International Journal of Applied Engineering Research ISSN 0973-4562 Volume 12, Number 23 (2017) pp. 13917-13926

Association Rule Hiding for Privacy Preserving Data Mining : A Survey on Algorithmic Classifications

Gayathiri. P
Research Scholar, Department of Computer Science,
Research and Development Centre,Bharathiar University, Coimbatore - 641046.

Orcid: 0000-0001-9842-3455

Dr. B Poorna
Principal, Shri Shankarlal Sunsarabai Shasun Jain College for Women,
TNagar,Chennai - 600017, Tamil Nadu, India.

Abstract

The increased collection, storage and analysis of person-specific data cause serious challenges to the protection of the identities to which such data correspond. In various conditions, the extracted knowledge is highly confidential and it needs to be sanitized before published in order to address privacy concerns. Data mining technology is capable of extract huge amount of knowledge with minimal time period. The knowledge extracted by intelligent data mining algorithm may reveal most sensitive information belongs to a person or an organization. Data belongs to a person or an organization may have different sensitive levels. These data are made available only for authorized persons. So ensuring the protection of sensitive data by access restriction is not a complete method. This may affect the utility of the data mining result and with help of the knowledge the user may re-identify sensitive data items from non-sensitive data is known as 'Inference Problem'. The privacy preserving data mining is to provide a solution for protecting sensitive information by developing a data mining techniques which could be applied on databases without affecting the accuracy of data mining result. At the same time without violating the privacy of individuals. This paper states a detailed study on various algorithms for association rule hiding methods.

Keywords: data security; data mining; association rules ; privacy preserving data mining; sensitive items; association rule hiding.

INTRODUCTION

Association rule mining is a data mining technique was first introduced in 1993. The association rule mining has become one of the core data-mining tasks and has attracted tremendous interest among researchers and practitioners since its inception. The technique adapted for data mining in association rule mining is to identify the symmetry found in huge database. This technique may help to re-identify necessary information that is private to a person or an organization.

The medical research service may be interested in developing personalized medicine for genetically disorder diseases. The data needed for their research is derived from a wide variety of sources such as hospital, laboratory, pharmacy and general government statistics. These records includes personal information such as name, age, gender, passport number, DNA sequence, disease and expenditure data such as bank transaction, transfers and purchase. Exposing these private information intentionally or unintentionally belongs to an individual is against the law in most countries. In order to preserve privacy, the information can be de-identified before the information is shared. This can be accomplished by deleting unique identity fields, such as name and passport number from the dataset. DNA sequence is one of the most reliable person-specific data like finger print, retina and iris in human. Compromise of DNA privacy via re-identification, the implication of explicit identity of the individual from which the DNA was derived, is dependent on unique features that may be implication from a DNA sequence. Deleting this field from data mining is meaningless. To avoid such exposure of sensitive information, algorithm for privacy preservation in association rule mining becomes a must.

The research in this field is tackling the problem in different angle using diverse techniques. The techniques used in the research are based on hiding strategy, data modification strategy, rules hidden per iteration and algorithmic nature**.** Most of the methods result in affecting the data utility, violating the privacy and various side effects. The side effects include wrongly hiding non sensitive rules and falsely introduce forged rules. So the research papers are categories in such a way to evaluate the merits and demerits of various rule hiding techniques for privacy preserving.

This paper is organized to association rule mining strategies, inference control in various level of transactions in section 2; various goals of association rule hiding methodologies in section 3; various approaches association rule mining techniques in section 4; literature surveys of privacy

preserving association rule mining techniques in section 5; recent trends in association rule hiding algorithm in section 6; conclusion and future research directions in section 7.

ASSOCIATION RULE MINING STRATEGIES AND INFERENCE CONTROL

Association rule mining is one of the most important and well researched techniques of data mining, was first introduced in Agrawal et al. in 1993[1]. It aims to extract interesting correlations, frequent patterns, associations or casual structures among sets of items in the transaction databases or other data repositories with the help of association rules of the databases. These rules are an important class of symmetries within data which have been extensively studied by the data mining community. The problem formulation of association rules are stated as follows: Let $I=I_1, I_2, \ldots, I_m$ be a set of m distinct attributes, T be a transaction that contains a set of items such that $T \subseteq I$, D be a database with different transaction records Ts. An association rule is an implication in the form of $X \Rightarrow Y$, where $X \subset I$, $Y \subset I$ are sets of items called itemsets, and $X \cap Y = \emptyset$. X is called antecedent (Left hand side) while Y is called consequent (right hand side), the rule means X implies Y. The antecedent of a rule can consist either of a single item or of a whole set of items this applies for the consequent as well. There are two important basic measures for association rules, support and confidence. In general only those itemsets that fulfill a certain support requirement are taken into consideration. In association rule mining always a compromise has been made between discovering all item set and computation time.

