Review and Comparative Analysis of Data Clustering Algorithms

Ugonna Victor Okolichukwu[1], Beatrice Adenike Sunday[2], Friday E. Onuodu[3]

[1] *Department of Computer Science Education, Federal College of Education, Eha-Amufu, Enugu State, Nigeria*

[2] *Department of Computer Science, School of Postgraduate Studies, Ignatius Ajuru University of Education, Rivers State, Nigeria*

[3] *Department of Computer Science, University of Port Harcourt, River State, Nigeria*

Abstract---- **Data mining is a process with an objective of information extraction from huge datasets. Data mining involves extracting useful data from a huge quantity of raw data to solve a given problem of clustering. Thus, it is otherwise called Knowledge Discovery of Data (KDD). Clustering is an aspect of machine learning that is of great importance in the area of data mining analysis. Clustering involves the grouping of a set of similar data objects into the same group (clusters) considering their unique qualities and similarity. A good clustering algorithm will result to an increased rate of intra-grouped similarity and a decreased rate of inter-grouped similarity. Clustering algorithm are grouped into Hierarchical, Partitioning and Density-based clustering algorithm. The Partitioning clustering algorithm splits the data objects into a number groups called partition and each partition represents a cluster. Hierarchical clustering techniques creates a hierarchy or tree of clusters for the data objects. Density-based algorithms groups its data objects based on a particular neighbourhood and locates the cluster in regions with high density. The purpose of this paper was to do a comparison between hierarchical, partitioning and density-based clustering algorithms based on their observed features and functions, and the metrics used is ability to deal with or handle noise and/or outliers. We conclude our findings using a summary table on their performance, stating that Density-based algorithm is highly sensitive in dealing with outliers and/or noise easily than hierarchical and partitioning clustering algorithm.**

Keywords--- **Data mining, Clustering, Data clustering algorithm, Outlier.**

I. INTRODUCTION

The term “Data” is defined as an information in an unprocessed state. In every field of human endeavor, data is continuously encountered. Mining simply means to extract or bring out or dig out something from something. Data mining deals with information extraction from huge sets of data. It can otherwise be called Knowledge Discovery of Data (KDD). Data mining is geared towards the extraction of useful data from a huge quantity of raw data to solve a given problem of clustering.

Mann and Kaur, 2013 posited that Association, Outer detection, Classification and Clustering are techniques in data mining. Clustering is an important tool for data mining and analysis which is focused on assigning of combination of objects in sets such that the objects in the identical or similar cluster (group) are allied or alike to objects in that cluster and unrelated to the objects in the other clusters (groups).

A. Clustering

Clustering is an aspect of machine learning that is of great importance in the area of data mining analysis. Tayal and Raghuwanshi, 2010 defined clustering as the process of grouping samples so that the samples are similar within each group. Data mining is categorized into two levels of learning: supervised and unsupervised learning, and it undergoes different process cycle. Bala, Sikka and Singh, 2014 opined that clustering is an unsupervised learning in data mining and its application is known in the area of artificial intelligence, image and pattern recognition, bioinformatics, segmentation and machine learning.

Furthermore, clustering involves grouping sets of related data objects into the same group (clusters) considering their unique qualities and similarity. Most often data being mined are usually large, containing noise and outliers. Thus, this noise and outliers need to be accurately identified or detected or handled by different algorithms employed.

B. Outlier

Outlier is the data object pattern that is dissimilar from the other pattern of data object, which is often cause by human error, change in system behavior, experimental error, etc (Barai and Dey, 2017). Noise is defined as a mislabeled data object (group or cluster) or errors in the values of attributes while outliers are abnormalities, deviants and odd data object in a cluster (Salgado et al., 2016). Hence, it needs to be detected or handled by any method of clustering chosen.

C. Cluster

Defreitas and Bernard, 2015 is of the view that clustering algorithms are techniques applied in data mining for the purpose of grouping set of data objects into subsets. These subsets or clusters represents organized similar objects within in a given cluster and are diverse to objects contained in the other clusters (Defreitas & Bernard, 2014, Han et al., 2012).

Cluster is opined as an organised list of data with resembling characteristics [13]. A cluster is also described as a collection of objects which are “similar” between them and are “dissimilar” to the objects contained in other clusters [20].

