International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056
Volume: 03 Issue: 06 | June-2016 www.irjet.net p-ISSN: 2395-0072

Geometric Data Perturbation Techniques in Privacy Preserving On Data Stream Mining

Nimpal Patel, Shreya Patel

Student, Dept. Of Computer Engineering, Grow More Faculty of Engineering Himatnagar, Gujarat, India
Asst. Professor, Dept. of Computer Engineering, Grow More Faculty of Engineering Himatnagar, Gujarat, India

Abstract-*Data mining is the information technology that extracts valuable knowledge from large amounts of data. Due to the emergence of data streams as a new type of data, data stream mining has recently become a very important and popular research issue. Privacy preservation issue of data streams mining is very important issue, in this dissertation work, an approach based on Geometric data perturbation has been proposed, which extends the existing process of data streams clustering to achieve privacy preservation. Our objective is to reduce the tradeoff between mining accuracy while minimizing information loss when data undergoing the process of perturbation. Experimental results will show that the method not only can preserve data privacy but also can mine data streams accurately. An effective approach of geometric transformation based data perturbation of data stream for mining has been proposed. It aims privacy of sensitive information before release while obtaining accuracy of data streamclustering with minimum information loss.*

Keywords : Data mining; Privacy preserving; data perturbation ; Geometric Data Perturbation; Gaussian noise; Cluster membership matrix.

1. INTRODUCTION

Databases today can range in size into the terabyte. Within these masses of data lies hidden information of strategic importance. The newest answer is data mining, which is being used both to increase revenues and to reduce costs. The potential returns are enormous. Innovative organizations worldwide are already using data mining to locate and appeal to higher-value customers, to reconfigure their product offerings to increase sales, and to minimize losses due to error or fraud. data mining as a powerful data analysis tool has made tremendous contributions in many areas and has the wide applications prospect. With the development of database technology and network technology,[1] a large number of useful data, which contains much individual privacy information, has been accumulated in various fields, such as patient's condition, customer preferences, personal background information, etc. produce privateinformation disclosure.[2] However data owners are not interested to share their original data sets due to privacy reason. So we need to do some procedure on their original data for privacy purpose before it is going to be released for mining [3]. We are going to present a new data perturbation technique that provides the privacy to outsourced data sets while we mining the data. A Data perturbation approach is simply work in the following manner. Before any data owner is going to publish their original data, they change the original data in such a way that it is becomes very difficult to get the original one [4]. The goal of perturbation technique is twofold. Preserving the accuracy of specific data mining models which are data utility and preserving the privacy of original data [5][6].

1.1 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[7]: (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. 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.

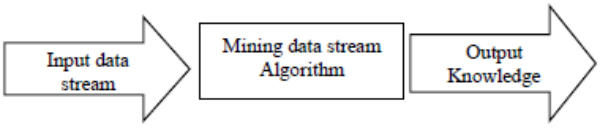

Fig. 1 Data Stream Mining Process

1.2 Need For Privacy in Data Mining

With more and more information accessible in electronic forms and available on the web, and with increasingly powerful data mining tools being developed and put into use, there are increasing concerns that data mining may pose a threat to privacy and data security. However, it is important to note that most of the major data mining applications focus on the development of scalable algorithms and also do not involve personal data. Data mining can be valuable in many applications, but due to no sufficient protection data may be abused for other goals. The main factor of privacy beaching in data mining is data misuse. In fact, if the data consists of critical and private characteristics and/or this technique is abused, data mining can be hazardous for individuals and organizations. Therefore, it is necessary to prevent revealing not only the personal confidential information but also the critical knowledge.

A. Privacy Preserving Data Mining (PPDM)

PPDM has been emerged to address the privacy issues in data mining. Embedding privacy into data mining has been an active and an interesting research area. Several data mining techniques, incorporating privacy protection mechanisms, have been proposed based on different approaches. Recent research in the area of PPDM has devoted much effort to determine a trade-off between privacy and the need for knowledge discovery, which is crucial in order to improve decision-making processes and other human activities. PPDM helps to protect personal, proprietary or sensitive information, to enable collaboration between different data owners and also to comply with legislative policies.

