Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 10, October-2020

A Comprehensive Study of Clustering Algorithms in Data Stream

Swathi Agarwal[1]
Research Scholar and Assistant Professor in IT,
CVR College of Engineering – Hyderabad.

Dr. C. R. K. Reddy[2]
Professor and Head, Department of CSE,
MGIT–Hyderabad.

***Abstract*:- Clustering algorithms have been developed as an excellent algorithms to precisely break down the massive volume of information produced by modern applications. Specifically, their primary objective is to classify information into bunches to such an extent that data points gathered in a similar group are comparable. There is a huge volume of information in the zone of grouping and there have been activities to classify them into different types of groups. In any case, the main challenge of any clustering algorithm is, understanding of data dimensions, interpretation, combination of related data and taking relevant decisions. The main objective of any clustering algorithm is to reduce the issues listed. This paper gives a brief overview of existing bunching algorithms both from a hypothetical and an experimental point of view. From a theoretical viewpoint, we built up a clustering algorithm depending on the primary properties of the data like Data Size, Speed, Time Complexity, Data type, Cluster Shape, Partition Accuracy. Then we test the huge data collected, by applying a suitable classification method. Bunching is most widely used technique to partition data spaces and discover patterns. The behavior of different clustering algorithms is estimated through various interior and outer validity measurements, steadiness, runtime, and adaptability tests. Furthermore, in this paper a comprehensive study of clustering algorithms is performed to help the users in choosing the appropriate algorithm based on user's requirement.**

Keywords: Big Data, Clustering Algorithms, Data Stream, High-Dimensional Data.

1. INTRODUCTION

In the current scenario, as indicated by massive advancement to the improvement of web and online environmental innovations. For example, from vast and incredible information, user needs tremendous bulk of data and information step by step from various assets and administrations which were not accessible to humanity only a couple of decades back. Enormous amounts of information are created by and about individuals, things, and co-operations. Different gatherings contend concerning the expected benefits and expenses of breaking down knowledge from the Twitter, Verizon, Google, Facebook, 23andMe, Wikipedia, and each area wherever vast gatherings of persons who follows & store data. The data is originated to accessible, numerous web assets & administrations that are created up to serve their shoppers. Administrations and assets like sensing element Networks, Cloud Storages, Social Networks, then forth., turn out a massive volume {of data of data of knowledge} and, besides ought to administrate and use that information or some diagnostic components of the fabric. Though this vast of data is useful for persons & enterprises, the data may be risky too. During this manner, the foremost valuable quantity {of data of data of knowledge} or vast information have their inadequacies conjointly. They need immense stockpiles, and this volume makes activities, for instance, clarifying tasks, process activities, recovery activities, exceptionally troublesome, and massively tedious. One approach to conquer these difficult issues is to have enormous information bunched in a minimal organization that is yet an educational variant of the full knowledge. Such bunching methods mean to deliver a decent nature of groups. In this way, they would massively profit everybody from typical clients to scientists and individuals, as they could furnish an efficient device to manage enormous information, for example, simple frameworks (to identify digital assaults).

2. DIFFERENT CATEGORIES OF CLUSTERING ALGORITHMS

As there are such many clustering algorithms, this area presents an ordering structure that bunches the different clustering algorithms found in writing into classifications. The proposed classification structure will be established from the algorithm point of view that centers around the specialized categories of the general methodology of the clustering process. The characteristics of various clustering algorithms are as follows:

2.1. Partitioning algorithm: In such algorithms, all groups are decided instantly. Beginning gatherings are determined, collected and combine the data. In the end, the partitioning algorithms information objects into several allotments, where each parcel speaks to a group. These groups ought to satisfy the accompanying prerequisites:
(1) Each gathering must contain at any rate one item, and
(2) Each article must have a place with precisely one meeting. K-Means [29],[30], [35] algorithm, for example, by using the core data the remaining data is equally divided and arranged to the number juggling mean. In the K-medoids [30] algorithm, the object which is close to the inside to the bunches. There are numerous other partitioning algorithms, for example, partitioning algorithms such as K-modes [30], PAM [32], CLARA [32], CLARANS [32], and FCM [12],[23].

