PERFORMANCE OF CLASSIFICATION CONFIDENCE MEASURES IN DYNAMIC CLASSIFIER SYSTEMS

David Štefka,[*] *Martin Holeňa* [†]

Abstract: Classifier combining is a popular technique for improving classification quality. Common methods for classifier combining can be further improved by using dynamic classification confidence measures which adapt to the currently classified pattern. However, in the case of dynamic classifier systems, the classification confidence measures need to be studied in a broader context – as we show in this paper, the degree of consensus of the whole classifier team plays a key role in the process. We discuss the properties which should hold for a good confidence measure, and we define two methods for predicting the feasibility of a given classification confidence measure to a given classifier team and given data. Experimental results on 6 artificial and 20 real-world benchmark datasets show that for both methods, there is a statistically significant correlation between the feasibility of the measure, and the actual improvement in classification accuracy of the whole classifier system; therefore, both feasibility measures can be used in practical applications to choose an optimal classification confidence measure.

Key words: *classifier combining, dynamic classifier systems, classification confidence*

Received: March 13, 2013
Revised and accepted: June 27, 2013

1. Introduction

In the literature of pattern recognition and machine learning in general, methods which combine information from multiple "weak learners", in order to build a better and more robust learning model, are increasingly more popular. In the field of classification, such approaches are usually called classifier combining, or classifier aggregation methods [16, 18, 19, 21, 22, 25].

Quite often, the aggregation method uses some kind of confidence measure to estimate the quality of a given classifier, which determines the classifier's weight in

[*]David Štefka
Department of Mathematics, FNSPE, CTU Prague, david.stefka@gmail.com
[†]Martin Holeňa
Institute of Computer Science AS CR, martin@cs.cas.cz

the aggregation process. Traditionally, the confidence measures assess the classifier from a global point of view, i.e., the resulting confidence is a constant of the classifier [10, 14]. As the computational power of today's computers grows, more complex methods, which compute the classifier's confidence dynamically (i.e., in the context of the currently classified pattern), are more and more popular [3, 5–7, 13]. If there are enough validation data, the confidence measure can express the quality of the classification better, which leads to better approximation properties of the resulting aggregated classifier system [1, 8, 12, 17, 20, 23, 29, 30].

However, as we will show in this paper, in classifier aggregation, the dynamic confidence of classification has to be studied in a broader context. An important novel feature, which needs to be taken into account, is the degree of consensus among the individual classifiers in the team. For instance, if most of the classifiers agree on the class prediction for a given pattern, the confidences of the individual classifiers are not relevant because it is very hard to change the prediction of the team, anyway. On the other hand, if, for a given pattern, the predictions of the classifiers in the team are more diverse, the (dynamic) confidences of the individual classifiers begin to play a key role in the aggregation process.

In this paper, we first discuss the properties which should hold for a "good" confidence measure, and we also present examples of the performance of two individual confidence measures in more detail. This discussion leads to the definition of two different methods for estimating the feasibility of a given confidence measure to a given classifier team and given data. The presented methods, called Similarity to the Oracle confidence measure (SOR), and Area Under ROC curve for OK/NOK histogram (AUC), both incorporate the concept of restriction of the validation set to patterns with low degree of consensus of the classifier team.

In the experimental section, we empirically study the correlation between the feasibility of a given confidence measure, and the actual improvement in classification accuracy if this measure is used in a dynamic classifier system (compared to a confidence-free classifier system). The experiments were performed on 6 artificial and 20 real-world benchmark datasets, using one static and four dynamic confidence measures. The results show a statistically significant correlation in most cases, and thus suggest that the proposed feasibility measures can be used in practical applications to choose an optimal confidence measure for a given application setup.

The paper is structured as follows. Section 2. briefly presents the formalism of dynamic classifier systems, and provides examples of the most common confidence measures. In Section 3., we formally define the degree of consensus in a classifier team, and we present the SOR and AUC methods for predicting the feasibility of a given confidence measure. Section 4. contains the experimental results, and, finally, Section 5. concludes the paper.

2. Formalism of Dynamic Classifier Systems

In this section, we recall the formalism of dynamic classifier systems, as proposed in [27]. Let $\mathcal{X} \subseteq \mathbf{R}^n$ be n-dimensional *feature space*, let $C_1, \ldots, C_N \subseteq \mathcal{X}$, $N \geq 2$ be disjoint sets called *classes*. A *pattern* is a tuple $(\mathbf{x}, c_\mathbf{x})$, where $\mathbf{x} \in \mathcal{X}$ are *features* of the pattern, and $c_\mathbf{x} \in \{1, ..., N\}$ is the index of the class the pattern belongs to.

Given an unclassified pattern $\mathbf{x}$, a *classifier* $\phi : \mathcal{X} \rightarrow [0,1]^N$ predicts the *degree of classification* (d.o.c.) to each class, $\phi(\mathbf{x}) = (\gamma_1(\mathbf{x}), \ldots, \gamma_N(\mathbf{x}))$. The d.o.c. are then transformed to a crisp class label (with maximal d.o.c.) to provide the final class prediction.

The reliability of the classifier's prediction for the current pattern is expressed by a *confidence measure* $\kappa_\phi : \mathcal{X} \rightarrow [0,1]$ (the closer to 1, the more confidence is given to the prediction), which can be either *static* (i.e., a constant of the classifier) [10, 14], or *dynamic* (i.e., the confidence measure is adapted to the currently classified pattern) [7, 12, 17, 23, 27, 29], e.g., the accuracy of the classifier, measured on a set of k nearest neighbors of the classified pattern x from a validation set.

In classifier combining, instead of using a single classifier, a team of r classifiers is trained, and the outputs of the team are aggregated into the final prediction. Given an unclassified pattern $\mathbf{x}$, the outputs of the classifiers are structured to a matrix $\Gamma(\mathbf{x}) \in [0,1]^{r \times N}$, called *decision profile*,

$$\Gamma(\mathbf{x}) = \begin{pmatrix} \phi_1(\mathbf{x}) \\ \phi_2(\mathbf{x}) \\ \vdots \\ \phi_r(\mathbf{x}) \end{pmatrix} = \begin{pmatrix} \gamma_{1,1}(\mathbf{x}) & \gamma_{1,2}(\mathbf{x}) & \ldots & \gamma_{1,N}(\mathbf{x}) \\ \gamma_{2,1}(\mathbf{x}) & \gamma_{2,2}(\mathbf{x}) & \ldots & \gamma_{2,N}(\mathbf{x}) \\ & & \ddots & \\ \gamma_{r,1}(\mathbf{x}) & \gamma_{r,2}(\mathbf{x}) & \ldots & \gamma_{r,N}(\mathbf{x}) \end{pmatrix}, \tag{1}$$

and the confidences to a *confidence vector*

$$\mathcal{K}(\mathbf{x})^T = (\kappa_{\phi_1}(\mathbf{x}), \ldots, \kappa_{\phi_r}(\mathbf{x}))^T. \tag{2}$$

We restrict ourselves to the most common *class-conscious aggregation* [18], where each column of the decision profile (represeting the d.o.c. to a particular class given by all the classifiers in the team) is aggregated individually by an aggregation operator $\mathcal{A}$, and the aggregation operator is usually parametrized by the confidence vector. An example is the well-known weighted mean aggregation:

$$\gamma_j(\mathbf{x}) = \frac{\sum_{i=1,\ldots,r} \kappa_{\phi_i}(\mathbf{x}) \gamma_{i,j}(\mathbf{x})}{\sum_{i=1,\ldots,r} \kappa_{\phi_i}(\mathbf{x})}, j = 1, \ldots, N. \tag{3}$$

The resulting classifier system behaves as a single classifier Φ to the outside. Depending on the confidence measures and the aggregation operator, the classifier system can be *confidence-free* (no classification confidence is used), *static* (only static classification confidence is used), and *dynamic* (the aggregation is adapted to $\mathbf{x}$ by utilizing the dynamic classification confidence) [27]. The different approaches are shown in Fig. 1. Our main interest in this paper lies in studying dynamic classifier systems (and thus dynamic confidence measures).

