Volume: 3 Issue: II Month of publication: February 2015

DOI:

International Journal for Research in Applied Science & Engineering Technology (IJRASET)

Privacy Preserving in Association Rule Mining

Sneha V. Dhande[1], Priyanka K. Patil[2]
Computer Science Department, H.VPM's COET Amravati

Abstract **— Data mining is the computer oriented process to evaluate vast database and extract the meaning of the data. Privacy is inter-related with secrecy and anonymity. Thus privacy preserving in data mining means maintaining secrecy which is concerned with the information that the others can retrieve from us. Then the loss of privacy means leakage of that information. So to mine the information from huge database number of techniques have developed that assures mining with privacy preserving. These technique are used to secure the sensitive items while extracting the appropriate knowledge from database. the methodology used to preserve the privacy in data mining using association rule is used because association rule mining is one of the important aspect in data mining. In this technique, secure multiparty computation is used which assures security using cryptography is also discussed. This technique ensures better privacy preserving with high efficiency.**

Keywords **— Data Mining, Elliptic Curve Cryptography, Uniform Randomization Privacy, Privacy Preserving Association Rule Mining, Secure Multiparty Computation (SMC)**

I. INTRODUCTION

The process of extracting significant information from the very large amount of database is called data mining. In many business organizations, data mining has emerged as one of the key feature. So the privacy has become an important issue in data mining. Due to the increased demand for knowledge discovery in all industrial domains, it is necessary to store all the raw data and to provide useful patterns with respective to the user needs. Generally, the storage of all raw data will be done in a database maintained by concerned organizations. Data mining techniques are available to retrieve useful information from large database. Prediction and description are the two fundamental goals of data mining. To full fill these goals many data mining techniques exists such as association rules, classification, clustering and so on.

Among these, association rule has wide applications to discover interesting relationship among attributes in large databases. Association rule mining greatly helps in protecting the secrecy and confidentiality of each database. From a general point of view, privacy information may be transmitted and illegally used. These problems and issues can be divided into two separate categories: one is data hiding and the second one is knowledge hiding. Data hiding tries to eliminate secret, confidential, private information from the huge data before its exposure . Association rule mining is used to find the rules which satisfy the user specified minimum support and minimum confidence. In the process of finding association rules, the set of frequent item sets are computed as the first step and then association rules are generated based on these frequent item sets.

II. BACKGROUND ON PRIVACY-AWARE DATA MINING

Privacy protection is a basic right, stated of the universal declaration of human rights. it is also an important concern in today's digital world. data security and privacy are two concepts that are often used in conjunction; however, they represent two different facets of data protection and various techniques have been developed for them privacy is not just a goal or service like security, but it is the people's expectation to reach a protected and controllable situation, possibly without having to actively look for it by themselves. therefore, privacy is defined as "the rights of individuals to determine for themselves when, how, and what information about them is used for different purposes" in information technology, the protection of sensitive data is a crucial issue, which has attracted many researchers. in knowledge discovery, efforts at guaranteeing privacy when mining and sharing personal data have led to developing privacy preserving data mining (ppdm) techniques. ppdm have become increasingly popular because they allow publishing and sharing sensitive data for secondary analysis. different ppdm methods and models (measures) have been proposed to trade of the utility of the resulting data/models for protecting individual privacy against different kinds of privacy attacks.

III. UNIFORM RANDOMIZATION

A simple approach for randomizing transactions would be to generalize Warner's \randomized response" method, Before transfer a transaction to the server, the client takes each item and with probability replaces it by a new item not originally present in this transaction. Let us call this process uniform randomization. estimate true (nonrandomized) support of an item set is nontrivial even for uniform randomization. Randomized support of, say, a 3-itemset depends not only on its true support, but also on the supports of its subsets. certainly, it is much more likely that only one or two of the items are inserted by chance than all three. So, almost all \false" occurrences of the item set are due to (and depend on) high subset supports. This requires estimating the supports of all subsets consecutively. For large values of p, most of the items in most randomized transactions will be \false", so we seem to have obtained a logical privacy security. Also, if there are enough clients and transactions, then frequent item sets will still be \visible", though less frequent than originally. For instance, after uniform randomization with p = 80%, an item set of 3 items that originally occurred in 1% transactions will occur in about 1% _ (0:2)3 = 0:008% transactions, which is about 80 transactions per each million. The opposite effect of \false" item sets becoming more frequent is comparatively inconsequential if there are many possible items: for 10,000 items, the probability that, say, 10 randomly inserted items contain a given 3-itemset is less than 10□7%. Unfortunately, this randomization has a problem. If we know that our 3-itemset escapes randomization in 80 per million transactions, and that it is unlikely to occur even once because of randomization, then every time we see it in a randomized transaction we know with near certainty of its presence in the nonrandomized transaction. With even more certainty we will know that at least one item from this item set is \true": as we have mentioned, a chance insertion of only one or two of the items is much more likely than of all three. In this case we can say that a privacy breach has occurred. while privacy is preserved on average, personal information leak through uniform randomization for some fraction of transactions, despite the high value of p.

IV. ASSOCIATION RULE MINING

Association Rule Mining is a popular technique in data mining for discovering interesting relations between items in large databases. It is purposeful to identify strong rules discovered in the databases using different available measures. Described association rules for discovering similarities between products in large-scale transaction data in supermarkets. For example, the rule {Bread, Butter} =>{Milk} found in the sales data of a shop would indicate that if a customer buys bread and butter together, he or she is likely to also buy milk. Such information can be used in decision making about marketing policies such as, e.g., product offers, product sales and discount schemes. In addition to the above mentioned example association rules are used today in many application areas including Web usage mining, Intrusion detection, As opposed to sequence mining, association rule learning typically does not consider the order of items either within a transaction or across transactions.

The problem of association rule mining [3] is defined as: Let I= {i1, i2,…, in} be a set of n binary attributes called *items*. Let D={t1,t2,…,tm} be a set of transactions called the *database*. Each transaction in database D has a unique transaction identity ID and contains a subset of the items in I [3]. A *rule* is defined as an implication of the form X=>Y where X,Y is subset of I and X intersection Y = Null Set. The sets of items (for short *item sets*) X and Y are called *antecedent* (left-hand-side or LHS) and *consequent* (right-hand-side or RHS) of the rule respectively.

Support count: The support count[3] of an item set X, denoted by X. count, in a data set T is the number of transactions in T that contain X. Assume T has n transactions. Then

$$support = \frac{(X \cup Y).count}{n}$$

$$confidence = \frac{(X \cup Y).count}{X.count}$$

Association Rules are helpful in many fields like Telecommunication and Medical records for retrieving some desired results. Association rules has been used in mining web server log files to discover the patterns that accesses different resources continuously or accessing particular resource at regular interval. Association rules are also useful in mining census data, text document, health insurance and catalog design [4].

The most famous application of association rules is its use for Market Basket Analysis. Consider a supermarket setting where the database records items purchased by a customer at a single time as a transaction. The planning department may be interested in finding “associations" between sets of items with some minimum specified confidence. Such associations might be helpful in designing promotions and discounts or shelf organization and store layout. Privacy preserving association rule mining technique commonly can be divided into two categories

A. Heuristic-Based Techniques
B. Cryptography-Based Techniques

These approaches can be further divided into two groups based on data modification techniques: data distortion techniques and data blocking techniques.

1) *Data distortion techniques :* Data distortion techniques try to hide association rules by decreasing or increasing support (or confidence). To increase or decrease support (or confidence), they replace 0’s by 1’s or vice versa in selected transactions. So they can be used to address the complexity issue. But they produce undesirable side effects in the new database, which lead them to suboptimal solution. M.Attallah et al. [1] were the first proposed heuristic algorithms. The proof of NP-hardness of optimal sanitization is also given in [1]. Verykios et al. [2] proposed five assumptions which are used to hide sensitive knowledge in database by reducing support or confidence of sensitive rules. Y-H Wu et al. [5] proposed method to reduce the side effects in sanitized database, which are produced by other approaches [2]. K. Duraiswamy et al. [6] proposed an efficient clustering based approach to reduce the time complexity of the hiding process.
2) *Data blocking techniques:* Data blocking techniques replace the 0’s and 1’s by unknowns (“?”) in selected transaction instead of inserting or deleting items. So it is difficult for an adversary to know the value behind “?”. Y.Saygin et al. [7][8] were the first to introduce blocking based technique for sensitive rule hiding. The safety margin is also introduced in [7] to show how much below the minimum threshold new support and confidence of a sensitive rule should. Wang and Jafari [9] proposed more efficient approaches than other approaches presented in [7][8].

The cryptography approach is very popular for the following two reasons:

i. It has a well established and well defined model meant for privacy which can actually provide good number of methodologies for verifying and validating purpose.
ii. Cryptography branch has a wide variety of tool set to incorporate privacy in data mining.

Cryptography-based approaches have been proposed in the context of privacy preserving data mining technique. Cryptography-based approaches like SMC are secure at the end of the computations. No party knows anything except its own input and the results.

V. RELATED WORK

Frequent item sets are detected using priory technique. For finding global support and confidence without privacy leakage secure computation is used. For satisfying result, knowledge hiding technique has been improved. To understand the background of privacy preserving in association rule mining, we present different techniques and in the following subsections

A. Secure Multiparty Computation (Smc) With Trusted Third Party

This technique worked as a client server system where one site is a server responsible for the generating global result and all remaining sites are client sites which sends its encrypted data to the server to retrieve global result. This technique has mainly two servers: Data Mining Server and Cryptosystem Management Server .

A disadvantage of this technique was that the failure of third party fails the communication.

B. Secure Multiparty Computation (Smc) With Semi Honest Model

This technique assumes all the sites as honest. One site acts as an initiator and all others as sites. All the sites send their encrypted data to the next site in queue. Finally the last site sends all data to initiator which finds the global result.

