Neurocomputing 160 (2015) 120–131

Contents lists available at ScienceDirect

Neurocomputing

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

Relating ensemble diversity and performance: A study in class noise detection

Borut Sluban [a,*], Nada Lavrač [a,b]

[a] Jožef Stefan Institute, Jamova 39, 1000 Ljubljana, Slovenia
[b] University of Nova Gorica, Vipavska 13, 5000 Nova Gorica, Slovenia

ARTICLE INFO

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

Keywords:
Class noise
Label noise
Noise detection
Ensemble methods
Diversity measures

ABSTRACT

The advantage of ensemble methods over single methods is their ability to correct the errors of individual ensemble members and thereby improve the overall ensemble performance. This paper explores the relation between ensemble diversity and noise detection performance in the context of ensemble-based class noise detection by studying different diversity measures on a range of heterogeneous noise detection ensembles. In the empirical analysis the majority and the consensus ensemble voting schemes are studied. It is shown that increased diversity of ensembles using the majority voting scheme does not lead to better noise detection performance and may even degrade the performance of heterogeneous noise detection ensembles. On the other hand, for consensus-based noise detection ensembles the results show that more diverse ensembles achieve higher precision of class noise detection, whereas less diverse ensembles lead to higher recall of noise detection and higher *F*-scores.

1. Introduction

In data mining, the success of learning and knowledge discovery from the data depends on various factors, including data quality. The quality of real-life data is frequently degraded due to errors and other data irregularities that are usually referred to as *noise*. The presence of noise has adverse effects on the quality of information retrieved from the data, models created from the data and decisions made based on the data [1]. Given that identifying noisy instances in the data and removing or correcting them proved to be beneficial in various applications, noise identification and filtering became an established area of machine learning and data mining research [2].

Noise in the data manifests itself as attribute noise (errors or unusual attribute values), class noise (wrong instance labels), or a combination of both. Noise detection algorithms are designed to identify erroneous data instances, which are typically found as those deviating from the expected distribution or not following a general pattern or model describing the data. Since every noise detection approach may perform best on a certain domain or on a certain type of noise, the overall noise detection performance can be improved by using ensembles of noise detection algorithms.

Ensemble learning methods are algorithms that construct a set of prediction models (an ensemble) and combine their outputs to a single prediction [3]. Ensembles are typically used with the purpose of improving the performance of simple base learning methods. The strength of ensemble methods lies in their ability to correct errors made by some of their members [4]. Therefore, ensemble members have to be diverse in terms of the errors they make, so that their combination can reduce the total prediction error [5]. Ensembles with greater diversity among their members tend to result in higher predictive accuracy [6].

Diversity among the members of an ensemble can be achieved in different ways, resulting in homogeneous or heterogeneous ensembles. On one hand, in homogeneous ensembles all ensemble members use the same learning algorithm. Popular methods based on boosting [7] and bagging [8], which construct homogeneous ensembles, diversify ensemble members by training them on differently selected subsets of the training data. Some approaches prefer to use different parameter settings of algorithms in the training phase to obtain different classifiers. Other approaches, like the Random Forest algorithm [9], use different feature subsets for training the base classifiers. On the other hand, heterogeneous ensembles are constructed from different base algorithms. It was shown that heterogeneous ensembles are more diverse [10] and that they provide better results than homogeneous ensembles [11]. Heterogeneous ensembles can be constructed by *ensemble selection* [12–14] or *ensemble pruning* [15,16], or be used for meta-learning called *stacking* [17,18]. Ensemble selection and ensemble

* Corresponding author.
E-mail address: borut.sluban@ijs.si (B. Sluban).

http://dx.doi.org/10.1016/j.neucom.2014.10.086
0925-2312/

pruning try to select the base classifiers by balancing the diversity and the performance of the ensemble, while stacking constructs a higher-level predictive model based on the predictions of the first-level base models.

Various measures for assessing the diversity of classifiers have been proposed in the literature [19,5,20,16,21]. The influence of diversity on ensemble performance has been extensively explored for classification problems by observing classification accuracy and classification error rates [5,22–25]. Some studies observed a positive correlation between diversity and classification accuracy [6,26], whereas others doubted that diversity measures can be used as means for improving classification performance [22,27].

In contrast with the above studies of the effects of ensemble diversity on classification accuracy, this paper focuses on the effects of ensemble diversity on the performance of explicit noise detection, which can be used for data cleaning, improved data understanding, and semi-supervised outlier identification, as studied in [28–32]. In these tasks, the main goal is to achieve high performance of explicit noise detection, rather than to increase the classification accuracy of learning algorithms applied after the noise filtering step. In the paper we explore the relation between different diversity measures and the performance of explicit noise detection, achieved by heterogeneous noise detection ensembles.

