International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-8 Issue-11, September 2019

Data Stream Clustering Algorithms: Challenges and Future Directions

G. Sunitha, C. Jaswitha

Abstract*: In the fast growing world applications are generating data in enormous volumes called data streams. Data stream is imaginably large, continual, rapid flow of information and in data mining the important tool is called clustering, hence data stream clustering (DSC) can be said as active research area. Recent attention of data stream clustering is through the applications that contain large amounts of streaming data. Data stream clustering is used in many areas such as weather forecasting, financial transactions, website analysis, sensor network monitoring, e-business, telephone records and telecommunications. In case of data stream clustering most popularly used heuristic is K-means and other algorithms like K-medoids and the popular BIRCH are developed. The aim of the abstract is to review the developments and trends of data stream clustering methods and analyze typical DSC algorithms proposed in recent years, such as BIRCH, STREAM, DSTREAM and some more algorithms.

***Keywords* : Data stream, Clustering, BIRCH, K-Means.**

I. INTRODUCTION

Recently large amount of data is being generated from both hardware and software. The challenge here is to deal with the immense volume of data, this is because of the physical constraints of the current computational resources. An expanding attention towards dealing with this large, unbounded flow of data objects that are constantly producing at fast rates, the so-called data streams.

Clustering can be defined as grouping the objects that are obtained in such a way that similar characteristics exhibit from the objects that belong to the same group than the objects that belong to different groups. Clustering focuses on many research areas like statistical data analysis, data mining, pattern recognition, machine learning, information retrieval and image analysis. To solve the clustering issue particularly one algorithm is not enough hence it needs several distinct algorithms that vary appreciably in its own view of what really makes a cluster and the way to rapidly identify them.

II. OVERVIEW OF DATA STREAM CLUSTERING

The enormous amount of data generated by meteorological Analysis, sensor networks, computer network traffic monitoring and stock market analysis are collected and mined. These are all the applications that are the source to form data streams, just to name a few [1]. The datasets obtained from these applications are huge to fit in main memory hence moved to external storage device. Based on the above reason, it is expensive to perform random access on the datasets; hence the accessing method is to provide regular scans of the data to achieve performance efficiency.

Creating data clusters with the help of mining data streams eventually be a provocation because of several reasons: (i) single-scan clustering: clustering of data need to be achieved in a single pass because of the continuously arriving data streams (ii) limited memory: only limited memory is equipped to the clustering algorithm for dealing with never-ending, incoming, infinite data streams (iii) limited time: within a restricted time frame data clusters should be created in real time (iv) shape of clusters and cluster count: these features of data stream are unknown before processing (v) noisy data: clustering results was affected by the noise hence clustering algorithms has to overcome the noise that presents in the data stream (vi) Evolving data: the algorithm need to withstand for the changes happening in the data stream.

In recent times different approaches and characteristics of data stream clustering became an important discussion; based on it various algorithms and methods are discussed.

A. Frameworks of Data Stream Clustering

Many algorithms as well as methods are recently developed and many particulars on data stream clustering have been discussed. A review was conducted on recent and definitive DSC algorithms and given a detailed survey of thirteen DSC algorithms along with its features depends on two classes called object-based and another one is attribute-based [2]. One more approach of reviewing the DSC algorithms belongs to two different perspectives called as clustering by example as well as clustering by variable is even discussed.

The steps involved in object-based DSC algorithms are data abstraction step alias online component and clustering step alias offline component. The first step categorizes the data stream by using the data structures that handle memory and space limitation of data stream applications. One more point about data abstraction step is these streaming algorithms make use of outlier detection mechanisms to differentiate among actual outliers as well as cluster evolution, by thinking that the data stream distribution might differ with time. All DSC algorithms are interested to carry out object clustering; still some of the jobs go for attribute clustering which is indeed named as variable clustering.

Manuscript published on 30 September 2019.
***Correspondence Author(s)**
G. Sunitha, professor of CSE Department at Sree Vidyanikethan Engineering College, Tirupati, India**.**
Jaswitha, PG Scholar from the Department of CSE at Sree Vidyanikethan Engineering College, Tirupati, India.

Here traditional clustering algorithms are used on top of the transposed data matrix which is the process behind batch offline procedure and does not takes place in online processing of data streams. To find the attributes in the data streams that are alike to each other with respect to time is the main cause behind introducing attribute clustering algorithms. In a limited time how to organize and evaluate such a massive data is one of the major issue in data stream clustering. There exist some strategies for handling and analysing the data streams. The strategies include: 1) Processing in single-pass, 2) Evolving, 3) Online-offline phases.

III. DATA STREAM CLUSTERING METHODS

The five different methods of DSC are hierarchical, partitioning, grid-based, density-based and model-based [3].

A. Hierarchical Method (HM)

Hierarchical clustering techniques are sub divided into two methods specifically agglomerative and divisive. First method combines a set of 'n' objects to form a lot of general categories and the second method produces smaller clusters consecutively by dividing 'n' objects. The main functionality of HM is to combine data objects to form a hierarchical tree of clusters. In hierarchical clustering algorithms objects in the clusters are connected based on their distance. Various clusters will form at various distances. By using cluster proximity as a measure hierarchical clustering techniques determine where to combine and where to split different clusters.

Advantages:

- It is capable of dealing with any kind of similarity or distance.

Disadvantages:

- It is very ambiguous in case of termination criteria.
- High complexity.

B. Partitioning Method (PM)

K-median and K-means are the two techniques of the partitioning based data stream clustering. Partitioning algorithm categorise datasets into n clusters, where n is a parameter that is predefined. It continuously allocates objects from one group to another group to reduce the objective function. The widely used traditional clustering methods are k-means and k-medians. In partitioning clustering data is grouped as single partition rather than presenting it as a nested structure, therefore it is highly used when the data set is large.

Advantages:

- Implementation is easy.
- It creates clusters in an iterative manner.

Disadvantages:

- User need to predefine the number of clusters.
- Only spherical shaped clusters are determined here.

C. Grid-Based Method (GBM)

The processing time in the grid-based algorithms is independent to the total number of data points hence these algorithms are faster. Here the object space splits into limited number of cells that indeed gives the grid structure where the clustering operations take place. The main intention of these algorithms is to compute the data sets into cells and then working on the objects that comes under these cells.

Advantages:

- This method can handle noises.
- Fast processing time.

Disadvantages:

- Here the grid sizes need to be predefined.
- GBM is not preferred for high dimensional data.

D. Density-Based Method (DBM)

To perform clustering, DBM construct a density profile of data. Clusters occupy the higher density regions and the objects in the sparse regions helps in differentiating the clusters that are generally noisy. The main idea here is within the given radius each cluster sample has to maintain certain number of objects. DBSCAN is one of the recommended density-based clustering algorithms.

Advantages:

- This method is able to detect arbitrary shaped clusters.
- It can easily handle noises.

Disadvantages:

- Various parameters need to be defined in advance.
- This method is not efficient to work well in multi density data.

E. Model-Based Method (MBM)

Model-based method is one of the clustering types that run a hypothesized model for each cluster to decide which data is going to fit the model exactly. Model-based algorithms identify the best model that fit to the data. This MBM makes use of a tree structure created with the help of category function. It gives raise to a classification tree. Hence every node has a notion and its description which gives the details about the objects that comes under this nodes.

Advantages:

- Number of clusters can be defined automatically with the help of standard statistics.
- Efficient to deal with noises.

Disadvantages:

- Always it depends on hypothesized model or structure.

Retrieval Number: K19900981119/19©BEIESP
DOI: 10.35940/ijitee.K1990.0981119
Journal Website: www.ijitee.org

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

IV. DISCUSSION

At many areas now a days data

Table- I: Analysis of seven Data Stream Clustering Algorithms

S.No.	Algorithm	Cluster Algorithm	Cluster Shape	Cluster Problem	Outlier Detection	Data Structure	Reference. N0.
1	Birch	K-means	Hyper-sphere	Object	Density-based	Feature vector	[4]
2	Scalable K-means	k-means	Hyper-sphere	Object	-	Feature vector	[5]
3	Single-pass k-means	k-means	Hyper-sphere	Object	-	Feature vector	[6]
4	Stream	k-median	Hyper-sphere	Object	-	Prototype array	[7]
5	Stream LSearch	k-means	Hyper-sphere	Object	-	Prototype array	[8]
6	CluStream	k-means	Hyper-sphere	Object	Statistical-based	Feature vector	[9]
7	DenStream	Dbscan	Arbitraty	Object	Density-based	Feature vector	[10]

streams are generated from applications like network traffic, sensor networks, web click streams etc. To extract the hidden knowledge from the patterns is quite simple with the help of data mining algorithms if the data set is static and need to face many Issues if the dataset is large. To overcome those issues author conducts the detailed survey on all the data stream mining algorithms [1]. Hence it gives the relationship between the mining algorithms and find out the mining constraints and introduces a general model.

Clustering algorithms need to operate fast and continuous processing of data objects need to be done and they need to focus on time and memory also. Author conducts survey on all the algorithms that are related to data stream clustering. These algorithms cover all the details and give the overview about experimental methodologies [2]. Applications that use these algorithms are mentioned.

To extract the knowledge from the real time continuous data streams clustering is the best method to achieve this and numerous applications are generating data streams. Some challenges do exist they are limited time with less memory space and single scan clustering. Along with them the algorithm need to identify noise, detect arbitrary shape clusters. Author proposed five data stream methods called partitioning, hierarchical, density-based, grid-based and model-based [3].

A. Hierarchical-Based Clustering Algorithms

BIRCH was initially intended to implement mining on traditional data and data streams hence it is capable enough to work with large amount of data, which indeed produce micro clustering and macro clustering concepts. With the help of these concepts this algorithm get over from two major disadvantages occur in the hierarchical agglomerative clustering (HAC) algorithm called as scalability and failure to perform undo operation on the data which has been executed earlier. If the dataset is large, clustering the data is becoming an inefficient task. To overcome this issue author introduces an algorithm called BIRCH(Balanced Iterative Reducing and Clustering using Hierarchies) which follows CF-tree structure [4]. Based on the parameters like memory, Cluster quality the performance of the algorithm has been achieved as well as it deals with two real world issues.

One of the time series data stream clustering techniques is online divisive agglomerative clustering (ODAC). With the help of both agglomerative and divisive HM, this algorithm is capable to hold concept drift. To maintain the clusters with tree-like hierarchy it employs top-down strategy. To split each node a measure is used called correlation-based Dissimilarity and to enlarge the recognition of concept drift the agglomerative strategy is deployed with in the time series data. The Online Divisive Agglomerative Clustering (ODAC) system uses a top-down strategy to keep track of all the levels of clusters [11]. This algorithm frequently updates the time and memory consumption of clusters. Author implements this algorithm in such a way that at each arrival it figures out the difference between time series and updates the statistics. The time and memory consumption reduces when the structure of the tree grows hence it overcome the issue of solving all the dissimilarities at lower level.

B. Partitioning-Based Clustering Algorithms

One of the K-median-based clustering algorithms is called StreamLSearch algorithm which is useful for clustering effective data streams. It includes a two-step process that initiates with the discovery of sample with the help of STREAM algorithm and incase if the sample size is huge comparing with the output obtained from the predefined equation then LSEARCH algorithm is used. Each data chunk undergoes for processing and for the obtained cluster centres the LSEARCH algorithm is applied. Comparing to BIRCH, this algorithm is said to be upper-level incase of the sum of squared distance.

Table- II: Analysis of six Data Stream Clustering Algorithm

S.No.	Algorithm	Cluster Algorithm	Cluster Shape	Cluster Problem	Outlier Detection	Data Structure	Reference. N0.
8	ODAC	Hierarchical clustering	Hyper-ellipsis	Attribute	-	Correlation matrix	[11]
9	D-Stream	Dbsacn	Arbitrary	Object	Density-based	Grid	[12]
10	SWClustering	k-means	Hyper-sphere	Object	-	Feature vector	[13]
11	DGClust	k-means	Hyper-sphere	Attribute	-	Grid	[14]
12	ClusTree	k-median/dbscan	Arbitrary	Object	-	Feature vector	[15]
13	StreamKM++	k-means	Hyper-sphere	Object	-	Coreset tree	[16]

Among many one algorithm that follows k-median problem is called stream algorithm [7]. Clustering generally follows data stream model of computation that produces order of Points and continuously provide a better sequence of clustering just consuming little space and memory. Many applications use the data stream model namely web click stream analysis and multimedia data analysis.

If we take two parameters namely cluster quality and speed, then Birch algorithm is there to perform operations fast and if quality comes to picture it is not so efficient. So Author proposed an algorithm called StreamLSearch that helps in intrusion detection or target marketing [8]. The performance details of the algorithm are also verified empirically on real data streams.

Using the online micro clustering and an offline macro clustering component CluStream clustering method clusters the data. Collecting summary statistics among the data stream is the primary step, which indeed done by the online micro clustering component. To produce clusters the second component uses these statistics moreover as alternative inputs. Author proposed an algorithm that is better than stream called CluStream that uses pyramidal time frame to calculate any time horizon as familiar as required [9]. Cluster quality is high in case of this algorithm even for large data sets. This algorithm gives the characteristics and functionality of various varieties of clusters in different time frames. This feature has been achieved because of online statistical data collection component as well as offline analytical component. SWClustering algorithm finds out the clusters among the data streams upon the sliding window model. Making use of the exponential histogram cluster feature (EHCF) this algorithm is effective to capture the in-cluster evolution. EHCF generates synopses that help in calculating clusters. Hence SWClustering algorithm was designed based on K-means algorithm, it is not so efficient to find out the arbitrary shaped clusters apart from that poor in dealing with outliers. To sense the advancement of clusters over sliding windows author proposed an algorithm called SWClustering using this data structure is invented called the Exponential Histogram of Cluster Features (EHCF). Because of this data structure the quality of the clusters gets increased and by using the in-cluster synopsis cluster evolution can do better. This algorithm helps in reducing the computing resources that are needed in data stream clustering. This algorithm gives a better performance than the CluStream algorithm in the scenario of sliding window [13].

In the Euclidean space to cluster the data streams a k-means algorithm called STREAMKM++ is preferable. This algorithm is said to be best incase of producing quality of results when the number of cluster centres are large and it is not preferred if the running time is considered. One of the k-means clustering algorithms that the author introduces is called StreamKM++ that takes a small weighted sample belongs to data stream and solve it using the StreamKM++ algorithm [16]. Author stated with comparing to BIRCH this algorithm is better and with respect to quality it is equal to StreamLS algorithm. Comparisons between different algorithms have been done and the efficiency of StreamKM++ algorithm has been figure out.

C. Grid-Based Clustering Algorithms

To look after the data streams generated in sensor networks a distributed grid clustering algorithm was proposed called as DGClust. To summarize the data streams it makes use of the grid structure. Here local sensors are used to keep track the online discretization of the streaming data as well as it works with the fixed update of both time and space that helps in reducing dimensionality and communication burden. Network generated data points are collected and maintained an cluster structure for them generally it is done by multivariate stream and the author here introduces new algorithm called DGClust that reduces the dimensionality and the communication burdens [14]. Hence the cluster centres changes regularly with the help of this algorithm all the local site changes have been updated to the central server. Author done the experiment on synthetic data and achieves the robustness.

D. Density-Based Clustering Algorithms

Density-based algorithm with a two-phase scheme called as DenStream has been evolved to perform clustering on data streams. To form the summary of the data this algorithm makes use of the fading window model in the first phase.

Retrieval Number: K19900981119/19©BEIESP
DOI: 10.35940/ijitee.K1990.0981119
Journal Website: www.ijitee.org

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

Clustering result in second phase can be obtained by considering the summary of the data taken from first phase. Time complexity is high for this algorithm because of the huge time vector calculations but it is efficient to deal with arbitrary shaped clusters. To discover clusters in evolving data streams author invented an algorithm called DenStream. To handle clusters with arbitrary shape this algorithm uses dense micro-cluster as well as potential micro-cluster and outlier micro-cluster is even used. This algorithm guarantees the precision and it reduces the consumption of memory [10].

D-Stream algorithm has the ability to make instinctive and dynamic adjustments to the data clusters in the absence of user specification with reference to the number of clusters and the target time horizon. To map new incoming data this algorithm generates separated grids. To differentiate which data is old and which is recent a decay factor will be used and incapable of dealing with high-dimensional data. To deal with the issues of finding arbitrary shaped clusters and to find outliers author introduces an algorithm called D-Stream [12]. This algorithm follows density based approach. This algorithm captures the dynamic changes of the clusters with the help of density decaying technique. To get the better results among space and time efficiency of the process sound technique is designed to find and remove sporadic grids that are directed to the outliers.

To reduce multiple scanning of large databases author implements scalable clustering framework that applies to iterative clustering [5]. Here one scan is sufficient and it follows the k-means algorithm to justify the framework. This method identifies the regions in the memory and state which to keep and which to discard. This algorithm follows sampling based approach. This scheme stores the data based on the memory allocated and based on fitting the present clustering model.

For improving the scalability of the k-means algorithm that works on huge databases author proposed single-pass k-means algorithm [6]. Important concept here is the usage of buffers and preserving the points in the data set as compressed form. Basically time taken to clustering is very less, hence the data set used here is the real-world data that have been used in KDD data mining contest and the dataset is large.

Author introduces a parameter free index-based approach that handles fluctuating speed and supports any time clustering called as ClusTree. It keeps the values that help in finding the mean and variance of micro-clusters. Regardless of the speed this algorithm aggregates for faster or slower streams. From the experiments it is proved that ClusTree can maintain constant micro clusters and the granularity is exponential [15].

Author mainly focuses on analytical tools that help the data stream algorithms to remove the outliers that indeed cause the incorrect clustering model. It gives the clear idea about the designing of the algorithms that can handle numerical and categorical data. Hence reviewed the requirements that are needed to build the DSC algorithms and some recent algorithms have been discussed [17].

Huge data base mining can be achieved through two ways one is data stream clustering and another one is by directly using significant algorithms. Author stated that best technique is by using data stream clustering rather than mining the whole database [18].To extract the hidden information clustering techniques are used and unsupervised methods are used here. Various problem definitions that are tie up with these clustering algorithms have been discussed and solutions are provided.

In the fast moving world the data that applications are generating is huge hence the goal of the author is to cluster that data in a single pass or with less pass count. Algorithms that can handle large data streams have been discussed. From this survey author concluded for web click streams to do faster BIRCH algorithm is preferable and for sensitive data to ensure quality Stream algorithm is preferable [19].

Author compared ten popular algorithms and conducted the experiments on them; the data set used here is real and synthetic data. Inspecting streaming data grabs lot more attention now a days hence empirical comparison between all these streaming algorithms had been presented. Discussion on D-Stream and DenStream algorithms takes place here [20].

V. CONCLUSION

Various data stream clustering methods along with the 13 mostly used data stream clustering algorithms are confer in this work. Extraction of data more precisely in real time plays a crucial role hence the scope of discovering the data stream clustering algorithms is high. In future research can take place in developing the most efficient Data Stream Clustering algorithms that help in solving the problems of data stream clustering.

REFERENCES

1. Nguyen, Hai-Long, Yew-KwongWoon, and Wee-Keong Ng. "A survey on data stream clustering and classification." *Knowledge and information systems* 45, no. 3 (2015): 535-569.
2. Silva, Jonathan A., Elaine R. Faria, Rodrigo C. Barros, Eduardo R. Hruschka, Andre CPLF De Carvalho, and João Gama. "Data stream clustering: A survey." *ACM Computing Surveys (CSUR)* 46, no. 1 (2013): 13.
3. Mousavi, Maryam, Azuraliza Abu Bakar, and Mohammadmahdi Vakilian. "Data stream clustering algorithms: A review." *Int J Adv Soft ComputAppl* 7, no. 3 (2015): 13.B. Smith, "An approach to graphs of linear forms (Unpublished work style)," unpublished.
4. Zhang, Tian, Raghu Ramakrishnan, and MironLivny. "BIRCH: A new data clustering algorithm and its applications." Data Mining and Knowledge Discovery 1, no. 2 (1997): 141-182.
5. Bradley, Paul S., Usama M. Fayyad, and Cory Reina. "Scaling Clustering Algorithms to Large Databases." In KDD, vol. 98, pp. 9-15. 1998.
6. Farnstrom, Fredrik, James Lewis, and Charles Elkan. "Scalability for clustering algorithms revisited." ACM SIGKDD Explorations Newsletter 2, no. 1 (2000): 51-57.
7. Guha, Sudipto, Nina Mishra, Rajeev Motwani, and Liadan O'Callaghan. "Clustering data streams." In Foundations of computer science, 2000. proceedings. 41st annual symposium on, pp. 359-366. IEEE, 2000.
8. O'callaghan, Liadan, Nina Mishra, Adam Meyerson, SudiptoGuha, and Rajeev Motwani. "Streaming-data algorithms for high-quality clustering." In Data Engineering, 2002. Proceedings. 18th International Conference on, pp. 685-694. IEEE, 2002.
9. Aggarwal, Charu C., Jiawei Han, Jianyong Wang, and Philip S. Yu. "A framework for clustering evolving data streams." In Proceedings of the 29th international conference on Very large data bases-Volume 29, pp. 81-92. VLDB Endowment, 2003.

