Available online at: http://www.ijmtst.com/vol6issue04.html

International Journal for Modern Trends in Science and Technology

ISSN: 2455-3778 :: Volume: 06, Issue No: 04, April 2020

# Incremental Semi-Supervised Clustering Ensemble for High Dimensional Data Clustering

M Pavithra[1] | Dr. R M S Parvathi[2]

[1]Assistant Professor, Department of CSE, Jansons Institute of Technology, Coimbatore, India.
[2]Professor, Department of CSE, Sri Ramakrishna Institute of Technology, Coimbatore, India.

**To Cite this Article**

M Pavithra, Dr. R M S Parvathi, "Incremental Semi-Supervised Clustering Ensemble for High Dimensional Data Clustering", *International Journal for Modern Trends in Science and Technology*, Vol. 06, Issue 04, April 2020, pp.:41-48.

**Article Info**

Received on 03-March0-2020, Revised on 23-March-2020, Accepted on 25-March-2020, Published on 31-March-2020.

## ABSTRACT

*Traditional cluster ensemble approaches have three limitations: (1) they do not make use of prior knowledge of the datasets given by experts. (2) Most of the conventional cluster ensemble methods cannot obtain satisfactory results when handling high dimensional data. (3) All the ensemble members are considered, even the ones without positive contributions. In order to address the limitations of conventional cluster ensemble approaches, we first propose an incremental semi-supervised clustering ensemble framework (ISSCE) which makes use of the advantage of the random subspace technique, the constraint propagation approach, the proposed incremental ensemble member selection process, and the normalized cut algorithm to perform high dimensional data clustering. The random subspace technique is effective for handling high dimensional data, while the constraint propagation approach is useful for incorporating prior knowledge [2]. The incremental ensemble member selection process is newly designed to judiciously remove redundant ensemble members based on a newly proposed local cost function and a global cost function, and the normalized cut algorithm is adopted to serve as the consensus function for providing more stable, robust, and accurate results. Then, a measure is proposed to quantify the similarity between two sets of attributes, and is used for computing the local cost function in ISSCE. Next, we analyze the time complexity of ISSCE theoretically [3]. It works well on datasets with very high dimensionality, and outperforms the state-of-the-art semi-supervised clustering ensemble approaches.Clustering techniques are applied to partition the transaction data values. High dimensional support, prior knowledge usage and equal membership priority are the key factors in the traditional cluster ensemble approach. Incremental Semi Supervised Cluster Ensemble (ISSCE) approach is built to solve the limitations of conventional cluster ensemble approaches [4]. The ISSCE approach uses the steps in random subspace technique, the constraint propagation approach, the incremental ensemble member selection process and the normalized cut algorithm to perform high dimensional data clustering. The random subspace technique is effective for handling high dimensional data. The constraint propagation approach is useful for incorporating prior knowledge. The incremental ensemble member selection process is applied to judiciously remove redundant ensemble members based on a local cost function and a global cost function.The normalized cut algorithm is adopted to serve as the consensus function for providing more stable, robust and accurate results [5]. A measure is applied to quantify the similarity between two sets of attributes, and is used for computing the local cost function in ISSCE. The incremental semi supervised clustering ensemble framework (ISSCE) approach is enhanced to support structure based parameter selection process. Datasets complexity is also integrated with the parameter selection process. Membership rearrangement mechanism is adapted to handle the incremental membership selection process. Member and ensemble weight measure is also applied to discover the importance of the cluster*

*ensembles [6]. The cluster ensemble model is integrated with the Partition around Medoids (PAM) clustering scheme. The system also increases the clustering accuracy and scalability levels.*

***KEYWORDS:*** *Cancer gene expression profile, Cluster ensemble, Clustering analysis, Random subspace, Semi-supervised clustering*



## I. INTRODUCTION

Clustering is an important technique of exploratory data mining, which divides a set of objects into several groups in such a way that objects in same group are more similar with each other in some sense than with the objects in other groups. It has been widely used in different disciplines and applications, such as machine learning, pattern recognition, data compression, image segmentation [8], time series analysis [7], information retrieval, spatial data analysis [1] and biomedical research [3]. Moreover, as data's variety and scale increase rapidly, and the prior knowledge about the data is usually limited, clustering has been a challenging task. The most popular example of density-based clustering is DBSCAN in which only the objects whose density is greater than the given thresholds are connected together to form a cluster. However, the proper thresholdsetting varies with different data sets; there is still no effective method to preassign these thresholds. The spectral clustering based algorithm does not make assumptions on the forms of the clusters; it utilizes the spectrum of the similarity matrix of the data to map the data into a lower dimensional space in which the objects can be easily clustered by traditional clustering techniques. Comparing to the traditional algorithms, such as kMeans and single-linkage, this kind of clustering algorithm is useful in non-convex boundaries and performs empirically very well [4]. The first few eigenvalues can be used to determine the number of clusters and reduce the dimension of data these first eigenvectors cannot successfully cluster objects that contain structures with different sizes and densities.

Cluster ensemble, also referred to as consensus clustering, is one of the important research directions in the area of ensemble learning, which can be divided into two stages: the first stage aims at generating a set of diverse ensemble members, while the objective of the second stage is to select a suitable consensus function to summarize the ensemble members and search for an optimal unified clustering solution. To attain these objectives, [1] first proposed a knowledge reuse framework which integrates multiple clustering solutions into a unified one. After that, a number of researchers followed up Strehl's work, and proposed different kinds of cluster ensemble approaches [5]-[1]. While there are different kinds of cluster ensemble techniques, few of them consider how to handle high dimensional data clustering, and how to make use of prior knowledge. High dimensional datasets have too many attributes relative to the number of samples, which willlead to the overfitting problem. Most of the conventional cluster ensemble methods do not take into account how to handle the overfitting problem, and cannot obtain satisfactory results when handling high dimensional data. Our method adopts the random subspace technique to generate the new datasets in a low dimensional space, which will alleviate this problem. There are also other research works which study the properties of the cluster ensemble theoretically, such as the stability of k-means based cluster ensemble [2], the efficiency of the cluster ensemble [2], the convergence property of consensus clustering [3], the scalability property of the cluster ensemble [4], the effectiveness of cluster ensemble methods [5], and so on. Cluster ensemble approaches have been applied to different areas, such as bioinformatics [6][7], image segmentation [8], language processing [4], Internet security [3], and so on. Recently, some researchers realized that not all the ensemble members contribute to the final result, and investigate how to select a suitable subset of members to obtain better results [1]-[5]. For example, Yu et al. [3]-[4] treated the ensemble members as features, and explored how to use suitable feature selection techniques to choose the ensemble members. In summary, most of the cluster ensemble approaches only consider using a similarity score or feature selection technique to remove the redundant ensemble members, and few ofthem study how to apply an optimization method to search for a suitable subset of ensemble members.

## II. RELATEDWORK

Generally, data stream mining refers to the mining tasks that are conducted on a sequence of rapidly arriving data records. As the environment where the data are collected may change dynamically, the data distribution may also change accordingly. This phenomenon, referred to as concept drift [3], [4], is one of the most important challenges in data stream mining. A data stream mining technique should be capable of constructing and dynamically updating a model in order to learn dynamic changes of data distributions, i.e., to track the concept drift. Concept drift is formally defined as the change of joint distribution of data, i.e., p(x, y), where x is the feature vector and y is the class label. Over the past few decades, concept drift has been widely studied [5]. The majority of the previous works focus on the concept drift caused by the change in class-conditional probability distribution, i.e., p(xjy). In comparison, class evolution, which is another factor that induces concept drift, has attracted relatively less attention. Briefly speaking, class evolution is concerned with certain types of change in the prior probability distribution of classes, i.e., p(y) and usually corresponds to the emergence of a novel class and the disappearance of an outdated class. Class evolution occurs frequently in practice. For example, new topics frequently appear on Twitter and outdated topics are forgotten with time. The number of classes may change when class evolution happens; the model needs to be adapted not only to capture the distribution of existing classes, but also to identify that of the novel classes. At the same time, the effects of disappeared classes need to be removed from the model. Hence, in comparison to the change of classconditional probability, class evolution brings additional challenges to data stream mining. In literature, a few approaches have been proposed to address class evolution problems, e.g., Learn++.NC, ECSMiner [2] and CLAM [6]. Although they have shown promising performance, they implicitly assume that classes emerge or disappear in a transient manner. 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 [1], [2], [7], [1], [5]. 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. An ensemble clustering approach is constructed with sparse graph representation and probability trajectory analysis.

## III. CLUSTER ENSEMBLE APPROACHES

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 [2].Conventional cluster ensemble approaches have several limitations: (1) they do not consider how to make use of prior knowledge given by experts, which is represented by pairwise constraints. Pairwise constraints are often defined as the must-link constraints and the cannot-link constraints [3]. The must-link constraint means that two feature vectors should be assigned to the same cluster, while the cannot-link constraints means that two feature vectors cannot be assigned to the same cluster. (2) Most of the cluster ensemble methods cannot achieve satisfactory results on high dimensional datasets. (3) Not all the ensemble members contribute to the final result [4].

## IV. PROPOSED WORK

### 4.1 INCREMENTAL SEMI-SUPERVISED CLUSTERING ENSEMBLE FRAMEWORK (ISSCE)

The incremental semi-supervised clustering ensemble framework (ISSCE) is designed to remove the redundant ensemble members. When compared with traditional semi supervised clustering algorithm, ISSCE is characterized by the incremental ensemble member selection process based on a global objective function and a local objective function, which selects ensemble members progressively. The local objective function is calculated based on a newly designed similarity function which determines how similar two sets of attributes are in the subspaces [3]. Next, the computational cost and the space consumption of ISSCE are analyzed theoretically. Multiple semi-supervised clustering ensemble approaches are analyzed over different datasets. The experiment results show the improvement of ISSCE over traditional semi supervised clustering

ensemble approaches or conventional cluster ensemble methods. 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 [4]. Non-parametric tests are used to compare multiple semi supervised clustering ensemble approaches over different datasets.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 [5]. 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).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 [6]. 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 [7].

### 4.2 INCREMENTAL ENSEMBLE MEMBERSHIP SELECTION (IEMS)

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 [1]. IEMS considers the ensemble members one by one and calculates the objective function (Ib) for each clustering solution Ib generated by E2CP with respect to the subspace Ab in the first step. In the second step, it sorts all the ensemble members in b in ascending order according to the corresponding values.

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

Where d(pi, ) denotes the Euclidean distance between the feature vectors pi and 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 [3]. Given the original ensemble I~ and the new ensemble I~ the local objective function xb for the local b-th ensemble member (Ab, xb) £ with respect to the ensemble member (At, xt) £ I~ is defined as follows:

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

where Δ(Ib) denotes the global objective function for the clustering solution Ib and S(Ab,At) denotes the similarity function between two subspaces Ab and At. Given the subspaces Ab and At, the set of attributes in these subspaces can be represented by Gaussian mixture models (GMMs) [5].

### 4.3 PARTITION AROUND MEDOIDS (PAM) 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 [2]. 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 [4]. 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. Dp, the dissimilarity between p and the closest object in S and Ep,the dissimilarity between p and the second closest object in S. The Partition around Medoids (PAM) algorithm is used for the clustering

process [6]. 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 [5].

### 4.4 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[2]. 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. The clustering methods are enhanced with the ensembles based model to increase the accuracy levels [4]. 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 [5]. The Dynamic Ensemble Membership Selection (DEMS) scheme is adapted to enhance the ensemble selection process with structure and data independent models.

### 4.5 DYNAMIC ENSEMBLE MEMBERSHIP SELECTION (DEMS)

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 [1]. 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 [3]. 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 [4]. 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 [6].

## V. DATASETS

The data sets used in our experiments include six UCI data sets1. Here is some basic information of those data sets. Table 5 summarizes the basic information of those data sets.

• Balance. This data set was generated to model psychological experimental results. There are totally
625 examples that can be classified as having the balance scale tip to the right, tip to the left, or be balanced.

• Iris. This data set contains 3 classes of 50 instances each, where each class refers to a type of iris
plant.

• Ionosphere. It is a collection of the radar signals belonging to two classes. The data set contains
351 objects in total, which are all 34-dimensional.

• Soybean. It is collected from the Michalski's famous soybean disease databases, which contains
562 instances from 19 classes.

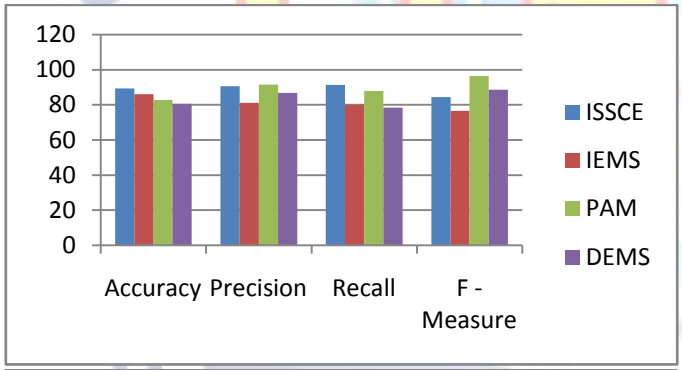
| Datasets | Size | Classes | Dimensions |
|---|---|---|---|
| Balance | 625 | 3 | 4 |
| Iris | 150 | 3 | 4 |
| Ionosphere | 351 | 2 | 34 |
| Soybean | 562 | 19 | 35 |

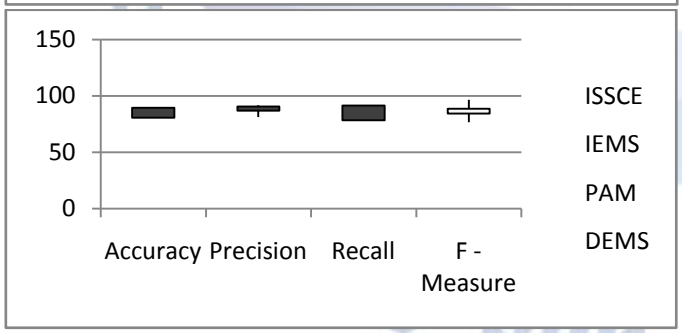
## VI. EXPERIMENTAL RESULTS

### 6.1 BALANCE DATASET RESULTS

| Balance Dataset | | | | |
|---|---|---|---|---|
| Algorithm | Accuracy | Precision | Recall | F - Measure |
| ISSCE | 89.45 | 90.67 | 91.45 | 84.45 |
| IEMS | 86.05 | 81.23 | 79.98 | 76.67 |
| PAM | 82.77 | 91.56 | 87.88 | 96.56 |
| DEMS | 80.56 | 86.78 | 78.34 | 88.67 |

The above graph shows that performance of Balance dataset. The Accuracy of ISSCE algorithm is 89.45 which is higher when compare to other three (IEMS, PAM, DEMS) algorithms. The Precision of PAM algorithm is 91.56 which is higher when compare to other three (ISSCE, IEMS, DEMS) algorithms. The Recall of ISSCE algorithm is 91.45 which is higher when compare to other three (IEMS, PAM, DEMS) algorithms. The F-Measure of PAM algorithm is 96.56 which is higher when compare to other three (ISSCE, IEMS, DEMS) algorithms.

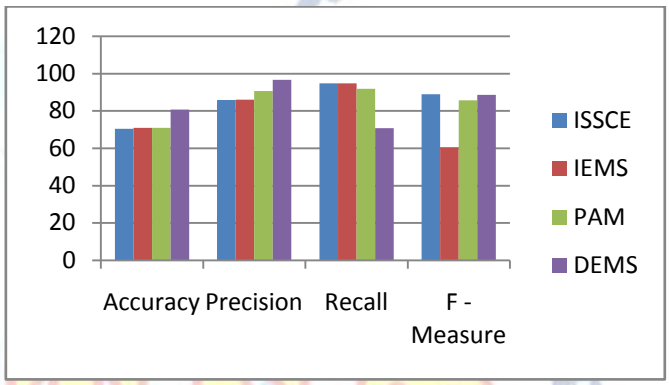



### 6.2 IRIS DATASET RESULTS

| Iris Dataset | | | | |
|---|---|---|---|---|
| Algorithm | Accuracy | Precision | Recall | F - Measure |
| ISSCE | 70.45 | 85.91 | 94.77 | 88.89 |
| IEMS | 70.91 | 86.08 | 94.78 | 60.56 |
| PAM | 70.92 | 90.67 | 91.89 | 85.78 |
| DEMS | 80.67 | 96.67 | 70.78 | 88.67 |

The above graph shows that performance of Iris dataset. The Accuracy of DEMS algorithm is 80.67 which is higher when compare to other three (ISSCE, IEMS, PAM) algorithms. The Precision of DEMS algorithm is 96.67 which is higher when compare to other three (ISSCE, IEMS, PAM) algorithms. The Recall of IEMS algorithm is 94.78 which is higher when compare to other three (ISSCE, DEMS, PAM) algorithms. The F-Measure of ISSCE algorithm is 88.89 which is higher when compare to other three (DEMS, IEMS, PAM) algorithms.



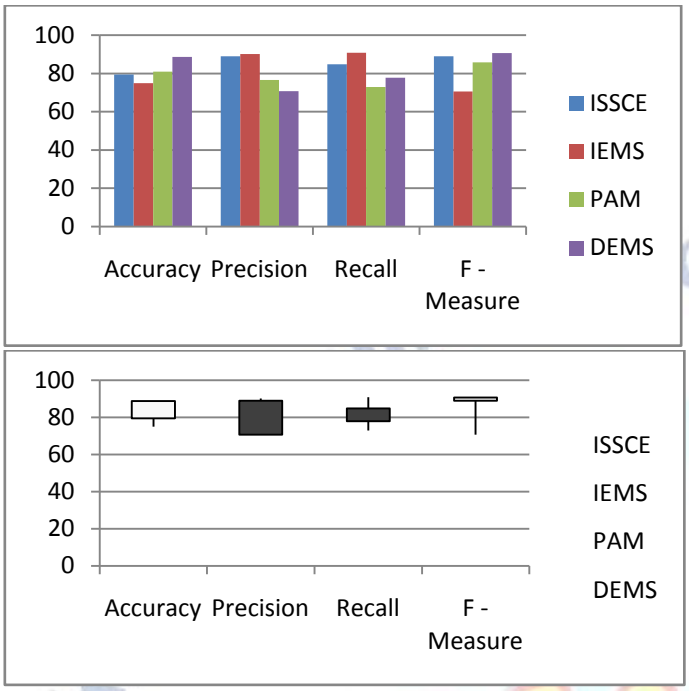


### 6.3 IONOSPHERE DATASET RESULTS

| Ionosphere Dataset | | | | |
|---|---|---|---|---|
| Algorithm | Accuracy | Precision | Recall | F - Measure |
| ISSCE | 79.45 | 88.91 | 84.77 | 88.89 |
| IEMS | 74.91 | 90.08 | 90.78 | 70.56 |
| PAM | 80.98 | 76.67 | 72.89 | 85.78 |
| DEMS | 88.67 | 70.67 | 77.78 | 90.67 |

The above graph shows that performance of Ionosphere dataset. The Accuracy of DEMS algorithm is 88.67 which is higher when compare to other three (ISSCE, IEMS, PAM) algorithms. The Precision of IEMS algorithm is 90.08 which is higher when compare to other three (ISSCE, PAM, DEMS) algorithms. The Recall of IEMS algorithm is 90.78 which is higher when compare to other three (ISSCE, DEMS, PAM) algorithms. The F-Measure of DEMS algorithm is 90.67 which is higher when

compare to other three (ISSCE, IEMS, PAM) algorithms.