The formula for computing support of the rule $X \Rightarrow Y$ is the percentage of transactions in T that contain $X \cup Y$. It determines how frequent the rule is applicable to the transaction set T.

$$\text{Support}(X) = \frac{\text{Support\_count}(X)}{n} \times 100$$

(Where n is number of transactions in dataset D)

The confidence of a rule describes the percentage of transactions containing X which also contain Y. The Confidence measure for rule $X \longrightarrow Y$ in dataset D is defined

$$\text{Confidence}(X \rightarrow Y) = \frac{\text{Support}(XY)}{\text{Support}(X)} \times 100$$

Support is a measure of the frequency of a rule; the confidence is a measure of the strength of the relation between itemsets. Association rule mining algorithms scan the transaction database and calculate the support and confidence of the candidate rules to determine if they are sensitive or not. A rule is sensitive if its support and confidence is higher than the user specified minimum support and minimum confidence threshold. The process of association rule mining is to identifying all the itemsets contained in the data that are adequate for mining association rules. These combinations are to exhibit certain frequency and are called frequent itemsets. Then generating all possible rules out of the discovered frequent itemsets with greater than minimum confidence is considered as sensitive.

Most of the sensitive data in a database are not available for public access but with the help of public interface is allowed to perform aggregate query. This lead to the violation of privacy, that an intelligent user may pose a sequence of queries through which he or she may infer sensitive facts about data [2, 3]. This type of inference is known as full discloser. The user may determine the exact values of the data attributes. Another type of inference is known as partial discloser; in this user may be able to narrow down the value to a range not the exact value of the data attribute. This inference problem is a natural threat in association rule mining.

GOAL OF ASSOCIATION RULE HIDING METHODOLOGIES

Association rule hiding methodologies aim at sanitizing the original database in order to achieve the following goals [4];

1. No rule that is considered as sensitive from the owner's perspective and can be mined from the original database at pre-specified thresholds of confidence and support; can be also revealed from the sanitized database, when this database is mined at the same or at higher thresholds.

2. All the non-sensitive rules that appear when mining the original database at pre-specified thresholds of confidence and support can be successfully mined from the sanitized database at the same thresholds or higher.

3. No rule that was not derived from the original database when the database was mined at pre-specified thresholds of confidence and support can be derived from its sanitized counterpart when it is mined at the same or at higher thresholds.

The first goal requires that all the sensitive rules disappear from the sanitized database, when the database is mined under the same or higher levels of support and confidence as the original database.

The second goal states that there should be no lost rules in the sanitized database. That is, all the non-sensitive rules that were mined from the original database should also be mined from its sanitized counterpart at the same or higher levels of confidence and support.

The third goal states that no false rules also known as ghost rules should be produced when the sanitized database is mined at the same or higher levels of confidence and support. A false (ghost) rule is an association rule that was not among the rules

mined from the original database.

A solution that achieves the first goal is called as feasible and it accomplished the hiding task. A solution that addresses all the three goals is called exact. Exact hiding solutions that cause the least possible modification to the original database are called ideal or optimal. Non-exact but feasible solutions are called approximate. All the three goals need to be attained by every association rule hiding algorithm. The first goal is try to hide as many sensitive rules as possible. The second and third goals are try to minimize the possible side effects while hiding the sensitive rules. Different hiding algorithms give different priorities to the satisfaction of these goals.

The privacy preserving data mining classify into two broad categories as data hiding and knowledge hiding [8]. The first method tries to remove or transform confidential information from the data before its disclosure. The second method is refining of confidential information from the knowledge. Privacy preserving association rule algorithm for data mining is adapted knowledge hiding method. At the same time, the hiding algorithm may hide either sensitive frequent itemsets or sensitive rules or both [9].

APPROACHES OF ASSOCIATION RULE HIDING METHODOLOGIES

Association rule hiding approaches is based on **four major dimensions and three classes**. The dimensions are *1 hiding strategy decides whether it is support based or confidence based or both, 2 data modification strategy tells whether data distortion or data blocking is used for data sanitization, 3 rules hidden per iteration decide whether it hide single rule or multiple rule per iteration, and 4 algorithmic nature tells the broad sense can be either heuristic or exact.* The three classes namely *heuristic approaches*, border-based approaches and exact approaches**.** The first class of approaches is efficient and fast that selectively sanitize a set of transaction from original database. The second class of approaches is achieved to hide the sensitive association rules by tracking the border of non-sensitive frequent item set. The third class of approaches contains non-heuristic which conceives the hiding process as a constraints satisfaction problem that is an optimization problem.

SURVEY ON PRIVACY PRESERVING ASSOCIATION RULE MINING (PPAM)

Let D be the original database, R be a set of significant association rules that can be mined from D, and let Rs be a set of sensitive rules in R. Privacy preserving association rule hiding algorithms sanitized database D into a database D'. All rules in R can still be mined from sanitized database D', except for the rules in Rs. In the sanitized database D', no rules (ghost) other than R should be mined and no rules (lost) from R-Rs should be hide. This paper survey on privacy preserving association rule mining based on the privacy preserving technologies and algorithmic nature used.

Heuristic based approaches: this approach determine how appropriate datasets for modification has been selected. The optimal selection of datasets for modification or sanitization is an NP-Hard problem; heuristics is used to address the complexity and various issues. The methods of Heuristic based modification include data distortion scheme, which is accomplished by the alteration of an attribute value by a new value (i.e., changing a 1-value to a 0- value, or adding noise) [10], swapping of values between transaction [5,6], deletion of specific items from the database[11]and blocking scheme, which is the replacement of an existing attribute value with an unknown (i.e. replace by a "?").

Atallah et al. [5] were the first to propose an algorithm for hiding association rule using heuristic hiding approach in support based technique, by decreasing the support of generating itemsets. This work contributed proof for NP-hardness of finding an optimal sanitization of dataset. This method not supported for large item set, it remains frequent in sanitizes database.

V. S. Verykios et al. [10] proposed five algorithms namely 1.a, 1.b, 2.a, 2.b and 2.c based on some assumption, they are hide only rules that are supported by disjoint large itemsets, hide association rules by decreasing either their support or their confidence, select to decrease either the support or the confidence based on the side effects on the information that is not sensitive, hide one rule at a time, decrease either the support or the confidence, one unit at a time. The first three algorithms are rule-oriented. They decrease either the confidence or the support of a set of sensitive rules, until the rules are hidden. This can happen either because the large itemsets that are associated with the rules are becoming small or because the rule confidence goes below the threshold. The last two algorithms are item set-oriented. They decrease the support of a set of large itemsets until it is below a user-specified threshold, so that no rules can be derived from the selected itemsets. The computation time and side effect i.e. lost rule and false rule are high.

Dasseni et al. [9] were proposed three single rule heuristic hiding approach, by decreasing either support or confidence based but not the both. These algorithms consider the hiding of both sensitive frequent itemsets and sensitive rule. Contribution of this work aims to hide all sensitive knowledge appearing in the dataset. The drawback of this scheme is the strong assumption that all the items appearing in a sensitive rule do not appear in any other sensitive rule. This assumption makes no difference in hiding of the rule one at a time or altogether and it fail to avoid undesired side effect of lost rule and false rule.

The work proposed in [9] was extended by Verykios et al [10]

by improving and evaluating the association rule hiding algorithms in different sizes of input datasets and different set of sensitive rules. The author proposed two algorithms based on heuristic approach, the first algorithm hide the item having maximum support with minimum length transaction. The second algorithm sorts the generated itemsets based on their size and support, and then hide in a round-robin method.

Multiple rules hiding approach is first introduced by Oliveria and Zaiane [11]. This algorithm requires two scans of the database. During the first scan, an index file is created for sensitive transactions it provide an efficient retrieval of the data. The algorithm sanitize the database by selectively removing the least amount of individual items those are represent the sensitive knowledge in second scan. During sanitization this algorithm takes into account not only hiding the sensitive patterns, also the issue related to the hiding of non-sensitive patterns. Three item restriction-based algorithms known as MinFIA, MaxFIA, and IGA are proposed to selectively remove items from transactions that support the sensitive rules. In MinFIA algorithm, identify item having smallest support in the pattern called the victim item for all supporting transaction for every sensitive pattern. With help of the user-supplied threshold it sorts the identified transaction in ascending order and then selects the number of transitions that need to be sanitized. Finally the algorithm removes the victim item from each transaction. In MaxFIA it works same as MinFIA but it selecting the victim item that has the maximum support in the sensitive association rule. In IGA algorithm it clusters the sensitive patterns which share the same itemsets. Based on the sensitive itemsets those are shared by the clusters of the algorithm hides the corresponding sensitive patterns at once. This reduces the amount of distortion needed for the database to hide the sensitive knowledge.

Oiveira and Zaiane [12] introduced a new algorithm known as SWA based on the work [11, 9, 13, 14]. The proposed algorithm SWA based on one scan heuristic it aims to provide a balance between privacy and utility in association rule hiding. It achieves to hide multiple rules in only one pass through the datasets, irrespective of its size and number of rules need to be hiding. SWA is an efficient algorithm for hiding all sensitive rules using heuristic approach, while maintaining high data utility.

Three effective, multiple association rule hiding algorithm that outperform SWA with higher data utility and lower data distortion with increased computational speed by A.Amiri[15]. The first algorithm compute the union of supporting transactions for all sensitive itemsets, from the transaction which supports most sensitive and least non sensitive is hidden from the database. This process is repeated until all sensitive itemsets are hidden. The second algorithm removes individual itemsets from transaction instead of removing the entire transaction. It computes the union of all transactions supporting sensitive itemsets. Calculate highest number of sensitive and non-sensitive itemsets affected by the item which is selected for removing from the transaction. The third algorithm combines both first and second algorithms to produce Hybrid algorithm. The first algorithm is used to identify sensitive transaction and second algorithm is used to selectively delete items from these transactions, repeat this process until all sensitive knowledge is hide.

Using set theory, Wu, et al [7] standardized a set of constrains related to possible side effect of the association rule hiding process and also enforced set of constrains to data sanitization. Relaxation of constrains enforced to hiding all sensitive itemsets without the user's approval is a drawback of this approach. A distortion based heuristic approach to selectively hide the association rules is proposed by Pontikakis, et al [16]. Two algorithms are proposed based on this approach namely Priority based distortion algorithm and weight based sorting distortion algorithm. Both algorithms are introducing a number of side effects that is lost rule and false rule. These side effects are minimized by using priority values assigned to transactions based on weight [14].

Wang and Jafari [17, 18] proposed two data modification algorithms to hide rules containing the sensitive items on their left hand side are known as predictive association rules. Both algorithms are performing based on decreasing the confidence of sensitive association rules. This approach needed multiple scan in the database. The first algorithm decreases the confidence of the rule by increasing the support of the itemset in left hand side as known as ISL. The second algorithm is decreases the confidence of the rule by increasing the support of the itemset in right hand side as known as DSR. Based on the item ordering effect, that is the order the sensitive items are hidden, both the algorithms are produced different sanitized databases for the same datasets. The DSR requires less running times then ISL, due to the size candidate transactions to be modified. The side effect of DSL algorithm shows 0% hiding failure, 5% of new rules and 11% of lost rule. The side effect of ISL algorithms shows 12.9% hiding failure, 33% of new rules and 0% of lost rules. The ISL algorithm requires more running time, due to more transitions are modified in the database then DSR. The DSR algorithm is effective for sensitive items are in high support. However, when the supports of sensitive item are low, ISL approach is effective.

