IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

*International Journal of Advanced Research in Computer and Communication Engineering*
*Vol. 4, Issue 11, November 2015*

# A Review On Clustering Of Streaming Data

**Madhuri Vilas Gohad[1], Prashant Yawalkar[2]**

PG Student, Computer Engineering, MET BKC Adgaon, Nashik, India [1]

Professor, Computer Engineering, MET BKC Adgaon, Nashik, India [2]

**Abstract**: Data stream clustering is an active research area that has recently used to discover knowledge from continuously generated large amounts of data. There are various data stream clustering algorithms have been developed and proposed to perform clustering on data stream. Clustering is the task of arrangement a set of objects so that objects in the identical group are more related to each other than to those in other groups (clusters). The data stream clustering imposes several challenges that need to be addressed; some of them are dealing with dynamic data, capable of performing processing on fast incoming objects, also capable to perform incremental processing of data objects, and ability to address time, memory and cost limitations.

**Keywords**: Affinity Propagation, Autonomic Computing, Clustering, Data stream, Grid Monitoring.

## I. INTRODUCTION

Clustering or discovering meaningful partitions of data based on a measure of similarity is a critical step in scientific data analysis and a fundamental problem in computer science. Clustering is a widely studied research problem in the data mining. It is the assignment of combination a set of objects such that objects in one group are more related to each other whereas unrelated objects are in other groups. Cluster is an ordered list of data having similarity and dissimilarity. Cluster analysis can be done by finding similarities between data and grouping similar data objects into clusters and dissimilar data objects into reservoir or other clusters. However, it is very difficult to adapt clustering algorithms having arbitrary shape to data streams because of data set has one pass constraints. A common approach within the machine learning community involves unsupervised learning of parameters that describe clusters and partitioning the data by associating every point or region with one or more clusters. In many situations, data is better and more easily characterized by a measure of pair wise similarities rather than defaulting to negative squared Euclidean distance, and in this case, clusters can instead be represented by an exemplar data point rather than domain-specific parameters.

Data streaming, one major task of Data Mining, aims to providing a compact description of the data flows generated by various sources for e.g., telecommunications, Internet traffic or sensor networks [3], [5], [8]. The challenge is to achieve a good clustering behavior, specifically enforcing a low distortion with low computational complexity or cost, while adapting to the changes in the underlying distribution. An additional requirement is that a cluster should be represented by an original data item, as opposed to an average item, and need to remove noise or outliers. The application domain hardly considers the artifacts and enables them.

The autonomic computing is emblematic of a vast and tangled hierarchy of natural self-governing systems. Autonomic Computing, rooted in the fact where the complexity of computational systems calls increases for new approaches to system management, enforcing self-configuring, self-modeling, self-healing and self-optimizing facilities. Autonomic systems are the collection of autonomic elements; individual system contains the resources and service to other elements or the humans. The autonomic systems infrastructure verifies the identities of other entities with which they were communicating, verify that a message has not been altered in transit, and also ensures that unauthorized parties do not got the message in transit or do not able to read it [8], [12].

The EGEE/EGI Grid is the world's largest e-science grid infrastructure. The EGEE/EGI grid system analyses the flow of jobs, submitted to the system and processed by the grids. Grid monitoring includes the two functionalities they are acquisition and usage of the relevant or related information. The goal is to provide dashboard of job stream to the system administrator. The processed log file describes the flow of jobs submitted and processed by grid or gLite that is major EGEE/EGI middleware [8], [9], [12].

The task of exemplar-based clustering is to identify a subset of the data points as exemplars and assign every other data point to one of those exemplars. Data streaming algorithms is the discipline concerned with handling large scale data sets in an online fashion many data stream algorithms have been adapted for clustering large data. This focuses on learning or understanding a data stream generative model, having some specific features:

- The generative model is expressed through actual data items (i.e. set of exemplars).
- It is available at any time, for monitoring application.
- Changes are detected through statistical hypothesis testing in the underlying distribution [12].

