International Journal of Engineering & Technology, 7 (4.36) (2018) 147-153

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

A Survey on Clustering Density Based Data Stream algorithms

Mayas Aljibawi*, Mohd Zakree Ahmed Nazri , Zalinda Othman

Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia 43600 Bangi, Selangor Darul Ehsan, Malaysia
**Corresponding author E-mail: mayasaljibawi@gmail.com*

Abstract

With the rapid evolution of technology, data size has increased as well. Thus, open the door to a new challenge of finding patterns such as the limitation of memory and time and the one pass to the whole data. Many clustering techniques has been developed to overcome these issues. Streaming data evolve with time, and that makes it almost impossible to define clusters number in that data. Density-based algorithm is one of the significant data clustering class to overcome this issue due to it doesn't require an advance knowledge about the number of clusters. This paper reviewed some of the existing density-based clustering algorithms for the data stream with the measurement used to evaluate the algorithm.

Keywords *data mining, clustering, density-based clustering, grid-based clustering, micro-clustering, stream data clustering.*

1. Introduction

The rapid development in the technology make the data size collected from various sources very large. For example, the genome of a single human been can hold up to 4 gigabytes of data space [1], and the amount of data that we create every day reach up to 2.5 quintillion bytes [2].Another huge amount of data can be continually generated from the streaming via different applications. Stream data mining which is referring to extract the structure of the knowledge from the stream, is attracting many researchers because of growing of data stream generation and its application importance [3]. Traditional approaches used to analysis the data are not suitable anymore to be used with the massive amount of the new data. Therefore, demands for new approaches to extract the important information from that data are needed, with a robust techniques for examining, explaining data the get the relevant knowledge that assists in the decision making.

2. Data mining and data clustering

2.1 Data mining

It is the method of extracting the unidentified relevant pattern such as unusual records (anomaly detection), cluster analysis and dependencies [4, 5]. Many definitions for the data mining mentioned in the literature are discussed below:
[6] Defines Data mining as the approach of finding essential connections, patterns, by moving through the data stored in depository. [4] Says, it is the process of processing voluminous data stored in the database, seeking for patterns and affiliation within that data. [7] Gives another definition for the data mining as the process of picking, discovering, and modeling huge amounts of data to discover previously anonymous patterns of a business advantage.

2.2 Data clustering:

Clustering is most suitable techniques to distribute the data into groups of similar objects which are closely related and different with other groups' objects. The clustering approaches smoothly arrange a set of patterns into the group or clusters on the basis of similarity measures. Cluster techniques are based on an unsupervised approach where data items are unlabeled to group them into valid clusters [4, 5], while in unsupervised approaches, the dataset is given in the form of pre-classified item set. If the dataset is already labeled it help us to create a new label.

Figure 1 data mining steps

- **Clustering:** is the process where the data points been partitioning into smaller groups. Each of the formed groups represent a cluster where the objects are similar to each other, while dissimilar to other cluster's objects. The results from this process referred to as a clustering [3].
- **Requirements for Cluster Analysis**

➢ **Scalability**: a lot of literature algorithms can handle small datasets, while databases nowadays consist of millions of objects, that makes high scalability is a must in the clustering algorithm.

➢ **Handling different types of attributes**: algorithms normally developed to deal with one type of data (numeric, binary, nominal, etc.). However, many applications start to require clustering algorithm for complex types of data.

➢ **Discover clusters with different shapes**: clustering algorithms usually use either the Euclidean or Manhattan for measuring the distance, then determine the shape of the clusters which normally will be a similar size and density spherical shape cluster. However, the shape of the clusters could be various (e.g.

sensors). That means clustering algorithm need to be able to cluster datasets with different shape of clusters.

- **Handle the noisy data**: outlier, noise and erroneous data are common things in the real-world datasets. Therefore, clustering algorithm need to be able to handle the noisy data in order to get a good quality clustering.
- **Handling high-dimensionality data**: datasets are various in dimensions and attributes. Thus make the dimensionality is one of the clustering challenges which must be handled by the clustering algorithms.

