Perturbation Based Privacy Preserving Data Mining

[1]M.Mohanrao, [2]Dr.S.Karthik
[1]Research Scholar, Bharathiar University Coimbatore, India
[2]Professor, SNS College of Technology, Coimbatore, India

***ABSTRACT**: Data mining process is used to find patterns among the dozens of fields in the large database. The main challenge in data mining is to maintain the privacy of confidential information. In order to share data while protecting privacy, data owner must achieve the goal of privacy preservation. Data perturbation has been used to protect the privacy of sensitive information. The proposed approach uses tuple values of the sensitive attribute to generate a normalized value which produces perturbed data. The resulting data records look very different from the original records and the distribution of data values is also different from the original distribution. The proposed mechanism gives low error rate compared with existing methods.*

***KEYWORDS**: Privacy Preserving, Data Mining, Data Perturbation, Sensitive Attribute.*

I. INTRODUCTION

Data mining technology has a goal of providing tools for intelligently transforming a large amount of data in knowledge relevant to users [1]. The derive Knowledge often convey in form of association rules, decision trees or clusters, allows one to find attractive patterns and regularities deeply buried in the data that are meant to promote decision making processes. Such a knowledge revelation process, however, can also return sensitive information about individuals, compromising the individual's right to privacy. Moreover, data mining techniques can reveal critical information about business transactions, compromising the free competition in a business setting. Thus, there is a strong need to restrict disclosure not only of confidential personal information but also of knowledge which is considered sensitive in a given context. For this reason, recently much research effort has been dedicated to addressing the problem of privacy preserving in data mining.

II. BACKGROUND

A number of newly proposed methods address the problem of privacy preservation by perturbing the data and reconstructing the distributions at an aggregate level in order to implement the mining. The typical additive reconstruction technique is column-based additive randomization. This type of techniques depend on the facts that Data owners may not want to uniformly protect all values in a record, thus a column-based value distortion can be used to reconstruct some sensitive columns. The condensation approach [2] is a typical multidimensional reconstruction technique, which intention at preserving the covariance matrix for multiple columns. Thus, some geometric properties such as the shape of decision boundary are well protected. Different from the randomization approach, it regenerates multiple columns as a whole to generate the entire reconstructed data set. As the reconstructed data set retains the covariance matrix, many existing data mining algorithms can be used directly to the reconstructed data set without requiring any change or new improvement of algorithms. In data swapping, technique confidentiality protection can be achieved by selectively exchanging a subset of attributes values between selected record pairs. Data swapping preserves the privacy of original sensitive information available at the record level. If the records are picked at random for each swap then it is called random swaps. It is difficult for an infiltrator to recognize distinct person or entity in the database because all the records are modified to the maximum level. The enviable properties of swapping technique are that it is simple and can be used only on sensitive data without disturbing non-sensitive data.

The present method simple additive noise (SAN) method [3] is adding the noise parameter which has mean zero and variance proportion parameter determined by the user to the initial confidential attribute then the result is a reconstructed value of a confidential attribute. The disadvantage of simple additive noise method is that the noise is independent of the scale of a confidential attribute.

To defeat the SAN method drawback next advanced approach is multiplicative noise (MN) [4], in this method, the confidential attribute is multiplied with the noise with mean one to get reconstructed value of a 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)[2],[5] the MA reconstructs 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. Related to SAN and MN, UMA produces bias in the variance of the confidential attribute, as well as in the relationships among attributes. Multivariate micro aggregation (MMA)[5],[6] groups data using a clustering technique that is based on a multidimensional distance measure.

As a result, the relationships within attributes are supposed to be better protected. 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 an approach based on the reconstruction trees[9], a kd-tree is a data structure for partitioning the and storing data.

III. PROPOSED ALGORITHM

The proposed approach, aim to achieved better result for privacy on the database than the existing system. The Proposed approach uses value distortion method for data perturbation. The proposed mechanism of reconstruction tree will handle the data partitioning on the data sets and subsets. The each subset must satisfy some minimum conditional values which store and from as leaf of the tree. This subset partitioning is a combination of the confidential and non confidential data. The proposed mechanism works and implements the approach of reconstruction tree [9], as one of the general methods like divide and conquer method.

