An Efficient Method for Preserving Privacy of Classification Data used For Decision Tree Learning

B. Anu Sheeba, Asst.Prof of CSE
Kottayam Institute of Technology and Science,
Kottayam Dist, Kerala, India.
anuja_cse@ymail.com

Dr. V. Suresh Kumar, Principal
Kottayam Institute of Technology and Science,
Kottayam Dist, Kerala, India.
principalkitscollege@gmail.com

Abstract - **In this research paper a proposed model for privacy preservation in data was implemented. At first a large amount of data can be collected. The collected data can be arranged in randomized manner based on the priority given to the data in database. A set of dummy data can be created for the selected tuples in the database. Then the dummy data can be arranged in randomized manner based on the priority given to the original data. This dummy data look like original data which cannot be identified by information hackers, only can identify the original data by the server who design.**

IndexTerm – Privacy Preserving , DummyData, Priority, Randmoized method.

I. INTRODUCTION

Privacy preserving in Data Mining is the important concept in this modern world. The privacy protection of personal information recently, due to the popularity of electronic data held by commercial corporations and we want to protect the information from the information hackers. Safeguarding confidential data of a database has been a challenging issue in the past and emerges. It is a one of the most critical information technologies today. The problem that arises when confidential information can be derived from released data by unauthorized users is commonly called the database inference problem. Data mining is a well-known technique for automatically and intelligently extracting information or knowledge from a large amount of data, however, it can also disclose sensitive information about individuals compromising the individual's right to privacy. Therefore, privacy preserving data mining has becoming an increasingly important field of research. Privacy preserving data mining is a novel research direction in data mining. In recent years, with the rapid development in Internet, data storage and data processing technologies, privacy preserving data mining has been drawn increasing attention.

Data are values of qualitative or quantitative variables, belonging to a set of items. In recent years, advances in hardware technology have led to an increase in the capability to store and record personal data about consumers and individuals.

This has led to concerns that the personal data may be misused for a variety of purposes. Data explains a business transaction, a medical record, bank details, educational details etc., Use of technology for data collection and analysis has seen an unprecedented growth in the last couple of decades. Such information includes private details, which the owner doesn't want to disclose, such data are the sources for data mining. Data mining gives us "facts" that are not obvious to human analysts of the data. When such sensitive data are given directly for mining, the security of the individual is highly affected. So the data are modified and presented for data mining. But the problem is that the altered data should also produce a similar mining result. This has led an area called privacy preservation in data mining, this is important in intersection of data in data mining and provide information security for highly sensitive data.

In today's scenario, the World Wide Web is filled with huge amount of data, which approximately doubles in every 50 days, huge amount of data requires special consideration for its storage, retrieval and reasoning. Such data should be efficiently handled with sufficient security. This paper is concentrated on the methods that consider incremental data for which there are huge chances of data expansion. In this paper three main steps can be carried out, the first step is the collection of large amount of data and arranged in priority based randomized manner. The second step is to create dummy data set for the original data, which can be arranged in randomized manner based on the priority given to the each dummy data in the database. The third step is an efficient Decision Tree creation for the dummy data sets, such that the dummy data look like original data. The dummy data can be reconstructed to original data, which can be done only by the

ISSN: 2278-0181

developer. The goal of privacy preserving data mining is to develop data mining methods without increasing the risk of misuse of the data used to generate those methods.

II. METHODOLOGY

In Privacy Preserving Data Mining Models and Algorithms, Aggarwal and Yu classify privacy preserving data mining techniques, including data modification and cryptographic, statistical, query auditing and perturbation-based strategies. Statistical, query auditing and most cryptographic techniques are subjects beyond the focus of this paper. In this section, we explore the privacy preservation techniques for storing data from privacy attacks.

Data modification techniques maintain privacy by modifying attribute values of the sample data sets. Essentially, data sets are modified by eliminating or unifying uncommon elements among all data sets. These similar data sets act as masks for the others within the group because they cannot be distinguished from the others, every data set is loosely linked with a certain number of information providers.

Even in microcontroller and microprocessor programming opcode is provided for each mnemonics instruction, which can be understand only by the machines. Like that for each data in a database can be converted in to a separate code which cannot be hacked by others should be done.

Our method in privacy preserving is item set mining. Let assume that data are stored separately. We apply distributed set operations to distributed data. We propose dummy-based set operations to process the itemset mining from distributed data. We evaluate them by distinguish ability based criterion based on the assumption that data are represented based on priority .The priority of each data can be give based on the sensitive of the data. Our dummy dataset based method is classified a secure and computational method based on anonymizing with dummy data.

