C²IMUFS: Complementary and Consensus Learning-Based Incomplete Multi-View Unsupervised Feature Selection

Yanyong Huang, Zongxin Shen, Yuxin Cai, Xiwen Yi, Dongjie Wang, Fengmao Lv, and Tianrui Li, Senior Member, IEEE

Abstract—Multi-view unsupervised feature selection (MUF S) has been demonstrated as an effective technique to reduce the dimensionality of multi-view unlabeled data. The existing methods assume that all of views are complete. However, multi-view data are usually incomplete, i.e., a part of instances are presented on some views but not all views. Besides, learning the complete similarity graph, as an important promising technology in existing MUS methods, cannot achieve due to the missing views. In this paper, we propose a complementary and consensus learning-based incomplete multi-view unsupervised feature selection method (C²IMUFS) to address the aforementioned issues. Concretely, C²IMUFS integrates feature selection into an extended weighted non-negative matrix factorization model equipped with adaptive learning of view-weights and a sparse $\ell_2,q$-norm, which can offer better adaptability and flexibility. By the sparse linear combinations of multiple similarity matrices derived from different views, a complementary learning-guided similarity matrix reconstruction model is presented to obtain the complete similarity graph in each view. Furthermore, C²IMUFS learns a consensus clustering indicator matrix across different views and embeds it into a spectral graph term to preserve the local geometric structure. Comprehensive experimental results on real-world datasets demonstrate the effectiveness of C²IMUFS compared with state-of-the-art methods.

Index Terms—Complementary and consensus information, incomplete multi-view data, unsupervised feature selection, weighted non-negative matrix factorization.

I. INTRODUCTION

MULTI-VIEW data widely exist in many real applications, where the same instances are described by multiple heterogeneous feature sets from different aspects. For example, in the video classification task, each video can be depicted by the visual features, audio inputs and text representations [1]. In the credit risk assessment task, Bad Credit History, Loan Description and Extra Credit Information are often used to describe each loan application from different perspectives [2]. In these applications, different views of data are usually represented in high dimensional feature space containing redundant and noisy features, which will result in the problem of “curse of dimensionality” and degrade the performance of clustering or classification tasks [3]. In addition, it is difficult or even infeasible to obtain a large amount of labeled data in practice. Therefore, how to reduce the dimensionality of multi-view unlabeled data for the improvement of performance of the subsequent tasks has become an urgent problem to be solved in various practical applications.

Multi-view unsupervised feature selection (MUS) provides an effective solution to address the above problem by choosing a compact subset of representative features from the original feature space. In recent years, a variety of MUS approaches have been developed, which can be mainly classified into two categories. In the first category are the traditional single-view unsupervised feature selection methods, which first concatenate all features derived from different views and then take the combined features as input straightforward [4], [5], [6], [7]. There are some typical single-view based unsupervised feature selection methods, like Laplacian Score (LPScore) [4], Robust Spectral Feature Selection (RSFS) [5] and Embedded Graph learning and Constraint-based Feature Selection (EGCFS) [8]. These methods usually characterize the local geometric structure of data by constructing various graphs and select the top ranked features according to their importance. Although these methods can help perform better feature selection to a certain extent, they treat each view independently and do not consider the underlying correlations among different views. Rather than combining features of different views, the second category of MUS methods directly construct models from multi-view data for feature selection [9], [10], [11]. Typical methods in this category include Adaptive Collaborative Similarity Learning (ACSL) [12].

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
adaptively learns a collaborative similarity structure of data by utilizing of complementary information across different views and integrates it into a sparse regression model for feature selection. Besides, Tan et al. have proposed the method referred to as Cross-view Local structure Preserved Diversity and Consensus Learning (CVLP-DCL), which employs a cross-view similarity graph learning to preserve the local geometric structure of data and learns the consensus and diversity label spaces to select features [13]. Furthermore, Bai et al. have developed a unified learning framework for multi-view unsupervised feature selection, named Nonnegative Structured Graph Learning (NSGL) [14]. It learns a common adaptive structured graph among different views and embeds it into the sparse linear regression model to learn the consensus pseudo labels and perform feature selection simultaneously. Although these proposed methods have been demonstrated promising performance in feature selection, they have a common assumption, i.e., each instance appears in all views. In real-world applications, it is more often the case that a part of instances are presented on some views but not all views [15], [16]. For example, in web image retrieval, the visual information, links, and textual tags of each image can be regarded as three views, while some images may lack the link or text view [17]. Hence, the above-mentioned methods could not be used in the incomplete multi-view data directly. There exists the following two key challenges in feature selection on incomplete multi-view data: (i) How to obtain the complete similarity graph from incomplete multi-view data? Learning or constructing the complete similarity graph of each view is an important promising technology in MIFS. The existing methods can not obtain the complete graphs since some views are missing for part of instances in the incomplete multi-view scenario. (ii) How to simultaneously use the consensus and complementary information among different views to improve the performance of MIFS? Most of existing MIFS methods consider using either consensus or complementary information, and just fuse these multi-view information with equal weights ignoring the difference of different views. As a result, these will limit their performance in incomplete multi-view unsupervised feature selection.

In order to address the aforementioned issues, we present a novel MIFS method for multi-view unlabeled data with incomplete views, named as Complementary and Consensus Learning-based Incomplete Multi-view Unsupervised Feature Selection (C2IMUFS). Specifically, our proposed method C2IMUFS embeds the process of feature selection into an extended weighted non-negative matrix factorization model (WNMF), which can automatically determine the importance of different views and consider the difference between the missing and non-missing instances. To obtain the complete similarity graph in the incomplete-view scenario, we propose a complementary learning-guided similarity matrix reconstruction model. It can adaptively learn the similarity matrix by sparse linear combinations of multiple similarity matrices derived from other views. Meanwhile, C2IMUFS learns a common adaptive structured indication matrix among multiple views and integrates it into a spectral graph term to preserve the common geometric structure of different views. Furthermore, a sparse constraint with $\ell_2, p$-norm is imposed on the feature selection matrix, which can offer better flexibility and satisfy different sparsity requirements. Finally, an alternative iterative optimization algorithm is developed to solve the proposed model and comprehensive experiments are carried out to demonstrate the effectiveness of the proposed method by comparing with several state-of-the-art single-view and multi-view unsupervised feature selection methods.

To sum up, the major contributions of this paper are summarized as follows: (i) To best of our knowledge, this is the first study to address the problem of unsupervised feature selection on incomplete multi-view data by simultaneously employing the complementary and consensus information across different views. (ii) We propose a novel model that incorporates an extended WNMF module and a complementary and consensus learning-based module for incomplete multi-view unsupervised feature selection. It can learn the complete similarity graph of each view by using the complementary graph structures acquired from other views and integrate different views together via learning a consensus indicator matrix with adaptive view weights. (iii) We develop an efficient alternative optimization algorithm to solve the proposed model and extensive experimental results on eight real-world datasets show the superiority of the proposed method compared with the state-of-the-art methods.

The remainder of this paper is organized as follows. In Section II, we briefly review related work about MIFS. In Section III, we introduce in detail the proposed method C2IMUFS. Section IV provides an effective solution to this method and the convergence and complexity of the proposed algorithm are discussed in the following Section V. In Section VI, a series of experiments are carried out to demonstrate the effectiveness of the proposed method. And finally, the conclusions are presented in Section VII.

II. RELATED WORK

In this section, we briefly review some representative works in terms of single-view-based unsupervised feature selection and multi-view unsupervised feature selection methods, respectively.

A. Single-View Unsupervised Feature Selection Methods

Single-view unsupervised feature selection methods are roughly categorized into three classes, i.e., filter-based, wrapper-based and embedded-based methods. The first kind of methods select the representative features according to certain evaluation criterion measuring the importance of features. LapScore [4] selects features in terms of the metric Laplacian Score, which describes the capability of features in preserving locality. Zhao et al. employed the spectral graph theory to identify the redundant features [18]. Filter-based methods ignore the relation of selected features and are independent of any learning algorithms, which will result in that they could not obtain the satisfied performance. The second kind of methods usually use a learning algorithm to evaluate the selected features until obtaining the most suitable features. Maldonado et al. developed a wrapper-based feature selection method by employing support vector machines with kernel functions [19]. The main drawback of
wrapper methods is their expensive computational cost arising by trying to find the features with the highest performance. The last kind of methods embed the procedure of feature selection into the learning phase of a specific algorithm, which can obtain the important features with low computational cost. RSFS [5] constructs a graph Laplacian to capture the local geometric structure of data and incorporates the feature selection procedure into a robust spectral regression model. Yang et al. proposed a unified framework for unsupervised feature selection by combining the discriminant analysis and $\ell_2$-norm minimization [7]. EGCFs [8] embeds the construction of the similarity graph into the procedure of maximizing the between-class scatter matrix, which can adaptively learn the graph structure and select discriminative features simultaneously. While using these single-view unsupervised feature selection methods to select features from multi-view data, they directly concatenate all features of different views together and ignore the latent correlations among multiple views. This will result in the inferior performance.

