International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 9 Sep 2021, pp: 84-89 www.ijaem.net ISSN: 2395-5252

Data Clustering Algorithms

H V Sampad,Student, Dr.Kiran V,
Department of Electronics and Communication, RVCE Bangalore
Associate Professor,Department of Electronics and Communication, RVCE Bangalore

Submitted: 30-08-2021 Revised: 03-09-2021 Accepted: 05-09-2021

ABSTRACT:This research paper is about one of the Unsupervised Machine learning techniques that is Clustering Algorithms types. The application and how good the results are obtained for the data set how they classify the data and data are predicted. A case study of Different algorithms is done concerning certain data set and the results are analysed. Comparative analysis of different algorithms is conducted and inferred.

KEYWORDS: Data Clustering,Affinity Propagation,Agglomerative Clustering,Birch, DBSCAN,K-Means,Mini-Batch K-Means, Mean Shift, Optics, Spectral Clustering, Gaussian Mixture Model

I. INTRODUCTION

In this paper, Different types of Clustering Algorithms are studied and the application of those algorithms on the different applications is studied. The clustering algorithm is one of the algorithms used in machine learning to segregate the data into different sets depending on the pattern of data available. We don't have a single good method to segregate the data. Cannot use a single algorithm for sorting all the data sets. For application, the data available pattern, data sets how accurate the data is. How accurate the application is Real-time system or the Reactive System, Data Available Reliability. Best Suited Clustering Method is applied.

Clustering is an unsupervised machine learning technique. This involves automatically finding the group in data. A cluster is in general a set of data in a particular observing a certain set of the data pattern. Clustering can help a lot in data analysis to know more about the problem. Once Data is Clustered and for further analysis, we may need a domain expert to analyze the data and work on it.

Most Clustering Algorithms generally use Similarity or the distance measure between datasets in space to cluster, i.e how close the data is present. Some Clustering Algorithms need us to specify the number of clusters to be formed on the given datasets,Eachalgorithm uses its model to work on the data and give the results, In this Paper, we have worked on analyzing the different clustering algorithm types by working on datasets. Here We work on 10 popular algorithms which are as follows a. Affinity Propagation b. Agglomerative Clustering c. BIRCH d. DBSCAN e. K – Means e. Mini -Batch K-Means f. Mean Shift g. OPTICS h. Spectral Clustering I. Mixture of Gaussians

II. ALGORITHM ANALYSIS

Algorithms and their respective results are reported below.

A. Affinity Propagation

The Cluster Algorithm is based on the concept of passing messages between the given data points. Unlike other algorithms, it does not require the total number of clusters to be formed on the given data or the estimated number of clusters to be mentioned before running the Affinity Propagation Clustering Algorithm.

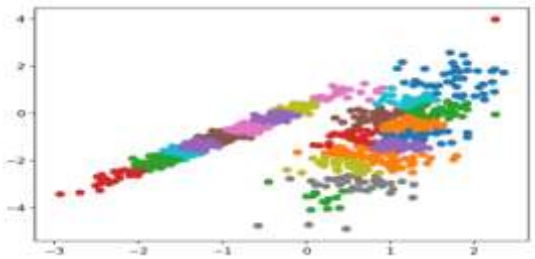

Fig b. Propagation Clustering Algorithm Results

International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 9 Sep 2021, pp: 84-89 www.ijaem.net ISSN: 2395-5252

B. BIRCH

BIRCH stands for Balanced Iterative Reducing and Clustering using Hierarchies. This involves building a tree-like structure and once built with these structures cluster centroids will be extracted. This is a scalable method that only needs to scan the datasets once which makes it fast for working on large data sets. This algorithm divides the data in terms of nodes. These clusters will also have some sets of sub-clusters.

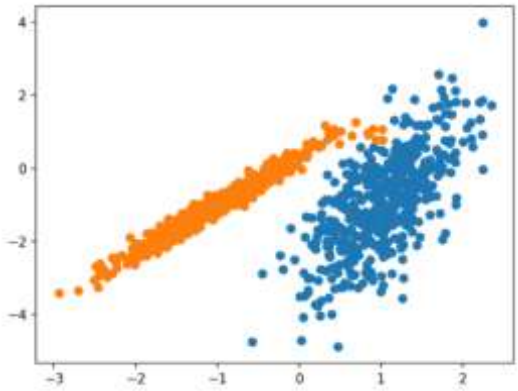

Fig c. BIRCH algorithm Results

C. DBSCAN

DBSCAN stands for Density-Based Spatial Clustering of Applications with Noise. In these algorithms,a High-Density area is found and they are expanded further for high accuracy and more clear analysis. This algorithm uses two important parameters a. minPts: ie, what are the minimum number of points that has to be considered for it to consider as a cluster. b.Eps: This is the measure of distance that will be used to locate the points beside points.

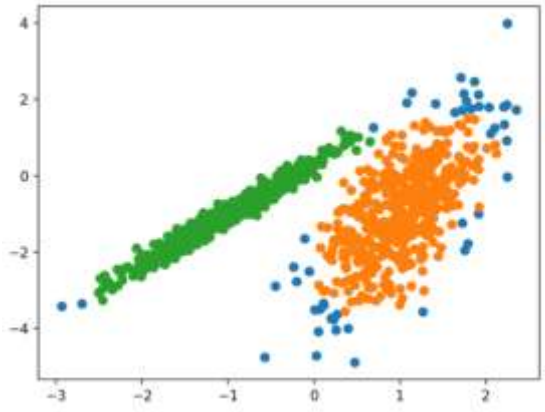

Fig d. DBSCAN algorithm Results

D. K-Means

This algorithm is widely is used. Here clusters are assigned with examples to minimizer the variance with the clusters formed. It tries to partition the data into k partitions, where each data points belong to only one group. This tries to make cluster points as similar as possible.

International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 9 Sep 2021, pp: 84-89 www.ijaem.net ISSN: 2395-5252

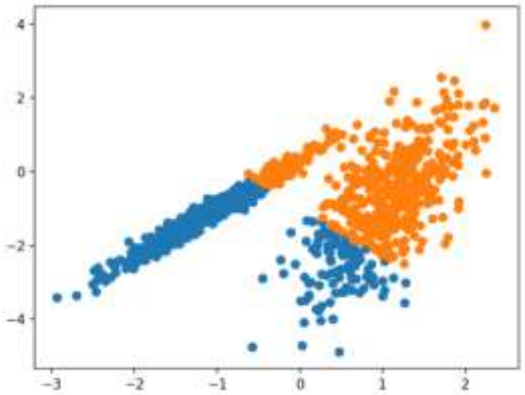

Fig e. DBSCAN algorithm Results

E. Mini-Batch K-Means

This is the modified version of the K-Means Algorithm where the clusters are formed using K-Means techniques but not on the whole sample. The sample is divided into sub batches and is applied on batches. Each iteration is new and is update in the clusters and this will be updated till we get the required Convergence.

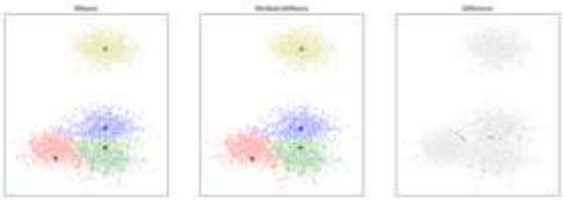

Fig f. K-Means, Mini-Batch K-Means and difference Results

F. Mean Shift

This Algorithm involves to find and adapt the centroids based on the density of the dataset clusters. This algorithm involves shifting of points towards the mode. Mode is the region where we have highest density of points. This can be also called as Mode Seeking Algorithm.

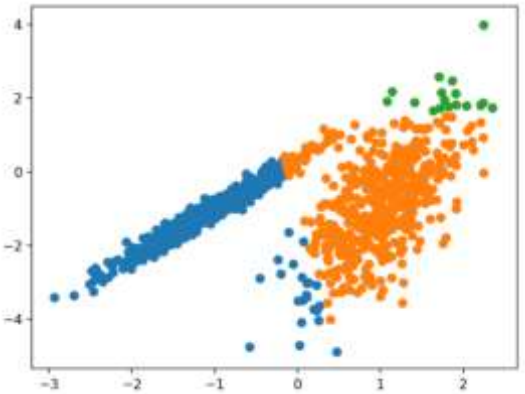

Fig g. Mean Shift Algorithm Results

G.OPTICS

OPTICS Stands for Ordering Points To Identify the Clustering Structure This is the modified type of DBSCAN algorithm. This new algorithm was introduced which does not produce the clustering set explicitly. But will create the augmented ordering of the database which will represent the density-based clustering structure. This does not extensively segment the data into clusters but gives out the visualization of reachable distances and uses this visualization to cluster these data.

