Multiparty Communication Using ID3 and Randomized Response Techniques for Preserving Privacy

Mr. Rajdeep Brar
M. Tech. *(Student)
Department of Computer Science & Engineering
Marudhar Engineering College, Bikaner (Raj.), India.
Email: rajdeep33@gmail.com

Mrs. Monika Soni
Asst. Prof. Department of Computer Science & Engineering,
Marudhar Engineering College, Bikaner (Raj.)
Email: 12.monika@gmail.com

Abstract— Data mining is the process of analyzing data from different perspectives and summarizing it into useful information used to feedback, increase revenue, cuts costs, or all. A number of freeware and shareware data mining software resources are available for analyzing data. Data Mining allows users or organizations to analyze the extracted data from many different dimensions or angles, systematically categorize it, and summarize the relationships identified. For multiparty collaboration the data's privacy is very important: all the parties of the collaboration promise to provide their private data to the collaboration, but for privacy concern of the private data each party wants to preserve their private data. But the issue that accompany with the huge collection or repository of data is confidentiality. Performance of privacy preserving collaborative data using secure multiparty computation is evaluated with ID3 algorithm and as a result the privacy of preserved data will is increased up to 65%.

Keywords- *Data Mining, Privacy, Security, data, Multiparty,Computation.*

*****

I. INTRODUCTION

Here the term data mining implies mining the data. Mining means to extract. The verb usually refers to mining operations that extract from the Earth her hidden, precious resources. The combination of this word with data suggests an in-depth search to find multiple information which previously not noticed in the huge available data. From the viewpoint of scientific research, data mining is a relatively new discipline that has developed mainly from studies carried out in other disciplines such as computing, marketing, statistics, medical, business etc [1].

II. PROPOSED APPROACH

This work has proposed an approach for secure multiparty computation in data mining using ID3 algorithm. This work uses ID3 algorithm to enhance the privacy of the secret data in privacy preserving data mining. It display how can we effectively computed the data mining problem of decision tree learning, where any party did not know anything other than the output itself. We demonstrate this on ID3, a well-known and significant algorithm for the task of decision tree learning.

The problem with the previous work for multiparty computation is that it is not clearly describes on privacy preserving data mining and it was not checking the group performance at every step and the privacy level was not very high. This approach uses three different data sets. The proposed work increases the level of privacy by using ID3 algorithms. Previous work was giving an overall result whereas this work is implementing it in step by step manner.

Here ID3 algorithm is applied on three different data sets, which is an enhancement over previous related work where it is using randomized response techniques for making the groups between parties. In previous work it is not using randomized response techniques with ID3 algorithm. So this work proposed an approach where using the randomized response techniques with ID3 algorithm on different data sets. This proposed work enhances the privacy level for multiparty communication using different algorithm than previous work and also work on the accuracy of datasets.

III. ID3 ALGORITHM

ID3 algorithm used to create a decision tree from a fixed set of data. The concluding tree is used to classify future samples. The example has several attributes and they belong to a class like yes or no. The leaf nodes of the decision tree involve the class name whereas a non-leaf node is a decision node. The decision node is an attribute test with each branch (to another decision tree) being a possible value of the attribute. Information gain is uses in ID3 algorithm to help it decide which attribute goes into a decision node. The advantage of learning a decision tree is that a program, rather than a knowledge engineer, bring out the knowledge from an expert[9].

J. Ross Quinlan originally developed ID3 at the University of Sydney. He first presented ID3 in 1975 in a book, Machine Learning, vol. 1, no. 1. ID3 is based off the Concept Learning System (CLS) algorithm. The basic CLS algorithm over a set of training instances C:

- **Step 1:** If all instances in C are positive, then create YES node and halt.
  If all instances in C are negative, create a NO node and halt.
  Otherwise select a feature, F with values v1, ..., vn and create a decision node.
- **Step 2:** Partition the training instances in C into subsets C1, C2, ..., Cn according to the values of V.
- **Step 3:** apply the algorithm recursively to each of the sets Ci [9].

