Robust and sparse canonical correlation analysis based $L_{2,p}$-norm

Zhong-rong Shi[1], *Sheng Wang*[2], *Chuan-cai Liu*[1]

[1]*School of Computer and Engineering, Nanjing University of Science and Technology, Nanjing 210094, People's Republic of China*
[2]*Institute of Image Processing and Pattern Recognition, Henan University, Kaifeng 475004, People's Republic of China*
E-mail: shizrong@163.com

Published in *The Journal of Engineering*; Received on 14th October 2016; Accepted on 18th January 2017

Abstract: The objective function of canonical correlation analysis (CCA) is equivalent to minimising an L_2-norm distance of the paired data. Owing to the characteristic of L_2-norm, CCA is highly sensitive to noise and irrelevant features. To alleviate such problem, this study incorporates robust feature extraction and group sparse feature selection into the framework of CCA, and proposes a feature fusion method named robust and sparse CCA (RSCCA). In RSCCA, $L_{2,p}$-norm is adopted as the distance measurement of paired data, which can alleviate the effect of noise and irrelevant features and achieve robust performance. The experimental results show that our method outperforms CCA and its variants for feature fusion.

1 Introduction

In real application, we are involved in dealing with high-dimensional data which needs more storage and more computational time. Lots of approaches are proposed to address such problem, and dimension reduction is one of the simplest one. Most of traditional dimension reduction methods are based on the sum of L_2-norm. Owing to excellent performance for feature fusion, canonical correlation analysis (CCA) has been widely used to extract the discriminant feature. By simple algebraic derivation, we find that objective function of CCA is equivalent to minimising an L_2-norm distance of the paired data [1–4]. In [5], Kwak pointed out that the objective function based on L_2-norm will be prone to outliers, because outliers with large norms dominate the objective function owing to the use of L_2-norm. Furthermore, CCA suffers from the fact that each vector of mapping is a linear combination of all the original variables, thus it cannot select the most discriminant feature and discard these discriminant irrelevant features.

For feature selection, Nie *et al.* [6] proposed a robust feature selection based on $L_{2,1}$-norm. Peng and Fan [7] proposed a feature selection method based on $L_{2,p}$-norm which can get more sparse solution than $L_{2,1}$-norm. For multi-instance learning, Wang *et al.* [8] proposed a novel class specific distance metrics which is based on $L_{2,1}$-norm. Experiments show that the objective based on $L_{2,p}$-norm is more robust to outliers and the regularisation based on $L_{2,p}$-norm also can select the most discriminant features.

In this paper, we propose a method named robust and sparse CCA (RSCCA) based on $L_{2,p}$-norm (RSCCA) which has the following characteristics: (i) RSCCA is much more robust to outlier samples; (ii) RSCCA can select the most discriminant feature; (iii) the performance of RSCCA will be more stable, because the different column vectors of mappings are constrained to be orthogonal.

2 Robust and SCCA

Given N pairs of samples $\{\boldsymbol{x}_i, \boldsymbol{y}_i\}_{i=1}^N$, we assume that the data of two views are both centred. Given a matrix $\boldsymbol{A} \in R^{m\times n}$, the $L_{2,p}$-norm is defined as

$$\|\boldsymbol{A}\|_{2,p} = \left(\sum_{i=1}^{m}\left(\sum_{j=1}^{n}\left|\boldsymbol{A}_{ij}\right|^2\right)^{(p/2)}\right)^{(1/p)} \qquad (1)$$

CCA aims to get a pair of mappings by maximising the correlation of the projected data of two views. By simple algebraic operations, the objective function can be transformed into the following minimising problem [1, 2]

$$\left\{\boldsymbol{W}_x, \boldsymbol{W}_y\right\} = \arg\min_{\boldsymbol{W}_x,\boldsymbol{W}_y} \frac{\sum_{i=1}^N \left\|\boldsymbol{x}_i\boldsymbol{W}_x - \boldsymbol{y}_i\boldsymbol{W}_y\right\|^2}{\sum_{i=1}^N \left\|\boldsymbol{x}_i\boldsymbol{W}_x\right\|^2 + \sum_{i=1}^N \left\|\boldsymbol{y}_i\boldsymbol{W}_y\right\|^2} \qquad (2)$$

