A Review on Data Stream Classification

IOP Conf. Series: Journal of Physics: Conf. Series **1018** (2018) 012019 doi:10.1088/1742-6596/1018/1/012019

A Review on Data Stream Classification

Haneen A. A[1], A. Noraziah[1, 2], Mohd Helmy Abd Wahab[3]

[1]Faculty of Computer Systems & Software Engineering, Universiti Malaysia Pahang, 26300 Kuantan, Pahang.
[2]IBM Centre of Excellence, Universiti Malaysia Pahang, 26300 Kuantan, Pahang.
[3]Department of Computer Engineering, Faculty of Electrical and Electronic Engineering
Universiti Tun Hussein Onn Malaysia, 86400 Batu Pahat, Johor, Malaysia
Email: noraziah@ump.edu.my

Abstract. At this present time, the significance of data streams cannot be denied as many researchers have placed their focus on the research areas of databases, statistics, and computer science. In fact, data streams refer to some data points sequences that are found in order with the potential to be non-binding, which is generated from the process of generating information in a manner that is not stationary. As such the typical tasks of searching data have been linked to streams of data that are inclusive of clustering, classification, and repeated mining of pattern. This paper presents several data stream clustering approaches, which are based on density, besides attempting to comprehend the function of the related algorithms; both semi-supervised and active learning, along with reviews of a number of recent studies.

Keywords: Clustering, Data Mining, Data Streams, Computational Intelligence

1. Introduction

The dramatic growth in information technology and the vast volume of generated data have led to several fresh challenging discovery tasks in data processing. In fact, the notion “data stream” refers to the data sequence that is embedded continuously and uncertainly in a certain system for storage or for processing[1]. Moreover, these data streams share similar characteristics from the aspects of massiveness, temporal order, rapid changes, as well as potential length infinite [2-4].

In addition, Aggarwal et al. has offered some reasons that disport data streams from the traditional data mining approach[5] which are: (i) size of data streams is potentially boundless; (ii) elements of stream arriving on-line; (iii) limitations in memory space, prior to processing an element, and system rejection or summary; (4) failure of system in controlling or determining the arrival of data elements.

A number of algorithms have been opted relevant to this study. Nevertheless, these selected algorithms fail when clustering data with varied densities due to usage of common parameters meant for clustering that is based on density, thus generating poor setting for densities that are of single and multiple types. Clustering algorithms with density-based data stream.

2. Data stream clustering

Clustering, the task of information focuses to (ordinarily *k*) groups to such an extent that focuses inside each groups are more like each other than to focuses in different groups, is an extremely essential unsupervised information mining assignment. For static informational collections, strategies like *k*-means, *k*-medoids, hierarchical clustering and density-based techniques have been produced among others. A number of these methods are available in tools like *R*. However, the standard algorithms

require access to all information and all data points and typically iterate over the data multiple times. This prerequisite makes these algorithms unsuitable for large data streams and led to the development of data stream clustering algorithms. Figure 1 shows the high level view of the stream architecture.

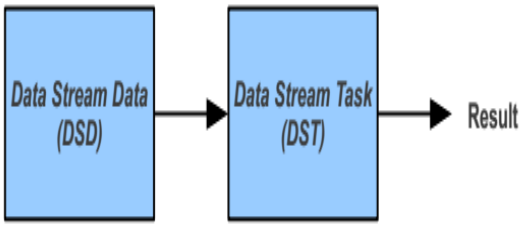

Fig.1. High level view of the stream architecture

Many algorithms for data streams over the last 10 years have been proposed for clustering. Most data stream clustering algorithms deal with the problems of unbounded stream size, and the requirements for real-time processing in a single pass by using the following two-stage online/offline approach:

1. Online: Summarize the information by a set of k' micro-clusters groups composed in a space efficient information structure which also enables fast look-up. Micro-clusters were introduced for CluStream[3] based on the idea of cluster features developed for clustering large data sets with the BIRCH algorithm. Micro-clusters are representatives for sets of similar data points and are created using a single pass over the data (normally progressively when the information of data stream arrives). Micro-clusters are frequently represented by cluster centres and additional statistics such as weight (local density) and dispersion (variance). Each new data point is assigned to its closest (in terms of a similarity function) micro-cluster. Some algorithms use a grid instead and micro-clusters are represented by non-empty grid cells [4]. If a new data point cannot be assigned to an existing micro-cluster, a new micro-cluster is created. The algorithm might also perform some housekeeping (merging or deleting micro-clusters) to keep the number of micro-clusters at a manageable size or to remove information outdated due to a change in the stream's data generating process [5].
2. Offline: When the user or the application requires a clustering, the k' micro-clusters are clustered into k « k' final clusters sometimes referred to as macro-clusters. Since the offline part is typically not respected time critical, most researchers utilize a regular clustering algorithm where micro-cluster centers are viewed as pseudo-points. Average clustering techniques involve k-means or clustering based on the concept of reachability introduced by DBSCAN. This algorithms is often modified to take also the weight of micro-clusters into account.

