Global Journal of Pure and Applied Mathematics.
ISSN 0973-1768 Volume 11, Number 3 (2015), pp. 1635-1639

http://www.ripublication.com

A Random Noise Based Perturbation Approach For Achieving Privacy In Data Mining

G.Manikandan[1], N.Sairam[2], N.Abitha[3], G.Sarada[3], R.Pranesh[3], M.Vigneshwaran[3]

[1]Assistant Professor, School of Computing, SASTRA University, Thanjavur, India
[2]Associate Dean / IT, School of Computing, SASTRA University, Thanjavur, India
[3] Student / B.Tech ICT,School of Computing, SASTRA University, Thanjavur, India
[1]E-mail:manikandan@it.sastra.edu

Abstract

Data mining refers to extracting appealing patterns or facts from enormous quantity of data. This data usually contains large amounts of secretive and susceptible data. One of the biggest concerns in data mining is privacy preservation. In this paper, the privacy issue in data mining is addressed by a new privacy preserving data mining method that utilizes a random noise for perturbation. A random value when added to the original data set results in a new customized data set. The customized data set has precisely the similar dimension as the original data set. Experiments were performed on real life dataset and the outcome demonstrates that random noise perturbation proficiently preserves the intended information and also maintains the accuracy of data mining algorithm after data deformation.

Keywords: Data Privacy, Pseudo random generator, Clustering, Data Perturbation, Data Accuracy

Introduction

With the rapid growth of technology there is a massive growth in the amount of data collection. The data collection includes shopping habits of customers, medical history of patients, credit card usage records etc. Data is an imperative positive feature to business organizations for decision making process. The investigation of data opens new intimidation to seclusion and independence of the human being if not done properly. The risk to privacy becomes authentic since data mining techniques are capable of obtaining exceedingly responsive information from uncategorized data. [2]

Data mining is a promising field which connects diverse areas like statistics, database and artificial intelligence [5].In recent days organizations are tremendously reliant on data mining outcome to offer better quality of service, achieving enhanced

profit, and improved decision-making[1][9]. Data mining seeks to ascertain unrecognized relations between data items in an existing database. It is the practice of extracting legitimate, formerly unseen or unidentified, intelligible information from huge databases.

There is a bigger concern on how the private information can be confined while performing data mining operation. This problem is addressed by Privacy preserving data mining algorithms. At present privacy preserving data mining turns out to be an eloquent topic for research. There are many methods proposed in the literature for privacy preserving data mining and it can be classified into two broad categories. The first category consists of modifying the data mining algorithms by mining the datasets without being aware of the precise values of the data. The second category consists of changing the values of the original dataset so as to ensure its confidentiality. Privacy preserving data mining gives legitimate data mining results and also guarantees privacy for sensitive data.

The constraint in most of the traditional techniques is that the modified data predominantly depends on the single noise value that is being added as shown in Table 1. From the Table 1 we infer that when a single value noise 5 is added to the original data say 50 at different occurrences, a modified value of 55 occurs at all instances of 50.

Table 1: Perturbation using single noise value

Original Data	50	31	32	21	33	50	31	32	21	33
Noise	5	5	5	5	5	5	5	5	5	5
Modified Data	55	36	37	26	38	55	36	37	26	38

Literature Survey

To provide healthcare privacy in medical cyber-physical systems, in [3] two additive perturbation algorithms have been proposed namely RDD and RACC. To enhance the anti-attack capability, matrix multiplicative perturbation is combined with those two additive perturbation algorithms. Matrix multiplicative perturbation preserves Euclidean distance of perturbed data with either small or no error.

An advanced algorithm for finding global frequent item sets with negligible communication overhead and time complexity was proposed by Rajalakshmi and Purusothaman [8].

In [4] the combination of K-anonymity with randomization makes it difficult for the attacker to identify background and homogeneity attack. Private data is protected with better accuracy and gives no loss of information which increases data utility.

A genetic algorithm for data mining privacy preservation using K-anonymity method with differing k-levels was used to demonstrate that increase in k-anonymity levels ensures a decrease in classifier performance within acceptable levels [7].

In [6] a combination of fuzzy membership functions, fuzzy data transformation approach and Random Rotation Perturbation (RRP) are used to transform the original dataset for privacy preserving clustering in centralized database environment.

