Data Hiding Techniques to Maintain Privacy in Revealed Numeric Database

[1]Prof. Vishal Shinde, [2]Mr. Chandanshiv Hanumant, [3]Miss. More Parvati, [4]Miss. Patil Rakhi

[1]Asst. Professor, [2,3,4]UG Student, [1,2,3,4]Computer Engg. Dept. Shivajirao S.Jondhle College of Engineering & Technology, Asangaon, Maharshatra, India.

[2]chandanshiv45@gmail.com, [3]parvatimore9@gmail.com, [4]rakhip12@gmail.com

Abstract- In many organizations large amount of data are collected. These data are sometimes used by the organizations for data mining tasks. However, the data collected may contain private or sensitive information which should be protected. Privacy protection is an important issue if we release data for the mining or sharing purpose. Privacy preserving data mining techniques allow publishing data for the mining purpose while at the same time preserve the private information of the individuals. Many techniques have been proposed for privacy preservation but they suffer from various types of attacks and information loss. In this paper we proposed an efficient approach for privacy preservation in data mining. Our technique protects the sensitive data with less information loss which increase data usability and also prevent the sensitive data for various types of attack. Data can also be reconstructed using our proposed technique.

Keywords- Data mining, Privacy preserving, Sensitive data, Randomization, K- anonymity, Quasi-identifier.

I. INTRODUCTION

Many organizations like credit card companies, real estate companies, search engines, hospitals collects and hold large volume of data. The data are further used by the data miner for the analysis purpose which helps the organizations for gaining useful knowledge. These data may contain sensitive or valuable information of any individuals, For example, organizations such as hospitals contain medical records of the patients, They provide these database or records to the researchers or data miner for the purpose of research. Data miner analyzes the medical records to gain useful global health statistics. However, in this process the data miner may able to obtain sensitive information and in combination with an external database may try to obtain personal attribute of an individual So privacy is become an important issue when data that involves sensitive information.

To solve this, an interesting new direction in the field of data mining has been emerged known as privacy preserving data mining (PPDM) [1, 2]. The objective of this technique is extraction of useful knowledge from large amount of data, while protecting the sensitive information simultaneously. Privacy preserving data mining techniques are divided into two broad areas, data hiding and knowledge hiding. Data hiding is removal or modification of confidential information from the data before disclosing to others. Knowledge hiding is focus on hiding the sensitive knowledge which can be mined from the database using any data mining algorithm [3].

Several techniques have been evolved for the privacy preservation in data mining such as non-cryptographic and cryptographic techniques. Non-cryptographic techniques contain K-anonymity, I-diversity, t-closeness, perturbation and association rules, The cryptographic technique supports multi - party computation. The problem with non-cryptographic method is information loss. On the other hand cryptographic method provides accurate results but it suffers from high computation and communication cost. In this paper we mainly focus on non-cryptographic techniques.

There is a tradeoff between the privacy and information loss. To maintain this tradeoff, an efficient privacy preservation method is proposed. In our proposed method, first we apply randomization on original data and then after randomization we categorize the sensitive attribute values into high sensitive and low sensitive class. Secondly we apply k-anonymization on those tuples who belongs to high sensitive class and those tuples who belongs to low sensitive remain as it is. So it reduces the information loss and improves the usability of data. The combination of anonymization with randomization technique is made difficult for the attacker to attack on database.

The paper is progress as follows. Section II illustrates the literature review. Section III defines the problem definition. Section IV discuss about the proposed approach, Section V concludes the work and future scope is presented.

II. RELATED WORK

The existing method simple additive noise (SAN) method is adding the noise parameter which have mean zero and variance proportion parameter determined by the user to the original confidential attribute then the result is perturbed

value of confidential attribute. The drawback of simple additive noise method is that the noise is independent of the scale of confidential attribute. To overcome the SAN method drawback next proposed approach is multiplicative noise (MN) ,in this method the confidential attribute is multiplied with the noise with mean one to get perturbed value of confidential attribute. These two methods are causes the bias in the variance of the confidential attribute, as well as in the relationships between attributes. Another proposed method is micro aggregation (MA), the MA perturbs data by aggregating Confidential values, instead of adding noise. For a data set with a single confidential attribute ,univariate micro aggregation (UMA) involves sorting records by the confidential attribute, grouping adjacent records into groups of small sizes, and replacing the individual confidential values in each group with the group average. Similar to SAN and MN, UMA causes bias in the variance of the confidential attribute, as well as in the relationships between attributes.Multivariate micro aggregation (MMA)differs from UMA in that it groups data using a clustering technique that is based on a multi-dimensional distance measure. As a result, the relationships between attributes are expected to be better preserved. However, this benefit comes with a higher computational time complexity, which could be inefficient for large data sets. So in order to provide privacy to the large data sets we are going to proposing approach based on the perturbation trees, a kd-tree is data structure for partitioning the and storing data. A kd-tree recursive partitioning technique to divide a data set into subsets that contain similar data. The partitioned data are perturbed using the subset average. Since the data are rtitioned based on the joint properties of multiple confidential and non-confidential attributes , the relationships between attributes are expected to be reasonably preserved.

