International Journal of Scientific Research in Computer Science, Engineering and Information Technology

ISSN : 2456-3307

OPEN ACCESS

Available Online at : www.ijsrcseit.com
doi : https://doi.org/10.32628/CSEIT2410251

Performance Analysis and Evaluation of Clustering Algorithms using WEKA

Shital Patel, Pooja Pancholi, Arpita Chaudhury

Department of Computer Science, Ganpat University, Mehsana, Gujarat, India

ARTICLEINFO

Article History:

Accepted: 10 April 2024
Published: 22 April 2024

Publication Issue
Volume 10, Issue 2
March-April-2024

Page Number
677-684

ABSTRACT

Clustering, an unsupervised learning technique, to find inherent groupings in un-labelled data. It seems to be referring to a study or research paper that examines and uses a number of clustering algorithms, including the canopy method, k-Means clustering, hierarchical clustering, density-based clustering, and EM algorithm. WEKA, a clustering program, is used for the examination of these techniques. and the effectiveness of these algorithms is evaluated through experiments using social network Ads datasets. The goal of this research paper or study seems to be to assess how well these clustering algorithms perform in grouping data within social network Ads datasets. Such analyses can help identify the most suitable algorithm for a specific type of data or problem domain and may lead to insights into the underlying structure of the data.

Keywords : K-Means Clustering, Hierarchical Clustering, Density Based Clustering, EM Algorithm, Canopy Algorithm, WEKA tool.

I. INTRODUCTION

1.1 Clustering

Data is divided up into categories of objects with similarities using clustering. Each grouping called cluster. In the data analysis and data mining application, the main task is to do clustering. It is the task to mixture a set of objects so that objects in the similar group are more associated to each other than so those in other groupA cluster is a structured group of data that is ordered in a particular sequence and has similar characteristics. Cluster analysis involves identifying similarities among data points based on their inherent characteristics and then organizing these akin data objects into distinct clusters or groups. Since clustering is an unsupervised learning process, the technique requires no inspection and works with unlabeled data.

A set of unsupervised machine learning techniques called clustering algorithms is used to arrange related data points into clusters or segments. They are these algorithms help discover hidden patterns, structures, or relationships within datasets[5]. There are several clustering algorithms available, each with its own strengths and weaknesses. The WEKA tool has been utilized to contrast various clustering methods. It is employed because, when compared to other data mining tools, it has a superior user interface. Here are some of the most commonly used clustering algorithms: hierarchical-based algorithms, density-based cluster algorithms, K-means algorithms, EM algorithms and canopy algorithms. density-based cluster algorithms, K-means algorithms, EM algorithms and canopy algorithms.

The following diagram shows the process of working of the clustering algorithm. Imagine you have a collection of various shapes like circles, triangles, and squares, each with its own unique properties such as size, color, and orientation. You want to organize these shapes into groups or clusters based on their similarities. Here we can see that the different Shapes are divided into different groups with similar properties.[12] (In Figure. 1).

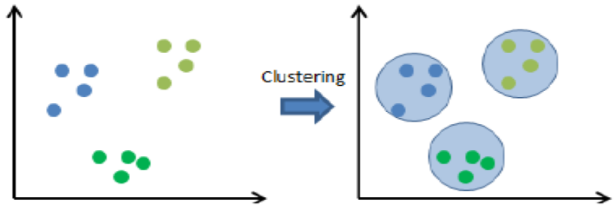

Fig. 1. Clustering Diagram

One single group may be identified among the objects grouped together in figure 1. It is able to recognize the clusters and count the three distinct clusters. The remainder of the essay is structured as follows. Section 2 presents the clustering algorithm, while Section 3 displays the experimental findings. And section 4 of the paper comes to an end.

Clustering Algorithm

2.1 K-means Clustering

Using K-means, an anonymous machine learning technique, you may classify data points according to how similar they are. This clustering algorithm is complete. The k-mean is determined by the exact assignment of each unit to a single cluster. We'll start deciding how much clustering from our dataset we want for the various clustering algorithms. Therefore, we shall use to denote the value of k.

