Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia Technology 4 (2012) 350 – 354

Procedia Technology

C3IT-2012

Non-Additive Random Data Perturbation for Real World Data

Geetha Mary A[a] , N.Ch.S.N.Iyengar[a]

[a]*School of Computing Science and Engineering, VIT University, Vellore, Tamil Nadu – 632014, India*

Abstract

Privacy preservation is the major concern while real datasets are handled. A specific topic- privacy preserving data mining (PPDM), completely deals with data modification but also limits rule loss. Data perturbation is one of the PPDM techniques, which mostly deals with numerical data and concentrates on the statistical analysis of the data. Perturbation is of two types, additive perturbation and multiplicative perturbation, where generated random data is either added or multiplied with the data, which results in a random modified data. In this paper we have proposed a model in which the perturbation is done by randomization, where the data is generated in intervals based on the level of privacy specified by each customer. Our model is proved by applying classification algorithm on the perturbed data set and the accuracy is still maintained the same.

Keywords: Perturbation ; Random data ; Classification

1. Introduction

In Health Insurance Portability and Accountability Act (HIPAA) [1] , signed in 1996 by President of United States, security rule is updated on 2006, states specifically about technical safe guard of patient details. The Patient Safety and Quality Improvement Act of 2005 (PSQIA) Patient Safety Rule specifies about confidential data but allows minimum disclosure. 8.3 and 10 million people were affected by security breaches on March and April of 2011. Almost 3 months once, privacy attacks happen and personal health information gets stolen. Nowadays patient details like transcription information are outsourced from USA to many of the countries like India, while doing so patient details need to be safeguarded.

Privacy Preserving Data Mining (PPDM) is the main focus of scientists due to privacy concern of people. Starting from health care, banking, customer details, all data concerned with an individual person or a company has level of disclosure. Nowadays Health care information of patient's faces major attack, but this information needs to be analyzed for research purposes. To balance between the disclosure of data and research, PPDM techniques plays a major role. Zhang[2] states that PPDM process is divided into three tiers, Data providers tier – Data collection takes place, Data warehouse tier

2212-0173 © 2012 Published by Elsevier Ltd.
doi:10.1016/j.protcy.2012.05.053

– Data is converted into OLAP for easier processed data like aggregates, sum, average etc, the top tier – Data Mining server – Analysis is done according to the requirement. The main concern always is in the Data Provider tier where Collection of data happens. In the collected data though the key fields like patient number, Social security number, name are discarded and given for analysis are also prone to identify individual persons by record linkage which is discussed primarily with many methods in [3]. Again, these disclosures are handled by different methods like k-anonymity, l-diversity and t-closeness. T-closeness is the improved version of k-anonymity and l-diversity. These methods leads to generalization and suppression of attribute which leads to major loss of data.

Perturbation is also a major technique followed in PPDM introduced by Agarwal et al [4] in 2000. Random noise is added, now the noisy data along with the distribution is given to the data miner, who reconstructs the data for analysis. Reconstruction algorithm should be effective in such a way that loss of accuracy is as low as possible. Perturbation is mostly implemented by adding random data – additive data perturbation and multiplying with a matrix – multiplicative data perturbation. Perturbation are mostly implemented by two phase [4, 5, 6, 7], first adding noise to original data, then check for distribution, i.e. maintaining mean of zero and allowed variance.

Each individual (patient or employee) may go for different level of privacy which was not considered so far, but Liu et al. [8] discussed about two phase perturbation where random data is added with original data, then again random data is generated within the desired interval. Interval is decided by the level of security selected by each individual. Then they have gone for a reconstruction mechanism for getting back the original data with slight deviation.

Islam et al. [9] considers both numerical and categorical data, uses a clustering method to cluster the similar data and replace them by values generated by any of the perturbation method.

2. Non-Additive data perturbation model

We have proposed a model, which gets the level of privacy from user, where the level is not given as discrete set but the user specifies the value according to his demand of security. Random number is generated as per the demand of the owner. The random number is generated as per the pseudo code specified below. The pseudo code also maintains the sum since the deviation between the generated and the original data is also considered.

