A NOVEL ALGORITHM FOR PPDM TO HIDE ASSOCIATION RULE

[1]Parth Prajapati, [2]Prof. Mahesh Panchal
[1]KIRC,Kalol,Gujarat,India
[2]Head CE Dept.,KIRC, Kalol

***Abstract*-Data mining is the field that use the available data sets and extract important unknown pattern and knowledge which help in decision making or strategic design. Because of increasing data storage capacities and sharing data among parties, those may be untrusted parties, Privacy Preserving in data mining (PPDM) is important topic on which lots of research going on recent years. So there are many approaches and techniques are available for PPDM. In this paper a method is proposed which is more efficient to hide association rule. The efficiency is compare with previous algorithms in terms on CPU Time, no of modified entries and false generation of rules.**

***Keywords:* Data mining, PPDM, Association Rule, Sensitive Item, Association Rule Hiding.**

I. INTRODUCTION

Data mining is the process of extracting hidden information from the datasets. the mining of these datasets with existing data mining tools [2] can reveal invaluable knowledge that was unknown to the data holder beforehand. The extracted knowledge patterns can provide insight to the data holders as well as be invaluable in tasks such as decision-making and strategic business planning. A number of successful techniques have been proposed to obtain valid result while privacy preserving important data. So in this paper we review existing techniques. In huge applications data mining technologies have been raised about securing information against unauthorized access is an important goal of database security and privacy. Privacy is the term that is associated with mining task so we are able to hide sensitive information those are directly affect identity, which we don’t want to disclose to the public. So Privacy Preserving in data mining is the process of protect private information or sensitive knowledge from leaking in the mining process, meanwhile obtains more accurate result. As and when data mining techniques are developed and increase capacity to store private data, Privacy Preserving becomes an important issues. So issue of Privacy Preserving in Data Mining emerged globally.

In our work we are concern of hiding certain association rules which contain some sensitive information which are on the Right hand side or left hand side of the rule, so that rules containing confidential item can’t be discover. Our approached is based on modifying the database in a way that confidence of the association rule can be reduce with the help increase or decrease the support value of RHS or LHS correspondingly. As the confidence of the rule is reduce below a specified threshold, it is hidden or we can say it will not be disclosed.

Our method is based on [7] proposed two algorithms namely ISL (Increase Support of Left hand side) and DSR (Decrease Support of Right hand side) to hide useful association rule from transactions data with binary attributes. In ISL method, confidence of a rule is decreased by increasing the support value of Left Hand Side (L.H.S.) of the rule. For this purpose, only the items from L.H.S. of a rule are chosen for modification. In DSR method, confidence of a rule is decreased by decreasing the support value of Right Hand Side (R.H.S.) of a rule. For this purpose, only the items from R.H.S. of a rule are chosen for modification. The reminder of this paper is organized as follows. Section 2 presents the statement of the problem and the notation

Used in the paper. Section 3 presents the proposed algorithms for hiding informative association rule. Section 4 shows example of the proposed algorithms. Section 5 shows the experimental results of the proposed algorithms. Section 6 shows the analysis of the proposed algorithm. Conclusion and future works are described in Section 7.

II. BACKGROUND AND RELATED WORK

The problem of mining association rules was introduced in [8]. The problem of mining association rules is to find all rules that are greater than the user-specified minimum support threshold and minimum confidence threshold.

Association rule using support and confidence can be defined as follows. Let I = {i1, i2… in} be a set of literals, called items. Database D= {T1, T2,T3,…,Tn) is a set of transactions, where each transaction T is a set of items such that T ⊂ I, an association rule is an expression, X → Y where X ⊂ I, Y ⊂ I and X ∩ Y = Ø. The X and Y are called correspondingly the body (left hand side) and head (right hand side) of the rule. An example of such a rule is that 90% of customers buy milk also buys bread. The 90% here is called the confidence of the rule, which means that 90% of transaction that contains X also contains Y. The confidence c is calculated as $|X \cup Y|/|X| \geq c$. The support s of the rule is the percentages of transactions that contain both X and Y, which is calculated as $|X \cup Y|/|D| \geq s$. In other words, the confidence of a rule measures the degree of the correlation between item sets, while the support of a rule measures the significance of the correlation between item sets. We consider user specified thresholds for support and confidence, MST (minimum support threshold) and MCT (minimum confidence threshold).

