International Journal On Engineering Technology and Sciences – IJETS™
ISSN(P): 2349-3968, ISSN (O): 2349-3976
Volume IV, Issue V, May - 2017

Privacy Preserving Secured Mining of Heterogeneous Data with Attack Prohibition System

***S.Sindhu Biravi*[1], *M.Kalaiselvi*[2]**
M.E (Scholar), Dept. Computer Science and Engineering, Vivekanandha College of Technology for Women[1]
Asst. Professor, Dept. Computer Science and Engineering, Vivekanandha College of Technology for Women[2]
Email id:Sindhubiravi8@gmail.com, Kalaiselvi.mayilsamy@gmail.com

ABSTRACT

Spurred by developments such as secured mining, there has been considerable recent interest in paradigm of data mining-as-a-service. The company lacking in expertise or computational resources can outsource it's data to the server. Transaction Items and association rules of outsourced database are considered private property of data owner. To protect privacy, data owner transforms and ships the data to server, by sending mining queries to server and true patterns can be recovered. In this paper, problem of outsourcing association rule mining task within the corporate privacy-preserving framework is focused. Proposed an attack model based on background knowledge and optimized distributed association rule mining (ODARM) for privacy preserving outsourced mining. ODARM algorithm ensures that each transformed item was indistinguishable, with respect to Attacker's background knowledge, from at least k-1 other transformed items. Our comprehensive experiments on the very large and real transaction database demonstrate that techniques protect heterogeneous data effectively.

Key words **- Association rule mining, Frequent item set mining, ODARM, Data anonymization.**

I.INTRODUCTION

Now a days there was large amount of data proceed in every day from different sources. Large amount of data are stored in different database. These data are stored in storage devices in the form of row data. Data mining was process of discovering interesting pattern and knowledge from large amount of data. Data mining techniques are used for text analysis, bioinformatics, Direct mail marketing, credit card fraud detection and market basket analysis. Extracting knowledge from raw data, Privacy preserving in data mining was one of technique that deal with security of data that are extracted from large dataset. There are various Data Mining Tasks: Classification Clustering, Association Rule Mining and Sequential Pattern Mining Regression.

PRIVACY PRESERVING DATA MINING

Data mining refers to extracting or mining knowledge from large amounts of data. Also known as Knowledge Discovery in Databases (KDD).Limitation with data mining output was that it will discloses some information, which was considered to be private and personal. Unauthorized access to such personal data causes the peril to individual privacy.

Figure 1-Data mining cycle

Recent research in area of privacy preserving data mining has effort to determine the trade-off between need for knowledge discovery and privacy, which was necessary in order to improve decision-making processes and other human activities. Privacy preserving data mining cope with problem of learning accurate models over aggregate data, while protecting privacy at level of individual records. Main purpose of PPDM was to design competent frameworks and algorithms that can extract relevant knowledge from the large amount of data without revealing of any sensitive information [9]. It protects sensitive information by providing sanitized database of original database on internet or the process was used in such the way that private data and private knowledge remain private even after mining process. It was PPDM due to which benefits of data mining be enjoyed, without compromising privacy of concerned individuals.

ASSOCIATION RULE MINING

Association rule mining one of task of data mining. Association rule mining was important field to under privacy preserving data mining. R. Agrawal was first proposed basic concept of Association rule mining. Association rule basically use the concept of IF-THEN relationship among different data. Following example of shows concept of Association rule. "If customer buys the laptop, then he/she was 85% likely to also purchase anti-virus ".

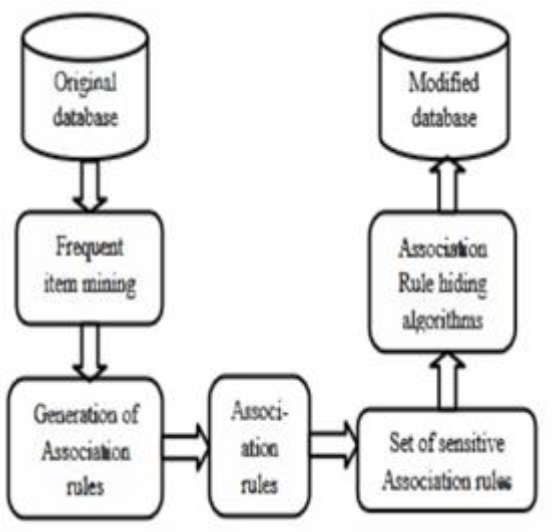

Figure 2-architecture of association rule hiding

From above example laptop was somewhat related to anti-virus because every time customer buy the computer then he/she buy anti-virus was observed. Association rule was used for market basket analysis. Let 1= {II, I2... In} be the set of item and D be the database of transactions where each transaction T was the set of item such that T belongs to I. For every transaction was associated to an identifier, called TID. An association rule was applied of form Every association rule must satisfy two criteria's support and confidence. Support of rule A-B was transaction database that contain support count of AUB and (A-B) can be calculated.

