ATLANTIS PRESS
Atlantis Studies in Uncertainty Modelling, volume 3
Joint Proceedings of the 19th World Congress of the International Fuzzy Systems Association (IFSA), the 12th Conference of the European Society for Fuzzy Logic and Technology (EUSFLAT), and the 11th International Summer School on Aggregation Operators (AGOP)

Fuzzy Co-clustering Algorithm for Multi-source Data Mining

Le Thi Cam Binh[a] and *__Pham Van Nha__[b] and **Pham The Long**[c]
[a]Hanoi University of Culture, 418 La Thanh, Hanoi, Vietnam, cambinhlt@gmail.com
[b]Academy of Military Science and Technology, 17 Hoang Sam, Hanoi, Vietnam, famvannha@gmail.com
[c]Le Quy Don University, 236 Hoang Quoc Viet, Hanoi, Vietnam, longpt@mta.edu.vn

Abstract

The development of information and communication technology has motivated multi-source data to become more common and publicly available. Compared to traditional data that describe objects from a single-source, multi-source data is semantically richer, more useful, however many-feature, more uncertain, and complex. Since traditional clustering algorithms cannot handle such data, multi-source clustering has become a research hotspot. Most existing multi-source clustering methods are developed from single-source clustering by extending the objective function or building combination models. In fact, the fuzzy clustering methods handle the uncertainty data better than the hard clustering methods. Recently, fuzzy co-clustering has proven effective in the many-feature data processing due to the possibility of isolating the uncertainty present in each feature. In this paper, a novel multi-source data mining algorithm based on a modified fuzzy co-clustering algorithm and two penalty terms is proposed, which is called Multi-source Fuzzy Co-clustering Algorithm (MSFCoC). Experimental results on various multi-source datasets indicate that the proposed MSFCoC algorithm outperforms existing state-of-the-art clustering algorithms.

Keywords: Data mining, multi-source, fuzzy co-clustering, multi-view, multi-subspace.

1 Introduction

Clustering plays an important role in data analysis, such as datamining, computer vision, pattern recognition, etc [1]. Some well-known clustering algorithms such as k-means, spectral clustering, fuzzy c-means, fuzzy co-clustering etc. Most of the clustering algorithms available to solve this problem can be grouped into two categories, including the hard clustering methods and the soft clustering methods. In this paper, we mainly focus on the soft clustering methods because of their outstanding advantages in clustering accuracy.

After years of development, the study of traditional single-source clustering has almost come to the bottleneck. The main cause of this situation is that the datasets are aggregated from a single source and describe the objects from a single perspective, so they do not accurately capture comprehensive information of objects. With the rapid development of information and communication technologies, multi-source data has begun to appear in large numbers, meaning that the same objects are described from different perspectives. Basically, multi-source data consists of data subsets that come from different sources or information receiving stations. Depending on the method of data collection, there are two types of multi-source data such as multi-view data and multi-subspace data. Thus multi-source data is a general concept, multi-view data and multi-subspace data are separate cases of multi-source data. According to such a concept, multi-source data becomes more complicated than single-source data because of diversity and heterogeneity. To analyze multi-source data, there are two branches of clustering techniques corresponding to two types of multi-source data, including multi-view clustering and multi-subspace clustering.

In multi-view clustering problems, the objects are distributed equally in the data subsets. Data subsets have different weights and feature numbers. The multi-view clustering algorithm finds functional matrices that belong to the optimal object common to all data subsets by learning similar matrices between data subsets. Some methods of multi-view clustering such as multi-view k-means [2], collaborative multi-view

fuzzy clustering [3], multi-view clustering via simultaneous weighting on sources and features [4], Multi-view clustering via deep concept factorization [5], Multi-view spectral clustering via sparse graph learning [6].

In multi-subspace clustering problems, the objects are scattered in subsets of data. Subsets of data have the same weights and feature numbers. The multi-subspace clustering algorithm finds function matrices that belong to the optimal objects of all data subsets by learning a matrix of characteristic functions between subset data. Some methods of clustering multi-subspace such as Simultaneously learning feature-wise weights and local structures for multi-view subspace clustering [7], adaptive multi-view subspace clustering for high-dimensional data [8], feature concatenation multi-view subspace clustering [9].

