基于 VPRS 多变量决策树优化算法①

邱云飞　王　光　关晓林　邵良杉　(辽宁工程技术大学系统工程研究所 辽宁 葫芦岛 125105)

摘　要： 噪声数据降低了多变量决策树的生成效率和模型质量，目前主要采用针对叶节点的剪枝策略来消除噪声数据的影响，而对决策树生成过程中的噪声干扰问题却没有给予关注。为改变这种状况，将基本粗糙集(rough set，RS)理论中相对核的概念推广到变精度粗糙集(variable precision rough set，VPRS)理论中，并利用其进行决策树初始变量选择；将两个等价关系相对泛化的概念推广为两个等价关系多数包含情况下的相对泛化，并利用其进行决策树初始属性检验；进而给出一种能够有效消除噪声数据干扰的多变量决策树构造算法。最后，采用实例验证了算法的有效性。

关键词： 单变量决策树；多变量决策树；噪声数据；变精度粗糙集；相对核

Optimization Algorithm for Multivariate Decision Trees Based on VPRS

QIU Yun-Fei, WANG Guang, GUAN Xiao-Lin, SHAO Liang-Shan (Institute of System Engineering, Liaoning Technical University, Huludao 125105, China)

Abstract: When construct multivariate decision trees, noise data reduced the training efficiency and quality of model, most of the present pruning methods aimed at leaf node to eliminate the influence of noise data, but not pay attention to the disturbed problem of noise data when selected testing attribute. In order to solve the problem, extends the relative core of attributes in rough sets theory to variable precision rough set(VPRS), and uses it for selection of initial variables for decision tree; extends the concept of generalization of one equivalence relation with respect to another one, to relative generalization equivalence relation under mostly-contained condition, and uses it for decision tree initial attribute check;propose an algorithm for multivariate decision tree that can avoid disturbance of noisy data. Finally, validated the algorithm by an experiment.

Keywords: univariate decision trees; multivariate decision trees; noisy data; variable precision rough set; relative core of attributes

决策树学习是以实例为基础的归纳学习，着眼于从一组无次序、无规则的实例中推理出决策树表示形式的分类规则。现有的大多数决策树构造算法被限制在每个结点上只检验单个属性，如 ID3[1]，AQ11[2]，ASSISTANT[3]和 GREEDY3 &GROVE[4]等，这样的决策树被称之为单变量决策树。这一限制使得对很多复杂概念的表达变得困难或无法表达[4]，为了解决这个问题人们提出了许多多变量决策树构造算法，即在树的结点上可以同时检验多个属性，如苗夺谦、王珏[5]利用粗糙集中条件属性相对于决策属性的相对核的概念构造多变量决策树，Brodley C E[6]等人采用初始属性的线性组合来构造多变量决策树。分析发现，这些多变量决策树构造算法虽然可以降低树的复杂度，提供更简洁的类别描述，但其与单变量决策树构造算法一样都是噪声敏感的，对训练集的数据质量要求较高，在有噪声数据干扰的情况下，或者树的规模过大[7]，或者不能正确对训练集中的数据分类。关于什么是噪声，Quinlan 的定义是训练例子中的错误就是噪声。

① 基金项目:国家自然科学基金(70971059)；辽宁省创新团队项目(2009T045)；辽宁省科技攻关项目(2007308003)
收稿时间:2010-05-03;收到修改稿时间:2010-06-21

它包含两方面：一是特征值取错，二是类别取错[8]。我们分析，既然是噪声，那么它就是数据集中的少量的错误的数据，错误比例不应太高，如果错误比例过高，也就失去了学习的意义。那么，能否提出或是改进现有的决策树学习算法，将带有一定噪声干扰的数据库直接作为各种学习算法的训练集呢？如果可以的话，这将使我们算法的应用领域得到大幅度的扩展。

Ziarko[9]等人提出的变精度粗糙集(variable precision rough set，VPRS)模型是对Pawlak.Z的粗糙集(rough set，RS)模型的一种推广，相对于RS模型来说VPRS模型能有效处理噪声数据，进而用于知识的约简及知识相依性的分析。因此,可以使用VPRS模型解决数据集中的噪声数据对决策树学习算法的干扰问题。

