PRIVACY PRESERVING DATA MINING IN FOUR GROUP RANDOMIZED RESPONSE TECHNIQUE USING ID3 AND CART ALGORITHM

Monika Soni[1], Vishal Shrivastava[2]

[1]M. Tech. Scholar, [2]Associate Professor, Arya College of Engineering and IT, Rajasthan, India,
12.monika@gmail.com**, **vishal500371@yahoo.co.in

Abstract

Data mining is a process in which data collected from different sources is analyzed for 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. 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 biased answers. The proposed work 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 are applied on the randomized data. The result shows that by increasing the group, the privacy level will increase. This work shows that as compared with three group scheme with four groups scheme the accuracy decreases 6% but the privacy increases 65%.

----------***----------

1. INTRODUCTION OF PROPOSED APPROACH

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. The proposed work increases the level of privacy by using ID3 and CART algorithms. Previous work was giving an overall result whereas this work is implementing it in step by step manner.

1.1 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. ID3 algorithm is given below:

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 a_i 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).

1.2 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. A possible question could be: ”Is age greater than 50?” or ”Is sex male?”. 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 process is then repeated for each of the resulting data fragments.

Step 1: Find each predictor’s best split.

Step 2: Find the node’s best split.

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

2. RANDOMIZED RESPONSE TECHNIQUES

Randomized Response (RR) techniques were developed for the purpose of protecting survey's privacy and avoiding answer bias mainly. 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.

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.

- 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. In this way the respondents' privacy is preserved. Since the interviewer's wants to know the answer to the first question, and the answer to the second question is exactly the opposite to the answer for the first one, if the respondent chooses to answer the first question, then it say that he/she is telling the truth; if the respondent chooses to answer the second question, then it say that he/she is telling a lie. To estimate the percentage of people who has the attribute A. The following equations can be used:

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

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

Where P*(A=yes) (resp. p*(A=no)) is the proportion of the "yes" (resp. "no") responses obtained from the survey data, and P (A=yes) (resp. P (A=no)) is the estimated proportion of the "yes" (resp. "no") responses to the sensitive questions.

2.1 One-Group 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 $(1 - \theta)$.

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

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

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

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

$$\begin{pmatrix} \mathrm{p}^*(\mathrm{E}) \\ \mathrm{P}^*(\overline{\mathrm{E}}) \end{pmatrix} = \mathrm{M1} \begin{pmatrix} \mathrm{P}(\mathrm{E}) \\ \mathrm{P}(\overline{\mathrm{E}}) \end{pmatrix}, \text{ where } \mathrm{M1} = \begin{bmatrix} \theta & (1-\theta) \\ 1-\theta & \theta \end{bmatrix}$$

2.2 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.

Then it has the following equation:

$$\mathrm{P}^*(\mathrm{E}_1\,\mathrm{E}_2) = \mathrm{P}(\mathrm{E}_1\,\mathrm{E}_2).\theta^2 + \mathrm{P}(\mathrm{E}_1\overline{\mathrm{E}_2}).\theta(1-\theta) + \mathrm{P}(\overline{\mathrm{E}_1}\mathrm{E}_2).\theta(1-\theta) + \mathrm{P}(\overline{\mathrm{E}_1\mathrm{E}_2}).(1-\theta)^2$$

There are four unknown variables in the above equation:

$$(\mathrm{P}(\mathrm{E}_1\,\mathrm{E}_2), \mathrm{P}(\mathrm{E}_1\overline{\mathrm{E}_2}), \mathrm{P}(\overline{\mathrm{E}_1}\mathrm{E}_2), \mathrm{P}(\overline{\mathrm{E}_1\mathrm{E}_2})) \quad (9)$$

To solve the above equation, three more equations are needed , derive them using the similar method.

$$\begin{pmatrix} \mathrm{P}^*(\mathrm{E}_1\,\mathrm{E}_2) \\ \mathrm{P}^*(\mathrm{E}_1\,\overline{\mathrm{E}_2} \\ \mathrm{P}^*(\overline{\mathrm{E}_1}\mathrm{E}_2) \\ \mathrm{P}^*(\overline{\mathrm{E}_1\mathrm{E}_2}) \end{pmatrix} = \mathrm{M}_2 . \begin{pmatrix} \mathrm{P}(\mathrm{E}_1\,\mathrm{E}_2) \\ \mathrm{P}(\mathrm{E}_1\overline{\mathrm{E}_2}) \\ \mathrm{P}(\overline{\mathrm{E}_1}\mathrm{E}_2) \\ \mathrm{P}(\overline{\mathrm{E}_1\mathrm{E}_2}) \end{pmatrix}$$

