Reconstruction Methods for Providing Privacy in Data Mining

Vishal R. Shinde
Asst. Prof.,Computer Dept.
SSJCET,Asangaon,
Thane,India
mailme.vishalshinde@rediffmail.com

Rahul P. Shukla
Student (B.E. Comp)
SSJCET,Asangaon,
Thane,India
rahulshukla1290@gmail.com

Jagdish V. Patil
Student (B.E. Comp)
SSJCET,Asangaon,
Thane,India
jagdish9764@gmail.com

Nitin S. Pawar
Student (B.E. Comp)
SSJCET,Asangaon,
Thane,India
nitus.pawar@gmail.com

***Abstract*—** Data mining is the process of finding correlations or patterns among the dozens of fields in large database. A fruitful direction for data mining research will be the development of techniques that incorporate privacy concerns. Since primary task in our paper is that accurate data which we retrieve should be somewhat changed while providing to users. For this reason, recently much research effort has been devoted for addressing the problem of providing security in data mining. We consider the concrete case of building a decision tree classifier from data in which the values of individual records have been reconstructed. The resulting data records look very different from the original records and the distribution of data values is also very different from the original distribution. By using these reconstructed distribution we are able to build classifiers whose accuracy is comparable to the accuracy of classifiers built with the original data.

Keywords- *reconstruction trees, privacy, Data mining, divide and conquer method*

*****

I. INTRODUCTION

Data mining technology has been developed with the goal of providing tools for automatically and intelligently transforming large amount of data in knowledge relevant to users. The extracted Knowledge, often expressed in form of association rules, decision trees or clusters, allows one to find interesting patterns and regularities deeply buried in the data that are meant to facilitate decision making processes. Such a knowledge discovery 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 prevent 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 devoted to addressing the problem of privacy preserving in data mining.

As a result, several data mining techniques, incorporating privacy protection mechanisms, have been developed based on different approaches. For instance, various sanitization techniques have been proposed for hiding sensitive items or

Patterns that are based on removing reserved information or inserting noise into data. Privacy preserving classification methods, instead, prevent a miner from building a classifier able to predict sensitive data. Additionally, privacy preserving clustering techniques have been recently proposed, which distort sensitive numerical attributes, while preserving general features for clustering analysis.

Many algorithms tried to extract the data without directly accessing the original data and guarantees that the mining process does not get information to reconstruct the original data.This proposed work considers a tree based data reconstruction approach to provide the privacy to individual disclosure information or sensitive data. It is challenging to provide privacy to individual sensitive data when organization releases the data to the third party to mine the data or data mining. To this many privacy preserving approaches are proposed in recent years but those algorithms or approaches are deals with the small data sets and failed in maintaining the relationships between the attributes.

II. RELATED WORK

A. NADDITIVE RECONSTRUCTION

The typical additive reconstruction technique (Agrawal and Srikant, 2000) is column-based additive randomization. This type of techniques relies on the facts that 1) Data owners may not want to equally protect all values in a record, thus a column-based value distortion can be applied to reconstruct

some sensitive columns. 2) Data classification models to be used do not necessarily require the individual records, but only the column value distributions (Agrawal and Srikant, 2000) with the assumption of independent columns. The basic method is to disguise the original values by injecting certain amount of additive random noise, while the specific information, such as the column distribution, can still be effectively regenerated from the reconstructed data. A typical random noise addition model (Agrawal and Srikant, 2000) [1] can be precisely described as follows. Treat the original values (x1, x2, . . . , xn) from a column to be randomly drawn from a random variable X, which has some kind of distribution. The randomization process changes the original data by adding random noises R to the original data values, and generates a reconstructed data column Y, Y = X + R. The resulting record (x1+r1, x2+r2, . . . , xn+rn) and the distribution of R are published. The key of random noise addition is the distribution reconstruction algorithm (Agrawal and Srikant, 2000; Agrawal and Aggarwal, 2002) that recovers the column distribution of X based on the reconstructed data and the distribution of R. While the randomization approach is simple, several researchers have recently identified that reconstruction-based attacks are the major weakness of the randomization approach. In particular, the spectral properties of the randomized data can be utilized to separate noise from the private data. Furthermore, only the mining algorithms that meet the assumption of independent columns and work on column distributions only, such as decision-tree algorithms, and association-rule mining algorithms, can be revised to utilize the regenerated column distributions from reconstructed datasets.

