Data Clustering using Modified Genetic Algorithm in Data Mining

Rahul Malviya [1],Prof. S. R. Yadav[2]

[1]*M.Tech Student* ,[2]*Assistant Professor and Head*

Department of CSE, MITS, Bhopal, India

Abstract: Digital world is growing as much faster beyond the thinking. Every digital process produces and uses lots of data on daily basis. It must need to store that data production into future uses format. So many digital equipment and sensor network produce the very precious data stream. Data stream consist variant parameters that are correlated with each other. In this paper, It proposed the method to cluster the unsupervised data stream to uncorrelated supervised cluster based on Principle component analysis. It mines dynamically changing sensor streams for states of system. It can be used for detecting the current state as predicting the coming state of system. So, in this way, It can monitor the behavior of the system so that any interesting events, such as anomalies or outliers can be found out.

Keywords: Clustering, Data Mining, Data stream, Unsurprised, Principle Component Analysis, Clustering Algorithm.

I. INTRODUCTION

Recent advancement in both hardware and software has alloyed large scale data acquisition. Typical statistical and data mining methods (e.g., clustering, regression, classification and frequent pattern mining) work with "static" data sets, meaning that the complete data set is available as a whole to perform all necessary computations. It will known methods like k-means clustering, linear regression, decision tree induction and the APRIORI algorithm to find frequent item sets scan the complete data set repeatedly to produce their results. However, in recent years more and more applications need to work with data which are not static, but are the result of a continuous data generating process. Data streams are ordered and potentially unbounded sequences of data points created by a typically non-stationary data generating process**.**

Nevertheless, dealing with huge amount of data poses a challenge for researchers, due to the limitations of current computational resources. For the last decade, It have seen an increasing interest in handling and managing these massive, unbounded sequences of data that are continuously generated at rapid speed over the period of time, the so-called data streams. More formally, a data stream S is a large sequence of data objects$X1, X^2,\ldots, X^N$, that is, $S=\{X^i\}$ where i =1 to N, which is potentially unbounded ($N\longrightarrow \infty$). Each data object is described by an n-dimensional attribute vector$Xi = [x^i_1, x^i_1, x^i_2, \ldots, x^i_n]$, it is an attribute space omega that can be continuous, categorical, or mixed [6].

Data stream mining is very popular research area that has recently emerged to find knowledge from large amounts of continuously generated data. Such as from network flows, sensor data and click stream. Examination and mining of such kind of data has been become very hot topic. Detecting hidden patterns in data stream give a great challenge for cluster analysis.

The main goal of clustering is to group the stream data and give some meaningful classes. Clustering problem needs to be defined carefully in this area. This is because a data stream should be view as an infinite process of continuously evolving data over time. For example, in network monitoring data, the TCP connection records of LAN (or WAN) network traffic, form a data stream. User network connection pattern often change gradually over time.

Evolving data streams have following requirements for clustering:

1. It cannot make any assumption on the number of clusters, because the number of clusters is unknown. Furthermore, in evolving data stream, the number of clusters usually keeps changing.
2. Discovery of clusters with any arbitrary shape: This is very important for every data stream applications. For example, in network monitoring data, the distribution of connection is usually irregular. In real time environment, the layout of area could be in any shape.

Several clustering algorithms are proposed to perform unsupervised learning. But data stream clustering poses several challenges, such as dealing with the unbounded data which comes in online fashion. The inherent nature of stream data requires algorithm that is capable of processing of data objects, suitably addresses memory and time limitations. Many data stream algorithm have been

proposed in recent years but they are not efficient when It talk about context detection.

Context in machine learning can be described as any information that helps in finding entity's behavior and context detection is finding hidden patterns in dataset. Context-aware systems offer entirely new opportunities for end user and application developer by gathering stream data and adapting system behaviour accordingly. Context awareness, described as an idea constantly observes an entity's context over a period of time for some applications. For example, an application of context awareness is task of finding data-leakage and its prevention for smart phones. In this, locomotion (e.g. walking or running) is the tracked context and outgoing emails are the behavior of interest. By tracking the context, a machine learning algorithm assumes that It cannot send an email while the user is running. Another example is of network monitoring dataset. In this example context is finding whether a TCP record is under attack or normal [18].

II. RELATED WORK

Detecting changes in multidimensional data streams is an important and challenging task. In this paper they present the new frame work to detecting abrupt changes in multidimensional data streams. The framework is based on projecting data on selected principle components. On each projection, densities in reference and test windows are estimated and compared. Then a change-score value is calculated by one of the divergence metrics. By treating all selected PCs with equal importance, the maximum change-score among different PCs is considered as the final change-score [27]. (Abdulhakim Qahtan, Basma Alharbi, Suojin Wang, Xiangliang Zhang; 2015)

Dealing with non-stationary distributions is a key issue for many application domains, e.g., sensor network or traffic data monitoring This method focuses on learning a generative model of a data stream, with some specific features: the generative model is expressed through a set of exemplars, i.e., actual data items as opposed to centroids in order to accommodate complex (e.g., relational) domains; the generative model is available at any time, for the sake of monitoring applications; the changes in the underlying distribution are detected through statistical hypothesis testing. The proposed algorithm is known as STRAP. In this proposed algorithm STRAP aims at clustering data streams with evolving data distributions [2]. (Xiangliang Zhang, Cyril Furtlehner; 2014)

The DenStream algorithm consists of an online part, which maintains a set of micro-clusters as a summarized representation of the data distribution, and an offline part, which generates the final clusters on demand. The DenStream algorithm uses the damped window model [4] to express the idea that recent data objects more accurately mirror the current data distribution than older ones. (Paul Voigtlaender; 2013)

They propose a parameter-free stream clustering algorithm ClusTree that is capable of processing the stream in a single pass, with limited memory usage. It always maintains an up-to-date cluster model and reports concept drift, novelty, and outliers. Moreover, their approach makes no apriori assumption on the size of the clustering model, but dynamically self-adapts. For handling of varying time allowances, It propose an anytime clustering approach. Anytime algorithms are capable of delivering a result at any given point in time, and of using more time if it is available to refine the result. (Philipp Kranen; 2011)

III. METHODOLOGY

Main approach of the core PCMGAA algorithm is to follow the stream's data distribution. After that It get the fuzzy membership scores by calculating similarity scores betIten a known context and point. When It see that an arriving point is inside a distribution of an existing context, It assign that point to that context. So It can see that all the points now belong to that context only. When the stream's distribution doesn't fit in any existing context, It has to define a new context. It has contexts which have limited memory for allowing concept drifts. Allocated memory space should be filled; It have to use one of the two methods for merging context model is performed. First method is method is to equally interleave the primary m observations and other method is to merge the memories in the order of their timestamps. Point anomalies are detected as short-term drift away from all present contexts that means our point deviates from all present contexts. There are mainly four parameters that affect the contexts, parameter for PCMGAA: the sensitivity threshold φ, the context drift size $t_{\min}$, the model memory size m, and at last the percent of variance to retain in projections ρ.

The pseudo code of methodology is as follows:

Input parameters $\{\varphi, t_{\min}, m, \rho\}$

Anytime Outputs: $\{c_t, d_c(\overrightarrow{x_t})\}$

1. $\boldsymbol{C} \leftarrow$ init(S, t_{min}, m, ρ)

2. $c_t \leftarrow c_1$

3. $D \leftarrow \emptyset$

4. *loop*

    4.1. $\overrightarrow{x_c} \leftarrow$ next(S)

    4.2 $scores \leftarrow d_c(\overrightarrow{x_c})$

4.3. $i \leftarrow$ Index_Of_Min($scores$)

4.4. if scores(i)< φ

4.4.1. Upd_Model(c_i, $Dump(D)$)

4.4.2. Upd_Model(c_i, $\overrightarrow{x_c}$)

4.4.3. $c_t \leftarrow c_i$

4.5. Else

4.5.1. insert($\overrightarrow{x_c}$, D)

4.5.2. if $length(D)$== $tmin$

4.5.2.1. $c \leftarrow CreateModel(Dump(D), m, \rho)$

4.5.2.2. $AddModel(c, \boldsymbol{C})$

4.5.3. $c_t \leftarrow c_i$

5. $end\ loop$

IV. RESULT AND ANALYSIS

The proposed method implements in MATLAB R2013a and tested with very reputed data set from UCI machine learning research center. In the research work, I have measured CR (Correct Rate), ER (Error Rate), PPV (Positive Predictive Value), NPV (Negative Predictive Value), PL (Positive Likelihood) and NL (Negative Likelihood) of Clustering. To evaluate these performance parameters, It has been used three datasets from UCI machine learning repository namely glass dataset, banana dataset and forest fire dataset.

