Fuzzy Optimization Multi-objective Clustering Ensemble Model for Multi-source Data Analysis

Le Thi Cam Binh[a] and *__Pham Van Nha__[b] and **Ngo Thanh 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

In modern data analysis, multi-source data appears more and more in real applications. Different data sources provide information about different data. Therefore, multi-source data linking is important to improve the processing performance. However, in practice multi-source data is often heterogeneous, uncertain, and large. This issue is considered a major challenge from multi-source data. Ensemble is a universal machine learning model in which learning techniques can work in parallel, with big data. Clustering ensemble has been shown to outperform any standard clustering algorithm in terms of accuracy and robustness. However, most of the traditional clustering ensemble approaches are based on single-objective function and single-source data. In this paper, we propose a new clustering ensemble method for multi-source data analysis. We call the fuzzy optimized multi-objective clustering ensemble method - FOMOCE. Firstly, a clustering ensemble mathematical model based on the structure of multi-objective clustering function, multi-source data, and dark knowledge is introduced. Then, rules for extracting dark knowledge from the input data, clustering algorithms, and base clusterings are designed and applied. Finally, a clustering ensemble algorithm is proposed for multi-source data analysis. Experiments were performed on benchmark data sets. The experimental results demonstrate the superior performance of the FOMOCE method compared with the existing clustering ensemble methods and multi-source clustering methods.

Keywords: Clustering ensemble, multi-source, multi-objective, fuzzy clustering.

1 Introduction

Clustering is an unsupervised learning technique that is used to discover the underlying structure or knowledge within data sets. Data objects grouped into the same cluster have similar properties based on their features and characteristics using difference measurement [1]. Depending on the relationship of each data object to the clusters, clustering algorithms can also be divided into hard and fuzzy clustering algorithms. In fuzzy clustering, data objects are assigned to all clusters with different degrees [2]. Fuzzy clustering algorithms often achieve better clustering quality than some hard clustering algorithms [3]- [5]. Clustering plays an important role in a number of machine learning, pattern recognition, decision support, image analysis, information retrieval, medical and biological applications [6]. However, many clustering approaches still have limitations in terms of computational complexity and large data processing capabilities [5]. To overcome these problems, the clustering ensemble methods were proposed.

Ensemble is a general machine learning approach based on divide-and-conquer. It is formed by a set of single techniques operating in parallel, whose outputs are combined by the decisive consolidation strategy to produce unique results for each particular problem [7]. Ensemble models are used to solve problems such as classification, prediction, regression or clustering [8].

Ensemble clustering aims to combine many different clustering techniques to produce better results than the individual clustering algorithms in terms of clustering consistency and quality [10]. Since ensemble clustering was proposed, it has quickly attracted the attention of researchers. There are several recent studies on cluster ensemble such as mining industry [11], health and biology [12], pattern identification [13], data classification [14], image processing [15], environmental management [17], and large data processing [18].

In general, ensemble clustering is divided into two categories: single-objective ensemble clustering [16, 18] and multi-objective clustering ensemble models [19]. Single-objective clustering ensemble models depend mainly on clustering techniques and the change of parameters on each base clusterings. Therefore, single-objective clustering ensemble models are suitable for clustering single or homogeneous datasets. Multi-objective clustering ensemble models allow the selection of different clustering objective functions based on the data characteristic of each base clustering. Therefore, multi-objective clustering ensemble models are suitable for clustering multi-source, heterogeneous datasets [19]. This paper was inspired by the idea of Wenting Ye et al. in [19] by identifying the dark knowledge of different data sources. This means determining the degree of uncertainty and the number of features in each data source to select each clustering technique (such as K-means, FCM [3], FCCI [4] or IVFCoC [5]) for each base clustering.

In the area of information technology, multi-source data is becoming more and more popular. Multi-view data is a special case of multi-source data where different data sources act role as different projections on the original data set. Compared with traditional data obtained from a single source, multi-source data is semantically richer, more useful, but also more complex. Since traditional clustering algorithms cannot process such data, multi-view clusterings were proposed and developed [20]. However, most of them are single-objective multi-view clustering methods. Samples between different views have a one-to-one relationship regardless of whether the current data is multi-view or single-view. In addition, the relationship of complex mapping between views has not been considered. This motivates us to work further in this direction.

In this paper, we are motivated by the idea in [19] to identify the dark knowledge in multi-source datasets to improve the quality of multi-objective ensemble clustering. We propose a fuzzy multi-objective clustering model ensemble based on the structure of multi-objective clustering function, multi-source data and dark knowledge. In the FOMOCE model, we combine many concurrent clustering algorithms to analyze data coming from different sources. In addition, to the best of our knowledge, this is the first time that dark knowledge has been combined in the input data, clustering algorithms, and base clusters. Experiments were conducted on single data sets and multi-source data sets. Experimental results have proven the effectiveness of our proposed algorithm in comparison to traditional ensemble clustering and multi-view single-objective clustering methods.

This article is organized as follows. Section 2 gives an overview of the ensemble clustering methods, the fuzzy clustering techniques, and the dark knowledge redefinition used in this paper. Section 3 describes in detail the FOMOCE model and ensemble algorithms, comparing the differences between ensemble models. Section 4 evaluates the experimental results of the proposed method. Section 5 presents the conclusions of this article and presents ideas for future work.

2 Related works

In this section, we present main concepts and definitions related to the ensemble clustering: multi-source data, traditional ensemble clustering methods, dark knowledge in ensemble clustering.

2.1 Multi-source data

Definition 2.1 *Single-source data is homogeneous data collected from one or more of the same receiver.*

Single-source data can be aggregated into a common data set to serve traditional clustering techniques that approach single-data processing or be divided into smaller sub-sets to serve traditional clustering ensemble model.

Definition 2.2 *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, $X = \{X_1, X_2, ..., X_M\}$, $N = \|X\|$, $X_i = \{x_{i,1}, x_{i,2}, ..., x_{i,Ni}\}$, $x_{i,qi} \in R^{Di}, i = \overline{1,M}, qi = \overline{1,N_i}$. 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 2.3 *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, $N_r = N_i = N_j$, $N = M * N_r$, $D_i \neq D_j$, $x_{iq} \Leftrightarrow x_{jq}$, $\forall i \neq j; i, j = \overline{1,M}$, $q = \overline{1,N_r}$.*

2.2 Traditional cluster ensemble model

The clustering ensemble model is described as follows.

Definition 2.4 *Clustering ensemble model: Given a data set X consisting of N data objects and M different clustering algorithms or an algorithm with M different sets of parameters. The M base clustering module is formed by implementing M clustering algorithm with M corresponding sub data sets to group each data subset into C different clusters. Clustering results of M clustering modules $\Pi_1, \Pi_2, \ldots, \Pi_M$ are combined by a* *consensus function to obtain the final result Π^*. The cluster ensemble model is depicted in Fig. 1.*

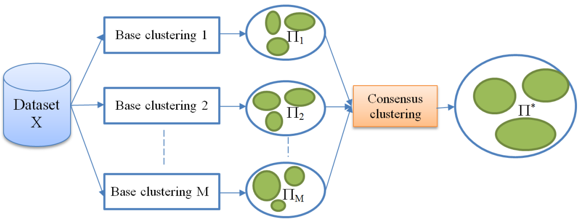

Figure 1: Traditional cluster ensemble model [22]

Definition 2.5 *Base clustering is a clustering module in clustering ensemble models. There the clustering process takes place, using a clustering technique and corresponding parameters.*

Each clustering ensemble model uses the single-objective algorithms to design the base clusterings. In this paper, we use K-means, FCM, FCCI or IVFCoC algorithms for each base clustering in order to set up multi-objective ensemble clustering models.

Definition 2.6 *The clustering consensus function is one of the main components of the clustering ensemble model, where the process of unifying the results from basis clusters into the final clustering result of ensemble clustering models takes place.*

2.3 Dark knowledge in machine learning

Definition 2.7 *Dark knowledge in machine learning is useful information which is hidden or hasn't been used in the data sets or basic elements in machine learning models.*

For instance, in multi-objective ensemble clustering models, dark knowledge can be the information related to the type of input data source or the type of objective function used in the base clusterings.

3 The proposed method

In this section, we present a clustering ensemble method on multi-source data. Including, FOMOCE's mathematical model, dark knowledge determination technique in multi-source data, clustering consensus functions and clustering ensemble algorithm, FOMOCE.

3.1 The clustering ensemble model FOMOCE

In this section, we present the clustering ensemble method based on multi-source datasets, FOMOCE. These include, the general mathematical model of FOMOCE, the technique of determining the dark knowledge in multi-source data, the base clustering linking function, and the FOMOCE algorithm. The purpose of this section is to clarify all components and mechanisms of action, the link between the components in the FOMOCE model as a basis for implementing experiments to demonstrate the superior potential of the proposed ensemble clustering model.

Definition 3.1 *The fuzzy multi-objective clustering ensemble model, denoted Ω, is characterized by data coming from M different sources S_1, S_2,..., S_M. The data is passed through P filter to classify the data source. The base clustering $\Pi = \{\Pi_1, \Pi_2, \ldots, \Pi_M\}$ is linked by L parallel processing fuzzy clustering techniques. An F classification technique is used to consensus the base clustering results and evaluate the quality of clusters to produce X clustering results of M data sources. The model Ω is represented by the following expression (1):*

$$\Omega = \{S, P, \Pi, L, F, X, V\} \tag{1}$$

The functional diagram of the FOMOCE model is summarized in Fig. 2.

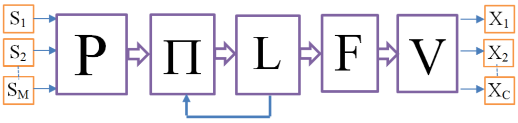

Figure 2: The functional diagram of the FOMOCE model

3.1.1 Data input S

The input data set is described by Eq, (2),

$$S = \{M, S, D\} \tag{2}$$

Where, M is the number of input data sources, S_1, S_2,..., S_M, $S_i = \{x_{i,j}\}$, $x_{i,j} \in R^{Di}$, $i = 1 \div M$, $j = 1 \div ||S_i||$, $||S_i||$ is the number of data objects from the S_i source. $D = \{D_1, D_2, \ldots, D_M\}$ is the number of features of M data sources. When M = 1, input data is collected from a single source, M> 1, data is collected from multiple sources.

3.1.2 Input data classifier P

The P input data classifier is described by equation (3),

$$P = \{S, R_P, f\} \tag{3}$$

Where, S is the input data set described by Eq. (2), R_P is the rule set for classifying the data source, f is the value set that classifies the input data source. R_P is a function of f with parameter S. That is, $f = R_p(S)$ or $f_i = R_P(S_i)$, $i = 1 \div M$.

3.1.3 Set of base clusterings Π

The set of base clusterings is described by Eq. (4), which includes the input data for each base cluster, the set of clustering algorithm selection rules, and the termination conditions of the base clustering.

$$\Pi = \{M, S, f, A, R_\Pi, E, O_\Pi\} \tag{4}$$

Where, $\Pi = \{\Pi_1, \Pi_2, \ldots, \Pi_M\}$ includes M base clusterings. $S = \{S_1, S_2, \ldots, S_M\}$ are M input data sets corresponding to the set of categorical values $f = \{f_1, f_2, \ldots, f_M\}$. $A = \{A_1, A_2, \ldots, A_M\}$ is the set of objective functions used for M base clusterings. R_Π is a set of rules for selecting the objective function for the base clusters based on the f classification value set. Mean, $A = R_\Pi(f)$ or $A_i = R_\Pi(f_i)$, $i = 1 \div M$.

Components A and R_Π are described in detail in section 3.3. down here.

$E = \{E_1, E_2, \ldots, E_M\}$ is the set of stop conditions for the base clusters. The stop condition E_i depends on the A_i clustering algorithm to determine whether the ith base clustering continues or ends the learning loop. That mean, $E = E(A)$ or $E_i = E(A_i)$, $i = 1 \div M$.

$O_\Pi = \{O_{\Pi 1}, O_{\Pi 2}, \ldots, O_{\Pi M}\}$ are the clustering results of the base clustering with $O_{\Pi 1} \cup O_{\Pi 2} \cup \ldots \cup O_{\Pi M} = S_1 \cup S_2 \cup \ldots \cup S_M$. Where, $O_{\Pi i} = \{O_{\Pi i,1}, O_{\Pi i,2}, \ldots, O_{\Pi i,C}\}$, $O_{\Pi i,1} \cup O_{\Pi i,2} \cup \ldots \cup O_{\Pi i,C} = S_i$, $i = 1 \div M$ is the output of the ith base clustering.

3.1.4 The link module of base clusterings L

L is described by Eq. (5), which includes the clustering results of the base clusters in the learning loops, the global best clustering results, and the knowledge exchange links from and to the base clusterings.

$$L = \{O_\Pi, O_{\Pi G}, R_L\} \tag{5}$$

Where, O_Π is the set of clustering results obtained from the base clusterings in the learning loops. Clustering results include the membership function of data objects, list of objects and cluster center of each cluster in each data set. $O_{\Pi G}$ is the result of globally optimal clustering at the iteration steps of the base clusterings. $O_{\Pi G}$ is called dark knowledge in the base clusterings. R_L is the rule for defining and exchanging dark knowledge between the base clusterings.

3.1.5 Consensus module F

The consensus module of the FOMOCE model is described by Eq. (6), which includes input data, consensus algorithm, and final clustering result.

$$F = \{O_\Pi, A^*, X^*\} \tag{6}$$

Where, O_Π is the set of output clusters of base clusterings, that is $O_\Pi = \{O_{\Pi 1}, O_{\Pi 2}$ which is the input data of consensus function. A^* is the objective function of the consensus algorithm. X^* is the output result of the consensus module.

Let $O_{\Pi i} = \{O_{\Pi i,1}, O_{\Pi i,2}, \ldots, O_{\Pi i,C}\}$, $i = 1 \div M$ is the result of the base clustering Π_i. Then, $O_{\Pi i,j} = \left\{x_{i,j,1}, x_{i,j,2}, \ldots, x_{i,j,Mij}\right\}$, $j = 1 \div C$ is the jth cluster of the base cluster Π_i, that is $O_\Pi = \{O_{\Pi ij}\}_{MxC}$ includes MxC result clusters from M the base clusterings. Each cluster $O_{\Pi ij} = \left\{\bar{x}_1, \bar{x}_2, ..., \bar{x}_{M_{ij}}\right\}, j = \overline{1,C}$ consists of objects derived from S_i which are represented by a cluster center $\overline{C_{ij}} \in R^{D_i}$. So, $O_{\Pi ij}$ can be seen as a super-object. We perform $O_\Pi = \{O_1, O_2, ..., O_{MxC}\}_{MxC}$ is a set of super-objects that are represented by $\overline{C} = \left\{\overline{C}_1, \overline{C}_2, ..., \overline{C}_{MxC}\right\}$ is a set of cluster centers.

X^* is the cluster consensus result, X^* includes C clusters. That is, X^* = argmin $(A^*(O_{Pi}))$.

To get the final clustering result, a consensus function in Eq. (6) is used to group M results of base clusterings into C different clusters. Several consensus functions have been developed to produce the final data clustering result. Recently, we have introduced a clustering tendency assessment method SACT [23] applied in hyperspectral image classification. The SACT is viewed as a consensus function based on graph-based approaches. In the FOMOCE model, we use SACT as a consensus function to classify the partitions obtained from the base clusterings into the final clustering result. We first aggregate the partitions obtained from the base clusterings into a set of MxC partitions. Next, we represent the partitions as super-objects that are represented by cluster centers and data object lists. Then, the SACT algorithm is used to group the set of MxC super objects into C clusters which is the final clustering result.

3.2 The clustering results evaluating module

The clustering results evaluation module is described by Eq. (7), including data sets of each cluster, cluster center and cluster quality assessment index.

$$V = \{X^*, C^*, I^*\} \tag{7}$$

Where, $X^* = \{X_1, X_2, \ldots, X_C\}$, $X_i = \left\{x_{i,j}\right\}$, $x_{i,j} \in R^D$, $i = 1 \div C$, $j = 1 \div ||X_i||$ are sets of data objects in result clusters corresponding to cluster centers $C^* = \{C_1, C_2, \ldots, C_C\}$. $I^* = \{I_1, I_2, \ldots, I_Q\}$ is the set of values of the final cluster quality assessment indicators.

3.3 Schematic of the FOMOCE clustering ensemble model

Schematic diagram of the FOMOCE clustering ensemble model including S input data, data classification module, base clustering module, clustering consensus module, cluster quality assessment module, and clustering results and their detailed components are depicted in Fig. 3.

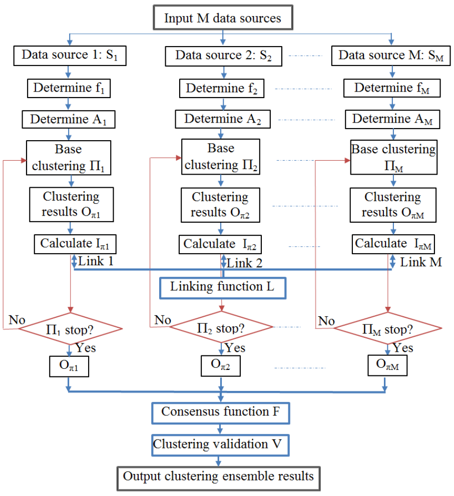

Figure 3: Detailed schematic of the FOMOCE model

3.4 The dark knowledge and derivation rules of the FOMOCE model

3.4.1 Dark knowledge in data sources

Definition 3.2 *The dark knowledge in input data sources is their degree of uncertainty and number of features.*

Dark knowledge in data sources includes: the degree of uncertainty and the degree of feature. Where, the degree of uncertainty includes: No noise and uncertainty, noise and uncertainty, a lot of noise and uncertainty; the degree of feature includes: Small number of features, high number of features and very high number of features. In the FOMOCE model, the type of data source is defined by the following R_P rule.

Definition 3.3 *The rule determines the type of data source:*

```
$R_P$: ClassifySourceFunction (S, M)
BEGIN
    Initialize the set $f = \{f_1, f_2, \ldots, f_M\}$ to store the set of data types.
    For (i = 1 to M)
        If ($S_i$ is no "noisy and uncertainty") and ($S_i$ is "small number of features") then ($f_i$ = "Simple");
        If ($S_i$ is "noisy and uncertainty") and ($S_i$ is "small number of features") then ($f_i$ = "Not simple");
        If ($S_i$ is much "noisy and uncertainty") and ($S_i$ is "small number of features") then ($f_i$ = "Quite complicated");
        If ($S_i$ is "noisy and uncertainty") and ($S_i$ is "high number of features") then ($f_i$ = "Complicated");
        If ($S_i$ is much "noisy and uncertainty") and ($S_i$ is "high number of features") then ($f_i$ = "Very complicated").
    End For
END
```

In the R_P rule, the two main parameters, the degree of the "noisy and uncertainty" and the "number of features" need to be supplied. In fact, the parameter "number of features" can be determined based on the input data. The "noisy and uncertainty" parameter is very difficult to determine which is usually qualitative and provided by the user.

3.4.2 Dark knowledge in clustering techniques

Each algorithm to be accepted must identify the challenges facing and demonstrate the ability to solve those challenges. However, the more and more advanced technology, the demand of people is higher and higher, the challenges are increasing. Hence an algorithm cannot solve all the problems that arise in the relevant field. In data mining, a clustering algorithm can only solve a few challenging problems. They may be suitable for this data type but not for other data types. In summary, each clustering algorithm has its own advantages and can be applied to solve a specific type of problem.

Definition 3.4 *Dark knowledge in clustering techniques is the ability to solve a certain type of clustering problem.*

In the FOMOCE model, we use five different algorithms for the base clustering A = {K-means, FCM, IT2FCM, FCCI, IVFCoC}. Each algorithm has its own strengths and weaknesses and is applied to solve specific problems. These clustering techniques all use a similar quantitative method based on the Euclidean distance measurement. Each algorithm has different advantages and disadvantages in terms of data type, computational complexity, and clustering accuracy. Usually, algorithms with low computational complexity are fast but with low accuracy, and vice versa, algorithms with high complexity are slow but with higher accuracy. In A set, K-means is is the simplest, FCM is not simple, IT2FCM is quite complicated, FCCI is complicated and IVFCoC is very complicated. We designed the rules for defining the objective function for the base clusterings as follows.

Definition 3.5 *The selection rule of objective function for the base clusterings:*

```
R_Π: SelectObjectiveFunction(f, A)
BEGIN
  For(i=1 to M)
    If ($f_i$=Simple) then return $A_i$(K-means);
    If ($f_i$=Not simple) then return $A_i$(FCM);
    If ($f_i$=Quite complicated) then return $A_i$(IT2FCM);
    If ($f_i$=Complicated) then return $A_i$(FCCI);
    If ($f_i$=Very complicated) then return $A_i$(IVFCoC);
  End For
END
```

Where, f is the set of multi-source data types that are defined and provided by module P, A is the set of clustering objective functions that are selected by the user.

3.4.3 Dark knowledge in the base clusterings

One of the fundamental differences between the FOMOCE model and the traditional cluster ensemble models is the existence of the *L* base clusterings linking module. In *L*, the base clusterings can exchange knowledge to each other during the learning loop. So what is the dark knowledge in the base clusterings, please refer to the definition below.

Definition 3.6 *Dark knowledge in base clusterings is the parameters of clustering algorithms, cluster centers and membership functions obtained in learning loops.*

The rule for determining and exchanging dark knowledge in base clusterings is designed as follows:

Definition 3.7 R_L: *If* $(L(O_{\Pi i}) \geq L(O_{\Pi G})$ *then* $O_{\Pi G} = O_{\Pi i}$
Else $O_{\Pi i} = O_{\Pi G}$.

Where, $L(.)$ is a function of clustering performance which was introduced by X. Zhao et al. [14]. $O_{\Pi i}$ is the result of the *ith* base clustering, $L(O_{\Pi i})$ is the performance of the *ith* base clustering. $O_{\Pi G}$ is the best clustering result globally, $L(O_{\Pi G})$ is the best clustering performance. The expression $O_{\Pi G}$= $O_{\Pi i}$ implies the knowledge sharing process from the best base clustering to the *ith* base clustering. Similarly, the expression $O_{\Pi i}$=$O_{\Pi G}$ implies the process of knowledge acquiring from the *ith* base clustering.

3.5 The FOMOCE algorithm

In the format of the FOMOCE mathematical model and the derivation rules described above, the procedure of the FOMOCE algorithm is shown as follows.

Algorithm 1 Pseudocode of the FOMOCE algorithm

```
Input data: $M$ data sources $S_1, S_2, .., S_M$.
Output: The result of clustering $X^*$.
1. Initialization: $f = \{f_1, f_2, \ldots, f_M\}$ and $A = \{A_1, A_2, \ldots, A_M\}$.
2. Determines the type of M data source using $f_i = ClassifySource(S_i, M)$ follows 3.3.
3. Initialize M base clusterings.
4. Select the clustering objective function for each base clustering using $A_i = SelectObjectiveFunction(f_i, A)$.
5. DO Execute parallel learning loops of base clusters.
   5.1. Update the parameters U, V of each base clustering.
   5.2. Update the cluster centers C of the base clusterings.
   5.3. Update and share knowledge according to $R_L$ rules.
   5.4. WHILE The stop condition of all base clusterings is satisfied.
6. Collect the results of the base clusterings and represent them as super-objects.
7. Consensus on the results of the base clusterings to get the result of final clustering $X^*$.
8. Evaluate the clustering quality of $X^*$.
9. Output the $O_{Pi}$ clustering results.
```

4 Experimental results

In this section, we will present some experimental results to simulate the working mechanism of the FOMOCE model and demonstrate the effectiveness of the proposed clustering consensus method. For fair comparison, we have installed clustering experiment along with the state-of-the-art clustering methods such as MKCE [13], eFCoC [16], and FCME [21]. These are single-source clustering ensemble models. To experiment on multi-source datasets, we implement experimentally on multi-view clustering methods which is a special case of multi-source data. Multi-view clustering methods include WCoFCM [8], Co-FKM and Co-FCM [10].

Table 1: Three Single-source data sets from the UCI Repository

Data set	#Clusters	#Objects	#fFeatures
Avila	12	20867	10
Chess	18	28056	6
Farm-Ads	2	4143	54877

Table 2: Clustering performance (ACC, PC and DBI) of different clustering algorithms on three single-source datasets

Data sets	Algorithms	ACC	PC	DBI
Avila	eFCoC	0.92	0.96	0.62
	MKCE	0.79	0.71	4.23
	FCME	0.82	0.77	3.95
	FOMOCE	***0.98***	***0.98***	***0.45***
Chess	eFCoC	0.93	0.92	0.73
	MKCE	0.80	0.74	3.15
	FCME	0.83	0.81	2.87
	FOMOCE	***0.98***	***0.98***	***0.52***
FarmAds	eFCoC	0.95	0.96	0.72
	MKCE	0.63	0.70	5.96
	FCME	0.66	0.71	5.24
	FOMOCE	***0.97***	***0.98***	***0.59***

To evaluate the clustering performance of the proposed approach, we use three standard evaluation metrics, i.e., clustering accuracy (ACC), Davies–Bouldins index (DBI), and partition coefficient (PC). These indicators have been widely used in the field of clustering [9, 12]. 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.

4.1 Single-source data clustering

In this subsection, we collect the results of experiments on three single-source datasets Avila, Chess, and Farm Ads. These data sets come from the UCI Machine Learning Repository. Table 1 provides some of the basic properties of clustering data sets. Three clustering ensemble algorithms, i.e. MKCE, FCME, eFCoC are involved in the experimental process. To use clustering consensus models on single-source datasets, we divide these datasets into five data subsets. Since the five data subsets come from a single-source, in the FOMOCE model we consider them the same. We design five base clusterings corresponding to five algorithms K-mean, FCM, IT2FCM, FCCI, and IVFCoC. Each base clustering corresponds to a single algorithm and a data subset. Experimental results are reported in Table 2.

4.2 Multi-view data clustering

In this subsection, we collect experimental results on two multi-view datasets such as 6Dims and Multiple Features. Where, 6Dims includes six data subsets from the Computing University of Eastern Finland. Each subset consists of 1024 objects evenly distributed and ordered in 16 clusters and the number of features are 32, 64, 128, 256, 512, and 1024, respectively. We assume that six data subsets 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 data set comes from the UCI Machine Learning Repository. Table 3 provides some of the basic properties of two multi-view datasets, where, $D_1 \div D_6$ is the number of features in each source. Three multi-view clustering methods, i.e., WCoFCM [8], Co-FKM and Co-FCM [10] were involved in the experimental process. To use FOMOCE model on multi-view datasets, we initialize the number of base clusterings corresponding to the number of views of each data set. Each base clustering chooses an algorithm using the R_Π rule. Experimental results are reported in Table 4.

4.2.1 Discussion

Tables 2 and 4 report the means of the ACC, PC, and DBI values obtained by the clustering algorithms in fifty runs on single-source and multi-view dataset. These results clearly indicate that the FOMOCE model is the best among all these methods. The experimental results of the proposed algorithm indicate that multi-objective together with dark knowledge is an effective way to enhance the performance of the clustering algorithm.

5 Conclusion

In this paper, based on the perspective of multi-objective and dark knowledge in multi-source data and linking base clusters a new clustering ensemble model has been proposed. Compared with traditional clustering ensemble models and multi-view clustering methods, FOMOCE model expands the ability to consider data to be robust with multi-source data. Experimental results indicate that the proposed FOMOCE algorithm outperforms some existing clustering ensemble methods and multi-source clustering methods.

Although the performance of the proposed FOMOCE 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 be-

Table 3: Two multi-view datasets

Dataset	#Objects	#Sources	#Clusters	D_1	D_2	D_3	D_4	D_5	D_6
6Dims	1024	6	16	32	64	128	256	512	1024
MF	2000	6	10	216	76	64	6	240	47

Table 4: Clustering performance (ACC, PC and DBI) of different clustering algorithms on two multi-view datasets

Data sets	Algorithms	ACC	PC	DBI
6Dims	Co-FKM	0.79	0.65	3.85
	Co-FCM	0.80	0.71	3.53
	WCoFCM	0.84	0.79	3.25
	FOMOCE	***0.97***	***0.98***	***0.55***
MF	Co-FKM	0.83	0.85	2.82
	Co-FCM	0.84	0.85	2.66
	WCoFCM	0.86	0.87	2.56
	FOMOCE	***0.93***	***0.94***	***0.65***

tween clustering algorithms. In the future, we will focus on these topics. In addition, there are several research aspects of FOMOCE that deserve further study. For example, developing a fast version of the proposed FOMOCE 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] S. Miyamoto, H. Ichihashi, K. Honda, Algorithms for Fuzzy Clustering, Springer: Studies in Fuzziness and Soft Computing, Vol. 229, 2008.

[2] D. Xu, Y. Tian, A Comprehensive Survey of Clustering Algorithms, Annals of Data Science, Vol. 2(2), 2015, pp 165–193.

[3] J.C. Bezdek, R. Ehrlich, W. Full , The fuzzy C-means clustering algorithm, Computers & Geosciences, Vol. 10(2–3), 1984, pp. 191–203.

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

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

[6] S. Wierzchon, M. Kłopotek, Modern Algorithms of Cluster Analysis, Springer: Studies in Big Data, Vol. 34, 2018.

[7] X. Dong, Z. Yu, W. Cao, Y. Shi, Q. Ma, A survey on ensemble learning, Frontiers of Computer Science, Vol. 14, 2020, pp. 241-258.

[8] O. Okun, G. Valentini, M. Re, Ensembles in Machine Learning Applications, Springer: Studies in Computational Intelligence, Vol. 373, 2011.

[9] X. Wu, T. Ma, J. Cao, Y. Tian, A. Alabdulkarim, A comparative study of clustering ensemble algorithms, Computers & Electrical Engineering, Vol. 68, 2018, pp. 603-615.

[10] T. Boongoen, N. Iam-On, Cluster ensembles: A survey of approaches with recent extensions and applications, Computer Science Review, Vol. 28, 2018, pp. 1-25.

[11] Y.Y. Yang, D.A. Linkeos, A.J. Trowsdale, J. Tenner, Ensemble neural network model for steel properties prediction, Metal Processing, 2000, pp. 401-406.

[12] H. Wang, B. Zheng, S.W. Yoon, H.S. Ko, A support vector machine-based ensemble algorithm for breast cancer diagnosis, European Journal of Operational Research, Vol. 267, 2018, pp. 687-699.

[13] L. Baia, J. Lianga, F. Cao, A multiple k-means clustering ensemble algorithm to find nonlinearly separable clusters, Information Fusion, Vol. 61, 2020, pp. 36-47.

[14] X. Zhao, J. Liang, C. Dang, Clustering ensemble selection for categorical data based on internal validity indices, Pattern Recognition, Vol. 69, 2017, pp. 150-168.

[15] M. Han, B. Liu, Ensemble of extreme learning machine for remote sensing image classification, Neurocomputing, Vol. 149, 2015, pp. 65-70.

[16] C.B. Le, L.T. Ngo, V.N. Pham, L.T. Pham, A new ensemble approach for hyper-spectral image segmentation, Conference on Information and Computer Science (NICS), 2018.

[17] J. Heinermann, O. Kramer, Machine learning ensembles for wind power prediction, Renewable Energy, Vol. 89, 2016, pp. 671-679.

[18] H. Yu, Y. Chen, P. Lingras, G. Wang, A three-way cluster ensemble approach for large-scale data, International Journal of Approximate Reasoning, Vol. 115, 2019, pp. 32-49.

[19] W. Ye, H. Wang, S. Yan, T. Li, Y. Yang, Nonnegative matrix factorization for clustering ensemble based on dark knowledge, Knowledge-Based Systems, Vol. 163, 2019, pp. 624-631.

[20] S. Lin, G. Zhong, T. Shu, Simultaneously learning feature-wise weights and local structures for multi-view subspace clustering, Knowledge-Based Systems, Vol. 205, 2020.

[21] P. Baraldi, R. Razavi-Far, E. Zio, Bagged ensemble of Fuzzy C-Means classifiers for nuclear transient identification, Annals of Nuclear Energy, Vol. 38(5), 2011, pp. 1161-1171.

[22] P. Panwong,T. Boongoen, N. Iam-On, Improving consensus clustering with noise-induced ensemble generation, Expert Systems with Applications, Vol. 146, 2020, 113138

[23] V. N. Pham, L. T. Pham, D.T. Nguyen, L. T. Ngo, A new cluster tendency assessment method for fuzzy co-clustering in hyperspectral image analysis, Neurocomputing, Vol. 307, 2018, pp. 213-226.