International Journal of Computer Trends and Technology
ISSN: 2231 – 2803 / https://doi.org/10.14445/22312803/IJCTT-V70I3P105

Volume 70 Issue 3, 27-33, March 2022

Review Article

Comparative Analysis of Clustering Approaches for Big Data Analysis

Satish S. Banait [1], Shrish S. Sane [2]

Department of Computer Engineering, K.K. Wagh Institue of Engineering Education & Research, Nashik, Savitribai Phule Pune University, Pune-Maharashtra.

Received Date: 16 February 2022
Revised Date: 03 April 2022
Accepted Date: 06 April 2022

Abstract *- This paper* performs *a comparative study of the most popular big data clustering techniques. Clustering is an unsupervised classification of patterns (observations, data items or feature vectors) into teams (clusters). The drawbacks of clustering have been noticed in several contexts by researchers in many disciplines and react to its broad charm and quality in concert with the steps in exploratory data analysis. K-means clustering algorithm falls underneath the category of centroid-based clustering. Hierarchical clustering is a cluster analysis technique that seeks to construct a hierarchy of clusters. Agglomerative clustering is a form of hierarchical clustering that uses the backside-up technique. Density-based Spatial Clustering of Algorithms with Noise (DBSCAN) is a clustering algorithm that organisations collectively point near every other primarily based on a distance dimension (Euclidean distance) and a minimal quantity of factors. Map-reduce is a programming paradigm for huge datasets which may be processed speedily by processing them on distributed clusters in parallel. This paper compares k-means, hierarchical agglomerative clustering, DBSCAN and k-means with map-reduce strategies for clustering big data.*

Keywords *- Big Data, Clustering Strategies, Density-Based Spatial Clustering, Hierarchical Agglomerative Clustering, K-Means.*

I. INTRODUCTION

The traditional k-means method set of rules may be very sensitive to selecting clustering centres and calculating distances, so the algorithm without problems converges to a domestically premier answer. In addition, the conventional algorithm has a slow convergence speed and low clustering accuracy and memory bottleneck problems when processing huge statistics. Therefore, a stepped forward k-means set of rules is proposed [1]. In this algorithm, the choice of the initial factors inside the conventional clustering algorithm is progressed first, and then a brand new worldwide measure, the powerful distance degree, is proposed. Its fundamental concept is to calculate the powerful distance among information samples employing sparse reconstruction. Sooner or later, on the MapReduce framework's idea, the algorithm's efficiency is similarly stepped forward by adjusting the Hadoop cluster. Based totally on the real purchaser records from the signs to evaluate the clustering effects of various algorithms. The consequences display that the proposed algorithm has exact convergence and accuracy and achieves higher performances than the ones of other compared algorithms.

The huge facts generation has brought about the fast improvement of machines gaining knowledge of technology. As one of the maxima generally used in traditional clustering algorithms, the k-means approach has been successfully carried out in many regions due to its simplicity, practicality, and efficiency. Consultant applications encompass report clustering, marketplace segmentation, image segmentation and characteristic gaining knowledge [2-5]. Generally, the k-means method includes 3 degrees: characteristic choice, feature extraction, and facts clustering based on the calculated similarities between records factors. Clustering aims to divide statistics into multiple instructions or clusters. The facts in the equal cluster possess high similarity, and the similarity between data in unique clusters is low [6]. Typically, clustering algorithms fall into categories: hierarchical clustering and partitional clustering. Hierarchical clustering algorithms construct an excessive-stage hierarchy of clusters referred to as a dendrogram consistent with the similarities among statistics factors. A dendrogram can be constructed by means of one of a kind techniques: agglomerative clustering (merging clusters backside-up) and divisive clustering (splitting clusters pinnacle-down). Then again, partitional clustering algorithms require predefining the number of clusters and the preliminary cluster centroids. These algorithms divide a dataset into more than one cluster without overlap with the aid of minimising a specific loss characteristic [7]. Seasoned-posed in 1967, k-means clustering is one of the maximum widely used clustering algorithms. It has been widely employed in various packages due to its simplicity and advanced overall performance compared to different clustering algorithms.

