Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17)

Multi-View Clustering and Semi-Supervised Classification with Adaptive Neighbours

Feiping Nie,[1] **Guohao Cai,**[1] **Xuelong Li**[2]
[1]School of Computer Science and Center for OPTical IMagery Analysis and Learning (OPTIMAL),
Northwestern Polytechnical University, Xi'an 710072, Shaanxi, P. R. China
[2]Center for OPTical IMagery Analysis and Learning (OPTIMAL),
Xi'an Institute of Optics and Precision Mechanics,
Chinese Academy of Sciences, Xi'an 710119, Shaanxi, P. R. China
{feipingnie@gmail.com, caiguohao1@gmail.com, xuelong_li@opt.ac.cn}

Abstract

Due to the efficiency of learning relationships and complex structures hidden in data, graph-oriented methods have been widely investigated and achieve promising performance in multi-view learning. Generally, these learning algorithms construct informative graph for each view or fuse different views to one graph, on which the following procedure are based. However, in many real world dataset, original data always contain noise and outlying entries that result in unreliable and inaccurate graphs, which cannot be ameliorated in the previous methods. In this paper, we propose a novel multi-view learning model which performs clustering/semi-supervised classification and local structure learning simultaneously. The obtained optimal graph can be partitioned into specific clusters directly. Moreover, our model can allocate ideal weight for each view automatically without additional weight and penalty parameters. An efficient algorithm is proposed to optimize this model. Extensive experimental results on different real-world datasets show that the proposed model outperforms other state-of-the-art multi-view algorithms.

Introduction

In many real world applications, such as multi-camera surveillance system, abundant data collected from different views are available. Usually each view captures partial information but they together admit the same clustering structure. Nowadays, we have easier access to data that contain heterogeneous features representing samples from different views in many scientific fields, such as pattern recognition, computer vision, genetics, data mining, etc. For example, in visual data, an image could be represented by different descriptors, such as SIFT (Lowe 2004), HOG (Dalal, Triggs, and Schmid 2006), GIST (Oliva and Torralba 2001), LBP (Ojala, Pietikäinen, and Mäenpää 2002); in ResearchIndex network, the keywords of a specific paper and its citations can be regarded as two separate views; in biological data, each human gene can be measured by gene expression, Array-comparative genomic hybridization (arrayCGH), Single-nucleotide polymorphism (SNP) and methylation.

Numerous clustering methods have been proposed in the past decades, it might be satisfying for an individual view of data to accomplish some clustering work, but methods which properly combine various views containing different fractional information will improve the clustering performance. Recently, varieties of multi-view clustering algorithms have been proposed. (Selee et al. 2007) introduced a new tensor decomposition called Implicit Slice Canonical Decomposition (IMSCAND) in which each similarity matrix is stored as a slice in a tensor. (Chaudhuri et al. 2009) proposed multi-view clustering method via Canonical Correlation Analysis (CCA), it computes two sets of variables and maximizes the correlation between them in the embedded space. (Kumar, Rai, and III 2011) proposed a co-regularized approach for multi-view spectral clustering in which they co-regularize the clustering hypotheses to make different graphs agree with each other. (Cai et al. 2011) proposed multi-modal spectral clustering (MMSC) algorithm to integrate heterogeneous image feature, it learns a commonly shared Laplacian matrix by unifying different modals and add a non-negative relaxation to improve the robustness of image clustering. (Li et al. 2015) proposed a new large-scale multi-view spectral approach (MVSC) based on bipartite graph. In general, graph-based methods are pretty conspicuous for efficiency and excellent clustering performance.

For multiple learning in semi-supervised learning, Co-training (Blum and Mitchell 1998) is a representative paradigm. It firstly trains two classifiers with labeled data, and classifies the unlabeled data separately. Next some predicted data that are of most confidence are added to the other classifier's training set, then the procedure repeats. (Tian and Kuang 2010) proposed an alignment-based semi-supervised learning model to classify gene expression data samples by seeking an optimal alignment between different samples' probe series. Under the manifold assumption, graph-based methods trade labeled and unlabeled examples as vertices of a graph and utilize edges to propagate information from labeled ones to unlabeled ones. (Cai et al. 2013) introduced an adaptive multi-modal semi-supervised classification (AMMSS) algorithm which considers each type of feature as one modality, it learns a shared class indicator matrix and weights for different modalities. (Karasuyama and Mamitsuka 2013) use sparse weights to linearly combine different graphs to implement label propagation (SMGI).

Although graph-based multi-view learning methods achieve state-of-the-art performance, there still exist some limits. For one thing, such methods conduct the following procedure base on the constructed similarity matrix from original data but rarely modify it. Real world datasets always contain noise and outlying entries that result in the unreliable similarity matrix which will impair the finally performance. For another, those methods combining different views often have additional weight parameters to set, which is unsatisfactory especially in unsupervised clustering task.

