International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169
Volume: 4 Issue: 11 316 - 318

A Study on privacy for Sensitive Data by DM algorithms

Dr. Halkar Rachappa
Asst. Prof. & Head Dept. of Computer Science
SGRCM Govt. Commerce & Management College, BALLARI
(Karnataka)

***Abstract*—** Whenever big data term is concerned the most important concern is privacy of data. One of the most common methods use random permutation techniques to mask the data, for preserving the privacy of sensitive data. Randomize response (RR) techniques were developed for the purpose of protecting surveys privacy and avoiding biased answers. The proposed work is to enhance the privacy level in RR technique using four group schemes. First according to the algorithm random attributes a, b, c, d were considered, then the randomization have been performed on every dataset according to the values of theta. Then ID3 and CART algorithm are applied on the randomized data.

Keywords: *Data Mining, PPDM, Privacy, Four Groups, Multi Groups, Privacy Issue, ID3, CART.*

*****

1. INTRODUCTION

As it is known that business data and business information both are very important terms in data mining. Day by day these are getting new steps has build upon previous one.

A. Process of Data Mining

There are basic for steps composed in Data Mining [6]. This consists of transforming the already summarized data found in a data warehouse into information producing useful results through:

- Selection of Data
- Transformation of Data
- Extracting the Data
- Results Interpretation

For selecting the data first of all data is gathered for analysis. As shown in Figure 2, the relevant information will be extracted by data-mining tool from the data warehouse environment

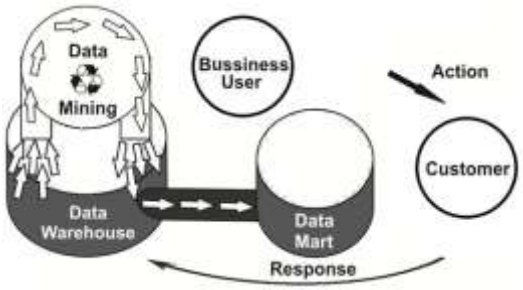

Fig. 1. Data Mining Process

According to [7] the following terms explain data mining:

- Data, Information, and Knowledge
- Need of data mining
- Working of data mining

2. LITERATURE SURVEY

This survey included various types of techniques used in data mining. Mainly the techniques are divided in two sections:

- Classical Techniques: This technique includes, Statistics Neighborhoods and Clustering
- Next Generation Techniques: This technique includes Trees, Networks and Rules [8]

A. Privacy Preserving Data Mining (PPDM)

PPDM is a novel research direction in data mining and statistical databases, where data mining algorithms are analyzed for the side-effects they incur in data privacy. The main consideration in privacy preserving data mining is twofold. First, sensitive raw data like identifiers, names, addresses and the like should be modified or trimmed out from the original database, in order for the recipient of the data not to be able to compromise another person's privacy. Second, sensitive knowledge which can be mined from a database by using data mining algorithms should also be excluded, because such knowledge can equally well compromise data privacy, as it will be indicated.

B. PROPOSED APPROACH

This work has proposed an approach for privacy preserving data mining using randomized response technique. This work uses ID3 and CART algorithm to enhance the privacy of the secret data. The problem with the previous work for three groups of data sets using ID3 algorithm was that it was not checking the group performance at every step and the privacy level was not very high [11].

A. *The Basic idea of ID3 Algorithm*

ID3 algorithm uses information entropy theory to select attribute values with maximum information gain in the current sample sets as the test attribute. The division of the sample sets is based on the value of the test properties, the numbers of test attributes decide the number of subsample sets; at the same time, new leaf nodes grow out of corresponding nodes of the sample set on the decision tree. Since simpler the decision tree structure the easier is to summarize the law of things in nature. The average paths from the non-leaf nodes to the descendant nodes are expected the shortest, i.e. the average depth of the generated decision tree should be minimum, which requires choosing a good decision on each node. Therefore, the algorithm selects the greatest attribute with the most information gain as a test property on each non-leaf node. ID3 algorithm divides sample sets into the training sample sets and the test sample sets. The decision tree construction produced on learning the training sample sets. After decision tree generation, the test sample sets are used to prune the tree off the non-eligible leaf nodes with inaccuracy of the inductive learning prediction [12].

B. *CART Algorithm*

Classification and Regression Trees is a classification method which uses historical data to construct so-called decision trees. Decision trees are then used to classify new data. In order to use CART the number of classes should be known a priori. Decision trees are represented by a set of questions which splits the learning sample into smaller and smaller parts. CART asks only yes/no questions. The process is then repeated for each of the resulting data fragments [13].