However, the k-means method has some barriers. First, the number of clusters k wishes to be predefined. Moreover, the initial cluster centroids of k-means are normally decided on randomly. Ultimately, the overall performance of the k-means method can be inspired by way of outliers inside the facts. To cope with the above troubles regarding k-means, researchers in extraordinary fields have proposed diverse stepped forward algorithms [8]. The k-means clustering set of rules is a difficult dynamic set based totally on the similarities among static statistics gadgets.

Compared with other clustering algorithms in phrases of complexity, the okay-way clustering algorithm is straightforward to put in force. It has low linear time complexity, so it is broadly utilised in statistics, technological know-how, industrial utility, and other fields. But, the k-means clustering algorithm also has some shortcomings. Its incapability to determine the proper wide variety of clusters, the excessive randomness of its clustering effects, and its great dependence on the selection of the preliminary clustering middle. The clustering results are significantly influenced by the initial clustering centres, which causes the clustering algorithm to fall into the local most reliable solution instead of the worldwide optimal solution. Moreover, pretreatment is just too steeply-priced in instances of big records analyses, and this impacts the general overall performance of the algorithm [9-11].

Because of the shortcomings and defects of k-means, many scholars have improved and optimised the k-means set of rules. These stepped forward algorithms are extensively utilised in specific fields. This additionally furnished a path for later upgrades to the conventional k-means algorithm. Primarily based on the connection between the clustering variety k and the sum of squared mistakes SSE, [12] selected the k fee similar to the elbow factor because the premier clustering quantity according to the variant trends of the SSE for specific k values. To clarify the problem of vague dots in the relation between k and the SSE, [13] decided on the most effective k cost by combining parameters, including the exponential function parameter, weight term, and bias term. For the trouble wherein the most efficient clustering number k wishes to be determined by way of manually analysing the choice graph, in combination with a statistical technique, [14] used linear regression to healthy the points inside the selection graph and decide the most reliable k value and the preliminary clustering middle in step with the differences among the observed values and the real values.

The density-based spatial clustering of applications with noise (DBSCAN) is a pioneering set of rules of the density-based clustering technique. It seasoned-vides the potential to handle outlier items, locate clusters of various shapes, and disregard the want for earlier information approximately current clusters in a dataset. Those features, along with its simplistic approach, helped it become broadly relevant in many regions of technological know-how. However, for all its accolades, the DBSCAN nevertheless has obstacles in performance phrases, it is potential to locate clusters of varying densities, and its dependence on consumer enter parameters. A couple of DBSCAN-inspired algorithms have been, in the end, proposed to relieve those and greater issues of the algorithm. In [15], the implementation, functions, strengths, and drawbacks of the DBSCAN are very well tested. The successive algorithms proposed to offer development at the unique DBSCAN are classified based on their motivations and are mentioned. Experimental exams have been carried out to understand and evaluate the changes supplied by means of a C++ implementation of these algorithms at the side of the unique DBSCAN algorithm. Finally, the analytical evaluation is provided primarily based on the effects determined.

The formal definitions of the clustering model and the DBSCAN rules were first introduced in 1996 at the understanding Discovery in Databases (KDD) data mining conference publication. The core idea behind the density-primarily based clustering approach assumes that a cluster is a region in data space with a high density of facts items. This clustering model brought specific features distinct from the sooner stated implementations. It added the capacity to form clusters of abnormal shapes, hit upon outliers within the data space, and pick out clusters without an earlier understanding of the lesson's gift in the dataset. It's miles a sensible algorithm, and the DBSCAN has been implemented in several fields of observation which include civil engineering, chemistry, spectroscopy, social sciences, clinical diagnostics, faraway sensing, pc imaginative and prescient, automated identification systems (AIS), and anomaly detection [16-20]. Its successful implementation on actual-world applications has led it to acquire the special hobby group on KDD test-of-time award. The DBSCAN has served as a stepping-stone to several different clustering algorithms aiming to improve on the unique algorithm and its obstacles.

