A Discretization Algorithm Based on a Heterogeneity Criterion

Xiaoyan Liu, *Student Member*, *IEEE Computer Society*, and Huaiqing Wang

Abstract—Discretization, as a preprocessing step for data mining, is a process of converting the continuous attributes of a data set into discrete ones so that they can be treated as the nominal features by machine learning algorithms. Those various discretization methods, that use entropy-based criteria, form a large class of algorithm. However, as a measure of class homogeneity, entropy cannot always accurately reflect the degree of class homogeneity of an interval. Therefore, in this paper, we propose a new measure of class heterogeneity of intervals from the viewpoint of class probability itself. Based on the definition of heterogeneity, we present a new criterion to evaluate a discretization scheme and analyze its property theoretically. Also, a heuristic method is proposed to find the approximate optimal discretization scheme. Finally, our method is compared, in terms of predictive error rate and tree size, with Ent-MDLC, a representative entropy-based discretization method well-known for its good performance. Our method is shown to produce better results than those of Ent-MDLC, although the improvement is not significant. It can be a good alternative to entropy-based discretization methods.

Index Terms—Data mining, data preparation, discretization, entropy, heterogeneity.

1 Introduction

Data mining is an extremely powerful approach to extracting meaningful information from large and unwieldy databases. However, the successful application of data mining tools relies heavily on the quality of the databases. Therefore, data preparation is a crucial research topic in the data mining field [24], [25]. Discretization, as one of the basic data preparation techniques, has received more and more research attention. The reason is that many existing data mining algorithms focus on learning only in nominal feature space [12], while real-world classification and data mining tasks often involve continuous features. Those continuous features have to be discretized before using such algorithms.

Discretization is a process that transforms continuous attributes into a finite number of intervals, where each interval is associated with a numerical discrete value. Discretized intervals can then be treated in a similar way to nominal values during induction and deduction. There are many advantages of using discrete values over continuous ones. The most important is that data can be reduced and simplified through discretization. In general, results obtained through decision trees or induction rules using discretized data are usually more compact, shorter, and more accurate than results derived using continuous values [15]. Discretization of data thus has the effect of increasing the speed and accuracy of machine learning.

Much research has taken place in the area of discretization. In the early days, simple discretization techniques were used such as equal-width and equal-frequency methods. In these methods, continuous ranges are divided into subranges by user-specified width (range of values) or frequency (number of instances in each interval). Since they do not make use of class information, these are usually called *unsupervised* methods. In contrast, discretization methods that consider class information are called *supervised* methods. Among the various supervised methods, there are two prominent classes of criteria. One is the Chi-square-based criteria that focus on the viewpoint of statistics. The other is the entropy-based criteria that focus on the viewpoint of information.

Statistics-based algorithms include ChiMerge [10], Chi2 [16], Modified Chi2 [22], StatDisc [19], Khiops [2], and so forth. The ChiMerge method [10], [16], [22] searches for the best merge of adjacent intervals by minimizing the Chi-square criterion applied locally to two adjacent intervals that are merged according to statistical similarity. Like ChiMerge, StatDisc [19] considers merging up to N adjacent intervals at a time where N is a user-specified parameter. Khiops [2] evaluates all merges between adjacent intervals and selects the best one according to the Chi-square criterion applied to the whole set of intervals.

There are also many entropy-based discretization methods. Representative algorithms include Maximum entropy [23], D2 [3], and Entropy-MDLC [8], etc. Maximum entropy [23] discretizes the continuous attributes using a minimum loss of information criterion. D2 [3] chooses a threshold T to partition the set of values into two subsets that maximize the information gain after binary partition. Entropy-MDLC [8] uses the class information entropy of candidate partitions to select threshold boundaries for discretization. It finds a single threshold that minimizes the entropy function over all possible thresholds and recursively applies this strategy to both induced partitions. The Minimum Description Length criterion is employed to determine a stopping rule for the recursive discretization strategy. We refer to this as Ent-MDLC in this paper. Other algorithms use class-attributes interdependency information as the criterion [5], [13], [14]. These methods try to maximize the interdependence between the discretized attributes and class labels based on information theory.

• *The authors are with the Department of Information Systems, City University of Hong Kong, Kowloon, Hong Kong.*
E-mail: 50007212@student.cityu.edu.hk and iswang@cityu.edu.hk.

Manuscript received 20 Nov. 2004; revised 27 Mar. 2005; accepted 1 Apr. 2005; published online 19 July 2005.
For information on obtaining reprints of this article, please send e-mail to: tkde@computer.org, and reference IEEECS Log Number TKDESI-0434-1104.

Discretization methods may be categorized according to whether they are *supervised* or *unsupervised*, or they may be grouped in terms of *dynamic* versus *static*, *local* versus *global*, *splitting* versus *merging*, or *direct* versus *incremental* methods [15]. A dynamic method would discretize continuous values when a classifier is being built, such as C4.5 [18]. The static approach, on the other hand, is used prior to the classification task, such as in the methods mentioned above: equal-width, equal-frequency, entropy-based methods [3], [8], [17], [23], Chimerge [10], and so forth. For a comparison between dynamic and static methods, see Dougherty et al. [7].

The difference between local and global discretization methods is in how many attributes are discretized simultaneously. The methods that restrict discretization to a single continuous attribute are termed *local*, whereas methods that simultaneously convert all continuous attributes are considered *global*, such as the Zeta method [9]. Dougherty et al. list some of the global algorithms in [7], and later Chmielewski [6] proposed a method of transforming any local discretization method into a global one.

Splitting versus merging methods are differentiated by their search strategy and can therefore also be termed top-down or bottom-up methods. ChiMerge is one well-known example of a merging method. Another dimension of discretization methods is whether they are direct or incremental. If the number of intervals is defined prior to discretization, such methods are called *direct*, for example, equal-width and equal-frequency methods. Incremental methods begin with a simple discretization and pass through an improvement process, requiring an additional criterion to know when to stop discretizing [4]. A hierarchical framework for discretization methods is proposed in [15]. In this paper, we focus on entropy-based, supervised discretization methods.

In the discretization problem, a tradeoff must be made between information quality and statistical equality that indicates good predictive accuracy and sufficient sample size in every interval. The entropy-based criteria considers the information quality. Entropy [21] is used to measure the class homogeneity of intervals, which is a symmetric function of the class probability. When there are only two classes, it is undoubtedly an excellent indicator of the class homogeneity of intervals. But, when there are more than two classes, entropy is sometimes unable to discriminate the class homogeneity between two intervals. (We will give an example of this in the next section.) Therefore, we propose a new measure of class heterogeneity of intervals from the viewpoint of the class probability itself. Based on the definition of heterogeneity, we present a new criterion to evaluate a discretization scheme and give a heuristic method to find the approximate optimal discretization scheme.

The rest of this paper is organized as follows: Section 2 presents our measurement of the heterogeneity of an interval and analyzes its property theoretically, then gives a criterion to evaluate a discretization scheme. The detailed discretization algorithm is described in Section 3. In Section 4, we compare our method with Ent-MDLC using the benchmark data sets. C4.5 [18] is chosen for evaluation and comparison. Finally, Section 5 concludes the paper with a brief discussion and future extension of this study.

2 Discretization Criterion

The quality of discretization methods involves a tradeoff between simplicity and predictive accuracy. The goal of our proposed method is to reduce the number of intervals while maximizing the accuracy of the information. The basic presentation and definitions used in this paper are introduced first in this section, and then the new discretization criterion is described.

2.1 Basic Presentation and Definitions

Suppose for a supervised classification task with s class labels, the training data set consists of M instances, where each instance belongs to only one of s classes. Let A be any of the continuous attributes from the mixed-mode data. Next, there exists a discretization scheme D on the attribute A, which discretizes the continuous range of the attribute A into n discrete intervals bounded by the pairs of numbers:

$$D : \{[b_0, b_1], (b_1, b_2], \ldots, (b_{n-1}, b_n]\},$$

where b_0 and b_n, respectively, are the minimal and the maximal values of the attribute A, and the values in D are arranged in ascending order. The discretization scheme D is called an n-scheme. These values constitute the boundary set $B_D = \{b_0, b_1, \cdots, b_n\}$ of the discretization scheme D. Since each boundary set corresponds to a discretization scheme, we use either B_D or D to indicate a discretization scheme in this paper. For a given n-scheme discretization, D, we can obtain a 2D discretization quanta matrix as follows.

In Table 1, q_{ji} denotes the total number of observed instances belonging to class c_j whose values of the attribute A fall into the ith interval. The sum of the jth row q_{j+} denotes the total number of observed instances belonging to c_j. The sum of the ith column q_{+i} denotes the total number of observed instances whose values of the attribute A fall into the ith interval.

2.2 Heterogeneity Discretization Criterion

For the ith interval $(b_{i-1}, b_i]$, we can get a conditional class probability $\boldsymbol{p}^{(i)} = (p_1^{(i)}, \ldots, p_j^{(i)}, \ldots, p_s^{(i)})$, where $p_j^{(i)} = q_{ji}/q_{+i}$ and satisfy $\sum_{j=1}^{s} p_j^{(i)} = 1$.

This is the conditional probability distribution of the class in the ith interval. The Shannon entropy of this interval can be expressed as follows:

$$E(\text{interval } i) = -\sum_{j=1}^{s} p_j^{(i)} \log p_j^{(i)}. \tag{1}$$

When each class label occurs at an equal probability in the interval, entropy of this interval gets the maximum value. When all the instances falling into the interval belong to one class, the entropy gets the minimum value *0*. Therefore, entropy can be used as the measurement of class homogeneity. In general, the smaller the entropy value, the worse the homogeneity, and the better the classification ability of an interval. Among the various discretization methods, entropy-based criteria are used in a large class of algorithms. When there are only two classes in a classification task, entropy undoubtedly is an excellent measurement of class homogeneity. However, when there are more than two classes in a classification problem, entropy sometimes cannot accurately reflect the class homogeneity, or the classification ability of an interval. As an example, for a

