AN ENSEMBLE MODEL APPROACH FOR MANY-FEATURE DATA CLUSTERING

Le Thi Cam Binh*, Ngo Thanh Long*, Pham Van Nha*, Pham The Long*
* Information Technology Institute, 17 Hoang Sam, Ha Noi
* Le Quy Don University, 236 Hoang Quoc Viet, Ha Noi

Abstract— The ensemble is a popular machine learning technique based on the principle of divide and conquer. In data clustering, the ensemble aims to improve performance in terms of processing speed and clustering quality. Most existing ensemble methods face inherent complex challenges such as uncertainty, ambiguity, and overlap. Fuzzy clustering has recently been developed to handle data with many-feature, heterogeneity, uncertainty, and big data. In this paper, we propose an ensemble feature-reduction clustering model (EFRC) using advanced machine learning techniques. The EFRC model consists of three phases. First, the data is feature-reduced by a random projection. Then, the data is divided into subsets based on the likelihood of overlap and quantification of noise. Various clustering techniques are used to cluster the subset of data. Finally, the results of the clustering modules are consensus using the classification technique to produce the final clustering result. Several tests were performed on the benchmark datasets. The test results show the superior performance of the EFRC model compared to the previous models.

Keywords— Clustering, classification, ensemble model, feature reduction, many-feature, big-data.

I. INTRODUCTION

A data clustering groups data objects in such a way that each object is assigned to the same group (called a cluster) with other objects similar to it [1]. It is a popular technique for statistical data analysis, used in many fields, including data mining, machine learning, pattern recognition, image analysis, information retrieval, bioinformatics, data compression, and computer graphics [2]. Some common clustering techniques, used for small-scale datasets, are fuzzy C-means (FCM) and K-means (KM). In [3], fuzzy co-clustering (FCoC) is used to classify high-dimensional data (for example, HSI). Unfortunately, those conventional clustering techniques are often not very efficient when dealing with complex, heterogeneous, high-volume, and rapidly generated data. New efficient clustering methods and tools are needed to be able to extract valuable information from huge amounts of data.

An ensemble is a popular machine learning technique based on the principle of divide and conquer. It is built with a set of independent and parallelizable individual models, whose outputs are combined with a decision synthesis strategy to produce a single outcome for a problem. certain [4]. Models can be classification, prediction, regression, or clustering, which the set is designed to perform [5]. Clustering ensemble is a machine learning method for data clustering. It combines multiple clustering models to produce better results than individual clustering algorithms in terms of consistency and quality [6]. Since clustering ensemble was proposed, it has rapidly gained much attention. There are some recent research on the ensemble in machine learning fields such as the mining industry [7], biology and medicine [8], pattern recognition [9], categorical data [10], image processing [11, 14], environmental management [12], and big data processing [13]. Generally, the clustering ensemble is very effective in unsupervised learning. It is suitable for more datasets than traditional single clustering, and it is also robust against noise and outliers. However, most existing ensemble algorithms are based on a static model, they become more difficult due to the inherent complexities such as uncertainty, vagueness, and overlapping. In this paper, we propose a many-feature data clustering model using advanced machine learning techniques, called the ensemble feature-reduction clustering model (EFRC). It consists of three stages. First, data is reduced-feature using a random projection. Then, second, we divide the data into smaller data subsets by qualifying the noise or the overlap. And then, the different objective functions are used to cluster data subsets in parallel. Finally, the results from the clustering modules are combined with a classification technique to create the final classification result. Experimental results on benchmark datasets demonstrate the superior performance of the EFRC model compared to the previous models.

The rest of the paper is organized as follows. Section II presents the recent work done in the areas of clustering ensemble. Section III introduces the main concepts and

Contact author: Pham Van Nha
Email: famvannha@gmail.com
Manuscript received: 14/11/2020, revised: 29/7/2021, accepted: 31/8/2021.

methods used in the study. The proposed EFRC clustering model is also presented in this section. Then the data used along with the experimental settings are described in Section IV. Section V is the conclusion and future work.

II. RELATED WORKS

In this section, we will present a summary of the main theoretical issues related to the clustering model. Includes model structure, base clusterings, clustering consensus, and clustering ensemble quality assessment.

A. Clustering ensemble model

A clustering ensemble model usually consists of three stages performed in order: data preprocessing, clustering, and clustering ensemble quality assessment. X. Wu et al. [6] have defined the clustering ensemble as follows: There is a dataset $X=\{x_1, x_2, \ldots, x_n\}$ that has n data point. Data X is divided into m different data subsets $X=\{X_1, X_2, \ldots, X_m\}$. Then, m clustering algorithms (base clusterings) are used to clustering these data subsets $X_i\left(i=1,\ 2,\ \ldots, m\right)$ and generate m different partitions $P=\{P_1, P_2, \ldots, P_m\}$. A consensus function is used to ensemble the result partitions $P=\{P_1, P_2, \ldots, P_m\}$ to obtain the clustering result P^*. Finally, the estimated indexes are used to evaluate the clustering quality and give the final clustering results. The traditional clustering ensemble model is shown in Fig. 1. The component modules of the clustering ensemble model are presented in sections B, C, and D below.

