A novel semi-supervised consensus fuzzy clustering method for multi-view relational data

Hoang Thi Canh[1,2], Pham Huy Thong[3], Phung The Huan[2], Vu Thuy Trang[4], Nguyen Nhu Hieu[5], Nguyen Tien Phuong[6], Nguyen Nhu Son[6]

[1]Graduate University of Science and Technology, Vietnam Academy of Science and Technology, Hanoi, Vietnam
[2]Faculty of Information Technology, Thai Nguyen University of Information and Communication Technology, Thai Nguyen, Vietnam
[3]Faculty of Information Technology, Hanoi University of Industry, Hanoi, Vietnam
[4]Faculty of Mathematics, Mechanics and Informatics, Hanoi University of Science, Vietnam National University, Hanoi, Vietnam
[5]Department of Information Technology, ViettelPost Corporation, Hanoi, Vietnam
[6]Institute of Information Technology, Vietnam Academy of Science and Technology, Hanoi, Vietnam

Article Info

Article history:

Received Mar 5, 2024
Revised Jul 22, 2024
Accepted Aug 6, 2024

Keywords:

Multi-view relational data
Clustering
Semi-supervised
Fuzzy clustering
Multi-view clustering

ABSTRACT

Multi-view data is widely employed in various domains, highlighting the need for advanced clustering methodologies to efficiently extract knowledge from these datasets. Consequently, multi-view clustering has emerged as a prominent research topic in recent years. In this paper, we propose a novel approach: the semi-supervised consensus fuzzy clustering method for multi-view relational data (SSCFMC). This method combines the advantages of fuzzy clustering and consensus clustering to address the challenges posed by multi-view data. By leveraging available labeled information and the relational structure among views, our method aims to enhance clustering performance. Extensive experiments on benchmark datasets demonstrate that our method surpasses existing single-view and multi-view relational clustering algorithms in terms of accuracy and stability. Specifically, the SSCFMC algorithm exhibits superior clustering performance across various datasets, achieving an adjusted rand index (ARI) of 0.68 on the multiple features dataset and an F-measure of 0.91 on the internet dataset, highlighting its robustness and efficiency. Overall, this study advances multi-view clustering techniques for relational data and provides valuable insights for researchers in this field.

Corresponding Author:

Nguyen Nhu Son
Institute of Information Technology, Vietnam Academy of Science and Technology
18 Hoang Quoc Viet Road, Cau Giay District, Hanoi, Vietnam
Email: nnson@ioit.ac.vn

1. INTRODUCTION

Multi-view clustering plays a crucial role in data analysis, especially when confronted with diverse data from various sources. Unlike traditional methods, it improves flexibility and accuracy by integrating information from different perspectives, offering a comprehensive view of data structure and relationships. This not only optimizes classification capabilities but also deepens insights, reveals hidden information, and reduces the impact of noise. It delivers accurate and meaningful analytical results in various fields such as data mining, marketing, healthcare, and many other domains [1], [2]. “Multi-view data” encompasses information from diverse sources, methodologies, or perspectives regarding a specific entity, characterized by multiple views offering varied insights with varying precision and reliability. Multi-view clustering (MvC) is a cutting-edge approach in data analysis that addresses the complexities of such data [3]. The core principle of multi-view clustering is its ability to harness the complementary nature of diverse perspectives.

By integrating insights from multiple views, a more holistic and precise representation of entities can be obtained, thereby greatly improving data analysis and decision-making processes [3]. This innovative clustering technique is increasingly gaining prominence in various fields, offering a nuanced and holistic approach to understanding complex datasets and uncovering hidden patterns across different dimensions. The success of the multi-view clustering (MvC) algorithm relies on two crucial principles: the complementary and consensus principles [4]. The complementary principle asserts the use of multiple different views to comprehensively and accurately describe data objects. While each view may offer enough information for a particular knowledge discovery task, diverse views often encompass supplementary information that should be exploited. The consensus principle in multi-view clustering aims to minimize conflicts among data perspectives, enhancing coherence and accuracy in clustering algorithms by integrating information from various viewpoints consistently.

The multi-view clustering method brings numerous advantages over single-view clustering methods, such as [5], [6]: Utilizing diverse information, increasing coherence, enhancing accuracy, and improving knowledge discovery capabilities. Despite the numerous advantages of multi-view clustering over single-view clustering methods, it has also encountered and is currently facing various difficulties and challenges such as data standardization [7], feature selection [8], [9], the integration of clustering results [10], computational complexity and data accuracy [11]. Researchers worldwide have proposed various solutions to these challenges. Zhang *et al.* [12] introduced a multitask multiview clustering algorithm utilizing locally linear embedding (LLE) and Laplacian eigenmaps (LE) to address clustering issues in heterogeneous scenarios. Cao and Xie [8] proposed a unified framework for unsupervised multi-view feature selection, leveraging consensus label information to enhance performance. Liu *et al.* [10] introduced a multi-view clustering method using joint nonnegative matrix factorization (NMF) to integrate information and enhance clustering performance. Wang *et al.* [13] introduced MVC-LFA, a direct and effective approach to maximizing alignment between consensus and weighted base partition matrices through orthogonal transformation, leading to reduced computational complexity and superior clustering performance compared to state-of-the-art methods.

Clustering is a challenging task in data mining, particularly for complex data [14], [15]. Despite the widespread use of multi-view data in real-world applications, most existing clustering algorithms are tailored for single-view data [16]. Consequently, there is a strong demand for advanced clustering methods to effectively extract knowledge from multi-view datasets, making multi-view clustering a prominent research topic in recent times. To address the challenges associated with clustering data, the introduction of partial supervision, commonly known as semi-supervision, has been implemented. Supervision in clustering involves utilizing previously acquired knowledge to guide the process, typically through object labels. Unlike fully supervised methods where every object is labeled, semi-supervised approaches label only a subset of objects [17]. This is advantageous due to the scarcity and potential high cost of fully labeled datasets. As a result, semi-supervision techniques have garnered attention due to their ability to provide partial guidance in the clustering process, making them preferable over fully supervised methods [18], [19]. While semi-supervision techniques are extensively studied in data clustering, their application to semi-supervised multi-view clustering of feature data is relatively limited.

