Unsupervised Multi-View Subspace Learning via Maximizing Dependence

Meixiang Xu, Zhenfeng Zhu*, and Yao Zhao

Institute of Information Science, Beijing Jiaotong University, Beijing, 100044, China
Beijing Key Laboratory of Advanced Information Science and Network Technology,
Beijing, 100044, China
{15112068, zhfzhu, yzhao}@bjtu.edu.cn

Abstract. The recent years have witnessed the great significance of learning from multi-view data in real-world tasks, such as clustering, classification and retrieval. In this paper, we propose an unsupervised dependence (correlation) maximization model, referred to as UDM, for multi-view subspace learning. Our proposed model is based on Hilbert-Schmidt Independence Criterion(HSIC), a kernel-based technique for measuring dependence between two random variables statistically. In the proposed model, sparse constraint on the projection matrix for each view is imposed as regularizations, playing the role of feature selection, which enables to capture more discriminative subspace representations. To efficiently solve the formulated optimization problem, an iterative optimizing algorithm is designed. Experimental results on cross-modal retrieval have shown the superiority of UDM over the compared approaches and the rapid convergence speed of the optimizing algorithm.

Keywords: Subspace Learning, Multi-view Learning, Dependence

1 Introduction

In recent years, there has been rapid growth of multi-view data and much efforts have witnessed the great significance of learning from multi-view data in many real-world applications. Often, multi-view data, presented in diverse forms or derived from different domains, show heterogeneous characteristics, which is a big challenge for practical tasks such as cross-modal retrieval, machine translation, biometric verification, matching, transfer learning, etc. To address this challenge, two common strategies are mainly adopted. One is to learn distance metrics, the other is to learn a common space. In this paper we focus on the latter, that is, multi-view subspace learning.

Intrinsically, multiple views represent the same underlying semantic data object, therefore, they are inherently correlated to each other. Based on this fact, statical techniques such as canonical correlation analysis (CCA) [11], Kullback-Leibler (KL) divergence [1], mutual information [12] and Hilbert-Schmidt Independence Criterion (HSIC) [6] and so on, to measure correlation (dependence) of two random variables, have been investigated and used for multi-view learning. Especially, CCA is the most popular one among the aforementioned measures.

From the point of multi-view learning, CCA can be regarded as finding the projection matrix for each view of the data object, by which the data can be projected into a common subspace where the low dimensional embeddings are maximally correlated. Due to the encouraging success of CCA, CCA-based approaches have attracted much attention during the past decades. Substantial variants of CCA have been developed for multi-view subspace representations, including unsupervised ones [20, 18, 16], supervised ones [15, 19], sparsity-based ones [10, 4], DNN-based ones [13, 2, 22], etc. Like CCA, considerable attention has been gradually paid to the use of HSIC for the dependence-based tasks of multi-view classification [7] , clustering [3], dictionary learning [8, 9]. Concerning these methods, they are supervised ones, which conformably expect the dependence between multi-view data and the corresponding labels to be maximized. However, labels are unknown beforehand in most cases of multi-view learning tasks.

In this paper, we propose an unsupervised dependence maximization model for multi-view subspace learning, referred to as UDM. The proposed UDM is designed specifically for the case of two views, which can be extended to multiple views. Unlike the supervised HSIC-based approaches, UDM aims at maximizing the dependence between two views under the unsupervised setting. Simultaneously, it incorporates the imposed $\ell_{2,1}$-norm constraint on the projection matrix for each view as regulations, playing the role of feature selection, which enables more discriminative representations. To solve the optimization problem formulated by UDM, an efficient iterative optimizing algorithm is designed. Experimental results on two real-world cross-modal datasets demonstrate the effectiveness and efficiency of UDM, and show the superiority of UDM over the compared approaches. Convergence curves of the objective function demonstrate the rapid convergence speed of the optimization algorithm.

2 Notations and HSIC

2.1 Notations

To begin with, we introduce some notations used in this paper. For any matrix $\mathbf{A} \in \mathbb{R}^{n\times m}$, $\mathbf{A}^{\cdot\cdot i}$ and $\mathbf{A}^{:j}$ are used to represent its i-th row and j-th column, respectively. $\left\|\mathbf{A}\right\|_{2,1}$ is the $\ell_{2,1}$-norm of $\mathbf{A}$, defined as $\left\|\mathbf{A}\right\|_{2,1} = \sum\limits_{i=1}^{n} \left\|\mathbf{A}^{\cdot\cdot i}\right\|_2$. $\left\|\mathbf{A}\right\|_{HS}$ is the Hilbert-Schmidt norm of $\mathbf{A}$, defined as $\left\|\mathbf{A}\right\|_{HS} = \sqrt{\sum\limits_{i,j} a_{ij}^2}$. Besides, $tr\left(\cdot\right)$ represents the trace operator, $\otimes$ the tensor product and $\mathbf{I}$ an identity matrix with an appropriate size. Throughout the paper, matrices and vectors are represented in bold uppercase and lowercase letters respectively. Variables are represented by conventional letters.

