Neurocomputing 136 (2014) 30–40

Contents lists available at ScienceDirect

Neurocomputing

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

Improve the performance of co-training by committee with refinement of class probability estimations

Shuang Wang*, Linsheng Wu, Licheng Jiao, Hongying Liu

Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education of China, Xidian University, Xi'an 710071, China

ARTICLE INFO

Article history:
Received 17 June 2013
Received in revised form
3 January 2014
Accepted 15 January 2014
Communicated by M. Wang
Available online 17 February 2014

Keywords:
Co-training
Semi-supervised learning
Ensemble learning
Class probability
Distance metric
Data editing

ABSTRACT

Semi-supervised learning is a popular machine learning technique where only a small number of labeled examples are available and a large pool of unlabeled examples can be obtained easily. In co-training by committee, a paradigm of semi-supervised learning, it is necessary to pick out a fixed number of most confident examples according to the ranking of class probability values at each iteration. Unfortunately, the class probability values may repeat, which results in the problem that some unlabeled instances share the same probability and will be picked out randomly. This brings a negative effect on the improvement of the performance of classifiers. In this paper, we propose a simple method to deal with this problem under the intuition that different probabilities are crucial. The distance metric between unlabeled instances and labeled instances can be combined with the probabilities of class membership of committee. Two distance metrics are considered to assign each unlabeled example a unique probability value. In order to prove that our method can get higher-quality examples and reduce the introduction of noise, a data editing technique is used to compare with our method. Experimental results verify the effectiveness of our method and the data editing technique, and also confirm that the method for the first distance metric is generally better than the data editing technique.

1. Introduction

In many practical data mining applications such as web page classification [1] and content-based image retrieval [2], unlabeled instances have become abundant, but to get their labels is time consuming and expensive. Semi-supervised learning [3,4] has been developed to make full use of the large number of unlabeled samples as well as the labeled ones to improve the performance of supervised learning.

Self-training [5] is an iterative and single-view semi-supervised learning algorithm. The process is simple: first, a single classifier is trained using the original labeled instances. Second, the classifier classifies the unlabeled instances and gets their class probability estimations. Then the examples are ranked according to the class probability values, and a fixed number of most confident examples are added permanently into the training set. After that, the classifier is retrained with the updated training set. The process is repeated until no more unlabeled examples are available or some stopping conditions are met.

Co-training [6] is an iterative and multi-view semi-supervised learning algorithm. Two classifiers are trained separately on two views, i.e. two independent sets of attributes. Then each classifier classifies the unlabeled instances and picks out the most confident examples for the other one. The two classifiers are retrained with the augmented training set. The process is repeated until no more unlabeled examples are available. The algorithm can work well if two conditions are met, one is that the data set has two sufficient and redundant feature subsets and the other one is that the two subsets are sufficient for learning and independent given the class label. However, the two conditions are too strong for many real data sets. In [5], it is shown that the performance of co-training is sensitive to the two requirements. Co-EM algorithm that combines EM algorithm [7] with co-training is also introduced.

In recent years, the graph-based multi-view learning has been proposed for semi-supervised learning [8–13]. In [8], various crucial factors of video annotation can be represented by different graphs. Multiple graphs are integrated into a regularization framework. Then semi-supervised learning is conducted on the fused graph. In [9], based on graph-based semi-supervised learning, the presented method learns the multi-view distance metrics from three visual feature sets and the labels of unlabeled cartoon characters simultaneously. The method proposed in [10] encodes different features from different views in a subspace. The information from the labeled cartoon characters is used to construct local patches where the manifold structure revealed by unlabeled cartoon characters is employed to capture the geometry information. In [11], a web image search reranking method is introduced to explore multiple

* Correspondence to: Mailbox 224, Xidian University, No. 2 South Taibai Road, Xi'an, 710071, China. Tel.: +86 29 88202661.
E-mail addresses: shwang.xd@gmail.com (S. Wang), 867208786@qq.com (L. Wu).

http://dx.doi.org/10.1016/j.neucom.2014.01.039
0925-2312

modalities in a graph-based learning scheme. In [12], based on patch alignment framework, a novel semi-supervised dimensionality reduction method for multi-view data is proposed. The method introduced in [13] deals with the problem of multi-view dimensionality reduction by learning a unified low-dimensional subspace to fuse the multi-view features. Both intra-class and inter-class geometries are taken into consideration. In addition, many other works on semi-supervised learning have been proposed in recent years [14–17]. The method presented in [14] constructs multiple hypergraphs for a set of 3-D objects based on their 2-D views. Retrieval and recognition are performed based on the hypergraphs. This method does not use the distance between objects. In [15], the proposed method simultaneously utilizes both visual and textual information to estimate the relevance of user tagged images. The relevance estimation is determined with a hypergraph learning approach. In [16], a discriminative probabilistic object modeling approach is proposed to estimate the distance between two objects. 3D object retrieval and recognition are accomplished based on the distance measures. In [17], a 3D object retrieval scheme with Hausdorff distance learning is proposed. Relevance feedback information is employed to select positive and negative view pairs, and a view-level Mahalanobis distance metric is learned. This Mahalanobis distance metric is adopted to estimate the Hausdorff distances between objects.

