ISSN Print: 2394-7500
ISSN Online: 2394-5869
Impact Factor: 8.4
IJAR 2021; 7(4): 178-181
www.allresearchjournal.com
Received: 18-02-2021
Accepted: 24-03-2021

Pranav Shetty
Department of Computer Science, All India Shri Shivaji Memorial Society's College of Engineering Savitribai Phule Pune University, Pune, Maharashtra, India

Suraj Singh
Department of Computer Science, All India Shri Shivaji Memorial Society's College of Engineering Savitribai Phule Pune University, Pune, Maharashtra, India

Corresponding Author:
Pranav Shetty
Department of Computer Science, All India Shri Shivaji Memorial Society's College of Engineering Savitribai Phule Pune University, Pune, Maharashtra, India

Hierarchical Clustering: A Survey

Pranav Shetty and Suraj Singh

DOI: https://doi.org/10.22271/allresearch.2021.v7.i4c.8484

Abstract
There is a need to scrutinise and retrieve information from data in today's world. Clustering is an analytical technique which involves dividing data into groups of similar objects. Every group is called a cluster, and it is formed from objects that have affinities within the cluster but are significantly different to objects in other groups. The aim of this paper is to look at and compare two different types of hierarchical clustering algorithms. Partition and hierarchical clustering are the two main types of clustering techniques. Hierarchical clustering algorithm is one of the algorithms discussed here. The aforementioned algorithms are described and analysed in terms of factors such as dataset size, data set type, number of clusters formed, consistency, accuracy, and efficiency. Hierarchical clustering is a cluster analysis technique that aims to create a hierarchy of clusters. A hierarchical clustering method is a set of simple (flat) clustering methods arranged in a tree structure. These methods create clusters by recursively partitioning the entities in a top-down or bottom-up manner. We examine and compare hierarchical clustering algorithms in this paper. The intent of discussing the various implementations of hierarchical clustering algorithms is to assist new researchers and beginners to understand how they function, so they can come up with new approaches and innovations for improvement.

Keywords: hierarchical clustering, clustering, divisive hierarchical clustering, agglomerative hierarchical clustering, partitioning clustering

Introduction
Clustering is a core concept that has attracted a lot of attention from pattern recognition, statistics researchers and machine learning. Clustering is an example of unsupervised learning, in which no training samples are available from which to learn and create model. Clustering creates clusters of samples that are all linked under certain ways. As a result, the similarities between samples belonging to the same cluster are greater than those belonging to different clusters. It's also known as unsupervised classification because it achieves the same results as classification algorithms without the need for predefined groups. The aim of clustering algorithms, in its most primitive form, is to take a dataset and find the distinct clusters that prevail within it. Clustering is a popular algorithm in a variety of fields, including psychology, business and retail, computational biology, social media network analysis, and so on.

Clustering approaches include hierarchical, partitioning, grid, and density-based clustering, each of which employs a different induction theory. In a nutshell, the hierarchical approach generates a series of clustering, each of which is nested into the next clustering in the series.

The dataset is partitioned into k partitions, with each partition representing a cluster. Based on the characteristics and similarities of the data, this clustering approach divides the information into multiple classes. The number of clusters that must be created for the clustering methods is defined by the data analysts. In the partitioning method when database (D) that contains multiple (N) objects then the partitioning method constructs user- specified (K) partitions of the data in which each partition represents a cluster and a particular region.

In this paper, we are comparing the new approaches discussed in [2] with the traditional approach

Clustering
Cluster analysis is the process of grouping a series of patterns (usually represented as a vector of measurements or a point in a multidimensional space) based on their similarity [3].

Patterns within a cluster are more closely related to each other than data from neighbouring clusters. It is essential to understand the distinction between unsupervised and supervised classification, as well as clustering and discriminate research. We are given a collection of pre-classified objects in the supervised approach; the task is to mark a newly encountered, but unlabelled object. The descriptions of the groups are given for the objects that have already been labelled, which will aid us in labelling a new object. We will be provided a set of unlabelled objects to categorise into valid clusters in an unsupervised approach. Clustering is the process of grouping data into clusters with high intra-cluster and low inter-cluster similarity. A strong clustering algorithm should be capable of detecting clusters of any type. Clustering is often used for a variety of purposes, including determining the internal structure of data (e.g., gene clustering) and partitioning data (e.g., market segmentation). Driver and Kroeber developed cluster analysis in anthropology in 1932, and Joseph Zubin and Robert Tryon applied it to psychology in 1938 and 1939, respectively. Cattell famously used it for trait theory classification in personality psychology starting in 1943.

