An Efficient Privacy Preservation for Data Outflow

D. Raja Vigneshwar[1]

Abstract -Privacy preserving data mining methods apply a transformation which reduces the effectiveness of data. The goal of privacy preserving data mining is to develop algorithms to preserve and such, no confidential information remains preserved as a result of applying data mining tasks. In existing system the ID3 decision tree learning algorithm along with AIDA is used. The problem in the existing system has insufficient storage mechanism and this ID3 only can be implemented for discrete valued attributes only. The proposed work covers a new privacy preserving approach via dataset complementation which confirms the utility of training dataset for decision tree learning along with C4.5 algorithm. It converts the sample datasets into some unreal datasets such as any original dataset is not reconstructable if an unauthorized party were to steal some portion of data. Meanwhile, there remains only a low probability of random matching of any original dataset to the stolen datasets. Leakage of unreal dataset can be overcome by detecting the leakage user. Data Distribution strategy is proposed to improve the distributor's chances of identifying a leaker.

Index Terms –Monotone Framework, Unrealized Training Set, Unrealized Data Set, Entropy Determination, Security and Privacy Protection, Data Tree Generation.

INTRODUCTION

Data mining is widely used by researchers for science and business purposes. Data collected (referred to as "sample data sets" or "samples" in this paper) from individuals (referred to in this paper as "Information providers") are important for decision Making or pattern recognition. Therefore, privacy preserving processes have been developed to sanitize private information from the samples while keeping their utility. A large body of research has been devoted to the protection of sensitive information when samples are given to third parties for processing or computing [1], [2], [3], [4], [5]. It is in the interest of research to disseminate samples to a wide audience of researchers, without making strong assumptions about their trustworthiness. Even if information collectors ensure that data are released only to third parties with non- malicious intent (or if a privacy preserving approach can be applied before the data are released, see Fig. 1a), there is always the possibility that the information collectors may inadvertently disclose samples to malicious parties or that the samples are actively stolen from the collectors (see Fig. 1b). Samples may be leaked or stolen anytime during the storing process [6], [7] or while residing in storage [8], [9].

[1]PG Scholar, Department of Computer Science & Engineering, M.A.M. College of Engineering, Tiruchirappalli-621 105, Tamil Nadu, India.
E-mail: rajavigneshwar@aol.com

This paper focuses on preventing such attacks on third parties for the whole lifetime of the samples. Contemporary research in privacy preserving datamining mainly falls into one of two categories: 1) perturbation and randomization-based approaches, and 2) secure multiparty computation (SMC)-based approaches [10]. SMC approaches employ cryptographic tools for collaborative data mining computation by multiple parties. Samples are distributed among different parties and they take part in the information computation and communication process. SMC research focuses on protocol development for protecting privacy among the involved parties or computation efficiency; however, centralized processing of samples and storage privacy is out of the scope of SMC. We introduce a new perturbation and randomization-based approach 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, our approach does not affect the accuracy of data mining results. 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. The following assumptions are made for the scope of this paper: first, as is the norm in data collection processes, a sufficiently large number of sample data sets have been collected to achieve significant data mining results covering the whole research target second, the number of datasets leaked to potential attackers constitutes a small portion of the entire sample database. Third, identity attributes (e.g., social insurance number) are not considered for the data mining process because such attributes are not meaningful for

decision making. Fourth, all data collected are discredited; continuous values can be represented via ranged- value attributes for decision tree data mining. The rest of this paper is structured as follows: the next section describes privacy preserving approaches that safeguard samples in storage. Section 3 introduces our new privacy preservation approach via data set complementation. Section 4 provides the decision-tree building process applied for the new approach. Section 5 and 6 describe the

Fig.1. Two forms of information release to a third Party:
(a) The data collector sends the preprocessed Information (which was sanitized through extra Techniques, such as cryptographic approaches or Statistical database) at will

(Or)

(b) Hackers steal the original samples in storage without notifying the data collector.

