Neurocomputing 176 (2016) 14–25

Contents lists available at ScienceDirect

Neurocomputing

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

Noise detection in the meta-learning level

Luís P.F. Garcia [a,*], André C.P.L.F. de Carvalho [a], Ana C. Lorena [b]

[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] Instituto de Ciência e Tecnologia, Universidade Federal de São Paulo, Talim St. 330, São José dos Campos, São Paulo 12231-280, Brazil

ARTICLE INFO

Article history:
Received 18 March 2014
Received in revised form
18 November 2014
Accepted 5 December 2014
Available online 8 May 2015

Keywords:
Noise identification
Characterization measures
Complexity measures
Meta-learning

ABSTRACT

The presence of noise in real data sets can harm the predictive performance of machine learning algorithms. There are several noise filtering techniques whose goal is to improve the quality of the data in classification tasks. These techniques usually scan the data for noise identification in a preprocessing step. Nonetheless, this is a non-trivial task and some noisy data can remain unidentified, while safe data can also be removed. The bias of each filtering technique influences its performance on a particular data set. Therefore, there is no single technique that can be considered the best for all domains or data distribution and choosing a particular filter is not straightforward. Meta-learning has been largely used in the last years to support the recommendation of the most suitable machine learning algorithm(s) for a new data set. This paper presents a meta-learning recommendation system able to predict the expected performance of noise filters in noisy data identification tasks. For such, a meta-base is created, containing meta-features extracted from several corrupted data sets along with the performance of some noise filters when applied to these data sets. Next, regression models are induced from this meta-base to predict the expected performance of the investigated filters in the identification of noisy data. The experimental results show that meta-learning can provide a good recommendation of the most promising filters to be applied to new classification data sets.

1. Introduction

Many Machine Learning (ML) textbooks define noise as any unwanted anomaly in the data [1]. This anomaly can be interpreted as errors in the predictive features caused by imprecisions during information recording, errors in the output values introduced by imprecisions in the data labelling process and, in extreme cases, errors arising from absent information. The presence of noise adds random components into the overall data structure, since some points will be shifted in the input space. When corrupted data sets are used in the induction of supervised ML models, the predictive performance of these models for new data may be harmed. Moreover, the induced models can be overly complex and require long processing times. In order to overcome these problems, various techniques have been developed to identify and treat noisy data [2].

The identification of noise in classification data sets has been the subject of several studies. These studies follow two main approaches: (1) designing classification techniques more tolerant and robust to noise [2] and (2) data cleaning in a previous preprocessing step [3].

Several data preprocessing techniques for noise identification can be found in the literature [4–9]. These techniques, referred as filters, look for suspicious examples in a data set, which are usually removed afterwards. Some of these techniques use neighborhood or density information to determine whether an example is noisy [4,5], while others use ensembles of classifiers [6–8]. There are also techniques based on descriptors extracted from the data [9]. Despite the technique employed, it is usually not possible to guarantee if a given instance is really a noisy instance.

The bias of each filtering technique influences its performance on a particular data set. Thus there is no single technique that can be considered the best for all domains or data distributions and choosing a particular filter technique is a difficult task.

This paper is concerned with noise identification in classification data sets, particularly noise introduced in the class label of the examples. It also investigates the use of meta-learning (MTL) to recommend the most promising filter technique for label noise identification in a new data set. Although MTL is frequently associated with the recommendation of a ML algorithm for a new classification data set [10–12], it has been recently used in different types of problems [13–16]. In this study, meta-regressors are induced to estimate the performance of noise filter techniques

[*] Corresponding author. Tel.: +55 16 3373 8161; fax: +55 16 3373 9633.
E-mail addresses: lpfgarcia@gmail.com, lpgarcia@icmc.usp.br (L.P.F. Garcia), andre@icmc.usp.br (A.C.P.L.F.d. Carvalho), aclorena@unifesp.br (A.C. Lorena).

http://dx.doi.org/10.1016/j.neucom.2014.12.100
0925-2312/© 2015 Elsevier B.V. All rights reserved.

regarding label noise identification. We believe that a good predictive performance in the estimation of the filter performance will lead to better label noise identification in new data sets.

In order to induce these meta-regressors, a set of diverse classification data sets was collected. Next, controlled label noise, with different noise degrees, was inserted into these data sets, resulting in several noisy data sets. Afterwards, well known filters found in the literature were applied to these noisy data sets and their performance regarding noise identification was recorded. A meta-base was created by extracting meta-features from these noisy data sets as predictive features and the performance of the filters as label features. Each noisy data set is represented by an example in the meta-base.

The meta-features used here describe various characteristics for each data set, including its expected complexity level [17]. The examples in this meta-base are labelled with the *F*-score performance achieved by the filters in noise identification. Using such meta-base as input, ML techniques from different paradigms are employed to induce meta-regressors. Therefore, they are trained to predict the expected *F*-score performance of some filters in noise identification. Overall, our hypothesis is that the choice of a filter can be based on the main characteristics of a data set, which include descriptors of data conformation, dimensionality and complexity, among others.

