IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 2, April 2022
Impact Factor: 6.252

Different Clustering Algorithms in Data Mining

Apurva Vashist
Delhi Skill and Entrepreneurship University, New Delhi, India
apurva.vashist@gmail.com

Abstract: *Clustering is the grouping together of similar data items into clusters. Clustering analysis is one of the main analytical methods in data mining; the method of clustering algorithm will influence the clustering results directly. This paper discusses the various types of algorithms like Hierarchical Clustering Algorithms Partitioning Method Nearest Neighbor algorithm K-Mean (A centroid based Technique) Density-Based clustering etc. This paper also deals with the problems of clustering algorithm such as time complexity and accuracy to provide the better results based on various environments.*

Keywords: Clustering; Datasets; Machine-Learning; Data Mining

I. INTRODUCTION

In clustering, an assembly of different data objects is classified as similar items. One group means a cluster of records. Data sets are separated into different groups in the cluster analysis, which is based on the similarity of the data. After the classification of data into several groups, a label is assigned to the group[1]. It helps in adapting to the modifications by doing the classification. Cluster Analysis in Data Mining means that to find out the group of items which are similar to each other in the group but are different from the item in other groups. The simplest definition is shared among all and includes one important concept: the grouping together of similar data items into clusters. Cluster analysis is the organization of a assembly of patterns (usually represented as a vector of sizes, or a point in a multidimensional space) into clusters based on resemblance. It is important to recognize the difference between clustering (unsupervised classification) and discriminant analysis (supervised classification). In supervised classification, we are delivered with a collection of labeled (preclassified) patterns [1]; the problem is to label a newly encountered, yet unlabeled, pattern.

II. APPLICATIONS

There are several uses of Data clustering analysis such as image processing, data analysis, pattern recognition, market research and many more. Using Data clustering, companies can discover fresh groups in the database of consumers. Classification of information can also be done based on patterns of buying.

Clustering in Data Mining supports in the classification of animals and plants are done using similar functions in the field of biology. Areas are recognized using the clustering in data mining. In the database of earth observation, lands are identified which are similar to each other.

Requirements of Clustering in Data Mining

- Interpretability
- Helps in dealing with messed up data
- High Dimensional
- Attribute shape clusters are discovered
- Algorithm Usability with multiple data kind
- Clustering Scalability

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 2, April 2022
Impact Factor: 6.252

III. CLUSTERING ALGORITHMS

3.1 Hierarchical Clustering Algorithms

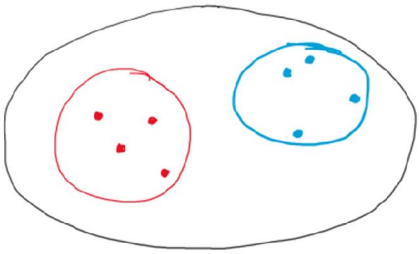

Figure 1: Hierarchical Clustering Algorithms

Hierarchical clustering is a different unsupervised machine learning algorithm, which is used to group the unlabeled datasets into a cluster and also famous as **hierarchical clustering analysis** or HCA.

In this algorithm, we improve the order of clusters in the form of a tree, and this tree-shaped arrangement is known as the **dendrogram.**

Sometimes the results of K-means clustering and hierarchical clustering may look alike, but they both differ depending on how they work together. As there is no requirement to predetermine the quantity of clusters as we did in the K-Means algorithm. Among the most used variations of the hierarchical clustering based on different distance measures are:

1. **Average Linkage Clustering:** The dissimilarity between clusters is calculated using average values. The average distance is calculated from the distance between every point in a cluster and all other points in another cluster. The 2 clusters with the lowest average distance are combined together to form the fresh cluster[13].
2. **Centroid Linkage Clustering:** This variation uses the group centroid as the average. The centroid is well-defined as the center of a cloud of points.
3. **Complete Linkage Clustering (Maximum or Furthest-Neighbor Method):** The dissimilarity between two groups is equal to the greatest dissimilarity between a member of cluster i and a member of cluster j. This procedure tends to produce very tight clusters of similar cases.
4. **Single Linkage Clustering (Minimum or Nearest-Neighbor Method):** The dissimilarity between two clusters is the minimum dissimilarity between members of the two clusters. This process produces long chains which form loose, straggly clusters.
5. **Ward's Method:** Cluster membership is assigned by computing the total sum of squared deviations from the mean of a cluster. The criterion for merging is that it should produce the minimum possible increase in the error sum of squares.

