International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2017 IJSRCSEIT | Volume 2 | Issue 6 | ISSN : 2456-3307

A Survey on Sanitizing Methods in Association Rule Hiding Technique

Apoorva Joshi, Pratima Gautam
Career College Bhopal, Aisect University, Bhopal, Madhya Pradesh, India

ABSTRACT

In current years, the use of data mining techniques and related applications has enlarged a lot as it is used to extract important knowledge from large amount of data. Now a days the incredible growth of data in every field[1]. This increment of the data created lots of challenges in privacy. Privacy preserving in data mining becomes too essential due to share this data for our benefit purpose[2].This shared data may contain sensitive attributes, Database containing sensitive knowledge must be protected against illegal access. Therefore this it has become necessary to hide sensitive knowledge in database. Privacy preserving data mining (PPDM) try to conquer this problem by protecting the privacy of data without sacrificing the integrity of data. A number of techniques have been proposed for privacy-preserving data mining. To address this problem, Privacy Preservation Data Mining (PPDM) include association rule hiding method to protect privacy of sensitive data against association rule mining. In this paper, we survey different existing approaches to association rule hiding, along with some open challenges. We have also summarized few of the recent evolution. and a review of different techniques for privacy preserving data mining along with merits and demerits.

Keywords : Privacy Preservation Data Mining, Association rule hiding, Data Mining

I. INTRODUCTION

Privacy preserving data mining (PPDM) is a dynamic research area in Data Mining (DM), where DM algorithms are analyzed and compared the impacts which occur in data privacy. The aim of PPDM is to transform the existing dataset in some way that the confidentiality of the data and knowledge remains intact even after the mining process. In DM, the users provided the data and they are free to use their own tools. So, the manipulation for privacy has to be applied on the data itself before the mining process[7].Protecting sensitive information in the context of our research surrounded with two main goals: knowledge protection and privacy preservation. The former is related to privacy preserving association rule mining, while the latter refers to privacy-preserving clustering. An attractive aspect between knowledge protection and privacy preservation is that they have a general characteristic. For instance, in knowledge protection, an organization is the owner of the data so it must protect the sensitive knowledge discovered from such data, while in privacy preservation individuals are the owner of their personal information.[2]

Association rule Mining-Association rules mining is a significant branch of data mining. Association rule knowledge is a popular and attractive researched technique for discovering elegant relations between variables in outsized databases[10] Association rule mining firstly proposed by Agrawal et al in 1993[6]. An association rule is an implication expression of the form X ∩ Y, where X and Y are disjoint item sets, i.e., X ∩ Y =∅.[1] Support and confidence are two important parameter of association rule mining. Definition of support and confidence is defined below [2]: Support is percentage of transactions in dataset that contain XUY.

.....

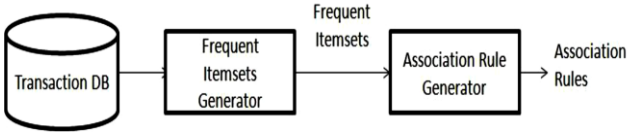
$$\text{Support}(XY): \quad \frac{\text{Total no of } (XY)}{\text{Total no. of transaction in D}}$$

Confidence is the percentage of transactions in dataset containing X that also contain Y. Confidence show the conditional probability.

$$\text{Confidence (XY):} \quad \frac{\text{Support (XY)}}{\text{Support (X)}}$$

Based on Minimum Support Threshold (MST) and Minimum Confidence Threshold (MCT) value, frequent item set and association rules are generated using different algorithm like apriori, FP growth. If we want to hide the rules then we should try to decrease the confidence value of that rule compare to MCT. we can do this by decreasing the value of confidence by increasing the value of denominator or by decreasing the value of numerator. And the value of denominator and numerator can be changed by altering the value of support count of Item sets. Modyfiying the values of support count are based on different methodologies .[10] Association rule mining works in two-step process:

- ✓ First find all frequent item sets- itemset which arise at least as frequently as a pre-determined minimum support count.
- ✓ Generate tough association rules- based on user defined minimum support and minimum confidence.



Different techniques of association rule mining for finding frequent item sets are available like Apriori algorithm, Partition algorithm, Pincher-search algorithm, Dynamic item set counting algorithm, FP-tree growth algorithm, etc [3]. Apriori algorithm is one of the most popular and best-known algorithm to mine association rule. It makes user of prior knowledge of frequent itemset properties, which is a two-step process: join step and prune step. It moves upward in the lattice starting from level1 till level k, where no candidate set remains after pruning. Apriori algorithm uses breadth first search strategy.[10]

Association rule hiding Techniques : Privacy Preserving Data Mining (PPDM) is used to extract related knowledge from large amount of data and protects the sensitive information from the data miners concurrently. Privacy preserving data mining is a attractive field in data mining. Privacy Preserving Data Mining (PPDM) solves the issues of designing precise models about combined data without access to exact information in individual data record. Association Rule Hiding is a PPDM technique use with Association Rule Mining method in transactional database. We can understand it by the following way.

