International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169
Volume: 4 Issue: 1 99 - 102

An Efficient Approach to Privacy Preserving Association Rule Mining

Abhinav Anurag
M.tech Student
Suresh GyanViharUniversity,Jaipur

Dinesh Goyal
Professor &Principal(Engg.)
Suresh GyanViharUniversity,Jaipur

Abstract:-The vulnerabilities associated with large databases is increasing with the passage of time and sharing of data over a network becomes a critical issue for every organization. When we talk about data mining approaches,there has been a tremendous success. But when we see the other side of the coin, it has put the databases and its sensitive information on the verge of being modified or altered by unwanted sources. The major problem is still out in there in the middle and we need to create a balance between the data mining results with the appropriate time management to hide the data. The main focus should be on how we can keep our sensitive data private and the sensitive information could not be revealed through data mining techniques with ease. In this thesis, we focus on hiding the sensitive data with a much faster pace as compared to hiding counter algorithm.

Keywords*: Association Rule, hiding counter, support, confidence*

*****

1. INTRODUCTION:

In terms of definition, data mining refers to a significant process of recognizing logical, novel, potentially useful and ultimately lucid pattern in the data. If we look at the bigger picture, data mining comprises of number of fields which includes tool learning, database, data visualization, information theory and statistics. Data mining techniques can further be classified as classification,clustering, association, sequential patterns and prediction etc.

Various researchers and organization are using data sharing method in data mining and this sharing of data also comes with some complexities. The researchers or organization should be aware of the fact that the sensitive information from the large storehouse of data should not be used for unwanted purposes. The data mining techniques has inspired numerous requirements in data sharing, knowledge discovery and privacy preserving data mining which simultaneously saw various research work in data mining and database security fields.

When we talk about databases, we see them as complex processes and one of the process is knowledge discovery. Knowledge discovery can be subdivided into various steps:

Data Gathering- Data gathering can be done from different diversified data sources. It can be from various sources like databases, non-electronic sources etc.

Preprocessing of Data: The data which we collect from different sources can have incorrect or missing information and may have numerous types and metrics. So, these incorrect data are corrected and the missing one's are predicted or given at that time.

Transformation of Data: Since the data gathered can be of different types and metrics, it should be transformed suitable format which can be useful to all.

Data Mining: In this step, we apply the relevant algorithm to the converted data and discover the results.

Pattern Evaluation: It is one of the most important step of knowledge discovery because it decides how the data mining results can be represented to the users. The evaluation of pattern is very useful in deciding how efficiently the results can be used.

Data mining are also classified into various categories. In this thesis, we are working with privacy preserving data mining (PPDM) which is rather a new study of statistical databases and data mining. Two factors come into play in terms of PPDM. The first one states that the sensitive information should be transformed in order to maintain other person's privacy which can be tampered by the receiver. Secondly, the sensitive information should be kept out so that it cannot be mined with the help of data mining algorithms.

2. BACKGROUND AND RELATED WORK:

A data mining problem is that most studies found association rules from large databases process. Association rule mining is found set, often co-occur in the transaction database to generate processing on the data retention rules significantly associated projects. Most of the association rules existing algorithms rely on support - confidence framework.

Formally, the association rules are defined as follows: Let I = {I1, I2 ... Instant} is a set of items. Let D task-related data, it is a set of database transactions, where each transaction T is a set of such pieces of T⊆I. Each transaction has an

identifier, called TRXID related projects. Let X be a group project. Transaction T is said to contain if X⊆T. In the form of association rules is X => Y, where X⊂I, Y⊂I and X⋂Y = Φ meaning. Rule X => Y holds transaction set D support s, where s is the percentage of transactions in D. Rule X => Y set D with confidence c in the transaction, the percentage of A also contains a measure of B. While support is a regular frequency of D if c is included in the transaction, the strength of the confidence measures between sets of projects relationship.

Support (S) association rules are defined as comprising (X∪Y) to the percentage of the total number of records in the database records / min.

$$\text{Support}(X=>Y)= \frac{\text{SupportCount}(X\cup Y)}{\text{Number of transaction}}$$

Total confidence association rules are defined as a percentage of the transaction amount included / component (X∪Y) the record contains A.

$$\text{Confidence}(X=>Y)= \frac{\text{SupportCount}(X\cup Y)}{\text{SupportCount}(X)}$$

Clifton[10] provides an example, the database data mining algorithms which displays key information, business competitors. Clifton proposed a technique to prevent the disclosure of sensitive information by releasing only the raw sample data. This technique can be applied independent of specific data mining algorithms to be used. In later work, Clifton raised through distributed mining technology can be combined with business competitors database to extract association rules, the parties without violating data privacy data. The problem solved by the Lindell peace under Fidel [6], when the classification rules are to be extracted. Another way to extract association rules without violating privacy is to reduce support and / or these rules [2,4] confidence.

