Review on Association Rule Mining for Privacy-Preserving

Ms.Rasika khairnar[1], Prof.P.D.Lambhate[2]
[1, 2] Department of Computer Engineering
[1, 2] JSPM College, Savitribai Phule Pune University, Pune, India

Abstract- *Cloud computing makes use of the conceptual model of data mining as a service, by making use of these it is by all accounts a conspicuous decision for organizations saving money on the cost of adding security, manage and keep up an IT framework. An association with no mining capacity can outsource its mining needs to specialist service provider on a cloud server. Notwithstanding, both the association rules and thing set of the outsourced database are seen as private property of the organization. The owner of information encrypts the information and forewords it to the server to save the corporate privacy. User forewords mining queries to server, and afterward server conduct data mining and sends encrypted pattern to the customer. To get true pattern user decrypts encoded pattern. The survey provides a study on various techniques for the association rule mining.*

Keywords- Cloud Computing, Association rules mining, Privacy-preserving outsourcing.

I. INTRODUCTION

The latest couple of years Data Mining have ended up being progressively famous. Together with the information age, the digital revolution made it vital to use a couple of heuristics to have the ability to look at the expansive measure of data that has ended up being available. Data Mining has especially ended up being prominent in the fields of forensic science, fraud analysis, and healthcare, for it minimizing costs in time and money.

Data mining can be defined in various ways one can define data mining as "The process which has applying data analysis as well as algorithms for discovery that has nominal computational efficiency constraints, generates a specific enumeration pattern on the information." Or "The enlistment of justifiable models and patterns from databases". As such at the start, we have a huge amount of data as well as possible models. Information Mining ought to bring about those models that describe the information best, the models that fit.

Data mining has been distinguished as the methodology that offers the potential outcomes of discovering the hidden knowledge from these collected databases. Research efforts endeavors dedicated to information mining, which concentrated on classification and prediction accuracy, have as of late been experiencing a huge change. The nonstop advancement of more complex classification models through commercial and software packages have ended up providing a few advantages just in particular issue domains where some earlier background information or new evidence can be misused to additionally enhance classification performance. All in all, in any case, related research demonstrates that no individual information mining method has appeared to deals well with a wide range of classification problems.

The objective of Association rule mining is the investigation of item sets that co-occur as often as possible in value-based information. The issue has a gigantic worst-case complexity, a reality that spurs business to use the mining methodology to service providers, who have created proficient, particular solutions. The information owner, aside from the mining cost relief, has additional motives for outsourcing. Firstly, it requires negligible computational resources, since the owner is just required to create and to send the exchanges to the miner. This makes the outsourcing model likewise attractive to applications in which information owners create exchanges as streams and they have constrained resources to look after them. Second, expect that the owner has different production sources of transactions, e.g., consider a chain of grocery stores which produce exchanges at particular areas. For mining affiliation runs all exchanges can be sent to a single provider. For the entire organization the supplier could estimate association rules that are local to the single stores or worldwide standards. Along these lines, the cost of transferring exchanges between the sources and executing the global mining in a distributed way is stored.

Section II gives the Literature review for the Association Rule Mining for Privacy-Preserving.

II. LITERATURE REVIEW

Li, Lichun focused on privacy-preserving mining [1] on vertically partitioned databases. In such a case, data owners wish to learn the association rules or frequent item sets from a collective data set and exhibit as little information about their (sensitive) raw data as possible to other data owners and third parties. To assure data privacy, design an efficient

homomorphic encryption scheme and a secure comparison scheme. Then implement a cloud-aided frequent item set mining solution, which is used to build an association rule mining solution. The solutions are designed for outsourced databases that allow multiple data owners to efficiently transfer their data securely without compromising on data privacy. These solutions escape less data about the raw data than most existing solutions. In comparison to the only known solution achieving a similar privacy level as the proposed solutions, the performance of our proposed solutions is three to five orders of magnitude higher. Depend on that experiment findings using different parameters and data sets, demonstrate that the run time in each of solutions is only one order higher than that in the best non-privacy-preserving data mining algorithms. Since both data and computing work are outsourced to the cloud servers, the resource consumption at the data owner end is very low.

