International Journal of Research in Advent Technology, Vol.6, No.9, September 2018
E-ISSN: 2321-9637
Available online at www.ijrat.org

Privacy Preserving Association Rule Mining for Customer Relationship Management

Tamanna kachwala[1], L. K. Sharma[2]
Ph. D. Scholar, Rai University, Saroda, Ahmedabad, Gujarat, India[1]
National Institute of Occupational Health, Ahmedabad, Gujarat, India[2]
Email: tamanna1828@yahoo.co.in[1] , lksharmain@gmail.com[2]

Abstract- Organizations have a huge customer base and thus they use data mining tools to study their customers. However there is risk of sensitive information about individuals which can be gained also during this process. Hence data that is used for data mining has to be protected. By developing technology and competition in different fields, preserving sensitive data is considered as a problematic issue for users. Many Industry, Defense ,Public Sector and Organization facing risk or having security issue while sharing their data so it is very crucial concern how to protect their sensitive information due to legal and customer concern. Many strategies have been proposed to hide the information containing sensitive data. Privacy preserving data mining is an answer to such problems. Association rule hiding is one of the PPDM techniques to protect the sensitive association rule.

Index Terms- Customer relationship management, Association Rule Mining, sensitive information

1. INTRODUCTION

Organizations such as customer relationship management, the telecommunication industry, financial sector investment trends, web technologies, demand and supply analysis, direct marketing, health industry, e-commerce, stocks and real estate, understanding consumer research marketing, e-commerce, and product analysis generate huge amounts of data that often contain useful information (i.e. name, address, age, salary, social security number, type of disease, and the like). Through data mining, we are able to extract useful and previously unknown information that organizations or individuals do not want to disclose to the public. Therefore, privacy preserving data mining (PPDM) techniques are applied to preserve such confidential information from any type of mining algorithm [1-4]. Hence, the basic objective of PPDM is to protect data against serious adverse effect. In addition, the privacy regarding data mining is divided into 2 types. The first type of privacy, termed as output privacy, is where the data are altered so that the mining result will conserve certain privacy. Many modification techniques such as perturbation, blocking, aggregation, swapping, and sampling are used for this type of privacy [5-10]. The second type of privacy, labeled as input privacy, is where the data are manipulated so that the mining result is not affected or is less affected. Cryptography- and reconstruction-based techniques are used for this type of privacy [11-15].

2. RELATED STUDY

In [16] presented two algorithms to hide association rules. The first algorithm called ISL decreases confidence rule by increasing support in sensitive rule in the left hand side elements as they select rules that have sensitive items in their left hand and insert sensitive items in transactions that don't contain this rule. This algorithm has lots of failure in hiding and making new rules. The second algorithm called DSR decreases support rule by decreasing support in right hand side elements as they select rules that have sensitive items in their right hand and removes sensitive items from those transactions which contain this rule. Failure in this algorithm is close to zero and many non-sensitive rules will be lost [16, 17].
In [18] presented an algorithm called DSRRC that applied clustering for right hand common items in hiding. The drawback of this algorithm is in hiding those items which have one element in their right side, it's dependent on arrangement in transactions and shows different results by changing the orders of transactions in database, it requires arrangement after deleting every item and it is not appropriate for large databases. There are lots of lost rules in this algorithm.
In [19] presented an algorithm called ADSRRC for improving DSRRC algorithm. This algorithm also hides those rules which have single RHS and arrangement is made once only. In addition, in this article, an algorithm called RRLR is suggested that hides those rules which have single LHS. In this algorithm, to hide sensitive rules both support and confidence are decreased as in transaction with high degree of sensitivity, left hand item is deleted and insert a transaction which has partially sensitive rule.

