A Local Method for Canonical Correlation Analysis

Tengju Ye(✉), Zhipeng Xie, and Ang Li

School of Computer Science, Fudan University, Shanghai, China
{yet13,xiezp,angli12}@fudan.edu.cn

Abstract. Canonical Correlation Analysis (CCA) is a standard statistical technique for finding linear projections of two arbitrary vectors that are maximally correlated. In complex situations, the linearity of CCA is not applicable. In this paper, we propose a novel local method for CCA to handle the non-linear situations.We aim to find a series of local linear projections instead of a single globe one. We evaluate the performance of our method and CCA on two real-world datasets. Our experiments show that local method outperforms original CCA in several realistic cross-modal multimedia retrieval tasks.

Keywords: Local linearity · Multivariate analysis · Cross-modal multimedia retrieval

1 Introduction

Canonical correlation analysis (CCA) is a statistical method of correlating linear relationships between two parts of multidimensional variables [11,17]. CCA can be regarded as the problem of finding two basis directions onto which the correlation between the projections of two variables is maximized. CCA is broadly used in unsupervised analysis since it does not require labeled data. The applications are therefore cross various areas, including natural language processing [9,18], neuronal data analysis [6], computer vision [20] and cross modal multimedia retrieval [8,15]. However, because of its linearity, when strong nonlinear relation occurs, CCA is often not applicable. Several methods were thus proposed to find nonlinear projections. However, most of the state-of-the-art non-linear improvements still aim to maximize the correlation of two variables in a single uniform projection.

In this paper, we propose a local linear model for CCA. Unlike the methods which aim to find uniform projections, we consider to construct several local projections, each of them maximizes correlation in a particular region of the dataset. The local projection relaxes the global linearity objective of CCA. In order to construct local CCA projection, we make use of the techniques from non-parametric kernel smoothing. The final correlation between two variables is smoothed combination of local correlation. Our results show that the method based on local linear projection outperforms the standard CCA in various real world information retrieval tasks.

2 Related Work

Rasiwasia [15] used canonical correlation analysis to solve the cross-modal multimedia retrieval problem. Instead of classical text-based information retrieving, Rasiwasia

J. Li et al. (Eds.): NLPCC 2015, LNAI 9362, pp. 428–435, 2015.
DOI: 10.1007/978-3-319-25207-0_39

made use of the rich multiple modalities information. Take text-image cross retrieving for example, they used Latent Dirichlet Allocation (LDA) method to process text corpus into a group of vectors and use SIFT method to extract images' features into another group of vectors. By applying canonical correlation analysis on these two kinds of vectors, it is able to retrieve related images from a text query or retrieve related texts from a given image.

Improvement of canonical correlation analysis for handling non-linear projections has been researched intensively. Kernel Canonical Correlation Analysis [4] maximizes correlation in higher dimension space with help of kernel trick. Neuronal networks are also introduced for solving the linearity drawback of CCA [2,19].

Another part of related work is the local models. Local principal component analysis by Kambhatla[13] proposed local models for PCA. Kambhatla's work partition the train data into disjoint regions and within each of which they construct linear models by PCA. Local Linear Embedding(LLE)[16] provides local linear model for dimensionality reduction. Local methods is also common in non-parametric analysis. Lee, Joonseok, et al.[14] proposed a local low rank model for matrix completion.

2.1 Background

Canonical Correlation Analysis(CCA)

Consider a pair of training vectors (we call they are in different views), $(\mathrm{X},\mathrm{Y}) \in \mathbb{R}^{\mathrm{m}\times \mathrm{r}_1} \times \mathbb{R}^{\mathrm{m}\times \mathrm{r}_2}$ with corresponding covariance pair $(\Sigma_{11}, \Sigma_{22})$. Let Σ_{12} denote the cross-covariance of(X,Y). CCA aims to find a pair of directions (a,b) (called canonical components) on which the vectors' projection is maximally correlated, i.e.

$$(a,b) = \arg\max_{a,b} \frac{\mathrm{cov}(a^T X, b^T Y)}{\sigma_{a^T X}\sigma_{b^T Y}} = \arg\max_{a,b} \frac{a^T \Sigma_{12} b}{\sqrt{a^T \Sigma_{11} a}\sqrt{b^T \Sigma_{22} b}} \tag{1}$$

Because the choice of re-scaling is arbitrary, the optimization of Eq. (1) is equivalent to maximizing the numerator subject to

$$a^T \Sigma_{11} a = 1, b^T \Sigma_{22} b = 1 \tag{2}$$

The optimization thus is transformed into

$$(a,b) = \mathrm{argmax}_{a^T \Sigma_{11} a=1, b^T \Sigma_{22} b=1}\, a^T \Sigma_{12} b \tag{3}$$

For simplicity, we denote $u = a^T X, v = b^T Y$. Then by constructing the corresponding Lagrangian we get,

$$\mathcal{L} = \sum_{i=1}^{N} a^T (x_i - \bar{u})(y_i - \bar{v}) b - \frac{\lambda}{2}(a^T \Sigma_{11} a) - \frac{\theta}{2}(b^T \Sigma_{22} b) \tag{4}$$

Taking derivatives to a, b, we obtain the following equations,

$$\frac{\partial \mathcal{L}}{\partial a} = \sum_{i=1}^{N} (x_i - \overline{u})^T (y_i - \bar{v}) b - \lambda \Sigma_{11} a = 0 \tag{5}$$

$$\frac{\partial \mathcal{L}}{\partial b} = \sum_{i=1}^{N} (y_i - \overline{v})^T (x_i - \overline{u}) a - \lambda \Sigma_{22} b = 0 \tag{6}$$

By using a^T times the Eq.(5) and using b^T times Eq.(6) and accompanied the constraint Eq.(2) we obtain,

$$\lambda = \theta = a^T \Sigma_{12}^W b \tag{7}$$

Obviously, the maximal λ is in fact the maximal correlation. By simplifying Eq.(7) and assuming $\Sigma_{11}$and Σ_{22} is invertible we obtain,

$$\Sigma_{11}^{-1} \Sigma_{12} b = \lambda a \tag{8}$$

$$\Sigma_{21}^{-1} \Sigma_{22} b = \lambda b \tag{9}$$

We then transform the Eq. (8) and Eq.(9) into matrix format,

$$\begin{pmatrix} \Sigma_{11}^{-1} & 0 \\ 0 & \Sigma_{22}^{-1} \end{pmatrix} \begin{pmatrix} 0 & \Sigma_{12} \\ \Sigma_{21} & 0 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \lambda \begin{pmatrix} a \\ b \end{pmatrix} \tag{10}$$

Let B denotes$\begin{pmatrix} \Sigma_{11} & 0 \\ 0 & \Sigma_{22} \end{pmatrix}$, A denotes$\begin{pmatrix} 0 & \Sigma_{12} \\ \Sigma_{21} & 0 \end{pmatrix}$ and r denotes$\begin{pmatrix} a \\ b \end{pmatrix}$, finally, we transform Eq. (9) into the following form,

$$B^{-1} A r = \lambda r \tag{11}$$

The problem left in Eq. (11) is a generalized eigenvalue problem.

Local Approach and Kernel Methods

In non-parametric statistics, kernel methods can be used to specify the local neighborhood by assigning weights to the points around a given point [10]. Let $K_h(p_0, p)$ denote a kernel function, h is the bandwidth parameter h > 0, p_0 is a query point. A large value of h implies that$K_h(p_0,)$ spread widely while a small value means it spread narrowly. The kernel function will calculate weight for each point within the neighborhood. Points within the local neighborhood will be assigned weights by the kernel function. The following is three popular kernels [14], uniform kernel, triangular kernel and Epanechnikov kernel.

$$K_h(p_0, p) \coloneqq \mathbf{1}[d(p_0, p) < h]$$

$$K_h(p_0, p) \coloneqq \left(1 - h^{-1} d(p_0, p)\right) \mathbf{1}[d(p_0, p) < h]$$

$$K_h(p_0, p) \coloneqq (1 - d(p_0, p)^2) \mathbf{1}[d(p_0, p) < h]$$

After the neighborhood is determined, result is calculated by Nadaraya–Watson [10] kernel weighted average.

$$\hat{f}(p_0) = \sum_i^N \frac{K_h(p_0, p_i) p_i}{\sum_j^N K_h(p_0, p_j)} \tag{12}$$

3 Local Linear Model

Most of the real-world data do not contain strong linearity globally. In order to handle non-linear situation. Our local model focuses on maximizing the correlation in a particular local region. Therefore, we aim to find a strategy to construct such a series of local projections and calculate distance between query vectors to the candidate vectors based on these local projections.

- **Weighted CCA:** In order to make use of the weights in the local region, we change the standard CCA to a new weighted form. Similarly ,we start with defining the pair of two projected vectors by $u = a^T X, v = b^T Y$ with the corresponding weighted averages:

$$\bar{u}^w = \frac{\sum_i w_i X_i}{\sum_i w_i}, \bar{v}^w = \frac{\sum_i w_i Y_i}{\sum_i w_i} \tag{13}$$

  For simplicity, we normalized the weights, let $\sum_i w_i = 1$. The weighted variance and the weighted covariance are,

$$var^w(u) = \sum_i^N w_i (a^T X - a^T \bar{u}^w)^2 = a^T \sum_i^N w_i (X_i - \bar{u}^w)^2 \, a = a^T \Sigma_{11}^w a \tag{14}$$

$$var^w(v) = b^T \Sigma_{22}^w b \tag{15}$$

$$cov^w(u, v) = \frac{1}{\sum w} a^T \sum_{i=1}^N w_i (X_i - \bar{u}^w)(Y_i - \bar{v}^w) b = a^T \Sigma_{12}^w b \tag{16}$$

  We now aim to maximize the weighted correlation in each region,

$$(a, b) = \text{argmax}_{a,b} \frac{a^T \Sigma_{12}^w b}{\sqrt{a^T \Sigma_{11}^w a} \sqrt{b^T \Sigma_{22}^w b}} \tag{17}$$

  The later derivation is similar as the standard CCA, we maximize the numerator, by constraining the weighted variances into unit ones.

$$var^w(u) = 1, var^w(v) = 1 \tag{18}$$

  The remaining steps are as same as standard CCA that we mentioned in Section 2.1 and we find then left a generalized eigenvalue problem again.

- **Local Method by Anchor Pairs:** In order to construct local projections, we firstly select q anchor pairs. Each anchor pair has its own region which is specified by the kernel function. Pairs in each region are also assigned weights by kernel function according to the distance to the anchors. Moreover, for each region, we apply the aforementioned weighted CCA to obtain a local projection. Fig.1.illustrates this idea.

- **Combining Strategy:** By local method with anchor pairs, suppose we have already selected q anchor pairs which are denoted by $(e_1, f_1) \ldots (e_q, f_q), e \in \mathbb{R}^{m \times r_1}, f \in \mathbb{R}^{m \times r_2}$, we now obtain q local canonical components$< a_1, b_1 > \cdots < a_q, b_q >$. Evaluation state of CCA is often the problem of retrieving top k nearest-data in different view. We denote $Q_1 \in \mathbb{R}^{r_1}$ as a query vector from view one.

The problem of calculating correlation from Q_1 to the vectors in $\mathbb{R}^{r_2}$ from q local projections is known as non-parametric regression. We propose using the aforementioned Nadaraya-Waston regression. Let $D(Q_1)$denotes the final distance between the query vectors Q_1 to all other candidate vectors in$\mathbb{R}^{r_2}$. Let $\widehat{D}^i(Q_1)$ denotes the distance from Q_1 to all other candidate vectors which are calculated in the i_{th} local projection. Therefore, with Nadaraya-Waston regression, we obtain the final distance by Eq. (19).

$$D(Q_1) = \sum_i^q \frac{K_h(e_i,Q_1)}{\sum_j^q K_h(e_j,Q_1)} \widehat{D}^i(Q_1) \tag{19}$$

- **Local Canonical Correlation Analysis(LCCA):** We denote the i_{th} pair of data in the training set using (X_i, Y_i) in which$X_i$contains $r_1$dimensions and$Y_i$contains r_2 dimensions. Let h_1 represent the kernel width for view$X$ and $h_2$represent the kernel width for view Y. An anchor pair is denoted by(e, f). $K_{h_1}$is a vector that stores the values from the kernel function with $h_1$kernel width and K_{h_2} has the similarly meaning. We use a vector $w \in \mathbb{R}$ to store weights for training pairs. Distance function is denoted as d ().$< a_t, b_t >$ is the canonical components maximize the correlation in the local region around the t_{th}anchor pair. Algorithm 2-1 describes the training state of LCCA. Obviously, during the training state, the q local region are independent. Therefore, the q iteration can be computed parallel which great increases the algorithm speed. In the predicating state , we calculate the distances between the query vector with candidate vectors in each local spaces thus we get q distance vectors. With Eq. (19). , we combine the local distance vectors into the final distance vectors. Then we select the top k nearest candidate as the results for the query.

Algorithm 2-1 LCCA training state

Input: $X \in \mathbb{R}^{n \times r_1}, Y \in \mathbb{R}^{n \times r_2}, h_1, h_2, q$

for all $t = 1 \dots q$in parallel do

$(e_t, f_t) :=$ randomly selected traiing pair

for $i = 1 \to n$do

$$K_{h_1}^{e_t}[i] := (1 - d(e_t, X_i)^2)\mathbf{1}_{d(e_t,X_i)<h_1}$$

$$K_{h_2}^{f_t}[i] := (1 - d(f_t, Y_i)^2)\mathbf{1}_{d(f_t,Y_i)<h_2}$$

$$w_i = K_{h_1}^{e_t}[i]\, K_{h_2}^{f_t}[i]$$

end for

$$(a_t, b_t) = \underset{a_t,b_t}{\operatorname{argmax}} \frac{a_t^T \Sigma_{12}^w b_t}{\sqrt{a_t^T \Sigma_{11}^w a_t}\sqrt{b_t^T \Sigma_{22}^w b_t}}$$

end for

Output:$< a_t, b_t >, t = 1 \dots q$

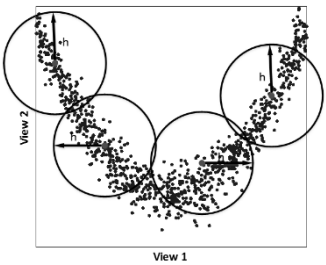

Fig. 1. Local Method by Anchor Pairs

4 Experiment

In this section, we perform experiments on two real world datasets to illustrate the performance of our algorithm. The task we set is cross modal information retrieval

task. Cross-modal information retrieval can match queries from one modality to database entries from another modality. Each dataset contains training set and testing set, either set consists of paired vectors which are categorized into several classes. Based on LCCA and CCA, we measure the distance between a query of one modality to candidate result of the other modality. The performances are measured with mean average precision (MAP) which is widely used in information retrieval task. Since the Norm Correlation distance metric achieve the best performance in the experiments of Rasiwasiaet al. [15], we use this metric as our distance function in our later experiments.The kernel we used is the Epanechnikov kernel which is mentioned above. In each of the following dataset, we also compare the influence of different region sizes for LCCA.

4.1 Datasets

- "Wikipedia" is a dataset assembled from Wikipedia's "featured articles" by Rasiwasia et al. [15]. It contains 2866 documents which are random split into a training set with 2173 documents and a test set with 600 documents. The documents are categorized into 10 categories. Each text is represented as a topic histogram over 10 topics by LDA topic model, while each image is represented by a SIFT codebook of 128 codewords.
- "Chinese Web Portal" is collected by ourselves from several popular web portal sites in China. It contains 7033 web documents of paired texts and images, which belong to 11 categories. These documents are randomly divided into two parts: 70% for training set, and 30% for testing set. We use a popular Chinese segment tool IKAnalyzer[1] to separate the documents into "bags of words" and use LDA to extract 30 latent features from each text. Images are also represented as SIFT histograms by a SIFT codebook of 128 codewords.

4.2 Results and Analysis

Fig. 2. graphs the MAP performances achieved by LCCA and CCA. We set LCCA with two different local region size for each experiment. In "Wikipedia" dataset, we set the local region size to 2000 and 2100 for both image query evaluation and text query evaluation. The main tendency of LCCA grows with the increasing number of anchor points. Either in image queries or in text queries, LCCA always outperforms the CCA when there are more than ten anchors. Similar results appear in the "Chinese Web Portal" datasets, the main tendency of MAP scores increases with the increasing number of anchors.

[1] From http://code. google. com/p/ik-analyzer

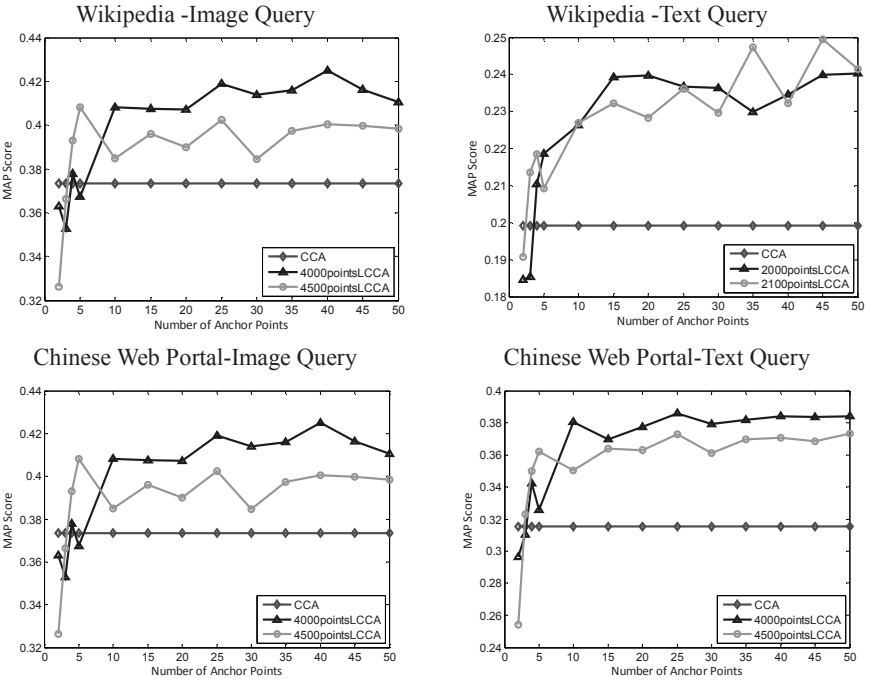

Fig. 2. MAP scores against the number of anchor points in different datasets.

5 Conclusions

We presented a novel local approach for canonical correlation analysis. Our proposed algorithm is called Local Canonical Correlation Analysis (LCCA) which can be easily implemented for parallel computing. The performance is evaluated in two different real world datasets. Our experiments indicate that LCCA outperforms the standard CCA in the cross modal information retrieval task. We also analyze LCCA's performance in terms of its locality (required training points in each local region) and number of required anchor points. Since the basic idea of LCCA is construct local regions and apply weighted CCA in the local region, our future work is plan to investigate the performance by applying other existing non-linear CCA(e.g. kernel CCA) in the local region.