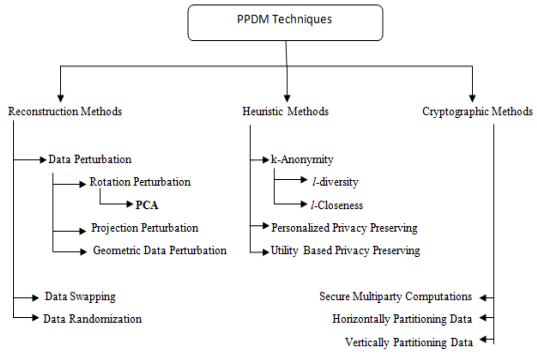

Fig. 2 PPDM Technique

(I). Reconstruction Based Techniques

A number of recently proposed techniques address the issue of privacy preservation by perturbing the data and reconstructing the distributions at an aggregate level in order to perform the mining. The work presented in [7] addresses the problem of building a decision tree classifier in which the values of individual records have been perturbed using randomization method. While it is not possible to accurately estimate original values in individual data records, but reconstruction procedure to accurately estimate the distribution of original data values is very much possible. The work presented in [8] is an improvement over the *Bayesian-based reconstruction* procedure by using an *Expectation Maximization (EM)* algorithm for distribution reconstruction. More specifically, the authors prove that the *EM* algorithm converges to the maximum likelihood estimate of the original distribution based on the perturbed data. Evfiemievski et al. [9] proposed a *select-a-size randomization* technique for privacy preserving mining of association rules. Du et al. [10] suggested randomized response techniques for PPDM and constructed decision trees from randomized data. Other such reconstruction based works are discussed in [11][12][13].

(II). Heuristic Based Techniques

There have been several methods developed by researchers in the database community that process records in a "group-based" manner, using information about specific local records globally to transform the records in a way which preserves specific privacy metrics. These modified records can then be published without fear of reconstruction by attacks[14]. There is an assumption that certain fields of a record contain *quasi-identifier* attributes that uniquely identify an individual associated with the record, as well as *sensitive* attributes that must not be linked to the individual by an untrusted third party. Three variants of grouping-based methods (*k-anonymity*, *ℓ-diversity*, and *t-closeness*) have been proposed that rely on achieving the final state where *k* records look exactly the same[15].

(III). Cryptographic Based Techniques

Cryptographic based methods become hugely popular for two main reasons: Firstly, cryptography offers a well defined model for privacy, which includes methodologies for proving and quantifying it. Secondly, there exists a vast toolset of cryptographic algorithms to implement PPDM algorithms. However, cryptography does not protect the output of a computation. Instead, it prevents privacy leaks in the process of computation. The data sources may be distributed across the network and hence pooling them at a centralized location may not be possible due to limitations in computational and communication resources[16]. Distributed applications store data using two models: vertically partitioned data model and horizontally partitioned data model. Distributed data mining

provides algorithms to perform data mining in a distributed setting without pooling the data into one location privacy considerations can prohibit organizations from sharing their data with each other[17].

Techniques from secure multiparty computation [18] form one approach to privacy preserving in distributed data mining.[19] introduced first secure multiparty computation for secure circuit evaluation, in theory, to compute any function over data partitioned between two parties, without revealing anything to either party beyond the computed output. However, because data mining usually involves millions or billions of data items, the communication costs of these protocols render them impractical for these purposes. This has led to the search for problem specific protocols that have efficient communication complexity. Lindell and Pinkas [20] used cryptographic techniques in their protocol. Their work differs from general secure multiparty computation in the sense that most computation is done locally by the individual parties[21].

2. GEOMETRIC TRANSFORMATION BASED ON DATA PERTURBATION

A. PROPOSED METHOD

The Goal Behind this method is how we can achieve better privacy while mining. The first step of it understood of various mining techniques used to work with Stream data perturbation to gathering sensitive data. So our focus is on transformation of original data sets that generates satisfactory result and preserves as much data as possible for future analysis task. We are expanding Geometric data perturbation method that is totally distance based. Geometric data perturbation is more powerful and enhanced technique than rotation perturbation. Geometric perturbation addresses the weakness of Rotation Perturbation by additional parameters Gaussian Noise and Random translation. Fig. 3 describes framework for data perturbation to protect privacy of sensitive attribute. Sensitive attribute has been extracted and given as an input to proposed algorithm for random noise addition. Perturbed sensitive attribute the replace sensitive attribute in original dataset. Data clustering algorithm with same parameter settings has been used with original datset D and perturbed dataset D'. Clustreing results R and R' from original dataset D and perturbed dataset D' respectively have been stored in cluster membership matrix.

