Neurocomputing 156 (2015) 221–230

Contents lists available at ScienceDirect

Neurocomputing

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

A least squares formulation of multi-label linear discriminant analysis

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

[a] College of Information Science and Technology, Nanjing Agricultural University, Nanjing, 210095, China
[b] Department of Computer Science, City University of Hong Kong

ARTICLE INFO

Article history:
Received 6 June 2014
Received in revised form
16 October 2014
Accepted 23 December 2014
Communicated by Deng Cai
Available online 5 January 2015

Keywords:
Multi-label linear discriminant analysis
Least squares
Dimension reduction
Spectral regression

ABSTRACT

The classical linear discriminant analysis has been recently extended to the multi-label dimensionality reduction. However, Multi-label Linear Discriminant Analysis (MLDA) involves dense matrices eigen-decomposition that is known to be computationally expensive for the large-scale problems. In this paper, we present that the formulation of MLDA can be equivalently casted as a new least-squares framework so as to significantly mitigate the computational overhead and scale to the data collections with higher dimension. Further, it is also found that appealing regularization techniques can be incorporated into the least-squares model to boost generalization accuracy. Experimental results on several popular multi-label benchmarks not only verify the established equivalence relationship, but also corroborate the effectiveness and efficiency of our proposed algorithms.

© 2015 Elsevier B.V. All rights reserved.

1. Introduction

Multi-label learning, handling data associated with multiple labels, has naturally gained considerable attention in many potential applications such as multi-topic document categorization [45,23], protein function prediction [2,29] and automatic image annotation [1,24]. More recently, a wide range of approaches have been developed for multi-label learning. According to [43], existing algorithms can be roughly divided into two categories: algorithm adaption and problem transformation. Algorithm adaption approaches attempt to adapt existing single-label classification algorithms to address the multi-label issue. Some notable examples include neural network [55], lazy learning [48,56,32], Adaboost MR [31,9] and rank SVM [11]. On the other hand, the multi-label classification problem can be transformed into several single-label classification problems so that existing single-label learning schemes can be easily utilized. Specifically, binary relevance method [43], pair-wise [13,51] and label embedding methods [18,39,28] fall into this transformation category.

However, multi-label learning frequently involves high-dimensional data which could make multi-label classification infeasible due to the curse of dimensionality. Therefore, dimension reduction (DR) becomes a necessary preprocessing step in the subsequent clustering or classification task. The state-of-the-art DR approaches include unsupervised methods such as PCA [19], ISOMAP [40], LLE [30], Laplacian Eigenmap [3], LPP [25], together with many supervised methods, i.e. linear discriminant analysis (LDA) [10,12], canonical correlation analysis (CCA) [14] to name a few. Additionally, it has been shown that LDA bears strong connections with least-squares problem [10] for binary classification. Many researchers have established a similar relationship between the least-squares and LDA in multi-class classification [15,53,7,42].

In essence, one main drawback of some existing DR algorithms is that they are designed for single-label multi-class classification, which means that they cannot be directly employed in multi-label learning. To deal with this problem, the classical LDA has been extended to handle multi-label DR [27] without taking into account label correlations. Wang et al. [46] develop a new multi-label linear discriminant analysis (MLDA) by defining class-wise scatter matrices in the conventional LDA. Though MLDA has been applied successfully in image annotation [46], it requires expensively computing the Moore–Penrose pseudo-inverse of the class-wise within-class scatter matrix to overcome the issue of singularity. Hence, an efficient implementation of MLDA awaits further consideration.

Inspired by the recent least-squares studies on the CCA [36] and spectral regression discriminant analysis (SRDA) [7], we establish that MLDA can be formulated as a least-squares formulation. On the basis of the equivalence relationship, the projection functions of MLDA can be gained by solving a set of linear equations and avoiding the costly computation of large size eigen-decomposition. In particular, the least-squares problem can be efficiently tackled by means of the conjugate gradient algorithms like LSQR [26], resulting in the superior performance in terms of large-scale data corpora. In summary, the primary contributions of this work are the following:

- The MLDA [46] is well analyzed, showing that MLDA is of time complexity $\mathcal{O}(mnt+t^3)$, where m is the number of features, n is the number of samples and $t=\min(m,n)$. When m and n are

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

http://dx.doi.org/10.1016/j.neucom.2014.12.057
0925-2312/© 2015 Elsevier B.V. All rights reserved.

large, such as high-dimensional text datasets, it is infeasible to directly apply MLDA [46] to learn the projection matrix.
- We extend the multi-label LDA (MLDA) to its least-squares formulation (LSMLDA). Consequently, the iterative conjugate gradient algorithm can be employed to efficiently handle the least-squares problem with large size in order to considerably reduce the computational cost.
- Besides, with least-squares as a building block, a group of attractive regularization techniques can be easily integrated with LSMLDA to control the complexity of a learning model and substantially improve the generalization performance.

The remainder of this paper is organized as below: Section 2 briefly reviews the related work. In Section 3, we give a brief review of MLDA and least squares. In Section 4, a detailed computational analysis of MLDA is rigorously presented. Section 5 introduces our least-squares MLDA (LSMLDA). Extensive experiments obtained by several noted image, biology, text data corpora are reported in Section 6, followed with conclusions in Section 7.

2. Related work

Many multi-label classification learning algorithms use dimension reduction as a preprocessing step. In this section, we discuss some closely related multi-label dimension reduction algorithms.

Multi-label informed latent semantic indexing (MLSI) was proposed in [54] 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. Canonical correlation analysis (CCA) [36] projects two sets of variables onto a lower-dimensional subspace in which they are maximally correlated. For multi-label problems, one variable represents the data sample and the other variable is derived from the label set. Similar to CCA, multi-label dimensionality reduction via dependence maximization (MDDM) [58] attempts to find a lower-dimensional subspace in which the dependence between the input features and the associated class labels is maximized. Unlike CCA, partial least squares (PLS) [49] maximizes the covariance of the two sets of variables in the transformed space. An equivalent relationship between CCA and PLS has been established in [38]. However, the above-mentioned algorithms cannot capture high order correlation information among different labels. As a result, a least squares formulation of hypergraph spectral learning has been proposed in [35] 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 [47] recently. Ref. [21] presents a novel multi-label dimensionality reduction using the variable pairwise constraints. Multi-label dimensionality reduction has also been studied in the context of semi-supervised learning [20]. A more comprehensive review of multi-label dimensionality reduction as well as multi-label learning algorithms can be found in [37,57].

3. Brief review of MLDA least squares

We present a brief review of MLDA [46] and least squares in this section. Some important notations have been first described in Table 1.

3.1. MLDA

Given a dataset with n samples $\{x_i, y_i\}_{i=1}^n$ and c classes, where $x_i \in \mathbb{R}^m$ and $y_i \in \{0,1\}^c$. $y_i(j)=1$ if x_i belongs to the j-th class and 0 otherwise. For convenience, we write $X=[x_1, x_2, \ldots, x_n] \in \mathbb{R}^{m\times n}$ and $Y=\{y_1, y_2, \ldots, y_n\}^T=[y_{(1)}, y_{(2)}, \ldots, y_{(c)}] \in \mathbb{R}^{n\times c}$. With these notations, the between-class scatter matrix, within-class scatter matrix and total scatter matrix for MLDA are defined as follows [46]:

$$S_b=\sum_{k=1}^{c} S_b^{(k)}, \quad S_b^{(k)}=\sum_{i=1}^{n} Y_{ik}(\mu_k-\mu)(\mu_k-\mu)^T$$

$$S_w=\sum_{k=1}^{c} S_w^{(k)}, \quad S_w^{(k)}=\sum_{i=1}^{n} Y_{ik}(x_i-\mu_k)(x_i-\mu_k)^T$$

$$S_t=\sum_{k=1}^{c} S_t^{(k)}, \quad S_t^{(k)}=\sum_{i=1}^{n} Y_{ik}(x_i-\mu)(x_i-\mu)^T$$

where μ_k is the mean of the k-th class and μ is the multi-label global mean, which are defined as follows:

$$\mu_k=\frac{\sum_{i=1}^{n} Y_{ik}x_i}{\sum_{i=1}^{n} Y_{ik}}, \quad \mu=\frac{\sum_{k=1}^{c}\sum_{i=1}^{n} Y_{ik}x_i}{\sum_{k=1}^{c}\sum_{i=1}^{n} Y_{ik}}$$

Similar to classical LDA, the objective function of MLDA is defined as follows:

$$a^*=\max_a \frac{a^T S_b a}{a^T S_w a} \tag{1}$$

Notice that $S_t=S_b+S_w$, the optimization problem (1) is equivalent to

$$a^*=\max_a \frac{a^T S_b a}{a^T S_t a}$$

When K projective functions $A=[a_1, a_2, \ldots, a_K]$ are needed, the objective function of MLDA can be rewritten as

$$A^*=\max_A \frac{\mathrm{tr}(A^T S_b A)}{\mathrm{tr}(A^T S_t A)} \tag{2}$$

where $\mathrm{tr}(\cdot)$ denotes the trace operator of a matrix. The optimization problem (2) is equivalent to finding all the eigenvectors that satisfy $S_b a=\lambda S_t a$, where $\lambda \neq 0$. The solution can be obtained by utilizing an eigen-decomposition on the matrix $S_t^{-1}S_b$, if S_t is nonsingular. When S_t is singular, a stable way to solve this eigen-problem is first to compute the pseudo-inverse of S_t, then solve the eigen-decomposition $S_t^{\dagger}S_b a=\lambda a$. One main shortcoming of this approach is the expensive computation of pseudo-inverse, especially for high-dimensional datasets.

Table 1
Notations.

Notations	Descriptions
n	The number of training samples
m	The dimension of data point
c	The number of labels
x_i	The i-th data point
S_b	The class-wise between-class scatter matrix
S_w	The class-wise within-class scatter matrix
S_t	The class-wise total scatter matrix
X	The data matrix
Y	The indicator matrix
$\tilde{X}$	The centered data matrix
a	The transformation vector
A	The transformation matrix

3.2. Least squares

Least squares is one of the most popular techniques for both regression and classification. Given a training dataset $X=[x_1, x_2, \ldots, x_n]$, $Y=[y_1, y_2, \ldots, y_n]$, where $x_i \in \mathbb{R}^m$ is the observation and $y_i \in \mathbb{R}^c$ is the corresponding response. Suppose that both X and Y are centered, i.e., $Xe=0$, $Ye=0$, where e is a all-ones vector with

Download English Version:

https://daneshyari.com/en/article/406254

Download Persian Version:

https://daneshyari.com/article/406254

Daneshyari.com