The k-medoids algorithm is like k-means; however, in the instance of k-means, the typical separation is determined; though, on account of the k-medoids algorithm, k-medoids of the group focuses are determined. Augmentations of the k-implies calculation on information streams have been talked about beforehand, for example, where the whole information stream is bunched in the STREAM [1],[2] algorithms like the LSEARCH [14]

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 10, October-2020

calculations/strategy which utilizes a k-median methodology and is utilized for information streams. Aggarwal et al. proposed a calculation called CluStream [8], [31], [38] that applies a k-means approach for bunching advancing information streams. In CluStream the online–disconnected system for bunching information stream is developed, this has been received for most of the information stream bunching algorithms.

2.2. Hierarchical algorithm: Data are sorted out in a hierarchical manner using different nodes starting from the root node. A dendrogram presentation to the datasets, where leaf hubs introduce singular information. The underlying bunch step by step isolates into a few groups as the chain of command proceeds. A hierarchical clustering strategy frames a chain of importance or tree of the bunches. There are two sorts of progressive techniques: (i) Agglomerative and (ii) Divisive. In the agglomerative methodology, a base up system is utilized, which at first thinks about each article as a gathering, at that point progressively, as the grouping process advances upwards, objects are gathered dependent on closeness. Along these lines, the top-down methodology, which is a troublesome methodology that at first gathers all articles into one gathering, and at that point recursively, as the calculation advances, it parts the meeting dependent on a closeness measure between the items. BIRCH [3],[20], CURE [4],[21], ROCK [5], and Chameleon [6],[22],[39] are the best reasonable calculations. ClusTree [7] produces and keeps up the chain of command of smaller scale bunches at various levels. The procedure proceeds until a halting measure is reached Habitually, the mentioned number k of groups. The different leveled technique has a significant downside. however, which identifies with the way that when a stage union is played out, this cannot be fixed. BIRCH, CURE, ROCK, and Chameleon are a portion of the notable algorithm of this class.

2.3. Density-algorithm: Here, information objects are isolated dependent on their locales of thickness, network, and limit. They are firmly identified with point-closest neighbors. A group, characterized as an associated thick part, develop towards any path that thickness prompts. This way, Density-based algorithms are equipped for finding bunches of self-assertive shapes. Likewise, this gives unsurpassed security against exceptions. Along these lines, the general thickness of a point is dissected to decide the elements of datasets that impact a specific information point. DBSCAN [26], [37] OPTICS [31], DBCLASD [26], DENCLUE [28], and DenStream [28], [31] are algorithms that utilize such a strategy to sift through outliers and find groups of self-assertive shape.

2.4. Grid-algorithm: The space of the information objects is partitioned into lattices. The primary preferred position of this methodology is its quick preparation time since it experiences the dataset once to figure the factual qualities for the matrix. The aggregated framework information makes network-based bunching strategies autonomous of the number of information protests that utilize a uniform lattice to gather provincial scientific knowledge, and afterward play out the grouping on the matrix, rather than the database legitimately. The presentation of a network put together technique depends on respect to the size of the environment, which is usually substantially less than the size of the database. In any case, for profoundly unpredictable information conveyances, utilizing a solitary uniform lattice may not be adequate to acquire the necessary bunching quality or satisfy the time prerequisite. Wave-Cluster [24], [32]and STING [24], [32], [37] are run of the mill instances of this classification.

The instance space has been divided between the limited number of cells called a network structure. All bunch activities are performed on the network structure (i.e., quantized space). As it processes measurable qualities for the lattices, the speed of the technique increments considerably. It is a result of aggregated lattice information, and matrix put together. Bunching is not needy in taking into consideration the number of information objects. Provincial information, determined over information objects mapped consistently on the environment, is additionally utilized for group development rather than information stream straightforwardly. Matrix-based strategies' exhibition is corresponding to the size of lattice, which needs less space in contrast with the actual information stream essentially. Algorithms, for example, Wave-Cluster and STING chips away at this strategy. A portion of the matrix-based approaches is utilizing thickness measures for bunching information streams. Such tactics have been referred to in the paragraph as the thickness matrix-based grouping techniques. Information contained in the focuses have turned out to be drawn into frameworks, and the following are lattices classified utilizing the thickness of the information point as reference. In such a technique that information focuses, you will require to think of being drawn into structures, and matrices are clustered dependent on the depth of data focuses. The CLIQUE algorithm [37][42] as the capability of automatically identifying subspaces of high dimensional data. Algorithms like D-Stream [8] and MR-Stream [8] have a place with texture framework-based algorithms.

2.5. Model-based algorithm: In model-based clustering the data is generated by a mixture of probability distribution in which each component represents a different cluster. A model is hypothesized for each cluster and find the best fit of data for the given model. EM [33] algorithm attempts to approximate the observed distribution of values based on mixtures of different distributions in different clusters. Each reflection belongs to each cluster with a certain probability.

3. COMPREHENSIVE STUDY

Table 1 gives the comprehensive study of different categories like partitioning algorithm, hierarchical algorithm, density-based algorithm, and gird-based algorithm. In each of the various classifications, different algorithms are being used. Partitioning algorithms consist of k-means, k-medoids, etc. Hierarchical algorithms such as CURE, BIRCH, etc. Density algorithms involve OPTICS, D-Stream, etc. and grid based. Different algorithms use different parameters like data set size, dimensional capability, input parameters, algorithm complexity and big data handling strategy. Also, in Table1 advantages and disadvantages of each algorithm are discussed.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 10, October-2020

Table 1
Comparison of various Clustering Algorithms

Categories	Name of Algorithm	Data Set Size	Dimens-ionality Capacity	No. of Input Param-eters	Algorithm Complexity	Advantages	Disadvantages	Applications
Partitioning	k-Means	Huge	Low	1	O (N pkt)	simple to implement, guarantees convergence, scales to large data sets.	Choosing K manually, dependent on initial values, clustering data of varying size and density, sensitive towards outliers.	Market segmentation, image compression, document clustering, pattern recognition, neural network, AI.
	k-Medoids	Small	High	1	O (k (N-k)^2)	easy to implement, converges in a fixed number of steps, less sensitive to outliers.	not suitable for clustering non-spherical groups and may obtain different results for different runs on the same dataset because the first k medoids are chosen randomly	
	FCM	Huge	Low	1	O(Npk^2)	gives best results for overlapped datasets, data point may belong to more than one cluster	problem in handling high dimensional datasets, sensitive to initialization, gives high membership values for outliers	
	CLARA	Huge	Low	1	O (kn^2+k (N - k))	deals with large datasets then PAM	Performance depends upon the size of dataset. A biased sample data may mislead into poor clustering of whole dataset.	
	CLARANS	Huge	Low	2	O(N^2)	Easy to handle outliers, more effective compared to PAM and CLARA.	Does not guarantee to give search to a localized area, not much efficient for large datasets, uses randomized samples for neighbors.	
	STREAM	Huge	Low	4	O (N pkt)	memory efficient, can handle outliers better than k-means	time granularity and data evolving, does not work well with high dimensional data	
	Parallel k-Means	Huge	Low	1	O (N pkt)	simple to implement, guarantees convergence, scales to large data sets.	Choosing K manually, dependent on initial values, clustering data of varying size and density, sensitive towards outliers.	
	CluStream	Huge	Low	5	O (N pkt)	Can distinguish clusters of different timestreams, fixed number of micro-clusters maintained, it used CF's to summarize data.	offline clustering is critical	
	k-Means (Single Pass)	Huge	Low	2	O(Npk)	scans the data only once	predefined number of clusters and centroids	
	k-Means (Mini Batch)	Huge	Low	3	O (n pkt)	a bit faster than k-means, small random batches of data are chosen as clusters	Increasing the number of clusters decreases the similarity	
Hierarchical	CURE	Huge	High	2	O (n^2 log(n))	robust to the presence of outliers and appropriate for handling large datasets.	ignores the information about the aggregate inter-connectivity of objects in two clusters	Applied science psychology, AI Social science, image compression, pixel, classification in images.
	BIRCH	Huge	High	2	O(Np)	finds a good clustering with a single scan and improves the quality with a few additional scans.	handles only numeric data, requires cluster count k and threshold T to compute the clusters, favors only clusters of spherical shapes and similar sizes	
	ROCK	Huge	Low	2	O (n^{2}log(n))	run on synthetic and real datasets, suitable to cluster data that have Boolean and categorical attributes	static modeling of the clusters to be merged, does not work on dynamic modelling.	

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 10, October-2020

	Chameleon	Huge	Low	2	O (n log(n))	can find clusters of diverse shapes, densities and sizes.	does not work on high dimensional data.	
Density-based	OPTICS	Huge	Low	2	O (N log(N))	does not need to maintain the epsilon parameter	requires more memory to maintain priority queue, requires more computational power.	Images of Satellite, crystallography of x-ray, Scientific literature.
	DBSCAN	Huge	Low	2	O (N log(N))	Can handle arbitrary shaped clusters, outliers. No need to define number of clusters in advance.	does not work well with high dimensional data, parameter selection is tricky, has problems of identifying clusters of varying densities.	
	DENCLUE	Huge	High	2	O (log (Np))	handle enormous data well, invariant against noise, clusters of arbitrary shapes can be found.	density parameters need to be selected carefully, not suitable for high dimensional data, cannot find efficient clusters for changeable density data.	
	D-Stream	Huge	Low	5	O (N log(N))	extension of DENCLUE, removes sparse grid cells to save memory and accelerate the mining process.	difficult to detect clusters with different densities	
	DenStream	Huge	Low	5	O (N log(N))	effectively handles the evolving data stream creates new micro-clusters of outliers	the pruning phase to remove the outliers is a time-consuming process, does not release memory space for micro-clusters.	
Grid-based	Wave Cluster	Huge	High	2	O(N)	handles large datasets efficiently, insensitive to the order of input, does not require specification of input parameters, discovers clusters with arbitrary shapes.	depends upon the granularity of cells	Social n/w, Biological n/w, image database exploration, medical imaging
	STING	Huge	High	2	O(K), k<<N K-grid cells, N-data points	Query independent, easy to parallelize, incremental update	The clustering result sensitive to the granularity (the mesh size), the high calculation efficiency at the cost of reducing the quality of clusters and reducing the clustering accuracy	
	CLIQUE	Huge	High	2	O (k^2+ np*k)	automatically finds subspaces of the highest dimensionality, insensitive to the order of records, has good scalability as the number of dimensions increases with data.	The accuracy of the clustering result may be degraded at the expense of simplicity of the method.	
Model-based	EM	Huge	High	3	O(npk)	It assures likelihood increases with each iteration, implementation of E-step and M-step are easy.	It has slow convergence and makes convergence to local optima only.	

In the algorithm complexity: t stands for Iterations, N stands for Data Size, n is Sample Size, p is Dimension, k stands for number of fuzzy clusters.

4. EXISTING CHALLENGES OF CLUSTERING DATA STREAM

To deal with unknown streaming data without user specific parameters. To save the memory space with a more concise summary of data. Number of clusters vary as the new data samples arrive. The separation of highly overlapped clusters and outlier detection. To handle data streams of mixed type (categorical, ordinal etc.) with several real-world application domains. Majority of existing algorithms depend on prior information, but structure of the data stream is unknown. The performance of most of the algorithms is tested on synthetic and real datasets with assumed noise, there are limitations in order to cater evolving data-streams in real-time. Need to investigate the context-based adaptive clustering methods to identify trends necessary for prediction of data. Handling data streams of social network and mobile applications is challenge in terms of processing capability and memory space optimization.

5. CONCLUSION

Over the period, a wide spectrum of clustering methods has been developed in the field of Data Mining, Statistics and Machine Learning. Each of the Clustering algorithm is tightly related to each other and exerts great challenge to the scientific disciplines regarding their selection for a given application. Also, these algorithms experience the ill effects

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 10, October-2020

of the security issues. To moderate such a problem, group bunching ought to be thought of. This paper discusses the basic and core idea of each commonly used clustering algorithms in the areas listed above. A brief analyses of advantages and disadvantages are listed in Table1. Clustering algorithms are categorized into: Partitioning, Hierarchical, Density-based, Grid-based, Model-based. Several algorithms of each category are explained, to give readers a systematical and clear understanding. This will help the user to select appropriate algorithm for specific domain. The classifying system is created by means of theoretical perspective that naturally suggest the appropriate algorithm.

Investigation permits us to make the accompanying inferences for vast information. No clustering algorithm performs well for all the assessment rules, and future work ought to be committed to address the downsides of each algorithm for taking care of meaningful information. EM and FCM clustering algorithms show excellent execution regarding the nature of the bunching yields, except for high-dimensional information. These algorithms experience the ill effects of high computational time complexity.

Clustering over streams is challenging as the underlying data distribution might evolve over time and number of clusters may vary. DENCLUE, OptiGrid, and BIRCH are reasonable bunching algorithms for managing enormous datasets, particularly DENCLUE and OptiGrid, which can likewise manage high dimensional information. There are lot of directions for future work in the stream clustering area.

REFERENCES

[1] Xu, R.Wunsch, D. “Survey of clustering algorithms”. IEEE Trans. Neural Networks, 645–678. [CrossRef] [PubMed] , 2005-16.

[2] O'Callaghan, L.Mishra, N.Meyerson, A.S.Guha and R. Motwani “Streaming-data algorithms for high-quality clustering”, proceedings of the 18th International Conference on Data Engineering, Washington, DC, USA, pp. 685–694,26 February–1 March 2002;.

[3] Zhang, T. Ramakrishnan, R. Livny and M. BIRCH, “A New Data Clustering Algorithm and Its Applications” Data Min. Knowl. Discov. pp.141–182. [CrossRef] ,1997.

[4] Guha, S. Rastogi and R. Shim, K., “CURE: An efficient clustering algorithm for large databases” ACM Sigmod Rec., pp.73–84, 1998, 27 [CrossRef].

[5] Guha, S.Rastogi and R.Shim, K. “ROCK: A robust clustering algorithm for categorical attributes”, Proceedings of the 15th International Conference on Data Engineering (Cat. No.99CB36337), Sydney, NSW, Australia, pp. 512–521, 23–26 March 1999.

[6] Karypis, G.Han and E.Kumar, V. “Chameleon: Hierarchical clustering using dynamic modeling”. computer 32, pp.68–75, 1999. [CrossRef]

[7] Philipp, K.Assent, I.Baldauf, C.Seidl, T. “The clustree: Indexing micro-clusters for anytime”,stream mining”, Knowl. Inf. Syst.29, pp.249–272, 2011.

[8] Guha, S.Meyerson, A.Mishra, N.Motwani and R.O'Callaghan, L. “Clustering data streams: Theory and practice” IEEE Trans. Knowl. Data Eng. 2003, 15, 515–528. [CrossRef]

[9] A. A. Abbasi and M. Younis, "A survey on clustering algorithms for wireless sensor networks," *Comput. Commun.*, vol. 30, no. 14, pp. 2826-2841, Oct. 2007.

[10] C. C. Aggarwal and C. Zhai, "A survey of text clustering algorithms," *Mining Text Data.*, pp. 77-128, 2012.

[11] A. Almalawi, Z. Tari, A. Fahad, and I. Khalil, "A framework for improving the accuracy of unsupervised intrusion detection for SCADA systems," *Proc. 12th IEEE Int. Conf. Trust Security Privacy Comput. Commun. (TrustCom)*, pp. 292-301, Jul. 2013.

