ISSN:2231-0711
www.ijcset.net

A New Approach To Maintain Privacy And Accuracy In Classification Data Mining

G.Rama Krishna[#1], G.V.Ajaresh*[2], I.Jaya Kumar Naik*[3], Parshu Ram Dhungyel*[4], D.Karuna Prasad*[5]

[#]*Professor, Vaddeswaram, K.L.C.E, Andhra Pradesh, India*
[1]ramakrishna_10@yahoo.com
[*]*IV/IV B.Tech Students, Vaddeswaram, K.L.C.E, Andhra Pradesh, India*
[2]ajaresh240@gmail.com
[4]sharmad99@gmail.com

***Abstract*- Privacy preserving classified data mining is one of the latest technical fields of data mining in recent years. Transforming the original data in order to maintain accuracy if we apply classification mining on original and transformed data is the key point of the privacy preservation. This paper proposes a privacy preserving method based on the random perturbation matrix. This method is suitable to the data of the character type, the Boolean type, the hierarchical type etc.**

***Key words* - privacy preservation, decision tree; Random perturbation matrix, privacy breaches**

I. INTRODUCTION

It is well known that data mining is a powerful data analysis tool enabling discovery of useful patterns in several applications. The availability of large data warehouses and associated technologies reveal that the usefulness of data requires preservation of individual key attributes such as patient's condition information, customer preferences, personal background information etc. If we release the original data directly to the miner, it will inevitably produce private information of the customer. Therefore, how to do mining without sacrificing privacy is the main issue in data mining. Privacy preserving data mining (PPDM) deals with this issue. Currently, there are many approaches of privacy preserving data mining to transform the original data. The methods of privacy preserving data mining are evaluated based on applicability, privacy protection metric, the accuracy, computation, etc.

The existing methods of privacy preserving data mining focus only on the privacy protection metric or the accuracy of mining results. This paper presents a new approach based on the random perturbation matrix, which is applicable to any data distribution. It can be applied to a variety of data types through the data fields by encoding strategy. A random perturbation matrix is generated for each property to increase the intensity of privacy protection, introducing r-amplifying method to improve the random perturbation matrix which prevents from privacy breaches after the generation of transformed data. The choice of single-attribute random perturbation matrix can reconstruct the original data set distribution, which increases the accuracy of mining results for the computation of decision tree.

II. LITERATURE SURVEY

In recent years, data mining is considered as a threat to privacy because of the rapid growth of electronic data maintained by corporations. This has lead to increased concerns about the privacy of the individuals. A number of techniques have been proposed for transforming the data in such a way so as to preserve privacy. Privacy-preserving data mining finds numerous applications which are naturally supposed to be "privacy-violating" applications. The key is to design methods which continue to be effective, without compromising privacy. Most methods for privacy computations use some form of transformation on the data in order to perform the privacy preservation. Some examples of such techniques are as follows:

a) *The randomization method*:

The randomization method is a technique for privacy-preserving data mining in which noise is added to the data in order to mask the attribute values of records. The noise added is sufficiently large so that individual record values cannot be recovered. Therefore, techniques are designed to derive aggregate distributions from the perturbed records. Subsequently, data mining techniques can be developed in order to work with these aggregate distributions.

b) *The k-anonymity model*:

The k-anonymity model solves the problem of indirect identification of records from public databases. This is because combinations of record attributes can be used to exactly identify individual records. In the k-anonymity method, we reduce the granularity of data representation with the use of techniques such as generalization and suppression. This granularity is reduced sufficiently that any given record maps onto at least k other records in the data.

c) *Distributed privacy preservation*:

In many cases, individual entities may wish to derive aggregate results from data sets which are partitioned across these entities. Such partitioning may be horizontal (when the records are distributed across multiple entities) or vertical (when the attributes are distributed across multiple entities). While the individual entities may not desire to share their entire data sets, they may consent to limited information sharing with the use of a variety of protocols. The overall effect of such methods is to maintain privacy for each individual entity, while deriving aggregate results over the entire data.

III PRIVACY PRESERVING CLASSIFICATION DATA MINING BASED ON RANDOM PERTURBATION

Problem Definition

This paper concentrates on developing a solution to the problem of how to convert the original data set into transformed data set such that the results of data mining on the original data set and the perturbated data set are almost similar reflecting accuracy.

Data preprocessing

This method can deal with character type, Boolean type, hierarchical type and number types of discrete data, and to facilitate conversion of data sets, it is necessary to preprocess the original data set. The data preprocessing is done in three steps, forming discrete data set, attribute coding, coded data set. This paper uses the method of average region to disperse the continuous data.