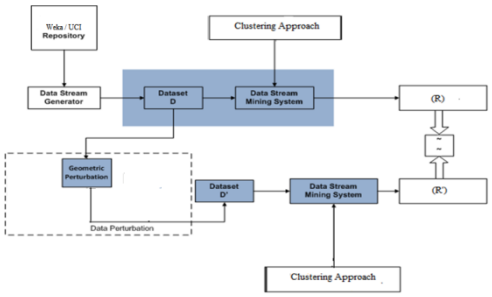

Fig.3.Framework for privacy preserving in data stream clustering Approach.

B. ALGORITHM:

Procedure: Geometric Transformation Based On 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 $[\boldsymbol{S}]_{\boldsymbol{n\times 1}}$.

[2]. Rotate $[\boldsymbol{S}]_{\boldsymbol{n\times 1}}$ using cosθ and sinθ function and generate $[\boldsymbol{R_S}]_{\boldsymbol{n\times 1}}$.

[3]. Multiply elements of [S] with $[R_S]$ transformed sensitive attribute values will be

$$[\boldsymbol{X}]_{\boldsymbol{n\times 1}} = [\boldsymbol{S}]_{\boldsymbol{n\times 1}} \times [\boldsymbol{R_S}]_{\boldsymbol{n\times 1}}$$

[4]. Calculate translation T as mean of sensitive attribute $[\boldsymbol{S}]_{\boldsymbol{n\times 1}}$.

[5]. Generate transformation $[\boldsymbol{S}]_{\boldsymbol{n\times 1}}$ by applying translation T to $[\boldsymbol{S}]_{\boldsymbol{n\times 1}}$.

[6]. Calculate Gaussian distribution $P(x)$ as a probability density function for Gaussian noise

$$\boldsymbol{P(x)} = \frac{\boldsymbol{1}}{\boldsymbol{\sigma\sqrt{2\pi}}} \boldsymbol{e}^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

where, μ=Mean, **σ**=Variance

[7]. Geometric data perturbation of sensitive attribute $[\boldsymbol{Gs}]_{\boldsymbol{n\times 1}} = [\boldsymbol{X}]_{\boldsymbol{n\times 1}} + [\boldsymbol{St}]_{\boldsymbol{nx1}} + \boldsymbol{P(x)}$.

[8]. Crate perturbed dataset D' by replacing sensitive attribute $[S]_{n\times 1}$ in original dataset D with $[\boldsymbol{Gs}]_{\boldsymbol{n\times 1}}$.

[9]. Apply **Hierarchical *k-Mean*** clustering algorithm with different values of k on original dataset D having sensitive attribute S.

[10]. Apply **Hierarchical *k-Mean*** clustering Algorithm with different values of k on Perturbed dataset D' having perturbed sensitive attribute Gs.

[11]. Create cluster membership matrix of results from step 9 and step 10 and analyse.

3. EXPERIMENTS AND RESULTS

A. Experimental Setup:

We are going to perform algorithm operations on Bank management dataset, Adult data set. These data sets are available at the UCI machine learning Repository[22]. There are number of attributes in datasets, some of them are numerical on which we are applying geometric data perturbation technique then passing it from classification algorithm and we compare result of both original and modified. To have accurate and proper result we must have proper dataset and well setup software and hardware environmental setup. Here we will have the best environment to get perfect result.

B. Cluster Membership Matrix (CMM)

Cluster Membership Matrix identifies how closely each cluster in the perturbed dataset matches its corresponding cluster in the original Dataset in table I. Rows represent the clusters in the original dataset, while Columns represent the clusters in the perturbed dataset, and $Freq_{i,j}$ is the number of points in cluster C_i that falls in cluster C_i' in the perturbed dataset. Table II shows the percentage of accuracy obtained when selected attribute are perturbed using our algorithm in each dataset.

