International Journal of Contemporary Research in Computer Science and Technology (IJCRCST) ISSN: 2395-5325
Volume 3, Special Issue 1 (February '2017)
Special Issue on 1st National Conference on Advanced Computing and Communication Technologies,
Conference held at Siri PSG Arts and Science College for Women, Sankari,Tamilnadu,India.

PLOT CLUSTER STRUCTURE OVER TIME USING STREAMOPTICS ALGORITHM

I.Aruna,
Research Scholar,
Department of Computer Science,
Theivanai Ammal College for Women,
Villupuram,Tamilnadu,India.

K.Manohari,
Assistant Professor,
Department of Computer Science,
Theivanai Ammal College for Women,
Villupuram,Tamilnadu,India.

Abstract - Data streams are generated by many real time systems. Data stream is fast changing and massive. Traditional methods are not efficient so that many methodologies developed to stream data processing. Clustering data streams has become an important technique for data and knowledge engineering. Data stream clustering aims at studying large volumes of data to build a good clustering of the stream, using a small amount of memory and time. In this paper, we propose StreamOptics algorithm. 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. StreamOptics is extends OPTICS algorithm for data streams using concepts of micro clusters. 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. This algorithm provides three dimensional plots which show evolution of cluster structure with respect to time.

Keyword : Clustering data streams, StreamOptics, micro-cluster.

I. INTRODUCTION

Nowadays, data clustering techniques are widely used in many different fields of application, such as image processing [1], pattern recognition [2], data mining [3], biological data analysis [4] and so on. Due to the high number of existing applications, many different approaches to data clustering have been proposed in literature. Usually, different techniques are more or less suitable for different types of application. For instance, for applications that require real-time data acquisition, the input to the clustering algorithm is in the form of online data streams. Thus, n-dimensional data observations are obtained, one by one, at a specific sampling rate.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. 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 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 [14]. 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 this paper, we develop and evaluate a new method to address this problem for micro-cluster-based algorithms. We introduce the concept of a Density based clustering method. It has 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]. StreamOptics is introduced for data streams using concepts of micro clusters. StreamOptics is an extension of OPTICS algorithm. OPTICS is stand for Ordering Points to Identify the Clustering Structure. 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.

II. RELATED WORK

This section briefly presents previous works on data stream clustering, and highlights the most relevant algorithms proposed in the literature. Table I summarizes the main features of different algorithms for data stream clustering: the basic clustering algorithm, whether the algorithm provides a hierarchical and/or topological structure, whether the links (if they exist) between clusters are weighted, how many phases an algorithm adopts (online and offline), the network adaptation (remove, merge, and split clusters), and whether a fading function is used.

Algorithms	based on	hierarchy	topology	WL	phases	remove	merge	split	fade
AING	NG	✗	√	✗	online	✗	√	✗	✗
CluStream	k-means	✗	✗	✗	2 phases	√	offline	✗	✗
DenStream	DbScan	✗	✗	✗	2 phases	√	offline	✗	√
SOStream	DbScan, SOM	✗	✗	✗	online	√	√	✗	√
E-Stream	k-means	✗	✗	✗	2 phases	√	√	√	√
SVStream	SVC, SVDD	✗	✗	✗	online	√	√	√	√
StreamKM++	k-means	√	✗	✗	2 phases	√	√	√	√
StrAP	AP	√	✗	✗	2 phases	√	✗	✗	√
ODAC	top-down tree	√	✗	✗	online	✗	✗	√	✗
G-Stream	NG	✗	√	√	online	√	✗	✗	√
GH-Stream	NG	√	√	√	online	√	✗	✗	√

Table I: Comparison between algorithms (WL: weighted links, 2 phases: online + offline).

