第28卷 第9期
2007年9月

仪 器 仪 表 学 报
Chinese Journal of Scientific Instrument

Vol. 28 No. 9
Sep. 2007

基于遗传算法的决策表连续属性离散化方法

陈 果
（南京航空航天大学民航学院 南京 210016）

摘 要：决策表连续属性离散化是用粗糙集理论处理连续问题的关键，本文进行了决策表连续属性离散化的遗传算法研究。首先对候选断点进行二进制编码，每个断点分别对应于二进制码的每位，其状态"1"和"0"分别对应断点的"取"和"舍"；然后通过构造适应度函数及交叉和变异算子，充分保证了决策表的分辨关系不变和断点数最少；最后，利用模拟数据和UCI机器学习数据对算法进行了验证，并与其他离散化方法进行了比较，结果充分验证了本文方法的有效性。

关键词：粗糙集理论；遗传算法；离散化
中图分类号：TP18 **文献标识码**：A **国家标准学科分类代码**：460.1540

Discretization method of continuous attributes in decision table based on genetic algorithm

Chen Guo
(*College of Civil Aviation, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China*)

Abstract: Discretization of continuous attributes is the key step in processing continuous problem with rough set theory. In this paper, genetic algorithm is used to carry out this task. Firstly, candidate cut points are encoded into binary code, in which a bit represents a cut point, and the value "1" and "0" denote "Adopted" and "Abandoned" respectively; Secondly, fitness function, crossover and mutation operators are constructed, which fully assure that discernible relationship of decision table is not changed and the number of cut points is minimum. Finally, the simulation data and UCI machine learning data are used to verify the new method, and the new method is compared with the other discretization algorithms. The results fully show the correctness and effectiveness of the proposed discretization method based on genetic algorithm.

Key words: rough set; genetic algorithm (GA); discretization

1 引 言

粗糙集理论只能对离散化数据进行处理，而样本数据一般为连续量，因此，决策表连续属性离散化是粗糙集理论处理连续问题的关键步骤。离散化是否成功直接关系到后续的知识获取。目前，已有多种连续属性值离散化方法，如等距离法[1]、等频率法[1]、L－方法[2]、W－方法[2]、P－方法[2]和C－方法[2]等，其中等距离和等频率方法算法简单、应用方便，但是会导致数据分布不均，丢失部分信息；L－方法、W－方法、P－方法和C－方法是基于统计学的离散化方法，若数据不充分，则结果意义不大。这些方法均不能保证离散化后的决策表分辨关系不变，而由Nguyen S. H. 和Skowron提出的布尔逻辑与粗糙集理论相结合的离散化方法[3]是粗糙集理论中离散化思想的重大突破，其基本思想是首先在保持信息系统的不可分辨关系不变的前提下，尽量以最小数目的断点把所有实例的分辨关系区分开，Nguyen S. H 和 Nguyen H. S. 在此基础上提出了贪心算法[4]，大大降低了计算的空间和时间复杂度，文献[5]又提出了几种改进的贪心算

收稿日期：2006-11 Received Date：2006-11

法,另外,还有将基于云模型的离散化方法[6]、以及将模糊性引进到离散化中的方法[7]。这些算法均有一定的合理性和适用范围,但算法复杂性相对较高。

本文以离散化后保持信息系统的不可分辨关系不变的前提下使断点数目最小为目标,引入遗传算法,对候选断点进行优化,同时对遗传算法的交叉和变异算子进行了改进,大大提高了遗传算法的收敛速度。与其他算法相比,本文方法概念简单、清晰、容易实现且精度很高。

2　离散化问题的描述

决策表 $S=\{U,R,V,f\}$,$R=C\cup\{d\}$ 是属性集合,子集 C 和 $\{d\}$ 分别称为条件属性集和决策属性集,$U=\{x_1,x_2,\cdots,x_n\}$ 是有限的对象集合即论域。设决策种类的个数为 $r(d)$。属性 a 的值域 V_a 上的一个断点可以记为 (a,c),其中 $a\in R$,c 为实数集。在值域 $V_a=[l_a,r_a]$ 上的任意一个断点集合 $\{(a,c_1^a),(a,c_2^a),\cdots,(a,c_{k_a}^a)\}$ 定义了 V_a 上的一个分类 P_a:

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

$$l_a=c_0^a<c_1^a<c_2^a<\cdots<c_{k_a}^a<c_{k_a+1}^a=r_a \tag{2}$$

$$V_a=[c_0^a,c_1^a)\cup[c_1^a,c_2^a)\cup\cdots\cup[c_{k_a}^a,c_{k_a+1}^a] \tag{3}$$

因此,任意的 $P=\bigcup_{a\in R}P_a$ 定义了一个新的决策表 $S^p=(U,R,V^p,f^p)$,$f^p(x_a)=i\Leftrightarrow f(x_a)\in[c_i^a,c_{i+1}^a)$,对于 $x\in U$,$i\in\{0,\cdots,k_a\}$,即经过离散化后,原来的信息系统被一个新的信息系统所代替。

离散化本质上可归结为利用选取的断点来对条件属性构成的空间进行划分的问题,把这个 n(n 为条件属性的个数)维空间划分成有限个区域,使得每个区域中的对象的决策值相同。假设某个属性有 m 个属性值,则在此属性上就有 $m-1$ 个断点可取,随着属性个数的增加,可取的断点数将随着属性值的个数呈几何增长,选取断点的过程也是合并属性值的过程,通过合并属性值,减少属性值的个数,减少问题的复杂度,这也有利于提高知识获取过程中所得到的规则知识的适应度。

3　基于遗传算法的连续属性离散化算法

由离散化问题的描述可知,连续属性离散化本质上是对初始断点的选择过程,同时也是断点的合并过程,显然选择不同的断点将会对决策表的分辨关系产生很大的影响,从而影响到后续的属性约简和规则提取。因此,需要研究如何选取最少的断点来保证决策表分辨关系不改变的问题,这显然是一个约束优化问题。

由于遗传算法[8]具有天生的隐含并行性和强大的全局搜索能力,它通过模拟生物适者生存的遗传进化原理来得到解空间的全局最优解。目前,遗传算法作为具有系统优化、适应和学习的高性能计算和建模方法,已经日趋成熟,并广泛应用于解决各工程应用中的优化问题。因此,本文将遗传算法引入来进行决策表断点的优化选择,构造了决策表连续属性离散化的遗传算法。

3.1　初始断点的选择

决策表 $S=\{U,R,V,f\}$,$R=A\cup\{d\}$ 是属性集合,子集 $A=\{a_1,a_2,\cdots,a_m\}$ 和 $\{d\}$ 分别称为条件属性集和决策属性集,$U=\{x_1,x_2,\cdots,x_n\}$ 是有限的对象集合即论域。则初始断点的计算步骤为:

Step1:选择属性 $a_i(i=1,2,\cdots,m)$;

Step2:对 $a_i(U)$ 中的值从小到大排序得到 $a_i(U)=\{a_i(1),a_i(2),\cdots,a_i(n)\}$;

Step3:设断点号 $j=0$,进行下列循环,即:

for ($k=1$ To n) { if($a_i(k)\neq a_i(k+1)$),得到断点 $p_i^j=\dfrac{[a_i(k+1)+a(k)]}{2}$ 同时 $j=j+1$ } 最终得到属性 a_i 的断点集合 $P_i=\{p_i^1,p_i^2,\cdots,p_i^{k_i}\}$,$k_i$ 为属性 a_i 的断点数;

Step4:如果决策表所有属性均计算完成,则输出断点集 $P_i(i=1,2,\cdots,m)$,计算结束。否则重复 Step1 到 Step4。

3.2　初始断点的遗传编码

由于离散化本质是选择初始断点,因此直接对初始断点的选择状态进行编码,而不必对断点值本身编码。因此,本文采用二进制编码方案,二进制码中的每一位对应一个断点,其值"1"和"0"分别代表该断点的"取"和"舍"。

设得到的初始断点集为 $P_i=\{p_i^1,p_i^2\cdots p_i^{k_i}\}$,$(i=1,2,\cdots,m)$,其中 k_{a_i} 为属性 a_i 的断点数。则对所有断点编码后得到的染色体如图 1 所表示。

