Analysis of clustering Algorithm for outlier Detection in Data Stream

[1]Hemaxi P. Jani, [2]Prof. D. J. Ramoliya
[1]Research Scholar, [2]Assistant Professor,
Department of Computer Engineering, Darshan Institute of Engineering & Technology Rajkot, Gujarat, INDIA

***Abstract* - Outlier detection is an important data mining task, aiming at the discovery of elements that show significant diversion from the expected behavior. Data stream mining has poses different challenges for outlier detection like concept drift, huge size and evolutionary data from data streams. Clustering techniques for data stream which helps to create a similar group of data are used to cluster the similar data items in data streams and also used to detect the outliers from data stream, so they are called as cluster based outlier detection. Which provides advantages like less memory requirement, less time consumption and it results exact outliers. In data streams if an object does not obey the behavior of normal data object is called as outlier. We proposed a new framework for outlier detection in data streams, which is combination of Neighbour based outlier detection approach and clustering based approach for outlier detection in data streams which provides better output in terms of true outliers from data streams.**

***Keywords*- application area of outlier detection; data stream; K-Means algorithm; outlier detection algorithm in data stream.**

I. INTRODUCTION

Nowadays, more data is generated from the different applications as number of user is increasing day by day. Data is generated and stored in database which is increasing at fast rate due to technology and hardware improvements. For example humans store different types of data like documents, images, songs, movies, scientific data and many other data into database. There is need to find meaningful information in the form of useful patterns, association, relationships among these data because these large database may contain both useful data and non-useful data. Data mining is the process of mining meaningful information, discovering new patterns, identifying relationship among data etc. from databases, flat files, spatial database, data repositories, temporal database, data stream, transactional database and World Wide Web.

Data mining is the process of analyzing the data from different perspective and summarizing it into useful information. Different applications like internet traffic, communication network data, sensor network data, online banking transaction, scientific data social data like emails, web click streams generates infinite volume of data in continuous and incremental manner which is called as data stream. Data stream is different than traditional data as data stream is having characteristics of being evolutionary in nature, massive, fast changing and potentially infinite. Traditional data processing methods don't work well with data streams so there is need to use different techniques for processing stream data. Outliers are the data points Which shows significant diversion from other data points or which is different from the regular or normal data points.

Outlier data points show different behavior than expected behaviour or same behaviour as other data points. . They are generated due to different reasons like malicious activity in network, instrumental error, environmental changes, and errors by human. Sadik, Shiblee, and Le Gruenwald [9] classified outliers into three categories. Type 1, outliers are the data point which is different and isolated individual data points with respect to all other data points in data set. It is easy to find type 1 outliers. Type 2 outliers are the point which is isolated from other data points in the same context. Context of data point refers to semantic relationship among data points. Difference between type 1 and type 2 outliers is that type 1 outlier is isolated from all other data points in dataset rather than same context. Type 3 outliers are a subset or a group of data points which appears as outliers with respect to entire dataset. Data points are not outlier with respect to data point of the same subset or group. There are various algorithms which are used for clustering data sets.

II. RELATED WORK

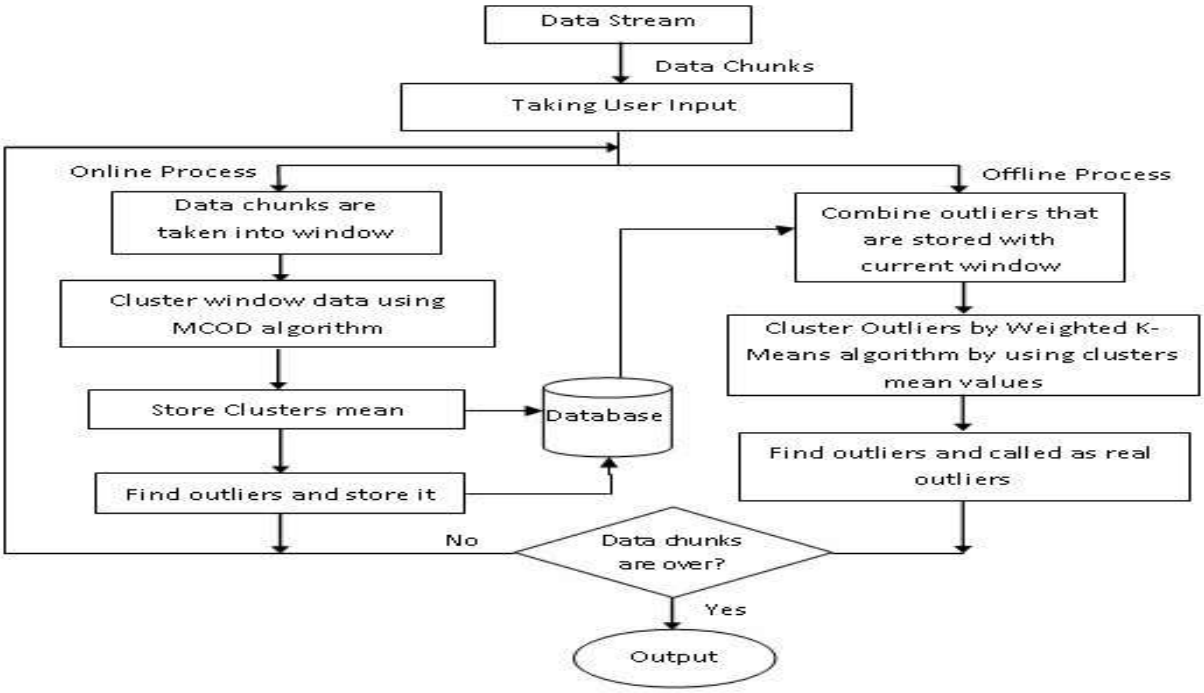
Many researchers have proposed different studies in recent years in order to detect outlier in data stream. Some of the research work is overviewed in the next few paragraphs.

In this paper,we have developed an effective and efficient method ,called Clustreamfor clustering large data streams. The method has clear advantages over recent techniques which try to cluster the whole stream at one time rather than viewing the stream as a changing process over time. has been proposed by Charu C.Aggarwal, Jiawei Han, Jianyong Wang [1].

Daniel Barbara et al [2] , we have summarized a simple set of conditions for data stream clustering algorithms. An important requirement ,often ignored by algorithm designers, is the need for a clear separation of outliers in the data stream, as sufficient number of these might indicate that a change in the clustering model is needed.

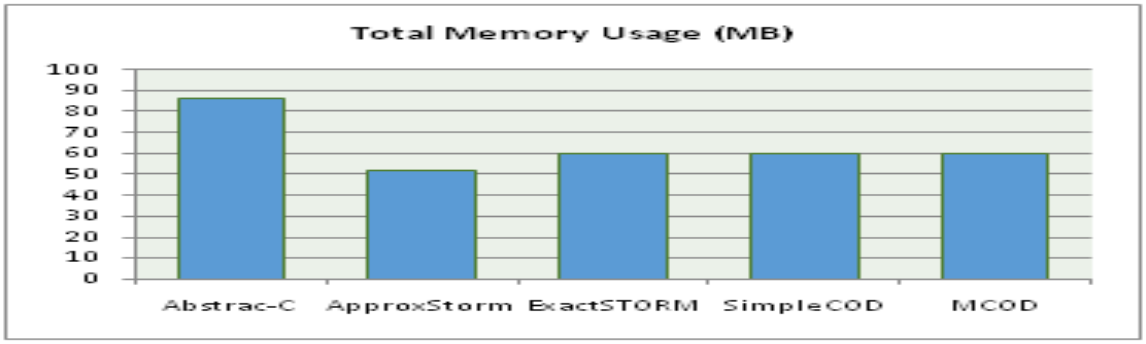
Renxia Wan, Lixin Wang et al [7] in 2011, we present the algorithm MStream which is designed for clustering over mixed evolving data stream. It uses micro-clusters to store the clustering statistical information, and then combines these micro-clusters into macro-clusters and the final clusters by iterations

