Preserving Privacy Using Data Perturbation in Data Stream

[1]Neha Gupta, [2] IndrJeet Rajput
[1](PG-CE Student) Department of Computer Engineering, Gujarat Technological University, Gujarat,
[2](Asst.Prof) Department of Computer Engineering, Gujarat Technological University, Gujarat,

***Abstract*: - Data stream can be conceived as a continuous and changing sequence of data that continuously arrive at a system to store or process. Examples of data streams include computer network traffic, phone conversations, web searches and sensor data etc. The data owners or publishers may not be willing to exactly reveal the true values of their data due to various reasons, most notably privacy considerations. To preserve data privacy during data mining, the issue of privacy preserving data mining has been widely studied and many techniques have been proposed. However, existing techniques for privacy preserving data mining is designed for traditional static data sets and are not suitable for data streams. So the privacy preservation issue of data streams mining is need for the time. This paper focused on describing a method that extends the process of data perturbation on data sets to achieve privacy preservation. The technique mainly exploits a combination of isometric transformations i.e. translation and rotation transformations used with a secure random function in order to provide secrecy of user-specified attributes without losing accuracy in results.**

***Keywords:* Data Stream, Data Perturbation, Data Perturbation, Random Function**

I. INTRODUCTION

In the field of information processing, data mining refers to the process of extracting the useful knowledge from the large volume of data. Widely used data mining techniques in such area of application includes Clustering, Classification, Regression analysis and Association rule / Pattern mining.

The data stream paradigm has recently emerged in response to the issues and challenges related with continuous data [1]. Mining data streams is concerned with extracting knowledge structures represented in models and patterns in non-stopping, continuous streams (flow) of information. Algorithms written for data streams can naturally cope with data sizes many times greater than memory, and can be extended to challenge real-time applications not previously tackled by machine learning or data mining.

But nowadays, in the field of information processing, an emergence of applications that do not fit this data model [2] Instead, information naturally occurs in the form of a sequence (stream) of data values. A data stream is a real-time, continuous, and ordered sequence of items. It is not possible to control the order in which items arrive, nor feasible to locally store a stream in its entirety. Likewise, queries over streams run continuously over a period of time and incrementally return new results as new data arrive.

II. PRIVACY CONCERN FOR DATA STREAM

Mining data streams is concerned with extracting knowledge structures represented in models and patterns in non-stopping streams of information.

Motivated by the privacy concerns on data mining tools, a research area called privacy-preserving data mining has been emerged.

Verykios et al. [3] classified privacy- preserving data mining techniques based on five dimensions – data distribution, data modification, data mining algorithms, data or rule hiding, and privacy preservation. In the dimension of data distribution, some approaches have been proposed for centralized data and some for distributed data.

Du and Zhan [4] utilized the secure union, secure sum and secure scalar product to prevent the original data of each site from revealing during the mining process. The disadvantage is that the approach requires multiple scans of the database and hence is not suitable for data streams, which flows in fast and requires immediate response.

In the dimension of data modification, the confidential values of a database to be released to the public are modified to preserve data privacy. Adopted approaches include perturbation, blocking, aggregation or merging, swapping, and sampling. Agrawal and Srikant [5] used the random data perturbation technique to protect customer data and then constructed the decision tree. For data streams, because data are produced at different time, not only data distribution will change with time, but also the mining accuracy will decrease with perturb data.

From the review of previous research, it can be seen that existing techniques for privacy-preserving data mining are designed for static databases with an emphasis on data security. These existing techniques are not suitable for data streams.

Perturbation techniques are often evaluated with two basic metrics: level of privacy guarantee and level of model-specific data utility preserved, which is often measured by the loss of accuracy for data classification and data clustering. An ultimate goal for all data perturbation algorithms is to optimize the data transformation process by maximizing both data privacy and data utility achieved. Data privacy is commonly measured by the difficulty level in estimating the original data from the perturbed data. Given a data perturbation technique, the higher level of difficulty in which the original values can be estimated from the perturbed data, the higher level of data privacy this technique supports. Data utility typically refers to the amount of mining-task/model specific critical information preserved about the data set after perturbation.

