A IMPROVED PRIVACY PRESERVING ALGORITHM USING ASSOCIATION RULE MINING IN CENTRALIZED DATABASE

Archana Tomar[1], Vineet Richhariya[2], Mahendra Ku. Mishra[3]
Department of Information & Technology, Lakshmi Narain College of Technology (LNCT), Bhopal
1archanatomar.lnct@gmail.com, 2vineet_rich@yahoo.co.in, 3mahendramishra.mishra84@gmail.com

Abstract

The recent advancement in data mining technology to analyze vast amount of data has played an important role in several areas of Business processing. Data mining also opens new threats to privacy and information security if not done or used properly. The main problem is that to hide sensitive information, including personal information, fact or even patterns which are generated by any algorithm of data mining from the others. In order to focusing on privacy preserving association rule mining, the simplistic solution to address the problem of privacy is presented. To overcome these problems, we propose a algorithm named improved Privacy Preserving Algorithm using Association Rule Mining which is based on the random Perturbation technique which is best in efficiency and performance. This method is suitable to the any type of data. Our algorithm is a good way to apply data mining techniques with security that hides our logical instances from others.

Keywords— Data Mining, Association Rule Mining, Privacy Preserving.

Introduction

The association rule mining has received a great deal of attention. It is still one of most popular pattern discovery methods in the field of data mining. Various proposals and algorithms have been designed for it in recent years. Simultaneity, Data mining algorithms are analyzed for the side-effects which incur in data privacy. Thus, several privacy-preserving techniques for association rule mining have also been proposed in the past few years. Various proposed algorithms have been developed for centralized data, while others refer to a distributed data scenario. Distributed data scenarios can also be classified as horizontal data distribution and vertical data distribution. Data mining technology can analyze massive data. Although it plays vital role in many domains, if it is used improperly it can also cause some new problem of information security. There are some new problems in the application of data mining recently. By studying deep in some special algorithms with association rule mining, some techniques also can be applied to other data mining computations, such as decision tree inducers, association rule mining algorithms, clustering algorithms, rough sets and Bayesian networks etc.

Fast increasing of a series of digitized data causes people of the world attend the privacy problem of information more and more. Because the data mining technology of traditional centralized database must collect all the data together to process, it will cause the individual information abused or misused easily. Therefore more and more people will not to provide individual privacy data and suspect the using of data mining. Some people mine the privacy information pattern of the database owner from the original data. It has harmed the database owner's benefit. In order to solve the privacy preserving problem of association rule in centralized database, before publishing database we should hide the privacy or the sensitive information pattern of the database owner including the sensitive association rule information. Usually we use disturbing data method to change the data of original database to hide association rule. But the data disturbance may generate some information pattern that is not existed at all or reduce the accuracy of the original database. Before executing the privacy preserving algorithm, we should analyze the information pattern of association rule and the data structure of the database and find the preferred plan to keep the balance between the accuracy of database information and the privacy of sensitive information. However, data mining also brings some problems. For example, credit card centres may intentionally or unconsciously make sensitive information of clients leak while mining relating information of clients. With the Internet popularity, because more and more information can be obtained in electronic form, that people have their own privacy confidential is becoming increasingly urgent. We provide here an overview of privacy preserving association rule mining. The rest of this paper is arranged as follows: Section 2 introduces Association rule mining strategies; Section 3 describes about Privacy Preserving Algorithm; Section 4 shows the evolution and recent scenario; Section 5 describes the proposed method. Section 6 describes conclusion and prospect.

