WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Cluster Analysis: A Case Study

Dhruvi Prakash Karu[1], Rutu Pradipkumar Pansaniya[2], Mayank Devani[3]

Computer Engineering Department, SAL College of Engineering
Computer Engineering Department, SAL College of Engineering
Information Technology Department, SAL College of Engineering

ABSTRACT:

Only a limited number of traits or characteristics can be accurately segmented by hand, which can be restrictive. It is deficient in its ability to grow. Cluster analysis with a larger number of attributes can be executed rapidly using Mearns analytics and machine learning. Furthermore, it employs a data-driven methodology, which enhances its accuracy and reliability. While there are various techniques for categorizing individuals into banking segments based on multivariate survey data, clustering is still the most preferred and popular technique. Clustering is a well-liked and widely applied technique for determining or creating data-based banking divisions.

A cluster, by itself, is a group that has been assigned a specific purpose based on certain characteristics.

When data contains hidden linkages that need more investigation and analysis, cluster analysis can be used to create discrete groups.
For this reason, by placing a lot of relative points in singularity, clustering analysis is employed to find pairs of groupings. Furthermore, to simplify the grouping process, divide the least relative points into blocks with singularity.

Key words: clustering, K-means, hierarchical, Unsupervised Machine Learning, partition.

1] Introduction

Among the unsupervised machine learning fields and statistical data analysis, clustering analysis is a crucial method employed to identify patterns or even natural groupings within a dataset. The process of breaking down a collection of data objects into smaller groups is called cluster analysis. Since there is no available class label information, clustering is commonly known as unsupervised learning.
Numerous disciplines, such as marketing, biology, and social science, extensively employ this method. Web search has also made extensive use of grouping. For instance, due to the abundance of web pages, a keyword search can frequently yield a large number of results (i.e., search for relevant pages). Clustering can be employed to group the search results and present them in a compact and easily accessible format.
A group of data objects can be considered an implicit class because it consists of objects that share similarities within the group but differ from objects in other groups. Clustering is also referred to as automatic classification in this context.
In some applications, clustering is also called data segmentation since it separates large data sets into groups according to their similarity.
Outliers can also be found via cluster analysis. Recognizing fraudulent credit card transactions and monitoring illicit activities in online transactions are two instances where outlier detection is employed.

2] Clustering requirements for data mining:

i] Scalability: clustering in a sample of a given large data set may lead to biased results. It is crucial to have scalable cluster algorithms.
ii] The ability to manage several attribute types: most algorithms are specifically designed to group numerical data. In addition to numerical data, other forms of data, such as binary, categorical, ordinal, or combinations of these kinds of data, may need to be clustered for certain uses. For complex data types like photos, graphs, sequences, etc., additional clustering techniques are necessary.
iii] Domain knowledge requirements to specify input parameters: Users must supply domain information in the form of input parameters, such as the desired number of clusters, for the majority of clustering algorithms. Requiring users to possess domain knowledge not only adds to their burden but also makes it challenging to enforce quality standards in clustering.
iv] Noisy data handling ability: Clustering algorithms may generate clusters of poor quality and be sensitive to noise. Hence, we require sound-resistant clustering algorithms.
v] Constraint-based clustering: In real-world scenarios, clustering may be required to be performed under different constraints. Identifying data groups that satisfy specific criteria and exhibit high clustering quality is a particularly difficult task.
vi] Interpretability and usability: People desire interpretable, understandable, and beneficial clustering results. In essence, it may be necessary to connect

clustering with particular meanings and uses.

3] Clustering Methods:

3.1] Partitioning methods:

The most fundamental and straightforward type of cluster analysis is partitioning. A partitioning algorithm divides a collection of n objects into k clusters, with a subset of the objects in each cluster. To put it another way, it divides the data into k groups and makes sure that each group has at least one object. Each item should be allocated to a single group. Most partitioning methods rely on proximity. A partitioning method generates an initial grouping for a specific k, the total number of groups to be formed. Following that, it employs an iterative relocation algorithm to transfer objects from one group to another, with the aim of achieving the most efficient grouping. Some typical ways of dividing the data are given below.