2.2 Hilbert-Schmidt Independence Criteria

Let C_{xy} be the cross-covariance function between x and y, $\varphi(x)$ and $\phi(y)$ two mapping functions with $\varphi(x) : x \in \mathcal{X} \to \mathbb{R}$ and $\phi(y) : y \in \mathcal{Y} \to \mathbb{R}$, $\mathcal{G}$ and $\mathcal{H}$

two Reproducing Kernel Hilbert Spaces (RKHSs) in $\mathcal{X}$ and $\mathcal{Y}$. The associated positive definite kernels k_x and k_y is defined as $k_x(x, x^T) =< \Phi(x), \Phi(x) >_{\mathcal{G}}$ and $k_y(y, y^T) =< \Phi(y), \Phi(y) >_{\mathcal{H}}$. Then cross-covariance C_{xy} is defined as:

$$C_{xy} = E_{xy}\left[\left(\varphi\left(x\right) - u_x\right) \otimes \left(\phi\left(y\right) - u_y\right)\right] \ . \tag{1}$$

where u_x and u_y is the expectation of $\varphi(x)$ and $\phi(y)$ respectively, i.e. $u_x = E\left(\varphi\left(x\right)\right)$ and $u_y = E\left(\phi\left(y\right)\right)$.

Given two independent RKHSs $\mathcal{G}$, $\mathcal{H}$ and the joint distribution p_{xy}, HSIC is the Hilbert-Schmidt norm of C_{xy}, defined as:

$$HSIC\left(p_{xy}, \mathcal{G}, \mathcal{H}\right) := \left\|C_{xy}\right\|_{HS}^2 \ . \tag{2}$$

In practical applications, the empirical estimate of HSIC is commonly used. Given n finite number of data samples $Z := \{(x_1, y_1), \cdots, (x_N, y_N)\}$, the empirical expression of HSIC is formulated as:

$$HSIC\left(Z, F, G\right) = \left(n-1\right)^{-2} tr\left(\mathbf{K}_1 \mathbf{H} \mathbf{K}_2 \mathbf{H}\right) \ . \tag{3}$$

where $\mathbf{K}_1$ and $\mathbf{K}_2$ are two Gram matrices with $k_{1,ij} = k_1(x_i, x_j)$ and $k_{2,ij} = k_2(y_i, y_j)$ $(i, j = 1, \cdots, N)$. $\mathbf{H} = \mathbf{I} - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T$, is a centering matrix, and $\mathbf{1}_n \in \mathbb{R}^n$ is a full-one column vector.

More details about HSIC can be found in literatures [6].

3 Multi-view Subspace Learning Model via Kernel Dependence Maximization

3.1 The Proposed Subspace Learning Model

Multi-view subspace learning approaches aim to project different high-dimensional heterogeneous views into a coherent low-dimensional common subspace in linear or nonlinear ways, where samples with the same or similar semantics have the coherent representation, as illustrated in Fig. 1.

In the following, the case of two views is mainly considered. Suppose that there are n pairs of observation samples $\left\{\mathbf{x}_1^i, \mathbf{x}_2^i\right\} \in \mathbb{R}^{1\times d_1} \times \mathbb{R}^{1\times d_2}$, where $\left\{\mathbf{x}_1^i\right\}_{i=1}^n$ and $\left\{\mathbf{x}_2^i\right\}_{i=1}^n$ are from view $\mathbf{X}_1 = \left[\mathbf{x}_1^1, \cdots, \mathbf{x}_1^n\right]^T \in \mathbb{R}^{n\times d_1}$ and view $\mathbf{X}_2 = \left[\mathbf{x}_2^1, \cdots, \mathbf{x}_2^n\right]^T \in \mathbb{R}^{n\times d_2}$ respectively. $\left\{\mathbf{x}_1^i, \mathbf{x}_2^i\right\}$ denotes the i-th pair samples in the sample set $\left\{\mathbf{x}_1^i, \mathbf{x}_2^i\right\}_{i=1}^n$. The goal of this paper is to learn the projection matrix $\mathbf{P}_v(v = 1, 2)$ for views $\mathbf{X}_v(v = 1, 2)$ simultaneously. Through $\mathbf{P}_v$, heterogeneous views $\mathbf{X}_v$ are projected into a common subspace $\mathbf{S}$, where samples $\mathbf{x}_1^i$ and $\mathbf{x}_2^j(i, j = 1, \cdots, n)$ with the same and similar semantics have the coherent representation. Correspondingly, the new representation for $\mathbf{X}_1$ and $\mathbf{X}_2$ in the shared subspace is $\mathbf{X}_1^S = \mathbf{X}_1\mathbf{P}_1$ and $\mathbf{X}_2^S = \mathbf{X}_2\mathbf{P}_2$. Adopting linear kernel as the kernel measure, kernel matrices $\mathbf{K}_{\mathrm{X}_1}$ and $\mathbf{K}_{\mathrm{X}_2}$ can be denoted as $\mathbf{K}_{\mathrm{X}_1} = \left\langle \mathbf{X}_1^S, \mathbf{X}_1^S \right\rangle = \mathbf{X}_1\mathbf{P}_1\mathbf{P}_1^T\mathbf{X}_1^T$ and $\mathbf{K}_{\mathrm{X}_2} = \left\langle \mathbf{X}_2^S, \mathbf{X}_2^S \right\rangle = \mathbf{X}_2\mathbf{P}_2\mathbf{P}_2^T\mathbf{X}_2^T$. Since multi-view data describe the same semantic object from different levels, they are