The basic idea of tree growing is to choose a split among all the possible splits at each node so that the resulting child nodes are the "purest". In this algorithm, only univariate splits are considered. A tree is grown starting from the root node by repeatedly using the following steps on each node.

Step 1: Find each predictor's best split.

Step 2: Find the node's best split.

Among the best splits found in step 1, choose the one that maximizes the splitting criterion.

Step 3: Split the node using its best split found in step 2 if the stopping rules are not satisfied.

C. *Randomized Response Techniques*

Randomized Response (RR) techniques were developed for the purpose of protecting survey's privacy and avoiding answer bias mainly. They were introduced by Warner (1965) [14] as a technique to estimate the percentage of people in a population U that has a stigmatizing attribute A. In such cases respondents may decide not to reply at all or to incorrectly answer defined the Warner's original method different RR procedure. The usual problem faced by researchers is to encourage participants to respond, and then to provide truthful response in surveys. The RR technique was designed to reduce both response bias and non-response bias, in surveys which ask sensitive questions. It uses probability theory to protect the privacy of an individual's response, and has been used successfully in several sensitive research areas, such as abortion, drugs and assault. The basic idea of RR is to scramble the data in such a way that the real status of the respondent cannot be identified [15

In Related-Question Model, instead of asking each respondent whether he/she has attribute 0, the interviewer asks each respondent two related questions, the answers to which are opposite to each other. For example, the questions could be like the following. If the statement is correct, the respondent answers "yes"; otherwise he/she answers "no". Similar as described in [16].

- I have the sensitive attribute A.
- I do not have the sensitive attribute A.

Respondents use a randomizing device to decide which question to answer, without letting the interviewer know which question is answered. The probability of choosing the first question is θ, and the probability of choosing the second question is 1- θ. Although the interviewer learns the responses (e.g., "yes" or "no"), he/she does not know which question was answered by the respondents.

To estimate the percentage of people who has the attribute A. The following equations can be used:

$$P^*(A = yes) = P(A = yes).\theta + \mathrm{P}(\mathrm{A} = \mathrm{no}).(1-\theta)$$
$$P^*(A = no) = P(A = no).\theta + \mathrm{P}(\mathrm{A} = \mathrm{yes}).(1-\theta) \quad (4)$$

- *Groups Scheme*

In the one-group scheme, all the attributes are put in the same group, and all the attributes are either reversed together or keeping the same values. In other words, when sending the private data to the central database, users either tell the truth about all their answers to the sensitive questions or tell the lie about all their answers. The probability for the first event is θ and the probability for the second event is $(1-\theta)$.. To simplify the presentation, P (001) is used to represent

$$\mathrm{P}(\mathrm{A}_1 = 1 \wedge \mathrm{A}_2 = 1 \wedge \mathrm{A}_3 = 0)$$

Because the contributions to P*(110) and P*(001) partially come from P (110) and partially come from P (001), the following equations can be used:

$$P^*(110) = P(110).\theta + P(001).(1-\theta)$$

$$P^*(001) = P(001).\theta + P(110).(1-\theta) \quad (5)$$

By solving the above equations, the P(110) provide the information needed to build a decision tree. The general

model for the one-group scheme is described in the following:

$$P^*(E) = P(E).\theta + P(\overline{E}).(1-\theta)$$
$$P^*(\overline{E}) = P(\overline{E}).\theta + P(E).(1-\theta) \quad (6)$$

Using the matrix form, let M1 denote the coefficiency matrix of the above equations, the Matrix

$$\begin{pmatrix} p^*(E) \\ P^*(\overline{E}) \end{pmatrix} = M_1 \begin{pmatrix} P(E) \\ P(\overline{E}) \end{pmatrix}, \text{ where } M_1 = \begin{bmatrix} \theta & (1-\theta) \\ 1-\theta & \theta \end{bmatrix} \quad (7)$$

Similarly we can calculate for two group, tree group and four groups.

Building Decision Trees

The decision tree is one of the classification methods. A decision tree is a class discriminator that recursively partitions the training set until each partition entirely or dominantly consists of examples from one class. Algorithm is described below where y represents the training samples and AL represents the attribute list:

ID3(S, AL)

Step 1. Create a node V.

Step 2. If S consists of samples with all the same class C then return V as a leaf node labeled with class C.

Step 3. If AL is empty, then return V as a leaf-node with the majority class in y.

Step 4. Select test attribute (TA) among the AL with the highest information gain.

