ISSN:0975-9646

Performance Evaluation of K-Means and Heirarichal Clustering in Terms of Accuracy and Running Time

Nidhi Singh,.Divakar Singh
Department of computer Science & Engg.
BUIT,BU,Bhopal.India.(M.P)

***Abstract*: As in today's world large number of modern techniques is evolving for scientific data collection, large number of data is getting accumulated at various databases. Systematic data analysis methods are in need to gain/extract useful information from rapidly growing databanks. Clustering analysis method is one of the main analytical methods in data mining; in which k-means clustering algorithm is most popularly/widely used for many applications. Clustering algorithm is divided into two categories: partition and hierarchical clustering algorithm. This paper discusses one partition clustering algorithm (k-means) and one hierarchical clustering algorithm (agglomerative). K-means algorithm has higher efficiency and scalability and converges fast when dealing with large data sets. Hierarchical clustering constructs a hierarchy of clusters by either repeatedly merging two smaller clusters into a larger one or splitting a larger cluster into smaller ones. Using WEKA data mining tool we have calculated the performance of k-means and hierarchical clustering algorithm on the basis of accuracy and running time.**

Keywords: Data mining, K-means clustering, Hierarchical clustering, Agglomerative clustering, divisive clustering.

I. INTRODUCTION

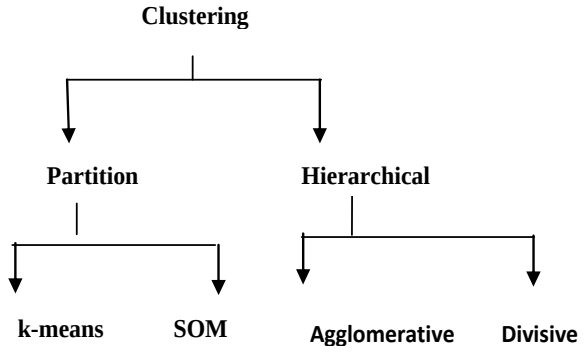
As a result of modern methods for scientific data collection, huge quantities of data are getting accumulated at various databases. Cluster analysis [4] is one of the major data analysis methods which helps to identify the natural grouping in a set of data items. The K-Means clustering algorithm is proposed by Mac Queen in 1967 which is a partition-based cluster analysis method. Clustering is a way that classifies the raw data reasonably and searches the hidden patterns that may exist in datasets [6]. It is a process of grouping data objects into disjointed clusters so that the data's in the same cluster is similar, yet data's belonging to different cluster differ. K-means is a numerical, unsupervised, non-deterministic, iterative method. It is simple and very fast, so in many practical applications, the method is proved to be a very effective way that can produce good clustering results. The demand for organizing the sharp increasing data's and learning valuable information from data, which makes clustering techniques are widely applied in many application areas such as artificial intelligence, biology, customer relationship management, data compression, data mining, information retrieval, image processing, machine learning, marketing, medicine, pattern recognition, psychology, statistics [2]and so on [1].

Clustering methods can be divided into two general classes, designated supervised and unsupervised clustering. In this paper, we focus on unsupervised clustering which may again be separated into two major categories: partition clustering and hierarchical clustering. There are many algorithms for partition clustering category, such as k-means clustering (MacQueen 1967), k-medoid clustering, genetic k-means algorithm (GKA), Self-Organizing Map (SOM) and also graph-theoretical methods (CLICK, CAST). Among those methods, K-means clustering is the most popular one because of simple algorithm and fast execution speed[3].

Hierarchical clustering methods are among the first methods developed and analyzed for clustering problems [9]. There are two main approaches. (i) The agglomerative approach, which builds a larger cluster by merging two smaller clusters in a bottom-up fashion. The clusters so constructed form a binary tree; individual objects are the leaf nodes and the root node is the cluster that has all data objects. (ii) The divisive approach, which splits a cluster into two smaller ones in a top-down fashion. All clusters so constructed also form a binary tree.

II. CATEGORIZATION OF CLUSTERING ALGORITHM

In this paper clustering algorithm we have used for comparison is categorized as follows:



1) THE K-MEANS CLUSTERING ALGORITHM

The process of k-means algorithm:

This part briefly describes the standard k-means algorithm. K-means is a typical clustering algorithm in data mining and which is widely used for clustering large set of data's. It was one of the most simple, non-supervised learning algorithms, which was applied to solve the problem of the well-known cluster[5]. It is a partitioning clustering algorithm, this method is to classify the given date objects into k different clusters through the iterative, converging to a local minimum. So the results of generated clusters are compact and independent. The algorithm consists of two separate phases. The first phase selects k centers randomly,

where the value k is fixed in advance. The next phase is to take each data object to the nearest center[7].Euclidean distance is generally considered to determine the distance between each data object and the cluster centers. When all the data objects are included in some clusters, the first step is completed and an early grouping is done. Recalculating the average of the early formed clusters. This iterative process continues repeatedly until the criterion function becomes the minimum. Supposing that the target object is x, xi indicates the average of cluster Ci, criterion function is defined as follows (eq. 1.):

