C(\psi_l|x) = \frac{\sum_{x_k \in \mathcal{S}} C(\psi_l|x_k) exp(-dist(x, x_k)^2)}{\sum_{x_k \in \mathcal{S}} exp(-dist(x, x_k)^2)}, \tag{8}$$

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

2.4 Measure of Diversity of Classifiers Ensemble

As it was mentioned previously, the diversity of a classifier ensemble Ψ_E is considered as an independency of the errors made by the member classifiers. Hence the method in which diversity measure is calculated as a variety of class-dependent error probabilities is fully justified.

Similarly, as in competence measure, we assume that at a learning point $x_k \in \mathcal{S}$ the conditional error probability for the class $j \neq \jmath_k$ of the base classifier ψ_l is equal to the appropriate probability of the equivalent RRC, namely:

$$Pe^{(\psi_l)}(j|x_k) = Pe^{(RRC)}(j|x_k). \tag{9}$$

Next, these probabilities can be extended to the entire feature space $\mathcal{X}$ using Gaussian potential function (8):

$$Pe^{(\psi_l)}(j|x) = \frac{\sum_{x_k \in \mathcal{S}, \jmath_k \neq j} Pe^{(\psi_l)}(j|x_k) exp(-dist(x, x_k)^2)}{\sum_{x_k \in \mathcal{S}, \jmath_k \neq j} exp(-dist(x, x_k)^2)}. \tag{10}$$

According to the presented concept, using probabilities (10) first we calculate pairwise diversity at the point $x \in \mathcal{X}$ for all pairs of base classifiers ψ_l and ψ_k from the pool Ψ:

$$D(\psi_l, \psi_k|x) = \frac{1}{M} \sum_{j \in \mathcal{M}} |Pe^{(\psi_l)}(j|x) - Pe^{(\psi_k)}(j|x)|, \tag{11}$$

and finally we get diversity of ensemble of n $(n \leq L)$ base classifiers $\Psi_E(n)$ at a point $x \in \mathcal{X}$ as a mean value of pairwise diversities (11) for all pairs of member classifiers, namely:

$$D(\Psi_E(n)|x) = \frac{2}{n \cdot (n-1)} \sum_{\psi_l, \psi_k \in \Psi_E(n); l \neq k} D(\psi_l, \psi_k|x). \tag{12}$$

3 Dynamic Ensemble Selection System

3.1 Method

The proposed DES competence and diversity based classification system (DES-CD) is constructed in the procedure consisting of two steps:

1. For the test object $x \in \mathcal{X}$ and for given ensemble size n and the competence threshold α first the ensemble of classifiers $\Psi_E^*(n)$ is found as a solution of the folowing optimization problem:
$$D(\Psi_E^*(n)|x) = max_{\Psi_E(n)} D(\Psi_E(n)|x) \tag{13}$$
subject to $C(\psi_l|x) \geq \alpha$ for $\psi_l \in \Psi_E^*$. This step eliminates incompetent (inaccurate) classifiers and keeps the ensemble maximally diverse.
2. The selected classifiers are combined by weighted majority voting where the weights are equal to the competence values. The weighted vector of class supports of DES-CD system is given by
$$d_j^{(DES-CD)}(x) = \sum_{\psi_l \in \Psi_E^*(n)} C(\psi_l|x)\, d_{jl}(x) \tag{14}$$
and final decision is made according to the maximum rule (2).

3.2 Solution of Optimization Problem

The key moment in the method developed is the optimization problem (13). As a solution method we propose suboptimal procedure which is followed sequential forward feature selection method [13]. In this method first the set of competent classifiers (better than threshold α) is created and next classifiers are sequentially selected from this set: at first the classifier with the highest competence is chosen, next to the already selected classifier we add another one so as to create the couple with the best diversity, then the three classifiers with the highest diversity, including the selected first and second ones are chosen and so one. This procedure is continued up to n classifiers are selected.

The pseudo-code of the algorithm is as follows:

Input data: $\mathcal{S}$ - learning set; Ψ_L - the pool of classifiers;
n - the size of enesemble; $x \in \mathcal{X}$ - the testing point;
α - the threshold of competence