构造能够抵抗噪声数据干扰的多变量决策树的关键是多变量检验的构造问题。它涉及2个方面：一是在有噪声干扰的情况下选择什么样的初始属性包含在多变量检验中；二是在有噪声干扰的情况下如何利用选择的属性来构造多变量检验？本文提出了一种能够有效抵抗噪声数据干扰的多变量决策树构造方法。本文第1节将Pawlak粗糙集模型中条件属性相对于决策属性的核的概念推广到变精度粗糙集模型中，给出了变精度粗糙集模型下相对核的定义，解决了有噪声数据干扰下的多变量检验中初始属性的选择问题；从容忍数据中存在一定噪声现象的角度出发，定义了两个等价关系多数包含情况下的相对泛化，并将它用于构造多变量检验；第2节描述了有噪声干扰的情况下构造多变量决策树的算法；第3节通过一个实例给出了基于VPRS的多变量决策树的构造方法。第4节给出了本文的结论及进一步研究的问题。

1 决策树中检验属性的选择及相对泛化

1.1 变精度粗糙集模型的相对核及属性的选择

设 X 和 Y 表示有限论域 U 的非空子集，令 $0.5<b\le 1$，定义多数包含关系(majority inclusion relation)[9]：$Y\overset{b}{\supseteq}X \Leftrightarrow D(Y\mid X)\ge b$

其中，

$$D(Y\mid X)=\begin{cases}|X\ \mathbf{I}\ Y|/|X|, & |X|>0\\ 0, & |X|=0\end{cases} \quad (1)$$

$|X|$ 表示集合 X 的基数，即集合 X 中元素的个数。称 $D(Y\mid X)$ 为集合 Y 关于集合 X 的相对包含度。即只要 X 中被包含于 Y 中的元素占到 X 的比例不小于 b 时，即可认为 Y 包含 X。特别的，当 $b=1$ 时，变精度粗糙集模型就退化为Pawlak粗糙集模型。

VPRS通过设置阈值参数 $(0.5<b\le 1)$，放松了RS理论对近似边界的严格定义，即允许一定程度的错误分类率的存在。随着 b 变大，VPRS模型的近似边界区域变宽，即变精度粗糙集意义下的不确定区域变大。特别的，当 $b=1$ 时，变精度粗糙集模型就退化为Pawlak粗糙集模型，因此RS模型是VPRS模型的一个特例。相对RS模型来说VPRS模型对数据的噪声有一定的容忍度。

设 $S=(U,A,V,f)$ 是一个信息系统，U 为论域，A 为属性的非空有限集，V 表示所有属性可取值的集合，f 是一个信息函数，该函数为论域中某个数据实体的某个属性赋予一个特定的值；设 $P,Q\subseteq A$ 为条件属性集和决策属性集，则 $ind(P)$, $ind(Q)$ 表示由 P,Q 决定的不可区分关系($ind(P)=\mathbf{I}P$ 表示所有属于 P 的等价关系的交，也是一个等价关系)，关系 $ind(P)$ 的等价类的集合称为条件类，用 U/P 表示，关系 $ind(Q)$ 的等价类的集合称为决策类，用 U/Q 表示。

设 $\{X_1,X_2,\cdots,X_i,\cdots,X_n\}$ 是论域 U 在等价关系 $ind(P)$ 上的划分，则满足 $\mathbf{U}\{X_i\mid X_i\in U/P,\ i\in 1,2,\cdots,n\}=U$，且 $X_i\ \mathbf{I}\ X_j=\varnothing, i\ne j$。$\{Y_1,Y_2,\cdots,Y_j,\cdots,Y_m\}$ 是论域 U 在等价关系 $ind(D)$ 上的划分。我们称 Y_j 为一个概念，是我们要描述的对象；称 X_i 为已知的知识块，用来近似描述 Y_j；而把用 X_i 对 Y_j 的描述称为规则。

对于决策等价类 Y_j 在阈值参数 b 时的上下近似集分别定义为：

$$\underline{R}_P^b(Y_j)=\mathbf{U}\{X_i\in U/P\mid D(Y_j\mid X_i)\ge b, i=1,2,\cdots,n\} \quad (2)$$

$$\overline{R}_P^b(Y_j)=\mathbf{U}\{X_i\in U/P\mid D(Y_j\mid X_i)>1-b, i=1,2,\cdots,n\} \quad (3)$$

$\underline{R}_P^b(Y_j)$ 是由 U 中那些在现有知识 P 下在正确分类水平 b 时属于概念 Y_j 的元素组成的集合；$\overline{R}_P^b(Y_j)$ 是在正确分类水平 $1-b$ 时属于概念 Y_j 的元素组成的集合。

