ATLANTIS PRESS

Advances in Computer Science Research, volume 59

7th International Conference on Education, Management, Computer and Medicine (EMCM 2016)

Multi-kernel Partial Least Squares for Multi-Modal Data Analysis

Ping Wang[1, 2, a] and Hong Zhang[1, 2, b]

[1]College of Computer Science and Technology, Wuhan University of Science and Technology

[2]Hubei Province Key Laboratory of Intelligent Information Processing and Real-time Industrial System

[a]491791519@qq.com, [b]2514384194@qq.com

Keywords: Partial least squares regression; Multi-kernel learning; Multi-modal classification; Multi-modal retrieval; Canonical correlation analysis

Abstract. In recent years, multi-modal data analysis has enjoyed an increasing attention. Multi-modal data mean the different modal data representing the same semantics. Moreover, many subspace learning methods are proposed to measure the correlation between different modal data. As the most representative subspace learning method, canonical correlation analysis (CCA) and its variants project different modal data into a common space where the Pearson correlation is maximized. Yet CCA often causes information loss when switching the modals, and as a result, the partial least squares regression (PLSR) model is adopted to handle the problem. Subsequently, considering the nonlinearity of data, the kernel partial least squares regression (KPLSR) is proposed. Besides, KPLSR mostly relies on the kernel parameters. Hence, we propose to apply multi-kernel partial least squares regression (MKPLSR) for multi-modal data analysis. To evaluate the proposed approach, massive experiments are carried out. Compared with previous methods, the experimental results on two benchmark datasets composed of images and texts pairs, show the effectiveness of our approach, when applied to multi-modal data retrieval and classification.

Introduction

Over the past decade, massive explosion of multimedia data have emerged on the Internet such as image, text, audio and video. Yet most multimedia data processing methods focus on the uni-modal data [1-3]. However, in today's information field, the data are usually comprised of different modalities describing the same events or topics. For example, in a search article, there are not only the text data, but also the image data for illustrating the same topic. Such different types of data are so-called multi-modal data, which have a characteristic of heterogeneity. Consequently, there are increasing needs to process multi-modal data. Recently, cross-modal retrieval and multi-modal classification have drawn considerable researchers' attention [4-8].

The key issue for multi-modal data analysis is how to measure the correlations between different modalities to facilitate heterogeneous modalities [6]. As a representative solution, canonical correlation analysis (CCA) builds the relationship between modalities by projecting different modal data into a common space with maximum correlation [9]. However, in this way, some important information may be lost and the matching precision will accordingly be affected because of rough processing. Although some variants of the CCA are later proposed, this problem cannot be avoided. As a result, partial least squares regression (PLSR) is proposed to handle the problem, which combines CCA, principal component analysis (PCA) and multiple linear regression [10]. Subsequently, in order to better deal with non-linear data, the multi-kernel partial least squares regression (MKPLSR) is proposed [11].

Recently, a continuum regression (CR) model is applied for modal switch, which unifies those previous linear regression expressions into a simplified mathematic form in terms of regularization of the variances and shrinkage properties [10]. Costa Pereira et al. [12] proposed a method called semantic correlation matching (SCM) which first projected image space and text space into a common space by CCA. Then it applied logistic regression in the common space to compute their posterior probability distributions and finally cross-modal retrieval was completed by means of similarity measurement

methods. Inspired by the above methods, we apply MKPLSR for cross-modal retrieval and multi-modal data classification, and we name it multi-kernel original matching (MKOM). Our current study just involves two different-modal media data, i.e., text and image. On two benchmark multimedia datasets, encouraging experimental results demonstrate the superiority and effectiveness of the proposed method over most existing algorithms.

The rest of this paper is organized as follows. In Section 2, we describe explicitly the improved modal switch algorithm. Then, in Section 3, we present the comparative results in multi-modal data retrieval and classification on two benchmark multimedia datasets. Finally, Section 4 concludes the paper.

Improved Modal Switch Methods

In this section, we introduce a novel approach of MKPLSR for modal switch. Let $I \in R^{d_i}$ and $T \in T^{d_t}$ be image and text matrices respectively, where d_i and d_t are the corresponding dimensionalities of image and text, the number of which is n. Because orders of magnitude are apparently different between features, we need first to center the data matrices I and T.

Using PLSR method, the two data matrices can be correlated by

$$T = IP + E \tag{1}$$

Where $P \in R^{d_i \times d_t}$ is the regression coefficient matrix and $E \in R^{d_t}$ is the residual matrix. The aim of the PLSR model is to find the maximum correlation between subspace U and V, which is implemented by

$$U = VQ + F \tag{2}$$

Where $Q \in R^{m \times m}$ denotes the subspace regression coefficient matrix and $F \in R^{m \times n}$ is the subspace residual matrix. As a result, Eq. 1 can be rewritten as

$$T = II^T U(V^T II^T U)^{-1} V^T T + E \tag{3}$$

At this moment, P becomes the regression coefficient matrix of original space [10].

