Knowledge-Based Systems 90 (2015) 153–164

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Using the One-vs-One decomposition to improve the performance of class noise filters via an aggregation strategy in multi-class classification problems

Luís P.F. Garcia [a,*], José A. Sáez [b], Julián Luengo [c], Ana C. Lorena [d], André C.P.L.F. de Carvalho [a], Francisco Herrera [e,f]

[a] Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, Trabalhador São-carlense Av. 400, São Carlos, São Paulo 13560-970, Brazil
[b] ENGINE Centre, Wrocław University of Technology, Wybrzeże Wyspiańskiego 27, Wrocław 50–370, Poland
[c] Department of Civil Engineering, Computing Systems and Languages, University of Burgos, Burgos 09006, Spain
[d] Instituto de Ciência e Tecnologia, Universidade Federal de São Paulo, Talim St. 330, São José dos Campos, São Paulo 12231-280, Brazil
[e] Department of Computer Science and Artificial Intelligence, University of Granada, CITIC-UGR, Granada 18071, Spain
[f] Faculty of Computing and Information Technology, King Abdulaziz University, North Jeddah, Saudi Arabia

ARTICLE INFO

Article history:
Received 12 March 2015
Revised 25 August 2015
Accepted 22 September 2015
Available online 9 October 2015

Keywords:
Noisy data
Class noise
Noise filters
Decomposition strategies
Classification

ABSTRACT

Noise filters are preprocessing techniques designed to improve data quality in classification tasks by detecting and eliminating examples that contain errors or noise. However, filtering can also remove correct examples and examples containing valuable information, which could be useful for learning. This fact usually implies a margin of improvement on the noise detection accuracy for almost any noise filter. This paper proposes a scheme to improve the performance of noise filters in multi-class classification problems, based on decomposing the dataset into multiple binary subproblems. Decomposition strategies have proven to be successful in improving classification performance in multi-class problems by generating simpler binary subproblems. Similarly, we adapt the principles of the One-vs-One decomposition strategy to noise filtering, making the noise identification process simpler. In order to integrate the filtering results achieved in the binary subproblems, our proposal uses a soft voting approach considering a reliability level based on the aggregation of the noise degree prediction calculated for each binary classifier. The experimental results show that the One-vs-One decomposition strategy usually increases the performance of the noise filters studied, which can detect more accurately the noisy examples.

1. Introduction

Real-world data usually contain errors or noise [1–4]. In classification problems, a classification model must be induced from labeled examples and this classifier should be capable of reliably predicting the true class of new examples. The correct assignment of class labels to the training examples has a strong impact on the predictive quality of the induced classifiers. Thus, errors in the class labeling of the training examples may severely harm the predictive performance and complexity of the induced classifiers [1,5,6]. This type of error is known in the literature as *class noise* or *label noise* [2].

In the case of multi-class classification problems, binary decomposition strategies [7] are usually employed to allow the usage of well-known algorithms originally proposed for binary classification problems, such as *Support Vector Machines* (SVM) [8], in multi-class tasks. These strategies decompose the original problem into several binary subproblems of a lower complexity. The most popular decomposition schemes are *One-vs-One* (OVO) [9], which induces a classifier to distinguish between each pair of classes, and *One-vs-All* (OVA) [9], which induces a classifier to distinguish each class from all other classes.

The behavior of the OVO strategy in presence of noise was studied by Sáez et al. in [10]. In order to analyze whether OVO was able to reduce the harmful effects of noise in the classification results, several classification algorithms with and without the usage of this decomposition were compared. The experimental results showed that, in the presence of noisy data, decomposition generally offers better classification performance than solving the original multi-class

* Corresponding author. Tel.: +551633738161; fax: +551633739633.
E-mail addresses: lpgarcia@icmc.usp.br, lpfgarcia@gmail.com (L.P.F. Garcia), jose.saezmunoz@pwr.edu.pl (J.A. Sáez), jluengo@ubu.es (J. Luengo), aclorena@unifesp.br (A.C. Lorena), andre@icmc.usp.br (A.C.P.L.F. de Carvalho), herrera@decsai.ugr.es (F. Herrera).

http://dx.doi.org/10.1016/j.knosys.2015.09.023
0950-7051/

problem. These improvements are mainly attributed to the distribution of the noisy examples in the binary subproblems. Furthermore, the separability of the classes is increased, while it is also possible to collect information from different classifiers.

Another alternative to overcome the problems resulting from the presence of class noise is the usage of noise filtering techniques, which remove potentially noisy examples in a preprocessing step [11,12]. Several studies show the benefits from their usage regarding improvements in the classification predictive performance and the reduction in the complexity of the classifiers built [5,13–15]. Noise filters can use different information to detect noise, such as those employing neighborhood or density information [11,16,17], descriptors extracted from the data [13,18] and noise identification models induced by classifiers [13] or ensembles of classifiers [5,14,19,20]. In other papers, they are also used to remove predictive noise [21] and investigate the presence of noise in imbalanced datasets [22,23]. Since each filter has a bias, it may have a distinct performance depending on the data used [24,25]. Thus, it is common the existence of a margin of improvement on the noise detection accuracy of filtering methods.