II. RELATED WORKS

Vijayarani and Jothi, 2014 in their work titled hierarchical clustering and partitioning grouping algorithms for recognizing outliers in data streams, planned for playing out the clustering process and distinguishing the anomalies in data streams. They utilized two sorts of clustering algorithms to be specific BIRCH with K-Means and CURE with K-means, for finding the anomalies in data streams. Two execution factors, for example, clustering exactness and anomaly detection precision were utilized for observation. Through looking at the experimental outcomes, it is seen that the CURE with K-Means grouping algorithm execution was more precise than the BIRCH with K-Means algorithm. Their investigation was restricted to just numerical dataset; hence this study intends to compare the single performance of the clustering algorithm on data objects through their observable differences.

Chris and Xiaofeng, 2002 did a work on hierarchical clustering algorithm which involved the process of merging and splitting. They gave an exhaustive investigation of choice techniques and proposes a few new strategies that decide how to best choose the following cluster for split or union activity on cluster. The researchers carried out a broad clustering examinations to evatuate eight choice strategies, and concluded with average comparability as being the most appropriate technique in divisive clustering while in agglomerative clustering the most appropriate technique is Min-Max linkage. Cluster equalization was a key factor there to accomplish great execution. They likewise presented the idea of target work saturation and clustering objective separation to successfully evaluate the quality and nature of clustering. This investigation didn't consider noise and outlier detection in applying these algorithms to analyze the data.

Popat and Emmanuel, 2014 worked on the review and comparative study of clustering techniques. They focused on Document clustering, which aimed at using an unsupervised method to group a specified document set into clusters, where the documents within each clusters are suchlike than those in dissimilar clusters. In their work, a survey on various clustering techniques such as “Partitional algorithms, Hierarchical algorithms, Density based, and comparison of various algorithm was done”, and findings shows that Hierarchical Clustering is more preferred than other techniques. They study did not look into the performance of this algorithms with regards to noise and outliers.

A. Issues Encountered in Cluster Analysis

Gupta, 2006 proposed the following as issues encountered in clustering analysis:

- Knowledge of the characteristics of cluster analysis
- The approach to measure similarities in analyzed dataset
- Knowing the appropriate method for cluster analysis for a specific dataset
- Efficiently cluster large data set with high attributes
- The how of evaluating cluster analysis result

B. Properties of Data Mining Clustering Algorithms

Firdaus and Uddin, 2015 stated the properties of data mining clustering algorithms as follows:

- Scalability (with respect to time and space)
- Capable of handling varieties of attributes
- Finding arbitrary shaped clusters
- It can handle noise and outliers
- Able to cluster data of high dimensionality
- Insensitive to order of input records
- Ease of use interpretation

III. TYPES OF DATA CLUSTERING ALGORITHM

Data Clustering algorithm is a useful technique in data mining as it helps to locate data that are alike to each other and are different in other cluster (group). Ashok et al., 2013 described data mining as “the exploration and analysis of large quantities of data in order to discover valid, novel, potentially useful and ultimately understandable patterns in data and the successful achievement of this requires a clustering algorithm”. A good clustering algorithm will result to an increased rate of intra-grouped similarity and a decreased rate of inter-grouped similarity. Data clustering algorithms are broken down into:

1. Hierarchical clustering algorithm
2. Partitioning clustering algorithm
3. Density-based clustering algorithm

A. Hierarchical Clustering Algorithm

Hierarchical clustering algorithm creates a hierarchy or tree of clusters for the data objects. This algorithm is a type of cluster analysis that tries to create or develop a hierarchy of clusters or a tree of clusters, also known as a Dendrogram, where each child clusters is contained in each cluster node, and the sibling clusters partition the points covered by their common parent [1].

The hierarchical clustering algorithm can either begin with one cluster and then divide into smaller clusters (called Divisive or Top down) or begin an individual cluster with an object and then attempt to bring together similar clusters into larger and larger clusters (called Agglomerative or Bottom up).

In agglomerative or bottom up clustering algorithm, it is rigid, based on the fact that all split or merges are final when done and cannot be reversed, hence, no back-tracking (Rokach and Maimon, 2002), and this can also cause trouble for noisy, high-dimensional data, such as document data, thus “Balanced Iterative Reducing and Clustering Using Hierarchies” (BIRCH) is an example of hierarchical clustering algorithm that tries to minimize the problem [8]. Also, well outlined steps are used to guide the merge or split decisions to avoid having low-quality clusters. This clustering algorithm does not scale properly because of time complexity and as such detecting noise and outliers is poor (Saket and Pandya, 2016).

Hence, the present day data mining deals on larger dataset mined at a high level of speed, time, accuracy and detecting noisy data and outliers, hence this algorithm cannot function effectively.

