REVIEW ARTICLE OPEN ACCESS

Study of Density Based Clustering Techniques on Data Streams

Darshali Thoriya*, Madhu Shukla**
*(Department of Computer Science, Marwadi Education Foundation Group of Institutes, Rajkot-3)
** (Department of Computer Science, Marwadi Education Foundation Group of Institutes, Rajkot-3)

ABSTRACT
Data streams are generated by many real time systems. Data stream is fast changing and massive. In stream data mining traditional methods are not efficient so that many methodologies developed to stream data processing. Many applications require data into groups based on its characteristics. So clustering on data streams is applied. Clustering of non liner data density based clustering is used. Review of clustering algorithm and methodologies is represented and evaluated if they meet requirement of users. Study of density based clustering algorithm is presented here because of advantages of density based clustering method over other clustering method.

***Keywords* -** Data Streams, Clustering, Density Based Clustering, Algorithms, Non Linear Data Base component

I. Introduction

Nowadays organization and scientific fields have very large data base. Fields like astronomy, telecommunication operations, stock market application, social media, website analysis, banking, e-commerce, network data, network intrusion detection, weather monitoring, planetary remote sensing, meteorological data, phone records this all are examples of large data . This kind of data is known as stream data. Stream data is ordered, massive, fast changing, continuous and likely infinite database. To find out pattern, find changes in data, making better decision and discovering new facts mining of stream data is necessary. Traditional data mining methods is not useful for mining of data streams. So for that many algorithm and methodologies is developed to mine the stream data. For processing stream data effectively we need new techniques, data structure and algorithms because we do not have enough space to store this large amount of data. Random sampling, sliding window, histograms, multi resolution methods, sketches and randomized algorithm are basic data structure and methodologies for mining data streams [13]. Classification of stream data is not efficiently possible with the simple classification algorithm of data mining. Classification of stream data mining is possible with Hoeffding tree algorithm, very fast decision tree (VFDT), concept adaptive very fast decision tree (CVFDT) and classifier ensemble approach. Web clickstream, stock market analysis and network intrusion detection clustering of stream data is required.

There are some requirements for any clustering algorithm. Algorithm’s representation must be compact because lengthy representation is not always affordable. Processing of new data points must be fast. Identification of outliers must be clear and fast. What to do with the outliers this decision should be taken simultaneously [1]. There are some challenges and issues in data stream clustering. Accuracy, efficiency, compactness, separateness, space limitation and cluster validity are important issue in the aspect of quality of clusters. In data streams data is uncertain so this also becomes a challenge for clustering. Different data type should be treated differently this is also an issue. Arbitrary shape of clusters makes hard to distinguish the accurate shape of cluster [2]. Clustering methods are as follows: partitioning methods, hierarchical methods, model based methods, density based methods, grid based methods, constraint based methods and evolutionary methods.

II. Algorithms of Clustering

1. BIRCH

BIRCH algorithm builds CF-tree hierarchical data structure, which is a height balance tree. In CF-tree each node contains CF-vector which contains information of node which is under that node [1]. BIRCH algorithm gives best cluster which has available main memory and it decreases I/O requirement. Complexity of this algorithm is O(N) because algorithm is one pass algorithm. Birch is not useful for arbitrary shapes because here node of CF-tree have limited space. BIRCH algorithm does not provide compact representation because more memory is required to process the data. It takes considerable time amount to minimize the I/O requirements. BIRCH algorithm have some reserved space to handle outliers and it timely re-evaluates outlier to check if they can absorbed in the current tree without increasing the size [1]. Fig 1 shows how the BIRCH algorithm works. Here in BIRCH Algorithm there are 4 phases. In first phase of BIRCH algorithm data are loaded into main memory by building CF tree. Subsequent phases become fast, accurate and less order sensitive. Second phase is optional. Here data is condensed and build a smaller CF-tree. Third phase is the global clustering phase where clustering algorithm is applied on CF-leafs. Forth phase is optional and offline phase. In fourth phase additional passes are applied over the dataset & reassign data points to the closest centroid from phase 3. Refining is optional. In this way BIRCH algorithm works. Figure shows flow diagram of BIRCH algorithm. Where four phases are displayed with its input and working of individual phase. As per diagram we give data as input and get final cluster as output.

