International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056
Volume: 04 Issue: 02 | Feb -2017 www.irjet.net p-ISSN: 2395-0072

CANCER DATA PARTITIONING WITH DATA STRUCTURE AND DIFFICULTY INDEPENDENT CLUSTERING SCHEME

K.R.Kavitha[1], G. Angeline Prasanna[2]

[1]*Research Scholar, Department of Computer Science, Kaamadhenu Arts and Science College, Tamilnadu, India*
[2]*Head and Assistant Professor, Dept. of Computer Application&IT, Kaamadhenu Arts and Science College, Sathy,*

Abstract - Hidden knowledge extraction is the main operation of the data mining applications. Decision making processes are carried out with the support of the discovered knowledge. Relevant records are grouped by using the clustering methods. Cancer diagnosis data values are maintain in high dimensional model. Micro array data models are adapted to process the high dimensional data values. Distance measures are estimated to identify the record relationship levels. The cluster representative elements are referred as cluster ensembles. All the relationship analysis is carried out through the ensemble analysis mechanism.

Cluster ensemble consolidates the transactions of the individual cluster results. Distributed Computing, Knowledge Reuse and Quality and Robustness are the key features of the cluster ensemble models. The ensemble members are fetched using the Incremental Ensemble Membership Selection (IEMS) scheme. The clustering operations are performed with Incremental Semi-Supervised Cluster Ensemble (ISSCE) framework. The cancer expressions are compared using the Similarity Functions (SF). Data and structure dependency is incased in the ISSCE scheme.

The cancer data partitioning process uses the breast cancer data values. Noisy data removal and missing value replacement operations are carried out under the data preprocess. The Dynamic Ensemble Membership Selection (DEMS) scheme is build to support data structure and complexity independent clustering process. Data clustering operations are performed through the Partition Around Medoids (PAM) clustering technique. The PAM clustering technique and DEMS scheme are combined to handle the ensemble based data partitioning process. The clustering accuracy level is increased in the healthcare data partitioning process.

Key Words: ISSCE (Incremental Semi Supervised Cluster Ensemble, IEMS (Incremental Ensemble Membership Selection), SF (Similarity Function), DEMS (Dynamic Ensemble Membership Selection),PAM **(**Partition Around Medoids) .

1. INTRODUCTION

1.1 Clustering Concepts

Clustering is the classification of objects into different groups, or more precisely, the partitioning of a data set into subsets, so that the data in each subset share some common trait - often proximity according to some defined distance measure. Data clustering is a common technique for statistical data analysis, which is used in many fields, including machine learning, data mining, pattern recognition, image analysis and bioinformatics.

It is possible to guarantee that homogeneous clusters are created by breaking apart any cluster that is unhomogeneous into smaller clusters that are homogeneous.

- ✓ Used mostly for consolidating data into a high-level view and general grouping of records into like behaviours. Space is defined as default n-dimensional space, or is defined by the user, or is a predefined space driven by part.
- ✓ Besides the term data clustering, there are a number of terms with similar meanings, including cluster analysis, automatic classification, numerical taxonomy, botryology and typological analysis.
- ✓ The clustering technique is called an unsupervised learning technique. It is a technique that when they are run, there is not a particular reason for the creation of the models to perform predication. In clustering, there is no particular sense of why certain records are near each other or why they all fall into the same cluster.

Use of Clustering in Data Mining

Clustering is often one of the first steps in data mining analysis. It identifies groups of related records that can be used as a starting point for exploring further relationships. This technique supports the development of population segmentation models, such as demographic-based customer segmentation. A company that sale a variety of products may need to know about the sale of all of their products in order to check that what product is giving extensive sale and which is lacking. This is done by data mining techniques. But if the system clusters the products that are giving fewer sales then only the cluster of such products would have to be checked rather than comparing the sales value of all the products. This is actually to facilitate the mining process.

K-Means Algorithm

The K Means clustering algorithm is applied for the postulated in the nineteen sixties. For a *m* attribute problem, each instance maps into a *m* dimensional space. The cluster centroid describes the cluster and is a point in *m* dimensional space around which instances belonging to the cluster occur. The distance from an instance to a cluster center is typically the Euclidean distance though variations such as the Manhattan distance are common. As most implementations of K-Means clustering use Euclidean distance.

- ✓ **Strength of the K-Means**
  - Relatively efficient: O (t kn), where n is of objects, k is of clusters and t is of iterations. Normally, k, t << n
  - Often terminates at a local optimum
- ✓ **Weakness of the K-Means**
  - Applicable only when mean is defined; what about categorical data?
  - Need to specify k, the number of clusters, in advance
  - Unable to handle noisy data and outliers
- ✓ **Variations of K-Means usually differ in**
  - Selection of the initial k means
  - Dissimilarity calculations
  - Strategies to calculate cluster means
- ✓ **Partitioning Methods**
  - Reallocation method - start with an initial assignment of items to clusters and then move items from cluster to cluster to obtain an improved partitioning
  - It involves movement or "reallocation" of records from one cluster to other to create best clusters. It uses multiple passes through the database fastly.
  - Single Pass method - simple and efficient, but produces large clusters and depends on order in which items are processed
  - The database must be passed through only once in order to create clusters

1.2 Types of Clustering Methods

There are many clustering methods available and each of them may give a different grouping of a dataset. The choice of a particular method will depend on the type of output desired, the known performance of method with particular types of data, the hardware and software facilities available and the size of the dataset. In general, clustering methods may be divided into two categories based on the cluster structure which they produce. The non-hierarchical methods divide a dataset of N objects into M clusters, with or without overlap.

These methods are sometimes divided into partitioning methods, in which the classes are mutually exclusive and the less common clumping method overlap is allowed. Each object is a member of the cluster with which it is most similar, the threshold of similarity has to be defined.. Some of the important Data Clustering Methods are described below.

Partitioning Methods

The partitioning methods generally result in a set of M clusters, each object belonging to one cluster. Each cluster may be represented by a centroid or a cluster representative; this is some sort of summary description of all the objects contained in a cluster. If the number of the clusters is large, the centroids can be further clustered to produces hierarchy within a dataset.

Single Pass: A very simple partition method, the single pass method creates a partitioned dataset as follows:

1. Make the first object the centroid for the first cluster.
2. For the next object, calculate the similarity, S with each existing cluster centroid, using some similarity coefficient.
3. If the highest calculated S is greater than some specified threshold value, add the object to the corresponding cluster and re determine the centroid; otherwise, use the object to initiate a new cluster. If any objects remain to be clustered, return to step 2.

Hierarchical Agglomerative methods

The hierarchical agglomerative clustering methods are most commonly used. The construction of an hierarchical agglomerative classification can be achieved by the following general algorithm.

1. Find the 2 closest objects and merge them into a cluster
2. Find and merge the next two closest points, where a point is either an individual object or a cluster of objects.
3. If more than one cluster remains , return to step 2

Individual methods are characterized by the definition used for identification of the closest pair of points and by the means used to describe the new cluster when two clusters are merged.

The Single Link Method (SLINK)

The single link method is probably the best known of the hierarchical methods and operates by joining, at each step, the two most similar objects, which are not yet in the same cluster. The name single link thus refers to the joining of pairs of clusters by the single shortest link between them.

The Complete Link Method (CLINK)

The complete link method is similar to the single link method except that it uses the least similar pair between two clusters to determine the inter-cluster similarity. This method is characterized by small, tightly bound clusters.

The Group Average Method

The group average method relies on the average value of the pair wise within a cluster, rather than the maximum or minimum similarity as with the single link or the complete link methods. Since all objects in a cluster contribute to the inter –cluster similarity object is average

more like every other member of its own cluster then the objects in any other cluster.

Text Based Documents

In the text based documents, the clusters may be made by considering the similarity as some of the key words that are found for a minimum number of times in a document. Now when a query comes regarding a typical word then instead of checking the entire database, only that cluster is scanned which has that word in the list of its key words and the result is given. The order of the documents received in the result is dependent on the number of times that key word appears in the document.

2. REVIEW OF THE LITERATURE

2.1 Speaker Diarization with Eigengap Criterion And Cluster Ensembles

Nowadays, a rapid increase in the volume of recorded speech is manifested. For example, archives of television and audio broadcasting, meeting recordings and voice mails have become a commonplace. A growing need for automatically processing such archives has arisen. Their enormous size hinders content organization, navigation, browsing and search. Speaker segmentation and speaker clustering alleviate the management of huge audio archives.

In the latter case, single microphone recordings have been used and the reference speech/nonspeech segmentation has been exploited in order to focus on a single source of the diarization error rate, namely the speaker error that is associated to the portion of the total length of the speech segments that are clustered into wrong speaker groups.

2.2 Robust Ensemble Clustering Using Probability Trajectories

The ensemble clustering technique has recently been drawing increasing attention due to its ability to combine multiple clusterings to achieve a probably better and more robust clustering. The relationship between objects lies not only in the direct connections, but also in the indirect connections. The key problem here is how to exploit the global structure information in the ensemble effectively and efficiently and thereby improve the final clustering results. Microcluster Similarity Graph (MSG) is constructed with the MCA matrix. Then, the ENS strategy is performed on the MSG and the sparse graph K-ENG is constructed by preserving a small number of probably reliable links. The random walks are conducted on the K-ENG graph and the PTS similarity is obtained by comparing random walk trajectories. Having computed the new similarity matrix, any pair-wise similarity based clustering methods can be used to achieve the consensus clustering. Typically, the system uses two novel consensus functions, termed PTA and PTGP, respectively. Note that PTA is based on agglomerative clustering, while PTGP is based on graph partitioning. The PTA and PTGP methods exhibit a significant advantage in clustering robustness over the baseline methods.

2.3 Constraint Neighborhood Projections for Semi-Supervised Clustering

Patterns are discovered using clustering, there exists known prior knowledge about the problem. Recently, semi-supervised clustering has emerged as an important variant of the traditional clustering paradigm.

The semi supervised clustering scheme is constructed with less training labels and better results that are also the goal of semi supervised learning. The system uses a semi-supervised clustering method based on Constraint Neighborhood Projections (CNP), where the constrained pairwise data points and their neighbors are used to transform the input data into a low dimensional space. The method requires fewer labeled data points for semi supervised learning and can naturally deal with the constraint conflicts. Consequently, the method has better generalization capability and more flexibility than some state-of-the-art methods.

2.4 Hierarchical Cluster with Co Association Based Cluster Ensembles

Clustering is the process of identifying the underlying groups or structures in a set of patterns without the use of class labels. While there have been a large set of clustering algorithms all have their limitations in terms of data characteristics that can be processed and types of clusters that can be found. The performance of many clustering algorithms also strongly depends on proper choices of parameters and/or initializations. As a result, the choice of appropriate clustering algorithms and/or parameters is highly problem dependent and often involves lots of heuristic choices or trial and error.

2.5 Double Selection based Ensemble for Tumor Clustering

The continuous improvement of microarray techniques and their applications in cancer research provides a new avenue to the diagnosis and the treatment of cancer. For example, the self organizing feature map is applied to identify Acute Myeloid Leukemia (AML) and Acute Lymphoblastic Leukemia (ALL) from microarray data. The combination of hierarchical and probabilistic clustering techniques is adopted distinguish different subtypes of lung adenocarcinoma from cancer gene expression profiles.

Three semi-supervised clustering ensemble frameworks Feature Selection based Semi Supervised Clustering Ensemble framework (FS-SSCE), Double Selection based Semi-Supervised Clustering Ensemble framework (DS-SSCE) and modified DS-SSCE (MDSSSCE)are employed to perform tumor clustering on bio-molecular data. (2) The clustering solutions are also viewed in an ensemble as new attributes of the original dataset, adopt the feature selection techniques to remove noisy genes and prune redundant clustering solutions in the ensemble under the same framework. The double selection approach is applied to perform tumor clustering from bio molecular data under the cluster ensemble framework. (3) To consider multiple clustering solution selection strategies at the same time.

3 PROBLEM ANALYSES

3.1 Existing Methodology

Cluster ensemble approaches are gaining more and more attention, due to its useful applications in the areas of pattern recognition, data mining, bioinformatics and so on. When compared with traditional single clustering algorithms, cluster ensemble approaches are able to integrate multiple clustering solutions obtained from different data sources into a unified solution and provide a more robust, stable and accurate final result.

The contributions of the system are fourfold. An Incremental Ensemble framework for Semi-Supervised Clustering in high dimensional feature spaces. A local cost function and a global cost function are applied to incrementally select the ensemble members. The similarity function is adopted to measure the extent to which two sets of attributes are similar in the subspaces. Non-parametric tests are used to compare multiple semi supervised clustering ensemble approaches over different datasets.

3.2 PROPOSED METHODOLOGY

The data clustering process is initiated to partition the data sets with its relevancy levels. The similarity measures are used to estimate the relationship between the transactions. The clustering operations are carried out in a supervised manner. The data preprocessing methods are initiated to remove the noise from the data sets. Redundant data filtering and missing value assignment tasks are carried out under the data preprocess tasks. Data partitioning is carried out with the user input cluster count values.

The PAM cluster method accepts a dissimilarity matrix. It is more robust because it minimizes a sum of dissimilarities instead of a sum of squared euclidean distances. The PAM clustering method allows selecting the number of clusters using mean. The clustering process supports all type of data values.

The Dynamic Ensemble Membership Selection (DEMS) mechanism is applied to select the cluster ensembles with global information. Structure independent ensemble selection is supported in the DEMS mechanism. The data complexity levels are also considered in the DEMS model. The Partition Around Medoids (PAM) clustering scheme is integrated with the Dynamic Ensemble Membership Selection (DEMS) mechanism. The DEMS based PAM clustering algorithm increases the cluster accuracy levels.

Cluster Ensemble Selection Implementations

Ensembles are most effective when constructed from a set of predictors whose errors are dissimilar. To a great extent, diversity among ensemble members is introduced to enhance the result of an ensemble. Particularly for data clustering, the results obtained with any single algorithm over much iteration are usually very similar. In such a circumstance where all ensemble members agree on data set should be partitioned, aggregating the base clustering results will show no improvement over any of the constituent members.

Besides its efficiency, this ensemble generation method has the potential to lead to a high-quality clustering result.

Incremental Semi-Supervised Clustering Ensemble Framework

Semi-Supervised Clustering Ensemble approaches have been successfully applied to different areas, such as data mining, bioinformatics and so on. The semi-supervised clustering ensemble approach achieves good performance on UCI machine learning datasets. The prior knowledge provided by experts as pair wise constraints and the knowledge based cluster ensemble method and the double selection based semi-supervised clustering ensemble approach. Both of them are successfully used for clustering gene expression data. Few of them consider how to handle high dimensional datasets. The system uses the Random subspace based Semi-Supervised Clustering Ensemble approach (RSSCE).

Incremental Ensemble Member Selection

The Incremental Ensemble Member Selection (IEMS) scheme uses the input as the original ensemble, while the output is a newly generated ensemble with smaller size. Algorithm 2 provides an overview of the Incremental Ensemble Member Selection (IEMS) process. IEMS considers the ensemble members one by one and calculates the objective function (I_b) for each clustering solution I_b generated by E2CP with respect to the subspace A_b in the first step. In the second step, it sorts all the ensemble members in b⍰ in ascending order according to the corresponding values.

$$\mu_h = \frac{\sum_{i=1}^{n} |\theta(y_i=h) p_i}{\sum_{i=1}^{h} \theta(y_i=h)}$$

Where d(p_i, μ_h) denotes the Euclidean distance between the feature vectors p_i and μ_h denotes an indicator function, ø(true) = 1 and ø(false) = 0. The objective of the cost function is to optimize the squared distances of the feature vectors from the centers, such that as many constraints are satisfied as possible.

Given the original ensemble I~ and the new ensemble I~ the local objective function x_b for the local b-th ensemble member (A_b, x_b) £ with respect to the ensemble member (A_t, x_t) £ I~ is defined as follows:

$$\tau_b = \sum_{\forall A_t \in \tau} \frac{S(A_b, A_t)}{\Delta(I^b)}$$

where Δ(I^b) denotes the global objective function for the clustering solution I^b and S(A_b,A_t) denotes the similarity function between two subspaces A_b and A_t. Given the subspaces A_b and A_t, the set of attributes in these subspaces can be represented by Gaussian mixture models (GMMs).

$$\Omega^b = \{\varphi_I^b = (w_1^b, \mu_1^b, \sum_1^b \quad), \varphi_2^b = (w_2^b, \mu_2^b, \sum_{k1}^b \quad)\}$$

.

The similarity is estimated to analyze the relation values. Algorithm 3 provides a flowchart of the similarity function (SF) for $S(A_b,A_t)$. The input of SF is two Gaussian mixture models b and t, while the output is the similarity value $S(A_b,A_t)$ between two subspaces A_b and A_t. Specifically, the similarity function first considers the similarity of all the pairs of components in b and t. The Bhattacharyya distance is used to calculate the similarity between two components $\Omega^b{}_{h1}$ in b and $\Omega^t{}_{h2}$ in t, which is as follows:

$$\varphi(\varphi_{h1}^t.\varphi_{h2}^t) = \frac{1}{8}(\mu_{h1}^b - \mu_{h2}^t) \quad T \left(\frac{\sum_{h1}^b + \sum_{h2}^t}{}\right) - 1(\mu_{h1}^b - \mu_{h2}^t)$$

SF sorts all the component pairs in ascending order according to the corresponding Bhattacharyya distance values and inserts them into a queue. Next, it sets $S(A_b,A_t) = 0$, performs a de-queue operation and considers the component pair one by one. If $w^b{}_h\ 1 > 0$ and $w^t{}_h\ 2 > 0$, the minimum weight w between the two is selected in the first step, which is as follows:

$$w = \min (w^b{}_{h1}, w^t{}_{h2})$$

Partition Around Medoids (PAM) Clustering Algorithm

PAM stands for "Partition Around Medoids". The algorithm is intended to find a sequence of objects called medoids that are centrally located in clusters. Objects that are tentatively defined as medoids are placed into a set S of selected objects. If O is the set of objects that the set U = O – S is the set of unselected objects. The goal of the algorithm is to minimize the average dissimilarity of objects to their closest selected object. Equivalently, the system can minimize the sum of the dissimilarities between object and their closest selected object. The algorithm has two phases:
(i) In the first phase, BUILD, a collection of k objects is selected for an initial set S.
(ii) In the second phase, SWAP, one tries to improve the quality of the clustering by exchanging selected objects with unselected objects.

The goal of the algorithm is to minimize the average dissimilarity of objects to their closest selected object. The system can minimize the sum of the dissimilarities between object and their closest selected object. For each object p the system maintains two numbers. D_p, the dissimilarity between p and the closest object in S and E_p, the dissimilarity between p and the second closest object in S.

DEMS scheme based PAM Clustering Framework

The Partition Around Medoids (PAM) clustering scheme is applied with transaction relationship based model. The build and swap functions are used in the PAM clustering scheme. The build function selects the K objects. The swap function performs the transaction reassignment task to improve the cluster results. The build and swap function operations are carried out with the DEMS scheme. The similarity function is called to estimate the relationship levels. The data values are partitioned with the DEMS based PAM clustering method.

Advantages of the Proposed Methodology

The clustering methods are enhanced with the ensembles based model to increase the accuracy levels. The Incremental Semi-Supervised Cluster Ensemble (ISSCE) scheme is adapted to support clustering process with ensemble analysis model. The Incremental Ensemble Membership Selection (IEMS) scheme is used to fetch the ensemble members incrementally. The data relationship is estimated with the Similarity Function (SF) model. The Partition Around Medoids (PAM) clustering algorithm is used to perform the data clustering with transaction similarity values. The Dynamic Ensemble Membership Selection (DEMS) scheme is adapted to enhance the ensemble selection process with structure and data independent models.

4. IMPLEMENTATION

4.1 Module Description

The cancer data clustering system is designed to perform data partitioning on the cancer diagnosis data values. The Incremental Semi-Supervised Cluster Ensemble (ISSCE) scheme is applied for the clustering process. Incremental Ensemble Membership Selection (IEMS) scheme is used for the cluster ensemble selection process. The relationship levels are estimated with the similarity functions. The Dynamic Ensemble Membership Selection (DEMS) is used to perform the ensemble selection for structure and complexity independent data values. The DEMS scheme is integrated with Partition Around Medoids (PAM) clustering algorithm.

The data cleaning module is designed to update noise data values. The ensemble selection module is designed to identify the cluster initial ensembles. The local similarity estimation process is carried out with the ensembles that are identified with the incremental model. The global similarity estimation process is carried out with the dynamic ensemble member selection model based data values. The Incremental Semi-Supervised Cluster Ensemble (ISSCE) approach is used in the ISSCE clustering process. The DEMS based PAM clustering approach is adapted in the dynamic membership based clustering process.

Data Cleaning Process

The cancer diagnosis details are imported from textual data files. The textual data contents are parsed and categorized with its property. Redundant and noise records are identified and maintained separately. The data values are parsed and transferred into the Oracle database. Redundant data values are removed from the database. Missing elements are assigned using aggregation based substitution mechanism. Cleaned data values are referred as optimal data sets.

Ensemble Selection

Cluster ensembles are selected from the transaction collections. Cluster count is collected from the user. The ensemble selection is carried out with the Incremental Ensemble Membership Selection (IEMS) scheme. The Similarity Function is used to compare the transactions. The ensembles are identified for each cluster levels.

Local Similarity Analysis

The local similarity analysis module is designed to perform the transaction similarity estimation process. Incremental ensemble based model is adapted to the similarity estimation process. The continuous data values are converted into categorical data. Median values are used in the conversion process. Similarity function is used in the relationship analysis. The similarity values are updated into the dissimilarity matrix.

Global Similarity Analysis

The similarity analysis is performed to estimate the transaction relationship. Independent data similarity is designed for binary, categorical and continuous data types. Similarity function is tuned to find similarity for all type of data values. Vector and link models are integrated for relationship analysis. The Dynamic Ensemble Membership Selection (DEMS) scheme is used in the ensemble member identification process. The similarity estimation is performed with the finalized ensemble values. Structure independent ensemble membership selection is used in the system.

ISSCE Clusters

The Incremental Semi-Supervised Cluster Ensemble (ISSCE) approach is adapted to perform the data clustering process. The Incremental Ensemble Membership Selection (IEMS) algorithm is used in the ensemble member selection process. The Similarity Function (SF) is applied to estimate the transaction similarity values. Local relationships are considered in the similarity estimation process. The Similarity matrix is composed with incomplete similarity details. The similarity intervals are used to partition the data values. The clustering process is performed with the user provided cluster count values. The cluster list shows the list of clusters with the transaction count. The cluster details form shows the cluster name and its associated transactions.

PAM Clusters with DEMS

The Partition Around Medoids (PAM) algorithm is used for the clustering process. The dissimilarity is minimized in the PAM algorithm. The Dynamic Ensemble Membership Selection (DEMS) scheme is employed to select the ensemble members with structure independent mechanism. Data set complexity is also considered in the DEMS scheme. The Similarity Functions is also tuned for the dynamic ensemble member selection process. The Dynamic Ensemble Membership Selection (DEMS) scheme is integrated with the PAM clustering algorithm. The clustering process is carried out with the cluster count specified by the user.

4.2 Implementation Procedure

The system implementation process replaces the existing system with the proposed system. Different methods are considered in the system implementation process. Parallel running, pilot running, staged changeover and direct changeover methods are considered for implementation

"PAM Cluster " is used as the system data source name for the project. The tables are created in the database. The user interface directly connected with the back end software.

4.3 Datasets

The clustering system is analyzed using the breast cancer patient data sets collected from the University of California Irwin (UCI) machine learning repository. The dataset is downloaded from http://archive.ics.uci.edu/ml/datasets.html. The diagnosis details are collected from patients from different countries. The dataset contains 1000 transactions with 15 attributes. Missing values are replaced in preprocess. Aggregation based data substitution mechanism is used for the data preprocess. Redundant transactions are removed from the datasets during the preprocess.

Table -1: Attribute details for breast cancer data set

S.No	Attribute Name	Description
1	Pid	***Patient identification number***
2	CT	***Clump Thickness***
3	UCS	***Uniformity of Cell Size***
4	CS	**1) Uniformity of Cell Shape**
5	MA	***Marginal Adhesion***
6	SECS	***Single Epithelial Cell Size***
7	BN	***Bare Nuclei***
8	BC	***Bland Chromatin***
9	NN	***Normal Nucleoli***
10	MI	***Mitoses***
11	Class	***Class***

