IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.918 ⁜ Vol. 11, Issue 12, December 2022

DOI: 10.17148/IJARCCE.2022.111211

Review Paper on Data Mining Clustering Algorithms

Harshali R. Tapase[1], Vijay M. Rakhade [2], Lowlesh N. Yadav[3]

Final Year Student, Computer Science Engineering, Shri Sai College of Engineering & Technology, Bhadrawati, India[1]

Professor, Computer Science Engineering, Shri Sai College of Engineering & Technology, Bhadrawati, India[2]

Professor, Computer Science Engineering, Shri Sai College of Engineering & Technology, Bhadrawati, India[3]

Abstract: Data mining is the method of extracting and discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database systems. Clustering performs an important role in the reference composition of data analysis. Clustering is a famous data analysis and data mining problem. Symmetry can be studied as a pre-attentive feature, which can enhance shapes and objects, as reconstruction and recognition. Clustering, recognized as a crucial issue of unsupervised learning, deals with the segmentation of the data structure in an unknown region and is the basis for more understanding. This paper explains the different types of clustering and methods in data mining.

Keywords: Data Mining, Clustering, Algorithm, k-means Clustering.

I. INTRODUCTION

Data mining refers to extracting information from large amounts of data and transforming that information into an understandable and meaningful structure for more use. Clustering can be considered a key problem in data analysis and data mining. It tries to group similar data objects into sets of disjoint classes or clusters. Clustering analysis has been an emerging research issue in data mining due to its quality of applications. The advent of many data clustering algorithms in the recent few years and their extensive use in a wide variety of applications, including image processing, computational biology, mobile communication, medicine, and economics, has led to the popularity of these algorithms.

II. DATA MINING

Data mining is the system of sorting through large data sets to identify patterns and relationships that can help explain business problems through data analysis. Data mining techniques and tools approve enterprises to predict future trends and produce more-informed business decisions. There are some steps in the data mining process: Data Cleaning, Data Integration, Data Reduction, Data Transformation, Data Mining, Pattern Evaluation, and Knowledge Representation.

1. Data Cleaning: This step involves identifying and removing any incomplete, inaccurate, or irrelevant data from the dataset.
2. Data Integration: This step involves combining data from multiple sources into a single dataset.
3. Data Reduction: This step involves reducing the size of the dataset by removing redundant or irrelevant data.
4. Data Transformation: This step involves transforming the data into a format that is more suitable for data mining.
5. Data Mining: This step involves applying algorithms and techniques to the dataset to identify patterns and relationships.
6. Pattern Evaluation: Pattern evaluation is used to assess the quality of the discovered patterns. This can include assessing the accuracy, stability, and scalability of the patterns.
7. Knowledge Representation: Knowledge representation is the process of representing the discovered patterns in a form that can be used by decision makers. This can include visualization.

III. CLUSTERING

In clustering, a group of dissimilar data objects is classified as similar objects. One group means a cluster of data. Data sets are divided into various groups in the cluster analysis, which is hanged on the similarity of the data. After the classification of data into various groups, a tag is imputed to the group. It supports in modifying to the changes by doing the classification.

Clustering is an essential technique in data mining and it is the procedure of partitioning data into a set of clusters such that each object in a cluster is near to another object in the same cluster, and dissimilar to every object not in the same cluster.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.918 ⁜ Vol. 11, Issue 12, December 2022

DOI: 10.17148/IJARCCE.2022.111211

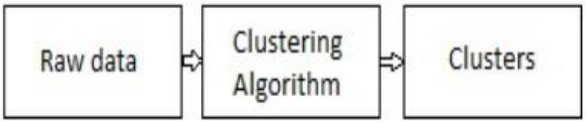

Fig. Stages of Clustering

Types of clustering are: -

1. Partitioning Clustering Method

In this method, let us say that "m" partition is complete on the "p" objects of the database. A cluster will be defined by each partition and m < p. K is the symbol of groups after the classification of objects. There are some demands which need to be satisfied with this Partitioning Clustering Method and they are: –

1. One objective should only belong to only one group.
2. There should be no group without even a single purpose.

There are certain points which should be remind in this category of Partitioning Clustering Method which are:

1. There will be an initial partitioning if we still give no. of a partition (say m).
2. There is one technique called iterative relocation, which means the object will be shifted from one group to another to improve the partitioning.