Consider glass dataset input the chunk size as 100 then following window has been displayed

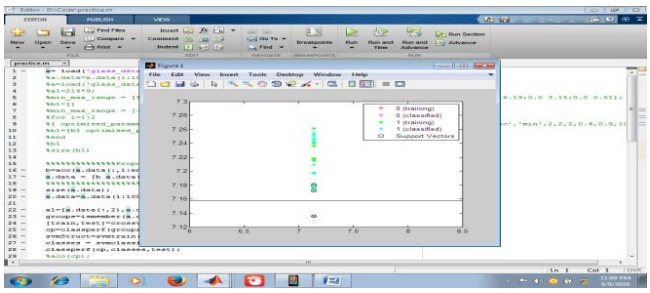

Figure 1: Data clustering using PCMGAA method for the given chunk size 100 in glass dataset

Comparison table for the PCA and PCMGAA method on the basis of given different chunk size and find the value of CR, ER, PPV and PL.

Table 1: Show that the calculated result of feature selection on different data size on the basis of two methods PCA and PCMGAA for glass data set.

Input Data Size	PCA				PCMGAA			
	CR	ER	PPV	PL	CR	ER	PPV	PL
80	51.28	48.72	0.221	1.63	82.05	17.94	1	1.77
90	56.8	43.18	0.406	1.56	70.45	29.54	0.5	2.38
100	63.26	36.73	0.41	1.22	65.31	34.69	0.57	2.29
110	53.7	46.3	0.47	1.24	55.56	44.44	0.45	1.12

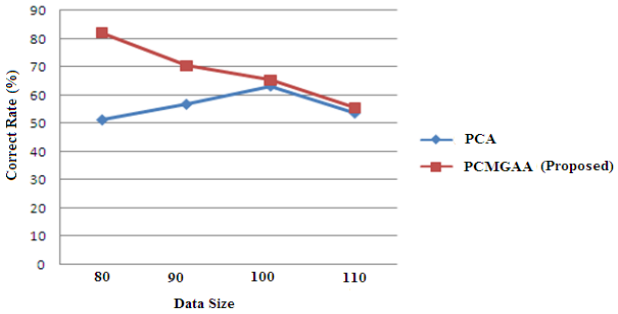

Figure 2: shows that the comparison graph correct rate of both methods PCA and PCMGAA for glass data set

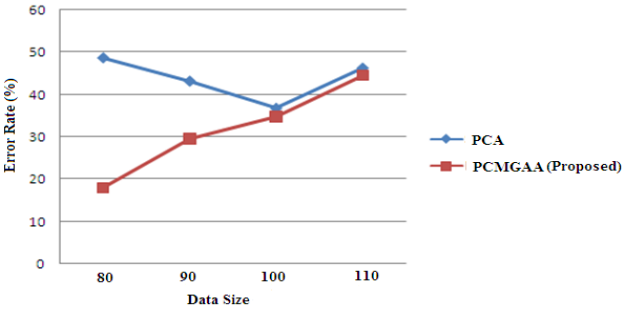

Figure 3: Shows that the comparisons graph betIten error rate of both methods PCA and PCMGAA for glass data set.

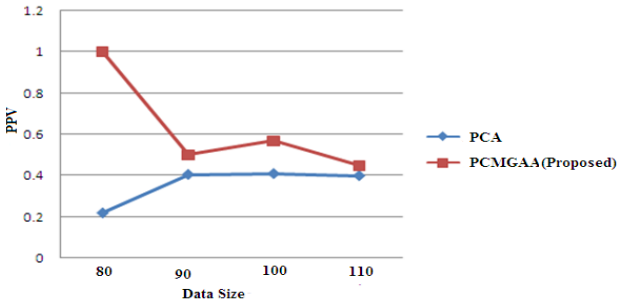

Figure 4: Shows that the comparisons graph betIten PPV of both methods PCA and PCMGAA for glass data set.

