



International Journal of Emerging Technologies in Computational and Applied Sciences (IJETCAS)

www.iasir.net

Decision Tree Classifier for Privacy Preservation

Tejaswini Pawar¹, Prof. Snehal Kamlapur²
^{1,2}Department of Computer Engineering,
K.K.W.I.E.E&R, Nashik,
INDIA

Abstract: In recent year's privacy preservation in data mining has become an important issue. A new class of data mining method called privacy preserving data mining algorithm has been developed. The aim of these algorithms is to protect the sensitive information in data while extracting knowledge from large amount of data. We focus the general classification in a secured manner and introduce a privacy-preserving decision tree classifier using C4.5 algorithm. The entire original dataset is replaced by unreal dataset for preserving the privacy via dataset complementation. This novel approach can be applied directly to the data storage as soon as the first sample is collected and applied at any time during the data collection process. This approach converts the original sample data sets into a group of unreal data sets, from which the original samples cannot be reconstructed without the entire group of unreal data sets. An accurate decision tree can be built directly from those unreal data sets. This paper discusses dataset complementation approach for converting sample dataset into sanitized or altered dataset. It covers system architecture and mathematical model of decision tree classifier for preserving privacy via dataset complementation.

Keywords: C4.5algorithm, Data Perturbation, Dataset Complementation, Decision tree classification, Privacy preserving data mining, unrealized dataset

I. Introduction

Data mining is a recently emerging field, connecting the three worlds of databases, statistics and artificial intelligence. Data mining is the process of extracting knowledge or pattern from large amount of data. It is widely used by researchers for science and business process. The data collection process takes time and efforts hence sample datasets are sometime stored for reuse. Data collected from information provider are important for pattern recognition and decision making. However attacks are attempted to steal these sample datasets and private information may be leaked from these stolen datasets. The privacy disclosure problem about company or individual is inevitably exposed while releasing or sharing data to mine useful decision information because of rapidly increased development of data analysis and processing technique. Some unauthorized parties try to access sample datasets and uses private information from stolen dataset. It gives the birth to the research field on privacy preserving data mining and knowledge.

Privacy preservation converts dataset containing private information into altered or sanitized versions in which private information is hidden from unauthorized parties. Privacy preserving data mining refers to the area of data mining that seeks to safeguard sensitive information from unsanctioned or unsolicited disclosure. Privacy Preservation Data Mining [1] [2] was introducing to preserve the privacy during mining process to enable conventional data mining technique. Many privacy preservation approaches were developed to protect private information of sample dataset. On the other hand privacy preserving process which hides information may reduce utility of these sanitized dataset. When there utility decreases to a certain level the downgraded information prevents accurate analysis. With the result that primary objective of data mining is compromised. Even when databases of samples with sensitive information are protected securely partial information of database can be lost through procedural mistakes or privacy attacks from anywhere within the network.

The rest of this paper is structured as follows: section-2 describes privacy preserving approaches that safeguard samples in storage and presents review on different privacy preservation technique through decision tree approach. Section 3 revised privacy preservation approach via data set complementation. Detail system design is included in Section-4. It includes system architecture and different functions of proposed system. Section-5 provides an overall summary of this paper. Bibliography covers all the references require to study this seminar.

II. Background and Related work

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]. Samples may be leaked or stolen anytime during the storing process [6], [7] or while residing in storage [8], [9]. Hence protections of such data samples become increasingly important. Different privacy preserving techniques are design to protect private information of

sample dataset. These techniques mainly divided into two types: first anonymization method and second randomization method. Anonymization operation design to prevent identification, hides some details in quasi identifier attribute and for categorical data, a specific value can be replaced with general value. The randomization method is privacy preserving technique in which attribute values of record are masked by adding noise or by substituting random value or by using data complementation approach. These techniques are developed to derive aggregate distribution from perturbed record because noise added or substituted is sufficiently large so that individual record values cannot be recovered. Noise Addition [27] uses a random number or noise which is use for adding into the numerical attribute for creating perturbed dataset. Random number is generally drawn from a small deviation and normal distribution with 0-mean. This strategy usually use for numeric values hence it has preserve little privacy. Random substitution [24] replaces sample by randomly replacing values of attributes. This technique uses invertible matrix M having size $n * n$ called as perturbation matrix where n is the no of possible values of an attribute that is being perturbed The optimal perturbation matrix is gamma diagonal matrix $M = x * G$ where $x = 1/(\lambda + n - 1)$. The information related to a particular attribute in the perturbation datasets is irrelevant to that of the original datasets. If we were to protect the privacy of other attributes by random-substituting their values, the accuracy of the final decision tree would decrease further. Pui K. Fong and Jens H. Weber-Jahnke [16], [17], [29], [30] introduce a new perturbation and randomization based approach that protects centralized sample data sets utilized for decision tree data mining. They introduced a new privacy preserving approach via data set complementation. This approach converts the original sample data sets into a group of unreal data sets. The decision tree can be built directly from the sanitized data sets, no need to be reconstructing the original dataset. This technique requires extra storage for storing perturbed and complement of sample data set. So optimizing the storage size of the unrealized samples needs to be explored.