A large number of tasks related to the association rules hiding must be done. Reducing support and sensitive association rules [8, 9 and 11] on the basis of trust, the largest researchers have worked. ISL and DSR are used to hide sensitive rules of common practice. Actually hide any given specific rules, a number of methods used to hide association, classification and clustering rules have been proposed. Some researchers have used data perturbation techniques to modify the value of confidential data in such a way, similar to data mining results can be obtained from the modification of the database version. Some researchers also recognize the need for a variety of data mining analysis.

3. PROPOSED METHOD:

Input:

1. database transactions
2. database rules
3. sensitive items set ->X
4. minimum support threshold (MST)
5. minimum confidence threshold (MCT)

Output:
A converted database of transactions where rules containing X will be hidden.

Procedure:
Step 1: Inputs are in the form of Transaction Data Base, Rule Data Base, MCT (Minimum Confidence Threshold) and
Step 2: Sensitive information to be entered

Step 3: Now we find out the rule in the database which contains the sensitive information on the right hand side and we also see that the confidence is greater than MCT.
Step 4: For each rule (X -> Y) which contains a sensitive item on RHS OR LHS

LOOP:
If (support < = 90 percent) and (support >= 75)
Set confidence(X → Y) = (minconf * 3/4);
Set support (X → Y) = (minsup * 3/4);
Else
Set confidence(X → Y) = (minconf * 4/5);
Set support (X → Y) = (minsup * 4/5);
Step 5: Exit

4. RESULT :

Suppose there is a dataset of transaction from a furniture store:

TRXID	ITEM
t1	bed, chair, table
t2	chair
t3	bed, lamp, table
t4	bed, chair
t5	bed, chair, table

We have also been given a MST OF 20% and MCT of 20% and further four association rules can be deduced from the above transaction.

bed→ chair (60%, 75%)
chair→ bed(60%, 75%)
bed→ table(60%, 75%)
table→ bed (60%, 100%)

Now there is a need to hide **table** and **chair**.

Using The ISL Method

With the help of **ISL** algorithm if anyone want to hide **table** and **chair**, then he can check it by modifying the transaction t2 from **chair** to {**chair,table}** (i.e. from 0100 to 0101).but still ISL cannot hide the rule **table→ bed**. Let us see by following example

TRXID	Item	BitMap
t1	bed, chair, table	1101
t2	chair	0100
t3	bed, lamp, table	1011
t4	bed, chair	1100
t5	bed, chair, table	1101

(Hiding table→ bed by ISL approach)

TRXID	Item	BitMap
t1	bed, chair, table	1101
t2	chair	**0101**
t3	bed, lamp, table	1011
t4	bed, chair	1100
t5	bed, chair, table	1101

So it can be concluded that rule **table→ bed**cannot be hidden by ISL approach because by modifying t2 from **chair** to {**chair,table}** (i.e. from 0100 to 0101) rule **table→ bed** will have support and confidence 60% and 75% respectively.

By DSR approach:

TRXID	Item	BitMap
t1	bed, chair, table	1101
t2	chair	0100
t3	bed, lamp, table	1011
t4	bed, chair	1100
t5	bed, chair, table	1101

(Hiding table→ bed by DSR approach)

TRXID	Item	BitMap
t1	bed, chair, table	**0101**
t2	chair	0100
t3	bed, lamp, table	1011
t4	bed, chair	1100
t5	bed, chair, table	1101

DSR way through rule **table → bed** has been hidden support and confidence, which are now 40% and 66%, but as a side effect of the rule **table → bed** is also hidden.

Hiding Counter Method:

TRXID	Item
t1	bed, chair, table
t2	chair
t3	bed, lamp, table
t4	bed, chair
t5	bed, chair, table

	S	C	Hiding Counter
bed→chair	60%	75%	0
chair→bed	60%	75%	0
bed→ table	60%	75%	0
table→bed	60%	100%	0

S-Support
C-Confidence

Support (*table*→ bed) = 3/5
After first pass, Support (*table*→ bed) = 3/6 = 50 %
After second pass, Support (*table*→ bed) = 3/7 = 43 %
After third pass, Support (*table*→ bed) = 3/8 = 37.5 %
After Fourth pass, Support (*table*→ bed) = 3/9 = 33 %
After Fifth pass, Support (*table*→ bed) = 3/10 = 30 %.....
After 10 pass, Support (*table*→ bed) = 3/16 = 19 %
So rule is hidden after ten passes.