B. Multi-View Unsupervised Feature Selection Methods

Different from the above-mentioned methods, multi-view unsupervised feature selection methods deal with multi-view data directly and explore the underlying information among different views to improve the performance. Shao et al. presented an online unsupervised multi-view feature selection method (OMVFS), which employs the non-negative matrix factorization model to simultaneously learn the feature selection matrices with sparse constraint and a consensus cluster indicator matrix across different views [20]. ACSL [12] integrates the adaptive collaborative graph learning and feature selection into a unified framework by employing the complementary information among different views and a sparse regression model. CvLP-DCL [13] learns a cross-view similarity graph equipped with a matrix-induced regularization term by linearly combining different view-specific similarity graphs. Besides, in CvLP-DCL, a common and view-specific label spaces are incorporated into a regression model for feature selection. Tang et al. embedded feature selection into a clustering process based on non-negative matrix factorization, which can both obtain a common clustering indicator matrix and these latent feature matrices of different views [21]. In NSGL [14], the common similarity graph learning among multiple views with adaptive view weights and pseudo label learning are combined into a linear regression model with $\ell_{2,1}$-norm regularization for feature selection. Guan et al. constructed the consistent and complementary correlations between the textual and visual modalities and embedded them into two parallel graph convolution networks for outfit compatibility modeling [22]. This work mainly focus on the fashion compatibility modeling. However, the aforementioned MUFs methods are confronted with three issues. First, they all assume that all of views are complete, which disable them to be used in the feature selection of incomplete multi-view data directly. Second, these methods can not obtain the complete similarity graph since the problem of missing views. This will result in the inferior performance. Last, they did not take the complementary and consensus information among different views into consideration simultaneously. These will limit their performance.

III. COMPLEMENTARY AND CONSENSUS LEARNING-BASED INCOMPLETE MULTI-VIEW UPS

In this section, we first summarize some notations and definitions used throughout the paper. Then, we introduce the proposed method C2IMUFs for incomplete multi-view data, which consists of two parts. The first part employs an extended WNMF model equipped with $\ell_{2,p}$-norm regularization term and adaptive view-weight learning to obtain the feature selection matrices and clustering indicator matrices. In the second part, complementary and consistency learning strategies are integrated into the extended model to reconstruct the complete similarity matrix of each view and learn the common clustering indicator matrix. The framework of C2IMUFs is shown in Fig. 1.

A. Notations

In this paper, matrices and vectors are written in boldface uppercase letters and boldface lowercase letters, respectively. Given a matrix $\mathbf{W} \in \mathbb{R}^{m \times n}$, $\mathbf{w}_i$ and $\mathbf{w}_j$ respectively indicate the $i$th row and $j$th column of matrix $\mathbf{W}$. The $(i, j)$th entry of $\mathbf{W}$ is denoted by $\mathbf{W}_{ij}$. Let $\text{Tr} (\mathbf{W})$, $\mathbf{W}^T$ and $\text{Rank} (\mathbf{W})$ respectively denote the trace of $\mathbf{W}$, the transpose of $\mathbf{W}$ and the rank of $\mathbf{W}$. The Frobenius norm of matrix $\mathbf{W}$ is defined as $\| \mathbf{W} \|_F = \sqrt{\sum_{i=1}^m \sum_{j=1}^n |\mathbf{W}_{ij}|^2}$. The $\ell_{2,p}$-norm is defined as $\| \mathbf{W} \|_{2,p} = \left( \sum_{i=1}^m \| \mathbf{w}_i \|_2^p \right)^{1/p}$, $0 < p \leq 1$ where $\| \mathbf{w}_i \|_2$ denotes the $\ell_2$-norm of vector $\mathbf{w}_i$. Especially, while $p = 1$, we have the $\ell_{2,1}$-norm: $\| \mathbf{W} \|_{2,1} = \sum_{i=1}^m \sqrt{\sum_{j=1}^n |\mathbf{W}_{ij}|^2} = \sum_{i=1}^m \| \mathbf{w}_i \|_2$. We use $\mathbf{I}$ to denote an identity matrix and use $\mathbf{1}$ to indicate a column vector with all elements being one.

Let $\mathbf{X} = \left\{ \mathbf{X}^{(v)}, v = 1, \ldots, l \right\}$ denote a multi-view data with $l$ views, where $\mathbf{X}^{(v)} \in \mathbb{R}^{d_v \times N}$ is the non-negative data matrix in view $v$ with $N$ instances and $d_v$ features. In real applications, some data instances may be missing in certain views. To describe the view-missing information, an indicator matrix $\mathbf{A} = (\mathbf{A}_{ij})_{N \times l}$ is defined as follows:

$$\mathbf{A}_{ij} = \begin{cases} 1 & \text{if the } i \text{th instance is in the } j \text{th view;} \\ 0 & \text{otherwise.} \end{cases}$$  

Then, it is easy to know $\sum_{i=1}^N \mathbf{A}_{ij} < N$. Our goal is to select $h$ most discriminative features from the incomplete multi-view data $\mathbf{X}$.

B. WNMF-Based Incomplete Multi-View Unsupervised Feature Selection

The traditional WNMF model [23] has been widely used to deal with incomplete data, which factorizes a data matrix into two non-negative matrices and introduces a weight matrix to describe the missing information. It can be formulated as follows:

$$\min_{\mathbf{U}, \mathbf{V}} (\mathbf{X} - \mathbf{U} \mathbf{V}^T) \mathbf{W} \|_F^2,$$  

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
where $X$ is an incomplete data matrix, $W$ is a weight matrix, and the weighted matrix entry $W_{ij}$ is 0, if the entry $X_{ij}$ of matrix $X$ is missing, otherwise $W_{ij} = 1$. To select features from the incomplete multi-view data, a direct method is that using (2) to approximately factorize each missing view data matrix. However, this way is equivalent to fill the missing values with 0 in different views. A more reasonable method is using the weighted averages to fill in the missing values and the weight is determined by the available instances in each view. To this end, a WNMF-based incomplete multi-view unsupervised feature selection is proposed as follows:

$$
\min_{\{U^{(v)}, V^{(v)}\}} \sum_{v=1}^{l} \|\{X^{(v)} - U^{(v)}V^{(v)T}\}W^{(v)}\|_F^2 + \lambda \|U^{(v)}\|_p
$$

s.t. $V^{(v)} \geq 0$, $V^{(v)T}V^{(v)} = I$, $U^{(v)} \geq 0$, $v = 1, 2, \ldots, l$,  

where $X^{(v)}$ is the $v$th view data matrix, $U^{(v)} \in \mathbb{R}^{d_x \times c}$ and $V^{(v)} \in \mathbb{R}^{N \times c}$ respectively denote the $v$th view feature selection matrix and clustering indicator matrix, $\|U^{(v)}\|_p$ is the regularization notation of $U^{(v)}$, $\lambda$ is the regularization parameter, and $c$ is the number of clusters. In (3), $W^{(v)} \in \mathbb{R}^{N \times N}$ is a diagonal weight matrix in the $v$th view, whose diagonal entry $W_{jj}^{(v)}$ is defined as follows:

$$
W_{jj}^{(v)} = \begin{cases} 
1 & \text{if instance } j \text{ exists in view } v; \\
\frac{1}{\sum_{i=1}^{N} A_{ij}} & \text{otherwise.}
\end{cases}
$$

(4) shows that the weights of non-missing instance and missing instance are 1 and the ratio of all available instances in the $v$th view, respectively. Hence, it can better measure the quality of available information in each view and give a more reasonable weighted average to impute the missing value.