TABLE 1
A 2D Discretization Quanta Matrix

Class	Intervals					Class total
	$[b_0, b_1]$	...	$(b_{i-1}, b_i]$	...	$(b_{n-1}, b_n]$	
c_1	q_{11}	...	q_{1i}	...	q_{1n}	q_{1+}
⋮	⋮	...	⋮	...	⋮	⋮
c_j	q_{j1}	...	q_{ji}	...	q_{jn}	q_{j+}
⋮	⋮	...	⋮	...	⋮	⋮
c_s	q_{s1}	...	q_{si}	...	q_{sn}	q_{s+}
Interval Total	q_{+1}	...	q_{+i}	...	q_{+n}	M

classification task with three classes c_1, c_2, c_3, there are three intervals I_1, I_2, I_3. The class probability vectors of the three intervals are respectively $\boldsymbol{p}^{(1)} = \left(\frac{1}{2}, \frac{1}{2}, 0\right)$, $\boldsymbol{p}^{(2)} = \left(\frac{1}{6}, \frac{2}{3}, \frac{1}{6}\right)$, and $\boldsymbol{p}^{(3)} = \left(\frac{1}{8}, \frac{3}{4}, \frac{1}{8}\right)$. If the base of the logarithm in (1) is specified as 2, then the entropy value of each interval is easily calculated as follows:

$$\begin{aligned} E(I_1) &= 1 \\ E(I_2) &= 1.25 \\ E(I_3) &= 1.06. \end{aligned} \tag{2}$$

From the viewpoint of entropy, the smaller entropy value is preferred. Therefore, the first interval is better than the last two in terms of classification ability. But, from the viewpoint of predictive accuracy, the last two intervals are better than the first. Based on the analysis above, we propose the following new criterion to measure the degree of class heterogeneity of an interval with better classification ability.

Let Q be the set of probability vector space with s dimensions,

$$\begin{aligned} Q = \{\boldsymbol{p} | \boldsymbol{p} = (p_1, p_2, \cdots, p_s) \in R^s, \sum_{i=1}^{s} p_i \\ = 1, 0 \le p_i \le 1, i = 1, 2, \cdots, s\}. \end{aligned} \tag{3}$$

This set consists of all the possible conditional class probability vectors in an interval. If $\boldsymbol{p}_i = \frac{1}{s}$ for each i, denoted by $\boldsymbol{p}_0 = \left(\frac{1}{s}, \frac{1}{s}, \cdots, \frac{1}{s}\right)$, it is called the *barycenter probability vector*. For the ith interval, if the conditional class probability vector $\boldsymbol{p}^{(i)} = (p_1^{(i)}, \cdots, p_j^{(i)}, \cdots, p_s^{(i)}) = \boldsymbol{p}_0$, then the class heterogeneity of the interval is the lowest, and so is the classification ability of the interval. Naturally, the distance between any conditional class probability vector $\boldsymbol{p}^{(i)}$ and the lowest point $\boldsymbol{p}_0$ can be used to indicate the class heterogeneity of an interval as follows.

Definition 2.1. *For a conditional class probability vector $\boldsymbol{p} = (p_1, \cdots, p_j, \cdots, p_s) \in Q$, its heterogeneity is defined as the distance between the vector $\boldsymbol{p}$ and the barycenter probability vector $\boldsymbol{p}_0$, denoted by $d(\boldsymbol{p})$,*

$$d(\boldsymbol{p}) = |\boldsymbol{p} - \boldsymbol{p}_0| = \sqrt{\sum_{j=1}^{s} \left(p_j - \frac{1}{s}\right)^2}. \tag{4}$$

Heterogeneity is symmetric, that is, the ordering of the probabilities $p_1, \cdots, p_s$ does not influence the value of heterogeneity. From the above definition, it is easy to get the following theorem:

Theorem 2.1. *For any s-dimensional class probability vector $\boldsymbol{p}$, $0 \le d(\boldsymbol{p}) \le \sqrt{\frac{s-1}{s}}$.*

It is apparent that heterogeneity reaches the minimal value when $\boldsymbol{p} = \boldsymbol{p}_0$ and reaches the maximal value when $\boldsymbol{p} = (0, \cdots, 0, 1, 0, \cdots 0)$, that is, all the instances falling into the interval belong to the same class. The larger the heterogeneity value of an interval, the stronger the classification ability of the interval.

Fig. 1 gives the geometrical explanation of heterogeneity when $s = 3$. The vertex of the triangle denotes the event in which only one class label occurs. The barycenter $O = \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)$ denotes that each class label occurs at the equal probability. The farther away the point moves from O, the higher the degree of the class heterogeneity. The heterogeneity value increases along each radial beginning with O. It is easy to see that the contours of heterogeneity are composed of the points on the circles centered at the barycenter O that fall into the interior of the triangle.

