Adaptive Collaborative Similarity Learning for Unsupervised Multi-view Feature Selection

Xiao Dong[1], **Lei Zhu**[1*], **Xuemeng Song**[2], **Jingjing Li**[3], **Zhiyong Cheng**[4]
[1] School of Information Science and Engineering, Shandong Normal University, China
[2] School of Computer Science and Technology, Shandong University, China
[3] University of Electronic Science and Technology of China, China
[4] School of Computing, National University of Singapore, Singapore
leizhu0608@gmail.com

Abstract

In this paper, we investigate the research problem of unsupervised multi-view feature selection. Conventional solutions first simply combine multiple pre-constructed view-specific similarity structures into a collaborative similarity structure, and then perform the subsequent feature selection. These two processes are separate and independent. The collaborative similarity structure remains fixed during feature selection. Further, the simple undirected view combination may adversely reduce the reliability of the ultimate similarity structure for feature selection, as the view-specific similarity structures generally involve noises and outlying entries. To alleviate these problems, we propose an adaptive collaborative similarity learning (ACSL) for multi-view feature selection. We propose to dynamically learn the collaborative similarity structure, and further integrate it with the ultimate feature selection into a unified framework. Moreover, a reasonable rank constraint is devised to adaptively learn an ideal collaborative similarity structure with proper similarity combination weights and desirable neighbor assignment, both of which could positively facilitate the feature selection. An effective solution guaranteed with the proved convergence is derived to iteratively tackle the formulated optimization problem. Experiments demonstrate the superiority of the proposed approach.

1 Introduction

With the advent of big data, multi-view features with high dimensions are widely employed to represent the complex data in various research fields, such as multimedia computing, machine learning and data mining [Liu *et al.*, 2016; 2017; Zhu *et al.*, 2017b; 2015; Cheng and Shen, 2016; Cheng *et al.*, 2016]. On the one hand, with multi-view features, the data could be characterized more precisely and comprehensively from different perspectives. On the other hand, high-dimensional multi-view features will inevitably generate expensive computation cost and cause massive storage cost. Moreover, they may contain adverse noises, outlying entries, irrelevant and correlated features, which may be detrimental to the subsequent learning process [Zhu *et al.*, 2016b; 2016a; 2017a]. Unsupervised multi-view feature selection [Wang *et al.*, 2016; Li and Liu, 2017] is devised to alleviate the problem. It selects a compact subset of informative features from the original features by dropping irrelevant and redundant features with advanced unsupervised learning. Due to the independence on semantic labels, high computing efficiency and well interpretation capability, unsupervised multi-view feature selection has received considerable attention in literature. It becomes a prerequisite component in various machine learning models [Li *et al.*, 2017].

The key problem of multi-view feature selection is how to effectively exploit the diversity and consistency of multi-view features to collaboratively identify the feature dimensions, which could retain the key characteristics of the original features. Existing approaches can be categorized into two major families. The first kind of methods first concatenates multi-view features into a vector and then directly imports it into the conventional single-view feature selection model. The candidate features are generally ranked based on spectral graph theory. Typical methods of this kind include Laplacian Score (LapScor) [He *et al.*, 2005], spectral feature selection (SPEC) [Zhao and Liu, 2007] and minimum redundancy spectral feature selection (MRSF) [Zhao *et al.*, 2010]. Commonly, the pipeline of these methods follows two separate processes: 1) Similarity structure is constructed with fixed graph parameters to describe the geometric structure of data. 2) Sparsity and manifold regularization are employed together to identify the most salient features. Although these methods are reported to achieve certain success, they treat features from different views independently and unfortunately neglect the important view correlations.

Another family of methods considers view correlation when performing feature selection. Representative works include adaptive multi-view feature selection (AMFS) [Wang *et*

*Corresponding Author

al., 2016], multi-view feature selection (MVFS) [Tang *et al.*, 2013] and adaptive unsupervised multi-view feature selection (AUMFS) [Feng *et al.*, 2013]. These methods first construct multiple view-specific similarity structures[1] and then perform the subsequent feature selection based on the collaborative (combined) similarity structure. These two processes are separate and independent. The collaborative similarity structure remains fixed during feature selection. The latently involved data noises and outlying entries in the view-specific similarity structures will adversely reduce the reliability of the ultimate collaborative similarity structure for feature selection. Furthermore, conventional approaches generally employ k-nearest neighbors assignment to construct the view-specific similarity structures and the simple weighted combination for ultimate similarity structure generation. This strategy can hardly achieve the ideal state for clustering that the number of connected components in the ultimate similarity structure is equal to the number of clusters [Nie *et al.*, 2014]. Thus, suboptimal performance may be caused under such circumstance.

In this paper, we introduce an adaptive collaborative similarity learning (ACSL) for unsupervised multi-view feature selection. The main contributions of this paper can be summarized as follows:

- Different from existing solutions, we integrate the collaborative similarity structure learning and multi-view feature selection into a unified framework. The collaborative similarity structure and similarity combination weights could be learned adaptively by considering the ultimate feature selection performance. Simultaneously, the feature selection can preserve the dynamically adjusted similarity structure.
- We impose a reasonable rank constraint to adaptively learn an ideal collaborative similarity structure with proper neighbor assignment which could positively facilitate the ultimate feature selection. An effective alternate optimization approach guaranteed with convergence is derived to iteratively solve the formulated optimization problem.

2 Related Work

One kind of unsupervised multi-view feature selection methods directly imports the concatenated features in multiple views into the single-view feature selection model. In [He *et al.*, 2005], Laplacian score (LapScor) is employed to measure the capability of each feature dimension on preserving sample similarity. [Zhao and Liu, 2007] proposes a general spectral theory based learning framework to unify the unsupervised and supervised feature selection. [Zhao *et al.*, 2010] adopts an embedding model to handle feature redundancy in the spectral feature selection. These methods generally rank the candidate feature dimensions with various graphs which characterize the manifold structure. They treat features from different views independently and unfortunately ignore the important correlation of different feature views. Another kind of methods directly tackles the multi-view feature selection. They consider view correlations when performing feature selection. Adaptive multi-view feature selection (AMFS) [Wang *et al.*, 2016] is an unsupervised feature selection approach which is developed for human motion retrieval. It describes the local geometric structure of data in each view with local descriptor and performs the feature selection in a general trace ratio optimization. In this method, the feature dimensions are determined with trace ratio criteria. Adaptive unsupervised multi-view feature selection (AUMFS) [Feng *et al.*, 2013] addresses the feature selection problem for visual concept recognition. It employs $l_{2,1}$ norm [Nie *et al.*, 2010] based sparse regression model to automatically identify discriminative features. In AUMFS, data cluster structure, data similarity and the correlations of different views are considered for feature selection. Multi-view feature selection (MVFS) [Tang *et al.*, 2013] investigates the feature selection for multi-view data in social media. A learning framework is devised to exploit the relations of views and help each view select relevant features.

