BUILDING AN ENSEMBLE CLASSIFIER USING ENSEMBLE MARGIN. APPLICATION TO IMAGE CLASSIFICATION

Li Guo[1,2], *Samia Boukir*[1]

[1]Bordeaux INP, G&E, EA 4592, F-33600 Pessac, France.
[2]Atos Worldline, ZIA rue de la pointe, 59113 Seclin, France.
E-mail: li.guo@worldline.com , samia.boukir@ipb.fr

ABSTRACT

Bagging is a simple and powerful ensemble method which relies on bootstrap sampling over training data to produce diversity. Indeed, ensembles generalise better when their members form a diverse and accurate set. In this paper, the margin theory is exploited to select training instances for bagging. The selection of training data is performed using a new iterative guided bagging algorithm exploiting low margin instances. This method has been successfully applied to image data. Results show that low margin instances have a major influence on forming an appropriate training set to build reliable ensemble classifiers, leading to a significant increase in both overall and per-class accuracies.

***Index Terms*—** Bagging, data selection, ensemble learning, margin, multiple classifier.

1. INTRODUCTION

Ensemble learning is a powerful learning paradigm, which builds a classification model by integrating multiple diversified component learners [1, 2]. The success of ensemble methods arises mainly from the fact that they improve the overall predictive performance. Typically, ensemble methods consist of two phases : the *production* of multiple classifiers and their *combination*. An ensemble can be composed of either *homogeneous* or *heterogeneous* classifiers. Homogeneous classifiers, which are the most widely used, derive from different executions of the same learning algorithm. Such classification models can be produced for example through the manipulation of the training instances or the input attributes. Popular methods for producing homogeneous models are *bagging* [3] and *boosting* [4]. A common method for combining an ensemble of classifiers is *majority voting* where the class with most votes is the one proposed by the ensemble.

The classification margin is considered as a measure of confidence of classification. Significant work has been published about bounding and reducing the generalization error based on the classification margin [5, 6, 7]. This work exploits this concept to design better ensemble classifiers that can efficiently learn from appropriate training sets, made up of the most informative instances, to successfully classify image data. We propose a new iterative guided bagging algorithm based on an unsupervised ensemble margin. This method emphasizes the role of lower margin samples in the learning process at the expense of highest margin samples, the latter having the least influence on ensemble classification performance. While in previous work [8, 9] we select lowest margin training instances in a single step data selection for SVM (Support Vector Machine), in this work we remove iteratively highest margin instances in a repetitive data selection for bagging.

2. ENSEMBLE LEARNING AND MARGIN THEORY

Margins, which were originally applied to explain the success of boosting [7], play a crucial role in modern machine learning research. The fact that it is the margin of a classification rather than the raw training error that matters has become a key tool in recent years when dealing with classifiers [10]. Indeed, the main issue in classification is to get the boundary between classes right [11]. The margin concept can be applied to both the theoretical analysis and design of algorithms [12, 13].

The ensemble margin [7] is a fundamental concept in ensemble learning. Several studies have shown that the generalisation performance of an ensemble classifier is related to the distribution of its margins on the training data [6, 7]. The margin can provide extra information for improving classification accuracy but how to use it is still not yet fully explored.

The decision by an ensemble for each instance is made by voting. This vote is the hypothesis of the ensemble decision function. The ensemble margin can be calculated as a difference between the votes, therefore, it is a distance between the hypothesises. The most popular ensemble margin is given by the difference between the fraction of classifiers voting correctly and incorrectly [7, 14].

We use the ensemble margin we previously defined in [8, 15], which combines the first and second most voted class labels under the model. This margin can be computed by equation 1, where c_1 is the most voted class for sample x and v_{c_1} the number of related votes, c_2 is the second most popular class and v_{c_2} the number of corresponding votes. This margin's range is from 0 to +1. It is an alternative to the classical definition of the margin [7] with an appealing property : *it does not require the true class label of an instance*, i.e. it is unsupervised. Thus, it is potentially more robust to noise as it is not affected by errors occurring on the class label itself. This margin has been used in previous work for data selection [8, 9], mislabeled data removal and correction [16, 17], data imbalance evaluation [18], and ensemble selection [15] where base classifiers are ordered according to a ranking measure on which lowest margins have the most influence and highest margins have the least influence.

$$\begin{aligned} margin(x) &= \frac{v_{c_1} - v_{c_2}}{\sum_{c=1}^{L}(v_c)} \\ &= \frac{\max_{c=1,\ldots,L}(v_c) - \max_{c=1,\ldots,L \cap c \neq c_1}(v_c)}{T} \end{aligned} \quad (1)$$

where T represents the number of base classifiers in the ensemble.