In general, traditional data describes objects from a single source, multi-source data is richer, semantically more useful, but more complex. Because traditional clustering algorithms cannot process such data, clustering multi-view and multi-subspace data have become a search hotspot. However, there is no technique that can be used to analyze both types of data. In fact, multi-source data is potentially more uncertain and the fuzzy clustering methods handle the uncertainty data better than the hard clustering methods. Recently, the weighted multi-view collaborative fuzzy C-means algorithm (WCoFCM) [10]has been shown to achieve high clustering performance compared with the existing state-of-the-art multiview clustering algorithms. However, WCoFCM tries to consider the effect separately from the views by calculating the weight of the views and the penalty of the fuzzy object membership function.

Fuzzy co-clustering (FCoC) [11] is an extension of fuzzy clustering that can simultaneously consider data in terms of object and feature based on the fuzzy object membership function and the fuzzy feature membership function. Recently, fuzzy co-clustering has been applied to solve some problems such as clustering documents and keywords [16]-[18], color segmentation [19], categorical multivariate data [11] and high-dimensional data [20, 21], hyper-spectral image analysis [11]. However, the new fuzzy cluster clustering algorithms are only applicable to single-source data processing problems. WCoFCM and FCoC motivated us to integrate these two algorithms to design a new algorithm that can cluster multi-view and multi-subspace data simultaneously.

In this paper, we are motivated by the advantages of fuzzy co-clustering in many-feature clustering and two penalty terms from the advanced ideas of WCoFCM. We further develop a framework for multi-source fuzzy co-clustering and thus propose a novel data mining algorithm called Multi-source Fuzzy Co-clustering Algorithm (MSFCoc) which can help automatically determine the type of multi-source data for high-performance clustering.

The rest of the paper is as follows. The theory of clustering and multi-source, the base of multi-source clustering are discussed in section 2. Section 3 introduces the proposed algorithm MSFCoC. Section 4 covers some experiments. Section 5 includes conclusions and future works.

2 Related works

This section presents the related works of MSFCOC by introducing the WCoFCM and FCoC techniques followed with discussing the multi-view clustering.

2.1 Multi-source data

Definition 2.1 *Multi-source data is a data set consisting of data objects distributed in M data subset coming from M data receiving stations located in different projection spaces. That mean,* $X = \{X_1, X_2, ..., X_M\}$, $N = \overline{\|X\|}$, $X_i = \{x_{i,1}, x_{i,2}, ..., x_{i,Ni}\}$, $x_{i,qi} \in R^{Di}, i = \overline{1,M}, qi = \overline{1,N_i}$. *Where, N is the size of the multi-source data set, M is the number of receiving stations,* X_i *is the data subset coming from the ith station,* $N_i = \|X_i\|$ *is the number of objects in the ith subset,* D_i *is the number of dimensions of the space containing the ith station.*

Definition 2.2 *Multi-view data is multi-source data consisting of* N_r *data objects projected by different spaces and equally distributed in M data subsets. Mean,* $N_r = N_i = N_j$, $N = M * N_r$, $D_i \neq D_j$, $x_{iq} \Leftrightarrow x_{jq}$, $\forall i \neq j; i, j = \overline{1,M}$, $q = \overline{1,N_r}$.

Definition 2.3 *Multi-subspace data is multi-source data consisting of* N_r *objects in the same D-dimensional space and distributed in M data subsets. Mean,* $X = X_1 \cup X_2 \cup ... \cup X_M$, $N = N_r = \sum_{i=1}^{M} N_i$, $D_i = D_j$, $x_{iq} \neq x_{jq}$, $q = \overline{1,N}$, $\forall i \neq j; i, j = \overline{1,M}$.

2.2 Weighted multi-view collaborative fuzzy C-means algorithm

Weighted multi-view collaborative fuzzy c-means algorithm (WCoFCM) [10] has been extended from the collaborative fuzzy c-means algorithm by adding weights to each view. The objective function of WCoFCM is shown in formula (1). Where, η is the penalty factor, and M is the number of views, N is the number of objects in each view, C is the real number

of clusters of the data set. $u_{ci,m}$ is the degree of membership of the ith object with the cth cluster. w_m is the weight of mth source. $d_{ci,m} = \|x_{i,m} - p_{cm}\|$ with $x_{i,m}$ is the ith object in the mth source, pc,m is the center of the cth cluster in the mth source. τ and β are fuzzy coefficients.

$$J_{WCoFCM} = \sum_{m=1}^{M} w_i^{\tau} \left(\sum_{c=1}^{C} \sum_{i=1}^{N} u_{ci,m}^{\beta} d_{ci,m}^2 + \frac{\eta}{M-1} \sum_{m'=1}^{M(m \neq m')} \sum_{c=1}^{C} \sum_{i=1}^{N} \left| u_{ci,m'}^{\beta} - u_{ci,m}^{\beta} \right| d_{ci,m}^{\beta} \right) \quad (1)$$