III. LITERATURE SURVEY

There are many approaches which have been adopted for privacy preserving data mining.

Classification is based on the following dimensions:

- Data distribution
- Data modification
- Data mining algorithm
- Data or rule hiding
- Privacy preservation

The first dimension refers to the distribution of data. Some of the approaches have been developed for centralized data, while others refer to a distributed data scenario. Distributed data scenarios can also be classified as horizontal data distribution and vertical data distribution. Horizontal distribution refers to these cases where different database records reside in different places, while vertical data distribution, refers to the cases where all the values for different attributes reside in different places. The second dimension refers to the data modification scheme. In general, data modification is used in order to modify the original values of a database that needs to be released to the public and in this way to ensure high privacy protection. It is important that a data modification technique should be in concert with the privacy policy adopted by an organization. Methods of modification include: Perturbation, which is accomplished by the alteration of an attribute value by a new value (i.e.changing a 1-value to a 0-value, or adding noise)

- blocking, which is the replacement of an existing attribute value with a ?
- Aggregation or merging which is the combination of several values into a coarser category.
- swapping that refers to interchanging values of individual records, and Sampling, this refers to releasing data for only a sample of a population

The third dimension refers to the data mining algorithm, for which the data modification is taking place. This is actually something that is not known beforehand, but it facilitates the analysis and design of the data hiding algorithm. various data mining algorithms have been considered in isolation of each other. Among them, the most important ideas have been developed for classification data mining algorithms, like decision tree inducers, association rule mining algorithms, clustering algorithms, rough sets and Bayesian networks.

The fourth dimension refers to whether raw data or aggregated data should be hidden. The complexity for hiding aggregated data in the form of rules is of course higher, and for this reason, mostly heuristics have been developed. The lessening of the amount of public information causes the data miner to produce weaker inference rules that will not allow the inference of confidential values. This process is also known as rule confusion. The last dimension which is the most important refers to the privacy preservation technique used forth selective modification of the data. Selective modification is required in order to achieve higher utility for the modified data given that the privacy is not jeopardized.

The techniques are:

- Heuristic-based techniques like adaptive modification that modifies only selected values that minimize the utility loss rather than all available values.
- Cryptography-based technique, secure multiparty computation where a computation is secure if at the end of the computation, no party knows anything except its input and the results.
- Reconstruction-based techniques where the original distribution of the data is reconstructed from the randomized data. It is important to realize that data modification results in degradation of the database performance. In order to quantify the degradation of the data, mainly two metrics are used. The first one, measures the confidential data protection, while the second measures the loss of functionality. [4]

IV.EXISTING SYSTEM

In existing system, there is no possibility to distributed and perturb the sensitive and secured data which is able to make the data less confidential. In the existing systems, the data of the original distributor will be changed which is known as data leakage which is incorporated with the firmly fixed data copies. The data leakers can be identified only when the data copy is within uncertified user hands and the watermarks will even be disrupted by the data recipients who turn the data into malicious data.

ALGORITHM 1 SAN METHOD

1: procedure SAN METHOD()Finding the Mean
2: for <i=0 to org.datalength> do
3: orgmean+ = Double:parseDouble(orgData[i][1])
4: end for
5: orgmean = orgmean=orgData:length
Generating Perturb Data
6: ptData = newdouble[orgData:length]
7: for <i=0 to org.datalength> do
8: orgtmp = Double:parseDouble(orgData[i][1])
9: orgtmp = orgtmp + orgmean
10: ptData[i] = orgtmp
11: end for
12: end procedure

ALGORITHM 2 MN METHOD

1: procedure MN METHOD( )Finding the Mean
2: for <i=0 to org.datalength> do
3: orgmean+ = Double:parseDouble(orgData[i][1])
4: end for
5: orgmean = orgmean=orgData:length
Generating Perturb Data
6: ptData = newdouble[orgData:length]
7: for <i=0 to org.datalength> do
8: orgtmp = Double:parseDouble(orgData[i][1])
9: orgtmp = orgtmp orgmean
10: ptData[i] = orgtmp
11: end for
12: end procedure