System Architecture

The System Architecture provides the details of how the components or modules are integrated. Figure 1 is indicating the system architecture of the tree based data perturbation process. This architecture will give the complete description of input and outputs of each process. This process has several modules. They are Query Handler, Privacy Preserving, Data Perturbation and Result Evaluation.

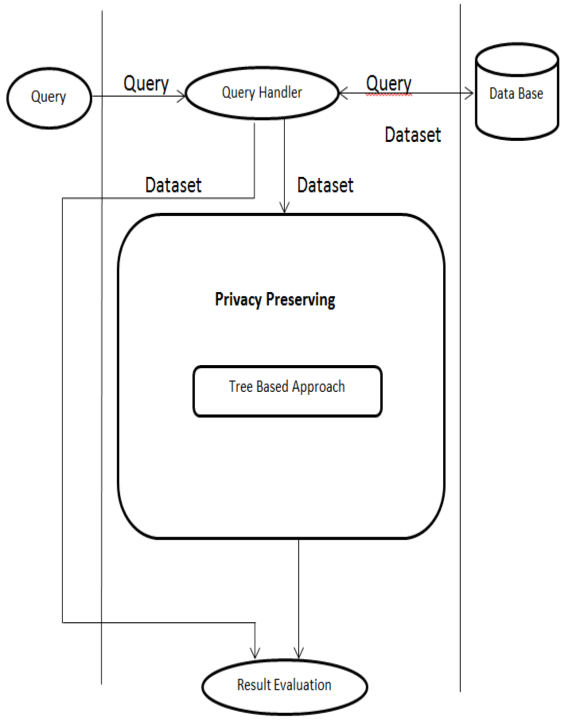


Fig.1.System Architecture

A. ***Query Handler***

The query handler is accepting the query data from the client and process the query to the database and fetching the datasets from the database.

B. ***Reconstruction Tree***

Reconstruction tree is proposed approach. This approach is using the divide and conquers technique. This technique will be using the following process; this approach accepts the datasets as input. These data sets will be divided into 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 fewer values. After completion of the division process, each leaf node attributes sensitive data will replacing with the average value and sending to the shareable person or other requested client.

C. Result Evaluation

The result evaluation is a process to find the error rate of different states in the data reconstruction of original data and the reconstructed data.

D. Splitting Criteria

It defines which attribute to implement for the splitting, for the numeric continuous attribute, and also determines which value is used for this splitting. Decision Tree algorithm ID3/ C4.5 uses information gain as a splitting pattern. The attribute with highest information gain will form the root of the tree and algorithm iteratively continues dividing the data to form a decision tree.

E. Dynamic Programming

This method will divide the data into the datasets and subsets, this datasets and subsets are conquering in the tree set approaching. This subset partitioning is a combination of the confidential and non-confidential data. The proposed approach works efficiently and effectively, due to the recursive divide and conquers technique adopted when dealing with the large data sets. A divide-and-conquer approach uses four basic steps to generating a super tree from an input dataset, S:

Step 1: Decompose the dataset into smaller, overlapping subsets.

Step 2: Construct trees on the subsets using the desired base reconstruction method.

Step 3: Merge the sub trees into a single tree on the entire dataset.

Step 4: Refine the resulting tree to produce a binary tree.

It is challenging to handle the sensitive data from the various private databases. Generally to solve this type of problem Data reconstruction technique can be used 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 the traditional system using various different approaches, this approaches are concentrating the mining of data as sensitive with confidential and non-confidential datasets.

To vary the confidential data from entire data is risk and the data of a confidential rules changes on the data access vendor. It is very costly operation on mining the data from databases and handling the sensitive data. To protect the data supplementary noise will be mixing with the actual data. To providing the results the encryption of data will be used with noise and decrypting the data to divide the actual data and noise data. The intended mechanism of reconstruction tree will manage the data partitioning of the data sets and subsets. The each subset must meet some minimum conditional values will stock and from as leaf of the tree. This subset partitioning is the union of the confidential and non-confidential data. The proposed mechanism performs the method of reconstruction tree, as one of the common method like divide and conquers technique. This technique will divide the data in the datasets and subsets, this datasets and subsets are conquering in the tree set approaching. This tree leaf sets are connected in from of average squared distances.

F. MATHEMATICAL MODEL

Reconstruction tree method consists numerous confidential and non-confidential datasets. The general idea of reconstruction tree is

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

Step 2: Let Q be the normalized data matrix at the current node. Compute the variance of each dimension, based on Q. Let j* be the dimension with the max. Variance.

Step 3: Find the median (mid-range) of attribute j*. Partition Q into two subsets (child nodes) based on the 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. Reconstruct the data by replacing these values with their Avg. Repeat this step for each leaf in the tree built in step-4. (If there are multiple attributes to be reconstructed, the Avg. of each attribute is used to replace the values of that attribute.)

Reconstruction Divide and conquer Algorithm

Input: original sensitive attribute value

Output: perturbed sensitive attribute value

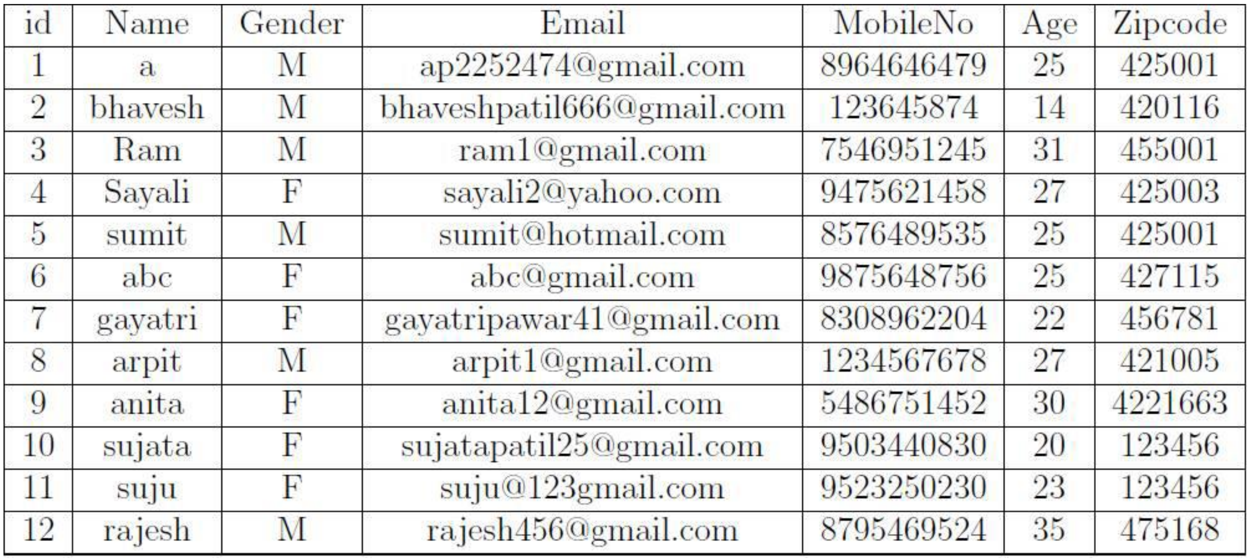
```
if (dataset.length> 3 &&dataset.length != 3)
For: w=0 to dataset.length do
wage = wage + Double.parseDouble(dataset[w][1])
end for
root = (wage/ (dataset.length))

left = new ArrayList<String>()
right = new ArrayList<String>()
For: d=0 to dataset.length
if( dataset[d][1])> root)
tmp = dataset[d][0]+dataset[d][1]+dataset[d][2]
left.add(tmp)
else
tmp = dataset[d][0]+dataset[d][1]+dataset[d][2]
right.add(tmp)
end for
conVar = makeStringArray(left)
divAndCon(conVar)
conVar = makeStringArray(right)
divAndCon(conVar)
```

The outcomes of this algorithm are the perturbed sensitive attribute value which replaces original sensitive attribute value in the database.

IV. SIMULATION RESULTS

Table 1 shows Original database. The database shows all available attribute information to the user without preventing sensitive attribute.

