ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

Improving the Robustness of Bagging with Reduced Sampling Size

Maryam Sabzevari, Gonzalo Martínez-Muñoz and Alberto Suárez *

C/Francisco Tomás y Valiente, 11, Escuela Politécnica Superior
Universidad Autónoma de Madrid, Madrid (28049), Spain

Abstract. Bagging is a simple and robust classification algorithm in the presence of class label noise. This algorithm builds an ensemble of classifiers by bootstrapping samples with replacement of size equal to the original training set. However, several studies have shown that this choice of sampling size is arbitrary in terms of generalization performance of the ensemble. In this study we discuss how small sampling ratios can contribute to the robustness of bagging in the presence of class label noise. An empirical analysis on two datasets is carried out using different noise rates and bootstrap sampling sizes. The results show that, for the studied datasets, sampling rates of 20% clearly improve the performance of the bagging ensembles in the presence of class label noise.

1 Introduction

Ensembles generally improve the performance of a single classifier by combining many randomized versions of the individual classifier. In bagging [1], the predictors are trained using different bootstrap samplings from the training data. Each bootstrap sample in bagging is extracted from the original training set with replacement. The standard procedure is to extract a number of samples equal to the size of the training set, that is a sampling proportion or ratio of 100%. This prescription produces samples containing on average 63.2% of different instances and the rest are repeated examples. The final combination of classifiers is given by majority voting. The combination of classifiers in bagging works by removing uncorrelated errors of individual predictors, decreasing their variance and, in consequence, the ensemble prediction error. Many studies have shown that bagging is a robust classification algorithm under different noise conditions [2, 3, 4, 5]. Noisy data affects the final error because of the increment of the variance of the individual classifiers in the ensemble [3]. In this context, variance reduction methods such as bagging is an option to consider in noisy datasets.

The prescription used in bagging for generating the bootstrap samples does not need to be optimal in terms of generalization error. In fact, when using sampling with equal size of the original training set, the performance of nearest neighbours is equal to that of bagged nearest neighbours [1]. However, if each bootstrap sample contains on average less than 50% unique instances from the training set, then the accuracy of bagged nearest neighbours improves. In addition, if the sampling ratio tends to 0 as the training set size tends to $+\infty$, then

*The authors acknowledge financial support from the Spanish Dirección General de Investigación, project TIN2010-21575-C02-02

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

its performance tends asymptotically to that of the optimal Bayes classifier [6]. Another study [7] shows that in general subbagging with low subsampling rates produce better results than bagging when combining stable classifiers. In [8] is proposed an out-of-bag estimation of the optimal sampling ratios for bagging and subbaging. The article analyses bagging in 30 datasets using sampling ratios from 2% to 120%. The results show that the optimal sampling rate is problem dependant and that out-of-bag error can be used to estimate the optimal sampling ratio.

An important aspect of using small sampling ratios is how they affect isolated instances, where by isolated instances we refer to those instances surrounded by instances of another class. When bootstrap samples contain less than 50% of the instances of the training set, then the classification given by the ensemble for those instances will tend to be dominated by the surrounding instances [6, 8]. Interestingly, incorrectly labelled instances can in general be considered as isolated instances. Hence, in these cases, using small sampling rates in bagging can help to reduce its influence in the training phase and, as a consequence, to build more robust ensembles.

In this paper, we will analyse the sensitivity of bagging ensembles in the presence of different class label noise and we will evaluate the effectiveness of different sampling ratios in the noisy datasets. The experimental study is performed on two datasets: a synthetic and real dataset.

2 Sampling ratio influence in bagging under class label noise

In Figure 1 it is shown the average percentage of unique instances contained on bootstrap samples with respect to the bootstrap sampling ratio. From this plot we can observe that, very large or very small sampling ratios are two non acceptable extremes. For very large sampling ratios, the number of unique instances in each sample tends to the number of instances of the original training set, and in consequence, all base learners would be rather similar –differences among classifiers would come from the effect of repeated examples on the base learners. On the hand, given a fixed number of instances in the training set, a small sampling ratio producing a single instance per sample would cause the base learners to output the class of that instance for the whole instance space. Hence, the final decision of the ensemble would be, for the whole instance space, equal to the class with the highest apriori probability in the training set. In general, the optimal generalization performance will be obtained at intermediate sampling ratios.

In addition, this figure shows that for the point where the number of unique samples is less than 50% of the total size, is for sampling ratios below $\approx 69.3\%$. For sampling ratios below this threshold, on average, every instance in the original training set is used for induction in less than half of the classifiers in the ensemble. This means that, the class assignment given by the ensemble to a given training example, takes into account not only its own labelling, but also the class

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

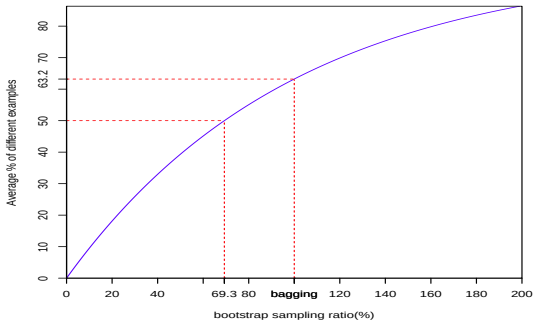

Fig. 1: Average percentage of the different selected examples with respect to the proportion of the bootstrap sample

labels of close by instances. In the case of mislabelled instances surrounded by instances of its *real* class, they would tend to be *correctly* classified (with respect to its true label) by the ensemble when sampling rates below $\approx 69.3\%$ are used. However, mislabeled instances close to the classification boundary would not be easily detected by using small sampling ratios.

