A COMPARATIVE ANALYSIS OF CLUSTERING ALGORITHMS IN DATA MINING

S.SANDHIYA
Ph.D Research Scholar
Department of Computer Science
Bharathiar University
Coimbatore – 641046
sandhiyasrcw@gmail.com

Dr.R.PORKODI
Associate Professor
Department of Computer Science
Bharathiar University
Coimbatore – 641046
porkodi_r76@buc.edu.in

Abstract— **The valuable information is extracted to identify patterns and trends from the large dataset known as data mining. Clustering is unsupervised data mining technique which groups the similar data under a single cluster. Identifying the similar objects plays major role in the study of disease mechanism. In this research work two clustering algorithms are evaluated with the gene expression dataset HG-U133A.22283 gene expressions for 24 samples were used for the experimental purpose. The dataset used in this research work is clustered into 3 clusters and average Silhouette Width is used to evaluate the performance of the algorithms used in this research. Based on the evaluation metrics k-means clustering proves it efficiency over gene expression data.**

Keywords— *Data Mining, Clustering, k-means, k-medoids, gene expression***.**

I. INTRODUCTION

This survey This survey aims to present a thorough analysis of several clustering techniques in data mining. Data is divided into groups of related objects using clustering. Each group, or cluster, is made up of things that are dissimilar from those in other groups yet similar to one another. While simplicity is achieved, fewer clusters must inevitably lose some fine details (similar to lossy data compression). It models data by its clusters and represents many data objects by a small number of clusters. Clustering is seen historically through data modeling, which is based on mathematics, statistics, and numerical analysis.[9] Clusters are hidden patterns from a machine learning perspective, finding clusters is unsupervised learning, and the resulting system is a data notion.

In several domains, including machine learning, data mining, pattern recognition, image analysis, and bioinformatics, clustering is a widely used statistical data processing tool. Other words with comparable connotations to clustering include automatic categorization, numerical taxonomy, botryology, and typological analysis. Clustering is hence the unsupervised learning of a hidden data idea[4].

The valuable information is extracted to identify patterns and trends from the large dataset known as data mining. It is one of the most useful techniques that help entrepreneurs, researchers, and individuals to extract data.It is an associative skill that uses machine learning, statistics, and AI to excerpt information upon assessing future events.

Machine learning is automation that enables computers to learn from past data automatically. It uses algorithms to build mathematical models and make predictions from historical data.[9]

Statistics uses representations, models, and mathematical analysis to summarize empirical data or real-world observations

The concept of artificial intelligence relies more on examining than on statistics. It tries to apply human thinking like processing to statistical problems. Several high-end commercial products as query optimization modules for relational database management systems (RDBMS) have approved specific AI concepts.The insights gained from data mining are used for marketing, fraud detection, scientific discovery, etc. It's also called Knowledge Discovery in Database[4]. KDD is a formulated process to allow valid, useful, and legible patterns from large and complex data sets. The KDD process entitles Data cleaning, Data Integration, Data Selection, Data transformation, Data mining, Pattern evaluation, and knowledge presentation.

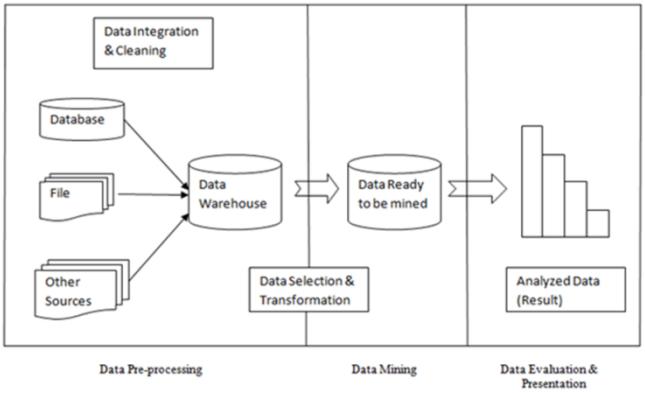

Fig.1 KDD Process

Figure 1 illustrates the Knowledge Discovery Process. The process begins with finding the KDD objectives and ends with the accomplishment of the discovered knowledge.

