An Efficient Fuzzy-based Method for Privacy Preserving in Data mining Using DRS

Radhamadhab Dalai, Prof. Kishore Kumar Senapati

***Abstract*— Analysis and investigating customer's prior choice or activity is becoming a general trend as these have been used for building business analytics and marketing tools. So for this purpose generally organizations are dependent on the customer's database to retrieve or collect the useful data the researchers and engineers need to examine. In some cases this has brought about a threat to privacy of customer's data. To avoid this, the method of collection and protection of data has seen a change in various directions. Here, useful attributes of data are encrypted to another domain so that the actual values are not disclosed. But as per database original association has to be maintained. Hence a number of algorithms have been designed to find out the useful association rules and then hide them but only in consideration of binary data, yet in real world databases mostly consist of quantitative or numerical values. Therefore, in this paper focus is on a fuzzy association rule concealing algorithm DRS (Decrease Rule Support) for obscure useful association rules located from a database containing quantitative values. The suggested algorithm incorporates the fuzzy set ideas to discover the useful or sensitive fuzzy association rule and then prevent them using privacy preserving approach. The Experimental results present that the proposed algorithm conceals more sensitive association rule with minimal modification of original data which has to be released. The experiment has wide area of application in health care system, stock market,e-payment,virtual market place,E-shop billing. The data from Wisconsin health care organization has been used for our proposed algorithm and corresponding results and comparison have been calculated using these data set.**

Index Terms*— *Fuzzy association rules, fuzzy set concepts, privacy preserving data mining, quantitative data, binary data, decrease support rule, sensitive association rule

I. Introduction

Data mining (called knowledge discovery) is the method of studying data from different prospective and outlining it into useful information that can be used to increase profit and decrease costs or both. Still, data mining also causes harm to data privacy and preservation, if it is unable to used carefully. For instance, association rule investigation is a popular mechanism for discovering useful associations from large database and by using this mechanism some useful or sensitive hidden information should be easily extracted. For this reason, the security of useful invisible information has become a risky problem to be fixed. The primary goal of privacy preserving data mining is to conceal sensitive information so that any association rule analysis technique or data mining technique can not able to mined it [2].There are two wide concepts founded for privacy preserving data mining [7][8][9][13]. The first concept known as input privacy is to handled skill fully the data using data distribution techniques where mining results is not affected or least affected. For instance, reconstruction based and cryptography based are some methods that have been presented for this type of input privacy [12].The second concept known as output privacy is to make unclear the data before conveyed to data miner so that original data is hided and mining result unable to reveal any privacy. For instance, data perturbation, blocking, merging, swapping and sampling are some methods that have been presented for this type of output privacy [12].Two algorithms were propounded by Wang et al. [7][8][9] specifically named as ISL (Increase Support of Left hand side) and DSR (Decrease Support of Right hand side) to conceal interesting association rules from transactions data with binary attributes database. In case of ISL approach, confidence of a rule is decreased by increasing the support value of L.H.S (Left Hand Side) of the rule by only considering the items from L.H.S. of a rule. In case of DSR approach, confidence of a rule is decreased by decreasing the support value of R.H.S (Right Hand Side) of a rule by only considering the items from R.H.S. of a rule. As stated above, most of the studies based on hiding Boolean association rules which are only involved with the presence or absence of an item rather than considering its quantity. However, transactions with quantitative values are commonly found in real life applications. For example at illness the sugar level of a person goes high at the same time people having sugar problem also have high level of sugar's quantity in blood, but it does no mean that they are sick. So only the presence or absence of a sugar does not decide that a this Two algorithms were propounded by Wang et al. [7][8][9] specifically named as ISL (Increase Support of Left hand side) and DSR (Decrease Support of Right hand side) to conceal interesting association rules from transactions data with binary attributes database. In case of ISL approach, confidence of a rule is decreased by increasing the support value of L.H.S (Left Hand Side) of the rule by only considering the items from L.H.S. of a rule. In case of DSR approach, confidence of a rule is decreased by decreasing the

***Radhamadhab Dalai**, Department of Computer Science and Engineering, Birla Institute of Technology Mesra, Ranchi, India, Phone/ Mobile No. 9766582174*

***Prof. Kishore Kumar Senapati**, Department of Computer Science and Engineering, Birla Institute of Technology Mesra,, Ranchi, India, Phone/ Mobile No. 8285583730*

