Knowledge and Information Systems
https://doi.org/10.1007/s10115-024-02197-1

REGULAR PAPER

Complementary incomplete weighted concept factorization methods for multi-view clustering

Ghufran Ahmad Khan[1] · Jalaluddin Khan[1] · Taushif Anwar[1] · Zaid Al-Huda[2] · Bassoma Diallo[3] · Naved Ahmad[4]

Received: 12 March 2024 / Revised: 14 June 2024 / Accepted: 30 July 2024

Abstract

The main aim of traditional multi-view clustering is to categorize data into separate clusters under the assumption that all views are fully available. However, practical scenarios often arise where not all aspects of the data are accessible, which hampers the efficacy of conventional multi-view clustering techniques. Recent advancements have made significant progress in addressing the incompleteness in multi-view data clustering. Still, current incomplete multi-view clustering methods overlooked a number of important factors, such as providing a consensus representation across the kernel space, dealing with over-fitting issue from different views, and looking at how these multiple views relate to each other at the same time. To deal these challenges, we introduced an innovative multi-view clustering algorithm to manage incomplete data from multiple perspectives. Additionally, we have introduced a novel objective function incorporating a weighted concept factorization technique to tackle the absence of data instances within each incomplete viewpoint. We used a co-regularization constraint to learn a common shared structure from different points of

✉ Ghufran Ahmad Khan
ghufraan.alig@gmail.com

Jalaluddin Khan
jalal4amu@gmail.com

Taushif Anwar
taushif21589@gmail.com

Zaid Al-Huda
zaid@stir.cdu.edu.cn

Bassoma Diallo
sanediallo2003@yahoo.fr

Naved Ahmad
nahmad@um.edu.sa

[1] Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Andhra Pradesh 522502, India

[2] Stirling College, Chengdu University, Sichuan 610106, China

[3] School of Computing and Artificial Intelligence, Southwest Jiaotong University, Chengdu 610756, China

[4] Department of Computer Science and Information Systems, College of Applied Sciences, AlMaarefa University, 13731 Riyadh, Saudi Arabia

Published online: 14 August 2024

view and a smooth regularization term to prevent view over-fitting. It is noteworthy that the proposed objective function is inherently non-convex, presenting optimization challenges. To obtain the optimal solution, we have implemented an iterative optimization approach to converge the local minima for our method. To underscore the effectiveness and validation of our approach, we conducted experiments using real-world datasets against state-of-the-art methods for comparative evaluation.

Keywords Clustering · Incomplete multi-view data · Concept factorization · Complementary information

1 Introduction

Multi-view learning is becoming increasingly popular as multi-view data finds applications across various real-world scenarios [1, 2]. The goal of this strategy is to utilize consistent and complementary information from several points of view. All the tasks in multi-view learning, multi-view clustering (MVC) is one of the most notable. MVC groups unlabeled data from several perspectives into clusters in order to produce reliable clustering results from all sides by utilizing a variety of perspectives. Many approaches have been developed in the field of multi-view clustering over the last ten years, as the literature [3–5] documents. Multi-view non-negative matrix factorization, or MultiNMF, is one such method [6]. This method incorporates a consensus constraint—which is essential for preserving consistent clustering results from many viewpoints—with a non-negative matrix factorization (NMF) process. Spectral clustering is used by another class of methods, such as centroid-based co-regularization and pairwise co-regularization [7], to produce clustering outcomes for every view. These techniques make use of a variety of procedures in order to align the clustering results from several viewpoints, guaranteeing consistency in the clustering results from multiple views.

The approaches for MVC discussed earlier presuppose that each of the examples' views is complete. However, circumstances where some viewpoints are absent are common in real-world scenario. When examining a web page, text and photographs, for instance, might be viewed as two distinct perspectives; nevertheless, some web sites may include neither text nor image data at all. Similar to this, a news story can be viewed from a variety of angles by reading news reports from numerous media sources, even though some outlets might not have covered the particular subject. When multi-view data are incomplete, traditional multi-view clustering methods fall short. In response to this challenge, a number of strategies have surfaced recently [8–10]. A crucial element of many clustering strategies, matrix factorization has been shown to be effective in a wide range of applications [11, 12]. These techniques aim to uncover latent representations of incomplete multi-view data by combining matrix factorization and regularization algorithms. The preservation of these representations was improved by Zhao et al. [13] with the addition of a graph Laplacian term to their optimization procedure.

Furthermore, the absence of information in missing views is the primary cause of incomplete multi-view clustering (IMC) shortcomings. IMC techniques can be split into non-inference and inference approaches, each of which uses a different strategy, to address this problem. Non-inference IMC [18–21] aims to achieve clustering with incomplete multi-view data while mitigating the effects of information loss. Wen et al. [22], for instance, used the samples that were accessible from each view to create an incomplete similarity graph, filling in the missing elements with zeros. Next, k-means clustering is used to a shared spectral embedding obtained from completed similarity graphs. Additionally, the local graph preservation approach was used by Wen et al. [23] to obtain a common representation from incomplete views. In order to efficiently fuse partial similarity graphs, Liang et al. [24] applied sample-level adaptive weights on the similarity graphs of all views that were available. Hu et al. [25] presented a matrix factorization-based method that aligns view-specific basis matrices to learn a shared representation from imperfect data. Next, k-means is utilized to cluster this shared representation. IMC techniques focus on organizing the reconstructed viewpoints and retrieving missing ones in order to prevent information loss [26–34]. Achieving high-quality retrieval of missing data is the primary goal. A straightforward method is to create unavailable samples by averaging characteristics. Zhou et al. [35], for instance, reduced the impact of missing cases by introducing a weighting method and adding average features for each view. But because all recovered samples from this method have the same properties, they are unable to provide enough useful information and may cause alignment issues between views. To mutually reinforce each other, integrating the inference and clustering processes is a more rational method [36–40]. Pairwise dimension graph preservation was used by Wen et al. [41] to recover the missing instances, and reverse graph regularization was used to guide finished views.

Concurrently, the various data perspectives show complementary and consensus behavior. Each view is important for computing clustering performance because it allows explicit information to be explored from incomplete multi-view data. The primary motivation of proposed study is the vital problem of completely using the data included in individual incomplete views for the analysis of the consensus structure of the heterogeneous perspective on the kernel space [7, 47, 48]. In order to achieve this objective, a new method called weighted concept factorization is introduced for clustering incomplete multi-view data. The proposed method seeks to reveal hidden structures, or clusters. First, each view of the data is normalized and missing data is imputed using the algorithm. Subsequently, three matrices are repeatedly refined: an association matrix that captures the relationships between data and clusters, a projection matrix that assigns a value to each feature within each view, and a consensus matrix that represents a single view that is shared by all data viewpoints. Using the correlation aspect, the disagreement factor is extracted. To further prevent the over-fitting of the view, the Frobenius norm is also used to pair up the projection matrix. This procedure is repeated until convergence is achieved or the maximum number of iterations is completed. The key aspects of the proposed algorithm are summarized as follows:

1. The missing issues in multi-view data are effectively handled by the proposed approach. In our suggested objective function, we employ the weighted concept factorization approach. For each incomplete view, a weight matrix is built so that the missing instances in each view have a lower weight than the examples that are provided.
2. To drive the latent feature matrix toward a consensus, we use the co-regularized technique. In order to avoid the view's over-fitting problem, and maintain the consistent information, the projection matrix, and the associated matrix are conjugated using the Frobenius norm. During the optimization process, each view's weight is automatically determined. To handle the related optimization problem efficiently and effectively, a new updating rule is created.
3. The outcomes of the carrying experiments on real-world datasets are displayed in terms of F-score, ACC, and NMI. According to the experimental study, the suggested approach outperforms other existing techniques in clustering.

We give a summary of the incomplete multi-view clustering methods that are currently in use in Sect. 2. Section 3 provides a detailed explanation of our suggested methodology. An examination of the experiments carried out using benchmark datasets is covered in Sect. 4. Our final remarks are presented in Sect. 5.

2 Related works

An overview of related work in incomplete multi-view clustering is given in this section.

2.1 Multi-incomplete view clustering (MIC)

The MIC method, as described in reference [14], is an IMC approach that utilizes weighted NMF. Its objective function can be formulated in the following manner:

$$\min_{U_f, V_f, V^* \geq 0} \sum_{f=1}^{F} \left\{ \left\| \left(X_f - U_f V_f\right) W_f \right\|_F^2 \right\} + \sum_{f=1}^{F} \left\{ \alpha \left\| \left(V_f - V^*\right) W_f \right\|_F^2 + \beta \left\| V_f \right\|_{2,1} \right\} \tag{1}$$

where α and β serve as the corresponding parameters for the respective terms. F denotes the total views, and the expression $\| \bullet \|_{2,1}$ encompass the $L_{2,1}$-norm. The matrix $X_f \in R^{m \times n}$ includes both present and missing values from the fth view, with the absent data filled by averaging corresponding viewpoints. $U_f \in R^{m \times c}$ and $V_f \in R^{c \times n}$ designated as the basis and coefficient matrices for the f^{th} view. The dimensions are outlined as follows: m signifies the original space dimension in the f^{th} view, c denotes the latent space dimension, and n represents the overall dataset size. V^* defines the common representation matrix, while W_f serves as the diagonal weighting matrix for the f^{th} view. If the i^{th} instance of the f^{th} view is available, $W_f^{ii} = z_v/n$, where z_v denotes the number of available instances in the f^{th} view. This technique undergoes iterative optimization.

2.2 Doubly aligned incomplete multi-view clustering (DAIMC)

A DAIMC is an incomplete clustering method for multi-view data based on weighted semi-NMF [25], and to express DAIMC's cost function:

$$\min_{V \geq 0} \sum_{f=1}^{F} \left\{ \left\| \left(X_f - U_f V\right) W_f \right\|_F^2 + \alpha \left(\left\| B_f^T W_f - I \right\|_F^2 \right) + \beta \left\| B_f \right\|_{2,1} \right\} \tag{2}$$

where α and β defined as the trade-off parameters for the respective terms. The input matrix is represented by $X_f \in R^{m \times n}$, the common coefficient matrix is V, and the diagonal weighting matrix for the f^{th} view is W_f. B_f is a regression coefficient matrix for the f^{th} view. W_f^{ii} equals 1 if the i^{th} instance of the f^{th} view is accessible, and W_f^{ii} equals 0 otherwise. By looking at Eq. (2), we can see that DAIMC aims to align several partial perspectives for V and U_f. This approach undergoes iterative optimization.

2.3 Incomplete multi-view clustering methods

This work addresses the incomplete multi-view clustering issue. Over the course of the last ten years, a variety of incomplete multi-view clustering approaches have been suggested.

In this regard, Yin et al. [42] presented incomplete multi-view clustering with cosine similarity. Enhancing the preservation of the data's manifold structure, this method computes cosine similarity directly in the original multi-view space. The need to include more variables is eliminated. Coherent method is achieved by merging the manifold structure preservation utilizing cosine similarity term with the matrix factorization component in the objective function. Chao et al. [43] offered a two-stage method to handle multi-view clustering in the event of any value missing that involved multiple imputation and ensemble clustering. The problem of missing values is addressed by multiple imputation, and multi-view clustering is implemented through the use of weighted ensemble clustering. Zhang et al. [44] introduced a novel approach that merges completed graphs into a single graph after filling in the gaps in the incomplete graphs based on agreement between various points of view. Further, the innovative method proposed by Xia et al. [45] involves information fusion in partition space to counteract consistency degradation, and adaptive weighting of all perspectives to represent their different contributions to clustering tasks. To create a desired similarity graph, the cluster structure information is used into the similarity learning process. In order to capture both the global and local structure of the data, Zhang et al. [46] suggested a novel incomplete multi-view clustering algorithm. By adding a distance regularization term to the model and applying a weighted fusion process, the suggested approach creates compact and discriminating representations from partial data.

3 Proposed method

This section provides a thorough explanation of the optimization step as well as a novel incomplete multi-view clustering technique that makes use of the CF approach. Additionally, we demonstrate the approach's convergent proof and establish the method's time complexity.

3.1 Concept factorization

CF serves as a potent method for matrix decomposition, particularly adept at handling datasets with negative values. Moreover, it demonstrates adaptability to altered data through kernel methodologies. Considering a data matrix $X = \left[x_1, x_2, ..., x_f\right] \in \Re^{m\times n}$, where each x_f is denoted by an f-dimensional feature vector, CF views every data point as an estimated linear formulation of all fundamental concepts. This approach provides a succinct representation of the data in the following manner:

$$x_f \approx \sum\nolimits_g w_g v_{fg} \tag{3}$$

In this context, v_{fg} represents the projection matrix of x_f onto the basis matrix u_f. Consequently, each basis matrix u_g is established through a linear combination involving all these data points. This concept is summarized as follows:

$$w_g \approx \sum\nolimits_f x_f u_{fg} \tag{4}$$

In the mathematical representation provided by Eq. (3) and (4), w_{fg} represents a positive association weight. This leads us to the subsequent mathematical formulation:

$$X_f \approx X_f U_f \left(V_f\right)^T \tag{5}$$

where U_f and V_f belong to the set of matrices with dimensions $n \times c$, the CF employs the Frobenius norm to approximate the data representation. This norm is utilized in minimizing the cost function through the subsequent objective function:

$$\min_{U_f, V_f} : O_{CF} = \left\| X_f - X_f U_f (V_f)^T \right\|_F^2 \ s.t. U_f, V_f \geq 0. \tag{6}$$

Following optimization, the variables adhere to the multiplicative update rule outlined below:

$$\left.\begin{aligned} U_f &\leftarrow U_f \frac{(K_f V_f)}{\left(K_f U_f V_f (V_f)^T\right)} \\ V_f &\leftarrow V_f \frac{(K_f U_f)}{\left(V_f (U_f)^T K_f U_f\right)} \end{aligned}\right\} \tag{7}$$

where $K_f = (X_f)^T X_f$ calculates the inner product within the initial data space.

3.2 Missing data completion

Given data matrix X_f with f-views, where each view is facing the problem of incompleteness. Since the missing instances could cause information to be incorrect for each view, we are unable to apply the clustering algorithm directly to partial data. In such a way, we introduce a weighted diagonal matrix for each incomplete view, which is filled through the following assumption:

$$W_f^s = \begin{cases} 1, & \text{if the } sth \text{ instances } in\ f^{th} \text{ view} \\ 0, & \text{otherwise} \end{cases} \tag{8}$$

3.3 Proposed objective function

In the input matrix $X = \{X_1, X_2, ..., X_F\} \in \Re^{m \times n}$, each row denotes a unique feature dimension, and each column represents an individual data instance, thereby defining the dataset with F views. CF satisfies the conditions for achieving the approximation through the three matrices, denoted as $X \approx XUV^T$. This is because $V \in \Re^{n \times c}$ acts as the projection matrix, displaying the projected values that correlate to the concepts, and $U \in \Re^{n \times c}$ acts as the association matrix, validating the relationship of data points to concepts. The following is the formulation of the objective function:

$$O_{WCFIMC} = \sum_{f=1}^{F} \left\{ \begin{array}{l} \left\| W_f \left(X_f - X_f U_f V_f^T \right) \right\|_F^2 + \alpha \omega_f \left\| W_f \left(V_f - V^* \right) \right\|_F^2 \\ +\beta \left\| U_f \right\|_F^2 + \gamma \left\| V_f \right\|_F^2 + \eta \left\| \omega_f \right\|^2 \end{array} \right. \tag{9}$$

$$s.t.\ \ U_f \geq 0,\ V_f \geq 0,\ \omega_f \geq 0,\ \sum_{f=1}^{F} \omega_f = 1.$$

- $\left\| W_f \left(X_f - X_f U_f V_f^T \right) \right\|_F^2$ is the mathematical representation of the concept factorization with weighted diagonal matrix.
- $\left\| W_f \left(V_f - V^* \right) \right\|_F^2$ represents the correlation between the coefficient matrix and consensus matrix.
- $\left\| U_f \right\|_F^2$ is used to represents the maintain the consistent information across the multiple views.
- $\left\| V_f \right\|_F^2$ defines to avoid the over-fitting issue among the views.

where α, β, γ and η are the trade-off parameters. We denote $Q_f = W_f^T W_f$ and Eq. (9) is rewritten as:

$$O_2 = \begin{cases} Q_f \left(tr \left(\left(X_f - X_f U_f V_f^T \right)^T \left(X_f - X_f U_f V_f^T \right) \right) \right) \\ +\alpha \omega_f Q_f \left(tr \left(\left(V_f - V^* \right)^T \left(V_f - V^* \right) \right) \right) + \beta tr \left(U_f U_f^T \right) \\ +\gamma tr \left(V_f V_f^T \right) + \eta \left\| \omega_f \right\|^2 \end{cases} \tag{10}$$

Defining the standard kernel matrix $K = X^T X$, which is used to calculate the data space's inner product. Next, we rewrite Eq. (10) in this way:

$$O_3 = \begin{cases} Q \left(tr\left(K\right) - 2tr\left(VU^T K\right) + tr\left(VU^T KUV^T\right)\right) \\ +\alpha \omega Q \left(tr(V - V^*)^T \left(V - V^*\right) \right) + +\beta tr \left(U_f U_f^T \right) + \gamma tr \left(V_f V_f^T \right) + \eta \|\omega\|^2 \end{cases} \tag{11}$$

To put it succinctly, the CIWCFMvC modifies the optimization approach to get the conventional comprehensive solution. Through the use of ω, each view is stated as follows: ω as $1/M$. The k-means algorithm returns W, V, and V^* primary values.

3.4 Optimization of the proposed function

Lagrange's multiplier (LM) is integrated during the optimization process to ascertain the most optimal local solution, which is accomplished through the use of the iterative updating technique. Karush–Kuhn–Tucker (KKT) criteria are then taken into consideration for the analysis of the final amended rules.

