ISSN : 2347 - 8446 (Online)
ISSN : 2347 - 9817 (Print)
Vol. 2 Issue Special 1 Jan-March 2014
International Journal of Advanced Research in Computer Science & Technology (IJARCST 2014)

# Privacy Preservation Decision Tree Learning of Functional Dependency Dataset

[I]J.Karthik, [II]J.Preethi
[I,II]Dept. of CSE, Anna University Regional Center, Coimbatore, Tamilnadu, India

**Abstract**
*Privacy-preserving is an important issue in the areas of data mining and security. The aim of privacy preserving data mining is to develop algorithms to modify the original dataset so that the privacy of confidential information remains preserved and as such, no confidential information could be revealed as a result of applying data mining tasks. In existing system they introduced a new privacy preserving approach via data set complementation which confirms the utility of training data sets for decision tree learning. This approach converts the original data sets, TS, into some unreal data sets such that any original data set is not able to reconstruct if an unauthorized party were steal some portion of unrealized datasets. Meanwhile, there remains only a low probability of random matching of any original data set to the stolen data sets, TL. This work covers the application of new privacy preserving approach with the ID3 decision tree learning algorithm. The problem in existing system is insufficient storage mechanism and this ID3 only can be implemented for discrete-valued attributes only. To support continuous-valued attributes c5.0 algorithm is used and to overcome the problem of privacy threat from the full functional dependency (FFD) that is used as part of adversary knowledge, proposed system formalizes the FFD based privacy attack and defines the privacy model to combat the FD-based attack.*

**Keywords**
*Classification, Data mining, Functional Dependency, Machine learning, Security and Privacy protection*

## I. Introduction

The problem of privacy-preserving in data mining has become more important in recent year because the ability to store personal data about users is increased, and the increasing knowledge about the data mining algorithms to control this information. There are number of techniques such as randomization and k-anonymity have been suggested in order to perform privacy- preserving data mining. Also this problem has been discussed in many communities such as the statistical disclosure control community, database community and the cryptography community[9], [14].

In some cases, the different communities have explored parallel lines of work which are quite similar. This paper will try to explore different topics from the perspective of different communities, and will try to give a fused idea of the work in different communities. Preserving privacy is more important for machine learning and data mining, but the measures designed to protect private information sometimes result in a degradation and reduced utility of the training samples.

Nowadays, worldwide networked society places great demand on the broadcasting and sharing of information, which is possibly becoming the most important and demanded storage but in the past released information was mostly in tabular and statistical form which is called as macro data. But today many situations call for the release of specific data which only belongs to the particular domain called as the micro data [23]. Micro data, contains the specific data in its original form, it said to be as contrast to macro data which reporting pre-computed statistics data, provide the convenience of allowing the final recipient to perform on them analysis as needed. Micro data domain such as public health and population studies there are many possibilities to violate individual privacy. This leads to concerns that the personal data may be misused for a variety of purposes. A field of research namely privacy preserving data mining works on techniques to alleviate these concerns. These techniques of privacy-preservation are drawn from a wide array of related topics such as data mining, cryptography and information hiding [1].Privacy preservation consists of two types:

1. Individual privacy preservation
2. Collective privacy preservation

The goal of Individual privacy preservation is to protection the personal identification information. Collective privacy preservation not only protecting the personal identification information and also some patterns and trends that are not supposed to be reveal. This work introduces an approach that can be applied to decision-tree learning, without concurrent loss of accuracy. It describes an privacy preservation approach for the collected data samples in cases when information of the sample database has been partially lost. This approach converts the original datasets into a group of unreal datasets [1], in which the original data cannot be reconstructed without the entire group of unreal datasets if some portion of the unreal datasets is stolen. This approach does not suitable when sample datasets have low frequency or low variance in the distribution of all samples. However, this problem can be resolved through a alternative implementation of the approach introduced later in this work, by using some extra storage. The various techniques of privacy preserving data mining [4] are shown in Fig. 1.

