Cross-View Local Structure Preserved Diversity and Consensus Learning for Multi-View Unsupervised Feature Selection

Chang Tang,[1] **Xinzhong Zhu,**[2,3] **Xinwang Liu,**[4] **Lizhe Wang**[1]
[1]School of Computer Science, China University of Geosciences, Wuhan 430074, China
[2]College of Mathematics, Physics and Information Engineering, Zhejiang Normal University, Jinhua 321004, China
[3]Research Institute of Ningbo Cixing Co., Ltd, Ningbo 315336, China
[4]School of Computer Science, National University of Defense Technology, Changsha 410073, China
tangchang@cug.edu.cn, zxz@zjnu.edu.cn, xinwangliu@nudt.edu.cn, lizhe.wang@gmail.com

Abstract

Multi-view unsupervised feature selection (MV-UFS) aims to select a feature subset from multi-view data without using the labels of samples. However, we observe that existing MV-UFS algorithms do not well consider the local structure of cross views and the diversity of different views, which could adversely affect the performance of subsequent learning tasks. In this paper, we propose a cross-view local structure preserved diversity and consensus semantic learning model for MV-UFS, termed CRV-DCL briefly, to address these issues. Specifically, we project each view of data into a common semantic label space which is composed of a consensus part and a diversity part, with the aim to capture both the common information and distinguishing knowledge across different views. Further, an inter-view similarity graph between each pairwise view and an intra-view similarity graph of each view are respectively constructed to preserve the local structure of data in different views and different samples in the same view. An $l_{2,1}$-norm constraint is imposed on the feature projection matrix to select discriminative features. We carefully design an efficient algorithm with convergence guarantee to solve the resultant optimization problem. Extensive experimental study is conducted on six publicly real multi-view datasets and the experimental results well demonstrate the effectiveness of CRV-DCL.

Introduction

With the rapid development of data acquisition sensors and data processing technologies, data are usually represented by various feature descriptors. For an instance, in image/video processing, different visual descriptors such as Local Binary Patterns (LBP) (Ojala, Pietikainen, and Maenpaa 2002), Scale Invariant Feature Transform (SIFT) (Lowe and Lowe 2004) and Histogram of Oriented Gradient (HOG) (Dalal and Triggs 2005) are often used to describe each image/video frame from different views. In biomedical research, both the chemical structure and chemical response in different cells can be used to represent a certain drug, while the sequence and gene expression values can represent a certain protein in different aspects (Li 2014; Li and Cai 2017). In general, data in these applications is termed multi-view data in data mining and machine learning communities (Liu et al. 2016; 2018; Zhang et al. 2018). In order to process the multi-view data, many multi-view learning methods have been put forward (Wang et al. 2018; Tang et al. 2018d). Among these algorithms, multi-view unsupervised feature selection (MV-UFS), which aims to select a feature subset from multi-view data without using the labels of samples, has obtained more and more attention since different views of data are usually with high dimensionality and processing these data is confronted with the curse of dimensionality problem (Friedman 1997). In addition, it is a challenging and laborious task to obtain the labels from large number of data instances.

In the past few years, a variety of MV-UFS methods have been proposed and these methods can be mainly grouped into two categories. The first category combines multiple features from different views into a single one and then applies traditional single-view unsupervised feature selection methods, including Laplacian score (He, Cai, and Niyogi 2005), trace ratio (Nie et al. 2008), spectral feature selection (Zhao and Liu 2007), minimum redundancy spectral feature selection (Zhao, Wang, and Liu 2010) and data representation (Zhu et al. 2017), into the concentrated data. This kind of methods does not well exploit the underlying correlations between different views. Instead of concentrating different views, the other category of MV-UFS methods aim to tackle multi-view data directly, and they often excavate the diversity and complementary information to promote the feature selection performance. Typical methods in this class include Adaptive Multi-View Feature Selection (AMFS) method (Wang et al. 2016), Adaptive Unsupervised Multi-View Feature Selection (AUMFS) (Feng et al. 2012), Robust Multi-View Feature Selection (RMFS) (Liu, Mao, and Fu 2017), Adaptive Similarity and View Weight (ASVW) learning for Multi-View Feature Selection (Hou et al. 2017) and Consensus Learning Guided Multi-view Unsupervised Feature Selection (CGMV-UFS) (Tang et al. 2018a). Since the diversity and complementary information are important for multi-view learning, MV-UFS methods in the second class often perform better than those in the first category. The work in this paper belongs to the second category.

Without labels of data instances, the local property of samples usually acts as a priori to regularize the feature selection process. Therefore, traditional methods usually utilize various similarity graphs to characterize the local geo-

metrical manifold structure of data and then rank the importance of each feature (Zhang et al. 2017). However, previous approaches often construct a similarity graph for each view separately, while the cross view local structure has been ignored. Furthermore, to capture the shared structure of different views, existing methods learn a certain consensus space from which different views are assumed to be projected (Z and J 2015). This does not considerably take the effect of the diversity and noises of different views on the projection into account. To overcome these two issues, in this paper, we propose a cross-view local structure preserved diversity and consensus semantic representation model for MV-UFS, referred to as CRV-DCL briefly. For capturing both the common information and distinguishing specificity of different views, we project each view of original data into a common semantic label space, and we relax this space to a consensus part and a diversity part. In such a way, different feature views are regularized to represent the same samples. Meanwhile, instead of using only an intra-view similarity graph of each view to preserve the local structure of different samples in the same view, we also construct an inter-view similarity graph between any two views to preserve the local structure of a certain sample in different views. The main contributions of this work are summarized as follows:

- We construct an intra-view similarity graph for each individual view and an inter-view similarity graph for each pairwise views to preserve the local structure of data for MV-UFS;
- Instead of projecting each view of data into a single common semantic label space, we relax the projected space into a consensus part and a diversity part. By this way, both the diversity and consensus information of different views can be better exploited;
- An efficient optimization algorithm is carefully designed to solve the proposed model, and the comprehensive experimental results on six publicly benchmark datasets demonstrate the effectiveness of the proposed method.