Matrix multiplication approach is proposed for reduced the support of the sensitive itemsets in by Lee, et al [19]. Based on this approach the author proposed three algorithms namely Hidden-First(HF), Non-Hidden-First (NHF) and Hiding sensitive Patterns Completely with Minimum side Effect on non-sensitive patterns (HPCME). The first algorithm aims to hide all sensitive rules from the original database, but the hiding solution suffers from lost rule. The second algorithm focus on preserving the non-sensitive rules in the sanitized

database, but fail to hide all sensitive rules. The third algorithm combines first two algorithms to hide all sensitive rules with minimum loss of non-sensitive one. The algorithm proposed factor restoration probability to achieve the desire result.

Another dimension of association rule hiding research in heuristic algorithm is using data blocking approach. Unlike data distortion approaches, blocking approaches do not add any false information to the original database and it is a better alternative for real life applications. Due to the unknowns are introduced in sanitized database, the support and confidence of association rules that are mined from sanitized database becomes fuzzified to an interval and no longer be a safety estimated.

The blocking approach is first introduced by Saygin et.al [13, 14] with heuristic approach to propose three simple algorithms. The first algorithm, depend on the reduction in the support of the generating itemsets of the sensitive association rules, other two algorithms based on the reduction of the confidence of the rule, below the minimum thresholds.

The main drawback of blocking scheme is except the blocked values other dataset is not distorted. This makes an adversary can disclose the hidden association rules simply by identifying those generating itemsets that contain question marks. This will lead to rules with a maximum confidence that lies above the minimum confidence threshold. If the number of these rules is small then the probability of identifying the sensitive ones among them becomes high. To avoid this problem the authors propose a blocking algorithm that purposely generates rules that were not existent in the original dataset (i.e., false rules) and that their generating itemsets contain unknowns. Thus, the identification of the sensitive association rules becomes harder, since the adversary is unable to tell which of the rules that have a maximum confidence above the minimum threshold are the sensitive and which are the false ones. The introduction of false rules leads to a decrement of data quality in sanitized database.

Border based approaches: This approach considers the task of sensitive rule hiding through modification of the original borders in the lattice of the frequent and in frequent patterns in the dataset. In this approach sensitive knowledge is hidden by enforcing the revised border in the sanitized database. The authors X. Sun and P. S. Yu [20] are first introduced the process of border revision for the hiding of sensitive association rule using heuristic approach. This approach first compute positive and negative borders in the lattice of all itemsets. Second it focuses on preserving the quality of the computed borders during the hiding process. The quality of the border directly affected the quality of sanitized database that it is produced. To reduce the support of a sensitive itemsets from the negative border, the algorithm calculates the impact of the possible item deletion and it deletes the item that will have minimum impact on the positive border.

The approaches [20] is followed by Moustakides an Verykios[21] to propose two heuristic approach that use revised positive and negative borders. The proposed algorithms try to remove all sensitive itemsets from the database that belong to the negative border. This algorithm provides a better solution than [20], in the most of the tested settings.

Exact Approach: Algorithms based on these approaches are capable of providing better solutions compared with heuristic approaches with a high computational cost. This approach formulates sanitization process as constrain satisfaction problem and by solving it using an integer and/or linear programming solver. The local minima experienced by the heuristic approaches are avoided by performing database sanitization as an atomic operation. The other side of this algorithm suffers from an increasing computational complexity cost.

Menon, et al [22] proposed a methodology to hide frequent item set, which consists of two parts viz. exact part and heuristic part. The exact part uses the original database to formulate a Constrain Satisfaction Problem (CSP) in the universe of sensitive itemsets. The itemsets may identify the minimum number of transactions that have to be sanitized for hiding of the sensitive knowledge. The CSP is defined by a set of variables and a set of constrains [23] and an integer programming solver is used to find the CSP, where each variable as a non-empty domain of potential values. The heuristic algorithm is using this information to sanitize the original database. The objective of generating CSP is to improve the data utility and produce sanitized database with minimum number of modifications.

The methodology adapted in inline algorithm [24] aims to hide sensitive frequent itemsets of original database. Like Menon's algorithm [22], the inline algorithms [24] to identify exact part and optimal hiding solutions to hide frequent itemsets. An important property of this proposed algorithm is the problem formulation leads to a CSP with a size is larger than the one of [22], the hiding algorithm achieves better efficiency.

A two-phase iterative algorithm [25] consists two phases that iterate until either an exact solution of the given problem instance is identified, or a pre-specified number of subsequent iterations are taken place. This pre-specified number of iteration must be low enough to allow for a computationally efficient solution. Inline algorithm is used to hide the sensitive knowledge in this approach and the two phase algorithm is better than inline algorithm.

A new notation of hybrid database generation was introduced in [26], is the first exact methodology to perform sensitive frequent itemset hiding. This approach extends the regular sanitization problem by applying an extension to the original database instead of modifying or rebuilds the dataset for

sensitive knowledge hiding. This algorithm provides least amount of side effect compared with inline algorithm.

