Types of Clustering Techniques

Kubra Iqbal

Abstract:

Clustering techniques have been discussed in this paper. They are following classifies in i) partition methods ii) Hierarchical methods iii) Density Based methods iv) Graph based methods.

The methods are discussed in detail that talk about why these methods are conducted and with what kind of data.

Keywords:

clustering, partition, hierarchical, density, graph, advantages, disadvantages

Introduction:

The main idea of this paper is to talk about Cluster Analysis and different ways it can be conducted. In today's world, Cluster Analysis has been used to do research with so many topics including, documents for browsing, to find genes and proteins that have similar functions and to provide a grouping of spatial locations after natural disasters. However, on the other hand, cluster analysis is sometimes only useful during the starting point, for example data compression or efficiency finding the nearest neighbors of points.

Cluster Analysis has been used in wide variety of fields: social sciences, biology, statistics, information retrieval, machine learning, data mining, pattern recognition. The main scope of this paper is to talk about Cluster Analysis and different kinds of cluster analysis in detail. The paper discusses some history of cluster analysis, why is it done, different ways it can be done and different clustering techniques that have been developed that are used for different kind of data.

The different kinds of techniques that will be covered are:

- Center-based partition clustering
- Hierarchical clustering
- Density Based Clustering
- Graph-Based Clustering

Partitioning methods are the most popular methods of clustering algorithms. They minimize a given clustering criterion by iteratively relocating data points between clusters until a optimal partition is attained.

Hierarchical clustering divides the given data set into smaller subsets. The hierarchical method groups the data instances into a tree of clusters. There are two major methods that are available under this category - agglomerative method and the divisive method.

Density Based Clustering works by detecting the areas where points are concentrated and where they are separated by areas that are empty or sparse. Parts that are not selected or part of the clusters are labeled as Nosie.

Graph-Based Clustering can provide more detailed information about the inner structure of the data set in terms of cliques, clusters, outliers etc. This kind of clustering method can be used for different kind of applications such as social network analysis, diffusion of information etc.

Thus, to summarize different kinds of clustering techniques and the idea of cluster analysis, it is a useful tool in many areas and can be used to figure out many data sets in every kind of industry.

Literature Review:

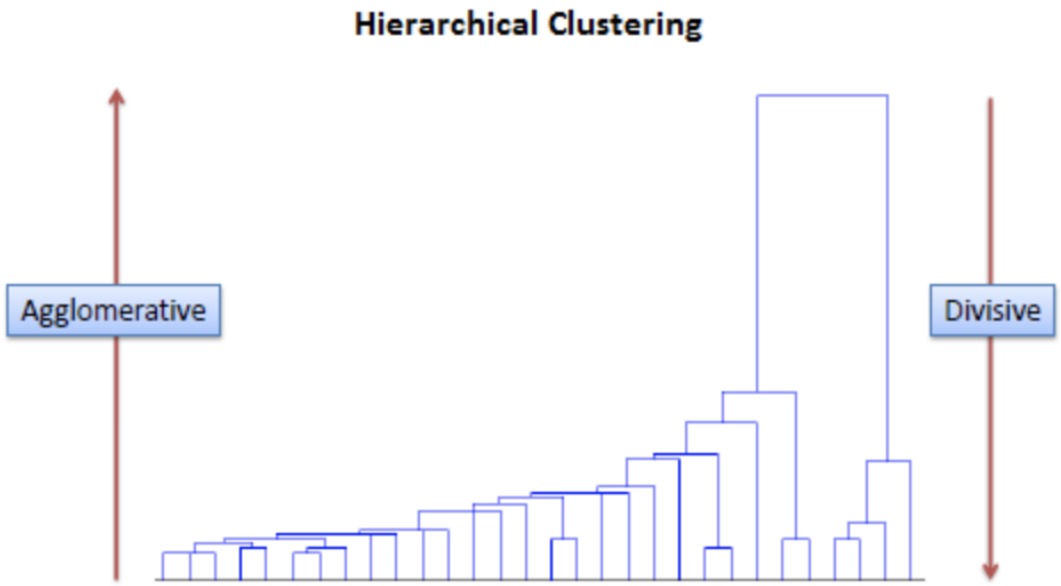
Cluster Analysis is a generic name for a variety of mathematical methods to find out which cases in a set are similar to each other. The main reason is that the data that represents physical characteristics of elements are sorted accordingly. The cluster analysis is an interdependence technique and simply defines the groups in a data set without specifying dependent or independent variables. The advantages of cluster analysis are that the technique makes it possible to objectively analyze thousands of cases in the brief time and it makes it easy for the researcher to understand the data as well. In general, there are six steps in the cluster analysis process. Though the steps are constant most of the time, there are a variety of methods to accomplish each step. To summarize the first step is to select the variables on which the respondents will be clustered together. The second step is to examine the data for outliers and multicollinearity. Third, the researcher determines the most appropriate clustering algorithm. Step 4 and 5 require the validation of the results from step 4. And the final step involves describing the clusters (Jurowski,2000).