Association Rule Mining

The association rule mining can be conceptualized as follows [4]: Let I= {i1,i2,…,im} be the set of all items. Let D, the task-relevant data, be a set of database transactions where each transaction T is a set of items such that T is subset of I. Each transaction is associated with an identifier, called TID. Let X be a set of items. A transaction is said to contain X if and only if X is subset of T. An association rule is an implication of the form X is subset of Y, where X is subset of T, Y is subset of T, X∩Y=NULL. The support count of an itemset is the number of transactions containing the itemset. An itemset is frequent if its support count is not less than the minimum support count.Rules with the support more than a minimum support threshold (smin) and the confidence more than a minimum confidence threshold (cmin) are called strong. Association rule mining is a two-step process: (1) Finding all frequent itemsets; (2) Generating strong association rules from the frequent itemsets.
The purpose of privacy preserving is to discover accurate patterns without precise access to the original data. The algorithm of association rule mining is to mine the association rule based on the given minimal support and minimal confidence. Therefore, the most direct method to hide association rule is to reduce the support or confidence of the association rule below the minimal support of minimal confidence. With regard to association rule mining, the proposed methodology that is effective at hiding sensitive rules is implemented mainly by depressing the support and confidence. The existing three algorithms, D_CONF1, D_CONF2 and D_SUPP, which are to hide the sensitive association rule all by reducing the support or confidence.

Privacy Preserving Algorithm

A lot of implementations of the confidentiality of data and knowledge are applied in association rule mining process. According to privacy protection technologies, at present, privacy preserving association rule mining algorithm commonly can be divided into three categories : a) Heuristic based techniques b) Reconstruction based approaches c) Cryptography based approaches Heuristic-based techniques are to resolve how to select the appropriate data sets for data modification. Heuristics can be used to address the complexity issues. The methods of Heuristic-based modification include perturbation, 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), and blocking, which is the replacement of an existing attribute value with a "?". There is a basic principle of choosing the transaction or the item of itemset to be modified that we should reduce the influence of the original database as far as possible. Those related works are given below.

A. Data Perturbation-Based Association Rule

Source database, R be a set of significant association rules that can be mined from D, and let Rh be a set of rules in R. How can we transform database D into a database D', the released database, so that all rules in R can still be mined from D', except for the rules in Rh. The heuristic proposed for the modification of the data was based on data perturbation, and in particular the procedure was to change a selected set of 1- values to 0-values, so that the support of sensitive rules is lowered in such a way that the utility of the released database is kept to some maximum value. Therefore, the key question of this algorithm is how to put D into D' with the use of heuristic thought.

Recommended font sizes are shown in Table 1.

B. Data Blocking-Based Association Rule

The approach of blocking is implemented by reducing the degree of support and confidence of the sensitive association rules. That is by replacing certain attributes of some data items with a question mark or a true value. In this regard, the minimum support and minimum confidence will be altered into a minimum support interval and a minimum confidence interval correspondingly. As long as the support and/or the confidence of a sensitive rule lie below the middle in these two ranges of values, then we expect that the confidentiality of data is not violated.

We should choose the algorithm according to the different situation that can reduce the influence for the original database as far as possible. D_CONF1 algorithm will increase one transaction item when it circulates one time. Usually there are much data and many association rules while it processing the problem of privacy preserving. Frequent using of D_CONF1 algorithm will increase the quantity of data and generate some association rules that do not exist. It has influenced the accuracy of database. If there are some important items that cannot be modified or deleted, D_CONF1 algorithm is suitable for this situation. The front parts of D_CONF2 and D_SUPP are same. D_CONF2 algorithm can only select sacrifice item in back-end itemset and D_SUPP algorithm can select sacrifice item in whole generated itemset. So if the influence for the original database of selecting sacrifice item in front-end is smallest we can choose D_CONF2 algorithm. If the influence for the original database of selecting sacrifice item in back-end is smaller, we

can choose algorithm through comparing the efficiency, the quantity and importance of selected sacrifice item and support of D_CONF2 and D_SUPP algorithm. We should analyze the transaction set of the original database and the sensitive association rule set to be hidden and find the relation of them.

So we could select the sensitive transaction and sacrifice item more efficient. The modified data will be fewer and the influence for the original database will be smaller.

Evolution and Recent Scenario

In [1] proposed a new algorithm for balance privacy preserving and knowledge discovery in association rule mining. The solution is to implement a filter after the mining phase to weed out or hide the restricted discovered association rules. Before implementing the algorithms, the data structure of database and sensitive association rule mining set have been analyzed to build the more effective model. In [3] proposed a privacy preserving association rule mining into three categories: heuristic-based techniques, reconstruction-based techniques, cryptography-based techniques. Finally, they conclude further research directions of privacy preserving algorithms of association rule mining by analyzing the existing work.

