Privacy Preserving in Data stream classification using different proposed Perturbation Methods

Kiran Patel, Hitesh Patel, Parin Patel

Assistant Professor
Department of Information Technology
Gandhinagar Institute of Technology, Gandhinagar
[1]Kiraninfo10@gmail.com

***Abstract* – Data mining is 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. Traditional algorithm is designed for the static database. If the data changes, it would be necessary to rescan the database, which leads to more computation time and inability to promptly respond to the user. The issue of privacy- preserving data mining has widely been 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 very important issue. We proposed methods and algorithms which extends the existing process of data streams classification to achieve privacy preservation.**

***Keywords* – MOA, Data Stream, Weka, Data perturbation**

I. INTRODUCTION

A data stream is a sequence of unbounded, real time data items with a very high data rate that can only read once by an application [2]. Imagine a satellite-mounted remote sensor that is constantly generating data. The data are massive (e.g., terabytes in volume), temporally ordered, fast changing, and potentially infinite. These features cause challenging problems in data streams field. Data Stream mining refers to informational structure extraction as models and patterns from continuous data streams. Data Streams have different challenges in many aspects, such as computational, storage, querying and mining.

Examples of data streams include computer network traffic, phone conversations, web searches and sensor data. These data sets need to be analyzed for identifying trends and patterns which help us in isolating anomalies and predicting future behavior. However, 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. Hence, some amount of privacy preservation needs to be done on the data before it can be made publicly available.

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 which extends the process of data streams classification to achieve privacy preservation.

II. PROBLEM STATEMENT AND PROPOSED FRAMEWORK FOR SOLUTION

Motivated by the privacy concerns on data mining tools, a research area called privacy-preserving data mining. The initial idea of it was to extend traditional data mining techniques to work with the stream data modified to mask sensitive information. The key issues were how to modify the data and how to recover the data stream mining result from the modified data. The solutions were often tightly coupled with the data stream mining algorithms under consideration.

The goal is to transform a given data set D into modified version D' that satisfies a given privacy requirement and preserves as much information as possible for the intended data analysis task.

We can compare the classification characteristics in terms of less information loss, response time, and more privacy gain so get better accuracy of different data stream algorithms against each other and with respect to the following benchmarks:

1. Original, the result of inducing the classifier on unperturbed training data without randomization.
2. Randomized, the result of inducing the classier on perturbed data (Perturbation based methods for privacy preserving perturb individual data values or the results of queries by swapping, condensation, or adding noise.) but without making any corrections for randomization.

Show the graphically represent of above defined work in figure.4.1.