Limitations of association rule hiding approaches: All the above discussed approaches have their own limitations. The heuristic algorithm approach may suffer from undesirable side-effects that lead them to identify approximate hiding solutions. This is due to fact that heuristics always aim at taking locally best decisions with respect to the hiding of the sensitive knowledge which, however, are not necessarily also globally best [4]. Heuristic algorithms may cause undesirable side effects to non-sensitive rules, e.g. lost rules and false rules. In border based approach, theory of border revision is critical for the understanding. Although border-based approaches provide an improvement over pure heuristic approaches, they are still reliant on heuristics to decide upon the item modifications that they apply on the original database. As a result, in many cases these methodologies are unable to identify optimal hiding solutions, although such solutions may exist for the problem at hand [4]. Algorithms in exact approaches have very high time complexity due to the time that is taken by the integer programming solver to solve the optimization problem [4].

Reconstruction Based Approaches: Mielikainen [27] was the first to analyze the computational complexity of inverse frequent set mining and showed the problem is computationally difficult. The author showed that finding a dataset compatible with a given collection of frequent itemsets is NP- complete. For privacy preservation the results state that publishing frequent set might not cause threat to privacy because the inverse frequent set mining is difficult. Y. Guo [28] proposed a FP tree based algorithm to reconstruct the original database by using non characteristic of database. It generates database using non-sensitive frequent itemsets. Compared with heuristic approaches, this approach is performed over the set of frequent itemsets which is much closer to the association rules than data. This algorithm provides good efficiency and a number of secure databases.

Cryptography Based Approaches: Vaidya and Clifton [29] proposed a secure approach for sharing association rules when data are vertically partitioned. Proposed approach uses the scalar product over the vertical bit representation of itemsets inclusion in transaction, in order to compute the frequency of the corresponding itemsets. The authors proposed a secure two party algorithm for discovering frequent itemset. This approach is quite effective in terms of communication cost but it is very expensive for large datasets. The authors in [30] addressed the secure mining of association rules over horizontal partitioned data. The proposed algorithm uses the secure set union to get the union of candidate association rules. Then summation and secure comparison is used to filter candidate items that are not supported globally. This approach mines association rules securely with reasonable communication cost and computation cost.

RECENT TRENDS IN ASSOCIATION RULE HIDING

In recent trend, numerous researches have been done in association rule hiding for privacy preserving data mining. Wang Yan et al. [31] proposed a privacy preserving association rule mining algorithm based on Secondary Random Response Column Replacement (SRRCR). It can achieve significant improvements in terms of privacy and efficiency.

Chunhua Su and Kouichi Sakurai [32] focus on the privacy issue of the association rules mining and propose a secure frequent-pattern tree (FP-tree) based scheme to preserve private information while doing the collaborative association rules mining. They apply frequent-pattern tree (FP-tree) structure to execute the association rules mining and extend it to distributed association rules mining framework. Also they use FP-tree to compress a large database into a compact FP-tree structure to avoid costly database scans.

Mohammad Naderi Dehkordi et al. [33] proposed a novel method for privacy preserving association rule mining based on genetic algorithms. It also makes sure that no non-sensitive rules are falsely hidden (lost rules) and no extra false rules (ghost rules) are introduced to sanitized database in rule hiding process using genetic algorithm. The algorithm sanitizes both rule and itemset with minimal side effects by introducing new sanitization strategies and proposing new fitness functions according to new types of sanitization.

Andr´es Gago-Alonso et al. [34] proposed a new property of the DFS code which is useful to remove all the duplicate candidates in FCSM during the candidate enumeration was introduced. This property allows defining boundaries between useful and duplicate candidates during the pattern growth process. A new FCSM algorithm called gdFil was designed using the new property. Besides, they introduce a new ES, called DFSE, to reduce the cost of SI tests.

S. Vijayarani et al. [35] uses tabu search optimization technique to modify the sensitive items for hiding the sensitive association rules. This approach has the advantage of modifying the sensitive rules accurately without affecting the non-sensitive rules and no false rules are generated. The disadvantage is that it needs several iterations for selecting the optimal transaction for modification. By developing new fitness functions and applying other optimization techniques the number of iterations can be minimized.

Haifeng Li and Ning Zhang [36] this paper focuses on mining maximal frequent itemsets approximately over a stream landmark model. They separate the continuously arriving transactions into sections and maintain them with 3-tuple lists indexed by an extended direct update tree; thus, an efficient algorithm named FNMFIMoDS is proposed. In this algorithm, they used Chernoff Bound to prune the infrequent itemsets

and classified the itemsets into categories to prune the un-maximal frequent itemsets, which still can guarantee to obtain the proper itemsets; thus, this algorithm was able to perform in an incremental manner. Furthermore, they employed an extended direct update tree to index the itemsets, which can raise the computing efficiency. Their experimental results showed that the algorithm was more efficient in memory cost and running time cost in comparison with the state-of-the-art maximal frequent itemset mining algorithm.

Muhammad Naeem et al. [37] proposes an architecture which hides the restricted association rules with complete removal of the known side effects like generation of unwanted, non-genuine association rules while yielding no hiding failure. This architecture uses other standard statistical measures instead of conventional framework of Support and Confidence to generate association rules, specifically a weighing mechanism based on central tendency is introduced.