Data Cleaning is the process of removing the noise and random data within a dataset. It may involve complex statistical techniques or the use of a data mining algorithm in these factors. This eliminates major errors and inconsistencies that are unavoidable when multiple data sources are pulled into a dataset. It is used to clean withData discrepancy detection and Data transformation tools.

Data integration is the process of combining data from multiple heterogeneous sources. It deals with many issues like data redundancy, randomness, spuriousness, and many more. This gives a consistent view of scattered data while maintaining data accuracy and data integration using Data Migration tools and Data Synchronization tools [9].

Data selection is defined as the process of determining the appropriate data type and source and the appropriate tools for collecting the data. Data selection precedes the actual practice of data collection. In that the process where data relevant to the analysis is decided and retrieved from the data collection. Data Selection using neural network,Decision Trees, Naive Bayes, Clustering, Regression

Data transformation is a technique used to transform raw data into a suitable format that effectively facilitates data mining and retrieves strategic information. Data transformation includes data cleaning techniques and data reduction techniques to convert the data into a suitable format. It is a two-step process is Data mapping and code generation

Data mining is the process of sorting through large data sets to identify patterns and relationships that help solve business problems through data analysis.[4] Data mining techniques and tools help companies predict future trends and make more informed business decisions. It transforms task-relevant data into patterns.

Model evaluation is defined as identifying strictly incremental patterns that represent knowledge based on given measures. It is responsible for the quantification of the investigation of the pattern using a threshold value. It collaborates with the data mining engine to focus the search on interesting patterns.Knowledge representation is defined as the technique of using visualization tools to represent data mining results. Knowledge representation is the provision of knowledge to the user for visualization in the form of trees, tables, rule diagrams, charts, matrices, etc.

In data mining, several techniques are used to solve different problems, and a method will be selected according to the problem being solved. The process can be divided into two basic parts, which are referred to as predictive data mining and descriptive data mining, respectively. Predictive data mining analysis as the name implies data that assists in predicting what will happen next (or in the future) in the business.[4] Predictive Data Mining can also be additionally split into four categories, namely Classification, Regression, Time Series, and Prediction. The primary goal of descriptive data mining tasks is to outline or transform given data into relevant information. There are four types of descriptive data mining tasks. These are as follows: Clustering, Summarization, Association Rules, and Sequence Discovery

II. CLUSTERING : AN OVERVIEW

Clusters are groups of objects belonging to the same class. A cluster consists of objects with similar characteristics, while a cluster consists of objects with dissimilar characteristics. Clustering helps divide the data into multiple subsets. Each of these subsets contains similar data to the other subsets, and these subsets are called clusters.[1] A clustering algorithm is an unsupervised, machine learning-based algorithm that consists of grouping data points into clusters based on their similarity. When performing cluster analysis, first divide the data set into groups form on data resemblance and then allocate labels to the groups.

Clustering analysis is frequently utilized in a variety of fields, including data analysis, market research, pattern identification, and image processing. Clustering can help marketers identify unique groups within their customer base. And they can segment their customer groups based on their buying patterns. In the field of biology, it is used to take plant and animal taxonomies, classify genes with similar functions, and gain insight into the intrinsic structures of populations. [2] Clustering also helps identify similar land use areas in a geospatial database. It also helps to identify groups of houses in a city based on house type, value, and geographic location. It also helps categorize documents on the Internet for information discovery and applications for outlier detection, such as those used to find credit card fraud. Cluster analysis provides a tool to accomplish intelligence in data distribution by observing the characteristics of each cluster.