This paper investigates a new approach to detect and remove label noise in multi-class classification tasks. This approach combines the OVO multi-class decomposition strategy with a group of noise filtering techniques. In this combination, each noise filter, instead of being applied to the original multi-class dataset, is applied to each binary subproblem produced by the OVO strategy. Each noise filter assigns to each training instance a degree of confidence of the example being noisy, named *noise degree prediction* (NDP), which is a real number. However, some noise filters only output two values: noisy and not noisy. If so, the noise filter is adapted to output NDPs. For each training instance, the NDPs obtained from all noise filters are combined using a soft voting strategy, producing a unique NDP for the instance. The strategy adopted in this paper is to remove a fixed number of the examples with highest NDP values.

The proposed approach has three main advantages: (i) it does not require any modification in the concept and the bias of the noise filters; (ii) it provides for each training instance a combined degree of confidence regarding noise identification and; (iii) it does not make any assumptions about the noise characteristics.

In order to evaluate the impact of using the OVO strategy for noise filtering in multi-class tasks, we present an empirical study using several well-known noise filters found in the literature that will be adapted for soft voting [5,13,14,16,20] and a large amount of datasets with different levels of class noise [1]. The differences between the filtering with and without decomposition will be analyzed based on the accuracy of the noise filters detecting the noisy examples in each scenario.

The rest of this paper is organized as follows. Section 2 points out the main motivations for this study, presenting an overview on noise filtering techniques and the motivations for the use of decomposition strategies in multi-class problems. Section 3 details the approach proposed for noise detection. Section 4 describes the experimental framework, whereas Section 5 analyzes the experimental results obtained by the noise filters with and without decomposition. Finally, Section 6 presents the main conclusions from this study. A website with additional information, such as the datasets employed and the results of each noise filter is available at http://www.biocom.icmc.usp.br/~lpfgarcia/ovo.

2. Preliminaries

This section presents the background to support our proposal. Section 2.1 describes the main aspects of class noise treatment with a brief overview of the noise filtering techniques employed. Then, Section 2.2 introduces the usage of binary decomposition strategies that are commonly employed in multi-class classification.

2.1. Class noise treatment by noise filtering

Noise filters [5,13–16,20] are preprocessing methods commonly used to identify and remove noise in a dataset [2]. Most of the existing filters focus on the elimination of examples with class noise, which has shown to be advantageous [18]. In contrast, the elimination of examples with feature noise is not as beneficial [1], since other attributes from these examples may be useful to build the classifier.

Most of the noise filters [5,14,20] adopt a *crisp decision* for noise identification, classifying each training example either as either noisy or safe. *Soft decision* strategies, on the other hand, assign a noise degree prediction to each example, NDP values. The soft decision helps to correctly identify examples, those whose identification as noisy is more difficult. Besides, it makes easier the combination of multiple filters, a strategy proposed in this paper.

Next, the noise filters used in the experiments performed for this study are briefly presented. Since they were all proposed for crisp noise detection, their adaptation to allow soft decision is also discussed. The following filtering methods were used in this study, each belonging to a different filtering paradigm:

1. **All-*k*-NN** (AENN) [16]. Distance-based approaches uses the *k*-NN decision rule [16,26] to identify noisy data. Techniques following this approach assume that an example is likely to be noisy if it is located close to other examples from a different class. These noise filters are able to remove examples with class noise and examples lying on the decision border, which increases the margin of separation between the classes. A well known technique from this group is *All-k-NN* (AENN) [16]. This filter applies, iteratively, the *k*-NN classifier with several increasing values of *k*. Examples misclassified by their neighbors are marked as noisy and eliminated from the dataset. The soft version of this technique estimates the NDP of an example as the percentage of times it is labeled as noisy in different iterations.
2. **Prune Saturation Filter** (PruneSF) [13]. Complexity-based approaches extract complexity measures from the training data [13,18]. For instance, the *Saturation Filter* (SF) [13] exhaustively looks for examples that reduce a metric called *Complexity of the Least Correct Hypothesis* (CLCH) associated with a dataset. The size of a Decision Tree (DT) without pruning is used to estimate the CLCH value [13]. If the removal of an example reduces the CLCH value, it is marked as noisy. Next, the method carries out a new search in the dataset without this example and repeats the same procedure until no example is marked as noisy or a stopping criterion is reached. PruneSF [13] is based on SF. It uses a DT with pruning in a previous step to overcome computation time restrictions. Therein, first a pruning step removes all examples misclassified by a pruned DT, which are regarded as noisy. Afterwards, the iterative procedure described for SF is performed. In our work, a soft decision is obtained by firstly ranking all examples removed in the pruning step as noisy with a probability of 1. Next, the examples are ranked according to their CLCH values, which are normalized to give their probability of being noisy.
3. **High Agreement Random Forest** (HARF) [20]. This is a well-known classifier-based filter that uses a *Random Forest* classifier [27]. This technique considers the rate of disagreement in the predictions made by the individual trees in the forest to detect the noisy examples: if this rate is high, the example is probably noisy; otherwise, it is considered to be clean. A soft decision for this filter can be obtained by the percentage of base trees that disagree on their predictions for a particular instance.
4. **Static Ensemble Filter** (SEF) [5]. Ensemble-based approaches employ ensembles of classifiers to identify the noisy examples [5,14,20]. Their motivation is that different classification models provide a better alternative for detecting mislabeled examples than using information from a single model only [5]. SEF