In [20] proposed MDSRRC algorithm to eliminate restriction in the number of left and right items. This algorithm selects the best item for deletion based on its frequency on the right side of the sensitive rule and supporting that item. This algorithm contains the minimum side effect compared with DSRRC. Failure in hiding is close to zero.
In [21] combined ISL and DSR algorithms together and made the main purpose on declining the number of changes in database and decreasing the time to hide sensitive rules.
In [22] presented a heuristic algorithm called ABS. in this algorithm selection of transaction is randomly done. Its idea is originated from the way honey bees looking for the source of food. In this algorithm a support-based method is used.
In [23] proposed two algorithms called Random, Round Robin that its base is on selecting an item to preserve that is done in order or randomly.
In [24] proposed an algorithm called SRH by which they could decrease time and memory complexity by computing the number of required transactions for hiding sensitive rule.
In [25] proposed an algorithm with an accurate focus called integer programming and blanket, intelligent strategies. The advantage of this algorithm was in hiding rate, assuring the best accuracy level, formulating for measurement and solving the problems in an optimizing way.
In [26] proposed WSDA and BA algorithms. WSDA hides sensitive rules by distorting technique and BA hides by blocking technique. WSDA algorithm concentrates on optimizing hiding techniques to minimize side effects and have the least complexity in hiding. This algorithm is not appropriate for large databases. The aim for BA algorithm is to hide rules that cannot be discoverable and to minimize the number of lost association rule and ghost rule.
In [27] proposed aggregate, disaggregate and hybrid algorithms that hide sensitive rules based on support-based method. In the first algorithm called aggregate supporting sensitive rule is decreased by deleting some transactions. The second algorithm called disaggregates declines supporting degree of sensitive rules by deleting some sensitive elements. The third algorithm called Hybrid determines the identified transactions by aggregate method and then specifies the required elements for deleting by disaggregate method.
Introduced a new method to preserve privacy based on genetic algorithm, to make sure no ghost rule or lost rule is made. This algorithm is based on rules and items with the least amount of side effect through hiding strategy. Three strategies for Selection of an item and three strategies for Crossover of an item are suggested in this algorithm [28].
In [26] presented two strategies and five algorithms that decline the degree of support for productive sensitive rules to reach to less than minimum amount of support. This is done in two ways: 1-deleting an item that contains maximum supporting degree of a transaction with the least length. 2- Sorting a group of sensitive productive rules according to their length and support, and hiding them by rotation [29].

3. METHODOLOGY

The hiding sensitive rules in the form of $M \rightarrow N$, can be performed by decreasing either the confidence or support of the rules to below the MST and MCT. In fact, the support of the rule $M \rightarrow N$ can be decreased by reducing the occurrences of item set MN. Similarly, the confidence of the rules can be reduced by one of the following techniques:

1. Increasing the support of the antecedent of the rule, i.e. LHS part, in transactions in which the consequence of rule is not present.

2. Decreasing the support of rule's consequence, i.e. RHS part, in transactions including both parts of the rule [30].

The hiding method aims at hiding sensitive rules with multiple items in LHS and multiple items in RHS. This type of rule is represented in the form of $aM \rightarrow bN$ where $a, b \in I$ and $M, N \subset I$. Here, 'a' and 'b' are single items selected by the algorithm to be inserted into or removed from LHS or RHS of the rule, respectively. The idea behind the proposed algorithm is to decrease both support and confidence measures to hide the sensitive rules. The support of the rules that are in the form of $aM \rightarrow bN$ can be decreased by reducing the support of the itemset aMbN. Also, to decrease the support of the large itemset aMbN, a suitable item is selected and removed from the LHS. items of the sensitive rule; and to decrease confidence of the sensitive rule, a selected item is inserted in the suitable transaction. To this end, the algorithm identifies a list of suitable transactions for modification that are called victim transactions. To identify a set of items to remove from the victim transactions, two objective parameters, namely α and β, are calculated for items that exist in the sensitive rules. They are also used to calculate the total sensitivity of the victim transactions.

These parameters are defined as follows:

1. Parameter α: The number of occurrence for LHS items of the sensitive rules in the whole set of non-sensitive rules. This parameter is used to construct the list $L\alpha$ in which LHS items of the

sensitive rules are sorted in ascending order, according to the value of α.

2. Parameter β: The frequency of an item in the set of sensitive rules, which is used to compute the sensitivity of each transaction.
3. Sensitivity of transaction: The sensitivity of a transaction is calculated as the sum of β values of all sensitive items included in that transaction.

