Chapter 4 Proposed Procedures

In this study, we propose a discritization algorithm, named Extended Chi2 algorithm for the VPRS model to discrete real value attributes for deriving classification rules. We also develop an effective approach to select β-reducts in the datasets.

4.1 A Discretization Algorithm

The modified Chi2 algorithm utilizes the quality of approximation (Chmielewski et al. 1996), in which it considers the effect of degrees of freedom. However, there are two shortcomings of this algorithm that should be overcome.

First, the rough sets approach is inspired by the notion of inadequacy in the available information to perform a complete classification of objects; that is, to perform a complete classification requires that the collected data must be fully correct or certain. Nevertheless, in real-world decision making, the objects of classes often overlap, suggesting that predictor information may be incomplete. Thus, we need a new method to determine the inconsistency rate to replace the quality of approximation in the RST.

Ziarko (1993) defined the measure of the inconsistency rate of the set X with respect to Y as:

$$c(X,Y)=\begin{cases}1-\dfrac{card(X\bigcap Y)}{card(X)} & \text{if} \quad card(X)>0\\ 0 & \text{if} \quad card(X)=0\end{cases}$$

Here, card denotes set cardinality.

We utilize a simple method to determine the inconsistency rate in the VPRS, which is based on the least upper bound $\xi(C,D)$ of the data set, where C is the equivalence

relation set, D is the decision set, and $C^* = \{E_1, E_2, ..., E_n\}$ is the equivalence classes. According to Ziarko (1993) the specified majority requirement, the admissible classification error β, must be within the range $0 \le \beta < 0.5$. Since we determine the β value in the VPRS model, which is based on the least upper bound $\xi(C, D)$ of the data set, if one chooses the max value in m_1 and the min value in m_2 then this leads to the calculated $\xi(C, D) < \beta^*$ (β^*: the exact classification error of data set), which cannot be discernible the data set. Therefore, we propose to choose the min value in m_1 and the max value in m_2. The following equality is used for calculating the least upper bound of the data set.

$$\xi(C, D) = \max(m_1, m_2), \tag{4.1}$$

where

$$m_1 = 1 - \min\{c(E, D) \mid E \in C^* \text{ and } \theta < c(E, D)\}$$

$$m_2 = \max\{c(E, D) \mid E \in C^* \text{ and } c(E, D) < \theta\}.$$

$$c(E, D) = 1 - \frac{card(E \cap D)}{card(E)}.$$

θ : denotes the threshold, which is determined by the decision maker based on the relative degree. Usually, θ is set at 0.5.

In the extended Chi2 algorithm, inconsistency checking (InConCheck (data)$< \delta$) of the Chi2 algorithm is replaced by the lease upper bound ξ after each step of discretization ($\xi_{discretized} < \xi_{original}$). By doing this, the inconsistency rate is utilized as the termination criterion.

Secondly, Tay, et al. (2002) proposed that the difference in degrees of freedom must be considered if there exists a χ^2 value calculated from the adjacent two intervals (I) and the threshold difference is greater than the other χ^2 value calculated from the adjacent two intervals and threshold difference. This means that the independence of the adjacent two intervals (I) is greater than the other adjacent intervals. In this case we

suggest that the adjacent two intervals (*I*) should be merged first.

Although the modified Chi2 algorithm considers the effect of the degrees of freedom, this algorithm only considers the difference in the χ^2 value and the threshold. It ignores the effect of variance in the two merging intervals. From the view of statistics, the compared baseline is not equal, and the interpretation is depicted as follows: Consider when we have a pair of two adjacent intervals. By formula (4.2), the first two adjacent intervals of the χ^2 value are 3.94, while the corresponding threshold is 7.344 (degrees of freedom $\nu = 8$; significant level $\alpha = 0.5$), the difference between the χ^2 value and the corresponding threshold is 3.404, the second two adjacent intervals of the χ^2 value are 0.54, while the corresponding threshold is 3.357 (degrees of freedom $\nu = 4$; significant level $\alpha = 0.5$), and the difference in the χ^2 value and the threshold is 2.817. If the variance in the two adjacent intervals is considered, the normalized difference ($= difference / \sqrt{2 * \nu}$) in the first two adjacent intervals is 0.851; the normalized difference in the second two adjacent intervals is 0.996. Therefore, the second two adjacent intervals should be merged.

The extended Chi2 algorithm

Step 1. Initialize:

Set the significant level as α=0.5; calculate the pre-defined inconsistency rate ξ.