Hierarchical clustering strategies seem to be an appealing way to discover anatomical subgroups from huge data as they may be inherently unsupervised therefore do now not require any earlier facts approximately the take a look at the populace and, in contrast to k-means clustering, do now not require specifying a predicted variety of subgroups [21]. Furthermore, clustering results may be graphically summarized in a dendrogram that depicts in a tree-like diagram how similar topics are grouped collectively whilst dissimilar topics are located on exceptional branches of the tree. But, evaluation of difficulty similarity or dissimilarity and clustering consequences closely depends on selecting each similarity or distance metric (with low inter-concern distance relating to higher similarity) and linkage function determining how subjects are connected collectively to shape a subgroup. Depending on the selected distance/linkage mixture, clustering outcomes can also vary

appreciably-doubtlessly, rendering meaningless effects [22]. Whilst previous studies have analysed clustering techniques based on universal shapes or 2-dimensional shape information [23], few have assessed hierarchical clustering overall performance in the use of actual huge information in a sensible setting. Density-based Distributed Clustering (DBDC) [24] is the most popular parallel density-based clustering algorithm; initially, the whole data set is divided and distributed between sites (machines). After that, each site runs the DBSCAN algorithm to form a local cluster and determine representatives data objects (i.e. clusters identifier) called a local model.

II. CLUSTERING STRATEGIES FOR BIG DATA

K-means clustering falls beneath the category of centroid-based clustering. A centroid is an information point in the cluster's centre and needs now not be a member of the dataset. This clustering algorithm is an iterative algorithm wherein the notion of similarity is derived by using how close a statistics factor is to the cluster's centroid. The steps concerned within the k-means method for huge records are: The input statistics are studied from the given massive statistics supply and processed as pair with the key being the gene_id and value being the attributes. The number of clusters is given as entering. The proper range of clusters for k-means may be located using the Elbow method. Then, k random centroids are first calculated, and the space of every statistics point from the centroids is calculated using Euclidean distance. The information point is assigned to its nearest cluster. Then the centroids of the brand new clusters are calculated, and this process runs iteratively.
Ultimately, the cluster project stops whilst the data factors are not reassigned to the unique cluster. The above steps are performed iteratively for all information points, and the cluster labels are retrieved in a listing. Outside Index calculation is finished using the floor truth statistics and cluster labels. The values of m11, m00, m01and m10 are updated by evaluating those floor reality labels and clustering labels. Jaccard coefficient and rand index are calculated in step with their formulation, and outcomes are displayed.
Jaccard coefficient: m11/(m11+m01+m10)
Rand Index: (m11+m00)/(m11+m00+m01+m10)

The authentic records are decreased to 2 main additives the usage of dimensionality reduction set of rules major component evaluation. This is performed with the use of the PCA module from sklearn. Decomposition library. Eventually, the result has visualised the use of a scatter plot.

Hierarchical clustering is a cluster analysis method that seeks to build a hierarchy of clusters. Agglomerative clustering is a kind of hierarchical clustering that uses the backside-Up method. In Hierarchical Agglomerative Clustering, each item starts to evolve in its cluster, and pairs of clusters, just like every different, are merged as one moves up the hierarchy until all items are merged right into a single cluster. The single link, one of the strategies to put in force hierarchical agglomerative clustering, is used right here. The space between two clusters is defined in a single linkage because of the minimal distance between any records factor inside the first cluster and any information point within the 2D cluster. At every degree, clusters with the smallest unmarried linkage distance are mixed.

The stairs worried in hierarchical agglomerative clustering of massive information are: The data is studied from the given big statistics supply and broken up into lists, one containing the information of the gene (all columns in the dataset beginning from column 3) on which the hierarchical clustering is done and different containing the floor truth facts (the second one column in the dataset). The variety of clusters into which the records should be partitioned was also given as entered. The distance matrix is calculated using the euclidean_distances module from sklearn. Metrics. Pairwise library.