To achieve feature selection, $\ell_{2,1}$-norm is usually used to impose row sparsity on the $v$th feature selection matrix $U^{(v)}$ in (3) [20], [24]. Based on $\ell_{2,1}$-norm, we can select the top ranked features according to the descending order of $\ell_{2,1}$-norm of the rows in $U^{(v)}$. However, many studies have demonstrated that $\ell_{2,p}$-norm ($0 < p < 1$) can give better sparsity in comparison with $\ell_{2,1}$-norm [25], [26], [27]. Hence, in (3), we impose $\ell_{2,p}$-norm regularization on $U^{(v)}$ to ensure its row-sparsity. Then, we can provide more flexible $p$ to satisfy the different sparsity requirements. Furthermore, considering that different views have different contributions to the final feature selection task, we use an adaptive weight determination strategy to assign the view-weight $\alpha^{(v)} (v = 1, 2, \ldots, l)$ for each view. Then, we can obtain the following novel incomplete multi-view unsupervised feature selection method based on WNMF:

$$
\min_{\{U^{(v)}, V^{(v)}\}, \{\alpha^{(v)}\}} \sum_{v=1}^{l} \sum_{u=1}^{l} \alpha^{(v)} \gamma \|\{X^{(v)} - U^{(v)}V^{(v)T}\}W^{(v)}\|_F^2 + \lambda \|U^{(v)}\|_{2,p}
$$

s.t. $V^{(v)} \geq 0$, $V^{(v)T}V^{(v)} = I$, $U^{(v)} \geq 0$, $0 \leq \alpha^{(v)} \leq 1$,

$$
\sum_{v=1}^{l} \alpha^{(v)} = 1, v = 1, 2, \ldots, l,
$$

where $\gamma$ is a regularization parameter.

In (5), the proposed model can select features from the incomplete multi-view unlabeled data. Nevertheless, it is limited by the inability to fully explore the inherent information of multi-view data. In what follows, we will exploit the complementary and
consistent information of multi-view data and integrate them into our model to improve its performance.

C. Complementary and Consensus Learning

Previous studies have confirmed that preserving local manifold structure by constructing a nearest neighbor graph is beneficial to improve the performance of unsupervised feature selection [28], [29], [30]. Hence, for the construction of graph in incomplete multi-view data, the key point is to construct the similarity matrix in each view. However, it is hard to obtain the reliable similarity matrix in each view since some instances are missing in certain views. Considering that there exists complementary information across different views [31], [32] and the similarity relation between two instances exists in at least one view, we can construct the similarity matrix by using the between-view interaction similarity information. Based on this, we propose the following complementary learning-guided similarity matrix reconstruction model, which can adaptively learn the similarity matrix by means of the similarity information derived from other views.

\[
\begin{align*}
\min_{\{s(v)\}, R(v)} & \sum_{v \in V} \|s(v) - \sum_{i=1, i \neq v} \|R(i)\|_F^2 + \|R\|_F^2 \\
\text{s.t.} & \ 0 \leq s(v) \leq 1, s(v)^T 1 = 1, S(v) = 0, \\
& 0 \leq R_{iv} \leq 1, \sum_{i=1, i \neq v} R_{iv} = 1, R_{vv} = 0,
\end{align*}
\]

where \(S(v) \in \mathbb{R}^{N \times N}\) is the similarity matrix of view \(v\) and \(R(v) \in \mathbb{R}^{l \times l}\) can be deemed as the regression coefficient matrix. The Frobenius norm imposed on \(R\) is beneficial to improve robustness to noise [33]. The coefficient matrix \(R\) measures the similarity degrees among the reconstructed and existent similarity matrices and is learned from data adaptively. It can reflect the difference of missing and existing views’ data distributions, namely, a larger divergence between the probability distributions of the missing and existing view data sources corresponds to a lower similarity value, and a smaller divergence between them corresponds to a higher similarity value. Then, from (6), we can see that the reconstructed similarity matrix of missing view is mainly affected by that of exiting views with high similarity degrees of data distributions, which will not result in a large deviation due to the inconsistency of distributions. Compared with using the predefined similarity matrix in incomplete multi-view data, (6) can obtain the more reliable similarity matrix by learning the complementary information from other views adaptively.

In addition, in order to exploit the consensus information across different views, we assume that each view shares the same clustering indicator matrix \(V\), i.e., \(V^{(1)} = V^{(2)} = \ldots = V^{(l)} = V\) in (5). Moreover, according to the spectral graph theory [34], [35], if two data points have high similarity in the intrinsic graph of data, then their corresponding clustering labels also have high similarity. Then, we can add the following regularization term to preserve the local geometric structure in each view.

\[
\min_{V} \sum_{v \in V} \text{Tr}(V^T L_s(v)V) \\
\text{s.t.} \ V \geq 0, V^T V = I,
\]

where \(L_s(v) = D_s(v) - S(v)\) denotes the Laplacian matrix and \(D_s(v) = (i,j) = \sum_{j=1}^N S(v)\).

By combining Eqs. (5), (6) and (7) together, the proposed incomplete multi-view unsupervised feature selection method (C²IMUFS) is summarized as follows:

\[
\min_{\Omega} \sum_{v \in V} (\alpha(v)^γ \|\|X(v) - U(v)V^T W(v)\|^2_F + \lambda_1 \|U(v)\|^2_F + \lambda_2 \|V\|^2_F) \\
+ \beta_1(\text{Tr}(V^T L_s(v)V) + \|S(v) - \sum_{i=1, i \neq v} \|R(i)\|_F^2 + \|R\|_F^2))
\]

\[
\text{s.t.} \ V \geq 0, V^T V = I, U(v) \geq 0, 0 \leq \alpha(v) \leq 1, \sum_{v \in V} \alpha(v) = 1 \\
0 \leq S(v) \leq 1, S(v)^T 1 = 1, S(v) = 0, 0 \leq R_{iv} \leq 1 \\
\sum_{i=1, i \neq v} R_{iv} = 1, R_{vv} = 0, v = 1, 2, \ldots, l,
\]

where \(\Omega = \{U(v), S(v), \alpha(v), V, R|v = 1, 2, \ldots, l\}\) and \(\beta \geq 0\) is a trade-off hyper-parameter.

As can be seen from (8), the proposed method C²IMUFS has two advantages: On one hand, the WNMF-based multi-view unsupervised feature selection model equipped with the adaptive view-weight learning and the \(l_2,p\)-norm regularization can better employ the available information of the incomplete data and automatically determine the view importance. It also provides a more flexible way to select informative features by adjusting \(p\) to control the sparsity according to the data. On the other hand, C²IMUFS can obtain the more precise similarity-induced graph matrix in each view and preserve the reliable local geometric structure by simultaneously exploiting the complementary and consensus information among different views.

IV. OPTIMIZATION AND ALGORITHM

Since the objective function in (8) is not convex to all variables simultaneously, we propose an alternative iterative algorithm to solve the optimization problem, namely, optimizing the objective function with regards to one variable while others fixed.

A. Update \(U(v)\) by Fixing Other Variables

When fixing other variables and removing the irrelevant terms, \(U(v)\) can be updated by solving the following problem:

\[
\min_{U(v)} (\alpha(v)^γ \|\|X(v) - U(v)V^T W(v)\|^2_F + \lambda_1 \|U(v)\|^2_F) \\
\text{s.t.} \ U(v) \geq 0.
\]
By introducing a Lagrange multiplier $\Psi$ for the non-negative constraint $U^{(v)} \geq 0$, we have the following Lagrange function:

$$L(U^{(v)}) = (\alpha^{(v)})^\gamma ||(X^{(v)} - U^{(v)}V^T)W^{(v)}||_F^2 + \lambda ||U^{(v)}||_{2,p}^p + Tr(\Psi^T U^{(v)}).$$

In order to obtain the derivative of $||U^{(v)}||_{2,p}^p$ w.r.t. $U^{(v)}$, we introduce a lemma as follows:

**Lemma 1:**

$$\lim_{u_i \to u_i^+} \frac{\partial ||U^{(v)}||_{2,p}^p}{\partial u_i} = \frac{p u_i^p}{||u_i^+||_2^p} \frac{\partial ||u_i^+||_2^p}{\partial u_i},$$

where $D$ is a diagonal matrix and the corresponding diagonal entry $D_{ii} = \frac{p ||u_i^+||_2^p}{2\sqrt{||u_i^+||_2^p}}$ and $\epsilon \to 0$ is a small value that prevents the denominator from being 0.

