Concise Papers

An Extended Chi2 Algorithm for Discretization of Real Value Attributes

Chao-Ton Su and Jyh-Hwa Hsu

Abstract—The Variable Precision Rough Sets (VPRS) model is a powerful tool for data mining, as it has been widely applied to acquire knowledge. Despite its diverse applications in many domains, the VPRS model unfortunately cannot be applied to real-world classification tasks involving continuous attributes. This requires a discretization method to preprocess the data. Discretization is an effective technique to deal with continuous attributes for data mining, especially for the classification problem. The modified Chi2 algorithm is one of the modifications to the Chi2 algorithm, replacing the inconsistency check in the Chi2 algorithm by using the quality of approximation, coined from the Rough Sets Theory (RST), in which it takes into account the effect of degrees of freedom. However, the classification with a controlled degree of uncertainty, or a misclassification error, is outside the realm of RST. This algorithm also ignores the effect of variance in the two merged intervals. In this study, we propose a new algorithm, named the extended Chi2 algorithm, to overcome these two drawbacks. By running the software of See5, our proposed algorithm possesses a better performance than the original and modified Chi2 algorithms.

Index Terms—VPRS model, RST, data mining, discretization.

1 Introduction

Deriving classification rules is an important task in data mining. As such, discretization is an effective technique in dealing with continuous attributes for rule generating. Many classification algorithms require that the training data contain only discrete attributes, and some work better on discretized or binarized data [9], [8]. However, for these algorithms, discretizing continuous attributes is a first step for deriving classification rules. The Variable Precision Rough Sets (VPRS) model is one example. The VPRS model is a powerful mathematical tool for data analysis and knowledge discovery from imprecise and ambiguous data. Although the theory of VPRS has been successfully applied to diverse areas, such as corporate failure prediction, identification of low-paying workplaces, and Web searching [2], [3], [19], it cannot conduct continuous data without discretization. Thus, this requires studies on appropriate discretization methods.

There are three different axes by which discertization methods can be classified: local versus global, supervised versus unsupervised, and static versus dynamic [5]. Local methods, such as ID3 (interactive dichotomizer 3, Quinlan 1983), produce partitions that are applied to localized regions of the instance space. By contrast, the global discertization method uses the entire instance space to discretize. Several discretization methods, such as equal width interval and equal frequency interval methods, do not utilize instance class labels in the discretization process. These methods are called unsupervised methods. Conversely, discretization methods that utilize the class labels are referred to as supervised methods.

- *C.-T. Su is with the Department of Industrial Engineering and Engineering Management, National Tsing Hua University, Hsinchi, Taiwan 101, Kuang Fu Road, Sec. 2, Hsinchu 300, Taiwan. E-mail: ctsu@mx.nthu.edu.tw.*
- *J.-H. Hsu is with the Department of Industrial Engineering and Management, National Chiao Tung University, 1001 Ta Hsueh Road, Hsinchu, Taiwan, Republic of China. E-mail: d9033809.iem90g@nctu.edu.tw.*

Manuscript received 29 Aug. 2003; revised 23 Mar. 2004; accepted 17 May 2004; published online 19 Jan. 2005.
For information on obtaining reprints of this article, please send e-mail to: tkde@computer.org, and reference IEEECS Log Number TKDE-0165-0803.

Many discretization methods require some parameter, m, indicating the maximum number of intervals that are produced in discretizing an attribute. Static methods, such as entropy-based partitioning, perform one discretization pass of the data for each attribute and determine the value of m for each attribute independent of the other attributes. Dynamic methods conduct a search through the space of possible m values for all attributes simultaneously, thereby capturing interdependencies in attribute discretization. A number of methods based on the entropy measure establish a strong group of works in the discretization domain. This concept uses class entropy as a criterion to evaluate a list of best cuts, which together with the attribute domain induce the desired intervals [14].