Correctly classified training instances with high margin values (i.e. close to 1), represent instances located away from class decision boundaries and can contain a high degree of redundant information in a classification problem. Conversely, training instances with low margin values (close to 0) are often located near class decision boundaries and are more informative in a classification task.

To demonstrate the relationship between the margin value of an instance and the characteristics of this instance, we used bagging to create an ensemble, involving Classification and Regression Trees (CART) [19] as base classifiers, to classify a synthetic data set *Sin*. Figure 1 clearly shows that the majority of the small margin instances of this data set are close to class boundaries while most of the high margin instances are away from class boundaries.

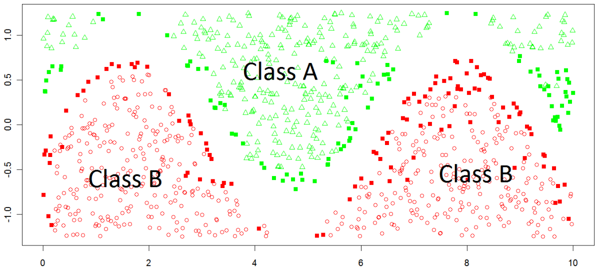

Fig. 1. Small margin instances (displayed in filled points) in data set *Sin*

In the following, while special emphasis is placed on bagging, the developed methods are also applicable to any bootstrap-aggregating ensemble learning approach like random forests [20].

3. ENSEMBLE MARGIN FOR TRAINING

In supervised learning, the training set is an essential element of the learning process. It is the single clue to finding the objective function. We focus here on forming an appropriate training set made of the most informative examples and cleaned out of any redundant data. We first construct an ensemble classifier at data level. A new bagging algorithm, relying on an unsupervised ensemble margin, is then applied to minimise data redundancy and improve classification accuracy.

3.1. Bagging

Bagging is one of the most successful ensemble methods. It relies on bootstrap sampling (with replacement) over training data to produce diversity [3]. Diversity is derived from the differences between the training sets of base classifiers [21]. The more differences in these training sets the more diversity can be achieved. However, bootstrapped training sets [22] become more and more similar as redundancy is increasing in training set. Redundancy not only slows down the training task but it can also significantly decrease diversity, thus degrading the performance of bagging, affecting the rarer and most difficult classes. In the training process, each example carries its own piece of contribution about the target. However, the contributions are different from each other. Obviously, the redundant instances' contribution is less significant than the contribution of other unmatched instances.

3.2. Margin-based bagging

Li et al. proposed to group the training instances into three categories: typical, critical and noisy [23]. Generally, especially in image classification, typical instances occur in larger numbers than critical instances such as class decision boundary instances. Consequently, typical instances are more likely to be similar. Thus, there is potentially more redundancy in these instances.

Our bagging algorithm is based on training instance ordering and relies on the unsupervised ensemble margin previously defined. This technique selects the most informative training instances based on their margins. The margin of each training instance is calculated using equation 1. Our method orders the training instances according to their margin values; the higher the margin of a training instance, the higher the probability this instance being typical and potentially redundant. It trains first the committee of bootstrapped base classifiers on the complete training set and then removes iteratively a fixed portion (training sample pruning parameter M) of the data points on which the committee most agrees according to

Algorithm 1 Margin-based bagging

```
Inputs:
    Whole training data set $S_0$ of size $N$.
    Ensemble creation method $E$.
    Base learning algorithm $B$.
    Validation set $V$.
    Amount $M$ of pruning at each iteration.
Initialise $S$ = $S_0$, $i = 0$.
repeat
    Create an ensemble classifier $EB_i$ with $S$.
    Compute the margin value of each training instance.
    Evaluate $EB_i$ on validation data set $V$ and obtain error rate $Er$ of $EB_i$.
    Remove $M$ first highest margin instances to compose a new training set $S_i$.
    Set training set $S$ to $S_i$, $i = i + 1$.
until Size of $S$ = 0.
Outputs:
    Ensemble $EB_b$ with lowest error rate on $V$.
    Training set $S_b$ of $EB_b$.
```

the unsupervised ensemble margin previously defined. This amounts to removing typical (highest margin) instances during the learning process. The training of the bagging ensemble takes place again with the resulting reduced training set composed of the lowest margin instances. New margin values are calculated for each remaining training instance. This iterative guided procedure is repeated until reaching a maximum training accuracy providing an optimal ensemble designed with a reduced and more informative training set.

