CLASSIFIER AGGREGATION USING LOCAL CLASSIFICATION CONFIDENCE

David Štefka and Martin Holeňa
Institute of Computer Science, Academy of Sciences of the Czech Republic
v.v.i., Pod Vodárenskou věží 2, Prague, Czech Republic

Keywords: Classifier aggregation, Classifier combining, Classification confidence.

Abstract: Classifier aggregation is a method for improving quality of classification. Instead of using just one classifier, a team of classifiers is created, and the outputs of the individual classifiers are aggregated into the final prediction. In this paper, we study the potential of using measures of local classification confidence in classifier aggregation methods. We introduce four measures of local classification confidence and study their suitability for classifier aggregation. We develop two novel classifier aggregation methods which utilize local classification confidence and we compare them to two commonly used methods for classifier aggregation. The results on four artificial and five real-world benchmark datasets show that by incorporating local classification confidence into classifier aggregation methods, significant improvement in classification quality can be obtained.

1 INTRODUCTION

Classification is a process of dividing *patterns* into disjoint sets called *classes*. Many machine learning algorithms for classification have been developed – for example naive Bayes classifiers, linear and quadratic discriminant classifiers, *k*-nearest neighbor classifiers, support vector machines, neural networks, or decision trees (Duda et al., 2000).

One comonly used technique for improving classification quality is called *classifier combining* (Kuncheva, 2004) – instead of using just one classifier, a team of classifiers is created, and their results are then combined. It can be shown that a team of classifiers can perform better than any of the individual classifiers in the team.

There are two main approaches to classifier combining: *classifier selection* (Woods et al., 1997; Aksela, 2003; Zhu et al., 2004) and *classifier aggregation* (Kittler et al., 1998; Kuncheva et al., 2001). If a pattern is submitted for classification, the former technique uses some rule to select one particular classifier, and only this classifier is used to obtain the final prediction. The latter technique uses some aggregation rule to aggregate the results of all the classifiers in the team to get the final prediction.

A common drawback of classifier aggregation methods is that they are global, i.e., they are not adapted to the particular patterns that are currently classified. However, if we use the concept of local classification confidence (i.e., the extent to which we can "trust" the output of the particular classifier for the currently classified pattern), the algorithm can take into account the fact that "this classifier is/is not good *for this particular pattern*".

Surprisingly, using local classification confidence is not very common in classifier combining. The goal of this paper is to provide basic introduction to local classification confidence measures, to study which particular measures are suitable for classifier aggregation, and to create novel classifier aggregation algorithms which improve the quality of classification by using local classification confidence.

The paper is structured as follows.After discussing motivations in Section 2, we provide a formalism of classifier combining in Section 3. In Section 4 we introduce four local classification confidence measures and we study their suitability with quadratic discriminant classifiers. In Section 5, classifier aggregation methods which utilize these measures are introduced and their performance is tested on 9 benchmark datasets. Section 6 then concludes the paper.

2 MOTIVATION

In the field of classification, several methods for assessing the quality of classification exist (Hand, 1997). Some of these methods try to measure the *classification confidence*, i.e., the probability that the

Štefka D. and Holeňa M. (2009).
CLASSIFIER AGGREGATION USING LOCAL CLASSIFICATION CONFIDENCE.
In *Proceedings of the International Conference on Agents and Artificial Intelligence*, pages 173-178
DOI: 10.5220/0001545101730178
Copyright © SciTePress

classifier predicts correctly, or the degree of trust we can give to this classifier.

Classification confidence measures can be divided into two main groups: *global measures*, which assess the classifier's predictive power as a whole (Hand, 1997; Duda et al., 2000), and *local measures*, which adapt themselves to the particular pattern submitted for classification (Woods et al., 1997; Cheetham and Price, 2004; Robnik-Šikonja, 2004; Delany et al., 2005; Tsymbal et al., 2006). Examples of global measures can be accuracy, precision, sensitivity, resemblance, etc.; these have been studied intensively since the formation of the theory of classification. Local classification confidence measures have not been studied as exhaustively as global measures so far; however, we believe they can express the classification confidence better than global measures in the context of classifier combining, where local properties are more important than global ones.