Given the context outlined above, the implementation of semi-supervised consensus fuzzy clustering for multi-view relational data is both crucial and valuable, and it forms the primary focus of this paper. In this study, we propose a novel method for semi-supervised consensus fuzzy multi-view clustering. This method is applied to complex and diverse datasets collected from various sources. With this approach, we first partition the data points into clusters for the unlabeled data in each view. Subsequently, we perform multi-view fuzzy semi-supervised clustering across all views. Finally, we integrate information from multiple views and execute consensus clustering across these views. This integration ensures increased cohesion and maximizes consensus across all views, thereby improving clustering performance, achieving more accurate clustering results, and minimizing the impact of noise and errors. The significance of the semi-supervised fuzzy consensus clustering algorithm for multi-view relational data lies in its ability to enhance clustering accuracy and efficiency by integrating information from multiple views and addressing data uncertainty. This algorithm leverages both labeled and unlabeled data to maximize the available information, which is particularly beneficial when fully labeled data is difficult or expensive to obtain. By using fuzzy logic, the algorithm effectively handles ambiguity and uncertainty in the data, resulting in more reliable clustering outcomes. Additionally, consensus clustering increases cohesion among different views, ensuring that the identified data clusters are consistent across various sources. This approach also allows for the detection of complex patterns and minimizes the impact of noise and errors, resulting in more accurate and robust clustering outcomes.

The following sections of this paper are organized as follows: Section 2 reviews the related work. Section 3 gives a detailed overview of the proposed methodology, and Section 4 showcases the experimental results, analysis, and comparative evaluations. In the conclusion section, key insights are highlighted, and possible directions for future research in upcoming publications are outlined.

2. RELATED WORK

Bickel and Scheffer [20] pioneered research in multi-view clustering, focusing on clustering vector data from multiple views. They extended k-means and expectation maximization (EM) to handle the multi-view setting effectively, particularly for text data with two conditionally independent views. Experimental results demonstrated significant improvements with the multi-view variant of k-means compared to its single-view counterpart. This foundational research has inspired the development of several multi-view clustering methods [3], [21].

Clustering can be applied to either vector data or relational data. Vector data is characterized by a matrix of quantitative or qualitative values that describe each object, represented by an n×l data matrix. In contrast, relational data encapsulates the relationships between objects through dissimilarity measures. Despite receiving less attention, relational data is more versatile and applicable in scenarios where vector data is insufficient for accurate description. Relational data also presents potential advantages in terms of confidentiality, as the attributes of objects can be kept confidential or remain inaccessible [22]. Multi-view relational data sets consist of multiple dissimilarity matrices, each representing a different aspect of the objects being described. To accurately cluster complex data, it is crucial to have multiview relational techniques that account for the varying influences of these dissimilarity matrices on the clustering process. The research detailed in reference [23] introduces a novel semi-supervised multi-view fuzzy clustering algorithm known as the semi-supervised approach for clustering and aggregating relational data (SSCARD). Specifically tailored for relational data, this algorithm integrates pairwise constraints from the AFCC as supervision. It operates as an iterative algorithm rooted in the clustering and aggregation of relational data with instance-level constraints (CARDr) algorithm, itself built upon the relational fuzzy c-means (RFCM) algorithm [22], [24]. The objective function of SSCARD encompasses two components: one addressing the unsupervised aspect and the other handling the violation of pairwise constraints (AFCC). Chen *et al.* [25] introduced the TW-k-means algorithm, which is a two levels weighted clustering method designed for multi view vector data. The algorithm facilitates the simultaneous computation of weights for various views and variables, incorporating two additional steps. Gusmão and Carvalho [26] introduced NERF c-means, a fuzzy clustering algorithm designed for non-Euclidean relational data. It extends the RFCM algorithm to handle arbitrary dissimilarity data, including transformed similarity data. The FANNY algorithm, by Kaufman *et al.* [24], is a fuzzy clustering method for data with similarities, enabling objects to belong to multiple clusters with varying degrees. Chen *et al.* [25] also introduces HCMdd, a variation of the K-medoids algorithm that focuses on medoid-based clustering and can effectively handle diverse data types. In [26], two hybrid clustering methods for multi-view relational data, named PSO_{RWL} and PSO_{RWG}, are introduced. These methods combine particle swarm optimization (PSO) with hard clustering algorithms utilizing multiple dissimilarity matrices. The PSO_{RWL} and PSO_{RWG} algorithms employ modified versions of the $MRDCA_{RWL}$ and $MRDCA_{RWG}$ algorithms, respectively, to update particle positions [27].

3. PROPOSED METHOD

This research proposes a novel approach to semi-supervised consensus fuzzy multi-view clustering (SSCFMC). We apply our method to a diverse multi-view dataset collected from various sources. This dataset is characterized by varying numbers of records within each view, different attribute counts across views, and complex many-to-many relationships between views.

3.1. Main idea

The SSCFMC algorithm involves two steps.

Step 1: Perform fuzzy clustering on the unlabeled data within each view. During this step, the fuzzy C-means (FCM) algorithm is applied independently to each view to divide the data points into clusters [28]. The outcome of this step comprises the cluster centroids and the corresponding membership degrees for each view. These membership degrees are denoted as $\overline{U}$.

Step 2: Multi-view fuzzy semi-supervised clustering. During this phase, the fuzzy membership degrees ($\overline{U}$) obtained from step 1 of the current view are combined with the membership degrees of the remaining views. The aim is to enhance the agreement between all views and reach the highest level of consensus. The output of this step consists of the final cluster centroid values and their corresponding membership degree values for each view. The proposed algorithmic idea is depicted in Figure 1.

3.2. Proposed method

The algorithm diagram is presented in Figure 2. This figure details each step of the proposed method, highlighting key components and their interactions. By examining Figure 2, a comprehensive understanding of the algorithm's workflow and fundamental principles can be gained.

