Neurocomputing 182 (2016) 322–331

Contents lists available at ScienceDirect

Neurocomputing

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

Kernel canonical correlation analysis via gradient descent ☆

Jia Cai [a,*], Yi Tang [b], Jianjun Wang [c]

[a] School of Mathematics and Statistics Guangdong University of Finance & Economics, Guangzhou, Guangdong 510320, China
[b] School of Mathematics and Computer Science, Yunnan University of Nationalities, Kunming, Yunnan 650500, China
[c] School of Mathematics and Statistics, Southwest University, Chongqing 400700, China

ARTICLE INFO

Article history:
Received 1 July 2015
Received in revised form
3 November 2015
Accepted 14 December 2015
Communicated by Shiliang Sun
Available online 24 December 2015

Keywords:
Kernel CCA
Gradient descent
Reproducing kernel Hilbert space
Content-based image retrieval

ABSTRACT

Kernel canonical correlation analysis (CCA) is a powerful statistical tool characterizing nonlinear relations between two sets of multidimensional variables. It has been widely used in many branches of science and technology, e.g. bioinfomatics, multi-media information retrieval, cross-language document retrieval, fMRI (functional magnetic resonance imaging). Previous algorithms focus on sparsity analysis of kernel CCA. In this paper, from another viewpoint, we address a new gradient descent kernel CCA algorithm, which is based on the relation between kernel CCA and linear systems of equations. Meanwhile, stability analysis of the algorithm is addressed by means of suitable error decomposition formula and compact operator theory. Theoretical analysis is elegantly investigated in terms of choices of regularization parameter and step size. Experimental results on real-world datasets demonstrate the effectiveness of the algorithm for content-based image retrieval task. The results indicate that the proposed algorithm is stable and the performance is comparable with several state-of-the-art CCA algorithms.

1. Introduction

Exploiting useful information from data is significant in the community of modern statistical data analysis. Canonical correlation analysis (CCA) is a powerful statistical tool for finding the correlation between two sets of multidimensional variables. Proposed by Hotelling [20], CCA aims at seeking a pair of linear transformations associated with the two sets of variables such that the projected variables in the lower-dimensional space are maximally correlated. It has many applications in, for instance, machine learning [18], cross-language document retrieval [30], genomic data analysis [32], multi-view learning [28]. The optimal pair of linear transformations can be solved via a generalized eigenvalue problem, which is computationally expensive for high-dimensional data. Moreover, CCA fails to capture non-linear relations due to its linearity. It is not adequate for studying relation among variables in a wide range of practical problems, especially when dealing with the data that are not in the form of vectors, such as images, microarray data and so on. Hence, detecting non-linear relations among data is crucial in the community of data analysis. Therefore, a natural extension of CCA, namely kernel CCA was introduced to explore and exploit nonlinear relations among data [1] by the frequently used kernel technique [26]. Let us review the kernel CCA problem firstly. Given two random variables x and y, non-linear mappings $f(x)$ and $g(y)$. Kernel CCA solves

$$\max_{f\in\mathcal{H}_X, g\in\mathcal{H}_Y} \frac{\mathrm{Cov}[f(x),g(y)]}{\mathrm{Var}[f(x)]^{1/2}\mathrm{Var}[g(y)]^{1/2}}. \tag{1.1}$$

Here $f\neq 0, g\neq 0$. $\mathcal{H}_X, \mathcal{H}_Y$ are RKHSs (reproducing kernel Hilbert spaces, see [11–13] and the references therein) of real-valued functions on measurable spaces $\mathcal{X}, \mathcal{Y}$ respectively, endowed with measurable positive semi-definite kernels k_X, k_Y. 'Cov' denotes the covariance between $f(x)$ and $g(y)$, and 'Var' means the variance of functions. In practice, when given an i.i.d. sample $\{(x_i,y_i)\}_{i=1}^m$ from some unknown probability measure ρ, an ERM (empirical risk minimization) estimation of (1.1) takes form

$$\max_{f\in\mathcal{H}_X, g\in\mathcal{H}_Y} \frac{\widehat{\mathrm{Cov}}[f(x),g(y)]}{\widehat{\mathrm{Var}}[f(x)]^{1/2}\widehat{\mathrm{Var}}[g(y)]^{1/2}}, \tag{1.2}$$

where

$$\widehat{\mathrm{Cov}}[f(x),g(y)] = \frac{1}{m}\sum_{i=1}^{m}\left(f(x_i)-\frac{1}{m}\sum_{j=1}^{m}f(x_j)\right)\left(g(y_i)-\frac{1}{m}\sum_{j=1}^{m}g(y_j)\right),$$

☆The work described in this paper is supported partially by National Natural Science Foundation of China (Nos. 11401112, 61462096, 61273020), Fundamental Research Funds for the Central Universities (No. XDJK2015A007), Natural Science Foundation of Guangdong (Nos. 2015A030313628, 2015A030310304), Foundation for Distinguished Young Talents in Higher Education of Guangdong (No. 2013LYM0032), Science and Technology Innovation Project of Guangdong (Nos. 2013KJCX0083, 2014KQNCX150).

