Multi-view Feature Learning with Discriminative Regularization

Jinglin Xu, Junwei Han, Feiping Nie
Northwestern Polytechnical University, Xi'an, 710072, P. R. China
{xujinglinlove, junweihan2010, feipingnie}@gmail.com

Abstract

More and more multi-view data which can capture rich information from heterogeneous features are widely used in real world applications. How to integrate different types of features, and how to learn low dimensional and discriminative information from high dimensional data are two main challenges. To address these challenges, this paper proposes a novel multi-view feature learning framework, which is regularized by discriminative information and obtains a feature learning model which contains multiple discriminative feature weighting matrices for different views, and then yields multiple low dimensional features used for subsequent multi-view clustering. To optimize the formulable objective function, we transform the proposed framework into a trace optimization problem which obtains the global solution in a closed form. Experimental evaluations on four widely used datasets and comparisons with a number of state-of-the-art multi-view clustering algorithms demonstrate the superiority of the proposed work.

1 Introduction

Nowadays, with the rapid development of information technology, a large number of data can be described by different kinds of features. These features usually are generated by various feature extraction ways. Taking the image data as an example, many heterogeneous visual features are widely used, which include SIFT [Lowe, 2004], HOG [Dalal and Triggs, 2005], LBP [Ojala *et al.*, 2002], GIST [Oliva and Torralba, 2001], CMT [Yu *et al.*, 2002] and CENT [Wu and Rehg, 2008]. They describe the rich content of image data from different viewpoints and capture their corresponding certain properties.

For processing such data, there are two challenges. The first is how to integrate different types of features effectively, which certainly can lead to more accurate and robust performance than by only using each individual type of features. The second is how to reduce dimensions of high dimensional heterogeneous features efficiently, which may lead to heavy complexity and curse of dimensionality.

A number of earlier efforts have been devoted to address these challenges. Focusing on the first challenge, many multi-view methods have been developed [Eaton *et al.*, 2010; Kumar *et al.*, 2011; Cai *et al.*, 2011; Jiang *et al.*, 2012; Cai *et al.*, 2013a; 2013b; Wang *et al.*, 2014] based on various techniques, such as Co-EM (which iteratively estimates the propagation with each view and transfers the constraints across views), Co-Regularized (which makes the clustering hypotheses on different views agree with each other), Minimax (which achieves multi-feature fusion via minimizing the maximum weighted disagreement costs) and so on. Although they can achieve heterogeneous features integration in a common space, there still exist some drawbacks, such as the lack of dimensionality reduction, which may result in the problems of heavy computational complexity, curse of dimensionality or over-fitting when dealing with high dimensional and complex data.

To address the second challenge, much progress based on multi-view setting has been made in recent years [Chaudhuri *et al.*, 2009; Han *et al.*, 2012; Tang *et al.*, 2013; Wang *et al.*, 2013; Zhao *et al.*, 2014; Cao *et al.*, 2015; Xu *et al.*, 2016]. These algorithms adopt a variety of methods to reduce data dimensionality, like CCA (which maximizes the total correlations between any two views to obtain one common space), Co-training (which exploits label learned automatically in one view to learn discriminative subspaces in another), HSIC (as a diversity term to explore the complementary of multi-view representation), LDA (which maximizes the between-class scatter matrix and minimizes the within-class scatter matrix) and so on. Although these algorithms can learn low dimensional features to reduce heavy computational complexity and curse of dimensionality, their performance for subsequent multi-view tasks (i.e., clustering/classification) is still not satisfactory.

Most previous works perform the multi-view feature learning in the manner of unsupervised learning, thus the discriminant information, e.g., label information, is not explicitly taken into account. Actually, nowadays in many real applications, the label information of the data is available already on the web or can be acquired from sensors easily. For example, image data and their label information, which are publicly available and can be downloaded them from ImageNet [1], CV

[1] http://image-net.org/

Datasets [2] and UCI Machine Learning Repository[3]. Thus, a promising idea is to develop supervised feature learning methods which may infer more discriminative information with the help of data labels.

Recently, two works of [Kan *et al.*, 2012] and [Arora and Livescu, 2014] were proposed by utilizing labeled data. The former, Multi-view Discriminative Analysis [Kan *et al.*, 2012], forms a generalized Rayleigh quotient by combining the between-class and within-class variations. Through jointly optimizing multiple transforms, it learns single unified discriminant common space for multiple views. Unfortunately, original objective function of [Kan *et al.*, 2012] in the form of trace ratio does not have the closed form solution, thus, authors reformulate their framework into a more tractable one: ratio trace. However, [Wang *et al.*, 2007] pointed out that transforming trace ratio into the corresponding ratio trace form is inexact, thus its feature learning models may not achieve satisfactory performance for multi-view tasks. The latter [Arora and Livescu, 2014] is mainly applied to bottleneck features, which can be learned if the training set is phonetically labeled. It includes a series of combinations of LDA and CCA. However, this simple concatenation of LDA and CCA renders its framework to be more separate and cannot tackle high dimensional multi-view data efficiently.

According to above mentioned analysis, considering that the cost of obtaining labeled data is no longer expensive and two challenges are not well addressed based on unsupervised framework, we design a novel framework, called Multi-View Feature Learning (MVFL). Concretely, the proposed MVFL method combines the regression-like objective function with discriminative regularization (utilizing training label information) to formulate a unified learning framework, where multiple view-specific projections (transformations) can be obtained and each projection learns much stronger discriminative features with very low dimensions for each view. The learned model has a good performance on test data for doing multi-view clustering.

