Information Sciences 294 (2015) 390–407

Contents lists available at ScienceDirect

Information Sciences

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

Support vector machine with manifold regularization and partially labeling privacy protection

Tongguang Ni [a,b], Fu-Lai Chung [c], Shitong Wang [a,c,*]

[a] School of Digital Media, Jiangnan University, Wuxi, Jiangsu, PR China
[b] School of Information Science and Engineering, Changzhou University, Changzhou, Jiangsu, PR China
[c] Department of Computing, Hong Kong Polytechnic University, Hong Kong

ARTICLE INFO

Article history:
Received 16 January 2014
Received in revised form 22 September 2014
Accepted 28 September 2014
Available online 7 October 2014

Keywords:
Large datasets
Classification
Support vector machine
Privacy protection
Manifold regularization

ABSTRACT

A novel support vector machine with manifold regularization and partially labeling privacy protection, termed as SVM-MR&PLPP, is proposed for semi-supervised learning (SSL) scenarios where only few labeled data and the class proportion of unlabeled data, due to privacy protection concerns, are available. It integrates manifold regularization and privacy protection regularization into the Laplacian support vector machine (LapSVM) to improve the classification accuracy. Privacy protection here refers to use only the class proportion of data. In order to circumvent the high computational burden of the matrix inversion operation involved in SVM-MR&PLPP, its scalable version called SSVM-MR&PLPP is further developed by introducing intermediate decision variables into the original regularization framework so that the computational burden of the corresponding transformed kernel in SSVM-MR&PLPP can be greatly reduced, making it highly scalable to large datasets. The experimental results on numerous datasets show the effectiveness of the proposed classifiers.

1. Introduction

A learning problem that has only recently gained attention in the machine learning community is that of learning a classifier from class labeling proportion information [21–23,28]. This type of learning problem appears in areas like politics, medicine, spam filtering and so on. For example in political election, the result of each vote is not open to everyone; but for each region, the vote distribution is publicly available and presents the class labeling proportion of each candidate in the region. The final vote results are closely related to every voter with his income, family class, and so on. As we know, the above elements will directly influence the distribution of votes in a region. Likewise, a similar situation also exists in medical diagnosis. For example, following the outbreak patterns of a new type of influenza virus is an important task, but revealing which patient actually got infected should be treated in a highly confidential manner. However, outbreak frequencies in certain risk groups are not sensitive information. There exist similar problems in the area of spam identification. The cost of data collection is also very important for real applications. As we know, the datasets of spam emails are likely to contain almost pure spam data (which is achieved, e.g., by listing e-mails as spam bait), while user's inboxes typically contain a mixture of spam and non-spam emails. The usual task is to use the inbox data to improve the estimation of spam.

[*] Corresponding author at: School of Digital Media, Jiangnan University, Wuxi, Jiangsu, PR China. Tel.: +86 510 85912136.
E-mail address: shwangst@aliyun.com (S. Wang).

http://dx.doi.org/10.1016/j.ins.2014.09.050
0020-0255/© 2014 Elsevier Inc. All rights reserved.

In many cases, it is possible to estimate the proportions of spam and non-spam in a user's inbox, which is much cheaper than estimating the actual labels. More importantly, collecting e-mail labels from a user's inbox may involve an invasion of personal privacy. A lot of individual information can be released in the form of label proportions in real life. As for the above examples, after election, the proportions of votes of each demographic area may be released by the government. In healthcare, the proportions of diagnosed diseases of each zip code area may be available to the public. Motivated by the above real-world applications, several classifiers using the class labeling proportion information have been proposed in [21,23,28]. However, since all these classifiers do not consider the intrinsic structure hidden between labeled and unlabeled samples, they become inappropriate for the semi-supervised application scenarios where a few labeled samples can be quite expensively acquired while huge amounts of unlabeled samples can be easily and/or cheaply collected and the class labeling proportion information is also available.