Relating to the clustering problems, we can list several algorithms such as: CluStream [6] and DenStream [5] use a temporal extension of the Clustering Feature vector [7] (called micro-clusters) to maintain statistical summaries about data locality and timestamps during the online phase. By creating two kinds of micro-clusters (potential and outlier micro-clusters), DenStream overcomes one of the drawbacks of CluStream, its sensitivity to noise. In the offline phase, the micro-clusters found during the online phase are considered as pseudo-points and will be passed to a variant of k-means in the CluStream algorithm (resp. to a variant of DbScan in the DenStream algorithm) in order to determine the final clusters. StreamKM++ [12] maintains a small outline of the input data using the merge-and-reduce technique. The merge step is performed by means of a data structure, named the bucket set. The reduce step is performed by a significantly different summary data structure, the coreset tree. ClusTree [8] is an anytime algorithm that organizes microclusters in a tree structure for faster access and automatically adapts micro-cluster sizes based on the variance of the assigned data points. Any clustering algorithm, e. g. k-means or DbScan, can be used in its offline phase. SOStream [15] is a density-based clustering algorithm inspired by both the principle of the DbScan algorithm and that of self-organizing maps (SOM) [9]. E-Stream [10] classifies the evolution of data into five categories: appearance, disappearance, self evolution, merge, and split. It uses another data structure for saving sUlmnary statistics, named a-bin histogram. StrAP [16], an extension of the Affinity Propagation algorithm for data streams, uses a reservoir for saving potential outliers.

III. SYSTEM ANALYSIS

System analysis is the act, process, or profession of studying an activity typically by mathematical means in order to define its goals or purposes and to discover operations and procedures for accomplishing them most efficiently.

Cluster Data Stream: Clustering data streams has become an important technique for data and knowledge engineering. Current reclustering approaches completely ignore the data density in the area between the micro-clusters (grid cells) and thus might join micro-clusters (cells) which are close together but at the same time separated by a small area of low density. To address this problem, Tu and Chen introduced an extension to the grid-based D-Stream algorithm based on the concept of attraction between adjacent grids cells and showed its effectiveness.

Limitations: The number of clusters varies over time for some of the datasets. This needs to be considered when comparing to CluStream, which uses a fixed number of clusters. Resulting in spikes of very low quality.D-Stream is slower in adapting to the structure and produces results inferior to DBSTREAM

IV. SYSTEM MODEL

In this paper, we propose Density based clustering method have capability to detect cluster in arbitrary shapes and it can also handle noise. StreamOptics is introduced for Density based clustering in data mining. OPTICS is stands for 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.

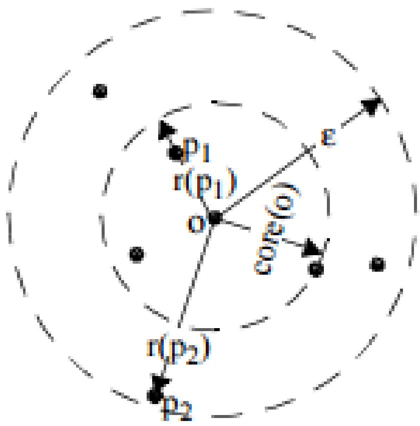

Figure 1: Core-distance(o), reachability-distances r(p1,o), r(p2,o) for MinPts=4

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. 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. This algorithm provides three dimensional plots which show 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.

ADVANTAGES

1. Cluster structure plot over time.
2. Provide the three-dimensional plot
3. It improves performance
4. Memory requirement is less when compare traditional technique

V. ALGORITHM OR TECHNIQUE

Our algorithm OPTICS creates an ordering of a database, additionally storing the coredistanceand a suitable reachability-distance for each object. We will see that this information is sufficient to extract all density-based clusterings with respect to any distance ε' which is smaller than the generating distance ε from this order. The main loop of the algorithm OPTICS is given below. At the beginning, we open a file OrderedFile for writing and close this file after ending the loop. Each object from a database SetOfObjects is simply handed over to a procedure ExpandClusterOrder if the object is not yet processed.

```
OPTICS (SetOfObjects, ε, MinPts, OrderedFile)
OrderedFile.open();
FOR i FROM 1 TO SetOfObjects.size DO
Object := SetOfObjects.get(i);
IF NOT Object.Processed THEN
ExpandClusterOrder(SetOfObjects, Object, ε, MinPts,
OrderedFile)
OrderedFile.close();
END; // OPTICS
```

VI. PERFORMANCE EVALUTION

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

Algorithms	***Parameters***	***Advantages/Disadvantages***	***Technology***	***Future Scope***
Streamoptics Algorithm	Potential Micro-Cluster List, Core Distance, Reachability Distance	***Adv:*** 1. Cluster structure plot over time 2. Provide the three-dimensional plot ***Dis:*** 1. It is not a supervised method for cluster extraction	WEKA and R tool	-automate the generation three dimensional plot and improve the efficiency.
MR-Stream Algorithm	Data Stream, Decay Factor, Dense Cell Threshold, Sparse Cell Threshold	***Adv:*** 1. This algorithm gives clusters in multiple resolutions. ***Dis:*** 1. MR-Stream keeps the sparse grids and merges them for consideration as a noise cluster. 2. Cannot work properly in high-dimensional data.	Apache hadoop	-Make algorithm effective for detecting noise by using various approaches.

D-Stream Algorithm	Data Stream, Decay Factor, Dense Grid Threshold, Sparse Grid Threshold	***Adv:*** 1. It has techniques for handling the outlier. 2. It provide clusters in arbitrary shapes. ***Dis:*** 1. for determining the time interval gap, the algorithm considers the minimum time but gap depends on many parameters. 2. It cannot handle the high-dimensional data	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

VII. 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 Streamoptics Algorithm 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.

VIII.REFERENCES

[1] V. Dehariya, S. Shrivastava, and R. Jain, “Clustering of image data set using k-means and fuzzy k-means algorithms,” in Computational Intelligence and Communication Networks (CICN), 2010 International Conference on, Nov 2010, pp. 386–391.

[2] X. Peng, C. Zhou, D. Hepburn, M. Judd, and W. Siew, “Application of k-means method to pattern recognition in on-line cable partial discharge monitoring,” Dielectrics and Electrical Insulation, IEEE Transactions on, vol. 20, no. 3, pp. 754–761, June 2013.

[3] T. C. Havens, J. C. Bezdek, C. Leckie, L. O. Hall, and M. Palaniswami, “Fuzzy c-means algorithms for very large data.” IEEE T. Fuzzy Systems, vol. 20, no. 6, pp. 1130–1146, 2012.

[4] L. M. O. Mesa, L. F. N. Vasquez, and L. L. Kleine, “Identification and analysis of gene clusters in biological data,” in 2012 IEEE International Conference on Bioinformatics and Biomedicine Workshops, BIBMW 2012, Philadelphia, USA, October 4-7, 2012, 2012, pp. 551–557.

[5] F. Cao, M. Ester, W. Qian, and A. Zhou, "Density-based clustering over an evolving data stream with noise," in SDM, 2006, pp. 328-339.

[6] c. C. Aggarwal, T. J. Watson, R. Ctr, J. Han, J. Wang, and P. S. Yu, "A framework for clustering evolving data streams," in In VLDB, 2003, pp. 81-92.

[7] T. Zhang, R. Ramakrishnan, and M. Livny, "Birch: An efficient data clustering method for very large databases," in SIGMOD Conference, 1996, pp. 103-114.

[8] P. Kranen, I. Assent, C. Baldauf, and T. Seidl, "The ClusTree: indexing micro-clusters for anytime stream mining," Knowledge and iriformation systems, vol. 29, no. 2, pp. 249-272, 2011.

[9] T. Kohonen, M. R. Schroeder, and T. S. Huang, Eds., Self-Organizing Maps, 3rd ed. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 200l.

[IO] K. Udommanetanakit, T. Rakthanmanon, and K. Waiyamai, "E-stream: Evolution-based technique for stream clustering," in ADMA, 2007, pp.605-615.

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

[12] M. R. Ackermann, M. Martens, C. Raupach, K. Swierkot, C. Lammersen, and C. Sohler, "StreamKM++: A clustering algorithm for data streams," ACM Journal of Experimental Algorithmics, vol. 17, no. 1, 2012.

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