There are many approaches have been proposed to preserve privacy for crucial knowledge or sensitive association rules in database. They can be classified in to following classes: Heuristic based, these approaches can be further divided in to two groups based on data modification techniques: data distortion techniques and data blocking techniques. Data distortion techniques try to hide association rules by decreasing or increasing support. To increase or decrease support, they replace 0’s by 1’s or vice versa in selected transactions. So they can be used to address the complication issue. But they produce undesirable side effects in the new database, which lead them to suboptimal solution [9]. The method of reduce the side effects in sanitized database, which are produced by other approaches [10]. An efficient clustering based approach [11] to reduce the time complexity of the hiding process. Data blocking techniques replace the 0 and 1 by unknowns “?” in selected transaction instead of inserting or deleting items. So it is difficult for an opponent to know the value behind “?”. First introduce blocking based technique [12] for sensitive rule hiding. Border based approaches, these use the notion of borders introduced in [13]. These approaches preprocess the sensitive rules so that minimum numbers of rules are given as input to hiding process. So, they maintain database quality while minimizing side effects. Hiding process in greedily [14] selects those modifications that lead to minimal side effects. Reconstruction based approaches generate [15] privacy aware database by extracting sensitive characteristics from the original database. These approaches generate minor side effects in database than heuristic approaches. Mielikainen [16] was the first analyzed the computational complexity of inverse frequent set mining and showed in many cases the problems are computationally difficult. Cryptography based approaches used in multiparty computation. If the database of one organization is distributed among several sites, then secure computation is needed between them. These approaches encrypt original database instead of distorting it for sharing. So they provide input privacy. Vaidya and Clifton [17] proposed a secure approach for sharing association rules when data are vertically partitioned. The secure mining of association rules over horizontal partitioned data. Many researchers have worked on the basis of reducing the support and confidence of sensitive association rule. ISL and DSR are the common approaches used to hide the sensitive rules. Some of the researchers have used data perturbation techniques to modify the confidential data value in such a way that the approximant data mining results could be obtained from the modified version of the database. Our work also has the basis of reduction of confidence using increase or decrease support value of generated sensitive rule.

III. PROPOSED ALGORITHM

In order to hide an association rule, X → Y, we can either decrease its support or its confidence to be smaller than user-specified minimum support transaction (MST) and minimum confidence transaction (MCT). To decrease the confidence of a rule, we can either (1) increase the support o of X, the left hand side of the rule, but not support of X ∪ Y, or (2) decrease the support of the item set X ∪ Y . For the second case, if we only

decrease the support of Y, the right hand side of the rule, it would reduce the confidence faster than simply reducing the support of X ∪ Y. To decrease support of an item, we will modify one item at a time by changing from 1 to 0 or from 0 to 1 in a selected transaction. Based on these two concepts, we propose a new association rule hiding algorithm for hiding sensitive items in association rules. In our algorithm, a rule X → Y is hidden by decreasing the support value of X ∪ Y and increasing the support value of X. That can increase and decrease the support of the LHS and RHS item of the rule correspondingly. This algorithm first tries to hide the rules in which item to be hidden i.e., X is in right hand side
and then tries to hide the rules in which X is in left hand side. For this algorithm t is a transaction, T is a set of transactions, R is used for rule, RHS (R) is Right Hand Side of rule R, LHS (R) is the left hand side of the rule R, Confidence (R) is the confidence of the rule R, a set of items H to be hidden.

ALGORITHM:

INPUT:

A source of Database D, Minimum Confidence, Minimum Support, Sensitive Item X.(Hide item X)

OUTPUT:

Sanitized database D', So that rules containing item X weather it is on L.H.S or R.H.S are not discovered.

```
Steps:
1. Generate all possible rules.
2. Select all rules containing sensitive item X on R.H.S
3. Calculate confidence of all selected rules.
4. If confidence R < Minimum Confidence.
        1. Go to step 2,
        2. Else go to step 5.
5. Decrease support of RHS
6. While (T is not Empty)
        1. Choose t form T.
        2. Modify value from 1 to 0.
        3. Save transaction data set.
7. For all rules contains sensitive item on LHS.
8. Calculate confidence of all selected rules.
9. If confidence R < Minimum Confidence.
        1. Go to step 10,
        2. Else go to step 13.
10. Increase support of L.H.S.
11. While (T is not Empty).
12. Choose t form T.
13. Modify value from 1 to 0.
14. Save transaction data set.
15. End while
```

IV. EXAMPLE

This section shows an example of the proposed algorithm in hiding sensitive item in association rule mining. Consider Table 1 as a database, MST=30%, MCT=70%, each element has value 1 if the corresponding item is supported by the transaction and 0 otherwise.

