Neurocomputing 160 (2015) 18–33

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Small margin ensembles can be robust to class-label noise

Maryam Sabzevari, Gonzalo Martínez-Muñoz *, Alberto Suárez

Universidad Autónoma de Madrid, Escuela Politécnica Superior, Dpto. de Ingeniería Informática, C/Francisco Tomás y Valiente, 11, 28049 Madrid, Spain

ARTICLE INFO

Article history:
Received 15 March 2014
Received in revised form
5 December 2014
Accepted 15 December 2014
Available online 11 February 2015

Keywords:
Label noise
Bagging
Small margin classifiers
Bootstrap sampling

ABSTRACT

Subsampling is used to generate bagging ensembles that are accurate and robust to class-label noise. The effect of using smaller bootstrap samples to train the base learners is to make the ensemble more diverse. As a result, the classification margins tend to decrease. In spite of having small margins, these ensembles can be robust to class-label noise. The validity of these observations is illustrated in a wide range of synthetic and real-world classification tasks. In the problems investigated, subsampling significantly outperforms standard bagging for different amounts of class-label noise. By contrast, the effectiveness of subsampling in random forest is problem dependent. In these types of ensembles the best overall accuracy is obtained when the random trees are built on bootstrap samples of the same size as the original training data. Nevertheless, subsampling becomes more effective as the amount of class-label noise increases.

1. Introduction

The success of large margin classifiers [46,33,21,20] has prompted many researchers to posit that large margins are a key feature in explaining the effectiveness of these methods. In the context of ensembles, the margin is defined as the weighted sum of votes for the correct class minus the weighted sum of votes for the most voted class other than the correct one. The effectiveness of boosting has been ascribed to the fact that it produces large margins on the training data. The margins increase as the ensemble grows because of boosting's progressive focus on instances that are difficult to classify [43]. Nonetheless, several empirical studies put in doubt the general validity of this view [9,35]. Furthermore, efforts to directly optimize the margin (or the minimum margin) have met with mixed results [40,41]. In contrast to boosting, bagging [7], random forest [11] and class-switching [10,31] ensembles do not tend to increase the classification margins. In this paper we show that subsampling can be used to generate bagging ensembles that are robust to class-label noise in spite of having small margins. By contrast, the effectiveness of subsampling in random forest is strongly problem dependent. Nevertheless, for both types of ensembles, subsampling becomes more effective as the amount of class-label noise increases.

As discussed in [53,18], class-label noise is generally more harmful for classification accuracy than noise in the feature values. Therefore, it is important to design classifiers that are robust to errors in the class labels of the training instances. The deterioration in performance caused by this type of noise is mainly due to an increase of the variance of the classifiers [36,1,39]. Bagging is robust to class-label noise because it is a variance reduction technique. As a result of its adaptive nature, boosting reduces the classification bias as well as the variance [4,48]. However, the excessive emphasis on incorrectly labeled examples makes standard boosting algorithms ill-suited for handling this type of noise. Nonetheless, it is possible to design robust versions of boosting to address this shortcoming [40,20].

A bagging ensemble is a collection of classifiers whose predictions are combined by majority voting. Each of the classifiers in the ensemble is built on a different bootstrap sample from the original training data. In standard bagging, bootstrap samples of the same size of the original training set are used to build the individual classifiers. However, this prescription need not be optimal. Several empirical studies have shown that the generalization capacity of bagging can significantly improve when smaller bootstrap samples are used [24,52,32]. Subsampling generally makes bagging more robust to label noise [42]. The key to this improvement is how smaller sampling ratios affect isolated instances. By an isolated instance we mean one that is located in a region where the majority of neighboring instances belong to a different class. Assume a sampling ratio such that the bootstrap samples used to build the individual classifiers contain less than 50% of the original training instances. This means that each instance is present in less than half of the ensemble classifiers. Therefore, the decision on the label of a given instance is dominated by classifiers trained on bootstrap samples that do not contain that particular instance [24,32]. If the instance in question is an isolated one, it is likely to receive the class label of its neighbors (i.e. the local

* Corresponding author.

http://dx.doi.org/10.1016/j.neucom.2014.12.086
0925-2312/

majority class). If the noise is uniform, most of the incorrectly labeled instances are far from the classification boundaries. They can therefore be viewed as isolated instances. In such cases, using smaller sampling ratios reduces the influence of these isolated noisy instances. Consequently, the ensemble becomes more robust.

In summary, this paper presents a comprehensive empirical assessment of the accuracy and robustness of bagging and random forest ensembles as a function of the bootstrap sampling ratio. This study extends our previous work [42] including more datasets, algorithms and experiments. In addition, we illustrate how small margin ensembles can be resilient to class-label noise.

The paper is organized as follows: Section 2 reviews previous work on label noise, focusing on classification ensembles. Section 3 is devoted to exploring the relation between margin and accuracy for different bootstrap sampling ratios and noise levels. In Section 4 we present the results of an extensive empirical evaluation of the performance of bagging and random forest ensembles built using subsampling. The experiments are carried out in a wide range of classification tasks with different amounts of class-label noise. Finally, the conclusions of this investigation are summarized in Section 5.