In the objective function J_{WCoFCM}, the first component $\sum_{c=1}^{C} \sum_{i=1}^{N} u_{ci,m}^{\beta} d_{ci,m}^2$. is the total distance in mth source. This component corresponds to the objective function of fuzzy C-means. When objects in the same source are grouped into real clusters, this component reaches the minimum value. The second component $\sum_{m'=1,m' \neq m}^{M} \sum_{c=1}^{C} \sum_{i=1}^{N} \left| u_{ci,m'}^{\beta} - u_{ci,m}^{\beta} \right| d_{ci,m}^{\beta}$ is the sum of the differences in the distance between the mth source and the remaining sources. When the objects in the sources are grouped into the correct real clusters, the total distance in the sources tends to come close. This component will then reach the minimum value.

The overall clustering result is expressed as:

$$U_{ci} = \sum_{m=1}^{M} w_m u_{ci,m} \quad (2)$$

Thus, the WCoFCM algorithm is suitable for multi-source data processing. In addition, the objective function J_{WCoFCM} is affected by the weights of the views. This adds strength to WCoFCM algorithm in data processing where the feature weight and the number of features are in different sources. However, because WCoFCM assumes the number of objects in the sources is equal, and the second component of the objective function J_{WCoFCM} is used only in cases where the number of objects in the sources are equal. Therefore, the WCoFCM algorithm is only suitable for multi-view data analysis problems, and other data types are not possible.

2.3 Fuzzy co-clustering algorithm

The fuzzy co-clustering algorithm (FCoC) [11] is extended from the fuzzy C-means algorithm approach for the purpose of processing data with many features and uncertainty. The objective function J_{FCoC} is expressed according to Eq. (3)

$$\begin{aligned} J_{FCoC}(U,V,P) = & \sum_{c=1}^{C} \sum_{i=1}^{N} \sum_{j=1}^{D} u_{ci} v_{cj} d_{cij} + \\ & T_U \sum_{c=1}^{C} \sum_{i=1}^{N} u_{ci} \log u_{ci} + T_V \sum_{c=1}^{C} \sum_{j=1}^{D} v_{cj} \log v_{cj} \end{aligned} \quad (3)$$

Where, N is the number of objects, C is the number of real clusters, D is the number of features. u_{ci} is a fuzzy object membership function. v_{cj} is a fuzzy feature membership function. T_u is the entropy weight of the fuzzy object membership function, T_v is the entropy weight of the fuzzy feature membership function. $d_{\text{cij}} = \left\| x_{\text{ij}}, p_{cj} \right\|^2$ where pc is cth cluster center, d_{cij} is the distance between jthfeatures of the ith object (x_{ij}) and the cth cluster center (p_{cj}). d_{cij} is calculated as follows,

$$d_{\text{cij}} = \left\| x_{\text{ij}}, p_{cj} \right\|^2 = \left(x_{\text{ij}} - p_{cj} \right)^2 \quad (4)$$

Compared with the fuzzy C-means algorithm, the FCoC algorithm not only considers the data by objects (x_i) but also considers their features (x_{ij}). That is, in addition to the object function membership u_{ci}, the objective function of FCoC also has the feature considers v_{cj}. In addition, the distance between the objects and the cluster center d_{ci} (see Eq. (1)) is divided into the distance between the corresponding features d_{cij}. This has impacted FCoC, making it a better performance than fuzzy C-means in terms of the precision and uncertainty perception of each individual feature. Thus, FCoC can replace fuzzy C-means for clustering data with many features and uncertainty. This motived us to continue extending FCoC towards fuzzy C-means for multi-source data clustering. Our research results presented in Section 3 presented a new algorithm that can not only process multi-view data but also multi-subspace data.

3 The proposed multi-source fuzzy co-clustering algorithm

In this section, we present a new method to improve the performance of the algorithm FCoC by applying the conception of multi-source, which called the MSFCoC algorithm. In this method, data consists of different data sets that are collected from different sources or projected by different spaces. To modify the FCoC to suit the multi-source data, we duplicate FCoC according to M sources while adding the object distance component and the feature distance component between the data sources.

