Agglomerative Hierarchical Clustering Algorithm- A Review

K.Sasirekha, P.Baby

Department of CS, Dr.SNS.Rajalakshmi College of Arts & Science

Abstract- Clustering is a task of assigning a set of objects into groups called clusters. In data mining, hierarchical clustering is a method of cluster analysis which seeks to build a hierarchy of clusters. Strategies for hierarchical clustering generally fall into two types:Agglomerative: This is a "bottom up" approach: each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy.Divisive: This is a "top down" approach: all observations start in one cluster, and splits are performed recursively as one moves down the hierarchy.

Index Terms- Agglomerative, Divisive

I. INTRODUCTION

Fast and robust clustering algorithms play an important role in extracting useful information in large databases. The aim of cluster analysis is to partition a set of N object into C clusters such that objects within cluster should be similar to each other and objects in different clusters are should be dissimilar with each other[1]. Clustering can be used to quantize the available data, to extract a set of cluster prototypes for the compact representation of the dataset, into homogeneous subsets.

Clustering is a mathematical tool that attempts to discover structures or certain patterns in a dataset, where the objects inside each cluster show a certain degree of similarity. It can be achieved by various algorithms that differ significantly in their notion of what constitutes a cluster and how to efficiently find them. Cluster analysis is not an automatic task, but an iterative process of knowledge discovery or interactive multi-objective optimization. It will often necessary to modify preprocessing and parameter until the result achieves the desired properties.

In Clustering, one of the most widely used algorithms is agglomerative algorithms. In general, the merges and splits are determined in a greedy manner. The results of hierarchical clustering are usually presented in a dendrogram.In the general case, the complexity of agglomerative clustering is $\mathcal{O}(n^3)$, which makes them too slow for large data sets. Divisive clustering with an exhaustive search is $\mathcal{O}(2^n)$, which is even worse. However, for some special cases, optimal efficient agglomerative methods (of complexity $\mathcal{O}(n^2)$) are known: SLINK[1] for single-linkage and CLINK[2] for complete-linkage clustering.

II. DISADVANTAGES

1) Very sensitive to good initialization
2) Coincident clusters may result

Because the columns and rows of the typicality matrix are independent of each other

Sometimes this could be advantageous (start with a large value of c and get less distinct clusters)

Cluster dissimilarity: In order to decide which clusters should be combined (for agglomerative), or where a cluster should be split (for divisive), a measure of dissimilarity between sets of observations is required. In most methods of hierarchical clustering, this is achieved by use of an appropriate metric (a measure of distance between pairs of observations), and a linkage criterion which specifies the dissimilarity of sets as a function of the pairwise distances of observations in the sets.

Metric:

The choice of an appropriate metric will influence the shape of the clusters, as some elements may be close to one another according to one distance and farther away according to another. For example, in a 2-dimensional space, the distance between the point (1,0) and the origin (0,0) is always 1 according to the usual norms, but the distance between the point (1,1) and the origin (0,0) can be 2, $\sqrt{2}$or 1 under Manhattan distance, Euclidean distance or maximum distance respectively.

Some commonly used metrics for hierarchical clustering are:[3]

Names	**Formula**
Euclidean distance	$\|a-b\|_2 = \sqrt{\sum_i (a_i - b_i)^2}$
squared Euclidean distance	$\|a-b\|_2^2 = \sum_i (a_i - b_i)^2$
Manhattan distance	$\|a-b\|_1 = \sum_i \|a_i - b_i\|$
maximum distance	$\|a-b\|_\infty = \max_i \|a_i - b_i\|$
Mahalanobis distance	$\sqrt{(a-b)^\top S^{-1}(a-b)}$ where *S* is the covariance matrix

cosine similarity $\frac{a \cdot b}{\|a\|\|b\|}$

For text or other non-numeric data, metrics such as the Hamming distance or Levenshtein distance are often used.A review of cluster analysis in health psychology research found that the most common distance measure in published studies in that research area is the Euclidean distance or the squared Euclidean distance.

The linkage criterion determines the distance between sets of observations as a function of the pairwise distances between observations.

Some commonly used linkage criteria between two sets of observations *A* and *B* are:

Names	**Formula**
Maximum or complete linkage clustering	$\max \{ d(a,b) : a \in A,\, b \in B \}.$
Minimum or single-linkage clustering	$\min \{ d(a,b) : a \in A,\, b \in B \}.$
Mean or average linkage clustering, or UPGMA	$\frac{1}{\|A\|\|B\|} \sum_{a \in A} \sum_{b \in B} d(a,b).$
Minimum energy clustering	$\frac{2}{nm} \sum_{i,j=1}^{n,m} \|a_i - b_j\|_2 - \frac{1}{n^2} \sum_{i,j=1}^{n} \|a_i - a_j\|_2 - \frac{1}{m^2} \sum_{i,j=1}^{m} \|b_i - b_j\|_2$

where *d* is the chosen metric. Other linkage criteria include:

- The sum of all intra-cluster variance.
- The decrease in variance for the cluster being merged (Ward's criterion)

A simple agglomerative clustering algorithm is described in the single-linkage clustering page; it can easily be adapted to different types of linkage (see below).

