广西科学院学报
Journal of Guangxi Academy of Sciences

2007,23(2):76～79
Vol. 23,No. 2 May 2007

一种基于粗糙集构造决策树的改进算法
An Improved Algorithm for Constructing Decision Tree Based on Rough Sets

王志强[1],吕跃进[2],操海燕[1],王　萌[1]
WANG Zhi-qiang[1],Lü Yue-jin[2],CAO Hai-yan[1],WANG Meng[1]

(1.广西大学电气工程学院,广西南宁　530004;2.广西大学数学与信息科学学院,广西南宁　530004)
(1. College of Electrical Engineering,Guangxi University, Nanning,Guangxi,530004,China;2. College of Mathematics and Information Sciences, Guangxi University, Nanning, Guangxi, 530004,China)

摘要:基于变精度粗糙集模型,对文献[3]提出的生成决策树方法进行改进,把变精度加权平均粗糙度作为属性选择标准,提出一种构造决策树新算法。新算法用变精度近似精度来代替近似精度,能有效地克服噪声数据在构造决策树过程中对刻画精度的影响,使生成的决策树复杂性降低,泛化能力更强。

关键词:决策树　粗糙集　变精度
中图法分类号:TP18　　**文献标识码**:A　　**文章编号**:1002-7378(2007)02-0076-04

Abstract: Based on Variable Precision Rough Sets Model, the decision tree inducing approach presented in Reference [3] is improved. The article presents a new algorithm for constructing decision tree with variable precision weighted mean roughness as the criteria for selecting attribute. The new algorithm effectively overcomes the influence of the noise data in structuring decision tree, reduces the complexity of decision tree and strengthens its extensive ability.

Key words: decision tree, rough sets, variable precision

决策树是一种有效的用于分类的数据挖掘方法。它的基本算法是贪心算法,采用自顶向下的递归方式构造决策树。构造一棵好的决策树关键是选择合适的属性作为划分结点。在各种决策树的构造算法中,比较有影响是 Quinlan 提出的基于信息熵的 ID3 算法,以后有它的改进版 C4.5、C5.0;其他常见的构造决策树方法还有 CART、CHAID、SLIQ、SPRINT 等[1]。

粗糙集理论是 Pawlak 教授于 1982 年提出来的,它是一种新的处理模糊和不确定性知识的数学工具[2]。目前,许多人提出了基于粗糙集理论的构造决策树算法,并与传统的基于信息熵构造决策树的方法相比较,指出其有效性。这些基于粗糙集构建决策树的属性选择标准有:基于加权平均粗糙度[3]、基于明确区域[4]、基于近似分类质量[5]和基于近似分类精度[6]等。另外也有基于粗糙集理论构造多变量决策树的方法[7]。这些构造决策树的方法都是基于经典 Pawlak 粗糙集模型。Pawlak 粗糙集模型所处理的分类是精确的,没有某种程度的近似,但在实际的数据中,尤其是在大型数据库中很可能存在许多噪声数据。这样基于 Pawlak 粗糙集模型构造决策树没能有效地处理噪声数据对生成决策树的影响,会使得分类变得过于细化,最终生成的决策树也比较复杂。变精度粗糙集模型是 Ziarko W[8]提出一种对粗糙集模型的扩展,它允许一定程度的错误分类率存在。文献[9]和[10]分别提出用变精度明确区域和变精度近似分类质量作为属性选择标准构造决策树,有效的减小了树的规模,提高了树的泛化能力。本文以变精度粗糙集模型为基础,对文献[3]提出的基于加权平均粗糙度属性选择方法加以改进,提出

收稿日期:2007-01-04
作者简介:王志强(1981-),男,硕士研究生,主要从事管理决策和数据挖掘研究。

以变精度加权平均粗糙度作为属性选择标准来构造决策树。

1 属性选择

构造决策树时,一般选择能够把实例尽可能正确划分到相应的类别中的属性作为分支结点。

定义 1[3] 加权平均粗糙度

$$\gamma_{Ri}=\sum_{j=1}^{m}1-\omega_j\mu_{Ri}(X_j),\qquad(1)$$

其中 $\mu_{Ri}(X_j)=|\underline{R_i}X_j|/|\overline{R_i}X_j|$,$\omega_j=|X_j|/|U|$,$Ri$ 表示第 i 个条件属性,j 表示决策属性的第 j 个等价类,m 是决策属性等价类的个数;X_j 表示决策属性的第 j 个等价类集合,U 表示非空有限集合(论域)。