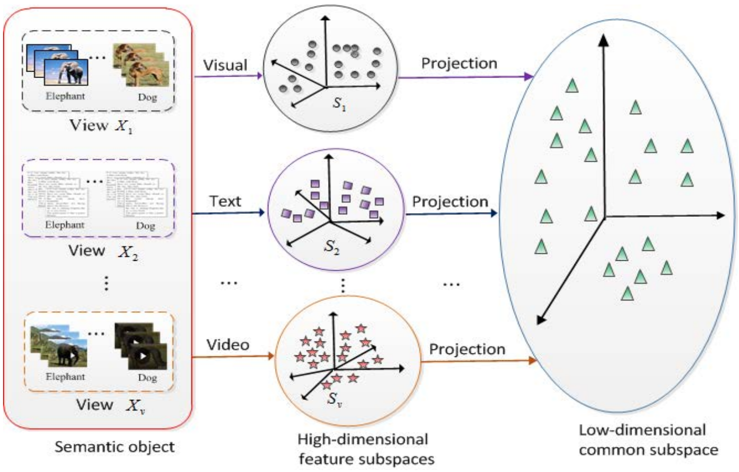


Fig. 1. Sketch map of subspace learning for multi-view data

inherently correlated to each other. Based on HSIC, the proposed unsupervised subspace learning model is formulated as:

$$\begin{aligned} &\max_{\mathbf{P}_1,\mathbf{P}_2} tr(\mathbf{H}\mathbf{X}_1\mathbf{P}_1\mathbf{P}_1^T\mathbf{X}_1^T\mathbf{H}\mathbf{X}_2\mathbf{P}_2\mathbf{P}_2^T\mathbf{X}_2^T) \\ &\quad s.t.\ \mathbf{P}_1^T\mathbf{P}_1 = \mathbf{I}_1; \mathbf{P}_2^T\mathbf{P}_2 = \mathbf{I}_2\ , \end{aligned} \tag{4}$$

where the orthogonal constraints imposed on $\mathbf{P}_v(v = 1, 2)$ is to avert the trivial solution of all zeros.

As demonstrated in literatures such as [14, 21], $\ell_{2,1}$-norm based learning models have capabilities of sparsity, feature selection and robustness to noise. Inspired by this, by imposing the $\ell_{2,1}$-norm constraint on the projection matrix $\mathbf{P}_v(v = 1, 2)$ as regularization terms to learn more discriminative representations for multi-view data, accordingly we have the following formulation:

$$\begin{aligned} &\max_{\mathbf{P}_1,\ \mathbf{P}_2} tr\left(\mathbf{H}\mathbf{X}_1\mathbf{P}_1\mathbf{P}_1^T\mathbf{X}_1^T\mathbf{H}\mathbf{X}_2\mathbf{P}_2\mathbf{P}_2^T\mathbf{X}_2^T\right) - \lambda_1\|\mathbf{P}_1\|_{2,1} - \lambda_2\|\mathbf{P}_2\|_{2,1} \\ &s.t.\ \ \mathbf{P}_1^T\mathbf{P}_1 = \mathbf{I}_1;\ \ \mathbf{P}_2^T\mathbf{P}_2 = \mathbf{I}_2\ , \end{aligned} \tag{5}$$

where λ_1 and λ_2 are the regularization parameters.

3.2 Optimization

Since the optimization objective function involved the $\ell_{2,1}$-norm, which is an intractable problem to handle. Consequently, here we employ the alternative optimization strategy to solve the optimization problem. With $\|\mathbf{A}\|_{2,1} = tr\left(\mathbf{A}^T\mathbf{D}\mathbf{A}\right)$

where $\mathbf{D} = diag\left(\frac{1}{\|A^{\cdot\cdot i}\|_2}\right)$, first let us re-express the formulation in Eq. (5) as:

$$\begin{aligned}&\max_{\mathbf{P}_1,\,\mathbf{P}_2} tr\left(\mathbf{H}\mathbf{X}_1\mathbf{P}_1\mathbf{P}_1^T\mathbf{X}_1^T\mathbf{H}\mathbf{X}_2\mathbf{P}_2\mathbf{P}_2^T\mathbf{X}_2^T\right) - \lambda_1 tr\left(\mathbf{P}_1^T\mathbf{D}_1\mathbf{P}_1\right) - \lambda_2 tr\left(\mathbf{P}_2^T\mathbf{D}_2\mathbf{P}_2\right)\\ &s.t.\ \mathbf{P}_1^T\mathbf{P}_1 = \mathbf{I}_1;\ \mathbf{P}_2^T\mathbf{P}_2 = \mathbf{I}_2\ .\end{aligned} \tag{6}$$

Specifically, according to the alternative optimization rules, the optimization problem formulated in Eq. (6) (i.e. Eq. (5)) can be decomposed into the following two sub-maximization ones:

1) **Solve $\mathbf{P}_1$, fixing $\mathbf{P}_2$**:

$$\begin{aligned}&\max_{\mathbf{P}_1} tr(\mathbf{H}\mathbf{X}_1\mathbf{P}_1\mathbf{P}_1^T\mathbf{X}_1^T\mathbf{H}\mathbf{X}_2\mathbf{P}_2\mathbf{P}_2^T\mathbf{X}_2^T) - \lambda_1 tr\left(\mathbf{P}_1^T\mathbf{D}_1\mathbf{P}_1\right)\\ \Leftrightarrow\ &\max_{\mathbf{P}_1} tr\left(\mathbf{P}_1^T(\mathbf{X}_1^T\mathbf{H}\mathbf{X}_2\mathbf{P}_2\mathbf{P}_2^T\mathbf{X}_2^T\mathbf{H}\mathbf{X}_1 - \lambda_1\mathbf{D}_1)\mathbf{P}_1\right)\\ &s.t.\ \mathbf{P}_1^T\mathbf{P}_1 = \mathbf{I}_1\ .\end{aligned} \tag{7}$$

Let $\mathbf{B}_1 = \mathbf{X}_1^T\mathbf{H}\mathbf{X}_2\mathbf{P}_2\mathbf{P}_2^T\mathbf{X}_2^T\mathbf{H}\mathbf{X}_1 - \lambda_1\mathbf{D}_1$, we can obtain $\mathbf{P}_1$ by solving the eigenvalue problem of $\mathbf{B}_1$, here $\mathbf{P}_1$ consists of the first d eigenvectors corresponding to the d largest eigenvalues of $\mathbf{B}_1$.

2) **Solve $\mathbf{P}_2$, fixing $\mathbf{P}_1$**:

$$\begin{aligned}&\max_{\mathbf{P}_2} tr(\mathbf{H}\mathbf{X}_2\mathbf{P}_2\mathbf{P}_2^T\mathbf{X}_2^T\mathbf{H}\mathbf{X}_1\mathbf{P}_1\mathbf{P}_1^T\mathbf{X}_1^T) - \lambda_2 tr\left(\mathbf{P}_2^T\mathbf{D}_2\mathbf{P}_2\right)\\ \Leftrightarrow\ &\max_{\mathbf{P}_2} tr\left(\mathbf{P}_2^T(\mathbf{X}_2^T\mathbf{H}\mathbf{X}_1\mathbf{P}_1\mathbf{P}_1^T\mathbf{X}_1^T\mathbf{H}\mathbf{X}_2 - \lambda_2\mathbf{D}_2)\mathbf{P}_2\right)\\ &s.t.\ \mathbf{P}_2^T\mathbf{P}_2 = \mathbf{I}_2\ .\end{aligned} \tag{8}$$

Likewise, let $\mathbf{B}_2 = \mathbf{X}_2^T\mathbf{H}\mathbf{X}_1\mathbf{P}_1\mathbf{P}_1^T\mathbf{X}_1^T\mathbf{H}\mathbf{X}_2 - \lambda_2\mathbf{D}_2$, we can obtain $\mathbf{P}_2$ by solving the eigenvalue problem of $\mathbf{B}_2$, here $\mathbf{P}_2$ consists of the first d eigenvectors corresponding to the d largest eigenvalues of $\mathbf{B}_2$.

To better understand the procedure for solving the proposed method, we summarize in detail the solver for solving the optimization problem in Eq. (5) as Algorithm 1.

3.3 Convergence Analysis

The convergence of the proposed UDM under the iterative optimization algorithm in Algorithm 1 can be summarized by the following Theorem 1.

Theorem 1. *Under the iterative optimizing rules in Algorithm 1, the objective function defined by Eq. (5) is increasing monotonically, and it can converge to its global maximum.*

Due to space limitation, here we omit the detailed proof of Theorem 1. The convergence curves in Section 4.5 can also demonstrate the good convergence behavior of the optimizing algorithm.

Algorithm 1: Multi-view Subspace Learning via Dependence Maximizing

```
Input: Multi-view data $\mathbf{X}_v \in \mathbb{R}^{n \times d_v}$, $v = 1, 2$; the regularization parameters $\lambda_1$
       and $\lambda_2$.
Output: The projection matrices $\mathbf{P}_v$, $v = 1, 2$.
Initializing: Initialize $\mathbf{P}_1$ and $\mathbf{P}_2$ randomly, let $t = 0$;
while not converge do
    Update $\mathbf{D}_1$ and $\mathbf{D}_2$: $\mathbf{D}_1^{(t)} = diag\left(\frac{1}{2\left\|\mathbf{P}_1^{\cdot\cdot i(t)}\right\|_2}\right)$, $\mathbf{D}_2^{(t)} = diag\left(\frac{1}{2\left\|\mathbf{P}_2^{\cdot\cdot i(t)}\right\|_2}\right)$;
    Update $\mathbf{P_1}$: obtain $\mathbf{P_1^{(t+1)}}$ by performing eigen-decomposition on
    $\mathbf{B}_1 = \mathbf{X}_1^T\mathbf{H}\mathbf{X}_2\mathbf{P}_2^{(t)}\left(\mathbf{P}_2^{(t)}\right)^T\mathbf{X}_2^T\mathbf{H}\mathbf{X}_1 - \lambda_1\mathbf{D}_1^{(t)}$. The $d$ eigenvectors
    corresponding to the first largest $d$ eigenvalues of $\mathbf{B}_1$ compose $\mathbf{P}_1$ ;
    Update $\mathbf{P}_2$: obtain $\mathbf{P}_2^{(t+1)}$ by performing eigen-decomposition on
    $\mathbf{B}_2 = \mathbf{X}_2^T\mathbf{H}\mathbf{X}_1\mathbf{P}_1^{(t)}\left(\mathbf{P}_1^{(t)}\right)^T\mathbf{X}_1^T\mathbf{H}\mathbf{X}_2 - \lambda_2\mathbf{D}_2^{(t)}$. The $d$ eigenvectors
    corresponding to the first largest $d$ eigenvalues of $\mathbf{B}_2$ compose $\mathbf{P}_2$;
    $t = t + 1$;
return $\mathbf{P}_1$, $\mathbf{P}_2$.
```

