IJRSET Volume 2, Issue 2 Pages: 59-64

ISSN 2394-739X

International Journal for Research in Science Engineering and Technology

ANALYSIS OF HIERARCHICAL CLUSTERING

[1] M. Premalatha, [2] T. Menaka,
[1] M.Phil Scholar, [2] Assistant Professor,
[1] Department of Computer Science (Aided), [2] Department of Computer Science,
[1, 2] NGM College, Pollachi, India.

Abstract:-

Data mining, the extraction of hidden predictive information from large databases, is a powerful, new technology with great potential to help companies that focus on the most important information in their data warehouses. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. The automated, prospective analyses offered by data mining move beyond the analyses of past events provided by retrospective tools typical of decision support systems. Data mining tools can answer business questions that traditionally are time consuming to resolve. The purpose of this article is to explain the notion of clustering and a concrete clustering method agglomerative and divisive hierarchical clustering.

Keywords: - Data Mining, Dendrogram, Agglomerative, Divisive, Hierarchical Clustering.

1. INTRODUCTION

Clustering can be considered the most important unsupervised learning problem. So, as every other problem of this kind, it deals with finding a structure in a collection of unlabeled data. A cluster is therefore a collection of objects which are “similar” between them and are “dissimilar” to the objects belonging to other clusters. Cluster Analysis, also called data segmentation, has a variety of goals. They all relate to grouping or segmenting a collection of objects into subsets or "clusters", such that those within each cluster are more closely related to one another than objects assigned to different clusters. Different types of similarity measures may be used to identify classes (clusters), where the similarity measure controls how the clusters are formed. Hierarchical clustering builds a cluster hierarchy known as a dendrogram. Hierarchical clustering methods are categorized into agglomerative and divisive. An agglomerative clustering starts with one-point clusters and recursively merges two or more most appropriate clusters. A divisive clustering starts with one cluster of all data points and recursively splits the most appropriate cluster. The process continues until a stopping criterion is achieved.

2. HIERARCHICAL CLUSTERING

Hierarchical clustering involves creating clusters that have a predetermined ordering from top to bottom. Hierarchical clustering builds a cluster hierarchy known as a dendrogram. Dendrogram is a convenient graphic to display a hierarchical sequence of clustering assignments. Every cluster node contains child clusters; sibling clusters partition the points covered by their common parent. Clustering obtained by cutting the dendrogram at a desired level: each connected component forms a cluster.

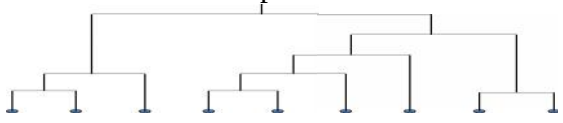

Figure 1: Dendrogram

The hierarchical, binary cluster tree created by the linkage function is most easily understood when viewed graphically. The Statistics Toolbox function dendrogram plots the tree as follows.

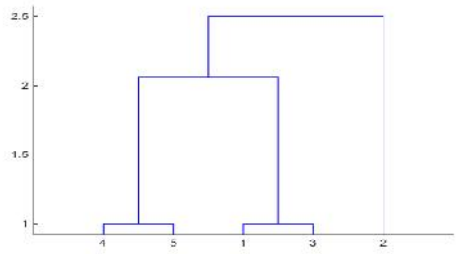

Figure 2: Statistics Toolbox function dendrogram

In the figure, the numbers along the horizontal axis represent the indices of the objects in the original data set. The links between objects are represented as upside-down U-shaped lines. The height of the U indicates the distance between the objects. For example, the link representing the cluster containing objects 1 and 3 has a height of 1. The link representing the cluster that groups object 2 together with objects 1, 3, 4, and 5, has a height of 2.5. The height represents the distance linkage computes between objects 2 and 8. For example, all files and folders on the hard disk are organized in a hierarchy. There are two types of hierarchical clustering, Divisive and Agglomerative.