The contributions of the proposed work can be summarized as follows. Firstly, as a feature learning framework, the proposed method simultaneously preserves the certain property of each view in its corresponding feature space and considers the view-wise consistency of multiple views via the discriminative regularization. The learned model realizes the integration of heterogeneous features and makes them complement with each other, which can generate a good performance in subsequent multi-view clustering. Secondly, the proposed method obtains multiple feature learning projections (transformations) for different views regularized by label information on training data, which has stronger discriminative ability to capture efficient and very low dimensional features, and then does well on test data. Thirdly, the proposed method can be rewritten as a trace optimization problem, which makes the original problem be solved in a closed form by using eigenvalue decomposition. Finally, experimental evaluations on four widely used datasets and comparisons with a number of state-of-the-art multi-view clustering algorithms demonstrate the superiority of the proposed work.

2 Related Work

In the setting of clustering, given n data samples $\{\mathbf{x}_i\}_{i=1}^n$, there is data matrix $\mathbf{X} = [\mathbf{x}_1, \cdots, \mathbf{x}_n] \in \mathbb{R}^{d \times n}$, where each column $\mathbf{x}_i \in \mathbb{R}^d$ is the input vector including all features. For a matrix $\mathbf{W} = [w_{ij}]$, we denote its i-th row as $\mathbf{w}^i$ and its j-th column as $\mathbf{w}_j$.

Previous work [Nie *et al.*, 2009] showed the following regression-like clustering objective, which is equivalent to the Discriminative K-Means [Ye *et al.*, 2008], obtains better results than K-Means or spectral clustering methods:

$$\begin{aligned} &\min_{\mathbf{W},\mathbf{F}} \|\mathbf{X}^T\mathbf{W} + \mathbf{1}_n\mathbf{b}^T - \mathbf{F}\|_F^2 \\ &s.t. \mathbf{F}^T\mathbf{F} = \mathbf{I} \end{aligned} \tag{1}$$

where $\mathbf{b} \in \mathbb{R}^{c \times 1}$ is the intercept vector, $\mathbf{1}_n$ is $n \times 1$ constant vector of all 1's, $\mathbf{F} = [\mathbf{f}_1, \cdots, \mathbf{f}_n]^T \in \mathbb{R}^{n \times c}$ is the cluster indicator matrix, and $\mathbf{f}_i \in \mathbb{R}^c$ is the cluster indicator vector for data point $\mathbf{x}_i$ with f_{ij} indicating how likely $\mathbf{x}_i$ belongs to the j-th cluster.

Recently, the work [Wang *et al.*, 2013] pointed out that although the traditional K-Means clustering or spectral clustering objectives can be extended for multi-view clustering, many multi-view clustering objectives still only learn one weight for all features from each sample. Thus, [Wang *et al.*, 2013] designed proper regularizers and learned the weight for each feature to capture the feature-wise importance. Its objective function is as follows:

$$\begin{aligned} &\min_{\mathbf{W},\mathbf{F}} \|\mathbf{X}^T\mathbf{W} + \mathbf{1}_n\mathbf{b}^T - \mathbf{F}\|_F^2 + \gamma_1\|\mathbf{W}\|_{G_1} + \gamma_2\|\mathbf{W}\|_{2,1} \\ &s.t. \mathbf{F}^T\mathbf{F} = \mathbf{I} \end{aligned} \tag{2}$$

where $\mathbf{x}_i \in \mathbb{R}^d$ is the input vector including all features from a total of K views and each view j has d_j features such that $d = \sum_{j=1}^k d_j$. Upon solution, [Wang *et al.*, 2013] learned the parameter matrix $\mathbf{W} = [\mathbf{w}_1^1, \cdots, \mathbf{w}_c^1; \cdots, \cdots, \cdots; \mathbf{w}_1^k, \cdots, \mathbf{w}_c^k] \in \mathbb{R}^{d \times c}$, where $\mathbf{w}_p^q \in \mathbb{R}^{d_q}$ indicated the weights of all features in the q-th view with respect to the p-th. $\|\mathbf{W}\|_{G1}$ and $\|\mathbf{W}\|_{2,1}$ were defined as the group ℓ_1-norm and the $\ell_{2,1}$-norm, respectively, where $\|\mathbf{W}\|_{G1} = \sum_{i=1}^c \sum_{j=1}^k \|\mathbf{w}_i^j\|_2$ and $\|\mathbf{W}\|_{2,1} = \sum_{j=1}^d \|\mathbf{w}^j\|$.

Although this recent work involves the interrelations among multi-view features, concatenating heterogeneous features (multiple views) directly ignores the view-specific information which has the certain property and denotes the specific physical significance and statistical property of each view. Thus it may lose their corresponding certain properties. Besides, this work is unsupervised, which may be lack of discriminative ability because of ignoring label information. Furthermore, the solution of problem (2) is not a closed form by iterative method and easily runs into the local extreme. In order to address above drawbacks, this paper proposes a novel multi-view framework to learn multiple feature weighting matrices (feature learning models) in a supervised way, which not only maintains the relative independence on heterogeneous view-specific properties but also keeps the consistency of multiple views through discriminative information.

