Genetic Algorithm for Privacy Preserving Data Publishing

V.Saranya[1], Dr.L.M.Nithya[2]

[1](*Department of Information Technology, SNS College of Technology, Coimbatore, India)*

[2](*Department of Information Technology, SNS College of Technology, Coimbatore, India)*

Abstract **—The collection of digital information and data sharing by governments, corporations and individuals has formed an environment that simplifies large-scale data mining and data analysis. Such information is stored in large-scale databases and easy access to these databases has resulted in a dramatic increase in the disclosure of private information about individuals. To keeping the privacy of individuals, we emphases on suggesting different anonymity algorithms for various data publishing scenarios and keep data utility at the same time. In this research work, it is proposed to implement novel method using Genetic Algorithm (GA) with Association rule mining. Association rule mining, which is a technique used to extract concealed data from great datasets. The rule is created by analyzing datasets based on calculation of support and confidence. Disclosure values above specified threshold are marked as delicate pattern. These changes may generate various non- restrictive patterns, known as ghost rules and also some patterns may be misplaced, called as lost rules. Genetic Algorithm (GA) is a population based meta heuristic search technique, popularly used to solve the optimization problems and also conquests over lost rules and ghost rule side effects. In the proposed framework, we also consider the collaborative data publishing problem for anonymizing horizontally partitioned data at multiple data providers.**

Keywords:**Privacy,Association rule mining,Sensitive Association rule hiding, Genetic Algorithm,Secure Multiparty Computation**.

I. INTRODUCTION

The developing of internet as a communication medium, there is an increasing need for dispensing data that contains personal information from large database. With the proliferation of information about individual's personal data available in the databases, data mining is considered as a threat to privacy of data. Extraction of hidden sensitive information from large databases with great potential to support corporations focus on the most important information in their data warehouses. The extracted knowledge, expressed as association rules, decision trees or clusters, permits locating patterns hidden in data but meant to facilitate decision making. This knowledge discovery process returns sensitive information about individuals and also reveals critical information about business, compromising free competition. Disclosures of confidential/personal information should be prevented in addition to knowledge considered sensitive in a given context. For Example in the Healthcare Database, A hospital has collected useful information about a group of patient records that would help medical researchers and would like to publish this data while preserving the privacy of the individuals involved. To share data among hospitals and other providers use of health information beyond direct patient attention with privacy protection. Through data mining, attacker can able to extract confidential and useful information of individual which do not want to disclose to public.

Privacy preserving data mining (PPDM) is widely used to address this issue. Several techniques for PPDM uses modified version of standard data mining algorithms, where the modifications usually using well known cryptographic techniques ensure the required privacy. In most cases, the constraints for PPDM are preserving accuracy of the data and the performance of the mining process while maintaining the privacy constraints. Generally, the techniques for PPDM are based on cryptography methods, knowledge discovery and data hiding. In general,

ISSN: 2278 – 7798

statistics-based and the crypto-based techniques are used to deal with Privacy preserving data mining. In the statistics-based technique, the data publisher's sanitize the data through perturbation before publishing. The statistics-based approach is that it powerfully handles large volume of datasets. In the crypto-based approach, data owners have to cooperatively implement specially designed data mining algorithms. Though these algorithms achieve verifiable privacy protection but they suffer from performance and scalability issues.

Usage of hiding association rules is introduced due to high performance which is generated from frequent item sets. Association rule hiding methodologies[1] aim at sanitizing the original database. Specific threshold is set at the disclosure risk rate and confidential data which is not disclosed to the public. The rule generation is based upon the calculation of support and confidence. All the non-sensitive rules that appear when mining the original database at pre-specified thresholds of confidence and support can be efficiently mined from the sanitized database at the same thresholds or higher. To avoid disclosure of sensitive information, algorithm for privacy preservation in association rule mining becomes a must. But this modification method can influence the original set of rules, that can be extracted from the original database, either by hiding rules which are not sensitive (lost rules), or by generating new rules in the mining of the reformed database, which were not maintained by the original database (ghost rules).In Current Research work, it is proposed to implement an Evolutionary Algorithm using Genetic Algorithm (GA) with Association rule mining. Genetic Algorithm is used to improve the PPDM techniques and to reduce the impact of lost rules and ghost rules side effects. GA is a meta search heuristic method that impersonates the genetics and natural selection and performs the fitness tests on new structures to select the best population.

