Chi-Square Discretization Algorithm of Continuous Attribute Based on Information Entropy

Wei Yanming[1,a], Gan Xusheng[2,b], Li Huaping[3,c],

[1.] XiJing College, Xi'an, Shaanxi, 710123, China;

[2.] Air Traffic Control and Navigation College, Air Force Engineering University, Xi'an, Shaanxi, 710051, China;

[3.] PLA, Air Force, Xi' an Flight Academy, Xi'an, Shaanxi, 710306

[a]350293122@qq.com; [b]ganxusheng123@163.com; [c]fyw2121@sohu.com

Keywords: Wavelet Neural Network, Chi2 Algorithm, Attribute Reduction, Discretization.

Abstract: Rough Set (RS) theory has been widely concerned and studied in the related fields of uncertain information processing with its unique ability of data reduction, and the discretization of continuous attributes is an important link in RS method and other induction learning system. For this reason, a Chi-square (Chi2) Discretization Algorithm Based on Information Entropy is proposed. In the algorithm, the information entropy is used to replace the inconsistency rate in the traditional Chi2 algorithm, and the combination algorithm based on RS and wavelet neural network is established. The simulation result shows that the proposed algorithm is effective and superior using UCI data set.

1. Introduction

In Rough Set (RS) theory, the knowledge can be seen as the division of universe of discourse, and the roughness of knowledge can be defined from the view point of indiscernibility relation, achieving the rigorous analysis and processing of the knowledge based on classification mechanism [1]. As the basis of knowledge discovery, the attribute reduction can reduce the overall number of attributes under the premise of the same classification ability, and has become the core content of RS theory [2].

RS can be used to analyze the discrete symbolic attribute values with the semantics. The data analysis method based on RS is mostly for discrete data. In the practical problem, the value range of some conditional attributes or decision attributes is usually a continuous value, which needs to be discretized before RS processing. This is an important subject in the application of RS theory. For this reason, there are many new discretization methods of RS for continuous attribute. These discretization methods rely mostly on human subjective experience. In order to solve the problem of determining the decision attribute before WNN modeling, an algorithm that does not rely on experience and automatically discretizes continuous attributes need to be selected. We first briefly describe the problem of continuous attribute discretization of RS, and then introduce a heuristic automatic continuous attribute discretization algorithm from the perspective of information quantity (entropy) of decision table.

RS with the entropy-based discretization algorithm is used to reduce the input dimension of Wavelet Neural Network (WNN) for good modeling performance.

2. Attribute Discretization Problem

Suppose the decision table $S=(U,A,V,f)$, the universe of discourse $U=\{x_1,x_2,\cdots,x_n\}$,

$D=\{d\}$, $R=C\bigcup\{d\}$, the value range of decision attribute $V_d=\{1,2,\cdots,r(d)\}$, the condition attribute $a\in C$, the value range $V_a=[l_a,r_a]$, and there is a set of points $l_a<c_1^a<c_2^a<\cdots<c_{m_a}^a<r_a$, these points are divided through

$$V_a=[l_a,c_1^a)\bigcup[c_1^a,c_2^a)\bigcup\cdots\ \bigcup[c_{m_a-1}^a,c_{m_a}^a)\bigcup[c_{m_a}^a,r_a] \qquad (1)$$

where each c_k^a is called the breakpoint, and the value of the attribute a is divided into m_a+1 equivalence classes by these breakpoints. The purpose of discretiza- tion is to determine the set of appropriate breakpoints for all continuous attributes. If

$$f^p(x,a)=i\Leftrightarrow f(x,a)\in[c_k,c_{k+1}) \qquad (2)$$

The new decision table can be gotten, that is, after the discretization, the new information system instead of the original information system.

Substantially, the discretization problem of condition attributes can boil down to the problem that the condition attribute space is divided by the selected breakpoint, that is, the *m*-dimension space (*m* is the number of condition attributes) is divided into finite intervals. Obviously, the decision table discretized using the different division method may be different from the original decision table in compatibility. Assuming that an attribute has *m* attribute values, the number of candidate breakpoints that can be considered is *m*-1. With the increase in the number of attributes and the sample size, the number of candidate breakpoints doubles, so the efficiency of the breakpoint selection algorithm is very important for discretization.

