Journal of Computational Science & Engineering 2 (2011) 85-91

Journal of Computational Science & Engineering

Available at http://www.asocse.org ISSN 1710-4068

A Novel Method of Continuous Attributes Discretization in Rough Sets Theory

Bin Jiang, Yunhong Zhang *

Information Engineering College of Qingdao University, Qingdao 266071, P. R China

Article Information

Article history:
Received 25 May 2011
Revised 18 June 2011
Accepted 12September 2011
Available online 15 October 2011

Keywords:
sets
continuous
attributes
cuts
discretization

Abstract

The discretization of continuous attributes is an important issue in the research of rough sets theory. It makes a great contribution to the process of knowledge acquirement. There are two steps selecting candidate cuts and confirming result cuts during the process of searching the cuts which are used to discretize continuous data. This paper puts forward a novel algorithm for selecting the sets of candidate cuts in the first step. In the second step, a top-down approach is adopted to confirm the importance of candidate cuts and a heuristic algorithm is proposed to figure out the cuts which are ultimately used in the discretization process. The simulation results demonstrate the validity of the new algorithm.

1. Introduction

As a powerful tool used to process uncertain knowledge, rough sets theory [1,2] has been widely applied in the fields of data mining, artificial intelligence, and pattern recognition. It was first proposed by the Polish mathematician Z.Pawlak. When processing practical data using a rough sets method, the attributes in the decision-making table are requested to be expressed by discrete value in a general way. If the value range of some attributes is continuous, we should discretize the value range. The discretization of continuous attributes in a decision-making table makes a great contribution to the following the process of knowledge acquirement and data mining. The problem of optimal discretization is a NP problem [3] ， and it has practical meaning to research efficiency and heuristic algorithm.

Nowadays, both foreign and domestic scholars have proposed many kinds of methods to discrete continuous attributes in rough sets. The methods can generally be divided into two kinds. One kind is the discrete algorithm to change the undistinguished relation of the original information system. The other kind is the discrete algorithm which does not change the undistingui-

*Corresponding author E-mail address: jiangbin80@126.com
zyh_qdu@126.com

2011 © The American Computational Science Society. All rights reserved.

shed relation of original information system. The primary difference between them is whether or not importing a new conflict base of the original information system after discretization. The first method [6] includes equidistant partition and equifrequent partition, etc. The methods should set parameters artificially. The most common method[7-10] of the second kind is the combination of Boolean logic and the rough sets theory method[3] proposed by Nguyen H.S. and Skowron. The benefit of the algorithm is that all possible cut sets can be solved according to the given information table and it will not import conflict no matter what kind cut set was adopted. The disadvantage is that time complexity and space complexity are all comparatively high and can't be used in the information table of large quantity data.

The article proposes a novel algorithm to select candidate cuts. The algorithm tries its best to reduce the number of candidate cuts, at the same time to guarantee the distinguished relation of original system. The traditional discretization algorithm based on cuts' importance was analyzed and top-down approach to confirm the most important cuts which were most optimal partition cuts was adopted. Based on above ideas, a novel algorithm of continuous attributes discretization was proposed and simulation test proved the validity of the algorithm.

2. The describing of rough sets' discretization problem

Decision-making table S can be expressed by S= (U, R, V, f) , here, U={x_1,x_2,…,x_m} is universe. R=C∪ {d} , C is named as condition attribute sets. {D} is named as decision attributes sets. V is the sets of attribute value. f: $U \times A \to V$ is an information function which used to appoint attribute value of every object x. Supposed that the number of decision kind is r(d). One cut in the value range Va of attribute a can be noted as (a, c). Here a ∈R and c is real number set. It is defined as a sort of Pa on Va to an arbitrary cuts sets in the value range of Va= [la, ra]. Real number set.It is defined as a sort of Pa on Va to an arbitrary cuts sets in the value range of Va= [la, ra].

$$P_a = \{[c_0^a, c_1^a), [c_1^a, c_2^a), \cdots, [c_{k_a}^a, c_{k_a+1}^a)\}$$

$$l_a = c_0^a < c_1^a < c_2^a < \cdots c_{k_a}^a < c_{k_a+1}^a = r_a$$

$$V_a = [c_0^a, c_1^a) \cup [c_1^a, c_2^a) \cup \cdots \cup [c_{k_a}^a, c_{k_a+1}^a]$$

So, arbitrary $P = \bigcup_{a \in R} P_a$ defines a new decision-making table SP=(U , R , VP ,f P), f P(x a)= i ⇔ f(x a)∈ $[c_i^a, c_{i+1}^a)$, x∈U, i ∈{0,…,Ka}. The table is named a discretization decision-making table.

Discretization is to make off the space which constituted by condition attributes using selected cuts naturally. It partitions n dimension space to limited regions and make decision values of every region's objects equal. There are two steps in the process of searching for the cuts. In the first step, sets of candidate cuts which used to continuous attributes' discretization. The second step is selecting result cuts set that select the possible smallest subset as cuts set adopted practically in discretization process candidate cuts.

3. Novel method of continuous attributes discretization based on rough sets theory

A Selecting algorithm of candidate cuts set

Selecting algorithm of candidate cuts set is the foundation of data discretization. On the precondition of assuring no change in the distinguished relation of the decision-making table, the number of candidate cuts should be selected to be the possible minimum. The approach that is usually adopted is that some continuous attributes in condition attributes are arrayed according to their value first, then the mean value of two neighborly attribute values is chosen as a candidate cut. All continuous condition attributes are operated as in the above steps and the candidate cuts set will be obtained at last. It is supposed that some attribute in the sample data has m values; we will obtain m-1

2011 © The American Computational Science Society. All rights reserved.

candidate cuts on the attribute using the above method. Following with the increasing of attributes numbers and the quantity sample, the number of candidate cuts will be increased geminately. Therefore, the method has low efficiency.

As explicated above, the decision-making table was supposed as S = (U, R, V, f) , and was supposed as continuous attribute, a∈C, d was supposed as decision attribute and its values were all discrete. a (x_i) denoted attribute value of object xi in a. d(x_i) denoted attribute value of object x_i in d*i*=1,2,…,|*U*|, |·| denoted cardinal number of set of·. The novel selecting algorithm of candidate cuts set was made out as follows:

Algorithm 1: Selecting algorithm of candidate cuts

Input: Decision-making table S;

Output: Candidate cuts set B;

Initialization: The data type of decision attribute value was changed to integer and candidate cuts set B=$\varnothing$;

Step 1: As to some attribute a∈C, object x∈U should be sorted in ascending order according to a(x). If the a(x) of several objects are equal, object x∈U should be sorted ascending order to d(x);

Step 2: i =1; while i<|U| {

As to objects xi and x_{i+1},

If d (x_i) ≠d (x_{i+1}) ,

a (x_{i+1}) would be selected as cut and added to B;

i++;

else i++;

}

B ***Top-down approach to confirm result cuts set***

The candidate cuts set B of decision-making table S can be obtained by the algorithm of section 3.1. In order to select result cuts from B to do discretization on S, we can measure the candidate cuts' resolving power on master according to the cuts' importance. First, the definition of a single cut and cut section are given in following.

Definition 1[9]: As to decision-making table S=(U, R, V, f), suppose there are j decision-making kinds which are denoted by D j (j=1,2,…,r(d)), suppose object set X⊆U and the I candidate cut of continuous attribute a is denoted as c_i^a, set X will be divided into two parts of X_L and X_R according to whether or not the object value of set X in a larger than c_i^a, here $X_L=\{x\in X : a(x)\le c_i^a\}$ $X_R=\{x\in X : a(x)> c_i^a\}$. The importance of cut c_i^a to set X is defined as:

$$W(c_i^a, X)=\left(\sum_{j=1}^{r(d)} L_j(c_i^a)\right)*\left(\sum_{j=1}^{r(d)} R_j(c_i^a)\right)-\sum_{j=1}^{r(d)} L_j(c_i^a)*R_j(c_i^a)$$

Here, $L_j(c_i^a)=card(D_j\cap X_L)$

$R_j(c_i^a)=card(D_j\cap X_R)$, $(j= 1,2,\ldots,r(d))$.

Definition 2[11]: Two cuts $c_L<c_R$ in an attribute can divide object set X⊆U into three parts as in figure 1. Supposing that L_1, … $L_{r(d)}$ expresses a(x)< c_L and belong to an object number distributing of each decision-making sort, M_1, … $M_{r.(d)}$ express $c_L\le a(x)< c_R$ and belong to the object number distributing of each decision-making sort, R_1, … $R_{r(d)}$ expresses a(x)≥c_R and belongs to the object number distributing each decision-making sort, the importance of the cut section [c_L, c_R] of attribute a to set X is defined as:

$$E(W(c_L,c_R,X))=\frac{W(c_L,X)+W(c_R,X)+conflict(c_L,c_R)}{2}$$

, here

$$conflict(c_L,c_R)=\sum_{i\neq j} M_i*M_j\cdot$$

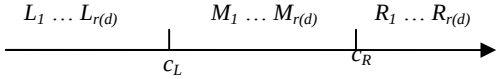

Figure1 Dividing to X by two cuts $c_L<c_R$ in attribute a

Discretization algorithm [9] is based on the

2011 © The American Computational Science Society. All rights reserved.

cuts' importance computes the importance of every candidate cuts and chooses the cut of the most important one, The most excellent cut which will be used to the divide set. The algorithm can approximate an optimal cut and its efficiency is better than a greedy algorithm [12], but the efficiency can be improved further. When confirming the most optimal partition cuts of set X in attribute a, the algorithm adopts a top-down approach. First, the candidate cut sequence after compositor in attribute is bisected and the importance of each section to set X is computed. The most optimal cut cannot be bisected but lies in the most important section [11]. Then the more important section will be bisected and the above work will be repeated until there is only one cut in a section. The cut will be the optimal partition cut.

Algorithm 2: Top-down confirming optimal partition cut of attribute a to set X

Input: Candidate cut set $C^a=(c_1,c_2,...,c_n)$ of attribute a. Set X⊆U;

Output: Most important cut of attribute a to set X, named optimal partition cut c best;

1) L=1;R=n;

2) while L<R {

ML=$\lfloor (Li+Ri)/2 \rfloor$; MR=$\lceil (Li+Ri)/2 \rceil$;

Compute $E(W(c_L,c_{ML},X))$ and $E(W(c_{MR},c_R,X))$;

If $E(W(c_L,c_{ML},X)) > E(W(c_{MR},c_R,X))$, then L=L;R=ML;

Else L=MR; R=R;

}

3) Output c_{best}=c;

Algorithm 2 will unnecessarily compute the importance of candidate cuts in other sections after confirming that the optimal partition cut lies in some cut section. So the computed complexity will be less than discretization algorithm based on the cuts' importance. Supposing the number of candidate cuts in attribute a is N, the computed complexity of the discretization algorithm based on cuts' importance is O(N) and it is O(logN) of algorithm 2 in the process of confirming optimal partition cut of attribute a to set X.

Algorithm 2 is the only optimal cut confirming method of the single attribute to single set that we are requested to confirm optimal cut of many attributes to many sets. Supposing that the decision-making table's equivalence class obtained by partition set P which has been selected cut, the importance of the cut section chosen in next step should be higher than the total importance to all equivalence class. Top-down confirming optimal partition cut of attribute to set X is given as follows.

Algorithm 3: Top-down confirming result cut

Input: Candidate cut set of all k continuous condition attributes and decision-making table S;

Output: Result cut set P;

Initialization: L= {U};P=∅;

1) The minimum and maximum of attribute a_i are corresponded to c_{Li}, and c_{Ri}. And the initial section set Cut={(a_1,L_1,R_1), (a_2,L_2,R_2), …(a_k,L_k,R_k)} will be obtained;

2) To every elements of set Cut , compute $Sum_i=\sum_{j=1}^{m} E(W(c_{L_i}^{a_i},c_{R_i}^{a_i},X_j))$, here, i=1,2,…,k, m is the number of equivalence class in L;

3) The cut section [L_i,R_i] of attribute ai which make the value of Sum maximal will be bisected. Supposing ML_i=$\lfloor (Li+Ri)/2 \rfloor$; MR_i=$\lceil (Li+Ri)/2 \rceil$, (a_i,L_i,ML_i) and (a_i,MR_i,R_i) will be obtained and added to the set Cut, and (a_i,L_i,R_i) will be deleted from the Cut;

4) If some attribute L of the Cut is equal toR, a result cut will be obtained and added to set P and deleted from cut set. Thenit will turn to step 5, or turn to step 2;

5) Divide L using the new result cut, to all Z∈L, if the new cut divides Z to Z_1 and Z_2, deletes Z from L and adds equivalence class Z_1 and Z_2 to L; If the decision attribute of all objects in the arbitrary equivalence class of L are equal, algorithm will end, output P;or else turn to step 1;

Supposing that the number of objects in universe is n , continuous condition attribute is k, and the average number of attribute's candidate cuts is N, the time complexity of algorithm 3 is O(nklogN) and its performance is better than a greedy algorithm [12]and algorithm based on the cuts' importance[9].

4 Test result

In order to testify the validity of the above

2011 © The American Computational Science Society. All rights reserved.

algorithms and to compare their performance, we experimented with classification rule mining on the machine learning database UCI. The result is listed in table 1. Firstly, the algorithms of literature [12] and literature [5] were compared with algorithm 1 of section 3.1 on their candidate cuts set selection results to the data set of table 1. Secondly, three discretization methods including greedy algorithm [12] which is expressed by A, and algorithm based on cuts' importance which is expressed by B, and algorithm 3 of section 3.2 which is expressed by C were used to discrete the sample. Then attributes reduction and attributes value reduction were processed to obtain the reasoning rule. Lastly, the test data was collected by identifying the test results and obtained information. During practical experimentation, half of the data of every data set was chosen randomly to learn and the other half of the data was gathered by the identifying test using the obtained reasoning rule. The basic deploy of the test computer was included by the CPU of AMD processor which had main frequency of 2 GHz and EMS memory of 512MB. The corresponding result are listed in tables 2, 3 and table 4.

