Multi-view classification via Multi-view Partially Common Feature Latent Factor Learning

Jian-Wei Liu[1], Hao-Jie Xie[1], Run-Kun Lu[1], Xiong-Lin Luo[1]

1. Department of Automation, China University of Petroleum, Beijing Campus, Beijing, 102249
E-mail: liujw@cup.edu.cn
E-mail: 1049123204@qq.com
E-mail: 983194327@qq.com
E-mail: luoxl@cup.edu.cn

Abstract: In the real world, multi-view data usually consists of different representations or views. There are two key factors in multi-view data: consistency and complementarity. Unlike most multi-view learning algorithms based on nonnegative matrix factorization (NMF), the proposed method can make full use of the consistency and complementarity of data. This paper presents a new semi-supervised multi-view learning algorithm, called Multi-View Partially Common Feature Latent Factor (MVPCFLF) Learning. MVPCFLF is an extended learning form based on Partially Shared Latent Factor (PSLF) learning, which can make full use of common and special features to obtain latent representation. The key idea of MVPCFLF is to increase the constraints on common feature matrix so as to maintain the consistency of common features. The experimental results show that MVPCFLF is more effective than the existing multi-view learning algorithms.

Key Words: Multi-view Learning, Common Feature, Special Feature, Latent Representation Nonnegative Matrix Factorization, Complementarity and Consistency

1 INTRODUCTION

In real life, the various representations for most of objects or entities can be extracted from different sources with multi-modal measurements and every representation can be seen as a view for certain object or entity. For example, a web page can be composed of different constituents such as text, pictures, sounds, videos, and the like. A book can be published in various countries in different languages. Different redundant and complementary feature representations can reflect the underlying aspect characteristics of the examples from different ways through different view. When learning multi-view datasets, we can't rely on only one view to learn, which will result in the loss of diverse multi-factor and profile information. Therefore, it is very important to integrate the information in multiple views. Some methods treat the information contained in each view by converting multi-view data into single view data. However, these approaches ignore the potential interconnection between different views in multiple views, which doesn’t reflect the actual situation possessed by multi-view data.

To leverage information from different views, consistency and complementarity are important issues that we need to consider. Consistency in different views refers to the compatible information contained in all views. Complementarity in different views refers to some particular information contained in each view that is not available in other views, so the considering both of consistency and complementarity will be the end of the future. Up to now existing multi-view learning paradigms can be mainly divided into three major categories: co-training style algorithms, co-regularization style algorithms and margin-consistency style algorithms. [1]

1) Co-training style algorithms are enlightened by co-training [2]. Co-training is one of the earliest methods in the multi-view learning scheme. In the vein, the classifier alternately trains on two different views, and the first k most trusted data and its classification results in the two views are added to the labeled data until the added data is the same as the number of the unlabeled data [3], [4], [5].

2) The key idea for co-regularization style algorithms is to treat the inconsistency between the discriminant functions or regression functions of the two views as a regularization term in the objective function [6], [7],[8],[9].

3) Margin-consistent algorithms have also recently been proposed to exploit the potential consistency of multi-view classification results [10], [11],[12],[13].

NMF-based multi-view learning algorithm belongs to the second category in the above categories：co-regularization style algorithms. NMF have been widely used in many fields such as image analysis, text mining and speech processing. In [14], D. Lee et al. first introduced the concept of NMF. NMF is an algorithm for multivariate analysis and linear algebra. We refer the readers to the literatures [15],[16],[17],[18],[19] and the reference therein for NMF-based multi-view learning algorithms. In addition, solving the convex clustering problem proposed by Eric C. Chi et al. can give us a good inspiration [20].

Since the NMF method of unsupervised learning cannot combine label information and multi-view information, the PSLF is proposed to address this deficiency and effectively solves this problem [16]. PSLF is a semi-supervised learning approach that operates by dividing the data matrix

This work is supported by the Science Foundation of China University of Petroleum Beijing (2462018QZDX02)

978-1-7281-5855-6/20/$31.00 ©2020 IEEE

$\mathbf{X}$ into two parts, i.e., labeled and unlabeled part. In order to conform to the consistency and complementarity principles, the latent representation matrix $\mathbf{U}$ and $\mathbf{V}$ which are acquired by NMF is further disjointedly divided into a special part and a common part. However, the constraints incorporating by PSLF on $\mathbf{V}_s$ and $\mathbf{V}_c$ is too stringent and couldn't reflect the actual situation underlying in multi-view data. To address the drawback, and inspired by Multi-NMF [15], we propose an extended learning approach based on PSLF, which we dub it as Multi-View Partially Common Feature Latent Factor (MVPCFLF). Our introducing constraints on $\mathbf{V}_c$,which follow the consistency principle, ensure that the shared information in all view is consistent. The main contributions of this paper are summarized as follows:

1) We combine the pros and discard the cons existing in Multi-NMF and PSLF to alleviate the limits in PSLF.

2) Different from most existing multi-view latent factor learning methods, MVPCFLF first divides the data matrix into special matrix and common matrix, and then carries out NMF operation. We have previously divided the $\mathbf{U}$ and $\mathbf{V}$ matrices, the purpose of this method is to ensure that the matrix can be independent.

3) We will operate on different parts of the eigenvalue matrix $\mathbf{V}$ .Firstly, we use $\mathbf{V}_c^*$ to combine the common features of $\mathbf{V}_c^1, \mathbf{V}_c^2, ..., \mathbf{V}_c^p$, which is the common feature matrix of each view, to highlight the consistency information.

4) We take the learning result from MVPCFLF as input to test the KNN classification on eight different datasets. The experimental results show that the MVPCFLF has better performance on the four data sets than the previous NMF algorithm. Especially, the test accuracy of the Reuters data set reached 78.4%, an obvious promotion of over 8.1% compared to the DICS.

2 OVERIEW

In this section, we will briefly introduce the algorithms which are related to MVPCFLF, including: Non-Negative Matrix Factorization (NMF), Multi-NMF and Partially Shared Latent Factor Learning (PSLF). NMF is the most basic algorithm, and Multi-NMF and PSLF are expansion algorithms based on NMF.

2.1 NMF

Supposed that $\mathbf{X} = \left[\mathbf{X}_{.,1}, ..., \mathbf{X}_{.,N}\right] \in \mathbb{R}_+^{M\times N}$ is denoted the data matrix. Each column of the matrix $\mathbf{X}$ represents a data point and each row is composed of feature components. The NMF method is to find two non-negative matrix factors $\mathbf{U} = \left[\mathbf{U}_{i,k}\right] \in \mathbb{R}_+^{M\times K}$ and $\mathbf{V} = \left[\mathbf{V}_{k,j}\right] \in \mathbb{R}_+^{K\times N}$ to satisfy the following equation:

$$\mathbf{X} \approx \mathbf{UV} \qquad (1)$$

where K is the dimension after dimensionality reduction and we called $\mathbf{U}$ as the base matrix and $\mathbf{V}$ as the coefficient matrix. We can use Frobenius norm to formulate the optimization problem:

$$\underset{\mathbf{U},\mathbf{V}}{\operatorname{argmin}} \left\|\mathbf{X} - \mathbf{UV}\right\|_F^2, s.t.\ \mathbf{U} \geq 0, \mathbf{V} \geq 0 \qquad (2)$$

Where $\|\cdot\|_F^2$ is the Frobenius norm and $\mathbf{U} \geq 0$ and $\mathbf{V} \geq 0$ represent every element in $\mathbf{U}$ and $\mathbf{V}$ must be non-negative. NMF is the basis of subsequent work.