The objective function of the MSFCoC algorithm is expressed according to Eq. (5)

$$J_{MSFCoC} = \sum_{m=1}^{M}\sum_{c=1}^{C}\sum_{i=1}^{N_m}\sum_{j=1}^{D_m} u_{cim} v_{cjm} d_{cijm} + \Delta_u + T_u \sum_{m=1}^{M}\sum_{c=1}^{C}\sum_{i=1}^{N_m} u_{cim}\log u_{cim} + \Delta_v + T_v \sum_{m=1}^{M}\sum_{c=1}^{C}\sum_{j=1}^{D_m} v_{cjm}\log v_{cjm} \quad (5)$$

where,

$$\Delta_u = \frac{\eta_u}{M-1}\sum_{m'=1}^{M(m\neq m')}\sum_{c=1}^{C}\sum_{i=1}^{N_{m'}}\sum_{j=1}^{D_m}\left|u_{ci,m'} - u_{ci,m}\right| d_{cij,m} \quad (6)$$

$$\Delta_v = \frac{\eta_v}{M-1}\sum_{m'=1}^{M(m\neq m')}\sum_{c=1}^{C}\sum_{j=1}^{D_m}\left|v_{cj,m'} - v_{cj,m}\right| d_{cij,m} \quad (7)$$

To get optimal clustering results, the objective function (5) is minimized subject to following constraints:

$$\begin{cases} \sum_{c=1}^{C} u_{ci,m} = 1, u_{ci,m} \in [0,1], \forall i = \overline{1,N_i}, \forall m = \overline{1,M} \\ \sum_{j=1}^{D} v_{cj,m} = 1, v_{cj,m} \in [0,1], \forall c = \overline{1,C}, \forall m = \overline{1,M} \\ R_1 : if\ \left((N_i = N_j)\,\&\,(x_{iq} \Leftrightarrow x_{jq}) : i \neq j; q = \overline{1,N_i}\right) \\ then \quad \eta_u \neq 0 \quad else \quad \eta_u = 0 \\ R_2 : if\ \left(D_i = D_j; i \neq j; \forall i, j = \overline{1,M}\right) \\ then \quad \eta_v \neq 0 \quad else \quad \eta_v = 0 \end{cases} \quad (8)$$

Where, $\eta_u$and η_v are parameters used to control the penalty associated with the disagreement. The terms $\frac{\eta_u}{M-1}\sum_{m'=1,m\neq m'}^{M}\sum_{c=1}^{C}\sum_{i=1}^{N_{m'}}\sum_{j=1}^{D_m}\left|u_{ci,m'} - u_{ci,m}\right| d_{cij,m}$ and $\Delta_v = \frac{\eta_v}{M-1}\sum_{m'=1}^{M(m\neq m')}\sum_{c=1}^{C}\sum_{j=1}^{D_m}\left|v_{cj,m'} - v_{cj,m}\right| d_{cij,m}$ are disagreement terms, which can be considered as the divergence between partitions from different views, i.e., the lower the value of $\left|u_{ci,m'} - u_{ci,m}\right|$, the lower the divergence between the object membership functions in views, the lower the value of $\left|v_{cj,m'} - v_{cj,m}\right|$, the lower the divergence between the feature membership functions in views.

The R_1 rule determines whether the multi-source data set is multi-view data. If it is not multi-view data, automatically assign the coefficient $\eta_u = 0$. That is, the objective function J_{MSFCoC} does not consider the relationship of the object membership functions $\left|u_{ci,m'} - u_{ci,m}\right|$ between the sources. Rule R_2 determines whether multi-source data is multi-subspace data. If it is not multi-subspace data, automatically assign the coefficient $\eta_v = 0$. That is, the objective function J_{MSFCoC} does not consider the relationship of the feature membership functions $\left|v_{cj,m'} - v_{cj,m}\right|$ between the sources. The remaining components of the J_{MSFCoC} objective function are similar to those in the J_{WCoFCM} objective function in Eq. (1) and J_{FCoC} in Eq. (3).