AUB support (A-B) belongs to D where D was total number of transaction in transaction database. Confidence of rule A-B was transaction database that contain the also contain B. confidence for rule (A-B) can be calculated using below formula in (2).

ASSOCIATION RULE HIDING

Association rule hiding was one technique to Privacy Preserving Data Mining. Association rule hiding methodology aim was to sanitize original data. so it may be applied to following condition:(1)Sanitized database was mining of all non-sensitive rules (2)Sanitized database was not reveal any sensitive rules.(3) Sanitized database was not add any new rules, not present in database D. Association rule hiding depends on support and confidence of rule, There was two ways to hide any rule (i) Decrease confidence up to certain threshold. (ii) Decrease support up to certain threshold.

There are many methodologies for maintaining privacy in transaction database. Before developing a tool it was necessary to determine time factor, company strength and economy. Once these things are satisfied, then next steps was to determine which operating system and language can be used for developing tool. Once programmers start building tool, programmers need lot of external support. These supports can be received from senior programmers, from book or websites. Before building system above considerations should be taken into account for developing proposed system.

Concept and meaning of privacy has been debated by philosophers, social scientists, academic lawyers and other scholars. All definitions are based on assumptions about individualism and about distinction between realms of civil society and state. However, many gloss over essential cultural, class-related and gender differences. Literature on privacy tends to give readers an overwhelming sense that privacy was the deeply contested concept, which often varies according to context and environment.

Westin also addresses specific functions that privacy plays. It promotes freedom of association. It shields scholarship and science from unnecessary interference by government, Permits use of the secret ballot and protects voting process by forbidding government surveillance of the citizen's past voting record. Restrains improper police conduct such as unreasonable search and seizure and also serves to shield those institutions, such as press, that operate to keep government accountable.

Privacy has been defined comprehensively:

Privacy was the concept related to solitude, autonomy, and secrecy, but it was not synonymous with these terms; for beyond purely descriptive aspects of privacy as curiosity, isolation from company and influence of others, it implies the normative element: right to privacy asserts sacredness of person, right to exclusive control of access to private realms,; any invasion of privacy constitutes an offence against rights of personality – against individuality, freedom and dignity.

Privacy can be divided into following facets Territorial privacy – concerning setting of limits on intrusion into domestic and other environments such as workplace or public space.

- Privacy of person – these was concerned with protecting the person against undue interferences such as physical searches and drug testing, and information that violates his or her moral sense;
- Privacy of communications, covering security and privacy of mail, telephones, email and other forms of communication;
- Privacy in information context – deals with compilation, gathering, and selective dissemination of personal information such as medical records and credit data.

The discourse on privacy as the policy issue has largely focused on information privacy and it was the fact of privacy that these research projects will focus on (Westin, 1967, p7). Increased concern with privacy of communications has caused some confusion between meanings of information security and information privacy and terms are often used interchangeably. It can be defined as "group or institution to determine how, when and what to extent information about them was communicated to others". As Clarke noted term 'privacy' was used by some people refers to security of data during transmission as protection against various risks, such as modified by unauthorized persons or data being accessed. However, rise to prominence of e-commerce and Internet communications has led to privacy of communications attracting more attention and concern. These aspects, however, are only the small fraction of considerations within field of 'information privacy'. That is, data security was the necessary but not sufficient condition for information privacy.