Moreover, the hierarchical clustering algorithm (agglomerative or divisive) is applied in analyzing data which has the basic hierarchical structure within the data [7]. Also, BIRCH is designed to reduce the numerous input and output operations as it incrementally and dynamically groups incoming data points and attempt to bring out the most appropriate quality clustering with the resources available such as time and memory (Rani and Rohil, 2013). One of the advantages of this algorithm is ease of managing different shapes of distance or similarity.

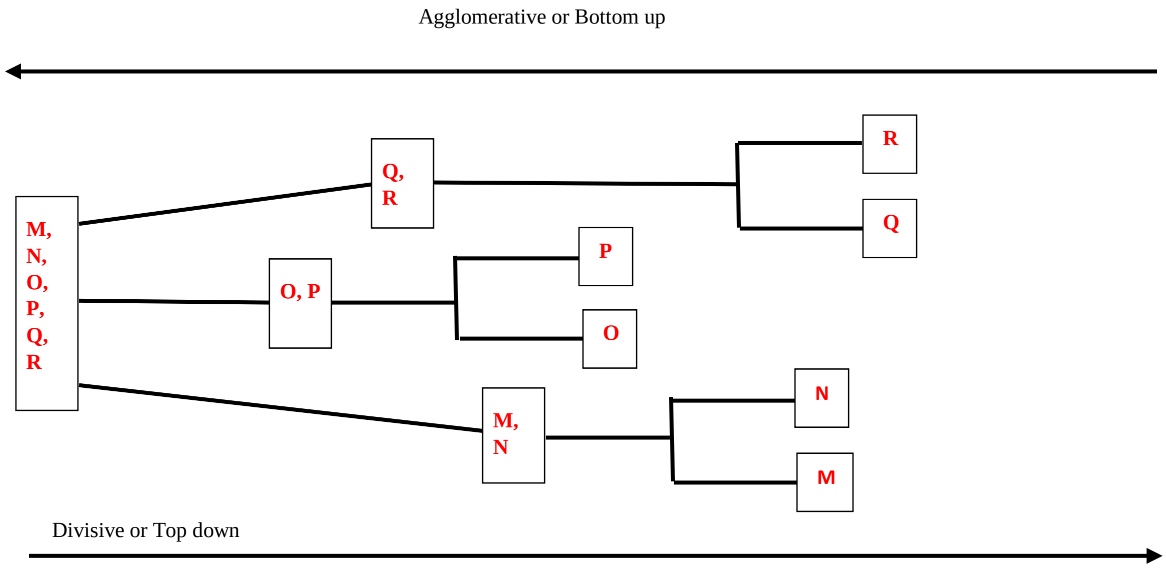

Fig. 1. Hierarchical Clustering Algorithm

B. Partitioning Clustering Algorithm

Partitioning clustering algorithms is an algorithm that groups objects into a specified partition where each partition reflects a cluster such that, the objects within that cluster are “similar” and “dissimilar” to objects in other clusters. Iterative relocation technique is then applied which tries to improve the grouping (partitioning) in the way of moving objects from a group to another hence partition of high quality is achieved suppose the objects in identical clusters are “close” or suchlike the other, and objects in distant clusters are “far apart” or very dissimilar [8].

In partitioning clustering algorithm, K-mean partition algorithm is mostly used. In 1967, James Macqueen developed K-Mean and defined it as a cluster represented by its centroid which represents the mean of points contained in a cluster (Rai and Singh, 2010). The k-means algorithm is simple to implement and its design is to deprecate the sum of distances, breaking up the data points in partitions called K, where one cluster represents one partition, whereas its weakness produces a poor or misleading result resulting from an overlap of data points especially when the gap between two points from the center are close to another cluster [19].

The k-means partitioning algorithm functions properly only with numerical attributes but a single outlier negatively affects it. From the ongoing, this cannot be a good algorithm for mining numerical and object with large dataset as it becomes a challenge for K-means to handle data with outliers, clusters of differing sizes, densities and non-globular shapes.