The degree of object membership, the degree of feature membership and clustering centers of each source can be obtained by using the Lagrangian multiplier method, with the expressions Eq.(9), Eq. (10) and Eq. (11).

$$u_{ci,m} = \frac{e^{\frac{-\sum_{j=1}^{D_m} v_{cj,m} d_{cij,m} - \frac{\eta_u}{M-1}\sum_{m'=1}^{M(m'\neq m)}\sum_{j=1}^{D_m}\left(u_{ci,m'}-1\right) d_{cij,m}}{T_u}} - 1}{\sum_{k=1}^{C} e^{\frac{-\sum_{j=1}^{D_m} v_{kj,m} d_{kij,m} - \frac{\eta_u}{M-1}\sum_{m'=1}^{M(m'\neq m)}\sum_{j=1}^{D_m}\left(u_{ki,m'}-1\right) d_{kij,m}}{T_u}} - 1} \quad (9)$$

$$v_{cj,m} = \frac{e^{\frac{-\sum_{j=1}^{D_m} u_{ci,m} d_{cij,m} - \frac{\eta_v}{M-1}\sum_{m'=1}^{M(m'\neq m)}\sum_{j=1}^{D_m}\left(v_{cj,m'}-1\right) d_{cij,m}}{T_v}} - 1}{\sum_{k=1}^{C} e^{\frac{-\sum_{j=1}^{D_m} u_{ki,m} d_{kij,m} - \frac{\eta_v}{M-1}\sum_{m'=1}^{M(m'\neq m)}\sum_{j=1}^{D_m}\left(v_{kj,m'}-1\right) d_{kij,m}}{T_v}} - 1} \quad (10)$$

$$p_{cj,m} = \frac{\sum_{i=1}^{N_m} u_{ci,m} x_{ij,m}}{\sum_{i=1}^{N_m} u_{ci,m}} \quad (11)$$

Thus, compared with J_{WCoFCM}, J_{MSFCoC} has no specific weights w_m of the sources. Instead, it is the feature membership function $v_{cj,m}$ and the corresponding penalty $\frac{\eta_v}{M-1}\sum_{m'=1,m\neq m'}^{M}\sum_{c=1}^{C}\sum_{j=1}^{D_m}\left|v_{cj,m'} - v_{cj,m}\right|$ of each source. Compared to J_{FCoC}, J_{MSFCoC}

adds a penalty of fuzzy membership functions such as $\frac{\eta_u}{M-1}\sum_{m'=1,m\neq m'}^{M}\sum_{c=1}^{C}\sum_{i=1}^{N_{m'}}\sum_{j=1}^{D_m}\left|u_{ci,m'}-u_{ci,m}\right|d_{cij,m}$ and $\frac{\eta_v}{M-1}\sum_{m'=1,m\neq m'}^{M}\sum_{c=1}^{C}\sum_{j=1}^{D_m}\left|v_{cj,m'}-v_{cj,m}\right|$.

Algorithm 1 Multi-source Fuzzy Co-Clustering algorithm MSFCoC

Input: M data sets $X_m=\left\{x_{im},x_{im}\in R^{D_m}\right\},i=\overline{1,N_m}$, the number of clusters C.
Output: Clustering result.
1. Initialize parameters $T_u,T_v,\eta_u,\eta_v,\varepsilon$, the maximum number of iterations τ_{max}.
2. Initialize $u_{ci,m}$ satisfying Eq. (8).
3. Using the rules R1 and R2 in Eq. (8) to define multi-view and multi-subspace data.
4. τ=1.
5. **While** (not convergent) **do**
6. Update $p_{cj,m}$ using Eq. (11).
7. Calculate $d_{cij,m}$ using Eq. (4).
8. Update $v_{cj,m}$ using Eq. (10).
9. Update $u_{ci,m}$ using Eq. (9).
10. If($\eta_u\neq 0$) update $\overline{u}_{ci}$ using Eq. (12).
11. If($\eta_v\neq 0$) update $\overline{v}_{cj}$ using Eq. (13).
12. τ=τ+1;
13. **End While**

In MSFCoC algorithm, the idea of fuzzy co-clustering ensemble is adopted to combine individual source fuzzy partitions $u_{ci,m}$, $v_{cj,m}$ and obtain the global clustering result $\overline{u}_{ci}$, $\overline{v}_{cj}$. The consensus function is defined as the geometric mean of $u_{ci,m}$, $v_{cj,m}$ for each source and expressed as follows:

$$\overline{u}_{ci}=\sqrt[M]{\prod_{m=1}^{M}u_{ci,m}} \tag{12}$$

$$\overline{v}_{cj}=\sqrt[M]{\prod_{m=1}^{M}v_{cj,m}} \tag{13}$$

MSFCoC improved the performance of multi-source clustering, and as indicated in Eq. (8), MSFCoC considered that each source and each feature contributed equally to clustering, which may decrease the clustering performance when the sources and features have different importance.

The MSFCoC algorithm diagram consists of the learning processes of membership function matrixs U, V that are shown as Algorithm 1.