3. Chi2 Discretization Algorithm

Chi-square (Chi2) algorithm is an improvement to ChiMerge method proposed by Kerber. From the Pearson theorem, it can be known that the asymptotic distribution of the statistics χ^2 is the χ^2 distribution with degrees of freedom $k-1$, that is, $\chi^2_{(k-1)}$ distribution. When the significance level is α, the corresponding critical value χ^2_α can be determined. In the ChiMerge algorithm, for a given significance level, it can determine whether the nodes are merged (the nodes with the largest difference is merged) according to the difference between the statistic χ^2 and the critical value. The difficulty is to artificially set the appropriate value. In the Chi2 algorithm, the significant level α keeps falling without fixing, and thus the threshold χ^2_α is increasing. The nodes with the largest difference $D=\chi^2_\alpha-\chi^2$ are merged, and judge whether to satisfy the termination condition by checking the inconsistency. This can greatly improve the calculation efficiency and information utilization. the basic concepts involved in Chi2 algorithm is as follows :

1. Interval and breakpoint

The initial interval of the data set is to treat each value as an interval, that is, the interval is the set of the attribute values. Two adjacent intervals need to be distinguished according to a breakpoint. Discretization of the continuous attributes is actually the process of eliminating the breakpoints and merging the adjacent intervals according to the certain criteria.

2. Statistics χ^2 and χ^2_α

In the discretization algorithm, χ^2 of the adjacent interval need to be calculated. The calculation method is

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

where *k* is the number of decision-making categories, the two intervals before and after the breakpoint are recorded as 1-th, 2-th interval. A_{ij} is the number of j class samples in the interval

i . E_{ij} is the expected frequency of A_{ij}, it can be calculated from

$$E_{ij} = \begin{cases} R_i \cdot C_j / N & R_i \cdot C_j \neq 0 \\ b & R_i \cdot C_j = 0 \end{cases} \tag{4}$$

where $R_i = \sum_{j=1}^{k} A_{ij}$ is the number of samples in the interval i. $C_j = \sum_{i=1}^{2} A_{ij}$ is the sum of the number of j class samples in the two intervals. $N = \sum_{i=1}^{2} R_i$ is the sum of the number of samples in the intervals at the two sides of the breakpoint. b is the constant, which can be taken as 0.1.

χ_α^2 is the critical value parameter, which is determined by the degree of freedom v and the significance level α of the adjacent interval. In statistics, the asymptotic distribution of the statistics χ^2 with k decision class is the χ^2 distribution with degrees of freedom $k-1$, i.e. $\chi_{(k-1)}^2$ distribution. When the significant level α is given, the corresponding critical value χ_α^2 can be determined, and $\int_0^{\chi_\alpha^2} f(x) = \alpha$, where $f(x)$ is the probability density function of χ^2 distribution with degrees of freedom $k-1$.

3. Inconsistency rate

The condition attributes of the object are the same, but the decision attributes are different, which shows that the classification information in the decision table has a certain inconsistency rate, the inconsistency rate is

$$Incon\_rate = 1 - \gamma_P \tag{5}$$

where γ_P is the approximation accuracy. $Incon\_rate$ is usually used to control the degree of merger and loss of information in the discretization process.

Chi2 algorithm conduct the discretization through two stages, and the degree of discretization is tested using inconsistency rate. The effect of discretization is better than that of Chimerge algorithm, but there are some shortcomings: 1. The lower limit of inconsistency rate need to be set in advance, but there is still a lack of uniform standards, it needs to try by multiple trials with an increase of the workload; 2. The inconsistency rate cannot fully reflect the classification characteristics, and these characteristics reflect the requirements on the degree of discretization of the sample data. When the multiple attributes are discretized, the effect of discreti- zation order doesn't be taken into account. On the basis of analyzing the relation knowledge of RS the information entropy can be used to replace the inconsistency rate, which can reflect the inherent characteristics of the sample data without selecting the parameter value, and can be automatically classified according to the classification characteristics of the sample itself. This can also be called Chi2 discretization algorithm based on information entropy.

In the modified algorithm, firstly, the information entropy measuring uncertainty is introduced, and the inconsistency rate of Chi2 algorithm is replaced, which can avoid the large amount of calculation caused by the parameter setting and can reflect the inherent characteristics of sample data. Then, the importance of each attribute is calculated and assigned to each of the breakpoints. In the ideas of backward optimization, according to the average importance of breakpoints, sort from small to large, and discretize the attributes. For attributes with the same average importance of breakpoints, the attributes with multiple breakpoints are preferred [3].