3.4.1 Optimization of U

For the restrictions $U_{a,b} \geq 0$, assume the LM $\phi_{a,b}$. In order to assess the function's optimal outcome in light of the limitations, the LM is applied. In the end, the formulation of the Lagrange's function L_1 is $L_1 = O - tr(\phi U)$. We address the relevant phrase up to U.

$$L_1 = Q\left(-2tr\left(VU^T K\right) + tr\left(VU^T KUV^T\right)\right) + \beta tr\left(UU^T\right) - tr\left(\phi U\right) \tag{12}$$

By applying the partial derivative of L_1 w.r.t U:

$$\frac{\partial L_1}{\partial U} = Q\left(-2KV + 2KUVV^T\right) + 2\beta U - \phi \tag{13}$$

Using the KKT condition $\phi_{ik} U_{ik} = 0$, the following optimize rule for U is:

$$U_{ik} = U_{ik} \frac{(QKV)_{ik}}{\left(QKUV^T V + \beta U\right)_{ik}} \tag{14}$$

3.4.2 Optimization of V

For the constraints $V_{a,b} \geq 0$, consider the LM $\psi_{a,b}$. Then, $L_2 = O - tr(\psi V)$ is the reformed Lagrange's function. We take into account only the required element of V.

$$L_2 = Q\left(-2tr\left(VU^T K\right) + tr\left(VU^T KUV^T\right)\right) + \alpha\omega Q\left(tr\left(V - V^*\right)^T\left(V - V^*\right)\right) \\ +\gamma tr\left(VV^T\right) - tr(\psi V) \tag{15}$$

By applying the partial derivative of L_2 w.r.t V:

$$\frac{\partial L_2}{\partial V} = Q\left(-2KU + 2VU^T KU\right) + Q\left(2\alpha\omega\left(V - V^*\right)\right) + 2\gamma V - \psi \tag{16}$$

Using the KKT condition $\psi_{i,k}V_{i,k} = 0$, the following optimize rule for H is:

$$V_{i,k} = V_{i,k}\frac{QKU + \alpha\omega QV^*}{VU^T QKU + \alpha\omega QV + \gamma V} \tag{17}$$

It is important to highlight that in order to avoid U_f from reaching excessively high values (which could result in extremely low values of V_f), it's typical to impose a constraints on each associate matrix U_f. However, the updated U_f might not satisfy the given constraints. Therefore, normalization is applied to matrices U and V in order to obtain the consistency constraint by the following scenario:

$$V \leftarrow V(N)^{\frac{-1}{2}}, U \leftarrow U(N)^{\frac{1}{2}} \tag{18}$$

While a diagonal matrix is implied by N and is expressed as:

$$N = diag\left(\sum\nolimits_z (V)_{z,1}, \sum\nolimits_z (V)_{z,2}, \ldots, \sum\nolimits_z (V)_{z,c}\right) \tag{19}$$

3.4.3 Optimization *V**

Assuming $\zeta_{a,b}$ as the LM, let $V^*_{a,b} \geq 0$ be the constraints. Then, $L_3 = O - tr(\zeta V^*)$ is the transformed Lagrange's function. We focus on terms that contain only V^*, and we use the partial derivation of Eq. (13) with respect to V^*.

$$L_3 = \alpha\omega Q\left(tr\left(\left(V - V^*\right)^T \left(V - V^*\right)\right)\right) \tag{20}$$

The above equation is solved, and the update rule for V^* is then drawn:

$$V^* = \frac{\sum_{j=1}^{M}\omega_j QV_j}{\sum_{j=1}^{M}\omega_j Q} \tag{21}$$

3.4.4 Optimization ω

The weights for distinct views are automatically computed based on the disagreement factor between each V, and V^*. Subsequently, the objective function is reformulated in the following manner:

$$O(\omega) = \sum_{j=1}^{J}\omega\left\|V - V^*\right\|_F^2 + \eta\left\|\omega\right\|_2^2 \tag{22}$$

where $\pi\left\|\omega\right\|_2^2$ is used to control the smoothen the weight distribution among the multiple views to avoid the futile solution. Equation (22) is effectively solved by the quadratic programing Matlab function, i.e., quadprog.

Algorithm 1: The CIWCFMvC Algorithm

```
Input:
The multi-view data $X_f$; Cluster number: $c$;
Initialize the view's weight $\omega = 1/F$ for individual view;
The values of the parameters $\alpha$, $\beta$, $\gamma$, and $\eta$.
Output:
Final association matrix $U_f$;
Final projection matrix $V_f$;
Consensus matrix $V^*$.
Initialization:
Use the average feature values to fill in the missing instances in each incomplete view;
Normalized each view of $X_f$ such that $\|X_f\| = 1$;
Initialize the values of $U_f$, and $V_f$;
repeat
    for i = 1 to F do
        Fix $V^*$, $V_f$, optimize $U_f$ by Eq. (14) ;
        Fix $V^*$, $U_f$, optimize $V_f$ by Eq. (17);
        Normalize $U_f$ and $V_f$ by Eq. (18)
    Fix $U_f$, $V_f$ , optimize $V^*$ by Eq. (21);
    Optimize weight $\omega$ by Eq. (22);
until convergence or maximum iteration achived.;
```

3.5 Computational complexity

We examine the complexity of the proposed method in this section. The kernel's computational complexity is $O(mn^2)$. The related cost for the multiplicative updating case is $O(tmn)$ if we assume that the multiplicative update ends after t iterations. Thus, $O(mn^2 + tmn)$ can be used to represent the overall computational complexity of the suggested approach.

4 Experiments and analysis

This section presents a comparison of seven state-of-the-art approaches on seven benchmark datasets with the proposed CIWCFMvC method. The normalized mutual information (NMI), F-score, and accuracy (ACC) are used to evaluate the clustering performances. The whole cases are arranged in these datasets. Next, in order to render the data as incomplete, we arbitrarily eliminate some representations from each view. In particular, the interval of 20% represents the ratio of incomplete occurrences, which ranges from 10 to 50%.

4.1 Dataset

To assess the efficacy of the recommended method, we conduct analysis on widely employed benchmark datasets, i.e., 3Sources,[1] NGs,[2] Wikipedia Articles,[3] BBCSport[4],

[1] http://mlg.ucd.ie/datasets/3sources.html.

[2] http://lig-membres.imag.fr/grimal/data.html.

[3] http://www.svcl.ucsd.edu/projects/crossmodal/.

[4] http://mlg.ucd.ie/datasets/segment.html.

Table 1 Important statistics of the benchmark datasets

Dataset	Size	# view	# cluster	#dimension
3Sources	169	3	6	{3560, 3631, 3068}
NGs	500	3	5	{2000, 2000, 2000}
Wikipedia Articles	693	2	10	{128, 10}
BBCSport	544	2	5	{3183, 3203}
BBC	685	4	5	{4659, 4633, 4665, 4684}
WebKB	203	3	4	{500, 500, 500}
Citeseer	3312	2	6	{3312, 3703}
Reuters	1200	5	6	{2000, 2000, 2000, 2000, 2000}

BBC,[5] WebKB,[6] Citeseer,[7] and Reuters.[8]. The description of the datasets is illustrated in Table 1.

- **3Sources** dataset is compiled from three reliable online news sources, each of which offers a distinct viewpoint. A selection of 169 distinct news stories has been made from these sources.
- **NGs** dataset is the subsets of the 20 newsgroup NGs dataset, which contains archives from a variety of newsgroups. Extracts of reports from various news groups have been used, with each group represented as an independent point of view.
- **Wikipedia Articles** dataset is made up of carefully picked parts of Wikipedia's featured articles that have been put together in reports. Since it was compiled in October 2009, 2,669 articles from 29 different categories have been included. The most popular ten categories are highlighted, including articles that have several sections and photographs.
- **BBCSport** is extracted from BCSport website which contains 544 records. A manual classification into one of five subject groups has been performed on each record, which has been divided into two sections.
- **BBC** website maintains a collection of articles that are organized into five primary categories: business, entertainment, politics, sports, and technology. These articles cover the years 2004 to 2005. Six hundred and eighty-five stories were selected from four different sources.
- **WebKB** dataset comprises 203 web pages organized into four divisions. The content of each webpage, including the title text and hyperlinks, define it.
- **Citeeer** are 3312 papers in this collection, linked by 4732 citations. Each of these publications is annotated using six different labels: DB, IR, ML, Agents, AI, and HC.
- **Reuters** is a compilation of English documents translated into four additional languages: Italian, French, Spanish, and German.

4.2 Evaluation indices

1. *ACC*: It determines which data point has the highest rate of accurate assignment to the correct cluster. Given that f_i represents the dataset x_i's actual label and g_i represents the

[5] http://mlg.ucd.ie/datasets/bbc.html.

[6] https://linqs.soe.ucsc.edu/data.

[7] http://lig-membres.imag.fr/grimal/data.html.

[8] http://ligmembres.imag.fr/grimal/data.html.

algorithm's label, the ACC can be computed as follows:

$$\mathrm{ACC} = \frac{\sum_{i=1}^{n} \rho(f_i, \mathrm{map}(g_i))}{n_d} \tag{23}$$

If the indicator function is defined by $\rho(x, y)$, the total number of points is n_d, and the mapping function $\mathrm{map}(g_i)$ is used to assess the clustering label in order to establish true labels.

2. *NMI*: It employ for collaborative analysis comparing the truth label of the dataset with the label generated by the proposed method.
   Given the actual label set $\Omega = \{S_1, S_2, ..., S_c\}$ and the clustering's label $\Omega' = \{S'_1, S'_2, ..., S'_k\}$, let m_i and m'_i represent the data points in clusters S_i and S'_i, respectively, and m_{xy} denotes the data points in the intersection of clusters S_x and S_y, the NMI between Ω and Ω' is computed as follows:

$$\mathrm{NMI} = \frac{\sum_{x=1}^{c} \sum_{y=1}^{k} \log\left(\frac{mm_{xy}}{m_x m'_y}\right)}{\sqrt{\left(\sum_{x=1}^{c} m_x \log \frac{m_x}{m}\right)\left(\sum_{y=1}^{k} m'_t \log \frac{m'_t}{m}\right)}} \tag{24}$$

3. *F-score*: The harmonic mean of the recall and precision is used to get the F-score. The definition of the calculation equation is:

$$F\mathrm{-score} = \frac{2 \times P_n \times R_l}{P_n + R_l} \tag{25}$$

   where P_n defines as precision and R_l defines as recall.

4.3 Baseline methods

We evaluate the CIWCFMvC against the existing techniques. Below is a summary of the techniques that have been compared in detail.

