Proceedings of the Fourth International Conference on Machine Learning and Cybernetics, Guangzhou, 18-21 August 2005

A GLOBAL DISCRETIZATION METHOD BASED ON ROUGH SETS

HONG SHI[1], JIN-ZONG FU[2]

[1]Department of Computer Science, Tianjin University, Tianjin 300072 China
[2]Beijing YUPONT Electric Power Technology Co., Ltd., Beijing
E-MAIL: shiscene@yahoo.com.cn, jinzongfu@sina.com

Abstract:

Since rough sets theory can unveil the dependency of data and implement data reduction, it has been increasingly researched in more and more fields. In rough sets theory and other induction learning systems, discretization is an important algorithm and can be viewed as a process of information generalization (or abstraction) and data reduction. In this paper, a global discretization algorithm is proposed based on rough sets. It modifies the criterion of selecting the best cut points, and introduces inconsistency checking to preserve the fidelity of the original data，which change the MDLP method into a global one. Thus the reduction of cut points can be performed while keeping the consistency level. The proposed algorithm is tested by using several data sets with ID3 and ROSETTA. Experiments results show that this method performs better than MDLP, and is also superior to those which process continuous data directly without discretization.

Keywords:

Discretization; Rough sets; Consistency; Reduction

1. Introduction

Continuous attributes discretization is an essential problem in empirical learning systems. The task of discretization is to divide continuous value of each attribute into a few numbers of sub-intervals, and then map each interval to a discrete symbol (categorical, nominal, and symbolic). These discrete symbols are used as new values of the original attributes.

It is obvious that discretization broadens the using of empirical learning algorithms (such as AQ*, ID3, and ROSETTA, etc.); and it will also decrease the amount of data and accelerate the learning speed by discarding some detail information. Furthermore, discretization can be viewed as a process of concept generalization or abstraction. For example, the continuous values of attribute "Age" can be replaced by three symbols: "Young", "Middle" and "Senior". Therefore, discretization will make data easier to use, comprehend and interpret; in other words, the discretized data can be viewed as a kind of representation much closer to the knowledge-level than those continuous values. Therefore, classifiers of a compact size can be achieved. Many studies show that induction tasks can benefit from discretization: rules with discrete values are normally shorter and more understandable and they always improve the predictive accuracy.

Discretization will make impact upon the accuracy of classification. A survey of the literature reveals a great number of existing successful discretization algorithms and their applications[1-4]. In recent years, some new rough sets related discretization methods [5-7] are proposed with the development of rough sets theory. People are still engaged in finding discretization approaches of universal purpose (i.e. applicable to various data and inductive learning algorithms).

At present, the selection of the best discretization scheme for given data is still a complex issue and depends on a particular application. Usually, there are three important evaluation dimensions [8]: (1)the total number of intervals—intuitively, the fewer the cut-points, the better the discretization result; (2)the number of inconsistencies caused by discretization—it should be no much higher than the number of inconsistencies of the original data before discretization; (3)predictive accuracy—how does the discretization help to improve accuracy. In short, at least three dimensions are needed: simplicity, consistency, and accuracy. In theory, the best discretization can score highest in all the three factors, but in reality, it may not be achievable or necessary. It is a wiser and more reasonable choice to find an appropriate compromise among the three conditions.

Trying to achieve this goal, in this paper, we propose a global discretization algorithm by improving the local MDLP method [9] based on rough sets theory [10].

It needs to be pointed out, according to many existing criteria, that the discretization methods can be classified into different categories, such as supervised vs. unsupervised, local vs. global, and parameterized vs. non-parameterized. However, the definitions of different

categories are not very strict and the classifications usually look slightly different in different articles. For example, in some literatures[3], the term "global"(or "local") means converting all continuous attributes simultaneously (or restricting discretization to single one attribute), while it is called "multivariate" discretization in some other papers[4,8]. In this paper, the former meaning of "global" is adopted.

The rest of the paper is organized as follows. Section 2 reviews the MDLP method; Section 3 presents the proposed global rough sets related discretization algorithm; and Section 4 presents the experiment and results; section 5 is conclusions.

2. The MDLP method

The MDLP method is a famous recursive local discretization algorithm proposed by Fayyad and Irani, in which an information criterion called "minimum description length principal" is used.