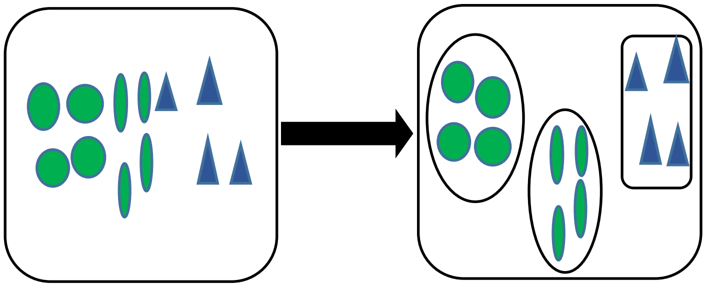

Fig. 2.Partitioning Clustering Algorithm

C. Density-Based Clustering Algorithm

Density-based Clustering algorithm groups its data objects based on a particular neighbourhood and locates the cluster in regions with high density. This algorithm is devised to create arbitrary-shaped and spherical-shaped clusters, therefore, a cluster is regarded as a region in which the density of data objects exceeds a threshold. Clusters are determined by checking areas with high dense region. Areas with high dense region points represents the existence of clusters while areas with low dense region points represents clusters of noise or outliers and this makes the algorithm suitable to deal with or handle large datasets, with outliers, noise and thus, it's capable of pointing out clusters with various shapes and sizes.

An example of this algorithm is the Density-based Clustering Algorithm Nesting (DBSCAN). The DBSCAN separates data points into Core points, Border points and Noise points (Chaudhari and Parikh, 2012). In density-based clustering algorithm nesting, Core points indicates the points located at the interior of a cluster while Border points shows the points that falls around the neighborhood of a core point but is not a core point whereas the Noise points represents any point that is not a border or core point. Its ability to identify outlier is a characteristic feature,thus, the basis for its high sensitivity towards detecting and being able to deal with or handle noise and outliers.

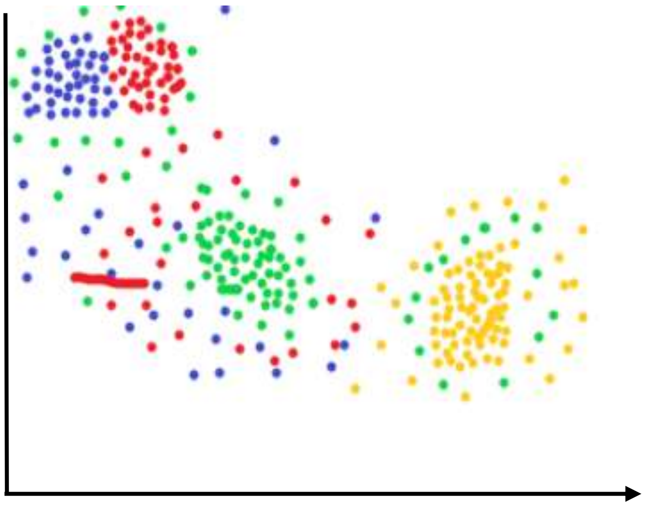

Figure 3:Density-based Clustering Algorithm

IV. SUMMARY OF COMPARATIVE ANALYSIS ON HIERARCHICAL, PARTITIONING AND DENSITY-BASED CLUSTERING ALGORITHM.

Algorithm Type	Comparison Metrics	Data Type	Performance
Hierarchical Clustering Algorithm (BIRCH)	Outliers and Noise	Numerical / Object	❖ A tree of clusters, simple to implement ❖ No back-tracking, hence it is rigid ❖ Limited to numerical data only ❖ Not robust to detecting noise and outlier
Partitioning Clustering Algorithm (K-means)	Outliers and Noise	Numerical / Object	❖ Group of k partition clusters ❖ Difficulty to group large datasets of different sizes, shapes and densities ❖ Works conveniently on numerical data ❖ Easily affected by a single outliers
Density-based Clustering Algorithm (DBSCAN)	Outliers and Noise	Numerical / Object	❖ Locates cluster in regions of high density ❖ Easily identify clusters with different sizes and shapes ❖ Robust to both numerical and object data ❖ Detects and Handles noise and outlier

Table 1. showing the comparative analysis on the hierarchical, partitioning and density-based clustering algorithm based on outliers and noise.

V. CONCLUSION

Data mining is a process with an objective of information extraction from huge datasets. Data mining involves extracting useful data from a huge quantity of raw data to solve a given problem of clustering. Thus, it is otherwise called Knowledge Discovery of Data (KDD). Clustering is an aspect of machine learning that is of great importance in the area of data mining analysis. Clustering involves the grouping of a set similar data objects into the same group (clusters) considering their unique qualities and similarity. A good clustering algorithm will result to an increased rate of intra-grouped similarity and a decreased rate of inter-grouped similarity. Clustering algorithm are grouped into Partitioning, Hierarchical and Density-based clustering algorithm. The Partitioning clustering algorithm splits the data objects into a number groups called partition and each partition represents a cluster. Hierarchical clustering techniques creates a hierarchy or tree of clusters for the data objects. Density-based algorithms groups its data objects based on a particular neighbourhood and locates the cluster in regions with high density. Grid-based algorithm operates on a grid structure formed as the data object space splits into a finite number of cells.

