Preserving Privacy Using Geometric Transformation in Data Stream

[1]Darshini U. Patel, [2]Maulik Dhamecha
[1]Student of Computer Engineering, [2]Assistant Professor , Department Of CE
[1] Department of Computer Engineering,
[1]R.K University, Rajkot, India

***Abstract* - Data stream can be characterize as a continuously changing sequence of data that come over the framework constantly for storing or handling. Illustrations identified with data streams incorporate PC system activity, web inquiries and sensor information and so on. The proprietors of the information may not be willing to uncover the accurate estimations of their data because of a few reasons, most likely privacy concern. In this way, for protecting information security amid data mining, prevacy preserving mining issue has been generally studied over and even numerous methods have been proposed. Yet at the same time the systems that have been intended for protection safeguarding data mining are for customary static information sets just and are not for data streams. So this issue for privacy preserving of data streams mining is required for the time. This paper essentially centered around strategies for Principal Component Analysis (PCA) based change for stream information utilizing Massive Online Analysis (MOA). The clustering accuracy is verging on equivalent to the first dataset utilizing perturbe data.**

***Index Terms* – Data Stream, Geometric Transformation, Data Perturbation, Random Function**

I. INTRODUCTION

In the field of information handling , data mining is describe as the system of isolating the important data from the boundless information. Data mining strategies are by and large use in scope of utilization which joins Clustering, Classification, Regression examination and Association oversee or Pattern mining.

The data stream has starting late been familiar as needs be with the issues and challenges related with nonstop data [1]. Data stream mining is use to focus discovering that is addressed in models and illustrations in a non ending and relentless stream of information. Algorithm for data streams can adjust to data sizes usually more unmistakable than memory and can be connected with constant applications not already taken care of by machine learning or information mining.

These days, in the field of data dealing with, different application don't fit in this information model [2] Instead, information happens as a course of action or stream of information. An information stream is portray as a consistent, relentless, and requested movement of things. There is scarcest possibility to control the solicitation in which the things arrive moreover it is not doable to for the most part store an entire stream. Thus in the same way, inquiries run continuously with the time of time over the stream and return new results according to the new data landing.

II. PRIVACY CONCERN FOR DATA STREAM

Data stream Mining is characterize as extracting knoweledge and is spoken to in models and examples in continuous floods of data. An exploration range called privacy-preserving data mining has been presented from the inspiration by the privacy concerns on data mining apparatuses.

Verykios et al. [3] ordered privacy protecting data mining systems are in view of five measurements i.e. data distribution, data modification, data mining algorithms, rule hiding, and privacy preservation. In data circulation measurement, some methodologies have been proposed for centralized data and some for distributed data.

Du and Zhan [4] utilize the secure union, secure sum and secure scalar product item to keep the first information of each dataset from revealing the mining procedure. The disadvantage of this methodology is that it requires various sweeps of the database which is not suitable for data streams which streams constantly quick and requires prompt reaction.

In data modification dimension, the private information of a database are adjusted first to protect information security before uncovering to open. The methodologies include perturbation, blocking, total or consolidating, swapping, and testing. Agrawal and Srikant [5] utilized the arbitrary information annoyance strategy for ensuring client information and later develop the choice tree. For information streams, as information are alterable that implies it arrives persistently at diverse time, its information dissemination will likewise be change and its exactness will likewise diminish with the annoy information.

Consequently from the past examination result it can be said that all the current strategies for privacy safeguarding data digging are intended for static databases just for data privacy. Along these lines, the current procedures are not suitable for data streams.

Perturbation procedures are for the most evaluated with two essential measurements: level of privacy guarantee and level of data utility saved, which can be offen measured by the accuracy misfortune for data classification and data clustering. Therefore, the fundamental objective for all data algorithm is to enhance the data evaluating the first information from the pertubated data change prepare by amplifying both data privacy and data utility.Data perseving is measured by the trouble in evaluating the first

information from the perturbe data. For a given data perturbation strategy, when the level of trouble is higher from which the first information can be assessed from theperturbated information, around then the level of information security is higher. Data utility generally allows to the measure of discriminating data saved from the first information set after perturbation.

