ISSN No. 0976-5697

Volume 2, No. 1, Jan-Feb 2011

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

A Novel Protocol For Privacy Preserving Decision Tree Over Horizontally Partitioned Data

Alka Gangrade*
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: 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. In this paper, we focus on the privacy preserving decision tree classification in multi-party environment when data are horizontally partitioned. We develop new and simple algorithm to classify the horizontally partitioned multi-party 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. With our algorithms, the computation cost and communication cost during tree building stage is reduced compared to existing algorithms.

Keywords: privacy preserving, secure multi-party computation, decision tree.

I. INTRODUCTION

Data mining, the extraction of hidden predictive information from huge databases, is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. The economic value of data mining is today well established. Most large organizations regularly practice on data mining techniques. One of the most popular techniques is classification rule mining. Data mining tools predict future trends and behaviors.

Classification is a very important problem in data mining. Decision tree is one of the most well known approaches for classification. Decision tree is a series of nested if/then/else rules. One of the most important features of decision tree classifier is their ability to break down a complex decision making process into a set of simpler decisions, thus providing a solution which is often easier to understand. Applications - Radar signal classification, character recognition, remote sensing, medical diagnosis, expert systems, and speech recognition etc.

There are numerous methods for finding the feature that best divides the training data such as information gain and gini index.

Nowadays people are more concerned with privacy protection while performing data mining task. In recent years privacy preserving data mining has emerged as a very active research area in data mining. With the rapid growth of the amount of information, two or more organizations often need to work together on data mining tasks nowadays. This leads to a lot of privacy concerns. The parties usually want to keep their own information private while still allowing certain data mining task performed on it. A lot of research has been done on privacy preserving data mining.

The objective of privacy preserving data classification is to build accurate classifiers without disclosing private information in the data being mined.

In this paper we address the issue of secure multi-party computation for classification rule mining. Specifically, we wish to run a classification algorithm using SMC protocol without revealing any original information. In this paper, we present a novel protocol based on secure multiparty computation for privacy preserving ID3 over horizontally partitioned data.

II. RELATED WORK

The first secure multi-party computation problem was described by Yao [1]. Secure Multi-party Computation (SMC) allows parties with similar background to compute result upon their private data, minimizing the threat of disclosure was explained [2].

A lot of work is going on by the researcher on similar problem related with 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 [3]. Various tools discussed and how they can be used to solve several privacy preserving data mining problem [4]. Cryptographic research on secure distributed computation and their applications to data mining demonstrated in [5].

Classification is one of the most widespread data mining problems come across in real life. Decision tree classification is the best solution approach. ID3 algorithm, particularly a well designed and natural solution, first proposed by Quinlan [6]. Proposed secure algorithm to build a decision tree using ID3 over horizontally partitioned data between two parties using secure multi-party computation [7]. 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 [8]. A generalized privacy preserving variant of the ID3 algorithm for vertically partitioned data distributed over two or more parties introduced in [9]. A decision tree algorithm over vertically partitioned data using secure scalar product protocol proposed in [10].

A novel privacy preserving distributed decision tree learning algorithm [11], that is based on Shamir [12] 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 classification algorithm over vertically partitioned data, which is based on idea of privacy preserving decision tree and passing control from site to site proposed by Weiwei Fang and Yang [14].

III. PRELIMINARIES

We start this section with a subsection summarizing the ID3 algorithm. Then we continue with a subsection describing of horizontally and vertically partitioned data. We present our algorithm in section 4 and experimental results in section 5. We discuss security of algorithm in section 6 and conclude in section 7.

A. The ID3 Algorithm

The basic algorithm from Quinlan [6] and Han and Kamber [15] for decision tree induction is a greedy algorithm that constructs decision trees in a top-down recursive divide-and-conquer manner. ID3 is a very popular decision tree building method in data mining. The ID3 algorithm summarized as follows.

[a] Algorithm 1 : The ID3 Algorithm

Require: R, a set of attributes.
Require: C, the class attribute.
Require: S, data set of tuples.

[i] if R is empty then
[ii] Return the leaf having the most frequent value in data set S.
[iii] else if all tuples in S have the same class value then
[iv] Return a leaf with the specific class value.
[v] else
[vi] Determine attribute A with the highest information gain in S.
[vii] Partition S in m parts $S(a_1),...,S(a_m)$ such that $a_1,...,a_m$ are the different values of A.
[viii] Return a tree with root A and m branches label as $a_1 ... a_m$, such that branch i contains $ID3(R-\{A\},C,S(a_i))$.
[ix] end if
[x] end.