3 Experimental Result

In this section we analyse the robustness of bagging ensembles in the presence of different rates of label noise with respect to the bootstrap sampling proportion. Two datasets were considered. One synthetic, *Threenorm* [9], and the real set, *Pima Indian Diabetes* [10].

For both datasets a similar experimental protocol was followed. The results reported are averages over one hundred partitions of the data into train and test sets. In *Pima* the partitions were obtained by 10×10-fold-cv. In *Threenorm* 100 random samplings from the true distribution were carried out, generating training sets of 300 instances and test sets of 2000 instances. For each data partition the following steps were carried out: (i) Noise was injected in the train set by changing the class to a given percentage of instances using the following values: 0% (no noise), 5%, 10% and 20%; (ii) For each noise level, six bagging ensembles composed of 500 CART unpruned trees [11] were built. The bootstrap sampling proportions used to train the bagging ensembles were: 20%, 40%, 60%, 80%, 100% (normal bagging) and 120%. CART trees were trained using its default parameters; (and iii) The generalization error of the ensembles was estimated in the test set. No noise was injected in the test sets, in order to make the performance of the different classifiers across different noise proportions comparable. Additionally, for the *Threenorm* dataset, the output of the ensemble was compared to the output of the optimal bayes classifier. This gives a total of $4 \times 6 \times 100 \times 500 = 1200000$ trees built for each dataset.

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

Table 1: Average error (in %) for various sampling sizes and label noise levels on *Threenorm*

	no noise		5% noise		10% noise		20% noise	
ratio	vs. Bay	Error	vs. Bay	error	vs. Bay	error	vs. Bay	error
20	**14.5**	**17.8**	**15.4**	18.7	**15.9**	**19.0**	**18.8**	**21.6**
40	14.6	18.0	**15.4**	**18.6**	16.1	19.3	19.2	22.0
60	14.8	18.2	15.7	18.8	16.8	19.8	19.9	22.5
80	15.3	18.6	16.0	19.3	17.2	20.1	19.9	22.3
100	15.7	19.1	16.9	20.1	17.6	20.5	20.6	23.0
120	16.1	19.3	17.2	20.1	18.1	21.0	21.6	23.9

Table 2: Average test error (in %) for various sampling sizes and label noise levels on *Pima*

ratios	no noise	5% noise	10% noise	20% noise
20	**23.8**	**23.8**	**24.8**	**25.5**
40	23.9	24.1	25.0	27.2
60	23.9	24.6	25.3	26.9
80	24.2	24.9	26.0	28.0
100	24.6	25.0	26.0	29.2
120	24.8	25.8	26.7	29.3

In Table 1, the average test error for bagging with different sampling ratios and noise levels, is shown for *Threenorm*. In addition, the table shows the mean difference of the ensembles' classification with respect to the optimal Bayes output (column "vs. Bayes"). Table 2 shows the same results for *Pima Indian Diabetes* except the comparison with the Bayes classifier since it is unknown.

In both datasets, the best results are obtained for small sampling ratios. Note that, the performance of bagging with a sampling ratio of 20% and with a noise injected of 10% in the train set, is equivalent to the performance of normal bagging trained in the noiseless data in both datasets. In addition, lower sampling ratios present lower differences with the oracle Bayes classifier (see Table 1). In *Pima Indian Diabetes*, bagging with sampling proportion of 20% presents a remarkably low deterioration of the classification error with respect to the level of noise injected. It goes from 23.8% error, when no noise is injected, to 25.5% when trained on the 20% label noise data (less than two percent points). By contrast, normal bagging presents a generalization deterioration of 5 percent points (from 24.6% with no noise to 29.2% with 20% noise).

Plots of Figures 2 show the performance of the ensembles with respect to the number of combined hypothesis for *Threenorm*. Similar plots are obtained for *Pima* dataset. What we observe in the figures is: (i) All ensembles over 69% sampling ratios (80%, standard bagging–100%– and 120%) independently of the noise injected, reach zero error in the training set. This indicates that the ensembles present some degree of over-fitting. Their performance in the test set seems to corroborate this observation. This is specially apparent when high noise levels are injected. (ii) The training error in ensembles with lower sampling ratios (20%, 40% and 60%), increases with the amount of injected noise. These