B. Condensation-based Reconstruction

The condensation approach [2] is a typical multidimensional reconstruction technique, which aims at preserving the covariance matrix for multiple columns. Thus, some geometric properties such as the shape of decision boundary are well preserved. Different from the randomization approach, it regenerate multiple columns as a whole to generate the entire reconstructed dataset. As the reconstructed dataset preserves the covariance matrix, many existing data mining algorithms can be applied directly to the reconstructed dataset without requiring any change or new development of algorithms. The condensation approach can be briefly described as follows. It starts by partitioning the original data into record groups. Each group is formed by two steps randomly selecting a record from the existing records as the centre of group, and then finding the nearest neighbours of the centre to be the other members.

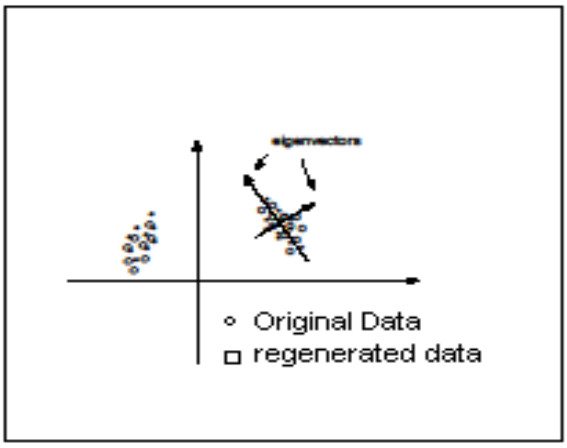

Fig 1: Condensation approach

The selected k records are removed from the original dataset before forming the next group. Since each group has small locality, it is possible to regenerate a set of k records to approximately preserve the distribution and covariance. The record regeneration algorithm tries to preserve the eigenvectors and eigen values of each group, as shown in Figure. The authors demonstrated that the condensation approach can well preserve the accuracy of classification models if the models are trained with the reconstructed data. However, it has been observed that the condensation approach is weak in protecting data privacy. As stated by the authors, the smaller the size of the locality is in each group, the better the quality of preserving the covariance with the regenerated k records is. However, the regenerated k records are confined in the small spatial locality.

C. Data swapping

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 record level. If the records are picked at random for each swap then it is called random swaps. It is difficult for an intruder to recognize particular person or entity in database, because all the records are altered 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 existing method simple additive noise (SAN) method [3] 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 reconstructed 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

 ISSN: 2321-8169
2891 – 2894

(MN), [4] in this method the confidential attribute is multiplied with the noise with mean one to get reconstructed 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)[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. 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) [5],[6] groups data using a clustering technique that is based on a multidimensional 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 reconstruction trees[9], 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 reconstructed using the subset average. Since the data are partitioned based on the joint properties of multiple confidential and non-confidential attributes, the relationships between attributes are expected to be reasonably preserved. Further, the proposed method is computationally efficient.

III. PROPOSED WORK

A. 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.

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 accept the data sets as input.
This data sets will be 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.

C. 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.

Those approaches are implemented by following sub modules of this mechanism. The proposed approach is Reconstruction

Tree and the existing systems are Simple additive noisy and multiplicative noisy.

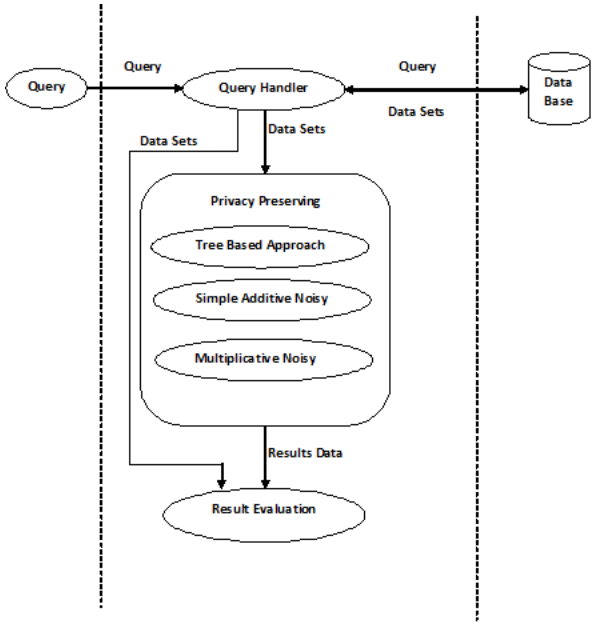

Fig 2: System Architecture

D. Simple additive Noisy

The simple additive noisy (SAN) method will be 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.

