Latent Multi-view Semi-Supervised Classification

Xiaofan Bo[*]
Glasgow college, University of Electronic Science and Technology of China

Zhao Kang[*] ZKANG@UESTC.EDU.CN
School of Computer Science and Engineering, University of Electronic Science and Technology of China

Zhitong Zhao
School of Computer Science and Engineering, University of Electronic Science and Technology of China

Yuanzhang Su[†] SYZ@UESTC.EDU.CN
School of Foreign Languages, University of Electronic Science and Technology of China

Wenyu Chen[†] CWY@UESTC.EDU.CN
School of Computer Science and Engineering, University of Electronic Science and Technology of China

Editors: Wee Sun Lee and Taiji Suzuki

Abstract

To explore underlying complementary information from multiple views, in this paper, we propose a novel Latent Multi-view Semi-Supervised Classification (LMSSC) method. Unlike most existing multi-view semi-supervised classification methods that learn the graph using original features, our method seeks an underlying latent representation and performs graph learning and label propagation based on the learned latent representation. With the complementarity of multiple views, the latent representation could depict the data more comprehensively than every single view individually, accordingly making the graph more accurate and robust as well. Finally, LMSSC integrates latent representation learning, graph construction, and label propagation into a unified framework, which makes each subtask optimized. Experimental results on real-world benchmark datasets validate the effectiveness of our proposed method.

Keywords: Semi-supervised classification; Multi-view learning; Latent space

1. Introduction

Thanks to its ability to take advantage of abundant unlabeled data, semi-supervised classification has been widely used for numerous problems Chapelle et al. (2009). A great number of semi-supervised classification methods have been developed in the past decades Zhu and Goldberg (2009). Among them, the graph-based semi-supervised classification technique has achieved state-of-the-art performance. Given a dataset with a limited number of initial labels, it labels the unlabeled ones according to the propagation of pairwise similarity. In particular, it consists of two steps. First, a graph is constructed based on certain similarity

. *Both authors contributed equally to this work
. †Corresponding authors

metrics. Each data point is represented by a node on the graph and the weight denotes the similarity between two points. Second, the labels of unlabeled points are inferred by label propagation. Therefore, numerous methods focus on either building graphs Jebara et al. (2009); Cheng et al. (2009); Zhuang et al. (2012); Li and Fu (2015); Li et al. (2017), or designing effective label propagation schemes Zhou et al. (2004); Wang and Zhang (2008); Zhu et al. (2003).

Although existing techniques have achieved promising performance in various real-world applications, they are limited in three aspects. First, they mainly deal with single-view data. Nowadays, data is often represented by multiple views brought by various sensors and feature descriptors Tao et al. (2017); Kang et al. (2019a). Moreover, leveraging multi-view data can boost the performance of single-view methods, due to the complementarity of heterogeneous information. Many machine learning, data mining, pattern recognition, and computer vision tasks have benefited from multi-view data Xu et al. (2013); Zhao et al. (2017); Fu et al. (2015). As a result, it is paramount to develop multi-view semi-supervised classification techniques. Though some methods have been developed to tackle multi-view data, they suffer other drawbacks.

Second, most graphs are constructed from the original data. Real-world data is often contaminated by various noise or outliers. Thus, the resulted graph may not be accurate to reflect the underlying relationships between data samples Kang et al. (2019d,e). As a matter of fact, many researchers have shown that graph quality is crucial to the performance of subsequent tasks Nie et al. (2017); Kang et al. (2017, 2019b, 2018b). Therefore, how to construct a robust and reliable graph is vital. For multi-view data, this is more challenging due to its heterogeneity nature.

Third, existing methods usually take graph construction and label propagation as two separate steps. Consequently, they are not jointly optimized. In particular, the resulted graph might not be optimal for a subsequent task. Furthermore, it is desired to exploit the partial label information to guide the graph construction process. Existing methods often fail to make use of this information.

Confronted with the aforementioned limitations, we propose a novel multi-view learning method, named as Latent Multi-view Semi-Supervised Classification (LMSSC). It is composed of three components. The first component extracts view-specific and shared latent factors from all the views. In specific, the shared latent factor can be treated as a view-independent data representation. Based on it, we can construct a common graph for all views. This is reasonable since points in different views indeed represent the same set of objects. Unlike traditional fixed graph, we learn it from data and update it iteratively according to the result of classification. After the graph is obtained, the label propagation function is updated accordingly. In this way, the latent factor, the graph, and the label propagation are jointly optimized.

The main merits of this paper can be summarized as follows:

- We propose a novel multi-view semi-supervised classification method, LMSSC. It integrates common representation learning, graph construction, and label prediction into a unified framework.

- The view consistency is ensured by extracting shared latent factor from multi-view data. A robust graph is built on the latent factor with structure guarantee.

- We conduct extensive experiments on benchmark datasets. Compared with the representative single-view and multi-view semi-supervised classification methods, our proposed method demonstrates its superiority.