International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 9 Sep 2021, pp: 84-89 www.ijaem.net ISSN: 2395-5252

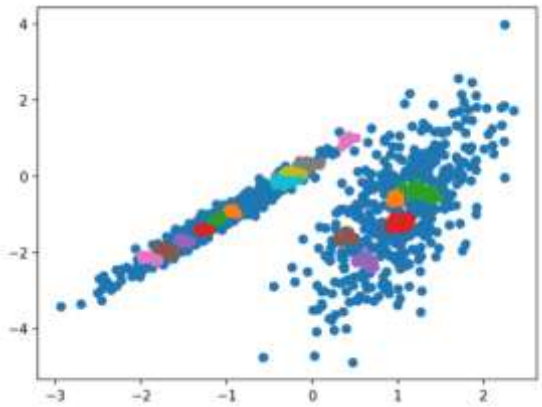

Fig h. OPTICS algorithm Results

H. Spectral Clustering

This clustering algorithm is derived from Linear Algebra. They make use of eigenvalues from the similarity matrix of data and dimensional reduction is done before clustering the data.Thes Similarity matrix will be provided as an input and this will consist of the quantitative assessment of the relative similarity of the individual pair of points present in the datasets.

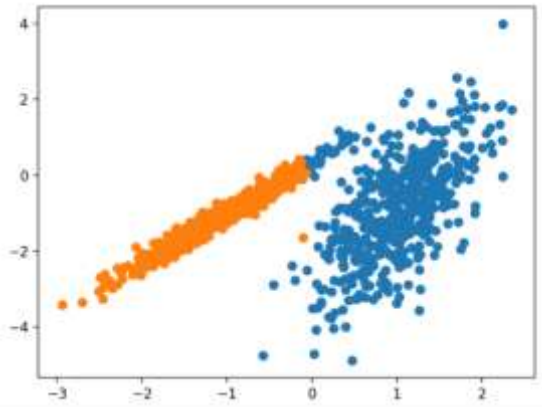

Fig i. Spectral clustering algorithm results

I. Gaussian Mixture Model

The Gaussian mixture model will summarize the multivibrator probability function including the mixture of the Gaussian Probability distribution. They will be used to cluster the unlabelled data similar to the K-mean algorithm with some advantages where K-means won't account for the variance.i.e K-mean cannot handle data when shapes are not perfect. But Gaussian model can handle every oblong cluster accurately. The other difference is K-mean algorithm tells us which data points will be present in which cluster but won't give us the probability that a given data point will be belonging to each cluster. i.e probability of datapoint belonging to each cluster.

International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 9 Sep 2021, pp: 84-89 www.ijaem.net ISSN: 2395-5252

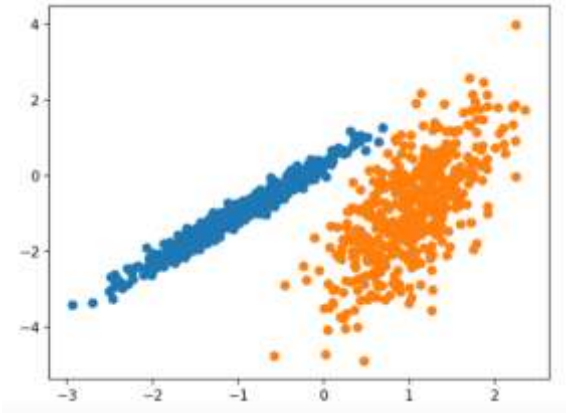

Fig j. Gaussian Mixture Model Results

III. CONCLUSION

We have different Clustering algorithms which will be implemented as discussed above and will be applied to the dataset and required application. These Clustering applications are effectively used in various applications which include customer segmentation, market research, medical imaging, biological data, recommendation engine, search result clustering,social network analysis, image processing, The data required for Ml techniques are also processed with clustering algorithms. With these applications, Data is clustered using the clustering algorithms mentioned above.

REFERENCES

[1]. K. M. A. Patel and P. Thakral, "The best clustering algorithms in data mining," 2016 International Conference on Communication and Signal Processing (ICCSP), 2016, pp. 2042-2046

[2]. S. Kapil, M. Chawla and M. D. Ansari, "On K-means data clustering algorithm with genetic algorithm," 2016 Fourth International Conference on Parallel, Distributed and Grid Computing (PDGC), 2016, pp. 202-206, doi: 10.1109/PDGC.2016.7913145.

[3]. Qinghe Zhang and Xiaoyun Chen, "Agglomerative hierarchical clustering based on affinity propagation algorithm," 2010 Third International Symposium on Knowledge Acquisition and Modeling, 2010, pp.250-253,doi:10.1109/KAM.2010.5646241 .