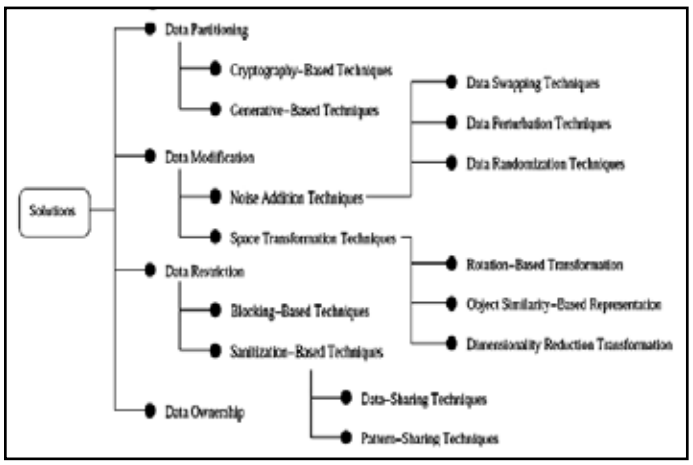


Fig. 1: Technique of PPDM: [6], [17]

ISSN : 2347 - 8446 (Online)
ISSN : 2347 - 9817 (Print)

The key directions in the field of privacy-preserving data mining [4], [5], [15] are as follows:

### 1. Privacy-Preserving Data Publishing

These techniques tend to study different transformation methods associated with privacy. These approaches include methods such as randomization[17], k-anonymity, and l-diversity where k-anonymity has been used in order to perform privacy- preserving data mining. A related issue is how the perturbed data can be used along with classical data mining methods such as association rule mining [13]. Other related problems include that of determining privacy preserving methods to keep the underlying data useful or the problem of studying the various privacy definitions, and how they compare in terms of effectiveness in different states.

### 2. Modifying the results of Data Mining Applications to preserve privacy

In many cases, the results of data mining applications such as association rule or classification rule mining can compromise the privacy of the data. This has generate a field of privacy in which the results of data mining algorithms such as association rule mining are modified in order to preserve the privacy of the data. A classic example of such techniques are association rule hiding methods [3], in which some of the association rules are suppressed in order to preserve privacy. Likewise many techniques are available to modify the results of the data mining applications.

3. Cryptographic techniques for Distributed Privacy

In many cases, the data may be distributed across many sites, and the owners of the data across these different sites may wish to compute a common function. In those cases, a variety of cryptographic protocols [2], [14] may be used to communicate among various sites, so that secure function computation is possible without revealing the sensitive information.

## II. Research Background and Objective

Even when databases of samples with sensitive information are protected securely, partial information of the databases can be lost through procedural mistakes or privacy attacks which can be from anywhere within a network. This work focuses on analyzing privacy preservation following the loss of some training datasets from the whole sample database used for decision-tree learning[11]. On this basis, we make the following assumptions for the scope of this work: first, as is the norm in data collection processes, a large number of sample datasets have been collected to achieve significant data mining results that cover the whole research target. Second, the number of datasets lost constitutes a small portion of the entire sample database. Third, for decision-tree data mining, no attribute is designed for distinctive values, because such values negatively affect decision classification.

The objective of this work is to introduce a new privacy preserving approach to the protection of sample datasets that are utilized for decision-tree data mining. Privacy preservation is applied directly to the samples in storage, so that privacy can be safeguarded even if the data storage were to be threatened by unauthorized parties. Although effective against privacy attacks by any unauthorized party, this approach does not affect the accuracy of data mining results. Moreover, this technique can be applied at any time during the data collection process, so that the protection of privacy can be in effect as early as the first sample is collected.

## III. Existing System

Iterative Dichotomiser 3 (ID3) selects the test attribute based on the information gain provided by the test outcome. Information gain measures the change of uncertainty level after a classification from an attribute. Fundamentally, this measurement is rooted in information theory. Privacy preservation in data mining activities is of significant importance for many applications. However, the privacy preserving process sometimes reduces the utility of training datasets, which causes inaccurate data mining results. Privacy preservation approaches focus on different areas of a data mining process, and data mining methods also vary. This thesis focuses on privacy protection of the training samples applied for decision tree data mining.