3. Clustering Data Stream

Clustering is an essential task for data mining [8-10] which the results are classified into what none as cluster. Data stream clustering came with many new challenges [2] like the unbound amount of data which makes it impossible to put it in memory, the high rate of arriving data which requires high processing speed. Furthermore, the one passes only to the data which means it's impossible to perform multi scan. There are so many algorithms proposed for clustering static datasets in the literature [11, 12] where some of these algorithm have been optimized to work for the data stream.
Many algorithms have been proposed for clustering static datasets [12] while other algorithms have been extended for data streams. Basically, there are five major clustering categories [3]: partitioning, hierarchical, model, grid and density. The clustering categories shown in figure 2.

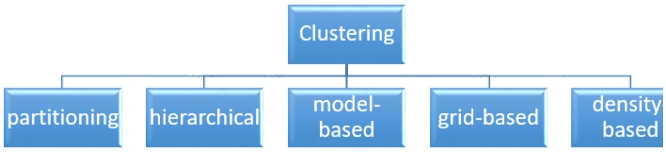

Figure 2: clustering categories

1- **Partitioning-based:** in this method the objects will be organized into a number of partitions which also known as cluster. Partitioning methods use a distance function to form the clusters which will leads to find only spherical shapers clusters. This kind of clusters will be influenced by the noise within the data. STREAM [9] and CluStream [13] which are extension of the k-mean are an example of the partitioning-based methods.
2- **Hierarchical-based:** in this method, the data will be grouped into a tree of clusters. This grouping process is very useful in the visualization and the summarization of the data. This method depends on (merge and split) steps. Once one of these two steps has been done, it cannot be undone. Chameleon [14] and BRICH [15] algorithms were proposed to improve the clustering quality of the hierarchical method. ClusTree [16] algorithm is another example of this method.
3- **Model-based:** this method tries to do enhance the compatibility between the data and the mathematical model. A good example of this method is the EM [17] algorithm, which can be considered as a k-mean extension, using the weight representing to assign the object to a cluster. SWEM [18] is another example of the Model-Based clustering, proposed for clustering the streaming data.
4- **Grid-based:** In this method, the space of the data will be partitioning into amount of cells grouped to form the grids. The processing time of this method is very fast since it is autonomous of the spreading of data points and not depends on the number of data points. STING [19], WaveClustern [20] and CLIQUE [21] (which is grid-density based) are example of grid-based algorithm. Grid method can be combined with the density method for streaming data clustering to form what known as density grid based algorithms like D-Stream [22] and MR-Stream [23].
5- **Density- based:** in this method, the space of the data will be partitioning for a number of dense area based on the density notation. The cluster will keep growing (with data points) if the neighborhood density override some threshold. This method is very useful in detecting noise and to find the arbitrary shape clusters. DBSCAN [24] OPTICS [25] and DENCLUE [26] are example of this method.

4. Analysis of Density Based Clustering Algorithms on Data Streams

Here we have compared density based clustering algorithms on data streams based on its time complexity, quality metric, memory usage, capability of clustering evolve data, capability of clustering high dimensional data, capability of handling the outliers, advantages and disadvantages. DenStream [27], StreamOptiics [28], C-DenStream [29], rDenStream [30], SDStream [31], HDenStream [32], SOStream [33], HDDStream [34], PreDeConStream [35], FlockStream [36], DUC-Stream [37], D-Stream [22], DD-Stream [38], D-Stream II [39], MR-Stream [23], DCUStream [40], PKS-Stream [41], DENGRIS-Strteam [42], SMOKE [43], LeaDen Stream [44], ExCC [45], HDC-Stream [46], MuDi-Stream [47], ADStream [48]and evoStream [49].are compared in this comparison table

Table 1: Algorithms analysis

algorithm	Time Complexity	Quality Metric	Memory usage	Evolving data	High dimensional data	Outlier Handling	Cons	Pros
DenStream	$O(m)$	Purity	m	Yes	No	Yes	The algorithm handles the data stream. Identifying the outlier from the potential clusters. Generate arbitrary shapes.	Does delete or merge the micro-clusters and that leads to consume memory. The algorithm also consume tine for the Pruning phase for removing outliers.
StreamOptics	$O(m * log(m))$	N/A	m	Yes	No	Yes	uses OPTICS algorithm to provide 3 dimensional plot	The generation of the 3 dimensional plot need a manual checking. Clustering extraction is not supervised.