Hierarchical agglomerative clustering has executed the usage of the following steps: Initially, every factor is stored as a single cluster. The two factors separated by a minimum distance that is more than zero are identified from the space matrix. The indices corresponding to this pair are merged into one row and column in the distance matrix. The gap matrix is up to date by re-computing the distance values for all other facts factors primarily based on their minimal distance to the merged factors. The access for one of the merged factors is deleted from the gap matrix, so one can no longer be considered again inside the next generation. The merged facts are added to the same cluster. This system is repeated till the preferred quantities of clusters are acquired.
An array called final clusters is computed with indices as each facts point, and an equal fee is assigned for all statistics factors belonging to the identical cluster. For example, all values belonging to cluster 1 might be assigned the price 1, and those belonging to cluster 2 can be assigned 2.

External Index calculation is done using the subsequent steps: The ground fact statistics and final clusters array are taken as entering. The values for m11, m00, m01and m10 are up to date through evaluating the ground truth labels and clustering labels. Jaccard coefficient and rand index are calculated as consistent with their formulas, and outcomes are displayed.

Jaccard coefficient: m11/(m11+m01+m10)
Rand Index: (m11+m00)/(m11+m00+m01+m10)

The original information is decreased to 2 fundamental components the usage of dimensionality reduction algorithm predominant factor evaluation. This is done through the use of the PCA module from sklearn. Decomposition library. In the end, the result is visualised using a scatter plot.

DBSCAN is a clustering algorithm that businesses collectively points that are close to every different based totally on a distance dimension (Euclidean distance) and a minimum number of points. It also marks as outliers the points that are in low-density regions. The algorithm calls for parameters: eps- the minimum distance among two factors. It approaches that if the gap between points is lower or the same as this value (eps), those factors are considered pals. Minutes- the minimal quantity of factors to form a dense region. For example, if we set the minPts parameter as five, we want at least 5 factors to shape a dense region.

The facts are examined from the given large statistics source and extracted as two lists, one containing the information of the gene (all columns within the dataset beginning from column three) on which the DBSCAN clustering is finished and different containing the ground fact facts (the second column within the dataset). The values of eps and min Pts are examined from the consumer to compute dense regions. Because the first step, the DBSCAN approach is called, reveals the buddies and checks the density? If the object's scale is much less than min Pts, it's far marked as noise. In any other case, it's far assigned to the next cluster, and the extended cluster approach is known as. Within the place query method, all of the points inside the eps-neighbourhood are retrieved via scanning all of the factors, computing the Euclidean distances and checking the cost with the eps.

Ultimately, the expand cluster technique is referred to as whilst the neighbour points are density accessible and introduced to the same cluster. Steps 2-7 are performed iteratively for all records factors, and the cluster labels are retrieved in a list. External Index calculation is consistent with-formed the use of the ground fact facts and cluster labels. The values of m11, m00, m01and m10 are up to date through evaluating those floor fact labels and clustering labels. Jaccard coefficient and rand index are calculated, and effects are displayed.

Jaccard coefficient: m11/(m11+m01+m10)
Rand Index: (m11+m00)/(m11+m00+m01+m10)

The original data is decreased to two foremost additives using dimensionality reduction algorithm primary component evaluation. Sooner or later, the result has visualised the use of a scatter plot.

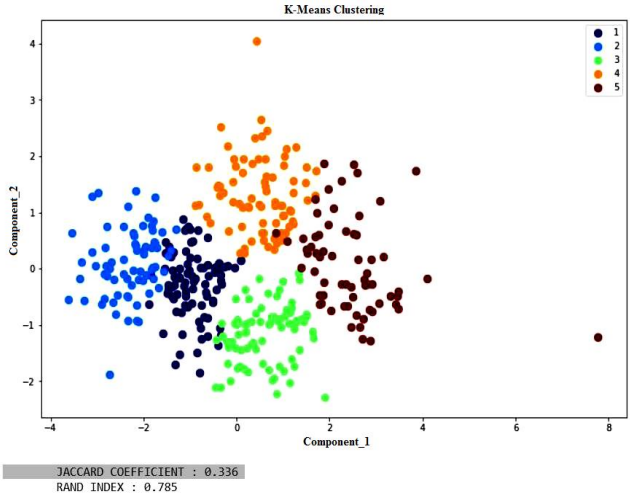
Map lessen is a programming paradigm for large datasets that may be processed speedily by processing them on disbursed clusters in parallel. It consists of two essential steps: Map Step- In performs a filtering or sorting operation and outputs the bring about pair form. Reduce Step- This takes the pair given by the map step and generally performs a summary operation. Apache Hadoop is a popular Map-lessen Framework.