Notations For a matrix M, we express its i-th row and j-th column element as M_{ij}. The Frobenius norm of matrix M is defined as the square root of summation of the square of every single element, i.e., $\|M\|_F = \sqrt{\sum_{ij} M_{ij}^2}$, while the ℓ_2-norm of vector m is denoted by $\|m\|_2 = \sqrt{m^T \cdot m}$. The trace operator is expressed as $Tr(\cdot)$. $\mathbf{1}$ represents a column vector whose elements are all ones.

2. Related Work

Recently, graph-based semi-supervised classification has attracted a lot of attention. For example, Zhu et al. Zhu et al. (2003) designed a semi-supervised classification algorithm based on Gaussian Field and Harmonic function (GFHF). It takes advantage of the harmonic property of Gaussian random field over the graph. Though it has gained huge popularity, its performance heavily depends on the input graph. Later, Nie et al. Nie et al. (2011) design a semi-supervised classification method by minimizing the ℓ_1-norm of spectral embedding; Sparse He et al. (2011); Yan and Wang (2009) and low-rank graphs Zhuang et al. (2012); Kang et al. (2018a, 2019c) have also been proposed. These methods all focus on single-view data and cannot make full use of the multi-view information.

Some semi-supervised classification techniques have been extended to the multi-view setting. Nie et al. propose Auto-weighted Multiple Graph Learning (AMGL) Nie et al. (2016). For multi-view data set $X = \{X^v\}, v = 1, 2, 3, \cdots, V$, where V represents the number of views. Each view contains N samples denoted as $\{x_1^v, x_2^v, ..., x_N^v\} \in \mathcal{R}^{d_{(v)} \times N}$, where $d_{(v)}$ denotes the feature number of the v-th view. Given graph matrix $S = \{s_{ij}\} \in \mathcal{R}^{N \times N}$, the corresponding degree matrix D ($d_{ii} = \sum_{j=1}^{N} s_{ij}$) is available. Then, $L = D - S$ is the so-called Laplacian matrix. Without loss of generality, all the points are rearranged and the front $l (l < N)$ points are labeled. AMGL solves

$$\min_F \sum_{v=1}^{V} w^v Tr(F^T L^v F) \quad s.t. \quad f_i = y_i, \quad \forall i = 1, 2, \cdots, l, \tag{1}$$

where the v-th view weight $w^v = 1 \Big/ \left(2\sqrt{Tr(F^T L^{(v)} F)}\right)$ is updated iteratively, $F = [f_1, \cdots, f_n]^T \in \mathcal{R}^{N \times c}$ is the class indicator matrix for c classes and y_i is the given indicator vector for the i-th point. $y_{ij} = 1$ only if the i-th point belongs to the j-th class, otherwise $y_{ij} = 0$. This approach suffers the graph construction issue.

To address the above problem, Multi-view Learning with Adaptive Neighbours (MLAN) Nie et al. (2017) is further developed. It learns the graph based on adaptive neighbours. Basically, it solves the following problem

$$\min_S \sum_{v=1}^{V} w^v \sum_{ij} \|x_i^v - x_j^v\|_2^2 s_{ij} + \alpha \|S\|_F^2 \quad s.t. \quad s_i^T \mathbf{1} = 1, \quad 0 \le s_{ij} \le 1, \tag{2}$$

where $\alpha > 0$ is a trade-off parameter and view weight is updated according to $w^v = 1\Big/\left(2\sqrt{\sum_{ij}\|x_i^v - x_j^v\|_2^2 s_{ij}}\right)$. The distance between x_i and x_j and the similarity are negatively correlated since adjacent samples have larger similarity. In traditional graph construction approaches, the neighbours are pre-determined and the similarity is fixed. Differing from this, MLAN assigns adaptive neighbours to each data point. In other words, the k nearest neighbours of any x_i are not steady and they change in every iteration. As a result, this strategy always shows better performance than previous heuristic approaches. Despite its promising performance, the graph is built on the raw data which can be easily contaminated by noise or outliers. In addition, the weight assignment is also quite arbitrary, which could lead to an unsatisfied solution.

3. Proposed Methodology

As demonstrated above, there exist two key problems that should be solved for multi-view semi-supervised classification. First, how to construct a robust graph based on multi-view data? Second, how to combine the graph construction with label propagation process? To address these problems, we propose a latent multi-view semi-supervised classification (LMSSC) method.

3.1. Formulation

The integrated formulation is as follows:

$$\min_{W,H,S,F} \Phi(X,W,H) + \beta\Omega(H,L,S) + \gamma\Theta(L,Y,F), \tag{3}$$

where $W = \left\{W^v \in \mathcal{R}^{d_{(v)}\times r}, v = 1,2,...,V\right\}$ and $H \in \mathcal{R}^{r\times N}$ are latent factors extracted from multi-view data X and r is the dimension of latent representation. S is the similarity graph shared by all views. L is the Laplacian matrix of S and Y is a prior label indicator matrix. F is the label indicator matrix that we aim to predict. β and γ are trade-off parameters. Next, we will discuss each term in problem (3) in detail.