Proposed System

Our approach aims at privacy preservation of sensitive attributes by using a pseudorandom number generator named multiply with carry. The uniqueness of this approach is that the data is modified by adding the unique numbers generated by pseudorandom number generators. The modified data is now validated by K-means clustering algorithm. Then misclassification error is calculated and tabulated. From our experimental results it is evident that the original data cannot be inferred from this adapted data.

Multiply with carry is a method for sequence generation of pseudorandom integers depending on the random seed and the carry value. Later carry values are the previous quotient values. The main expediency of this method is that it does simple computer integer arithmetic and results in rapid pseudorandom number generation.

Multiply with carry is based on the recurrence:

$x_i = ((s*x_{i-1}) + c_{i-1}) \% m$

$c_i = ((s*x_{i-1}) + c_{i-1}) / m$

Where

x_0 is the seed or start value

s is the multiplier

c is the carry

m is the modulus

The pseudorandom number generated by the above method is mapped to a range and then added with the original data to generate the sanitized data. At this point first the original data is clustered and then the modified data is clustered using K-means algorithm. Now the modified cluster is compared with the original clusters for misclassification error using the following formula and is tabulated in Table 2 for various instances.

$M = 1/n \Sigma_k | CLUSTER (D) | - | CLUSTER (D') |$

Where

M is the misclassification error

D represents original data/cluster

D' represents sanitised data/cluster

Table 2: Misclassification Error

K Value	2	3	4
M for 25 instances	0	0.16	1.12
M for 50 instances	0.6	0.4	0.96
M for 75 instances	0	0.58	1.01

Experimental Results

The data set used in this work is the Pima Indians diabetic data set which belongs to National Institute of Diabetes and Digestive and Kidney Diseases available on UCI Machine Learning Repository [10]. It has 768 records where each record contains information about at least 21 year old female person of Pima Indian heritage. Age attribute is used for clustering purpose. The above proposed approaches have been implemented using JAVA Programming language and the resulting observations are tested in Intel core i5 processor with 4GB RAM and Windows 8 operating system. We have tabulated few of the results below for comparison. From our experimental results it is evident that the original data cannot be inferred from the modified data. From the Table 3 we infer that when a random noise is added to the original data say 50 at different occurrences, a different modified value occurs is obtained.

Table 3: Perturbation Using Random Noise Value

Original Data	50	31	32	21	33	50	31	32	21	33
Random Noise	1	2	5	14	13	12	8	9	11	5
Modified Data	51	33	37	35	49	62	39	41	32	38

Conclusion

In this paper we have put forward a new approach for achieving privacy in data mining using a random noise based perturbation approach. From the experimental results it has been apparent that this system overcomes the restrictions of the traditional methods which utilize a single noise value for perturbation. The uniqueness of this approach is that the data is modified by adding the unique numbers generated by the pseudorandom number generator. In the future this work can be extended by using other pseudo random number generators.

References

[1] G.K.Gupta, Introduction to Data Mining with Case Studies, Prentice Hall of India, 2008.

[2] Jiawei Han and Micheline Kamber, Data Mining: Concepts and Techniques, Morgan Kauffman Publishers, 2006.

[3] Liu L, Yang K et al., Using noise addition method based on pre-mining to protect health care privacy, Journal of Control Engineering and Applied Informatics,14(2012), 58–64.

[4] Manish Sharma, Atul Chaudhary, Manish Mathuria, Shalini Chaudhary and Santosh Kumar, An Efficient Approach for Privacy Preserving in Data

Mining, International Conference on Signal Propagation and Computer Technology, 2014, 244 – 249.

[5] Margaret.H.Dunham, Data Mining: Introductory and advanced topics, Pearson Education, 2003.

[6] M. Nagalakshmi and K Sandhya Rani, Privacy Preserving Clustering Based on Fuzzy Data Transformation Methods, International Journal of Advanced Research in Computer Science and Software Engineering, 8(2013), 1027-1033.

[7] J. Paranthaman, Dr. T Aruldoss Albert Victoire, Genetic Optimization For Privacy Preserving in Data Mining, Journal of Theoretical and Applied Information Technology,57 (2013), 517-522.

[8] Rajalakshmi M and Purusothaman T , Privacy preserving distributed data mining using randomized site selection, European Journal Of Scientific Research, 64(2011), 610-624.

[9] K.P.Soman, Shyam Diwakar and V.Ajay, Insight into Data Mining: Theory and Practice, Prentice Hall of India, 2006.

[10] UCI Data Repository http://archive.ics.uci.edu/ml/datasets.html