We compare the contours of three criteria, entropy, heterogeneity, and predictive accuracy, in Fig. 2b. The

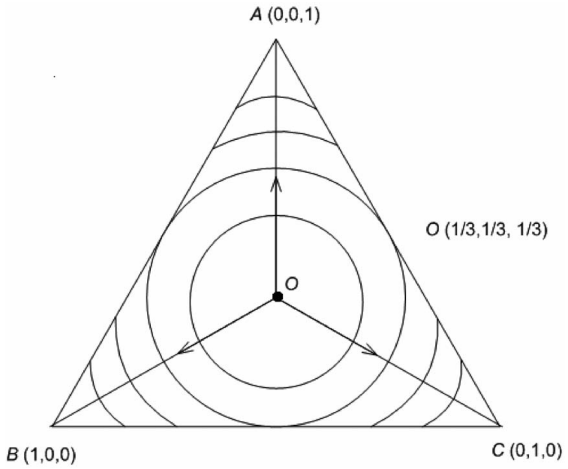

Fig. 1. The contours of heterogeneity.

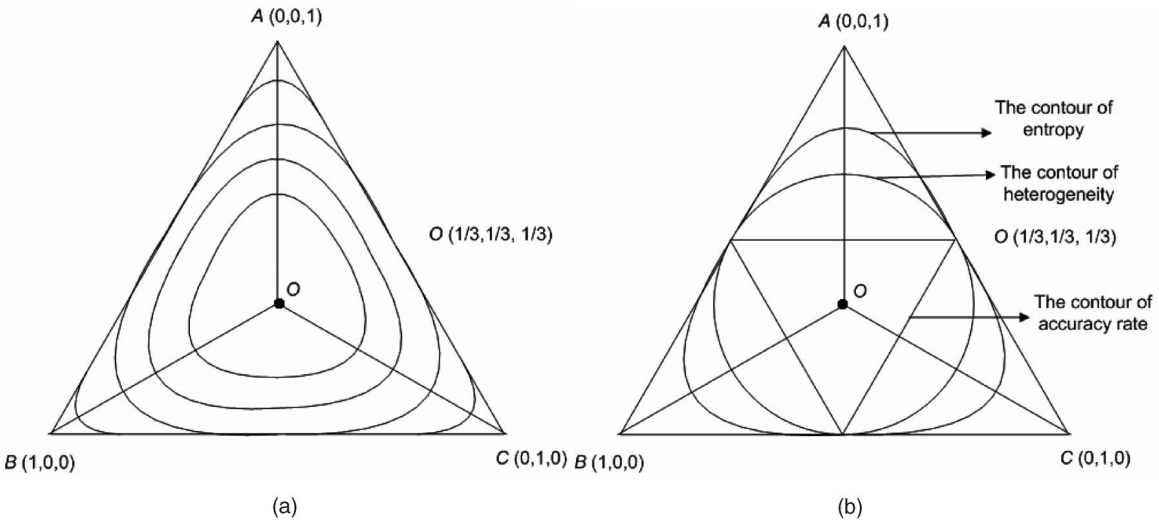

Fig. 2. Comparison of the contours of three criteria. (a) The contours of entropy. (b) The contours of three criteria.

criterion of predictive accuracy uses the maximal component of the conditional class probability as the measurement of interval quality [20]. The maximal component of the class probability indicates the predictive accuracy rate of the interval. The criterion of accuracy rate only considers the dominating class information in an interval, whereas the entropy criterion takes into account all the class information and evaluates the interval according to the whole class distribution. But, it seems to prefer the points that are close to the boundary of the triangle in Fig. 2a. For example, from the viewpoint of entropy, point $\left(\frac{1}{2},\frac{1}{2},0\right)$ is better than point $\left(\frac{1}{8},\frac{3}{4},\frac{1}{8}\right)$. But, from the viewpoint of predictive accuracy, the latter is much better than the former. The heterogeneity criterion also considers all the class information, but lifts the effect of the dominating class as compared to the entropy criterion. It is shown in Fig. 2b that the entropy criterion prefers the points that are close to the boundary of class probability vector space than the heterogeneity criterion. The heterogeneity criterion also has less of a computational workload than the entropy criterion.

After the definition of the heterogeneity of an interval is given, the following is the heterogeneity of a discretization scheme D:

Definition 2.2. *The heterogeneity $\bar{d}$ of an n-scheme D is defined as*

$$\bar{d}_D = \sum_{i=1}^{n} \frac{q_{+i}}{M} d(\boldsymbol{p}^{(i)}). \tag{5}$$

For any two discretization schemes D and D', we denote their corresponding boundary sets by B_D and $B_{D'}$. If $B_{D'} \supseteq B_D$, then we say D' can be generated from D. That is, D' can be obtained by adding some boundary points into the boundary set of D. The following theorem gives the relationship between the heterogeneity values of D and D'.