According to Lemma 1, the derivative of (10) with respect to $U^{(v)}$ is equivalent to the derivative of the following function:

$$L(U^{(v)}) = (\alpha^{(v)})^\gamma ||(X^{(v)} - U^{(v)}V^T)W^{(v)}||_F^2 + \lambda Tr(U^{(v)}D^{(v)}U^{(v)}) + Tr(\Psi^T U^{(v)}).$$

Then, by taking the partial derivative of $L(U^{(v)})$ w.r.t. $U^{(v)}$ in (13), we have

$$\frac{\partial L(U^{(v)})}{\partial U^{(v)}} = (\alpha^{(v)})^\gamma [2(U^{(v)}V^T W^{(v)})V - X^{(v)}W^{(v)}V^T W^{(v)} + 2\lambda D^{(v)}U^{(v)}] + \Psi.$$  

According to the Karush-Kuhn-Tucker (KKT) complementary condition $\Psi_{ij} U^{(v)}_{ij} = 0$ [36], $U^{(v)}$ can be updated as follows:

$$U^{(v)}_{ij} \leftarrow U^{(v)}_{ij} \sqrt{\frac{(X^{(v)}W^{(v)}W^{(v)}V)_{ij}}{(U^{(v)}V^T W^{(v)}W^{(v)}V)^{ij} + \lambda \sum_{l \neq i} D^{(v)}U^{(v)}_{lj}}}.  \tag{15}$$

While updating $U^{(v)}$, we set $\epsilon$ as a relatively small value, $\epsilon = 10^{-6}$ in our experiments.

**B. Update V by Fixing Other Variables**

When other variables are fixed, the problem of solving $V$ is degraded to minimize the following optimization problem by removing irrelevant terms:

$$\min_{V} \sum_{v=1}^{l} (\alpha^{(v)})^\gamma ||(X^{(v)} - U^{(v)}V^T)W^{(v)}||_F^2 + \beta Tr(V^T L_{\alpha^{(v)}} V)$$

s.t. $V \geq 0, V^T V = I$.  \tag{16}

Then, we can construct the following Lagrange function by introducing the Lagrange multipliers $\xi$ and $\Theta$ for the orthogonal constraint $V^T V = I$ and the non-negative constraint $V \geq 0$, respectively.

$$L(V) = \sum_{v=1}^{l} (\alpha^{(v)})^\gamma ||(X^{(v)} - U^{(v)}V^T)W^{(v)}||_F^2 + \beta Tr(V^T L_{\alpha^{(v)}} V) + \xi ||V^T V - I||_F^2 + Tr(\Theta^T V).$$

To guarantee the orthogonality of $V$, $\xi$ is set to a large constant ($\xi = 10^7$ in our experiments).

By taking the derivative of $L(V)$ w.r.t. $V$, we have

$$\frac{\partial L(V)}{\partial V} = \sum_{v=1}^{l} (\alpha^{(v)})^\gamma [2(U^{(v)}V^T W^{(v)})U^{(v)} + 2\lambda D^{(v)}V - 2W^{(v)}U^{(v)}V^T W^{(v)} + \xi (4VV^T V - 4V) + \Theta].$$

Based on the KKT complementary condition $\Theta_{ij} V_{ij} = 0$, we obtain the updating rule for $V$ as follows:

$$V_{ij} \leftarrow V_{ij} \sqrt{\frac{(E^{(v)} + 2\xi V)^{ij}}{(Q^{(v)} + 2\xi V^T V)^{ij}}}, \tag{19}$$

where $Q^{(v)} = \sum_{i=1}^{l} (\alpha^{(v)})^\gamma [W^{(v)}U^{(v)} + \beta D^{(v)} V]$, $E^{(v)} = \sum_{v=1}^{l} (\alpha^{(v)})^\gamma (W^{(v)}U^{(v)}V^T U^{(v)} + \beta D^{(v)} V)$.  \tag{18}

**C. Update S^{(v)} by Fixing Other Variables**

With other variables fixed, we can update $S^{(v)}$ by solving the following problem:

$$\min_{S^{(v)}} \sum_{i=1}^{l} (\alpha^{(v)})^\gamma Tr(V^T L_{\alpha^{(v)}} V) + ||S^{(v)} - \sum_{i=1}^{l} S^{(v)}_{ii} R_{i}||_F^2, \quad s.t. \ 0 \leq S^{(v)} \leq 1, S^{(v)} T 1 = 1, S^{(v)}_{ii} = 0. \tag{20}$$

We first introduce some notations for the convenience of following discussion. Let $T = \{S^{(v)} | 0 \leq S^{(v)} \leq 1, S^{(v)} T 1 = 1, S^{(v)}_{ii} = 0\}$, $a^{(v)} = (\alpha^{(v)})^\gamma$, $N^{(i)} = S^{(i)} - \sum_{j=1, j \neq i}^{l} S^{(j)} R_{ij}$, $B^{(v)} = \sum_{i=1, i \neq v}^{l} S^{(v)} R_{iv}$, and $H_{ij} = ||v^i - v^j||_2$. Then, (20) can be rewritten as follows:

$$\min_{T} \frac{a^{(v)}}{2} \sum_{i,j=1}^{l} ||v^i - v^j||_2^2 S_{ij} + a^{(v)} ||S^{(v)} - \sum_{i=1, i \neq v}^{l} S^{(v)}_{ii} R_{i}||_F^2$$

$$+ \sum_{k=1, k \neq v}^{l} a^{(k)} ||R_{ik} S^{(v)} - \sum_{j=1, j \neq v, j \neq k}^{l} S^{(v)} R_{jk}||_F^2$$

$$\Rightarrow \min_{T} \frac{a^{(v)}}{2} \sum_{i,j=1}^{l} H_{ij} S_{ij} + \sum_{k=1, k \neq v}^{l} a^{(k)} ||R_{ik} S^{(v)} - \sum_{j=1, j \neq v, j \neq k}^{l} S^{(v)} R_{jk}||_F^2 + a^{(v)} ||S^{(v)} - B^{(v)}||_F^2.$$
\[ \min \sum_{i,j=1}^{N} \left[ \frac{1}{2} H_{ij} S_{ij}^{(v)} + \sum_{k=1,k \neq v}^{l} a^{(k)}(R_{vk})^2 \left( S_{ij}^{(v)} - \frac{N_{ij}^{(k)}}{R_{vk}} \right) \right]^2 + a^{(v)} \left( S_{ij}^{(v)} - B_{ij}^{(v)} \right)^2 \]
\[ \min \sum_{i,j=1}^{N} \left[ S_{ij}^{(v)} + \sum_{k=1,k \neq v}^{l} a^{(k)}N_{ij}^{(k)} + a^{(v)}B_{ij}^{(v)} - \frac{1}{4} H_{ij} \right] \]

Since \( N^{(k)} \) and \( B^{(v)} \) are independent of \( S^{(v)} \), we can formulate (21) in an equivalent form as follows:
\[ \min \sum_{i,j=1}^{N} (S_{ij}^{(v)} - P_{ij}^{(v)})^2, \]
where \( P_{ij}^{(v)} = \frac{Q_{ij}^{(v)}}{\sum_{k=1,k \neq v}^{l} a^{(k)}(R_{vk})^2 + a^{(v)}}, \) and \( Q_{ij}^{(v)} = \sum_{k=1,k \neq v}^{l} a^{(k)}R_{vk}N_{ij}^{(k)} + a^{(v)}(B_{ij}^{(v)} - \frac{1}{4} H_{ij}) \). We can observe that all columns of \( S^{(v)} \) are independent in problem (21).

Then, we can obtain the following optimal solution of problem (20) according to [37], [38]:
\[ S_{ij}^{(v)} = \begin{cases} P_{ij}^{(v)} + \delta_i & \text{if } i \neq j; \\ 0 & \text{if } i = j, \end{cases} \]

where \( \delta_i = (1 - \sum_{j=1,j \neq i}^{n} P_{ij}^{(v)})/(n - 1) \) and the function \((Z)^+ = \max(0, Z)\) indicates a max function to make sure \( Z \) is non-negative.