Privacy preservation decision tree learning techniques [3], [10], [11] in which firstly data is modified by using different data modification and perturbation-based approaches [12], [13], [14], [15], and then decision tree mining [16], [17], [18], [19], [20], [21], [22] is applied to modified or sanitized dataset.

III. Data set complementation approach

In the following sections, we will work with sets that can contain multiple instances of the same element. We begin this section by defining fundamental concepts then data unrealization algorithm.

Universal set and Data set complementation[30]:

Definition1. T^U , the universal set of data table T , is a set containing a single instance of all possible datasets in data table T .

Example1. If data table T associates with a tuple of attributes $\langle \text{Wind, Play} \rangle$ where $\text{wind} = \{\text{Strong, Weak}\}$ and $\text{play} = \{\text{Yes, No}\}$, then $T^U = \{\langle \text{Strong, Yes} \rangle, \langle \text{Strong, No} \rangle, \langle \text{Weak, Yes} \rangle, \langle \text{Weak, No} \rangle\}$

Remark1. If data table T associated with a tuple of m attributes $\langle a_1, a_2, \dots, a_m \rangle$ where a_i has n_i possible values and $1 \leq i \leq m$ then $T^U = n_1 * n_2 * \dots * n_m$

Definition2. If T_D is subset of T and q is a positive integer, the q -multiple-of T_D , denoted as qT_D , is a set of data sets containing q instances of each dataset in T_D .

Remark2. $|qT^U| = q * |T^U|$

Definition3. If k is a possible value of attribute a and l is a possible value of attribute b in T , then $T_{(a=k)}$ denotes a subset of T that contains all data sets with attribute a equals k . Similarly, $T_{(a=k) \wedge (b=l)}$ denote a subset of T that contains all data sets with attribute a equals k and b equals l .

Theorem1. If k_i is a possible value of attribute a_i in T , then $|qT_{(a_i=k_i)}^U| = (q * n_1 * n_2 * \dots * n_m) / n_i$.

Proof: $|qT_{(a_i=k_i)}^U|$

$$= q * |T_{(a_i=k_i)}^U|$$

$$= q * n_1 * n_2 * \dots * n_{i-1} * n_{i+1} * \dots * n_m$$

$$= (q * n_1 * n_2 * \dots * n_m) / n_i$$

Corollary1. $|qT_{(a_i=k_i) \wedge \dots \wedge (a_j=k_j)}^U|$

$$= (q * n_1 * n_2 * \dots * n_m) / (n_i * \dots * n_j)$$

Definition4. If T_D is subset of T , then absolute complement of T_D , denoted as T_D^c , is equal to $T^U - T_D$, and a q -absolute-complement of T_D denoted as qT_D^c is equal to $qT^U - T_D$.

The size of qT^U can be computed from the sizes 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 perturbing set, T^P .

Theorem: If $T_S = q[T' + T^P]^C$ and $|T'| = |T_S|$ for some positive integer q , then $|qT^U| = 2|T'| + |T^P|$

Proof: $T_S = q[T' + T^P]^C$
 $\Rightarrow T_S = qT^U - (T' + T^P)$
 $\Rightarrow |T_S| = |qT^U - (T' + T^P)|$
 $\Rightarrow |T_S| = |qT^U| - |(T' + T^P)| \quad \because (T' + T^P) \subseteq qT^U$
 $\Rightarrow |T_S| = |qT^U| - |T'| - |T^P|$
 $\Rightarrow |T'| = |qT^U| - |T_S| - |T^P| \quad \because |T'| = |T_S|$
 $\Rightarrow |qT^U| = 2 * |T'| + |T^P|$

Algorithm for data unrealization:

Dataset Complementation approach [30] was designed for discrete value classification so continuous values are replaced with ranged values. The entire original dataset is replaced by unreal dataset for preserving the privacy via dataset complementation. 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 original accuracy of training dataset is preserved without linking the perturbed dataset to the information provider i.e. accurate data mining results are yields while preserving privacy of individual's records by dataset complementation approach.