Local classification confidence measures are used for example in case-based reasoning systems (Cheetham and Price, 2004), or in classification tasks where misclassification is less acceptable than refusing to classify the pattern (Delany et al., 2005). Surprisingly, local classification confidence is used scarcely in classifier combining, where we have a battery of different classifiers to use if one classifier refuses to classify the pattern. Even if a method for combining classifiers utilizes a concept of local classification confidence, as in (Woods et al., 1997; Robnik-Šikonja, 2004; Tsymbal et al., 2006), the authors usually choose one particular measure of local classification confidence, and this measure is tightly incorporated into the combining algorithm. However, any other local classification confidence measure could be used. Moreover, creating a unified framework for classifier combining with local classification confidence would be more systematic and general approach.

In this paper, we examine the potential of incorporating the concept of classification confidence to classifier combining algorithms. Firstly, we propose four local classification confidence measures and we examine if they actually predict the probability of correct classification. Secondly, we develop two algorithms for classifier aggregation which utilize local classification confidence and we test their performance on three benchmark datasets.

3 CLASSIFIER COMBINING

Throughout the rest of the paper, we use the following notation. Let $\mathcal{X} \subseteq \mathbf{R}^n$ be a n-dimensional *feature space*, an element $\vec{x} \in \mathcal{X}$ of this space is called a *pattern*, and let $C_1, \ldots, C_N \subseteq \mathcal{X}$ be disjoint sets called *classes*. The goal of classification is to determine to which class a given pattern belongs. We call a *classifier* any mapping $\phi : \mathcal{X} \rightarrow [0,1]^N$, where $\phi(\vec{x}) = (\mu_1, \ldots, \mu_N)$ are *degrees of classification* to each class.

In classifier combining, a team of classifiers $(\phi_1, \ldots, \phi_r)$ is created, each of the classifiers predicts independently, and then the classifiers' outputs are combined into the final prediction. While classifier selection methods use some techniques to determine which classifier is locally better than the others, such algorithms select only one classifier, discarding much potentially useful information, and thus reducing the robustness compared to classifier aggregation. This is the reason why we restrict ourselves to classifier aggregation only in the rest of the paper.

3.1 Ensemble Methods

If a team of classifiers consists only of classifiers of the same type, which differ only in their parameters, dimensionality, or training sets, the team is usually called an *ensemble of classifiers*. For this reason the methods which create a team of classifiers are sometimes called *ensemble methods*. Well-known methods for ensemble creation are *bagging* (Breiman, 1996), *boosting* (Freund and Schapire, 1996), *error correction codes* (Kuncheva, 2004), or *multiple feature subset* methods (Bay, 1999).

3.2 Classifier Aggregation

For classifier aggregation, the output of the team $(\phi_1, \ldots, \phi_r)$ for input pattern $\vec{x}$ can be structured to a $r \times N$ matrix, called *decision profile* (DP):

$$DP(\vec{x}) = \begin{pmatrix} \phi_1(\vec{x}) \\ \phi_2(\vec{x}) \\ \vdots \\ \phi_r(\vec{x}) \end{pmatrix} = \begin{pmatrix} \mu_{1,1} & \mu_{1,2} & \ldots & \mu_{1,N} \\ \mu_{2,1} & \mu_{2,2} & \ldots & \mu_{2,N} \\ & & \ddots & \\ \mu_{r,1} & \mu_{r,2} & \ldots & \mu_{r,N} \end{pmatrix} \quad (1)$$

Many methods for aggregating the team of classifiers into one final classifier have been proposed in the literature. A good overview of commonly used aggregation methods can be found in (Kuncheva et al., 2001). These methods comprise simple arithmetic rules (sum, product, maximum, minimum, average, weighted average, etc.), fuzzy integral, Dempster-Shafer fusion, second-level classifiers, decision templates, and many others.

4 LOCAL CLASSIFICATION CONFIDENCE

Suppose we have a classifier ϕ, and a pattern $\vec{x}$ to classify. The local classification confidence of classifier ϕ on pattern $\vec{x}$ is a real number in the unit interval $[0,1]$, and its value is obtained by a mapping $\kappa_\phi : \mathcal{X} \to [0,1]$.

Local classification confidence can be any property expressing our "trust" in the classifier's prediction for the current pattern $\vec{x}$. In the literature, many methods for assessing local classification confidence can be found (Woods et al., 1997; Wilson and Martinez, 1999; Avnimelech and Intrator, 1999; Delany et al., 2005). Among them, we selected four, which are described in the following subsections.

