ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 9, September 2017

Clustering Data Streams Using K-Means Based on Shared Density between Micro-Clusters

Swapna Bhausaheb Ukarde, Prof.S.G.Tuppad

currently pursuing M.E.(CSE), Matsyodari Shikshan Sansthas College of Engineering and Technology, Jalna, Maharashtra, India

Assistant Professor, Matsyodari Shikshan Sansthas College of Engineering and Technology, Jalna, Maharashtra, India

ABSTRACT: In recent years,data stream clustering gets more focus in the literature.Data stream resist many more challenges on clustering because of its memory usage and processing speed.Data stream can be summarize in real time with an online process which is called as micro-clusters.Micro clusters are summation of matter having large amount of items so as to constitute new definite state of matter which shows local density estimates in specific area by muster the information of many data points.In this paper,we present two stage clustering, in first stage K-means ,to enhance the performance and scalability.K-means clustering aims to minimize thedistance of objects from the centroid of each group.In second stage,DBSTREAM micro-clustering,implements a simple density based stream clustering algorithm that assigns data points to micro-clusters with a given radius and implements shared density based reclustering via a shared density graph.The graph having density information is then utilized for reclustering based on actual density between adjacent micro-clusters.In this way, shared desity graph very helpful for recognizing the huge data stream clustering.

KEYWORDS: Data mining, data stream clustering, k-means clustering, DBSTREAM clustering.

I. INTRODUCTION

Data mining is the survey of extracting useful information from large sets of data. A commonly apply data mining procedure is clustering, the classification of objects into different groups by dividing sets of data into a series of subsets (clusters). An interpretation of clustering is the tracking of data applied to study (identify) changes in traffic patterns and to detect possible interruption.Cluster analysis is the task of grouping a large amount of objects in the same group which called as cluster,more similar to each other than to those who present in other groups.Data stream clustering is the clustering of data that appear continuously such as telephone records,multimedia data, financial transactions etc. . Micro clustering can be used to generate more robust and stable clustering results compared to a single clustering approach, perform distributed computing under privacy or sharing constraints, or reuse existing knowledge.

The paper gives a large consideration of a new approach to clustering huge amount of data.This approach contains two stage clustering .In first stage,clustering using K-means algorithm,where main aim to partition n observations into k clusters in which each observation belongs to cluster with the nearest mean,which serving as a prototype of the cluster and In second stage,DBSTREM,micro-cluster based online clustering algorithm that represent the density between micro-clusters with the help of shared density graph .The shared density graph captures the density of original data between micro-clusters while clustering the data items and after this represents how the graph useful for reclustering the micro-clusters.For the best effects,here we use the concept of shared density based reclustering technique for data stream clustering.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 9, September 2017

II. LITERATURE REVIEW

In this paper we have provide a one place package for understanding the features of streaming data, challenges of streaming data clustering, applications and tools of data stream clustering, and current scenario of literature on clustering of streaming data. The paper we will focus on the challenges and necessary features of data stream clustering techniques, review and compare the literature for data stream clustering by example and variable, describe some real world applications of data stream clustering, and tools for data stream clustering.[1]

Propose to analyze call networks as a spatio-temporal evolutionary stream Also discussed sampling at a precise level of socio-centric and ego-centric network Also discussed some prospective aspects of influence analysis such as family influence.[2]

Propose an unsupervised learning neural network named Density Based Self Organizing Incremental Neural network (DenSOINN) for data stream clustering tasks. DenSOINN is a self organizing competitive network that grows incrementally to learn suitable nodes to fit the distribution of learning data, Combining on- line unsupervised learning and topology learning by means of competitive Hebbian learning ruleTo Finding underlying clusters in data when the data is organized in the form of a stream. To adopt a self-adaptive distance metric to approximate the effect ofdata normalization, making the algorithm works on raw data as well as on normalized data. The network structure of DenSOINN is automatically built during learning procedure by means of Competitive Hebbian Learning.[3]

Present a distributed online learning scheme to classify data captured from distributed and dynamic data sources. Propose a novel online ensemble learning algorithm to update the aggregation rule in order to adapt to the underlying data dynamics Each learner uses a local classifier to make a local prediction. The local predictions are then collected by each learner and combined using a weighted majority rule to output the final prediction.[4]

Present a new algorithm that learns a Mahalanobis metric using similarity and dissimilarity constraints in an online manner. This approach hybridizes a Mahalanobis distance metric learning algorithm and a k-NN data stream classification algorithm with concept drift Detection. Finally, our algorithm is evaluated on different datasets by comparing its results with one of the best instance-based data stream classification algorithm of the state of the art.[5]

Present a novel experimental framework for both tasks. It offers the means for extensive evaluation and visualization and is an extension of the Massive Online Analysis (MOA) software environment released under the GNU GPL License To build an experimental framework for clustering data streams similar to the WEKA framework, so that it will be easy for researchers to run experimental data stream clustering benchmarks.[6]

Describe a new approach to using ensembles for stream classification. While the core method is straightforward, it is specifically designed to adapt quickly with very little overhead to the dynamic and evolving nature of data streams generated from non-stationary function shows particularly strong gain over the baseline method when ground truth is of limited availability to the classifiers.[7]

This paper proposes a real-time distributed intrusion detection system (DIDS) for the AMI infrastructure that utilizes stream data mining techniques and a multi-layer implementation approach. Using unsupervised online clustering techniques, the anomaly-based DIDS monitors the data flow in the AMI and distinguish if there are anomalous traffics. Identifying distributed security solutions to maintain the Confidentiality, integrity, and availability of AMI devices' Traffic is an essential requirement that needs to be taken into account.[8]

In this paper we studied a simple existing data stream clustering algorithm DenStream based on DBScan. Based on DenStream a novel data stream clustering approach “DDenStream” is proposed. DDenStream is a modified data stream clustering algorithm of DenStream. Find out arbitrary specs and good quality of clusters with noise. [9]

Propose a parameter free algorithm that automatically adapts to the speed of the data stream. It makes best use of the time available under the current constraints to provide a clustering of the objects seen up to that point. Our approach incorporates the age of the objects to reflect the greater importance of more recent data. ClusTree can maintain the same amount of micro clusters at stream speeds that are faster by orders of magnitude and that for equal stream speeds our granularity is exponential w.r.t. competing approaches.[10]

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 9, September 2017

III. SYSTEM ARCHITECTURE

Cluster collection combine multiple clustering of a set of objects into a single build up clustering, often referred to as the *union* solution. Micro clustering can be used to generate more robust and stable clustering results compared to a single clustering approach, perform distributed computing under privacy or sharing constraints, or reuse existing knowledge. In this work use the share market dataset to represent the micro clusters. In above dataset, the attributes are open price, high price, low price, close price respectively. Proposed system to address the cluster, organizing them in conceptual categories that bring out the common threads and lessons learnt while simultaneously highlighting unique features of individual approaches. A DBSREAM algorithm is to achieve comparable results.

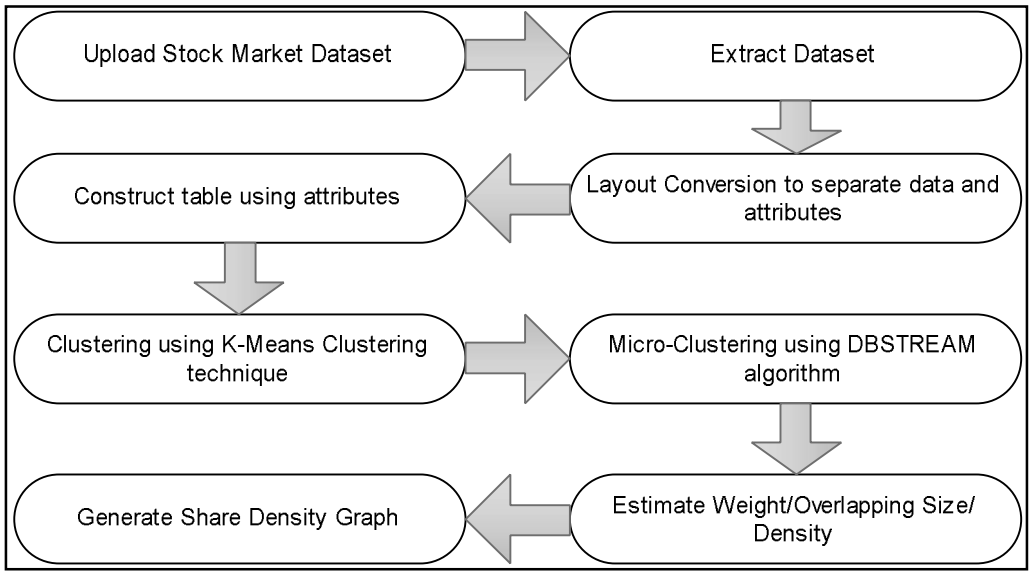

Fig. System Architecture

IV. METHODOLOGY OVERVIEW

Algorithm 1. Update DBSTREAM Clustering

Require: Clustering data structures initially empty or 0

MC - set of MCs

mcϵ MC has elements mc = (c, ω, t) - center, weight, last update time

S - Weighted adjacency list for shared density graph

$\boldsymbol{S_{ij}}\epsilon$ **S** has an additional field t - time of last update

t -current time step

Require: User-specified parameters

R - Clustering threshold

λ - Fading factor

$\boldsymbol{t_{gap}}$ - Cleanup interval

ω_{min} - Minimum weight

α - Intersection factor

1: function UPDATE(x)

2: $N \leftarrow$findFixedRadiusNN (x, *MC*, r)

3:if | N | < 1 **then** - create new MC

4: add(c=x, t=t, ω =1) to *MC*

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 9, September 2017

5:else - update existing MCs
6: **for each** iϵN **do**
7: $mc_i[\omega\ ] \leftarrow mc_i\ [\omega\ ]\ 2^{-\lambda(t-mc_i[t])} +1$
8: $mc_i\ [c] \leftarrow mc_i$ [c] + h(x, mc_i [c]) (x- mc_i [c])
9: $mc_i\ [t\ ] \leftarrow$ t - update shared density
10: **for each** j ϵN where j >I **do**
11: $S_{ij} \leftarrow S_{ij} 2^{-\lambda(t-S_{ij}[t])} +1$
12: S_{ij} [t] ← t
13: **end for**
14: **end for** - prevent collapsing clusters
15: **for each** (i, j) ϵNxN and j> i **do**
16: **if** dist ($mc_i\ [c]$, $mc_j\ [c]$) <r **then**
17: revert $mc_i\ [c]$, $mc_j\ [c]$ to previous positions
18: **end if**
19: **end for**
20: end if
21: t ← t+1
22: end function

Algorithm 2.Cleanup Process to Remove Inactive Micro- Clusters and Shared Density Entries from Memory.
Require:λ, ,t_{gap} , t, MC and S from the clustering.
1: functionCLEANUP ( )
2: $\omega_{weak} = 2^{-\lambda t_{gap}}$
3: **for each** *mc* ϵ *MC* **do**
4: **if** *mc*[ω] $2^{-\lambda(t-mc[t])}<\omega_{weak}$ **then**
5: remove weak *mc* from *MC*
6: **end if**
7: **for each** $S_{ij}\epsilon$ S **do**
8: **if** $S_{ij}2^{-\lambda(t-S_{ij}[t])}<\alpha\omega_{weak}$ **then**
9: remove weak shared density S_{ij} from S
10: **end if**
11: **end for**
12: **end for**
13: end function

Algorithm 3. Proposed Algorithm, K-means Clustering
Clustering is the process of partitioning a group of data points into a small number of clusters.
Clustering is an unsupervised or dividing pattern in groups or subgroups classification (i.e. clusters). Here the objects are grouped into classes of similar objects based on their location and connectivity within a space of dimension n. Mainly the principle of the grouping is to maximize similarity within a cluster, and to minimize the similarity between the groups.
Although there are many clustering algorithms available, one of the most used it is the k means
algorithm. Its aim is to minimize thedistance of objects from the centroid of each group. One of the most clustering algorithms used is k-means clustering, which is a partitioning method.For instance, the items in a supermarket are clustered in categories (butter, cheese and milk are grouped in dairy products). Of course this is a qualitative kind of partitioning. A quantitative approach would be to measure certain features of the products, say percentage of milk and others, and products with high percentage of milk would be grouped together. In general, we have *n* data points **x***i*,*i*=1...*n* that have to be partitioned in *k* clusters. The goal is to assign a cluster to each data point. K-means is a

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 9, September 2017

clustering method that aims to find the positions $\mu i, i=1...k$ of the clusters that minimize the *distance* from the data points to the cluster. K-means clustering solves

$\text{argmin}\mathbf{c}\sum i=1k\sum\mathbf{x}\in ci d(\mathbf{x},\mu i)=\text{argmin}\mathbf{c}\sum i=1k\sum\mathbf{x}\in ci\|\mathbf{x}-\mu i\|22$

where$\mathbf{c}i$ is the set of points that belong to cluster i. The K-means clustering uses the square of the Euclidean distance $d(\mathbf{x},\mu i)=\|\mathbf{x}-\mu i\|22$. This problem is not trivial (in fact it is NP-hard), so the K-means algorithm only hopes to find the global minimum, possibly getting stuck in a different solution.