3.1.1] K-Means [A Method Based on Centroids]:

A sort of machine learning algorithm called K-means clustering divides unlabeled data into k distinct clusters according to how similar they are.

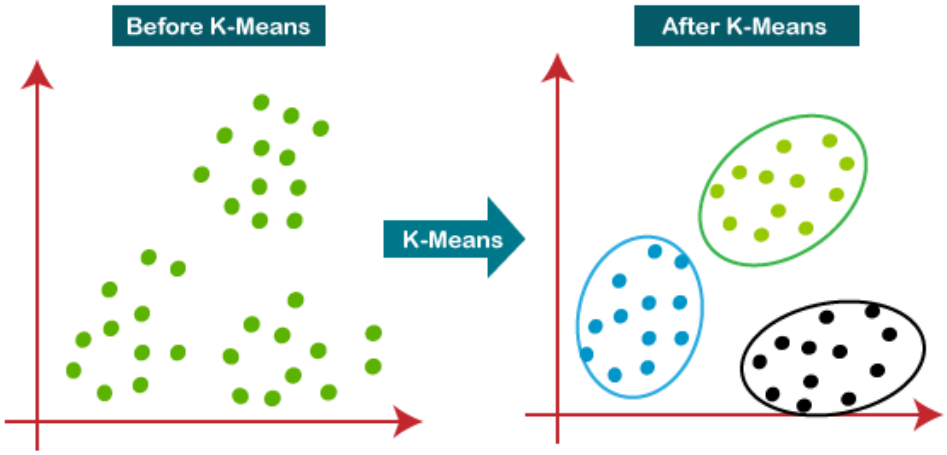
The algorithm first chooses a small number of random locations, or centroids, to serve as the beginning points for forming a cluster. The closest cluster is then assigned to each point. Once each point has been allocated to a cluster, the centroids are recalculated by calculating the average position of the points inside a cluster. This process is repeated as long as the centroids are still moving and forming clusters. Assuring that related data points are grouped in the same cluster is the aim of clustering, which groups data points together according to their commonalities.

Advantages:

- Commonly utilized and a simple method
- Guarantees convergence
- Offers a precise estimation of the initial locations of centroids

Disadvantages:

- It is required to specify the number of clusters.
- Depends on randomly chosen initial values, which can result in variability between different executions.
- Data normalization may be essential before clustering.



K-means cluster analysis is frequently employed in a variety of industries, from identifying urban traffic patterns for Uber drivers to dividing up customers according to their interests, past purchases, or purchasing patterns.

3.1.2] K-Medoids [A Typical Object-Based Method]:

A significantly altered version of the k-means algorithm, K-MEDOIDS is a clustering algorithm based on partitioning. While minimizing the squared error is the goal of both techniques, the k-medoids algorithm is more robust to noise than the k-means algorithm. The k-means method selects means as centroids, whereas the k-medoids algorithm selects points to serve as medoids (medians). A medoid is an object within a cluster whose mean differs from the overall mean of all objects in that cluster.

Benefits:

- Simple to grasp and apply
- Fast
- Reduced sensitivity to outliers

Drawbacks:

- Results may vary between runs due to the random selection of the initial k-medoids.
- Unsuitable for non-spherical clusters as it emphasizes data point proximity over their connectivity.

Facial recognition software employs the k-medoids algorithm due to its ability to handle real data and its resistance to outliers. In the business and marketing world, k-medoid cluster analysis is commonly used.

3.2] Hierarchical Methods:

Data objects are organized using the hierarchical clustering method into a hierarchical structure or "tree" of clusters. Initially, each data point is assigned to its own cluster in hierarchical clustering, which then repeatedly combines the nearest clusters until a preset end condition is satisfied. A dendogran, which resembles a tree and graphically depicts the hierarchical relationship between the groups, is the result of hierarchical clustering. There are two categories of hierarchical clustering.

