International Journal of Technical Research and Applications e-ISSN: 2320-8163,
www.ijtra.com Volume 3, Issue 2 (Mar-Apr 2015), PP. 257-259

A SURVEY ON GEOMETRIC DATA TRANSFORMATION FOR PRIVACY PRESERVING ON DATA STREAM

Krupali N. Vachhani, Prof. Dinesh Vaghela
Parul Institute of Technology
Computer science and engineering
krupalivachhani@gmail.com, dineshcsepit@gmail.com

***ABSTRACT*— Recently data stream mining is new emerging field of data. It is different from traditional static data. In stream data, data is changing dynamically. They have characteristics like timing preference, data distribution changes constantly with time, data flows in and out with speed, amount of data is enormous and immediate response is required. So, to preserve the privacy during data stream mining many privacy preserving techniques have been proposed. Many existing techniques for privacy preserving are suitable for traditional static database but not suitable for dynamic data. So in data stream privacy preservation is an important issue. To achieve the privacy and balancing the accuracy we will use geometric data transformation (perturbation) technique. This paper present survey about geometric data perturbation technique for privacy preserving data mining.**

Keywords- Data Stream Mining, Geometric Data Transformation, Privacy Preserving.

I. INTRODUCTION

Data mining is nothing but extracting meaningful knowledge from the large amount of data. We can classify data mining techniques as follows: classification, association rule mining, clustering, sequential pattern analysis, data visualization, prediction. In recent years, simple transactions like using credit card, browsing the web, phone database, sensor network lead to wide and automated data storage. All these have large flows of data continuously and dynamically. This type of large volume data leads to many mining and computational challenges.

A. Data Stream Mining

Data stream is new type of data that is different than traditional static database. Data stream is continuous and dynamic flow of data. It is sequence of real time data with high data rate and application can read once. The characteristics of data streams are different than traditional static database which are as follows[1]: (1)Data has timing preference (2) Data Distribution changes constantly with time (3) The amount o data is enormous (4)Data flows in and out with fast speed (5)Immediate response is required. Because of these, data stream mining is challenging.

We have many data mining algorithm for traditional database where data is static and continuous flow. Use of traditional data mining algorithm is not appropriate in data stream mining because of no control over dataflow. If data will change, then we have to rescan the database. This will take more computational time. In data stream mining data is not persistent but rapid and time varying.

In mining of data stream, solution are categorize in data based and task based. In data based solution, data transform horizontally or vertically. In task based solution, different techniques have been adopted to achieve time and space efficient solution. Figure 1.1 shows simple data stream mining process. Once element of data stream is processed, it is discarded. So, it is not easy to retrieve it unless if we explicitly store them in memory.

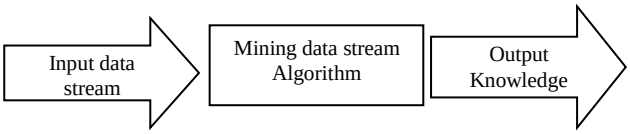

Fig. 1. Data Stream Mining Process

II. PRIVACY PRESERVING DATA MINING (PPDM)

The term privacy means keep information about me from being available to others. PPDM is producing valid model and pattern without disclosing sensitive information. The main aim to keep the information private is to prevent the misuse of private information. Once important data is disclosed then it is impossible to prevent the misuse of data. If data owner published their data, they have fear of misuse. So this prevents them to share their data. On other side, sharing of these data will useful in industries and business organization. They collect and analyze the data to know market policy. Sharing of data will help in improving the business strategy and to know customer behavior. If owner share the data, privacy must not breached.

Different people have different perspective of privacy, for some people personal information is privacy while for some people only some of the sensitive attribute is privacy. There are various techniques for privacy preserving for entire dataset modification or modification of some sensitive attributes. PPDM techniques are classified into four types: data partitioning, data modification, data restriction, data ownership. Because of personal data, privacy preserving becomes important in recent years. Now a days advanced technology provide the capacity to store large number of personal data. Suppose cancer research institutes in different areas need to collaboratively find the environmental factor related to certain type of cancer [2]. These distributed databases contain sensitive information.

The key direction in the field of privacy is: (1) Data perturbation: During pattern mining hiding private sensitive information. (2) Secure multi party computation: Building a model over multi party distributed databases without knowing others inputs. (3) Knowledge hiding: Hide sensitive rules or patterns. (4) Privacy aware knowledge sharing: do the data mining results violate their privacy? If we want to release database to public then confidential value of database are modified to preserve the privacy. There is research direction like privacy protection principles, general privacy preservation technology, and data mining privacy preservation technology. In general privacy preservation technology perturbation, randomization, swapping, encryptions are used. In data mining privacy preservation technology association rule mining, classification, clustering is used.

III. DATA PERTURBATION

Data perturbation is privacy preserving data mining technique. Data perturbation balances the privacy and

maintains data quality. Data perturbation has two categories (1) probability distribution (2) value distortion [6]. In probability distribution category, database is taken from given population and replaces original data with another sample from the same distribution. Fixed data perturbation method, database attributes that will use for statistic are perturbed once. Two methods for probability distribution. That is data swapping and probability distribution method. In swapping process, original database is replaced with random database which probability distribution is approximately same as the original database. But sometimes results of this method are unacceptable because it has up to 50% errors. In value distortion approach, sensitive information perturbed using additive noise or using some random processes. In privacy preserving transformation causes information loss that should be reduce during extracting meaningful knowledge from data. There are three types of data perturbation approaches: rotation perturbation, projection perturbation and geometric data perturbation.

IV. DATA PERTURBATION APPROACHES

A. Rotation Perturbation

Rotation perturbation is used for classification and clustering for privacy preservation. Rotation perturbation is used in geometric data transformation as an initial step. Rotation perturbation is shown as G(x) =RX, where R is rotational matrix that is randomly generate and X is original dataset that is to be perturbed. Distance preservation is unique advantage of rotation perturbation. Because of distance preservation many data mining model is invariant so major disadvantage of rotation perturbation is also distance preservation.

B. Projection Perturbation

In projection perturbation we project the set of original data from original space to random space. Suppose $p_{k\times d}$ is random projection matrix where P's rows are orthonormal. $G(X)=\frac{\sqrt{d}}{k}$ PX is applied to original dataset X for the perturbation [3].

C. Geometric Data Perturbation

Geometric data perturbation is very popular data perturbation technique. It is used in privacy preserving in collaborative data mining. As compared to other data perturbation techniques, geometric data perturbation has many advantages over privacy preservation so it is most widely used. Many popular data mining models are invariant in geometric perturbation. K- nearest neighbour classifier, linear classifier, support vector machine classifier are invariant means classifier with geometric data perturbation has almost same accuracy as the original data [2]. So this technique is valid for almost all clustering algorithms. Another advantage is geometric data perturbation can easily generated with low cost and main thing is that preserve same accuracy. As compared to other approaches, geometric data perturbation reduce the complexity in balancing data utility and data privacy guarantee [5].

Geometric data perturbation can apply to multi party collaborative data mining. Three main factor that affect the quality of perturbation is privacy guarantee, data utility and efficiency of perturbation unification protocol. The level of data utility also maintain in geometric data perturbation. The level of data utility means how much amount of critical information is preserved after data perturbation. While preserving the privacy guarantee some technique lose the data utility of sensitive information. This means that level of data utility must be high. One of the best multi party collaborative data perturbation.

Geometric perturbation method perturbed their own dataset before releasing to other of public use. We can describe a digital image b(x,y) in 2D space using analog image b(x,y) in continuous space. Whole 2D image divided into n rows and m columns. is applied to original dataset X for the perturbation [3].