III. PRIVACY PRESERVING DATA STREAM CLUSTERING

The data stream model of computation requires algorithms to make a single pass over the data, with bounded memory and limited processing time, whereas the stream may be highly dynamic and evolving over time. For effective clustering of stream data, several new methodologies have been developed, as follows: Compute and store summaries of past data: Due to limited memory space and fast response requirements, compute summaries of the previously seen data, store the relevant results, and use such summaries to compute important statistics when required.

The main idea of Perturbation- Based technique involves increasing a noise in the raw data in order to perturb the original data distribution and to preserve the content of hidden raw data. Geometric Data Transformation Methods (GDTMs) [6] is one simple and typical example of data perturbation technique, which perturbs numeric data with confidential attributes in cluster mining in order to preserve privacy.

Nonetheless Kumari et al. [7] proposed a privacy preserving clustering technique of Fuzzy Sets, transforming confidential attributes into fuzzy items in order to preserve privacy. Furthermore, the largest issue encountered when implementing a perturbation technique is the inaccurate mining result from a perturbed data.

Vaidya and Clifton [8] proposed the method of privacy preserving clustering technique over vertically partitioning data. In the vertical partitioning the attributes of the same objects are split across the partitions.

On the contrary, Meregu and Ghosh [9] proposed the method of privacy preserving cluster mining over horizontally data partitioning, whereas it is framework of "Privacy-preserving Distributed Clustering using Generative Model." In this approach, rather than sharing parts of the original data or perturbed data, the parameters of suitable generative models are built at each local site.

In [10] proposed a method of Privacy-Preserving Clustering of Data Stream (PPCDS), stressing the privacy-preserving process in a data stream environment while maintaining a certain degree of excellent mining accuracy. PPCDS is mainly used to combine Rotation-Based Perturbation, optimization of cluster enters and the concept of nearest neighbour, in order to solve the privacy-preserving clustering of mining issues in a data stream environment. In the phase of Rotation-Based Perturbation, rotation transformation matrix is employed to rapidly perturb with data streams in order to preserve data privacy. In the phase of cluster mining, perturbed data is primarily used to establish a micro-cluster through the optimization of a cluster center, then applying statistic calculation to update the micro-cluster.

IV. PROBLEM DESCRIPTION

The initial idea of it was to extend traditional data mining techniques to work with the perturbed stream data to mask sensitive information. The key issue is to get accurate stream mining results using perturb data. The solutions are often tightly coupled with the data stream mining algorithms under consideration.

The goal is to transform a given data set D into perturbed version D' that satisfies a given privacy requirement and loss minimum information for the intended data analysis task. In this paper data perturbation algorithms have been proposed for data set perturbation.

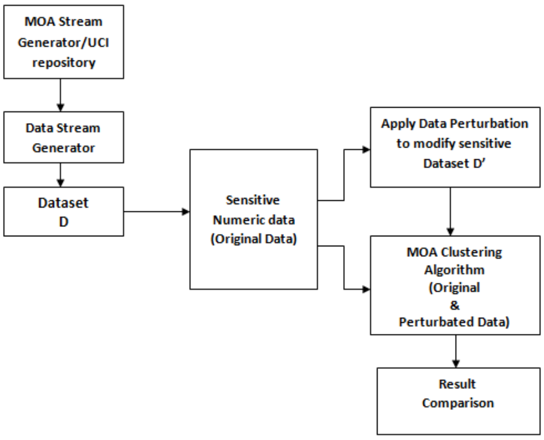
MOA Stream Generator/UCI repository
Data Stream Generator
Dataset D
Sensitive Numeric data (Original Data)
Apply Data Perturbation to modify sensitive Dataset D'
MOA Clustering Algorithm (Original & Perturbated Data)
Result Comparison

Fig 1. Framework for privacy preserving in data stream clustering

V. RELATED WORK BACKGROUND

A. Isometric Transformation

Transformations which leave the metric properties of the space unaltered are called isometric. Under these transformations the space is not stretched or twisted so that the distances between any pair of points remain unchanged upon transformation. Formally, an isometric transformation is defined as follows [11]:

Definition (Isometric Transformation). Let T be a transformation in the n-dimensional space, i.e., T : $\Re^n$->$\Re^n$ .T is said to be an isometric transformation if it preserves distances satisfying the following constraint: |T (p)−T (q)| =|p − q| for all p, q € ℜ.