2.2 Multi-NMF

In this subsection, we will introduce Multi-NMF, a joint NMF-based multi-view learning algorithm. The difference between Joint-NMF and NMF is that joint-NMF performs NMF on each view separately. Supposed that $\left\{\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, ..., \mathbf{X}^{(n_v)}\right\}$ indicates that there are n_v views, and each view $\mathbf{X}^{(v)}$ can be factorized into non-negative matrix $\mathbf{U}^{(v)}$ and $\mathbf{V}^{(v)}$,i.e., $\mathbf{X}^{(v)} \approx \mathbf{U}^{(v)}\mathbf{V}^{(v)}$ .When there are the same number of data points for different views, but for different data points there may possess different number of attributes, the dimensions of $\mathbf{V}^{(v)}$ for each view are the same, but the row dimensions of $\mathbf{U}^{(v)}$ for each view are different. The core idea of Multi-NMF is to use the common matrix $\mathbf{V}^* \in \mathbb{R}_+^{K\times N}$ to achieve the swift between different views. The goal is to enable $\mathbf{V}^*$ to contain more views' information than just a single view. It is mainly expressed by the following constraint:

$$D\left(\mathbf{V}^{(v)}, \mathbf{V}^*\right) = \left\|\mathbf{V}^{(v)} - \mathbf{V}^*\right\|_F^2 \qquad (3)$$

Therefore, the Multi-NMF method uses a common coefficient matrix $\mathbf{V}^*$ to predict the output. The loss function of the Multi-NMF is as follows:

$$\sum_{v=1}^{n_v}\left\|\mathbf{X}^{(v)} - \mathbf{U}^{(v)}\mathbf{V}^{(v)}\right\|_F^2 + \sum_{v=1}^{n_v}\lambda_v\left\|\mathbf{V}^{(v)} - \mathbf{V}^*\right\|_F^2 \qquad (4)$$
$$s.t.\ \mathbf{U}^{(v)}, \mathbf{V}^{(v)}, \mathbf{V}^* \geq 0$$

Where λ_v is the only trade-off parameter in the Multi-NMF that not only adjusts the relative weight between different views, but also determines the relative weight between the joint multi-view NMF reconstruction error and the inconsistent term $D\left(\mathbf{V}^{(v)}, \mathbf{V}^*\right)$.

2.3 PSLF

For multi-view datasets with P views $\left\{\mathbf{X}^p, \mathbf{Y}\right\}_{p=1}^P$,PSLF is a method proposed by Liu et al[16]in 2014. PSLF method mainly realizes the learning of latent features by partitioning matrices. PSLF is also a multi-view learning algorithm based on NMF. The core thought of the PSLF is to block the matrix， as we can see from Fig.1. PSLF partitions the coefficient matrix $\mathbf{V}^p$ in the p-th view into two parts: $\mathbf{V}_c$ and $\mathbf{V}_s^p$. $\mathbf{V}_c$ is a common information matrix and its shape is the same in each view , and $\mathbf{V}_c$ reflects the consistent information of all views. $\mathbf{V}_s^p$ is a special information matrix and Its dimensions are different in each view, and $\mathbf{V}_s^p$ reflects complementary information of all views. The PSLF also divides the base matrix $\mathbf{U}^p$ in the p-th view into two parts: $\mathbf{U}_s^p$ and $\mathbf{U}_c^p$, which represent special and common parts, respectively. In addition, since PSLF is a semi-supervised learning method, it is necessary to chop $\mathbf{V}_c$ and $\mathbf{V}_s^p$ into labeled and unlabeled parts.

PSLF updates the matrix through NMF, so the loss function is:

$$\underset{\mathbf{V}_l,\mathbf{W}}{\operatorname{argmin}} \sum_{p=1}^{P} \pi^p \left\|\mathbf{X}^p - \mathbf{U}^p\mathbf{V}^p\right\|_F^2 + \lambda\left\|\Pi\right\|_2^2 + \beta\begin{pmatrix}\left\|\mathbf{W}^T\mathbf{V}_l - \mathbf{Y}\right\|_F^2 \\ +\gamma\left\|\mathbf{W}\right\|_{2,1}\end{pmatrix}$$
$$s.t.\ \mathbf{U}^1,...,\mathbf{U}^p,\ \mathbf{V}^1,...,\mathbf{V}^p,\Pi = \left(\pi^1,\pi^2,...,\pi^p\right) \geq 0, \sum_{p=1}^{P}\pi^p = 1 \tag{5}$$

where β is a nonnegative tradeoff parameter , $\mathbf{V}_l = [\mathbf{V}_s^1,...,\mathbf{V}_s^p;\mathbf{V}_c]$, the parameter λ controls the smoothness of $\prod$. Composite norm term $L_{2,1}$ is to ensure the row sparsity of matrix $\mathbf{W}$.

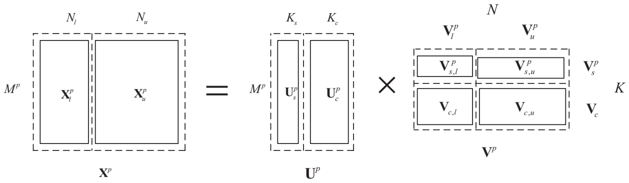

Fig.1. Matrix factorization in the p-th view

3 OUR PROPOSED APPROACH

In this section, we first present the splitting processes for each view. By incorporating $l_{2,1}$ Regularized term, we take advantage of the partition results to formulate the objective function of regression prediction with $l_{2,1}$ Regularized term for output.

3.1 Splitting for NMF

Unlike the traditional NMF method, for each view' base matrix $\mathbf{U}^p$, our method needs to divide the p- th view' base matrix $\mathbf{U}^p$ into disjoint two parts: $\mathbf{U}_\text{s}^p$ and $\mathbf{U}_\text{c}^p$,i.e., $\mathbf{U}^p = \left[\mathbf{U}_\text{s}^p,\mathbf{U}_\text{c}^p\right]$ .Similarly, we also partition the p -th view' coefficient matrix $\mathbf{V}^p$ into disjoint two parts: $\mathbf{V}_s^p$ and $\mathbf{V}_c^p$ i.e., $\mathbf{V}^P = [\mathbf{V}_s^p;\mathbf{V}_c^p]$. Unlike PSLF, this decomposition scheme leads to share a common matrix V_c^p for each view. At this time, the loss function for the traditional NMF needs to be revised to suit our aim. The loss function is defined as follows:

$$\left\|\mathbf{X}^p - \mathbf{U}_\text{s}^p\mathbf{V}_s^p - \mathbf{U}_\text{c}^p\mathbf{V}_c^p\right\|_F^2 \tag{6}$$