An itemset is a set of products and transaction maintains simultaneously for a given set of items. The support of an itemset I in a transaction database is the percentage of transactions having I in the whole database. An itemset is frequent when the support is higher than a minimum support threshold (MST).

For two itemsets X and Y where $X \cap Y = \emptyset$ .The confidence of an association rule $X \rightarrow Y$is the probability that number of times Y occurs given that X occurs is equal to $SupX \cup Y$ divided by $SupX$. When $X \rightarrow Y$ holds in the database if $X \cup Y$ is frequent and its confidence is higher than a minimum confidence threshold (MCT). This rule is called the strong association rule. Association rule mining is used to discover all strong rules in the database.[12]

Association Rule Hiding Approaches-

4.1 Heuristic Based Approaches

This approach is further divided into two techniques: i) Data distortion technique and ii) Data Blocking Technique.[10]

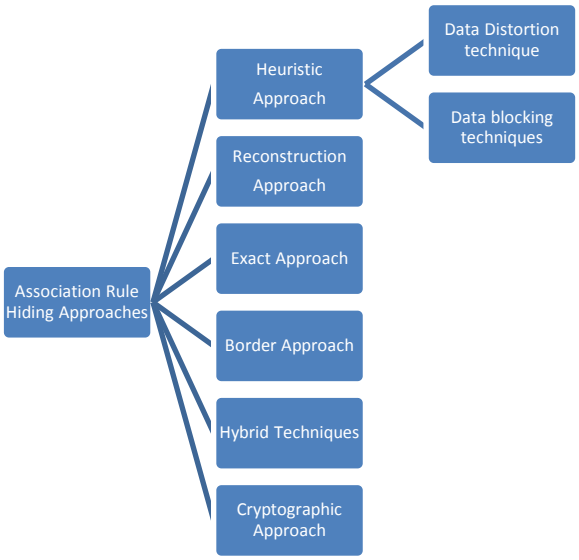
Heuristic Approach

Data Distortion technique : Data Distortion technique is a technique for modifying data using a random process. This technique apparently distorts sensitive data values by adding noise, data transpose matrix, or adding unknown values etc. This technique can handle different data types: character, Boolean, classification and integer. Discrete data need original data set to be processed. The processing of data is classified into attribute coding and obtaining sets coded data set. In most of the distortion techniques it is very difficult to hide the sensitive association rules through the reduction in the support of their generating item sets. The authors propose the construction of a lattice-like graph in the database.

Data blocking techniques

Blocking method works by reduction of the degree of support and confidence of sensitive association rules and replacing some attribute values of data items with

unknown values (?) or replace '1' by '0' or '0' by '1'. In this technique preserving privacy is done in two steps. First is to recognize transactions of sensitive rule and second is to replace the known values to the unknown, so the support of certain items goes



down to a certain level and rule mining algorithm not able to mine the sensitive rules [14]. One problem with block-based privacy preserving association rule mining is that it is too hard to calculate the support and confidence of a sensitive association rule since the some of the original data is replaced with unknown value [14], [15]. This can be solved by using uncertain symbols which then can be restored with actual The first approach, relies on the reduction in the support of the generating itemsets of the rule, while the other two rely on the reduction of the rule confidence of the rule, below the minimum thresholds.

4.2 Border Based Approaches

The process of border modification is introduced by X. Sun [12], The authors propose a heuristic approach that uses the notion of the border (improve effectiveness of the previous work in [9]) of the non-sensitive frequent itemsets to track the impact of altering transactions in the database. The proposed method first computes the positive and the negative borders in the lattice of all itemsets and then focuses on preserving the quality of the computed borders during the hiding process. The quality of database can be well maintained by greedily selecting the changes with minimal side effect. In the proposed heuristic, a weight is assigned to each element of the calculated positive border to quantitative the effect of deleting an item. During the sanitization process these weights are dynamically computed according to the current support of the equivalent itemsets in the database. To reduce the support of a sensitive itemset from the negative border, the algorithm calculates the effect of the possible item deletions by calculating the sum of the weights of the positive border elements that will be affected. Then, it proceeds to delete the items that will have the minimal impact on the positive border. In [5], authors improves the hiding solutions of [8],The proposed algorithms follow a similar approach and try to modify this item in such a way that the support of the max- min itemset is minimally affected. In case of multiple itemsets the hiding process starts with lower support itemset at one at a time base.