4.4 Purity

The purity of a cluster represents the fraction of the cluster corresponding to the largest class of documents assigned to that cluster; thus, the purity of the cluster j is defined as

$$Purity(j) = \frac{1}{n_j} \max_i (n_{ij})$$

The overall purity of the clustering result is a weighted sum of the purity values of the clusters as follows:

$$Purity = \sum_j \frac{n_j}{n} Purity(j)$$

Table-2: Purity analysis of ISSCE and DEMS based PAM Schemes

Transactions	ISSCE	DEMS based PAM
200	0.826	0.951
400	0.839	0.965
600	0.851	0.972
800	0.863	0.989
1000	0.875	0.997

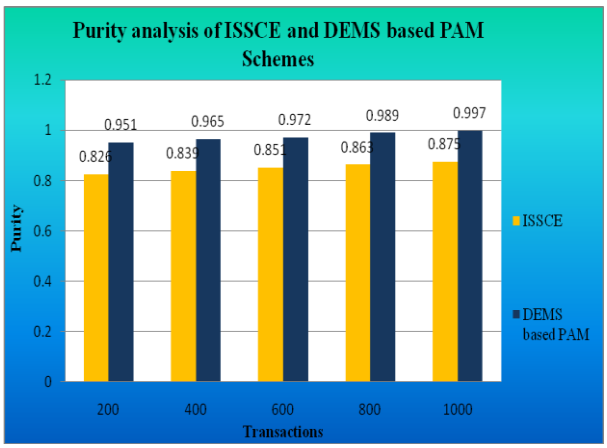

Fig-2: Purity analysis of ISSCE and DEMS based PAM Schemes

The Purity analysis between the Incremental Semi-Supervised Cluster Ensemble (ISSCE) and Dynamic Ensemble Membership Selection based Partition Around Medoids (DEMS based PAM) clustering schemes. The analysis result shows that the Incremental Semi-Supervised Cluster Ensemble (ISSCE) scheme increases the cluster accuracy level 10% than the Dynamic Ensemble Membership Selection based Partition Around Medoids (DEMS based PAM) (DEMS based PAM) clustering scheme.

4.5 Separation Index

Separation Index (SI) is another cluster validity measure that utilizes cluster centroids to measure the distance between clusters, as well as between points in a cluster to their respective cluster centroid. It is defined as the ratio of average within-cluster variance to the square of the minimum pairwise distance between clusters:

$$SI = \frac{\sum_{i=1}^{NC}\sum_{xj\in ci} dist(x_j, m_i)^2}{N_D \min_{1\le r,s\le N_c} dist(m_r, m_s)\}^2}$$

$$= \frac{\sum_{i=1}^{N_c}\sum_{xj\in ci} dist(x_j, m_i)^2}{N_D \cdot dist_{\min}^2}$$

Where m_i is the centroid of cluster c_i, and distmin is the minimum pairwise distance between cluster centroids.

The Incremental Semi-Supervised Cluster Ensemble (ISSCE) and Dynamic Ensemble Membership Selection based Partition Around Medoids (DEMS based PAM) clustering schemes. The analysis result shows that the Incremental Semi-Supervised Cluster Ensemble (ISSCE) scheme increases the inter cluster distance level 30% than the Dynamic Ensemble Membership Selection based Partition Around Medoids (DEMS based PAM) (DEMS based PAM) clustering scheme.

Table-3: Separation Index Analysis of ISSCE and DEMS based PAM Schemes

Transactions	ISSCE	DEMS based PAM
200	2.806	4.705
400	2.716	4.365
600	3.796	6.728
800	6.484	9.213
1000	8.719	11.457

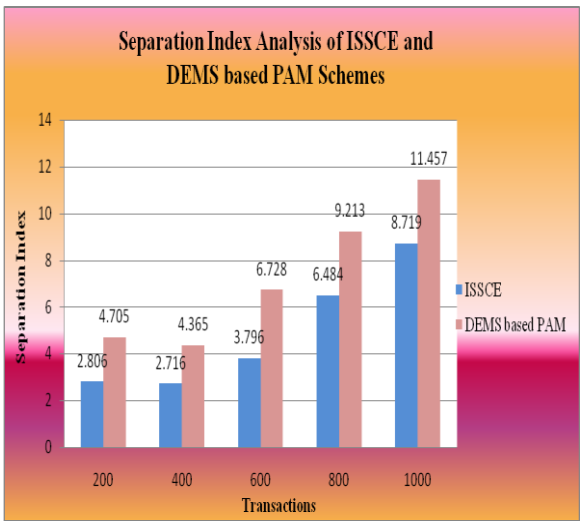

Fig-3: Separation Index Analysis of ISSCE and DEMS based PAM Schemes

5. Results and Discussion

The medical data analysis system is developed to partition the breast cancer patient diagnosis details. The Incremental Semi-Supervised Cluster Ensemble (ISSCE) approach is used for the data clustering process. The