Isometric transformations include:

(1) Translations, which shift points a constant distance in parallel directions

(2) Rotations, which have a center such that |T (p) − a| = |p −a| for all p

For the sake of simplicity, such a transformation is done in a 2D discrete space. It is shown in; any transformation of a space which leaves the metric properties unaltered can be reduced to translation, rotation to a certain combination of these transformations.

1) *Translation Based Perturbation*

In TBP method, the observations of confidential attributes are perturbed using an additive noise perturbation. Here we apply the noise term applied for each confidential attribute which is constant and value can be either positive or negative.

2) *Rotation Based Perturbation*

In this method a rotation matrix is used to rotate two attributes at a time. For the sake of simplicity a 2D rotation matrix is considered. The rotation of a point by an angle Θ in a 2D discrete space can be seen as a matrix representation V’= R(Θ)×V, where V is the column vector containing the original coordinates, and V’ is a column vector whose coordinates are rotated coordinates and R(Θ) is a 2×2 rotation matrix,

$$R(\Theta)=\begin{bmatrix} \cos\Theta & \sin\Theta \\ -\sin\Theta & \cos\Theta \end{bmatrix}.$$

B. Normalization

Objects (e.g. individuals, patterns, events) are usually represented as points (vectors) in a multi-dimensional space. Each dimension represents a distinct attribute describing an object. Thus, a set of objects is represented as an m × n matrix D, where there are m rows, one for each object, and n columns, one for each attribute. This matrix is referred to as a data matrix, represented as follows:

$$D=\begin{bmatrix} a_{11} & \cdots & a_{1k} & \cdots & a_{1n} \\ a_{21} & \cdots & a_{2k} & \cdots & a_{2n} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mk} & \cdots & a_{mn} \end{bmatrix}$$

The attributes in a data matrix are sometimes normalized before being used. The main reason is that different attributes may be measured on different scales .For this reason, it is common to standardize the data so that all attributes are on the same scale. There are many methods for data normalization.

We review only two of them in this section: min-max normalization and z-score normalization.

Min-max normalization performs a linear transformation on the original data. Each attribute is normalized by scaling its values so that they fall within a small specific range, such as 0.0 and 1.0. Min-max normalization maps a value V of an attribute A to V’ as follows:

$$V'=\frac{V-minA}{maxA-minA}\times(new_{maxA}-new_{minA})+new\_minA$$

where minA and maxA represent the minimum and maximum values of an attribute A, respectively, while new_minA and new_maxA are the new range in which the normalized data will fall.

When the actual minimum and maximum of an attribute are unknown, or when there are outliers that dominate the min-max normalization, z-score normalization (also called zero-mean normalization) should be used. In z-score normalization, the values for an attribute A are normalized based on the mean and the standard deviation of A. A value V is mapped to V’ as follows:

$$V'=\frac{V-\bar{A}}{\sigma A}$$

where A and σA are the mean and the standard deviation of the attribute A, respectively.

VI. PROPOSED METHOD

Assuming the data stream for processing includes multiple multi-dimensional numeric data X 1...X K ...,each data contains its proprietary timestamp T1…TK...,with multi-dimensional data represented by X i = (xi1...xid). When a data stream incoming, data is represented in an m x n data matrix Dm×n, while each row represents one entry and each column represents an attribute of data.

The proposed hybrid method distorts data points in the n dimensional space based on the following assumptions:

1) The mxn data matrix D, subjected to perturbation, contains only confidential numerical attributes.

2) We need the Attributes to get suppressed which are not subjected to perturbation and clustering.
3) Normalization helps prevent attributes with large range from outweighing attributes with smaller ranges. Here, we use z-score normalization

A. Data Perturbation Algorithm using Rotation

Here, from Original Dataset the data matrix *D*, *k* pairs of attributes are selected randomly. If number of attributes is odd, then last attribute is paired with an already selected attribute. If number of attributes is even, then during pairing one attribute is taken once only. Security administrator selects *k* pair-wise security threshold i.e. PST ($\rho 1$, $\rho 2$) for each attribute pair. The set of Θ which satisfy the constraints *Variance* $(Ai-A'i) > \rho 1$ and *Variance* $(Aj-A'j) > \rho 2$ is a interval which is called the security range. At $\Theta=0$ (i.e. at 2Π) both the variances are 0. To find the range we can compute $V'(A'i,A'j)=R(\Theta)\times V(Ai,Aj)$ for values of Θ increasing from 0 till the constraints are satisfied.

```
RBP ()
Input: An Original Dataset Vmxn (.ARFF or .CSV file)
Output: A perturbed Dataset V'mxn (.ARFF or .CSV file)
1) Read Original Data set V file.
2) Consider only numeric data type attribute from data set S.
3)  a. If n is even Select k=n/2otherwise k = (n + 1)/2
    b. Select k pairs of attributes randomly
    c. selects k pair-wise security threshold for each attribute pair
4) Consider k pairs of attributes selected in step 3 are distorted as follows:
    a. Compute $V'(A'i,A'j)=R(\Theta)\times V(Ai,Aj)$ for the different values of $\Theta$ to find the  security range
    b. From the security range select randomly a real value for $\Theta$.
   c. Compute $V'(A'i,A'j)=R(\Theta)\times V(Ai,Aj)$
   d. Store perturbed data set V' into new file.
```

B. Data Perturbation Algorithm using Translation

In this subsection, we report a security enhanced translation based perturbation algorithm. The major attraction of this algorithm is the use of a randomization function, F_R. F_R is initially used to generate a long list of random numbers i.e. say L_R, which is then normalized to generate, say L'_R. Next, depending on the number of selected attributes for perturbation, it selects random & normalized pairs from L_R, L'_R. Now, from the value of L'_R entry it is decided whether to add or subtract the corresponding L_R entry from the original data. Next, we present the TBP algorithm.

```
TBP ()
Input: An Original Dataset Tmxn (.ARFF or .CSV file)
Output: A perturbed Dataset T'mxn (.ARFF or .CSV file)
1). For each confidential attribute Aj ($1 \le j \le n$) in T do
a. Select the noise term $r_j$ and the corresponding r'j from $L_R$ and $L'_R$  respectively
 b. For each aij an instance of $A_j$ where $1 \le i \le m$ do
        If $r'_j > 0.5$ then
            aij ← aij + rj //Output the perturbed attribute value of T′
        else
            aij ← aij - rj //Output the perturbed attribute value of T′
 c. next i ;
 2). next j;
3) Store perturbed data set T' into new file
```

C. Data Perturbation Algorithm using Combine of Rotation & Translation

Instead of applying one method alone if we apply all the above mentioned two methods combined, then it will be more difficult for an intruder to get back the original data. To achieve this goal here Hybrid Data Perturbation Method is proposed. The Hybrid Data Perturbation Method, denoted by *RTDP ()*, combines the strength of the translation and rotation based transformation method.

```
RTDP ()
Input: An Original Dataset $Dm\times n$ (.ARFF or .CSV file)
Output: An perturbed Dataset $Dm\times n'$ (.ARFF or .CSV file)
1. Take user-specified p attributes for translation, q attributes for rotation such that $p+q=n$;
2. Call RBP() for q attributes;
3. Call TBP() for p attributes;
```

VII. EXPERIMENTAL EVALUATION

In this section, we empirically validate our proposed technique. The proposed technique is implemented in Java. We evaluate this technique from degree of privacy the experiments are run on a PC with 1.66GHz CPU, 1GB memory. Three real datasets are chosen which are obtained from UCI machine learning repository [12]. The brief information of chosen datasets is described in Table I.

Table I: Properties of Data Sets

	Ecoli	Diabetes	CMC
No. Of Records	336	768	1473
No. Of Attributes	7	9	9
No. Of Category	8	2	3

A. Degree of privacy

Traditionally, the privacy provided by a perturbation technique is measured by the variance between the actual and the perturbed values. We have also used this metric for measuring the degree of privacy that is provided with TBP, RBP and RTDP. This measure is given by *Var* (*X* - *X* ') where *X* represents a single original attribute and *X* ' the distorted attribute. This measure can be made scale invariant with respect to the variance of *X* by expressing security as *S* = *Var* (*X* - *X* ') /*Var*(*X*), the higher *S* shows the higher protection level. Table II shows the degree of privacy provided by these methods.

Table II: S Values for Transformed Datasets

	Ecoli	Diabetes	CMC
TDP	0.76	0.89	0.80
RBP	1.30	1.49	1.32
RTDP	1.45	1.63	1.40

B. Cracking Complexity

A brute force attack to crack RTDP method would require a great deal of computational power to get the original data.

Security of the RTDP Method based on the following factors:

-To which attribute which transformation is applied is unknown.

