Linked Spectral Graph based Cluster Ensemble Approach using Weighted Spectral Quality Algorithm for Medical Data Clustering

[1]Mrs.S.SARUMATHI, [2]Dr.N.SHANTHI, [3]Ms.M.SHARMILA
[1]Associate Professor, [2]Professor and Dean, [3]Assistant Professor
[1,3]Department of IT, [2]Department of CSE
[1]K.S.R College of Technology, Namakkal, Tamil Nadu, India.
[2]Nandha Engineering College, Erode, Tamil Nadu, India.
[3]M.Kumarasamy College of Engineering, Karur, Tamil Nadu, India.
rishi_saru20@rediffmail.com, shanthimoorthi@yahoo.com, sharmi28.it@gmail.com

Abstract: - Over the certain span of time, Cluster Ensembles have been emerged as an offspring for solving the problem of extracting the efficient clustering results. Although many efforts have been commenced, it is examined that these techniques adversely creates the final data partition based on imperfect information. The original Ensemble information matrix exposes only the cluster data object relations with many entries being left empty. This paper presents an investigation that provides a solution to the problem of degrading the quality of the final partition through a Linked Spectral Graph based Cluster Ensemble approach. In particular, an effective Weighted Spectral Quality algorithm is proposed for the underlying similarity measurement among the Ensemble Members which in turn can be highly used to avoid the local optimum and the ill-posed issues derived from the huge dimensional samples. Subsequently, to obtain the final ultimate clustering results a Spectral Clustering based Consensus Function is applied to the Distilled Similarity Matrix (DSM) that is formulated from the similarity assessment algorithm. The Experimental results projected on Medical datasets retrieved from the UCI repository demonstrate that the proposed approach outperforms the traditional ones in data clustering.

Key-Words: - Clustering algorithms, Cluster Ensemble, Spectral graph partitioning, Consensus Function, Data Mining, Similarity Measures.

1 Introduction

Recently Cluster Ensemble learning has been a blooming concern and also regarded as the most outstanding paradigm in several domains such as data mining, information retrieval, machine learning, image processing, World Wide Web, Pattern Recognition etc, Many individual clustering algorithms produces irrelevant clusters thereby occupies computer memory space which results in increasing the time complexity and execution time of the algorithm. Thus Cluster Ensemble mainly inspired by the design of classifier ensemble encountered in the supervised learning technique has appeared as a new approach for rectifying the problems associated with the solutions of the individual clustering algorithms. The basic scheme of Cluster Ensemble is the process of aggregating the results of number of possibly single clustering algorithms to produce the final clustering of the dataset which is far better than the individual clustering alone can extracts. It can also provide more robust and stable quality of the clustering results through the use of the consensus functions across multiple cluster solutions. Moreover, several well equipped clustering algorithms such as K-Means [1] and PAM [2] have been specifically designed for handling the numerical data whose logics are mainly intended for measuring the distance between feature vectors [3], [4]. Hence these cannot be directly applied for the clustering of categorical type of data in various domains. Apart from the above a large variety of clustering algorithms such as EM (Expectation Maximization) based on the spectral graph theory [5], K-modes [6], GAClust [7], CobWeb [8], STIRR [9], Robust Clustering Algorithm for Categorical Attributes ROCK [10,68], CLICK [11], Clustering Categorical Data Using Summaries CACTUS [12], COOLCAT [13], CLOPE [14], Squeezer [15], 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 [16] , single link, Fuzzy C-Means, Fuzzy C-Medoids [17],[18] etc are emerged over earlier periods. Subsequently it is proved that no single clustering algorithm appears to be the best in extracting the exact and accurate clustering results.

In cluster analysis the evaluation of the results are associated to the use of Cluster Validity Indices which is used to measure the quality of clustering results [18]. 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. This cluster ensemble approach involves two major tasks as generation phase and the consensus phase. In addition to clustering Medical Data, it is investigated herein that the anticipated framework is highly generic such that it can also be applied to other types of data.

The remaining part of this paper is framed as follows, Section 2 presents the general scheme of the cluster ensemble methodology in which it includes generation methods and consensus functions upon which this approach has been established. Section 3 exposes the Related Work and the Problem Formulation. Section 4 introduces the proposed Linked Spectral Graph based Cluster Ensemble approach including its working paradigm and the process of Weighted Spectral Quality algorithm for similarity assessment. Section 5 reveals the performance evaluation of this new technique by using several validity indices over the Medical datasets collected from UCI repositories. This paper is concluded in Section 6 along with the implication for future work in which it enhances the quality of this approach thereby further decreasing the execution time and improving its efficiency.

2 Cluster Ensemble Methodology

Clustering Ensemble was mainly proposed to overcome the lack of cluster quality resulted from the individual clustering [19] algorithms. This eminence leads to the emergence of several cluster ensemble techniques over the past decades. The cluster ensemble paradigm comprises of two main aspects, first phase is to produce the several clustering membership and the second phase is to merge the clusters into a global design of ultimate partition. The general process of the cluster ensemble methodology [20] is shown in the above Fig.1.

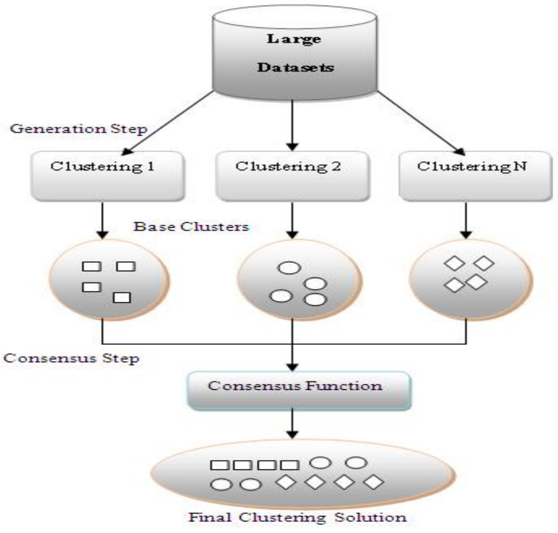


Fig.1 General Process of Cluster Ensemble

2.1 Clustering Membership Generation

The method of clustering ensemble process initiates by generating diverse population of the clustering partitions through several generative mechanisms. Ensembles are more efficient, when assembled from a set of forecaster whose errors are dissimilar [21]. To a massive extent, diversity among the ensemble methods will enhances the result of cluster ensemble. In particular the results obtained from clustering the dataset using any single clustering algorithm over much iteration are usually similar to each other. This condition leads all the ensemble members to concur with the process of partitioning the dataset. As a result several approaches have been projected to introduce the synthetic instability in clustering algorithms, which paves the way for multiplicity within the cluster ensemble. The following ensemble generation techniques give up different methods of producing the base clustering results.

1) *Homogeneous Ensembles*: Base clustering solutions are achieved through the repeated usage of the single clustering algorithm with various different parameter declarations such as cluster centre point of the K-Means algorithm [22].
2) *Heterogeneous Ensembles*: Number of different clustering algorithms is used mutually to generate the base clustering results [23], [24].
3) *Fixed–K*: This technique creates the fixed number of clusters (k) for each ensemble member by using K-Means clustering algorithm [25].
4) *Random–K*: This technique creates the randomized number of clusters (k) for each ensemble member by using K-means clustering algorithm [26], [27].
5)*Projection of Data on different Subspace/Sampling:* This ensemble approach can be achieved by producing base clusters from different object representations or subsets of objects [28-30] of the initial dataset. It can also be obtained from different subspaces, features and data sampling [31-33].
6) *Mixed Heuristics*: This technique [34], [35] results in using the any combination of the aforesaid techniques to generate the base clusters.