3.2.1] Agglomerative Hierarchical Clustering:

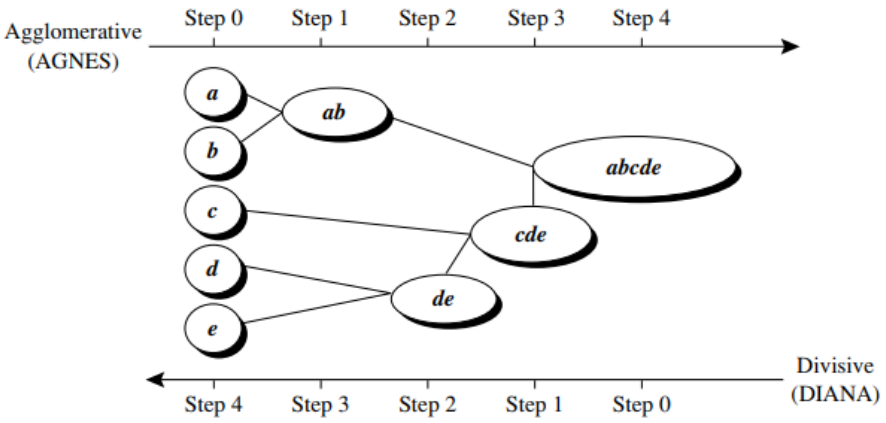
An algorithm for agglomerative hierarchical clustering employs a bottom-up approach. Each object typically forms its own cluster at the start of the process, which then progressively merges clusters into bigger ones until either all objects are merged into a single cluster or certain termination requirements are satisfied. The structure is composed of just one cluster. An agglomerative algorithm performs a maximum of n iterations, as each iteration combines two clusters, with each cluster containing at least one object.
The fundamental algorithm performs the following tasks:

- Determine the proximity matrix.
- group all points into clusters.
- merge the two closest clusters, update the proximity matrix and repeat.
- Keep repeating until there is just one cluster left.

3.2.2] Divisive Hierarchical Clustering:

Using a top- down methodology, a partitioning hierarchical clustering algorithm first groups all items into a single cluster, which serves as the basis for the hierarchy. The root cluster is further subdivided into many smaller subclusters, which are further subdivided into still smaller subclusters in a recursive fashion. The procedure keeps going until the objects in a cluster are sufficiently similar or until each of the final clusters is homogeneous enough to contain only one object.
The following is what the basic algorithm does.

- Take into account every data point as one cluster.
- Repeat and divide the cluster's disparate data points.
- Keep going until each data point is unique and considered a different cluster.



3.3] Density-Based Methods:

Hierarchical and partitioning approaches are designed to locate clusters that have a spherical shape. They struggle with recognizing clusters of non-random shapes, such as the "s" shape and oval-shaped clusters. As a result, In order to efficiently find clusters of non-spherical structures, we employ density-based clustering algorithms. These three primary methodologies comprise the basic density-based clustering algorithms.

3.3.1] Density-Based Spatial Clustering of Noised Applications, or DBSCAN:

DBSCAN is one of the most used density-based clustering techniques. The two main parameters off the DBSCAN method are the minimum number of neighbours (minpts) and the maximum distance (eps) between core data points.
Benefits of DBSCAN include

- its ability to detect clusters of any shape and its resistance to noise, in contrast to k-means.
- The number of clusters does not have to be predetermined.

Disadvantages:

- The selection of the Eps and MinPts parameters has an impact on it.
- Clusters with different densities do not function well with it.
- Finding every cluster in the data is not a given.

