

一种多变量决策树方法研究^{*})

梁道雷^{1,2} 黄国兴¹ 金 健¹

(华东师范大学计算机系 上海 200062)¹ (浙江理工大学先进纺织材料教育部重点实验室 杭州 310018)²

摘 要 单变量的决策树算法造成树的规模庞大,规则复杂,不易理解。本文结合粗糙集原理中的相对核及加权粗糙度的方法,提出了一种新的多变量决策树算法。通过实例表明,本文的多变量决策树方法产生的决策树比传统的 ID3 算法构造的决策树更简单,具有较好的分类效果。

关键词 多变量决策树,粗糙集,相对核,加权粗糙度

A New Multivariate Decision Tree Algorithm

LIANG Dao-Lei^{1,2} HUANG Guo-Xing¹ JIN Jian¹

(Computer Department, East China Normal University, Shanghai 200062)¹ (Zhejiang Sci-Tech University, Hangzhou 310018)²

Abstract Decision Tree Algorithm in univariate tests caused large-scale, complex rules that are difficult to understand. Based on the rough sets theory of attributes reduction, the core of condition attributions and the weighted roughness of condition attributions, a new multivariate decision tree algorithm is proposed. An example shows in this paper, the decision tree built by the method is more simple and has better classification result than that of ID3 algorithm.

Keywords Multivariate decision tree, Rough sets, Relative core of attributes, Weighted roughness

决策树是一种机器学习方法,因为具有分类规则容易理解、速度快、不需要背景知识等优点而广泛使用。该学习方法是产生一棵树,所以树的规模越小,知识越简洁,越容易理解和使用。目前的大多数决策树算法都是单变量决策树,也就是在每个结点只对单个属性进行检验,著名的算法有 ID3, C4.5, ASSISTANT。然而,单变量决策树算法的局限是很容易造成有些属性在一棵树中某一路径被多次检验和一棵决策树中的子树的重复。

为了克服这一局限,人们提出了多变量归纳学习方法,即在决策树的某一结点上可以同时检验多个属性,该方式产生新的、更相关的属性,以及修改或去掉初始提供的不相关的属性。目前对于多变量决策树算法有 Pagallo G^[1] 等人的 RLS 方法,它是通过属性 Boolean 组合来减少决策树的大小。方法为树根到该叶结点路径上的最后 2 个属性或其否定形式的合取构成,通过迭代形成有意义的多变量检验,改进了决策树的准确性,但其分析非常复杂。Brodley C E^[2] 等人采用线形组合的方式构造多属性组合,但要属性进行数值化处理。前两者都只适合 2 个决策类的情况。Rivest R^[3] 提供了一种改进方法,但其事先人为地规定了结点所包含的属性数量,不具有客观性。国内的苗夺谦^[4] 等采用粗糙集的核与信息熵集合的方法产生多条件决策树,并取得很好的效果,但决策树的规模仍较大。当该数据集没有核属性的条件下,其树的结果形式和 ID3 完全一样。

本文在研究多变量决策树时,主要利用了粗糙集的理论中相对核的概念以及加权粗糙度的概念,分析了初始属性的选择,然后使用了 2 个等价关系相对泛化的原理。本文第 1 节介绍本方法的处理过程,第 2 节介绍粗糙集原理,第 3 节利用粗糙集的相对核和加权粗糙度来选择组合属性,并进行等价关系相对泛化,产生树的分支,以及算法的描述,第 4 节通过例子与 ID3 算法进行比较。最后是本文的结论及进一步的

工作探讨。

1 本方法的处理过程

由于本研究算法主要是利用粗糙集的原理,对于连续型数据处理会造成属性值过多,计算非常复杂,因此事先将连续型数据进行离散化处理;对于数据样本的属性值非常多,可以对一些对决策值无关联的属性进行简约;找出样本数据的核属性,进行等价关系相对泛化;求出其他属性的加权粗糙度,选择结点属性,构成决策树。