V. PROBLEM STATEMENT

It is challenging to handling the sensitive data from the various private data bases. Generally to solve this type of problem they used Data perturbation technique with some specific mechanism in existing methods. The challenge comes from the individuals need to protection and privacy of sensitive and private data. To do this work traditional system using various different approaches, this approaches are concentrating the mining of data as sensitive with confidential and non-confidential data sets. To vary the confidential data from entire data is risk and the data of confidential rules changes on the data access vendor. It is very costly operation on mining the data from databases and handling the sensitive data. The existing methods failure on maintain performances and time complexity. To protect the data they used as additional noise will merging with the actual data. To producing the results they will uses the encryption of data with noise and decrypting the data to divide the actual data and noise data. The proposed mechanism of perturbation tree, the tree will handle the data partitioning the data sets and subsets. The each subset must satisfy the some minimum conditional values will store and from as leaf of the tree. This subset partitioning is combination of the confidential and non-confidential data

VI. PROPOSED SYSTEM

The proposed mechanism works and implements the approach of perturbation tree, as one of the general method like divide and conquer method. This method will divide the data in the data set and subsets, this datasets and subsets are conquering in the tree set approaching. This perturbation tree delivers the tree set having the leaf as a distance relation linkage. The distances of linkage will find based on the average square distances. The linkage of records is linked set having duplicate date and nearly sub linked having the actual data. This approach will accept the datasets as input and getting the sensitive data. The sensitive data sets will divide into subsets and storing in tree structured format. In the tree each leaf node having some set of attributes dependent on the data sets selections. In each tree node will replacing the sensitive original data with the average value of attributes. This data will send to the shareable author and decrypting the data by using the linear regression technique. This approach will evaluate based on the existing systems.

VII. WORKING OF ALGORITHM

Perturbation tree mechanism consist the various confidential and non-confidential data sets. The general idea of perturbation tree is as follows:

- By getting the data sets attributes, including the confidential attributes, in the data and normalized data (non-confidential).
- By computing the normalized data into Normalized data matrix in the tree current node. Now compute the matrix variances in each dimension.
- By finding the median value of matrix it id needed to compute the leaf of tree (sub sets of data) based on the median value range.
- By repeating the variances and median values computation until the each node having the less than specified records range.
- By perturbation the confidential values by specifying the average values base on the above step to attributes in the each and every node set of confidential data.

Mathematical Model

Perturbation tree mechanism consist the various confidential and non-confidential datasets. The general idea of perturbation tree is

• Step 1: Let J be the number of attributes, including confidential attributes in data. Normalize the data to the unit

$$MAE(MeanAbsoluteError) = \frac{1}{M}\sum_{i=1}^{M}|x_i - \hat{x}_i|$$

scale.

• Step2: Let Z be the normalized data matrix at the current node. Compute the variance of each dimension, based on Z.

$$ASD(AVG.SQ.DIST) = \frac{1}{N}\sum_{i=1}^{N}(y_i - x_i)^2$$

Let j* be the dimension with the max. Variance.

• Step 3: Find the median (mid-range) of attribute j*. Partition Z into two sub sets (child nodes) based on median.

• Step 4: Repeat step-2 and 3 for each of child nodes. Stop the process when the node contains less than a pre-specified number of nodes.

• Step 5: For a leaf t with nt records, let xt1....xtnt be the confidential values. Perturb the data by replacing these values with

Repeat this step for each leaf in the tree built in step-4. (If there are multiple attributes to be perturbed, the avg. of each attribute is used to replace the value so that attribute.)Smaller disclosure risk, higher info. Loss. Sx and Sy are Std. deviation of original and perturb confidential values resply. Smaller BIM and BISD are desirable. Measure univariate info. Loss due to perturbation.

Apply linear regression and C4.5 classifier on perturbed data to build regression and classification model and computed errors Smaller MAE value is desirable.Where M is the number of records in the test set. xi is the confidential value of the ith record in the test set, and *xi* is the estimate of xi based b on the regression model. Since MAE measures the distance between the predictions of the model built from the perturbed data and the (unperturbed) test data, a smaller MAE value is desirable.

