Presenting a Proper Ensemble Clustering (EC) Method Based on Hierarchical Methods and Classical Generative Algorithms

Zahra Sahebkaram[1*], Alireza Norouzi[2]
1- Department of Electrical Engineering, Majlesi Branch, Islamic Azad University, Isfahan, Iran.
Email: sahebkaram@gmail.com (Corresponding author)
2- Department of Electrical Engineering, Majlesi Branch, Islamic Azad University, Isfahan, Iran.
Email: norouzi.arz@gmail.com

Received: August 2020 Revised: October 2020 Accepted: November 2020

ABSTRACT:
Ensemble Clustering (EC) methods became more popular in recent years. In this methods, some primary clustering algorithms are considered to be as inputs and a single cluster is generated to achieve the best results combined with each other. In this paper, we considered three hierarchical methods, which are single-link, average-link, and complete-link as the primary clustering and the results were combined with each other. This combination was done based on correlation matrix. The basic algorithms were combined as binary and triplicate and the results were evaluated as well. the IMDB film dataset were clustered based on existing features. CH, Silhouette and Dunn Index criteria were used to evaluate the results. These criteria evaluate the clustering quality by calculating intra-cluster and inter-cluster distances. CH index had the highest value when all three basic clusters are combined. Our method shows that EC can achieve better results and present clusters with higher robustness and accuracy.

KEYWORDS: Clustering, Correlation Matrix, Single-Link Algorithm, Average-Link Algorithm, Full-Link Algorithm.

1. INTRODUCTION

Clustering is an unsupervised method that is considered as an important part of data mining. Its algorithms are more complicated compared to supervised classification methods. Clustering is done based on the similarity between the data, where the data with the most similarity are placed in same cluster. Moreover, their clustering is such that the data of each cluster have the greatest difference with the data of other clusters. There are no prior knowledge and classifications of data in clustering. Data can be clustered based on a variety of numeric, textual, binary, and even compound features.

By measuring the distance between the data, the similarity between them can be measured. Creating appropriate subspaces and visualizing the data can also partly improve clustering. Each clustering algorithm provides an appropriate response to a specific sample of data. Sometimes, it is necessary for an algorithm to be iterated repeatedly to reach optimal clustering. For example, sometimes k-means algorithm[1] , which is a basic class algorithm, is iterated so many times to present the best clustering. It shows that an algorithm cannot always provide with certainty a correct, optimal and suitable results for different types of data. Perhaps using a particular algorithm ends in misleading and meaningless results.

The goal of EC is to combine the results of different algorithms to reduce the faults of the basic algorithms. Among the advantages of EC, minimizing the possibility of errors, increasing the accuracy and quality of the final clustering, and reducing the dependence of the initial data on the final clustering can be cited. No clustering algorithm is optimal alone, various clustering algorithms produce different partitions, then each one applies a different structure to a set of data [2]. Another advantage of EC is its robustness, which, has a better performance on the existing dataset. EC uses a hybrid solution that most basic algorithms are unable to obtain. This approach is less sensitive to noise data. Parallelization and scalability are the other advantages. Clustering is applied as parallel on data subset and the results are combined with each other. Moreover, it integrates the solutions obtained from multiple distributed data

Paper type: Research paper
DOI: https://doi.org/10.52547/mjee.15.1.19
How to cite this paper: Z. Sahebkaram, A. R. Norouzi, **"Presenting a Proper Ensemble Clustering (EC) Method Based on Hierarchical Methods and Classical Generative Algorithms"**, *Majlesi Journal of Electrical Engineering*, Vol. 15, No. 1, pp. 19-24, 2021.

sources or features. In some EC approaches, scalability exists for many data [3]

EC is done in two steps: the first step is to perform initial clustering using basic or classical algorithms, and the second step is to reach an optimal consensus and create the final clustering. In other words, this method will reach a single consensus between several clustering solutions.

2. EC METHODS AND REVIEWING THE LITERATURE

Please use automatic hyphenation and check your EC approaches are divided into two categories Consensus Function(CF) methods and generative mechanism. Various solutions have been provided for each of them. Using various clustering algorithms, using an algorithm with initialization and different parameters, dividing data in various subsets, selecting different subset of features, and selecting different subcategories of data are examples of a generative mechanism. These approaches are used for generating the fundamental models [4] [2]

In the second step of running EC, it is necessary to apply a CF on the primary clustering to produce the final cluster. As it shown in Fig. 1, co-association, voting approach, mutual information, Hypergraph Partitioning, and Finite Mixture Model are among the methods that can be used as a consensus function for generating the final cluster.

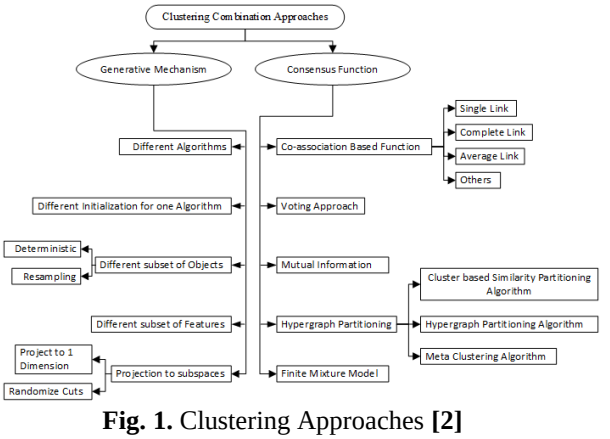

Fig. 1. Clustering Approaches **[2]**

The voting approach consists of two steps. First, a clustering algorithm is iterated several times, then a voting procedure gets these results and obtains the number of clusters in the dataset. Wendong et al. proposed a model for dividing social networking users. The model was based on the combination of heterogeneous data and EC. In this model, the behavior of individuals in social networks and the tags that users have defined based on their interest, along the user-defined texts have led to the initial division of social networking users. Then these partitions were combined according to the voting mechanism and ultimately the users were divided by clustering. The name of their proposed model was SLT. Their experiments were performed on real data and had better performance compared to k-means. In this study, a kind of soft clustering is used, as the individuals may be members of several groups and have different topics of interest [5] .

In soft clustering, if set X has n members as X = {x1, x2, ..., xn}, placed in k cluster, in clustering results by voting method, the data points are not assigned to a single cluster. Thus, the result will be a fuzzy partition. After the voting, obtained by N times iteration and running, for every x as the data point and for each j cluster, a D_{Nj} value determines the number of times that the data point has been recorded in cluster j [6]

$$k = \arg\max_j (D_{Nj}) \quad (1)$$

Tumer and Agogino used a method to combine multiple primary clustering into an EC that acts based on active cluster voting method. They called the method used VAC. This method is robust against the missing data and does not need to collect data in a specific position. When the quality of the clusters obtained was measured by the relevant measures, the maximum values are obtained compared with other traditional clusters ending in better performance. Moreover, VAC method has a high error tolerance and provides acceptable performance even if 50% of the votes are defective [7] .

In the theory of probabilities and information, mutual information for two random variables is the degree of interdependence between the two variables. Specifically, one can say that it specifies the quantity or amount of information obtained for a variable. The concept of mutual information is complicatedly connected to the entropy of a random variable. The consensus function for an EC can be calculated as mutual information for the candidate cluster and compound cluster. Assuming that partitions are independent, mutual information can be written as the sum of the pairs of interactions between the specified partition and the target partition[8] .

In graph-based methods, primary clustering data is shown as graph without direction. Then EC is obtained through graph division. The basic idea of EC is based on this graph to consider a graph as G = (V, E), where V is the number of vertices and E is the number of edges. The cut C = (S, T) generates a partition where S and T sets have no shared parts with each other. By making splits and graph partitions on the graph, an EC is obtained by the integration of primary clustering data. The problem of EC is that the probability of finding a minimum cut from a hypergraph is so low[9] .

Strehl and Ghosh proposed three consensus functions in this regard, including CSPA, HGPA, and MCLA, which have computational complexity of O (kN2H), O (kNH) and O (k2NH2), respectively [10].

Huang et al. proposed an approach for EC that uses a Factor Graph. Their proposed method is called ECFG.

Compared to existing approaches, this approach has several advantages: the number of clusters is automatically obtained and there is no need for considering original value for that. This is one of the advantages of EC. Another advantage is that the reliability of each basic cluster can be estimated in an unsupervised manner. This approach can be effective for data of a large size and large aggregate size. The results of testing this method on real datasets showed that this method is more effective and accurate than similar approaches. In this study, they have probably formulated EC as a binary linear programming problem and proposed the factor graph technique to solve this optimization problem[11] .

Zheng et al. proposed a framework for hierarchical EC that combines hierarchical clustering results with partition clustering. K-means algorithm is iterated several times to create partitions in partition clustering. Integrating the results from the implementation of the k-means algorithm is recorded in a matrix. Based on the dataset, there is a distance matrix. Data or datasets are clustered hierarchically based on this distance. These two results are combined in a matrix called ultra-matrix, and the result will be a hierarchical clustering. The results indicate the effectiveness of the proposed method and approach[12] .

Most EC methods are based on the premise that primary clustering and object are equally important, whereas in some approaches, weights related to clusters are also considered. Ren et al. proposed a method called WOEC approach, where the results of primary clustering are summarized in a correlation matrix. This related information can be used as the weight of the objects. Techniques for the proposed consensus that can be used for weighted objects and converts EC problem into a graph partition. This method has better stability and robustness compared to the weighted version of the k-means algorithm. This approach has been tested on 15 real dataset [13].

3. MATERIAL AND METHODS

It is assumed that our dataset has n members. preprocessing methods are used as a first step. The output of each algorithm is considered as the input of the consensus function. By using consensus function, the best clustering is obtained.

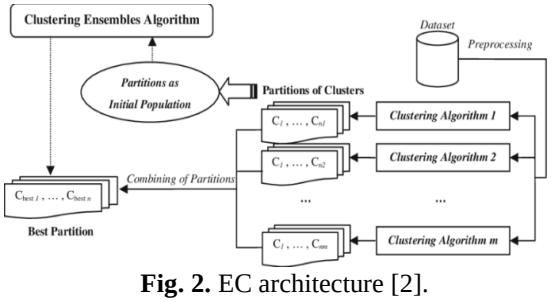

Fig. 2. EC architecture [2].

Then data points are partitioned by m clustering algorithms. The number of final clusters is not known in advance. This number depends on the scale of the data being checked in a dataset. It is assumed to be a set with six members.

$$V = \{v_1, v_2, v_3, v_4, v_5, v_6\}$$

These data have been clustered by m algorithm (m = 3) and data are partitioned into clusters C1, C2, and C3. The purpose is to create a final cluster that is compatible with early clustering as much as possible.

If $\{v_1, v_3, v_5\}$ of the three times of the implementation of these input algorithms is placed in a cluster three times each, then one can conclude that in the final clustering, the probability that these three data points will be placed in a cluster together is very high and may not be cost effective to separate the final clustering. Thus, the final clustering resulting from the combination should have the least degree of discrepancy with the three primary clustering. For two objects v and u, it is possible to calculate the difference between clustering based on a function whose output is zero and one [14]:

$$d_{u.v}(C_1.C_2) = \begin{cases} 1 & if\ C_1(u) = C_1(v)\ and\ C_2(u) \neq C_2(v). \\ & or\ C_1(u) \neq C_1(v)\ and\ C_2(u) = C_2(v). \\ 0 & else \end{cases} \quad (2)$$

The primary clustering algorithm can use any partitioning, hierarchical, density-based, and grid-based approaches [15]. In the partitioning method, if there is an n-member dataset, it is possible to divide the data into k partition, so that each cluster has at least one member and one element cannot belong to two clusters. When hierarchical methods are used, clustering sometimes initiates with the most intra-clustering similarity, then, they separate into distinct clusters based on the differences between the data, and sometimes each data-point initially becomes a member of a cluster, and then data is merged based on similarities to create larger clusters. Density-based algorithms can create non-spherical clusters. A Grid-Base method quantizes D dataset to a finite set of cells that have a Grid structure.

Single-link, full-link, and average-link algorithms used as primary and classical clustering have a hierarchical structure. The difference between these three algorithms is in calculating the distance. In single-link algorithm, the distance between each element of the cluster is calculated twice and the lowest value represents the distance between the two clusters. In complete-link, the distance between the clusters is calculated according to the farthest distance between the elements in it and in the average-link, the average distance between the two cluster elements is calculated [15]

$$SL: dist_{\min}(C_i, C_j) = \min_{p \in C_i, q \in C_j} \{| p - q |\}$$
$$CL: dist_{\max}(C_i, C_j) = \max_{p \in C_i, q \in C_j} \{| p - q |\} \quad (3)$$
$$AL: dist_{avg}(C_i, C_j) = \frac{1}{n_i n_j} \sum_{p \in C_i, q \in C_j} | p - q |$$

In methods based on the correlation matrix, basic cluster information is shown as similarity matrix. Then these results are aggregated through the mean matrix. EC acts based on similarity, so the final clustering is generated by agreement on a similarity matrix[4]

For instance, if there is a 7-member dataset clustered with four primary clustering:

$\lambda^{(1)} = (1,1,1,2,2,3,3)$ $\lambda^{(2)} = (2,2,2,3,3,1,1)$
$\lambda^{(3)} = (1,1,2,2,3,3,3)$ $\lambda^{(4)} = (1,2,3,1,2,4,5)$

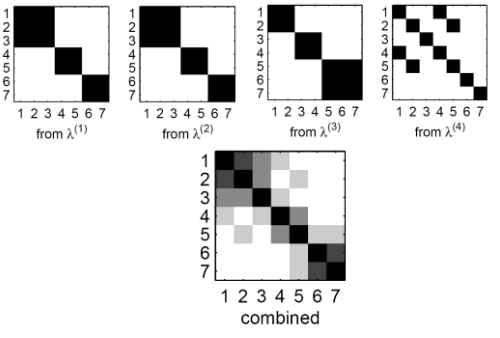

Fig. 3. Creating a correlation matrix[4]

Now, a correlation matrix is built, so that if n data-point is available, an n × n matrix is created, each entry of which is computed from C (i, j) with the following formula where N is the number of partitions and n_{ij} shows the number of times when i-th element is also clustered with j-th element[16] .

In this paper, single-link, full-link and average-link clustering algorithms were considered as primary clustering. The results of these three algorithms are recorded in a correlation matrix and EC runs based on that. For example, if data 1 is located in cluster 2 in all three algorithms and data 2 is clustered in the cluster 2 in all three algorithms, these two data will cluster together in EC. This comparison is calculated individually between the elements to obtain data in the same cluster. The number of clusters is unknown from the beginning. For each base cluster, the number of clusters is considered four.

3.1. Dataset and Evaluation Criteria

In this paper, the data of a film dataset has been evaluated. The datasets used in this project are related to video data taken from IMDB site. This data has 1659 tuples and 9 features, including film production year, film title, director's name, genre and length of the film, the name of the famous actors and actresses participated in the film, was it awarded a prize or not, and ultimately its popularity. The inclusion of minor features, such as movie playback time or its title has been ignored in clustering to achieve better results. In the clustering used in this paper, 71 data were randomly used to give a clearer view of the data.

Silhouette and CH and Dunn indices were used to evaluate the clustering. All three measures measure intra-cluster distance between clusters. As the intra-cluster distance is less and inter-cluster distance is more, it means a better clustering done. In other words, the larger these three numbers are, the better the result of clustering will be [17]

$$C(i, j) = \frac{n_{ij}}{N} \quad (4)$$

The formula for computing CH index is as follows:

$$CH = \frac{B(k)}{W(k)} \times \frac{N - k}{k - 1} \quad (5)$$

B (k) is the inter-clusters distance and W (k) is the distance between the members of a cluster. The larger the B (k) is and the smaller W (k) is, more favorable clustering is obtained.

Silhouette index is calculated based on the following formula:

$$Sil(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \quad (6)$$

b (i) determines the distance of the object from other clusters and a (i) shows the clustering density that i belongs to. This index is calculated for each individual data, and an average is obtained for it and the value recorded in the table is the same as the average of the Silhouette index[18] .

The formula below represents the Dunn index. $dist(c_i, c_j)$ is the distance between clusters ci and cj and k is the number of clusters. $diam(c_l)$ is the intra-cluster distance of c_l cluster. Increase in the fraction and decrease in the denominator will result in better clustering [19] .

$$D = \min_{1 \le i \le k} \left\{ \min_{\substack{i+1 \le j \le k \\ j \ne i}} \left\{ \frac{dist(c_i, c_j)}{\max_{1 \le l \le c} \{(diam(c_l)\}} \right\} \right\} \quad (7)$$

4. RESULTS AND DISCUSSION

The results of the implementation of the basic algorithms are as follows:

For each of the three single-link, full-link, and average-link algorithms, the data is clustered into four clusters. The number of clusters is specified beforehand. Data located at one level on the tree structure is in one

cluster. For example, data-points 1 and 2 are located in cluster 1.

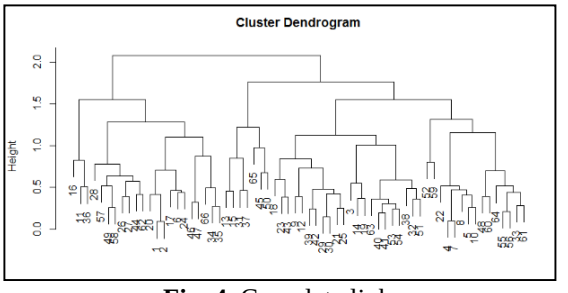

Fig. 4. Complete link.

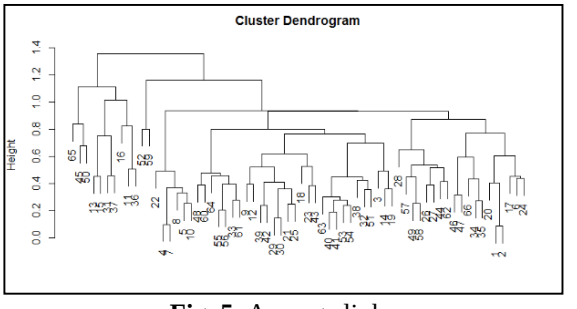

Fig. 5. Average link.

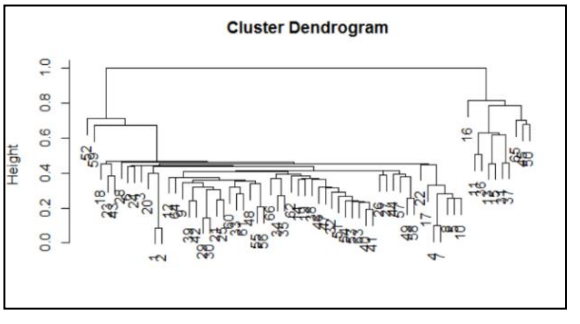

Fig. 6. single link.

Fig. 7 shows how many numbers are assigned to clusters 1, 2, 3, and 4 in each clustering algorithm. The number of data belonging to the cluster 1 is greater than the other clusters.

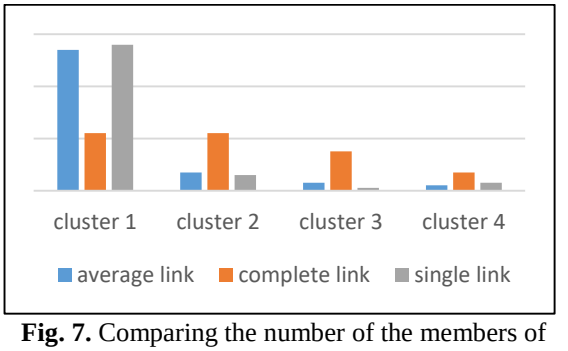

Fig. 7. Comparing the number of the members of primary clustering.

Table 1. Comparing the number of members of each cluster in the classical cluster.

Algorithm	Cluster1	Cluster2	Cluster3	Cluster4
Average link	54	7	3	2
Complete link	22	22	15	7
Single link	56	6	1	3

Each matrix column was examined after creating the correlation matrix. If the column's index is named with i and the row's index with j, for some data in the corresponding column i with that data-point, the rows with the maximum value show that (i, j) will also cluster in the final cluster. The result of EC was as follows:

Table 2. Number of members of each cluster in EC.

Cls1	Cls2	Cls3	Cls4	Cls5	Cls6	Cls7
19	22	13	3	4	3	2

Fig. 8 also shows the number of cluster members in the hybrid model. The number of final clusters from the combination is 7.

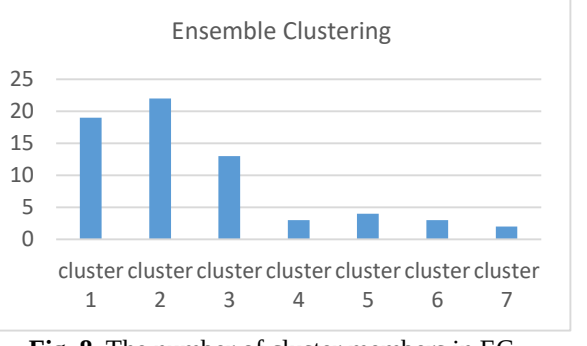

Fig. 8. The number of cluster members in EC

The results of cluster evaluation are shown based on the measurements given in the table.

CH index has the largest value possible when all three primary clustering models, one-link, full-link and average-link, are combined. This value is 79.44, which is a larger value both relative to single clustering and when clustering is combined into two. When the clustering was evaluated using Silhouette, the value obtained was greater than the total-link and average-link clustering when the three algorithms were combined. This value was 0.313 yet less than the value of this index for the single-link algorithm. The Dunn index is the highest when the two single-link and average-link clusters are combined equal to the average-link value. Given the values obtained, one can conclude that EC can improve clustering results drastically.

Table 3. Comparison of cluster evaluation indices.

Algorithm	CH Index	Silhouette Index	Dunn Index
Complete Linkage	53.12	0.304	0.212
Average Linkage	53.75	0.249	0.417
Single Linkage	43.56	0.373	0.415
ENSEMBLE CLUSTERING			
Complete & Average & Single	79.44	0.313	0.323
Complete & Average	67.077	0.297	0.264
Complete & Single	60.599	0.330	0.3009
Single & Average	39.800	0.247	0.417

5. CONCLUSION

This paper was conducted on EC on film data. At first, the data was clustered using single-link, average-link, and full-link hierarchical algorithms. In the next step, the obtained results were combined in a binary and triplicate using a correlation matrix. Finally, the clustering performance was evaluated using three clustering criteria: Silhouette, CH and Dunn. Although the criteria did not produce the same results, where all three clustering algorithms were combined, CH had the highest value. Thus, EC can improve the clustering quality drastically.

REFERENCES

[1] A. Norouzi et al., **"Medical image segmentation methods, algorithms, and applications,"** *IETE Technical Review*, Vol. 31, No. 3, pp. 199-213, 2014.

[2] S. N. Ghaemi Reza, Ibrahim Hamidah, Norwati Mustapha, **"A Survey: Clustering Ensembles Techniques,"** 2009.

[3] A. Topchy, A. K. Jain, and W. Punch, **"Clustering ensembles: models of consensus and weak partitions,"** *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 27, No. 12, pp. 1866-1881, 2005.

[4] J. G. Alexander Strehl, **"Cluster Ensembles {A Knowledge Reuse Framework for Combining Multiple Partitions,"** *Journal of Machine Learning Research 3 (2002)*, pp. 583-617, 2002.

[5] Y. Wendong, L. Hong, P. Na, and L. Zhenzhen, **"Social media user partitioning based on ensemble clustering,"** *in Service Systems and Service Management (ICSSSM), 2016 13th International Conference on*, 2016, pp. 1-6: IEEE.

[6] E. Dimitriadou, A. Weingessel, and K. Hornik, **"Voting-Merging: An Ensemble Method for Clustering,"** *in Artificial Neural Networks — ICANN 2001: International Conference Vienna, Austria, August 21–25, 2001 Proceedings, G. Dorffner, H. Bischof, and K. Hornik, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2001*, pp. 217-224.

[7] K. Tumer and A. K. Agogino, **"Ensemble clustering with voting active clusters,"** *Pattern Recognition Letters*, Vol. 29, No. 14, pp. 1947-1953, 2008.

[8] N. Nguyen and R. Caruana, **"Consensus Clusterings,"** *in Seventh IEEE International Conference on Data Mining (ICDM 2007)*, pp. 607-612, 2007.

[9] S. VEGA-PONS and J. RUIZ-SHULCLOPER, **"A Survey of Clustering Ensemble Algorithms,"** *International Journal of Pattern Recognition and Artificial Intelligence*, Vol. 25, No. 03, pp. 337-372, 2011.

[10] A. Strehl and J. Ghosh, **"Cluster ensembles --- a knowledge reuse framework for combining multiple partitions,"** *J. Mach. Learn. Res.*, Vol. 3, pp. 583-617, 2003.

[11] D. Huang, J. Lai, and C.-D. Wang, **"Ensemble clustering using factor graph,"** *Pattern Recognition*, Vol. 50, pp. 131-142, 2016.

[12] L. Zheng, T. Li, and C. Ding, **"A framework for hierarchical ensemble clustering,"** *ACM Transactions on Knowledge Discovery from Data (TKDD)*, Vol. 9, No. 2, p. 9, 2014.

[13] Y. Ren, C. Domeniconi, G. Zhang, and G. Yu, **"Weighted-object ensemble clustering: methods and analysis,"** *Knowledge and Information Systems*, Vol. 51, No. 2, pp. 661-689, 2017.

[14] H. M. Aristides Gionis, and Panayiotis Tsaparas, **"Clustering Aggregation,"** *ACM Transactions on Knowledge Discovery from Data (TKDD*) 2007.

[15] K. M. Han Jiawei, Pei Jian, **Data Mining Concepts and Techniques**, *3 ed*. 2012.

[16] A. L. N. Fred and A. K. Jain, **"Combining multiple clusterings using evidence accumulation,"** *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 27, No. 6, pp. 835-850, 2005.

[17] Y. Liu, Z. Li, H. Xiong, X. Gao, and J. Wu, "**Understanding of Internal Clustering Validation Measures,"** *in 2010 IEEE International Conference on Data Mining*, pp. 911-916, 2010.

[18] F. Kovács, C. Legány, and A. Babos, **"Cluster validity measurement techniques,"** *in 6th **International** symposium of hungarian researchers on computational intelligence*, 2005: Citeseer.

[19] Z. Ansari, M. Azeem, W. Ahmed, and A. V. Babu, "**Quantitative evaluation of performance and validity indices for clustering the web navigational sessions,"** *arXiv preprint arXiv:1507.03340*, 2015.