In the first step, association rules are mined from the original database by using the Apriori algorithm. Then, sensitive rules are selected from the mined rules (Rs) and are sorted in descending order, according to their confidence value. To make sure that all sensitive rules are hidden, a Boolean variable named "State" is defined to maintain the hiding status of each sensitive rule. The states of all sensitive rules are initially set to false. Now, parameters α, β and sensitivity of transaction are computed by the algorithm and the transactions are arranged in descending order, based on their sensitivity and length. Next, Lα is computed by the algorithm and the process of hiding sensitive rules starts from the first sensitive rule.

Among the LHS items of the first sensitive rule, an item with the least value of α (according to list Lα) is selected to be removed. Then, the selected item is removed from the first transaction that has the highest sensitivity and length.

After removing, the selected item is inserted in the transactions, which do not have the item (i.e. the large itemsets that partially support LHS of the rule and partially support, or do not support RHS of the rule). If the suitable transaction for the selected item is not found, the insertion will not be done. During the process of hiding, the sensitivity of transactions will not be updated.

After each removal and insertion, support and confidence of the sensitive rules existing in Rs will be updated. If they reach below the MST and MCT, the false state of the sensitive rule is changed to true, without being removed from the list Rs. The rule state is changed from true to false, if a sensitive rule becomes disclosed because of inserting an item. In this situation, there will be no insertion and the rule is hidden just by removing a suitable item of the left side ones. As shown in Figure 1, a Boolean variable, called Disclosed, is used for identifying disclosed rules and preventing the algorithm from calling item insertion process. The hiding goes on until the state values of all sensitive rules become true.

Finally, the transactions in the original database are modified and constitute a new database that is a sanitized version of the original database D. This preserves the privacy of the sensitive data and keeps data quality. The main steps of the proposed algorithm named ARRLR (Advanced Remove and Reinsert LHS of Rule) are shown below.

Input: Original database D, (MCT), (MST).

Output: The sanitized database D′.

Use Apriori, extract association Rules. Select a set of sensitive rules Rsensitive (Rs) Set the State and Disclosed variable of all sensitive rules to false Sort Rs in decreasing order based on their Confidence. Calculate α, β and sensitivity of transactions. Create Lα by sorting LHS items of the sensitive rules in ascending order based on their α value Arrange transactions in decreasing order of their sensitivity and length.

```
while (states of all Rsensitive are not true)
{
        Find the first rule Rk from Rsensitive such
        that Rk. State
        is false
        Select item I from LHS of rule Rk according
        to Lα
        //Antecedent Deletion Process
        for m = 1 to no. of transactions in database
{
if (Tm supports both parts of rule Rk)
        Remove selected antecedent item I from
        transaction Tm
if (Rk. Disclosed is false)
{
        // Antecedent Insertion Process
        for n = m to no. of transactions in database
{
if (Tn does not include item I and partially
        supports rule Rk)
        Insert selected LHS item I in transaction Tn
}
}
for each rule R in RS
{
        Recalculate Support & Confidence for R
if (R.Support < MST or R.Confidence < MCT)
        Set R.State to true.
else
{
If (R.State is true)
        Set R.Disclosed to true
        Set R.State to false.
}
}
}
```

4. RESULT AND DISCUSSION

The two data were obtained from the UCI machine-learning repository related to customers. The first data is associated with direct marketing campaigns of a Portuguese banking institution [13]. The phone calls were used for the marketing campaigns. This data set includes 20 various social, economic attributes of

customers. The second data set contains information on customers of an insurance company. The data consist of 86 variables and includes product usage data and socio-demographic data. The data were supplied by the Dutch data mining company Sentient Machine Research and is based on a real world business problem. The training set contains over 5000 descriptions of customers, including the information about whether or not they have a caravan insurance policy. A test set contains 4000 customers of whom only the organisers know if they have a caravan insurance policy [14]. The performance was measure by utilizing the performance metrics Hiding Failure (HF), Misses Cost (MC), Dissimilarity (Diss) and Artificial patterns (AP). The Table 1 shows the performance for the both data sets.