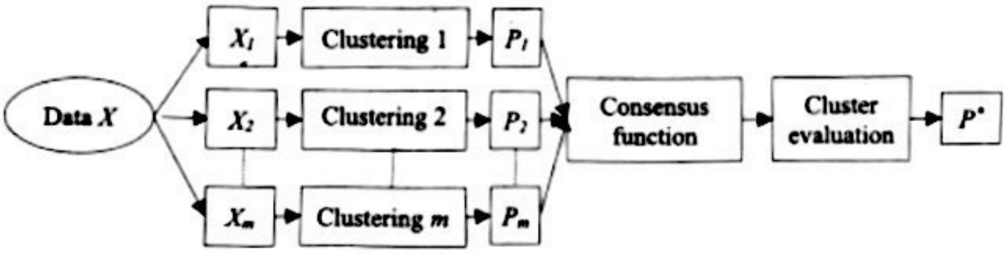

Fig. 1. Traditional clustering ensemble modelBase clustering modules

In the clustering ensemble model, the base clusterings are the basic components in the clustering stage. The base clusterings can be different clustering techniques to cluster the corresponding data. In this paper, we are motivated mainly by demonstrating the working mechanism and demonstrating the effectiveness of a clustering ensemble machine learning approach. Therefore, we will select some popular clustering algorithms such as KM [2], FCM [15], FCoC [16] and IVFCoC [17] for our purposes. These four clustering algorithms have different mathematical structures and different data objects. These algorithms have their advantages and disadvantages in terms of cluster processing time complexity and accuracy.

B. Consensus function module

In the clustering ensemble model, the consensus function module is implemented with a clustering or classification technique to consensus results obtained from the clustering stage. The result obtained by the consensus function module is the final clustering result of the original dataset.

To get the final clustering result, a consensus function is used to group *m* results of base clusterings into k 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 [3] applied in hyperspectral image classification. The SACT is viewed as a consensus function based on graph-based approaches. In the EFRC model, we use SACT as a consensus function to classify the partitions obtained from base clusterings into the final clustering result. We first aggregate the partitions obtained from the base clusterings into a set of *mxk* 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 *m*x*c* super-objects into *k* final clustering result clusters.

C. Cluster evaluation module

The cluster evaluation module is used to evaluate the clustering quality obtained from the consensus function module. This module will quantify cluster evaluation indicators. There are two groups of cluster evaluation indexes: supervised indexes and unsupervised indexes. Table 1 lists a list of cluster evaluation indicators.

Table 1. *Cluster evaluation indexes*

Type	Name	Denote	Best If	Range
Supervised	Accuracy rate [19]	AR	High	0,1
	Recall index [20]	RcI	High	0,1
	Precision index [20]	PcI	High	0,1
	F1 score [21]	F1	High	0,1
Unsupervised	Mean Squared Error [29]	MSE	Low	0,+∞
	Image Quality Index [30]	IQI	High	0,1

	Davies-Bouldins index [31]	DBI	Low	-∞,+∞
	Xie and Benis index [31]	XBI	Low	-∞,+∞

Supervised evaluation indexes are used to evaluate cluster quality on labeled datasets. Supervised evaluation indexes include Accuracy rate [19] (AR), Recall index [20] (Recall), Precision index [20] (Pre.) and Rand index [21] (RI). Where higher values indicate better clustering results. These indexes are often used in the clustering consensus module to evaluate the cluster quality of the final clustering results.

Unsupervised evaluation indexes are used to evaluate cluster quality on unlabeled datasets. Unsupervised evaluation indexes include Mean Squared Error [22] (MSE), Image Quality Index [23] (IQI), Davies-Bouldins index [24] (DBI). and Xie and Benis index [24] (XBI). Where, lower values of the MSE, DBI, and XBI indices, while higher values of the IQI indicate better clustering results. These indexes are often used in the clustering consensus module to evaluate the cluster quality of base clustering and clustering consensus modules.

III. EFRC MODEL

A. EFRC model

Let X be the input dataset in a d-dimensional space, k be the cluster number of the data, M be a set of unsupervised clustering modules and a classification module U. The clustering ensemble problem aims to form a 3-stage classification model: Firstly, the original X dataset is divided into m different data subsets $X=\{X_1, X_2, ..., X_m\}$; then the M is used to cluster each data subset into k different clusters. Thus we obtain $|C|= m * k$ component clusters; Finally, the U is used to classify the component clusters into k different classes. The model of EFRC is shown in Fig. 2.