2. Related work

Poor data quality and contamination by noise are unavoidable in many real-world classification problems [18,53]. This has a strong potential to mislead the learning algorithms used for automatic induction from these data. Two types of noise can be present in these problems: class-label noise and polluted feature values [18,53]. Class-label noise is the consequence of incorrect manual labeling, missing information or failures in the data measuring process. Feature noise is often the result of a faulty data gathering process [18,53]. Class-label noise typically has a more pronounced misleading effect than feature noise, except when most of the feature values are corrupted [53]. Frénay and Verleysen [18] identify three types of label noise, characterized by different statistical models: The Noisy Completely at Random Model (NCAR), in which the probability of a class-label error is independent of the values of the features, the actual class of the instance and the noise rate. To simulate this type of noise the class labels of randomly selected instances are changed to a different class label, also at random. The second model is Noisy at Random (NAR). Labelling errors in this model are assumed to occur with a different probability for each class. NAR is useful to characterize tasks in which some classes are more susceptible to mislabeling than others. The third model is Noisy Not at Random (NNAR). In this case, the probability of an error depends on the actual class label and on the values of the features. This model should be used when some regions of the feature space, such as boundaries or sparse regions, are more prone to noise than others. Noise can be handled in a preprocessing step (data cleansing) or during the learning process, assuming that the algorithms used for induction from the contaminated data are robust [18].

2.1. Data cleansing

To mitigate their harmful effects, noise and outliers can be eliminated in a preprocessing step, before the selected learning algorithm is applied. For instance, it is possible to use statistical models or clustering-based methods to detect outliers. Patterns and association rules can also be used in the cleansing process [27]. An example of a pattern-based data cleansing algorithm is described in [45,44]. In this method, local SVM's are used to identify and remove instances that are suspected to be noise. For each particular training instance, k-NN is applied to locate nearby instances. A SVM is then trained on these instances to find the optimal separating hyperplane in that neighborhood. If the label predicted by this locally trained SVM does not coincide with the actual label, the instance is identified as noisy and discarded. This cleansing method has been tested on real and artificial datasets, where it showed improvements over k-NN. In [51], noisy instances are removed based on wrappers of different classification methods. In this study, the best results were obtained by removing or cleaning instances based on the prediction of a SVM built with the rest of the training data. Noisy instances are often included in the set of support vectors by a SVM classifier. Based on this observation, Fefilatyev et al. [16] propose to manually remove support vectors that are identified as noise by an expert. Then, a new SVM is built on the cleansed dataset. This process is iterated until no more support vectors are identified as noisy instances.

2.2. Robust learning algorithms

Another strategy to deal with noise is the design of robust learning algorithms. For instance, pruning is used in decision trees to reduce overfitting: the presence of noise tends to increase the size of the decision trees induced from the contaminated training data. Pruning is thus an effective way to improve the robustness of decision trees [12,13]. Another robustifying strategy is to explicitly incorporate in the learning algorithm the fact that the values of the features and the class labels can be polluted by noise. This strategy is adopted in the construction of Credal Decision Trees [28]. These types of trees are grown using the Imprecise Info-Gain Ratio (IIGR) as a splitting criterion. In this method the values of the features and class labels are approximated using probabilities and uncertainty measures.

It is also possible to adapt the algorithms used to build Support Vector Machines to improve their robustness to class-label noise. For instance, in [47] the hinge loss is replaced by a related loss function that takes into account the amount of noise in the data. With this loss function the optimization problem becomes non-convex. Heuristic optimization methods are then used to search for the global minimum of this non-convex problem. Promising results were obtained by this robust SVM in problems with asymmetric class noise (NAR model). A drawback of this method is that it is necessary to estimate the amount of noise in the data. Another robust version of SVM, called P-SVM (Probabilistic SVM) is proposed in [37] to classify magnetic resonance medical images. The P-SVM takes as inputs not only class labels but also class probability estimates. These probabilities are used to estimate the confidence on the labeling of each instance. The lower the confidence on the label, the lower the weight of that instance in the learning process. A practical limitation of this method is that one needs both qualitative (class labels) and quantitative (class posterior probabilities) information on the classes.

The problem of induction from noisy data has also been extensively addressed in the area of ensemble learning. In [2], Ali and Pazzani analyze the behavior of multiple classifier systems in the presence class-label noise. They observed that the improvements of the ensemble with respect to a single learner are generally smaller when the training data are contaminated with class-label noise. However, the reduction is not uniform and depends on the type of ensemble used.

Noise is not always harmful. In fact, noise injection is a powerful regularization mechanism that has the potential of improving the generalization capacity and robustness of prediction systems. In particular, randomization is used to build diverse ensembles that have good generalization capacity [4,38,10,15,11,34,36,31,29,17,30,49]. Furthermore, randomized ensembles, such as bagging and random forests, have been shown to be robust classifiers. By contrast, adaptive ensembles, such as boosting, are