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

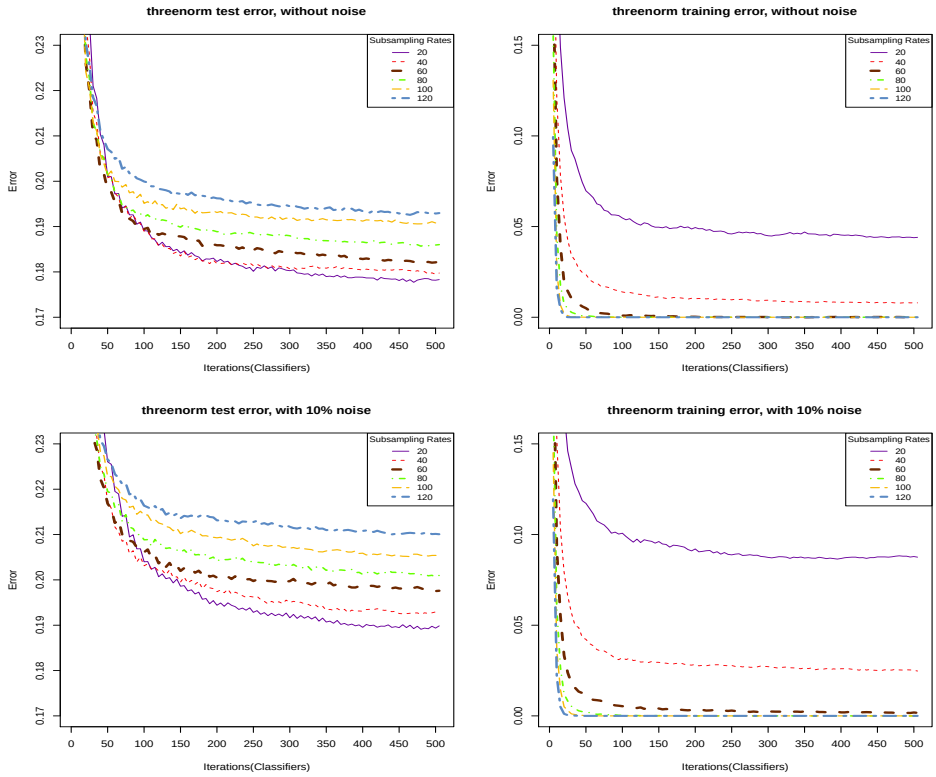

Fig. 2: Average test (left) and training (right) errors for different sampling ratios in presence of no-noise (top) and 10% (bottom) class label noise for *Threenorm* dataset.

seems to indicate the these ensembles present do not tend to over-fit to the injected noise as much as when 80%, 100% and 120% are used.

These observations can be explained by the fact that, for higher sampling ratios, each instance is used to train more than half of the base classifiers. And since we are using unpruned trees, all training instances will be correctly classified by the ensemble independently of the correctness of the class label. In these cases, regions in the attribute space around incorrectly labelled instances, will be miss classified. By contrast, when we use sampling ratios below 69%, the classification of each instance is influenced by the class of nearby instances. This is a positive property when a miss labelled instance is surrounded by correctly labelled examples. In this sense, the use of small sampling ratios in bagging can be seen as a regularization strategy.

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

4 Conclusion

In this article we have analysed the robustness of bagging for different sampling ratios under the presence of class label noise. The experiments carried out show that, in the two studied datasets, bagging ensembles of unpruned CART trees trained on bootstrap samples between 20% and 40% of the size of the original training set, are more robust than standard bagging. These are very promising results specially taking into account that standard bagging is considered as a robust classification algorithm. We have the intention to carry out a more exhaustive study on the effects of sampling size in the presence of noise using a larger and more diverse set of datasets and different base classifiers.

References

[1] Leo Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996.

[2] Thomas G. Dietterich. An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. *Machine Learning*, 40(2):139–157, 2000.

[3] Prem Melville, Nishit Shah, Lilyana Mihalkova, and Raymond J. Mooney. Experiments on ensembles with missing and noisy data. In *In: Proc. of the Workshop on MCS*, pages 293–302. Springer, 2004.

[4] Christian Pölitz and Ralf Schenkel. Robust Ranking Models Using Noisy Feedback. In *Workshop "Information Retrieval Over Query Sessions" (SIR 2012) at ECIR 2012*, pages 1 – 6, 2012.

[5] B. Frenay and M. Verleysen. Classification in the presence of label noise: a survey. *T. on IEEE Neural Networks and Learning Systems*, (99), 2013.

[6] Peter Hall and Richard J Samworth. Properties of bagged nearest neighbour classifiers. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(3):363–379, 2005.

[7] Faisal Zaman and Hideo Hirose. Effect of subsampling rate on subbagging and related ensembles of stable classifiers. In *Pattern Recognition and Machine Intelligence*, pages 44–49. Springer, 2009.

[8] Gonzalo Martínez-Muñoz and Alberto Suárez. Out-of-bag estimation of the optimal sample size in bagging. *Pattern Recognition*, 43(1):143–152, January 2010.

[9] Leo Breiman. Bias, Variance, and Arcing Classifiers. Technical Report 460, Department of Statistics, University of Berkeley, 1996.

[10] K. Bache and M. Lichman. UCI machine learning repository, 2013.

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