4.3 Exact Approaches

Exact approaches are usually able to offer better quality solutions compared to the heuristic approaches, but with a high intricacy cost. This is coming through represent the sanitization process as a constraint satisfaction problem and by solving it using linear or integer programming solver. Sanitization process is done as an atomic operation to prevent the local minima experienced by the heuristic approaches. It solves the problem as a Constraint Satisfaction Problem (CSP) with a goal to discover the minimum number of transactions that need to be sanitized for the suitable hiding of all the sensitive knowledge. It works with the sensitive itemsets only to reduce the problem size, apply for their support stays lower than the minimum support threshold. The optimization process is determined by a standard measure function that is glorious by the measure of accuracy. Moreover, the constraints obligatory in the CSP formulation catch the number of supporting transactions that need to be sanitized for the hiding of each sensitive itemset. The best
solution of the CSP can be identified by using integer programming solver to satisfy the objective[16].

4.4 Reconstruction Based Approach

Reconstruction approach has two ladder, first perform distortion of data and then reconstructing the distributions. There are several algorithms for reconstructing the distributions and data types. For distributed data, Bayesian reconstruction process is used which is based on EM algorithm. EM algorithm is robust and it can estimate the original distribution when a large amount of data is obtained. An additional way of data reconstruction is to keep the original data aside

and start from sanitizing knowledge base. The new data are reconstructed from the sanitized knowledge base [9].

4.5 Cryptography Based Approaches

Cryptography is a technique through which sensitive data can be encrypted. It is a good technique to protect the data. In [12], According to the auother cryptographic technique which is very general because it provides security and safety of sensitive attributes. There are different algorithms of cryptography available. But this technique has many disadvantages. It fails to protect the output of computation. It prevents privacy leakage of the computation. This algorithm does not give successful results when it talks about more parties. It is very complex to apply this algorithm for huge databases. Final data mining result may violate the privacy of the individual Record.

Table 1. Summary of association rule hiding approaches

Advantage	Limitation
Heuristic Based Approaches (Distortion technique)	
More efficient, scalable	Difficult to revert the changes made in database
Heuristic Based Approaches (Blocking technique)	
It maintains veracity of database, since instead of inserting false value it just block original value.	Suffer from various side effects like ghost rule, lost rule etc.
Border Based Approaches	
Maintains data quality by greedily selecting the modification with minimal side effects. Improvement over pure heuristic approach.	Unable to identify optimal hiding solution But still dependent on heuristic to decide upon the item modification.
Exact Approaches	
Guarantees quality for hiding sensitive information than other approaches.	But requires very high time complexity due to integer programming
Reconstruction Approaches	
Create privacy aware database by exacting sensitive characteristic from the original database. Lesser side effects in database than heuristic	The open problem is to restrict the number of trans-actions in the new database.
Cryptographic Approaches	
Secure mining of association rule over partitioned database.	Do not protect the output of a computation. Falls short of providing a complete answer to the problem of privacy preserving data mining. Communication and computation cost should be low.

II. EVALUATION METRICS

Following metrics are used to evaluating association rule hiding algorithms [15][16].

1) **Efficiency**- It is measured in terms of CPU-time, space requirements and communication required for hiding. In short, excellent performance in terms of resources allocated.

2) **Scalability**- It is measured in terms of high-quality performance for increasing sizes of input datasets.

3) **Data quality**- Data quality parameters are accuracy measure, completeness, consistency which are in relationship to preservation of original data values and of data mining results.

4) **Hiding failure**- It is the percentage of the piece of information that fails to be hidden. It is derived by, HF = |Rs(D')| / |Rs(D)| where, |Rs(D')| are the number of sensitive rules appearing in the sanitized database and |Rs(D)| are the number of sensitive rules in the original database.

5) **Privacy level**- It measures the degree of uncertainty according to which the protected information can still be predicted.

6) **Lost Rules cost**- It measures the number of nonsensitive association rules found in the original database but not in sanitized database.

7) **Ghost Rules**- It measures the percentages of rules that are not there in the original database but can be derived from sanitized database.

8) **Dissimilarity**- It quantify difference between original database and sanitized database.