Related Work

In this section, we give a briefly review about some recent related work on MV-UFS. AMFS (Wang et al. 2016) is a MV-UFS method proposed for human motion retrieval. In AMFS, multiple local feature descriptors are used to represent human motion data. For each view of data, a graph Laplacian matrix is generated, and these view-dependent Laplacian matrices are then linearly combined with weights to exploit complementary information of different views. Finally, trace ratio criteria is deployed to eliminate redundant features. To identify discriminative features, AUMFS (Feng et al. 2012) adopts a robust $l_{2,1}$-norm regularized sparse regression model to project original data into cluster labels. In AUMFS, the $l_{2,1}$-norm is used to impose row sparsity on the projection matrix for measuring feature importance. In addition, the local geometrical structure of data is also preserved by linearly combining view-dependent graph Laplacian matrices with weights. RMFS (Liu, Mao, and Fu 2017) applies robust multi-view k-means to obtain the robust and high quality pseudo labels for sparse feature selection in an efficient way. In RMFS, the pseudo labels are generated by utilizing the heterogeneous information from multiple views. By considering that previous methods such as AMFS and AUMFS ignore the underlying shared structure across different feature views, and the pre-computed similarity matrices are not accurate for characterizing the local structure of data, ASVW (Hou et al. 2017) leverages the learning mechanism to adaptively learn a common similarity matrix shared by different views. Recently, CGMV-UFS (Tang et al. 2018a) constructs a view-dependent graph Laplacian matrix for each view for intra-view local structure preservation to capture both the common and complementary information of different views. Meanwhile, CGMV-UFS learns a common label indicator matrix to regularize that different feature views represent the same samples. However, as aforementioned, almost all of previous methods are confronted with at least two issues, i.e., the cross-view local structure is not taken into consideration and the assumption of projecting multi-view data into a single label space is too strict since there usually are noises and specificity in each single view.

The Proposed CRV-DCL

Notations

Throughout this paper, matrices and vectors are denoted as boldface capital letters and boldface lower case letters, respectively. For an arbitrary matrix $\mathbf{M} \in \mathbb{R}^{m\times n}$, $\mathbf{M}_{ij}$ denotes its (i,j)-th entry, $\mathbf{m}^i$ and $\mathbf{m}_j$ denote its i-th row and j-th column, respectively. $Tr(\mathbf{M})$ is the trace of $\mathbf{M}$ if $\mathbf{M}$ is square and $\mathbf{M}^T$ is the transpose of $\mathbf{M}$. $\mathbf{I}_m$ is the identity matrix with size $m \times m$ (denoted by $\mathbf{I}$ if the size is obviously known). The $l_{2,1}$-norm of matrix $\mathbf{M}$ is defined as $||\mathbf{M}||_{2,1} = \sum_{i=1}^{m} ||\mathbf{m}^i|| = \sum_{i=1}^{m} \sqrt{\sum_{j=1}^{n} \mathbf{M}_{ij}{}^2}$. $||\mathbf{M}||_F = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} \mathbf{M}_{ij}{}^2}$ is the well-known Frobenius norm of $\mathbf{M}$. $||\mathbf{M}||_1 = \sum_{i=1}^{m}\sum_{j=1}^{n} |\mathbf{M}_{ij}|$ represents the l_1-norm of matrix $\mathbf{M}$, i.e., the absolute summation of its entries.

Supposed we have N data samples $\{\mathbf{x}_i\}_{i=1}^{N}$ belonging to c classes, and they are characterized by V views of features, the data matrix is denoted as $\mathbf{X} = [\mathbf{x}_1, \cdots, \mathbf{x}_N] \in \mathbb{R}^{d\times N}$. Let $\mathbf{x}_i^v$ denote the v-th view of the i-th sample, then the complete i-th sample $\mathbf{x}_i = [\mathbf{x}_i^1; \cdots; \mathbf{x}_i^V] \in \mathbb{R}^d$ is composed of features from V views, where the v-th view $\mathbf{x}_i^v \in \mathbb{R}^{d_v}$ has d_v features such that $d = \sum_{v=1}^{V} d_v$. Denote the data matrix of the v-th view as $\mathbf{X}^v = [\mathbf{x}_i^v, \cdots, \mathbf{x}_N^v] \in \mathbb{R}^{d_v\times N}$, then $\mathbf{X} = [\mathbf{X}^v; \cdots; \mathbf{X}^v]$. MV-UFS aims to select the top K discriminative features from those d features without using the labels of data instances.

Formulation of CRV-DCL

Although data consist of multi-view heterogeneous features, they still share the same semantic information. In order to capture this common information, we project different views of features into a common semantic label space, which rep-

resent original data in a relatively higher level manner. Considering that each single view contains both the common information and distinguishing specificity, we relax the common label space to a consensus part and a diversity part, this can be mathematically formulated as follows:

$$\min_{\mathbf{W}^v,\bar{\mathbf{Y}},\mathbf{Y}^v} \sum_{v=1}^{V} \mathcal{L}(\mathbf{X}^v,\mathbf{W}^v,\bar{\mathbf{Y}},\mathbf{Y}^v) + \alpha \sum_{v=1}^{V} \mathcal{R}(\mathbf{W}^v,\mathbf{Y}^v), \quad (1)$$

where $\mathbf{W}^v \in R^{d_v \times c}$ is the projection matrix for the v-th view, $\bar{\mathbf{Y}} \in R^{N\times c}$ and $\mathbf{Y}^v \in R^{N\times c}$ denote the consensus part and the diversity part of the common label space, respectively. Since $\bar{\mathbf{Y}}$ denotes the pure label indicator matrix of data, we constrain it as $\bar{\mathbf{Y}} \in \{0,1\}^{N\times c}$. However, the discrete constraint in Eq. (1) makes it difficult to solve. Instead we adopt the orthogonal constraint, i.e., $\bar{\mathbf{Y}}^T\bar{\mathbf{Y}} = \mathbf{I}, \bar{\mathbf{Y}} \geq 0$. $\mathcal{L}(\mathbf{X}^v,\mathbf{W}^v,\bar{\mathbf{Y}},\mathbf{Y}^v)$ is the projection operator for the v-th view, and $\mathcal{R}(\mathbf{W}^v,\mathbf{Y}^v)$ denotes certain regularization on $\mathbf{W}^v$ and $\mathbf{Y}^v$. α is a positive constant for balancing the two terms. In this work, we also use the regression model to formulate the projection process, which can be written as:

$$\mathcal{L}(\mathbf{X}^v,\mathbf{W}^v,\bar{\mathbf{Y}},\mathbf{Y}^v) = ||(\mathbf{X}^v)^T\mathbf{W}^v - (\bar{\mathbf{Y}}+\mathbf{Y}^v)||_F^2. \quad (2)$$

In Eq. (2), the projected semantic label space is decomposed into a consensus part for capturing the consensus label representation of different views and a diversity part for capturing the distinct diversity of each view. The consensus part represents the true labels of samples, while the diversity part is produced by the specificity and noises contained in each view.

