IJETCR Open Access

ISSN: 2348 - 2117

International Journal of Engineering Technology and Computer Research (IJETCR)
Available Online at www.ijetcr.org
Volume 3; Issue 1; Page No. 39-41

Comparison between Cluster Analysis Algorithms

Ved Prakash Jha[1], Soumya Mishra[2], Niharika Singh[3], Swati Vashisht[4]

[1]Scholar, Department of Computer Engineering, AMITY University Gr. Noida Campus, Knowledge Park-3, Uttar Pradesh, India
vedprk8@gmail.com

[2]Scholar, Department of Computer Engineering, AMITY University Gr. Noida Campus, Knowledge Park-3, Uttar Pradesh, India
sm.mania9@gmail.com

[3]Scholar, Department of Computer Engineering, AMITY University Gr. Noida Campus, Knowledge Park-3, Uttar Pradesh, India
niharika927@gmail.com

[4]Department of Computer Science and Engineering, AMITY University Greater Noida, Knowledge Park-3, Uttar Pradesh, India
svashisht@gn.amity.edu

ABSTRACT
Due to explosions in the number of autonomous data sources there is a growing need for effective approaches to distributive data clustering. Clustering is the process of grouping of data by finding similarities between data based on their characteristics. These groups are called Clusters.[4] Each cluster consists of objects that are similar between themselves and dissimilar compared to objects of other groups. In this place we are going to compare and analyse the performance of different clustering algorithm. All the algorithms are compared according to the number of clusters, size of dataset, type of dataset, and type of software used. We extract the conclusion based on performance, quality and accuracy of the clustering algorithm.

Keywords: Clustering, Clustering Algorithms, K-Means Algorithm, Hierarchical clustering algorithm, Self-organizing maps (SOM) Algorithm, Expectation Maximization Clustering Algorithm (EM).

INTRODUCTION:

Clustering is an unsupervised learning method that constitutes a Cornerstone of an intelligent data mining process. It is the process of grouping objects in to clusters such that Objects within a cluster have high similarity in comparison to one another, but differ with objects in other cluster. Clustering is mainly needed to organize the results provided by a search engine. It is also viewed as a special type of Classification. The Cluster forms as a result of clustering can be defined as a set of like elements. But the elements from different Cluster are not alike [1]. Cluster is similar to database Segmentation where like tuples in a database are grouped together.[4] When Clustering is applied to a real world database many problems occur there like interpreting the semantics of each cluster is difficult, handling outlier is difficult, no correct answer for a cluster problem and what data should be used for Clustering[5].

Some researchers worked on some data clustering algorithms, few implemented new ones and some compared different data clustering algorithms. Following are some of the previous studies that considered the effect of different factors on the performance of some data clustering algorithm and compared the result. In this paper we compared the different clustering algorithms on previous studies.

2. ALGORITHM FOR COMPARISION

Three different clustering algorithms are chosen to investigate, study, and compare them. The algorithms that are chosen are: k-means algorithm, hierarchical clustering and Expectation Maximization (EM) clustering algorithm. The general reasons for selecting these three algorithms are:

- Popularity
- Flexibility
- Applicability
- Handling high dimensionality

However, detailed reasons behind selecting every algorithm are listed in the context. [5]

A.K-means Algorithm

K-means is a well-known partitioning method. Cluster membership is determined by calculating the centroid for each group and assigning each object to the group with the closest centroid.[1] This approach minimizes the

overall within-cluster dispersion by iterative reallocation of cluster members.

The algorithm is presented with a set of n sample vectors and a number K for the expected number of clusters, which produces K centroids that attempt to minimize the function, which is the average distance of each sample vector to the centroid. Implementation of the algorithm starts with a random selection for the centroids,iteratively assigning each vector to the nearest centroid, and updating the new centroid positions until convergence is reached.[3]

K-means is one of the simplest algorithms known to perform well with many data sets, but its performance is limited mainly to compact groups. When the points are drawn from the mixture of Gaussian distributions, the K-means is a gradient descent algorithm which minimizes the quantization error. As with many gradient descent algorithms, one of the drawbacks of K-means is that it can reach a local minimum of the objective function instead of the desired global minimum, which means that convergence is reached but the solution is not optimal. One way to overcome this is by running the algorithm multiple times with different random seeds and selecting the partition that appears with the highest frequency [3].

K-means algorithm was chosen to study because of the following reasons:

- Its time complexity is O(nkl) , whereon refers to the number of patterns, k is the number of clusters, and l is the number of iterations taken by the algorithm to converge.
- Its space complexity is O(k+n). It requires additional space to store the data matrix.
- Its order-independent; it generates the same partition of data irrespective of the order in which the pattern are presented to the algorithm for a given initial seed set of cluster centers.

B. Hierarchical Clustering Algorithm

Hierarchical clustering creates a hierarchical tree of similarities between the vectors, called a dendrogram [3]. The usual implementation is based on agglomerative clustering, in which the algorithm is initialized by assigning each vector to its own separate cluster and defining the distance between each cluster based on a distance metric or similarity. Next, the algorithm merges the two nearest clusters and updates all the distances to the newly formed cluster via some linkage method. This is repeated until there is only one cluster left that contains all the vectors [4].

The pseudo code of the hierarchical clustering algorithm to explain how it works is explained in the following steps:

- Compute the proximity matrix containing the distance between each pattern pairs. Treat each pattern as a cluster.
- Find the most similar pair of clusters using the proximity matrix. Merge the two clusters into one cluster. Update the proximity matrix to show this merge operation.
- If all patterns are found to beclustered together in a single cluster, stop. Otherwise, go to step 2.

The advantages of the hierarchical clustering algorithms are the reason this algorithm was chosen for discussion. The advantages include:

- Embedded flexibility with regards to the level of granularity.
- Handling various forms of similarity or distance with ease.
- Consequent applicability to any attribute types.
- Hierarchical clustering algorithms are more flexible in nature.

C. Expectation Maximization

Expectation maximization clustering estimates the probability densities of the classes using the Expectation Maximization (EM) algorithm. The estimated result is a set of K multivariate distributions each defining a cluster, with sample vector assigned to each cluster with maximum conditional probability [3].

EM is selected to cluster data because of the following reasons among others.

- It has a strong statistical basis.
- It is linear in database size.
- It is robust to noisy data.
- It can accept any number of clusters as input.
- It can handle high dimensionality.
- It converges fast on being given a good initialization.

4. PARAMETERS FOR COMPARISION

The comparisons among the four clustering algorithms are performed based on the following factors:

- The size of the dataset.
- Number of the clusters.
- Type of dataset.
- Type of software.

According to the size of data, each of the four algorithms: k-means, Hierarchical Clustering, and EM is executed twice; first by trying a huge dataset and then by trying a small dataset. Table 1 explains how the four algorithms are compared. The total number of times the algorithms have been executed is 24. For each 8-runs group, the results of the executions are studied and compared. The conclusions are written down. This step is repeated for all the factors

Table 1: The factors according to which the algorithms are compared.

	Size of Dataset	Number of Clusters	Type of Dataset	Type of Software
K-means Algo.	Huge Dataset & Small Dataset	Large number of clusters & Small number of Clusters.	Ideal Dataset & Random Dataset.	LNKnet Package & Cluster and TreeView Package
HC Algo.	Huge Dataset & Small Dataset	Large number of clusters & Small number of Clusters.	Ideal Dataset & Random Dataset.	LNKnet Package & Cluster and TreeView Package
EM Algo.	Huge Dataset & Small Dataset	Large number of clusters & Small number of Clusters.	Ideal Dataset & Random Dataset.	LNKnet Package & Cluster and TreeView Package

Here, the performance of different algorithms for different k's is compared in order to test the performances that are related to k. To simplify the situation and to make the comparisons easier, k is chosen equal to 8, 16, 32, and 64.

Table 2: The relationship between number of clusters and the performance of the algorithms.

Number of Clusters (K)	K-means	Expectation Maximization	Hierarchical Clustering
8	63	62	65
16	71	69	74
32	84	84	87
64	89	89	92

5. CONCLUSION

After analyzing the results of testing the clustering algorithms and running them under different factors and situations, the following conclusions are obtained:

- The performance of k-means and EM algorithms is better than hierarchical clustering algorithm.[5]
- With the increasing value of k, the accuracy of hierarchical clustering becomes better.
- K-means and EM have less accuracy than the others.
- All the algorithms have some ambiguity in data when clustered.
- The accuracy of EM and k-means algorithms become very good when using huge dataset.
- Hierarchical clustering show good results when using small dataset.
- As a general conclusion, k-means & EM are recommended for huge dataset while hierarchical clustering algorithms for small dataset.
- K-means and EM algorithms are very sensitive for noise in dataset that makes it difficult to cluster an object into its suitable cluster.
- Running the clustering algorithms using any software gives almost the same results even when changing any of the factors because most software use the same procedures and ideas in any algorithm implemented by them.[5]

6. REFERENCES:

1. Eisen M., *Cluster and Tree View Manual*, Standard University, 1998.
2. Han J. and Kamber M., *Data Mining: Concepts and techniques,* Morgan Kaufnann Publishers, 2001.
3. Survey of Clustering Data Mining Techniques, Pavel Berkhin, Accrue Software Inc.
4. S. Revathi, Dr. T. Nalini, International Journal of Advanced Research in Computer Science & Software Engineering, 2013.
5. Osama Abu Abbas, the International Arab Journal of Information & Technology Vol.5 No. 3, Yarmouk University, Jordon, 2008.