In this paper, we propose a novel multi-view learning model, named Multi-view Learning with Adaptive Neighbours (MLAN). There are several benefits of our approach: The proposed approach performs multi-view clustering/semi-supervised classification and local manifold structure learning simultaneously, modifying similarity matrix during each iteration until reach to the optimal one; No explicit weight parameter in our model, it can learn the weight coefficient automatically after finite iterations, which has conspicuous advantage in unsupervised clustering work; Comprehensive experiments on several real-world data sets show the effectiveness of proposed approach, and demonstrate the advantage over other state-of-the-art methods.

Methodology

In this section, we will firstly introduce the assignment of adaptive neighbours. Then we will address the issue of acquiring optimal linear combination of multiple graphs, the weight coefficient and corresponding penalty parameter can all be omitted.

Notations are summarized here throughout the paper. All the matrices are written as uppercase. For a matrix $M \in \mathbb{R}^{n\times d}$, the i-th row and the (ij)-th element of M are denoted by m_i and m_{ij} , respectively. The transpose of matrix M is denoted by M^T. The trace of matrix M is denoted by $Tr(M)$. The ℓ_2-norm of vector v is denoted by $\|v\|_2$. $\mathbf{1}$ denotes a column vector with all the elements as one, and the identity matrix is denoted by I. $\overline{x}$ and $\sigma(x)$ denote the average value and standard deviation of vector x, respectively.

Adaptive Local Structure Learning

One important factor to the success of graph-based methods is the preserving local manifold structure, high-dimensional data is considered to contain low-dimensional manifold structure (Nie, Li, and Li 2016), so the obtained similarity matrix is crucial to the ultimate performance. Given a set of data points $\{x_1, x_2, \cdots, x_n\}$, denote data matrix $X \in \mathbb{R}^{n\times d}$, where n is the number of data points and d is the dimension of features, we adopt the data preprocessing proposed in (Liu et al. 2013). In details, $x_i \leftarrow (x_i - \overline{x}))/\sigma(x)$. For each data point x_i, it belongs to one of the c classes, and can be connected by all the data points with the probability s_{ij}, and such probability can be seen as the similarity between them. Closer samples should have larger probability, thus s_{ij} has the negative correlation with the distance between x_i and x_j. The determination of probability s_{ij} can be seen as solving following problem:

$$\begin{aligned} &\min_{s_i\in^{n\times 1}} \sum_{i,j}^{n} ||x_i - x_j||_2^2 s_{ij} + \alpha||S||_F^2 \\ &s.t. \quad \forall i, s_i^T\mathbf{1} = 1, 0 \le s_{ij} \le 1 \end{aligned} \tag{1}$$

where s_i is a vector with j-th element as s_{ij} in similarity matrix S. The second item is added for the consideration that there would be a trivial solution where only the nearest data point to the x_i is assigned probability 1 and all the other points' similarity would be 0 without such penalty item. In spectral analysis, $L_S = D_S - (S^T + S)/2$ is called Laplacian matrix, where the degree matrix D_S in $S \in \mathbb{R}^{n\times n}$ is the diagonal matrix whose i-th diagonal element is $\sum_j (s_{ij} + s_{ji})/2$. Given the class indicator matrix $F = [f_1, \cdots, f_n]$, classical spectral clustering can be written as

$$\min_{F\in\mathbb{R}^{n\times c}, F^TF=I} Tr(F^T L_S F) \tag{2}$$

Since the similarity matrix S does not has exact c connected components, previous methods have to resort to other discretization procedures like K-means performing on F to obtain the final results (Huang, Nie, and Huang 2013). There is an important property of Laplacian matrix (Mohar 1991), (Chung 1997)

Theorem 1. *The multiplicity c of the eigenvalue 0 of the Laplacian matrix L_S (nonnegative) is equal to the number of connected components in the graph with the similarity matrix S.*

In view of the above consideration, (Nie, Wang, and Huang 2014) added a rank constrain to the L_S in problem 1 according to the Theorem 1:

$$\begin{aligned} &\min_{s_i\in^{n\times 1}} \sum_{i,j}^{n} ||x_i - x_j||_2^2 s_{ij} + \alpha||S||_F^2 \\ &s.t. \quad \forall i, s_i^T\mathbf{1} = 1, 0 \le s_{ij} \le 1, rank(L_S) = n - c \end{aligned} \tag{3}$$

It assigns adaptive neighbours to each of samples, which means that the similarity between data points will change, so similarity matrix S will be modified until it contains exact c connected component. Namely, not only the indicator matrix F can be learned, different from the traditional spectral clustering methods, our model can also learn similarity matrix S simultaneously. The learned S can be used for clustering directly according to Tarjan's strongly connected components algorithm (Tarjan 1972).