2.2 Consensus Function

A Consensus Function is said to be the mutual function in which it merges the results of the several base clustering memberships. Each consensus technique exploits the specific form of information matrix in which it précis the base clustering results. However after producing the cluster ensemble a variety of consensus functions have been emerged over the past decades to finalize the ultimate partition. The consensus functions can be categorized as follows,

1) Feature based Approach: It deals with the cluster label generated as an outcome of each base clustering such that it was considered as a new feature describing each data point in which it is used to originate the vital cluster solution [36], [37].

2) Pairwise Similarity Approach: This approach [38] generates a matrix containing similitude measures among the paired data points through which any similarity based base clustering algorithm can be applied.

3) Graph-based Approach: Graphical representation of similarity measures of the data points is created from a Pairwise matrix. To achieve the final clustering result the graph is partitioned into finite number of estimated equal sized partitions using METIS [39] or Spectral graph partitioning technique [40].

4) Meta Clustering Approach: This technique [39] initially solves the cluster correspondence problem by grouping the clusters indentified in the individual clustering solutions. After that it uses the voting method to set the data points into final consensus cluster results.

5) Hyper-graph Partitioning Approach: In this approach [41] the formulation of the cluster ensemble problem is done as partitioning the hyper-graph by dividing the minimal number of hyper edges.

6) Cluster based Similarity Partitioning Approach: Here the similarity between the data objects are directly proportional to the number of ingredient clustering [34] of the ensemble in which they are aggregated together. The more similar data points are credited with higher chance to be placed in the same cluster. The computational and storage complexity of this method is quadratic in nature.

7) Direct Approach: This method [42] is based on the relabeling and searching for the final partition that has been the best match for all ensemble members. It generates the unique set of decision labels from the heterogeneous clustering decisions.

3 Related Work and Problem Formulation

There have been a lot of efforts and investigations evolved over the recent years in Cluster Ensemble approaches. Contrasted with the performance of single clustering algorithms Cluster Ensemble techniques have their superior capability to assimilate multiple clustering solutions, which in turn provides more accurate, robust and stable cluster results. As illustrated in the above section Cluster Ensemble approaches involves two major tasks such as ensemble generation and the consensus function. The main aspire of the first task is to promote the results of diverse clustering solutions in the ensemble, and the second task focuses on the consensus fusion of the individual clustering solutions in which it can further enhances the accuracy and constancy of the ultimate cluster results.

Clustering Ensemble approaches are also referred to as the Consensus clustering approaches in which it mainly gains more and more consideration due to its diverse applications in the areas of data mining, machine learning, pattern recognition, bioinformatics, information retrieval, image processing and analyzing, statistical data analysis. Clustering Ensemble techniques have the powerful ability to achieve the aggregation of the several partitions from different data sources and thus it improves the stability, compactness, and robustness of the traditional single clustering algorithms.

There are a several collection of different Clustering Ensemble approaches which are of designed on the emphasis of cluster data point and cluster to cluster similarity relationships. For example, the prediction based resampling cluster ensemble [43], [44] approach which was named as Clest to extract class discovery from the microarray cancer datasets. Another Consensus clustering approach based on the random subspace technique [45] and a cluster solidity measures to detect the number of clusters in the cancer dataset. Self-Organizing Map (SOM) based Clustering Ensemble approach [46] and the Hierarchical Clustering based Ensemble technique extracts the final clustering results through the process of cluster data point relations. In addition to the above cluster ensemble, Graph based Consensus Clustering algorithm [47] mainly discovers the cancer data subtypes from the genetic profiles.

Still there are also a number of tremendous collections on the existing Clustering Ensemble approaches. For example in the Hybrid Fuzzy Cluster Ensemble [48], the fuzzy theory is implemented into the cluster ensemble paradigm in order to accurately denote the samples corresponding to different types of cancer data. Knowledge based Cluster Ensemble technique [49] mainly integrates the prior knowledge of the information in the dataset into the cluster ensemble process. In particular the prior knowledge about the data is done by the Pairwise constrains in which it helps in enhancing the quality and the accuracy of the clustering results. Then Weighted Cluster Ensemble approach [19] denotes the weighted cluster in which it is a subset of data points together with a vector of weights such that the points in the cluster are close to each other. Here the similarity measures are evaluated on the basis of the data point's compactness in the datasets.

Apart from the above, there are also a number of excellent surveys on existing Cluster Ensemble techniques. Some of them are, Squared Error Adjacent Matrix Clustering Ensemble [50] in which it exposes the similarity matrix by considering the co-association matrix generated from the data points in multiple partitions. Bayesian Cluster Ensemble [51] approach mainly deals with the Bayesian theorem with two distinct interpretations. Projective Cluster Ensemble [52], [53] denotes the process of generation of the ultimate cluster results based on the subsets of several input objects having different subsets of features associated to them.

The proposed Linked Spectral Graph based Cluster Ensemble approach outperforms the above stated several traditional clustering ensemble methods. In the existing Cluster Ensemble techniques [20] the similarity matrix measures are assessed between the base clusters formed in each partition of the Ensemble. The traditional Ensemble data matrix expresses only the cluster to cluster associations [20] while it completely disregards those similarities among the Ensemble members in the Ensemble partition. It was observed that cluster to cluster similarity estimation produces more time consumption and not so appropriate to evaluate the quality of the final cluster solutions. As a result, the performance of the conventional cluster ensemble techniques may gradually degrades the computational efficiency.

In this paper, the Linked Spectral Graph based approach mainly applies the Spectral Graph partitioning technique [40], [54] to solve graph partitioning issues for large scale undirected graph formed from the Cluster Ensemble using the method of normalized-cuts [55]. After partitioning the Ensemble formation, the similarities between the Ensemble Members are evaluated rather than the base clusters through the process of Weighted Spectral Quality similarity assessment algorithm. Finally the ultimate cluster partition was extracted through the Spectral Clustering based Consensus Function. Hence this proposed Linked Spectral Graph based Cluster Ensemble framework mainly focuses on the improvement of the quality of existing Cluster Ensemble methods.

4 A Novel Linked Spectral Graph Based Cluster Ensemble Approach

Several existing cluster ensemble methods for clustering the Medical data analysis rely on the traditional methods and binary cluster association matrices [56], [57] which précis underlying ensemble information to a certain extent. Many matrix entries are left empty or simply recorded as "0". Despite the consensus function, the quality of the final clustering result may be degraded. As a result Linked Spectral Graph based Cluster Ensemble (LSGCE) approach has been established with the capability to discover unknown values and thus it improves the accuracy rates of the ultimate data partition. In spite of promising findings the initial framework is mainly based on the data objects Pairwise similarity matrix, which is largely exclusive to obtain. The traditional link-based similarity methods, SimRank [58], Approximate SimRank [59], are employed to evaluate the similarity among data points in which it is inappropriate to a several Medical datasets.

To conquer those above stated challenging factors, a new Linked Spectral Graph based Cluster Ensemble approach is introduced herein. It is more effective than the earlier methods, where a BM [20], [56] like matrix is largely used to formulate the ensemble information. Here, in this novel approach the focus has reallocated from revealing the similarity among the data points to estimating those among the Ensemble members of the Cluster Ensemble partition. A new Weighted Spectral Quality (WSQL) similarity assessment algorithm has been purposely established to generate the similarity measures in higher optimal and inexpensive manner.

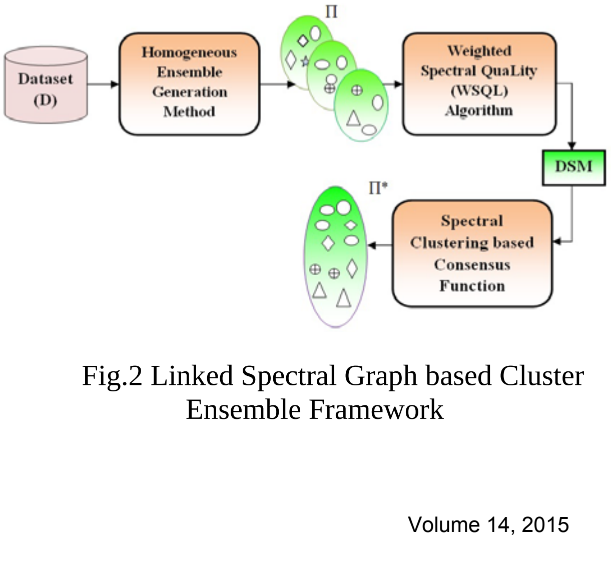

Dataset (D)
Homogeneous Ensemble Generation Method
Π
Weighted Spectral QuaLity (WSQL) Algorithm
DSM
Π*
Spectral Clustering based Consensus Function

Fig.2 Linked Spectral Graph based Cluster Ensemble Framework

The framework of LSGCE methodology is illustrated in the above Fig.2. It includes three major processes which are as follows,

1) Generation of base clustering to form Cluster Ensemble (π).
2) Producing Distilled Similarity Matrix (DSM) using Weighted Spectral Quality algorithm.
3) Extracting the ultimate data partition (π) by exploiting the Spectral Clustering based Consensus Function.*

4.1 Creating Cluster Ensemble

Consider the Dataset $X = \{x_1, \dots, x_n\}$ be a set of data points and π denotes the cluster ensemble such that $\pi = \{\pi_1, \dots, \pi_M\}$ are the ensemble members with base clustering. Each base clustering profits a set of clusters $\pi_i = \{C_1^i, C_2^i, \dots, C_k^i\}$ where as k_i is number of clusters in the clustering results. The following Fig.3 illustrates the Sample Cluster Ensemble and its corresponding clusters

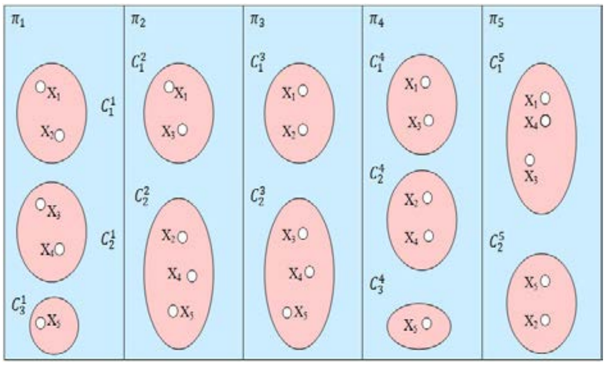
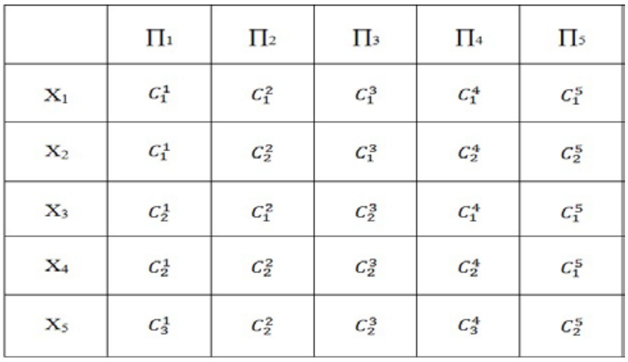


Fig.3 Sample Cluster Ensemble

Here, in this LSGCE approach Homogeneous Cluster ensemble generation method is used in which running Spectral Clustering algorithm n times obtain the base clustering results. In particular to a full space ensemble, the generations of the base clustering are extracted from the original dataset with all its instances and attributes. The systematic process of the Spectral Clustering algorithm are summarized as follows,

```
SPECTRAL CLUSTERING ALGORITHM
Input: A set of points $X = \{x_1, \dots, x_n\}$ in $R^k$.
Output: The Resulting set of Base clusters.
1) begin
2) Compute Affinity Matrix $A \in R^{n*n}$ defined by
   $A_{ij} = \frac{\exp(-\|x_i - x_j\|^2)}{2\sigma^2}$ If $i \neq j, A_{ij} = 0$
3) Define D be the Diagonal Matrix, and Build the
   Laplacian Matrix $L = D^{-1/2} * AD^{-1/2}$
4) Determine $e_1, e_2, \dots e_k$ such that k be largest
   eigenvectors of Matrix L.
5) Build the Matrix $E = [e_1, e_2, \dots e_k] \in R^{n*k}$
6) Build the Matrix B from E by stabilize each row
   to have a unit length $B_{ij} = \frac{E_{ij}}{(\sum_j E^2{}_{ij})^{1/2}}$
7) Apply K-Means clustering technique over each
   row of B as a point in $R^k$ to cluster them into k
   clusters.
8) end
```

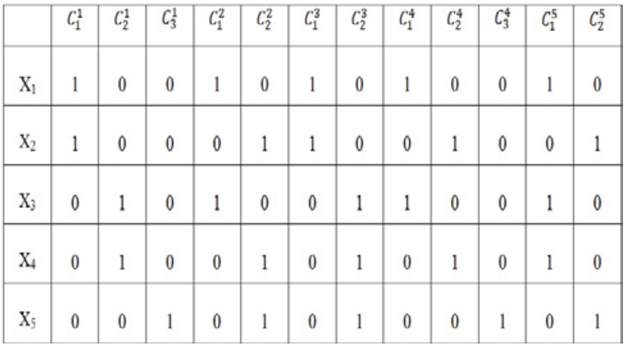
Having obtained the set of base clusters formed from the repeated runs of the Spectral Clustering algorithm, the Cluster Ensemble is employed. From the Sample Cluster Ensemble shown in the Fig.3, Label assignment Matrix of size was created as illustrated in the Fig.4. It specifically symbolizes the cluster labels that are assigned to each data points by different base clustering.



	Π₁	Π₂	Π₃	Π₄	Π₅
X₁	C_1^1	C_1^2	C_1^3	C_1^4	C_1^5
X₂	C_1^1	C_2^2	C_1^3	C_2^4	C_2^5
X₃	C_2^1	C_1^2	C_2^3	C_1^4	C_1^5
X₄	C_2^1	C_2^2	C_2^3	C_2^4	C_1^5
X₅	C_3^1	C_2^2	C_2^3	C_3^4	C_2^5

Fig.4 Label Assignment Matrix

Moreover the Binary Cluster Association Matrix [20] illustrated in Fig.5 exposes the cluster specific nature of the original label assignment matrix. Each entry in this matrix mainly denotes the crispy association degree between the data points and the clusters formed in the ensemble. The co-association degree is based on the occurrence of the data points in the extracted clusters. It fills the matrix entry by either "1" or "0" such that if the particular data point



	C_1^1	C_2^1	C_3^1	C_1^2	C_2^2	C_1^3	C_2^3	C_1^4	C_2^4	C_3^4	C_1^5	C_2^5
X₁	1	0	0	1	0	1	0	1	0	0	1	0
X₂	1	0	0	0	1	1	0	0	1	0	0	1
X₃	0	1	0	1	0	0	1	1	0	0	1	0
X₄	0	1	0	0	1	0	1	0	1	0	1	0
X₅	0	0	1	0	1	0	1	0	0	1	0	1

Fig.5 Binary Cluster Association Matrix

is occurred then its corresponding entry will be recorded as “1” otherwise “0”. It is clearly inspected that Binary Association Matrix [56] is commonly sparse, with large number of entries being filled as “0”. Spontaneously this particular feature that is commonly encountered with ensemble solutions of the hard clustering results may degrade or limit the quality of the ultimate data partition produced by the consensus function. In order to rectify this problem, few methods are proposed to obtain the Distilled Similarity Matrix.

4.2 Weighted Spectral Quality (WSQL) Algorithm

Spectral Clustering techniques mainly denote the usage of the spectrum (Eigen values) of the association matrix of the original data to perform dimensionality reduction before clustering in fewer dimensions. Basically Spectral Clustering treats the data clustering as a Spectral graph partitioning problem without assigning any assumption on the form of the data clusters.

Consider the given Sample Cluster Ensemble π with the set of data points $X = \{x_1, \dots, x_n\}$ a Linked Spectral Graph $LSG = (V, W)$ can be constructed, where V denotes the set of vertices representing the link between the clusters and W be the weight of the data points associated with every two clusters in the sample cluster ensemble. The following Fig.6 illustrates the Linked Spectral Graph of the cluster network obtained in the Ensemble partition.

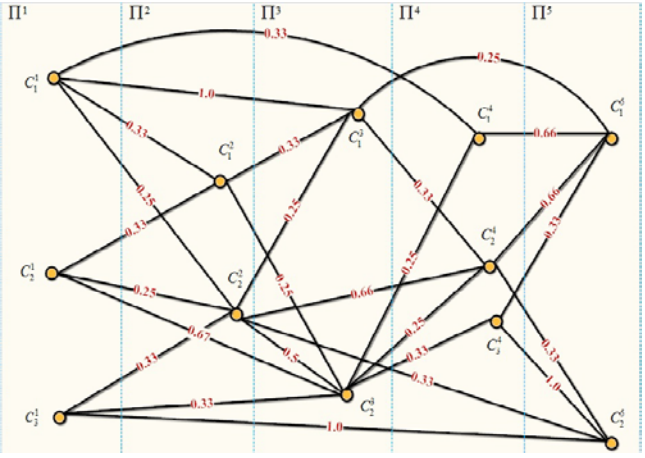


Fig.6 Linked Spectral Graph

Generally the weight estimated for each cluster $W_{ij} \in W$ that connects the clusters $C_i, C_j \in \pi$ is calculated by the proximity of the related data members in the clusters as shown below equation (1),

$$W_{ij} = \frac{d_i \cap d_j}{d_i \cup d_j} \tag{1}$$

where $d_i \subset X$ denotes the number of occurrences of the data points corresponds to the cluster $C_i \in \pi$. In the graph shown above, the circle node represents the clusters and the edge node will be present only when its appropriate weights are tends to be nonzero.

Having obtained the Spectral Graph with its weighted vertices between the cluster nodes, the most simplest and direct way to solve the graph partition problem is to make the Normalized cut, as it divides the best group of vertices based on the similarity. Here the Normalized Partition cut is measured through the following calculation in equation (2).

$$NP_{cut}(\pi_i, \pi_j) = \frac{MinCut(\pi_i, \pi_j)}{\sum(vol(\pi_i, \pi_j))} \tag{2}$$

where the Normalized Partition cut of every two partitions in the ensemble can be estimated through the ratio of minimum number of clusters obtained in π_i, π_j to the sum of weights associated with the clusters in the corresponding two partitions. This objective function is mainly used to normalize the graph cuts in terms of subsets of volumes in the Cluster Ensemble partition. And also it is highly computational efficient in nature.

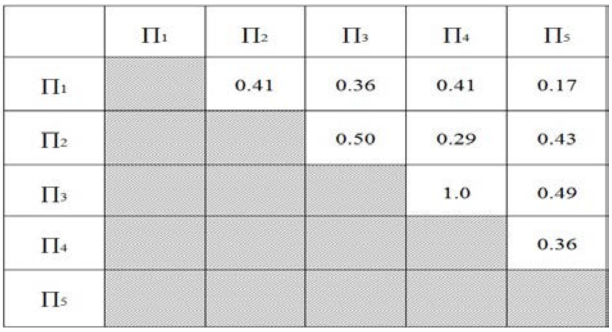
However, to evaluate the proximity between the Ensemble Members in the Partition, it is mandatory to take into account the composite characteristic feature of each sub partitions in the Ensemble. Inspired by this proposal, Weighted Spectral Quality (WSQL) factor is established. The Normalized Partition cut NP_{cut} measure estimated from equation (2) is obtained for the Ensemble members π_i, π_j . This in turn is applied over the below mentioned equation (3).

$$WSQL_{ij}^{c} = \frac{1}{n}\sum_{i=1}^{p} NP_{cut} \tag{3}$$

where n denotes the total number of clusters formed in the Ensemble members π_i, π_j and p tends to be fixed to two partitions. Following that the similarity between the two ensemble partitions can be defined in equation (4),

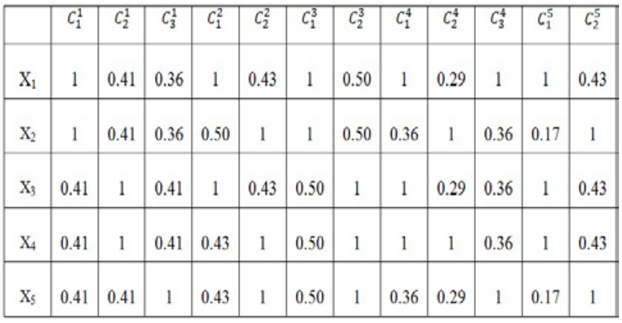
$$Sim_{WSQL}(\pi_i,\pi_j)=\frac{WSQL_{ij}^{c}}{Min\left[\left(\sum W_t(\pi_i)\right),\left(\sum W_t(\pi_j)\right)\right]}*DC \quad (4)$$

where $W_t(\pi_i)$ and $W_t(\pi_j)$ denotes the summation of total weights associated with the clusters that forms the triple in the Linked Spectral Graph. Formally, a triple $t=(V_t,E_t)$ is a sub graph of *LSG* containing two vertices $V_t=\{v_{i,}v_j\}\subset V$ and three cluster nodes termed as edges $E_t=\{e_{i,}e_{j,}e_k\}\subset E$. Hence the Similarity measure between the two ensemble members $\pi_i and \pi_j$ can be valued by considering minimum sum of weighted triples in the two partitions. Additionally to boost the confidence level of recognizing two non identical ensemble members being similar, a constant Decay factor $DC\in[0,1]$ is fixed. Following the Sample Cluster Ensemble shown in the Fig.3 the similarity measures between each ensemble members in the Cluster Ensemble are estimated with the decay factor fixed to 0.9. These measures are then formulated in Fig.7 and the Distilled Similarity Matrix is illustrated in the Fig.8. Consequently, from the empirical analysis it is recognized that the estimation of similarity measures between the ensemble members rather than the clusters in the Ensemble drastically improves the similarity degrees of the clustering results over the conventional Cluster Ensemble techniques.



	Π1	Π2	Π3	Π4	Π5
Π1	▨	0.41	0.36	0.41	0.17
Π2	▨	▨	0.50	0.29	0.43
Π3	▨	▨	▨	1.0	0.49
Π4	▨	▨	▨	▨	0.36
Π5	▨	▨	▨	▨	▨

Fig.7 Similarity Measures between the Ensemble Members where DC = 0.9



	C_1^1	C_2^1	C_3^1	C_1^2	C_2^2	C_1^3	C_2^3	C_1^4	C_2^4	C_3^4	C_1^5	C_2^5
X_1	1	0.41	0.36	1	0.43	1	0.50	1	0.29	1	1	0.43
X_2	1	0.41	0.36	0.50	1	1	0.50	0.36	1	0.36	0.17	1
X_3	0.41	1	0.41	1	0.43	0.50	1	1	0.29	0.36	1	0.43
X_4	0.41	1	0.41	0.43	1	0.50	1	1	1	0.36	1	0.43
X_5	0.41	0.41	1	0.43	1	0.50	1	0.36	0.29	1	0.17	1

Fig.8 Distilled Similarity Matrix (DSM)

The Weighted Spectral Quality (WSQL) algorithm is summarized below,

ALGORITHM: $WSQL(LSG,\pi_{i,}\pi_j)$

Input: *A Dataset with x dimensional data objects.*

Output*: Distilled Similarity Matrix.*

```
LSG = (V,W) a linked spectral graph where $C_i,C_j \in V$ ; ;
begin
    Compute Weight: $W_{ij}=\frac{d_i \cap d_j}{d_i \cup d_j}$ ;
    init $WSQL_{ij}^{c} \rightarrow 0$ ;
    for each $C \in \pi_i$
        If $C \in \pi_j$
            $NP_{cut}(\pi_i,\pi_j)=\frac{MinCut(\pi_i,\pi_j)}{\sum(vol(\pi_i,\pi_j))}$ ;
            $WSQL_{ij}^{c}=\frac{1}{n}\sum_{i=1}^{p} NP_{cut}$ ;
    Return $WSQL_{ij}^{c}$ ;
end
$Sim_{WSQL}(\pi_i,\pi_j)=\frac{WSQL_{ij}^{c}}{Min\left[\left(\sum W_t(\pi_i)\right),\left(\sum W_t(\pi_j)\right)\right]}*DC$ ;
end
```

4.3 Applying Spectral Clustering based Consensus Function to DSM

Having attained the DSM, a Spectral Clustering based Consensus Function [60] is applied to extract the final clustering results. This consensus technique requires the Distilled Similarity Measures through which it applies the spectral clustering algorithm to partition the similarity measures for exploiting the ultimate clustering solutions. In the first step, it builds the affinity matrix with the entries in the obtained DSM in which it represents the similarity degrees between the two ensemble members $\pi_i and\ \pi_j$. In the second step, after obtaining the affinity matrix, spectral clustering generates the diagonal matrix through the summation of entries in the diagonal. In the third step, it normalizes the affinity matrix in order to perform efficient dimensionality reduction. In the fourth step, it produces the Eigen vectors corresponding to the first six largest Eigen values of the affinity matrix and re-normalizes each rows of the matrix. Finally in the fifth step, K-Means algorithm is implemented to assign the samples in the newly formed data matrix to their corresponding clusters. Thus this consensus function proves to be the powerful and efficient method in obtaining absolute cluster results and also it attains the nearer optimal solutions.

5 Performance Evaluation

This section exposes the performance of proposed Linked Spectral Graph based Cluster Ensemble approach using few validity indices and variety of Medical datasets. The quality of each ensemble members in the total Cluster Ensemble acquired by this technique is evaluated against two different traditional clustering algorithms.

5.1 Examined Datasets

The experimental analysis is conducted over five medical datasets which are taken from the UCI Machine Learning Repository [61]. The details regarding the number of instances and attributes are summarized in the below Table I.

TABLE I
DATASETS USED IN THE EXPERIMENT

Datasets	Instances	Attributes
Arrhythmia	452	279
Dermatology	366	33
Heart Disease	303	75
Hepatitis	155	19
Lung Cancer	32	56

The descriptions about the experimented Medical datasets are as follows,

1) Arrhythmia- This dataset mainly denotes the presence and absence of cardiac disease. Among the 279 attributes, 206 are linear valued and the remaining are nominal.

2) Dermatology- The data comprises of clinical features of Erythematic disease observed in the patient. It also includes the age feature and the possibility of high effect of intermediate results of the disease.

3) Heart Disease- It includes the details of heart disease present in the patient which was taken from the Cleveland database.

4) Hepatitis- It represents the data regarding the inflammation of the liver and the inflammatory cells present in tissues of the organ.

5) Lung Cancer- This dataset describes about the three types of pathological lung cancer cells in the patient.

5.2 Evaluation Criteria

The experiment set out to observe the performance of the LSGCE in contrast to few conventional clustering algorithms. In order to analyze the efficiency of the proposed work, the final clustering results of each method is evaluated with its appropriate true labels by using the following performance validity metrics.

1) Classification Accuracy – It is the measure [62] of number of exactly classified data objects of the clustering results compared with the known true labels divided by the total number of data points in the datasets. This Classification accuracy measures can be estimated as given below,

$$CA(\pi^*) = \frac{\sum_{1=0}^{k} M_i}{D} \qquad (5)$$

where π^*denotes the final partition results, M_i illustrates the number of data objects with the majority of the cluster label points in the cluster i, then D is the total number of data objects in the dataset.

2) Error Rate- It is the term that describes the degree of errors or irrelevant data encountered during data clustering. This error rate is computed as given below,

$$E = 1 - CA \qquad (6)$$

where *CA* denotes the clustering accuracy calculated from the equation (5).

3) Rand Index- Generally Rand index *l*, is the measure of the similarity between the two data clustering. In other words it is stated that a measure [63] of number of object pairs that exist in the same and different clusters. More formally it can also be stated as a proportional measure of the quantity of agreements and disagreements between the two partitions. It can be calculated as below,

$$R = \frac{(x+y)}{(x+y)+(u+v)} \qquad (7)$$

where $(x + y)$ can be denoted as the number of agreements between the two clusters C_i and C_j similarly $(u + v)$ can be considered as the number of disagreements between the same two clusters.

4) Compactness- It measures [62] the average distances between the each pair of data points occurring in the same cluster. More specifically it is calculated as given below,

$$CP(\pi^*) = \frac{1}{D}\sum_{k=1}^{K} d_k \left(\frac{\sum_{x_i,x_j \in C_k} d(x_i,x_j)}{\frac{d_k(d_k-1)}{2}} \right) \qquad (8)$$

where K denotes then number of clusters formed finally, d_k is the number of data objects corresponding to that particular cluster, and $d(x_i, x_j)$ is the distance between the data points $x_i and\ x_j$, then D be the total number of data points in the dataset.

5) Dunn- Its main aspire [64] is to identify the closeness and the well separated clusters. It compares the size of the clusters with the distance between the clusters. Such that it is stated the distances between the clusters are expected to be large and the diameter of the clusters should be small. Hence it can be computed as given below,

$$Dunn(\pi^*) = \frac{\min d(C_i, C_j)}{\max \Delta(C_i)} \qquad (9)$$

where $d(C_i, C_j)$ denotes the distance computed between the two clusters $C_i\ and\ C_j$, and $\Delta(C_i)$ expresses the size of the cluster .

6) Davies Bouldin (DB)- This DB [65] measure mainly determines average of the similarity between the two clusters $C_i\ and\ C_j$ in which it is defined by the estimation of dispersion of a single cluster and the dissimilarity measure between the two clusters. It is evaluated as follows,

$$Sim_{ij} = \frac{s_i + s_j}{d_{ij}} \qquad (10)$$

in which s_i denotes the dispersion of C_i and d_{ij} shows the dissimilarity between the two clusters, these can be calculated as given below,

$$s_i = \frac{1}{|C_i|}\sum_{\forall x \in C_i} d(x, v_i) \qquad (11)$$

$$d_{ij} = d(v_i, v_j) \qquad (12)$$

where $|C_i|$ denotes the number of data points in the cluster C_i and $v_i and v_j$ shows the center points of the two clusters C_i and C_j respectively. Hence from the above the DB can be derived as,

$$DB(\pi^*) = \frac{1}{k}\sum_{i=1}^{k} Sim_i \qquad (13)$$

where $Sim_i = \max($ ▪$im_{ij})$ such that $i \neq j$.

The following Table II and Table III exemplify the results of each measure when evaluated with the LSGCE algorithm implemented in MATLAB environment to find its efficacy.

TABLE II
AVERAGE CLUSTERING ACCURACY AND ERROR RATES OF 10 RUNS

Datasets	Clustering Accuracy	Clustering Error Rate
Heart Disease	0.785	0.215
Lung Cancer	0.468	0.532
Hepatitis	0.750	0.250
Arrhythmia	0.565	0.435
Dermatology	0.588	0.412

TABLE III
PERFORMANCE COMPARISON AMONG EVALUATION INDICES

Datasets	Compactness	Rand Index	Davies-Bouldin	Dunn
Heart Disease	36.02	0.174	0.840	0.785
Lung Cancer	50.03	0.130	2.318	0.993
Hepatitis	66.04	0.595	0.640	1.470
Arrhythmia	62.83	0.420	2.805	0.824
Dermatology	15.20	0.212	1.818	1.574

The following Figures represent the graphical notation of the performance of LSGCE when examined with the Medical Datasets over several evaluation indices.

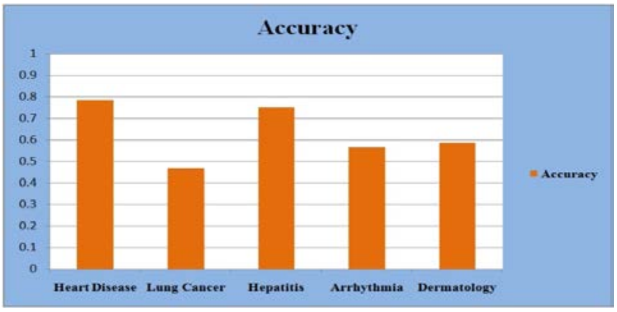


Fig.9. Average Accuracy Rates of 10 runs

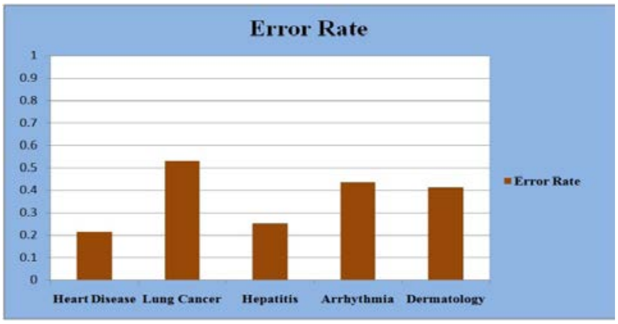


Fig.10. Average Error Rates of 10 runs

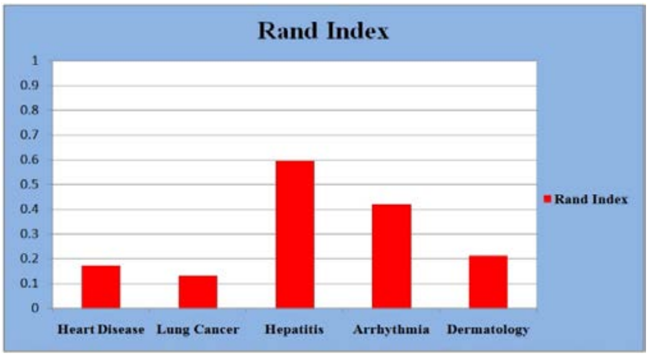


Fig.11 Rand Index Rates

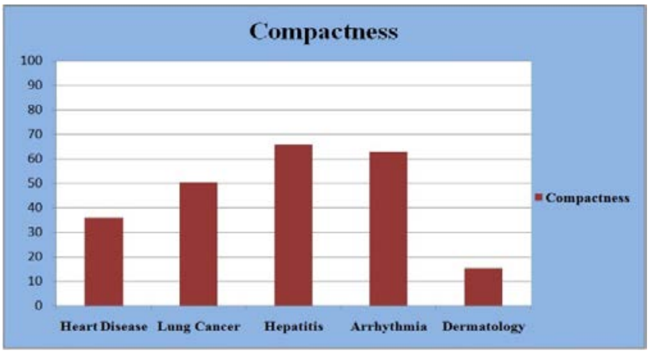


Fig.12 Compactness Rates

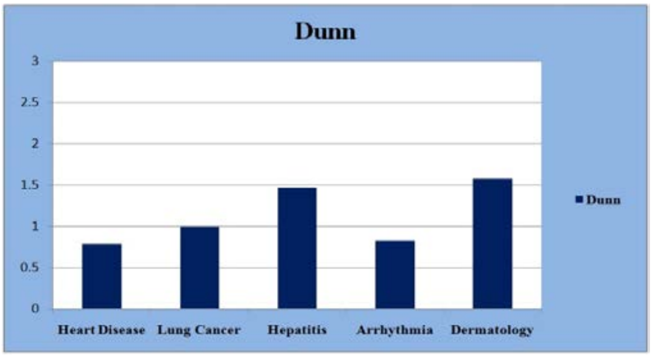


Fig.13 Dunn Rates

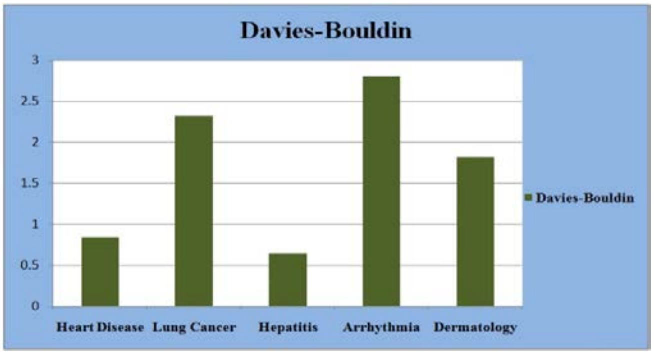


Fig.14 Davies-Bouldin Rates

5.3 Compared Traditional Clustering Methods

In order to estimate the efficiency of the newly proposed Linked Spectral Graph based Cluster Ensemble (LSGCE) approach, the following two traditional clustering methods have been contrasted over the Lung Cancer dataset.

5.3.1 K-Means Clustering Method

K-Means is the well-known conventional clustering algorithm [66] often used to cluster the numerical data. It is an algorithm mainly framed to find the K-center point of the dataset based on the distance between the other data points and the center point. The Euclidean distance function is most preferable in nature. It initiates the K number of value as a starting point to estimate the cluster centers. In this algorithm, the main issue is to reduce the distance between the data object and the corresponding cluster center point in the dataset. Moreover this K-Means algorithm [66] faces two main challenges such as its behavior mainly depends on the initial center point and it often converges to local minima. Different initial cluster center points provide different clustering solutions. This difficulty is most widely seen when initial center points are not well separated.

