Volume 4, Issue 5, May 2014 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Study and Comparison of Partition Based and Hierarchical Clustering

ARCHANA
STUDENT OF MASTER OF TECHNOLOGY
DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
SHOBHIT UNIVERSITY
MEERUT, INDIA

Abstract :Data mining is the mechanism of implementing patterns in large amount of data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems. Clustering is the very big area in which grouping of same type of objects in data mining. Clustering has divided into different categories – partitioned clustering and hierarchical clustering. In this paper we study. Clustering and show comparison between partitioned clustering and hierarchical clustering. which is better for categorical data.

Keywords :Web mining, Data mining, clustering, Partitioned, Hierarchical.

1. INTRODUCTION

Web mining is the use of data mining techniques to automatically discover and extracts information from Web documents and services. In other words, Web Mining is the extraction of interesting and potentially useful patterns and implicit information from artifacts or activity related to the World Wide Web.

Clustering is a unsupervised learing. Division of data into groups of similar objects is called Clustering. Certain fine details are lost by representing the data by fewer clusters but it achieves simplification. It models data by its clusters. Data modeling puts clustering in a historical perspective rooted in mathematics, statistics, and numerical analysis. According to machine learning perspective clusters correspond to hidden patterns, the search for clusters is unsupervised learning, and the resulting system represents a data concept. From a practical perspective clustering plays an important role in data mining applications such as scientific data exploration, information retrieval and text mining, spatial database applications, Web analysis, CRM, marketing, medical diagnostics, computational biology, and many others. Clustering can be shown with a simple graphical.

Clustering is a process which partitions a given data set into homogeneous groups based on given features such that similar objects are kept in a group whereas dissimilar objects are in different groups. It is the most important unsupervised learning problem. It deals with finding structure in a collection of unlabeled data. For better understanding please refer to **Fig I.**

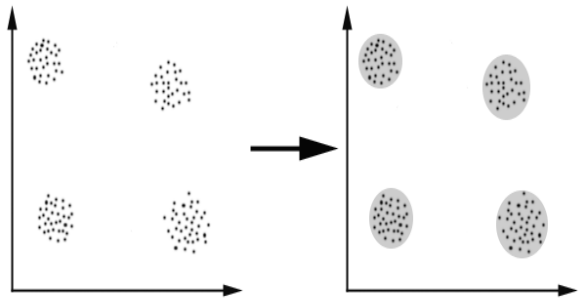

Figure 1: Example showing 4 cluster of data

In this case we easily identify the 4 clusters into which the data can be divided; the similarity criterion is distance*:* two or more objects belong to the same cluster if they are "close" according to a given distance (in this case geometrical distance).This is called distance-based clustering.
Another kind of clustering is conceptual clustering: two or more objects belong to the same cluster if this one defines a concept common to all that objects. In other words, objects are grouped according to their fit to descriptive concepts, not according to simple similarity measures. For clustering algorithm to be advantageous and beneficial some of the conditions need to be satisfied.

- Scalability - Data must be scalable otherwise we may get the wrong result.

- Clustering algorithm must be able to deal with different types of attributes.
- Clustering algorithm must be able to find clustered data with the arbitrary shape.
- Clustering algorithm must be insensitive to noise and outliers.
- Clustering algorithm must be able to deal with data set of high dimensionality.

2. CLUSTERING ALGORITHM APPLICATIONS

1) Clustering Algorithm in Identifying Cancerous Data

Clustering algorithm can be used in identifying the cancerous data set. Initially we take known samples of cancerous and non cancerous data set. Label both the samples data set. We then randomly mix both samples and apply different clustering algorithms into the mixed samples data set (this is known as learning phase of clustering algorithm) and accordingly check the result for how many data set we are getting the correct results (since this is known samples we already know the results beforehand) and hence we can calculate the percentage of correct results obtained. Now, for some arbitrary sample data set if we apply the same algorithm we can expect the result to be the same percentage correct as we got during the learning phase of the particular algorithm. On this basis we can search for the best suitable clustering algorithm for our data samples.

