Spectrum Journal of Innovation, Reforms and Development
Volume 15, May, 2023 | ISSN (E): 2751-1731
Website: www.sjird.journalspark.org

# DATA CLUSTERING ALGORITHMS

Abdujabborov Madaminjon Vohidjon o'g'li
Teacher at Andijan State University

Anvarbekova Hilola Numonovna
Magistr student at Andijan State University

**ABSTRACT**
This article provides information about clustering and clustering algorithms. The selection of an appropriate clustering algorithm depends on various factors, such as the nature of the data, the number of clusters required, the available computing resources, and the specific problem domain. Different algorithms may be more efficient for different types of datasets or clustering purposes.

**Keywords**: Clustering, Model, Method, Analysis, K-means, Algorithm, Hierarchical Clustering, DBSCAN, Cluster Visualization, Cluster Analysis, Agglomerative Clustering.

## Introduction

Clustering is the grouping of a set of objects such that objects in the same group are more similar to each other than to those in other groups. Given a set of data points, we can use a clustering algorithm to classify each data point into a specific group. In theory, data points in the same group should exhibit similar features and/or characteristics, while data points in different groups should possess distinct features and/or characteristics. Clustering is an unsupervised learning technique and is commonly employed for statistical data analysis across various fields.

The primary purpose of clustering is to ascertain whether objects in a set can be grouped based on their similarity. This facilitates the organization and categorization of data, while also assisting in the analysis and modeling processes.

**Clustering data offers several advantages:**

1. Aids in Analysis: Clustering facilitates data analysis and enables the definition of analysis outcomes. Through clustering, similarities, trends, and relationships within data sets can be identified.
2. Data simplifies learning and sorting: Clustering algorithms streamline data sorting and classification, making it easier to analyze by organizing the data into groups or segments based on observations.
3. Create a personalized customer experience: Clustering can be employed to delineate customer categories, enabling the provision of tailored services, customized advertising campaigns, and fostering personalized customer relationships.

4. Identifying similarities in non-uniform data: Clustering can be utilized to identify similarities in non-uniform data. For instance, in genetics, biology, and other related fields, clustering is employed to detect similarities in data and group them accordingly.
5. Improve decision-making and strategy: Clustering facilitates data-driven decision-making. By leveraging clusters, organizations can develop tailored strategies, analyze customer categories, identify new market segments, and enhance other decision-making processes.
6. Data saves time and resources: It accelerates clustering, data analysis, and prediction processes, thereby saving analysts time and enabling more efficient utilization of critical resources.

We can utilize clustering analysis to extract valuable insights from our data by observing how data points are grouped when applying clustering algorithms. In this article, we will explore five popular clustering algorithms along with their advantages and disadvantages!

The concept of a 'cluster' cannot be clearly defined, which is one of the reasons why there are numerous clustering algorithms. While a group of data objects serves as a common denominator, different researchers employ various cluster models, leading to a diversity of algorithms. The properties of clusters identified by different algorithms exhibit significant variation. Understanding these 'cluster models' is crucial for comprehending the distinctions among various algorithms. Typical cluster models include:

- Connectivity models: Hierarchical clustering constructs models based on the connectivity of distances.
- The centroid model: The k-means algorithm represents each cluster with a single mean vector.
- Distribution models: Clusters are modeled using the expectation-maximization algorithm, which employs statistical distributions similar to multivariate normal distributions.
- Group models: These models do not enhance the results of certain algorithms but solely offer grouping information

As mentioned above, clustering algorithms can be classified based on the cluster model. The examples provided above are just a few prominent ones, as there are over 100 published clustering algorithms. Not all algorithms provide explicit models for their clusters, making their classification less straightforward.

There is no objectively "correct" clustering algorithm, as clustering is subjective and user-driven. Selecting the most suitable clustering algorithm for a specific problem often requires experimental evaluation, unless there is a mathematical basis to favor one clustering model over another. An algorithm designed for one type of model will typically perform poorly on a dataset that follows a fundamentally different model.

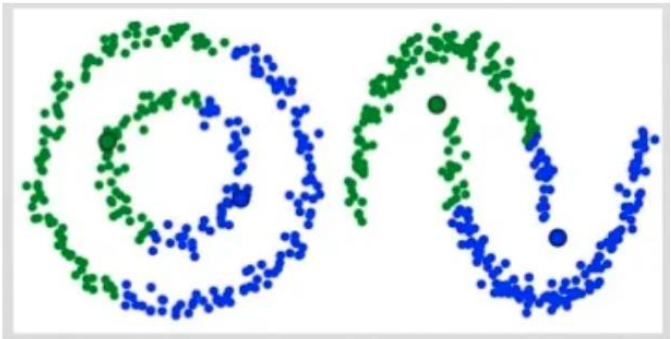
Clustering algorithms employ various methods to identify similarities in data. Some popular clustering algorithms include:

- K-means clustering: This algorithm seeks similarities among a set of objects and aids in grouping the data into a specified number of clusters.

- Hierarchical clustering: This algorithm operates by progressively clustering the data at different levels, grouping the data step by step.
- DBSCAN (Density-Based Spatial Clustering of Applications with Noise): This algorithm is utilized for clustering data with dense point sets. It detects similarities among neighboring points and introduces significant resistance to points where there is insufficient density.
- Gaussian Mixture Model (GMM) clustering: This algorithm partitions the data into clusters using a statistical model. It identifies clusters by leveraging Gaussian probability distributions.