From (2), we know that CCA equals to minimise the L_2-norm distance of the paired data. Thus, CCA could be influenced by outlier samples. To alleviate this problem, we adopt $L_{2,p}$-norm to measure the distance of paired data. Then, the objective function can be written as

$$\boldsymbol{W} = \arg\min_{\boldsymbol{W}} \frac{\|\boldsymbol{UW}\|_{2,p}}{\|\boldsymbol{DW}\|_{2,p}} \qquad (3)$$

where $\boldsymbol{U} = \left[\boldsymbol{X}^{\mathrm{T}} \quad -\boldsymbol{Y}^{\mathrm{T}}\right]$, $\boldsymbol{D} = \begin{bmatrix}\boldsymbol{X}^{\mathrm{T}} & 0 \\ 0 & \boldsymbol{Y}^{\mathrm{T}}\end{bmatrix}$ and $\boldsymbol{W} = \begin{bmatrix}\boldsymbol{W}_x \\ \boldsymbol{W}_y\end{bmatrix}$. After getting $\boldsymbol{W}$, we can get $\boldsymbol{W}_x$ and $\boldsymbol{W}_y$ by the following equation

$$\begin{aligned}\boldsymbol{W}_x &= \boldsymbol{W}(1{:}d_x, :); \\ \boldsymbol{W}_y &= \boldsymbol{W}(d_x+1{:}d, :)\end{aligned} \qquad (4)$$

where d_x denotes the dimension of $\boldsymbol{x}_i$ and d denotes the dimension of $\boldsymbol{u}_i$.

As we all know, the vector of mapping of CCA is combination of all features which contains discriminant irrelevant features. To discard these discriminant irrelevant features, we add a regularisation term $R(\boldsymbol{W})$ into the objective function of RSCCA. Many studies show that an $L_{2,p}$-norm could lead to the sparsest solution [7]. Thus, we adopt $L_{2,p}$-norm to regularise $\boldsymbol{W}$. Then, the objective function of RSCCA can be written as

$$\boldsymbol{W} = \arg\min_{\boldsymbol{W}} \frac{\|\boldsymbol{UW}\|_{2,p} + \beta\|\boldsymbol{W}\|_{2,p}}{\|\boldsymbol{DW}\|_{2,p}} = \frac{\|\tilde{\boldsymbol{U}}\boldsymbol{W}\|_{2,p}}{\|\boldsymbol{DW}\|_{2,p}} \qquad (5)$$

where $\tilde{\boldsymbol{U}} = \begin{bmatrix}\boldsymbol{U} \\ \beta\boldsymbol{I}\end{bmatrix}$.

From the definition of $\boldsymbol{U}$, we know that the ith row of $\boldsymbol{UW}$ denotes $\boldsymbol{W}_x^{\mathrm{T}}\boldsymbol{x}_i - \boldsymbol{W}_y^{\mathrm{T}}\boldsymbol{y}_i$ $(i \le N)$. If $i > N$, $\boldsymbol{UW}$ denotes the

$(i-N)$th row of $\boldsymbol{W}$. In RSCCA, we adopt $L_{2,p}$ as the distance measurement. From above, we know that $L_{2,p}$ can get more sparse solution than $L_{2,1}$-norm. From the computation of (3), we know that it is computed by giving a weight for each row. Thus, minimising (3) equals with minimising the weighted paired data of two views or the row of $\boldsymbol{W}$ by which the weight of outliers and the noise feature will be 0. Therefore, the performance of RSCCA gets more robust than CCA and its variants.