To relax the hard requirements of co-training, a lot of works [18–23], which will be introduced in detail in the following section, have been proposed. Among them, co-training by committee [23], which is similar to self-training [5], is also an iterative and single-view semi-supervised learning algorithm. It employs a set of diverse classifiers together to learn from labeled and unlabeled examples. The learning process is the same as that in self-training. In the work [24], different Bayesian classifiers are respectively used in the process of self-training, but the experimental results on UCI data sets [25] show that the performance of classifier is generally not better than that of classifier learned only from the labeled data. From this wok, we can see that a single classifier sometimes may not operate well, so we can construct a set of diverse classifiers by ensemble as that in the algorithm of co-training by committee, the experimental results in [23] verify the effectiveness of this algorithm.

The motivation of our paper is as follows. In co-training by committee [23], it is necessary to pick out a fixed number of most confident examples according to the ranking of class probability values at each iteration. However, the class probability values may repeat, which results in the problem that some unlabeled instances share the same probability and will be picked out randomly. This brings a negative effect on the improvement of the performance of classifiers. Therefore it motivates us to obtain a more accurate ranking of class probability values. Some measures can be taken to disturb the class probability estimation values of ensemble learner, thus a more accurate ranking is obtained and also a better classification performance can be expected.

The main contributions of this paper can be summarized as follows. First, a simple method is proposed to adjust the class probability values predicted by committee. We combine the probabilities of class membership of committee with the distance metric between unlabeled instances and labeled instances. Two distance metrics are considered, specifically, the mean squared Euclidean distance between unlabeled instance and labeled instances is used as the first distance metric and the distance between unlabeled instance and clustering center of labeled instances is employed as the second distance metric. Once two examples have the same class probability value, the one that owns smaller distance metric should have the larger chance to be picked out. Second, to prove that our method can get higher-quality examples and reduce the introduction of noise, a data editing technique is used to compare with our method. Third, experimental results from different aspects are obtained on 12 UCI data sets [25]. The results verify the effectiveness of our method and the data editing technique, and also confirm that the method for the first distance metric is generally better than the data editing technique.

The rest of this paper is organized as follows. Related work is described in Section 2. Section 3 presents our method and the data editing technique. Experimental results are reported in Section 4. Finally, conclusions are drawn in Section 5.

2. Related work

2.1. Co-training with a single view

Many algorithms have been proposed in semi-supervised learning [3,4], but in the following we mainly introduce the single-view co-training algorithms.

A single-view semi-supervised learning method called statistical co-learning was presented by Goldman and Zhou [18]. The method does not need two sufficient and redundant views, but the two different supervised learning algorithms used are required to partition the example space into a set of equivalence classes. For instance, each leaf of a decision tree defines an equivalence class. At each iteration 10-fold cross validation is used to select the most confident examples to label. When making the final prediction by combining the two hypotheses, 10-fold cross validation is also employed. However, it has some drawbacks: first, cross validation technique is very time consuming, so the efficiency of this algorithm is not high. Second, when the original labeled sample set is very small, cross validation technique is not reliable. Finally, it requires special learning algorithms, which limits its applicability. Zhou and Goldman [19] extended their previous work [18] and introduced another single-view method, which employs three or more classifiers to label data for each other. It reduces the need of statistical tests compared with statistical co-learning, but cross validation technique is still required to measure confidence intervals, which are exploited to select the most confident unlabeled instances and to combine the final hypotheses.

Recently, DCPE co-training was introduced by Xu et al. [20]. Two different learning algorithms instead of two sufficient and redundant views are employed to produce diversity during the learning process. First, two learners are trained using the original labeled instances. Then the learners independently predict the unlabeled instances and get the class probability estimations of them. For one learner, if unlabeled examples have the same prediction labels and the highest class probability estimation differences between the learner and the other one, then the corresponding unlabeled examples are added into the training set of the learner. Finally, the two learners are separately retrained with the augmented training set. The process is repeated until no more unlabeled examples are available or some stopping conditions are met.

Zhou and Li [21] proposed tri-training algorithm, which requires neither sufficient and redundant views nor special supervised learning algorithms that could partition the example space into a set of equivalence classes. For the first step, three initial learners are trained from data sets generated via bootstrap sampling [27] from the original labeled data set, which is just similar to construct an ensemble by Bagging [28]. Then learners are refined using unlabeled examples in the training process, an unlabeled example is added temporarily to the training set of a learner if the other two learners agree their labels under certain conditions, the learners are updated on their augmented data sets. The final prediction is produced by majority voting. Although the algorithm is more applicable than previous co-training style algorithms, better performance can be expected by employing more