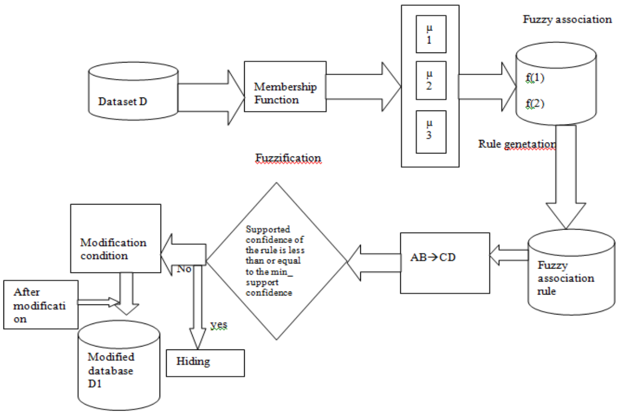
support value of R.H.S (Right Hand Side) of a rule by only considering the items from R.H.S. of a rule. As stated above, most of the studies based on hiding Boolean association rules which are only involved with the presence or absence of an item rather than considering its quantity. However, transactions with quantitative values are commonly found in real life applications. For example at illness the sugar level of a person goes high at the same time people having sugar problem also have high level of sugar's quantity in blood, but it does no mean that they are sick. So only the presence or absence of a sugar does not decide that a person is sick, quantity of elements in sugar is also important for verification of illness.

2. BACKGROUND

2.1 Some related works

Submit your manuscript electronically for review. A few work has been done to extract interesting fuzzy association rules from quantitative database using fuzzy set concepts by L.A zadeh, M Kaya, R Alhajj, F Polat, A Arslan, and T P hong [1][4][5][11]. Although, only one experiment has been performed in the field of hiding interesting fuzzy association rule in quantitative data by T. Berberoglu et al. [10] and he proposed an algorithm to conceal useful fuzzy association rule in quantitative data. The fundamental idea of this designed algorithm was to decrease the confidence of a rule by increasing support of L.H.S. of rule for hiding more useful rules. In this paper, we have tried to state a method for blocking extraction of useful or sensitive association rules from quantitative data by decreasing the support of the rule. The support of a rule, let take a rule AB□CD which contains more than one items on both side of the rule and it is decreased by decreasing the support count of item set A,B (lest side rules) and C,D (right side rules) which is achieved by decreasing the support value of either A,B (L.H.S) or C,D (R.H.S) or both at the same time and this is performed till either support or confidence value of the rule goes below than the assumed minimum support or minimum confidence value respectively. The rule is hidden from the user if either the support goes below the minimum support or the confidence goes below the minimum confidence values respectively.

3. *Diagrammatical Representation of the Proposed Model*



.

Figure 1: The diagram for proposed model

4. Methods & Algorithm

In order to hide an association rule, AB→CD, which contain more than item on both side of the rule, we can either decrease its support to be smaller than minimum support value or its confidence to be smaller than its minimum confidence value. To decrease the confidence of a rule, two strategies can be used. The first one is to increase the support count of AB i.e. L.H.S. of the rule, but not support count of ABUCD. The second one is to decrease the support count of ABUCD, while keeping the support count of AB i.e. L.H.S. of the rule constant. Based on first method mentioned above, we proposed an 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 the rule and then hide them using privacy preserving technique. For hiding purpose, the algorithm tries to decrease the support of rule ABUCD by decreasing the support count of item set 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 ABCD is decreased by decreasing the support count of either AB or CD 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 on. Abbreviations used in the proposed algorithm are given as follows:

D: Initial database with n transaction data; F: fuzzified database; TL: value of a L.H.S. item in transaction t; TR: value of a R.H.S. item in transaction t; U: An association Rule.

4.1 Notations and Problem Definition

Let I= {i_1, i2,..., i_m } be the entire itemset where each i_j (1<= j<= m) is a quantitative attribute of a database. Given a database { t_1,t_2 ,...,t_n } where t_n is number of transactions with columns I and the fuzzy sets associated with columns in I, we want to find out some interesting useful or sensitive association rules. Let X={x_1 ,x_2 ,...,x_p } , Y={ y_1,y_2 ,...,y_q }, Z={z_1,z_2,….,z_r} and S={s_1,s_2,….,s_t} are four large item sets. So the fuzzy association rule is stated as follows: AB→CD ,where A={ e_1,e_2 ,...,e_p } ,B={f_1 ,f_2 ,...,f_q },C={g_1,g_2,…,g_r} and D={h_1,h_2,…,h_t**}**

e_i э {the fuzzy sections related to column x_i}
f_j э {the fuzzy sections related to column y_j}
g_K э {the fuzzy sections related to column z_k}
h_l э {the fuzzy sections related to column q_l}

X, Y, Z, Q are subsets of I and are disjoint which signifies that they share no familiar attributes or columns. A, B, C, D contains the fuzzy sets associated with the corresponding attributes in X, Y, Z, Q. Here A, B is known as the body or Left Hand Side (L.H.S.) of the rule and C, D is known as the head or Right Hand Side (R.H.S.) of the rule. The importance of an association rule is estimated by its support and confidence value. Support is explained as the percentage of transaction that contain all A,B,C and D at a time while confidence is defined as the ratio of the support of AB → CD to the support of AB. In other words the support of a rule measures the importance of dependent relationship between itemsets while the confidence of a rule measures the degree of dependency between itemsets. A rule is called useful or interesting when the support is larger than or equals to the assumed minimum support and confidence value larger than or equals to the assumed minimum confidence value. To hide the interesting rule the value of the item in the left hand side and right hand side is subtracted from one then check the condition weather the support or confidence values goes below than the minimum support or confidence.

Input Parameters

- A Source Database D,
- A min _Support,
- A min _ Confidence,
- A set of items X

Output Parameter

- A changed database D′ so that useful fuzzy association rules cannot be mined.

4.2 Algorithm

```
1. Fuzzification of the database D to F;
2. After fuzzification modified database F, calculate every items Support count value;
3. IF all f(support count)< min _Support THEN
   EXIT; //there isn't any rule
4. Find all large items from F;
5. FOR EACH X"s large items generate all possible rules and take one of them rule U;
6. Compute the Support and Confidence value of the rule U;
7. IF Support (U)< min _ Support or Confidence(U)< min _ Confidence THEN
    The rule is already hided.
8. GOTO line 5;
9. ELSE choose the first transaction t from $T_{X;}$

10. max ($TL_1$, TL2…..,TLn)= 1- ( $TL_1$,$TL_2$…..,TLn);
     max ($TR_1$, TR2…..,TRn)= 1- ( $TR_1$,$TR_2$…..,TRn);
11. Recomputed the new _Support and new _Confidence value of rule U;
12.IF new _ Support (U)< min _ Support OR new _ Confidence< min _ Confidence ;
 THEN EXIT;
13.ELSE go to the next transaction from $T_X$ and repeat
    step   10;
14. GOTO the line % until all the rules become hided;
15. Transform the update database F→D and output updated to D as D′;
```

5. Steps of Fuzzy Association Rule Hiding Algorithm With Example

The dataset used is Wisconsin breast Cancer dataset from UCI Machine Learning Repository. The dataset consists of nine quantitative attributes and one categorical attribute. We used only six quantitative attributes and ignored categorical.

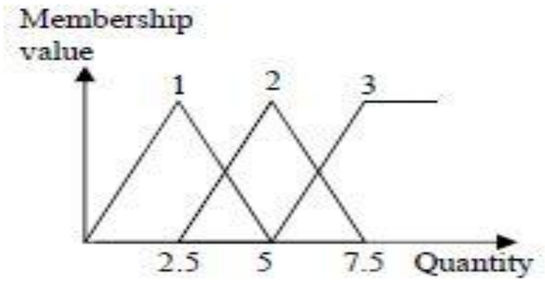
A	B	C	D	E	F
5	1	1	1	3	1
5	4	4	5	3	1
3	1	1	1	3	1
6	8	8	1	3	1

Table 1: Sample data from database[13]

5.1 Explanation

Step 1: The membership function as shown below called triangular membership function is used for converting all the six attributes to their corresponding fuzzy set and each attribute has three fuzzy sets.

μ = Max (min (x-a / b-a, c-x / c-b), 0).