[12] J. C. Bezdek, R. Ehrlich and W. Full, " FCM: The fuzzy c -means clustering algorithm, " *Comput. Geosci.*, vol. 10, no. 2, pp. 191-203, 1984.

[13] Xiao-Dong Wang, Rung-Ching Chen Fei Yan, Zhi-QiangZeng, and Chao-Qun Hong, “Fast Adaptive K-Means Subspace Clustering

[14] for High-Dimensional Data”, IEEE Access, vol 7, pp.42639-42651, 2019.

[15] TakanoriFujiwara, Jia-KaiChou, Shilpika and PanpanXu,LiuRen, “An Incremental Dimensionality Reduction method for Visualizing Streaming Multidimensional Data”, IEEE Transaction on Visualization and Computer Graphics, vol 26, pp.418-428, 2019.

[16] Cnor Fahy and Shexiang Yang (Senior member IEEE): “Dynamic Feature Selection for Clustering High Dimensional Streams”, IEEE Access, vol 7, pp.127-139,2019.

[17] PunitRathore, Dheeraj Kumar and James C.Bezdek, “A Rapid Hybrid Clustering Algorithm for Large Volumes of High Dimensional Data”, IEEE Transactions on knowledge and data engineering, vol 31, pp.641-654, 2019.

[18] K. Peng, V. C. M. Leung, and Q. Huang: Clustering approach based on mini batch kmeans for intrusion detection system over big data, IEEE Access 2018, vol 6, pp.11897-11906.

[19] UmeshKokate, ArvindDeshpande, Parikshit Mahalle and PramodPatil, “Data Stream Clustering Techniques, Applications, and Models: Comparative Analysis and Discussion”, Big data and cognitive computing, MDPI journal, pp.1-30, 2018.

[20] Bankov and Boris,” An approach for clustering social media text messages, retrieved from continuous data streams”, Research gate article 2018.

[21] Tian Zhang, Raghu Ramakrishnan and Miron Livny: “BIRCH An efficient data clustering Mehtod for very large datasbases”, ACM, p.103-114, 1996.

[22] Sudipto Guha, Rajeev Rastogi and kyuseok shim, “CURE An efficient clustering algorithm for large datasbases”, ACM, p.73-85,1996.

[23] George Karypis, Eui-Hang (Sam) Han and Vipin kumar, “Chameleon: Hirerachical clustering using dynamic modelling”, IEEE, pp.68-75, 1999.

[24] James C.Bezdek, Robert Enrlich and William Full, “FCM: The Fuzzy c-Means Clustering algorithm”, computer and geosciences , vol. 10, No.2-3, pp.191-203, 1984.

[25] Rui Xu and Donald Wunsch II, “Survey of Clustering Algorithms”, IEEE Transactions on Neural Networks, vol.16, No.3, pp.645-677, 2005.

[26] B.Biku and P.Praveen, “An Analysis of Clustering Algorithms for Big Data”, IJRITCC Journal, Vol.5, issue 6, ISSN:2321-8169, pp. 1294-1299.

[27] Rupanka Bhuyan and Samarjeet Borah, “A survey of some densitybased clustering techniques”, IEEE conference paper 2013.

[28] Dominik Sacha, Leishi Zhang, Michael Sedlmair and John A. Lee, “Visual interaction with dimensionality reduction:A struciured literature analysis”, IEEE transactions on visualization and computer gaphics,pp.1077-2626, 2016.

[29] Yangli-ao Geng, Qingyong and Chong-Yung Chi, “Local-Density subspace Distributed Clustering for High Dimensional Data”, IEEE Transaction on Parallel and Distributed Systems, Vol.31, No,8, pp.1799-1814, August 2020.

[30] Kristina P. Sinaga and Miin-Shen Yang, “Unsupervised K-Means Clustering Algorithm”, IEEE Access, Vol. 8, pp.80716-80727, 2020.

[31] Mohamed Cassim Alibuhtto and Nor Idayu Mahat, “Distance based k-means clustering algorithm for determining number of clusters for high dimensional data”, Direct Science Letters 9, pp.51-58, 2019.