2 粗糙集理论

采用粗糙集理论作为研究知识发现的工具具有许多优点。粗糙集理论将知识定义为不可区分关系的一个族集,这使得知识具有了一种清晰的数学意义,并可使用数学方法进行处理。

定义 1 设 $K=(U, AT, V, \rho)$ 是一个信息系统,我们记由属性集 $B \subseteq AT$ 所导出的等价关系为 R_B 。若 $R_{AT} = R_{AT \setminus \{a\}}$, 则称属性 a 是多余的;若在系统中没有多余属性,则称 AT 是独立的;子集 $B \subseteq AT$ 称为是 AT 的约简,常记作 $red(AT)$, 若 $R_B = R_{AT}$ 且 B 中没有多余属性; AT 的所有约简的交集称为 AT 的核,记作 $core(AT)$, 一般属性的约简不唯一而核是唯一的^[5]。

定义 2 (下近似和上近似) 对信息系统 $S=(U, A, V, f)$, 设 $B \subseteq A, X \subseteq U$, 则 $Ind(B)$ 是 $U \times U$ 上的等价关系, $B(x_i)$ 是按等价关系 $Ind(B)$ 得到的包含的等价类,称 $B(x_i)$ 为 B 的基本集。用属性 B 对 U 进行划分,获得的是一个等价类集。子集 X 的下近似集 $\underline{B}(x)$ 和上近似集 $\overline{B}(x)$ 分别定义如下:

$$\underline{B}(x) = \{x_i \in U \mid B(x_i) \subseteq X\}$$

$$\overline{B}(x) = \{x_i \in U \mid B(x_i) \cap X \neq \emptyset\}$$

定义 3 设 P 和 Q 是 U 上的 2 个等价关系族,令 $U/IND(p) = \{X_1, X_2, \dots, X_s\}$

^{*}) 华东师范大学 211 重点项目(521B0108)。梁道雷 博士生,讲师,研究方向:知识发现与数据挖掘、纺织品 CAD;黄国兴 教授,博士生导师,研究方向:智能信息系统、数据挖掘、知识库与专家系统等;金 健 博士生,研究方向:知识发现与数据挖掘。

$$U/IND(Q)\{Y_1, Y_2, \dots, Y_3\}$$

$$Z_i = \bigcup_{x_j \in IND(P)} (X_i; X_j \subset Y_j), i=1, 2, \dots, m$$

$$Z_{m+1} = \bigcup_{x_j \in IND(P)} (X_j; X_i \notin Y_i, \forall i)$$

则称 $\{Z_1, Z_2, \dots, Z_{m+1}\}$ 在 U 上确定的等价关系为 P 相对于 Q 的泛化,记作 $GEN_Q(p)^{[6]}$ 。

3 算法与实例

定义 4 对决策系统 $S=(U, A, V, f), X \subseteq U, B \subseteq A, \mu_B(X) = \text{card}(\overline{B(X)}) / \text{card}(B(X))$ 是 X 在空间 B 的精确度, $0 \leq \mu_B(X) \leq 1, X$ 关于 B 的加权粗糙度定义为

$$\beta_B(i) = 1 - (\sum_{j=1}^m \omega_j \mu_B(X_j))$$

其中 i 表示第 i 个条件属性; j 表示决策属性的第 j 个等价类。 $J=1, \dots, m$, 其中 m 是决策属性等价类的数目; 表示决策属性第 j 个等价类集合, $X_j \subseteq U; \omega_j$ 表示 X_j 在整个论域 U 上所占的比例, $\omega_j = \text{card}(X_j) / \text{card}(U)^{[7]}$ 。

由定义 1 构造可辨识矩阵, 获取当前数据集的核属性, 定义 2 和定义 3 进行等价关系相对泛化, 定义 4 给出加权粗糙度的计算方法, 由以上定义形成加权粗糙集决策树算法。

加权粗糙集决策树算法 WR 描述:

输入: Q 是数据集

Att 是属性集合

输出: 决策树

WR(Q, Att)

1) 使用 Naive 离散算法对数据集离散化处理。

2) 计算条件属性 C 相对于决策属性 D 的核, 记作 $\text{CORED} = \{a_1, a_2, \dots, a_k\}$, 若 $\text{CORED}(C)$ 为空集, 转步骤 4, 否则继续步骤 3。

3) 令 $P = a_1 \cap a_2 \cap \dots \cap a_k$, 计算 P 相对决策属性的泛化 $\text{GEND}(P)$, 作为决策树根结点。

