



Application of DSR on Quantitative Association Rules Using Rough Set for Privacy Preservation

Pankaj Shivalkar, Gurudatt Kulkarni, Darpan Sathe, Prof. Lavanya K.
School of Computer Science and Engineering,
VIT University, Vellore - 632014, Tamil Nadu, India

Abstract: The extraction of the information from the data is the main aim of Data Mining. There is always a threat to information privacy in Data Mining. Data mining privacy preservation technique is used by the information to hide the sensitive rules from public. The elements of the selected data set can be reduced by using attribute reduction technique. In this technique rough set is being used for reduction of attributes to created reduced sets. The paper proposes an approach in which decreasing the support of right hand side (DSR) techniques along with the reduct core is used for hiding the sensitive information from the selected Datasets.

Keywords: Rough Set, Fuzzification, attribute reduction, discernibility matrix, privacy preserving data mining, DSR, quantitative association rule.

I. Introduction

Reduction of attributes is the feature for selecting the necessary attributes from the whole set of the dataset which can provide us with the information which can be important and useful for us for the deriving the association rule with the help of Rough Set and Decreasing the Support of Right hand side (DSR). It helps in maintaining the attribute that are useful in characterizing the dataset from the selected database. With the help of this attribute reduction we get the improved quality of the dataset on the basis of which we compute the necessary privacy preserving association rules for data hiding [8]. The accuracy of the classification can be increased with application of attribute reduction by removing the noise, unnecessary attributes which just form redundant data in the dataset generated after the preprocessing of data. The generation of the association rules, trends and correlation [4] becomes easy when they are created out from a reduced dataset having less number of the attributes compared to the original dataset. A good algorithm for the attribute reduction helps reducing the unnecessary attributes that helps in creating both the efficient rule comprehension and rule prediction performance.

The Data Mining methodology used in this paper used is dealing in discovering the patterns from the large datasets from database, Hadoop, along with the useful qualities of the data mining techniques, statistics and database system. Data extraction is the goal of all the techniques used in the understandable formats which can be used as the knowledge base for decision making. The privacy preservation has become the important issue in few recent years due to the increased use of the sophisticated algorithms of the data mining to retrieve the personal information from the selected datasets. Data mining is possibly considered as the threat to data privacy because of the heavy use of the e-data maintenance done over the databases looked after by the corporate. This has triggered the concerns about the data privacy for the dataset retrieved from the databases. But our aim is to maintain the privacy preservation without compromising with the data security. [5]

II. Literature Survey

The attribute reduction (feature selection) which we are using in our paper is a main problem in field of Machine Learning. This technique is the process of selection of the necessary attributes which we require for creation of the necessary association rules. Here, it reduces the count of the columns which are selected for the necessary operations; it comparatively reduces the time required by the software entity to create the association rules. Rough Set based Attribute Reduction [5], get the columns from the dataset while excluding the unnecessary columns hence further reducing the time complexity of computing the confidence and support necessary to build the association rule. The Rough Set which we are using in our paper has depends on two different approximations; upper approximation and lower approximation which on computing we get the boundary region which represents the elements of the target set of the whole dataset which we are considering. Here we are using the Rough Set to calculate get the reduct cores from the set of the columns of data tables from select Hadoop database [8]. In the Rough Set we are making use of the core which we construct as followed. We are considering the decision table 'Tbl' such that $Tbl = (U, C, D)$ where U denotes the Universe of finite n attributes out of which C denotes the condition attributes where as D denotes the decision attributes such that $C \cap D = \Phi$. So our Information System 'InSy' is denoted as $InSy = (U, C \cup D) \forall M \subseteq A$. The Indiscernibility relation $IND(M)$ is defined as $IND(M) =$

$\{(z, y) \in U \times U: a(z) = a(y) \forall a \in M\}$. Given to us is a set of object and extracted set of attributes. The Lower approximation of Z with respect to M is $M_*(Z) = \{z \in U: [z]_M \subseteq Z\}$. The upper approximation of Z with respect to M is $M^*(Z) = \{z \in U: [z]_M \cap Z \neq \emptyset\}$. The positive region of decision $d \in D$ with respect to M is $POS M(d) = \cup \{M^*(Z): Z \in U/IND(d)\}$. [2, 3]

Wang et al. have proposed the algorithms for ISL (Increasing Support of LHS) and DSR (Decreasing Support of RHS) to make the useful association rules unavailable for the purpose of the transactions along with the help of binary data attributes. As mentioned earlier, almost all of the proposed information mainly concentrates on making the visible data hidden like Boolean association rule for which we consider the first which are concerned only with the presence of the given Item Set for the transactions without considering its quality [5].