在文献[3]中,作者用加权平均粗糙度来选择划分属性,这个概念涉及到决策属性的每一个取值情况,目的是使所选条件属性包含的确定性因素更多。条件属性的加权平均粗糙度越小,则其包含的确定性因素就越多,将其选为划分属性。

下面我们来考虑(1)式中的近似精度。在有些情况下,论域中少数元素会对在等价关系 R 下集合 X 的近似精度产生很大影响。举个简单的例子如下:$U/R=\{\{x_1,x_2,\cdots,x_{10}\},\{x_{11},x_{12},\cdots,x_{18}\},\{x_{19},\cdots,x_{30}\}\}$,$X=\{x_{12},x_{13},x_{14},x_{15},x_{16},x_{17},x_{18},x_{19}\}$,由等价关系 R 定义的集合 X 的近似精度 $\alpha_R(X)=|\underline{RX}|/|\overline{RX}|=0/20=0$。这样 X 在等价关系 R 下并不精确,比较粗糙。也就是根据其分类作出的决策是不准确的。事实上,从上面分类中可以看出等价关系 R 的分类对集合 X 描述的还是比较好,近似精度仅仅受到两个元素影响而变得很小。为此,我们引入变精度粗糙集模型,用变精度近似精度来代替近似精度。在本例中,设定分类误差 $\beta=0.2$,则由等价关系 R 定义的集合 X 的变精度近似精度 $\mu_R^\beta(X)=|\underline{R_\beta}X|/|\overline{R_\beta}X|=8/8=1$.

可以看到,用变精度近似精度克服了 x_{11} 和 x_{19} 对描述 X 在等价关系 R 下精确性的影响。在现实数据库中不可避免的存在许多噪声数据,这样用变精度近似精度能够在一定程度上消除噪声数据对刻画精度的影响。

定义 2 变精度加权平均粗糙度

$$\gamma_{Ri}^{\beta}=1-\sum_{j=1}^{m}\omega_j\mu_{Ri}^{\beta}(X_j),\qquad(2)$$

其中 $\mu_{Ri}^{\beta}(X_j)=|\underline{R_{i\beta}}X_j|/|\overline{R_{i\beta}}X_j|$,$\omega_j=|X_j|/|U|$,$Ri$ 表示第 i 个条件属性,j 表示决策属性的第 j 个等价类,m 是决策属性等价类的个数;X_j 表示决策属性第 j 个等价类集合,U 表示非空有限集合(论域)。

γ_{Ri}^{β} 的取值范围是[0,1],γ_{Ri}^{β} 越小则反映第 i 个属性包含的近似确定性越大。在构造决策树过程中,每次选择其最小的属性作为分支结点。这样生成决策树避免了对少量特殊数据的细化分类,大大提高了决策树的泛化能力。虽然在树的叶子结点中可能会有不一致的数据类别,但其错误分类率不会超过 β,从而保证了决策树分类具有一定的精度。

2 基于变精度加权平均粗糙度构造决策树算法

输入决策表和精度 β,即可输出一棵决策树。

算法步骤如下:

步骤 1:计算每一个条件属性的变精度加权平均粗糙度;

步骤 2:选择变精度加权平均粗糙度最小的属性 r 作为决策树划分属性;

步骤 3:用选择的属性 r 去划分训练集,相应于该属性的每一个取值产生一个分支(子表);

步骤 4:若子表中属于某一类别实例个数占表中总实例个数不小于(1 − β) 或表中没有可选的属性,则以该子表中占多数的实例类别标识该节点,并作为叶子结点;否则,将子表中的条件属性去掉已选划分属性 r,重复上述过程;

步骤 5:返回。

本算法与基于加权平均粗糙度构造决策树方法[3] 主要区别在于选取属性的标准不同。

3 实例分析

以表 1 中的数据为例,用本文提出的变精度加权平均度作为属性选择的标准构造决策树。表 1 中共有 26 个实例,条件属性{A,B,C,D},决策属性{E}。

根据各条件属性及决策属性我们可以得到如下等价分类:

$U/A=\{\{1,2,3,4\},\{5,6,7,8,9,10,11,21,22,23,24,25,26\},\{12,13,14,15,16,17,18\},\{19,20\}\}$;

