Neurocomputing 160 (2015) 3–17

Contents lists available at ScienceDirect

Neurocomputing

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

Comparison of two topological approaches for dealing with noisy labeling

Fabien Rico [a,d], Fabrice Muhlenbach [b,f,*], Djamel A. Zighed [a,c,e], Stéphane Lallich [a,c]

[a] Laboratoire ERIC, EA 3083 – 5, avenue Pierre Mendès-France – 69676 Bron Cedex, France
[b] Lab. Hubert Curien, UMR CNRS 5516 – 18 rue du Pr. Benoît Lauras – 42000 St-Étienne, France
[c] Institut des Sciences de l'Homme – ISH – USR 3385 CNRS – 14 av. Berthelot, 69007 Lyon, France
[d] Université Claude Bernard, Lyon I, Université de Lyon, France
[e] Université Lumière, Lyon II, Université de Lyon, France
[f] Université Jean Monnet, Saint-Étienne, Université de Lyon, France

ARTICLE INFO

Article history:
Received 17 March 2014
Received in revised form
10 October 2014
Accepted 11 October 2014
Available online 10 February 2015

Keywords:
Identification of mislabeled instance
Relaxation
Cut edges weighted
Topological learning
Separability index
Machine learning

ABSTRACT

This paper focuses on the detection of likely mislabeled instances in a learning dataset. In order to detect potentially mislabeled samples, two solutions are considered which are both based on the same framework of topological graphs. The first is a statistical approach based on Cut Edges Weighted statistics (CEW) in the neighborhood graph. The second solution is a Relaxation Technique (RT) that optimizes a local criterion in the neighborhood graph. The evaluations by ROC curves show good results since almost 90% of the mislabeled instances are retrieved for a cost of less than 20% of false positive. The removal of samples detected as mislabeled by our approaches generally leads to an improvement of the performances of classical machine learning algorithms.

1. Introduction

In machine learning, and more specifically in the framework of supervised learning, we more or less explicitly assume that the learning dataset might be noisy. Moreover, we suppose that this noise lies on the data related to the predictive variables. This noise may come from the lack of relevant predictors, from the small size of the learning sample, from the noise due to the observation/measurement tool of data acquisition, and so forth. On the other hand, we suppose that instances of the learning dataset are correctly labeled, therefore the predicted attribute is not corrupted. This is a major assumption rarely discussed in the literature of machine learning, in comparison to the noise in predictive variables. In this paper, we will deal with the noise in the labeling.

Nowadays, specifically in the world of big data, storing, managing and retrieving information in data warehouses require real time annotation processes for indexing and labeling the continuous flow of data. These processes, which may be automatic or sometimes manual, are often imperfect and therefore generate wrong annotations and mislabeled observations. For example, manual annotation in data stream of images, video, medical curves and so on, can generate mislabeling because of human limitations, especially with a high-speed work flow. Automatic annotation processes can also produce errors in labeling objects or situations because of artifacts or because of intrinsic limitations in the automatic process/devices. For more details about situations that cause mislabeling, see [1]. Classical machine learning algorithms are not designed to deal with such noise, even though some algorithms are considered to be robust to the noise in labels. Therefore, specific pre-processing must be carried out before the learning itself. Handling mislabeled data involves at least two tasks. The first identifies samples that are likely mislabeled and the second decides what to do with them. For this latter task, two options are possible: (i) each supposed mislabeled sample is withdrawn from the learning dataset, or (ii) the true label is restored for each of them according to a specific rule. Whatever the set of tasks accomplished in order to fix the noise in the labels, at the end of the day, what we expect is an improvement, or at least no deterioration of the performances of any classifier, in which "performance" is taken to mean the accuracy of the prediction on the test sample that is not noisy. However, we can observe that the performances of a classifier might decrease after the treatment of the noise. We will propose some explanations for this later on. For now, let us focus on the process of handling the mislabeled samples in a learning dataset.

In this paper, we will focus on the detection of likely mislabeled instances in the dataset, which is the keystone of our work. What to do next? Removing, restoring or doing something else with the

* Corresponding author.

http://dx.doi.org/10.1016/j.neucom.2014.10.087
0925-2312/

noisy samples that have been detected is another issue that we will discuss only briefly. We have designed two solutions for detecting potentially noisy labeling samples. Both are based on the same framework of topological graphs. The first is a statistical approach based on the Cut Edges Weighted statistics (CEW) in the neighborhood graph. The second is a relaxation technique (RT) that optimizes a local criterion in the neighborhood graph. Both solutions try to provide an estimate of the probability of the class Y for all points in the learning sample and, depending on this probability, the samples that likely belong to a class other than to the one given are declared suspicious (likely corrupted). To compare these two methods (CEW vs. RT), we use ROC (Receiver Operating Characteristic) curves. We have also carried out some evaluations with four classifiers: KSVM, Random Forest, 1-NN and AdaBoost. The goal is to check to what extent each classifier is improved by removing the suspicious samples from the learning dataset. Various levels of noise are tried.