2. RELATED WORK

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 storage 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. K-anonymity [15] is a data modification approach that aims to protect private information of the samples by generalizing attributes. K-anonymity trades privacy for utility. Further, this approach can be applied only after the entire data collection process has been completed. Perturbation-based approaches attempt to achieve privacy protection by distorting information from the original data sets. The perturbed data sets still retain features of the originals so that they can be used to perform data mining directly or indirectly via data reconstruction. Random substitutions [16] is a perturbation approach that randomly substitutes the values of selected attributes to achieve privacy protection for those attributes, and then applies data reconstruction when these data sets are needed for data mining. Even though privacy of the selected attributes can be protected, the utility is not recoverable because the reconstructed data sets are random estimations of the originals. Most cryptographic techniques are derived for secure multiparty computation, but only some of them are applicable to our scenario. To preserve private information, samples are encrypted by a function, f,(or a set of functions) with a key, k, (or a set of keys); meanwhile, original information can be reconstructed by applying a decryption function, f1, (or a set of functions) with the key, k, which raises the security issues of the decryption function(s) and the key(s).Building meaningful decision trees needs encrypted data to either be decrypted or interpreted in its encrypted form. The (anti)monotone framework [17] is designed to preserve both the privacy and the utility of the sample data sets used for decision tree data mining. This method applies a series of encrypting functions to sanitize the samples and decrypts them correspondingly for building the decision tree. However, this approach raises the security concerns about the encrypting and decrypting functions. In addition to protecting the input data of the data mining process, this approach also protects the output data, i.e., the generated decision tree. Still, this output data can normally be considered sanitized because it constitutes an aggregated result and does not belong to any individual information provider. In addition, this approach does not work well for discrete-valued attributes.

3. DATA SET COMPLEMENTATION APPROACH

In the following sections, we will work with sets that can contain multiple instances of the same element, i.e., with multisets (bags) rather than with sets as defined in the classical set theory. We begin this section by defining fundamental concepts (Section 3.1). We then introduce our data unrealization Algorithm in Section 3.2.

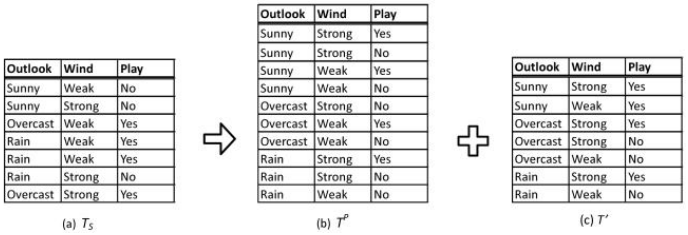


Fig 2.unrealizing training samples in (a) by calling unrealized training set. The resulting tables Tp and T′ are given in (b) &(c)

3.1 Dataset Complementation

In the following sections, we will work with sets that Can contain multiple instances of the same element, i.e., with multisets (bags) rather than with sets as Defined in the classical set theory. We then introduce our data unrealization algorithm in Section 3.2.

Outlook	Wind	Play
Sunny	Weak	No
Sunny	Strong	No
Overcast	Weak	Yes
Rain	Weak	Yes
Rain	Weak	Yes
Rain	Strong	No
Overcast	Strong	Yes

(a) T_S

Outlook	Wind	Play
Sunny	Strong	Yes
Sunny	Strong	No
Sunny	Weak	Yes
Sunny	Weak	No
Overcast	Strong	No
Overcast	Weak	Yes
Overcast	Weak	No
Rain	Strong	Yes
Rain	Strong	No
Rain	Weak	No

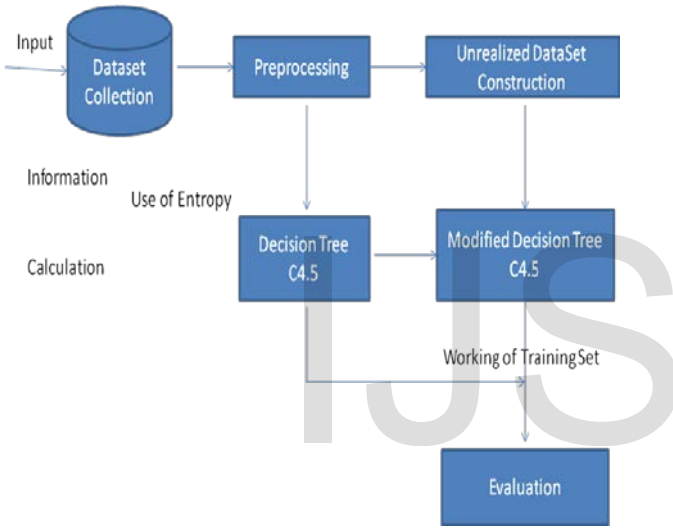
(b) T^P

Outlook	Wind	Play
Sunny	Strong	Yes
Sunny	Weak	Yes
Overcast	Strong	Yes
Overcast	Strong	No
Overcast	Weak	No
Rain	Strong	Yes
Rain	Weak	No