In order to select discriminative features, we impose row sparsity on $\mathbf{W}^v$ by using the $l_{2,1}$-norm regularization. In addition, although each view contains some view-specific information, they still represent the same data, the consensus semantic label representation should be the main part. Therefore, we wish each view contains a small quantity of distinct diversity, which means that the diversity part of the semantic label representation should be sparse. To this end, we impose l_1-norm regularization on $\mathbf{Y}^v$. As a result, $R(\mathbf{W}^v,\mathbf{Y}^v)$ can be formulated as:

$$\mathcal{R}(\mathbf{W}^v,\mathbf{Y}^v) = ||\mathbf{W}^v||_{2,1} + ||\mathbf{Y}^v||_1. \quad (3)$$

Since the local geometrical structure of data works as an crucial priori for unsupervised feature selection (Liu et al. 2014; Nie, Wei, and Li 2016; Tang et al. 2017; 2018c; 2018b). In this work, we also preserve the local geometrical structure of data by constructing an intra-view similarity graph of each view and an intra-view similarity graph between any two views. For different samples in the same v-th view, we constrain that similar data samples should share similar semantic label representation, and this can be regularized by the following formulation:

$$\min_{\mathbf{W}^v} \sum_{i=1}^{n}\sum_{j=1}^{n} \mathbf{S}_{ij}^v ||(\mathbf{W}^v)^T\mathbf{x}_i^v - (\mathbf{W}^v)^T\mathbf{x}_j^v||_2^2, \quad (4)$$

where $\mathbf{S}_{ij}^v \in \mathbb{R}^{N\times N}$ denotes the sample similarity matrix of the i-th view of data, of which the element is calculated by the Gaussian RBF kernel function as:

$$\mathbf{S}_{ij}^v = \begin{cases} \exp(-\frac{||\mathbf{x}_i^v - \mathbf{x}_j^v||_2^2}{2\sigma^2}) & \mathbf{x}_i^v \in \mathcal{N}_k(\mathbf{x}_j^v) \ or \ \mathbf{x}_j^v \in \mathcal{N}_k(\mathbf{x}_i^v) \\ 0, & \text{otherwise} \end{cases}, \quad (5)$$

where $\mathcal{N}_k(\mathbf{x}_i^v)$ is the set of k nearest neighbors of $\mathbf{x}_i^v$ in the v-th view.

For a certain sample in different views, we constrain that its semantic label representations from different views should be consistent, and this can be regularized by the following formulation:

$$\min_{\mathbf{W}^v} \sum_{v=1}^{V}\sum_{u=1,u\neq v}^{V}\sum_{i=1}^{n}\sum_{j=1}^{n} \mathbf{S}_{ij}^{vu} ||(\mathbf{W}^v)^T\mathbf{x}_i^v - (\mathbf{W}^u)^T\mathbf{x}_j^u||_2^2, \quad (6)$$

where $\mathbf{S}_{ij}^{vu} \in \mathbb{R}^{N\times N}$ denotes the cross-view sample similarity matrix of the v-th and u-th views, of which the element is defined as:

$$\mathbf{S}_{ij}^{vu} = \begin{cases} 1, & \mathbf{x}_i^v \ and \ \mathbf{x}_j^u \ belong \ to \ the \ same \ sample \\ 0, & otherwise \end{cases}. \quad (7)$$

Based on the above intra-view and inter-view similarities, we define an overall similarity matrix $\mathbf{S}$ as follows:

$$\mathbf{S} = \begin{bmatrix} \mathbf{S}^1 & \mathbf{S}^{12} & \cdots & \mathbf{S}^{1V} \\ \mathbf{S}^{21} & \mathbf{S}^2 & \cdots & \mathbf{S}^{2V} \\ \vdots & \cdots & \vdots & \\ \mathbf{S}^{V1} & \mathbf{S}^{V2} & \cdots & \mathbf{S}^V \end{bmatrix} \in \mathbb{R}^{VN\times VN}. \quad (8)$$

By some simple algebra and combining Eq. (4) and Eq. (4) with Eq. (8), we have the cross-view local geometrical structure preservation term as follows:

$$\min_{\mathbf{W}^v} \sum_{v=1}^{V}\sum_{u=1}^{V} Tr((\mathbf{W}^v)^T\mathbf{X}^v\mathbf{L}^{vu}(\mathbf{X}^u)^T\mathbf{W}^u), \quad (9)$$

where $\mathbf{L} = \mathbf{D} - \mathbf{S}$ is the cross-view Laplacian matrix and $\mathbf{D}$ is a diagonal matrix with its i-th diagonal entry calculated as the sum of the i-th row in $\mathbf{S}$, i.e., $\mathbf{D}_{ii} = \sum_{j=1}^{VN}\mathbf{S}_{ij}$. Then $\mathbf{L}$ can be written as following form:

$$\mathbf{L} = \begin{bmatrix} \mathbf{L}^{11} & \mathbf{L}^{12} & \cdots & \mathbf{L}^{1V} \\ \mathbf{L}^{21} & \mathbf{L}^{22} & \cdots & \mathbf{L}^{2V} \\ \vdots & \cdots & \vdots & \\ \mathbf{L}^{V1} & \mathbf{L}^{V2} & \cdots & \mathbf{L}^{VV} \end{bmatrix} \in \mathbb{R}^{VN\times VN}. \quad (10)$$

By putting Eq. (2), Eq. (3) and Eq. (9) together, we obtain our CRV-DCL model as follows:

$$\begin{aligned} &\min_{\mathbf{W}^v,\bar{\mathbf{Y}},\mathbf{Y}^v} \sum_{v=1}^{V} ||(\mathbf{X}^v)^T\mathbf{W}^v - (\bar{\mathbf{Y}}+\mathbf{Y}^v)||_F^2 \\ &+ \alpha \sum_{v=1}^{V} (||\mathbf{W}^v||_{2,1} + ||\mathbf{Y}^v||_1) \\ &+ \beta \sum_{v=1}^{V}\sum_{u=1}^{V} Tr((\mathbf{W}^v)^T\mathbf{X}^v\mathbf{L}(\mathbf{X}^u)^T\mathbf{W}^u), \\ &s.t. \ \ \bar{\mathbf{Y}}^T\bar{\mathbf{Y}} = \mathbf{I}, \bar{\mathbf{Y}} \geq 0. \end{aligned} \quad (11)$$

As can be seen from Eq. (11), our proposed model can capture both the shared information and diversity information of different views by using the relaxed semantic label representation. In addition, both the intra-view and inter-view local geometrical structure of data samples can be preserved via the cross-view graph Laplacian regularization term.