3.2 Partitioning Method

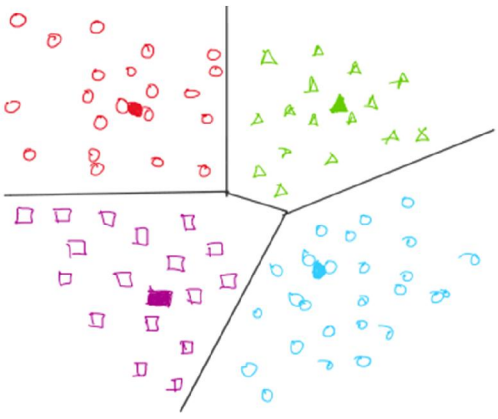

Figure 2: Partitioning Method

This clustering process classifies the data into several groups based on the characteristics and similarity of the data. It's the data analysts to specify the number of clusters that has to be generated for the clustering systems.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 2, April 2022
Impact Factor: 6.252

In the partitioning process when database(D) that contains multiple(N) objects then the partitioning process constructs user-specified(K) partitions of the data in which each partition represents a cluster and a particular region. There are several algorithms that come under partitioning process some of the famous ones are K-Mean, PAM(K-Mediods), CLARA algorithm (Clustering Large Applications) etc.

3.3 Nearest Neighbor Algorithm

The main idea of the algorithm is to find sets of clusters to merge by following paths in the nearest neighbour graph of the clusters. Every such path will finally terminate at a pair of clusters that are nearest neighbors of each other, and the algorithm selects that pair of clusters as the pair to merge. In order to save work by re-using as much as possible of every path, the algorithm uses a stack data structure to keep track of every path that it follows. By following paths in this way, the nearest-neighbor chain algorithm combines its clusters in a different order than procedures that always find and merge the closest pair of clusters. However, despite that difference, it always produces the same hierarchy of clusters.

The nearest-neighbor chain algorithm constructs a clustering in time proportional to the square of the number of points to be clustered. This is also proportional to the size of its input, when the input is provided in the form of an explicit distance matrix. The algorithm uses an amount of memory proportional to the number of ideas, when it is used for clustering procedures such as Ward's method that allow constant-time calculation of the distance between clusters. However, for some other clustering methods it uses a higher amount of memory in an auxiliary data structure with which it keeps track of the distances between sets of clusters.

3.4 K-Mean (A Centroid Based Technique)

The K means algorithm takes the input parameter K from the user and partitions the dataset containing N items into K clusters so that resulting similarity among the data objects inside the group (intracluster) is high but the similarity of data objects with the data objects from outside the cluster is low (intercluster). The similarity of the cluster is determined with respect to the mean value of the cluster[14].

It is a type of square error algorithm. At the start randomly k objects from the dataset are selected in which each of the objects represents a cluster mean(centre). For the rest of the data items, they are assigned to the nearest cluster based on their distance from the cluster mean. The new mean of each of the cluster is then calculated with the added data objects.

3.5 Mean-Shift Clustering

Mean shift clustering is a sliding-window-based algorithm that attempts to find dense areas of data points. It is a centroid-based algorithm meaning that the objective is to locate the center points of every group/class, which works by updating candidates for center points to be the mean of the points within the sliding-window. These candidate windows are then filtered in a post-processing stage to eliminate near-duplicates, forming the final set of center points and their equivalent groups.

3.6 Expectation–Maximization (EM) Clustering using Gaussian Mixture Models (GMM)

One of the main drawbacks of K-Means is its naive use of the mean value for the cluster center. We can see why this isn't the greatest way of doing things by looking at the image below. On the left-hand side, it looks quite obvious to the human eye that there are two circular clusters with changedradius'centered at the same mean. K-Means can't handle this because the mean values of the clusters are very near together. K-Means also fails in cases where the clusters are not circular, again as a result of using the mean as cluster center.

Figure 3: Expectation–Maximization (EM) Clustering

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 2, April 2022
Impact Factor: 6.252

3.7 Density-Based Clustering

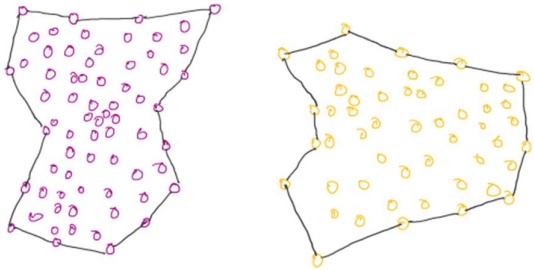

Figure 4: Density-Based clustering

The density-based clustering process connects the highly-dense areas into clusters, and the arbitrarily shaped distributions are made as long as the dense region can be joined together. This algorithm does it by identifying different clusters in the dataset and connects the areas of high densities into clusters. The dense areas in data space are separated from each other by sparser areas. These algorithms can face difficulty in clustering the data points if the dataset has varying densities and high dimensions.

IV. Conclusion

Clustering algorithms are the great way to learn different things from old data/information. Sometimes you'll be surprised by the resulting clusters that you acquire and it might help you make sense of a particular problem. One of the great things about using clustering for unsupervised learning is that you can use the effects in a supervised learning problem. The clusters could be your latest features that you use on a completely different data set! You can use clustering on just about any unsupervised machine learning problem, but make sure that you know how to analyze the outcomes for accuracy. In this paper, we have learnt different kinds of clustering methods. By using these methods, we can solve newer problem of the data set. In the future, we can combine multiple methods together.

Acknowledgment

I am really thankful to Mr. Ashutosh Vashist, who was highly cooperative and supportive during the completion of this paper.

References

[1]. L. Parsons, E. Haque, and H. Liu, "Subspace clustering forhigh dimensional data: a review," ACM SIGKDD Explorations Newsletter, vol. 6, pp. 90-105, 2004.

[2]. C. C. Aggarwal, J. Han, J. Wang, and P. S. Yu, "A framework for projected clustering of high dimensional data streams," in Proceedings of the Thirtieth international conference on Very large data bases-Volume 30, 2004, p.863.

[3]. R. Agrawal, J. E. Gehrke, D. Gunopulos, and P. Raghavan, "Automatic subspace clustering of high dimensional data fordata mining applications," Google Patents, 1999.

[4]. C. Domeniconi, D. Papadopoulos, D. Gunopulos, and S. Ma, "Subspace clustering of high dimensional data," 2004.

[5]. X. Z. Fern and C. E. Brodley, "Random projection for high dimensional data clustering: A cluster ensemble approach,"2003, p. 186.

[6]. H. P. Kriegel, P. Kröger, and A. Zimek, "Clustering high-dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering," 2009.

[7]. R. XU and I. Donald C. Wunsch, clustering: A Johnwiley& Sons, INC., Pub, 2008.

[8]. Guha, Meyerson, A. Mishra, N. Motwani, and O. C. ."Clustering data streams: Theory and practice. "IEEE Transactions on Knowledge and Data Engineering, vol. 15,pp. 515-528, 2003.

[9]. A. Jain , M. Murty , and p. Flynn " Data clustering: A review.," ACM Computing Surveys, vol. 31, pp. 264-323, 1999.

[10]. P. C. Biswal, Discrete Mathematics and Graph Theory. New Delhi: Prentice Hall of India, 2005. [11] M. Khalilian,

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 2, April 2022
Impact Factor: 6.252

Discrete Mathematics Structure. Karaj: Sarafraz, 2004.

[11]. K. J. Cios, W. Pedrycz, and R. M. Swiniarsk, "Data mining methods for knowledge discovery," IEEE Transactions on Neural Networks, vol. 9, pp. 1533-1534, 1998.

[12]. V. V. Raghavan and K. Birchard, "A clustering strategy based on a formalism of the reproductive process in natural systems," 1979, pp. 10-22.

[13]. D. Isa, V. P. Kallimani, and L. H. Lee, "Using the self organizing map for clustering of text documents," Expert Systems With Applications, vol. 36, pp. 9584-9591, 2009

[14]. Shi Na, Liu Xumin, "Research on k-means Clustering Algorithm", IEEE Third International Conference on Intelligent Information Technology and Security Informatics,2010

[15]. S. Mythili ,E. Madhiya," An Analysis on Clustering Algorithms in Data Mining", International Journal of Computer Science and Mobile Computing,2014.