D. Update \( R \) by Fixing Other Variables

When other variables are fixed and the irrelevant terms are removed, updating \( R \) is equal to solve the following problem:
\[ \min R \sum_{v=1}^{l} \|S^{(v)} - S^{(v)}(R_{iv})^2 + \|R\|^2 \]
\[ s.t. 0 \leq R_{iv} \leq 1, \sum_{i=1,i \neq v}^{l} R_{iv} = 1, R_{vv} = 0. \]

Let \( K = \{\text{vec}(S^{(1)}), \text{vec}(S^{(2)}), \ldots, \text{vec}(S^{(l)})\} \in \mathbb{R}^{N^2 \times l} \), where \( \text{vec}(\cdot) \) indicates a vec-operator stacking each column of \( S^{(v)} \) (\( v = 1, 2, \ldots, l \)) on top of the next. Then, problem (24) can be transformed as follows:
\[ \min R \sum_{v=1}^{l} \|K(:, v) - KR(:, v)\|^2 + \|R(:, v)\|^2 \]
\[ s.t. 0 \leq R_{iv} \leq 1, \sum_{i=1,i \neq v}^{l} R_{iv} = 1, R_{vv} = 0. \]

where \( K(:, v) \) and \( R(:, v) \) denote the \( v \)th column of \( K \) and \( R \), respectively. As seen from (25), it can be divided into \( l \) independent sub-problems. Then, we can obtain the following sub-problem w.r.t. \( R(:, v) \):
\[ \min \frac{0 \leq R_{iv} \leq 1, \sum_{i=1,i \neq v}^{l} R_{iv} = 1, R_{vv} = 0}{\sum_{i=1,i \neq v}^{l} R_{iv} = 1, R_{vv} = 0.} \]

Problem (26) can be solved by the accelerated projected gradient algorithm in [40].

E. Update \( \alpha^{(v)} \) by Fixing Other Variables

To update \( \alpha^{(v)} \) with other variables fixed, we first let \( d^{(v)} = \|(X^{(v)} - U^{(v)}V^T)W^{(v)}\|^2_F + \lambda \|U^{(v)}\|^2_F + \beta(\text{Tr}(V^T L^{(v)}V) + \|S^{(v)} - \sum_{i=1,i \neq v}^{l} S^{(i)}(R_{iv})^2 \|^2_F) \). Then, the problem of solving \( \alpha^{(v)} \) is reduced as follows:
\[ \min \sum_{v=1}^{l} (\alpha^{(v)})^2 d^{(v)}, s.t. \sum_{v=1}^{l} \alpha^{(v)} = 1, 0 \leq \alpha^{(v)} \leq 1. \]

We introduce the following Lagrange function to solve \( \alpha^{(v)} \):
\[ \mathcal{L}(\alpha^{(v)}) = \sum_{v=1}^{l} (\alpha^{(v)})^2 d^{(v)} - \psi \left( \sum_{v=1}^{l} \alpha^{(v)} - 1 \right), \]

where \( \psi \) is the Lagrange multiplier. Taking the partial derivative of \( \mathcal{L} \) w.r.t. \( \alpha^{(v)} \) and setting it to zero, we have
\[ \frac{\partial \mathcal{L}(\alpha^{(v)})}{\partial \alpha^{(v)}} = \gamma \alpha^{(v)} - \psi = 0 \]
\[ \Rightarrow \alpha^{(v)} = \left( \frac{\psi}{\gamma d^{(v)}} \right)^{1/(\gamma - 1)}, \]

Besides, since \( \sum_{v=1}^{l} \alpha^{(v)} = 1 \), we can get the final solution of \( \alpha^{(v)} \) as follows:
\[ \alpha^{(v)} = \left( \frac{d^{(v)}}{\sum_{v=1}^{l} d^{(v)}} \right)^{1/(1 - \gamma)}. \]

The overall optimization procedure of \( \text{C}^2\text{IMUF} \) is summarized in Algorithm I. Regarding the initialization, we use the \( k \)-nearest neighbor graph method to initialize the similarity matrix \( S^{(v)} \) according to [42]. \( V \) is initialized by using the eigenvalue decomposition on (7). \( \alpha^{(v)} \) is set as \( 1 \) for each view and \( U^{(v)} \) is initialized randomly.

V. CONVERGENCE AND COMPLEXITY ANALYSIS

In this section, we give the theoretical analysis on convergence and complexity of the proposed algorithm \( \text{C}^2\text{IMUF} \).

A. Convergence Analysis

Since the objective function in (8) is not convex for the variables \( U^{(v)}, V, S^{(v)} \), \( R \) and \( \alpha^{(v)} \) simultaneously, we divide (8) into five sub-objective functions, namely, (9), (16), (20), (24) and (27). Hence, we can prove the convergence of Algorithm I by proving the monotonic convergence of each sub-objective functions. We first introduce the following theorem to prove the
As to updating $\alpha^{(v)}$, it only involves the numerical division and addition operations in terms of (30). Hence, the total time complexity of Algorithm 1 is $O(N_{max}(d_v, N)cT)$, where $T$ is the number of iteration.

VI. EXPERIMENTS

In this section, we demonstrate the effectiveness of our method by comparing with several state-of-the-art unsupervised feature selection methods in terms of clustering performance.

A. Experimental Schemes

1) Datasets: In experiments, eight real-world multi-view datasets are used to investigate the performance of our method. The descriptions of these datasets are presented as follows.

BBCSport: It is a multi-view document dataset collected from the BBC Sport website. We follow the experimental settings in [41], where 116 news objects with four views are selected from the entire datasets.

3Sources: It contains 169 news objects with six topics, which are collected from three online news medias.

Washington Dataset: It is one of the famous WebKB dataset, which contains 203 web-pages described by three views including the content features, title features and citation features.

BBC4view: It consists of 685 news articles characterized by four views. These articles are divided into five topics including politics, entertainment, business, sports and technology.

Reuters: It contains 1200 documents, which are characterized by five different languages (views).

Caltech101-7: It is a multi-view image dataset containing 1474 pictures of objects. Each image is characterized by six visual features.

Handwritten 2 sources (HW2sources) [42]: It contains 2000 objects of ten handwritten digits (0-9) derived from two different information sources, namely, MNIST and USPS.

CiteSeer [43]: It is a collection of scientific publications, which contains 3312 documents classified into six categories.

2) Evaluation Metrics: Two widely used clustering metrics including clustering accuracy (ACC) and normalized mutual information (NMI) are employed to measure the performance of compared methods. They can be defined as follows, respectively.

\[
\text{ACC} = \frac{1}{n} \sum_{i=1}^{n} \delta(y_i, \text{map}(\hat{y}_i)),
\]

\[
\text{NMI} = \frac{2 \cdot \text{Mutual Information}(y, \hat{y})}{\text{Entropy}(y) + \text{Entropy}(\hat{y})}.
\]

The detailed statistics of these multi-view datasets are summarized in Table I. Besides, we simulate the incomplete multiview setting according to the proposed method in [44], namely, removing certain ratios of instances from each view as missing data. The missing ratio is set from 10% to 50% with a step of 10%.

A. Experimental Schemes

1) Datasets: In experiments, eight real-world multi-view datasets are used to investigate the performance of our method. The descriptions of these datasets are presented as follows.

BBCSport: It is a multi-view document dataset collected from the BBC Sport website. We follow the experimental settings in [41], where 116 news objects with four views are selected from the entire datasets.

3Sources: It contains 169 news objects with six topics, which are collected from three online news medias.

Washington Dataset: It is one of the famous WebKB dataset, which contains 203 web-pages described by three views including the content features, title features and citation features.

BBC4view: It consists of 685 news articles characterized by four views. These articles are divided into five topics including politics, entertainment, business, sports and technology.

Reuters: It contains 1200 documents, which are characterized by five different languages (views).

Caltech101-7: It is a multi-view image dataset containing 1474 pictures of objects. Each image is characterized by six visual features.

Handwritten 2 sources (HW2sources) [42]: It contains 2000 objects of ten handwritten digits (0-9) derived from two different information sources, namely, MNIST and USPS.

CiteSeer [43]: It is a collection of scientific publications, which contains 3312 documents classified into six categories.

The detailed statistics of these multi-view datasets are summarized in Table I. Besides, we simulate the incomplete multiview setting according to the proposed method in [44], namely, removing certain ratios of instances from each view as missing data. The missing ratio is set from 10% to 50% with a step of 10%.

2) Evaluation Metrics: Two widely used clustering metrics including clustering accuracy (ACC) and normalized mutual information (NMI) are employed to measure the performance of compared methods. They can be defined as follows, respectively.

\[
\text{ACC} = \frac{1}{n} \sum_{i=1}^{n} \delta(y_i, \text{map}(\hat{y}_i)),
\]

\[
\text{NMI} = \frac{2 \cdot \text{Mutual Information}(y, \hat{y})}{\text{Entropy}(y) + \text{Entropy}(\hat{y})}.
\]

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
where \( y_i \) and \( \hat{y}_i \) denote the ground truth label and clustering result of the \( i \)th instance, respectively, \( n \) is the number of samples, \( \delta(x, y) \) is an indicator function satisfying \( \delta(x, y) = 1 \) for \( x = y \), otherwise \( \delta(x, y) = 0 \), and \( \text{map}() \) denotes the permutation mapping function which maps each cluster index to the best ground true label using Kuhn-Munkres algorithm [45].

