Privacy Is Become With, Data Perturbation

Author :- Niky Singhai
M-Tech (Computer science and engineering)
TIT Bhopal
Bhopal, (M.P.) India
nikysinghai@gmail.com

Author :- Er. Niranjan Singh
M-Tech (Computer science and engineering)
TIT Bhopal
Bhopal, (M.P.) India
enggniranjan@gmail.com

Abstract— **Privacy is becoming an increasingly important issue in many data mining applications that deal with health care, security, finance, behavior and other types of sensitive data. Is particularly becoming important in counter-terrorism and homeland security-related applications. We touch upon several techniques of masking the data, namely random distortion, including the uniform and Gaussian noise, applied to the data in order to protect it. These perturbation schemes are equivalent to additive perturbation after the logarithmic Transformation. Due to the large volume of research in deriving private information from the additive noise perturbed data, the security of these perturbation schemes is questionable Many artificial intelligence and statistical methods exist for data analysis interpretation, Identifying and measuring the interestingness of patterns and rules discovered, or to be discovered is essential for the evaluation of the mined knowledge and the KDD process as a whole. While some concrete measurements exist, assessing the interestingness of discovered knowledge is still an important research issue. As the tool for the algorithm implementations we chose the "language of choice in industrial world" – MATLAB.**

.

Keywords- ***KDD, data mining, perturbation, additive perturbation, security, ogarithmic, artificial intelligence, statistical methods, data analysis.***

I. Introduction

Most of our daily activities are now routinely recorded and analyzed by the variety of governmental and commercial organization for the purpose of security and business related applications .From telephone calls to credit card purchases, from Internet surfing to medical prescription refills, we generate data with almost every action we take. Collecting and analyzing such data are causing a major concern about our privacy.

Recent Interest in the collection and monitoring of data using data mining technology for the purpose of security and business-related applications has raised serious concerns about privacy issues. For example, mining health care data for the detection of disease outbreaks may require analyzing clinical records and pharmacy transaction data of many individuals over a certain area. However releasing and gathering such diverse information belonging to different parties may violate privacy laws and eventually be a threat to civil liberties. Privacy preserving data mining strives to provide a solution to this dilemma. It aims to allow useful data patterns to be discovered without compromising privacy.

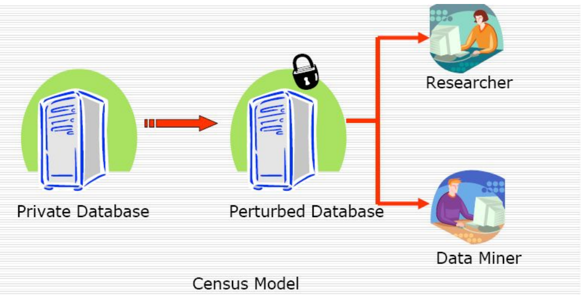

Figure 1 this is referred to as the census model

The problem can be stated as an organization has a private database and wishes to make it publicly available for data analysis while keeping the original data records private. To achieve this goal, this organization transforms its database into another form and only releases that. A third party data miner or researcher can analyze and discover useful patterns of the original data from only the transformed data.

II. PROPOSED TECHNIQUE

The main objective of data hiding is to transform the data so that the private data remains private during and/or after data mining operations. This section presents a classification and an extended description of the various techniques and methodologies that have been developed in this area.

A. Data Perturbation

Data perturbation techniques can be grouped into two main categories, which we call the value distortion technique and probability distribution technique. The value distortion technique perturbs data elements or attributes directly by either some other randomization procedures. On the other hand, the probability distribution technique considers the private database to be a sample from a given population that has a given probability distribution. In this case, the perturbation replaces the original database by another sample from the same (estimated) distribution or by the distribution itself.

There has been extensive research in the area of statistical databases (SDB) on how to provide summary statistical information without disclosing individual's confidential data. The privacy issues arise the summary statistics are derived from data of very few individuals. A popular disclosure control method is data perturbation, which alters individual data in a way such that the summary statistics remain approximately the same. However, problems in data mining become somewhat different from those in SDBs. Data mining techniques, such as clustering, classification, predication and association rule mining are essentially relying on more sophisticated relationships among data records or data attributes, but no just simple summary statistics. This project specifically focuses on data perturbation for privacy preserving data mining. In the following, we will primarily discuss different perturbation techniques in the data mining area. Some important perturbation approaches in SDBs are also covered for sake of completeness.

B. Multiplicative Perturbation

Two basic forms of multiplicative noise have been studied in the statistics community. One multiplies each data element by a random number that has a truncated Gaussian distribution with mean one and small variance. The other takes a logarithmic transformation of the data first, adds multivariate Gaussian noise, and then takes the exponential function exp (.) of the noise-added data. Neither of these perturbations preserves pair wise distance among data records.

Where the data is multiplied by a randomly generated matrix – in effect, the data is projected into a lower dimensional random space. This technique preserves distance on expectation. Oliveira and Zaiane, Chen and Liu discussed the use of random rotation for privacy preserving clustering and classification. These authors observed that the distance preserving nature of random rotation enables a third party to produce exactly the same data mining results on the perturbed data as if on the original data. However, they did not analyze the privacy limitations of random rotation. Liu *et al.* addressed the privacy issues of distance preserving perturbation (including rotation) by studying how well an attacker can recover the original data from the transformed data and prior information. They proposed two attack techniques: the first is based on basic properties of linear algebra and the second on principal component analysis. Their analysis explicitly illuminated scenarios where privacy can be breached. As such, valuable information was gained into the effectiveness of distance preserving transformation for privacy preserving data mining.

Mukherjee *et al.* considered the use of discrete fourier transformation (DFT) and discrete cosine transformation (DCT) to perturb the data. Only the high energy DFT/DCT coefficients were used, and the transformed data in the new domain approximately preserved the Euclidean distance. The DFT/DCT coefficients were further permutated to enhance the privacy protection level. However, the authors did not offer a rigorous analysis of the privacy. Also note that if no coefficients were dropped, their technique would be fundamentally the same as distance preserving transformation; therefore, the privacy issues could be analyzed using the model proposed by Liu.

C. Data Micro-aggregation

Data micro-aggregation is a popular data perturbation approach in the area of secure statistical databases (SDBs). For a data set with a single private attribute univariate micro-aggregation sorts data records by the private attribute, group's adjacent records into groups of small sizes, and replaces the individual private values in each group with the group average. Multivariate micro-aggregation considers all the attributes and groups data using a clustering technique. This approach primary considers the reservation of data covariance instead of the pair wise distance among data records.

Recently, researchers in the data mining area have proposed two multivariate micro-aggregation approaches. Agarwal and Yu presented a consideration approach to privacy data mining .This approach first partitions the original data into multiple groups of predefined size. For each group, a certain level of statistical information.(e.g. mean and covariance) about different data records is maintained. This statistical information is used to create anonymized data that has similar statistical characteristics to the original data set, and only the anonymized data is released for data mining applications. This approach preserves data covariance instead of the pair-wise distance among data records. Proposed kd-tree based perturbation method which recursively partitions a data set into smaller subset such that data records in each subset are more homogeneous after each partition, The private data in each subset are than perturbed using the subset average. The relationship between-attributes are expected to be preserved.

1) Data Anonymization:

Sweeney developed the k-anonymity frame work wherein the original data is transformed so that the information for any individual cannot be distinguished from (k-1) others. Generally speaking, anonymization is achieved by suppressing (deleting) individual values from data records.(e.g., name , and social security numbers are removed), and/or replacing every occurrence of certain attribute values with a more general value(e.g. the zip codes 21250-21259 might be replaced with 2125*) A variety of refinement of this frame work have been proposed since its initial appearance. Some of the work start from the original data set and systematically or greedily generalize it into one that is k-anonymous. Some start with a fully generalized data set and systematically specialize the data set into one that is minimally k-anonymous. The problem of k-anonymizaion is not simply to find any k-anonymization, but to, instead, finds one that is "good" or even "best" according to some quantifiable cost metric. Each of the previous work provides its own unique cost metrics for modeling desirable anonymization.

