Randomized Response Technique in Data Mining

Monika Soni
Arya College of Engineering and IT, Jaipur(Raj.)
12.monika@gmail.com

Vishal Shrivastva
Arya College of Engineering and IT, Jaipur(Raj.)
vishal500371@yahoo.co.in

Abstract - Data mining is a process in which data is collected from different sources and resume it in useful information. Data mining is also known as knowledge discovery in database (KDD).Privacy and accuracy are the important issues in data mining when data is shared. A fruitful direction for future data mining research will be the development of techniques that incorporate privacy concerns. Most of the methods use random permutation techniques to mask the data, for preserving the privacy of sensitive data.

Randomize response techniques were developed for the purpose of protecting surveys privacy and avoiding answers bias mainly. In RR technique it adds certain degree of randomness to the answer to prevent the data. The objective of this thesis 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 was applied on the randomized data. The result shows that by increasing the group, the privacy level will increase.

***Keywords*—Data Mining, Privacy Preserving, Randomized Response, Groups.**

*****

1. INTRODUCTION

Data Mining is a field of search and researches of data. Data Mining can be refers to as extracting the useful information from large amount of data. The goal of data mining is to improve the quality of the interaction between the organization and their customers.

According to Giudici [1], "data mining is the process of selection, exploration, and modeling of large quantities of data to discover regularities or relations that are at first unknown with the aim of obtaining clear and useful results for the owner of the database."

1.1 Privacy issues in data mining

It is well known that data mining is a powerful data analysis tool enabling discovery of useful patterns in several applications. The availability of large data warehouses and associated technologies reveal that the usefulness of data requires preservation of individual key attributes such as patient's condition information, customer preferences, personal background information etc. If we release the original data directly to the miner, it will inevitably produce private information of the customer. Therefore, how to do mining without sacrificing privacy is the main issue in data mining. Privacy preserving data mining (PPDM) deals with this issue. Currently, there are many approaches of privacy preserving data mining to transform the original data. The methods of privacy preserving data mining are evaluated based on applicability, privacy protection metric, the accuracy, computation, etc. [2]

1.2Privacy preserving data mining [3]

The problem of privacy preserving data mining has become more important in recent years because of the increasing ability to store personal data about users and the increasing sophistication of data mining algorithm to leverage this information. A number of techniques such as classification, k-anonymity, association rule mining, clustering have been suggested in recent years in order to perform privacy preserving data mining. Furthermore, the problem has been discussed in multiple communities such as the database community, the statistical disclosure control community and the cryptography community.

2. LITERATURE SURVEY

Literature survey included the various types of techniques used in data mining. Mainly we divide the techniques in two sections [4]:

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

2.1. Classification of Clustering Algorithms

Categorization of clustering algorithms is not easy. In reality, groups given below overlap. For convenience we provide a classification as given in [5]. We are considering mainly hierarchical and partitioning methods.

Hierarchical Methods

- Agglomerative Algorithms
- Divisive Algorithms

Partitioning Methods

- Relocation Algorithms
- Probabilistic Clustering
- K-medoids Methods
- K-means Methods
- Density-Based Algorithms
- Density-Based Connectivity Clustering
- Density Functions Clustering

2.2 CART -Classification and Regression Trees

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 CART algorithm will search for all possible variables and all possible values in order to find the best split – the question that splits the data into two parts with maximum homogeneity. The processes then repeated for each of the resulting data fragments.[6]CART can easily handle both numerical and categorical variables. Among other advantages of CART method is its robustness to outliers. Usually the splitting algorithm will isolate outliers in individual node or nodes. An important practical property of CART is that the structure of its classification or regression trees is invariant with respect to monotone transformations of independent variables. One can replace any variable with its logarithm or square root value, the structure of the tree will not change.[6]CART methodology consists of three parts:

1. Construction of maximum tree
2. Choice of the right tree size
3. Classification of new data using constructed tree

In building the CART tree each predictor is picked based on how well it teases apart the records with various predictions. For instance one measure that is used to determine whether a given split point for a give predictor is better than another is the entropy metric. The measure originated from the work done by Claude Shannon and Warren Weaver on information theory in 1949.They were concerned with how information could be efficiently communicated over telephone lines. Interestingly, their results also prove useful in creating decision trees.

2.3 Decision Trees and the ID3 Algorithm

A decision tree is a rooted tree containing nodes and edges. Each internal node is a test node and corresponds to an attribute; the edges leaving a node correspond to the possible values taken on by that attribute. For example, the attribute “Home-Owner” would have two edges leaving it, one for “Yes” and one for “No”. Finally, the leaves of the tree contain the expected class value for transactions matching the path from the root to that leaf.