* Corresponding author.
E-mail addresses: jiacai1999@gdufe.edu.cn (J. Cai), yitang4math@ynni.edu.cn (Y. Tang), wjj@swu.edu.cn (J. Wang).

http://dx.doi.org/10.1016/j.neucom.2015.12.039
0925-2312/

$$\widehat{\mathrm{Var}}[f(x)]=\frac{1}{m}\sum_{i=1}^{m}\left(f(x_i)-\frac{1}{m}\sum_{j=1}^{m}f(x_j)\right)^2,$$

$$\widehat{\mathrm{Var}}[g(y)]=\frac{1}{m}\sum_{i=1}^{m}\left(g(y_i)-\frac{1}{m}\sum_{j=1}^{m}g(y_j)\right)^2.$$

In the literature, Tikhonov regularization scheme was introduced in the denominator to overcome over-fitting problem [18], i.e. replacing $\widehat{\mathrm{Var}}[f(x)]$, $\widehat{\mathrm{Var}}[g(y)]$ with $\widehat{\mathrm{Var}}[f(x)]+\varepsilon_m\|f\|^2_{\mathcal{H}_\mathcal{X}}$ and $\widehat{\mathrm{Var}}[g(y)]+\varepsilon_m\|g\|^2_{\mathcal{H}_\mathcal{Y}}$, respectively. Here regularization coefficient ε_m is a positive constant. Fukumizu et al. [16] investigated the above analysis via the idea of cross-covariance operators, while Cai and Sun [5] addressed convergence rates of it under AC condition. Kernel CCA has been widely used in many fields of science and technology, including: independent component analysis [2], biology and neurology [19,29], bioinformatics [32], image retrieval [18], and cross-language document retrieval [30].

However, how to select regularization parameter ε_m theoretically and practically remains largely unsolved. Here we consider a novel least squares model of kernel CCA with l^2-penalty, and propose a gradient descent algorithm to implement it. Essential difficulties arise when one apply optimization methods such as gradient descent to solve Eq. (1.2), we overcome it by employing a new relation between kernel CCA and linear systems of equations. Our main contributions are stated as follows:

- Motivated by the ideas of [8] and [9], we discuss a new model of kernel CCA via the theory of least squares in terms of l^2-penalty. Theoretical consistency is justified elegantly.
- A gradient descent algorithm for kernel CCA is analysed, which is novel in the literature of kernel CCA. Choices of step size and regularization parameters are also addressed.

The rest of the paper is organized as follows. In Section 2, we describe a gradient descent algorithm for kernel CCA. Theoretical analysis will be given in Section 3. Proof of theoretical results goes to Section 4. Section 5 devotes to the numerical results of the newly proposed algorithm.

2. Preliminaries and gradient descent kernel CCA algorithm

2.1. Preliminaries

Before delving into the algorithm, let us rewrite (1.2) in another form. Kernel CCA constructs a feature map ϕ_x such that a data matrix $X=(x_1,\ldots,x_m)\in\mathbb{R}^{n_1\times m}$ can be converted to

$$\Phi_x=(\phi_x(x_1),\ldots,\phi_x(x_m))\in\mathbb{R}^{\mathcal{N}_1\times m},$$

here $\mathcal{N}_1$ is the dimension of $\mathcal{H}_\mathcal{X}$. Hence kernel function $k_\mathcal{X}(x_1,x_2)$ takes form $k_\mathcal{X}(x_1,x_2)=\langle\phi_x(x_1),\phi_x(x_2)\rangle$, $\langle,\rangle$ is an inner product in $\mathcal{H}_\mathcal{X}$. Similarly, ϕ_y maps $Y=(y_1,\ldots,y_m)\in\mathbb{R}^{n_2\times m}$ into $\mathcal{H}_\mathcal{Y}$ by

$$\Phi_y=(\phi_y(y_1),\ldots,\phi_y(y_m))\in\mathbb{R}^{\mathcal{N}_2\times m}.$$

Denote $K_x=\langle\Phi_x,\Phi_x\rangle=(k_\mathcal{X}(x_i,x_j))_{i,j=1}^m$, $K_y=\langle\Phi_y,\Phi_y\rangle=(k_\mathcal{Y}(y_i,y_j))_{i,j=1^m}$. K_x,K_y are called Gram matrices. Let $f=\sum_{i=1}^m\alpha_i\phi_x(x_i)=\Phi_x\alpha$, $g=\sum_{i=1}^m\beta_i\phi_y(y_i)=\Phi_y\beta$, where $\alpha=(\alpha_1,\ldots,\alpha_m)^T$, $\beta=(\beta_1,\ldots,\beta_m)^T\in\mathbb{R}^m$ are called dual vectors. Here without loss of generality, we assume that Φ_x and Φ_y have been centered. Simple calculations show that problem Eq. (1.2) can be rewritten as [18].