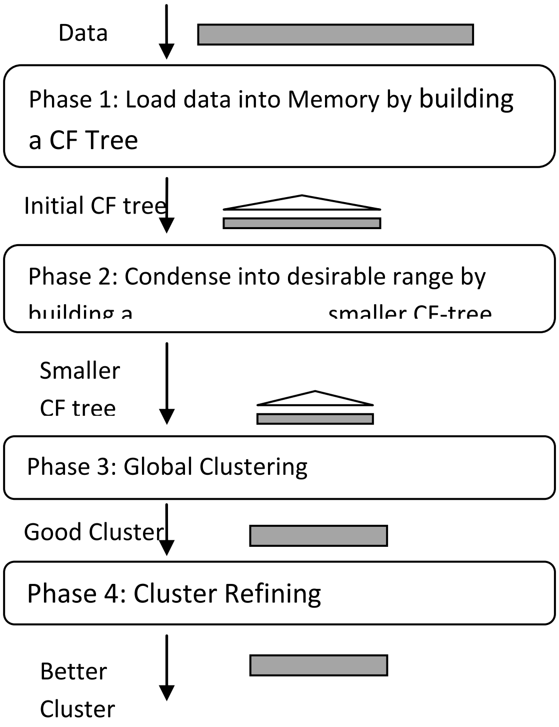
Data
Phase 1: Load data into Memory by building a CF Tree
Initial CF tree
Phase 2: Condense into desirable range by building a smaller CF-tree
Smaller CF tree
Phase 3: Global Clustering
Good Cluster
Phase 4: Cluster Refining
Better Cluster

Fig.1. Working of BIRCH Algorithm

2. COBWEB

COBWEB is clustering algorithm which uses incremental clustering technique. For classification and tree formation it uses category utility function. Hierarchical clustering model is used as classification tree. In this tree each node has description of objects classified under that node. These algorithms have certain disadvantages because COBWEB does not provide compact representation and time complexity of adding new point is high. But identification of outlier is effectively done by COBWEB [1].

3. STREAM

STREAM algorithm makes the batches of points which fit in main memory and then process the data streams. STREAM uses the LOCALSEARCH algorithm which runs in linear time in proportional to number of point [1]. STREAM algorithm provides compact representation but it has higher time complexity. Identification of outlier is not done by STREAM which is considerable disadvantage.

4. Fractal Clustering Algorithm

FC algorithm uses the parameter of fractal impact. This algorithm places the points in the cluster which have minimum fractal impact [1]. FC has capacity of finding clusters in arbitrary shapes. This algorithm has compact representation and lower time complexity. Identification of outlier is done effectively by this algorithm. So impact of FC algorithm is better than BIRCH, COBWEB and STREAM.

5. CluStream

In the CluStream idea of BIRCH and STREAM are used. In online and offline component concept of micro clustering and macro clustering is applied. This algorithm uses CF-vector to store data summary. It employs pyramid structure for organizing clusters. This algorithm has acceptable higher efficiency and accuracy.

6. SPKMEANS Algorithm

SPKMEANS is the improved version of K-Means, Spherical K-Means algorithm. In SPKMEANS algorithm all vectors are normalized and distance measure between them is cosine similarity [2]. In SPKMEANS there is concept to set cluster center such that it make both uniform and minimal the angle between components.

7. Density Based Clustering

Density based clustering method is useful because it can find clusters in arbitrary shapes and it can handle noisy data efficiently. And it is one scan algorithm because examination of raw data is done only once [3]. This all are advantages of density based clustering algorithms. In density based clustering clusters are defined as areas of higher density than remainder of the data sets. One cluster is separated from other clusters by lower density regions. There are many algorithms developed under the density based clustering methods DBSCAN, GDBSCAN, OPTICS and DenClu.

8. OCTS and OCTSM algorithm

OCTS is online clustering algorithm and OCTSM is extended version of OCTS. There are two phases in the OCTS algorithm 1) offline initialization phase 2) online clustering process. Topic signature model is generated in offline phase [2]. Xract tool is used in this algorithm's offline phase. OCTS algorithm provides good cluster quality. OCTSM is developed to improve the accuracy of the algorithm. In OCTSM process of merging is introduced. For that MergeFactor is used. Merge process gives second chance to inactive cluster.

9. New Weighted Fuzzy Cmeans Algorithm:

New weighted fuzzy CMeans algorihm (NWFCM) is developed to overcome disadvantages of FCM and FWCM. This algorithm is useful for unsupervised learning. Two concepts first weighted mean from the nonparametric weighted feature extraction (NWFE) and second cluster mean from discriminate analysis feature extraction (DAFE) is used in (NW-FCM) [2]. This algorithm has better classification accuracy and stability than other fuzzy logic algorithms. This algorithm is used for high-dimensional multiclass pattern reorganization problem.

III. DENSITY BASED CLUSTERING

Density based clustering method have capability to detect cluster in arbitrary shapes and it can also handle noise. Main algorithms of density based clustering are DBSCAN, OPTICS and DENCLUE [11]. Where DBSCAN is main algorithm which represent the density based clustering method effectively. DBSCAN is density based clustering algorithm which defines the high density regions into clusters. This algorithm is able to find clusters of arbitrary shapes and it can handle noisy data. Parameters of DBSCAN are radius of neighborhood and the number of objects in neighborhood (MinPts). Algorithm of DBSCAN is as follows: first this algorithm randomly selecting a point and checking whether the □-neighbourhood of the point contains at least MinPts points [3]. Than if yes than, it is considered as a core point and a new cluster is created and if not, it is considered as a noise point. DBSCAN iteratively adds the data points, which do not belong to any cluster and are directly density reachable. This algorithm continues until all points are visited and not any new point is added. In initial stage of DBSCAN, it performed on simplified dataset as shown in figure [3]. The dashed line displays a two dimensional distance radius and the minimum points threshold is 1.

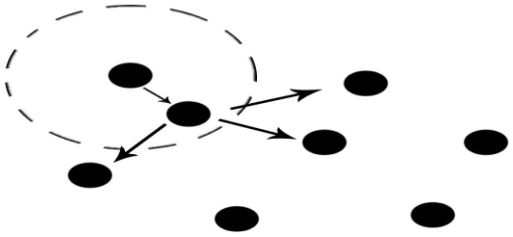

Fig 3 DBSCAN on simplified dataset

IV. DENSITY BASED CLUSTERING ALGORITHM ON DATA STREAMS

1. DenStream algorithm

This algorithm has ability to handle noise. This algorithm use fading window model for clustering the stream data. The algorithm expands the micro-cluster concept as core micro-cluster, potential micro-cluster, and outlier micro-cluster in order to distinguish real data and outliers [11]. It is based on the online-offline framework.

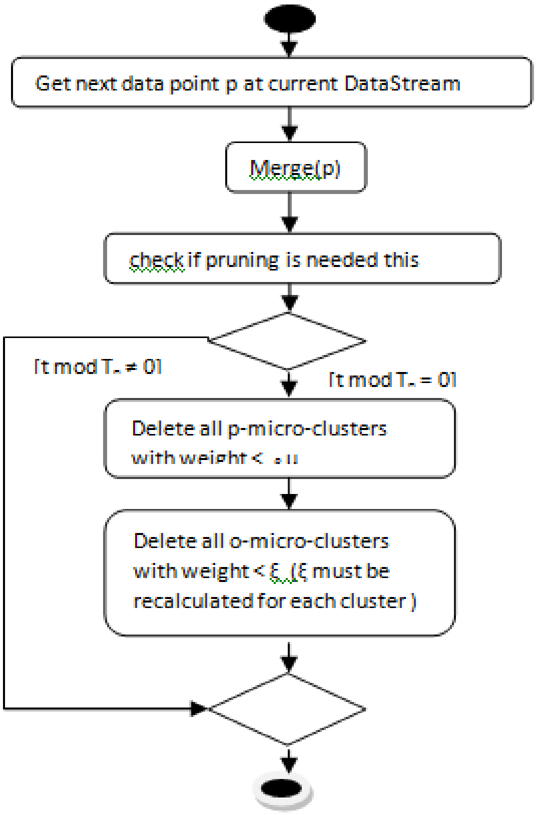
Get next data point p at current DataStream
Merge(p)
check if pruning is needed this
[t mod T. ≠ 0]
[t mod T. = 0]
Delete all p-micro-clusters with weight < βμ
Delete all o-micro-clusters with weight < ξ (ξ must be recalculated for each cluster)

Fig 3 Algorithm of DenStream

This algorithm use fading window model for clustering the stream data. Core-micro-cluster is defined as CMC (W, C, R), W is the weight, C is center and R is radius. Algorithm for DenStream is as follows: DenStream (DS,$\epsilon, \beta, \mu, \lambda$): first define the minimal time span for micro cluster than get the next point at current time from data streams than merging process is done on data streams. In merging process first we try to merge point into nearest micro cluster if it does not fit into micro cluster than we try to merge it with outliers and check the weight of current micro cluster [11]. This process gets repeated if request of cluster is arrived and generate the cluster. DenStream algorithm does not release any memory space by either deleting a micro-cluster or merging two old microclusters [11].