Incremental Ensemble Membership Selection (IEMS) mechanism is used to select the members for the cluster ensembles. The Similarity Function (SF) is used to find out the relationship between the transactions. The Dynamic Ensemble Membership Selection (DEMS) scheme is build to identify the ensembles with structure and complexity independency. The DEMS scheme is integrated with the Partition Around Medioids clustering algorithm to produce better cluster results. The system performance is evaluated with Incremental Semi-Supervised Cluster Ensemble (ISSCE) and Dynamic Ensemble Membership Selection based Partition Around Medoids (DEMS based PAM) clustering schemes. The system is tested with three performance parameters to measure the cluster quality levels. They are F-measure, purity and separation index levels. The system is tested with different data intervals.

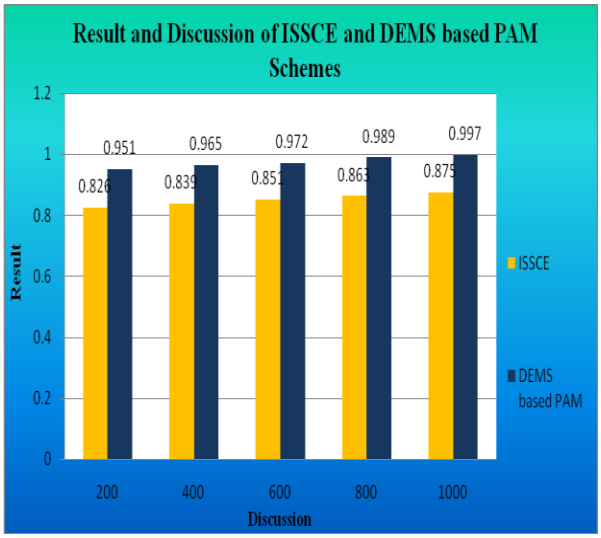

Fig-4: Result and Discussion of ISSCE and DEMS based PAM Schemes

6 .CONCLUSIONS AND FUTURE ENHANCEMENT

The cancer data clustering system is build to partition the breast cancer diagnosis data values. The Incremental Semi-Supervised Cluster Ensemble (ISSCE) approach is used for the data clustering process with ensemble models. The ensemble identification process is performed with Incremental Ensemble Membership Selection (IEMS) scheme. The Similarity Function (SF) is applied to estimate the relationship values. The Dynamic Ensemble Membership Selection (DEMS) mechanism is applied to identify the cluster ensembles with structure and data complexity independent models. The Partition Around Medoids (PAM) clustering scheme is integrated with Dynamic Ensemble Membership Selection (DEMS) mechanism. The system can be enhanced with the following features.

- The clustering scheme can be improved to support clustering under distributed database environment.
- The clustering model can be adapted to perform clustering on data stream based data source model.
- The system can be adapted to support hierarchical clustering process.
- The fuzzy logic and genetic algorithm models can be integrated with the system to improve the cluster accuracy levels.

REFERENCES

1. N. Bassiou, V. Moschou and C. Kotropoulos, “Speaker Diarization Exploiting the Eigengap Criterion and Cluster Ensembles”, IEEE/ACM Transactions on Audio, Speech, and Language Processing, Vol. 18, No. 8, pp. 2134-2144, 2010.
2. Dong Huang, Jian-Huang Lai and Chang-Dong Wang, “Robust Ensemble Clustering Using Probability Trajectories”, Journal of Latex Class Files, Vol. 13, No. 9, September 2014
3. H. Wang, T. Li, T. Li and Y. Yang, “Constraint Neighborhood Projections for Semi-Supervised Clustering”, IEEE Transactions on Cybernetics, Vol. 44, No. 5, pp. 636-643, 2014.
4. T. Wang, “CA-Tree: A Hierarchical Cluster for Efficient and Scalable Co Association-based Cluster Ensembles”, IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, Vol. 41, No. 3, pp. 686-698, 2011.