Holte [6] proposed a one-level decision tree algorithm, called 1RD (One Rule Discretizer), which attempts to greedily divide the attribute range into a number of intervals, using a constraint that each interval must include at least the user-specified minimum number of values. It starts with an initial partition into intervals, each containing the minimum number of values, and then moves the initial partition boundary (cut), by adding the attribute values, so that each interval contains a strong majority of objects from one decision class. Nguyen and Skowron [13] offered an approach dealing with the discretization problem, which is based on a rough set and Boolean reasoning and the computational complexity of which is $O(n^3k)$, where n is the number of objects and k is the number of attributes. The main results state that the problem of optimal discretization of real value attributes is polynomially reducible to the problem of minimal reduct finding, and so it is NP-hard.

Nguyen [12] considered a general genetic strategy-based algorithm of searching for an optimal set of separating hyperplanes by a genetic algorithm. In the case of a consistent decision table (where the misclassification rate equals 0), the algorithm will be continued until the hyperplants cut the space R^k into regions containing objects from one decision class only. Nguyen [14] wrote of the relationship between the reduct problem in the rough set and the problem of real value attribute discretizaition, which searches for a minimal set of cuts on attribute domains that preserve the discernibility of objects with respect to any chosen attributes' subset of cardinality t (t denotes a parameter given by the user). Such a discretization procedure assures that one can keep all reducts consisting of at least t attributes.

The ChiMerge algorithm introduced by Kerber [8] is a supervised global discertization method. The user has to provide several parameters such as the significance level α, and the maximal intervals and minimal intervals during the application of this algorithm. ChiMerge requires α to be specified. Nevertheless, too big or too small, a α will overdiscretize or underdiscretize an attribute. Liu and Setiono [10] proposed a Chi2 algorithm that uses a ChiMerge algorithm as a basis, whereby the Chi2 algorithm improves the ChiMerge algorithm in that the value of α is calculated based on the training data itself.

Tay and Shen [17] indicated that, although the Chi2 algorithm automates the ChiMerge algorithm by calculating a significance value α based on the training data set, it still has two drawbacks: 1) The Chi2 algorithm requires the user to provide an inconsistency rate to stop the merging procedure. This is unreasonable since an inappropriate threshold will result in overmerging. 2) This merging criterion does not consider the degrees of freedom, but rather only the fixed degrees of freedom (the classes' number minus one). According to the statistical point of view, this is inaccurate [11], since the power of a statistical test is affected by the degrees of freedom of a test. They utilize the quality of approximation to replace the inconsistency checking of the Chi2 algorithm and consider the degrees of freedom of each two adjacent intervals, in which the two adjacent intervals when it has a maximal difference in the calculated χ^2 value and the threshold should be merged first.

The rough sets approach is inspired by the notion of inadequacy of 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. However, in real-world decision making, the objects of classes often overlap, suggesting that predictor information may be incomplete.

In this study, we propose a method to determine the predefined inconsistency rate (δ) based on the least upper bound of data misclassification error. In addition, the effect of variance in the two merging intervals is considered. These two remedies can conquer the drawbacks of the Chi2 algorithm. The effectiveness of our proposed discretization method, named the extended Chi2 algorithm, is demonstrated by three numerical data sets. Comparing the implementation results with the original Chi2 algorithm and modified Chi2 algorithm using See5, the extended Chi2 algorithm performs better than the original and modified Chi2 algorithms.

2 Variable Precision Rough Sets Model

The variable precision rough sets model was introduced by [18] and is an extension of the original Rough Set Theory (RST) as a tool for classification of objects. This is an important extension since, as noted by Kattan and Cooper [7], "In real-world decision making, the patterns of classes often overlap, suggesting that predictor information may be incomplete... This lack of information results in probabilistic decision making, where perfect prediction accuracy is not expected."

VPRS deals with partial classification by introducing a precision parameter β (in the original rough set, the β value is zero). The β value represents a bound on the conditional probability of a proportion of objects in a condition, where the objects are classified to the same decision class. Ziarko [18] defined the β value as a classification error with a range in the domain $[0.0, 0.5)$. However, An et al. [1] and Beynon [2] considered β to denote the proportion of correct classifications, in which case the appropriate range is $(0.5, 1.0]$. In this study, we use the Ziarko notion.

VPRS operates on what may be described as a knowledge representation system or information system. An information system (S) consisting of four parts is shown as:

$$S = (U, A, V, f),$$

where U is a nonempty set of objects:

- A is the collection of objects; we have $A = C \bigcup D$ and $C \bigcap D = \phi$, where C is a nonempty set of condition attributes, and D is a nonempty set of decision attributes.
- V is the union of attribute domains, i.e., $V = \bigcup_{a \in A} V_a$, where V_a is a finite attribute domain and the elements of V_a are called values of attribute a;
- F is an information function such that $f(u_i, a) \in V_a$ for every $a \in A$ and $u_i \in U$.

Every object that belongs to U is associated with a set of values corresponding to the condition attributes C and decision attributes D.

2.1 β-lower and β-upper Approximations

Suppose that information system $S = (U, A, V, f)$, with each subset $Z \subseteq U$ and an equivalence relation R that is referred to as an indiscernibility relation, corresponds to a partitioning of U into a collection of equivalence classes $R^* = \{E_1, E_2, \ldots, E_n\}$. We will assume that all sets under consideration are finite and nonempty [19]. The variable precision rough sets approach to data analysis hinges on two basic concepts, namely, the β-lower and the β-upper approximations of a set. The β-lower and the β-upper approximations can also be presented in an equivalent form as shown below:

The β-lower approximation of the set $Z \subseteq U$ and $P \subseteq C$:

$$\underline{C_\beta}(D) = \bigcup_{1 - P_r(Z|x_i) \leq \beta} \{x_i \in E(P)\}.$$

The β-upper approximation of the set $Z \subseteq U$ and $P \subseteq C$

$$\overline{C_\beta}(D) = \bigcup_{1 - P_r(Z|x_i) < 1 - \beta} \{x_i \in E(P)\},$$

where $E(\bullet)$ denotes a set of equivalence classes (in the above definitions, they are condition classes based on a subset of attributes P).

$$Z \subset E(D); \quad P_r(Z|x_i) = \frac{card(Z \bigcap x_i)}{card(x_i)}.$$

2.2 Majority Inclusion Relation

The heart of the VPRS model is the generalization of the notion of the standard set inclusion relation. The extended notion should be able to allow for some degree of misclassification in the largely correct classification.

Let X and Y be nonempty subsets of a finite universe U. We say that X is included in Y, or $X \subseteq Y$, if every $a \in X$ implies $a \in Y$. Clearly, there is no room for even the slightest misclassification according to this definition. Therefore, before a more general definition is presented, it is convenient to introduce the measure $c(X, Y)$ of the relative degree of misclassification of the set X with respect to set Y, which is defined as:

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

Here, card denotes set cardinality.

If we classify all elements of the set X into set Y, then in $c(X, Y) * 100\%$ of the cases, we would make a classification error. Consequently, the quantity $c(X, Y)$ will be referred to as the relative classification error.

3 Modified Chi2 Algorithm

The modified Chi2 algorithm introduced by Shen and Tay [16] can be sectioned into two phases: The first phase of the modified Chi2 algorithm can be regarded as a generalization version of the ChiMerge algorithm. Instead of specifying a χ^2 threshold, the modified Chi2 algorithm provides a wrapping that automatically increments the χ^2 threshold (decreasing the significant level α). A consistency check is used as a stopping criterion to make sure that the modified Chi2 algorithm automatically determines a proper χ^2 threshold while still keeping the fidelity of the original data.

The second phase is a finer process of the first phase, beginning with the significant level α_0 determined in the first phase, where each attribute i is associated with a $\mathrm{sigLvl}[i]$ and they take turns for merging. A consistency check is conducted after each attribute's merging. If the inconsistency rate does not exceed the predefined inconsistency rate (δ), then $\mathrm{sigLvl}[i]$ is decreased for attribute i's next round of merging. Otherwise, the attribute i will not be involved in further merging. This process is repeated until no attribute's value can be merged.

