Hindawi Publishing Corporation
Journal of Applied Mathematics
Volume 2013, Article ID 350123, 8 pages
http://dx.doi.org/10.1155/2013/350123

Research Article

An Algorithm for Discretization of Real Value Attributes Based on Interval Similarity

Li Zou,[1,2] Deqin Yan,[1] Hamid Reza Karimi,[3] and Peng Shi[4,5]

[1] *School of Computer and Information Technology, Liaoning Normal University, Dalian 116029, China*
[2] *State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210093, China*
[3] *Department of Engineering, Faculty of Engineering and Science, University of Agder, 4898 Grimstad, Norway*
[4] *College of Engineering and Science, Victoria University, Melbourne, VIC 8001, Australia*
[5] *School of Electrical and Electronic Engineering, The University of Adelaide, Adelaide, SA 5005, Australia*

Correspondence should be addressed to Hamid Reza Karimi; hamid.r.karimi@uia.no

Received 3 November 2012; Accepted 25 February 2013

Academic Editor: Xiaojing Yang

Discretization algorithm for real value attributes is of very important uses in many areas such as intelligence and machine learning. The algorithms related to Chi2 algorithm (includes modified Chi2 algorithm and extended Chi2 algorithm) are famous discretization algorithm exploiting the technique of probability and statistics. In this paper the algorithms are analyzed, and their drawback is pointed. Based on the analysis a new modified algorithm based on interval similarity is proposed. The new algorithm defines an interval similarity function which is regarded as a new merging standard in the process of discretization. At the same time, two important parameters (condition parameter α and tiny move parameter c) in the process of discretization and discrepancy extent of a number of adjacent two intervals are given in the form of function. The related theory analysis and the experiment results show that the presented algorithm is effective.

1. Introduction

The intelligent information processing is researching hot spot in today's information science theory and application. In machine learning and data mining, many algorithms have already been developed according to processing discrete data. Discretization of real value attributes is an important method of compression data and simplification analysis and also is an indeterminable in pattern recognition, machine learning, and rough set analysis domain. The key of discretization lies with dividing the cut point. At present, there are five different axes by which the proposed discretization algorithms can be classified [1–4]: supervised versus unsupervised, static versus dynamic, global versus local, top-down (splitting) versus bottom-up (merging), and direct versus incremental. Continuous attributes need to be discretized in many algorithms such as rule extraction and tag sort, especially rough set theory in research of data mining. In view of an algorithm for discretization of real value attributes based on rough set, people have conducted extensive research and proposed a lot of new discretization method [5], one kind of thought of which is that the decision table compatibility is not changed during discretion. Rough set and Boolean logical method proposed by Nguyen and Skowron are quite influential [6]. Moreover, there are two quite influential discretization methods which are the algorithms of the correlation based on information entropy and the algorithms of the correlation of Chi2 algorithm based on statistical method for supervised discretization. Reference [7] is an algorithm for discretization of real value attributes based on decision table and information entropy, which belongs to a heuristic and local algorithm that seeks the best results. Reference [8] proposed a discretization algorithm for real value attributes based on information theory, which regards class-attribute interdependence as an important discretization criterion and selects the candidate cut point which can lead to the better correlation between the class labels and the discrete intervals. But this algorithm has the following

disadvantages. It uses a user-specified number of intervals when initializing the discretization intervals. The significance test used in the algorithm requires training for selection of a confidence interval. It initializes the discretization intervals using a maximum entropy discretization method. Such initialization may be the worst starting point in terms of the CAIR criterion. And it is very easy to cause the lower degree of discretization which is not immoderate. Huang has solved the above problem, but at the expense of very high-computational cost [9]. Kurgan and Cios have improved in the discretization criterion and attempted to cause class-attribute interdependence maximization [10]. But this criterion merely considered dependence between the most classes in the interval and the attribute, which will cause the excessive discretization and the result is not to be precise. References [3, 4, 11, 12] are the algorithms of the correlation of Chi2 algorithm based on the statistics. The ChiMerge algorithm introduced by Kerber in 1992 is a supervised global discretization method [11]. The method uses χ^2 test to determine whether the current point is merged or not. Bondu et al. [13] proposed a Chi2 algorithm in 1997 based on the ChiMerge algorithm. In this algorithm, the authors increase the value of the χ^2_α threshold dynamically and decide the intervals' merging order according to the value of D, where $D = \chi^2_\alpha - \chi^2$ and χ^2_α is a fractile decided by the significance level α. Tay and Shen further improved the Chi2 algorithm and proposed the modified Chi2 algorithm in [4]. The authors showed that it is unreasonable to decide the degree of freedom by the number of decision classes on the whole system in the Chi2 algorithm. Conversely, the degree of freedom should be determined by the number of decision classes of each two adjacent intervals. In [3], the authors pointed out that the method of calculating the freedom degrees in the modified Chi2 algorithm is not accurate and proposed the extended Chi2 algorithm, which replaced D with $D/\sqrt{2v}$.

