Performance Comparison of Hard and Fuzzy Clustering Algorithms on ESTs of Human Genes

Abhilasha Chaudhuri[1], Asha Ambhaikar[2]

[1]Assistant Professor, Department of Computer Science and Engineering, Rungta Engineering College, Raipur, Chhattisgarh, India

[2]Professor & Dean (R&D), Department of Computer Science and Engineering
Rungta College of Engineering & Technology, Bhilai, Chhattisgarh, India

Abstract: ***In biological data analysis sequences discovered in laboratory experiments are not properly identified. Biologists attempt to group genes based on the temporal pattern of their expression levels. Clustering algorithms could provide a structure to the data. Hard clustering methods such as K-means or Hierarchical clustering assign each gene to a single cluster, whereas in fuzzy clustering methods a gene possesses varying degrees of membership with more than one cluster. Performances of both type of clustering algorithms are analyzed in this paper.***

Keywords: Clustering, Hard clustering, K-means clustering, Hierarchical clustering, Fuzzy clustering, EST, Fuzzy C-means Clustering.

1. Introduction

Clustering, is one of key analysis tools for gene expression data sets, attempts to discover groups of genes having similar expression patterns [1]. Elucidating the patterns hidden in gene expression data offers a tremendous opportunity for an enhanced understanding of functional genomics. However, the large number of genes and the complexity of biological networks greatly increase the challenges of comprehending and interpreting the resulting mass of data, which often consists of millions of measurements. A step toward addressing this challenge is the use of clustering techniques, which is an essential process to reveal natural structures and identify interesting patterns in the underlying data. Cluster analysis seeks to partition a given data set into groups based on specified features so that the data points within a group are more similar to each other than the points in different groups. Such information provides helps with EST identification; it is also useful, for example, when developing a new drug aimed at one particular protein as it is important to be aware of its effect on related proteins to prevent cross reactivity [4].

In this paper we consider the application of clustering techniques in biological sequence analysis. At present a significant amount of unrecognized DNA sequences exist in human genome. It is probable that important information about the human genome is hidden in these unrecognized sequences. Any method that can aid their identification is thus extremely valuable. The investigation is based on an inter-disciplinary approach using domain expertise. Such algorithms require data reduction and sampling processes to be performed before they can be applied. “Biologically interesting” clusters have been derived.

A DNA molecule is made up of two strands, these strands are held together with weak bonds consisting of pairs of bases referred to as base pairs, *bp*. The order of the bases along each strand in any particular case is called the DNA sequence.

A DNA molecule contains many genes. The human genome is estimated to comprise at least 100,000 genes which vary considerably in length. The partial sequences (termed *Expressed Sequence Tags* or ESTs) serve as markers but can also identify expressed genes. Such a system therefore gives an efficient method of identifying most human genes [1]. Pharmaceutical companies have enormous databases of ESTs of which about 30% have been identified. Computational tools exist that match unidentified ESTs against known sequences with certain similar characteristics. Nevertheless, these techniques do not give a clear picture of where the EST in question fits into the database as a whole - i.e. groups of sequences that are related in varying degrees. To provide information such as which chemical will react with what kind of protein, the EST data needs to be given some structure by clustering.

2. Methodology and Technique

All the clustering algorithms used here take the input data in matrix form, the matrix containing some numerical value. Numerical value is necessary because the distances between data objects need to be calculated. Now the problem arises that the DNA sequence are coded in terms of four bases a, t, c, and g. the data we have is in text format therefore we need to encode the data.

2.1 Preprocessing EST sequence

EST sequence data needs some preprocessing i.e. encoding before applying it to the clustering algorithm. There are two ways to encode the data binary encoding and decimal encoding. In binary encoding the four bases of the DNA sequence are represented by a two digit code, while in case of decimal encoding bases are represented by only one digit code. Binary encoding needs double space compared to decimal encoding system in which each of the four bases (a,t,c,g) are represented by only one digit thus it allows just double the number of attributes as in case of binary encoding.