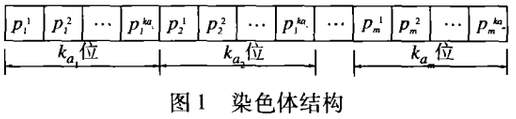

图 1　染色体结构

Fig. 1 The structure of chromosome

3.3　适应度函数

由于选择断点的原则是在不改变决策表分辨关系的前提下断点数最少。因此适应度函数应该由断点数和分辨关系来确定。

(1)求断点数 N_1:

Step1:$N_1=0$;

Step2:选择属性 $a_i(i=1,2,\cdots,m)$,并计算属性 a_i 的初始断点数 N_{i0};

Step3:对染色体解码后,计算属性 $a_i(i=1,2,\cdots,m)$ 的断点数目 N_i;

Step4:计算属性 $a_i(i=1,2,\cdots,m)$ 舍去的断点数目 $\Delta N=N_{i0}-N_i$;

Step5：$N_1 = N_1 + \Delta N$；

Step6：如果所有属性均计算完成，则输出 N_1，结束，否则转向 Step2 继续计算。

(2)求分辨关系的改变程度 N_2：

设决策表 $S = \{U,R,V,f\}$，$R = A \cup \{d\}$ 是属性集合，子集 $A = \{a_1, a_2, \cdots, a_m\}$ 和 $\{d\}$ 分别称为条件属性集和决策属性集，$U = \{x_1, x_2, \cdots, x_n\}$ 是有限的对象集合即论域。为了计算决策表的分辨关系需要引入如下概念：

定义1 不可分辨关系：对于每个属性子集 $B \subseteq R$，定义不可分辨二元关系 $IND(B)$，即：$IND(B) = \{(x,y) \mid (x,y) \in U^2, \forall b \in B(b(x) = b(y))\}$。

定义2 上下近似集：知识表达系统 $S = < U,R,V,f >$，对于每个子集 $X \subseteq U$ 和不可分辨关系 B，X 的上近似集和下近似集分别可以由 B 的基本集定义：

$$B_-(X) = \cup \{Y_i \mid (Y_i \in U \mid IND(B) \wedge Y_i \subseteq X)\} \quad (4)$$

$$B^-(X) = \cup \{Y_i \mid (Y_i \in U \mid IND(B) \wedge Y_i \cap X \neq \varphi)\} \quad (5)$$

式中：$U \mid IND(B) = \{X \mid (X \subseteq U \wedge \forall x \forall y \forall b(b(x) = b(y)))\}$ 是不分明关系 B 对 U 的划分，也是论域 U 的 B 基本集的集合。

在定义2中，取 $D = X = U \mid IND(\{d\}) = \{(x,y) \mid ((x,y) \in U^2 \wedge (d(x) = d(y)))\}$，$B$ 为条件属性集合，即 $B = A$；由定义2可以计算得到 D 的下近似集 $A_-(D)$ 和 D 的上近似集 $A^-(D)$。

下面计算分辨关系的改变。

Step1：对原决策表计算 D 的下近似集 $A_-(D)$ 和 D 的上近似集 $A^-(D)$；

Step1：对染色体解码后，得到所有属性 a_i($i = 1,2, \cdots, m$)的所有断点数目集 P_1；

Step2：用断点集对原决策表离散化，得到新的决策表 $S^p = (U,R,V^p,f^p)$；

Step3：对新决策表 $S^p = (U,R,V^p,f^p)$ 计算 D 的下近似集 $A_-(D)'$ 和 D 的上近似集 $A^-(D)'$；

Step4：如果 $A_-(D)' = A_-(D)$ 并且 $A^-(D)' = A^-(D)$，则 $N_2 = 1$(分辨关系不变)；否则 $N_2 = 0$(分辨关系改变)。