Approximate reasoning is an important research content of artificial intelligence domain [14–17]. It needs measuring similarity between the different pattern and the object. Similarity measure is a function that is used in comparing similarity among information, data, shape, and picture etc. [18]. In some domain such as picture matching, information retrieval, computer vision, image fusion, remote sensing, and weather forecast, similarity measure has the extremely vital significance [13, 19–22]. The traditional similarity measure method often directly adopts the research results in statistics, such as the cosine distance, the overlap distance, the Euclid distance, and Manhattan distance.

Using χ^2 statistic and significance level codetermines whether that cut point can be merged is the main role of algorithms related to Chi2 algorithm. In this paper, we point out that using the importance of nodes determined by the distance, divided by $\sqrt{2v}$, for extended Chi2 algorithm of reference [3] lacks theory basis and is not accurate. It is unreasonable to merge first adjacent two intervals which have the maximal difference value. At the same time, based on the study of applied meaning of χ^2 statistic, the drawback of the algorithm is analyzed. To solve these problems, a new modified algorithm based on interval similarity is proposed. The new algorithm defines an interval similarity function which is regarded as a new merging standard in the process of discretization. At the same time, two important parameters (condition parameter α and tiny move parameter c) in the process of discretization and discrepancy extent of a number of adjacent two intervals are given in the form of function. Besides, two important stipulations are given in the algorithm. The related theory analysis and the experiment results show that the presented algorithm is effective.

2. Correlative Conception of Chi2 Algorithm

At first, a few of conceptions about discretization are introduced as follows.

(1) Interval and cut point. A single value of continuous attributes is a cut point; two cut points produce an interval. Adjacent two intervals have a cut point. Discretization algorithm of real value attributes actually is in the process of removing cut point and merging adjacent intervals based on definite rules.

(2) χ^2 and χ^2_α. χ^2 is a statistic in probability.

The formula for computing the χ^2 value is

$$\chi^2 = \sum_{i=1}^{2} \sum_{j=1}^{k} \frac{\left(A_{ij} - E_{ij}\right)^2}{E_{ij}}, \tag{1}$$

where

k: number of system classes;

A_{ij}: number of patterns in the ith interval, jth class;

$C_j = \sum_{i=1}^{2} A_{ij}$: number of patterns in jth class;

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

$N = \sum_{i=1}^{2} R_i$: total number of patterns;

$E_{ij} = R_i \times C_j / N$: expected frequency of A_{ij}.

χ^2_α is threshold determined significance level α. In statistics, the asymptotic distribution of χ^2 statistic with k degrees of freedom is χ^2 distribution with $k - 1$ degrees of freedom, namely, $\chi^2_{(k-1)}$ distribution. χ^2_α is determined by selecting a desired significance level α.

(3) Inconsistency rate. When condition attribute values of objects are the same and decision attribute value is not the same, the classified information of the decision table has definite inconsistency rate (error rate), where

$$\text{Inconsistency rate: Incon\_rate} = 1 - \gamma_C \tag{2}$$

γ_C is approximate precision. Rectified Chi2 algorithm proposed in this paper controls merger extent and information loss in the discretization process with Incon_rate.

Extended Chi2 algorithm is as shown in Algorithm 1 [1].

```
Step1: Initialization. Set significance level α = 0.5. Calculate inconsistency rate of information
       systems: Incon_rate.
Step2: Sort data in ascending order for each attribute and calculate χ² value of
       each adjacent two intervals according to (1), then using a table to obtain
       the corresponding χ² threshold. Calculate difference D' = (χ²_α − χ²) / √(2v).
Step3: Merge.
       While (mergeable cut point)
       {Search cut point that has the maximal difference D', then merging it;
          If Incon_rate change
             {Withdraw merging;
                goto Step4;}
          else goto Step2;
       }
Step4: If α can not be decreased
                Exit procedure;
       Else {α_0 = α;
             Decreasing the significance level by one level;
             goto Step2; }
Step5: Do until no attribute can be merged
       {For each mergeable attribute i
          {Calculate difference D';
             α = α_0;
             sign flag=0;
             While (flag= =0)
             {While (mergeable cut point)
               {Search cut point that has the maximal difference D', then merging it;
                  If Incon_rate change
                  {Withdraw merging;
                     flag=1;
                     break;}
                  Else update difference D';
                     }
                  If α can not be decreased
                        Break;
                  Else {Decreasing the significance level by one level;
                     Update difference D';}
             }
          }
       }
```

Algorithm 1

3. Interval Similarity Function

3.1. Insufficiency of Chi2 Correlation Algorithm. (1) In formula (1), C_j/N is the proportion of a number of patterns in jth class accounting for a total number of patterns, and $E_{ij} = R_i \times C_j/N$ is a number of patterns in the ith interval. Therefore, statistical χ^2 indicates the equality degree of the jth class distribution of adjacent two intervals. The smaller the χ^2 value is, the more the similar is class distribution, and the more unimportant the cut point is. It should be merged.