Hiding Failure (HF): It specifies the number of sensitive rules which can still be explored by the rule extraction algorithm. It can be calculated by the relation between the number of sensitive rules in sanitized database and the number of sensitive rules in the original database. It can be calculated as follows:

$$HF = \frac{|R_s(D')|}{|R_S(D)|}$$

Where $|Rs(D')|$ and $|Rs(D)|$are the number of sensitive rules extracted from modified database D' and the original database D, respectively.

Misses Cost (MC): This performance measure is used to show the percentage of the non-sensitive rules that are hidden as a side-effect of the sanitization process. The misses cost is calculated as follows

$$MC = \frac{|{\sim}R_s(D)| - |{\sim}R_s(D')|}{|{\sim}\pi R_s(D)|}$$

Where ~ R*s* (*D*) denotes the numbers of nonsensitive rules discovered from the original database *D*, and ~ R*s* (*D'*) denotes the number of non-sensitive rules discovered from modified database (D′).

Dissimilarity (Diss): the dissimilarity measure is calculated according to the formula as follow:

$$Diss(D, D') = \frac{1}{\sum_{i=1}^{n} f_D(i)} \times \sum_{i=n}^{n} [f_D(i) - f_{D'}(i)]$$

Where *fD* (*i*) denotes repetition of the *i*-th item in the database *D*, and *n* is the number of different items in the initial database *D*.

Artificial Patterns (AP): This performance factor is used to measure the percentage of the extracted rules that are ghost. It can be calculated as follows [4]:

$$AP = \frac{|R'| - |R \cap R'|}{|R'|}$$

Where, $|R'|$ and $|R|$ are the numbers of rules extracted from D' and D, respectively.

Table 1. Evaluation of data hiding method for both customer data sets.

	Metric							
	HF %		MC %		DISS %		AP %	
Rules	Data 1	Data 2	Data 1	Data 2	Data 1	Data 2	Data 1	Data 2
25	2	1	70.3	75.2	0.25	0.15	0	0
20	1	0	68.9	72.3	0.25	0.10	0	0
15	1	0	68.8	70.2	0.20	0.10	0	0
10	1	0	65.5	68.1	0.15	0.10	0	0
05	1	0	60.4	65	0.10	0.10	0	0

5. CONCLUSION

This paper implemented a data hiding algorithm for the customer relationship data set. The result shows that algorithm is rule-oriented and enables to hide the sensitive association rules. The result shows that missing cost is higher when the number of data is large compare to less data.

REFERENCES

[1] R. Agrawal, R. Srikant, Privacy preserving data mining", ACM SIGMOD International Conference on Management of Data, Vol. 29, pp. 439-450, 2000.

[2] Y. Lindell, B. Pinkas, Privacy preserving data mining", Proceedings of the CRYPTO, pp. 36-54, 2000.

[3] V. Verykios, E. Bertino, I.G. Fovino, L.P. Provenza, Y. Saygin, and Y. Theodoridis, State-of-the-art in Privacy Preserving Data Mining", SIGMOD Record, Vol. 33, pp. 50-57, 2004.

[4] D. Agrawal, C. Aggarwal, On the design and quantification of privacy preserving data mining algorithms", Proceedings of the 20th Conference on Principles of Database Systems, pp. 247-255, 2001.

[5] C. Clifton, M. Kantarcioglu, X. Lin, M. Zhu, Tools for privacy preserving distributed data mining", Proceedings of the SIGKDD Explorations, Vol. 4, pp. 28-34, 2002.

[6] E. Dasseni, V.S. Verykios, A. Elmagarmid, E. Bertino, Hiding association rules by using con_dence and support", Proceedings of 4th Information Hiding Workshop, pp. 369-383, 2001.

[7] S. Oliveira, O. Zaiane, Privacy preserving frequent itemset mining", Proceedings of the IEEE 14th

International Conference on Data Mining, Vol. 14, pp. 43-54, 2002.

