Information Sciences 328 (2016) 42–59

Contents lists available at ScienceDirect

Information Sciences

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

On the stopping criteria for *k*-Nearest Neighbor in positive unlabeled time series classification problems

Mabel González [a], Christoph Bergmeir [b,*], Isaac Triguero [c,d], Yanet Rodríguez [a], José M Benítez [e]

[a] *Department of Computer Science, Universidad Central "Marta Abreu" de Las Villas, Cuba*
[b] *Faculty of Information Technology, Monash University, P.O. Box 63, Victoria 3800, Melbourne, Australia*
[c] *Department of Respiratory Medicine, Ghent University, Gent 9000, Belgium*
[d] *Data Mining and Modelling for Biomedicine group, VIB Inflammation Research Center, Zwijnaarde 9052, Belgium*
[e] *Department of Computer Science and Artificial Intelligence, E.T.S. de Ingenierías Informática y de Telecomunicación, University of Granada, Spain*

ARTICLE INFO

Article history:
Received 5 February 2015
Revised 18 June 2015
Accepted 29 July 2015
Available online 22 August 2015

Keywords:
k-Nearest Neighbor
Self-training
Positive unlabeled learning
Time series classification
Transductive learning

ABSTRACT

Positive unlabeled time series classification has become an important area during the last decade, as often vast amounts of unlabeled time series data are available but obtaining the corresponding labels is difficult. In this situation, positive unlabeled learning is a suitable option to mitigate the lack of labeled examples. In particular, self-training is a widely used technique due to its simplicity and adaptability. Within this technique, the stopping criterion, i.e., the decision of when to stop labeling, is a critical part, especially in the positive unlabeled context. We propose a self-training method that follows the positive unlabeled approach for time series classification and a family of parameter-free stopping criteria for this method. Our proposal uses a graphical analysis, applied to the minimum distances obtained by the *k*-Nearest Neighbor as the base learner, to estimate the class boundary. The proposed method is evaluated in an experimental study involving various time series classification datasets. The results show that our method outperforms the transductive results obtained by previous models.

1. Introduction

In this paper we tackle positive unlabeled learning (PU) [12] problems in the time series classification context. PU is a suitable approach when we dispose of a limited number of examples from a given class (the positive class) and a great amount of unlabeled data. In the time series field, unlabeled data are often easy to obtain [6,19,32] but obtaining the labels may require a lot of time and attention of a skilled domain expert. In such a situation, PU becomes a proper solution for binary classification problems.

The PU learning takes advantage from both labeled and unlabeled examples. It can be considered a particular case of semi-supervised learning (SSL) [9,38], in which there are only labeled examples from a determined class. Thus, the PU learning becomes an extension of one-class classification [36] to the semi-supervised context. Despite the close relation between PU and SSL, classical methods of the latter approach cannot be directly applied to PU learning [7].

* Corresponding author. Tel.: +61 3 990 59555.
E-mail address: christoph.bergmeir@monash.edu (C. Bergmeir).

http://dx.doi.org/10.1016/j.ins.2015.07.061
0020-0255/

In the specialized literature, different PU learning approaches have been proposed for time series classification. Two main approaches are used. The first one is based on the use of clustering techniques [30,31]. The second one is focused on traditional supervised learning that is the most widely used approach in time series classification [3,10,33,42]. Most of these solutions adapt the self-training technique [44] that was initially proposed for standard SSL. It is a wrapper method that iteratively classifies the unlabeled examples, assuming that its more accurate predictions tend to be correct. As the base learner, it is common to focus on the k-Nearest Neighbor (k-NN) [1] because it has shown to be particularly effective for time series classification tasks [35]. As dissimilarity measures for time series in the self-training context, we can find the Euclidean distance, Dynamic Time Warping (DTW) [34] and DTW-D [10].