[2] http://www.cvpapers.com/datasets.html

[3] https://archive.ics.uci.edu/ml/datasets.html

3 The Proposed Method

In this section, we propose a novel supervised multi-view learning framework, which can be obtained by exploring some heterogeneous training data, and then use this framework to learn multiple discriminative and low dimensional features by testing data, and further to improve multi-view clustering.

3.1 Formulation

Inspired by problem (1), we construct a multi-view framework from a new perspective, which considers the heterogeneous features from both view-wise and individual viewpoints. Thus, our framework can be formulated as follows:

$$\min_{\substack{\mathbf{W},\mathbf{b},\\ \mathbf{C},\mathbf{Z}}}\sum_{i=1}^{K}\|\mathbf{X}_i^T\mathbf{W}_i+\mathbf{1}_n\mathbf{b}_i^T-\mathbf{Z}_i\|_F^2+\gamma\|[\mathbf{Z}_1,\cdots,\mathbf{Z}_K]-\mathbf{Y}\mathbf{C}\|_F^2$$
$$s.t.\mathbf{Z}_i^T\mathbf{Z}_i=\mathbf{I},\mathbf{Z}_i^T\mathbf{1}=\mathbf{0},i=1,\cdots,K \tag{3}$$

where the index i of variable means the i-th view and there are K views. $\mathbf{X}_i\in\mathbb{R}^{d_i\times n}$ is the data matrix, $\mathbf{W}_i\in\mathbb{R}^{d_i\times m_i}$ transforms original features from d_i-dimension to m_i-dimension in each view, $\mathbf{b}_i\in\mathbb{R}^{m_i\times 1}$ is the intercept vector, $\mathbf{Z}_i\in\mathbb{R}^{n\times m_i}$ denotes the learned features of each view, $\mathbf{C}$ is defined as the cluster centroid in multiple discriminative feature spaces, and $\mathbf{Y}\in\mathbb{R}^{n\times c}$ is the cluster indicator matrix of all views. Because problem (3) is a supervised model, $\mathbf{Y}$ essentially is the label matrix, where each row of $\mathbf{Y}$ denotes the label vector of each sample and its c-th element is 1 and other elements are 0 if the sample belongs to the c-th class.

In this proposed model, we can achieve two goals. On one hand, for the residual term, we use current learned features $\mathbf{Z}_i$ in the i-th discriminative subspace to approximate the regression-like objective $\mathbf{X}_i^T\mathbf{W}_i+\mathbf{1}_n\mathbf{b}_i^T$. This maintains their own specific properties of different views. On the other hand, for the regularized term, we utilize the label information $\mathbf{Y}$ to keep the consistency of clustering results of different views. This regularization essentially is like a K-Means objective but the labeling knowledge is available here, which makes multiple feature weighting matrices (models) more discriminative (calculated by multiple discriminative features $[\mathbf{Z}_1,\cdots,\mathbf{Z}_K]$ of training data), and can improve the performance of subsequent multi-view clustering.

3.2 Optimization

The problem (3) comprises two constraints for $\mathbf{Z}_i$, which requires $\mathbf{Z}_i$ to be orthogonal and requires the sum of each column of $\mathbf{Z}_i$ to be zero. Thus, directly solving problem (3) is difficult and we need a series of mathematical transformations as follows.

Theorem 1. *Solving problem (3) is equivalent to solving the following objective function:*

$$\min_{\mathbf{Z}}\sum_{i=1}^{K}Tr(\mathbf{Z}_i^T\mathbf{M}_i\mathbf{Z}_i)+\gamma Tr\big([\mathbf{Z}_1,\cdots,\mathbf{Z}_K]^T\mathbf{N}[\mathbf{Z}_1,\cdots,\mathbf{Z}_K]\big)$$
$$s.t.\mathbf{Z}_i^T\mathbf{Z}_i=\mathbf{I},\mathbf{Z}_i^T\mathbf{1}=\mathbf{0},i=1,\cdots,K \tag{4}$$

where $\mathbf{H}=\mathbf{I}-\frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T$ is the centering matrix and $\mathbf{M}_i=\mathbf{H}\mathbf{X}_i^T(\mathbf{X}_i\mathbf{H}\mathbf{X}_i^T)^{-1}\mathbf{X}_i\mathbf{H}-\mathbf{H}$, and $\mathbf{N}=\mathbf{I}-\mathbf{Y}(\mathbf{Y}^T\mathbf{Y})^{-1}\mathbf{Y}^T$ are *Laplacian*-like matrices.

Proof. Using the properties of matrix trace, the objective of problem (3) can be rewritten as follows:

$$\begin{aligned}&\sum_{i=1}^{K}\big[Tr(\mathbf{W}_i^T\mathbf{X}_i\mathbf{X}_i^T\mathbf{W}_i)+2Tr(\mathbf{b}_i^T\mathbf{W}_i^T\mathbf{X}_i\mathbf{1}_n)\\&-2Tr(\mathbf{W}_i^T\mathbf{X}_i\mathbf{Z}_i)-2Tr(\mathbf{b}_i^T\mathbf{Z}_i^T\mathbf{1}_n)+Tr(\mathbf{b}_i\mathbf{1}_n^T\mathbf{1}_n\mathbf{b}_i^T)\big]\\&+\gamma\big[Tr\big([\mathbf{Z}_1,\cdots,\mathbf{Z}_K]^T[\mathbf{Z}_1,\cdots,\mathbf{Z}_K]\big)\\&-2Tr\big(\mathbf{C}^T\mathbf{Y}^T[\mathbf{Z}_1,\cdots,\mathbf{Z}_K]\big)+Tr\left(\mathbf{C}^T\mathbf{Y}^T\mathbf{Y}\mathbf{C}\right)\big]\end{aligned} \tag{5}$$