- **MIC** [14]: For each incomplete view, the average feature values are filled in to address missing occurrences using the MIC approach. It then tackles this problem by applying $L_{2,1}$-Norm regularization and weighted NMF.
- **DAIMC** [25]: Considering both basis matrix alignment and instances aligned, DAIMC aims to obtain a common latent feature matrix for all perspectives. It presents a corresponding weight matrix for every incomplete view, giving each view's supplied instances one weight and its missing instances zero.
- **OMVC** [15]: OMVC enforces sparsity in the acquired latent feature matrices through lasso regularization, thereby enhancing resilience to noise and outliers. Noteworthy is OMVC's memory efficiency, as it circumvents the need to store the entire data matrix, resulting in reduced space complexity. The method processes data incrementally, simultaneously learning latent features and updating the basis matrix.
- **OPIMC** [16]: OPIMC tackles the challenge of large-scale incomplete multi-view clustering by incorporating information about missing instances through weighted and matrix factorization. It introduces two global statistics that facilitate direct clustering outcomes and effectively determine the conclusion of the iteration process.

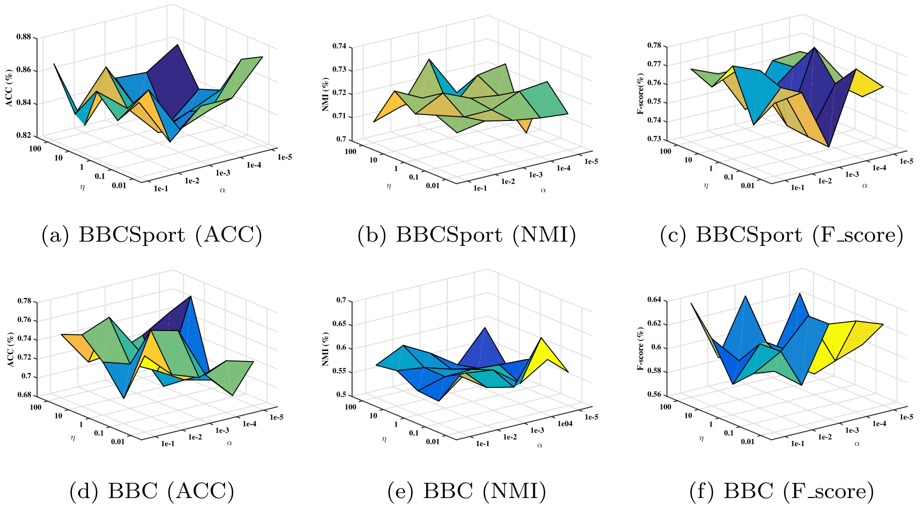
- **CFTIMC** [17]: The common local graph is learned from the completed multiple views by CFTIMC, which models the inter-view alignment relation to infer the missing samples. Lastly, CFTIMC generates the spectral embedding for k-means clustering using the common local graph.
- **GIMC-FLSD** [24]: With the help of local graph regularization, GIMC-FLSD determines the common representation from imperfect data and gives each view a learnable weight.
- **UEAF** [41]: By using dimension graph regularization, UEAF assures that missing data are recovered, treating them as errors. Through the use of reserve graph constraints, it additionally guarantees the consensus structure of completed views. The multi-view data that UEAF generates are then used to extract a common representation.
- **IMC-LRAGR** [46]: To build graphs that capture both global and local data structures, the suggested approach combines non-negative restrictions with distance regularization terms inside low-rank representations. The low-dimensional representation of the graph is then obtained by using spectral clustering.
- **EEOMVC** [47]: This method creates low-dimensional latent features, makes a single partition representation, and breaks down larger similarity graphs from anchor graphs for every view. The binary indicator matrix is directly generated via a label discretization process. Clustering results are improved by the method by combining latent information fusion and clustering into a unified framework.
- **EERIMVC** [49]: This technique presents a regularization technique to enhance the effectiveness of clustering in spite of missing data. The technique generates a single clustering result by combining data from all accessible views, even if some views are lacking.
- **UOMvSC** [50]: In this method, the unified graph is produced by utilizing the relationship between the graph and the inner product of the embedding matrix. Information from every view is combined into one single graph. It is a one-step technique where the clustering labels are obtained directly from this unified network.

4.4 Parameter study

This section analyses the sensitivity of the manually adjusted parameters α, β (=γ), and η under average clustering performance. The parameter α is chosen as $\left\{1e-1, 1e-2, 1e-3, 1e-4, 1e-5\right\}$, β is selected as {10, 20, 30, 40, 50}, and η from {0.01, 0.1, 1, 10, 100}. The performance evaluation on variable values of η and α is shown in Fig. 1 and for β is discussed in Fig. 2. These figures clearly show that the proposed method maintains consistent performance across a diverse set of parameters. These experiments provide strong evidence for the robustness of the proposed methods against parameter variations.

4.5 Convergence study

The objective function meets the convergence for a missing rate of 0.1, 0.3, and 0.5 in Fig. 3. It is noteworthy to emphasize that the method optimizes the given function while continuously meeting the convergence requirements. Our method finds the most optimized values for the variables through iterative updates. The function's values steadily decline as the number of repetition rises, finally attaining convergence after 30 iterations, according to analysis of Fig. 3.



(a) BBCSport (ACC)

(b) BBCSport (NMI)

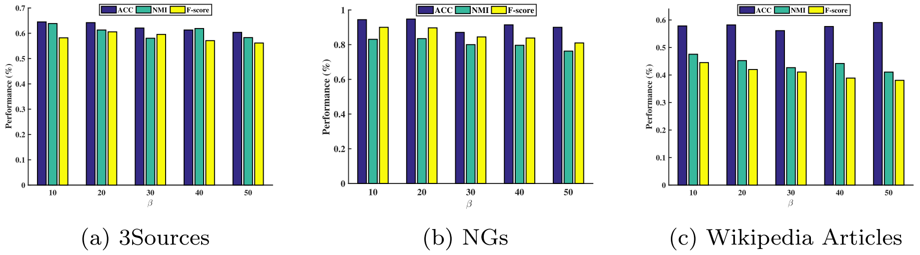
(c) BBCSport (F_score)

(d) BBC (ACC)

(e) BBC (NMI)

(f) BBC (F_score)

Fig. 1 Parameter sensitivity on the compared datasets

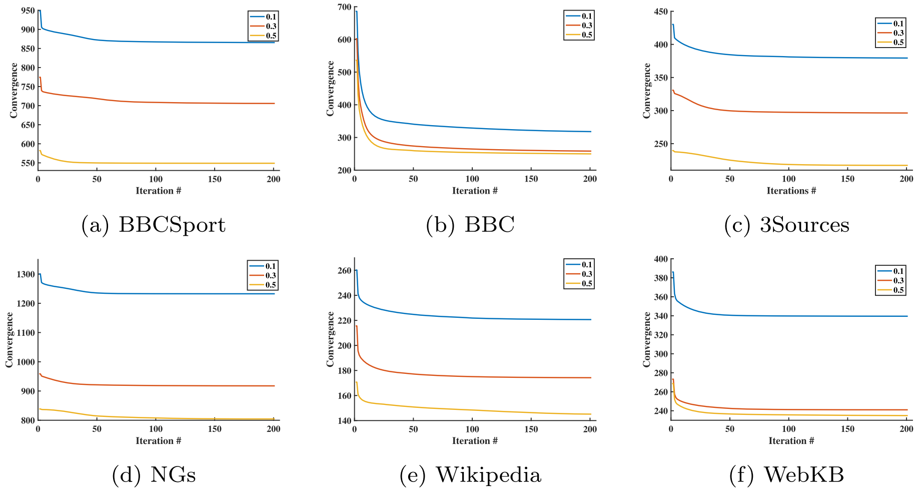


(a) 3Sources

(b) NGs

(c) Wikipedia Articles

Fig. 2 Parameter sensitivity on the compared datasets



(a) BBCSport

(b) BBC

(c) 3Sources

(d) NGs

(e) Wikipedia

(f) WebKB

Fig. 3 Convergence rate of the benchmark datasets on 0.1, 0.3, 0.5 missing rates

4.6 Experiment results

The proposed approach's clustering performance will be assessed using commonly employed metrics such as F-score, ACC, and NMI. The corresponding results are presented in Tables 2, 3, and 4, with bold numbers highlighting the top performances. Drawing conclusions from the evaluated performance, we arrive at the following findings:

To replace missing values in each data view with the matching average value, MIC used a weighted NMF method. It did not, however, outperform the recommended technique, demonstrating the superiority and efficacy of our suggested approach in improving performance. DAIMC outperformed other techniques when it came to clustering performance on the Wikipedia dataset. Likewise, our suggested approach produced better assessment outcomes on other datasets, verifying the efficacy of our technique.

Even though OMVC handled missing multi-view data, on average it performed the worst out of all the algorithms that were compared. On the other hand, our suggested approach demonstrated better clustering performance and handled incompleteness in multi-view data, achieving over 70% performance on datasets like BBCSport, NGs, and BBC.

Using NMF and Frobenius norm, OPIMC obtained the second-best results for all metrics. Alternatively, our approach demonstrated the best clustering performance when compared to other methods by utilizing weighted idea factorization and a co-regularization expression to create the common consensus matrix.

Average clustering performance was achieved by CFTIMC by combining the NMF approach with common latent subspace and manifold learning. On the other hand, our weighted idea factorization-based approach demonstrated better clustering performance in all of the datasets that were evaluated. While GIMC_FLSD improved over IMG in addressing missing instances, it was not able to outperform our suggested approach, which outperformed other state-of-the-art methods by achieving over 45% average performance across all criteria.

