A Survey on Privacy Preserving Decision Tree Classifier

Tejaswini Pawar*, Prof. Snehal Kamalapur **
* Department of Computer Engineering, K.K.W.I.E.E&R, Nashik,
**Department of Computer Engineering, K.K.W.I.E.E&R, Nashik,

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 protecting the sensitive information in data while extracting knowledge from large amount of data. The extracted knowledge is generally expressed in the form of cluster, decision tree or association rule allow one to mine the information. Several data modification technique like randomization method, anonymization method, distributed privacy technique have been developed to incorporating privacy mechanism and allow to hide sensitive pattern or itemset before data mining process is executed. This paper mainly focuses on general classification technique decision tree classifier for preserving privacy. It presents a survey on decision tree learning on various privacy techniques.

Keywords - Data Mining, Decision Tree Classifiers, Perturbation, Privacy Preserving Data Mining, Splitting Criteria

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. Data collected from information providers are important for pattern reorganization and decision making. The data collection process takes time and efforts hence sample datasets are sometime stored for reuse. However attacks are attempted to steal these sample datasets and private information may be leaked from these stolen datasets. Therefore privacy preserving data mining are developed to convert sensitive datasets into sanitized version in which private or sensitive information is hidden from unauthorized retrievers.

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.

The rest of this paper is structured as follows: the next section describes Decision tree classifier and gives brief introduction about decision tree algorithm. Section 3 introduces different privacy preservation technique through decision tree approach. Section 4 provides an overall summary of this paper, and suggests directions for further research on this topic.

II. DECISION TREE CLASSIFIER

A decision tree[3][4][5] is defined as "a predictive modeling technique from the field of machine learning and statistics that builds a simple tree-like structure to model the underlying pattern of data".

Decision tree is one of the popular methods is able to handle both categorical and numerical data and perform classification with less computation. Decision trees are often easier to interpret. Decision tree is a classifier which is a directed tree with a node having no incoming edges called root. All the nodes except root have exactly one incoming edge. Each non-leaf node called internal node or splitting node contains a decision and most appropriate target value assigned to one class is represented by leaf node.

Decision tree classifier is able to break down a complex decision making process into collection of simpler decision. The complex decision is subdivided into simpler decision on the basis of splitting criteria. It divides whole training set into smaller subsets. Information gain, gain ratio, gini index are three basic splitting criteria to select attribute as a splitting point. Decision trees can be built from historical data they are often used for explanatory analysis as well as a form of supervision learning. The algorithm is design in such a way that it works on all the data that is available and as perfect as possible.

According to Breiman *et al.* [6] the tree complexity has a crucial effect on its accuracy performance. The tree complexity is explicitly controlled by the pruning method employed and the stopping criteria used. Usually, the *tree complexity* is measured by one of the following metrics:

- The total number of nodes;
- Total number of leaves;
- Tree depth;

- Number of attributes used.

Decision tree induction is closely related to rule induction. Each path from the root of a decision tree to one of its leaves can be transformed into a rule simply by conjoining the tests along the path to form the antecedent part, and taking the leaf's class prediction as the class value. The resulting rule set can then be simplified to improve its accuracy and comprehensibility to a human user [7].

Flowchart for tree based classification is shown in Fig.1

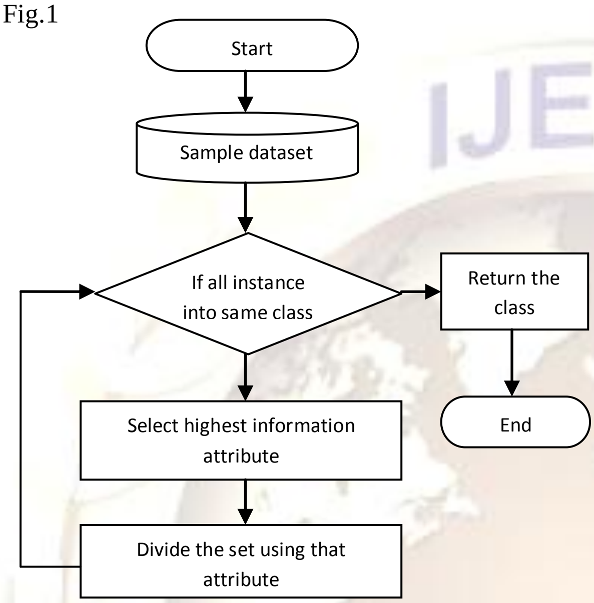
Start
Sample dataset
If all instance into same class
Return the class
End
Select highest information attribute
Divide the set using that attribute