This paper provides an update on approaches dealing with class noise that are based on topological graphs. It provides a synthesis of experiments conducted to evaluate and compare the two methods, both regarding the quality of the filtered learning set and the performance of various machine learning methods on this filtered learning set. This paper is organized in three main parts. The first introduces the concepts of topological learning using proximity graphs. It leads to a valuable tool that is a statistical test for assessing the separability of classes into a multidimensional representation space. We show that this statistic can be used to estimate the accuracy of any machine-learning algorithm. An evaluation of the relevance of this approach is carried out and discussed. In this first part, we suppose that there is no noise in the labeling. The second part deals with the noise in the labeling. It presents and compares two methods, CEW and RT using ROC curves. The third part assesses to what extent, removing likely mislabeled (suspicious) samples may improve the performance of well-known classifiers. The conclusion includes some future paths of research.

2. Definitions and notation

Throughout this paper we use the following notation and conventions.

Considering a global population Ω, the supervised learning methods aim to produce a model that predicts the unknown belonging class label $Y(i)$ of an instance i extracted from the global population Ω using its vector representation $X(i)$ associated with various real predictive attributes. The construction of the model requires a set of labeled data, called the learning set, denoted by Ω_L. We denote the size of the learning set by n, p the number of descriptive attributes, and k the number of categories of the class variable Y. The learning dataset Ω_L is a set of pairs $(X(i), Y(i)), i = 1, 2, \ldots, n$, where $Y(i)$ is the class label of i and $X(i) = (X_1(i), X_2(i), \ldots, X_p(i))$ is the p-dimensional vector corresponding to the representation of the instance i in the p-dimensional space according to the different predictive attributes. The quality of the model obtained is assessed on a test set, denoted by Ω_T, another dataset of labeled data which was not used during the learning step.

The learning ability of a given method is strongly associated with its class separability degree in $X(\Omega)$. We consider that the classes will be easier to separate, therefore to learn, if they fulfill the following conditions:

- the instances of the same class appear mostly gathered in the same subgroup in the representation space;

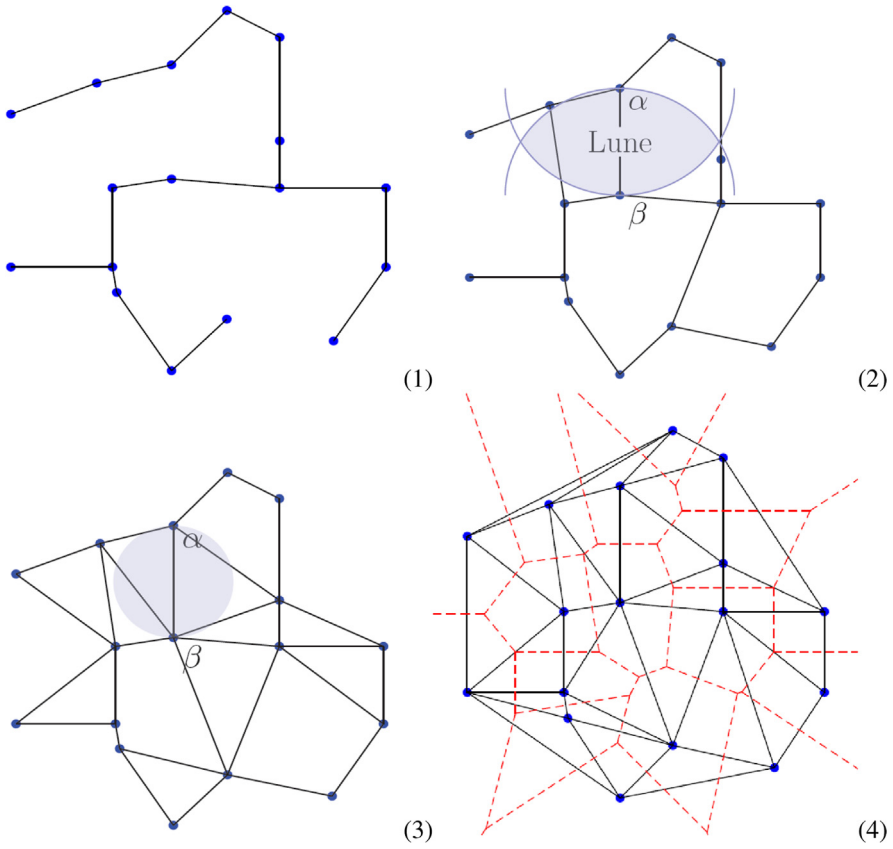

Fig. 1. Graphs and regions of influence: MST (1), RNG (2), GG (3), and DT (4).