$U/B=\{\{3,9,12,13,14,15,16,17,18,22,23\},\{1,5,6,10,11,20\},\{2,7,8,19,21,24\},\{4,25,26\}\}$;

$U/C=\{\{1,3,4,5,6,7,9,11,12,18,21,22,23\},\{10,13,14,15,16,17,26\},\{2,19,20\},\{8,24,25\}\}$;

U/D = {{1,5,8,10,11,13,18,19,20,24},{7,14,15,21,25,26}, {9,12,16,22}, {2,3,4,6,17,23}};

U/E = {{1,2,5,7,8,10,11,13,17,19,21,24,25,26}, {3,4,9,12,16,18,20,22,23}, {6,14,15,}}。

表1 决策表

U	A	B	C	D	E
1	1	2	1	1	L
2	1	3	3	4	L
3	1	1	1	4	R
4	1	4	1	4	R
5	2	2	1	1	L
6	2	2	1	4	B
7	2	3	1	2	L
8	2	3	4	1	L
9	2	1	1	3	R
10	2	2	2	1	L
11	2	2	1	1	L
12	3	1	1	3	R
13	3	1	2	1	L
14	3	1	2	2	B
15	3	1	2	2	B
16	3	1	2	3	R
17	3	1	2	4	L
18	3	1	1	1	R
19	4	3	3	1	L
20	4	2	3	1	R
21	2	3	1	2	L
22	2	1	1	3	R
23	2	1	1	4	R
24	2	3	4	1	L
25	2	4	4	2	L
26	2	4	2	2	L

在本例中设定 $\beta = 0.25$，计算各条件属性的变精度加权平均粗糙度，结果为：$\gamma_A^\beta = 1$，$\gamma_B^\beta = 0.7846$，$\gamma_C^\beta = 0.9379$，$\gamma_D^\beta = 0.6168$。可以看到属性 D 的变精度加权平均粗糙度最小，所以选择属性 D 作为根结点，并由它产生四个分支。每个分支对应着决策表的一个子表，在每个子集上重复以上操作，直到分支子表中某一类的数据个数占子表总数据个数的比例不小于 3/4 或表中没有可选的属性，以该类别名标识此节点作为叶子结点。图1就是用本文提出的方法构造的决策树，图2是以加权平均粗糙度[3]为标准构造的决策树。

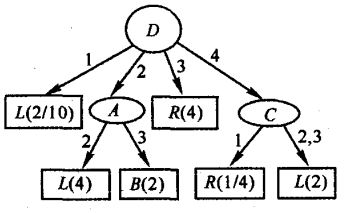

图1 基于变精度加权平均粗糙度的决策树

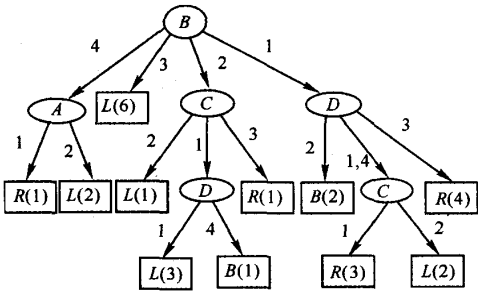

图2 基于加权平均粗糙度的决策树

比较图1与图2，我们可以看出用变精度加权平均粗糙度为标准构造决策树可以使树的复杂性大大降低。图1中共有9个结点，6个叶子结点，也即对应6条决策规则，而图2中共有17个结点，11个叶子结点，也即对应11条决策规则。运用变精度加权平均粗糙度构造的决策树允许小部分数据分类到其他类别中去，例如，在属性 D 值取1的10个数据中，大部分数据属于 L 类，只有2个数据属于其他类。这两个数据很可能是被误分的数据，我们把它认为噪声数据，将不对其细分。而运用文献[3]中的方法构造的决策树对决策表中的所有数据都正确分类，没有考虑到其中可能得噪声数据，这样的分类过于细化，造成对训练集过度拟合，泛化能力不强。

4 结束语

本文提出以变精度加权平均粗糙度作为构造决策树的属性选择标准，通过一个具体的实例，可以看出，用本文提出的改进方法构造的决策树，可以有效的弱化少数实例对决策树造成的不良影响，虽然决策中存在一定的误差，但决策树总体分类是比较好的，最终生成的决策树也比较简洁，且有效的处理了噪声数据，泛化能力也大大提高。

参考文献：

[1] 栾丽华，吉根林. 决策树分类技术研究[J]. 计算机工程，2004，30(9)：94-96.