For the newest extended Chi2 algorithm, it is very possible to have such two groups of adjacent intervals: the number of classes of one is more than another, then, the difference of class distribution of adjacent two intervals which have the greater number of classes is bigger and the corresponding χ^2 value is greater. Yet, the difference of class distribution of adjacent two intervals which have the less number of classes is smaller and the corresponding χ^2 value is smaller. Moreover, degree of freedom of adjacent two intervals with the greater number of classes is bigger. Then, quantile $\chi^2_\alpha(v_1)$ is possibly much more than $\chi^2_\alpha(v_2)$ (see Figure 1). Therefore, even if $\chi^2_1 > \chi^2_2$, $v_1 > v_2$, we still have such situation: $D_1/\sqrt{2v_1} > D_2/\sqrt{2v_2}$. But in fact, adjacent two intervals with the bigger difference of class distribution and the greater number of classes should not be first merged. This merging standard in the computation is not precise. So it is unreasonable to merge first the adjacent two intervals with the maximal difference.

(2) In algorithms of the series of Chi2 algorithm, expansion to χ^2 is as follows:

$$\chi^2 = \sum_{i=1}^{2}\sum_{j=1}^{k}\frac{\left(A_{ij}-E_{ij}\right)^2}{E_{ij}} = \sum_{i=1}^{2}\sum_{j=1}^{k}\frac{N\cdot\left(A_{ij}-R_i\cdot C_j/N\right)^2}{R_i\cdot C_j}. \tag{3}$$

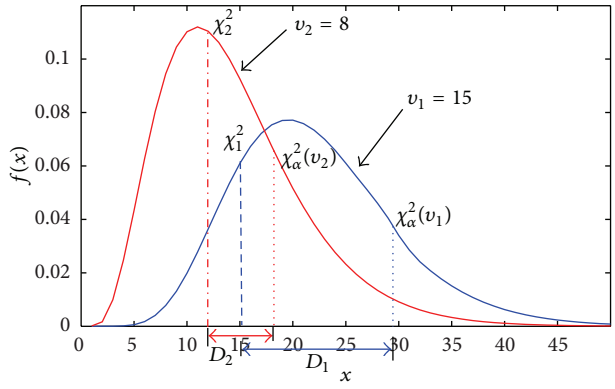

Figure 1: Comparison of χ^2 distribution with different degrees of freedom.

In formula (3), under certain situations χ^2 is not very accurate: there are adjacent two intervals of class distribution adjoin. When the number of some class C_j increases (two intervals both have this class, N and R_i are invariable, A_{ij} value of one of two intervals is invariable); the numerator and the denominator of expansion to χ^2 formula are increasing at the same time. Regarding $(A_{ij} - R_i \cdot C_j/N)^2$, its value may be increased first and then turn to be decreased. In other words, when $R_i \cdot C_j/N$ is quite bigger than A_{ij}, χ^2 value will increase (degree of freedom not to change) and probability of interval merging will be reduced. In fact, when the number of some class C_j increases, this class has stronger independence with intervals, and it has leader's class status. Therefore, compared with not increased, this time should have the same opportunity of competition and even should merge first these two intervals.

(3) The situation when χ^2 value is 0 is as follows.

There exists the case that class distribution of adjacent two intervals is completely uniform, namely, $A_{ij} = E_{ij} \Rightarrow \chi_{ij}^2 = 0$. Thus, D' is very big relatively and the two intervals are possibly first merged. But in fact, it is possibly unreasonable that they are first merged. For example (see Table 1), a, b, and c are condition attributes and d is decision attribute. Observing attribute a: the same value is in the identical interval. The number of samples of two intervals is the same. Classification in A is completely uniform, Namely, $A_{ij} = E_{ij} \Rightarrow \chi_{ij}^2 = 0$; D' is quite big relatively. Even if degree of freedom in A is bigger than B, but because the difference of degree of freedom between A and B is very small, it is possible that the difference D' of A is bigger than the difference D' of B. From the computation with Table 1, we get $\chi^2 = 0$ and $v = 2$ in A, then $\alpha = 0.9$. We can see $\chi_\alpha^2 = 4.61$ and get $D' \approx 2.3$. Regarding B in Table 1, $\chi^2 = 0.45$ and $v = 1$, then $\alpha = 0.9$. We can see $\chi_\alpha^2 = 2.71$ and get $D' \approx 1.6$. Thus, two intervals of attribute a in A will be first merged, and then the sample 3, 4 and the sample 1, 5 in A could have the conflict, but it is not the case in B. So, when χ^2 value is equal to 0, using difference D' as the standard of interval merging is inaccurate.