Due to solving the minimum, we get the derivative of Eq.(5) with respect to $\mathbf{b}_i$ and $\mathbf{C}$, respectively. Ignoring irrelevant terms and using the rules of matrix derivative, there is:

$$\begin{cases}\mathbf{W}_i^T\mathbf{X}_i\mathbf{1}_n-\mathbf{Z}_i^T\mathbf{1}_n+\mathbf{b}_i\mathbf{1}_n\mathbf{1}_n^T=0\\ \mathbf{Y}^T[\mathbf{Z}_1,\cdots,\mathbf{Z}_K]-\mathbf{Y}^T\mathbf{Y}\mathbf{C}=0\end{cases} \tag{6}$$

$$\Leftrightarrow\begin{cases}\mathbf{b}_i=\frac{1}{n}(-\mathbf{W}_i^T\mathbf{X}_i\mathbf{1}_n+\mathbf{Z}_i^T\mathbf{1}_n)\\ \mathbf{C}=(\mathbf{Y}^T\mathbf{Y})^{-1}\mathbf{Y}^T[\mathbf{Z}_1,\cdots,\mathbf{Z}_K]\end{cases} \tag{7}$$

where $\mathbf{C}$ is the cluster centroid of multiple discriminative feature spaces. Through Eq.(7), the objective of problem (3) becomes:

$$\sum_{i=1}^{K}\|\mathbf{H}(\mathbf{X}_i^T\mathbf{W}_i-\mathbf{Z}_i)\|_F^2+\gamma Tr\big([\mathbf{Z}_1,\cdots,\mathbf{Z}_K]^T\mathbf{N}[\mathbf{Z}_1,\cdots,\mathbf{Z}_K]\big) \tag{8}$$

where $\mathbf{H}=\mathbf{I}-\frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T$ is the centering matrix and $\mathbf{N}=\mathbf{I}-\mathbf{Y}(\mathbf{Y}^T\mathbf{Y})^{-1}\mathbf{Y}^T$ are *Laplacian*-like matrices.

Along the similar way, taking derivative of Eq.(8) with respect to $\mathbf{W}_i$, we have:

$$\mathbf{W}_i=(\mathbf{X}_i\mathbf{H}\mathbf{X}_i^T)^{-1}\mathbf{X}_i\mathbf{H}\mathbf{Z}_i \tag{9}$$

Substituting $\mathbf{W}_i$ into Eq.(8), we can achieve problem (4) and complete the proof. □

In the analysis of transforming problem (3) to problem (4), it can be seen that $\mathbf{W}_i$ can be solved completely depending on $\mathbf{Z}_i$ and $\mathbf{H}$. Thus, solving $\mathbf{W}_i$ in problem (3) can be transformed into solving $\mathbf{Z}_i$ in problem (4) which has only one variable $\mathbf{Z}_i$. In addition, problem (4) contains not only a data-driven term of $\mathbf{Z}_i$ but also a regularization term of $\mathbf{Z}_i$. So, the proposed method will not lead to an over-fitting result and its solution can be calculated in a closed form by eigen decomposition to avoid running into the local extreme.

Furthermore, according to the optimization theory, it is known that $\min[A+B]$ equals to $\min[A]+\min[B]$. Thus, problem (4) can be decoupled to the following problem:

$$\min_{\mathbf{Z}_i}Tr\big(\mathbf{Z}_i^T(\mathbf{M}_i+\gamma\mathbf{N})\mathbf{Z}_i\big)$$
$$s.t.\mathbf{Z}_i^T\mathbf{Z}_i=\mathbf{I},\mathbf{Z}_i^T\mathbf{1}=\mathbf{0} \tag{10}$$

Algorithm 1 : The algorithm for solving MVFL.

```
Input: Training set
Multi-view data $\mathbf{X}=\{\mathbf{X}_i\}_{i=1}^K$, $\mathbf{X}_i \in \mathbb{R}^{d_i \times n}$, label information $\mathbf{Y} \in \mathbb{R}^{n \times c}$, the number of clusters $c$, the reduced dimension $m_k$ and parameter$\lambda$.
for $i \leftarrow 1, K$ do
    Learn features $\mathbf{Z}_i \in \mathbb{R}^{n \times m_i}$ of each view by solving problem (11).
    Calculate $\mathbf{W}_i$,$\mathbf{b}_i$ and $\mathbf{C}$ using Eq.(9) and Eq.(7).
end for
Output: Trained feature learning model $\{\mathbf{W}_i,\mathbf{b}_i\}_{i=1}^K$.
Use trained model on testing set $\tilde{\mathbf{X}}=\{\tilde{\mathbf{X}}_i\}_{i=1}^K$, and obtain testing features $\tilde{\mathbf{Z}}=\{\tilde{\mathbf{Z}}_i\}_{i=1}^K$.
Perform multi-view K-Means clustering on $\tilde{\mathbf{Z}}$ to evaluate the accuracy of trained feature learning model $\{\mathbf{W}_i,\mathbf{b}_i\}_{i=1}^K$.
```

It can be seen that the problem (10) is equivalent to the following problem for a large enough value of λ:

$$\begin{aligned}&\min_{\mathbf{Z}_i} Tr\left(\mathbf{Z}_i^T(\mathbf{M}_i+\gamma\mathbf{N})\mathbf{Z}_i\right)+\lambda Tr(\mathbf{Z}_i^T\mathbf{1}\mathbf{1}^T\mathbf{Z}_i)\\ &s.t. \mathbf{Z}_i^T\mathbf{Z}_i=\mathbf{I}\end{aligned} \tag{11}$$

When parameter λ is large enough, for every i, the optimal solution $\mathbf{Z}_i$ to the problem (11) will make the second term $Tr(\mathbf{Z}_i^T\mathbf{1}\mathbf{1}^T\mathbf{Z}_i)$ be zero, and thus the constraint $\mathbf{Z}_i^T\mathbf{1}=0$ in problem (10) could be satisfied.