Our method is illustrated by algorithm 1.

3.3. Experimental results

In this work, bagging was used to create an ensemble using pruned Classification and Regression Tree (CART) [19] as base classifier. Statistical analysis is based on R-project [24], a software environment for statistical computing and graphics. Bagging ensembles were constructed with 100 trees, a typical moderate size ensemble [25]. The size M of the subset of instances to remove at each pruning step was set to 5% of the whole training set. We ran experiments on 10 representative data sets from the UCI repository [26]. The top five data sets are image data and the bottom five, including *Waveform*, a popular synthetic data, stem from various classification problems. Each data set has been divided into two parts: training set and test set, as shown on Table 1. We used the training set as validation set for each experiment. All the reported results are mean values of a 10-time calculation.

Data set	Training	Test	Attributes	Classes
Letter	10000	10000	16	26
Optdigits	2810	2810	64	10
Pendigit	5496	5496	16	10
Segment	1155	1155	19	7
Vehicle	423	423	18	4
Connect-4	33779	33778	42	3
Krvskp	1598	1598	36	2
Pima	384	384	8	2
TicTac	479	479	9	2
Waveform	2500	2500	21	3

Table 1. Data sets.

3.3.1. *Overall classification performance*

Fig. 2 shows the classification accuracy curve as function of the size of the training sub-set selected by our method on $Pendigit$'s test set. The classification accuracy curve increases monotonically as the percentage of instances of higher margin removed from the whole training set increases until reaching a peak (best accuracy) and then decreases monotonically. Less than 20% of the whole training set was used but the accuracy increased by about 3.5%.

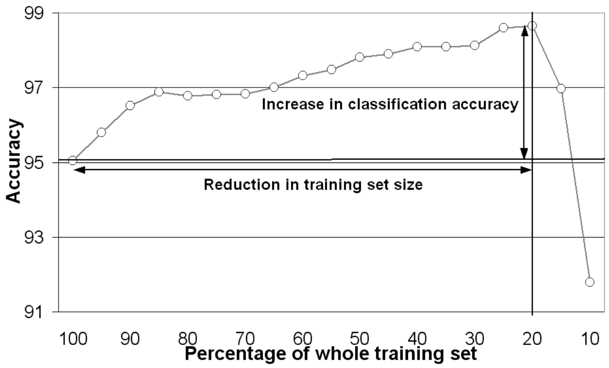

Fig. 2. Classification accuracy as function of training set size for data set $Pendigit$ (on test set).

Table 2 presents the average and standard deviation of the classification accuracy obtained on test set by the chosen ensemble that led to maximum classification accuracy on training set, as well as the pruned training set size (percentage of lowest margin instances of the whole training set) of this optimal ensemble. This table shows that our bagging method outperforms classic bagging, which relies on the whole training set, in terms of classification accuracy (with an improvement of up to 3.7%). The training data sizes were significantly reduced (size reduction rate greater that 50% for half the data sets), especially in case of data set *Pendigit* for which the training set was reduced to about **20%** of the full training set. Hence, the **80%** highest margin samples that have

Data set	Acc. W (%)	Acc. R (%)	Size R (%)
Letter	91.01 ± 0.16	93.95 ± 0.52	47.0 ± 4.21
Optdigits	93.27 ± 0.15	96.93 ± 0.22	31.0 ± 5.16
Pendigit	95.05 ± 0.16	98.62 ± 0.16	20.5 ± 2.83
Segment	96.88 ± 0.34	97.40 ± 0.23	26.0 ± 3.94
Vehicle	70.87 ± 1.10	74.56 ± 0.59	65.0 ± 4.71
Connect-4	78.32 ± 0.12	81.00 ± 0.45	46.5 ± 2.41
Krvskp	97.81 ± 0.22	98.68 ± 0.40	66.5 ± 5.29
Pima	76.04 ± 0.77	75.65 ± 0.73	57.0 ± 3.49
TicTac	97.18 ± 0.55	98.35 ± 0.31	73.0 ± 6.74
Waveform	83.64 ± 0.37	84.88 ± 0.36	59.0 ± 2.10

Table 2. Accuracy on test sets by the ensembles with whole training set (W), reduced training set (R) and associated reduced size.