Optimization Algorithm

Since the variables including the projection matrices $\mathbf{W}^v$, semantic label matrices $\bar{\mathbf{Y}}$ and $\mathbf{Y}^v$ in Eq. (11) are related to each other, it is difficult to solve them at one step. Hence, we develop an alternative iterative algorithm to solve the optimization problem. At each time, we optimize the objective function w.r.t one variable with others fixed and the procedure repeats until convergence.

Optimize $\bar{\mathbf{Y}}$ by fixing other variables

When $\mathbf{Y}^1, \cdots, \mathbf{Y}^v, \mathbf{W}^1, \cdots, \mathbf{W}^v$ are fixed, optimizing $\bar{\mathbf{Y}}$ is equal to solve the following problem:

$$\min_{\bar{\mathbf{Y}}} \sum_{v=1}^{V} ||(\mathbf{X}^v)^T\mathbf{W}^v - (\bar{\mathbf{Y}} + \mathbf{Y}^v)||_F^2, \;\; s.t. \;\; \bar{\mathbf{Y}}^T\bar{\mathbf{Y}} = \mathbf{I}, \bar{\mathbf{Y}} \geq 0. \tag{12}$$

Then, Eq. (12) can be rewritten as the following trace form:

$$\begin{aligned} &\min_{\bar{\mathbf{Y}}} \sum_{v=1}^{V} Tr(-2(\mathbf{W}^v)^T\mathbf{X}^v\bar{\mathbf{Y}} + 2(\mathbf{Y}^v)^T\bar{\mathbf{Y}}) + Tr(\bar{\mathbf{Y}}^T\bar{\mathbf{Y}}), \\ &s.t. \;\; \bar{\mathbf{Y}}^T\bar{\mathbf{Y}} = \mathbf{I}, \bar{\mathbf{Y}} \geq 0. \end{aligned} \tag{13}$$

We add an extra penalty term $\rho||\bar{\mathbf{Y}}^T\bar{\mathbf{Y}} - \mathbf{I}||_F^2$ and introduce a Lagrange multiplier Φ to eliminate the orthogonal constraint and remove the inequality constraint, respectively. Then we have the following Lagrange function:

$$\begin{aligned} \mathcal{F}(\bar{\mathbf{Y}}, \Phi) = &\sum_{v=1}^{V} Tr(-2(\mathbf{W}^v)^T\mathbf{X}^v\bar{\mathbf{Y}} + 2(\mathbf{Y}^v)^T\bar{\mathbf{Y}}) \\ &+ Tr(\bar{\mathbf{Y}}^T\bar{\mathbf{Y}}) + \rho||\bar{\mathbf{Y}}^T\bar{\mathbf{Y}} - \mathbf{I}||_F^2 - Tr(\Phi^T\bar{\mathbf{Y}}). \end{aligned} \tag{14}$$

By taking the derivative of $\mathcal{F}(\bar{\mathbf{Y}}, \Phi)$ w.r.t $\bar{\mathbf{Y}}$, and setting it to zero, we have

$$\begin{aligned} \frac{\partial F(\bar{\mathbf{Y}}, \Phi)}{\partial \bar{\mathbf{Y}}} = &\sum_{v=1}^{V} 2\mathbf{Y}^v - 2(\mathbf{X}^v)^T\mathbf{W}^v \\ &+ 2\bar{\mathbf{Y}} + 4\rho\bar{\mathbf{Y}}(\bar{\mathbf{Y}}^T\bar{\mathbf{Y}} - \mathbf{I}) - \Phi = 0. \end{aligned} \tag{15}$$

Then, we can get Φ:

$$\Phi = 2\bar{\mathbf{Y}} + 4\rho\bar{\mathbf{Y}}(\bar{\mathbf{Y}}^T\bar{\mathbf{Y}} - \mathbf{I}) + \sum_{v=1}^{V} 2\mathbf{Y}^v - 2(\mathbf{X}^v)^T\mathbf{W}^v \tag{16}$$

According to the Karush-Kuhn-Tucker condition (Hanson 1999; Boyd and Vandenberghe 2004), i.e., $\Phi_{ij}\bar{\mathbf{Y}}_{ij} = 0$, we get the following equation:

$$[2\bar{\mathbf{Y}} + 4\rho\bar{\mathbf{Y}}(\bar{\mathbf{Y}}^T\bar{\mathbf{Y}} - \mathbf{I}) + \sum_{v=1}^{V} 2\mathbf{Y}^v - 2(\mathbf{X}^v)^T\mathbf{W}^v]_{ij}\bar{\mathbf{Y}}_{ij} = 0. \tag{17}$$

Then, $\bar{\mathbf{Y}}$ can be updated via following strategy:

$$\bar{\mathbf{Y}}_{ij} \leftarrow \bar{\mathbf{Y}}_{ij} \frac{[2\rho\bar{\mathbf{Y}} + \sum_{v=1}^{V} (\mathbf{X}^v)^T\mathbf{W}^v]_{ij}}{[\bar{\mathbf{Y}} + 2\rho\bar{\mathbf{Y}}\bar{\mathbf{Y}}^T\bar{\mathbf{Y}} + \sum_{v=1}^{V} \mathbf{Y}^v]_{ij}}. \tag{18}$$

In this work, in order to constrain the orthogonality of $\bar{\mathbf{Y}}$, we set ρ a relatively large value, $\rho = 10^6$ in our experiments.

Optimize $\mathbf{Y}^v$ by fixing other variables

When we fix $\bar{\mathbf{Y}}$, $\mathbf{W}^1, \cdots, \mathbf{W}^V$, $\mathbf{Y}^1, \cdots, \mathbf{Y}^{v-1}$, $\mathbf{Y}^{v+1}, \cdots, \mathbf{Y}^V$, $\mathbf{Y}^v$ can be updated by solving following problem:

$$\min_{\mathbf{Y}^v} ||(\mathbf{X}^v)^T\mathbf{W}^v - (\bar{\mathbf{Y}} + \mathbf{Y}^v)||_F^2 + \alpha||\mathbf{Y}^v||_1, \tag{19}$$

which can be solved by using the soft-thresholding operator (Cai et al. 2008) and $\mathbf{Y}^v$ can be obtained as follows:

$$\mathbf{Y}^v = sign((\mathbf{X}^v)^T\mathbf{W}^v - \bar{\mathbf{Y}})max(|(\mathbf{X}^v)^T\mathbf{W}^v - \bar{\mathbf{Y}}| - \frac{\alpha}{2}, \mathbf{0}). \tag{20}$$

$\mathbf{Y}^1, \cdots, \mathbf{Y}^{v-1}, \mathbf{Y}^{v+1}$ can be updated in a similar way.