The most popular approach to adapt to concept drift (changes of the data generating process over time) is to use the exponential fading strategy introduced first for DenStream by [4]. Micro-cluster weights are faded in every time step by a factor of $2^{-\lambda}$, where $\lambda > 0$ is a user-specified fading factor. This way, new data points have more impact on the clustering and the influence of older points gradually disappears. Alternative models use sliding or landmark windows. Details of these methods as well as other data stream clustering algorithms are discussed in[4].

Clustering algorithms that are based on density are meant to identify random-shaped clusters, as these clusters are determined by looking into the areas of density. For example, when a couple of points are relatively near at a dense area, these points could link to generate one cluster. Some instances of this approach are DBSCAN [6], OPTIC[7], and DENCLUE [8]. Moreover, such methods have been employed within these recent times to cluster data streams that are evolving[9-11].

Besides, the DenStream[12] has extended the notion of micro-cluster [13], besides introducing both the potential and the outlier for these micro-clusters in order to differentiate the actual data from the ones

that are not. In fact, this technique embeds the stages of online and offline. As for the online stage at the initiation phase, the DBSCAN algorithm is employed by the DenStream for the few points that are initial, thus forming potentially the first micro-clusters. Nonetheless, any new point of data is attached to the closest micro-cluster with potential or even to a micro-cluster that is an outlier. However, upon failure of attachment, a fresh micro-cluster that is outlier in nature is built and located at the outlier region. On the other hand, the DBSCAN is used in the offline stage in order to identify clusters that are final found on micro-clusters with potential. Furthermore, DenStream promotes a method of pruning to determine the weight of micro-clusters that are outlier found at the outlier region. Unfortunately, this technique fails in clustering data stream that has varied densities because density is dismissed when forming micro-clusters at the online stage result is in Figure 2.

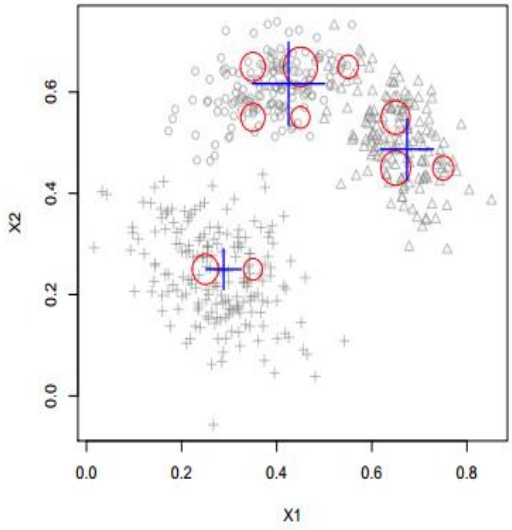

Fig.2. Data stream clustering result of D-stream on a simple simulated data set.

Meanwhile, the offline stage also leads to the formation of final clusters with the use of DBSCAN[14] that fails in determining data with density. In addition, the process of attaching takes up a lot of time. On the other hand, the FlockStream [15] refers to a type of clustering algorithm for density and it adopts the bio-inspired model, which reflects the model of flocking [16], where micro-clusters function as independent agents that generate formation of clusters. In fact, an agent is considered for the points of data via mapping on virtual space. Furthermore, these agents levitate for a certain time within a predetermined range of visibility. However, upon visiting an agent with similar attributes, cluster is formed, thus generating the stages of online and offline for cluster formation occurs without prediction.

Additionally, a clustering framework that is based on density grid is suggested by[17] meant for real time data streams, or also known as D-Stream I, which possesses both stages of online and offline. At the online stage, a fresh point of data is read, mapped onto a grid, and lastly, the grid attribute vector is updated. In oppose, the clusters are modified at offline stage at interval time. This gap of interval time is reflective of the minimal transformation time for varied grids. Besides, this D-Stream I initially keeps the grid density updated, and later, uses a common technique to cluster based on density. Significantly, this particular model adopts outliers are grids of sporadic, which refers to a grid that is sparse with limited data and fails converting to a grid that is dense.