10. Cao, Feng, Martin Estert, WeiningQian, and Aoying Zhou. "Density-based clustering over an evolving data stream with noise." In Proceedings of the 2006 SIAM international conference on data mining, pp. 328-339. Society for Industrial and Applied Mathematics, 2006.
11. Rodrigues, Pedro Pereira, Joao Gama, and Joao Pedroso. "Hierarchical clustering of time-series data streams." IEEE transactions on knowledge and data engineering 20, no. 5 (2008): 615-627.
12. Chen, Yixin, and Li Tu. "Density-based clustering for real-time stream data." In Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 133-142. ACM, 2007.
13. Zhou, Aoying, Feng Cao, WeiningQian, and Cheqing Jin. "Tracking clusters in evolving data streams over sliding windows." Knowledge and Information Systems 15, no. 2 (2008): 181-214.
14. Gama, Joao, Pedro Pereira Rodrigues, and Luís Lopes. "Clustering distributed sensor data streams using local processing and reduced communication." Intelligent Data Analysis 15, no. 1 (2011): 3-28.
15. Kranen, Philipp, Ira Assent, CorinnaBaldauf, and Thomas Seidl. "The ClusTree: indexing micro-clusters for anytime stream mining." Knowledge and information systems 29, no. 2 (2011): 249-272.
16. Ackermann, Marcel R., Marcus Märtens, ChristophRaupach, KamilSwierkot, Christiane Lammersen, and Christian Sohler. "StreamKM++: A clustering algorithm for data streams." Journal of Experimental Algorithmics (JEA) 17 (2012): 2-4.
17. Barbará, Daniel. "Requirements for clustering data streams." *ACM sIGKDD Explorations Newsletter* 3, no. 2 (2002): 23-27.
18. Khalilian, Madjid, and Norwati Mustapha. "Data stream clustering: Challenges and issues." *arXiv preprint arXiv:1006.5261* (2010).
19. Guha, Sudipto, Adam Meyerson, Nina Mishra, Rajeev Motwani, and Liadan O'Callaghan. "Clustering data streams: Theory and practice." *IEEE transactions on knowledge and data engineering* 15, no. 3 (2003): 515-528.
20. Carnein, Matthias, Dennis Assenmacher, and Heike Trautmann. "An empirical comparison of stream clustering algorithms." In *Proceedings of the Computing Frontiers Conference*, pp. 361-366. ACM, 2017.

AUTHORS PROFILE

Dr. Gurram Sunitha is a professor of CSE Department at Sree Vidyanikethan Engineering College, Tirupati, India. She received her Ph.D. degree in CSE from S.V.University, Tirupati. She has 19 years of experience in academia. Her research interests include Data Mining, Spatio-Temporal Analytics, Machine Learning and Artificial Intelligence. She has 7 patents and 3 books published to her credit. She has published around 30 papers in reputed journals and conferences. She has been serving as reviewer for several journals; and has served on the program committees and co-chaired for various international conferences.

C. Jaswitha is a PG Scholar from the Department of CSE at Sree Vidyanikethan Engineering College, Tirupati, India. She received B.Tech degree in Computer Science and Engineering from JNT University, Anantapur in 2014. Her research interests include Computer Networks and Machine Learning.