A Comparative Analysis of Different Categorical Data Clustering Ensemble Methods in Data Mining

S.Sarumathi
Associate Professor
Department of IT
K.S.Rangasamy College of Technology, Tamil Nadu, India.

N.Shanthi, Ph.D
Professor and Dean
Department of CSE
Nandha Engineering College, Erode, Tamil Nadu, India

M.Sharmila
PG Scholar
Department of IT
K.S.Rangasamy College of Technology, Tamil Nadu, India.

ABSTRACT

Over the past decades, a prevalent amount of work has been done in the data clustering research under the unsupervised learning technique in Data mining. Moreover a myriad of algorithms and methods has been proposed focusing on clustering different data types, representation of cluster models, and accuracy rates of the clusters. However no single clustering algorithm proves to be the most efficient in providing best results. Accordingly in order to find the solution to this issue a new technique, called Cluster ensemble method was bloomed. This cluster ensemble is a good alternative approach for facing the cluster analysis problem. The main aspire of the cluster ensemble is to combine different clustering solutions in such a way to achieve accuracy and to improve the quality of individual data clustering. Due to the substantial and unremitting development of the new methods in the sphere of data mining, it is obligatory to make a critical analysis of the existing techniques and the future novelty. This paper reveals the comparative study of different cluster ensemble methods along with their features, systematic working process and the average accuracy and error rates of each ensemble methods. Consequently this theoretical and comprehensive analysis will be very useful for the community of clustering practitioners and also helps in deciding the most suitable one to rectify the problem in hand.

Keywords

Cluster Ensemble methods, Co-association matrix, Consensus function, Median partition.

1. INTRODUCTION

Clustering is one of the most crucial and an underpinning process in Data Mining. It also plays an imperative role in the other fields such as Machine Learning process, Pattern Recognition, Information retrieval, Spatial Data Extraction, Image Processing and World Wide Web. Data clustering mainly concerns with how to group a set of objects based on their proximity in vector space. The main objective of the cluster analysis is finding similarities between data according to the uniqueness found in the data and grouping related data objects into clusters. An excellent clustering method produces a high superiority clusters with high intra class similarity and low inter class similarity. A large assortment of clustering algorithms which are of well established such as K-Means, EM (Expectation Maximization) based on the spectral graph theory [1], K-modes, GAClust [2], CobWeb [3]. STIRR [4], Robust Clustering Algorithm for Categorical Attributes ROCK [5], CLICK [6], Clustering Categorical Data Using Summaries CACTUS [7], COOLCAT [8], CLOPE [9], Squeezer [10], Differential fuzzy clustering, Standard Deviation of Standard deviation Roughness algorithm, Frequency of attribute value combination algorithm and some hierarchical clustering algorithms like Divisive algorithm, LIMBO [11] , single link, Fuzzy C-Means, Fuzzy C-Medoids [12] [13] [14] etc are emerged over earlier periods. Conversely it is known that there is no single clustering method is capable of providing accurate and appropriate cluster results [14]. Since by applying a clustering algorithm to the data set it works on the basis of the internal criteria i.e. similarity or dissimilarity measures used in that algorithm. At the same time if two different clustering algorithms were applied to the same data set consequently it will results in very different clusters solutions. Therefore this critical concern is very difficult to evaluate the exact clustering results. In cluster analysis the evaluation of the results are associated to the use of Cluster Validity Indexes which is used to measure the quality of clustering results [14]. Nevertheless to overcome this serious issue combining multiple clustering approaches in an ensemble framework may allow one to take advantage of the strengths of individual clustering approaches.

The general outlier of the cluster ensemble is done by achieving the solutions from the different base clustering which are then aggregated to form a final partition [13]. This Meta level approach involves these two major tasks of generating a cluster ensemble and then producing a final partition normally referred as the consensus function [15] [13]. Precisely the great challenge in clustering ensemble is the definition of most suitable consensus function which is capable of improving the consequences of single clustering algorithm. Accordingly the rest of this paper is followed with the methodical process of the different ensemble methods and concludes with the hope of that this comparative study will be very useful for the evaluation of future clustering ensemble methods.

2. CLUSTER ENSEMBLE PARADIGM

Cluster ensembles are supposed to be a robust and most perfect alternative to single clustering runs. It is the process of grouping up of multiple clustering solutions to obtain a consensus result by merging different partitions based upon well defined rules. It also provides for a visualization tool to examine cluster number, membership, and boundaries. In this sense ensemble clustering is a potential approach to generate more accurate clusters than might be possible using an individual clustering approach [15]. It generally involves two major tasks as Generation step in which generating several clustering solutions by applying clustering algorithm is done and the Consensus step through which final cluster partition is produced. The general basic construction of the cluster ensemble method was shown in Figure1 [13].

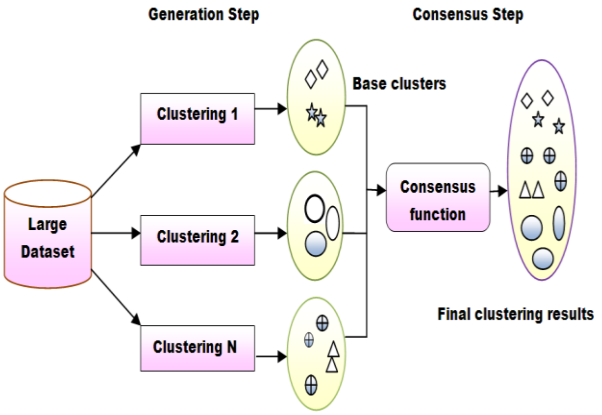

Fig 1. Basic Process of Cluster Ensembles

2.1 Generation Steps

In this generation step there are no constrains about how the partitions must be obtained. Since during the creation process [14] different clustering algorithms or the same algorithm with different parameters initialization, different object representations, and subsets of objects or projections of the objects on different subspaces can be used to produce the different base cluster solutions [14]. In spite of this process even a weak clustering algorithms are capable of producing high quality consensus clustering in concurrence with the proper consensus function.

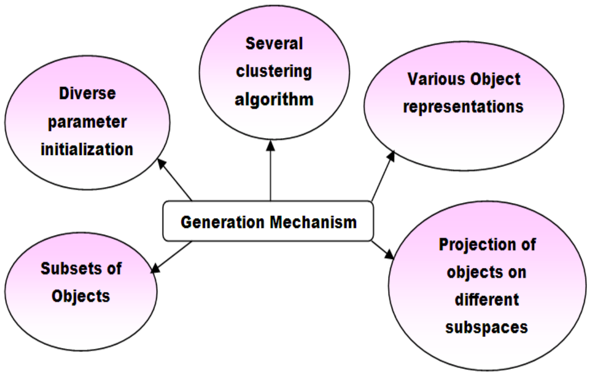
Several clustering algorithm
Diverse parameter initialization
Various Object representations
Generation Mechanism
Subsets of Objects
Projection of objects on different subspaces

Fig 2. Primary Cluster Ensemble Generation Steps

2.2 Consensus Steps

In this step consensus functions are developed and are made available for gaining the ultimate data partition from the different base clustering results. This consensus function has the large capability of improving the results of the single clustering algorithms. It involves two approaches such as object co-occurrence and median partition. In the first approach it deals with the measuring the number of occurrences of an object in a single cluster and also it analysis how many times two objects belongs together in the same cluster. In the second approach it deals with the partition that maximizes the similarity with all partitions in the cluster ensemble. The complexity of this median partition method is the improper analysis of the dissimilarity measures. Even though these approaches are evolved still there are several questions raised such as, Which clustering algorithms should be used? , Which are the correct parameters? , Which are the exact dissimilarity measures? , Which is the best heuristic approach to solve the problem or to come close to the solution? [14]. Therefore a bunch of clustering ensemble methods is projected over recent years to answer those questions.

3. DIFFERENT CLUSTER ENSEMBLE METHODS

The following sections will present the some diverse collection of cluster ensemble methods. And also for each method its systematic working process and features are elucidated.

3.1 Weighted Cluster Ensemble (WCE)

A Weighted cluster is a subset of data points together with a vector of weights such that the points in the cluster are close to each other. In this ensemble method [16] Locally Adaptive Clustering algorithm was used and it discovers clusters in subspaces spanned by different combinations of dimensions through local weightings of features. The major benefit of this Locally Adaptive clustering was that it avoids the risk of loss of information encountered in global dimensionality reduction techniques. This ensemble method consists of two approaches as follows.

3.1.1. Weighted Similarity Partitioning Algorithm (WSPA)

This method [16] starts initially by running locally adaptive clustering algorithm m times with different h values. Then for each data point x_i the weighted distance from the cluster c_{ls} is calculated by the below formula as,

$$d_{il} = \sqrt{\sum_{s=1}^{D} wls\ (x_{is} - c_{ls})^2} \quad (1)$$

where d_{il} is the larger corresponding capability credited to the cluster c_{ls} and wls is the weighted clusters. Then the probabilistic estimation for embedding the clustering result is given by,

$$P\,(c_l \,|\, x_i) = \frac{D_i - d_{il} + 1}{k\,D_i + k - \sum_l d_{il}} \quad (2)$$

After that to compute the similarity between the data points x_i and x_j both cosine similarity measure and Kullback-Leibler (KL) divergence measures were applied as given below,

$$S(x_i, x_j) = \frac{P^i_t\ P_j}{\|P_i\|\ \|P_j\|} \quad (3)$$

The above formula denotes the cosine similarity measure in which it detects the probability vectors associated to x_i *and* x_j.

Then the distance between *xi and* x_j was computed using KL divergence formula as follows,

$$d(x_i, x_j) = (1/2 \sum^{k}_{l=1} P_{il} \log 2\, P_{il} / P_{jl}) + (1/2 \sum^{k}_{l=1} P_{jl} \log 2\, P_{jl} / P_{il}) \quad (4)$$

Finally a consensus function that guides the computation of the consensus partition is define by the formula $\psi = 1/ m \sum^{m}_{l=1} S_l$. After this complete graph G = (V,E) where |V| = n and $V_i \,|||\, x_i$ was constructed. Main aim and feature of this method is to generate robust and stable cluster solutions.

3.1.2. Weighted Bipartite Partitioning Algorithm (WBPA)

This approach mainly maps the problem of finding a consensus partition to a bipartite graph partitioning problem. It overcomes the shortcomings of Weighted Similarity Partitioning Algorithm [16] in which it assigns only low similarity values to both pairs of a data set where as Weighted Bipartite Partitioning Algorithm has the ability to differentiate the two cases by modeling both instance-based and cluster-based similarities. The starting process of this approach was similar to the Weighted Similarity Partitioning algorithm. Only additional measure in this method is the formation of the matrix using the vectors of posterior probabilities. Hence based on that matrix a bipartite graph to which the consensus partition problem maps. Thus the bipartite graph was constructed with number of vertices and each represents the cluster of the ensemble.

3.2 K-Means Cluster Ensemble based on center matching scheme (KCE)

In this method center matching scheme [17] is projected for constructing a consensus function in the K-Means cluster ensemble learning. The well known K-Means algorithm has a striking characteristic feature due to its computational simplicity. Here it was chosen for the ensemble. The working process of this method starts by extracting the output sequence of K-Means cluster centers using the K-Means clustering. Then it randomly selects the cluster sequence as a reference one and rearranges the other cluster sequences according to the reference sequence. Let $C_r = \{c^1{}_{r1,} c^2{}_{r2,} c^3{}_{r3}\}$ be the reference sequence and $C_p = \{c^1{}_{p1,} c^2{}_{p2,} c^3{}_{p3}\}$ be the any cluster sequence. Then a weight matrix between the two sequences is constructed as follows

$$W^{rp} = \begin{pmatrix} 2.3 & 2.8 & 2.7 \\ 4.6 & 3.9 & 1.7 \\ 2.0 & 0.9 & 3.3 \end{pmatrix}$$

To find an efficient center matching, Hungarian algorithm is used through the formula given below,

$$\min = \sum_{i=1}^{k} \sum_{j=1}^{k} W^{rp}{}_{ij} B_{ij} \quad (5)$$

where B_{ij} denotes the indicator variables to determine the center matching between the two sequences. Labeling the data using these matched cluster sequences [17] is done. Hence it results in producing multiple partitions or clustering which do not need matching again. Finally these multiple clustering is combined to consensus clustering using some combinational rules such as voting rules [18].

3.3 Extended Evidence Accumulation Clustering Ensemble method (EEAC)

This method is highly employed to select the more robust cluster in the final ensemble. It generally selects the best performing cluster results rather than choosing all the generated cluster solutions for the ensemble. Those clusters which satisfy the stability criteria can participate in the cluster ensemble which was measured using Normalized Mutual information (NMI). A stable cluster [19] is the one that has high likelihood of reoccurrence across multiple applications of the clustering method. After applying the stability threshold to the each cluster then selected clusters are used to construct the co-association matrix. The stability of the cluster Ci is measured as given below,

$$\text{Stability (Ci)} = 1/ M \sum^{M}_{i=1} NMI_i \quad (6)$$

where M is the number of data partitions available in reference set and i denotes the i^th^ partition in that same reference set. In the next step for truly recognize the pair wise similarity a co-association matrix was computed by,

$$C(i,j) = n_{ij} / \max(n_i, n_j) \quad (7)$$

where ni and nj are the number present in remaining (after stability threshold) clusters for the i-th and j-th data points, respectively. Also, nij counts the number of remaining clusters which are shared by both data points indexed by i and j, respectively. Finally hierarchical method is applied over the generated matrix to mine the final partition. Hence the main outstanding aspects of this Extended Evidence Accumulation clustering Ensemble approach [19] is the stability measurement for each clusters and the accuracy in deciding the final partition.

3.4 Squared Error Adjacent Matrix Clustering Ensemble method (SEAM)

This new method mainly focus on how to combine the multiple data partitions to get a consistent partition for a given data set using the information obtained in the different clustering results. This Squared Error Adjacent Matrix algorithm [20] [21] is mainly based upon the similarity matrix which is defined as the co-association matrix. It has the high potential of finding the final data partition without predefining the number of clusters or any value of the thresholds when similarity matrix is given. This matrix is constructed by measuring the co-occurred times of the data pairs in the same cluster, the N data partitions of n data objects are mapped into an n x n co-association matrix which is expressed below,

$$S(i,j) = n_{ij} / N \quad (8)$$

where n_{ij} is the number of times the pair (i, j) is located in the similar cluster among the N data partitions. The value of S (i, j) represents the similarity of the data objects x_i and x_j. Thus the Squared Error Adjacent Matrix ensemble method can find the final partition of the data set over the given similarity matrix with low complexity.

3.5. Adaptive Spectral Clustering Ensemble Selection Method (ASCE)

This method can adaptively access the number of component members which is not owned by many of the ensemble methods. In this, system spectral clustering [22] [23] is used as basic learner of the ensemble system. Spectral clustering ensemble approach is based on re-sampling technique and Population Based Incremental Learning algorithm [24]. Hence this search approach is more stable and faster to solve more complex optimization problems. It mainly denotes that random variables are independent. The distribution density was computed through the product of the random variables. Updated probability measure was given below,

$$P_{l+1}(x) = \prod\nolimits^{n}_{i=1} P_{l+1}(x_j) \qquad (9)$$

However Population Based Incremental Learning algorithm is mainly used to detect the optimum clustering ensemble for its plainness and robustness. After that re-sampling the clustering set in accordance to the probability vector is done to compute the consensus partition. Finally the clustering set which posses the probability of being selected above the threshold level is picked for ensemble. The key feature of this method is that it is highly effective when the ensemble size is large.

3.6. Link based Clustering Ensemble Method (LCE)

This link based cluster ensemble method denotes the discovery of unknown values in the cluster co-association matrix [25]. The matrix analyses the pair wise-similarity between the objects and if similarity occurs it enter the value as “1” otherwise the entries are left unknown and simply record as “0”. This Link based clustering ensemble methodology [13] involves three stages as

a) Creating base clustering to form a cluster ensemble.
b) Generating the Refined cluster association Matrix RM using a link based similarity algorithm.
c) Producing final data partition by exploring special graph partitioning technique.

Refined Matrix (RM) [13] is the enhanced variation of the co-association matrix. For each clustering $\prod_t$, t =1….M and their corresponding clusters $C_1^t \ldots C^t_{kt}$ where kt is the number of clusters in the clustering $\prod_t$. The association degree RM (*xi*, cl) €[0,1] that data point xi € X has with each cluster cl € { $C_1^t \ldots C^t_{kt}$ } is estimated as follows:

$$\text{RM}(xi, cl) = \begin{cases} 1 & \text{if } cl = Ct\,{*}(xi), \\ \text{sim}(cl, Ct\,{*}(xi)), & \text{otherwise} \end{cases} \qquad (10)$$

where Ct *(*xi*) is a cluster label to which data point xi belongs. In addition, sim (Cx,Cy) € [0,1] denotes the similarity between any two clusters Cx, Cy, which can be discovered using the following link-based algorithm. The process of the link based algorithm entirely depends on the Weighted Triple Quality factor [13] in which it mainly denotes the construction of weighted graphs G = (V,W) where V represents the set of vertices denoting each cluster and W represents the set of weighted edges between the clusters. To determine the quality of the clusters it’s mandatory to find the rarity of links connected with each cluster in a network. Hence the WTQ measure of cluster Cx, Cy € V with respect to each triple C_k € V is estimated by,

$$WTQ^{k}_{xy} = \frac{1}{W_k} \qquad (11)$$

The accumulative WTQ score from all triples (1..q) between clusters Cx , Cy can be found using the below measure,

$$WTQ = \sum_{k=1}^{q} WTQ^{k}_{xy} \qquad (12)$$

Then the similarity between the clusters Cx , Cy can be estimated by,

$$\text{Sim}(Cx, Cy) = \frac{WTQ_{xy}}{WTQ_{max}} * DC \qquad (13)$$

where WTQ_{xy} is the value of any two clusters and WTQ_{max} is the maximum of WTQ_{xy} and DC € [0,1] is a constant delay factor. Finally by applying consensus function to the RM a final clustering partition can be exploited. Thus the main key feature is that it is a powerful method for decomposing an undirected graph with good performance being exhibited in diverse application areas.

3.7. Selective Spectral Clustering Ensemble Method (SELSCE)

This approach is introduced to construct the selective ensemble in order to explore the diverse and qualified final cluster partition. To generate the selective ensemble the initial step is to pick the good and efficient base clustering solution through spectral clustering technique [26] and also it produces the individual learner based on the approach given in reference to [27]. Here NMI (Normalized Mutual Information) is used to measure the diversity of the component clustering as given below,

$$\text{NMI}(\prod\nolimits_a, \prod\nolimits_b) = \frac{\sum_{h=1}^{k_a} \sum_{l=1}^{k_b} n_{hl} \log 2\, [N{*}\, n_{hl} / n_l * n_h]}{\sqrt{\left(\sum_{h=1}^{k_a} n_h \log 2 \frac{n_h}{N}\right) * \left(\sum_{l=1}^{k_b} n_l \log 2 \frac{n_l}{N}\right)}} \qquad (14)$$

where $\prod_a$, and $\prod_b$ are the two clustering then k_a and k_b are the number of clusters in $\prod_a$, and $\prod_b$ respectively. n_{hl} represents the number of instances in the hth cluster of $\prod_a$ and lth cluster of $\prod_b$ concurrently. In order to find the greater diversity between the two clustering the NMI measure was slightly changed and denoted it as Div [26].

$$\text{Div} = 1 - \text{NMI} \qquad (15)$$

However besides diversity of the cluster accuracy also an important factor to be considered. The function which takes into account both accuracy and diversity simultaneously is given below,

$$\mathrm{Sim} = -(\mathrm{Div}^* \ln\mathrm{Div}+(1-\mathrm{Div})^*\ln(1-\mathrm{Div})) \qquad (16)$$

After the above process the final selection of best cluster for ensemble is achieved by two steps such as,

a) Computing the pair-wise distance between the component clusters thereby discarding the nearest one as determined by its distance.
b) Repeated progress for the remaining clustering until all of them is either selected or discarded.

Therefore this ensemble technique achieves better performance among other traditional clustering algorithms. And an efficient feature in this method is that the computational cost of the selection process is low.

3.8 Bayesian Cluster Ensemble Method (BCE)

Bayesian cluster ensemble method was emerged for being a mixed membership model for learning cluster ensembles [28]. It basically denotes the Bayesian approach which deals with Bayes' theorem with two distinct interpretations. This Bayesian Cluster Ensemble method generates a Bayesian graph model from the base clustering solutions. From the generative model it is assumed that θ_i is sampled from Dirirchlet distribution with the parameter α and the consensus cluster h for each x_{ij} selected from θ_i separately. After this generation process in order to estimate the mixed-membership of each object to the consensus clusters Variation inference [28] is calculated as follows,

$$q(\theta_{i\,,}\, z_i \,|\, \gamma_{i\,,}\, \varphi_i\,) = q(\theta_i \,|\, \gamma_i\,) \prod_{j=1}^{m} q(z_{ij} \,|\, \varphi_{ij}\,) \qquad (17)$$

where γ_i is the Dirichlet distribution parameter and $\varphi_i = \{ \varphi_{ij}$,[j] ${}^{M}{}_{1}$ } are said to be as the discrete distribution parameters. Then Generalized Bayesian Cluster Ensemble algorithm [28] was proposed in which it deals with combining both the base clustering results and feature vectors of original data points to yield a consensus clustering. Hence the outstanding feature of this Generalized Bayesian Cluster Ensemble method is its versatile nature due to its applicability to several variants of the cluster ensemble problem including missing value cluster ensembles, row distributed and column distributed cluster ensembles.

3.9 Three Staged Cluster Ensemble Method (TSCE)

This ensemble method is mainly used for clustering the mixed data points in which the datasets contain both numerical and categorical attributes. The main aim of this technique is to find relatively high quality cluster and then to utilize an aggregation method to produce the final clustering result that minimizes the number of disagreements [29]. As the name implies this technique is composed of the following three stages of the process.

a) Building BASE clusters and this process repeats until it detects that no samples are left in the data sets.
b) Refining the Initial cluster is started by selecting the BASE of the second cluster obtained and calculates its similarity with all the samples in the first cluster.
c) Verification is done by refining the BASE cluster to focus whether the solution can be further improved or not.

However, three staged ensemble method was mainly constructed as a core modeling method and are used for generating a series of clustering results with diverse conditions for a given dataset. The systematic functioning process of this three staged ensemble method is illustrated in the below figure [29]·

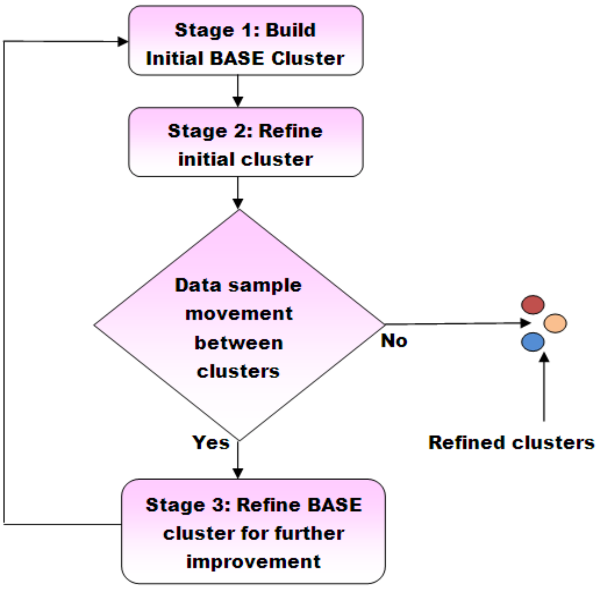
Stage 1: Build Initial BASE Cluster
Stage 2: Refine initial cluster
Data sample movement between clusters
No
Yes
Refined clusters
Stage 3: Refine BASE cluster for further improvement

Fig 3. Framework of Three Staged Ensemble Technique

3.10 Exact Method based Cluster Ensembles (EXAMCE)

This method was mainly proposed to produce the high quality ensemble solutions better than the local search methods and it also to outperform the best known technique for the Minimum Sum of Squares Clustering (MSSC) problems [31] on several benchmark data sets. Exact Method based Cluster Ensemble technique seeks to optimally recombine the partially generated solutions of different base clustering results to extract better feasible solutions to the original problem.

This process was iteratively made through local search heuristics until it finds no more further improvement can be done. The recombination step involves the search for the globally optimal solution of a restricted Set-Covering Problem [31] with a side constraint on the number of clusters in the final solution. Solving the set covering problem ($\mathrm{SCP_R}$) [32] optimally is still a NP-Hard problem but practically it can be solved quite easily. The Set covering problem contains the matrix A_B (having only q columns) that only involves the groups returned as solutions by the base clusters such as given below,

$$(SCP_R) \quad \min \sum_{x \quad i=1}^{q} c([A_B]_i)x_i$$

$$\text{s.t} \quad \begin{cases} A_B{}^x \geq e \\ \sum\limits_{i=1}^{q} x_i = k, \\ x_i € B \quad i = 1 \ldots q \end{cases} \qquad (18)$$

After this measure the duplicates are eliminated from the clusters selected by x which in turn produces a new set of clusters that are of highly feasible. Then the newly formed cluster is localized to evaluate the cost and then expanded to return the final partitioning solution. The major striking feature of this ensemble algorithm is its capability to solve the problems involving large number of clusters especially in the application area of fraud detection. It also performs well on illuminating the clustering structure as measured by the Adjacent Rand index and in other combinatorial optimization problems.

3.11 Effects of Resampling method and Adaptation on clustering ensemble efficacy

In this approach, Non-adaptive and Adaptive Resampling schemes for the integration of the multiple independent and dependent clustering solutions were proposed. In this adaptive technique [33] the individual partitions in the cluster ensembles are linearly produced by clustering specially selected subsamples of the given dataset. This adaptive scheme involves the process of Resampling, Relabeling, and finally as an upshot of the relabeling the consistency index of the cluster partitions are computed. In Non-adaptive Resampling scheme [34] [35] [36] the main goal is to obtain a reliable clustering with measurable uncertainty from a set of different *k*-means partitions. The key idea of the approach is to aggregate multiple partitions produced by clustering of pseudo-samples of a dataset. Furthermore the non-adaptive technique involves two methods such as Bootstrap in which sampling the subsets of data is done with replacement and Sub sampling method in which it deals with sampling of the data without replacement. To generate the similar labels of the clusters throughout the ensemble partitions a new technique called Relabeling is applied to each partition in the ensemble using some fixed reference partitions. The most inherent feature of this technique is the Resampling process of the original data.

3.12 Projective Clustering Ensembles Method (PCE)

In this respect, the Projective Clustering Ensembles (PCE) [37] is defined to deal with the high dimensionality and multiple clustering issues. PCE is formulized as an optimization problem and is designed to satisfy the desirable requirements on independence from the specific cluster ensemble algorithm and the skill to handle the hard and soft data clustering. These projective clusters [38] [39] [40] are mainly referred as the subsets of several input data having different subsets of features associated to them. The formal definition of the problem of projective clustering ensembles (PCE) [41] is presented here. The main aspire of this PCE is to define methods that exploit the information provided by an ensemble of projective clustering solutions (i.e., *projective ensemble*) to compute a projective consensus clustering. The information provided by any projective ensemble is two-fold which are as follows,

a) Data are grouped in clusters

b) Features assigned to clusters

After the two-fold method the techniques applied in this projective clustering approach is Multi-Objective Evolutionary algorithm [42] based Projective clustering and the Expectation Maximization based projective clustering Ensemble process. Hence the main salient features of this method are the capability of handling the high dimensionality and multi view data issues.

3.13 An Improved method for Multi-Objective Clustering Ensemble Algorithm (IMOCLE)

In this approach, Improvement of the multi-objective cluster ensemble algorithm which is expressed as IMOCLE [43] was proposed. This method mainly shows the superiority of the other techniques and the capability of finding the optimum number of clusters and accuracy. It refers to both multi-objective methods [44] and cluster ensemble techniques in optimization process. The major systematic procedure of this algorithm is as follows,

a) Initial base cluster results are obtained by applying several different clustering algorithms on the given dataset.
b) Several objective functions are optimized in the development process. This objective function can be obtained through the calculation of the similarity between the cluster partitions as follows,

$$\mathrm{Sim}(\prod_i) = 1/n \sum_{j=1}^{n} S(\prod_{i,} \prod_j) \qquad (19)$$

c) In addition to the above step special crossover [45] is applied to combine two parents using cluster ensemble technique.
d) Finally set of cluster ensembles are generated.

3.14. Generalized Adjusted Rand Index for Cluster Ensemble (ARImp)

In this approach a new method called Adjusted Rand Index [46] was proposed between similarity matrix and cluster partition to measure the consistency between the different set of clustering results and their associated consensus matrix in a cluster ensemble. ARI measure is highly defined as the adjusted form of Rand Index used mainly for the purpose of grouping the elements in the dataset.

From the mathematical point of view it is stated that this measure is related to the accuracy evaluation even if the class labels are not applicable. This measure is highly meaningful in analyzing the cluster performance without the underlying labels rather than with few similarity matrices between the partitions. The Adjusted Rand Index (ARI) measure [47] is define as follows,

$$S_0 = \sum_{i=1}^{Kp} \sum_{j=1}^{Kq} \binom{Nij}{2}, \quad S_1 = \sum_{i=1}^{Kp} \binom{Ni}{2}$$

$$S_2 = \sum_{j=1}^{Kq} \binom{N.j}{2}, \quad S_3 = \frac{2_{s1\ s2}}{N\,(N-1)}$$

$$ARI(P,Q) = \frac{S_0 - S_3}{0.5(S_1 + S_2) - S_3} \qquad (20)$$

where $P = \{P_1, P_2, \ldots P_{Kp}\}$ and $Q = \{Q_1, Q_2, \ldots Q_{Kq}\}$ be the two partitions on a data set X with N objects and the Nij are the number of objects in each cluster partitions. After finding the ARI measure in addition to preserving the desirable properties of ARI, filtering method to serve for identifying less effective cluster ensemble method was applied. This approach was experimented on the most popular UCI data sets.

3.15 Fuzzy Clustering Ensemble Algorithm for Partitioning Categorical Data (FCE)

In this approach, the fuzzy clustering ensemble algorithm [48] is proposed mainly to make use of the relationship degree between different attributes for pruning a part of the features in the data set. Pruning is highly mandatory as it prevents the surplus and unwanted attributes from reducing the efficiency of the algorithm through declining accuracy rates. The systematic process of this Fuzzy clustering ensemble algorithm was as follows,

a) By setting the initial parameters numbers of base clusters are generated.

b) Pruning the redundant attributes is done.

c) Searching for the subsets of Descartes.

d) Choosing one object from each of the subsets as initial cores.

e) Compute the membership degree of the cluster and value of the objective function.

f) Finally search for the nearest object from to the clusters from the initial core and sets the collection of cluster ensembles.

Thus the main key feature of this fuzzy clustering ensemble is to obtain the optimal number of clusters and also it establishes the relationship between the objects in the dataset under the unsupervised circumstances.

4. COMPARISON OF CLUSTER ENSEMBLE METHODS

This section exemplifies the comparison of the previously described different ensemble methods based on different parameters. The main thought of this contrast is not to examine which is the best clustering ensemble method but to differentiate the methods based on its behavioral performance and its features in which it helps the users to select the appropriate cluster ensemble method for solving their problem on hand. In below Table.1 we summarized the previously denoted ensemble methods in relate to its highlighting features and limitations of each technique which are as follows,

4.1 Ensemble Size

Ensemble is the method of cumulating the cluster partitions together in order to improve the individual clustering algorithms thereby it produces efficient results in accuracy. This Ensemble size denotes the number of clusters obtained in the ensemble through merging of the different base clustering solutions to form the final partition. This size varies in two forms as fixed size in which the cluster length is defined previously where as in variable size the ensemble size has no limitation.

4.2 Types of Consensus Function used

Consensus function comprises of two types such as Object Co-occurrence method and Median Partition method. First type deals with measuring the number of Co-occurrences of an object in a single cluster and the second type deals with the partition that maximizes the similarity with all partitions in the cluster ensemble.

4.3 Dimensionality

This property denotes the capacity of the datasets used for the experimental analysis of the ensemble methods. Capacity of the datasets are classified into small and large by analyzing through the number of data points, attributes values, classes, features and patterns occurring in the dataset.

4.4 Type of Datasets used

Datasets used for the experimental setup comprised of three types such as Numerical Datasets and Categorical Datasets and Mixed numerical & categorical datasets. First type consists of only a bunch of numerical data points, the second type involves the text data points related to the particular domain whereas the third type of datasets deals with combination of the first and second type.

4.5 Algorithm used for Base clustering

Base clustering algorithms are selected and used in each method mainly for the repeated runs of that single clustering algorithm with several sets of parameter initializations. This base clustering is mainly used for the generation of cluster ensembles. Apart from this a different clustering algorithms can also be used as a base clustering to perform heterogeneous ensemble creation.

4.6 Examined Datasets

In the previously mentioned several cluster ensemble techniques, the experimented datasets are classified into real, artificial and UCI datasets such as Iris, Zoo, Lymphography, Breast Cancer, Mushroom, 20Newsgroup, KDDCup99, WDBC, Vote, Soybean, Ionosphere, Wine, Vehicle, Glass, Bupa, Yeast, E.Coli, Segmentation, Waveform, Ionosphere, Liver disorder, LON,Star/galaxy, Three Gaussian, Yellow-small, Lung, heart, Sonar, Isolet, SatImage and Credit Approval.

Table1. Summarized Cluster Ensemble Methods

Clustering Ensemble Methods	Ensemble Size	Type of Consensus Function used	Dimensionality (size of the dimensions used in the datasets)	Type of Dataset used	Base Clustering Algorithm	Salient Features	Average Accuracy Rates
WSPA	Fixed	Object Co-occurrence	Small and Large	Categorical	Locally Adaptive Clustering	Generation of Robust and Stable Clusters	0.726
KCE	Variable	Median Partition	Small	Categorical	K-Means	Computational Simplicity	0.715
EEAC	Fixed	Object Co-occurrence	Large	Categorical	K-Means	Higher Stability and accuracy in clusters	0.690
SEAM	Fixed	Object Co-occurrence	Small	Categorical	K-Means	Low Complexity	0.850
ASCE	Variable	Median Partition	Small and Large	Categorical	Spectral Clustering	Effective for Complex optimization problems	0.721
LCE	Fixed	Object Co-occurrence	Small and Large	Categorical	K-Modes	Efficient in discovery of unknown values in Cluster matrix	0.873
SELSCE	Variable	Object Co-occurrence	Small	Categorical	Spectral Clustering	Computational cost of Selection process is low	0.742
BCE	Fixed	Object Co-occurrence	Small and Large	Categorical	K-Means	Versatile Nature due to its applicability	0.675
TSCE	Variable	Object Co-occurrence	Small	Mixed numerical and categorical	K-Means	Spotting most likely number of Clusters automatically.	0.893
EXAMCE	Variable	Object Co-occurrence	Small and Large	Categorical	K-Means	Efficient clustering in the area of fraud detection system	0.596
RMACE	Fixed	Object Co-occurrence	Small	Categorical	K-Means	Resampling of the original data	0.680
PCE	Variable	Median Partition	Small and Large	Categorical	Projective Clustering	Handling high dimensionality and multi view data issues	0.641
IMOCLE	Fixed	Object Co-occurrence	Small and Large	Categorical	K-Means	Capability of finding optimal number of clusters	0.645
ARImp	Variable	Object Co-occurrence	Small	Categorical	K-Means	Expression of consistency between the clusters.	0.432
FCE	Fixed	Median Partition	Small	Categorical	K-Means	Maintains Relationships between objects in datasets	0.635

5. CONCLUSION

Cluster Ensembles have been came into sight as a recent offspring for rectifying the negative aspects of the individual clustering consequences. This technique was mainly emerged as a high-flying method to enhance the stability, robustness, individuality, and accuracy of unsupervised learning solutions. It involves grouping up of multiple clustering solutions to obtain a consensus result by merging different partitions based upon well defined rules. This integration process of the ensemble method is really helpful and acts as bedrock for detecting and compensating the possible errors in single clustering algorithms. Consequently this proportional study reveals some of the different categorical cluster ensemble approaches including their systematic functioning process and salient features of each method along with the average accuracy and error rates of each technique. Hence the original contribution of this paper is the methodical work flow of each techniques and the comparative table denotes differential analysis, characteristics, and limitations of the diverse ensemble methods along with the graphical representation of the accuracy levels of different ensemble methods. Here in this review we compared clustering accuracy and error rates on different datasets of the each ensemble methods. The comparison result proves that the many of the proposed works in cluster ensemble technique faces accuracy problem on different real world and artificial datasets. This investigation makes better understanding for the readers and also hopes to be more legible and useful for the society of clustering researchers to innovate more remarkable and efficient clustering ensemble methods. And hence most of the ensemble approach needs to improve their accuracy level therefore further progressing of accuracy can be an imperative research in future.

REFERENCES

[1] Sandro Vega-pons & Jose reuiz Shulcloper. “A Survey of Clustering Ensemble algorithms.” International Journal of Pattern Recognition and Artificial Intelligence Vol. 25, No. 3 337_372 , 2011.

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

[3] Fisher.D.H .” Knowledge Acquisition via Incremental Conceptual Clustering. Machine Learning,” vol. 2, pp. 139-172, 1987.

[4] Gibson.D, Klein.J & Raghavan.R, “Clustering Categorical Data: An Approach Based on Dynamical Systems.” Very Large Data Base Endowment Journal .vol. 8, nos. 3-4, pp. 222-236, 2000

[5] Guha.S, Rastogi.R, & Shim.K,. “ROCK: A Robust Clustering Algorithm for Categorical Attributes.” Information Systems, vol. 25, no. 5, pp. 345-366, 2000

[6] Zaki.M.J & Peters.M. Clicks:” Mining Subspace Clusters in Categorical Data via Kpartite Maximal Cliques”. Proc. International Conference on Data Engineering (ICDE), pp. 355-356, 2005.

[7] Ganti.V, Gehrke.J, & Ramakrishnan.R “CACTUS: Clustering Categorical Data Using Summaries.” Proc. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), pp. 73-83, 1999.

[8] Barbara.D, Li.Y, & Couto.J “COOLCAT: An Entropy-Based Algorithm for Categorical Clustering.” Proc. International Conference on Information and Knowledge Management pp. 582-589, 2002.

[9] Yang.Y, Guan.S, & You.J. “CLOPE: A Fast and Effective Clustering Algorithm for Transactional Data.” Proc. ACM SIGKDD Int’l Conf. Knowledge Discovery and Data Mining (KDD), pp. 682- 687, 2002.

[10] He.Z, Xu.X, & S. Deng. Squeezer: “An Efficient Algorithm for Clustering Categorical Data.” J. Computer Science and Technology vol. 17, no. 5, pp. 611-624, 2002.

[11] Andritsos.P & Tzerpos.V. “Information Theoretic Software Clustering. “ IEEE Transactions on Software Engineering., Vol. 31, no. 2, pp. 150-165, 2005.

[12] Indrajit Saha, Ujjwal Maulik, & Nilanjan. “Differential Fuzzy Clustering for Categorical Data.” International Conference on Methods and Models in Computer Science, 2009.

[13] Natthakan Iam-On, Tossapon Boongoen, Simon Garrett, & Chris Price. “A Link based cluster ensemble approach for categorical data clustering.“ IEEE Transactions on knowledge and data engineering, Vol. 24, No. 3, 2012.

[14] Sandro Vega-pons & Jose reuiz Shulcloper. “A Survey of Clustering Ensemble algorithms. “International Journal of Pattern Recognition and Artificial Intelligence Vol. 25, No. 3 (2011) 337_372.

[15] Harun Pirim, Dilip Gautam, Tanmay , Bhowmik, Andy D. Perkins, Burak Ekşioglu, & Ahmet Alkan, “ Performance of an ensemble clustering algorithm on biological datasets”. Mathematical and Computational Applications, Vol. 16, No. 1, pp. 87-96. 2011

[16] Domeniconi.C & Al-Razgan.M, “ Weighted cluster ensembles: methods and analysis.”ACM Transaction on. Knowledge Discovery Data 2(4) 1_40. 2009

[17] Li Zhang*a, Weida Zhoua, Caili Wua, Jieting Huoa, Haishuang Zoua, & Licheng Jiaoa. “Center matching scheme for K-means cluster ensembles. “ MIPPR Pattern Recognition and Computer Vision, edited by Mingyue Ding, Bir Bhanu, Friedrich M. Wahl, Jonathan Roberts, Proc. of SPIE Vol. 7496, 749614 SPIE. 2009

[18] Weingessel, A, Dimitriadou, E., & Hornik, K. “An ensemblemethodforclustering.”Workingpaperhttp://www.Ci.tuwien.ac.at/conferences/DSC-2003, 51. 2003

[19] Hamid Parvin, Hamid Alinejad-Rokny, & Sajad Parvin. ” A New Clustering Ensemble Framework.” International Journal of Learning Management Systems, J. Learn. Man. Sys. 1, No. 1, 19-25. 2013

[20] Yang Lili, Yu Jian, & JIA Caiyan. “A New method for Cluster Ensembles”, Programs Foundation of Ministry of Education of China. 2013.

[21] Yu J. & Lin Z C. “ Squared error adjacency matrix clustering.” Technical report on Dept. of Computer Science, Beijing Jiaotong University 2008.

[22] Fowlkes C, Belongie S, & Chung F, et al..” Spectral grouping using the Nyström method.” IEEE Transactions on Geoscience and Remote Sensing (2): 214-225 2004.

[23] Ng A, Jordan M, & Weiss Y. “On spectral clustering: Analysis and an algorithm[C].” Advances in Neural Information Processing Systems (NIPS). Boston: MIT Press, 849-857. 2002

[24] XU Yuanchun, JIA Jianhua**.** “Adaptive Spectral Clustering Ensemble Selection via Re-sampling and Population Based Incremental Learning Algorithm.” Journal of Natural Sciences, Vol.16 No.3, 228-236 2011

[25] Al-Razgan.M, Domeniconi.R, & Barbara.D. “Random Subspace Ensembles for Clustering Categorical Data. Supervised and Unsupervised Ensemble Methods and Their Applications,” pp. 31-48, Springer. 2008.

[26] Jianhua Jia, Xuan Xiao, & Binxiang Liu, “Similarity-based Spectral Clustering Ensemble Selection.” 9th IEEE International Conference on Fuzzy Systems and Knowledge Discovery.2012

[27] Zhang.X.R, JiaoL.C, & Liu.F et.al. “Spectral clustering ensemble applied to SAR image segmentation.” IEEE Transactions on Geoscience and Remote Sensing, 46 (7)2126-2136 2008

[28] Hongjun Wang, Hanhuai Shan & Arindam Banerjee. “Bayesian Cluster Ensembles.” Wiley Periodicals, Inc. 2011

[29] Jamil Al-Shaqsi & Wenjia Wang, “A Clustering Ensemble Method for Clustering Mixed Data.” IEEE International conference 978-1-4244-8126-2/10/$26.00. 2010

[30] Al Shaqsi J. & Wang W. “A Novel Three Staged Clustering Algorithm. AIDES European Conference on Data Mining,” A. P. Abraham, Ed. Ed. Algarve, Portugal, pp. 19-26 2009.

[31] Ioannis T. Christou, Member IEEE “ Coordination of Cluster Ensembles via Exact Methods. “ IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 33, No. 2.2010

[32] O. du Merle, P. Hansen, B. Jaumard, and N. Mladenovich. “An Interior Point Algorithm for Minimum Sum of Squares Clustering.” SIAM J. Scientific Computing, vol. 21, no. 4, pp. 1484-1505, Mar. 2000.

[33] Topchy A, Jain AK, Punch WF “A mixture model for clustering ensembles.” In: Proceedings of SIAM international conference on data mining, SDM 04, pp 379–390 2004

[34] Fred ALN, Jain AK “Combining multiple clustering using evidence accumulation.” IEEE Trans Pattern Anal Mach Intell 27(6)2005

[35] Strehl A, Ghosh J “Cluster ensembles-a knowledge reuse framework for combining multiple partitions.” J Mach Learn Res 3:583–617 2003

[36] Topchy A, Jain AK, Punch WF “Combining multiple weak clusterings.” In: Proceedings of 3rd IEEE international conference on data mining, pp 331–338 2003

[37] Gullo F, Domeniconi C, Tagarelli A “Projective clustering ensembles.” In: Proceedings of the international conference on data mining (ICDM), pp 794–799 2009

[38] Ka Ka Ng E, Wai-Chee Fu A, Chi-Wing Wong R “Projective clustering by histograms.” IEEE Trans Knowl Data Eng (TKDE) 17(3):369–383 2005

[39] Yiu ML, Mamoulis N “Iterative projected clustering by subspace mining. “ IEEE Trans Knowl Data Eng (TKDE) 17(2):176–189 2005

[40] Achtert E, Böhm C, Kriegel H-P, Kröger P, Müller-Gorman I, Zimek A ” Finding hierarchies of subspace clusters.” In: Proceedings of the European conference on principles and practice of knowledge discovery in databases (PKDD), pp 446–453 2006

[41] Domeniconi C, Gunopulos D,MaS,YanB,Al-Razgan M, PapadopoulosD ”Locally adaptive metrics for clustering high dimensional data.” Data Min Knowl Disc 14(1):63–972007

[42] Deb K “Multi-objective optimization using evolutionary algorithms”. Wiley, New York. 2001

[43] Ruochen Liu, Member, IEEE, Yong Liu, Yangyang Li，Member, IEEE, “An Improved Method for Multi-Objective clustering Ensemble Algorithm.” IEEE World Congress on Computational Intelligence June, 10-15, 2012 - Brisbane, Australia 2012

[44] A. Strehl, J. Ghosh, “Cluster ensembles-a knowledge reuse framework for combining multiple partitions,” Journal of Machine Learning Research 3 (2002) 583–618. 2002

[45] K. Faceli, A. Carvalho, M. de Souto.” Multi-objective clustering ensemble for gene expression data analysis,” Neurocomputing 72(2009)2753-2774.

[46] Shaohong Zhang, Hau-San Wong, “ARImp A Generalized Adjusted Rand Index for Cluster Ensembles.” International Conference on Pattern Recognition, IEEE Computer Society. 2010

[47] L. Hubert and P. Arabie.” Comparing partitions.” Journal of Classification, 2:193–218, 1985.

[48] Taoying Li, Yan Chen “Fuzzy Clustering Ensemble Algorithm for Partitioning Categorical Data.” International Conference on Business Intelligence and Financial Engineering IEEE Computer Society. 2009