An example to SMC with semi honest model was Fast Private Association Rule Mining for Securely Sharing technique. The detailed description is mentioned in.The limitation of this a technique was the increase in computation time with the increase in the number of sites.

VI. PROPOSED NEW TECHNIQUE

A. Basic Concepts Of New Technique

Suppose database D is distributed among n sites (S1,S2,..,Sn) in such a way that database Di containing site Si consists of same set of attributes but different number of transactions. All sites are considered as semi honest. Now the problem is to mine valid global association rules satisfying given minimum support threshold (MST) and minimum confidence threshold (MCT) in unsecured environment, which should fulfill following privacy and security issues.

1) No any involving party should be able to know the contents of the transaction of any other involving parties.
2) Adversaries should not be able to affect the privacy and security of the information of involving parties by reading communication channel between involving parties.

B. Elliptic Curve Cryptography

Elliptic curve provides public cryptosystem based on the discrete algorithm problem over integer modulo a prime. Elliptic curve cryptosystem requires much shorter key length to provide a security level with larger key length. Elliptical curve cryptography is a method of encoding data files so that only specific individuals can decode them. ECC is based on the mathematics of elliptic curves and uses the location of points on an elliptic curve to encrypt and decrypt information.

The main goals of any privacy preserving association rule mining technique should insist on following factors : An technique using association rule mining for privacy preserving should prevent the finding of sensible information; The technique should not restrict the use and access of non sensitive data items/information; The technique should not have large computational complexity; It should be challenging to the various data mining techniques.; The technique should be equally efficient for very large database. This is very important factor; all mentioned technique does not satisfy all the goals only some of them satisfy these goals. Considering above mentioned goals, the technique can be evaluated using

Efficiency: The efficiency of technique is measured with its ability to execute with good performance using all required resources

Scalability: The technique should work with good performance even when storage requirement is very large along with communication costs of the distributed system when data sizes are increased.

Data quality: If the data quality is not relevant, the knowledge extraction is of no use.

VII. CONCLUSIONS

For evaluating privacy preserving association rule mining technique is proposed. To avoid the data leakage while sharing the data secure multiparty computation technique is used. To get good quality result some modification is also suggested. Drawback is that secure computation will cause high communication cost for huge database.

The proposed evaluation methodologies can be applied in new set of privacy preservation like Elliptic curve cryptography-technique.

VIII. ACKNOWLEDGMENT

We take this opportunity to express our gratitude and indebtedness to our guide **Prof. P.D.Kaware** as well as H.O.D, **Prof. A. B. Raut**, Computer Science and Engineering Department, they are a constant source of guidance and inspiration in preparing this work. Their constant help and encouragement helped us to complete our paper. We are grateful to Principal **Dr. A.B. Marathe,** for his encouragement and support. We are also thankful to all the staff members of Computer Science and Engineering department, whose suggestions helped us to complete the paper work and those who have directly and indirectly helped for completion of the paper.

REFERENCES

[1] M. Atallah, A. Elmagarmid, M. Ibrahim, E. Bertino, and V. Verykios, “Disclosure limitation of sensitive rules,” in Proceedings of the 1999 Workshop on Knowledge and Data Engineering Exchange, ser. KDEX "99. Washington, DC, USA: IEEE Computer Society, 1999, pp. 45–52
[2] V.S. Verykios, A.K. Elmagarmid, E. Bertino, Y. Saygin, and E. Dasseni, “Association rule hiding,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 16(4), pp. 434–447, April 2004.
[3] Rakesh Agrawal and Ramakrishnan Srikant, “Privacy-preserving data mining, In Proceedings of the ACM SIGMOD Conference on Management of Data

(2000)", 439–450.

[4] Vassilios S. Verykios, Elisa Bertino, Igor Nai Fovino Loredana Parasiliti Provenza, Yucel Saygin, Yannis Theodoridis: "State-of-the-art in Privacy Preserving Data Mining", March 2004.

[5] Y. H. Wu, C.M. Chiang and A.L.P. Chen, "Hiding Sensitive Association Rules with Limited Side Effects," *IEEE Transactions on Knowledge and Data Engineering*, vol.19(1), pp. 29–42, Jan. 2007.

[6] K. Duraiswamy, and D. Manjula, "Advanced Approach in Sensitive Rule Hiding" *Modern Applied Science*, vol. 3(2), Feb. 2009.

[7] Y.Saygin, V. S. Verykios, and C. Clifton, "Using Unknowns to PreventDiscovery of Association Rules," *ACM SIGMOD*, vol.30(4), pp. 4554, Dec. 2001

[8] Y. Saygin, V. S. Verykios, and A. K. Elmagarmid, "Privacy preserving association rule mining," *In Proc. Int'l Workshop on Research Issues in Data Engineering (RIDE 2002)*, 2002,pp. 151–163.

[9] S.L.Wang and A. Jafari, "Using unknowns for hiding sensitive predictive association rules," *In Proc. IEEE Int'l Conf. Information Reuse and Integration (IRI 2005)*, pp. 223–228, Aug. 2005.