3.2. Factors in Latent Space

In order to build a robust graph, we learn a shared latent factor from all views. Since points in different views indeed represent the same set of objects, the shared latent factor is considered to be view-independent features.

Specifically, we decompose the samples in v-th view as $X^v = W^v H$, where W^v can be treated as view-specific factor and H is the latent representation shared by all views. Consequently, the first term in problem (3) can be formulated as

$$\Phi(X,W,H) = \sum_{v=1}^{V}\|X^v - W^v H\|_F^2 \quad s.t. \quad W^v \geq 0. \tag{4}$$

Each h_i denotes a new representation of object i.

3.3. Graph Construction

With the shared latent representation H, we can build a graph on it. Hence, this graph is also shared by all views. Since adaptive neighbours strategy can capture the local manifold structure of data Nie et al. (2017), we utilize this approach to construct similarity graph S. Then, the second term in problem (3) is formulated as follows:

$$\Omega(H,L,S)=\frac{1}{2}\sum_{i,j=1}^{N}\|h_i-h_j\|_2^2 s_{ij}+\alpha\|S\|_F^2 \quad s.t. \quad s_i^T\mathbf{1}=1, \quad 0\le s_{ij}\le 1. \tag{5}$$

Moreover, we have the following equality

$$\frac{1}{2}\sum_{i,j=1}^{N}\|h_i-h_j\|_2^2 s_{ij}=Tr(HLH^T). \tag{6}$$

So formulation (5) can be simplified as

$$\Omega(H,L,S)=Tr(HLH^T)+\alpha\|S\|_F^2 \quad s.t. \quad s_i^T\mathbf{1}=1, \quad 0\le s_{ij}\le 1. \tag{7}$$

Though this graph is built on latent space, there is no guarantee that it would be optimal for subsequent classification. Ideally, graph S should have exact c connected components, i.e., the data points are already identified as c classes. However, the current solution can hardly satisfy such a condition. To tackle this problem, we can resort to the following theorem Mohar et al. (1991):

Theorem 1 *The number of connected components c of the graph S is equal to the multiplicity of zero eigenvalue of its Laplacian matrix L.*

We denote $\sigma_i(L)$ as the i-th smallest eigenvalue of L. Since L is a positive semidefinite matrix, its eigenvalues $\sigma_i(L)\ge 0$. Theorem 1 indicates that if $\sum_{i=1}^{c}\sigma_i=0$, then our requirement can be met. Hence, we can achieve a desired graph by minimizing $\sum_{i=1}^{c}\sigma_i$. Taking this into account, we can formulate Eq. (7) as

$$\Omega(H,L,S)=Tr(HLH^T)+\alpha\|S\|_F^2+\gamma\sum_{i=1}^{c}\sigma_i \quad s.t. \quad s_i^T\mathbf{1}=1, \quad 0\le s_{ij}\le 1. \tag{8}$$

3.4. Label Propagation

In fact, Eq. (8) is hard to handle due to the involvement of graph structure term. Fortunately, Ky Fan's theorem Fan (1949) gives the following equality:

$$\sum_{i=1}^{c}\sigma_i=\min_{F,F^TF=I}Tr(F^TLF), \tag{9}$$

For semi-supervised learning, F can be decomposed as $F=[F_l;F_u]=[Y_l;F_u]$, where $Y_l=[y_1,y_2,...,y_l]^T$ represents the known label matrix and $l(u)$ is the number of labeled(unlabeled) data points. With these notations, the right part of above equation becomes the objective function of semi-supervised classification Nie et al. (2017, 2016). Therefore, the requirement for the graph structure and label propagation are the same in essence. Then, the third term in problem (3) can be expressed as:

$$\Theta(L,Y,F) = Tr(F^T LF) \quad s.t. \quad F_l = Y_l. \tag{10}$$

3.5. Unified Objective Function

Based on Eqs. (4), (7), and (10), our objective function (3) can be explicitly written as

$$\begin{aligned} &\min_{W,H,S,F} \sum_{v=1}^{V} \|X^v - W^v H\|_F^2 + \beta(Tr(HLH^T) + \alpha \|S\|_F^2) + \gamma Tr(F^T LF) \\ &s.t. \quad W^v \geq 0, s_i^T \mathbf{1} = 1, 0 \leq s_{ij} \leq 1, F_l = Y_l. \end{aligned} \tag{11}$$

We can observe that Eq. (11) integrates latent representation learning, graph construction, and label prediction into a unified framework. Joint optimization of H, S, and F will facilitate an overall optimal solution. Furthermore, graph S is built in latent space, thus it is robust to noise and outliers in general.