C-DenStream	$O(m+m_c)$	Rand Index	$m+m_c$	Yes	No	Yes	Domain information added to the micro-clusters as constraints. For the application with a prior knowledge, this algorithm consider very useful. The creationg of clusters is blocked if it's included in the semantics application.	Other than the same limitation of DenStream, difining the constraints need an expert.
rDenStream	$O(m) + T_h$	Purity	$m + S_{hb}$	Yes	No	Yes	Useful for mining the information pattern from the first arriving data streams.	Memory consuming. Needs a large amount of outlier to perform well.
SDStream	*N/A*	Purity	n_{sw}	Yes	No	Yes	Processing the new data by using sliding window as well as keep summarization of the old data.	The usage of the exponential histogram for the algorithm is not clarify very well.
HDenStream	$O(m)$	Purity	m	Yes	No	Yes	can handle continuous and categorical data	The algorithm didn't clarify the way of saving the categorical features in an efficient way for data stream environment.
SOStream	$O(n2 \log n)$	Purity	m	Yes	No	Yes	SOStream use a threshold can be adapted for data streaming.	Not suitable for clustering data stream due to the high consumption of time.
HDDStream	$O(m) + O(mp)$	Purity	m	yes	Yes	Yes	Cluster high dimensional data.	Checks only micro-cluster (vanishes over time) weights during the pruning process.
PreDeConStream	$O(m)+O(m_{ip})+O(m_{dp})$	Purity	m	Yes	yes	Yes	Clustering high dimensional data using density method.	Time consuming algorithm
FlockStream	$O(m) + O(n_{agent})$	Purity/NMI	$m+n_{agent}$	Yes	No	Yes	Limited the number of comparisons compared to DenStream, and offline phase will not perform frequently.	Removing the discovered outlier has no clear stratigy in this algorithm.
DUC-Stream	$O(c_b)$	SSQ	n_d	No	No	No	The density of each unit is decreased if that unit does not receive any new data over time and eventually that unit will not be	data chunk's size must be controlled by the user of the algorithm.

							consider for the clustering.	
D-Stream	*O(1) + O(g)*	SSQ	*g*	Yes	No	Yes	Compared to CluStream the quality and time complexity has been improved.	It consider the time interval gap always minimum time. But the gap can be changed depends on some parameters. The algorithm unable of clustering high dimensional data.
DD-Stream	*O(g2)*	N/A	*g*	Yes	No	Yes	The quality ig the clustering has been inproved due to the taking out the boundary points from the grid.	Consume lots of Time, sporadic grids removing process is not clear.
D-Stream II	$O(log\ log\frac{1}{\lambda}\ g)$	SSQ	$log\frac{1}{\lambda}\ g$	Yes	No	Yes	Improve the quality of the clustering by take out the boundary points from the grid. Based on the density of the spares and dense grid the algorithm recognize them.	Consume time for extracting the boundary points. The algorithm has no clear strategy for sporadic grids removing
MR-Stream	*O(g × H) + O(2g × H) +* *O(g × log(N))*	Purity	*g*H*	yes	No	Yes	Improve the performance of the clustering by introducing the memory sampling method which indicate the right time for running the offline phase.	Merging the sparse grids as noise. Can't handle high dimensional data.
DCUStream	*O(g)*	Average Quality of Cluster	*g*	Yes	No	Yes	handles the uncertain data stream environment	Consumes lots of time searching for the core dense grids and their neighbors.
PKS-Stream	*O(log k), O(k)*	Purity	log^{g}_{k}	Yes	Yes	Yes	Can cluster the stream of high dimensional data.	The tree keep adding new data points without any pruning process. The depending on *k* affecting the results of the clustering as well as the *k*-cover, which describes the cluster resolution.
DENGRIS-Stream	*O(g)*	N/A	*g*	Yes	No	Yes	Used sliding model to capture the most recent records distribution. Save time and memory by removing expired grids before any kind of processing.	there is no comparison with other state of art algorithms to show its effectiveness