Author of this paper, proposed a novel Protocol for Outsourced Rule Mining (PORM)[2]. PORM excites rule mining in a cloud environment where data are both encrypted and outsourced. They formally proved that PORM is both correct and secure, and also extended PORM to the multiple-user scenario Rule mining, for discovering valuable relations among items in large databases, has been a popular and well researched technique for years. However, such old but important methods faces huge challenges and difficulties in the era of cloud computing although which affords both storage and computing scalability: 1) data are outsourced to a cloud due to data explosion and high storage and management cost, 2) moreover, data are usually encrypted first before being outsourced for privacy's sake. Existing privacy-preserving rule mining methods only assume a distributed model where every data owner holds the self data without encryption and together follow a secure protocol to perform rule mining. To address this limitation, proposed a novel Protocol for Outsourced Rule Mining (PORM).

Cheung , implemented a protocol for secure mining of association rules in horizontally distributed databases. The current leading protocol is that of Kantarcioglu and Clifton. This protocol, like theirs, is depending on the Fast Distributed Mining (FDM) algorithm of Cheung et al [3]., which is an unsecured distributed version of the Apriori algorithm. The main components in this protocol are two novel secure multi-party algorithms-one that estimate the union of private subsets that each of the interacting players hold, and another that tests the inclusion of an element held by one player in a subset held by another. This protocol provides enhanced privacy with respect to the protocol. Additionally, it is simpler and is significantly more efficient in terms of communication rounds, communication cost and computational cost.

Samet, Saeed, and Ali Miri ,present a new protocol for privacy-preserving association rule mining [4], to overcome the security flaws in existing solutions, with better performance, when data is vertically partitioned among two or more parties. Two sub-protocols for secure binary dot product and cardinality of set intersection for binary vectors are also designed which are used in the main protocols as building blocks. Association rule mining offers useful knowledge from raw data in various applications such as health, insurance, marketing and business systems. Similarly, many real world applications are divided between two or more parties, each of which wants to keep its sensitive information private, while them collaboratively gaining some knowledge from their data. Therefore, secure and distributed solutions are needed that do not have a central or third party accessing the parties' original data.

Zhao, Chunye, et al ,designed user behavior (UB)-based scheme [5], by evaluating log data from cloud servers. First of all, present the description rules of UB for operating system logs. Then put forward an association rule mining algorithm depend upon the Long Sequence Frequent Pattern (LSFP) to extract UB. At last, the result of experiment proves that this solution can implement the track and forensic of data leakage efficiently for the cloud security auditing. Cloud Storage offers external data storage service by combing and coordinating distinct types of storage devices in the network, so that they can collectively work together. However it always exists a trust game relationship between users and service providers, therefore building a healthy, fair and secure cloud data service environment is necessary, especially for the security auditing on the state of cloud data and operation processes.

Building a real system is one of the major challenges of privacy-preserving data mining (PPDM). In this paper,[6] developed a CRYPPAR, a novel, full-fledged framework for privacy preserving association rule mining based on a cryptographic approach. They use secure scalar product protocols and public-key cryptosystems in CRYPPAR to efficiently mine association rules over vertically partitioned data. Also introduce a partial topology to lower communication cost as much as possible. Empirically results display that the schema is efficient in privacy-preserving association rules and may become a general framework for PPDM systems.

Ren, Jinghan, and Baowen Zhang , proposes an improvement on an existed substitution cipher encryption algorithm: non-deterministic one-to-n item mapping [7]. Encrypting transactional data is essential to outsource association rule mining for the purpose of privacy preserving.

However encryption transformation would affect the efficiency which is another concerned aspect of the delivering process. The new transformation has greater efficiency while it is still valid and secure not be covered by one-to-one mappings. Both theoretical analysis and experiments validate this work.