There are specifically two files that run iteratively until the desired situations are met: they may be mapper.py and reducer.py. There is a major.py to govern the wide variety of iterations until the converge condition happens, and it calls the mapper and reducer repeatedly. The cost of k is set right here. The randomly initialised centroid values are passed to the mapper as entering. Inside the mapper.py, the k centroids are given as entering, and the data point is given as entering. Each data point finds the Euclidean distance between the data point and the centroid. The data point is assigned to the cluster whose centroid it has the minimal distance. It outputs the cluster_id and the data point to the reducer. Within the reducer.py, it takes as input the cluster_id and datapoint given to it, and for every cluster_id, it computes the suggestion of all the data points in that cluster and writes the new fee of the centroids to the output.

III. EVALUATION OF BIG DATA CLUSTERING STRATEGIES

The k-means clustering is easy. The data factors are assigned to the nearest cluster. The clusters are round in shape. Some factors lie long from the clusters; however, they belong to a cluster. These may be outliers, and k-means is sensitive to outliers. Typically, Cluster length decreases as the price of k increases. The scatter plots obtained for the datasets Cho and iyer are as proven in figure 1.

Hierarchical clustering with unmarried linkage is implemented. So, the points closest to each different be part of the first to form a cluster. This ends in forming non-elliptical formed clusters, which can be determined from the consequences. The outcomes can also be noticed that the dendrogram is cut at an excessive point to create the restrained range of clusters requested. So this ends in forming one dominant cluster that incorporates a maximum of the statistics points, and most effective, the farther apart factors shape small separate clusters.



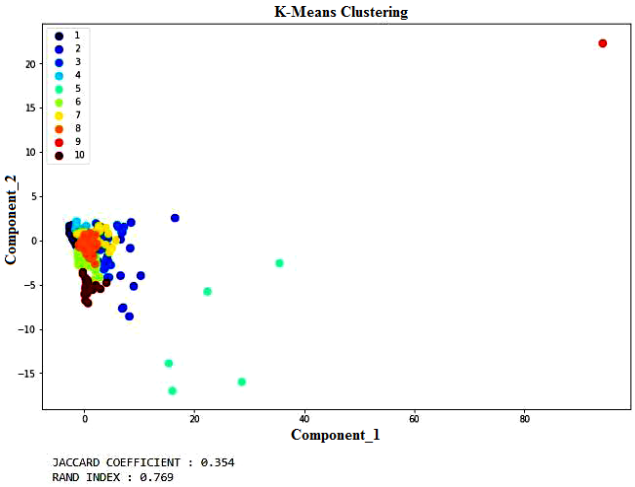


Fig. 1. K-means clustering results on datasets Cho and iyer for k=5 and k =10, respectively, with Jaccard coefficient and index

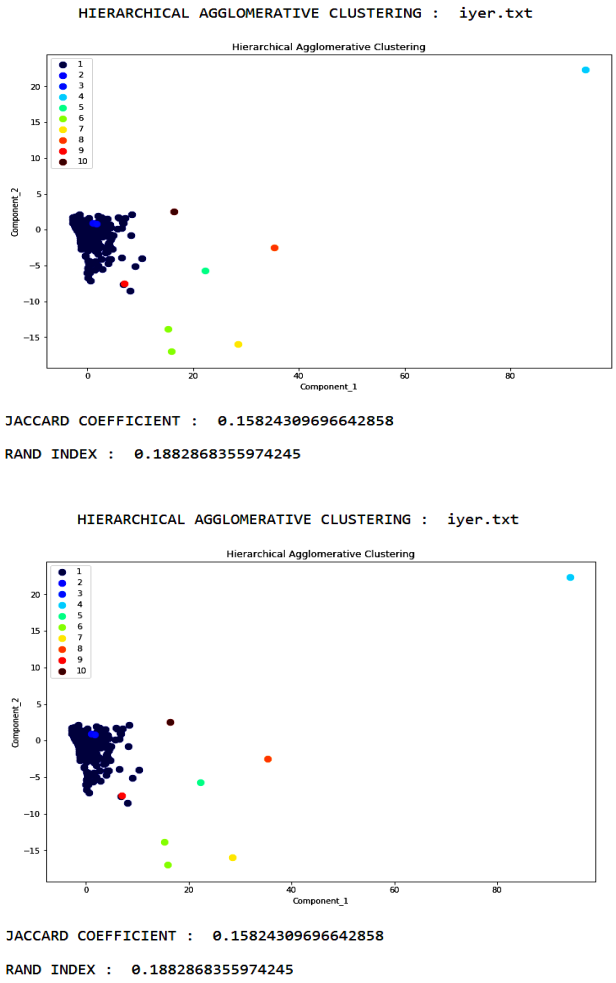
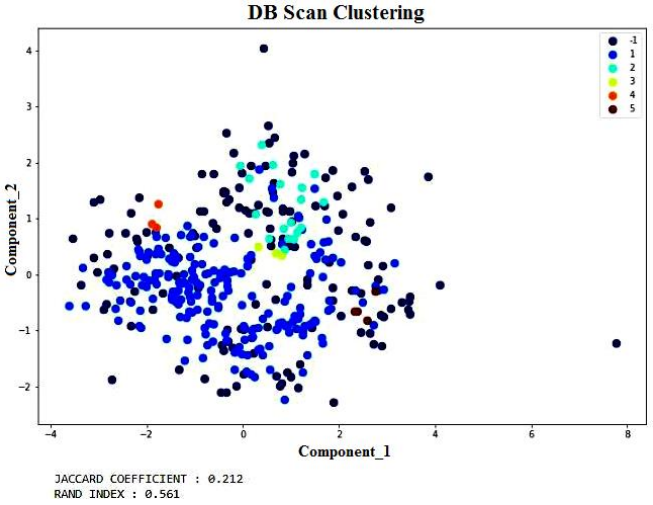


Fig. 2. Hierarchical agglomerative clustering results on datasets Cho and iyer for the number of clusters 5 and 10 respectively with Jaccard coefficient and rand index

It could be visible from the consequences that few factors present inner a large cluster don't belong to that cluster. This is probably a consequence of dimensionality reduction. The one's points would possibly, in reality, be at a very a ways distance from the huge cluster, but because of dimensionality reduction, that may not seem so. It can also be observed that hierarchical clustering doesn't offer evenly dispensed clusters and it's far adversely encouraged with the aid of outliers ensuing in low outside index values (Jaccard Coefficient and Rand Index) in comparison to the other algorithms. The scatter plots received for the datasets Cho and iyer and the usage of hierarchical agglomerative clustering are proven in figure 2.

Experimenting with one-of-a-kind values of eps and minPts shows that DBSCAN is robust to outliers. Suppose we pick a very low eps fee. In that case, a huge part of the information can be considered noise as they don't satisfy the number of points to create a dense place. Still, if we pick out a high value, then most of the statistics points could be within the identical cluster. The fee of minPts has to be selected primarily based on the scale of the dataset and the size of the dataset. With a few area know-how, if we choose the right parameter values, i.e. eps and minPts, we will decide on any arbitrary shaped clusters and even discover clusters that can be surrounded by using clusters. The used datasets are very small. Hence, jogging parallel k means takes longer than the serial one resulting in higher runtime. That is due to the setup overhead involved. Strolling parallel k-means is computationally in-depth for smaller datasets compared to the serial k-means approach. The scatter plots obtained for the datasets Cho and iyer the use of DBSCAN clustering are shown in figure 3.



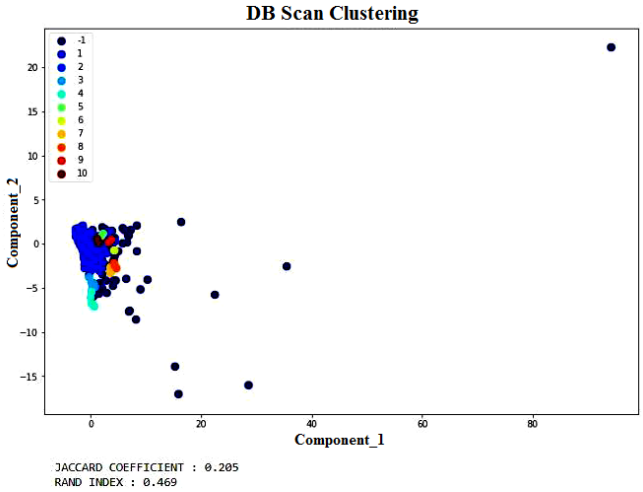


Fig. 3. DBSCAN clustering results on datasets Cho and iyer for eps: 1.14, minPts : 3 and eps: 1.42 and minPts: 2, respectively, with Jaccard coefficient and rand index.

IV. CONCLUSION

The major task in this field is how to process the collected data to understand and obtain novels insights in a reasonable time. Clustering algorithms are the most popular, powerful and commonly used to deal with these data. K-means works nicely with round fashioned clusters. It can produce tighter clusters as compared to different algorithms. This algorithm is efficient based on runtime and easy to put into effect. It is simple to interpret the clustering effects. K-means requires the number of clusters to be exact initially. Without domain knowledge or ground fact values, hard to expect the okay value. The order of the statistics has influenced the final consequences. It is sensitive to outliers and does now not successfully deal with non-round clusters. Hierarchical agglomerative clustering does not require any prior records about a range of clusters. Any preferred cluster number can be received by 'reducing' the dendrogram at the right stage. It produces meaningful taxonomies. It produces an ordering of informative objects for statistics show (Dendrogram). It is sensitive to noise and outliers. It can't undo any formerly made selections, including combining clusters. It isn't easy to address clusters with exceptional sizes and convex shapes. It has high time complexity.

DBSCAN can deal with any form of cluster. It does not require any prior statistics on the approximate quantity of clusters. It requires just two parameters and does not rely on the ordering of the points. It's miles proof against outliers. DBSCAN is extraordinarily dependent on the parameters eps and min Pts, and it's miles hard to determine the proper values. It fails to perceive clusters if density varies or the dataset is too sparse. It relies upon the gap measure – Euclidean distance. Excessive dimensional statistics perform poorly due to diverse phenomena that arise whilst analysing and arranging statistics. The dataset can't be sampled as sampling would affect the density measures. In the okay approach with the map lessen method, the logical partitioning of mapper and reducer features makes the code extra readable. For extremely big datasets, parallel and disbursed processing of ok-manner clustering runs appreciably faster than the iterative serial implementation. The setup overhead for the parallel okay approach clustering could be very large for small datasets. For smaller datasets, serial okay-manner runs quicker than the okay parallel approach.

REFERENCES

[1] Yang Liu, Shuaifeng Ma, and Xinxin Du, A Novel Effective Distance Measure and a Relevant Algorithm for Optimising the Initial Cluster Centroids of K-means IEEE Access Early Access, DOI: 10.1109/ACCESS.2020.3044069, (2021).

[2] Dhanachandra, N., Manglem, K., & Chanu, Y. J. Image Segmentation Using K -means Clustering Algorithm and Subtractive Clustering Algorithm, Procedia Computer Science, 54 764–771.

[3] Habib, S. T., & Zahid, A. An Analysis of MapReduce Efficiency in Document Clustering using Parallel K-Means Algorithm, Future Computing & Informatics Journal, (2018).

[4] Siddiqui, F. U., & Mat Isa, N. A., Enhanced moving K-means (EMKM) algorithm for image segmentation, IEEE Transactions on Consumer Electronics, 57(2) (2011) 833–841.

[5] Tleis, M., Callieris, R., & Roma, R., Segmenting the organic food market in Lebanon: an application of K-means cluster analysis, British Food Journal, 119(7) (2017) 1423–1441.