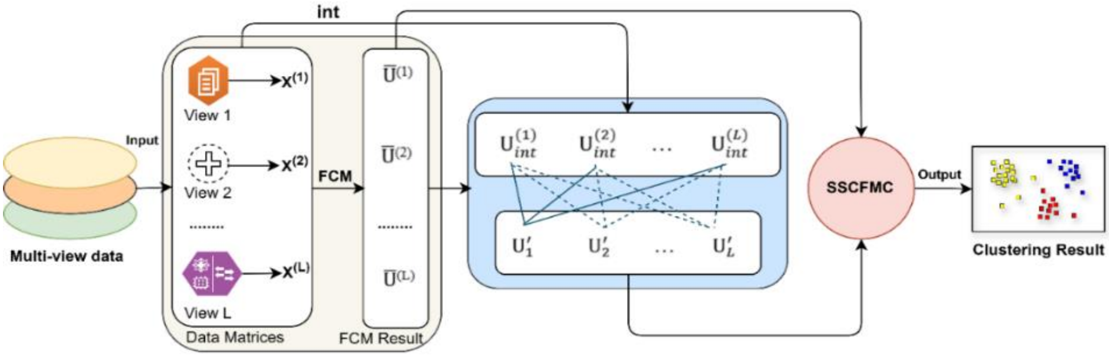

Figure 1. The main idea of SSCFMC

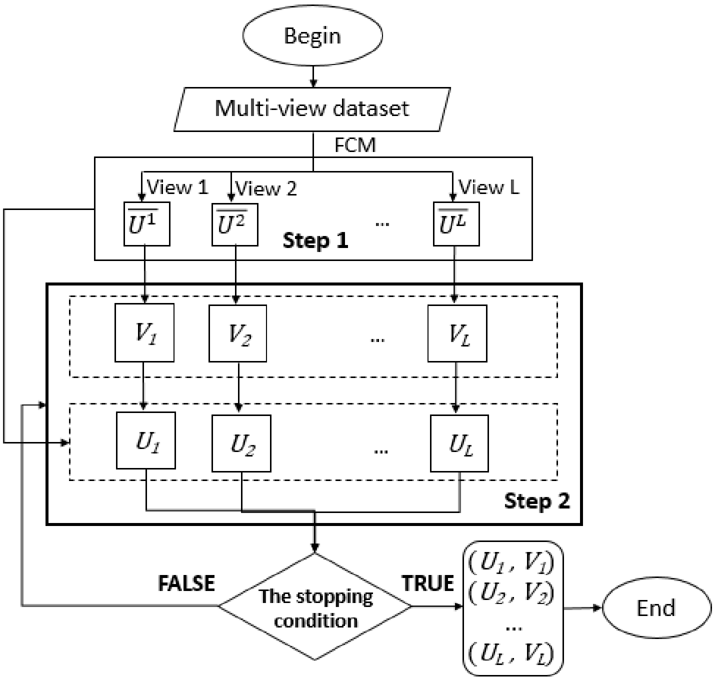

Figure 2. Algorithm diagram

- Mathematical notation: consider a dataset $X = \{x_1, x_2, x_3, \dots, x_N\}$ consisting of N data samples, where x_i denotes the i-th sample. In a multi-view dataset, each data sample is characterized by features derived from various views. Therefore, the multi-view dataset can be expressed as $X = \{X_1, X_2, X_3, \dots, X_L\}$, where $X_l \in R^{N \times d_l}$ denotes the data matrix of the l-th view. Here, L is the total number of views, and d_l is the dimensionality of the l-th view. C is the total cluster count, d_{lkj} is the Euclidean distance from data point k to cluster j in view l, u_{lkj} represents the degree to which data point k belongs to cluster j in view l, x_{lk} is data point k in view l, and v_{lj} is the center of cluster j in view l.
- The objective function: Following the ideas introduced in the previous section, this part will provide a detailed explanation of the modeling for the proposed approach. Specifically, the objective function of this method is defined by three key components, as detailed in (1):

$$F=\sum_{k=1}^{N}\sum_{j=1}^{C}\sum_{l=1}^{L}u_{lkj}^{2}\left|\left|x_{lk}-v_{lj}\right|\right|^{2}+\sum_{k=1}^{N}\sum_{j=1}^{C}\sum_{l=1}^{L}\left(u_{lkj}-\overline{u}_{lkj}\right)^{2}\left|\left|x_{lk}-v_{lj}\right|\right|^{2}+\sum_{k=1}^{N}\sum_{j=1}^{C}\sum_{l=1}^{L}\sum_{i=1,i\neq l}^{L}\left(u_{lkj}-u_{ikj}\right)^{2}\left(\left|\left|x_{lk}-v_{lj}\right|\right|^{2}+\left|\left|x_{ik}-v_{ij}\right|\right|^{2}\right)\rightarrow Min \quad (1)$$

– With constraint conditions:

$$\begin{cases}\sum_{j=1}^{C}u_{lkj}=1\\ u_{lkj}\in(0,1)\end{cases}$$

The objective function (1) consists of three components, detailed as follows:

– The component represents fuzzy clustering:

$$\sum_{k=1}^{N}\sum_{j=1}^{C}\sum_{l=1}^{L}u_{lkj}^{2}\left|\left|x_{lk}-v_{lj}\right|\right|^{2}$$

– The component represents semi-supervised clustering:

$$\sum_{k=1}^{N}\sum_{j=1}^{C}\sum_{l=1}^{L}\left(u_{lkj}-\overline{u}_{lkj}\right)^{2}\left|\left|x_{lk}-v_{lj}\right|\right|^{2}$$

– The component represents consensus multi-view clustering:

$$\sum_{k=1}^{N}\sum_{j=1}^{C}\sum_{l=1}^{L}\sum_{i=1,i\neq l}^{L}\left(u_{lkj}-u_{ikj}\right)^{2}\left(\left|\left|x_{lk}-v_{lj}\right|\right|^{2}+\left|\left|x_{ik}-v_{ij}\right|\right|^{2}\right)$$

In this component, our goal is to integrate the attributes of the object across all perspectives to determine the optimal membership degree. This membership degree should minimize the differences across all views.