SOMKE	$O(N\ell)+O(MN)$	MISE	*N/A*	Yes	yes	No	Gives a good performance. Can cluster non-stationary data efficiently and effectively.	It cannot cluster the imbalance data.
LeaDen Stream	*N/A*	purity	*N/A*	Yes	No	No	It has a good quality clustering and less time complexity compared to DenStream.	Cannot cluster high dimensional data.
ExCC	$O(gxk)$	Purity	$g + S_{Pool} + S_{HQ}$	yes	No	Yes	Can cluster numeric and categorical data stream.	Consume lots of time and memory because it uses the pool for keeping the denes grids. Moreover, using the hold queue strategy which is defined for each dimension needs more time and memory.
HDC-Stream	$o(\log \log 1/\lambda N)$	Purity, NMI	O(mi+g)	No	No	Yes	Improve the computation time and quality.	Cannot cluster multi-density data which makes it not suitable for distributed environments.
MuDi-Stream	$O(r_{mc})+O(\log \log N) + O(1) + O(mc) + O(\log N)$	Purity, rand index, adjusted rand iindex, NMI, F measure, Fowlkes–Mallow index , Jaccard index	*cmc*	yes	No	Yes	A hybrid approach using grid based method (to calculating mini-core distance, handle the outliers and reducing the merging time), and micro clustering method to condense the clusters with arbitrary shapes	The empty grids increased by increasing the dimensionality which make the algorithm is not unable to cluster high dimensional data.
ADStream	O(s(S+C) CDM).	Purity	*N/A*	No	Yes	Yes	Using sliding window model to analyze the data streams, and an enhanced similarity propagation clustering is applied to adaptively calculate the initial micro-clusters. Used density grid clustering to generate and update the results of different time granularities.	Need to increase the strength of the algorithm and remove the bad impact of noise in complex data streams on the clustering;
evoStream	*N/A*	SSQ, Adjusted Rand index,	*N/A*	Yes	No	No	Build and refine the final clusters in online-phase	Not suitable for clustering multi-objective data streams. Can’t

		silhouette width, purity, precision, recall, F1 and NMI					by using an evolutionary optimization method. Removes the computational overhead of the offline phase without affecting the speed of processing.	cluster high dimensional datasets.

5. Conclusion

The density-based clustering method has many advantages like special features. Density-based algorithms can handle the noice as well as it has the ability to detect arbitrary shape clusters. Therefore, many clustering algorithms used density method for clustering data stream. In this paper, we studied a number of density based for data streams clustering. This paper gives a comprehensive overview of the density-based algorithms for clustering data stream and analyzied information of time complexity, quality metric, memory usage, capability of clustering evolve data, capability of clustering high dimensional data, capability of handling the outliers, advantages and disadvantages.