In this paper, the method uses binary transactional dataset as an input and modifies the original dataset based on the concept of genetic algorithm. All the Sensitive rules become hide and minimum modification performed in original dataset. The data is transformed into Boolean format such as 1 and 0, 1 represents the availability of data items and 0 represents non-availability of data items. The modification process can impact the original set of rules, which can be extracted from the original database, either by hiding rules which are not sensitive or by generating new rules in the mining of the reformed database, which were not maintained by the original database. We have tried to minimize these effects by minimum and suitable modification of original dataset. Hence the Genetic Algorithm transforms the original database into sanitize database. At first, the confidence of Sensitive Association Rules (SAR) obtains from original dataset and compared with Minimum Confidence Threshold (MCT). If confidence of SAR is larger than or equal to MCT then the fitness of each transaction is calculated. Finally, the data leakage avoidance system provides a protective way, which will control the accessibility of the distributed data. It also presents algorithms for distributing objects to agents, in a way that improves our chances of identifying a leaker. we have proposed a transformation framework that allows us to systematically transform normal computations to secure multi-party computations.

II. RELATED WORK

Nowadays, PPDM has caught more consideration when outsourcing computational resources to service provider. Privacy preserving data mining uses various methods to modify unique dataset using data mining techniques. The problem of privacy preserving in Data Publishing was first addressed in [2].After this,researchers conduct so many techniques to solve theprivacy issue of mining results.Aggarwal et al.[3] proposed the process of modification method can be categorized into data blocking and data distortion techniques. The major concept of data distortion techniques, are the replacement of designated values with "false" values (i.e., replacing 1's by 0's and vice versa). He suggested that the process of changing original database in such a way that some restrictive patterns hide without really affecting the information and the non-restrictive patterns. Adding false values to actual transaction database which cause the problems of lost rules and ghost rules side effects.

Work of Weng et al. [4] proposed the blocking technique which is the replacement of an existing attribute value with "unknown" or "?".

ISSN: 2278 – 7798

In blocking technique, the algorithms do not add false value to the transaction database. In addition, to restore a value by an unknown value instead of placing a false value. By using blocking method to transform original database *D* into release database *D`* by increasing support of the rule antecedent by changing 0s to ?or by decreasing support of rule consequent by changing 1s to ?. Hence, comparing to other techniques, blocking base technique do not distort the database but only change some known values to unknown. The main drawback of this technique is the privacy violation of the transformed database. Verykios et al. [5] presented five algorithms. These algorithms run on the strategy which is based on decreasing the support and confidence of association rules. Specifically, these methods are not overcome all the side effects caused by preserving the privacy of association rules and also the time taken by each algorithm to hide a set of rules is also high. Atallah et al. showed that optimal sanitization is an NP-hard problem and need to standardization. In this research, they proposed a heuristic based on support reduction, to selectively hide some frequent itemsets from large databases with as little as possible impact on other, non-frequent itemsets.Clifton et al discussed the security issues and inference of data mining. He investigated the idea of limiting access to the database; supplementing data, remove needless combination and fuzzy data.

In the same direction, Chih-Chia et al.[6] proposed algorithm Fast Hiding Sensitive association Rules (FHSAR) which hides sensitive association rules successfully by establishing association between transaction and sensitive association rules. Normally, this technique assigns a weight W to each transaction. Weight demonstrates the dependency of transaction on restrictive patterns. The bottleneck of this algorithm is the number of lost rule and performance in term of W, which is computed again after each item modified.

Naeem et al.[7]projected five methods namely Confidence, All-Confidence, Conviction, Leverage and Lift in order to mine association rules from large databases. The weighting mechanisms uses in this approach are Sum, Mean, Median and Mode. This technique is only applicable on dataset whose attributes not more than 26 and also generate high side effect in term of lost rule. Similarly, Dehkordi et al.[8] suggested genetic algorithm in the domain of privacy preserving in association rules. This method divides the original database into safe transaction and critical transaction. Safe transactions are those which do not contain any sensitive item and no need to modify while critical transactions are those which contain sensitive items and need to sanitize. Furthermore, the side effect in term of lost rules and ghost not define clearly.

