iJETRM
International Journal of Engineering Technology Research & Management

UNCOVERING INSIGHTS WITH CLUSTERING: A DATA SCIENCE PERSPECTIVE ON MACHINE LEARNING

Hussein Alhroot*[1]
Dr.Suhailan Safei [1]
[1]Faculty of Informatics and Computing, University Sultan Zainal Abidin, Malaysia

ABSTRACT

Clustering is a Machine Learning technique that groups data points together. Using a clustering technique, we can assign each data point to a particular group given a set of data points. Theoretically, data points belonging to the same group should have similar features and/or characteristics, whereas those belonging to other groups should have vastly different features and/or characteristics. Clustering is a method of unsupervised learning and a common tool for statistical data analysis employed in various fields. In Data Science, we can use clustering analysis to get valuable insights from our data by applying a clustering algorithm and observing which categories the data points fall into.

Keywords:
Clustering, Machine learning, Data science

INTRODUCTION

Clustering is a sort of unsupervised machine learning in which similar data points in a dataset are grouped. How clustering algorithms function is finding patterns in the data and grouping similar data points together. There are various clustering methods, such as k-means, hierarchical clustering, and density-based clustering[1]. Each method has its advantages and disadvantages, and the selection of an algorithm depends on the individual objective and dataset.

What is Clustering?

Clustering is the process of grouping several objects into clusters, so each cluster has data that is as similar as possible and distinct from other clusters' objects. Hierarchical clustering and partitioning are two well-known techniques for clustering. Hierarchical clustering comprises complete linkage clustering, single linkage clustering, average linkage clustering, and centroid linkage clustering. The partitioning method itself is comprised of k-means and fuzzy C-means[2].

Furthermore, Clustering is a data grouping technique or method. The absence of goal variables in clustering differentiates clustering from classification. Clustering does not classify or determine the value of the target variable. Nevertheless, this technique attempts to partition all data into homogeneous groups[3]. Numerous fields employ clustering techniques significantly. For example, clustering can be used in the medical field to categorize disease types based on the type and symptoms of patients.

What is Cluster Analysis?

Cluster analysis is a data mining method that enables researchers to group a set of observations into similar (homogeneous) groups based on criteria used to categorize the observations. In other words, cluster analysis groups related observations into homogeneous subgroups. [4].

Cluster analysis is used to identify similarities in groups of objects. Moreover, using a proximity matrix, cluster analysis maximizes both homogeneity within and heterogeneity between clusters of objects.

cluster analysis is an important technique in machine learning and data mining that has many applications in a variety of domains, such as computer vision, natural language processing, and bioinformatics.

Clustering aims to group objects into sets ("clusters") in which objects belonging to the same cluster are more similar to one another than those within other clusters [5].

Cluster analysis is an essential technique in machine learning and data mining that groups related objects. It provides help with data exploration, dimensionality reduction, pattern recognition, and data compression. Cluster analysis is vital for studying the structure of data, identifying similarities, and making sense of massive, complex datasets.

TYPES OF CLUSTERING

There are various types of clustering, each having its strengths and drawbacks, which can be commonly stated as follows:

- **Hierarchical Clustering:** For this type of clustering, data is grouped according to different levels of similarity, illustrated by a dendrogram's tree-like structure. Generally, hierarchical division follows two methods: divisive and agglomeration[6].
  For divisive clustering, clusters are generated via top-down recursive hierarchical data splitting. Each data item initially belongs to a single cluster. This single cluster is subdivided until a termination requirement is reached or each data item becomes a separate cluster.
  In addition, agglomerative clustering is being implemented from the bottom up, combining data points hierarchically to build clusters. Each data item represents itself initially as a cluster before being merged into larger clusters until a termination requirement is reached or a single cluster including all data items is formed[7].
- **Partitional Clustering:** Partitional clustering is more popular and recommended than hierarchical clustering because of its computing efficiency, especially for large datasets.
  The concept of similarity acts as the measurement metric in this clustering method. Generally, partitional clustering places data items into clusters based on a particular objective function so that data items inside a cluster are more similar to one another than data items in other clusters[8].
  Moreover, the objective function in partitional clustering often is described as the minimization of the within-cluster similarity criterion, which would be calculated using Euclidean distance.
- **Centroid-based Clustering:** Centroid-based clustering is a widely used technique in machine learning and data mining for grouping similar objects into clusters. The main principle behind centroid-based clustering is to represent each cluster by a center point, or centroid, which is the mean of the data points in the cluster. The k-means algorithm is considered one of the most widely used centroid-based algorithms[9].
  The main goal of the k-means algorithm is to partition a set of data points into k clusters, where k is the number of clusters desired. Every cluster is represented by its own centroid, which will be the mean of data points inside it[10].
- **Model-based clustering:** Model-based clustering is a machine learning and data mining technique for clustering similar objects. Modeling the structure of data with a probabilistic model and then utilizing the model to assign data points into clusters is the main idea of model-based clustering.
  Model-based clustering seems to be more flexible and efficient than centroid-based clustering because it can capture complicated and nonlinear data correlations. The algorithms can also work with various data types, including categorical and continuous variables, and can model correlation relationships between variables.

However, model-based clustering methods are more computationally costly and more difficult to implement and analyze than centroid-based techniques. The selection of a model and the number of clusters are other crucial factors, as they can significantly affect the clustering results[11].

TYPES OF CLUSTERS

- **Well-Separated Clusters:** Well-separated clusters are clusters that are discrete from one another, with an obvious boundary separating them. The data points inside the cluster are more similar than those in other clusters. generally, well-separated clusters are easier to interpret and also can reveal deep insights into the data than not well-separated clusters. This is due to the fact that well-separated clusters make it simpler to clearly distinguish between groups of similar objects as well as to understand relationships between them[9].

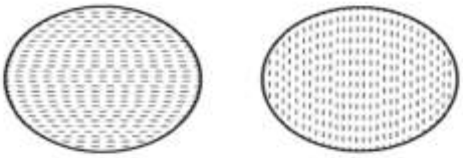

Fig. 1: Well-Separated Clusters

- **Center-based Clusters:** Center-based clustering, commonly identified as centroid-based clustering, is a technique for clustering similar objects depending on their proximity to a centroid or center point.

One of the main advantages of center-based clustering is efficiency and simplicity, which makes it suitable for large and complex data sets. In addition, center-based clustering algorithms may also generate spherical clusters, which can be simpler to analyze and interpret than clusters of more complex shapes.

Nevertheless, center-based clustering algorithms might be sensitive to the initial selection of the centroids and could converge to insufficient solutions if initial centroids are not selected carefully[12].

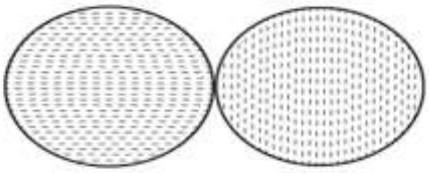

Fig. 2: Center-based Clusters

- **Contiguous Clusters:** Contiguous clustering is a sort of clustering technique that groups similar objects based on their data-space proximity to each other. It involves forming clusters with connected or contiguous regions, in which objects are grouped together if they are close to one another, regardless of the presence of a particular set of neighbors or a center point.

  Contiguous clustering aims to highlight structures in the data that reflect underlying relationships between objects. This type of cluster is commonly used in data analysis to identify patterns and trends. It is helpful for data sets which non-spherical and complex clusters[13].

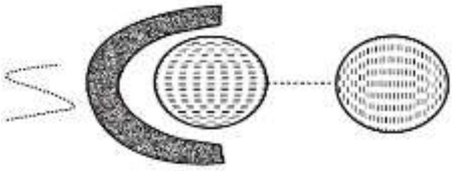

Fig. 3: Contiguous Clusters

- **Conceptual Clusters:** Conceptual clustering is a clustering technique that groups objects according to their features and characteristics rather than proximity to each other in the data space. Conceptual clustering aims to discover concepts or classes that represent the underlying relationships between data objects.

  Moreover, conceptual clustering is a technique that focuses on identifying meaningful and understandable relationships between data objects. This sort of clustering is also well-suited for knowledge discovery and data sets in which identifying meaningful concepts or classes is the goal, and domain knowledge can be incorporated into the clustering process.

  However, conceptual clustering is not well-suited for large and complex data sets since they can be computationally costly and may require a significant amount of memory and processing resources[14].

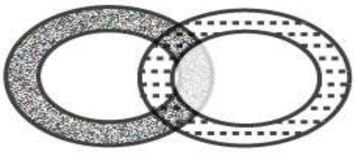

Fig. 4: Conceptual Clusters

CONCLUSION

Clustering is a powerful tool that can help to discover hidden patterns and structures in the data. It can be used to gain insights into the data and make better decisions based on those insights.

In conclusion, the various clustering types have unique features, strengths, and weaknesses. The selection of the suitable clustering technique depends on the particular data set and clustering aims. Data scientists

can select the proper approach to a particular problem if they fully understand the various types of clustering and their advantages and disadvantages.
The following table shows a comparison between different types of clustering techniques:

Table 1: Comparison table for Clustering Techniques

Clustering technique	**Characteristics**	**Advantages**	**Disadvantage**
Hierarchical Clustering	Groups objects into a tree-like structure, where objects at the bottom of the tree are grouped into more general clusters at the top.	Can handle non-spherical clusters, and can produce a clear representation of the relationships among the objects.	Can be computationally expensive, and may not scale well for large data sets.
Partition Clustering	Divides the data into a fixed number of non-overlapping clusters.	Can be computationally efficient, can handle large data sets, and can produce clear and interpretable clusters.	May not be well-suited for non-spherical clusters, and may not produce optimal clusters for complex data sets.
Centroid-based Clustering	Groups objects based on their proximity to a set of central points, or centroids.	Can be computationally efficient, can handle large data sets, and can produce clear and interpretable clusters.	May be sensitive to the initial placement of the centroids, and may not be well-suited for non-spherical clusters.
Model-based Clustering	Groups objects based on a set of underlying probabilistic models.	Can handle complex relationships among the objects, and can incorporate prior knowledge into the clustering process.	Can be computationally expensive, and may not be well-suited for large data sets.

REFERENCES

[1] G. Chao, S. Sun, and J. Bi, "A survey on multiview clustering," *IEEE transactions on artificial intelligence*, vol. 2, no. 2, pp. 146-168, 2021.

[2] S. M. Kim *et al.*, "An evaluation of different clustering methods and distance measures used for grouping metabolic pathways," in *2016 international conference on bioinformatics and computational biology. ISCA*, 2016, pp. 115-122.

[3] R. K. Raman and L. R. Varshney, "9 Universal Clustering," *Information-Theoretic Methods in Data Science*, p. 263, 2021.

[4] K. Denaro, B. Sato, A. Harlow, A. Aebersold, and M. Verma, "Comparison of cluster analysis methodologies for characterization of Classroom Observation Protocol for Undergraduate STEM (COPUS) data," *CBE—Life Sciences Education*, vol. 20, no. 1, p. ar3, 2021.

[5] M. Golzadeh, A. Decan, D. Legay, and T. Mens, "A ground-truth dataset and classification model for detecting bots in GitHub issue and PR comments," *Journal of Systems and Software*, vol. 175, p. 110911, 2021.

[6] S. Malik, A. Rana, and M. Bansal, "Analysis of Current Recommendation Techniques and Evaluation Metrics to Design an Improved Book Recommendation System," in *Electronic Systems and Intelligent Computing: Proceedings of ESIC 2021*: Springer, 2022, pp. 507-524.

[7] T. Li, A. Rezaeipanah, and E. M. T. El Din, "An ensemble agglomerative hierarchical clustering algorithm based on clusters clustering technique and the novel similarity measurement," *Journal of King Saud University-Computer and Information Sciences*, vol. 34, no. 6, pp. 3828-3842, 2022.

[8] H. Mittal, A. C. Pandey, M. Saraswat, S. Kumar, R. Pal, and G. Modwel, "A comprehensive survey of image segmentation: clustering methods, performance parameters, and benchmark datasets," *Multimedia Tools and Applications*, pp. 1-26, 2021.

[9] A. E. Ezugwu *et al.*, "A comprehensive survey of clustering algorithms: State-of-the-art machine learning applications, taxonomy, challenges, and future research prospects," *Engineering Applications of Artificial Intelligence*, vol. 110, p. 104743, 2022.

[10] B. Nepal, M. Yamaha, H. Sahashi, and A. Yokoe, "Analysis of building electricity use pattern using k-means clustering algorithm by determination of better initial centroids and number of clusters," *Energies*, vol. 12, no. 12, p. 2451, 2019.

[11] A. Ahmad and S. S. Khan, "Survey of state-of-the-art mixed data clustering algorithms," *Ieee Access*, vol. 7, pp. 31883-31902, 2019.

[12] H. Bangui, M. Ge, and B. Buhnova, "Exploring Big Data Clustering Algorithms for Internet of Things Applications," in *IoTBDS*, 2018, pp. 269-276.

[13] R. J. Campello, P. Kröger, J. Sander, and A. Zimek, "Density-based clustering," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 10, no. 2, p. e1343, 2020.

[14] A. Saxena *et al.*, "A review of clustering techniques and developments," *Neurocomputing*, vol. 267, pp. 664-681, 2017.