ENGROSSMENT OF STREAMING DATA WITH AGGLOMERATION OF DATA IN ANT COLONY

Jagan K, Parthiban E and Manikandan B

ABSTRACT

A data stream is an ongoing process to appear a series of data and clustering data streams essential supplementary analysis to standard clustering. A stream is possible unlimited, data points appear online and each data point can be surveyed only once. This inflicts constraints on accessible memory and processing time. Moreover, streams can be strident and the number of clusters in the data and their statistical estate can change over time. This operation presents an online approach to clustering energetic data streams. A stochastic method is employed to find these rough clusters; this is shown to crucially speeding up the data with only a minor cost to performance, as compared to a deterministic approach. The rough clusters are then filtered using a method inspired by the discover sorting behavior of ants. Ants pick-up and drop items based on the correlation with the surrounding items. Artificial ant sort clusters by probabilistically picking and dropping micro clusters based on local density and local similarity in these operations.

Keywords - Data Stream, Local Density, Ants, Clusters and Data Points.

Reference to this paper should be made as follows:

Jagan K, Parthiban E Manikandan B,(2021), "Engrossment of Streaming Data with Agglomeration of Data in Ant Colony" Int. J. of Electronics Engineering and Applications, Vol. 9, No. 1, pp. 19-27, doi 10.30696/IJEEA.IX.I.2021.19-27.

Biographical notes:

Jagan K *is a student of Computer Science and Engineering from PERI Institute of Technology,in .*
His Email id : jagankannan22@gmail.com

Parthiban E *is a student of Computer Science and Engineering from PERI Institute of Technology,in .*
His Email id : parthielan1999@gmail.com

Manikandan B *is an Associate professor of Computer Science and Engineering from PERI Institute of Technology,in .*
His Email id : bmanibala@gmail.com

[I] INTRODUCTION

A data stream is a potentially unbounded sequence of data and in dynamic environments, the properties of this data can change over time in unforeseen ways. As a stream progresses, the performance of traditional classifiers and predictive Models target objects change [1-3]. This change can be gradual, known as concept-drift, sudden as a concept-shift, or in the form of concept-evolution when entirely new classes appear in the stream. When dealing with a continuous sequence of information, it is only possible to examine the data once. Clustering needs to be performed quickly to prevent bottlenecks and potential loss of data. A stream can be potentially infinite but only a limited amount of memory is available, necessitating the summarization of identified clusters in a meaningful way. Density-based clustering defines clusters, as high-density areas of the feature spaceseparated by areas of low density. It can identify arbitrarily shaped clusters, is robust to outliers and, crucially, does not require the number of clusters [4-6].

The fundamental problems in dealing with streams, the problem of being unable to revisit evolving data. A stream clustering should consist of two components: 1) an online component and 2) an offline component. Data arriving online should be summarized and the offline component should perform clustering of the summarized data. The performance of the k Nearest Neighbor (kNN) algorithm depends extremely on its being given a good measure over the input capacity. We construct that objects belonging to the cloned cluster usually share some common traits even though their commutative distance might be large. We, therefore, decided to define a measure based on clustering.

[2] LITERATURE REVIEW AND PREVIOUS WORK

Data Mining is applied to analyze large data sets and found useful patterns in the data. Data Mining is used in various fields to recognize patterns which are used in analyzation and prediction. Many studies describe how to analyze data by using classification, regression, correlation, clustering and machine learning [7]

The work by Aggarwal et al. was the first attempt to address one of the fundamental problems in dealing with streams, the problem of being unable to revisit evolving data [8-9]. The authors suggested that a stream clustering algorithm should consist of two components:

1) an online component and

2) an offline component.