Table 1: Decision table.

		A					B		
U	a	b	c	d	U	a	b	c	d
1	0	0	2	1	1	0	0	2	1
2	0	—	—	2	2	0	—	—	1
3	0	1	1	3	3	0	1	1	1
4	1	1	1	1	4	1	1	1	1
5	1	0	2	2	5	1	0	2	1
6	1	—	—	3	6	1	—	—	2

3.2. Interval Similarity Function

Definition 1. Let B be a database, or an information table, and let $t_i, t_j \in B$ be two arrays then their similar degree $\mathrm{SIM}(t_i, t_j)$ is defined as a mapping to the interval $[0, 1]$.

A good similarity measure should have the following characteristic:

for all $t_i \in B$, $\mathrm{SIM}(t_i, t_i) = 1$;

for all $t_i, t_j \in B$, if t_i and t_j are completely different, then $\mathrm{SIM}(t_i, t_j) = 0$;

for all $t_i, t_j, t_k \in B$, compared with t_j, if t_k closes to t_i, then $\mathrm{SIM}(t_i, t_j) < \mathrm{SIM}(t_i, t_k)$.

The traditional similarity measure method often directly adopts the research results in statistics, such as the cosine distance, the overlap distance, the Euclid distance, and Manhattan distance. Based on the analysis to the drawback of the correlation of Chi2 algorithm, we propose the similarity function as follows.

Definition 2. Given two intervals (objects), let a_i be a class label according to the ith value in the first interval, and let b_j be a class label according to the jth value in the second interval. Then, the difference between a_i and b_j is

$$d_{ij} = \begin{cases} 0 & \text{if } a_i = b_j, \\ 1 & \text{if } a_i \neq b_j, \end{cases} \tag{4}$$

where $i = 1, 2, \ldots, s$, $j = 1, 2, \ldots, t$.

Definition 3. Similarity function of adjacent two intervals t_k, t_{k+1} is defined as

$$\mathrm{SIM}(t_k, t_{k+1}) = \begin{cases} 1 & \text{if } C' = 1, \\ 1 - \dfrac{2}{\pi} \arctan\left(\dfrac{\sum_{i=1}^{s} \sum_{j=1}^{t} d_{ij}}{(s+t)^\alpha} \right) & \text{if } C' > 1. \end{cases} \tag{5}$$

In the formula (5), α is a condition parameter:

$$\alpha = \begin{cases} 1 & \text{if } \left(|s - t| < \left(\sqrt{\dfrac{\sum_{i=1}^{A} V_i}{A}} \pm c \right) \right), \\ \dfrac{1}{2} & \text{other}, \end{cases} \tag{6}$$

where C' is the number of classes of two adjacent intervals, A is the number of condition attribute, "| |" denotes absolute value, V_i is the number of cut points of attribute i before discretizing, $1 \le i \le A$, c is tiny move parameter ($c = 1, 2, 3$), $1 \le k \le V_i$, and $\alpha \in \{1, 1/2\}$.

Considering any adjacent two intervals t_k and t_{k+1}, $\sum_{i=1}^{s}\sum_{j=1}^{t} d_{ij}$ can express the difference degree between adjacent two intervals. But, because the number of each group of adjacent intervals is different, it is unreasonable to merely take $\sum_{i=1}^{s}\sum_{j=1}^{t} d_{ij}$ as a difference measure standard. In order to obtain a uniform standard of difference measure and a fair compete opportunity among each group of adjacent intervals, it is reasonable to take $\sum_{i=1}^{s}\sum_{j=1}^{t} d_{ij}/(s+t)^{\alpha}$ as a difference measure standard. In formula (5), when the number of adjacent two intervals has only one ($C' = 1$), similar degree between them is the biggest obviously. In order to enable similar degree among various intervals to compare in the uniform situation, we can take arc tangent function to normalized processing, making similar value mapped in $[0, 1]$. The formula $\sqrt{\sum_{i=1}^{A} V_i/A}$ expresses the average normative value of cut points before discretizing. And we take it as benchmark of distance of the number between two intervals, carrying on tiny move in the c scope.