It has been found through experiment that cancerous data set gives best results with unsupervised non linear clustering algorithms and hence we can conclude the non linear nature of the cancerous data set.

2) Clustering Algorithm in Search Engines

Clustering algorithm is the backbone behind the search engines. Search engines try to group similar objects in one cluster and the dissimilar objects far from each other. It provides result for the searched data according to the nearest similar object which are clustered around the data to be searched. Better the clustering algorithm used, better are the chances of getting the required result on the front page. Hence, the definition of **similar object** play a crucial role in getting the search results, better the definition of similar object better the result is. Most of the brainstorming activities needs to be done for defining the criteria to be used for similar object.

3) clustering algorithm in academics

The ability to monitor the progress of students' academic performance has been the critical issue for the academic community of higher learning. Clustering algorithm can be used to monitor the students' academic performance. Based on the students' score they are grouped into different-different clusters (using k-means, fuzzy c-means etc), where each clusters denoting the different level of performance. By knowing the number of students' in each cluster we can know the average performance of a class as a whole.

4) clustering algorithm in wireless sensor network's based application

Clustering Algorithm can be used effectively in Wireless Sensor Network's based application. One application where it can be used is in Landmine detection. Clustering algorithm plays the role of finding the Cluster heads (or cluster center) which collects all the data in its respective cluster.

3. CLASSIFICATION OF CLUSTERING ALGORITHMS

As given in fig 2. We are considering Categorization of clustering algorithms is not easy. In reality, groups given below overlap. For convenience we provide a classification mainly hierarchical and partitioning methods.

- **Hierarchical Methods**
- **Partitioning Methods**

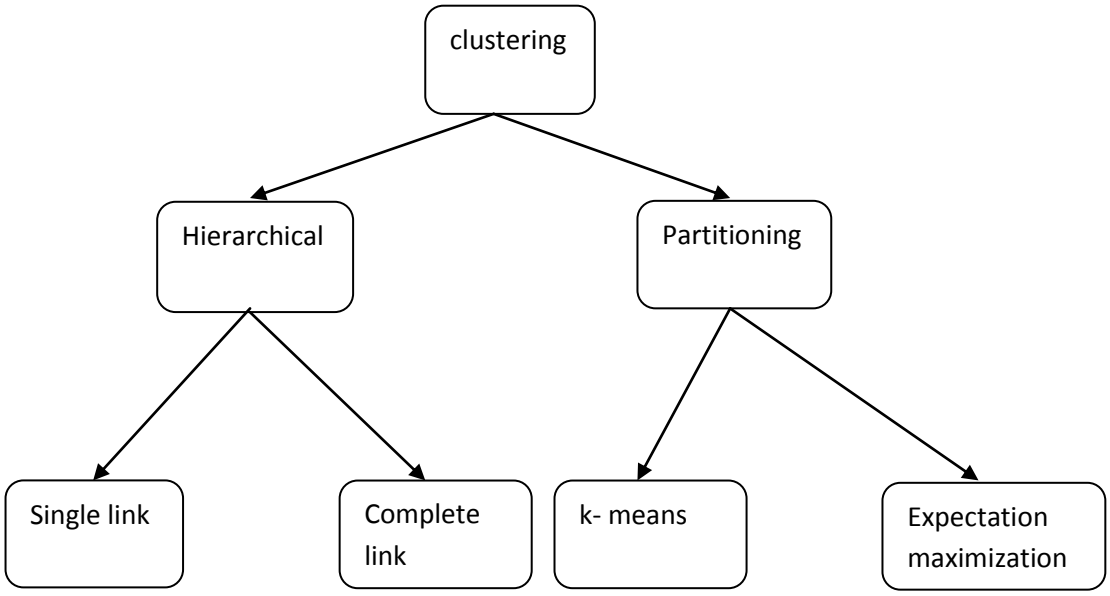

Fig 2. Classification of clustering

HIERARCHICAL METHODS

Hierarchical clustering constructs a hierarchy of clusters that can be illustrated in a tree structure which is also known as a *dendrogram*. Each node of the dendrogram, including the root, represents a cluster and the parent-child relationship among them enables us to explore different levels of clustering granularity.
There are mainly two types of algorithms for hierarchical clustering:
one is in an **agglomerative (bottom-up manner**)
the other is in a **divisive (top-down manner)**