In [4] proposed the algorithm which is designed to solve the shortage of low privacy protection of the geometric transform algorithm. The algorithm first gives four parameters, corresponding to the probability of four different types of geometric transformations. According to the various random number generated, different geometric transformation method is selected, which serves the dual effect of privacy protection. In [5] proposed a framework involves several components designed to anonymize data while preserving meaningful or actionable patterns that can be discovered after mining. In contrast with existing works for traditional data-mining, this framework integrates domain ontology knowledge during DGH creation to retain value meanings after anonymization. In [6] proposed an algorithm provides privacy and security against involving parties and other parties (adversaries) who can reveal information by reading unsecured channel between involving parties.

In [7] proposed a privacy preserving association rule mining algorithm based on SRRCR is presented, which can achieve significant improvements in terms of privacy and efficiency. Finally, they present experimental results that validate the algorithms by applying it on real datasets.

Proposed Work

In this section, we describe the proposed method. We propose a algorithm named improved Privacy Preserving Mining. The entire system architecture consists of five phases: 1) Check for Authentication. 2) Encoded the data by using the random Perturbation technique 3) On the basis of decryption key we read the transaction 4) Perform Pruning 5) On the basis of decryption key we generate association rules Our algorithm is a good way to apply data mining techniques with security that hides our logical instances from others. Our algorithm shows good performance in different operating environment.

Algorithm: IPPM (Improved Privacy Preserving Mining)
Input:
A. Set of rules to hide the data values
B. The source database
C. A Key for visualizing the authentication.
Output:
D. The database (DB) transformed so that the set of rules are properly applied and produce the result with security.

```
IPPM(R, DB, Key)
        Begin
1. Check the Authentication
        a. Enter uid & pwd
        b. If ( uid == udb && password == pdb)
        {
c.      Welcome in the database
            SIPM (DB)
              User(entry)
                {
                    Log(id)
                }
        }
d.      Else
        {
            Not an authorized user
        }
        Exit (0)
2.  IPPM (DB)
    a. While (object. read ()! =-1)
        {
          [Start Reading]
          [Generate Tokens]
          TK1, Tk2…..Tkn
          [Token is generated according to the
           alphabet entered]
        If(,)
          {
            TK1, Tk2……Tkn
          }
          Else
```

```
        {
        [Enter the character]
         String a=Object.nextLine();
         STK1, STk2……STkn
            }
      }
b.  [compute the occurrences]
      For i=1 to n iterations do
      {
      Itemset[i]=count;
      Count++;
      }
c.  [Enter the minimum support]
      Check for authentication again
      Enter the min-sup key
      If(min-sup==msdb)
      {
      Prune (db, key)
      }
      Else
      {
      [Enter the value again]
      }
3.  Prune (db, key)
    a. enter the min-sup
      For i=1 to n do
      If(count[i]>min-sup)
      List=itemset[i];
      Else
      Remove from the list
4.  Add the final result
      Finish
```

Following example illustrates proposed privacy preserving mining algorithm. A sample transaction database D is shown in Table1. TID shows unique transaction number. Binary valued item shows whether an item is present or absent in that transaction. Suppose MST and MCT are selected 3 and 75% respectively. Table 2 shows frequent itemsets satisfying MST, generated from sample database D Table 1. Sample Transaction Database D.

Table 1
Sample Transactions Database D

TID	Items	Encrypted Items
1	a b c d e	1 2 3 4 5
2	a c d	1 3 4
3	a b d f g	1 2 4 6 7
4	b c d e	2 3 4 5
5	a b d	1 2 4
6	c d e f h	3 4 5 6 8
7	a b c g	1 2 3 7
8	a c d e	1 3 4 5
9	a c d h	1 3 4 8

In following, the possible number of association rules satisfying MST and MCT, generated by Apriori algorithm [2]: 2⇒1, 1⇒4, 4⇒1, 2⇒4, 3⇒4, 4⇒3, 5⇒3, 5⇒4, 12⇒4, 24⇒1, 13⇒4, 35⇒4, 45⇒3, 5⇒43. Suppose the rules 2⇒1, 2⇒4 and 3⇒4 specified as sensitive and should be hidden in sanitized database.