[1] In this paper, view-specific similarity structure is constructed with the corresponding view-specific feature.

3 The Proposed Methodology

3.1 Notations and Definitions

Throughout the paper, all the matrices are written in uppercase with boldface. For a matrix $\mathbf{M} \in \mathcal{R}^{N\times d}$, its i_{th} row is denoted by $\mathbf{M}_{i'} \in \mathcal{R}^{N\times 1}$, its j_{th} column is denoted by $\mathbf{M}_j \in \mathcal{R}^{d\times 1}$. The element in the i_{th} row and j_{th} column is represented as M_{ij}. The trace of the matrix $\mathbf{M}$ is denoted as $Tr(\mathbf{M})$. The transpose of matrix $\mathbf{M}$ is denoted as $\mathbf{M}^{\mathrm{T}}$. The $l_{2,1}$ norm of the matrix $\mathbf{M}$ is denoted as $||\mathbf{M}||_{2,1}$, which is calculated by $\sum_{i=1}^{N}\sqrt{\sum_{j=1}^{d} M_{i,j}^2}$. The Frobenius norm of $\mathbf{M}$ is denoted by $||\mathbf{M}||_F = \sqrt{\sum_{i=1}^{N}\sum_{j=1}^{d} M_{i,j}^2}$. $\mathbf{1}$ denotes a column vector whose all elements are one. $\mathbf{I}_{k\times k}$ denotes $k \times k$ identify matrix.

The feature matrix of data in the v_{th} view is denoted as $\mathbf{X}^v = [\mathbf{x}_1^v, \mathbf{x}_2^v, ..., \mathbf{x}_N^v]^{\mathrm{T}} \in \mathcal{R}^{N\times d_v}$, $\mathbf{x}_1^v \in \mathcal{R}^{d_v\times 1}$, d_v is the dimension of feature in the v_{th} view, N is the number of data samples. We pack the feature matrices in V views $\{\mathbf{X}^v\}_{v=1}^V$ and the overall feature matrix of data can be represented as $\mathbf{X} = [\mathbf{X}^1, \mathbf{X}^2, ..., \mathbf{X}^V] \in \mathcal{R}^{N\times d}$, $\sum_{v=1}^{V} d_v = d$. The objective of unsupervised multi-view feature selection is to identify l most valuable features with only $\mathbf{X}$.

3.2 Formulation

The importance of feature dimensions are primarily determined by measuring the their capabilities on preserving the similarity structures in multiple views. In this paper, we develop a unified learning framework to learn an adaptive collaborative similarity structure with automatic neighbor assignment for multi-view feature selection. In our model, the neighbors in the collaborative similarity structure could be adaptively assigned by considering the feature selection performance, and simultaneously the feature selection could preserve the dynamically constructed collaborative similarity structure. Given V similarity structures constructed in multiple views $\{\mathbf{S}^v\}_{v=1}^V$, V is the number of views, we can auto-

matically learn a collaborative similarity structure $\mathbf{S}$ by combining $\{\mathbf{S}^v\}_{v=1}^V$ with V weights.

$$\begin{aligned} &\arg\min_{\mathbf{S},\mathbf{W}} \sum_{j=1}^{N} ||\mathbf{S}_j - \sum_{v=1}^{V} w_j^v \mathbf{S}_j^v||_F^2 \\ &s.t.\ \forall j\ \mathbf{1}_N^\top \mathbf{S}_j = 1, \mathbf{S}_j \geq \mathbf{0}, \mathbf{W}_j^\top \mathbf{1}_V = 1 \end{aligned} \tag{1}$$

where $\mathbf{S}_j \in \mathcal{R}^{N\times 1}$ characterizes the similarities between any data points with j, it should be subjected to the constraint that $\mathbf{1}^\top \mathbf{S}_j = 1, \mathbf{S}_j \geq \mathbf{0}$, $\mathbf{W}_j = [w_j^1, w_j^2, ..., w_j^V]^\top \in \mathcal{R}^{V\times 1}$ is comprised of view weights for the j_{th} column of similarities, it is constrained with $\mathbf{W}_j^\top \mathbf{1}_V = 1$, $\mathbf{W} = [\mathbf{W}_1, \mathbf{W}_2, ..., \mathbf{W}_N] \in \mathcal{R}^{V\times N}$ is view weight matrix for all columns in the similarity structures. As indicated in recent work [Nie *et al.*, 2014], a theoretically ideal similarity structure for clustering should have the property that the number of connected components is equal to the number of clusters. The similarity structure with such neighbor assignment could benefit the subsequent feature selection. Unfortunately, the similarity structure learned from Eq.(1) does not have such desirable property.

To tackle the problem, in this paper, we impose a reasonable rank constraint on the Laplacian matrix of the collaborative similarity structure to enable it to have such property. Our idea is motivated by the following spectral graph theory.

Theorem 1. *If the similarity structure **S** are nonnegative, the multiplicity of eigen-values* 0 *corresponding to its Laplacain matrix is equal to the number of components of **S**. [Alavi, 1991]*

As mentioned above, the data points can be directly partitioned into k clusters if the number of components in the similarity structure $\mathbf{S}$ is exactly equal to k. **Theorem** 1 indicates that this condition can be achieved if the rank of Laplacian matrix is equal to $n-k$. With the analysis, we add a reasonable rank constraint in Eq.(1) to achieve the condition. The optimization problem becomes

$$\begin{aligned} &\arg\min_{\mathbf{S},\mathbf{W}} \sum_{j=1}^{N} ||\mathbf{S}_j - \sum_{v=1}^{V} w_j^v \mathbf{S}_j^v||_F^2 \\ &s.t.\ \forall j \quad \mathbf{1}_N^\top \mathbf{S}_j = 1, \mathbf{S}_j \geq \mathbf{0}, \mathbf{W}_j^\top \mathbf{1}_V = 1, rank(\mathbf{L}_S) = n-k \end{aligned} \tag{2}$$

where $\mathbf{L}_S = \mathbf{D}_S - \frac{\mathbf{S}^\top + \mathbf{S}}{2}$ is the Laplacain matrix of similarity structure $\mathbf{S}$, $\mathbf{D}_S = \mathbf{S1}$ is diagonal matrix. As shown in Eq.(2), directly imposing the rank constraint $rank(\mathbf{L}_S) = n-k$ will make the above problem hard to solve. Fortunately, according to Ky Fan's Theorem [K., 1949], we can have $\sum_{i=1}^{k} \delta_i(\mathbf{L}_S) = \arg\min_{\mathbf{F}\in\mathcal{R}^{n\times k}, \mathbf{F}^\top\mathbf{F}=\mathbf{I}_k} Tr(\mathbf{F}^\top \mathbf{L}_S \mathbf{F})$, where $\delta_i(\mathbf{L}_S)$ is the i_{th} smallest eigen-values of $\mathbf{L}_S$ and $\mathbf{F} \in \mathcal{R}^{N\times k}$ is the relaxed cluster indicator matrix. Obviously, the rank constraint $rank(\mathbf{L}_S) = n-k$ can be satisfied when $\sum_{i=1}^{k} \delta_i(\mathbf{L}_S) = 0$. To this end, we reformulate the Eq.(2) as the following simple equivalent form

$$\begin{aligned} &\arg\min_{\mathbf{F},\mathbf{S},\mathbf{W}} \sum_{j=1}^{N} ||\mathbf{S}_j - \sum_{v=1}^{V} w_j^v \mathbf{S}_j^v||_F^2 + \alpha Tr(\mathbf{F}^\top \mathbf{L}_S \mathbf{F}) \\ &s.t.\ \forall j \quad \mathbf{1}_N^\top \mathbf{S}_j = 1, \mathbf{S}_j \geq \mathbf{0}, \mathbf{W}_j^\top \mathbf{1}_V = 1, \mathbf{F} \in \mathcal{R}^{N\times k}, \mathbf{F}^\top\mathbf{F} = \mathbf{I}_k \end{aligned} \tag{3}$$

As shown in the above equation, when $\alpha > 0$ is large enough, the term $Tr(\mathbf{F}^\top \mathbf{L}_S \mathbf{F})$ is forced to be infinitely approximate 0 and the rank constraint can be satisfied accordingly. By simply transforming the rank constraint to trace in objective function, the problem in Eq.(2) can be tackled more easily.

The selected features should preserve the dynamically learned similarity structure. Conventional approaches separate the similarity structure construction and feature selection into two independent processes, which will potentially lead to sub-optimal performance. In this paper, we learn the collaborative similarity structure dynamically and further integrate it with feature selection into a unified framework. Specifically, based on the collaborative similarity structure learning in Eq.(3), we employ sparse regression model to learn a projection matrix $\mathbf{P} \in \mathcal{R}^{d\times k}$, so that the projected low-dimensional data $\mathbf{XP}$ can approximate the relaxed cluster indicator $\mathbf{F}$. To select the features, we impose $l_{2,1}$ norm penalty on $\mathbf{P}$ to force it with row sparsity. The importance of features can be measured by the l_2 norm of each row feature in $\mathbf{P}$. The overall optimization formulation can be derived as

$$\begin{aligned} \arg\min_{\mathbf{P},\mathbf{F},\mathbf{S},\mathbf{W}} \ &\mathbf{\Omega}(\mathbf{P},\mathbf{F},\mathbf{S},\mathbf{W}) = \sum_{j=1}^{N} ||\mathbf{S}_j - \sum_{v=1}^{V} w_j^v \mathbf{S}_j^v||_F^2 + \\ &\alpha Tr(\mathbf{F}^\top \mathbf{L}_S \mathbf{F}) + \beta(||\mathbf{XP} - \mathbf{F}||_F^2 + \gamma||\mathbf{P}||_{2,1}) \\ s.t.\ \forall j \quad &\mathbf{1}_N^\top \mathbf{S}_j = 1, \mathbf{S}_j \geq \mathbf{0}, \mathbf{W}_j^\top \mathbf{1}_V = 1, \mathbf{F} \in \mathcal{R}^{N\times k}, \mathbf{F}^\top\mathbf{F} = \mathbf{I}_k \end{aligned} \tag{4}$$

With $\mathbf{P}$, the importance of features are measured by $||\mathbf{P}_{i'}||_2$. The features with the l largest values can be finally determined.

3.3 Alternate Optimization

As shown in Eq.(4), the objective function is not convex to three variables simultaneously. In this paper, we propose an effective alternate optimization to iteratively solve the problem. Specifically, we optimize one variable by fixing the others.

Update P. By fixing the other variables, the optimization for $\mathbf{P}$ can be derived as

$$\arg\min_{\mathbf{P}} ||\mathbf{XP} - \mathbf{F}||_F^2 + \gamma||\mathbf{P}||_{2,1} \tag{5}$$

This equation is not differentiable. Hence, we transform it to following equivalent equation [Nie *et al.*, 2010]

$$\arg\min_{\mathbf{P}} ||\mathbf{XP} - \mathbf{F}||_F^2 + \gamma Tr(\mathbf{P}^\top \mathbf{\Gamma} \mathbf{P}) \tag{6}$$

$\mathbf{\Gamma} \in \mathcal{R}^{d\times d}$ is diagonal matrix whose i_{th} diagonal element is $\frac{1}{2\sqrt{\mathbf{P}_{i'}\mathbf{P}_{i'}^\top}+\epsilon}$. ϵ is small enough constant. It is used to avoid the condition that $||\mathbf{P}_{i'}||_2$ is zero. By calculating the derivations of the objective function with $\mathbf{P}$ and setting it to zeros, we can obtain the updating rule for $\mathbf{P}$ as

$$\mathbf{P} = (\mathbf{X}^\top \mathbf{X} + \gamma \mathbf{\Gamma})^{-1} \mathbf{X}^\top \mathbf{F} \tag{7}$$

Note that $\mathbf{\Gamma}$ is dependent on $\mathbf{P}$. We develop an iterative approach to solve $\mathbf{P}$ and $\mathbf{\Gamma}$ until convergence. Specifically, we fix $\mathbf{\Gamma}$ to solve $\mathbf{P}$, and vice versa.

Update F. By fixing the other variables, the optimization for $\mathbf{F}$ can be derived as

$$\arg\min_{\mathbf{F}} Tr(\mathbf{F}^\top \mathbf{L}_S \mathbf{F}) + \beta(||\mathbf{XP} - \mathbf{F}||_F^2 + \gamma Tr(\mathbf{P}^\top \mathbf{\Gamma} \mathbf{P})) s.t. \mathbf{F}^\top \mathbf{F} = \mathbf{I}_c \tag{8}$$

By substituting Eq.(7) into the objective function in Eq.(8), we arrive at

$$\begin{aligned} &Tr(\mathbf{F}^{\mathrm{T}}\mathbf{L}_S\mathbf{F}) + \beta(||\mathbf{X}\mathbf{P}-\mathbf{F}||_F^2 + \gamma Tr(\mathbf{P}^{\mathrm{T}}\mathbf{\Gamma}\mathbf{P})) \\ &= Tr(\mathbf{F}^{\mathrm{T}}(\mathbf{L}_S + \beta\mathbf{I}_N - \beta\mathbf{X}\mathbf{Q}^{-1}\mathbf{X}^{\mathrm{T}})\mathbf{F}) \end{aligned} \tag{9}$$

where $\mathbf{Q} = \mathbf{X}^{\mathrm{T}}\mathbf{X} + \gamma\mathbf{\Gamma}$. With the transformation, the optimization for updating $\mathbf{F}$ can be solved by simple eigen-decomposition on the matrix $\mathbf{L}_S + \beta\mathbf{I}_N - \beta\mathbf{X}\mathbf{Q}^{-1}\mathbf{X}^{\mathrm{T}}$. Specifically, the columns of $\mathbf{F}$ are comprised of the k eigenvectors corresponding to the k smallest eigenvalues.

Update S. By fixing the other variables, the optimization for $\mathbf{S}$ becomes

$$\begin{aligned} &\arg\min_{\mathbf{S}} \sum_{j=1}^{N} ||\mathbf{S}_j - \sum_{v=1}^{V} w_j^v \mathbf{S}_j^v||_F^2 + \alpha Tr(\mathbf{F}^{\mathrm{T}}\mathbf{L}_S\mathbf{F}) \\ &s.t.\ \forall j\ \mathbf{1}_N^{\mathrm{T}}\mathbf{S}_j = 1, \mathbf{S}_j \geq \mathbf{0} \end{aligned} \tag{10}$$

The above equation can be rewritten as

$$\begin{aligned} &\arg\min_{\mathbf{S}_j} \sum_{j=1}^{N} ||\mathbf{S}_j - \sum_{v=1}^{V} w_j^v \mathbf{S}_j^v||_F^2 + \alpha \sum_{i,j=1}^{N} ||\mathbf{f}_i - \mathbf{f}_j||_2^2 S_{i,j} \\ &s.t.\ \forall j\ \mathbf{1}_N^{\mathrm{T}}\mathbf{S}_j = 1, \mathbf{S}_j \geq \mathbf{0} \end{aligned} \tag{11}$$

where $S_{i,j}$ denotes the element in the i_{th} row and j_{th} column of $\mathbf{S}$. The optimization processes for the columns of $\mathbf{S}$ are independent with each other. Hence, they can be optimized separately. Formally, $\mathbf{S}$ can be solved by

$$\arg\min_{\mathbf{S}_j} ||\mathbf{S}_j - \sum_{v=1}^{V} w_j^v \mathbf{S}_j^v||_F^2 + \alpha\mathbf{A}_j^{\mathrm{T}}\mathbf{S}_j \quad s.t.\ \forall j\ \mathbf{1}_N^{\mathrm{T}}\mathbf{S}_j = 1, \mathbf{S}_j \geq \mathbf{0} \tag{12}$$

Let $\mathbf{A}_j$ be row vector with $N \times 1$ dimensions. Its i_{th} element is $||\mathbf{f}_i - \mathbf{f}_j||_2^2$. The above optimization formula can be transformed as

$$\arg\min_{\mathbf{S}_j} ||\mathbf{S}_j + \frac{\alpha}{2}\mathbf{A}_j - \sum_{v=1}^{V} w_j^v \mathbf{S}_j^v||_F^2 \quad s.t.\ \forall j\ \mathbf{1}_N^{\mathrm{T}}\mathbf{S}_j = 1, \mathbf{S}_j \geq \mathbf{0} \tag{13}$$

This problem can be solved by an efficient iterative algorithm [Huang *et al.*, 2015].

Update W. Similar to $\mathbf{S}$, the optimization processes for the columns of $\mathbf{W}$ are independent with each other. Hence, they can be optimized separately. Formally, its j_{th} column $\mathbf{W}_j$ is solved by

$$\arg\min_{\mathbf{W}_j} ||\mathbf{S}_j - \sum_{v=1}^{V} w_j^v \mathbf{S}_j^v||_F^2 \quad s.t.\ \mathbf{W}_j^{\mathrm{T}}\mathbf{1}_V = 1 \tag{14}$$

The objective function in Eq.(14) can be rewritten as

$$\begin{aligned} &||\mathbf{S}_j - \sum_{v=1}^{V} w_j^v \mathbf{S}_j^v||_F^2 = ||\sum_{v=1}^{V} w_j^v \mathbf{S}_j - \sum_{v=1}^{V} w_j^v \mathbf{S}_j^v||_F^2 \\ &= ||\sum_{v=1}^{V} w_j^v (\mathbf{S}_j - \mathbf{S}_j^v)||_F^2 = ||\mathbf{B}_j\mathbf{W}_j||_F^2 = \mathbf{W}_j^{\mathrm{T}}\mathbf{B}_j^{\mathrm{T}}\mathbf{B}_j\mathbf{W}_j \end{aligned} \tag{15}$$

where $\mathbf{B}_j^v = \mathbf{S}_j - \mathbf{S}_j^v$, $\mathbf{B}_j = [\mathbf{B}_j^1, ..., \mathbf{B}_j^v, .., \mathbf{B}_j^V]$.