Let a sample set S be composed of k classes $c_1, c_2, \cdots, c_k$, having probabilities $p_1, p_2, \cdots, p_k$ respectively. Then the entropy of S is defined as:

$$Ent(S) \triangleq -\sum_{i=1}^{k} p_i \log_2 p_i \tag{1}$$

In a classification problem, we need to evaluate the entropy of a given set partitioned by the selected attribute. Let an attribute A divides S into n disjoint subsets $S_1, S_2, \cdots, S_n$. Then, the entropy $E(A,S)$ of S partitioned by A is defined as the weighted average of entropies of subsets S_i $(i = 1, 2, \ldots n)$. That is:

$$E(A,S) \triangleq \sum_{i=1}^{n} P_i Ent(S_i) \tag{2}$$

Where,

$$Ent(S_i) = -\sum_{j=1}^{k} p_{ij} \log_2 p_{ij};$$

$$p_{ij} \triangleq \left|S_i \cap c_j\right| / \left|S_i\right|, j = 1, 2, \ldots, k;$$

$$P_i \triangleq \left|S_i\right| / \left|S\right|, i = 1, 2, \ldots, n.$$

And then, an entropy-based criterion for evaluating the importance of attributes on classification called Information Gain is defined as:

$$Gain(A) = Ent(S) - E(A,S) \tag{3}$$

Let d the definition field of the attribute $X(.)$, S the set of training samples. Among the possible cut points $d_j (j = 1, \ldots, k)$, we look for the d_t which leads to the "bipartition" on the training sample set S, that is $S = S_1 \cup S_2$, and this bipartition should satisfy the following MDLP criterion:

$$\begin{cases} Gain(d_t) > \dfrac{\log_2(n-1)}{n} + \dfrac{\delta(d_t)}{n} \\ Gain(d_t) = Ent(S) - \sum_{i=1}^{2} P_i Ent(S_i) \\ \delta(d_t) = \log_2(3^m - 2) - mEnt(S) + \sum_{i=1}^{2} m_i Ent(S_i) \end{cases} \tag{4}$$

Where, $n = \left|S\right|$ is the sample size; m is the number of the classes in S; m_i is the number of the classes in the subset S_i.

In turn, run the same process on the two subintervals, and so on. The process stops when the formula (4) can't be satisfied.

Considering the possible cut points $d_j (j = 1, \ldots, k)$, some reference show that it is not necessary to place initial interval boundaries before and after each example, possible cut points are between the attribute values taken by two examples of different classes. If several classes are superposed on a same value of $X(.)$, then the associated interval will be reduced to this unique value and, unlike other intervals, this interval will contain a mixing of the classes.

3. The proposed rough sets based method

The original MDLP method is a local discretization; it divides the value of an individual attribute independently into intervals with appropriate size according to the criterion denoted by the formula (4).

But it is likely to increase the level of inconsistency and result in degradation of fidelity of the original data set. Because in local methods, the effects of other attributes are ignored in the discretization process of one attribute, and valuable information in the data may be lost and critical relations may be destroyed. In contrast, global methods can usually get better results, because they take interacts of all attributes into account. On the other hand, since the significance of each attribute is not equal for preserving the information of original data, redundancy may be produced in global discretizations.

Therefore, we propose this improved two-stage global discretization algorithm. With rough sets theory, firstly, it tries to make the consistency of the result data not less than that of the original data; then the algorithm detects and eliminates potential redundancies produced in the first

stage.

3.1. Keeping consistency level in discretization

In order to guarantee that the discrete data accurately represents the original one, in stage one of the algorithm, a consistency checking is introduced as a global stopping criterion. The measure of consistency level is defined by the quality of classification in rough set. It is expressed as $Cnst(C_\delta) = \gamma(C_\delta, D)$, where $\gamma(C_\delta, D)$ represents the degree of the deterministic dependency between conditional and decision attributes, i.e.:

$$\gamma(C_\delta, D) = \sum_i^m \left| C_{\delta-}(Y_i) \right| \Big/ |U| , \quad Y_i \in U/D \qquad (5)$$

Where, δ is an information gain threshold, which can be adjusted according to the level of consistency; C_δ is the set of discrete condition attributes obtained with a certain value of δ .

Each continuous attribute is discretized one by one using the MDLP method, but with the criterion of formula (4) replaced by the following:

$$\begin{cases} \max(Gain(d_j)) - \min(Gain(d_j)) \geq \delta, \\ Gain(d_t) = \max(Gain(d_j)), \quad \text{j=1,…,k}_s \end{cases} \qquad (6)$$

Where, k_s is number of candidate cut points in recursive sub-interval.

The above process is repeated with a decreased threshold δ' changed according to $Cnst(C_\delta)$ until the consistency level $Cnst(C_\delta)$ restores to the original level (or exceeds a specified level).