$$E= \sum_{i=1}^{x} \sum_{x \in ci} (x-xi)^2 \qquad (1)$$

E is the sum of the squared error of all objects in database. The distance of criterion function is Euclidean distance, which is used for determining the nearest distance between each data objects and cluster center. The Euclidean distance between one vector x=(x1 ,x2 ,…xn) and another vector y=(y1 ,y2 ,…yn), The Euclidean distance d(xi, yi) can be obtained as follow:

$$d(xi, yi)= [\sum_{i=1}^{n} (x_i-y_i)^2]^{1/2} \qquad (2)$$

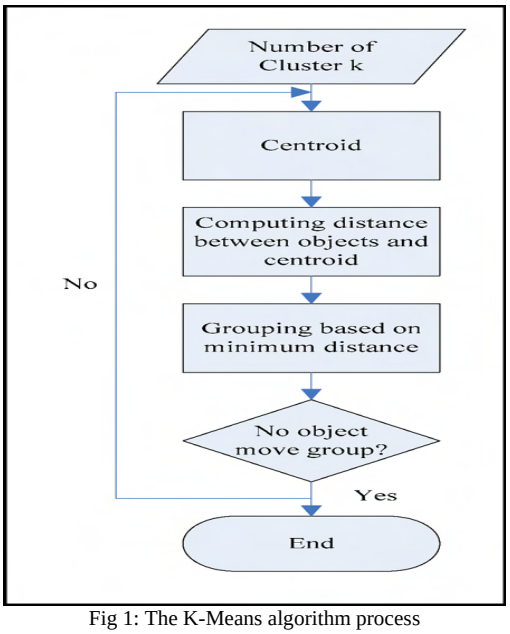


Fig 1: The K-Means algorithm process

k-means algorithm :
Input:
Number of desired clusters, k, and a database D={d1, d2,…dn} containing n data objects.
Output:
A set of k clusters
Steps:
1) Here we have to select randomly k data objects from dataset D as initial cluster centers.
2) Repeat;
3) Then calculate the distance between each data object di (1 <= i<=n) and all k cluster centers cj(1<=j<=k) and assign data object di to the nearest (closest) cluster.
4) For each cluster j , recalculate the cluster center.
5) until no changing in the center of clusters.

The k-means clustering algorithm always converges to local minimum. Before the k-means algorithm converges, calculations of distance and cluster centers are done while loops are executed a number of times, where the positive integer t is known as the number of k-means iterations. The precise value of t varies depending on the initial starting cluster centers[8]. The distribution of data points has a relationship with the new clustering center, so the computational time complexity of the k means algorithm is O(nkt). n is the number of all data objects, k is the number of clusters, t is the iterations of algorithm. Usually requiring k <<n and t <<n[1].
The reason behind choosing K-means algorithm to study is its popularity for the following reasons:

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

2) HIERARCHICAL CLUSTERING

The k-means algorithm gives us what's sometimes called a simple or at partition, because it just gives us a single set of clusters, with no particular organization or structure within them. But it could easily be the case that some clusters could, themselves, be closely related to other clusters, and more distantly related to others. So sometimes we want a hierarchical clustering, which is depicted by a tree or dendrogram .There are two approaches to hierarchical clustering: we can go from the bottom up", grouping small clusters into larger ones, or \from the top down", splitting big clusters into small ones. These are called Agglomerative and Divisive clustering's, respectively. We will return to divisive clustering later, after we have tools to talk about the over-all pattern of connections among data points.

The basic algorithm is very simple:

1. Start with each point in a cluster of its own
2. until there is only one cluster
   (a) Find the closest pair of clusters
   (b) Merge them
3. Return the tree of cluster-mergers

The above algorithm is simple concept of hierarchical algorithm. In this paper we have used agglomerative clustering algorithm, in this algorithm the cluster have sub-clusters, and again sub-clusters have sub-clusters and so on. Agglomerative algorithm starts with every single object in a single cluster. Then in each successive iteration, it agglomerates (merges) the closest pair of cluster by satisfying some similarity measures, until all the data is in one single cluster. Any such procedure is greedy, like our feature-selection algorithm and deterministic (no random initial conditions, unlike k-means). It returns a sequence of nested partitions, where each level up merges two cells of the lower partition.

The advantages of using hierarchical algorithm are as follows:

- Embedded flexibility regarding a level of granularity.
- Ease of handling of any form of similarity or distance.
- Consequently applicability to any attributes types.
- Hierarchical clustering algorithm is more versatile.

.

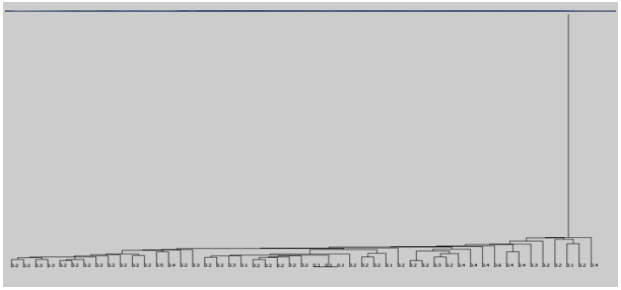


Fig 2: Showing dendrogram formed from the iris data set.

III. EXPERIMENTS

For testing accuracy and efficiency of simple k-means and hierarchical clustering algorithm, various datasets with known clustering available at UCI repository of machine learning databases[10].This paper uses Iris[10] and Diabetes[10] datasets and a brief description of datasets used in experiment evaluation Table 1 shows some characteristics of datasets as the test datasets.

TABLE 1. DESCRIPTION OF DATASETS.

Datasets	Number of Attribute	Number of records/Instances
IRIS	05	150
DIABETES	09	768

In this paper, we have used weka data mining tool version 3.7 for testing accuracy and running time of simple K-means and Hierarchical clustering algorithm on given datasets. The clustering results for cluster k=3 is shown in Table 2.

TABLE 2.CLUSTERING RESULTS FOR DATASETS.

Datasets	k-means running time(sec)	Hierarchical clustering running time	k-means Accuracy %	Hierarchical clustering Accuracy %
IRIS	0.03	0.17	88.667	66
DIABETES	0.06	2.14	51.6927	65.1042

The results for accuracy and running time are shown in Figure 3 and Figure 4.

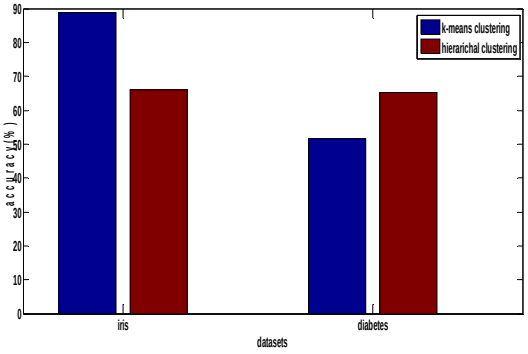


Figure 3. Accuracy v/s datasets

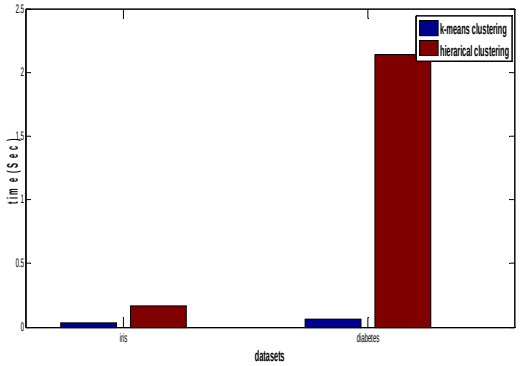


Figure 4. Running time v/s datasets

IV. CONCLUSION

K-means is a typical clustering algorithm and it is widely used for clustering large sets of data. From experiments, we can conclude that the accuracy of k-means for iris dataset having "real" attributes is much than the hierarchical clustering and for diabetes dataset having "integer, real" attributes accuracy of hierarchal clustering is more than the k-means algorithm (fig 3). Though the time taken to cluster the data sets is less in case of k-means (fig 4). Hierarchical clustering results are usually presented in dendrogram, the dendrogram for iris dataset is shown in (fig 2).A good clustering method produces high-quality clusters to ensure that the inter-cluster similarity is low and the intra-cluster similarity is high, in other words, members of a cluster are more like each other than they are like members of a different cluster. As we have discussed in this paper k-means algorithm is good for large datasets.

REFERENCES

[1]. Shi Na , Liu Xumin, Guan yong, "Research on k-means Clustering Algorithm", IEEE 2010,978-0-7695-4020-7/10$26.00.

[2]. Sun Shibao, Qin Keyun, "Research on Modified k-means Data Cluster Algorithm" I. S. Jacobs and C. P. Bean, "Fine particles, thin films and exchange anisotropy," Computer Engineering, vol.33, No.13, pp.200–201,July 2007

[3]. Bernard Chen, Phang C. Tai, R. Harrison and Yi Pan, "Novel Hybrid Hierarchical-K-means Clustering Method (H-K-means) for Microarray Analysis" ,IEEE 2005(CSBW'05)

[4]. Jiawei Han M. K, Data Mining Concepts and Techniques, Morgan Kaufmann Publishers, An Imprint of Elsevier, 2006

[5]. Sun Jigui, Liu Jie, Zhao Lianyu, "Clustering algorithms Research ",Journal of Software ,Vol 19,No 1, pp.48-61,January 2008

[6]. Huang Z, "Extensions to the k-means algorithm for clustering large data sets with categorical values," Data Mining and Knowledge Discovery, Vol.2, pp:283–304, 1998

[7]. Fahim A M,Salem A M,Torkey F A, "An efficient enhanced k-means clustering algorithm" Journal of Zhejiang University Science A, Vol.10,pp:1626-1633,July 2006

[8]. K.A.Abdul Nazeer, M.P.Sebastian, "Improving the Accuracy and Efficiency of the k-means Clustering Algorithm",Proceeding of the World Congress on Engineering, vol 1,london, July 2009

[9]. A.K. Jain and R.C. Dubes. Algorithms for clustering data. Prentice Hall, 1988

[10]. Merz C and Murphy P, UCI Repository of machine learning Databases, Available: ftp://ftp. ics.uci.edu/pub/machine-learningdatabase