2.1 Static Confidence Measures

Static confidence measures estimate the classifier's predictive power from a global point of view (the confidence is a constant of the classifier). These methods include accuracy, precision, sensitivity, resemblance, etc. [10, 14]. In this paper, we will use the (most common) Global Accuracy measure.

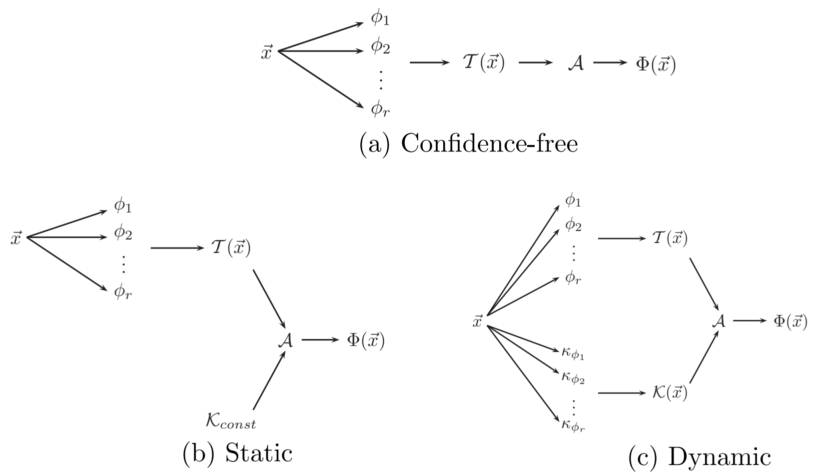


Fig. 1 *Schematic comparison of confidence-free, static, and dynamic classifier systems.*

Global Accuracy (GA) of a classifier ϕ is defined as the proportion of correctly classified patterns from the validation set:

$$\kappa_{\phi}^{(GA)} = \frac{\sum\limits_{(\mathbf{y},c_{\mathbf{y}})\in\mathcal{V}} I(\phi^{(cr)}(\mathbf{y}) \stackrel{?}{=} c_{\mathbf{y}})}{|\mathcal{V}|}, \tag{4}$$

where $\mathcal{V} \subseteq \mathcal{X} \times \{1, \ldots, N\}$ is the validation set and $\phi^{(cr)}(\mathbf{y})$ is the crisp output of ϕ on $\mathbf{y}$.

2.2 Dynamic Confidence Measures

Dynamic confidence measures adapt the estimate to the currently classified pattern $\mathbf{x}$. The most straightforward way is to restrict a global confidence measure to some neighborhood of $\mathbf{x}$. Let $N(\mathbf{x}) \subseteq \mathcal{V}$ denote a set of neighboring patterns from the validation set (e.g., using the Euclidean metric). We define two dynamic confidence measures based on $N(\mathbf{x})$:

Euclidean Local Accuracy (ELA), used in [29], measures the local accuracy of ϕ in $N(\mathbf{x})$:

$$\kappa_{\phi}^{(ELA)}(\mathbf{x}) = \frac{\sum\limits_{(\mathbf{y},c_{\mathbf{y}})\in N(\mathbf{x})} I(\phi^{(cr)}(\mathbf{y}) \stackrel{?}{=} c_{\mathbf{y}})}{|N(\mathbf{x})|}, \tag{5}$$

where $\phi^{(cr)}(\mathbf{y})$ is the crisp output of ϕ on $\mathbf{y}$.

Euclidean Local Match (ELM), based on the ideas in [7], measures the proportion of patterns in $N(\mathbf{x})$ from the same class as ϕ is predicting for $\mathbf{x}$:

$$\kappa_{\phi}^{(ELM)}(\mathbf{x}) = \frac{\sum_{(\mathbf{y},c_{\mathbf{y}})\in N(\mathbf{x})} I(\phi^{(cr)}(\mathbf{x}) \stackrel{?}{=} c_{\mathbf{y}})}{|N(\mathbf{x})|}, \tag{6}$$

where $\phi^{(cr)}(\mathbf{x})$ is the crisp output of ϕ on $\mathbf{x}$. The difference between (5) and (6) is that in the latter case, there is $\phi^{(cr)}(\mathbf{x})$ instead of $\phi^{(cr)}(\mathbf{y})$ in the indicator.

In [12], the authors suggest that if the classifier is a member of a team of classifiers, the set of nearest neighbors $N(\mathbf{x})$ should be restricted to patterns which are similar to $\mathbf{x}$ in the way how often the individual classifiers in the team classify the patterns into the same class. This is very similar to the approach of Robnik-Šikonja and Tsymbal et al. [20, 23] for random forests [4]. In this paper, we use this approach to modify the ELA and ELM confidence measures as follows.

Let $\{\phi_1, \ldots, \phi_r\}$ be a set of classifiers, and let $\mathbf{x}$ and $\mathbf{y}$ be two patterns. The similarity of the patterns is defined as

$$S(\mathbf{x}, \mathbf{y}) = \frac{1}{r} \sum_{i=1}^{r} I(\phi_i^{(cr)}(\mathbf{x}) \stackrel{?}{=} \phi_i^{(cr)}(\mathbf{y})), \tag{7}$$

where $\phi_i^{(cr)}(\mathbf{x})$ and $\phi_i^{(cr)}(\mathbf{y})$ are crisp outputs of the i-th classifier on $\mathbf{x}$ and $\mathbf{y}$. Let $N(\mathbf{x})$ be a set of k nearest neighbors of $\mathbf{x}$ under Euclidean metric. Then we define a set $\widetilde{N(\mathbf{x})}$ of neighboring patterns of $\mathbf{x}$ similar to $\mathbf{x}$, as a restriction of $N(\mathbf{x})$ to patterns with $S(\mathbf{x}, \mathbf{y})$ higher than a fixed similarity threshold $T \in (0, 1]$:

$$\widetilde{N(\mathbf{x})} = \{\mathbf{y} \in N(\mathbf{x}) \mid S(\mathbf{x}, \mathbf{y}) \geq T\}. \tag{8}$$

This allows us to modify ELA and ELM confidence measures:

Restricted Euclidean Local Accuracy (RELA), same as ELA, but using $\widetilde{N(\mathbf{x})}$ instead of $N(\mathbf{x})$

Restricted Euclidean Local Match (RELM), same as ELM, but using $\widetilde{N(\mathbf{x})}$ instead of $N(\mathbf{x})$

The aforementioned confidence measures defined in this section need to compute neighboring patterns of $\mathbf{x}$, which can be time-consuming, and sensitive to the similarity measure used. There are also dynamic confidence measures which compute the classification confidence directly from the degrees of classification [2, 28], e.g., the highest degree of classification, the ratio of the highest d.o.c. to the sum of all d.o.c.s, etc. However, our preliminary experiments with such measures with quadratic discriminant classifiers and random forests show that such confidence measures give very poor results [26]. This may be caused by the fact that in these approaches, the d.o.c.s must be good approximations of the posterior probabilities that the pattern belongs to a given class, which is often hard to accomplish.

Another reason that these approaches fail to improve the prediction of a classifier system may be that the same information (d.o.c.s) is used both to compute the confidence, and also in aggregation of the results of the individual classifiers in the team, which means there is little useful information added to the classifier aggregation process.

2.3 The Oracle Confidence Measure

For reference purposes, we also define a so-called *Oracle confidence measure*, which represents the "best-we-can-do" approach.

Oracle (OR) confidence is equal to 1 iff the pattern is classified correctly, 0 otherwise:

$$\kappa_{\phi}^{(OR)}(\mathbf{x}) = I(\phi^{(cr)}(\mathbf{x}) \stackrel{?}{=} c_{\mathbf{x}}) \tag{9}$$

Of course, in practical applications, we cannot use the Oracle confidence measure because we do not know the actual class the pattern belongs to ($c_{\mathbf{x}}$). However, the Oracle confidence measure can give us upper bound for performance of a classifier system using classification confidence, and it can also be used to assess the feasibility of a given confidence measure (cf. Sec. 3.2).