[8] S. Oliveira, O. Zaiane, Algorithms for balancing privacy and knowledge discovery in association rule mining", Proceedings of the 7th International Database Engineering and Applications Symposium, pp. 54-63, 2003.

[9] S. Oliveira, O. Zaiane, Protecting sensitive knowledge by data sanitization", Proceedings of the IEEE 3rd International Conference on Data Mining, pp. 613-616, 2003.

[10] Y. Saygin, V.S. Verykios, C. Clifton, Using unknowns to prevent discovery of association rules", SIGMOD Record, Vol. 30, pp. 45-54, 2001.

[11] Evfimievski, R. Srikant, R. Agrawal, J. Gehrke, Privacy preserving mining of association rules", Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 217-228, 2002.

[12] Evfimievski, Randomization in privacy preserving data mining", Proceedings of the SIGKDD Explorations, Vol.4, pp. 43-48, 2002.

[13] Evfimievski, J. Gehrke, R. Srikant, Limiting privacy breaches in privacy preserving data mining", PODS, pp. 211-222, 2003.

[14] M. Kantarcioglu, C. Clifton, Privacy-preserving distributed mining of association rules on horizontally partitioned data", ACM SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery, 2002.

[15] J. Vaidya, C.W. Clifton. Privacy preserving association rule mining in vertically partitioned data", Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 639-644, 2002.

[16] Wang SL, Jafari A. Hiding sensitive predictive association rules. Systems, Man and Cybernetics. 2005; 1:164–9.

[17] Wang SL, Parikh B, Jafari A. Hiding informative association rule sets. Expert Systems with Applications. 2007; 33:316– 23.

[18] Modi CN, Rao UP, Patel DR., Maintaining privacy and data quality in privacy preserving association rule mining, Computing Communication And Networking Technologies. 2010; 32:1–6.

[19] Shah K, Thakkar A, Ganatra A. Association rule hiding by heuristic approach to reduce side effects and hide multiple R.H.S. items. International Journal of Computer Applications. 2012; 45:1–7.

[20] Domadiya NH, Rao UP. Hiding sensitive association rules to maintain privacy and data quality in database, Advance Computing Conference (IACC). 2012; 32:1306–10.

[21] Jain YK, Yadav VK, Panday GS. An efficient association rule hiding algorithm for privacy preserving data mining, International Journal on Computer Science and Engineering. 2011; 3:2792–8.

[22] Vijayarani S, Prabha MS. Association rule hiding using artificial bee colony algorithm, International Journal of Computer Applications. 2011; 33:41–7.

[23] Oliveira SM, Za¨ıane OR. Algorithms for balancing privacy and knowledge discovery in association rule mining. Seventh International Database Engineering and Applications Symposium, 2003. Proceedings. 2003; 56:54– 63.

[24] Duraiswamy K, Manjula D, Maheswari NA. New approach to sensitive rule hiding. Stud Comp Intell. 2008; 1:107–11.

[25] Menon S, Sarkar S, Mukherjee S. Maximizing accuracy of shared databases when concealing sensitive patterns. Information System Research. 2005; 16:256–570.

[26] Verykios VS, Pontikakis ED, Theodoridis Y, Chang L. Efficient algorithms for distortion and blocking techniques in association rule hiding, Distributed and Parallel Databases. 2007; 22:85–104.

[27] Amiri A. Dare to share: Protecting sensitive knowledge with data sanitization. Decision Support Systems. 2007; 43:181–91.

[28] Dehkordi MN, Badie K, Zadeh AK. A novel method for privacy preserving in association rule mining based on genetic algorithms. Journal of Software. 2009; 4:555–62.

[29] Ramakrishnan M. Switch pattern encryption based WBAN security in an IOT environment. Indian Journal of Science and Technology. 2015; 8:67–98. DOI: 10.17485/ijst/2015/ v8i34/85274.

[30]Farsad Zamani Boroujeni and Doryaneh Hossien Afshari,An Efficient Rule-Hiding Method For Privacy Preserving in Transactional Databases, Journal of Computing and Information Technology, Vol. 25, No. 4, December 2017, 279–290