Journal of Visual Communication and Image Representation 49 (2017) 361–370

Contents lists available at ScienceDirect

Journal of Visual Communication and Image Representation

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

Speed up kernel dependence maximization for multi-label feature extraction☆

Xin Shu[a,*], Jing Qiu[b]

[a] College of Information Science and Technology, Nanjing Agricultural University, Nanjing 210095, China
[b] Suzhou University of Science and Technology, China

ARTICLE INFO

Keywords:
Multi-label dimensionality reduction
Dependence maximization
Least squares
Hilbert-Schmidt independence criterion

ABSTRACT

Kernel dependence maximization for multi-label dimensionality reduction (kMDDM) has been proposed recently to cope with high-dimensional multi-label data. In order to produce discriminant projection vectors, kMDDM utilize the Hilbert-Schmidt independence criterion to capture the dependence between the feature description and the associated labels. However, the computation of kMDDM involves dense matrices eigen-decomposition that is known to be computationally expensive for large scale problems. In this paper, we reformulate the original kMDDM as a least-squares problem, so as to significantly lessen computational burden by utilizing the conjugate gradient algorithms. Further, appealing regularization techniques can be incorporated into the least-squares model to boost the generalization performance. Extensive experiments conducted on benchmark data collections verify the effectiveness of our proposed model.

1. Introduction

Multi-label classification has recently received considerable attention in various applications such as automatic image annotation [1–3], multi-topic document categorization [4,5] and protein function prediction [6,7]. Different from traditional single-label classification where each instance belongs to only one class, multi-label classification tackles problems where each instance may associate with more than one classes. A large body of algorithms have been developed in the literature. According to [8], existing approaches can be roughly split 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 [9,10], lazy learning [11–13], Adaboost MR [14,15], rank SVM [16]. 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 [8], pair-wise method [17,18] and label embedding method [19–21]. Madjarov et al. [22] extends this categorization of multi-label methods with a third group of methods, namely, ensemble methods. Algorithms belonging to this group include RAkEL [23], ensembles of classifier chains [24]. Recently, multi-view feature learning algorithms are introduced to deal with multi-label problem [25,26]. In addition, sparse and low rank representation has also been introduced to develop robust multi-label learning algorithms [27–30]. In particular, Yu et al. [29] employ the low rank constraints to deal with multi-label learning problem with missing labels. The work in [30] combines the merits of privileged information and low-rank constraints to explore and exploit the relationship between labels in multi-label learning problems. When labels are arrived on the fly, You et al. [31] have proposed streaming label learning (SLL) framework which is capable of modelling newly-arrived labels with the help of the knowledge learned from past labels.

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 [32] 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 linear discriminant analysis has been extended by Park and Lee [33] to handle multi-label data samples. However, it does not take label correlation into account. Wang et al. [34] proposed a novel multi-label linear discriminant analysis (MLDA) to take advantage of label correlation and explore the powerful discrimination ability to copy with multi-label DR. Zhang and Zhou [35] developed a multi-label dimensionality reduction

☆ This paper has been recommended for acceptance by Zicheng Liu.

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

http://dx.doi.org/10.1016/j.jvcir.2017.10.006
Received 26 December 2016; Received in revised form 1 September 2017; Accepted 13 October 2017
Available online 14 October 2017

via dependence maximization (MDDM). MDDM uses Hilbert-Schmidt independence criterion (HSIC) to capture the strong dependence between the feature description and the associated labels. MDDM involves generalized eigenvalue decomposition which requires expensive computation cost especially for high-dimensional data. Canonical correlation analysis (CCA) [36] and partial least squares (PLS) [37] have been utilized for multi-label feature extraction. CCA aims to find a pair of vectors, one for each variable, such that the data are maximally correlated in the transformed space. Unlike CCA, PLS maximizes the covariance of the two sets of variables in the transformed space. Further, an equivalent relationship between CCA and PLS has been established in [38]. In addition to the above linear multi-label dimensionality reduction approaches, nonlinear algorithm have also been studied in the literature. Kernel CCA has been presented in [36]. A kernel extension of MDDM (kMDDM) has been developed in [35]. However, both linear and nonlinear approaches involve eigenvalue decomposition on dense matrix which is expensive for large scale dataset.

The above multi-label feature extraction approaches usually need to solve an eigenvalue problem which is usually computational expensive and can not scale to large problems. Recently, some works have devoted to solve the scalability issue in multi-label feature extraction. The authors in [39] develop a least squares formulation for a class of generalized eigenvalue problems and employ conjugate gradient algorithm to speed up the training process. Specifically, in [36], CCA and kernel CCA have been reformulated as a least squares problem. Shu et al. [40] reformulate MLDA as a least squares problem whose solution can be efficiently solved via conjugate gradient algorithm which has linear time complexity in terms of dimensionality. The authors in [41] further show that MDDM can be formulated as a least squares problems. However, the least squares formulation of kMDDM is still an open problem.

In this paper, we develop an efficient algorithm for solving kMDDM. Our work builds on the recent work of kernel MDDM [35]. We first give a computational analysis for kMDDM and show that the original kMDDM has $O(n^3)$ complexity. We then propose an equivalent least squares formulation for kMDDM to reduce the computational cost. In summary, the key contributions of this article are highlighted as follows.

- We propose an efficient SVD based approach for computing the optimal solution of kMDDM [35]. Compared with original formulation, whose time complexity requires $\frac{5}{6}n^3 + \frac{9}{2}n^3$ flam, the new algorithm requires only $\frac{9}{2}n^3$ flam which is smaller than the original formulation.
- We further show that kMDDM can be reformulated as a least squares problems. Based on this equivalent relationship, the solution of kMDDM can be efficiently derived via conjugate gradient algorithms. Further, appealing regularization techniques can be incorporated into the least squares framework to boost the generalization performance.
- 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 kMDDM. Section 3 presents a computational analysis for kMDDM. The equivalent least squares formulation of kMDDM and its extension are presented in Section 4. We report experimental results in Section 5. Followed with conclusion in Section 6.

2. HSIC and kMDDM

In this section, we give a brief review of kMDDM [35]. Before presenting kMDDM, we first introduce the Hilbert-Schmidt independence criterion (HSIC) [42] since kMDDM is based on HSIC.

Table 1
Summary of relevant matrices and their associated computational complexity.

Matrix	Size	Computation	Complexity
K	$n \times n$	SVD	$\frac{9}{2}n^3$
B	$c \times t$	SVD	$\frac{9}{2}c^3$

Table 2
Summary of statistics of the data sets. d is the dimensionality, n is the number of data samples, c is the number of labels.

Dataset	d	n	c
scene	294	2407	6
yeast	103	2414	14
Arts	17973	7441	19
Business	16621	9968	17
Computers	25259	12371	23
Education	20782	11817	14
Health	18430	9109	14
Reference	26397	7929	15
Recreation	25095	12797	18
Science	24002	6345	22
Social	32492	11914	21
Society	29189	14507	21
RV1	47236	6000	101
Pascal07	512	9963	21

Table 3
Performance of kMDDM and LSkMDDM in terms of AUC, macro F1 and micro F1 on scene and yeast datasets.

Dataset	Method	AUC	macro F1	micro F1
Scene	kMDDM	0.6038	0.2243	0.2491
	LSkMDDM	0.6039	0.2247	0.2510
Yeast	kMDDM	0.5980	0.4212	0.6063
	LSkMDDM	0.6005	0.4229	0.6062

2.1. Hilbert-Schmidt independence criterion

Given two random variables $x \in \mathscr{X}$ and $y \in \mathscr{Y}$ with joint probability p_{xy}. HSIC measures the dependence between x and y by the computing the norm of the cross-variance operator over the domain $\mathscr{X}$ and $\mathscr{Y}$ in the Hilbert space. Specifically, let us denote by $\mathscr{F}$ the reproducing kernel Hilbert space (RKHS) on $\mathscr{X}$ with feature map $\phi: \mathscr{X} \to \mathscr{F}$ and kernel $K_x: \mathscr{X} \times \mathscr{X} \to R$. Let $\mathscr{G}$ be another RKHS with feature map ψ and kernel K_y. The cross-covariance is defined as

$$C_{xy} = E_{xy}[(\phi(x)-\mu_x) \otimes (\psi(y)-\mu_y)]$$

where $\mu_x = E(\phi(x)), \mu_y = E(\psi(y))$ and $\otimes$ denotes the tensor product. Given that $\mathscr{F}$ and $\mathscr{G}$ are separable RKHSs, the square of the Hilbert-Schmidt norm of the cross-covariance is called HSIC. With kernel representation, HSIC can be expressed as

$$\begin{aligned} HSIC = \; & E_{xx'yy'}[K_x(x,x')K_y(y,y')] + E_{xx'}[K_x(x,x')]E_{yy'}[K_y(y,y')] \\ & - 2E_{xy}[E_{x'}[K_x(x,x')]E_{y'}[K_y(y,y')]] \end{aligned}$$

where (x,y) and (x',y') are two independent pairs drawn independently from p_{xy}, $E_{xx'yy'}$ is the expectation over these pairs. Given a finite set of data pairs $Z = \{(x_i,y_i)\}_{i=1}^n$ independently drawn from P_{xy}, the empirical estimate for HSIC is given by [42]

$$HSIC(Z,\mathscr{F},\mathscr{G}) = (n-1)^{-2} tr(K_x H K_y H)$$

where $H = I - \frac{1}{n}ee^T$ is the centering matrix, I is the $n \times n$ identity matrix, e is a $n \times 1$ vector with all elements are ones and n is the number of samples.