In the modified Chi2 algorithm, inconsistency checking (InConCheck (data) $< \delta$) of the original Chi2 algorithm is replaced by the quality of approximation L_c after each step of discretization ($L_{c\text{-}discretized} \leq L_{c\text{-}original}$). This inconsistency rate is utilized as the termination criterion. The quality of approximation coined from the Rough Sets Theory is defined as follows:

$$L_c = \frac{\sum card(\underline{B}X_i)}{card(U)}, \tag{1}$$

where U is the set of all objects of the data set:

- X can be any subset of U.
- $\underline{B}X$ is the lower approximation of X in B ($B \subseteq A$).
- A is the set of attributes. The card denotes set cardinality.

The merge criterion of the original Chi2 algorithm does not consider the degrees of freedom, as it only used the fixed degrees of freedom (the classes' number minus one). The original Chi2 algorithm merges the pair of adjacent intervals with the lowest x^2 value being the critical value. The merge criterion of modified Chi2 considers the degrees of freedom of each of the two adjacent intervals. When two adjacent intervals have a maximal difference in the calculated χ^2 value, the threshold should be merged first.

The Chi2 algorithm is shown as follows:

```
Phase 1:
    Set α = 0.5;
    do while (InConCheck (data)< δ)
    {  for each numeric attribute
        {    Sort (attribute, data); /* sort data
                    on attribute */
             Chi-sq-init (attribute, data);
             do
             { Chi-sq-calculation (attribute, data);
              } while (Merge (data))
        }
        α0 = α;
        α = decreSigLevel(α);
    }
Phase 2:
    Set all SigLvl[i] = α0 for attribute i;
    do until no attribute can be merged
    {  for each mergeable attribute i
        {    Sort (attribute, data); /* sort data on
                    attribute */
             Chi-sq-init (attribute, data);
             do
             { Chi-sq-calculation (attribute, data);
              } while (Merge (data))
              if (InConCheck (data) < δ)
              SigLvl [i]=decreSigLevel (SigLvl [i]);
              else attribute i is no mergeable;
        }
    }
```

4 Formulation of Extended Algorithm

The modified Chi2 algorithm utilizes the quality of approximation [4], in which it considers the effect of degrees of freedom. 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 [18] defined the measure of the inconsistency rate of the set X with respect to Y as:

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

Here, card denotes set cardinality.

In this study, 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 $C^* = \{E_1, E_2, \ldots, E_n\}$ is the equivalence classes. According to [18], the specified majority requirement (the admissible classification error β) must be within the range $0 \leq \beta < 0.5$. 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 in 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.

$$\xi(C, D) = \max(m_1, m_2), \tag{2}$$

where

$$m_1 = 1 - \min\{c(E, D) | E \in C^* \text{ and } 0.5 < c(E, D)\},$$

$$m_2 = \max\{c(E, D) | E \in C^* \text{ and } c(E, D) < 0.5\}.$$

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

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.

Second, Tay and Shen [17] 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 regards 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 (3), 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, then 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 $\alpha = 0.5$; calculate the predefined inconsistency rate ξ.

Step 2. Calculate the chi-square value:

For each numeric attribute, sort data on the attribute and use (3) 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 normalized difference and the computed x^2 value is smaller than the corresponding threshold. If no two adjacent

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 predefined inconsistency rate, then discard the merger. Go to Step 5. Otherwise, go to Step 2.

Step 5. Decrease the significance level:
Decrease $\alpha \rightarrow \alpha_0$.

Step 6. Calculate finer the chi-square value:
For each numeric attribute, sort data on the attribute and use (3) 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 two adjacent 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 predefined inconsistency rate, then discard the merger. 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:

$$\chi^2 = \sum_{i=1}^{n} \sum_{j=1}^{k} \frac{(A_{ij} - E_{ij})^2}{E_{ij}}, \quad (3)$$

where

- $n = 2$ (two intervals being compared);
- k = number of classes;
- A_{ij} = 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 degrees of freedom of the χ^2 statistic are one less than the number of classes.

5 Experimental Results

Five data sets are demonstrated to present the effectiveness of the proposed extended Chi2 algorithm. The five data sets are taken from the University of California, Irvine's repository of machine learning databases (http:// www.ics.uci.edu/~mlearn/ MLSummary.html).

5.1 The Data

The five data sets used in the experiment are the Bupa Liver Disorders, the Glass Types, the Heart Disease, the Iris Plants, and the Breast Cancer. They have different types of attributes. The Bupa Liver Disorders data, Glass Types data, and the Iris Plants data are of the type with continuous attributes, the Breast Cancer data are of ordinal discrete ones, while the Heart Disease data shows mixed attributes (numeric and discrete). The five data sets are described below:

1. **The Bupa Liver Disorders Data**. This data set contains 345 instances (145 instances that are normal; 200 instances of a liver malfunction), where each instance is described using six numeric attributes: MCV, ALKPHOS, SGPT, SGOT, GAMMAGT, and DRINKS.
2. **The Glass Types Date**. This data set contains 214 instances (70 instances of building windows that are float processed, 76 instances of building windows that are nonfloat processed, 17 instances of vehicle windows float processed, 13 instances of containers, 9 instances of tableware, 29 instances of headlamps), each instance is described using nine numeric attributes: RI, NA, MG, AL, SI, K, CA, BA, and FE.
3. **The Iris Plants Data**. This data set contains 150 instances (50 instances of setosa, 50 instances of versicolor, 50 instances of verginica); each instance is described using four numeric attributes: sepal-length, sepal-width, petal-length, and petal-width.
4. **The Breast Cancer Data**. This data set contains 699 instances, where 16 instances have missing attributes values. Removing instances with missing attributes values, we use 683 instances (444 instances of benign, 239 instances of malignant), where each instance is described using nine attributes: clump thickness, uniformity of cell size, uniformity of cell shape, marginal adhesion, single epithelial cell size, bare nuclei, bland chromatin, normal nucleoli, and mitoses.
5. **The Heart Disease Data**. This data set contains 297 instances (160 instances of 0, 54 instances of 1, 35 instances of 2, 35 instances of 3, 13 instances of 4), where each instance is described using eight nominal attributes: SEX, CP, FBS, RESTECG, EXANG, SLOPE, CA, and THALPUL; and five numeric attributes: AGE, TRESTBPS, CHOL, THALACH, and OLDPEAK.

5.2 Experimental Results

We ran See5 on both the original data sets and the discretized data sets. The parameters of See5 utilize its default setting. The 10-fold cross-validation test method is applied to all data sets. The data set is divided into 10 parts of which nine parts are used as training sets and the remaining one part as the testing set. The experiments were repeated 10 times. The final predictive accuracy is taken as the average of the 10 predictive accuracy values.

The extended Chi2 algorithm is compared with the original Chi2 algorithm and modified Chi2 algorithm with the predefined inconsistency rate (δ) value equal to 0 in the experiment. The experimental process includes two steps:

Step 1: **Discretization**. All five data sets are discretized using the original Chi2 algorithm, the modified Chi2 algorithm, the extended Chi2 algorithm, and Boolean Reasoning algorithm.

Step 2: **Comparison**: The discretized data sets are sent into See5. The predictive accuracy and its standard deviation of these methods are listed in Table 1. From Table 1, we know that the predictive accuracy of the extended Chi2 algorithm outperforms other discretization algorithms.

The tree sizes using See5 with different discretization methods shown in Table 2. From Table 2, we know that, although the extended Chi2 algorithm has no significant difference in tree size compared to the original and modified Chi2 algorithms, it is in fact significantly smaller than when using the original data with See5.

6 Conclusion

Many classification algorithms developed in the data mining community can only acquire knowledge on the nominal attributes' data sets. However, many real-world classification tasks exist that involve continuous attributes, such that these algorithms cannot be applied unless the continuous attributes are discretized. The VPRS model is a powerful mathematical tool for data analysis and knowledge discovery from inconsistent and ambiguous data. It cannot be applied to extract rules from the continuous attributes unless they are first discretized.

In this study, we propose an extended Chi2 algorithm that determines the predefined misclassification rate (δ) from the data itself. We also consider the effect of variance in the two adjacent intervals. With these modifications, the extended Chi2 algorithm

TABLE 1
The Predictive Accuracy Using See5 with the Discretization Algorithm

Data Set	See5				
	Continuous	Original Chi2 Algorithm	Modified Chi2 Algorithm	Extended Chi2 Algorithm	Boolean Reasoning Algorithm
Bupa	$67.5 \pm 2.4\%$	$65.2 \pm 3.2\%$	$67.5 \pm 1.9\%$	$68.4 \pm 2.7\%$	$68.1 \pm 2.3\%$
Glass	$68.6 \pm 2.5\%$	$93.1 \pm 2.1\%$	$93.4 \pm 2.3\%$	$93.5 \pm 1.3\%$	$71.9 \pm 2.8\%$
Iris	$94.0 \pm 2.1\%$	$94.0 \pm 2.1\%$	$93.3 \pm 2.2\%$	$94.0 \pm 2.1\%$	$96.0 \pm 1.8\%$
Breast Cancer	$94.9 \pm 0.8\%$	$95.5 \pm 1.0\%$	$96.0 \pm 0.9\%$	$96.5 \pm 0.8\%$	$95.2 \pm 0.8\%$
Heart dissease	$51.9 \pm 1.4\%$	$52.5 \pm 2.3\%$	$53.2 \pm 2.7\%$	$54.2 \pm 1.7\%$	$55.9 \pm 2.6\%$

TABLE 2
The Tree Size Comparison of the Five Methods

Data Set	See5				
	Continuous	Original Chi2 Algorithm	Modified Chi2 Algorithm	Extended Chi2 Algorithm	Boolean Reasoning Algorithm
Bupa	27.1 ± 1.7	15.9 ± 1.0	12.7 ± 0.6	9.9 ± 0.6	30.3 ± 2.1
Glass	24.0 ± 0.7	9.9 ± 0.1	9.8 ± 0.1	9.2 ± 0.2	23.5 ± 0.9
Iris	4.6 ± 0.2	3.7 ± 0.2	3.0 ± 0.0	3.0 ± 0.0	3.9 ± 0.1
Breast Cancer	10.3 ± 0.9	8.6 ± 0.7	8.8 ± 0.8	9.1 ± 0.8	9.2 ± 0.3
Heart dissease	46.0 ± 0.9	34.8 ± 1.7	34.1 ± 0.7	36.0 ± 1.1	34.7 ± 1.2

not only handles misclassified or uncertain data, but also becomes a completely automated discretization method and its predictive accuracy is better than the original Chi2 algorithm.

Five real-world data set experiments were conducted to demonstrate the feasibility of the proposed algorithm. The experimental results show that our proposed algorithm could acquire a higher predicted accuracy than the original and modified Chi2 algorithm. Furthermore, the tree size is significantly smaller than using the original data with See5.

For m attributes, the computational complexity of original Chi2 algorithm at phase 1 has $O(Kmn \log n)$, where n is the number of objects in the data set, and K is the number of incremental steps. A similar complexity can be obtained for phase 2. Although our proposed algorithm adds one step (i.e., to select the merging intervals), it does not increase the computational complexity as compared to the original Chi2 algorithm. The computational complexities of the original Chi2 algorithm, modified Chi2 algorithm, and our proposed algorithm are the same.

Further work should develop a merger criterion to effectively reduce the computational complexity of the extended Chi2 algorithm.

REFERENCES

[1] A. An, N. Shan, C. Chan, N. Cercone, and W. Ziarko, "Discovering Rules for Water Demand Prediction: An Enhanced Rough-Set Approach," *Eng. Applications in Artificial Intelligence,* vol. 9, no. 6, pp. 645-653, 1996.

[2] M. Beynon, "Reducts within the Variable Precision Rough Sets Model: A Further Investigation," *European J. Operational Research,* vol. 134, pp. 592-605, 2001.

[3] M. Beynon, "The Identification of Low-Paying Workplaces: An Analysis Using the Variable Precision Rough Sets Model," *Proc. Third Int'l Conf. Rough Sets and Current Trend in Computing,* pp. 530-537, 2002.

[4] R. Chmielewski and W. Grzymala-Busse, "Global Descretization of Continuous Attributes as Preprocessing for Machine Learning," *Int'l J. Approximate Reasoning,* vol. 15, no. 4, pp. 319-331, 1996.

[5] J. Dougherty, R. Kohavi, and M. Sahami, "Supervised and Unsupervised Discretization of Continuous Features," *Machine Learning: Proc. 12th Int'l Conf.,* pp. 194-202, 1995.

[6] R.C. Holte, "Very Simple Classification Rules Perform Well on Most Commonly Used Datasets," *Machine Learning,* vol. 11, no. 1, pp. 63-91, 1993.

[7] M.W. Kattan and R.B. Cooper, "The Predictive Accuracy of Computer-Based Classification Decision Techniques. A Review and Research Directions," *Omega-Int'l J. Management Science,* vol. 26, no. 4, pp. 467-482, 1998.

[8] R. Kerber, "ChiMerge: Discretization of Numeric Attributes," *Proc. 10th Int'l Artificial Intelligence,* pp. 123-128, 1992.

[9] R.P. Li and Z.O. Wang, "An Entropy-Based Discretization Method for Classification Rules with Inconsistency Checking," *Proc. First Conf. Machine Learning and Cybemetics,* pp. 243-246, 2002.

[10] H. Liu and R. Setiono, "Feature Selection via Discretization," *IEEE Trans. Knowledge and Data Eng.,* vol. 9, no. 4, pp. 642-645, July/Aug. 1997.

[11] D.C. Montgomery and G.C. Runger, *Applied Statistics and Probability for Engineers.* John Wiley & Sons, 1999.

[12] H.S. Nguyen, "Discretization of Real Value Attributes: A Boolean Reasoning Approach," PhD thesis, Warsaw Univ. 1997.

[13] H.S. Nguyen and A. Skowron, "Quantization of Real Value Attributes: Rough Set and Boolean Reasoning Approach," *Bull. Int'l Rough Set Soc.,* vol. 1, no. 1, pp. 5-16, 1997.

[14] H.S. Nguyen, "Discretization Problem for Rough Sets Methos," *Proc. First Int'l Conf. Rough Sets and Current Trend in Computing,* pp. 545-552, 1998.

[15] H.S. Nguyen and S.H. Nguyen, "Discretization Methods in Data Mining," *Rough Sets in Knowledge Discovery,* Heidelberg: Physica-Verlag, pp. 451-482, 1998.

[16] L. Shen and E.H. Tay, "A Discretization Method for Rough Sets Theory," *Intelligent Data Analysis,* vol. 5, pp. 431-438, 2001.

[17] E.H. Tay and L. Shen, "A Modified Chi2 Algorithm for Discretization," *IEEE Trans. Knowledge and Data Eng.,* vol. 14, no. 3, pp. 666-670, May/June 2002.

[18] W. Ziarko, "Variable Precision Rough Set Model," *J. Computer and System Science,* vol. 46, pp. 39-59, 1993.

[19] W. Ziarko, "VPRSM Approach to WEB Searching," *Proc. Third Int'l Conf. Rough Sets and Current Trend in Computing,* pp. 514-521, 2002.