ISSN: 2277-9655
(I2OR), Publication Impact Factor: 3.785

IJESRT

INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

CLUSTERING TECHNIQUES FOR DNA MICROARRAY BASED GENE EXPRESSION DATA SURVEY

K.Sathishkumar*, Dr.V. Thiagarasu, E.Balamurugan
* Assistant Professor,Dept. of Information Technology. Gobi Arts & Science College (Autonomous), Gobichettipalayam,India
Associate Professor,Dept. of Computer Science,Gobi Arts & Science College(Autonomous),Gobichettipalayam,India
Associate Professor,Department of (IT & IS) BlueCrest College, Accra-North, Ghana

ABSTRACT

The advent of DNA microarray technology has enabled biologists to monitor the expression levels (MRNA) of thousands of genes simultaneously. Cluster analysis seeks to partition a given data set into groups based on specified featuresso that the data points within a group are more similar to each other than the points in differentgroups. Many conventional clustering algorithms have been adapted or directly applied to gene expression data, and also new algorithms have recently been proposed specifically aiming at gene expression data. These clustering algorithms have been proven useful for identifying biologicallyrelevant groups of genes and samples.In this survey, the various approaches to gene expression data analysis using clustering techniques are addressed. Then the performances of various existing clustering algorithms under each of these approachesare discussed. The specific challenges pertinent toeach clustering category and introduce several representative approaches are presented. And also theproblems of cluster validation are discussed and review various methods to assess the qualityand reliability of clustering results. Finally, this paper is concluded and suggests the promisingtrends in this field.

KEYWORDS: microarray technology, gene expression data, clustering.

INTRODUCTION

Microarray is a comparatively new technology and a chip-based highthroughput technology to investigate the expression levels of thou-sands of genes simultaneously, compared with the traditional approach to genomic research. Expression data are collected via eitherexperiments in a time series during a biological process or experiments of different tissue samples [1]. A gene expression datasetis organized in a expression matrix, where the rows represent theexpression profiles or patterns of genes, and the columns representthe expression profiles of samples. Data preprocessing is requiredbefore any clustering analysis, because the original gene expression matrix contains noise, missing values and systematic variations.By analyzing the gene expression across multiple experiments, co-regulated gene with similar biological functions and their interactions, or their characteristic of the states, diseases, or phenotypesrepresented by groups of samples may be discovered.

The eminence of DNA microarray technology is the aptitude to be used to simultaneously monitor and study the expression levels of thousands of genes, relationship between genes, their functions and classifying genes or samples that perform in a parallel or synchronized manner during imperative biological processes[2]. Functional genomics can be better implicit when the veiled patterns in gene expression data is elucidated, however, it is very challenging to comprehend and construe this due to the complexity of biological networks and large number of genes.

The most important area of microarray bioinformatics is possibly the data clustering analysis. Clustering is an exceptional preference for initial data analysis and data mining processes.To perceive and identify appealing patterns of expression across multiple genes and experiments, reveal natural structures and compress high-dimensional array data clustering must be ascertained to allow easier management of data set. This data reduction method is a simple tool yet powerful method of organizing genes based on their interdependence behaving similarly over the different conditions in different mutants, patients or at different time points in a time series during an experiment with similar

expression patterns and properties into a set of disjoint groups based on specific features so that the underlying structures can be acknowledged and explored. Clustering, also known as unsupervised learning has been usedfor decades in many fields, such as image processing, data miningand artificial intelligence and in recent years, has benefited microarray gene expression data analysis in genomic research[3][4]. Thegoal of the clustering analysis is to group individual objects or samples in a population within which the objects are more similar toeach other than those in other clusters.

The rest of the survey is organized as follows: Section II describes the formulation of the problem and its complexity. In Section III focuses to analyze various available clustering techniques to determine the best suitable and effective clustering approaches for gene expression data. Section IV describes the inference from existing clustering techniques. Section V concludes the paper.

