International Journal of Computer and Information Technology (2277 – 0764)
Volume 01– Issue 01, September 2012

Privacy Preserving Two-Layer Decision Tree Classifier for Multiparty Databases

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

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

Abstract— **Privacy protection is one of the important problems in data mining. The growth of the Internet has triggered incredible opportunities for cooperative computation, where people are jointly conducting computation tasks based on the private inputs they each supplies. These computations could occur between mutually un-trusted parties or even between competitors. Today, to conduct such computations, one entity must usually know the inputs from all the participants, however if nobody can be trusted enough to know all the inputs, privacy will become a primary concern. Our two layer protocol uses an Un-trusted Third Party (UTP). We study 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. In our protocol, an UTP allows well-designed solutions that meet privacy constraints and achieve acceptable performance.**

Keywords- Privacy preserving; Un-trusted Third Party; decision tree.

I. INTRODUCTION

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 [1]. For example, different superstores with sensitive sales data may wish to coordinate among themselves in knowing aggregate trends without leaking the trends of their individual stores. This requires secure protocols for sharing the information across the different parties. The data may be distributed in two ways across different sites: Horizontal partition and Vertical partition. Horizontal partition means, where different sites have different sets of records containing the same attributes. Vertical partition means, where different sites have different attributes of the same sets of records [2].

In this paper, we particularly focus on applying privacy preserving data mining method on the decision tree over horizontally partitioned data using UTP.

A. Decision Tree Classification

Classification is often seen as the most useful form of data mining. Decision trees are the most useful, 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.

Decision trees are built by choosing an attribute and a value for that attribute which splits the dataset. The attribute and value are chosen to minimize diversity of class label in the two resulting sets (an alternative way of looking at this is to maximize information gain or to minimize entropy). The first split is unlikely to be perfect, so we recursively split the sets created until all the sets we have consist of only one class. Creating the decision tree is simply a matter of collating the splits in the correct order. The trick in data mining (where we may be dealing with large datasets; possibly even too big to fit into memory) is to find that attribute and value with the minimum number of passes through the database.

B. Secure Protocols with an Un-trusted Third Party

A straightforward solution for privacy preserving data mining is to use a trusted third party to gather data from all data sources and then send back results after running the desired data mining algorithms. However, the level of trust is not acceptable in this scheme since the privacy of the data sources cannot be protected from the third party. There have been several approaches to support privacy preserving data mining over multiple data bases without using third parties [3, 4]. 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. Typically the third party is given some information in intermediate result form. We simply mean that the third party cannot make any sense of the data given to it without the assistance of the local parties involved in the protocol.

The third party 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. Our Contributions

Our main contributions in this paper are as follows:

- We present a novel two-layer privacy preserving decision tree classifier for real world data set.
- It proposes a new protocol to construct a decision tree on horizontally partitioned data in distributed manner.
- We carry out an extensive study of our protocol. It uses UTP.

D. Organization of the paper

The paper is organized as follows: In Section 2, we discuss the related work. Section 3, describes our two layer privacy preserving classification model. Section 3.1 describes architecture of our model. Section 3.2 and 3.3 describe informal algorithm and formal algorithms of our proposed work respectively. In Section 4, we present our experimental results that are conducted by using our privacy preserving two-layer decision tree classifier on real-world data sets. In Section 5, we conclude our paper with the discussion of the future work.

II. Related Work

The first Secure Multiparty Computation (SMC) problem was described by Yao [5]. SMC allows parties with similar background to compute result upon their private data, minimizing the threat of disclosure was explained [6].

Privacy preserving data mining has been an active research area for a decade. A lot of work is going on by the researcher on privacy preserving classification in distributed data mining.

An overview of the new and rapidly emerging research area of privacy preserving data mining, also classify the techniques, review and evaluation of privacy preserving algorithms presented in [7]. Various tools discussed and how they can be used to solve several privacy preserving data mining problem [8]. Cryptographic research on secure distributed computation and their applications to data mining were demonstrated by Pinkas Benny [1].