Data set	Accuracy W (%)	Accuracy R (%)
Letter	82.78 ± 1.09	88.28 ± 1.68
Optdigits	88.23 ± 1.08	93.79 ± 0.22
Pendigit	86.65 ± 0.92	96.35 ± 0.49
Segment	92.04 ± 1.26	91.32 ± 1.45
Vehicle	46.30 ± 3.23	46.80 ± 2.25
Connect-4	4.20 ± 0.32	16.52 ± 1.29
Krvskp	96.81 ± 0.35	98.02 ± 0.65
Pima	53.77 ± 1.48	54.44 ± 2.47
TicTac	94.96 ± 1.48	96.62 ± 1.08
Waveform	78.29 ± 0.81	79.48 ± 0.84

Table 3. Minimum classification accuracy per class by the ensembles with whole training set (W) and reduced training set (R), on test sets.

been discarded from the training set are redundant. These results show that classification data, particularly image data, are highly redundant and that only a fraction of them are useful for training.

3.3.2. Classification performance per class

Table 3 presents the minimum accuracy per class obtained on test set by the chosen ensemble with best overall accuracy on training set. This table shows that our method largely outperforms classic bagging in terms of handling complex and rare classes. Moreover, the ensemble evaluation function which has been used is overall accuracy on training set; if we use the accuracy of the hardest class as evaluation function instead, the accuracy of the most difficult class on test set would further increase. Thus, our method not only improves the overall classification accuracy compared to classic bagging (see table 2), but also significantly increases the accuracy of the hardest class (up to **12%**). The effectiveness of our method in handling difficult classes encourages its application to cost-sensitive learning and imbalanced data.

3.3.3. Unsupervised margin versus supervised margin

We assess here the performance of our bagging method when applying different margin definitions. Schapire's margin [7] is one of the most popular margin formulations. It is calculated as the difference between the votes to the correct class label and the maximal votes to any single incorrect class label. In figure 3, axis x represents the overall accuracy achieved by our method applying our margin definition [8, 15] (equation 1), and axis y represents the overall accuracy obtained using Schapire's margin. We can notice that, except for one dataset, our margin-based bagging method, which relies on an unsupervised ensemble margin, outperforms Schapire's margin-based bagging, which relies on a supervised ensemble margin. The involvement of the unsupervised margin in our bagging algorithm significantly improves the ensemble performance for about half the data sets compared to the use of the classic supervised margin (the performances are similar for the other half data sets).

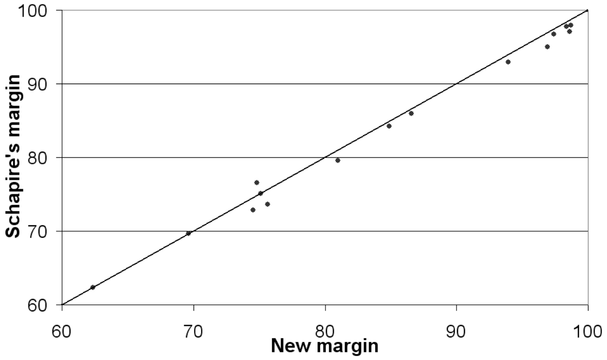

Fig. 3. Unsupervised margin versus supervised margin performances in margin-based bagging

4. CONCLUSION

In this paper, we have presented an original contribution for efficient ensemble design based on an unsupervised version of ensemble margin. A simple bagging method, which relies on critical instances (class decision boundaries or difficult or rare classes), which have low margins, to improve the learning process, has been introduced. This method increases the ensemble performance, particularly in case of difficult or rare classes. Hence, targeting lower margin instances (which represent samples closer to class decision boundaries and/or more difficult than higher margin samples) demonstrates improved ensemble performance. This strategy reduces data redundancy and increases information significance (e.g. class decision boundary instances are more informative). Therefore, it designs stronger ensemble classifiers with an increased capability for handling hard or rare classes. The automatic tuning of pruning parameter M will be investigated in future work.

5. REFERENCES

[1] T.G. Dietterich, "Ensemble methods in machine learning," *Proceedings of the First International Workshop on Multiple Classifier Systems*, pp. 1–15, 2000.