B. Horizontally And Vertically Partitioned Data

The method of privacy preserving data mining depends on the data mining task i.e. association rule, classification, clustering, etc. The data sources distribution manner may be the following:

Centralize, where all transactions are stored in only one party;

Horizontally, where every involving party has only a subset of transaction records, but every record contains all attributes. "Fig. 1" shows horizontally partitioned data.

Vertically, where every involving party has the same numbers of transaction records, but every record contains partial attributes. "Fig. 2" shows vertically partitioned data.

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

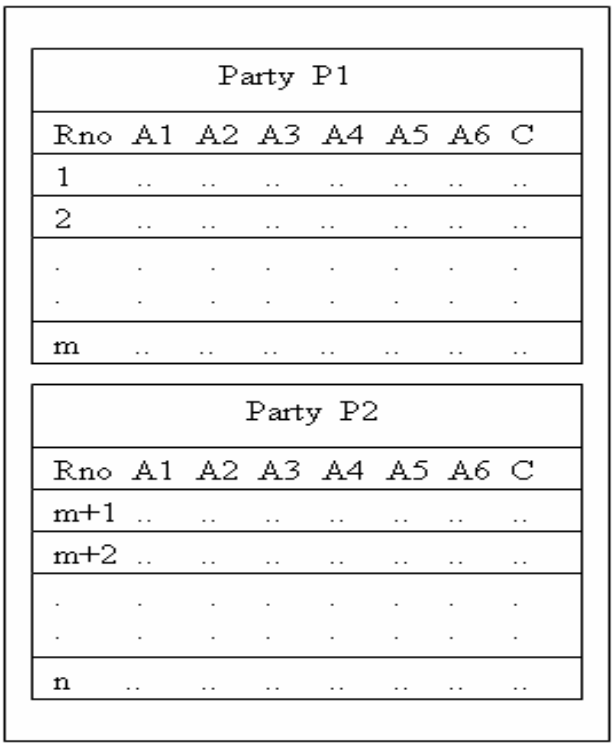

Figure 1. Horizontally Partitioned Data

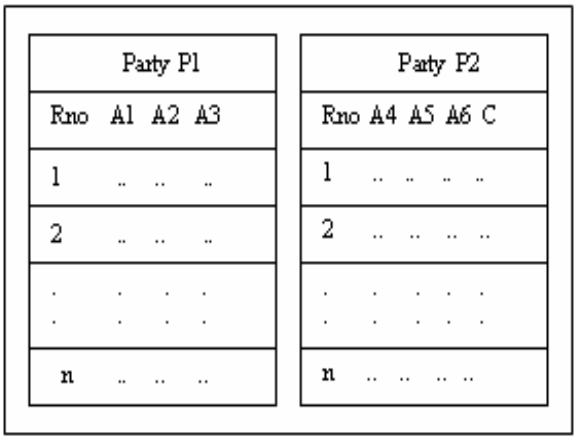

Figure 2. Vertically Partitioned Data

IV. PROPOSED WORK

A. Privacy Preserving Decision Tree Over Horizontally Partitioned Data

[a] Systematic Approach

Our protocol is based on secure multi-party computation for privacy preserving ID3 over horizontally partitioned data. Every party separately calculates information gains for each and every attribute then calculates total information gain by using secure sum protocol and finally find out the maximum information gain. "Fig. 3" shows the proposed approach and "Fig. 4" explains the protocol.

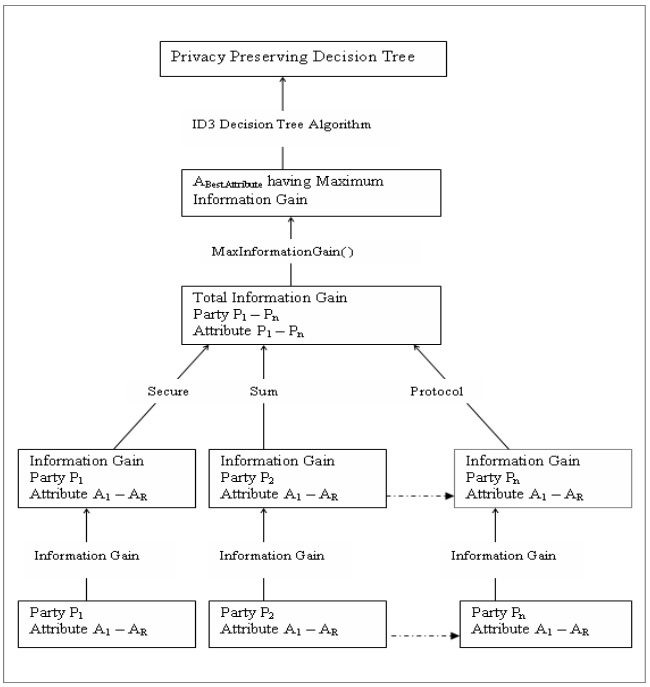

Figure 3. Proposed Approach for Privacy Preserving Decision Tree Over Horizontally Partitioned Data