The Discrete data is calculated using the formula:

A (max) - A (min)/n = length

A is continuous attributes, n is the number of discrete values, length is the length of the discrete interval. When the interval length is a decimal, round to the nearest integer, the first interval of discrete begin from A(min), the last interval is A(max). In this paper, the attributes of integer type are taken as continuous attributes, Considering Table I as an example, the continuous attributes are Employee-id, year-of-joining and salary.

Table 1
Employee Data Set

EID	DNo	Year 0f joining	Job	sex	Salary
20	B01	1973	Manager	M	41250
100	E21	1979	Manager	M	46150
120	B00	1963	Clerk	M	29250
140	C01	1971	Analyst	F	28420
150	D11	1972	Designer	F	25280
220	D11	1968	Designer	F	29840
230	D21	1966	Clerk	M	12180
240	D21	1969	Clerk	M	19180
290	E11	1980	Operator	M	15240

Table 2
Discrete Data Set

EID	Dno	Year	Job	Sex	Salary
20-50	B01	1972-74	Manager	M	38606-42380
82-111	E21	1978-80	Manager	M	42381-46155
112-141	B00	1963-65	Clerk	M	27281-31055
112-141	C01	1969-71	Analyst	F	27281-31055
142-171	D11	1972-74	Designer	F	23506-27280
202-231	D11	1966-68	Designer	F	27281-31055
202-231	D21	1966-68	Clerk	M	12180-15955
232-261	D21	1969-71	Clerk	M	15956-19730
262-291	E11	1978-80	Operator	M	12180-15955

Table 3
Attribute Coding

EID	Code	Year	Code	Salary	Code
20-50	1	1972-74	1	38606-42380	1
82-111	2	1978-80	2	42381-46155	2
112-141	3	1963-65	3	27281-31055	3
142-171	4	1969-71	4	23506-27280	4
202-231	5	1966-68	5	12180-15955	5
232-261	6			15956-19730	6
262-291	7				

Table 4
Attribute Coding

D No	Code	Job	Code	Sex	Code
B00	1	Manager	1	Male	1
B01	2	Clerk	2	Female	2
C01	3	Analyst	3		
D11	4	Designer	4		
D21	5	Operator	5		
E11	6				
E21	7				

Table 5
Coded Data Set

E ID	D No	Year	Job	Sex	Salary
1	2	3	5	1	4
2	3	1	1	2	2
1	5	3	2	1	5
1	2	3	3	1	2
2	7	2	3	2	3
3	5	3	1	2	3
2	3	4	3	2	4
3	4	1	5	1	1
1	6	3	4	2	5

RANDOM PERTURBATION METHOD

1) Related definitions

a) The prior probability:

The probability of guessing the original value without any information about the randomized value is known as the prior probability. Denoted as P[Q(x)], where Q(x) is a property of the original value.

b) The posterior probability:

The probability of guessing the original value after knowing the randomized value is known as the posterior probability. Denoted as P[Q(x)/R(x)=y], where R(x) is the randomization operator.

c) Privacy breaches:

It is a situation when, for some property Q (x_i), the disclosure of y_i to others significantly increases the probability of this property. If it is important that property Q (x_i) of its private information is not disclosed, then a significant increase in probability may be a violation of privacy

Let p_1 and p_2 be two probabilities, such that p_1 corresponds to the notion of "very unlikely" (e.g., p_1=0.01) whereas p_2 corresponds to "likely" (e.g., p_2=0.5). Let $Q_1(x)$ and $Q_2(x)$ be two properties. 1.

Straight (p_1-to-p_2) privacy breach: 2.

We say that there is a straight privacy breach with respect to Q_1 property if

$$P[Q_1(x)] \le p_1 \text{ and } P[Q_1(x)/R(x)=y] \ge p_2$$

Inverse (p2-to-p1) privacy breach:

We say that there is a inverse privacy breach with respect to Q_2 property if

$$P[Q_2(x)] \ge p_2 \text{ and } P[Q_2(x)/R(x)=y] \le p_1$$

d) Amplification:

The test is based on comparing the operator's transitional probabilities p[x->y] for the same y belongs to Vy but different x belongs to Vx. If all of the x-values are reasonably likely to be randomized into a given y, then revealing "R(x)=y" does not tell too much about x.

A randomization operator R(x) is at most r-amplifying for y if

For all x_1, x_2 $p[x_1\text{->}y] / p[x_2\text{->}y] \le r$, $r \ge 1$

We will select r such that the privacy breaches do not occur.

$$\frac{p2(1-p1)}{p1(1-p2)} > r$$

2) Set the single-attribute random perturbation

The value of single attribute random perturbation matrix means the probability that each value transform into other value in the attribute domain, which determine the intensity of privacy protection of privacy preserving mining and the accuracy of mining results, directly related to the method good or bad, which is key of privacy preserving classification data mining based on random perturbation.

