Preserving the Privacy and Sharing the Data Using Classification on Perturbed Data

B. V. HARITA'
M. Tech Student, Computer Science,
GITAM University, Rushikonda, Vishakapatnam,
Andhra Pradesh, INDIA
venkataharitabillapati@gmail.com

P.G.Chaitanya,
Mtech student, Computer Science,
GITAM University, Rushikonda, Vishakapatnam,
Andhra Pradesh, INDIA
chytucool@gmail.com

R.L.Diwakar
Assistant professor,Computer Science,
LENDI College,Visakapatnam,
Andhra Pradesh,INDIA

Abstract:

Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. Even though data mining successfully extracts knowledge to support a variety of domains, it is still a challenge to mine certain kinds of data without violating the data owner's privacy. Due to increasing concerns related to privacy, various privacy-preserving data mining techniques have been developed to address different privacy issues. These techniques usually operate under various assumptions and employ different methods such as perturbation - randomization and secure multi-party computation approaches.

A novel method has been proposed to preserve the privacy by perturbing the original data at owner's site using "Multiplicative Randomized Data Perturbation" Privacy Preserving Data Mining (PPDM) technique. At the miner's site, we perform secure multiparty computation on the perturbed data from different owners and then we construct a decision tree classifier on the perturbed data. We also construct an improved decision tree classifier in order to overcome the demerits of ID3 algorithm. The experiment results show that the proposed algorithm can overcome ID3's shortcoming effectively and get more reasonable and effective rules.

***Keywords*— Data perturbation, Data mining, Decision tree, Privacy preservation, sensitive data, ID3**

1. INTRODUCTION

Data mining, the extraction of hidden predictive information from large databases, is a powerful technology with great potential to help companies focus on the most important information in their data warehouses. Data Privacy and Protection has become an important concern in the data mining world. The major privacy concerns of the data owners include public disclosure of data relating to lifestyle, financial information like bank transactions, assets and bank balances, medical and political issues. All such data Collected from public domain when processed would result in discovering the identity of the individual. The data owners are reluctant to share all such information which is turning to be an obligation for the data miners. Hence, the Data Mining world is

looking forward for solutions that not only enable data mining operations but also preserve privacy. As a result the Privacy Preserving Data Mining (PPDM) algorithms have gained importance.

At present, the decision tree has become an important data mining method. The basic learning approach of decision tree is greedy algorithm, which use the recursive top-down approach of decision tree structure. Quinlan in 1979 put forward a well-known ID3 algorithm, which is the most widely used algorithm in decision tree. But this algorithm has a defect of tending to use attributes with many values. In this paper we propose a novel method to preserve the privacy by perturbing the original data using randomized data perturbation privacy preserving data mining technique and then constructing a decision tree classifier on the perturbed data. Aiming at the shortcomings of the ID3 algorithm, an association function is introduced to improve ID3 algorithm. The result of experiment shows that the presented algorithm is effective.

2. PREVIOUS WORK

Recently the application of data mining is increased in various domains like business, academia, communication, bioinformatics, medicine field, etc. The data mining not only gives the valuable results hidden in these databases, but sometimes reveals private information about individuals. The difficulty is that by means of linking different attributes data mining process extracts the individual data which is considered as private. The true problem is not data mining, but the way data mining is done. PPDM is an emerging technique in data mining where privacy and data mining can coexist. It gives the summarized results without any loss of privacy throughout the data mining process.

In general there are two main approaches in PPDM:
i) Data transformation based Approach
ii) Cryptographic-based Methods.

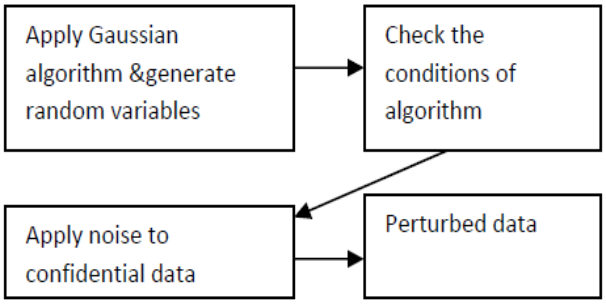
The data transformation based approach modifies sensitive data in such a way that it loses its sensitive meaning. In this process statistical properties of interest can be retained but exact values cannot be determined during the mining process. Various data modification techniques are noise addition [4], noise multiplication [5], data swapping [6], aggregation [7], suppression and signal transformation.

