浙江科技学院学报,第 15 卷第 3 期,2003 年 9 月
Journal of Zhejiang University of Science and Technology
Vol.15 No.3, Sep. 2003

An algorithm of discretization of continuous attributes in rough sets based on cluster

XIANG Xin-jian[1], Stolle. M[2]

(1. Dept. of Information and Electrical Engineering, Zhejiang University of Science and Technology, Hangzhou 310012, China;
2. Fachhochschule Hannover, Hannover, Germany)

Abstract: Rough set theory is a new mathematical tool to deal with imprecise, incomplete and inconsistent data. The rough set theory can only deal with the discrete attributes. It can not deal with the continuous attributes. In order to overcome the limitation, by using the character of in size order for continuous attributes and error theory of statistics, this paper proposes an algorithm of discretization based on cluster. The result of comparison between this algorithm and some existing methods of discretization is encouraging.

Key words: cluster; error theory; rough set; discretization of continuous attributes

CLC number: TP273.5 **Document code**: A **Article ID**: 1671 - 8798(2003)03 - 0154 - 04

1 Foreword

In machine leaning and KDD(Knowledge Discovery Databases) studying, many algorithm demand attributes must be discrete. But in reality there are many continuous attributes. Therefore, to be discrete is necessary for continuous attributes.

Rough Set (RS) that was put forward by Pawlak is a new mathematical theory. It can deal with imprecise, incomplete and inconsistent data. But it is a pity that RS can't deal with continuous attributes. This defect limits its application range. It is the practicing need of application to improve RS theory able to deal with continuous attributes.

The task of discretization of continuous attributes is first to set up a number of demarcation points in the range of value region of continuous attributes, the next to divide the range of value region of continuous attributes into a number of small areas. The key of discretization of continuous attributes is how to determine the number and the position of demarcation points. At present, people have put forward many discretization algorithms. They almost based on no supervision discretization algorithm, such as demesne knowledge discretization algorithm, equal width discretization algorithm or equal frequency discretization algorithm and so on. These methods are simple and easy to achieve. But some of them demand abundant region knowledge. Some of them have not thought about the classification information of attributes over. So it is difficult to achieve the satisfactory discretization result. On the contrary, the methods of supervision discretization algorithm that have considered the classification information of attributes are more complex, such as information entropy method, blurring cluster method and Bayes decision algorithm. The main problem is that these methods have not thought

Received date: 2003 - 04 - 22

Foundation item: Funds granted by Zhejiang Provincical Natural Science Foundation(602007)

Biography: XIANG Xin-jiang, male, born in 1964 in Yongkang, Zhejiang, vice-professor, specializes in intelligent control.

about the sequence character of continuous attributes value region. In fact, the value of continuous attributes is a value sequence that is linear, especially the continuous attributes in rough sets.

Therefore, combining the merits of no supervision discretization algorithm with the merits of supervision discretization algorithm, and using the character of in size order for continuous attributes and error theory of statistics, we proposes a discretization algorithm based on cluster.

2 A discretization algorithm based on cluster of in size order

N independent samples are divided into m kinds. The number of classification is very large, equal to m^n . So it is very difficult to find the best classification method. But if n samples of in size order are divided into m kinds, only (m − 1) demarcation points need to be set up. So the number of classification is equal to (m_{-1}^{n-1}) . Obvious the number is smaller . It is easier to find the best classification method. The best classification method is more similar in kinds and less similar between the kinds. We use the sum of the distances between the elements in kind to the heart of the kind to express the similar character in kind (kind diameter).

$$A = \sum_{i=1}^{k} (X_i - X_g)^2 \tag{1}$$

Where

$$X_g = \frac{1}{k} \sum_{i=1}^{k} X_i$$

The error theory of statistics point out, the total square sum of every distance (between the elements in kind to the heart of the kind) can be divided into two parts. The first part is the square sum of every distance in kinds. The second part is the square sum of every distance between kinds. When the first part is smallest (i.e. the similar character in kinds is the biggest) the second part will become the biggest (i.e. the similar character between kinds is the smallest). Therefore we can define kind diameter as the loss function.

N sequence samples x_1, $x_2 \cdots x_n$, for convenience sake, we use their array number $1, 2, \cdots, n$ to express them, are divided into m kinds. One kind is shown as follows:

$$\{i_0, \cdots i_1 - 1\}, \{i_1, \cdots i_2 - 1\}, \{i_m - 1, \cdots i_m - 1\} \text{ where } 1 = i_0 < i_1 < \cdots < i_m = n + 1$$

We name $i_1, i_2, \cdots i_{m-1}$ as demarcation points.

M kinds in fact determine by (m − 1) demarcation points. The classification method is written as:

$f(n, m, i_1, \cdots i_{m-1})$ Or simply written: $f(n, m)$

If $A(i, j)$ is used to express kind diameter of region $\{i, j\}$, the loss function of above m-cluster is:

$$L[f(n, m, i_1, \cdots, i_{m-1})] = \sum_{j=0}^{m-1} (i_j, i_{j+1} - 1) \tag{2}$$

We define the cluster $f^*(n, m)$ which enable the loss function least as the best sequence m-cluster. Specific count as follows:

We first carry on the best sequence two-cluster:

$$L[f(n,2)] = \min_{2 \le i_j \le n} L[f(n,2,i_j)] = \min_{2 \le i_j \le n} \{A(1, i_{j-1}) + A(i_j, n)\} \tag{3}$$

The formula express: N sequence samples are divided into two kinds:

$$\{1, 2, \cdots i_{j-1}\}; \{i_j \cdots n\} = \{B_{11}\}; \{B_{12}\} \tag{4}$$

So produce two new sequence samples $\{B_{11}\}$, $\{B_{12}\}$. Using this method as above, we carry on the best sequence two-cluster once more to $\{B_{11}\}$, $\{B_{12}\}$. We have:

$$B_{11} = \{B_{21}\} \{B_{22}\}, B_{12} = \{B_{23}\} \{B_{24}\} \tag{5}$$

Using the same principle, we carry on the best sequence two-cluster. Then we can obtain 2^i samples. They are:

$$B_{i1};\ B_{i2} \cdots B_{ii}$$

When $i = \log 2m$, we can obtain the best sequence m -cluster. We define this method as the best sequence two-cluster algorithm.

3 The main step and algorithm of the best sequence two-cluster

According to the thinking above, the solution of the step of the best sequence m-cluster is shown as follows:

3.1 Calculate kind diameter of every sample classification

Appling formula(1), calculate kind diameter of every sample classification $A\ (1, i_{j-1})$ and $A\ (i_j,\ n)\ (i_j = 2,3,\cdots n)$, form a group of number of two-dimensional.

3.2 Calculate demarcation point of the best sequence two-cluster and value of loss function

Appling formula (3), find out the demarcation point i_j of the best sequence two-cluster which enable the value of loss function to be least; and find out the two new sample kinds B_{11} and B_{12} after demarcation.

Where $$B_{11} = (1,2,\cdots i_{j-1}\},\ B_{12} = \{i,\cdots n\}.$$

3.3 Form the new demarcation and sample kinds of the best sequence two-cluster to counter the sample kinds B_{ij}

Based on the sample kinds which is obtained in former demarcation, calculate kind diameter and value of loss function using formula (1) and formula (3) continuously, find out the new demarcation points of the best sequence two-cluster and form the new sample kinds.

3.4 The condition and the calculation to stop demarcation

The value of m in the best m-cluster waits to determine for many actual problems. Generally speaking, if n don't vary, $L\ [f(n,m)]$ will reduce with increase in m. I. e. the value of loss function will reduce with increase in the number of demarcation. When will stop demarcation to obtain the best cluster? Generally doing is: give a positive integer $\sum$, while counting the best two-cluster check the value of loss function. When $L\ [f\ (n,\ m)] < \sum$, the value of m this time is just the number of demarcation. Of course, the value of m can determine by region experts according to their experience.

However it's difficult to define the value of $\sum$. A new condition to stop demarcation is put forward in this paper. Generally speaking, $L\ [f\ (n,\ m)]$ will reduce with increase in m . But their reducing speed will vary with the difference of m . When m get in front of the best cluster, $L\ [f(n,m)]$ will reduce rapidly. When m get behind the best cluster, $L\ [f(n,m)]$ will reduce slowly. A clear dividing line will be formed between them two. According to this, a function of the reducing rate is lead into:

$$J(m) = L[f^*(n,m)]/L[f^*(n,m+1)] \tag{6}$$

Calculate the value of $J\ (m)$ according to the sequence $m = 1,2,\cdots$. Then fix the value of m to be the number of kinds, which enable $J\ (m) < 2$.