Recently, Machanavajjhala pointed that simple k-anonymity id vulnerable to strong attacks due too lack of diversity in the sensitive attributes. They proposed a new privacy definition called l-diversity. The main idea behind I-diversity is the requirement that the values of the sensitive attributes are well represented in each group. Other enhanced k-anonymity models have been proposed elsewhere.

2) Data Swapping:

This technique transforms the database by switching subsets of attributes record entries are unmatched, but the statistics (e.g., marginal distributions of individual attributes) are maintained across the individual fields this technique was first proposed by Dalenius and Reiss. A variety of refinements and applications of data swapping have been addressed since its initial appearance.

III. TEST RESULTS AND ANALYSIS

I considered the cardiology database from the Internet without loss of generality, I selected the first 303 rows of the data with only 14 attributes.(age, cholesterol, maximum heart rate, peak, slope, thal).

Age	cholesterol	maximum heart rate	peak	Slope	thal
60	206	132	2.4	2	130
49	266	171	0.6	1	130
64	211	144	1.8	2	110
63	254	147	1.4	2	130
53	203	155	3.1	3	140
58	224	173	3.2	1	132
63	233	150	2.3	3	145
67	229	129	2.6	2	120

Table 1 Part of Original Data before Perturbation

1.4647	3.0103	0.63018	2.0014	0.0051443	1.4647
0.49949	0.56027	0.3205	1.4227	1.0147	0.49949
0.72158	0.35545	0.91954	0.075245	0.31358	0.72158
0.11508	0.03844	1.0261	0.93879	0.43129	0.11508
0.27168	0.28699	1.2543	0.67087	0.3837	0.27168
0.78424	0.30111	0.14083	2.3763	4.878	0.78424
3.9898	0.83962	0.46775	3.8305	0.42442	3.9898
0.19674	0.45529	1.4222	1.8355	0.056322	0.19674

Table 2 Part of Random projection matrix

Above table shows the random number generated by probability distribution with parameter mean and variance and the rotation of that matrix take place to some angle

54.366	246.26	149.65	2.14851	2.9835	131.62

Table 3 Mean of the original dataset

54.366	246.26	149.65	2.14851	2.9835	131.62

Table 4 Mean of the perturbated dataset

Above tables gives the mean (statistical property) of the original data and of the perturbed data respectively.

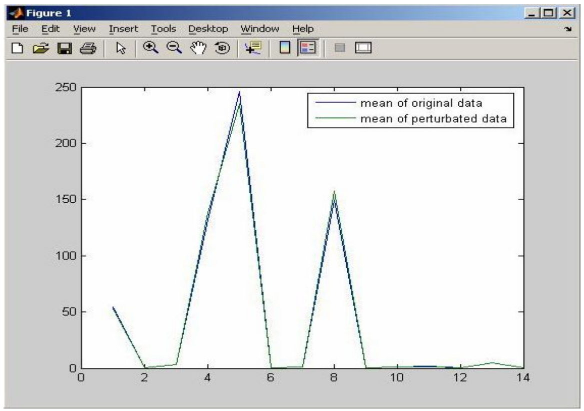

Fig 2 Mean of original dataset and perturbed

82.212	2677.6	522.91	3.752	2.343	306.57

Table 5 Variance of the original dataset:

3031.6	61005	31548	4.34	3.5367	26167

Table 6 Variance of the perturbated dataset:

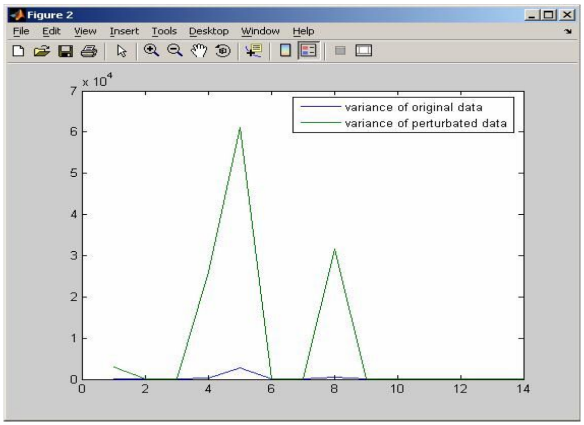

Fig 3 Variance of original dataset and perturbed dataset:

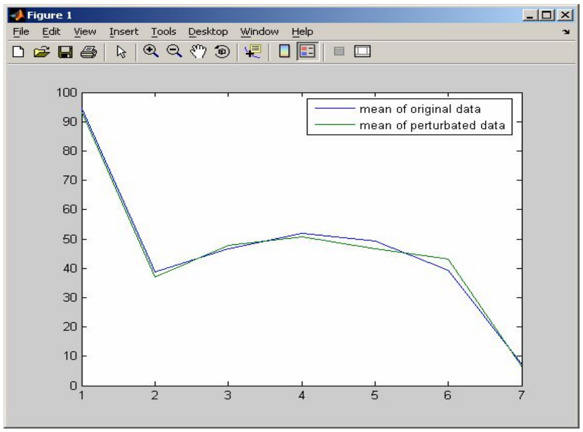

Fig 4 Mean of original dataset and perturbed dataset:

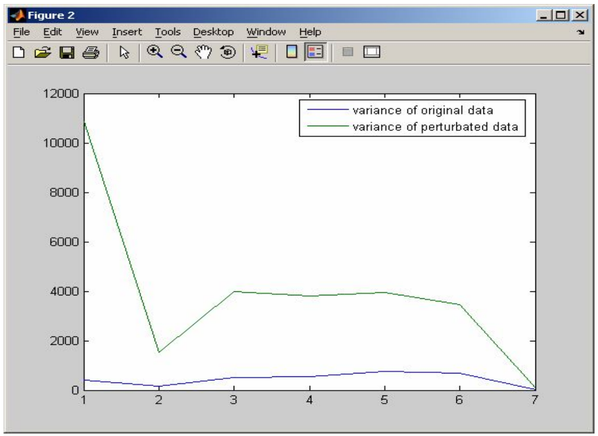

Fig 5 Variance of original dataset and perturbated dataset:

IV. CONCLUSION

This paper may require creating profiles, constructing social network models, and detecting terrorists' communications. All of them involve the collection and analysis of private sensitive data. However, releasing and combining such diverse data sets belonging to different parties may violate privacy laws. Although health organizations allowed to release the data as long as the identifiers (e.g., name, SSN, address, etc.,) removed, it is not considered safe enough because re-identification attacks may be constructed for linking different public data sets to identify the original subjects . This calls for well-designed techniques that pay careful attention to hiding privacy sensitive information while preserving the inherent patterns of the original data. Privacy preserving data mining (PPDM) strives to provide a solution to this problem. It aims to allow useful data patterns to be extracted without compromising privacy. This paper specifically investigates the characteristics of different multiplicative data perturbation techniques for PPDM. First, we have briefly reviewed two traditional multiplicative data perturbation techniques that have been well studied in the statistics community.

These perturbation schemes are equivalent to additive perturbation after the logarithmic Transformation. Due to the large volume of research in deriving private information from the additive noise perturbed data, the security of these perturbation schemes is questionable.