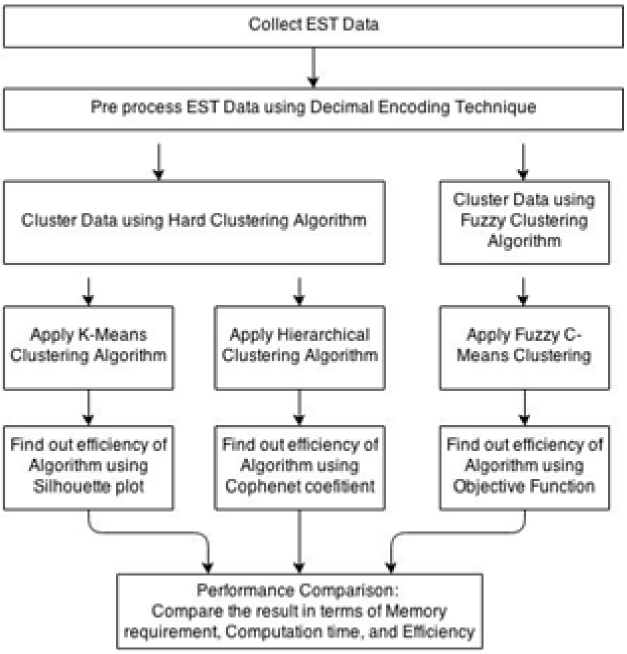

Collect EST Data
Pre process EST Data using Decimal Encoding Technique
Cluster Data using Hard Clustering Algorithm
Cluster Data using Fuzzy Clustering Algorithm
Apply K-Means Clustering Algorithm
Apply Hierarchical Clustering Algorithm
Apply Fuzzy C-Means Clustering
Find out efficiency of Algorithm using Silhouette plot
Find out efficiency of Algorithm using Cophenet coefitient
Find out efficiency of Algorithm using Objective Function
Performance Comparison: Compare the result in terms of Memory requirement, Computation time, and Efficiency

Figure 1: Shows the overall block diagram of methodology used.

After preprocessing basically two types of clustering technique is used:

1. Hard clustering
   - K-means clustering
   - Hierarchical clustering
2. Fuzzy clustering
   - C-means clustering

These clustering algorithms have been proven useful for identifying biologically relevant groups of genes and samples.

A hard clustering algorithm assigns each sample in a data set to a single cluster, whereas a fuzzy clustering method assigns various degrees of membership in several clusters. A fuzzy clustering can be converted to a hard counterpart by assigning all samples to the clusters in which they have the strongest membership. Hard clustering algorithms can be further divided into hierarchical and partitional clustering algorithms. In this paper we have used hierarchical clustering algorithm, K-means clustering algorithm from partitional family and Fuzzy C-means algorithm from the Fuzzy algorithms category.

3. Simulation Models

All three clustering algorithms have been evaluated on same data set. The source of data is “National Center for Biotechnology Information” [8]

3.1 Algorithms used

3.1.1 K-means clustering algorithm

Given the dataset to be clustered, first select the initial number of clusters k to proceed with the algorithm described below [5]:

1. Select an initial partition of k clusters.
2. Assign each object to the cluster with the closest centroid.
3. Compute the new centroid of the clusters.
4. Repeat step 2 and 3 until no object changes cluster.

3.1.2Hierarchical clustering algorithm

Given a set of N items to be clustered, and an NxN distance (or similarity) matrix, the basic process hierarchical clustering is this [6,7]:

1. Start by assigning each item to its own cluster, so that if you have N items, you now have N clusters, each containing just one item.
2. Find the closest (most similar) pair of clusters and merge them into a single cluster, so that now you have one less cluster.

3.1.3Fuzzy C-means clustering algorithm

Fuzzy c-means (FCM) is a method of clustering which allows one piece of data to belong to two or more clusters [9]. This method is frequently used in pattern recognition. It is based on minimization of the following objective function:

$$J_m = \sum_{i=1}^{N}\sum_{j=1}^{C} u_{ij}^{m}\left\|x_i - c_j\right\|^2 \quad , \quad 1 \le m < \infty$$

where m is any real number greater than 1, u_{ij} is the degree of membership of x_i in the cluster j, x_i is the ith of d-dimensional measured data, c_j is the d-dimension center of the cluster, and ||*|| is any norm expressing the similarity between any measured data and the center. Fuzzy partitioning is carried out through an iterative optimization of the objective function shown above, with the update of membership u_{ij} and the cluster centers c_j by:

$$u_{ij} = \frac{1}{\sum_{k=1}^{C}\left(\frac{\left\|x_i - c_j\right\|}{\left\|x_i - c_k\right\|}\right)^{\frac{2}{m-1}}} \quad , \quad c_j = \frac{\sum_{i=1}^{N} u_{ij}^{m} \cdot x_i}{\sum_{i=1}^{N} u_{ij}^{m}}$$

This iteration will stop when $\max_{ij}\left\{\left|u_{ij}^{(k+1)} - u_{ij}^{(k)}\right|\right\} < \varepsilon$, where ε is a termination criterion between 0 and 1, whereas k are the iteration steps. This procedure converges to a local minimum or a saddle point of J_m.