4) 计算剩余条件属性相对于决策属性每个划分集合的上、下近似, 求每一个条件属性的加权平均粗糙度, 选择加权平均粗糙度最小的属性 T 作为当前属性。

5) 如果没有到达叶子结点, 就继续调用 $\text{WR}(Q, \text{Att}-P-\{T\})$ 。

6) 返回。

	A	B	C	D	E	F	Dec
1	0	0	0	0	0	0	2
2	1	1	0	0	0	0	1
3	2	0	0	0	0	0	1
4	2	0	0	0	0	0	1
5	1	1	1	1	1	0	2
6	0	0	0	0	0	0	2
7	0	0	0	0	0	0	2
8	1	0	0	0	0	0	2
9	2	1	0	0	0	0	2
10	1	2	0	0	0	1	2
11	0	0	0	0	0	1	2
12	2	0	0	0	0	1	1
13	2	2	0	0	1	1	2
14	0	2	0	0	0	0	2
15	1	0	0	0	0	1	2

图 1 liver-disorders 数据离散结果

实验例子以及与 ID3 的比较

本文以 UCI 中的 liver-disorders 测试数据做为实验样本, 从其样本中随机提取 15 行数据做为训练样本, 该数据为连续型数据, 先将其数据离散化处理, 结果为图 1。

根据可辨识矩阵获取该样本的核为 A 和 B , 以这两个属性的合取做为决策树的根结点, 并可以算出 $A \cap B$ 相对于 Dec 的泛化在 U 上导出的划分为 $\{\{2, 5, 9, 10, 13\} \{1, 6, 7, 8, 11, 14, 15\}, \{3, 4, 12\}\}$, 再求出属性 C, D, E, F 的平均粗糙度, $E=0.8666 D=0.9333 C=0.9333 F=1$ 平均粗糙度大小从大到小排列为: E, D, C, F, A 为下一个条件选择属性, 根据算法依次构造决策树, 如图 2 所示。

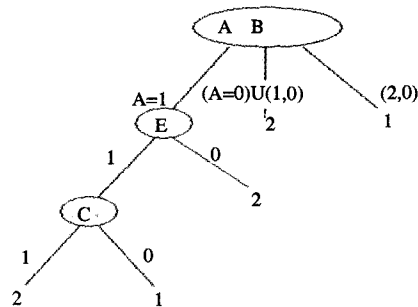


图 2 加权粗糙集决策树

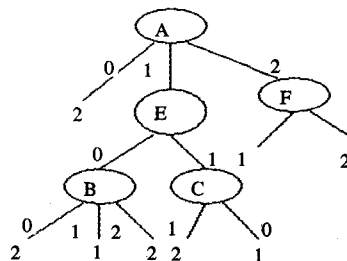


图 3 ID3 算法决策树

图 3 为 ID3 算法根据信息熵作为属性选择的依据产生的决策树, 图中有树叶树为 8。而同样的样本的前提下, 加权粗糙集的树叶树为 5, 规则简单了很多。将 liver-disorders 数据样本共 345 条记录做为测试样本, 加权粗糙集方法的正确率为 81%, ID3 在树层数为 3 的正确率为 78.5%, 所以加权粗糙集决策树方法具有一定的价值。

结论及进一步研究方向 本文基于粗糙集的原理构建多属性决策树, 该方法对原数据集进行属性简约, 按照核属性及加权粗糙度的大小来选择条件属性作为树结点, 通过实验表明该方法比 ID3 算法产生的树结点少, 规则简单, 准确率较高的。但本算法在考虑除核属性以外的条件属性时, 都是单属性来检验, 所以进一步工作方向为考虑除核属性以外的条件属性的相关性, 在一定的相关度阈值的情况下, 对其属性进行组合, 使得产生的多变量决策树更小, 规则更容易理解。

参考文献

- Pagallo G, Haussler D. Boolean feature discovery in empirical learning [J]. Machine Learning, 1990, 5: 71~99
- Brodley C E, Utgoff P E. Multivariate decision trees [J]. Machine Learning, 1995, 19: 45~77
- Rivest R. Learning decision lists [J]. Machine Learning, 1987, 2: 229~246
- 苗夺谦, 王珏. 基于粗糙集的多变量决策树的构造方法 [J]. 软件学报, 1997, 8(6): 425~431
- 沙慧新, 叶东毅. 基于知识粗糙度的多变量决策树的构建 [J]. 福州大学学报(自然科学版), 2004, 32(2): 138~141
- 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
- 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