2. StreamOptics

OPTICS- Ordering Points To Identify The Clustering Structure.

StreamOptics extends OPTICS algorithm for data streams using concepts of micro clusters. It computes augmented cluster ordering for automatic and interactive analysis. This algorithm maintains proper ordering of the objects in given database and also store the core-distance and reach ability distance for each object respectively [12]. Stream algorithm also uses potential micro-cluster and outlier micro-cluster. StreamOptics algorithm are as follows: first the neighbourhood of each potential micro-cluster is determined than ordered list of potential micro-clusters is made based by using reachability distance and then produces a reachability plot that represents the micro-cluster structure using OPTICS algorithm [12]. This algorithm provides three dimensional plot which shows evolution of cluster structure with respect to time. But it is not useful for cluster extraction and manual checking is needed in three dimensional plots.

3. *MR-Stream Algorithm*

MR-stream is algorithm which applies density based cluster on multiple resolution data stream. This algorithm improves the performance by running offline component in steady time [10]. MR-Stream algorithm uses tree like data structure. In MR-Stream algorithm there are two phases online phase and offline phase. Algorithm for online component of MR-Stream is as follows: first step is to initialize tree than check if data stream is active or not. Second step is to read the records than using parameters two methods updateTree and pruneTree is applied [10]. Next step is to sample number of nodes in T and after that update and prune method applied again. After the online component offline component generate the cluster cell. In MR-Stream algorithm a memory sampling method is proposed to trigger right time to offline component [10]s. This algorithm makes relation between odes of tree and evaluated cluster. This algorithm improves the performance by introducing memory sampling method which define the right time for online and offline clusters. MR-Stream put the sparse grids and merges them for consideration as a noise cluster, this we can consider a disadvantage.

4. *D-Stream Algorithm*

This algorithm is developed to clustering data in real time. There are mainly two challenges for which algorithm is developed. First it is not desirable to treat the data stream as a long sequence of data. And second, it is not possible to retain the density information for every data in database [3]. Algorithm which displays overall process of D-Stream algorithm is as follows. In algorithm first initialize empty hash table to grid list than read record of data stream than determine the density of the grid. After that check whether density grid g is in the grid list or not. If density grid is not in the grid list than insert density grid in the grid list. Then update the characteristic vector of density grid. Then method of initial clustering and adjust clustering is called. Timestamp we should increase continuously. In the process of initial clustering, they update the density of all grids which are active to the current proposed time. After than cluster procedure become same as standard method of density based clustering. Now the process of dynamically adjusting cluster is as follows: Here grid list is given as input. If density grid g is sparse grid than delete g from its cluster and label g as NO_CLASS. If cluster becomes unconnected than split into two clusters [3]. If g is dense grid than among all neighbouring grids of g find out grid whose cluster has the largest size than check for grid h, if it is sparse, dense or traditional grid. In this way process of adjust clustering takes place [3]. Figure 4 shows the illustration of density grid. Here data is assigned to grids. Grids are weighted by regency of points added to it and each grid is associated with one label. After the offline processing we get the final cluster. This algorithm effectively handles the outliers. It proposes density decaying function to adjusting the clusters in real time data. But for finding time interval gap it gives minimum time but time interval gap depends upon many parameters which is disadvantage of the D-Stream algorithm.

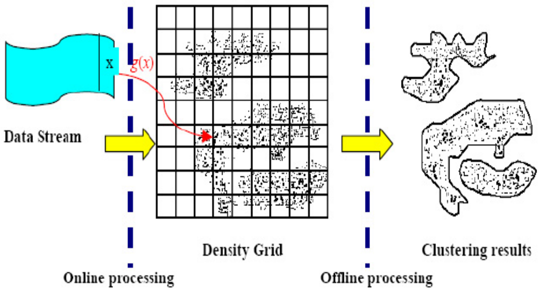

Fig 4 Illustration of the use of density grid

5. *HDDStream Algorithm*

This algorithm is for clustering high dimensional data streams. In the online phase of this algorithm keep the summary of the points and dimensions and offline phase generates the final cluster based on given projected cluster [8]. Main three parts of HDDStream algorithm: first initial set of microcluster is extracted. After that online microcluster maintenance as new point has arrived and old points are expire due to ageing. Adding method is used in this phase where update dimension preference and find the closest micro cluster are there and final cluster is extracted on demand [8]. HDDStream can cluster high dimensional data effectively but in the pruning time it checks only micro cluster weights whereas micro cluster weight should be checked also. This is the disadvantage of the HDDStream algorithm.