1. For each $\psi_l \in \Psi_L$ calculate competence $C(\psi_l|x)$ at the point x
2. Create temporal set of competent classifiers at the point x
   $\Psi(x) = \{\psi_l \in \Psi_L \,:\, C(\psi_l|x) \geq \alpha\}$
3. $\Psi_E^*(n) = \{\psi_{(1)}\}$ and $\Psi(x) = \Psi(x) - \psi_{(1)}$ where
   $\psi_{(1)} \,:\, C(\psi_{(1)}|x) = max_{\psi \in \Psi(x)} C(\psi|x)$
4. For $i = 2$ to n do
   a) Find $\psi_{(i)} \in \Psi(x)$ for which
      $D(\Psi_E^*(n) \cup \psi_{(i)}|x) = max_{\psi \in \Psi(x)} D(\Psi_E^*(n) \cup \psi|x)$
   b) $\Psi(x) = \Psi(x) - \psi_{(i)}$
   c) $\Psi_E^*(n) = \Psi_E^*(n) \cup \psi_{(i)}$

   endfor

4 Experiments

4.1 Databases and Experimental Setup

The benchmark databases used in the experiments were obtained from the UCI Machine Learning Repository (*Breast Cancer Wisconsin*, *Glass*, *Iris*, *Sonar*, *Ionosphere*) and StatLib statistical datasets (*Biomed*). The experiments were conducted in MATLAB using PRTools, which automatically normalizes feature vectors for zero mean and unit standard deviation and for a given $x \in \mathcal{X}$ produces classifying functions for all base classifiers according to the paradigms of their activity [9]. The training and testing datasets were extracted from each database using two-fold cross-validation. The base classifiers and both competence and diversity measures were trained using the same training dataset.

The DES-CD system was compared against three multiclassifier systems: (1) SB system – this system selects the single best classifier in the pool; (2) MV system – this system is based on majority voting of all classifiers in the pool; (3) DES-SC system – this system defines the competence of the classifier ψ for the test object x according to (8) and next the ensemble of competent (better-than-random) classifiers is selected [17] – the final decision is made as in (14).

Two types of classifier ensembles were used in the experiments: homogeneous and heterogeneous. The homogeneous ensemble consisted of 20 pruned decision tree classifiers with Gini splitting criterion. Each classifier was trained using randomly selected 70% of objects from the training dataset.

The pool of heterogeneous base classifiers using in the experiments, consisted of the following nine classifiers [8]: (1-2) linear (quadratic) discriminant classifier based on normal distributions with the same (different) covariance matrix for each class; (3) nearest mean classifier; (4-6) k-NN - k-nearest neighbours classifiers with $k = 1, 5, 15$; (7-8) Parzen classifier with the Gaussian kernel and the optimal smoothing parameter h_{opt} (and the smoothing parameter $h_{opt}/2$); (9) pruned decision tree classifier with Gini splitting criterion.

The following values of parameters of DES-CD system were adopted in the experiments: $\alpha = 1/M$ and $n = max\{\frac{1}{2}|\Psi(x)|, 2\}$.

4.2 Results and Conclusion

Classification accuracies (i.e. the percentage of correctly classified objects) were calculated for the MCSs as average values obtained over 10 runs (5 replications of two-fold cross validation). Statistical differences between the performances of the DES-CD system and the three MCSs were evaluated using Dietterich's 5x2cv test [7]. The level of $p < 0.05$ was considered statistically significant. The results obtained for the MCSs using heterogeneous and homogeneous ensembles are shown in Table 1. For each database and for DES systems the mean sizes of classifier ensembles are given under the classification accuracy.

These results imply the following conclusions:

1. The DES-CD system outperformed the SB and MV classifiers by 7.82% and 5.78 % for heterogeneous ensemble and by 3.99% and 0.47% for homogeneous ensemble, respectively;

Table 1. Classification accuracies of the MCSs using heterogeneous/homogeneous ensembles. The mean sizes of classifier ensembles and statistically significant differences found are given under the classification accuracies. The best result for each database is bolded.

Database	SB (1)	MV (2)	DES-CS (3)	DES-CD (4)
Breast C.W.	95.39/94.99	96.45/95.99	98.03/96.18 9.13/19.58 1, 2/1	**98.05**/**96.22** 5.21/9.82 1, 2/1
Biomed	84.10/83.29	87.62/86.90	**90.09**/**86.91** 8.70/17.31 1, 2/1	**90.09**/86.81 5.96/9.61 1, 2/1
Glass	71.80/61.56	69.55/71.03	**76.46**/**73.18** 9.46/19.44 1, 2, 4/1, 2	73.26/72.06 5.13/9.75 1, 2/1, 2
Iris	96.00/91.07	96.80/90.80	96.67/90.8 8.87/20.00	**97.33**/**91.13** 4.98/10.00 1/
Sonar	74.48/70.19	76.44/76.06	**82.29**/**77.12** 8.63/19.76 1, 2/1, 2	81.52/**77.12** 5.21/10.00 1, 2/1, 2
Ionosphere	84.84/88.15	86.50/89.63	**90.14**/89.74 9.06/19.85 1, 2/1	89.80/**89.88** 5.48/9.95 1, 2/1, 2
Average	81.02/81.54	83.06/85.06	**88.94**/**85.56** 8.97/19.32	88.84/85.53 5.32/9.85