5.3.2 Fuzzy C-Means Clustering Method

Fuzzy C-Means clustering algorithm [67] was mainly established to smoothen the hard nature of K-Means algorithm in which a data alone can assign to the cluster. It mainly makes use of the fuzzy partitioning to let the data objects to assign to all the clusters generated with the membership grade between 0 and 1 and the sum of it is 1. By considering the highest grade the data is recorded to its appropriate cluster. Adversely the same challenging factors of K-Means happened to Fuzzy C-Means algorithm as it cannot ensure the global optima. The Clustering result is highly dependent to the randomly assigned initial membership grades.

The following Table IV compares the average clustering accuracy rates of LSGCE with traditional clustering techniques over 10 rums.

Furthermore the forthcoming Fig.15 represent the graphical illustration of the performance of newly proposed Linked Spectral Graph based Cluster Ensemble (LSGCE) approach examined with Medical datasets.

TABLE IV
ACCURACY COMPARISON OF TRADITIONAL CLUSTERING METHODS

Datasets	LSGCE	KM	FCM
Heart Disease	0.785	0.761	0.691
Lung Cancer	0.468	0.431	0.432
Hepatitis	0.750	0.713	0.732
Arrhythmia	0.565	0.554	0.488
Dermatology	0.588	0.428	0.419

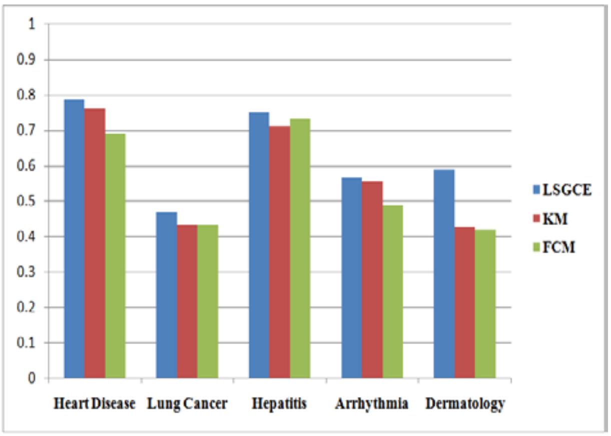


Fig.15 Accuracy Mean of Different Medical Datasets

6 Conclusion and Future Work

The main contribution in this paper is to exemplify the novel Linked Spectral Graph based Cluster Ensemble approach for providing efficiency in clustering Medical data and also in reducing cluster degradation problem. It greatly aims to explore and makes use of the relationship degree between the generated base clustering solutions. Additionally LSGCE performs the similarity assessment among the ensemble members of the Cluster Ensemble. This allows formation of Distilled Similarity Matrix (DSM) to be refined from the traditional Binary cluster association Matrix. The challenging issue of generating DSM is expertly resolved by Weighted Spectral Quality (WSQL) algorithm. With the results of the extracted similarity measures, Spectral based Consensus Function is applied to finalize the ultimate cluster solutions. Hence the experimental investigation of conventional clustering algorithms tested over the Medical datasets suggests that newly proposed LSGCE approach highly overwhelms the traditional ones. Beyond these accomplishments, the future work includes the extension of LSGCE in Text data clustering. Furthermore this new approach can also be applied to other business related dataset with huge dimensions and also it further improves its efficiency in execution time.

References:

[1] D.S. Hochbaum and D.B. Shmoys, “A Best Possible Heuristic for the K-Center Problem,”*Math.of Operational Research*,vol. 10, no. 2, pp. 180-184, 1985.

[2] L. Kaufman and P.J. Rousseeuw,”Finding Groups in Data: An Introduction to Cluster Analysis”.*Wiley Publishers*, 1990.

[3] A.K. Jain and R.C. Dubes,Algorithms for Clustering.Prentice-Hall, 1998.

[4] P. Zhang, X. Wang, and P.X. Song, “Clustering Categorical Data Based on Distance Vectors,”*The J. Am. Statistical Assoc.*,vol. 101, no. 473, pp. 355-367, 2006.

[5] A. Ng, M. Jordan, and Y. Weiss, “On Spectral Clustering: Analysis and an Algorithm,” *Advances in Neural Information Processing Systems*,vol. 14, pp. 849-856, 2001.

[6] Z. Huang, “Extensions to the K-Means Algorithm for Clustering Large Data Sets with Categorical Values,” *Data Mining and Knowledge Discovery*, vol. 2, pp. 283-304, 1998.

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

[8] D.H. Fisher, “Knowledge Acquisition via IncrementalConceptual Clustering,”*Machine Learning*,vol. 2, pp. 139-172, 1987.

[9] D. Gibson, J. Kleinberg, and P. Raghavan, “Clustering Categorical Data An Approach Based on Dynamical Systems,”*VLDB J.*, vol. 8, nos. 3-4, pp. 222-236, 2000.

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

[11] M.J. Zaki and M. Peters, “Clicks: Mining Subspace Clusters in Categorical Data via Kpartite Maximal Cliques,”*Proc. Int’l Conf. Data Eng. (ICDE)*,pp. 355-356, 2005.

[12] V. Ganti, J. Gehrke, and R. Ramakrishnan, “CACTUS: Clustering Categorical Data Using Summaries,”*Proc. ACM SIGKDD Int’l Conf. Knowledge Discovery and Data Mining* (KDD),pp. 73-83, 1999.

[13] D. Barbara, Y. Li, and J. Couto, “COOLCAT: An Entropy-Based Algorithm for Categorical Clustering,”*Proc. Int’l Conf. Information and Knowledge Management (CIKM)*,pp. 582-589, 2002.

[14] Y. Yang, S. Guan, and J. You, “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.

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

[16] P. Andritsos and V. Tzerpos, “Information-Theoretic Software Clustering,”*IEEE Trans. Software Eng.*,vol. 31, no. 2, pp. 150-165, Feb. 2005.

[17] S. Indrajit, M. Ujjwal , & Nilanjan. “Differential Fuzzy Clustering for Categorical Data”. *International Conference on Methods and Models in Computer Science*, 2009.

[18] 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.

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

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

[21] J. Kittler, M. Hatef, R. Duin, and J. Matas, “On Combining Classifiers,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 20, no. 3, pp. 226-239, Mar. 1998.

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

[23] X. Hu and I. Yoo, “Cluster Ensemble and Its Applications in Gene Expression Analysis,” *Proc. Asia-Pacific Bioinformatics Conf.*, pp. 297-302, 2004.

[24] M. Law, A. Topchy, and A.K. Jain, “Multiobjective Data Clustering,” *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, vol. 2, pp. 424-430, 2004.

[25] Huang. Z, “Extensions to the K-Means Algorithm for Clustering Large Data Sets with Categorical Values,” *Data Mining and Knowledge Discovery*, vol. 2, pp. 283-304, 1998.

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

[27] L.I. Kuncheva and D. Vetrov, “Evaluation of Stability of K-Means Cluster Ensembles with Respect to Random Initialization,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 28, no. 11, pp. 1798-1808, Nov. 2006.

[28] X.Z. Fern and C.E. Brodley, “Random Projection for High Dimensional Data Clustering: A Cluster Ensemble Approach,” *Proc. Int’l Conf. Machine Learning (ICML)*, pp. 186-193, 2003.