(c) T'

Fig.2. Unrealizing training samples in (a) by calling Unrealizing-Training Set (T_S, T^U, {}, {}) the resulting tables T^P and T′ are given in (b) and (c).

System Architecture



3.2 Unrealized Training Set

Traditionally, a training set, T_s, is constructed by inserting sample data sets into a data table. However, a dataset complementation approach as presented in this paper requires another data table; Tp.Tp is a perturbing set that generates unreal datasets which are used for converting the sample data into unrealized training set T′.To unrealize the samples,T_s,we, we initialize both T′ and T^P as empty sets. The resulting unrealized training set contains some dummy datasets expecting the ones in T_s.The elements in the resulting datasets are unreal individually, but meaningful when they are used together. To calculate the information required by a modified C4.5 algorithm, this will be covered in section 4.

4. DECISION TREE GENERATION

The well-known C4.5 algorithm shown above builds a decision tree by calling algorithm Choose Attribute recursively. This algorithm selects a test Attribute (with the smallest entropy) according to the Information content of the training set T_s. Decision tree builds classification or regression models in the form of tree structure. It breaks down data structure into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. Result is a tree with decision nodes and leaf nodes. A decision (e.g., outlook) has two or more branches (e.g., overcast, sunny and rainy). Leaf node (e.g., play) represents a classification or decision. The topmost decision node in which corresponds to the best predictor called root node. Decision trees can handle both categorical and numerical data

$$H_{a_i}(T_S) = -\sum_{e \in K_i} (|T_{S_{(a_i=e)}}|/|T_S|) \log_2 (|T_{S_{(a_i=e)}}|/|T_S|), \qquad (1)$$

and

$$H_{a_i}(T_S|a_j) = \sum_{f \in K_j} (|T_{S_{(a_j=f)}}|/|T_S|) H_{a_i}(T_{S_{(a_j=f)}}), \qquad (2)$$

Where K_i and k_j are the sets of possible values for the decision attribute, a_i , and the test attribute, a_j ,in T_S, respectively, and the algorithm Majority-Value retrieves the most frequent value of the decision attributes of T_S.

Algorithm

The core algorithm for building decision trees called C4.5 is an extension of ID3. As decision trees generated by C4.5 can be used for classification is often referred to as statistical classifier. It uses the concept of information entropy. The training data is a set $S=s_1,s_2,\ldots$ of already classified samples. Each sample S_i consists of p-dimensional vector$(X_{1,i},X_{2,i},\ldots X_{p,i})$,where the X_j represent attribute or features of the sample, as well as the class in which S_i falls.

Improvement

C4.5 has been improved by overcoming the obstacles of ID3 are as follows. Handling both continuous and discrete attributes, i.e., In order to handle continuous attributes, C4.5 creates a threshold and then splits the list into those whose attribute value is above the threshold and those that are less than or equal to it. Handling training data with missing attribute values. Handling attributes with differing costs. Pruning trees after creation, i.e., C4.5 goes back through the tree once it's been created and attempts to remove branches that do not help by replacing them with leaf nodes.

4.1 Information Entropy Determination

From the algorithm Unrealizing –Training-set, it is obvious that the size of T_S is the same as the size of T′. Furthermore, all data set in $(T'+T^P)$ are based on the data sets in TU, expecting the ones in T_S is the q-absolute-complement of $(T'+T^P)$ for some positive integer q. According to theorem 2, the size of qT^U can be computed from the size of T′ and T^P , with $qT^U = 2*|T'| + |T^P|$. Therefore, entropies of the original data sets, T_S , with any decision attribute and any test attribute, can be

determined by the unreal training set, T', and pertaining set, T^P.