5. PROPOSED APPROACH:

After first pass, Support (*table*→ bed) = 50 * 4/5 = 40
After second pass, Support (*table*→ bed) = 40 * 4/5= 32
After third pass, Support (*table*→ bed) = 32 * 4/5 = 25
After fourth pass, Support (*table*→ bed) = 25* 4/5 = 20
After Fifth pass, Support (*table*→ bed) = 20 * 4/5 = 16
So rule is hidden in just 5 passes as compared to the 10 passes of the hiding counter method.

6. CONCLUSION:

To solve the problem of privacy preserving data mining problems, we have introduced a new algorithm. We have summarized various contribution in this thesis. We have proposed a new item based restriction technique which hides sensitive information. Along with this, we have compared hiding counter, DSR, ISL methods of association rule mining with their proposed algo. After comparing the above association rule mining approaches, we came to a conclusion that the item-sets or a particular sensitive information can be hidden faster than the hiding counter approach.

REFERENCES

[1] R. Agrawal, T. Imielinski, and A. Swami. *Mining association rules between sets of items in large databases*. In Proceedings of the ACM SIGMOD

Conference on Management of Data, pages 207–216, New York, NY, USA, May 1993. ACM Press.

[2] S. Goldwasser. *Multi-party computations: Past and present*. In Proceedings of the 16th Annual ACM Symposium on the Principles of Distributed Computing, pages 1–6, Santa Barbara, California, USA,1997. ACM Press.

[3] E. Dasseni, V. S. Verykios, A. K. Elmagarmid, and E. Bertino. *Hiding association rules by using confidence and support*. In I. S. Moskowitz,editor, Proceedings of the 4th Information Hiding Workshop, volume2137, pages 369–383, Pittsburg, PA, USA, April 2001. Springer VeralgLecture Notes in Computer Science.

[4] Y. Saygin, V. S. Verykios, and C. Clifton. *Using unknowns to prevent discovery of association rules*. In ACM SIGMOD Record, volume 30(4), pages 45–54, New York, NY, USA, December 2001. ACMPress.

[5] W. Du and M. J. Atallah. *Secure multi-party computation problems and their applications: A review and open problems*. In V. Raskin,S. J. Greenwald, B. Timmerman, and D. M. Kienzle, editors, Proceedings of the New Security Paradigms Workshop, pages 13–22, Cloudcroft, New Mexico, USA, September 2001. ACM Press.

[6] Y. Lindell and B. Pinkas. *Privacy preserving data mining*. InCRYPTO-00, volume 1880, pages 36–54, Santa Barbara, California,USA, 2000. Springer Verlag Lecture Notes in Computer Science.

[7] C. Clifton and D. Marks. *Security and privacy implications of data mining*. In Workshop on Data Mining and Knowledge Discovery,pages 15–19, Montreal, Canada, February 1996. University of BritishColumbia, Department of Computer Science.

[8] V. S. Verykios, A. K. Elmagarmid, B. Elisa, Y. Saygin, and D. Elena.*Association rule hiding*. In IEEE Transactions on Knowledge andData Engineering, volume 16(4), pages 434–447, Los Alamitos, CA,USA, April 2004. IEEE Computer Society.

[9] M. Atzori, F. Bonchi, F. Giannotti, and D. Pedreschi. *Blocking anonymity threats raised by frequent itemset mining*. In Proceedings of the 5th IEEE International Conference on Data Mining(ICDM'05), pages 561–564, Houston, Texa, USA, November 2005.IEEE Computer Society.

[10] Shyue-Liang Wang, Yu-Huei Lee, Steven Billis, AyatJafari *Hiding Sensitive Items in Privacy Preserving Association Rule Mining*, 2004. IEEE International Conference on Systems.

[11] Vi-Hung Wu, Chia-Ming Chiang, and Arbee L.P. Chen, SeniorMember, IEEE Computer Society *Hiding Sensitive Association Rules with Limited Side Effects*, VOL. 19, NO.1,JANUARY 2007. IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING