Pattern Recognition 48 (2015) 1724–1733

Contents lists available at ScienceDirect

Pattern Recognition

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

Penalized partial least square discriminant analysis with ℓ_1−norm for multi-label data

Huawen Liu [a,b], Zongjie Ma [a], Shichao Zhang [c], Xindong Wu [d,e,*]

[a] Department of Computer Science, Zhejiang Normal University, China
[b] NCMIS, Academy of Mathematics and Systems Science, CAS, China
[c] Department of Computer Science, Zhejiang Gongshang University, China
[d] School of Computer Science and Information Engineering, Hefei University of Technology, China
[e] Department of Computer Science, University of Vermont, USA

ARTICLE INFO

Article history:
Received 11 February 2014
Received in revised form
25 October 2014
Accepted 10 November 2014
Available online 18 November 2014

Keywords:
Partial least squares
Discriminant analysis
Dimension reduction
Sparse learning
Multi-label learning

ABSTRACT

Multi-label data are prevalent in real world. Due to its great potential applications, multi-label learning has now been receiving more and more attention from many fields. However, how to effectively exploit the correlations of variables and labels, and tackle the high-dimensional problems of data are two major challenging issues for multi-label learning. In this paper we make an attempt to cope with these two problems by proposing an effective multi-label learning algorithm. Specifically, we make use of the technique of partial least square discriminant analysis to identify a common latent space between the variable space and the label space of multi-label data. Moreover, considering the label space of the multi-label data is sparse, a l_1-norm penalty is further performed to constrain the Y-loadings of the optimization problem of partial least squares, making them sparse. The merit of our method is that it can capture the correlations and perform dimension reduction at the same time. The experimental results conducted on eleven public data sets show that our method is promising and superior to the state-of-the-art multi-label classifiers in most cases.

1. Introduction

Multi-label learning is a typical classification application of supervised learning. Unlike traditional supervised learning techniques involving class labels exclusively, multi-label learning mainly concerns data associated with more than one class label simultaneously. This kind of data is known as multi-label data, which are prevalent in many real-world applications [1]. For example, the movie *Avatar* can be tagged with *action*, *horror* and *science fiction* types; a piece of *financial storm* news may be linked with *market*, *economics* and *politics*; an image portraying *Pyramids* may be involved with *pyramid*, *Egypt*, *architecture* and *Africa*; a gene in bioinformatics may be associated with a number of functional classes, such as *metabolism* and *protein synthesis*.

Since multi-label learning has a great number of potential applications, it has now been receiving more and more attentions from various fields [2]. Many multi-label learning algorithms, such as IBLR_ML [3], RA*k*EL [4] and Rank-CVM [5], have been witnessed during the past decades, and widely applied in many domains, including text categorization [6], image and video annotation [7], content annotation [8], music processing [9], bioinformatics [10], and so on. Generally they can be grouped into two categories, i.e., algorithm adaption and problem transformation [1,11]. The former extends traditional classifiers (e.g., *k*NN and SVM) to cope with multi-label data by exerting some proper constraint conditions, while the latter technique transforms multi-label data into corresponding single-label ones. Typical examples include BR*k*NN and LP*k*NN [4].

How to effectively capture the correlations of the labels of data is still an open issue in multi-label learning. As we know, the output of the multi-label learning models is a set of class labels, no longer a single label, for a given data at a time. Indeed, the labels are often relevant to each other in practice. It requires that the correlations of the labels should be considered when building a multi-label learning model [12]. The existing learning algorithms either treat the class labels independently, or take each set of the labels occurred in training data as a new label. However, they cannot work effectively especially when there are a large number of class labels in the data, albeit the correlations of the labels can be captured to some extent.

Another challenging issue often encountered in multi-label learning is the high dimensionality of data collected from large-scale applications. As the dimensionality is getting larger, it gives rise to problems like over-fitting, multi-collinearity and "curse of dimensionality" [13,14,16]. Dimensionality reduction is an effective technique to address the high-dimensional problems [15]. For

* Corresponding author.
E-mail addresses: huaw.liu@gmail.com (H. Liu), zhangsc@gxnu.edu.cn (S. Zhang), xwu@cs.uvm.edu (X. Wu).

http://dx.doi.org/10.1016/j.patcog.2014.11.007

example, Zhang and Zhou [16] projected the original data into a lower-dimensional space with the Hilbert–Schmidt Independence Criterion, while Ji et al. [17] extracted a common subspace shared among multiple labels by virtue of ridge regression. Liu et al. [18] made use of logistic regression to construct a multi-label classifier. However, one limitation of the above methods is that their outputs are weighted combinations of the original spaces by linear transformation, making the interpretation of results difficult.

With these motivations, in this paper we address the above problems and propose a new multi-label learning algorithm called PPLS-MD (Penalized Partial Least Squares for Multi-label Data) according to the inherent properties of multi-label data. Specifically, we adopt the technique of partial least square to explore the correlations of variables and labels and perform dimension reduction for multi-label data. Furthermore, a constraint condition of the l_1-norm penalty is imposed on the optimization problem of partial least square discriminant analysis. Imposing the constraint is consistent with the sparse character of the labels for multi-label data, that is, the class labels are often sparse.