Let \( Y = \{ Y_i \}_{i=1}^{t} \) and \( Y' = \{ Y'_j \}_{j=1}^{t} \) denote the ground-truth partition and clustering result, respectively. Then, NMI is given by

\[
\text{NMI}(Y, Y') = \frac{\sum_{i=1}^{t} \sum_{j=1}^{t} |Y_i \cap Y'_j| \log \frac{|Y_i \cap Y'_j|}{|Y_i||Y'_j|}}{\max \left( \sum_{i=1}^{t} |Y_i| \log \frac{|Y_i|}{\pi}, \sum_{j=1}^{t} |Y'_j| \log \frac{|Y'_j|}{\pi} \right)},
\]

where \( \cdot \) indicates the cardinal operation.

For these two metrics, the larger value indicates the better performance.

3) Comparison Methods: We compare our method C\(^2\)IMUFS with several state-of-the-art unsupervised feature selection methods including four single-view based methods (i.e., LPscore, EGCFS, UDPFS and RUSLP) and six multi-view based methods (i.e., OMVFS, ACSL, CGMV-UFS, NSGL, CVFS and CvLP-DCL). These compared methods are briefly introduced as follows:

- **AllFea** uses all original features to be compared with.
- **LPscore** [4] selects features based on the Laplacian score, which measures the locality preserving power.
- **EGCFS** [8] incorporates the procedure of maximizing the between-class scatter matrix and the adaptive graph learning into a joint framework to select discriminative features.
- **UDPFS** [46] learns an unsupervised discriminative project for feature selection by introducing the fuzziness learning and sparse learning simultaneously.
- **RUSLP** [47] selects features by using the non-negative matrix factorization with \( \ell_2 \)-norm regularization and the preservation of local geometric structure.
- **OMVFS** [20] designs an incremental non-negative matrix factorization model to select features.
- **ACSL** [12] incorporates the adaptive collaborative similarity structure learning and multi-view feature selection into a unified framework.

- **CGMV-UFS** [21] learns a consensus cluster indicator matrix among different views and embeds the feature selection into a non-negative matrix factorization framework.
- **NSGL** [14] integrates the adaptive similarity graph learning and multi-view feature selection into a unified framework by employing the complementary information across views.
- **CVFS** [48] embeds the HilbertSchmidt Independence Criterion into a non-negative matrix factorization model for feature selection of streaming data.
- **CvLP-DCL** [13] constructs two modules including consensus learning and view-specific local structure learning to select discriminative features.

We use these single-view based methods to select features by combining all features derived from different views into a single view. These parameters of compared methods are set according to the original papers for the guarantee of obtaining the optimal results. Meanwhile, we use the same grid search strategy to tune parameters according to the following ranges: \( \lambda \) and \( \beta \) in \( \{10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 10^2, 10^3\} \), \( \gamma \) in \( \{2, 3, 4, 5, 6, 7, 8\} \) and \( p \in \{0.001, 0.01, 0.1, 1\} \). Since the optimal number of selected features is hard to determine in each dataset, we set the percentage of selected features from 10% to 50% with an interval of 10% for all datasets. Then, we run the incomplete multi-view clustering algorithm proposed in [49] 30 times on selected features and report the average results with standard deviation.

B. Comparisons Over Clustering Performance

In this section, we compare the clustering performance of our method with other competing methods in terms of ACC and NMI. Table II shows the clustering result ACC of different methods on eight datasets, where the best performance is highlighted with bold-face type. In addition, two-sample t-test is used to verify whether our proposed method C\(^2\)IMUFS is significantly superior to other methods. The values with \( \bullet / \circ \) in Table II indicate that whether C\(^2\)IMUFS is statically better / inferior than other compared methods with the significance level of 0.05. Due to the limited space, the clustering result NMI of different feature selection methods can be founded in Table I in the supplementary material, available online.

From the following Table II and the supplementary material Table I, we can see that our method C\(^2\)IMUFS is better than other methods in most cases. As to BBCS, Reu, BBC4 and HW2, C\(^2\)IMUFS gains over 13% and 11% average improvement compared to all other methods in terms of ACC and NMI, respectively. As to 3S and CS, C\(^2\)IMUFS also obtains more than 5% improvement in average ACC and NMI. On the Cal7 dataset, C\(^2\)IMUFS is still superior to the baseline in terms of both ACC and NMI. On the Washington dataset, C\(^2\)IMUFS outperforms other methods in ACC and gets the second best performance in NMI. In addition, when comparing with the baseline method AllFea, C\(^2\)IMUFS outperforms AllFea over all datasets both in ACC and NMI, which demonstrates the effectiveness of the proposed method. Moreover, C\(^2\)IMUFS

---

**TABLE I**

A DETAIL DESCRIPTION OF DATASETS

| Datasets       | Abbr | Views | Instances | Features | Classes |
|----------------|------|-------|-----------|----------|---------|
| BBCSport       | BBCS | 4     | 116       | 1991/2063/2113/2158 | 5       |
| 3Sources       | 3S   | 3     | 169       | 3560/3631/3068     | 6       |
| Washington     | DC   | 3     | 203       | 1703/230/230       | 4       |
| BBCView        | BBC4 | 4     | 685       | 4659/4665/4684     | 5       |
| Reuters        | Reu  | 3     | 1200      | 2000/2000/2000/2000 | 6       |
| Caltech101-7   | Cal7| 6     | 1474      | 46/40/254/184/512/928 | 7       |
| HW2sources     | HW2  | 2     | 2000      | 764/236            | 10      |
| CiteScore      | CS   | 2     | 3312      | 3312/3703          | 6       |

---

**Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.**
TABLE II  
MEANS AND STANDARD DEVIATION (%) OF ACC FOR DIFFERENT METHODS ON EIGHT DATASETS WITH MISSING RATIO 0.3 WHILE SELECTING 20% OF ALL FEATURES

| Methods  | Datasets | ACC (%) |
|----------|----------|---------|
|          | BBCS     | 65.13 ± 3.37 |
|          | 3S       | 65.13 ± 3.37 |
|          | Washington | 65.13 ± 3.37 |
|          | BBC4     | 65.13 ± 3.37 |
|          | Reu      | 65.13 ± 3.37 |
|          | Cal7     | 65.13 ± 3.37 |
|          | HW2      | 65.13 ± 3.37 |
|          | CS       | 65.13 ± 3.37 |

Fig. 2. ACC of different methods with different selected features ratios.

entirely outperforms all single-view based methods and achieves almost 10% average improvement on most of datasets in terms of ACC and NMI. This demonstrates the effectiveness of our method using the consensus and complementary information among different views in comparison with these single-view based methods combining each data view by stacking.

Since it is hard to determine the optimal number of selected features in each dataset, we also show the clustering performances of different methods vary with the number of selected features. Due to the page limitation, we only report the experimental results of ACC. Fig. 2 plots the ACC values with different ratios of selected features while fixing the missing ratio at 0.3. As can be seen, C²IMUFS outperforms other methods with the variation of feature selection ratio from 10% to 50% in most cases. Furthermore, we also show the ACC of different methods with variation of missing ratio while the feature selection ratio is set to 20%. From Fig. 3, we can observe that C²IMUFS is still better than other compared methods on all datasets in most of the time. The experimental results of NMI with regards of different selected feature ratios and missing ratios can be founded in Figs. 1 and 2 in the supplementary material, respectively. In these two figures, we also can find that our method is still better than other methods in most of the cases. The superior performance of the proposed method is attributed to simultaneously learn the complementary and consensus information across different views and embed them into an extended WNMF framework with adaptive view-weight learning and a flexible sparse constraint.
C. Parameter Sensitivity

