Privacy Preserving Data Mining: Comparion of Three Groups and Four Groups Randomized Response Techniques

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: 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 biased answers. In randomized response technique adds certain degree of randomness to the answer to prevent the biased data. 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

Data Mining can be referred 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, "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."

2. 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 checking the group performance at every step and the privacy level was not very high. This proposed work is increase the level of privacy by using ID3 and CART algorithms. Previous work was giving an overall result whereas this work is implementing in step by step manner.

Here randomized algorithm is applied on four groups, which is an enhancement over previous related work where it has been applied on only three groups. In previous work only ID3 algorithm was used for privacy preserving data mining and according to that if work is done on more groups the privacy will not be affected. So this work proposed an approach where using the CART algorithm with ID3 algorithm on four groups of data sets. The previous work shows that when data set group is increase up to three levels, the privacy of the data is increased only 29%. This proposed work enhances the privacy level for four groups using different algorithm than previous work and also work on the accuracy of datasets.

The principle of the ID3 algorithm is as follows. The tree is constructed top-down in a recursive fashion. At the root, each attribute is tested to determine how well it alone classifies the transactions. The “best” attribute (to be discussed below) is then chosen and the remaining transactions are partitioned by it. ID3 is then recursively called on each partition (which is a smaller data set containing only the appropriate transactions and without

the splitting attribute). Figure 1 shows the description of the ID3 algorithm.

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]**

Figure 1: The ID3 Algorithm for Decision Tree Learning

2.1 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.

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. That is, each split depends on the value of only one predictor variable. All possible splits consist of possible splits of each predictor. If X is a nominal categorical variable of I categories, there are $2^{I-1} - 1$ possible splits for this predictor. If X is an ordinal categorical or continuous variable with K different values, there are K - 1 different split on X. 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.

For each continuous and ordinal predictor, sort its values from the smallest to the largest. For the sorted predictor, go through each value from top to examine each candidate split point (call it v, if $x \leq v$, the case goes to the left child node, else, goes to the right.) to determine the best. The best split point is the one that maximize the splitting criterion the most when the node is split according to it. For each nominal predictor, examine each possible subset of categories (call it A, if $x \in A$, the case goes to the left child node, else, goes to the right.) to find the 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.

2.2 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) 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. 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 . Warner used RR technique to solve the following survey problem: To estimate the percentage of people in a population that has attribute A, queries are sent to a group of people. Since the attribute 0 is related to some confidential aspects of human life, respondents may decide not to reply at all or to reply with incorrect answers. Two models, Related-Question Model and Unrelated-Question Model, have been proposed to solve this survey problem.

3. Methodology

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. If r<=Theta, the record is copied to G without any change; if r>Theta, then the opposite of the record is copied to G, namely each attribute value of the record put into G is exactly the opposite of the value in the original record. This randomization step is performed for all the records in the training data set D and generate the new data set G. For the two-group, three-group and four-group scheme, here D is randomly split into two, three or four groups, and conduct the above randomization for each group, finally obtain disguised data set G.

Now after the randomization, the gains and variance are calculated in the datasets

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 is find out how the accuracy and privacy is affected

4. Comparison of accuracy between three groups and four groups

The figure 2 shows the accuracy using three group scheme and figure 3 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.

