Survey of Privacy Preserving algorithm for Protect Sensitive Information

Pratiksha Sethi [1] **Nishi Thakur** [2]
[1,2]Department of Computer Science & Engineering
[1, 2] Sanghvi Institute of Science & Management, Indore, India

Abstract— Privacy in the present era has become a more rational issue either it may be in the real world or in cyber world. The field of privacy has seen rapid advances in recent years because of the increases in the ability to store data but due to recent advances in the data mining field have led to increased concerns about privacy. Privacy preserving data mining is a research area concerned with the privacy driven from personally identifiable information when considered for data mining. This project addresses the privacy problem by considering the privacy and algorithmic requirements simultaneously. The proposed work has the basis of reduction of support and confidence of sensitive rule. This project is to implement an association rule hiding algorithm for privacy preserving data mining which would be efficient in providing confidentiality and improve the performance at the time when the database stores and retrieves huge amount of data.

Key words: Data Mining, Association Rule, Support, Confidence, Privacy Preserving.

I. INTRODUCTION

Data Mining refers to "*extracting*" or "*mining" knowledge from large amounts of data* [10]. Data mining process is to find the patterns that are hidden among the huge sets of data and interpret them to useful knowledge and information. Privacy preserving data mining [8] is a novel research direction in data mining where data mining algorithms are analyzed for the side-effects they incur in data privacy. One known fact which is very important in data mining is discovering the association rules from database of transactions where each transaction consists of set of items. Two important terms support and confidence are associated with each of the association rule. The work we have proposed here has the basis of reduction of support and confidence of sensitive rules but this work is not editing or disturbing the given database of transactions directly .In this project the proposed algorithm uses some modified definition of support and confidence so that it would hide any desired sensitive association rule without any side effect.

II. ASSOCIATION RULE HIDING ALGORITHM

The objective of association rule hiding algorithm is to hide certain confidential data so that they cannot be discovered through data mining techniques [10].Association rules are if/then statements that help uncover relationships between seemingly unrelated data in a relational database or other information repository. An example of an association rule would be "If a customer buys a dozen eggs, he is 80% likely to also purchase milk [9]."

An association rule has two parts, an antecedent (if) and a consequent (then). An antecedent is an item found in the data. A consequent is an item that is found in combination with the antecedent. *Support* is an indication of how frequently the items appear in the database. *Confidence* indicates the number of times the if/then statements have been found to be true. The problem of association rule mining is defined as: Let I={i1,i2,i3………in} be a set of n binary attributes called *items*. Let D={t1,t2,t3……tn} be a set of transactions called the *database*. Each transaction in D has a unique transaction ID and contains a subset of the items in I. A *rule* is defined as an implication of the form X=>Y where, $X, Y \subseteq I_{X,Y}$ and $X \cap Y = \emptyset$. The sets of items (for short *item sets*) X and Y are called *antecedent* (left-hand-side or LHS) and *consequent* (right-hand-side or RHS) of the rule respectively.

To select interesting rules from the set of all possible rules, constraints on various measures of significance and interest can be used. The best-known constraints are minimum thresholds on support and confidence. The *support* supp(X) an itemset X is defined as the proportion of transactions in the data set which contain the itemset. The *confidence* of a rule is defined

$$conf(X \Rightarrow Y) = supp(X \cup Y)/supp(X).$$

Association rules [9] are usually required to satisfy a user- specified minimum support and a user-specified minimum confidence at the same time. Association rule generation is usually split up into two separate steps:

1) First, minimum support is applied to find all frequent item sets in a database.
2) Second, these frequent item sets and the minimum confidence constraint are used to form rules. Given a rule r and calculate

Minconf(r), maxconf(r) as minconf(r) = minsup(r)*100 / maxsup (lr) -------------- (1)

Maxconf(r) = maxsup(r)*100 / minsup (lr) ---- (2)

Where, lr denotes the rule antecedent.

Considering the support interval and the minimum support threshold have the following cases for an itemset A: (i) A is hidden when maxsup(A) is smaller than MST (ii) A is visible with an uncertainty level when minsup (A) ≤ MST ≤ maxsup(A) (iii) A is visible if minsup(A) is greater than or equal to MST.

III. RULE HIDING PROCESS

A rule hiding process takes place according to two different strategies:

1) Decreasing its support.
2) Decreasing its confidence.

In this method, the adopted alternative strategies aim at introducing uncertainty in the frequency or the importance of the rules to hide [9]. The two strategies reduce the minimum support and the minimum confidence of the item sets generating these rules below the Minimum Support Threshold (MST) and Minimum Confidence Threshold (MCT) correspondingly by a certain Safety Margin Threshold (SMT) fixed by the user. In order to reduce the support of the large itemset generating a sensitive rule, Algorithm 1 replaces 1‟s by " ? " for the items in transactions supporting the itemset until its minimum support goes below the minimum support threshold MST by the fixed safety margin SM. The first rule decreases the minimum support of the generating itemset of a sensitive rule by replacing items of the rule consequent with unknown values. Whereas the second rule increases the maximum support value of the antecedent of the rule to hide via placing question marks in the place of the zero values of items in the antecedent. All the algorithms hide a sensitive rule with an uncertainty level by decreasing the minimum support or confidence values below the resulting thresholds, MST-SM and MCT-SM [9].