Hierarchical Clustering

Organizing optimization algorithms by determining the number of clusters at the start of the process before clustering. Hierarchical clustering algorithms, on the other hand, combine or divide existing groups and specify the order in which clusters are divided or combined. A tree or dendrogram is used to display hierarchical clusters. Hierarchical clustering can be accomplished in two ways. They can be bottom-up or top-down. Large clusters are divided into small clusters, and small clusters of large clusters are combined together. Hierarchical method can be subdivided as following:

A. Divisive hierarchical clustering

Divisive clustering (Butler, 2003) is a "reverse" approach to agglomerative clustering that starts with a single cluster or model with all data points and splits it recursively. The procedure is repeated until a stopping criterion (a predetermined number K of clusters or models) is met. The "poorest-fit" cluster gives the lowest probability to the items in this cluster will be split after each iteration of division. This process is repeated until the clusters become singletons or a stop criterion is met. This, like agglomerative clustering, has high computational costs and model selection issues. Moreover, it is quite sensitive to initialization, due to the possible divisions of data into two clusters at the first step.

The steps to form divisive (top-down) clustering are:

Step 1: Start with all data points in the cluster.

Step 2: After each iteration, remove the "outsiders" from the least cohesive cluster.

Step 3: Stop when each example is in its own singleton cluster, else go to step 2.

B. Agglomerative hierarchical clustering

A bottom-up method in which each entity represents its own cluster, which is then iteratively merged until the desired cluster structure is achieved. This N-sample algorithm starts with N clusters, each containing a single sample. Following that, two clusters with the greatest similarity will combine until the number of clusters is reduced to one or the user specifies. The minimum, maximum, average, and centre distances are the parameters used in this algorithm.

The steps for forming agglomerative (bottom-up) clustering are:

Step 1: Start by considering each data point as its own singleton cluster.

Step 2: After each iteration of calculating Euclidian distance, merge two clusters with minimum distance.

Step 3: Stop when there is a single cluster of all examples, else go to step 2

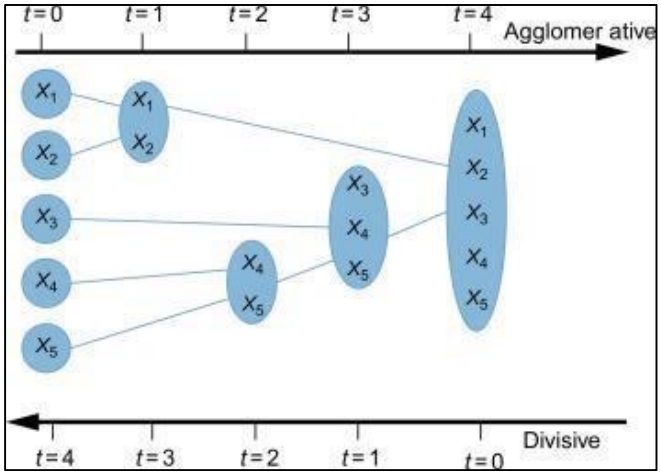

Fig 1: Hierarchical Clustering

4. Partitioning

Partitioning clustering is the most basic form of clustering, in which a given dataset is divided into k (an arbitrary number) partitions, each of which represents a cluster. Partitioning algorithms divide data points into one-level (un-nested) partitions. If k is the desired number of clusters, partitioning algorithms find all k clusters at the same time, as opposed to conventional hierarchical approaches, which divide a cluster into two sub-clusters or combine two sub-clusters into one cluster. This clustering approach employs a number of greedy heuristics schemes in the form of iterative optimization, which entails various relocation schemes that reassign points between the k clusters iteratively. Clustering results are steadily improved by relocating algorithms. Clusters must have two properties in this method: (a) each cluster must contain at least one object, and (b) each object must belong to exactly one cluster. Many partitioning clustering methods exist, including K-means, Bisecting k-means, PAM (Partitioning Around Medoids), CLARA, and Probabilistic Clustering.

5. Euclidian Distance

The length of a line segment connecting two points in Euclidean space is the Euclidean distance between them. It is often referred to as the Pythagorean distance because it can be determined from the Cartesian coordinates of the points using the Pythagorean theorem.

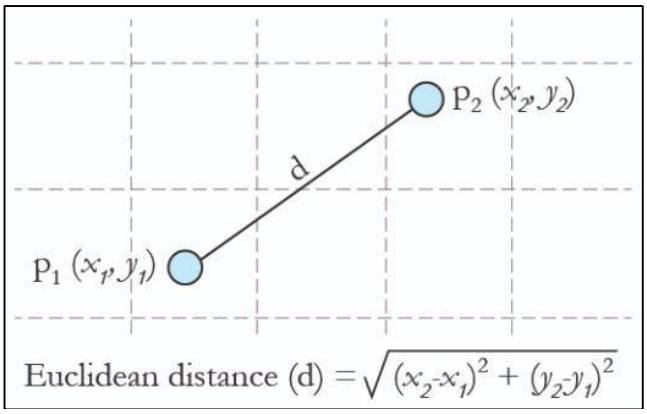

Fig 2: Euclidean Distance

6. Arithmetic Mean

The arithmetic mean, also known as the mean or average is the sum of a set of numbers divided by the number of values in the set.

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{X_1 + X_2 + \lambda + X_n}{n}$$

Fig 3: Formula for Arithmetic Mean

7. Analysis of Related Work

In this section, we will discuss, what most of the related works mentioned above provide to us and what were the drawbacks in the schemes proposed by these papers

In paper [1] the distance between each point in the data set and every other point is determined, and the two points with the shortest distance are combined to form a single cluster. These two are now combined as a single point or vector, and the distance calculation process is repeated. This process will be continued till all points are combined to form a single cluster.