Classification is one of the most widespread data mining problems come across in real life. General classification techniques have been extensively studied for over twenty years. Decision tree classification is the best solution approach. Algorithm ID3 particularly a well designed and natural solution, first proposed by Quinlan [9]. Lindell and Pinkas proposed a secure algorithm to build a decision tree using ID3 over horizontally partitioned data between two parties using SMC [3]. Data perturbation method used to solve the problem that Alice is allowed to conduct data mining operation on private database of Bob, how Bob prevents Alice from accessing private information in his database while Alice is still able to conduct the data mining operation defined in [10]. A generalized privacy preserving variant of the ID3 algorithm for vertically partitioned data distributed over two or more parties introduced in [11]. A decision tree algorithm over vertically partitioned data using secure scalar product protocol proposed in [4].

A novel privacy preserving distributed decision tree learning algorithm [12], that is based on Shamir [13] and the ID3 algorithm is scalable in terms of computation and communication cost, and therefore it can be run even when there is a large number of parties involved and eliminate the need for third party and propose a new method without using third parties.

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 altered by a malicious party. Privacy preserving decision tree algorithm over vertically partitioned data, which is based on idea of passing control from site to site proposed by Weiwei Fang and Yang [15]. The main purpose of data classification is to build a model (i.e., classifier) to predict the (categorical) class labels of records based on a training data set where the class label of each record is given. The classifier is usually represented by classification rules, decision trees, neural networks, or mathematical formulae that can be used for classification.

III. Proposed Work

In this paper, we address the issue related to privacy preserving data mining in a distributed manner. In particular, we focus on privacy preserving two-layer decision tree classifier on horizontally partitioned data. The objective of privacy preserving data classification is to build accurate classifiers without disclosing private information in the data being mined. The performance of privacy preserving techniques should be analyzed and compared in terms of both the privacy protection of individual data and the predictive accuracy of the constructed classifiers.

A. Architecture

1) Input Layer : Input layer comprises of all the parties that are involved in the computation process [16]. All participating party individually calculate the Information Gain of each attribute and send Information Gain as an intermediate result form to the UTP. This process is done at every stage of decision tree.

2) Output Layer : The UTP exists at the 2nd layer i.e. the computation layer of our protocol. UTP collects only intermediate results i.e. Information Gain of all attributes from all parties not data and calculate the total information gain of each attribute. Then find the attribute with highest information gain and then create the root of decision tree with this attribute and send this attribute to all parties for further calculation. This process is also done at every stage of decision tree.

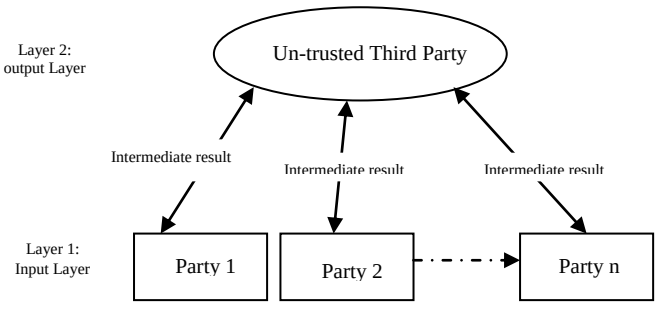

Figure 1. Two-layer Architecture.

B. Formula for Calculating Information Gain

Information gain calculation from Han and Kamber [17] and Pujari [18].

The next step is to compute the best attribute with the maximum information gain. The information gain when an attribute A is used to partition the data set S is:

$$Gain(S, A) = Entropy(S) - \sum ((|S_v|/|S|) * Entropy(S_v)) \quad (1)$$

where vϵA

The entropy of a dataset S is given by:

$$Entropy(S) = - \sum (N_j/N) * \log (N_j/N) \quad (2)$$

where j = 1 to p

Where N_j is the number of transactions having class c_j in S and N is the number of transactions in S. As we see, this again becomes a problem of counting transactions: the number of transactions that reach the node N, the number in each class N_j, and the same two after partitioning with each possible attribute value vϵA. Algorithm 1 shows our two-layer privacy preserving horizontally partitioned ID3 algorithm. Algorithm 2 calculates the information gain; Algorithm 3 calculates the total information gain and Algorithm 4 computes the maximum information gain.

The best attribute A is the one that has the maximum gain, i.e., minimum entropy, among all considered attributes. Once the best attribute has been determined, execution proceeds and creates an interior node for the split, and then recursively do till no attribute left.

C. Informal Algorithm

1) Input Layer :

- Party individually calculates Expected Information of every attribute.
- Party individually calculates Entropy of every attribute.
- Party individually calculates Information Gain of each attribute(InformationGain( )).

2) Output Layer :

- All party send Information Gain of each attribute to the UTP
- UTP compute the sum of Information Gain of all parties of all attributes (TotalInformationGain( )) [19].
- UTP find out the attribute with the largest Information Gain by using MaxInformationGain( ).
- Create the root with largest Information Gain attribute and edges with their values, then send this attribute to all parties at Input Layer for further development of decision tree.

Recursively do when no attribute is left.

D. Assumptions

The following assumptions have been set:

- UTP computes the final result from the intermediate results provided by all parties at every stage of decision tree.
- UTP computes attribute with highest information gain and send to all party at every stage of decision tree.
- UTP has the ability to announce the final result of the computation publicly.
- Each party is not communicating their input data to other party.
- The communication networks used by the input parties to communicate with the UTP are secure.

E. Formal Algorithms

1) Algorithm 1: TLPPHPID3() – Two Layer Privacy Preserving Horizontally Partitioned ID3.

a) Input Layer

- Define P_1, P_2, …., P_n Parties.(Horizontally Partitioned).
- Each Party contains R set of attributes A_1, A_2, …., A_R.
- C the class attributes contains c class values C_1, C_2, …., C_c.
- For party P_i where i = 1 to n do
- If R is Empty Then
- Return a leaf node with class value
- ElseIf all transaction in $T(P_i)$ have the same class Then
- Return a leaf node with the class value
- Else
- Calculate Expected Information classify the given sample for each party P_i individually.

- Calculate Entropy for each attribute (A_1, A_2, …., A_R) of each party P_i.
- Calculate Information Gain for each attribute (A_1, A_2,…., A_R) of each party P_i
- Send Information Gain to UTP
- End If
- End If
- End For

b) Output Layer – Computation is done by UTP

- Calculate Total Information Gain for each attribute of all parties (TotalInformationGain( )).
- $A_{BestAttribute}$ ← MaxInformationGain( )
- Let V_1, V_2, …., V_m be the value of attributes. $A_{BestAttribute}$ partitioned P_1, P_2,…., P_n parties into m parties
- $P_1(V_1)$, $P_1(V_2)$, …., $P_1(V_m)$
- $P_2(V_1)$, $P_2(V_2)$, …., $P_2(V_m)$
- . . .
- . . .
- $P_n(V_1)$, $P_n(V_2)$, …., $P_n(V_m)$
- Return the Tree whose Root is labelled $A_{BestAttribute}$ and has m edges labelled V_1, V_2, …., V_m. Such that for every i the edge Vi goes to the Tree
- TLPPHPID3(R – $A_{BestAttribute}$, C, ($P_1(V_i)$, $P_2(V_i)$, …., $P_n(V_i)$))
- End.

2) Algorithm 2 : InformationGain() - To calculate Information Gain for attribute A.

- T ← Total number of transactions at this node
- InfoGain ← Entropy(T)
- for each attribute value a_i do
- T_{ai} ← Total number of transaction for attribute value a_i
- InfoGain ← InfoGain - $|T_{ai}| / |T| * Entropy(T_{ai})$
- end for
- return InfoGain

3) Algorithm 3 : TotalInformationGain( ) - To compute the Total Information Gain [19] for every attribute.

- TotalInfoGain = 0
- For i = 1 to n do {Parties P_1, P_2 ,…., P_n }
- TotalInfoGain=TotalInfoGain+InformationGain(A_i)
- End For
- Return (TotalInfoGain)
- End For

4) Algorithm 4 : MaxInformationGain( ) – To find out the attribute with highest Information Gain [19] for horizontally partitioned data.

- MaxInfoGain = -1
- For j = 1 to R do {Attribute A_1, A_2,…., A_R }
- Gain = TotalInformationGain(A_j)
- If MaxInfoGain < Gain then
- MaxInfoGain = Gain
- $A_{BestAttribute} = A_j$
- End If
- End For
- Return ($A_{BestAttribute}$)

IV. EXPERIMENTAL RESULTS

For the real word data experiment results, we have generated 400 records, and randomly choose 200 records for training sample, and remaining 200 records for testing purpose. We used WEKA [20] data mining software to run basic ID3 decision tree and two-layer privacy preserving horizontally partitioned ID3 decision tree classifiers on our reconstructed data, and reported the experiment results on the test data. Experiment result shows total time taken to generate the decision tree by two parties with five attributes. Class attribute has two values. Number of parties as well as the number of attributes could be extended. We have compared the performance of these two classifiers on various scenarios like execution time, mean absolute error, relative absolute error. Based on our experimental results, using this real world data set, the performance of privacy preserving two-layer horizontally partitioned ID3 decision tree classifier is better than the basic ID3 decision tree classifiers.

A. Experimental Results of number of instances vs execution time : We used two different decision tree classifiers on different number of instances. "Fig. 2" shows the comparison between them. We find that privacy preserving two-layer decision tree horizontally partitioned ID3 classifier is faster than basic ID3 decision tree classifier.

TABLE I. EXECUTION TIME CALCULATION

Number of Instances	ID3 Execution Time(ms)	2-Layer PPHP ID3 Execution Time(ms)
14	78	15
25	93	15
50	110	16
100	125	31
200	150	32

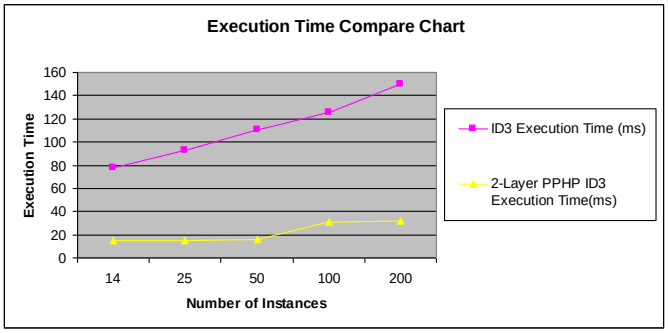

Figure 2. Number of Instances vs Execution time.

B. Experimental Results of number of instances vs Mean absolute error : "Fig. 3" shows the comparison between two classifiers. We find that mean absolute error is less in privacy preserving two-layer horizontally partitioned ID3 decision tree classifier.

TABLE II. MEAN ABSOLUTE ERROR CALCULATION

Number of Instances	ID3 Mean Absolute Error	2-Layer PPHP ID3 Mean Absolute Error
14	0.2857	N.P
25	0.24	0.237
50	0.24	0.24
100	0.23	0.22
200	0.235	0.23

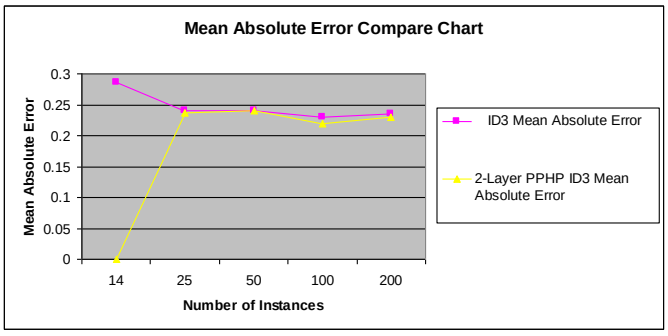

Figure 3. Number of Instances vs Mean absolute error.

C. Experimental Results of number of instances vs Relative absolute error : "Fig. 4" shows the comparison between two classifier's relative absolute error in percentage. We find that relative absolute error is less in privacy preserving horizontally partitioned ID3 decision tree classifier.

TABLE III. RELATIVE ABSOLUTE ERROR CALCULATION

Number of Instances	ID3 Relative Absolute Error	2-Layer PPHP ID3 Relative Absolute Error
14	60%	N.P
25	63.22%	60.13%
50	64.53%	63.24%
100	64.28%	63.63%
200	65.06%	64.28%

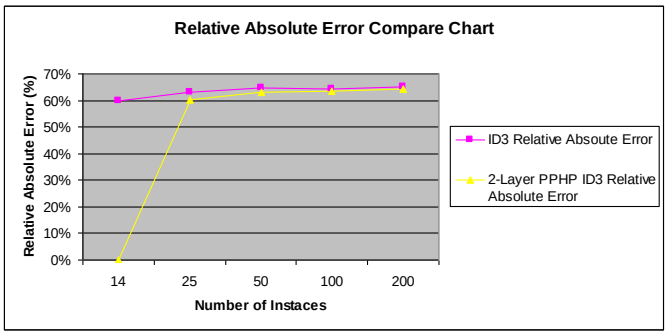

Figure 4. Number of Instances vs Relative absolute error.

V. CONCLUSION AND FUTURE WORK

We believe that it is feasible to build a privacy preserving decision tree classifier with SMC techniques. In this paper, we proposed a new classifier using two-layer architecture that enables SMC by hiding the identity of the parties taking part in the classification process using UTP. Further we may describe that intermediate result is calculated by every party individually and send only intermediate result to UTP not the input data. Through the communication between UTP and all party final result is carried out. It requires less memory space. Also provides fast and easy calculations. Using this protocol, classification will almost secure and privacy of individual will be maintained. Further development of the protocol is expected in the sense that for joining multi-party attributes using a trusted third party can be used. We are continuing work in this field to develop new classifier for building privacy preserving decision tree classifier using grid partitioned data and to analysis new as well as existing classifiers.

As part of future work, we are actually implementing the entire protocol in JAVA on huge databases, which should be the first working code in the area of privacy preserving decision tree classifier on horizontally partitioned data using un-trusted third party.

ACKNOWLEDGMENT

We are thankful to the University and the College for their support. We thank my colleagues for their technical support and the referees for their constructive suggestions.

REFERENCES

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

[2] Charu C. Aggarwal, Philip S. Yu., "Privacy-Preserving Data Mining:Models and Algorithms", Kluwer Academic Publishers Boston/Dordrecht/London.

[3] Yehuda Lindell, Benny Pinkas, "Privacy preserving data mining," Journal of Cryptology vol. 15, no. 3, 2002, pp. 177–206.

[4] Wenliang Du, Zhijun Zhan, "Building decision tree classifier on private data," In CRPITS, 2002, pp. 1–8.

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

[6] Wenliang Du, Mikhail 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.

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

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

[9] 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.

[10] R. Agrawal, 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.

[11] J. Vaidya, C. Clifton, M. Kantarcioglu, 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.

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

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

[14] J. Shrikant Vaidya, "Privacy preserving data mining over vertically partitioned data," Ph. D. Thesis of Purdue University, August 2004, pp. 28-34.

[15] Weiwei Fang, Bingru Yang, "Privacy Preserving Decision Tree Learning Over Vertically Partitioned Data," In Proceeding of the 2008 International Conference on Computer Science & Software Engineering.

[16] Alka Gangrade, Ravindra Patel, "A novel protocol for privacy preserving decision tree over horizontally partitioned data," International Journal of Advanced Research in Computer Science, 2 (1), Jan–Feb, 2011, 305-309.

[17] Jiawei Han, Micheline Kamber, "Data Mining: Concepts and Techniques," Indian Reprint ISBN-81-8147-049-4, Elsevier.

[18] Arun K Pujari, "Data Mining Techniques," Universities Press(India) 13Th Impression 2007.

[19] Jaideep Vaidya, Chris Clifton, "Leveraging the "Multi" in secure multi-party computation".

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