3. Assessing Confidence Measures

In [26, 27], we have experimentally shown that dynamic classifier systems of Random Forests [4] and Quadratic Discriminant Classifiers [10] using the ELA and ELM confidence measures can significantly improve the quality of classification, compared to confidence-free, or static classifier systems.

However, in these experiments, the performance of the dynamic classifier systems varied from dataset to dataset. For some datasets, the ELM confidence measure obtained better results, for others the ELA was more successful, and for some datasets, neither of them improved the classification. In other words, the performance of a dynamic classifier system is heavily influenced by the particular confidence measure used and by the particular data.

Given a particular dataset to classify, and given a set of classifiers which form a classifier team, there are several questions which come into one's mind:

- Will a dynamic classifier system yield improvement in the classification quality compared to confidence-free or static classifier system?
- Which confidence measure will perform the best for the given classifiers and the given dataset?
- Are the benefits of a dynamic classifier system worth the higher computational complexity?

To answer these questions, we could, of course, build the classifier systems and compare their performance using crossvalidation or other standard machine learning techniques. However, it would be more convenient if we had some criterion of feasibility of a given confidence measure which could answer these questions *prior* to building and crossvalidating the aggregation models.

Suppose we have several confidence measures which can be used with given classifiers on a given data. If we had such a feasibility criterion, we could experimentally measure the feasibilities of the different confidence measures, and we could choose the one with the highest feasibility value. Using this approach, we can build a classifier team in which the confidence measures are well-suited for the given classifier type and for the given data. The last step is to add a team aggregator to create a dynamic classifier system. Or, alternatively, if none of the confidence measures obtains sufficiently high feasibility value, we can decide to create a static, or confidence-free classifier system instead (in accordance with the Occam's razor principle).

In this paper, we introduce two such feasibility criteria. Before that, we summarize the properties which should hold for a "good" confidence measure. Intuitively, if $\kappa_\phi(\mathbf{x})$ estimates the degree of trust we can give to the classifier ϕ when classifying a pattern $\mathbf{x}$, the following should be satisfied:

- With increasing confidence $\kappa_\phi(\mathbf{x})$, the probability of correct classification of the classifier's prediction $\phi^{(cr)}(\mathbf{x})$ should increase as well
- If the errorness of $\phi^{(cr)}(\mathbf{x})$ increases, the classification confidence $\kappa_\phi(\mathbf{x})$ should decrease to zero

For example, if $\kappa_\phi(\mathbf{x})$ is an estimate of the probability of correct classification of $\mathbf{x}$ by ϕ (for example the ELA confidence measure), both these implications are satisfied, if the estimate is good enough. According to these two properties, the ideal confidence measure is the Oracle confidence measure.

In this paper, we propose an approach in which the feasibility of a confidence measure is measured empirically, on a set of validation patterns. Let ϕ be a classifier, κ_ϕ a confidence measure, and $\mathcal{V} \subseteq \mathcal{X} \times \{1, \ldots, N\}$ the validation set. We will model the feasibility as a number in the unit interval – the more the confidence measure satisfies the above-mentioned properties, the closer to 1 the feasibility is. The feasibility of κ_ϕ for classifier ϕ, measured empirically on data $(\mathbf{x}, c_\mathbf{x}) \in \mathcal{V}$ will be denoted as $\mathcal{F}(\phi, \kappa_\phi, \mathcal{V}) \in [0, 1]$. Two particular methods how $\mathcal{F}(\phi, \kappa_\phi, \mathcal{V})$ can be defined will be shown in Sec. 3.2 and 3.3.

However, in classifier combining, we do not have a single classifier and its corresponding confidence measure – we have a set of classifiers Γ, and a set of corresponding confidence measures $\mathcal{K}$. Therefore, we define $\mathcal{F}(\Gamma, \mathcal{K}, \mathcal{V}) \in [0, 1]$ as the average feasibility of $\kappa_\phi \in \mathcal{K}$ for the corresponding classifier $\phi \in \Gamma$, measured on $\mathcal{V}$:

$$\mathcal{F}(\Gamma, \mathcal{K}, \mathcal{V}) = \frac{\sum\limits_{\phi \in \Gamma} \mathcal{F}(\phi, \kappa_\phi, \mathcal{V})}{|\Gamma|}. \tag{10}$$

3.1 Restricting the Validation Set

There is an important aspect which needs to be taken into account when assessing the feasibility of a confidence measure in the context of classifier systems. If we measure $\mathcal{F}(\phi, \kappa_\phi, \mathcal{V})$ on the whole validation set $\mathcal{V}$, we have an estimate how κ_ϕ

predicts the classification confidence *for a single classifier*. However, if we want to assess a confidence measure's performance in the context of dynamic classifier systems, we need to know something different: can this particular confidence measure improve the generalization of the classifier system?

What is the difference between these two kinds of information? A typical situation in classifier aggregation is as follows: for most patterns, the crisp outputs of the individual classifiers in a classifier system show consensus on a certain class (i.e., a vast majority of the classifiers predicts one particular class), and the team aggregator usually does not break this consensus, even when incorporating the classification confidences (for example, if we have a system of ten classifiers in which nine of them predict class C_1 with confidence 0.1, and one classifier predicts class C_2 with confidence 0.8, then if we use the weighted mean aggregation, the prediction of C_2 is discarded). Therefore, the behavior of the confidence measures on such patterns is irrelevant. On the other hand, for patterns where there is no such consensus, the behavior of the confidence measure is *much* more important. Therefore, we need to identify such patterns, and restrict $\mathcal{V}$ to a such subset.

Let $0 \leq s \leq r$, where $r = |\Gamma|$ is the number of classifiers. Let $U(s) \subseteq \mathcal{V}$ be the set of patterns $(\mathbf{x}, c_{\mathbf{x}})$, for which for all classes C_j, $j = 1, \ldots, N$, we have

$$|\{i; i = 1, \ldots, r,\ \phi_i^{(cr)}(\mathbf{x}) = j\}| \leq s. \tag{11}$$

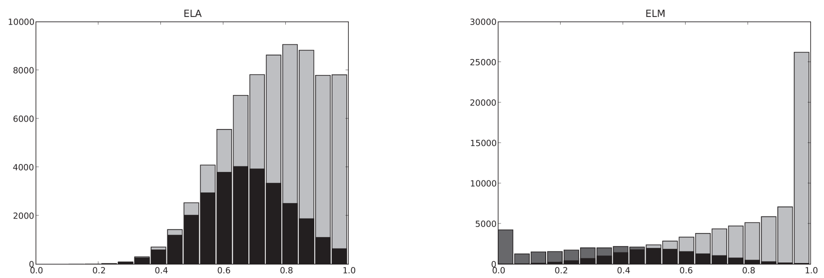
$U(s)$ therefore denotes a set of patterns for which at most s classifiers vote for any particular class. For lower s, this means that there is no consensus on a particular class, and so the team aggregator can easily use the classification confidence to improve the prediction – this suggests that the restricted validation sets for lower s are more important for the analysis. However, the smaller s, the smaller $|U(s)|$, which leads us to the fact that we need s big enough so the feasibility is measured on enough data, but also small enough to preserve the focus to the patterns with small consensus. To solve the dilemma, we use the following heuristic: choose smallest s, for which $U(s)$ covers a given portion (5-10%) of the validation data, i.e.,

$$s = \min\{\bar{s}; |U(\bar{s})| \geq \alpha|\mathcal{V}|\}, \text{ where } \alpha \in (0, 1]. \tag{12}$$

3.2 Similarity to the Oracle Confidence Measure

The first approach how $\mathcal{F}(\phi, \kappa_\phi, \mathcal{V})$ can be measured is to compute the similarity of values $\kappa_\phi(\mathbf{x})$ to the values of the Oracle confidence $\kappa_\phi^{(OR)}(\mathbf{x})$ for patterns $(\mathbf{x}, c_{\mathbf{x}}) \in \mathcal{V}$, where $\mathcal{V}$ is the (restricted) validation set. In this paper, we measured the similarity with Mean Absolute Error (average absolute value of the differences of the confidences):

$$\mathcal{F}^{(SOR)}(\phi, \kappa_\phi, \mathcal{V}) = 1 - \frac{\sum\limits_{(\mathbf{x}, c_{\mathbf{x}}) \in \mathcal{V}} |\kappa_\phi(\mathbf{x}) - \kappa_\phi^{(OR)}(\mathbf{x})|}{|\mathcal{V}|}. \tag{13}$$



