# A SURVEY PAPER ON SECURE MINING FOR VERTICAL DISTRIBUTED DATABASE

[1]Jyoti Srivastava
[1] Lecturer,
[1]Dept. of Computer Engineering,

[2] Shefali Rana
[2] Lecturer,
[2] Dept of Computer Engineering,
Parul University, Vadodara, India

[3]Jigisha Patel
[3]Lecturer
[3] Dept of Computer Engineering

***Abstract*** *- Data mining is a process of extracting the large database, these databases are mostly scattered among various websites where security is required. These databases are divided into various categories such as horizontally distributed databases, vertically distributed databases and hybrid distributed databases. Privacy concept occurs when the data is distributed in environment and association rule also depend on the distribution of data in environment. Privacy preserving data mining (PPDM) is most important in data security field and it has become serious concern in the secure transformation of personal data. Association rule mining (ARM) algorithms are used to discover important knowledge from databases. There are Some problem in secure mining when transactions are vertically distributed. Each site holds some attributes of each transaction, and sites wish to collaborate. Various algorithms have been designed for it. In this paper, summarize them and survey of current algorithms, and analyse the mining methods for vertical distributed database.*

***Keywords –*** *Distributed database, Privacy Preserving Data Mining, Association Rule Mining*

## I. INTRODUCTION

Data mining technology provides the number of advantages using automated tools to analyse corporate, research and development, biological, Financial data, retail industry, telecommunication industry, and other scientific applications can help to find way to increase efficiency of organization, industry, or in medical applications. Privacy preserving data mining [1, 2], is a novel research direction in data mining, where data mining algorithms are analysed for the side-effects they incur in data privacy. The problem of mining the vertical distributed databases with the algorithm based on association rules with security has been studied here for the several sites which are holding the heterogeneous databases. The goal is to find association rules based on support and confidence. An association rule has two parts, an antecedent (if)and a consequent (then). An antecedent is an item found in data. A consequence is item that found in combination with the antecedent. Association rules are created by analysing the data for frequent patterns and using the criteria support and confidence to identify the most important relationships. Association rule mining is a process of mining the databases based on rules. This association rule is purely based on support and confidence. The concept of privacy preserving data mining involves in preserving personal information from data mining algorithms.

PPDM technique [3] is a research area in data mining and statistical databases where mining algorithms are analysed for the side effect they acquire in data privacy. The objective of privacy preserving in data mining is to build algorithms for transforming the original information in secured/unsecured way, so that the private data and private knowledge remains confidential even after the mining process [4].

Here some problem in secure mining of association rules in vertically distributed databases. In such a setting, there are several databases where transactional attributes are distributed across the databases. With vertical approach, some of the columns of a relation are apportioned into a base relation at one of the databases, and other columns are assigned into a base relation at another database. The relations at each of the sites must share a common domain so the original table can be reconstructed. In "Market-Basket" example, one database may contain grocery items and other one have clothing purchases. Using a key such as transaction date, transaction id or credit card details, we can join these to identify relationships between purchases of clothing and groceries. Horizontal partitions support an organizational design in which functions are repeated, often on a regional basis, whereas vertical partitions are typically applied across organizational functions with reasonably separate data requirements.

## II. ASSOCIATION RULE MINING STRATEGY

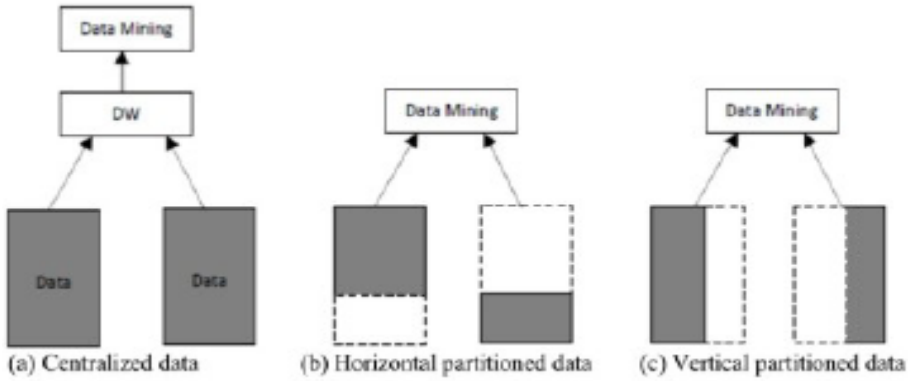
Association rule mining (ARM) is a technique in data mining that identifies the regularities in large volume of data. Such a technique may identify and reveal hidden information that is private for an individual or organization. Privacy preserving association rule mining needs to prevent disclosure not only the confidential or a personal information from the original or aggregated data also to prevent data mining techniques from discovering sensitive knowledge from large datasets.