Clustering itself can be classified into two types. Hard clustering and soft clustering. In hard clustering, a data point belongs to only one cluster. But in soft clustering, the output delivered is the possibility of a data point belonging to each of several predefined clusters. Clustering analysis is frequently utilized in a variety of fields, including data analysis, market research, pattern identification, and image processing. There are many clustering methods like partition clustering, hierarchical clustering, density-based clustering, distribution model clustering, and fuzzy clustering.

III. TECHNIQUES FOR CLUSTERING ALGORITHM

A. PARTITITONING BASED CLUSTERING

Partitioning objects into k clusters, where each partition structure/serves a cluster, and these clusters hold some properties, such as each cluster carrying at least one data object and each data object classified as exactly one cluster. Based on centroids and data points, a cluster is assigned based on its nearness to the cluster centroid.[10] These techniques are generally categorized for optimizing the benchmark similarity function, making distance an important factor. The partitioning-based clustering algorithm consists of k-means, k-medians, and k-modes. Partitioning clustering algorithms are a type of non-hierarchical that often handle static sets to investigate the groups displayed in the data using techniques for the objective function optimization and improve the quality of the partition frequently. Partitioning-based clustering computes all achievable clusters synchronously and is very effective in terms of simplicity, proficiency, and ease of deployment.[10] The disadvantage of partitioning clustering is that you have to predetermine the number of centroids also, the clusters that are produced have erratic densities and sizes noise- and outlier-affected. The resulting clusters have the following properties: Each cluster must include precisely one stone object and there may not be any overlap between any two clusters.

Partitioning techniques are divided into two types namely Medoids Algorithms, and Centroid Algorithms: Each cluster contains the instances that are closest to the gravity center called medoids algorithms. Centroid Algorithms are

the gravity center of the instances used to represent each cluster. Example: K –Means clustering technique, where the data set is partitioned into k subsets in such a manner that all points in a given subset are closest to the same gravity center. The effectiveness of the k-means technique depends on the objective function that is being used to calculate the distance between the instances. k- means technique has a requirement that all the data must be available prior.

$$\arg\min_{\mathbf{S}} \sum_{i=1}^{k} \sum_{\mathbf{x}_j \in S_i} \|\mathbf{x}_j - \boldsymbol{\mu}_i\|^2$$

Time Complexity:
Space Complexity: O(k+n)
The steps for the K Means Clustering Method are as follows:

Step 1: Choose K random points as cluster centers called centroids.

Step 2: Assign each x(i) to the closest cluster by implementing euclidean distance

Step 3: Identify new centroids by taking the average of the assigned points.

Step 4: Keep repeating step 2 and step 3 until convergence is achieved

B. HIERARCHICAL-BASED CLUSTERING

These clustering techniques construct a cluster with a tree-like structure depending on the hierarchy, where each newly formed cluster is created utilizing previously established clusters. To generate clusters, use the data points top-to-bottom hierarchical order. It was divided into the Agglomerative (bottom-up method) and Divisive categories (top-down approach). The agglomerative clustering approach locates each point in a cluster by first combining the two points that are closest to it, where each point represents a single object or a group of objects. In divisive clustering, the entire population is first viewed as one cluster before being divided into smaller groupings [12]. The hierarchical-based clustering algorithm consists of the Agglomerative Hierarchical clustering algorithm (AGNES), and Divisive Hierarchical clustering algorithm (DIANA). The benefits of this clustering of dendrograms are simple to comprehend, and the number of clusters need not be predetermined. The negative effects of this grouping Cluster assignment are rigid and cannot be changed. It is also time-consuming and ineffective for larger datasets.There are two sorts of hierarchical techniques namely the Agglomerative Method and the Divisive Method

In the agglomerative Method, the clusters all data instances until they are part of the same cluster. The closest cluster pair is combined. The three types of closeness are single-link, full-link, and average-link.Single-link closeness is the resemblance between two instances that are in different clusters, according to its definition. Although it handles clusters of elliptical forms effectively, it is sensitive to noise and errors.Complete-link closeness is the similarity between the most typical occurrences, each of which belongs to a different cluster. Although it is less sensitive to noise, mistakes, or outliers, it is not ideal for convex-shaped clusters.