a) HIERARCHICAL AGGLOMERATIVE CLUSTERING

An **agglomerative bottom-up manner** that the algorithm starts with all the objects and successively combine them into clusters.This algorithm works by grouping the data one by one on the basis of the nearest distance measure of all the pair wise distance and dissimilarity between the data point

Algorithmic steps for Agglomerative Hierarchical clustering

Let $X = \{x_1, x_2, x_3, ..., x_n\}$ be the set of data points.
1) Begin with the disjoint clustering having level $L(0) = 0$ and sequence number $m = 0$
2) Find the least distance pair of clusters in the current clustering, say pair (r), (s), according to $d[(r),(s)] = \min d[(i),(j)]$ where the minimum is over all pairs of clusters in the current clustering.
3) Increment the sequence number: $m = m + 1$.Merge clusters (r) and (s) into a single cluster to form the next clustering m. Set the level of this clustering to $L(m) = d[(r),(s)]$.

4) Update the distance matrix, D, by deleting the rows and columns corresponding to clusters (r) and (s) and adding a row and column corresponding to the newly formed cluster. The distance between the new cluster, denoted (r,s) and old cluster(k) is defined in this way: $d[(k), (r,s)] = \min (d[(k),(r)], d[(k),(s)])$.

b) DIVISIVE ANALYSIS (DIANA)

DIANA is a hierarchical clustering technique, but its main difference with the agglomerative method (AGNES) is that it constructs the hierarchy in the inverse order. that the algorithm starts with one whole cluster which includes all objects and recursively splits the clusters into smaller ones.

Algorithmic steps for Agglomerative Hierarchical clustering

1. Find the object, which has the highest average dissimilarity to all other objects. This object initiates a new cluster– a sort of a splinter group.
2. For each object i outside the splinter group compute
3. D_i = [average d(i,j) j $\notin R_{\text{splinter group}}$] - [average d(i,j) j $\in R_{\text{splinter group}}$]
4. Find an object h for which the difference D_h is the largest. If D_h is positive, then h is, on the average close to the splinter group.
5. Repeat Steps 2 and 3 until all differences D_h are negative. The data set is then split into two clusters.
6. Select the cluster with the largest diameter. The diameter of a cluster is the largest dissimilarity between any two of its objects. Then divide this cluster, following steps 1-4.
7. Repeat Step 5 until all clusters contain only a single object.

Advantages of hierarchical clustering

- ✓ Embedded flexibility regarding the level of granularity.
- ✓ Ease of handling of any forms of similarity or distance.
- ✓ Consequently, applicability to any attributes types.

Disadvantages of hierarchical clustering

- ✓ Vagueness of termination criteria.
- ✓ The fact that most hierarchical algorithms do not revisit once constructed (intermediate) clusters with the purpose of their improvement.

PARTITIONING METHODS

The partitioning methods generally result in a set of K clusters, each object belonging to one cluster. Each cluster may be represented by a centroid or a cluster representative; this is some sort of summary description of all the objects contained in a cluster. The precise form of this description will depend on the type of the object which is being clustered. In case where real-valued data is available, the arithmetic mean of the attribute vectors for all objects within a cluster provides an appropriate representative; If the number of the clusters is large, the centroids can be further clustered to produces hierarchy within a dataset.

Original Points Partitioned Points

Fig 3. Partition Clustered points in a domain The *k*-Means Algorithm

In briefly, *k*-means clustering is a top-down algorithm that classifies the objects into *k* number of groups with regard to attributes or features, where *k* is a positive integer number and specified apriori by users. The grouping is done by minimizing the sum of squares of distances between object and the corresponding cluster centroid.

Algorithmic steps for k-means clustering

Let $X = \{x_1,x_2,x_3,\ldots\ldots..,x_n\}$ be the set of data points and $V = \{v_1,v_2,\ldots\ldots.,v_c\}$ be the set of centers.

1) Randomly select *'c'* cluster centers.

2) Calculate the distance between each data point and cluster centers.

3) Assign the data point to the cluster center whose distance from the cluster center is minimum of all the cluster centers..

4) Recalculate the new cluster center using:

$$v_i = (1/c_i) \sum_{j=1}^{c_i} x_i$$

where, *'c_i'* represents the number of data points in i^{th} cluster.

5) Repeat steps 2 and 3 until all differences D_h are negative. The data set is then split into two clusters.

6) If no data point was reassigned then stop, otherwise repeat from step 3).

Advantages of k-means clustering

- ✓ Fast, robust and easier to understand.
- ✓ Gives best result when data set are distinct or well separated from each other.

Disadvantages of k-means clustering

- ✓ The learning algorithm requires apriori specification of the number of cluster centers.
- ✓ The use of Exclusive Assignment - If there are two highly overlapping data then k-means will not be able to resolve that there are two clusters.
- ✓ Randomly choosing of the cluster center cannot lead us to the fruitful result.
- ✓ Applicable only when mean is defined i.e. fails for categorical data.
- ✓ Unable to handle noisy data and outliers.

4. COMPARISION

Finally, the result carried out from the above study is listed in (Table 1) in the form of comparison of both the clustering algorithm which highlights the realistic approach as well as desirable features of the algorithm.

		Hierarchical	Partition
1	Running time	Slower	Faster
2	Assumptions	Needs only a similarity Measure	Needs stronger Assumptions
3	Input parameter	Not require	Need a number of cluster
4	Output	meaningful and subjective division of clusters.	k clusters.
5	Complexity	O (n^2)	O (i k n)
6	Suited	suited for categorical and non liner data	Not suited for categorical and non liner data
7	Efficiency	Comparatively less	Comparatively more

Table 1. Comparison b/w hierarchical and partition clustering

5. CONCLUSION

From the above study, it can be concluded that partition algorithm suited for large data sets and its performance increases as number of clusters increases. But its use is limited to numeric values. Therefore Agglomerative and Divisive Hierarchical algorithm was adopted for categorical data, but due to its complexity a new approach for assigning rank value to each categorical attribute using partition can be used in which categorical data is first converted into numeric by assigning rank. So that performance of partition algorithm is better than Hierarchical Clustering Algorithm.

REFERENCES

[1] M. Marin, A. van, Deursen, and L. Moonen. Identifying Aspects Using Fan-in Analysis. In Proceedings of the 11th Working Conference on Reverse Engineering (WCRE2004), pages 132-141. IEEE Computer Society, 2004.

[2] Pradeep Rai, Shubha Singh, A Survey of Clustering Techniques, International Journal of Computer Applications

(0975 – 8887), Volume 7– No.12, October 2010.

[3] Orlando Alejo Mendez Morales. Aspect Mining Using Clone Detection. Master's thesis, Delft University of Technology, The Netherlands, August 2004.

[4] D. Shepherd and L. Pollock. Interfaces, Aspects, and Views. In Proceedings of Linking AspectTechnology

[5] Evolution Workshop(LATE 2005), March 2005.

[6] P. Tonella and M. Ceccato. Aspect Mining through the Formal Concept Analysis of Execution Traces. In Proceedings of the IEEE Eleventh Working Conference on Reverse Engineering (WCRE 2004), pages 112_121, November 2004.