The association rule mining is as follow [5]: Let I= $\{i_1,i_2,\ldots,i_m\}$ be the set of all items. Let D, the transactional data, be a set of database transactions where each transaction T is set of items such that T⊆I. Each transaction is associated with identifier, called TID. Let X be a set of items. A transaction is said to contain X if and only if X⊆T. An association rule is an implication of the form X⇒Y, where X⊆T, Y⊆T, X∩Y=∅, the support S and confidence C of the rule X⇒Y are defined as: S = Count(X)/|D|, C = Count(X⇒Y)/ Count(X).An item set that contains k items is a k-item set. The support count of an item set is the number of transactions containing the item set. An item set is frequent if its support count is not less than the minimum support count. Rules with the support more than a minimum support threshold ($S_{min}$) and the confidence more than a minimum confidence threshold ($C_{min}$) are called strong.

Association rule mining is a two-step process: (1) Finding all frequent item sets; (2) Generating strong association rules from the frequent item sets. The purpose of privacy preserving is to discover accurate patterns without precise access to the original data. The algorithm of association rule mining is to mine the association rule based on the given minimal support and minimal confidence. Therefore, the most direct method to hide association rule is to reduce the support or confidence of the association rule below the minimal support of minimal confidence. With regard to association rule mining, the proposed methodology that is effective at hiding sensitive rules is implemented mainly by depressing the support and confidence. The existing tree algorithms, D_CONF1, D_CONF2 and D_SUPP, are simply introduced in [6], which are to hide the sensitive association rule all by reducing the support or confidence.

## III. DATABASE ENVINROMENT

In centralized environment, all the datasets are collected at central site (data warehouse) and then mining operation is performed, as shown in Fig (a), where in distributed environment, data may be distributed among different sites which are not allowed to send their data to find global mining result. There are two types of distributed data considered. One is horizontally partitioned data and another is vertically partitioned data. As shown in Fig. (b) and Fig. (c) Data are distributed among two sites which wish to find the global mining result. The horizontal partitioned data shown in Fig. (b) Where Fig. (c) Shows vertical partitioned data. In horizontal partitioned data, each site contains same set of attributes, but different number of transactions wherein vertical partitioned data each site contains different number of attributes but same number of transactions [7].




Different Database Environment

## IV. PRIVACY PRESERVING IN ASSOCIATION RULE MINING

Privacy preserving association rule mining needs to prevent disclosure not only the confidential or a personal information from the original data also to prevent data mining techniques from discovering sensitive knowledge. Association rule hiding is widely researched along two principal directions. The first includes approaches that aims at hiding specific association rules among those mined from the original database. The second includes approaches that hides specific frequent item sets from those frequent item set found by mining original database. By ensuring that the item sets that lead to the generation of a sensitive rule become insignificant in disclosed database, the data owner can be certain that his or her sensitive knowledge is adequately protected from untrusted third parties.

Advantages:

1) Association rule hiding methods support for the sensitive item is unchanged. Instead, only the position of the sensitive item set is changed.

2) It provide the use of a different technique for modifying the database transaction, to reduce the confidence of sensitive rules without making any change in the sensitive item.

Disadvantage:

1) One of the main disadvantages of these approaches is that the approach tries to hide every single rule from a given set of rules without checking if some of the rules could be pruned after modification of some transactions from the set of all transactions.
2) Association rule hiding approach hides only rules that have sensitive items either in the right side or in the left side.

Privacy preserving association rule mining into three categories as follow:

**1) Heuristic-based techniques:**

This method include perturbation, blocking, which is the replacement of an existing attribute value.The goal of Heuristic algorithm is to modify data for selected data sets and take into account the effectiveness of data security and privacy. The advantage of heuristic algorithms is that it can be used to solve problem, which needs exponentially increased computations. The algorithms have yet to be optimized.

A. Data Perturbation-Based Association Rule: The algorithms can be described as following one. The approach proposed for the modification of data based on data perturbation, and in particular the procedure was to change a selected set of 1-values to 0-values, so that support of sensitive rules is lowered in such a way that utility of the released database is kept to some maximum value.

B. Data Blocking-Based Association Rule: The approach of blocking is implemented by reducing degree of support and confidence of sensitive association rules. By replacing certain attributes of some data items with a question mark or a true value. In this regard, the minimum support and minimum confidence will be altered into a minimum support interval and a minimum confidence interval.

a. Replacement-Based Techniques: After original data is replaced the value of some data with unknown value, the support and confidence of sensitive association rules will not be able to determine, which may be arrange of arbitrary values.
b. Anonymity Techniques: The distribution reconstruction technique uses the Expectation Maximization method to maintain the level of information loss.

**2) Reconstruction-Based Techniques:**

These algorithms are implemented by perturbing the data first and then reconstructing the distributions at an aggregate level in order to perform the association rules mining. According to the different methods of reconstructing the distributions and data types such as numerical data, binary data and categorical data, the algorithm is not work the same.