4 Experimental results

In this section, we conduct some experiments and adopt three mainstream evaluation metrics to observe the performance of above algorithms, and then make analyses based on the real experimental data.

4.1 Data sets

In this paper, we use two many-feature multi-source datasets:

6Dims: This data set includes six data subsets from Computing University of Eastern Finland. Each subset consists of 1024 objects evenly distributed and ordered in 16 clusters and the number of features is 32, 64, 128, 256, 512, and 1024, respectively. We assume that six data subset as copies of an original data set by projecting on six spaces having the number of dimensions 32, 64, 128, 256, 512, and 1024 respectively. Therefore, we consider the 6Dims as the multi-view data set.

Multiple Features (MF): This data set consists of six subset of data from the UCI Machine Learning Repository. Each subset consists of 2000 objects evenly distributed and ordered in 10 clusters, and the number of features is 76, 216, 64, 240, 47 and 6., respectively. this is a copy of an original data set by projecting on six views with dimensions 76, 216, 64, 240, 47 and 6., respectively. We consider the Multiple Features as a multi- view data set.

Image segmentation (IS) data set: This data set is composed of 2310 outdoor images which have 7 classes. Each image is represented by 19 features. The features can be considered as two views which are shape view and RGB view. The shape view consists of 9 features which describe the shape information of each image. The RGB view consists of 10 features which describe the RGB values of each image.

Landsat Satellite (LS) data set: The database consists of the multi-spectral values of 6435 pixels which have 7 classes (red soil, cotton crop, grey soil, damp grey soil, soil with vegetation stubble, mixture class, very damp grey soil). Each pixel is represented by 36 features. The LS data set is divided into two subsets: The sat.trn training subset has 4435 pixels, the sat.tst test subset has 2000 pixels. We consider LS as a multi-subspace data set.

Handwritten Digits (HD) data set: This data set is composed of 5620 handwritten digits which have 10 classes. Each digit is represented by 64 features. The HD data set is divided into two subsets: The "optdigits.trn" training subset has 3823 digits, the "optdigits.tst" test subset has 1797 digits. We consider LS as a multi-subspace data set.

Table 1 describes the basic information of the seven data sets. Where, $D_1\div D_6$ is the number of features in each source.

Table 1: Description of data sets

Dataset	No. objects	No. sources	No. clusters	D_1	D_2	D_3	D_4	D_5	D_6
Multi-view data									
6Dims	1024	6	16	32	64	128	256	512	1024
MF	2000	6	10	216	76	64	6	240	47
Multi-subspace									
IS	2310	2	7	19	19	-	-	-	-
LS	6435	2	7	36	36	-	-	-	-
HD	5620	2	10	64	64	-	-	-	-

4.2 Evaluation metrics

Three standard evaluation metrics, i.e., clustering accuracy (ACC), Davies–Bouldins index (DBI), and partition coefficient (PC), widely used for clustering evaluations are selected in this paper following the experimental settings in [11, 15]. Please refer to [1] for the detailed information and computation formulas of these metrics. Please note that, the larger the values of ACC and PC metrics and the smaller the value of DBI metric, the better the performance of the algorithm.

4.3 Experimental results

In this subsection, we will present the empirical results on five multi-source datasets divided into two groups of multi-view data and multi-subspace data. In each experiment, we compare the results obtained from the MSFCoC algorithm with the single clustering methods and the corresponding multi-source clustering methods.

4.3.1 Experiments on multi-view data sets

This subsection collects experimental results on two datasets 6Dims and MF. We use single clustering methods, that is, FCM and FCCI [11], and the multi-view clustering methods, i.e., WCoFCM [10], Co-FKM and Co-FCM [12] are better than the other single-view clustering methods. For each data set, the reported results were averaged from 50 experimental results. For the single-view clustering methods, we first collected the average results of fifty experiments on each view of the datasets. Then, we take the average of the views again, which is the reported clustering result. The experimental results are reported in Table 2 below.

4.3.2 Experimental on multi-subspace data sets

This subsection collects experimental results on three data sets IS, LS and HD. We use the single clustering methods, that is, FCM and FCCI [11], and the multi-subspace clustering methods, i.e., LMSC [13], CSMSC [14], and JFLMSC [15]. For each data set, the reported results were averaged from 50 experimental results. For the single clustering methods, we first join the subset of data into a complete data set. We then collected the results of fifty experiments on the complete data set. Finally, we report the average results of these 50 experimental results. The experimental results are reported in Table 3 below.