The affinity propagation algorithm is simple to implement and customize; it is also computationally efficient, scaling linearly in the number of similarities or quadratic ally in the number of data points if all possible pair wise similarities are used. It investigates cluster detection from data streams, i.e. stream clustering. Because of the huge size and evolving property of data streams, a good stream clustering algorithm should meet the following requirements:


DOI 10.17148/IJARCCE.2015.41193

- Single scan of data: This is a natural constraint of streaming data because of its very large or infinite size. There is no time or may even be impossible to reread the stream for the computation.
- Ability to filter out noise in continuously evolving streams: Random noise can occur anywhere in the data stream and filtering the noise can help getting a good clustering result.
- The system should be able to process very large or infinite streams in main memory of limited size.

Data streaming is a data that gathers through telephone records, webcams, and online transactions. This kind of data is continuous, so to maintain that data the need is to select best representatives from clusters of streaming data.

## II-RELATED WORK

Data streaming is the major task of data mining, faces additional challenges compared to traditional data. Recently, the data stream clustering has been attracting a lot of research attention.

Clustering is the unsupervised learning task of organizing or partitioning data into meaningful groupings. Data streaming faces two difficulties compared to traditional data streaming algorithm. The difficulties are related to the non-stationarity of the data distribution. One is, functionally streaming algorithm must have to maintain a trade-off between noise filtering (outlier must be discarded) and model updating (the model must be appropriately updated when the changes are detected in the data distribution). Second is streaming algorithm must adjust its own parameters [8], [12].

## DIFFERENT TECHNIQUES:

### A. BIRCH

The BIRCH method is suitable for very large databases and used data streaming. BIRCH makes use of all available memory for generating finest possible sub-clusters / clusters. In this method the concept of Clustering Feature (CF) tree is used to generate the clusters, where CF is a height balanced tree. The BIRCH clustering algorithm works as: in first step it scans whole data and builds the initial tree. And in second step it refines the clustering information of the dataset by removing sparse nodes as outliers and concrete original clusters. The I/O complexity is little more than one scan of data. BIRCH is used to overcome the difficulties of agglomerative clustering: scalability and inability to undo what was performed in previous steps. The disadvantages of this method is it has limited capacity for leaf and if clusters are not spherical in shape than algorithm does not perform well [1], [13].

### B. STREAM

In this method K-Medians is leveraged to cluster objects base on sum of squared distances (SSQ) criterion for error measuring. In the first scan of data, the objects are grouped and medians of each group are gathered and weights are assigned to it based on the number of data objects in the cluster. In next step the medians are clustered until top tree. In this the LSERACH method is described based on local search, in which it start with an initial solution and then refine it by making local improvements. The LSEARCH is consistently slower in terms of running time, although its running time has very low variance. There are two main disadvantages for STREAM method they are: time granularity and data evolving [2], [7], [13].

### C. CLUSTREAM

The CluStream method is used for clustering large evolving data streams. This technique tries to cluster the whole stream at single time rather than viewing the stream as a changing process over time. The idea of this technique is to divide the clustering process into online components which stores detailed summary statistics periodically and offline components uses this summary statistics. The online micro clustering process does not depend on any user input. The only aim is to maintain statistics at sufficiently higher level of granularity so that it can be used by offline component.

Firstly the need is to create initial q micro clusters. This initialization is done offline at very beginning of data stream computation process. For creating initial q micro clusters k-means clustering algorithm is used. After the initialization of q micro clusters the online process of updating micro clusters is initiated. When new data object arrives the micro clusters are updated to reflect the changes. The new set of higher level clusters is created by applying macro-cluster creation algorithm separately to each of this set of micro- clusters. The macro clustering process does not use the data stream instead compactly stored summary statistics of the micro clusters are used. These methods are not suitable for discovering clusters of very different size or of non-convex shapes. The CluStream algorithm needs to periodically store the current snapshot of micro-clusters and the snapshots are maintained in disk [3], [5], [7], [13].

### D. CLUSTERING ON DEMAND (COD)

Clustering On Demand Framework is used to dynamically cluster multiple data streams. The COD framework mainly has two advantageous features one is one online statistics collection has one data scan and another is to compact multi resolution approximation which are designed respectively for time and space constraint in data stream environment.

The COD framework mainly consists of two phases, one is the online maintenance phase and the second is offline clustering phase. The online maintenance phase gives an effective and efficient algorithm to maintain the data streams summary hierarchies with multiple resolutions in time linear in both the number of data points in each stream and the number of streams. And for offline phase an adaptive clustering algorithm is devised to retrieve the approximations of the desired sub streams from the summary hierarchies, according to the user specified clustering queries. In the online maintenance phase summaries are maintained according to the processing of data streams. The offline clustering phase deals with user specified queries for clustering.