Step 5. Label node V with TA.

Step 6. For each known value ai of TA

a) Grow a branch from node V for the condition TA=a_i

b) Let s_i be the set of samples in S for which TA=a_i.

c) If s_i empty then attach a leaf labeled with the majority class in S.

d) Else attach the node returned by ID3(s_i,AL-TA).

According to ID3 algorithm, each non-leaf node of the tree contains a splitting point, and the main task for building a decision tree is to identify an attribute for the splitting point based on the information gain. Information gain can be computed using entropy. In the following, it is assumed that there are m classes in the whole training data set. Entropy(S) and Gain defined as follows:

$$\text{Entropy}(s) = -\sum_{j=1}^{m} Qj(S) \log_2 Qj(s), \quad (16)$$

$$\text{Gain}(S, A) = \text{Entropy}(s) - \sum_{v \in A} \left(\frac{|Sv|}{|S|} \text{ Entropy } (Sv)\right), \quad (17)$$

C. Conclusion and Future Scope

This work gives a different approach for enhancing the privacy of the sensitive data in datamining. In this thesis, the four group randomized response technique is used in privacy preserving algorithm. To support the work CART and ID3 algorithm are used. In this experiment first applied randomized response techniques on one, two, three and four groups. The ID3 and CART algorithm are applied on the randomized data.

Reference

[1] Giudici, P, "Applied Data-Mining: Statistical Methods for Business and Industry." John Wiley and Sons (2003) West Sussex, England.

[2] American Association for Artificial Intelligence Advances in Knowledge Discovery and Data Mining. Press/ The MIT Press. 1996.

[3] Y. Ramamohan, K. Vasantharao, C. Kalyana Chakravarti, A.S.K.Ratnam, "A study of data mining tools in knowledge discovery process", IJSCE, Volume-2, Issue-3, July 2012.

[4] http://www.pilotsw.com/dmpaper/dmindex.htm; "An Introduction to Data Mining". Pilot Software Whitepaper. Pilot Software. 1998.

[5] "Data Mining: An Introduction", SPSS Whitepaper. SPSS. 2000.

[6] Walter Alberto, Data Mining Industry: Emerging Trends and New Opportunities: MIT, 2000 Springer book.

[7] G.Rama Krishna, G.V.Ajaresh, I.Jaya Kumar Naik, Parshu Ram Dhungyel, D.Karuna Prasad "A New Approach to Maintain Privacy And Accuracy In Classification Data Mining" IJCSET Volume 2, Issue 1, January 2012 Y.

[8] Gayatri Nayak, Swagatika Devi, "A Survey On Privacy Preserving Data Mining Approaches And Techniques", IJEST, Vol. 3 No. 3 March 2011.

[9] Malik, M.B.,"Privacy Preserving Data Mining Techniques: Current Scenario and Future Prospects", ICCCT, Nov. 2012

[10] Wu Xiaodan, Chu Chao-Hsien, Wang Yunfeng, Liu Fengli, Yue Dianmin, Privacy Preserving Data Mining Research: Current Status and Key Issues, Computional Science-ICCS 2007,4489(2007), 762-772.

[11] Carlos N. Bouza1, Carmelo Herrera, Pasha G. Mitra,"A Review Of Randomized Responses Procedures The Qualitative Variable Case", Revista Investigación Operacional VOL., 31, No. 3, 240-247 2010

[12] ftp://.ftp.ics.uci.edu/pub/machine-learning-databases.

[13] Raj Kumar, Dr. Rajesh Verma, "Classification Algorithms for Data Mining: A Survey", IJIET, Vol. 1 Issue 2 August 2012.

[14] Zhouxuan Teng, Wenliang Du,"A Hybrid Multi-Group Privacy-Preserving Approach for Building Decision Trees",

[15] Gerty J. L. M. Lensvelt-Mulders, Joop J. Hox And Peter G. M. Van Der Heijden

[16] Carlos N. Bouza1, Carmelo Herrera, Pasha G. Mitra,"A Review Of Randomized Responses Procedures The Qualitative Variable Case", Revista Investigación Operacional VOL., 31 , No. 3, 240-247 2010

[17] Zhijun Zhan, Wenliang Du, "Privacy-Preserving Data Mining Using Multi-Group Randomized Response Techniques" 2010.

[18] Monika Soni, Vishal Shrivastva, "Privacy Preserving Data Mining: Comparion of Three Groups and Four Groups Randomized Response Techniques", IJRITCC, Volume 1 Issue 7, July 2013.