As we all know, it is hard to solve the objective function which is based on $L_{2,p}$-norm. In [8], Wang *et al.* proposed an efficient iterative algorithm to solve the objective function based on $L_{2,1}$-norm. Inspired by such algorithm, we propose an algorithm to optimise our objective function. Then, we introduce three intermediate variables into our objective function and it can be transformed into

$$\begin{aligned} &\min_{\boldsymbol{W},\boldsymbol{A},\boldsymbol{B},\boldsymbol{S}} \ \mathrm{tr}\left\{\boldsymbol{W}^{\mathrm{T}}\tilde{\boldsymbol{U}}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{A}\tilde{\boldsymbol{U}}\boldsymbol{W}\right\} \\ &\text{s.t}\ \ \mathrm{tr}\left\{\boldsymbol{W}^{\mathrm{T}}\boldsymbol{D}^{\mathrm{T}}\boldsymbol{B}\boldsymbol{S}\right\} = 1 \end{aligned} \tag{6}$$

$\boldsymbol{A}$ and $\boldsymbol{B}$ denote two diagonal matrixes where their ith diagonal elements are, respectively, defined as

$$\begin{aligned} \boldsymbol{a}_{ii} &= \frac{1}{\|\tilde{\boldsymbol{U}}_i\boldsymbol{W}\|_2^{1-p/2}} \\ \boldsymbol{b}_{ii} &= \frac{1}{\|\boldsymbol{D}^i\boldsymbol{W}\|_2^{1-p/2}} \end{aligned} \tag{7}$$

$\boldsymbol{S}$ is defined as follows

$$\boldsymbol{S} = \left[\frac{\left(\boldsymbol{D}^1\boldsymbol{W}\right)^{\mathrm{T}}}{\|\boldsymbol{D}^{\boldsymbol{1}}\boldsymbol{W}\|_2^{1-p/2}}, \ \frac{\left(\boldsymbol{D}^2\boldsymbol{W}\right)^{\mathrm{T}}}{\|\boldsymbol{D}^2\boldsymbol{W}\|_2^{1-p/2}}, \ \ldots, \ \frac{\left(\boldsymbol{D}^c\boldsymbol{W}\right)^{\mathrm{T}}}{\|\boldsymbol{D}^c\boldsymbol{W}\|_2^{1-p/2}}\right]^{\mathrm{T}} \tag{8}$$

Fixed $\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{S}$ and $\boldsymbol{W}$ can be calculated using Lagrange method. Then, the Lagrangian of (6) is

$$L(\boldsymbol{W}) = \mathrm{tr}\left\{\boldsymbol{W}^{\mathrm{T}}\tilde{\boldsymbol{U}}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{A}\tilde{\boldsymbol{U}}\boldsymbol{W}\right\} - \lambda\left(\mathrm{tr}\left\{\boldsymbol{W}^{\mathrm{T}}\boldsymbol{D}^{\mathrm{T}}\boldsymbol{B}\boldsymbol{S}\right\} - 1\right) \tag{9}$$

$$\frac{\mathrm{d}L}{\mathrm{d}\boldsymbol{W}} = 2\tilde{\boldsymbol{U}}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{A}\tilde{\boldsymbol{U}}\boldsymbol{W} - \lambda\boldsymbol{D}^{\mathrm{T}}\boldsymbol{B}\boldsymbol{S} = 0 \tag{10}$$

Then, W can be got by the following equation

$$\boldsymbol{W} = \left(\tilde{\boldsymbol{U}}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{A}\tilde{\boldsymbol{U}}\right)^{-1}\frac{1}{2}\lambda\boldsymbol{D}^{\mathrm{T}}\boldsymbol{B}\boldsymbol{S} \tag{11}$$

We can calculate λ using the following equation

$$\lambda = \frac{\mathrm{tr}\left\{\boldsymbol{W}^{\mathrm{T}}\tilde{\boldsymbol{U}}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{A}\tilde{\boldsymbol{U}}\boldsymbol{W}\right\}}{\mathrm{tr}\left\{\boldsymbol{W}^{\mathrm{T}}\boldsymbol{D}^{\mathrm{T}}\boldsymbol{B}\boldsymbol{S}\right\}} \tag{12}$$

Fixed $\boldsymbol{W}$, $\boldsymbol{A}$, $\boldsymbol{B}$, λ and $\boldsymbol{S}$ can be easily computed using (7), (8) and (12).