TABLE I. CLUSTER MEMBERSHIP MATRIX

	C1'	C2'	……	Cn'
C1	$Freq_{1,1}$	$Freq_{1,2}$	……	$Freq_{1,n}$
C2	$Freq_{2,1}$	$Freq_{2,2}$	……	$Freq_{2,n}$
....	……	……	……	….
Cn	$Freq_{n,1}$	$Freq_{n,2}$	……	$Freq_{n,n}$

TABLE II. ACCURACY OBTAINED

Dataset Name	Total instances	Attribute	K=2	K=3	K=4	K=5
Adult	32561	Age	0.75	0.72	0.69	0.68
		Income	0.98	0.96	0.92	0.89
		Age, Income	0.79	0.73	0.71	0.68
		Elevation	0.98	0.85	0.80	0.84
Cover Dataset	100000	Price(aspect)	0.74	0.71	0.70	0.70
		Slope	0.96	0.95	0.94	0.91
		Ele, Price	0.73	0.71	0.71	0.69
		Ele, slope	0.94	0.90	0.89	0.87
		Price, slope	0.76	0.72	0.70	0.68
		Ele,Price, Slope	0.89	0.83	0.76	0.72
Bank Mgmt.	45211	Income	0.97	0.96	0.92	0.88
		Duration	0.92	0.89	0.88	0.83
		Income, Duration	0.93	0.91	0.87	0.82
Letter Reorganization	15327	Lno	0.97	0.94	0.94	0.88
Average			0.879286	0.841429	0.816429	0.790714

Hierarchical k-Mean clustering algorithm has been applied on original dataset D and perturbed dataset D' generated using proposed algorithm. Results in table II shows that for all tested cases almost 90% mining accuracy has been achieved. Algorithm has been tested against different values of *k* and it has been observed that accuracy has been decreasing as *k* value increases. This justifies that probability of tuple to fall into original cluster will be decreasing as number of clusters increases.

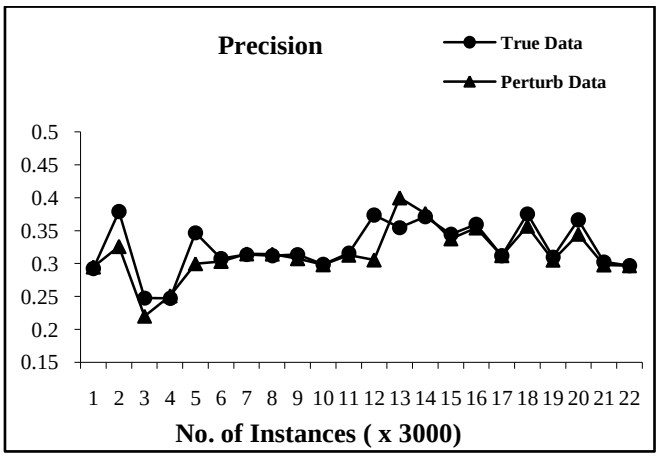

Fig.4 Accuracy on attribute Elevation in Cover type dataset using precision

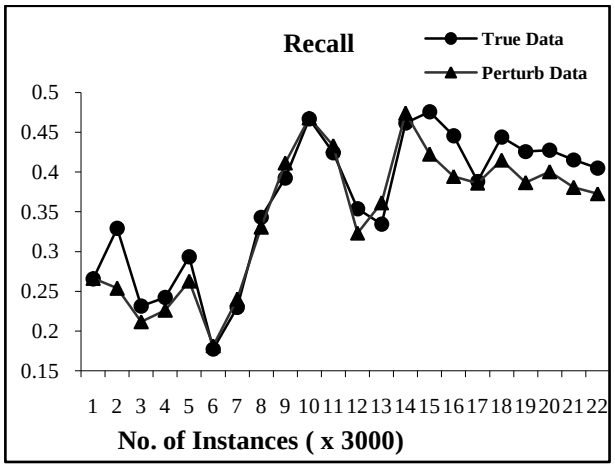

Fig. 5 Accuracy on attribute Elevation in Cover type dataset using recall

4. CONCLUSION

Our approach was motivated by reconstruction method commonly used in privacy preservation of sensitive data. The method followed the statistical relationship among the tuple attributes. Our approach considered sensitive attribute as dependent attribute and remaining attributes of a tuple. We applied Hierarchical *K-Mean*s algorithm over defined sliding window size on perturbed data stream in order to estimate the accuracy of clustering results. Our experiments show that proposed method is an efficient, flexible and easy-to-use method in PPDM. tuple except class attribute as independent attributes. We used independent attributes of tuple to calculate random noise. This noise is specific to Perturbation of the dataset can be implemented independently to individual variable and can also be implemented simultaneously to a group of variables. We quantified the privacy of our scheme using the concept of misclassification error. Information loss due to data perturbation was quantified by a loss of accuracy, which can be quantified by percentage of instances of data stream are misclassified using cluster membership matrix. Proposed method shows reasonably good results tested against evaluation measures. our experiments to only numeric type attributes.