Our goal is to provide dummy-based methods for itemset mining using non-secure methods for set union and set intersection used in the mathematical models. The cost is expected to be lower than SMC-based method.

Dummies are generated and they satisfy our criterion based on the data are represented by ZDDs. Several databases are stored in distributed places and are regarded as a large database DB = DB1 U DB2 U … U DBn. ID Users of DB1 are allowed to make query like item sets which are DB1 and occur at least three times: {a, b, ab} and item sets which are DB1 and occur at least three times and have at least two items {ab}.

In distributed item set mining users of DBk cannot make such queries (k≠1). And users of DB1 sent them to Users of DBk for mining. Several databases are stored in distributed places and the user data have some characteristics, they are candidate sets (like Apriori) and Set of closed item sets (like Pasquier).

In the cases of two database i.e., n=2, to keep privacy is difficult but the cases are useful in analyzing dummy-based methods. In identifiable data method the data hackers can easily get the original data, which contain n=2 basis, because the size of the database is too small and the data contain is more sensitive to the hacker. Mostly in medical database field the data about the patient is more sensitive but because of lack insecurity the data can be easily hacked.

But in indistinctive method more than two distributed database is used and we can easily provide dummy data for distributed data item set. In previous works, there is no assumption about data structures, while we restrict ourselves to use ZDDs for representing item sets, and our privacy criterion is specialized for ZDDs.

The modules defined below flowchart represent the overall function of the privacy preservation method used in this paper. At first the dataset from database is collected and each data in the data set can be analyzed and other than the identity like name, date of birth, age etc should not be considered for identification.

The identity number is only taken under consideration. The ID number provided to each patient in medical field can be taken as dataset for creating dummy dataset.

After the data is altered, the amount of compromise to privacy is more important. More the percentage of data leakage less efficient is the algorithm. A record is said to be private when its sensitive attribute value cannot be identified even after having the best knowledge of all the QI attributes.

As the data is modified, the amount of data distortion should be minimal. This can be verified by using a data mining tool and compare the table before and after data modification. It concerns with accuracy, completeness and consistency of the data.

The algorithm that is used for implementing privacy has no right to alter the result of data mining that is done for the actual data. The data from different database can be collected for this purpose and similar attributes can be taken under consideration.

The data in the database contain attributes like ID and items. The dummy data are used to reconstruct the original data back if necessary in any situation occur.

In SMC method privacy preservation is done only for data which is already collected but in dummy data method even update data can be protected by creating new dummy data related to the original data. The dummy data created for the original data should be based on special codes created by the server .In this method all the sensitive data has high priority and creating dummy data will be created very carefully and very effectively. In numerology also they are providing specific numbers for each alphabet.

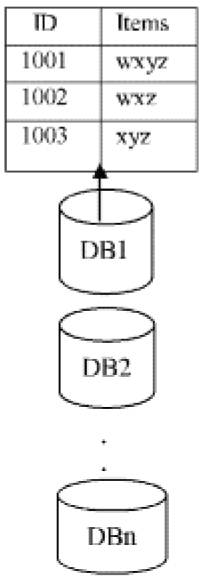


Fig.1 Distributed database

ISSN: 2278-0181

The Computational cost is reduced in this paper. Since the procedure is an additional task done on the useful data for providing privacy, the total computational cost should be low. Higher the computational cost lesser efficient is the methodology. The method used in this paper reduces the computational cost in providing privacy.

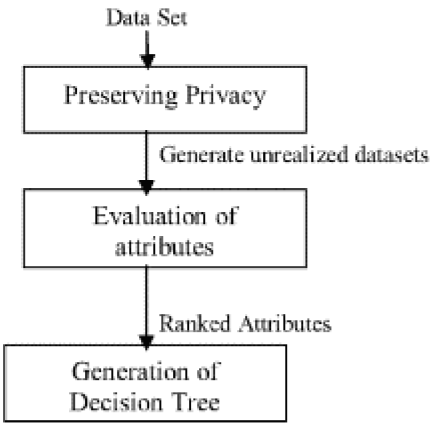


Fig.2 Flow diagram for Privacy Preserving method

III. MODULE CLASSIFICATION

There are three module classification done in this paper. They are

a) Preserving privacy
b) Evaluation of Attributes
c) Generation of Decision

a) Preserving Privacy

A new perturbation and randomization based approach is used that protects centralized sample data sets utilized for decision tree data mining. Privacy preservation is applied to sanitize the samples prior to their release to third parties in order to mitigate the threat of their inadvertent disclosure or theft. In contrast to other sanitization methods, this approach does not affect the accuracy of data mining results. The output of the preserving privacy is generation of unrealized datasets. These unrealized datasets are the dummy datasets but which can be look like original datasets.

b) Evaluation of Attributes

The unrealized datasets can be evaluating for the user to get the required information easily. In here each attributes can be classified according to their priority. Based on their priority each attributes can be ranked. The output of evaluation of attributes is ranked attributes. The priority is provided based on the sensitive of the data. If the data has higher sensitive then that data has high priority.

c) Generation of Decision Tree

A decision tree is a predictive model that can be used to represent both classifiers and regression models. The decision tree can be built directly from the sanitized data sets, such that the originals do not need to be reconstructed. Moreover, this approach can be applied at any time during the data collection process so that privacy protection can be in effect even while samples are still being collected.

A. Algorithm for Dummy Data Creating

Input: T_S, a set of input sample data sets
T^U, a universal set
T^1, a set of output training data set
T^P, a perturbing set

Output: (T^1, T^P)

If T_S is empty then return (T^1, T^P)

$t \leftarrow$ a data set in T_S

if t is not an element of T^P or $T^P = \{t\}$ then

$T^P \leftarrow T^P + T^U$ and

$T^P \leftarrow T^P - \{t\}$

$t^1 \leftarrow$ the most frequent dataset in T^P

return Dummy data Set

$\{(T_S - \{t\}, T^U, T^1 + \{t^1\}, T^P - \{t^1\})\}$

B. Description of Algorithm

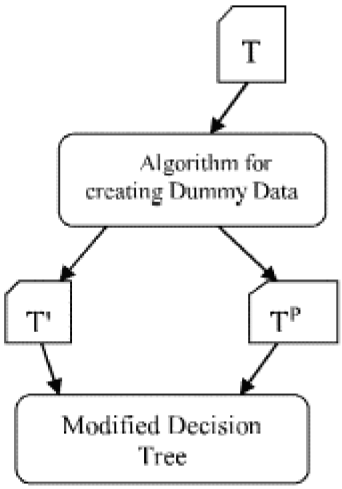
A training set, T_S, is constructed by inserting sample data sets into a data table. However, a data set complementation approach, is used to requires an extra data table, T^P. T^P is a perturbing set that generates unreal data sets which are used for converting the sample data in to an unrealized training set, T '. To create dummy data for the samples, T_S, we initialize both T' and T^P as empty sets, i.e., we invoke the above algorithm with Dummy data Set (T_S, T^U, {}, {}).The resulting dummy dataset contains some data sets excepting the ones in T_S. The elements in the resulting data sets are unreal individually, but meaningful when they are used together to calculate the information required by a modified ID3 algorithm. The ID3 algorithm assumes that each attribute is categorical, that is containing discrete data only, in contrast to continuous data such as age, height etc.[1]

The basic usage scenario for the data mining is the following: imagine a data custodian (e.g., a health authority) holds a large amount of sensitive data about individuals. The custodian wishes to release information to a third party for information processing; for instance, it wishes to release information to a data analysis firm that will assemble a classification schema that determines which patients are likely to avoid paying their medical bills.

Since the data custodian does not have staff members capable of performing the analysis, it must find a method of releasing data to a third party in a manner that protects privacy interests. While it could attempt to protect the data by contract, contractual protections are only useful if the parties trust one another to honor them, or if defections from the contract can be detected. In the case of secondary uses of data sets, detection of unauthorized uses is next to impossible.

The dummy data approach allows the data custodian to take the database T, and process it into an unreal data set T′ and a perturbing data set T^P. Instead of giving T to the third party, it may select a subset $t \subseteq T$, process it by using the algorithm and release $t' \subseteq T'$ and $t^P \subseteq T^P$. The third party may use the modified data mining algorithm to build the same decision tree from t′ and t^P that it would have constructed from t. The problem of protecting the information in t has now become the problem of ensuring that the third party cannot use t′ and t^P to make inferences about the original data subset t, apart from what it can learn by constructing a decision tree. Although a reverse transformation scheme can allow a third party to reconstruct T from T′ and T^P it only works if the data custodian has released a perturbing and unreal data set corresponding to the full contents of T. For $t' \subset T'$ and $t^P \subset T^P$ (proper subsets), the reverse transformation will not work. In particular, the dummy data approach enables secure multiparty computation schemes. If the custodian doesn't trust one party with both data sets T′ and T^P, it may give the perturbing set to one party, and the unrealized set to another. Protocols can then be developed to facilitate data mining using an exchange, where no one party is privy to both the full unrealized and perturbing data sets.