4. Optimization

Eq. (11) is not convex with respect to all variables. Thus we solve problem (11) based on an alternating strategy, i.e., we solve one variable while considering other variables stationary.
A. *Update View-Specific Latent Factor W^v*
When H, S, and F are fixed, problem (11) turns into

$$\min_{W^v} \sum_{v=1}^{V} \|X^v - W^v H\|_F^2 \quad s.t. \quad W^v \geq 0. \tag{12}$$

This can be solved column-wisely, i.e.,

$$\min_{W_{i,:}^v} \|X_{i,:}^v - W_{i,:}^v H\|_2^2 \quad s.t. \quad W_{i,:}^v \geq 0. \tag{13}$$

It is a quadratic programing problem which can be easily solved by many existing packages.
B. *Update Shared Latent Factor H*
We set variables other than H fixed, then we have the following subproblem

$$\min_{H} \sum_{v=1}^{V} \|X^v - W^v H\|_F^2 + \beta Tr(HLH^T). \tag{14}$$

By setting the derivative w.r.t. H to zero, we obtain

$$\sum_{v=1}^{V} (W^v)^T W^v H + \beta HL = \sum_{v=1}^{V} (W^v)^T X^v. \tag{15}$$

The above equation is the so-called Sylvester equation and can be easily solved by the Bartels-Stewart algorithm Bartels and Stewart (1972). It has a unique solution since $\sum_{v=1}^{V}(W^v)^T W^v$ and $-\beta L$ have no common eigenvalues Zhang et al. (2017).

C. *Update Similarity Graph S*

After ignoring non-relevant variables, we get

$$\min_{S} \beta(Tr(HLH^T) + \alpha \|S\|_F^2) + \gamma Tr(F^T LF) \quad s.t. \quad s_i^T \mathbf{1} = 1, 0 \leq s_{ij} \leq 1. \tag{16}$$

Remember that L is a function of S and $Tr(HLH^T) = \frac{1}{2}\sum_{i,j=1}^{N} \|h_i - h_j\|_2^2 s_{ij}$. Let $d_{ij}^h = \|h_i - h_j\|_2^2$ and $d_{ij}^f = \|f_i - f_j\|_2^2$, then we can reformulate problem (16) column-wisely as

$$\min_{s_i} \sum_{j=1}^{N} \beta(\frac{1}{2} d_{ij}^h s_{ij} + \alpha s_{ij}^2) + \frac{1}{2}\gamma d_{ij}^f s_{ij} \quad s.t. \quad s_i^T \mathbf{1} = 1, 0 \leq s_{ij} \leq 1. \tag{17}$$

Denote $d_i \in R^{N\times 1}$ a column vector with $d_{ij} = \beta d_{ij}^h + \gamma d_{ij}^f$. Then above problem can be simplified as

$$\min_{s_i} \left\| s_i + \frac{1}{4\alpha\beta} d_i \right\|_2^2 \quad s.t. \quad s_i^T \mathbf{1} = 1, \quad 0 \leq s_{ij} \leq 1. \tag{18}$$

We will show its closed-form solution in the next section.

D. *Update Label Indicator F*

For F, the remaining terms are

$$\min_{F} Tr(F^T LF) \quad s.t. \quad F_l = Y_l. \tag{19}$$

We first split L into blocks as $L = \begin{bmatrix} L_{ll} & L_{lu} \\ L_{ul} & L_{uu} \end{bmatrix}$. We take the derivative of Eq. (19) in terms of F and set its first-order derivative to zero, that is,

$$\begin{bmatrix} L_{ll} & L_{lu} \\ L_{ul} & L_{uu} \end{bmatrix} \begin{bmatrix} Y_l \\ F_u \end{bmatrix} = 0. \tag{20}$$

We can see that

$$F_u = -L_{uu}^{-1} L_{ul} Y_l. \tag{21}$$

We iteratively update above variables until maximum 100 times are reached or the relative change of F is less than 10^{-5}. The complete procedures are shown in Algorithm 1.

Finally, the labels for unlabeled points can be assigned according to the following decision function:

$$y_i = \arg\max_{j} F_{ij}, \quad \forall i = l+1, l+2, \cdots, N. \quad \forall j = 1, 2, \cdots, c. \tag{22}$$

Algorithm 1 The algorithm of LMSSC

```
Input: Data matrices: $X^{(1)}, \cdots, X^{(V)}$, label matrix $Y_l$, parameters $\beta > 0$, $\gamma > 0$.
Initialize: Random matrix $H$ and $S$, $F_u = 0$.
REPEAT
    Calculate $W$ by (13).
    Update $H$ by solving (15).
    Calculate $S$ by solving (18)
    Update $F$ using (21).
UNTIL stopping criterion is met.
```