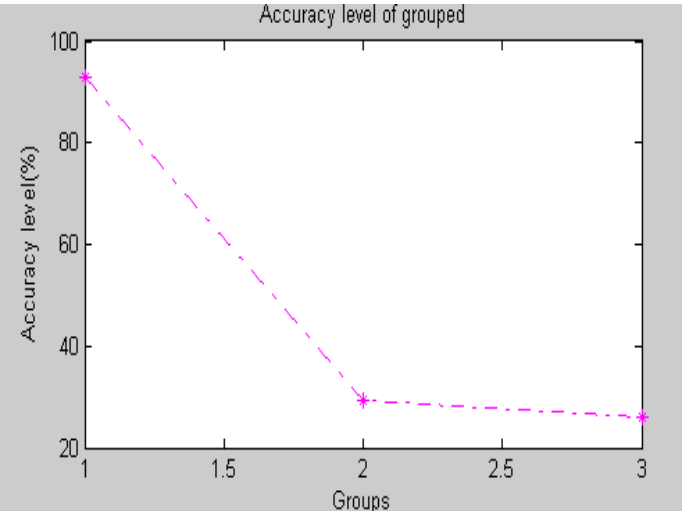

Figure 2 : Accuracy in percentage using three groups scheme

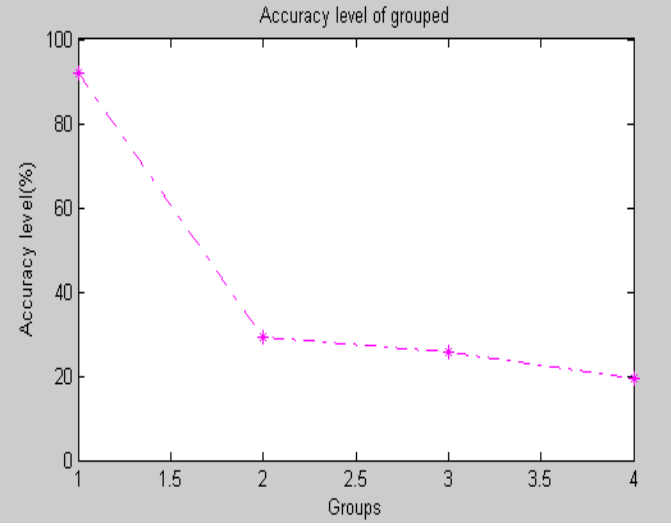

Figure 3 : Accuracy in percentage using four groups scheme

5. Comparison of privacy between three groups and four groups

The figure 4 shows the privacy using three group scheme and figure 5 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 decreasing of accuracy can be neglected.

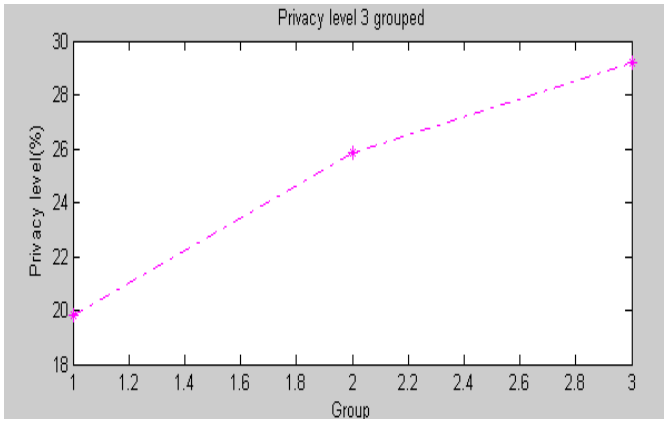

Figure 4 : Privacy using three group schemes

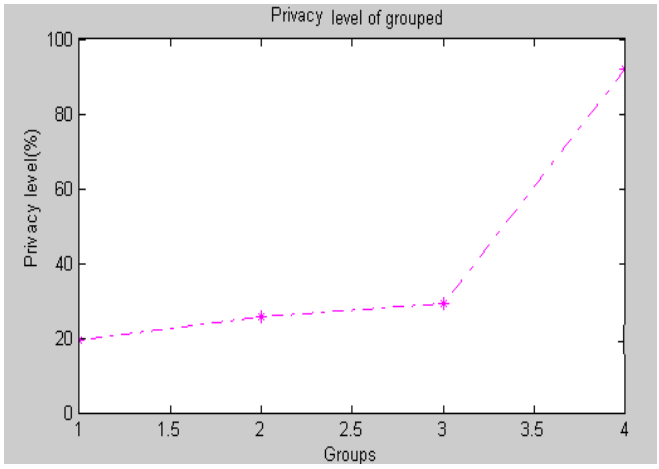

Figure 5 : Privacy using four group scheme

6. Conclusion

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.

7. Reference

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

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

[3] 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

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

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

[6] Gerty J. L. M. Lensvelt-Mulders, Joop J. Hox And Peter G. M. Van Der Heijden "How To Improve The Efficiency of Randomised Response Designs ", Springer 2005