3.2 Experiment Results

3.2.1K-means clustering algorithm

K-means clustering can best be described as a partitioning method. That is, k-means partitions the observations in your data into K mutually exclusive clusters. k-means treats each observation in your data as an object having a location in space. It finds a partition in which objects within each cluster are as close to each other as possible, and as far from objects in other clusters as possible.

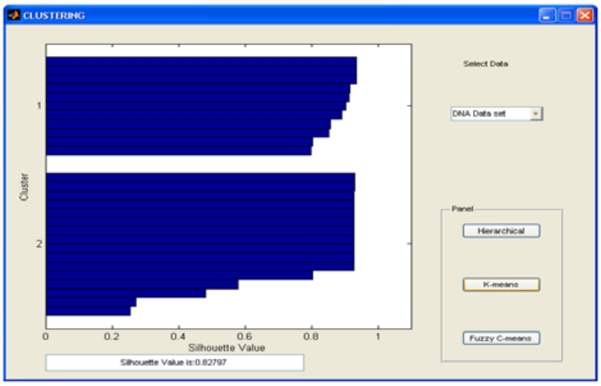
Each cluster in the partition is defined by its member objects and by its centroid, or center. The centroid for each cluster is the point to which the sum of distances from all objects in that cluster is minimized.

Measurement of efficiency of K-means clustering algorithm:

To get an idea of how efficient the algorithm is or in other words how well-separated the resulting clusters are. You can make a silhouette plot. The silhouette plot displays a measure of how close each point in one cluster is to points in the neighboring clusters [2]. This measure ranges from +1, indicating points that are very distant from neighboring clusters, through 0, indicating points that are not distinctly in one cluster or another, to -1, indicating points that are probably assigned to the wrong cluster.

From the silhouette plot, you can see that if most points in any particular cluster have a large silhouette value, greater than 0.6, then it indicates that the cluster is somewhat separated from neighboring clusters. A more quantitative way to compare the two solutions is to look at the average or mean silhouette values.

(A) Silhouette plot obtained after applying k-means algorithm to the EST sequence dataset is shown below.



Silhouette Value is:0.82797

Figure 2: silhoutte plot for DNA sequence dataset

(B) Number of clusters present in this dataset is two.
(C) Silhouette value is 0.82797.
(D) Efficiency of the algorithm is: 82.797%

3.2.2Hierarchical clustering algorithm

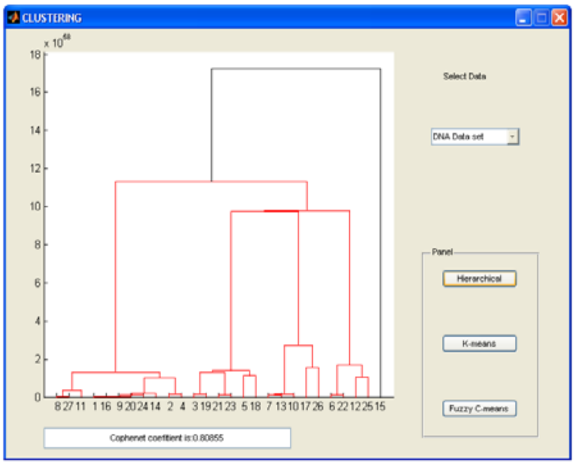
The tree is not a single set of clusters, but rather a multilevel hierarchy, where clusters at one level are joined as clusters at the next higher level. This allows you to decide what level or scale of clustering is most appropriate in your application.

Results of hierarchical clustering are produced as a dendrogram.

Measurement of efficiency of hierarchical clustering algorithm:

Cophenetic correlation coefficient is the measure of efficiency of hierarchical clustering algorithm. The closer the value of cophenetic coefitient to one the more the efficiency of algorithm. In a hierarchical cluster tree, any two objects in the original data set are eventually linked together at some level. The height of the link represents the distance between the two clusters that contain those two objects. This height is known as the "*cophenetic distance"* between the two objects. One way to measure how well the cluster tree generated by algorithm reflects your data is to compare the cophenetic distances with the original distance data. If the clustering is valid, the linking of objects in the cluster tree should have a strong correlation with the distances between objects in the distance vector. The cophenet function compares these two sets of values and computes their correlation, returning a value called the cophenetic correlation coefficient. The closer the value of the "*cophenetic correlation coefficient"* is to 1, the more accurately the clustering solution reflects your data [3] .

a) The dendrogram obtained after applying the hierarchical clustering algorithm to the EST sequence data is shown below.



Cophenet coeftient is:0.80855

Figure 3: Dendrogram of EST sequence dataset

b) The cophenet correlation coefitient is 0.80855 which close to 1 that means DNA sequence data is well reflected by hierarchical clustering algorithm.
c) Efficiency of the algorithm is: 80.855%
d) Maximum distance between any two objects is 18 E+58.

3.2.3Fuzzy C-means clustering algorithm

Fuzzy c-means (FCM) is a data clustering technique wherein each data point belongs to a cluster to some degree that is specified by a membership grade. It provides a method that shows how to group data points that populate some multidimensional space into a specific number of different clusters[10].

Fuzzy c-means assigns every data point a membership grade for each cluster. By iteratively updating the cluster centers and the membership grades for each data point, fcm iteratively moves the cluster centers to the right location within a data set. This iteration is based on minimizing an objective function that represents the distance from any given data point to a cluster center weighted by that data point's membership grade.

Measurement of efficiency of fuzzy c-means clustering algorithm:

The objective function should no longer be decreasing much at all. Graph of the objective function must not show a plot of decreasing value, it must go parallel to the x axis.

Capital O and the sign of cross shows the two cluster centroids. Rest of the points show the data samples. DNA sequence data samples are of very big size as we can see from the y axis value.

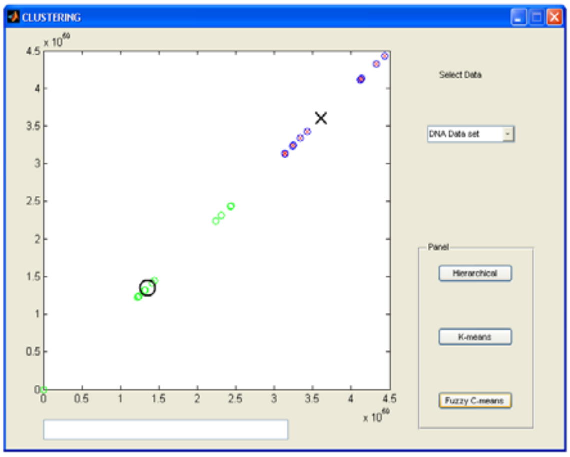


Figure 4: Cluster centroids of EST Sequence dataset

Objective function plot is shown below:

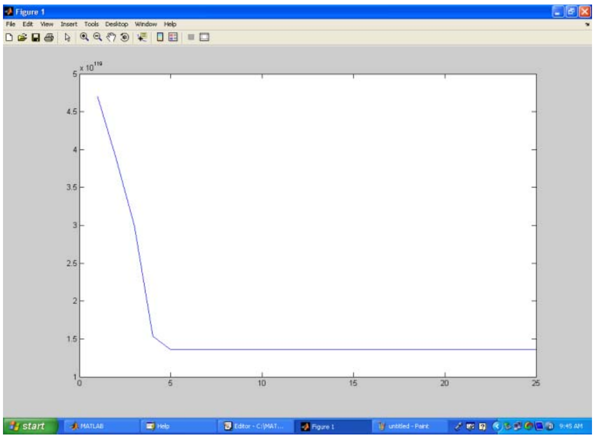


Figure 5: Objective function plot for EST sequence dataset

Objective function's value is: 1.35719808 E+119. As we can see from the graph that graph has became parallel to the x axis i.e. objective function value is not decreasing much at all. It shows that fuzzy C-means algorithm works well for DNA sequence dataset.

3.3 Performance comparison of three algorithms

Table 1: Performance comparison of clustering algorithm for EST sequence data

	Hierarchical clustering	K-means clustering	Fuzzy C-means clustering
Computation time	O(mn2log(n))	O(ktm n)	Near O(n)
Memory requirements	O(mn + n2)	O(mn+kn)	Near O(n)
Number of clusters	27	10	2
Efficiency	80.85%	**82.79%**	81%

4. Conclusion

We summarize and conclude the paper with the mention of the important issue and research trends for cluster algorithms. There is no clustering algorithm that can be universally used to solve all problems. Usually, algorithms are designed with certain assumptions and favor some type of biases. In this sense, it is not accurate to say "best" in the context of clustering algorithms, although some comparisons are possible.

There is a broad variation in number of clusters formed by different algorithms. K-Means algorithm has the highest efficiency 82.79% in clustering human genes. Fuzzy C-Means clustering defines only two clusters and assigns varying degree of membership to EST sequences to each member.

For future work it is suggested that an Adaptive Neuro Fuzzy Inference System can be trained to select the best algorithm according to the application area of the data. Some variants of the classical algorithms can also be used/ developed in order to improve the performance further.

References

[1] M.B. Eisen, P.T. Spellman, P.O. Browndagger, and D. Botstein, "Cluster Analysis and Display of Genome-wide Expression Patterns," *Proc. the National Academy of Sciences of the United States of America* (*PNAS*), vol. 95, no. 25, pp. 14863-14868, 1998.

[2] Rousseeuw, P.J., *Silhouettes: a graphical aid to the interpretation and validation of cluster analysis.* J Comp App Math, 1987. **20**: p. 53-65.

[3] K. Yeung, D. Haynor, and W. Ruzzo, "Validating clustering for gene expression data," *Bioinformatics*, vol. 17, no. 4, pp. 309–318, 2001.

[4] F. Abascal and A. Valencia, "Clustering of proximal sequence space for the identification of protein families," *Bioinformatics*, vol. 18, pp. 908–921, 2002.

[5] S. Gupata, K. Rao, and V. Bhatnagar, "K-means clustering algorithm for categorical attributes," in *Proc. 1st Int. Conf. Data Warehousing and Knowledge Discovery (DaWaK'99)*, Florence, Italy, 1999, pp. 203–208.

[6] Hastie, T., R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. Springer series in statistics. 2001, New York: Springer-Verlag.

[7] Salvador, S. and P. Chan, *Determining the Number of Clusters/Segments in Hierarchical Clustering / Segmentation Algorithms.* ICTAI'04,2004.**00**: p. 576 – 584

[8] Data source "National Center for Biotechnology Information" http://www.ncbi.nlm.nih.gov

[9] A Possibilistic Fuzzy c-Means Clustering Algorithm Pal, N.R. ; Electron. & Commun. Sci. Unit, Indian Stat. Inst., Calcutta, India ; Pal, K. ; Keller, J.M. ; Bezdek, J.C.

[10] M. Roubens, "Pattern classification problems and fuzzy **sets"**, *Fuzzy Sets and System.* vol.1, 1978, pp.239-253.

Author Profile

Abhilasha Chaudhuri received the B.E. degree in Computer Science & Engineering from Pt. Ravishankar Shukla University Raipur and M.Tech degree in Software Engineering from ABV-IIITM Gwalior. During 2008-2009 she worked as SME for Amdocs DVCI, Pune, then from 2009 - 2012 she served as Assistant Professor at KIT Raigarh. Now she is working as Assistant professor at Rungta Engineering College, Raipur. Her areas of interest include Soft computing, Neural Networks and Data Mining.

Dr. Asha Shripad Ambhaikar is working as a professor and Dean (R&D) in the Department of Computer Science and Engineering in Rungta College of Engineering and Technology, Bhilai. She has more than 19 year's academic experience. She has published more than 40 research papers in reputed International and national Journals. She is also a Chairman Board of Studies of IT, Member of Academic Council, Member of Examination Committee & Syllabus revision in CSVTU Bhilai. She has published two books on Adhoc Networking by Germany Publications. She is also the Editorial Board member and reviewer of International Journals' like IJCSI, IJSER and IJSR. She is also the member of various societies like IEEE, CSI, ISTE, IAENG, CSTA, and IACSIT. She has guided more than 25 M.Tech Scholars and guiding 6 Ph.D. Scholars of various universities. Her area of specialization is Computer Networking, Data warehouse and Data Mining, Cloud Computing.