Multi-view Data Fusion

For multi-view data, denote $X_1, X_2, \cdots, X_v$ be the data matrix of each view. $X_v \in \mathbb{R}^{n\times d^v}$, where n is the number of data and d^v is the feature dimension of the v-th view. As for graph-based methods, each view can construct similarity graph and maximize the performance quality on its own. In the context of multi-view clustering, there is an inherent problem that all methods have to deal with elaborately: when maximizing the within-view clustering quality, the clustering consistency across different views should be taken into

consideration. The rough way that combining multiple views directly through similarity matrix addition or feature concatenation would not help improve the clustering performance, for fallible similarity matrix could lead to suboptimal result. A more reasonable manner is to integrate these views with suitable weights $w_v(v = 1, \cdots, V)$, and an extra parameter γ is needed to keep weights distribution smooth. Basically, there are two kind of models, if adding such parameters to Eq. (3), it tunes to be:

$$\begin{aligned} \min_{S,w_v} \quad & \sum_v (w_v)^\gamma \sum_{i,j} ||x_i^v - x_j^v||_2^2 s_{ij} + \alpha ||S||_F^2 \\ s.t. \quad & s_i^T \mathbf{1} = 1, 0 \le s_{ij} \le 1, w_v^T \mathbf{1} = 1, 0 \le w_v \le 1, \\ & rank(L_S) = n - c \end{aligned} \tag{4}$$

where γ is the non-negative scalar, it could be regularization parameter in another model:

$$\begin{aligned} \min_{S,w_v} \quad & \sum_v (w_v \sum_{i,j} ||x_i^v - x_j^v||_2^2 s_{ij}) + \gamma ||w_v||_2^2) + \alpha ||S||_F^2 \\ s.t. \quad & s_i^T \mathbf{1} = 1, 0 \le s_{ij} \le 1, w_v^T \mathbf{1} = 1, 0 \le w_v \le 1, \\ & rank(L_S) = n - c \end{aligned} \tag{5}$$

For unsupervised learning methods, the less parameter to be set, the strong robustness they possess. On the other hand, since parameters can be searched in a large range, methods with parameters like the above form often show better result than parameter-free methods. It's really elusive to pursue good performance while rely less on parameter searching. However, we will propose one to alleviate such challenging problem in the next section.

Multi-view Learning with Adaptive Neighbours

In this paper, we propose a novel multi-view learning with adaptive neighbours method as the following form:

$$\begin{aligned} \min_{S} \quad & \sum_v \sqrt{\sum_{i,j} ||x_i^v - x_j^v||_2^2 s_{ij}} + \alpha ||S||_F^2, \\ s.t. \quad & s_i^T \mathbf{1} = 1, 0 \le s_{ij} \le 1, rank(L_S) = n - c \end{aligned} \tag{6}$$

where each view shares the same similarity matrix, thus the goal of assigning each data point to the most suitable cluster in each view and ensuring clustering consistency across views is achieved. There is no weight hyperparameter explicitly defined in our model. The Lagrange function of Eq. (6) can be written as

$$\sum_v \sqrt{\sum_{i,j} ||x_i^v - x_j^v||_2^2 s_{ij}} + \alpha ||S||_F^2 + \mathcal{G}(\Lambda, S) \tag{7}$$

where Λ is the Lagrange multiplier, $\mathcal{G}(\Lambda, S)$ is the formalized term derived from constraints. Taking the derivative of Eq. (7) w.r.t S and setting the derivative to zero, we have

$$\sum_v w_v \frac{\partial \sum_{i,j} ||x_i^v - x_j^v||_2^2 s_{ij}}{\partial S} + \frac{\alpha \partial ||S||_F^2}{\partial S} + \frac{\partial \mathcal{G}(\Lambda, S)}{\partial S} = 0 \tag{8}$$

where

$$w_v = 1 \Big/ 2\sqrt{\sum_{i,j} ||x_i^v - x_j^v||_2^2 s_{ij}} \tag{9}$$

we can see that w_v is dependent on the target variable S, so that Eq. (8) cannot be directly solved. But if w_v is set to be stationary, Eq. (8) can be considered accounting for following problem

$$\begin{aligned} \min_{S} \quad & \sum_v w_v \sum_{i,j} ||x_i^v - x_j^v||_2^2 s_{ij} + \alpha ||S||_F^2, \\ s.t. \quad & s_i^T \mathbf{1} = 1, 0 \le s_{ij} \le 1, rank(L_S) = n - c \end{aligned} \tag{10}$$

Under the assumption that w_v is stationary, the Lagrange function of Eq. (6) also apply to Eq. (10), if we calculate S from Eq. (10), the value of w_v can be updated correspondingly, which inspires us to optimize Eq. (6) in an alternative way. After optimization, S tune to be $\widehat{S}$, according to Eq. (8), $\widehat{S}$ is as least a local optimal solution to problem (6). Similarly, w_v tune to be $\widehat{w_v}$, and they are exactly the learned weights which linearly combining different graphs.

Optimization Algorithm

To solve the challenging problem (6), we should solve problem (10) iteratively. In the iterative procedure, parameters are updated one by one. The specific parameter updated in the last step could be seen as a constant during current step.