In recent paper which has formally proven that the encoding system can be broken without using context-specific information. The success of attacks mainly relies on the existence of unique, common and fake items. Tai et al. supposed the attacker knows the exact frequency of single items. They use privacy model which requires that each real item must have the same frequency count as k-1 other items in the outsourced dataset. Furthermore they do not offer any theoretical analysis of anonymity of itemsets. In current research work, genetic algorithm is used to triumph over ghost and lost rule side effects. Also, this technique can be applied for small as well as large dataset in domain of healthcare, military and business datasets.

III. PROPOSED SOLUTION

In the Proposed Solution we will explain most important work such as preprocessing, Association rule mining, the specification of our fitness function in Genetic Algorithm method and Secure Multiparty Computation Methods.

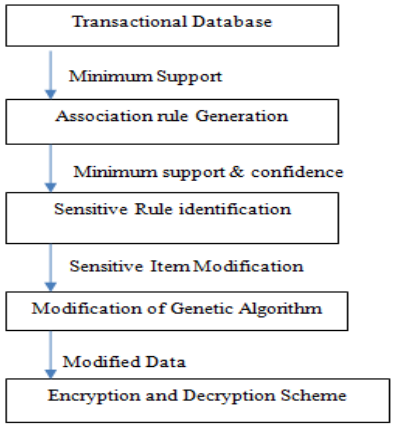

Figure 1:System Architecture

ISSN: 2278 – 7798

A. Preprocessing of Original Dataset

In this Phase, first we introduced the method is Dataset Pre-Sanitization Process (DPSP) which involves preprocess the original database. In preprocessing of dataset, that sensitive items are limited to some transactions, hence there is no need to modify all of the transactions. For the reason that, we select all the transactions that support sensitive items. With this critical phase, we can reach to better performance of sanitization speed and less number of modification needed in hiding process. Furthermore, by preprocessing of original dataset which shows that the size of each chromosome decreases significantly.

B. *Association Rule Mining*

Association Rule Mining is used to finding frequent patterns, associations, correlations, or causal structures among sets of items or objects in transaction databases, relational databases, and other information repositories. An Association rule is an implication of two item sets.(ie) X⇒Y. This rule is based on calculation of support and Confidence framework. Here, the support is a measure of the frequency of a rule, the confidence is a measure of the strength of the relation between sets of items. Association rule mining algorithms scan the database of transactions and calculate the support and confidence of the candidate rules to decide if they are significant or not. A rule is considerable if its support and confidence is greater than the user specified minimum support and minimum confidence threshold. In this way, algorithms do not recover all possible association rules that can be derivable from a dataset, but only a very small subset that satisfies the minimum support and minimum confidence requirements set by the users.