The purpose of this paper was to do a comparison between hierarchical, partitioning and density-based clustering algorithms based on their observed features and functions, and the metrics used is ability to handle or deal with or noise and/or outliers. We conclude our findings using a summary table on their performance, stating that Density-based algorithm is highly sensitive in detecting and able to deal with noise and outliers easily, and handles large datasets than hierarchical and partitioning clustering algorithm.

REFERENCES

[1] Ashok, A. R., Prabhakar, C. R. and Dyaneshwar, P. A.,(2013). Comparative Study on Hierarchical and Partitioning Data Mining Methods. International Journal of Science and Research (IJSR), India, 2(3), pp. 211-215.

[2] Bala, R., Sikka, S. and Singh, J., (2014). A Comparative Analysis of Clustering Algorithms. International Journal of Computer Applications, India, 100(15), pp. 35-39.

[3] Baviskar, C. T. and Patil, S. S., (2016). Dealing with Overlapping of Clusters by usingFuzzy K-Means Clustering. International Conference on Global Trends in Engineering, Technology and Management (ICGTETM), India, pp. 45-51.

[4] Chaudhari, B. and Parikh, M., (2012). A Comparative Study of Clustering Algorithms UsingWeka Tools. International Journal of Application or Innovation in Engineering & Management (IJAIEM), 1(2), pp. 154-158.

[5] Chris, D. and Xiaofeng, H., (2002). Cluster Merging and Splitting in Hierarchical Clustering Algorithms.

[6] DeFreitas, K. and Bernard, M., (2015). Comparative Performance Analysis of Clustering Techniques in Educational Data Mining. International Journal on Computer Science and Information Systems,Trinidad and Tobago, 10(2), pp. 65-78.

[7] DeFreitas, K. and Bernard, M., (2014). A Framework for Flexible Educational Data Mining. In The 2014 International Conference on Data Mining. Las Vegas, USA, pp. 176–180.

[8] Firdaus, S. and Uddin Md. A., (2015). A Survey on Clustering Algorithms and ComplexityAnalysis. International Journal of Computer Science Issues (IJCSI), Bangladesh, 12(2), pp. 62-85.

[9] Gupta, G. K., (2006). Introduction to data mining with case studies.PHI Learning Pvt. Ltd.

[10] Han, J., Kamber, M. and Pei, J., (2012). Data Mining: Concepts and Techniques 3rd ed., Massachusetts, USA: Morgan Kaufmann.

[11] Jain, S., Aalam, M. A. and Doja, M. N., (2010). K-means Clustering Using Weka Interface.Proceedings of the 4th National Conference, India.

[12] Mann, K. A. and Kaur, N., (2013). Review Paper on Clustering Techniques. Global Journal of Computer Science and Technology Software & Data Engineering, USA, 13(5), pp. 42-48.

[13] Popat, S. K. and Emmanuel, M., (2014). Review and Comparative Study of Clustering Techniques. International Journal of Computer Science and Information Technologies (IJCSIT), 5(1), pp. 805-812.

[14] Rai, P. and Singh, S., (2010). A Survey of Clustering Techniques. International Journal of Computer Applications, 7(12), pp. 1-5.

[15] Rani, Y. and Rohil, H., (2013). A Study of Hierarchical Clustering Algorithm. International Journal of Information and Computation Technology, India, 3(11), pp. 1225-1232.

[16] Rokach, L. and Maimon, O., (2002). Clustering Methods. Department of Industrial Engineering Tel-Aviv University.

[17] Salgado, C. M., Azevedo, C., Proenca, H. and VSM., (2016). Secondary Analysis of ElectronicHealth Records: Noise versus Outliers. pp 163 -183.

[18] Saket, S. J. and Pandya, S., (2016). An Overview of Partitioning Algorithms in Clustering Techniques. International Journal of Advanced Research in Computer Engineering &Technology (IJARCET), 5(6), pp. 1943-1946.

[19] Tayal, M. A. and Raghuwanshi, M. M., (2010). Review on Various Clustering Methods for the Image Data. Journal of Emerging Trends in Computing and Information Sciences, India, 2, pp. 34-38.

[20] Vijayarani, S. and Jothi, P., (2014). Hierarchical and Partitioning Clustering Algorithms for Detecting Outliers in Data Streams. International Journal of Advanced Research in Computer and Communication Engineering, 3(4), pp. 6204-6207.