Information Gain

Information gain IG(A) is the measure of the difference in entropy from before to after the set S is split on an attribute A. In other words, how much uncertainty in S was reduced after splitting set S on attribute A.

IG(A) = H(S) - \sum_{t \in T} p(t)H(t)

Where,

H(S) - Entropy of set S

T - The subsets created from splitting set S by attribute A such that S = \bigcup_{t \in T} t

p(t) - The proportion of the number of elements in t to the number of elements in set S

H(t) - Entropy of subset t

The attribute with the largest information gain is used to split the set S on this iteration.

4.2 Modified Decision Tree Generation Algorithm

As entropies of the original data sets,T_S, can be determined by the retrievable information- content of the unrealized training set,T' , and perturubating set,T^P- the dicision tree T_S can be generated by the following algorithm.

Algorithm. Generate-Tree' (size,T',T^P,attribs,default)

Input: Size of qT^U

T', the set of unreal training data sets

T^P, the set of perturbing data sets

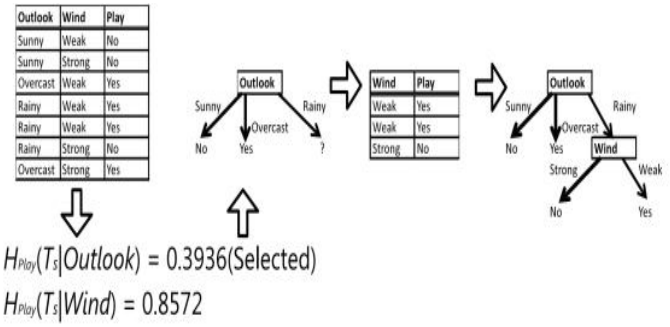


Fig 3.Illustration of generate-tree process by applying the conventional C4.5 approach with original samples T_S.

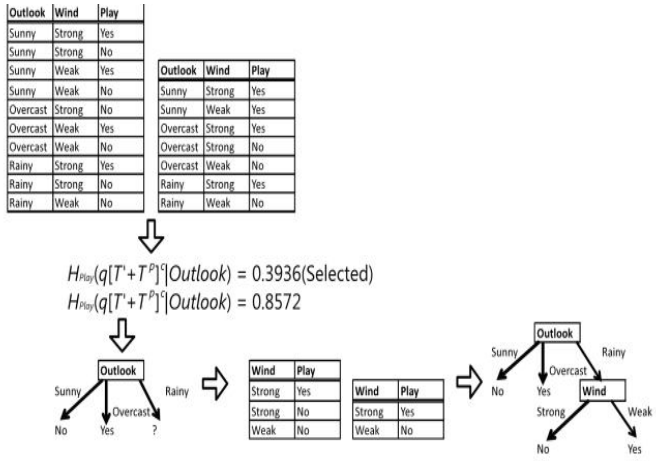


Fig 4.Illustration of generate-tree process by applying the modified C4.5 approach with unrealized samples(T'+T^P).for each step the entropy values and resulting sub trees are exactly same as the result of traditional approach

Similar to the traditional C4.5 approach, algorithm choose-attribute selects the test attribute using the c4.5 criteria, based on information entropies, i.e., selecting the attribute with greatest information gain. Algorithm minority-value retrieves the least frequency value of the decision attribute of (T'+T^P),which performs the same function as algorithm majority-value of the tradition c4.5 approach, i.e., receiving the most frequent value of the decision attribute of T_S.

Fig 4 shows the resulting decision tree of our new c4.5 algorithm with unrealized sample inputs shown in

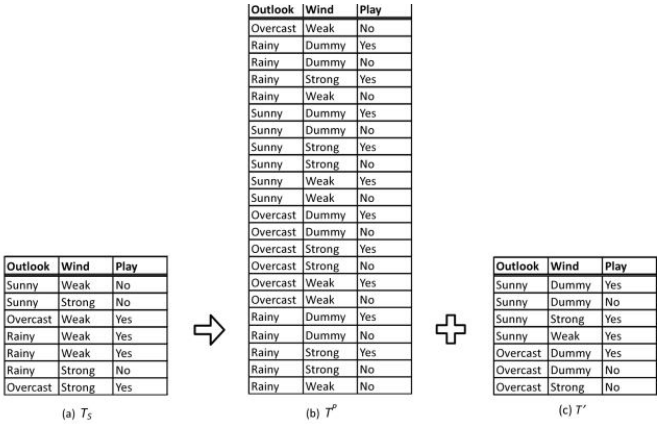
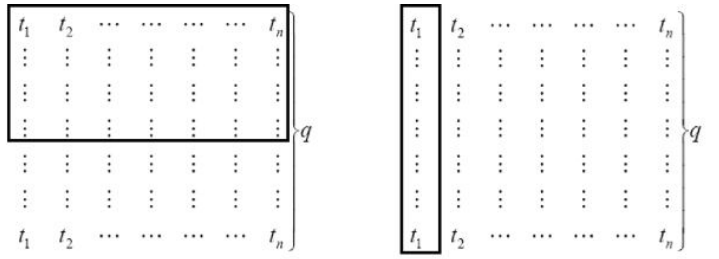


Fig 5.Unrealizing training samples in(a)with a dummy value .the resulting value Tp and T'are shown in (b) and (c)



(a) Every dataset in T_S has frequency ($|T_S|/T^U|$). (b) One dataset in T_S has frequency $|T_S|$.

Fig 6. (a) The case with lowest variance distribution
(b) The case with highest variance distribution

4.3 Data set Reconstruction

Section 4.2 introduced a modified decision tree learning algorithm by using the unrealized training set, T′, and the perturbing set, T_P.alternatively, we could have reconstructed the original sample data sets, T_s, from T′ and Tp, followed by application of the conventional c4.5 algorithm for generating decision tree from T_P.The reconstruction process is dependent upon the full information of T′ and T_P.

4.4 Enhanced Protection with Dummy Values

Dummy values can be added for any attribute such that the domain of the perturbed sample data sets will be expanded while the addition of dummy values will have no impact on T_s. For example, we can expand the possible values of attribute wind from [strong, weak] to [dummy, strong, weak] where dummy represents a dummy attribute value that plays no role in the data collection process. In this way we can keep the same resulting decision tree.

5. EXPERIMENTS

5.1 Output Accuracy

In All cases decision tree(s) generated from the unrealized samples is the same as the decision tree(s), generated from the original sample by the regular method.

5.2 Storage Complexity

From the experiment, the storage requirement for the data set complementation approach increases from$|T_L|$ to $(2|T_u|-1)*|T_{s.}|$, while the required storage may be doubled if the dummy attribute values technique is applied to double the sample domain. The best case happens when the samples are evenly distributed, as the storage requirement is the same as for the originals. The worst case happens when the samples are distributed extremely unevenly. Based on the randomly picked tests, the storage requirement for our approach is less than five times (without dummy values) and eight times (with dummy values, doubling the sample domain) that of original samples.

5.3 Privacy Risk

Without the dummy attribute technique, the average privacy loss per leaked unrealized data set is small, except for even distribution case. By doubling the sample domain, the average privacy loss for a single leaked data set is zero, as the unrealized samples are not linked to any information provider. The randomly picked tests show that the data set complementation approach eliminates the privacy risk for most cases and always improves privacy security significantly when dummy values are used.

6. CONCLUSION

We introduced a new privacy preserving approach via data set complementation which confirms the utility of training data sets for decision tree learning. This converts the sample data set Ts, into some unreal data sets (T′+T_P).such that any original data is not reconstuctable if an unauthorized party were to steal some portion of (T′+T_P). Meanwhile there remains only a low probability of random matching of any original data set to the stolen data sets, T_L.

REFERENCES

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

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

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

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

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

[6] N. Lomas, "Data on 84,000 United Kingdom Prisoners is Lost," Retrieved Sept. 12, 2008, http://news.cnet.com/8301- 1009_3-10024550-3.html, Aug. 2008.

[7] 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.

[8] 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-Missouridatabase/article/34964/, May 2007.

[9] D. Goodin, "Hackers Infiltrate TD Ameritrade client Database," Retrieved Sept. 2008, http://www.channelregister.co.uk/2007/09/15/ameritrade_database_burgled/, Sept. 2007

[10] L. 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.