Author of this paper], proposed an efficient algorithm EP-DMA [8]. Compared with DMA, the EP-DMA is a global hashing and cryptographic technique. Although EP-DMA has solved the issue of security, messages for communication raised. Association rules mining is one of the most important and fundamental issue in data mining. Recently, in need of security, more and more people are studying privacy-preserving association rules mining in distributed database.

III. PROPOSED SYSTEM

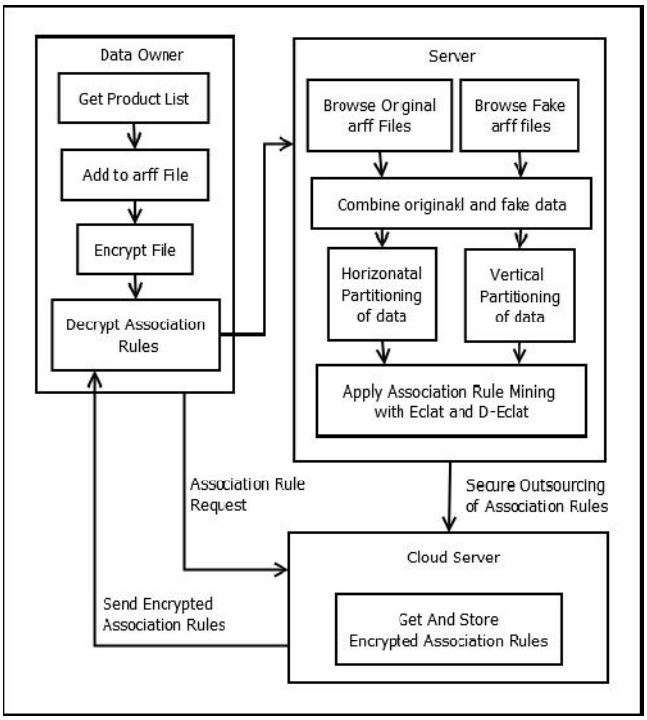

Fig1: Proposed system Architecture

To take care of the issues in present system we propose an idea of the system in which will have three important parts those are data owner, system server and cloud server. In this system, multiple data owners send their private databases to server. For maintaining the security, databases are encrypted before storing to server. At server side, all collected databases from multiple data owners are combined. These combined databases may contain original or fake data. This database is partitioned horizontally and vertically. On these partitions association rules mining is performed by using Eclat as well as D-Eclat algorithm. These encrypted association rules are then outsourced to cloud server. On data owner's request, cloud server provides the encrypted rules which are decrypted locally at data owner side.

Mathematical Model

Let S be as system such that,
S = {Input, process Output}

Input: Dynamic dataset

D = {TID, Transaction, ERV}
Where, D is the database of owner containing product list of multiple customers. Database has three fields such as:

TID = transaction ID
Transaction = this contains number of products selected by customer

ERV = Encrypted realness value
ERV = {0, 1}

Where, 0 = Fake data
1 = Original data

Process:

1. At data owner
   P = {P1, P2, P3, p4, p5}
   Where,

   P1 = Get Product List of multiple customers
   L = {items selected by customers}

   P2 = Generate arff File
   F = {pl1, pl2, …, pln}
   Where,
   F is the arff file contains n product list of customers. After each transaction, product list is added to arff file.

   P3 = Encryption of arff file
   E = {e1, e2, e3}
   Where,
   e1 = cryptographic hash function
   Used to encrypt TID
   e2 = Encryption with substitute cipher
   Used to encrypt transactions
   e3= Probabilistic homomorphic encryption function
   used to encrypt the ERV
   p4 = Send encrypted arff file to server

p5 = Decryption of received association rules from cloud server.
AR = {ar1, ar2, …., arn}
Where AR is the association rules received from cloud server.

2. At Server
P = {p1, p2, p3, p4}

P1 = Browse arff files and Merging
F = {f1, f2}
Where, F1 is the original arff file and f2 is the fake arff file. F is the merged database of f1 and f2.

P2 = Partitioning of databases
DP = {H, V}
Where, DP is the partitions of combined database.
H = Horizontal partitioning of database
V = Vertical partitioning of database

P3 = Association Rule Mining
Association rules mining is perform on encrypted database by using two algorithms such as:
A = {a1, a2}
Where,
a1 = Association rule mining with Eclat
a2 = Association rule mining with D-Eclat

P4 = Outsourcing of encrypted generated association rules

3. At Cloud Server
P = {Req, Resp}

Req = Cloud server receive the request from data owners for association rules
DOR = {r1, r2,…., rn}
Where, DOR is the set of n number of data owners request to the cloud server.

Resp = In return, cloud server send the encrypted association rule to the data owners

Output: Encrypted Association rules received at data owner. Data owners can decrypt these rules by using the decryption key.

AR = {ar1, ar2… arn}

Where AR is the association rules received from cloud server.

Algorithm

$Input: E((i_1, t_1), \dots (i_n, t_n))|P), s_{min}$
$Output: F(E, s_{min})$
1: $for\ all\ i_j\ occuring\ in\ E\ do$
2: $P \coloneqq P \cup i_j$ // add i_j to create a new prefix
3: $init(E')$ // initialize a new equivalence class with the new prefix P
4: $for\ all\ i_k\ occuring\ in\ E\ such\ that\ k > j\ do$
5: $t_{tmp} = t_j \cap t_k$
6: $if\ |t_{tmp}| \geq s_{min}\ then$
7: $E' \coloneqq E \cup (i_k, t_{tmp})$
8: $F = F \cup (i_k \cup P)$
9: $end\ if$
10: $end\ for$
11: $if\ E' \neq \{\}\ then$
12: $Eclat(E', s_{min})$
13: $end\ if$
14: $end\ for$

IV. RESULTS AND DISCUSSION

A. Experimental Setup

The system is built using Java framework on Windows platform. The Net beans IDE is used as a development tool. The system doesn't require any specific hardware to run; any standard machine is capable of running the application.

B. Database

Database contain arff file. This file is dynamically created by data owners. This file contains transaction ID, transactions and ERV. For each customer, unique ID is allocated. Transaction contain list of all products purchase by customer. ERV represent the reality of data that is whether it is original or fake. 1 represent original transaction and 0 represent fake transaction.

C. Expected Results

Table I describes the comparative analysis of 4 different combinations of systems for association rules mining. In this system, dataset is either horizontally or vertically partitioned and for association rule mining Eclat or D-Eclat algorithm is used. Table represents the accuracy of mined rules and time requires generating the rules. Accuracy is measured in percentage and time is measured in milliseconds.

Table I Result comparison

System No.	Association Rule Mining Algorithm	Accuracy in %	Time in ms
1	Eclat on Horizontal data partitioning	75	1470
2	Eclat on Vertical data partitioning	82	1350
3	D-Eclat on Horizontal data partitioning	84	1200
4	D- Eclat on Vertical data partitioning	92	990

Figure 2 represent the graphical view of comparison of accuracy obtained in 4 systems described in table 1. D-Eclat on vertically portioned data is more accurate that other 3 systems.

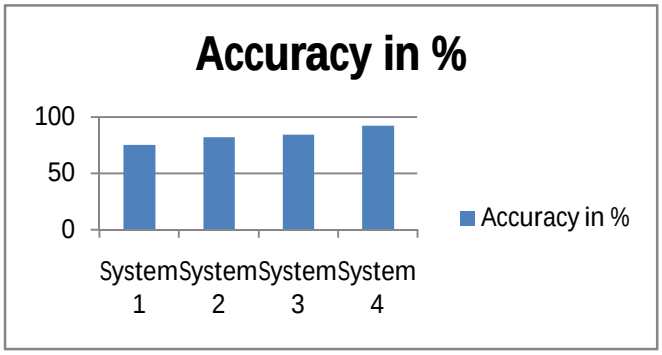

Figure 2: Accuracy Comparison

Figure 3 represent the graphical view of comparison of time required to generate association rules for 4 systems described in table 1. D-Eclat on vertically portioned data is more efficient that other 3 systems.