**3) Cryptography-Based Techniques:**

The ways of cryptography are used to data encryption and decryption. Many Cryptography approaches have been proposed in context of privacy preserving data mining algorithms.Cryptography-based approaches like Secure Multi-party Computation (SMC) are secure at the end of the computations. No parties knows anything except its own input and the end results. The [8] presents four secure multiparty computations based on the methods that can support privacy preserving data mining. The described methods include the secure sum, the secure set union, the secure size of set intersection, and the scalar product. SMC is mainly uses in distributed environment. The purpose of SMC is that it is necessary to guarantee the correctness of the calculation, but also to protect their respective input and output data from leaking when two or more participants who are carrying out the cooperation calculation.

A. Vertically Partitioned Distributed Data: Using the idea of "secure sum" for the secure calculation of inter-site and calculating the sum of support degree every sub-itemsets which are distributed in different sites, that is, the global support degree of the item sets. And then determining whether the item set is the global frequent item set or not by comparing whether it is greater than the threshold of the support degree or not. The security of the scalar product protocol is based on the inability of either side to solve k equations in more than k unknowns. Some of the unknowns are randomly chosen, and can safely be assumed as private. A similar approach has been proposed by Ioannidis et al. who present an extremely efficient and sufficiently secure protocol for computing the dot-product of two vectors by using

linear algebraic techniques and demonstrate superior performance in terms of computational overhead, numerical stability, and security by using analytical as well as experimental results [9]. Another way for computing the support count utilizes the secure size of set intersection method described in [10]. If the transactions are vertically partitioned across the sites, this problem can be solved by generating and computing a set of independent linear equations [11]. The work in [12] develops a log-linear model approach for strictly vertically partitioned databases and a more general secure logistic regression for problems involving partially overlapping data bases with measurement error.

B. Horizontally Partitioned Distributed: The key idea is to find global frequent item sets, while ensure inter-site information will not be leaked. Compared with vertical distribution algorithm, the algorithm is relatively simple. It only calculates the secure sum of support degree inter-sites. Kantarcioglu and Clifton in [13] use a secure multi-party computation to model the horizontal partitioning of transactions across sites, and present algorithms that incorporate cryptographic techniques to minimize the shared information without incurring much overhead in the mining process. The paper in [14] proposes an efficient distributed algorithm FDM (Fast Distributed Mining of association rules) for mining association rules. Some interesting properties between local and global large item sets are observed, which leads to an effective technique for the reduction of candidate sets in the discovery of large item sets. Two powerful pruning techniques, local and global pruning, are proposed. The algorithm finds global frequent item sets through two steps. In first step, it finds the candidate sets by using the methods of exchange encryption. Each site encrypt the respective frequent item sets, and then transfers the results to the next site. It removes repeated sets during transmitting; the process lasted until all of the sites encrypt all item sets, and then each site uses its own key to decrypt the result, and finally gets a public itemsets. In the second step, the purpose is to find global frequent item sets which can satisfy the conditions. First of all, the first site calculates the difference between local support degree of item sets which is acquired in the first step and the threshold of minimum support degree. Secondly, it adds a random number *R*, transfers the results to the next site. The second site does the same work with the first site, and adds the value which comes from the first site and then transfers the results to the next site. This process will continue until all the sites are transferred. The final value is transferred back to the first site. Finally, it compares the results with *R* in the first site, if the value not less than *R*, which illustrates the item sets is the global frequent item sets. The paper in [15] extends the former work to quantify association rule mining which can support the continuous property (via discretization).

## V. LITERATURE SURVEY

Feng Zhang, Chunming Rong (2013) proposed Privacy preserving distributed association rules mining protocols have been developed for horizontally partitioned data scenarios with more than two participating parties. It depend on a secure multi-party summary and union computation, which cannot guarantee security while the number of participating parties is two. Literature survey

Tamir Tassa (2013) proposed two novel secure multi-party algorithms, one that computes 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. It is simpler and is significantly more efficient in terms of communication rounds, communication cost and computational cost. The techniques in this paper are not implemented for the vertical setting [16].

Gayatri K et al (2014) proposed Secure Mining of Association Rules in Horizontally Distributed Databases Using FDM and K&C Algorithm. In their work they used FDM algorithm and Unifi-KC algorithm. Also the efficiency, computational cost and communication cost were compared in that paper. Future work is to devise an efficient protocol for inequality verifications that uses the existence of semi-honest third party and another in the implementation of the techniques to the problem of distributed association rule mining in vertical setting.

Sonal Patil et al (2014) proposed Overview of secure mining of association rules in horizontally distributed databases. This paper explains the overview of the topic, secure mining of association rules in Horizontally Distributed Databases. Distributed database is a database in which the storage devices are not all attached by a distributed DBMS. It may be stored in multiple computers located in the same physical location over a network of interconnected computers. Association rule is a method to find the relation between variables in large databases these horizontally distributed databases are stored where the association rule is applied to provide secure mining.