Clustering

Denote $\sigma_i(L_S)$ is the i-th smallest eigenvalue of L_S, because L_S is positive semi-definite, $\sigma_i(L_S) \ge 0$. So the constraint $\text{rank}(L_S) = n - c$ will be ensured if $\sum_{i=1}^c \sigma_i(L_S) = 0$. According to Ky Fan's Theorem (Fan 1949), we have

$$\sum_{i=1}^c \sigma_i(L_S) = \min_{F \in \mathbb{R}^{n \times c}, F^T F = I} Tr(F^T L_S F) \tag{11}$$

Then problem (10) is equivalent to the following problem

$$\begin{aligned} & \min_{S,F} \sum_v w_v \sum_{i,j} ||x_i^v - x_j^v||_2^2 s_{ij} + \alpha ||S||_F^2 + 2\lambda Tr(F^T L_S F) \\ & s.t. \ s_i^T \mathbf{1} = 1, 0 \le s_{ij} \le 1, F^T F = I \end{aligned} \tag{12}$$

where λ is a very large number, the optimal solution to the problem (12) will make equation $\sum_{i=1}^k \sigma_i(L_S) = 0$ hold.

Fix S, update w_v and F When S is fixed, we can easily calculate the value of w_v by Eq. (9). So the first and second item of problem (12) could be seen as constant, then it transforms into:

$$\min_{F \in \mathbb{R}^{n \times c}, F^T F = I} Tr(F^T L_S F) \tag{13}$$

the optimal solution F is formed by the c eigenvectors corresponding to the c smallest eigenvalues of L_S.

Fix w_v and F, update S Since w_v is fixed, the first item of Eq. (10) can be replaced as $\sum_{i,j}\sum_{v} w_v||x_i^v - x_j^v||_2^2 s_{ij}$. Denote $d_{ij}^x = \sum_v w_v||x_i^v - x_j^v||_2^2$, which represents the weighted distance between data points x_i and x_j. Then the problem (10) becomes

$$\begin{aligned} \min_{S} \quad & \sum_{i,j} (d_{ij}^x s_{ij} + \alpha s_{ij}^2) + 2\lambda Tr(F^T L_S F) \\ s.t. \quad & s_i^T \mathbf{1} = 1, 0 \le s_{ij} \le 1 \end{aligned} \tag{14}$$

There is an elementary but very important equation in spectral analysis

$$\sum_{i,j} ||f_i - f_j||_2^2 s_{ij} = 2Tr(F^T L_S F) \tag{15}$$

Denote $d_{ij}^f = ||f_i - f_j||_2^2$, note that the problem (14) is independent between different i, we can deal with following problem individually for each i:

$$\begin{aligned} \min_{s_i} \quad & \sum_{j=1}^{n} (d_{ij}^x s_{ij} + \alpha s_{ij}^2 + \lambda d_{ij}^f s_{ij}) \\ s.t. \quad & s_i^T \mathbf{1} = 1, 0 \le s_{ij} \le 1 \end{aligned} \tag{16}$$

Denote $d_i \in \mathbb{R}^{n\times 1}$ is a vector with the j-th element as $d_{ij} = d_{ij}^x + \lambda d_{ij}^f$, then the above problem can be written as follow:

$$\min_{s_i} \quad ||s_i + \frac{1}{2\alpha} d_i||_2^2 \quad s.t. \quad s_i^T \mathbf{1} = 1, 0 \le s_{ij} \le 1 \tag{17}$$

The intermediate variable α can be determined using the number of adaptive neighbours, by saying adaptive, we mean that the k nearest neighbours to any data point x_i are not steady, they change in every iteration since the weighted distance d_{ij}^x between every pair of x_i and x_j is updated. The determination of the α value will be described in the next section.

Extend to Semi-supervised Classification

Denote l and u are the number of labeled and unlabeled points. Denote $Y_l = [y_1, \cdots, y_l]^T$, where $y_i \in \mathbb{R}^{c\times 1}$ is the known indicator vector for the i-th sample, y_i is one-hot and the element $y_{ij} = 1$ means that the i-th sample belongs to the j-th class. Without loss of generality, we rearrange all the points and let the front l points be labeled. We split L_S and F into blocks, so they could be expressed respectively as $L_S = \begin{bmatrix} L_{ll} & L_{lu} \\ L_{ul} & L_{uu} \end{bmatrix}$ and $F = [F_l; F_u]$, $F_l = Y_l$. The optimization procedure is just the same as clustering depicted above, the only difference is updating the class indicator matrix F. When λ is a very large number, problem (10) is equivalent to the following problem

$$\begin{aligned} &\min_{S,F} \sum_v w_v \sum_{i,j} ||x_i^v - x_j^v||_2^2 s_{ij} + \alpha||S||_F^2 + 2\lambda Tr(F^T L_S F) \\ &s.t. \ s_i^T \mathbf{1} = 1, 0 \le s_{ij} \le 1, F_l = Y_l \end{aligned} \tag{18}$$

Algorithm 1 Multi-view Learning with Adaptive Neighbours

Input:
$X = \{X_1, X_2, \cdots, X_v\}, X_v \in \mathbb{R}^{n\times d^v}$, number of classes c, parameter λ, label matrix Y_l.