On top of that, D-Stream I lowers the threshold of density in accordance to its function. Hence, a grid is known as sporadic if the density of the sparse grid is lower that the threshold. Besides, the pruning stage occurs at interval time, where modification of clusters and removal of sporadic grid occur. Furthermore, a hash table is used in D-Stream I to retain the list of grids. Next, D-Stream I was

expanded [18], thus called as D-Stream II, which disregards the position of data in a grid for clustering algorithm with multi-density.

On the other hand, [19] developed MR-Stream, which is a clustering algorithm based on density meant for streams of data with numerous resolutions[20] mainly to enhance its function by ensuring that the offline phase is continuously run because this algorithm projects the accurate time for clusters generation. This works by storing space partitioning by providing partition of space data in cells, as well as data structure that resembles a tree, which stores clustering data in varied resolutions. Hence, every node possesses data concerning its parent and children. Furthermore, the MR-Stream works based on online and offline stages.

At the stage of online, the fresh and newly arrived point of data is mapped on the grid. Meanwhile, if there is no sub-node at the tree structure, a new one is developed to store detail regarding the parent until the tree root. Besides, pruning is done at interval. Next, clusters are developed based on that defined by users at the offline stage to mark reachable cells that are dense at a certain distance as a cluster. As such, clusters that are noisy are eliminated after comparing size and densities with those of the thresholds. Moreover, the MR-Stream adopts the method of memory sampling to identify the exact time to run the offline element, thus enhancing its performance.

Meanwhile, HDC-Stream [21] refers to a clustering algorithm based on density for data stream applied for Internet of Things (IoT). Nonetheless, this algorithm projects processing time that is low, hence appropriate devices of IoT with real-time usage. Besides, clustering is generated by its hybrid approach with three steps, under stages of online and online. The HDC-Stream online stage reads fresh data in a continuous manner, thus adding them into grid or mini-cluster. As for outliers, HDC-Stream prunes them often. Meanwhile, the final clusters are generated at the offline stage, based on users. In fact, this method is high in quality, but has intricacy at lower time for data stream clustering. Nevertheless, its drawback is that it cannot cluster data with multiple densities.

Besides, a comprehensive study about clustering algorithms with density was conducted by dismissing the offline clustering stage [22]. In fact, comparison of new point via FlockStream with other aspects within the visibility distance has decreased neighbourhood comparison for points. This distance of visibility refers to the threshold determined by user. Hence, the related algorithm is a failure for data with multiple densities that possess high quality due to its single distance.

Moreover, each agent has to adhere to some regulations over time to get located within the virtual space, for instance, separation, alignment, and cohesion [23]. With that, there are three types of agents in FlockStream: fresh points of data have fundamental representative agents, while potential and outlier micro-clusters are represented by p- and o-representative agents. In precise, a representative agent is generated as agents with similar attributes combine, which relies on their weights. Nonetheless, FlockStream possesses higher time for computation and decreases the amount of comparisons, when compared to DenStream.Although the procedure for clustering is similar for both D-Stream and D-Stream I; correlation could occur between grids that are dense in D-Stream II, whereby these grids are deemed as strongly correlated if they have higher attraction between grids, in comparison to that of predetermined threshold. Furthermore, the list of grids is kept in tree, and not table, to enable rapid processing and to save memory space.

3. Clustering algorithms with multi-density data set

The GMDBSCAN[24] refers to a clustering algorithm for multiple densities that employs the grid method. In fact, this algorithm identifies the local parameters of MinPts through the use of density grid, which are clustered via DBSCAN by the selected MinPts local parameters. Besides, similarity in density determines the integration among this two-pass clustering; therefore inapt for use of data streams. On the other hand, MSDBSCAN [25] refers to a clustering algorithm based on density in the light of the notion of essential points found in DBSCAN, depending on the location of the neighbours. Furthermore, this suggests a new approach for the essential points, which are known as local core distance (lcd), to reflect the MinPts objects distance. Besides, MSDBSCAN determines the value of lcd for the available points of data to identify the essential, which shares similar value with the lcd vector. Moreover, these essential points are used by the algorithm to generate huge clusters.

Nevertheless, MSDBSCAN has high time intricacy, thus unsuitable for streams of data. Generally, this multiple DBSCAN[26] blankets the issue related to multiple density, which is determined by various values of ϵ. Thus, "must link constraint" and "kth" are employed at the nearest distance to obtain values for ϵ at varied densities. Besides, the multiple DBSCAN selects the best value of ϵ for distribution of density by applying the algorithm that detects outlier. Next, DBSCAN is carried out based on the data by determining the values of ϵ.