Compared with the original problem (10), the problem (11) is much easier to be solved. For a large enough value of λ, the optimal solution $\mathbf{Z}_i$ of problem (11) is formed by m_i eigenvectors of $\mathbf{M}_i+\gamma\mathbf{N}+\lambda\mathbf{1}\mathbf{1}^T$ corresponding to m_i smallest eigenvalues. Besides, because of $(\mathbf{M}_i+\gamma\mathbf{N})\mathbf{1}=\mathbf{0}$, m_i smallest eigenvectors of $\mathbf{M}_i+\gamma\mathbf{N}+\lambda\mathbf{1}\mathbf{1}^T$ from the first to the m_i-th are actually m_i smallest eigenvectors of $\mathbf{M}_i+\gamma\mathbf{N}$ from the second to the m_i+1-th.

Thus, we further calculate $\mathbf{W}_i$, $\mathbf{b}_i$ and $\mathbf{C}$ according to Eq.(9) and Eq.(7), respectively, which makes us obtain K feature learning models for K views.

To sum up, the whole procedure of our work can be described as follows. Given the training data $\{\mathbf{X}_i\}_{i=1}^K$ and its corresponding label $\mathbf{Y}$, we can obtain the feature learning model for K views, i.e., feature weighting matrix $\{\mathbf{W}_i\}_{i=1}^K$ and intercept vector $\{\mathbf{b}_i\}_{i=1}^K$. Such model will be evaluated on testing data. That is to say, we use the trained model to learn some low dimensional features based on testing data and utilize the multi-view K-Means clustering to verify the effectiveness of the trained model. Concretely, we learn new features $\{\tilde{\mathbf{Z}}_i\}_{i=1}^K$ on testing data $\{\tilde{\mathbf{X}}_i\}_{i=1}^K$ using the model $\{\mathbf{W}_i,\mathbf{b}_i\}_{i=1}^K$, and then cluster these features $\{\tilde{\mathbf{Z}}_i\}_{i=1}^K$ by performing multi-view K-Means clustering according to [Xu *et al.*, 2016]. We describe this process in Algorithm 1.

3.3 Convergence Analysis

As mentioned above, it is obvious that problem (3) can be transformed into problem (4) which is further equivalent to problem (11) for a large λ, which makes the original problem be solved by a trace optimization problem and obtains the global optimal solution in a closed form.

Table 1: Descriptions of testing datasets.

View	MSRCv1	Caltech101-7	Handwritten	Yale
1	CENT(1302)	CENT(1302)	FOU(76)	v1(4096)
2	CMT(48)	CMT(48)	FAC(216)	v2(3304)
3	GIST(512)	GIST(512)	KAR(64)	v3(6750)
4	HOG(100)	HOG(100)	PIX(240)	-
5	LBP(256)	LBP(256)	ZER(47)	-
6	SIFT(210)	SIFT(441)	MOR(6)	-
Images	210	441	2000	165
Classes	7	7	10	15

4 Experiments

In this section, we evaluate the proposed framework through clustering task on four widely used datasets in terms of four standard clustering evaluation metrics, namely Accuracy (ACC) [Cai *et al.*, 2005], Normalized Mutual Information (NMI) [Cai *et al.*, 2005], Jaccard Index (Jaccard) [Varshavsky *et al.*, 2005] and Purity [Varshavsky *et al.*, 2005].

4.1 Datasets

In our experiments, by following [Cai *et al.*, 2013b; Cao *et al.*, 2015; Xu *et al.*, 2016], four datasets including MSRCv1 [Lee and Grauman, 2009; Cai *et al.*, 2013b], Caltech101-7 [Dueck and Frey, 2007], Handwritten [Asuncion and Newman, 2007] and Yale [Cao *et al.*, 2015] datasets are adopted for evaluations. Tabel 1 summarizes the information of each dataset including the number of samples and classes, heterogeneous features and the dimensionality of each type of feature in the parenthese. Besides, we need to normalize and center all data values before experiments.

4.2 Experiment Setup

Following [Cai *et al.*, 2013b; Xu *et al.*, 2016], firstly, we applied the proposed MVFL (Multi-view Feature Learning) on each type of features to form a baseline SVFL (Single-view Feature Learning) to demonstrate that the multi-view method MVFL has advantages of multi-view with respect to the single-view method SFL. Next, we compared MVFL with AVFL (All-view Feature Learning), where AVFL concatenates all heterogeneous features directly as one view and performs the proposed feature learning algorithm. However, AVFL dose not consider specific properties of different views and relationships between them. Thus, AVFL is not as good as the proposed MVFL method.