The cluster centers are determined using (2):

$$v_{lj}=\frac{\sum_{k=1}^{N}A_{lkj}x_{lk}+\sum_{k=1}^{N}\sum_{i=1,i\neq l}^{L}B_{lkj}x_{lk}}{\sum_{k=1}^{N}A_{lkj}+\sum_{k=1}^{N}\sum_{i=1,i\neq l}^{L}B_{lkj}} \quad (2)$$

In which:

$$\begin{cases}A_{lkj}=u_{lkj}^{2}+\left(u_{lkj}-\overline{u}_{lkj}\right)^{2}\\ B_{lkj}=\left(u_{lkj}-u_{ikj}\right)^{2}\end{cases}$$

The Lagrange multiplier technique is employed to ascertain the value of u_{lkj}:

$$u_{lkj}=\frac{1-\sum_{h=1}^{C}H_{lkh}}{\sum_{h=1}^{C}\frac{Y_{lkj}}{Y_{lkh}}}+H_{lkj} \quad (3)$$

In which:

$$\begin{cases}Z_{lkj}=\sum_{i=1,i\neq l}^{L}u_{ikj}\left(d_{lkj}+d_{ikj}\right)\\ T_{lkj}=\overline{u}_{lkj}d_{lkj}\end{cases}\text{ and }\begin{cases}Y_{lkj}=(L+1)d_{lkj}+\sum_{i=1,i\neq l}^{L}d_{ikj}\\ H_{lkj}=\frac{Z_{lkj}+T_{lkj}}{Y_{lkj}}\end{cases}$$

The algorithm SSCFMC is shown in algorithm 1. The proposed method involves two distinct iterative steps. In the first step, we use the FCM algorithm, which has a complexity of $O(N.C^{2}.nstep)$. In the second step, the proposed algorithm has a complexity of $O(N.C^{2}.L^{2}.nstep)$. Consequently, the overall computational complexity of the algorithm is $O(N.C^{2}.L^{2}.nstep)$.

Algorithm 1. SSCFMC

```
Input $X=\{X_1,X_2,X_3,\dots,X_L\}$, where $X_l \in R^{N\times d_l}$, $C$ is the number of clusters, $Maxstep$ is the number
of iterations, $\epsilon$ is the allowable error.
```

```
Output Cluster labels Y
BEGIN
      Step 1: Fuzzy clustering for unlabeled data in each view
             $\overline{U} = \overline{U_1}, \overline{U_2}, \dots, \overline{U_L}$
      Step 2: Multi-view fuzzy semi-supervised clustering
             2.1. Init t = 0
          Repeat:
             2.2. Update v using formula (2)
             2.3. Update u using formula (3)
      Until: $Max\{\left\|U_{(l)}^{(t+1)} - U_{(l)}^{(t)}\right\|\} \le \epsilon$ or $t \ge$ Maxstep
END
```

4. EXPERIMENTAL RESULTS

This section evaluates the performance of the proposed algorithm. To achieve this, we compare it with other single-view and multi-view clustering algorithms. The comparative analysis provides insights into the strengths and weaknesses of the proposed method relative to existing techniques.

4.1. Environmental configuration

The experiments were conducted on an Apple MacBook Air M1 from 2020, equipped with 8 GB of RAM, 256 GB of storage, and a 7-core GPU. The Python programming language version 3.10 was utilized throughout the study. All tests were executed within the same standardized environment. To evaluate the performance, the following metrics were employed: adjusted rand index (ARI) and F-measure [29]. To introduce randomness into the experiments, we execute all algorithms 20 times and present the average results and standard deviation for the evaluation measures.

To evaluate the effectiveness of the proposed approach, the research team conducted simulations on seven datasets from the UCI machine learning repository. Table 1 provides a comprehensive overview of the characteristics and details of these datasets. This information offers a clear understanding of the data used for performance evaluation.

Table 1. Multi-view datasets

Dataset	Classes	Views	Variables	Objects
3_Sources	06	03	10,259	169
Water_Treatment	04	13	38	527
Flowers	17	04	-	1,360
Multiple_Features	10	06	649	2,000
Phoneme	05	03	150	2,000
Image	07	16	19	2,310
Internet	02	06	1,492	2,359

4.2. Results and discussion

In this section, we conducted experiments on seven multi-view datasets to evaluate the performance of the proposed method and compared the obtained results with those of eleven other methods. Include: TW_Kmeans and EW_Kmeans algorithms [25], PSO_R [30], Hard C-Medoids (HCMdd) [24], NERF [26], FANNY [24], CARDR [23], $MRDCA\_RWL$ and $MRDCA\_RWG$ [27], PSO-RWL and PSO-RWG [26]. The SSCFMC algorithms were executed 20 times, with each execution having a maximum iteration limit of 100.

The experimental results are presented in Table 2, with evaluations conducted using the ARI and F-measure metrics, respectively. In the context of semi-supervised consensus fuzzy multi-view clustering, the adjusted rand index (ARI) and F-measure provide comprehensive performance evaluations. ARI measures the similarity between predicted and true clustering, adjusted for chance, highlighting accuracy by considering both correct pairings and correct separations of data points. A higher ARI value indicates more accurate clustering. ARI ranges from -1 to 1, where 1 represents perfect clustering, 0 indicates random clustering results, and negative values denote clustering worse than random. F-measure is a metric that combines Precision and Recall. The F-measure ranges from 0 to 1, where a value of 1 indicates perfect Precision and Recall, meaning the clusters are both accurate and comprehensive. Together, these metrics offer a holistic view of clustering performance, enabling comparisons with other algorithms and assessing effectiveness across diverse datasets. Table 2 does not display the results of the TW_Kmeans and EW_Kmeans algorithms on the Flowers dataset due to the absence of necessary attributes in this dataset. Hence, only algorithms capable of handling relational data can be tested using this dataset.