Reason of parameter α selected: when the distance of the number between adjacent two intervals reaches a certain extent, $\sum_{i=1}^{s}\sum_{j=1}^{t} d_{ij}$ will be small relatively, but $s + t$ will quite possibly be big. Thus, $\sum_{i=1}^{s}\sum_{j=1}^{t} d_{ij}/(s+t)$ will be relatively very small and not be easily merged. In fact, considering the relations of containing and being contained between two adjacent intervals, they still have the greater merged opportunity and it is unfair. Therefore, parameter α as condition parameter can play a fair role: when the distance of the number between adjacent two intervals reaches the certain extent, we select $\sum_{i=1}^{s}\sum_{j=1}^{t} d_{ij}/\sqrt{s+t}$ as standard. In addition, by considering that the size of intervals has relation with the number of initial attribute value (the number cut point), we can take $\sqrt{\sum_{i=1}^{A} V_i/A} \pm c$ as the distance of the number of adjacent two intervals, carrying on the tiny move in the c scope on benchmark of $\sqrt{\sum_{i=1}^{A} V_i/A}$. In brief, interval similarity definition not only can inherit the logical aspects of χ^2 statistic but also can resolve the problems about algorithms of the correlation of Chi2 algorithm, realizing equality.

4. Discretization Algorithm for Real Value Attributes Based on Interval Similarity

In this section we propose a new discretization algorithm for real value attributes based on interval similarity (the algorithm is called SIM for short). The new algorithm defines an interval similarity function which is regarded as a new merging standard in the process of discretization. In the algorithm we adopt two operations.

(1) With formula (5) there are many maximal similar values calculated among groups of adjacent intervals; we will merge the adjacent two intervals with the smallest number of classes.

(2) When there are many maximal similar values calculated and the number of classes among groups of adjacent intervals is the same, we will merge the adjacent two intervals with the smallest number of samples of adjacent intervals (namely, $s + t$ is the smallest).

The two operations can reduce the influence of merge degree to other intervals or attributes, and the inconsistency rate of system cannot increase beforehand. The algorithm SIM is as shown in Algorithm 2.

Table 2: Data Information.

Datasets	Continuous attributes	Discrete attributes	Number of Classes	Examples
Iris	4	0	3	150
Glass	9	0	7	214
Breast	9	0	2	683
Wine	13	0	3	178
Auto	5	2	3	392
Bupa	6	0	2	345
Machine	7	0	8	209
Pima	8	0	2	768
Ionosphere	34	0	2	351

5. The Experimental Results and Analysis

We adopt the datasets of UCI machine learning database (see Table 2). The UCI machine learning datasets are commonly used in data mining experiment.

Nine datasets were discrete respectively by the algorithm proposed in this paper (SIM) and the EXT algorithm, the Boolean algorithm. We ran C4.5 on the discreted data. Choosing randomly, 80 percent of examples are training sets; the rest are testing sets. The average predictive accuracy, the average numbers of nodes of decision tree, and the average numbers of rules extracted are computed and compared by different algorithms (see Table 3). Meanwhile, discreted data is classified by multiclass classification method [23–26] of SVM. 80 percent of examples are randomly chosen as training sets; the rest are testing sets. Model type is C-SVC. Kernel function type is RBF function. Search range of penalty C is $[1, 100]$. Kernel function parameter γ is 0.5. The predictive accuracy (acc) and the number of support vector (svs) are computed and compared for the above three algorithms (see Table 4).

From Table 3, we can see that compared with extended Chi2 algorithm and Boolean discretization algorithm, the average predictive accuracy of decision tree of SIM algorithm for discretization of real value attributes based on interval similarity has been rising except Bupa and Pima datasets for 9 datasets. In particular promotion scope of Glass, Wine, and Machine datasets is very big. The average numbers of nodes of decision tree and the average numbers of rules extracted

```
Step1:  Compute inconsistency rate of information system;
Step2:  Sort data in ascending order for each attribute and calculate the similar
        value SIM of each adjacent intervals according to (5) and (6);
Step3:  Merge
While (merge-able cut point)
{
  Search cut point that has the maximal similar value, then merging it;
  If (many maximum values)
   {
     Merge adjacent two intervals with the smallest number of classes;
     If (Incon_rate increases)
     {
        Withdraw merging;
        Exit procedure;
      }
     Else {break; goto Step2;}
   }
  If (some several maximum values and the same class number of classes
      among groups of adjacent intervals)
    {
        Merge the adjacent two intervals with the smallest number of samples
        of adjacent intervals;
        If (Incon_rate increases)
       {
        withdraw merging;
        exit procedure;
        }
      Else {break; goto step2;}
      }
}
```

Algorithm 2

Table 3: Comparison of C4.5 performance on 9 UCI real datasets.

Datasets	Predictive accuracy (%)			Number of nodes			Number of rules		
	EXT	SIM	Boolean	EXT	SIM	Boolean	EXT	SIM	Boolean
Iris	91.67	93.67	90.0	20.85	20	20.43	14.35	13.7	13.03
Glass	51.16	55.12	49.74	121.5	125.6	127.8	79.55	112.4	105.6
Breast	92.55	94.12	92.0	89.1	79.1	88.3	57.5	45.1	58.2
Wine	80.28	91.53	88.9	62.8	28.65	48.35	36	21.4	28.56
Auto	77.15	78.73	72.2	139.65	122.9	148.6	99.35	91	104.5
Bupa	45.29	38.41	38.24	236.4	248.7	249.3	183.55	201.9	196.9
Machine	77.38	83.69	66.7	64.15	62	72.35	42.45	42.9	48.36
Pima	61.82	59.25	63.6	448.4	485.9	445.3	342.1	427.6	356.8
Ionosphere	83.94	88.48	85.9	90.25	74.3	82.23	1	1	1

of algorithm for discretization of real value attributes based on interval similarity have been decreased for most of the data. These results show the superiority of algorithm for discretization of real value attributes based on interval similarity.

From Table 4, we can see that under 1-V-1 classification method the predictive accuracy with SIM algorithm is higher than that of extended Chi2 algorithm and Boolean discretization algorithm except for Breast and Pima datasets.

Figures 2 and 3 visually describe predictive accuracy of decision tree and SVM with different discretization algorithms.

From the experiments we can see that the algorithm for discretization of real value attributes based on interval similarity proposed in this paper can obtain very good discretization effect.

We give a further analysis about the algorithms.

(1) In regard to data set with the greater number of classes, it is very possible that the difference of the number

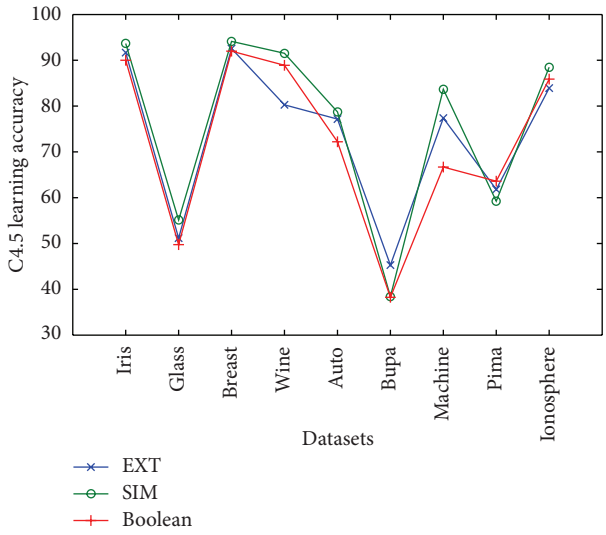

Figure 2: Comparison of C4.5 performance on 9 UCI real datasets.

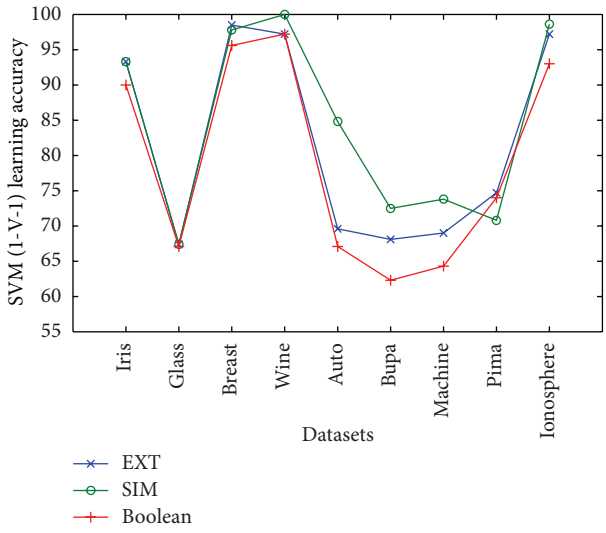

Figure 3: Comparison of SVM (1-V-1) performance on 9 UCI real datasets.

of classes of each group of adjacent intervals is very big. Thus, if extended Chi2 discretization algorithm was used, it is not accurate and unreasonable to merge first adjacent two intervals which have the maximal difference value. The method proposed in this paper has avoided the above situation. This is also the main reason that recognition effect of Glass and Machine datasets is effective.

