Expert Systems With Applications 84 (2017) 171–185

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Multiple-rank supervised canonical correlation analysis for feature extraction, fusion and recognition

Xizhan Gao[a], Quansen Sun[a,*], Haitao Xu[b]

[a] School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China
[b] School of Mathematical Sciences, Liaocheng University, Liaocheng, Shandong, China

ARTICLE INFO

Article history:
Received 20 November 2016
Revised 16 April 2017
Accepted 6 May 2017
Available online 8 May 2017

Keywords:
Two-dimensional canonical correlation analysis (2D-CCA)
Supervised information
Feature extraction
Feature fusion
Pattern recognition

ABSTRACT

The traditional CCA and 2D-CCA algorithms are unsupervised multiple feature extraction methods. Hence, introducing the supervised information of samples into these methods should be able to promote the classification performance. In this paper, a novel method is proposed to carry out the multiple feature extraction for classification, called two-dimensional supervised canonical correlation analysis (2D-SCCA), in which the supervised information is added to the criterion function. Then, by analyzing the relationship between GCCA and 2D-SCCA, another feature extraction method called multiple-rank supervised canonical correlation analysis (MSCCA) is also developed. Different from 2D-SCCA, in MSCCA k pairs left transforms and k pairs right transforms are sought to maximize the correlation. The convergence behavior and computational complexity of the algorithms are analyzed. Experimental results on real-world databases demonstrate the viability of the formulation, they also show that the classification results of our methods are higher than the other's and the computing time is competitive. In this manner, the proposed methods proved to be the competitive multiple feature extraction and classification methods. As such, the two methods may well help to improve image recognition tasks, which are essential in many advanced expert and intelligent systems.

1. Introduction

High-dimensional data can be found in many applications, such as face recognition, image analysis, text classification, web page classification, and so on. Processing high-dimensional data needs a lot of memory and time, moreover, using high-dimensional data directly may lead to the curse of dimensionality. To overcome these problems, feature extraction strategy is used. The major target of feature extraction is to find the meaningful low-dimensional representations of high-dimensional data such that the inherent data structures and relations are revealed (Yuan & Sun, 2014). The famous feature extraction techniques include principal component analysis (PCA) (Truk & Pentland, 1991), linear discriminant analysis (LDA) (Belhumeur, Hespanha, & Kriegman, 1997), locality preserving projections (LPP) (He & Niyogi, 2003), etc.

All above methods extract features from original high-dimensional data rather than multiset data. Hence, they are not suitable for multiple feature extraction. Canonical correlation analysis (CCA) (Hotelling, 1936) can measure the inter-correlation between two variable sets, and recently, it is exploited for image recognition. Sun et al. proposed a new CCA (Sun, Zeng, Liu, Heng, & Xia, 2005) method, which can extract canonical correlation features from two sets of variables. For easy classification, the discriminant version of CCA was further developed, called generalized CCA (GCCA) (Sun, Liu, Heng, & Xia, 2005). The two papers above first extract two groups of low-dimensional feature vectors from two high-dimensional data sets, then fuse them using feature fusion method. There are two famous feature fusion methods: serial feature fusion and parallel feature fusion. Serial feature fusion groups two sets of feature vectors into one union-vector, while parallel feature fusion combines two sets of feature vectors by a complex vector. Since CCA is a linear subspace learning method, it fails to discover the nonlinear relationship. In order to overcome this problem, kernel CCA (KCCA) (Hardoon, Szedmak, & Shawe-Taylor, 2004; Melzer, Reiter, & Bischof, 2003) and locality preserving CCA (LPCCA) (Sun & Chen, 2007) were proposed. KCCA uses the kernel trick to find the nonlinear relation, while LPCCA uses local information to discover the local manifold structure of data. Sparse CCA (Chu, Liao, Ng, & Zhang, 2013) seeks a sparse solution of CCA from a solution subset. In order to decrease the computing time, complete canonical correlation analysis (C3A) (Xing, Wang, yan, & Lv, 2016) first reformulates the traditional CCA, then transforms the

* Corresponding author.
E-mail addresses: gaoxizhan123@126.com (X. Gao), sunquansen@njust.edu.cn (Q. Sun), xuhaitao@live.com (H. Xu).