4.1. Determine Value of α

In fact, α is closely related to the number of neighbours when we construct the graph. It would be easier to set the nearest neighbours number k than tuning α. Suppose we assign k different neighbors to each data point, we can set γ to be the average of $\{\alpha_i\}_{i=1}^N$. For any x_i , the Lagrangian function of problem (18) can be written as:

$$\mathcal{L}\left(s_i, \phi, \varphi_i\right) = \frac{1}{2}\left\|s_i + \frac{1}{4\alpha_i\beta}d_i\right\|_2^2 - \phi(s_i^T\mathbf{1} - 1) - \varphi_i^T s_i, \tag{23}$$

where ϕ, $\varphi_i \geq 0$ are Lagrangian multipliers. Applying KKT condition(Lemaréchal 2006), we get the optimal solution of s_i:

$$s_{ij} = (-\frac{d_{ij}}{4\alpha_i\beta} + \phi)_+. \tag{24}$$

Considering the constraint $s_i^T\mathbf{1} = 1$, we get

$$\sum_{j=1}^{k}(-\frac{d_{ij}}{4\alpha_i\beta} + \phi) = 1 \Rightarrow \phi = \frac{1}{k} + \frac{1}{4k\alpha_i\beta}\sum_{j=1}^{k} d_{ij}. \tag{25}$$

If the optimal s_i has only k nonzero elements, then $s_{i,k} > 0$ and $s_{i,k+1} = 0$. That is,

$$\begin{cases} -\frac{d_{i,k}}{4\alpha_i\beta} + \phi > 0, \\ -\frac{d_{i,k+1}}{4\alpha_i\beta} + \phi \leq 0. \end{cases} \tag{26}$$

Combining (25) and (26), we have the following inequality for α_i:

$$\frac{1}{2\beta}(\frac{k}{2}d_{ik} - \frac{1}{2}\sum_{j=1}^{k} d_{ij}) < \alpha_i \leq \frac{1}{2\beta}(\frac{k}{2}d_{i,k+1} - \frac{1}{2}\sum_{j=1}^{k} d_{ij}), \tag{27}$$

where $d_{i1}, d_{i2}, \cdots, d_{iN}$ are sorted in ascending order. We set α_i to its maximum and we set the overall α as the mean of α_i. Then, the value of α becomes

$$\alpha = \frac{1}{2N\beta}\sum_{i=1}^{N}(\frac{k}{2}d_{i,k+1} - \frac{1}{2}\sum_{j=1}^{k} d_{ij}). \tag{28}$$

With this equation, we can tune the value of k instead of α since k is an integer and has a small range, which is trivial to find an appropriate value.

Table 1: Statistics of datasets used in experiments

Dataset	Dimension of Views	# of Instances	# of Class
BBC	4659/4633/4665/4684	145	2
Sonar	20/20/20	208	2
HW	216/76/64/6/240/47	2000	10
Reuters	2000/2000/ 2000/ 2000/ 2000	1200	6

5. Experiments

5.1. Datasets

We assess our approach on 4 publicly available datasets. We summarize the information of these datasets in Table 1.

BBC dataset is derived from the BBC sport website corresponding to sports news articles. Each document is split into four segments by separating the document into paragraphs. Each segment is at least 200 characters long and is logically associated with the original document from which it was obtained.

Sonar is a dataset with 208 observations on 60 variables. The features represent the energy within a particular frequency band, integrated over a certain period of time. There are two classes 0 if the object is a rock, and 1 if the object is a mine (metal cylinder). The 60 variables are divided into three parts, each representing a view of 20 variables.

Handwritten numerals (HW) dataset consists of 2000 data points evenly distributed in 0 to 9 digit classes with 200 data points in each class. We choose six types of features of all data points: profile correlations, Fourier coefficients of the character shapes, Karhunen-Love coefficients, morphological features, pixel averages in 2×3 windows, and Zernike moments.

Reuters is a textual dataset which is written in five different languages (English, French, German, Spanish, and Italian). All the documents are categorized into 6 classes. There are altogether five views and each of them represents documents written in one certain language. From each view, we randomly sample 1200 documents in a balanced manner, with each of the 6 classes having 200 documents.

5.2. Comparison Methods

As a baseline, we apply the classic single-view semi-supervised classification method GFHF Zhu et al. (2003) to each view of the datasets. Moreover, we compare our method with two other popular multi-view semi-supervised classification methods: (a) Multi-view Learning with Adaptive Neighbours (MLAN) Nie et al. (2017), (b) Auto-weighted Multiple Graph Learning (AMGL) Nie et al. (2016). For each dataset, we randomly choose 10%, 20%, 30%, and 50% of points as labeled. This procedure is repeated 20 times. Finally, we report the

Table 2: Semi-supervised classification accuracy (%) on BBC

Dataset	BBC			
rate	0.1	0.2	0.3	0.5
AMGL	49.50(2.68)	50.68(4.10)	49.22(3.97)	52.32(4.74)
MLAN	52.02(6.45)	55.73(3.43)	56.12(5.68)	57.40(5.28)
HFPF(1)	92.77(0.88)	92.24(0.97)	93.12(1.45)	92.64(1.92)
HFPF(2)	92.38(0.66)	92.76(1.10)	92.43(1.34)	92.15(2.71)
HFPF(3)	92.42(0.87)	92.72(1.49)	92.23(1.30)	93.47(2.78)
HFPF(4)	92.54(0.75)	92.67(1.17)	92.18(1.01)	92.00(1.85)
LMSSC	**92.86**(**1.22**)	**93.59**(**1.85**)	**94.42**(**1.57**)	**95.14**(**3.00**)

