Secure Protocol for Mining in Horizontally Scattered Database Using Association Rule

R. Vidyadevi
PG scholar
Computer science and engineering
KLN College of Engineering
vidyaramaraj@gmail.com

D. Pravinkumar
Assistant Professor1
Computer science and Engineering
KLN College of Engineering
Pravinpillai21@yahoo.co.in

Abstract:- Data mining is the Data mining is the analysis step of the "Knowledge Discovery in Databases" process or KDD. In this paper, produced protocol for secure protocol for mining in scattered database using association rule. Here frequent pattern tree used to find a frequent item sets. The primary part in this protocol is secure multi party algorithm in which one compute the union of private subsets that each of the interacting players hold, and another one that test the inclusion of an element hold by one player in a subset which another subset has. Our protocol provides privacy more securely than previous protocols. In addition, it is simpler and it is improved in terms of communication rounds, communication cost and computational cost than other protocols.

Index Terms—*Data Mining, Databases, frequent pattern tree, Frequent itemset, Association Rule*

*****

1. Introduction

Data mining is the non-trivial process of identifying valid, novel, potentially useful and ultimately understandable pattern in data. Data mining represents the integration of several fields, including machine learning, database systems, data visualization, statistics and information theory. Several major data mining techniques have been developed and used in data mining projects recently including association, classification, clustering, prediction and sequential patterns [1].

Knowledge discovery in databases is a complex process, which covers many interrelated steps. Key steps in the knowledge discovery process are-

DATA SELECTION- The data needed for the data mining process may be obtained from many different and heterogeneous data sources. This first step obtains the data from various databases, files and non-electronic sources.

DATA PREPROCESSING- The data to be used by the process may have incorrect or missing data. There may be anomalous data from multiple sources involving different data types and metrics. There may be many different activities performed at this time. Erroneous data may be corrected or removed, whereas missing data must be supplied or predicted.

DATA TRANSFORMATION- Data from different sources must be converted into common format for processing. Some data may be encoded or transformed into more usable formats. Data reduction may be used to reduce the number of possible data values being considered.

DATA MINING- This step applies algorithms to the transformed data to generate the desired results.

EVALUATION/INTERPRETATION -How the data mining results are presented to the users is extremely important because the usefulness of the results is dependent on it. Various and GUI strategies are used at this last step.

Knowledge Discovery in Databases (KDD) is an automated extraction of novel, understandable and potentially useful patterns implicitly stored in large databases, data warehouse and other massive information repositories. KDD is a multi-disciplinary field drawing work from areas including database technology, artificial intelligence, machine learning, neural networks, statistics, pattern recognition, information retrieval, high performance computing and data visualization [1].

Association rule mining finds interesting association or correlation relationship among large set of data items and it satisfy both minimum support threshold and minimum confidence threshold. It is used to predicate and analysis the customer behavior

Association rule there are two step approaches

1. Frequent Itemset Generation
   - Generate all itemsets whose support ≥ min support
2. Rule Generation
   - Generate high confidence rules from each frequent itemset, where each rule is a binary partitioning of a frequent itemset
   - Frequent itemset generation is still computationally expensive

Scattered database is multiple a blocks are read into a database block buffer that are scattered throughout memory

Frequent pattern tree is one way to find frequent item sets without candidate item set generations and improving their performance and easy to compute.

Advantages of FP:

The most significant advantage of the FP-tree

- Scan the DB only twice and twice only.

Completeness:

- the FP-tree contains all the information related to mining frequent patterns (given the min-support threshold). Why?

Compactness:

- The size of the tree is bounded by the occurrences of frequent items
- The height of the tree is bounded by the maximum number of items in a transaction

Kantarcioglu and Clifton studied that problem in [2] and devised a protocol for its solution .The main part of the protocol is a sub-protocol for the secure computation of the union private subset that are held by different palyers.That is most costly part of the protocol and its implementation depends an cryptographic primitives such as commutative encryption, oblivious transfer and hash function .It also lead to information leakage and that solution not perfectly secure

Here propose protocol implements the a parameterized family function, which we call threshold functions to compute the union and intersection of private subsets and easily find the frequent itemset in large database