In another case, [27] developed a clustering algorithm that is hierarchical to generate clusters with multiple density. This algorithm applies an approach that has multiple levels in order to identify clustered structures that are hierarchical from the data. As such, the agglomerative k-means is employed to build a tree for cluster with varied densities. Moreover, this approach of cluster validation can be employed to determine both clusters of composite and atomic. For instance, clusters that are atomic disregards further division. Meanwhile, division is performed using agglomerative k-means for clusters that are composite, where the repeated process generates the cluster tree, which is also reflective of a clustering algorithm with dual-pass.

Moving on, the IS-DBSCAN [28] refers to an approach that enhances the algorithm of DBSCAN for limited parameters, as well as its ability in clustering data with multiple densities. This approach introduces a fresh notion called 'space ranking', where the points of data are ranked within space depending on metrics density. On top of that, the ISDBSCAN applies the concept of Influence Space (IS) initiated in INFLO (INFLuenced Outlierness)[27]. Basically, IS offers more accurate prediction for the distribution of neighbourhood density, besides enhancing clusters division with varied densities. Moreover, the conventional technique of determining ϵ-neighbourhood is substituted by a technique that applies IS meant for clustering based on density.

Meanwhile, SCDM2 [21] refers to the extension of SCDM[22], thus a clustering algorithm that is semi-supervise. The SCDM2 was meant for data with multiple density for it applies constriction in the process of clustering. Moreover, in SCDM, the parameter ϵ in DBSCAN is determined based on "must link constraint". Hence, when a cluster does not have any constriction, it would not emerge in the final cluster. As for SCDM2, those clusters that cannot be linked are also embedded to the ϵ representative. On top of that, the DBSCAN-DLP (multiple density DBSCAN based on Density Levels Partitioning)[28] ensures that the parameters used for every cluster to identify the multiple densities clusters via division of density level. Initially, the data sets are separated into various levels of densities. After that, the value of ϵ is calculated for every level of density.

Finally, the DBSCAN is employed to carry out clustering for every level of density, along with values of ϵ to obtain results of clustering. In fact, the DBSCAN-DLP is reflective of clustering algorithm with dual-pass that possesses computation time that is rather high to be used for streams of data. Other than that, GDCLU[24] refers to a clustering algorithm based on density employing the grid approach. Besides, this algorithm, which is scale independent, suggests that this grid is density-based in relation to the neighbours.

Next, the DSCLU[29] employs the approach of micro-clustering for data stream with multiple density. This particular algorithm identifies the micro-clusters that are dominant by taking weights of neighbours into consideration. Hence, these micro-clusters become dense, in fact, share similar density with their neighbours. As for the offline stage, the dominant micro-clusters are used for clustering. Besides, this technique employs the same radius for micro-clusters formation.

4. Conclusions

This particular study has looked into several clustering algorithms that are based on single and multi-density meant for data stream. Furthermore, these grid-based algorithms have their points of data marked on grids, which undergo systematic process. Besides, the values of density for the points of data indicate if the grids are sparse or dense. In fact, structures with grid help in addressing drawback related to algorithm with legacy partition, which is inapt when used for data streams. Moreover,

density enhances clustering to be more rapid and more reliable. Thus, issues related to clustering of data stream are investigated, thus seeking solutions to these glitches. The following are the main issues found within this area: Constraint in Memory: Memory has emerged as a huge limitation due to the huge data stream size. Since it is not possible to keep all data in memory, performing clustering is rather tricky. Therefore, a portion of the data is placed in a time window for clustering purpose, in which the data are later passed to free memory, and lastly, moved forward.

Constraint in Time: Time is also a hindrance for clustering of data stream because rapid stream and data with continuous flow do not wait for the processing. Therefore, clustering is not only done within a certain time, but the system has to continue accepting data without any problem. Hence, one solution to this is the enhancement of algorithm in both response time and efficiency in handling data with a rapid pace.

Acknowledgments

Appreciation conveyed to Ministry of Higher Education, Malaysia for financial assistance under Fundamental Research Grant Scheme RDU140101. In addition, our appreciation convey to the Universiti Malaysia Pahang for project financing under Research Grant Scheme RDU170398 and PGRS170304.