(a) ELA – bad separation (b) ELM – relatively good separation

Fig. 2 *The OK (light) and NOK (dark) histograms of the ELA and ELM confidence measures of a Random Forest ensemble for the Waveform dataset.*

3.3 Area Under ROC curve for OK/NOK Histogram

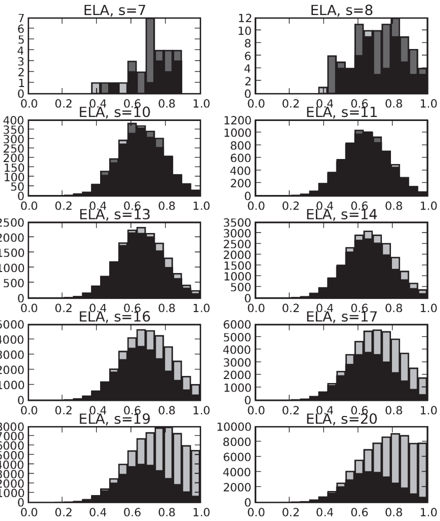
The second approach how $\mathcal{F}(\phi, \kappa_\phi, \mathcal{V})$ can be measured is to analyze histograms of $\kappa_\phi(\mathbf{x})$ for patterns classified correctly by ϕ (*OK patterns*) and for patterns classified incorrectly by ϕ (*NOK patterns*). Values of $\kappa_\phi(\mathbf{x})$ for the NOK patterns should be concentrated near 0, while for the OK patterns, $\kappa_\phi(\mathbf{x})$ should concentrate near 1. Moreover, these two distributions should not overlap.

Let $\mathcal{V}$ be the (restricted) validation set, and let $\mathcal{V}_i \subseteq \mathcal{V}$ for $i = 1, \ldots, N$ denote the sets of validation patterns from class C_i. For two arbitrary classes C_k, C_j, we define the multiset

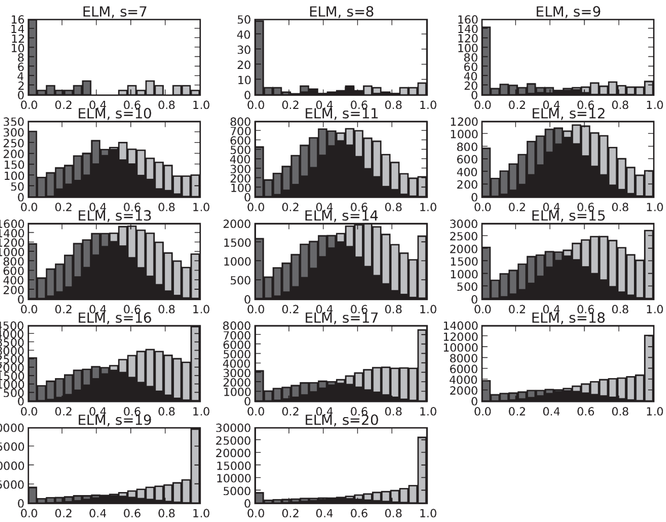
$$H_{kj} = \{\kappa_\phi(\mathbf{x}) | (\mathbf{x}, c_\mathbf{x}) \in \mathcal{V}_k,\ \phi^{(cr)}(\mathbf{x}) = j\}, \tag{14}$$

as a multiset of classification confidence values for all validation patterns from class C_k which have been classified to class C_j by ϕ. Using this notation, we can define the *OK histogram* as the histogram computed from $\bigcup_k H_{kk},\ k = 1, \ldots, N$ and the *NOK histogram* as the histogram computed from $\bigcup_{k \neq j} H_{kj},\ k, j = 1, \ldots, N$.

As an example, the OK and NOK histograms of the ELA and ELM confidence measures for a Random Forest ensemble for the Waveform dataset [9] are shown in Fig. 2 (the figure is computed using all the patterns in the dataset, i.e., the validation set is not restricted). Figs. 3a and 3b show the evolution of the histograms for the restricted validation set. The data have been collected from the experiment described in the following section. Observe that for lower s, the histograms are very different from the histograms for higher values of s. More specifically, for the ELA confidence measure, the histograms for small values of s are totally overlapping, which indicates that the performance of the confidence measure in a dynamic classifier system will be poor (for patterns with no consensus, it does not predict the degree of trust in the classification, and for patterns with consensus, the breaking of the consensus is very hard, anyway). On the other hand, for the ELM confidence measure, the OK and NOK histograms for small values of s are separated, which means that this confidence measure will perform much better in a dynamic classifier systems.



(a) ELA



(b) ELM

Fig. 3 *The restricted OK (light) and NOK (dark) histograms of the ELA and ELM confidence measures of a Random Forest ensemble for the Waveform dataset for* $s = 7, \ldots, 20$.

Altough the OK/NOK (restricted) histograms give us visual information about the feasibility of the confidence measure, we would like to evaluate the degree of overlapping using a single number. This is possible, if we represent the OK/NOK confidence values by a ROC curve, and then we compute the area under the ROC curve (for the sake of simplicity, we will use the well-known area under ROC curve in this paper, regardless of its criticism given in [15]; on the other hand, any other measure of OK/NOK classifier performance could be used, including the modification of the AUC measure presented in [15]).

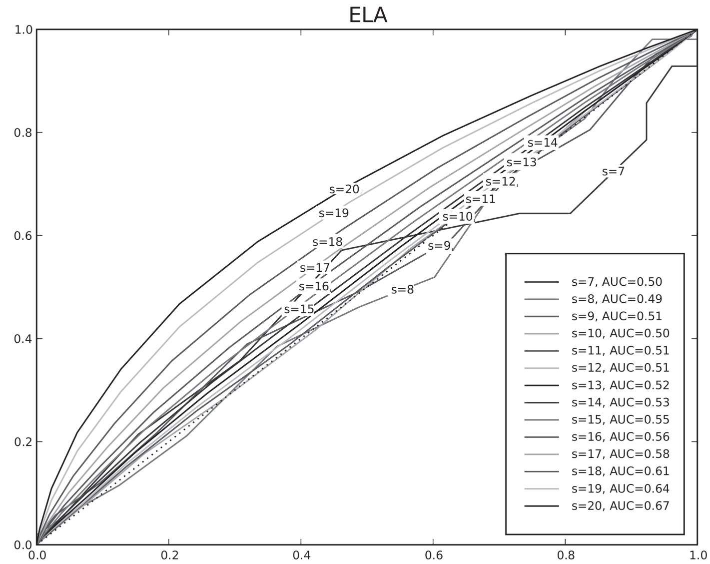
Receiver operating characteristic (ROC) curves [11] are a standard tool in data mining and machine learning. ROC is basically a plot of the fraction of true positives vs. the fraction of false positives of a binary classifier, as some parameter is being varied (e.g., the discrimination threshold of the classifier). If a classifier assigns patterns to classes entirely at random, its ROC curve is the diagonal. On the other hand, for an ideal classifier, the ROC curve consists only of one point $(0, 1)$. The closer we are to the ROC of the ideal classifier (i.e., the farther the ROC curve is from the diagonal (above the diagonal)), the better discrimination of the classifier. The strong point of the ROC curve aprroach is that we can summarize the ROC curve into a single number – area under ROC curve (AUC) – which can be used as a criterion of the quality of a binary classifier. For a random classifier, AUC=0.5; for the ideal classifier, AUC=1. The higher the AUC, the better discrimination of the classifier. Classifiers with AUC below 0.5 are actually *worse* than a random classifier.