Theorem 2.2. *For any two discretization schemes D and D', if $B_{D'} \supseteq B_D$, then $\bar{d}_{D'} \geq \bar{d}_D$.*

Proof. We only consider the case that $B_{D'}$ contains one more point than B_D. Suppose D has n intervals $\{[b_0, b_1], (b_1, b_2], \cdots, (b_{n-1}, b_n]\}$, and D' is generated by splitting the ith interval $(b_{i-1}, b_i]$ into two subintervals, which are denoted by interval i_1 and i_2, then

$$\bar{d}_D = \sum_{j=1, j\neq i}^{n} \frac{q_{+j}}{M} d(\boldsymbol{p}^{(j)}) + \frac{q_{+i}}{M} d(\boldsymbol{p}^{(i)}),$$

$$\bar{d}_{D'} = \sum_{j=1, j\neq i}^{n} \frac{q_{+j}}{M} d(\boldsymbol{p}^{(j)}) + \frac{q_{+i_1}}{M} d(\boldsymbol{p}^{(i_1)}) + \frac{q_{+i_2}}{M} d(\boldsymbol{p}^{(i_2)}).$$

Therefore, $\bar{d}_{D'} \geq \bar{d}_D$ is equivalent to the following inequation:

$$\frac{q_{+i_1}}{M} d(\boldsymbol{p}^{(i_1)}) + \frac{q_{+i_2}}{M} d(\boldsymbol{p}^{(i_2)}) \geq \frac{q_{+i}}{M} d(\boldsymbol{p}^{(i)}). \tag{6}$$

Square both sides of the above inequation, then the right side is equal to

$$\begin{aligned} \left(\frac{q_{+i}}{M} d(\boldsymbol{p}^{(i)})\right)^2 &= \left(\frac{q_{+i}}{M} \sqrt{\sum_{j=1}^{s} \left(\frac{q_{ji}}{q_{+i}} - \frac{1}{s}\right)^2}\right)^2 \\ &= \left(\frac{1}{Ms}\right)^2 \cdot \sum_{j=1}^{s} (sq_{ji} - q_{+i})^2 \\ &= \left(\frac{1}{Ms}\right)^2 \cdot \left(s^2 \sum_{j=1}^{s} q_{ji}^2 - sq_{+i}^2\right), \end{aligned} \tag{7}$$

and the left side is

$$\begin{aligned} &\left(\frac{q_{+i_1}}{M} d(\boldsymbol{p}^{(i_1)}) + \frac{q_{+i_2}}{M} d(\boldsymbol{p}^{(i_2)})\right)^2 = \\ &\left(\frac{1}{Ms}\right)^2 \cdot \left(s^2 \sum_{j=1}^{s} q_{ji_1}^2 - sq_{+i_1}^2 + s^2 \sum_{j=1}^{s} q_{ji_2}^2 - sq_{+i_2}^2\right) \\ &+ \left(\frac{1}{Ms}\right)^2 \cdot 2 \sqrt{\sum_{j=1}^{s} (sq_{ji_1} - q_{+i_1})^2} \cdot \sqrt{\sum_{j=1}^{s} (sq_{ji_2} - q_{+i_2})^2}. \end{aligned} \tag{8}$$

Subtract (7) from (8) and make use of $q_{+i} = q_{+i_1} + q_{+i_2}$ and $q_{ji} = q_{ji_1} + q_{ji_2}, j = 1, \cdots, s$, then the following expression is obtained:

$$\sqrt{\sum_{j=1}^{s} (sq_{ji_1} - q_{+i_1})^2} \cdot \sqrt{\sum_{j=1}^{s} (sq_{ji_2} - q_{+i_2})^2} - s^2 \sum_{j=1}^{s} q_{ji_1} q_{ji_2} - sq_{+i_1} q_{+i_2}. \tag{9}$$

According to Cauchy's inequation,

$$\left(\sum_{i=1}^{n} a_i b_i\right)^2 \leq \left(\sum_{i=1}^{n} a_i^2\right) \cdot \left(\sum_{i=1}^{n} b_i^2\right),$$

then

$$\begin{aligned} &\sqrt{\sum_{j=1}^{s} (sq_{ji_1} - q_{+i_1})^2} \cdot \sqrt{\sum_{j=1}^{s} (sq_{ji_2} - q_{+i_2})^2} \\ &\quad \geq \sum_{j=1}^{s} (sq_{ji_1} - q_{+i_1})(sq_{ji_2} - q_{+i_2}) \\ &\quad = s^2 \sum_{j=1}^{s} q_{ji_1} q_{ji_2} - sq_{+i_1} q_{+i_2}. \end{aligned}$$

Thus,

$$\bar{d}_{D'} - \bar{d}_D \geq 0.$$

The equation holds in the above inequation if and only if $q_{ji_1} = kq_{ji_2}$ for all $j = 1, \cdots, s$, where k is a constant. □

Considering that we prefer the smaller number of intervals, we give the evaluation criterion for the quality of a discretization scheme.

Definition 2.3. *Let D, D' be different discretization schemes with n and m intervals, respectively. If*

$$\frac{\frac{\bar{d}_D}{n}}{\frac{\bar{d}_{D'}}{m}} = \frac{\bar{d}_D}{\bar{d}_{D'}} \cdot \frac{m}{n} \geq 1, \tag{10}$$

it is said that the discretization scheme D is better than D' and $\frac{m}{n}$ is called the effect factor of the number of intervals.

The goal of our discretization criterion is to find the best one among all possible discretization schemes. When two discretization schemes have the same number of intervals, then the larger the heterogeneity value, the better the discretization scheme. But, if they have different numbers of intervals, in general, the heterogeneity is higher when the number of interval is larger. When each instance falls into an interval in a discretization scheme, the heterogeneity value is the highest among all the schemes. However, the number of intervals should be reduced through discretization. Therefore, the ratio of the numbers of intervals between two discretization schemes, i.e., $\frac{m}{n}$, is added into Definition 2.3 to control the number of intervals. In general, the effect factor of the number of intervals can be specified as a constant or other function of the number of intervals.

3 Heterogeneity-Based Discretization Algorithms

In our discretization criterion, the effect factor of the number of intervals is the ratio of the numbers of intervals between two discretization schemes. So, our discretization criterion can be transformed into the following criterion function:

$$CF(D) = \frac{\bar{d}_D}{n}, \tag{11}$$

where n is the number of intervals in the discretization scheme D. The optimal discretization scheme can be found by searching the space of all possible discretization schemes to find the one with the highest $CF(D)$ value. Since the space of all possible discretization schemes grows exponentially with the number of possible boundary points, a heuristic method to find an approximate optimal scheme is a natural choice.

Let M, s, A, and m denote the number of instances in a data set, the number of class labels, the attribute needed to be discretized, and the number of distinct values of A, respectively. First, sorting the m distinct attribute values of the attribute A in ascending order, $a_0, a_1, \cdots, a_m$, we denote it by $B = \{a_0, a_1, \cdots, a_m\}$. If the instances that fall into the intervals $(a_{i-1}, a_i]$ and $(a_i, a_{i+1}]$ belong to the same class, remove a_i from the set B until there are instances that fall into two adjacent intervals but do not belong to the same class. Then, we get a boundary set $BS = \{b_0, b_1, \cdots, b_n\}$, where $n \leq m$, $b_i < b_j$ for $i < j$. Arbitrary boundary subset $S \subseteq BS$ corresponds to a discretization scheme D. If scheme D' is obtained by adding a boundary point from BS into the boundary subset with respect to scheme D, then we say D' is generated from D. We denote the scheme set generated from D by GD, and the set composed of the schemes satisfying our criterion is called the candidature set, denoted by CD. The *Globalopt* is the current optimal scheme value. Then, our discretization algorithm has the following structure:

Algorithm 3.1 Heterogeneity-based discretization Algorithm (Heter-Disc)

Initial boundary set $BS = \{b_0, b_1, \cdots, b_n\}$;
Initial discretization scheme $D = \{[b_0, b_n]\}$;
$CD = D$;
$Globalopt = CF(D)$;
While $CD \neq \{\ \}$;
  Generate GD from CD;
  $CD = \{D | D \in GD, CF(D) > Globalopt\}$
  $Globalopt = \underset{D \in CD}{\text{Max}}\, CF(D)$;
End

If the class distribution of the whole data set is homogeneous, then in the first step, all the 2-schemes enter into CD. That is the worst case. After the optimal 2-scheme value is obtained from CD, it can be used as a threshold to generate the new CD in the second step. The cardinality of CD will decrease considerably because of the high threshold. Then, the GD generated from the new CD in the third step also has a small cardinality. That is to say, whatever the initial class distribution of the data

TABLE 2
Data Sets Description

Data set	Instances	Total attributes	Continuous attributes	Classes
Anneal	798	38	6	5
Glass	214	9	9	6
Hepatitis	155	19	6	2
Iris	150	4	4	3
Page Block	5473	10	10	5
Pima	768	8	8	2
Wine	178	13	13	3

TABLE 3
C4.5 Predictive Error Rates and Standard Deviations

Data set	Continuous	Ent-MDLC	Heter-Disc
Anneal	8.35±1.60	10.35±1.00	10.28±0.95
Glass	1.86±2.28	2.31±3.08	2.30±3.35
Hepatitis	21.94±2.77	24.52±1.94	23.73±2.42
Iris	4.34±2.84	4.25±4.58	4.00±3.85
Page Block	3.16±1.13	4.86±2.51	4.83±2.12
Pima	26.22±2.65	25.21±4.23	24.4±3.13
Wine	6.22±6.84	7.95±7.27	7.96±6.84
Average	10.29	11.35	11.07

set is, the number of the schemes we search in our algorithm decreases considerably. In the whole process, it is assumed that the global best k-scheme can be generated from $(k-1)$-schemes in the set of CD with higher CF values. It is clear that the proposed algorithm is reasonably heuristic. Thus, our searching space can be reduced remarkably compared to the whole scheme space.

4 Experiment Results

In this section, we compare a number of discretization methods as a preprocessing step to C4.5 [18]. The reason for our choice is that C4.5 is a state-of-the-art decision-tree learner algorithm, and decision-tree learners are the most commonly used type of machine learning algorithm.