Given a decision tree, one can predict the class of a new transaction t as follows. Let the attribute of a given node v (initially the root) be A, where A obtains possible values a1,….,am. Then, as described, the m edges leaving v are labeled a1,….,am respectively. If the value of A in t equals ai, we simply go to the son pointed to by ai. We then continue recursively until we reach a

leaf. The class found in the leaf is then assigned to the transaction.

We use the following notation:

R: the set of *attributes*

C: the *class* attribute

T: the set of *transaction*

ID3(*R*, *C*, *T*)

1. **[Algorithm Starts]**
2. **If *R* is empty, return a leaf-node with the class value assigned to the most transactions in *T*.**
3. **If *T* consists of transactions which all have the same value *c* for the class attribute, return a leaf-node with the value *c* (finished classification path).**
4. **Otherwise,**
   a. **Determine the attribute that *best* classifies the transactions in *T*, let it be *A*.**
   b. **Let $a_1,\ldots,a_m$ be the values of attribute *A* and let $T(a_1),\ldots,T(a_m)$ be a partition of *T* such that every transaction in $T(a_i)$ has the attribute value a_i.**
   c. **Return a tree whose root is labeled A (this is the test attribute) and has edges labeled a1,….,am such that for every i, the edge ai goes to the tree ID3(R-{A},C,T(ai)).**
5. **[End]**

---

3. GROUP SCHEMES

3.1 One-Group Scheme [7]

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 $(1-\theta)$

we use P (001) to represent

$P(A_1 = 1 \wedge A_2 = 1 \wedge A_3 = 0)$ to present$P(A_1 = 0 \wedge A_2 = 0 \wedge A_3 = 1)$. (1)

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

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

$$P^*(001) = P(001).\theta + P(110).(1-\theta) \quad (2)$$

By solving the above equations, we can get P(110)the information needed to build a decision tree. The general model for the one-group scheme is described in the following:

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

$$P^*(\overline{E}) = P(\overline{E}).\theta + P(E).(1-\theta) \quad (3)$$

Using the matrix form, let M_1 denote the coefficiency matrix of the above equations, the Matrix

$$\begin{pmatrix} p^*(E) \\ P^*(\overline{E}) \end{pmatrix} = M_1 \begin{pmatrix} P(E) \\ P(\overline{E}) \end{pmatrix}, \text{ where } M_1 = \begin{bmatrix} \theta & (1-\theta \\ 1-\theta & \theta \end{bmatrix} \quad (4)$$

3.2 Two-Group Scheme

In the one-group scheme, if the interviewer somehow knows whether the respondents tell a truth or a lie for one attribute, he/she can immediately obtain all the true values of a respondent's response for all other attributes. To improve data's privacy level, data providers divide all the attributes into two groups. They then apply the randomized response techniques for each group independently. For example, the users can tell the truth for one group while telling the lie for the other group. With this scheme, even if the interviewers know information about one group, they will not be able to derive the information for the other group because they are disguised independently.

To show how to estimate $P(E1\ E2)$ we look at all the contributions to $P^*(E1\ E2)$ .There are four partsthat contribute to $P^*(E1\ E2)$.

$$P^*(E_1\ E_2) = P(E_1\ E_2).\theta^2 + P(E_1\overline{E_2}).\theta(1-\theta) + P(\overline{E_1}E_2).\theta(1-\theta) + P(\overline{E_1E_2}).(1-\theta)^2$$

There are four unknown variables in the above equation:

$$(P(E_1\ E_2), P(E_1\overline{E_2}), P(\overline{E_1}E_2), P(\overline{E_1E_2}))$$

To solve the above equation, we need three more equations. We can derive them using the similar method.

$$\begin{pmatrix} P^*(E_1\ E_2) \\ P^*(E_1\ \overline{E_2} \\ P^*(\overline{E_1}E_2) \\ P^*(\overline{E_1E_2}) \end{pmatrix} = M_2 . \begin{pmatrix} P(E_1\ E_2) \\ P(E_1\overline{E_2}) \\ P(\overline{E_1}E_2) \\ P(\overline{E_1E_2}) \end{pmatrix} \quad (5)$$

The final equations are described in the following

Where

$$M_2 = \begin{bmatrix} \theta^2 & \theta(1-\theta)\theta(1-\theta) & (1-\theta)^2 \\ \theta(1-\theta) & \theta^2 & (1-\theta)^2 & \theta(1-\theta) \\ \theta(1-\theta) & (1-\theta)^2 & \theta^2 & \theta(1-\theta) \\ (1-\theta)^2 & \theta(1-\theta)\theta(1-\theta) & \theta^2 \end{bmatrix} \quad (6)$$

3.3 Three-Group Scheme

To further preserve the data's privacy, we can partition the attributes into three groups, and disguised each group independently. The model can be derived using the similar way as we did for the two-group model. The model for the three-group scheme is as follows:

$$\begin{pmatrix} P^*(E_1\,E_2\,\overline{E_3}) \\ P^*(E_1E_2\overline{E_3}) \\ P^*(E_1\overline{E_2}E_3) \\ P^*(E_1\overline{E_2E_3}) \\ P^*(E_1\overline{E_2}\,\overline{E_3}) \\ P^*(\overline{E_1}E_2\overline{E_3}) \\ P^*(\overline{E_1E_2}E_3) \\ P^*(\overline{E_1E_2E_3}) \end{pmatrix} = M_3 = \begin{pmatrix} P(E_1\,E_2\,\overline{E_3}) \\ P(E_1E_2\overline{E_3}) \\ P(E_1\overline{E_2}E_3) \\ P(E_1\overline{E_2E_3}) \\ P(E_1\overline{E_2}\,\overline{E_3}) \\ P(\overline{E_1}E_2\overline{E_3}) \\ P(\overline{E_1E_2}E_3) \\ P(\overline{E_1E_2E_3}) \end{pmatrix} \qquad (7)$$

$$M_3 = \begin{bmatrix} \theta^3 & \theta^2(1-\theta) & \theta^2(1-\theta) & \theta(1-\theta)^2 & \theta^2(1-\theta) & \theta(1-\theta)^2 & \theta(1-\theta)^2 & (1-\theta)^3 \\ \theta^2(1-\theta) & \theta^3 & \theta(1-\theta)^2 & \theta^2(1-\theta) & \theta(1-\theta)^2 & \theta(1-\theta)^2 & (1-\theta)^3 & \theta(1-\theta)^2 \\ \theta^2(1-\theta) & \theta(1-\theta)^2 & \theta^3 & \theta^2(1-\theta) & \theta(1-\theta)^2 & (1-\theta)^3 & \theta^2(1-\theta) & \theta(1-\theta)^2 \\ \theta(1-\theta)^2 & \theta^2(1-\theta) & \theta^2(1-\theta) & \theta^3 & (1-\theta)^3 & \theta(1-\theta)^2 & \theta(1-\theta)^2 & \theta^2(1-\theta) \\ \theta^2(1-\theta) & \theta(1-\theta)^2 & \theta(1-\theta)^2 & (1-\theta)^3 & \theta^3 & \theta^2(1-\theta) & \theta^2(1-\theta) & \theta(1-\theta)^2 \\ \theta(1-\theta)^2 & \theta^2(1-\theta) & (1-\theta)^3 & \theta(1-\theta)^2 & \theta^2(1-\theta) & \theta^3 & \theta(1-\theta)^2 & \theta^2(1-\theta) \\ \theta(1-\theta)^2 & (1-\theta)^3 & \theta^2(1-\theta) & \theta(1-\theta)^2 & \theta^2(1-\theta) & \theta(1-\theta)^2 & \theta^3 & \theta^2(1-\theta) \\ (1-\theta)^3 & \theta(1-\theta)^2 & \theta(1-\theta)^2 & \theta^2(1-\theta) & \theta(1-\theta)^2 & \theta^2(1-\theta) & \theta^2(1-\theta) & \theta^3 \end{bmatrix} \qquad (8)$$

Where $M_3$the coefficiency matrix and similar techniques is is can be employed to extend the above schemes to four-group scheme, five-group scheme, and so on. In this paper, we only describe our results for up to three-group scheme because the undesirable performance for schemes beyond the three-group scheme makes them not very useful. We will compare the performance of various schemes.

3.4 Four group scheme

To further preserve privacy we can partition the attributes into four groups, and disguised each group independently. The model for the four group scheme is as follows:

$$\begin{pmatrix} P^*(E_1E_2E_3E_4) \\ P^*(E_1E_2E_3\overline{E_4}) \\ P^*(E_1E_2\overline{E_3}E_4) \\ P^*(E_1E_2\overline{E_3E_4}) \\ P^*(E_1\overline{E_2}E_3E_4) \\ P^*(E_1\overline{E_2}E_3\overline{E_4}) \\ P^*(E_1\overline{E_2E_3}E_4) \\ P^*(E_1\overline{E_2E_3E_4}) \\ P^*(\overline{E_1}E_2E_3E_4) \\ P^*(\overline{E_1}E_2E_3\overline{E_4}) \\ P^*(\overline{E_1}E_2\overline{E_3}E_4) \\ P^*(\overline{E_1}E_2\overline{E_3E_4}) \\ P^*(\overline{E_1E_2}E_3E_4) \\ P^*(\overline{E_1E_2}E_3\overline{E_4}) \\ P^*(E_1\overline{E_2E_3}E_4) \\ P^*(\overline{E_1E_2E_3E_4}) \end{pmatrix} = M_4 \begin{pmatrix} P(E_1E_2E_3E_4) \\ P(E_1E_2E_3\overline{E_4}) \\ P(E_1E_2\overline{E_3}E_4) \\ P(E_1E_2\overline{E_3E_4}) \\ P(E_1\overline{E_2}E_3E_4) \\ P(E_1\overline{E_2}E_3\overline{E_4}) \\ P(E_1\overline{E_2E_3}E_4) \\ P(E_1\overline{E_2E_3E_4}) \\ P(\overline{E_1}E_2E_3E_4) \\ P(\overline{E_1}E_2E_3\overline{E_4}) \\ P(\overline{E_1}E_2\overline{E_3}E_4) \\ P(\overline{E_1}E_2\overline{E_3E_4}) \\ P(\overline{E_1E_2}E_3E_4) \\ P(\overline{E_1E_2}E_3\overline{E_4}) \\ P(E_1\overline{E_2E_3}E_4) \\ P(\overline{E_1E_2E_3E_4}) \end{pmatrix} \qquad (9)$$

Where M_4 is the coefficiencymatrix

ID3 algorithm uses the information gain to select the test attribute. Information gain can be computed using entropy. In the following, we assume there are m classes in the whole training data set. We know

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

where Qj(S) is the relative frequency of class j in S. We can compute the information gain for any candidate attribute A being used to partition S:

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

3.5 Privacy Analysis

We conduct privacy analyze from two aspects one is the case where we fix the scheme other is the case where we fix theta value.

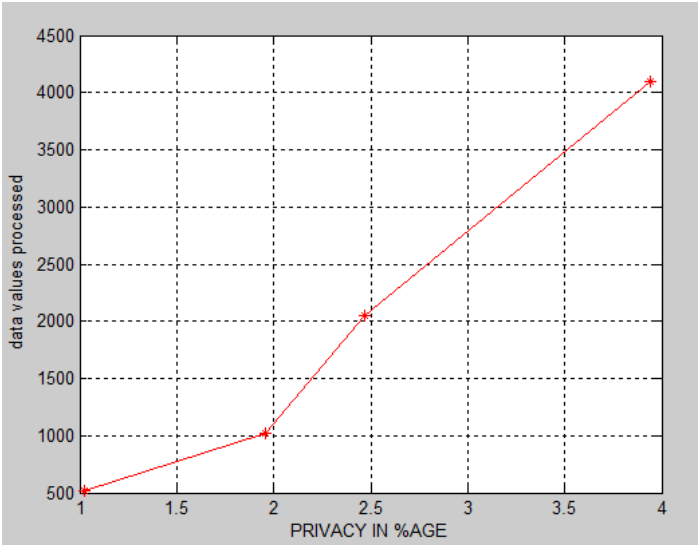

Figure : Privacy of dataset using four groups.

4. CONCLUSION

This thesis gives a different approach for enhancing the privacy of the sensitive data in data mining. In this thesis, we use the four group randomized response technique in privacy preserving algorithm. To support our work we used CART and ID3 algorithm.

In our experiment we first applied randomized response techniques on one, two, three and four groups. We have applied ID3 and CART algorithm on the randomized data. After that we have calculated the gain and compared the randomized data with original undisguised data to test the accuracy level.

5. REFERENCES

[1] Giudici, P., *Applied Data-Mining: Statistical Methods for Business and Industry*. West Sussex, England: John Wiley and Sons (2003). Page no 2

[2] A New Approach to Maintain Privacy And Accuracy In Classification Data Mining

[3] A Survey On Privacy Preserving Data Mining Approaches And Techniques

[4] An Overview of Data Mining Techniques Excerpted from the book by Alex Berson, Stephen Smith, and Kurt Thearling Berkhin Pavel,

[5] A Survey of Clustering Data Mining Techniques, Springer Berlin Heidelberg, 2006A

[6] Hybrid Approach Using C Mean and CART for Classification in Data Mining

[7] Privacy-Preserving Data Mining Using Multi-Group Randomized Response Techniques