Next, we have examined the effectiveness of distance preserving perturbation. Theoretical and experimental results have shown the following. This type of perturbation is essentially a series of rotations and reflections of the data. It exactly preserves the Euclidean distances and inner products in the original data. Therefore, many interesting data mining algorithms can be applied directly, to the perturbed data and produce an error-free result. However, this perturbation is vulnerable to many attacks such as known input-output attacks, known sample attacks and independent signals attacks. Finally, I have explored a random projection-based perturbation. This technique projects the data into a lower dimensional subspace while maintaining the pair wise distances of the original records with high probabilities. I have shown that.

We believe that the privacy issues are intrinsically complex because they representation intersection of legal, governmental, commercial, ethical and personal positions. It is not easy to produce on universal solution that addresses all these perspectives when the very definition of privacy is still open to debate .But the pressure is on to take more positive steps to encourage privacy protection while doing data mining to benefit the society.

V. Reference

[1] A. L. Penenberg, “The end of privacy,”

[2] B. Thuraisingham, “Data mining, national security, privacy and civil liberties,”

[3] S. E. Committee, ““Data Mining” is NOT against civil liberties,” http://www.acm.org/sigs/sigkdd/civil -liberties.pdf, June 30, 2003.

[4] R. Agrawal and R. Srikant, “Privacy preserving data mining,”

[5] H. Kargupta, S. Datta, Q. Wang, and K. Sivakumar, “On the privacy preserving properties of random data perturbation techniques,” in *Proceedings of the IEEE International Conference on Data Mining*.

[6] Z. Huang, W. Du, and B. Chen, “Deriving private information from randomized data,” in *Proceedings of the 2005 ACM SIGMOD Conference*, Baltimroe, MD, June 2005, pp. 37–48.

[7] S. Guo and X. Wu, “On the use of spectral filtering for privacy preserving data mining,” in *Proceedings of the 21st ACM Symposium on Applied Computing*, Dijon, France, April 2006, pp. 622–626.

[8] N. R. Adam and J. C. Worthmann, “Security-control methods for statistical databases: a comparative study,” *ACM Computing Surveys (CSUR)*, vol. 21, 145 146 no. 4, pp. 515–556, 1989. [Online]. Available: http://portal.acm.org/citation.cfm? id=76895

[9] G. T. Duncan and S.Mukherjee, “Optimal disclosure limitation strategy in statistical databases: Dterring tracker attacks through additive noise,” *Journal of The American Statistical Association*, vol. 95, no. 451, pp. 720–729, 2000.

[10] R. Gopal, R. Garfinkel, and P. Goes, “Confidentiality via camouflage: The cvc approach to disclosure limitation when answering queries to databases,” *Operations Research*, vol. 50, no. 3, pp. 501–516, 2002.

[11] D. Agrawal and C. C. Aggarwal, “On the design and quantification of privacy preserving data mining algorithms,” in *Proceedings of the twentieth symposium on Principles of Database Systems*, Santa Barbara, CA, 2001, pp. 247–255. [Online]. Available: http://portal.acm.org/citation.cfm?id=375602

[12] S. Guo, X. Wu, and Y. Li, “On the lower bound of reconstruction error for spectral filtering based privacy preserving data mining,” in *Proceedings of the 10th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD’06)*, Berlin, Germany, 2006.

[13] K. Muralidhar, R. Parsa, and R. Sarathy, “A general additive data perturbation method for database security,” *Management Science*, vol. 45, no. 10, pp. 1399–1415, 1999.

[14] K. Muralidhar and R. Sarathy, “A theoretical basis for perturbation methods,” *Statistics and Computing*, vol. 13, no. 4, pp. 329–335, 2003.

[15] A. Weingessel, E. Dimitriadou, K. Hornik. An Ensemble Method for Clustering. DSC 2003 Working Papers.

[16] S. Evfimievski. Randomization techniques for privacy preserving association rule mining.

[17] E. Dimitriadou, A. Weingessel, K. Hornik. Voting