4 Experiments

To test the performance of UDM, we conducted experiments on cross-modal retrieval between image and text, i.e. using image to query text (I2T) and using text to query image (T2I), adopting Mean Average Precision (MAP) as the evaluation metric and the normalized correlation (NC) as the distance measure [15].

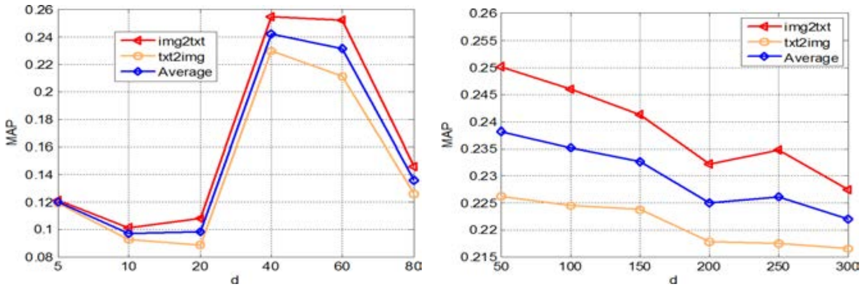
4.1 Datasets

The follow-ups are brief descriptions on the used two datasets i.e. Wikipedia [23] and NUS-WIDE [5].

- **Wikipedia:** This dataset consists of 2866 image-text pairs labeled with 10 semantic classes in total. For each image-text pair, we extract 4096-dimensional visual features by convolutional neural network to represent the image view, and 100-dimensional LDA textual features to represent the text view. In the experiment, the dataset is partitioned into two parts, one for training (2173 pairs) and the other for testing (693 pairs).
- **NUS-WIDE:** This dataset is a subset from [5], including 190420 image examples totally, each with 21 possible labels. For each image-text pair, we extract 500-dimensional SIFT BoVW features for image and 1000-dimensional text annotations for text. To reduce the computational complexity, further we sample a subset with 8687 pairs of image-text. Likewise, the dataset is divided into two parts, one for training (5212 pairs) and the other for testing (3475 pairs).

4.2 Benchmark Approaches and Experimental Setup

The proposed UDM is unsupervised, kernel-based, correlation-based and sparsity-based. Accordingly, the compared approaches include CCA, KPCA [17], KCCA [16], SCCA [10]. The parameters involved in the compared approaches are kept default following the literatures. For details, please refer to the corresponding literatures. Next, we will present the specific settings for the parameters involved in UDM. First, by fixing λ_1 and λ_2 we determine the optimal d. Specifically, we tune d from the range of $\{5, 10, 20, 40, 60, 80\}$ and $\{50, 100, 150, 200, 250, 300, 350\}$ on Wikipedia and NUS-WIDE respectively, as shown in Fig. 2. It can be seen from Fig. 2, with $d = 40$ on Wikipedia and $d = 50$ on NUS-WIDE, UDM obtains the best performance. Therefore, in the following experiments we set $d = 40$ and $d = 50$ for Wikipedia and NUS-WIDE. Then, with d fixed we decide the optimal λ_1 and λ_2 by tuning them from $\{10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 10^2, 10^3, 10^4, 10^5\}$ with d fixed. Empirically, we determine $\lambda_1 = 10^{-3}$ and $\lambda_2 = 10^3$ on Wikipedia, as well as $\lambda_1 = \lambda_2 = 10^{-5}$ on NUS-WIDE. In the subsequent section, we will give the parameter sensitivity analysis on λ_1 and λ_2.