http://dx.doi.org/10.1016/j.eswa.2017.05.017

singular generalized eigensystem computation of CCA into two stable eigenvalue decomposition problems. In 2005, Francis et al. presented a probabilistic interpretation of CCA (Bach & Jordan, 2005), which enables the use of local CCA models as components of a larger probabilistic model, and suggests generalizations to members of the exponential family other than the Gaussian distribution. E-CCA (Arandjelovic, 2014) is an extension of CCA, which aims to extract the 20 most similar modes of variability within two sets. CCA also has been applied in many fields, for instance, statistical analysis (Anderson, 2003), text mining (Vinokourov, Shawe-Taylor, & Cristianini, 2003), facial expression recognition (Zheng, Zhou, Zou, & Zhao, 2006), genomic data analysis (Yamanishi, Vert, Nakaya, & Kanehisa, 2003), image dehazing (Wang, Xiao, & Wei, 2015), closed-loop data identification (Chou & Verhaegen, 2015) and machine learning (Hardoon et al., 2004; Sun, Ji, & Ye, 2011).

The above-mentioned work on feature extraction are all one-dimensional (1d) based methods, they must reshape image matrices into vectors, but such reshaping might break the spatial structure of the images and increase the computational complexity. To solve these problems, some two-dimensional (2d) based methods have been proposed. Two-dimensional PCA (2DPCA) (Yang, Zhang, Frangi, & Yang, 2004) was proposed by Yang et al. in 2004, in which an image covariance matrix is constructed directly using the original image matrices, and its eigenvectors are derived for image feature extraction. As opposed to CCA, two-dimensional canonical correlation analysis (2D-CCA) (Lee & Choi, 2007; Sun, Ji, Zou, & Zhao, 2010) is based on 2d image matrices rather than 1d vectors so the image matrix does not need to be transformed into a vector prior to feature extraction. Based on the manifold learning method, local 2D-CCA (L2DCCA) (Wang, 2010) uses local information to identify nonlinear correlation between two sets of images. Probabilistic 2D-CCA (P2DCCA) (Afrabandpey, Safayani, & Mirzaei, 2014) is a probabilistic framework of 2D-CCA, which is robust to noise and is able to cope with the missing data problem. 2D-CCA based on pseudoinverse (2DCCAP) (Wu, 2009) uses the pseudoinverse technique to compute the singular matrix. We notice that 2D-CCA, P2DCCA and 2DCCAP only extract two groups of feature, but they don't fuse them. L2DCCA first combines 2d method with feature fusion strategy, and has achieved good results. In essence, 2D-CCA is an unsupervised subspace learning method. From the viewpoint of classification, the supervised information of samples should be used.

Motivated by above and the algorithm GCCA, a new method, called two-dimensional supervised canonical correlation analysis (2D-SCCA), is proposed to carry out the multiple feature extraction for classification, in which the supervised information is added to the criterion function. Inspired by the method multiple rank multi-linear SVM (MRMLSVM) (Hou, Nie, Zhang, Yi, & Wu, 2014) and paper (Gao, Fan, & Xu, 2016), by analyzing the relationship between GCCA and 2D-SCCA, another feature extraction method called multiple-rank supervised canonical correlation analysis (MSCCA) is also developed. Different from 2D-SCCA, in MSCCA k pairs left transforms and k pairs right transforms are sought to maximize the correlation. The detailed process of the algorithms we proposed is described as follows: we firstly extract two feature matrices from the same patterns; then two groups of canonical correlation features are extracted by using 2D-SCCA or MSCCA; finally, two feature fusion strategies are used to fuse the canonical correlation features. Besides, the convergence behavior and computational complexity of the algorithms are also analyzed. Plenty of experiments on different kinds of data sets are presented for illustration.

The rest of this paper is organized as follows. In Section 2, we will give overviews of CCA and 2D-CCA. Section 3 presents the 2D-SCCA method and its relevant theory and algorithm for classification. In Section 4, we first analyze the relation between GCCA and 2D-SCCA, then a new method MSCCA is proposed in details. Convergence analysis and computational complexity are carried out in Section 5. Experiments and results analysis on various kinds of databases are performed in Section 6. Finally, Section 7 provides the conclusion and future work.

2. Background of related work

This section includes brief reviews of CCA and 2D-CCA.