The final equations are described in the following

$$\text{Where } \mathrm{M}_2 = \begin{bmatrix} \theta^2 & \theta(1-\theta)\theta(1-\theta) & (1-\theta)^2 \\ \theta(1-\theta) & \theta^2 \quad (1-\theta)^2 & \theta(1-\theta) \\ \theta(1-\theta) & (1-\theta)^2 \quad \theta^2 & \theta(1-\theta) \\ (1-\theta)^2 & \theta(1-\theta)\theta(1-\theta) & \theta^2 \end{bmatrix}$$

2.3 Three-Group Scheme

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

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

$$M_3 = \begin{bmatrix} \theta^3 & \theta^2(1-\theta) & \theta^2(1-\theta) & \theta(1-\theta)^2 & \theta^2(1-\theta) & \theta(1-\theta)^2 & \theta(1-\theta)^2 & (1-\theta)^3 \\ \theta^2(1-\theta) & \theta^3 & \theta(1-\theta)^2 & \theta^2(1-\theta) & \theta(1-\theta)^2 & \theta(1-\theta)^2 & (1-\theta)^3 & \theta(1-\theta)^2 \\ \theta^2(1-\theta) & \theta(1-\theta)^2 & \theta^3 & \theta^2(1-\theta) & \theta(1-\theta)^2 & (1-\theta)^3 & \theta^2(1-\theta) & \theta(1-\theta)^2 \\ \theta(1-\theta)^2 & \theta^2(1-\theta) & \theta^2(1-\theta) & \theta^3 & (1-\theta)^3 & \theta(1-\theta)^2 & \theta(1-\theta)^2 & \theta^2(1-\theta) \\ \theta^2(1-\theta) & \theta(1-\theta)^2 & \theta(1-\theta)^2 & (1-\theta)^3 & \theta^3 & \theta^2(1-\theta) & \theta^2(1-\theta) & \theta(1-\theta)^2 \\ \theta(1-\theta)^2 & \theta^2(1-\theta) & (1-\theta)^3 & \theta(1-\theta)^2 & \theta^2(1-\theta) & \theta^3 & \theta(1-\theta)^2 & \theta^2(1-\theta) \\ \theta(1-\theta)^2 & (1-\theta)^3 & \theta^2(1-\theta) & \theta(1-\theta)^2 & \theta^2(1-\theta) & \theta(1-\theta)^2 & \theta^3 & \theta^2(1-\theta) \\ (1-\theta)^3 & \theta(1-\theta)^2 & \theta(1-\theta)^2 & \theta^2(1-\theta) & \theta(1-\theta)^2 & \theta^2(1-\theta) & \theta^2(1-\theta) & \theta^3 \end{bmatrix} \quad (13)$$

Where $M_3$the coefficiency matrix and similar techniques can be employed to extend the above schemes to four-group scheme

2.4 Four group scheme

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

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

Where M_4 is the co efficiency matrix.

3. PROCESSING THE STEPS ARE GIVEN BELOW FOR IMPLEMENTING THE WORK:

1. Since it is considered that the dataset contains only binary data. First transformed the non binary data into binary data. For this there are many methods available in MATLAB.
2. Divided the data randomly in groups as follows:
   - Four random attributes are taken namely a, b, c and d.
   - The dataset are grouped in to 1, 2, 3 and 4 level according to the following manner

     group1=all data attributes are considered as single group and performed the above operations

group2=data set divided into ab and cd and performed the randomization and id3 algorithm group3=data set divided into ab, c, d group4=data set divided into a, b, c, d

Following steps are used on each dataset and on each group for different values of Θ. Here related question model of randomized response technique is used.

Θ = [0.0, 0.1, 0.2, 0.3, 0.45, 0.51, 0.55, 0.6, 0.7, 0.8, 0.9 1];

For randomization it generates a random number r from 0 to 1 using uniform distribution.

a. Randomization

   For one-group scheme, a disguised data set G is created. For each record in the training data set D, A random number r from 0 to 1 is generated using uniform distribution.

b. Build the decision tree
For building the decision tree CART and ID3 algorithms are used with disguised data set G.

c. Testing
Training dataset is used for testing. Test with the original data set.

d. Repeat the steps a to c for 50 times. Then compute the mean and variance.

On after the calculation of the gain, it is checked against the original data to find out how the accuracy and privacy is affected

4. RESULT ANALYSIS

For each group and each data set mean and variance are calculated. Mean value of dataset are calculated for finding the accuracy for different groups of different dataset.

4.1 Accuracy between Three Groups and Four Groups

The figure 1 shows the accuracy using three group scheme and figure 2 shows the accuracy using four group schemes. The result shows that by using three groups scheme the accuracy is 25.2% but the accuracy will decrease up to 19.2% while using the four group scheme. The accuracy will decrease 6% when the data is divided from three groups to four groups. When this work is compared with previous work then the results shows that the accuracy is negligible decreased but the privacy of datasets are increased which is described in next section

.