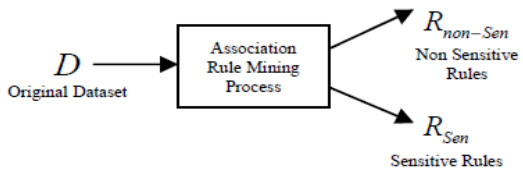

Figure 2: Input and outputs of Association Rule Mining Process

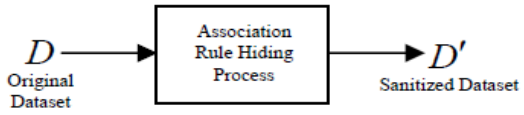

Figure 3: Association Rule Hiding Process

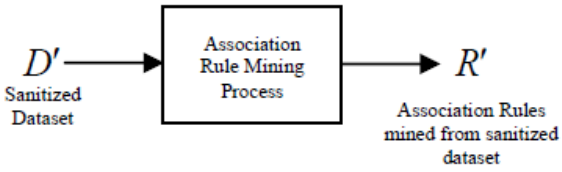

Figure 4: Association Rule Mining after Association Rule Hiding

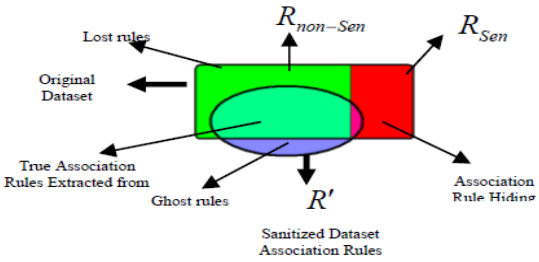

Figure 5: Side Effects of Sensitive Rule Hiding

In this Rule Mining, Reducing the support of the large itemset via eradicating items from transactions or adding fake item into the transactions are an NPhard problem. Hence, we are looking for a special modification of D (the source dataset) in D* (sanitized dataset) that exploits the number of rules in Rnon-sen (minimizing number of lost rules) that can still be mined. Therefore, we must hide the sensitive association rule, thus it is necessary to modify the dataset and in the other side we should keep the utility of modified dataset to extracting useful information and rules. Therefore we have designated the Evolutionary algorithm using Genetic algorithm approach to solving this optimization problem. Problem devising elements are depicted in Figures 2 to 4 and side effects of Sensitive rule hiding problem is shown in Figure 5.

C. Genetic Algorithm

In proposed solution, we use Genetic Algorithm (GA) for minimizing side effects. GAs is a family of development inspired computational models. Genetic algorithm (GA) is an Evolutionary and meta heuristic method that impersonates the genetics and natural selection. GA encodes a

ISSN: 2278 – 7798

potential solution for a specific problem on chromosome-like data structure applying recombination operators to these structures to preserve critical information. It develops solution to optimization problem using the techniques of inheritance, mutation, selection and cross over. This method converts the database recursively until the support or confidence of the preventive patterns drop below the user specified threshold and this technique is only applicable on binary dataset. Optimality of solution depends on the complexity of fitness function. The potential strength of fitness function ensures a desirable level of optimal solution.

In GA population encompasses the group or collection of individuals called chromosome that specify a complete solution to problem. The first set of population is intermittently generated set of individuals. Each transaction is denoted as a chromosome and occurrence of ith item is represented as 1 and non-occurrence as 0. Population contains many number of chromosomes and best is used to generate next population that is based on the survival fitness.GA is used to triumph over the lost rules and ghost rule side effects. The potential strength of fitness function ensures a predictable level of optimal solution. Hence GA is used to hide restrictive patterns, X->Y, by decreasing support of Y or by increasing the support of X. The fitness function assigns a value to each transaction in the database.

Calculation of Fitness Function

Initially fitness strategy depend on both hiding all sensitive rules and minimum number of modification in original dataset. This fitness function based on weighted sum function as follows:

Minimize: Cost_function_1=W1× Rules Hiding Distances + W2×Number of Modifications

Where,

- W1+W2=1 (It is the essential condition for weighted sum optimization problem and their values specified based on their costs).

- Rules Hiding Distances $=\sum_{i=1}^{number\ of\ Sensitive\ Rules} Rules^{\text{Hiding}}$
- Number of Modifications$=\sum_{j=1}^{|critical\ \ Transactions\ \ |\times|I|} D' \oplus D_{\text{j}}$

Selection

After calculation of fitness strategy, the next step is to select the chromosomes for reproduction in a population. Satisfied fitness chromosomes are selected for reproduction and lower fitness chromosomes may be selected a few or not at all. There are various selection methods, such as: "Roulette-Wheel" selection, "Rank" selection and "Tournament" selection. Tournament selection method which is used in this paper. In this, we select two chromosomes randomly from the population. For a predefined probability p, the more fitness of these two is selected and with the probability (1-p) the other chromosome with lower fitness is selected.

Crossover

After performing selection, crossover operation in Genetic algorithm is combination of two chromosomes composed to generating new offspring or child. Crossover arises only with some probability and chromosomes are not subjected to crossover remain unmodified. More fitness chromosomes have a prospect to be selected, so good solution always alive to the next generation. Single point crossover and multi-point are the most famous crossover operators. In this paper single-point crossover has been applied to make new offspring.

Mutation

Mutations are global searches and mutation probability is predetermined before starting the algorithm and applied to every individual bit of each offspring chromosome for determining if it is to be inverted. Mutation is used to keep the genetic diversity from one generation of a population of genetic algorithms chromosome to the next. After execution of crossover operation, the new developed generation will only have the character of the parents. This activity can lead to a problem where no new genetic material is presented in the offspring and finding better population has been stopped. Mutation operator permits new genetic patterns to be presented in the new chromosomes. Mutation introduces a new sequence of genes into a chromosome but there is no guarantee that mutation will produce necessary features in the new chromosome. The purpose of mutation in GA is preserving and introducing diversity.

ISSN: 2278 – 7798

D. Secure Multiparty Computation

Two Competing financial organizations might jointly invest in aproject that must satisfy both organizations' private and valuable constraints. To conduct computations, one entity must usually know the inputs fromall the participants; however if nobody can be trusted enoughto know all the inputs. Secure Multiparty Computation (SMC)is to enable parties to carry out such distributed computing tasks in a secure manner. SMC is used to create methods that enable parties to jointly compute a function over their inputs, while at the same time keeping these inputs private.Here, we have two organizations with separate input sets X and Y. The goal of the Organizations is to jointly compute the median of the union of their sets X U Y without revealing anythingabout each other's set that cannot be derived from the output itself. In order to acquirethis output, they run an interactive protocol which involves those sending messages to eachother according to some prescribed specification, which in turn should result in them learningthe output as desired.

IV. CONCLUSION

Organizations frequently distribute data in order to accomplish mutual profits. However, sharing of these data, most of the time discloseconfidential information of individuals. Hence, data preprocessing techniques are applied to preserve the confidentiality of their confidential data or preventive pattern in the form of sensitive association rules.In this research work we have used privacy preserving data mining technique to emphasis on inferences originating from the application of data mining algorithm to large public databases. We also investigated how sensitive rules should be protected from malicious data miner and proposed an algorithm to hide this sensitive rule, known as genetic algorithm. In genetic algorithm, a new fitness function is calculated, based on this value transactions are selected and sensitive items of this transactions are reformed with cross over and mutation operations without any loss of data. Finally all sensitive rules are hidden, no false values could be produced and non-sensitive values are unaffected. Secure Multiparty Computation enable parties to carry out such distributed computing tasks in a secure manner.

REFERENCES

[1] F.Giannotti,V.S.Lakshmanan,Anna Monreale, Dino Pedreschi and Hui Wang, "Privacy-Preserving Mining of Association Rules From Outsourced Transaction Databases",IEEE Systems Journal,Vol 7,No3,Sep 2013.

[2]M.Atallah, E.Bertino, A.Elmagarmid, M.Ibrahim and V.Verykios."Disclosure limitation of Sensitive Rules"Proc.of IEEE knowledge and Data Engineering Exchange Workshop, Nov 1990.

[3] C.C.Aggarwal,Yu "Privacy preserving Data mining :Models and Algorithm," Springer,2008.

[4] S.L.Wang,"Hiding Sensitive Predictive Association Rules", Systems, Man and Cybernetics,IEEE International Conference on Vol.1.pp.164-169,Oct 2005.

[5]V.SVerykios,A.Elmagarmid,E.Bertino,Y.Saygin,E .Dasseni,"Assoction Rule Hiding",IEEE Transactions on Knowledge and Data Engineering,Vol 16.pp.434-447,Apr.2004

[6]W.Chih-Chia, C.Shan-Tai, L.Hung-Che,"A Novel Algorithm for completely hiding Sensitive association Rules", IEEE 8th International Conference on Intelligent Systems Design and Applications, Vol.3, pp.202-208, Nov.2008.

[7] M.Naeem, S.Asghar,"A Novel Architecture for Hiding Sensitive Association Rules", In Proceedings of DMIN,pp 380-385,Jul.2010.

[8]M.N.Dehkordi,K.Badie,A.k.Zadeh,"A Novel Method for Privacy Preserving in Association Rule Mining Based on Genetic Algorithms", Journal of Software,Vol.4,pp.555-562,Aug.2009.

ISSN: 2278 – 7798