A Modified Approach of OPTICS Algorithm for Data Streams

M. Shukla
Department of Computer Engineering
Marwadi Education Foundation &
R. K. University, Rajkot, India
madhu.ce@gmail.com

Y. P. Kosta
Department of Computer Engineering
Marwadi Education Foundation
Rajkot, India
ypkosta@gmail.com

M. Jayswal
Department of Computer Engineering
Marwadi Education Foundation
Rajkot, India,
mjmeghnesh@gmail.com

***Abstract*-Data are continuously evolving from a huge variety of applications in huge volume and size. They are fast changing, temporally ordered and thus data mining has become a field of major interest. A mining technique such as clustering is implemented in order to process data streams and generate a set of similar objects as an individual group. Outliers generated in this process are the noisy data points that shows abnormal behavior compared to the normal data points. In order to obtain the clusters of pure quality outliers should be efficiently discovered and discarded. In this paper, a concept of pruning is applied on the stream optics algorithm along with the identification of real outliers, which reduces memory consumption and increases the speed for identifying potential clusters.**

Keywords-two phase; cluster quality; clustering technique; pruning; time and space complexity; threshold value

I. INTRODUCTION

Traditional data mining methods are not that successful in case of huge data streams, as off-line mining is not applicable. There are some requirements for clustering algorithms. Fast processing of data points and identification of outliers must be clear and precise [1]. Data uncertainty is an added issue. Different data type should be treated differently which is also an issue. An arbitrary shape of clusters makes hard to distinguish the accurate shape of the cluster [2]. There are many different applications like network traffic analysis, sensor network, internet traffic etc. that produce stream data [20]. Random sampling, sliding window, histograms, multi-resolution methods, sketches and randomized algorithm are some basic data sampling techniques for mining data streams [13]. Classification of stream data is possible with algorithms such as the Hoeffding tree, the Concept Adaptive Very Fast Decision tree (CVFDT), the Very Fast Decision Tree (VFDT), and the classifier ensemble approach.

Accuracy, efficiency, compactness, separateness, purity, space limitation and cluster validity are an important issue in the aspect of clusters quality. Different types of clustering methods like partitioning, hierarchical methods, model based, density based, grid based, constraint-based and evolutionary methods etc. are used for clustering stream data. Various algorithms are developed for clustering in data streams. Micro-clustering algorithms for data streams are as follows: Den Stream Algorithm, Stream Optics Algorithm, HDD Stream Algorithm [19] etc. Density grid-based algorithms for data streams are as follows: D-Stream Algorithm, MR-Stream Algorithm, DENGRIS Algorithm[19] etc. Table I gives a basic mapping of several existing algorithms that contains the description of their advantages and disadvantages.

Clustering is a key task in data mining. There are various other additional challenges by data streams on clustering such as one pass clustering, limited time and limited memory. Along with this, finding out clusters with arbitrary shapes is very much necessary in data stream applications. Density-based clustering method is of significant importance in clustering data streams, as it has the tendency to discover arbitrary shape clusters as well as outliers. In density based clustering, a cluster is defined as areas of higher density than the remaining data set. Clustering algorithms requires very tedious calculations for detecting the outliers. Handling noisy data, limited time and memory, handling evolving data, and handling high dimensional data are also to be considered. An outlier is defined as a data point which shows abnormal behaviour to the system and it is application dependent. In stream data mining the data points are huge in number thus it may be possible that during clustering of such data points few of the data points which does not belong to clusters or does not take part in clustering due to its distance from the cluster will be termed as outliers. It is required to remove such data points. As data generation is continuous and fast, a structure should be established to handle the mining process. Clustering provides a solution to such types of issues in stream data mining

II. PROPOSED ARCHITECTURE

In this paper a modification is applied on the stream optics algorithm by applying a pruning method and setting a threshold value cut off points for data dynamically. The extension of the most basic density algorithm (DBScan) is OPTICS which is based on the ordering point in the stream data mining. Its concept is to continuously increase the given cluster till the density in the neighborhood cross some of the threshold value. For each data point in a given cluster from the group of the cluster, the neighborhood of given radius has to contain at least a minimum number of points.

One of the important advantages of this method is that it can find clusters of arbitrary shapes and can be used to filter out

noise. It considers clusters as dense regions of the objects in the data space which are separated by low-density regions. For interactive and automatic cluster analysis this algorithm determines an augmented cluster ordering. An ordering of clusters is done to obtain basic clustering information and deliver the intrinsic clustering architecture. The core distance represents the smallest value from the core. The reachability distance considers the greater value of the core distance of the second object and the Euclidean distance between the two objects. This creates an ordering of the objects in a database and stores the suitable reachability distance and core distance for each object. The major drawback of the OPTICS method is that if there is no core object the reachability distance between two objects is undefined. The same basic scheme is applied for the stream OPTICS algorithm modified with the addition of an iterative property of the threshold value and with the concept of pruning in order to optimize and time complexity.