Acknowledgements. This work is supported by National High-tech R&D Program of China (863 Program) (No. SS2015AA011809), Science and Technology Commission of Shanghai Municipality (No. 14511106802), and National Natural Science Foundation of China (No. 61170007). We are grateful to the anonymous reviewers for their valuable comments.

References

1. Akaho, S.: A kernel method for canonical correlation analysis (2006). arXiv preprint cs/0609071
2. Andrew, G., Arora, R., Bilmes, J., Livescu, K.: Deep canonical correlation analysis. In: Proceedings of the 30th International Conference on Machine Learning, pp. 1247–1255 (2013)
3. Asoh, H., Takechi, O.: An approximation of nonlinear canonical correlation analysis by multilayer perceptrons. In: ICANN 1994, pp. 713–716. Springer (1994)
4. Bach, F.R., Jordan, M.I.: Kernel independent component analysis. The Journal of Machine Learning Research **3**, 1–48 (2003)
5. Barnard, K., Duygulu, P., Forsyth, D., De Freitas, N., Blei, D.M., Jordan, M.I.: Matching words and pictures. The Journal of Machine Learning Research **3**, 1107–1135 (2003)
6. Bießmann, F., Meinecke, F.C., Gretton, A., Rauch, A., Rainer, G., Logothetis, N.K., Müller, K.R.: Temporal kernel CCA and its application in multimodal neuronal data analysis. Machine Learning **79**(1–2), 5–27 (2010)
7. Bishop, C.M.: Pattern recognition and machine learning. Springer (2006)
8. Costa Pereira, J., Coviello, E., Doyle, G., Rasiwasia, N., Lanckriet, G.R., Levy, R., Vasconcelos, N.: On the role of correlation and abstraction in cross-modal multimedia retrieval. IEEE Transactions on Pattern Analysis and Machine Intelligence **36**(3), 521–535 (2014)
9. Dhillon, P., Foster, D.P., Ungar, L.H.: Multi-view learning of word embeddings via CCA. In: Advances in Neural Information Processing Systems, pp. 199–207 (2011)
10. Friedman, J., Hastie, T., Tibshirani, R.: The elements of statistical learning. Springer series in statistics, vol. 1. Springer, Berlin (2001)
11. Hardoon, D.R., Szedmak, S., Shawe-Taylor, J.: Canonical correlation analysis: An overview with application to learning methods. Neural Computation **16**(12), 2639–2664 (2004)
12. Hsieh, W.W.: Nonlinear canonical correlation analysis by neural networks. Neural Networks **13**(10), 1095–1105 (2000)
13. Kambhatla, N., Leen, T.K.: Dimension reduction by local principal component analysis. Neural Computation **9**(7), 1493–1516 (1997)
14. Lee, J., Kim, S., Lebanon, G., Singer, Y.: Local low-rank matrix approximation. In: Proceedings of the 30th International Conference on Machine Learning, pp. 82–90 (2013)
15. Rasiwasia, N., Costa Pereira, J., Coviello, E., Doyle, G., Lanckriet, G.R., Levy, R., Vasconcelos, N.: A new approach to cross-modal multimedia retrieval. In: Proceedings of the International Conference on Multimedia, pp. 251–260. ACM (2010)
16. Roweis, S.T., Saul, L.K.: Nonlinear dimensionality reduction by locally linear embedding. Science **290**(5500), 2323–2326 (2000)
17. Thompson, B.: Canonical correlation analysis. Encyclopedia of Statistics in Behavioral Science (2005)
18. Vinokourov, A., Cristianini, N., Shawe-Taylor, J.S.: Inferring a semantic representation of text via cross-language correlation analysis. In: Advances in Neural Information Processing Systems, pp. 1473–1480 (2002)
19. Wand, M.P., Jones, M.C.: Kernel smoothing. CRC Press (1994)
20. Zheng, N., Loizou, G., Jiang, X., Lan, X., Li, X.: Computer vision and pattern recognition (2007)