E. Multiplicative Noisy

The multiplicative noisy (MN) also existing 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.

F. 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. 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 first 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 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 reconstructed data.

By using approach the information loss of reconstruction will be found.

To calculate the bias in standard deviation proposed system is using of the original data and reconstructed data. By finding this value, get the loss of information in reconstructed 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 reconstruction.

G. Splitting Criteria

It decides which attribute to use for the splitting, and 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 splitting criterion. The attribute with highest information gain will form the root of the tree and algorithm iteratively continues splitting the data to form a decision tree.

H. Dynamic Programming

This method will divide the data in the datasets and subsets, this datasets and subsets are conquering in the tree set approaching. This subset partitioning is 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 method uses four basic steps to construct a super tree from a given 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 data bases. 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 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 additional noise will be merging with the actual data. To producing the results the encryption of data will be used with noise and decrypting the data to divide the actual data and noise data. The proposed mechanism of reconstruction tree is that 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. The proposed mechanism works and implements the approach of reconstruction tree, as one of the general method like divide and conquer method.

This method 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 linked in from of average squared distances.

IV. MATHEMATICAL MODEL

Reconstruction tree mechanism consist the various confidential and non-confidential data sets. 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 Z be the normalized data matrix at the current node. Compute the variance of each dimension, based on Z. 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. 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.)

ISSN: 2321-8169
2891 – 2894

ASD (AVG.SQ.DIST) = $1/N \sum_{i=1}^{N} (yi - xi)^2$

Smaller disclosure risk, higher info. Loss.

BIM (BIAS IN MEAN) = $(\overline{yi}-\overline{xi})/\ \overline{x}$

BISD (Bias in std. Deviation) =$(Sy - Sx)/Sx$

Sx and Sy are Std. deviation of original and reconstruct confidential values reply.
Smaller BIM and BISD are desirable. Measure univariate info. Loss due to reconstruction
Apply linear regression and C4.5 classifier on reconstructed data to build regression and classification model and computed errors.

MAE (Mean Absolute Error) = $1/M \sum_{i=1}^{M} |xi - \hat{}xi|$
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 on the regression model.

Since MAE measures the distance between the predictions of the model built from the reconstructed data and the (unreconstructed) test data, a smaller MAE value is desirable.

Finding Mean error

```
x_mean = x_mean + x[i]
y_mean = y_mean + y[i]

x_mean = x_mean/x.length
y_mean = y_mean/y.length

BIM = y_mean - x_mean
BIM = BIM/x_mean
```

Finding Standard deviation

```
x_mean += x[i]
y_mean += y[i]

x_mean = x_mean/x.length
y_mean = y_mean/y.length
BIM = (y_mean - x_mean)/x_mean
x[i] = (x[i]-x_mean)
y[i] = (y[i]-y_mean)

sdxx = sdxx + x[i]*x[i]
sdyy = sdyy + y[i]*y[i]

sdxx =  sdxx/x.length
sdyy =  sdyy/y.length

BISD = sdyy-sdxx
BISD = BISD/sdxx
```

Algo. Reconstruction Divide and conquer

```
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)
```

Algo for MAE

```
for  i =0 to maxN
        sumx +=  x[i]
        sumy +=  y[i]
    sumx2 = sumx * sumx
end for

xbar = sumx/maxN
ybar = sumy/maxN

for  i =0 to maxN
    mae += (x[i]-xbar)
        end for
```

ISSN: 2321-8169
2891 – 2894

V. RESULTS AND DISCUSSION

Table 1: Error Rate Analysis: SAN, MN and P-TREE

EmpId	Wages	Age	SAN	MN	P-tree
430	10.30	53	22.30	123.64	10.29
431	18.39	39	30.39	220.75	18.26
432	25.0799	54	37.084	301.06	25.08
433	12.636	28	24.054	144.651	12.06
434	22.399	35	34.404	268.895	22.455
435	8.75	48	20.754	105.037	8.690
436	22.199	32	34.204	266.49	22.2
437	17.25	29	29.254	207.07	17.375
438	6.410	26	18.414	76.947	6.375
439	8.060	36	20.064	96.75	8.095