In order to avoid the straight and inverse privacy breaches, we will select 'r' such that

$$\frac{p[x1\to y]}{p[x2\to y]} \le r < \frac{p2(1-p1)}{p1(1-p2)}$$

Now the method chooses r positive definite symmetric matrix as the single attribute perturbation matrix. Firstly it requires user to give the threshold the rate before the test and after of every attribution, request $0< p_l< p_2<1$.

Taking a random r from $p_2(1\text{-}p_1)/p_1(1\text{-}p_2)>r>=1$, the random perturbation matrix is generated as follows.

A_{ij} = rx if i=j

= x if i! =j where $x = \frac{1}{r+(|Su|-1)}$

Su is the domain of property A.

That is:

$$A_{ij} = x\begin{bmatrix} r & 1 & 1 & \dots \\ 1 & r & 1 & \dots \\ 1 & 1 & r & \dots \\ \dots & \dots & \dots & \dots \end{bmatrix}$$

The value of Aij denotes the probability of changing i to j. The matrix has following characters:

1) The sum of the element each is 1.

2) Matrix elements satisfy the conditions of r-amplifying

3) The generation of multi-property combined Random perturbation matrix

a) The ideology of generating multi-property combined Random perturbation matrix

Taking the generation of two properties combined perturbation matrix as an example to illustrate the generation of multi-property joint random perturbations matrix.

Al attribute has n different values, the random perturbation matrix of attribute Al is the n-order square matrix R (A_1), A_2 attribute has m different values, the random perturbation matrix of attribute A_2 is the m-order square matrix R (A_2), the idea of generating A_1, A_2 joint disturbance matrix R (A_1, A_2) of n * m-order attributes is that each element aij of R (A_2) is multiplied by R (A_1) as the i-n line in R (A1, A2), j-n column elements. Random perturbation matrix of attributes A1, A2 is in Figure 1, the joint disturbance matrix of property A1, A2 is in Figure 2.

$$R(A_1) = x_1\begin{bmatrix} r_1 & 1 & 1 & \dots \\ 1 & r_1 & 1 & \dots \\ 1 & 1 & r_1 & \dots \\ \dots & \dots & \dots & \dots \end{bmatrix}$$

$$R(A_2) = x_2\begin{bmatrix} r_2 & 1 & 1 & \dots \\ 1 & r_2 & 1 & \dots \\ 1 & 1 & r_2 & \dots \\ \dots & \dots & \dots & \dots \end{bmatrix}$$

Figure I. Random perturbation matrix of attributes A" A,

b) The nature of multi-attribute combined random perturbation matrix

Nature 1: The inverse of Multi-attribute joint random perturbation matrix is equal to the joint of inverses of single property random perturbation matrix.

$$P(A_1,A_2,.....,A_k)^{-1} = P(A_1^{-1},A_2^{-1} ... ,A_k^{-1})$$

Nature 2:

$$T(A_1,A_2,....,A_k) * P(A_1,A_2,.....,A_k) = D(A_1,A_2 ... ,A_k)$$

T (A_1,A_2, ... ,A_k) means the attribute of the original data {A_1,A_2, ... ,A_k},which is the line matrix of support count of different joint value, D(A_1, A_2, ... ,A_k) means the attribute of the data {A_1,A_2, ... ,A_k} after transform, which is the line matrix of support count of different joint value.

4) The data perturbation

The data perturbation is that the value of each property is transformed into other value of the property domain by given probability. This paper takes independent perturbation method on each attribute. Assume that attribute A is perturbed on the data set, at first, give coded data set, the records number s of the data set perturbed and the perturbation matrix R (A) of property A, then the data set is perturbed by the following perturbation algorithm.

Perturbation Algorithm
input: data set records |s| the disturbance R(A) of attribute A
output: attribute A's data field v[n] after perturbation

```
n= |s| ;
for i=1 to n
 {
   U=u[i]; /*u[i]is the original date of the record i at attribute A in the data set*/
   a=random (0,1); //a is a random number from 0 to 1
   for j=1 to R(A).length
      f(u,j)={R(Aul),R(Au2), ... ,R(Auj)} ;
/* f(U , j) is Probability distribution of U convert into j*/
      for k=1 to R(A).length
       {
          If (F(k-1)<a<=F(k))
       v[i]=k; //F(k)is the distillation function of f(U,j)
          Break;
       }
Break;
}
```

C. Privacy preserving classification data mining algorithm

1) Building decision tree by the perturbation data set

In the classification mining, the most typical classification method bases on the decision tree, which is a tree structure similar to the flowchart. Each internal node represents a test on an attribute, each branch represents a test output, and leaf node of tree represents a class or class distribution. The top node is the root node.

This paper uses the classical mining algorithm ID3 to construct the decision tree.

The key of building decision tree is to find the largest property information as a branch node from each data set to which each branch corresponding. With the transformed data set and multi-property joint disturbance matrix, it can calculate attribute information. The method is as follows:

Set a data set S, the attribution set of S is {A_1, A2,...,A_k}, A_k is the label attribution.

a) Request the attribution of Maximum information gain of the root Through the formula $T(A_k)*P(A_k)=D(A_k)$, we can calculate the entropy E(S) of label attribution A_k.

Through the formula $T(A,A_k)*P(A,A_k)=D(A,A_k)$, we can calculate the entropy E(S,A) of every label attribution.

Through the formula Gain(S,A)=£(S)-£(S,A) , we can calculate information gain of the attribution.

b) As is known that the root node is A, the value ofAl is the aI data set SI, calculate information gain of the attribution ofaI branch Split nodes.

Through the formula $T(A(a_1),A_k)*P(A(a_1),A_k)=D(A(a_1),A_k)$, $A_1(a_1)$ denotes the value of attribution Al is a_1, we can calculate data set S1'Iable attribution A_k, entropy is $E(S_1)$.

Through the formula

$T(A_1(a_1),A,A_k)*P(A_1 (a_1),A,A_k)=D(A_1 (a_1),A,A_k)$, we can calculate the entropy E(SI,A) of every label attribution over the data set S1.

Through the formula Gain (S_1, A) =E (s_1)-E (S_1, A), we can calculate information gain of the attribution.

c) In the same way, we can calculate lower node information gain of the attribution. Until all the records of the tag attributes in Generated data set are the same or All properties split off is over.

2) Decision tree pruning

When the decision Tree is created, due to data noise and outliers, many branches reflect the abnormal of data set. Pruning algorithms deal with this over-adaptation problem. Typically, this method uses a statistical measure, cutting the least reliable branches, to improve the speed and accuracy of classification.

There are usually two pruning algorithms.

a) Pre-pruning algorithm

Pre-pruning algorithm is completed before the tree growth process carried out. Friedman proposed the method of restrictions of the smallest node, when the number of nodes is less than the threshold value k, the growth of the node will be stopped.

b) Post-pruning algorithm

Post-pruning algorithm is to be done when the decision tree growth process is completed, which allow decision tree grow excessively, then according to certain rules, cutting the leaf nodes or branch of the decision tree which are not generally representative.

This paper uses the post-pruning algorithm.

3) Extraction rules from decision tree

The classification knowledge which indicated by decision tree can be extracted and explained as IF-THEN type. The path from the tree's root to any leaf nodes forms a classification rule; the conjunction of property value which is formed by a path along the tree constitutes the antecedent of classification rule
("IF" section), the category that labeled leaf nodes constitutes the consequent of classification rules
("THEN" part). As the extraction of classification rules are encoded, and then the code is translated into attribute names according to the attribution code tables.

IV CONCLUSION

The privacy preserving data mining method mainly depends on the privacy protection and the accuracy of mining results. The methods of most literatures sacrifice the privacy in exchange of a high accuracy or sacrifice accuracy for a high privacy protection metric. This paper's new approach allows ensuring limitations on privacy breaches for a randomization operator, without any knowledge about the prior distribution. Also it is applicable to any property of client's private information.

The experimental result shows that the method balances the intensity of privacy protection and the accuracy of mining results. Additionally computation is reduced greatly.

ACKNOWLEDGMENT

We would like to thank Prof S.Venkateswarlu, Head of the Department, Computer Science & Engineering, Kluniversity, Vaddeswaram for his encouragement and motivation to write this paper. Also we are grateful to Prof G.Rama Krishna, (CSE), Kluniversity, Vaddeswaram for guiding us in writing this paper.

REFERENCES

[1]. Alexandre Evmievski, Johannes Gehrke, Ramakrishnan Srikant. "Limiting Privacy Breaches In Privacy Preserving Data Mining" In: PODS. 2003.

[2]. Shipra Agrawal, Jayant R. Haritsa, "A Framework For High-Accuracy Privacy-Preserving Mining" In: The 21st International Conference on Data Engineering (ICDE). 2005, pp. I 93-204.

[3]. "Privacy-Preserving Data Publishing: A Survey Of Recent Developments" BENJAMIN C. M. FUNG Concordia University, Montreal KE WANG Simon Fraser University

[4]. "The Impact of Data Perturbation Techniques on Data Mining" by Rick L. Wilson, Department of Management Science and Information Systems, Oklahoma State University

[5]. "Privacy Preserving Decision Tree Mining from Perturbed Data" by Li Liu Global Information Security eBay Inc.

[6]. "Privacy-Preserving Data Mining:Models And Algorithms" Edited by CHARU C. AGGARWAL IBM T. J. Watson Research Center, Hawthorne, NY 10532 PHILIP S. YU University of Illinois at Chicago, Chicago, IL 60607