This work presents a new privacy preserving approach via dataset complementation[1], in which the universal set is generated from the original samples. It removes each sample from a set of perturbing datasets. During the privacy preserving process, this set of perturbed datasets is dynamically modified. As the sanitized version of the original samples, these perturbed datasets are stored to enable a modified decision tree data mining method. This method guarantees to provide the same data mining outcomes like the originals, which is mathematically proved and also by a test using one set of sample datasets in this thesis. From the viewpoint of privacy preservation, the original datasets can only be reconstructed in their entirety if someone has all perturbed datasets, which is not supposed to be the case for an unauthorized party.

### 1. Unrealized Training Set Completion

To unrealized the samples, we initialize both set of input sample dataset(T') and perturbing dataset($T^p$)as empty sets, i.e. Unrealized training set ($T^u$) is called.

| workclass | race | sex |
|---|---|---|
| Private | Black | Female |
| Private | White | Female |
| State gov | White | Male |
| Local gov | Black | Female |
| State gov | white | Male |

Figure 2 – Original Table (T')

| workclass | race | sex |
|---|---|---|
| Private | White | Female |
| Private | Black | Male |
| State gov | Black | Female |
| Local gov | Black | Male |
| State gov | White | Male |
| Local gov | White | Female |

Figure 3 – Perturb Table ($T^p$)

| workclass | race | sex |
|---|---|---|
| Private | Black | Male |
| Private | White | Male |
| State gov | White | Male |
| Local gov | White | Female |
| State gov | Black | Male |

Figure 4 –unrealized Table ($T^u$)

Universal set is generated by using the single instance of all the possible values of the original data set. Consistent with the procedure described above, universal dataset is added as a parameter of the function because reusing pre-computed universal datasetis more efficient than recalculating universal dataset [1]. The recursive function unrealized training-set takes one dataset in input sample dataset in a recursion without any special requirement; it then updates perturbing dataset [11]and set of output training data sets correspondent with the next recursion. Therefore, it is obvious that the unrealized training set process can be executed at any point during the sample collection process. The following figures show the steps for the unrealized training set construction.

## IV. Proposed System

FFDs enable de-generalization of the anonymized data and thus lead to privacy breaches. Based on the impact of FFDs [21], [24] to privacy, we distinguish "safe" FFDs that cannot enable any FFD-based attack from the "unsafe" ones that can. The overall proposed work is shown in Fig. 5. The research work presented here uses the C5.0 Algorithm for data mining.

### 1. Micro Data Attributes

Micro data are analyses for privacy preservation shows that there are various dependencies in data like functional dependencies (FDs), conditional functional dependencies (CFDs), matching dependencies (MDs) which can be used as part of adversary knowledge to violate privacy. The aim of this paper is to study the various techniques in finding and preservation privacy breaches in these data. . Classification of Attributes in Micro data: [21]
The attributes in the micro data can be classifies as follows based on the information contained in them:

- Key attributes/primary data/identifier
- Quasi identifier
- Sensitive attribute

Examples of Key attribute: Name, address, phone number - uniquely identifying.

Examples of Quasi-identifiers: ZIP code, gender, birth date uniquely and which can be used for linking anonymized dataset with other datasets

Examples of Sensitive attributes: Medical records, salaries, etc. Sensitive attributes is need for the researchers, so they mostly released the micro data with these attributes.
The various attributes representation in the micro data is shown in Table.1. It shows three types of attributes as shown in the above list.

Table 1: Micro data table with attribute representation

| Key attribute | Quasi-identifier | | | Sensitive attribute |
|---|---|---|---|---|
| Name | DOB | Gender | Zipcode | Disease |
| Andre | 1/21/76 | Male | 53715 | Heart Disease |
| Beth | 4/13/86 | Female | 53715 | Hepatitis |
| Carol | 2/28/76 | Male | 53703 | Brochitis |
| Dan | 1/21/76 | Male | 53703 | Broken Arm |
| Ellen | 4/13/86 | Female | 53706 | Flu |
| Eric | 2/28/76 | Female | 53706 | Hang Nail |

### 2. Dependencies in Micro Data

Micro data contain specific data in its original form. There may be number of dependencies within micro data, which may lead to privacy breach.
The types of dependencies in micro data are:

- Full Functional dependencies (FFD)
- Matching dependencies (MDs)