Table 3: Semi-supervised classification accuracy (%) on Handwritten

Dataset	Handwritten			
rate	0.1	0.2	0.3	0.5
AMGL	92.80(0.54)	93.85(0.58)	94.17(0.55)	94.89(0.62)
MLAN	**97.83**(**0.23**)	**97.81**(**0.44**)	**98.00**(**0.38**)	98.10(0.41)
HFPF(1)	88.94(0.78)	91.28(0.77)	92.52(0.48)	93.12(0.44)
HFPF(2)	80.06(0.85)	81.90(0.99)	82.98(0.86)	83.39(1.05)
HFPF(3)	94.21(0.59)	95.79(0.38)	96.26(0.48)	96.59(0.47)
HFPF(4)	42.85(1.23)	44.88(0.90)	45.92(0.95)	47.62(1.30)
HFPF(5)	94.96(0.54)	96.60(0.44)	96.75(0.33)	97.25(0.44)
HFPF(6)	79.72(0.77)	82.03(0.62)	82.52(0.85)	82.93(0.88)
LMSSC	96.59(0.46)	97.10(0.38)	97.59(0.37)	**98.70**(**0.33**)

average classification accuracy and deviation. For our proposed LMSSC[1], all data points are assumed to have 15 nearest neighbours, i.e., we use $k = 15$ to calculate α.

5.3. Results

All results are shown in Tables 2-5. As expected, the accuracy of each method monotonously increases with the increase of label rate. We can see that the performance of HFPF heavily depends on the datasets and the specific view. On the other hand, our LMSSC consistently outputs high accuracy. This validates the advantage of multi-view learning. On BBC, Reuters, and Sonar datasets, our method outperforms AMGL and MLAN by a large margin. For HW dataset, our accuracy is comparable to MLAN method and outperforms others. Furthermore, we can see that AMGL and MLAN perform pretty well on Handwritten and

[1] The code for our implementation is available: https://github.com/sckangz/LMVL

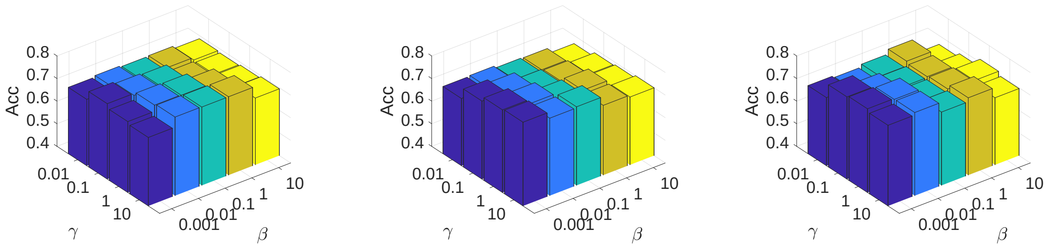
Table 4: Semi-supervised classification accuracy (%) on Reuters

Dataset	Reuters			
rate	0.1	0.2	0.3	0.5
AMGL	35.44(4.95)	39.34(5.00)	42.70(3.48)	46.97(2.04)
MLAN	57.66(9.45)	63.83(5.83)	66.12(3.69)	70.66(1.77)
HFPF(1)	31.18(3.00)	35.22(3.76)	37.40(4.55)	41.13(4.79)
HFPF(2)	30.32(4.83)	33.50(3.35)	39.00(3.76)	42.29(1.92)
HFPF(3)	27.88(3.91)	35.76(2.86)	36.94(4.77)	39.70(3.93)
HFPF(4)	30.67(3.51)	36.03(2.74)	38.53(2.76)	42.18(2.29)
HFPF(5)	30.54(3.88)	35.38(1.94)	37.54(2.58)	41.08(1.96)
LMSSC	**60.66**(**2.52**)	**68.08**(**2.74**)	**76.48**(**1.97**)	**87.28**(**1.38**)

Table 5: Semi-supervised classification accuracy (%) on Sonar dataset

Dataset	Sonar			
rate	0.1	0.2	0.3	0.5
AMGL	63.32(6.43)	71.85(3.44)	72.60(3.76)	74.95(5.97)
MLAN	61.78(4.93)	67.71(4.19)	69.11(3.80)	72.26(3.33)
HFPF(1)	66.50(4.35)	67.86(4.48)	69.17(3.60)	69.90(3.97)
HFPF(2)	55.51(4.00)	56.54(4.24)	57.97(4.19)	58.08(4.06)
HFPF(3)	57.03(4.68)	58.70(3.54)	59.72(3.69)	61.97(2.39)
LMSSC	**69.84**(**5.54**)	**80.21**(**4.65**)	**90.14**(**2.15**)	**95.48**(**1.29**)

Sonar, bad on BBC and Reuters. This instability is caused by the built graph which heavily depends on the quality of data. By contrast, LMSSC builds a graph on latent representation, which makes the graph robust to noise and outliers.



