A new ensemble approach for hyper-spectral image segmentation

Le Thi Cam Binh
Academy of Military S&T
17 Hoang Sam, Cau Giay
Hanoi, Vietnam
lcbinh@yahoo.com

Pham Van Nha
Academy of Military S&T
17 Hoang Sam, Cau Giay
Hanoi, Vietnam
nhapv@mod.gov.vn

Ngo Thanh Long
Le Quy Don University
236 Hoang Quoc Viet
Hanoi, Vietnam
ngotlong@mta.edu.vn

Pham The Long
Le Quy Don University
236 Hoang Quoc Viet
Hanoi, Vietnam
longpt@mta.edu.vn

***Abstract*—The ensemble is an universal machine learning method that is based on the divide-and-conquer principle. In data clustering, ensemble aims to improve performance in terms of processing speed and clustering quality. Most existing ensemble methods become more difficult due to the inherent complexities such as uncertainty, vagueness and overlapping. In this paper, we proposed a new ensemble method that improve the ability to identify uncertainty issues, deal with the noise, and accelerate hyperspectral image data clustering. We called fuzzy co-clustering ensemble algorithm (eFCoC). eFCoC uses fuzzy co-clustering algorithm (FCoC) to clustering data and silhouette-based assessment of cluster tendency algorithm (SACT) to ensemble the final clustering result. Experiments were conducted on synthetic data sets and hyper-spectral images. Experimental results demonstrated the key properties, rationality, and practicality of the proposed method.**

***Index Terms*—Fuzzy co-clustering, clustering ensemble, assessment of cluster tendency, hyper-spectral image, image segmentation**

I. Introduction

Ensemble is an universal machine learning method that is based on the divide-and-conquer principle. It is constructed with a set of individual models working in parallel, whose outputs are combined with a decision fusion strategy to produce a single answer for a given problem [1]-[4]. The models can be classification, prediction, regression or clustering, that the ensemble is designed to do. Clustering ensemble is a machine learning method for data clustering. Clustering ensemble aims to combine multiple clustering models to produce a better result than that of the individual clustering algorithms in terms of consistency and quality [1]. Since clustering ensemble has been proposed, it has rapidly attracted much attention. Some recent research on ensemble in machine learning fields such as mining industry [5], biology and medicine [6]-[8], pattern recognition [9]-[11], categorical data [12]-[14], image processing [15]-[17], environmental management [18], [19], and big data processing [20]. Generally, clustering ensemble has been shown to be very effective in unsupervised learning. It fits more data sets than clustering and it is also robust against noise and outliers. However, most existing ensemble algorithms are based on static model [27], they become more difficult due to the inherent complexities such as uncertainty, vagueness and overlapping. In this paper, we proposed a new ensemble method that improve the ability to identify uncertainty issues, deal with the noise, and accelerate hyperspectral image data clustering. We called fuzzy co-clustering ensemble algorithm eFCoC. First, the data is divided into different parts. Then, FCoC algorithm is used to clustering these data parts. Finally, SACT algorithm is used to fuse the obtained results from FCoC algorithms. eFCoC is seen as a dynamically ensemble clustering model that consists of three main stages. The dynamic characteristic of this model is that each FCoC module can be adjusted to deal with the degree of noise and uncertainty on each data subsets through the fuzzy parameters of the objective function FCoC. This paper is organized as follows. Section 2 presents in detail the proposed method including review of the techniques related for our proposed approach. In Section 3 a set of experimental results is presented to demonstrate the effectiveness of the proposed method. The discussion of the results is also given in the end of Section 3. The conclusions are given in Section 4.

II. Fuzzy co-clustering ensemble methods

A. The concept of clustering ensemble

The definition of clustering ensemble [3] is as follows: there is a data set $X = \{x_1, x_2, ..., x_N\}$ that has N data point. Data X is divided into M

different data subsets $X = X1, X_2, \ldots, X_M$. Then, M clustering algorithms are used to clustering these data subsets $X_i\,(i = 1, 2, \ldots, M)$ 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 final clustering result $P*$. In addition, for clustering quality evaluation, eFCoC is added to the evaluation phase. The intuitive illustration of clustering ensemble is shown in Fig. 1.

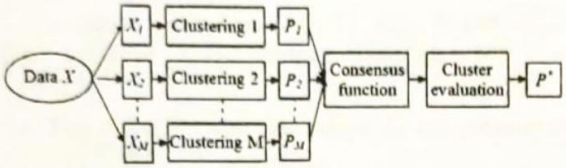

Figure 1. Framework of clustering ensemble

B. Proposed fuzzy co-clustering ensemble algorithm

In this section we will present and analyze the components of proposed fuzzy co-clustering ensemble algorithm, which is called eFCoC algorithm.

1) Splitting input data: The first task of the ensemble algorithm is to divide the data into different M subsets. The purpose of this task is parallel processing to accelerate overall processing. In addition, the data splitting is also used to isolate the noise and uncertain data components which can be handled separately more effectively.

2) Clustering algorithm: Selecting the clustering algorithm is one of the most important tasks for the ensemble methods. The most existing ensemble algorithms are based on static model. Such as the method in [28] uses the k-means algorithm for sentiment analysis, in [11] uses k-nearest neighbor for attack samples classification, in [8] uses support vector machine for breast cancer diagnosis. The drawbacks of these models are sensitivity to data size and uncertainty. FCoC is an unsupervised learning technique used to solve multi-dimensional clustering problems such as clustering documents and keywords [21, 23], multi-dimensional data classification [24, 25], color image segmentation [22], multi-spectral and hyper-spectral clustering [26]. FCoC uses T_u and T_v fuzzy parameters to adjust the detection ability of noise and uncertainty. The drawback of improved fuzzy co-clustering methods is high computational complexity, especially when there is an increase in data size. In this paper we have used FCoC in the ensemble model to overcome some limitations of FCoC on computational complexity and uncertainty.

3) Partitions: $P_1, P_2, ..., P_M$ are the partitions obtained from the clustering modules that contain the clusters of the corresponding data subsets. The consensus function treat these data clusters as super-objects. Let $N_1, N_2, ..., N_M$ be the corresponding number of clusters of $P_1, P_2, .., P_M$, so we have a set of super-objects X_{New} with sizes $N_{New} = N_1 + N_2 + ... + N_M$. Thus, instead of clustering the original X data set with the size N, we can clustering X_{New} data set with size of N_{New}.

4) Consensus function: In the clustering ensemble model, the consensus function serves as a fusion technique of clustering results. Normally, the number of clusters is unknown, so the consensus function must also have the function of determining the appropriate cluster solution for each particular data set. A variety of consensus functions have been developed and can be classified to four major categorizations: direct, feature-based, pairwisesimilarity based and graph-based approaches [2]. Where graph-based approaches are more interested [3]. SACT algorithm [29] is one graph-based approache that have proven effective in cluster trend evaluation. In this paper, SACT algorithm is used as a consensus function to fusion the clustering results and determine the appropriate number of data clusters.

5) Cluster evaluation: In this phase, the aim is to evaluate the quality of the final clustering result. This is an important task to determine whether the proposed method is effective. In this paper, we use the assessment indices as Partition Coefficient index (PC) [30], Mean Squared Error index (MSE) [31], Image Quality Index (IQI) [32], Recall and Precision [33]. Note that, PC, IQI, Recall, Precision are larger and MSE is smaller, the clustering quality is better.

6) Schema of the proposed algorithm eFCoC: eFCoC algorithm uses FCoC for data clustering and SACT to fusion the final clustering result. Schema of eFCoC algorithm is shown in Fig. 2. In this algorithm, FCoC is described in detailed in [22] and SACT is described in detailed in [29]. eFCoC algorithm demonstrates flexibility in dealing with the noise and uncertainty by the ability to adjust T_u and T_v fuzzy parameters for individual data subsets. This is very meaningful for data sets that have distributed distribution. Such as image data, so splitting data can help us isolate noise for more efficient processing.

Table III
CLUSTERING RESULTS USE EFCOC AND FCOC SINGLE ON HYPER-SPECTRAL IMAGE DATA SETS

Data set	Algorithm	PC	MSE	IQI	τ
GoMWs image	FCoC	0,92	76,8	0,94	14
	eFCoC	0,97	73,8	0,98	71
A&Vs image	FCoC	0,87	159,2	0,91	9
	eFCoC	0,97	136,5	0,95	74

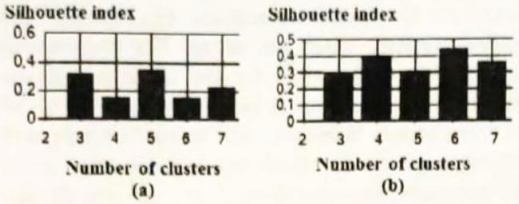

Figure 2. Chart of silhouettes index by the number of clusters using SACT a) on GoMWs image; b) on A&Vs image

According to the chart of silhouettes and number of clusters, the SACT algorithm has determined the number of appropriate clusters of the respective data sets 5, 6. The clustering result is expressed as a layered image as shown in Fig. 3.

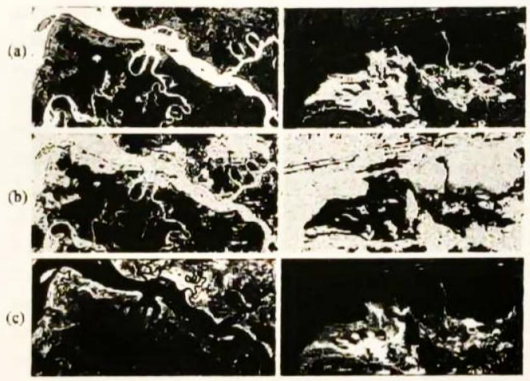

Figure 3. Hyper-spectral image clustering results (a) using FCoC single; (b) using eFCoC; (c) Three-band color composite using ENVI software

According to the experimental results in Table 3, the proposed algorithm achieves performance better than FCoC algorithm. However, eFCoC's the average number of loops is higher than FCoC. The visual result of eFCoC algorithm is basically similar to that of ENVI image and FCoC.

Experiments show that the original results demonstrate the potential of the proposed algorithm. Although the experiments were just only conducted on several multivariate data sets, the performance comparison was performed only with the single FCoC algorithm. But with the results of this study, this may be the basis for deeply further research on ensemble fuzzy co-clustering in multi-dimensional clustering such as the hyper-spectral image.

C. Discussion

1) Comparison against base clustering: The purpose of ensemble clustering is to combine multiple base clusterings into a probably better and more robust clustering. In this section, we compare the eFCoC approach against the base clusterings on the benchmark data sets. We run the eFCoC approach 100 times on each data set. For each run, an ensemble of M base clusterings is randomly generated and validity indexes PC, MSE, IQI and the consensus function by SACT are computed, respectively. The average validity indexes over 100 runs of eFCoC and the base clusterings are listed in Table III-A. Basically, eFCoC approach produces significantly better clustering results than the base clusterings in terms of validity indexes. As shown in Table III-A, the proposed approach produces overall better and more robust clusterings than the base clusterings on the benchmark data sets. To evaluate the reliability and correctness of the proposed algorithm. We look at the clustering results in Fig. 3. According to this result, we can see in the clustering image obtained from the data set "Gulf of Mexico Wetland Sample". Although data were splited, but when pairing the resulting images there is no disruption in the color areas. This shows the correctness and reliability of the proposed method.

2) Comparison against non-automatic ensemble clustering approaches: In this section, we compare the proposed eFCoC approach against non-automatic ensemble clustering approach, that is single FCoC. Note that the proposed eFCoC approach is an automatic approach. This non-automatic approach lack the ability to estimate the cluster number automatically and need to take the cluster number of the final clustering as input. In our experiments, for comparison, the cluster number of non-automatic approach is set to be the same as the automatically estimated cluster number by eFCoC using SACT algorithm [29].

3) Robustness to ensemble size M: In this section, we further test the robustness of our approach w.r.t. varying ensemble size M. For each ensemble size M, we run eFCoC 100 times, respectively, and report their average performance w.r.t. varying M from 2

to 16 on six high-dimensional synthetic data sets and M from 2 to 120 on two hyper-spectral image data sets. Experimental results show that there are only a few values of M where clustering performance of eFCoC is highest. There are even more values of M where the performance of eFCoC is lower than single FCoC. If M is too large, data subsets are too small, which cause the clustering performance to decrease and the ability of parallel processing of the system will overload. In addition,the noise and uncertainty components will appear on many different subsets, which results in reduced clustering performance. If M is too small, the goal is not achieved as desired. The performance of the proposed algorithm is not much different than the single clustering algorithm. Specifically, for six high-dimensional synthetic data sets M = 4 and for two hyper-spectral images data sets M = 60 are appropriate.

4) Robustness to select parameters: One of the advantages of FCoC algorithms is the ability to recognize noise and uncertainty through adjusting T_u and T_v fuzzy parameters. In this section we investigated the effect of these parameters on eFCoC algorithm. For each ensemble size M = 4, we run eFCoC 100 times on six high-dimensional synthetic data sets, with parameter set T_u = 2, T_v = 10^6 and C = 16. The clustering results are quantified through the mean of values of PC, MSE, and IQI validity indices. There are always some data subset modules that results in poor clustering because of the effects of noise and uncertainty. After a few adjustments to the value of T_u and T_v on these modules we have received better clustering results. We did the same survey for two hyper-spectral images and found that many number of modules had poor results. This can be concluded that the two hyper-spectral images exhibit more noise and uncertainty than six high-dimensional synthetic data sets.

Similar to traditional fuzzy co-clustering algorithms, selecting the initial value of the T_u and the T_v parameters is very important, which greatly affects the quality of the final clustering. In these experiments, the parameters T_u and T_v are known parameters. The value of these parameters is determined using the parameter search method that was used in [25].

The main objective of selecting parameter M is to accelerate clustering. The determination of the M parameter depends on the size of the data and the number of actual processors that can be executed in parallel. The M is the larger, the size of sub-data is the smaller, the quality of clustering on each sub-data is more inaccurate. So we have to choose M such that the size of the subset of subsets is not too small and M must be commensurate with the parallel ability of the respective microprocessors.

5) Execution time: In this section, we test the execution time of eFCoC and the baseline approaches. We first use a fixed ensemble size M= 4 for six high-dimensional synthetic data sets and M=60 for two hyperspectral image data sets to evaluate the time expenses of different approaches. Then we compare the efficiency of these ensemble clustering approaches with varying ensemble sizes. All experiments are conducted in Visual studio 2010, 64-bit on a workstation (Windows 7 64-bit, HP EliteBook 8560w with Intel® Core™ i7 - 2820QM, VGA Nvidia Quadro 1000M). Experimental results in Tables 1 show that for high-dimensional synthetic data sets, the average number of loops of FCoCs in eFCoC and the total of consumption time of eFCoC is always lower than that of single FCoC. Experimental results in Tables 3 show that for hyper-spectral image data sets, the average number of loops of FCoCs in eFCoC is higher but the total of consumption time of eFCoC is always lower than the single FCoC.

IV. Conclusion

In this paper we have proposed an new fuzzy co-clustering algorithm following the recent ensemble trend to improve clustering performance on large data. First, the data is divided into different parts. Then, fuzzy co-clustering algorithm is used to clustering these data parts. Finally, the SACT cluster trend estimation algorithm is used to ensemble the results obtained from fuzzy co-clustering algorithms. The results of SACT algorithm are the appropriate number of clusters and cluster distribution of the original data set. The experiments were conducted on 6 sets of multivariate data and two hyper-spectral images. The experiment results demonstrate the potential of the proposed algorithm in variable data clusters.

Some research directions in the future are as follows: Research to improve eFCoC for big data clustering; Research on the application of advanced eFCoC for object detection and target recognition in the hyper-spectral image.

Acknowledgment

This research is funded by Vietnam National Foundation for Science and Technology Development (NAFOSTED) under Grant number 102.05-2016.09.

REFERENCES

[1] T. Alqurashi, W. Wang, "Clustering ensemble method," International Journal of Machine Learning and Cybernetics, pp. 1-20, 2018.

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

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

[4] B. Krawczyk, L.L.Minku, J. Gama, J. Stefanowski, M. Woźniak, "Ensemble learning for data stream analysis: A survey," Information Fusion, Vol. 37, pp. 132-156, 2017.

[5] 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.

[6] Y. Peng, A novel ensemble machine learning for robust microarray data classification, Computers in Biology and Medicine, Vol. 36(6), 2006, pp. 553-573.

[7] 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.

[8] 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, 687–699, 2018.

[9] L. Franek, X. Jiang, "Ensemble clustering by means of clustering embedding in vector spaces," Pattern Recognition, Vol. 47, pp. 833-842, 2014.

[10] B. Krawczyk, M. Galar, M. Woźniak, H. Bustince, F. Herrer, "Dynamic ensemble selection for multi-class classification with one-class classifiers," Pattern Recognition, Vol. 83, pp. 34-51, 2018.

[11] Y. Zhang, G. Cao, B. Wang, X. Li, "A novel ensemble method for k-nearest neighbor," Pattern Recognition, Vol. 85, pp. 13–25, 2019.

[12] I. Saha, J.P. Sarkar,U. Maulik, "Ensemble based rough fuzzy clustering for categorical data," Knowledge-Based Systems, Vol. 77, pp. 114-127, 2015.

[13] X. Zhao, F. Cao, J. Liang, "A sequential ensemble clusterings generation algorithm for mixed data," Applied Mathematics and Computation, Vol. 335, pp. 264–277, 2018.

[14] J. Rodríguez, M.A. Medina-Pérez and et. al., "Cluster validation using an ensemble of supervised classifiers," Knowledge-Based Systems, Vol. 145, pp. 134–144, 2018.

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

[16] B. Ayerdi, I. Marqués, M. Graña, "Spatially regularized semisupervised Ensembles of Extreme Learning Machines for hyperspectral image segmentation," Neurocomputing, Vol. 149, Part A, pp. 373-386, 2015.

[17] Y. Song, S. Zhang and et. al., "Gaussian derivative models and ensemble extreme learning machine for texture image classification," Neurocomputing, Vol. 277, pp. 53–64, 2018.

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

[19] S. Sun, S. Wang, G. Zhang, J. Zheng, "A decomposition-clustering-ensemble learning approach for solar radiation forecasting," Solar Energy, Vol. 163, pp. 189-199, 2018.

[20] S. Huang, B. Wang, J. Qiu, J. Yao, G. Wang, G. Yu, "Parallel ensemble of online sequential extreme learning machine based on MapReduce," Neurocomputing, Vol. 174, pp. 352-367, 2016.

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

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

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

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

[25] 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.

[26] V.N. Pham, L.T. Ngo, D.T. Nguyen, Feature-Reduction Fuzzy Co-Clustering algorithm for hyperspectral image clustering, IEEE Conference on Fuzzy Systems, pp. 1-6, 2017.

[27] W. Xiao, Y. Yang, H. Wang, T. Li, H. Xing, "Semi-supervised hierarchical clustering ensemble and its application, " Neurocomputing, Vol. 173(3), pp. 1362-1376, 2016.

[28] M.T. AL-Sharuee, F. Liu, M. Pratama, "Sentiment analysis An automatic contextual analysis and ensemble clustering approach and comparison," Data & Knowledge Engineering, Vol. 115, pp. 194-213, 2018.

[29] 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, pp. 213-226, 2018.

[30] J.C. Bezdek, Cluster validity with fuzzy sets, J. Cybernet., Vol. 3, 1974, pp. 58-73.

[31] Z. Wang, A.C. Bovik, Mean squared error: love it or leave it? A new look at signal fidelity measures, IEEE signal processing magazine, 2009, pp. 98-117.

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

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