In this subsection, we investigate the parameter sensitivity of the proposed method \( C^2 \)IMUFS with regards to \( p, \beta, \lambda \), and \( \gamma \). \( p (0 < p \leq 1) \) is the parameter to control the sparsity and convexity of \( C^2 \)IMUFS, namely, the smaller \( p \) is, the more sparse solution is. We first study the effect of \( p \) value on our model performance. Fig. 4 shows ACC on BBCS, Cal7 and Reu datasets when \( p \) varies in \( \{0.001, 0.01, 0.1, 1\} \) and other parameters are tuned to obtain the optimal result. There yields the similar results on the other datasets. From Fig. 4, we can see that the performance of \( C^2 \)IMUFS with \( p < 1 \) is superior that of \( C^2 \)IMUFS with \( p = 1 \) in most of cases. Hence, we can flexibly select the value of \( p \) on different datasets to obtain the more accurate sparse, which is beneficial to improve the performance of \( C^2 \)IMUFS. In addition, to investigate the sensitivity of \( C^2 \)IMUFS w.r.t. parameters \( \beta, \lambda \), and \( \gamma \), we show the performance of \( C^2 \)IMUFS by varying a parameter and fixing the rest two parameters. Due to the limited space, we only report the ACC of \( C^2 \)IMUFS on BBCS dataset, as shown in Fig. 5. We can see that ACC show a slight fluctuations w.r.t. \( \beta \) and \( \lambda \) in subfigures (a) and (b), and a relatively large fluctuations w.r.t. \( \gamma \) in subfigure (c). Thus, \( C^2 \)IMUFS is not very sensitive to \( \beta \) and \( \lambda \) and relatively sensitive to \( \gamma \). We need tune the parameter \( \gamma \) to obtain a better performance by using grid search method in practice.

D. Convergence Analysis

We have theoretically proven the convergence of the proposed algorithm \( C^2 \)IMUFS in Section V-A. In this subsection, we
discuss its convergence speed experimentally. Fig. 6 shows the convergence curves of C^2IMUFS on BBCS, Cal7 and Reu datasets, where the x-coordinate pertains to the number of iterations and the y-coordinate pertains to the objective function values. The convergence curves of C^2IMUFS on other five datasets (3S, Washington, BBC4, HW2 and CS) can be founded in the supplementary material Fig. 3. From these figures, we can observe that the convergence curve decreases sharply within a few iterations and achieves stable almost within 100 iterations.

E. Ablation Study

In this section, we carry out an ablation study to verify the effectiveness of two important modules in the proposed C^2IMUFS. To do so, two variants of the proposed C^2IMUFS are conducted as follows:

1) C^2IMUFS-I: It is only composed of the extended WNMF module in C^2IMUFS.

2) C^2IMUFS-II: It is only composed of the consensus and complementary learning module in C^2IMUFS and a

$$
\min_{\alpha(v), \gamma} \sum_{v=1}^{l} (\alpha(v))^\gamma \left\| \left( X^{(v)} - U^{(v)} V^{(v)^T} \right) W^{(v)} \right\|_F^2 + \lambda \left\| U^{(v)} \right\|_{p,2}^2
$$

(33)
The constraints of C²IMUFII and C²IMUFIII are the same with (8).

Table III shows the ablation experiment results by comparing the proposed C²IMUFS with its two variants. From this table, we can find that C²IMUFS has the best performance on all datasets in terms of ACC and NMI. Compared with C²IMUFS-I, C²IMUFS significantly improves the performance. This demonstrates the effectiveness of the proposed consensus and complementary learning module, which can learn the complete similarity graph in incomplete multi-view data and capture the common clustering structure of different views. In addition, C²IMUFS is also better than C²IMUFII, which demonstrates the proposed extended WNMF module is beneficial to improve the performance. The reason is that the extended WNMF module can automatically determine both the view and instance importance and flexibly control the sparsity of feature selection matrix.

VII. CONCLUSION

In this paper, we proposed a novel multi-view unsupervised feature selection method C²IMUF for incomplete multi-view data. Different from the existing methods limited the complete-view scenario, our method integrates the feature selection of incomplete multi-view data and the learning of complementary and consensus information across different views into a unified framework. An extended WNMF model equipped with the adaptive view-weight learning and a flexible sparse constraint was presented to select features from the incomplete multi-view data. Meanwhile, the complementary and consensus learning-guided module was developed to obtain the complete similarity graph in each view and the common clustering indicator matrix across of different views. We also developed an iterative optimization algorithm to solve the proposed problem with proven convergence. Experimental results on eight real-world multi-view datasets demonstrated the superiority of the proposed method compared with the state-of-the-art methods.

In many real applications, we may obtain a small amount of labeled instances, which can provide valuable discriminative and semantic information to guide the procedure of multi-view feature selection. Hence, our future work will extend the proposed unsupervised framework to conduct semi-supervised feature selection on multi-view incomplete data.

REFERENCES

[1] Z. X. Wu, Y. G. Jiang, J. Wang, J. Pu, and X. Y. Xue, “Exploring inter-feature and inter-class relationships with deep neural networks for video classification,” in Proc. ACM Int. Conf. Multimedia, 2014, pp. 167–176.

[2] L. X. Cui, L. Bai, Y. C. Wang, X. Jin, and E. R. Hancock, “Internet financing credit risk evaluation using multiple structural interacting elastic net feature selection,” Pattern Recognit., vol. 114, 2021, Art. no. 107853.

[3] Z. Y. Guan, L. J. Zhang, J. Y. Peng, and J. P. Fan, “Multi-view concept learning for data representation,” IEEE Trans. Knowl. Data Eng., vol. 27, no. 11, pp. 3016–3028, Nov. 2015.

[4] X. F. He, D. Cai, and P. Niyogi, “Laplacian score for feature selection,” in Proc. Adv. Neural Inf. Process. Syst., 2005, pp. 507–514.

[5] L. Shi, L. Du, and Y. D. Shen, “Robust spectral learning for unsupervised feature selection,” in Proc. IEEE Int. Conf. Data Mining, 2014, 977–982.

[6] A. H. Yuan, M. B. You, D. J. He, and X. L. Li, “Convex non-negative matrix factorization with adaptive graph for unsupervised feature selection,” IEEE Trans. Cybern., vol. 52, no. 6, pp. 5522–5534, Jun. 2022.

[7] Y. Yang, H. T. Shen, Z. G. Ma, Z. Huang, and X. F. Zhou, “\(\ell_2,1\)-norm regularized discriminative feature selection for unsupervised learning,” in Proc. Int. Joint Conf. Artif. Intell., 2011, pp. 1589–1594.

[8] R. Zhang, X. Y. Zhang, and X. L. Li, “Unsupervised feature selection via adaptive graph learning and constraint,” IEEE Trans. Neural Netw. Learn. Syst., vol. 33, no. 3, pp. 1355–1362, Mar. 2022.

[9] Z. Wang, Y. F. Feng, T. Qi, X. S. Yang, and J. J. Zhang, “Adaptive multi-view feature selection for human motion retrieval,” Signal Process., vol. 210, pp. 691–701, 2016.

[10] L. Li, Z. Q. Wan, and H. B. He, “Incomplete multi-view clustering with joint partition and graph learning,” IEEE Trans. Knowl. Data Eng., vol. 29, no. 9, pp. 1998–2011, Sep. 2017.

[11] C. Tang et al., “Cross-view locality preserved diversity and consensus learning for multi-view unsupervised feature selection,” IEEE Trans. Knowl. Data Eng., vol. 34, no. 10, pp. 4705–4716, Oct. 2022.

[12] X. P. Bai, L. Zhu, C. Liang, J. J. Li, X. S. Nie, and X. J. Chang, “Multi-view feature selection via nonnegative structured graph learning,” Neurocomputing, vol. 387, pp. 110–122, 2020.

[13] J. Wen et al., “Unified tensor framework for incomplete multi-view clustering and missing-view inferring,” in Proc. AAAI Conf. Artif. Intell., 2021, pp. 10273–10281.

[14] L. Li, Z. Q. Wan, and H. B. He, “Incomplete multi-view clustering with adaptive view-weight learning and a flexible sparse constraint was presented to select features from the incomplete multi-view data. Meanwhile, the complementary and consensus learning-guided module was developed to obtain the complete similarity graph in each view and the common clustering indicator matrix across of different views. We also developed an iterative optimization algorithm to solve the proposed problem with proven convergence. Experimental results on eight real-world multi-view datasets demonstrated the superiority of the proposed method compared with the state-of-the-art methods.

[15] Y. D. Kim and S. Choi, “Weighted nonnegative matrix factorization,” in Proc. IEEE Int. Conf. Acoust. Speech Signal Process., 2009, pp. 1541–1544.