(a) $r = 5$ (b) $r = 10$ (c) $r = 15$

Figure 1: Parameter sensitivity for classification accuracy on Sonar dataset.

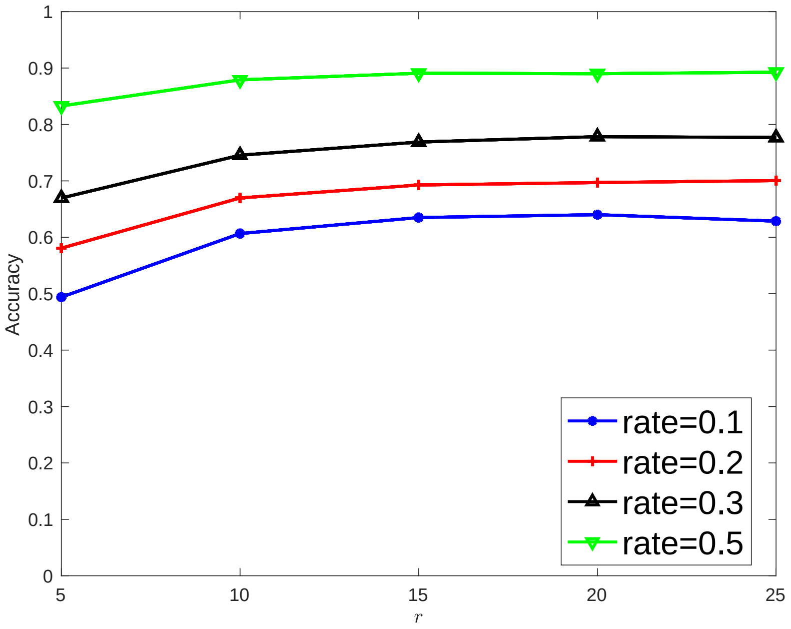


Figure 2: The influence of latent dimension r on accuracy of Sonar dataset.

5.4. Parameter Analysis

In our model, there are altogether four parameters: α, β, γ, and latent space dimension r. α is explicitly calculated. We use the Sonar dataset with 20% label rate as an example to demonstrate the effects of β, γ, and r. As shown in Fig. 1, accuracy changes slightly with a very wide range of parameters.

To explicitly demonstrate the influence of latent dimension, we vary the dimension r from 5 to 25 with interval 5. For each r value, we obtain its optimal accuracy. From Fig. 2, we can see that after r increases to 10, the performance becomes quite stable, which indicates that H contains redundant information.

6. Conclusion

In this paper, we propose a novel multi-view semi-supervised classification method. It first learns a unique latent representation from original multi-view data. Based on this latent representation, a graph with structure guarantee is constructed using adaptive neighbours strategy. Eventually, latent representation learning, graph construction, and label prediction are seamlessly integrated together, which enjoys the benefits of joint optimization. Extensive experiments on real datasets verify the effectiveness of our proposed approach.

Acknowledgments

This paper was in part supported by Grants from the Natural Science Foundation of China (Nos. 61806045 and 61572111), two Fundamental Research Fund for the Central Universities of China (Nos. ZYGX2017KYQD177 and A03017023701012) and a 985 Project of UESTC (No. A1098531023601041).

References

Richard H. Bartels and George W Stewart. Solution of the matrix equation ax+ xb= c [f4]. *Communications of the ACM*, 15(9):820–826, 1972.

Olivier Chapelle, Bernhard Scholkopf, and Alexander Zien. Semi-supervised learning (chapelle, o. et al., eds.; 2006)[book reviews]. *IEEE Transactions on Neural Networks*, 20 (3):542–542, 2009.

Hong Cheng, Zicheng Liu, and Jie Yang. Sparsity induced similarity measure for label propagation. In *2009 IEEE 12th international conference on computer vision*, pages 317–324. IEEE, 2009.

Ky Fan. On a theorem of weyl concerning eigenvalues of linear transformations i. *Proceedings of the National Academy of Sciences of the United States of America*, 35(11):652, 1949.

Yanwei Fu, Timothy M Hospedales, Tao Xiang, and Shaogang Gong. Transductive multi-view zero-shot learning. *IEEE transactions on pattern analysis and machine intelligence*, 37(11):2332–2345, 2015.

Ran He, Wei-Shi Zheng, Bao-Gang Hu, and Xiang-Wei Kong. Nonnegative sparse coding for discriminative semi-supervised learning. In *CVPR 2011*, pages 2849–2856. IEEE, 2011.

Tony Jebara, Jun Wang, and Shih-Fu Chang. Graph construction and b-matching for semi-supervised learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 441–448. ACM, 2009.