C. Block diagram for Algorithm



T → Training Data set
T' → Unrealized training set
T^p → Perturbing data set

D. Description of Block Diagram of Algorithm

The training set T is the set of data's which is the sample data fed as input to the dummy data algorithm and can be converted in to two types of data they are perturbation data T^P and unrealized data T'. Then both the data can be given to the decision tree algorithm. In the modified decision tree algorithm the data can be arranged in priority based, which is useful for the user for data mining. Thus privacy preservation can be developed for sampled data sets. The high sensitive data can be protected from the hackers and the dummy data can be created for all original data.

E. Description about Dummy data Algorithm

In Distributed Itemset Mining we find frequent itemsets from DB = DB1 U DB2.

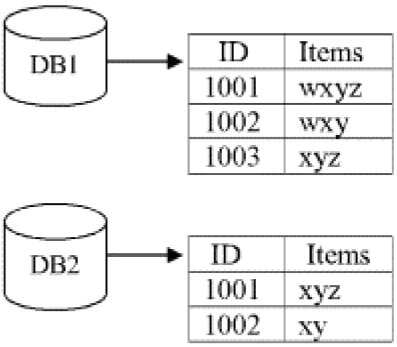


Fig.3 Separate database DB1 and DB2

The above itemsets occurring at least 3 times in DB1∪DB2 is {w,x, y,z, ...}

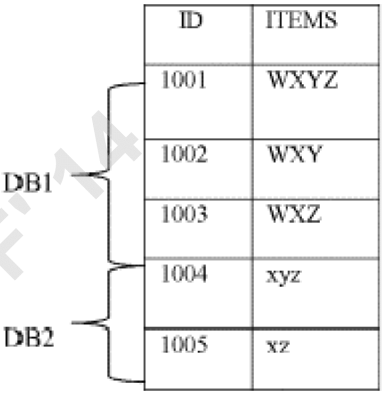


Fig.4 *Database DB after union operation*

We find frequent item sets from DB = DB1 U DB2. And Image of our targeting database Set of items *I*. Then the condition of frequency w.r.t. σ is X.sup≥ σ x (|DB1| + |DB2|).Then find the candidates based on closed itemsets by set union and intersection operations, and then check the condition of frequent. We use natural numbers in an interval [1,R] as identifiers of items, so we use *I* = {1,2,…,R} instead of {w,x,y,…}. The interval [m,n] represents the itemset {m,m+1,…,n}. We sometimes represent the database as binary table (bij). If DB contains an item j on ID i, then bij=1. Our proposal is based on dummy insertion like below table.[1]

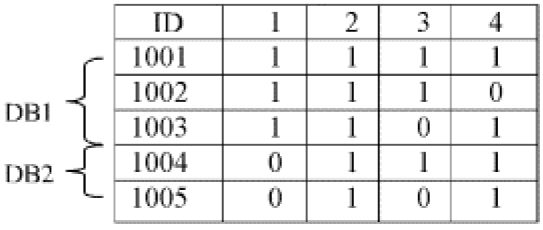


Fig.5 Database after dummy data insertion

To conceal original records and to conceal information on items three main steps want to be performed, they are

a) Expanding the set of increasing variation of items
b) Inserting many records- Increasing number of records
c) Removing dummy data.

The above three main steps can be carried out in this paper. The rarely used tuples used in the medical database can be selected as training data set and dummy data set can be created based on all possibilities present in that tuples. Inserting dummy data for two sets M, N, p(M, N) becomes a new set. By using this method the information hackers cannot distinguish which items are original data or dummy data.

$|DB| < |D|$

$|DB|$: the size of original data

$|D|$: the size of dummy data ,

A user u1 of DB1 and a user u2 of DB2 insert dummy data in to their databases respectively by mixing function p, and represent and the users calculate the set union then the users remove dummy data with screening function e, and get the final result.

IV. RESULTS AND DISCUSSION

The privacy preservation method used in this paper is in progress for applying medical database especially for cardiac patients. The different medical test taken for the cardiac patients