Fig.1. Flowchart for tree based classification

Hyafil and Rivest proved that getting the optimal tree is NP-complete [8]. Most algorithms employ the greedy search and the divide-and-conquer approach to grow a tree. In particular, the training data set continues to be split in small. ID3 [4] and C4.5 [4] [9] adopt a greedy approach in which decision trees are constructed in top down recursive divide and conquer manner.

ID3 was one of the first Decision tree algorithms. It works on wide variety of problems in both academia and industry and has been modified improved and borrowed from many times over. ID3 picks splitting value and predicators on the basis of gain in information that the split or splits provide. Gain represents difference between the amount of information that is needed to correctly make a prediction both before and after the split has been made. Information gain is defined as the difference between the entropy of original segment and accumulated entropies of the resulting split segment.

C4.5 [9] is an extension of ID3, presented by the same author (Quinlan, 1993). It uses gain ratio as splitting criteria.

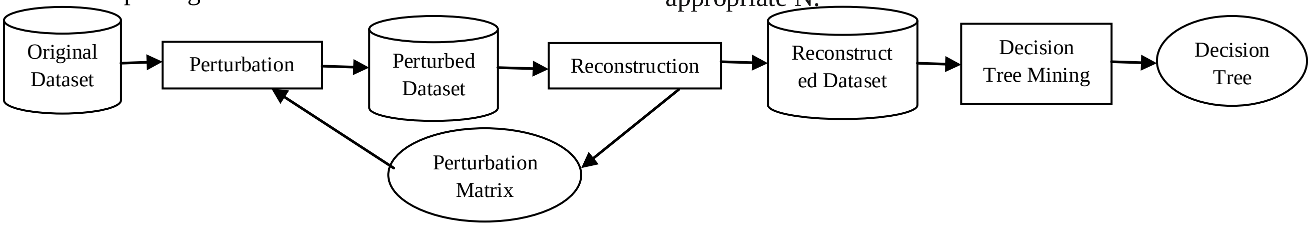

Fig. 2 A framework for privacy preserving decision tree mining

The splitting ceases when the number of instances to be split is below a certain threshold. C4.5 can handle numeric attributes. It perform error based pruning after growing phase. It can use corrected gain ratio induce from a training set that incorporates missing values.

III. PRIVACY PRESERVING DECISION TREE LEARNING

This section explores the privacy preservation decision tree learning techniques in which firstly data is modified by using different data modification and perturbation-based approaches and then decision tree mining is applied to modified or sanitized dataset.Fig.2 shows a framework for privacy preserving decision tree mining

A. Privacy-Preserving Decision Tree Mining Based on Random Substitutions:

Inspired by the fact that the pioneering privacy-preserving decision tree mining method of [1] was flawed [11], Jim Dowd, Shouhuai Xu, and Weining Zhang [10] explored a random substitution perturbation technique for privacy-preserving decision tree mining methods. This perturbation technique is based on a different privacy measure called ρ1-to-ρ2 privacy breaching [12] and a special type of perturbation matrix called the γ-diagonal matrix [13].They makes several contributions towards privacy-preserving decision tree mining. A novel error reduction technique is introduced for data reconstruction, so that it not only prevents a critical problem caused by a large perturbation matrix, but also guarantees a strictly better accuracy. In addition, the resulting privacy-preserving decision tree mining method has immune to the relevant data-recovery and repeated-perturbation attacks which were not accommodated in the model of [12]. It ensures that the decision trees learned from the perturbed data are quite accurate compared with those learned from the original data. The parameters include: (1) Privacy assurance metric γ, (2) Dimension size N of perturbation matrix, which affects the accuracy of reconstructed data as well as performance, and (3) Entropy of a perturbation matrix, which, to some extent, can encompass the effect of both γ and N. It is showed that one can achieve the desired trade-off between accuracy, privacy, and performance by selecting an appropriate N.