Komal Shah et al. [38] propose two algorithms, ADSRRC (Advanced Decrease Support of R.H.S. items of Rule Cluster) and RRLR (Remove and Reinsert L.H.S. of Rule), for hiding sensitive association rules. Both algorithms are developed to overcome limitations of existing rule hiding algorithm DSRRC (Decrease Support of R.H.S. items of Rule Cluster). Algorithm ADSRRC overcomes limitation of multiple sorting in database as well as it selects transaction to be modified based on different criteria than DSRRC algorithm. Algorithm RRLR overcomes limitation of hiding rules having multiple R.H.S. items.

Chun-Wei Lin et al. [39] are proposed a greed-based algorithm is used to insert newly transactions into the original database for efficiently hiding the sensitive itemsets. The number of newly inserted transactions and the length of each inserted transaction can be thus determined by empirical rules in standard normal distribution. The large itemsets in the original database are respectively added into the inserted transactions, for reducing the side effects of missing rules.

Peng CHENG et al. [40] are proposed an association rule hiding based on evolutionary multi-objective (EMO) optimization algorithm for removing items. The proposed EMO-based algorithm is to solve the association rule hiding problem. Optimized version of Genetic Algorithms, NSGA-II and SMSEMO were utilized to drive the evolution process forward. Comparative experiments demonstrated that EMO-based methods can effectively hide all sensitive rules with fewer side effects.

Priyanka k. Dhongade et al. [41] are propose a heuristic based algorithm named MDSRRC (Modified Decrease Support of R.H.S. item of Rule Clusters) to hide the sensitive association rules with multiple items in consequent (R.H.S) and antecedent (L.H.S). Proposed algorithm selects the items and transactions based on certain criteria which modify transactions to hide the sensitive information. 1. Sensitivity of Item: is number of sensitive rules which contain this item. 2. Sensitivity of Transaction: is the total of sensitivities of all sensitive items which are presented in that transaction. This algorithm overcomes the limitation of existing rule hiding algorithm DSRRC**.**

Saad M. Darwish et al. [42] proposed a methodology for building a sanitizing algorithm for hiding association rules at multiple concept levels. Employing multi-level association rule mining may lead to the discovery of more specific and concrete knowledge from datasets. The proposed system uses genetic algorithm as a biogeography-based optimization strategy for modifying multi-level items in database in order to minimize sanitization's side effects.

Maryam Fouladfar et al. [43] are proposed an algorithm, which uses distortion techniques based on reducing the confidence of sensitive rules. In this method, there is no limitation for hiding association rules with each number of items on the left and right hand sides of the base(=rule). Reduction of database scans and calculating the rate of changes before starting the hiding process would significantly reduce the amount of required operations for hiding process that shows the most efficiency on large databases. Also, in order to reduce the lost rules, victim item is calculated in each rule (=base) and according to that the leading rules would be specified for hiding.

R. Sugumar, et al. [44] proposed a Weight Based Sorting Distortion Algorithm for Privacy Preserving Data Mining. The distortion (WBSD) algorithm is to distort certain data which satisfies a particular sensitive rule. First hide that transaction which support a sensitive rule and assigns them a priority and sorts them in ascending order according to the priority value of each rule. Second it uses these weights to compute the priority value for each transaction according to how weak the rule is that a transaction supports. Data distortion is one of the important methods to avoid this kind of scalability issues.

Gayathiri. P. Poorna.B [45] proposed an Effective Gene Patterned Association Rule Hiding Algorithm for Privacy Preserving Data Mining on Transactional Database. This algorithm is based on perturbation technique, it show an notable ratio in true positive privacy rate using genetic algorithm.

CONCLUSION AND FUTURE DIRECTION

In this paper, a classification of privacy preserving techniques is presented and major algorithms in each class are surveyed. The merits and demerits of different techniques were pointed out. The optimal sanitization is proved to be NP- Hard and always there is a tradeoff between privacy and accuracy. All the proposed methods provide only approximate solution for the goal of privacy preservation. To address this, following

issues should be studied.

The rule hiding techniques based on fuzzy methods requires membership function to be specified by an expert. These algorithms either use ISL or DSR approaches to hide sensitive association rules. Hybrid technique can be applied with which side effects of rule hiding can be reduced. Metrics for measuring the side effects can also be developed. The algorithms for hiding sensitive association rules like privacy preserving rule mining using genetic algorithm, Tabu search based algorithms are limited to binary data, which can be extended to quantitative data. Although the personalized generalization approaches are flexible, the definitions of sensitive attributes are the same as other approaches. Thus, specifying sensitive information dynamically needs to be researched further. Rule sensitiveness measures in privacy preserving algorithms are generally limited to support and confidence. Depending on the nature of application, different measures can be used to measure the sensitiveness of quantitative rules. Semantic relation between attributes can be exploited in order to hide sensitive association rules with fewer side effects.

As each user may have different concern over privacy, user-oriented privacy preserving techniques can be developed. Parallel algorithms could be developed to prevent revealing of sensitive association between items and to improve the performance of the algorithm for large and dynamic datasets.

Most of the proposed research works are concentrating on side effects (lost rule and false rule) and numbers of sensitive rules are hidden from sanitized database. Those are not clearly stated about number of rules are hidden in each iteration, number of levels in multi-level sensitive rule hiding, number of scan needed for the database, computational efficiency in terms of memory and CPU time. In future, these objectives are also being considered and new techniques are to be proposed for hiding the sensitive association rules in privacy preserving data mining.