PROBLEM FORMULATION

Gene expression data are generated by DNA chips andother microarray techniques and they are often presented as matrices of expression levels of genes under different conditions (including environments, individuals, and tissues). One of the major objectives of gene expression dataanalysis is to identify groups of genes having similar expression patterns over the full space or subspace of conditions. It may result in the discovery of regulatory patterns or condition similarities.

Generally co-expressed genes, which are members of the same clusters, are expected to have similarfunctions.

Most researchers address this issue by: (i) using either partitional, hierarchical, density-based, model-based or subspace clustering algorithms, based on the proximity between genes or conditions in the expression matrix or (ii) giving equal weights to all conditions or all genesin the computation of gene similarity and vice versa. However, if the proximity measure is not properly selected, it may lead to the discovery of some similar groups of genes at the expense of obscuring other similar groups. Hence, it may be necessary to recover information lost due to over-simplificationof similarity and grouping computation, which may reveal the involvement of a gene or a condition in more than one group.

LITERATURE SEURVEY

In this section, the existing clustering techniques are discussed. And also the advantage and disadvantages of clustering techniques discussed.

Divisive Correlation Clustering Algorithm (DCCA) for grouping of genes: detecting varying Patterns in expression profiles

De K.R and Bhattacharya .A et al., had presented the Divisive Correlation Clustering Algorithm (DCCA) that is suitable for finding a group of genes having similar pattern of variation in their expression values[5]. To detect clusters with high correlation and biological significance used the correlation clustering concept introduced by Bansal et al. In this algorithm DCCA produces a clustering solution without taking number of clusters to be created as an input. DCCA uses the correlation matrix in such a way that all genes in a cluster have highest average correlation with genes in that cluster. The performance of the DCCA, have applied DCCA and some well-known conventional methods to an artificial dataset, and nine gene-expression datasets, and compared the performance of the algorithms. The clustering results of the DCCA are found to be more significantly relevant to the biological annotations than those of the other methods. All these facts show the superiority of the DCCA over some others for the clustering of gene-expression data.

Advantage

- It has higher degree of accuracy than other conventional clustering algorithms.
- The computational cost of the DCCA is less.

Disadvantage

- The method is not suitable for low-dimensional data.
- It cannot guarantee absence of repulsion inside a cluster or highest average attraction between genes inside clusters.

Phylogenetic reconstruction using an unsupervised neural network that adopts the topology of a phylogenetic tree

Dopazo .J and Carazo .J et al., had proposed a new type of unsupervised, growing, self-organizing neural network that expands itself by following the taxonomic relationships that exist among the sequences being classified. The binary tree topology of this neutral network, contrary to other more classical neural network topologies, permits an efficient classification of sequences[6]. The growing nature of this procedure allows stopping it at the desired taxonomic level without the necessity of waiting until a complete phylogenetic tree is produced. This novel approach presents a number of other interesting properties, such as a time for convergence which is, approximately, a lineal function of the number of sequences. Computer simulation and a real example show that the algorithm accurately finds the phylogenetic tree that relates the data.

Advantage

- The algorithm is highly accurate.
- The neural network presented here an excellent tool for phylogenetic analysis of a large number of sequences.

Disadvantage

- The execution time high.
- The efficiency of algorithm is less.

Fuzzy C means method for clustering microarray data Bioinformatics

DoulayeDembele and Philippe Kastner et al., had explained clustering analysis of data from DNA microarray hybridization studies is essential for identifying biologically relevant groups of genes[7]. Partitional clustering methods such as K-means or self-organizing maps assign each gene to a single cluster. However, these methods do not provide information about the influence of a given gene for the overall shape of clusters. Here a fuzzy partitioning method, Fuzzy C-means (FCM) is applied to attribute cluster membership values to genes. We proposed an empirical method, based on the distribution of distances between genes in a given data set, to determine an adequate value for m. By setting threshold levels for the membership values, genes which are tigthly associated to a given cluster can be selected. Using a yeast cell cycle data set as an example, we show that this selection increases the overall biological significance of the genes within the cluster.