6. *DENGRIS Algorithm*

DENGRIS-Stream (a DENsity GRId-based algorithm for clustering data streams over sliding window). This algorithm maps each input into grids. It has online and offline components like other algorithms. Online components read data streams into grids and maps the data to related grid cells [4]. It updates the feature vector. Offline component adjust cluster in sliding window. Algorithm for DENGRIS is as follows: First initialize empty tree for grid list then for active data stream read record from the data streams and determine the grid cell. If grid cell not in the grid list than insert it into grid list otherwise update feature vector of grid cell after that generate initial cluster, adjust cluster and remove expired grids is applied [4]. RemoveExpiredGrids process detects and removes expired grid based on the grid's timestamp. Two other processes are used in the DENGRIS algorithm. Which are Ganerate Initial Cluster and Adjust Cluster. Generate Initial Clusters initializes the clusters by considering each dense grid as unique cluster identification [4]. Adjust-Clusters the sparse grids are removed from the grid list while clustering the dense grids. In adjust cluster method neighbouring grids are checked in respect to large cluster [4]. Then assign grid into the cluster with respect to its size. Still DENGRIS algorithm is not compared with any other current algorithm so for efficient evaluation comparison is needed.

7. *SOMKE Algorithm*

SOMKE algorithm is developed for kernel density estimation over streaming data and data streams which is based on sequence of self organizing maps. SOMKE algorithm creates SOM sequence and merges SOM sequence [5]. SOMKE algorithm has good performance compared to other Algorithms. SOMKE consist main three steps first creates the SOM sequence which is built to summarize the information of **W**t (weight) needed for KDE analysis and second merging the SOM sequence. SOM sequences are merged based on the kullback-Leibler divergence [5].

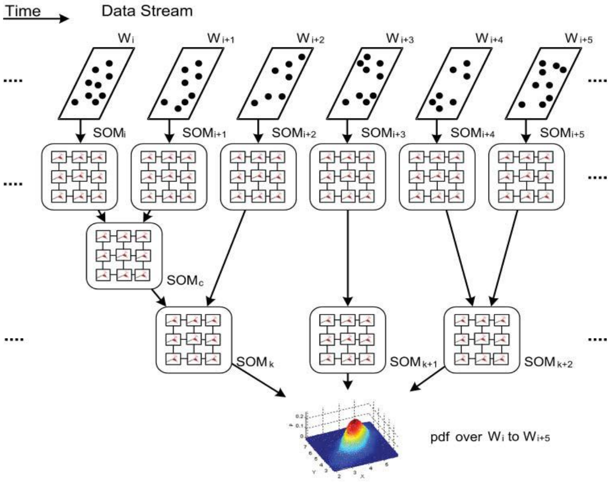

Fig. 5 SOMKE: the proposed system architecture of the SOM-based KDE method (example of 2-D data).

Final estimation which is the last step of SOMKE where SOM sequence to estimate the PDFs over arbitrary time periods on data stream. Advatages of SOMKE algorithm is that it can be used in the context of nonstationary data streams effectively and efficiently. In figure 5 proposed system architecture of SOM based KDE method is displayed.

8. *LeaDen-Stream Algorithm*

LeaDen-Stream algorithm is Leader Density based Clustering algorithm over data streams. LeaDen-Stream algorithm is two phase clustering method where online phase selects the appropriate mini-micro cluster leader or micro cluster leader depended on distribution of data points in the micro cluster. Offline phase makes final cluster.