4. Wavelet Neural Network

The structure of WNN used is composed of one input layer (N_i node), one hidden layer (N_h node) and one output layer (1 node), as shown in Figure 1. The formula of WNN can be expressed as

$$g(\mathbf{x}) = \sum_{j=1}^{N_h} \omega_j \Phi_j(\mathbf{x}) + \sum_{k=1}^{N_i} c_k x_k \tag{6}$$

where $\Phi_j(\mathbf{x}) = \prod_{k=1}^{N_i} \phi(z_{jk})$, $z_{jk} = (x_k - a_{jk})/b_{jk}$, $\phi(z_{jk}) = \phi[(x_k - a_{jk})/b_{jk}]$. In the formula (6), $\Phi_j(\mathbf{x})$ is the output of j-th wavelet node, ω_j is the weight which between the wavelet node and the output node, and c_k is the connection weight between input node and output node, a_{jk}, b_{jk} are the translation factor and dilation factor respectively, $\mathbf{x} = [x_1, x_2, \cdots, x_{N_i}]^{\mathrm{T}}$ is the input vector.

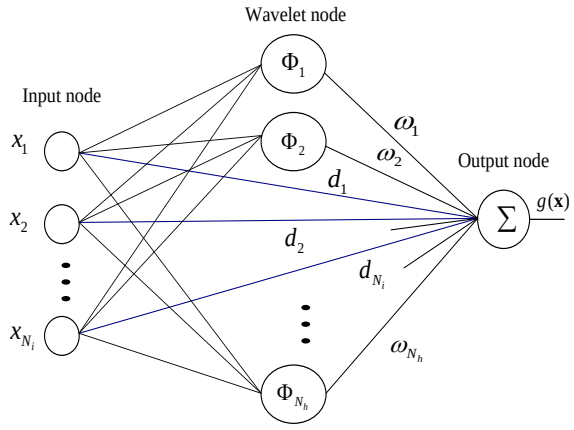

Fig. 1. Structure of WNN

In the architecture of Fig.1, with the different of wavelet basis function, the learning algorithms of WNN have many differences[4][5]. In the algorithm, the wavelet basis function can be taken as the real part of Morlet wavelet function.

5. Experiment Simulation

To verify the effectiveness of the method, RS-WNN proposed is used to establish the model and predict for three data sets with continuous attributes in UCI machine learning database. The experiment process is divided into four steps: 1. Discretize the data set; 2. Eliminate the redundant attributes according to RS reduction algorithm; 3. The data set after attribute reduction is divided into the training set and test set; 4. The training set is used to build the WNN model, and then the test set is used to validate the WNN model. For the sake of comparison, BP algorithm can be selected as the learning algorithm of WNN. According to the above steps, 10 experiments are performed for each data set, and the average value of experiments is taken as the final result. The selected UCI data sets and some of the experimental design information are shown in Table 1.

Table 1 UCI Data Set and Some Experiment Design Information

Data set	Housing	Triazines	R_wpbc
Brief description	Predict housing prices of Boston	Predict the biological activity of compounds	Predict the recurrence time of breast cancer
Sample number	506	186	194
Attribute number	13	60	32
Training set	400	140	46
Test set	106	46	44

For example, the proposed algorithm is used to discretize the decision table of the data set: Housing. Considering the attribute: *TAX full-value property-tax rate per $ 10,000*, some non-repeat sample value and its corresponding number of the attribute are shown in Table 2. After discretization, three division points can be obtained: namely 187.5, 332.0 and 688.5, and the four intervals [0, 187.5], [187.5, 332.0), [332.0, 688.5) and [688.5, ∞] can be constructed by discretization and assigned as 1, 2, 3, and 4 respectively. The discretize- ation results for the continuous attributes of the data set are given in Table 3.

Table 2 Non-Repetitive Sample Value and its Number of a Certain Attribute

Non-repetitive sample value	187	188	193	……	469	666	711
Number	1	7	8	……	1	132	5

Through the attribute reduction process of RS, the minimum condition attribute set can be determined as: *CRIM per capita crime rate by town*, *INDUS proportion of non-retail business acres per town*, *NOX nitric values concentration* and *RM average number of rooms per dwelling*. The original data set with the above four attributes is divided into training set and test set according to Table 1, which are used for training and test of WNN model respectively.

Table 3 Discretization Results of Continuous Attributes for the Data Set: Housing

Attribute name	Points number	Division point
CRIM per capita crime rate by town	8	0.010, 0.012, 0.014, 0.015, 0.090, 0.107, 5.848, 9.774
ZN proportion of residential land zoned for lots over 25,000 sq.ft.	1	83.5
INDUS proportion of non-retail business acres per town	6	0.600, 1.495, 1.900, 18.84, 23.77, 26.70
NOX nitric oxides concentration	5	0.387, 0.391, 0.393, 0.412, 0.755
RM average number of rooms per dwelling	3	4.001, 7.300, 8.753
AGE proportion owner-occupied units built prior to 1940	5	4.450, 6.700, 7.300, 9.350, 92
DIS weighted distances to five Boston employment centres	8	1.422, 1.506, 1.585, 1.800, 6.328, 6.602, 9.222, 9.905
RAD index of accessibility to radial highways	1	3.5
TAX full-value property-tax rate per $10,000	3	187.5, 332.0, 688.5
PTRATIO pupil-teacher ratio by town	3	12.80, 14.55, 21.15
B1000(Bk-0.63)^2 where Bk is the proportion of blacks by town	1	8.5
LSTAT % lower status of the population	6	2.225, 3.145, 3.935, 4.735, 5.545, 8.845