Optimize $\mathbf{W}^v$ by fixing other variables

When we fix $\bar{\mathbf{Y}}$, $\mathbf{Y}^1, \cdots, \mathbf{Y}^V$, $\mathbf{W}^1, \cdots, \mathbf{W}^{v-1}$, $\mathbf{W}^{v+1}, \cdots, \mathbf{W}^V$, $\mathbf{W}^v$ can be updated by solving following problem:

$$\begin{aligned} &\min_{\mathbf{W}^v} ||(\mathbf{X}^v)^T\mathbf{W}^v - (\bar{\mathbf{Y}} + \mathbf{Y}^v)||_F^2 + \alpha||\mathbf{W}^v||_{2,1} \\ &+ \beta\sum_{v=1}^{V}\sum_{u=1}^{V} Tr((\mathbf{W}^v)^T\mathbf{X}^v\mathbf{L}^{vu}(\mathbf{X}^u)^T\mathbf{W}^u). \end{aligned} \tag{21}$$

By taking the derivative of objective function in Eq. (21) w.r.t $\mathbf{W}^v$ and setting it to zero, we obtain

$$\begin{aligned} &\mathbf{X}^v(\mathbf{X}^v)^T\mathbf{W}^v - \mathbf{X}^v(\bar{\mathbf{Y}} + \mathbf{Y}^v) + \alpha\mathbf{G}^v\mathbf{W}^v \\ &+ \beta(\mathbf{X}^v\mathbf{L}^v(\mathbf{X}^v)^T\mathbf{W}^v + \sum_{l\neq v}^{V} \mathbf{X}^l\mathbf{L}^{vl}(\mathbf{X}^l)^T\mathbf{W}^l) = 0, \end{aligned} \tag{22}$$

where $\mathbf{G}^v$ is a diagonal matrix with its i-th diagonal entry calculated as:

$$\mathbf{G}_{ii}^v = \frac{1}{2||(\mathbf{W}^v)^i||_2}. \tag{23}$$

According to Eq. (22), $\mathbf{W}^v$ can be updated as:

$$\begin{aligned} \mathbf{W}^v = &(\mathbf{X}^v(\mathbf{X}^v)^T + \alpha\mathbf{G}^v + \beta(\mathbf{X}^v\mathbf{L}^v(\mathbf{X}^v)^T)^{-1} \\ &(\mathbf{X}^v(\bar{\mathbf{Y}} + \mathbf{Y}^v) - \beta\sum_{l\neq v}^{V} \mathbf{X}^l\mathbf{L}^{vl}(\mathbf{X}^l)^T\mathbf{W}^l) \end{aligned} \tag{24}$$

At this step, $\mathbf{G}^v$ and $\mathbf{W}^v$ can be updated iteratively via Eq. (23) and Eq. (24). We summarize the optimization procedure of CRV-DCL in Algorithm 1.

Theoretical Analysis of Algorithm 1

In this section, we give a brief theoretical analysis of Algorithm 1, including convergence analysis and complexity analysis.

Algorithm 1 Iterative algorithm for solving CRV-DCL

Input: Multi-view data matrices $\{\mathbf{X}^v \in \mathbb{R}^{d_v \times N}\}_{v=1}^V$, parameters: α, β.
Initialize: $\mathbf{Y}^1, \cdots, \mathbf{Y}^V, \mathbf{W}^1, \cdots, \mathbf{W}^V, \varepsilon, t = 0$.
while not converged do
1. Update $\bar{\mathbf{Y}}$ via Eq. (18);
2. Update $\mathbf{Y}^v$ via Eq. (20);
3. Update $\mathbf{W}^v$ by solving Eq. (21);
4. Check convergence condition: $\frac{obt^t - obj^{t+1}}{obj^{t+1}} < \varepsilon$.
end while
Output: $\bar{\mathbf{Y}}, \mathbf{Y}^1, \cdots, \mathbf{Y}^V, \mathbf{W}^1, \cdots, \mathbf{W}^V$.
Feature selection: Sort the l_2-norm of the rows of $\{\mathbf{W}^v\}_{v=1}^V$ in decent order and select the largest K values. The corresponding feature indexes form the selected feature index set.

Convergence Analysis

Although it is not easy to theoretically proof the convergence of Algorithm 1, the convergence of each step of Algorithm 1 can be guaranteed. In step 1 of Algorithm 1, since we use the Karush-Kuhn-Tucker condition to update $\bar{\mathbf{Y}}$, the objective value of Eq. (13) can be ensured to monotonically decrease. In step 2 of Algorithm 1 for updating $\mathbf{Y}^v$, the soft-thresholding operator can ensure the global optimal solution of Eq. (19). As to step 3 of Algorithm 1, $\mathbf{W}^v$ and $\mathbf{G}^v$ are iteratively updated via the iterative re-weighted least-squares algorithm, of which the convergence can be guaranteed. In addition, in the experimental section, we will also empirically validate that the objective value of Eq. (11) can be ensured to decrease monotonically with numbers of iteration.

Time Complexity Analysis

For updating $\bar{\mathbf{Y}}$, the main computation lies in calculating Eq. (18), which only consists of some matrix multiplication operations. As to updating $\mathbf{Y}^v$, there also only consists of a matrix multiplication operation, i.e., $(\mathbf{X}^v)^T\mathbf{W}^v$. The main computational cost of Algorithm 1 lies in solving $\mathbf{W}^v$ since we need to compute the inverse of a $d_v \times d_v$ matrix, of which the computational complexity is $\mathcal{O}(d_v^3)$.

Experiments

Datasets

In this work, six publicly available multi-view benchmark datasets are used in our experiments.

Handwritten is a dataset which consists of handwritten digits of 0 to 9 from UCI machine learning repository (Bache and Lichman 2013). It consists of 2000 data samples. All of the 6 published features including 76 Fourier coefficients of the character shapes (FOU), 216 profile correlations (FAC), 64 Karhunen-love coefficients (KAR), 240 pixel averages in 2×3 windows (Pix), 47 Zernike moment (ZER) and 6 morphological (MOR) features are used in our experiments. **Caltech101-7** is an image dataset which consists of 101 categories of images for object recognition problem (Li, Fergus, and Perona 2005). Following previous works (Dueck and Frey 2007; Li et al. 2015), we select the widely used 7 classes, i.e. Face, Motorbikes, Dolla-Bill, Garfield, Snoopy, Stop-Sign and Windsor-Chair and get 1474 images. Six features are extracted from all the images: i.e. 48 dimension Gabor feature, 40 dimension wavelet moments (WM), 254 dimension CENTRIST feature, 1984 dimension HOG feature, 512 dimension GIST feature, and 928 dimension LBP feature. **Reuters** is a dataset consists of documents that are written in five different languages and their translations (Amini, Usunier, and Goutte 2009). There are 6 classes of all the documents. We use the subset that are written in English and all their translations in all the other 4 languages (French, German, Spanish and Italian). **NUSWIDEOBJ** is a dataset for object recognition which consists of 30000 images in 31 categories (Chua et al. 2009). Five features including 65 dimension color Histogram (CH), 226 dimension color moments (CM), 145 dimension color correlation (CORR), 74 dimension edge distribution and 129 wavelet texture are used in our experiments. **MSRCV1** is an image dataset which consists of 240 images and 8 object classes (Xu, Han, and Nie 2016). We select 7 classes, i.e., tree, building, airplane, cow, face, car and bicycle, and extract 6 types of features: 1302 dimensional CENT feature, 48 dimensional CMT feature, 512 dimensional GIST feature, 100 dimensional HOG feature, 256 dimensional LBP feature and 210 dimensional SIFT feature from each image to construct different view features. **BBCSport** consists the documents from the BBC Sport website corresponding to sports news in 5 topical areas, which is associated with 2 views which are 3183 and 3203 dimension, respectively (Xia et al. 2014).