In addition, if $\max(Gain(d_j)) - \min(Gain(d_j)) = 0$, the criterion of formula (4) is adopted.

3.2. Redundancy reduction

The discretization obtained through the stage one maintains the fidelity of data, but the total cut points number of all attributes is probably very large. Yet because the significance of each attribute is not equal for preserving the information of original data, it is not necessary to find too many cut points for all of attributes. As long as the more significant attributes are discretized appropriately, even though other less important attributes are under-discretized (i.e. getting fewer cut points), this discretization scheme will still preserve the consistency well. That means there are redundant discrete points existing in the result of the stage one.

In stage two, the theory of reduct and core of attributes in rough sets are employed to detect redundancy and reduce the redundant cut points, which won't change the discrimination degree.

Let $PT = \{d_1, d_2, \cdots d_n\}$ be the discrete cut points obtained in stage one; C_{PT} is the discrete attributes set corresponding to PT ; so that, PT can be viewed as a set of new conditional attributes; then the reduct of PT , denoted as $PTred$ ($PTred \subseteq PT$), can be defined. It satisfies the following two conditions:

(a) $Cnst(C_{PTred}) = Cnst(C_{PT})$;

(b) $\forall d_i \in PTred, Cnst(C_{PTred-\{d_i\}}) \neq Cnst(C_{PT})$;

And the core of PT , denoted as $PTcore$, is defined as $PTcore = \left\{ d_i \left| Cnst(C_{PT-d_i}) < Cnst(C_{PT}) \right. \right\}$.

3.3. Steps of the algorithm

According to the algorithm principle addressed above, steps of this algorithm are summarized:

1.1 Specifying an initial value to δ （and α ）, then calculating original data consistency C_0 ;

1.2 For each continuous attribute,

{

Supposing $d_j (j = 1, \ldots, k)$ are initial candidate cut points;

Using the formula (6) to decide recursively which candidate cut point is the desired discrete point, and adding it to PT .

}

1.3 Calculating the current $Cnst(C_\delta)$; if $Cnst(C_\delta) = C_0$ （or $C_0 - Cnst(C_\delta) \leq \alpha$ ）, go to 1.5;

1.4 Let $\delta' = (1-\Delta)\delta$, Δ is step length, go to 1.2;

1.5 Output the set of discrete point PT ;

2.1 Let $PTred = \varnothing$, and calculating the $PTcore$ from PT ;

2.2 Let $PTred = PTred \cup PTcore$,
$PTleft = PT - PTred$;

2.3 Calculating $Cnst(C_{PTred})$,
if $Cnst(C_{PTred}) = Cnst(C_{PT})$, go to 2.5;

2.4 Among $d_j \in PTleft$, choosing the d_t which satisfies

$Cnst(C_{PTred\cup d_t}) = \max(Cnst(C_{PTred\cup d_j}))$,

and let $PTred = PTred \cup d_t$,

$PTleft = PTleft - d_t$, then go to 2.3;

2.5 Output $PTred$.

The phase 2 uses a heuristic strategy, it does not ensure the obtained reduct is minimal, but in general, the result is satisfying. If necessary, we could Let $PTred$ as the initial PT and run the above process iteratively to get the minimal reduct.

4. Experiment

In order to compare the improved algorithm with the original MDLP algorithm, five data sets from the UCI machine learning repository have been selected, because each of them only includes numerical attributes and has no inconsistent objects. Tab.1 gives the summary of these data sets.

All of the data sets were discretized respectively using the proposed improved global discretization algorithm (denoted as IGD) and the MDLP algorithm. The discrete results are also listed in Tab.1.

Table 1. Data information and discrete results

Data set	No. of objects	No. of attributes	No. of classes	Consistency		No. of intervals	
				IGD	MDLP	IGD	MDLP
Bupa	345	6	2	1.0	0.0	33	7
Glass	214	9	7	1.0	0.38	29	22
Iris	150	4	3	1.0	0.84	12	12
Pima	768	8	2	1.0	0.32	36	17
Breast Cancer	699	9	2	1.0	0.99	22	29

Then we use two methods to evaluate the performance of the IGD algorithm. One is the rule generator based on rough sets provided by the tool kit of ROSETTA [11]; the other is ID3 algorithm, which is a well-known method and worked well for many decision-making problems.

The ten-fold cross-validation test method was applied to all the data sets. The data set was divided into 10 parts of which nine parts were used as training sets and the remaining one part as the test set. The experiments were repeated 10 times. The final predictive accuracy was taken as the average of the 10 predictive accuracy values.