In the context of classification confidence, we will study the AUC of a so-called *OK/NOK classifier*, which assigns a pattern to the class "correctly classified" if the classification confidence is higher than some threshold T, and to the class "incorrectly classified" instead. By varying T between 0 and 1, we obtain the ROC curve of a particular classifier, representing the quality of the separation of the OK/NOK histograms. The AUC of the OK/NOK classifier measured on a validation set $\mathcal{V}$ (or, on a restricted set $U(s)$) can be used as an empirical property expressing the degree of overlapping of the OK and NOK distributions. Now we can define $\mathcal{F}^{(AUC)}(\phi, \kappa_\phi, \mathcal{V})$ as the AUC of the OK/NOK classifier for the confidence κ_ϕ, measured on $\mathcal{V}$. Figs. 4a and 4b show an example of the ROCs for the ELA and ELM confidence measures for a Random Forest ensemble for the Waveform dataset.

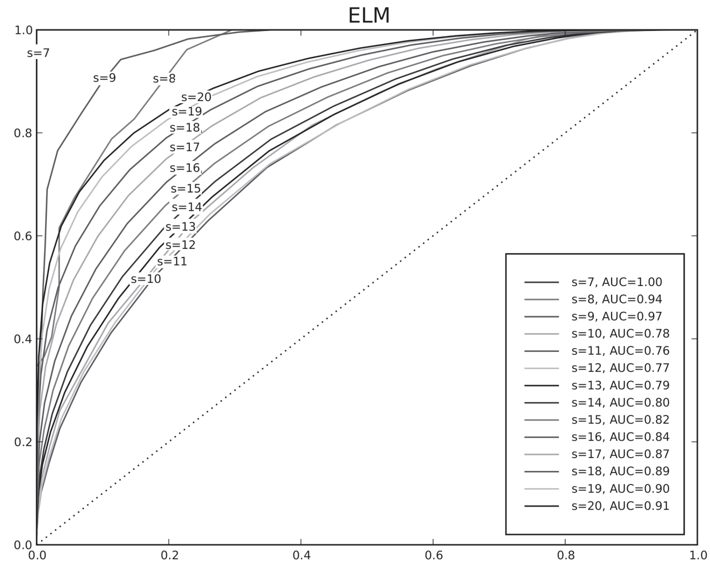
4. Experiment: Measuring the Feasibility of a Confidence Measure

To find out whether the methods for assessing confidence measures described in the previous sections can really predict the improvement in the classification quality of a dynamic classifier system, we designed the following experiment.

Suppose we have a classifier team $(\Gamma, \mathcal{K})$. Given a validation dataset $\mathcal{V}$, we put apart 20% of the data, denoted as $\mathcal{V}^1$, to measure $\mathcal{F}(\Gamma, \mathcal{K}, \mathcal{V}^1)$ using 5-fold crossvalidation on $\mathcal{V}^1$. After that, we use the remaining 80% of the data from $\mathcal{V}$, denoted as $\mathcal{V}^2$, to measure the relative improvement of the error rate of a dynamic classifier system (aggregated by the weighted mean aggregator with the



(a) ELA



(b) ELM

Fig. 4 *The ROC curves and the AUCs of the OK/NOK classifiers of the ELA and ELM confidence measures for the Waveform dataset, measured on* $U(s)$, $s = 7, \ldots, 20$, *for a Random Forest ensemble.*

particular dynamic confidence measure) compared to the error rate of a confidence-free classifier system (aggregated by the mean value aggregator), using 10-fold crossvalidation on $\mathcal{V}^2$. The relative improvement in the mean error rate will be computed as:

$$\mathcal{I}(S_1, S_2) = \frac{Err(S_1) - Err(S_2)}{Err(S_1)}, \tag{15}$$

where $Err(S_1)$ denotes the error rate of the reference classifier system (using the mean value aggregator), and $Err(S_2)$ denotes the error rate of the dynamic classifier system (using weighted mean aggregator). However, if the dataset $\mathcal{V}$ was too small (consisted of less than 500 patterns), we did not divide $\mathcal{V}$ to $\mathcal{V}^1$ and $\mathcal{V}^2$, i.e., $\mathcal{V}^1 = \mathcal{V}^2 = \mathcal{V}$, and thus both $\mathcal{F}$ and $\mathcal{I}$ were measured on the whole dataset $\mathcal{V}$.

Our goal in this experiment is to study the correlation between $\mathcal{F}$ and $\mathcal{I}$. We performed the experiment on 6 artificial and 20 real-world datasets from the Elena database [24] and from the UCI repository [9] (cf. Tab. II). The classifier teams were created using the Random Forest method [4], and as the classification confidences we used ELA, ELM, RELA, and RELM. For reference purposes, we also used the Oracle confidence measure (for which $\mathcal{F} = 1$ by definition). For assessing the confidence measures, we used methods described in the previous section, i.e., similarity to the Oracle confidence (SOR) and the area under ROC curve of the OK/NOK classifier (AUC), measured on the restricted validation set $U(s)$, for s such that $U(s)$ covers 10% of the data. As the similarity threshold parameter for RELA and RELM, we used a constant value $T = 0.5$. All the methods were run using the same random seed, so when a pattern was classified, all the methods were using the same data.

In the experiment, we classified the data using the following models:

- single-best classifier (SB) – result of the best single classifier in the classifier team, representing a non-combined classifier
- mean value aggregation (MV) – representing a confidence-free classifier system
- (static) weighted mean using global accuracy (WM-GA) – representing a static classifier system
- (dynamic) weighted mean (WM) using ELA, ELM, RELA, RELM, OR confidence measures – representing a dynamic classifier system

Classification error rates (mean value and standard deviation of the error rates from 10-fold crossvalidation) are shown in Appendix A, Tab. A. Tab. I shows a comparison of the performance of dynamic classifier systems (aggregated using WM) using different dynamic confidence measures, compared to the performance of confidence-free classifier systems (aggregated using MV). From these results, we can see that, in general, dynamic classifier systems outperform confidence-free classifier systems. Another interesting result is that the restricted versions of ELA and ELM confidence measures obtained better results than the ordinary ELA and ELM confidence measures.

However, the main goal of the experiment was to study the correlation between the feasibility of a particular confidence measure ($\mathcal{F}$) and the improvement in the

Conf. measure	WM better	MV better	Tie
ELA	13	6	7
ELM	15	10	1
RELA	16	6	4
RELM	19	7	0
OR	26	0	0

Tab. I *Comparison of the performance of dynamic classifier systems vs. confidence-free classifier systems. The table shows the number of datasets for which a WM aggregator using a particular dynamic confidence measure obtained better/worse/same mean error rate as the MV aggregator (ties are defined as the same error rate up to first decimal place).*

performance of a dynamic classifier system (aggregated by WM using dynamic confidence measure) compared to the performance of a confidence-free classifier system (aggregated by MV) ($\mathcal{I}$).

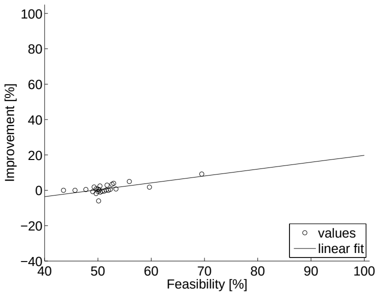
For each feasibility measure, we obtained a scatterplot of ($\mathcal{F}$,$\mathcal{I}$) values (for the 26 datasets) which are shown in Figs. 5 and 6, including the least-squares linear approximation of the scatterplot. To measure the correlation between $\mathcal{F}$ and $\mathcal{I}$, we computed Pearson's and Spearman's rank correlation coefficients, and tested their significance. The results of the correlation tests are shown in Tabs. 5b and 6b.