- Divisive Method: It divides data sets into smaller clusters until each cluster contains just one data instance. It operates top-down.

Time Complexity: O (m* time taken to find points in the neighborhood)
Space Complexity: O (m)
The steps to solve the Hierarchical Clustering Method are as follows:

Step 1: Create each data point as a single cluste

Step 2: Take two closest data points or clusters and merge them to form one cluster

Step 3: Take the two closest clusters and merge them together to form one cluster

Step 4: Repeat Step 3 until only one cluster left.

Step 5: After all the clusters are combined into one big cluster, develop the dendrogram to divide the clusters as per the problem.

dense regions that have some characteristics and are different from the space's low-density parts. These techniques can integrate two clusters with a high degree of precision. Using the data point density, also referred to as model-based clustering these techniques use distance measurements between the objects to group them. Since clusters created using this method are typically spherical, it can be challenging to distinguish between arbitrary-shaped clusters.[6] The benefits of this clustering can tolerate noise and outliers, has no initial cluster number specification, produces very homogeneous clusters, and has no limits on cluster shape. This clustering's disadvantage is that its slow, complicated method cannot handle bigger amounts of data.

The DBSCAN (Density-Based Spatial Clustering of Applications with Noise) and OPTICS algorithms are used in the hierarchical clustering algorithm (Ordering Points to Identify Clustering Structure). The data points in DBSCAN are either categorized as noise or belong to a cluster.

The three types of data points are noise, border, and core points. The points that make up the cluster's interior are referred to as core points. If there are more than a specific number of data points nearby, a point is regarded as being inside the cluster.[16] Border points are locations that are close to core points but are not the core points themselves. A point that is neither a core point nor a border point is referred to as a noise point.
Time Complexity: O(m* time taken to find points in the neighborhood)
Space Complexity: O(m)
Density-based spatial clustering of applications with noise, or DBSCAN, in data mining. The steps for the DBSCAN Clustering Method are as follows:

Step1:Arbitrary select a point PArbitrary select a point P.

Step 2. Retrieve all point density reachable from P wrt ε and MinPts.

Step 3.If P is a core point a cluster is formed.

Step 4. If P is a border point, then there is no point that is density-reachable and DBSCAN moves to the next point.

Step 5. This process is continued until all the points are processed.

These techniques use a predetermined mathematical model to fit the data and then optimize it, assuming that the data is a hybrid of probability distributions, and then compute the number of clusters using conventional statistics. This cluster is derived using different metrics, such

as mean, variance, etc., depending on the data's probability distribution. When generating the standard statistics for robust clustering, noise and outliers are taken into consideration. These clustering techniques are divided into the statistical approach and the neural network approach to create clusters [20]. This clustering has the advantage of working with real-time data, not requiring apriori cluster number specifications, and metrics that are simple to comprehend and tune. Grid-based clustering's drawback is that its intricate process is slow and unable to handle bigger amounts of data. Two separate algorithms, Gaussian Mixed Models and DBCLASD are utilized in this clustering.
Fundamental Grid-based Algorithm

Step 1:Establish a group of grid cells.
Step 2:Assign the right grid cell to an object and calculate the density of that cell.
Step 3: Remove cells with densities below a predetermined threshold (t).
Step 4: Create clusters from adjacent, contiguous sets of dense cells

E. FUZZY BASED CLUSTERING

The partitioning approach serves as the foundation for fuzzy clustering, however, data points can belong to several clusters.[13] It falls within the category of soft method clustering techniques, as opposed to hard method clustering approaches, which include all of the techniques discussed above. According to this clustering method, points near the center may be a component of the other cluster to a greater extent than points towards the periphery of the same cluster. This clustering has advantages using substantially overlapping data and a higher rate of convergence. The drawback of this clustering is that the number of centroids must be specified a priori, Noise and outliers have an impact, are Incapable of scaling, and slow algorithms. Utilizing the algorithm Uncertain C Means and rough k-means

Step 1: Randomly group the data points into the required number of clusters.
Step 2: Determine the centroid.
Step 3: Calculate each point's distance from the centroid.
Step 4: Updating membership values.
Step 5: Repeat steps (2-4) until the membership numbers are constant or the difference is less than the tolerance value
Step 6: Defuzzify the membership values obtained.

IV. LITERATURE SURVEY

A. Ghosal et.al.[1] a category's proportion of DBCLAs for each class was examined. The existence of multi-class algorithms demonstrates how DBCLAs have evolved by utilizing a variety of strategies to overcome difficulties in high-dimensional clustering. Out of all the algorithms covered in this survey, it offers an empirical analysis of a few of the representative DBCLAs. It demonstrated meaningful DBCLA behavior that was supported by experimental data. Finding the conceptual dependency and citation percentage chart of DBCLAs was done to identify the group of algorithms that have significantly impacted the density-based clustering field.

P. Felcy Judith et.al.,[2] Describe all the approaches for clustering's implementation and its methodologies. Additionally, a comprehensive examination of clustering and research of various methodologies' comparisons are involved. However, compared to other methods, ANFIS and FCM methods can be applied in real-world circumstances. Various applications use different methods depending on the level of tolerance required.

MohiuddinAhmedet.al.,[3] Data analysis activities heavily rely on clustering across a wide variety of application disciplines. The popular k-means approach was highlighted, along with startup problems and the system's inability to handle data with mixed feature types. To illustrate how various k-means versions perform, this work includes both a critical examination of the body of current research and an experimental investigation of six benchmark datasets. The results of the experimental research revealed that the k-means algorithm's issues cannot be solved universally; rather, each of the variants of the algorithm now in use is either application- or data-specific.

Dafir, Z.et.al.,[4] offered a thorough analysis of the most recent parallel clustering methods arranged by the Big Data platforms utilized. Platforms that can manage massive data processing fall into two basic groups. The clustering techniques based on MapReduce, Spark, and peer-to-peer networks were discussed in the first category. The horizontal scaling platforms include these platforms. The clustering algorithms developed with multi-core CPUs, GPUs, and FPGA are the emphasis of the second group, which is referred to as the vertical scaling platforms. Each algorithm under consideration was examined in light of the methods used to guarantee parallelism. This paper also contains a thorough comparison of the clustering methods that have been described based on certain standard benchmarks for validating clustering outcomes in the Big Data context.

Panthadeep BHATTACHARJEE et.al.,[5] Due to the vast amounts of data kept in various areas, clustering analysis has become essential for identifying patterns in enormous amounts of data, facilitating the process of meaningful knowledge discovery. Examining the many clustering techniques and the applications they have in various industries. Some methods, such as k-means and DBSCAN, have been extensively studied and used, but others, such as the Sparse Subspace Clustering Algorithm, are still in their infancy and are the subject of intensive research. The fact that no particular clustering technique has been discovered to predominate in all implementation areas is remarkable.

Shapol M. Mohammed et.al.,[6] to understand which density-based algorithm types are most frequently utilized for GloVe word embedding-specific document clustering. Additionally, to understand what metrics and similarity measuresGloVe word embedding and density-

based methods are employed. Very few density-based algorithms have employed GloVe word embedding. DBSCAN and DPC are the two types of density-based algorithms that were most frequently utilized in our survey. Cosine similarity and F-measure are the two metrics most frequently employed in our study for comparing two metrics and assessing the effectiveness and precision of the two density-based methods.

CanAtilganaet.al.,[7] The framework for density-based incremental clustering, which uses fuzzy clustering for primary clustering, is presented in this research. The system comprises two innovative algorithms: MVSA for adaptively finding the outliers thresholds and selecting the final clusters, and FLCA for allocating incoming points to micro-clusters. Both conventional and streaming datasets have been used to test the performance of the proposed system.

