



Comparison of different Clustering Algorithms for Gene Prediction

Sandeep Kaur¹, Sandeep Kaur²

¹Student M.Tech. (CSE), ²Assistant Professor

^{1,2}Department of Computer Science and Engineering,
GNDU Regional Campus,
Gurdaspur, Punjab, INDIA.

Abstract: Clustering algorithms are used to classify different objects and their behavior and properties with other objects. These algorithms are mainly used to analyze microarray data and gene comparisons. In this paper we compared some algorithms i.e. hierarchical algorithms, K-means and Fuzzy C-means algorithms according to their performance, size, software and dataset used and their applications. These clustering algorithms are very valuable in recognizing the genes and its sample data. Thus we can easily analyze gene expression data of different genes using these clustering algorithms. Through these comparisons between different algorithms, we compare which algorithm works more efficiently considering different metrics without degrading its performance.

Keywords: Clustering, Centroid, K-Means Algorithm, Fuzzy C-means algorithm.

I. Introduction

Clustering is a technique to group together similar entities or objects such that objects in the same group have similar behavior and properties as compare to a group of other objects [4][7][8]. Thus grouping of different objects together that poses similar behavior and properties are called clusters. Clusters are basically used to analyze the microarray data or testify and compare genes of unlabeled data to form a structure [8].

Example:

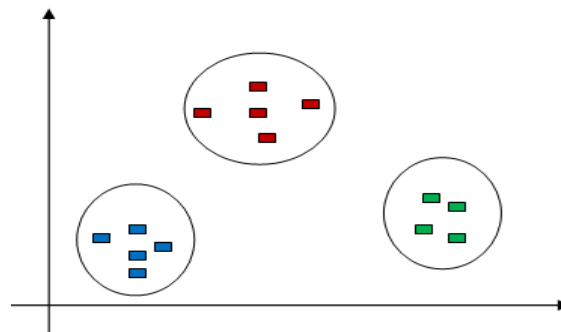


Fig. 1 Group of clusters [8].

In fig. 1 we can see that similar data is closer than the dissimilar data.

Clustering can be defined mathematically by the equation:

$$C = X_1 \cup X_2 \cup X_3 \cup \dots \cup X_n$$

Where,

C is the dataset, $X_1, X_2, X_3, \dots, X_n$ are the clusters in the dataset C and n is total number of clusters in the dataset [9].

A. How to group a cluster [8]:

- Selection: Identify the data or pattern having original features and common behavior.
- Extraction: Alteration of input to introduce a new significant feature or behavior.
- Resemblance: Measured to define similar pattern using distance function.
- Grouping: Group similar pattern to form a cluster.

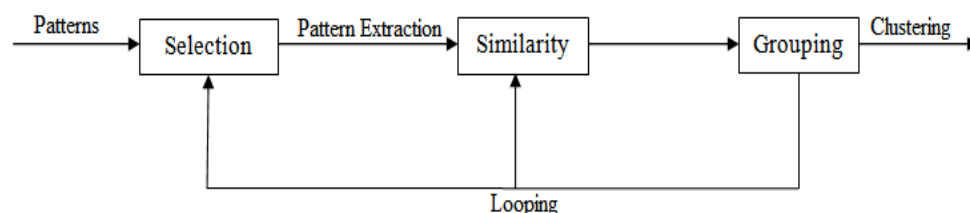


Fig. 2 Procedure to group clusters [8].

II. Reviewed Algorithms and their Comparisons

A. Hierarchical Clustering Algorithm

Hierarchical clustering is a mathematical technique to group together genes into small clusters and then group these cluster into further high level system [8]. Hierarchical clustering as shown in fig.3 organizes elements into a tree, rather than forming elements into clusters. In this case, the genes/data are characterized as the leaves of a tree. The edges of the trees are allocated lengths and the distance between leaves is the path in the tree that joins two leaves [7][12]. Such structure is used to analyze both expression data and molecular evolution studies.

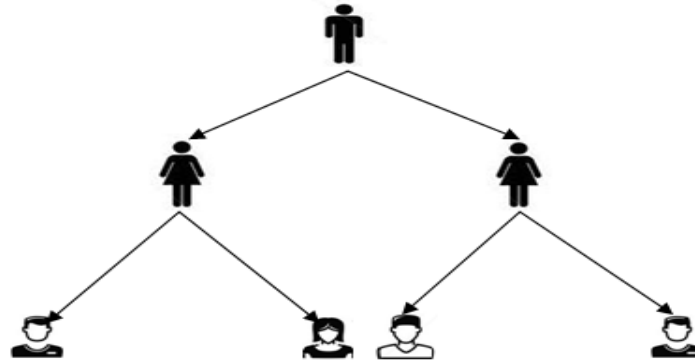


Fig. 3 Hierarchical clustering.

Hierarchical clustering is considered as an efficient tool for microarray datasets by demonstrating the construction in the form of tree structure called dendograms. The hierarchical clustering algorithm contain following attributes [7]:

- an $n \times n$ distance matrix
- Input d of n partitions that is generated from tree's output data.

HierarchicalClustering Algorithm [7][8] (d, n)

1. Create clusters of n number, each having one element
2. Allocate a unique vertex to every cluster by constructing a graph G
3. **While** if number of cluster is more than 1
4. Select 2 nearest clusters C_1 and C_2 and combine C_1 and C_2 in newly formed cluster C
5. Calculate distance from new cluster C to every other clusters
6. Add new vertex in newly formed vertex C to graph G and join with vertices C_1 and C_2
7. Correspond to C_1 and C_2 , eliminate rows and columns of d
8. In newly formed cluster C , insert row and column to d
9. Return G

B. K-Means Clustering

K-means clustering also known as hard C-means clustering is a partitioning method used to analyze and observe data and data objects that lies at different distance and locations between different data points. Partition of objects and data is done in such a way that each and every cluster remains closer to one other as well as far from other cluster objects [8]. It mainly comprises of a central point called centroid. K-means algorithm finds a desired set of these centroids along with their corresponding clusters.

Algorithm [7][8]:

1. Select K as number of partitioned clusters.
2. Choose k as initial starting points which will be used as initial cluster centroids.
3. Examine each point and allocate it to the cluster whose centroid is closer to it.
4. When every point is allocated to the clusters in dataset, further recomputed new centroids k .
5. Repeat steps [3] and [4] till the centroids don't change.

C. Fuzzy C-Means Clustering

FCM is a clustering method which permits an object to lie in two or more clusters. Fuzzy C-means clustering method was introduced in 1981 by Bezdek. This method was extended from K-means clustering. Unlike K-means, FCM is an unsupervised algorithm [10] that is implemented to number of problems related with feature analysis, clustering, pattern recognition etc. FCM is an analysis algorithm based on distance between different data points. It is actually a technique to group data set into cluster of n numbers with data point in related dataset cluster, and thus having a higher connection degree to data point and cluster that lies away from centroid of cluster that have lower connection degree [1][8]. Since FCM is an unsupervised, it is taken as an advantage over K-means algorithm and is not bounded to a certain set of data [10].

Algorithm:

FCM is an iterative algorithm [2][5], which is described in the following.

1. Initialize the matrix

$U = [u_{ij}], U^{(0)}$

2. At k-step: calculate the centers vectors $C^{(k)} = [c_j]$ with $U^{(k)}$

$$c_j = \frac{\sum_{i=1}^N u_{ij}^m \cdot x_i}{\sum_{i=1}^N u_{ij}^m}$$

3. Update $U^{(k)}, U^{(k+1)}$

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

4. If $\|U^{(k+1)} - U^{(k)}\| < \epsilon$ then STOP; otherwise return to step 2.

D. Comparative Analysis

Comparison between various clustering algorithms can be done by following features [12]:

1. Size: Algorithms are compared on the basis of size of the raw datasets as in some algorithms like FCM where clustering is not possible of the datasets that are larger than the main memory.
2. Cluster: Clusters' number is another factor while comparing the algorithms.
3. Software: Category of Software used is one of the features to compare the clustering algorithms.
4. Dataset: Comparison can be done on the basis of dataset type.

➤ **Comparison between Hierarchical and K-means Clustering**

Earlier basically two clustering algorithms were known: Hierarchical and Partitional clustering. Hierarchical Clustering form a layered series of partitions while Partitional Clustering (K-means) produce a single level partitioning [4][11]. Hierarchical Clustering techniques basically emphasizes on the schemes that how to select the clusters in order to merge them to get similarity. Osama Abu Abbas [12] compare various algorithms on the basis of above mentioned 4 factors and gave following results:

As per in table 1, with the increase in the number of cluster k, performance of K-means enhances as compare to Hierarchical Clustering algorithm.

Table 1. Relation between no. of clusters and performance of different algorithms

Number of clusters(K)	Performance	
	Hierarchical Clustering	K-means
8	65	63
16	74	71
32	87	84
64	92	89

According to quality relationship as in table 2, accuracy of hierarchical algorithm becomes improved but has ambiguity to some extents. K-means shows low quality due to some noisy data.

Table 2. Relation between no. of clusters and quality of different algorithms

Number of clusters(K)	Quality	
	Hierarchical Clustering	K-means
8	1090	1112
16	960	1089
32	850	910
64	760	840

Table 3 shows with larger dataset size k-means has better quality but hierarchical algorithm shows better results with small datasets.

Table 3. Influence of size of dataset

Data Size	K=32	
	Hierarchical Clustering	K-means
36000	850	910
4000	91	95

According to dataset type (see table 4), Random and Ideal datasets are used in which ideal dataset is part of software and appropriate for testing. Hierarchical algorithm gives better results as compare to k-means while using random dataset and vice versa.

Table 4. Influence of type of data

K=32		
Data Type	Hierarchical Clustering	K-means
Random	850	910
Ideal	829	810

➤ **Comparative analysis between K-means and Fuzzy C-means clustering**

According to Soumi Ghosh [13] time complexity of K-means is $O(ncdi)$ and time complexity of Fuzzy C-means is $O(ndc^2i)$. In order to guess our results, we keep the data point numbers constant and assume $n=100$, $d=3$ and $i=20$, following results are driven:

Table 5. Time Complexity

Algorithm	Time Complexity	Elapsed Time(seconds)
K-Means	$O(ncdi)$	0.44387
FCM	$O(ndc^2i)$	0.78167

where n = number of data points, c = number of cluster, d = number of dimension and i = number of iterations. Now above table 5 shows the result of comparison when number of clusters are constant. For variable number of clusters [13], following result is found:

Table 6. when number of clusters varies

Number of Clusters	Time Complexity of K-means	Time Complexity of Fuzzy C-means
1	6000	6000
2	12000	24000
3	18000	54000
4	24000	96000

In above table 6, it is clearly determined that with the increase in the number of clusters, time complexity of both algorithms increase but as compare to Fuzzy C-means, time complexity of K-means algorithm is reasonably low [13].

III. Conclusion

With thoroughly reviewing the three clustering algorithms, it is concluded that although performance of K-means algorithm is far better than the Hierarchical algorithm still it has less quality and accuracy along with ambiguity to noise. K-means is sensitive for noise in dataset, thus hierarchical algorithms provide better results in random datasets. As a general scenario, K-means algorithms are used for large datasets while small datasets are used for hierarchical algorithm as K-means is an unsupervised algorithm. As in case between K-means and FCM, reviewed time complexity of K-means $O(ncdi)$ and time complexity of Fuzzy C-means is $O(ndc^2i)$ result s that K-means is better than FCM algorithm as the time complexity of FCM is require more time for computation. This property of K-means algorithm makes it superior over the FCM algorithm.

References

- [1] A.Fahad et al. (2014), "A Survey of Clustering Algorithms for Big Data: Taxonomy & Empirical Analysis", IEEE Transactions On Emerging Topics In Computing, pp. 1-12.
- [2] Chih Cheng Hung et al. (2011), "A New Weighted Fuzzy C-Means Clustering Algorithm for Remotely Sensed Image Classification", IEEE Journal Of Selected Topics In Signal Processing, Vol. 5, No. 3, pp. 543-553.
- [3] Daxin Jiang et al. (2004), "Cluster Analysis for Gene Expression Data: A Survey", IEEE Transactions on Knowledge and Data Engineering, Vol. 16, No. 11, pp. 1370-1384.
- [4] Gen Nowak and Robert Tibshirani (2007), "Complementary hierarchical clustering", Biostatistics, Vol. 9, No. 3, pp. 467-483.
- [5] J. C. Bezdek et al. (1999), "Will the real Iris data please stand up?", IEEE Transaction Fuzzy System, Vol. 7, No. 3, pp. 368-369.
- [6] Kamran Kowsari(2014), "Comparison three methods of clustering: K-means, Spectral Clustering and Hierarchical Clustering", George Washington University, Dept. of Computer Science.
- [7] Lopresti (2007), "Bioinformatics: Issues and algorithms", Lehigh University, pp. 1-53.
- [8] Luis Tari (2004), "Clustering", pp. 1-36.
- [9] Marjan Kuchaki Rafsanjani et al. (2011), "A survey of hierarchical clustering algorithms", The Journal of Mathematics and Computer Science Vol. 5 No. 3, pp. 229-240.
- [10] Maryam Bakhshi et al. (2012), "Review and Comparison between Clustering Algorithms with Duplicate Entities Detection Purpose", International Journal of Computer Science Emerging Technology, Vol-3 No 3, pp. 108-114.
- [11] Michael Steinbach et al. (2001), "A Comparison of Document Clustering Techniques", Department of Computer Science and Engineering, University of Minnesota, pp. 1-20.
- [12] Osama Abu Abbas (2008), "Comparison between Data Clustering Algorithms", The International Arab Journal of Information Technology, Vol. 5, No. 3, pp. 320-325.
- [13] Soumi Ghosh, Sanjay Kumar Dubey (2013), "Comparative Analysis of K-Means and Fuzzy C Means Algorithms", International Journal of Advanced Computer Science and Applications, Vol. 4, No.4, pp. 35-39.