Furthermore, we compared MVFL method with some state-of-the-art multi-view clustering methods: NMVKM (Naive Multi-view K-Means), RMVKM (Robust Multi-view K-Means) [Cai *et al.*, 2013b] and DEKM (Discriminatively Embedded K-Means) [Xu *et al.*, 2016]. Besides, we compared our method with MvDA (Multi-view Discriminant Analysis) [Kan *et al.*, 2012] which projects multi-view data

Table 2: Comparison of MVFL and SVFL on MSRCv1 and Caltech101-7 datasets.

Method	MSRCv1				Caltech101-7			
	ACC	NMI	Jaccard	Purity	ACC	NMI	Jaccard	Purity
SVFL(view1)	0.7905	0.6934	0.4930	0.6479	0.8778	0.7893	0.6893	0.8181
SVFL(view2)	0.6381	0.5343	0.3010	0.4144	0.4977	0.3564	0.2077	0.3220
SVFL(view3)	0.8571	0.7581	0.5789	0.7139	0.8100	0.6497	0.5432	0.7179
SVFL(view4)	0.3429	0.2059	0.1148	0.1807	0.6742	0.4708	0.3181	0.4443
SVFL(view5)	0.6381	0.4864	0.2761	0.4142	0.5928	0.4432	0.3307	0.4283
SVFL(view6)	0.5048	0.3673	0.1979	0.2941	0.7240	0.5853	0.4639	0.6647
MVFL	**0.9619**	**0.9274**	**0.8541**	**0.9188**	**0.9367**	**0.8747**	**0.8359**	**0.9143**

Table 3: Comparison of MVFL and SVFL on Handwritten and Yale datasets.

Method	Handwritten				Yale			
	ACC	NMI	Jaccard	Purity	ACC	NMI	Jaccard	Purity
SVFL(view1)	0.7110	0.6219	0.4013	0.5656	0.7667	0.8343	0.4844	0.6200
SVFL(view2)	0.9750	0.9434	0.9054	0.9500	0.9000	0.8948	0.6307	0.7449
SVFL(view3)	0.9200	0.8444	0.7374	0.8473	0.9111	0.9168	0.7041	0.8174
SVFL(view4)	0.9095	0.8185	0.6966	0.8196	-	-	-	-
SVFL(view5)	0.7850	0.7159	0.5309	0.6797	-	-	-	-
SVFL(view6)	0.5600	0.6370	0.3237	0.4001	-	-	-	-
MVFL	**0.9870**	**0.9704**	**0.9494**	**0.9738**	**0.9667**	**0.9654**	**0.8607**	**0.9170**

Table 4: Clustering results of compared methods on MSRCv1 and Caltech101-7 datasets.

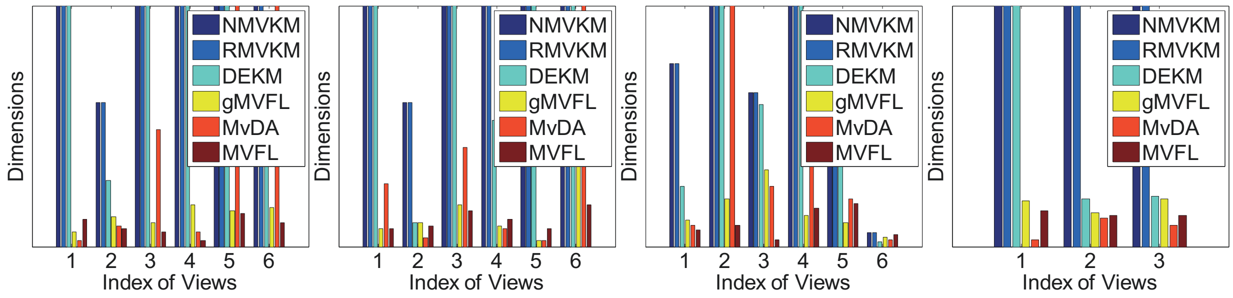
Method	MSRCv1				Caltech101-7			
	ACC	NMI	Jaccard	Purity	ACC	NMI	Jaccard	Purity
NMVKM	0.7810	0.7122	0.4356	0.5737	0.7143	0.7337	0.5575	0.7195
RMVKM	0.9048	0.8463	0.6301	0.7412	0.7846	0.7145	0.6065	0.7516
DEKM	0.9238	0.8649	0.7477	0.8471	0.8503	0.8231	0.7553	0.8624
MvDA	0.5524	0.4559	0.2656	0.3869	0.7376	0.5548	0.4697	0.6479
AVFL	0.9238	0.8787	0.7432	0.8470	0.8552	0.8135	0.7641	0.8773
gMVFL	0.9429	0.8917	0.7934	0.8812	0.8733	0.8285	0.7462	0.8471
MVFL	**0.9619**	**0.9274**	**0.8541**	**0.9188**	**0.9367**	**0.8747**	**0.8359**	**0.9143**

Table 5: Clustering results of compared methods on Handwritten and Yale datasets.