Hongjing Zhang et.al.,[8] Simple pairwise together and apart constraints often constitute the majority of constraints for constrained partitional clustering. In this work, we show that deep clustering may be extended to a wide range of fundamentally distinct constraint types, such as instance-level (specifying hardness), cluster-level (specifying cluster sizes), and triplet-level. The framework is capable of handling both new constraints produced by an ontology graph and standard constraints produced from labeled side information. First of all, it outperforms well-known k-means, mixture models, and spectral restricted clustering in terms of experimental performance in both academic and real-world settings. By learning a representation that satisfies the constraints and discovers a good clustering, that technique avoids the detrimental impacts of constraints.

Abdullahi[9] a thorough illustration of data mining methods. Big data refers to enormous quantities of intricate data sets. The extraction of helpful rules or intriguing patterns from historical data is known as data mining. To use some of the techniques listed in this paper, such as clustering, decision, tree prediction, and neural networks, high-performance computation is necessary.

Swarndeep Sake et.al.,[10]In this study, four significant partitioning algorithms—k-means, k-medoids, CLARA, and CLARANS—are analyzed. The study provides a comparative table to help readers understand the benefits and drawbacks of each method. According to the investigation, CLARA and CLARANS are comparably more effective and scalable than other algorithms. However, it is possible to further alter algorithms like k-means and k-medoid so that they are equally effective and scalable. To examine the efficiency parameters of each partitioning algorithm, more investigation is necessary.

V. RESULT &DISCUSSION

For this study, two clustering algorithms namely-means and k-medoids are used. According to the literature, k-medoids clustering and k-means clustering produce better prediction outcomes when compared to other techniques. The dataset HG-U133A is collected from NCBI database and it is experimented with k-means and k-medoids clustering algorithms. The HG-U133A set includes 2 arrays with a total of 44928 entries. Out of 44928 entries 22283 records are taken for further studies.the optimal number of clustering has been chosen as 3 based on the Silhouette and Elbow method. Figure2 represent the optimal number of cluster obtained using Elbow method. The silhouette coefficient has a value between [-1, 1]. When a data point receives a score of 1, it means that its cluster is the largest and that it is the furthest apart from all other clusters. The worst number is one. Values close to 0 indicate clusters that overlap.

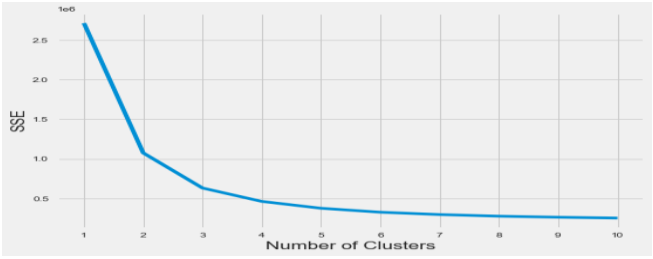

Figure 2: Graphical representation of Elbow method

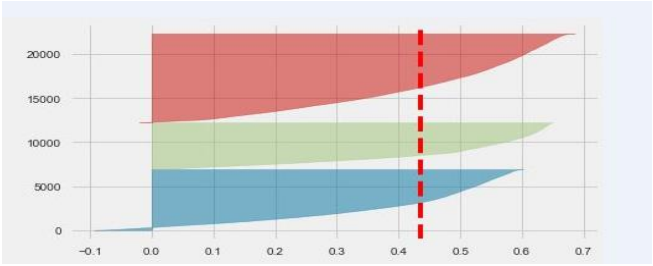

Figure 3: Graphical representation of Average Silhouette Width plot

k-medoids algorithm groupred6980 genes for cluster 1 for Cluster 2 & Cluster 3 k-medoids algorithm grouped 8902 genes and 6368. Each cluster is characterized by human genes that can be used to study about the disease mechanism as depth as possible. The k-medoids algorithm successfully produced three clusters and the number of genes in each cluster are represented in table 1 and figure 4 shows the graphical representation of the same.