B. A Noise Addition Scheme in Decision Tree for Privacy Preserving Data Mining [14]:

Mohammad Ali Kadampur, Somayajulu D.V.L.N. propose a strategy that protects the data privacy during decision tree analysis of data mining process. It is basically a noise addition framework specifically tailored toward classification task in data mining. They propose to add specific noise to the numeric attributes after exploring the decision tree of the original data. The obfuscated data then is presented to the second party for decision tree analysis. The decision tree obtained on the original data and the obfuscated data are similar but by using this method the proper data is not revealed to the second party during the mining process and hence the privacy will be preserved. The method also preserves averages and few other statistical parameters thus making the modified data set useful for both data mining and statistical purposes.

This paper uses Quinlan's [15] C5.0 decision tree builder on the selected data set [16] and obtain the decision tree of the original data set and a unique method of listing the nodes (attributes) that we touch in the path from the root of the tree to the leaf, then use a noise addition strategy for each of the attributes. The approach taken in this paper integrates both categorical and numeric data types and focuses on privacy preserving during decision tree analysis. The noise addition methods used are effective in preserving the privacy of the data proper and producing prediction accuracies on par with the original dataset. Noise addition technique is the ability to maintain good data quality and ensure individual privacy. This approach deals only with the classification task so this approach is addressing the issue of privacy preserving data mining partially. More experiments are to be conducted on security level measurement and data quality.

C. Privacy-Preserving Decision Trees over Vertically Partitioned Data:

Jaideep Vaidya, Chris Clifton, Murat Kantarcioglu and A. Scott Patterson [17] introduce a Privacy-Preserving Decision Trees over Vertically Partitioned Data, generalized privacy-preserving variant of the ID3 algorithm for vertically partitioned data distributed over two or more parties. The algorithm as presented in this paper is working program is a significant step forward in creating usable, distributed, privacy preserving data-mining algorithms. This paper presents a new protocol to construct a decision tree on vertically partitioned data with an arbitrary number of parties where only one party has the class attribute. It presents a general framework for constructing a system in which distributed classification would work. It serves to show that the methods can actually be built and are feasible. This work provides an upper bound on the complexity of building privacy preserving decision trees. Significant work is required to find a tight upper bound on the complexity.

D. Privacy Preserving Decision Tree Mining from Perturbed Data:

In this paper, Li Liu, Murat Kantarcioglu and Bhavani Thuraisingham [18] propose a new perturbation based technique to modify the data mining algorithms. They build a classifier for the original data set from the perturbed training data set. This paper proposed a modified C4.5 [9] decision tree classifier which is suitable for privacy preserving data mining can be used to classify both the original and the perturbed data. This technique considers the splitting point of the attribute as well as the bias of the noise data set as well. It calculates the bias whenever try to find the best attribute, the best split point and partition the training data. This algorithm is based on the perturbation scheme, but skips the steps of reconstructing the original data distribution. The proposed technique has increased the privacy protection with less computation time. Privacy as a security issue in data mining area is still a challenge.

E. Privacy Preserving Decision Tree Learning Using Unrealized Data Sets:

Pui K. Fong and Jens H. Weber-Jahnke [19] 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. Dataset complementation confirms the utility of training data sets for decision tree learning. This approach converts the original sample data sets into a group of unreal data sets. 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 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. In order to mitigate the threat of their inadvertent disclosure or theft, privacy preservation is applied to sanitize the samples prior to their release to third parties. In contrast to other sanitization methods, this technique does not affect the accuracy of data mining results. The decision tree can be built directly from the sanitized data sets, no need to be reconstructing the original dataset. The data set reconstruction algorithm is generic hence privacy preservation via data set complementation fails if all training data sets were leaked. This limitation need to be overcome by applying technique like encryption. 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.

IV. CONCLUSION

In this paper, we present a decision tree classification technique. We have surveyed different approaches used in evaluating the effectiveness of privacy preserving data mining algorithms using decision tree classifier. The work presents in this paper indicates the ever increasing interest of researchers in the area of securing sensitive data and knowledge from malicious users. Many privacy preserving algorithms of decision tree mining are proposed by researchers however, privacy preserving technology needs to be further researched because of the complexity of the privacy problem. We conclude privacy preserving decision tree mining algorithms by analyzing the existing work in the table.1and make some remark. Future research need to be developed to work on these remarks.

Sr. No	Technique	Purpose	Remark
1	Privacy-Preserving Decision Tree Mining Based on Random Substitutions	Author presented a data perturbation technique based on random substitutions & showed that the resulting privacy-preserving decision tree mining method is immune to attacks that are seemingly relevant.	Extra storage required for storing perturbation matrix, do not make strict tradeoffs between preservation of data sample utility and privacy
2	A Noise Addition Scheme in Decision Tree for Privacy Preserving Data Mining	Propose to add specific noise to the numeric attributes after exploring the decision tree of the original data. Proper data is not revealed to the second party during the mining process	It works on numerical data only, need to work on data quality and security level measurement.
3	Privacy-Preserving Decision Trees over Vertically Partitioned Data	Tackled the problem of classification & introduced a generalized privacy preserving variant of the ID3 algorithm for vertically partitioned data distributed over two or more parties	Required to find a tight upper bound on the complexity
4	Privacy Preserving Decision Tree Mining from Perturbed Data	A new perturbation based technique proposed a modified C4.5 decision tree classifier	Need various ways to build classifiers which can be used to classify the perturbed data set.
5	Privacy Preserving Decision Tree Learning Using Unrealized Data Sets	New perturbation and randomization based approach via data set complementation	Requires extra storage for storing perturbed and complement of sample data set and fails if all training data sets were leaked

REFERENCES

[1] R. Agrawal and R. Srikant. Privacy-preserving data mining. In *ACM SIGMOD International Conference on Management of Data*, pages 439–450. ACM, 2000.

[2] 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.

[3] 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

[4] Jiawei Han, Micheline Kamber, “Data Mining: Concepts and Techniques”, Morgan Kaufmann, 2001.

[5] Lior Rokach, Oded Maimon Data Mining and Knowledge Discovery Handbook Second edition pages 167-192, Springer Science + Business Media,2010

[6] L. Breiman, J. Friedman, R. Olshen, and C. Stone. Classification and Regression Trees. Wadsworth Int. Group, 1984.

[7] J.R. Quinlan, Simplifying decision trees, International Journal of Man- Machine Studies, 27, 221-234, 1987.

[8] L. Hyafil and R. L. Rivest, "Constructing Optimal Binary Decision Trees is {NP}-Complete," *Inf. Process. Lett*, vol. 5, pp. 15-17, 1976.

[9] J. R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, 1993.

[10] 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.

[11] 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.

[12] A. 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.

[13] Shipra Agrawal and Jayant R. Haritsa. A framework for high-accuracy privacypreserving mining. *In IEEE International Conference on Data Engineering*, 2005.

[14] 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

[15] 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.

[16] Charu C Agrwal.,Philip S Yu., “ Privacy preserving data mining models and Algorithms.”, *Springer Science+Business media.,LLC.*.2008

[17] J. Vaidya and C. Clifton. Privacy-preserving decision trees over vertically partitioned data. In *Proceedings of the 19th Annual IFIP WG 11.3 Working Conference on Data and Applications Security*, Storrs,Connecticut, 2005. Springer. 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.

[18] 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.