Expert Systems with Applications 42 (2015) 5065–5075

Contents lists available at ScienceDirect

Expert Systems with Applications

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

A biclustering approach for classification with mislabeled data

Fabrício O. de França[a], André L.V. Coelho[b,*]

[a] Center of Mathematics, Computing and Cognition (CMCC), Federal University of ABC (UFABC), Brazil
[b] Graduate Program in Applied Informatics, Center of Technological Sciences, University of Fortaleza (UNIFOR), Brazil

ARTICLE INFO

Article history:
Available online 6 March 2015

Keywords:
Classification
Noisy labels
Biclustering
Feature learning

ABSTRACT

Labeling samples on large data sets is a demanding task prone to different sources of errors. Those errors, denoted as noise, can significantly impact the performance of a classification algorithm due to overfitting of wrongly labeled data. So far, this problem has been treated by avoiding the overfitting and correcting mislabeled data through similarity analysis. The former approach can be affected by the curse of dimensionality and some mislabeled data will not be corrected. In this paper, we investigate the use of a biclustering approach to capture local models of coherence across subsets of instances and attributes. Those models are used to replace and augment the attributes of the original dataset. Through a systematic series of experiments, we have assessed the performance of the proposed approach, referred to as *BicNoise*, by considering different rates and types of label noise, and also different types of classifiers, binary datasets, and evaluation metrics. The good results achieved suggest that the transformed data can alleviate the dimensionality problem, reduce the redundancy of correlated features and improve the separability of the data, thus improving the classifier performance (most noticeably, in the highest noise settings).

1. Introduction

The investigation of the effects of noisy data on the performance of classification algorithms is a research line that has attracted a great deal of interest in machine learning (Atla, Tada, Sheng, & Singireddy, 2011; Nettleton, Orriols-Puig, & Fornells, 2010; Wu & Zhu, 2008; Zhu, Wu, Khoshgoftaar, & Shi, 2007). This is because datasets derived from real-world problems are usually plagued with several types of noise, bringing much uncertainty to the classifier induction process. The noise due to data mislabeling, in particular, which entails the modification (either random or not) of the observed labels assigned to the data instances (objects), can be potentially harmful and very difficult to cope with, since it can severely misconfigure the underlying relationships between the input (instance) and output (class) spaces (Frénay & Verleysen, 2014; Zhu & Wu, 2004).

A large body of work on the topic of supervised classification with label noise has emerged in the preceding years (Frénay & Verleysen, 2014). On one hand, there are approaches aiming at improving the quality of the noisy training data by modeling, detecting and then correcting, or simply removing, the affected cases. These methods are usually referred to as data cleansing methods (Brodley & Friedl, 1999; Guan, Yuan, Lee, & Lee, 2011; Zhu, Wu, & Chen, 2006). On the other hand, there are approaches, called noise-tolerant (or noise-robust), that can deal intrinsically with label noise while inducing the classifier models and, thus, do not depend on data preprocessing (Abellán & Moral, 2003; Abellán & Masegosa, 2012; Bootkrajang & Kabán, 2012, 2014). Finally, there are also some approaches, referred to here as hybrid ones, which combine features and properties of the abovementioned classes, e.g. by creating probabilistic models of label noise and then using this information to improve the noise-tolerance of the classifier during its training (Bouveyron & Girard, 2009; Rebbapragada & Brodley, 2007; Tabassian, Ghaderi, & Ebrahimpour, 2012a, Tabassian, Ghaderi, & Ebrahimpour, 2012b; Wang et al., 2012).

In this paper, we report on an empirical study investigating a novel approach for tackling the label noise problem. The approach, referred to as *BicNoise*, centers on the notion of biclusters (Cheng & Church, 2000; Madeira & Oliveira, 2004), i.e., submatrices of the training dataset showing high coherence of values across subsets of instances, attributes, and possibly class labels. By resorting to the local correlation models captured by the biclusters, we show that it is possible to elicit (learn) good discriminative features for improving the generalization performance of the induced classifiers. Different BicNoise variants are presented and assessed, which vary according to the way they modify the original dataset as well as to whether the label information of the training instances is used or not (supervised/unsupervised modes). Through a systematic series of experiments, we have assessed the performance of

[*] Corresponding author.
E-mail addresses: folivetti@ufabc.edu.br (F.O. de França), acoelho@unifor.br (A.L.V. Coelho).

http://dx.doi.org/10.1016/j.eswa.2015.02.045
0957-4174/© 2015 Elsevier Ltd. All rights reserved.

the BicNoise variants by considering different rates and types of label noise, and also different types of classifiers, binary datasets, and evaluation metrics.

The rest of the paper is structured as follows. In Section 2, we provide a brief survey on recent related work and a contrast on the label noise taxonomies independently conceived by Frénay and Verleysen (2014) and Rider, Johnson, Davis, Hoens, and Chawla (2013), which were both considered in the experiments reported in this work. Also in this section, we overview the main concepts associated with the biclustering task, giving special emphasis to the bicluster models and biclustering algorithm used in our experiments. In Section 3, we present in detail the BicNoise approach and its variants. Further, in Section 4, we outline the way the computational experiments were set up, and then present and discuss the several results achieved. Finally, Section 5 concludes the paper and provides remarks on future work.

2. Background

In the first subsection that follows, recent papers investigating the data mislabeling problem are overviewed. Then, two alternative classifications of the types of label noise are reviewed and contrasted. Finally, we focus specifically on the main concepts related to the biclustering task.

2.1. Related work

In a recent survey, Frénay and Verleysen (2014) have provided a comprehensive characterization of the problem of classification in the presence of label noise. Different definitions and sources of label noise were considered as well as several data cleansing, noise-tolerant and hybrid approaches were reviewed. Moreover, the authors have analyzed different statistical measures for validating the performance of algorithms within the label noise scenario.

Some noise-tolerant approaches for dealing with the data mislabeling problem are based on kernelized machines, such as the label-noise robust Kernel Logistic Regression classifier proposed by Bootkrajang and Kabán (2014). In this work, the authors have employed a multiple kernel learning setting, jointly with a Bayesian regularisation scheme, in order to determine the complexity parameters of the kernelized logistic regression models when no trusted validation set is available. Empirical results on 13 benchmark data sets and two real-world applications have demonstrated the success of the proposed approach.

Other approaches are based on the notion of classifier ensembles (Tabassian et al., 2012a; Guan, Yuan, Ma, & Lee, 2014). In this context, several classifiers or several variations of the same classifier are trained under the assumption that, for each mislabeled instance, only a minor set of the classifiers will learn their incorrect label through overfitting. Although interesting, one weakness of this approach has to do with the fact that whenever a mislabeled instance is located at the frontier of two or more classes, the majority of the classifiers will incorrectly learn a mistaken boundary for those classes.

Other approaches are variants of the Bagging and Boosting techniques (Abellán & Masegosa, 2012; Cantador & Dorronsoro, 2005; Cao, Kwong, & Wang, 2012). Here, the same classifier is trained by using different samples from the dataset, leading to different class boundaries that are combined afterwards. The main argument behind the Bagging schemes devised by Abellán and Masegosa (2012) in particular is that it is expected that each mislabeled instance will be part of just a few of the generated training sets, thus a majority of the boundaries will not be influenced by such instance. Even though the proposed schemes usually improve the performance, their success still strongly depends on which instances were mislabeled.

Finally, a more conceptually elaborated approach published in Expert Systems with Applications (Mantas & Abellán, 2014b; Mantas & Abellán, 2014a) makes use of the theory of Imprecise Probabilities, which deals with vague and conflicting information, for modeling the probability of each class. By using such theory, the authors have shown that the performance of the decision classifier C4.5 could be significantly improved when under the influence of noisy labels.

One should notice that the aforementioned approaches have two issues in common: (i) they do not entirely discard/transform the mislabeled data; and (ii) in most cases their performance is much dependent on the particular locations of the incorrect instances. These issues, however, do not appear in the approach investigated in the present paper, which aims at improving the separability of the classes by extracting novel features directly from the noisy data without resorting to any previous information (probabilistic or not) regarding the training set labels.

2.2. Statistical models of label noise

In most of the studies involving the classification of mislabeled instances, there is an implicit assumption that data are mislabeled completely at random. This assumption may be unrealistic in some real-world scenarios where multiple sources of systematic biases may happen during experimentation and data collection.

By mirroring the types of mechanisms usually considered in the missing value literature (Allison, 2002; Little & Rubin, 2002), Frénay and Verleysen (2014) and Rider et al. (2013) came up with two alternative taxonomies for modeling the different types of biases underlying the mislabeling process. According to the taxonomy of Frénay and Verleysen (2014), the three statistical models of label noise are defined as follows:

- *Noisy completely at random* (NCAR), whereby the mislabeling of an instance is viewed as a completely random process;
- *Noisy at random* (NAR), whereby the probability of mislabeling depends (solely) on the true class; and
- *Noisy not at random* (NNAR), whereby the mislabeling of an instance depends both on the true class and the particular values assumed by the instance attributes.

The authors emphasize that the first model can be considered as a special case of the second, while the third model is at the same time the more generic, complex, and realistic one.

On the other hand, the taxonomy adopted by Rider et al. (2013) was devised having in mind binary classification problems with a poorly defined negative class; that is, problems where it is known beforehand that only one of the classes can have its label flipped. The three types of biases accounting for the mislabeling of the positive class instances were defined as:

- *Biased completely at random* (BCAR), whereby the label change of the positive class occurs uniformly at random;
- *Biased at random* (BAR), whereby the mislabeling can be (completely) explained within the data (e.g. due to some property of the class and/or an explicit attribute of the dataset); and
- *Biased not at random* (BNAR), whereby the mislabeling happens as a consequence of some aspect not explicitly available in the dataset (e.g. due to some property of a latent attribute).

One should notice that this second taxonomy takes into account the possibility of having labeling errors due to external (latent) factors.