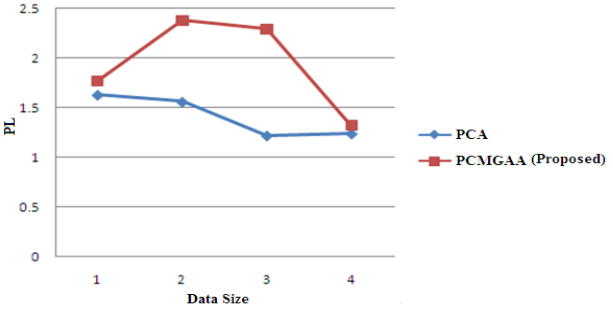

Figure 5: Shows that the comparisons graph betIten PL of both methods PCA and PCMGAA for glass data set.

V. CONCLUSIONS

Traditional stream clustering techniques also make impractical assumptions about the availability of labeled data. Most techniques assume that the true label of a data point can be accessed as soon as it has been classified by the clustering model. Thus, according to their postulation, the existing model can be updated without delay using the labeled instance. In reality, one would not be so lucky in obtaining the label of a data instance immediately, since manual labeling of data is time consuming and costly. It claims two major contributions in novel class detection for data streams. First, it propose a dynamic selection of boundary for outlier detection by allowing a slack space outer the decision boundary. This space is restricted by a threshold, and the threshold is modified all the time to reduce the risk of false alarms and missed novel classes. Principle Component with Modified Genetic Algorithm Analysis is very efficient data mining tool for data clustering. Data clustering is challenging task in the field of clustering. Evaluation of new feature creates a problem in feature selection during the clustering process of modified Principle Component Analysis. In this dissertation reduces these problems using genetic algorithm, it is used to control new feature evolution problem. Genetic algorithm creates a feature prototype for cluster used in clustering. The controlled feature evaluation process proposed a Principle Component with Genetic Algorithm Analysis is called PCMGAA.

The empirical evaluation of modified algorithm is better in comparison of PCA algorithm. The error rate of modified algorithm decreases in comparison of PCA algorithm. Also improved the rate of Correct, PPV and PL for evolution of result, after these improvement still some problem is still remain such as infinite length and data drift. Infinite length and data drift problem are not considered in this dissertation.

6. SCOPE OF FUTURE WORK

The proposed method Principle Component with Modified Genetic Algorithm Analysis solved the problem of feature evaluation and concept evaluation. The controlled feature evaluation process increases the value of correct rate and reduces the error rate. The genetic algorithm optimization cluster faced a problem of right number of cluster, in future used self optimal clustering technique along with other optimization technique is as particle of swarm optimization, min max ant colony optimization etc.

REFERENCES

[1] Abdulhakim Qahtan, Basma Alharbi, Suojin Wang, Xiangliang Zhang, “A PCA-Based Change Detection Framework for Multidimensional Data Streams”, KDD’15, August 10-13, 2015, Sydney, NSW, Australia. ACM 978-1-4503-3664-2/15/08.

[2] Xiangliang Zhang, Cyril Furtlehner, Cecile Germain-Renaud, Mich`ele Sebag. “Data Stream Clustering with Affinity Propagation”. IEEE Transactions on Knowledge and Data Engineering, Institute of Electrical and Electronics Engineers, 2014.

[3] Paul Voigtlaender, “DenStream”, Rheinisch-Itstfälische Technische Hochschule Aachen Lehrstuhl für Date management and –exploration, January 22, 2013.

[4] Philipp Kranen, Ira Assent, Corinna Baldauf, Thomas Seidl “The ClusTree: indexing micro-clusters for anytime stream mining ” Springer-Verlag London Limited, November 2011, Volume 29, Issue 2, pp 249–272, DOI 10.1007/s10115-010-0342-8.

[5] Yixin Chen, “Density-based clustering for real-time stream data”, KDD '07 Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining Pages 133-142 San Jose, California, USA — August 12 - 15, 2007.

[6] T. Dasu, S. Krishnan, S Venkata subramanian, “An Information-Theoretic Approach to Detecting Changes in Multi-Dimensional Data Streams”, In Proc. Symp. on the Interface of Statistics, Computing Science, and Applications, 2006.

[7] Daniel Kifer, Shai Ben-David, Johannes Gehrke, “Detecting Change in Data Streams” Proceedings of the 30th VLDB Conference, Toronto, Canada, 2004.