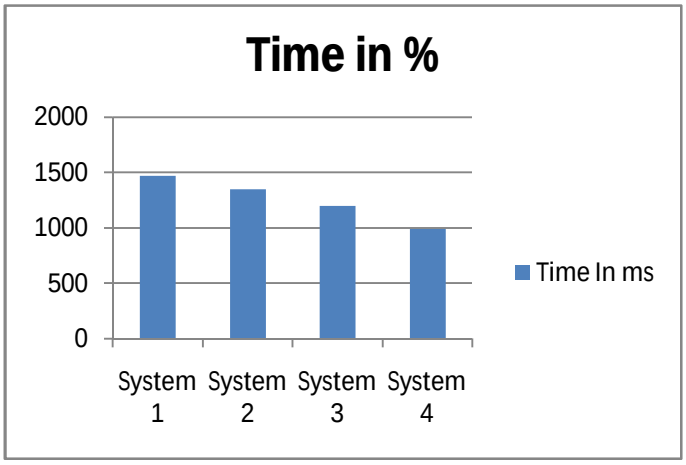

Figure 3: Time Comparison

REFERENCES

[1] Li, Lichun, et al. "Privacy-Preserving-Outsourced Association Rule Mining on Vertically Partitioned Databases." IEEE Transactions on Information Forensics and Security 11.8 (2016): 1847-1861.

[2] Liu, Fang, Wee Keong Ng, and Wei Zhang. "Encrypted Association Rule Mining for Outsourced Data Mining." 2015 IEEE 29th International Conference on Advanced Information Networking and Applications. IEEE, 2015.

[3] Tassa, Tamir. "Secure mining of association rules in horizontally distributed databases." IEEE Transactions on Knowledge and Data Engineering 26.4 (2014): 970-983.

[4] Samet, Saeed, and Ali Miri. "Secure two and multi-party association rule mining." 2009 IEEE Symposium on Computational Intelligence for Security and Defense Applications. IEEE, 2009.

[5] Zhao, Chunye, et al. "Efficient association rule mining algorithm based on user behavior for cloud security auditing." Online Analysis and Computing Science (ICOACS), IEEE International Conference of. IEEE, 2016.

[6] Tran, Duc H., Wee Keong Ng, and Wei Zha. "CRYPPAR: An efficient framework for privacy preserving association rule mining over vertically partitioned data." TENCON 2009-2009 IEEE Region 10 Conference. IEEE, 2009.

[7] Ren, Jinghan, and Baowen Zhang. "An Improvement on a Non-deterministic One-to-n Substitution Scheme in Outsourcing Association Rule Mining." Computer Science and Information Engineering, 2009 WRI World Congress on. Vol. 4. IEEE, 2009.

[8] Liu, Jie, Xiufeng Piao, and Shaobin Huang. "A privacy-preserving mining algorithm of association rules in distributed databases." First International Multi-Symposiums on Computer and Computational Sciences (IMSCCS'06). 2006.

[9] Mr. Nitin J.Ghatge, Prof. Poonam D. Lambhate "An Effective Use of Meta Information for Text Mining" International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 4, Issue 6, June 2015.

[10] Creighton, Chad, and Samir Hanash. "Mining gene expression databases for association rules." Bioinformatics 19.1 (2003): 79-86.

[11] Keshavamurthy, Bettahally N., Asad M. Khan, and Durga Toshniwal. "Privacy preserving association rule mining over distributed databases using genetic algorithm." Neural Computing and Applications 22.1 (2013): 351-364.

[12] Wong, Wai Kit, et al. "Security in outsourcing of association rule mining." Proceedings of the 33rd international conference on Very large data bases. VLDB Endowment, 2007.

[13] Dong, Boxiang, Ruilin Liu, and Hui Wendy Wang. "Result integrity verification of outsourced frequent item set mining." IFIP Annual Conference on Data and Applications Security and Privacy. Springer Berlin Heidelberg, 2013.

[14] Rozenberg, Boris, and Ehud Gudes. "Association rules mining in vertically partitioned databases." Data & Knowledge Engineering 59.2 (2006): 378-396.