The advantage of our proposed partition method is that it can divide the public and private parts of each view in advance and better discover the underlying characteristics of each view data. To be able to describe the problem more clearly, we first define a constant ratio $\eta = (K_c / K_s + K_c)$ as proportion of public information, where K_c represents the latent feature dimension of the public part and K_s represents the latent feature dimension of the private part. When the value of η is larger, consistency attributes are dominant, and when the value of η is smaller, complementarity attributes are dominant, thus the corresponding objective function is:

$$\underset{\substack{\mathbf{U}_\text{s}^p,\mathbf{V}_s^p,\mathbf{U}_\text{c}^p,\mathbf{V}_c^p \\ p=1,...,P}}{\operatorname{argmin}} \sum_{p=1}^{P}\alpha^p\left\|\mathbf{X}^p - \mathbf{U}_\text{s}^p\mathbf{V}_s^p - \mathbf{U}_\text{c}^p\mathbf{V}_c^p\right\|_F^2$$
$$s.t.\quad \mathbf{U}^1,...,\mathbf{U}^P,\mathbf{V}^1,...,\mathbf{V}^P \geq 0, \sum_{p=1}^{P}\alpha^p = 1 \tag{7}$$

Where α^p represents the weight of p-th view. To adapt to semi-supervised learning scenarios, we also further partition the submatrices $\mathbf{V}_s^p$ and $\mathbf{V}_c^p$ of view' coefficient matrix $\mathbf{V}^p$ into labeled and unlabeled subparts which is similar to the partitioning process of PSLF. The four parts of $\mathbf{V}^p$ are shown in Fig.2.

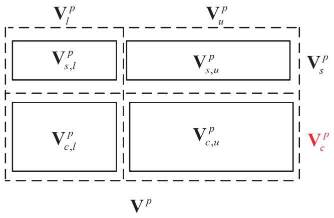

Fig.2. Matrix factorization of $\mathbf{V}^p$ in the p-th view

3.2 Formulating the objective function of regression prediction with $l_{2,1}$ Regularized term for output

In this subsection, we incorporate the same regularization term as PSLF in objective function of regression prediction, but comparing with PSLF, since each view' coefficient matrix $\mathbf{V}^p$ contains one more submatrix $\mathbf{V}_c$. So in MVPCFLF each view' coefficient matrix has a specific form $\mathbf{V}_l^p = \left[\mathbf{V}_s^1,...,\mathbf{V}_s^P;\mathbf{V}_c^p\right]$ .when MVPCFLF gets some of the label information $\mathbf{Y}$, by minimizing the following problem:

$$\underset{\mathbf{W}^p,p=1,\cdots,P}{\operatorname{argmin}}\ \beta\left(\sum_{p=1}^{P}\left\|(\mathbf{W}^p)^T\mathbf{V}_l^p - \mathbf{Y}\right\|_F^2 + \gamma\left\|\mathbf{W}^p\right\|_{2,1}\right) \tag{8}$$
$$s.t.\quad \mathbf{V}_l^1,...,\mathbf{V}_l^P \geq 0$$

we can predict the output labels. Note that each view has a regression coefficient matrix $\mathbf{W}^p$. The purpose of introducing $l_{2,1}$ norm regularization term is to ensure that the $\mathbf{W}^p$, rows are sparse in each view.

By introducing above constrained optimization problem, our proposed method can be utilized in semi-supervised learning scenarios.

3.3 Shared consistency matrix $\mathbf{V}_c^*$

In MVPCFLF, information in all views is integrated by adding a shared consistency matrix $\mathbf{V}_c^*$ to the public part of each view. The difference between the public coefficient matrix $\mathbf{V}_c^p$ and $\mathbf{V}_c^*$ in each view is measured by the following loss function:

$$D(\mathbf{V}_c^p, \mathbf{V}_c^*) = \left\| \mathbf{V}_c^p - \mathbf{V}_c^* \right\|_F^2 \tag{9}$$

Through this difference function, we can extract the public information $\mathbf{V}_c^*$ from all views, which is consistent with the principle of consistency in multiple views.

3.4 Unified Objective Function for MVPCFLF

In order to be able to combine consistency and complementarity information in multiple views, according to above discussion, we formulate the unified objective function for MVPCFLF as follows:

$$\begin{aligned}
&\underset{\substack{\mathbf{V}_l^p, \mathbf{W}^p \\ p=1,\dots,P}}{\operatorname{argmin}} \sum_{p=1}^{P} \alpha^p \left\| \mathbf{X}^p \text{-} \mathbf{U}_{\mathrm{s}}^p \mathbf{V}_s^p \text{-} \mathbf{U}_{\mathrm{c}}^p \mathbf{V}_c^p \right\|_F^2 + \sum_{p=1}^{P} \mu \left\| \mathbf{V}_c^p - \mathbf{V}_c^* \right\|_F^2 \\
&+\beta \left(\sum_{p=1}^{P} \left\| (\mathbf{W}^p)^T \mathbf{V}_l^p - \mathbf{Y} \right\|_F^2 + \gamma \left\| \mathbf{W}^p \right\|_{2,1} \right) \qquad (10) \\
&s.t. \quad \mathbf{U}^1, \dots, \mathbf{U}^P, \mathbf{V}^1, \dots, \mathbf{V}^P \ge 0, \sum_{p=1}^{P} \alpha^p = 1, \mathbf{V}_c^* \ge 0
\end{aligned}$$

where β , μ , γ are nonnegative tradeoff parameters.

3.5 Optimization

Number citations consecutively in square brackets [1]. The sentence punctuation follows the brackets [2].

Please note that the references at the end of this document are in the preferred referencing style. Give all authors' names; do not use "et al." unless there are six authors or more. Use a space after authors' initials.

We define the joint optimization function :

$$\begin{aligned}
&\mathcal{O}\left(\mathbf{U}^1, \dots, \mathbf{U}^P, \mathbf{V}^1, \dots, \mathbf{V}^P, \mathbf{W}^1, \dots, \mathbf{W}^P, \mathbf{V}_c^* \right) \\
&= \sum_{p=1}^{P} \alpha^p \left\| \mathbf{X}^p \text{-} \mathbf{U}_{\mathrm{s}}^p \mathbf{V}_s^p \text{-} \mathbf{U}_{\mathrm{c}}^p \mathbf{V}_c^p \right\|_F^2 + \sum_{p=1}^{P} \mu \left\| \mathbf{V}_c^p - \mathbf{V}_c^* \right\|_F^2 \qquad (11) \\
&+\beta \left(\sum_{p=1}^{P} \left\| (\mathbf{W}^p)^T \mathbf{V}_l^p - \mathbf{Y} \right\|_F^2 + \gamma \left\| \mathbf{W}^p \right\|_{2,1} \right)
\end{aligned}$$

Then, the optimization problem (10) can be iteratively solved by following four steps. We summarize the updating process in Algorithm 1.

(1) updating rule for $\mathbf{U}_{\mathrm{s}}^p$ and $\mathbf{U}_{\mathrm{c}}^p$

Since we divide the $\mathbf{U}^p$ matrix into the public part $\mathbf{U}_{\mathrm{c}}^p$ and the private part $\mathbf{U}_{\mathrm{s}}^p$, without loss of generality, where we will give the update process of the p -th view in detail, and the update process is equally applicable to other views. First, we define $\mathcal{L}(\mathbf{U}^p)$ as the part of $\mathcal{O}$ that is related to $\mathbf{U}^p$, The Lagrange function $\mathcal{L}(\mathbf{U}^p)$ is defined as:

$$\mathcal{L}(\mathbf{U}^p) = \left\| \mathbf{X}^p \text{-} \mathbf{U}_{\mathrm{s}}^p \mathbf{V}_s^p \text{-} \mathbf{U}_{\mathrm{c}}^p \mathbf{V}_c^p \right\|_F^2 + \mu \left\| \mathbf{V}_c^p \mathbf{Q}^p - \mathbf{V}_c^* \right\|_F^2 + \mathrm{Tr}(\psi \mathbf{U}^p) \tag{12}$$

where $\psi = [\Psi_s, \Psi_c] = [\psi_{mk}]$, ψ_{mk}^p is the Lagrange multiplier for $\mathrm{U}^p \ge 0$. Paying attention to the regularization term $\left\| \mathbf{V}_c^p \mathbf{Q}^p - \mathbf{V}_c^* \right\|_F^2$, the specific form of $\mathbf{Q}^p$ is as follows:

$$\mathbf{Q}^p = Diag\left(\sum_{i=1}^{M} \mathbf{U}_{i,1}^p, \sum_{i=1}^{M} \mathbf{U}_{i,2}^p, \dots, \sum_{i=1}^{M} \mathbf{U}_{i,K_c}^p \right) \tag{13}$$

the derivative of $\mathbf{U}_{\mathrm{c}}^p$ is described as follows:

$$\begin{aligned}
&\frac{\partial \mathcal{L}\left(\mathbf{U}^p \right)}{\partial \mathbf{U}_{\mathrm{c}}^p} = \text{-}2\mathbf{X}(\mathbf{V}_c^p)^T + 2\mathbf{U}_{\mathrm{c}}^p \mathbf{V}_c^p (\mathbf{V}_c^p)^T + 2\mathbf{U}_{\mathrm{s}}^p \mathbf{V}_s^p (\mathbf{V}_c^p)^T \\
&+2\mu\left(\sum_{l=1}^{M} \mathbf{U}_{l,K_c}^p \sum_{j=1}^{M} (\mathbf{V}_{c\,j,K_c}^p)^2 - \sum_{j=1}^{M} \mathbf{V}_{c\,j,K_c}^p \mathbf{V}_{c\,j,K_c}^* \right) + \psi_{mK_c} = 0
\end{aligned} \tag{14}$$

Using the Karush-Kuhn-Tucker condition $\psi_{mK_c} \mathbf{U}_c^p = 0$, we can acquire the following updating rule for $\mathbf{U}_{\mathrm{c}}^p$:

$$\mathbf{U}_{\mathrm{c}}^p \leftarrow \mathbf{U}_{\mathrm{c}}^p \frac{\mathbf{X}^{\mathrm{p}} (\mathbf{V}_{\mathrm{c}})^T + \mu \sum_{j=1}^{M} \mathbf{V}_{c\,j,K_c}^p \mathbf{V}_{c\,j,K_c}^*}{\left(\begin{array}{l} \mathbf{U}_{\mathrm{c}}^p \mathbf{V}_c^p (\mathbf{V}_c^p)^T + \mathbf{U}_{\mathrm{s}}^p \mathbf{V}_s^p (\mathbf{V}_c^p)^T \\ + \mu \sum_{l=1}^{M} \mathbf{U}_{l,K_c}^p \sum_{j=1}^{M} (\mathbf{V}_{c\,j,K_c}^p)^2 \end{array} \right)} \tag{15}$$

Similarly, when we update U_s^p , we don't need to consider items that contain $\mathrm{U}_{\mathrm{c}}^p$, so the updating rule for U_s^p is:

$$\mathbf{U}_s^p \leftarrow \mathbf{U}_s^p \frac{\mathbf{X}^{\mathrm{p}} (\mathbf{V}_s)^T}{\mathbf{U}_s^p \mathbf{V}_s^p (\mathbf{V}_s^p)^T + \mathbf{U}_c^p \mathbf{V}_c^p (\mathbf{V}_s^p)^T} \tag{16}$$

(2) updating rule for $\mathbf{W}^p$

Unlike PSLF, we only extract the items from $\mathcal{O}$ in the p -th view which related to $\mathbf{W}^p$, the loss function related to $\mathbf{W}^p$ in $\mathcal{O}$ is as follows:

$$\mathcal{L}(\mathbf{W}^p) = \beta \left(\left\| (\mathbf{W}^p)^T \mathbf{V}_l^p - \mathbf{Y} \right\|_F^2 + \gamma \left\| \mathbf{W}^p \right\|_{2,1} \right) \tag{17}$$

the derivative of objective function $\mathcal{L}(\mathbf{W}^p)$ about W^p is:

$$\frac{\partial \mathcal{L}\left(\mathbf{W}^p \right)}{\partial \mathbf{W}^p} = 2\beta \left(\mathbf{V}_l^p \left((\mathbf{W}^p)^T \mathbf{V}_l^p - \mathbf{Y} \right) \right)^T + \gamma \mathbf{E}^p \mathbf{W}^p \tag{18}$$

where $\mathbf{E}$ is a diagonal matrix with $e_{kk} = 1/2 \left\| \mathbf{w}_k^{\mathrm{p}} \right\|_2$. Setting $\partial \mathcal{L}\left(\mathbf{W}^p \right) / \partial \mathbf{W}^p = 0$, we can get the following updating rule:

$$\mathbf{W}^p = \left(\mathbf{V}_l^p (\mathbf{V}_l^p)^T + \gamma \mathbf{E}^p \right)^{-1} \mathbf{V}_l^p \mathbf{Y}^T \tag{19}$$

(3) updating rule for $\mathbf{V}^p$

In order to derive the updating rule for $\mathbf{V}^p$, we need to describe the specific splitting process for the coefficient matrix $\mathbf{V}^p$ in the p -th view. Since the matrix $\mathbf{V}^p$ contains the common and special part, the labeled and the unlabeled part, so $\mathbf{V}^p$ can be divided into four parts: $\mathbf{V}_{s,l}^p, \mathbf{V}_{s,u}^p, \mathbf{V}_{c,l}^p$ and $\mathbf{V}_{c,u}^p$. $\mathbf{X}^p$ will also be divided into $\mathbf{X}_l^p$ and $\mathbf{X}_u^p$ .Correspondingly, $\mathbf{U}^p$ is also divided into $\mathbf{U}_{\mathrm{c}}^{p}$ and $\mathbf{U}_{s}^{p}$ .Let ϕ_{kn} is the Lagrange multiplier for $\mathbf{V}_{kn} \geq 0$ and $\Phi=[[\phi_{s,l},\phi_{s,u}];[\phi_{c,l},\phi_{c,u}]]=[\phi_{kn}]$.

