A Comparative Study of Multiplicative Data Perturbation Techniques for Privacy Preserving Data Mining

Bhupendra Kumar Pandya
Institute of Computer Science
Vikram University, Ujjain

Umesh kumar Singh
Institute of Computer Science
Vikram University, Ujjain

Keerti Dixit
Institute of Computer Science
Vikram University, Ujjain

Abstract: Data perturbation techniques are one of the most popular models for privacy preserving data mining. It is especially useful for applications where data owners want to participate in cooperative mining but at the same time want to prevent the leakage of privacy-sensitive information in their published datasets. The goal of privacy preserving data mining is to develop data mining methods without increasing the risk of misuse of the data used to generate those methods. The topic of privacy preserving data mining has been extensively studied by the data mining community in recent years. This research paper systematically investigated different multiplicative data perturbation techniques for privacy preserving data mining. These types of perturbation distort the private data by multiplying some random noise and only the perturbed version is released for data mining analysis. We have analyzed these techniques on the basis of Utility, Privacy and accuracy and we have deduced the pros and cons of each technique.

Keywords: Multiplicative Data Perturbation Techniques

1. Introduction:

Data mining and knowledge discovery in databases are two new research areas that investigate the automatic extraction of previously unknown patterns from large amounts of data. However, many data owners may be reluctant to share their data with others due to privacy and confidentiality concerns. This is a serious impediment to perform mutually beneficial data mining tasks.

PRIVACY is becoming an increasingly important issue in many data-mining applications that deal with healthcare, security, financial, behavioural, and other types of sensitive data. It is particularly becoming important encounter terrorism and homeland defence-related applications [1].

Data perturbation refers to a data transformation process typically performed by the data owners before publishing their data. The goal of performing such data transformation is two-fold. On one hand, the data owners want to change the data in a certain way in order to disguise the sensitive information contained in the published datasets, and on the other hand,

ISSN: 2278 – 1323

the data owners want the transformation to best preserve those domain-specific data properties that are critical for building meaningful data mining models, thus maintaining mining task specific data utility of the published datasets.

2. APPLICATIONS OF PRIVACY PRESERVING TECHNIQUES

The problem of privacy-preserving data mining has numerous applications in homeland security, medical database mining, and customer transaction analysis. Some of these applications such as those involving bio-terrorism and medical database mining may intersect in scope.

Medical Databases: There are systems designed for de identification of clinical notes and letters which typically occurs in the form of textual data. Clinical notes and letters are typically in the form of text which contain references to patients, family members, addresses, phone numbers or providers. Traditional techniques simply use a global search and replace procedure in order to provide privacy. However clinical notes often contain cryptic references in the form of abbreviations which may only be understood either by other providers or members of the same institution. Therefore traditional methods can identify no more than 30-60% of the identifying information in the data. This system was designed to prevent identification of the subjects of medical records which may be stored in multidimensional format. The multi-dimensional information may include directly identifying information such as the social security number, or indirectly identifying information such as age, sex or zip-code. The system was designed in response to the concern that the process of removing only directly identifying attributes such as social security numbers was not sufficient to guarantee privacy.

Bioterrorism Applications: In typical bioterrorism applications, analysis of medical data for privacy preserving data mining purposes takes place. Often a biological agent produces symptoms which are similar to other common respiratory diseases such as the cough, cold and the flu. In the absence of prior knowledge of such an attack, health care providers may diagnose a patient affected by an anthrax attack of have symptoms from one of the more common respiratory diseases. The key is to quickly identify a true anthrax attack from a normal outbreak of a common respiratory disease, In many cases, an unusual number of such cases in a given locality may indicate a bio-terrorism attack. Therefore, in order to identify such attacks it is necessary to track incidences of these common diseases as well. Therefore, the corresponding data would need to be reported to public health agencies. However, in the

ISSN: 2278 – 1323

event of suspicious activity, it allows a drill-down into the underlying data. This provides more identifiable information in accordance with public health law.