Algorithm for Lea Den Stream is as follows: Here in this algorithm we give data stream as input and arbitrary shape cluster is output. After that read whole data stream and Adjust Leader Cluster and Puring Leader Cluster this two methods is applied on the data streams [6]. Adjust Leader Cluster method list mini micro leader cluster and micro leader cluster we got as a output. Here all data points are analyzed and fine the nearest micro leader cluster. If the distance is less than l_m than find distance with mini micro leader cluster if it is less than merge the data point with cluster otherwise form a new mini micro leader cluster or micro leader cluster as per condition [6]. Here all the MLC and MMLC (micro mini leader cluster) are checked. If all the MMLC are sparse then delete the MLC (micro leader cluster), if all the MMLC are dense then add the MLC center to the list. If some of MMLC are dense and some sparse then add all the DMMLC (dense mini micro leader cluster) center and remove SMMLC (sparse mini micro leader cluster) [6]. In this way we got list of centers. After that by using this information final cluster is formed. However, we need to use a clustering algorithm to get the final clusters. When a clustering request arrives, DBSCAN algorithm is used on the micro and mini-micro leader cluster centers to get the final results. Each mini- micro and micro leader center is used as a virtual point to be used for clustering. LeaDen Stream algorithm.

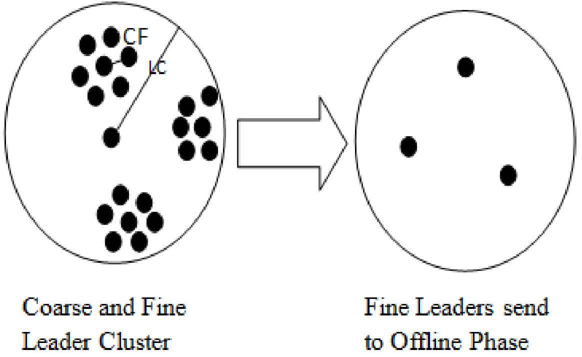
CF
LC
Coarse and Fine Leader Cluster
Fine Leaders send to Offline Phase

Fig. 3 Mini micro and micro leader clusters

The efficiency is measured by the execution time. The quality of LeaDen- Stream is higher than CluStream with lower execution time. The LeaDen-Stream clustering quality is equal to DenStream while it runs faster than DenStream.

V. ANALYSIS OF DENSITY BASED CLUSTERING ALGORITHMS ON DATA STREAMS

Here we have compared density based clustering algorithms on data streams based on its parameters, advantages, disadvantages, technology and future scope. Denstream algorithm, streamoptics algorithm, MR-Stream algorithm, D-Stream algorithm, HDDStream algorithm DENGRIS algorithm, SOMKE algorithm and LeaDen Stream algorithm are compared in this comparison table

Table 1: Comparison of Algorithms

Algorithms	Parameters	Advantages	Disadvantages	Technology	Future Scope
Denstream Algorithm [11]	Cluster Radius, Cluster Weight, Outlier Threshold, Decay Factor	-Gives Arbitrary shape clusters -Handles the data stream effectively by recognizing the potential clusters from the real outliers.	-Does not release any memory space by either deleting a micro-cluster or merging two old micro-clusters. -Pruning phase for removing outliers is a time consuming process in the algorithm	MOA tool	-The discovery of clusters with arbitrary shape at multiple levels of granularity. -Dynamic adaption of the parameters in data streams. - Investigation of our framework for outlier detection and density-based clustering in other stream models, particular in a sliding window model.
Streamoptics Algorithm [12]	Potential Micro-Cluster List, Core Distance, Reachability Distance	-Cluster structure plot over time -Provide the three-dimensional plot	-It is not a supervised method for cluster extraction -It needs manual checking of the generated three-dimensional plot	WEKA and R tool	-automate the generation three dimensional plot and improve the efficiency.
MR-Stream Algorithm [10]	Data Stream, Decay Factor, Dense Cell Threshold, Sparse Cell Threshold	-Memory sampling method in order to define the right time for running the offline component -Improves the performance of the clustering. -This algorithm gives clusters in multiple resolutions. -It improves performance	-MR-Stream keeps the sparse grids and merges them for consideration as a noise cluster. -Cannot work properly in high-dimensional data.	Apache hadoop	-Make algorithm effective for detecting noise by using various approaches.
D-Stream Algorithm [3]	Data Stream, Decay Factor, Dense Grid Threshold, Sparse Grid Threshold	-It proposes a density decaying to adjust the clusters in real time and captures the evolving behavior of data streams	-for determining the time interval gap, the algorithm considers the minimum time but gap depends on many	R-tool and VC++ 6.0	-Algorithm would define the time gap based on only the conversion of dense grids to sparse ones, since the conversion of sparse grid to dense one has already been considered in the weight of the grid

