International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 3 Issue 11, November-2014

A Review: Hierarchial Clustering Method

Rati Saxena
Computer Science Dept.
VIT University, Vellore

Palak Agrawal
Electronics and Communication Dept.
VIT University, Vellore

Radhika Vaid
Electrical and Electronics Dept.
VIT University, Vellore

Abstract:- **With advancing experience and; explosive growth of data to be handled from megabytes to terabytes, it's become necessary to own a mechanism for higher handling of data giving a complete fresh scope for the growth in this field. Data processing is precisely outlined as the extraction of data and the process to find the fascinating patterns of data. In this paper a brief outline of the hierarchical clustering algorithms used extensively now-a-days is described which is followed by the fundamental methodology of hierarchical clustering; its numerous varieties and; the complexness of the algorithmic program of hierarchical clustering. Lastly, the challenges of hierarchical clustering for huge information are mentioned.**

Keywords- Hierarchical; clustering; supervised learning; unsupervised learning; metric.

1. INTRODUCTION

Clustering is the task of collection of objects in such a way that objects within the same group is similar in a way. Since no cluster formula is exactly outlined hence we've got numerous cluster algorithms, with completely different techniques. [1] Clustering is beneficial in many alpha pattern-analysis, grouping, machine-learning things, image segmentation and process, and finally in pattern classification.

Clustering is unsupervised classification of information which implies to group together unlabeled data into purposeful clusters. These labels measure knowledge, that's obtained by the pattern of the information. Clustering mechanism is completely different from discrimination analysis mechanism that is supervised classification which implies that we've got a group of labeled patterns. These labeled patterns further describe classes which are used to label a new pattern. [2]

This paper starts with the definition of a number of the common terms that are utilized in the paper and are necessary for additional understanding. The section 3 explains the distance metric that is employed to live the similarity between two clusters. The section 4 gives introduction regarding hierarchal cluster by demonstrating an example. The section 5 tells about the types of hierarchical clustering and also compares its two types based on the complexity of each algorithm. Finally section 6 tells about clustering of large set of data

2. COMMON TERMS

The following terms have been used in the paper –

Pattern--Data instances of similar structure are grouped together into subsets called as a pattern.

Features-The individual component of a pattern is called as a feature of the pattern.

Metric-A metric is a measure of distance between pairs of observations which is used to decide similarity or dissimilarity between clusters

Linkage Criteria-The linkage criterion determines the distance between sets of observations.

3. DISTANCE METRICS

To classify the objects in step with similarity and difference, we have a proximity index. One of the proximity indexes that is used is a distance metric [7]. We have got three points x, y and z and a distance metric D that would satisfy the subsequent conditions:

A) Non-negativity: $D(x, y) \geq 0$
B) Symmetry :$D(x, y)=D(y, x)$
C) Triangular Inequality:
$D(x, z) \leq D(x, y) + D(y, z)$

The linkage criterion determines the distance between sets of observations as a function of the pair wise distances between observations. The linkage criteria which are used between two sets of observation A and B in the hierarchical clustering are –

Complete Linkage Clustering- This method is also called as the maximum linkage and it considers the maximum distance between any object of one cluster to any other object of second cluster

Max {D (a, b), a € A and b € B}

Single Linkage Clustering- This method is also called as the minimum linkage and it considers the minimum distance between any object of one cluster to any other object of second cluster.

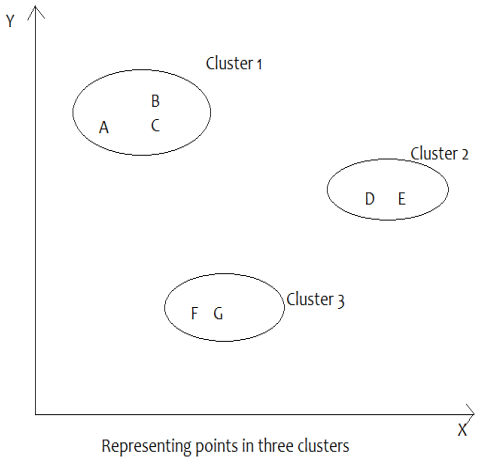
Min {D (a, b), a € A and b € B}

Average Linkage Clustering- This method is also called as the unweighted Pair Group Method with Arithmetic Mean and it considers the average distance from any member of one cluster to any member of the other cluster.