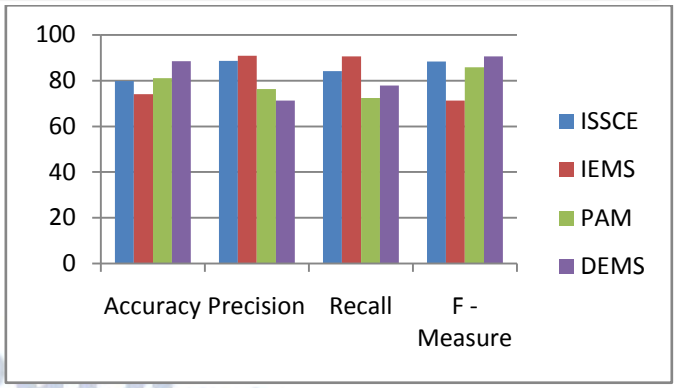


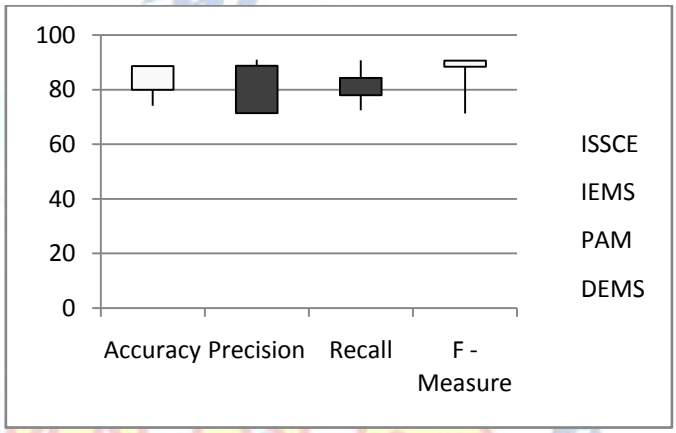


### 6.4 SOYBEAN DATASET RESULTS

| Soybean Dataset | | | | |
|---|---|---|---|---|
| Algorithm | Accuracy | Precision | Recall | F - Measure |
| ISSCE | 79.89 | 88.65 | 84.23 | 88.34 |
| IEMS | 74.03 | 90.89 | 90.67 | 71.23 |
| PAM | 81.08 | 76.32 | 72.45 | 85.9 |
| DEMS | 88.54 | 71.32 | 77.89 | 90.56 |

The above graph shows that performance of Soybean dataset. The Accuracy of DEMS algorithm is 88.54which is higher when compare to other three (ISSCE, IEMS, PAM) algorithms. The Precision of IEMS algorithm is 90.89 which is higher when compare to other three (ISSCE, PAM, DEMS) algorithms. The Recall of IEMS algorithm is 90.67 which is higher when compare to other three (ISSCE, PAM, DEMS) algorithms. The F-Measure of DEM Salgorithm is 90.56 which is higher when compare to other three (ISSCE, IEMS, PAM) algorithms.




## VII. CONCLUSION

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 [2]. The Partition around Medoids (PAM) clustering scheme is integrated with Dynamic Ensemble Membership Selection (DEMS) mechanism [4]. 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.

## VIII. FUTURE ENHANCEMENTS

In this paper, we propose a new semi-supervised clustering ensemble approach, which is referred to as the incremental semi supervised clustering ensemble approach (ISSCE). Our major contribution is the development of an incremental ensemble member selection process based on a global objective function and a local objective function [3]. In order to design a good local objective function, we also propose a new similarity function to quantify the extent to which two sets of attributes in the subspaces are similar to each other [5]. We conduct experiments on 6 real-world datasets from the UCI machine learning repository and 12 real-world datasets of cancer gene expression profiles, and obtain the following observations:

1. The incremental ensemble member selection process is a general technique which can be used in different semi-supervised clustering ensemble approaches.
2. The prior knowledge represented by the pairwise constraints is useful for improving the performance of ISSCE.
3. ISSCE outperforms most conventional semi-supervised clustering ensemble approaches on a large number of datasets, especially on high dimensional datasets.

In the future, we shall perform theoretical analysis to further study the effectiveness of ISSCE, and consider how to combine the incremental ensemble member selection process with other semi-supervised clustering ensemble approaches [6]. We shall also investigate how to select parameter values depending on the structure/complexity of the datasets.

## REFERENCES

[1] Yang, Y. and K. Chen, "Temporal data clustering via weighted clustering ensemble with different representations," IEEE Trans. on KDE, 23(2): 307-320, 2011.

[2] Iam-On, N., T. Boongoen, S. Garrett and C. Price, "A link-based approach to the cluster ensemble problem," IEEE Trans. on PAMI, 33(12): 2396-2409, 2011.

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

[4] Dennis Ebenezer and M. Suganya, "A Survey On Data Partitioning And Cluster Ensemble Techniques", Elysium Journal of Engineering Research and Management, 3(6): 1-6, 2016.

[5] 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.

[6] 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.

[7] 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.

[8] 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.