VOL. 11, NO. 13, JULY 2016 ISSN 1819-6608
ARPN Journal of Engineering and Applied Sciences
©2006-2016 Asian Research Publishing Network (ARPN). All rights reserved.
www.arpnjournals.com

A RULE BASED APPROACH FOR ENSURING PRIVACY IN DATA MINING

G. Manikandan[1], N. Sairam[2], M. SathyaPriya[1], Sri RadhaMadhuri[1], V. Harish[1] and Nooka Saikumar[1]
[1]Department of Information and Communication Technology, SASTRA University, Thanjavur, India
[2]Department of Information Technology, SASTRA University, Thanjavur, India
E-Mail: manikandan@it.sastra.edu

ABSTRACT

Data mining or the knowledge discovery process is examining data in several views and sum up it into practicable data that can be utilized to raise gross, cuts costs, or both. The ultimate goal of data mining is to attain the patterns and the trends that are concealed in the stimulant data. For this purpose it utilizes the large data sets that are available as part of data publishing. Data Privacy has to be ensured while extracting information from large data sets. In recent times achieving privacy in data mining plays a vital role. Data Perturbation can be used for generating the sanitized data from the original data. Most of the existing systems in the literature use a single noise value for the perturbation process. Here we are adding noise by generating If-then-rules for various attributes by grouping them into various classes for attaining privacy.

Keywords: data privacy, clustering, misclassification, accuracy.

1. INTRODUCTION

With the escalation of data investigation and dispensation technique, various private organizations, and government agencies are increasingly publishing micro data for data mining purposes. While the unconstrained data sets provide precious information to researchers, at the same time they may also contain some sensitive information about individuals whose confidentiality may be at risk[1-2].

Data mining is a practice of analyzing data from diverse point of views to figure out useful information, which can be utilized in various domains like bioinformatics, weather forecasting, financial forecasting and telecommunication [3-4].

While the reimbursement realized by making statistics available for the use of data mining are enormous, latest results have revealed that privacy is a noteworthy requirement that must be considered all along with the mining results. For moral and technological reasons, data owners who afford their data for data mining purposes are apprehensive that their data will be used for inferring some private information.

Privacy preservation plays a vital role in data mining since distribution of information is a prime prerequisite for the success of data mining process. It can construe some unknown and volatile facts which may provide high interestingness to the data miners. These algorithms are intended to afford data mining results that conceals access to user identity [5-6].

The techniques available for privacy preservation of data are classified into: Data Perturbation, Cryptographic technique and Query auditing. Generally in data perturbation technique, the sanitized data is generated by adding single value noise to the original data. Due to this, privacy is not ensured completely, as the data must be exposed easily. Cryptographic technique deals with the generation of keys for encryption of the plain text and decryption of cipher text. The key is the noise by using which the sanitized data is generated. As part of cryptographic process, the cipher text generated ensures data security but compromises on the utility of data. If the key value is exposed, then it can be deciphered easily. This leads to data breach.

One way of query inspection involves looking into the queries that were responded previously and ascertaining if the solutions to those queries were exploited by someone to acquire admission to the secret selective information banned by revealing policies.

In the second approach of query inspection, the online editor limits the queries of exploiter that could despoil privateness. In the third approach, noise is added up to the data that alters the results to the queries thereby precluding the exploiter from unauthorized admission.

The rest of the paper is organized as follows: Section 2 gives us the literature review of the techniques used in our proposed system. Section 3 provides an insight to the proposed system. The formula used for computing misclassification error of our experiment is given in Section 4. Section 5 is a brief experimental analysis of the approach used in the experiment. Section 6 gives the conclusion to the approach defined in this paper.

2. LITERATURE REVIEW

In [7], for providing healthcare privacy in medical cyber-physical systems two additive perturbation algorithms have been proposed namely RDD and RACC. These two algorithms not only make the reconstruction with high computation unnecessary and keep the mining algorithm unmodified but also have higher accuracy. 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.

Min max normalization based transformation approach was applied in [8] for ensuring data secrecy. This approach transforms the original data to privacy-preserved data by maintaining the inter-relative distance among the data .The efficiency was calculated by the use of K-means algorithm.

©2006-2016 Asian Research Publishing Network (ARPN). All rights reserved.

An algorithm for identifying global frequent item sets with minimal communication overhead and time complexity was proposed by Rajalakshmi and Purusothaman [9].

In [10] the combination of K-anonymity with randomization makes it arduous for the attacker to find background and homogeneity attack.

Karthikeyan *et al*. [11] proposed a method to achieve privacy using fuzzy logic. For demonstrating their approach they have used s-shaped membership function for data sanitization. It is also proved that the clustering algorithm takes lesser number of iteration for clustering process with the modified data.

In [12] a genetic algorithm for privacy preservation utilizing K-anonymity method with varying k-levels was used to demonstrate that increase in k-anonymity levels ensures a decrease in classifier performance within acceptable levels.

In [13] fuzzy based membership functions like Z-shaped fuzzy membership function, Triangular fuzzy membership function, Gaussian fuzzy membership function and combination of fuzzy data transformation approach and Random Rotation Perturbation (RRP) are used to transform the original dataset for secrecy ensuring clustering in centralized database environment.

In [14] a survey on the merits and demerits of secrecy ensuring data mining approaches and techniques are discussed.

