Neurocomputing 168 (2015) 356–364

Contents lists available at ScienceDirect

Neurocomputing

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

Learning shared subspace for multi-label dimensionality reduction via dependence maximization

Xin Shu [a,*], Darong Lai [b], Huanliang Xu [a], Liang Tao [c]

[a] College of Information Science and Technology, Nanjing Agricultural University, China
[b] School of Computer Science and Engineering, Southeast University, China
[c] Department of Computer Science, City University of Hong Kong, Hong Kong

ARTICLE INFO

Article history:
Received 1 February 2015
Received in revised form
18 April 2015
Accepted 26 May 2015
Communicated by Jiayu Zhou
Available online 4 June 2015

Keywords:
Multi-label dimensionality reduction
Dependence maximization
Shared subspace
Least squares

ABSTRACT

Multi-label Dimensionality reduction via Dependence Maximization (MDDM) has been proposed recently to cope with high-dimensional multi-label data. MDDM projects the original data onto a lower-dimensional feature space in which the dependence between the feature and the associated class labels is maximized. However, the computation of MDDM involves dense matrices eigen-decomposition that is computationally expensive for the high-dimensional data. In addition, MDDM cannot be guaranteed to capture the correlation between multiple labels, which are highly beneficial to multi-label learning. To efficiently solve MDDM, in this paper we propose a novel framework that does not require any eigen-decomposition of a matrix. Specifically, our algorithm has linear time complexity in the dimensionality of the data set. Further, we show that MDDM can be reformulated as a least-squares problem, enabling us to integrate the shared subspace that can effectively uncover multiple label interactions. Extensive experiments conducted on benchmark data collections verify the effectiveness of our proposed model.

1. Introduction

Multi-label classification has recently gained significant attention in many applications such as multi-topic document categorization [1,2], protein function prediction [3,4] and automatic image annotation [5,6]. Unlike traditional single-label classification where each instance belongs to only one class, multi-label classification deals with problems where each instance may associate with more than one class. A large number of algorithms for multi-label classification have been developed in the literature. According to [7], existing multi-label classification methods can be roughly divided into two categories: algorithm adaption and problem transformation. Algorithm adaption approaches attempt to extend existing single-label classification algorithms to handle multi-label problems. Typical examples include neural network [8,9], lazy learning [10–12], Adaboost MR [13,14], and rank SVM [15] . For the transformation approaches, one usually transforms the multi-label classification problem into several single-label classification problems so that existing single-label approaches can be easily employed. Some prominent examples include binary relevance method [7], pair-wise method [16,17] and label embedding method [18–20]. Recently, Madjarov et al. [21] extend this categorization of multi-label methods with a third group of methods, namely, ensemble methods. Algorithms belonging to this group include RAkEL [22] and ensembles of classifier chains [23].

However, multi-label classification frequently involves high-dimensional data which makes existing approaches impractical due to the curse of dimensionality. As a result, a large number of multi-label dimension reduction approaches have been developed in the literature. Multi-label informed latent semantic indexing (MLSI) was proposed in [24] for multi-label dimension reduction. MLSI employs the label information to guide the learning of the transformation and has been applied successfully in multi-label text classification. Classical LDA has been extended by Park and Lee [25] to handle multi-label data samples. However, it does not take label correlation into account. Wang et al. [26] proposed a novel multi-label linear discriminant analysis (MLDA) to take advantage of label correlation and explore the powerful discrimination ability to cope with multi-label DR. Zhang and Zhou [27] developed a multi-label dimensionality reduction via dependence maximization. However, it involves generalized eigenvalue decomposition which requires expensive computation cost especially for high-dimensional data. Sun et al. [28] investigated the relationship between canonical correlation analysis and least squares and proposed a least squares canonical correlation analysis for multi-label classification. Unlike CCA, partial least squares (PLS) [29] maximize the covariance of the two sets of variables in the transformed space. An equivalent relationship between CCA and PLS has been established in [30]. However, the above mentioned algorithms cannot capture high

* Corresponding author.
E-mail address: xinshu@outlook.com (X. Shu).

http://dx.doi.org/10.1016/j.neucom.2015.05.090
0925-2312/

order correlation information among different labels. As a result, a least squares formulation of hypergraph spectral learning has been proposed in [31] to capture the correlation information contained in different labels. To further incorporate the data and label correlation, a hypergraph canonical correlation analysis for multi-label classification has been presented in [32] recently. Li et al. [33] present a novel multi-label dimensionality reduction using the variable pairwise constraints. A more comprehensive review of multi-label dimensionality reduction as well as multi-label learning algorithms can be found in [34,35].

As indicated in [36,37], it is reasonable to assume that there is a certain common information shared among data samples and uncover this shared structure may improve learning performance. It has been claimed in [38] that there should be a shared subspace across multiple tasks and uncovering this shared subspace can improve classification performance. Yang et al. [39] have assumed that there is shared subspace among different labels and proposed a semi-supervised learning framework for multi-label image annotation. Based on the assumption that different related tasks may share common structures, Yang et al. [40] proposed a novel feature selection approach for multimedia analysis. Recently, Shu and Lu [41] proposed a trace norm regularized discriminant analysis for dimension reduction and simultaneously uncover the shared information among data samples. However, their algorithm is essentially devised for single-label problem which means it does not deal with multi-label data directly.

Motivated by the consideration that there should exist a common subspace to be shared among multiple labels, we attempt to extract a shared subspace for multi-label dimensionality reduction. Our work builds on the recent work of multi-label dimensionality reduction via dependence maximization (MDDM) [27]. We propose an efficient approach for computing the optimal solution of MDDM which requires much smaller computation time. Further analysis shows that MDDM can be reformulated as a least squares problem which enables us to capture the shared information among different labels in the least squares framework. In summary, the key contributions of this article are highlighted as follows:

- We propose an efficient algorithm for computing the optimal solution of MDDM [27] which avoids the direct eigendecomposition on the large scale matrix. For high-dimensional data set, the time complexity of the new algorithm is $O(mn^2)$ which is smaller than the original formulation $O(m^2d)$, where m is the number of features, n is the number of samples and d is the dimensionality of the lower-dimensional subspace.
- We further show that MDDM can be reformulated as least squares problems. Based on this equivalent relationship, we develop a shared subspace MDDM for multi-label dimensionality reduction.
- We have conducted extensive experiments on several benchmark datasets to demonstrate the effectiveness of the proposed formulation.

The rest of the article is organized as follows. Section 2 reviews MDDM. Section 3 presents the new technique to compute the optimal solution of MDDM. The shared subspace MDDM is presented in Section 4. We report experimental results in Section 5. Followed with conclusion in Section 6.

2. A brief review of MDDM

In this section, we give a brief review of MDDM. We focus on the linear version of MDDM. Some important notations have been first described in Table 1.

Suppose we are given a training data set $X=[x_1,x_2,\ldots,x_n]\in\mathbf{R}^{m\times n}$. Each x_i is associated with c labels which can be represented as a c-dimensional binary vector y_i, where $y_i(j)=1$ if x_i is associated with jth label, and $y_i(j)=0$ otherwise. Let $Y=[y_1,y_2,\ldots,y_n]\in\mathbf{R}^{c\times n}$ be the indicator matrix.

Table 1
Notations.

Notations	Descriptions
n	the number of training samples
m	the dimensionality of data point
c	the number of labels
d	the dimensionality of lower-dimensional subspace
x_i	the i-th data point
X	the data matrix
Y	the indicator matrix
H	the centering matrix
P	the transformation matrix

Since there should exist some relationships between the feature space and the label space associated with the same object, MDDM attempts to find a lower-dimensional feature space where the dependence between the input and the output is maximized. Denote the projection matrix as $P\in\mathbf{R}^{m\times d}$. The original data instance x_i is first projected into a new space by $\phi(x_i)=P^Tx_i$ and the reduced kernel function is given by

$$k(x_i,x_j)=\langle\phi(x_i),\phi(x_j)\rangle=\langle P^Tx_i,P^Tx_i\rangle$$

where $\langle a,b\rangle$ denotes the inner product defined as $\langle a,b\rangle=a^Tb$. For the label space, the corresponding kernel function can be simply defined as $\ell(y_i,y_j)=\langle y_i,y_j\rangle$.

Given (X,Y) with a joint distribution P_{XY}, the feature kernel matrix $K(K_{ij}=k(x_i,x_j))$ and the label kernel matrix $L(L_{ij}=\ell(x_i,x_j))$, the empirical Hilbert–Schmidt independence criterion (HSIC) is estimated by the trace of kernel matrices product as [42]

$$\mathrm{HSIC}(X,Y,P_{XY})=(n-1)^{-2}\mathrm{tr}(HKHL)$$

where H is the centering matrix defined as $H=I-(1/n)ee^T$. HSIC is proposed for measuring the statistical dependence of random variables and has been applied successfully for data clustering [43] and supervised feature selection [44].

In order to achieve the maximum dependence between the input and output space, MDDM aims to maximize the following expression to compute the optimal P:

$$\mathrm{HSIC}(X,Y,P_{XY})=(n-1)^{-2}\mathrm{tr}(HKHL)$$

Notice that $K=\langle\phi(X),\phi(X)\rangle=X^TPP^TX$, we obtain

$$P^*=\arg\max_P\mathrm{HSIC}(X,Y,P_{XY})=\mathrm{tr}(HX^TPP^TXHL)$$

where we drop the constant $(n-1)^{-2}$. To avoid trivial solution, an additional constraint for P is introduced which leads to the following expression:

$$\begin{aligned}\max_P\quad&\mathrm{trace}(HX^TPP^TXHL)\\ \text{s.t.}\quad&P^T\left(\mu XX^T+(1-\mu)I\right)P=I\end{aligned}\tag{1}$$

where μ is a parameter to control the importance between two constraints. When $\mu=0$, the projection vectors are orthonormal which is called orthogonal projection [27]. For $\mu=1$, the projection vectors are uncorrelated on the training data and the method is called uncorrelated subspace dimensionality reduction [27].

It is easy to verify that the optimal solutions of (1) are characterized by the following generalized eigenvalue problem:

$$S_2p=\lambda S_1p\tag{2}$$

where $S_1=\mu XX^T+(1-\mu)I$ and $S_2=XHLHX^T$. If S_1 is nonsingular, the optimal P is given by the eigenvectors of $S_1^{-1}S_2$ corresponding to the d largest eigenvalues, which requires $O(dm^2)$ computational cost. When