We examine the effect of discretization on C4.5 through comparisons before and after discretization. The classification quality is measured using predictive error rate and tree size, i.e., the number of nodes. Since the evaluation function in our algorithm is proposed based on the heterogeneity compared to the entropy criterion, our method would be compared with Ent-MDLC, which is a well-known representative of entropy-based discretization methods. Its good performance has often been shown in previous research [7], [12], [15], [22]. It is recommended as the first choice when other things are equal, such as user need, class information, and other considerations [15]. Here, our method is denoted by Heter-Disc. The data sets are taken from the University of California at Irvine repository of machine learning data sets [1]. Table 2 gives a summary of data sets used in our experiments.

To compare the efficacy of these methods, the predictive error rate and the tree size of C4.5 on the undiscretized data sets is presented, denoted by "continuous" in Tables 3 and 4. To get more reliable results, the 10-fold cross-validation test method was applied to all data sets. Each data set was divided into 10 parts of which nine parts were used as the training set and the remaining one part as the test set. The experiments were repeated 10 times. The final predictive error rate was taken as the average of the 10 predictive error rate values.

All seven data sets were discretized using the original discretization algorithms, and the discretized data sets were classified using C4.5. The predictive error rate and its standard deviation of those methods are presented in Table 3. The tree size using C4.5 with different discretization methods is presented in Table 4. Averages in the bottom rows of Tables 3 and 4 give an indication of how the discretization methods affect predictive error rate and tree size.

From Table 3, it can be seen that, in terms of predictive error rate, the results obtained by Heter-Disc are more accurate than those of Ent-MDLC on average. In six out of

TABLE 4
Number of Nodes in C4.5 Before and After Discretization

Data set	Continuous	Ent-MDLC	Heter-Disc
Anneal	75	68	66
Glass	11	11	10
Hepatitis	23	21	19
Iris	9	7	8
Page Block	90	82	80
Pima	43	27	30
Wine	9	15	16
Average	37.14	33.00	32.71

seven data sets, our method obtained better results than Ent-MDLC. For data set Wine, the result of Heter-Disc is slightly less accurate than that of Ent-MDLC. However, the results of our method are more reliable than those of Ent-MDLC according to standard deviations. It can be seen that our method and Ent-MDLC both reduce the tree size compared to C4.5 except for data set Wine in Table 4. Our method only outperformed Ent-MDLC in four out of seven data sets with regard to tree size. But, in the average of the tree sizes, our method is still a little better than Ent-MDLC.

5 Conclusions

In this paper, a new method, Heter-Disc, for discretization of continuous values has been introduced. Our method has been compared, in terms of predictive error rate and tree size, with Ent-MDLC, the entropy-based method that is known for its good performance [7], [12], [15], [22].

In order to evaluate the discretization scheme, an evaluation function has been proposed in our method, which is based on the measure of heterogeneity. The heterogeneity measurement has the stronger ability of discriminating between two intervals than the entropy criterion, where the class probability vectors of an interval are close to the boundary of the class probability vector space (see Section 2.2). It should be noted that such points, closing to the boundary of the class probability vector space, occupy a very small percentage in the whole space. And, for other points, our approach gets results similar to the entropy criterion. However, overall, the results of our method are more accurate than the results of Ent-MDLC. Therefore, the proposed method in this paper offers a good alternative to entropy-based discretization methods.

We are aware that there is much research yet to be done regarding our method. To begin with, from Fig. 1, it can be seen that the degree of heterogeneity increases along the radial beginning with the barycenter. But, the increasing speed of each radial is different. Therefore, we plan to consider the effect of the increasing speed in a future study. In addition, because of the high combinatorial complexity of the discretization problem, we use a hill-climbing heuristic method to search the optimal discretization scheme. Though the final results are satisfactory, this kind of search strategy cannot guarantee the global optimal solution. Sometimes they fall into a local optimal discretization scheme. Therefore, we will consider a random search strategy in the future to avoid such a problem. Moreover, our method deals with only one continuous attribute at a time, so the globalization of our discretization method is also of future interest. Furthermore, in the future, we will apply the method in this paper to real-world, real-time, financial data analysis. We believe that such application will be the best test bed for our method.

Acknowledgments

The authors thank Ms. Qiudan Li of the Department of Information Systems, City University of Hong Kong, for her support on this research. The paper is supported by a UGC Research Grant (No. CityU 1234/03E) from the Hong Kong Government.

References

[1] C.L. Blake and C.J. Merz, UCI Repository of Machine Learning Databases, http://www.ics.uci.edu/~mlearn/MLRepository.html, UC Irvine, Dept. of Information and Computer Science, 1998.

[2] M. Boulle, "Khiops: A Statistical Discretization Method of Continuous Attributes," *Machine Learning*, vol. 55, pp. 53-59, 2004.

[3] J. Catlett, "On Changing Continuous Attributes into Ordered Discrete Attributes," *Machine Learning-EWSL-91, Proc. European Working Session on Learning*, pp. 164-178, Mar. 1991.

[4] J. Cerquides and R.L. Mantaras, "Proposal and Empirical Comparison of a Parallelizable Distance-Based Discretization Method," *Proc. Third Int'l Conf. Knowledge Discovery and Data Mining (KDD-97)*, pp. 139-142, 1997.