Data arriving online should be summarized and the offline component should perform clustering on the summarized data. Their algorithm, CluStream, introduced the concept of micro-clusters as a method to summarize data. Micro clusters are a temporal extension of the cluster-feature-vector proposed in . In CluStream, only a certain number of micro-clusters can be stored in memory at any one time so when a new micro is formed, two existing micro-clusters must be merged or one deleted [10-13]. The offline clustering of the micro- clusters are based on the k-means algorithm Partitioned clustering algorithms have been extended for single-pass and stream clustering. In proposed to enable traditional soft partitioned clustering algorithms to deal with streaming data. The data stream is split into chunks and each chunk is partitioned into a set of cluster centroids. The centroids are weighted with the amount of samples they represent and in orderto maintain the history of the stream, previously identified centroids are added to the newly arriving chunk of data to be clustered. In , a constant factor approximation algorithm for a K- median approach to data streaming is described. These algorithms offer a fast, accurate single pass clustering of data but could be sensitive to changes in the underlying distribution or a shifting number of natural clusters. Many data mining systems often require the use of a clustering algorithm: this is the case for instance when one wants to reduce a huge data collection to a few clusters which can easily be studied, or when one wants to discover a structure within unstructured data [14]. Natural systems have evolved in order to solve many problems that can be related to clustering. Different species have developed social behaviors to tackle the problem of gathering objects or individuals. For instance, we can cite the brood sorting or cemetery organization of ants [FRA 92] or the collective movements in different species such as the ability of bacteria to form surprising spatial patterns and aggregations [CAM 01] [15-17]. Many researchers in computer science have been inspired by real ants [HOL 90] and have defined artificial ants paradigms for dealing with optimization or machine learning problems (see a review in [BON 99]). In this paper, we propose the adaptation of a new biological model which, as far as we know, has never been used before to solve computer science problems. We model the ability of ants to build live structures with their bodies [LIO 01] in order to discover, in a distributed and unsupervised way, a tree-structured organization of the data set. Each ant represents a data and is initially placed on a fixed point, called the support, which corresponds to the root of the tree. The behavior of an ant consists of moving on already fixed ants to fix itself to a convenient location in the tree.

[3] PROBLEMS IN EXISTING SYSTEM

In this existing system, it collects the data from the user and stored in the database in later part thee absorption of those data from the data stream that leads stores in the offline base system. Where combinations of those data which has high- density of those data feature is separated by low-density areas. The nature of an evolving stream implies that massing of data that can drift, new data can appear with the usability of data that can disappear and reappear cyclically. The tumbling window model is used to read a stream and rough clusters are incrementally formed during a single pass of a window. The levels of cluster purity are comparable across each dataset, but purity, in isolation, is not a very revealing evaluation metric as it does not consider the true topology of the data, A stochastic method is employed to find these rough clusters, this is shown to significantly speeding up the method with only a minor cost to performance, as compared to a deterministic approach. The results depend upon the distance measure. Fails to identify the data in thecombinationprocess. If the data density varies and the dataset it leads to sparse. The data that could not handle the data points with various values that have been stored with various data densities. Partitionable for multiprocessor systems is not applicable in this process. Datasets with altering densities are tricky while receiving the data.

[4] PROPOSED SYSTEM

In this paper, we proposed an adaptive data for the combination of alternating the data with local parameters. The admin has the accessibility to view the data which has been entered by the user. In the whole process of clustering, the global threshold is determined by the density distribution of all samples, and then the local threshold is self- adaptive. All sample density is sorted to search automatically for the data from the highest point of density of the distribution of all samples.

The user adds the details and the data which has been stored in a base system is received while the admin wants to share the data with the user. By getting the approval from the bank it gets approved to upload the details and it sends those details to the admin to get the sanction of the loan which has been applied by the user.

Once the admin gets the details of the user it allocates the amount which has to approve for the particular applicant on the basis of their application after the approval of the loan, the admin can view the payment details which have been paid by the user and it also has the details of that particular applicant. Build a secure connection between the user's mail transfer agent and the mail user agent. To produce better results in complex domains it can be applied to the data from any distribution.