The remainder of this paper is organized as follows section 2.survey about predicating a frequent itemsets and section 3.methodology about compute union and private and intersection of private subsets section 4.conclusion section

2. Survey about predicating a frequent itemsets

Murat Kantarcioglu and Chris Clifton[2] suggested a two phase approach which is used to find a frequent item sets that is first phase is discovering a candidate item sets and another phase is determining which of the candidate item sets meet the global support/confidence threshold but it occur collision among padded item sets would seem as information leakage and it implementation cost is high

Jaideep Vaidya and Christopher W. Clifton, [3] suggested to define scoring protocol is used to predicate a frequent patterns and prevent the revealing information. To use Fagin's algorithm discover a radius separating the kth element while minimizing disclosing of information required to efficiently the perform task and their information was leaked when top k results send to the third party so performance can be low

Shuguo Han and Li Wan, and Vincent C.S. Lee [4] suggested a Gradient descent method to provide privacy and there are two approaches such as least square approach ,stochastic approaches main aim is to minimize a target function in order to reach a local minimum .but least square approach dose not support multi party computation

Tamir Tassa and Dror J. Cohen[5] suggested a sequential clustering algorithm is used to prevent the revealing information that is do given a set of sequences , create groups such that similar sequences in the same group and provide privacy is not reliable because discovering a optimum value repeatedly

3. Methodology

To compute union and intersection of private database using UNFI-KC protocol (unifying list of locally frequent item sets) and it is perfectly secure so not reveal any information. This protocol is find locally frequent item sets and then find a globally frequent item set. here threshold function invoke their binary vectors such as{0,1} each and every player can find locally frequent itemsets using this threshold function and finally identify association rule pattern while support count can be calculated.

3.1 Initialization

Let D be a transaction database which is viewed as binary matrix of N rows and L columns. The database D is partitioned horizontally between M players , denoted as $p_1, p_2 \dots p_M$.players p_m holds the partial database D_m that contains $N_m = |D_m|$ of the transaction in D, $1 \leq m \leq M$ .Get the data set and load the data set into memory and finally horizontally spilt the data between players so all players get data sets .players are easily find the locally frequent itemset

3.2 Generate a locally frequent item sets

Frequent pattern tree is used to find a locally frequent itemsets without candidate item generations. It have a two properties that is node link property and prefix path property. It contain table consist of item name and head of the node.

General idea (divide-and-conquer) for Recursively grows frequent patterns using the FP-tree: looking for shorter ones recursively and then concatenating the suffix:

For each frequent item, construct its conditional pattern base, and then its conditional FP-tree;

Repeat the process on each newly created conditional FP-tree until the resulting FP-tree is empty, or it contains only

 ISSN: 2321-8169
676 - 679

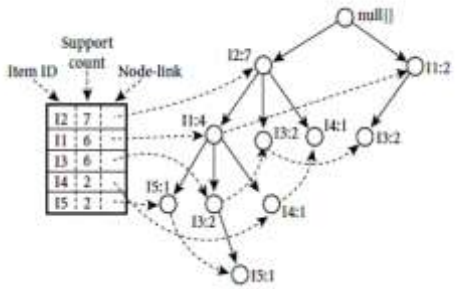
one path (single path will generate all the combinations of its sub-paths, each of which is a frequent pattern)

FP tree constructs two steps

- Scan the transaction DB for the first time, find frequent items (single item patterns) and order them into a list L in frequency descending order.
  In the format of (item-name, support)

- For each transaction, order its frequent items according to the order in L; Scan DB the second time, construct FP-tree by putting each frequency ordered transaction onto it.

Diagrammatic representation of FP tree used to predicate a frequent item set



Sets of locally frequent k -item sets denoted as $c_s^{k,m}$,$1 \leq m \leq M$, are subset of $FP(F_s^{k-1})$, they may be encoded as binary vectors of length $n_k = |FP(F_s^{k-1})|$. The binary vector that encodes the union $c_s^k = \cup_{m=1}^{M} c_s^{k,m}$ is the OR of the vector that encodes the sets $c_s^{k,m}$, $1 \leq m \leq M$.hence the players can compute the union by invoking threshold function on their binary vectors. finally all players can predicate a locally frequent item denoted as c_s^k sets but it provide item sets will be secure because item set should be encrypted that is each player add to his private subset $c_s^{k,m}$ fake itemsets, in order to hide its size. Then the players jointly to compute the encryption of their private subsets by applying on those subsets a encryption, where each player adds, in his turn, his own layer of encryption using their private secret key. At the end of the stage, every item set in each subset is encrypted by all of the players; the usage of encryption scheme ensures that all item sets are, eventually encrypted in the same manner. Then they compute the union of those subsets in their encrypted form. Finally they decrypt the union subset and it remove from itemset identified as fake

3.3 Generate globally frequent item sets

Each player jointly to compute the union of private database using protocol UNFI-KC yield the set c_s^k that consist of all item sets that are locally s-frequent in at least one site Those are the k-item sets that have potential to be also globally s-frequent. In order to reveal which of those item sets is globally s-frequent there is a need to securely compute the support of each those item set. That computation must not reveal the local support in any of the sites. Here semi honest players can be used to verify the inclusion so collision not occurs. The idea is to verify that inequality by starting an implementation of the secure union of the protocol [6] on the private inputs. In that protocol, all players jointly compute random additive shares of the required sum of the candidate item sets, it is globally s-frequent and then, by sending all shares to, say,p_1, he may add them and then reveal the sum. If, however, p_M withholds his share of sum, then $p_1$will have one random share and similarly p_M will have corresponding share and then two players execute generic secure circuit evaluation in order to verify whether the sum the item set equal or not .

3.4 Identifying all (s, c) association rules

Identifying association rule after finding a s-frequent item sets for example An implication expression of the form X → Y, where X and Y are itemsets Support (s)-percentage of transactions that contain both X and Y .Confidence (c)-Measures how often items in Y appear in transactions that contain X .It satisfy both minimum support threshold and minimum confidence threshold. Identify association rules(s , c) in all subsets

4. Conclusion

To prevent information disclosure in horizontally scattered database using association rules, a secure protocols was developed that improves significantly in terms of preserving privacy and efficiency because when there are no one collusion is occur. This protocol allows parties to share data in a private way with no one data can be leaked and with no loss of their data. One of the implemented protocols is secure multi party computation protocol which is used for computing the union of private subsets that each interacting player holds. Those protocols exploit the fact that the underlying problem is refined, when the number of players is greater than two. This protocol provides privacy while retrieving the data from scattered database and improves their computation speed, communication rounds so providing information is perfectly secure.

References

[1] Aggarwal, Ch Ph. Yu, "Privacy-Preserving Data Mining: Models and Algorithms", Kluwer Academic publishers,2007

[2] Kantarcioglu, M. Nix, R. and Vaidya, J."An Efficient Approximate Protocol for Privacy-Preserving Association Rule Mining ,"Proc. 13th Pacific-Asia

Conference. Advances in Knowledge Discovery and Data Mining (PAKDD), pp. 515-524, 2009

[3] Jaideep Vaidya and Christopher W. Clifton, "Privacy-Preserving Kth Element Score over Vertically Partitioned Data" ieee transactions on knowledge and data engineering, vol. 21, no. 2 , February 2009

[4] Shuguo Han and Li Wan, and Vincent C.S. Lee, "Privacy-Preserving Gradient-Descent Methods" ieee transactions on knowledge and data engineering, vol. 22, no. 6, June 2010

[5] Tamir Tassa and Dror J. Cohen, "Anonymization Of Centralized And Distributed Social Networking By Sequential Clustering " ieee transactions on knowledge and data engineering, vol. 25, no. 2, February 2013

[6] Benaloh, J.C "Secret Sharing Homeomorphisms: Keeping Shares of a Secret ," Proc. Advances in Cryptology (Crypto), pp. 251-260, 1986.

[7] Cheung, D.W.L. Ng, V.T.Y. Fu, A.W.C. and Fu, Y. "Efficient Mining of Association Rules in Distributed Databases," IEEE Transaction Knowledge and Data Eng., vol. 8, no. 6, Dec. 1996

AUTHORS:

R.Vidyadevi
ME-Computer Science and Engineering
KLN College of Engineering
Pottappalayam, Sivagangai(dist)

MR D.PRAVINKUMAR
Associate professor1
KLN College of Engineering
Pottappalayam, Sivagangai(dist)