Due to its iterative nature, a critical part of the self-training technique is to decide when the learning should be stopped. The aim of this mechanism is to avoid enlarging the positive set by unlabeled instances with a low confidence level. This is known in the literature as the stopping criterion problem. In comparison to the standard SSL context, this issue is even more important here, because the presence of only positive labeled examples facilitates the erroneous inclusion of negative examples during the iterative process.

The pioneer work of Wei and Keogh [42] (in what follows denoted as WK) proposes a simple heuristic to stop the iterative learning based on the minimal distance decrease in the k-NN method. Frequently, this solution causes an incomplete learning of the trained classifier. An improvement of this work is addressed in Ratanamahatana and Wanichsan [33] (in what follows denoted as RW). This criterion takes advantage of the significative changes in the minimal k-NN distance between actual and previous iterations. A recent contribution [3,37] to the stopping criterion problem tries to learn the intrinsic structure of the data using Minimum Description Length (MDL) [20]. The MDL concept has been widely applied in other domains and the work of Begum et al. [3] (in what follows denoted as BHRK) uses it to develop a novel stopping criterion. The stopping criteria proposed for the self-training technique will be analyzed in more detail in Section 2.

In this paper we delve into the stopping criterion problem in the self-training context. To do so, we base our proposal on the k-NN method and the time series measure DTW. We define different stopping criteria based on the minimum distances achieved by the k-NN in each iteration. The principal novelty of this work is the use of a specialized graphic analysis technique [13] to identify the boundary between classes. The resulting procedure is parameter-free and without additional computational efforts to obtain the stopping point.

To evaluate the performance of our proposal, we conduct experiments involving various UCR [23] classification datasets, starting the learning with only one positive labeled instance. We test the stopping criteria with different parameters of the measure selected to compute the dissimilarity between time series. The experimental study includes a statistical analysis based on non-parametric statistical tests [18]. Finally, we select the most competitive stopping criterion proposed.

The rest of this paper is organized as follows. In Section 2, we provide definitions and the notation of PU learning. In Section 3, we define the stopping criteria proposed and some theoretical considerations. In Section 4, we include the experiment design to test the criteria proposed. In Section 5, we present the results obtained in the experiments. In Section 6, we offer some conclusions.

2. Background and preliminaries

In this section we define the PU learning problem and the principal proposals in this topic (Section 2.1). Then, we review the self-training technique for PU time series classification (Section 2.2) and address the existing stopping criteria for this technique (Section 2.3).

2.1. Positive unlabeled learning

This section presents the definition and notation for SSL, which is applicable to the PU learning approach. In semi-supervised classification, the dataset can be divided into two parts, L and U. Let L be the set of instances $X_l = (x_1, \ldots, x_l)$ for which the labels $Y_l = (y_1, \ldots, y_l)$ are provided. Let U be the set of instances $X_u = (x_{l+1}, \ldots, x_{l+u})$ for which the labels are not known.

In particular, for PU learning the set L contains only positive instances, and is hence also denoted as P. SSL will be most useful whenever there are far more unlabeled data than labeled, therefore, it is a classical assumption for this learning scheme [9].

There are various domains where the PU learning becomes an appropriate option, for example in electronic text classification [28] and human genes identification associated with diseases [8]. The key feature of PU classification is that there are no negative labeled examples, which renders traditional supervised and semi-supervised techniques not applicable directly. Two main approaches have been followed in the past to deal with this situation.

The first approach is based on adapting the supervised traditional techniques to the PU classification [7,27,29,46]. In the work of Zhang and Zuo [46] the classic k-NN method is used to rank the unlabeled examples according to their similarity with respect to the positive training instances. Next, the r first ranked instances are labeled as positive examples. Unfortunately, there hasn't been proposed a method to approximate the value of r.

The latter approach is focused on adapting the standard SSL techniques [11,45]. The work of Denis et al. [11] is based on the classic co-training method, whereas the proposal of Yu and Li [45] uses a graph-based SSL technique.