Expectation-Maximization (EM) clustering using Gaussian Mixture Models (GMM)

One of the major drawbacks of K-means is its simplistic use of the mean as the cluster center. This limitation becomes evident when examining the image below. On the left side, it is visually apparent that there are two circular clusters with different radii sharing the same mean center. However, K-means fails to capture this distinction accurately as the average values of the clusters are too close to each other. Additionally, K-means struggles when dealing with non-circular clusters, primarily due to its reliance on the mean as the cluster center.



K-vositalar uchun ikkita nosozlik holati

Gaussian mixture models (GMMs) offer more flexibility compared to k-means. With GMMs, we make the assumption that the data points follow Gaussian distributions, which is a less restrictive assumption than assuming circularity based on the mean. Consequently, there are two parameters to describe the shape of the clusters: mean and standard deviation. For instance, in two dimensions, this implies that clusters can have elliptical shapes, as we account for standard deviations in both the x and y directions. Each Gaussian distribution is then assigned to a specific cluster.

To determine the Gaussian parameters for each cluster (e.g., mean and standard deviation), we employ an optimization algorithm known as expectation-maximization (EM). The following graphic serves as an illustration of clustered Gaussians. Subsequently, we can proceed with the expectation-maximization clustering process using GMMs.

1. We begin by selecting the number of clusters, similar to k-Means, and randomly initializing the parameters of the Gaussian distribution for each cluster. A preliminary examination of the data can offer a reasonable estimate for the initial parameters. However,

as depicted in the graph above, this step is not entirely crucial since the initial Gaussians may result in suboptimal performance but quickly optimize.

2. Having established these Gaussian distributions for each cluster, we can calculate the probability of each data point belonging to a specific cluster. The closer a point is to the center of the Gaussian, the higher the likelihood of it belonging to that cluster. This intuitively aligns with the assumption that, with a Gaussian distribution, the majority of the data lies closer to the cluster center.

3. Based on these probabilities, we compute a new set of parameters for the Gaussian distributions to maximize the likelihood of data points within clusters. We determine these new parameters using a weighted sum of the positions of the data points, where the weights correspond to the probability of a data point belonging to a specific cluster. To illustrate this visually, let's refer to the above graphic, particularly the yellow cluster. In the initial iteration, the distribution starts randomly, but we observe that most of the yellow points lie on the right side of this distribution. When we calculate the weighted sum of the probabilities, the majority of them contribute to the right side, even if there are some points near the center. Consequently, the mean of the distribution naturally converges towards this group of points. Furthermore, we notice that a significant portion of the points exhibits a "top right to bottom left" pattern. Consequently, we adjust the standard deviation to form an ellipse that better fits these points, thereby maximizing the probability-weighted sum.

4. Steps 2 and 3 are repeated until convergence is achieved, indicated by minimal changes in the distributions across iterations. The utilization of GMMs offers two significant advantages. Firstly, GMMs provide flexibility in terms of cluster covariance compared to k-means. By incorporating the standard deviation parameter, clusters are capable of adopting elliptical shapes rather than being restricted to circles. In fact, k-means can be considered as a special case of GMM where the covariance of each cluster approaches zero on all dimensions. Secondly, GMMs leverage probabilities, enabling multiple memberships per data point. Consequently, in the case of a data point lying between two overlapping clusters, we can assign its class as a combination of percentages belonging to class 1 and class 2. In other words, GMMs support mixed membership.

**References**

1. "Pattern Recognition and Machine Learning" - Christopher M. Bishop: This book explains clustering algorithms and basic analysis techniques.
2. "Data Mining: Concepts and Techniques" - Jiawei Han, Micheline Kamber, Jian Pei: This book covers the basic principles of clustering and clustering algorithms.
3. "Introduction to Data Mining" - Pang-Ning Tan, Michael Steinbach, Vipin Kumar: This book covers the basic procedure of clustering and clustering algorithms.
4. "Data Science for Business" - Foster Provost, Tom Fawcett: This book explains data analysis, data-driven decision making, and clustering in business.
5. "Machine Learning: A Probabilistic Perspective" - Kevin P. Murphy: This book provides an extensive review of clustering algorithms and analysis methods based on Bayesian methods.

6. Mulaydinov, F. (2021). Digital Economy Is A Guarantee Of Government And Society Development. Ilkogretim Online, 20(3), 1474-1479. [Please verify the relevance of this source to the topic of clustering algorithms.]
7. Mulaydinov, F. M. (2019). Econometric Modeling of the Innovation Process in Uzbekistan. Forum molodyx uchenyx, (3), 35-43. [Please verify the relevance of this source to the topic of clustering algorithms.]
8. Ahmadjanov, O., Abdullayev, A., Mamayusupov, M., & Umarjanov, O. (2021). Management problems in the digital economy. Science and Education, 2(10), 636-642. [Please verify the relevance of this source to the topic of clustering algorithms.]
9. Mulaydinov, F., Kadirova, A., Melibaeva, G., & Akhmadjonov, O. (2020). Advantages of the transition to a digital economy in the innovative development of Uzbekistan. Journal of Advanced Research in Dynamical and Control Systems, 12(6), 1226-1232. [Please verify the relevance of this source to the topic of clustering algorithms.]
10. Mulaydinov, F., & Nishonkulov, S. (2021). The role of information technologies in the development of the digital economy. [Please verify the relevance of this source to the topic of clustering algorithms.]