Tid	Item	Size	abc
T1	abc	3	111
T2	abc	3	111
T3	abc	3	111
T4	a	1	100
T5	ac	2	101
T6	ab	2	110

Table 1: Binary Dataset

Consider all possible rules with confidence are: A→B (66%), A→C (66%), B→A (100%), B→C (75%), C→A (100%), C→B (75%). Suppose we first want to hide item A, first take rule in which A is in RHS. These rules are B→A and C→A both has greater confidence from MCT. First take rule B→A search for transaction which support both B and A, B=A=1. There are four transactions T1, T2, T3, T4 with A=B=1. Now update table put 0 for item A in all four transactions. Now calculate confidence of B→A, it is 0% which is less than MCT so now this rule is hidden. Now take rule C→A, search for transaction in which A=C=1, only transaction T6 has A=C=1, update transaction by putting 0 instead 1 in place of A. Now take the rules in which A is in LHS. There are two rules A→B and A→C but both rules have confidence less than MCT so there is no need to hide these rules. So Table 2 shows the modified database after hiding item A.

After Changing database entry

Tid	Size	abc
T1	2	011
T2	2	011
T3	2	011
T4	0	000

T5	2	101
T6	2	110

Table 2 Modify data entries in Original dataset

V. RESULT & DISSCUSION

We have perform all the experiments on a system with core i3 processor, 2GB RAM, under Windows 7. Here a binary dataset is used named “LUng Cancer”[20]. We perform some experiments to compare the performance of Propose work with previous algorithm’s result. For all association rules are generate by various minimum support and confidence. Minimum support is range from 20% to 40% and minimum confidence range in 60% to 80%. The experiments results in CPU Time in millisecond(s), number of modified entries. In experiments Minimum Confidence=80% and support values are 20,30 and 40 for transactions 1000,1500 and 2000.

First Experiments done with minimum confidence 70%. Following 2 chart 1 & 2 shows comparison

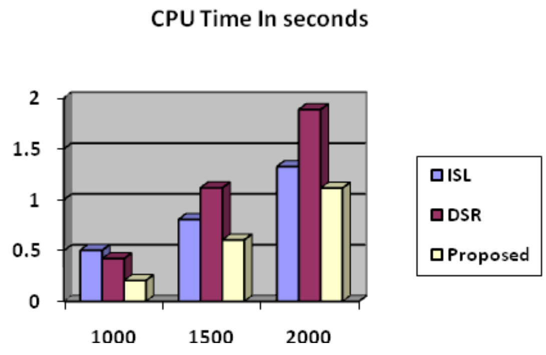

Chart -1

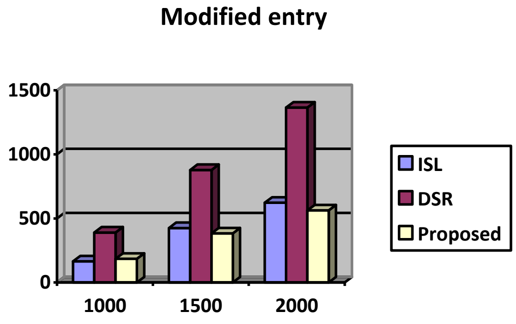

Chart-2

Second experiments done with minimum support 30% and chart 3 & 4 compare in terms of CPU Time & modified entry**.**

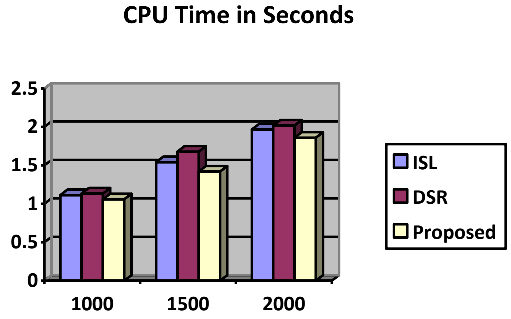

Chart-3

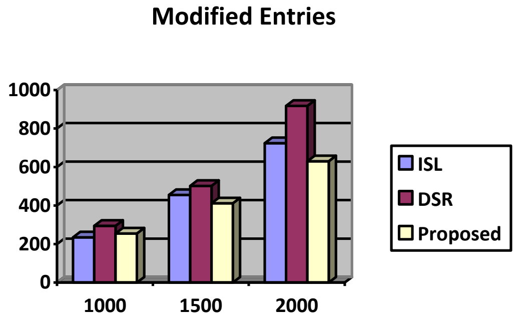

Chart-4

Finally following table shows false rules generated

Conclusion

The increasing ability to track and collect large amounts of data with the use of current hardware technology has lead to an interest in the development of data mining algorithms which preserve user privacy. With the development of data analysis and processing technique, the privacy disclosure problem about individual or company is inevitably exposed when releasing or sharing data to mine useful decision information and knowledge, then give the birth to the research field on privacy preserving data mining. In this paper, we carries out a wide 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 At present a variety of privacy preserving data mining algorithms are still some shortcomings, and are targeted at specific applications. and data sets, rather than to be extended to the general. The premise of ensuring the privacy of how to reduce the loss of accuracy, how to further improve the algorithm efficiency and privacy preserving generality in different types, distribution characteristics of different data sets are the direction of the future worthy of further study.

