面向连续值属性的模糊粗糙集决策树模型研究

张曙红[1,2], 张金隆[2], 孙建勋[2]

(1.中国地质大学 管理学院; 2.华中科技大学 管理学院, 湖北 武汉 430074)

摘　要: 决策树分类方法是一种应用广泛的分类算法, 但难于处理连续值属性的决策分析问题。给出了一种面向连续值属性的模糊粗糙集决策树分析方法。该方法基于模糊聚类对连续属性进行离散化, 并通过计算模糊隶属度矩阵中条件属性和类属性之间的模糊依赖性量度来确定属性的重要性和发现冗余属性, 进而构造基于粗糙集的决策树分析模型。

关键词: 决策树; 模糊聚类; 粗糙集; 分类

中图分类号: C939　　文献标识码: A　　文章编号: 1001-7348(2006)05-0161-02

0　前言

决策树归纳算法是一种应用广泛的分类算法, 具有不局限于领域知识、分类速度快、生成的规则易于理解等优点。特别是在Quilan提出了ID3算法[1]以后, 在机器学习、知识发现等领域得到了进一步应用及巨大的发展。ID3算法中, 采用信息熵的概念来判断属性的重点程度。根据划分前后信息熵的缩减, 即信息增益来选择测试属性。这种方法的缺点是, 算法倾向于选择取值较多的属性, 而取值较多的属性未必是重要属性。另外, 属性集中可能存在冗余, 需要在分类前排除掉, 使得到的规则更加简洁, 而ID3算法无法对此作出判断。在粗糙集理论中, 属性的重要性可以通过属性之间的依赖衡量, 通过这种方法还可能发现冗余属性。因此, 可以采用粗糙集的方法来选择测试属性, 从而构造决策树[2,3]。但在面向数据库的决策分析时, 由于数据库中经常包含连续属性, 而粗糙集方法只能处理离散属性决策表, 因此基于粗糙集的决策树分类方法具有一定的局限性。本文给出了一个基于模糊聚类的粗糙集决策树分类方法。该方法利用模糊聚类方法对连续属性进行离散化, 将连续值变换到不同的模糊子集, 得到一个模糊隶属度矩阵决策表, 并通过计算模糊隶属度矩阵中条件属性和类属性之间的模糊依赖性量度来确定属性的重要性和发现冗余属性, 进而构造决策树分析模型。

1　粗糙集决策分类方法

粗糙集(Rough Set)理论是由波兰数学家Z.Pawlak提出的一种算是不完整性、不确定性问题的数学工具[4,5]。它从新的角度定义知识, 把知识看作是关于等价类的划分, 从而把知识和分类紧密联系起来。粗糙集理论的主要特点是它仅利用数据本身提供的信息, 不需要其它先验知识, 目前已被广泛地应用于知识分类、模式识别、机器学习、图像处理等领域[6]。在粗糙集决策树分类方法中, 属性的重要性可以通过属性之间的依赖程度来衡量, 通过这种方法也可以发现冗余属性。

定义1　设 $P\subseteq R, P\neq\emptyset$, 则 IP(P中所有等价关系的交集)也是一个等价关系, 称为P上的不可区分关系, 词类 ind(P), 且有:

$$[x]_{ind(P)}=\mathop{I}_{R\in P}[x]_R$$

这样U/ind(P)就表示与等价关系族P相关的知识, 称为知识库中关于U的P基本知识。为简便起见, 可用U/P代替U/ind(P)。不可区分关系ind(P)的等价类称为知识的基本概念或基本范畴, 是构成知识的基本模块。

定义2　令 $X\subseteq U$, R为U上的一个等价关系, 称 $\underline{R}X=\{x\in U|[x]_R\subseteq U\}$ 和 $\overline{R}X=\{x\in U|[x]_R I X\neq\emptyset\}$ 为X的R下近似集和R上近似集。集合 $bn_R(X)=\overline{R}X-\underline{R}X$ 称为X的R边界域, $pos_R(X)=\underline{R}X$ 称为X的R正域, $neg_R(X)=\overline{R}X$ 称为X的R负域。

定义3　令P和Q为U中的等价关系, 称 $pos_P(Q)=\mathop{I}_{X\in U/Q}\underline{P}X$ 为Q的P正域。它表示论域中通过分类U/P表达的知识能确定地划入U/Q类的对象的集合。

定义4　若 $pos_{ind(P)}(ind(Q))=pos_{ind(P-\{r\})}(ind(Q))$, 则称 $r\in P$ 为P中Q可省略的, 否则称

收稿日期: 2005-09-15

基金项目: 湖北省教育厅重点科技基金(2004X014)

作者简介: 张曙红(1974-), 男, 副教授, 研究方向为决策支持系统与数据挖掘、系统工程; 张金隆, 男, 华中科技大学管理学院院长, 教授, 博士生导师。

为 P 中 Q 不可省略的。当 P 中每一个 r 都为 Q 不可省略的，则称 P 为 Q 独立的。当 S 为 P的 Q 独立子族，且 $pos_S(Q)=pos_P(Q)$，则称 S 为 P 的 Q 简化。P 中所有 Q 不可省略关系的集合称为 P 的 Q 核，记作：

$$core_Q(P)=I\ red_Q(P)$$

令 $r_P(Q)=card(pos_P(Q))/card(U)$，可将 $r_P(Q)$ 看作 P 和 Q 之间依赖性的量度。其中 card 表示集合的基数。对一个知识表达系统，其中的属性可看作一个等价类。不同的属性对分类的重要程度不同，可以通过前面定义的依赖性量度来衡量。具体地讲，令 C 为条件属性集，D 为分类属性集，a, b∈C，若：

$$r_C(D)-r_{C-a}(D)>r_C(D)-r_{C-b}(D)$$

则称属性 a 比属性 b 更重要。这是由于去掉 a 后，分类 U/D 的正域受到的影响更大。若某个属性是 C 中 D 可省略的，即去掉该属性后，分类的正域不变，则称该属性为冗余属性，可在分类前去除。

由以上的讨论可以看出，基于粗糙集构造决策树的基本方法是：首先通过检查冗余属性，得到属性集的简化，然后根据属性的重要程度选择测试属性，形成决策树的分支，依次递归构造出决策树。

2 面向连续值属性的粗糙集决策树分类方法

由于数据库中经常包含连续属性，而粗糙集方法只能处理离散属性，因此可利用模糊聚类方法，首先对属性进行离散化。此时得到的是一个模糊隶属度矩阵，其中每一个属性都对应着一个模糊集合族。由于传统粗糙集分类方法是面向普通集合的，因此需要对其进行改造，以适合模糊集合的情况。

给定论域 U，模糊集合 A，截取阈值 λ，可以得到 A 的一个 λ 截集 A_λ：

$$A_\lambda=\{u\in U|A(u)\geq\lambda\}$$

式中 A_λ 是普通集合。

粗糙集是基于等价类和不可区分关系对论域进行划分的。在知识表中，每个等价类对应一个属性。当属性取值为模糊集合时，可将每个属性看作一个模糊等价关系。$\forall R\in A$，则 R_λ 可认为是 U 上的普通等价关系。

定义 5　给定论域 U，U 上的一个模糊等价关系 R 和 R_λ，A 是 U 的一个模糊集合，则 A 关于 (U, R_λ) 的一对下近似 $\underline{R}_\lambda A$ 和上近似 $\overline{R}_\lambda A$ 定义为 U 上的一对模糊集合，其隶属函数分别定义为：

$$\underline{R}_\lambda A(x)=\inf\{A(y)|y\in[x]_{R_\lambda}\},\ x\in U,$$

$$\overline{R}_\lambda A(x)=\sup\{A(y)|y\in[x]_{R_\lambda}\},\ x\in U。$$

记

$$pos_{R_\lambda}(A)=\sum_{x\in U}\underline{R}_\lambda A(x)$$

$$pos_{P_\lambda}(Q_\lambda)=\sum_{x\in U_\lambda}pos_{R_\lambda}(X)$$

P_λ 和 Q_λ 之间的依赖性量度可表示为：

$$r_{P_\lambda}(Q_\lambda)=pos_{P_\lambda}(Q_\lambda)/pos_{P_\lambda}(U)$$

根据以上概念，就可以对由模糊隶属度矩阵表达的知识表进行属性化简和选择，从而构造出决策树，算法基本过程如下：

(1) 对连续属性知识表进行模糊聚类划分，将原始数据转换成模糊隶属度矩阵；

(2) 对隶属度矩阵进行模糊等价类划分，计算各个条件属性和类属性之间的依赖性量度；

(3) 求出属性集的简化和核，从一个简化开始，构造决策树；

(4) 选择其中最重要的属性作为根结点；

(5) 对属性的每个模糊概念形成一个分支，对样本进行划分；

(6) 递归地对每个分支进行以上过程，从而构造出整个决策树模型。

3 实例研究

表 1 中的数据来自 UCI 数据库的汽车知识 Auto 测试数据集。其中有 3 个汽车连续值属性：mpg(英里/加伦)，horsepower(马力)，weight(重量)，分别简记为 m, h 和 w。现在要从中导出分类规则，构造决策树分析模型。

表 1　Auto 测试集部分汽车知识数据

No.	mpg	horsepower	weight
1	24	95	2 372
2	22	95	2 833
3	18	97	2 774
4	21	85	2 587
5	28	90	2 123
6	27	88	2 130
7	25	87	2 672
8	24	90	2 430

首先利用模糊聚类 PCM 算法[7]对连续属性进行模糊聚类划分。本例取聚类中心个数 C=2；表 2 是经过模糊聚类后得到的汽车知识表隶属度矩阵。

取阈值 λ=0.5，记条件属性集为 C，分类属性集为 D，可得：

表 2　汽车知识表的隶属度矩阵

No.	mpg		horsepower		weight		class	
	少	多	小	大	轻	重	1	2
1	0.1	0.9	0.4	0.6	0.8	0.2	0.8	0.2
2	0.8	0.2	0.4	0.6	0.1	0.9	0.2	0.8
3	0.9	0.1	0.1	0.9	0	1	0.1	0.9
4	1	0	0.9	0.1	0.1	0.9	0.1	0.9
5	0.1	0.9	1	0	1	0	1	0
6	0	1	1	0	1	0	1	0
7	0	1	1	0	0	1	1	0
8	0.1	0.9	1	0	0.6	0.4	0.7	0.3

$U/C_\lambda=\{\{1\},\{2,3\},\{4\},\{5,6,8\},\{7\}\}$;

$U/D_\lambda=\{\{1,5,6,7,8\},\{2,3,4\}\}$;

$pos_{C_\lambda}(U)=0.6+0.6\times2+0.9+0.6\times3+1=5.5$;

$pos_{C_\lambda}(D_\lambda)=0.6+0.6\times2+0.9+0.6\times3+1=5.5$;

可根据定义 5 计算依赖性量度：

$$r_{C_\lambda}(Q_\lambda)=5.5/5.5=1$$

同样可得依赖性量度 $r_{(C-\{m\})_\lambda}(Q_\lambda)=0.67$, $r_{(C-\{h\})_\lambda}(Q_\lambda)=1$, $r_{(C-\{w\})_\lambda}(Q_\lambda)=1$, $r_{(C-\{h,w\})_\lambda}(Q_\lambda)=1$。

可见，只有属性 m 是重要的，而 h 和 w 都是冗余属性，属性集的简化和核都是 m，因此只需利用属性就可对原数据集进行分类，建立决策树，从而大大简化了原知识表。得到的分类规则集只包含两条规则：①if "mpg=少", then "class=2"；②if "mpg=多", then "class=1"。

如果选择不同的模糊子集个数和截取阈值 λ，也会得到不同的决策树和决策规则集。

参考文献：

[1] Quinlan J R. Induction on Decision Trees. Machine Learning,1986,(1):81-106.

[2] 张文修，吴伟志等.粗糙集理论与方法[M].北京：科学出版社，2001.

[3] Pawlak.Z. Rough Sets and Decision Tables. Lecture Notes in Computer Science 1985.

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

[5] Pawlak.Z. Rough Sets:Theoretical Aspects of Reasoning about Data. Kluwer Academic Publishers, Boston,1991

[6] 曾黄麟.粗集理论及其应用[M].重庆：重庆大学出版社，1996.

[7] J.Bezdek. Pattern Recognition with Fuzzy Objective Function Algorithms, Plenum Press, New York,1981.

（责任编辑：汪智勇）