4.1 Results Discussion

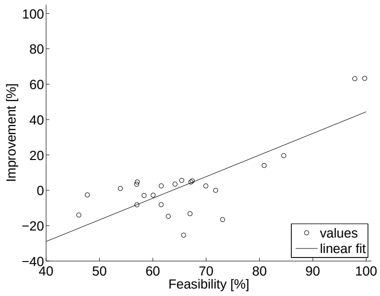
For $\mathcal{F}^{(SOR)}$, the scatterplot shows a statistically significant correlation between $\mathcal{F}$ and $\mathcal{I}$ for the ELA, ELM, and RELM confidence measures. For the RELA confidence measure, Spearman's test was not statistically significant; however, the least-squares fit in the figure shows a well-fitting linear dependency, and Pearson's test was also statistically signigicant. On the other hand, Pearson's and Spearman's tests for the ELA confidence measure are highly significant ($< 1\%$), but because the $\mathcal{F}^{(SOR)}$ values are clustered in the area around $\mathcal{F}^{(SOR)} = 50\%$, the least-squares fit does not indicate strong linear dependency.

For $\mathcal{F}^{(AUC)}$, the results are quite similar – for the ELM, RELA, and RELM confidence measures, both Pearson's and Spearman's tests, and also the least-squares fit indicate a strong correlation between $\mathcal{F}^{(AUC)}$ and $\mathcal{I}$. For the ELA confidence measure, the correlation is not clear, neither from the tests, nor from the least-squares fit (again, the $\mathcal{F}^{(AUC)}$ values are clustered around $\mathcal{F}^{(AUC)} = 50\%$).

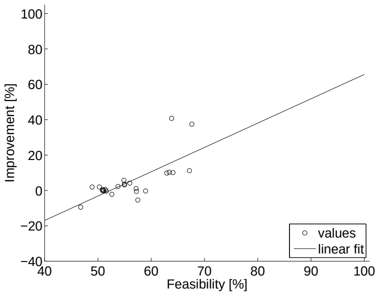
In general, we can say that these results indicate that the methods for assessing confidence measures described in the previous section (SOR, AUC), computed on the restricted validation sets of 10% most-unconsensed values, could be used for predicting whether using a dynamic classifier system instead of a confidence-free system would bring improvement in the error rate. Moreover, the methods can also be used for predicting which confidence measure will perform the best for a given classifier type and a given dataset.



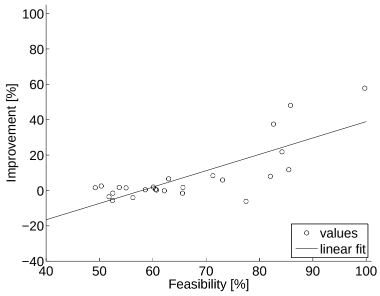
Scatterplot of $\mathcal{I}$ versus $\mathcal{F}$, ELA



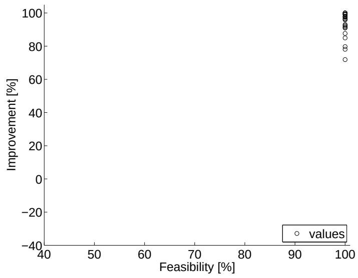
Scatterplot of $\mathcal{I}$ versus $\mathcal{F}$, ELM



Scatterplot of $\mathcal{I}$ versus $\mathcal{F}$, RELA



Scatterplot of $\mathcal{I}$ versus $\mathcal{F}$, RELM

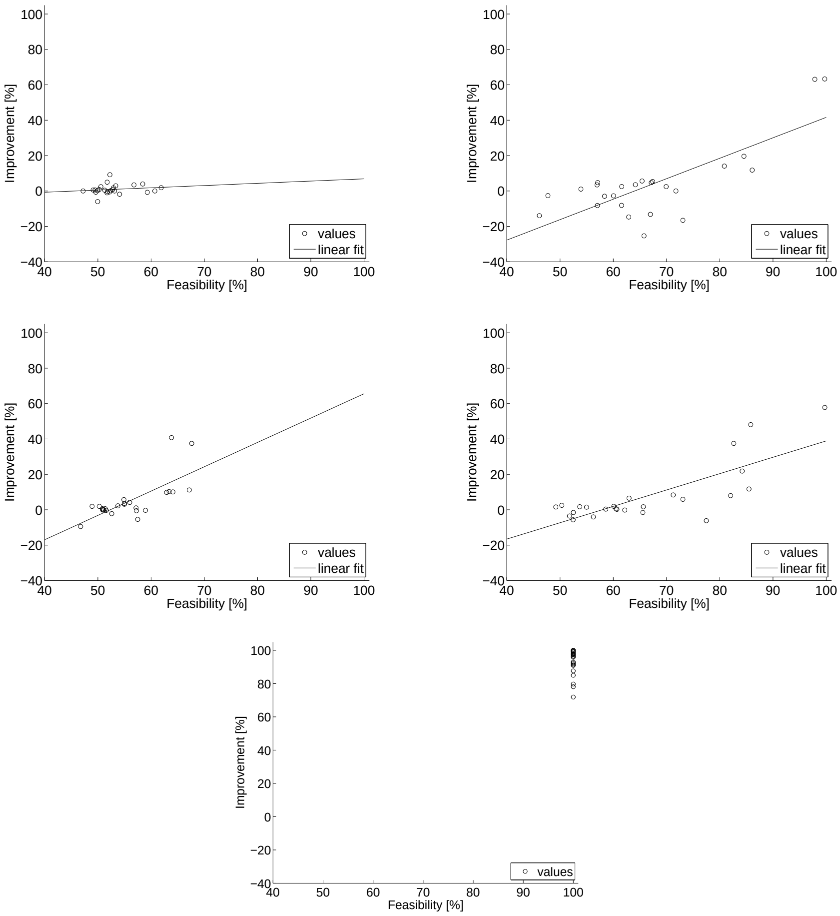


Scatterplot of $\mathcal{I}$ versus $\mathcal{F}$, OR

	Pearson			Spearman		
Conf. measure	ρ	p [%]	significant	ρ	p [%]	significant
ELA	0.68	0.01	yes	0.55	0.4	yes
ELM	0.67	0.02	yes	0.41	4.4	yes
RELA	0.45	2.3	yes	0.33	10.4	no
RELM	0.72	0.003	yes	0.54	0.5	yes

(b) Pearson's correlation and Spearman's rank correlation tests. ρ denotes the correlation coefficient of the sample and p denotes the statistical significance of the test. The significance is evaluated at 5% level.

Fig. 5 *Experimental results for the Similarity to Oracle (SOR) method, for restricted validation set $U(s)$, covering 10% of the validation data for the ELA, ELM, RELA, RELM, and OR dynamic confidence measures.*



Conf. measure	Pearson ρ	Pearson p [%]	Pearson significant	Spearman ρ	Spearman p [%]	Spearman significant
ELA	0.17	41	no	0.19	37.5	no
ELM	0.76	0.0008	yes	0.51	0.9	yes
RELA	0.72	0.005	yes	0.58	0.2	yes
RELM	0.78	0.0004	yes	0.63	0.1	yes

(b) Pearson's correlation and Spearman's rank correlation tests. ρ denotes the correlation coefficient of the sample and p denotes the statistical significance of the test. The significance is evaluated at 5% level.

Fig. 6 *Experimental results for the Area Under ROC Curve (AUC) method, for restricted validation set $U(s)$, covering 10% of the validation data for the ELA, ELM, RELA, RELM, and OR dynamic confidence measures.*

5. Summary

In this paper, we dealt with dynamic classification confidence measures in classifier aggregation. We discussed the properties which should hold for a good confidence measure, and we studied the performance of dynamic confidence measures in the context of the degree of consensus in the classifier team. As the results show, the properties of the confidence measures are important mainly for patterns with a small degree of consensus only. This lead to the definition of two measures of feasibility of a given classification confidence measure to a given classifier team and given data.

In the experimental section, we have empirically shown that for both methods, there is a statistically significant correlation between the feasibility, and the actual improvement of the accuracy of the classifier system. This suggests that both proposed feasibility measures can be used in practical applications to choose an optimal confidence measure for a given application setup.

Acknowledgment

The research reported in this paper has been supported by the Czech Science Foundation (GA ČR) grant 13-17187S.

References

[1] Matti Aksela. Comparison of classifier selection methods for improving committee performance. In: Multiple Classifier Systems, pages 84–93, 2003.

