Privacy Preserving Data Stream Mining using Two Phase Geometric Data Perturbation

Sanket P. Modi[1] Ashil R. Patel[2]
[1]M.E Scholar [2]Assistant Professor
[1,2]Department of Information Technology
[1,2]L. D College of Engineering, Ahmedabad

Abstract— Data mining is an information technology that extracts valuable knowledge from large amounts of data. Recently, data streams are emerging as a new type of data, which are different from traditional static data. The characteristics of data streams are: Data has timing preference; data distribution changes constantly with time; the amount of data is enormous. Data flows in and out with fast speed; and immediate response is required 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 are designed for traditional static databases and are not suitable for data streams. So the privacy preservation issue of data streams mining is a very important issue. This work is about proposing a Method and algorithms for the process of Geometric Data Perturbation or Geometric Data Transformation to achieve privacy preservation. Geometric Data Perturbation is a kind of data perturbation techniques. In this report, we describe the geometric transformations including translation, scaling, rotation, which can transform data in the protection of privacy while maintaining the similarity between data objects.

Key words: data mining, multiplicative data perturbation, privacy preserving data mining, geometric data perturbation

I. INTRODUCTION

Databases today can range in size into the hundreds of GB. Within these masses of data lies hidden information of strategic importance. But when there are so many trees, how do you draw meaningful conclusions about the forest? Data mining is a process that uses a variety of data analysis tools to discover patterns and relationships in data that may be used to make valid predictions.

A. Data Stream Mining:

Here we not talk about static data mining. Our goal is to mine the data on streaming mean the flow of the data are continuous and we have to mining the data. Traditional algorithm is designed for the static database. If the data changes, it would be necessary to rescan the database, which leads to long computation time and inability to promptly respond to the user

B. Data Mining Techniques:

1) Association Analysis:

Association analysis is the discovery of what are commonly called association rules. It studies the frequency of items occurring together in transactional databases, and based on a threshold called support, identifies the frequent item sets. Another threshold, confidence, which is the conditional probability than an item appears in a transaction when another item appears, is used to pinpoint association rules. Association analysis is commonly used for market basket analysis.

2) Classification:

Classification analysis is the organization of data in given classes. Also known as supervised classification, the classification uses given class labels to order the objects in the data collection. Classification approaches normally use a training set where all objects are already associated with known class labels. The classification algorithm learns from the training set and builds a model. The model is used to classify new objects. For example, after starting a credit policy, the Our Video Store managers could analyze the customers' behaviors vis-à-vis their credit, and label accordingly the customers who received credits with three possible labels "safe", "risky" and "very risky". The classification analysis would generate a model that could be used to either accept or reject credit requests in the future.

3) Clustering:

Similar to classification, clustering is the organization of data in classes. However, unlike classification, in clustering, class labels are unknown and it is up to the clustering algorithm to discover acceptable classes. Clustering is also called unsupervised classification, because the classification is not dictated by given class labels.

II. NEED FOR PRIVACY IN DATA MINING

Generally when people talk of privacy, they say "keep information about me from being available to others". However, their real concern is that their information not be misused. The fear is that once information is released, it will be impossible to prevent misuse. Utilizing this distinction ensuring that a data mining project won't enable misuse of personal information opens opportunities that "complete privacy" would prevent. To do this, we need technical and social solutions that ensure data will not be released. Another view is corporate privacy – the release of information about a collection of data rather than an individual data item. I may not be concerned about someone knowing my birth date, mother's maiden name, or social security number; but knowing all of them enables identity theft. This collected information problem scales to large, multi-individual collections as well. A technique that guarantees no individual data is revealed may still release information describing the collection as a whole. Such "corporate information" is generally

A. Privacy versus Security:

Privacy and Security are most confusing terms for people. Most of the people use this two terms alternative. But it is not like that. This two are so much differ from each other. Security is always in the top level where privacy is below level of security. Here I show simple difference between them:

Privacy	Security
It involves discriminating access between individual users to a system or network	It involves limiting or controlling access to a specific system or network.
For Example: The Facebook privacy system prevents people you haven't acknowledged as friends from viewing your posts unless you intend them to.	For Example: The Facebook security system prevents people from posting or interacting with comments unless they're logged in.