Cluster analysis is also described as an interdependence statistical technique, it is one of the most effective methods of identifying customer segments and exposing differences between them. This technique has also been used for both product development and marketing decisions. Though it is primarily used by marketing department, cluster analysis can help managers improve tactics and strategies in other functional areas as well(Jurowski,2000).

Since Clustering is used for so many other industries it is also helpful to be used in the economic development. In the recent years due to changing nature of competition among the forms, cluster analysis has been very useful. Using this technique helps in providing a richer and more meaningful representation of local industry drivers and regional dynamics than do traditional methods (McNee).
In general, cluster analysis is a valuable but underused tool for identifying and understanding customer segments for hotels, restaurants, casinos, resorts and other hospitality firms.

There are different kinds of Clustering methods as mentioned in the introduction and this literature review will focus on five of them and explain why and how these methods are used and how are they useful. Further on it will explain how the beneficial and what kind of datasets are or real-life scenarios can be solved by these methods.

Hierarchical Clustering - which is also known as the hierarchical cluster analysis, is an algorithm that groups all the similar objects into groups that are called clusters. The end point is a set of clusters where each cluster is distinct from each other cluster and the objects within each cluster are broadly similar to each other. Hierarchical clustering can be performed with two types of data: Distance matrix or Raw data (Bock,2018).
Hierarchical clustering further branches into types, Divisive and Agglomerative.



The **Divisive Method** also called the top-down clustering method is where all the observations are assigned to a single cluster and then partition the cluster to two least similar clusters. Following this method there should be only one cluster left for each observation. There is evidence that divisive algorithms produce more accurate hierarchies than agglomerative algorithms in circumstances, but it is conceptually more complex.

The **Agglomerative Method** also called bottom-up clustering method is until the data instances belong to the same cluster. From there they compute the similarity(distance) between each of the clusters and join the two most similar clusters. This step carries two or three times until a single cluster is left. The related algorithm is shown below (Sayed).

Given:
A set X of objects $\{x_1,...,x_n\}$
A distance function $dist(c_1,c_2)$
for $i = 1$ to n
    $c_i = \{x_i\}$
end for
$C = \{c_1,...,c_n\}$
$l = n+1$
while C.size > 1 **do**
- (c_{min1},c_{min2}) = minimum $dist(c_i,c_j)$ for all c_i,c_j in C
- remove c_{min1} and c_{min2} from C
- add $\{c_{min1},c_{min2}\}$ to C
- $l = l + 1$

end while

Even though it is a well-known method it has some key issues attached to it. Firstly, there is a **lack of Global objective function**. Since Hierarchical clustering cannot be viewed as globally optimizing an objective function. Hierarchical clustering techniques use various criteria to decide at each step which clusters should be joined or split. For agglomerative hierarchical techniques, the criteria are typically to merge the "closest" pair of clusters, where close is defined by a specified measure of cluster proximity. This yield clustering algorithms that avoid the difficulty of trying to solve a hard-combinatorial optimization problem. Furthermore, such approaches do not have problems with difficulties in choosing initial points. The **Impact of the Cluster Size** is another aspect of hierarchical agglomerative clustering that should have considered us how to treat the relative sizes of the pairs of clusters that may be merged. It comes down to two schemes. Weighted and unweighted. Unweighted schemes seem to be more popular when it comes to comparing. Lastly, **Merging decisions are final hierarchical clustering tends** to make decisions that are good about combining two clusters since they have the proximity matrix available. However, once

a decision is made to merge two clusters, the hierarchical scheme does not allow for that decision to be changed again (Fung, Benjamin & Wang, ke & Ester, Martin, 2006).