Santhana Joyce, Nirmalrani et al (2015) proposed Protocol is based on Fast Distributed Mining (FDM) Algorithm, which is the current leading protocol, which overcomes the disadvantages of various other algorithms such as apriori, FP tree etc. This protocol improves simplicity and efficiency as well as privacy. AES (Advanced Encryption standard) algorithm is used to ensure

security and to encrypt and decrypt the data while inserting and retrieving. Thus, the proposedprotocol is built for the horizontally distributed databases based on various association rules which rely on FDM algorithm. This protocol is high in efficiency and performance and highly focusing on the security of the databases [17].

## VI. CONCLUSION

Privacy preserving data mining has become increasingly popular because it allows sharing of the privacy-sensitive data for analysis purposes. Privacy preserving is one of the major problem of data mining. The problem of association rule mining where operation is in distributed environments when database exits in different multiple resources presents in the environment, so it's very difficult to find the global relationship among the attributes in the environments. There is some problem in secure mining of association rules in vertically partitioned databases.

In future work, for finding association rules in vertically distributed database the Apriori algorithm using Matrix is used which will increase the efficiency. In local sites, it calculate local support counts that minimizes the number of candidate sets and exchange messages by local and global pruning using matrix. For security purpose the scalar product protocol is useful in vertical distributed database.

## REFRENCES

[1] Chris Clifton and Donald Marks, Security and privacy implications of data mining, In Proceedings of the ACM SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery (1996), 15–19.

[2] Daniel E. O'Leary, Knowledge Discovery as a Threat to Database Security, In Proceedings of the 1st International Conference on Knowledge Discovery and Databases (1991), 107–516.

[3] Evfimievski, A., R. Srikant, R. Agrawal and J. Gehrke. "Privacy preserving mining of association rules". Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, July 23-25, ACM Press, Edmonton, AB., Canada, pp. 1-12,2002.

[4] Pingshui WANG "Survey on Privacy Preserving Data Mining" International Journal of Digital Content Technology and its Applications Volume 4, Number 9, December 2010

[5] R. Agrawal, R. Srikant, "Fast algorithms for mining association rules," In: Proc. 20th Int'l Conf. Very Large Data Bases, 1994, pp. 487–499.

[6] Shaofei Wu, Hue Wang, "Research on the Privacy Preserving Algorithm of Association Rule Mining In Centralized Database," In:2008 International Symposiums on Information Processing, 2008,pp.131–134.

[7] 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

[8] Chris Clifton, Murat Kantarcioglou, XiadongLin, and Michaed Y.Zhu,"Tools for privacy preserving distributed data mining," SIGKDD Explorations 4 (2002), no. 2.

[9] Ioannidis, I.; Grama, A, Atallah, M., "A secure protocol for computingdot-products in clustered and distributed environments," ParallelProcessing, 2002. Proceedings. International Conference on 18-21 Aug.2002, pp.379–384.

[10] Chris Clifton, Murat Kantarcioglou, XiadongLin, and Michael Y. Zhu,"Tools for privacy preserving distributed data mining, "IGKDDExplorations 4 (2002), no. 2.

[11] Vaidya, J. & Clifton, C.W., "Privacy preserving association rule miningin vertically partitioned data," In Proceedings of the eighth ACMSIGKDD international conference on knowledge discovery and data mining, Edmonton, Canada, July 2002.

[12] ZongBo Shang; Hamerlinck, J.D., "Secure Logistic Regression of Horizontally and Vertically Partitioned Distributed Databases," Data Mining Workshops, ICDM Workshops 2007. Seventh IEEE International Conference on 28-31 Oct. 2007, pp.723–728.

[13] M. Kantarcioglu, C. Clifton, "Privacy-preserving distributed mining of association rules on horizontally partitioned data," The ACM SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery (DMKD'02). ACM SIGMOD'2002 [C]. Madison,Wisconsin, 2002, pp.24–31.

[14] David W. Cheung, Jiawei Han, Vincent T. Ng,Ada W. Fu, and YongjianFu, "A fast distributed algorithm for mining association rules," In Proceedings of the 1996 International Conference on Parallel and Distributed Information Systems (1996).

[15] Li Naiqian, Shen Junyi, "Cross-table association rule mining based on privacy preserving," pattern recognition and artificial intelligence,dec.2003, pp.418–422.

[16] Tassa, Tamir. "Secure mining of association rules in horizontally distributed databases." Knowledge and Data Engineering, IEEE Transactions on 26.4 (2014): 970-983.
[17] Privacy in Horizontally Distributed Databases Based on Association Rules, Santhana Joyce M, Nirmalrani V, IEEE, 2015