The method will partition the data into K clusters; K is a user-defined hyperparameter that specifies the desired number of clusters. These k values, which could be quite huge, are fictitiously the biggest integers, such as 2, 3, or 4. Therefore, we go back and decide which values of k will be chosen from the available possibilities. [1]

The main goal of K-means is to maximize cluster dissimilarity while minimizing variation within each cluster. Numerous industries, including image segmentation, customer segmentation, anomaly detection, and pattern recognition, have found widespread use for this technique. K-means is one of the core tools for exploratory data analysis and data mining since it provides a straightforward and effective method of clustering.[2]

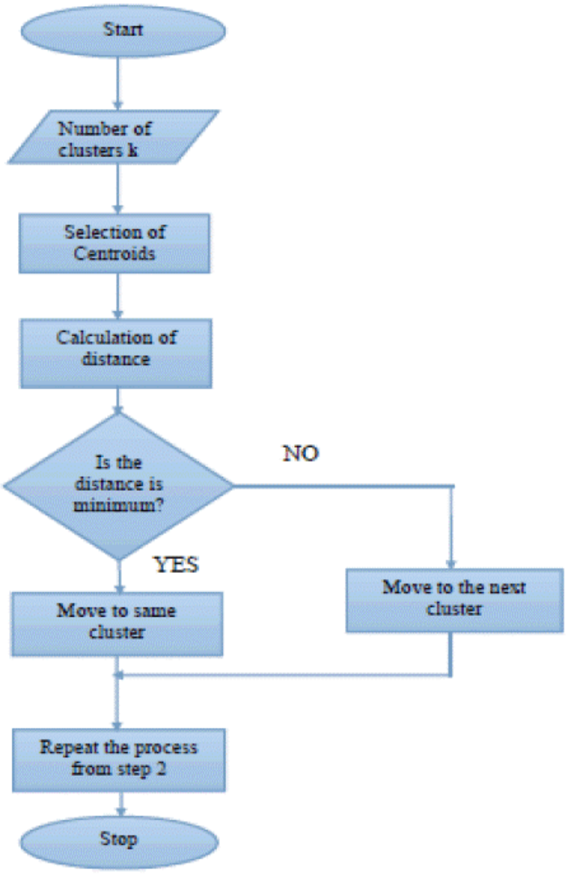

Fig. 2. K-mean clustering process.

The K-means algorithm operates through an iterative process that converges to a locally optimal solution. The steps involved are as follows [5]

Step 1. Start: Typically select K data points from the datset as the starting cluster centroids.

Step 2. Calculate the distance between each centroid and each data point, then assign each point to the group that the closest centroid most closely resembles.

Step 3. Recalculate each group's mean by averaging all the data points that belong to that group.

Step 4. Repeat. Until convergence occurs, the centroids remain constant, or the number of calculations stays the same, steps 2 and 3 are repeated..

Table 1 show the advantages and disadvantages of K-Means algorithm.

Advantages	**Disadvantages**
Frequently used and easy to implement.	Only do it for the shaped clusters.
The fastest way is to calculate	It could be challenging to forecast K's behaviour due to the fixed number of groups.
It can be expanded	Ignores non-globular data of varying size and density.
Faster for low resolution data	Ignores emissions and noise
Creates fixed groups	Limited to data whose mean centre (centroid)
If a large data group number is found, repeat the search for smaller groups.	Assumes spherical clusters

Table. 1. Advatages and disadvantages of K-mean clustering process.

2.2 Hierarchical Clustering:

Hierarchical clustering is a popular and versatile technique in the field of unsupervised machine learning, used to classify data into a hierarchical structure of nested clusters Hierarchical clustering creates a dendrogram, which represents the relationships between data points and groupings at various degrees of granularity, as opposed to K-means clustering, which assigns data points to a set number of clusters.[2] This approach allows us to explore the underlying hierarchical organization of the data, making it particularly useful for datasets where the number of clusters is not known in advance or when insights into subgroups within clusters are desired. Hierarchical clustering can be employed in various domains, such as biology for gene expression analysis, social sciences for clustering individuals based on their behaviour and data exploration to gain a deeper understanding of complex datasets. By revealing the inherent structure of data, hierarchical clustering serves as a valuable tool for pattern recognition and data-driven decision-making.[3]

This algorithm's output is typically displayed as a dendrogram. Based on how the hierarchies are created, it is typically divided into agglomerative and divisive techniques [2].

Agglomerative: This method uses a "bottom-up" strategy. Each object is initially placed in its own cluster. Then, until all of the items are in a single cluster or until specified termination conditions are met, combines these tiny clusters into larger and larger clusters. Because to its O (n3) complexity, it is too sluggish for large data sets.

• Divisive: It follows a "top down" methodology. Beginning with a cluster of all the objects. Recursive splits are then carried out as one descends the hierarchy. Its O (2n) complexity makes it worse.

Table 2 show the advantages and disadvantages of Hierarchical clustering.

Advantages	Disadvantages
The idea is simple	Cluster splitting or merging is permanent, making it impossible to keep track of subsequent errors.
Suitable for small data	Sensitivity to noise and external objects
It does not require a prior number of k nodes.	Difficulty working with different sets and convex shapes.
Addition/subtraction of a cluster is constant, minus is the opposite.	Difficulties may arise in calculating allocation methods.
The data set's noise has no effect on it.	Large files cannot be accommodated by the approaches.
Completed/read status required	• None of the objective functions are given properly.

Table. 2. Advantages and disadvantages of Hierarchical clustering process.

2.3 Density based algorithm

Hinneburg and Keim, in their 1998 work, shifted the focus away from calculating densities associated with individual data points. Instead, they introduced the concept of computing density functions across the underlying property space. Their innovation led to the creation of the DENCLUE (Density-based Clustering) algorithm, which relies on a solid mathematical foundation known as DBCLASD. DENCLUE operates by employing a density function to detect clusters, with its core ideas centered around the notions of density and connectedness. These two fundamental concepts are determined by assessing the localized distribution of nearest neighbors.

In the realm of density-based clustering, the DBSCAN (Density-Based Spatial Clustering Application with Noise) algorithm stands as a prominent example. It excels in clustering low-dimensional spatial data. To employ DBSCAN effectively, several input parameters must be defined.

1) ε -environment Nε (x) = {yεX |d (x, y) ≤ ε }

A point x designates a data point's neighborhood, with two points being regarded neighbors if their distances are less than or equal to "eps."

2) Significant (item with many MinPts points)

3) A dense point y is an idea that may be generated from a significant x (a limited series of important factors between x and y, each of which corresponds to the community of its predecessor)

4) The two points' coordinates, x, and y (they must be fixable from a single origin). [2][3].

Table 3 show the advantages and disadvantages of density-based clustering.

Advantages	Disadvantages
It's good if the clusters are not normal	Databases are complex due to varying densities.
Returns the divisor and quantity	Contains minPoints and EPS parameters.
Used if the data is noisy	The sample includes solid measurements.
Used when external to data	Can be computationally expensive
Gives the closest result to K-means algorithms.	Can have difficulty identifying clusters in high-dimensional data

Table.-3. Advatages and disadvantages of density-based clustering process.

2.4 EM Algorithm

The maximum likelihood or maximum a posteriori (MAP) estimates of parameters in statistical models with latent or unobservable variables are computed using the EM Algorithm, an iterative process. The method is made up of a series of steps that go back and forth between the expectation step (E-step) and the maximization step (M-step).

1. Expectation (E) Step: Using the most recent parameter estimations, the technique computes the expected log-likelihood of the model's parameters in this step. Given the observed data and the existing parameter estimations, it entails figuring out the distribution of the latent variables. Because it determines the predicted values of the latent variables based on the existing model, this phase is known as the "expectation" step.

2. Maximization (M) Step: Using the expected log-likelihood calculated in the E-step as a maximum, the technique computes fresh parameter estimates in the maximization phase. The model's parameters are updated using these new parameter estimates. The "maximization" process seeks to identify the parameter values that, given the predicted values of the latent variables, make the observed data the most plausible.[2][4]

3. Iterative Process: Until convergence is attained, the EM algorithm repeats these first two phases recursively. When the difference in parameter estimations between iterations is less than a predetermined threshold or when the maximum number of iterations is reached, convergence usually happens.[4]

When there are latent variables that cannot be directly observed, the main objective of the EM algorithm is to identify the parameter values that maximize the likelihood of the observed data.It is commonly used in various statistical and machine learning applications, such as clustering, mixture models, and hidden Markov models, where latent variables play a crucial role in the modeling process.

Table 4 show the advantages and disadvantages of EM clustering.

Advantages	**Disadvantages**
Provides the best solution for a real data set.	The algorithm is very complex.
This algorithm is used when the amount of data is small or the available region of interest is poor.	Can be slow to converge:

Table.-4. Advatages and disadvantages of EM clustering process.

2.5 Canopy

Developed by Andrew McCallum, Kamal Nigam, and Lyle Unger in 2000, the Canopy clustering algorithm, serves as a valuable preprocessing technique in the realm of clustering. Its primary role is to enhance the efficiency of subsequent clustering algorithms, notably K-Means and Hierarchical clustering. When using these clustering algorithms directly on large datasets would be impossible due to the amount of the data, this solution fills the gap.

In essence, Canopy clustering acts as a preliminary filter, helping to cull and organize data points before applying more resource-intensive clustering techniques. It provides a means to quickly identify and group data points that exhibit certain similarities, allowing subsequent algorithms to focus on a reduced subset of the data, which can significantly improve computational efficiency. This makes Canopy clustering a valuable tool when dealing with large datasets, as it streamlines the clustering process and makes it more feasible to work with extensive and complex data collections. set.[1] IEEE.

Table 5 show the advantages and disadvantages of Canopy clustering.

Advantages	Disadvantages
The Canopy algorithm is a very fast clustering algorithm, especially for large datasets.	The performance of the Canopy algorithm is sensitive to the canopy threshold, which needs to be tuned carefully.
The Canopy algorithm is scalable to very large datasets, as it only requires pairwise distance calculations between data points and canopy centers.	The Canopy algorithm can produce overlapping canopies, which can lead to ambiguity in cluster assignments.

Table.-5. Advatages and disadvantages of Canopy clustering process.

II. EXPERIMENTS AND RESULTS

3.1 Data set Description

The information gathered for this study is Social Network ads. Its obtain from online kaggle datasets site to do an experiment and evaluate the effectiveness of clustering techniques. The dataset have five number of attributes , 400 number of instances, 0 % missing value and Numeric category. The dataset chosen for the experiment has a variety of traits and can be used in a variety of contexts.

3.2 Result and Discussion

Various Clustering techniques discussed in section 3 have been compared with the help of WEKA 3.8.6 Toolkit. Steps followed in the analysis are:

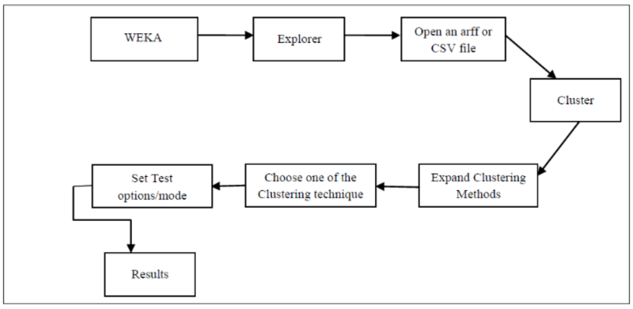

Fig. 3. Clustering process used in WEKA 3.8.6 Tool

For the purpose of data mining, this tool comprises of a number of machine learning algorithms. It is free, open-source Java program that provides excellent user interfaces for a wide range of activities, such as pre-processing, classification, clustering, etc. Five clustering techniques, including k-means, density-based cluster, hierarchical, EM, and canopy, are used in the course of the study.

Additionally, datasets for Social Networking Ads were selected for an experiment. Performance analysis takes into account factors including the number of clustered nodes, cluster instances, iterations, within-cluster sum of squared errors, and modeling time.

The outcome of the experiment using the dataset and other parameters is shown in table 6 below.

Algorithm Name	Total Number of clusters	Detail of Cluster Instances	Perform Number of Iterations	Within clusters sum of squared errors	To Build model how much time require.

K-Means Algorithm	3	204(51 %)
		66(17 %)
		130(33 %)
			6	125.5384	0.01 Sec

Density Based Cluster	3	204(51 %)	6	125.5384	0.02 Sec

		67(17%)
		129(32%)			

Hierarchical Clustering	3	257(64%)
		77(19%)
		66(17%)
			-	-	0.23 Sec

EM Algorithm	3	18(5%)
		257(64%)
		125(31%)
			2	-	0.05 Sec

Canopy Algorithm	3	127(32%)
		148(37%)
		125(31%)
			-	-	0.02 Sec

Table.-6. Comparison & Results using Weka Clustering Tool

According to the findings, the k-means clustering algorithm performs well with accuracy compared to other algorithms and takes less time to generate the model. The make density-based algorithm also performs well, but it takes a long time to build the model. Additionally, it has been noted that the EM clustering technique and hierarchical clustering both took longer and produced less accurate results. Although the canopy method produced less accurate results, less time was required to generate the model. The conclusion drawn from the results is that all three types of datasets are successfully clustered using the k-means approach.

III. CONCLUSION

This paper's main objective is to give a thorough overview of various clustering techniques. The many clustering approaches that are accessible in data mining are therefore briefly covered in this study. This study compares the effectiveness of various clustering algorithms using various parameters, including the density-based clustering algorithm and the k-means clustering algorithm, both of which perform well on words in all dimensions. This finding suggests that our algorithm should be improved in order to achieve better results.

IV. FUTURE WORK

In this paper, we aim to compare the six algorithms described earlier, and we have given some results above. However, we cannot cover all the factors for comparing these five algorithms.

Future work could compare these algorithms (or other algorithms) using different parameters than those proposed in this paper. Another important thing is normalization. Comparing the results of algorithms using normal data and unaffected data gives different results. Of course, normalization affects the performance of the algorithm and the quality of the results.

V. REFERENCES

[1]. D Swasti Singhal," A Study on WEKA Tool for Data Preprocessing, Classification and Clustering", International Journal of Innovative Technology and Exploring Engineering (IJITEE) ,Vol. 2(6), May 2013.

[2]. Team, E. "What is machine learning? a definition." online].(https://www. expertsystem. com/machine-learning-definition/ (2019).

[3]. Fung, Glenn. "A comprehensive overview of basic clustering algorithms." (2001): 1-37.

[4]. Namratha, M., and T. R. Prajwala. "A comprehensive overview of clustering algorithms in pattern recognition." IOSR Journal of Computer Engineering 4.6 (2012): 23-30.

[5]. N Patel, Meghna, Shital Patel, and Sonal Patel. "Data Analysis in Shopping Mall data using K-Means Clustering." 2022 4th International Conference on Advances in Computing,

Communication Control and Networking (ICAC3N). IEEE, 2022.

[6]. Devi, R. Delshi Howsalya, and P. Deepika. "Performance comparison of various clustering techniques for diagnosis of breast cancer." 2015 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC). IEEE, 2015.

[7]. Chitra, K., and D. Maheswari. "A comparative study of various clustering algorithms in data mining." International Journal of Computer Science and Mobile Computing 6.8 (2017): 109-115.

[8]. Kameshwaran, K., and K. Malarvizhi. "Survey on clustering techniques in data mining." International Journal of Computer Science and Information Technologies 5.2 (2014): 2272-2276.

[9]. Yadav, Krishna Mohan, et al. "Clustering Techniques and Algorithms of Data Mining–A Review."

[10]. Surya Narayana, G., and D. Vasumathi. "An attributes similarity-based K-medoids clustering technique in data mining." Arabian Journal for Science and Engineering 43.8 (2018): 3979-3992.

[11]. Wegmann, Marc, et al. "A review of systematic selection of clustering algorithms and their evaluation." arXiv preprint arXiv:2106.12792 (2021).

[12]. Popat, S. K., & Emmanuel, M. (2014). Review and comparative study of clustering techniques. International journal of computer science and information technologies, 5(1), 805-812.