III. Architectural Framework

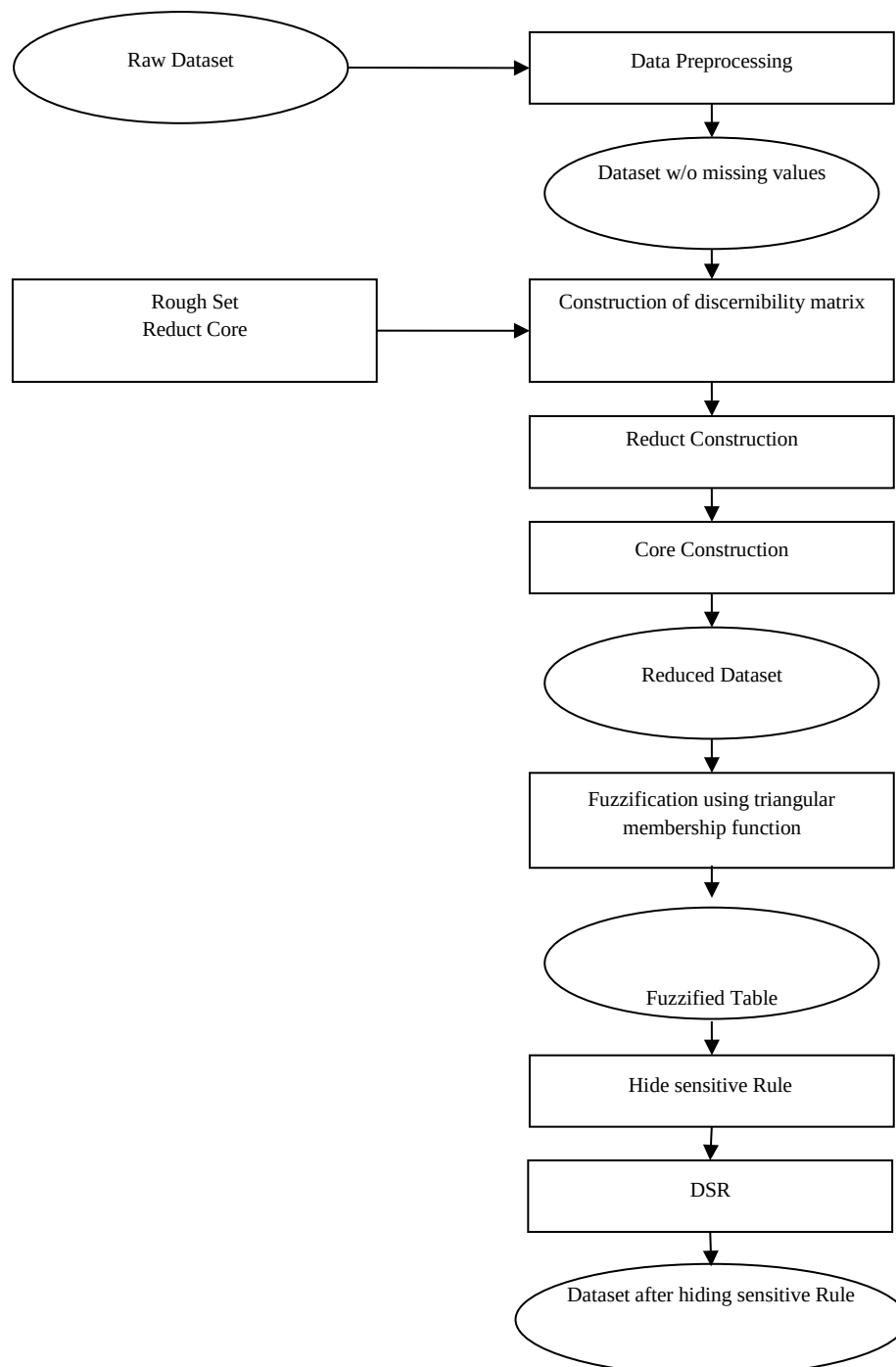


Figure 1: Workflow of the proposed Approach

A. Data Preprocessing

Preprocessing of the data is an important step in data mining. Data is the real world entity and it is collected from different sources. This real world data is often noisy, incomplete and inconsistent. And to resolve such issues the methodology of Data preprocessing is used. This data preprocessing makes the raw data that will be used in proceeding steps. There are different tasks that come under the data preprocessing. Those are: Data cleaning, data integration, data transformation, data reduction, data discretization. Filling the missing value, solving the inconsistencies problem come under the data cleaning process. Data integration combines data from multiple database and store in a coherent location. In the data transformation smoothing is performed by removing noise from the data and performing the normalization and summarization in data. In the data reduction technique reduction of the volume is done by using the data cube aggregation approach, binning method, clustering method. Data discretization is also a part of data reduction in which replacing the value takes place by nominal ones.

B. Attribute Reduction

Attribute Reduction (also called as attribute selection) is a technique used to reduce the dimensionality of dataset by selecting the relevant attributes. The selection of attributes is done by using the algorithm i.e. Rough set. Rough Set based attribute reduction (RSAR) is takes a dataset as an input and calculates the equivalence class and indiscernibility matrix in order to find optimized set. It removes the redundant data from dataset and identifies the core of the final reduct set.

The Rough Set theory is divided into some categories: positive region based methods, information entropy-based method, and discernibility matrix-based method. Our work uses a discernibility matrix-based method for finding a reduced dataset.

C. Fuzzification and Data hiding using DSR

The *fuzzy sets* concepts proposed by *L. Zadeh* resembled the *human reasoning* for the utilization of information approximately and uncertainty in the decision generation around in 1965, also known as *type-I fuzzy sets* [9].

A characterization of *fuzzy set (Class) A* in *X* is done by a *membership function* $f_A(x)$ which is associated with every point in *X* which is a real number inside the *interval* $[0,1]$, with the value of $f_A(x)$ at *x* representing the “grade of membership” of *x* in *A*. So if the value of $f_A(x)$ is nearer to unity, the grade of membership of *x* in *A* is also higher.

Every input and output variable has, two or more membership functions defined but can be normally *three*. The qualitative categories are defined for each function, like low, normal or high. Functions being of diversified shapes, mostly the triangular and trapezoidal are used.

The fuzzy set concepts are used in determining the fuzzy association rules from the quantitative data [9, 10]. The integration of the fuzzy set concepts and Apriori mining algorithm is done in finding fuzzy association rules and hiding them with the help of privacy preserving technique. For the purpose of data hiding, the support is decreased of to-be-hidden association rule by decreasing the support value of data in Right Hand Side (R.H.S) of this rule [4].

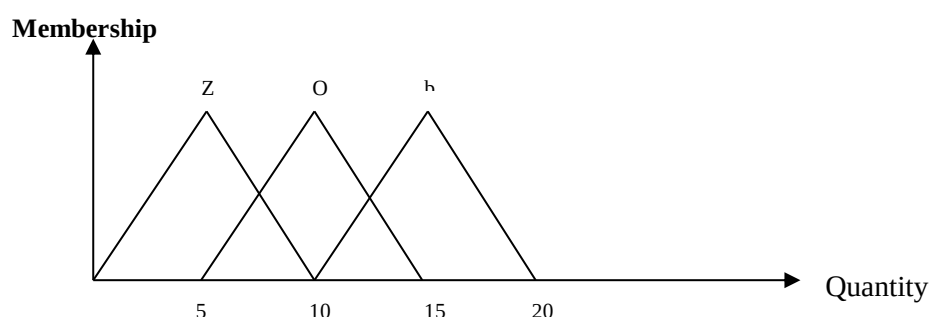


Figure 2: Triangular Membership Function

IV. Algorithm

In this the algorithm of Rough Set for reducing the dataset and hiding the sensitive rules using DSR is defined.