IV. PRIVATE DATA MINING.

Now discuss issues precise to the case of two-party communication where the inputs x and y are databases. They denote the two parties P1 and P2 and P1 has its private database D1 and P2 has its private database D2. First, we assume that D1 and D2 have the same structure and that the names of the attributes are public. This is necessary for carrying out any joint communication in this setting. There is a somewhat delicate issue when it comes to the names of the possible values for each attribute. On the one hand, universal names must clearly be agreed upon in order to compute any joint function. On the other hand, even the existence of a certain attribute value in a database can be sensitive information. This problem can be determine by a pre-processing phase in which random value names are assigned to the values such that they are consistent in both databases. Doing this smoothly is in itself a non-trivial problem. However, in our work we assume that the attribute-value names are also public (as would be after the above-described random mapping stage). Each party should receive the output of some data mining algorithm on the union of their databases, D1UD2. In actuality it considers a combining of the two databases so that if the same transaction appears in both databases, then it appears twice in the merged database. Finally, we assume that an upper-bound on the size of | D1UD2 | is known and public.

A. Generic Constructions

There are generic protocols that utensil secure computation for any probabilistic polynomial-time function. These protocols are different for a scenario in which there are two parties, and for the multiparty scenario where there are m > 2 parties.

B. The Two-Party Case

Secure computation in the two-party case can be efficiently implemented by a generic protocol due to Yao [2]. The protocol (or rather, simple variants of it) are proved to be secure, according to Definitions 1 and 2, against both semi-honest and malicious adversaries [3, 4].

Denote the two parties participating in the protocol as Alice (A) and Bob (B), and denote their respective inputs by x and y. Let f be the function that they wish to compute (for simplicity, assume that Bob alone learns the value f(x, y)). The protocol is based on expressing f as a combinatorial circuit with gates expressing any function g:{0,1}X {0,1} → {0,1} (including simple or, and not gates).

The protocol is based on evaluating this circuit. The number of rounds of the protocol is constant. Its communication overhead depends on the size of the circuit, while its computation overhead depends on the number of input wires (more specifically, it requires running one oblivious transfer protocol for every input wire of party B, and, in addition computing efficient symmetric encryption/decryption functions for each gate of the circuit). A more detailed analysis of the overhead of the protocol is given below. More details on the protocol and a proof of security can be found in [3]. It provides a high level description of Yao's protocol.

V. SECURE MULTIPARTY COMPUTATION (CONSTRUCTIONS)

A. Basic Building Blocks

This approach describes some simple protocols that are often used as basic building blocks, or primitives, of secure computation protocols. The protocols describe here include oblivious transfer and oblivious polynomial evaluation, which are multi-party protocols, and homomorphic encryption, which is an encryption system with special properties.

B. Oblivious Transfer

Oblivious transfer is a simple functionality involving two parties. It is a basic building block of many cryptographic protocols for secure computation

It will use a specific variant of oblivious transfer, 1-out-of-2 oblivious transfer, which was suggested by Even, Goldreich and Lempel [5] (as a variant of a different but equivalent type of oblivious transfer that has been suggested by Rabin [6]). The protocol involves multi parties, a sender and a receiver, and its functionality is defined as follows:

Input: The sender's input is a pair of strings (x0, x1) and the receiver's input is a bit σ ∈ {0, 1}.

Output: The receiver's output is xσ (and nothing else), while the sender has no output.

An efficient oblivious transfer protocol. It now presents the protocol of [7] and proves that it achieves privacy.

Computed Protocol

- **Input:** The sender has a pair of strings (m0; m1) and the receiver has a bit.
- **Auxiliary input:** The parties have the description of a group G of order n, and a generator g for the group; the order of the group is known to both parties.