Similar to our approach, UEAF sought to remedy missing instances; however, it did not outperform it. Comparing our approach to other state-of-the-art techniques, it showed an average performance of over 50% across all parameters, demonstrating its efficacy in filling missing instances. IMC_LRAGR produced average clustering performance by combining the NMF approach with common latent subspace and manifold learning. By using weighted idea factorization, on the other hand, our suggested approach showed better clustering performance on all of the comparable datasets.

For the majority of datasets, the EEOMVC, EERIMVC, and UOMvSC algorithms perform better when clustering multi-view data. However, across all datasets, our suggested algorithm outperforms the competition and shows the best results..

In summary, our approach performs better than the current methods on real-world datasets, as shown by the comparison between Tables 2, 3 and 4. This highlights the superior performance of our method, which leverages a smooth regularization term to reduce over-fitting problems between views and a co-regularization term to reveal the shared consensus structure in the data.

5 Conclusion

In this study, we explore the challenge of dealing with incomplete views in multi-view clustering, where each view is affected by the absence of certain instances. By using the weighted concept factorization theory, which reduces disagreement between many viewpoints and

Table 2 Average and Standard Deviation of ACC (%) of different approaches

Datasets	Missing values (%)	MIC	DAIMC	OMVC	OPIMC	CFTIMC	GIMC_FLSD
BBCSport	0.1	0.7345(0.0092)	0.7665(0.0183)	0.5678(0.0234)	0.4567(0.0045)	0.3784(0.0224)	0.7123(0.0132)
	0.3	0.6734(0.0321)	0.6892(0.0216)	0.6056(0.0461)	0.4871(0.0231)	0.3456(0.0112)	0.7243(0.0112)
	0.5	0.6143(0.0090)	0.6121(0.0228)	0.4673(0.0219)	0.4476(0.0234)	0.3987(0.0134)	0.6854(0.0254)
BBC	0.1	0.5671(0.0034)	0.6178(0.0217)	0.4891(0.0012)	0.4567(0.0001)	0.7092(0.0234)	**0.8242(0.0213)**
	0.3	0.5134(0.0123)	0.5894(0.0342)	0.4671(0.0143)	0.4156(0.0004)	0.6623(0.0198)	**0.7675(0.0167)**
	0.5	0.4978(0.0081)	0.5167(0.0215)	0.4256(0.0156)	0.3982(0.0001)	0.6182(0.0216)	0.7176(0.0219)
3Sources	0.1	0.5534(0.0067)	0.5391(0.0023)	0.4456(0.0076)	0.5739(0.0056)	0.4365(0.0231)	0.6473(0.0023)
	0.3	0.5009(0.0012)	0.5145(0.0067)	0.4178(0.0111)	0.5207(0.0023)	0.6134(0.0461)	0.6156(0.0198)
	0.5	0.4651(0.0012)	0.4078(0.0045)	0.3787(0.0045)	0.4987(0.0167)	0.5764(0.0156)	0.5767(0.0321)
NGs	0.1	0.6089(0.0075)	0.8154(0.0033)	0.4765(0.0001)	0.5989(0.0045)	0.5987(0.0341)	0.8656(0.0289)
	0.3	0.5672(0.0127)	0.7945(0.0034)	0.4467(0.0009)	0.5711(0.0012)	0.5534(0.0112)	0.8156(0.0214)
	0.5	0.5327(0.0119)	0.7534(0.0056)	0.4156(0.0023)	0.5234(0.0010)	0.3891(0.0119)	0.7774(0.0263)
Wikipedia Articles	0.1	0.4876(0.0187)	0.5716(0.0111)	0.4456(0.0178)	0.5298(0.0114)	0.5078(0.0231)	0.5467(0.0035
	0.3	0.4634(0.0191)	0.5478(0.0173)	0.4138(0.0139)	0.4982(0.0034)	0.4327(0.0110)	0.4971(0.0342)
	0.5	0.4134(0.0101)	0.5034(0.0147)	0.3789(0.0169)	0.4267(0.0169)	0.3879(0.0132)	0.4999(0.0231)
WebKB	0.1	0.4467(0.0134)	0.6410(0.0012)	0.6738(0.0110)	0.6934(0.0045)	0.4767(0.0276)	–
	0.3	0.4318(0.0178)	0.6036(0.0113)	0.6234(0.0111)	0.6454(0.0018)	0.3987(0.0245)	–
	0.5	0.4115(0.0109)	0.5819(0.0137)	0.5981(0.0161)	0.6278(0.0051)	0.4351(0.0152)	–
Citeseer	0.1	0.4778(0.0324)	0.4598(0.0256)	0.4423(0.0045)	0.3271(0.0124)	0.4981(0.0167)	0.5127(0.0026)
	0.3	0.4434(0.0034)	0.4271(0.0124)	0.4110(0.0067)	0.3025(0.0111)	0.4632(0.0067)	0.4871(0.0173)

Table 2 continued

Datasets	Missing values (%)	MIC	DAIMC	OMVC	OPIMC	CFTIMC	GIMC_FLSD
	0.5	0.4154(0.0045)	0.4003(0.0117)	0.3987(0.0034)	0.2789(0.0152)	0.4334(0.0034)	0.4471(0.0235)
Reuters	0.1	**0.4278(0.0218)**	0.4134(0.0254)	0.4254(0.0016)	0.3876(0.0135)	0.2867(0.0236)	0.3091(0.0178)
	0.3	**0.4012(0.0156)**	0.3981(0.0067)	0.4008(0.0116)	0.3523(0.0032)	0.2753(0.0193)	0.2854(0.0193)
	0.5	**0.3876(0.0034)**	0.3687(0.0234)	0.4187(0.0352)	0.3423(0.0176)	0.2554(0.0024)	0.2667(0.0265)

Datasets	Missing values (%)	UEAF	IMC-LRAGR	EEOMVC	EERIMVC	UOMvSC	Proposed method
BBCSport	0.1	0.7845(0.0222)	0.7843(0.0243)	0.7834(0.0217)	0.7981(0.0181)	0.7956(0.0002)	**0.8465(0.0012)**
	0.3	0.7723(0.0178)	0.7687(0.0217)	0.7435(0.0221)	0.7634(0.0118)	0.7745(0.0021)	**0.8134(0.0123)**
	0.5	0.6581(0.0261)	0.7423(0.0199)	0.7234(0.0211)	0.7323(0.0121)	0.7435(0.0018)	**0.7981(0.0034)**
BBC	0.1	0.7581(0.0231)	0.7435(0.0312)	0.7881(0.0213)	0.7643(0.0222)	0.7582(0.0031)	0.7681(0.0087)
	0.3	0.7265(0.0217)	0.7289(0.0231)	0.7443(0.0216)	0.7401(0.0189)	0.7311(0.0031)	0.7356(0.0023)
	0.5	0.7076(0.0217)	0.7088(0.0291)	0.7014(0.0185)	0.7256(0.0267)	0.7003(0.0091)	**0.7234(0.0012)**
3Sources	0.1	0.5763(0.0213)	0.6634(0.0125)	0.6534(0.0241)	0.6767(0.0169)	0.6423(0.0023)	**0.6956(0.0213)**
	0.3	0.5234(0.0156)	0.6325(0.0222)	0.6226(0.0127)	0.6327(0.0217)	0.6223(0.0015)	**0.6587(0.0167)**
	0.5	0.5367(0.0217)	0.6123(0.0295)	0.6172(0.0209)	0.6098(0.0261)	0.6057(0.0089)	**0.6281(0.0231)**
NGs	0.1	0.8534(0.0231)	0.8435(0.0214)	0.8178(0.0121)	0.8367(0.0251)	0.7954(0.0045)	**0.8734(0.0023)**
	0.3	0.8156(0.0345)	0.8134(0.0332)	0.7834(0.0345)	0.8003(0.0216)	0.7663(0.0012)	**0.8434(0.0045)**
	0.5	0.7853(0.0245)	0.7634(0.0316)	0.7663(0.0211)	0.7724(0.0211)	0.7225(0.0018)	**0.7878(0.0011)**
Wikipedia Articles	0.1	0.5678(0.0245)	0.5536(0.0117)	0.5453(0.0210)	0.5661(0.0231)	0.5611(0.0009)	**0.5865(0.0132)**
	0.3	0.5123(0.0034)	0.5234(0.0178)	0.5221(0.0281)	0.5221(0.0193)	0.5178(0.0001)	**0.5398(0.0119)**
	0.5	0.5119(0.0213)	0.5009(0.0221)	0.4999(0.0222)	0.5001(0.0222)	0.4934(0.0003)	**0.5166(0.0034)**