3. PROPOSED SYSTEM

The Proposed system consists of a three step process. We have formulated various If-then-rules by using attributes such as Age, Gender and Income in the first step. In the Second step we assign a class for all the rules then a noise value is assigned for each class as part of the third step. The sensitive data is modified using the noise value of the class. For example a sample rule may be as follows

IF AGE IS BETWEEN 15-30 AND GENDER = MALE AND INCOME <= 50K.

The records satisfying the above rule will be placed in class 1. For sanitization the noise 7 is added to the age attribute of all the records of class 1.

The other rules, the class and the corresponding noise are shown in Table-1.

Table-1. If-Then rules

Age	Gender	Income	Noise	Class
15-30	M	<=50K	7	1
15-30	F	<=50K	8	2
15-30	M	>50K	6	3
15-30	F	>50K	9	4
31-59	M	<=50K	3	5
31-59	F	<=50K	5	6
31-59	M	>50K	4	7
31-59	F	>50K	2	8
>=60	M	<=50K	4	9
>=60	F	<=50K	3	10
>=60	M	>50K	2	11
>=60	F	>50K	1	12

©2006-2016 Asian Research Publishing Network (ARPN). All rights reserved.

Table-2. Example.

Age	Gender	Income	Noise	Class	Modified age
25	0	0	7	1	32
19	1	0	8	2	21
71	0	0	4	9	75
24	1	0	8	2	32
37	0	1	4	7	41
42	0	1	4	7	46
49	1	0	5	6	54
52	0	1	4	7	56
31	1	1	2	8	33
39	0	0	3	5	42
61	1	1	1	12	62
62	1	0	3	10	65
30	0	1	6	3	36
24	1	1	9	4	33
76	0	1	2	11	78

Table-3. Original dataset clustering.

Original dataset	Clusters	
	Cluster 1	Cluster 2
25,19,71,24,37,42,49,52,31,39,61,62,30,24,76	71,49,52,61,62,76	25,19,24,37,42,31,39,30,24

Table-4. Sanitized dataset clustering.

Sanitized dataset	Clusters	
	Cluster 1	Cluster 2
32,21,75,32,41,46,54,56,33,42,62,65,36,33,78	75,54,56,62,65,78	32,21,32,41,46,33,42,36,33

4. MISCLASSIFICATION ERROR

The original data is first clustered using K-means algorithm. Then the sanitized data is clustered using K-means algorithm. Now the sanitized clusters are compared with the original clusters for misclassification using the following formula:

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

Where

M is the misclassification error
r represents original data/cluster
r' represents sanitized data/cluster

5. EXPERIMENTAL RESULTS

In this paper, we have considered various attributes such as Age, Gender and Income from the adult data set available in the UCI repository. The data set consists of about 32561 records, out of which we have taken close to 10 different records and demonstrated the proposed entity. 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 found out the misclassification error for the original and modified data set for the age attribute which has turned out to be close to 0.124811. From our observations, it has been proved that the original data can't be recovered from the altered data by homogeneity attack and background knowledge attack.

6. CONCLUSIONS

Thus the original data becomes difficult to be known to the people who use these published data .From the observational analysis, it is elucidate that these techniques gets over the restrictions of the conventional techniques. The unicity of this technique is that the data is altered by adding the unique noise generated using various

©2006-2016 Asian Research Publishing Network (ARPN). All rights reserved.

rules. The limitation is that these approaches can be used only for numeric data. This work can be extended in future for categorical data also. These approaches can be used for achieving privacy in data publishing in several organizations like banks, hospitals etc.

REFERENCES

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

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

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

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

[5] Manish Sharma*et al.*2014. An Efficient Approach for Privacy Preserving in Data Mining, International Conference on Signal Propagation and Computer Technology. pp. 244-249.

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

[7] Benjamin C.M. Fung, Ke Wang, Ada Wai-Chee Fu; Philip S. Yu. 2010. Introduction to Privacy-Preserving Data Publishing: Concepts and Techniques, Chapman and Hall.

[8] Jaideep Vaidya, Christopher W. Clifton, Yu Michael Zhu.2005. Privacy Preserving Data Mining, Springer.

[9] Karthikeyan B *et al.* 2011 A fuzzy based approach for privacy preserving clustering. Journal of Theoretical and applied information Technology. 32(2): 118-122.

[10] Liu L *et al.* 2012. Using noise addition method based on pre-mining to protect health care privacy, Journal of Control Engineering and Applied Informatics. 14(2): 58-64.

[11] Manikandan G *et al*. 2013. Achieving privacy in data mining using normalization. Indian Journal of Science and Technology. 6(4): 4268-4272.

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

[13] J. Paranthaman and Dr. T Aruldoss Albert Victoire. 2013. Genetic Optimization for Privacy Preserving in Data Mining. Journal of Theoretical and Applied Information Technology. 57(3): 517-522.

[14] Rajalakshmi M and Purusothaman T. 2011. Privacy preserving distributed data mining using randomized site selection. European Journal of Scientific Research. 64(2): 610-624.

[15] W.T.Chembian and Dr. J.Janet. 2010. A Survey on Privacy Preserving Data Mining Approaches and Techniques. Proceedings of the International Conference on Information Science and Applications ICISA.