Table 2: Experimental Results

Approach	SAN	MN	P-Tree
Time Complexity	0	47	0
BIM	1.0	11.783	0.0
BISD	-1.1 E-16	11.7	-0.02
MAE	1.7 E-15	1.77 E-15	1.7 E-15

The experiment was conducted on data set to evaluate the proposed algorithm. Both regression and classification analysis is performed, where the confidential attribute serves as the dependent or class variable. Data set is randomly

divided into two parts: approximately 75 percent for training, and 25 percent for testing. The training set serve as the

original set for reconstruction, while the testing set are not reconstructed. For classification analysis, two categories of the confidential attribute are formed by dividing its sorted numeric values at the median. Linear regression and the C4.5 classifier were run on reconstructed data sets to build regression and classification models, and then computed errors using the reserved test sets. The error measure for classification is the usual test error rate. The classification results differ more substantially, both when compared to that from the original data, and across the different reconstruction methods. Reconstruction trees yield the lowest error rate on data set.

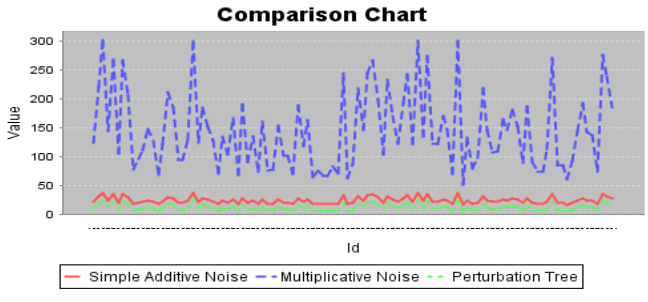

Fig : Comparative Graph By Selecting Field Occupation-Professional

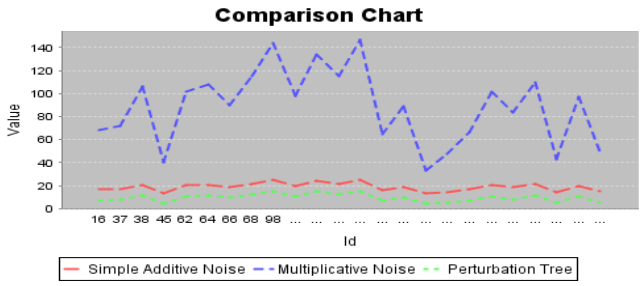

Fig : Comparative Graph By Selecting Field Sector-Manufacturing

VI. CONCLUSION AND FUTURE SCOPE

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. Reconstruction tree performances rate is

comparing with the existing method on various levels like regression, classification, bias in mean (BIM) and bias in standard variances (BISD). Reconstruction tree is having low error rate on providing the confidential data. Reconstruction tree will be providing 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 reconstruction tree.

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 andHeterogeneous k-Anonymity through Micro aggregation," Data Mining and Knowledge Discovery, vol. 11, no. 2, pp. 195-212, 2005.

[7] Lambodar Jena, Ramkrushna Swain, IEEE,"Comparative study on Privacy Pre- serving Association Rule Mining Algo,"I nternational Journal of Internet Comput- ing, Vol.1, 2011.

[8] . Verykios, Ahmed K. Elmagarmid, Bertino Elisa, Yucel Saygin, and Dasseni Elena, "Association Rule Hiding,"IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING,2008

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

[10] Aggarwal C. C., Yu P. S.:,"A Condensation approach to privacy preserving data mining.",EDBT Conference, 2008.

[11] Aggarwal C. C,"On Randomization, Public Information and the Curse of Dimensionality," ICDE Conference, 2007.

[12] Keke Chen1, Ling Liu2,"Geometric Data Reconstruction for Privacy Preserving Outsourced Data Mining," Oct 23, 2010

[13] P.Kamakshi, Dr. A.Vinaya Babu,"A Novel Framework to Improve the Quality of Additive Reconstruction Technique," International Journal of Computer Applications (0975 8887) Volume 30 No.6, September 2011

[14] Xiao-Bai Li and Sumit Sarkar, A Tree-Based Data Reconstruction Approach for Privacy-Preserving Data Mining, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING,VOL. 18, NO. 9,

[15] S. L. Hansen and S. Mukherjee, A Polynomial Algorithm for Optimal Univariate Microaggregation, IEEE Trans. Knowledge and Data Eng., vol. 15, no. 4, pp. 1043-1044, July/Aug. 2003.