To summarize the problems with hierarchical clustering:

- No global objective function is being optimized
- Merging decisions are final
- Good local merging decisions may not result in good global results.

Even though it has some disadvantages, but this method also comes with other advantages that make it beneficial to use. It is easy to implement. With a large number of variables, it is also computed easily and lastly it is easier to decide the number of clusters by looking at the dendrogram (Ishani,2004).

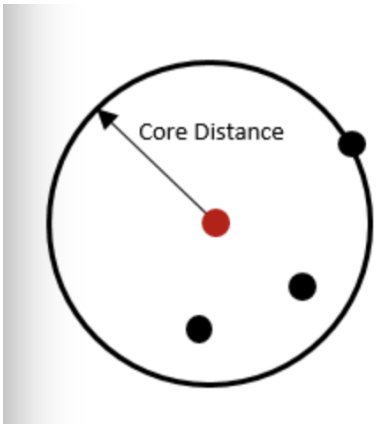
Density Based Clustering tool works by detecting areas where points are concentrated and where they are separated by the areas that are empty. Points that do not end up becoming the part of the cluster are labeled as noise. This tool uses unsupervised machine learning algorithms what then automatically detect patterns based on the location and the distance specified to number of neighbors. There are many ways that this method can be used to solve real problems. An example can be: Geo-locating tweets following natural hazards or terror attacks can be clustered and rescue and evacuation needs to be informed based on the size and the location of the clusters that are being identified. Another example can be studying a pest-borne disease and have a point in the dataset which is representing households in your particular study area and some of them are infested and some are not. By using the Density-based Clustering tool, you can find out what are the largest clusters of infested households to help pinpoint an area to begin treatment and extermination of the pests ArcGis).

The Density Based clustering tool provides three different clustering methods with which to find clusters in your point area:

- Defined Distance
- Self-Adjusting
- Multi Scale

The Defined Distance uses a specified distance to separate dense clusters from the noise found in the data. It is the fastest of the clustering methods, but it is only appropriate to use if there is a very clear search distance. The data should also have meaningful clusters. If the data is not very clear and there is a lot of noise, this method might not be very appropriate to use.
Self-adjusting method uses a range of distances to separate clusters of varying densities from noise. This kind of method requires the least user input. And lastly, Multi Scale use the difference between the neighboring features and reach the plots which are then used to separate clusters of varying densities from noise ArcGis).

The Minimum Feature Per Cluster parameter is also important in the calculation of the core-distance and it is a measurement used by all three methods to find clusters. The core-distance or each point given is the measurement of the distance that is required that travels from each point to the defined minimum number of features. If a large **Minimum Features per Cluster** is chosen, the corresponding core-distance will be larger. On the other hand, if a large Minimum Features per Cluster is chosen, then the corresponding distance will be smaller(ArcGis).

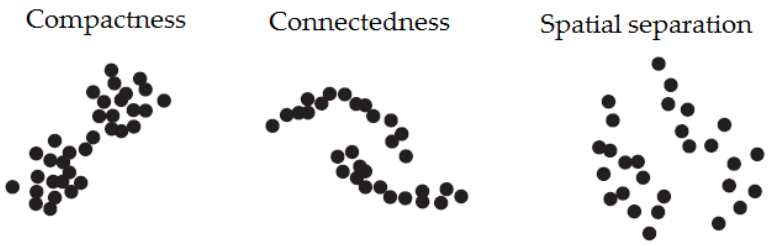


To summarize Density Based Clustering – there are a lot of advantages it comes with. It is commonly used and easily implemented. It can also be computed quickly so it is considered a faster method. The results that are produced are tight clusters and if finds more sub-clusters in the data set that is being used is large.

Graph Based Clustering is another method that provides detailed information about the inner structure of the data set in terms of cliques, clusters and outliers. This kind of clustering is very commonly used in applications such as social network analysis, diffusion of information etc. When Graph Based Clustering is applied the analysis is easy to interpret since it shows up in graphs. The main idea is that the objects are represented as nodes in complete or connected graph. There is X and Y and the weight is defined by the distance between the nodes when this is conducted (Novak, 2010).

In clustering algorithm, generally the data set is partitioned according to the similarities of the data points and an objective function. When the clusters are different sizes, shapes and densities, the similarities are not enough to produce a satisfactory clustering result. There are some problems that are encountered when Graph Based Clustering is used. Traditional clustering algorithms can be roughly categorized into hierarchical, partitioning, density-based and model-based. The cluster problems are classified into two groups: separated problems and torching problems.