In association rule hiding, most of the research work has been concentrated on developing heuristics algorithms based on distortion concepts. In future, the other concept like blocking can also be used for sensitive association rule protection. Most of research mainly focuses on heuristic hiding approaches. In future, the other classes of hiding approaches, such as border based hiding approaches and the exact hiding approaches can also be considered and new techniques are to be proposed for hiding the sensitive association rules in privacy preserving data mining.

REFERENCES

[1] Agrawal, R., Imielinski, T., and Swami, A. N., " Mining association rules between sets of items in large databases", In Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, P. Buneman and S. Jajodia, Eds. Washington, D.C., 207-216.

[2] S. Giessing., "Survey on methods for tabular data protection in argues", In J. Domingo-Ferrer and V. Torra, editors, Privacy in Statistical Databases, volume 3050 of Lecture Notes in Computer Science, pages 1–13, Berlin Heidelberg, 2004. Springer.

[3] L.Willenborg and T. DeWaal, " Elements of Statistical Disclosure Control", Springer-Verlag, New York, 2001.

[4] Gkoulalas-Divanis, Aris, Verykios, Vassilios S. “Association Rule Hiding for Data Mining”, Springer Series: Advances in Database Systems, Vol. 41, 1st Edition., 2010, p.13.

[5] M. Atallah, E. Bertino, A. Elmagarmid, M. Ibrahim, and V. S. Verykios, " Disclosure limitation of sensitive rules", In Proceedings of the 1999 IEEE Knowledge and Data EngineeringvExchange Workshop (KDEX), pages 45–52, 1999.

[6] E. Dasseni, V. S. Verykios, A. K. Elmagarmid, and E. Bertino. "Hiding association rules by using confidence and support", In Proceedings of the 4th International Workshop on Information Hiding, pages 369–383, 2001.

[7] 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, 19(1):29–42, 2007.

[8] Vassilios S. Verykios and Aris Gkoulalas-Divanis, “A Survey of Association Rule Hiding Methods for Privacy”, Privacy-Preserving Data Mining : Model and algorithm Volume 34, 2008, pp 267-289.

[9] E. Dasseni, V. S. Verykios, A. K. Elmagarmid, and E. Bertino. “Hiding association rules by using confidence and support”, In Proceedings of the 4th International Workshop on Information Hiding, pages 369–383, 2001.

[10] V. S. Verykios, A. K. Emagarmid, E. Bertino, Y. Saygin, and E. Dasseni. “Association rule hiding”, IEEE Transactions on Knowledge and Data Engineering, 16(4):434–447, 2004.

[11] S. R. M. Oliveira and O. R. Zaïane. “Privacy preserving frequent itemset mining”, In Proceedings of the 2002 IEEE International Conference on Privacy, Security and Data Mining (CRPITS), pages 43–54, 2002.

[12] S. R. M. Oliveira and O. R. Zaïane. “Protecting sensitive knowledge by data sanitization”, In Proceedings of the 3rd IEEE International Conference

on Data Mining (ICDM), pages 211– 218, 2003.

[13] Y. Saygin, V. S. Verykios, and C. W. Clifton. "Using unknowns to prevent discovery of association rules", ACM SIGMOD Record, 30(4):45–54, 2001.

[14] Y. Saygin, V. S. Verykios, and A. K. Elmagarmid. "Privacy preserving association rule mining", In Proceedings of the 2002 International Workshop on Research Issues in Data Engineering: Engineering E–Commerce/E–Business Systems (RIDE), pages 151–163, 2002.

[15] A. Amiri. "Dare to share: Protecting sensitive knowledge with data sanitization", Decision Support Systems, 43(1):181–191, 2007.

[16] E. D. Pontikakis, A. A. Tsitsonis, and V. S. Verykios. "An experimental study of distortion– based techniques for association rule hiding", In Proceedings of the 18th Conference on Database Security (DBSEC), pages 325–339, 2004.

[17] S. L.Wang and A. Jafari. "Using unknowns for hiding sensitive predictive association rules", In Proceedings of the 2005 IEEE International Conference on Information Reuse and Integration (IRI), pages 223–228, 2005.

[18] S.-L.Wang, B. Parikh, and A. Jafari. "Hiding informative association rule sets", Expert Systems with Applications, 33(2):316–323, 2007.

[19] G. Lee, C. Y. Chang, and A. L. P. Chen. Hiding sensitive patterns in association rules mining. In Proceedings of the 28th International Computer Software and Applications Conference (COMPSAC), pages 424–429, 2004.

[20] X. Sun and P. S. Yu. A border-based approach for hiding sensitive frequent itemsets. In Proceedings of the Fifth IEEE International Conference on Data Mining (ICDM 2005), pages 426–433, 2005.

[21] G. V. Moustakides and V. S. Verykios. A max-min approach for hiding frequent itemsets. In Workshops Proceedings of the 6th IEEE International Conference on Data Mining (ICDM 2006), pages 502–506, 2006.

[22] S. Menon, S. Sarkar, and S. Mukherjee. Maximizing accuracy of shared databases when concealing sensitive patterns. Information Systems Research, 16(3):256–270, 2005.