$$\frac{1}{|A||B|}\sum_{a\in A}\sum_{b\in B} D(a,b)$$

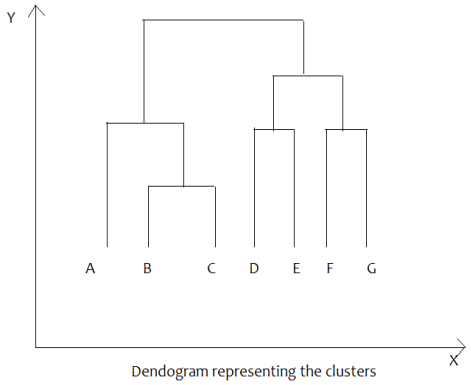
4. HIERARCHIAL CLUSTERING

Hierarchical clustering is additionally referred as the connectivity based clustering and it is outlined as the clustering technique where a hierarchy of clusters is constructed. This methodology follows the recursive partitioning strategy that can be either in a top-down or bottom-up fashion. The results of hierarchical clustering are described by using a dendogram. The concept is illustrated using figure 1. We have seven labeled patterns which is described on a dendogram, shown in figure 2. [9]



Representing points in three clusters

Figure 1



Dendogram representing the clusters

Figure 2

5. TYPES OF HIERARCHIAL CLUSTERING

The hierarchal clustering is of two types which can be defined as follows-

1) Agglomerative hierarchical clustering is a bottom-up agglomeration methodology where clusters have sub-clusters, and so on. It starts where every object form its own cluster and iteratively merges cluster into larger and bigger clusters, till all the clusters are in turn united and therefore the desired cluster structure is obtained.[4]The single cluster becomes the hierarchy's root. For the merging step, it finds the two clusters that are closest to each other, and combines the two to create one cluster.The process of agglomerative clustering can be described as –
   1) Assign each object to a cluster.
   2) Calculate pair-wise distance between each cluster using distance metric.
   3) Construct a distance matrix.
   4) Take the pair of clusters having shortest distance and then remove it from the matrix and merge them.
   5) Use this cluster to measure the distance to other clusters and then update the matrix.
   6) Continue the process until, the desired cluster is obtained.

2) Divisive is a top-down clustering methodology. This methodology starts with one cluster containing all objects, then in turn splits ensuing clusters till only clusters of individual objects stay and also the desired cluster is obtained. This methodology is almost like the agglomerative approach except that it starts from the only cluster that consists of all the objects. However, this methodology is less normally used.[6] Examples for this algorithms are LEGCLUST [5], BRICH [8] (Balance iterative Reducing and clustering using Hierarchies), CURE (Cluster using Representatives) [10], and Chemeleon [3].

The complexity of agglomerative clustering is O(n^3) , and for divisive clustering is O(2^n), that is even worse. However, we have got an optimal efficient agglomerative methods referred to as single-linkage and complete-linkage clustering which have a complexity of O(n^2).

The complete-link algorithm produces tightly bound clusters whereas the single-link algorithm has chaining result and therefore it produces elongated clusters. There are two clusters in Figures three and four that are suffering from noisy patterns. Figure three shows the clusters made by the single link algorithm and figure four shows the clusters made by the complete link algorithm. It can be observed from Figure three and four that the clusters obtained by the complete link algorithm are a lot compact than those obtained by the single-link algorithm; the cluster obtained using the single-link algorithm is elongated due to the noisy patterns. Thus, single-link algorithm is additional versatile than the complete-link algorithm.

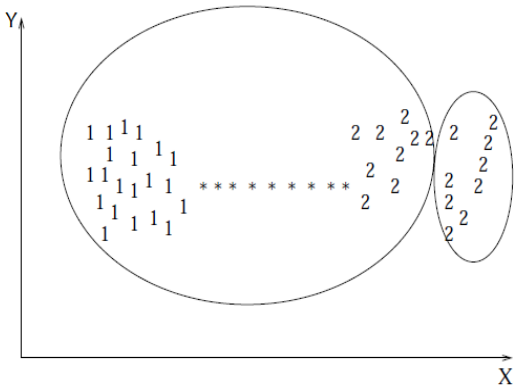


Figure 3

A Single link clustering pattern under noisy patters which are shown by '*'.