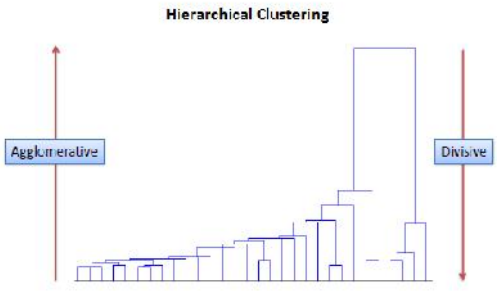

Figure 3: Hierarchical clustering

In data mining, hierarchical clustering (also called hierarchical cluster analysis or HCA) is a method of cluster analysis which seeks to build a hierarchy of clusters. In hierarchical clustering the data are not partitioned into a particular cluster in a single step. Instead, a series of partitions takes place, which may run from a single cluster containing all objects to N clusters each containing a single object. Hierarchical Clustering is subdivided into agglomerative methods, which proceed by series of fusions of the N objects into groups, and divisive methods, which separate N objects successively into finer groupings. Agglomerative techniques are more commonly used, and this is the method implemented in the free version of XLMiner™ which is the Microsoft Office Excel add-in. Connectivity based clustering, also known as hierarchical clustering, is based on the core idea of objects being more related to nearby objects than to objects farther away. These algorithms connect "objects" to form "clusters" based on their distance. A cluster can be described largely by the maximum distance needed to connect parts of the cluster. At different distances, different clusters will form, which can be represented using a dendrogram, which explains where the common name "hierarchical clustering" comes from: these algorithms do not provide a single partitioning of the data set, but instead provide an extensive hierarchy of clusters that merge with each other at certain distances. In a dendrogram, the y-axis marks the distance at which the clusters merge, while the objects are placed along the x-axis such that the clusters don't mix.

Connectivity based clustering is a whole family of methods that differ by the way distances are computed. Apart from the usual choice of distance functions, the user also needs to decide on the linkage criterion to use. Popular choices are known as single-linkage clustering (the minimum of object distances), complete linkage clustering (the maximum of object distances) or UPGMA ("Unweighted Pair Group Method with Arithmetic Mean"). It also known as average linkage clustering. Furthermore, hierarchical clustering can be agglomerative (starting with single elements and aggregating them into clusters) or divisive (starting with the complete data set and dividing it into partitions). These methods will not produce a unique partitioning of the data set, but a

hierarchy from which the user still needs to choose appropriate clusters. They are not very robust towards outliers, which will either show up as additional clusters or even cause other clusters to merge In the general case, the complexity is which makes them too slow for large data sets. For some special cases, optimal efficient methods are known:
SLINK for single-linkage and CLINK for complete-linkage clustering.

A. GENERAL STEPS OF HIERARCHICAL CLUSTERING

Given a set of N items to be clustered, and an N*N distance (or similarity) matrix, the basic process of hierarchical clustering is this:

- Start by assigning each item to a cluster, so that if you have N items, you now have N clusters, each containing just one item. Let the distances between the clusters the same as the distances between the items they contain.
- Find the closest pair of clusters and merge them into a single cluster, so that now you have one cluster less.
- Compute distances between the new cluster and each of the old clusters.
- Repeat steps 2 and 3 until all items are clustered into K number of clusters
-

B. HIERARCHICAL CLUSTERING: TIME AND SPACE REQUIREMENTS

- O(N2) space since it uses the proximity matrix.
- N is the number of points.
- O(N3) time in many cases
- There are N steps and at each step the size, N2, proximity matrix must be updated and searched
- Complexity can be reduced to O(N2 log(N)) time for some approaches

3. AGGLOMERATIVE (BOTTOM-UP) CLUSTERING

Start individual clusters, at each step, merge the closest pair of clusters until only one cluster (or k clusters) left. Agglomerative is more popular and simpler than divisive but less accurate.

Typical alternatives to calculate the distance between clusters:

A. Single link

The distance between groups is defined as the distance between the closest pair of objects, where only pairs consisting of one object from each group is considered i.e. the distance between two clusters is given by the value of the shortest link between clusters. At each stage the two clusters for which the distance is minimum are merged Smallest distance between an element in one cluster and an element in the other, i.e., dis(Ki, Kj) = min(tip, tjq) .

B. Complete link

Complete linkage clustering (farthest neighbor) is the opposite of the single linkage i.e. distance between groups is defined as the distance between the most distant pair of objects, one from each group. At each stage the two clusters for which the distance is minimum are merged; largest distance between an element in one cluster and an element in the other, i.e., dis (Ki, Kj) = max (tip, tjq)

C. Average

The distance between two clusters is defined as the average of distances between all pairs of objects, where each pair is made up of one object from each group. At each stage the two clusters for which the distance is minimum are merged;
Average distance between an element in one cluster and an element in the other, i.e., dis (Ki, Kj) = avg (tip, tjq)

D. Centroid

In this method, groups once formed are represented by their mean values for each variable, that is their mean vector and inter-group distance is defined in terms of distance between two such mean vectors. At each stage the two clusters for which the distance is minimum are merged. In this case, those two clusters are merged such that the newly formed cluster, on average, will have minimum pairwise distances between the points in it; distance between the centroids of two clusters, i.e., dis(Ki, Kj) = dis(Ci, Cj)

E. Ward's hierarchical clustering

Ward (1963) proposed a clustering procedure seeking to form the partitions P ,..., P1 n in a manner that minimizes the loss associated with each grouping and to quantify that loss in aform that is readily interpretable. At each step the union of every possible cluster pair is considered and the two clusters whose fusion results in minimum increase in "information loss" are combined. Information loss is defined by Ward in terms of an error sum-of-squares criterion.

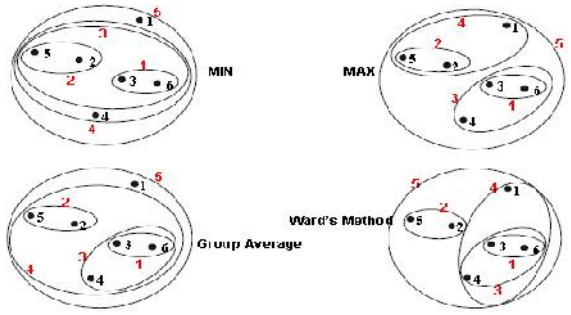

Figure 4: Typical Alternatives to Calculate the Distance Between Clusters

AGGLOMERATIVE CLUSTERING ALGORITHM

Agglomerative clustering algorithm is more popular hierarchical clustering technique

- Compute the proximity matrix
- Let each data point be a cluster
- Repeat
- Merge the two closest clusters
- Update the proximity matrix
- Until only a single cluster remains
- Key operation is the computation of the proximity of two clusters
- Different approaches to defining the distance between clusters distinguish the different algorithms

4. DIVISIVE (TOP-DOWN) CLUSTERING

Start with one cluster, at each step, split a cluster until each cluster contains a point. So far we have only looked at agglomerative clustering, but a cluster hierarchy can also be generated top-down. This variant of hierarchical clustering TOP-DOWN is called top-down clustering or divisive clustering. We start at the top with all CLUSTERING documents in one cluster. The cluster is split using a flat clustering algorithm. This procedure is applied recursively until each document is in its own singleton cluster. Top-down clustering is conceptually more complex than bottom-up clustering since we need a second, flat clustering algorithm as a "subroutine". It has the advantage of being more efficient if we do not generate a complete hierarchy all the way down to individual document leaves. For a fixed number of top levels, using an efficient flat algorithm like K-means, top-down algorithms are linear in the number of documents and clusters. So they run much faster than HAC algorithms, which are at least quadratic. There is evidence that divisive algorithms produce more accurate hierarchies than bottom-up algorithms in some circumstances. See the references on bisecting K-means in Section 17.9. Bottom-up methods make clustering decisions based on local patterns without initially taking into account the global distribution. These early decisions cannot be undone. Top-down clustering benefits from complete information about the global distribution when making top-level partitioning decisions.

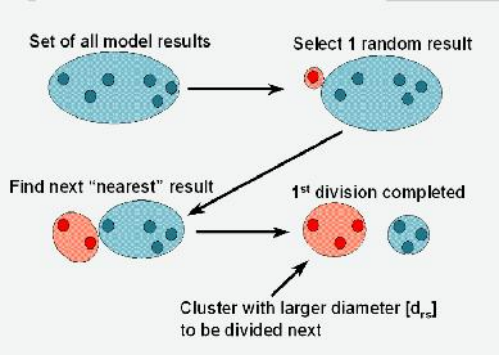

Figure 5: Divisive Clustering process

DIVISIVE CLUSTERING ALGORITHM

1. Put all objects in one cluster
2. Repeat until all clusters are singletons

a) choose a cluster to split

- what criterion?

b) replace the chosen cluster with the sub clusters

- split into how many?
- how split?