In Cryptographic techniques the data is encrypted with encryption methods and still allow the data mining operation. These methods use certain set of protocols such as secured multiparty computation (SMC). Secure multi-party computation is a computation process performed by group of parties with distributed data set where each party has in its control a part of the input data needed to perform the computation. In SMC the participating parties should only learn the final result of the computation and no additional information is supposed to be revealed at the end of computation. Perfect privacy in the SMC [8] [9] is achieved because no information is released to any third party. The basic SMC PPDM techniques are secure sum, secure set union, secure size of set union etc.

2.1 Overview of randomization perturbation technique

In randomization perturbation approach the privacy of the data can be protected by perturbing [10] sensitive data with randomization algorithms before releasing to the data miner. The perturbed data version is then used to mine patterns and models. The algorithm is so chosen that combined properties of the data can be recovered with adequate accuracy while individual entries are considerably distorted. In this method privacy of confidential data [11] can be obtained by adding small noise component which is obtained from the probability distribution .The method of randomization can be described as follows. Consider a set of data records denoted by $X = \{x_1 \ldots x_N\}$. For record $x_i = X$, we add a noise component which is drawn from the probability distribution $f_y(y)$. Commonly used distributions are the uniform distribution over an interval $[-\alpha, \alpha]$ and Gaussian distribution with mean $\mu = 0$ and standard deviation σ. These noise components are drawn independently, and are denoted $y_1 \ldots y_N$. Thus, the new sets of distorted records are denoted by $x_1 * y_1 \ldots x_N * y_N$. We denote this new set of records by $z_1 \ldots z_N$. In general, it is assumed that the variance of the multiplied noise is large enough, so that the original record values cannot be easily guessed from the distorted data. Thus, the original records cannot be recovered, but the distribution of the original records can be recovered.

Fig 1. Block diagram for implementing perturbation technique



Implementation of perturbation technique is shown in figure 1. In the first step, we apply Gaussian algorithm to generate the random numbers. In the second step, it verifies the required data from the clients and based on these in the third step, we apply noise to the confidential data. In the fourth step, we get the perturbed data as output.

One key advantage of the randomization method is that it is relatively simple, and does not require knowledge of the distribution of other records in the data. Our experiment is performed on numerical database by applying Gaussian technique to all the attributes in a given database. The same technique can be applied to only selected attributes, which the database administrator considers as more sensitive.

2.2 Overview of decision tree

In the decision tree method, information gain approach is generally used to determine suitable property for each node of a generated decision tree. Thus, we can select the attribute with the highest information gain (entropy reduction in the level of maximum) as the test attribute of current node. In this way, the information needed to classify the training sample subset obtained from later on partitioning will be the smallest. That is to say, the use of this property to partition the sample set contained in current node will make the mixture degree of different types for all generated sample subsets reduce to a minimum. Therefore, the use of such an information theory approach will effectively reduce the required dividing number of object classification.

Set S is set including *s* number of data samples whose type attribute can take *m* potential different values corresponding to m different types of C_i i $(1,2,3, ..., m)$. Assume that S_i is the sample number of C_i. Then, the required amount of information to classify a given data is shown in eq.1.

$$I(S_1, S_2, ..., S_m) = -\sum_{i=1}^{m} p_i \log(P_i) \qquad (1)$$

Where $P_i = S_{ij}/|S_j|$ is the probability that any subset of data samples belonging to categories C_i. Suppose that A is a property which has v different values $\{a_1, a_2, \ldots, a_v\}$. Using the property of A, S can be divided into v number of subsets $\{S_1, S_2, \ldots S_v\}$, in which S_j contains data samples whose attribute A are equal a_j in S set. If property A is selected as the property for test, that is, used to make partitions for current sample set, suppose that S_{ij} is a sample set of type C_i in subset S_i , the required information entropy is as shown in eq.2.

$$E(A) = \sum_{j=1}^{v} \frac{S_{1j} + S_{2j} + ... + S_{mj}}{S} I(S_{1j}, ...S_{mj}) \qquad (2)$$

Such use of property A on the current branch node corresponding set partitioning samples obtained information gain is as shown in eq.3.

$$Gain\,(A) = I(S_1, S_2, ..., S_m) - E(A) \qquad (3)$$

ID3 algorithm traverses possible decision-making space using top-down greedy search strategy, and never trace back and reconsider previous selections. Information gain is exactly the metrics for selecting the best attribute in each step of the growth tree in ID3 algorithm.