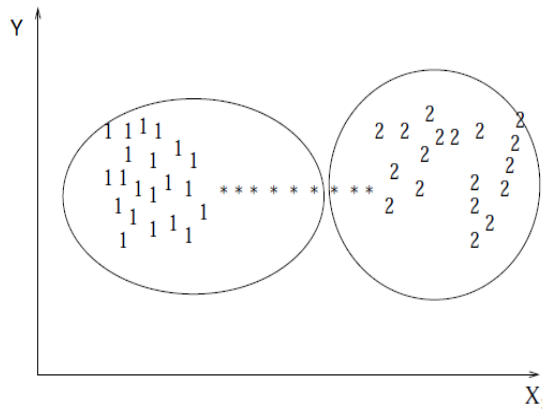


Figure 4
A Complete link clustering pattern under noisy patters which are shown by '*'.

6. CLUSTERING LARGE DATA SETS

There are many applications where it is necessary to cluster an outsized collection of patterns. In document retrieval, countless instances with a dimensionality of more than one hundred need to be clustered to realize data abstraction. A majority of the approaches and algorithms proposed within the literature cannot handle such giant knowledge sets. The hierarchical clustering has the following advantages over other clustering other methods –

1) Versatility — The single-link strategies, maintain good performance on information sets containing non-isotropic clusters, as well as well separated, chain-like and coaxial clusters.
2) Multiple partitions — hierarchical ways produce not one partition, but multiple nested partitions, which permit completely different users to choose different partitions.

7. RESULTS AND CONCLUSIONS

The paper describes hierarchical clustering and its sorts and parameters associated in larger data sets. It additionally explains the complexity of the strategy used and compares the two kinds of hierarchical clustering thus giving the conclusion that complete line agglomerative clustering is better for pragmatic purpose. Hierarchical clustering has been adopted for categorical knowledge and its useful because of its versatility. In summary, hierarchical clustering is an appealing and challenging problem having use in numerous applications like image segmentation, object recognition and data filtering and retrieval.

8. REFERENCES

[1] S.Anitha Elavarasi and Dr. J. Akilandeswari and Dr. B. Sathiyabhama, January 2011, A Survey On Partition Clustering Algorithms.
[2] Manish Verma, Mauly Srivastava, Neha Chack, Atul Kumar Diswar, Nidhi Gupta," A Comparative Study of Various Clustering Algorithms in Data Mining," International Journal of Engineering Reserch and Applications (IJERA), Vol. 2, Issue 3, pp.1379-1384, 2012
[3] Jiawei Han, Micheline Kamber, "Data Mining Concepts and Techniques" Elsevier Publication.
[4] Improved Outcome Software, Agglomerative Hierarchical Clustering Overview. Retrieved from:http://www.improvedoutcomes.com/docs/WebSiteDocs/Clustering/Agglomerative_Hierarchical_Clustering_Overview.htm [Accessed 22/02/2013].
[5] S. Guha, R. Rastogi, and K. Shim, 1998. CURE: An Efficient Clustering Algorithm for Large Databases. Proc. ACM International Conf. Management of Data: 73-84.
[6] Han, J., Kamber, M. 2012. Data Mining: Concepts and Techniques, 3rd edition, 443-491.
[7] Periklis Andritsos, University of Toronto Department of Computer Science, March 11, 2002: "Data Clustering Techniques".
[8] M. Livny, R.Ramakrishnan, T. Zhang, 1996. BIRCH: An Efficient Clustering Method for Very Large Databases. Proceeding ACMSIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery: 103-114.
[9] Hierarchial Clustering Analysis .Retreived from: "http://www.econ.upf.edu/~michael/ Stanford/ maeb7 .pdf"
[10] S. Guha, R. Rastogi, and K. Shim, 1998. CURE: An Efficient Clustering Algorithm for Large Databases. Proc. ACM Int'l Conf. Management of Data: 73-84.