In paper [2] the new hierarchical clustering algorithm is a bottom-up agglomerative hierarchical clustering approach. Consider set of points X = {a1, a2….an} in Zm is given and we want to cluster them. The first step is to find the data point's nearest neighbour to form pairs, then search for those pairs that share a point to form primary clusters. The next step is to calculate the mean value for each primary cluster and then measure the distance between the mean and all cluster data points to determine which cluster has the greatest distance (D). The distance between data points in different clusters is measured as the final step. If the distance between two data points from different clusters is equal to (D of cluster 1) or (D of cluster 2), then these two clusters should be combined.

8. Discussion

The method used in paper [1] for creating clusters using hierarchical clustering is calculate the distance between each pair of patterns in the distance matrix. Assume that each pattern belongs to a cluster. Using the data matrix, find the most related pair of clusters. Make a single cluster out of these identical pairs of patterns. Stop if all of the points are in a single cluster; otherwise, go to the previous step.

Similarly, the methodology proposed in paper [2] has a few extra steps as follows calculate the Euclidean distance between all data points to find the nearest neighbor for and data point in order to make pairs. To render primary clusters, combine certain pairs that have a point in common. For each primary cluster, calculate the mean (μ). Find the distance between the mean (μ) and all of the data points in a cluster. In each cluster, find the maximum value of (d) and mark it D.

Determine the distance between data points from different clusters; if the distance between two points (point 1 from cluster 1 and point 2 from cluster 2) is equal to (D of cluster 1) or (D of cluster 2), then combine the two clusters.

Since the new approach ensures that all nearby points are grouped together. We may conclude that the new approach is more accurate than others based on the results of k-means, agglomerative clustering, and the new clustering method published in paper [2]. However, the computational time is longer, especially for larger datasets.

9. Advantages

i. As we are using mean to calculate the distance again the accuracy is higher.
ii. Can easily handle all types of distances. III It is robust for noisy data.
iii. It can accept definite number of clusters as input.
iv. It can also handle high dimensionality.
v. It converges fast if we give the desired data.

10. Disadvantages

i. The algorithm can never undo any previous steps. So, for example, the algorithm clusters 2 points, and later on we see that the connection was not a good one, the program cannot undo that step.
ii. The time complexity for the clustering can result in very long computation times, in comparison with efficient algorithms, such k-Means.
iii. If we have a large dataset, it can become difficult to determine the correct number of clusters by the dendrogram.

11. Conclusion

After comparing the outcomes based on all of the above factors as aforementioned in the discussion, we conclude that using mean to calculate the distance yields better results.

Future work can focus on reducing the algorithm's computational time to make it more suitable for high-dimensional datasets. In addition, further clustering problems will be tested, and its efficiency will be compared to that of other clustering methods.

12. References

1. Ahalya G, Pandey HM. “Data Clustering Approaches Survey and Analysis” 1st International Conference on Futuristic trend in Computational Analysis and Knowledge Management (ABLAZE-2015) 2015
2. Zahra Nazari, Dongshik Kang, Reza Asharif M, Yulwan Sung, Seiji Ogawa. A New Hierarchical Clustering Algorithm ICIIBMS Track2: Artificial Intelligence, Robotics, and Human-Computer Interaction, Okinawa, Japan 2015.
3. Kaur, Maninderjit, Sushil Kumar Garg. "Survey on Clustering Techniques in Data Mining for Software Engineering." International Journal of Advanced and Innovative Research 2014;3:238-243.

4. Singh, Nidhi, Divakar Singh. "Performance Evaluation of K-Means and Hierarchal Clustering in Terms of Accuracy and Running Time." *IJCSIT)* International Journal of Computer Science and Information Technologies 2012;3(3):4119-4121.
5. Sathya R, Annamma Abraham. "Comparison of supervised and unsupervised learning algorithms for pattern classification." Int J Adv Res Artificial Intell 2013;2(2):34-38.
6. Cichosz P. Data Mining Algorithms Explained Using R, John Wiley & Sons, Ltd 2015, 349-362
7. Mann AK, Kuar N. "Review paper on clustering techniques," Global Journal of Computer Science and Technology Software & Data Engineering 2013;13(5):43-46, version. 1.0,
8. Kaur M, Kaur U. "Comparison between k- means and hierarchical algorithm using query redirection," International Journal of Advanced Research in Computer Science and Software Engineering 2013;3(7):1454-1455.
9. Zhao Y, Karypis G. "Hierarchical clustering algorithms for document datasets," pp. 141-142, Data Mining and Knowledge Discovery, Springer Science + Business Media, Inc 2005;10:141-168.
10. Masciari E, Mazzeo GM, Zaniolo C. "A new, fast and accurate algorithm for hierarchical clustering on euclidean distances" Springer-Verlag Berlin Heidelberg 2013, 111-114, LNAI 7819.