Methodology

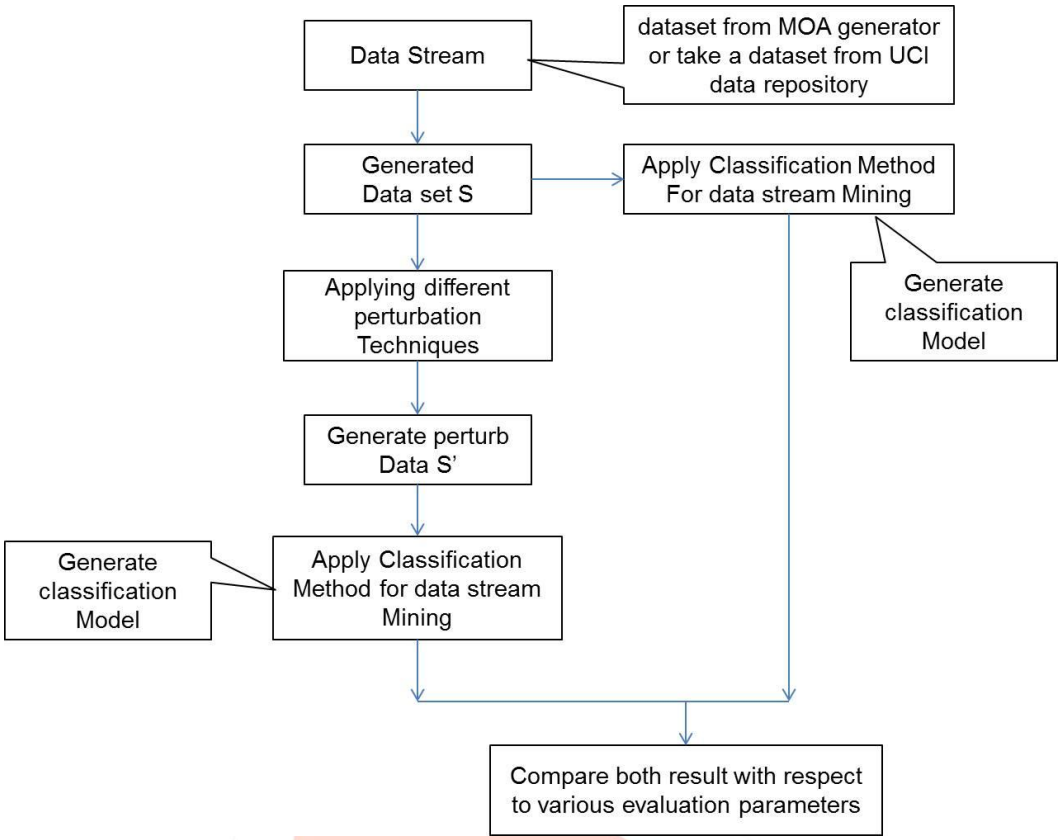

Figure 1 Framework for privacy preserving in data stream classification

III. PROPOSED DATA PERTURBATION TECHNIQUES

3.1 Proposed following method for data perturbation

1. **This option is only for numeric values**
   Add the any value in attribute's values (input from the user) suppose any attribute have value 12, 14, 11, 15, 9 etc. user give input 5, so add 5 to each value and output will be 17, 19, 16, 20, 14 etc.
2. **This option is only for non-numeric values**
   Change the non-numeric value of selected attribute by any other non-numeric value. (Suppose values is **car1** so replace by selected value suppose **p1** or other) (Used ASCII in programming)
   1. Select non numeric attribute.
   2. Find distinct values of selected attribute.
   3. Generate distinct values mapping to each value of distinct values of selected attribute.
   4. Replace old distinct values with generated/new distinct values.
3. **This option is only for numeric and non-numeric values**
   Interchange the values of the same attribute (by randomly choose value only from that attribute)
   1. Select Attribute from the data file.
   2. Loop through all instances, **I=0 To numberOfInstance**
      a. Randomly select instance/row and get Value of selected attribute of that instance/row.
      b. Set randomly selected value to selected attribute of instance '**I**'.
4. **This option is only for numeric values**
   Find mean of numeric value of any particular row's numeric attribute and replace chosen attribute value by this answer.
   1. Select non numeric attribute.
   2. Loop through all instances.
      a. Find mean of numeric attribute of all each instance/row.
      b. Set mean to selected attribute.

3.2 Brief Description of above techniques

Option 2 is only for non-numeric values

Table 1 for Non NumericValues

Name	Age	Gender	Salary	Car	Education Level
James	25	M	25000	Car1	4563.45
Bob	22	M	34000	Car3	2314.34
Alice	24	F	23400	Car1	3498.56
Prince	28	M	34500	Car2	4467.00
…..	**…..**	F	**…..**	Car3	**…..**

…..	…..	M	…..	Car2	…..
…..	…..	F	…..	Car1	…..

See above table in this table suppose I select non-numeric attribute **Gender.** So it contain only two distinct value in whole column that is **M** and **F.** so create to random value for this two value. Suppose For **M is P** and for **F is Q.** then replace value M by P and F by Q.

Same like for **car** attribute. It contains 3 distinct values that are car1, car2, and car3. So generate three random values. For car1 is suppose p1, for car2 is p2 and for car3 is p3, and replace according to that.

Option 3 is only for numeric and non-numeric values

See following table. Suppose selected attribute is Gender so randomly select any value from Gender attribute and replace first values by that. Again randomly select new value and replace second values by that selected value. Continue for all value of selected row.

Table 2 for numeric and non-numeric values

Name	Age	Gender	Salary	Education Level
James	25	M	25000	4563.45
Bob	22	M	34000	2314.34
Alice	24	F	23400	3498.56
Prince	28	M	34500	4467.00

Option 4is only for numeric values

See following table in that table there are 3 numeric attribute (Age, Salary, and Education Level) and 2 non-numeric attribute (Name and Gender)

Table 3 for numeric values

Name	Age	Gender	Salary	Education Level
James	25	M	25000	4563.45
Bob	22	M	34000	2314.34
Alice	24	F	23400	3498.56
Prince	28	M	34500	4467.00

Suppose selected attribute is salary.
Numeric attribute is 3
First for first row:

- So add all this ex. 25 + 25000 + 4563.45 = 29588.45
- Mean= 29588.45/3= 9862.81
- So replace salary attribute values 25000 by mean value that is 9862.81

Then second row:

- So add all this ex. 22 + 34000 + 2314.34 = 36334.34
- Mean= 36334.34/3= 12112.11
- So replace salary attribute values 34000 by mean value that is 12112.11