We can obtain the Lagrangian function of problem (14)

$$\mathcal{L}(W_j, \psi) = \mathbf{W}_j^{\mathrm{T}}\mathbf{B}_j^{\mathrm{T}}\mathbf{B}_j\mathbf{W}_j + \psi(1 - \mathbf{W}_j^{\mathrm{T}}\mathbf{1}_V) \tag{16}$$

ψ is also Lagrangian multiplier. By calculating the derivative of (16) with $\mathbf{W}_j$ and setting it to 0, we obtain the updating rule of $\mathbf{W}_j$ as

$$\mathbf{W}_j = \frac{(\mathbf{B}_j^{\mathrm{T}}\mathbf{B}_j)^{-1}\mathbf{1}_V}{\mathbf{1}_V^{\mathrm{T}}(\mathbf{B}_j^{\mathrm{T}}\mathbf{B}_j)^{-1}\mathbf{1}_V} \tag{17}$$

The main steps for solving problem (4) are summarized in Algorithm 1.

Algorithm 1 Multi-view feature selection via collaborative similarity structure learning with adaptive neighbors.

Input:

The pre-constructed similarity structures in v views $\{\mathbf{S}^v\}_{v=1}^V$, the number of clusters k, the parameters α, β, γ.

Output:

The collaborative similarity structure $\mathbf{S}$, the projection matrix $\mathbf{P}$ for feature selection, l identified features.

1: Initialize $\mathbf{W}$ with $\frac{1}{V}$, the collaborative similarity structure $\mathbf{S}$ with the weighted sum of $\{\mathbf{S}^v\}_{v=1}^V$. We also initialize $\mathbf{F}$ with the solution of problem (8) by substituting the Laplacian matrix calculated from the new $\mathbf{S}$.
2: **repeat**
3: Update $\mathbf{P}$ with Eq.(7).
4: Update $\mathbf{F}$ by solving the problem in Eq.(8).
5: Update $\mathbf{S}$ with Eq.(13).
6: Update $\mathbf{W}$ with Eq.(17).
7: **until** Convergence

Feature Selection

8: Calculate $||\mathbf{P}_{i'}||_2, (i = 1, 2, ..., d)$ and rank them in descending order. The l features with the top rank orders are finally determined as the features to be selected.

3.4 Convergence Analysis

The convergence of solving problem (6) can be proven by the following theorem.

Theorem 2. *The iterative optimization process for solving Eq.(5) will monotonically decrease the objective function value until convergence.*

Proof. Let $\widehat{\mathbf{P}}$ be the newly updated $\mathbf{P}$, we can obtain the following inequality

$$||\mathbf{X}\widehat{\mathbf{P}} - \mathbf{F}||_F^2 + \gamma Tr(\widehat{\mathbf{P}}^{\mathrm{T}}\mathbf{\Gamma}\widehat{\mathbf{P}}) \leq ||\mathbf{X}\mathbf{P} - \mathbf{F}||_F^2 + \gamma Tr(\mathbf{P}^{\mathrm{T}}\mathbf{\Gamma}\mathbf{P}) \tag{18}$$

By adding $\gamma\sum_{i=1}^{d} \frac{\epsilon}{2\sqrt{\mathbf{P}_{i'}^{\mathrm{T}}\mathbf{P}_{i'}+\epsilon}}$ to the both sides of the inequality (18) and substituting $\mathbf{\Gamma}$, the inequality can be rewritten as

$$\begin{aligned} &||\mathbf{X}\widehat{\mathbf{P}} - \mathbf{F}||_F^2 + \gamma\sum_{i=1}^{d} \frac{\widehat{\mathbf{P}}_{i'}^{\mathrm{T}}\widehat{\mathbf{P}}_{i'} + \epsilon}{2\sqrt{\mathbf{P}_{i'}^{\mathrm{T}}\mathbf{P}_{i'} + \epsilon}} \leq \\ &||\mathbf{X}\mathbf{P} - \mathbf{F}||_F^2 + \gamma\sum_{i=1}^{d} \frac{\mathbf{P}_{i'}^{\mathrm{T}}\mathbf{P}_{i'} + \epsilon}{2\sqrt{\mathbf{P}_{i'}^{\mathrm{T}}\mathbf{P}_{i'} + \epsilon}} \end{aligned} \tag{19}$$

On the other hand, according to the **Lemma** 1 in [Nie *et al.*, 2010], we can obtain that for any positive number u and v, we can have

$$\sqrt{u} - \frac{\sqrt{u}}{2\sqrt{v}} \leq \sqrt{v} - \frac{\sqrt{v}}{2\sqrt{v}} \tag{20}$$

Dataset	Feature dimension	LapScor	SPEC	MRSF	MVFS	AUMFS	AMFS	ACSL
MSRC-V1	100	0.2867	**0.2952**	0.2838	0.2762	0.2810	0.28571	**0.3000**
	200	0.2952	0.2905	**0.3152**	0.2905	**0.3143**	0.2895	0.3124
	300	0.2905	**0.3119**	0.2895	0.2833	0.2833	0.2952	**0.3124**
	400	0.2952	**0.3181**	0.3057	0.3000	0.2952	0.2924	**0.3219**
	500	0.3038	0.2976	0.3038	**0.3095**	0.3048	0.2990	**0.3400**
Handwritten Numeral	100	0.5844	0.4795	**0.6207**	0.5938	0.3345	0.3302	**0.6106**
	200	**0.6148**	0.5520	0.6002	0.5820	0.4225	0.4226	**0.6389**
	300	**0.5980**	0.5384	**0.6028**	0.5737	0.4757	0.4497	0.5930
	400	0.6068	**0.6102**	0.5890	0.5808	0.4909	0.4755	**0.6327**
	500	**0.5909**	0.5666	0.5795	0.5888	0.4889	0.5006	**0.5969**
Youtube	100	**0.2873**	**0.2873**	0.2851	0.2717	0.1305	0.2165	0.2861
	200	**0.2896**	0.2840	0.2754	0.2774	0.1274	0.2313	**0.2924**
	300	0.2835	0.2832	**0.2862**	0.2828	0.1357	0.2374	**0.2906**
	400	0.2862	**0.2889**	0.2779	0.2807	0.1329	0.2433	**0.2993**
	500	**0.2857**	0.2853	0.2802	0.2854	0.1329	0.2546	**0.3003**
Outdoor Scene	100	0.3687	0.3327	0.3707	0.2044	0.4231	**0.4313**	**0.5845**
	200	0.3619	0.3295	0.3501	0.2104	0.4656	**0.4816**	**0.5616**
	300	0.3634	0.3740	0.3576	0.2150	**0.4949**	0.4854	**0.5801**
	400	0.3804	0.3653	0.3679	0.2153	**0.5061**	0.4926	**0.5927**
	500	0.3574	0.3620	0.3687	0.2255	0.5003	**0.5045**	**0.6103**

Table 1: ACC of different methods with different numbers of selected features by using K-means for clustering.

Then, we can obtain that

$$\begin{aligned}&\gamma\sum_{i=1}^{d}\sqrt{\widehat{\mathbf{P}}_{i'}^{\mathrm{T}}\widehat{\mathbf{P}}_{i'}+\epsilon}-\gamma\sum_{i=1}^{d}\frac{\widehat{\mathbf{P}}_{i'}^{\mathrm{T}}\widehat{\mathbf{P}}_{i'}+\epsilon}{2\sqrt{\mathbf{P}_{i'}^{\mathrm{T}}\mathbf{P}_{i'}+\epsilon}}\\&\leq\gamma\sum_{i=1}^{d}\sqrt{\mathbf{P}_{i'}^{\mathrm{T}}\mathbf{P}_{i'}+\epsilon}-\gamma\sum_{i=1}^{d}\frac{\mathbf{P}_{i'}^{\mathrm{T}}\mathbf{P}_{i'}+\epsilon}{2\sqrt{\mathbf{P}_{i'}^{\mathrm{T}}\mathbf{P}_{i'}+\epsilon}}\end{aligned}\tag{21}$$

By summing the above inequalities (19) and (21), we arrive at

$$\begin{aligned}&||\mathbf{X}\widehat{\mathbf{P}}-\mathbf{F}||_F^2+\gamma\sum_{i=1}^{d}\sqrt{\widehat{\mathbf{P}}_{i'}^{\mathrm{T}}\widehat{\mathbf{P}}_{i'}+\epsilon}\\&\leq||\mathbf{X}\mathbf{P}-\mathbf{F}||_F^2+\gamma\sum_{i=1}^{d}\sqrt{\mathbf{P}_{i'}^{\mathrm{T}}\mathbf{P}_{i'}+\epsilon}\end{aligned}\tag{22}$$

We can derive that

$$||\mathbf{X}\widehat{\mathbf{P}}-\mathbf{F}||_F^2+\gamma||\widehat{\mathbf{P}}||_{2,1}\leq||\mathbf{X}\mathbf{P}-\mathbf{F}||_F^2+\gamma||\mathbf{P}||_{2,1}\tag{23}$$

□

The convergence of solving Algorithm 1 can be proven by the following theorem.

Theorem 3. *The iterative optimization in Algorithm 1 can monotonically decrease the objective function of problem (4) until convergence.*

Proof. As shown in **Theorem** 2, updating $\mathbf{P}$ will monotonically decrease the objective function in problem (4) (t is number of iterations).

$$\mathbf{\Omega}(\mathbf{P}^{(t)},\mathbf{F},\mathbf{S},\mathbf{W})\geq\mathbf{\Omega}(\mathbf{P}^{(t+1)},\mathbf{F},\mathbf{S},\mathbf{W})\tag{24}$$

By fixing other variables and updating $\mathbf{F}$, the objective function in Eq.(8) is convex (The Hessian matrix of the Lagrangian function of Eq.(8) is positive semidefinite [Alavi, 1991]). Therefore, we can obtain that

$$\mathbf{\Omega}(\mathbf{P},\mathbf{F}^{(t)},\mathbf{S},\mathbf{W})\geq\mathbf{\Omega}(\mathbf{P},\mathbf{F}^{(t+1)},\mathbf{S},\mathbf{W})\tag{25}$$

By fixing other variables and updating $\mathbf{S}$, optimizing the Eq.(13) is a typical Quadratic programming problem. The Hessian matrix of the Lagrangian function of problem (13) is also $\mathbf{1}_N\mathbf{1}_N^{\mathrm{T}}$ that is positive semidefinite. Therefore, we can obtain that

$$\mathbf{\Omega}(\mathbf{P},\mathbf{F},\mathbf{S}^{(t)},\mathbf{W})\geq\mathbf{\Omega}(\mathbf{P},\mathbf{F},\mathbf{S}^{(t+1)},\mathbf{W})\tag{26}$$

By fixing other variables and updating $\mathbf{W}$, the Hessian matrix of Eq.(16) is $\mathbf{B}_j^{\mathrm{T}}\mathbf{B}_j$. It is positive semidefinite as $\mathbf{W}_j^{\mathrm{T}}\mathbf{B}_j^{\mathrm{T}}\mathbf{B}_j\mathbf{W}_j)=||\mathbf{B}_j\mathbf{W}_j||_F^2\geq 0$. Hence, the objective function for optimizing $\mathbf{W}$ is also convex. Then, we arrive at

$$\mathbf{\Omega}(\mathbf{P},\mathbf{F},\mathbf{S},\mathbf{W}^{(t)})\geq\mathbf{\Omega}(\mathbf{P},\mathbf{F},\mathbf{S},\mathbf{W}^{(t+1)})\tag{27}$$

□

4 Experiments

4.1 Experimental Datasets

1) **MSRC-v1** [Winn and Jojic, 2005]. The dataset contains 240 images in 8 class as a whole. Following the setting in [Grauman and Darrell, 2006], we select 7 classes composed of tree, building, airplane, cow, face, car, bicycle and each class has 30 images. We extract 5 visual features from each image: color moment with dimension 48, GIST with 512 dimension, SIFT with dimension 1230, CENTRIST feature with 210 dimension, and local binary pattern (LBP) with 256 dimension. 2) **Handwritten Numeral** [van Breukelen *et al.*, 1998]. This dataset is comprised of 2,000 data points from 0 to 9 digit classes. 6 features are used to represent each digit. They are 76 dimensional Fourier coefficients of the character shapes, 216 dimensional profile correlations, 64 dimensional Karhunen-love coefficients, 240 dimensional pixel averages in 2×3 windows, 47 dimensional Zernike moment and 6 dimensional morphological features. 3) **Youtube** [Liu *et al.*, 2009]. This real-world dataset is collected from Youtube. It contains intended camera motion, variations of the object scale, viewpoint, illumination and cluttered background. The dataset is comprised of 1,596 video sequences in 11 actions. 4) **Outdoor Scene** [Monadjemi *et al.*, 2002]. The outdoor scene dataset contains 2,688 color images that belong to 8 outdoor scene categories. 4 visual features are extracted from each image: color moment with dimension 432, GIST with dimension 512, HOG with dimension 256, and LBP with dimension 48.

Dataset	Feature dimension	LapScor	SPEC	MRSF	MVFS	AUMFS	AMFS	ACSL
MSRC-v1	100	**0.1653**	**0.1930**	0.1555	0.1362	0.1146	0.12681	0.1635
	200	0.1730	0.1518	0.1754	0.1502	**0.1799**	0.1591	**0.1875**
	300	0.1632	0.1637	**0.1713**	0.1358	0.1341	0.1609	**0.1912**
	400	0.1815	**0.2195**	0.1787	0.1407	0.1716	0.1595	**0.1905**
	500	0.1672	**0.2027**	0.1813	0.1798	0.1735	0.1670	**0.2146**
Handwritten Numeral	100	**0.5967**	0.4751	0.5927	0.5485	0.2738	0.2744	**0.6403**
	200	**0.6050**	0.5413	0.5943	0.5538	0.3720	0.3718	**0.6513**
	300	0.5962	**0.6068**	**0.6051**	0.5584	0.4101	0.4013	0.5932
	400	0.6014	0.6010	**0.6015**	0.5690	0.4436	0.4423	**0.6025**
	500	**0.6078**	0.5799	**0.5983**	0.5974	0.4796	0.4831	0.5926
Youtube	100	**0.2690**	0.2683	0.2610	0.2531	0.0121	0.1280	**0.2705**
	200	**0.2693**	0.2688	0.2561	0.2604	0.0108	0.1474	**0.2699**
	300	0.2627	**0.2673**	**0.2677**	0.2605	0.0152	0.1597	0.2570
	400	**0.2670**	0.2647	0.2606	**0.2736**	0.0142	0.1817	**0.2743**
	500	0.2641	0.2696	0.2635	**0.2771**	0.0123	0.1982	**0.2736**
Outdoor Scene	100	0.2228	0.1933	0.2203	0.0595	**0.3314**	0.3267	**0.4717**
	200	0.2174	0.2023	0.2021	0.0522	**0.3801**	0.3772	**0.4860**
	300	0.2264	0.2414	0.2142	0.0562	**0.4157**	0.3975	**0.4838**
	400	0.2358	0.2337	0.2211	0.0588	**0.4152**	0.4023	**0.5111**
	500	0.2163	0.2316	0.2196	0.0769	**0.4210**	0.4159	**0.5211**

Table 2: NMI of different methods with different numbers of selected features by using K-means for clustering.

4.2 Experimental Setting

Baselines. We compare ACSL with several representative unsupervised multi-view feature selection methods on clustering performance. The compared methods include, three single view feature selection approaches (Laplacian score (LapScor) [He *et al.*, 2005], spectral feature selection (SPEC) [Zhao and Liu, 2007] and minimum redundancy spectral feature selection (MRSF) [Zhao *et al.*, 2010]), and three multi-view feature selection approach (adaptive multi-view feature selection (AMFS) [Wang *et al.*, 2016], multi-view feature selection (MVFS) [Tang *et al.*, 2013] and adaptive unsupervised multi-view feature selection (AUMFS) [Feng *et al.*, 2013]). **Evluation Metrics.** We employ standard metrics: clustering accuracy (ACC) and normalized mutual information (NMI), for performance comparison. Each experiment is performed 50 times and the mean results are reported. **Parameter Setting.** In implementation of all methods, the neighbor graph is adopted to construct the initial affinity matrices. The number of neighbors is set to 10 in all methods. In ACSL, α, β, γ are chosen from 10^{-4} to 10^{4}. The parameters in all compared approaches are carefully adjusted to report the best results.

4.3 Comparison Results

The comparison results measured by ACC and NMI are reported in Table 1 and Table 2, respectively. For these metrics, the higher value indicates the better feature selection performance. Each metric penalizes or favors different properties in feature selection. Hence, we report results on these diverse measures to perform a comprehensive evaluation. The obtained results demonstrate that ACSL can achieve superior or at least comparable performance than the compared approaches. The promising performance of ACSL is attributed to the reason that the proposed collaborative similarity structure learning with proper neighbor assignment could positively facilitate the ultimate multi-view feature selection.

4.4 Parameter and Convergence Experiment

We investigate the impact of parameters α, β and γ in Eq.(4) on the performance of ACSL. Specifically, we vary one parameter by fixing the others. Figure 1 presents the main results on **MSRC-V1**. The obtained results clearly show that ACSL is robust to the involved three parameters. Figure 2 records the variations of the objective function value in Eq.(4) with the number of iterations on **MSRC-V1** and **Handwritten Numeral**. We can easily observe that the convergence curves become stable within about 5 iterations. The fast convergence ensures the optimization efficiency of ACSL.

5 Conclusion

In this paper, we propose an adaptive collaborative similarity structure learning for multi-view feature selection. Different from existing approaches, we integrate collaborative similarity learning and feature selection into a unified framework. The collaborative similarity structure with the ideal neighbor assignment and similarity combination weights are adaptively learned to positively facilitate the subsequent feature selection. Simultaneously, the feature selection can supervise the similarity learning process to dynamically construct the desirable similarity structure. Experiments show the superiority of the proposed approach.

References

[Alavi, 1991] Y. Alavi. The laplacian spectrum of graphs. *Graph theory, combinatorics, and applications*, 2(12):871–898, 1991.

[Cheng and Shen, 2016] Zhiyong Cheng and Jialie Shen. On very large scale test collection for landmark image search benchmarking. *Signal Processing*, 124:13 – 26, 2016.

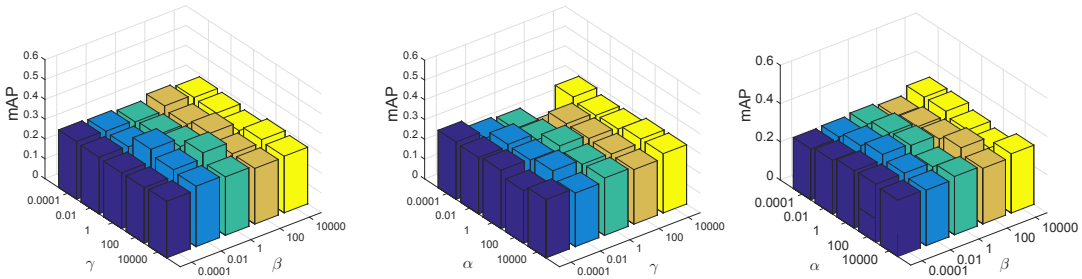
[Cheng *et al.*, 2016] Zhiyong Cheng, Jialie Shen, and Haiyan Miao. The effects of multiple query evidences on social image retrieval. *Multimedia Systems*, 22(4):509–523, 2016.

[Feng *et al.*, 2013] Yinfu Feng, Jun Xiao, Yueting Zhuang, and Xiaoming Liu. Adaptive unsupervised multi-view feature selection for visual concept recognition. In *ACCV*, 2013.

[Grauman and Darrell, 2006] K Grauman and T Darrell. Unsupervised learning of categories from sets of partially matching image features. In *CVPR*, pages 19–25, 2006.

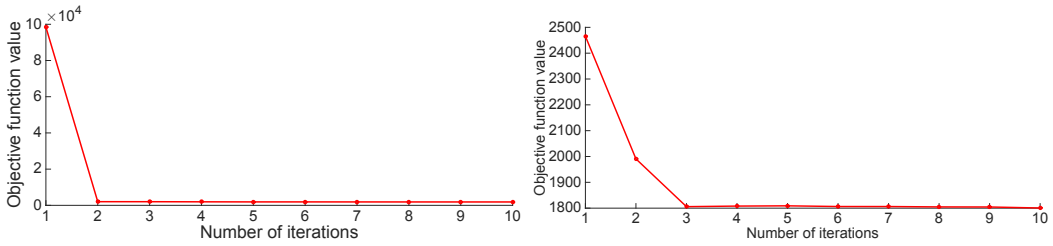
[He *et al.*, 2005] Xiaofei He, Deng Cai, and Partha Niyogi. Laplacian score for feature selection. In *NIPS*, pages 507–514, 2005.

[Huang *et al.*, 2015] Jin Huang, Feiping Nie, and Heng Huang. A new simplex sparse learning model to measure data similarity for clustering. In *IJCAI*, pages 3569–3575, 2015.



(a) α is fixed to 10^4 (b) β is fixed to 10^{-2} (c) γ is fixed to 10^3

Figure 1: Clustering accuracy variations with parameters α, β, γ in Eq.(4) on MSRC-V1.



(a) MSRC-V1 (b) Handwritten Numeral

Figure 2: Variations of the objective function value in Eq.(4) with the number of iterations on MSRC-V1 and Handwritten Numeral.

[K., 1949] Fan K. On a theorem of weyl concerning eigenvalues of linear transformations. *Proc Natl Acad Sci U S A*, 11(35):652–655, 1949.

[Li and Liu, 2017] Jundong Li and Huan Liu. Challenges of feature selection for big data analytics. *IEEE Intelligent Systems*, 32(2):9–15, 2017.

[Li *et al.*, 2017] Yun Li, Tao Li, and Huan Liu. Recent advances in feature selection and its applications. *Knowledge and Information Systems*, 53(3):551–577, 2017.

[Liu *et al.*, 2009] J. Liu, Yang Yang, and M. Shah. Learning semantic visual vocabularies using diffusion distance. In *CVPR*, pages 461–468, 2009.

[Liu *et al.*, 2016] An-An Liu, Wei-Zhi Nie, Yue Gao, and Yu-Ting Su. Multi-modal clique-graph matching for view-based 3d model retrieval. *TIP*, 25(5):2103–2116, 2016.

[Liu *et al.*, 2017] A. A. Liu, Y. T. Su, W. Z. Nie, and M. Kankanhalli. Hierarchical clustering multi-task learning for joint human action grouping and recognition. *TPAMI*, 39(1):102–114, 2017.

[Monadjemi *et al.*, 2002] A. Monadjemi, B. T. Thomas, and M. Mirmehdi. Experiments on high resolution images towards outdoor scene classification. *Computer Vision Winter Workshop*, 2002.

[Nie *et al.*, 2010] Feiping Nie, Heng Huang, Xiao Cai, and Chris H. Q. Ding. Efficient and robust feature selection via joint l21-norms minimization. In *NIPS*, pages 1813–1821, 2010.

[Nie *et al.*, 2014] Feiping Nie, Xiaoqian Wang, and Heng Huang. Clustering and projected clustering with adaptive neighbors. In *KDD*, pages 977–986, 2014.

[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.

[van Breukelen *et al.*, 1998] M P. W van Breukelen, D M. J Tax, and J E den Hartog. Handwritten digit recognition by combined classifiers,. *Kybernetika*, 34:381–386, 1998.

[Wang *et al.*, 2016] Zhao Wang, Yinfu Feng, Tian Qi, Xiaosong Yang, and Jian J. Zhang. Adaptive multi-view feature selection for human motion retrieval. *Signal Processing*, 120(C):691 – 701, 2016.

[Winn and Jojic, 2005] J. Winn and N. Jojic. Locus: learning object classes with unsupervised segmentation. In *ICCV*, pages 756–763, 2005.

[Zhao and Liu, 2007] Zheng Zhao and Huan Liu. Spectral feature selection for supervised and unsupervised learning. In *ICML*, pages 1151–1157, 2007.

[Zhao *et al.*, 2010] Zheng Zhao, Lei Wang, and Huan Liu. Efficient spectral feature selection with minimum redundancy. In *AAAI*, 2010.

[Zhu *et al.*, 2015] Lei Zhu, Jialie Shen, Hai Jin, Ran Zheng, and Liang Xie. Content-based visual landmark search via multimodal hypergraph learning. *TCYB*, 45(12):2756–2769, 2015.

[Zhu *et al.*, 2016a] L. Zhu, J. Shen, L. Xie, and Z. Cheng. Unsupervised topic hypergraph hashing for efficient mobile image retrieval. *TCYB*, 47(11):3941–3954, 2016.

[Zhu *et al.*, 2016b] Lei Zhu, Jialie She, Xiaobai Liu, Liang Xie, and Liqiang Nie. Learning compact visual representation with canonical views for robust mobile landmark search. In *IJCAI*, pages 3959–3965, 2016.

[Zhu *et al.*, 2017a] L. Zhu, Z. Huang, X. Liu, X. He, J. Sun, and X. Zhou. Discrete multimodal hashing with canonical views for robust mobile landmark search. *TMM*, 19(9):2066–2079, 2017.

[Zhu *et al.*, 2017b] Lei Zhu, Jialie Shen, Liang Xie, and Zhiyong Cheng. Unsupervised visual hashing with semantic assistant for content-based image retrieval. *TKDE*, 29(2):472–486, 2017.