Where a is the left end of the triangle,
b is the peak of the triangle
c is the right end of the triangle

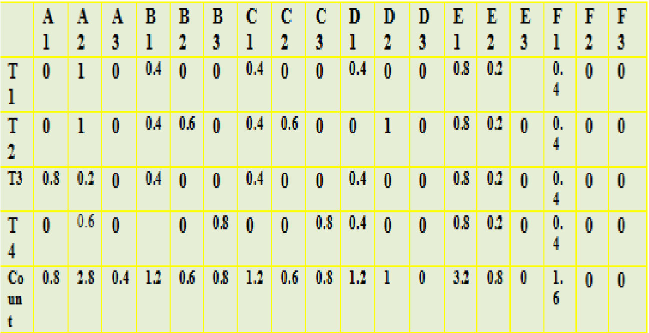
A1	A2	A3	B1	B2	B3	C1	C2	C3	D1	D2	D3	E1	E2	E3	F1	F2	F3
0	1	0	0.4	0	0	0.4	0	0	0.4	0	0	0.8	0.2	0	0.4	0	0
0	1	0	0.4	0.6	0	0.4	0.6	0	0	1	0	0.8	0.2	0	0.4	0	0
0.8	0.2	0	0.4	0	0	0.4	0	0	0.4	0	0	0.8	0.2	0	0.4	0	0
0	0.6	0	0	0	0.8	0	0	0.8	0.4	0	0	0.8	0.2	0	0.4	0	0

Table 2: shows after modification of each attribute after fuzzification

Step 2.Calculate the support count of each attribute region on the transactions data which is performed by adding each transaction values of each attribute.

Step 3.Assume that the minimum support is (0.8) 20% and minimum confidence is 75%

Step 4.Now we will consider the rule (A2, B1 →E1, F1) and calculate its support and confidence value and compare with the minimum confidence and minimum support value. If its support and confidence value is grater or equals to minimum support and confidence value then consider useful the rule for modification.



	A1	A2	A3	B1	B2	B3	C1	C2	C3	D1	D2	D3	E1	E2	E3	F1	F2	F3
T1	0	1	0	0.4	0	0	0.4	0	0	0.4	0	0	0.8	0.2		0.4	0	0
T2	0	1	0	0.4	0.6	0	0.4	0.6	0	0	1	0	0.8	0.2	0	0.4	0	0
T3	0.8	0.2	0	0.4	0	0	0.4	0	0	0.4	0	0	0.8	0.2	0	0.4	0	0
T4	0	0.6	0		0	0.8	0	0	0.8	0.4	0	0	0.8	0.2	0	0.4	0	0
Count	0.8	2.8	0.4	1.2	0.6	0.8	1.2	0.6	0.8	1.2	1	0	3.2	0.8	0	1.6	0	0

Table 3: Table for support count

Calculate the support and confidence

support(A2,B1→E1,F1) =

The support (A2, B1 → D1, E1)
----
n

= 1.0 / 4 = 25%

The confidence (A2, B1 → D1, E1) =

support (A2,B1→E1,F1)
----
support (A2 , B1)

= 1.0/ 1.2 = 83.3333%

	A1	B1	E1	F1	Support
T1	1	0.4	0.8	0.4	0.4
T2	1	0.4	0.8	0.4	0.4
T3	0.2	0.4	0.8	0.4	0.2
T4	0.6	0	0.8	0.4	0
Count	2.8	1.2	3.2	1.6	1

Table 4: calculation of support count and confidence

Step 5.As the support and confidence is greater so we will modifying the table by applying the condition

max (TL$_1$, TL2…..,TLn) = 1- (TL$_1$,TL$_2$…..,TLn)
max (TR$_1$, TR2…..,TRn) = 1- (TR$_1$,TR$_2$…..,TRn);

first take the first transaction and apply the above formula and after modification if the support and confidence value goes below than the minimum support and minimum confidence value then stop here otherwise go for the second one and so on.

Now the support and confidence is below than the minimum support and minimum confidence

	A1	B1	E1	F1	Support
T1	0	0.4	0.2	0.4	0
T2	1	0.4	0.8	0.4	0.4
T3	0.2	0.4	0.8	0.4	0.2
T4	0.6	0	0.8	0.4	0
Count	1.8	1.2	2.6	1.6	0.6

Table 5: calculation of new support and new confidence value

The new_ support (A2, B1→D1, E1) =

support(A2,B1àE1,F1)

----
n

= 0.6 / 4 = 15%

The new_confidence =

(A2, B1 → D1, E1)
----
support (A2 , B1)

= 0.6/ 1.2 = 50%

Since the new _ support and new _ confidence is less than min _ support and min _ confidence value respectively ,the rule is hidden from the user and we stop our process of modifying data here otherwise we will further proceed for modifying entry values of next transaction .