Table 2: Clustering performance (ACC, PC and DBI) of different clustering algorithms on two multi-view datasets

Data sets	**Algorithms**	**ACC**	**PC**	**DBI**
6Dims	FCM	0.53	0.49	5.85
	FCCI	0.92	0.95	0.64
	Co-FKM	0.79	0.65	3.85
	Co-FCM	0.80	0.71	3.53
	WCoFCM	0.84	0.79	3.25
	MSFCoC	**0.96**	**0.98**	**0.42**
MF	FCM	0.62	0.70	5.21
	FCCI	0.90	0.92	1.45
	Co-FKM	0.83	0.85	2.82
	Co-FCM	0.84	0.85	2.66
	WCoFCM	0.86	0.87	2.56
	MSFCoC	**0.93**	**0.95**	**0.52**

4.3.3 Discussion

Tables 2 and 3 report the means of the ACC, PC, and DBI values obtained by the six clustering algorithms in fifty runs on multi-source datasets. These results clearly indicate that the MSFCoC algorithm is the best among all these algorithms. The experimental results of the proposed algorithm indicate that collaborative learning together with feature weighting is an effective way to enhance the performance of the clustering algorithm.

We can draw the following conclusions. First of all, the multi-source methods have better effects than single-source methods on most datasets, which prove that using information complementarity between multiple sources can actually improve clustering effects. Certainly, some algorithms do not achieve better results on

Table 3: Clustering performance (ACC, PC and DBI) of different clustering algorithms on three multi-subspace datasets

Data sets	Algorithms	ACC	PC	DBI
IS	FCM	0.70	0.72	2.67
	FCCI	0.90	0.91	0.94
	LMSC	0.77	0.74	2.38
	CSMSC	0.79	0.75	2.18
	JFLMSC	0.90	0.89	0.98
	MSFCoC	**0.95**	**0.93**	**0.69**
LS	FCM	0.53	0.50	5.97
	FCCI	0.91	0.91	0.90
	LMSC	0.72	0.66	3.15
	CSMSC	0.75	0.68	2.90
	JFLMSC	0.83	0.76	1.55
	MSFCoC	**0.96**	**0.96**	**0.57**
HD	FCM	0.49	0.43	6.26
	FCCI	0.91	0.91	0.75
	LMSC	0.65	0.62	4.47
	CSMSC	0.66	0.63	3.25
	JFLMSC	0.71	0.69	2.75
	MSFCoC	**0.96**	**0.96**	**0.59**

some datasets. For example, on the dataset 6Dims, the methods FCM and FCCI are less effective than MSFCoC. Then, when the data present a high-dimensional situation, the fuzzy co-clustering-based methods, such as FCCI and MSFCoC outperform the other algorithms, indicating that these methods can indeed learn a good subspace representation or potential representation to many-feature data clustering.

5 Conclusion

In this paper, based on a proposed new objective function that explicitly combines two penalty terms, a new multi-source fuzzy clustering algorithm called MSFCoC is proposed. Compared with traditional multi-source clustering algorithms, MSFCoC expands the ability to consider data to be robust with many-feature multi-source data. Also, adding rules (R_1 and R_2 in expression (8)) to define the input data format (multi-view or multi-subspace) makes MSFCoC adaptable to different data types. Experimental results indicate that the proposed MSFCoC algorithm outperforms the existing state-of-the-art multi-view and multi-subspace clustering algorithms.

Although the performance of the proposed MSFCoC is promising, there are still many opportunities for further research. For example, the experiments in this paper are still limited to small datasets, so we do not have a fair basis to compare the time consumed between clustering algorithms. Currently, there is no algorithm that can handle both multi-view and multi-subspace data simultaneously, so there is no basis to compare the ability to automatically process other types of data. The new paper only proposes a new clustering objective function and demonstrates its effectiveness experimentally. Therefore, the algorithm MSFCoC needs to be proven in theory of convergence cases. In the future, we will focus on these topics. In addition, there are several research aspects of MSFCoC that deserve further study. For example, developing a fast version of the proposed MSFCoC algorithm so that they are scalable for large multi-source datasets as clustering based on large datasets is becoming more and more important in real-world applications.

References

[1] L. Fu, P. Lin, A.V. Vasilakos, S. Wang, An overview of recent multi-view clustering, Neurocomputing, Vol. 402, 2020, pp. 148-161.