Method	Handwritten				Yale			
	ACC	NMI	Jaccard	Purity	ACC	NMI	Jaccard	Purity
NMVKM	0.7810	0.7661	0.4927	0.6290	0.4606	0.4990	0.1659	0.2408
RMVKM	0.9125	0.8539	0.6020	0.7014	0.6000	0.6377	0.2689	0.3729
DEKM	0.9530	0.9080	0.8333	0.9069	0.5576	0.6107	0.2500	0.3719
MvDA	0.9790	0.9536	0.9199	0.9579	0.8889	0.8946	0.6439	0.7716
AVFL	0.9710	0.9368	0.8913	0.9420	0.9111	0.9080	0.6752	0.7906
gMVFL	0.9520	0.9089	0.8289	0.9033	0.8111	0.8510	0.5294	0.6667
MVFL	**0.9870**	**0.9704**	**0.9494**	**0.9738**	**0.9667**	**0.9654**	**0.8607**	**0.9170**

into a discriminative common space by using multiple transformations and performs multi-view K-Means clustering on this space.

Finally, we performed a simple version of the proposed MVFL method, gMVFL ($\gamma = 0$), to evaluate the effectiveness of its regularized term in multi-view framework. Obviously, we only use the first term of problem (3), which is equivalent to linear regression to learn features, and then performed multi-view K-Means clustering on these learned features.



(a) MSRCv1 (b) Caltech101-7 (c) Handwritten (d) Yale

Figure 1: Dimensions of different views on (a) MSRCv1, (b) Caltech101-7, (c) Handwritten and (d) Yale datasets.

4.3 Experiment Results

The comparison results measured by ACC, NMI, Jaccard and Purity are reported in Tables 2 to 5, respectively. For these metrics, the higher value indicates better clustering quality. Each metric penalizes or favors different properties in the clustering, and hence we report results on these diverse measures to perform a comprehensive evaluation.

The experimental results, in Tables 2 and 3, show that the proposed method (MVFL) performs generally better than that on each view (SVFL), which validates the superiority of heterogeneous features integration. For example, for MSRCv1 and Caltech101-7 datasets shown in Table 2, the results of SVFL method (for each view) are lower than that of MVFL method. Because in multi-view setting, each view of the data may contain some knowledge that other views do not have. It is straightforward to demonstrate the superiority of multi-view.

In Tables 4 and 5, MVFL method on multiple views outperforms AVFL on the concatenation of all views. Taking Handwritten dataset as an example, the performance of AVFL is not as good as that of MVFL. Because MVFL not only preserves the certain property of each view but also considers the intercoordinations of different views while the simple features concatenation does not have such capability. Sometimes, AVFL is unable to appropriately cope with multiple views, and may even degrade the performance of multi-view clustering compared with some single views.

Moreover, compared with other four multi-view methods (NMVKM, RMVKM, DEKM and MvDA), the proposed MVFL method can achieve the best performance. This is consistent with our theoretical analysis in the above sections. Concretely, compared with NMVKM and RMVKM, MVFL method obtains the low dimensional and discriminative learning model by feature weighting matrices (projections), such as the results of Handwritten dataset in Table 5. In addition, although DEKM also can achieve feature projections using dimensionality reduction, without help of training data label information, the learned features of DEKM are lack of discriminative ability. Besides, compared with MvDA which is a supervised method, MVFL obtains better results on different datasets. It is because although MvDA utilizes the label information in training stage, its original objective function dose not have the closed form and its approximate objective function is inexact, which affects the performance of MvDA to some extent. In contract, MVFL can be solved analytically in a closed form. Finally, we implement a simple version of the proposed method, i.e., gMVFL, which only has a regression-like objective function. Actually, gMVFL is actually an unsupervised method and its performance is degenerated without the discriminative regularization.

4.4 Parameters Setup

For comparison methods, we performed them according to their corresponding original works and selected their optimal parameters by using grid search. For the proposed method, we set the regularized parameter $\gamma = 1$ in (3), and tuned the dimension parameters $\{m_i\}_{i=1}^{K}(m_i < \min(d_i, n))$ heuristically by searching the grid with proper step-size. Besides, if $\mathbf{X}_i\mathbf{H}\mathbf{X}_i^T$ is nearly singular, we can regularize it as $\mathbf{X}_i\mathbf{H}\mathbf{X}_i^T + \epsilon\mathbf{I}_{d_i}$ by introducing a small perturbation $\epsilon(\epsilon = 10^{-4})$. It is demonstrated that if $\epsilon \to 0$, minimizing the perturbation version is reduced to problem (3). Besides, we compared total dimensions of different multi-view methods (NMVKM, RMVKM, DEKM, gMVFL, MvDA and MVFL) which is shown in Figure 1. It is observed that the proposed MVFL method only learns a few of discriminative features to achieve satisfactory experimental results.

5 Conclusion

In this paper, we proposed a novel supervised multi-view learning framework, which efficiently learns multiple discriminative feature weighting matrices for different views with the help of label information. In this framework, the regression-like objective and discriminative regularization were utilized to yield multiple feature learning models with stronger discriminative abilities, and then learned low dimensional features for subsequent processing. Besides, our method can be transformed into a trace optimization problem, which obtains the global solution in a closed form. Comparisons with existing state-of-the-art multi-view clustering methods, our framework captured the efficient features and achieved better performance in multi-view clustering tasks.

Acknowledgments

This work was supported in part by the National Science Foundation of China under Grants 61522207 and 61473231.

References

[Arora and Livescu, 2014] Raman Arora and Karen Livescu. Multi-view learning with supervision for transformed bottleneck features. In *ICASSP*, pages 2499–2503, 2014.

[Asuncion and Newman, 2007] Arthur Asuncion and David Newman. Uci machine learning repository, 2007.

[Cai *et al.*, 2005] Deng Cai, Xiaofei He, and Jiawei Han. Document clustering using locality preserving indexing. *TKDE*, 17(12):1624–1637, 2005.