Table 1: Clusters by k-medoids

Algorithm	Cluster 1	Cluster 2	Cluster 3
k-medoids	6980	8902	6367

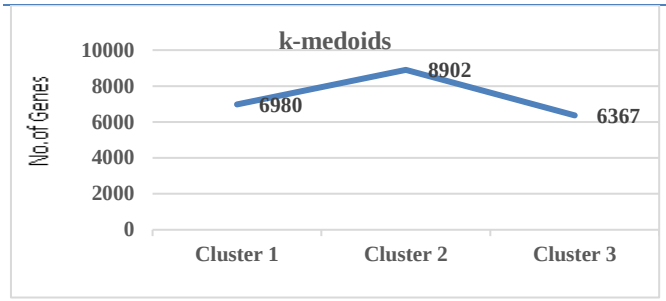

Figure 4: Graphical representation by k-medoids

The k-means clustering has produced three clusters. Cluster 1 grouped 5302 genes Similarly, Cluster 2 and Cluster 3 grouped 6905 genes and 10076 genes. The k-medoids algorithm successfully produced three clusters and the number of genes in each cluster is represented in Table 2 and Figure 5 shows the graphical representation as the same.

Table 2: Clusters by k-means

Algorithm	Cluster 1	Cluster 2	Cluster 3
k-means	5302	6905	10076

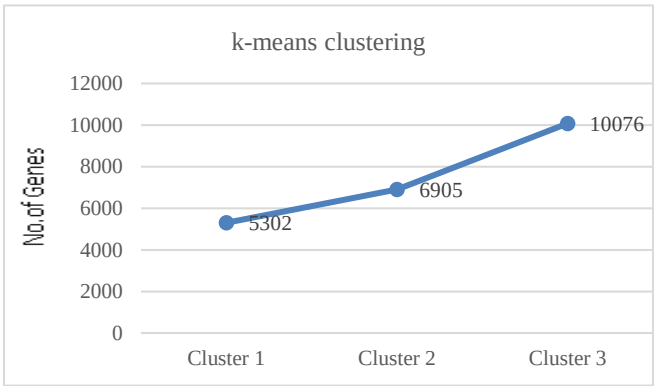

Figure 5: Graphical representation by k-means

Accuracy of the clustering algorithm is used in this research are validated with average Silhouette Width and table 3 represent the Average Silhouette Width value achieved by the 2 algorithms used in this research work. From the table it is inferred that k-means algorithm has the highest average Silhouette Width while comparing with k-medoids algorithm.

Table 3: Average Silhouette Width

Method	k-means	k-mediod
Average	0.267	0.183

VI. Conclusion

Clustering is one of the most used techniques in data mining. Clustering separates the group based on the similarities among the entities. In this study comparative analysis of two clustering algorithms namely k-means and k-medoids are analyzed using HG-U133Adataset. From the analysis k-means algorithm outperformed k-medoids algorithm based on the average Silhouette Width value. In future different clustering algorithms can be applied to analyze the efficiency of clustering algorithms over gene expression data set.

References

[1]. AttriGhosal, ArunimaNandy, Amit Kumar Das, SaptarsiGoswami, and MrityunjoyPanday,A Short Review on Different Clustering Techniques and Their Applications ,Emerging Technology in Modelling and Graphics, Advances in Intelligent Systems and Computing 937, https://doi.org/10.1007/978-981-13-7403-6_9

[2]. M.JayaprabhaDr.P. FelcyJudith,A Review of Clustering, its Types, and Techniques International Journal of Innovative Science and Research Technology ISSN No:-2456-2165 Volume 3, Issue 6, June – 2018

[3]. Mohiuddin Ahmed, Raihan SerajThe k-means Algorithm: A Comprehensive Survey and PerformanceEvaluation Electronics 2020, 9(8), 1295; https://doi.org/10.3390/electronics9081295