Table 1: Privacy and Security

III. Privacy in Multiplicative Data Perturbations

The randomization approach is particularly well suited to privacy-preserving data mining of streams, since the noise added to a given record is independent of the rest of the data. However, streams provide a particularly vulnerable target for adversarial attacks with the use of principal component analysis (PCA) based techniques because of the large volume of the data available for analysis. In [4], an interesting technique for randomization has been proposed which uses the auto-correlations in different time series while deciding the noise to be added to any particular value. It has been shown in [5] that such an approach is more robust since the noise correlates with the stream behavior, and it is more difficult to create effective adversarial attacks with the use of correlation analysis techniques.

IV. Survey on Geometric Data Perturbation

The work proposes in [5], about privacy concern for data stream mining that we discuss in 2.4.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 work proposes in [2][6][7], about different techniques on privacy preserving data mining as below:

A. Reconstruction Based Approach:

Reconstruction based approaches generate privacy aware database by extracting sensitive characteristics from the original database. These approaches generate lesser side effects in database than heuristic approach. Reconstruction based techniques perturb the original data to achieve privacy preserving.

B. Types of Geometric Data Perturbation:

The work proposes in [4][5][8][9][10], about different methods of geometric data perturbation. Geometric transformation method is a kind of data perturbation techniques. In this paper, we describe the geometric transformations including translation, scaling, rotation and reflection The data is considered as a matrix Dmn, where each of the m rows is an observation Oi ($1 \leq i \leq m$), and each observation contains values for each of the n attributes Ai ($1 \leq i \leq n$).

Primary there are three types of geometric perturbation methods:

- Translation Data Perturbation
- Scaling Data Perturbation
- Rotation Data Perturbation

C. Survey on Issue of Geometric Data Perturbation:

The issue proposes in [11], there exist attacks that can utilize the published information of perturbation and the perturbed data to approximately reconstruct the original dataset. Author primary suggest possible attacks which can be happen on geometric data perturbation:

D. ICA based Attack:

Naive estimation is the basic attack trying to find the original value directly from the perturbed data, which will be ineffective to carefully perturbed data. In this section, we introduce a high-level attack based on data reconstruction. The basic method trying to reconstruct X from the perturbed data RX would be Independent Component Analysis (ICA) technique derived from signal processing. The ICA model can be applied to estimate the independent components (the row vectors) of the original dataset X, from the perturbed data, if the following conditions are satisfied:

- The source row vectors are independent.
- All source row vectors should be non-Gaussian with possible exception of one row.
- The number of observed row vectors must be at least as large as the independent source row vectors.
- The transformation matrix R must be of full column rank.

E. Distance Inference Attack:

The attacker manages to get more knowledge about the original dataset: s/he also knows at least d +1 original data records, {x1, x2,....., xd+1}. S/he then tries to find the mapping between these points and their images in the perturbed dataset, denoted by {o1, o2,....., od+1}, to break the rotation and translation perturbation.

With the known points, it is possible to find their images in the perturbed data. Particularly, if a few (>=¸ d + 1) original points, such as the \outliers", are known, their images in the perturbed data can be found with high probability for low-dimensional small datasets (< 4 dimensions).

F. Attacks to Rotation Center:

The basic rotation perturbation uses the origin as the rotation center. Therefore, the points around the origin will be still close to the origin after the perturbation, which leads to weaker privacy protection over these points. The attack to rotation center is another kind of naive estimation. We address this problem with random translation perturbation.

G. Main Goal:

Our main goal is to preserve privacy in Geometric Data Perturbation with Less data losses More response time to construct a classification model. More privacy gain. Maintain the accuracy of the classification model. The Definition of this research is entitled by “Privacy Preserving Data Stream Mining Using Two Phase Geometric Data

Perturbation". so now we clear understanding about why it is called two phases. We divide work on two phases as below: The individual work of translation, scaling and rotation perturbation called as Phase-I. Mixing all method together means to make hybrid perturbation called as Phase-II. The propose algorithm may be looks like as below: Here the operation which we perform translation, scaling and rotation to make it. hybrid it is not necessary it is in order. It can be any kind of the order possible. We can say that there are three methods so there are 3! Mean six possibility made. Among of them which of the combination give us best privacy, that is we take it granted as final in Hybrid algorithm.