Research show that orthogonal projective system is more robust to noises [9, 10]. Thus, in RSCCA, we constrain that the different columns of mappings are orthogonal. Therefore, the objective function can be written as

$$\begin{aligned} \boldsymbol{W} &= \arg\min_{\boldsymbol{W}} \frac{\|\tilde{\boldsymbol{U}}\boldsymbol{W}\|_{2,p}}{\|\boldsymbol{D}\boldsymbol{W}\|_{2,p}} \\ \text{s.t.} &\quad \boldsymbol{W}^{\mathrm{T}}\boldsymbol{W} = I \end{aligned} \tag{13}$$

Table 1 Recognition rates (%) and their corresponding dimension

Method	CCA	OCCA	SCCA	RSCCA
N = 100	78.82 (21)	87.72 (34)	84.32 (138)	89.08 (29)
N = 200	85.66 (23)	90.54 (44)	88.18 (138)	91.90 (26)
N = 300	87.72 (24)	92.02 (43)	89.58 (137)	93.16 (35)

After getting $\boldsymbol{W}$, we compute the orthogonal mapping $\boldsymbol{W}$ by Gram–Schmidt orthonormalisation. As we all know, the matrix $\tilde{\boldsymbol{W}}$ by Gram–Schmidt orthonormalisation equates to $\tilde{\boldsymbol{W}} = \boldsymbol{W}\boldsymbol{R}$. Thus, we know that $\boldsymbol{W}$ is also sparse after Gram–Schmidt orthonormalisation. Then, $\boldsymbol{W}_x$ and $\boldsymbol{W}_y$ can be got using (4). The whole algorithm of our RSCCA is given in Algorithm 1.

```
Algorithm 1: Framework of RSCCA

Input: The sets of two views $\{\boldsymbol{x}_i, \boldsymbol{y}_i\}_{i=1}^N$;
      The initial mapping $\boldsymbol{W}$, the initial value of objective function $J_0$;
      The maximal iteration times maxIter;
  Output: $\boldsymbol{W}_x$, $\boldsymbol{W}_y$
  1: Construct $\tilde{\boldsymbol{U}}$ and $\boldsymbol{D}$;
  2: for each $i \in [1, \text{maxIter}]$ do
  3:    Compute the value of objective function $\lambda$ using (12);
  4:    if $|J(\boldsymbol{W}) - J_0| < 1e-3\,\&\,(i \neq 1)$ then
  5:       break;
  6:    else
  7:       $J_0 \leftarrow \lambda$;
  8:    end if
  9:    Compute $\boldsymbol{A}$ and $\boldsymbol{B}$ using (7);
  10:    Compute $\boldsymbol{S}$ using (8);
  11:    Compute $\boldsymbol{W}$ using (11);
  12: end for
  13: Compute the orthogonal mapping $\boldsymbol{W}$ by Schmidt orthogonalisation;
  14: Compute $\boldsymbol{W}_x$ and $\boldsymbol{W}_y$ by (4);
  15: return $\boldsymbol{W}_x$, $\boldsymbol{W}_y$.
```

3 Experiments and analyses

During the experiments, we select CENPAMI dataset to evaluate our RSCCA. As we all know, the letter images of CENPAMI dataset are not totally aligned. Therefore, there are parts of samples that are illegible and hardly distinguished in CENPAMI

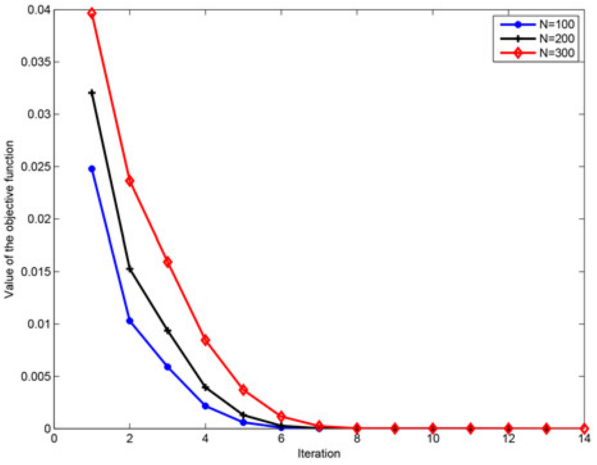

Fig. 1 *Value of the objective function versus the iteration on CENAPRMI dataset*