TABLE I. APPROACH, ADVANTAGES, AND DISADVANTAGES OF EXISTING ALGORITHMS

Algorithm	Approach	Advantages	Disadvantages
STREAM [1]	Partitioning	Space and time complexity is low.	It does not have flexibility of computing the clusters at user defined time periods.
Clu-STREAM [2]	Partitioning & Hierarchical	It provides flexibility of computing the clusters at user defined time periods.	Does not support concept drift and gives the clusters of arbitrary shapes.
HPStream [3]	Partitioning & Hierarchical	High dimensional projected clustering of data stream	An Average number of projected dimension parameters and number of clusters requires detail domain knowledge.
DEN-Stream [4]	Density Based	Gives arbitrary shape clusters.	Deleting and merging of the micro clusters does not allow release of any memory space.
D-Stream [5]	Density and Grid Based	It supports density decaying and monitors evolving behaviour for real time data streams.	Does not support handling of multiple dimensional data
E-Stream [6]	Hierarchical	High performance then other algorithm	User complexity of setting high number of parameters
DBSCAN [7]	Density Based	Can analyze the cluster for large dataset	Inputting the Parameter setting is very much difficult
Stream-Optics [8]	Density Based	Plotting of cluster structure based upon time.	For cluster extraction it is not supervised technique
MR-Stream [9]	Density Based	Improves the performance of the clustering.	In high dimensional data, it lacks its working
HDD-Stream [10]	Density Based	High Dimensional data is clustered.	It takes more amount of time in searching the neighbor clusters.
DENGRIS [11]	Density Based	Using sliding window model the distribution of recent most records are captured precisely	No evaluation to show its effectiveness compared with other state off art algorithms.
SOMKE [12]	Density Based	Use in non-stationary data efficiently and effectively.	Cannot handle unbalance data.
LeaDen-Stream [13]	Density and Grid Based	It supports density decaying and monitors evolving behaviour for real time data streams.	Does not support handling of multiple dimensional data
POD – Clus [14]	Model Based	Supports Concept Evolution and data fading	Computation and updating of pairwise distance takes a lot of time in data streams
BIRCH [15]	Hierarchical	Overcomes the inability to undo what was done in previous step	Does not perform well if the cluster is not in spherical shape.
SPE-Cluster [16]	Partitioning Based	Solve the problem of specifying the number of cluster.	Cannot Discover Arbitrary Shape clusters.
COBWEB [17]	Model Based	It identifies the outliers effectively.	It does not provide compact representation.

Data stream input in the form of small chunks is called data chunks. Here a windowing concept is used because stream data are huge. Data are fitted into a window frame and then passes ahead. Different parameters like window size, threshold value, and radius are set by the user. After that, an one-way online process is used in which data chunks are fitted into the window and then the clustering process is applied. In the online phase, micro-clustering is performed and basic DBScan algorithm is used. The micro clustering is done by the selection of the centroid with nearest data point's i.e. cluster mean value of object. The output of these will be the data points with k cluster and n objects. The clustering is done based on distance. These will now be input to the offline phase with the parameters like core distance, epsilon and fixed value of min point, generating distance. The proposed scheme is depicted in Figure 1.

In the offline phase, the Stream Optics algorithm is used for clustering. Macro clustering takes place and the data points form clusters of good quality. Two phase work is required because due to the nature of data streams. Points which have not yet been part of any clusters are distinguished by giving them weight. For every new iteration, this value will be incremented to make sure it is part of real outlier. Based on the application of a threshold value, points termed as real outliers are detected. Thus the node or the data points, which are outliers, will be pruned off which will reduce the memory consumption and the time taken for generation of the potential clusters. These will improve the

purity of the potential clusters. For each threshold value set the whole dataset is been checked with the prime motto of maintaining the quality of the clusters. The algorithm is broken down in steps in Figure 2.

III. RESULTS & DISCUSSIONS

Various parameters like cluster purity, number of clusters, SSQ, threshold, memory and time consumption are evaluated. Then these evaluated parameters are used for comparison and performance evaluation. Net Beans is used for the simulation studies in our research work. The Forest Cover Type data set and sensor data set are used for evaluation and algorithms are computed for 50000 and 100000 data records respectively. Different threshold values were tested and an optimum value was chosen in each case (3 for the the Forest Cover Type data set and 14 for the sensor data set). Tables II to V sum up the results.