[23] S. Russell and P. Norvig. Artificial Intelligence: A Modern Approach. Prentice–Hall, 2nd edition, 2003.

[24] A. Gkoulalas-Divanis and V. S. Verykios. An integer programming approach for frequent itemset hiding. In Proceedings of the 15th ACM International Conference on Information and Knowledge Management (CIKM), pages 748–757, 2006.

[25] A. Gkoulalas-Divanis and V. S. Verykios. Hiding sensitive knowledge without side effects. Knowledge and Information Systems, 20(3):263–299, 2009.

[26] A. Gkoulalas-Divanis and V. S. Verykios. Exact knowledge hiding through database extension. IEEE Transactions on Knowledge and Data Engineering, 21(5):699–713, 2009.

[27] Mielikainen, T.: On inverse frequent set mining. In: Proc. 3rd IEEE ICDM Workshop on Privacy Preserving Data Mining, pp. 18–23. IEEE Computer Society, Los Alamitos (2003)

[28] Guo, Y.: Reconstruction-Based Association Rule Hiding. In: Proc. of SIGMOD 2007 Ph.D. Workshop on Innovative Database Research (2007).

[29] Vaidya, J., Clifton, C.: Privacy preserving association rule mining in vertically partitioned data. In: Proc. Int'l. Conf. Knowledge Discovery and Data Mining, pp. 639–644 (2002).

[30] Kantarcioglu, M., Clifton, C.: Privacy preserving distributed mining of association rules on horizontally partitioned data. IEEE Transactions on Knowledge and Data Engineering 16(9), 1026–1037 (2004).

[31] Wang Yan, Le Jiajin and Huang Dongmei, " A Method for Privacy Preserving Mining of Association Rules Based on Web Usage Mining", International Conference on Web Information Systems and Mining,IEEE 2010, pp.33-37.

[32] Chunhua Su_ and Kouichi Sakurai, A Distributed Privacy-Preserving Association Rules Mining Scheme Using Frequent-Pattern Tree, Tang et al. (Eds.): ADMA 2008, LNAI 5139, pp. 170–181.

[33] Mohammad Naderi Dehkordi, Kambiz Badie, Ahmad Khadem Zadeh, " A Novel Method for Privacy Preserving in Association Rule Mining Based on Genetic Algorithms", Journal of software, vol. 4, no. 6, August 2009

[34] Andr´es Gago-Alonso, Jes´us Ariel Carrasco-Ochoa, Jos´e Eladio Medina-Pagola1, and Jos´e Fco. Mart´ınez-Trinidad, "Duplicate Candidate Elimination and Fast Support Calculation for Frequent Subgraph Mining", E. Corchado and H. Yin (Eds.): IDEAL 2009, LNCS 5788, pp. 292–299.

[35] S. Vijayarani, A. Tamilarasi, R. SeethaLakshmi, "Tabu Search based Association Rule Hiding", International Journal of Computer Applications 19(1):12-18, April 2011.

[36] Haifeng Li and Ning Zhang, A False Negative Maximal Frequent Itemset Mining Algorithm over

Stream*, J. Tang et al. (Eds.): ADMA 2011, Part I, LNAI 7120, pp. 29–41, 2011.

[37] Muhammad Naeem, Sohail Asghar, Simon Fong, "Hiding Sensitive Association Rules Using Central Tendency",Advanced Information Management and Service(IMS), 2010 6th International conference on, On page(s):478 - 484, Nov.30 2010 -Dec. 2 2010.

[38] Komal Shah, Amit Thakkar, Amit Ganatra, Association Rule Hiding by Heuristic Approach to Reduce Side Effects & Hide Multiple R.H.S. Items, International Journal of Computer Applications (0975 – 8887) Volume 45– No.1, May 2012,

[39] Chun-Wei Lin, Tzung-Pei Hong, Chia-Ching Chang, and Shyue-Liang Wang, A Greedy-based Approach for Hiding Sensitive Itemsets by Transaction Insertion, Journal of Information Hiding and Multimedia Signal Processing, Volume 4, Number 4, October 2013.

[40] Peng CHENG, Jeng-Shyang PAN, Association Rule Hiding Based on Evolutionary Multi-Objective Optimization by Removing Items, Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence,2014.

[41] Priyanka k. Dhongade ,Prof. Yogesh Nagargoje, DATA PRIVACY USING MDSRRC, International Journal of Engineering Research and General Science Volume 2, Issue 6, October-November, 2014.

[42] Saad M. Darwish, Magda M. Madbouly, and Mohamed A. El-Hakeem, A Database Sanitizing Algorithm for Hiding Sensitive Multi-Level Association Rule Mining, International Journal of Computer and Communication Engineering, Vol. 3, No. 4, July 2014.

[43] Maryam Fouladfar , Mohammad Naderi Dehkordi, A heuristic algorithm for quick hiding of association rules, Advances in Computer Science: an International Journal, Vol. 4, Issue 1, No.13 , January 2015.

[44] R. Sugumar, A. Rengarajan and C. Jayakumar, Design a Weight Based Sorting Distortion Algorithm for Privacy Preserving Data Mining, Middle-East Journal of Scientific Research 23 (3): 405-412, 2015.

[45] Gayathiri, P., and B. Poorna, Effective Gene Patterned Association Rule Hiding Algorithm for Privacy Preserving Data Mining on Transactional Database, Cybernetics and Information Technologies 17, no. 3, 2017.