II. PROPOSED WORK

The particular problem attacked was outsourcing of pattern mining within the corporate privacy-preserving framework. The key distinction between these problem and abovementioned PPDM problems was that, not only underlying data but also mined results are not intended for sharing and must remain private. In particular, when server possesses background knowledge and conducts attacks on that basis, it should not be able to guess correct candidate item or item set corresponding to the given cipher item or item set with the probability above the given threshold. Proposed a protocol to solve these problem by using k-privacy, i.e., each item in outsourced dataset should be indistinguishable from at least k − 1 items regarding their support. In addition, claimed that the information that our protocol may leak is less sensitive than the excess information leaked by the protocol. It uses Generalization and specialization technique to hide the data by making an abstract view on the details stored. In order to make the data more precise avoid showing the sensitive data to the service provider.

This work was to devise encryption schemes such that formal privacy guarantees can be proven against attacks conducted by server using background knowledge, while keeping there source requirements under control. Proposed an ODARM algorithm for the secure computation of the union of private subsets. The proposed algorithm improves upon that in terms of simplicity and efficiency as well as privacy. In particular, algorithm does not depend on oblivious transfer and commutative encryption (what simplifies it significantly and contributes towards much reduced communication and computational costs).The algorithm that proposed here computes a parameterized family of functions, which is called threshold functions, in which the two extreme cases correspond to the problems of computing the union and intersection of private subsets.

ENCRYPTION/DECRYPTION SCHEME

Encryption:

In this section, introduce encryption scheme, which transforms the TDB D into its encrypted version D∗. Our scheme was parametric w.r.t. k > 0 and consists of three main steps: (1) using 1-1 substitution ciphers for each plain item; (2) using the specific item k-grouping method; (3) using the method for adding new fake transactions for achieving k-privacy.
The constructed fake transactions are added to D) to form D∗, and transmitted to server.

Decryption:

When client requests execution of the pattern mining query to server, specifying the minimum support threshold σ, server returns computed frequent patterns from D∗. Clearly, for every item set S and its corresponding cipher item set E, have that supp D(S) ≤ supp D_(E).For each cipher pattern E returned by server together with supp D_(E), E/D module recovers corresponding plain pattern S. It needs to reconstruct exact support of S in D and decide on these bases if S was the frequent pattern. To achieve these goals, E/D module adjusts support of E by removing effect of fake transactions. Supp D(S) = supp D_(E)−supp D\D(E). These follows from fact that support of an item set was additive over the disjoint union of transaction sets. Finally, pattern S with adjusted support was kept in output if supp D(S) ≥ σ. calculation
of supp D_\D(E) was performed by E/D module using synopsis of fake transactions in D∗ \ D.

III. SYSTEM MODELS

A. THE PATTERN MINING TASK

The reader was assumed to be familiar with basics of association rule mining. Let I = i1... in be set of items and D = t1... tm the transaction database (TDB) of transactions, each of which was the set of items. Denote support of an item set S ⊆ I as supp D(S) and frequency by freq D(S). Recall, freq D(S) =supp D(S)/|D|. For each item i, supp D (i) and freq D(i) denote respectively individual support and frequency of i. Function supp(D) projected over items, was also called *item support table* . Well-known frequent pattern mining problem: given the TDB D and the support threshold σ, find all item sets whose support in D was at least σ. In this paper, confine to study of the (corporate) privacy preserving outsourcing framework for frequent pattern mining.

International Journal On Engineering Technology and Sciences – IJETS™
ISSN(P): 2349-3968, ISSN (O): 2349-3976
Volume IV, Issue V, May - 2017

B. PRIVACY MODEL

Let D denote original TDB that owner has. To protect identification of individual items, owner applies an encryption function to D and transforms it to D∗, encrypted database. Refer to items in D as *plain items* and items in D∗ as *cipher items.*

Term item shall mean plain item by default. notions of plain item sets, plain transactions, plain patterns, and their cipher counterparts are defined in obvious way. Use I to denote set of plain items and E to refer to set of cipher items.

C. ODARM MODEL

Propose an ODARM algorithm for the secure computation of the union of private subsets. The proposed algorithm improves upon that in terms of simplicity and efficiency as well as privacy. In particular, our algorithm does not depend on commutative encryption and oblivious transfer (what simplifies it significantly and contributes towards much reduced communication and computational costs).The algorithm that proposed here computes a parameterized family of functions, which is called threshold functions, in which the two extreme cases correspond to the problems of computing the union and intersection of private subsets. Those are in fact general-purpose algorithm that can be used in other contexts as well. The ODARM also gives a high security and accuracy. For privacy here used general cryptographic functions. The performance of ODARM algorithms is more efficient for various reasons. It requires n number of database scans to generate a frequent n-item set. Furthermore, it recognizes transactions in the data set with identical item sets if that data set is not loaded into the main memory. Therefore, it unnecessarily occupies resources for repeatedly generating item sets from such identical transactions.