VIII. SYSTEM ARCHITECTURE

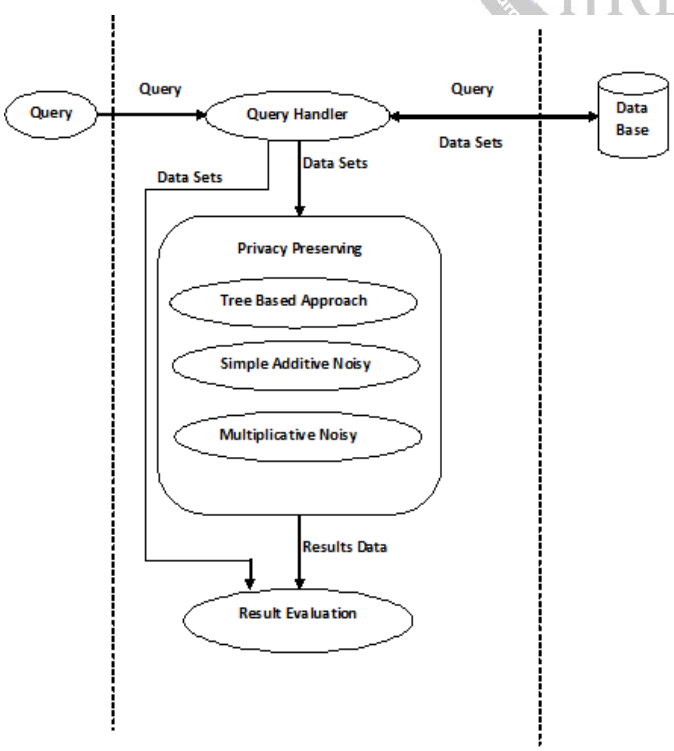

Fig. 1 System Architecture

Indicating the system architecture of the tree based data perturbation process. This architecture will give complete description of input and outputs of each process. This process we have several modules. They are

1) Query Handler
2) Privacy Preserving
    1. Perturbation Tree
    2. Simple Adaptive Noisy
    3. Multiplicative Noisy
3) Result Evaluation

Query Handler

The Query handler is accepting the query data from the client and process the query with the data base and fetching the datasets from the data base.

Privacy Preserving

The privacy preserving is a process of providing the security for sensitive data. The sensitive data like an employee salary, annual income of company, transferring the money from one account to another account etc. providing the security to this data is very important. To do this work we have some existing system and we proposed one system. Those approaches are implemented by following sub modules of this mechanism. The proposed approach is Perturbation Tree and the existing systems are Simple adaptive noisy and multiplicative noisy. We implemented both existing and proposed approaches because to evaluate out proposed approach results.

Perturbation Tree

Perturbation tree is our proposed approach. In this approach we using the divide and conquer technique. This technique will be using the following process, this approach accept the data sets as input. This data sets will divided in subsets by using above mention technique and storing in tree format up to in tree each child of leaf node having the attributes as the user mention equals or less values. After completion of the division process, each leaf node attributes sensitive data will replacing with the average value and sending to sharable person or other requested client.

Simple Adaptive Noisy

The simple adaptive noisy (SAN) exists system this approach we are adding the random number to the original data and replacing the original data with the noisy data. The random number will get by the client.

Multiplicative Noisy

The multiplicative noisy (MN) also exist system this approach calculating the mean of the original data. The mean value will be multiplicity with the original data and replacing the original data with multiplicity result data.

Result Evaluation

The result evaluation is a process to finding the error rate of different states in the data perturbation of original data and the perturbed data. In the result evaluation we considering several processes those are time complexity, record linkage, RASD, bias in mean, bias in standard deviation, regression

and classification. The time complexity will evaluated based on the processing time delay of the input acceptance to producing the output to client. To do this work we find the start time and end time of process, subtracting the end time and start time we get the time of evaluation process in milliseconds. The time will convert to seconds by dividing the milliseconds with the thousand.

The Record linkage will be finding based on the perturbed data and the original data. The distance between the perturbed data and original data is a disclosure risk of the data. To finding the distances we used Ebulliences Distances between the perturbed data and the original data. The RL and square root of average squared distances (ASD) used to calculate the disclosure risk. The disclosure risk is evaluating the information loss will be measured. To calculate the bias in mean value we are using the mean of the original data and the perturbed data. By using approach we find the information loss of perturbation. To calculate the bias in standard deviation we are using of the original data and perturbed data. By finding this value we get the loss of information in perturbed data. The regression error rate will finding based on the mean average error rate. These values will give the information of error rate of this approach on the data perturbation.

IX. DESIGN DETAILS

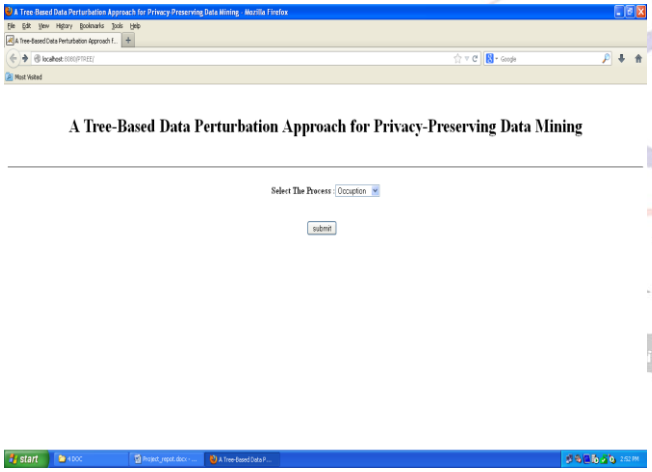

Figure 2 Application Index

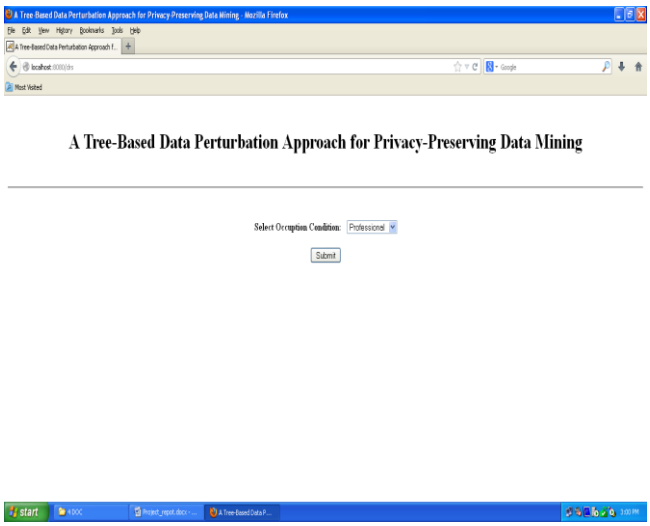

Figure 3 Occupation Page

X. CONCLUSION

We have tried to implement paper "DATA HIDING TECHNIQUES TO MAINTAIN PRIVACY IN REVEALED NUMERIC DATABASE", ref paper IEEE 20014 **'**An Efficient Approach for Privacy Preserving in Data Mining**'.** According to implementation conclusion is Perturbation tree will provide the privacy preserving on the sensitive data with high effectively and efficiently. Typical challenge of mining the confidential data (sensitive data) from datasets problem will be solved by perturbation tree. The proposed mechanism will give very high performances and low error rate compared with existing methods. To evaluate the mechanism, few test cases can be performed on real/demographic data for providing the protection and privacy on confidential data.

REFERENCES

[1] Agarwal, R. and Shrikant, R. "Privacy Preserving Data Mining", Proceeding of Special Interest Group on Management of Data. pp. 439 - 450, 2000.

[2] Jian Wang, Yong Cheng Lou, Yen Zh Jiajin Le, "A Survey on Privacy Preserving Data Mining", International Workshop on Database Technology and Application pp. III - 114, 2009.

[3] V.S Verykios, A.K Elmagarmid, E. Bertino, Y. Saygin and E. Dasseni, "Association Rule Hiding", IEEE Transaction Knowledge and Data Engineering, 16(4): 434 - 447, 2004.

[4] L. Sweeny, "K-Anonymity: A Model for Protecting Privacy", International Journal on Uncertainty, Fuzziness and Knowlwdge based Syatem, pp. 557 -570,2002.

[5] S. Vijayrani, A. Tamilarasi, M. Sampoorna, "Analysis of Privacy Preserving k-anonymity Methods and Techniques", Proceeding of the International Conference on Communication and Computational Intelligence, pp. 540 - 545, December 2010.

[6] K. Wang, P.S. Yu and S. Chakraborty, "Bottom Up Generalization: A Data Mining Solution to Privacy Protection", In International Conference on Data Mining, pp. 249 - 256,2004.

[7] B. Fung, K. Wang, P. Yu "Top Down Specialization" For International Conference on Data Engineering (lCDE' 05), pp. 205 - 216.

[8] E. Poovamal, M. Ponnavaikkp, "Task Independent Privacy Preserving Data Mining on Medical Data Set", International Conference on Advance Computing, Control and Telecommunication Technologies, pp. 815- 818,2009.

[9] H. Karagupta, S.Datta, Q. Wang and K. Sivakumar, "Random Data Peturbation Techniques and Privacy Preserving Data Mining", IEEE International Conference on Data Mining 2003.

[10] X. Zhang, H. Bi, " Research on Privacy Preserving Classification Data Mining on Random Peturbation", International Conference on Information Networking and Automation (ICINA), pp. 173 - 178,2010.