The induced meta-regressors achieved good predictive performance in an extensive set of practical experiments with various data sets. The results obtained show that MTL can support the recommendation of a noise identification filter. Moreover, this confirms our hypothesis that each filter technique is best suited for data sets with specific conformations and has a particular area of competence, which should be taken into account when choosing a filter for a new data set. The proposed MTL-based approach allows the identification of the most suited filter for a new data set.

The main contributions from this study can be summarized as follows:

- Comparison of the performance of several popular filters from the literature for various data sets with distinct noise levels.
- Proposal of a new MTL approach based on the induction of meta-regressors able to predict the expected performance of those filters in the identification of noisy data.
- Show the relevance of MTL as a decision support tool for the recommendation of a suitable noise filter technique for a new classification data set.

The paper is organized as follows. Section 2 points out the main motivations of this study and presents an overview of the noise filter techniques investigated in this study. Section 3 presents the proposed MTL-based approach and the meta-features used to describe the data sets. Section 4 describes the experiments carried out to validate our proposal, while Section 5 reports and analyzes the experimental results obtained. Finally, Section 6 summarizes the main conclusions from this study and points out directions for future work.

2. Noise detection

In classification data sets, noise can be either present in the data predictive features or in the class labels [18]. The study in this paper considers only the label noise scenario. Label noise can be more harmful to ML algorithms, since many classification techniques use an objective function based on class labels, like the predictive error (Artificial Neural Networks [19]) and the homogeneity of the examples within the classes (Decision Tree Induction Algorithms [20]). Moreover, high noise levels in predictive features can ultimately lead to a wrong labelling of an example, when the example is moved to the wrong side of the decision border.

Many works directly related to noise detection and elimination show the effectiveness of filter techniques for label noise identification [4–9,21]. A filter technique, here named simply filter, is a preprocessing technique that can be applied to any given data set, outputting the potential noisy examples. Different strategies can be used to assess if an example is noisy or not: ensembles of classifiers [6,8,21], cross validation [7], descriptors of the complexity of the hypothesis [9] or even the distance between examples [4,5].

Several authors use ensembles of classifiers for noise identification [6,8]. The motivation for using ensembles is that if distinct classifiers disagree on their predictions for an instance, the instance is probably incorrectly labelled. In [6], for instance, the authors describe strategies to combine the prediction of distinct classifiers for noise identification. According to the authors, the majority vote of the predictions made by *k*-nearest neighbor (*k*-NN) [22], C4.5 [20] and Support Vector Machines (SVM) [23] with 10 fold cross validation presented the best predictive performance. Here we will refer to this technique as Static Ensemble Filter (SEF), because the set of classifiers composing the ensemble is fixed.

In [8], the authors propose a Dynamic Ensemble Filter (DEF). In DEF, the set of classifiers to be combined is specific for each data set and is chosen based on a criterion that considers the agreements in the predictions made by the individual classifiers. Thus, the set of classifiers combined is dynamically adapted for each data set. A majority vote of the predictions is used to assess whether an example is noisy, similar to [6]. This ensemble approach is used here, combining the classifiers with the best 10 fold cross validation predictive performance on training data. As shown in the experimental results (Section 5), this simple criterion created ensembles with noise identification predictive capabilities.

Another recent ensemble is the High Agreement Random Forest Filter (HARF) method [21], which uses Random Forest (RF) classifiers for noise identification. The algorithm considers the rate of disagreement in the predictions made by the individual forest trees using 10 fold cross validation to detect the noisy examples: if the rate is relatively high (70–90%), the example is probably noisy; otherwise, it is considered to be clean.

The Cross-validated Committees Filter (CVCF) algorithm, proposed in [7], induces a classification model using 10 fold cross validation. Any ML technique can be used to induce the classification model. Examples from the training fold wrongly classified by this model are considered as potential noise. The number of times an example is marked as noisy is used to assess its reliability. If the example was marked as noisy most of the time, CVCF will consider the example to be noisy.

Gamberger et al. proposed Saturation Filter (SF), which uses the first order language representation of a data set to filter noisy examples [9]. For such, SF exhaustively looks for examples that reduce a value named Complexity of the Least Correct Hypothesis (CLCH) associated with a data set. To estimate the CLCH value, the problem is first represented in first order language. Next, this formalized data set is fed to a filter, which randomly selects one example, from a set of examples previously highlighted as potential noise, and verifies if the CLCH value decreases when this example is removed. If so, the example is marked as noisy. This procedure continues until no example is marked as noisy or until a stop criterion is reached. Since the number of iterations required by SF is usually high, Sluban et al. [3] proposed the Pruned SF (PruneSF) algorithm to reduce this number. PruneSF uses a Decision Tree (DT) to estimate the CLCH value with and without pruning.

Some noise filtering techniques are based on the distance between examples and employ the *k*-NN algorithm [24,5]. They consider an example to be consistent if it is close to other examples from its class. Otherwise, it is either probably incorrectly labelled or