Table 2 continued

Datasets	Missing values (%)	UEAF	IMC-LRAGR	EEOMVC	EERIMVC	UOMvSC	Proposed method
WebKB	0.1	0.6371(0.0251)	0.6545(0.0233)	0.6634(0.0121)	0.6834(0.0201)	–	**0.7084(0.0156)**
	0.3	0.5734(0.0165)	0.6423(0.0129)	0.6321(0.0213)	0.6505(0.0189)	–	**0.6639(0.0162)**
	0.5	0.5167(0.0265)	0.6356(0.0265)	0.6115(0.0290)	0.6289(0.0319)	–	**0.6345(0.0187)**
Citeseer	0.1	0.3718(0.0118)	0.3145(0.0000)	0.3881(0.0098)	0.5219(0.0017)	0.4218(0.0102)	**0.6134(0.0173)**
	0.3	0.3522(0.0043)	0.3027(0.0119)	0.3665(0.0023)	0.4999(0.0197)	0.4016(0.0024)	**0.5934(0.0110)**
	0.5	0.3215(0.0326)	0.2777(0.0287)	0.3221(0.0278)	0.4513(0.0218)	0.3792(0.0018)	**0.5545(0.0223)**
Reuters	0.1	0.3156(0.0111)	0.3626(0.0000)	0.2956(0.0089)	0.2967(0.0087)	0.3378(0.0298)	0.4189(0.0215)
	0.3	0.2963(0.0067)	0.3441(0.0271)	0.2771(0.0342)	0.2653(0.0314)	0.3162(0.0220)	0.3956(0.0098)
	0.5	0.2724(0.0215)	0.3281(0.0165)	0.2573(0.0216)	0.2745(0.0111)	0.3045(0.0134)	0.3745(0.0352)

Table 3 Average and standard deviation of NMI (%) of different approaches

Datasets	Missing values (%)	MIC	DAIMC	OMVC	OPIMC	CFTIMC	GIMC_FLSD
BBCSport	0.1	0.4278(0.0125)	0.6934(0.0223)	0.2723(0.0116)	0.4287(0.0123)	0.2356(0.0323)	0.7201(0.0024)
	0.3	0.3709(0.0109)	0.6609(0.0023)	0.2689(0.0222)	0.3767(0.0267)	0.1675(0.0235)	0.6443(0.0423)
	0.5	0.3989(0.0111)	0.6023(0.0217)	0.2132(0.0134)	0.3156(0.0193)	0.2686(0.0308)	0.5654(0.0324)
BBC	0.1	0.4067(0.0213)	0.4876(0.0213)	0.3376(0.0216)	0.2996(0.0123)	0.5987(0.0225)	**0.6456(0.0123)**
	0.3	0.3709(0.0106)	0.4235(0.0218)	0.2515(0.0213)	0.2763(0.0193)	0.5585(0.0337)	**0.5867(0.0234)**
	0.5	0.3178(0.0167)	0.3546(0.0023)	0.2289(0.0189)	0.2191(0.0181)	0.2004(0.0368)	**0.5634(0.0156)**
3Sources	0.1	0.4424(0.0178)	0.5521(0.0189)	0.3467(0.0218)	0.4436(0.0217)	0.2845(0.0367)	0.6134(0.0265)
	0.3	0.4123(0.0222)	0.5234(0.0198)	0.3376(0.0243)	0.3934(0.0254)	0.4836(0.0373)	0.5272(0.0231)
	0.5	0.3536(0.0219)	0.4903(0.0234)	0.3765(0.287)	0.4152(0.0187)	0.4928(0.0330)	0.5643(0.0267)
NGs	0.1	0.0834(0.0213)	0.7854(0.0276)	0.0526(0.0321)	0.2865(0.0167)	0.2584(0.0245)	0.8095(0.0231)
	0.3	0.0634(0.0278)	0.7534(0.0139)	0.0546(0.0356)	0.1945(0.0216)	0.5271(0.0227)	0.6873(0.0123)
	0.5	0.0698(0.0265)	**0.7345(0.0245)**	0.0509(0.0301)	0.1723(0.0117)	0.3986(0.0317)	0.4671(0.0222)
Wikipedia Articles	0.1	0.3987(0.0136)	**0.5123(0.0148)**	0.3216(0.0191)	0.4561(0.0227)	0.4156(0.0110)	0.3412(0.0254)
	0.3	0.3847(0.0149)	**0.4778(0.0218)**	0.2952(0.0117)	0.3771(0.0246)	0.3782(0.0318)	0.3145(0.0234)
	0.5	0.3306(0.0103)	**0.4528(0.0278)**	0.2387(0.0137)	0.3217(0.0324)	0.3555(0.0019)	0.2267(0.0156)
WebKB	0.1	0.1762(0.0013)	0.2765(0.0171)	0.2167(0.0218)	0.2541(0.0167)	0.3072(0.0265)	–
	0.3	0.1572(0.0178)	0.2451(0.0281)	0.1598(0.0119)	0.2357(0.0194)	0.2994(0.0262)	–
	0.5	0.1634(0.0116)	0.2171(0.0119)	0.1067(0.0257)	0.1771(0.0163)	0.2573(0.0317)	–
Citeseer	0.1	0.2678(0.0254)	0.2934(0.0218)	0.2817(0.0081)	0.2019(0.0125)	0.2987(0.0109)	0.2811(0.0132)
	0.3	0.2524(0.0216)	0.2791(0.0156)	0.2634(0.0118)	0.1934(0.0216)	0.2716(0.0118)	0.2416(0.0156)
	0.5	0.2381(0.0217)	0.2617(0.0310)	0.2381(0.0318)	0.1723(0.0218)	0.2589(0.0193)	0.2258(0.0208)

Table 3 continued

Datasets	Missing values (%)	MIC	DAIMC	OMVC	OPIMC	CFTIMC	GIMC_FLSD
Reuters	0.1	0.2345(0.0265)	0.3156(0.0209)	0.2817(0.0081)	0.2414(0.0193)	0.1989(0.0118)	0.2582(0.0078)
	0.3	0.2134(0.0234)	0.2845(0.0165)	0.2665(0.0218)	0.2376(0.0162)	0.1754(0.0222)	0.2381(0.0265)
	0.5	0.2009(0.0216)	0.2672(0.0234)	0.2381(0.0218)	0.2005(0.0242)	0.1543(0.0324)	0.2131(0.0228)
Datasets	Missing values (%)	UEAF	IMC-LRAGR	EEOMVC	EERIMVC	UOMvSC	Proposed method
BBCSport	0.1	0.6984(0.0234)	0.7067(0.0221)	0.7165(0.0187))	0.7067(0.0024)	0.6965(0.0076)	**0.7335(0.0213)**
	0.3	0.6393(0.0231)	0.6534(0.0332)	0.6745(0.0125)	0.6634(0.0091)	0.6723(0.0023)	**0.6834(0.0111)**
	0.5	0.5087(0.0431)	0.6111(0.0218)	0.6235(0.0111)	0.6209(0.0023)	0.6406(0.0023)	**0.6526(0.0218)**
BBC	0.1	0.6156(0.0345)	0.6245(0.0221)	0.6221(0.0023)	0.6117(0.0076)	0.5981(0.0218)	0.6045(0.0231)
	0.3	0.5876(0.0234)	0.5773(0.0200)	0.5835(0.0132)	0.5734(0.0013)	0.5627(0.0182)	0.5612(0.0134)
	0.5	0.5187(0.0314)	0.5226(0.0201)	0.5164(0.0036)	0.5218(0.0009)	0.5234(0.0123)	0.5345(0.0243)
3Sources	0.1	0.5987(0.0278)	0.5956(0.0167)	0.5884(0.0034)	0.5745(0.0023)	0.5812(0.0067)	**0.6157(0.0167)**
	0.3	0.5318(0.0216)	0.5529(0.0111)	0.5436(0.0065)	0.5421(0.0013)	0.5534(0.0047)	**0.5678(0.0114)**
	0.5	0.4456(0.0112)	0.5335(0.0111)	0.5234(0.0019)	0.5242(0.0012)	0.5138(0.0018)	**0.5813(0.0217)**
NGs	0.1	0.8154(0.0115)	0.7834(0.0221)	0.7756(0.0034)	0.7676(0.0013)	0.7456(0.0023)	**0.8214(0.0213)**
	0.3	0.7645(0.0216)	0.7423(0.0319)	0.7381(0.0019)	0.7456(0.0018)	0.7145(0.0046)	**0.7799(0.0132)**
	0.5	0.6945(0.0421)	0.7034(0.0119)	0.6984(0.0056)	0.7045(0.0045)	0.6834(0.0231)	0.7134(0.0032)
Wikipedia Articles	0.1	0.3078(0.0289)	0.4563(0.0229)	0.4456(0.0112)	0.4551(0.0123)	0.4325(0.0167)	0.4756(0.0189)
	0.3	0.2345(0.0326)	0.4367(0.0391)	0.4203(0.0013)	0.4381(0.0078)	0.4182(0.0109)	0.4572(0.0214)
	0.5	0.1376(0.0318)	0.4003(0.0299)	0.3982(0.0027)	0.4111(0.0118)	0.3982(0.0098)	0.4236(0.0167)
WebKB	0.1	0.3271(0.0024)	0.4334(0.0123)	0.4045(0.0198)	0.4201(0.0012)	–	**0.4345(0.0143)**
	0.3	0.2715(0.0234	0.3781(0.0281)	0.3764(0.0271)	0.3991(0.0076)	–	**0.3861(0.0276)**
	0.5	0.3172(0.0118)	0.3444(0.0201)	0.3325(0.0181)	0.3712(0.0067)	–	**0.3581(0.0067)**

Table 3 continued

Datasets	Missing values (%)	UEAF	IMC-LRAGR	EEOMVC	EERIMVC	UOMvSC	Proposed method
Citeseer	0.1	0.2001(0.0098)	0.1281(0.0000)	0.1634(0.0101)	0.2091(0.0115)	0.2581(0.0078)	**0.3323(0.0156)**
	0.3	0.1881(0.0215)	0.1118(0.0293)	0.1472(0.0317)	0.1992(0.0209)	0.2381(0.0156)	**0.3076(0.0256)**
	0.5	0.1523(0.0318)	0.0924(0.0318)	0.1287(0.0321)	0.1638(0.0278)	0.1782(0.0118)	**0.2767(0.0221)**
Reuters	0.1	0.2976(0.0000)	0.2265(0.0209)	0.2634(0.0023)	0.2317(0.0090)	0.2367(0.0197)	**0.3293(0.0213)**
	0.3	0.2773(0.0129)	0.2134(0.0215)	0.2493(0.0273)	0.2185(0.0218)	0.2006(0.0312)	**0.3052(0.0217)**
	0.5	0.2576(0.0262)	0.1945(0.0215)	0.2334(0.0283)	0.1972(0.0263)	0.2041(0.0203)	**0.2745(0.0316)**