Pseudocode:

```
1.   set deviation to 0
2.   for each row in the table, Get the row
3.   set PrivacyLevel = valueOfColumn('level') + 1;
4.   set Data = valueOfColumn('OriginalData'); //Data to be perturbed
5.   set LowerLimit = Data - (Data Mod PrivacyLevel);
6.   set UpperLimit = LowerLimit + PrivacyLevel - 1;
7.   If (Deviation < -3)
7.1. newData = rand(LowerLimit, Data);
8.   else If(Deviation > 3)
8.1. newData = rand(Data, UpperLimit);
9.   Else
9.1. newData = rand(LowerLimit, UpperLimit);
10.  Deviation = Deviation + (Data - Newdata);
11.  Update the Column('Originaldata'), SetValue = NewData;
12.  Goto line 2
```

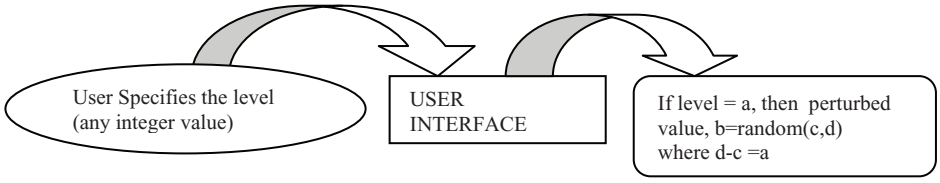

Fig.1. Level to Perturbed value mapping

3. Analysis

Real world data set from UCI repository is used for analysis of the method. The datasets were selected with different attribute types and of varying instance and attribute sizes. Classification algorithm – c5.0 in Clementine11.1 software is used for analysis. Then random number were generated according to the proposed method and again classification is done considering 50% of the data set as training dataset and other 50% for testing. After completion of the classification, accuracy is checked for each dataset, the accuracy is maintained the same for medium and small data set, while a little degradation happened for a large dataset. Information about dataset and the analysis is discussed in detail in this session. The data sets were taken which contradicts each other with number, type and size of attributes and number of instances. These data sets were selected to show that the algorithm functions similar with varied situations. Two measures were considered loss due to privacy and information loss which are discussed in the later of the session.

Privacy setting level is not constrained to a set of value, rather the data owner can specify the need of privacy for his data, and some may choose a high value so that the random data generated is a way more distant from the original one. The different level of privacy considered for each dataset is specified in Table.2.

Table 1. Over all dataset description

Dataset	Number of instances	Number of attributes	Attribute characteristics	Dataset characteristics
Credit approval dataset	690	15	Categorical, real, integer	multivariate
Statlog heart dataset	270	13	Categorical, real	multivariate
Shuttle dataset	58000	9	integer	multivariate

Table.2. Privacy level distribution according to dataset

Dataset	Minimum privacy level	Maximum privacy level	Privacy level type
Credit approval dataset	0	67	Continuous
Statlog heart dataset	3	7	set
Shuttle dataset	1	4	set

We have used c5.0, a classification algorithm to check the effect of perturbed data on the data mining algorithm. Clementine data mining tool is used to perform classification analysis, c5.0 algorithm is used. Original data is first given as a input to the algorithm, 50% of the data is taken as training data and the rest 50% is used for testing of the data. All the three data sets were used and accuracy is checked for each classified datasets. Though there are differences in the classification trees of original and perturbed dataset, the accuracy of the classification is maintained. Though high privacy level is

selected in credit approval data set which leads to highly dispersed values from the original data, the data classification is not affected by the level of the privacy considered since the classification done on the perturbed data has also lead to similar accuracy level.

50% of the data is taken for training and other 50% is taken for testing, while testing the data accuracy of the correct classification of original data and the perturbed data is specified in Fig.2.

In [10] Sumit has used various measures for checking privacy disclosure and loss due to perturbation. To measure privacy disclosure, Average Squared Distance is used.

$$ASD = \frac{1}{n}\sum_{i=1}^{n}(y_i - x_i)^2 \quad (1)$$

Where, x_i = Original Value at position i

y_i = Perturbed Value at position i

Loss due to privacy is measured by,

1. Bias in mean (BIM) as

$$BIM = \frac{|\bar{y} - \bar{x}|}{\bar{x}} \quad (2)$$

where $\bar{y}$ = *Mean of perturbed values* and $\bar{x}$ = *Mean of original values*

2. Bias in Standard Deviation (BIS) as

$$BIS = \frac{SDY - SDX}{SDX} \quad (3)$$

where SDY = Standard Deviation of Perturbed Values

SDX = Standard Deviation of Original Values

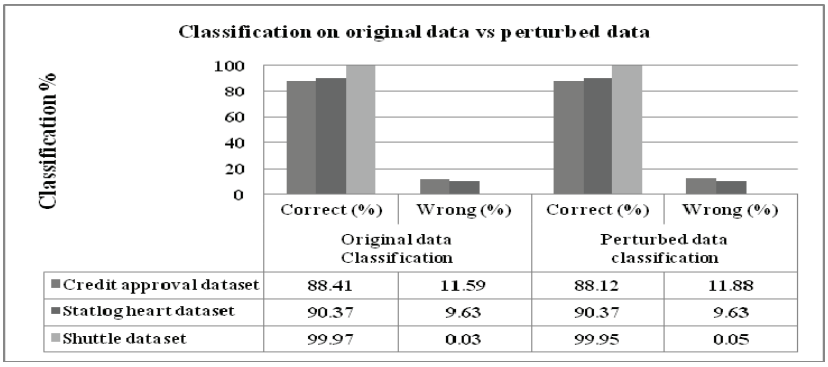

Fig.2. Correct Classification on original data vs perturbed data

Table.3. Measures of Privacy disclosure and Loss due to perturbation

Dataset	Mean		Standard Deviation		Average Squared Distance	Bias in Mean	Bias in Standard Deviation
	Original	Perturbed	Original	Perturbed			

Credit approval dataset	180.5479	180.7844	173.8442	173.7295	1.1536	0.0013	-6.5978×10^{-4}
Statlog heart dataset	249.6593	249.6502	51.5904	51.5747	3.9446	-3.6450×10^{-5}	-3.0432×10^{-4}
Shuttle dataset	48.2039	48.3223	12.1943	12.0915	0.4501	0.0025	-0.00843

4. Conclusion

From the experimental analysis we could understand that the accuracy of data is independent of the size of the dataset since the shuttle dataset's accuracy is also maintained similar, before and after perturation when compared with that of statlog heart dataset and credit approval dataset. This method need not go for reconstruction of the values, when comapared to other methods[5,8], since the values are more or less maintained similar which is proved by measures of loss due to perturbation and privacy disclosure.

References

1. http://www.hipaa.com/
2. Nan Zhang, Wei Zhao, "Privacy-Preserving Data Mining Systems," IEEE - Computer, vol. 40, no. 4, p no: 52-58, Apr. 2007
3. Ninghui Li, Tiancheng Li, Venkatasubramanian S, "t-Closeness: Privacy Beyond k-Anonymity and l-Diversity",2007 IEEE 23rd International Conference on Data Engineering,April 2007,P.no:106–115
4. R. Agrawal, R. Srikant, "Privacy-preserving data mining" ,SIGMOD Conference-2000, p.no: 439-450.
5. Kun Liu, Hillol Kargupta, Senior Member, IEEE, and Jessica Ryan, "Random Projection-Based Multiplicative Data Perturbation for Privacy Preserving Distributed Data Mining", IEEE Transactions On Knowledge And Data Engineering, Vol. 18, No. 1, P.no: 92-106, January 2006
6. D. Agrawal, C. C. Aggarwal,"On the design and quantification of privacy preserving data mining algorithms", In Proceedings of the Twentieth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems. Santa Barbara, California, USA. May 21-23 2001, ACM.
7. H. Kargupta, S. Datta, Q. Wang, K. Sivakumar," On the privacy preserving properties of random data perturbation techniques",ICDM, p.no: 99-106, IEEE Computer Society, 2003.
8. L. Liu, M. Kantarcioglu, B.Thuraisingham, "The applicability of the perturbation based privacy preserving data mining for real-world data", Data & Knowledge Engineering ,Vol.65, Issue:1,P.no: 5-21 , April 2008
9. Islam, M. Z., & Brankovic, L. (2011). Privacy preserving data mining: A noise addition framework using a novel clustering technique. Knowledge-Based Systems, 24(8), 1214-1223
10. Li, X.-B, Sumit Sarkar, "A Tree-Based Data Perturbation Approach for Privacy-Preserving Data Mining", Knowledge and Data Engineering, Volume: 18, Issue:9, P.no:1278 - 1283, September 2006.