The objective of online maintenance phase is to provide one scan algorithm for statistics collection of the incoming

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

*International Journal of Advanced Research in Computer and Communication Engineering*
*Vol. 4, Issue 11, November 2015*

multiple data streams. A summary hierarchy is incrementally maintained to provide multi resolution approximations for a stream. In offline clustering phase, the users want to examine clusters with window size w, and at most p windows will be observed. The window size maintained in the summary hierarchy could be different from desired window size. In this situation, there is need to select the fitting models to get approximate desired windows from appropriate levels of the hierarchy. The partitioning of data stream does not support the clustering with flexible time ranges [4].

### E. DEN-STREAM

The Den-stream algorithm was implemented in an evolving data stream for discovering clusters. The dense micro clusters are also called as core micro cluster and used to summarize the clusters of arbitrary shape. Den-stream clustering algorithm is divided into two phases one is micro cluster maintenance and other is generating final clusters. Micro cluster maintenance is done in an online fashion and clusters are generated in an offline fashion.

The micro cluster maintenance phase maintains the group in online fashion of micro clusters which is called as p-micro cluster and q-micro cluster. The q-micro clusters are used to store the outliers and separate memory is used for it. The merging and pruning (Den-stream) algorithms are used. In merging algorithm, if any new object arrives say x, then it needs to be merge in to its nearest p-micro clusters. If radius of p-micro cluster is less than or equal to some threshold $\epsilon$ then merge x into p-micro clusters. Else it is merge in to q-micro clusters $c_o$. Then check the new weight (w) of q-micro cluster, if w is greater than $\beta\mu$(value is between 0 to 1) it means that q-micro clusters grows into potential c-micro clusters. Hence remove $c_o$ from outlier memory and create new p- micro cluster by $c_o$. Else create new q-micro cluster $c_o$ by p and place $c_o$ into outlier memory.

The initialization of P points is done with the help of DBSCAN algorithm. By scanning the P points they have initialized the group of p-micro clusters. In second phase of generating final clusters the online micro clusters captures the density area of data streams. The variant of DBSCAN algorithm is used to get the final clustering result which is applied on set of p-micro clusters. The problem of this method is that it cannot discover clusters with arbitrary shape at multiple levels of granularity [5], [15].

### F. D-STREAM

The D-Stream clustering algorithm is used for data stream clustering using density based approach. This algorithm uses the online component for mapping each input data into the grid and offline component calculates the grid density and clusters the grid based on density. In D-Stream algorithm, it assume discrete time step model (time stamp has integer value 0,1,2…n). The online component continuously reads the new data record at each time step and place that data record into the multidimensional data place with respective dis-cretized density grid in the multi-dimensional space and updates the characteristics vector of the density grid. The offline components dynamically adjust the clusters at every gap time step, at first gap (gap is the integer parameter) algorithm generates the initial clusters. After that it periodically regulates the clusters and removes sporadic grids. It is not able to cluster categorical and text data. One more disadvantage is that for finding time interval gap it gives minimum time but time interval gap depends on many parameters [6], [13], [15].

### G. K-AP

The K-AP technique is used to generate given number optimal set of exemplars using affinity propagation. The K-AP method offers the good guarantee of AP optimality and generates K number of clusters specified by user. They have define the responsibilities as r(i, j) = log$\rho$ij , the availabilities as a(i, j) = log$\alpha$ij , the new messages confidences ($\eta^{out}$ and $\eta^{in}$) as $\eta^{out}$ (i) = $logh_{ii}{}^{out}$ and $\eta^{in}$ (i) = $logh_{ii}{}^{in}$ . The update of responsibilities r(i,j) and r(i, i) differ as K-AP uses _ $\eta^{out}$ (i) instead of s(i, i). The disadvantage is that the computational complexity of K-AP is same as that of traditional AP [10].

### H. FAST AND ACCURATE K-MEANS

To improve the streaming approximation where k is not known as input for Euclidean k-means and data point are sequentially read to form the clusters. This approach can provide better approximate cluster guarantee and it is efficient with complexity O(nk). To compute the facility assignments of each point the approximate nearest-neighbor algorithms is applied on it. For this purpose fast streaming k-means algorithm is used. The approximate nearest neighbor process is very time consuming. These approach is based on solving k-means problem (thus at each point in the algorithm, the memory contains a current set of facilities). The algorithm is little more parametrizable [11].