4.1 Local Accuracy (LA)

The *local accuracy* (LA) (Woods et al., 1997) measures the accuracy of a classifier on a set of neighbors of $\vec{x}$. These neighbors are obtained using a k-NN algorithm, i.e., the accuracy is measured on the set of k neighbors from the validation set closest to $\vec{x}$ with respect to some metric – we will denote this set as $NN_k(\vec{x})$. In this paper, we use Euclidean metric. The LA of a classifier ϕ on $\vec{x}$ is computed as

$$\kappa_\phi(\vec{x}) = \frac{\sum_{y \in NN_k(\vec{x})} c_\phi(\vec{y})}{\#NN_k(\vec{x})} = \frac{\sum_{y \in NN_k(\vec{x})} c_\phi(\vec{y})}{k}, \quad (2)$$

where $c_\phi(\vec{y}) = 1$ if $\vec{y}$ is classified correctly by ϕ, and 0 otherwise, and $\#A$ denotes number of patterns in A.

LA is a representative of local confidence measures which compute some standard global measure of classification quality on neighborhood of the currently classified pattern $\vec{x}$. Of course, any other global measure of confidence could be used.

4.2 Local Match (LM)

Delany et al., 2005, describe several methods for determining local classification confidence in their spam filtering application. Most of the methods are based on similarity of the currently classified pattern $\vec{x}$ to neighboring training patterns. The main idea is that if $\vec{x}$ is near the decision boundary, the prediction may not be accurate, i.e., the local classification confidence should be low.

Based on the ideas behind these methods, we propose a local classification confidence measure called *local match* (LM). Let $NN_k(\vec{x})$ denote a set of k training patterns nearest to $\vec{x}$ (again, we used Euclidean metric), and let $NLN_k(\vec{x})$ denote a set of patterns from $NN_k(\vec{x})$ which belong to the same class as predicted by ϕ for pattern $\vec{x}$. LM of ϕ on $\vec{x}$ is computed as

$$\kappa_\phi(\vec{x}) = \frac{\#NLN_k(\vec{x})}{\#NN_k(\vec{x})} = \frac{\#NLN_k(\vec{x})}{k}. \quad (3)$$

4.3 Confidence Measures based on Degrees of Classification

It is also possible to use directly the output of the classifier, i.e., the degrees of classification, to compute the classification confidence. Let $\phi(\vec{x}) = (\mu_1, \ldots, \mu_N)$, and $\mu_{(1)}$, $\mu_{(2)}$ be the highest and the second highest degrees of classification. Wilson and Martinez, 1999, define local classification confidence as

$$\kappa_\phi(\vec{x}) = \frac{\mu_{(1)}}{\sum_{i=1}^{N} \mu_i}. \quad (4)$$

We will call this measure *degree of classification ratio* (DCR). Avnimelech and Intrator, 1999, define local classification confidence as

$$\kappa_\phi(\vec{x}) = (\mu_{(1)} - \mu_{(2)})^s, \quad (5)$$

where $s \geq 1$. We will call this measure *two-best margin* (TBM). These measures do not need to compute neighboring patterns of $\vec{x}$, and therefore they are very fast compared to LA and LM.

4.4 Experiment 1 - Performance of the Proposed Measures

To get a general idea to which extent the proposed local classification confidence measures really express the probability that the classification of the currently classified pattern is correct, we examined the histograms of the local classification confidence values for correctly classified (OK) and for misclassified (NOK) patterns.

We tested the measures with a quadratic discriminant classifier (Duda et al., 2000) implemented in Java programming language. 10-fold crossvalidation was performed to obtain the results on four artificial (Clouds, Concentric, Gauss_3D, Waveform) and five real-world (Balance, Breast, Phoneme, Pima, Satimage) datasets from the Elena database (UCL MLG, 1995) and from the UCI repository (Newman et al., 1998). The parameters of the measures were set to $k = 20$ for LA and LM, and $s = 1$ for TBM (based on preliminary testing; no fine-tuning or optimization was done).

Ideally, the OK distribution should be concentrated near one, while the NOK distribution should be concentrated near zero, and the distributions should be clearly separated. If the distributions overlap, or if the NOK distribution has high values near one, it means that the measure does not really express the probability that the classification of the currently classified pattern is right.

Unfortunately, due to space constraints, we can not show here the results for all the datasets. Generally speaking, for most of the datasets, the OK and NOK distributions for LA and LM are quite separated, but for DCR and TBM, the OK and NOK distributions are similar and overlapping, and the NOK distributions have high values near one. This suggests that the DCR and TBM measures do not really express the probability of correct classification. Additional preliminary experiments showed that DCR and TBM give very poor results in classifier combining. Therefore we did not further study DCR and TBM.

