Knowledge-Based Systems 98 (2016) 172–184

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

A multi-label feature extraction algorithm via maximizing feature variance and feature-label dependence simultaneously

Jianhua Xu*, Jiali Liu, Jing Yin, Chengyu Sun

School of Computer Science and Technology, Nanjing Normal University, Nanjing, Jiangsu 210023, China

ARTICLE INFO

Article history:
Received 30 June 2015
Revised 21 January 2016
Accepted 21 January 2016
Available online 1 February 2016

Keywords:
Multi-label classification
Dimensionality reduction
Feature extraction
Principal component analysis
Hilbert–Schmidt independence criterion
Eigenvalue problem

ABSTRACT

Dimensionality reduction is an important pre-processing procedure for multi-label classification to mitigate the possible effect of dimensionality curse, which is divided into feature extraction and selection. Principal component analysis (PCA) and multi-label dimensionality reduction via dependence maximization (MDDM) represent two mainstream feature extraction techniques for unsupervised and supervised paradigms. They produce many small and a few large positive eigenvalues respectively, which could deteriorate the classification performance due to an improper number of projection directions. It has been proved that PCA proposed primarily via maximizing feature variance is associated with a least-squares formulation. In this paper, we prove that MDDM with orthonormal projection directions also falls into the least-squares framework, which originally maximizes Hilbert–Schmidt independence criterion (HSIC). Then we propose a novel multi-label feature extraction method to integrate two least-squares formulae through a linear combination, which maximizes both feature variance and feature-label dependence simultaneously and thus results in a proper number of positive eigenvalues. Experimental results on eight data sets show that our proposed method can achieve a better performance, compared with other seven state-of-the-art multi-label feature extraction algorithms.

1. Introduction

Traditional supervised classification solves problems in which one instance has one label only [1], which is regarded as single-label classification. However, in many real-world applications, one instance are possibly associated with multiple labels simultaneously. For example, a sunrise picture is annotated by sky, sun and sea at the same time [2]; a piece of news belongs to environment protection, haze, and weather [3]; a protein is located in different sub-cellular locations simultaneously [4]. Such a learning task is referred to as multi-label classification. Recently multi-label classification has been paid much attention to in machine learning, pattern recognition, data mining and statistics. Thus a variety of multi-label classification methods were proposed, such as, kNN-type methods [5,6], SVM-like techniques [7–10], ensemble classifiers [11–13], and collective learning techniques [14,15], most of which have been reviewed extensively in [16–20].

On the other hand, recent technological innovations allow us to collect massive amount of data with a large number of features. Bellman [21] coined a well-known term, curse of dimensionality. For classification issue, such a term implies that for a given instance size, there is a maximum number of features above which the performance of a classifier will degrade rather than improve [22]. Additionally, this cure also results in a high computational complexity in practice. To alleviate its possible effect, dimensionality reduction becomes an effective way to remove the irrelevant, redundant and noisy features, which generally covers feature extraction [22], and feature selection [23]. In this paper, we only focus on feature extraction for dimensionality reduction.

Regardless of single-label and multi-label classification, feature extraction techniques mainly could be grouped into two categories: unsupervised and supervised. What differentiates them is whether class label information is exploited or not.

Unsupervised methods extract a small number of features without using label information to retain as much discriminant information as possible. A representative linear method is principal component analysis (PCA), which could be described from two different viewpoints. One is to find an orthogonal low-dimensional sub-space via maximizing feature variance [1,24,25], and the other is via minimizing a squared reconstruction error [25–29]. In addition, some complicated methods, e.g., locally linear embedding(LLE) [30], Laplacian eigenmap [31], locality preserving projections(LPP) [32] and ISOMAP [33], aim at finding a nonlinear low-dimensional sub-space through preserving the data

* Corresponding author. Tel.: +86 2596306209; fax: +86 2585891990.
E-mail address: xujianhua@njnu.edu.cn, xujianhua99@tsinghua.org.cn (J. Xu).

http://dx.doi.org/10.1016/j.knosys.2016.01.032
0950-7051/© 2016 Elsevier B.V. All rights reserved.

manifold structures. In principle, these unsupervised approaches primarily in single-label classification could be directly applied to multi-label classification.

Conversely, supervised methods sufficiently and elaborately exploit label information in feature extraction procedures. Linear discriminant analysis (LDA) [1,25,34] is a typical single-label technique which obtains an optimal low-dimensional sub-space by maximizing the between-class scatter measure and minimizing the within-class scatter one at the same time. However, LDA cannot be directly applied to multi-label classification, since a multi-label instance belongs to several different classes simultaneously, and how much it should contribute to two scatter measures remains ambiguous. Therefore three generalized versions have been proposed. In [35], a multi-label instance is duplicated into several single-label ones and then a multi-label data set is converted into a single-label one, to fit LDA directly. To correct the over-counting problem in [35], multi-label LDA (MLDA) [36] is proposed as a weighted form, where the weights of each instance to all labels are estimated by label correlation information. Since the dimensions of low-dimensional sub-space from the above two methods does not exceed the number of classes minus 1, the between-class scatter matrix is changed in direct multi-label LDA (DMLDA) [37], to distinguish training instances that do not have a specific label from the mean vector of the instances belonging to this label.

Latent semantic indexing (LSI) turns to be a successful linear single-label feature extraction approach in document analysis and information retrieval [38]. Multi-label informed latent semantic indexing (MLSI) [39] extends LSI to obtain a low-dimensional sub-space which maximizes feature variance and binary label variance via a linear combination way. Theoretically, canonical correlation analysis (CCA) could be directly used in multi-label situation [40]. Further a least-squares formulation and its several regularized variants are extended for CCA in [41], which means that CCA could be converted into slightly different least-squares problems. Hilbert–Schmidt independence criterion (HSIC) [42] is a non-parametric dependence measure which considers all modes of dependencies between all variables. Multi-label dimensionality reduction via dependence maximization (MDDM) [43] attempts to search for a low-dimensional sub-space by maximizing feature-label dependence using HSIC with orthonormal projection directions and orthonormal projected features respectively (MDDMp and MDDMf in detail).

It is worth noting that most of the aforementioned feature extraction methods can fall in a least-squares framework, including PCA, LDA, CCA, LPP, LLE and Laplacian eigenmaps [28], MDDMf [44] and MLDA [45]. But whether MDDMp could be associated with a least-squares issue is still an open problem. Additionally, PCA is concerned with feature data structure only, MLSI deals with feature and label data structures at the same time, CCA and MDDM consider dependence from features to labels, and multi-label LDA forms implicitly express the relationship between features and labels. According to matrix theory [46], PCA, MLSI and DMLDA could produce many small positive eigenvalues, whereas CCA, MDDM and MLDA result in a few large positive eigenvalues, both of which could deteriorate the multi-label classification performance because of an improper number of projection directions.

In this study, we will take feature data structure and feature-label dependence simultaneously into consideration explicitly, and derive a proper number of positive eigenvalues or projection directions. We prove that MDDMp also falls in the least-squares framework and then propose a novel multi-label feature extraction method to integrate two least-squares formulations in PCA and MDDMp linearly, which both maximizes feature variance and maximizes feature-label dependence at the same time. Therefore our method is referred to as MVMD simply. Based on multi-output linear ridge regression [1,25] as our multi-label baseline classifier, our experimental results on eight benchmark data sets illustrate that, with a proper number of projected features, our MVMD is overall superior to the aforementioned multi-label feature extraction techniques including PCA, two MDDM versions, CCA, MLSI, and MLDA and DMLDA, according to six instance-based performance evaluation measures (Hamming loss, accuracy, F1, precision, recall and subset accuracy) and computational time.

Summarily, the main contributions of this paper are highlighted as follows: (a) we propose a least squares formulation for multi-label dimensionality reduction via dependence maximization with orthonormal projection directions (MDDMp); (b) via combing such a formulation with that of PCA linearly, a novel multi-label feature extraction approach is presented, which maximizes both feature variance and feature-label dependence simultaneously; (c) the extensive experiments on eight data sets demonstrate the effectiveness and efficiency of our proposed method.

The rest of this paper is organized as follows. Multi-label feature extraction setting is introduced in Section 2. In Section 3, PCA and its least-squares form are briefly introduced. Two MDDM versions are reviewed firstly and then a least-squares formulation for MDDMp is derived in Section 4. In Section 5 we propose and analyze our novel feature extraction algorithm (MVMD). The related work is formally summarized in Section 6. Section 7 is devoted to experiments with eight benchmark data sets. Finally this paper ends with some conclusions in Section 8.

2. Multi-label feature extraction setting

Let $Q = \{1, 2, \ldots, q\}$ be a finite set of q class labels, and 2^Q all possible subsets of Q. We denote a multi-label training data set of size l drawn identically and independently from an unknown probability distribution in a D-dimensional real space by,

$$\{(\mathbf{x}_1, L_1), \ldots, (\mathbf{x}_i, L_i), \ldots, (\mathbf{x}_l, L_l)\}, \tag{1}$$

where $\mathbf{x}_i \in R^D$ and $L_i \in 2^Q$ represent the ith instance and its relevant label set. Additionally, the complement of L_i, i.e., $\bar{L}_i = Q \backslash L_i$, is referred to as the irrelevant label set of $\mathbf{x}_i$.

The goal of multi-label classification is to learn a classifier $f(\mathbf{x})$: $R^D \to 2^Q$ which can predict the relevant labels for unseen instances [5,8,9,16–20].

The multi-label linear feature extraction is to derive a projection matrix $\mathbf{P} \in R^{D \times d}$ that maps any original instance vector $\mathbf{x}$ in the D-dimensional space into a projected one $\mathbf{x}'$ in the lower d-dimensional sub-space ($d < D$), to preserve the information and structure of original data based on a certain criterion [22],

$$\mathbf{x}' = \mathbf{P}^T \mathbf{x}, \tag{2}$$

where T indicates the transpose operation of matrix or vector. For the convenience of representation, we also adopt a binary vector $\mathbf{y}_i = [y_{i1}, y_{i2}, \ldots, y_{iq}]^T$ to label the instance $\mathbf{x}_i$, where $y_{ik} = 1$ if the kth label is in L_i, and -1 otherwise. Furthermore, we utilize the following two matrices to depict feature and binary label data,

$$\begin{aligned} \mathbf{X} &= [\mathbf{x}_1, \ldots, \mathbf{x}_i, \ldots, \mathbf{x}_l]^T, \\ \mathbf{Y} &= [\mathbf{y}_1, \ldots, \mathbf{y}_i, \ldots, \mathbf{y}_l]^T, \end{aligned} \tag{3}$$

where the ith rows in $\mathbf{X}$ and $\mathbf{Y}$ correspond to the ith training instance. In this study, such two matrices are centered for all feature extraction methods but LDA-type techniques, i.e.,

$$\begin{aligned} \mathbf{X} &\Leftarrow \mathbf{HX}, \\ \mathbf{Y} &\Leftarrow \mathbf{HY}, \end{aligned} \tag{4}$$

through the centered matrix,

$$\mathbf{H} = \mathbf{I} - \mathbf{u}\mathbf{u}^T / l, \tag{5}$$

where $\mathbf{I}$ is an identity matrix of size l and $\mathbf{u}$ denotes an all-one column vector of length l. Without the loss of generality, we still use