We extract the items from $\mathcal{O}$ in the p -th view which related to $\mathbf{V}_{s,l}^{p},\mathbf{V}_{s,u}^{p},\mathbf{V}_{c,l}^{p}$ and $\mathbf{V}_{c,u}^{p}$, and the same process is also suitable to the other views, Given space limitations, no further details are given here. The corresponding loss function is defined as:

$$\mathcal{L}\left(\mathbf{V}^{p}\right)=\alpha^{p}\left\|\mathbf{X}^{p}\text{-}\mathbf{U}_{\mathrm{s}}^{p}\mathbf{V}_{s}^{p}\text{-}\mathbf{U}_{\mathrm{c}}^{p}\mathbf{V}_{c}^{p}\right\|_{F}^{2}+\beta\left\|(\mathbf{W}^{p})^{T}\mathbf{V}_{l}^{p}-\mathbf{Y}\right\|_{F}^{2}$$
$$+\mu\left\|\mathbf{V}_{c}^{p}\mathbf{Q}^{p}-\mathbf{V}_{c}^{*}\right\|_{F}^{2}+\mathrm{Tr}(\Phi^{T}\mathbf{V}^{p}) \tag{20}$$

the derivative of objective function $\mathcal{L}(\mathbf{W}^{p})$ about $\mathbf{V}_{s,l}^{p},\mathbf{V}_{s,u}^{p},\mathbf{V}_{c,l}^{p}$ and $\mathbf{V}_{c,u}^{p}$ are:

$$\frac{\partial\mathcal{L}}{\partial\mathbf{V}_{s,l}^{p}}=2\alpha^{p}\begin{pmatrix}-\left(\mathbf{U}_{s}^{p}\right)^{T}\mathbf{X}_{l}^{p}+\left(\mathbf{U}_{s}^{p}\right)^{T}\mathbf{U}_{s}^{p}\mathbf{V}_{s,l}^{p}\\+\left(\mathbf{U}_{s}^{p}\right)^{T}\mathbf{U}_{c}^{p}\mathbf{V}_{c,l}^{p}\end{pmatrix}+2\beta\mathbf{B}_{s}^{p}+\phi_{s,l}=0$$

$$\frac{\partial\mathcal{L}}{\partial\mathbf{V}_{s,u}^{p}}=2\alpha^{p}\begin{pmatrix}-\left(\mathbf{U}_{s}^{p}\right)^{T}\mathbf{X}_{u}^{p}+\left(\mathbf{U}_{s}^{p}\right)^{T}\mathbf{U}_{s}^{p}\mathbf{V}_{s,u}^{p}\\+\left(\mathbf{U}_{s}^{p}\right)^{T}\mathbf{U}_{c}^{p}\mathbf{V}_{c,u}^{p}\end{pmatrix}+\phi_{s,u}=0$$

$$\frac{\partial\mathcal{L}}{\partial\mathbf{V}_{c,l}^{p}}=2\alpha^{p}\begin{pmatrix}-\left(\mathbf{U}_{c}^{p}\right)^{T}\mathbf{X}_{l}^{p}+\left(\mathbf{U}_{c}^{p}\right)^{T}\mathbf{U}_{c}^{p}\mathbf{V}_{c,l}^{p}\\+\left(\mathbf{U}_{c}^{p}\right)^{T}\mathbf{U}_{s}^{p}\mathbf{V}_{s,l}^{p}\end{pmatrix}+2\beta\mathbf{B}_{c}^{p}+\phi_{c,l}$$
$$+2\mu\left(\mathbf{V}_{c,l}^{p}-\mathbf{V}_{c,l}^{*}\right)=0$$

$$\frac{\partial\mathcal{L}}{\partial\mathbf{V}_{c,u}^{p}}=2\alpha^{p}\begin{pmatrix}-\left(\mathbf{U}_{c}^{p}\right)^{T}\mathbf{X}_{u}^{p}+\left(\mathbf{U}_{c}^{p}\right)^{T}\mathbf{U}_{c}^{p}\mathbf{V}_{c,u}^{p}\\+\left(\mathbf{U}_{c}^{p}\right)^{T}\mathbf{U}_{s}^{p}\mathbf{V}_{s,u}^{p}\end{pmatrix}+\phi_{c,u}$$
$$+2\mu\left(\mathbf{V}_{c,u}^{p}-\mathbf{V}_{c,u}^{*}\right)=0 \tag{21}$$

where $\mathbf{B}^{p}=\mathbf{W}^{p}\left(\mathbf{W}^{p}\right)^{T}\mathbf{V}_{l}-\mathbf{W}^{p}\mathbf{Y}=\left[\mathbf{B}_{s}^{1};\ldots;\mathbf{B}_{s}^{p};\mathbf{B}_{c}^{p}\right]$, using Karush Kuhn Tucker (KKT) conditions, we have:

$$\mathbf{V}_{s,l}^{p}\leftarrow\mathbf{V}_{s,l}^{p}\frac{\alpha^{p}\left(\mathbf{U}_{s}^{p}\right)^{T}\mathbf{X}_{l}^{p}+\beta\left(\mathbf{B}_{+}\right)_{s}^{p}}{\alpha^{p}\left(\mathbf{U}_{s}^{p}\right)^{T}\mathbf{U}_{s}^{p}\mathbf{V}_{s,l}^{p}+\alpha^{p}\left(\mathbf{U}_{s}^{p}\right)^{T}\mathbf{U}_{c}^{p}\mathbf{V}_{c,l}^{p}+\beta\left(\mathbf{B}_{-}\right)_{s}^{p}}$$

$$\mathbf{V}_{s,u}^{p}\leftarrow\mathbf{V}_{s,u}^{p}\frac{\alpha^{p}\left(\mathbf{U}_{s}^{p}\right)^{T}\mathbf{X}_{u}^{p}}{\alpha^{p}\left(\mathbf{U}_{s}^{p}\right)^{T}\mathbf{U}_{s}^{p}\mathbf{V}_{s,u}^{p}+\alpha^{p}\left(\mathbf{U}_{s}^{p}\right)^{T}\mathbf{U}_{c}^{p}\mathbf{V}_{c,u}^{p}} \tag{22}$$

$$\mathbf{V}_{c,l}^{p}\leftarrow\mathbf{V}_{c,l}^{p}\frac{\alpha^{p}\left(\mathbf{U}_{c}^{p}\right)^{T}\mathbf{X}_{l}^{p}+\beta\left(\mathbf{B}_{+}\right)_{c}^{p}+\mu\mathrm{V}_{c,l}^{*}}{\begin{pmatrix}\alpha^{p}\left(\mathbf{U}_{c}^{p}\right)^{T}\mathbf{U}_{c}^{p}\mathbf{V}_{c,l}^{p}+\alpha^{p}\left(\mathbf{U}_{c}^{p}\right)^{T}\mathbf{U}_{s}^{p}\mathbf{V}_{s,l}^{p}\\+\beta\left(\mathbf{B}_{-}\right)_{c}^{p}+\mu\mathrm{V}_{c,l}^{p}\end{pmatrix}}$$