3. Traditional Multiplicative Data Perturbation:

Perturbation Scheme I:

Perturbation Scheme[2,3,4]: Let x_i be the ith attribute of a private database. Let x_i be the private value for the i^{th} attribute of the jth record in the database, i = 1, . . . , n, j=1, . . . , m. Let r_{ij} denote the random noise corresponding to xij. The perturbed data y_{ij} is

$$y_{ij} = x_{ij} * r_{ij} ,$$

where r_{ij} is independent and identically chosen from a Gaussian distribution with mean 1 (usually μ_i =1) and variance σi2. In other words, all r_{ij}'s for a given I follow the same distribution.

Perturbation Scheme II:

Let xij be the value for the i-th attribute of the j-th record in the database as before i=1... n,j=1…m.

Let We generate the random noise following the multivariate Gaussian Distribution N (0, c ∑ u), where 0 < c < 1 and ∑ u is the covariance matrix of variables u1, u2… un. We denote the noise as eij. Let

zij = uij + eij,

yij = exp(zij)

=exp(In xij+eij)

=xij exp(eij)

= xij hij.

This perturbed data yij is released then. Note scheme assumes that all xij are positive.

4. Euclidean Distance Preserving Perturbation

Orthogonal transformation-based data perturbation [5,6,7] can be implemented as follows. Suppose the data owner has a private database $X_{n\times m}$, with each column of X being a record and each row an attribute. The data owner generates an n × n orthogonal matrix M_T , and computes

$$Y_{n\times m} = M_{Tn\times n}X_{n\times m}$$

The perturbed data $Y_{n\times m}$ is then released for future usage.

ISSN: 2278 – 1323

5. Projection Based Data Perturbation

Random projection refers to the technique of projecting a set of data points from a high dimensional space to a randomly chosen lower dimensional space[8,9,10]. Mathematically, let $X \epsilon R^{n \times m}$ be m data points in n-dimensional space. The random projection method multiplies X by a random matrix $R \epsilon R^{k \times n}$, reducing the n dimensions down to just k. It is well known that random projection preserves pairwise distances in the expectation.

$$X^{*}_{q \times n} = R^{*}_{q \times p} * X_{p \times n}$$

Thus we can see that transforming the data to a random projection space is a simple matrix multiplication with the guarantees of distance preservation.

6. CAMDP Technique

CAMDP (Combination of Additive and Multiplicative Data Perturbation) Method combines the strength of the translation and distance preserving method. Translation and Orthogonal transformation-based data perturbation can be implemented as follows. Suppose the data owner has a private database $D_{m \times n}$, with each column of D being a record and each row an attribute. The data owner generates a n × n noise matrix O_R , and computes

$$D'_{m \times n} = D_{m \times n} * OR_{n \times n}$$

Where $OR_{n \times n}$ is generated by Translation and Orthogonal Transformation.

The perturbed data $D'_{m \times n}$ is then released for future usage.

7. Comparative Study of Multiplicative Data Perturbation techniques For Privacy Preserving Data Mining on the basis of Euclidean Distance, K-means Clustering and KNN Classification.

In this study we have taken Students result database of Vikram University, Ujjain. We have randomly selected 7 rows of the data with only 7 attributes(Marks of Foundation, Marks of Mathematics, Marks of Physics, Marks of Computer Science, Marks of Physics Practical, Marks of Computer Science Practical and Marks of Job Oriented Project).

ISSN: 2278 – 1323

Multiplicative Data Perturbation Techniques	Euclidean Distance	K-means Clustering	KNN Classification
Traditional Perturbation Scheme I & II	Not Preserve	Not Applicable	Not Applicable
Euclidean Preserving Perturbation	Exactly Preserve	Exact Clustering	Exact Classification
Projection Based Perturbation	Expected Preserve	Expected Clustering	Expected Classification
CAMDP(Combination of additive and Multiplicative Data Perturbation	Exactly Preserve	Exact Clustering	Exact Classification