VI. REFERANCE

[1] Evfimievski, A., R. Srikant, R. Agrawal and J. Gehrke. “Privacy preserving mining of association rules”. Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, July 23-25, ACM Press, Edmonton, AB., Canada, pp. 1-12, 2002.

[2] Saygin, Y., V.S. Verykios and A.K. Elmagarmid. “Privacy preserving association rule mining”. Proceedings of the 12th International Workshop on Research Issues in Data Engineering: Engineering E-Commerce/E-Business Systems, Feb. 24-25, IEEE Xplore Press, San Jose, CA. USA., pp. 151-158, 2002.

[3] Oliveira, S., & Zaiane, O. “Privacy preserving frequent itemset mining”. In Proceedings of IEEE international conference on data mining,
November pp. 43–54, 2002.

[4] Ali Amiri, “Dare to share: Protecting sensitive knowledge with data sanitization”, Decision Support System archive vol. 43, issue 1, pp.181- 191, 2007.

[5] Stanley Robson de Medeiros Oliveira. “Data Transformation For Privacy-Preserving Data Mining”. University of Alberta,2005

[6] Jian Wang, Yongcheng Luo, Yan Zhao and Jiajin Le. “A Survey on Privacy Preserving Data Mining”. First International Workshop on Database Technology and Applications, DOI 10.1109/DBTA.2009.147, 2009.

[7] Aris Gkoulalas-Divanis and Vassilios S. Verykios. “An Overview of Privacy Preserving DataMining”. Summer 2009/Vol. 15, No. 4,Page-23-26

[8] Wei Zheng and Qingshui Li. ” Privacy Preserving Research Based on Association Rule Algorithm”. IEEE 978-1-61284-486-2/111 ©2011.

[9] Tinghuai Ma, Sainan Wang and ZhongLiu “Privacy Preserving Based on Association Rule Mining”. 3rd International Conference on Advanced Computer Theory and Engineering(ICACTE) 2010.

[10] Yongcheng Luo, Yan Zhao and Jiajin Le. “A Survey on the Privacy Preserving Algorithm of Association Rule Mining”. Second International Symposium on Electronic Commerce and Security,2009.

[11] Lambodar Jena, Ramakrushna Swain “A Comparative Study on Privacy Preserving Association Rule Mining Algorithms” International Journal of Internet Computing, Volume-I, Issue-1, 2011.

[12] Dr. K. Duraiswamy, N. Maheswari. “Identification of Sensitive Items in Privacy Preserving - Association Rule Mining”.Computer & Information Science. Vol 1 No. 2 ,May 2008.

[13] Haisheng Li.” Study of Privacy Preserving Data Mining”. Third International Symposium on Intelligent Information Technology and Security Informatics.2010 IEEE.

[14] Li Xiaohui.” The Study on Privacy Preserving Data Mining for Information Security”. 2011 International Conference on Future Information Technology IPCSIT vol.13 (2011) © (2011) IACSIT Press, Singapore

[15] Guanling Lee, Yi Chun Chen. “Protecting sensitive knowledge in association patterns mining”. WIREs Data Mining and Knowledge Discovery. Volume 2, January / February 2012

[16] Elisa Bertino, Dan Lin, and Wei Jiang A Survey of Quantification of Privacy Preserving Data Mining Algorithms”

[17] Elisa Bertino, Igor Nai Fovino. “A Framework for Evaluating Privacy Preserving Data Mining Algorithms” Data Mining and Knowledge Discovery, 11, 121–154, 2005

Books

[18] Privacy-Preserving Data mining: Models and Algorithm by Charu C. Aggarwal and Philip S. Yu.

[19] Data Mining Concepts and Techniques by Jiawei Han and Micheline Kamber.

Websites

[20] http://researchers.lille.inria.fr/~freno/datasets.html

[21] http://www.cs.umb.edu/~laur/ARtool/

[22] www.ics.uci.edu/~mlearn/

Dissertation

[23] Ahmed HajYasien PH.D Thesis “Preserving Privacy in Association Rule Mining” Griffith University June 2007

[24] Stanley Robson de Medeiros Oliveira “Data Transformation For Privacy-Preserving Data Mining” University of Alberta 2005