Hierarchical Clustering

Andrei Karpau
Institute of Technology Blanchardstown
Dublin 16

December 4, 2016

Abstract

Hierarchical clustering is powerful clustering technique that provides the hierarchical structure of grouping inside the data set. This paper makes an overview of this methodology and main principles behind it. Also, it describes the top down and bottom up approaches of hierarchical grouping as well as different techniques of computing the proximities between clusters.

Keywords: hierarchical clustering, agglomerative clustering, divisive clustering, linkage, unsupervised learning

1 Introduction

Clustering is the unsupervised classification of data items, observations or feature vectors into the groups named clusters (Jain et al., 1999, p. 264). The objects in these groups are similar to each other in some sense, which usually refers to similarity measure, a real-valued function of similarity between objects. Nowadays, clustering is widely used in different data mining tasks. In contrast with classification it is the unsupervised learning technique, which means that the initial data is unlabeled and there is no initial knowledge of how many groups exists or what actually causes the grouping. There are plenty of different clustering techniques, so it may be implemented on hierarchy, partition, density, grid, constraint, subspace and so on (Li et al., 2012, p. 44).
Hierarchical clustering is an example of powerful and widely used technique that is based on finding the successive clusters using the previously established ones. So, it uncovers the hierarchical structure of groups with the single all-inclusive group at the top and multiple single-point groups at the bottom (Li et al., 2012, p. 45). After the clustering is done the results are often presented in a dendrogram, which is a tree-structured graph that shows the whole hierarchy. Since it visualizes all the possible numbers of clusters, it helps to build an overview of the items that can be grouped together and of how big the clusters are. So, hierarchical clustering is not only a powerful technique, but also a useful first step used for understanding the hidden structure of data set.

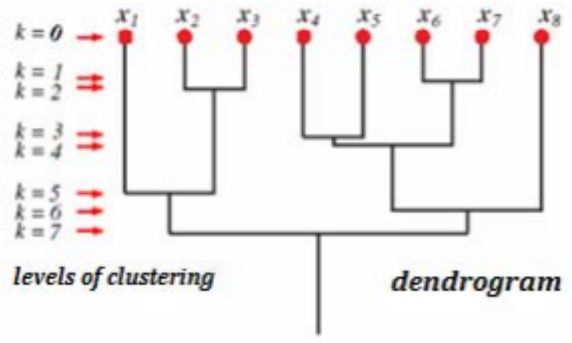

Figure 1: Shows the dendrogram of hierarchical model (Rafsanjani et al., 2012, p. 232)

1.1 Contribution and Structure

This paper addresses the hierarchical clustering technique. It makes the overview of the main idea of hierarchical clustering and of the principles that are behind it. Also, it considers different types and the advantages and disadvantages of this model. Finally, it describes in detail some popular algorithms of building it. This paper is structured as follow: in Section 1 is the introduction part; Section 2 provides the overview of the model with the main principles of how it works; Section 3 takes a look into both divisive clustering and agglomerative clustering algorithms and into different methods of computing the merge and split points of the clusters; and the last section 4 makes the final conclusions.

2 Hierarchical Clustering

Hierarchical clustering tends to build a tree or a graph that groups the most similar points on each level of hierarchy. As a result, on one side of the graph is one cluster that includes all the points. On the other are the clusters that include only one item. The example of the final structure is presented in the dendrogram in the figure 1.

The building of the model can be done either in top-down or in bottom-up direction. So, there are two main types of methods used for constructing it (Maimon and Rokach, 2005, pp. 330–331):

1. Agglomerative hierarchical clustering it is a bottom up approach where each point is initially considered as its own cluster. Then the merging of pairs is made recursively on each level of hierarchy until the one all-inclusive cluster is built on the top.

2. Divisive hierarchical clustering it is a top-down approach where all points are initially considered as a single cluster. Then the splitting is made

recursively as one moves down the hierarchy. The process is stopped when the all the clusters contain only one point.

Both techniques provide approximately the same results, however both have high computational complexity. For example, the nave implementation of agglomerative clustering is $O(n^3)$, where n is the number of items (Feng et al., 2010, p. 1217). Some implementations of divisive algorithms tend to provide even higher complexity $O(2^n)$. This means that when the number of items increases, the number of calculations that should be made increases with the same speed as the complexity function. As a result, it is inefficient and sometimes even impossible to use these techniques with the vast amount of data. Although, there are some optimized implementations of agglomerative algorithms their complexity is still pretty high $O(n^2 \log n)$ (Feng et al., 2010, p. 1217). The space complexity of agglomeration algorithms is $O(n^2)$, which is also high (Jain et al., 1999, p. 293). The reason of it is the need of the computation of proximity matrix that is of size n^2 and that should be stored. There is a possibility of computing the matrix based on the need, but this makes the computational complexity even higher (Jain et al., 1999, p. 293).
The final results of hierarchical clustering are usually presented in a dendrogram that shows the nested grouping of objects and similarity levels at which it changes. The final clustering of the data can be obtained by cutting the dendrogram at the desired level of similarity (Maimon and Rokach, 2005, p. 331).

3 Overview of hierarchical clustering algorithms

3.1 Divisive clustering in minimum spanning trees

There are many different techniques that relates to divisive clustering. Although, divisive techniques tend to be less popular than the agglomerative ones it is worth to consider them. Besides the basic top-down approach there is another example of a simple divisive clustering algorithm that is based on Minimum Spanning Trees (MST). Generally, MST is a graph where all points are connected by the edges, and total length of all edges is as small as possible. The example of MST is shown in the figure 2.

The MST based hierarchical clustering algorithm is shown below (Kogan et al., 2006, p. 43):

1. Build MST for all points in the set
2. Until only single point clusters remain repeat:
   (a) Break the link with the largest distance in order to divide one cluster into two smaller ones

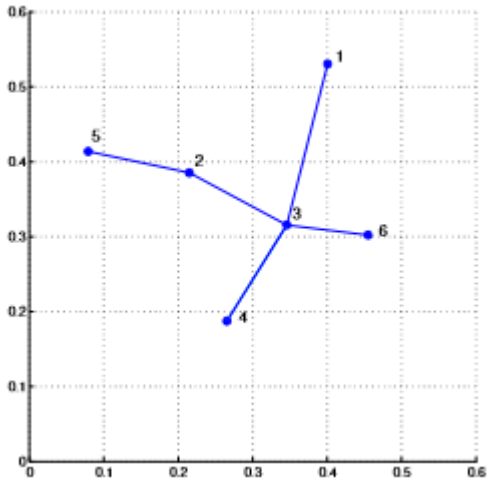

Figure 2: Minimum Spanning Tree (Kogan et al., 2006, p. 43)

This method is pretty simple for implementation. However, it still suffers from the high computational complexity and is inefficient when dealing with big data. All in all, it can be treated as a divisive implementation of a single linkage technique, which will be discussed in the next sub sections.

3.2 Agglomerative Hierarchical Clustering

Agglomerative hierarchical clustering is a popular technique, which in contrast with divisive algorithms is a bottom up approach. One of the versions of the basic algorithm is shown below (Jain et al., 1999, p. 277):

1. Consider each point as a single cluster.
2. Compute the proximity matrix for all clusters (the matrix that contains all the distances between each pair of clusters).
3. Until all points are in the same cluster repeat:
   (a) Detect the closest clusters in the proximity matrix and merge them into new cluster.
   (b) Update the proximity matrix.

The computation of proximities plays the central role in this algorithm. There are a lot of possibilities of doing it. Among them are such techniques as single linkage, complete linkage, average linkage and other. In more details these algorithms are discussed in the next subsection.

3.3 Computation of proximities

The method used for computation of proximities can strongly influence the final results returned by the hierarchical clustering algorithms. In basic hierarchical clustering approaches, different linkage methods can be used, which actually means using a different formula for calculating proximity matrix on the initial stage and for updating it after each merge or splitting is made. Some of the commonly used methods are (Kogan et al., 2006, pp 44-46):

1. MIN or Single Linkage
2. MAX or Complete Linkage
3. Average Linkage
4. Mean Distance or Centroid Linkage
5. Wards method

In single linkage the proximity between the clusters is computed as the distance between two closest points, which belong to these clusters. This method is also known as Nearest-neighbor clustering (Lance and Williams, 1967, p. 374). This technique makes decision based on only two values in the cluster, ignoring all the other points. As a result, it tends to provide long thin clusters, where distances between the nearest elements are small but the intervals between the opposite sides can be higher than the distances to other clusters. It is also known as a chaining effect (Maimon and Rokach, 2005, p. 331). So, single linkage fits good the scenarios where the clusters have non-elliptic shapes but is sensitive to outliers and noise.
Complete linkage, which is also known as Furthers-neighbor approach (Lance and Williams, 1967, p. 374), is another technique where the computation of proximity between two clusters is based on the distance between two points. It is the antithesis of Nearest-neighbor. In contrast with it the distance between the furthest points is considered as the distance between the clusters. So, the clusters are combined based on the least similar points. It helps to avoid the chaining effect that can be observed in the previous method. This technique is less susceptible to noise and outliers than single linkage (Kogan et al., 2006, p. 45) but it can break large clusters and tends to provide globular shapes.
In the average linkage approach the distance between two clusters is computed as the average of all pair-wise proximities of all points from these clusters (Punj and Stewart, 1983, p. 139). So, it is some kind of intermediate approach between single and complete linkage. In this technique, all the points of both clusters are considered, which makes it not sensitive to outliers and noise. From the other side, it is still prone to creation of globular shapes. The averages are widely used for computing the proximity matrix. There are several different variants of defining it. For example, in WPGMA (weighted pair group method with averaging) the distance between clusters is calculated as a simple average, and the number of elements in the cluster has no influence on it. Thus, the size

of the clusters weighs the result distances. The opposite approach is UPGMA (unweighted pair group method with averaging). According to it the averages are weighted by the number of items in each cluster at each step, so the final results become unweighted.
Centroid linkage is sometimes considered as a type of average linkage (Punj and Stewart, 1983, p. 139). According to it the proximity between clusters is calculated as the distance between their centroids. The definition centroid refers to the middle of the cluster, which is actually a vector of means of all the attributes. This method takes all points into consideration. So, outliers and noise do not influence it. The drawback of this method is the possibility of inversions. This refers to the case when the clusters merged on the first step are less similar than the clusters merged on the second step (Kogan et al., 2006, p. 46). As well as the previous method, it is prone to provide global clusters.
Wards (1963, p. 239) proposes the method where the decision of choosing the pair of clusters for merge is based on the minimization of the change of objective functions value, which usually refers to error sum of squares. In other words, in Wards method the proximity between clusters is defined as the increase in the variance between all points and cluster mean that results when two clusters are merged (Kogan et al., 2006, p. 46). This approach is sometimes considered as a special case of average linkage, because the objective function is based on the deviations from the mean and as a result it tries to minimize the average distance within the cluster (Punj and Stewart, 1983, p. 139). Wards method behaves the same way as the average one. It is not sensitive to outliers and noise, but is biased towards globular shapes.
Even more linkage methods can be found in the literature. Some of them are subtypes of the techniques that are discussed above. The other methods are less popular and are not considered in this paper. Overall, the right choice of the proximity calculation technique has a strong influence on the final results of the hierarchical clustering. Despite this, there is no simple rule or criterion, which method is better to use. Thus, the final choice of the linkage method should be considered based on the concrete use case.

4 Conclusion

This paper discusses hierarchical clustering approach, which is the powerful and popular technique. Its roots are going back to 1960s, but it is still often used for solving data mining problems. This paper takes a look into the main idea of hierarchical clustering and into its advantages and disadvantages. Also, it provides the overview of basic algorithms and main methods of computing the proximities between the clusters. All in all, hierarchical clustering can be used as a first step in clustering analysis. It does not need the predefinition of groups count and provides the dendrogram at the end, which makes the overview of the grouping possibilities of the data set. Even though, nowadays the computers tend to have increasingly computational power, the high computational complexity of the hierarchical clustering algorithms makes it difficult to use them

on the big datasets. To deal with it some studies provide the optimized versions and parallelization techniques, which help to make hierarchical clustering more computationally efficient.

References

Feng, L., Qiu, M.-H., Wang, Y.-X., Xiang, Q.-L., Yang, Y.-F., and Liu, K. (2010). A fast divisive clustering algorithm using an improved discrete particle swarm optimizer. *Pattern Recognition Letters*, 31(11):1216–1225.

Jain, A. K., Murty, M. N., and Flynn, P. J. (1999). Data clustering: a review. *ACM computing surveys (CSUR)*, 31(3):264–323.

Kogan, J., Nicholas, C., Teboulle, M., et al. (2006). *Grouping multidimensional data*. Springer.

Lance, G. N. and Williams, W. T. (1967). A general theory of classificatory sorting strategies ii. clustering systems. *The computer journal*, 10(3):271–277.

Li, D., Wang, S., Gan, W., and Li, D. (2012). Data field for hierarchical clustering. *Developments in Data Extraction, Management, and Analysis*, page 303.

Maimon, O. and Rokach, L. (2005). *Data mining and knowledge discovery handbook*, volume 2. Springer.

Punj, G. and Stewart, D. W. (1983). Cluster analysis in marketing research: Review and suggestions for application. *Journal of marketing research*, pages 134–148.

Rafsanjani, M. K., Varzaneh, Z. A., and Chukanlo, N. E. (2012). A survey of hierarchical clustering algorithms. *The Journal of Mathematics and Computer Science*, 5(3):229–240.

Ward Jr, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301):236–244.

5 Student Self-Evaluation Form

Is it evident from your paper that you understand what you have written?

To my mind, the paper generally shows my understanding on hierarchical clustering. It describes in detail the model, two simple algorithms and most popular types of linkage, as well as provides figures of hierarchical clustering and minimum spanning trees.

Criteria 2: Depth, scope & bibliography
Is it evident from your paper that you have researched your topic in depth, from a number of sources, and included content not covered in lecture notes?

During the work on this paper I have investigated a lot of papers related to this topic. I cannot argue that the full list of possible sources is investigated, but from the other side all main topics in the paper has the references to scientific studies.

Criteria 3: Structure & flow of the paper
Does the paper flow well, and is it properly structured.

I find my paper well-structured and all main topics are covered.

Criteria 4: Authors comments & analysis
Have you included your own analysis on what you have read?

I do not think that the points, that I have included, can be classified as an analysis from the point of view given in the example. But I do not feel that I should provide my own analysis in such a paper. From the other side, my points are highly based on the literature that is read by me, so the actual analysis of the literature is really made and more or less exists in the paper.