(3)求适应度函数：

$$Fitness = N_1 \times N_2 \quad (6)$$

显然，如果分辨关系改变，则适应度函数值为0，在分辨关系不变的情况下，断点数越少，适应度函数值越大。

3.4 交叉算子

遗传算法中的交叉运算是指对2个相互配对的染色体按某种方式相互交换部分基因，从而形成两个新的个体。交叉算法是遗传算法区别于其他进化算法的重要特征，它在遗传算法中起关键作用，是产生新个体的主要方法。一般要求它既不能太多地破坏个体编码串中表示优良性状的优良模式，又要能够有效地产生出一些较好的新个体模式。本文通过对各种交叉运算的比较分析，选择了文献[8]的均匀交叉算法，其主要操作步骤为：

(1)随机产生一个与个体编码串长度等长的屏蔽字 $W = w_1\ w_2 \cdots\ w_i \cdots\ w_l$，其中 l 为个体编码串长度；

(2)由下述规则从 A、B 两个父代个体中产生出两个基本点新的子代个体 A'、B'：

①若 $w_i = 0$，则 A' 在第 i 个基因座上的基因值继承 A 的对应基因值，B' 在第 i 基因座上的基因值继承 B 的对应基因值；

②若 $w_i = 1$，则 A' 在第 i 个基因座上的基因值继承 B 的对应基因值，B' 在第 i 基因座上的基因值继承 A 的对应基因值.

3.5 变异算子

在遗传算法中，交叉运算决定了全局搜索能力，而变异运算决定了遗传算法的局部搜索能力，同样非常重要。本文针对离散点优化问题，由于基因座中位的状态表示了离散点的存在与否，为了得到个数最小的断点集，本文对基本位变异算法进行了改进，改进算法为：

(1)对个体的每一个基因座，依变异概率 p_m 指定其为变异点；

(2)对每个指定的变异点，如果其基因值为“1”，则变为“0”，如果为“0”，则不变。从而产生新的个体。

通过实验，表明该改进的变异算子有效地改进了遗传算法的局部搜索能力，不仅加快了收敛速度，而且所得到的断点集最小。

3.6 遗传算法离散化流程

本文连续属性离散化的流程如图2所示。

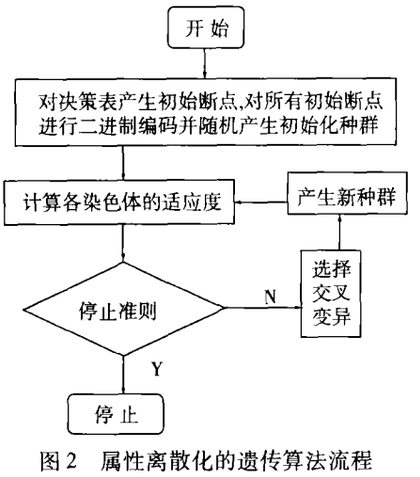

图2 属性离散化的遗传算法流程

Fig. 2 Flowchart of the discretization based on GA

(1)产生决策表的初始断点，按2.2节对初始断点进

行二进制编码。

(2)设定种群数目 n,种群数目太小,遗传算法的性能将变得很差或根本找不出问题的解,太大,则会增加计算量,使收敛时间增长,种群数目一般取为100~300个。

(3)对种群中的染色体解码,得到每条染色体代表的断点集,计算染色体的适应度函数值。

(4)将适应度最大的个体,即种群中最好的个体无条件地复制到下一代新种群中,然后对对父代种群进行选择、交叉和变异等遗传算子运算,从而繁殖出下一代新种群其他 $n-1$ 个染色体。本文采用转轮法作为选取方法,适应度大的染色体被选择的机会大,从而被遗传到下一代的机会大,相反,适应度小的染色体被选择的机会小,从而被淘汰的机率大。按3.4节和3.5节进行交叉和变异操作,交叉率一般取为0.5~0.9。变异率太大,将使遗传算法变为随机搜索,太小则不会产生新个体,一般取为0.01~0.1。