The Lloyd's algorithm, mostly known as k-means algorithm, is used to solve the k-means clustering problem and works as follows. First, decide the number of clusters k. Then:

1. Initialize the center of the clusters	$\mu i=$ some value $, i=1,...,k$
2. Attribute the closest cluster to each data point	$\mathbf{c}i=\{j:d(\mathbf{x}j,\mu i)\leq d(\mathbf{x}j,\mu l), l\neq i, j=1,...,n\}$
3. Set the position of each cluster to the mean of all data points belonging to that cluster	$\mu i=1\|ci\|\sum j\in ci\mathbf{x}j,\forall i$
4. Repeat steps 2-3 until convergence	
Notation	$\|\mathbf{c}\|=$ number of elements in $\mathbf{c}$

The algorithm eventually converges to a point, although it is not necessarily the minimum of the sum of squares. That is because the problem is non-convex and the algorithm is just a heuristic, converging to a local minimum. The algorithm stops when the assignments do not change from one iteration to the next.

ACKNOWLEDGEMENT

I would like to thank Principal Dr. S.K.Biradar,MSSCET and HOD Prof. G.P.Chakote,ComputerScience,MSSCET,Jalna,Maharashtra for the support and guidance throughout the project.

V. CONCLUSION

Clustering data streams has keeps lots of attention in the last few years because of its ever-growing presence. Data stream clustering resist with lots of challenges due to the space complexity and time complexity.For achieving great results in terms of data stream clustering we propose this paper.

In this paper, we have developed and evaluate the data stream clustering, K-means algorithm which contains the smaller parts of clusters and then micro-clustering algorithm which records density in the area which is commonly stored by micro-clusters and keep this for reclustering.We include the shared density graph together this in order to achieve the comparable results.This helps us to improve the performance as well as scalability with better quality.

REFERENCES

[1] J. A. Silva, E. R. Faria, R. C. Barros, E. R. Hruschka, A. C. P. L. F. d. Carvalho, and J. A. Gama, "Data stream clustering: A survey," ACM Comput. Surveys, vol. 46, no. 1, pp. 13:1–13:31, Jul. 2013.
[2] M. Hahsler and M. H. Dunham, "Temporal structure learning for clustering massive data streams in real-time," in Proc. SIAM Conf. Data Mining, Apr. 2011, pp. 664–675.
[3] C. Isaksson, M. H. Dunham, and M. Hahsler, "Sostream: Self organizing density-based clustering over data stream," in Proc. Mach. Learn. Data Mining Pattern Recog., 2012, vol. 7376, pp. 264–278.
[4] A. Bifet, G. de Francisci Morales, J. Read, G. Holmes, and B. Pfahringer, "Efficient online evaluation of big data stream classifiers," in Proc. 21th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2015, pp. 59–68.
[5] J. Gama, R. Sebasti~ao, and P. P. Rodrigues, "On evaluating stream learning algorithms," Mach. Learn., vol. 90, pp. 317–346, 2013.
[6] H. Kremer, P. Kranen, T. Jansen, T. Seidl, A. Bifet, G. Holmes, and B. Pfahringer, "An effective evaluation measure for clustering on evolving data streams," in Proc. 17th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2011, pp. 868–876.
[7] A. Bifet, G. Holmes, R. Kirkby, and B. Pfahringer, "MOA: Massive online analysis," J. Mach. Learn. Res., vol. 99, pp. 1601–1604, Aug. 2010.
[8] M. Hahsler, M. Bolanos, and J. Forrest, Stream: Infrastructure for Data Stream Mining, 2015, R package version 1.2-2, r-project.org/package=stream.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 9, September 2017

[9] A. Amini and T. Y. Wah, "Leaden-stream: A leader density-based clustering algorithm over evolving data stream," J. Comput. Commun., vol. 1, no. 5, pp. 26–31, 2013.
[10] P. Kranen, I. Assent, C. Baldauf, and T. Seidl, "The clustree: Indexing micro-clusters for anytime stream mining," Knowl. Inf. Syst., vol. 29, no. 2, pp. 249–272, 2011.