2.1. CCA

Given two sets of random vectors $\{x_i \in \mathbb{R}^{pq\times 1}\}_{i=1}^N$ and $\{y_i \in \mathbb{R}^{mn\times 1}\}_{i=1}^N$. The goal of CCA is to find a pair of projection vectors w_x and w_y, such that the relation between $w_x^T x$ and $w_y^T y$ is maximized. That is the following objective function is maximized

$$\rho = \frac{cov(w_x^T x, w_y^T y)}{\sqrt{var(w_x^T x) var(w_y^T y)}}.$$

Assume $\{\tilde{x}_i\}_{i=1}^N$ and $\{\tilde{y}_i\}_{i=1}^N$ are centered data, denote $C_{xy} = \frac{1}{N}\sum_{i=1}^N \tilde{x}_i \tilde{y}_i^T$, $C_{xx} = \frac{1}{N}\sum_{i=1}^N \tilde{x}_i \tilde{x}_i^T$, $C_{yy} = \frac{1}{N}\sum_{i=1}^N \tilde{y}_i \tilde{y}_i^T$, then the objective function can be simplified as

$$\max \quad \frac{w_x^T C_{xy} w_y}{\sqrt{(w_x^T C_{xx} w_x)(w_y^T C_{yy} w_y)}}. \tag{1}$$

Since the objective function of the optimization problem in (1) is invariant with respect to scaling of w_x and w_y, problem (1) can be reformulated as follows:

$$\begin{aligned} \arg \max_{w_x, w_y} \quad & w_x^T C_{xy} w_y \\ s.t. \quad & w_x^T C_{xx} w_x = 1, \\ & w_y^T C_{yy} w_y = 1. \end{aligned} \tag{2}$$

Using the Lagrange multiplier method, we can get the following generalized eigenvalue equation

$$\begin{bmatrix} 0 & C_{xy} \\ C_{yx} & 0 \end{bmatrix} \begin{bmatrix} w_x \\ w_y \end{bmatrix} = \lambda \begin{bmatrix} C_{xx} & 0 \\ 0 & C_{yy} \end{bmatrix} \begin{bmatrix} w_x \\ w_y \end{bmatrix}.$$

Via solving this problem, we can get the d largest eigenvalues of w_x and w_y.

2.2. 2D-CCA

Now we consider two sets of image data, $\{ X_t \in \mathbb{R}^{p\times q}, t = 1, \cdots, N \}$ and $\{ Y_t \in \mathbb{R}^{m\times n}, t = 1, \cdots, N \}$, which are realizations of random variable matrices X and Y, respectively. The goal of 2D-CCA is to seek the left transform vectors l_x, l_y and the right transform vectors r_x, r_y to maximize the correlation between the projections $l_x^T X r_x$ and $l_y^T Y r_y$. In other words, the objective function to be maximized is given as follows

$$\rho = \frac{cov(l_x^T X r_x, l_y^T Y r_y)}{\sqrt{var(l_x^T X r_x)}\sqrt{var(l_y^T Y r_y)}}. \tag{3}$$

Using the same solving method with CCA, we can get the following two generalized eigenvalue problems, which are used to solve the transform vectors l_x, l_y and r_x, r_y,

$$\begin{bmatrix} 0 & \sum_{xy}^r \\ \sum_{yx}^r & 0 \end{bmatrix} \begin{bmatrix} l_x \\ l_y \end{bmatrix} = \lambda \begin{bmatrix} \sum_{xx}^r & 0 \\ 0 & \sum_{yy}^r \end{bmatrix} \begin{bmatrix} l_x \\ l_y \end{bmatrix}, \tag{4}$$

$$\begin{bmatrix} 0 & \sum_{xy}^l \\ \sum_{yx}^l & 0 \end{bmatrix} \begin{bmatrix} r_x \\ r_y \end{bmatrix} = \lambda \begin{bmatrix} \sum_{xx}^l & 0 \\ 0 & \sum_{yy}^l \end{bmatrix} \begin{bmatrix} r_x \\ r_y \end{bmatrix}, \tag{5}$$

where the definitions of $\sum_{xy}^r, \sum_{yx}^r, \sum_{xx}^r, \sum_{yy}^r, \sum_{xy}^l, \sum_{yx}^l, \sum_{xx}^l, \sum_{yy}^l$ are same as in (Lee & Choi, 2007).