[2] C. Chen, Y. Wang, W. Hu, Z. Zheng, Robust multi-view k-means clustering with outlier removal, Knowledge-Based Systems, Vol. 210, 2020, 106518. DOI: https://doi.org/10.1016/j.knosys.2020.106518.

[3] CoFKM. a centralized method for multiple-view clustering, 2009 Ninth IEEE International Conference on Data Mining, 2009. DOI: 10.1109/ICDM.2009.138.

[4] B. Jiang, F. Qiu, L. Wang, Multi-view clustering via simultaneous weighting on sources and features, Applied Soft Computing, Vol. 47, 2016, pp. 304-315.

[5] S. Chang, J. Hu, T. Li, H. Wang, B. Peng, Multi-view clustering via deep concept factorization, Knowledge-Based Systems, Vol. 217, 2021, 106807. DOI: https://doi.org/10.1016/j.knosys.2021.106807.

[6] Z. Hu, F. Nie, W. Chang, S. Hao, R. Wang, X. Li, Multi-view spectral clustering via sparse graph learning, Neurocomputing, Vol. 384, 2020, pp. 1-10.

[7] S.X. Lin, G. Zhong, T. Shu, Simultaneously learning feature-wise weights and local structures for multi-view subspace clustering, Knowledge-Based Systems, Volume 205, 12 October 2020, 106280. DOI: https://doi.org/10.1016/j.knosys.2020.106280.

[8] F. Yan, X. Wang, Z. Zeng, C. Hong, Adaptive multi-view subspace clustering for high-

dimensional data, Pattern Recognition Letters, Vol. 130, 2020, pp. 299-305.

[9] Q. Zheng, J. Zhu, Z. Li, S. Pang, J. Wang, Y. Li, Feature concatenation multi-view subspace clustering, Neurocomputing, Vol. 379, 2020, pp. 89-102.

[10] W. Yiping et al., An improved multi-view collaborative fuzzy C-means clustering algorithm and its application in overseas oil and gas exploration, Journal of Petroleum Science and Engineering, Vol. 197, 2021, 108093. DOI: https://doi.org/10.1016/j.petrol.2020.108093.

[11] N.V. Pham et al., Feature-reduction fuzzy co-clustering approach for hyperspectral image analysis, Knowledge-Based Systems, Vol. 216, 2021, 106549. DOI: https://doi.org/10.1016/j.knosys.2020.106549.

[12] Y. Wanga, L. Chen, Multi-view fuzzy clustering with minimax optimization for effective clustering of data from multiple sources, Expert Systems with Applications, Vol. 72, 2017, pp. 457-466.

[13] C. Zhang, H. Fu, Q. Hu, X. Cao, Y. Xie, D. Tao, D. Xu, Generalized latent multi-view subspace clustering, IEEE Trans. Pattern Anal. Mach. Intell., Vol. 42 (1), 2020, pp. 86-99.

[14] S. Luo, C. Zhang, W. Zhang, X. Cao, Consistent and specific multi-view subspace clustering, Thirty-Second AAAI Conference on Artificial Intelligence, 2018.

[15] S.-X. Lin, G. Zhong, T. Shu, Simultaneously learning feature-wise weights and local structures for multi-view subspace clustering, Knowledge-Based Systems, Vol. 205, 2020, https://doi.org/10.1016/j.knosys.2020.106280.

[16] K. Kummamuru, A. Dhawale, R. Krishnapuram, Fuzzy Co-clustering of Documents and Keywords, IEEE International Conf. on Fuzzy Systems, Vol. 2, 2003, pp. 772-777.

[17] W. C. Tjhi, L. Chen, Possibilistic fuzzy co-clustering of large document collections, Pattern Recognition 40 (12), 2007, 3452-3466.

[18] Y. Yan, L. Chen, W. C. Tjhi, Fuzzy semi-supervised co-clustering for text documents, Fuzzy Sets and Systems, Vol. 215, 2013, 74-89.

[19] M. Hanmandlua, O. P. Verma, S. Susan, V. Madasu, Color segmentation by fuzzy co-clustering of chrominance color features, Neurocomputing, Vol. 120, 2013, pp. 235-249.

[20] W. C. Tjhi, L. Chen, A heuristic-based fuzzy co-clustering algorithm for categorization of high-dimensional data, Fuzzy Sets and Systems, Vol. 159, 2008, pp. 371-389.

[21] V. N. Pham , L. T. Ngo, W. Pedrycz, Interval-valued fuzzy set approach to fuzzy co-clustering for data classification, Knowledge-Based Systems, Vol. 107, 2016, pp. 1-13.