Geometric data perturbation is combination of rotation perturbation and translation perturbation and noise addition.

$$G(X)=RX+T+\Delta$$

Where R is rotation perturbation, X is original dataset, T is translation perturbation and Δ noise addition. Addition of noise is protecting privacy and maintains high data quality. Addition of noise protects data from being disclosed. Noise can be added to both data value: categorical data and numerical data. First we add noise to sensitive class attributed known as label then we add noise to non class attributes. Here we are adding random noise. Use of random noise has some disadvantage. Because of random noise we may lose the data or value of sensitive attributes then we have to calculate that value as null or average.

Translation transformation: Translation is task of moving any point with co-ordinates(x,y) to the new location (x1,y1). Translation T is easily shown in the form of matrix. In figure 3.1 translation matrix Tv is shown where T is translation matrix and v is original co-ordinates and v' is new transformed co-ordinates.

$$\begin{bmatrix} 1 & 0 & x1 \\ 0 & 1 & y1 \end{bmatrix}$$

Fig. 2. Translation matrix

Rotation transformation: Pair of attributes are chosen arbitrarily and then regards them in two dimension. Then rotate them as per given angel θ with respect to origin. Rotation may be done in either clock wise or anti clock wise. Rotation is based on the value of θ. If θ is positive then we rotate anti clock wise and if θ is negative then we rotate in clock wise direction. Figure 3.2 shows rotation matrix.

$$\begin{bmatrix} cos\theta & sin\theta \\ -sin\theta & cos\theta \end{bmatrix}$$

Fig. 3. Rotation matrix

V. CONCLUSION

In Privacy-Preserving applications correspond to designing data management and mining algorithms in such a way that the privacy remains preserved. There are much privacy preserving technique as above but as we require more privacy and without more data lose and also maintain accuracy these all have some problems. So to overcome with this problem we will use geometric perturbation technique with noise addition, not random noise but Gaussian noise with probability density function to perturbed data. Its utility needs to be preserved. So, the data mining and privacy transformation techniques need to be designed effectively like such as it preserve the utility of the results. Preserve privacy using Multiplicative Data

Perturbation provide Less losses in data and more response time and Privacy must be increase. We have concluded that we have taken a brief review of privacy and data perturbation in data stream which is new increasing area. Till now we are adding random noise to the original data to get the perturbed data but in future we will use Gaussian noise rather than random noise which has more benefits like less data loss than random noise addition. We will use probability density function in that. After that we will apply classification or clustering algorithm to both, original and perturbed data. By using clustering algorithm we can also know about miss placed elements which means the element which were in cluster c1 and after perturbation , they are in cluster c2.

REFERENCES

[1] Golab, L. And Ozsu, M., "Issues in Data Stream Management," ACM SIGMOD Record, Vol. 32, pp. 5-14(2003).
[2] Keke chen, Ling Liu, Privacy Preserving Multiparty Collaborative Mining With Geometric Data Perturbation, IEEE TRANSACTION ON PARALLEL AND DISTRIBUTED COMPUTING, VOL.XX, NO. XX. JANUARY 2009.
[3] H. Chhinkaniwala and S. Garg, "Tuple Value Based Multiplicative Data Perturbation Approach to preserve privacy in data stream mining", IJDKP, Vol3, No.3 May 2013.
[4] Xinjun Qi, Mingkui Zong, "An Overview of Privacy Preserving Data Mining", Science Direct, 2011.
[5] R. Agrawal and R. Shrikant, "Privacy preserving data mining" in Proceeding of ACM SIGMOD Conference, 2000.
[6] A. Evfimievski, R. Shrikant, and J. Gehrke, "Limitimg Privacy Breaches in privacy preserving data mining", in Proceedings of ACM Conference on Principles of Database System(PODS), 2003.
[7] Md Zahidul Islam, Ljiljana Brankovic, "Privacy Preserving Data Mining: A Noise Addition Framework Using A Novel Clustering Technique", Science Direct, 2011.