[6] Sridharan, K., & Sivakumar, P., A Systematic Review On Feature Selection and Classification Techniques for Text Mining, International Journal of Business Information Systems, 28(4) (2018) 504–518 .

[7] Tal, G. Dend Extend An R Package for Visualising, Adjusting and Comparing Trees Of Hierarchical Clustering. Bioinformatics, 22 (2015) 3718–3720.

[8] Premkumar, M. S., & Ganesh, S. H., A Median Based External Initial Centroid Selection Method for K-Means Clustering, 143–146, 2017.

[9] Cohen-Addad, V., Approximation Schemes for Capacitated Clustering in Doubling measures, (2018).

[10] Friggstad, Z., Khodamoradi, K., & Salavatipour, M. R., Exact Algorithms and Lower Bounds for Stable Instances of Euclidean k-means, Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms, (2019)2958–2972..

[11] Stemmer, U., Locally Private k -Means Clustering, 2020.

[12] Chakraborty, S., & Das, S., k means Clustering with a New Divergence-Based Distance Measure: Convergence and Performance Analysis, Pattern Recognition Letters, (2017).

[13] Celebi, M. E., Kingravi, H. A., & Vela, P. A., A comparative study of Efficient Initialisation Methods For The K-Means Clustering Algorithm, Expert Systems with Applications: An International Journal, 40 (2013).

[14] Lei, J., Jiang, T., Wu, K., Du, H., Zhu, G., & Wang, Z., Robust K-Means Algorithm with Automatically Splitting and Merging Clusters and its Applications for Surveillance Data, Multimedia Tools And Applications, 75(19) (2016) 12043–12059.

[15] Adil Abdu Bushra and Gangman Yi, Comparative Analysis Review of Pioneering DBSCAN and Successive Density-Based Clustering Algorithms, IEEE Access, 9 (2021) 87918 – 8793.

[16] T. N. Tran, K. Drab, and M. Daszykowski, Revised DBSCAN Algorithm to cluster data with dense adjacent clusters, Chemometrics Intell. Lab. Syst., 120 (2013) 92-96.

[17] H. Chebi, D. Acheli, and M. Kesraoui, Dynamic detection of Abnormalities in Video Analysis of Crowd Behavior with DBSCAN and Neural Networks, Adv. Sci., Technol. Eng. Syst. J., 1(5) (2016) 56-63.

[18] H. Li, J. Liu, K.Wu, Z. Yang, R.W. Liu, and N. Xiong, Spatio-temporal vessel Trajectory Clustering Based on Data Mapping and Density, IEEE Access, 6 (2018) 58939-58954.

[19] H. Li, J. Liu, Z. Yang, R. W. Liu, K. Wu, and Y. Wan, Adaptively constrained dynamic time warping for time series classification and clustering, Inf. Sci., 534 (2020) 97-116.

[20] R.W. Liu, J. Nie, S. Garg, Z. Xiong, Y. Zhang, and M. S. Hossain, Datadriven trajectory quality improvement for promoting intelligent vessel traffic services in 6G-enabled maritime IoT systems, IEEE Internet Things J., 8(7) (2021) 5374-5385.

[21] A. K. Jain, Data clustering: 50 years beyond K-means, Pattern Recognit. Lett., 31(8) (2010) 651–666.

[22] L. Dalton et al., Clustering algorithms: On learning, Validation, Performance, and Applications to Genomics, Current Genomics, 10(6) (2009) 430–445.

[23] A. Srivastava et al., Statistical shape analysis: Clustering, Learning, and Testing, IEEE Trans. Pattern Anal. Mach. Intell., 27(4) (2005) 590–602.

[24] T. Wu, S. A. N. Sarmadi, V. Venkatasubramanian, A. Pothen and A. Kalyanaraman, Fast svd computations for synchrophasor algorithms, IEEE Transactions on Power Systems, 31(2) (2015) 1651-1652.

[25] Chris Ding and Xiaofeng He, K-Means Clustering via Principal Component Analysis, In proceedings of the 21st International Conference on Machine Learning, Banff, Canada, (2004).