Output:
Clustering: similarity matrix $S \in \mathbb{R}^{n\times n}$ with exact c connected components
Classification: the predicted label matrix $F \in \mathbb{R}^{n\times c}$ for all data points.

Initial the weight for each view, $w_v = \frac{1}{v}$, then each row s_i of S can be initialized by solving the following problem: $\min_{s_i^T\mathbf{1}=1, 0\le s_{ij}\le 1} \sum_{j=1}^{n} (\frac{1}{w_v}\sum_v ||x_i^v - x_j^v||_2^2 s_{ij} + \alpha s_{ij}^2)$.

repeat
Update w_v by using Eq. (9)
Clustering: update F by solving the problem (13)/ Semi-supervised Classification: update the unlabeled fraction of F by Eq. (20)
Update each row of S by solving the problem (17)

until converge
Semi-supervised Classification: Assign the single class label to unlabeled point by Eq. (21).

It could be written as

$$\min_{F\in\mathbb{R}^{n\times c}, F_l = Y_l} Tr(F^T L_S F) \tag{19}$$

According to the (Zhu, Ghahramani, and Lafferty 2003), the optimal solution to problem (19) can be calculated as

$$F_u = -L_{uu}^{-1} L_{ul} Y_l \tag{20}$$

After iteration, the final single class label could be assigned to unlabeled data points by following decision function:

$$\begin{aligned} &y_i = \arg\max_j F_{ij}, \\ &\forall i = l+1, l+2, ..., n. \forall j = 1, 2, ..., c \end{aligned} \tag{21}$$

By iteratively solving problem (10), the final S and F in the objective function Eq. (6) can be obtained and could be used for clustering and classification respectively. The Algorithm is summarized in Alg. 1.

Convergence Analysis

The proposed algorithm can find a local optimal solution, to prove its convergence, we need to utilize the lemma introduce by (Nie et al. 2010)

Lemma 1 For any positive real number u and v, the following inequality holds:

$$\sqrt{u} - \frac{u}{2\sqrt{v}} \le \sqrt{v} - \frac{v}{2\sqrt{v}}. \tag{22}$$

Theorem 2. *In Alg. 1, updated S will decrease the objective value of problem (6) until converge.*

Proof. Suppose the updated S is $\widetilde{S}$ in each iteration, it's easy

to know that:

$$\begin{aligned}&\sum_v \frac{\sum\limits_{i,j}||x_i^v-x_j^v||_2^2\widetilde{s_{ij}}}{2\sqrt{\sum\limits_{i,j}||x_i^v-x_j^v||_2^2 s_{ij}}}+\alpha||\widetilde{S}||_F^2\\ \leq &\sum_v \frac{\sum\limits_{i,j}||x_i^v-x_j^v||_2^2 s_{ij}}{2\sqrt{\sum\limits_{i,j}||x_i^v-x_j^v||_2^2 s_{ij}}}+\alpha||S||_F^2\end{aligned} \tag{23}$$

According to Lemma 1, we have

$$\begin{aligned}&\sum_v \sqrt{\sum_{i,j}||x_i^v-x_j^v||_2^2\widetilde{s_{ij}}}-\sum_v \frac{\sum\limits_{i,j}||x_i^v-x_j^v||_2^2\widetilde{s_{ij}}}{2\sqrt{\sum\limits_{i,j}||x_i^v-x_j^v||_2^2 s_{ij}}}\\ \leq &\sum_v \sqrt{\sum_{i,j}||x_i^v-x_j^v||_2^2 s_{ij}}-\sum_v \frac{\sum\limits_{i,j}||x_i^v-x_j^v||_2^2 s_{ij}}{2\sqrt{\sum\limits_{i,j}||x_i^v-x_j^v||_2^2 s_{ij}}}\end{aligned} \tag{24}$$

Sum over Eq. (23) and Eq. (24) in the two sides, we arrive at:

$$\begin{aligned}&\sum_v \sqrt{\sum_{i,j}||x_i^v-x_j^v||_2^2\widetilde{s_{ij}}}+\alpha||\widetilde{S}||_F^2\\ \leq &\sum_v \sqrt{\sum_{i,j}||x_i^v-x_j^v||_2^2 s_{ij}}+\alpha||S||_F^2\end{aligned} \tag{25}$$

which completes the prove. □

Determine α using Adaptive Neighbours

The value of regularization parameter α could be from zero to infinite, it's difficult to tune in experiment. Let us recall the original intention of introducing parameter α. In problem (1), it determines number of the neighbour to data point x_i: neighbour number will be one if α equal to zero, $n-1$ if α becomes infinite. We assign k nearest neighbours to each point, for any x_i, the Lagrangian Function of problem (17) is:

$$\mathcal{L}(s_i,\phi,\varphi_i)=\frac{1}{2}||s_i+\frac{1}{2\alpha_i}d_i||_2^2-\phi(s_i^T\mathbf{1}-1)-\varphi_i^T s_i \tag{26}$$

where $\phi,\varphi_i \geq 0$ are Lagrangian multipliers and $d_{ij}=\sum\limits_v w_v||x_i^v-x_j^v||_2^2+\lambda||f_i-f_j||_2^2$. According to KKT condition (Lemaréchal 2006), the optimal solution of s_i is:

$$s_{ij}=(-\frac{d_{ij}}{2\alpha_i}+\phi)_+ \tag{27}$$

where $\phi=\frac{1}{k}+\frac{1}{2k\alpha_i}\sum_{j=1}^k d_{ij}$ (Nie, Wang, and Huang 2014). That x_i have k neighbours can be translate into $s_{ij}>0, \forall 1\leqslant j\leqslant k$ and $s_{i,k+1}=0$. According to Eq. 27 and substitution ϕ, we have

$$\frac{k}{2}d_{ik}-\frac{1}{2}\sum_{j=1}^k d_{ij}<\alpha_i\leq\frac{k}{2}d_{i,k+1}-\frac{1}{2}\sum_{j=1}^k d_{ij} \tag{28}$$

where $d_{i1},d_{i2},\cdots,d_{in}$ are sorted in ascending order. Hence, to make most of s_i has exact k non-zeros elements, we let α_i equal to the right item and set the final α be the average of them:

$$\alpha=\frac{1}{n}\sum_{i=1}^n\alpha_i=\frac{1}{n}\sum_{i=1}^n(\frac{k}{2}d_{i,k+1}-\frac{1}{2}\sum_{j=1}^k d_{ij}) \tag{29}$$

Experiment

Since our MLAN is kind of graph-based learning model, we will perform the proposed methods on four benchmark data sets, compared with other related graph based state-of-the-art multi-view clustering and semi-supervised classification methods.

Data Set Descriptions

MSRC-v1 data set contain 240 images in 8 class as a whole. Following (Cai et al. 2011), we select 7 classes composed of tree, building, airplane, cow, face, car, bicycle and each class has 30 images. We extract three visual features from each image: colour moment (CM) with dimension 24, GIST with 512 dimension, CENTRIST feature with 254 dimension, and local binary pattern (LBP) with 256 dimension.

Handwritten numerals (HW) data set is comprised of 2,000 data points for 0 to 9 digit classes, 200 data points for each class. There are Six public features are available: 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.

Caltech101 is an object recognition data set containing 101 categories of images. We follow previous work (Li et al. 2015) and select the widely used 7 classes, i.e. Dolla-Bill, Face, Garfield, Motorbikes, 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.

NUS-WIDE is a real-world web image dataset for object recognition problem. We select the front 25 from the all 31 categories in alphabetical order (bear, bird,...,tower), and choose the first 120 images for each class. Five low-level features are extracted to represent each image: 64 color histogram, 144 color correlogram, 73 edge direction histogram, 128 wavelet texture, and 225 block-wise color moment.

Experiment Setup

The classic single-view approach: Spectral Clustering (SC) (Ng, Jordan, and Weiss 2001) and Label Propagation (LP) (Zhu, Ghahramani, and Lafferty 2003) are conducted on each of view as the baseline. Then we compare the proposed methods with six other state-of-art clustering and semi-supervised classification approaches: (a) Co-trained spectral clustering (Co-train) (Kumar and III 2011), (b) Co-regularized Spectral Clustering (Co-reg) (Kumar, Rai, and III 2011), (c) Multi-view Spectral Clustering (MVSC) (Li et al. 2015), (d) Multi-Modal Spectral Clustering (MMSC)(Cai et al. 2011), (e) Adaptive Multi-Model Semi-Supervised classification (AMMSS) (Cai et al. 2013), (f) Sparse Multiple Graph Integration (SMGI) (Karasuyama and Mamitsuka 2013).

There is only one parameter λ in our model, brought by the Laplacian matrix rank constrain. Considering of simpleness and accelerating the convergence procedure, we can initialize λ to a random positive value, 1 to 30 in our experiment, and decrease it ($\lambda=\lambda/4$) if the connected components of S

is greater than class number c or increase it ($\lambda = \lambda * 4$) if smaller than c in each iteration. For other compared methods, we set their parameters to the optimal value if they have. To all dataset, each sample is assigned 9 nearest neighbours to construct graph. In terms of semi-supervised classification, we choose the front 20% data for as labeled sample to mimic the real situation ($l \ll u$). Since the compared clustering methods are spectral related, the performance varies because the required k-means procedure is dependent on the choose of initial centroids, so we perform 50 times experiments for all methods on each dataset.

Table 1: Clustering result in terms of accuracy (mean and standard deviation).

Data set	MSRC-v1	Caltech101	HW
SC(1)	0.364(0.009)	0.346(0.025)	0.726(0.058)
SC(2)	0.542(0.046)	0.448(0.041)	0.658(0.051)
SC(3)	0.566(0.046)	0.529(0.049)	0.690(0.056)
SC(4)	0.575(0.045)	0.607(0.053)	0.670(0.042)
SC(5)		0.672(0.040)	0.715(0.056)
SC(6)		0.591(0.029)	0.218(0.024)
Co-train	0.634(0.014)	0.620(0.004)	0.824(0.010)
Co-reg	0.724(0.049)	0.657(0.032)	0.889(0.070)
MVSC	0.623(0.047)	0.725(0.046)	0.756(0.074)
MMSC	0.688(0.028)	0.745(0.012)	0.934(0.016)
MLAN	**0.738(0.000)**	**0.780(0.000)**	**0.973(0.000)**

Table 2: Clustering result in terms of NMI (mean and standard deviation).

