Performance Analysis of SMC Protocols for Decision Tree Classification Rule Mining

Alka Gangrade
Deptt. of M.C.A.
Technocrats Institute of Technology
Bhopal, India
alkagangrade@yahoo.co.in

Ravindra Patel
Deptt. of M.C.A.
U.I.T., R.G.P.V.
Bhopal, India
ravindra@rgtu.net

***Abstract*—This research paper is a generous report on experimental setup, implementation and result analysis of SMC protocols for decision tree classification rule mining. The main focus of the paper is on distributed data mining, where data in horizontally distributed. We are comparing the performance of our two privacy preserving horizontally partitioned decision tree protocols with basic ID3. The result analysis shows that our algorithms execution time is less than the basic ID3 decision tree execution time since in our protocols, all parties individually calculate their information gain as an intermediate result and transfer only these results for further calculations. Accuracy (correctly classified tuple) of test data is almost same because designed decision tree of training data is exactly same. Our protocols are very easy to understand, can be interpreted quickly with minimum efforts, fast and preserved privacy.**

Keywords - Decision tree, Privacy preserving, horizontally partitioned database, SMC, UTP

I. INTRODUCTION

The classification is very important step in data mining for interpretation of useful information. In recent times, there have been growing interests on how to preserve the privacy in data mining when sources of data are distributed across multi-parties. Extractions of useful knowledge from huge amount of data need different techniques and strategies. These techniques are preferred to be faster, more accurate and above all very intelligent. Privacy preserving data mining is one of the most demanding research areas within the data mining community. In many cases, multiple parties may wish to share aggregate private data, without leaking any sensitive information at their end. Cryptographic research on secure distributed computation and their applications to data mining were demonstrated by Pinkas Benny [1].

Classification analysis is an example of supervised learning and is used to partition the data into disjoint groups discriminated by different class labels. In other words, classification analysis develops a model based on a set of training data that is henceforth used to predict the class or category of an unseen example belongs to test data. First, a training set consisting of records whose class labels are known. The training set is used to build a classification model, which is then applied to the test set that consists of records with unknown class labels. Applications are Radar signal classification, character recognition, medical diagnosis, expert systems, and speech recognition etc.

A. Decision Tree Classification

Classification rule mining is the most useful form of data mining. Decision trees are popular and powerful tools for classification and prediction. This may be because they form rules which are easy to understand, or perhaps because they can be converted easily into SQL. While not as "robust" as neural networks and not as statistically "tidy" as discriminate analysis, decision tree often show very good generalization capability. It is an attractive classification method for data mining. It has collection of decision nodes, connected by branches, extending downward from the root node until terminating in leaf nodes. Beginning at the root node, which by convention is placed at the top of the decision tree diagram, attributes are tested at the decision nodes, with each possible outcome resulting in a branch. Each branch then leads either to another decision node or to a terminating leaf node [2, 3]. The basic ID3 algorithm from Quinlan [4] is a greedy algorithm that constructs the decision tree in a top-down recursive divide-and-conquer manner.

B. Secure protocol using Un-trusted Third Party (UTP)

There have been several approaches to support privacy preserving data mining over multiple databases without using third parties [5, 6]. The existence of an Un-trusted Third Party (UTP) enables efficient protocols without revealing private information. The idea of an UTP is that it is willing to perform some computation for the parties in the protocol. It is not trusted with the data or the results. The trust placed in this party is that it does not join with any of the participating parties to violate information privacy and correctly executes the protocol. Correct execution of the protocol is only required to guarantee correct results even a dishonest third party is unable to learn private information in the absence of collusion. In our protocol, UTP is given some information in intermediate result form. We simply mean that the UTP cannot make any sense of the data given to it without the assistance of the local parties involved in the protocol. The UTP performs a computation on the intermediate result, possibly exchanging information with the other parties in the process. And only the final decision tree is revealed to the local parties.