Table 2
Frequent itemsets with support count.

Frequent Itemsets with Support Count
1:7,2:5,3:7,4:8,5:4, 12:4,13:5,23:3, 14:6, 24:4, 34:6, 35:4, 45:4,124:3, 134:4, 345:4

Result

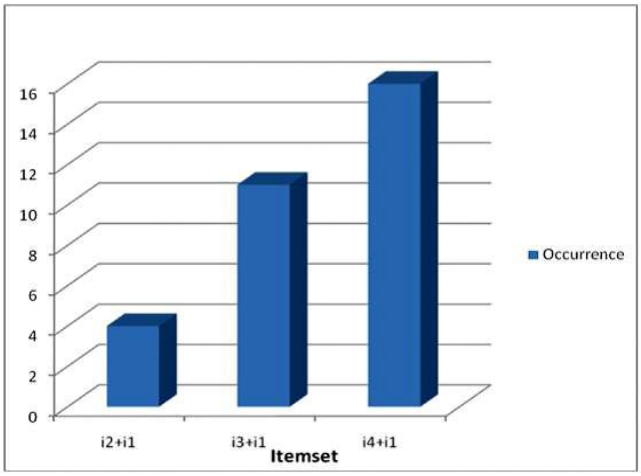

Figure 1 Item Associate V/s Occurrence

In this graph we can check the combination of item selling through its occurrence and maintain our stock

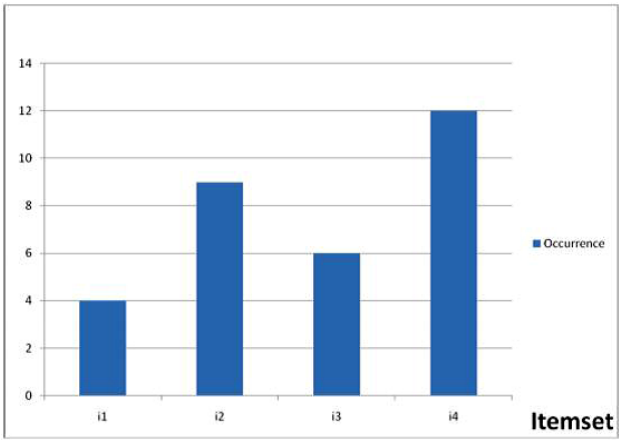

Figure 2 Itemset V/s Occurrence

In this graph we can check the single item selling through its occurrence and maintain our stock

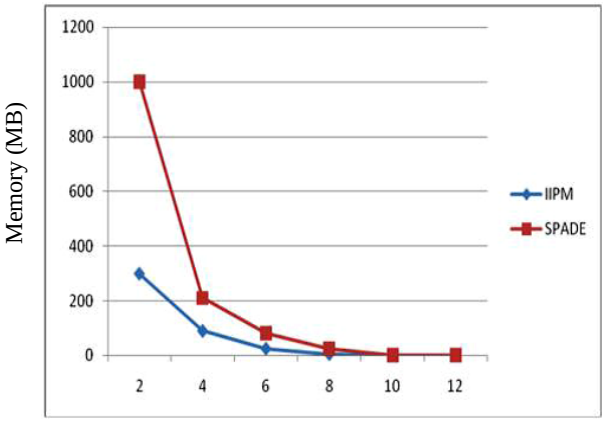

Figure 3 Memory Based graph

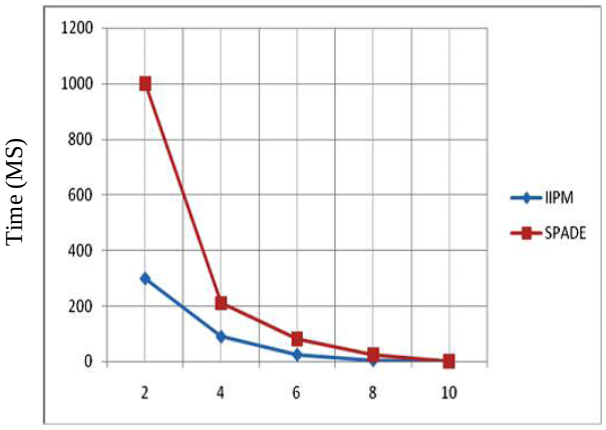

Figure 4 Time Based graph

Figure 3 and 4 shows that our algorithm takes less memory and less time because its work on post while Spade work on pre and post both.

