J. Vis. Commun. Image R. 36 (2016) 69–79

Contents lists available at ScienceDirect

J. Vis. Commun. Image R.

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

Kernel propagation strategy: A novel out-of-sample propagation projection for subspace learning ☆

Shuzhi Su, Hongwei Ge *, Yun-Hao Yuan *

Key Laboratory of Advanced Process Control for Light Industry (Ministry of Education), School of Internet of Things Engineering, Jiangnan University, Wuxi 214122, China

ARTICLE INFO

Article history:
Received 22 July 2015
Accepted 9 January 2016
Available online 16 January 2016

Keywords:
Kernel matrix optimization
Propagation
Out-of-sample projection
Semi-supervised learning
Canonical correlation analysis
Dimensionality reduction
Subspace feature extraction
Multi-view learning

ABSTRACT

Kernel matrix optimization (KMO) aims at learning appropriate kernel matrices by solving a certain optimization problem rather than using empirical kernel functions. Since KMO is difficult to compute out-of-sample projections for kernel subspace learning, we propose a kernel propagation strategy (KPS) based on data distribution similar principle to effectively extract out-of-sample low-dimensional features for subspace learning with KMO. With KPS, we further present an example algorithm, i.e., kernel propagation canonical correlation analysis (KPCCA), which naturally fuses semi-supervised kernel matrix learning and canonical correlation analysis by means of kernel propagation projections. In KPCCA, the extracted correlation features of out-of-sample data not only incorporate integral data distribution information but also supervised information. Extensive experimental results have demonstrated the superior performance of our proposed method.

1. Introduction

Subspace learning is a prevalent research field in pattern recognition and machine learning. Feature extraction and feature selection are two important topics in subspace learning. The former is intended to find a set of projection directions using a certain criterion to extract low-dimensional features from high-dimensional data, while the latter aims at discovering an appropriate subset from original feature set. In this paper, we focus on feature extraction. Typical feature extraction algorithms include principal component analysis (PCA) [1], linear discriminant analysis (LDA) [2], locality preserving projection (LPP) [3], and canonical correlation analysis (CCA) [4].

PCA, LDA and LPP are three typical single-view subspace learning algorithms. PCA is an unsupervised method for seeking a linear subspace where the projections of data possess maximum variance. Generally, PCA is difficult to well cater complex nonlinear data, and thus kernel PCA (KPCA) [5,6] was proposed to solve the problem. KPCA firstly maps original data into a higher (even infinite) dimensional kernel space, and then implements linear PCA in the kernel space. By utilizing class labels, LDA learns a discriminant subspace where the classes of objects can be properly separated. Similar to PCA, LDA is also a representative work of globally linear subspace learning, and a kernel discriminant analysis method [7] was proposed to extract nonlinear low-dimensional face features. In addition, LPP can preserve local information hidden in data as much as possible. Likewise, kernel-based LPP algorithms [8] have also been presented for better capturing nonlinear relationships among data.

CCA is an important multi-view learning method. The method aims at seeking a linear transformation for each of two views, which was proposed by Hotelling [9] as early as 1936. Up to now, CCA-related algorithms have been applied to many scientific fields, including genomic data analysis [10], information forecast [11], feature fusion [4], etc. CCA is linear and difficult to cater complex nonlinear data in many real-world applications. To extract nonlinear correlation features, some CCA variants [12–14] exploit different graph structures of data to capture nonlinear information among data. Recently, Shen et al. [15] proposed a unified multiset CCA framework based on graph embedding for dimensionality reduction (GbMCC-DR), which provides a unified viewpoint to embed different graphs into correlation analysis algorithms. With this framework, Shen et al. respectively borrowed the idea from the graphs in LDA, local discriminant embedding (LDE) [16], and marginal Fisher analysis (MFA) [17], and further developed three example algorithms, i.e. GbMCC-LDA, GbMCC-LDE, and GbMCC-MFA. In addition, Kernel CCA (KCCA) [18] is also a frequently used algorithm for extracting nonlinear correlation features, which has

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

* Corresponding authors.
E-mail addresses: ghw8601@163.com (H. Ge), yyhzbh@163.com (Y.-H. Yuan).

http://dx.doi.org/10.1016/j.jvcir.2016.01.007
1047-3203/

been applied into many scientific fields, such as bioinformatics [10], image retrieval [19], expression analysis [20], and blind identification [21]. Later, Zhu et al. [22] proposed a new kernel-based CCA algorithm called mixed KCCA, which projects the original data into a reproducing kernel Hilbert space with mixed kernels, i.e. a linear combination between local and global kernels. To better analyze complex but structured data, Arthur et al. [23] proposed a kernel generalized CCA (KGCCA). Ashad et al. [24] presented a higher-order regularized KCCA to overcome ill-posed solutions of KCCA on some specific cases. To enhance the discriminative power of nonlinear correlation features, Sun et al. [25] proposed kernel discriminative CCA (KDCCA) that maximizes within-class correlations of inter-view data and minimizes between-class correlations of inter-view data in kernel spaces. Recently, Jing et al. [26] presented a novel multi-view subspace learning algorithm called kernel intra-view and inter-view supervised correlation analysis (KI^{2}SCA). KI^{2}SCA fully utilizes the within-class and between-class correlation information from inter-view and intra-view data.

Most of kernel-based algorithms utilize empirical kernel methods (EKMs) in feature extraction. That is, kernel matrices are computed based on empirical kernel functions. Recently, researchers have developed a new way [27–31] to obtain kernel matrices instead of using EKMs, i.e., learning kernel matrices by an optimization problem. In this paper, we refer to the new way as kernel matrix optimization (KMO). KMO is data-dependent and often draws support from supervised information. The learned kernel matrices by KMO are usually able to better reveal authentic nonlinear relationships of data than those by EKMs. Due to the high data-dependent property and the lack of out-of-sample supervised information, KMO is usually difficult to effectively map out-of-sample data into kernel spaces. Therefore, subspace learning algorithms with KMO have some trouble in extracting subspace features of out-of-sample data.

To solve the above problem, we propose a kernel propagation strategy (KPS) based on data distribution similar principle for out-of-sample projections. The main idea of KPS is that out-of-sample projections of the kernel space should be similar to those of its neighbor samples. With the help of KPS, we further present a typical example algorithm, called kernel propagation CCA (KPCCA), which integrates KMO into CCA. KPCCA can obtain the subspaces with the well class separate property, i.e. intraclass compactness and interclass separability. In addition, out-of-sample correlation features extracted by KPCCA not only contain data distribution information of training and testing samples, but also inherit supervised information from the learned kernel matrices of training data, which are well beneficial for final recognition tasks. To evaluate our proposed algorithm, we design extensive experiments on four real-world image datasets. Promising experimental results have showed the effectiveness of our algorithm in image recognition tasks.

The rest of the paper is organized as follows. In the next section, we provide a brief review of KCCA and some important concepts. We introduce KPS, KPP and PCP in Section 3, and then present our KPCCA algorithm in Section 4, In Section 5, extensive experiments are designed for evaluating our algorithm. We conclude the paper in the last section.

2. Preliminary

2.1. Kernel canonical correlation analysis

In many real-world applications, kernel-based subspace learning algorithms can cater complex real-world data to some extent, and EKMs are commonly used for extracting nonlinear subspace features. In the section, we briefly review KCCA [18] that is a typical subspace learning algorithm with EKMs.

KCCA is a two-view joint feature extraction method, and aims at seeking a nonlinear transformation for each of two view datasets, so we suppose that two view datasets are $X=\{x_1,x_2,\ldots,x_n\}\in R^{p\times n}$ and $Y=\{y_1,y_2,\ldots,y_n\}\in R^{q\times n}$, and a pair of samples $\{x_i,y_i\}(i=1,2,\ldots,n)$ come from the same object, where n is the number of samples, and p (or q) is the sample dimension. In this paper, X and Y are treated as two training sample sets, and the corresponding testing sample sets are $\widetilde{X}=\{\tilde{x}_1,\tilde{x}_2,\ldots,\tilde{x}_N\}\in R^{p\times N}$ and $\widetilde{Y}=\{\tilde{y}_1,\tilde{y}_2,\ldots,\tilde{y}_N\}\in R^{q\times N}$, where N denotes the number of testing samples. Through two mappings $x_i\mapsto\varphi(x_i)\in R^{p_\varphi}$ and $y_i\mapsto\phi(y_i)\in R^{q_\phi}\quad(i=1,2,\ldots,n)$, X and Y can be mapped into the higher (even infinite) dimensional kernel spaces Ω^{p_φ} and Ω^{q_ϕ}, i.e. $\varphi(X)=[\varphi(x_1),\varphi(x_2),\ldots,\varphi(x_n)]\in R^{p_\varphi\times n}$ and $\phi(Y)=[\phi(y_1),\phi(y_2),\ldots,\phi(y_n)]\in R^{q_\phi\times n}$, where p_φ (or q_ϕ) is the dimension. By means of the two mappings, $\widetilde{X}$ and $\widetilde{Y}$ can be also mapped into the kernel spaces, i.e.$\varphi(\widetilde{X})=[\varphi(\tilde{x}_1),\varphi(\tilde{x}_2),\ldots,\varphi(\tilde{x}_N)]\in R^{p_\varphi\times N}$ and $\phi(\widetilde{Y})=[\phi(\tilde{y}_1),\phi(\tilde{y}_2),\ldots,\phi(\tilde{y}_N)]\in R^{q_\phi\times N}$. In addition, we assume $\varphi(X)$ and $\phi(Y)$ have been mean-normalized, and one can refer to [5] for detailed introduction of the normalized method. For simplifying the notation system of this paper, we uniformly utilize these notations in all the sections.

A pair of canonical projection directions $\alpha_\varphi\in R^{p_\varphi}$ and $\beta_\phi\in R^{q_\phi}$ can be obtained by maximizing the correlations between $\alpha_\varphi^T\varphi(X)$ and $\beta_\phi^T\phi(Y)$. *More specifically*, the optimization problem of KCCA can be formulated as follows:

$$\begin{aligned}&\max_{\alpha_\varphi,\beta_\phi}\quad \alpha_\varphi^T\varphi(X)\phi(Y)^T\beta_\phi\\ &s.t.\quad \alpha_\varphi^T\varphi(X)\varphi(X)^T\alpha_\varphi=1,\quad \beta_\phi^T\phi(Y)\phi(Y)^T\beta_\phi=1\end{aligned}\tag{1}$$

By the kernel trick [5], it is assumed that $\alpha_\varphi=\varphi(X)\alpha$ and $\beta_\phi=\phi(Y)\beta$, where $\alpha\in R^{n\times1}$ and $\beta\in R^{n\times1}$. Therefore, Eq. (1) can be rewritten as

$$\begin{aligned}&\max_{\alpha,\beta}\quad \alpha^TS_{xy}\beta\\ &s.t.\quad \alpha^TS_{xx}\alpha=1,\quad \beta^TS_{yy}\beta=1\end{aligned}\tag{2}$$

where $S_{xy}=K^{(x)}K^{(y)}$, $S_{xx}=K^{(x)}K^{(x)}$, and $S_{yy}=K^{(y)}K^{(y)}$. In addition, $K^{(x)}=\varphi(X)^T\varphi(X)=\left[k_x(x_i,x_j)\right]_{i,j=1}^{n,n}\in R^{n\times n}$ and $K^{(y)}=\phi(Y)^T\phi(Y)=\left[k_y(y_i,y_j)\right]_{i,j=1}^{n,n}\in R^{n\times n}$ are kernel matrices, and the kernel functions k_x and k_y may be any kernel satisfying the Mercer's condition [5], such as Gaussian kernel, linear kernel, and polynomial kernel.

Eq. (2) can be solved by using the Lagrangian multiplier method. The Lagrangian $\mathcal{L}$ is given by

$$\mathcal{L}=\alpha^TS_{xy}\beta+\frac{\lambda_x}{2}(1-\alpha^TS_{xx}\alpha)+\frac{\lambda_y}{2}(1-\beta^TS_{yy}\beta)\tag{3}$$

By setting the derivative of $\mathcal{L}$ with respect to α and β to zero, we have

$$\frac{\partial\mathcal{L}}{\partial\alpha}=S_{xy}\beta-\lambda_xS_{xx}\alpha=0\tag{4}$$

$$\frac{\partial\mathcal{L}}{\partial\beta}=S_{xy}^T\alpha-\lambda_yS_{yy}\beta=0\tag{5}$$

Multiplying both sides of Eqs. (4) and (5) by α^T and β^T respectively, we obtain

$$\alpha^TS_{xy}\beta=\lambda_x\alpha^TS_{xx}\alpha=\lambda_x$$

$$\beta^TS_{xy}^T\alpha=\lambda_y\beta^TS_{yy}\beta=\lambda_y$$

So $\lambda_x=\lambda_x^T=(\alpha^TS_{xy}\beta)^T=\beta^TS_{xy}^T\alpha=\lambda_y$. Let $\lambda=\lambda_x=\lambda_y$, then Eqs. (4) and (5) can be equally transformed into the following generalized eigenvalue problem: