INTERNATIONAL JOURNAL FOR ADVANCE RESEARCH IN ENGINEERING AND TECHNOLOGY
WINGS TO YOUR THOUGHTS.....

DRS Method for Privacy Preserving In Quantative Data

Tushar Aggarwal[1], Harish Saini[2]

[1]Research Scholar, [2]Associate Professor
[1]N.C. College of Engineering, Kurukshetra University,
Israna, Panipat-132107, India
tushar.aggarwal88@yahoo.com

[2]N.C. College of Engineering, Kurukshetra University,
Israna, Panipat-132107, India
erharishsaini@yahoo.co.in

Abstract: *Data mining is the process of extracting useful patterns or knowledge from large databases. However, data mining also poses a threat to privacy and information protection if not done or used properly).Association rule mining is a method to find out the correlation among huge set of data items. As techniques for hiding association rules are limited to binary items, but real world data is quantative.In this paper, a technique to hide fuzzy association rule is proposed in which fuzzified data is mined using modified apriori algorithm in order to extract rules and identify sensitive rules. To hide sensitive rules, we are decreasing the support value of R.H.S.of the rule. However; results show the efficient information hiding with some side effects*

Keywords: *fuzzy association rule, fuzzy set concepts, decrease rule support, quantative data, frequent itemsets*

1. INTRODUCTION

Data mining and knowledge discovery in databases are two research areas that deal with automatic extraction of previously unknown and potentially useful information from large amount of data and have played an important role in various domains such as web commerce, marketing, medical analysis and so on. Despite benefit in such domains, data mining also poses a threat to data and information privacy if not done or used properly. For example, association analysis is a powerful tool for discovering patterns hidden in large data sets and some useful hidden information could be easily discovered using this kind of tool. It involves the discovery of association rules, showing attribute values and conditions that occur frequently together in a given set of data. One rule is categorized as sensitive/useful if its disclosure risk is above some given threshold. Therefore, the protection of sensitive hidden information has become a critical issue to be resolved. The objective of privacy preserving data mining is to hide certain information so that they cannot be discovered through data mining techniques such as association rule analysis [1]. There have been two broad approaches for privacy preserving data mining [2]-[5]. The first approach, called output privacy, is to alter the data before delivery to data miner so that real data is obscured and mining result will not disclose certain privacy. For example, perturbation, blocking, merging, swapping and sampling are some methods that have been proposed for this type of output privacy [6]. The second approach, called input privacy, is to manipulate the data using data distribution methods. In this approach, mining result is not affected or minimally affected. The problem of mining quantitative association rule was first introduced in [7]. The basic idea was to map the categorical attribute values into corresponding binary attribute values. Some work has been done to discover fuzzy association rules from quantitative data using fuzzy set concepts [8]-[11]. However, only one work has been done in the field of hiding fuzzy association rule in quantitative data [12]. He proposed an algorithm to hide fuzzy association rule in quantitative data. The basic idea of this algorithm was to decrease the confidence of a rule by increasing support of L.H.S. of rule. In this paper, we attempt to present a method for preventing extraction of useful association rules from

INTERNATIONAL JOURNAL FOR ADVANCE RESEARCH IN ENGINEERING AND TECHNOLOGY
WINGS TO YOUR THOUGHTS.....

quantitative data by decreasing the support of the rule. The support of a rule *A®B* is decreased by decreasing the support count of item set AB which is achieved by decreasing the support value of either A or B i.e. item in L.H.S. or R.H.S. of the rule and this is done until either support or confidence value of the rule goes below minimum support or minimum confidence value respectively.

2. LITERATURE REVIEW

Privacy preserving data mining is a novel research direction where data mining algorithms are analyzed for the side effects they incur in data privacy. The main objective in privacy preserving data mining is to develop algorithms for modifying the original data in some way, so that the private data and private knowledge remain private even after the mining process [13]. The problem of privacy preserving rule hiding can be formally stated as below [9]:

Let D be the source database, R be a set of significant association rules that can be mined from D, and let R_h be a set of sensitive 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 R_h.

For transformation of database D into a released database D', a number of data modification methods exists and are discussed as follows.

To preserve the privacy of the data, the real data is modified by using a number of different methods of data modification discussed as follows [14]:

- *Perturbation:* In perturbation, privacy is preserved by replacing the original attribute value by a new value (i.e. changing a 1-value to 0-value or vice-versa) or by adding some noise.
- *Blocking:* In blocking, privacy is preserved by replacing the original attribute value with a question mark ("?").
- *Aggregation or Merging:* In aggregation or merging, privacy is preserved by combining several values into a common category.
- *Swapping:* In swapping, privacy is preserved by interchanging the values of the individual records.

3. FUZZY ASSOCIATION RULES

Association rules are the interesting associations or correlations that appear frequently together in a given dataset. Association rule mining finds association or correlation relationships among a large set of data items. With massive amount of data continuously being collected and stored, many industries are becoming interested in mining association rules from their databases. The discovery of interesting association relationship among large amount of business transaction records can help in many business decision making processes, such as catalog design, and customer shopping behavior analysis [14].

Association rule mining is a two-step process. The two steps are:

1. *Find all frequent itemsets***:** In this step, all those itemsets which occur at least as frequently as a pre-defined minimum support count (considered as frequent itemsets) are calculated.
2. *Generate strong association rules from the frequent itemsets***:** Those rules which satisfy minimum support and minimum confidence (considered as strong association rules) are generated.

Let $I=\{i_1, i_2, ..., i_m\}$ be the complete item set where each $i_j\ (1\le j\le m)$ is a quantitative attribute.

Given a database $D=\{t_1, t_2, ..., t_n\}$ with attributes I and the fuzzy sets associated with attributes in I, we want to find out some useful association rules. Let $X=\{x_1, x_2, ..., x_p\}$ and $Y=\{y_1, y_2, ..., y_q\}$ are two large item sets.

Then, the fuzzy association rule is given as follows:

IF X is A THEN Y is B

Or simply, $A\rightarrow B$

Where $A=\{f_1, f_2, ..., f_p\}$ and

$B=\{g_1, g_2, ..., g_q\}$ and

$f_i \in$ {the fuzzy regions related to attribute x_i}

$g_j \in$ {the fuzzy regions related to attribute y_j}

X and Y are subsets of item set I and are disjoint which means that they share no common attributes. A and B contain the fuzzy sets associated with the corresponding attributes in X and Y. Here A is called as the antecedent or body or Left Hand Side (L.H.S.) of the rule and B is called as the consequent or head or Right Hand Side (R.H.S.) of the rule. For example,

[Age = Young] → [Income = High]

is a fuzzy association rule, where young and high are the fuzzy terms associated with age and income

INTERNATIONAL JOURNAL FOR ADVANCE RESEARCH IN ENGINEERING AND TECHNOLOGY

WINGS TO YOUR THOUGHTS.....

attribute respectively. The significance of an association rule is measured by its support and confidence [14]. Support is defined as the percentage of transactions that contain both A and B, while confidence is defined as the ratio of the support of $A \cup B$ to the support of A. In other words, the support of a rule measures the significance of the correlation between item sets, while the confidence of a rule measures the degree of the correlation between item sets. If a rule is useful/interesting, it should have support larger than or equal to minimum support value and confidence larger than or equal to minimum confidence value.

4. NEW PROPOSED SCHEME

There are two methods to hide a fuzzy association rule $(A \rightarrow B)$ as follows:

(1) Decrease the support of the rule to be smaller than minimum support value

(2) Decrease the confidence of the rule to be smaller than minimum confidence value.

To decrease the confidence of a rule, two strategies can be used. The first one is to increase the support of count A i.e. L.H.S. of the rule, but not support count of $A \cup B$. The second one is to decrease the support count of $A \cup B$, while keeping the support count of A i.e. L.H.S of the rule constant

Based on first method mentioned above, we propose the first algorithm namely Decrease Rule Support (DRS).

This algorithm first finds the useful fuzzy association rules which consist of only one item on both sides of it and then hide them using privacy preserving technique. For hiding purpose, the algorithm tries to decrease the support of rule $A \rightarrow B$ by decreasing the support count of itemset AB until either support or confidence value of the rule goes below minimum support or minimum confidence value respectively. To achieve this, the support count of item set AB is decreased by decreasing the support count of either A or B i.e. item in L.H.S. or R.H.S. of the rule. For this purpose, the value of item in L.H.S. or R.H.S. is subtracted from one, in case one minus value of item in L.H.S. or R.H.S. is less than the value of item in R.H.S or L.H.S respectively.

Abbreviations used in the proposed algorithms are given as follows:

D: Initial database with *n* transaction data; F: fuzzified database; T_L: value of a L.H.S. item in transaction t; T_R: value of a R.H.S. item in transaction t; U: An association Rule; Tx: transactions belong to the rule U.

Input:

(1) A source database D,
(2) A minimum support value (min_support),
(3) A minimum confidence value (min_confidence).

Output:

A transformed database D' so that useful association rules cannot be mined. Our algorithm is given in the following sub-section.

Algorithm DRS:

1. Fuzzi fication of the database, D → F;
2. In fuzzi fied database F, calculate every item's support value where f ε F;
3. IF all f (support) < min_support THEN
4. EXIT; // there isn't any rule
5. Find large 2-itemsets from F;
6. FOR EACH X's large 2-itemset {//find all rules
7. Find R = {Rules from itemset X};
   //for X= {i1, i2},two possible rules are \
   1 2 *i* ®*i*
   //and 2 1 *i* ®*i*
8. IF R is empty THEN
9. GO TO next large 2-itemset; //i.e. line 6
10. Select and remove a rule U from R;
11. Compute confidence of the rule U;
12. IF confidence (U) < min_confidence THEN
13. Add the rule U to Rh;
14. GO TO line 8;
15.} //end of FOR EACH line 6
   //Hides all rules in Rh
16. REPEAT {//until no more rule can be hidden
17. Select the next rule U from Rh;
18. IF confidence (U) < min_confidence OR
   support < min_support THEN
19. GO TO line 16;
20. Find Tx = {t| t ε U such that 1-max (TL, TR) < min (TL, TR)};
21. Sort transactions in Tx in descending order by value TL + TR -1;
// for maximum decrease in support value of rule
22. WHILE (confidence (U) ≥ min_confidence and support (U) ≥ min_support and Tx is not empty) {
23. Choose the first transaction t from Tx;
24. IF TR > 0.5 and TL = TR THEN
25. TR = 1 – TR;
26. ELSE

INTERNATIONAL JOURNAL FOR ADVANCE RESEARCH IN ENGINEERING AND TECHNOLOGY

WINGS TO YOUR THOUGHTS…..

27. max (TL,TR) = 1 – max (TL,TR);
28. Remove and save the transaction t from Tx;
29. Re-compute support and confidence of rule U
30.} // end WHILE line 22
31. IF Tx is empty THEN
32. Cannot hide rule U and restore F;
33.} UNTIL (No rule in Rh is modified)//end line 16 and support (U) ≥ min_support and Tx is not empty)
{
23. Choose the first transaction t from Tx;
24. IF TR > 0.5 and TL = TR THEN
25. TR = 1 – TR;
26. ELSE
27. max (TL,TR) = 1 – max (TL,TR);
28. Remove and save the transaction t from Tx;
29. Re-compute support and confidence of rule U
30.} // end WHILE line 22
31. IF Tx is empty THEN
32. Cannot hide rule U and restore F;
33.} UNTIL (No rule in Rh is modified)//end line 16
34. Transform the updated database F→D', output updated D';

5. CONCLUSIONS AND FUTURE WORK

We focused our attention on the problem of privacy preserving fuzzy association rules hiding in quantitative database. We proposed one fuzzy association rules hiding algorithm for hiding useful fuzzy association rules. Unlike previous approaches which mainly deal with association rules in binary database, our approaches deal with hiding the association rules in quantitative database. For this purpose, fuzzy set concepts are used for converting quantitative value into its corresponding fuzzy sets. Numerical experiments have been performed to measure the performance of the algorithms according to three metrics: hidden failure, database effects and side effects of the algorithm. In terms of hidden failure, we observed that DRS algorithm gives the best performance in comparison to all the other algorithms. In terms of database effects, we observed that DRS algorithm modify a small number of entries in comparison to previous work for different dataset sizes.

5.1 Suggestions for Future work

The proposed algorithms give adequate amount of scope for extension. Some suggestions for future work are as follows:

1 In our algorithms, a single member function is used for mapping all the attributes to their corresponding fuzzy sets. Instead of single member function, different member functions can be used for different attributes based upon attribute data characteristic.
2 The proposed algorithms can be extended to work in the context of fuzzy association rules privacy preserving data mining blocking family.

REFERENCES

[1] D.E.O' Leary," Knowledge Discovery as a Threat to Database Security", Proceedings of first International Conference Knowledge Discovery and Databases, (1991, pp. 507-516.

[2] Wang, S.L., Jafari, A., "Using Unknown for Hiding Sensitive Predictive Association Rules", In Proceedings of the 2005 IEEE International Conference on Information Reuse and Integration, USA, August 2005, pp. 223-228.

[3] S. L. Wang, B. Parikh and A. Jafari, "Hiding Informative Association Rule Sets", Expert Systems with Applications, Volume 33, Issue 2, August 2007, pp. 316-323.

[4] S. L. Wang, Y. Lee, S. Billis and A. Jafari, "Hiding Sensitive items in Privacy Preserving Association Rule Mining", IEEE International Conference on Systems, Man and Cybernetics, Vol. 4, Oct 2004, pp. 3239- 3244.

[5] S. L. Wang and A. Jafari, "Hiding Sensitive Predictive Association Rules", IEEE International Conference on Systems, Man and Cybernetics, Vol. 1, Oct 2005, pp. 164- 169.

[6] V. Verkios, E. Bertino, I. G. Fovino, L. P. Provenza, Y. Saygın and Y. Theodoris, "State-of-the-art in Privacy Preserving Data Mining", SIGMOD Record, Vol. 33, No. 1, March 2004, pp. 50-57

[7] R. Srikant and R. Agrawal, "Mining Quantitative Association Rules in Large Relational Tables", Proceedings of ACM SIGMOD, 1996, pp. 1-12.

[8] L.A. Zadeh, "Fuzzy Sets," Information and Control, Vol. 8, 1965, pp. 338-353.

INTERNATIONAL JOURNAL FOR ADVANCE RESEARCH IN ENGINEERING AND TECHNOLOGY

WINGS TO YOUR THOUGHTS.....

[9] T. P. Hong, C. S. Kuo, S. C. Chi, "Mining association rules from quantitative data", Intell. Data Anal 3 (5), 1999, pp. 363–376.

[10] M. Kaya, R. Alhajj, F. Polat and A. Arslan, "Efficient Automated Mining of Fuzzy Association Rules," Proc. of DEXA, 2002.

[11] Chen G., Yan P., Kerre E.E,"Computationally Efficient Mining for Fuzzy Implication-Based Association Rules in Quantitative Databases", International Journal of General Systems, Vol. 33, No. 2-3, 2004. Pp. 163-182.

[12] T. Berberoglu and M. Kaya, "Hiding Fuzzy Association Rules in Quantitative Data",The 3rd International Conference on Grid and Pervasive Computing Workshops, May 2008, pp. 387-392.

[13] J. Han and M. Kamber, "Data Mining: Concepts and Techniques", 2/e, Morgan Kaufmann Publishers, March 2006, pp. 230.

[14]http://mlearn.ics.uci.edu/databases/breast-cancerwisconsin/breast-cancer-wisconsin.da