B. Informal Algorithm

[a] To compute Expected Information classify the given sample for each party individually.
[b] To compute Entropy of individual attribute of all parties.
[c] To compute Information Gain for each attribute of each party.
Calculation of information gain from Han and Kamber [15] and Pujari [16]:
[d] Assume there are two classes, *P* and *N*
[e] Let the set of examples *S* contain *p* elements of class *P* and *n* elements of class *N*
[f] The amount of information, needed to decide if an arbitrary example in *S* belongs to *P* or *N* is defined as

$$I(p,n) = -\frac{p}{p+n}\log_2\frac{p}{p+n} - \frac{n}{p+n}\log_2\frac{n}{p+n}$$

[g] Assume that using attribute A set *S* will be partitioned into sets $\{S_1, S_2, \ldots, S_v\}$
[h] If S_i contains p_i examples of *P* and n_i examples of *N*, the entropy, or the expected information needed to classify objects in all subtrees S_i is

$$E(A) = \sum_{i=1}^{v}\frac{p_i + n_i}{p+n} I(p_i, n_i)$$

[i] The encoding information that would be gained by branching on *A*

$$\text{GAIN}(A) = I(p, n) - E(A)$$

[j] To compute the sum of Information Gain of all parties of all attributes by using Secure Sum Protocol (TotalInformationGain( )).
[k] To compute the attribute with the largest Information Gain by using MaxInformationGain( )
[l] Create the root with largest Information Gain attribute and edges with their values.
[m] Recursively do when no attribute is left.

C. Formal Algorithm

[a] Algorithm 2 : Novel Privacy Preserving ID3 (NPPID3( ))

a) Define P_1, P_2, …., P_n Parties.(Horizontally partitioned).
b) Each Party contains R set of attributes A_1, A_2, …., A_R.
c) C the class attributes contains c class values C_1, C_2, …., C_c.
d) For party P_i where i = 1 to n do
e) If R is Empty Then
f) Return a leaf node with class value
g) Else If all transaction in $T(P_i)$ have the same class Then
h) Return a leaf node with the class value
i) Else
j) Calculate Expected Information classify the given sample for each party P_i individually.
k) Calculate Entropy for each attribute (A_1, A_2, …., A_R) of each party P_i.
l) Calculate Information Gain for each attribute (A_1, $A_2 \ldots A_R$) of each party P_i
m) End If.
n) End For
o) Calculate Total Information Gain for each attribute of all parties by using Secure Sum Protocol (TotalInformationGain( )).
p) $A_{BestAttribute} \leftarrow$ MaxInformationGain( )
q) Let V_1, V_2, …., V_m be the value of attributes. $A_{BestAttribute}$ partitioned $P_1, P_2, \ldots, P_n$ parties into m parties
r) $P_1(V_1), P_1(V_2), \ldots, P_1(V_m)$
s) $P_2(V_1), P_2(V_2), \ldots, P_2(V_m)$
t) . .
u) . .
v) $P_n(V_1), P_n(V_2), \ldots, P_n(V_m)$
w) 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
x) NPPID3(R – $A_{BestAttribute}$, C, ($P_1(V_i)$, $P_2(V_i)$, …., $P_n(V_i)$))
y) End.

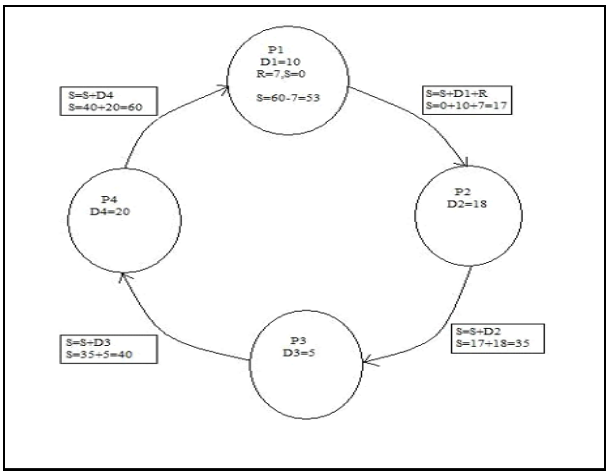

Figure 4. Secure Sum Protocol

[b] Algorithm 3 : TotalInformationGain( ) - To compute the Total Information Gain for every attribute by using Secure Sum Protocol "Fig. 4" explains the protocol.

a) For j = 1 to R do {Attribute A_1, A_2,…., A_R }
b) Total_Info_Gain(A_j) = 0
c) Party P_1 choose Random Number r
d) For i = 1 to n do {Parties P_1, P_2,…., P_n }
e) Total_Info_Gain(A_j) = Total_Info_Gain(A_j) + Info_Gain(A_{ij})
f) End For
g) Total_Info_Gain(A_j) = Total_Info_Gain(A_j) – r
h) End For
i) End.

[c] Algorithm 4 : MaxInformationGain( ) – To compute the highest Information Gain for horizontally partitioned data.

a) MaxInfoGain = -1
b) For j = 1 to R do {Attribute A_1, A_2,…., A_R }
c) Gain = TotalInformationGain(A_j)
d) If MaxInfoGain < Gain then
e) MaxInfoGain = Gain
f) $A_{BestAttribute} = A_j$
g) End If
h) Return ($A_{BestAttribute}$)
i) End For
j) End.

V. COMPUTATION AND COMMUNICATION ANALYSIS

Our novel privacy preserving decision tree classification algorithm over horizontally partitioned data is memory efficient because major calculation is done individually by every party in a distributed manner. We have introduced a new method to find out the best attribute without encryption. That is why it is fast and privacy is preserved in all respect. The "Fig. 5" shows the chart for total time taken to generate the decision tree by two parties with six attribute. Number of parties as well as the number of attributes could be extended.

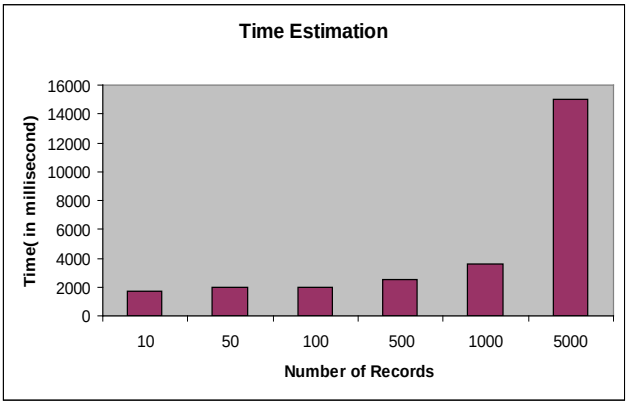

Figure 5. No. of Record vs Time Chart

VI. SECURITY DISCUSSION

Our privacy preserving decision tree classification system contains several components. We explain how to correctly compute Entropy, Information Gain. In our protocol, encryption is not required because every party calculate only for their own data. Input data is not communicated, only the intermediate result is communicated via protocol. We show how to calculate the Information Gain for each attribute for every party securely. Then calculate the total Information Gain for each party using Secure Sum Protocol. We then describe how to obtain the attribute with the highest Information Gain. We maintain the privacy and the correctness of the computation at every stage of the algorithm and it is also guaranteed.

VII. CONCLUSIONS 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 protocol that enables SMC by hiding the identity of the parties taking part in the classification process. Further we may describe that every party individually compute their result before final computation, making it nearly fast and easy. 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 and an untrusted third party can be used. We are continuing work in this field to develop new classifier for building privacy preserving decision tree using grid partitioned data and to analysis new as well as existing classifier.

VIII. ACKNOWLEDGMENT

We are also 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.

IX. REFERENCES

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

[2] 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.

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

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

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

[6] 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.

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

[8] R. Agrawal, R. Srikant "Privacy Preserving Data mining," In proc. of the ACM SIGMOD on Management of data, Dallas, TX USA, May 15-18, 2000, pp. 439-450.

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

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

[11] 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.

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

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

[14] Weiwei Fang, Bingru Yang, "Privacy Preserving Decision Tree Learning Over Vertically Partitioned Data," In Proc. of the 2008 Int. Conf. on Computer Science & Software Engineering.

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

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