### I.STRAP ALGORITHM

The STRAP algorithm is extended online version of Affinity Propagation. The STRAP algorithm continues by incrementally updating the current model, if the current data object fits the model; otherwise putting it in a reservoir. A Change Point Detection (CPD) test, Page Hinkley test detects the changes of data distribution by monitoring the data items sent to reservoir. When CPD test triggered, the new model is rebuilt based on the current model and data items in the reservoir. This algorithm investigates the real time adapted test, instead of using empirically setting or model based optimization of CPD. The CPD test plays an important role for catching the evolving distribution. The adaptive threshold can be used to achieve better quality, cause less outlier and less computing time. STRAP extracts exemplars from data streams for building summary model. CPD test enables STRAP to catch drifting exemplars that significantly deviate away [8], [12].

The online updating clustering model is maintained by STRAP through i) when a data item arrives, checking its fitness against the model; ii) if fitting, simply updating the corresponding cluster in the model, otherwise putting it into the reservoir. Restart criteria are used to monitor the


DOI 10.17148/IJARCCE.2015.41193

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
*International Journal of Advanced Research in Computer and Communication Engineering*
*Vol. 4, Issue 11, November 2015*

changes of stream distribution. If changes are detected, the stream model is re-build by applying WAP on the current model and the data in the reservoir [8], [12]. The STRAP algorithm has quadratic computational complexity w.r.t the number of outliers and exemplars. For adjusting the parameters of the change detection test an adaptive procedure is used [9]. The limitation is that the number of exemplars is not easily controlled from the penalty parameter and its computational cost [8], [9], [12].

## ALGORITHM

**Steps**:
Input: Data Stream $x_1, x_2,\ldots,x_t$, threshold value ε
Begin
$AP(x_1,\ldots,x_t)\rightarrow$STRAP model;
Reservoir={ };
For t >=T do
Compute $e_i$=nearest exemplar to $x_t$
if $d(x_t,e_i) <=\varepsilon$ then
Update STRAP model
else
Reservoir← $x_t$
end if
If restart criterion then
Rebuild model
Reservoir = { }
End if
End for

The algorithm works as follows:
1. The data set flow into the system is used by AP to identify the first exemplar and initialize the stream model.
2. Each data item is compared with the exemplar, if it is near to the nearest exemplar, the model is updated otherwise it is put in the reservoir
3. The data distribution is checked for change point detection test using statistical test, called the Page Hinkley test.
4. When the change detection triggered a test or if the number of outlier exceeds the reservoir size, the model is rebuilt based on the items in reservoir and current model [8], [9], [12], [14].

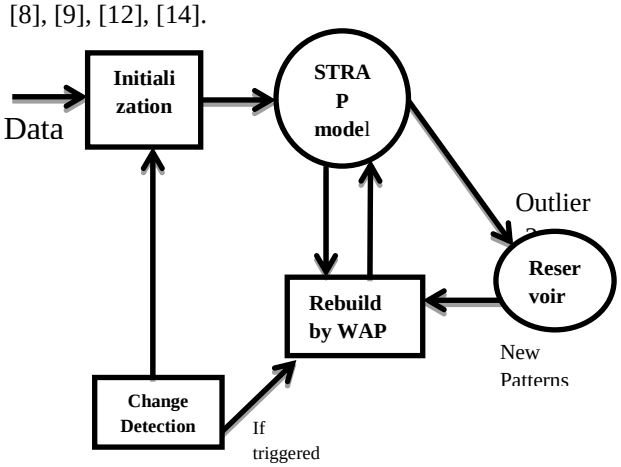

Fig.1: STRAP Algorithm Diagram.