Table 2. Results comparison: single-view and multi-view algorithms

Datasets	Index	3_Sources	Water_Treatment	Flowers	Multiple_Features	Phoneme	Image	Internet
PSO_R	Ari	0.32 (0.02)	0.01 (0.00)	0.35 (0.01)	0.50 (0.00)	0.71 (0.00)	0.41 (0.01)	0.50 (0.00)
	F_Measure	0.54 (0.01)	0.20 (0.00)	0.53 (0.02)	0.68 (0.00)	0.85 (0.00)	0.63 (0.01)	0.89 (0.00)
PSO-RWL	Ari	0.11 (0.05)	0.03 (0.00)	0.27 (0.01)	0.66 (0.03)	0.72 (0.01)	0.50 (0.00)	0.53 (0.00)
	F_Measure	0.41 (0.04)	0.25 (0.00)	0.46 (0.01)	0.82 (0.03)	0.85 (0.00)	0.67 (0.00)	0.90 (0.00)
PSO-RWG	Ari	0.36 (0.02)	0.03 (0.00)	0.28 (0.02)	0.62 (0.00)	0.72 (0.00)	0.49 (0.00)	0.50 (0.00)
	F_Measure	0.57 (0.02)	0.27 (0.01)	0.47 (0.02)	0.80 (0.00)	0.85 (0.00)	0.66 (0.00)	0.89 (0.00)
HCM-dd	Ari	0.18 (0.03)	0.02 (0.00)	0.25 (0.03)	0.39 (0.04)	0.48 (0.08)	0.33 (0.05)	0.40 (0.13)
	F_Measure	0.46 (0.02)	0.29 (0.01)	0.43 (0.03)	0.58 (0.05)	0.68 (0.07)	0.56 (0.05)	0.85 (0.05)
NERF	Ari	0.39 (0.03)	0.02 (0.00)	0.14 (0.02)	0.32 (0.00)	0.20 (0.01)	0.35 (0.05)	0.29 (0.00)
	F_Measure	0.58 (0.00)	0.21 (0.01)	0.28 (0.02)	0.48 (0.00)	0.47 (0.01)	0.57 (0.05)	0.82 (0.00)
FANNY	Ari	0.07 (0.02)	0.01 (0.00)	0.20 (0.05)	0.33 (0.01)	0.39 (0.09)	0.34 (0.05)	0.29 (0.00)
	F_Measure	0.38 (0.02)	0.21 (0.02)	0.35 (0.04)	0.50 (0.02)	0.63 (0.07)	0.56 (0.06)	0.82 (0.00)
CARDR	Ari	0.39 (0.03)	-0.02 (0.00)	0.10 (0.03)	0.22 (0.05)	0.21 (0.04)	0.40 (0.01)	0.20 (0.04)
	F_Measure	0.59 (0.00)	0.33 (0.01)	0.24 (0.03)	0.39 (0.05)	0.51 (0.03)	0.59 (0.02)	0.77 (0.03)
MRDCA_RWL	Ari	0.23 (0.06)	0.03 (0.02)	0.27 (0.02)	0.58 (0.06)	0.62 (0.10)	0.40 (0.09)	0.27 (0.04)
	F_Measure	0.49 (0.04)	0.28 (0.02)	0.46 (0.02)	0.74 (0.05)	0.77 (0.08)	0.59 (0.06)	0.80 (0.08)
MRDCA_RWG	Ari	0.25 (0.06)	0.03 (0.02)	0.28 (0.03)	0.56 (0.06)	0.60 (0.11)	0.38 (0.09)	0.29 (0.21)
	F_Measure	0.51 (0.05)	0.28 (0.02)	0.47 (0.03)	0.72 (0.06)	0.75 (0.09)	0.58 (0.07)	0.82 (0.06)
TW_Kmeans	Ari	0.01 (0.01)	0.04 (0.01)	-	0.35 (0.06)	0.71 (0.05)	0.36 (0.04)	0.12 (0.25)
	F_Measure	0.01 (0.01)	0.31 (0.02)	-	0.53 (0.06)	0.84 (0.04)	0.57 (0.03)	0.76 (0.09)
EW_Kmeans	Ari	0.01 (0.10)	0.02 (0.00)	-	0.43 (0.03)	0.41 (0.28)	0.33 (0.07)	0.20 (0.25)
	F_Measure	0.36 (0.04)	0.23 (0.01)	-	0.60 (0.04)	0.62 (0.20)	0.57 (0.06)	0.76 (0.11)
SSCFMC	Ari	0.45 (0.00)	0.03 (0.00)	0.35 (0.00)	0.68 (0.00)	0.51 (0.01)	0.42 (0.00)	0.60 (0.01)
	F_Measure	0.29 (0.00)	0.41 (0.00)	0.31 (0.00)	0.52 (0.00)	0.53 (0.00)	0.67 (0.00)	0.91 (0.00)

In Table 2, the proposed SSCFMC method, evaluated using the ARI metric, demonstrated the best performance in 4 out of 7 cases (flowers, internet, multiple features, 3-sources). The performance comparison based on the ARI measure demonstrates the outstanding capabilities of the SSCFMC algorithm across various datasets. In the Flowers dataset, SSCFMC achieved an ARI score of 0.35, matching the performance of the PSO_R algorithm, indicating strong competitive ability. For the Image dataset, SSCFMC attained an ARI score of 0.42, surpassed only by PSO_RWL (0.50) and PSO_RWG (0.49), highlighting its high efficacy in clustering image data. Notably, in the internet dataset, SSCFMC achieved the highest ARI score of 0.60 among all compared algorithms, underscoring its superior performance in handling internet-related data. Similarly, in the multiple features dataset, SSCFMC achieved an ARI score of 0.68, outperforming all other methods, including the top performers PSO_RWL (0.66) and PSO_RWG (0.62), showcasing its excellent capability in dealing with complex datasets. While in the Phoneme dataset, SSCFMC scored an ARI of 0.51, which is lower than some algorithms like PSO_R (0.71) and PSO_RWL (0.72), it still shows considerable potential for improvement. For the water treatment dataset, although SSCFMC achieved a modest ARI score of 0.03, its performance was consistent with other algorithms, reflecting the inherent challenges of this type of data. Finally, in the 3-Sources dataset, SSCFMC achieved an impressive ARI score of 0.45, significantly outperforming other methods, emphasizing its strength in multi-source data integration.