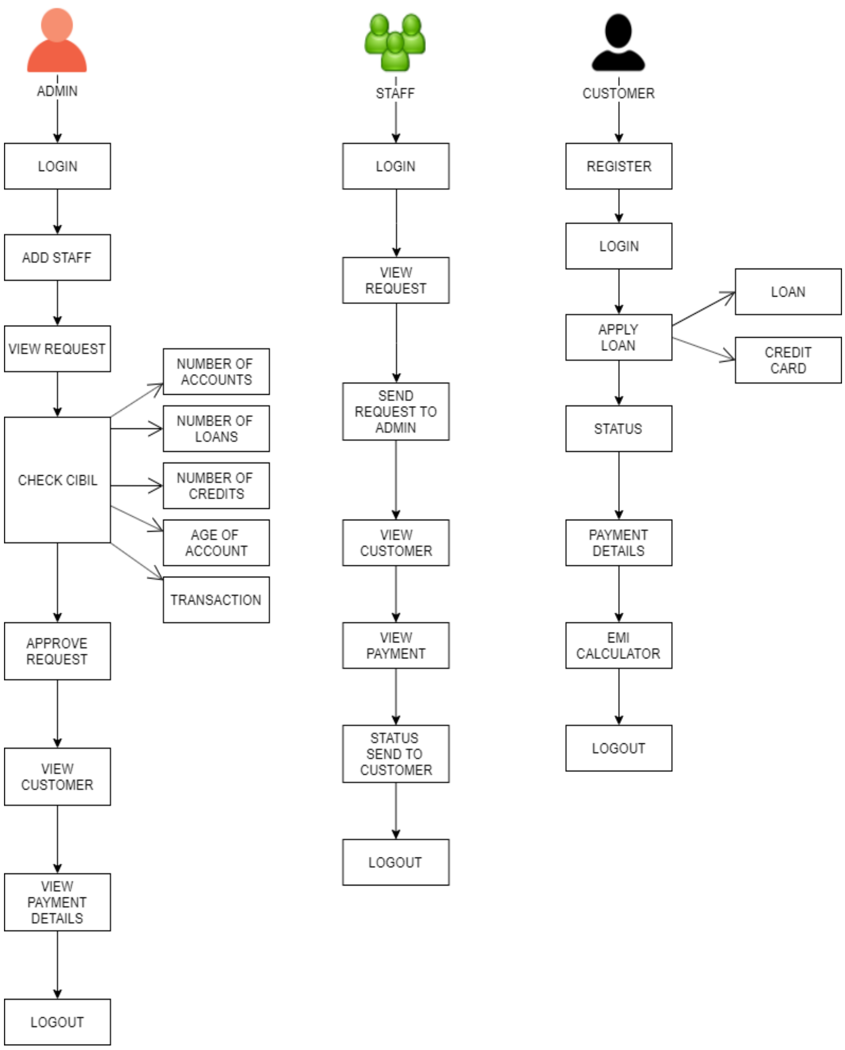

Fig.1. System Architecture

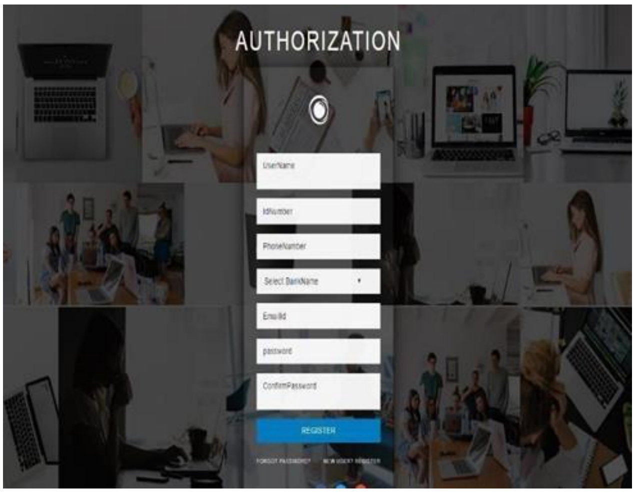

Fig.2. Customer Registration

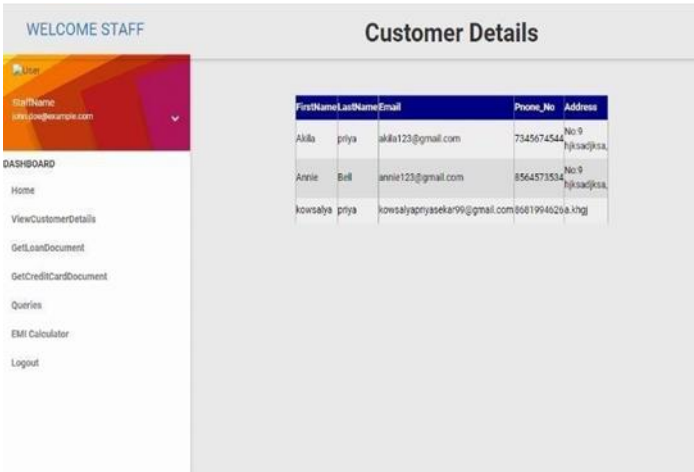

Fig.3. Customer Database

[5] AUTHENTICATION AND AUTHORIZATION

Authentication is about verification of your credentials such as Username/User ID and password to verify your personality. The system analysis, whether you are using your credentials or not. Usually, authentication is done with a username and password, although there are various ways to be authenticated. Authorization is the process to determine whether the authenticated user has access to the particular resources. It verifies your rights to grant you access to resources Such as information verified successfully then only it get permission to access the application.

[6] ARCHIEVE RESIGNATION

Archive resignation is the process of Apply loan and Credit card. Customer wants to register themselves to use this application. Once customer registered can able to access the features like EMI Calculator. Customer can able to know the status and the Payment history information's. If they need any Loan or Credit Card, they can raise the request. So that the request will be sent for the Staff to process further. They can get the status about the application it will be getting through mail.

[7] FAVOR UTILIZATION

Archive resignation is the process of Apply loan and Credit card. Customer wants to register themselves to use this application. Once customer registered can able to access the features like EMI Calculator. Customer can able to know the status and the Payment history information's. If they need any Loan or Credit Card, they can raise the request. So that the request will be sent for the Staff to process further. They can get the status about the application it will be getting through mail.

[8] TRACKER

Tracker is the process of get civil score. Staff will send the Check CIBIL request to the Admin if the Customer have already owned a Credit Card or availed the Loan from the particular bank. . In this tracker will give the details of customer like Number of account, Number of loans and credit card, Age of account and Name of Bank. Admin can approve or decline the request based on the conditions.

[9] DESCRIPTION CONTROL

Description control based on the customer Aadhar number and PAN number will check the process of customer details. In this all process done by the admin. Customer has the process of raise the query to staff and staffs will response to the query. If the application once approved then continue the further process of loan or credit card.

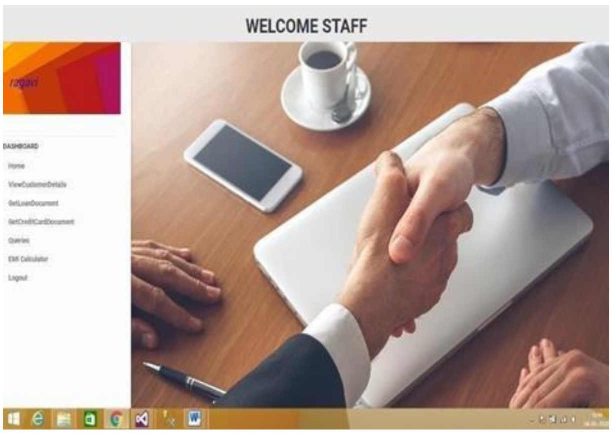

Fig.4. Staff Homepage

Fig.5. Staff Database

Density-based clustering defines clusters as high-density areas of the feature space separated by areas of low density. It can identify arbitrarily shaped clusters, is robust to outliers and, crucially, does not require the number of clusters to be known a priori. In our proposed algorithm, dense areas are described using micro-clusters: n dimensional spheres with center c and radius r. Micro-clusters have a maximum radius _ where r ≤ _. A data point is assigned to a micro-cluster if the point falls within its radius. The set of micro-clusters that are connected form the macro-cluster. Generally, there are more micro-clusters than there are actual clusters but significantly fewer micro-clusters than there are data points. This serves a dual purpose both as the clustering mechanism and as a summarization technique because a number of local data points can be represented by a single micro-cluster. Clusters identified by the algorithm are summarized by their constituent micro clusters and these summaries are stored offline for evaluation by the user.