As can be seen in table 2, the confirmed candidate cuts capability of algorithm 1 of section 3.1 was far excel than the algorithm of literature [12], and its effect were similar to the algorithm of literature [5] while its algorithm description was more compact than the literature [5]. Table 3 showed that the cut set obtained by algorithm 3 were closer to the optimal cuts set and the algorithm's carry time were less than the greedy algorithm and the algorithm based on cuts' importance.

Table2：data set of experiment

Dataset name	Number of objects	Number of condition attribute	Number of decision sort
iris	150	4	3
ecoli	336	7	8
glass	214	9	7
austra	690	14	2

Table3: number of candidate cuts

Dataset	Algorithm of literature [12]	Algorithm of literature [5]	Algorithm 1
iris	112	53	53
ecoli	339	282	294
glass	739	484	615
austra	872	658	674

Table4：recognition capability of discretization

Dataset	Number of residual continuous condition attribute		Number of result cuts				Processing time(s)		
	A	B	A	B	C	C	A	B	C
iris	3	3	7	9	9	3	2	1	1
ecoli	5	5	18	33	32	6	15	1	1
glass	8	6	13	26	32	7	23	2	1
austra	9	10	15	17	15	10	186	2	2

2011 © The American Computational Science Society. All rights reserved.

Table5：recognition capability of discretization result

Dataset	Correct recognition rate(%)			Fault recognition rate(%)			Denial recognition rate(%)		
	A	B	C	A	B	C	A	B	C
iris	97.2	92.4	95.1	2.8	7.6	4.9	0	0	0
ecoli	78.1	60.3	66.2	21.7	36.8	28.5	0.2	2.9	5.3
glass	66.5	64.5	65.4	33.5	32.2	30.4	0	3.3	4.2
austra	77.3	79.4	75.1	20.4	18.0	21.2	2.3	2.6	3.7

5 Conclusions

The target of data discretization based on rough sets was selecting candidate cuts of possible minimum on the precondition of assuring no change in the distinguished relation of original system. The problem of seeking least cuts was is a NP complete problem. So the efficiency of the cuts selecting algorithm was the key which decided whether or not solve large scale problem could be solved. The algorithm proposed in this article could reduce the candidate cuts' quantity. The efficiency of top-down method and bisection seeking optimal result cuts was higher than some reference algorithms. The experiment's results proved the validity of the algorithm.

References

[1]. Pawlak Z., "Rough Sets", International Journal of Computer and Information Sciences, 11 (1982) 341-356

[2]. Pawlak Z. and Skowron A., Information Sciences, 177, no.1, pp. 3-27, 2007.

[3]. Nguyen H. S. and Skowron A, "Quantization of real values attributes, rough set and Boolean reasoning approaches", In : Proceedings of the 2nd Joint Annual Conference on Information Science , Wright sville Beach , NC ,1995 , pp. 34-37

[4]. Wang Guo-yin. "Rough Set Theory and Knowledge Acquiring". Xi'an: Xi'an Traffic University Publishing House, 2001.

[5]. Zhao Jun, Wang Guo-yin, Wu Zhong-fu,etc. "Discretization Method Based on Rough Set Theory", Mini-Micro Systems,vol.25, no.1, pp. 60-64,2004.

[6]. He Ya-qun, Hu Shou-song. "A New Method for Continuous Value Attribute Discretization in Rough Set Theory". Journal of Nanjing University of Aeronautics & Astronautics, vol. 35, no.3, pp.213-215, 2003.

[7]. Nguyen S. H. and Nguyen H. S., "Some efficient algorithms for rough set methods", In : Proceedings of the Conference of Information Processing and Management of Uncertainty in Knowledge-Based Systems , Granada , Spain , 1996 , pp. 1451～1456.

[8]. Susmaga R. , "Analyzing discretizations of continuous attrib- utes given a monotonic discrimination function", Intelligent Data Analysis, vol. 1, no.4, pp.157-179, 1997.

[9]. Pawlak Z. and Skowron A., "Rough sets and boolean reason -ing", Information Sciences, vol. 177, no.1, pp. 41-73, 2007.

[10] Xie Hong, Cheng Hao-zhong, Niu Che-xiao. "Discretization of Continuous Attributes in Rough Set Theory Based on Information Entropy", Chinese Journal of Computers, vol. 28, no.9, pp.1570-1574, 2005.

[11] H.S. Nguyen, "On efficient handling of continuous attributes in large data bases", Fundamenta Informaticae, vol. 48 , no.1, pp.

2011 © The American Computational Science Society. All rights reserved.

61–81, 2001.

[12] Hou Li-juan, Wang Guo-yin, Nie Neng,etc. "Rough Set Discretization Indiscernible Relation Cut Core", Computer Science, vol. 27, no.12, pp. 89-94, 2000.

2011 © The American Computational Science Society. All rights reserved.