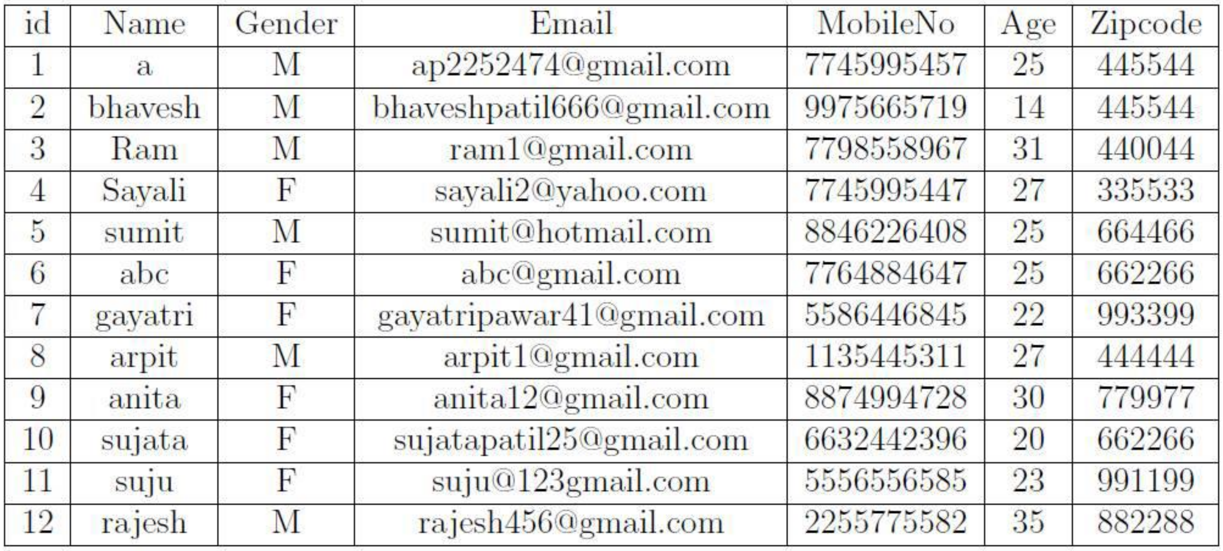
TABLE.1. ORIGINAL DATABASE



id	Name	Gender	Email	MobileNo	Age	Zipcode
1	a	M	ap2252474@gmail.com	8964646479	25	425001
2	bhavesh	M	bhaveshpatil666@gmail.com	123645874	14	420116
3	Ram	M	ram1@gmail.com	7546951245	31	455001
4	Sayali	F	sayali2@yahoo.com	9475621458	27	425003
5	sumit	M	sumit@hotmail.com	8576489535	25	425001
6	abc	F	abc@gmail.com	9875648756	25	427115
7	gayatri	F	gayatripawar41@gmail.com	8308962204	22	456781
8	arpit	M	arpit1@gmail.com	1234567678	27	421005
9	anita	F	anita12@gmail.com	5486751452	30	4221663
10	sujata	F	sujatapatil25@gmail.com	9503440830	20	123456
11	suju	F	suju@123gmail.com	9523250230	23	123456
12	rajesh	M	rajesh456@gmail.com	8795469524	35	475168

Table 2 represents perturbed database. This database applies perturbation method on sensitive attribute (Mobile No., Zip Code) and replaces original value with perturbed value. User able to access only perturbed database values, whenever admin can access original as well as perturbed database values.