Experimental Setup

Similar to previous single view and multi-view unsupervised feature selection methods, we use the selected feature subsets to perform K-means clustering for evaluating the performance of the proposed CRV-DCL. As two widely used evaluation metrics, i.e., accuracy (ACC) and normalized mutual information (NMI), are employed to evaluate the quality of clustering results. Larger ACC and NMI values represent better performance. Meanwhile, We also compare the proposed CRV-DCL with other seven different single view and multi-view unsupervised feature selection methods, they are as follows:

- **Baseline**: k-means is employed to cluster original multiple view data by simply combining all features into a single view.
- **LS** (He, Cai, and Niyogi 2005) and **SPEC** (Zhao and Liu 2007): Two representative and classical single view feature selection methods. Samples with combined features are taken as input. In this paper, we employ them to show the effectiveness of multi-view feature selection.
- **AMFS** (Wang et al. 2016), **ASVW** (Hou et al. 2017), **RMFS** (Liu, Mao, and Fu 2017) and **CGMV-UFS** (Tang et al. 2018a): Four representative multi-view feature selection approaches which are used to compare with our proposed CRV-DCL for demonstrating its effectiveness.

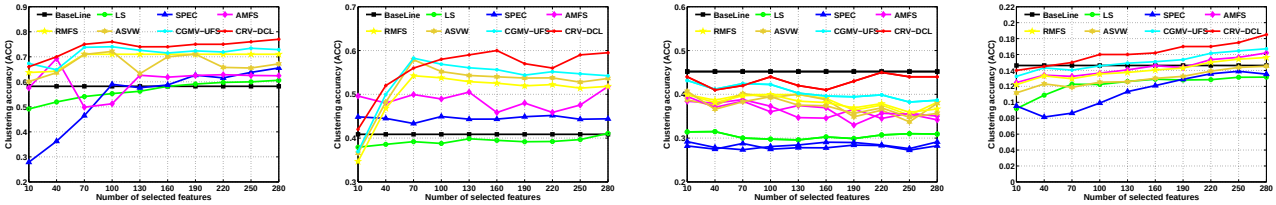
There are several parameters need to be set in CRV-DCL and other methods. For LS, SPEC, CGMV-UFS and CRV-DCL, the neighborhood size for constructing the intra-view

Table 1: Clustering results (ACC% $\pm$ std%) of different feature selection algorithms on different datasets.

Datasets	handwritten	Caltech101-7	Reuters	NUSWIDEOBJ	MSRCV1	BBCSport
Baseline	58.20±4.89	40.86±3.70	**45.20±2.51**	14.62±0.43	47.67±2.87	53.37±1.41
LS	60.71±5.32	41.17±3.37	31.42±1.01	13.26±0.31	52.21±5.65	43.04±4.25
SPEC	65.53±6.47	45.15±2.67	27.20±0.00	14.06±0.46	36.74±5.41	36.05±0.10
SGOFS	63.46±4.43	44.48±3.87	33.65±1.67	15.26±0.50	54.52±6.92	47.98±4.22
AMFS	69.41±1.81	52.37±2.86	39.84±1.31	16.10±0.38	58.41±4.96	48.02±1.12
RMFS	71.04±3.21	54.37±2.64	39.94±1.24	16.23±0.53	62.94±5.27	48.32±1.07
ASVW	72.13±4.91	56.24±5.18	41.48±1.97	16.52±0.49	65.41±4.62	51.77±1.21
CGMV-UFS	75.45±5.99	58.25±5.46	43.16±2.33	17.25±0.40	68.93±6.22	54.03±1.05
CRV-DCL	**76.47±4.23**	**59.23±5.25**	45.07±2.14	**17.86±0.39**	**69.58±5.72**	**54.95±1.16**

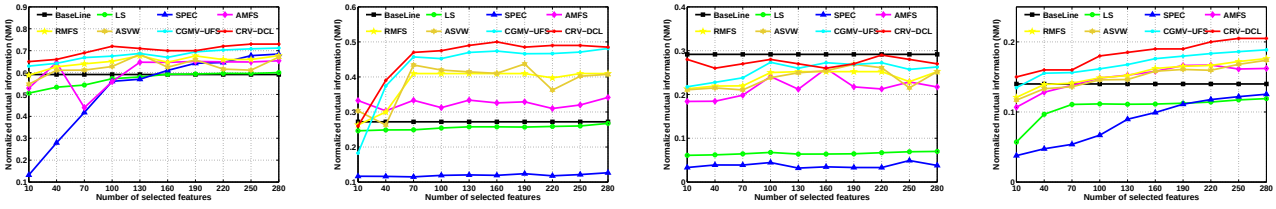
Table 2: Clustering results (NMI% $\pm$ std%) of different feature selection algorithms on different datasets.

Datasets	handwritten	Caltech101-7	Reuters	NUSWIDEOBJ	MSRCV1	BBCSport
Baseline	59.11±1.89	27.19±1.00	**29.16±2.51**	14.00±0.17	39.69±2.40	30.10±1.28
LS	59.97±1.44	26.36±1.07	7.63±0.91	12.13±0.18	42.63±4.01	16.79±6.54
SPEC	68.45±3.98	12.35±1.06	6.04±0.00	12.82±0.19	22.30±5.14	13.24±0.06
SGOFS	60.69±1.72	27.61±1.25	22.12±0.86	14.25±0.20	47.56±3.47	17.31±4.89
AMFS	65.09±0.64	35.53±2.03	24.30±0.94	16.51±0.17	50.37±4.80	19.86±3.37
RMFS	67.75±1.60	40.97±1.69	25.21±1.19	16.58±0.26	56.61±3.17	23.62±1.23
ASVW	68.92±1.37	46.41±1.92	26.75±1.27	16.87±0.21	57.20±3.61	27.29±2.54
CGMV-UFS	71.83±2.18	48.71±3.33	27.76±1.06	18.96±0.19	60.50±5.46	31.94±1.39
CRV-DCL	**72.65±2.20**	**49.86±3.14**	29.14±1.02	**19.76±0.23**	**62.36±5.38**	**32.41±1.35**



(a) handwritten (b) Caltech101-7 (c) Reuters (d) NUSWIDEOBJ

Figure 1: The clustering accuracy (ACC) of using different selected features by different methods on different datasets.