In a nutshell, the key contributions of this paper are highlighted as follows:

- We propose a multi-label learning framework for multi-label data, which can not only capture the correlations of variables and labels, but also perform the operation of dimension reduction simultaneously.
- We reveal the correlations by using the multivariate analysis method of partial least square discriminant analysis, which is an effective technique to analyze the relationships between two sets of variables.
- We impose the l_1-norm penalty on the discriminant model to regularize the Y-loadings, yielding a sparse model. This is consistent with the sparse property of the label space of multi-label data.

The rest of this paper is organized as follows. Section 2 briefly reviews the state-of-the-art on multi-label learning. We formulate the multi-label learning problem and give the basic concept of partial least squares in Section 3. Section 4 presents the proposed learning framework of PPLS-MD for multi-label data and discusses the relationship to several popular learning methods. The experimental settings are provided in Section 5, followed by the experimental results of our method with the comparing algorithms on eleven data sets in Section 6. Section 7 concludes the paper.

2. Related work

In this section, we briefly review the most related and representative multi-label learning methods. Please refer to good surveys (see, e.g., [1,11,19]) and references therein to get more details of multi-label learning algorithms.

As mentioned in Section 1, multi-label learning algorithms fall into two major categories: problem transformation and algorithm adaptation. The characteristic of the former one is fitting data to algorithms. The central idea is to partition each multi-label observation (i.e., sample) into several corresponding single-label ones, on which a classification model can be constructed by using ensemble strategies. Binary relevance (BR) [11] and DBR [20] are representative examples of this kind. However, BR does not involve the correlations of class labels during the learning stage.

To alleviate this problem, Label Powerset (LP) constructs classification models by concerning pairwise or subset correlations. For example, calibrated label ranking (CLR) [21] pays attention on the pairwise correlations of the labels, while RAndom *k*-labELset (RA*k*EL) [4] takes *k* labels as a whole at a time. Although the LP methods take the correlations into account to some extent, they have relatively high complexities and cannot work well especially when the number of the class labels is large [22].

The second kind of multi-label learning methods is algorithm adaptation. It extends traditional learning algorithms, such as C4.5, *k*NN and SVM, to solve the multi-label problems. BR*k*NN [4] and ML*k*NN [23] are representative examples of this kind, and both of them adopt *k*NN to predict the output results. IBLR_ML [3] integrates *k*NN and logistic regression to construct a classification model for multi-label data, while Rank-CVMz [24] makes use of a core vector machine to train a multi-label classifier. Usually the adaptation learning methods have not involved the correlations of the class labels.

How to effectively tackle high-dimensional data is still an open issue in multi-label learning. The existing multi-label learning algorithms discussed above place less attention on the high-dimensional problems, which are prevalent in real-world applications, such as computer vision, information retrieval and bioinformatics. Dimensionality reduction, seeking a succinct and low-dimensional variable space to preserve intrinsic characteristics of the original high-dimensional data, is an effective technique to cope with the high-dimensional data in machine learning.

Recently several works have also discussed the high-dimensional problems and applied the techniques of dimension reduction to high-dimensional multi-label data. For example, Hsu et al. [25] encoded and decoded the label space using compressive sensing. Bi and Kwok [26] represented the labels as a tree or directed acyclic graph, while Tai and Lin [27] transformed the label space into a small linear space by mapping all possible label sets to vertices of a hypercube. Ji et al. [17] extracted a common subspace from the label and variable spaces to capture their correlations. In addition, canonical correlation analysis has also been applied to measure the correlations of the variables to the class labels in the literature [28,29]. However, the common limitation of dimension reduction is that the derived results are hard to be interpreted, because they are weighted combinations of the original space. Especially when the dimensionality of data is very high, the interpretability becomes impossible. Besides, they have not fully explored the characteristic of multi-label data.

3. Preliminary concepts

3.1. Problem formulation

Assume that $\mathcal{X}=\{X_1,X_2,\ldots,X_n\}$ represents n independent observations (or samples), where each observation $X_i \in R^p$ is a vector in a p-dimension variable space. $\mathcal{Y}=\{y_1,y_2,\ldots,y_q\}$ are q class labels associated with the observations. The data collection $\mathcal{D}=\{(X_i,Y_i)\}_{i=1}^{n}$ is called a set of multi-label data, if $Y_i \subseteq \mathcal{Y}$. Without loss of generality, the subset Y_i of the labels is often represented as a vector form like $Y_i=\{Y_{ik}\}_{k=1}^{q}$, where $Y_{ik} \in \{0,1\}$. $Y_{ik}=1$ indicates that the k-th label y_k is associated to X_i; Otherwise y_k is irrelevant to X_i.

Multi-label learning refers to the process of constructing a classification model $h:\mathcal{X}\rightarrow 2^{\mathcal{Y}}$ from the data set $\mathcal{D}$, such that it can effectively help users to understand the data easily and make right decisions in future prediction. From this definition, we know that the model h is a mapping of the variable space $\mathcal{X}$ to the label space $\mathcal{Y}$. Unlike traditional learning models, the output of a multi-label model h is a subset $Y_k \subseteq \mathcal{Y}$ of the labels. Once the model h is available, we can make use of it to determine a proper label subset for any new observation.

3.2. Partial least square discriminant analysis

Since partial least squares (PLS), a well known method for modeling the relationships between two sets of observed variables [30], was introduced, it has been gaining a lot of attention from various domains. In practice, PLS has often been used as a tool of