1. **The protocol:** The receiver R chooses a; b; c∈R {0,……,n-1} and computes γ as follows:

   If σ = 0 then γ = (ga, gb, gab, gc)

   If σ = 1 then γ = (ga, gb, gc, gab)

   R sends γ to S.

2. Denote the tuple γ received by S by (x, y, z0, z1). Then, S checks that z0≠ z1. If they are equal, it aborts outputting. Otherwise, S chooses random u0, u1, v0, v1 ∈R{0,….,n-1} and computes the following four values:

   w0 = $x^{u_0} . g^{v_0}$ k0 = $z^{u_0} . y^{v_0}$

   w1 = $x^{u_1} . g^{v_1}$ k1 = $z^{u_1} . y^{v_1}$

3. S then encrypts m0 under k0 and m1 under k. For the sake of simplicity, assume that one-time pad type encryption is used. That is, assume that m0 and m1 are mapped to elements of G. Then, S computes co = m1 . k1 and c1 = m1 . k1 where multiplication is in the group G. S sends R the pairs (w0, c0) and (w1, c1).

4. R computes kσ = (wσ)b and outputs mσ = cσ . (kσ)-1.

VI. PRIVACY PRESERVING DISTRIBUTED COMPUTATION OF ID3

The setting considers include multi parties, each with a database of various different transactions, where all transactions have the same set of attributes (this scenario is also denoted as a “horizontally partitioned" database). The parties wish to compute a decision tree by applying the ID3 algorithm to the union of their databases. An efficient privacy preserving protocol for this problem was described in [8].

VII. MAIN TECHNIQUES

A review of the specialized constructions that were described in this section shows that they were based on some basic principles:

- A protocol can reveal intermediate results to the parties, if these intermediate results are computable from the final output. This principle was used in the construction of the protocols for computing ID3 and for computing the median.
- Homomorphic encryption can be used to perform operations on encrypted data. This is useful for analyzing data while preserving privacy (as was done in the set intersection protocol).
- Oblivious polynomial evaluation is another useful tool for analyzing and manipulating data while preserving privacy.

VIII. RESULT

The figure 1 shows the accuracy using communication between two parties and figure 2 shows the accuracy using communication between three parties and multiparty. The result shows that by using two party the accuracy is 91% but the accuracy will decrease up to 6% while using the three party computation. The accuracy will decrease 3% when the communication is done between four parties.

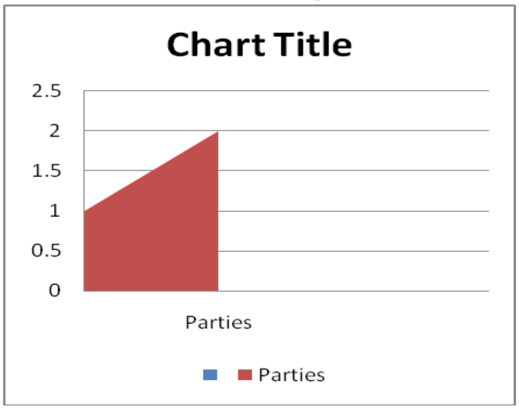

Figure 1 Accuracy in percentage using three party communication

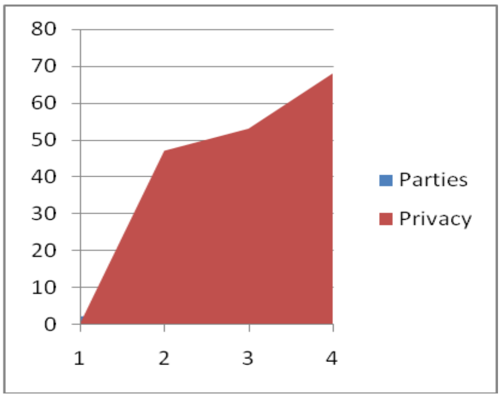

Figure 2 Accuracy in percentage using four party communication

The figure 3 shows the privacy using two party communication and figure 4 shows the privacy using four party communications. The result shows that by using two party communications the privacy is 47% and the privacy will increase up to 53% while using the three party communications. The privacy will increase 68% when the there is four party communication. When this work is compared with previous work then the results shows that the privacy is increased when we using multi party communication.