[2] PAWLAK Z. Rough sets [J]. International Journal of Information and Computer Science，1982，11(5)：314-356.

[3] 蒋芸，李战怀，张强，等. 一种基于粗糙集构造决策树的新方法[J]. 计算机应用，2004，24(8)：21-23.

[4] 王名扬，卫金茂，伊卫国. 基于RST的决策树生成与剪枝方法[J]. 计算机工程与科学，2005，27(10)：69-70.

[5] 赵翔，向一丹，刘同明，等. 一种基于粗糙集的决策树生成算法[J]. 华东船舶工业学院学报，2005，19(4)：73-76.

[6] 关晓蕾,刘煜伟.一种基于粗糙集的决策树构造方法[J].科技情报开发与经济,2006,16(13):136-137.
[7] 苗夺谦,王钰.基于粗糙集的多变量决策树构造算法[J].软件学报,1997,8(6):425-430.
[8] ZIARKO W. Variable precision rough set model[J]. Journal of Computer and System Sciences, 1993, 46(1):39-59.
[9] 王名扬,卫金茂,伊卫国.变精度粗集模型在决策树生成过程中的应用[J].计算机工程与科学,2005,27(1):96-98.
[10] 常志玲,周庆敏,杨清莲.基于变精度粗糙集的决策树优化算法研究[J].计算机工程与设计,2006,27(17):3175-3177.

(责任编辑:韦廷宗)

(上接第 75 页)

了当扩散粒子距离种粒子的距离较远时,增大扩散粒子跳动的距离,并且该距离随着扩散粒子离种粒子的距离的增加而成倍增加,且在此模型的基础上采用回转半径法计算分形维数。从模拟的过程中,可以发现粒子跳动的距离将会对分形生长造成较大的影响,显然与真实的情况是有一定出入的,导致了模拟结果只是一个近似值。

在最近邻的条件下,由于粘附的粒子数少且平均半径小,所以生长的难度大,成核粒子数少。在次近邻的情况下,粘附的粒子数多,相对生长容易,在释放粒子数相同的条件下,成核的粒子数多。但是两种情况凝聚体的结构具有稳定且确定的分形维数,生长界面具有多重生长的分形的变化,生长体的半径和粒子数很好地符合幂律关系[6]。即使随机扩散粒子的总数从一万到几万变化,按 DLA 模型生长的分形凝聚体都满足标度不变性和自相似性,其分形维数均在 2.48 左右,而且在两种不同的点阵中,结果是一致的,这也说明了 DLA 模型其分形维数与点阵的种类关系不大(在小范围粒子数内),在生长体的内部,空间被占有程度一样,结构致密度也一样。

3 结束语

本文建立了三维凝聚生长模型及其近邻条件,并在两种近邻条件下进行了计算机模拟,计算了在不同情形下的分行维数,得出了具有一定意义的结果。尽管近邻条件不同,长成了不同的凝聚体外貌,但是,却有相同的分形维数,这说明凝聚体的分形维数与点阵的关系不大(在小范围粒子数内)。但对于种子粒子的移动和不同生长概率下的情况,还有待进一步的研究。

参考文献:

[1] MANDELBORT B B. Fractals, form, chance and dimension[M]. San Francisco: Freeman, 1977.
[2] WITTEN T A, SANDER L M. Diffusion-limited aggregation a kinetic critical phenomenon[J]. Phys Rev Let, 1981, 47: 1400.
[3] WITTEN T A, SANDER L M. Diffusion-limited aggregation[J]. Phys Rev, 1983, B27: 5686.
[4] 吴锋民,王衍,吴自勤,等.一维随机成核生长模型[J].物理学报,1996,45(12):1960-1969.
[5] 谢钢,张郑,陈书荣,等.点阴极下金属电沉积过程枝晶二维生长的计算机模拟[J].科学技术与工程,2003(4):343-346.
[6] MEAKIN PAUL. Diffusion-controlled cluster formation in 2~6 dimensional space[J]. Phys Rev Lev Let, 1983, 27: 1495-1507.
[7] MUTHUKUMAR M. Mean-field theory for diffusion-limited cluste formation [J]. Phys Rev Lev Let, 1983, 14: 839-842.
[8] HENTSCHEL H G E. Fractal dimension of diffusion-limited aggregates[J]. Phys Rev Lev Let, 1984, 16: 212-214.

(责任编辑:韦廷宗)