(2) In regard to data set with the greater number of real value attribute, although the difference of the number of classes of each group of adjacent intervals is small, it may appear very unreasonable many situations in extended Chi2 algorithm in the discretization process: merged standard is not precise in computation, $\chi^2 = 0$ and so on. Regarding such situation, the method proposed in this paper has superiority very well (e.g. Ionosphere and Wine datasets).

Table 4: Comparison of SVM (1-V-1) performance on 9 UCI real datasets.

Datasets	EXT		SIM		Boolean	
	acc (%)	svs	acc (%)	svs	Acc (%)	svs
Iris	93.3	36	93.3	26	90.0	27
Glass	67.4	120	67.4	149	67.1	146
Breast	98.5	50	97.8	129	95.6	79
Wine	97.2	37	100.0	72	97.2	103
Auto	69.6	148	84.8	138	67.1	173
Bupa	68.1	180	72.5	167	62.3	197
Machine	69.0	101	73.8	70	64.3	130
Pima	74.7	339	70.8	357	74.0	370
Ionosphere	97.2	178	98.6	221	93.0	272

Under the comparison for two methods, the difference of recognition and forecast effect of Auto and Iris datasets (each of them has three classes) is small. But the method proposed in this paper is good.

(3) In regard to Auto and Iris datasets (each of them has two classes) class distribution difference of each adjacent two intervals is not big. It is improbable to appear unreasonable factors. This time, merged standard of extended Chi2 algorithm is possibly more accurate in computation. However, as the data of Breast is with less attribute and more samples, the intervals are massive in process of discretizing and inconsistency rate will increase easily. At this time, extended Chi2 algorithm produces the lower discretization effect; SIM algorithm proposed in this paper gets better discretization results by means of two important parameters' choice.

However, from the experiments we can see that SIM algorithm does not outperform extended Chi2 algorithm and Boolean discretization algorithm for all datasets. The characteristic of the data set on which SIM algorithm does not perform well is that it has lesser classes. That is the data set has not enough information of class.

6. Conclusions and Next Step of Work

Study of discretization algorithm of real value attributes operates an important effect for many aspects of computer application. Series of algorithms correlative to Chi2 algorithm based on probability statistics theory offer a new way of thinking to discretization of real value attributes. Based on the study for these algorithms a new algorithm using interval similarity technique is proposed. The new algorithm defines an interval similarity function which is regarded as a new merging standard in the process of discretization. At the same time, two important parameters (condition parameter α and tiny move parameter c) which embody equilibrium in the process of discretization and discrepancy of adjacent two intervals are given in the function. The new algorithm gives fair standard and can discrete the real value attributes exactly and reasonably, and not only can it inherit the logical aspects of χ^2 statistic, but also it can avoid the problems with the correlation of Chi2 algorithm. The theory analysis and

the experiment results show that the presented algorithm is effective.

Acknowledgments

This work is partly supported by the National Natural Science Foundation of China (Grant nos. 61105059, 61175055, and 61173100), International Cooperation and Exchange of the National Natural Science Foundation of China (Grant no. 61210306079), China Postdoctoral Science Foundation (Grant no. 2012M510815), Liaoning Excellent Talents in University (Grant no. LJQ2011116), Sichuan Key Technology Research and Development Program (Grant no. 2011FZ0051), Radio Administration Bureau of MIIT of China (Grant no. [2011] 146), China Institution of Communications (Grant no. [2011] 051), and Sichuan Key Laboratory of Intelligent Network Information Processing (Grant no. SGXZD1002-10).

References

[1] J. Dougherty, R. Kohavi, and M. Sahami, "Supervised and unsupervised discretization of continuous feature," in *Proceedings of the 12th International Conference on Machine Learning*, pp. 194–202, 1995.

[2] X. Liu and H. Wang, "A discretization algorithm based on a heterogeneity criterion," *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, no. 9, pp. 1166–1173, 2005.

[3] C. T. Su and J. H. Hsu, "An extended Chi2 algorithm for discretization of real value attributes," *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, no. 3, pp. 437–441, 2005.

[4] F. E. H. Tay and L. Shen, "A modified Chi2 algorithm for discretization," *IEEE Transactions on Knowledge and Data Engineering*, vol. 14, no. 3, pp. 666–670, 2002.

[5] Y. He and S. Hu, "New method for continuous value attribute discretization in rough set theory," *Journal of Nanjing University of Aeronautics and Astronautics*, vol. 35, no. 2, pp. 212–215, 2003.

[6] H. S. Nguyen and A. Skowron, "Quantization of real values attributes, rough set and Boolean reasoning approaches," in *Proceedings of the 2nd Joint Annual Conference on Information Science*, pp. 34–37, Wrightsville Beach, NC, USA, 1995.