IV. DECREASING SUPPORT AND CONFIDENCE

Considering the case of decreasing support value, the support S of a rule X => Y is given by

$$|X\ U\ Y| * 100 / N \text{ --------------------- } (1)$$

Since N is constant (as it is the number of transactions in the given database), the only option left for this is to change the numerator value (option (a)) and decrease the support of any rule by decreasing the support of the generating item set of the rule. Considering the case of decreasing confidence value, Confidence C of a rule X => Y is given by

$$|X\ U\ Y| * 100 / |X| \text{ ------------------- } (2)$$

The first option implies to decrease the numerator (which is the support) of the generating item set of the rule, while the support of the item set in the left hand side of the rule remains fixed. The second option implies to increase the denominator (which is the support of the item set in the antecedent) of the rule, while the support of the generating item set of the rule remains fixed.

V. PROPOSED APPROACH

Internet communication technology has made this world very competitive. In their struggle to keep customers, to approach new customers or even to enhance services and decision making, data owners need to share their data for a common good. Privacy concerns have been influencing data owners and preventing them from achieving the maximum benefit of data sharing. Data owners usually sanitize their data and try to block as many inference channels as possible to prevent other parties from finding what they consider sensitive. Data sanitization is defined as the process of making sensitive information in non-production databases safe for wider visibility. However, sanitized databases are presumed secure and useful for data mining, in particular, for extracting association rules.

To hide any specified association rule X→ Y this algorithm works on the basis of confidence (X→ Y) and support (X→ Y). To hide any sensitive rule X→ Y, this algorithm first finds the value of support (sup) and confidence (conf) in the available set of rules and then it computes the support and confidence of the sensitive rule using following:

Confidence (X→ Y) = (conf * 1/3);

Support (X→ Y) = (sup * 1/3);

A. Procedure

```
//find value of support and confidence
Select confidence into conf from database.
Select support into supp from database.
For each X
{
// Now check all the rules containing sensitive element x.
For each rule R which contain X on LHS OR RHS.
{
While (conf(R) >= MCT)
{
Set confidence(X→ Y) = (conf * 1/3);
Set support (X→ Y) = (sup * 1/3);
}}}
End of procedure
```

VI. CONCLUSION

The work presented in here, indicates the ever increasing interest of researchers in the area of securing sensitive data and knowledge from unauthorized users. The conclusion that we have reached from reviewing this area, that privacy issues can be effectively considered only within the limits of certain data mining algorithms. The inability to generalize the results for classes of categories of data mining algorithms might be a tentative threat for disclosing information.

REFERENCES

[1] Shyue-Liang Wang, Yu-Huei Lee, Steven Billis, AyatJafari "Hiding Sensitive Items in Privacy Preserving Association Rule Mining" 2004 IEEE International Conference on Systems, Man and Cybernetics

[2] Vassilios S. Verykios, Ahmed K. Elmagarmid, Elisa Bertino, YucelSaygin and Elena Dasseni "Association Rule Hiding", IEEE Transactions on Knowledge and Data Engineering, Vol. 16No. 4, April 2004.

[3] R. Agrawal and R. Srikant, "Privacy preserving data mining", In ACM SIGMOD Conference on Management of Data, pages 439450, Dallas, Texas, May 2000.

[4] i-Hung Wu, Chia-Ming Chiang, and Arbee L.P. Chen, Senior Member, IEEE Computer Society Hiding Sensitive Association Rules with Limited Side Effects IEEE transactions on knowledge and data engineering, vol. 19, no.1, JANUARY 2000

[5] S. Oliveira, o. Zaiane, "Algorithms for Balancing Privacy and Knowledge Discovery in Association Rule Mining", Proceedings of 71th International Database Engineering and Applications Symposium (IDEAS03), Hong Kong, July 2003.

[6] C. Clifton and D. Marks, “Security and Privacy Implications of Data Mining,” Proc. 1996 ACM Workshop Data Mining and Knowledge Discovery, 1996.

[7] C. Clifton, “Protecting against Data Mining through Samples,” Proc. 13th IFIP WG11.3 Conf. Database Security, 1999.

[8] T. Johnsten and V.V. Raghavan, “Impact of Decision-Region Based Classification Mining Algorithms on Database Security,” Proc. 13th IFIP WG11.3 Conf. Database Security, 1999

[9] International Journal of Advanced Research in Computer and Communication Engineering Vol. 1, Issue 7, September 2012.

[10] International Journal of Advanced Technology & Engineering Research (IJATER)

[11] Jiawei Han and Micheline Kamber, “Data Mining Concepts and Techniques- second edition”