To the best of our knowledge, this is the first study that directly addresses the relation between different diversity measures and the performance of heterogeneous noise detection ensembles. Note that ensemble-based approaches to noise detection found in the literature recognize the diversity among ensemble members as a requirement for good ensemble performance, however they cope with ensemble diversity only indirectly. Commonly a heterogeneous set of presumably diverse approaches to noise detection is used [32–36], or the diversity of noise identification models is achieved by sampling of the training space and by random selection of features [37–41], or a combination of both approaches is adopted [42,43]. The reason for not explicitly measuring ensemble diversity may lie in the absence of a uniformly accepted definition of diversity. To fill this void, this work studies the relation between different commonly used diversity measures and the performance of various noise detection ensembles. In further work, these results can be used as guidance in the construction of noise detection ensembles.

The rest of the paper is structured as follows. Section 2 introduces the noise detection algorithms, the performance measures used in the evaluation of explicit noise detection, and the measures used for measuring the diversity of ensembles of noise detection algorithms. In Section 3 the aim of the paper is further clarified by presenting the research hypothesis and the goals, followed by the proposed methodology and experimental setting used in evaluating the relation between ensemble diversity and noise detection performance. The experimental results are presented in Section 4. The paper concludes in Section 5 with a discussion of the obtained results and directions for further work.

2. Preliminaries

This section introduces the basic methods and measures required for studying the relation between ensemble diversity and noise detection performance. First, class noise detection is described, second the performance measures for noise detection evaluation are specified, and finally, a selection of commonly used ensemble diversity measures is presented.

2.1. Noise detection

Class noise denotes errors in the labels assigned to data instances. From a wide variety of noise handling techniques [2], we chose a popular class noise detection approach proposed in [33], which became to be later known as *classification noise filtering*. This approach uses classification algorithms to identify wrongly labeled data instances. It works in a k-fold cross-validation manner, where in k repetitions $k-1$ folds of the dataset are used for training of a classification algorithm and the complementary fold is used for classifier validation. The instances that are misclassified on the validation folds are identified as noisy. The concept of classification noise filtering is illustrated in Fig. 1.

In the experiments we will investigate the performance of heterogeneous ensembles of classification noise filters, employing different learning algorithms as base classifiers for noise detection. A noise detection ensemble E of size L is formed of a set of algorithms $\{A_1, \ldots, A_L\}$ that are used for noise identification. The individual classifiers can be combined to the final ensemble prediction using different combination rules [19]. Predictions of algorithms that return label outputs (like 'noise' and 'non-noise') can be combined using different voting schemes. Two most commonly used voting schemes for combining ensemble predictions are the following.

- *Majority* (*plurality*) *voting*: If more than half of the algorithms A_i from E identify an instance $\mathbf{x}$ as noisy, then the ensemble declares it as noisy.
- *Consensus* (*unitary*) *voting*: If all the algorithms A_i from E identify the instance $\mathbf{x}$ as noisy, then the ensemble declares it as noisy.

Let function δ be 1 for 'noisy' labels and 0 otherwise. Then the formal notation of the condition for noise identification of instance $\mathbf{x}$ by ensemble E using the majority voting scheme can be written as $\sum_{i=1}^{L} \delta(A_i(\mathbf{x})) > L/2$, and using the consensus voting scheme as $\sum_{i=1}^{L} \delta(A_i(\mathbf{x})) = L$.

2.2. Performance measures

Quantitative evaluation of noise detection methods requires to know which are the noisy instances in a dataset. In real-life

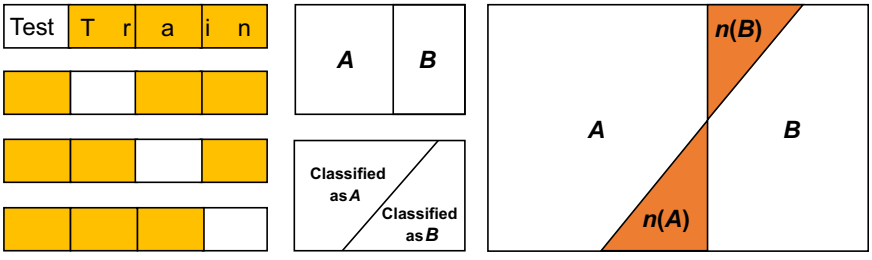

Fig. 1. Classification filtering using cross-validation. A and B are the class labels of instances in the test fold. The misclassified instances of A and B, denoted with $n(A)$ and $n(B)$, present the noise detected by the classification filter.