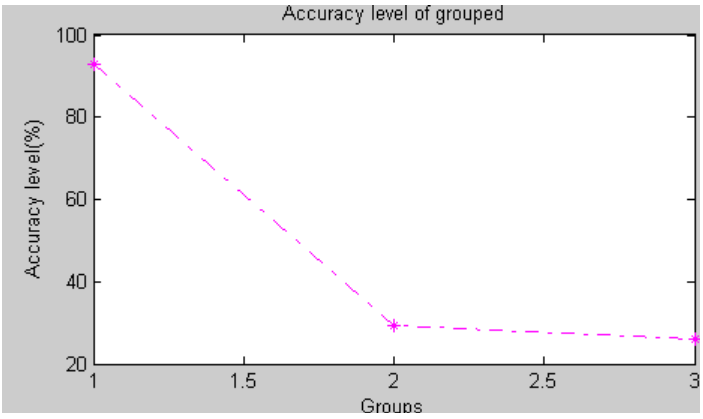

Figure 1: Accuracy in percentage using three groups scheme

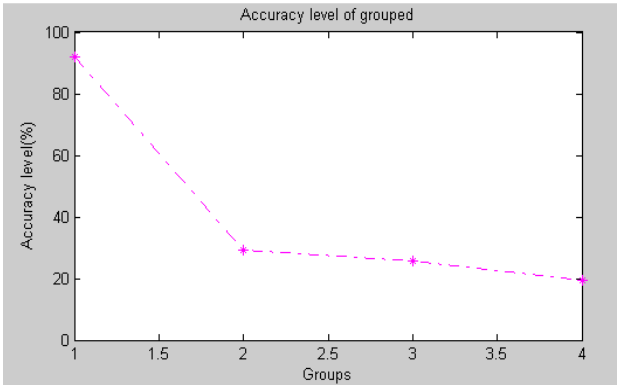

Figure 2: Accuracy in percentage using four groups scheme

4.2 Privacy between Three Groups and Four Groups

The figure 3 shows the privacy using three group scheme and figure 4 shows the privacy using four group schemes. The result shows that by using three groups scheme the privacy is 29% and the privacy will increase up to 94% while using the four group scheme. The privacy will increase 65% when the data is divided from three groups to four groups. When this work is compared with previous work then the results shows that the privacy is increased at very high level and the decrease in accuracy is negligible.

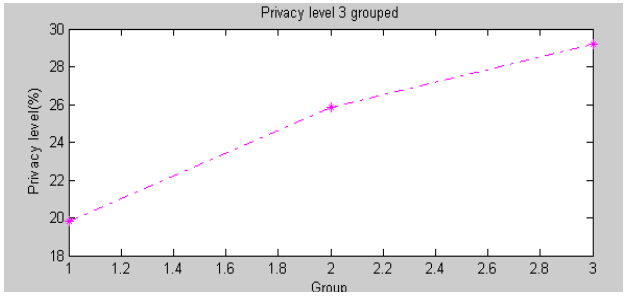

Figure 3: Privacy using three group schemes

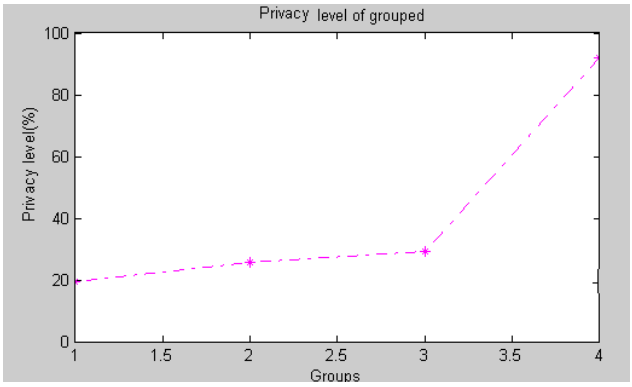

Figure 4: Privacy using four group schemes

CONCLUSIONS

This thesis 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. This work shows that as compared with three group scheme with four groups scheme the accuracy decreases 6% but the privacy increases 65%.

By the experiment it is concluded that with increasing the level of grouping the privacy of the dataset will be enhanced and the accuracy level decreases but it is negligible.

REFERENCES

[1] 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

[2] Jasbir Malik, Rajkumar, "A Hybrid Approach Using C Mean and CART for Classification in Data Mining", IJCSMS, Vol. 12, Issue 03, Sept 2012

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

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

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

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

BIOGRAPHIES:

Mrs. Monika Soni Pursuing M. Tech. in Computer Science She has published many national and international research papers. She has written 3 books for engineering and engineering diploma.

Mr. Vishal Shrivastava working as Assistant Professor in Arya College & IT He has published many national and international research papers. He has very depth knowledge of his research areas.