-For rotation the angle Θ for each pair is selected randomly in a continuous interval (the security range). And the Θ value is different for each pair of attribute. The lower the pair wise-security threshold selected by a security administrator results in broader the security range.

-For translation a random noise is generated which may be positive or negative.

From the factors mentioned above it is clear that the computational difficulty becomes progressively harder as the number of attributes in a database increases. Apart from that, it is not trivial for an attacker to guess the angle Θ for rotation for a particular pair since the security range is a continuous interval and the random noise for translation. More important point here is that attacker is unknown to which transformation is applied to an attribute.

VIII. EXPERIMENTAL SETUP AND RESULTS

We have conducted experiments to evaluate the performance of data perturbation algorithms. For experiment we use Massive Online Analysis (MOA) – an open source framework for data stream mining [13]. Applying the clustering algorithm CluStream on all dataset with parameter is decay horizon: 10, evaluation frequency: 200, decaythresold: 0.05.

A. Quality of clustering can be measured using CMM, SSQ and purity.

Cluster Mapping Measure: With the mapping from found clusters to ground truth classes, we can now determine the set *F* ⊆ *O*+ of points that cause *faults*, i.e. missed points, misplaced points, or included noise points.

SSQ: SSQ is the sum square of the distance between each point in the cluster and the center of the cluster. SSQ is used to measure concentration of a cluster and the lower the SSQ, the higher the concentration of the cluster.

SSQ calculation is:

$$SSQ=\frac{\sum_{j=1}^{K}\sum_{i=1}^{N}|xji-\bar{x}j|2}{K}$$

In equation xji is the ith data point of the jth cluster .*x j* is the center of the *jth* cluster. The average SSQ can be calculated by sum the SSQ of each cluster and divided by the number of clusters.

Purity: Purity is an indicator used to measure the accuracy of a cluster. We can compare the clustering result with the corrected label to calculate the purity of the cluster. The purity calculation formula is:

$$purity=\frac{\sum_{i=1}^{K}|\frac{c_i^d}{c_i}|}{K}*100\%$$

Where *k* is the number of clusters, *dci* denotes the number of points with the corrected label in cluster *i*. *Ci* denotes the number of points in cluster *i*.

B. Evaluating Quality of clustering on Datasets

Table III: Quality of clustering on Dataset Diabetes

Quality Measure	Original	RDP	TDP	RTDP
CMM	0.52	0.91	0.94	0.73
SSQ	1.65	0.75	0.52	1.08
Purity	0.33	0.93	1.06	0.63

Table IV: Quality of clustering on Dataset CMC

Quality Measure	Original	RDP	TDP	RTDP
CMM	0.97	1.10	0.88	1.03
SSQ	0.95	0.82	0.92	0.79
Purity	1.02	1.34	1.22	1.12

Table V: Quality of clustering on Dataset Ecoli

Quality Measure	Original	RDP	TDP	RTDP
CMM	0.37	0.55	0.60	0.45
SSQ	2.28	1.98	1.82	1.53
Purity	0.96	1.04	1.12	1.09

IX. CONCLUSION

In the step of data streams pre-processing, we proposed hybrid algorithms for data perturbation that are the data perturbation for privacy preserving in data stream clustering.

Perturbation techniques are often evaluated with two basic metrics: level of privacy guarantee and level of model-specific data utility preserved, which is often measured by the loss of accuracy for data clustering. The experimental results have shown that the proposed technique provides a proper degree of privacy. By using this technique, data owners can share their data with data miners to find accurate clusters without any concern about violating data privacy.

Using data perturbation algorithm, we generate different perturbed data set. And in the second step we apply the clustering algorithm on perturbed data set. We carried out set of experiments to generate clustering model of original data set and perturbed data set. Clustering results have been evaluated on accuracy parameters. Proposed algorithms can perturb sensitive attributes with numerical values.

REFERENCES

[1] A. 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 andDataMining*,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).

[11] H. T. Croft, K. J. Falconer, and R. K. Guy. Unsolved Problems in Geometry: v.2. New York: Springer Verlag, 1991

[12] Asuncion A, Newman D. UCI Machine Learning Repository [EB/OL].

[13] A. Bifet, R. Kirkby, P. Kranen, P. Reutemann, MOA: Massive Online Analysis Manual, *Journal of Machine Learning Research (JMLR)*, 2010.