C. Organization of the paper

This paper compares the performance of two privacy preserving horizontally partitioned protocols based on decision tree classification rule mining. The paper is organized as follows: In Section 2, we discuss the related work. Section 3, describes experimental result analysis of our Novel Privacy Preserving Horizontally Partitioned ID3 (NPPHPID3). Section 4, describes experimental result analysis of 2-Layer Privacy Preserving Horizontally Partitioned ID3 (2-Layer PPHPID3). Section 5, we conclude our paper with the discussion of the future work.

II. Related Work

Privacy preserving data mining has been very exciting research area for a decade. A lot of work is going on by the researcher on privacy preserving classification rule mining. The first Secure Multiparty Computation (SMC) problem was described by Yao [7]. SMC allows parties with related background to compute result upon their private data, minimizing the threat of disclosure was explained [8]. An outline of the new and rapidly emerging research area, classify the techniques, review and evaluation of privacy preserving algorithms are presented in [9]. Various tools and how they can be used to solve several privacy preserving data mining problem is explained in [10]. Classification rule mining is one of the most widespread data mining functionality used in real life. General classification techniques have been extensively studied for two decades. Decision tree classification is the best solution approach. Basic ID3 algorithm is a well designed and natural solution of classification rule mining method, first proposed by Quinlan [4]. Lindell and Pinkas proposed ID3 over horizontally partitioned data between two parties using SMC [5]. Data perturbation method is used for data classification is defined in [11]. Privacy preserving ID3 algorithm for vertically partitioned data distributed over two or more parties introduced in [12]. Du and Zhan proposed decision tree algorithm over vertically partitioned data using secure scalar product protocol [6]. Algorithms on building decision tree, however, the tree on each party doesn't contain any information that belong to other party [13]. The drawback of this method is that the resulting class can be distorted by a malicious party. A novel privacy preserving distributed decision tree learning algorithms [14, 15] eliminate the need of third party and proposed a new method for multi-parties. Privacy preserving horizontally partitioned decision tree algorithm using UTP is proposed in [16].

III. Experimental Result Analysis of Novel Privacy Preserving Horizontally Partitioned ID3

A. Introduction

In this section, we focus on performance analysis of privacy preserving decision tree classification in multi-party environment using secure sum protocol where data are horizontally partitioned. We developed new and simple algorithm to classify the horizontally partitioned data. The main advantage of our work over the existing one is that each party cannot gather the other's private data and it is simple and its performance is unmatched by any previous algorithm. Every party separately calculates information gains for each and every attribute then calculates total information gain by using secure sum protocol and finally finds out the maximum information gain. Reference [15] is used for system architecture and details of Algorithms/Protocol. With our algorithms, the execution time required to build the decision tree is reduced compared to existing algorithms and the accuracy of test data set is almost same.

B. Security Analysis

We initially analyzed the security of the primary algorithms, then the security of the complete algorithm. Some of the primary algorithms are executed by the party itself so there is no question of privacy leakage. Master or driving party is used secure sum protocol for calculating total information gain. Protocol secured the information transfer by other parties, thus overall privacy is maintained.

C. Experiment and Results

The experiment was conducted with i3 II generation processor with 2GB RAM having 500GB hard disk. For implementing of algorithms, we use the software Net-Beans IDE (version 6.9). It is an open-source integrated development environment. Net-Beans IDE supports development of all Java applications and integrated these algorithms into Weka version 3.6. Weka is a data mining tool that is used to perform various data mining algorithms. It consists of various classification algorithms and a number of tools to evaluate the data mining algorithm performance. Here we have integrated the API of Weka into java such that by using various functions of the Weka we can develop various applications. It is software through which we can analyze the data on different datasets [17].