III. CONCLUSION

Association rule hiding is an important concept in the area of privacy preserving data mining. It protects the privacy of sensitive information in databases against the association rule mining approaches. In this paper, represented survey of the different approaches for

privacy preserving data mining, and analyses the major algorithms available for each method and points out the existing drawback. While all the purposed techniques are only approximate to our goal of privacy preservation, we need to further perfect those approaches or develop some efficient methods. To address this issue, the following problems should be considered: Privacy and accuracy is a couple of contradiction; improving one usually incurs a cost in the other. How to apply various optimizations to achieve a trade-off should be deeply researched. Lot of Side-effects are there with data sanitization process. How to reduce their negative impact on privacy preserving needs to be considered carefully. We also need to define some metrics for measuring the side-effects resulted from data processing. In distributed privacy preserving data mining areas, efficiency is an essential issue. We should attempt to develop more efficient algorithms and achieve a balance between identity disclosure cost, computation cost and communication cost. It also presents a comprehensive survey on the list of existing association rule hiding techniques to hide sensitive item set without revealing pattern. Existing approaches provide only the approximate solution to hide sensitive knowledge. There is need of finding exact solution to the privacy problem in database disclosure.

IV. REFERENCES

[1]. Khyati B. Jadav, Jignesh Vania , Dhiren R. Patel "A Survey on Association Rule Hiding Methods," .In Proc. of the International Journal of Computer Applications (0975 – 8887) Volume 82 – No 13, November 2013

[2]. Nidhi Jain, Prof.Angad Singh, "A Survey On Privacy Preserving Mining Various Techniques With Attacks". In Proceedings Of International Journal Of Research In Computer Applications And Robotics,Volume 5,Issue 5,Issn 2320-7345

[3]. Peng Cheng1,3 • John F. Roddick2 • Shu-Chuan Chu2 • Chun-Wei Lin1, "Privacy Preservation Through A Greedy, Distortion-Based Rule-Hiding Method". In Proc. Of The Springer Science+Business Media New York 2015

[4]. George V. Moustakides a, Vassilios S. Verykios b,*"A MaxMin approach for hiding frequent itemsets", Data & Knowledge Engineering 65 (2008) 75–89

[5]. 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

[6]. Neelkamal Upadhyay, Kuldeep Tripathi, Prof. Ashish Mishra" A Survey Of Association Rule Hiding Approachesiracst - International Journal Of Computer Science And Information Technology & Security (Ijcsits), Issn: 2249-9555 Vol. 5, No1, February 2015

[7]. Umesh Kumar Sahu, Anju Singh" Approaches For Privacy Preserving Data Mining By Various Associations Rule Hiding Algorithms – A Survey" International Journal Of Computer Applications (0975 – 8887) Volume 134 – No.11, January 2016

[8]. Kasthuri S1 And Meyyappan T2" Hiding Sensitive Association Rule Using Heuristic Approach" International Journal Of Data Mining & Knowledge Management Process (Ijdkp) Vol.3, No.1, January 2013 Doi

[9]. 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

[10]. Divya C. Kalariya, Vinita Shah; Jay Vala" Association Rule Hiding Based On Heuristic Approach By Deleting Item At R.H.S. Side Of Sensitive Rule" International Journal Of Computer Applications (0975 – 8887) Volume 122 – No.8, July 2015

[11]. Komal Shah, Amit Thakkar & Amit Ganatra "A Study On Association Rule Hiding Approaches" International Journal Of Engineering And Advanced Technology (Ijeat) Issn: 2249 – 8958, Volume-1, Issue-3, February 2012, Pp. 72-76.

[12]. Gayathiri P ,Dr. B Poorna"Association Rule Hiding Techniques For Privacy Preserving Data Mining: A Study, (Ijacsa) International Journal Of Advanced Computer Science And Applications, Vol. 6, No. 12, 2015

[13]. Mohamed Refaat Abdellah, H. Aboelseoud M. Khalid Shafee Badran, M. Badr Senousy" Privacy Preserving Association Rule Hiding Techniques: Current Research Challenges" International Journal Of Computer Applications

(0975 – 8887) Volume 136 – No.6, February 2016

[14]. Vassilios S. Verykios, Aris Gkoulalas-Divanis," A Survey Of Association Rule Hiding Methods For Privacy"Springer Volume 34

[15]. Vassilios S. Verykios∗"Association Rule Hiding Methods" Wires Data Mining Knowl Discov 2013, 3: 28–36 Doi: 10.1002/Widm.1082

[16]. Supriya Borhade" A Survey On Privacy Preserving Data Mining Techniques" Issn 2250-2459, Iso 9001:2008 Certified Journal, Volume 5, Issue 2, February 2015