(a) Wikipedia (b) NUS-WIDE

Fig. 2. MAP vs. varying d on Wikipedia and NUS-WIDE with λ_1 and λ_2 fixed

4.3 Results

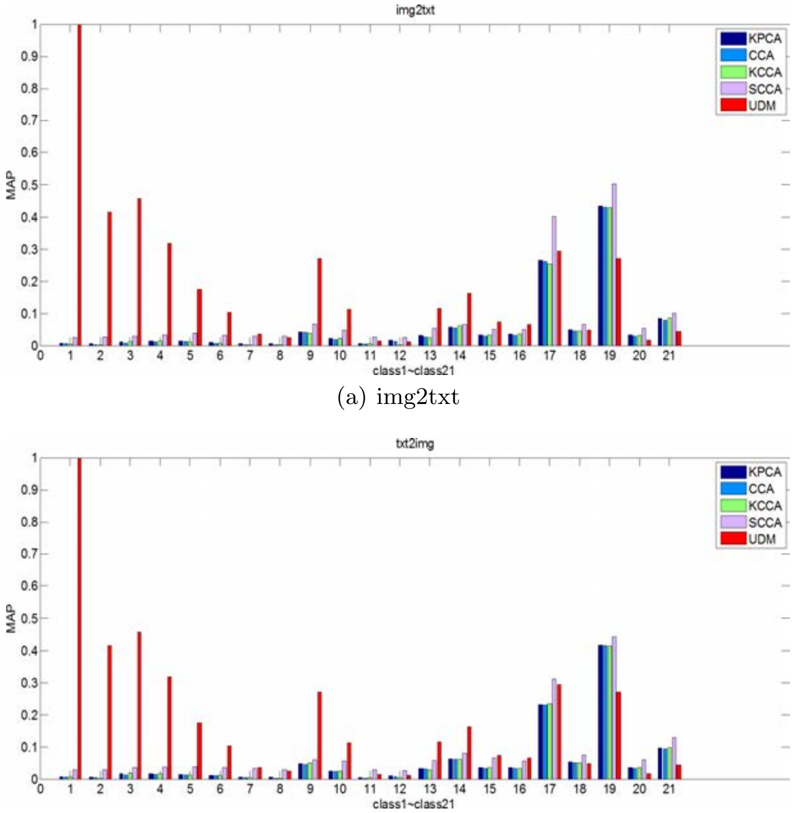
Table 1 and Table 2 displays the comparison results on two datasets, respectively. As can be seen from Table 1 and Table 2, the proposed UDM performs best, followed by KCCA, on both Wikipedia and NUS-WIDE. Besides, Fig. 3 shows the per-class MAP scores of all the compared approaches on NUS-WIDE. From Fig. 3, we can observe that UDM achieves better results on most categories, but it is not always the best on each category. More specifically, it achieves the best result on the first sixteen categories while it is the worst among the five approaches on category 20 and category 21. Therefore, incorporating label supervision information will be considered to improve UDM for each category.

Table 1. MAP Comparison on Wikipedia

Approaches	Image as query	Text as query	Average
KPCA	0.1983	0.1826	0.1905
CCA	0.1222	0.1189	0.1206
KCCA	0.3337	0.3031	0.3184
SCCA	0.2270	0.1961	0.2116
UDM	**0.4204**	**0.4394**	**0.4299**

Table 2. MAP Comparison on NUS-WIDE

Approaches	Image as query	Text as query	Average
KPCA	0.2326	0.2215	0.2171
CCA	0.2441	0.2356	0.2399
KCCA	0.2554	0.2451	0.2503
SCCA	0.2415	0.2145	0.2145
UDM	**0.2904**	**0.2498**	**0.2702**



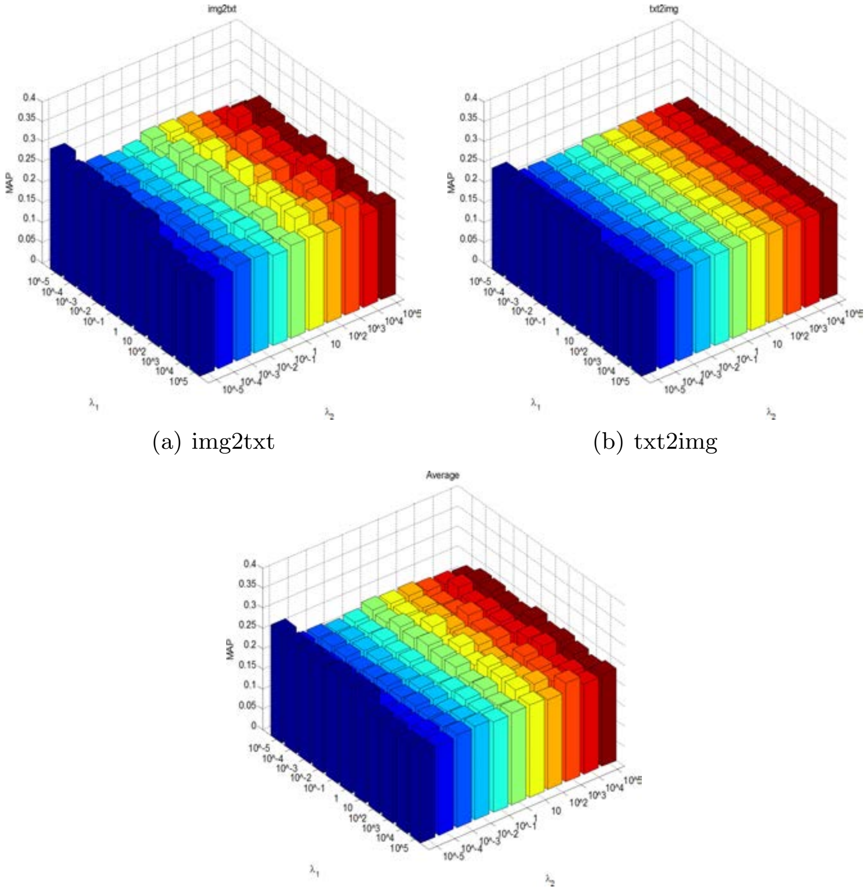
(a) img2txt

(b) txt2img

Fig. 3. Per class MAP on NUS-WIDE

4.4 Parameter Sensitivity Analysis

To show the impacts of λ_1 and λ_2 on UDM, we have carried out experiments on Wikipedia and NUS-WIDE respectively, by tuning them from the same range set as Subsection 4.2. Fig. 4 and Fig. 5 shows the retrieval MAP scores versus different values of λ_1 and λ_2 on NUS-WIDE and Wikipedia respectively. From Fig. 4 and Fig. 5, we can see that the performance of UDM varies as λ_1 and λ_2 changes. By contrast, the proposed UDM on Wikipedia is much more sensitive to two parameters than on NUS-WIDE.

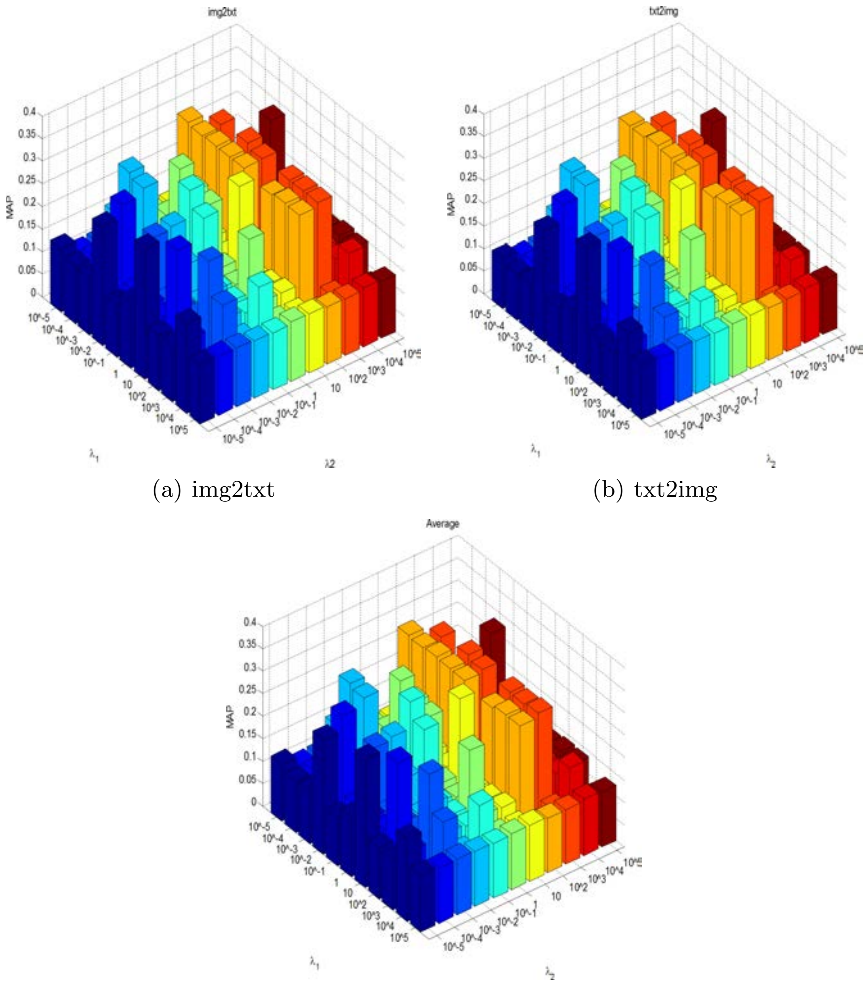


(a) img2txt

(b) txt2img

(c) Average

Fig. 4. MAP vs. varying λ_1 and λ_2 on NUS-WIDE



(a) img2txt

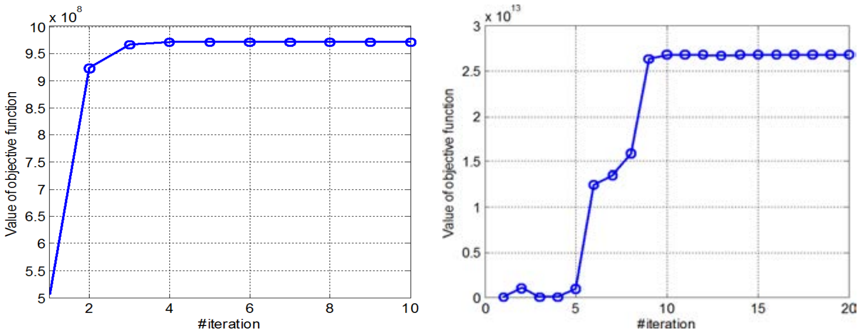
(b) txt2img

(c) Average

Fig. 5. MAP vs. varying λ_1 and λ_2 on Wikipedia

4.5 Convergence Study

Fig. 6 displays the relationship between the objective function and the number of iteration on Wikipedia and NUS-WIDE, respectively. As can be observed from Fig. 6, for each dataset, the objective function defined in Eq. (5) can rapidly converge to its maximum within about ten iterations, which demonstrates the efficiency of the designed iterative optimization algorithm.