References

[1] Wong, K.-C., K.-S. Leung, and M.-H. Wong. Effect of spatial locality on an evolutionary algorithm for multimodal optimization. in European Conference on the Applications of Evolutionary Computation. 2010. Springer.
[2] Amini, A., T.Y. Wah, and H. Saboohi, On density-based data streams clustering algorithms: A survey. Journal of Computer Science and Technology, 2014. 29(1): p. 116-141.
[3] Han, J., J. Pei, and M. Kamber, Data mining: concepts and techniques. 2011: Elsevier.
[4] Hruschka, E.R., R.J. Campello, and A.A. Freitas, A survey of evolutionary algorithms for clustering. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 2009. 39(2): p. 133-155.
[5] Berkhin, P., A survey of clustering data mining techniques, in Grouping multidimensional data. 2006, Springer. p. 25-71.
[6] Filippone, M., et al., A survey of kernel and spectral methods for clustering. Pattern recognition, 2008. 41(1): p. 176-190.
[7] Borland, J., J. Hirschberg, and J. Lye, Data reduction of discrete responses: an application of cluster analysis. Applied Economics Letters, 2001. 8(3): p. 149-153.
[8] Aggarwal, C.C., Data streams: models and algorithms. Vol. 31. 2007: Springer Science & Business Media.
[9] O'callaghan, L., et al. Streaming-data algorithms for high-quality clustering. in Data Engineering, 2002. Proceedings. 18th International Conference on. 2002. IEEE.
[10] Ackermann, M.R., et al., StreamKM++: A clustering algorithm for data streams. Journal of Experimental Algorithmics (JEA), 2012. 17: p. 2.4.
[11] Jain, A.K. and R.C. Dubes, Algorithms for clustering data. 1988.
[12] Mohammed, M.A., Ghani, M.K.A., Arunkumar, N., Obaid, O.I., Mostafa, S.A., Jaber, M.M., Burhanuddin, M.A., Matar, B.M. and Ibrahim, D.A., 2018. Genetic case-based reasoning for improved mobile phone faults diagnosis. Computers & Electrical Engineering, 71, pp.212-222.
[13] Aggarwal, C.C., et al. -A Framework for Clustering Evolving Data Streams. in Proceedings 2003 VLDB Conference. 2003. Elsevier.
[14] Karypis, G., E.-H. Han, and V. Kumar, Chameleon: Hierarchical clustering using dynamic modeling. Computer, 1999. 32(8): p. 68-75.
[15] Zhang, T., R. Ramakrishnan, and M. Livny. BIRCH: an efficient data clustering method for very large databases. in ACM Sigmod Record. 1996. ACM.
[16] Mostafa, S.A., Mustapha, A., Mohammed, M.A., Ahmad, M.S. and Mahmoud, M.A., 2018. A fuzzy logic control in adjustable autonomy of a multi-agent system for an automated elderly movement monitoring application. International journal of medical informatics, 112, pp.173-184.
[17] Dempster, A.P., N.M. Laird, and D.B. Rubin, Maximum likelihood from incomplete data via the EM algorithm. Journal of the royal statistical society. Series B (methodological), 1977: p. 1-38.
[18] Dang, X.H., et al. An EM-based algorithm for clustering data streams in sliding windows. in International Conference on Database Systems for Advanced Applications. 2009. Springer.
[19] Ghani, M.K.A., Mohammed, M.A., Ibrahim, M.S., Mostafa, S.A. And Ibrahim, D.A., 2017. Implementing An Efficient Expert System For Services Center Management By Fuzzy Logic Controller. Journal of Theoretical & Applied Information Technology, 95(13).
[20] Sheikholeslami, G., S. Chatterjee, and A. Zhang, WaveCluster: a wavelet-based clustering approach for spatial data in very large databases. The VLDB Journal—The International Journal on Very Large Data Bases, 2000. 8(3-4): p. 289-304.
[21] Agrawal, R., et al., Automatic subspace clustering of high dimensional data for data mining applications. Vol. 27. 1998: ACM.
[22] Chen, Y. and L. Tu. Density-based clustering for real-time stream data. in Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining. 2007. ACM.
[23] Wan, L., et al., Density-based clustering of data streams at multiple resolutions. ACM Transactions on Knowledge discovery from Data (TKDD), 2009. 3(3): p. 14.
[24] Ester, M., et al. A density-based algorithm for discovering clusters in large spatial databases with noise. in Kdd. 1996.
[25] Mostafa, S.A., Mustapha, A., Hazeem, A.A., Khaleefah, S.H. and Mohammed, M.A., 2018. An Agent-Based Inference Engine for Efficient and Reliable Automated Car Failure Diagnosis Assistance. IEEE Access, 6, pp.8322-8331.
[26] Hinneburg, A. and D.A. Keim. An efficient approach to clustering in large multimedia databases with noise. in KDD. 1998.
[27] Cao, F., et al. Density-based clustering over an evolving data stream with noise. in Proceedings of the 2006 SIAM international conference on data mining. 2006. SIAM.
[28] Tasoulis, D.K., G. Ross, and N.M. Adams. Visualising the cluster structure of data streams. in International Symposium on Intelligent Data Analysis. 2007. Springer.
[29] Mutlag, A.A., Ghani, M.K.A., Arunkumar, N., Mohamed, M.A. and Mohd, O., 2019. Enabling technologies for fog computing in healthcare IoT systems. Future Generation Computer Systems, 90, pp.62-78..
[30] Liu, L.-x., et al. A three-step clustering algorithm over an evolving data stream. in Intelligent Computing and Intelligent Systems, 2009. ICIS 2009. IEEE International Conference on. 2009. IEEE.
[31] Ren, J. and R. Ma. Density-based data streams clustering over sliding windows. in Fuzzy Systems and Knowledge Discovery, 2009. FSKD'09. Sixth International Conference on. 2009. IEEE.
[32] Lin, J. and H. Lin. A density-based clustering over evolving heterogeneous data stream. in Computing, Communication,