[2] L. Rokach, "Ensemble-based classifiers," *Artificial Intelligence Review*, vol. 33, no. 1, pp. 1–39, 2010.

[3] L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, no. 2, pp. 123–140, 1996.

[4] Y. Freund and R.E. Schapire, "Experiments with a new boosting algorithm," in *The 13th International Conference on Machine Learning, ICML'96*, 1996, pp. 148–156.

[5] V. Koltchinskii and D. Panchenko, "Complexities of convex combinations and bounding the generalization error in classification," *The Annals of Statistics*, vol. 33, no. 4, pp. 1455–1496, 2005.

[6] L. Mason, P.L. Bartlett, and J. Baxter, "Improved generalization through explicit optimization of margins," *Machine Learning*, vol. 38, no. 3, pp. 243–255, 2000.

[7] R.E. Schapire, Y. Freund, P. Bartlett, and W.S. Lee, "Boosting the margin: a new explanation for the effectiveness of voting methods," *The Annals of Statistics*, vol. 26, no. 5, pp. 1651–1686, 1998.

[8] L. Guo, S. Boukir, and N. Chehata, "Support vectors selection for supervised learning using an ensemble approach," in *ICPR'2010, 20th IAPR International Conference on Pattern Recognition*, 2010, pp. 37–40.

[9] L. Guo and S. Boukir, "Fast data selection for svm training using ensemble margin," *Pattern Recognition Letters*, , no. 51, pp. 112–119, 2015.

[10] P.J. Bartlett, B. Schölkopff, D. Schuurmans, and A.J Smola, Eds., *Advances in Large Margin Classifiers*, Neural Information Processing. The MIT Press, 1 edition, 2000.

[11] L. Breiman, "Half & half bagging and hard boundary points," Tech. Rep. 534, Statistics department, University of California, 1998.

[12] R. Gilad-Bachrach, A. Navot, and N. Tishby, "Margin based feature selection - theory and algorithms," in *The twenty-first international conference on Machine learning, ICML'04*, 2004, pp. 43–50.

[13] Q. Hu, L. Li, X. Wu, G. Schaefer, and D. Yu, "Exploiting diversity for optimizing margin distribution in ensemble learning," *Knowledge-Based Systems*, vol. 67, no. 0, pp. 90–104, 2014.

[14] M. N. Kapp, R. Sabourin, and P. Maupin, "An empirical study on diversity measures and margin theory for ensembles of classifiers," in *The10th International Conference on Information Fusion, Fusion'07*, 2007, 1-8.

[15] L. Guo and S. Boukir, "Margin-based ordered aggregation for ensemble pruning," *Pattern Recognition Letters*, , no. 6, pp. 603–609, 2013.

[16] L. Guo and S. Boukir, "Ensemble margin framework for image classification," in *ICIP'2014, 21st IEEE International Conference on Image Processing*, 2014, pp. 4231–4235.

[17] W. Feng and S. Boukir, "Class noise removal and correction for image classification using ensemble margin," in *ICIP'2015, 22nd IEEE International Conference on Image Processing*, 2015, pp. 4698–4702.

[18] A. Mellor, S. Boukir, A. Haywood, and S. Jones, "Using ensemble margin to explore issues of training data imbalance and mislabeling on large area land cover classification," in *ICIP'2014, 21st IEEE International Conference on Image Processing*, 2014, pp. 26–29.

[19] L. Breiman, J.H. Friedman, R.A. Olshen, and C.J. Stone, *Classification and Regression Trees*, Publisher: Wadsworth, 1984.

[20] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, Oct. 2001.

[21] L.I. Kuncheva and C.J. Whitaker, "Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy," *Machine Learning*, vol. 51, pp. 181–207, 2003.

[22] B. Efron and R. Tibshirani, *An Introduction to the Bootstrap*, Chapman & Hall/CRC, 1994.

[23] L. Li, A. Pratap, H. Lin, and Y. S. Abu-mostafa, "Improving generalization by data categorization," in *Knowledge Discovery in Databases, PKDD'05*, 2005, pp. 157–168.

[24] R Development Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2013, ISBN 3-900051-07-0.

[25] G. Tsoumakas, I. Partalas, and I. Vlahavas, *Applications of Supervised and Unsupervised Ensemble Methods*, vol. 245/2009, chapter An Ensemble Pruning Primer, pp. 1–13, Springer, 2009.

[26] A. Asuncion and D.J. Newman, "UCI machine learning repository," 2007.