The predictive accuracies of ten-fold cross-validation tests using different classification techniques are presented in Tab.2. In addition, since C4.5 algorithm can deal with the continuous attributes directly, the predictive accuracies of C4.5 on the continuous data sets were presented.

From Tab.1, it can be seen that the consistent degrees of the five sets are kept unchanged after the discretization by IGD, while degraded much by MDLP. The number of discrete intervals of each set does not increase much relative to the number of attributes, which will lead to appropriate size classifiers.

Table 2. The accuracy of classification by different techniques

Data set	ROSETTA%		ID3%				C4.5%
			No pruning		Pruned		
	IGD	MDLP	IGD	MDLP	IGD	MDLP	Continuous
Bupa	63.76	62.85	86.71	63.12	71.62	61.15	71.20
Glass	64.34	46.85	89.96	91.28	72.06	73.08	71.50
Iris	97.32	94.98	98.13	96.73	96.60	95.20	95.33
Pima	76.80	73.40	88.97	86.69	78.13	75.02	73.50
Breast Cancer	96.83	93.24	98.17	96.42	95.38	96.00	95.00

From Tab.2, it can be concluded that the performance of IGD is superior to that of MDLP when ID3 algorithm is employed; and the discretization by IGD is also better than processing continuous data by C4.5; for ROSETTA, the predictive accuracies are also more or less enhanced with the IGD algorithm.

5. Conclusions

In this paper, a global discretization algorithm is proposed. It modifies the criterion in selecting the best cut points of the MDLP, and makes the MDLP globalized by introducing the inconsistency checking based on rough sets theory in the first stage, which preserves the fidelity of the original data and overcomes the drawback of local discretization method. Then the reduction of cut points is performed, which will not change the consistency level and lead to small size learning model. For three evaluating criteria——the total number of intervals，the number of inconsistencies, and predictive accuracy, simulation results show that the proposed global algorithm is superior to the MDLP method.

Acknowledgements

The research work in this paper is supported by the fund of Tianjin University for young teachers.

References

[1] Susmaga, Robert. "Analyzing discretizations of continuous attributes given a monotonic discrimination function", Intelligent Data Analysis, Vol. 1, No.1-4, pp 157-179, 1997.

[2] Xindong Wu. "Correspondence — — fuzzy interpretation of discretized interval", IEEE transaction on fuzzy systems, Vol. 7, No. 6, pp. 753-759, 1999.

[3] Chmielewski. M. R, Grzymala-Busse. J. W, "Global Discretization of Continuous Attributes as Preprocessing for Machine Learning", International Journal of Approximate Reasoning, Vol. 15, No. 4, pp. 319-331, 1996.

[4] Chmielewski. M. R, Grzymala-Busse. J. W, "Global Discretization of Continuous Attributes as Preprocessing for Machine Learning", The Third International Workshop on Rough Sets and Soft Computing, San Jose, CA, pp. 474-480, Nov. 1994.

[5] Chao-Ton Su, Jyh-Hwa 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, March 2005.

[6] Meng-Xin Li, Cheng-Dong Wu, Zhong-Hua Han, Yong Yue, "A hierarchical clustering method for attribute discretization in rough set theory", Proceedings of 2004 International Conference on Machine Learning and Cybernetics, Vol. 6, pp. 3650-3654, Aug. 2004.

[7] Cai-Yun Chen, Zhi-Guo Li, Sheng-Yong Qiao, Shuo-Pin Wen. "Study on discretization in rough set based on genetic algorithm", 2003 International Conference on Machine Learning and Cybernetics, Vol.3, pp. 1430-1434, Nov. 2003.

[8] Huan Liu, Farhad Hussain, al, "Discretization: An Enabling Technique", Data Mining and Knowledge Discovery, Vol. 6, No. 4, pp. 393-423, 2002.

[9] Fayyad. U. M, Irani. K. B, "Multi-interval discretization of continuous valued attributes for classification learning", Proceedings of the 13th International Joint Conference on Artificial Intelligence (IJCAI '93), pp. 1022-1027, 1993.

[10] Pawlak. Z, et al, "Rough sets", Communications of the ACM, Vol. 38, No. 11, pp.89-95, 1995.

[11] Φhrn. A, Komorowski. J, "ROSETTA-A rough set toolkit for analysis of data", P.P.Wang (ed.), The Fifth International Workshop on Rough Sets and Soft Computing(RSSC'97) at Third Annual Joint Conference on Information Sciences(JCIS'97), Rough Set & Computer Science 3, pp. 403-407, 1997.