Algorithm for generating a decision tree according to a given data sets.

Input: training samples, each attribute taking discrete value, a candidate attribute set available for induction is attribute_list.
Output: a decision tree.
Begin
If S is empty, return a single node with value Failure;
If S consists of records all with the same value for the target attribute, return a single leaf node with that value;
If R is empty, then return a single node with the value of the most frequent of the values of the target attribute that are found in records of S;
Select test-attribute, the attribute among attribute-list with highest information gain;
Let A be the attribute with largest Gain (A, S) among attributes in R;
Let {aj | j=1, 2... m} be the values of attribute A;
Let {Sj | j=1, 2... m} be the subsets of S consisting respectively of records with value aj for A;
Return a tree with root labeled A and arcs labeled a1, a2... am going respectively to the trees (ID3(R-{A}, C, S1), ID3 (R-A}, C,S2)… ID3(R-{A}, C, Sm);
Recursively apply **ID3** to subsets {Sj | j=1, 2... m} until they are empty
End.

This is a greedy algorithm which use recursive manner of top-down, divide and conquer to construct a decision tree. The termination condition of recursion is: all samples within a node are of the same category. If no attribute can be used to divide current sample set, then voting principle is used to make it a Compulsory leaf node, and mark it with the category of having the most number of sample types. If no sample satisfies the condition of test-attribute = *i a* , then a leaf node is created, and mark it with the category of having the most number of sample types.

2.3 The shortcoming of ID3 algorithm

The principle of selecting attribute A as test attribute for ID3 is to make E (A) of attribute A, the smallest. Study suggest that there exists a problem with this method, this means that it often biased to select attributes with more taken values [12, 13], however, which are not necessarily the best attributes. In other words, it is not so important in real situation for those attributes selected by ID3 algorithm to be judged firstly according to make value of entropy minimal. Besides, ID3 algorithm selects attributes in terms of information entropy which is computed based on probabilities, while probability method is only suitable for solving stochastic problems. Aiming at these shortcomings for ID3 algorithm, some improvements on ID3 algorithm are made and a improved decision tree algorithm is presented.

3. THE IMPROVED ID3 ALGORITHM

To overcome the shortcoming stated above, attribute related method is firstly applied to compute the importance of each attribute. Then, information gain is combined with attribute importance, and it is used as a new standard of attribute selection to construct decision tree. The conventional methods for computing attribute importance are sensitivity analysis (SA) [14], information entropy based joint information entropy method (MI) [15], Separation Method (SCM)[16], Correlation Function Method (AF) [17,18], etc. SA needs not only to compute derivatives of output respect to input or weights of neural network, but also to train the neural network. This will increase computational complexity. MI needs to compute density function and it is not suitable for continuous numerical values. SCM computes separation property of input-output and the correlation property of input and output attributes and is suitable for both continuous and discrete numerical values, but computation is complex. AF not only can well overcome the ID3's deficiency of tending to take value with more attributes, but also can represent the relations between all elements and their attributes. Therefore, the obtained relation degree value of attribute can reflect its importance.