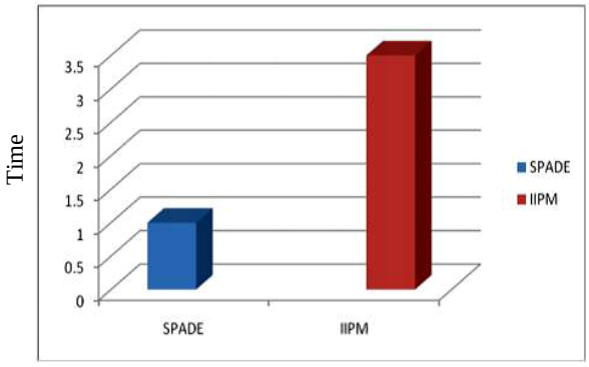

Figure 5 Comparison of Spade algorithm and IPPM algorithm at security point of view

The above graph shows that there is very less security in spade. But in our approach we apply three type of security:

1. Authentication
2. Session Key
3. Random perturbation (hide data at three levelstransaction, frequent item , association rule)

Conclusion and future work

The work presents in here, which indicates the ever increasing interest of researchers in the area of securing sensitive data and knowledge from malicious users. We propose a algorithm named Improved Privacy Preserving Mining (IPPM). The entire system architecture consists of five phases: 1) Check for Authentication. 2) Encoded the data by using random perturbation technique 3) On the basis of decryption key we read the transaction 4) Perform Pruning. 5) On the basis of decryption key we generate association rules. Our algorithm is a good way to apply data mining techniques with security that hides our logical instances from others. At present we are using the Perturbation technique for data encryption. In future we can use the SHA algorithm for encryption and also include the simulation result for showing the efficiency and performance.

References

[1] Agrawal R, Imielinski T, Swami A, "Mining association rules between sets of items in large databasesWashington, DC, 1993, pp.207–216.

[2] Chen M S, Yu P S. Data Mining:An Overview from a Database Perspective [J]. IEEE Trans on Knowledge and Data Engineering, 2004,8(6) :866-883.

[3] M K Reiter. Crowds:Anonymity for Web Transactions[J]. The ACM Transactions on Information and System Security,2005.S. Zhang, C. Zhu, J. K. O. Sin, and P. K. T. Mok, "A novel ultrathin elevated channel low-temperature poly-Si TFT," *IEEE Electron Device Lett.*, vol. 20, pp. 569–571, Nov. 1999.

[4] R. Agrawal, R. Srikant, "Fast algorithms for mining association rules,"In: Proc. 20th Int'l Conf. Very Large Data Bases, 1994.

[5] Shaofei Wu, Hui Wang, "Research On The Privacy Preserving Algorithm Of Association Rule Mining InCentralized Database,"International Symposiums on Information Processing, 2008.

[6] Vassilios S. Verykios, Elisa Bertino ， et al., " State-of-the-art in Privacy Preserving Data Mining," March 2004, pp.50-57.

[7] Elena Dasseni, Vassilios S. Verykios, Ahmed K. Elmagarmid, and Elisa Bertino, "Hiding Association Rules by using Confidence and Support," 2001.

[8] Stanley R. M. Oliveira and Osmar R. Zaiane, "Privacy preserving frequent itemset mining, In Proceedings of the IEEE ICDM Workshop on Privacy, Security and Data Mining 2002.

[9] E.T. Wang, G. Lee, Y.T. Lin, "A novel method for protecting sensitive knowledge in association rules mining," 2005.

[10] E.T. Wang, G. Lee, "An efficient sanitization algorithm for balancing information privacy and knowledge discovery in association patterns mining," 2008.

[11] S.R.M. Oliveira, O.R. Zaıane, Y. Saygin, "Secure association rule sharing, advances in knowledge discovery and data mining, in:PAKDD2004.Jun Lin Lin, Yung Wei Cheng, "Privacy preserving

[12] itemset mining 2009.

[13] Evfimievski, A., Srikant, R., Agrawal, R., & Gehrke, J., "Privacy preserving mining of association rules," 2002.

[14] Rizvi S J, Haritsa J R., "Maintaining data privacy in association rule mining," August 2002.

[15] Yucel Saygin, Vassilios Verykios, and Chris Clifton, "Using unknowns to prevent discovery of association rules," 2001.

[16] S Yucel Saygin, Vassilios S. Verykios, and Ahmed K. Elmagarmid,"Privacy preserving association rule mining," 2002.

[17] L. Sweeney, "k-anonymity: a model for protecting privacy",2002. Xiao X, Tao Y, "Personalized privacy preservation", 2006.

[18] LIU Ming, Xiaojun Ye, "Personalized K-anonymity", Computer Engineering and Design, Jan.2008.

[19] Rakesh Agrawal and Ramakrishnan Srikant, "Privacy-preserving data mining 2000.

[20] Dakshi Agrawal and Charu C. Aggarwal, 2001.

[21] Chris Clifton, Murat Kantarcioglu, Jaideep Vaidya,Xiaodong Lin, Michael Y. Zhu, 2002.

[22] Chris Clinton, „Privacy Preserving Distributed Data Miting", 2001.

[23] Murat Kantarcioglu, Chris Clinton, IEEE 2003.

[24] Rakesh Agrawal, Ramakrishnan Srikant, "Privacy- Preserving Data Mining", 2000.

[25] V. S. Verykios, E. Bertino, I. N. Fovino, L. P. Provenza, Y. Saygin, Y.Theodoridis, "State-of-the-art in Privacy Preserving DataMining",2004.

[26] Shipra Agrawal, Vijay Krishnan, Jayant Haritsa, "On Addressing Efficiency Concerns in Privacy Preserving Data Mining",DB/0310038.

[27] Rakesh Agrawal and Ramakrishnan Srikant, "Privacy-preserving data mining," 2000.

[28] Dakshi Agrawal and Charu C. Aggarwal, "On the design and quantification of privacy preserving data mining algorithms 2001.

[29] Rizvi S J, Haritsa J R., "Maintaining data privacy in association rule mining," August 2002.

[30] Alexandre Evfimievski, Ramakrishnan Srikant, Rakesh Agrawal, and Johannes Gehrke, 2002.

[31] Ioannidis, I.; Grama, A, Atallah, M., "A secure protocol for computing dot-products in clustered and distributed environments," 2002.

[32] Chris Clifton, Murat Kantarcioglou, XiadongLin, and Michael Y. Zhu, "Tools for privacy preserving distributed data mining 2002.

[33] Vaidya, J. & Clifton, C.W., "Privacy preserving association rule mining in vertically partitioned data," July 2002.

[34] ZongBo Shang; Hamerlinck, J.D., "Secure Logistic Regression of Horizontally and Vertically Partitioned Distributed Databases,"2007. Seventh IEEE

[35] GENG Bo,ZHONG Hong,PENG Jun,WANG Da-gang Temporal Rule Distribution Mining of Privacy-preserving_ 2008.

[36] Shaofei Wu and Hui Wang, IEEE International Symposiums on Information Processing, 2008.

[37] Yongcheng Luo,Yan Zhao and Jiajin Le ,Second International Symposium on Electronic Commerce and Security,2009 IEEE .

[38] Jie Liu and Yifeng XU,2009 Fourth International Conference on Internet Computing for Science and Engineering,IEEE.

[39] Brian, C.S. Loh and Patrick, H.H. Then,2010 Second International Symposium on Data, Privacy, and E-Commerce,IEEE.

[40] Chirag N. Modi, Udai Pratap Rao and Dhiren R. Patel,2010 International Conference on Advances in Communication, Network, and Computing,IEEE.

[41] Wang Yan,Le Jiajin and Huang Dongmei,2010 International Conference on Web Information Systems and Mining,IEEE.