Step 2. Calculate the chi-square value:

For each numeric attribute, sort data on the attribute and use formula (4.2) to compute the x^2 value.

Step 3. Merge:

For a Comparison, compute the x^2 value and corresponding threshold; merge the adjacent two intervals which have the maximal normalize

difference and the computed x^2 value is smaller than the corresponding threshold. If no adjacent two intervals satisfy this condition, then go to Step 5.

Step 4. Check inconsistency rate for merger:

Check the merged inconsistency rate, and if the merged inconsistency rate exceeds the pre-defined inconsistency rate, then discard the merger. Go to step 5. Otherwise, go to step 2.

Step 5. Decrease the significance level:

Decrease $\alpha \to \alpha_0$.

Step 6. Calculate finer the chi-square value:

For each numeric attribute, sort data on the attribute and use formula (4.2) to compute the x^2 value.

Step 7. Finer merge:

For a comparison, compute the x^2 value and corresponding threshold; merge the adjacent two intervals which have a maximal normalize difference and the computed x^2 value is smaller than the corresponding threshold. If no adjacent two intervals satisfy this condition, then go to Step 9.

Step 8. Check the inconsistency rate much finer for a merger:

Check the merged inconsistency rate; if the merged inconsistency rate exceeds the pre-defined inconsistency rate, then discard merge. Go to step 9. Otherwise, go to step 6.

Step 9. Decrease finer the significance level:

Decrease the significance level; then stop.

The formula for computing the χ^2 value is:

$$\chi^2 = \sum_{i=1}^{n} \sum_{j=1}^{k} \frac{(A_{ij} - E_{ij})^2}{E_{ij}}, \qquad (4.2)$$

where $n = 2$ (two intervals being compared);

k = number of classes;

Aij = number of objects in the ith interval, jth class;

R_i = number of objects in the ith interval $= \sum_{j=1}^{k} A_{ij}$;

C_j = number of objects in the jth class $= \sum_{i=1}^{n} A_{ij}$;

N = total number of objects $= \sum_{i=1}^{n} R_i$;

E_{ij} = expected frequency of $A_{ij} = \frac{R_i * C_j}{N}$.

If either R_i or C_j is 0, then E_{ij} is set to 0.1. The degree of freedom of the χ^2 statistic is one less than the number of classes

4.2 An Approach for the β-reducts

When performing a VPRS analysis, how the β-reducts are selected is a key point of the process. The precision parameter value can control the choice of β-reducts. Previous related research studies lacked an effective method to determine a precision parameter value. Ziarko (1993) defined the measure of the relative degree of misclassification of the set X with respect to Y as:

$$c(X,Y) = \begin{cases} 1 - \dfrac{card(X \cap Y)}{card(X)} & \text{if} \quad card(X) > 0 \\ 0 & \text{if} \quad card(X) = 0 \end{cases}$$

where card denotes set cardinality.

Let X and Y be non-empty subsets of U. We say that X is included in Y, if for every element that belongs to X, then that also belongs to Y. The measure of relative misclassification can define the inclusion relationship between X and Y without explicitly using a general quantifier:

$$Y \supseteq X \Leftrightarrow c(X,Y) = 0.$$

According to the specified majority requirement, the admissible β must be within the range $0 \le \beta < 0.5$. Based on this assumption the majority inclusion relation is defined as:

$$Y \overset{\beta}{\supseteq} X \Leftrightarrow c(X,Y) \le \beta.$$

The above definition covers the entire family of β-majority relations.

In this study we propose an effective approach to select the β-reducts, which involves two steps:

Step 1: Obtain the candidates of β-reducts using precision parameter (β)

The determination of the β value in the VPRS model is based on the least upper bound $\xi(C,D)$ of the data set, where C is the condition attributes set, D is the decision attributes set, and $C^* = \{E_1, E_2, ..., E_n\}$ is the equivalence classes.

Step 2: Find the β-reducts

(1) For each candidate of β-reducts (subset P), calculate the quality of classification based on (2.2).

(2) Remove redundant attributes

Let subset $X \subseteq P$. For each subset ($\gamma(C,D,\beta) = \gamma(P,D,\beta)$), if $\gamma(X,D,\beta) = \gamma(C,D,\beta)$, then remove the attributes $P \setminus X$. Otherwise, do not remove any attribute from subset P.

(3) Find the β-reducts

Any subset X, which has $\gamma(X,D,\beta) = \gamma(C,D,\beta)$ is a β-reduct.