#### (i). Fully Functional Dependencies

It is a type of integrity constraint. Given two attribute X and Y, R (A1, A2... An) : a relation schema , X and Y are attributes in R. r(R): a specific relation of type R, which satisfies the functional dependency (fd) $X \rightarrow Y$. if each specific relation (relational VALUE) r(R) satisfies $X \rightarrow Y$. A relation value r satisfies $X \rightarrow Y$. if each X value in r is associated with a unique Y value in r. In other words, a relation value r satisfies $X \rightarrow Y$ if for any two tuples t1 and t2 in r, $t1[X] = t2[X] \rightarrow t1[Y] = t2[Y]$.
The full functional dependency (FFD) [22], can be used as part of adversary knowledge. FFD expose the cross-attribute correlations among the data. For example FFD :Phone→Zipcodestates the fact that any two same phone numbers must correspond to the same zip code and imagine that the attacker having the knowledge of F in a micro data can bring potential vulnerability to privacy

#### (ii). Matching Dependencies

The concept of matching dependencies (MDs) [20],has recently been proposed for specifying matching rules for object identification. Similar to the functional dependencies, matching dependencies can also be applied to various data quality applications such as detecting the violations of integrity constraints. Consider a relation with schema R (A1, A2... An). Following similar syntax of FDs, we define MDs as following.A matching dependency (MD) φ has the form $(X \rightarrow Y, \lambda)$, where X and Y are two sets of attributes in relation R, and λ is a threshold pattern of similarity thresholds on attributes in $X \cup Y$,
e.g., $\lambda [A]$ denotes the similarity threshold on attribute $A \in X \cup Y$.
The MDs can be regarded as a generalization of FDs, which are based on the equality of values (i.e., having matching similarity equal to 1.0 exactly). Thus, FDs can be represented by the syntax of MDs as well.
It specifies the dependency between two set of attributes according to their matching quality measured by some similarity matching operators, such as Euclidean distance and cosine similarity [20]. we may have an MD as([Street] → [City]) which states that for any two tuples from Contacts, if they agree on attribute Street then the corresponding City attribute should match as well. The high value of MDs will help to attacker to threat the privacy of the individual.

ISSN : 2347 - 8446 (Online)
ISSN : 2347 - 9817 (Print)

### 4. d-closeness

Given two sensitive values s1 and s2, let f1 and f2 be their frequency, then s1 and s2 are considered as d-close if $|f1 - f2| \leq d$. Given a QI-group G that consists of a set of distinct sensitive values, G is d-close [21] if ∀ sensitive values si, sj∈ G, si and sj are d-close.

### 5. ℓ-diversity

Given a micro data D, let D∗ be its anonymized version. Then D∗ is ℓ-diverse [22], [23] if ∀ sensitive attribute S of D, each QI-group G ∈ D∗ consists of atleast ℓ distinct sensitive values on S that are d-close.

The attributes having functional dependencies are selected for the process. Then the l-diversity is applied to those attributes. This can be done in two ways here, first the attributes with functional dependency is selected manually and the l-diversity technique is applied. Second is the attributes with functional dependency is find out using the information gain and the entropy.

The equations for the calculation of information gain and entropy is given in this paper in future. The information gain is calculated using the entropy and it is done for all the attributes in the dataset. Then the decision tree is generated from the output of this process for the performance evaluation.

### 6. Decision Tree Algorithm C5.0

The research work presented here considers the C5.0 Algorithm for data mining. The enhancement and the optimization of the C4.5 emerge as algorithm C5.0, which exhibits the better performance as compared to the other existing mining algorithms. C5.0 algorithm to build either a decision tree or a rule set. In C5.0 model the sample is split based on the field that provides the maximum information gain. Again each subsample defined by the first split is split based on a different field, and this process repeats until the subsamples cannot be split anymore. Finally, the lowest-level splits are evaluated again, and those that do not provide significantly to the value of the model are removed or pruned. C5.0 can produce two varieties of models.

The decision tree generated here is a straightforward description of the splits found by the algorithm. Each terminal (or "leaf") node describes a particular subset of the training data and each case in this training data belongs to exactly one terminal node in the tree.