2. Hierarchical Clustering Methods

Between the multiple different types of clustering in data mining, in this hierarchical clustering method, the given set of an object of data is created into a kind of hierarchical decomposition. The formation of hierarchical decomposition will decide the causes of classification. There are two categories of approaches for the creation of hierarchical decomposition, which are: –

2.1. Divisive Approach

Another name for the isolating approach is a top-down approach. At the beginning of this approach, all the data objects are retained in the same cluster. Smaller clusters are created by splitting the group by using the continuous iteration. The constant iteration method will keep on going until the condition of termination is met. One cannot undo after the group is divide or merged, and that is why this approach is not so flexible.

2.2. Agglomerative Approach

Another name for this method is the bottom-up approach. All the groups are divided in the beginning. Then it keeps on merging until all the groups are merged, or condition of termination is met.

There are two methods which can be used to enhance the Hierarchical Clustering Quality in Data Mining which are:

1. One should carefully analyse the similarity of the object at every partitioning of hierarchical clustering.
2. One can use a hierarchical agglomerative algorithm for the integration of hierarchical agglomeration. In this approach, first, the objects are banded into micro-clusters. After grouping data objects into micro clusters, macro clustering is performed on the micro cluster.

3. Density-Based Clustering Method

In this approach of clustering in Data Mining, density is the major focus. The notion of mass is used as the foundation for this clustering method. In this clustering approach, the cluster will keep on developing continuously. At least one number of points should be there in the radius of the band for each point of data.

4. Grid-Based Clustering Method

In this type of Grid-Based Clustering Method, a grid is found using the object together. A Grid Structure found by quantifying the object space within a finite number of cells.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.918 ⁜ Vol. 11, Issue 12, December 2022

DOI: 10.17148/IJARCCE.2022.111211

Advantage of Grid-based clustering method: –

1. Faster time of processing: The processing time of this method is much quicker than another way, and thus it can save time.
2. This method depends on the no. of cells in the space of quantized each dimension.

5. Model-Based Clustering Methods

In this type of clustering method, each cluster is hypothesized so that it can find the data which is appropriate for the model. The density function is clustered to detect the group in this method.

6. Constraint-Based Clustering Method

Application or user-oriented constraints are merged to perform the clustering. The assumption of the user is indicated to as the constraint. In this process of grouping, communication is very connected, which is brought by the restrictions.

IV. CLUSTERING ALGORITHM

Clustering is a procedure which dividing a given data set into homogeneous groups based on given features such that similar objects are kept in a group whereas dissimilar objects are in other groups. It is the most essential unsupervised learning problem. It deals with detecting structure in a collection of unlabelled data.

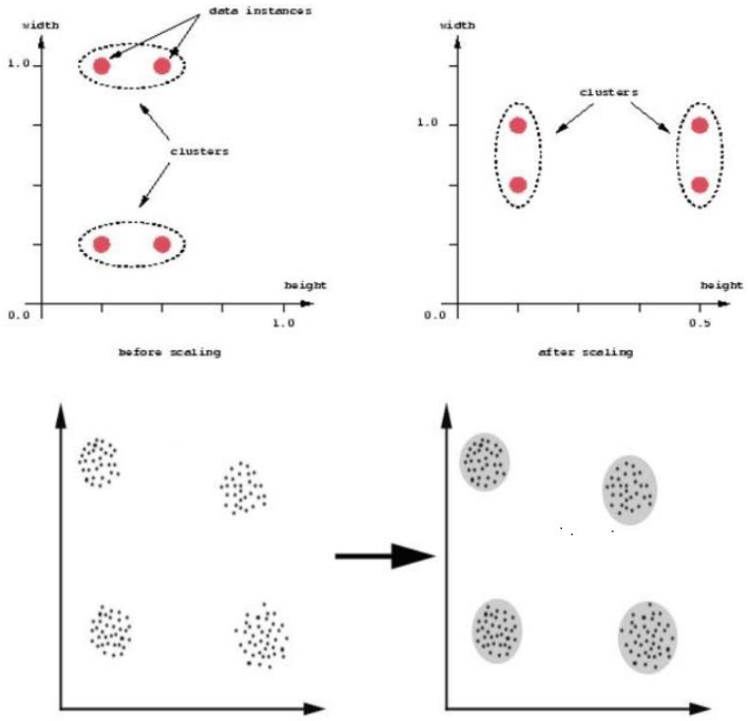

Fig: showing example where scalability may lead to wrong result

For clustering algorithm to be fortunate and beneficial some of the conditions need to be satisfied.

1) Scalability - Data must be scalable else we may get the wrong result. Fig shows simple graphical example where we may get the wrong result.
2) Clustering algorithm must be able to deal with various types of attributes.
3) Clustering algorithm must be able to find clustered data with the arrogant shape.
4) Clustering algorithm must be inconsiderate to noise and outliers.
5) Interpret-ability and Usability - Result gained must be interpretable and usable so that maximum knowledge about the input parameters can be gained.
6) Clustering algorithm must be able to deal with data set of high dimensionalities.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.918 ⁜ Vol. 11, Issue 12, December 2022

DOI: 10.17148/IJARCCE.2022.111211

Clustering algorithms can be broadly classified into two categories:

1) Unsupervised linear clustering algorithms and
2) Unsupervised non-linear clustering algorithms

The most commonly used clustering algorithms are:

1. K-Means: K-Means is the most popular and widely used clustering algorithm. It is used to partition data set into K clusters. It is a simple and efficient algorithm. K-means clustering algorithm is based on the idea of minimizing the within-cluster variance. It assigns each data point to a cluster such that the sum of squares of the distances of each data point from the cluster centroid is minimized.

2. Hierarchical Clustering: Hierarchical clustering is an unsupervised learning algorithm which divides the dataset into a set of clusters. In this algorithm, the clusters are formed by merging the similar data points. This algorithm is used to group the data points which have similar characteristics.

3. DBSCAN: DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is a clustering algorithm which groups the data points into clusters. In this algorithm, the data points which are densely packed together are considered to be in.

4. OPTICS: OPTICS Clustering stands for Ordering Points to Identify Cluster Structure. It draws inspiration from the DBSCAN clustering algorithm. It adds two more terms to the concepts of DBSCAN clustering.

5. Mean-Shift: Mean shift is an unsupervised learning algorithm that is mainly used for clustering. It is widely used in real-world data analysis (e.g., image segmentation) since it's non-parametric and doesn't require any predefined shape of the clusters in the feature space.

6. Affinity Propagation: In contrast to other traditional clustering methods, Affinity Propagation does not require you to specify the number of clusters. In layman's terms, in Affinity Propagation, each data point sends messages to all other points informing its targets of each target's relative attractiveness to the sender.

7. Spectral Clustering: Spectral clustering is a technique with roots in graph theory, where the approach is used to identify communities of nodes in a graph based on the edges connecting them. The method is flexible and allows us to cluster non graph data as well.

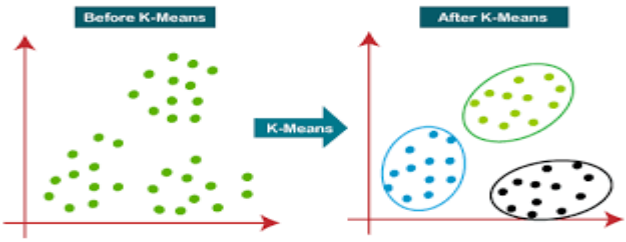

Fig. k-means Algorithm

V. K-MEANS CLUSTERING ALGORITHM

The K-means is a common clustering algorithm for data mining used in many real-life applications, such as healthcare, environment and air pollution, and industry data. K-means clustering is the most accepted partitioning algorithm. K-means reallocated each data in the dataset to basically one of the new clusters found. A record or data point is imputed to the nearest cluster using a measure of distance or similarity.

The k-means algorithm generates the input parameter, k, and division a group of n objects into k. clusters so that the resulting inter cluster similarity is large but the inter cluster analogy is low. Cluster likewise is computed respecting the mean value of the objects in a cluster, which can be considered at as the cluster's centroid or centre of gravity.

There are the following steps used in the K-means clustering −

- It can choose K initial cluster centroid c1, c2, c3…. ck.
- It can allocate each instance x in the S cluster whose centroid is nearest to x.

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.918 ⁜ Vol. 11, Issue 12, December 2022

DOI: 10.17148/IJARCCE.2022.111211

- For each cluster, recalculate its centroid based on which elements are contained in that cluster.
- Go to (b) until intersection is completed.
- It can split the object (data points) into K clusters.
- It is used to cluster centre (centroid) = the average of all the data points in the cluster.
- It can allocate each point to the cluster whose centroid is nearest (using distance function).

The original values for the means are randomly authorized. These can be imputed randomly or maybe can use the values from the first k input items themselves. The intersection element can be based on the squared error, but they are appropriate not to be. For example, the algorithm is imputed to different clusters. Other conclusion techniques have simply locked at a fixed amount of emphasis. A maximum number emphasis of can be included to ensure shopping even without merging.

VI. CONCLUSION

This paper aims to provide an overview of the algorithm used in different clustering techniques and their respective advantages and disadvantages. clustering is the dynamic field of research in data mining the ability to discover highly correlated regions of the object becomes desirable when the data set groans in this paper detailed literature about various data clustering algorithms for categorical data mentions different data clustering techniques have their own advantage and disadvantage.

REFERENCES

[1]. S. M. Metev and V. P. Veiko, Laser Assisted Microtechnology, 2nd ed., R. M. Osgood, Jr., Ed. Berlin, Germany: Springer-Verlag, 1998.
[2]. J. Breckling, Ed., The Analysis of Directional Time Series: Applications to Wind Speed and Direction, ser. Lecture Notes in Statistics. Berlin, Germany: Springer, 1989, vol. 61.
[3]. S. Zhang, C. Zhu, J. K. O. Sin, and P. K. T. Mok, "A novel ultrathin elevated channel low-temperature poly-Si TFT," IEEE Electron Device Lett., vol. 20, pp. 569–571, Nov. 1999.
[4]. M. Wegmuller, J. P. von der Weid, P. Oberson, and N. Gisin, "High resolution fiber distributed measurements with coherent OFDR," in Proc. ECOC'00, 2000, paper 11.3.4, p. 109.
[5]. R. E. Sorace, V. S. Reinhardt, and S. A. Vaughn, "High-speed digital-to-RF converter," U.S. Patent 5 668 842, Sept. 16, 1997.
[6]. (2002) The IEEE website. [Online]. Available: http://www.ieee.org/
[7]. M. Shell. (2002) IEEEtran homepage on CTAN. [Online]. Available: http://www.ctan.org/tex-archive/macros/latex/ contrib. /supported/ IEEEtran/
[8]. FLEXChip Signal Processor (MC68175/D), Motorola, 1996.
[9]. "PDCA12-70 data sheet," Opto Speed SA, Mezzovico, Switzerland.
[10]. A. Karnik, "Performance of TCP congestion control with rate feedback: TCP/ABR and rate adaptive TCP/IP," M. Eng. thesis, Indian Institute of Science, Bangalore, India, Jan. 1999.
[11]. J. Padhye, V. Firoiu, and D. Towsley, "A stochastic model of TCP Reno congestion avoidance and control," Univ. of Massachusetts, Amherst, MA, CMPSCI Tech. Rep. 99-02, 1999.
[12]. Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specification, IEEE Std. 802.11, 1997.