D. Selection of attributes using Rough Set

1. Take a random set.
2. Calculate the Equivalence class and Indiscernibility matrix.
Equivalence Class = Equi
Indiscernibility matrix= M
3. Remove the redundant column.
4. Consider X is 0th column of indiscernibility matrix.
5. While (Count (column) >2) {

- i. Create indiscernibility matrix.
- ii. Remove column X.
- iii. Compute equivalence class as “Equi”
- iv. If (Equi != Equi”) then
 - a. Restore the column X
 - b. Increment X.

6. Stop.

E. Illustrative Example

Consider the dataset which contain the records and columns named as Rec and Col. In the dataset the equivalence class $U/Col = \{Rec-1, Rec-2, Rec-3, Rec-4, Rec-5\}$ where U is a universal set.

Table I Raw Dataset

	Col-0	Col-1	Col-2	Col-3	Col-4	Col-5	Col-6	Col-7	Col-8	Col-9
Rec-1	6	4	3	2	7	8	5	6	3	4
Rec-2	3	2	3	1	3	1	5	2	4	3
Rec-3	2	2	3	1	3	1	5	2	1	3
Rec-4	6	4	3	2	7	1	5	6	2	4
Rec-5	3	2	3	1	2	1	5	3	1	1

In the Table I it is clearly seen that the column 2 and column 6 are redundant. For calculating the value of $U/(Col - (Col-2 + Col-6))$ by applying the rough set methodology we get is:

$$U/(Col - (Col-2 + Col-6)) = \{Rec-1, Rec-2, Rec-3, Rec-4, Rec-5\}$$

Hence it is seen that the equivalence class is same as U/Col . It means that these column are redundant and it does not effort the equivalence class. So, the obtained indiscernible matrix as follows:

Table II Reduced Dataset

	Col-0	Col-1	Col-3	Col-4	Col-5	Col-7	Col-8	Col-9
Rec-1	6	4	2	7	8	6	3	4
Rec-2	3	2	1	3	1	2	4	3
Rec-3	2	2	1	3	1	2	1	3
Rec-4	6	4	2	7	1	6	2	4
Rec-5	3	2	1	2	1	3	1	1

After deleting Col-1 from the Table II, the equivalence class obtained is:

$$U/(Col - Col-1) = \{Rec-1, Rec-2, Rec-3, Rec-4, Rec-5\}$$

This obtained equivalence classes similar to that of U/Col . It shows that Col-1 is redundant.

Table III Equivalence Classes

	Col-0	Col-3	Col-4	Col-5	Col-7	Col-8	Col-9
Rec-1	6	2	7	8	6	3	4
Rec-2	3	1	3	1	2	4	3
Rec-3	2	1	3	1	2	1	3
Rec-4	6	2	7	1	6	2	4
Rec-5	3	1	2	1	3	1	1

After on deleting Col-3 and Col-4 the equivalence class obtained is:

$$U/(Col - (Col-1+Col-2+Col-3+Col-4)) = \{Rec-1, Rec-2, Rec-3, Rec-4, Rec-5\}.$$

It shows that Col-3 and Col-4 are redundant or less necessary to represent the association rules which we require.

Table IV Reduced Equivalence Classes

	Col-0	Col-5	Col-7	Col-8	Col-9
Rec-1	6	8	6	3	4
Rec-2	3	1	2	4	3
Rec-3	2	1	2	1	3
Rec-4	6	1	6	2	4
Rec-5	3	1	3	1	1

On removal of Col-5, Col-7 and Col-9, the equivalence class obtained is similar to our first, calculated equivalence class as:

$$U/(Col - (Col-1+Col-2+Col-3+Col-4+Col-5+Col-7+Col-9)) = \{Rec-1, Rec-2, Rec-3, Rec-4, Rec-5\}.$$

It means Col-5, Col-7 and Col-9 are redundant and hence we get the discernible matrix as shown in Table V which shows the core reduct of our selected dataset.

Table V Core- Reduct

	Col-0	Col-8
Rec-1	6	3
Rec-2	3	4
Rec-3	2	1
Rec-4	6	2
Rec-5	3	1

F. DSR algorithm hiding the sensitive association fuzzy rules

In the database, if the value of confidence is greater than the threshold value, our sensitive information is disclosed. So to hide the information we have to decrease the value for our confidence until its value becomes smaller enough than minimum threshold value. There are different ways to minimize the value of confidence of which we have tried using one.