AF algorithm: Suppose *A* is an attribute of data set *D*, and *C* is the category attribute of *D*. the relation degree function between *A* and *C* can be expressed as follows:

$$AF(A) = \frac{\sum_{i=1}^{n} | x_{i1} - x_{i2} |}{n} \qquad (4)$$

Where x_{ij} (j = 1, 2 represents two kinds of cases) indicates that attribute *A* of *D* takes the *i*-th value and category attribute *C* takes the sample number of the *j*-th value, *n* is the number of values attribute *A* takes. Then, the

normalization of relation degree function value is followed. Suppose that there are *m* attributes and each attribute relation degree function value are *AF*(1), *AF*(2),…, *AF*(*m*) , respectively. Thus, there is

$$V(k) = \frac{AF(k)}{AF(1) + AF(2) + ... + AF(m)} \qquad (5)$$

Which $0 < k \leq m$. Then, equation (3) can be modified as

$$Gain'(A) = (I(S_1, S_2, ..., S_m) - E(A)) * V(A) \quad (6)$$

Gain'(*A*) can be used as a new criterion for attribute selection to construct decision tree according to the procedures of ID3 algorithm. Namely, decision tree can be constructed by selecting the attribute with the largest *Gain*'(*A*) value as test attribute. By this way the shortcomings of ID3 can be overcome. It construct the decision tree, this tree structure will be able to effectively overcome the inherent drawbacks of ID3 algorithm.

4. OUR FRAMEWORK

In this novel framework as shown in figure 2, we use two key components, the data perturbation component at data provider site and classifier component in the data miner site. Our scheme is a Four -step process. In the first step, the data miner negotiates with different data provider depending on the query submitted by the user. In The second step the randomized perturbation technique is applied on the data set which satisfies the user query. In the third step data miner obtains the perturbed data from the data provider. In the fourth step a classifier is built on the perturbed data set.

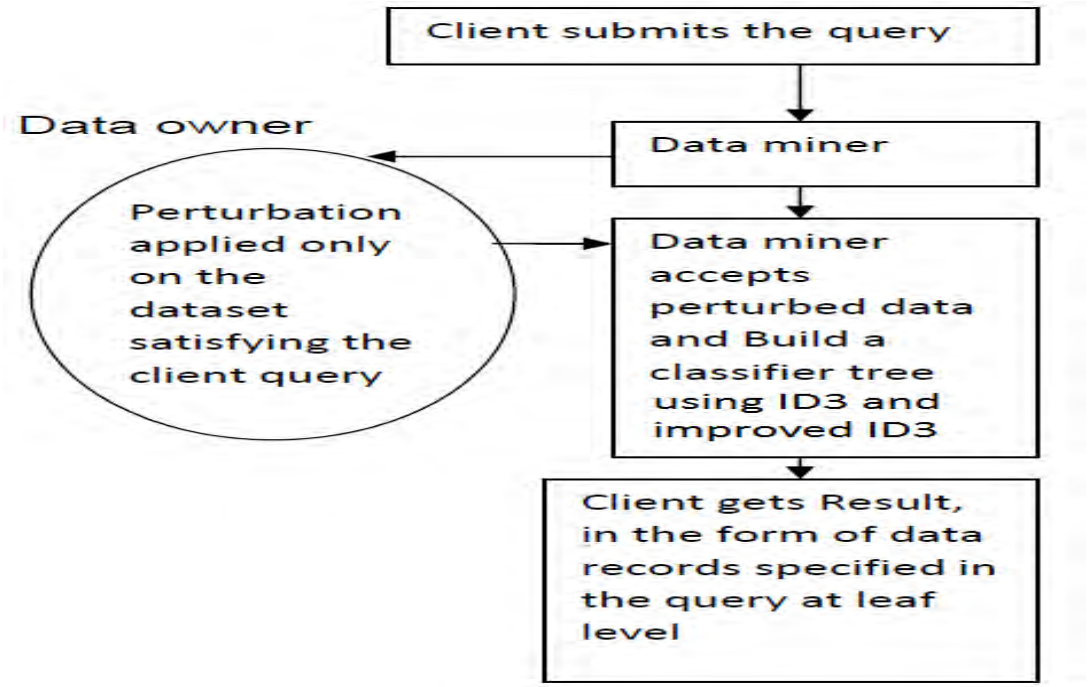


Fig 2. A framework to integrate perturbation and classification techniques

This framework guarantees the privacy because the record on which the classifier is constructed is in the perturbed form. Confidentiality is also achieved because the data owner or provider does not learn anything about the classifier which has been constructed. The parameter like attribute selected at the root node, attribute used as class attribute and the records selection criteria remain hidden from the data owner. Figure 3 gives the example of classifier tree on perturbed data.