$$\begin{aligned}&\max\alpha^TK_xK_y\beta\\&\text{s.t.}\alpha^TK_x^2\alpha=1,\\&\beta^TK_y^2\beta=1.\end{aligned}\tag{2.1}$$

In practice, one would face a large amount of information retrieval tasks. Considering problem (2.1) is not enough, especially in the coming of big data era. Therefore multiple version of kernel CCA are introduced [8]:

$$\begin{aligned}&\max_{W_x,W_y}\quad\mathrm{Trace}(W_x^TK_xK_yW_y)\\&\text{s.t.}\quad W_x^TK_xW_x=I,\\&W_y^TK_yW_y=I,\end{aligned}\tag{2.2}$$

where $W_x=(\alpha_1,\ldots,\alpha_l),W_y=(\beta_1,\ldots,\beta_l)$, $1\le l\le\mathrm{rank}(K_xK_y)$. Theoretical analysis suggests that the solutions of Eq. (2.2) are the singular values of K_xK_y. Firstly, we will characterize a new representation of kernel CCA which was described in [9], for CCA, see [8].

2.2. A new representation of kernel CCA

Define $r_1=\mathrm{rank}(K_x),r_2=\mathrm{rank}(K_y)$ and $r_3=\mathrm{rank}(K_xK_y)$. Let the eigenvalue decomposition of K_x and K_y be

$$K_x=U\begin{pmatrix}D_1&0\\0&0\end{pmatrix}U^T=\begin{pmatrix}U_1&U_2\end{pmatrix}\begin{pmatrix}D_1&0\\0&0\end{pmatrix}\begin{pmatrix}U_1&U_2\end{pmatrix}^T=U_1D_1U_1^T,\tag{2.3}$$

and

$$K_y=V\begin{pmatrix}D_2&0\\0&0\end{pmatrix}V^T=\begin{pmatrix}V_1&V_2\end{pmatrix}\begin{pmatrix}D_2&0\\0&0\end{pmatrix}\begin{pmatrix}V_1&V_2\end{pmatrix}^T=V_1D_2V_1^T,\tag{2.4}$$

respectively, where $U\in\mathbb{R}^{m\times m},U_1\in\mathbb{R}^{m\times r_1},U_2\in\mathbb{R}^{m\times(m-r_1)},D_1\in\mathbb{R}^{r_1\times r_1}$, $V\in\mathbb{R}^{m\times m},V_1\in\mathbb{R}^{m\times r_2},V_2\in\mathbb{R}^{m\times(m-r_2)},D_2\in\mathbb{R}^{r_2\times r_2}$. U,V are orthogonal matrices, D_1,D_2 are non-singular and diagonal matrices. Let the SVD (singular value decomposition) of $U_1^TV_1$ be

$$U_1^TV_1=P_1D_3P_2^T,\tag{2.5}$$

where $P_1\in\mathbb{R}^{r_1\times r_1},P_2\in\mathbb{R}^{r_2\times r_2},D_3$ is a diagonal matrix, $r_3\le\min\{r_1,r_2\}$. Chu et al. [9] stated that

Lemma 1. *Any (W_x,W_y) of the following forms:*

$$\begin{cases}W_x=U_1D_1^{-1}P_1(1:l)+U_2\mathcal{E},\\W_y=V_1D_2^{-1}P_2(1:l)+V_2\mathcal{F},\end{cases}$$

where $P_1(1:l),P_2(1:l)$ are the first $l(1\le l\le r_3)$ columns of P_1 and P_2, respectively. $\mathcal{E}\in\mathbb{R}^{(d_1-r_1)\times l}$ and $\mathcal{F}\in\mathbb{R}^{(d_2-r_2)\times l}$ are arbitrary matrices, is a solution of optimization problem (2.2).

By applying the above lemma, one immediately have a new representation of kernel CCA problem.

$$K_xW_x=U_1P_1(1:l),\quad K_yW_y=V_1P_2(1:l)\tag{2.6}$$

Our aim is to find a solution (not all) that satisfy Eq. (2.6). Therefore without loss of generality, let $W_x^*=U_1D_1^{-1}P_1(1:l)$, $W_y^*=V_1D_2^{-1}P_2(1:l)$, then (W_x^*,W_y^*) is a solution pair that satisfy Eq. (2.6). We will use a gradient descent approach to find it. However, in practice, K_x,K_y are not invertible. Ordinary kernel CCA method fails to detect mutual information between two variables x and y for general kernels. This is the so called over-fitting problem [18]. Taking Gaussian kernel

$$K(s,t)=\exp\left(-\frac{1}{2\sigma^2}\|s-t\|^2\right)$$

as an example. The Gram matrix K_x given by $(K_x)_{ij}=\exp\left(-1/2\sigma^2\|x_i-x_j\|^2\right)$ has full rank provided that the sample points $\{x_i\}_{i=1}^m$ are distinct (similarly for K_y) [24]. Then we have

$$\mathrm{rank}(K_x)=m-1,\quad\mathrm{rank}(K_y)=m-1,$$

after centering. Hence the canonical correlations returned by kernel CCA will be 1 even though they did not have any mutual