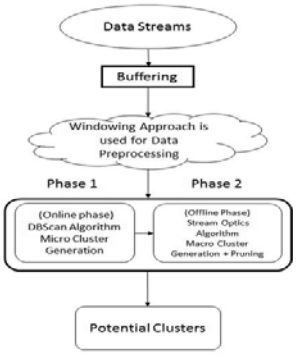

Fig. 1. The proposed scheme

Potential Cluster ()
Step 1: Input Data Streams in the form of data chunks.
Step 2: Automatically takes input like radius, window size.
Step 3: Online Process ()
  Step 3.1: Data chunks are inserted into window.
  Step 3.2: Cluster window data using clustering algorithm (Micro-Clustering Using DBscan algorithm based upon partitioning method).
  Step 3.3: Store mean value of clusters of this window.
  Step 3.4: Find outliers
  Step 3.5: End.
Step 4: Offline Process ()
  Step 4.1: k clusters with n objects are obtained and input to offline phase is the parameters like fixed min points, generating distance and Epsilon for performing Stream Optics (Macro Clustering) with current window and user need to give the threshold value.
  Step 4.2: Cluster window frame using Stream Optics algorithm
  Step 4.3: Assign weight to outliers (If belongs to cluster set 0 else set 1) and set threshold value as per data set provided and check practically for different values.
  Step 4.3: Discover outliers and discard outliers if total weight of data point is equal to the threshold value.
  Step 4.4: Apply pruning on the generated clusters
  Step 4.5: Find out Potential clusters and store potential cluster
  Step 4.6: End.
Step 5: Continue step 3 and 4 until all data chunks are processed.
Step 6: End

Fig. 2. The proposed algorithm

TABLE II. OVERALL RESULT SUMMARY OF FOREST COVER DATASET WITH DIFFERENT PARAMETER AT VARIOUS THRESHOLD VALUES BEFORE PRUNING.

Threshold Value	Maximum Time (Sec)	Maximum Memory (Mb)	Number of Cluster	Sum of Squared Error	No. of Noise Points	Purity
1	44427	171.46	22	1.31	19	93.02
2	44327	163.09	23	1.21	19	92.13
3	**44162**	**138.99**	**21**	**1.46**	**12**	**91.29**
4	43966	128.56	25	1.35	15	91.35
5	44636	137.21	24	1.42	10	93.82
6	44172	138.18	21	1.38	17	92.95

TABLE III. OVERALL RESULT SUMMARY OF FOREST COVER DATASET WITH DIFFERENT PARAMETER AT VARIOUS THRESHOLD VALUES AFTER PRUNING

Threshold Value	Maximum Time (Sec)	Maximum Memory (Mb)	Number of Cluster	Sum of Squared Error	No. of Noise Points	Purity
1	43783	145.46	16	1.15	23	95.53
2	43643	132.09	16	1.02	22	95.68
3	**43580**	**110.99**	**13**	**1.02**	**17**	**95.87**
4	43480	100.56	14	1.00	21	95.17
5	43939	109.21	14	1.06	16	95.42
6	43172	111.18	9	1.18	20	95.34

TABLE IV. OVERALL RESULT SUMMARY OF SENSOR DATASET WITH DIFFERENT PARAMETER AT VARIOUS THRESHOLD VALUES BEFORE PRUNING

Threshold Value	Maximum Time (Sec)	Maximum Memory (Mb)	Number of Cluster	Sum of Squared Error	No. of Noise Points	Purity
2	15499	53.06	21	1.27	17	92.33
10	15370	57.06	23	1.47	20	93.80
13	15434	52.06	25	1.33	13	91.08
14	**15498**	**58.06**	**24**	**1.09**	**16**	**93.61**
15	15156	54.06	25	1.29	12	93.33
2	15499	53.06	21	1.27	17	92.33

TABLE V. OVERALL RESULT SUMMARY OF SENSOR DATASET WITH DIFFERENT PARAMETER AT VARIOUS THRESHOLD VALUES AFTER PRUNING

Threshold Value	Maximum Time (Sec)	Maximum Memory (Mb)	Number of Cluster	Sum of Squared Error	No. of Noise Points	Purity
2	14938	27.06	17	1.02	19	95.74
10	14797	26.21	9	1.09	24	94.94
13	14704	27.12	7	1.07	19	95.52
14	**14938**	**27.06**	**8**	**1.02**	**19**	**96.29**
15	14578	27.06	6	1.08	16	95.31
16	14704	26.08	6	1.07	15	94.96

IV. CONCLUSION

Handling data streams shows increased complexity due to their constant, huge and potentially infinite nature. Working with data streams challenges the memory, space, time and handling changes along with speed and multiple source of data generation. Thus, the algorithms used for offline data mining and management may prove insufficient in such application and variations may be required. Such a variation of the OPTICS algorithm is proposed in this paper. Simulations are performed, results are discussed and an overall improvement is documented.