$$\mathbf{V}_{c,u}^{p}\leftarrow\mathbf{V}_{c,u}^{p}\frac{\alpha^{p}\left(\mathbf{U}_{c}^{p}\right)^{T}\mathbf{X}_{l}^{p}+\mu\mathrm{V}_{c,u}^{*}}{\alpha^{p}\left(\mathbf{U}_{c}^{p}\right)^{T}\mathbf{U}_{c}^{p}\mathbf{V}_{c,u}^{p}+\alpha^{p}\left(\mathbf{U}_{c}^{p}\right)^{T}\mathbf{U}_{s}^{p}\mathbf{V}_{s,u}^{p}+\mu\mathrm{V}_{c,u}^{p}} \tag{23}$$

where $\mathbf{B}$ needs to be transformed according to identities: $A^{+}=\left(|A|+A\right)/2\ \mathit{and}\ A^{-}=\left(|A|-A\right)/2$, thus we can get $\mathbf{B}_{+}^{p}$ and $\mathbf{B}_{-}^{p}$:

$$\mathbf{B}_{+}^{p}=\left(\left(\mathbf{W}^{p}\left(\mathbf{W}^{p}\right)^{T}\right)^{+}\mathbf{V}_{l}+\left(\mathbf{W}^{p}\right)^{-}\mathbf{Y}\right)$$
$$=\left[\left(\mathbf{B}_{+}\right)_{s}^{1};\ldots;\left(\mathbf{B}_{+}\right)_{s}^{p};\left(\mathbf{B}_{+}\right)_{c}^{p}\right]$$
$$\mathbf{B}_{-}^{p}=\left(\left(\mathbf{W}^{p}\left(\mathbf{W}^{p}\right)^{T}\right)^{-}\mathbf{V}_{l}+\left(\mathbf{W}^{p}\right)^{+}\mathbf{Y}\right)$$
$$=\left[\left(\mathbf{B}_{-}\right)_{s}^{1};\ldots;\left(\mathbf{B}_{-}\right)_{s}^{p};\left(\mathbf{B}_{-}\right)_{c}^{p}\right] \tag{24}$$

(4) updating rule for $\mathbf{V}^{*}$

Let the derivative of objective function $\mathcal{O}$ about $\mathbf{V}^{*}$ equals to zero:

$$\frac{\partial\mathcal{O}}{\partial\mathbf{V}^{*}}=\frac{\partial\sum_{p=1}^{P}\mu\left\|\mathbf{V}_{c}^{p}\mathbf{Q}^{p}-\mathbf{V}^{*}\right\|_{F}^{2}}{\partial\mathbf{V}^{*}}=0 \tag{25}$$

We can acquire the following updating rule for $\mathbf{V}^{*}$:

$$\mathbf{V}^{*}=\frac{\sum_{p=1}^{P}\mu\mathbf{V}_{c}^{p}\mathbf{Q}^{p}}{\sum_{p=1}^{P}\mu} \tag{26}$$

4 EXPERIMENT

In this section, we run our experiments to perform classification tasks on eight real world datasets. The experimental results are used to validate the effectiveness of the proposed MVPCFLF algorithm. and In addition, we also studied the effect of parameter changes on the accuracies and F1 scores of the classification task for MVPCFLF on the same data set. Our experimental environment is i7-4700mq and 16g memory. The algorithm is written by MATLAB, and the time consumed to complete 200 epoch calculation on different data sets is no more than 15 minutes.

The results derived from MVPCFLF and unsupervised comparison baselines are served as input of classifiers. More specifically, we will get a latent feature matrix $\mathbf{V}_{r}=[\mathbf{V}_{s}^{1},\mathbf{V}_{s}^{2},\ldots,\mathbf{V}_{s}^{P};\mathbf{V}_{c}^{*}]$ from MVPCFLF, utilize $\mathbf{V}_{r}$ as the input of the KNN classifiers to classify test examples.

4.1 Datasets

In our experiments, we used eight real-world multi-view datasets to verify the performance of MVPCFLF, including the Leaves, Yale-Face， Cornell, Texas, Washington, Wisconsin, Reuters and BBC datasets.

Leaves (one-hundred plant species leaves data set) is a data set containing 100 kinds of leaf information, each sample represents all the information of a leaf. The data set has 3 views, 6 categories, and each view's each sample is 64 dimensional.

Reuters (RCV1 / RCV2 multilingual data set) is a text data set containing a large number of Reuters news stories for research and development of natural language processing, information retrieval and machine learning. This data set has 5 views and 6 category, each sample of each view is 2000 dimensions.

YaleFace (Extended Yale Face Database B) is a commonly used face dataset, which contains 16,128 images of 28 human objects under 9 poses and 64 lighting conditions. The dataset has 2 views and 8 categories, and each sample of each view is 2016 dimensions.

The BBC dataset is a set of multi-view text datasets. It is formed by segmenting each article in the BBC news corpus, and according to different segmentation and construction methods, a total of 6 different datasets are generated. The dataset has 4 views and 6 categories. The dimensions of each view are: 4659,4633,4665,4684.

The WebKB dataset is a dataset containing webpage information from computer science departments at four different universities (Cornell, Texas. Washington, Wisconsin). WebKB's four sub-data sets each have two views, the categories are: 5,5,5,8, the feature dimensions of the four sub-data sets are: 1703/585, 1703/561, 1703/690, 1703/795.

4.2 Tricks

In order to make the MVPCFLF to achieve better results, it is necessary to adjust the four parameters α, β, γ ,and μ . Where α is uniformly set to $1/n_v$, the purpose of this setup is to enable each view to make full use of its own feature information, so that the extracted information is more compatible with the consistency and complementarity of multi-view data.

On the other hand, we set the parameters β , γ , μ according to different multi-view datasets to ensure optimal performance on different datasets. The value range of β is [10,10000], the range of γ is [1 ,100], and the range of μ is [0.01 ,100].

Another thing to note in the experiment configuration is that we set the latent factors K to 50 in MVPCFLF, and the sharing ratio η is adjusted to 0.3. The parameter settings of MVPCFLF can be seen in Table 1.

Table 1. Hyperparameters on each data sets

Data Sets	**MVPCFLF parameter setting**		
	β	γ	μ
Leaves	10	100	0.01
Reuters	1000	100	100
YaleFace	100	10	100
BBC	100	100	100
Cornell	1000	10	1
Texas	10000	100	10
Washington	10000	1	0.1
Wisconsin	10000	10	1

4.3 Compared Algorithms

We evaluated MVPCFLF in the classification task by comparing MVPCFLF with several baseline multi-view learning algorithms including GMVNMF [17], Multi-NMF, MVCC [18], DICS [19], and PSLF. In order to make our experiment comparison fair and equitable, we apply the parameter settings suggested in the corresponding multi-view learning algorithms, and select the latent factor K within the range of [5:5:20] with the stepping increment 5 for all NMF-based algorithms, as to PSLF we need additional settings. Before conducting the classification task, we need to obtain the latent representation from all multi-view learning algorithms and choose KNN (k=1) as the classifier, i.e., 1-nearest neighbor classifier.