dataset. To some extent, these samples can be viewed as outliers. Considering the dimension of features, we just select 'GAB' and 'LEG' features. Besides CCA, orthogonal regularised CCA (OCCA) [10] and sparse CCA [11] are also selected for comparison. The parameters of OCCA and sparse CCA are set by the same way of the corresponding refers. $N(N = 100,\ 200,\ 300)$ samples from each class are selected as train samples, the remaining samples for testing ones. During the experiments, we set $p = 0.9$, $\beta = 0.01$. After getting mappings, the data can be embedded into lower-dimensional space. We construct the combined feature as follows (which is a common strategy [10] for CCA): $\boldsymbol{Z} = \begin{bmatrix} \boldsymbol{W}_x^{\mathrm{T}}\boldsymbol{x} \\ \boldsymbol{W}_y^{\mathrm{T}}\boldsymbol{y} \end{bmatrix}$. 1 nearest neighbour (1NN) classifier is employed to classify the combined data. Each experiment is repeated ten times, and Table 1 reports the average recognition rates where the best performances are highlighted in bold.

From this table, our method performs better than the other methods. Thus, we conclude that RSCCA is more robust than the other three methods. In Fig. 1, we draw the values of the objective function during iterations which show that our optimised algorithm can be converged quickly.

4 Conclusion

In this paper, we proposed a method named RSCCA-based $L_{2,p}$-norm for feature fusion. Furthermore, we propose an algorithm to optimise our objective function. Experimental results show RSCCA has better performance than CCA and its variants.

5 Acknowledgment

This work was supported by the National Natural Science Foundation of China under grant 61373063.

6 References

[1] Melzer T., Reiter M., Bischof H.: 'Appearance models based on kernel canonical correlation analysis', *Pattern Recognit..*, 2003, **36**, (9), pp. 1961–1971

[2] De la Torre F.: 'A least-squares framework for component analysis', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2012, **34**, (6), pp. 1041–1055

[3] Yuan Y.H., Sun Q.S., Ge H.W.: 'Fractional-order embedding canonical correlation analysis and its applications to multi-view dimensionality reduction and recognition', *Pattern Recognit.*, 2014, **473**, pp. 1411–1424

[4] Yuan Y.H., Sun Q.S., Zhou Q., *et al.*: 'A novel multiset integrated canonical correlation analysis framework and its application in feature fusion', *Pattern Recognit.*, 2011, **44**, (5), pp. 1031–1040

[5] Kwak N.: 'Principal component analysis based on L_1-norm maximization', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2008, **30**, (9), pp. 1672–1680

[6] Nie F., Huang H., Cai X., *et al.*: 'Efficient and robust feature selection via joint $L_{2,\ 1}$-norms minimization', Advances in Neural Information Processing Systems 23 (NIPS 2010), Vancouver, Canada, December 2010, pp. 1813–1821

[7] Peng H., Fan Y.: 'Direct l(2, p)-norm learning for feature selection', arXiv preprint arXiv:1504.00430, 2015

[8] Wang H., Nie F., Huang H.: 'Robust and discriminative distance for multi-instance learning'. 2012 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2012, pp. 2919–2924

[9] Cai D., He X., Han J., *et al.*: 'Orthogonal Laplacian faces for face recognition', *IEEE Trans. Image Process.*, 2006, **15**, (11), pp. 3608–3614

[10] Hou S., Sun Q.: 'An orthogonal regularized CCA learning algorithm for feature fusion', *J. Vis. Commun. Image Represent.*, 2014, **25**, (5), pp. 785–792

[11] Sun L., Ji S., Ye J.: 'A least squares formulation for canonical correlation analysis'. Proc. of the 25th Int. Conf. on Machine Learning, 2008, pp. 1024–1031