Advantage

- The C items of the fuzzy categories and minimize the objective functions.
- FCM algorithm has much stronger dependence on the initial cluster centers and membership function.

Disadvantage

- It has high dimension of the text vector.
- The clustering result is not stable.

An Improved Fuzzy Clustering Method for Text Mining

Jiabin Deng, JuanLi Hu et al., had proposed an improved fuzzy clustering-text clustering method based on the fuzzy C-means clustering algorithm and the edit distance algorithm[8]. Text clustering is a fully automatic process to divide the similar text into a group. We use the feature evaluation to reduce the dimensionality of high-dimensional text vector. Because the clustering results of the traditional fuzzy C-means clustering algorithm lack the stability, we introduce the high-power sample point set, the field radius and weight. Due to the boundary value attribution of the traditional fuzzy C-means clustering algorithm, we recommend the edit distance algorithm.

Advantage

- It produces the clustering with more stable and accurate than the traditional FCM clustering algorithm.
- The algorithm is feasible, and can effectively improve the precision and stability of the text clustering.

Disadvantage

- It will cause the slow speed to deal directly with these high dimensional vectors.
- The price is too high to find the optimal solution.

Dimension reduction for classification with gene expression microarray data Jian J. Dai, Linh Lieu et al., had explained an important application of gene expression microarray data is classification of biological samples or prediction of clinical and other outcomes[9]. One necessary part of multivariate statistical analysis in such applications is dimension reduction. It provides a comparison study of three dimension reduction techniques, namely partial least squares (PLS), sliced inverse regression (SIR) and principal component analysis (PCA), and evaluates the relative performance of classifi- cation procedures incorporating those methods. A five-step assessment procedure is designed for the purpose. Predictive accuracy and computational efficiency of the methods are examined. Two gene expression data sets for tumor classification are used in the study.

Advantage

- The PLS and SIR based classification procedures performed consistently better than the PCA based procedure in prediction accuracy.
- The PLS and SIR components are more likely to be good predictors than those from PCA.

Disadvantage

- These would be difficult to compare the performance of dimension reduction methods.
- The processing time is high.

K-Means Clustering Algorithm with Improved Initial Center MadhuYedla, SrinivasaRao et al., hadproposed for finding the better initial centroids and to provide an efficient way of assigning the data points to suitable clusters with reduced time complexity. Cluster analysis is one of the primary data analysis methods and k-means is one of the most well known popular clustering algorithms[10]. The k-means algorithm is one of the frequently used clustering methods in data mining, due to its performance in clustering massive data sets. The final clustering result of the kmeans clustering algorithm greatly depends upon the correctness of the initial centroids, which are selected randomly. The original k-means algorithm converges to local minimum, not the global optimum. Many improvements were already proposed to improve the performance of the k-means, but most of these require additional inputs like threshold values for the number of data points in a set. In experimental results, the proposed algorithm has the more accuracy with less computational time comparatively original k-means clustering algorithm.

Advantage

- It has the more accuracy with less computational time.
- It provides an efficient way of assigning the data points to suitable clusters with reduced time complexity.

Disadvantage

- The cost for repairing from any misplacement is also high.
- The proposed algorithm required to be desired number of clusters given as an input.

A New Method for Dimensionality Reduction using K-Means Clustering Algorithm for High Dimensional Data Set Napoleon .D, Pavalakodi .S et al., had explained clustering is the process of finding groups of objects such that the objects in a group will be similar to one another and different from the objects in other groups[11]. Dimensionality reduction is the transformation of high-dimensional data into a meaningful representation of reduced dimensionality that corresponds to the intrinsic dimensionality of the data. K-means clustering algorithm often does not work well for high dimension, hence, to improve theefficiency, apply PCA on original data set and obtain a reduced dataset containing possibly uncorrelated variables. Here the principal component analysisand linear transformation is used for dimensionality reduction and initial centroid is computed, then it is applied to K-Means clustering algorithm.

Advantage

- It obtain better clustering with reduced complexity for high dimensional datasets
- It provides better accuracy and efficiency for high dimensional datasets.

Disadvantage

- This algorithm is generally inefficient in the case of categorical datasets which lack inherent distancemeasure.
- The empty clusters can be obtained if no points are allocated to a cluster during the assignment step.

A Possibilistic Fuzzy c-Means Clustering Algorithm Pal N.R, Pal K, Keller J.M. et al., had proposed the fuzzy-possibilistic c-means(FPCM) model and algorithm that generated both membershipand typicality values when clustering unlabeled data. FPCMconstrains the typicality values so that the sum over all datapoints of typicalities to a cluster is one[12]. The row sum constraintproduces unrealistic typicality values for large data sets. PFCM produces memberships and possibilitiessimultaneously, along with the usual point prototypes or clustercenters for each cluster. PFCM is a hybridization of possibilisticc-means (PCM) and fuzzy c-means (FCM) that often avoidsvarious problems of PCM, FCM and FPCM. PFCM solves thenoise sensitivity defect of FCM, overcomes the coincident clustersproblem of PCM and eliminates the row sum constraints of FPCM.The first-order necessary conditions for extrema of thePFCM objective function is derived, and use them as the basis for a standardalternating optimization approach to finding local minima of thePFCM objective functional. Several numerical examples are giventhat compare FCM and PCM to PFCM. The example show thatPFCM compares favorably to both of the previous models. SincePFCM prototypes are less sensitive to outliers and can avoid coincident clusters, PFCM is a strong candidate for fuzzy rule-basedsystem identification.

Advantage

- It is computationally intensive.
- It is still better than supervising the clusteringalgorithm, especially when labeling knowledge is unavailable.

Disadvantage

- It provides robustness to noise and an intuitive interpretation of the membership values
- ThePCM in its original form is not very suitable for clustering due to the undesirable coincident centroid.

Gene Clustering via Integrated Markov Models Combining Individual and Pairwise Features MatthieuVignes and Florence Forbes et al., had proposed a probabilistic model that has the advantage to account for individual data (eg. expression) and pair wise data (eg. Interaction information coming from biological networks) simultaneously[13]. This model is based on hidden Markov random field models in which parametric probability distributions account for the distribution of individual data. Data on pairs, possibly reflecting distance or similarity measures between genes, are then included through a graph where the nodes represent the genes and the edges are weighted according to the available interaction information. As a probabilistic model, this model has many interesting theoretical features. Also, preliminary experiments on simulated and real data show promising results and points out the gain in using such an approach.

Advantage

- It produces clusters associated to pathways with possible coordinated change in gene expression.
- Consists statistically well founded approach which does not require choosing a distance or a kernel function.

Disadvantage

- It doesn't account for dependencies between observations and is well suited to deal with mixture models.
- To incorporate a priori knowledge regarding class proportions or strength of interactions to put more weight on network data.

Constrained Co-clustering of Gene Expression Data Ruggero G. Pensa et al., had explained the expert interpretation of co- clustering is easier than for mono-dimensional clustering. Co-clustering aims at computing a bi-partition that is a collection of co-clusters: each co-cluster is a group of objects associated to a group of attributes and these associations can support interpretations[14]. Many constrained clustering algorithms have been proposed to exploit the domain knowledge and to improve partition relevancy in the mono-dimensional case (e.g., using the so-called must-link and cannot-link constraints). Here, considered constrained co-clustering not only for extended must-link and cannot-link constraints (i.e., both objects and attributes can be involved), but also for interval constraints that enforce properties of co-clusters when considering ordered domains. Here an iterative co- clustering algorithm is proposed which exploits user-defined constraints while minimizing the sum-squared residues, i.e., an objective function introduced for gene expression data clustering. Then the added value is illustrated of our approach in two applications on gene expression data.

Advantage

- It does not look for overlapping co-cluster.

- It can be easily extended towards other kinds of objective functions.

Disadvantage

- Its accuracy in classification is low.
- Computation complexity is high.

Enhancement of Fuzzy Possibilistic C-Means Algorithm using EM Algorithm (EMFPCM) Shanthi .R and Suganya .R et al., had proposed an effective clustering techniques called Enhancement of Fuzzy Possibilistic C-Means Algorithm using EM Algorithm (EMFPCM). The major difficulties that arise in several fields, comprising pattern recognition, machine learning and statistics, is clustering. The basic data clustering problem might be defined as finding out groups in data or grouping related objects together. A cluster is a group of objects which are similar to each other within a cluster and are dissimilar to the objects of other clusters[15]. The similarity is typically calculated on the basis of distance between two objects or clusters. Two or more objects present inside a cluster and only if those objects are close to each other based on the distance between them. Thus with the help of EMFPCM, noise is reduced, provides more accuracy and thus provides better result in predicting the user behavior. The algorithm was implemented and the experiment result proves that this method is very effective in predicting user behavior. This approach is suitable for applications in business, such as to design personalized web service. The performance of the proposed approaches is evaluated on the UCI machine repository datasets such as Iris, Wine, Lung Cancer and Lymphograma.

Advantage

- The evaluation of clustering accuracy is high and mean square error rate is less.
- The execution time is less.

Disadvantage

- For large data sets the row sum constraint produces unrealistic typicality values.
- Computational cost is high.

Possibilistic Fuzzy c-Means Clustering Model Using Kernel Methods Xiao-Hong et al., presented a novel approach on Possibilistic Fuzzy C-Means Clustering Model Using Kernel Methods. The author insisted that fuzzy clustering method is based on kernel methods. This technique is said to be Kernel Possibilistic Fuzzy C-Means model (KPFCM). KPFCM is an improvement in Possibilistic Fuzzy C-Means model (PFCM) which is superior to FC M model[16]. The KPFCM model is different from PFCM and FCM which are based on Euclidean distance. The KPFCM model is based on non-Euclidean distance by using kernel methods. In addition, with kernel methods the input data can be mapped implicitly into a high-dimensional feature space where the nonlinear pattern now appears linear. KPFCM can deal with noises or outliers better than PF CM. The KPFCM model is interesting and provides good solution.

Advantage

- It can deal with noises or outliers better than PFCM.
- The performance is high.

Disadvantage

The accuracy is less for high dimensional data set.

- Processing time is high.

Similarity Based Fuzzy and Possibilistic c-means Algorithm Chunhui et al., presented a similarity based fuzzy and possibilistic c-means algorithm called SFPCM. It is derived from original fuzzy and possibilistic-means algorithm (FPCM) which was proposed by Bezdek[17]. The difference between the two algorithms is that the proposed SFPCM algorithm processes relational data, and the original FPCM algorithm processes propositional data. Experiments are performed on 22 data sets from the UCI repository to compare SFPCM with FPCM. The results show that these two algorithms can generate similar results on the same data sets. SFPCM performs a little better than FPCM in the sense of classification accuracy, and it also converges more quickly than FPCM on these data sets.

Advantage

- It is especially useful for processing large data sets in practical applications.

- SFPCM converges more quickly than FPCM.

Disadvantage

- It is not suitable for more data sets with different choices of input parameters.
- Which type of data is suited to use SFPCM is not analyzed.

INFERENCE FROM EXISTING SOLUTION

The main drawbacks of existing clustering techniques are given below.

- Divisive Correlation Clustering Algorithm is not suitable for low-dimensional data and it cannot guarantee absence of repulsion inside a cluster or highest average attraction between genes inside clusters
- Phylogenetic reconstruction using an unsupervised neural network of execution time high and efficiency is less.
- Fuzzy C means method for clustering has high dimension of the text vector and clustering result is not stable.
- An Improved Fuzzy Clustering Method for Text Mining will cause the slow speed to deal directly with these high dimensional vectors and price is too high to find the optimal solution.
- Dimension reduction for classification the performance of dimension reduction is difficult to compare and processing time is high.
- K-Means Clustering Algorithm with Improved Initial Center cost for repairing from any misplacement is also high and it required to be desired number of clusters given as an input.
- K-Means Clustering Algorithm for High Dimensional Data Set is inefficient in the case of categorical datasets and empty clusters can be obtained.
- Possibilistic Fuzzy c-Means Clustering provides robustness to noise and an intuitive interpretation of the membership values and is not very suitable for clustering.
- Gene Clustering via Integrated Markov Models Combining Individual and Pairwise Features doesn't account for dependencies between observations and is well suited to deal with mixture models.And to incorporate a priori knowledge regarding class proportions or strength of interactions to put more weight on network data.
- Constrained Co-clustering of Gene Expression Data accuracy in classification is low. And Computation complexity is high.
- Enhancement of Fuzzy Possibilistic C-Means Algorithm using EM Algorithm (EMFPCM) For large data sets the row sum constraint produces unrealistic typicality values. And computational cost is high.
- Possibilistic Fuzzy c-Means Clustering Model Using Kernel Methods the accuracy is less for high dimensional data set. And Processing time is high.
- Similarity Based Fuzzy and Possibilistic c-means Algorithm it is not suitable for more data sets with different choices of input parameters. And which type of data is suited to use SFPCM is not analyzed.

A Solution To Overcome These Issues

The clustering analysis of DNA micro arrays based gene expression databased above methods had some issues. To overcome these issues the hybridization method is used. Hybridization generally refers to a molecular biology technique that measures the degree of genetic similarity between pools of DNA sequences. It is usually used to determine the genetic distance between two organisms.

CONCLUSION

This paper describes different clustering techniques in DNA micro arrays based gene expression data. A clustering algorithm's suitability to cluster biological data depends upon certain desirable features such as speed, minimum number of input parameters, robustness to noise and outliers, redundancy handling and independence of object order input. Though the features of many clusteringalgorithms match these requirements, they have not yet been applied to clustering biological data. Moreover, not all validity measures are suitable for all gene datasets; hence a judicious choice of the applicability of the validity measurehas to be made. It is well known that most clustering methods are highly sensitive to input data and a slight variation or change in the data may result in very different gene clusters. In this survey, an attempt has been made to provide a comprehensiveand precisesurvey of various clustering approaches in thecontext of pattern identificationand recognitionin the gene expression data. Effectiveness of aclustering technique is highly influenced by the selection of algorithmand criterion used by the technique.A short list of clustering approaches availablefor gene data clusteringis provided. Each algorithm is analyzed

effectively and their shortcomings are also enumerated. Finally, discussion about the challenges faced by the clustering schemes available for the effective clustering of gene expression is also provided. In future research is based on Hybridization to measures the degree of genetic similarity between pools of DNA sequences.

REFERENCES

[1] D. M. Dziuda, Data Mining for Genomics and Proteomics: Analysis of Gene and Protein Expression Data, Wiley, 2010.
[2] M.B. Eisen and P.O. Brown, "DNA arrays for analysis of gene expression", Methods Enzymol, vol. 303, pp. 170-205, P.O. 1999.
[3] R. Xu and D. II Wunsch, "Survey of clustering algorithms,"IEEE Trans. Neural Networks, vol. 16, No. 3, pp. 645– 678, 2005.
[4] D. X. Jiang, C. Tang, and A. D. Zhang, "Cluster analysis for gene expression data: A survey,"IEEE Trans. Know. and Data Eng., Vol.16, No. 11, pp. 1370–1386, 2004.
[5] De K.R and Bhattacharya .A, "Divisive Correlation Clustering Algorithm (DCCA) for grouping of genes: detecting varying Patterns in expression profiles," bioinformatics, Vol. 24, pp.1359- 1366, 2008.
[6] Dopazo.J andCarazo.J, "Phylogenetic reconstruction using an unsupervised neural network that adopts the topology of a phylogenetic tree," J MolEval, Vol. 44, pp. 226–233, 1997.
[7] DoulayeDembele and Philippe Kastner, "Fuzzy C means method for clustering microarray data", Bioinformatics, Vol.19, No.8, pp.973- 980, 2003.
[8] Jiabin Deng, JuanLi Hu, Hehua Chi and Juebo Wu, "An Improved Fuzzy Clustering Method for Text Mining", Second International Conference on Networks Security Wireless Communications and Trusted Computing (NSWCTC), Vol. 1, pp. 65–69, 2010.
[9] Jian J. Dai, Linh Lieu, and David Rocke, "Dimension reduction for classification with gene expression microarray data," Statistical Applications in Genetics and Molecular Biology, Vol. 5, No. 1, pp. 1–21, 2006.
[10] MadhuYedla, Srinivasa Rao Pathakota, Srinivasa T M, "Enhancing K-Means Clustering Algorithm with Improved Initial Center", MadhuYedla et al. / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 1 (2), pp121-125, 2010.
[11] Napoleon .D, Pavalakodi .S, "A New Method for Dimensionality Reduction using K-Means Clustering Algorithm for High Dimensional Data Set", International Journal of Computer Applications Vol. 13, No.7,pp. 41-46, January 2011.
[12] Pal N.R, Pal K, Keller J.M. and Bezdek J.C, "A Possibilistic Fuzzy c-Means Clustering Algorithm", IEEE Transactions on Fuzzy Systems, Vol. 13, No. 4, pp. 517–530, 2005.
[13] MatthieuVignes and Florence Forbes," Gene Clustering via Integrated Markov Models Combining Individual and Pairwise Features:, m transactions on computational biology and bioinformatics, Vol. 6, No. 2, april-june 2009.
[14] Ruggero G. Pensa, and Jean-François Boulicaut," Constrained Co-clustering of Gene Expression Data", SIAM International Conference on Data Mining - SDM, pp. 25-36, 2008.
[15] Shanthi .R and Suganya .R, "Enhancement of Fuzzy Possibilistic C-Means Algorithm using EM Algorithm (EMFPCM)", International Journal of Computer Applications (0975 – 8887) Volume 61– No.12, January 2013
[16] Xiao-Hong Wu and Jian-Jiang Zhou, "Possibilistic Fuzzy c-Means Clustering Model Using Kernel Methods", International Conference on Intelligent Agents, Web technologies and Internet Commerce, Vol. 2, Publication Year: 2005 , pp. 465 – 470
[17] Chunhui Zhang, Yiming Zhou and Trevor Martin, "Similarity Based Fuzzy and Possibilistic c-means Algorithm",Proceedings of the 11th Joint Conference on Information Sciences (2008), pp. 1-6.

Table 1 Performance Comparison of Existing Clustering Techniques

S.No	Authors	Techniques	Advantages	Disadvantages
1	De K.R and Bhattacharya.A et al., [5]	Divisive Correlation Clustering	It has higher degree of accuracy than other conventional clustering algorithms The computational cost of the DCCA is less.	The method is not suitable for low-dimensional data It cannot guarantee absence of repulsion inside a cluster or highest average attraction between genes inside clusters.
2	Dopazo .J and Carazo .J et al., [6]	Phylogenetic reconstruction using an unsupervised neural network	The algorithm is highly accurate. The neural network presented here an excellent tool for phylogenetic analysis of a large number of sequences.	The execution time high. The efficiency of algorithm is less.
3	DoulayeDembele and Philippe Kastner et al [7]	Fuzzy C means method for clustering	The C items of the fuzzy categories and minimize the objective functions. FCM algorithm has much stronger dependence on the initial cluster centers and membership function.	It has high dimension of the text vector. The clustering result is not stable.
4	Jiabin Deng, JuanLi Hu et al [8]	An Improved Fuzzy Clustering Method for Text Mining	It produces the clustering with more stable and accurate than the traditional FCM clustering algorithm. The algorithm is feasible, and can effectively improve the precision and stability of the text clustering.	It will cause the slow speed to deal directly with these high dimensional vectors. The price is too high to find the optimal solution.
5	Jian J. Dai, Linh Lieu et al [9]	Dimension reduction for classification	The PLS and SIR based classification procedures performed consistently better than the PCA based procedure in prediction accuracy. The PLS and SIR components are more likely to be good predictors than those from PCA.	These would be difficult to compare the performance of dimension reduction methods. The processing time is high.

6	MadhuYedla, Srinivasa Rao et al [10]	K-Means Clustering Algorithm	It has the more accuracy with less computational time. It provides an efficient way of assigning the data points to suitable clusters with reduced time complexity.	The cost for repairing from any misplacement is also high. The proposed algorithm required to be desired number of clusters given as an input.
7	Napoleon .D, Pavalakodi .S et al [11]	K-Means Clustering Algorithm for High Dimensional Data Set	It obtain better clustering with reduced complexity for high dimensional datasets It provides better accuracy and efficiency for high dimensional datasets.	This algorithm is generally inefficient in the case of categorical datasets which lack inherent distance measure. The empty clusters can be obtained if no points are allocated to a cluster during the assignment step.
8	Pal N.R, Pal K, Keller J.M. et al [12]	Possibilistic Fuzzy c-Means Clustering	It is computationally intensive. It is still better than supervising the clustering algorithm, especially when labeling knowledge is unavailable.	It provides robustness to noise and an intuitive interpretation of the membership values The PCM in its original form is not very suitable for clustering due to the undesirable coincident centroid.
9	MatthieuVignes and Florence Forbes [13]	Gene Clustering via Integrated Markov Models Combining Individual and Pairwise Features	It produces clusters associated to pathways with possible coordinated change in gene expression. Consists statistically well founded approach which does not require choosing a distance or a kernel function.	It doesn't account for dependencies between observations and is well suited to deal with mixture models. To incorporate a priori knowledge regarding class proportions or strength of interactions to put more weight on network data.
10	Ruggero G. Pensa, and Jean-François Boulicaut[14]	Constrained Co-clustering of Gene Expression Data	It does not look for overlapping co-cluster. It can be easily extended towards other kinds of objective functions.	Its accuracy in classification is low. Computation complexity is high.
11	Shanthi .R and Suganya .R [15]	Enhancement of Fuzzy Possibilistic C-Means Algorithm using EM Algorithm (EMFPCM)	The evaluation of clustering accuracy is high and mean square error rate is less. The execution time is less.	For large data sets the row sum constraint produces unrealistic typicality values. Computational cost is high.
12	Xiao-Hong Wu and Jian-Jiang Zhou [16]	Possibilistic Fuzzy c-Means Clustering Model Using Kernel Methods	It can deal with noises or outliers better than PFCM. The performance is high.	The accuracy is less for high dimensional data set. Processing time is high.

ISSN: 2277-9655
(I2OR), Publication Impact Factor: 3.785

13	hunhui Zhang, Yiming Zhou and Trevor Martin [17]	Similarity Based Fuzzy and Possibilistic c-means Algorithm	It is especially useful for processing large data sets in practical applications. SFPCM converges more quickly than FPCM	It is not suitable for more data sets with different choices of input parameters. Which type of data is suited to use SFPCM is not analyzed.