In KPLSR method, the two data matrices first are mapped into a high dimension space with a nolinear transformation, and consequently the kernel matrix $K(I, I^T)$ will be obtained. Then the procedures are similar with PLSR. The final original space is

$$T = KU(V^T KU)^{-1} V^T T + E \tag{4}$$

The nature of linear multi-kernel learning is a combination of different kernels. In the paper, the radial basis function (RBF) kernel and polynomial kernel are adopted, which is formulated as

$$\begin{aligned} &K = \lambda K_{poly} + (1-\lambda) K_{rbf} \\ &K_{poly}(X, X^T) = (XX^T + 1)^d, d = 1, 2, ..., n \\ &K_{rbf}(X, X^T) = \exp(\frac{-\| X - X^T \|}{2\sigma^2}) \end{aligned} \tag{5}$$

The adaptive genetic algorithm (AGA) is applied to select parameters $\{d, \lambda, \sigma\}$, which can adjust probabilities of crossover and mutation according to fitness function automatically [13]. The AGA algorithm aims to minimize error between the output and real value of KPLSR by parameters selection. The fitness function is as follow:

$$Fitness(\lambda, d, \sigma) = \sqrt{\frac{1}{n}\sum_{i=1}^{n} \| y_i - \hat{y}_i \|^2} \tag{6}$$

Where n is the number of iteration and $\hat{y}$ is the prediction value. The parameters selection demonstrates both effectiveness and flexibility of MKOM method for multi-modal data analysis.

Experiments

In this section, we perform massive experiments to certify the effectiveness of the proposed approach in cross-modal retrieval and multi-modal classification, and compare our approach with most existing methods.

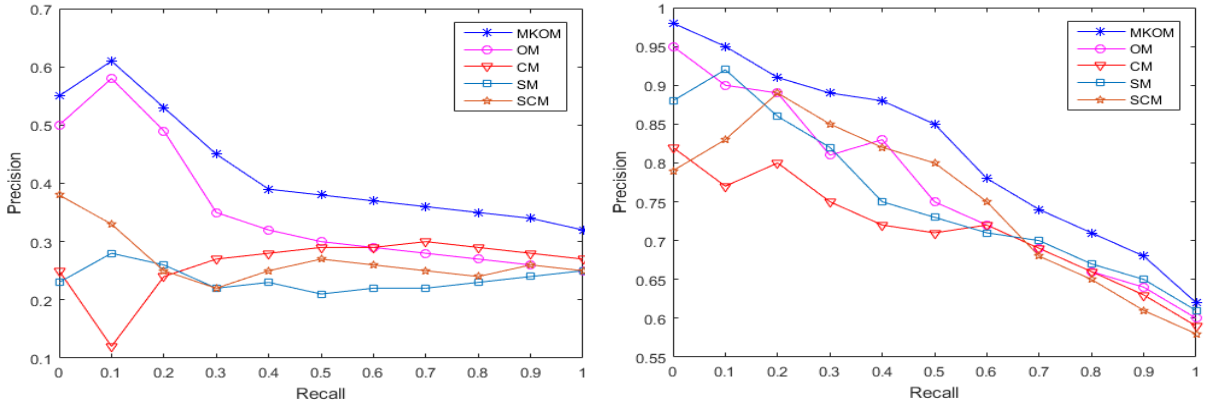
Evaluation Datasets. To evaluate our method, Wiki and Pascal datasets [8] are adopted in the experiments, where Wiki is a popular dataset extracted from " Wikipedia featured articles", in which one image is surrounded by at least 70 words. We choose 2866 pairs of image and text with labels from 10 categories, where 2173 are used for training and 693 for testing. As for Pascal, 700 for training and 300 for testing. In addition, latent Dirichlet allocation (LDA) and scale-invariant feature transform (SIFT) features are extracted for representing text and image, respectively.

Image-Text Retrieval. Image-text retrieval indicates that retrieving text when given image and retrieving image when given text. Precision-recall (PR) curve and mean average precision (MAP) are used to evaluate retrieval performances. We compare our method with previous representative cross-modal retrieval methods. SCM is the combination of CM and SM, where the CCA modeling is first applied to learn two maximally correlated subspace, and then logistic regression is used in each of these subspace. Similarly, OM first applies PLSR to switch image features into text features, and then learns abstract semantics by logistic regression.

Table 1 Cross-modal MAP scores on WIKI dataset

Matching methods	Image query	Text query	Average
MKOM	**0.427**	**0.845**	**0.636**
OM	0.334	0.782	0.588
CM	0.265	0.711	0.488
SM	0.235	0.743	0.489
SCM	0.253	0.673	0.463

Table 1 shows the MAP scores achieved by CM, SM, SCM [12] OM [10] and our MKOM on Wiki dataset. The best results are in boldface. Due to the advantages of the feature, the text query has apparent superiority over the image query. It's obvious that our method achieves promising performances on cross-modal retrieval, which demonstrates the significance of use of multi-kernel learning and non-linearity of data. Therefore, our method can better achieve modal switch and learn abstract semantics, and thereby obtain excellent performances on cross-modal retrieval.



(a) PR curves for image query

(b) PR curves for text query

Figure 1. Precision-Recall curves of cross-modal retrieval using image and text queries

Fig. 1 shows the PR curves on Wiki dataset. It can be seen that the proposed method consistently outperforms the other methods across all views. There's no doubt that our method improves the modal switch and then boosts the retrieval precision. It exhibits the benefits of multi-kernel learning and non-linearity of data.

Classification. All of the above methods are originally used for cross-modal retrieval. Hence, to solve the classification problem, a classifier is learned in the semantic space. In our experiments, the k-nearest neighbor (KNN) classifier is applied to find the closest semantics to the test data. For CCA and kernel canonical correlation analysis (KCCA), we first use them to project image features and text features into a sharing space. Then support vector machine (SVM) is adopted for classification in the common space. Here, Wiki and Pascal datasets are adopted for multi-modal classification.

Table 2 describes the average classification accuracy in both the Pascal and Wiki datasets. We can see that our method has clear advantages over other methods. The results show that our approach can effectively enhance modal switch and accordingly promote multi-modal classification.

Table 2 Average classification accuracy on Wiki dataset.

Datasets	Methods					
/	CCA	KCCA	SM	SCM	OM	MKOM
Pascal	0.421	0.458	0.464	0.482	0.526	0.578
Wiki	0.453	0.492	0.508	0.527	0.559	0.621

Conclusion

In this paper, we proposed to apply MKPLSR model for cross-modal retrieval and multi-modal classification. Extensive comparing experiments on two benchmark datasets validate that MKPLSR is a great model for multimodal data analysis. In the future, we will do some research on dictionary learning and deep correlation learning for multi-modal data analysis.

Acknowledgments

This work is supported by National Natural Science Foundation of China (No.61373109, No.61003127).

References

[1] J. Chen, Q. Li, Q. Peng and K.H. Wong: Pattern Analysis and Applications, Vol. 18 (2015) No.2, p.441.

[2] Y.S. Lin, J.Y. Jiang and S.J. Lee: IEEE Transactions on Knowledge and Data Engineering, Vol. 26 (2014) No.7, p.1575.

[3] J. Johnson, R. Krishna and M. Stark: *IEEE Conference on Computer Vision and Pattern Recognition* (Boston, MA, USA, June 7-12, 2015), p. 3668.

[4] G. Song, S. Wang, Q. Huang and Q. Tian: *IEEE International Conference on Computer Vision* (Santiago, Chile, December 7-13, 2015), p. 4050.

[5] J. Masci, M.M. Bronstein, A.M. Bronstein and J. Schmidhuber: IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 36 (2014) No.4, p.824.

[6] H. Zhang, L. Chen, J. Liu and J. Yuan: *IEEE International Conference on Image Processing* (Paris, France, October 27-30, 2014), p. 5916.

[7] J. Wang, Y. Zhou, K. Duan and J.Y. Wang: *IEEE International Conference on Systems, Man, and Cybernetics* (Hong Kong, October 9-12, 2015), p. 1882.

[8] A.K. Menon, D. Surian and S. Chawla: *Proceedings of the 2015 SIAM International Conference on Data Mining* (Society for Industrial and Applied Mathematics, USA 2015), p. 199.

[9] W. Zhang and H. Zhang: *Advanced Intelligent Computing Theories and Applications* (Springer International Publishing, Germany 2015), p. 221.

[10] Y. Chen, L. Wang, W. Wang and Z. Zhang: *IEEE International Conference on Image Processing* (Orlando, Florida, USA, September 30-October 3, 2012), p. 1949.

[11] S.W. Liu, J. Tang and D. Yan: *Automation, Mechanical Control and Computational Engineering* (Ji'nan, China, April 24-26, 2015), p. 139.

[12] J. Costa Pereira, E. Coviello, G. Doyle and N. Rasiwasia: IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 36 (2014) No.3, p.521.

[13] M. Srinivas and L. M. Patnaik: IEEE Transactions on SMC, Vol. 24 (1994) No.4, p.656.