对于条件属性集 P 和决策属性集 Q，有 Q 的 P -正区域定义为

$$POS_P^b(Q)=\mathop{\mathbf{U}}_{Y_j\in U/Q}\{X_i\in U/P\mid D(Y_j\mid X_i)\ge b, i=1,2,\cdots,n, j=1,2,.....m\} \quad (4)$$

设 U 是一个论域，P 和 Q 是定义在 U 上的2个等价关系族，称一个等价关系 $R\in P$ 是 $Q-$ 不必要的(或多

余的)，如果式

$$POS_P^b(Q) = POS_{(P-\{R\})}^b(Q) \quad (5)$$

成立。否则，R 在 P 中是 $Q-$ 必要的。

P 中所有 $Q-$ 必要的等价关系组成的集合，称为 P 的 $Q-$ 核，记作 $CORE_Q^b(P)$。

当 P,Q 分别表示信息系统的条件属性集和决策属性集时，若一个属性 $R\in P$ 是 $Q-$ 不必要的，则从 P 中去掉属性 R 不会改变原来信息系统的决策，而去掉 P 的 $Q-$ 核中的属性将改变原信息系统的决策。所以，P 的 $Q-$ 核中的属性对于决策来说是至关重要的。我们将选择相对核中的属性作为构造多变量检验的属性。

1.2 相对泛化的定义

如何用所选的属性来构造多变量检验呢？通过分析，我们知道用所选属性的简单合取作为多变量检验，可能会导致对数据的过拟合问题。为此,我们定义了两个等价关系在多数包含情况下的泛化的概念。

定义 1.设 P 和 Q 是 U 上的两个等价关系族，且 $U/P=\{X_1,X_2,.....,X_n\}$，$U/Q=\{Y_1,Y_2,.....,Y_m\}$

令

$$Z_i = \bigcup_{X_j\in U/P}\{X_j : X_j \subseteq^b Y_i\}, j=1,2,.....n, i=1,2,.....m. \quad (6)$$

$$Z_{m+1} = \bigcup_{X_j\in U/P}\{X_j : X_j \not\subset^b Y_i, j=1,2,.....n, \forall i\} \quad (7)$$

则称 $\{Z_1,Z_2,.....,Z_{m+1}\}$ 在 U 上确定的等价关系为 P 相对于 Q 的泛化，记作 $GEN_Q^b(P)$。

上述定义的合理性可通过如下命题说明。

命题 1. $\{Z_1,Z_2,.....,Z_{m+1}\}$ 构成了 U 上的一个划分，其中 Z_i，i=1,2,……,m+1,由(6)(7)定义。

证明：从 Z_i 的定义可知，$\bigcup_{i=1}^{m+1} Z_i = \bigcup_{i=1}^{n} X_i = U$。下证 $Z_i \cap Z_j = \varnothing, i\neq j, i,j=1,2,...,m+1$。由定义显然有 $Z_i \cap Z_{m+1} = \varnothing$，对任意的都成立。

对于的情况，用反证法证明。假设 $Z_i \cap Z_j \neq \varnothing$，$i\neq j, i,j=1,2,...,m$。则至少存在一个 $x\in U$，使得 $x\in Z_i \cap Z_j$。

$\Rightarrow x\in Z_i$ 且 $x\in Z_j$，从而有 $x\in Y_i$ 且 $x\in Y_j$

$\Rightarrow x\in Y_i \cap Y_j$，所以 $Y_i \cap Y_j \neq \varnothing$，$i\neq j$

这与 $\{Y_1,Y_2,...,Y_m\}$ 是 U 的划分矛盾！所以，$\{Z_1,Z_2,...,Z_{m+1}\}$ 构成了 U 上的一个划分。我们知道，U 上的划分与其上的等价关系是一一对应的。因此，这一划分唯一地确定了 U 上的一个等价关系。

相对泛化的概念将用于多变量检验。

2 基于VPRS的多变量决策树优化算法

决策树作为一种示例学习算法，是从标有类别信息的示例中，采用自顶向下的递归方式构造决策树的。在决策树的内部节点进行属性值的比较并根据不同的属性值判断从该节点向下的分支，在决策树的叶结点得到结论。

一个自顶向下的决策树算法是，首先根据某种划分度量准则选择最佳检验属性，然后，用选择的检验属性去划分训练集，并且，相应于该检验的每一个结果产生一个分支，这一过程递归的应用到该检验导出的每一个分类上，如果某一分类中的所有事例都来自一个类别，那么就产生一个标有该类别名的叶结点。在决策树的构造过程中，人们希望在每个节点上都选择能把事例划分到他们类中的最佳检验。

