Information Sciences 349–350 (2016) 216–228

Contents lists available at ScienceDirect

Information Sciences

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

Semi-supervised classification method through oversampling and common hidden space

Aimei Dong[a,c], Fu-lai Chung[b], Shitong Wang[a,b,*]

[a] School of Digital Media, Jiangnan University, Wuxi, JiangSu, China
[b] Department of Computing, Hong Kong Polytechnic University, Hong Kong, China
[c] School of Information, Qilu University of Technology, Jinan, ShanDong, China

ARTICLE INFO

Article history:
Received 17 July 2015
Revised 15 February 2016
Accepted 21 February 2016
Available online 27 February 2016

Keywords :
Semi-supervised classification
Oversampling
Common hidden space
Dimensionality augmentation

ABSTRACT

Semi-supervised classification methods attempt to improve classification performance based on a small amount of labeled data through full use of abundant unlabeled data. Although existing semi-supervised classification methods have exhibited promising results in many applications, they still have drawbacks, including performance degeneration, due to the introduction of unlabeled data and partially false labels in a small amount of labeled data. To circumvent such drawbacks, a new semi-supervised classification method OCHS-SSC through oversampling and a common hidden space is proposed in the paper. The primary characteristics of the proposed method include two aspects. One is that unlabeled data are only used to generate new synthetic data to extend the minimal amount of labeled data. The other is that the final classifier is learned in the extended feature space, which is composed of the original feature space and the common hidden space found between labeled data and the synthetic data instead of the original feature space. Extensive experiments on 23 datasets indicate the effectiveness of the proposed method.

1. Introduction

Pattern classification is a research technique in machine learning and has been applied to many related fields, such as web-page classification and spam filtering. There are an increasing number of applications [4,12,19,22,23,28,29] with a great deal of unlabeled training data and a small amount of labeled data. Abundant unlabeled data are quite easy to acquire, whereas a small amount of labeled data is usually costly and difficult to acquire. To solve these problems, semi-supervised learning [5,24,36,38] is proposed to learn an effective pattern classifier from a small amount of labeled data with the aid of abundant unlabeled data.

To date, many semi-supervised classification methods have been developed using different approaches. The most distinguished achievements among all of the semi-supervised classification methods include (1) semi-supervised support vector machines [15], which maximize the margin of labeled and unlabeled data through adaption of the hyperplane of SVM and the labels of unlabeled data; (2) generative methods [20–21], in which all of the training data (labeled and/or unlabeled) are assumed to be generated from the same generation model and its parameters. As the bridge linking labeled and unlabeled data, the generation model and its parameters are estimated through maximizing the corresponding likelihood with EM-based algorithms. It is important but difficult for such methods to find an appropriate model. If the training data are

* Corresponding author at: School of Digital Media, Jiangnan University, Wuxi, JiangSu, China. Tel.: +86 13182791468.
E-mail address: wxwangst@aliyun.com (S. Wang).

http://dx.doi.org/10.1016/j.ins.2016.02.042
0020-0255/

not subject to the hypothetical model, unlabeled data will become harmful for the final classifier; (3) graph-based methods [3,18,33,37], in which the training data are mapped into a graph embodying the relationship among the training data. The label information is then propagated based on the graph. If the constructed graph is not consistent with the inherent law of the training data, the introduction of abundant unlabeled data will be dangerous for the final classifier; and (4) self-labeled [26–27] methods, which accept that their own predictions tend to be correct in an iterative process through different mechanisms such as self-training [16,32], tri-training [34] or disagreement-based models [35]. If unlabeled data are falsely predicted during the iterative self-labeling process, harmful results will perhaps occur. Thus, how to control false predictions for unlabeled data is critical for such methods.