[29] S. Dudoit and J. Fridyand, “Bagging to Improve the Accuracy of a Clustering Procedure,” *Bioinformatics*, vol. 19, no. 9, pp. 1090-1099, 2003.

[30] B. Minaei-Bidgoli, A. Topchy, and W. Punch, “A Comparison of Resampling Methods for Clustering Ensembles,” *Proc. Int’l Conf.Artificial Intelligence*, pp. 939-945, 2004.

[31] A.P. Topchy, A.K. Jain, and W.F. Punch, “Clustering Ensembles: Models of Consensus and Weak Partitions,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 27, no. 12, pp. 1866-1881, Dec. 2005.

[32] B. Fischer and J.M. Buhmann, “Bagging for Path-Based Clustering,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 25, no. 11, pp. 1411-1415, Nov. 2003.

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

[34] N. Iam-On, T. Boongoen, and S. Garrett, “Refining Pairwise Similarity Matrix for Cluster Ensemble Problem with Cluster Relations,” *Proc. Int’l Conf. Discovery Science*, pp. 222-233, 2008.

[35] L. Getoor and C.P. Diehl, “Link Mining: A Survey,” *ACM SIGKDD Explorations Newsletter*, vol. 7, no. 2, pp. 3-12, 2005.

[36] D. Cristofor and D. Simovici, “Finding Median Partitions Using Information-Theoretical-Based Genetic Algorithms,” *J. Universal*

Computer Science, vol. 8, no. 2, pp. 153-172, 2002.

[37] N. Nguyen and R. Caruana, “Consensus Clusterings,” *Proc. IEEE Int’l Conf. Data Mining (ICDM)*, pp. 607-612, 2007.

[38] Z. Yu, H.-S. Wong, and H. Wang, “Graph-Based Consensus Clustering for Class Discovery from Gene Expression Data,” *Bioinformatics*, vol. 23, no. 21, pp. 2888-2896, 2007.

[39] G. Karypis and V. Kumar, “Multilevel K-Way Partitioning Scheme for Irregular Graphs,” *J. Parallel Distributed Computing*, vol. 48, no. 1, pp. 96-129, 1998.

[40] A. Ng, M. Jordan, and Y. Weiss, “On Spectral Clustering: Analysis and an Algorithm,” *Advances in Neural Information Processing Systems*, vol. 14, pp. 849-856, 2001.

[41] X.Z. Fern and C.E. Brodley, “Solving Cluster Ensemble Problems by Bipartite Graph Partitioning,”*Proc. Int’l Conf. Machine Learning (ICML)*,pp. 36-43, 2004.

[42] A. Gionis, H. Mannila, and P. Tsaparas, “Clustering Aggregation,” *Proc. Int’l Conf. Data Eng. (ICDE)*, pp. 341-352, 2005.

[43] S. Dudoit and J. Fridlyand, “A Prediction-Based Resampling Method to Estimate the Number of Clusters in a Data Set,”*GenomeBiology*,vol. 3, no. 7, pp. 0036.1-0036.21, 2002.

[44] S. Dudoit and J. Fridlyand, “Bagging to Improve the Accuracy of a Clustering Procedure,”*Bioinformatics*,vol. 19, no. 9, pp. 1090-1099, 2003.

[45] M. Smolkin and D. Ghosh, “Cluster Stability Scores for Microarray Data in Cancer Studies,”*BMC Bioinformatics*,vol. 4, article 36, 2003.

[46] S. Monti, P. Tamayo, J. Mesirov, and T. Golub, “Consensus Clustering: A Resampling-Based Method for Class Discovery and Visualization of Gene Expression Microarray Data,”*Machine Learning*,vol. 52, pp. 9 118, 2003.

[47] Z. Yu and H.-S. Wong, “Knowledge Based Cluster Ensemble for Cancer Discovery from Biomolecular Data,”*IEEE Trans. NanoBioscience*, vol. 10, no. 2, pp. 76-85, June 2011.

[48] Zhiwen Yu Member, IEEE, Hantao Chen Jane You Member, IEEE, Guoqiang Han Le Li “ Hybrid Fuzzy Cluster Ensemble Framework for Tumor Clustering from Bio-molecular Data” *IEEE Transactions on computational biology and bioinformatics* 2013.

[49] Zhiwen Yuℵ, Member, IEEE, Hau-San Wongb, Member, IEEE, Jane You, Member, IEEE, Qinmin Yang, Member, IEEE, and Hongying Liao “ Knowledge based Cluster Ensemble for Cancer Discovery From Biomolecular Data” *IEEE Transactions on Nanobioscience*, Vol 10 No. 2, june 2011.

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

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

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

[53] Ka Ka Ng E, Wai-Chee Fu A, Chi-Wing Wong R “Projective clustering by histograms”. *IEEE Trans Knowl Data Eng (TKDE)* 17(3):369–3832005.

[54] Yangzihao Wang, “Spectral Clustering: A Graph Partitioning Point of View”, *ECS231 Course Report*, CSE University of California, Davis.

[55] Inderjit S.Dhillon, Yuqiang Guan, and Brian Kulis, “Kernel K-Means, Spectral Clustering and Normalized cuts”, *ACM 1-58113-888-1/04/0008 KDD* 04 August 2004.

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

[57] Z. He, X. Xu, and S. Deng, “A Cluster Ensemble Method for Clustering Categorical Data,” *Information Fusion*, vol. 6, no. 2, pp. 143-151, 2005.

[58] Jeh G, Widom J “SimRank: A Measure of Structural-Context Similarity." In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 538{543. ACM, New York 2002.

[59] Weiguo Zheng, Lei Zou,YanSong Feng, Le Chen, Dongyan, “Efficient SimRank based Similarity Join over Large graphs” *Proceedings of the VLDB endowment*, Vol 6 No.7 2011.

[60] Zhiwen yu, Le Li, Jane You, Hau-San Wong, and Guoqiang Han, “SC3: Triple Spectral Clustering Based Consensus Clustering Framework for Class Discovery from Cancer Gene Expresion Profiles”, *IEEE/ACM Transaction on Computational Biology and Bioinformatics*, Vol 9, No. 6, December 2012.

[61] A. Asuncion and D.J. Newman, “UCI Machine Learning Repository,” *School of Information and Computer Science*, Univ. of California,http://www.ics.uci.edu/~mlearn/ML Repository. html.

[62] Nguyen N, Caruana R ,”Consensus Clusterings." In *Proceedings of IEEE International Conference on Data Mining*, pp. 607{612. IEEE Computer Society, Washington, DC 2007.

[63] Rand WM ,”Objective Criteria for the Evaluation of Clustering Methods." *Journal of the American Statistical Association*, 66, 846 850 1971.

[64] Dunn JC, “Well Separated Clusters and Optimal Fuzzy Partitions." *Cybernetics and Systems,* 4(1), 95104 1974.

[65] Davies DL, Bouldin DW, “A Cluster Separation Measure." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1(2), 224227 1979.

[66] G. Hammerly, C. Elken, “Alternatives to the K-means algorithm that find better clusterins”, in: *Proceedings of the 11 th International Conference on Information and Knowledge Management,* 2002, pp. 600-607.

[67] R.L. Canon, J.Dave and J.C. Bezdek, “Efficient implementation of the fuzzy cmeans clustering algorithms”. *IEEE Trans Pattern Arial Machine,* Intell 8, 248-255.

[68] Yuzhen Zhao, Xiyu Liu, and Wenping Wang, “ROCK Clustering algorithm based on the P System with active membranes”, *WSEAS Transactions on Computers,*Vol 13 2014.