TABLE I. Dataset of Student Education

Attribute Name	No. of values	Category
Age	3	<=30, 31..40, >40
Income	3	High, Medium, Low
Technical	3	Best, Better, Good
Student	2	Yes, No
Credit_rating	2	Fair, Excellent
Buy_computer (Class)	2	Yes, No

In this section we cover the details of implementation of our Novel Privacy Preserving Horizontally Partitioned ID3 (NPPHPID3) algorithm on modern hardware and show the experimental results on Student Education datasets shown by Table I. In our experiments, we

showed how collaboration reduces execution time to build the decision tree. We also show that accuracy of our algorithm on test data is almost same with compare to basic ID3 algorithm. Here we use 50% of the dataset for training to build the decision tree and use 50% for measuring the accuracy of classification. Here we are discussing two-party and multi-party cases.

1) For Two-party case: Execution time and accuracy comparison are shown below.

TABLE II. EXECUTION TIME COMPARISON TABLE

Number of Tuples	Size of Dataset	ID3 Execution Time (ms)	NPPHPID3 Two-Party Execution Time (ms)
10500	277 KB	384	167
21000	555KB	709	259
42000	1.08MB	1301	451
84000	2.16MB	2500	923
100000	2.58MB	3180	1194

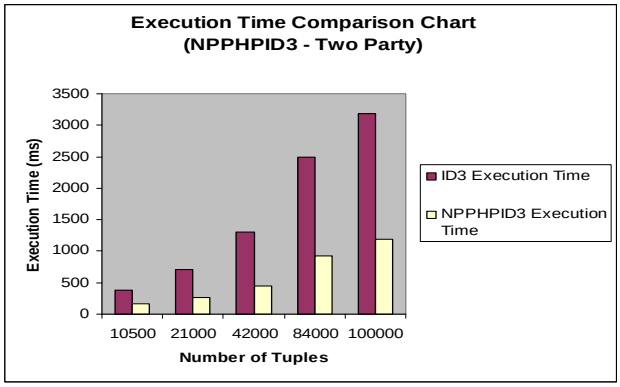

Figure 1. Execution time comparison between ID3 and NPPHPID3.

Table II and Fig. 1 show the comparative analysis of execution time of the existing ID3 based decision tree and the proposed privacy preserving horizontal partitioned based decision tree. It is found that our proposed algorithm takes very much less time to build a tree.

TABLE III. COMPARISON OF ACCURACY OF TEST DATA

Number of Tuples	ID3 Correctly Classified Tuples (%)	NPPHPID3 Correctly Classified Tuples (%)
10500	73.9524 %	73.8476%
21000	73.9524 %	73.9524%
42000	73.9524 %	73.9429%
84000	73.9524 %	73.95%
100000	73.943 %	73.946%

Table III and Fig. 2 show the accuracy comparison for classifying the test data. We find that accuracy of existing ID3 based decision tree and the proposed privacy preserving horizontal partitioned based decision tree is same since decision tree generated by both the algorithm is exactly same.

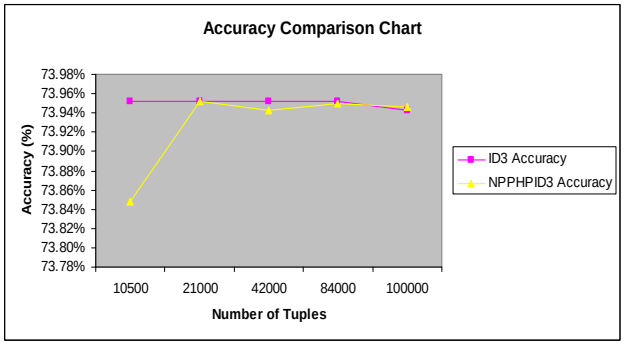

Figure 2. Accuracy comparison between ID3 and NPPHPID3.

2) For Multi-party case: Table IV and Fig. 3 show the comparison of time to build the decision tree while using multi-party i.e. 3 party and 4 party with 2 party. The time to build the decision tree decreases if the number of parties involved increases. Accuracy is almost same.