The clustering ensemble model consists of ten basic components that are shown in Eq. (1).

$$\Sigma=\{m, R, X, P, D, S, M, T, U, H\} \quad (1)$$

1. Size of model m

m is the number of base clusterings [6, 10, 14] of the EFRC model. m is called the size of the model, m is a positive integer.

2. The data space R

R is a real number field.

3. Input data X

X is the input dataset.

$$X=\{d,\ s,\ \{x_i\}_{i=\overline{1,n}}\}, \quad (2)$$

Where, d is the number of features of the data, s number of data sources. X be the input dataset in a d-dimensional space $X=\{x_1, x_2, ..., x_n\}, x_i \in R^d, i=\overline{1,n}$. X can be a single-source dataset, or X can also be aggregated from different s-source datasets $X=\{X_1, X_2, ..., X_s\}$.

When $s = m$, each data subset for each module is taken from a separate data source. When $s< m$, some large input data sources can be separated into small datasets to provide enough for each processing module. When $s> m$, some small input data sources can be merged into a larger dataset to provide enough for each processing module.

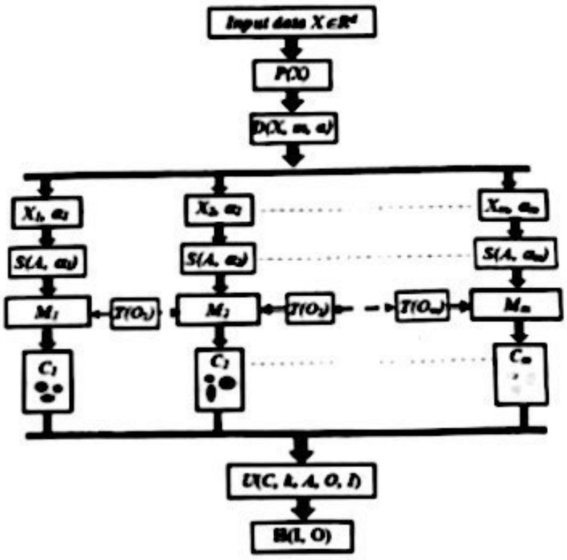

Fig. 2. The ensemble feature-reduction clustering model EFRC

4. Data preprocessing module P

$P=P(X)$ is the pre-processing techniques such as dimensional reduction techniques (PR, PCA, or Sammon) or noise filtering techniques, feature selection, etc.

Dimensionality reduction techniques have been studied and applied in many fields of data mining such as data classification and clustering [25]. In data clustering, especially for datasets with a large number of dimensions, dimensional reduction techniques are used as a preprocessing step before clustering to produce the main clustering results more accurately and to improve clustering times. The selection of an appropriate dimension reduction technique can help to enhance the processing speed and reduce the time and effort required to extract valuable information. Currently, there are many different dimensionality reduction methods such as Principal Components Analysis (PCA) [26], Random Projection (RP) [27], Sammon [28], FRFCoC [29]. In this paper, we use the Random projection algorithm to reduce the clustering data feature because it is a powerful method of dimensionality reduction that is noted for its simplicity [30]. Random projection is a powerful dimension reduction technique that uses random projection matrices to project data from a high-dimensional subspace to a low-dimensional subspace [31].

5. Data spliting module D

$$D=D(X, m, \alpha) \quad (3)$$

D is the data splitting module that divides X data into m different data subsets, $X=\{X_1, X_2, \ldots, X_m\}$, satisfying $X=\{X_1 \cup X_2 \cup \ldots \cup X_m\}$ and $X_1 \cap X_2 \cap \ldots \cap X_m = \varnothing$.

m_i is the number of items in X_i, that is $m_i = |X_i|, x_{ij} \in R^d, X_i=\{x_{i1}, x_{i2}, \ldots, x_{im i}\}$.

$\alpha=\{\alpha_1, \alpha_2, \ldots, \alpha_m\}$ is a set of influence factors, α_i is the influence coefficient of the data module X_i. α_i determined by the ruleset (4).

$$R_1 = \begin{cases} \textit{if } X_i \text{ is small and clearly, } \textit{then } \alpha_i = 0 \\ \textit{if } X_i \text{ is low-dimensions and uncertainty, } \textit{then } \alpha_i = 1 \\ \textit{if } X_i \text{ is high-dimensions or uncertainty, } \textit{then } \alpha_i = 2 \\ \textit{if } X_i \text{ is high-dimensions and high-uncertainty, } \textit{then } \alpha_i = 3 \end{cases} \quad (4)$$

6. Clustering algorithm selection module

$$S = S(A, \alpha) \quad (5)$$

S is a clustering algorithm selection function. $S=\{S_1, S_2, \ldots, S_m\}$.

Where,

A={KM, FCM, FCoC, IVFCoC} is the list of clustering algorithms that have been presented in section 2.

$\alpha=\{\alpha_1, \alpha_2, \ldots, \alpha_m\}$ is the set of influence factors. S_i is a clustering algorithm for module *ith*, S_i is determined by the ruleset (6).

$$R_2 = \begin{cases} \textit{if } \alpha_i = 0 \textit{ then } S_i = K-\textit{means algorithm} \\ \textit{if } \alpha_i = 1 \textit{ then } S_i = F-\textit{cmeans algorithm} \\ \textit{if } \alpha_i = 2 \textit{ then } S_i = \textit{FCoC algorithm} \\ \textit{if } \alpha_i = 3 \textit{ then } S_i = \textit{IVFCoC algorithm} \end{cases} \quad (6)$$

Currently, there are many different efficient algorithms. But using the four algorithms KM, FCM, FCoC, and IVFCoC is only a specific illustration of a proposed model, not a rigid one. We can integrate any algorithm so that it fits our actual needs.

7. Data clustering modules

$$M = \mathrm{M}(X, k, A, I, C) \quad (7)$$

Where, input dataset of clustering modules: $X=\{X_1, X_2, \ldots, X_m\}$ are data subsets; k is the number of data clusters; $A=\{A_1, A_2, \ldots, A_m\}$ is a set of clustering algorithms. Algorithm A_i is used to cluster each subset X_i into k different clusters. I is a set of indicators used to evaluate cluster quality (set of quality evaluation indicators), $I=\{MSE, IQI, DBI, XBI\}$ (see Table 1). Through I, during the learning repetition, if a module has a better clustering quality, the clustering results of that module can be shared with the remaining modules. Set of clustering results: C is the result set of clustering modules $C=\{C_1, C_2, \ldots, C_m\}$. $C_i=\{C_{i1}, C_{i2}, \ldots, C_{ik}\}$. $C_{ij} \in R^D, i=\overline{1,m}, j=\overline{1,k}$.

Each module $M_i = \mathrm{M}_i(X_i, k, A_i, I, C_i)$ is used to group data subsets X_i into k clusters. Where A_i is the clustering algorithm, I is the index set shared for the clustering modules, and C_i is the clustering result set of the clustering module M_i.

8. Knowledge exchange module T

$T=T(C)$ is a function that converts clustering knowledge output at each clustering iteration between clustering modules.

9. The base clustering consensus module

U is a module to consensus the cluster results to obtain global cluster results.

$$U = U(C, k, A, O, I) \quad (8)$$

a) The input of clustering consensus module

C is the input of clustering consensus module, $C=\{C_1, C_2, \ldots, C_m\}$, $C_i=\{C_{i1}, C_{i2}, \ldots, C_{ik}\}$, $C_{ij} \in R^D, i=\overline{1,m}, j=\overline{1,k}$. That is, $|C| = m * k$.

b) Number of data layers: k is the number of data clusters.

c) Clustering consensus technique A:

A is a clustering consensus technique. Algorithm A is used to classify $m*k$ items of C into k different clusters. The consensus technique could be a classification technique such as Logistic regression or Naive Bayes classifier or Support vector machines. Then, the EFRC model is called the supervised clustering ensemble model. The consensus technique can also be an unsupervised clustering technique such as KM, FCM, FCoC, or IVFCoC. Then, the EFRC model was called the unsupervised clustering ensemble model.

d) The output of clustering consensus module:

O is the output of clustering consensus module, O={O$_1$, O$_2$, …, O$_k$}, $i=\overline{1,k}$ are the result clusters, $O_i=\{x_{i,1}, x_{i,2}, \ldots, x_{i,ni}\}$, $n_i = |O_i|$ is the number of data objects in the resulting cluster O_i, $x_{i,j} \in R^d$ is the jth data object in the resulting cluster O_i.

e) Clustering quality evaluation indexes:

I is a set of indicators used to evaluate the quality of clustering result unifying. If A is a classification technique, then I is a set of supervised indexes, $I=\{MSE,\ IQI,\ DBI,\ XBI\}$. Else if A is an unsupervised clustering technique, then I is a set of unsupervised indexes, $I=\{AR,\ RcI,\ PcI,\ F1\}$ (see Table 1).

10. The global clustering result display module

$$H=H(I,O) \qquad (9)$$

H is the global clustering result display module, I is a set of the consensus quality evaluation indexes. O is the set of final results of the EFRC model, includes the class center and the distribution of the items in each class.

B. Compare EFRC with some other machine learning models

To have a better view of the EFRC model, let us compare EFRC with some other data processing models that are similar to the parallel processing model and the swarm intelligence model. The processing modules in these models are called individuals. The parallel and swarm models consisting of similar individuals so they can be called swarms. EFRC consists of different individuals, so EFRC can be called the combination swarm or the population. In Table 2, the basic characteristics of the EFRC model are compared with the swarm models and parallel processing models. According to the comparison results in Table 2, we can easily see that the EFRC model has outstanding advantages compared to other models.

C. EFRC algorithm

Based on the EFRC model presented and analyzed in Section 2 and the association of the components in the EFRC model are depicted in Fig. 2, we build a clustering algorithm. We call the algorithm EFRC as indicated in algorithm 1 below.

Table 2. Compare EFRC model with swarm models and parallel processing models

Features	EFRC model	Swarm models	Parallel processing models
Data	Dividing for the individuals	Sharing for the individuals	Dividing for the individuals
Data effects	Estimating coefficients that affect sub datasets	No	No
Objective function	Multi-objective	Multi-objective	Single-objective
Processing	Parallel processing between the individuals on different sub datasets	Parallel processing between the individuals on the same dataset	Parallel processing between the individuals on different sub datasets
Knowledge	Exchanging between the individuals	Exchanging between the individuals	No
Processing strategies	3 steps: Step 1 reduce data features, step 2 clustering on the individuals, and step 3 aggregating clustering results	1 step: Searching and selecting the best result	2 steps: Step 1 clustering on the individuals, step 2 clustering results of the individuals into the final result

Algorithm 1. Pseudocode of the ensemble feature-reduction clustering algorithm EFRC

```
Input: Dataset X
Output: The clustering results
1. Initialize parameters of $\Sigma=\{m,R,X,P,D,S,M,T,U,H\}$
2. Reading structured input data $I=\{X, d, s\}$.
3. Reduce features of data.
4. Split data into m components using $D=D(X, m, a)$ and determine the rule $R_1$
5. Clustering algorithm selection using function $S=S(A,a)$ and the rule $R_2$
6. Begin repeat
7. Clustering on modules $M(X,k,A,I,C)$;
8. Quantify indexes $I=\{MSE, IQI, DBI, XBI\}$.
9. Quantify the knowledge on each module $T=T(C)$.
10. Compare the knowledge on each module. If it is better, share it with the other individuals.
```

```
11. Check the stop condition on each clustering module.
12. End repeat: All clustering modules complete?
13. Consensus clustering results using $U=\{C,k,A,O,I\}$.
14. Quantify the estimating indicators $I=\{$Pre., Rec., F1, Acc.$\}$.
15. Output the clustering result $H=\{I,O\}$.
```

In algorithm 1, the stopping condition in step 11 ensures that all basic clusterings converge on the criterion that the membership function does not change after a few iterations. Mean,

$$\left\|\left|U_{i=1}^{m}(\tau)\right| - \left|U_{i=1}^{m}(\tau-1)\right|\right\| < \varepsilon \qquad (10)$$

The algorithm based on the EFRC model has some improvements compared to traditional clustering ensemble models such as feature reduction in step 3, multi-objective base clusterings in step 5, quantification of cluster quality index in step 8, select the best centroid in step 9 and share the best centroid for other base clusterings.

IV. EXPERIMENT RESULTS

In this section, we present some experimental results to simulate the working mechanism of the EFRC model and demonstrate the effectiveness of the proposed clustering ensemble method. The EFRC model is a combination of four single algorithms KM, FCM, FCoC, and IVFCoC for four base clusterings. In the EFRC model, we use a random projection algorithm to reduce the features of the data. We divide the original dataset into four equal parts and give each base clustering one part of the data. We used the Silhouette-Based Assessment of Cluster Tendency algorithm [3] to assess the clustering tendency of clusters obtained from four clustering modules.

For a fair comparison, we have installed clustering experiments along with state-of-the-art methods such as single clustering (KM, FCM, FCoC, and IVFCoC) and single-objective clustering ensemble (eFCoC) methods. However, in the single clustering experiments, the experimental results of KM on multi-feature data are of too low quality compared with other FCM, FCoC and IVFCoC single fuzzy algorithms. Therefore, we do not state the test results with KM.

To quantify the clustering quality of different algorithms, we use the indices Accuracy rate [19], Recall index [20], Precision index [20], and F1 score [21]. The higher the index value, the better the corresponding cluster quality (see Table 1).

Experiments are implemented on Windows 7 of HP Elitebook 8560W, Core i7-2670QM, 8 GB RAM, NVIDIA Quadro 2000M, and C#.Net development environment.

Firstly, we present experimental results on the many-feature datasets and small size. Includes three datasets Dim256, Dim512, and Dim1024 which were downloaded from the clustering data repository of the School of the Computing University of Eastern Finland[1]. These datasets have many features d from 256 to 1024 and 1024 data objects are evenly distributed over 16 Gaussian clusters. Each cluster has 64 data objects in sequential order. The statistics of these datasets are summarized in Table 3.

***Table 3**. The details of used benchmark datasets.*

Name	Size	#Clusters	#Features
Dim256	1024	16	256
Dim512	1024	16	512
Dim1024	1024	16	1024
PEMS-SF	440	7	138672
Radar	325834	7	175

The goal of these experiments is to prove that the clustering quality of the EFRC algorithm is superior to its single algorithm. The use of small and labeled datasets will help us to easily control the operation progress of our experiments. The experimental results are quantified by the validity indexes in Table 4. In Table 4, we highlight the best results in bold.

***Table 4**. Clustering results of algorithms FCM, FCoC, IVFCoC, eFCoC and EFRC on datasets Dim128, Dim256 and Dim1024*

Datasets	Alg.	Pre.	Rec.	F1	Acc.	Time (Sec)
Dim256	FCM	0.846	0.830	0.830	0.860	14.8
	FCoC	0.951	0.950	0.950	0.953	12.2
	IVFCoC	0.982	0.981	0.981	0.982	26.9
	eFCoC	0.982	0.967	0.973	0.980	**1.24**
	EFRC	**0.998**	**0.997**	**0.997**	**0.997**	4.12
Dim512	FCM	0.828	0.818	0.819	0.850	19.6
	FCoC	0.960	0.959	0.959	0.961	13.5
	IVFCoC	0.987	0.986	0.986	0.987	42.6
	eFCoC	0.990	0.984	0.990	0.992	**1.83**
	EFRC	**0.996**	**0.995**	**0.995**	**0.995**	3.45
Dim1024	FCM	0.817	0.812	0.810	0.848	25.3
	FCoC	0.956	0.955	0.955	0.957	22.8
	IVFCoC	0.990	0.990	0.990	0.990	59.3

[1] http://cs.joensuu.fi/sipu/datasets/

	eFCoC	0.989	0.991	0.990	0.990	**2.35**
	EFRC	**0.998**	**0.997**	**0.997**	**0.997**	5.21

In Table 3, we easily see that the value of the indexes obtained from our method is better than the previously proposed algorithm. Meanwhile, the time consumed by the eFCoC algorithm is the smallest. The results in Table 3 show that the time consumption of the EFRC algorithm is higher than that of the eFCoC algorithm. This can be explained as follows: Theoretically, the computational complexity of the EFRC algorithm is higher than single algorithms and the eFCoC algorithm because EFRC adds a few functions such as feature reduction, multi-object, and optimal centroid sharing. These improvements make the EFRC algorithm more accurate than traditional algorithms. For time consumption: Although the computational complexity of EFRC is higher than other algorithms, the base clusterings are installed in parallel on 25% of the original data, so the total time consumption is lower than other single algorithms. However, since the base clusterings of the eFCoC are also installed in parallel on 25% of the data, the consumption time of the eFCoC is lower than the EFRC algorithm. This is completely logical. In the next experiments, we cluster on the many-feature and labeled datasets. The goal of these experiments is to demonstrate the potential of the EFRC algorithm on real datasets. The two datasets are downloaded from the UCI Machine Learning Repository[2]. The data set PEMS-SF is 400 MB in size and includes 440 data objects. This dataset has 138672 features and 440 data objects that are grouped into seven different clusters. This data describes the occupancy rate, between 0 and 1, of different car lanes of San Francisco bay area freeways. Measurements cover the period from Jan. 1st, 2008 to Mar. 30th, 2009 and are sampled every 10 minutes. We treat each day in this database as a single time series of dimension 963 (the number of sensors that functioned consistently throughout the studied period) and length 6 x 24=144. This results in a database of 440 time series correspond to 440 data objects. Each data object is labeled with an integer in {1,2,3,4,5,6,7} corresponding to a day of the week from Monday to Sunday.

The radar dataset is 411MB in size and 325834 data objects with the number of features is 175. Radar dataset is a fused bi-temporal optical-radar data for cropland classification. The images were collected by RapidEye satellites (optical) and the Unmanned Aerial Vehicle Synthetic Aperture Radar (UAVSAR) system (Radar) over an agricultural region near Winnipeg, Manitoba, Canada in 2012. There are 2 * 49 radar features and 2 * 38 optical features for two dates: 05 and 14 July 2012. Seven crop type classes exist for this dataset as follows: 1-Corn; 2-Peas; 3- Canola; 4-Soybeans; 5- Oats; 6- Wheat; and 7-Broadleaf. The statistics of these datasets are summarized in Table 3. The experimental results are quantified by the validity indexes in Table 5.

Table 5. *Clustering results of algorithms FCM, FCoC, IVFCoC, eFCoC and EFRC on datasets PEMS-SF and Radar*

Datasets	Alg.	Prec.	Rec.	F1	Acc.	Time (min)
PEMS-SF	FCM	0.853	0.840	0.846	0.875	118
	FCoC	0.946	0.946	0.946	0.949	96
	IVFCoC	0.965	0.964	0.964	0.965	138
	eFCoC	0.953	0.951	0.953	0.954	34
	EFRC	***0.978***	***0.978***	***0.977***	***0.981***	***10***
Radar	FCM	0.846	0.842	0.846	0.845	142
	FCoC	0.932	0.931	0.933	0.933	116
	IVFCoC	0.947	0.943	0.946	0.944	205
	eFCoC	0.965	0.962	0.967	0.966	39
	EFRC	***0.988***	***0.986***	***0.988***	***0.986***	***20***

Table 4 also shows us, the value of the indexes obtained from the proposed algorithm is better than the previously proposed algorithm. In addition, the time consumption of the proposed algorithm is smaller than that of the previously proposed algorithms. Once again, we can learn from Table 4 that the EFRC algorithm has obvious advantages over the other four methods in many-feature data clustering. The average correct clustering rate of the EFRC algorithm is higher than that of other methods. The results in Table 4 show that the time consumption of the EFRC algorithm is lower than that of the eFCoC algorithm. This can be explained as follows: In Table 4, the PEMS-SF dataset has a very high number of features (138672), Radar dataset has a rather large size (325834). Therefore, feature reduction is very significant for the EFRC algorithm, which significantly reduces data size. Reducing data size helps EFRC accelerate clustering faster than the eFCoC.

General, the clustering results in Table 3 and Table 4 are demonstrated that the EFRC algorithm is more accurate than single algorithms KM, FCM, FCoC, IVFCoC, and eFCoC.

V. CONCLUSION

In this paper, an ensemble mathematical model and a clustering algorithm EFRC based on the EFRC model are proposed. The model of the clustering ensemble consists of ten basic components which are analyzed in detail to make EFRC more explicit than classic ensemble clustering models. Based on the EFRC model, the EFRC algorithm is formed to cluster the many-feature data. The EFRC model is an intelligent multi-objective clustering model that takes full advantage of clustering techniques to contribute two valuable rule sets R_1 and R_2. The EFRC algorithm

[2] https://archive.ics.uci.edu/ml/datasets.php

combines feature reduction algorithm, based on the divide-and-conquer principle, and EFRC model's preeminent techniques to demonstrate potential in the many-feature data processing. The experimental results showed that the EFRC algorithm obtained better clustering accuracy and consumption time than the single clustering algorithms.

Hyperspectral images have wide observability, high resolution, and feature numbers from hundreds to thousands. The hyperspectral image data plays an important in quantitative remote sensing, military, environmental management, mineral mining, biological and medical, precision agriculture applications. In the future, we will apply the EFRC algorithm to conduct further applications of classification, target detection, and change detection.

REFERENCES

[1] S. Miyamoto, H. Ichihashi, K. Honda, "Algorithms for Fuzzy Clustering," Springer: Studies in Fuzziness and Soft Computing, Vol. 229, 2008.

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

[3] Pham Van Nha, Pham The Long, Nguyen Duc Thao, Ngo Thanh Long, "A new cluster tendency assessment method for fuzzy co-clustering in hyperspectral image analysis," Neurocomputing, Vol. 30713, pp. 213-226, 2018.

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

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

[6] 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.

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

[8] Y. Kazemi, S. Abolghasem, Mirroshandel, "A novel method for predicting kidney stone type using ensemble learning," Artificial Intelligence in Medicine, Vol. 84, pp. 117-126, 2018.

[9] L. Bai, J. Liang, F. Cao, "A multiple k-means clustering ensemble algorithm to find nonlinearly separable clusters," Information FusionIn press, In press, pp. 1-38, 2020.

[10] X. Zhao, F. Cao, J. Liang, "A sequential ensemble clusterings generation algorithm for mixed data," Applied Mathematics and Computation, Vol. 33515 2018, pp. 264-277.

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

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

[13] 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.

[14] Le Thi Cam Binh, Ngo Thanh Long, Pham Van Nha, Pham The Long, "A new ensemble approach for hyper-spectral image segmentation," Conference on Information and Computer Science (NICS), 2018.

[15] J. Dunn, "A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters," *Journal of Cybernatics*, Vol. 3, pp. 32-57, 1973.

[16] C.H. Oh, K. Honda, H. Ichihashi, "Fuzzy Clustering for Categorical Multivariate Data," *IFSA World Congress and 20th NAFIPS International Conference*, pp. 2154-2159, 2001.

[17] Pham Van Nha, Ngo Thanh Long, W. Pedrycz, "Interval-valued fuzzy set approach to fuzzy co-clustering for data classification," Knowledge-Based Systems, Vol. 107, pp. 1-13, 2016.

[18] P. Fränti, O. Virmajoki, V. Hautamäki, "Fast agglomerative clustering using a k-nearest neighbor graph," IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 28 (11), pp. 1875-1881, 2006.

[19] J.C. Bezdek, "Cluster validity with fuzzy sets," Journal Cybernetics, Vol. 3, pp. 58-73, 1974.

[20] I.S. Dhillon, S. Mallela, D.S. Modha, "Information-theoretic co-clustering," ACM IC KDDM, pp. 89–98, 2003.

[21] M.W.P. David, "Evaluation: From Precision, Recall, and F-Measure to ROC, Informedness, Markedness & Correlation," Journal of Machine Learning Technologies, Vol. 2(1), pp. 37-63, 2011.

[22] Z. Wang, A.C. Bovik, "A universal image quality index," IEEE signal processing letters, Vol. 9(3), pp. 81-84, 2002.

[23] W. Wang, Y. Zhang, "On fuzzy cluster validity indices," Fuzzy Sets Systems, Vol. 158, pp. 2095–2117, 2007.

[24] Chris Ding, "Dimension Reduction Techniques for Clustering," Encyclopedia of Database Systems, 2009

[25] I. Jolliffe, "Principal Component Analysis," Springer, 2002.

[26] Achlioptas, "Database-friendly random projections, Johnson-Lindenstrauss with binary coins," Journal of Computer and System Sciences 66, pp. 671–687, 2003.

[27] B. Lerner, H. Guterman, M. Aladjem, and I. Dinstein, "On the initialisation of Sammon's nonlinear mapping," Pattern Analysis and Applications, vol. 3, pp. 61–68, 2000.

[28] Pham Van Nha, Pham The Long, Witold Pedrycz, and Ngo Thanh Long, "Feature-Reduction Fuzzy Co-Clustering approach for hyper-spectral image analysis", EEE International Conference on Fuzzy Systems (FUZZ-IEEE), 2017.

[29] E. Bingham, H. Mannila, "Random projection in dimensionality reduction: Applications to image and text data," Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining, 2001.

[30] N. Goel, G. Bebis, A. Nefian, "Face recognition experiments with random projection," SPIE, Vol. 5779, pp. 426–437, 2005.

[31] Z. Wang and A.C. Bovik, "Mean squared error: love it or leave it? A new look at signal fidelity measures," IEEE Signal Processing Magazine, Vol. 26(1), 98–117, 2009.

TIẾP CẬN MÔ HÌNH ĐỒNG THUẬN ĐỂ PHÂN CỤM DỮ LIỆU NHIỀU ĐẶC TRƯNG

Tóm tắt: Đồng thuận là một mô hình học máy phổ biến dựa trên nguyên tắc chia để trị. Trong phân cụm dữ liệu, đồng thuận nhằm mục đích cải thiện hiệu suất về tốc độ xử lý và chất lượng phân cụm dữ liệu. Hầu hết các phương pháp đồng thuận hiện có đang đối mặt với những thách thức phức tạp như không chắc chắn, không rõ ràng và lấp chồng. Kỹ thuật phân cụm mờ gần đây đã được phát triển để xử lý dữ liệu nhiều đặc trưng, không đồng nhất, không chắc chắn và kích thước lớn. Trong bài báo này, chúng tôi đề xuất một mô hình đồng thuận phân cụm giảm đặc trưng (EFRC) sử dụng các kỹ thuật học máy tiên tiến. Mô hình EFRC bao gồm ba giai đoạn. Đầu tiên, dữ liệu được giảm bớt một số đặc trưng bằng phép chiếu ngẫu nhiên. Sau đó, dữ liệu được chia thành các tập con dựa trên mức độ chồng chéo và định lượng nhiễu. Các kỹ thuật phân cụm khác nhau được sử dụng để phân cụm các tập hợp con dữ liệu. Cuối cùng, kết quả của các mô-đun phân cụm được đồng thuận bằng cách sử dụng kỹ thuật phân loại để tạo ra kết quả phân cụm cuối cùng. Một vài thực nghiệm được thực hiện trên các bộ dữ liệu mẫu chuẩn. Kết quả thử nghiệm cho thấy hiệu suất vượt trội của mô hình EFRC so với các mô hình trước đó.

Từ khóa: Phân cụm, phân loại, mô hình đồng thuận, giảm đặc trưng, nhiễu đặc trưng, dữ liệu lớn.

Pham The Long received the PhD and DSc degrees from Belorussian State Univ., Minsk, in 1982 and 1987, respectively. He is currently a Professor with the Department of Information Technology, Le Quy Don Technical University, Hanoi, Vietnam. His current research interests include optimization, fuzzy logic and virtual reality.

Science and Technology, Hanoi, Vietnam. Her research interests include pattern recognition, image processing, and data mining.

Le Thi Cam Binh received the B.S degree and the M.S degree in Computer Science from VNU University of Science, Hanoi, Vietnam, in 1997 and 2005. She is currently working toward the PhD degree in the Academy of Military

Ngo Thanh Long received the M.SC, Ph.D degrees from Le Quy Don Technical University, Vietnam, in 2003 and 2009, respectively. He is currently a Associate Professor with the Department of Information Technology and Institute of Simulation Technology, Le Quy Don Technical University, Hanoi, Vietnam. His current research interests include computational intelligence, type-2 fuzzy logic, pattern recognition and image processing.

Pham Van Nha received the M.SC, Ph.D degrees from Le Quy Don Technical University, Vietnam, in 2009 and 2018, respectively. He is currently work for Information Technology Institute, Hanoi, Vietnam. His research interests include pattern recognition, image processing, and data mining.