REFERENCES

[1] U. M. Fayyad , G. P. Shapiro and P. Smyth, From Data Mining to Knowledge Discovery in Databases. 0738-4602-1996, AI Magazine (Fall 1996). pp: 37–53.

[2] J. Han and M. Kamber, Data Mining: Concepts and Techniques. Second edition Morgan Kaufmann Publishers.

[3] Majid, M.Asger , Rashid Ali, "Privacy preserving Data Mining Techniques: Current Scenario and Future Prospects", IEEE 2012.

[4] L.Golab and M.T.Ozsu, Data Stream Management issues-" A Survey Technical Report", 2003.

[5] Majid , M. Asger, Rashid Ali, "Privacy preserving Data Mining Techniques: Current Scenario and Future Prospects", IEEE 2012.

[6] Aggrawal, C.C, and Yu.PS. ," A condensation approach to privacy preserving data mining". Proc. Of In .conf. on extending Database Technology (EDBT)(2004).

[7] R. Agrawal and R. Srikant, "Privacy preserving data mining," In Proceedings of SIGMOD Conference on Management of Data, pp. 439-450, 2000.

[8] D. Agrawal and C. Aggarwal, "On the design and quantification of privacy preserving data mining algorithm," In Proceedings of ACM SIGMOD, pp. 247-255, 2001.

[9] A. Evfimieski, R. Srikant, R. Agrawal and J. Gehrke, "Privacy preserving mining of association rules," In Proceedings of the 8th ACM SIGKDD, pp. 217-228, 2002.

[10] W. Du and Z. Zhan, "Using randomized response techniques for PPDM," In Proceedings of the 9th ACM SIGKDD, pp. 505-510, 2003

[11] K. Liu, H. Kargupta and J. Ryan, "Random projection-based multiplicative perturbation for privacy preserving distributed data mining," IEEE Transactions on Knowledge and Data Engineering, vol. 18, no. 1, pp. 92-106, 2006.

[12] J. Ma and K. Sivakumar, "Privacy preserving Bayesian network parameter learning," 4th WSEAS International Conference on Computational Intelligence, Man-machine Systems and Cybernetics, Miami, Florida, November, 2005.

[13] J. Ma and K. Sivakumar, "A PRAM framework for privacy -preserving Bayesian network parameter learning," WSEAS Transactions on Information Science and Applications, vol. 3, no. 1, 2006.

[14] M. Kantarcioglu and J. Vaidya, "Privacy preserving naive Bayes classifier for horizontally partitioned data," In Workshop on Privacy Preserving Data Mining," The 3rd IEEE International Conference on Data Mining, pp 19-22, 2003.

[15] M. Kantarcioglu and C. Clifton, "Privacy-preserving distributed mining of association rules on horizontally partitioned data," In Proceedings of ACM SIGMOD

Workshop on Research issues on Data Mining and Knowledge Discovery, pp 24-31, 2002.

[16] J. Vaidya and C. Clifton, "Privacy preserving association rule mining in vertically Partitioned data." In Proceedings of the 8th ACM SIGKDD international conference on Knowledge discovery in data mining, pp 639-644, 2002.

[17] R. Agrawal, A. Evfimievski and R. Srikant, "Information sharing across private databases," In Proceedings of ACM SIGMOD International Conference on Management of Data, San Diego, CA, 2003.

[18.] O. Goldreich, "Foundations of Cryptograph," vol. 2,Cambridge University Press, 2004.

[19.] A. C. Yao, "How to generate and exchange secrets," In Proceedings of the 27th IEEE Symposium on Foundations of Computer Science, pp 162-167, 1986.

[20] Y. Lindell and B. Pinkas, "Privacy preserving data mining, Journal of Cryptology ," vol. 15, issue 3, pp 177-206, 2002.

[21] J. Vaidya and C. Clifton, "Privacy-preserving naïve Bayes classifier for vertically partitioned data," In Proceedings of 4th SIAM International Conference on Data Mining, pp 522-526, 2004.

[22] UCI Machine Learning Repository http://archive.ics.uci.edu/ml/datasets.