Zhao Kang, Chong Peng, and Qiang Cheng. Twin learning for similarity and clustering: A unified kernel approach. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

Zhao Kang, Xiao Lu, Jinfeng Yi, and Zenglin Xu. Self-weighted multiple kernel learning for graph-based clustering and semi-supervised classification. In *IJCAI*, pages 2312–2318, 2018a.

Zhao Kang, Chong Peng, Qiang Cheng, and Zenglin Xu. Unified spectral clustering with optimal graph. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18). AAAI Press*, 2018b.

Zhao Kang, Zipeng Guo, Shudong Huang, Siying Wang, Wenyu Chen, Yuanzhang Su, and Zenglin Xu. Multiple partitions aligned clustering. In *IJCAI*, pages 2701–2707, 2019a.

Zhao Kang, Yiwei Lu, Yuanzhang Su, Changsheng Li, and Zenglin Xu. Similarity learning via kernel preserving embedding. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence (AAAI-19). AAAI Press*, 2019b.

Zhao Kang, Haiqi Pan, Steven C. H. Hoi, and Zenglin Xu. Robust graph learning from noisy data. *IEEE Transactions on Cybernetics*, pages 1–11, 2019c. ISSN 2168-2267. doi: 10.1109/TCYB.2018.2887094.

Zhao Kang, Liangjian Wen, Wenyu Chen, and Zenglin Xu. Low-rank kernel learning for graph-based clustering. *Knowledge-Based Systems*, 163:510–517, 2019d.

Zhao Kang, Honghui Xu, Boyu Wang, Hongyuan Zhu, and Zenglin Xu. Clustering with similarity preserving. *Neurocomputing*, 2019e. doi: 10.1016/j.neucom.2019.07.086.

Sheng Li and Yun Fu. Learning balanced and unbalanced graphs via low-rank coding. *IEEE Transactions on Knowledge and Data Engineering*, 27(5):1274–1287, 2015.

Sheng Li, Hongfu Liu, Zhiqiang Tao, and Yun Fu. Multi-view graph learning with adaptive label propagation. In *2017 IEEE International Conference on Big Data (Big Data)*, pages 110–115. IEEE, 2017.

Bojan Mohar, Y Alavi, G Chartrand, and OR Oellermann. The laplacian spectrum of graphs. *Graph theory, combinatorics, and applications*, 2(871-898):12, 1991.

Feiping Nie, Hua Wang, Heng Huang, and Chris Ding. Unsupervised and semi-supervised learning via ℓ 1-norm graph. In *2011 International Conference on Computer Vision*, pages 2268–2273. IEEE, 2011.

Feiping Nie, Jing Li, Xuelong Li, et al. Parameter-free auto-weighted multiple graph learning: A framework for multiview clustering and semi-supervised classification. In *IJCAI*, pages 1881–1887, 2016.

Feiping Nie, Guohao Cai, and Xuelong Li. Multi-view clustering and semi-supervised classification with adaptive neighbours. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

Zhiqiang Tao, Hongfu Liu, Sheng Li, Zhengming Ding, and Yun Fu. From ensemble clustering to multi-view clustering. In *IJCAI*, 2017.

Fei Wang and Changshui Zhang. Label propagation through linear neighborhoods. *IEEE Transactions on Knowledge and Data Engineering*, 20(1):55–67, 2008.

Chang Xu, Dacheng Tao, and Chao Xu. A survey on multi-view learning. *arXiv preprint arXiv:1304.5634*, 2013.

Shuicheng Yan and Huan Wang. Semi-supervised learning by sparse representation. In *Proceedings of the 2009 SIAM International Conference on Data Mining*, pages 792–801. SIAM, 2009.

Changqing Zhang, Qinghua Hu, Huazhu Fu, Pengfei Zhu, and Xiaochun Cao. Latent multi-view subspace clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4279–4287, 2017.

Jing Zhao, Xijiong Xie, Xin Xu, and Shiliang Sun. Multi-view learning overview: Recent progress and new challenges. *Information Fusion*, 38:43–54, 2017.

Dengyong Zhou, Olivier Bousquet, Thomas N Lal, Jason Weston, and Bernhard Schölkopf. Learning with local and global consistency. In *Advances in neural information processing systems*, pages 321–328, 2004.

Xiaojin Zhu and Andrew B Goldberg. Introduction to semi-supervised learning. *Synthesis lectures on artificial intelligence and machine learning*, 3(1):1–130, 2009.

Xiaojin Zhu, Zoubin Ghahramani, and John D Lafferty. Semi-supervised learning using gaussian fields and harmonic functions. In *Proceedings of the 20th International conference on Machine learning (ICML-03)*, pages 912–919, 2003.

Liansheng Zhuang, Haoyuan Gao, Zhouchen Lin, Yi Ma, Xin Zhang, and Nenghai Yu. Non-negative low rank and sparse graph for semi-supervised learning. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2328–2335. IEEE, 2012.