The can be further divided into distance separated problems and density separated problems. The figure below shows that the two classes have similar shape, size and density, Since the distance of a point pair in the same cluster is less that of a point pair in different clusters, the data set is said to be distance separated (Zohng,2010).



Center Based Partition clustering – It is one of the most popular methods of clustering algorithms. They minimize and give clustering criteria by iteratively relocating data points between clusters until an optimal partition is attained. Partition methods are father divided into two sub categories; one is centroid and the other one is called medoids algorithms. Centroid algorithms represent each cluster by using the gravity center of the instances. The medoid algorithms represents each cluster by means of the instances closest to the gravity center. The centroid algorithm is the k-means. The K-means method works on partitioning the data set into K subsets such that all points in a given subset are closest to the same center. Further on, K-means then computes the new centers by taking the mean of all data points belonging to the same cluster. The process is performed until there is no change in gravity centers. If K cannot be known ahead of time, a lot of values of K can be evaluated until the most suitable one is found. The effectiveness of this method mainly relies heavily on the objective function used in measuring the distance between instances. This method has the following important characteristic; it is very efficient in processing large data sets. It terminates at a local optimum even though it is sensitive to noise.
Compared to other types of clustering algorithms, this method is very efficient for clustering large databases and high-dimensional databases. Usually, they have their own objective functions which define how good a clustering solution is. The goal of a center-based algorithm is to minimize its objective function (Can, 2007)

There are some limitations to this method – K-means attempts to minimize the squared or absolute error of points with respect to their clusters. While this is a method that sometimes that works out it leads to a simple algorithm. K-means has a number of limitations and problems. The figures below show the result when clusters have widely different sizes or have convex shapes. One of the biggest difficulties in these two stations is that K-means objective function is a mismatch for the kind of clusters we are trying to find through this method.

Conclusion:

Cluster analysis is still an active field of development. In areas of statistics, computer science, pattern recognition and vector quantization, there is still a lot of work that is being done. Many cluster analysis techniques do not have a have a strong form basis. While some of the techniques make sure of formal mathematical methods, they often do not work better than more informal methods. Since there are a wide variety of clustering techniques, some would argue that the wide range of subject matter, size and type of data and differing user goals make this inevitable and the cluster analysis is a collection of different problems that require a variety of techniques for their solutions.

It has hard to compare and see which of the methods is better. All of the methods are better for different kinds of situations and it is hard to judge how well the technique will do for a particular kind of dataset. Even though there are some problems with each kind of technique, many people still use cluster analysis for a wide variety of useful tasks.

.

References:

Jurowski, C., & Reich, A. Z. (2000). An Explanation and Illustration of Cluster Analysis for Identifying Hospitality Market Segments. *Journal of Hospitality & Tourism Research, 24*(1), 67-91. doi:10.1177/109634800002400105

McNee, M. (n.d.). Untitled Document. Retrieved from http://www.umich.edu/~econdev/Cluster/

Bock, T. (2018, April 23). What is Hierarchical Clustering? Retrieved from https://www.displayr.com/what-is-hierarchical-clustering/

Sayed, S. (n.d.). Retrieved from http://www.saedsayad.com/clustering_hierarchical.htm

Fung, Benjamin & Wang, ke & Ester, Martin. (2006). Hierarchical Document Clustering. 1. 10.4018/9781591405573.ch105.

Ishani Follow, M. (2014, August 23). Hierarchical clustering. Retrieved from https://www.slideshare.net/ishmecse13/hierarchical-clustering-38276870

ArcGIS Pro. (n.d.). Retrieved from http://pro.arcgis.com/en/pro-app/tool-reference/spatial-statistics/how-density-based-clustering-works.htm

Novák, P., Neumann, P., & Macas, J. (2010, July 15). Graph-based clustering and characterization of repetitive sequences in next-generation sequencing data. Retrieved from https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-11-378

Zohng, C. (2010, March). A graph-theoretical clustering method based on two rounds of minimum spanning trees. Retrieved from https://dl.acm.org/citation.cfm?id=1660180.1660647

Can, G. (2007). Data Clustering: Theory, Algorithms, and Applications. Retrieved from https://epubs.siam.org/doi/abs/10.1137/1.9780898718348.ch9