Table 4 Average and standard deviation of F_score (%) of different approaches

Datasets	Missing values (%)	MIC	DAIMC	OMVC	OPIMC	CFTIMC	GIMC_FLSD
BBCSport	0.1	0.4567(0.0123)	0.7491(0.0023)	0.3891(0.0111)	0.5067(0.0109)	0.6456(0.0234)	0.6953(0.0245)
	0.3	0.4381(0.0217)	0.7102(0.0143)	0.3516(0.0178)	0.5237(0.0189)	0.6578(0.0197)	0.6356(0.0167)
	0.5	0.4456(0.0016)	0.6732(0.0156)	0.3199(0.0241)	0.4723(0.0098)	0.5578(0.0098)	0.5467(0.0345)
BBC	0.1	0.4561(0.0132)	0.6098(0.0231)	0.4156(0.0213)	0.3892(0.0165)	0.6345(0.0314)	0.6145(0.0342)
	0.3	0.4099(0.0154)	0.5067(0.0147)	0.3817(0.0143)	0.4327(0.0167)	0.6098(0.0345)	0.5346(0.0213)
	0.5	0.3876(0.0265)	0.3467(0.0178)	0.3078(0.0112)	0.3954(0.0265)	0.5892(0.0119)	0.4476(0.0178)
3Sources	0.1	0.3978(0.0262)	0.5073(0.0321)	0.3654(0.0178)	0.4871(0.0267)	0.5378(0.0345)	0.6243(0.0216)
	0.3	0.4139(0.0111)	0.4723(0.0165)	0.3245(0.0287)	0.5423(0.0167)	0.5987(0.0231)	0.5673(0.0286)
	0.5	0.3627(0.0218)	0.4616(0.0265)	0.3189(0.0219)	0.4523(0.0178)	0.4465(0.0019)	0.5167(0.0234)
NGs	0.1	0.5812(0.0167)	0.8234(0.0231)	0.3524(0.0194)	0.3762(0.0218)	0.8865(0.0345)	0.8754(0.0234)
	0.3	0.5581(0.0209)	0.7745(0.0109)	0.3267(0.0231)	0.3221(0.0189)	0.7834(0.0234)	0.7475(0.0415)
	0.5	0.5123(0.0222)	0.7534(0.0161)	0.3218(0.0328)	0.3125(0.0188)	0.6345(0.0167)	0.6378(0.0291)
Wikipedia Articles	0.1	0.3676(0.0211)	**0.4977(0.0219)**	0.3113(0.0217)	0.4315(0.0219)	0.4153(0.0234)	0.4089(0.0421)
	0.3	0.3425(0.0116)	**0.4423(0.0177)**	0.2776(0.0281)	0.3718(0.0138)	0.4089(0.0234)	0.3511(0.0118)
	0.5	0.3075(0.0297)	**0.3918(0.0210)**	0.2232(0.0101)	0.3178(0.0205)	0.3546(0.0215)	0.3486(0.0217)
WebKB	0.1	0.4213(0.0165)	0.5216(0.0111)	0.5776(0.0219)	0.6116(0.0209)	0.6234(0.0119)	–
	0.3	0.4178(0.0169)	0.5543(0.0243)	0.5021(0.0111)	0.5782(0.0163)	0.5467(0.0342)	–
	0.5	0.4436(0.0239)	0.5317(0.0176)	0.4637(0.0194)	0.6027(0.0199)	0.4456(0.0190)	–
Citeseer	0.1	0.3869(0.0189)	0.3617(0.0201)	0.3156(0.0209)	0.2767(0.0101)	0.3312(0.0197)	0.3516(0.0193)
	0.3	0.3671(0.0034)	0.3567(0.0217)	0.3072(0.0324)	0.2554(0.0231)	0.3256(0.0165)	0.3371(0.0216)
	0.5	0.3325(0.0215)	0.3215(0.0216)	0.2734(0.0167)	0.2371(0.0216)	0.2978(0.0203)	0.3111(0.0253)

Table 4 continued

Datasets	Missing values (%)	MIC	DAIMC	OMVC	OPIMC	CFTIMC	GIMC_FLSD
Reuters	0.1	0.3256(0.0210)	**0.3367(0.0212)**	0.3125(0.0214)	0.3208(0.0035)	0.2156(0.0205)	0.2834(0.0276)
	0.3	0.2983(0.0202)	**0.3145(0.0251)**	0.2976(0.0289)	0.3056(0.0222)	0.2065(0.0245)	0.2881(0.0298)
	0.5	0.2781(0.0289)	**0.3067(0.0291)**	0.2864(0.0217)	0.2863(0.0111)	0.1943(0.0234)	0.2654(0.0314)

Datasets	Missing values (%)	UEAF	EEOMVC	EERIMVC	UOMvSC	Proposed method
BBCSport	0.1	0.7778(0.0231)	0.7645(0.0215)	0.7523(0.0055)	0.7413(0.0034)	**0.7812(0.0012)**
	0.3	0.7363(0.0234)	0.7455(0.0178)	0.7310(0.0193)	0.7167(0.0012)	**0.7523(0.0167)**
	0.5	0.6934(0.0345)	0.7124(0.0252)	0.7045(0.034)	0.6954(0.0054)	**0.7281(0.0093)**
BBC	0.1	0.6024(0.0231)	0.6445(0.0217)	0.6569(0.0081)	0.6456(0.0342)	**0.6753(0.0231)**
	0.3	0.5234(0.0256)	0.6234(0.0023)	0.6134(0.0056)	0.6124(0.0012)	**0.6427(0.0211)**
	0.5	0.4376(0.0318)	0.6013(0.0012)	0.5974(0.0034)	0.5964(0.0103)	**0.6123(0.0222)**
3Sources	0.1	0.5976(0.0237)	0.6172(0.0128)	0.6216(0.0067)	0.6145(0.0003)	**0.6445(0.0215)**
	0.3	0.5342(0.0067)	0.5823(0.0116)	0.5913(0.0021)	0.5991(0.0003)	**0.6115(0.0145)**
	0.5	0.4457(0.0316)	0.5467(0.0019)	0.5678(0.0017)	0.5613(0.0015)	**0.5695(0.0289)**
NGs	0.1	0.8745(0.0234)	0.8678(0.0121)	0.8436(0.0121)	0.8534(0.0013)	**0.9034(0.0123)**
	0.3	0.7745(0.0236)	0.8245(0.0191)	0.8156(0.0117)	0.8115(0.0083)	**0.8554(0.0163)**
	0.5	0.6985(0.0034)	0.7834(0.0003)	0.7935(0.0110)	0.7954(0.0018)	**0.8127(0.0287)**
Wikipedia Articles	0.1	0.4367(0.0342)	0.4145(0.0093)	0.4076(0.0231)	0.4272(0.0101)	0.4408(0.0231)
	0.3	0.3896(0.0034)	0.3998(0.0117)	0.3867(0.0146)	0.4113(0.0131)	0.4289(0.0341)
	0.5	0.3489(0.0129)	0.3678(0.0046)	0.3625(0.0012)	0.3765(0.0015)	0.3887(0.0111)
WebKB	0.1	0.6345(0.0341)	0.6345(0.0023)	0.6434(0.0134)	–	**0.6631(0.0203)**
	0.3	0.5563(0.0453)	0.5987(0.0012)	0.6127(0.0111)	–	**0.6113(0.0189)**
	0.5	0.4986(0.0256)	0.5734(0.0073)	0.5934(0.0181)	–	**0.6274(0.0218)**

Table 4 continued

Datasets	Missing values (%)	UEAF	EEOMVC	EERIMVC	UOMvSC	Proposed method
Citeseer	0.1	0.2916(0.0019)	0.3016(0.0000)	0.2706(0.0023)	0.2871(0.0034)	**0.4012(0.0098)**
	0.3	0.2854(0.0126)	0.2853(0.0098)	0.2554(0.0119)	0.2664(0.0205)	**0.3879(0.0163)**
	0.5	0.2665(0.0217)	0.2554(0.0218)	0.2338(0.0111)	0.2445(0.0312)	**0.3427(0.0211)**
Reuters	0.1	0.2251(0.0119)	0.2845(0.0000)	0.2481(0.0071)	0.2891(0.0132)	0.3311(0.0065)
	0.3	0.2076(0.0261)	0.2782(0.0318)	0.2338(0.02187)	0.2676(0.0209)	0.3082(0.0271)
	0.5	0.1954(0.0245)	0.2581(0.0291)	0.2171(0.0239)	0.2472(0.0218)	0.2873(0.0229)

a common consensus matrix in addition to utilizing matrix factorization, the CIWCFMvC model is proposed. Moreover, the weight of the view is automatically adjusted throughout the optimization process. Lastly, the innovative iterative technique is used to maximize the suggested objective function of the CIWCFMvC. Comprehensive tests on benchmark datasets confirm that the CIWCFMvC is better than the current methods.

Acknowledgements The authors would like to express their gratitude to AlMaarefa University, Saudi Arabia, for supporting this research.

Author Contributions Ghufran Ahmad Khan was involved in the conception and design of study. Ghufran Ahmad Khan and Jalaluddin Khan acquired and curated the data. Ghufran Ahmad Khan and Diallo Bassoma performed the analysis and/or interpretation of data. Ghufran Ahmad Khan and Naved Ahmad drafted the manuscript. Zail Al-Huda and Taushif Anwar contributed to critical revision.

Data availability No datasets were generated or analyzed during the current study.

Declarations

Conflict of interest The authors declare no conflict of interest.

References

1. Diallo B, Hu J, Li T, Khan GA, Liang X, Wang H (2023) Auto-attention mechanism for multi-view deep embedding clustering. Pattern Recognit 143:109764
2. Khan GA, Hu J, Li T, Diallo B, Wang H (2023) Multi-view subspace clustering for learning joint representation via low-rank sparse representation. Appl Intell 53:1–20
3. Chen M, Huang L, Wang C, Huang D (2020) Multi-view clustering in latent embedding space. In: Proceedings of AAAI conference on artificial intelligence, pp 3513-3520
4. Chaudhuri K, Kakade SM, Livescu K, Sridharan K (2009) Multi-view clustering via canonical correlation analysis. In: Proceedings of international conference on machine learning, pp 129-136
5. Peng X, Huang Z, Lv J, Zhu H, Zhou JT (2019) COMIC: multi-view clustering with- out parameter selection. In: Proceedings of international conference on machine learning, pp 5092-5101
6. Liu J, Wang C, Gao J, Han J (2013) Multi-view clustering via joint non-negative matrix factorization. In: Proceedings of SIAM international conference on data mining, pp 252-260
7. Khan GA, Hu J, Li T, Diallo B, Wang H (2023) Multi-view clustering for multiple manifold learning via concept factorization. Digital Signal Process 140:104118
8. Yin Q, Wu S, Wang L (2017) Unified subspace learning for incomplete and unlabeled multi-view data. Pattern Recogn 67:313–327
9. Liu X, Zhu X, Li M, Wang L, Tang C, Yin J, Shen D, Wang H, Gao W (2019) Late fusion incomplete multi-view clustering. IEEE Trans Pattern Anal Mach Intell 41(10):2410–2423
10. Zong L, Zhang X, Liu X, Yu H (2020) Multi-view clustering on data with partial instances and clusters. Neural Netw 129:19–30
11. Li SY, Jiang Y, Zhou ZH (2014) Partial multi-view clustering. In: Proceedings of AAAI conference on artificial intelligence, pp 1968-1974
12. Xu W, Gong Y (2021) A survey on concept factorization: from shallow to deep rep- resentation learning. Inf Process Manage 5(3):102534
13. Zhao H, Liu H, Fu Y (2016) Incomplete multi-modal visual data grouping. In: Proceedings of international joint conference on artificial intelligence, pp 2392-2398
14. Shao W, He L, Yu Philip S (2015) Multiple incomplete views clustering via weighted non-negative matrix factorization with $l_{2,1}$ regularization. In: Joint European conference on machine learning and knowledge discovery in databases, pp 318–334
15. Shao W, He L, Lu CT, Philip SY (2016) Online multi-view clustering with incomplete views. In: 2016 IEEE International conference on big data, pp 1012-1017
16. Hu M, Chen S (2019) One-pass incomplete multi-view clustering. Proc AAAI Conf Artif Intell 33(01):3838–3845
17. Liang N, Yang Z, Li L, Li Z, Xie S (2021) Incomplete multi-view clustering with cross-view feature transformation. IEEE Trans Artif Intel 3(5):749–762

18. Liu C, Wu Z, Wen J, Xu Y, Huang C (2022) Localized sparse incomplete multi-view clustering. IEEE Trans Multimed 25:1–13
19. Li L, Wan Z, He H (2023) Incomplete multi-view clustering with joint partition and graph learning. IEEE Trans Knowl Data Eng 35(1):589–602
20. Lv Z, Gao Q, Zhang X, Li Q, Yang M (2022) View-consistency learning for incomplete multiview clustering. IEEE Trans Image Process 31:4790–4802
21. Shang M, Liang C, Luo J, Zhang H (2023) Incomplete multi-view clustering by simultaneously learning robust representations and optimal graph structures. Inf Sci 640:119038
22. Wen J, Xu Y, Liu H (2020) Incomplete multiview spectral clustering with adaptive graph learning. IEEE Trans Cybern 50(4):1418–1429
23. Wen J, Zhang Z, Zhang Z, Fei L, Wang M (2021) Generalized incomplete multiview clustering with flexible locality structure diffusion. IEEE Trans Cybern 51(1):101–114
24. Liang N, Yang Z, Xie S (2022) Incomplete multi-view clustering with sample-level auto-weighted graph fusion. IEEE Trans Knowl Data Eng 35:1–7
25. Hu M, Chen S (2018) Doubly aligned incomplete multi-view clustering. In: Proceedings of international joint conference on artificial intelligence, pp 2262–2268
26. Gao H, Peng Y, Jian S (2016) Incomplete multi-view clustering. In: Proceedings of the international conference on intelligent information processing, pp 245-255
27. Liu X, Zhu X, Li M, Tang C, Zhu E, Yin J, Gao W (2019) Efficient and effective incomplete multi-view clustering. In: Proceedings of the AAAI conference on artificial intelligence, pp 4392-4399
28. Liu X, Zhu X, Li M, Wang L, Tang C, Yin J, Shen D, Wang H, Gao W (2019) Late fusion incomplete multi-view clustering. IEEE Trans Pattern Anal Mach Intell 41(10):2410–2423
29. Niu G, Yang Y, Sun L (2021) One-step multi-view subspace clustering with incomplete views. Neurocomputing 438:290–301
30. Zhang C, Cui Y, Han Z, Zhou JT, Fu H, Hu Q (2022) Deep partial multi-view learning. IEEE Trans Pattern Anal Mach Intell 44(5):2402–2415
31. Yin J, Cai R, Sun S (2022) Anchor-based incomplete multi-view spectral clustering. Neurocomputing 514:526–538
32. Yin M, Liu X, Wang L, He G (2023) Learning latent embedding via weighted projection matrix alignment for incomplete multi-view clustering. Inf Sci 634:244–258
33. Xia D, Yang Y, Yang S, Li T (2023) Incomplete multi-view clustering via kernelized graph learning. Inf Sci 625:1–19
34. Yang JH, Fu LL, Chen C, Dai HN, Zheng Z (2023) Cross-view graph matching for incomplete multi-view clustering. Neurocomputing 515:79–88
35. Zhou W, Wang H, Yang Y (2019) Consensus graph learning for incomplete multi-view clustering. In: Proceedings of the advances in knowledge discovery and data mining, pp 529-540
36. Cui J, Fu Y, Huang C, Wen J (2022) Low-rank graph completion-based incomplete multiview clustering. IEEE Trans Neural Netw Learn Syst 35:1–11
37. Yin J, Sun S (2023) Incomplete multi-view clustering with reconstructed views. IEEE Trans Knowl Data Eng 35(3):2671–2682
38. Li Z, Tang C, Zheng X, Liu X, Zhang W, Zhu E (2022) High-order correlation preserved incomplete multi-view subspace clustering. IEEE Trans Image Process 31:2067–2080
39. Jue He W, Zhang Z, Wei Y (2023) Scalable incomplete multi-view clustering with adaptive data completion. Inf Sci 649:119562
40. Liang N, Yang Z, Li L, Li Z, Xie S (2021) Incomplete multi-view clustering with cross-view feature transformation. IEEE Trans Artif Intell 3(5):749–762
41. Wen J, Zhang Z, Xu Y, Zhang B, Fei L, Liu H (2019) Unified embedding alignment with missing views inferring for incomplete multi-view clustering. In: Proceedings of the AAAI conference on artificial intelligence, Vol 33, pp 5393-5400
42. Yin J, Sun S (2022) Incomplete multi-view clustering with cosine similarity. Pattern Recogn 123:108371
43. Chao G, Wang S, Yang S, Li C, Chu D (2022) Incomplete multi-view clustering with multiple imputation and ensemble clustering. Appl Intell 52(13):14811–14821
44. Zhang H, Chen X, Zhang E, Wang L (2023) Incomplete multi-view learning via consensus graph completion. Neural Process Lett 55(4):3923–3952
45. Xia D, Yang Y, Yang S (2022) Incomplete multi-view clustering via auto-weighted fusion in partition space. Tsinghua Sci Technol 28(3):595–611
46. Zhang K, Liu B, Du S, Yu Y, Song J (2023) Incomplete multi-view clustering based on low-rank representation with adaptive graph regularization. Soft Comput 27(11):7131–7146
47. Wang J, Tang C, Wan Z, Zhang W, Sun K, Zomaya AY (2023) Efficient and effective one-step multiview clustering. IEEE Trans Neural Netw Learn Syst

48. Chen MS, Wang CD, Lai JH (2022) Low-rank tensor based proximity learning for multi-view clustering. IEEE Trans Knowl Data Eng 35(5):5076–5090
49. Liu X, Li M, Tang C, Xia J, Xiong J, Liu L, Kloft M, Zhu E (2020) Efficient and effective regularized incomplete multi-view clustering. IEEE Trans Pattern Anal Mach Intell 43(8):2634–2646
50. Tang C, Li Z, Wang J, Liu X, Zhang W, Zhu E (2022) Unified one-step multi-view spectral clustering. IEEE Trans Knowl Data Eng 35(6):6449–6460

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.