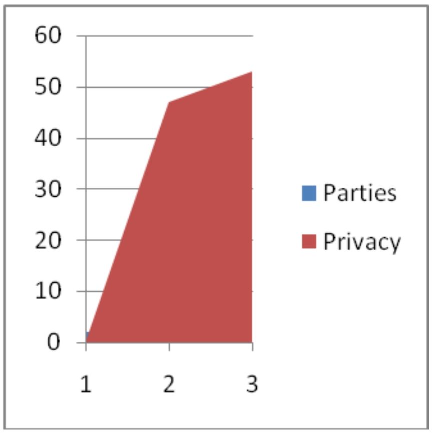

Figure 3 Privacy in percentage using three party communication

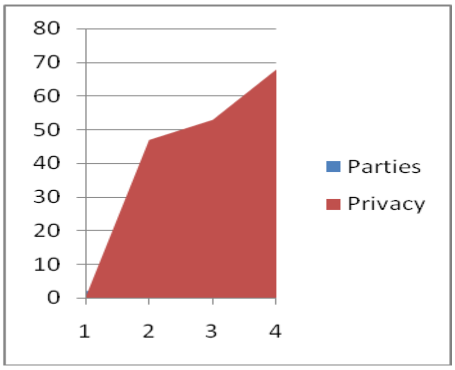

Figure 4 Privacy in percentage using four party communication

IX. Conclusion

This thesis gives a different approach for enhancing the privacy of the multi party communication in data mining. This approach uses three different data sets. And for making the parties or groups, it is using randomized response technique in privacy preserving data mining. ID3 algorithm is used to calculate the accuracy and privacy. In this experiment the ID3 algorithm is applied on the randomized data. This work shows that the privacy can be enhanced when data sets are divided in different groups. When it is compared with two party communications then it is seen that privacy increases 21% and accuracy decreases 9%.

X. References:

[1] Giudici Paolo, “Applied Data-Mining: Statistical Methods for Business and Industry” 28 January 2003 page no. 1.

[2] A. Yao, “How to Generate and Exchange Secrets”. In 27th FOCS, pages 162 to167, 1986.

[3] Y. Lindell and B. Pinkas, “A Proof of Yao's Protocol for Secure Two-Party Computation. To appear in the Journal of Cryptology”. Also appeared in the Cryptology ePrint Archive, Report 2004/175, 2004.

[4] Y. Lindell and B. Pinkas, “An Efficient Protocol for Secure Two-Party Computation in the Presence of Malicious” Adversaries In EUROCRYPT 2007, Springer-Verlag (LNCS 4515), pages 52 to78, 2007.Ramamohan, K. Vasantharao, C. Kalyana Chakravarti, A.S.K.Ratnam, “A study of data mining tools in knowledge discovery process”, IJSCE, Volume-2, Issue-3, July 2012.

[5] S. Even, O. Goldreich and A. Lempel, “A Randomized Protocol for Signing Contracts”, Communications of the ACM, pages 637 to647, 1985.

[6] M. O. Rabin, “How to Exchange Secrets by Oblivious Transfer”, Technical Memo TR-81, Aiken Computation Laboratory, 1981.

[7] M. Naor and B. Pinkas, “Efficient Oblivious Transfer Protocols”, In the 12th SIAM Symposium on Discrete Algorithms (SODA), pages 448 to 457, 2001.

[8] L. Kissner and D. Song, “Privacy-Preserving Set Operations”, In CRYPTO 2005, Springer-Verlag (LNCS 3621), pages 241 to 257, 20

[9] Training on Information and Communication Technologies through Personalised ODL “Implementation of the Personalised ODL System” Technical Report Bucharest August 2001.