## III- CONCLUSION

This paper reviews about various data stream clustering techniques and algorithms. In data mining and machine learning, to deal with non-stationary distribution of data and to handle large scale data sets in online fashion; many clustering algorithms are used. Data streams are dynamic ordered, fast changing, massive, limitless and infinite sequence of data objects. Data streams clustering technique are highly helpful to handle those data and outlier detection is one of the challenging areas in data stream. The clustering algorithms for data streams should be adaptive in the sense that up to date clusters are obtainable at any time, taking new data items into account as soon as they arrive. To overcome the drawbacks of previous techniques the STRAP clustering algorithm was used. It shows that how the online version of AP is useful in most of the clustering problems. The STRAP algorithm is to form a cluster of streaming data with the data distribution. The performance of this algorithm is measured in terms of the quality and purity of the clusters. The computation time depends on the complexity of underlying data distribution. It separate out the noise or outlier from the data items arrived in the system, and stores the outlier into the reservoir. STRAP provides the streaming model at any time. Clustering data by identifying exemplar data points rather than parametric methods allows for rich domain-specific models that can achieve superior result. Hence the need is to design new algorithms or to do improvements in existing ones. An alternative is to design distributed version of STRAP (e.g. sharing the reservoir), or to define the natural number of clusters or for AP to define the best value of preference penalty parameter. Another is to organize the STRAP model in a hierarchical manner so that retrieval phase in STRAP can be speed up.

## ACKNOWLEDGEMENTS

The author is thankful to MET's Institute of Engineering Bhujbal Knowledge City Nashik, HOD of computer department, guide, parents and friends for their blessing, support and motivation behind this work.

## REFERENCES

[1] Zhang, Ramakrishnan, and L. M., "BIRCH: An efficient data clustering method for very large databases'', presented at ACM SIGMOD Conference on Management of Data, 1996.
[2] L. O Callaghan, N. Mishra, A. Meyerson, S. Guha, and R. Motwani, "Streaming-data algorithms for high-quality clustering," 2002.
[3] C. Aggarwal, J. Han, J. Wang, and P. S. Yu, "A framework for clustering evolving data streams," in Proc. 29th Int. Conf. VLDB, Berlin, Germany, pp. 81–92, 2003
[4] B.-R. Dai, J.W. Huang, M.-Y. Yeh, and M.-S. Chen, "Adaptive clustering for multiple evolving streams," IEEE Trans. Knowl. Data Eng., vol. 18, no. 9, pp. 1166–1180, Sept. 2006.
[5] F. Cao, M. Ester, W. Qian, and A. Zhou, "Density-based clustering over an evolving data stream with noise," in Proc. SDM, pp. 326–337, 2006
[6] Y. Chen and L. Tu, "Density-based clustering for real-time stream data," in Proc. 13th ACM SIGKDD, New York, NY, USA, pp. 133–142, 2007
[7] C. Aggarwal, Data Streams: Models and Algorithms. Berlin, Germany: Springer, 2007.
[8] X. Zhang, C. Furtlehner, and M. Sebag, "Data streaming with affinity propagation," in Proc. ECML/PKDD, pp. 628–643., 2008
[9] X. Zhang, C. Furtlehner, J. Perez, C. Germain-Renaud, and M. Sebag, "Toward autonomic grids: Analyzing the job flow with affinity streaming," in Proc. 15th ACM SIGKDD, Paris, France, 2009.

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

*International Journal of Advanced Research in Computer and Communication Engineering*
*Vol. 4, Issue 11, November 2015*

[10] X. Zhang, W. Wang, K. Norvag, and M. Sebag, "K-AP: Generating specified K clusters by efficient affinity propagation," in Proc. IEEE 10th ICDM, Sydney, NSW, Australia, pp. 1187–1192., 2010
[11] M. Shindler, A. Wong, and A. Meyerson, "Fast and accurate Kmeans for large datasets," in Proc. NIPS, pp. 2375–2383., 2011
[12] X. Zhang, C. Furtlehner, C.G. Renaud, and M. Sebag, "Data Stream Clustering with Affnity Propagation ", IEEE Transaction on Knowledge and Data Engineering, vol. 26, No. 7, July 2014
[13] Madjid Khalilian, Norwati Mustapha "Data Stream Clustering: Challenges and Issues ", in Proc. International Multi-conference of engineers and scientists, vol-1, 2010.
[14] X. Zhang, C. Furtlehner, and M. Sebag, "Distributed and Incremental Clustering Based on Weighted Affinity Propagation", Jun 2008.
[15] Amini A,Wah TY, Saboohi H. "On density-based data streams clustering algorithms: A survey", Journal Of Computer Science And Technology, vol-29, No. 1, Jan. 2014


DOI 10.17148/IJARCCE.2015.41193