[8] Charu C. Aggarwal, JiaIti Han, Jianyong Wang, Philip S. Yu “A Framework for Clustering Evolving Data Streams”, VLDB’03 Paper ID: 312, Proceedings of the 29th VLDB Conference, Berlin, Germany, 2003.

[9] Martin Ester, Hans-Peter Kriegel, Jiirg Sander, XiaoIti Xu, A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise, KDD-96 Proceedings. Copyright © 1996, AAAI (www.aaai.org).

[10] Tian Zhang Raghu Ramakrishnan Miron Livny”, BIRCH: An Efficient Data Clustering Method for Very Large Databases, SIGMOD ’96 6/96 Montreal, Canada IQ 1996 ACM 0-89791 -794-4/96/0006.

[11] Harries, M.B. et al.: Extracting hidden context. Machine learning.32, 2, 101–126 (1998).

[12] [12]. Hubert, L., Arabie, P.: Comparing partitions. Journal of classification.2, 1, 193–218 (1985).

[13] Jolliffe, I.: Principle Component Analysis. Encyclopedia of Statistics in Behavioral Science.John Wiley & Sons, Ltd (2005).

[14] Ka Yee Yeung, Walter L.Ruzzo :Detail of the Adjust Rand Index and Clustering algorithms Supplement to the paper “An empirical study on Principle Component Analysis for clustering gene expression data” .

[15] Liu, W. et al.: A survey on context awareness. Computer Science and Service System (CSSS), 2011 International Conference on.pp.144–147 IEEE (2011).

[16] Maesschalck, R.D. et al.: The Mahalanobis distance. Chemometrics and Intelligent Laboratory Systems.50, 1, 1–18 (2000).

[17] Makris, P. et al.: A Survey on Context-Aware Mobile and Wireless Networking: On Networking and Computing Environments' Integration. Communications Surveys & Tutorials, IEEE.15, 1, 362–386 (2013).

[18] Padovitz, A. et al.: Towards a theory of context spaces. Pervasive Computing and Communications Workshops, 2004.Proceedings of the Second IEEE Annual Conference on.pp.38–42 IEEE (2004).

[19] Shlens, J.: A tutorial on principle component analysis. arXiv preprint arXiv:1404.1100. (2014).

[20] Wold, S., Sjostrom, M.: SIMCA: a method for analyzing chemical data in terms of similarity and analogy. Presented at the (1977).

[21] Yang, Y. et al.: Mining in Anticipation for Concept Change: Proactive-Reactive Prediction in Data Streams. Data Mining and Knowledge Discovery.13, 3, 261–289 (2006).

[22] [22]. Amineh Amini,The Ying Wah: Density Micro Clustering Algorithm on Data Stream. IMECS 2011 Hong Kong.

[23] J. Gama. Knowledge Discovery from Data Streams.Chapman & Hall/CRC, 2010.

[24] J. Gama, I. Zliobaite, A. Bifet, M. Pechenizkiy, and A. Bouchachia.A survey on concept-drift adaptation.ACMComputing Surveys, 46(4), 2014.

[25] Daniel Kifer, Shai Ben-David, Johannes Gehrke, "Detecting Change in Data Streams" Proceedings of the 30th VLDB Conference, Toronto, Canada, 2004.

[26] Tian Zhang, Raghu Ramakrishnan, Miron Livny", BIRCH: An Efficient Data Clustering Method for Very Large Databases, SIGMOD '96 6/96 Montreal, Canada IQ 1996 ACM 0-89791 -794-4/96/0006

[27] Abdulhakim Qahtan, Basma Alharbi, Suojin Wang, Xiangliang Zhang, "A PCA-Based Change Detection Framework for Multidimensional Data Streams", KDD'15, August 10-13, 2015, Sydney, NSW, Australia. ACM 978-1-4503-3664-2/15/08. DOI: http://dx.doi.org/10.1145/2783258.2783359.

[28] Philipp Kranen, Ira Assent, Corinna Baldauf, Thomas Seidl "The ClusTree: indexing micro-clusters for anytime stream mining " Springer-Verlag London Limited, November 2011, Volume 29, Issue 2, pp 249–272, DOI 10.1007/s10115-010-0342-8.