GMVNMF is an unsupervised learning algorithm based on manifold learning and NMF，which employs the matrix factorization objective function by constructing a nearest neighbor graph to integrate local geometrical information of each view. Extracted features for GMVNMF take into account the correlation between different view data. We set λ_f , to 0.01 and μ to 10 according to the author’s suggestion.

MultiNMF is an NMF-based multi-view clustering algorithm that integrates information from different views. Here we refer to the parameter setup given by the author, and take the value of trade-off parameter λ in the range of [0.001, 1].

MVCC is a new multi-view clustering method based on conceptual factorization and local manifold regularization, which drives the common consensus representation of multiple views. Here we set the parameter α to 100 according to the requirements in the original paper, while the values of β and γ are in the range of [50, 1000].

DICS is a new robust multi-view learning algorithm that explores the discriminant and non-discriminant information existing in different views by joint non-negative matrix factorization. DICS can learn the potential shared and special subspaces of different views, and further extract discriminant and non-discriminant information from all subspaces to secure better classification performance. In this experiment, we set the range of parameters α and β to [0,1] and the parameter γ to 1.

In addition, PSLF is a semi-supervised multi-view learning algorithm based on NMF. PSLF obtains the shared latent representation factor by setting the sharing ratio parameter and combining the NMF method. The slogan for PSLF is not only to pay attention to the consistency but also to its complementarity information on multi-view data. In order to achieve the best experimental results, we will refer to the parameter setup given by the author. More specifically, the range of parameter β is in [1 10 100 1000 10000], the range of parameter γ is in [0.01 0.1 1 10 100], the sharing ratio η is set to 0.3, and the latent factors K is set to 50.

4.4 Result

All datasets were divided into training and test sets in a ratio of 0.7:0.3. For MVPCFLF and all comparison baseline algorithms, we need to first select the most appropriate hyperparameter on each dataset for each algorithm to achieve the best accuracy on each dataset. The main goal is to ensure that the latent representations we obtain are both consistent and complementary. In order to obtain stable prediction results, all our algorithms need to run ten times on each dataset, and use derived ten latent representation matrices and corresponding real labels as input of the 1-nearest neighbor classifier, then obtain the average of the accuracy and F1 scores.

The classification results for all eight data sets are summarized in Tables 2 and 3. As we can see, MVPCFLF achieved better accuracies and F1 scores than the other algorithms on the Reuters, Cornell, Texas, Washington datasets. In addition, although MVPCFLF did not achieve the best results in Leaves and Wisconsin, the gap with the corresponding best algorithm is small.

But on YaleFace dataset, MVPCFLF is not as good as DICS. In generally speaking, compared with DICS, MVPCFLF achieves good results on remaining five different datasets. When MVPCFLF is compared with PSLF and MVCC, MVPCFLF achieves good results on six different data sets.

The reason of the overall improvement comes from the following aspects: The base matrix $\mathbf{U}$ and the coefficient matrix $\mathbf{V}$ are partitioned on the original joint-NMF method. We introduce $\mathbf{V}_c^*$ to ensure that $\mathbf{V}_c^*$ can extract features from $\mathbf{V}_c^p$ in different views. We incorporate $\mathbf{V}_c^*$ into $\mathbf{V}_r$, making $\mathbf{V}_r$ as a new latent representation, so that $\mathbf{V}_r$ contains information from different views, which promotes classification performance on the KNN classifier.

Table 2. Accuracies and F1-score of different methods (1)

Method	**ACC(%)/F1-score(%)**			
	Leaves	Reuters	YaleFace	BBC
GNMF	95.0±0 96.1±0	40.8±1.2 41.3±1.3	50.0±2.5 51.4±2.6	38.0±1.5 32.1±3.6
MultiNMF	95.0±0 94.4±0.3	52.7±0.2 53.3±0.1	64.2±4.2 62.7±4.8	73.1±0.2 71.0±0.2
PSLF	97.93±2.8 98.6±1.9	74.9±1.8 75.1±1.7	86.5±3.0 86.3±3.1	86.3±1.7 85.4±1.8
MVCC	**100±0** **100±0**	54.4±1.9 54.4±2.0	33.3±6.9 33.1±6.8	**95.8±2.6** **93.7±5.8**
DICS	97.9±2.5 98.0±2.6	70.3±4.0 70.4±3.7	**89.1±3.2** **88.6±3.4**	90.2±2.4 89.1±2.6
MVPCFLF	96.9±2.8 97.9±1.9	**78.4±1.4** **78.4±3.7**	81.8±3.8 81.7±1.4	90.7±1.9 90.7±1.8

Table 3. Accuracies and F1-score of different methods (2)

Method	**ACC(%)/F1-score(%)**			
	Cornell	Texas	Washington	Wisconsin
GNMF	41.0±1.8 25.3±0.9	57.9±1.84 50.9±1.5	69.6±2.2 47.7±5.6	52.8±1.4 39.9±2.7
Multi-NMF	49.7±7.7 33.2±5.8	68.7±3.4 48.9±5.3	59.3±2.6 27.6±2.4	50.3±3.5 38.3±3.6
PSLF	58.3±4.7 51.5±4.8	66.1±5.5 50.0±8.5	60.9±3.5 39.0±4.9	67.8±3.7 45.2±3.7
MVCC	60.8±5.0 42.9±4.5	64.7±5.5 53.4±9.1	62.8±3.8 41.1±2.9	64.3±2.7 51.3±2.5
DICS	72.8±6.1 59.8±8.3	81.6±4.0 62.0±11.2	77.4±6.0 55.2±7.8	**85.1±4.5** **69.7±9.9**
MVPCFLF	**76.4±4.6** **73.2±5.5**	**83.7±3.7** **76.3±5.5**	**79.7±2.8** **61.4±5.7**	84.6±2.7 67.6±6.3

4.5 Parameter Sensitive Analysis

In our experiments, there are six parameters in the PSLF algorithm, the weight parameter α, the hyperparameters β, γ, and μ, the dimensions K of latent factor and the sharing ratio η. Among them, α, K, and η are the same on eight different data sets. The purpose of this setup is to be able to compare with PSLF more fairly. Below we select the Reuters dataset as example, the analyses on remaining datasets are similar. When we change one of the three parameters, the other two take the optimal value. Next, we compare the effects of hyperparameters β, γ, μ on the accuracies of the KNN classifier and the F1 score under the Reuters dataset, as shown in Fig.3. As we can see, when β is set to 1000, γ is set to 100, and μ is set to 100, MVPCFLF can perform well on Reuters dataset. In addition, MVPCFLF is more sensitive to changes in parameter γ, while MVPCFLF is relatively insensitive to changes in parameter β and μ.

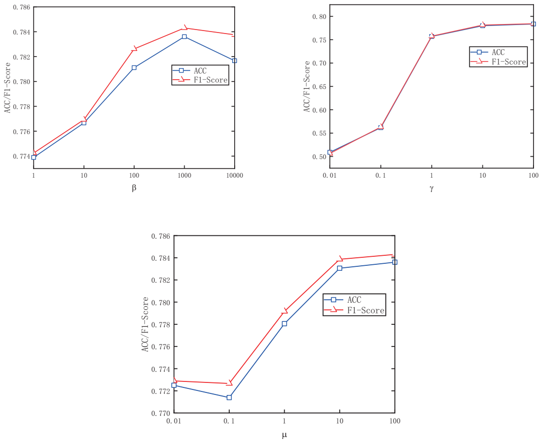

Fig.3. Parameter Sensitive analyses of MVPCFLF on Reuters datasets

4.6 Visualization of $\mathbf{V}_r$

t-SNE is a shorthand term for the t-distributed stochastic neighbor embedding algorithm. The main idea t-SNE is that when a point in a high-dimensional space is projected into a low-dimensional space, a good representation of a point in a low-dimensional space should maintain its neighborhood relationships in a high-dimensional space. In our visualization experiments, we use $\mathbf{V}_r$ as the feature matrix to classify. In order to show $\mathbf{V}_r$ more clearly, we use t-SNE to derive the 2D latent feature matrix and depict it through scatter plot. Figure 4 shows the visualization results of $\mathbf{V}_r$ on the BBC and Leaves datasets (left panel is result of BBC datasets, right panel is result of Leaves datasets). We can see from the above Fig.4. that the visualization results for $\mathbf{V}_r$ of MVPCFLF is easy to distinguish, so MVPCFLF can find effective latent representation.

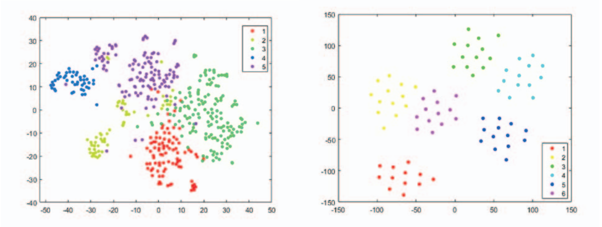

Fig.4. Visualization of $\mathbf{V}_r$ on BBC and Leaves

5 CONCLUSIONS AND FUTURE WORK

In this article, we present a new multi-view learning algorithm MVPCFLF. MVPCFLF aims to construct new latent representations for classification or clustering problems through the consistency and complementarity principles of multi-view. Concretely, we pre-block the matrix to make it possible to take consistent and complementary information into consideration in the base matrix $\mathbf{U}^p$, then we incorporate the coefficient matrix $\mathbf{V}_c^*$ to combine the coefficient matrices of multiple views, finally, we merge $\mathbf{V}_c^*$ into latent feature matrix $\mathbf{V}_r$. The experimental results of our proposed method on eight real-world datasets demonstrate that our algorithm is more efficient than the PSLF method, and has achieved promising results on four datasets compared to the DICS and MVCC algorithms. Of course, existing isolated cases, our algorithm does not outperform in some data sets, but it is on par with the effects of the most state-of-the-art algorithms.

In future work, we will try to compare the submatrix $\mathbf{V}_s^p$ of each view to make the complementarity between the different views more obvious. It is possible to achieve better results on multi-view data. In addition, how to construct the different weighting parameters for different views and how to select the shared child data for multi-view data will be a new topic.

REFERENCES

[1] Zhao J, Xie X, Xu X, et al. Multi-view learning overview: Recent progress and new challenges. Information Fusion, 2017, 38: 43-54.

[2] A. Blum, T. Mitchell, Combining labeled and unlabeled data with co-training, Proceeding of the 11th Annual Conference on Computational Learning Theory, 1998, pp. 92–100.

[3] K. Nigam, R. Ghani, Analyzing the effectiveness and applicability of co-training, Proceedings of the 9th International Conference on Information and Knowledge Management, 2000, pp. 86-93.

[4] Muslea I, Minton S, Knoblock C A. Active learning with multiple views. Journal of Artificial Intelligence Research, 2006, 27, pp. 203-233.

[5] Sun S, Jin F. Robust co-training. International Journal of Pattern Recognition and Artificial Intelligence, 2011, 25(07), pp. 1113-1126.

[6] Sun S, Shawe-Taylor J. Sparse semi-supervised learning using conjugate functions. Journal of Machine Learning Research, 2010, 11(Sep), pp. 2423-2455.

[7] Xie X, Sun S. Multi-view twin support vector machines. Intelligent Data Analysis, 2015, 19(4), pp. 701-712.

[8] Sun S. Multi-view Laplacian support vector machines. Proceedings of the International Conference on Advanced Data Mining and Applications. Springer, Berlin, Heidelberg, 2011, pp. 209-222.

[9] Xie X, Sun S. Multi-view Laplacian twin support vector machines. Applied intelligence, 2014, 41(4), 1059-1068.

[10] S. Sun, G. Chao, Multi-view maximum entropy discrimination, Proceedings of the 23rd International Joint Conference on Artificial Intelligence, 2013, pp. 1706-1712.

[11] G. Chao, S. Sun, Alternative multi-view maximum entropy discrimination, IEEE Trans. Neural Netw. Learn. Syst. 27 (2016) 1445-1456.

[12] L. Mao, S. Sun, Soft margin consistency based scalable multi-view maximum entropy discrimination, Proceedings of the 25th International Joint Conference on Artificial Intelligence, 2016, pp. 1839-1845.

[13] G. Chao, S. Sun, Consensus and complementarity based maximum entropy dis-crimination for multi-view classification, Information Sciences, 2016, 367, pp. 296-310.

[14] Lee D D, Seung H S. Learning the parts of objects by non-negative matrix factorization. Nature, 1999, 401(6755) 788-791

[15] Liu J, Wang C, Gao J, et al. Multi-view clustering via joint nonnegative matrix factorization, Proceedings of the 2013 SIAM International Conference on Data Mining. Society for Industrial and Applied Mathematics, 2013, pp. 252-260

[16] Liu, J., Jiang, Y., Li, Z., Zhou, Z. H., & Lu, H. (2014). Partially shared latent factor learning with multi-view data. IEEE transactions on neural networks and learning systems, 26(6), 1233-1246.

[17] Wang Z, Kong X, Fu H, et al. Feature extraction via multi-view non-negative matrix factorization with local graph regularization, 2015 IEEE International conference on image processing (ICIP). IEEE, 2015, pp. 3500-3504.

[18] Wang H, Yang Y, Li T. Multi-view clustering via concept factorization with local manifold regularization, 2016 IEEE 16th International Conference on Data Mining (ICDM). IEEE, 2016, pp. 1245-1250.

[19] Zhang Z, Qin Z, Li P, et al. Multi-view discriminative learning via joint non-negative matrix factorization, International Conference on Database Systems for Advanced Applications. Springer, Cham, 2018, pp. 542-557.

[20] Chi E C, Lange K. Splitting methods for convex clustering. Journal of Computational and Graphical Statistics, 2015, 24(4), pp. 994-1013