REFERENCES

[1] L. O'Callaghan, N. Mishra, A. Meyerson, S. Guha, R. Motwani "Streaming-Data Algorithms for High-Quality Clustering", 18th

International Conference on Data Engineering, pp. 685-694, February 26-March 1, 2002

[2] C. C. Aggarwal, J. Han, J. Wang, P. S. Yu, "A Framework for Clustering Evolving Data Streams", International Conference on Very Large Databases, Vol. 29, pp. 81-92, 2003

[3] C. C. Aggarwal, J. Han, J. Wang, P. S. Yu, "A Framework for Projected Clustering of High Dimensional Data Streams", Thirtieth International Conference On Very Large Data Bases, Vol. 30, pp. 852-863, 2004

[4] F. Cao, M. Ester, W. Qian, A. Zhou, "Density-based Clustering over an Evolving Data Stream with Noise", SIAM International Conference on Data Mining and Secure Data Management (SDM), Vol. 6, pp. 328-339, 2006

[5] Li-xiong, H. Hai, G. Yun-fei, and C. Fu-cai, "rDenStream: A Clustering Algorithm over an Evolving Data Stream", International Conference on Information Engineering and Computer Science, pp. 1-4, December 19-20, 2009

[6] K. Udommanetanakit, T. Rakthanmanon, K. Waiyamai, "E-Stream: Evolution-Based Technique for Stream Clustering", Lecture Notes in Computer Science, Vol. 4632, pp. 606-616, 2007

[7] C. Dharni, M. Bnasal, "An improvement of DBSCAN Algorithm to analyze cluster for large datasets", IEEE International Conference on MOOC Innovation and Technology in Education (MITE), pp. 42-46, 2013

[8] M. Ankerst, M. M. Breunig, H. Kriegel, J. Sander, "OPTICS : Ordering Points To Identify the Clustering Structure", ACM SIGMOD, Vol. 28, No. 2, pp. 49-60, 1999

[9] L. Wan, W. K. Ng, X. H. Dang, P. S. Yu, and K. Zhang, "Density-based clustering of data streams at multiple resolutions", ACM Transactions on Knowledge Discovery from Data (TKDD), Vol. 3, No. 3, pp. 1-28, 2009

[10] I. Ntoutsi, A. Zimek, T. Palpanas, P. Kröger, H. Kriegel, "Density-based projected clustering over high dimensional data streams", Society of Industrial and Applied Mathematics (SIAM) International Conference on Data Mining, pp. 987-998, 2012

[11] A. Amini, T. Y. Wah, "DENGRIS-Stream: A density-grid based clustering algorithm for evolving data streams over sliding window", International Conference on Data Mining Computer Engineering, pp. 206-211, 2012

[12] Y. Cao, H. He, H. Man, "SOMKE: Kernel density estimation over data streams by sequences of self-organizing maps", IEEE Transactions on Neural Networks Learning Systems., Vol. 23, No. 8, pp. 1254-1268, 2012.

[13] A. Amini, T. Y. Wah, "LeaDen-Stream: A Leader Density-Based Clustering Algorithm over Evolving Data Stream", Journal of Computer Communication, Vol. 1, No. 5, pp. 26-31, 2013

[14] P. P. Rodrigues, J. Gama, J. P. Pedroso, "ODAC: Hierarchical Clustering of Time Series Data Streams", IEEE Transaction on Knowledge Data Engineering , Vol. 20, No. 5, pp. 615-627, 2008

[15] T. Zhang, R. Ramakrishnan, M. Livny, "BIRCH: An Efficient Data Clustering Databases Method for Very Large", ACM SIGMOD Record, Vol. 25, No. 2, pp. 103-114, 1996

[16] E. Keogh, S. Chu, D. Hart, M. Pazzani, "An online algorithm for segmenting time series", International Conference on Data Mining, pp. 289-296, 2001

[17] Kavita, P. Bedi, "Clustering of Categorized Text Data Using Cobweb Algorithm", International Journal Computer Science and Information Technology Research, Vol. 3, No. 3, pp. 249-254, 2015

[18] M. Khalilian, N. Mustapha, "Data Stream Clustering: Challenges and Issues", International Multi Conference of Engineers and Computer Scientists, Vol. 1, Hong Kong, March 17-19, 2010

[19] A. Amini, T. Y. H. Saboohi, "On Density-Based Data Streams Clustering Algorithms:A Survey", Journal of Computer Science and Technology, Vol. 29, No. 1, pp.116-141, 2014

[20] M. Shukla, Y. P. Kosta, P. Chauhan, "Analysis and evaluation of outlier detection algorithms in data streams", IEEE International Conference on Computer, Communication and Control (IC4), pp. 1-8, September 10-12, 2015

[21] P. Chauhan, M. Shukla, "A review on outlier detection techniques on data stream by using different approaches of K-Means algorithm", IEEE International Conference on Advances in Computer Engineering and Applications (ICACEA), pp. 580-585, 2015

[22] M. Kamber, J. Han, Data Mining: Concepts and Techniques, Second edition, Elsevier, 2001