III. PRIVACY PRESERVING DATA STREAM CLUSTERING

The computation of data stream model orequires algorithms to make a solitary go for the data, with constrained memory and restricted processing time in light of the fact that the stream may be exceptionally dynamic and developing after some time. A few new approachs have been created for successful grouping of stream idata, for example, Compute and store synopses of past information, because of restricted memory space and quick reaction necessity, Compute outlines of the past information, store the individual results, and utilize such rundowns to figure essential measurements when needed.

The primary thought behind Perturbation-Based procedure is to build a noise in the accessible raw data andperturbate the original data distribution and to protect the genuine content of original crude data. Geometric Data Transformation Methods (GDTMs) [6] is one basic run of the mill illustration of data perturbation techniques. It will bother numeric information wit h private credits in group mining to save security.

In any case Kumari et al. [7] proposed a protection safeguarding grouping system of Fuzzy Sets, it changes the classified credits into fuzzy items to preseve privacy. Futher, the issue that happen is when executing a bother method is the inacurate mining result from an perturbation data.

Vaidya and Clifton [8] proposed the strategy for privacy preserving clustering system base on vertically partitioning data. In the vertical partitioning the attributes of the same objects are split across the partitions.

Unexpectedly, Meregu and Ghosh [9] proposed the strategy for privacy preserving cluster mining base on the Horizontally data partitioning.It is the system Of "Privacy- preserving Distributed Clustering using Generative Model." In this methodology, as opposed to sharing parts of the first data or perturbe information, the parameters of suitable generative models are manufactured at every neighborhood site.

In [10] proposed a technique for Privacy-Preserving Clustering of Data Stream (PPCDS), it focuses on the privacy preserving process in an data stream environment and keep up a certain level of excellent mining accuracy. PPCDS is for the most part used to comibine Rotation-Based Perturbation, omtimizing the cluster center and the idea of closest neighbor to explain the security protecting grouping of mining issues in an information stream environment. In Rotation-Based Perturbation, rrotation transformation matrix is use to continuously perturb the data streams in order to preserve data privacy. In the period of cluster mining, perturbe data is fundamentally used to build up a smaller scale group through the enhancement of a cluster center and afterward updatecluster by applying measurement calculation.

IV. PROBLEM DESCRIPTION

The introductory idea was to investigate the traditional data mining procedures so it can work with the perturb stream data to veil sensitive data. The key issue is to get great precision for the consequence of stream mining utilizing the perturb information. The arrangements are frequently firmly coupled with the data stream mining calculations under thought.

The fundamental objective is to change a given information set D into perturtube dataset D' so it can fulfill a given privacy requiremnet with the loss of least data for the information investigation assignment. In this paper data preservation calculations have been proposed for information set perturbe.

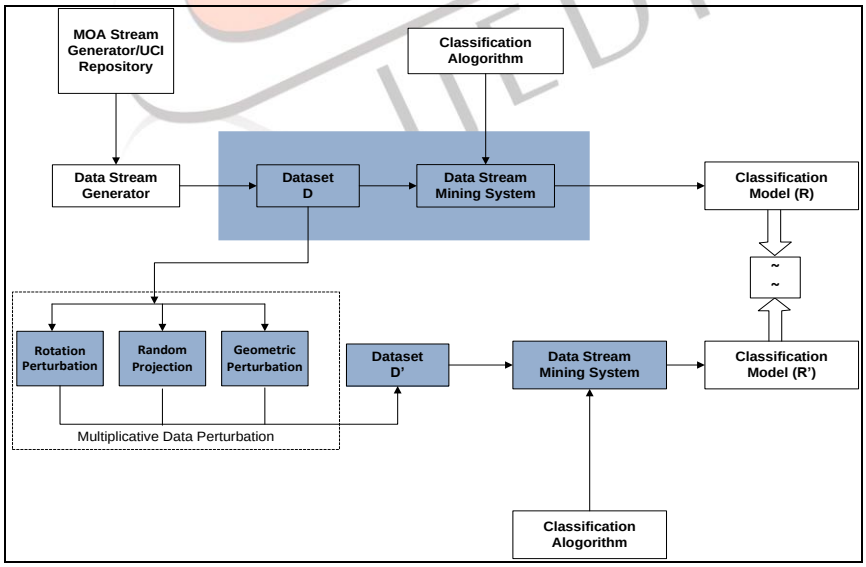

Fig 1. Framework for privacy preserving in data stream clustering

V. PROPOSED METHOD

Accepting that the data stream for handling has different multi-dimensional numeric information X 1...X K..., and every information contains timestamp T1… TK..., and its multi-dimensional information is spoken to by X i = (xi1...xid). At the point when andata stream is prepared, information is presented to in m x n information framework Dm×n, where every line speaks to