References

[1]. Aggarwal, C.C., *Data streams: models and algorithms*. Vol. 31. 2007: Springer Science & Business Media.
[2]. Rajaraman, A. and J.D. Ullman, *Mining of massive datasets. 2012.* Cited on, 2011: p. 139.
[3]. Kholghi, M., H. Hassanzadeh, and M. Keyvanpour. *Classification and evaluation of data mining techniques for data stream requirements*. in *Computer Communication Control and Automation (3CA), 2010 International Symposium on*. 2010. IEEE.
[4]. Gama, J., *Knowledge discovery from data streams*. 2010: CRC Press.
[5]. Aggarwal, C.C., et al. *A framework for clustering evolving data streams*. in *Proceedings of the 29th international conference on Very large data bases-Volume 29*. 2003. VLDB Endowment.
[6]. Aggarwal, C.C. and C.K. Reddy, *Data clustering: algorithms and applications*. 2013: CRC press.
[7]. Amini, A., H. Saboohi, and T.Y. Wah. *A multi density-based clustering algorithm for data stream with noise*. in *Data Mining Workshops (ICDMW), 2013 IEEE 13th International Conference on*. 2013. IEEE.
[8]. Amini, A., et al., *A fast density-based clustering algorithm for real-time internet of things stream.* The Scientific World Journal, 2014. 2014.
[9]. Amini, A. and T.Y. Wah. *Density micro-clustering algorithms on data streams: A review*. in *Proceeding of the International Multiconference of Engineers and Computer scientists (IMECS)*. 2011.
[10]. 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.
[11]. Amini, A., et al. *A study of density-grid based clustering algorithms on data streams*. in *Fuzzy Systems and Knowledge Discovery (FSKD), 2011 Eighth International Conference on*. 2011. IEEE.
[12]. Ankerst, M., et al. *OPTICS: ordering points to identify the clustering structure*. in *ACM Sigmod record*. 1999. ACM.
[13]. 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.
[14]. Cassisi, C., et al., *Enhancing density-based clustering: Parameter reduction and outlier detection.* Information Systems, 2013. 38(3): p. 317-330.
[15]. Chen, L., L.-J. Zou, and L. Tu, *A clustering algorithm for multiple data streams based on*

spectral component similarity. Information Sciences, 2012. 183(1): p. 35-47.

[16]. Chen, X., et al., *An improved semi-supervised clustering algorithm for multi-density datasets with fewer constraints.* Procedia Engineering, 2012. 29: p. 4325-4329.

[17]. 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.

[18]. Esfandani, G. and H. Abolhassani, *MSDBSCAN: multi-density scale-independent clustering algorithm based on dbscan.* Advanced Data Mining and Applications, 2010: p. 202-213.

[19]. Nakata, Y., T. Hochin, and H. Nomiya, *2012 13th ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel & Distributed Computing (SNPD).*

[20]. Boden, B., M. Ester, and T. Seidl. *Density-based subspace clustering in heterogeneous networks*. in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. 2014. Springer.

[21]. 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: p. 1-26.

[22]. Huang, T.-q., et al. *Reckon the parameter of dbscan for multi-density data sets with constraints*. in *Artificial Intelligence and Computational Intelligence, 2009. AICI'09. International Conference on*. 2009. IEEE.

[23]. Li, X., et al., *On cluster tree for nested and multi-density data clustering.* Pattern Recognition, 2010. 43(9): p. 3130-3143.

[24]. Namadchian, A. and G. Esfandani. *DSCLU: a new Data Stream CLUstring algorithm for multi density environments*. in *Software Engineering, Artificial Intelligence, Networking and Parallel & Distributed Computing (SNPD), 2012 13th ACIS International Conference on*. 2012. IEEE.

[25]. 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.

[26]. Xiong, Z., et al., *Multi-density dbscan algorithm based on density levels partitioning.* JOURNAL OF INFORMATION &COMPUTATIONAL SCIENCE, 2012. 9(10): p. 2739-2749.

[27]. Hu, W., et al. *Research on Parallel Data Stream Clustering Algorithm Based on Grid and Density*. in *Computer Science and Mechanical Automation (CSMA), 2015 International Conference on*. 2015. IEEE.

[28]. Sun, Z., et al. *Knowledge-based evolving clustering algorithm for data stream*. in *Service Systems and Service Management (ICSSSM), 2014 11th International Conference on*. 2014. IEEE.

[29]. Hahsler, M. and M. Bolaños, *Clustering Data Streams Based on Shared Density Between Micro-Clusters.* IEEE Transactions on Knowledge and Data Engineering, 2016. 28(6): p. 1449-1461.