Additionally, the experimental results presented in Table 2 demonstrate the performance of the SSCFMC algorithm across various datasets using the F-measure metric. The SSCFMC method achieved the highest performance in 3 out of 7 cases (image, internet, water treatment) when evaluated with the F-measure metric. Notably, SSCFMC excels in the internet dataset with an F-measure of 0.91, outperforming all other algorithms and demonstrating its superior capability in handling internet-related data. In the Image dataset, SSCFMC achieved an impressive F-measure of 0.67, matching the top scores by PSO_RWL (0.67) and PSO_RWG (0.66), which underscores its effectiveness in clustering image data. Additionally, SSCFMC showed robustness in the Water Treatment dataset with an F-measure of 0.41, surpassing several other methods in this challenging context. However, the algorithm's performance in the Flowers dataset, with an F-measure of 0.31, suggests there is significant room for improvement compared to higher-scoring algorithms like PSO_R (0.53), PSO_RWG (0.47) and MRDCA_RWG (0.47). Similarly, in the Multiple Features dataset, SSCFMC attained a competitive F-measure of 0.52, but still lagged behind the top performers such as PSO_RWL (0.82) and PSO_RWG (0.80), indicating areas for further enhancement. For the Phoneme dataset, SSCFMC's F-measure of 0.53, while respectable, trails behind leading algorithms such as PSO_R (0.85), PSO_RWL (0.85) and PSO_RWG (0.85). In the 3-Sources dataset, SSCFMC scored an F-measure of 0.29, indicating a need for further refinement to better integrate multi-source data.

These results highlight both the strengths and limitations of the algorithm, suggesting specific areas for optimization to enhance performance across diverse datasets. Notably, SSCFMC excels in internet data clustering, demonstrating significant potential in complex and internet-related datasets. While there is room for further improvement, SSCFMC also shows high and stable performance across various data types, along with remarkable flexibility and adaptability, making it a highly useful and effective clustering tool.

Figure 3 showcases the runtime performance of the SSCFMC algorithm across seven multi-view relational datasets. The results reveal significant differences in runtime among these datasets, offering valuable insights into the algorithm's efficiency and potential areas for enhancement. The multiple features (Mfeat) dataset shows the longest runtime, approximately 1,400 seconds, suggesting that its complexity requires substantial computational resources. In contrast, the Water and 3-Sources datasets have the shortest runtimes, both under 200 seconds, demonstrating SSCFMC's efficiency with simpler or smaller datasets. The Image and internet datasets also exhibit relatively high runtimes, around 900 and 600 seconds respectively, highlighting the computational intensity required for processing image and internet data. The Flower and Phoneme datasets have moderate runtimes of about 600 and 300 seconds, reflecting the algorithm's effectiveness in dealing with moderately complex data. These findings underscore the importance of considering computational resources when applying the SSCFMC algorithm to different datasets. While SSCFMC demonstrates high and stable performance across various data types, optimizing runtime is crucial, especially for complex datasets like multiple features and image. Overall, Figure 3 provides valuable insights into the computational efficiency of SSCFMC and suggests areas where algorithmic improvements could enhance performance.

In general, the proposed method achieves superior results compared to other single-view and multi-view algorithms. Additionally, it exhibits minimal variation in the quality measures of the generated partitions across all runs and consistently demonstrates higher mean values for these metrics, as shown in the results table. These findings highlight the stable performance and capability of our method to produce higher-quality clusters consistently. To further enhance efficiency, especially for complex datasets, optimizing the algorithm's processing steps or employing more advanced computational techniques is essential.

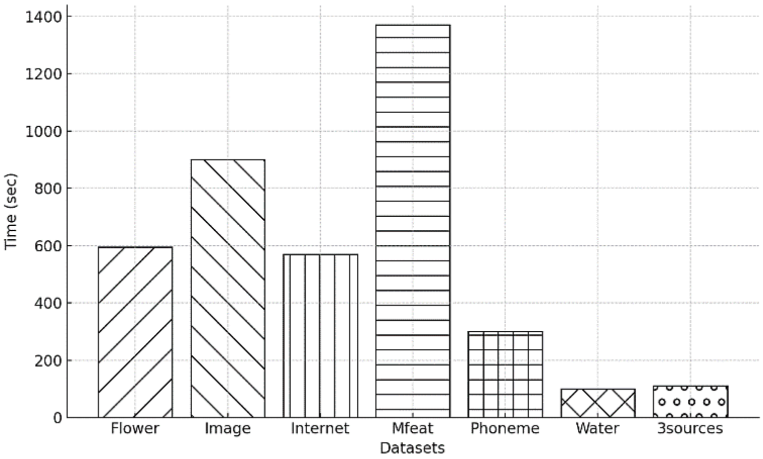

Figure 3. Execution time of multi-view relational methods

5. CONCLUSION AND FUTURE WORK

In this study, we introduced a novel semi-supervised consensus fuzzy clustering method (SSCFMC) specifically designed for multi-view relational data. Our method effectively combines the strengths of semi-supervised learning, consensus clustering, and multi-view data integration to provide a comprehensive solution for complex clustering tasks. Our research has demonstrated several significant advancements. First, SSCFMC achieves higher clustering accuracy by leveraging both labeled and unlabeled data, outperforming traditional single-view and unsupervised methods. This results in more reliable and robust clustering outcomes. Second, the flexibility and adaptability of SSCFMC enable it to handle multiple views and integrate diverse data sources, providing a more holistic understanding of the data. This capability is particularly beneficial for complex datasets often encountered in real-world applications. Lastly, SSCFMC maintains computational efficiency while processing multi-view data, making it suitable for large-scale applications. This addresses a significant limitation of many existing methods that struggle with scalability.

Our findings have several important implications for both the research field and the broader community. First, SSCFMC advances multi-view clustering techniques by introducing a robust method capable of efficiently handling the complexity of multi-view relational data. This innovation pushes the boundaries of current clustering techniques. Additionally, the flexibility and accuracy of SSCFMC make it applicable to various domains, such as bioinformatics, social network analysis, and market research, where multi-view data is prevalent. One of the outstanding questions addressed by our research is the effective integration of multi-view data for clustering purposes. The SSCFMC method offers a solution that significantly improves clustering accuracy while enhancing result interpretability through the integration of multiple data perspectives. These promising findings pave the way for further research in multi-view clustering. However, despite its clear effectiveness in handling multi-source data, SSCFMC has certain limitations. One drawback of SSCFMC is its slower running time, as it is designed to better explore the search space. Additionally, the proposed approach may not be suitable for handling very large datasets due to its complexity. Moreover, it requires maintaining several dissimilarity matrices, which can pose memory constraints. Effectively tackling the complexities of multi-view data, especially when sourced from diverse perspectives, necessitates the ongoing development of innovative algorithms in future research efforts. Future work could focus on further optimizing the algorithm's processing steps or exploring more advanced computational techniques to enhance its performance.

In conclusion, the novel SSCFMC method we have proposed offers significant improvements over existing clustering techniques. It not only addresses the challenges associated with multi-view data but also sets a foundation for future advancements in the field. By providing a robust, efficient, and flexible clustering solution, SSCFMC has the potential to make a substantial impact on both research and practical applications.

ACKNOWLEDGEMENTS

We are grateful for the support from the Grant CS24.04 funded by the Institute of Information Technology, Vietnam Academy of Science and Technology.

REFERENCES

[1] C. Zheng, T. Liu, A. Abd-Elrahman, V. M. Whitaker, and B. Wilkinson, “Object-detection from multi-view remote sensing images: A case study of fruit and flower detection and counting on a central Florida strawberry farm,” *International Journal of Applied Earth Observation and Geoinformation*, vol. 123, Sep. 2023, doi: 10.1016/j.jag.2023.103457.

[2] J. Zhang, Y. Zhou, K. Xia, Y. Jiang, and Y. Liu, “A novel automatic image segmentation method for Chinese literati paintings using multi-view fuzzy clustering technology,” *Multimedia Systems*, vol. 26, no. 1, pp. 37–51, Jul. 2020, doi: 10.1007/s00530-019-00627-7.

[3] Y. Yang and H. Wang, “Multi-view clustering: a survey,” *Big Data Mining and Analytics*, vol. 1, no. 2, pp. 83–107, Jun. 2018, doi: 10.26599/BDMA.2018.9020003.

[4] J. Liu, Y. Jiang, Z. Li, Z. H. Zhou, and H. Lu, “Partially shared latent factor learning with multiview data,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 26, no. 6, pp. 1233–1246, Jun. 2015, doi: 10.1109/TNNLS.2014.2335234.

[5] G. Chao, S. Sun, and J. Bi, “A survey on multiview clustering,” *IEEE Transactions on Artificial Intelligence*, vol. 2, no. 2, pp. 146–168, Apr. 2021, doi: 10.1109/TAI.2021.3065894.

[6] D. Huang, C. D. Wang, and J. H. Lai, “Fast multi-view clustering via ensembles: Towards scalability, superiority, and simplicity,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 11, pp. 11388–11402, Nov. 2023, doi: 10.1109/TKDE.2023.3236698.

[7] Q. Wang, Z. Ding, Z. Tao, Q. Gao, and Y. Fu, “Partial multi-view clustering via consistent GAN,” in *2018 IEEE International Conference on Data Mining (ICDM)*, Nov. 2018, pp. 1290–1295, doi: 10.1109/ICDM.2018.00174.

[8] Z. Cao and X. Xie, “Structure learning with consensus label information for multi-view unsupervised feature selection,” *Expert Systems with Applications*, vol. 238, Mar. 2024, doi: 10.1016/j.eswa.2023.121893.

[9] R. Zhang, F. Nie, X. Li, and X. Wei, “Feature selection with multi-view data: a survey,” *Information Fusion*, vol. 50, pp. 158–167, Oct. 2019, doi: 10.1016/j.inffus.2018.11.019.

[10] J. Liu, C. Wang, J. Gao, and J. Han, “Multi-view clustering via joint nonnegative matrix factorization,” in *Proceedings of the 2013 SIAM International Conference on Data Mining, SDM 2013*, May 2013, pp. 252–260, doi: 10.1137/1.9781611972832.28.

[11] L. Fu, P. Lin, A. V. Vasilakos, and S. Wang, “An overview of recent multi-view clustering,” *Neurocomputing*, vol. 402, pp. 148–161, Aug. 2020, doi: 10.1016/j.neucom.2020.02.104.

[12] Y. Zhang, Y. Yang, T. Li, and H. Fujita, “A multitask multiview clustering algorithm in heterogeneous situations based on LLE and LE,” *Knowledge-Based Systems*, vol. 163, pp. 776–786, Jan. 2019, doi: 10.1016/j.knosys.2018.10.001.

[13] S. Wang *et al.*, “Multi-view clustering via late fusion alignment maximization,” in *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, Aug. 2019, pp. 3778–3784, doi: 10.24963/ijcai.2019/524.

[14] P. T. Huan *et al.*, “TS3FCM: trusted safe semi-supervised fuzzy clustering method for data partition with high confidence,” *Multimedia Tools and Applications*, vol. 81, no. 9, pp. 12567–12598, Feb. 2022, doi: 10.1007/s11042-022-12133-6.

[15] A. Murugan, D. Gobinath, S. Ganesh Kumar, B. Muruganantham, and S. Velusamy, “A time efficient and accurate retrieval of range aggregate queries using fuzzy clustering means (FCM) approach,” *International Journal of Electrical and Computer Engineering*, vol. 10, no. 1, pp. 415–420, Feb. 2020, doi: 10.11591/ijece.v10i1.pp415-420.

[16] W. Alomoush, A. Alrosan, A. Almomani, K. Alissa, O. A. Khashan, and A. Al-Nawasrah, “Spatial information of fuzzy clustering based mean best artificial bee colony algorithm for phantom brain image segmentation,” *International Journal of Electrical and Computer Engineering*, vol. 11, no. 5, pp. 4050–4058, Oct. 2021, doi: 10.11591/ijece.v11i5.pp4050-4058.

[17] T. M. Tuan *et al.*, “A new approach for semi-supervised fuzzy clustering with multiple fuzzifiers,” *International Journal of Fuzzy Systems*, vol. 24, no. 8, pp. 3688–3701, Aug. 2022, doi: 10.1007/s40815-022-01363-3.

[18] D. P. Diogo and F. de A. Francisco de, “Medoid based semi-supervised fuzzy clustering algorithms for multi-view relational data,” *Fuzzy Sets and Systems*, vol. 469, Oct. 2023, doi: 10.1016/j.fss.2023.108630.

[19] H. Cai, B. Liu, Y. Xiao, and L. Y. Lin, “Semi-supervised multi-view clustering based on orthonormality-constrained nonnegative matrix factorization,” *Information Sciences*, vol. 536, pp. 171–184, Oct. 2020, doi: 10.1016/j.ins.2020.05.073.

[20] S. Bickel and T. Scheffer, “Multi-view clustering,” in *Fourth IEEE International Conference on Data Mining (ICDM’04)*, 2004, pp. 19–26, doi: 10.1109/ICDM.2004.10095.

[21] F. Ye, Z. Chen, H. Qian, R. Li, C. Chen, and Z. Zheng, "New approaches in multi-view clustering," in *Recent Applications in Data Clustering*, InTech, 2018.

[22] H. Frigui, C. Hwang, and F. Chung-Hoon Rhee, "Clustering and aggregation of relational data with applications to image database categorization," *Pattern Recognition*, vol. 40, no. 11, pp. 3053–3068, Nov. 2007, doi: 10.1016/j.patcog.2007.02.019.

[23] H. Frigui and C. Hwang, "Fuzzy clustering and aggregation of relational data with instance-level constraints," *IEEE Transactions on Fuzzy Systems*, vol. 16, no. 6, pp. 1565–1581, Dec. 2008, doi: 10.1109/TFUZZ.2008.2005692.

[24] S. Modak, "Finding groups in data: an introduction to cluster analysis, authored by Leonard Kaufman and Peter J. Rousseeuw, John Wiley and Sons, 2005, ISBN: 0-47-1-73578-7," *Journal of Applied Statistics*, vol. 51, no. 8, pp. 1618–1620, Jun. 2024, doi: 10.1080/02664763.2023.2220087.

[25] X. Chen, X. Xu, J. Z. Huang, and Y. Ye, "TW-(k)-means: automated two-level variable weighting clustering algorithm for multiview data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 25, no. 4, pp. 932–944, Apr. 2013, doi: 10.1109/TKDE.2011.262.

[26] R. P. de Gusmão and F. de A. T. de Carvalho, "Clustering of multi-view relational data based on particle swarm optimization," *Expert Systems with Applications*, vol. 123, pp. 34–53, Jun. 2019, doi: 10.1016/j.eswa.2018.12.053.

[27] F. D. A. T. De Carvalho, Y. Lechevallier, and F. M. De Melo, "Partitioning hard clustering algorithms based on multiple dissimilarity matrices," *Pattern Recognition*, vol. 45, no. 1, pp. 447–464, Jan. 2012, doi: 10.1016/j.patcog.2011.05.016.

[28] K. V. Rajkumar, A. Yesubabu, and K. Subrahmanyam, "Fuzzy clustering and fuzzy C-Means partition cluster analysis and validation studies on a subset of CiteScore dataset," *International Journal of Electrical and Computer Engineering*, vol. 9, no. 4, pp. 2760–2770, Aug. 2019, doi: 10.11591/ijece.v9i4.pp2760-2770.

[29] H. Wang, Y. Yang, and B. Liu, "GMC: graph-based multi-view clustering," *IEEE Transactions on Knowledge and Data Engineering*, vol. 32, no. 6, pp. 1116–1129, Jun. 2020, doi: 10.1109/TKDE.2019.2903810.

[30] R. P. De Gusmão and F. D. A. T. De Carvalho, "Particle swarm optimization applied to relational data clustering," in *2016 IEEE International Conference on Systems, Man, and Cybernetics, SMC 2016 - Conference Proceedings*, Oct. 2017, pp. 1690–1695, doi: 10.1109/SMC.2016.7844480.

BIOGRAPHIES OF AUTHORS

Hoang Thi Canh is a researcher and Ph.D. student at the Graduate University of Science and Technology, Vietnam Academy of Science and Technology. She received a bachelor's in information technology at Thai Nguyen University (ICTU), Vietnam in 2009. She got a master's degree on computer science at ICTU in 2012. Her research interests include artificial intelligence, data mining, soft computing, and fuzzy computing. She can be contacted at email: htcanh@ictu.edu.vn.

Pham Huy Thong received the Ph.D. degree in mathematics and informatic from VNU University of Science, Vietnam National University, Hanoi, in 2020. He is currently a teacher and researcher of Hanoi University of Industry. His research interests include artificial intelligence, data mining, soft computing, and fuzzy computing. He can be contacted at email: thongph@haui.edu.vn.

Phung The Huan received the Ph.D. degree in computer science at University of Information and Communication Technology, Thai Nguyen University (ICTU) in 2023. He received a bachelor's in information technology at ICTU, Vietnam in 2009. He got a master's degree in computer science at ICTU in 2012. His research interests include geographic information systems, pattern recognition, data mining, soft computing, and fuzzy computing. He can be contacted at email: pthuan@ictu.edu.vn.

Vu Thuy Trang is currently a student at the Hanoi University of Science, Vietnam National University, Vietnam. Her research interests include artificial intelligence, data mining, soft computing, and fuzzy computing. She can be contacted at email: vuthuytrang_t65@hus.edu.vn.

Nguyen Nhu Hieu is currently a researcher at ViettelPost Corporation and a master student at University of Engineering and Technology, Vietnam National University, Vietnam. His research interests include artificial intelligence, soft computing, and deep learning. He can be contacted at email: hieunn13@viettelpost.com.vn.

Nguyen Tien Phuong received the Ph.D. degree in mathematics from the Vietnam Academy of Science and Technology in 2016. He is currently a principal researcher and member of Scientific Council, at Institute of Information Technology (IOIT) - Vietnam Academy of Science and Technology. His research interests include artificial intelligence, data mining, geographical information system and soft computing. He can be contacted at email: phuongnt@ioit.ac.vn.

Nguyen Nhu Son received the Ph.D. degree in computer science from The University of Queensland, Australia, in 2007. He is currently a Vice Chair of Scientific Council, Head of Department at Institute of Information Technology (IOIT) - Vietnam Academy of Science and Technology. His research interests include artificial intelligence, data mining, soft computing, and fuzzy computing. He can be contacted at email: nnson@ioit.ac.vn.