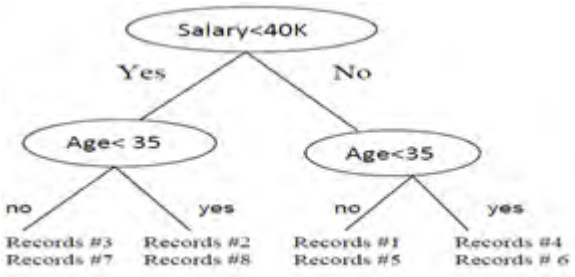


Fig 3. Example of classifier tree on perturbed data

5. EXPERIMENTAL RESULTS

A customer database of some shopping mall is shown in Table 1 (a training sample set). The category attribute of the sample set is "buying-computer", which can take two different values: buying-computer or No buying-computer

Table 1. Shopping mall customer database

Case	Age	Color-cloth	Income	Student	Buy-computer
1	>40	Red	High	No	No
2	<30	Yellow	High	No	No
3	30-40	Blue	High	No	Yes
4	>40	Red	Medium	No	Yes
5	<30	White	Low	Yes	No
6	>40	Red	Low	No	No
7	30-40	Blue	Low	No	Yes
8	<30	Yellow	Medium	Yes	Yes
9	<30	Yellow	Low	No	No
10	>40	White	Medium	No	No

In order to illustrate the effectiveness of our present algorithm, the improved ID3 algorithm and ID3 algorithm are applied on this example to construct decision trees and comparison is made. Figure 4 and figure 5 shows the generated decision trees using the ID3 algorithm and the improved ID3 algorithm respectively.

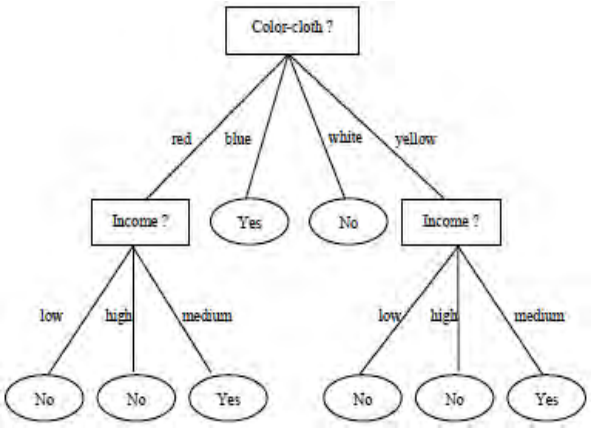


Fig 4. The obtained decision tree using ID3 algorithm.

The two results of the experiment shows that ID3 algorithm choose attribute color-cloth as root node to generate decision tree, but the importance of attribute color-cloth is lower than the other attributes, and it is just the shortcoming of ID3 which tends to take attributes with many values. However the improved ID3 algorithm decreases the importance of attribute color-cloth in classification and comparatively enhanced the importance of

attributes such as age, income, and student, etc. in classification. It well solves the problem that ID3 algorithm tends to take attributes with many values and it can obtain more reasonable and effective rules.