All of these methods have the same goal of taking full advantage of unlabeled data. Although semi-supervised classification methods have shown promising performance in many applications, several scholars [2,25] have found unavoidable disadvantages immanent to these methods. In other words, the performance of the final classifier might degrade due to the introduction of unlabeled data. This discovery inspires us to consider how to use unlabeled data carefully. On this topic, Zhou [17] and Chen [30] proceeded by developing the S4VM and SA-SSCCM methods, respectively. Both authors advance safe strategies for using unlabeled data in unique ways and obtained outstanding performance. The common idea of the two approaches is to use labeled and unlabeled data directly. When labeled data are partially falsely labeled and/or unlabeled data contain outliers, the performance of these methods might degrade more seriously.

This paper focuses on solving semi-supervised classification problems with the characteristic of partially false labels of labeled data and the presence of outliers among unlabeled data. The overall approach to addressing these problems contains two aspects. One is not to use unlabeled data directly but rather to generate new synthetic data with the help of unlabeled data. The other is not to learn the final classifier in the original feature space but rather in the extended feature space. The latter is composed of the original feature space and the augmented feature space, which is shared by both the small amount of labeled data and the synthetic data. To this end, a new semi-supervised classification method OCHS-SSC through oversampling and the common hidden space is proposed in this paper. Extensive experiments on six benchmark datasets and seventeen UCI datasets have confirmed the efficiency of the proposed method.

The remainder of this paper is organized as follows. The preliminaries of the proposed method are introduced in Section 2. Section 3 describes the proposed method. Extensive experimental results are provided in Section 4 to validate the effectiveness of the proposed method. In the last section, the conclusions are provided, and future works are discussed.

2. Preliminaries of the proposed method

2.1. Definitions of semi-supervised classification

In a semi-supervised classification scenario, all of the data can be split into the labeled part and the unlabeled part. Let the labeled dataset be $L = \{\boldsymbol{x}_i, y_i\}_{i=1}^{n_l}$ and the unlabeled dataset be $U = \{\boldsymbol{x}_j\}_{j=1}^{n_u}$, in which $\boldsymbol{x}_i \in \mathbb{R}^{\mathrm{d}}$ and $\boldsymbol{x}_j \in \mathbb{R}^{\mathrm{d}}$, $y_i \in \{+1, -1\}$, and $n_u \gg n_l$. The goal of semi-supervised classification is to learn a better classifier for future unseen data (inductive semi-supervised learning) or for the unlabeled data (transductive semi-supervised learning) [7]. In this study, we perform experiments from the perspective of both inductive semi-supervised learning and transductive semi-supervised learning.

2.2. Motivation for using oversampling technology

In traditional semi-supervised classification methods, unlabeled data are self-labeled by using certain mechanisms such as manifold assumption and graph-based models to expand the labeled dataset. In a self-labeling process, there might be wrong unlabeled data included in the labeled dataset. The condition occurs due to the following:

(1) There might be noise in the unlabeled dataset. The existence of noise in the unlabeled dataset can easily cause the wrong unlabeled examples to be added to the labeled dataset.
(2) The size of the labeled dataset is so small that the labeled dataset cannot reflect its genuine data distribution. The shortage of labeled data will cause wrong unlabeled examples to be added to the labeled dataset. This problem is more obvious when some labeled data are very close to the decision boundary.

The first reason shows that it is better not to use the unlabeled dataset directly in semi-supervised classification; the second reason shows that it is necessary to expand the labeled dataset. Considering these points, the usage of oversampling technology for the labeled dataset based on the labeled dataset and the unlabeled dataset is a good choice in semi-supervised classification.

2.3. Motivation for using the common hidden space

In traditional semi-supervised classification methods, the unlabeled dataset is labeled in a repeated process until a certain criterion is satisfied with the direct usage of the labeled dataset. In the repeated labeling process, if labeled data are attacked by certain external factors and some of the labels of the labeled dataset are consequently wrong, then more wrong unlabeled data are included into the training set. To overcome this disadvantage, a common hidden space between the labeled dataset and the generated synthetic dataset is found such that the proposed method in this study is performed