World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:15, No:6, 2021

A Review and Comparative Analysis on Cluster Ensemble Methods

S. Sarumathi, P. Ranjetha, C. Saraswathy, M. Vaishnavi, S. Geetha

Open Science Index, Computer and Information Engineering Vol:15, No:6, 2021 publications.waset.org/10012100.pdf

Abstract**—**Clustering is an unsupervised learning technique for aggregating data objects into meaningful classes so that intra cluster similarity is maximized and inter cluster similarity is minimized in data mining. However, no single clustering algorithm proves to be the most effective in producing the best result. As a result, a new challenging technique known as the cluster ensemble approach has blossomed in order to determine the solution to this problem. For the cluster analysis issue, this new technique is a successful approach. The cluster ensemble's main goal is to combine similar clustering solutions in a way that achieves the precision while also improving the quality of individual data clustering. Because of the massive and rapid creation of new approaches in the field of data mining, the ongoing interest in inventing novel algorithms necessitates a thorough examination of current techniques and future innovation. This paper presents a comparative analysis of various cluster ensemble approaches, including their methodologies, formal working process, and standard accuracy and error rates. As a result, the society of clustering practitioners will benefit from this exploratory and clear research, which will aid in determining the most appropriate solution to the problem at hand.

Keywords**—**Clustering, cluster ensemble methods, consensus function, data mining, unsupervised learning.

I. Introduction

CLUSTERING is one of the vital and widely used techniques in Data Mining. It plays a crucial role in the other fields such as Spatial Data Extraction, World Wide Web, Machine Learning Process, Pattern Recognition, Image Processing and Information Retrieval. Data clustering mainly deals with the process of grouping a collection of objects based on their proximity in vector space. The destination of the cluster analysis is to find similarities among data objects according to the uniqueness found in the data and to group associated data objects collectively as clusters. A great number of clustering algorithms have been proposed from earlier stages [1], [2]. On the divergent, there is no single clustering method that is able to give accurate and suitable cluster outcomes. Similarity or dissimilarity distances between the instances in the dataset are determined using an efficient clustering algorithm. However, if two similar clustering algorithms are applied to the same data set, diverse cluster solutions are produced which is used to estimate the accurate clustering outcomes. This estimation is related to the use of cluster validity indexes that are used to determine the quality of clustering outcomes. On the other hand, to overcome this severe concern, combining multiple clustering approaches in an ensemble framework may allow one to take advantage of the strengths of individual clustering approaches. The general sketch of the cluster ensemble is done by attaining the solutions from the diverse base clustering, which are then aggregated to form a final partition. This meta level approach involves the following two major tasks, namely generation of a cluster ensemble and then creating a final partition usually referred to as the consensus function. The challenges in cluster ensembles are the definition of the most suitable consensus function that is capable of improving the performance of single clustering algorithm [1].

II. General Idea on Cluster Ensemble Techniques

Cluster ensemble is a process for getting consensus solutions that can be formed by grouping up with various clustering results. The consensus solution is depending upon combining several partitions which contain well-defined rules. Therefore, the cluster ensembles are considered to be more robust. The cluster membership, the number and boundaries are determined by using the visualization tool. For creating most ideal clusters it has an ensemble clustering as a major approach and it may be possible by the individual clustering approach. Generation step and Consensus step are the two major tasks in cluster ensembles [2]. The general structure of the cluster ensemble was shown in Fig. 1.

A. Generation Step

In generation step no confines are available for the partition that should be acquired [1]. For generating numerous base cluster solutions, different clustering algorithms or the same algorithm with different parameter initialization, different object representations, and subsets of objects or projections of the objects on different subspaces can be used to create the different base cluster solutions [2]. In spite of this development even a weak clustering algorithm is proficient of producing high quality consensus clustering in concurrence with the proper consensus function [2]. In the cluster ensemble method, the generation phase is depicted in Fig. 2.

Dr. S.Sarumathi, Professor, Ms. P.Ranjetha, Assistant Professor, Ms. M. Vaishnavi, Assistant Professor, and Ms.S. Geetha, Assistant Professor, are with the Department of Information Technology, K. S. Rangasamy College of Technology, Tamil Nadu, India (e-mail: rishi_saru20@rediffmail.com, ranjetha0405@gmail.com, vaishnavi.munusamy@gmail.com, geethas@ksrct.ac.in).

Ms. C.Saraswathy, Associate professor, is with the Department of Electronics and Communication Engineering, K. S.Rangasamy College of Technology, Tamil Nadu, India (e-mail: csaraswathy66@gmail.com)

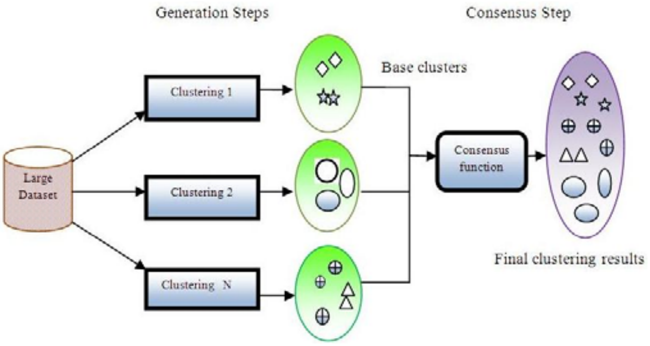

Fig. 1 Structure of Cluster Ensemble

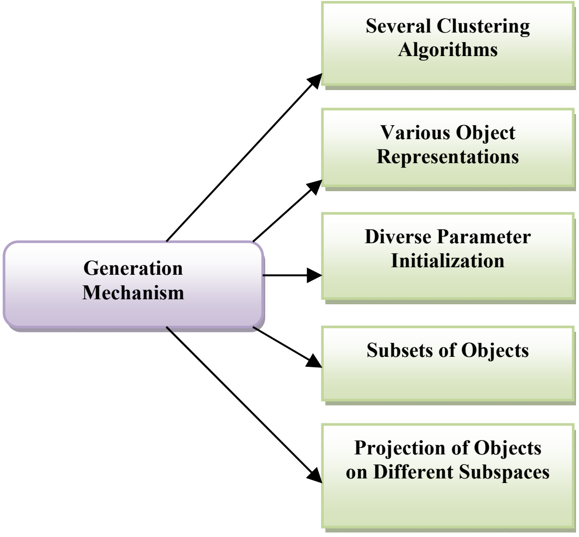
Generation Mechanism
Several Clustering Algorithms
Various Object Representations
Diverse Parameter Initialization
Subsets of Objects
Projection of Objects on Different Subspaces

Fig. 2 Generation Steps in Cluster Ensembles

B. Consensus Step

Different consensus functions are formed in the consensus process, which is also useful for obtaining the final data partition from several base clustering results [2]. The result of the single clustering algorithm has been enhanced by means of the consensus function. This comprises two systems such as median partition and also objects co-occurrences. It holds the partition in the cluster ensembles that exploit their similarity with every other partition. The improper analysis of the difference measures provides the complexity of this median partition in the first method. In the second method, it contracts with an individual cluster for determining the number of incidences of an object and in similar cluster [2].

III. Different Cluster Ensemble Methods

The next segments will present the few varied collections of cluster ensemble processes. Along with the features, methodology of each process is explained here.

A. Framework for Active Clustering with Ensembles (FACE)

In [3], a semisupervised framework for clustering face patterns into individual groups using minimal human interaction is proposed. This method merges concepts from ensemble clustering and active learning to get better clustering accuracy. The system asks the user for a soft connection limitation between each pair of neighboring faces that are uncertainly balanced by the ensemble. With the most comprehensive examination of active face clustering algorithms to date, the efficacy of our technique is proved. The tests look for data that can be used for human-in-the-loop face recognition, such as fuzzy point-and-shoot videos, photos of women before and after applying makeup, and twin photographs. The findings show that ensemble-based constrained clustering algorithms are more noise-resistant than other approaches [3]. The first stage is to detect or track the faces to determine when and where they emerge. This method results in a collection of cropped face images or sequences $F = \{f_1, f_2, \cdots, f_{n_f}\}$. The main aim is to achieve an arbitrary self-labeling for the faces $L : F \to \mathbb{Z}$ that points out, which face observations correspond into the similar person.

1) Clustering Faces

The FACE technique joins the idea of clustering and active learning. In ensemble clustering the multiple partitions are produced by discrete clustering algorithms by several algorithmic parameterizations or else from diverse views of the data [4], [5]. A high-class clustering of the data can result from a consensus vote on which pairs of samples belong to the similar cluster. In addition, an ensemble is used to get well randomly shaped clusters. During supervised active learning, query-by-committee is a known method for creating queries for user labeling. The partitioning ensemble extends the semi-supervised clustering. The FACE algorithm iteratively clusters faces to identity-specific clusters by a diverse ensemble of grouping computed with a discrete algorithm and parameterizations.

Soft Hierarchical Agglomerative Clustering with Constraints (SHACC) algorithm is the expansion of the clustering method at the core of the Active HACC algorithm (AHACC) [6]. Both are employed with the constrained distance between its patterns. Linear Constrained Vector Quantization Error (LCVQE) expands the classic k-means method to update the cluster process. The constrained distance is given by

$$d_c(x_i, x_j) = d(x_i, x_j)^{\sigma_{ij}} \tag{1}$$

where $d(x_i, x_j)$ is the min-max normalized Euclidean distance among the prototypes.

LCVQE is based on k-means by means of shifting cluster centers into contain violated constraints. The original k-means algorithm begins by selecting k cluster centers $\{\mu_a \, for \, a = 1 \cdots k\}$. The prototype x_i is allocated to a cluster which minimizes all the objective function, the distortion

$$\sum_{a=1}^{k} \sum_{x_i \in C_a} \left\| x_i - \mu_a \right\|^2 \tag{2}$$

calculates the distances between patterns and cluster centers every iteration.

The centers are rationalized after the prototypes are allocated to clusters

$$\mu_a = \frac{1}{|C_a|} \sum_{x_i \in C_a} x_i \tag{3}$$

Officially the LCVQE objective functions $\sum_{a=1}^{k} J_a$ where

$$J_a = \frac{1}{2} \sum_{x_i \in C_a} \left\| x_i - \mu_a \right\|^2 + \frac{1}{2} \sum_{m_{ij} \in V_M, L_t(x_i) = a} \left\| x_i - \mu_a \right\|^2 + \frac{1}{2} \sum_{m_{ij} \in V_M, L_t(x_j) = a} \left\| x_i - \mu_a \right\|^2 + \frac{1}{2} \sum_{c_{ij} \in V_c, N(c_{ij}) = a} \left\| F_{L_t(x_i)}(c_{ij}) - \mu_a \right\|^2 \tag{4}$$

B. Hierarchical Cluster Ensemble Selection (HCES)

In data mining, clustering ensemble approach is widely adopted in the cluster's research to improve the quality and robustness of clustering results. The choice of a subset of obtainable ensemble members depends on diversity plus quality frequently directs to a great accurate ensemble solution. Cluster-based Similarity Partition Algorithm (CSPA) plus Hypergraph-Partitioning Algorithm (HGPA) are working on HCES technique for getting the all ensembles plus cluster ensemble selection solution [7].

1) Consensus Function

Consensus function is an algorithm for collecting dissimilar clustering to get last clusters. We presume that H has L ensemble members where $H = \{h_1, h_2, \cdots, h_L\}$ the consensus function ф unites each ensemble member of H as $h^* = ф(h_1, h_2, \cdots, h_L)$. In the cluster ensemble selection, the consensus function has an effect on a subset of ensemble members. The cluster ensemble selection of the consensus function is defined as $h_s^* = ф(H_s)$ such as $H_s \subset H$. This contains various approaches that are separated into voting, pairwise, feature-based and graph-based approaches.

Voting approach is known as direct approach or else relabeling approach. In the feature-based approach, the output of each clustering algorithm is considered as a categorical feature [8]. The pairwise approach creates the co-association matrix in which the similarity between points is the number of times that points are in the same clusters of clustering results. Usually, hierarchical algorithms such as single-link, average-link, and complete-link are used for combining results by co-association matrix.

The graph-based approach contains instance-based, hybrid and cluster-based approaches. In instance-based approach, the objects are measured as vertices and a similarity measure between the objects in clusters are calculated as the weight of the edges. The CSPA as an instance-based approach builds a hypergraph in where the amount of frequency of two vertices that are accumulated in the similar clusters is regarded as weight of every edge [7].

2) Diversity and Quality Measures

The Normalized Mutual Information (NMI) and Adjusted Rand Index (ARI) are commonly employed to measure the diversity or quality of separation.

$$NMI(h_a, h_b) = \frac{-2\sum_{i=1}^{k_a}\sum_{j=1}^{k_b}\left(n_{ij}\log\frac{n.n_{ij}}{n_{ia}\cdot n_{bj}}\right)}{\sum_{i=1}^{k_a}\left(n_{ia}\log\frac{n_{ia}}{n}\right)+\sum_{j=1}^{k_b}\left(n_{bj}\log\frac{n_{bj}}{n}\right)} \tag{5}$$

where $h_a = \{C_1^a, C_2^a, \cdots, C_{k_a}^a\}$ and $h_b = \{C_1^b, C_2^b, \cdots, C_{k_b}^b\}$ with k_a and k_b clusters are the two clusters on dataset D with n sample.

$$ARI(h_a, h_b) = \frac{\sum_{i=1}^{k_a}\sum_{j=1}^{k_b}\frac{n_{ij}}{2} - t_3}{1/2(t_1 + t_2) - t_3} \tag{6}$$

where $t_1 = \sum_{i=1}^{k_a}\binom{n_{ia}}{2}, t_2 = \sum_{j=1}^{k_b}\binom{n_{bj}}{2}, and\ t_3 = \frac{2t_1t_2}{n(n-1)}$.

Diversity measures can be classified into external and internal diversities. With known class labels, the external diversity measure is defined based on a quality measure such as NMI or ARI, as follows:

$$diversity(\bar{h}, h_i) = 1 - quality(\bar{h}, h_i) \tag{7}$$

where $\bar{h}$ is the given class label and h_i, such as $i = 1,2,\cdots,L$ are clustering. The average of diversity is

$$D_e = \frac{1}{L}\sum_{i=1}^{L} diversity(\bar{h}, h_i) \tag{8}$$

Internal diversity is classified into pair-wise and non-pair-wise diversities. In pair-wise diversity every cluster is selected as a class label implicitly plus another clustering is measured using the selected class label. The diversity is evaluated as

$$diversity(h_i, h_i) = 1 - quality(h_i, h_i) \tag{9}$$

where $i \neq j = 1,2,\cdots,L$. The average of diversity measure is

$$D_p = \frac{1}{L(L-1)}\sum_{i=1}^{L}\sum_{j=1,i\neq j}^{L} diversity(h_i, h_i) \tag{10}$$

The non-pair-wise diversity measure is defined as:

$$diversity(h^*, h_i) = 1 - quality(h^*, h_i) \tag{11}$$

where $i = 1,2,\cdots,L$ plus h^* is the outcome got by a consensus function. The average of diversity measure is

$$D_{np} = \frac{1}{L}\sum_{i=1}^{L} diversity\left(h^{*}, h_i\right) \qquad (12)$$

3) Cluster Ensemble Extraction Approach

a) Generate dissimilar clustering
b) Find consensus clustering solution h^* using consensus function
c) Calculate pair-wise diversity measure matrix in which every element of the matrix is a diversity measure among two clusters
d) Partition all clustering results as a dendrogram implicitly, using a hierarchical clustering algorithm on the diversity measure matrix
e) Select the one solution from every cluster with highest quality by means of NMI quality measure for identifying a new subset of clustering
f) Find an ensemble solution using a consensus function on the new subset
g) Select the preeminent ensemble solution between ensembles outcomes depend on their quality.

C) K-Means-Based Consensus Clustering (KCC): A Unified View

KCC offers an essential and enough state for utility functions. Based on this fact, the KCC utility function can be easily derived from a continuously differentiable convex function, which facilitates to create a unified framework for KCC, and makes it an efficient solution. Next, the computations of the utility functions and distance functions are adjusted so as to enlarge the appropriate scope of KCC to the cases where data incompleteness occurs [9]. Finally, the major factors which may affect the performances of KCC are empirically explored, and acquire some useful guidance from specially designed experiments on various datasets.

1) Consensus Clustering

In common, the already existing consensus clustering methods can be divided into two classes, i.e., the techniques with or without global objective functions. Here, the former technique is considered that are classically formulated as a combinatorial optimization problem. Given r basic separations of $\chi = \{x_1, x_2, \cdots, x_n\}$ in $\prod = \{\pi_1, \pi_2, \cdots, \pi_r\}$ the aim is to identify a consensus partitioning π such that

$$\Gamma(\pi, \Pi) = \sum_{i=1}^{r} w_i \cup\left(\pi, \pi_i\right) \qquad (13)$$

is maximum. Such as $\Gamma = \mathbb{Z}_{++}^{n} \times \mathbb{Z}_{++}^{nr} \longmapsto \mathbb{R}$ is a consensus function where $\cup\colon \mathbb{Z}_{++}^{n} \times \mathbb{Z}_{++}^{nr} \longmapsto \mathbb{R}$ is a utility function plus $w_i \epsilon [0,1]$ is a user specified weight for π_i by means of $\sum_{i=1}^{r} w_i = 1$.

2) K-means Clustering

K-means is a prototype-based partitioning method to identify user-specified K crisp clusters. Such clusters are standing for their centroid [10]. K-means is out looked as a heuristic to optimize the objective function as

$$\min \sum_{k=1}^{k} \sum_{x\in C_k} f\left(x, m_k\right) \qquad (14)$$

Such as m_k is the centroid of the kth cluster C_k, f is the distance as of a data point to a centroid. The clustering procedure of k-means is a two-phase iterative heuristic by means of the data assignment and centroid update staggering successively. The Bregman divergence is well-known since a family of distances fits the classic k-means. That is $\emptyset : \mathbb{R}^d \times \mathbb{R}^d \longmapsto \mathbb{R}$ is defined as for k-means clustering.

$$f(x, y) = \Phi(x) - \Phi(y) - \left(x - y\right)^{T} \nabla\Phi(y) \qquad (15)$$

The sternness of the convexity of ф is unrestricted if the unique minimizer assumption diminishes towards the non-unique case. This shows the way to the more common "point-to-centroid distance" derived as of convex other than essential strictly convex ф.

D) Revisiting Link-Based Cluster Ensembles (LCE) for Microarray Data Classification

Novel techniques that make use of cluster ensembles which summarize information matrix transform data for the classification. The LCE approach offers a highly accurate clustering [11]-[13]. Two steps in LCE are creating an ensemble Π and aggregating base clusterings $\pi_g \in \Pi$, g = 1 ... M where M is a meta-level data matrix.

1) Creating Cluster Ensembles

a) *Fixed-k*: Every clustering $\pi_g \in \Pi$ is produced by the data set $X \in \mathbb{R}^{N\times D}$ with each D attributes. The number of clusters in all base clustering is fixed to $k = \lceil\sqrt{N}\rceil$.
b) *Random-k*: Every π_g is build by the data set with each attribute plus the number of clusters is arbitrarily selected among $\{2, \cdots, \lceil\sqrt{N}\rceil\}$ [14]-[16].

2) Aggregating Base Clustering Results

In the cluster ensemble Π, in which base clustering results are aggregated into an information matrix $\Theta \in [0,1]^{\mathrm{N\times P}}$, the last partition π^* is produced. $\Theta(x_i, cl)$ denotes the association degree that the sample $x_i \epsilon X$ with every cluster $cl \in \{C_1^g, \cdots, C_{kg}^g\}$ as

$$\Theta\left(x_i, cl\right) = \begin{cases} 1 & cl = C_*^g\left(x_i\right) \\ sim\left(cl, C_*^g\left(x_i\right)\right) & otherwise \end{cases} \qquad (16)$$

where $C_*^g(x_i)$ is a label for cluster to samplex_i. $sim(C_x, C_y) \in [0,1]$ Denotes the similarity among two clusters $C_x, C_y \in \pi_g$ which discovered by the link-based algorithm

3) Weighted Connected Triple (WCT) Algorithm

WCT expands the Connected-Triple technique that is developed to recognize ambiguous names inside publication databases [17]-[20]. The initial method is created on a social network describe as n undirected graph $G = (V, E)$.

$$WCT^{z}_{xy} = \min\left(\left|w_{xz}\right|, \left|w_{yz}\right|\right) \tag{17}$$

here $|w_{xz}|$ and $|w_{yz}|$ is weight connecting the vertices v_x and v_z, and vertices v_y and v_z.

$$WCT_{xy} = \sum_{z=1}^{\lambda} WCT^{z}_{xy} \tag{18}$$

$$s_{WCT}\left(v_x, v_y\right) = \frac{WCT_{xy}}{WCT_{\max}} \times DC \tag{19}$$

where $S_{WCT}\left(v_x, v_y\right)$ is the similarity between the vertices and DC is is the constant decay factor.

E) Temporal Data Clustering via Weighted Clustering Ensemble with Different Representations (WCE)

Temporal data clustering offers fundamental methods for finding out the intrinsic formation plus compacting information over temporal data. A temporal data clustering framework offers a weighted clustering ensemble of different separations created by initial clustering analysis on similar temporal data representations [21]. This approach mainly proposed a novel weighted consensus functions directed by clustering validation criteria towards settling initial separations to candidate consensus separation from different viewpoints.

1) Weighted Clustering Ensemble

The fundamental design of weighted consensus function makes use of the pairwise comparison which measures the clustering quality with similar clustering validation criteria [14]. A dendrogram is built depending on every similarity matrix to produce candidate consensus separations.

a) Partition Weighting Scheme

A partitioning weighting system allocates a weight w_m^π to every P_m in expressions of a clustering validation criterion along with the weights of each separation depending on the criterion together from a weight vector $W^\pi = \{w_m^\pi\}_{m=1}^{M}$ [22], [23]. A weight in the weighting scheme is defined as

$$w^{\pi}_{m} = \frac{\pi\left(p_m\right)}{\sum_{m=1}^{M} \pi\left(p_m\right)} \tag{20}$$

b) Weighted Similarity Matrix

In $H_m = \{0,1\}^{N \times K^m}$ a row represents one data with a column represents a binary encoding vector for one specific cluster in the partition P_m. K_m represents the number of clusters in P_m.

$$s_m = H_m H_m^{T} \tag{21}$$

A weighted similarity matrix S^π about every separation in P is built with a linear mixture of their similarity matrix S_m by means of their weight w_m^π since

$$S^{\pi} = \sum_{m=1}^{M} w^{\pi}_{m} s_m \tag{22}$$

c) Candidate Consensus Partition Generation

The quantity of clusters in a candidate consensus partition P^π is strong-minded mechanically by cutting the dendrogram plagiaristic from S^π to shape clusters. Three candidate consensus partitions P are created, P^π where $\pi = \{MHT, DVI, NMI\}$ using the DSPA method.

2) Agreement Function

Concatenating H^π matrices guides to an adjacency matrix containing every data in a known data set against candidate consensus partitions $H = [H^{MHT}|H^{DVI}|H^{NMI}]$, then the pairwise similarity matrix $\bar{S}$ is obtained by

$$\bar{s} = \frac{1}{3} H H^{T} \tag{23}$$

The DSPA technique is used to generate a dendrogram from S, as well as the final partition P [24], [25].

F) Visual Analytics for Comparison of Ocean Model Output with Reference Data: Detecting and Analyzing Geophysical Processes Using Clustering Ensembles (VAA)

A novel visual analytics approach is proposed, that expands the scope of the analysis, reduces subjectivity, and assists comparison of the two data sets. It comprises of three steps: In the first step, it permits modelers to consider various aspects of the temporal activities of geophysical processes by performing multiple clusterings of the temporal profiles in each data set. Modelers can opt for diverse features, express the temporal behavior of relevant processes, clustering algorithms, and parameterizations. The results of the clusterings are combined into a single clustering in the second stage, using a clustering ensemble methodology. The aggregated clustering presents an outline of the geospatial distribution of temporal behavior in a data set. Third, a graphic interface allows modelers to evaluate the two consolidated clusterings. It facilitates them to detect clusters of temporal profiles that represent geophysical processes and to investigate differences and similarities between two data sets [26], [27].

1) Advantages

a) They are no longer confined to single statistical measures for the detection of geophysical processes, but instead have access to a variety of temporal behavior aspects.

b) For the detection of geophysical processes, we are no longer limited to single statistical measurements, but instead have access to a variety of temporal behavior features.

c) The interactive tool allows modelers to get a more complete image of model and reference data differences and similarities.

d) It offers a lot of potential for speeding up the model creation process because it allows for rapid initial evaluation of new model versions.

2) Visual Analytics Concept

The three major objectives for a visual analytics approach are to facilitate comparison of model data with reference data.

a) Every set of clustering is merged into one consolidated clustering.
b) Modelers interactively investigate and contrast the two consolidated clustering.
c) A visual analytics approach permits modelers to develop many spatial clustering of the temporal profiles in the model result and reference data, then to combine the several clustering for every dataset by means of an ensemble approach and to interactively investigate contrast and similarity between the two datasets.

a. Analytical Requirements (AR)

The analytical requirements include two requirements for the calculation of many clustering as AR1 and AR2 plus two requirements by means of clustering ensembles AR3 and AR4 [28]-[30].

a) AR1 Wide variety of features
b) AR2 Multiple discrete cluster parameterizations
c) AR3 Supple configuration of the consolidation
d) AR4 Quantitative measures to support the appraisal of consolidated clustering

b. Visualization Requirements

a) VR1 Overview of consolidated clustering
b) VR2 Examination of cluster properties
c) VR3 Detailed contrast of clusters.

G) Sc-GPE: A Graph Partitioning-Based Cluster Ensemble Method for Single-Cell

A novel cluster ensemble method integrating five single-cell graph partitioning-based clustering algorithms are proposed namely, SNN-cliq, PhenoGraph, SC3, SSNN-Louvain, and MPGS-Louvain in cluster ensemble processes [31]. Sc-GPE, a consensus matrix, is developed based on the five clustering solutions by calculating the chance that cell pairs are classified into the same cluster. The problem was solved using a hypergraph-based ensemble technique, which took into account the various cluster labels assigned in the individual clustering methods, and it was challenging to locate the relevant cluster labels across all approaches. Then, to differentiate the different significance of each method in a clustering ensemble, a weighted consensus matrix was constructed by designing an importance score strategy. Ultimately, hierarchical clustering was performed on the weighted consensus matrix to cluster cells. To assess the performance, Sc-GPE is compared with the individual clustering methods and the state-of-the-art SAME-clustering on 12 single-cell RNA-seq datasets. The result shows that Sc-GPE obtained the best average performance, and achieved the highest NMI and ARI value in five datasets.

H) Clustering Ensemble of Massive High Dimensional Data Based on BLB and Stratified Sampling Framework

A novel clustering ensemble algorithm based on BLB and stratified sampling framework for massive high-dimensional data is proposed [32]. From two aspects of sample and feature, BLB (Bag of Little Bootstrap) algorithm is used to divide the original data set into several small-scale data subsets, then use the feature stratified sampling to obtain a low-dimensional subset. Then, using link-based consensus functions, basic clustering results are generated on several small-scale low-dimensional subsets, and finally, cluster integration results are achieved. The investigations result in synthetic data sets and UCI real data sets show that the algorithm proposed in this paper is effective for clustering massive high-dimensional data.

I) Clustering Ensemble Based on Hybrid Multiview Clustering

The ensemble approaches in the clustering algorithm incorporate different clustering solutions into a final one, thus improving the clustering efficiency [33]. The key to propose the clustering ensemble algorithm is to progress the diversities of base learners and optimize the ensemble strategies. To address these issues, a clustering ensemble framework that consists of three parts is proposed. In the first approach, three view transformation schemes, namely random principal component analysis, random nearest neighbor, and modified fuzzy extension model, are used as base learners to learn different clustering views. A random transformation and hybrid multiview learning-based clustering ensemble method (RTHMC) is then considered to synthesize the multiview clustering results. In the second method, a new random subspace transformation is incorporated into RTHMC to increase its performance. In the last step, a view-based self-evolutionary strategy is developed to further improve the proposed method of optimizing random subspace sets. Experimentation and comparisons reveal the efficiency and superiority of the proposed method for clustering different kinds of data.

J) Ensemble-Based Clustering of Large Probabilistic Graphs Using Neighborhood and Distance Metric Learning

Graphs are normally used to articulate the link between various data. Here, to handle indecisive data, the probabilistic graph method is proposed [34]. As a primary problem of such graphs, clustering is used in many applications to analyze uncertain data. To tackle the challenges with individual clustering, a unique method called ensemble clustering is utilized for huge probabilistic graphs. To create ensemble clusters, a set of probable, possible worlds of the initial probabilistic graph is developed. Then, a probabilistic co-association matrix as a consensus function is presented to integrate base clustering results. It depends on co-occurrences of node pairs, based on the probability of the corresponding common cluster graphs. Also, two steps, before and after of ensembles generation, are applied. In the before step, neighborhood information is appended based on node features to the initial graph to attain a more accurate estimation of the probability between the nodes. In the after step, supervised metric learning-based Mahalanobis distance is used to automatically learn a metric from ensemble clusters. It aims to increase essential features of the base clustering results. The work is assessed using five real-world datasets and three evaluation metrics, namely the Dunn index, Davies–Bouldin index, and Silhouette coefficient. The output shows the notable performance of clustering large probabilistic graphs.

K) High-Performance LCE Approach for Categorical Data Clustering

In recent days, the clustering ensembles come forward as a problem solver for mine the data objects into clusters in a well-organized way. However, still clustering creates a serious matter due to the occurrence of imperfect information while partitioning the data objects into clusters. This causes a severe issue in creating a proficient cluster with cluster ensembles. In this paper, a solution to solve the degradation in clusters during data partitioning is proposed [35]-[37]. The initial base clusters are generated using firefly algorithm. A LCE method uses similarity measurement using multi-viewpoint and weighted triple quality using entropy measurements that ensembles the data objects into clusters. These methods avoid the problem of local optimum and also avoid the issues happened from high-dimensional datasets and improve the quality of clustering. Here, the data partitioning is completed with bipartite spectral algorithm and similarity measurement. Finally, to generate classified results from the optimized clustered datasets the artificial neural network is used. The research was carried out using data from the UCI repository, and the results show that the suggested technique performs effective ensemble clustering with higher clustering accuracy than the predictable methods [38]-[41].

L) Consensus Function Based on Cluster-Wise Two-Level Clustering

Ensemble clustering attempts to combine several fundamental clusterings in order to produce a more consistent, robust, and high-performing consensus clustering result. A novel ensemble clustering algorithm is proposed in [42] to enhance the quality of the final clustering results. The suggested method, referred to as a consensus function supported two level clustering (CFTLC), introduces a replacement consensus clustering task during which a mean hierarchical clustering is applied to a cluster–cluster similarity matrix created using an explicit similarity metric. A set of meta clusters was created using the average hierarchical clustering technique. It assigns each data point to a meta cluster based on object-cluster similarity, with each meta cluster being treated as a consensus cluster in the result. CFTLC first converts the primary partitions to a binary cluster representation, in which the primary ensemble is divided into a number of basic binary clusters (BC). The basic BCs with the highest cluster–cluster similarity are combined first via CFTLC. This phase is repeated until a predetermined number of meta clusters have been created. It then assigns each data point to exactly one meta cluster in the next step. In terms of accuracy and resilience, the suggested method has been tested against state-of-the-art clustering methods [42].

M) An Ensemble of Locally Reliable Cluster Solutions

Clustering ensemble refers to a method that involves performing a number of (usually weak) base clusterings and then using the consensus clustering as the final clustering. Knowing that popular decisions are preferable than dictatorial decisions, it appears straightforward and straightforward that ensemble (here, clustering ensemble) decisions are preferable to simple models (here, clustering) decisions. However, not every ensemble is guaranteed to outperform a simple model. If the members of an ensemble are valid or high-quality, and they participate in consensus clustering according to their attributes, the ensemble is regarded to be a better ensemble. For creating base clusters, this research employs a clustering ensemble framework that employs a simple clustering algorithm based on the kmedoids clustering method. The validity of the discovered clusters is guaranteed by our simple clustering methodology. It is also ensured that the clustering ensemble framework employs a method that prioritizes the use of each found cluster based on its quality. To implement this approach, different k-means clustering methods are used to produce an auxiliary ensemble known as the reference set. The proposed ensemble clustering method is compared to many current ensembles clustering algorithms and three powerful fundamental clustering algorithms on a collection of simulated and real-world benchmark datasets in the empirical study. According on the empirical data, the suggested ensemble clustering algorithm outperforms state-of-the-art ensemble clustering approaches significantly [43]-[45].

N) Ensemble Clustering Based Semi-Supervised Learning for Revenue Accounting Workflow Management

Amadeus is the world's biggest IT solutions supplier for the travel and tourism industry. Amadeus develops software that enables airlines, airports, hotels, trains, search engines, travel agents, tour operators, and other stakeholders to manage travel globally. The process of managing and dispatching the amount obtained from the customer's payment for their travel is referred to as revenue accounting. This procedure entails several iterations of the data in the input, which is represented as a ticket calculation code sequence for each journey. A semi-supervised ensemble clustering approach is described here for discovering important multi-level clusters in big datasets with relation to application objectives and translating them to application classes for predicting the class of incoming instances. This framework builds on the MultiCons closed sets-based multiple consensus clustering strategy, although it can be easily extended to other ensemble clustering methods. It was created to make the Amadeus Revenue Management application more efficient. Revenue accounting in the travel sector is a difficult undertaking when trips involve many modes of transportation and associated services, all of which are provided by various operators and take place in different geographical locations with different taxes and currencies, for example. The proposed methodology for automating the Amadeus Revenue Management workflow optimizes anomaly fixes, according to statistics [46].

O) An Explainable and Statistically Validated Ensemble Clustering Model Applied to the Identification of Traumatic Brain Injury Subgroups

In the United States, traumatic brain injury (TBI) is a primary cause of mortality and disability. It may cause long-term deficits in a person's physical (movement, eyesight, hearing), emotional (depression, personality changes), and/or cognitive (memory loss) capabilities. In the United States, nearly 3 million TBI-related occurrences result in emergency department visits, hospitalizations, or deaths each year [47]. TBI is a brain condition with a wide range of causes, severity, pathology, and prognosis [48]. It can be brought on by a variety of factors, including car accidents, falls, attacks, and trauma. TBI patients are a diverse group with a wide range of pathologies, prognoses, and recovery times. It is difficult to sort through all of this diversity. A verifiable and explainable model, on the other hand, has the potential to disclose insights that can help clinicians. A methodology is proposed here for identifying phenotypic traits that separate patients into more homogeneous subgroups and describing them in terms of injury severity and recovery. The ensemble clustering model employs four distinct algorithms (k-means, spectral, Gaussian mixture, and agglomerative clustering with Ward's linkage), as well as robust consensus decision metrics to support a principled integration method [49]-[52]. The model employs two ensemble finishing strategies: Mixture Model (MM) and Graph Closure (GC). The MM methodology uses a maximum likelihood approach to obtain the final partition. Seven commonly used internal clustering validation metrics, including the Silhouette Index (SI), Dunns index, Xie-Beni index (XB), I index, S Dbw index, CH index, and Davies-Bouldin index (DB), are used to determine the optimal clustering configuration from a different perspective in order to evaluate the results of cluster analysis in a quantitative and objective manner [53].

TABLE I
COMPARISON OF CLUSTER ENSEMBLE METHODS

Clustering ensemble methods	Ensemble size	Type of consensus function used	Dimensionality	Type of dataset used	Algorithm used to build base clustering	Features
FACE	Fixed	GET-QUERY	Small	Complex	LCVQE SHACC	Accurate, robust and parsimonious
HCES	Variable	CSPA, HGPA	Small & Large	Mixed	k-Means	Scalable, accurate
KCC	Fixed	KCC	Small & Large	Mixed & Complex	k-Means	High robustness, highly efficient
LCE	Fixed	k-Means	Small &Large	Mixed	k-Means	High classification accuracy, better performance
WCE	Fixed	DSPA	Small & Large	Mixed & Complex	k-Means	Easy-to-use technique, does not suffer from a tedious parameter tuning and a high computational complexity
VAA	Fixed	CSPA, HGPA, MCLA	Large	Mixed	k-Means, hierarchical clustering, DBscan	Increasing acceptance of visual analytics in the ocean modelling community, interactive visual analysis can be of high value
Sc-GPE	Fixed	weighted consensus matrix	Large	Mixed	SNN-cliq, PhenoGraph, SSNN-Louvain, MPGS-Louvain, and SC3	Achieved the highest NMI and ARI value in datasets
BLB	Fixed	Link-based	Small & Large	Mixed	Bag of Little Bootstrap	Effective for clustering massive high-dimensional data.
HMC	Fixed	random transformation and hybrid multiview learning-based clustering ensemble method	Small & Large	Mixed	random principal component analysis, random nearest neighbor, and modified fuzzy extension model	Improves optimization
EBCLP	Fixed	probabilistic co-association matrix	Small & Large	Mixed	Probalistic graph	Provides better performance
HPLCE	Fixed	firefly	Small & Large	Mixed	bipartite spectral algorithm, artificial neural network	Offers higher clustering accuracy
CFTLC	Fixed	two level clustering	Small & Large	Mixed	average hierarchical clustering	Provides better accuracy and robustness.
LRC	Fixed	normalized spectral clustering algorithm	Large	Mixed	k.medoids	Provides improved performance
MCC	Fixed	Semi-MultiCons	Large	Mixed	Semi-supervised K-means, semi-supervised metric learning, semisupervised spectral clustering, semi-supervised ensemble clustering, collaborative clustering, declarative clustering, semi-supervised evolutionary clustering and constrained expectation-maximization	Provides better optimization
ESVCEM	Fixed	Mixture Model (MM) and Graph Closure (GC)	Large	Mixed	k-means, spectral, Gaussian mixture and agglomerative clustering	Achieved effective quality clustering results

IV. Comparison of Cluster Ensemble Methods

Table I shows a comparison of various cluster ensemble approaches based on their compatibility characteristics and application domains. The aim of this comparison is not to determine which cluster ensemble method is the best, but to demonstrate the usage model and understanding of ensemble methods in a variety of fields.

V.Conclusion

Cluster ensembles are a modern derivative that can be used to address issues created by individual clustering effects. The accuracy, individuality, robustness, and stability of unsupervised learning outcomes were all improved with this new technique. This clustering ensemble technology is useful in that it serves as a foundation for identifying and recompensing issues that may arise in solo clustering algorithms. As a result, the overall learning uncovers a variety of mixed data Cluster Ensemble methods, each with its own operating procedure and key features. As a result, the paper's innovative approach is to express each method's structured workflow, and the corresponding table reveals each technique's unique features and limitations. This expose improves the readers' perceptions of cluster ensembles strategies and is also useful for the clustering followers' society to innovate in many research activities in the future.

References

[1] S. Sarumathi, N. Shanthi, M. Sharmila, “A Review: Comparative Analysis of Different Categorical Data Clustering Ensemble Methods,” International Journal of Computer, Information Science and Engineering, Vol.7, no.12, 2013.

[2] S. Sarumathi, N. Shanthi, S.Vidhya, M.Sharmila "A Comprehensive Review on Different Mixed Data Clustering Ensemble Methods," World Academy of Science, Engineering and Technology, International journal of computer Information Science and Engineering, Vol.8, No.8, Jan 2014.

[3] Jeremiah R. Barr, Kevin W. Bowyer, and Patrick J. Flynn, "Framework for Active Clustering With Ensembles," IEEE Transactions On Information Forensics And Security, Vol. 9, No. 11, Nov. 2014.

[4] A. P. Topchy, M. H. C. Law, A. K. Jain, and A. L. Fred, “Analysis of consensus partition in cluster ensemble,” in Proc. 4th IEEE Int. Conf. Data Mining, pp. 225–232, Nov. 2004.

[5] A. L. N. Fred and A. K. Jain, “Data clustering using evidence accumulation,” in Proc. 16th Int. Conf. Pattern Recognit, vol. 4, pp. 276–280, 2002.

[6] D. Pelleg and D. Baras, “K-means with large and noisy constraint sets,” in Proc. 18th Eur. Conf. Mach. Learn, pp. 674–682, 2007.

[7] Ebrahim Akbaria, Halina Mohamed Dahlan, Roliana Ibrahim, Hosein Alizadeh, "Hierarchical cluster ensemble selection," ELSEVIER, Engineering Applications of Artificial Intelligence, 2014.

[8] Topchy, A., Jain, A.K., Punch, W," A mixture model of clustering ensembles," In: Proceedings of the International Conference on Data Mining, pp. 379–390, 2004.

[9] Junjie Wu, Hongfu Liu, Hui Xiong, Jie Cao, Jian Chen, "K-means-based Consensus Clustering: A Unified View," IEEE Transactions on Knowledge and Data Engineering, Vol.XXX, No.XXX, Dec. 2013.

[10] J. MacQueen, L. L. Cam and J. Neyman, Eds, “Some methods for classification and analysis of multivariate observations,” Statistics. University of California Press, Vol. 1, 1967.

[11] S. Chen, B. Mulgrew, and P. M. Grant, “A clustering technique for digital communications channel equalization using radial basis function networks,” World Academy of Science, Engineering and Technology Trans. Neural Networks, vol. 4, pp. 570–578, July 1993.

[12] Andritsos P. and Tzerpos V., “Information Theoretic Software Clustering,” IEEE Transactions on Software Engineering, vol. 31, no. 2, pp. 150-165, 2005.

[13] Asuncion A. and Newman D.J, “UCI Machine Learning Repository,” School of Information and Computer Science, University of California, 2007.

[14] Ayad H. and Kamel M., “Finding Natural Clusters Using Multi cluster Combiner Based on Shared Nearest Neighbours,” in Proceeding of International Workshop Multiple Classifier Systems, Guildford, pp. 166-175, 2003.

[15] Barbara D, Li Y, and Couto J., “COOLCAT: An Entropy-Based Algorithm for Categorical Clustering,” in Proceeding of The Eleventh International Conference on Information And Knowledge Management, Virginia, pp. 582-589, 2002.

[16] Boulis C. and Ostendorf M, “Combining Multiple Clustering Systems,” in Proceeding of European Conference on Principles and Practice of Knowledge Discovery in Databases, Pisa, pp. 63-74, 2004.

[17] Cristofor D. and Simovici D., “Finding Median Partitions Using Information Theoretical Based Genetic Algorithms,” Journal of Universal Computer Science, vol. 8, no. 2, pp. 153-172, 2002.

[18] Domeniconi C. and Al-Razgan M, “Weighted Cluster Ensembles: Methods and Analysis,” ACM Transaction on. Knowledge Discovery Data, vol. 2, no. 4, pp. 1-40, 2009.

[19] Fern X. and Brodley C., “Solving Cluster Ensemble Problems by Bipartite Graph Partitioning,” in Proceeding of International Conference on Machine Learning, Banff, pp. 36-43, 2004.

[20] Fern X. and Brodley C., “Random Projection for High Dimensional Data Clustering: A Cluster Ensemble Approach,” in Proceeding of International Conference on Machine Learning, Washington, pp. 186-193, 2003.

[21] Yun Yang, Ke Chen,” Temporal Data Clustering via Weighted Clustering Ensemble with Different Representations”, Transactions on Knowledge and Data Engineering 23(2):307 – 320

[22] M. Halkidi, Y. Batistakis, and M. Varzirgiannis, “On Clustering Validation Techniques,” J. Intelligent Information Systems, vol. 17, pp. 107-145, 2001.

[23] A. Strehl and J. Ghosh, “Cluster Ensembles—A Knowledge Reuse Framework for Combining Multiple Partitions,” J. Machine Learning Research, vol. 3, pp. 583-617, 2002.

[24] A. Fred and A. Jain, “Combining Multiple Clusterings Using Evidence Accumulation,” IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 27, no. 6 pp. 835-850, June 2005.

[25] E. Keogh, Temporal Data Mining Benchmarks, http://www.cs.ucr.edu/~eamonn/time_series_data, 2010.

[26] Patrick K¨othur, Mike Sips, Henryk Dobslaw, and Doris Dransch, "Visual Analytics for Comparison of Ocean Model Output with Reference Data: Detecting and Analyzing Geophysical Processes Using Clustering Ensembles," IEEE Transactions On Visualization And Computer Graphics, Vol. 20, No. 12, Dec. 2014.

[27] R. W. Lucky, “Automatic equalization for digital communication,” Bell Syst. Tech. J., vol. 44, no. 4, pp. 547–588, Apr. 1965.

[28] S. P. Bingulac, “On the compatibility of adaptive controllers (Published Conference Proceedings style),” in Proc. 4th Annu. Allerton Conf. Circuits and Systems Theory, New York, 1994, pp. 8–16.

[29] G. R. Faulhaber, “Design of service systems with priority reservation,” in Conf. Rec. 1995 World Academy of Science, Engineering and Technology Int. Conf. Communications, pp. 3–8.

[30] W. D. Doyle, “Magnetization reversal in films with biaxial anisotropy,” in 1987 Proc. INTERMAG Conf., pp. 2.2-1–2.2-6.

[31] Zhu X, Li J, Li H-D, Xie M and Wang J,”Sc-GPE: A Graph Partitioning-Based Cluster Ensemble Method for Single-Cell”, Front. Genet. 11:604790. doi: 10.3389/fgene.2020.604790,2020.

[32] Jiaxuan Zhao and Suqin Ji, “Clustering ensemble of massive high dimensional data based on BLB and stratified sampling framework”, CAIH2020: 2020 Conference on Artificial Intelligence and Healthcare, Pages 154–160,2020

[33] Z. Yu, D. Wang, X. -B. Meng and C. L. P. Chen, "Clustering Ensemble Based on Hybrid Multiview Clustering," in IEEE Transactions on Cybernetics, doi: 10.1109/TCYB.2020.3034157.

[34] Malihe Danesh, “Ensemble-based clustering of large probabilistic graphs using neighborhood and distance metric learning”, The Journal of Supercomputing,2021

[35] Yuvaraj, N., Suresh Ghana Dhas, C. High-performance link-based cluster ensemble approach for categorical data clustering. J Supercomput 76, 4556–4579 (2020). https://doi.org/10.1007/s11227-018-2526-z

[36] Tianshu Yang, Nicolas Pasquier, Frederic Precioso, “Ensemble Clustering based Semi-supervised Learning for Revenue Accounting Workflow Management”, 9th International Conference on Data Science, Technology and Applications,2020

[37] Arko Banerjee, Arun K. Pujari, Chhabi Rani Panigrahi, Bibudhendu Pati, Suvendu Chandan Nayak & Tien-Hsiung Weng, "A new method for weighted ensemble clustering and coupled ensemble selection", Connection Science, DOI: 10.1080/09540091.2020.1866496,2021

[38] Xue H, Chen S, Yang Q,"Discriminatively regularized least-squares classification", Pattern Recognition, 42(1):93–104,2009

[39] Yang X-S,"Firefly algorithms for multimodal optimization. Stochastic algorithms: foundations and applications", Springer, pp 169–178,2009

[40] Zaki JM, Peters M, "CLICKS: mining subspace clusters in categorical data via K-partite maximal cliques", In: 21st International Conference on Data Engineering, IEEE Proceedings, pp 355–356,2005

[41] Iam-On N, Boongeon T, Garrett S, Price C,"A link-based cluster ensemble approach for categorical data clustering", IEEE Trans Knowl Data Eng 24(3):413–425,2012.

[42] Mohammad Reza Mahmoudi, Hamidreza Akbarzadeh, Hamid Parvin, Samad Nejatian, Vahideh Rezaie, Hamid Alinejad-Rokny, "Consensus function based on cluster-wise two level clustering", Artificial Intelligence Review,2021.

[43] Huan Niu, Nasim Khozouie, Hamid Parvin, Hamid Alinejad-Rokny, Amin Beheshti,Mohammad Reza Mahmoudi, "An Ensemble of Locally Reliable Cluster Solutions", Applied Sciences, 2020.

[44] Khoshnevisan, B, Rafiee, S, Omid, M, Mousazadeh, H, Shamshirband, S, Hamid, S.H.A, "Developing a fuzzy clustering model for better energy use in farm management systems" , Renew. Sustain. Energy Rev.2015.

[45] Bagherinia, A, Minaei-Bidgoli, B, Hossinzadeh, M, Parvin, H, "Elite fuzzy clustering ensemble based on clustering diversity and quality measures", Appl. Intell. 2019.

[46] Tianshu Yang, Nicolas Pasquier, Frederic Precioso, "Ensemble Clustering based Semi-supervised Learning for Revenue Accounting Workflow Management", Data, 2020.

[47] A. B. Peterson, L. Xu, J. Daugherty, and M. J. Breiding, ''Surveillance report of traumatic brain injury-related emergency department visits, hospitalizations, and deaths, United States, 2014,'' Center Disease Control Prevention, Atlanta, GA, USA, Tech. Rep., 2019.

[48] A. J. Masino and K. A. Folweiler, ''Unsupervised learning with GLRM feature selection reveals novel traumatic brain injury phenotypes,'' , arXiv:1812.00030. [Online]. Available: http://arxiv.org/abs/1812.00030 , 2018.

[49] K. Al-Jabery, T. Obafemi-Ajayi, G. Olbricht, and D. Wunsch, Computational Learning Approaches to Data Analytics in Biomedical Applications. New York, NY, USA: Academic, 2019.

[50] U. von Luxburg, ''A tutorial on spectral clustering,'' Statist. Comput., vol. 17, no. 4, pp. 395–416, Dec. 2007.

[51] J. Shi, J. Malik, ''Normalized cuts and image segmentation",IEEE Trans. Pattern Anal. Mach. Intell., vol. 22, no. 8, pp. 888– 905,2000.

[52] H. Zeng and Y.-M. Cheung, ''Iterative feature selection in Gaussian mixture clustering with automatic model selection,'' in Proc. Int. Joint Conf.Neural Netw., pp. 2277–2282, 2007.

[53] Dacosta Yeboah, Louis Steinmeister, Daniel B. Hier, Bassam Hadi, Donald C. Wunsch II, Gayla R. Olbricht,Tayo Obafemi-Ajayi, "An Explainable and Statistically Validated Ensemble Clustering Model Applied to the Identification of Traumatic Brain Injury Subgroups", IEEE Access, 2020.

Sarumathi Sengottaian received B.E degree in Electronics and Communication Engineering from Madras University, Madras, Tamil Nadu India in 1994 and the M.E degree in Computer Science and Engineering from K.S.Rangasamy College of Technology, Namakkal, Tamil Nadu, India in 2007. She has completed Ph.D degree in Anna University, Chennai in Data Mining. She has a teaching experience of about 22.10 years. At present she is working as Professor in Information Technology department at K.S.Rangasamy College of technology. She has published 19 papers in the reputed International Journals and 2 papers in the reputed National journals. And also she has presented papers in five International conferences and four national conferences. She has received many cash awards for producing cent percent results in university examination. She is a life member of ISTE.

Ms. P. Ranjetha received B.Tech degree in Information Technology from K.S.Rangasamy College of Technology, affiliated to Anna University Chennai, Tamil Nadu, India in 2014 and M.Tech degree in Information Technology from K.S.Rangasamy College of Technology, affiliated to Anna University, Chennai, Tamil Nadu, India in 2016. Now she is working as Assistant Professor in Information Technology department at K.S.Rangasamy College of Technology. She has presented two papers in National level technical symposium. Her Research interests include Mining Medical data, Opinion Mining and Web mining.

Saraswathy Chinnasamy completed B.E degree in Electronics and Communication Engineering from K.S.Rangasamy College of Technology, Tiruchengode, Tamil Nadu, India in 2001 and the M.E degree in Applied Electronics from Government College of Technology, Coimbatore, Tamil Nadu, India in 2005. She has a teaching experience of about 15.10 years. At present she is working as Associate Professor in Electronics and Communication Engineering department at K.S.Rangasamy College of technology. She has published 7 papers in the International Journals and 1 paper in the National journals. And also she has presented papers in eight International conferences and two national conferences. She has received many cash awards for producing cent percent results in university examination. She is a life member of IETE.

Vaishnavi Munusamy completed B.E degree in Computer Science and Engineering from Anna University, Thiruchirapalli in the year of 2011 then she got M.Tech degree in Database Systems from SRM University, Chennai in the year of 2013. She has a teaching experience around 5.6 years. Currently, she is working as an Assistant Professor in Department of Information Technology at K.S.Rangasamy College of technology. She has presented 2 papers in International Conference. She has received many cash awards for producing cent percent results in university examination.

S. Geetha holds a B.Tech degree in Information Technology from K.S.Rangasamy College of Technology, affiliated to Anna, University, Chennai, Tamil Nadu, India in 2014 and M.Tech degree in Information Technology from K.S.Rangasamy College of Technology, affiliated to Anna University, Chennai, Tamil Nadu, India in 2016. Now she is working as Assistant Professor in Information Technology department at K.S.Rangasamy College of Technology. She has published 3 papers in international journals and 3 papers in reputed National Journals. Also, she has presented 2 papers in International Conference and one paper in the National Conference. Her research interests include Image Processing, Mobile Ad hoc Networks and Security.