3.3.2] OPTICS [Clustering Structure Identification by Ordering Points]:

A clustering process called optics uses the idea of density to produce a graph that shows the relationships between data points.
Each data piece is connected to a directed graph that its closest neighbours within a certain distance limit is known as a reachability graph. The weight of the reachability graph's edges is determined by the distance between the connected data points. The algorithm divides the reachability graph into clusters, with each cluster having a density threshold, resulting in a hierarchical clustering graph.

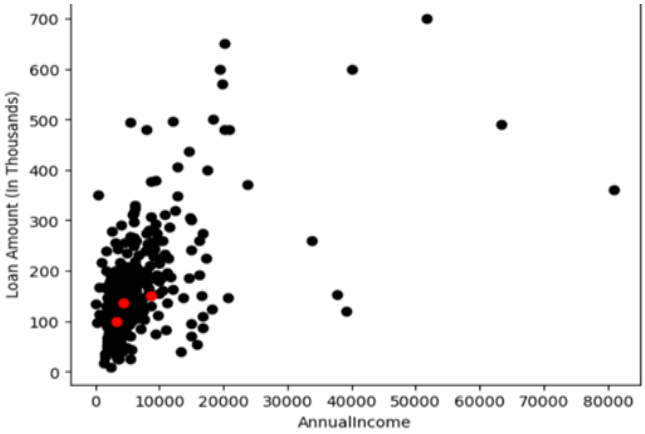
3.3.3] High-Density-Based Spatial Clustering of Applications with Noise, or HDBSCAN:

HDBSCAN creates a structure of clusters by analysing a mutual-reachability graph, a graph where each point is a node and the edges between them are assigned a value representing their distance or similarity. When two points are connected by an edge in a graph, and the distance between them is less than a specific threshold, it indicates that they are reachable from each other. The greatest of two points' reachability distances—a measure of how readily one point may be reached from the other—is the mutual reachability distance. The reachability distance is the distance between two points that is the furthest from the point that has the fewest points between them.

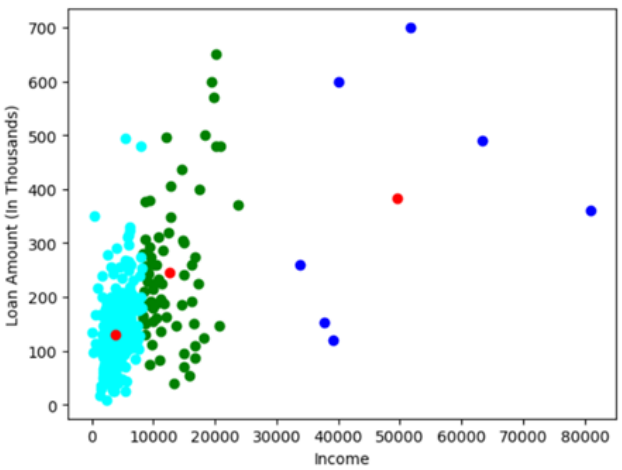
4] Case study:

The Credit / Home Loans prediction dataset, which we obtained from https://www. Altruistdelhite04/loan-prediction-problem-dataset on kaggle.com. Standard Bank plans to use innovative technologies to provide their clients with a full range of services from the comfort of their mobile devices as they embrace the wave of digital transformation. The bank, which is the largest lender in Africa in terms of the assets, wants to enhance the way prospective borrowers currently apply for home loans. The dataset is available here.
Pre-processing has already been done on this data.
just two variables from the data—"LoanAmount" and "ApplicantIncome"—and using a scatter plot to visualize it.