one section and every segment speaks to a quality of data.
Here in our work for changing the multidimensional information into lower measurements, Principal Component Analysis (PCA) is utilized. In cutting edge information examination, PCA is a standard device. PCA expect that every one of the variables in a procedure ought to be utilized for the investigation with the goal that it gets to be hard to recognize the essential variable from the variable that are less critic.

Algorithm: Geometric Transformation Based Multiplicative Data Perturbation.

Input: Data Stream **D,** Sensitive attribute **S**.
Intermediate Result: Perturbed data stream **D'**.
Output: Clustering results **R** and **R'** of Data stream **D** and **D'** respectively**.**

Steps:

1. Given input data **D** with tuple size **n**, extract sensitive attribute $\mathbf{[S]_{n\times 3}}$.
2. Rotate $[\boldsymbol{S}]_{n\times 1}$ into 180o clock-wise direction and generate $[\boldsymbol{R_S}]_{n\times 1}$.
3. Multiply elements of [S] with $[R_S]$, transformed sensitive attribute values will be $[\boldsymbol{X}]_{n\times 1} = [\boldsymbol{S}]_{n\times 1} \times [\boldsymbol{R_S}]_{n\times 1}$
4. Calculate translation T as mean of sensitive attribute $[\boldsymbol{S}]_{n\times 1}$.
5. Generate transformation $[St]_{n\times 1}$ by applying translation T to $[\boldsymbol{S}]_{n\times 1}$.
6. Calculate Gaussian distribution P(x) as a probability density function for Gaussian noise $\boldsymbol{P(x)} = \frac{1}{\sigma\sqrt{2\pi}} \boldsymbol{e}^{-\frac{(x-\mu)^2}{2\sigma^2}}$
   Where, μ=Mean, σ=Variance
7. Crate perturbed dataset D' by replacing sensitive attribute $[S]_{n\times 1}$ in original dataset D with $[\boldsymbol{Gs}]_{n\times 1}$.
8. Apply ***k-Mean*** clustering algorithm with different values of k on original dataset **D** having sensitive attribute **S**.
9. Apply ***k-Mean*** clustering algorithm with different values of k on perturbed dataset **D'** having perturbed sensitive attribute **P**.
10. Create cluster membership matrix of results from step 10 and step 11 and analyze.

VI. RESULT AND DISCUSSION

For evaluating the clustering accuracy, many experiment were carried out over sliding window size (w). Our evaluation mainly focus on the quality of cluster generated after the dataset perturbation.
Steps for the experiment to be performed.

- Firstly, set each and every dataset as stream in MOA framework
- Secondly, to evaluate measure and cluster membership matrix define the sliding win- dow (w) over the data stream.
- For protecting the sensitive attribute val-ue, apply the proposed data pertur-bation method to all the instances in sliding win-dow.
- Here, K-Means Clustering algorithm is used for finding clusters for evaluation. For 2 reason, K-Means was chosen i.e. 1) K-Means is one of the well known Clustering algorithm and even scalable.2) No. of cluster to be obtained from original and perturbed dataset have been taken same as no. of cluster.
- Lastly, compare the each cluster in perturbed dataset to its match in original dataset. The quality of the cluster generated is computed by F-Measure.

The entire examination were performed for measuring the accuracy along the insurance of sensitive data. Here two diverse result have been displayed, one is demonstrating the group accuracy in membership matrix term that was gotten from the clustering result and second one speak to the chart for F1_P[precision] and F1_R[Recall].
Table 6.1 beneath present datasets to focus on accuracy in view of membership Matrix. Each datasets have been designed to get the 3 and 5 cluster utilizing the K-Means Clustering algorithm. Underneath given Table 6.2 and Table 6.3 present the membership lattice that was acquired by clustering perturbate data of Bank Management dataset. Every grid speak to 3 and 5 groups for Original dataset and perturbate dataset. The original dataset clustering gives data of the no. of occurrence that are really classifeid in each of the cluster while the perturbeded dataset clustering gives data of the aftereffect of the right task after the data perturbation furthermore gives the rate of precision acquired

Dataset	Total instances	Instances processed	Attributes protected
Bank Management	45210	45k	Age,Balance,Duration

Table 6.1: Dataset design to focus accuracy in view of Membership Matrix

k-Mean clustering algorithm has been applied on unique dataset D and perturbate dataset D' produced utilizing proposed algorithm. Results in table 6.2 and 6.3 demonstrates that for every single tried cas very nearly 90% mining exactness has been accomplished. Calculation has been tried against diverse estimations of k and it has been watched that precision has been diminishing as k worth increments. This legitimizes that likelihood of tuple to fall into unique cluster will be diminishing as number of groups increments.

Table 6.2: Resultant accuracy of 3-Cluster

Dataset	Attributes	No. of Clusters	Stream Data	K-Means
Bank Management	Age	5	2000	89.51
	Balance			90.75
	Duration			88.10
	Age		3000	84.64
	Balance			89.05
	Duration			84.49

Table 6.3: Resultant accuracy of 5-Cluster

VII. CONCLUTION

PCA based multiplicative data perturbation approach has been proposed for random noise addition to preserve privacy of sensitive attributes. Proposed approach has tried to keep statistical relationship among the sensitive attributes intact to mine favorable results with perturbed data. It considers sensitive attribute as dependent attribute and remaining attributes of dataset except class attribute as independent attributes. Only dependent attribute of dataset has been used to calculate tuple specific random noise. *K-Mean* clustering algorithm on perturbed dataset has been used in order to estimate the accuracy and effectiveness of clustering results over four standard datasets. Results show fairly good level of privacy has been achieved with reasonable accuracy in almost all tested cases. Privacy of original data after applying perturbation has been quantified using correlation analysis. Data mining accuracy due to data perturbation has been quantified by percentage of instances of dataset that are been misclassified with clustering results with original dataset. We limited experiments to protect numeric attribute only but work can be extended to nominal type attributes also.

VIII. REFERENCE

[1] Bifet, G. Holmes, R. Kirkby and B. Pfahringer, *Data Stream Mining-A Practical approach*, 2011.
[2] L. Golab and M. T. Ozsu, Data Stream Management Issues -A Survey Technical Report,2003.
[3] V.S. Verykios, K. Bertino, I. N. Fovino, L.P. Provenza, Y.Saygin and Theodoridis, State-of-the-Art in Privacy Preserving Data Mining, *ACM SIGMOD Record*, Vol. 33, pp. 50-57, 2004.
[4] W. Du and Z. Zhan, Building Decision Tree Classifier on Private Data, *Proceedings of IEEE International Conference on Privacy Security and Data Mining*, pp. 1-8, 2002.
[5] R. Agrawal and R. Srikant, Privacy-Preserving Data Mining, *Proceedings of ACM SIGMOD International Conference on Management of Data*, pp. 439-450, 2000.
[6] S. R. M. Oliveira and O. R. Zaiane. Privacy Preserving Clustering By Data Transformation. In *Proc. of the 18th Brazilian Symposium on Databases*, pages 304–318, Manaus, Brazil, October 2003.
[7] V. Estivill-Castro and L. Brankovic. Data Swapping: Balancing Privacy Against Precision in Mining for Logic Rules.In *Proc. of Data Warehousing and Knowledge Discovery DaWaK-99*, pages 389–398, Florence, Italy, August 1999.
[8] Vaidya, J. and Clifton, C., "Privacy-Preserving KMeans Clustering over Vertically Partitioned Data,"*Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and DataMining*, Washington, D.C., U.S.A.,pp.206_215 (2003).
[9] Meregu, S. and Ghosh, J., "Privacy-Preserving Distributed Clustering Using Generative Models,"*Proceedings of the 3th IEEE International Conference on Data Mining*, Melbourne, Florida, U.S.A.,pp. 211_218 (2003).
[10] Ching-Ming Chao, Po-Zung Chen and Chu-Hao Sun, *Privacy-Preserving Clustering of Data Streams*, Tamkang Journal of Science and Engineering, Vol.13, No. 3, pp.349 - 358(2010).