[4]. X. Liu and J. Xu, "Based on Multiple Time Series Affinity Propagation Algorithm," 2018 IEEE 4th Information Technology and Mechatronics Engineering Conference (ITOEC), 2018, pp. 1500-1503, doi: 10.1109/ITOEC.2018.8740379.

[5]. H. Du and Y. Li, "An Improved BIRCH Clustering Algorithm and Application in Thermal Power," 2010 International Conference on Web Information Systems and Mining, 2010, pp. 53-56, doi: 10.1109/WISM.2010.123.

[6]. Y. Zhang and S. Li, "Privacy Preserving BIRCH Algorithm under Differential Privacy," 2017 10th International Conference on Intelligent Computation Technology and Automation (ICICTA), 2017, pp. 48-53, doi: 10.1109/ICICTA.2017.18.

[7]. S. Jebari, A. Smiti and A. Louati, "AF-DBSCAN: An unsupervised Automatic Fuzzy Clustering method based on DBSCAN approach," 2019 IEEE International Work Conference on Bioinspired Intelligence (IWOBI), 2019, pp. 000001-000006, doi: 10.1109/IWOBI47054.2019.9114411.

[8]. L. Yang, X. Guangqiang, L. Xiaomei and L. Hua, "Dependent function interval parameters training algorithm based on DBSCAN clustering," Proceedings of the 31st Chinese Control Conference, 2012, pp. 7709-7712.

[9]. R. M. Esteves, T. Hacker and C. Rong, "Competitive K-Means, a New Accurate and Distributed K-Means Algorithm for Large Datasets," 2013 IEEE 5th International Conference on Cloud Computing Technology and Science, 2013, pp. 17-24, doi: 10.1109/CloudCom.2013.89.

[10]. V. K. Dehariya, S. K. Shrivastava and R. C. Jain, "Clustering of Image Data Set Using K-Means and Fuzzy K-Means Algorithms," 2010 International Conference on

Computational Intelligence and Communication Networks, 2010, pp. 386-391, doi: 10.1109/CICN.2010.80.

[11]. S. R. Fitriyani and H. Murfi, "The K-means with mini batch algorithm for topics detection on online news," 2016 4th International Conference on Information and Communication Technology (ICoICT), 2016, pp. 1-5, doi: 10.1109/ICoICT.2016.7571914.

[12]. E. Slapak, J. Gazda, G. Bugar, M. Volosin, D. Horvath and T. Maksymyuk, "HetNet Spatial Topology Design Using Mini-Batch K-means Clustering," 2019 IEEE 15th International Conference on the Experience of Designing and Application of CAD Systems (CADSM), 2019, pp. 1-4, doi: 10.1109/CADSM.2019.8779242.

[13]. S. Babichev, B. Durnyak, V. Zhydetskyy, I. Pikh and V. Senkivskyy, "Application of Optics Density-Based Clustering Algorithm Using Inductive Methods of Complex System Analysis," 2019 IEEE 14th International Conference on Computer Sciences and Information Technologies (CSIT), 2019, pp. 169-172, doi: 10.1109/STC-CSIT.2019.8929869.

[14]. A. Omrani, K. Santhisree and Damodaram, "Clustering sequential data with OPTICS," 2011 IEEE 3rd International Conference on Communication Software and Networks, 2011, pp. 591-594, doi: 10.1109/ICCSN.2011.6014339.

[15]. D. Huang, C. Wang, J. Wu, J. Lai and C. Kwoh, "Ultra-Scalable Spectral Clustering and Ensemble Clustering," in IEEE Transactions on Knowledge and Data Engineering, vol. 32, no. 6, pp. 1212-1226, 1 June 2020, doi: 10.1109/TKDE.2019.2903410.

[16]. F. Nie, Z. Zeng, I. W. Tsang, D. Xu and C. Zhang, "Spectral Embedded Clustering: A Framework for In-Sample and Out-of-Sample Spectral Clustering," in IEEE Transactions on Neural Networks, vol. 22, no. 11, pp. 1796-1808, Nov. 2011, doi: 10.1109/TNN.2011.2162000.

[17]. Y. Zhang, X. Liu, W. Liu and C. Zhu, "Hybrid Recommender System Using Semi-supervised Clustering Based on Gaussian Mixture Model," 2016 International Conference on Cyberworlds (CW), 2016, pp. 155-158, doi: 10.1109/CW.2016.32.

[18]. H. Ye, L. Zheng and P. Liu, "Color detection and segmentation of the scene based on Gaussian mixture model clustering," 2017 7th IEEE International Conference on Electronics Information and Emergency Communication (ICEIEC), 2017, pp. 503-506, doi: 10.1109/ICEIEC.2017.8076615.