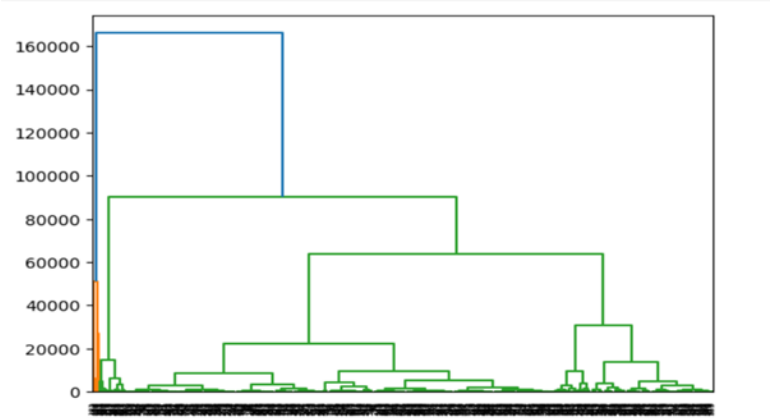
Steps 1 and 2 of the K-means algorithms were to determine the number of clusters (k) and select random centroids for each cluster. the centroids will be selected at random from the data, and three clusters will be selected.



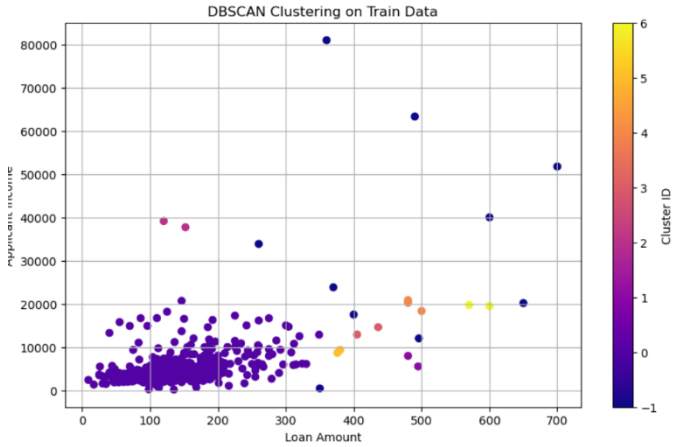
The red dots represent the three centroids of each cluster. These are randomly chosen points. Next, implement the K-Means Clustering algorithm.
Find out how the centroids from the prior iteration differed from one another and the current iteration by first defining the diff as 1 and then computing it throughout the entire loop. The training is terminated when this difference is zero. Look at the clusters now.

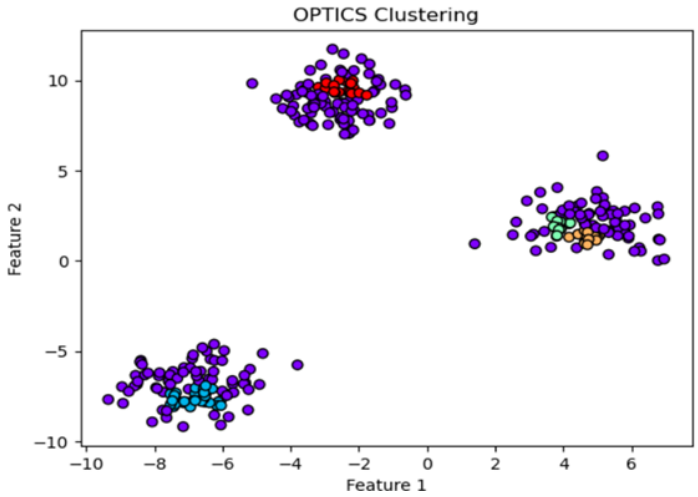


The centroid of each of the three clusters, which are clearly visible, is represented by the red dots. We are now using our dataset to illustrate hierarchical clustering.



We are now using DBSCAN clustering and optic clustering and "LoanAmount" and "ApplicantIncome".





REFERENCES:

[1] Morgan Kaufmann's Data Mining Concepts and Techniques, Third Edition.
[2] https://medium.com/@adeleke.joe/clustering-analysis-in-machine-learning-a-case-study-cf72a7b69096 is a case study on clustering analysis in machine learning.
[3] cluster analysis from the following resources: https://www.surveymonkey.com/market-research/resources/how-cluster-analysis-identifies-market-and-customer-segments/
[4] loan prediction problem dataset[Kaggle.com]