As for LA and LM, for most datasets, the OK and NOK distributions are quite separated, which suggests good predictive power, cf. Fig. 1(a). For some datasets (Gauss_3D, Breast, Pima), the distributions for LA and LM are overlapping, which suggests bad predictive power, cf. Fig. 1(b).

5 CLASSIFIER AGGREGATION WITH CLASSIFICATION CONFIDENCE

In this section, we describe two commonly used algorithms for classifier aggregation, and two modifications of these algorithms which utilize the concept of local classification confidence. Recall that $(\phi_1, \ldots, \phi_r)$ is a team of classifiers, and (1) is the output of the team for a pattern $\vec{x}$. Let μ_j denote the aggregated degree of classification to class C_j. Then we define the following aggregation algorithms:

Mean Value Aggregation (MV) computes the final aggregated degree of classification to class C_j as the average of all degrees of classification to class C_j by all the classifiers $\phi_1, \ldots, \phi_r$ in the team:

$$\mu_j = \frac{1}{r}\sum_{i=1}^{r} \mu_{i,j}. \qquad (6)$$

Weighted Mean Aggregation (WM) uses weighted mean to compute the final prediction:

$$\mu_j = \frac{\sum_{i=1}^{r} \omega_i \mu_{i,j}}{\sum_{i=1}^{r} \omega_i}. \qquad (7)$$

The weights $\omega_1, \ldots, \omega_r$ are defined as global confidences (e.g., validation accuracies) of the classifiers in the team.

Local Weighted Mean Aggregation (LWM) replaces the weights in the weighted mean by local classification confidences of the classifiers in the team on the currently classified pattern $\vec{x}$:

$$\mu_j = \frac{\sum_{i=1}^{r} \kappa_{\phi_i}(\vec{x})\mu_{i,j}}{\sum_{i=1}^{r} \kappa_{\phi_i}(\vec{x})}. \qquad (8)$$

Filtered Mean Aggregation (FM) is a modification of MV, the difference being that prior to computing the mean value, classifiers with local classification confidence on the current pattern lower than some threshold T are discarded. If $T = 0$, FM coincides with MV. If there are no classifiers with local classification confidence higher than T, then T is lowered to the value of the maximal local classification confidence in the team.

5.1 Experiment 2 - Performance of the Proposed Aggregation Algorithms

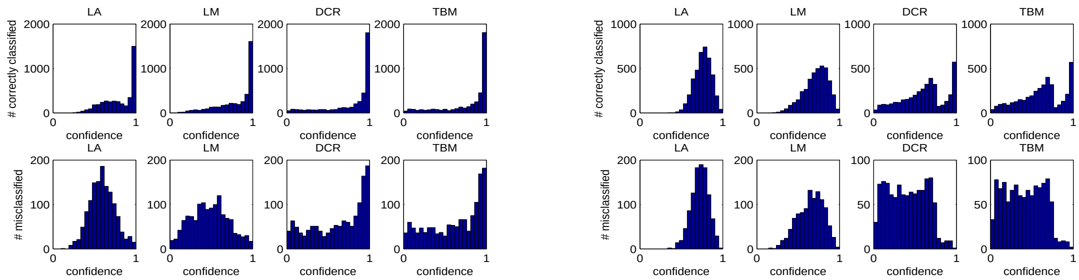
In the second experiment, we tested the performance of the classifier aggregation algorithms described in Section 5, in order to determine possible benefits of incorporating local classification confidence to classifier aggregation methods.

We designed an ensemble $(\phi_1, \ldots, \phi_r)$ of quadratic discriminant classifiers (Duda et al., 2000), and we aggregated the ensemble using the methods described in this section (MV, WM, which do not use local classification confidence, and LWM and FM using LA and LM local classification confidence measures). We also compared the algorithms' performance with the so-called *non-combined classifier* (NC), i.e., a common quadratic discriminant classifier (the NC classifier represents an approach which we had to use if we could use only one classifier). The 7 individual methods will be denoted NC, MV, WM, LWM-LA, FM-LA, LWM-LM, FM-LM. The algorithms' performance was tested on the same datasets as in Exp. 1.

The ensemble was created either by the bagging algorithm (Breiman, 1996), which creates classifiers trained on random samples drawn from the original training set with replacement, or by the multiple feature subset method (Bay, 1999), which creates classifiers using different combinations of features, depending on which method was more suitable for the particular dataset.

All the methods were implemented in Java programming language, and 10-fold crossvalidation was performed to obtain the results. The same parameter values as in Exp. 1 were used, and we set $T = 0.8$ or $T = 0.9$, depending on the particular dataset (based on some preliminary testing; no fine-tuning or optimization was done).

The results of the testing are shown in Table 1. Mean error rate and standard deviation of the error



(a) Phoneme – LA, LM good separation; DCR, TBM bad separation

(b) Gauss_3D – all measures bad separation

Figure 1: Histograms of local classification confidence values (LA - Local Accuracy, LM - Local Match, DCR - Degree of Classification Ratio, TBM - Two Best Margin) for correctly classified and misclassified patterns.

Table 1: Comparison of the classifier aggregation methods – non-combined classifier (NC), mean value (MV), weighted mean (WM), local weighted mean (LWM) using two confidence measures (LA, LM), and filtered mean (FM) using two confidence measures (LA, LM). Mean error rate (in %) ± standard deviation of error rate from 10-fold crossvalidation was calculated. The (B/M) after dataset name means whether the ensemble was created by Bagging or Multiple feature subset algorithm.

Dataset	NC	MV	WM	LWM-LA	FM-LA	LWM-LM	FM-LM
Clouds (M)	24.9 ± 1.4	24.9 ± 1.8	24.7 ± 1.8	23.4 ± 2.2	**22.2 ± 1.7** *†	23.1 ± 1.9	**21.9 ± 1.9** *†
Concentric (B)	3.5 ± 1.1	3.6 ± 1.1	3.4 ± 1.5	3.6 ± 1.4	**1.7 ± 0.7** *†	2.8 ± 1.4	**1.6 ± 0.7** *†
Gauss_3D (B)	21.4 ± 1.8	21.3 ± 2.2	21.4 ± 1.5	21.5 ± 2.7	21.7 ± 1.7	21.4 ± 1.6	21.6 ± 1.9
Waveform (B)	14.8 ± 1.4	14.8 ± 1.4	15.1 ± 2.0	15.0 ± 1.6	14.7 ± 0.7	14.5 ± 1.5	14.2 ± 1.9
Balance (M)	8.3 ± 3.6	11.0 ± 4.7	15.5 ± 4.2	9.0 ± 1.9	9.5 ± 3.9	8.3 ± 3.5	9.5 ± 2.5
Breast (M)	4.7 ± 3.0	4.7 ± 2.7	3.5 ± 2.6	2.9 ± 1.0	2.9 ± 1.5	3.1 ± 2.5	3.1 ± 2.6
Phoneme (M)	24.5 ± 2.0	23.7 ± 0.9	23.8 ± 2.7	**21.4 ± 2.0** *	**16.8 ± 1.9** *†	**21.0 ± 1.0** *†	**16.2 ± 1.7** *†
Pima (M)	27.0 ± 3.0	25.5 ± 6.8	26.1 ± 5.7	24.5 ± 5.0	24.2 ± 3.6	23.3 ± 4.2	25.0 ± 4.6
Satimage (B)	15.5 ± 1.6	15.5 ± 1.0	15.6 ± 1.7	15.5 ± 1.1	15.4 ± 1.0	15.1 ± 1.7	14.4 ± 1.5

*Significant improvement to NC
†Significant improvement to MV

rate of the aggregated classifiers from 10-fold crossvalidation was measured. We also measured statistical significance of the results – results which are significantly better than NC classifier or MV aggregator are marked by * or † and are displayed in boldface. The significance was measured at 5% level by the analysis of variance using Tukey-Kramer method (by the 'multcomp' function from the Matlab statistics toolbox).

The results show that for most datasets, the four aggregation methods which use local classification confidence (LWM-LA, FM-LA, LWM-LM, FM-LM) outperform the two aggregation methods which do not use local classification confidence (MV, WM). For three datasets, these results were statistically significant. FM usually gives better results than LWM, and if we compare the two confidence measures, we can say that LM gives usually slightly better results than LA. Generally speaking, the FM-LM was the most successfull algorithm in this experiment.

To summarize the results from both Exp. 1 and Exp. 2, we can say that by incorporating local classification confidence measures into classifier aggregation algorithms, significant improvement in classification quality can be obtained. However, the measures of local classification confidence sometimes do not express the probability that the classification of the currently classified pattern is right (see Fig. 1(b)), and therefore they do not improve classifier aggregation.

The experimental results from this paper are relevant to quadratic discriminant classifiers only, because for any other classifier types (k-NN, SVM, decision trees, etc.), the measures could give quite different results. However, it should be noted that in our not yet published experiments with Random Forests, we obtained similar results.

6 SUMMARY & FUTURE WORK

In this paper, we studied the concept of local classification confidence and we introduced four measures of local classification confidence. We compared the distribution of the values of the measures for correctly classified and misclassified patterns for quadratic discriminant classifier. This experiment showed that the DCR and TBM measures are not suitable for using in aggregation of ensembles of quadratic discriminant classifiers.

We showed a possible way how local classification confidence can be used in classifier aggregation. The performance of these methods was compared to commonly used methods for classifier aggregation of an ensemble of quadratic discriminant classifiers on four artificial and five real-world benchmark datasets. The results show that incorporating local classifier confidence into classifier aggregation can bring significant improvements in the classification quality.

In our future work, we plan to study local classification confidence measures for other classifiers than quadratic discriminant classifier, mainly decision trees and support vector machines, and to incorporate local classification confidence into more sophisticated classifier aggregation methods.

ACKNOWLEDGEMENTS

The research presented in this paper was partially supported by the Program "Information Society" under project 1ET100300517 (D. Štefka) and by the Institutional Research Plan AV0Z10300504 (M. Holeňa).

REFERENCES

Aksela, M. (2003). Comparison of classifier selection methods for improving committee performance. In *Multiple Classifier Systems*, pages 84–93.

Avnimelech, R. and Intrator, N. (1999). Boosted mixture of experts: An ensemble learning scheme. *Neural Computation*, 11(2):483–497.

Bay, S. D. (1999). Nearest neighbor classification from multiple feature subsets. *Intelligent Data Analysis*, 3(3):191–209.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, 24(2):123–140.

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

Delany, S. J., Cunningham, P., Doyle, D., and Zamolotskikh, A. (2005). Generating estimates of classification confidence for a case-based spam filter. In Muñoz-Avila, H. and Ricci, F., editors, *Case-Based Reasoning, Research and Development, 6th International Conference, on Case-Based Reasoning, ICCBR 2005, Chicago, USA, Proceedings*, volume 3620 of *LNCS*, pages 177–190. Springer.

Duda, R. O., Hart, P. E., and Stork, D. G. (2000). *Pattern Classification (2nd Edition)*. Wiley-Interscience.

Freund, Y. and Schapire, R. E. (1996). Experiments with a new boosting algorithm. In *International Conference on Machine Learning*, pages 148–156.

Hand, D. J. (1997). *Construction and Assessment of Classification Rules*. Wiley.

Kittler, J., Hatef, M., Duin, R. P. W., and Matas, J. (1998). On combining classifiers. *IEEE Trans. Pattern Anal. Mach. Intell.*, 20(3):226–239.

Kuncheva, L. I. (2004). *Combining Pattern Classifiers: Methods and Algorithms*. Wiley-Interscience.

Kuncheva, L. I., Bezdek, J. C., and Duin, R. P. W. (2001). Decision templates for multiple classifier fusion: an experimental comparison. *Pattern Recognition*, 34(2):299–314.

Newman, D., Hettich, S., Blake, C., and Merz, C. (1998). UCI repository of machine learning databases. http://www.ics.uci.edu/∼mlearn/MLRepository.html.

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

Tsymbal, A., Pechenizkiy, M., and Cunningham, P. (2006). Dynamic integration with random forests. In Frnkranz, J., Scheffer, T., and Spiliopoulou, M., editors, *ECML*, volume 4212 of *Lecture Notes in Computer Science*, pages 801–808. Springer.

UCL MLG (1995). Elena database. http://www.dice.ucl.ac.be/mlg/?page=Elena.

Wilson, D. R. and Martinez, T. R. (1999). Combining cross-validation and confidence to measure fitness. In *Proceedings of the International Joint Conference on Neural Networks (IJCNN'99), paper 163*.

Woods, K., W. Philip Kegelmeyer, J., and Bowyer, K. (1997). Combination of multiple classifiers using local accuracy estimates. *IEEE Trans. Pattern Anal. Mach. Intell.*, 19(4):405–410.

Zhu, X., Wu, X., and Yang, Y. (2004). 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. IEEE Computer Society.