Fig. 2, Fig. 3 and Table 4 show some experiment results. Fig. 2 shows the training convergence effect of WNN for the data set: Housing; Fig. 3 shows the training convergence performance of RS-WNN for the data set: Housing, where *n* is the number of training times. The average experimental results for the selected 3 UCI datasets are compared in Table 1. It can be seen from the experimental results that the training and test accuracy of RS-WNN proposed is obviously better than that of WNN, and the convergence performance is faster than that of the latter. This shows that, the redundant attributes are firstly removed by RS and then conduct the modeling and predicting of WNN, which can effectively improve the performance of WNN model. More importantly, extracting the minimum conditional attributes by RS can simplify the structure and solution scale of WNN model, which greatly saves the space and time resources, mainly in the training time index. This is very important to solve the problems in reality.

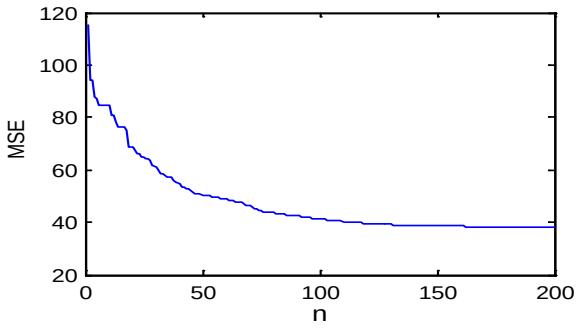

Fig. 2. Training convegence curve of WNN

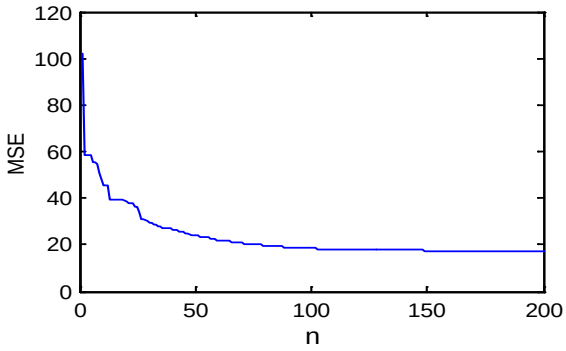

Fig. 3. Training convegence curve of RS-WNN

Table 4 Experimental Results of UCI Dataset and Their Comparison

Data set		Housing	Triazines	R_wpbc
WNN	Training MSE	38.3938	0.0277	1010.4017
	Test MSE	163.2305	0.0434	2352.0122
	Training time (s)	313.4	104.5	109.0
RS-WNN	Reduced attributes	4	6	3
	Training MSE	17.2643	0.0171	868.5014
	Test MSE	53.7699	0.0193	944.6320
	Training time (s)	266.3	944.6320	94.7

6. Conclusion

For the discretization problem in RS application, an improved Chi2 discretization algorithm is proposed, and WNN method based on RS pre-processing is constructed to verify the effectiveness of the proposed discretization algorithm. In the discretization algorithm, the information entropy measuring uncertainty is introduced to replace the inconsistency rate in Chi2 algorithm, avoiding the large amount of calculation caused by the parameter setting. The experiment results validate its feasibility.

References

[1] Z. Pawlak, “Rough sets,” International Journal of Computer and Information Sciences, vol. 1, No. 11, 1982, pp. 341-356.

[2] B. B. Qu, Y. S. Lu, “Rough set-based algorithm for attribute reduction,” Journal of Huazhong University of Science and Technology, vol. 33, No. 8, 2005, pp. 30-33.

[3] Y. B. Meng, J. H. Zou, G. H. Liu, X. S. Gan, “Optimization metnod for hidden layer nodes of WNN based on rough set reduction,” Control and Decision, vol. 29, No. 6, 2014, pp. 1091-1096.

[4] Q. H. Zhang, A. Benveniste, “Wavelet networks,” IEEE Transaction on Neural Networks, vol. 3, pp. 889-898, 1992..

[5] Q. H. Zhang, “Wavelet network in nonparametric estimation,” IEEE Transaction on Neural Networks, vol. 8 , pp. 227–236, 1997.