(a) handwritten (b) Caltech101-7 (c) Reuters (d) NUSWIDEOBJ

Figure 2: The normalized mutual information (NMI) of using different selected features by different methods on different datasets.

similarity graph is set to 5 and the kernel parameter σ in the Gaussian RBF kernel function is set to 1. For AMFS, parameter r is set to 2 as suggested in the corresponding paper. For ASVW, the regularization parameter λ is tuned from $\{10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 10^2, 10^3\}$ and other parameters are set to default values as in the original paper to obtain the optimal results. As to the α and β in CRV-DCL, we also tune their values by a "grid-search" strategy from $\{10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 10^2, 10^3\}$. For every method, the best results by tuning the parameters are reported for comparison.

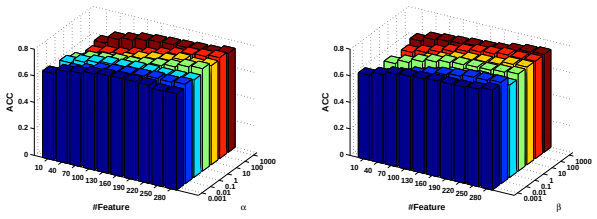
Since it is hard to determine the optimal number of se-

lected features for a certain dataset, we set different number of selected features for all datasets. As to each dataset, the best clustering results from the optimal parameters and selected feature numbers are reported for all the methods. For all of the datasets, we vary the selected feature numbers from $\{10, 40, 70, \cdots, 250, 280\}$. After obtaining the feature subsets, K-means algorithm is run 20 times on the selected feature subsets with random starting points for eliminating the bias of initialization. Then, the average results of the 20 times running of K-means are recorded and reported.

Experimental Results

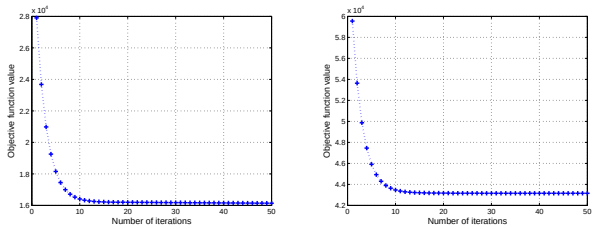
The experimental results of different methods in terms of ACC and NMI on different datasets are summarized in Table 1 and Table 2, respectively. The best results are highlighted in bold fonts. As can be seen, on handwritten, Caltech101-7, NUSWIDEOBJ, MSRCV1 and BBCSport datasets, the proposed CRV-DCL performs the best when compared with other methods.As to Reuters dataset, although our method does not achieve the best clustering results, it still outperforms all of other feature selection methods, which demonstrates that the proposed CRV-DCL can obtain better clustering results with a small subset of selected features when compared with other methods. In addition, compared with traditional single view unsupervised feature selection methods, the multi-view methods perform significantly better. We can see that CRV-DCL can get more than 10% improvements in average when compared to the best result of all the other single-view methods. This is caused by the fact that single view methods characterize the structures of each data view independently and combine them by simply stacking.

Since the optimal number of selected features is hard to determine, in order to illustrate the effect of feature selection to clustering, we show the detailed performance of all algorithms with respect to different selected numbers of features on different datasets. Figure 1 and Figure 2 plot the ACC values and the NMI values with respect to the number of selected features on different datasets, respectively (due to the page limitation, only results of four datasets are shown). The results also show that the proposed method can steadily perform better than other methods over a range of selected features. It is worth noting that when using fewer features, our method can obtain higher clustering accuracy than the baseline excluding the Reuters dataset, which demonstrates



(a) $\beta = 1$ (b) $\alpha = 1$

Figure 3: ACC of CRV-DCL with different α, β, and feature numbers on handwritten dataset.



(a) handwritten (b) Caltech101-7

Figure 4: Convergence curves of Algorithm 1 on different datasets.

that the selected subset of the features can not only reduce the computation cost, but also improve the clustering performance. As to Caltech101-7 dataset, when the number of selected features is fewer than 50, our method dose not perform the best. However, when we select more than 50 features, the proposed CRV-DCL can steadily perform better than other methods.

Parameter Sensitivity and Convergence Analysis

There are two parameters in our model (i.e., α and β). To further demonstrate the performance of the proposed method, we study its sensitivity w.r.t. the parameters in Eq. (11). Due to the page space limitation, we only report the ACC of handwritten dataset here. First, we fix $\alpha = 1$ and vary β. Then we fix $\beta = 1$ and vary α. Figure 3 plot the ACC and NMI values given by CRV-DCL for different λ, β and selected features. The experimental results shown that our CRV-DCL is not very sensitive to parameters α and β, but it is relatively sensitive to the number of selected features. However, this is a common problem for most unsupervised feature selection methods.

In Figure 4, we plot the objective function values of Eq. 11 with varying iteration times on handwritten and Caltech101-7 datasets, the results show that the objective value of Eq. (11) decreases very fast within the first 10 iterations.

Conclusions

This paper introduces a novel MV-UFS method via cross-view local structure preserved diversity and consensus learning. The proposed method captures both the common information and distinguishing knowledge across different views by projecting each view of original data into a common semantic label space, which is composed of a consensus part and a diversity part. Meanwhile, in order to preserve the local structure of a certain sample in different views and different samples in the same view, an cross-view Laplacian regularization term is designed. Experiments on real-world multi-view datasets are conducted to demonstrate the efficacy of the proposed method.

Acknowledgments

This work was supported in part by NSFC 61701451 and 61773392, and in part by the Fundamental Research Funds

for the Central Universities, China University of Geosciences (Wuhan) under Grant CUG170654 and the Open Research Project of The Hubei Key Laboratory of Intelligent Geo-Information Processing under Grant KLIGIP-2017B04. The authors would like to thank Prof. Huiying Xu from Zhejiang Normal University for her help in the proofreading of this paper and NVIDIA Corporation for the donation of a Titan Xp GPU card used for computing acceleration in this research. Xinzhong Zhu and Xinwang Liu are the corresponding authors of this paper.

References

Amini, M. R.; Usunier, N.; and Goutte, C. 2009. Learning from multiple partially observed views – an application to multilingual text categorization. In *NIPS*, 28–36.

Bache, K., and Lichman, M. 2013. Uci machine learning repository.

Boyd, S., and Vandenberghe, L. 2004. *Convex Optimization*. Cambridge University Press.

Cai, J. F.; Cand; S, E. J.; and Shen, Z. 2008. A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization* 20(4):1956–1982.

Chua, T. S.; Tang, J.; Hong, R.; Li, H.; Luo, Z.; and Zheng, Y. 2009. Nus-wide: a real-world web image database from national university of singapore. In *ACM ICIVR*, 48.

Dalal, N., and Triggs, B. 2005. Histograms of oriented gradients for human detection. In *CVPR*, 886–893.

Dueck, D., and Frey, B. J. 2007. Non-metric affinity propagation for unsupervised image categorization. In *ICCV*, 1–8.

Feng, Y.; Xiao, J.; Zhuang, Y.; and Liu, X. 2012. Adaptive unsupervised multi-view feature selection for visual concept recognition. In *ACCV*, 343–357.

Friedman, J. H. 1997. On bias, variance, 0/1–loss, and the curse-of-dimensionality. *Data Mining & Knowledge Discovery* 1(1):55–77.

Hanson, M. A. 1999. Invexity and the kuhn–tucker theorem. *Journal of Mathematical Analysis and Applications* 236(2):594–604.

He, X.; Cai, D.; and Niyogi, P. 2005. Laplacian score for feature selection. In *NIPS*, 507–514.

Hou, C.; Nie, F.; Tao, H.; and Yi, D. 2017. Multi-view unsupervised feature selection with adaptive similarity and view weight. *IEEE TKDE* 29(9):1998 – 2011.

Li, L., and Cai, M. 2017. Drug target prediction by multi-view low rank embedding. *IEEE/ACM TCBB* PP(99):1–1.

Li, Y.; Nie, F.; Huang, H.; and Huang, J. 2015. Large-scale multi-view spectral clustering via bipartite graph. In *AAAI*, 2750–2756.

Li, F. F.; Fergus, R.; and Perona, P. 2005. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *CVPRW*, 178–178.

Li, L. 2014. Mpgraph: multi-view penalised graph clustering for predicting drug-target interactions. *Iet Systems Biology* 8(2):67–73.

Liu, X.; Wang, L.; Zhang, J.; Yin, J.; and Liu, H. 2014. Global and local structure preservation for feature selection. *IEEE TNNLS* 25(6):1083–1095.

Liu, X.; Dou, Y.; Yin, J.; Wang, L.; and Zhu, E. 2016. Multiple kernel k -means clustering with matrix-induced regularization. In *AAAI*, 1888–1894.

Liu, X.; Zhu, X.; Li, M.; Wang, L.; Tang, C.; Yin, J.; Shen, D.; Wang, H.; and Gao, W. 2018. Late fusion incomplete multi-view clustering. *IEEE TPAMI*.

Liu, H.; Mao, H.; and Fu, Y. 2017. Robust multi-view feature selection. In *ICDM*, 281–290.

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

Nie, F.; Xiang, S.; Jia, Y.; Zhang, C.; and Yan, S. 2008. Trace ratio criterion for feature selection. In *NCAI*, 671–676.

Nie, F.; Wei, Z.; and Li, X. 2016. Unsupervised feature selection with structured graph optimization. In *AAAI*, 1302–1308.

Ojala, T.; Pietikainen, M.; and Maenpaa, T. 2002. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE TPAMI* 24(7):971–987.

Tang, C.; Cao, L.; Zheng, X.; and Wang, M. 2017. Gene selection for microarray data classification via subspace learning and manifold regularization. *MBEC* (6871):1–14.

Tang, C.; Chen, J.; Liu, X.; Li, M.; Wang, P.; Wang, M.; and Lu, P. 2018a. Consensus learning guided multi-view unsupervised feature selection. *KBS* 160:49–60.

Tang, C.; Liu, X.; Li, M.; Wang, P.; Chen, J.; Wang, L.; and Li, W. 2018b. Robust unsupervised feature selection via dual self-representation and manifold regularization. *KBS* 145:109–120.

Tang, C.; Zhu, X.; Chen, J.; Wang, P.; Liu, X.; and Tian, J. 2018c. Robust graph regularized unsupervised feature selection. *ESWA* 96:64–76.

Tang, C.; Zhu, X.; Liu, X.; Li, M.; Wang, P.; Zhang, C.; and Wang, L. 2018d. Learning a joint affinity graph for multiview subspace clustering. *IEEE TMM*.

Wang, Z.; Feng, Y.; Qi, T.; Yang, X.; and Zhang, J. J. 2016. Adaptive multi-view feature selection for human motion retrieval. *Signal Processing* 120:691–701.

Wang, Y.; Wu, L.; Lin, X.; and Gao, J. 2018. Multiview spectral clustering via structured low-rank matrix factorization. *IEEE TNNLS* PP(99).

Xia, R.; Pan, Y.; Du, L.; and Yin, J. 2014. Robust multi-view spectral clustering via low-rank and sparse decomposition. In *AAAI*, 2149–2155.

Xu, J.; Han, J.; and Nie, F. 2016. Discriminatively embedded k-means for multi-view clustering. In *CVPR*, 5356–5364.

Z, L., and J, T. 2015. Unsupervised feature selection via nonnegative spectral analysis and redundancy control. *IEEE TIP* 24(12):5343.

Zhang, S.; Fang, Y.; Lei, C.; Li, Y.; Hu, R.; and Li, Y. 2017. Unsupervised spectral feature selection with local structure learning. In *ICBK*, 303–308.

Zhang, C.; Fu, H.; Hu, Q.; Cao, X.; Xie, Y.; Tao, D.; and Xu, D. 2018. Generalized latent multi-view subspace clustering. *IEEE TPAMI*.

Zhao, Z., and Liu, H. 2007. Spectral feature selection for supervised and unsupervised learning. In *ICML*, 1151–1157.

Zhao, Z.; Wang, L.; and Liu, H. 2010. Efficient spectral feature selection with minimum redundancy. In *AAAI*.

Zhu, X.; Zhu, Y.; Zhang, S.; Hu, R.; and He, W. 2017. Adaptive hypergraph learning for unsupervised feature selection. In *IJCAI*, 3581–3587.