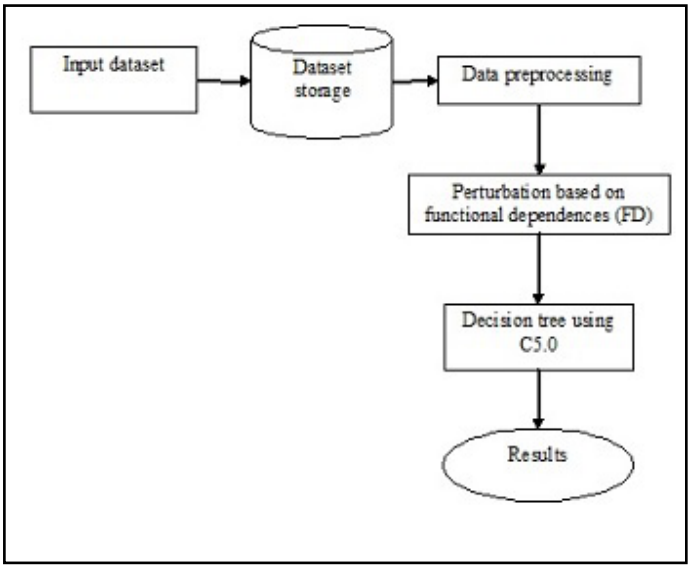


Figure 5- Overall process diagram

The C5.0 algorithm needs to calculate the entropy and the information gain of the attributes in the table. For calculating the entropy, the equations (1) is used, where $p_i$ is the probability of the class $c_i$ in the table D.

$$\text{Entropy (D)} = -\left(\sum_{i=1}^{m} p_i * log_2 p_i\right) \quad (1)$$

The Information gain is calculated using the entropy value, the equation (2) is used.

$$\text{Gain (D)} = \text{Entropy (D)} - \sum_{i=1}^{n} \frac{|D_i|}{|D|} * Entropy(D_i) \quad (2)$$

The information gain is calculated for all the attributes in the table, to find out the attribute with maximum information. Based on the attribute with maximum gain, the sample is splitted.

### V. Conclusion

This paper covers the new approach for privacy preservation using l-diversity, and decision tree mining using C5.0 algorithm which supports both discrete and continuous value attributes. It optimizes the storage size of the output data. The future work should concerns data with fully functional dependencies.

### Reference

[1] *PuiK.Fong and JensH.Weber-Jahnke, "Privacy preserving Decision tree Learning Using Unrealized Data sets", IEEE Trans. Knowledge and Data Eng., vol.24 No. 2,Feb 2012.*

[2] *S.Ajmani, R.Morris, and B.Liskov, "A Trusted Third- Party Computation Service,"Technical Report MIT-LCS-TR-847, MIT 2001.*

[3] *S.L.Wang and A.Jafari,"Hiding Sensitive Predictive Association Rules," Proc.IEEE Int'l Conf. Systems, Man and Cybernetics, pp.164-169,2005.*

[4] *R.Agrawal and R.Srikant,"Privacy Preserving Data Mining," Proc. ACM SIGMOD Conf. Management of Data (SIGMOD'00), pp.439-450,May 2000.*

[5] *Q.Ma and P.Deng,"Secure Multi-Party Protocols for Privacy Preserving Data mining," Proc. Third Int'l Conf.Wireless Algo-rithms, Systems, and Applications (WASA'08), pp.526-537, 2008.*

[6] *J.Gitanjali, J.Indumathi, N.C.Iyengar, and N.Sriman, "A Pristine Clean Cabalistic Foruity Strategize Based approach for Incre-mental Data Stream Privacy Preserving Data Mining," Proc. IEEE Second Int'l Advance Computing Conf. (IACC), pp.410-415,2010.*

[7] *N.Lomas,"Data on 84,000 United Kingdom Prisonersis Lost, "Retrieved Sept.12,2008, http://news.cnet.com/8301-1009_3-10024550-83.html, Aug.2008.*

[8] *BBC News Brown Apologises for Records Loss. Retrieved Sept. 12, 2008, http://news.bbc.co.uk /2/hi/uk_news/ politics/7104945.stm,Nov.2007.*

[9] *D.Kaplan, Hackers Steal 22,000 Social Security Numbers from Univ. of Missouri Database, Retrieved Sept.2008, http://www.scmaga- zineus.com/Hackers-steal-22000-Social-Security- numbers-from-Univ.-of-Missouri- database/ article/34964/ May2007.*