		-It has techniques for handling the outlier. -It provide clusters in arbitrary shapes. -improves the quality	parameters. -It cannot handle the high-dimensional data		
HDDStream Algorithm [8]	Cluster Radius, Cluster Weight, Out-Lier Threshold, Decay Factor	-Arbitrary shape clusters -Clustering high-dimensional data -allows noise handling -does not require the information of number of clusters -monitors the behavior of data.	-searching the affected neighboring clusters is a time consuming process.	java and c++	-improve the time complexity by using various concepts.
DENGRIS Algorithm [4]	Data Stream, Sliding Window Size	-it can capture the distribution of recent records precisely using sliding window model. -it removes the expired grids before any processing on the grid list which leads to save time and memory	No evaluation to show its effectiveness compared with other state-of-the-art algorithms	KEEL software	-Compare with other density based clustering approaches to show its effectiveness.
SOMKE Algorithm [5]	Window Size, Neurons, Topological Neighborhood	-Algorithm has good performance compared to other algorithm. -Use in non stationary data efficiently and effectively.	- Cannot handle imbalance data.	WEKA	-Compare with other SOM based time series approach like patch cluster method, SOMAR and GSOMAR model and observe and analyze the results. So new insight and findings we can get. -Study of magnification effect on SOM-based technique. -Apply this algorithm on imbalanced data.
Leaden Stream algorithm [6]	MMLC And MLCweight, MMLC And MLC Center, MMLC And MLC Radius.	-Increase cluster quality -Decrease time complexity	- Cannot handle high dimensional data. - Quality is same as Denstream Algorithm	MOA framework	-Automate the parameters of this algorithm and examine algorithm in sliding window model.

VI. CONCLUSION

The density-based clustering method has many advantages like special characteristics, which has the ability to detect arbitrary shape clusters and handle noise. Therefore, so many clustering algorithms on data stream used density method. In this paper, we surveyed a number of density based clustering algorithms over data streams. The main advantage of this paper is that it gives a comprehensive overview of the density-based data stream clustering algorithms and the evaluation table gives information of parameters, advantages and disadvantages.

REFERENCES

Journal Papers

[1] Daniel barbara,"Requirements of clustering data streams"in SIGKDD Explorations, Volume 3, Issue 2 - page 23-27

[2] Madjid Khalilian, Norwati Mustapha," Data Stream Clustering: Challenges and Issues",IMCES 2010, march 17-19,2010, hong kong.

[3] Yixin Chen and Li Tu," Density-Based Clustering for Real-Time Stream Data".

[4] Amineh Amini1 and Teh Ying Wah2," DENGRIS-Stream: A Density-Grid based Clustering Algorithm for Evolving Data Streams over Sliding Window", (ICDMCE'2012) December 21-22, 2012 Bangkok (Thailand), 206-210.

[5] Yuan Cao, Haibo He and Hong Man," SOMKE: Kernel Density Estimation Over Data Streams by Sequences of Self-Organizing Maps", IEEE Transactions On Neural Networks And Learning Systems, Vol. 23, No. 8, August 2012, 1254-1267.

[6] Amineh Amini, Teh Ying Wah," Leaden-Stream: A Leader Density-Based Clustering Algorithm over Evolving Data Stream", Journal of Computer and Communications, 2013, 1, 26-31

[7] David Breitkreutz and Kate Casey," Clusterers: a Comparison of Partitioning and Density-Based Algorithms and a Discussion of Optimisations".

[8] Irene Ntoutsi1, Arthur Zimek1, Themis Palpanas2, Peer Kröger1, Hans-Peter Kriegel1," Density-based Projected Clustering over High Dimensional Data Streams".

[9] Amineh Amini, Teh Ying Wah, and Hadi Saboohi," On Density-Based Data Streams Clustering Algorithms: A Survey", Journal Of Computer Science And Technology 29(1): 116–141 Jan. 2014. DOI 10.1007/s11390-013-1416-3.

[10] Wan, Li, Wan, L. "Density-based clustering of data streams at multiple resolutions" Presented at Discover URECA @ NTU poster exhibition and competition, Nanyang Technological University, Singapore.2008 March.

[11] Feng Cao, Martin Estery, Weining Qian, Aoying Zhou, "Density-Based Clustering over an Evolving Data Stream with Noise".

[12] Ankerst M, Breunig M M, Kriegel H P, Sander J. Optics: Ordering points to identify the clustering structure. ACM SIGMOD Record, 1999, 28(2): 49-60.

Books:

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