(a) Wikipedia (b) NUS-WIDE

Fig. 6. The objective function vs. the number of iteration

5 Conclusions and Future Work

In this paper, we have proposed a HSIC-based unsupervised learning approach for discovering common subspace representations shared by multi-view data, which is a kernel-based, correlation-based and sparsity-based projection method. To solve the optimization problem, we develop an efficient iterative optimizing algorithm. Cross-modal retrieval results on two benchmark datasets have shown the superiority of the proposed UDM over the compared approaches. Inspired by CCA-like methods, nonlinear extensions of UDM will be considered by incorporating nonlinear kernel and neutral network to expect a better common representation for multi-view data in future work.

Acknowledgments. This work was jointly supported by the National Natural Science Foundation of China (NO.61572068, NO.61532005), the National Key Research and Development of China (NO.2016YFB0800404) and the Fundamental Scientific Research Project (NO.KKJB16004536).

References

1. C. A and Y. H. H. A new learning algorithm for blind signal separation. *NIPS*, 3:757–763, 1996.

2. G. Andrew, R. Arora, J. Blimes, and K. Livescu. Deep canonical correlation analysis. In *ICML*, pages 1247–1255, 2013.
3. X. Cao, C. Zhang, H. Fu, S. Liu, and H. Zhang. Diversity-induced multi-view subspace clustering. In *CVPR*, pages 586–594, 2015.
4. D. Chu, L. Liao, M. K. Ng, and X. Zhang. Sparse canonical correlation analysis:new formulation and algorithm. *TPAMI*, 35(12):3050–3065, 2013.
5. T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y. Zhang. Nus-wide: a real-world web image database from national university of singapore. In *CIVR*, 2009.
6. J. CP. Information theory, machine learning, and repreoducing kernel hilbert spaces. *Information Science and Statistics*, pages 1–45, 2010.
7. Z. Fang and Z. Zhang. Simultaneously combining multi-view multi-label learning with maximum margin classification. In *ICDM*, pages 864–869, 2012.
8. M. J. Gangeh, P. Fewzee, A. Ghodsi, Mohamed, and Fakhri. Kernelized supervised dictionary learning. *TSP*, 61(19):4753–4767, 2013.
9. M. J. Gangeh, P. Fewzee, A. Ghodsi, Mohamed, and Fakhri. Multi-view supervised dictionary learning in speech emotion recognition. *ACM Transactions on Audio, Speech, and Language Processing*, 22(6):1056–1068, 2014.
10. S.-T. J. Hardoon DR. Sparse canonical correlation analysis. *Machine Learning*, 83(3):331–353, 2011.
11. H. Hotelling. Relations between two sets of variates. *Biometrika*, 28(3/4):321–377, 1936.
12. T. K. Feature extraction by non-parametric mutual information maximization. *Journal of machine learning research*, 3(3):1415–1438, 2003.
13. J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Ng. Multimodal deep learning. In *ICML*, pages 689–696, 2011.
14. F. Nie, H. Huang, X. Cai, and C. Ding. Efficient and robust feature selection via joint l2,1-norms minimization. *NIPS*, pages 1813–1821, 2010.
15. N. Rasiwasia, J. C. Pereira, E. Coviello, G. Doyle, G. R. G. Lanckriet, and R. L. ahd Nuno Vasconcelos. A new appraoch to cross-modal multimedia retrieval. In *ICMM*, pages 251–260, 2010.
16. A. S. A kernel method for canonical correlation analysis. In *IMPS2001*, 2007.
17. S. A. R. G. M. K. R. SchAolkopf B, Mika S. Kernel pca pattern reconstruction via approximation preimages. In *ICANN*, 1998.
18. A. Sharma and D. W. Jacobs. Bypassing synthesis: Pls for face recognition with pose, lowresolution and sketch. In *CVPR*, pages 593–600, 2011.
19. K. Tae-Kyun, J. Kittler, and R. Cipolla. Discriminative learning and recognition of image set classess using canonical correlation. *TPAMI*, 29(6):1005–1018, 2007.
20. J. B. Tenenbaum and W. T. Freeman. Separating style and content with bilinear models. *Neural Computation*, 12(6):1247–1283, 2000.
21. K. Wang, R. He, L. Wang, W. Wang, and T. Tan. Joint feature selection and subspace learning for cross-modal retrieval. *TPAMI*, 38(10):2010–2023, 2016.
22. W. Wang, R. Arora, K. Livescu, and J. Bilmes. On deep multi-view representation learning. In *ICML*, 2015.
23. Y. Wei, Y. Zhao, C. Lu, S. Wei, L. Liu, Z. Zhu, and S. Yan. Cross-modal retrieval with cnn visual features: A new baseline. *TCB*, 47(2):449–460, 2017.