本文使用条件属性对应于决策属性的区分能力作为度量标准，也就是利用变精度粗糙集模型中条件属性 P 相对于决策属性 Q 的核的概念给出构造多变量决策树的步骤如下：

(1) 在变精度粗糙集模型的概念下删除冗余属性，进而计算条件属性集 P 相对于决策属性集 Q 的核，即 $CORE_Q^b(P)$。若 $CORE_Q^b(P)=j$，则转(2)；否则，不妨设 $CORE_Q^b(P)=\{a_1,a_2,.....,a_k\}$,转(3)。

(2) 用 ID3 的方法选择一个最佳属性，作为该节点的检验。

(3) 令 $P=a_1\wedge a_2\wedge.....\wedge a_3$，计算 P 相对于 Q 的泛化 $GEN_Q^b(P)$，将它作为该节点的检验。

3 实例分析

根据以上的讨论我们给出一个基于变精度粗糙集模型的多变量决策树构造方法，以表 1 中含有噪声数据的信息系统为训练集来进行决策树的构造。

表 1 中的噪声数据是我们在苗夺谦等的文章中给出的信息系统中加入三条实例数据 15、16 和 17 之后得到的，这三条数据在条件属性部分与实例数据 1 相同，15、16 的决策属性与 1 相同，而 17 的决策属性与 1、15、16 不同，也就是说，17 是一条噪声数据。在这 4 条数据中的正确率为 0.75，也就是说在

变精度粗糙集模型中的 $b \le 0.75$ 才能消除这条噪声数据对分类结果的影响。下文中我们选 $b = 0.7$ 来进行决策树的构造，其中 C 表示条件属性，D 表示决策属性，则有：

$U / IND(C) = \{\{1,15,16,17\},\{2\},\{3\},\{4\},\{5\},\{6\},\{7\},\{8\},\{9\},\{10\},\{11\},\{12\},\{13\},\{14\},\}$

$U / IND(D) = \{\{1,15,16,2,6,8,14\},\{3,4,5,7,9,10,11,12,13,17\}$

有 $POS^b_{IND(C)} = \{1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17\}$

表 1　含有噪声数据的信息系统

U	条件属性（C）				决策属性（D）
	Outlook (a_1)	Temperature (a_2)	Humidity (a_3)	Windy (a_4)	
1	Sunny	Hot	High	False	N
2	Sunny	Hot	High	True	N
3	Overcast	Hot	High	False	P
4	Rain	Mild	High	False	P
5	Rain	Cool	Normal	False	P
6	Rain	Cool	Normal	True	N
7	Overcast	Cool	Normal	True	P
8	Sunny	Mild	High	False	N
9	Sunny	Cool	Normal	False	P
10	Rain	Mild	Normal	False	P
11	Sunny	Mild	Normal	True	P
12	Overcast	Mild	High	True	P
13	Overcast	Cool	Normal	False	P
14	rain	mild	high	true	N
15	Sunny	Hot	High	False	N
16	Sunny	Hot	High	False	N
17	Sunny	Hot	High	False	P

考察 $a_i (i = 1,2,3,4)$ 在 C 中以 b 的正确率相对 D 来说是否必要。为此，从 C 中去掉 a_1，得到

$U / IND(C - \{a_1\}) = \{\{1,3,15,16,17\},\{2\},\{4,8\},\{5,9,13\},\{6,7\},\{10\},\{11\},\{12,14\}\}$

$POS^b_{IND(C-\{a_1\})}(D) = \{2,5,9,13,10,11\} \neq POS_{IND(C)}(D)$ 可知，a_1 在 C 中是必要的。

同理可知，a_2, a_3 在 C 中是不必要的，a_4 在 C 中是必要的。

因此，我们得到 $CORE^b_D(C) = \{a_1, a_4\}$。

把重要属性选出来之后，接下来要利用选择的属性构造多变量检验。首先，令 $P = a_1 \wedge a_4$，则有

$U / P = \{\{1,8,9,15,16,17\},\{2,11\},\{3,13\},\{4,5,10\},\{6,14\},\{7,12\}\}$

然后，可以算出 P 相对于 D 的泛化在 U 上导出的划分为 $\{\{6,14\},\{3,13,4,5,10,7,12\},\{1,8,9,15,16,17,2,11\}\}$ 由于划分与属性之间存在着一一对应关系，所以，这一划分唯一地确定了 U 上的一个新属性，即我们构造的多变量检验 GEN^b_D。

经过以上的建树过程得到如下的未分类数据，如表 2：

令

$P = \{a_2, a_3\}$　　$U / IND(P) = \{\{1,2,15,16,17\},\{8\},\{9\},\{11\}\}$

$U / IND(D) = \{\{1,2,8,15,16\},\{9,11,17\}\}$

由公式，有 $POS^b_{IND(P)} = \{1,2,8,9,11,15,16,17\}$

表 2　剩余的未分类数据

1	Hot	High	N
2	Hot	High	N
8	Mild	High	N
9	Cool	Normal	P
11	Mild	Normal	P
15	Hot	High	N
16	Hot	High	N
17	Hot	High	P

考察 $a_i (i = 2,3)$ 在 P 中相对 D 来说是否必要，同理可知是 a_2 不必要的，a_3 是必要的。因此，我们得到 $CORE^b_D(P) = \{a_3\}$，所以有由公式，有 $POS^b_{IND(P)} = \{1,2,8,9,11,15,16,17\}$

$U / \{a_3\} = \{\{1,2,8,15,16,17\},\{9,11\}\}$，进一步得出 $\{a_3\}$ 以 b 的正确率相对于 D 的泛化在 U 上导出的划分为 $\{\{1,2,8,15,16,17\},\{9,11\}\}$，从而得到了最终的决策树如图 1 所示：

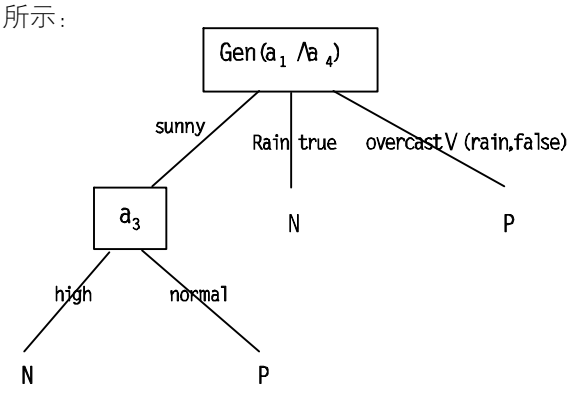

图 1　基于 VPRS 的多变量决策树

很明显，已经有效的消除了噪声数据对决策树分类的影响，从而解决了多变量决策树构造过程中噪声数据的干扰问题。可以看出与[3]中提出的多变量决策树构造算法相比采用 VPRS 构造多变量决策树拓宽了算法的适用范围，而[10]中的算法是一种单变量决策树构造算法，与本文中的多变量决策树算法解决问题不同。

4 结论

本文提出了一种基于 VPRS 的多变量决策树优化算法，我们利用变精度粗糙集理论允许一定程度的错误分类率存在的性质，消除了多变量决策树生成过程中噪声数据的影响，在此基础上构造了变精度粗糙集理论中条件属性集相对于决策属性集的核的概念，解决了噪声干扰情况下多变量检验中属性的选择问题，进而定义了两个等价关系在多数包含情况下的泛化的概念。利用这一概念，使得多变量检验并不是被选属性的简单合取，而是由其导出的一个新属性，由于使用了属性的不可区分性作为划分度量标准，这一方法避免了在决策树的一条路径上多次检验某一属性的问题。

参考文献

1 Cestnik B, Kononenko I, Bratko I. ASSISTAN T 86:a know ledge elicitation tool for sophisticated users. Proceedings of EW SL-87, 1987.31－45.

2 Pagallo G, Haussler D. Boolean feature discovery in empirical learning. Machine Learning, 1990,5:71－99.

3 苗夺谦,王珏.基于粗糙集的多变量决策树构造方法.软件学报, 1997,8(6):425－431.

4 Brodley CE, Utgoff P E. Multivariate decision trees. Machine Learning, 1995,19:45－77.

5 史忠植. 知识发现.清华大学出版社. 2002/1.

6 陈文伟.决策支持系统及开发（第二版）.清华大学出版社, 2000/2.

7 Quinlan JR. Induction of decision trees. Machine Learning,1986,1:81－106.

8 Michalski RS, Larson JB. Selection of the most representative training examples and incremental generation of VL 1 hypotheses. Rept. No. 782867, Urbana-Champaign: Department of Computer Science, University of Illinois, 1978.

9 Ziarko. Variable precision rough set model. Journal of Computer and System Sciences, 1993,46(1):39－59.

10 张瑞玲,都彦格,张克勇.基于VPRS的ID3算法改进.陕西理