Suppose we have merged the two closest elements *b* and *c*, we now have the following clusters {*a*}, {*b*, *c*}, {*d*}, {*e*} and {*f*}, and want to merge them further. To do that, we need to take the distance between {a} and {b c}, and therefore define the distance between two clusters. Usually the distance between two clusters $\mathcal{A}$and $\mathcal{B}$is one of the following:

- The maximum distance between elements of each cluster (also called complete-linkage clustering):

$\max\{ d(x,y) : x \in \mathcal{A},\, y \in \mathcal{B} \}.$

- The minimum distance between elements of each cluster (also called single-linkage clustering):

$\min\{ d(x,y) : x \in \mathcal{A},\, y \in \mathcal{B} \}.$

- The mean distance between elements of each cluster (also called average linkage clustering, used e.g. in UPGMA):

$\frac{1}{|\mathcal{A}| \cdot |\mathcal{B}|} \sum_{x \in \mathcal{A}} \sum_{y \in \mathcal{B}} d(x,y).$

- The sum of all intra-cluster variance.
- The increase in variance for the cluster being merged (Ward's method[6])
- The probability that candidate clusters spawn from the same distribution function (V-linkage).

Each agglomeration occurs at a greater distance between clusters than the previous agglomeration, and one can decide to stop clustering either when the clusters are too far apart to be merged (distance criterion) or when there is a sufficiently small number of clusters (number criterion).

Divisive Hierarchical Clustering

- A top-down clustering method and is less commonly used. It works in a similar way to agglomerative clustering but in the opposite direction. This method starts with a single cluster containing all objects, and then successively splits resulting clusters until only clusters of individual objects remain. GeneLinker™ does not support divisive hierarchical clustering.

Disadvantages

- No provision can be made for a relocation of objects that may have been 'incorrectly' grouped at an early stage. The result should be examined closely to ensure it makes sense.
- Use of different distance metrics for measuring distances between clusters may generate different results. Performing multiple experiments and comparing the results is recommended to support the veracity of the original results.

III. CONCLUSION

Agglomerative hierarchical clustering is a bottom-up clustering method where clusters have sub-clusters, which in turn have sub-clusters, etc. The classic example of this is species taxonomy. Gene expression data might also exhibit this hierarchical quality (e.g. neurotransmitter gene families). Agglomerative hierarchical clustering starts with every single object (gene or sample) in a single cluster. Then, in each successive iteration, it agglomerates (merges) the closest pair of clusters by satisfying some similarity criteria, until all of the data is in one cluster.

Advantages:It can produce an ordering of the objects, which may be informative for data display.

Smaller clusters are generated, which may be helpful for discovery. determine the similarity between prototypes and data points, and it performs well only in .

IV. FUTURE WORK

This paper was intened to compare between two algorithms.Through my extensive search I was unable to find any study that attempts to comapre between all algoritms under investigation.

As a future work comparison between these algoritms can be attempted according to different factors other than those considered in this paper.Comparing between the results of algorithms using normalized data or non-normalizes data will give different results.Ofcourse normalization will affect the performance of the algorithm and quality of the results.

Another approach may consider using data clustering algorithms in applications such as object and character recognition or information retrieval which is concerned with automatic documents.

REFERENCES

[1] M.S.Yang," A Survey of hierarchical clustering" Mathl. Comput. Modelling Vol. 18, No. 11, pp. 1-16, 1993.

[2] A. vathy-Fogarassy, B.Feil, J.Abonyi"Minimal Spanning Tree based clustering" Proceedings of World academy of Sc., Eng & Technology, vol-8, Oct-2005, 7-12.

[3] Pal N.R, Pal K, Keller J.M. and Bezdek J.C, "A Possibilistic Clustering Algorithm", IEEE Transactions on Fuzzy Systems, Vol. 13, No. 4, Pp. 517–530, 2005.

[4] R. Krishnapuram amd J.M. Keller, "A possibilistic approach to clustering", IEEE Trans. Fuzzy Systems, Vol. 1, Pp. 98-110, 1993.

[5] J. C. Dunn (1973): "A Agglomerative Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters", Journal of Cybernetics 3: 32-57

AUTHORS

First Author – K.Sasirekha MCA, M.Phil., Assistant Professor, Dr.SNS.Rajalakshmi College of Arts & Science,Chinnavedampatti,Coimbatore.
Email-id:sasirekharamesh1985@gmail.com
Second Author – P.Baby, MCA, M.Phil., Assistant Professor , Dr.SNS.Rajalakshmi College of Arts & Science,Chinnavedampatti, Coimbatore.
Email-id:cb.ridhu@gmail.com