TABLE.2. PERTURBED DATABASE

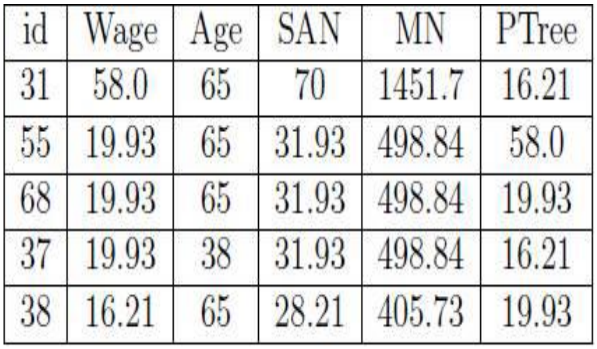


id	Name	Gender	Email	MobileNo	Age	Zipcode
1	a	M	ap2252474@gmail.com	7745995457	25	445544
2	bhavesh	M	bhaveshpatil666@gmail.com	9975665719	14	445544
3	Ram	M	ram1@gmail.com	7798558967	31	440044
4	Sayali	F	sayali2@yahoo.com	7745995447	27	335533
5	sumit	M	sumit@hotmail.com	8846226408	25	664466
6	abc	F	abc@gmail.com	7764884647	25	662266
7	gayatri	F	gayatripawar41@gmail.com	5586446845	22	993399
8	arpit	M	arpit1@gmail.com	1135445311	27	444444
9	anita	F	anita12@gmail.com	8874994728	30	779977
10	sujata	F	sujatapatil25@gmail.com	6632442396	20	662266
11	suju	F	suju@123gmail.com	5556556585	23	991199
12	rajesh	M	rajesh456@gmail.com	2255775582	35	882288

Error Rate Analysis

Following Table 3 shows error rate analysis between SAN, MN and perturbation tree method:

TABLE3.ERROR RATE ANALYSIS



id	Wage	Age	SAN	MN	PTree
31	58.0	65	70	1451.7	16.21
55	19.93	65	31.93	498.84	58.0
68	19.93	65	31.93	498.84	19.93
37	19.93	38	31.93	498.84	16.21
38	16.21	65	28.21	405.73	19.93

The operation was carried on data set to evaluate the proposed algorithm. Both regression and classification analysis is conducted, where the confidential attribute serves as the dependent or class variable. Dataset is randomly divided into two parts: approximately 75 percent for training, and 25 percent for testing.

The training set serves as the original set for reconstruction, while the testing set is not reconstructed. For classification analysis, two divisions of the confidential attribute are created by dividing its sorted numeric values at the median. Linear regression and the classifier were run on reconstructed data sets to build regression and classification models, and then computed errors using the reserved test sets. The error criterion for classification is the usual test error rate. The classification results differ more substantially, both when compared to that of the original data, and across the different reconstruction methods. Reconstruction trees produce the lowest error rate on data set.

V. CONCLUSION AND FUTURE WORK

Perturbation mechanism provides the privacy preserving on sensitive data with low error rate compared with existing methods. To evaluate the mechanism, few test cases can be performed on real data for providing the protection and privacy on confidential data. The typical challenge of mining the confidential data from datasets problem solved by perturbation tree.

Future work along this direction is to study how to apply two ways perturbation method, so user can able to retrieve original values from perturbed values. Also, perturbation method can apply on character attribute values to protect sensitive attribute values.

REFERENCES

[1]Kun Liu, Hillol Kargupta, IEEE, "Random Projection Based Multiplicative Data Reconstruction for Privacy Preserving Distributed Data Mining," IEEE Trans Knowl. Data Eng, VOL. 18, NO. 1, JANUARY 2008

[2]C.C. Aggarwal and P.S. Yu, “A Condensation Approach to Privacy Preserving Data Mining,” Proc. Ninth Int’l Conf. Extending Database Technology, pp. 183-199, 2004.

[3] N.R. Adam and J.C. Wortmann, “Security-Control Methods for Statistical Databases: A Comparative Study,” ACM Computing Surveys, vol. 21, no. 4, pp. 515-556, 1989.

[4]R. Agrawal and R. Srikant, “Privacy-Preserving Data Mining,” Proc. 2000 ACM SIGMOD Int’l Conf. Management of Data, pp. 439- 450, 2000

[5] J. Domingo-Ferrer and J.M. Mateo-Sanz, “Practical Data-Oriented Micro aggregation for Statistical Disclosure Control,” IEEE Trans. Knowledge and Data Eng., vol. 14, no. 1, pp. 189-201, 2002.

[6] J. Domingo-Ferrer and V. Torra, “Ordinal, Continuous and Heterogeneous k-Anonymity through Micro aggregation,” Data Mining and Knowledge Discovery, vol. 11, no. 2, pp. 195-212, 2005.

[7] Agrawal R., Srikant R, "Privacy-Preserving Data Mining," ACM SIGMOD Con- ference, 2009