As we may know well, semi-supervised learning (SSL), which exploits the huge amounts of unlabeled data jointly with the limited labeled data for learning, has attracted considerable attention in recent years [1,4,12,14,15,18,36,39]. A popular SSL approach [2] is learning with the manifold assumption which states that each class lies on a separate low-dimensional manifold embedded in a higher dimensional space. Many forms of real-world data, such as handwritten digits [10], webpages [40], images [17] and so on, have been demonstrated to exhibit such an intrinsic geometric structure, which can be commonly approximated by a weighted graph. In this study, we try to integrate such a SSL learning mechanism with class labeling proportion to develop a novel support vector machine with manifold regularization (MR) and partially labeling privacy protection (PLPP). The difference between the proposed classifiers and the often-used semi-supervised learning methods is illustrated in Fig. 1.

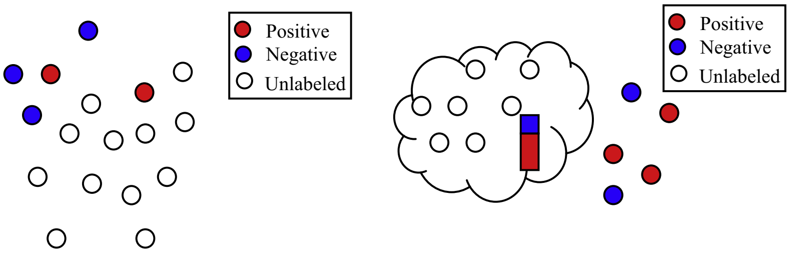
In order to achieve our goal, we first use the Laplacian support vector machine (LapSVM) [2] as the basic framework to construct a novel learning framework with manifold regularization and partially labeling privacy protection, which makes good use of the labeling proportion of unlabeled data, e.g., the proportion of patients having the diseases. Such proportion information can be helpful for classification and at the same time protects the privacy information. Based on the principle of support vector machine (SVM) [16,25,29,32–34,38] by using an ε-insensitive loss function, a support vector machine with manifold regularization and partially labeling privacy protection, termed as SVM-MR&PLPP, is then proposed. The learning task can be solved using the classical quadratic programming (QP) solver. Moreover, the traditional SSL approaches based on manifold regularization framework are limited to small scale datasets and hence inappropriate for large data as a result of the matrix inversion operation involved in the dual problem [2]. As the computation of matrix inversion, likewise in LapSVM, is very expensive for a large dataset [6,13,20,35], the objective function of the proposed classifier, i.e., SVM-MR&PLPP, is reconstructed by introducing intermediate decision variables in the manifold regularization framework, and hence a scalable version of SVM-MR&PLPP, called SSVM-MR&PLPP, is developed accordingly. Since the proposed classifiers consider both the labeled and unlabeled data as well as the labeling proportion of the unlabeled data in a learning task, they not only inherit the advantages of manifold learning, but also can effectively correct its decision boundary with the given labeling proportion.

The main contributions of this work can be highlighted below.

(1) A novel classifier SVM-MR&PLPP which considers the labeled data, the unlabeled data and the labeling proportion of the unlabeled data is proposed. We also prove that the training of SVM-MR&PLPP can be equivalently transformed as a classical QP problem.

(2) By introducing intermediate variables into the learning framework, the proposed classifier SVM-MR&PLPP is extended into its scalable version called SSVM-MR&PLPP for large datasets. We show that the training of SSVM-MR&PLPP can also be transformed as a classical QP problem and consequently SSVM-MR&PLPP can be efficiently solved by a QP solver. SSVM-MR&PLPP inherits the same sparsity as in SVM, though it has a kernel matrix different from SVM-MR&PLPP.

(3) Extensive experiments on synthetic and real-world datasets demonstrate that the proposed classifiers outperform or are at least comparable to several state-of-the-art methods.



(a) semi-supervised learning

(b) semi-supervised learning with proportion

Fig. 1. Difference between the proposed classifiers and the often-used SSL (colors encoding class labels): (a) semi-supervised learning with labeled and unlabeled data explicitly given; (b) semi-supervised learning with labeling proportion where labeled data, unlabeled data and a proportion of unlabeled data are given.