[4]. Dafir, Z., Lamari, Y. &Slaoui, S.C. A survey on parallel clustering algorithms for Big Data. ArtifIntell Rev 54, 2411–2443 (2021). https://doi.org/10.1007/s10462-020-09918-2

[5]. Shapol M. Mohammed1, Karwan Jacksi2, Subhi R. M. ZeebareeIndonesian,A state-of-the-art survey on semantic similarity for document clustering using GloVe and density-based algorithms, Journal of Electrical Engineering and Computer Science Vol. 22, No. 1, April 2021, pp. 552~562 ISSN: 2502-4752, DOI: 10.11591/ijeecs.v22.i1.pp552-562

[6]. PanthadeepBhattacharjee, PinakiMitra,A survey of density-based clustering algorithms Department of Computer Science and Engineering, Indian Institute of Technology, Guwahati 781039, India, Front. Comput. Sci., 2021, 15(1): 151308 https://doi.org/10.1007/s11704-019-9059-3

[7]. CanAtilgana, BarisTekinTezel, EfendiNasiboglu,Efficient implementation and parallelization of fuzzy density-based clustering, Information SciencesVolume 575, October 2021, Pages 454-467

[8]. Hongjing Zhang, Tianyang Zhan, SugatoBasu&Ian Davidson,A framework for deep constrained clustering, Data Mining and Knowledge Discovery 593–620 (2021)

[9]. AbdullahiSidowOsman,Data Mining Techniques: Review, IJDSR, Volume 2, Issue 1 June2019, Al-Madinah International Universit

[10]. SwarndeepSaket J, Dr. Sharnil,An Overview of Partitioning Algorithms in Clustering Techniques, Pandyainternational Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 5, Issue 6, June 2016, ISSN: 2278 – 1323

[11]. Faizan, Shahrinaz-Ismail,Applications of Clustering Techniques in Data Mining: A Comparative Study Muhammad-December 2020 International Journal of Advanced Computer Science and Applications 11(12)DOI:10.14569/IJACSA.2020.0111218

[12]. Pranav Shetty and SurajSingh,Hierarchical Clustering: A Survey ResearchInternational Journal of Applied DOI: https://doi.org/10.22271/allresearch.2021.v7.i4c.8484 2021; 7(4): 178-181

[13]. Jiamin Li; Harold W. Lewis,Fuzzy Clustering Algorithms — Review of the Applications https://doi.org/10.1109/SmartCloud.2016.14 IEEE Xplore

[14]. Dhara Patel, KetanSarvakar, A Comparative Study of Clustering Data Mining: Techniques and Research Challenges, Volume III, Issue IX, September 2014 IJLTEMAS ISSN 2278 – 2540

[15]. Pradeep Kumar Singh, Clustering Techniques in Data Mining: A Comparison 2nd International ConferenceonComputing for Sustainable Global Development (INDIACom978-9-3805-4416-8/15/$31.00

[16]. M. Ester, H. Kriegel, J. Sander, X. Xu, A Density- Based Algorithm for discovering clusters in large spatial databases with noise, KDD-96 Proceedings, pp. 226-231

[17]. Hilles,S. M. (2018, July). Sofm And Vector Quantization For Image Compression By Component.In2018 International Conferencon Smart Computing and Electronic Enterprise (ICSCEE)(pp. 1-6). IEEE

[18]. 18.Altrad,A.M., Amphwan,A.&Hilles S. M. Adaptive Shuffled Frog Leaping Algorithm For Optimal Power Rate Allocation: Power Line.In 2018 International Conference on Smart Computing and Electronic Enterprise (ICSCEE) (pp. 1-5). IEEE.

[19]. EDUCBA, "Data Mining Techniques for Successful Business (Tools, Software)," Educba.com [Online].Available: https://www.educba.com/data-mining-techniques/. [Accessed 3 December 2018]

[20]. M.Brown,"Dataminingtechniques,"[Online].Available:https://www.ibm.com/developerworks/library/ba-data-mining-techniques/index.html. [Accessed 3 December 2018]