Example: there is a sample, it's:

{1.0,1.1,1.2,1.3,1.4,2.0,2.1,2.2,2.3,2.4,50,51,52,53,54,100,101,102,103,104}

There are 20 elements in all. According to the algorithm above, we can obtain the loss function, the function of the reducing rate, the demarcation points, kinds and decision to stop demarcation. They are shown in the table Ⅰ as follows.

Table Ⅰ the condition and the calculation to stop demarcation

Loss function	Reducing rate	Demarcation points	New kinds	Stop decision
$L_1 = 6272.7$	—	11	B_{11}; B_{12}	—
$L_2 = 6271.0$	$J = L_1/L_2 = 1.00$	5 11	B_{21}; B_{22}; B_{12}	B_{11} stop demarcation
$L_3 = 12.7$	$J = L_1/L_3 = 493.3$	11 16	B_{11}; B_{23}; B_{24}	—
$L_4 = 10.2$	$J = L_3/L_4 = 1.25$	11 14 16	B_{11}; B_{35}; B_{36}; B_{24}	B_{23} stop demarcation
$L_5 = 10.2$	$J = L_3/L_5 = 1.25$	11 16 19	B_{11}; B_{23}; B_{37}; B_{38}	B_{24} stop demarcation

The best classification is three-cluster: B_{11}; B_{23}; B_{24}, demarcation points are 11 and 16.

i. e. 20 sequence elements above are divided into three kinds as follows:

{1.0,1.1,1.2,1.3,1.4,2.0,2.1,2.2,2.3,2.4}

{50,51,52,53,54}

{100,101,102,103,104}

The conclusion is identical with our observation conclusion.

4 Experiment comparison

For test the performance of the algorithm above, we name it for BASE2. We compare it with the traditional C 4.5 algorithm and Greedy algorithm. The data come from a collection of machine learning test data.

The experiment method is: take 60% data from every data collection as train collection, the rest 40% data as test collection. 20 experiments are made in every data collection.

BASE2 algorithm is proved to be reliable and efficacious from table Ⅱ.

Table Ⅱ experiment comparison of the rate of mistakes

Test data collection	The rate of mistakes		
	BASE2	C 4.5	Greedy
Australian	13.5	14.7	14.3
Heart	18.6	20.6	19.7
Iris	5.1	4.5	4.0
Vebicle	26.8	27.4	27.2

5 Summary

Using the character of in size order for continuous attributes and error theory of statistics, a discretization algorithm based on cluster is proposed.

This algorithm is simple, reliable, efficacious, thinking clear, easy to understand. It can solve the problem of the best demarcation for the sequence samples. It is a good choose to disperse Rough Set continuous attributes. We have realized the algorithm using VC ++ in PC computer. As the same time the algorithm has been tested and compared in the collection of typical experiment data. We have achieved pleased result. The method of fault diagnosis in power transformer based on Rough Set theory just utilized the algorithm. It affords the algorithm strong proofs. Therefore the algorithm has nice application and spread value.

References:

[1] Pawlak Z. Rough stets-theoretical aspects of reasoning about data [M]. Kluwer: Kluwer Academic pub., 1991.

[2] Kerber R. Discretizaton of numeric attributes[A]. Proceedings 10th national conference on artificial intelligence[C]. New York: MIT Press, 1992.

[3] MIAO Duo-qian. A new method of discretization of continuous attributes in rough sets[J]. Acta automatica sinica, 2001, 27(3): 320-326.

一种基于聚类的粗糙集连续属性的离散化算法

项新建[1], Stolle. M[2]

(1. 浙江科技学院 信息与电气工程学系,浙江 杭州 310012;2. 汉诺威技术学院,汉诺威 德国)

摘　要:粗糙集理论是一种新的处理不精确、不完全与不相容知识的数学工具。粗糙集理论只能对离散属性进行处理,而不能处理连续属性。文中针对这一缺陷,利用连续数值属性有序性的性质和统计方差理论,提出了一种基于聚类的连续属性离散化算法。运用典型数据将本算法与现有方法进行了比较分析,得到了满意的结果。

关键词:聚类; 方差理论; 粗糙集; 连续属性离散化