[10] CONCLUSION

This application is performed to get the amount that has been approved by the admin. The user gets the details of the applicants and sends those details to the admin by the approval of the data which has been shared from the database. The user can view the details of the applicant and also the payment which has been paid by the applicant after receiving the approval from the admin the user can process by paying the amount on their basis of the loan. By getting particular id of the applicant the admin and the user can view the details of the applicant in different basis with the details they have registered and the payment details. Further the whole part can be done by the admin by approving the details of the applicant and sending the due payment notification and to remind the applicant to pay this amount on time by giving prior message to the applicant which is helpful in time consumption.

REFERENCES

[1] C. C. Aggarwal, J. Han, J. Wang, and P. S. Yu, “A framework for clustering evolving data streams,” in Proc. 29th Int. Conf. Very Large Data Bases, vol. 29. Berlin, Germany, 2003, pp. 81–92.

[2] H. Azzag, N. Monmarche, M. Slimane, and G. Venturini, “AntTree: A new model for clustering with artificial ants,” in Proc. IEEE Conf. Evol. Comput., vol. 4. Canberra, ACT, Australia,2003,pp.2642–2647.

[3] R. D. Baruah and P. Angelov, “DEC: Dynamically evolving clusteringand its application to structure identification of evolving fuzzy models,” IEEE Trans. Cybern., vol. 44, no. 9, pp. 1619–1631, Sep. 2014.

[4] J. C. Bezdek, R. Ehrlich, and W. Full, “FCM: The fuzzy C-means clustering algorithm,” Comput.Geosci., vol. 10, nos. 2–3, pp. 191–203, Jan. 1984.

[5] A. Bifet, G. Holmes, R. Kirby, and B. Pfahringer, “MOA: Massive online analysis,” J. Mach. Learn. Res., vol. 11, pp. 1601–1604, May 2010.

[6] U. Boryczka, “Finding groups in data: Cluster analysis with ants,” Appl. Soft Comput., vol. 9, no. 1, pp. 61–70, Jan. 2009.

[7] JharmaMajumdar, Sneha Naraseeyappa, ShilpaAnakalaki,”Analysis of agriculture using datamining techniques: application of big data”, Journal of Big Data, Dec (2017).

[8] A Mucherino, PapajorgjiPetraq, P. M. Pardalos, "A survey of data mining techniques applied to agriculture”, Springer-verlag, 2009.

[9] Ramesh A. Medar, Vijay. S. Rajpurohit, "A Survey of data mining techniques for crop yield prediction", IJARCSMS International Journal of Advance Research in Computer Science and Management Studies, vol. 2, no. 9, pp. 59-64, September 2014.

[10] Rakesh Kumar, M.P. Singh, Prabhat Kumar, J.P. Singh, "Crop Selection Method to Maximize Crop Yield Rate using Machine Learning Technique", International Conference on Smart Technologies and Management for Computing Communication Controls Energy and Materials (ICSTM), 2015.

[11] Abu Salim, Sachin Tripathi and Rajesh Kumar Tiwari “A secure and timestamp-based communication scheme for cloud environment” Published in International Journal of Electronic Security and Digital Forensics, Volume 6, Issue 4, 319-332.

[12] Rajesh Kumar Tiwari and G. Sahoo, “A Novel Watermark Scheme for Secure Relational Databases” Information Security Journal: A Global Perspective, Volume 22, Issue 3, July 2013.

[13] V. Ramesh and K. Ramar, 2011. “Classification of Agricultural Land Soils: A Data Mining Approach”, Agricultural Journal, 6: 82-86.

[14] N.Hemageetha,”A Survey on application of datamining techniques to analyse the soil for agriculture purpose”, INDIA Com, 2016.

[15] UnmairAyub, Syed AtifMosqurrab, ”Prediction cross diseases using data mining approaches:Classification”,ICPESG,2018.

[16] D.Ramesh, VishnuVaradhan, “Data Mining Techniques and Application to Agricultural Yield Data”, IJARCCE, 2013.

[17] Anoop Joyti Sahoo, and Rajesh Kumar Tiwari “A Novel Approach for Hiding Secret data in Program Files” International Journal of Information and Computer Security. Volume 8 Issue 1, March 2016,