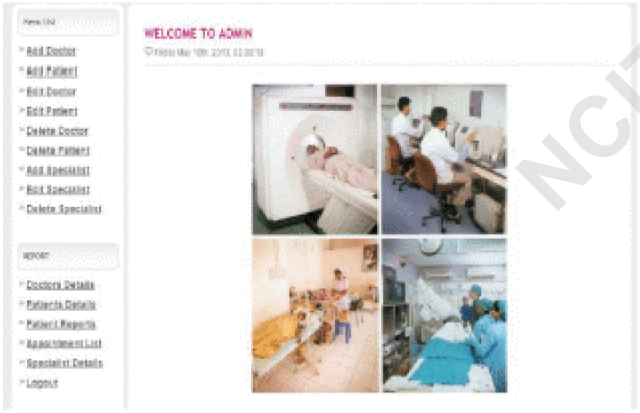


Fig.6. *Log In Design*

In this paper the front end designing work is completed and the method used for privacy preservation and the algorithm is designed. Eclipse tool is used for front end design and in back end MysqlR2 2008 is used. The algorithms want to be implemented in the selected database and selection of attributes. Selection of best tuples are the work under process. The main aim of this method is to provide privacy to the database and to reduce the computational cost.

Our method mainly depends on the size and property of dummy data, while SMC depends on cryptographic methods like encryption. Reduction of computational cost is derived by how to treat large-scale dummy data efficiently. [5] We proposed the method of inserting dummy data for set operations in item set mining from distributed databases. We achieve well-mixedness based on the size of a ZDD. If we use many dummy data, the criterion is satisfied. We require more analyses of other data structures, more complex cases where there exist more than two users or colluding user. We also require more precise comparisons between the dummy-based method and the SMC-based method. We have to evaluate how many data are needed for satisfying our criterion. This method is efficient than other method like noise adding, cryptography and other randomized methods. Selection of attributes and analyzing the possibilities are the main work done in this paper.

Fig.7. Patient Detail Design

V. CONCLUSION

In this paper a detailed study of the method used for privacy preservation by inserting dummy data for set operations is done, which is used in mining from distributed databases. We achieve well-mixedness based on the size of a ZDD and if we use many dummy data, the criterion is satisfied. The algorithm and the methodology is created and implementation of this privacy preservation method in medical database is in process and we require more analyses of other data structures, more complex cases where there exist more than two users or colliding user. And also we worked in a detail comparisons between the dummy-based method and SMC-based method. By doing comparing analyzing we find out that dummy based data privacy method best than SMC-based method, because in dummy based method we can provide a better privacy in updating data but in SMC method before providing privacy all data to be collected and it is difficulty to provide privacy for updated data.

VI. REFERENCES

[1] Pui K. Fong and Jens H. Weber-Jahnke," Privacy Preserving Decision Tree Learning Using Unrealized Data Sets, " IEEE Trans. On Knowledge and Data engineering, vol. 24, no. 2, february 2012.

[2] Jim Dowd, Shouhuai Xu, and Weining Zhang," Privacy-Preserving Decision Tree Mining Based on Random Substitutions, " IEEE Trans on Knowledge and Data engineering ETRICS 2006, LNCS 3995, pp. 145–159, 2006.

[3] Alka Gangrade, Ravindra Patel,"Building Privacy-Preserving C4.5 Decision Tree Classifier On MultiParties ," International Journal on Computer Science and Engineering Vol.1(3), 2009, 199-205.

ISSN: 2278-0181

[4] Selim V. Kaya, Thomas B. Pedersen, Erkay Sava¸, and Yücel Saygın," Efficient Privacy Preserving Distributed Clustering Based on Secret Sharing ," PAKDD 2007 Workshops, LNAI 4819, pp. 280–291, 2007. @Springer-Verlag Berlin Heidelberg 2007.

[5] Keng-Pei Lin and Ming-Syan Chen, "On the Design and Analysis of the Privacy Preserving SVM Classifier," IEEE Trans on Knowledge And Data Engineering, vol. 23,no. 11,November 2011.

[6] Oded Maimon and Rokach Lior, "Introduction to knowledge discovery in Databases", In Data Mining and Knowledge Discovery Handbook, page 1. Springer Berlin / Heidelberg, 2005.

[7] Rakesh Agrawal and Ramakrishnan Srikant, "Privacy-preserving data mining", SIGMOD Rec., 29(2):439–450, 2000.

[8] Joseph Albert, "Algebraic properties of bag data types", In VLDB '91: Proceedings of the 17th International Conference on Very Large Data Bases, pages 211–219, San Francisco, CA, USA, 1991. Morgan Kaufmann Publishers Inc.

[9] Alastair R. Beresford and Frank Stajano, "Location privacy in pervasive computing". IEEE Pervasive Computing, 2(1):46–55, 2003.

[10] Paul W. P. J. Grefen and Rolf A. de By," A multi-set extended relational algebra a formal approach to a practical issue", In Proceedings of the Tenth International Conference on Data Engineering, pages 80–88, Washington, DC, USA, 1994. IEEE Computer Society.