Control, and Management, 2009. CCCM 2009. ISECS International Colloquium on. 2009. IEEE.

[33] Isaksson, C., M.H. Dunham, and M. Hahsler. SOStream: Self organizing density-based clustering over data stream. in International Workshop on Machine Learning and Data Mining in Pattern Recognition. 2012. Springer.

[34] Ntoutsi, I., et al. Density-based projected clustering over high dimensional data streams. in Proceedings of the 2012 SIAM International Conference on Data Mining. 2012. SIAM.

[35] Hassani, M., et al. Density-based projected clustering of data streams. in International Conference on Scalable Uncertainty Management. 2012. Springer.

[36] Forestiero, A., C. Pizzuti, and G. Spezzano, A single pass algorithm for clustering evolving data streams based on swarm intelligence. Data Mining and Knowledge Discovery, 2013. 26(1): p. 1-26.

[37] Gao, J., et al. An incremental data stream clustering algorithm based on dense units detection. in Pacific-Asia Conference on Knowledge Discovery and Data Mining. 2005. Springer.

[38] Jia, C., C. Tan, and A. Yong. A grid and density-based clustering algorithm for processing data stream. in Genetic and Evolutionary Computing, 2008. WGEC'08. Second International Conference on. 2008. IEEE.

[39] Tu, L. and Y. Chen, Stream data clustering based on grid density and attraction. ACM Transactions on Knowledge Discovery from Data (TKDD), 2009. 3(3): p. 12.

[40] 20. Mostafa, S.A., Ahmad, M.S., Mustapha, A. and Mohammed, M.A., 2017. Formulating layered adjustable autonomy for unmanned aerial vehicles. International Journal of Intelligent Computing and Cybernetics, 10(4), pp.430-450.

[41] Ren, J., B. Cai, and C. Hu, Clustering over data streams based on grid density and index tree. Journal of Convergence Information Technology, 2011. 6(1).

[42] Amini, A. and T.Y. Wah. DENGRIS-Stream: A density-grid based clustering algorithm for evolving data streams over sliding window. in Proc. International Conference on Data Mining and Computer Engineering. 2012.

[43] Cao, Y., H. He, and H. Man, SOMKE: Kernel density estimation over data streams by sequences of self-organizing maps. IEEE transactions on neural networks and learning systems, 2012. 23(8): p. 1254-1268.

[44] Amini, A. and T.Y. Wah, Leaden-stream: A leader density-based clustering algorithm over evolving data stream. Journal of Computer and Communications, 2013. 1(05): p. 26.

[45] Bhatnagar, V., S. Kaur, and S. Chakravarthy, Clustering data streams using grid-based synopsis. Knowledge and information systems, 2014. 41(1): p. 127-152.

[46] Amini, A., et al., A fast density-based clustering algorithm for real-time internet of things stream. The Scientific World Journal, 2014. 2014.

[47] Amini, A., et al., MuDi-Stream: A multi density clustering algorithm for evolving data stream. Journal of Network and Computer Applications, 2016. 59: p. 370-385.

[48] Ding, S., et al., An adaptive density data stream clustering algorithm. Cognitive Computation, 2016. 8(1): p. 30-38.

[49] Carnein, M. and H. Trautmann, evoStream–Evolutionary Stream Clustering Utilizing Idle Times. Big Data Research, 2018.