[Cai *et al.*, 2011] Xiao Cai, Feiping Nie, Heng Huang, and Farhad Kamangar. Heterogeneous image feature integration via multi-modal spectral clustering. In *CVPR*, pages 1977–1984, 2011.

[Cai *et al.*, 2013a] Xiao Cai, Feiping Nie, Weidong Cai, and Heng Huang. Heterogeneous image features integration via multi-modal semi-supervised learning model. In *ICCV*, pages 1737–1744, 2013.

[Cai *et al.*, 2013b] Xiao Cai, Feiping Nie, and Heng Huang. Multi-view k-means clustering on big data. In *IJCAI*, pages 2598–2604, 2013.

[Cao *et al.*, 2015] Xiaochun Cao, Changqing Zhang, Huazhu Fu, Si Liu, and Hua Zhang. Diversity-induced multi-view subspace clustering. In *CVPR*, pages 586–594, 2015.

[Chaudhuri *et al.*, 2009] Kamalika Chaudhuri, Sham M Kakade, Karen Livescu, and Karthik Sridharan. Multi-view clustering via canonical correlation analysis. In *ICML*, pages 129–136, 2009.

[Dalal and Triggs, 2005] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *CVPR*, pages 886–893, 2005.

[Dueck and Frey, 2007] Delbert Dueck and Brendan J Frey. Non-metric affinity propagation for unsupervised image categorization. In *ICCV*, pages 1–8, 2007.

[Eaton *et al.*, 2010] Eric Eaton, Marie Desjardins, and Sara Jacob. Multi-view clustering with constraint propagation for learning with an incomplete mapping between views. In *CIKM*, pages 389–398, 2010.

[Han *et al.*, 2012] Yahong Han, Fei Wu, Dacheng Tao, Jian Shao, Yueting Zhuang, and Jianmin Jiang. Sparse unsupervised dimensionality reduction for multiple view data. *TCSVT*, 22(10):1485–1496, 2012.

[Jiang *et al.*, 2012] Yu Jiang, Jing Liu, Zechao Li, Peng Li, and Hanqing Lu. Co-regularized plsa for multi-view clustering. In *ACCV*, pages 202–213, 2012.

[Kan *et al.*, 2012] Meina Kan, Shiguang Shan, Haihong Zhang, Shihong Lao, and Xilin Chen. Multi-view discriminant analysis. In *ECCV*, pages 808–821, 2012.

[Kumar *et al.*, 2011] Abhishek Kumar, Piyush Rai, and Hal Daume. Co-regularized multi-view spectral clustering. In *NIPS*, pages 1413–1421, 2011.

[Lee and Grauman, 2009] Yong Jae Lee and Kristen Grauman. Foreground focus: Unsupervised learning from partially matching images. *IJCV*, 85(2):143–166, 2009.

[Lowe, 2004] David G Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004.

[Nie *et al.*, 2009] Feiping Nie, Dong Xu, Ivor W Tsang, and Changshui Zhang. Spectral embedded clustering. In *IJCAI*, pages 1181–1186, 2009.

[Ojala *et al.*, 2002] Timo Ojala, Matti Pietikainen, and Topi Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *TPAMI*, 24(7):971–987, 2002.

[Oliva and Torralba, 2001] Aude Oliva and Antonio Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *IJCV*, 42(3):145–175, 2001.

[Tang *et al.*, 2013] Jiliang Tang, Xia Hu, Huiji Gao, and Huan Liu. Unsupervised feature selection for multi-view data in social media. In *SDM*, pages 270–278, 2013.

[Varshavsky *et al.*, 2005] Roy Varshavsky, Michal Linial, and David Horn. Compact: A comparative package for clustering assessment. In *ISPDPA*, pages 159–167, 2005.

[Wang *et al.*, 2007] Huan Wang, Shuicheng Yan, Dong Xu, Xiaoou Tang, and Thomas Huang. Trace ratio vs. ratio trace for dimensionality reduction. In *CVPR*, pages 1–8, 2007.

[Wang *et al.*, 2013] Hua Wang, Feiping Nie, and Heng Huang. Multi-view clustering and feature learning via structured sparsity. In *ICML*, pages 352–360, 2013.

[Wang *et al.*, 2014] Hongxing Wang, Chaoqun Weng, and Junsong Yuan. Multi-feature spectral clustering with minimax optimization. In *CVPR*, pages 4106–4113, 2014.

[Wu and Rehg, 2008] Jianixn Wu and James M. Rehg. Where am i: Place instance and category recognition using spatial pact. In *CVPR*, pages 1–8, 2008.

[Xu *et al.*, 2016] Jinglin Xu, Junwei Han, and Feiping Nie. Discriminatively embedded k-means for multi-view clustering. In *CVPR*, pages 5356–5364, 2016.

[Ye *et al.*, 2008] Jieping Ye, Zheng Zhao, and Mingrui Wu. Discriminative k-means for clustering. In *NIPS*, pages 1649–1656, 2008.

[Yu *et al.*, 2002] Hui Yu, Mingjing Li, Hong-Jiang Zhang, and Jufu Feng. Color texture moments for content-based image retrieval. In *ICIP*, pages 929–932, 2002.

[Zhao *et al.*, 2014] Xuran Zhao, Nicholas Evans, and Jean-Luc Dugelay. A subspace co-training framework for multi-view clustering. *PRL*, 41:73–82, 2014.