Formula for defuzzification is given below

$$X = \frac{\sum_{i=1}^{n} Xi \,.\, \mu(Xi)}{\mu(Xi)}$$

A	B	C	D	E	F
0	1	1	1	3	1
5	4	4	5	3	1
3	1	1	1	3	1
6	8	8	1	3	1

Table 6: After defuzzification

6. Results And Table comparison

We have tested the proposed algorithm with the system configuration as Processor is Intel core 2 duo with Speed 2.20 GHz having RAM 4 GB and Hard Disk 298 GB ,Operating System is Windows 7 used Language is VB.NET and IDE is Microsoft Visual Studio 2008. It can be easily seen that our algorithm generates less side effects and modify only a small number of entries. The reason is that our algorithm makes minimum modification of data because we used a method for selection of transaction for modification. The method used is to select transaction in the order which results in maximum decrease in support value of the rule. Therefore, higher data quality of the released database is maintained by our algorithm

A	B	C	D	E	F
5	1	1	1	3	1
5	4	4	5	3	1
3	1	1	1	3	1
6	8	8	1	3	1

Table 7: Before modification

A	B	C	D	E	F
0	1	1	1	3	1
5	4	4	5	3	1
3	1	1	1	3	1
6	8	8	1	3	1

Table 8: After modification

7. CONCLUSION

In this paper, Our proposed algorithm hides the useful fuzzy association rule which contains more than one item on both side of the rule and it integrates the fuzzy set concepts and Apriori mining algorithm to find useful fuzzy association rules from a quantitative database and then hide them using privacy preserving technique. It performs minimum number of modifications to the original Database. Our experimental results conclude that the proposed scheme hides more sensitive rules with minimum number of modifications and maintains quality of the released data .We have used numerical data for our experimentation purpose and similarly this method can be extended to categorical data. We can improve the working of this algorithm by applying a different fuzzy membership function by comparing with the nature of membership function used in our paper. It is applicable in different domains such as medical diagnosis, temperature control, predicting travel times and predicting genetic behaviours.

8. REFERENCES

[1] 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.

[2] 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.

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

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

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

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

[7] 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.

[8] 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.

[9] 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.

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

[11] 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

[12] 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.

[13] Online resource for winconsin bc data for machine learning http://mlearn.ics.uci.edu/databases/breastcancerwisconsin/breast-cancer-wisconsin.data

[14] De Cock,M.,Cornelis,C.,Kerre,E.E:Fuzzy Association Rules:A two sided Approach In:FIP,Pages(s)385-390,2003

[15] Nabar S. Marathi B,Kenthapadi K,Mishra N,Motwani R,"Towards Robustness in Query Auditing"VLDB Conference,2006.

[16] R.R Rajalaxmi,A.M Natarajan "An Effective Data Transformation Approach for privacy crets

[17] Zadeh L "Fuzzy sets",Inf. Control .Vol.8,pp,338-353,1995.

[18] Yingyi Bu,Ada wai-chee Fuet al,"privacy preserving serial data pulishinh by role composition',in VLDB,2008 pp.845-856

[19] Z.huang "Extension to the K-means algorithm for clustering large data sets with categorical values" data mining and knowledge discovery,1998.pp.283-304

[20] Ms.M.Sandhiyah, Mrs.V.Shanmuga priya," An Effective Association Rule Mining Using Fuzzy Confabulation Rare Item set Selection Algorithm", International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 4 Issue 9, September 2015

Radhamadhab Dalai PhD Scholar in Computer Science and Engineering Department, BIT Mesra, MTech(CSE), Research Areas: Pattern recognition and image analysis, Machine learning, Machine intelligence and soft computing

Prof. Kishore Kumar Senapati Assistant Professor , Computer Science and Engg, M.Tech(.CSE)-UTKAL, Ph.D(Engg)-BIT_MESRA,Research Areas : Pattern Matching through Image. Soft computing, Efficient Data Structure for sorting and searching. Prediction Algorithm Life member of the society CSI (Computer Society of India)