[32] M.Shukla, Y.P.Kosta and M.Jayswal “A modified approach of OPTICS algorithm for Data Streams”, Engineering Technology & Applied Science Research, vol. 7, No. 2, pp.1478-1481, 2017.

[33] Preeti Baser, Dr. Jatinder kumar and R. Saini, “A Comparative Analysis of Various Clustering Techniques used for Very Large Datasets”, International Journal of Computer Science& Communication Networks, vol.3(4), pp. 271-275, 2013.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 10, October-2020

[34] Y. Naser Eldin, Hythem Hashim,Ali Satty and Samani A. Talab, "A Comparative Analysis between K-Means and EM Clustering Algorithms", vol.6, issue 7, pp. 12809-12816, July 2017.
[35] Jiang Wang, Cheng JHU, Yun Zhou and WEIMING Zhang, "From partition-based clustering to density based: Fast finding clusters with diverse shapes and densities in spatial databases", IEEE Access on Advanced data analytics for large-scale complex data environments, vol.6,pp. 1718-1729,2017.
[36] Kai Peng, Victor C. M. Leung and QINGJIA Huang, "Clustering approach based on Mini batch K-means for intrusion detection system over Big data", IEEE Access on Cyber-Physical social computing and networking, Vol.6, pp.11897-11906, 2018.
[37] XUYANG YAN, RAZEGHI-JAHROMI and EDWARD TUNSTEL:" A Novel streaming data clustering algorithm based on proportionate sharing", IEEE Access, Vol.7, pp.184985-184999, 2019.
[38] Bo Wu and Bogdan M. wilamowski:" A Fast Density and Grid based clustering Method for data with arbitrary shapes and noise", IEEE transaction on industrial informatics, Vol.13, No.4, pp.1620-1928, 2017.
[39] Ashish kumar, Ajmer Singh and Rajvir Singh, "An efficient hybrid-clustream algorithm for stream mining", 13th international conference on SITIS, 978-1-5386-4283-2/17, IEEE 2017.
[40] Georgr Karypis, Eui-Hong and Vipin kumar, "CHAMELEON: A Hierarchical Clustering Algorithm using Dynamic Modeling", IEEE Computer, Vol.32, Issue 8, Aug 1999.
[41] Stratos Mansalis, Eirini Ntoutsi, Nikos Pelekis and Yannis Theodoridis, "An Evaluation of Data Streams Clustering Algorithms", Article Statistical Analysis and Data Mining, pp.1-26, June 2018.
[42] Amineh Amini Member IEEE, Teh Ying Wah and Hadi Saboohi Member ACM, IEEE, "Density-Based Data Streams Clustering Algorithms: A Survey" Article in Journal of Computer Science and Technology, pp.116-141, Jan 2014.
[43] Suman and Pinki Rani Kurukshetra Iniversity, "A survey on STING and CLIQUE Grid Based Clustering Methods", International Journal of Advanced Research in Computer Science", Vol.8, No.5, pp. 1510-1512, May-June 2017.

AUTHOR PROFILE

Swathi Agarwal is a Research Scholar, Department of Computer Science and Engineering, University College of Engineering, Osmania University, Hyderabad. Presently she is working as Assistant Professor in Department of IT, CVR College of Engineering, Hyderabad, India. She received her M.Tech (CSE) from affiliated college of JNTU Hyderabad. Her main research interests are Data Mining, Scaling High Dimensional Data and High-speed Data Streams, Security, Privacy and Data Integrity, Machine Learning algorithms for prediction analysis in data mining.

Dr. C.R.K. Reddy is working as a Professor and Head, Department of CSE, Mahatma Gandhi Institute of Technology, Hyderabad, India. He received M.Tech (CSE) from JNTU Hyderabad and Ph.D. in Computer Science and Engineering from Central University of Hyderabad, Telangana. He has published and presented wide range of Research and Technical Papers in National and International Journals and Conferences. His main research interests are Program Testing, Software Metrics & Architecture, Mobile Adhoc Networks and Data Mining.