[2] Ran Avnimelech and Nathan Intrator. Boosted mixture of experts: An ensemble learning scheme. Neural Computation, 11(2):483–497, 1999.

[3] Zoran Bosnic and Igor Kononenko. Estimation of individual prediction reliability using the local sensitivity analysis. Appl. Intell., 29(3):187–203, 2008.

[4] Leo Breiman. Random forests. Machine Learning, 45(1):5–32, 2001.

[5] W. Cheetham and J. Price. Measures of Solution Accuracy in Case-Based Reasoning Systems. In P. A. Gonzalez Calero and P. Funk, editors, Proceedings of the European Conference on Case-Based Reasoning (ECCBR-04), pages 106–118. Springer, 2004. Madrid, Spain.

[6] William Cheetham. Case-based reasoning with confidence. In EWCBR '00: Proceedings of the 5th European Workshop on Advances in Case-Based Reasoning, pages 15–25, London, UK, 2000. Springer-Verlag.

[7] Sarah Jane Delany, Padraig Cunningham, Dónal Doyle, and Anton Zamolotskikh. Generating estimates of classification confidence for a case-based spam filter. In Héctor Muñoz-Avila and Francesco Ricci, editors, Case-Based Reasoning, Research and Development, 6th Int. Conf., ICCBR 2005, Chicago, USA, volume 3620 of LNCS, pages 177–190. Springer, 2005.

[8] Luca Didaci, Giorgio Giacinto, Fabio Roli, and Gian Luca Marcialis. A study on the performances of dynamic classifier selection based on local accuracy estimation. Pattern Recognition, 38(11):2188–2191, 2005.

[9] C.L. Blake D.J. Newman, S. Hettich and C.J. Merz. UCI repository of machine learning databases, online. `http://www.ics.uci.edu/~mlearn/MLRepository.html`.

[10] Richard O. Duda, Peter E. Hart, and David G. Stork. Pattern Classification (2nd Edition). Wiley-Interscience, 2000.

[11] Tom Fawcett. An introduction to ROC analysis. Pattern Recogn. Lett., 27(8):861–874, 2006.

[12] Giorgio Giacinto and Fabio Roli. Dynamic classifier selection based on multiple classifier behaviour. Pattern Recognition, 34(9):1879–1881, 2001.

[13] S. I. Gurov. Reliability estimation of classification algorithms I-III. Computational Mathematics and Modeling, 15,15,16(4,2,3):365–376,169–178,279–288, 2004-2005.

[14] David J. Hand. Construction and Assessment of Classification Rules. Wiley, 1997.

[15] David J. Hand. Measuring classifier performance: a coherent alternative to the area under the roc curve. Mach. Learn., 77(1):103–123, October 2009.

[16] Josef Kittler, Mohamad Hatef, Robert P. W. Duin, and Jiri Matas. On combining classifiers. IEEE Trans. Pattern Anal. Mach. Intell., 20(3):226–239, 1998.

[17] Albert H. R. Ko, Robert Sabourin, and Alceu Souza Britto, Jr. From dynamic classifier selection to dynamic ensemble selection. Pattern Recogn., 41(5):1718–1731, 2008.

[18] Ludmila I. Kuncheva. Combining Pattern Classifiers: Methods and Algorithms. Wiley-Interscience, 2004.

[19] Ludmila I. Kuncheva, James C. Bezdek, and Robert P. W. Duin. Decision templates for multiple classifier fusion: an experimental comparison. Pattern Recognition, 34(2):299–314, 2001.

[20] Marko Robnik-Šikonja. Improving random forests. In J. Boulicaut, F. Esposito, F. Giannotti, and D. Pedreschi, editors, ECML, volume 3201 of Lecture Notes in Computer Science, pages 359–370. Springer, 2004.

[21] Lior Rokach. Taxonomy for characterizing ensemble methods in classification tasks: A review and annotated bibliography. Comput. Stat. Data Anal., 53(12):4046–4072, 2009.

[22] Dymitr Ruta and Bogdan Gabrys. An overview of classifier fusion methods. Computing and Information Systems, 7:1–10, 2000.

[23] Alexey Tsymbal, Mykola Pechenizkiy, and Padraig Cunningham. Dynamic integration with random forests. In J. FĂĽrnkranz, T. Scheffer, and M. Spiliopoulou, editors, ECML, volume 4212 of Lecture Notes in Computer Science, pages 801–808. Springer, 2006.

[24] UCL MLG. Elena database, online. `http://www.dice.ucl.ac.be/mlg/?page=Elena`.

[25] Giorgio Valentini. Ensemble Methods Based on Bias-Variance Analysis. PhD thesis, DISI, Universita di Genova, 2003.

[26] David Štefka and Martin Holeňa. Classifier aggregation using local classification confidence. In: Proceedings of the International Conference on Agents and Artificial Intelligence, ICAART 2009, Porto, Portugal, pages 173–178, 2009.

[27] David Štefka and Martin Holeňa. Dynamic classifier systems and their applications to random forest ensembles. In: Proceedings of the 9th International Conference on Adaptive and Natural Computing Algorithms, ICANNGA'09, pages 458–468, Berlin, Heidelberg, 2009. Springer-Verlag.

[28] D. Randall Wilson and Tony R. Martinez. Combining cross-validation and confidence to measure fitness. In: Proceedings of the International Joint Conference on Neural Networks (IJCNN'99), paper 163, 1999.

[29] Kevin Woods, Jr. W. Philip Kegelmeyer, and Kevin Bowyer. Combination of multiple classifiers using local accuracy estimates. IEEE Trans. Pattern Anal. Mach. Intell., 19(4):405–410, 1997.

[30] Xingquan Zhu, Xindong Wu, and Ying Yang. Dynamic classifier selection for effective mining from noisy data streams. In: ICDM '04: Proceedings of the Fourth IEEE International Conference on Data Mining (ICDM'04), pages 305–312, Washington, DC, USA, 2004. IEEE Computer Society.

A Detailed Results of the Experiment

Dataset	ref	patterns	classes	dim.
Artificial				
clouds	[24]	5000	2	2
concentric	[24]	2500	2	2
gauss 3D	[24]	5000	2	3
gauss 8D	[24]	5000	2	8
twonorm	[9]	3000	2	20
waveform	[9]	5000	3	21
Real-world				
balance	[9]	625	3	9
breast	[9]	699	2	9
glass	[9]	214	7	9
iris	[9]	150	3	4
letter-recg.	[9]	20000	26	16
pendigits	[9]	10992	10	16
phoneme	[24]	5427	2	5
pima	[9]	768	2	8
poker	[9]	4828	3	10
satimage	[24]	6435	6	4
segmentation	[9]	2310	7	16
sonar	[9]	208	2	10
texture	[24]	5500	11	10
transfusion	[9]	748	2	4
vehicle	[9]	946	4	18
vowel	[9]	990	11	10
wine	[9]	178	3	13
wineq-red	[9]	1600	3	11
wineq-white	[9]	4898	3	11
yeast	[9]	1484	4	8

Tab. II *Datasets used in the experiment.*

Dataset	Non-Combined SB	Conf.-free MV	Static κ	SWM	Dynamic κ	DWM
Artificial						
clouds	13.3 ± 1.4	11.9 ± 1.9	GA	11.9 ± 2.0	ELA	12.0 ± 1.9
					ELM	$\mathbf{11.7 \pm 2.0}$
					RELA	11.8 ± 2.0
					RELM	11.8 ± 2.2
					OR	1.5 ± 0.8
concentric	7.0 ± 1.4	2.6 ± 1.4	GA	2.6 ± 1.4	ELA	2.5 ± 1.4
					ELM	$\mathbf{2.1 \pm 1.1}$
					RELA	2.3 ± 1.3
					RELM	2.4 ± 1.2
					OR	0.1 ± 0.2
gauss_3D	28.7 ± 2.4	23.9 ± 1.4	GA	23.9 ± 1.4	ELA	23.9 ± 1.2
					ELM	$\mathbf{22.8 \pm 1.5}$
					RELA	23.9 ± 1.3
					RELM	23.6 ± 1.5
					OR	2.1 ± 0.4