2. The DES-CD system achieved the highest classification accuracy for 3 datasets and 4 datasets for heterogeneous and homogeneous ensembles, respectively; it produced statistically significant higher scores in 19 out of 36 cases.
3. There are no statistically significant difference between classification accuracies of the DES-CS and the DES-CD systems;
4. The relative difference between mean ensemble sizes for the DES-CS and the DES-CD systems is on average equal to 40.6% and 49% for heterogeneous and homogeneous ensembles, respectively.

5 Conclusion

In this paper a novel procedure for dynamic ensemble selection is proposed using probabilistic measures of competence and diversity of member classifiers. Results of experimental investigations indicate, that the proposed method can eliminate weak classifiers and keep the ensemble maximally diverse. This approach leads to the final classification accuracy which, on average, was very close to the accuracy of DES system using only the competence measure but achieved by means of smaller number of classifiers in the ensemble.

References

1. Aksela, M.: Comparison of classifier selection methods for improving committee performance. In: Windeatt, T., Roli, F. (eds.) MCS 2003. LNCS, vol. 2709, pp. 84–93. Springer, Heidelberg (2003)
2. Aksela, M., Laaksonen, J.: Using diversity of errors for selecting members of a committee classifier. Pattern Recognition 39, 608–623 (2006)
3. Berger, J.O.: Statistical Decision Theory and Bayesian Analysis. Springer, New York (1987)
4. Brown, G., Wyatt, J., Harris, R., Yao, X.: Diversity creation methods: a survey and categorisation. Information Fusion 6, 5–20 (2005)
5. Canuto, A., Abreu, M., et al.: Investigating the influence of the choice of the ensemble members in accuracy and diversity of selection-based and fusion-based methods for ensemble. Pattern Recognition Letters 28, 472–486 (2007)
6. Didaci, L., Giacinto, G., Roli, F., Marcialis, G.: A study on the performances of dynamic classifier selection based on local accuracy estimation. Pattern Recognition 38, 2188–2191 (2005)
7. Dietterich, T.G.: Approximate statistical tests for comparing supervised classification learning algorithms. Neural Computation 10, 1895–1923 (1998)
8. Duda, R., Hart, P., Stork, G.: Pattern Classification. John Wiley and Sons, New York (2000)
9. Duin, R., Juszczak, P., Paclik, P., et al.: PRTools4. In: A Matlab Toolbox for Pattern Recognition, Delft University of Technology (2007)
10. Eulanda, M., Santos, D., Sabourin, R., Maupin, P.: A dynamic overproduce-and-choose strategy for the selection of classifier ensembles. Pattern Recognition 41, 2993–3009 (2008)
11. Giacinto, G., Roli, F.: Dynamic classifier selection based on multiple classifier behaviour. Pattern Recognition 34, 1879–1881 (2001)
12. Huenupan, F., Yoma, N., et al.: Confidence based multiple classifier fusion in speaker verification. Pattern Recognition Letters 29, 957–966 (2008)
13. Kanal, L.: Patterns in pattern recognition. IEEE Trans. Information Theory 20, 697–722 (1974)
14. Ko, A., Sabourin, R., Britto, A.: From dynamic classifier selection to dynamic ensemble selection. Pattern Recognition 41, 1718–1733 (2008)
15. Kuncheva, I.: Combining Pattern Classifiers: Methods and Algorithms. Wiley Interscience, Hoboken (2004)
16. Smits, P.: Multiple classifier systems for supervised remote sensing image classification based on dynamic classifier selection. IEEE Trans. on Geoscience and Remote Sensing 40, 717–725 (2002)
17. Woloszynski, M., Kurzynski, M.: A measure of competence based on randomized reference classifier for dynamic ensemble selection. In: 20th Int. Conf. on Pattern Recognition, pp. 4194–4197. IEEE Computer Press, Istanbul (2010)
18. Woloszynski, T.: Matlab Central File Enchange (2010), `http://www.mathwork.com/matlabcentral/fileenchange/28391-classifier-competence-based-on-probabilistic-modeling`
19. Woods, K., Kegelmeyer, W., Bowyer, W.: Combination of multiple classifiers using local accuracy estimates. IEEE Trans. on Pattern Analysis and Machine Learning 19, 405–410 (1997)