[16] M. J. Qian and C. X. Zhai, “Unsupervised feature selection for multi-view clustering on text-image web news data,” in Proc. ACM Int. Conf. Multimedia, 2014, pp. 229–230.

[17] C. P. Hou, F. P. Nie, H. Tao, and D. Y. Yi, “Multi-view unsupervised feature selection with nontrivial solution,” in Proc. IEEE Int. Conf. Data Mining, 2016, pp. 1293–1208.

[18] K. H. Guo, R. F. Zhang, Z. R. Zhou, Y. Y. Tang, and L. Kuang, “Combined retrieval: A convenient and precise approach for internet image retrieval,” Inf. Sci., vol. 358, pp. 151–163, 2016.

[19] Z. Zhao and H. Liu, “Spectral feature selection for supervised and unsupervised learning,” in Proc. Int. Conf. Mach. Learn., 2007, pp. 1151–1157.

[20] Z. Zhao and H. Liu, “Spectral feature selection for supervised and unsupervised learning,” in Proc. Int. Conf. Mach. Learn., 2007, pp. 1151–1157.

[21] S. Moldouano and R. Weber, “A wrapper method for feature selection using support vector machines,” Inf. Sci., vol. 179, no. 13, pp. 2208–2217, 2009.

[22] W. X. Shao, L. F. He, C. T. Lu, X. K. Wei, and P. S. Yu, “Online unsupervised multi-view feature selection,” in Proc. IEEE Int. Conf. Data Mining, 2016, pp. 1203–1208.

[23] C. Tang et al., “Consensus learning guided multi-view unsupervised feature selection,” Knowl. Based Syst., vol. 160, pp. 49–60, 2018.

[24] W. L. Guan, H. K. Wen, X. M. Song, C. H. Yeh, X. J. Chang, and L. Q. Nie, “Multimodal compatibility modeling via exploring the consistent and complementary correlations,” in Proc. ACM Int. Conf. Multimedia, 2021, pp. 229–230.

[25] Z. Zhao and H. Liu, “Spectral feature selection for supervised and unsupervised learning,” in Proc. Int. Conf. Mach. Learn., 2007, pp. 1151–1157.

[26] S. Moldouano and R. Weber, “A wrapper method for feature selection using support vector machines,” Inf. Sci., vol. 179, no. 13, pp. 2208–2217, 2009.
[28] H. J. Wang, Y. H. Zhang, J. Zhang, T. R. Li, and L. X. Peng, “A factor graph model for unsupervised feature selection,” Inf. Sci., vol. 480, pp. 144–159, 2019.

[29] Y. Y. Huang, Z. X. Shen, F. X. Cai, T. R. Li, and F. M. Lv, “Adaptive graph-based generalized regression model for unsupervised feature selection,” Knowl. Based Syst., vol. 227, pp. 107–156, 2021.

[30] Z. C. Li, J. Liu, Y. Yang, X. F. Zhou, and H. Q. Lu, “Clustering-guided sparse structural learning for unsupervised feature selection,” IEEE Trans. Knowl. Data Eng., vol. 26, no. 9, pp. 2138–2150, Sep. 2014.

[31] H. Yu, X. C. Wang, G. Y. Wang, and X. H. Zeng, “An active three-way clustering method via low-rank matrices for multi-view data,” Inf. Sci., vol. 507, pp. 823–839, 2020.

[32] C. Q. Zhang et al., “Generalized latent multi-view subspace clustering,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 42, no. 1, pp. 86–99, Jan. 2020.

[33] C. Y. Lu, H. Min, Z. Q. Zhao, L. Zhu, D. S. Huang, and S. C. Yan, “Robust and efficient subspace segmentation via least squares regression,” in Proc. Eur. Conf. Comput. Vis., 2012, pp. 347–360.

[34] P. F. Zhu, W. C. Zhu, Q. H. Hu, C. Q. Zhang, and W. M. Zuo, “Subspace clustering guided unsupervised feature selection,” Pattern Recognit., vol. 66, pp. 364–374, 2017.

[35] Z. C. Li and J. H. Tang, “Unsupervised feature selection via nonnegative spectral analysis and redundancy control,” IEEE Trans. Image Process., vol. 24, no. 12, pp. 5343–5355, Dec. 2015.

[36] G. Gordon and R. Tibshirani, “Karush-Kuhn-Tucker conditions,” Optimization, vol. 10, no. 36, 2012, Art. no. 725.

[37] F. P. Nie, X. Q. Wang, M. Jordan, and H. Huang, “The constrained Laplacian rank algorithm for graph-based clustering,” in Proc. AAAI Conf. Artif. Intell., 2016, pp. 1969–1976.

[38] J. Wen et al., “Adaptive graph completion based incomplete multi-view clustering,” IEEE Trans. Multimed., vol. 23, pp. 2493–2504, 2020.

[39] C. Ding, T. Li, and M. I. Jordan, “Convex and semi-nonnegative matrix factorizations,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 32, no. 1, pp. 45–55, Jan. 2010.

[40] J. Huang, F. P. Nie, and H. Huang, “A new simplex sparse learning model to measure data similarity for clustering,” in Proc. Int. Joint Conf. Artif. Intell., 2015, pp. 3569–3575.

[41] J. Wen, Y. Xu, and H. Liu, “Incomplete multiview spectral clustering with adaptive graph learning,” IEEE Trans. Cybern., vol. 50, no. 4, pp. 1418–1429, Apr. 2020.

[42] H. Wang, Y. Yang, B. Liu, and H. Fujita, “A study of graph-based system for multi-view clustering,” Knowl. Based Syst., vol. 163, pp. 1009–1019, 2019.

[43] C. L. Giles, K. D. Bollacker, and S. Lawrence, “CiteSeer: An automatic citation indexing system,” in Proc. ACM Conf. Digit. Libraries, 1998, pp. 89–98.

[44] X. W. Liu et al., “Late fusion incomplete multi-view clustering,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 41, no. 10, pp. 2410–2423, Oct. 2019.

[45] A. Strehl and J. Ghosh, “Cluster ensembles-a knowledge reuse framework for combining multiple partitions,” J. Mach. Learn. Res., vol. 3, pp. 583–617, 2002.

[46] R. Wang, J. T. Bian, F. P. Nie, and X. L. Li, “Unsupervised discriminative projection for feature selection,” IEEE Trans. Knowl. Data Eng., vol. 34, no. 2, pp. 942–953, Feb. 2022.

[47] C. Luo, J. Zheng, T. R. Li, H. M. Chen, Y. Y. Huang, and X. Peng, “Orthogonally constrained matrix factorization for robust unsupervised feature selection with local preserving,” Inf. Sci., vol. 586, pp. 662–675, 2022.

[48] Y. Y. Xu et al., “Unsupervised cross-view feature selection on incomplete data,” Knowl. Based Syst., vol. 234, 2021, Art. no. 107595.

[49] J. Wen, Z. Zhang, Z. Zhang, L. K. Fei, and M. Wang, “Generalized incomplete multiview clustering with flexible locality structure diffusion,” IEEE Trans. Cybern., vol. 51, no. 1, pp. 101–114, Jan. 2021.

Yanyong Huang received the double PhD degrees from the Southwest Jiaotong University and FernUniversi- stat, in Hagen in 2018 and 2020, respectively. He is currently a professor in the School of Statistics, Southwestern University of Finance and Economics. His research interests include data mining, machine learning and granular computing.

Fengmao Lv received the bachelor’s and PhD degrees in computer science from the University of Electronic Science and Technology of China, Chengdu, China, in 2013 and 2018, respectively. He is currently an associate professor with Southwest Jiaotong University, Chengdu. His research interests include anomaly detection, generative adversarial network, reinforcement learning and spatio-temporal data mining.

Tianrui Li (Senior Member, IEEE) received the PhD degree from the Southwest Jiaotong University, Chengdu, China, in 2015. He is currently a Data Scientist of IET Smart Cities journal. He has published more than 20+ research papers in refereed conferences and journals.

Yuxin Cai received the PhD degree from Southwest Jiaotong University, in 2018. He was a postdoctoral Researcher with Tsinghua University from 2019 to 2020, a visiting scholar with Penn State University during 2017-2018, and an research intern with MSR from 2014 to 2017. He is currently a Data Scientist of JD Intelligent Cities Research. His research interests include: Urban Computing, Data Mining, and Deep Learning. He serves as associate editor of IET Smart Cities journal.