Dataset	Non-Combined SB	Conf.-free MV	Static κ	SWM	Dynamic κ	DWM
gauss_8D	24.7 ± 1.6	$\mathbf{14.5 \pm 0.8}$	GA	14.6 ± 0.7	ELA	14.6 ± 0.9
					ELM	16.5 ± 1.8
					RELA	14.6 ± 1.4
					RELM	14.7 ± 1.2
					OR	0.1 ± 0.2
twonorm	21.3 ± 2.7	8.6 ± 0.8	GA	8.5 ± 0.8	ELA	8.3 ± 0.9
					ELM	$\mathbf{3.2 \pm 1.0}$
					RELA	7.6 ± 0.7
					RELM	6.7 ± 0.8
					OR	0.0 ± 0.0
waveform	27.1 ± 1.3	18.0 ± 1.6	GA	18.0 ± 1.6	ELA	17.8 ± 1.4
					ELM	$\mathbf{15.4 \pm 1.5}$
					RELA	17.8 ± 1.4
					RELM	16.8 ± 0.8
					OR	0.1 ± 0.2
Real-world						
balance	20.7 ± 5.2	13.5 ± 6.4	GA	13.6 ± 6.3	ELA	13.5 ± 6.1
					ELM	$\mathbf{11.9 \pm 5.7}$
					RELA	14.7 ± 5.0
					RELM	13.9 ± 5.0
					OR	2.7 ± 2.5
breast	6.6 ± 3.6	3.6 ± 3.4	GA	3.6 ± 3.4	ELA	3.5 ± 3.2
					ELM	4.6 ± 2.6
					RELA	3.7 ± 2.0
					RELM	$\mathbf{3.2 \pm 1.9}$
					OR	0.5 ± 0.8
glass	24.9 ± 5.8	19.5 ± 8.3	GA	20.2 ± 8.2	ELA	18.8 ± 8.5
					ELM	18.8 ± 9.6
					RELA	$\mathbf{18.7 \pm 9.5}$
					RELM	19.4 ± 9.3
					OR	0.7 ± 2.1
iris	9.3 ± 6.1	6.7 ± 6.7	GA	6.7 ± 6.7	ELA	6.7 ± 6.7
					ELM	$\mathbf{4.7 \pm 4.3}$
					RELA	7.3 ± 7.0
					RELM	5.3 ± 5.0
					OR	0.0 ± 0.0
letter-recg	21.7 ± 1.4	7.4 ± 1.0	GA	7.4 ± 1.0	ELA	7.3 ± 1.0
					ELM	7.2 ± 1.0
					RELA	7.0 ± 0.9
					RELM	$\mathbf{6.8 \pm 1.0}$
					OR	0.6 ± 0.2
pendigits	7.6 ± 0.7	2.0 ± 0.4	GA	2.0 ± 0.4	ELA	2.0 ± 0.4
					ELM	2.3 ± 0.7
					RELA	$\mathbf{1.8 \pm 0.4}$
					RELM	$\mathbf{1.8 \pm 0.6}$
					OR	0.1 ± 0.1
phoneme	19.2 ± 1.3	13.3 ± 0.9	GA	13.3 ± 0.9	ELA	13.0 ± 1.0
					ELM	13.7 ± 1.0
					RELA	$\mathbf{12.9 \pm 0.9}$
					RELM	13.4 ± 0.8
					OR	0.6 ± 0.4
pima	30.3 ± 5.9	24.7 ± 3.4	GA	24.7 ± 3.4	ELA	25.0 ± 3.5
					ELM	24.5 ± 3.5
					RELA	24.7 ± 3.2
					RELM	$\mathbf{24.4 \pm 3.6}$
					OR	0.7 ± 0.9
poker	50.1 ± 2.3	45.9 ± 1.8	GA	45.9 ± 1.7	ELA	45.7 ± 2.5
					ELM	$\mathbf{43.7 \pm 2.4}$
					RELA	45.0 ± 2.1
					RELM	44.8 ± 2.2
					OR	3.7 ± 1.2

Dataset	Non-Combined SB	Conf.-free MV	Static κ	Static SWM	Dynamic κ	Dynamic DWM
satimage	16.8 ± 1.4	$\mathbf{13.9 \pm 1.1}$	GA	13.9 ± 1.2	ELA	14.1 ± 1.2
					ELM	$\mathbf{13.9 \pm 1.1}$
					RELA	$\mathbf{13.9 \pm 1.4}$
					RELM	14.1 ± 1.3
					OR	3.0 ± 0.6
segmentation	13.2 ± 2.9	7.7 ± 1.9	GA	$\mathbf{7.6 \pm 1.9}$	ELA	7.7 ± 1.8
					ELM	8.8 ± 2.6
					RELA	8.1 ± 1.6
					RELM	8.2 ± 1.6
					OR	0.5 ± 0.6
sonar	35.2 ± 7.0	$\mathbf{24.1 \pm 13.6}$	GA	24.6 ± 12.6	ELA	25.5 ± 13.9
					ELM	26.1 ± 14.3
					RELA	$\mathbf{24.1 \pm 13.2}$
					RELM	25.1 ± 14.0
					OR	0.0 ± 0.0
texture	13.1 ± 2.4	2.5 ± 0.7	GA	2.5 ± 0.7	ELA	2.4 ± 0.6
					ELM	$\mathbf{0.9 \pm 0.3}$
					RELA	2.2 ± 0.6
					RELM	1.0 ± 0.4
					OR	0.0 ± 0.0
transfusion	25.0 ± 3.9	23.8 ± 3.2	GA	23.8 ± 3.2	ELA	23.9 ± 3.2
					ELM	$\mathbf{22.5 \pm 4.0}$
					RELA	23.7 ± 3.6
					RELM	23.4 ± 3.5
					OR	6.7 ± 1.7
vehicle	35.5 ± 6.0	27.1 ± 6.8	GA	26.9 ± 6.4	ELA	27.0 ± 6.5
					ELM	27.8 ± 6.4
					RELA	$\mathbf{26.5 \pm 6.3}$
					RELM	28.6 ± 7.0
					OR	0.6 ± 0.6
vowel	45.2 ± 3.8	16.5 ± 3.2	GA	16.4 ± 3.6	ELA	15.0 ± 3.9
					ELM	15.7 ± 4.1
					RELA	9.8 ± 1.7
					RELM	$\mathbf{8.6 \pm 2.8}$
					OR	0.1 ± 0.4
wine	14.0 ± 10.6	4.4 ± 6.0	GA	3.9 ± 5.6	ELA	4.4 ± 6.0
					ELM	5.0 ± 4.6
					RELA	$\mathbf{2.8 \pm 4.5}$
					RELM	$\mathbf{2.8 \pm 4.5}$
					OR	0.0 ± 0.0
wineq-red	40.7 ± 4.9	28.8 ± 5.5	GA	29.1 ± 5.3	ELA	28.6 ± 5.1
					ELM	31.1 ± 4.5
					RELA	$\mathbf{28.2 \pm 5.0}$
					RELM	$\mathbf{28.2 \pm 5.3}$
					OR	0.3 ± 0.5
wineq-white	45.9 ± 3.3	34.2 ± 2.4	GA	34.2 ± 2.5	ELA	34.2 ± 2.4
					ELM	35.1 ± 2.5
					RELA	$\mathbf{33.0 \pm 2.5}$
					RELM	34.1 ± 2.2
					OR	0.8 ± 0.6
yeast	46.9 ± 2.6	36.6 ± 3.4	GA	36.4 ± 3.4	ELA	36.4 ± 3.1
					ELM	$\mathbf{35.4 \pm 1.8}$
					RELA	37.5 ± 2.6
					RELM	36.4 ± 2.7
					OR	3.3 ± 1.1

Tab. III *Mean value ± standard deviation of the classifier error rates (in %) from 10-fold crossvalidation. The best method (lowest mean error rate, excluding DWM-OR) for each dataset is displayed in boldface.*