Table 1

8. Merits and Demerits of Multiplicative Data Perturbation techniques for Privacy Preserving Data Mining

Multiplicative Data Perturbation Techniques	Merit	Demerit
Traditional Perturbation Scheme I & II	This technique is used to mask the private data while allowing summary statistics to be estimated.	This technique distort each data element independently, therefore Euclidean distance and inner product among data records are usually not preserved, and the perturbed data can not be used for many data mining applications.
Euclidean Preserving Perturbation	In this technique the Euclidean distance is preserved after perturbation. Hence the data perturbed by this technique can be used in various data mining applications.	The attacker can get the original data after applying attack techniques. Hence the privacy of original data is vulnerable.
Projection Based Perturbation	The random projection perturbation approach project the data onto a lower dimensional random space and can dramatically change its original form while preserving much of its distance related characteristics	This technique can be used in various data mining application but with little loss of accuracy.
CAMDP(Combination of additive and Multiplicative Data Perturbation	Data owners can share their data with data miners to find accurate clusters without any concern about violating data privacy.	-

Table 2

ISSN: 2278 – 1323

9. Conclusion:

With the development of data analysis and processing technique, the privacy disclosure problem about individual or company is inevitably exposed when releasing or sharing data to mine useful decision information and knowledge, then give the birth to the research field on privacy preserving data mining. In this research paper we have presented a comparative study of Multiplicative Data Perturbation Techniques. Traditional Multiplicative Data Perturbation approach distort each data element independently, therefore Euclidean Distance among data records are usually not preserved, and the perturbed data can not be used for many data mining applications. This 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 this perturbation scheme is questionable. The Distance Preserving data perturbation approach is quite useful as it allows many interesting data mining algorithms to be applied directly to the perturbed data and produce an error-free result but cannot effectively protect data privacy. The attacker can get the original data after applying attack techniques. Hence the privacy of original data is vulnerable. The random projection perturbation approach project the data onto a lower dimensional random space and can dramatically change its original form while preserving much of its distance related characteristics but with little loss of accuracy. CAMDP (Combination of Additive and Multiplicative Data Perturbation) Technique for Privacy Preserving Data Mining can be applied for several categories of popular data mining models with better utility preservation and privacy preservation. So this technique can be used in various Data Mining applications with accurate result and provide robust privacy preservation as compare to previous technique.

10. Reference

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

[2] B. Pandya,U.K.Singh, K Bunkar and K. Dixit, "An Overview of Traditional Multiplicative Data Perturbation" International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 2, Issue. 3 pp 424-428, 2012.

ISSN: 2278 – 1323

[3] B. Pandya,U.K.Singh and K. Dixit, "Effectiveness of Multiplicative Data Perturbation for Perturbation for Privacy Preserving Data Mining" International Journal of Advanced Research in Computer Science, Vol. 5, No. 6 pp 112-115,2014.

[4] B. Pandya,U.K.Singh and K. Dixit, "Performance of Multiplicative Data Perturbation for Perturbation for Privacy Preserving Data Mining" International Journal for Research in Applied Science and Engineering Technology, Vol. 2, Issue VII, 2014.

[5] B. Pandya,U.K.Singh and K. Dixit, "An Analysis of Euclidean Distance Presrving Perturbation for Privacy Preserving Data Mining" International Journal for Research in Applied Science and Engineering Technology, Vol. 2, Issue X, 2014.

[6] B. Pandya,U.K.Singh and K. Dixit, "Performance of Euclidean Distance Presrving Perturbation for K-Means Clustering" International Journal of Advanced Scientific and Technical Research, Vol. 5, Issue 4, pp 282-289, 2014.

[7] B. Pandya,U.K.Singh and K. Dixit, "Performance of Euclidean Distance Presrving Perturbation for K-Nearest Neighbour Classification" International Journal of Computer Application, Vol. 105, No. 2, pp 34-36, 2014.

[8] B. Pandya,U.K.Singh and K. Dixit, "A Study of Projection Based Multiplicative Data Perturbation for Privacy Preserving Data Mining" International Journal of Application or Innovation in Engineering and Management, Vol. 3, Issue 11, pp 180-182,2014.

[9] B. Pandya,U.K.Singh and K. Dixit, "An Analysis of Projection Based Multiplicative Data Perturbation for K-Means Clustering" International Journal of Computer Science and Information Technologies, Vol. 5, Issue. 6, pp 8067-8069, 2014.

[10] B. Pandya,U.K.Singh and K. Dixit, "An Evaluation of Projection Based Multiplicative Data Perturbation for K-Nearest Neighbour Classification" International Journal of Science and Research, Vol. 3, Issue. 12, pp 681-684, 2014.

ISSN: 2278 – 1323