(5)如果达到设定的繁衍代数,返回最好的染色体,并以此获取最佳断点集,算法结束。否则,回到(3)继续下一代的繁衍。

4 算　例

4.1 算例1:模拟数据

为了与同类方法进行比较,本文首先选取一个来自文献[1]的决策表,并将本文算法与贪心算法[1]和属性重要性离散化方法[1]进行比较,离散化结果如表1所示,断点结果如表2所示。从表1和表2中可以看到,本文的遗传算法离散化方法与属性重要性得到了相同的离散化结果,具有3个断点,而贪心算法的断点数相对更多,为4个断点。

表1　不同算法的离散化结果比较

Table 1 Comparison of various discretization algorithms

原始决策表				贪心算法		属性重要性离散化		本文离散化方法	
U	a	b	D	a	b	a	b	a	b
x_1	0.8	2	1	1	3	1	2	1	2
x_2	1	0.5	0	1	1	1	1	1	1
x_3	1.3	3	0	1	4	2	2	2	2
x_4	1.4	1	1	2	2	2	1	2	1
x_5	1.4	2	0	2	3	2	2	2	2
x_6	1.3	1	1	1	2	2	1	2	1
x_7	1.6	3	1	2	4	3	2	3	2
x_8	4	3	1	2	4	3	2	3	2

表2　不同算法的离散化断点结果

Table 2 Discretization results of various discretization algorithms

离散化方法	属性 a 的断点集	属性 b 的断点集
贪心算法	1.35	0.75,1.50,2.50
属性重要性离散化	1.15,1.50	1.50
本文离散化方法	1.15,1.50	1.50

4.2 算例2:UCI机器学习数据

为了说明本文基于遗传算法的离散化方法的有效性,本文进行了规则知识获取实验,选择本文方法与贪心算法[1]和属性重要性离散化方法[1]进行比较。首先用不同的离散化方法将原始数据进行离散化处理,然后用文献[9]的归纳法进行属性约简和值约简,提取规则,最后用获取的知识进行测试。实验中采用了五组取自UCI机器学习数据库的不同实验数据,每组数据中,随机选一半用于学习,利用所得到的断点对其余数据进行测试,识别结果如表3所示,断点结果如表4。在本文算法中,遗传算法参数为:种群数为300,进化代数为100,交叉率为0.5,变异率为0.05。

规则推理采用前向推理机制,文献[8]的归纳法值约简方法得到的均为确定性规则,可信度均为1。定义正确识别为:

(1)测试样本在规则集中找到唯一与之相匹配的规则且规则结论正确;

(2)规则集中有多条规则与测试样本相匹配,规则结论均相同且正确。定义错误识别为:

(1)测试样本在规则集中找到唯一与之相匹配的规则且规则结论错误;

(2)规则集中有多条规则与测试样本相匹配,规则结论均相同且错误。拒识定义为:

(1)测试样本在规则集中找不到与之相匹配的规则;

(2)规则集中有多条规则与测试样本相匹配,匹配的规则结论不一样。

比较表3和表4可以发现,本文基于遗传算法的离散化方法从离散化后得到的断点数目来说基本上与贪心算法相当,而属性重要性算法所得到的断点数明显要多得多;从识别结果来看,本文算法要优于贪心算法,而属性重要性算法明显要差得多。从算法的复杂性来看,本文算法思路非常简单,容易实现,而贪心算法和属性重要性算法,均要对要构造新的信息表,原理更为复杂。由此可见,与其他方法相比,本文算法具有明显优势。

表 3 识别结果

Table 3 Recognition results

数据集	样本数	贪心算法			属性重要性算法			本文遗传算法		
		识别率	误识率	拒识率	识别率	误识率	拒识率	识别率	误识率	拒识率
Iris	150	0.949	0.051	0.00	0.949	0.051	0.00	0.949	0.051	0.00
Ecoli	336	0.648	0.247	0.105	0.175	0.077	0.748	0.641	0.204	0.155
Glass	214	0.358	0.265	0.377	0.282	0.282	0.436	0.368	0.376	0.256
HSV	122	0.433	0.183	0.384	0.450	0.250	0.30	0.466	0.154	0.380
Pima	768	0.438	0.108	0.454	0.387	0.142	0.471	0.443	0.136	0.421