TABLE IV. EXECUTION TIME COMPARISON USING MULTI-PARTY

Number of Tuples	NPPHPID3 Two Party Execution Time (ms)	NPPHPID3 Three Party Execution Time (ms)	NPPHPID3 Four Party Execution Time (ms)
10500	167	151	139
21000	259	239	221
42000	451	420	395
84000	923	861	813
100000	1194	1070	969

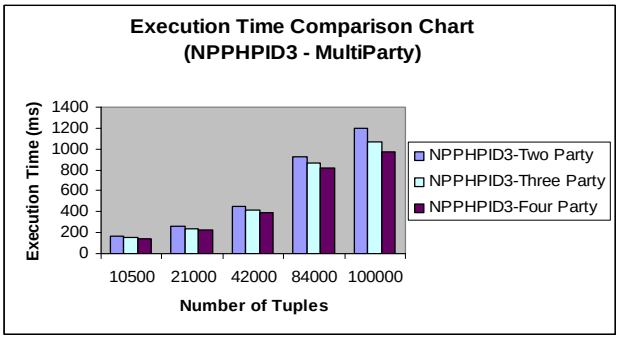

Figure 3. Execution time comparison NPPHPID3(Multi-party).

IV. EXPERIMENTAL RESULT ANALYSIS OF 2-LAYER PRIVACY PRESERVING HORIZONTALLY PARTITIONED ID3

A. Introduction

Our two layer protocol uses an Un-trusted Third Party (UTP). We have already studied how to build privacy preserving two-layer decision tree classifier, where database is horizontally partitioned and communicate their intermediate results to the UTP not their private data. Reference [16] is used for system architecture and details of Algorithms/Protocol. In our protocol, an UTP allows well-designed solutions that meet privacy constraints and achieve acceptable performance.

B. Security Analysis

We initially analyzed the security of the primary algorithms, then the security of the complete algorithm. Some of the primary algorithms are executed by the party itself so there is no question of privacy leakage. Here we are using the concept of Un-trusted Third Party (UTP). All participating parties send their intermediate results to UTP not the actual values, thus there is no privacy leakage. UTP calculates total information gain and find the attribute having highest information gain. Protocol secured the information transfer by other parties, thus overall privacy is maintained.

C. Experiment and Results

The experiments are conducted on the same H/W and S/W system mentioned in previous section (Refer Section 3). Mathematical calculations are almost same. The information gain calculations are done by the parties themselves but the intermediate calculations are done by UTP. First applying our 2-layer privacy preserving horizontally partitioned ID3 (2-Layer PPHPID3) algorithm on Student Education dataset (Refer Table I) on two party and then multi-party case.

1) For Two-party case: Execution time and accuracy comparison are shown below.

TABLE V. EXECUTION TIME COMPARISON TABLE

Number of Tuples	Size of Dataset	ID3 Execution Time (ms)	2-Layer PPHPID3 Two-Party Execution Time (ms)
10500	277 KB	384	156
21000	555KB	709	234
42000	1.08MB	1301	405
84000	2.16MB	2500	827
100000	2.58MB	3180	980