• "reversing" agglomerative => split in two
• cutting operation: cut-based measures seem to be a natural choice.
• focus on similarity across cut - lost similarity.
• not necessary to use a cut-based measure

5. COMPARISON BETWEEN AGGLOMERATIVE AND DIVISIVE:

Agglomerative	Divisive
Start with each document being a single cluster.	Start with all documents belong to the same cluster.
Eventually all documents belong to the same cluster.	Eventually each node forms a cluster on its own.
Recursively add two or more appropriate clusters	Recursively divide into smaller clusters
Until there is only one cluster repeatedly merge the two groups that have the smallest dissimilarity	Until all points are in their own cluster repeatedly split the group into two resulting in the biggest dissimilarity
Stop when k number of clusters is achieved.	Stop when k number of clusters is achieved.

6. MERITS OF HIERARCHICAL CLUSTERING

- Embedded flexibility regarding the level of granularity
- Ease of handling of any forms of similarity or distance
- Consequently, applicability to any attribute types
- No need to assume any particular number of clusters
- Any desired number of clusters can be obtained by 'cutting' the dendrogram at the proper level
- They may correspond to meaningful taxonomies, Example in biological sciences
- Traditional hierarchical algorithms use a similarity or distance matrix to merge or split one cluster at a time

7. DEMERITS 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
- Once a decision is made to combine two clusters, it cannot be undone
- No objective function is directly minimized
- Different schemes have problems with one or more of the following:
- Sensitivity to noise and outliers
- Difficulty handling different sized clusters and convex shapes
- Breaking large clusters

CONCLUSION

Data mining software allows users to analyze large databases to solve business decision problems. Data mining is, in some ways, an extension of statistics, with a few artificial intelligence and machine learning. Like statistics, data mining is not a business solution, it is just a technology. Clustering is important in data analysis and data mining applications. It is the task of grouping a set of objects so that objects in the same group are more similar to each other than to those in other groups. Clustering can be done by the different no. of algorithms such as hierarchical, partitioning, grid and density based algorithms. Hierarchical clustering is the connectivity based clustering. The two types of hierarchical clustering agglomerative and divisive and its merits and demerits are analysed in this article. The probabilistic scheme allows for automatic cluster and hierarchy level assignment for unseen data and further a natural technique for interpretation of the clusters.

REFERENCES

[1] J. Han, M. Kamber. Data mining: Concepts and Techniques, Morgan Kaufmann, 2000.
[2] C. Fraley, Algorithms for Model-Based Hierarchical Clustering, SIAM J. Sci. Comput. 20, 1 (1998) 279–281.
[3] Kaufman, L.; Rousseeuw, P.J. (1990). Finding Groups in Data: An Introduction to Cluster Analysis (1 ed.). New York: John Wiley. ISBN 0-471-87876-6.
[4] Pradeep Rai, Shubha Singh" A Survey of Clustering Techniques" International Journal of Computer Applications, October 2010.
[5] Pavel Berkhin, "A Survey of Clustering Data Mining Techniques", pp.25-71, 2002.
[6] M.Vijayalakshmi, M.Renuka Devi, "A Survey of Different Issue of Different clustering Algorithms Used in Large Data sets" , International Journal of Advanced Research in Computer Science and Software Engineering, pp.305-307, 2012.
[7] Anoop Kumar Jain, Prof. Satyam Maheswari "Survey of Recent Clustering Techniques in Data Mining", International Journal of Computer Science and Management Research, pp.72-78, 2012.
[8] P. S. Bradley, U. Fayyad, and C. Reina. Scaling clustering algorithms to large databases. In Proc. 4th
[9] L. Kaufman and P.J. Rousseeuw. (1990) Finding Groups in Data: an Introduction to Cluster Analysis, John Wiley & Sons
[10] A. Topchy, A. K. Jain, and W. Punch, "A mixture model for clustering ensembles," in Proc. SIAM Conf. Data Mining, 2004, pp. 379-390.
[11] A. K Jain, M. N. Murty, and P. J. Flynn "Data clustering: A review, "ACM Comput. Surveys, vol.31, pp. 264-323, 1999.
[12] A. Mirzaei, M. Rahmati, and M. Ahmadi, "A new method for hierarchical clustering combination," Intell. Data Anal., vol. 12, no. 6, pp. 549-571, 2008.