H. HDP Algorithm [8]:

Input: V, N
Output: V'

1) Step 1:

For each confidential attribute Aj in V , where $1 \leq j \leq d$ do

- Select the noise term ej in N for the confidential attribute Aj
- The j-th operation opj ← {Add,Mul,Rotation}

2) Step 2:

For each vi ϵ V do
For each aj in vi = (a1,…., ad), where aj is the observation of the j-th attribute do
1. a'j ← Transform(aj, opj, ej)
End

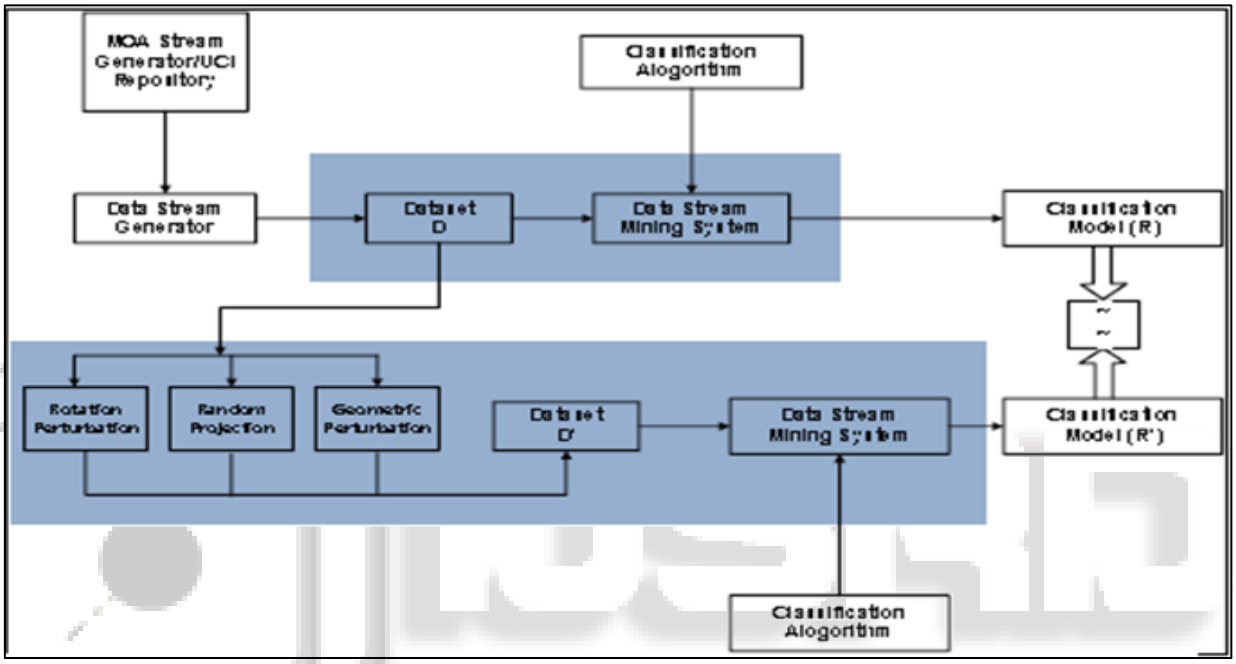

Fig. 1: HDP Algorithm

V. EVALUATION PARAMETERS

A. Kappa Statistics:

Introduced by Cohen. Kappa is a chance-corrected measure of agreement between the classification and the true classes. It's calculated by taking agreement expected by chance away from the observed agreement and dividing by maximum possible agreement. Kappa statistic is more sensitive measure for quantifying the predictive performance of streaming classifiers.[14] A common problem is that for unbalanced data streams with, for ex.90% of the instances in one class, the simplest classifiers will have high accuracies of at least 90%. To deal with this type of data stream, use the Kappa statistic, based on a sliding window, as measure for classifier performance in unbalanced class streams. Kappa needs to be estimated using some sampling procedure. In data streams, there are two basic evaluation procedures: one is holdout evaluation, and other is prequential evaluation (interleaved Test-Then-Train) Consider classifier h, A data set containing m examples and l classes, and contingency table where cell cij contain the number of example for which h(x) =I and the class is j. If h(x) correctly predicts all the data, then all non-zero counts will appear along the diagonal. If h misclassified some examples, then some off diagonal elements will be non-zero.

In problems where one class is much more common than the others, any classier can easily yield a correct prediction by chance, and it will hence obtain a high value for p0. To correct for this, the statistic is defined as follows:

B. Security Measurement:

We use the average squared distance (ASD) and the distance-based record linkage (DBRL) between the original data and the perturbed data to measure the security data perturbation algorithm. ASD uses the space distance formula to measure the difference between the original data and the perturbed data. In addition to calculating the distance between two collections of data, DBRL also takes the standard deviation into account. Therefore, it can measure the variance level between the original data and the perturbed data.[15]

C. Data error Measurement:

The data error of the mining results between the perturbed data and the original data. So uses the bias in mean (BIM) and the bias in standard deviation (BISD) between the original data and the perturbed data to measure the data error of the algorithm. [15]

D. Misclassification Error:

If the intended data usage is data classification; the information loss can be measured by the percentage of legitimate data set that is not well-classified after the sanitization process. As in [22], a misclassification error ME is defined to measure the information loss.

Me = |

Where, N =number of data in the original dataset, k =number of class under analysis, |Class (D)| &|Class (D′)| = number of legitimate data points of the ith class in the original dataset D and the sanitized dataset D′ respectively.

E. Confusion Matrix:

A confusion matrix (Kohavi and Provost, 1998) contains information about actual and predicted classifications done by a classification system. Performance of such systems is commonly evaluated using the data in the matrix. The following table shows the confusion matrix for a two class classifier.

The entries in the confusion matrix have the following meaning in the context of our study:

F. Prediction:

		Negative	**Positive**
Actual	Negative	A	B
	Positive	C	D

Table 2: Prediction

VI. CONCLUSION

We have reviewed the geometric data perturbation method as an alternative method to privacy preserving data mining. The design of this category of perturbation algorithms is based on an important principle: by developing perturbation algorithms that can always preserve the mining task and model specific data utility, one can focus on finding a perturbation that can provide higher level of privacy guarantee. We described four representative geometric data perturbation methods translation transformation, Scaling transformation, rotation transformation, and hybrid transformation. All aim at preserving the distance relationship in the original data, thus achieving good data utility for a set of classification and clustering models.

REFERENCES

[1] Jiawei Han, Jian Pei and Micheline Kamber, "Data Mining: Concepts and Techniques", Third Edition, The Morgan Kaufmann Series in Data Management Systems Elsevier, 2012.
[2] Hitesh Chhinkaniwala and Sanjay Garg, "Privacy Preserving Data Mining Techniques: Challenges & Issues", International Conference on Computer Science and information Technology CSIT, 2011.
[3] Kun Liu, Hillol Kargupta, and Jessica Ryan, "Random Projection-Based MultiplicativeData Perturbation for Privacy Preserving Distributed Data Mining", IEEE Transactions On Knowledge And Data Engineering-VOL. 18-NO. 1, 2006.
[4] Keke Chen and Ling Liu," Privacy-Preserving Multiparty Collaborative Mining with Geometric Data Perturbation", IEEE Transactions On Parallel and Distributed Computing, VOL.20, NO.12", 2009.
[5] Keke Chen and Ling Liu, "Geometric Data Perturbation for Privacy Preserving Outsourced Data Mining", IEEE Transactions Knowledge and Data Engineering,2012.
[6] Ompriya Kale and Prachi Patel, "A survey on Privacy Preserving Data Mining", Global Journal of Advanced Engineering Technologies Vol2 -Issue3, 2013.
[7] Neha Gupta and Indrjeet Rajput, "Preserving Privacy Using Data Perturbation in Data Stream", International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), 2013.
[8] Stanley R. M. Oliveira and Osmar R. Zaiane, "Privacy Preserving Clustering by Data Transformation", Journal of Information and Data Management Vol. 1 – No. 1, 2010.
[9] Stanley R. M. Oliveira and Osmar R. Zaiane, "Data Perturbation by Rotation for Privacy Preserving Clustering", Technical Report at University of Alberta, 2004.

