International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:9, No:11, 2015

Analysis of Diverse Cluster Ensemble Techniques

S. Sarumathi, N. Shanthi, P. Ranjetha

***Abstract*—**Data mining is the procedure of determining interesting patterns from the huge amount of data. With the intention of accessing the data faster the most supporting processes needed is clustering. Clustering is the process of identifying similarity between data according to the individuality present in the data and grouping associated data objects into clusters. Cluster ensemble is the technique to combine various runs of different clustering algorithms to obtain a general partition of the original dataset, aiming for consolidation of outcomes from a collection of individual clustering outcomes. The performances of clustering ensembles are mainly affecting by two principal factors such as diversity and quality. This paper presents the overview about the different cluster ensemble algorithm along with their methods used in cluster ensemble to improve the diversity and quality in the several cluster ensemble related papers and shows the comparative analysis of different cluster ensemble also summarize various cluster ensemble methods. Henceforth this clear analysis will be very useful for the world of clustering experts and also helps in deciding the most appropriate one to determine the problem in hand.

***Keywords*—**Cluster Ensemble, Consensus Function, CSPA, Diversity, HGPA, MCLA.

I. Introduction

THE real data mining job is to automatic or semi-automatic investigation of large quantities of data to excerpt earlier unknown interesting patterns such as groups of dependencies (association rule mining), unusual records (anomaly detection) and data records (cluster analysis). This usually includes via database methods like spatial indices. These designs can be realizing as a kind of summary of the input data, and possibly used in further analysis that is in predictive analytics and machine learning. For an example, the data mining step influence identifies multiple groups in the data can be used to obtain more accurate prediction results of a decision support system. Neither the data collection nor data preparation outcomes interpretation and reporting are a portion of the data mining step, but do be appropriate to the general KDD process as additional steps.

A. Process

The Knowledge Discovery in Databases (KDD) process generally defined by the five major stages:

- Selection
- Pre-processing
- Transformation
- Data Mining
- Interpretation/Evaluation

B. Data Mining Involves Six Common Classes

- *Anomaly detection (Outlier/change/deviation detection)* – The documentation of uncommon data records might be interesting or data inaccuracies that have need of further investigation.
- *Association rule learning (Dependency modeling)* – Searches for relationships in the middle of variables. For an example, a supermarket might collect data on customer purchasing habits. By using association rule learning, the supermarket can govern the products are commonly bought together and use this information for marketing purposes. This occasionally referred to as market basket analysis.
- *Clustering* – Clustering is the process of discovering groups plus structures in the data that are in several ways or extra "similar", without by using well-known arrangements in the data.
- *Classification* – Classification is the process of generalizing known structure to communicate to original data. An example, an e-mail programs influence attempt to classify an e-mail as "legitimate" or as "spam".
- *Regression* – Regression tries to discover a function, which modeling data with least error.
- *Summarization* – Summarization provided that a denser representation of the data set, with visualization and report generation.

C. Cluster Analysis

Cluster analysis or else clustering is the process of grouping a set of objects in such a way that objects in the same group (known as cluster) are more related (in certain sense or another) to each other than to those in additional groups (clusters). It is a chief task of exploratory data mining, and a common method for statistical data analysis, used in numerous fields, containing bioinformatics, pattern recognition, machine learning, information retrieval, and analysis. Cluster ensembles offer a framework for combining multiple bases clustering of a dataset hooked on a single consolidated clustering deprived of accessing the features of the data or else base clustering algorithms [2]. Cluster ensembles produce more robust and stable clustering outcomes compared to single clustering algorithms [4]. By computing the base clustering in an entirely distributed manner, cluster ensembles can leverage distributed computing [3]. Since cluster ensemble is essential to access the base clustering outcomes rather than the data. They deliver an expedient method to privacy

Mrs. S. Sarumathi, Associate Professor, is with the Department of Information Technology, K. S. Rangasamy College of Technology, Tamil Nadu, India (phone: 9443321692; e-mail: rishi_saru14@rediffmail.com).

Dr. N. Shanthi, Professor and Dean, is with the Department of Computer Science Engineering, Nandha Engineering College, Tamil Nadu, India (e-mail: shanthimoorthi@yahoo.com).

Ms. P. Ranjetha, PG Scholar, is with the Department of Information Technology, K. S. Rangasamy College of Technology, Tamil Nadu, India (e-mail: ranjethamtechit@gmail.com).

preservation also knowledge reuse [3]. Desirable aspects have made the study of cluster ensembles more important in the context of data mining. To produce a consensus clustering from a whole set of base clustering, it is highly desirable for cluster ensemble algorithms have numerous additional properties suitable for actual life applications. Present cluster ensemble algorithms are CSPA, HGPA, k-means based algorithms, etc.

- *From K-Means to Hierarchical Clustering*

Two properties of K-means clusterings are

✓ Exactly fits the K clusters, as mentioned
✓ The final clustering assignment based on the chosen initial cluster centers.

A Given pairwise dissimilarities d_{ij} among data points, hierarchical clustering produces a consistent outcome, without the need to select initial starting positions that are the number of clusters. To measure the dissimilarity between the groups one of the ways is linkage. The linkages known, the hierarchical clustering produces a sequence of clustering assignments. All the points are in their own cluster at one end and at the other end all points are in one cluster.

- *Agglomerative versus Divisive*

Two types of hierarchical clustering algorithms are:

✓ Agglomerative
✓ Divisive

Agglomerative:

i. Agglomerative is the bottom-up approach.
ii. It starts up with all points in their own group.
iii. Repeat until there is only one cluster; merge the two groups that have the smallest dissimilarity as measured by linkage.

Divisive:

i. Divisive is the top-down approach.
ii. It starts with all points in one cluster.
iii. Repeat until all points in their own cluster; split the group into two resulting in the biggest dissimilarity.

- *Dendogram*

Dendogram is an expedient graphic to display a hierarchical sequence of clustering projects. It is simply a tree, where

i. Every node represents a group
ii. Every leaf node is a singleton that is a group containing a single data point.
iii. The root node is the group that containing an entire data set
iv. Each internal node has two daughter nodes that representing the groups that were merge to form it.

- *Linkages*

There are three different types of linkages used in dendogram. They are:

- Single Linkage
- Average Linkage
- Complete Linkage

Agglomerative clustering method gives the linkages.

i. Agglomerative Clustering starts with all points in own group
ii. Till there is only one group, continuously merge the two clusters G, H such as $d(G,H)$ is smallest

- *Single Linkage*

Single linkage is the nearest neighbour linkage. The dissimilarity between G, H is the shortest distance flanked by the two points in the opposed group.

$$d_{single}(G,H) = \min_{i \in G,\, j \in H} d_{ij}$$

$d_{single}(G,H)$ is the distance between the closest pair.

- *Complete Linkage*

Complete linkage is the furthest neighbour linkage. The dissimilarity between G, H is the largest distance flanked by the two points in the opposed group.

$$d_{complete}(G,H) = \max_{i \in G,\, j \in H} d_{ij}$$

$d_{complete}(G,H)$ is the distance amid the furthest pair.

- *Average Linkage*

Average linkage is the dissimilarity among G, H which is the average similarity over all points in opposed group.

$$d_{average}(G,H) = \frac{1}{n_G \cdot n_H} \sum_{i \in G,\, j \in H} d_{ij}$$

$d_{average}(G,H)$ is the average distance amid all pair.

- *Common Properties of All Linkages*

i. All linkages operate on dissimilarities d_{ij} and they do not want the points $X_1, \cdots X_n$ to be in Euclidean distance.
ii. Running agglomerative clustering with some of these linkages creates a dendogram with no overturns.
iii. While running the algorithm, dissimilarities scores among merged group increases the height of the parent and is always higher than its daughter is.

D. Clustering Methods

Subsequently, there are many clustering methods have been developed, each of which uses a different induction principle. Farley and Raftery divide the clustering methods into two main groups: hierarchical and partitioning methods. Han and Kamber categorize the methods into additional three main categories: density-based methods, model-based clustering and grid-based methods.

- *Partition methods:*

i. Find mutually exclusive clusters of spherical shape
ii. Distance-based

iii. May use mean or medoid (etc.) to represent the cluster center
iv. Effective for small-to medium-size data sets

- *Hierarchical methods*:

i. Clustering is a hierarchical decomposition that is multiple levels.
ii. Cannot correct erroneous merges or splits
iii. May incorporate other techniques like micro clustering or consider object "linkages".

- *Density-based methods:*

i. Can find arbitrarily shaped clusters
ii. Clusters are compact regions of objects in space that are separated by low-density region
iii. Cluster density: Each point must have a minimum number of points within its "neighborhood"
iv. May filter out outliers

- *Grid-based methods:*

i. Use a multi resolution grid data structure
ii. Fast processing time (typically independent of the number of data objects, yet dependent on grid size)

- *Model-based methods:*

i. These methods attempt to optimize the fit among the known data and certain mathematical models.
ii. Not like conventional clustering, which identifies collections of objects; model-based clustering approaches also find characteristic descriptions for each group, where each group represents a concept or class.
iii. The furthermost frequently used induction methods are decision trees and neural networks

E. Different Clustering Algorithms

An excellent clustering method will produce high superiority clusters with high intra class similarity and low inter class similarity. A huge variety of clustering algorithms which are of well established such as K-Means, EM (Expectation Maximization) based on the spectral graph theory, K-modes, GAClust, CobWeb. STIRR, Robust Clustering Algorithm for Categorical Attributes (ROCK), CLICK, Clustering Categorical Data Using Summaries (CACTUS), COOLCAT, Differential fuzzy clustering, CLOPE, Squeezer, Standard Deviation of Standard Deviation Roughness algorithm, Frequency of attribute value grouping algorithm and certain hierarchical clustering algorithms like Divisive algorithm, LIMBO, single link, Fuzzy C-Means, Fuzzy C-Medoids etc. are emerging over earlier periods. Conversely, it known that there is no single clustering technique is capable of giving accurate and appropriate cluster results. By applying a clustering algorithm to the dataset, it works based on the internal criteria, i.e. similarity or dissimilarity measures used in that algorithm. Simultaneously, if two diverse clustering algorithms applied to the same data set consequently it will result in very different cluster solutions. Therefore, this critical concern is very difficult to evaluate the exact clustering results.

The clustering ensemble methods divided into two steps: Generation and Consensus Function [1].

Some of the clustering ensemble processes are:

i. *Robustness*: The grouping process has better average performance than the single clustering algorithms.
ii. *Consistency*: The outcome of the grouping should be similar to all combined single clustering algorithm outcomes.
iii. *Novelty*: Cluster ensembles must permit discovering solutions unachievable by single clustering algorithms.
iv. *Stability*: Outcome is lower sensitivity to noise and outliers.

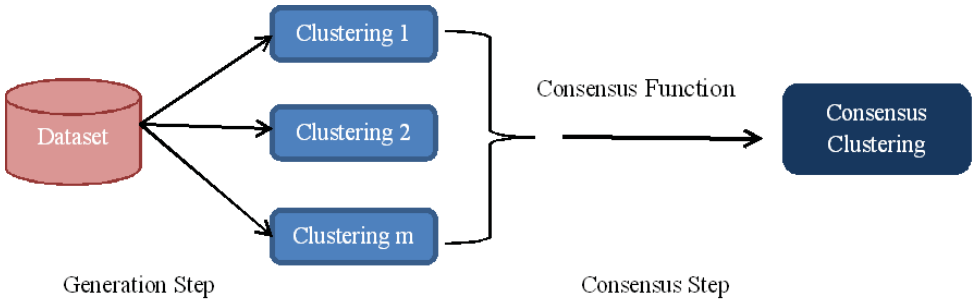

Fig. 1 General Process of cluster ensemble

The first step in Generation step is cluster ensemble methods, the set of clustering that will be combined is generated [1]. It is important to relate an appropriate generation process, for the reason that the outcome conditioned by the initial clustering obtained. In the generation step, there are no constraints around how the partitions initiated. Disparate clustering algorithms or the same algorithm with dissimilar parameter initialization applied in this process. A uniform dissimilar objects representation, diverse subsets of objects or projections of the objects on unlike subspaces used in Fig. 2.

The consensus function is the most important step in any clustering ensemble algorithm. Challenges in clustering ensemble are the definition of an appropriate consensus function is proficient of improving the outcomes of single clustering algorithms [1]. Here, the concluding data partition or else the consensus partition, which is the outcome of any clustering ensemble algorithms obtained. However, the consensus among a set of clustering not obtained in the same way in all cases. There are two main consensus function

approaches and they are objects co-occurrence and median partition. In the first consensus function approach, the hint is to define which must be the cluster label associated with all objects in the consensus partition. In the second approach, the consensus partitions attained by the solution of an optimization problem; find the median partition with respect to the cluster ensemble.

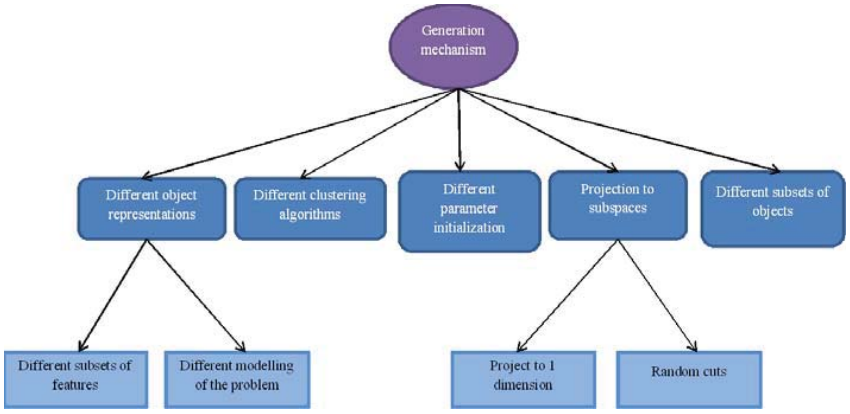

Fig. 2 The primary clustering ensemble generation mechanisms

II. Different Cluster Ensemble Methods

The next sections will discuss about some diverse collection of cluster ensemble approaches [5]. In addition, for all method its systematic working process explicated.

A. Cluster Ensemble Selection (CES)

Cluster ensemble selection studies the ensemble selection problem for unsupervised learning (clustering) [6]. A big library of dissimilar clustering solutions aim is to choose a subset of solutions to create a smaller so as better performing cluster ensemble than using completely available solutions.

1. Selection Based on Quality and Diversity

In supervised learning (classification), the quality and diversity are well-recognized concepts where diversity measures the dissimilarity in the predictions made using the ensemble members and quality measures the accuracy. For unsupervised learning, data are not defined perfectly [6]. Here they explain how to measure the diversity and quality of the clustering solutions.

a) Quality

For unsupervised clustering, a number of external objective functions like accuracy to measure the quality of the clustering solutions. In clustering literature, it pooled to use predefined class labels as a surrogate for the true causal structure and then measure the quality of a clustering solution depends on how it improves the class labels. The ensemble selects, as supervised information likes the class labels are not integrated. The internal quality measure is depending on an impartial function [3] for designing consensus functions.

$$SNMI(C,E)=\sum_{i=1}^{r} NMI(C,C_i) \quad (1)$$

where, an ensemble E for r clustering solution i, $E=\{C_1,C_2,\cdots,C_r\}$. $NMI(C,C_i)$ is the normalized mutual information among clustering C plus C_i. NMI is 0 while two clustering completely independent partitions and NMI value is 1 once two clustering describes the similar partition of the data. A giant library of clustering solutions $L=\{C_1,C_2,\cdots,C_r\}$, use $SNMI(C_i,L)$ to measure the quality of every clustering solution C_i.

b) Diversity

For cluster ensembles, a number of dissimilar diversity measures and the pairwise normalized mutual information [19] between clustering solutions were projected. The pairwise similarity of two clustering such as $NMI(C_i,C_j)$ and calculate the sum of all pairwise similarities $\sum_{i\neq j,C_i,C_j\in E} NMI(C_i,C_j)$ inside the ensemble as a measure of the ensemble diversity. The better quality is the diversity and the low-grade is the value.

2. Consensus Function

Once a cluster ensemble created through selective and we need a consensus function to merge the selected solutions to create a final consensus clustering. A variety of consensus functions proposed such as HBGF [13], CSPA [3] and hierarchical agglomerative approach [7]. They focus on the Cluster-based Similarity Partition Algorithm (CSPA). It creates a similarity matrix depend on the clustering solutions in the ensemble that measures for every pair of data points the frequency is being clustered together in the ensemble also known as co-association matrix. Relate a graph partition algorithm to the similarity matrix to form a final clustering. Subsequently, apply spectral graph partition [14] to create a final partition of the data points into number of known classes

in the data clusters.

3. Cluster and Select (CAS)

There are multiple possible ways to partition the clustering solutions. However, here they apply spectral clustering [14] to the pairwise NMI matrix considered as a similarity matrix describing the relationship between clustering solutions and this technique referred to as CAS.

B. Combining Multiple Clusterings Using Evidence Accumulation

Evidence Accumulation (EAC) is for combining the results of multiple clusterings. In EAC concept, every partition observed as an independent evidence of data organization, which separate data partition merged and depends on a voting mechanism to produce a new $n \times n$ similarity matrix among the n patterns [7].

Consider N partitions of data X, and P determine the set of N partitions that define as a clustering ensemble:

$$\begin{aligned} \mathrm{P} &= \left\{ p^1, p^2, \ldots p^N \right\} \\ p^1 &= \left\{ C_1^1, C_2^1, \ldots, C_{k_1}^1 \right\} \\ &\vdots \\ p^N &= \left\{ C_1^N, C_2^N, \ldots, C_{k_N}^N \right\} \end{aligned} \tag{2}$$

1. Evidence Accumulation Clustering

To address cluster ensemble combination problem, the concept of evidence accumulation clustering flourished. They compose no assumptions on the number of clusters k_i in every data partition p_i and the number of clusters k_1 in the merged data partition p_1. It is approximate that the merged data partition p_i will explain natural groupings of the data compared to the individual clustering outcomes p_i.

The inkling of evidence accumulation clustering is to merge the outcomes of multiple clustering into a single data partition via viewing every clustering outcome as an independent evidence of data organization [7]. The three issues are

i. How gather evidence or else how to create the clustering ensemble?
ii. How merge the evidence?
iii. How to citation a reliable data partitioning from merged evidence.

2. Producing Clustering Ensembles

Clustering ensembles are fashioned by means of the following approaches:

a) Selection of Data Representation

i. Employing dissimilar pre-processing and/or else feature extraction mechanisms, that eventually lead to dissimilar pattern representations like vectors, strings, graphs, etc. or dissimilar feature spaces.
ii. Discovering subspaces of the similar data representation like by subsets of features
iii. Perturbing the data, akin to in bootstrapping techniques (like bagging), or sampling approaches.

b) Selection of Clustering Algorithms or Algorithmic Parameters

i. Apply the disparate clustering algorithms
ii. Use the identical clustering algorithm with disparate parameters or else initializations.
iii. Explore diverse dissimilarity measures for assessing interpattern relationships, inside a known clustering algorithm.

3. Combining Evidence: The Co-Association Matrix

In order to manage the partitions with the dissimilar number of clusters they propose a voting mechanism to merge the clustering outcomes that leads to a new measure of similarity along with patterns. The primary supposition is the patterns belonging to a natural cluster are very probable to be collocated in the similar cluster in disparate data partitions [7]. The co-occurrences of pairs of patterns in the similar cluster as votes aimed at their association, the N data partitions of n patterns mapped into a $n \times n$ co-association matrix.

$$C(i, j) = \frac{n_{ij}}{N} \tag{3}$$

where n_{ij} is the number of times the pattern pair (i, j) is assigned to the identical cluster between N partitions.

C. Moderate Diversity for Better Cluster Ensembles (MDCE)

In cluster ensembles, Adjusted Rand index (ARI) is charity to measure diversity [8]. Select ARI because of their following properties [16]:

i. ARI has a fixed value of zero (0), if the two compared partitions are fashioned independently from one another.
ii. In the experiment, this index found to have a greater sensitivity to pick out good partitions likened to other indices.

Clustering algorithm correctness or a cluster ensemble accuracy measured using the match between the partition produced and specific known as ground-truth partition.

a) Procedure for Building Cluster Ensembles

i. Create K ensembles changing the random parameters of the clustering algorithm.
ii. Calculate diversity by using a preferred diversity measure.
iii. Identify the median of the diverse values and pick the equivalent ensemble.

b) Several Ways to Construct a Cluster Ensemble

i. Using dissimilar subsets of the feature (overlap or disjoint) known as feature-distributed clustering [3], [20], [21].

ii. Use dissimilar clustering algorithms within the ensemble [15]. Such ensembles are known as hybrid or heterogeneous. An ensemble with the similar clustering method obtaining, using varying a random parameter called homogeneous.
iii. Modify a random parameter of the clustering algorithm [19], [22].
iv. Use unlike data set for every ensemble member called object-distributed clustering [3].

1. Diversity Measures for Cluster Ensembles

The ARI needed for diversity and accuracy of the ensemble [16]. Let us take A and B be the two partitions on dataset Z with N objects. Suppose A have C_A clusters and B have C_B clusters. Represented by

i. N_{ij} is the number of objects in cluster i in partition A also in cluster j in partition B.
ii. $N_{.j}$ is the number of objects in clusters j in the partition B.
iii. $N_{i.}$ is the number of objects in clusters i in the partition A.

The ARI is

$$ar(A,B)=\frac{\sum_{i=1}^{C_a}\sum_{j=1}^{C_b}\binom{N_{ij}}{2}-t_3}{\frac{1}{2}(t_1+t_2)-t_3} \quad (4)$$

where

$$t_1=\sum_{i=1}^{C_a}\binom{N_{i.}}{2},\ t_2=\sum_{j=1}^{C_b}\binom{N_{.j}}{2},\ \text{and}\ t_3=\frac{2t_1t_2}{N(N-1)}$$

There are two approaches, to measure the ensemble diversity as well as pairwise and non-pairwise. The non-pairwise approach is separated into group diversity [17] and individual diversity.

In pairwise approach ARI is used and the ensemble diversity is,

$$D_p=\frac{2}{L(L-1)}\sum_{i=1}^{L-1}\sum_{j=i+1}^{L}\left(1-ar\left(P_i,P_j\right)\right) \quad (5)$$

D.Resampling-Based Selective Clustering Ensembles (RBSCE)

Cluster analysis classifies data items into clusters so that items in the similar cluster are more similar to each other and they are more dissimilar in different cluster [9]. The Cluster ensemble is an effective method to improve the robustness and stability of cluster analysis. The two important factors of cluster ensemble are:

i. An accurate and diverse ensemble group of the clustering ensembles is constructed.
ii. A proper consensus function to merge all clustering outcomes of the ensemble group designed.

1. Clustering Ensembles Using Part of all Available Clustering Results

The clustering ensemble method uses all of obtained clustering outcomes with the following steps,

i. A population of clustering outcomes acquired via executing dissimilar clusters on the similar data set.
ii. The ensemble group built by all acquired clustering outcomes [9].
iii. A consensus function espoused to merge all clustering outcomes of the ensemble group. Unlike classification problems where labels of data items are accepted, where data items in unsupervised clustering problems is unlabeled. As a result, there is no explicit correspondence among outcomes offered using dissimilar clusters.

2. Resampling-Based Selective Clustering Ensembles

The framework of resampling-based selective clustering ensembles that discriminating merges part of all acquired clustering outcomes. Resampling-based selective clustering ensembles work with gauging the qualities of all acquired clustering outcomes by using the resampling method and selecting part of promising clustering outcomes to generate the ensemble group [9]. The last solution attained via merging all the selected clustering outcomes of the ensemble group.

Algorithm1. Framework of resampling-based selective clustering ensembles

i. $\left\{I^{(1)},I^{(2)},\cdots,I^{(M)}\right\}\leftarrow$ A population clustering outcomes are acquired
ii. $\left\{q_{(1)}^{(a)},q_{(2)}^{(a)},\cdots,q_{(M)}^{(a)}\right\}\leftarrow$ Calculate the accuracy of all outcomes using resampling.
iii. $\left\{q_{(1)}^{(d)},q_{(2)}^{(d)},\cdots,q_{(M)}^{(d)}\right\}\leftarrow$ Calculating individual diversity factors as $q_{(i)}^{(d)}=1-\frac{\sum_{j=1,j\neq i}^{M}\left\|S^i,S^j\right\|}{M-1}\ i=1,\cdots,M$
iv. $\left\{fitness\left(I^{(1)}\right),\cdots,fitness\left(I^{(M)}\right)\right\}\leftarrow$ Calculate the fitness value as $fitness\left(I^{(i)}\right)=(1-\lambda).\frac{q_i^{(a)}}{Q^{(a)}}+\lambda.\frac{q_i^{(d)}}{Q^{(d)}}\ i=1,\cdots,M$
v. $\left\{'I^{(1)},'I^{(2)},\cdots,'I^{(M)}\right\}\leftarrow$ Reorder the cluster outcome, such as $\left\{fitness\left('I^{(1)}\right)\geq\cdots\geq fitness\left('I^{(M)}\right)\right\}$
vi. $\left\{I^{(1)},'I^{(2)},\cdots,'I^{(N)}\right\}\leftarrow$ Choose N best clustering outcomes.
vii. $I^{(final)}\leftarrow$ Merge it $\left\{I^{(1)},'I^{(2)},\cdots,'I^{(N)}\right\}$ together.

E. Bagging-Based Spectral Clustering Ensemble Selection (BBSCES)

They proposed a spectral clustering ensemble selection depend on the bagging method for ranking the cluster components. The input of the ensemble system created using the perturbation of spectral clustering (SC) [10]. The

researches on clustering ensemble concentrate on two challenges,

i. How the diverse components clusterings for a cluster ensemble are generated
ii. How the design consensus function.

1. Spectral Clustering (SC)

SC translates the clustering to a multi-way partition of a non-directional graph. They viewed samples as the nodes of the non-directional graph $G(V,E)$. The chief heuristic for the SC is to map data points into a new space with k-dimensions using eigenvector decomposition. The data points clustered effectively in the k-dimensional space.

2. Spectral Clustering Ensemble Selection

To build the selective ensemble learning system, first create the diverse components and select the best ones to choose. The framework of selective SC ensemble consists of three parts,

i. The generation of the diverse components
ii. Selection of the clustering
iii. The aggregation of the selected multiple clustering.

3. Generation of the Diverse Component Clustering

The diversity of the component clustering found using multi-perturbation embedded in the procedure of SC [10], [23] – [27]. Unlike component, clusterings created using the following methods.

i. Nystrom approximation [28]
ii. Random scaling parameter
iii. Random initialization

4. Consensus Function

To merge the component clustering, a consensus function wanted to merge them and produce a final partition. The [3] three methods of graphs are HGPA, CSPA, and MCLA. MCLA and CSPA approach are using as the consensus function. Meta clustering algorithm (MCLA) used for complexity.

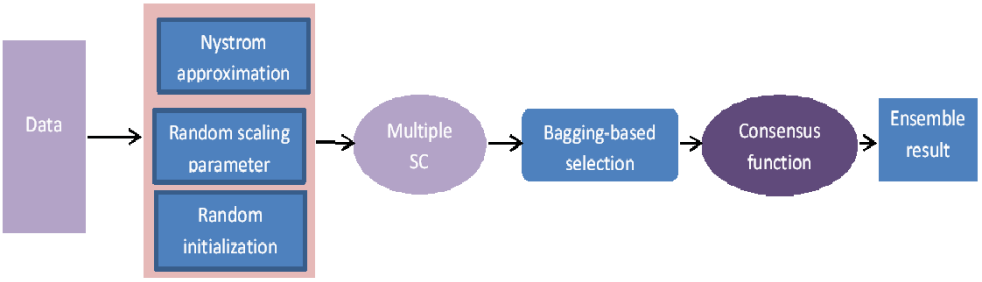

Fig. 3 Selective Spectral Clustering Ensemble Framework

In Cluster-based similarity partition algorithm (CSPA) the co-association matrix of component, clustering calculated. Then a graph engendered whose nodes are the samples and whose weights among pairwise points are the corresponding values in the co-association matrix [10]. Hypergraph-partition algorithm (HGPA) denotes every cluster using a hypergraph in a graph where the nodes corresponding to the known objects and the minimal cut algorithm like HMETIS used to get the consensus partition. MCLA relates the hyperedge collapsing operation to determine the object belongs to which cluster. However, the algorithms depend on the hypergraph method remove hyperedge as the entire and work best for balanced cluster.

MCLA is a collection of the following steps:

i. Constructing the meta-graph
ii. Partition the meta-graph.
iii. Computing cluster members

5. Clustering Selection Method

The consensus partition and the ensemble member are clustering. Various approaches proposed to evaluate the relevance among the two clustering. Normalized mutual information (NMI) used to measure the diversity of the component clustering. ARI is an alternative measure of the diversity and accuracy between two clusters. The relevance among each component clustering and the consensus partition is as,

$$Rank = 1 - NMI(p_i, P^*)$$
$$or$$
$$Rank = 1 - ARI(p_i, P^*) \quad (6)$$

F. An Efficient and Scalable Family of Algorithms for Combining Clusterings (ESFA)

In this paper, a new family of algorithms for merging multiple clustering is contributed. The chief motivation is to improve new techniques that solve the weakness of the related work. The complete experimental outcomes on a variety of datasets determine that these methods are fast, robust and require very less memory compared to the state-of-the-art [11].

1. Combining Multiple Clustering

Combining multiple clusterings takes a group of clustering and offers a final clustering with quality. Several state-of-the-art methods for combining multiple clusterings are Combining multiple clusterings using similarity graph (COMUSA), Link-based cluster ensemble (LCE), CSPA [32], HGPA [33], MCLA [3], Combining multiple clusterings using evidence

accumulation (EAC), Bipartite merger (BM) and Metis merger (MM).

i. COMUSA [29], [30] is a graph-based algorithm that makes use of the evidence accumulated in the group of input clusterings and creates a very good quality final clustering. A similar graph created by calculating the co-associations of objects in the input. COMUSA initiates a new cluster by using selecting a pivot vertex and out spreading the cluster along with its neighbors if they are most similar to the pivot. The advantage is that COMUSA correctly identify the number of clusters in the final clustering automatically.
ii. LCE [31] begins with a bipartite membership graph of objects and clusters and forms up a dense graph using implied similarities among all clusters and each object. LCE creates a final clustering on this structure by using a spectral graph partitioning technique. It is design to work on gene expression data sets, creates good outcomes on biological and non-biological data sets.
iii. EAC [7] amasses the evidence in all clusters to create a co-association matrix, SM. A similarity matrix (SM) is delivered to an agglomerative clustering algorithm
iv. BM and MM [34] are in merging clusters and denoted by sets of cluster centers. BM works on some clusterings all having n clusters. It collects the centroids according to their similarity and combines them to have a final clustering with n clusters. MM use of METIS and it is more flexible. In MM, clusterings can have different number of clusters.

2. Weaknesses of Related Work

CSPA, HGPA, EAC, and COMUSA, work at the object level and every method produces a dense graph of objects [11]. CSPA and HGPA make use of METIS package that partitions a graph having roughly the same number of vertices.

i. CSPA [32] and HGPA [33] work very fast and their result may not be accurate when especially clusters dissimilar in size.
ii. EAC [7] and COMUSA [29], [30] create a final clustering with a very high accuracy in a wide range of datasets. Nevertheless, they easily become inefficient for large data sets in terms of both run time and memory.
iii. LCE [31] not only make use of multiple clusterings as the input, but it needs the real dataset that is available in some cases. LCE calculates a bipartite membership graph of objects and clusters with implied similarities that requires a lot of computation.
iv. BM and MM [34]is prototype based cluster ensemble algorithms that are verify on only globular shape data sets.
v. MCLA [3] works at cluster level. Therefore, it creates a final clustering in a small amount of time and takes less amount of memory. Still, MCLA makes use of Jaccard measure that captures the syntactic similarity among clusters that downgrades the accuracy of the method. Median partition and genetic methods are also used for combining multiple clusterings

Our chief motivation of proposing new algorithms is that function at cluster level is to resolve the scalability problem. However, there is usually a trade-off between run time and final clustering validity. A new family of clustering algorithms creates a new final clustering having high accuracy in a less amount of time.

3. A New Family of Algorithms for Combining Clusterings

A novel family of algorithms for combining multiple clusterings that works at cluster level. A fast improvement of COMUSA remedies its execution time shortcoming. A scalable and efficient algorithm designed to work on a diverse set of data sets through a broad range of features. This algorithm improves both the accuracy and execution time.

G.Cluster Ensembles – A Knowledge Reuse Framework for Combining Multiple Partitions (KRF)

The problem of combining multiple partitioning of a set of objects into a single consolidated clustering devoid of accessing the features or algorithms that determined these partitions. Find some application scenarios for the resultant 'knowledge reuse' framework that is called as cluster ensembles [3]. In terms of shared mutual information, the cluster ensemble problem then formalized as a combinatorial optimization problem. In addition to a direct maximization, approach three effective and efficient techniques for obtaining high-quality consensus functions.

i. A similarity measure from the partitioning and then reclusters the objects.
ii. Combiner is depending on hypergraph partitioning.
iii. Merge groups of clusters into meta-clusters then contend for all objects to conclude the combined clustering.

The efficiency of cluster ensembles in three qualitatively dissimilar application scenarios:

i. The original clusters are designed depending on non-identical sets of features
ii. The original clustering algorithms functioned on non-identical sets of objects
iii. A general dataset is used to combine multiple clusterings is to improve the quality and robustness of the solution.

Not like classification or regression settings, there are very limited approaches planned for combining multiple clusterings. Distinguished exceptions include:

i. While strict consensus clustering for designing evolutionary trees that usually leading to a solution at a much lower resolution than that of the individual solutions.
ii. Combining the outcomes of some clusterings of a known data set where all solution exists in a publicly known as feature space.

1. Knowledge Reuse

A diversity of clusterings for the objects under attention may exist and requirements to either assimilate these clusterings into a single solution. For example, clustering of mortgage loan applications depends on the information in the application methods can be accompanied by segmentations of the applicants designated by external sources like FICO scores

delivered by Fair Isaac.

2. The Cluster Ensemble Problem

The problem of combining multiple clusterings, suggest an appropriate objective function for determining a single consensus clustering, and explore the feasibility of directly optimizing this objective function by using greedy approaches [3].

3. Objective Function for Cluster Ensembles

Let $I(X,Y)$ represent the mutual information among X and Y [35].

i. $H(X)$ represent the entropy of X .
ii. $H(Y)$ represent the entropy of Y .
iii. $I(X,Y)$ is a metric and there is no upper bound, so it is easy for interpretation as well for comparison. $I(X,Y)$ ranges from 0 to 1 is desirable. Some normalizations are possible depend on the observation that $I(X,Y) \le MIN(H(X),H(Y))$ and these contain normalizing by means of the arithmetic or geometric mean of $H(X)$ and $H(Y)$. $H(XI=(X,X)$ in Hilbert space, the geometric mean as of the analogy with a normalized inner product. NMI used is:

$$NMI(X,Y) = \frac{I(X,Y)}{\sqrt{H(X)H(Y)}} \quad (7)$$

4. Efficient Consensus Functions

Here they introduce three efficient heuristics to solve the cluster ensemble problem. All algorithms approach the problem by first transforming the set of clusterings into a hypergraph representation. They are CSPA, HGPA and MCLA.

H.A Framework for Hierarchical Ensemble Clustering (FHEC)

In Hierarchical Ensemble clustering (HEC) framework, the input may be in the form of both partitional and hierarchical clustering, [18] but the output is a consensus hierarchical clustering. The three different cases are

i. When the input clustering is partitional clustering that leads to the usual ensemble clustering. First, the aggregate consensus distance between the partitional clustering is constructed, and then a consensus clustering created by using the consensus distance. In HEC, a hierarchy structure is generating on top of the consensus clustering by using the consensus distance. The cluster hierarchy structure determination a problem in the normal ensemble clustering when the input partitional clustering has a diverse number of clusters.
ii. When the input clustering is hierarchical clusterings, a dendogram is demarcates as a nested family of partitions normally signified graphically as a rooted tree [36]. Dendogram represent a hierarchical decomposition of the underlying data set. First aggregate dendogram constructed among objects, and then a hierarchical clustering as the outcome created as the result.
iii. The input clustering contains both partitional as well as hierarchical clusterings. Construct the consensus distance from the partitional clustering as well the dendogram distance. These distances merged into a single distance. As an outcome, a hierarchical clustering constructed.

1. Hierarchical Clustering

In Hierarchical, clustering algorithms unsupervised methods used to create tree-like clustering solutions. Group the data points into a hierarchical tree structure by using agglomerative (bottom-up) or divisive (top-down) methods [37]. The agglomerative method starts taking every data point as a single cluster and forms bigger clusters by combining similar data points together until the complete dataset is encapsulated into a single cluster. The top-down approaches start with every data point in one cluster and split the larger clusters. Research efforts reported on algorithm-level developments in the hierarchical clustering process and on sympathetic hierarchical clustering [38]–[40].

2. Ensemble Clustering

In Ensemble, clustering finding a problem of combined clustering outcome depends on multiple input clusterings of a given dataset. Several methods are used to get multiple clusterings like applying diverse clustering algorithms by using resampling to become subsamples of the dataset and utilizing feature selection methods to get dissimilar feature spaces and exploiting the chance of the clustering algorithm. Various methods [12], [41] developed to solve ensemble-clustering problems. Present ensemble clustering methods designed for partitional clustering methods. The main difference is that a coherent algorithm to study the closest ultra-metric solution while the approach needs many parameters that selected in an ad hoc manner. A hierarchical ensemble-clustering framework combines both partitional clustering and hierarchical clustering outcomes.

3. Consensus Tree

The techniques for answering the consensus problem are established on agreement subtrees i.e., the substructure is common to all the trees. It is difficult for the consensus tree [42], [43] methods to sanctuary structural information while including all the existing leaves from the input trees. A framework depends on descriptor matrices are projected to reserve the general structures from the input clustering and produce a full consensus tree.

4. Cluster Ensemble Selection

Selecting a subset of the input clustering problem is to form a smaller but then improved performing cluster ensemble by using available solutions studied recently for partitional clustering [44]. A cluster ensembles selection method for hierarchical clustering depends on tree distance to combine both multiple hierarchical clustering and partitional clustering results. [18] The Dendrogram selection problem is to build a

technique for learning the ultra-metric distance from the aggregated distance.

5. Ultra-Metric and Dendrogram Reconstruction

A dendrogram graphically represent a rooted tree where leaves denote data objects as well internal nodes denote cluster at several levels and defined as a nested family of partitions. Pairwise cophenetic proximity measures and the level because of which two data objects are first merged into a cluster reserve the structural information [18]. The dendrogram job is to allocate distances among leaf nodes. Every single of these dendrogram distances is in fact an ultra-metric distance. It is significant because given an ultra-metric distance matrix $D = d_{ij}$ reconstructs the original tree.

6. Hierarchical Ensemble Clustering Algorithm Strategy

The algorithmic approach of our hierarchical ensemble clustering is

i. The Dendrogram distance measure used to produce an ultra-metric dendrogram distance for all input dendrogram and the consensus distance matrix for partitional clustering results.
ii. Aggregate the ultra-metric dendrogram distances and the consensus distance for partitional clusterings.
iii. Discover the closest ultra-metric distance from the aggregated distance.
iv. Build the final hierarchical clustering.

TABLE I
SUMMARIZED CLUSTER ENSEMBLE METHODS

Clustering Ensemble Methods	Ensemble Size	Type of Consensus Function used	Dimensionality (size of the dimensions used in the datasets)	Types of Dataset used	Algorithm used to build Base Clustering	Measures
CES	Fixed	CSPA, HBGF, Hierarchical Agglomerative Approach	Small & Large	Image & Mixed Dataset (numerical and categorical)	K-Means	SNMI
EAC	Fixed	CSPA, HPGA, and MCLA	Small & Large	Image & Mixed Dataset	K-Means, Single-Link, Average-Link, Spectral Clustering	NMI
MDCE	variable	K-Means	Small	Mixed Dataset	K-Means, Mean-Link	ARI
RBSCE	Fixed	CSPA, HPGA, and MCLA	Small & Large	Image & Mixed Dataset	K-Means	Resampling Techniques
BBSCES	Fixed	CSPA, HPGA, and MCLA	Small & Large	Image & Mixed Dataset	Spectral clustering	ARI & NMI
ESFA	Fixed & variable	COMUSA, LCE, CSPA, HPGA, EAC and MCLA	Small	Image & Mixed Dataset	K-Means	ARI & NMI
KRF	Fixed	CSPA, HPGA, and MCLA	Small & Large	Mixed Dataset	K-Means	NMI
FHEC	Fixed & variable	Not Applicable	Small & Large	Text &Mixed Dataset	K-Means, K-Mediod	CPCC

III. CONCLUSION

Cluster Ensembles appeared as a new descendant for correcting the negative aspects of the individual clustering consequences. This method appeared as a well-known method to improve the accuracy, individuality, robustness and stability of unsupervised learning solutions. The incorporation procedure of the ensemble method is helpful and performances as bedrock for detecting as well as recompense the possible errors in single clustering algorithms. Therefore, this proportional study discloses some of the diverse cluster ensemble approaches with their systematic functioning process and salient features of each method along with the average accuracy also error rates of each technique. This study makes better understanding of the person who reads and hopes to be more legible and useful for the society of clustering ensemble researchers to innovate more remarkable and efficient clustering ensemble approaches. Hence, most of the cluster ensemble approach needs to improve their accuracy level; consequently, further progressing of accuracy can be an imperative research in future.

REFERENCES

[1] Sandro Vega-Pons and José Ruiz-Shulcloper, "A Survey of Clustering Ensemble Algorithms", International Journal of Pattern Recognition and Artificial Intelligence, Vol. 25, No. 3, pp.337-372, 2011.
[2] Hongjun Wang, Hanhuai Shan, Arindam Banerjee, "Bayesian Cluster Ensembles", Statistical Analysis and Data Mining, 2011.
[3] A. Strehl and J. Ghosh, "Cluster ensembles - a knowledge reuse framework for combining multiple partitions", JMLR, 3: pp.583–617, 2002.
[4] A. Topchy, A. Jain, and W. Punch, "A mixture model for clustering ensembles", In SDM, pp. 379–390, 2004.111
[5] 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.
[6] Fern X. Z, Lin W, "Cluster ensemble selection", Stat. Anal.Data Mining 1(3), pp. 128–141, 2008.
[7] Fred A.L, Jain A.K, "Combining multiple clusterings using evidence accumulation", IEEE Trans. Pattern Anal. Mach. Intell. 27 (6), pp. 835–850, 2005.
[8] Hadjitodorov S.T, Kuncheva L.I, Todorova L.P, "Moderate diversity for better cluster ensembles", Inf. Fusion 7 (3), pp.264–275, 2006.
[9] Hong Y, Kwong S, Wang H, Ren Q, "Resampling-based selective clustering ensembles", Pattern Recognit. Lett.30 (3), pp.298–305, 2009.
[10] Jia J, Xiao X, Liu B, Jiao L, "Bagging-based spectral clustering ensemble selection", Pattern Recognit. Lett.32 (10), pp.1456–1467, 2011.
[11] Mimaroglu S, Erdil E, "An efficient and scalable family of algorithms for combining clusterings", Eng. Appl. Artif.Intell.26 (10), pp.2525–2539, 2013.
[12] A.Gionis, H.Mannil, P.Tsaparas, "Clustering aggregation", In Proceedings of the 21st International Conference on Data Engineering (ICDE'05), pp. 341 – 352, 2005.
[13] X. Z. Fern and C. E. Brodley. "Solving cluster ensemble problems by bipartite graph partitioning", In Proceedings of the Twenty First International Conference on Machine Learning, pp. 281–288, 2004.
[14] A. Ng, M. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm", In Advances in Neural Information Processing Systems 14, pp

849–856, 2002.
[15] X. Hu, I. Yoo, "Cluster ensemble and its applications in gene expression analysis", in: Y.-P.P. Chen (Ed.), Proc. 2-nd Asia-Pacific Bioinformatics Conference (APB2004), Dunedin, New Zealand, pp. 297–302, 2004.
[16] L. Hubert, P. Arabie, Comparing partitions, Journal of Classification 2, pp.193–218, 1985.
[17] D. Greene, A. Tsymbal, N. Bolshakova, P. Cunningham, "Ensemble clustering in medical diagnostics", in: R. Long et al. (Eds.), Proc. 17th IEEE Symp. on Computer-Based Medical Systems CBMS_2004, Bethesda, MD, National Library of Medicine/National Institutes of Health, IEEE CS Press, pp. 576– 581, 2004.
[18] Li Zheng, Tao Li, Chris Ding, "A Framework for Hierarchical Ensemble Clustering", ACM Transactions on Knowledge Discovery from Data, Vol. 9, No. 2, Article 9, 2014.
[19] X. Z. Fern and C. E. Brodley, "Random projection for high dimensional data clustering: A cluster ensemble approach", In Proceedings of the Twentieth International Conference on Machine Learning, pp. 186– 193, 2003.
[20] D. Greene, A. Tsymbal, N. Bolshakova, P. Cunningham, "Ensemble clustering in medical diagnostics", in: R. Long et al. (Eds.), Proc. 17th IEEE Symp. on Computer-Based Medical Systems CBMS_2004, Bethesda, MD, National Library of Medicine/National Institutes of Health, IEEE CS Press, pp. 576–581, 2004.
[21] A. Strehl, J. Ghosh, "Cluster ensembles—a knowledge reuse framework for combining partitionings", in: Proc. of 11-thNational Conf. on Artificial Intelligence, NCAI, Edmonton, Alberta, Canada, pp. 93–98, 2002.
[22] A. Topchy, A.K. Jain, W. Punch, "Combining multiple weak clusterings", in: Proceedings of IEEE Int. Conf. on Data Mining, Melbourne, Australia, pp. 331–338, 2003.
[23] Dudoit S, Fridlyand J, "Bagging to improve the accuracy of a clustering procedure", Bioinformatics 19 (9), 2003.
[24] Fern X.Z, Brodley C.E, "Random projection for high dimensional data clustering: A cluster ensemble approach", In: Proc. 20th Internat.Conf. Machine Learning, vol. 20, pp. 186–191, 2003.
[25] Fischer B, Buhmann J.M, "Bagging for path-based clustering", IEEE Trans. Pattern Anal. Machine Intell.25 (11), pp. 1411–1415, 2003.
[26] Minaei-Bidgoli B, Topchy A, Punch W.F, "A comparison of resampling methods for clustering ensembles", In: Internat.Conf. on Machine Learning, Models, Technologies and Applications (MLMTA 2004), pp. 939–945, 2004a.
[27] Minaei-Bidgoli B, Topchy A, & Punch W. F, "Ensembles of partitions via data resampling", In: Proc. Internat. Conf.on Information Technology: Coding and Computing (ITCC'04), vol. 2, pp. 188–192, 2004b.
[28] Fowlkes C, Belongie S, Chung F, Malik, J,. Spectral, "Grouping using the Nyström method", IEEE Trans. Pattern Anal. Machine Intell.26 (2), pp. 214–225, 2004.
[29] Mimaroglu S, Erdil E, "Asod: arbitrary shape object detection", Engineering Applications of Artificial Intelligence 24, pp. 1295–1299, 2011a.
[30] Mimaroglu S, Erdil E, "Combining multiple clusterings using similarity graph", Pattern Recognition 44, pp. 694–703, 2011b.
[31] Iam-on N, Boongoen T, Garrett S, "LCE: a link-based cluster ensemble method for improved gene expression data analysis", Bioinformatics 26, pp. 1513–1519, 2010.
[32] Karypis G, Kumar V, "A fast and high quality multilevel scheme for partitioning irregular graphs", SIAM Journal on Scientific Computing 20, 359, 1999.
[33] Karypis G, Aggarwal R, Kumar V, Shekhar S, "Multilevel hypergraph partitioning: application in VLSI domain", In: Proceedings of the 34th Annual Conference on Design automation, ACM New York, NY, USA, pp. 526–529, 1997.
[34] Hore P, Hall L, Goldgof D, "A scalable framework for cluster ensemble", Pattern Recognition 42, pp. 676–688, 2009.
[35] Thomas M. Cover and Joy A. Thomas, "Elements of Information Theory", Wiley, 1991.
[36] J. Podan, "Simulation of random dendrograms and comparison tests: Some comment", Journal of Classification 17, pp.123–142, 2000.
[37] P.N. Tan, M. Steinbach, V. Kumar, "Introduction to Data Mining (1st ed.)", Addison-Wesley Longman, Boston, MA, 2005.
[38] J. Wu, H. Xiong, J. Chen, "Towards understanding hierarchical clustering: A data distribution perspective", Neuro computing, 72, pp. 10–12, 2319–2330, 2009.
[39] Y. Zhao and G. Karypis, "Evaluation of hierarchical clustering algorithms for document datasets", In Proceedings of the 11th International Conference on Information and Knowledge Management (CIKM'02), ACM, New York, NY, pp. 515–524, 2002.
[40] L. Zheng, T. Li, C. H. Q. Ding, "Hierarchical ensemble clustering", In ICDM'10, pp.1199–1204, 2010.
[41] J. Azimi and X. Fern, "Adaptive cluster ensemble selection", In Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI'09), pp. 992–997, 2009.
[42] E. N. Adams, "N-trees as nestings: Complexity, similarity, and consensus", Journal of Classification 3, pp. 299–317, 1986.
[43] E. N. Adams III, "Consensus techniques and the comparison of taxonomic trees", Systematic Zoology 21, 4, pp. 390–397, 1972.
[44] N. Ailon and M. Charikar, "Fitting tree metrics: Hierarchical clustering and phylogeny", In Proceedings of the Symposium on Foundations of Computer Science, pp.73–82, 2005.

Mrs. S. Sarumathi received B.E. degree in Electronics and Communication Engineering from Madras University, Madras, Tamil Nadu India in 1894 and the M.E. degree in Computer Science and Engineering from K.S. Rangasamy College of Technology, Namakkal, Tamil Nadu, India in 1407. She is doing her Ph.D. programmer under the area Data Mining in Anna University, Chennai. She has a teaching experience of about 15 years. At present, she is working as Associate professor in Information Technology department at K.S. Rangasamy College of technology. She has published 5 papers in the reputed International Journals and 2 papers in the reputed National journals. And also she has presented papers in three 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.

Dr.N. Shanthi received the B.E. degree in Computer Science and Engineering from Bharathiyar University, Coimbatore, Tamil Nadu, India in 1894 and the M.E. degree in Computer Science and Engineering from Government College of Technology, Coimbatore, Tamil Nadu, and India in 1401. She has completed the Ph.D. degree in Periyar University, Salem in offline handwritten Tamil Character recognition. She worked as a HOD in department of Information Technology, at K.S.Rangasamy College of Technology, Tamil Nadu, India since 1894 to 143, and currently working as a Professor& Dean in the department of Computer Science and Engineering at Nandha Engineering College Erode. She has published 30 papers in the reputed International journals and 9 papers in the National and International conferences. She has published 2 books. She is supervising 13 research scholars under Anna University, Chennai. She acts as the reviewer for 4 international journals. Her current research interest includes Document Analysis, Optical Character Recognition, and Pattern Recognition and Network security. 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 1413. Now she is an M.Tech student of Information Technology department in 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.