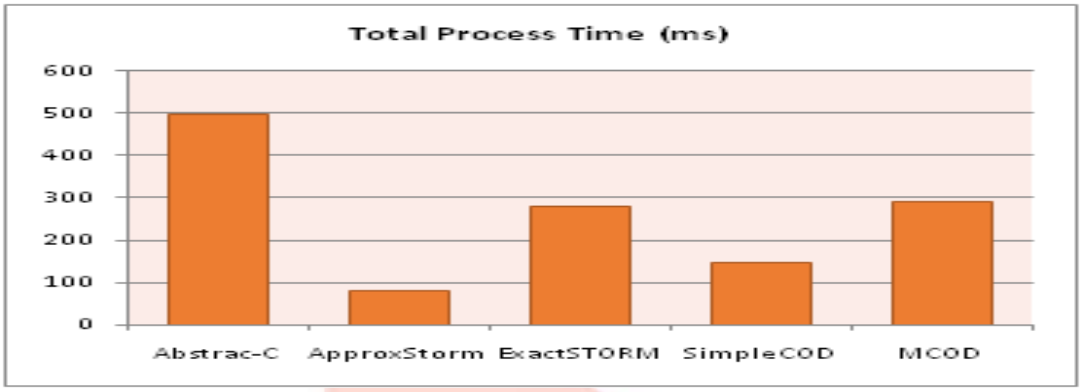
III. PROPOSED METHODOLOGY



- Step 1: Input Data Streams in the form of data chunks.
- Step 2: Take input from user like radius, window size and threshold value.
- Step 3: Online Process ()
  - Step 3.1: Data chunks are inserted into window.
  - Step 3.2: Cluster window data using clustering algorithm (Find Outliers by using MCOD algorithm).
  - Step 3.3: Store mean value of clusters of this window.
  - Step 3.4: Find outliers and store it.
  - Step 3.5: End.
- Step 4: Offline Process ()
  - Step 4.1: Outliers that are detected from all previous windows are combined with current window and is taken as input to offline process.
  - Step 4.2: Cluster outliers by weighted k-means algorithm.
  - Step 4.3: Find outliers and call it as real outliers.
  - Step 4.4: Store real outliers.
  - Step 4.5: End.
- Step 5: Continue step 3 and 4 until all data chunks are processed.
- Step 6: End.

IV. EVALUATION





V. CONCLUSION & FUTURE WORK

Outlier detection over streaming data is an important research problem in data mining community. Detecting outlier is important because it contains useful information which may lead for further research in domain. In this paper, we discussed about requirements for clustering data streams and application areas of outlier detection that uses outlier detection for various task.

FUTURE DIRECTION

- Design Proper GUI.
- Algorithm like Simple COD, Abstract-C, Approx-STROM, Exact-STROM,MCOD another outlier detection approaches perform well in outlier detection as well as data processing but there is a need to developing a process which considers minimum tradeoff between accuracy, time and space requirement.

VI. REFERENCES

1. Aggarwal, C.C., Han, J., Wang, J., Yu, P.S.,"A framework for clustering evolving data streams," In: Proceedings of the 29th international conference on Very large data bases, VLDB Endowment (2003) 81–92.

2. Aggarwal, C.C., Han, J., Wang, J., Yu, P.S, "A framework for projected clustering of high dimensional data streams," In: Proceedings of the Thirtieth international conference on Very large data bases - Volume 30. VLDB '04, VLDB Endowment (2004) 852–863.

3. Aggarwal, C.C., Han, J., Wang, J., Yu, P.S, "On high dimensional projected clustering of data streams," Data Mining and Knowledge Discovery 10 (2005) 251–273.

4. Babcock, B., Datar, M., Motwani, R., O'Callaghan, L. "Maintaining variance and k-medians over data stream windows," In: Proceedings of the twenty-second ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems. PODS '03, New York, NY, USA, ACM (2003) 234–243.

.

5. Ng, W., Dash, M.,"Discovery of frequent patterns in transactional data streams," In: Transactions on Large-Scale Data- and Knowledge-Centered Systems II. Volume 6380 of Lecture Notes in Computer Science. Springer Berlin / Heidelberg (2010) 1–30.

6. Wan, L., Ng, W.K., Dang, X.H., Yu, P.S., Zhang, K.: "Density-based clustering of data streams at multiple resolutions," ACM Transactions Knowledge Discovery Data 3(3) (2009) 1–28.

7. Zhou, A., Cao, F., Qian, W., Jin, C,"Tracking clusters in evolving data streams over sliding windows," Knowledge and Information Systems 15(May 2008) 181–214.

8. Zhu,Y.,Shasha,D,"Efficient elastic burst detection in data streams," IN: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining.KDD '03, New York, NY, USA,ACM (2003) 336-345.

9. A.Amini and W.Teh Ying, "A comparative study of density-based clustering algorithms on data streams Micro-clustering approaches," in Intelligent Control and Innovative Computing,ser. . Lecture Notes in Electrical Engineering, S. I. Ao, O. Castillo, and X. Huang, Eds. Springer US, 2012, vol. 110, pp. 275–287.

10. Rana Poonam, Deepika Pahuja, and Ritu Gautam,"A Critical Review on Outlier Detection Technique," International Journal of Science and Research, Volume 3 Issue 12, pp. 2394-2403, December 2014.

11. Tian Zhang, Raghu Ramakrishnan, Miron Livny. "BIRCH: An Efficient Data Clustering Method for Very Large Databases." ACM SIGMOD, pp.103-114, 1996.

12 Yogitaa, Durga Toshniwala. "A Framework for Outlier Detection in Evolving Data Streams by Weighting Attributes in Clustering." 2nd International Conference on Communication, Computing & Security, pp. 214–222, ICCCS-2012.

13. Yogita and Durga Toshniwal. "Unsupervised Outlier Detection in Streaming Data Using Weighted Clustering." World Academy of Science, Engineering and Technology, Vol:6, Nov 2012.

14 Jiang, Mon-Fong, Shian-Shyong Tseng, and Chih-Ming Su. "Two-phase clustering process for outliers detection." Pattern recognition letters 22, no. 6, pp. 691-700, 2001.