After complete the calculation of all row output dataset will be like following:

Table 4 complete data

Name	Age	Gender	Salary	Education Level
James	25	M	9862.81	4563.45
Bob	22	M	12112.11	2314.34
Alice	24	F	2342.33	3498.56
Prince	28	M	1231.22	4467.00

IV. EXPERIMENT SETUP AND DATA SET

We have setup experiments to evaluate the performance of data perturbation method. We choose generated Database. Generate a dataset from Massive Online Analysis (MOA) Framework [5]. And use the Agraval dataset generator. We use Waikato Environment for Knowledge Analysis (WEKA) [6] tool that is integrated with MOA to test the accuracy of Hoeffding tree algorithm. The both data perturbation algorithm implemented by a separate Java programme.

Following are the basic step for how to perform whole experiment.

Step 1. Generate a dataset or take a dataset. In this step we are generate the dataset from MOA generator or take a dataset from UCI data repository.

Step 2. Apply the algorithm on dataset and generate perturbed dataset. In this step we apply the both algorithm on dataset. (Data perturbation using window concept and multiplicative data perturbation using rotation perturbation)

Step 3. Take one classification algorithm (Hoeffding tree) and apply on perturbed dataset. Use WEKA (MOA integrated) tool

Step 4. Generate a Classification model.

Dataset

For experiment we take three dataset.

Agrawal dataset - Agraval dataset that is generated by using MOA Framework that contain 200000 instances and 10 attributes. Generator.AgrawalGenerator [5] Generates one of ten different pre-defined loan functions. It was introduced by Agrawal et al. in [7]. It was a common source of data for early work on scaling up decision tree learners. The generator produces a stream containing nine attributes, six numeric and three categorical. Although not explicitly stated by the authors, a sensible conclusion is that these attributes describe hypothetical loan applications. There are ten functions defined for generating binary class labels from the attributes. Presumably these determine whether the loan should be approved.

Adult Dataset - Extraction was done by Barry Becker from the 1994 Census database. A set of reasonably clean records was extracted using the following conditions: ((AAGE>16) && (AGI>100) && (AFNLWGT>1) && (HRSWK>0)). Prediction task is to determine whether a person makes over 50K a year. Adult dataset contain 48842 instances and 14 attributes. [9]

Bank Marketing Dataset - Bank marketing dataset [8] taken from UCI dataset repository is related with direct marketing campaigns of a Portuguese banking institution, and it contain 45211 instances and 17 attributes.

V. CONCLUSION

There is scope for further improvement in proposed methods and algorithms for preserving privacy in data stream classification. Output of such methods can be compared based on widely accepted evaluation metrics. Further we can also propose evaluation metrics to measure information gain / loss and privacy gain. Proposed work in data stream classification and privacy in data stream classification have their merits and demerits on key issues.

REFERENCES

[1] Chaudhry, N.A., Show, K., and Abdelgurefi, M. *STREAM DATA MANAGEMENT, Advances in Database system*. Vol. 30. Springer(2005)

[2] A. Bifet, G. Holmes, R. Kirkby, and B. Pfahringer. *Data stream mining: A Practical Approach* (2011).

[3] Albert Bifet, Richard Kirkby, Philipp Kranen, Peter Reutemann,*"Massive Online Analysis,Manual"* May 2011.

[4] Ching-Ming Chao, Po-Zung Chen and Chu-Hao Sun, *Privacy-Preserving Classification of Data Streams*, Tamkang Journal of Science and Engineering, Vol. 12, No. 3, pp. 321-330(2009).

[5] Albert Bifet, Richard Kirkby,Philipp Kranen, Peter Reutemann, "massive online analysis", manual, 2011.

[6] "The Weka Machine Learning Workbench", http://www.cs.waikato.ac.nz/ml/weka.

[7] R. Agrawal, T. Imielinski, and A. Swami. "Database mining: A performance perspective", IEEE Trans. on Knowl. And Data Eng., 5(6):914– 925, 1993.

[8] S. Moro, R. Laureano and P. Cortez, "Using Data Mining for Bank Direct Marketing: An Application of the CRISP-DM Methodology", In P. Novais et al. (Eds.), Proceedings of the European Simulation and Modelling Conference - ESM'2011, pp. 117-121, Guimarães, Portugal, 2011. EUROSIS (http://hdl.handle.net/1822/14838)

[9] Ronny Kohavi and Barry Becker, "Data Mining and Visualization", Silicon Graphics.

[10] A. Bifet, G. Holmes, R. Kirkby, and B. Pfahringer, "Data stream mining: A Practical Approach" , 2011.