International Journal of Electrical and Computer Engineering (IJECE)
Vol. 8, No. 4, August 2018, pp. 2351~2357
ISSN: 2088-8708, DOI: 10.11591/ijece.v8i4.pp2351-2357

Extensive Analysis on Generation and Consensus Mechanisms of Clustering Ensemble: A Survey

Y. Leela Sandhya Rani[1], V. Sucharita[2], K. V. V. Satyanarayana[3]
[1]Department of CSE, Scholar, Koneru Lakshmaiah Education Foundation (KLU), India
[2]Department of CSE, NEC, Gudur, India
[3]Department of CSE, Professor, Koneru Lakshmaiah Education Foundation (KLU), India

Article Info

Article history:

Received Oct 24, 2017
Revised Jan 7, 2018
Accepted Jan 15, 2018

Keyword:

Clustering
Clustering ensemble
Consensus method
Generation method
Unsupervised classification

ABSTRACT

Data analysis plays a prominent role in interpreting various phenomena. Data mining is the process to hypothesize useful knowledge from the extensive data. Based upon the classical statistical prototypes the data can be exploited beyond the storage and management of the data. Cluster analysis a primary investigation with little or no prior knowledge, consists of research and development across a wide variety of communities. Cluster ensembles are melange of individual solutions obtained from different clusterings to produce final quality clustering which is required in wider applications. The method arises in the perspective of increasing robustness, scalability and accuracy. This paper gives a brief overview of the generation methods and consensus functions included in cluster ensemble. The survey is to analyze the various techniques and cluster ensemble methods.

Corresponding Author:

Y. Leela Sandhya Rani,
Departement of Computer Science and Engineering,
Sir C R Reddy College of Engineering, Eluru, W G Dt, Andhra Prradesh, India.
Email: lsraniyalamarthi18@gmail.com

1. INTRODUCTION

Data mining refers to dredging of knowledge from extensive data. Data mining deals with large amount of data sets. Clustering is the major functionality in data mining. Clustering is a way of organizing the data into similar groups that have same features and it does not occur in other groups. Because of this property of clustering for large data sets, scalable algorithms are required. But for most applications of clustering they are not scaled well due to small data sets. Each algorithm has its pros and cons, for different algorithms with same data set or/and different data sets with same algorithm produces distinct solutions. It is tough to know which algorithm is suitable for given data set.

Clustering plays a major role in data mining, machine learning, bioinformatics image processing, information retrieval, market segmentation, big data analytics and many more areas. One of important tasks in cancer classification call class discovery by micro array is previously done using single clustering algorithms. Using clustering techniques we can identify the co-location patterns that are usually arise in spatial data bases using some data mining algorithms [1]. Grouping of unstructured data based on its content is done by document clustering which is one of the most popular machine learning techniques and it further analyse the data to understand patterns in it. Separating of pixels into clusters is done by clustering based image segment approach in image processing. Now days it is very difficult to search in internet as there are many documents available in the internet. Searching can be done effectively using some keywords by clustering algorithms. Text based clustering plays a major role in browsing and navigation process [2]. Clustering play a major role in social networks also. It is used to analyze psychology of humans and their relationships [3]. Clustering [4].

Some of the data sets contains any type of data such as numeric or categorical or both and differs from their attributes. Conventional clustering algorithms cannot perform well when the data sets are mixed type. In order to group the dissimilar data types there is a clustering ensemble approach in which a combined solution is obtained from a group of individual solutions to produce a quality clustering. Clustering ensemble improves the strength and stability of clustering solution due to its consolidating and dividing nature. Cluster Ensemble is an approach that consolidates various findings of dissimilar clusterings to bring out the final quality clustering of original data set.

Clustering ensembles are more advantageous than a single clustering algorithm in so many strands

Robustness: The clustering ensemble improves the average performance on different streams and datasets more than a single clustering.

Novelty: The combined solution gives unusual results which cannot be produced by one clustering algorithm.

Stability: Clustering ensemble works efficiently and can handle noise and outliers.

Parallelization and Scalability: Parallelization of clustering can be acquired by successive synthesis of results. It has the ability to amalgamate results from multiple heterogeneous sources of data.

Clustering ensembles are used in many areas. Such as bioinformatics, machine learning and information retrival. The ensembles are formulated with different types of optimization algorithms such as genetic algorithms, evolutionary algorithms, k-particle swarm optimization algorithm, k-muscles wandering optimization algorithms in different aspects in different areas are specified in the following sections clearly in accordance with some journal papers.

The difficulty with clustering ensemble is to perceive a consensus function. There are different consensuses functions are available but in order to increase the stability and robustness genetic algorithm with co-association matrix is used as a consensus function. The genetic algorithm composed of four phases includes fitness function, selection method, crossover method and finally mutation method. The co-association matrix values are used to obtain the intra and extra cluster fitness by evaluating average similarity between all clusters in first phase. In the second phase tournament selection is used in which two individuals are adopted arbitrarily and the individual with preferable fitness is elected for next population. In third phase the two off springs are generated by the individuals are exchanged with a random crossover point. Intelligence mutation is used in fourth phase [5].

The utilization of locally adaptive clustering algorithm provides an implementation to identify a partition that finds solutions to the clusters. It imparts set of clusters with some weights then assign a specified probability to each cluster. Using Jaccard coefficient find inter cluster similarity based on feature and object. This two-objective clustering ensemble complicate in setting parameter and in interpretation of results. So single objective clustering is composed both feature based and object based as a whole which increases accuracy [6].

For stream mining clustering ensemble is imparted. This integrates both clusters and classifiers together and employ genetic algorithm and has high propensity to handle optimization [7].

The clustering ensemble is designed as an optimization problem on multiple objects by adopting evolutionary algorithm on multiple objects. The first criteria in multi objective clustering ensemble are to maximize the similarity measure of final clustering from all input clusterings. The similarity measure is calculated using adjusted random index. The second criterion in multi objective clustering ensemble is to reduce the similarity measure [8].

The clustering ensemble is designed using three different algorithms k-means, k-particle swarm optimization and k-muscles wandering optimization. The combination of k-means with muscles wandering optimization overcomes the shortcomings of k-means algorithm. It implements similarity based clustering algorithm using weights on input data. Samples the dataset first and then apply clustering algorithms specified on subsamples which give clustering results. From that similarity matrices are generated. Based on various metrics of clustering best clustering can be derived. Reduce the weights of the samples. Repeat the process until best resultant clustering found [9].

The clustering ensemble is introduced based on particle swarm clustering. The particle swarm clustering is act as a base clusterer and as well as consensus function is a challenging element. The consensus function allows the base partitions with different number of clusters and permits both disjoint and overlapping partitions. Proposed ensemble produce statistically better partitions [10].

The next part of the paper is organized as follows. Section 2 gives concept of clustering ensemble, Section 3 explains taxonomy of generation methods, Section 4 specifies taxonomy of consensus methods and Section 5 presents conclusion and future work on clustering ensemble.

2. CLUSTERING ENSEMBLE

Clustering ensemble is an approach that concatenates the subset clustering solutions in to quality clustering for original data set. From Figure 1 the data set can be divided into N number of samples and applied clustering algorithm on each sample set, clustering ensemble generates N clustering solutions and finally combined the solutions to get final quality clustering using a consensus function. The clustering ensemble has been processed in two steps. One is generation step which generates the number of clustering solutions. Second one is consensus step which combines the solution into a final clustering.

2.1. Generation step

The way of combining all the individual clustering solutions of subsets generated from original data set is called ensembling. The first step is generation of all individual clustering solutions. The clustering ensemble is the combination of clustering results. Given a data sets of m objects $P = \{P_1, P_2, P_3 \ldots P_m\}$, the clustering ensemble generates n number of clusterings represented as $\beta = \{\beta_1, \beta_2, \beta_3 \ldots \beta_n\}$[1]. Each clustering solution β_i is one part of the original data set P into K_i dissimilar groups of objects, denoted as $\beta_i = a_{i1}, \ldots a_{ik}$.

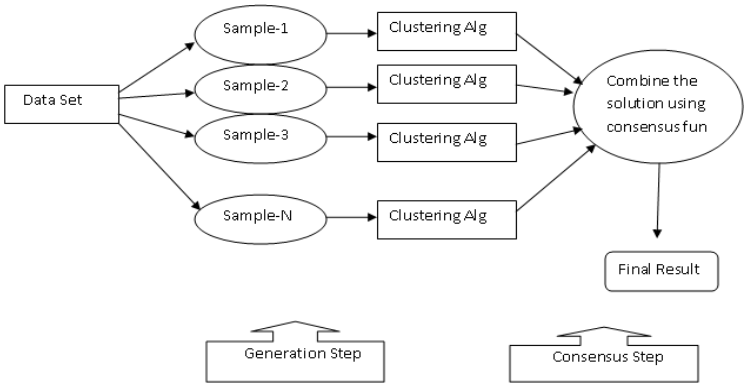

Figure 1. General mechanism of clustering ensemble

2.2. Consensus step

The consensus step is used to combine the solutions of clusterings and is the important step in any algorithm of clustering ensemble. It is the function which improves the results of single clustering algorithm. There are two ways to apply consensus function one is correlation between objects and optimization for partition. The first one is to analyze the number of one instance belonging to one cluster and number of two instances belonging to the same cluster. It is done through voting approach and Co-Association Matrix based methods.

In the second approach of consensus function, the feature partition is acquired in association with optimization problem [11]. The partition can be find by using some similarity between the features is the main problem with respect to the cluster ensemble. Formally, the feature partition is defined as:

$$X^* = arg \max_{x \epsilon X_i} \sum_{j=1}^{m} \Gamma(X, X_j)$$

Here ɼ is a similarity measure between partitions. The feature partition is the maximization problem which is given as the subgroup that increases the similarity with all subgroups in the cluster ensemble. The following are the examples use feature partition are kernel based methods and non-negative matrix factorization.

3. TAXONAMY OF CLUSTERING ENSEMBLE GENERATION METHODS

The clustering ensemble generates the set of clustering solutions by applying some clustering algorithm on set of samples and combines the clustering solutions to get final quality clustering. The main concept is to handle different types of features. This can be solved by randomly selecting the features on basis of cluster analysis. The clustering ensemble produces accurate results as it finds one final clustering

from set of clusterings based on different samples with different algorithms for generation process. We provide some of the generation methods which are reviewed from previous papers.

3.1. Similar ensembling

The ensemble generation process uses similar algorithms for all sample subsets. The k-means algorithm fails to perform class discovery effectively on data sets because it assumes that the data is in Gaussian distribution. As spectral clustering handles this problem, Spectral clustering (SC) algorithm is used for the ensemble generation in Knowledge based Cluster Ensemble (KCE).

The knowledge based cluster ensemble randomly generates d^x dimensions from d dimensions this process continues until A subspaces generated $\{D_1,D_2,\ldots D_a\}$. A spectral clustering algorithm is applied to each subspace and generates A clustering solutions. Spectral clustering partition the data points into K classes. First SC constructs an affinity matrix, and obtains a normalized matrix X, and then applies k-means for each row of X to get these points into K clusters. In this way SC is applied repeatedly to all samples to get solution for each sample subspace.

Next a confidence factor was calculated for all the clustering solutions by constructing adjacency matrix on pair wise constraints. If the most of the pair wise constraints are satisfied by the clustering solution it will have high confidence measure otherwise less confidence measure [12].

3.2. Random initialization of input parameters

The "Projective Clustering Ensemble" (PCE) is based on a set of heterogeneous gene-to-cluster assignments and sample-to-cluster assignments. Input to the PCE is taken from gene expression data G. Each entry of G represents a gene expression level of a particular gene. If we group samples into clusters use sample-to-cluster assignment. The probability of a sample that belongs to a cluster is nothing but sample-to-cluster assignment. If we group genes into clusters use gene-to-cluster assignment. The probability of gene belonging to a cluster is nothing but gene-to-cluster assignment.

These assignments are produced by applying continuous projective clustering N times with different random initializations for input parameters to produce N clustering solutions, which are used as main clustering for consensus clustering [13].

3.3. Feature selection for sampling

Set of sample subset can be generated based on random sampling techniques to generate set of clustering solutions. Now a days large dimensional data sets are used for data analysis. So feature selection plays an important role in generation of sample subsets. In "Double Selection Semi Supervised Clustering Ensemble" (DSSSCE) they used feature selection methods to remove noise and outliers.

The DSSSCE use input from gene expression data. It first applies a set of feature selection methods such as Mutual Information Maximization (MIM), Mutual Information Feature Selection (MIFS), Joint Mutual Information (JMI), Conditional Infomax Feature Extraction (CIFE), Conditional Redundancy (CONDRED), Interaction Capping (ICAP), Double Input Symmetrical Relevance (DISR), Max-Relevance Min-Redundancy (MRMR) to select set of sub samples.

Later DSSSCE applies PC-Kmeans to identify the labels of the cancer dataset. This algorithm considers the number of must-link and cannot-link constraints between pairs of cancer samples which leads to clustering solution. Using feature selection methods as a selection strategy it selects set of clustering solutions and aggregate all the solutions by building matrix in the first phase. Next, DSSSCE divides the aggregated solution into set of clustering solutions and calculate the confidence factors for the clustering solutions based on prior knowledge of the data set which is specified by pair wise constraints [14].

3.4. Incremental ensembling

In "Incremental Semi Supervised Clustering Ensemble" (ISSCE) first one original ensemble is generated. Then the final new ensemble is produced with the help of set of selection members. It generates two ensembles using random subspace generation method as a subspace generator, Constraint Propagation approach as a clustering algorithm.

The Double Selection Semi Supervised Clustering Ensemble feature selection methods are used as a subset selection and clustering applied in two phases. The ISSCE also used two ensembles in the design. To handle high dimensional data space use random subspace methodology to generate set of subspaces. Apply constraint propagation methodology on set of subspaces to produce set of clustering solutions. The ISSCE incorporated incremental member selection process based on local and global cost function and produced new ensemble with same algorithm used in first ensemble [15].

3.5. Dissimilar ensembling

The generation step in dissimilar ensembling involves different clustering algorithms and finally uses a base clustering. The number of clusters is generated by using different clustering algorithm with its different parameters that is randomization of the sample data. Different clustering algorithms may give different clustering results due to properties of data.

If we apply different number of clustering algorithms on data set then set of different clustering solutions may occur. Among them a compromised clustering solution should be identified. All the clusters can be identified by using any one of the candidate clustering algorithm. So it must follow that any data point must assigned to only one cluster and every data point in the set must assigned to any cluster. It is necessary to interpret all the partitions whether they follow above mentioned criteria or not. Use goodness function to evaluate the quality of the cluster. Select certain clustering solutions using goodness function [16].

4. TAXONAMY OF CLUSTERING ENSEMBLE CONSENSUS METHODS

There are various types of consensus functions they are Hyper graph Partitioning, Co-association based functions, Mutual Information Algorithm, Finite Mixture model and Voting Approach. We provide some of the consensus functions which are reviewed from previous papers.

4.1. Spectral graph partitioning

Spectral Clustering chooses a spectral graph partitioning algorithm, which used to optimize the cut scale. First KCE constructs a matrix $\emptyset$ by considering all the generated matrices of the clustering solutions and respective confidence factors simply concatenation of all matrices. Finally based on spectral clustering algorithm it partitions the new features into K classes. For the majority of the cancer datasets KCE outperforms the other clustering ensembles.

KCE constructs a matrix by specifying all the membership matrices of the clustering solutions and the respective confidence factors as follows:

$$\emptyset = \prod(\mathbf{T}^{\boldsymbol{A}}\mathbf{\eth}^{\mathbf{A}})$$

Where is T^A is the representation of all membership matrices of the clustering solutions, and $\eth^A$ is the set of confidence vectors of clustering solutions. $\prod$ is used to concatenate these two. Using spectral clustering partition the new features of concatenated result in to K classes [12].

4.2. Optimization algorithm

It is necessary for a clustering ensemble to find a consensus function that minimizes the distance from all clusters so the following function is optimized.

$$¥^{*} = \mathbf{argmin}_{\boldsymbol{I}} \left\{\boldsymbol{\psi_s}\,(\mathcal{I}, \mathcal{E})\boldsymbol{\psi_g}(\mathcal{I}, \mathcal{E})\right\}$$

Here ψ is used as a distance function for the clusterings. PCE optimize the $\mathcal{E}$ for two requirements gene-to-cluster and sample-to-cluster assignment. So Expectation Maximization of Projective Clustering Ensemble (EM-PCE) is used as a consensus function. The main aim of EM-PCE is to minimize the error that corresponds to both sample to cluster and gene to cluster assignment [13].

4.3. Graph partitioning

In "double selection semi supervised clustering ensemble" they designed consensus function by combining all the membership matrices of the clustering solutions and corresponding confidence factors to one matrix A. Based on the sample set Y a graph is constructed on Y and A. Using the normalized cut approach on the constructed graph, the final clustering of the original data set is obtained [14], [15].

4.4. Hill climbing

Based upon the goodness function the number of clustering solutions can be obtained. To generate clustering solutions there are two conflicts, one is absence conflict and other is coverage conflict. So the consideration of conflicts becomes NP hard. Based on hill climbing approach the optimization problem can be solved and finally gets one clustering for the given data set [16].

5. PROPOSED OBJECTIVE

Clustering ensemble is a framework that combines the solutions from individual clusterings to produce a qualified clustering. Our objective is to preprocess the data by using hybrid fuzzy logic feature selection method which is our next future work. For the resultant samples we apply different clustering algorithms and finally we get qualified clustering. From Figure 2, c-1 c-2 c-3 c-4 specifies different clustering algorithms.

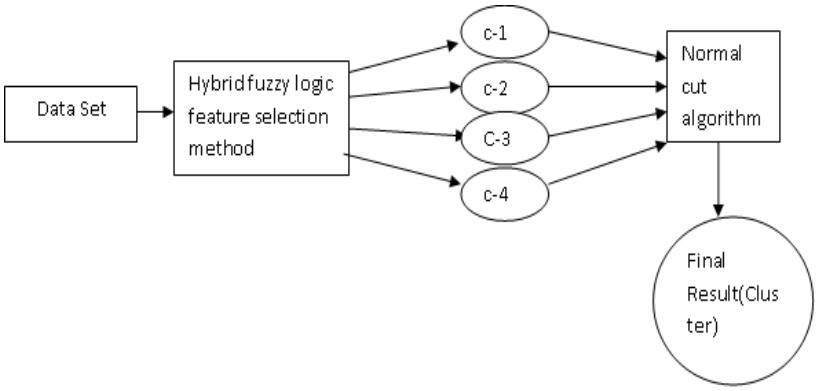

Figure 2. Extended clustering ensemble

6. CONCLUSION AND FUTURE WORK

Clustering ensemble is a framework that provides set of clustering solutions and merges solutions to get a qualified clustering output for the given data set. Definitely it produces more accurate results as with the single clustering and also improves the robustness, scalability and quality of the clustering. In this paper we reviewed some papers which use different generation methods and different consensus functions to get final clustering. The main aspect is generation mechanism. In some papers they used similar algorithms and in some they used dissimilar algorithms for generation process. For sampling of subspace some used feature selection and some used Random Sampling. Current trends handle large dimensional data sets so we use feature selection methods for reducing the dimensionality and increasing the performance. Later we apply different clustering algorithms for each subset generated from the application of feature selection methods. This new generation step of our new ensemble increases the performance of the final clustering solution as we applying hybrid fuzzy logic feature selection method and different clustering algorithms. If we remove the noise and redundant data from the data set it will increases the performance of data analysis. It is done with hybrid fuzzy logic feature selection method. If we apply different clustering algorithms different clustering solutions will be generated from them which are having the highest similarity those will be considered as best clustering solutions and also uncovered clusters from different solutions are amalgamated to get final clustering solution. This is the future scope of our work.

REFERENCES

[1] Naveen Kumar, S. Siva Sathya, "Clustering Assisted Co-location Pattern Mining for Spatial Data", *International Journal of Applied Engineering Research*, ISSN 0973-4562, vol. 11, no. 2, pp. 1386-1393, 2016.

[2] M. John Basha, K.P. Kaliyamurthie, "An Improved Similarity Matching based Clustering Framework for Short and Sentence Level Text", *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 7, no. 1, pp. 551-558, February 2017.

[3] Charu Virmani, Anuradha Pillai, Dimple Juneja, "Clustering in Aggregated User Profiles across Multiple Social Networks", *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 7, no. 6, pp. 3692-3699 December 2017.

[4] K.R. Nirmal, K.V.V. Satyanarayana, "Issues of K Means Clustering While Migrating to Map Reduce Paradigm with Big Data: A Survey", *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 6, no. 6, , pp. 3047-3051, December 2016.

[5] Javad Azimi, Mehdi Mohammadi, Ali movaghar, Morteza Analoui, "Clustering Ensembles Using Genetic Algorithm", *The International Workshop on Computer Architecture for Machine Perception and Sensing*, September 2006.

[6] Purushothaman B, "Clustering Ensembles Using Evolutionary Algorithm", *International Journal for Research in Applied Science & Engineering Technology*, ISSN: 2321-9653, vol. 3, no. 2, February 2015.

[7] Anutosh Pratap Singh, Jitendra Agrawal, Varsha Sharma, "An Efficient Approach to Enhance Classifier and Cluster Ensembles Using Genetic algorithms for Mining Drifting Data Streams Multi objecctive clustering", *International Journal of Computer Applications* ISSN 0975 – 8887, vol. 44, no. 21, April 2012.
[8] Sujoy Chatterjee, Anirban Mukhopadhyay, "Clustering Ensemble: A Multiobjective Genetic Algorithm based Approach", *International Conference on Computational Intelligence: Modeling, Techniques and Application*, 2013.
[9] Qi Kang, ShiYao Liu, Meng Chu Zhou, SiSi Li, "A weight-incorporated similarity-based clustering ensemble method based on swarm intelligence", *knowledge based systems, Elsevier*.
[10] Jos´e Valente de Oliveira, "Particle Swarm Clustering in clustering ensembles: exploiting pruning and alignment free consensus", *Applied Soft Computing*, Feb 13, 2016.
[11] Sandro Vega-Pons and Jose Ruiz-Shulcloper, "A Survey of Clustering Ensemble Algorithms", 2011.
[12] Zhiwen Yu, Hau-San Wongb, Jane You, Qinmin Yang and Hongying Liao, "Knowledge Based Cluster Ensemble for Cancer Discovery from Biomolecular Data", *IEEE Transactions on Nanobioscience* vol. 10, no 2, June 2011.
[13] Xlanxue Yu Guoxian Yu and JunWang, "Clustering cancer gene expression data by projective clustering ensemble", Research article, *Plos One*, 2017.
[14] Zhiwen Yu, Hongsheng Chen Jane Yu, Hau SanWong , Jiming Liu Fellow Le Li Guoqiang Han, "Double Selection Based Semi Supervised Clustering Ensemble for Tumor Clustering from Gene Expression Profiles", *IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2013*.
[15] Zhiwen Yu, Jane You, Hau SanWong, Jun Zhang, "Incremental semi supervised Clustering Ensemble for High Dimensional data Clustering", article, *IEEE Transactions on Knowledge and Data Engineering*, 2015.
[16] Martin H.C. Law, Alexander P. Topchy, Anil K. Jain "Multiobjective Data Clustering", *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 204.

BIOGRAPHIES OF AUTHORS

Y. L. Sandhya Rani is a Scholar in K L University and working as an assistant professor in Sir C R R College of Engineering, Eluru. She has interested in the areas like Clustering in data mining, Data Structures etc. Currently she was working with Clustering Ensemble Performance as a research work. She has published many papers in national and international journals.

Dr. V. Sucharita is a professor in Narayana Engineering College, Gudur. She has published many papers in national and international journals. Also she is reviewer and editorial board member for reputed journals

Dr. K. V. V. Satyanarayana is professor in K L university. He has published many papers in national and international journals. He has interested in the area like Bio informatics and cloud computing.