A data complementation approach requires an extra table T^P for converting sample dataset T_S into an unrealized training set T^1 . T^P is perturbing set that generates unreal dataset.

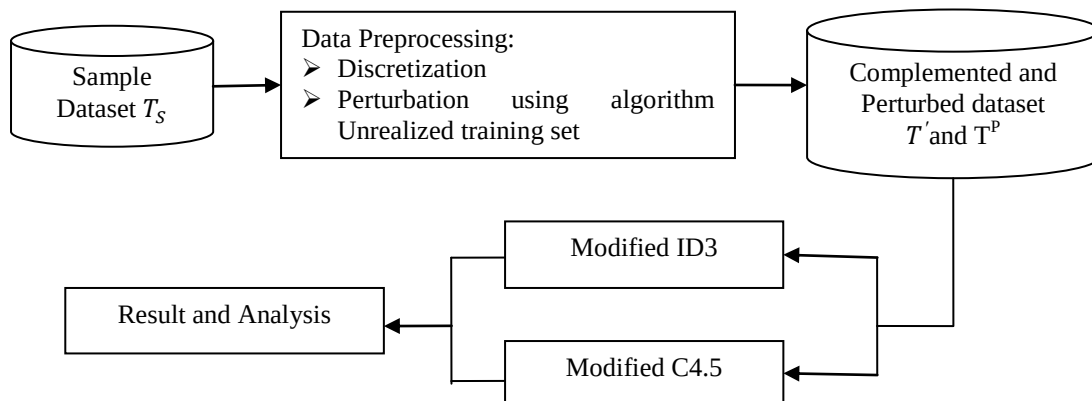
Initially T^1 and T^P are an empty set. When we get an T_S the T^P is constructed with universal set T^U by adding T^U into T^P . Whenever we get sample data item t in T_S we remove it from T^P and transfer one data item t_i^1 into T^1 . t_i^1 is the latest available frequent data item in T^P . When traversing T^P is finished and if sample data item t^1 is not available in T^P then again add universal set T^U into T^P .

To unrealized the samples, T_S , initialize both T' and T^P as empty sets, i.e., invoke the above algorithm with *Unrealize_Training_set*($T_S, T^U, \{\}, \{\}$). 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 and C4.5 algorithm. Text Font of Entire Document

IV. System design

Architectural design:

Figure 1. System Architecture



The system architecture of proposed system is shown in figure 6.1. It consist of two main module first is data preprocessing and second is decision tree generation. In data preprocessing module initially continuous value attribute dataset is converted into discrete value after that the dataset is converted into sanitized version by using algorithm unrealized training set. Then generated complemented dataset and perturbed dataset is given as a input to decision tree generation module in which decision tree is built by using ID3 and C4.5 and result generated by both the algorithm is compared to analyze the algorithm.

Functions and data dependency:

Decision tree learning based on dataset complementation approach for preserving privacy with assumptions that large number of sample data sets has been collected to achieve significant data mining results. Preprocessing of

sample dataset is based on discrete function and unrealized function. Decision tree is construction is based on optimal node selected after calculating information gain and gain ratio. Input to decision tree classifier will be perturbed dataset and complement of dataset given by function unrealized.

Dataset: Any sample dataset $T_S: \{a_1, a_2 \dots an\}$.

d : Discrete dataset, c : complement of sample dataset, p : Perturbed dataset, if : Information gain, S : Function which holds, gr : Gain ratio, dt_id3 : Tree_id3, $dt_c4.5$: Tree_C4.5

Let, $S: \{d, c, p, if, gr, dt1, dt2\}$

Let, $d: \{d_1, d_2 \dots dn\}$ where d consist of discrete value dataset

Let, $c: \{c_1, c_2, \dots cn\}$ where c consist complement of sample dataset

Let, $p: \{p_1, p_2, \dots ps\}$ where p consist of total perturbed dataset

Let, $if: \{if_{11}, if_{12}, \dots if_{nm}\}$ where if consist of information gain of terms present in all ci and pi

Let, $gr: \{gr_{11}, gr_{12}, \dots gr_{nm}\}$ where gr consist of gain ratio of terms present in all ci and pi

Let, $dt_id3: \{id_{11}, id_{12}, \dots id_{nm}\}$ where dt_id3 consist of decision tree node generated from ID3

Let, $dt_c4.5: \{cd_{11}, cd_{12}, \dots cd_{nm}\}$ where $dt_c4.5$ consist of decision tree node generated from C4.5

Function $f1$ load the sample dataset imported from user.

$f1(T_S) \rightarrow a$ for importing dataset

e.g. $f1(T_S) \rightarrow \{a_1, a_2, \dots an\} \in T_S$

Function $f2$ returns the discrete dataset designed from sample dataset

$f2(T_S) \rightarrow d$ for converting continuous value attribute into discrete

e.g. $f2(T_{S1}) \rightarrow \{d_1, \dots dn\} \in d$

Function $f3$ returns complement of dataset and perturbed dataset.

$f3(d) \rightarrow c, p$

e.g. $f3(d) \rightarrow \{c_1, c_2, \dots cn\} \cup \{p_1, p_2, \dots ps\}$

Function $f4$ consist information gain of each attribute.

$f4(c, p) \rightarrow if$

e.g. $f4(c_1, p_1) \rightarrow \{if_1, \dots if_n\} \in if$

Function $f5$ consist gain ratio of each attribute.

$f5(c, p) \rightarrow gr$

e.g. $f5(c_1, p_1) \rightarrow \{gr_1, \dots gr_n\} \in gr$

Function $f6(dt_id3)$ consist all the nodes that belong to decision tree build from ID3.

$f6(if) \rightarrow id$

e.g. $f6(if_1) \rightarrow \{id_1, \dots id_n\} \in dt_id3$

Function $f7(dt_c4.5)$ consist all the nodes that belong to decision tree build from C4.5.

$f7(gr) \rightarrow cd$

e.g. $f7(gr_1) \rightarrow \{cd_1, \dots cd_n\} \in dt_c4.5$

Functional Dependency of the above functions

Table1. Functional Dependency

	$F1$	$F2$	$F3$	$F4$	$F5$	$F6$	$F7$
$F1$	1	0	0	0	0	0	0
$F2$	1	1	0	0	0	0	0
$F3$	0	1	1	0	0	0	0
$F4$	0	0	1	1	0	0	0
$F5$	0	0	1	1	1	0	0
$F6$	0	0	0	1	0	1	0
$F7$	0	0	0	0	1	0	1

Function Dependency Graph:

There are seven main functions which are given below:

i. Getting dataset input from user ($F1$).

This function imports/load the dataset which is stored by user.

ii. Converting sample dataset into discrete dataset ($F2$).

This function is used to convert continuous value attribute into discrete value attribute.

iii. Converting discrete value dataset into perturbed and complemented dataset ($F3$).

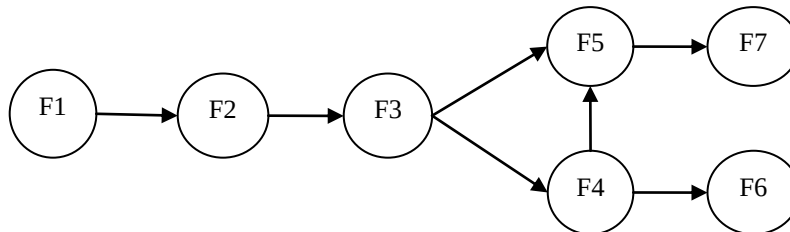
This function is used to hide the sensitive information of sample dataset hence it converts discrete value dataset into perturbed and complemented dataset.

iv. Finding the Information gain ($F4$).

This function is used to calculate information gain of each attribute.

- v. Finding the gain ratio (F5).
This function is used to calculate gain ratio of each attribute using entropy and information gain.
- vi. Building decision tree using ID3 (F6).
This function is used to build decision tree by selecting highest information gain as a splitting criteria.
- vii. Building decision tree using C4.5 (F7).
This function is used to build decision tree by selecting highest gain ratio as a splitting criteria.

Figure2. Function dependency graph



V. Conclusions

This paper covers the system architecture and mathematical model of decision tree classifier for preserving privacy via dataset complementation. The entire original dataset is replaced by using algorithm Unrealize_training_set. This approach converts the original sample data sets into a group of unreal data sets, from which the original samples cannot be reconstructed without the entire group of unreal data sets. 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.

References

- [1] S. Ajmani, R. Morris, and B. Liskov, "A Trusted Third-Party Computation Service," *Technical Report MIT-LCS-TR-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 Approach for 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_3-10024550-83.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.scmagazineus.com/Hackers-steal-22000-Social-Security-numbers-from-Univ.-of-Missouri-database/article/34964/>, May 2007.
- [9] D. Goodin, "Hackers Infiltrate TD Ameritrade client Database," Retrieved Sept. 2008, <http://www.channelregister>
- [10] Y. Lindell and B. Pinkas. Privacy preserving data mining. In *Advances in Cryptology*, volume 1880 of *Lecture Notes in Computer Science*, pages 36-53. Springer-Verlag, 2000.
- [11] Charu C Agrwal, Philip S Yu., " Privacy preserving data mining models and Algorithms.", *Springer Science+Business media, LLC*, 2008
- [12] L. Sweeney, "k-Anonymity: A Model for Protecting Privacy," *Int'l J. Uncertainty, Fuzziness and Knowledge-based Systems*, vol. 10, pp. 557-570, May 2002.
- [13] Bayardo, R. J., & Agrawal, R. (2005). Data Privacy Through Optimal k-Anonymization. In *Proc. of the 21st International Conference on Data Engineering (ICDE '05)*, pp. 217-228.
- [14] AGRAWAL, D. AND AGGARWAL, C. C. 2001. On the design and quantification of privacy preserving data mining algorithms. In *Proceedings of the 20th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems (PODS'01)*. ACM, New York, NY, 247-255.
- [15] KARGUPTA, H., DATTA, S., WANG, Q., AND SIVAKUMAR, K. 2003. On the privacy preserving properties of random data perturbation techniques. In *Proceedings of the 3rd IEEE International Conference on Data Mining (ICDM'03)*. IEEE Computer Society, Los Alamitos, CA, 99.

- [16] Lior Rokach and Oded Maimon "Top-Down Induction of Decision Trees Classifiers – A Survey", *IEEE TRANSACTIONS ON SYSTEMS, MAN AND CYBERNETICS: PART C, VOL. 1, NO. 11*, NOVEMBER 2002
- [17] Jiawei Han, Micheline Kamber, "Data Mining: Concepts and Techniques", Morgan Kaufmann, 2001.
- [18] Lior Rokach, Oded Maimon Data Mining and Knowledge Discovery Handbook Second edition pages 167-192, Springer Science + Business Media,2010
- [19] L. Breiman, J. Friedman, R. Olshen, and C. Stone. Classification and Regression Trees. Wadsworth Int. Group, 1984.
- [20] J.R. Quinlan, Simplifying decision trees, *International Journal of Man- Machine Studies*, 27, 221-234, 1987.
- [21] L. Hyafil and R. L. Rivest, "Constructing Optimal Binary Decision Trees is {NP}-Complete," *Inf. Process. Lett.*, vol. 5, pp. 15-17, 1976.
- [22] J. R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, 1993.
- [23] Hillol Kargupta, Souptik Datta, Qi Wang, and Krishnamoorthy Sivakumar. On the privacy preserving properties of random data perturbation techniques. In *IEEE International Conference on Data Mining*, 2003.
- [24] J. Dowd, S. Xu, and W. Zhang, "Privacy-Preserving Decision Tree Mining Based on Random Substitutions," *Proc. Int'l Conf. Emerging Trends in Information and Comm. Security (ETRICS '06)*, pp. 145-159, 2006.
- [25] Evfimievski, J. Gehrke, and R. Srikant. Limiting privacy breaching in privacy preserving data mining. In *ACM Symposium on Principles of Database Systems*, pages 211–222. ACM, 2003.
- [26] Shipra Agrawal and Jayant R. Haritsa. A framework for high-accuracy privacy preserving mining. In *IEEE International Conference on Data Engineering*, 2005.
- [27] Mohammad Ali Kadampur, Somayajulu D.V.L.N., "A Noise Addition Scheme in Decision Tree for Privacy Preserving Data Mining", *Journal of Computing*, Vol 2, Issue 1, January 2010, ISSN 2151-9617
- [28] Lian Liu.,Jie Wang.,Jun Zhang., "Wavelet based data perturbation for simultaneous privacy preserving and statistics preserving.," In *Proceedings of IEEE International Conference on Data Mining workshop.*, 2008.
- [29] S. Russell and N. Peter, Artificial Intelligence. A Modern Approach 2/E. Prentice-Hall, 2002.
- [30] Pui K. Fong and Jens H. Weber-Jahnke, "Privacy Preserving Decision Tree Learning Using Unrealized Data Sets" .” *IEEE Transl. on knowledge and data engineering*, vol. 24, no. 2, February2012.