[10] *D.Goodin, "Hackers In filtrate TD Ameritrade clientDatabase," Retrieved Sept.2008, http://www. channelregister.co.uk/2007/09/15/ ameritrade_database_ burgled/, Sept.2007.*

[11] *Liu, M.Kantarcioglu, and B.Thuraisingham, "Privacy Preserving Decision Tree Mining from Perturbed Data,"Proc.42nd Hawaii Int'l Conf.System Sciences*

*(HICSS'09), 2009.*

*[12] Y.Zhu, L.Huang, W.Yang, D.Li, Y.Luo, and F.Dong, "Three New Approaches to Privacy-Preserving Add to Multiply Protocol and Its Application,"Proc.Second Int'l Workshop Knowledge Discovery and Data Mining, (WKDD'09), pp.554-558, 2009.*

*[13] J.Vaidya and C.Clifton, "Privacy Preserving Association Rule Mining in Vertically Partitioned Data," Proc Eighth ACMSIGKDD Int'l Conf. Knowledge Discovery and data Mining (KDD'02), pp.23-26,July 2002.*

*[14] M.Shaneck and Y.Kim, "Efficient Cryptographic Primitives for Private Data Mining," Proc.43rd Hawaii Int'l Conf. System Sciences (HICSS), pp.1-9, 2010.*

*[15] C.Aggarwal and P.Yu, Privacy-Preserving Data Mining:, Models and Algorithms. Springer, 2008.*

*[16] L.Sweeney, "k-Anonymity: A Model for Protecting Privacy,"Int'lJ.Uncertainty, Fuzziness and Knowledge – based Systems, vol.10, pp.557-570, May 2002*

*[17] J.Dowd,S.Xu, and W.Zhang, "Privacy- Preserving Decision Tree Mining Based on Random Substitions," Proc. Int'l Conf. Emerging Trends in Information and Comm. Security (ETRICS'06), pp.145-159, 2006.*

*[18] S.Bu, L.Lakshmanan, R.Ng, and G.Ramesh, "Preservation of Patterns and Input-Output Privacy," Proc. IEEE 23rd Int'l Conf Data Eng., pp.696-705, Apr.2007.*

*[19] S.Russell and N.Peter, Artificial Intelligence.A Modern Approach2/ E. Prentice-Hall, 2002.*

*[20] P.K.Fong, "Privacy Preservation for Training Data Sets in Database: Application to Decision Tree Learning, master'sthesis, Dept.of Computer Science, Univ.of Victoria, 2008.*

*[21] Hui Wang and RuilinLui,"Privacy Preserving Publishing Micro data with Full Functional Dependencies", Data & Knowledge Engineering 70 (2011) 249–268*

*[22] N.Li, T.Li, t-Closeness: privacy beyondk-anonymity and l-diversity, Proceedings of the International Conference on Data Engineering (ICDE), 2007,pp.106–115.*

*[23] AshwinMachanavajjhala, ohannesGehrke, Daniel Kifer, MuthuramakrishnanVenkitasubramaniam, "L-Diversity: Privacy beyond K-Anonymity"*

*[24] V. Ciriani, S. De Capitani di Vimercati, S. Foresti, and P. Samarati, "k-Anonymity" UniversitμadegliStudi di Milano, 26013 Crema*

*J.Karthikreceived B.E degree in Computer Science and Engineering from RVS College of Engineering and Technology, Anna University, Coimbatore, India in 2012. Currently, He is pursuing M.E Computer Science and Engineering, Anna University, Regional Centre, Coimbatore.*

*J.Preethi received the B.E. degree in Computer Science and Engineering from Sri Ramakrishna Engineering College, Coimbatore, Anna University, Chennai, India, in 2003, the M.E. degree in Computer Science and Engineering from the Govt. college of Technology, Anna University, Chennai, India, in 2007 and obtained Ph.D. degree in the Department of Computer Science and Engineering from the Anna University Chennai in the year 2013. Currently, she works as a Assistant Professor in the Department of Computer Science and Engineering, Anna University, Regional Centre, Coimbatore. Her research interests include Soft Computing Techniques, Medical Image Processing, Data mining and Heterogeneous Wireless Networks.*