[5] J.Y. Ching, A.K.C. Wong, and K.C.C. Chan, "Class-Dependent Discretization for Inductive Learning from Continuous and Mixed Mode Data," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 17, no. 7, pp. 641-651, 1995.

[6] M.R. Chmielewski and J.W. Grzymala-Busse, "Global Discretization of Continuous Attributes as Preprocessing for Machine Learning," *Int'l J. Approximate Reasoning*, vol. 5, pp. 319-331, 1996.

[7] J. Dougherty, R. Kohavi, and M. Sahami, "Supervised and Unsupervised Discretization of Continuous Features," *Proc. 12th Int'l Conf. Machine Learning,* pp. 194-202, 1995.
[8] U.M. Fayyad and K.B. Irani, "Multi-Interval Discretization of Continuous-Valued Attributes for Classification Learning," *Proc. 13th Int'l Joint Conf. Artificial Intelligence,* pp. 1022-1027, 1993
[9] K.M. Ho and P.D. Scott, "Zeta: A Global Method for Discretization of Continuous Variables," *Proc. Third Int'l Conf. Knowledge Discovery and Data Mining (KDD97),* pp. 191-194, 1997.
[10] R. Kerber, "ChiMerge: Discretization of Numeric Attributes," *Proc. 10th Int'l Conf. Artificial Intelligence (AAAI-91),* pp. 123-128, 1992.
[11] R. Kohavi, "Bottom-Up Induction of Oblivious Read-Once Decision Graphs: Strengths and Limitation," *Proc. 12th Nat'l Conf. Artificial Intelligence,* pp. 613-618, 1994.
[12] R. Kohavi and M. Sahami, "Error-Based and Entropy-Based Discretization of Continuous Features," *Proc. Second Int'l Conf. Knowledge Discovery and Data Mining (KDD-96),* pp. 114-119, 1996.
[13] L.A. Kurgan and K.J. Cios, "Discretization Algorithm that Uses Class-Attribute Interdependence Maximization," *Proc. Int'l Conf. Artificial Intelligence (IC-AI-2001),* pp. 980-987, 2001.
[14] L.A. Kurgan and K.J. Cios, "CAIM Discretization Algorithm," *IEEE Trans. Knowledge and Data Eng.,* vol. 16, no. 2, pp. 145-153, Feb. 2004.
[15] H. Liu, F. Hussain, C.L. Tan, and M. Dash, "Discretization: An Enabling Technique," *Data Mining and Knowledge Discovery,* vol. 6, pp. 393-423, 2002.
[16] H. Liu and R. Setiono, "Feature Selection via Discretization," *IEEE Trans. Knowledge and Data Eng.,* vol. 9, no. 4, pp. 642-645, 1997.
[17] B. Pfahringer, "Compression-Based Discretization of Continuous Attributes," *Proc. 12th Int'l Conf. Machine Learning,* 1995.
[18] J.R. Quinlan, *C4.5: Programs for Machine Learning.* San Mateo, Calif.: Morgan Kaufmann, 1993.
[19] M. Richeldi and M. Rossotto, "Class-Driven Statistical Discretization of Continuous Attributes (extended abstract)," *Machine Learning: ECLM-95 (Proc. European Conf. Machine Learning, 1995),* N. Lavrac and S. Wrobel, eds., pp. 335-338, 1995.
[20] R.J. Roiger and M.W. Geatz, *Data Mining: A Tutorial Based Primer.* Addison Wesley, 2002.
[21] C. Shannon and W. Weaver, *The Mathematical Theory of Information.* Urbana: Univ. Illinois Press, 1949.
[22] F.E.H. Tay and L.X. Shen, "A Modified Chi2 Algorithm for Discretization," *IEEE Trans. Knowledge and Data Eng.,* vol. 14, no. 3, pp. 666-670, 2002.
[23] A.K.C. Wong and D.K.Y. Chiu, "Synthesizing Statistical Knowledge from Incomplete Mixed-Mode Data," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 9, pp. 796-805, 1987.
[24] S.C. Zhang, Q. Yang, and C.Q. Zhang, "Data Preparation for Data Mining," *Applied Artificial Intelligence,* vol. 17, nos. 5-6, pp. 375-382, 2003.
[25] S.C. Zhang, C.Q. Zhang, and Q. Yang, "Information Enhancement for Data Mining," *IEEE Intelligent Systems,* pp. 12-13, 2004.

Xiaoyan Liu received the BS and MS degrees in mathematics from Tongji University in 2000 and Tsinghua University in 2003, in China, respectively. She is currently a PhD candidate in the Department of Information Systems at the City University of Hong Kong. Her research interests include data mining, financial engineering, business intelligent systems and their applications. She is a student member of the IEEE Computer Society.

Huaiqing Wang received the PhD degree in computer science from the University of Manchester in 1987. He is a professor in the information systems department at the City University of Hong Kong. He specializes in research and development of business intelligence systems, intelligent agents, and their applications (such as multiagent supported financial information systems, virtual learning systems, knowledge management systems, and conceptual modeling).