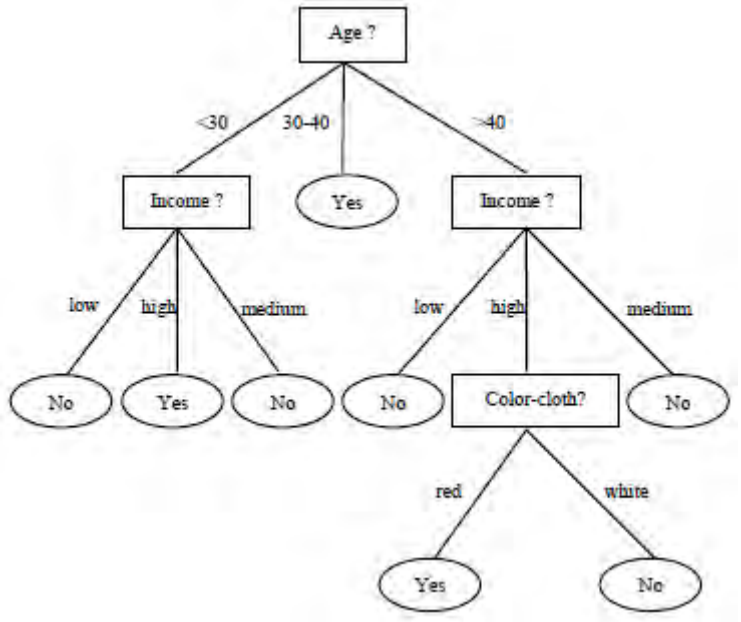


Fig 5. The obtained decision tree using improved ID3 algorithm

6. CONCLUSION

Data mining extracts useful patterns from large quantities of data stored in the data warehouse. The data mining process results valuable patterns to support decision making in different domains. But easy access to sensitive data poses threat to individual privacy. In this paper we presented a novel approach in which both data perturbation technique and classification are integrated to provide better data quality and individual privacy both at data owner site as well as at data mining site. The owner's data consists of both categorical and numeric data types. To preserve the privacy of data at owner's site perturbation technique is used in which small amount of noise is added to sensitive data such that the properties and the meaning of the original data is not changed. The problem with the randomization technique is that some privacy intrusion techniques can be used to reconstruct private information from the randomized data tuples. Hence to enhance the performance a decision tree is built on the perturbed data at data mining site, which reveals and gives only the required results and hides other information.

REFERENCES

[1] H. Witten, E. Frank, Data Mining Practical Machine Learning Tools and Techniques, China Machine Press, 2006.
[2] S. F. Chen, Z. Q. Chen, Artificial intelligence in knowledge engineering [M]. Nanjing: Nanjing University Press, 1997.
[3] Z. Z. Shi, Senior Artificial Intelligence [M]. Beijing: Science Press, 1998.
[4] K.Muralidhar.,R.Sarathy,"A General additive data perturbation method for data base security", journal of Management Science. ,45(10):1399-1415,2002.
[5] Privacy Is Become With, Data Perturbation, Niky Singhai, Er. Niranjan Singh
[6] Muralidhar K. and Sarathy R.," Data Shuffling- a new masking approach for numericadata"managements science, forthcoming, 2006.
[7] V.S. Iyengar."Transforming data to satisfy privacy constraints" In Proc. of SIGKDD'02, Edmonton, Alberta,Canada,2
[8] Lindell Y., Pinkas B."Privacy preserving Data Mining" CRYPTO 2000.
[9] Yu.H.,Vaidya J.,Jiang X."Privacy preserving SVM Classification on vertically partitioned data" PAKDD conference, 2006.
[10] R.Agarwal and R.Srikant, "Privacy preserving data mining", In Proceedings of the 19th ACM SIGMOD conference on
[11] Management of Data ,Dallas,Texas,USA, May2000.
[12] H. Kargupta, S. Datta, Q. Wang, and K. Sivakumar, "On the privacy preserving properties of random data perturbation techniques" In Proc. of 3rd IEEE Int. Conf. on Data Mining,
[13] Washington, DC, USA,, pages99–106, 2003.
[14] M. Zhu, Data Mining [M]. Hefei: China University of Science and Technology Press ,2002.67-72.
[15] D. Jiang, Information Theory and Coding [M]: Science and Technology of China University Press, 2001.

[16] P. Engelbrecht., A new pruning heuristic based on variance analysis of sensitivity information[J]. IEEE Trans on Neural Networks, 2001, 12(6): 1386-1399.
[17] N. Kwad, C. H. Choi, Input feature selection for classification problem [J],IEEE Trans on Neural Networks, 2002,13(1): 143- 159.
[18] X. J. Li, P. Wang, Rule extraction based on data dimensionality reduction using RBF neural networks [A]. ICON IP2001 Proceedings, 8th International Conference on Neural Information Processing [C]. Shanghai, China, 2001.149- 153.
[19] S. L. Han, H. Zhang, H. P. Zhou, correlation function based on decision tree classification algorithm for computer application in November 200.
[20] S. Y. Zhang, Z. Y. Zhu, Study on decision tree algorithm based on autocorrelation function. Systems Engineering and Electronic Jul. 2005 Vol.27 No.7.