Let, $A \rightarrow B$ is a sensitive rule whose confidence value is greater than the minimum confidence value and we have to hide this rule, which can be done applying following rules:

1. By increasing the value of support A at L.H.S. or
2. By decreasing the value of support B at R.H.S.

Here we are decreasing the value of confidence by decreasing the support B at R.H.S. and for the same we use the given abbreviations as follows:

Fuzz: after Fuzzification, we get the Fuzz dataset.

Pred_Dataset: set of calculated values.

Rul: Rule.

Sen_Rul: Sensitive Rules.

Val_ColL: Value of left Column.

Val_ColR: Value of Right Column.

Cln_Data: Cleaned dataset.

Input Data:

- i. Optimized dataset.
- ii. Value of Minimum support (min_sup).
- iii. Value of Minimum confidence (min_conf).

Output Data:

- i. The reduced Dataset D' with hidden sensitive information.

Decreased the support of right hand side (DSR) Algorithm:

1. Optimized dataset.
2. Fuzzification is applied on the redundant free database by using triangle membership function. $Cln_Data \rightarrow Fuzz$.
3. In Fuzz Dataset, compute the support of each value $S_FUZZ \rightarrow FUZZ$.
4. If the value of S_FUZZ (support) $< min_sup$, then EXIT;
5. Consider two large item set from FUZZ.
6. For each Pred_Dataset large two value set.
Find $Rul = \{Rules\}$ for Pred_Dataset.
//for Pred_Dataset = $\{i_1, i_2\}$, rules are $i_1 \rightarrow i_2, i_2 \rightarrow i_1$.

Find the value of confidence of the rule Rul.

If Confidence (Rul) $> min_conf$ THEN,

Add the Rul to Sen_Rul

End //If

End //For

7. For each Rul in Sen_Rul { until the sensitive rule is open}
For each Val_ColR of rule {
If $Val_ColR > 0.5$ and $Val_ColR > Val_ColL$
 $Val_ColR = 1 - Val_ColR$
End //If

- End //For
8. Re-Compute the value of confidence of rule *Rul*.
 If rule *Rul* (confidence) > min_conf
 For each Val_ColR of the rule
 If Val_ColR = 1.0
 Val_ColR = 0.0
 End //If
 End //For
 Else go to Step-9
 End //If
 9. Convert the database FUZZ to *D'* and output is updated to *D'*
 10. End

Step -1

The resultant of the rough set works as an input for DSR Algorithm.

Table VI Optimized dataset

	Col-0	Col-8
Rec-1	6	3
Rec-2	3	4
Rec-3	2	1
Rec-4	6	2
Rec-5	3	1

Step -2

Fuzzification is applied on the dataset (Table VI) by using the Triangular membership function i.e.

$$FUZZ = \text{Max} \left(\min \left(\frac{(x-a)}{b-a} + \frac{(c-x)}{c-b} \right), 0 \right)$$

After fuzzification calculate the value of support for each column by taking the summation of all the value of particular column as shown in Fuzzified dataset (Table VII).

Table VII Fuzzified Dataset

	Col-0			Col-8		
	Col-0 _z	Col-0 _o	Col-0 _b	Col-8 _z	Col-8 _o	Col-8 _b
Rec-1	0.8	0.2	0.0	0.6	0.0	0.0
Rec-2	0.6	0.0	0.0	0.8	0.0	0.0
Rec-3	0.4	0.0	0.0	0.2	0.0	0.0
Rec-4	0.8	0.2	0.0	0.4	0.0	0.0
Rec-5	0.8	0.0	0.0	0.2	0.0	0.0
Count	3.4	0.4	0.0	2.2	0.0	0.0

Step -3

The minimum support is considered to be 2.0 and minimum confidence to be 50%. On comparing the count value of each column with the minimum support, the observation made is that the support of Column Col-0_z and Col-8_z has greater value than 2. So the rules of the columns will be Col-0_z → Col-8_z, Col-8_z → Col-0_z. Consider the critical rule Col-0_z → Col-8_z has to be hidden.

Support of the rule is calculated as Support (Col-0_z → Col-8_z) = min (Col-0_z, Col-8_z) as shown in Table VIII.