Data set	MSRC-v1	Caltech101	HW
SC(1)	0.292(0.023)	0.142(0.011)	0.795(0.030)
SC(2)	0.499(0.040)	0.279(0.012)	0.700(0.029)
SC(3)	0.477(0.029)	0.338(0.027)	0.727(0.034)
SC(4)	0.487(0.027)	0.492(0.054)	0.681(0.023)
SC(5)		0.507(0.051)	0.786(0.025)
SC(6)		0.451(0.047)	0.143(0.022)
Co-train	0.553(0.011)	0.561(0.003)	0.798(0.004)
Co-reg	0.653(0.038)	0.549(0.017)	0.814(0.043)
MVSC	0.553(0.037)	0.586(0.067)	0.830(0.051)
MMSC	0.612(0.024)	0.605(0.014)	0.893(0.010)
MLAN	**0.734(0.004)**	**0.630(0.000)**	**0.939(0.001)**

Performance Evaluation

For clustering results, three evaluation metric are adopted, namely, accuracy, normalized mutual information (NMI), and purity; for semi-supervised classification, the evaluation metric is accuracy (the proportion of the correct-classified data points in all unlabeled data).

Table 1, Table 2, and Table 3 show the clustering accuracy, NMI, and Purity respectively, and Table 4 show the accuracy of semi-supervised classification. Generally, almost all multi-view clustering methods achieve superior result than the best of single-view approaches. We can see that the proposed method MLAN outperform other state-of-the-art methods in almost all experiments. In addition, MLAN is very robust

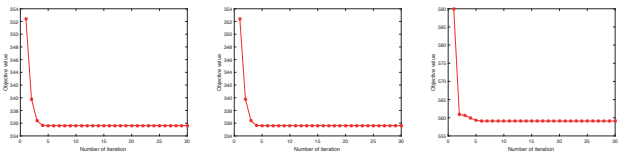
Table 3: Clustering result in terms of purity (mean and standard deviation).

Data set	MSRC-v1	Caltech101	HW
SC(1)	0.418(0.018)	0.622(0.021)	0.832(0.045)
SC(2)	0.624(0.040)	0.772(0.013)	0.702(0.038)
SC(3)	0.577(0.034)	0.792(0.024)	0.736(0.042)
SC(4)	0.617(0.035)	0.747(0.037)	0.708(0.033)
SC(5)		0.765(0.043)	0.823(0.044)
SC(6)		0.725(0.046)	0.237(0.026)
Co-train	0.652(0.012)	0.810(0.002)	0.846(0.007)
Co-reg	0.758(0.038)	0.792(0.010)	0.866(0.055)
MVSC	0.662(0.041)	0.826(0.062)	0.884(0.062)
MMSC	0.718(0.026)	0.876(0.001)	0.934(0.010)
MLAN	**0.805(0.000)**	**0.889(0.000)**	**0.973(0.000)**

Table 4: Semi-supervised classification performance (mean accuracy and deviation).

Data set	MSRC-v1	NUS-WIDE	HW
LP(1)	0.369(0.000)	0.172(0.000)	0.925(0.000)
LP(2)	0.304(0.000)	0.121(0.000)	0.804(0.000)
LP(3)	0.244(0.000)	0.104(0.000)	0.763(0.000)
LP(4)	0.226(0.000)	0.097(0.000)	0.693(0.000)
LP(5)		0.085(0.000)	0.691(0.000)
LP(6)			0.473(0.000)
SMGI	0.799(0.054)	0.226(0.030)	0.973(0.004)
AMMSS	0.821(0.000)	0.237(0.000)	0.976(0.000)
MLAN	**0.863(0.000)**	**0.238(0.000)**	**0.977(0.000)**

to the parameter λ, it nearly can be seen parameter-free approach. By contrast, the parameter-free method Co-train is robust to some extent, but the performance is not satisfactory. In MSRC-v1 dataset, Co-reg method beat MLAN in terms of accuracy, however its' result is not steady in duplicate test, with average 5% deviation in all results. Thus the proposed MLAN approach notably alleviates the challenging problem that parameter-free methods cannot achieve better result than those with one or two to be searched. The convergence curves of the objective value are shown in Figure 1.



(a) MSRC-v1 (b) Caltech101 (c) HW

Figure 1: Convergence speed of MLAN in terms of clustering task.

Conclusions

In this paper, we introduce a novel multi-view learning model named MLAN, which performs clustering/semi-supervised classification and local structure learning simultaneously.

With the reasonable rank constrain, the obtained optimal graph can be partitioned into specific clusters directly. Due to the robustness to the only parameter, MLAN nearly can be seen as parameter-free method, which is very commendable, especially for unsupervised clustering task. Extensive experimental results show that the proposed model achieve superiors performances. The future work can be the extension of the data fusion form, using cube root or other elementary functions.

References

Blum, A., and Mitchell, T. M. 1998. Combining labeled and unlabeled sata with co-training. In *Proceedings of the Eleventh Annual Conference on Computational Learning Theory, COLT 1998, Madison, Wisconsin, USA, July 24-26, 1998.*, 92–100.

Cai, X.; Nie, F.; Huang, H.; and Kamangar, F. 2011. Heterogeneous image feature integration via multi-modal spectral clustering. In *The 24th IEEE Conference on Computer Vision and Pattern Recognition*, 1977–1984.

Cai, X.; Nie, F.; Cai, W.; and Huang, H. 2013. Heterogeneous image features integration via multi-modal semi-supervised learning model. In *IEEE International Conference on Computer Vision*, 1737–1744.

Chaudhuri, K.; Kakade, S. M.; Livescu, K.; and Sridharan, K. 2009. Multi-view clustering via canonical correlation analysis. In *Proceedings of the 26th Annual International Conference on Machine Learning*, 129–136.

Chung, F. R. K. 1997. Spectral graph theory. In *CBMS Regional Conference Series in Mathematics*, number 92.

Dalal, N.; Triggs, B.; and Schmid, C. 2006. Human detection using oriented histograms of flow and appearance. In *The 9th European Conference on Computer Vision, Proceedings, Part II*, 428–441.

Fan, K. 1949. On a theorem of weyl concerning eigenvalues of linear transformations. In *Glass and Ceramics*, volume 35, 652–655.

Huang, J.; Nie, F.; and Huang, H. 2013. Spectral rotation versus k-means in spectral clustering. In *Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence, July 14-18, 2013, Bellevue, Washington, USA*.

Karasuyama, M., and Mamitsuka, H. 2013. Multiple graph label propagation by sparse integration. *IEEE Transactions on Neural Networks and Learning Systems* 24(12):1999–2012.

Kumar, A., and III, H. D. 2011. A co-training approach for multi-view spectral clustering. In *Proceedings of the 28th International Conference on Machine Learning*, 393–400.

Kumar, A.; Rai, P.; and III, H. D. 2011. Co-regularized multi-view spectral clustering. In *The 25th Annual Conference on Neural Information Processing Systems*, 1413–1421.

Lemaréchal, C. 2006. S. boyd, l. vandenberghe, convex optimization, cambridge university press, 2004 hardback, ISBN 0 521 83378 7. *European Journal of Operational Research* 170(1):326–327.

Li, Y.; Nie, F.; Huang, H.; and Huang, J. 2015. Large-scale multi-view spectral clustering via bipartite graph. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2750–2756.

Liu, Y.; Nie, F.; Wu, J.; and Chen, L. 2013. Efficient semi-supervised feature selection with noise insensitive trace ratio criterion. *Neurocomputing* 105:12–18.

Lowe, D. G. 2004. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision* 60(2):91–110.

Mohar, B. 1991. The laplacian spectrum of graphs. In *Graph Theory, Combinatorics, and Applications*, 871–898.

Ng, A. Y.; Jordan, M. I.; and Weiss, Y. 2001. On spectral clustering: Analysis and an algorithm. In *Advances in Neural Information Processing Systems*, 849–856.

Nie, F.; Huang, H.; Cai, X.; and Ding, C. H. Q. 2010. Efficient and robust feature selection via joint 2, 1-norms minimization. In *Advances in Neural Information Processing Systems*, 1813–1821.

Nie, F.; Li, J.; and Li, X. 2016. Parameter-free auto-weighted multiple graph learning: A framework for multiview clustering and semi-supervised classification. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI 2016,*, 1881–1887.

Nie, F.; Wang, X.; and Huang, H. 2014. Clustering and projected clustering with adaptive neighbors. In *The 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 977–986.

Ojala, T.; Pietikäinen, M.; and Mäenpää, T. 2002. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 24(7):971–987.

Oliva, A., and Torralba, A. 2001. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision* 42(3):145–175.

Selee, T. M.; Kolda, T. G.; Kegelmeyer, W. P.; and Griffin, J. D. 2007. Extracting clusters from large datasets with multiple similarity measures using imscand. *Technical Report, Sandia Natl Laboratories* (SAND2007-7977):87–103.

Tarjan, R. E. 1972. Depth-first search and linear graph algorithms. *SIAM J. Comput.* 1(2):146–160.

Tian, Z., and Kuang, R. 2010. Integrative classification and analysis of multiple arraycgh datasets with probe alignment. *Bioinformatics* 26(18):2313–2320.

Zhu, X.; Ghahramani, Z.; and Lafferty, J. D. 2003. Semi-supervised learning using gaussian fields and harmonic functions. In *Machine Learning, Proceedings of the Twentieth International Conference (ICML 2003), August 21-24, 2003, Washington, DC, USA*, 912–919.