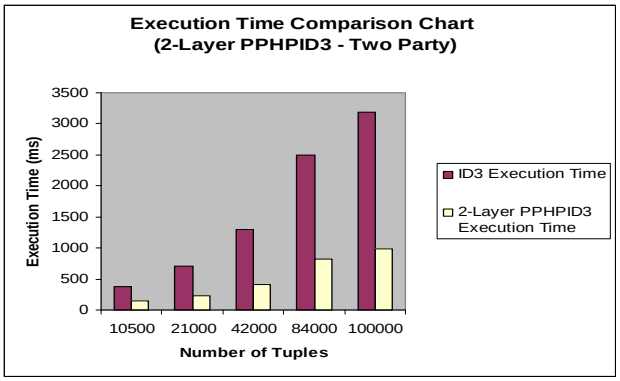

Figure 4. Execution time comparison between ID3 and 2-Layer PPHPID3

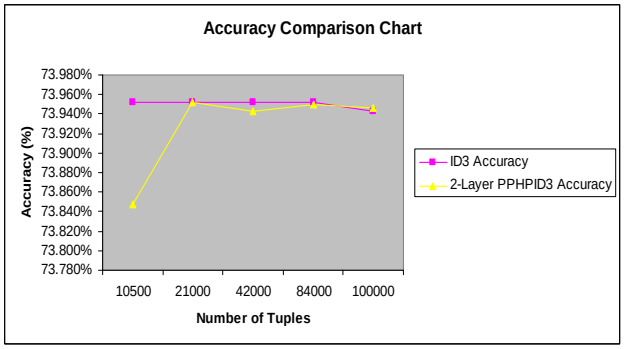

Figure 5. Accuracy comparison between ID3 and 2-Layer PPHPID3.

Above mentioned tables and figures show that execution time of our proposed 2-Layer PPHPID3 on Student Education dataset is much less than the existing ID3 and also less than the NPPHPID3 and accuracy is exactly same as the basic ID3 and NPPHPID3.

2) For Multi-party case: Table VI and Fig. 6 show the comparison of time to build the decision tree while using multiparty i.e. 3 party and 4 party with 2 party. Here we found that execution time decreases if the number of parties increases. Accuracy is almost same.

TABLE VI. EXECUTION TIME COMPARISON USING MULTI-PARTY

Number of Tuples	2-Layer PPHPID3 Two Party Execution Time (ms)	2-Layer PPHPID3 Three Party Execution Time (ms)	2-Layer PPHPID3 Four Party Execution Time (ms)
10500	156	138	129
21000	234	219	208
42000	405	386	371
84000	827	801	782
100000	980	951	929

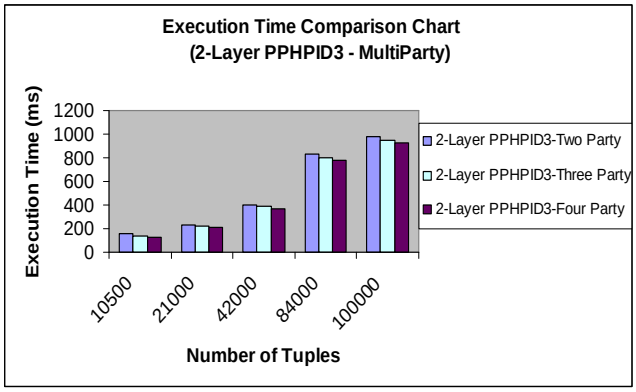

Figure 6. Execution time comparison 2-Layer PPHPID3(Multi-party).

V. CONCLUSION AND FUTURE WORK

In this paper we have presented performance of our decision tree algorithm which uses secure sum protocol to calculate the total information gain of each attribute of all involved parties. The first party drives the protocol and wishes to publish a decision tree while maintaining privacy of all participating parties.

We have also presented performance of our second protocol which uses Un-trusted Third Party (UTP). Here all party transfer their information gain as an intermediate results form to UTP only not the original data and UTP calculates total information gain and find the attribute having highest information gain recursively. Privacy is maintained since party does not know actual data of other party.

According to our experiments, our 2-Layer PPHPID3 algorithm is faster than the NPPHPID3 since in 2-Layer PPHPID3, all party transfer their information gain directly to UTP there is no internal transformation is required.

In our work we have shown how this simple and intuitive concept of privacy can be brought into the world of privacy preserving classification rule mining. However, the simplicity of the model comes at a cost.

In our protocols, all parties are involved for calculation of information gain. However, there can be other such mechanism needs to be addressed which minimize the involvement of parties for distributed classification rule mining.

We have addressed distributed privacy preserving classification rule mining methods where data are distributed horizontally. However, there can be other mechanism needs to be addressed where data are distributed horizontally as well as vertically both i.e. grid partitioned.

One of the major challenges in privacy preserving classification rule mining in these days is the definition of rigorous privacy models that will fit real world privacy needs of real world applications, while maintaining the elegance, simplicity and ease of use that characterize the model.

ACKNOWLEDGMENT

We are thankful to reviewer committee of journal for excellent comments to enhance the work. We are also thankful to the University and the College for their support.

REFERENCES

[1] B. Pinkas. “Cryptographic techniques for privacy-preserving data mining,” ACM SIGKDD Explorations Newsletter, vol. 4, no. 2, pp. 12-19, 2006.

[2] J. Han and Micheline Kamber. “Data Mining: Concepts and Techniques,” Indian Reprint ISBN-81-8147-049-4, Elsevier.

[3] A. K. Pujari. “Data Mining Techniques,” Universities Press(India) 13th Impression 2007.

[4] J. R. Quinlan. “Induction of decision trees,” in: Jude W. Shavlik, Thomas G. Dietterich, (Eds.), Readings in Machine Learning. Morgan Kaufmann, 1990, vol. 1, pp. 81–106.

[5] Y. Lindell and B. Pinkas. “Privacy preserving data mining,” Journal of Cryptology, vol. 15, no. 3, pp. 177–206, 2002.

[6] W. Du and Z. Zhan. “Building decision tree classifier on private data,” In CRPITS, 2002, pp. 1–8.

[7] A. C. Yao. “Protocols for secure computation,” In Proceeding of 23rd IEEE Symposium on Foundations of Computer Science (FOCS), 1982, pp. 160-164.

[8] W. Du and M. J. Attalah. “Secure multi-problem computation problems and their applications: A review and open problems,” Tech. Report CERIAS Tech Report 2001-51, Center for Education and Research in Information Assurance and Security and Department of Computer Sciences, Purdue University, West Lafayette, IN 47906, 2001.

[9] V. Verykios and E. Bertino. “State-of-the-art in Privacy preserving Data Mining,” SIGMOD, vol. 33, no. 1, 2004.

[10] C. Cliffton, M. Kantarcioglu and J. Vaidya. “Tools for privacy preserving distributed data mining,” ACM SIGKDD Explorations Newsletter, vol. 4, no. 2, pp. 28-34, 2004.

[11] R. Agrawal and R. Srikant. “Privacy preserving data mining,” In proceeding of the ACM SIGMOD on Management of data, Dallas, TX USA, May 15-18, 2000, pp. 439-450.

[12] J. Vaidya, C. Clifton, M. Kantarcioglu and A. S. Patterson. “Privacy-preserving decision trees over vertically partitioned data,” In the Proceeding of the 19th Annual IFIP WG 11.3 Working Conference on Data and Applications Security, 2008, pp. 139–152.

[13] A. Shamir. “How to share a secret,” Communications of the ACM 1979, vol. 22, no. 11, pp. 612-613, 1979.

[14] F. Emekci , O. D. Sahin, D. Agrawal and A. El Abbadi. “Privacy preserving decision tree learning over multiple parties,” Data and Knowledge Engineering 63, pp. 348-361, 2007.

[15] A. Gangrade and R. Patel. “A novel protocol for privacy preserving decision tree over horizontally partitioned data,” International Journal of Advanced Research in Computer Science, vol. 2, No. 1, pp. 305-309, Jan–Feb, 2011.

[16] A. Gangrade and R. Patel. “Privacy Preserving Two-Layer Decision Tree Classifier for Multiparty Databases”, International Journal of Computer and Information Technology (IJCIT), ISSN No. 2277-0764, Vol. 1, No. 1, pp. 77-82, September 2012.

[17] I. H. Witten and E. Frank. Data Mining: Practical Machine Learning Tools and Techniques, second ed., Morgan Kaufmann, San Francisco, 2005.