Table VIII Support of rules

	Col-0 _z	Col-8 _z	Support
Rec-1	0.8	0.6	0.6
Rec-2	0.6	0.8	0.6
Rec-3	0.4	0.2	0.2
Rec-4	0.8	0.4	0.4
Rec-5	0.8	0.2	0.2
Count	3.4		2.0

Step -4

Confidence of the rules is calculated as:

$$\text{Confidence (P} \rightarrow \text{Q)} = \frac{\text{Support (PQ)}}{\text{Support (P)}}$$

So, the confidence is calculated as:

$$\text{Confidence (A} \rightarrow \text{B)} = \frac{\text{Support (AB)}}{\text{Support (A)}}$$

Confidence ($\text{Col-0}_z \rightarrow \text{Col-8}_z$) = $2.0 / 3.4 = 58.8\%$

Step -5

To hide the sensitive information the confidence is decreased by reducing the value of right hand side. Thus to hide the rule $\text{Col-0}_z \rightarrow \text{Col-8}_z$ the support of ($\text{Col-0}_z, \text{Col-8}_z$) is reduced by subtracted the value of Col-8_z from 1 when the value is greater than 0.5 or its respective left hand side column. According to procedure the value of Rec-1, Rec-2 is reduced as shown in Table IX.

Table IX Modified Dataset.

	Col-0_z	Col-8_z	Support
Rec-1	0.8	0.4	0.4
Rec-2	0.6	0.2	0.2
Rec-3	0.4	0.2	0.2
Rec-4	0.8	0.4	0.4
Rec-5	0.8	0.2	0.2
Count	3.4		1.4

Now the confidence will be,

Confidence ($\text{Col-0}_z \rightarrow \text{Col-8}_z$) = $1.4 / 3.4 = 41\%$

V. Conclusion

We are using the attribute reduction (feature selection) method to get the sensitive Association Rules hidden. For the same, we considered a raw dataset and obtained preprocessed data from this using the techniques from the data mining. We applied the Rough set on this table to get a reduced reduct core which on generation helps in formation of the reduced dataset for the Fuzzification. This obtained detail is carried out over by triangular method of Fuzzification further applying the DSR algorithm give the minimum support and confidence of the derived rules. It shows that the hiding of sensitive Association Rules can be performed fast and efficiently with the application of Rough Set reduct core along with the DSR Algorithm. Further we will do try using this method along with the PSO (Particle Swarm Optimization) Algorithm.

References

- [1]. G.Sudha Sadasivam, S.Sangeetha, K.Sathya Priya, "Privacy Preservation with Attribute Reduction in Quantitative Association Rules using PSO and DSR", Special Issue of International Journal of Computer Applications (0975 – 8887) on Information Processing and Remote Computing – IPRC, August 2012.
- [2]. Pawlak. Z, "Rough Sets," International Journal of Informational and Computer Science, vol. 11, pp. 341-356, 1982.
- [3]. Pawlak. Z, "Rough sets theory and its application to data analysis," Cybernetics and System, vol. 29, pp. 661-688, 1998.
- [4]. Manoj Gupta, R. C. Joshi, "Privacy Preserving Fuzzy Association Rules Hiding in Quantitative Data", International Journal of Computer Theory and Engineering, Vol. 1, No. 4, October, 2009 1793-8201.
- [5]. G.Y. Wang, "Rough Set Theory and Data Mining". Xi'an Jiaotong University Press, Xi'an, 2001.
- [6]. R.Jenshen, Q.Shen, (2004b) "Semantics-Preserving Dimensionality Reduction: Rough and Fuzzy-Rough- Based Approaches", IEEE Transactions on Knowledge and Data Engineering, 2004, Vol.16, Issue:12, pp. 1457- 1471.
- [7]. G.Y. Wang; J. Zhao; J.J.An, et al.(2004) "Theoretical study on attribute reduction of rough set theory: comparison of algebra and information views." In: Proceedings of the Third IEEE International Conference on Cognitive Informatics, pp-148 – 155.
- [8]. Neil Mac Parthala'in; Qiang Shen, Richard Jensen (2010) , "A Distance Measure Approach to Exploring the Rough Set Boundary Region for Attribute Reduction", IEEE transaction on Knowledge and data engineering, Vol.22, No.3, pp. 305-317.
- [9]. L. Zadeh, "Fuzzy sets," Information Control, vol. 8, pp. 338–353, 1965.
- [10]. 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.