[7] H. Xie, H. Z. Cheng, and D. X. Niu, "Discretization of continuous attributes in rough set theory based on information entropy," *Chinese Journal of Computers*, vol. 28, no. 9, pp. 1570–1574, 2005.

[8] 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 Transactions on Pattern Analysis and Machine Intelligence*, vol. 17, no. 7, pp. 641–651, 1995.

[9] W. Huang, *Discretization of continuous attributes for inductive machine learning [M.S. thesis]*, Department Computer Science, University of Toledo, Toledo, Ohio, USA, 1996.

[10] L. A. Kurgan and K. J. Cios, "CAIM Discretization Algorithm," *IEEE Transactions on Knowledge and Data Engineering*, vol. 16, no. 2, pp. 145–153, 2004.

[11] R. Kerber, "ChiMerge: discretization of numeric attributes," in *Proceedings of the 9th National Conference on Artificial Intelligence—AAAI-92*, pp. 123–128, AAAI Press, July 1992.

[12] H. Liu and R. Setiono, "Feature selection via discretization," *IEEE Transactions on Knowledge and Data Engineering*, vol. 9, no. 4, pp. 642–645, 1997.

[13] A. Bondu, M. Boullé, and V. Lemaire, "A non-parametric semi-supervised discretization method," *Knowledge and Information Systems*, vol. 24, no. 1, pp. 35–57, 2010.

[14] L. Martínez, D. Ruan, and F. Herrera, "Computing with words in decision support systems: an overview on models and applications," *International Journal of Computational Intelligence Systems*, vol. 3, no. 4, pp. 382–395, 2010.

[15] J. Park, H. Bae, and S. Lim, "A DEA-based method of stepwise benchmark target selection with preference, direction and similarity criteria," *International Journal of Innovative Computing, Information and Control*, vol. 8, no. 8, pp. 5821–5834, 2012.

[16] Z. Pei, Y. Xu, D. Ruan, and K. Qin, "Extracting complex linguistic data summaries from personnel database via simple linguistic aggregations," *Information Sciences*, vol. 179, no. 14, pp. 2325–2332, 2009.

[17] Z. Pei, D. Ruan, J. Liu, and Y. Xu, *Linguistic Values Based Intelligent Information Processing: Theory, Methods, and Application*, vol. 1 of *Atlantis Computational Intelligence Systems*, Atlantis Press/World Scientific, Singapore, 2009.

[18] M. Arif and S. Basalamah, "Similarity-dissimilarity plot for high dimensional data of different attribute types in biomedical datasets," *International Journal of Innovative Computing, Information and Control*, vol. 8, no. 2, pp. 1275–1297, 2012.

[19] S. H. Ha, L. Zhuang, Y. Zhou et al., "Treatment method after discretization of continuous attributes based on attributes importance and samples entropy," in *Proceedings of the International Conference on Intelligent Computation Technology and Automation (ICICTA' 11)*, pp. 1169–1172, Guangdong, China, March 2011.

[20] X. Z. Wang and C. R. Dong, "Improving generalization of fuzzy IF-THEN rules by maximizing fuzzy entropy," *IEEE Transactions on Fuzzy Systems*, vol. 17, no. 3, pp. 556–567, 2009.

[21] X. Wang, Y. He, L. Dong, and H. Zhao, "Particle swarm optimization for determining fuzzy measures from data," *Information Sciences*, vol. 181, no. 19, pp. 4230–4252, 2011.

[22] J. Yu, L. Huang, J. Fu, and D. Mei, "A comparative study of word sense disambiguation of english modal verb by BP neural network and support vector machine," *International Journal of Innovative Computing, Information and Control*, vol. 7, no. 5A, pp. 2345–2356, 2011.

[23] C. W. Hsu and C. J. Lin, "A comparison of methods for multiclass support vector machines," *IEEE Transactions on Neural Networks*, vol. 13, no. 2, pp. 415–425, 2002.

[24] F. Liu and X. Xue, "Constructing kernels by fuzzy rules for support vector regressions," *International Journal of Innovative Computing, Information and Control*, vol. 8, no. 7A, pp. 4811–4822, 2012.

[25] J. Pahasa and I. Ngamroo, "PSO based Kernel principal component analysis and multi-class support vector machine for power quality problem classification," *International Journal of Innovative Computing, Information and Control*, vol. 8, no. 3A, pp. 1523–1539, 2012.

[26] J. C. Platt, N. Cristianini, and J. Shawe-Taylor, "Large margin DAG's for multiclass classification," in *Advances in Neural Information Processing Systems*, vol. 12, pp. 547–553, MIT Press, Cambridge, Mass, USA, 2000.