For example, if a data set has 10 identical transactions, the Apriori algorithm not only enumerates the same candidate item sets 10 times but also updates the support counts for those candidate item sets 10 times for each iteration. Moreover, directly loading a raw data set into the main memory won't find a significant number of identical transactions because each transaction of a raw data set contains both frequent and infrequent items.

To overcome these problems, don't generate candidate support counts from the raw data set after the first pass. This technique not only reduces the average transaction length but also reduces the data set size significantly, so it can accumulate more transactions in the main memory.

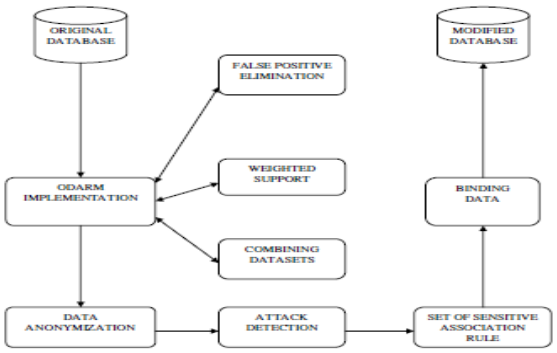

Figure 4-system architecture

Let D denote original TDB that owner has. To protect identification of individual items, owner applies an encryption function to D and transforms it to D∗, encrypted database. Refer to items in D as plain items and items in D∗ as *cipher items*. Term item shall mean plain item by default. Notions of plain item sets, plain transactions, plain patterns, and their cipher counterparts are defined in obvious way. Use I to denote set of plain items and E to refer to set of cipher items.

D. DATA ANONYMIZATION

The data will be made anonymised so that no unauthorized user can access the data. The data in the data set will be stored so that it will be associated only with the relevant data and relevant user. Data anonymization is a type of information sanitization whose intent is privacy protection. It is the process of removing personally identifiable information from data sets, so that the people whom the data describe remain anonymous.

Table 1-original table

Pid	*PName*	*Blood Group*	*Disease*	*Email*	*Mobile*	*City*	*DOB*	*Age*	*Address*	*Gender*	*Zip Code*
100	adm	B	cold	adm@gmail.com	8987656673	Ede	10.10.2010	30	Admire	Male	678678
101	new	B	Fever	new@gmail.com	9898989898	Ede	19.12.1992	26	Adm	Male	676676
102	Naveen	O	HeadAche	naveen@gmail.com	9087889908	Erode	06.06.1994	45	Admire	Male	675432
103	Java	J	Virus	java@gmail.com	8798765456	Erode	10.04.1989	40	Admire	Male	546789
104	sample	o positive	cold	sample@gmail.com	9879879879	erode	24.4.1986	23	Admire	Male	678678
105	sindhu	o	fever	sindhu@gmail.com	9876543210	erode	12/02/1993	23	tcs	Female	234567
106	sdhu	o	cold	sdhu@gmail.com	987654321	erode	09/08/1992	24	vibro	Female	680002

Generalization and perturbation are the two popular anonymization approaches for relational data Cipher Frequent Pattern based attack simulation.

The Privacy Technology Focus Group defines it as "technology that converts clear text data into a nonhuman readable and irreversible form, including preimage resistant hashes (e.g., one-way hashes) and encryption techniques in which the decryption key has been discarded."

Table 2-anonymized table

Age	Sex	Zip Code	Disease	Company	Bank
1-30	*	678678	cold	Admare	Barodа
1-30	*	234567	fever	tcs	cub
1-30	*	680002	cold	vibeo	iob
1-30	*	676676	Fever	Adm	icici
1-30	*	678678	cold	Admare	sbi
31-60	*	546789	Virus	Admare	SBI
31-60	*	675432	HeadAche	Admare	Axis

Data anonymization enables the transfer of information across a boundary, such as between two departments within an agency or between two agencies, while reducing the risk of unintended disclosure, and in certain environments in a manner that enables evaluation and analytics post-anonymization.

In the context of medical data, anonymized data refers to data from which the patient cannot be identified by the recipient of the information. The name, address, and full post code must be removed together with any other information which, in conjunction with other data held by or disclosed to the recipient, could identify the patient.[2] De-anonymization is the reverse process in which anonymous data is cross-referenced with other data sources to re-identify the anonymous data source.[3]

E. ATTACK MODEL

The server or an intruder who gains access to it may possess some background knowledge using which they can on encrypted database D∗. Generically refer to these agents as an *attacker*. Adopt the conservative model and assume that attacker knows exactly set of (plain) items I in original transaction database D and their true supports.

Assume service provider (who can be an attacker)is semi-honest in sense that although he does not know details of our encryption algorithm, he can be curious and thus can use his background knowledge to make inferences on encrypted transactions. Assume that attacker always returns (encrypted) item sets together with their exact support. Data owner (i.e., corporate) considers true identity of:

(1)Every cipher item,
(2)Every cipher transaction, and
(3)Every cipher frequent pattern as intellectual property which should be protected. Consider following attack Model.

• Item-based attack:

Semi honest service provider can attack owners data depend upon single item identity.

• **Set**-based attack:

Service provider attack owners data depend upon many item identities. In these method attacker can easily attacks data correctly but they can't use that data because that data's are in cipher text form.

IV. CONCLUSION

The proposed protocol for secure mining of association rules in horizontally distributed databases that improves significantly upon current leading protocol in terms of privacy and efficiency. One of main ingredients in our proposed protocol was the novel secure multi-party protocol for computing union (or intersection) of private subsets that each of interacting players holds. Another ingredient was the protocol that test inclusion of element held by one player in the subset to another. Those protocols exploit fact that underlying problem was of interest only when number of players was greater than two. One research problem that these study suggests. Namely, to devise an efficient protocol for inequality verifications that uses existence of the semi honest third party. Such the protocol might enable to further improve upon communication and computational costs of second and third stages of protocol. Other research problems that these study suggests was implementation of techniques presented here to problem of distributed association rule mining in vertical setting problem of mining generalized association rules, and problem of subgroup discovery in horizontally partitioned data.

REFERENCES

[1] 'Privacy-Preserving Outsourced Association Rule Mining on Vertically Partitioned Databases', Lichun Li, Rongxing Lu,Kim-Kwang, Raymond Choo, Anwitaman Datta, and Jun Shao ,2016.
[2] 'Towards semantically secure outsourcing of association rule mining on categorical data', Lai J,Li Y, Deng R.H, Weng J, Guan C, and Yan Q, Information Sciences, vol. 267, pp. 267–286,2014.
[3] 'A fast secure dot product protocol with application to privacy preserving association rule mining', Dong C and Chen L, in Advances in Knowledge Discovery and Data Mining - 18th Pacific-Asia Conference, PAKDD 2014, Tainan, Taiwan, May13-16. Proceedings, Part I, 2014, pp. 606–617.[Online]. Available: http://dx.doi.org/10.1007/978-3-319-06608-050,2016.
[4] 'Privacy preserving association rule mining over distributed databases using genetic algorithm', Keshavamurthy B. N, Khan A. M, and Toshniwal DNeural Computing and Applications, pp. 1–14,2013.
[5] 'Privacy-preserving mining of association rules from outsourced transaction databases', Giannotti F,Lakshmanan L,Monreale A,Pedreschi D and Wang H, IEEE Systems Journal, vol. 7, no. 3, pp. 385–395,2013.
[6] 'Result integrity verification of outsourced frequent item set mining', Dong B, Liu R, and Wang H.W, in Data and Applications Security and Privacy XXVII - 27th Annual IFIP WG 11.3 Conference, DBSec 2013, Newark, NJ, USA, July 15-17. Proceedings,2013,pp.258265.[Online].Available:http://dx.doi.org/10.1007/978-3-642-39256-617,2013.
[7] 'Secure two-party association rule mining', Kaosar M.G, Paulet R, and Yi X, in ACSW-AISC, 2011.
[8] 'On the (in) security and (im) practicality of outsourcing precise association rule mining', Molloy I, Li N, and Li T, in ICDM,2009.
[9] 'Privacy-preserving algorithms for distributed mining of frequent item sets', Zhong S, Information Sciences, vol. 177, no. 2, pp. 490–503,2007.
[10] 'Security in outsourcing of association rule mining', Wong W. K, Cheung, Hung E, Kao B, and Mamoulis N , in VLDB,2007.