表 4 断点结果

Table 4 Cut point results

数据集	原来条件属性数	初始断点数	贪心算法		属性重要性算法		本文遗传算法	
			剩余条件数	断点数	剩余条件数	断点数	剩余条件数	断点数
Iris	4	102	3	5	2	8	3	4
Ecoli	7	283	5	13	4	29	5	14
Glass	9	484	7	11	4	21	8	11
HSV	11	412	8	9	5	12	8	9
Pima	8	886	8	17	5	33	8	17

5 结 论

本文构造了决策表连续属性离散化的遗传算法，对候选断点进行了二进制编码，染色体的长度代表了候选断点的数目，染色体的位的状态表示了候选断点的取舍，并以断点数和对决策表的分辨关系改变作为计算适应度函数的依据。最后通过逐代遗传能够得到在不改变决策表的分辨关系的情况下使离散化的断点数目最小。通过与其他方法相比较，充分表明了本文方法不仅简单而且具有更好的离散化效果。

参考文献

[1] 王国胤. Rough 集理论与知识获取[M]. 西安：西安交通大学出版社，2001.
WANG G Y. Rough set theory and knowledge acquisition [M], Xi'an: Xi'an Jiaotong University Press, 2001.

[2] RYSZARD N. Evaluation of vibroaustic diagnostic symptoms by means of the rough sets theory [J]. Computers in Industry, 1992, 20:141-152.

[3] NGUYEN H S, SKOWRON A. Quantization of real values attributes, rough set and Boolean reasoning approaches [C]. Proceeding of the 2nd Joint Annual Conference on Information Science, Wrightsville Beach, Nc, 1995: 34-37.

[4] NGUYEN S H, NGUYEN H S. Some efficient algorithms for rough set methods [C]. In: Proc. of the Conference of Information Processing and Management of Uncertainty in Knowledge-based Systems. Granada, Spain, 1996: 1451-1456

[5] 侯利娟，王国胤，聂能，等. 粗糙集理论中的离散化问题[J]. 计算机科学，2000,27(12):89-94.
HOU L J, WANG G Y, NIE N, et al. Discretization problems in rough set theory [J]. Computer Science, 2000,27(12):89-94

[6] 李兴生. 一种基于云模型的决策表连续属性离散化方法 [J]. 模式识别与人工智能,2003,16(3): 33-38.
LI X SH. A kind of continuous attribute discretization method of decision table based cloud model[J]. PR & AI, 2003, 16(3):33-38.

[7] ROY A, PAL S K. Fuzzy discretization of feature space for a rough set classifier[J]. Pattern Recognition Letters, 2003, 24(6): 895-902.

[8] SYSWERDA D. Uniform crossover in genetic algorithm [C]. In: Proceeding of 3rd International Conference. on Genetic Algorithm, Morgan Kaufmann, 1989:2 ~9.

[9] 吴福保，李奇，宋文忠. 基于粗集理论知识表达系统的一种归纳学习方法[J]. 控制与决策，1999，14(3):206-211.
WU F B, LI Q, SONG W ZH. A induce learning method based on Rough set theory knowledge expression system [J]. Control & Decision, 1999, 14(3): 206-211.

作者简介

陈果,男,1972 年出生,分别于 1994、1997 和 2000 年在西南交通大学获得学士、硕士和博士学位,现为南京航空航天大学副教授,主要研究方向为航空发动机状态监测与故障诊断、智能诊断与专家系统、机器学习与知识获取、图像处理及模式识别、非线性转子动力学等。

地址:南京航空航天大学民航学院,210016

电话:+86-25-84891850;E-mail:cgzyx@263.net

Chen Guo, male, born in 1972, obtained bachelor degree in 1994, master degree in 1997, and doctor degree in 2000, all from Southwest Jiaotong University. Now he is an associate professor in Nanjing University of Aeronautics and Astronautics; his main research fields include aeroengine condition monitoring and fault diagnosis, intelligent diagnosis and expert system, machine learning and knowledge acquisition, image processing and pattern recognition, nonlinear rotor dynamics etc.

Address: College of Civil Aviation, Nanjing University of Aeronautics and Astronautics, Najing 210016, Jiangsu, China

Tel: +86-25-84891850; E-mail:cgzyx@263.net