第 20 卷 第 3 期
2000 年 9 月

河北大学学报(自然科学版)
Journal of Hebei University

Vol. 20 No. 3
Sep. 2000

决策树与模糊决策树的比较

黄冬梅[1],哈明虎[2],王熙照[2]
(1. 河北农业大学 基础科学系,河北 保定 071001; 2. 河北大学 数学系,河北 保定 071002)

摘 要: 对决策树与模糊决策树的异同进行了比较分析. 模糊决策树是决策树在模糊环境下的一种推广,它作为一种知识表示形式更符合人类的思维.

关键词: 决策树; 模糊决策树; ID3 算法; 模糊 ID3 算法
中图分类号: O 159 **文献标识码**: A **文章编号**: 1000 - 1565(2000)03 - 0218 - 04

决策树归纳学习是机器学习领域中最重要的内容之一,可用于知识的自动获取过程. 通过决策树归纳学习产生规则,是知识获取过程中最为常用而有效的方法,是建立专家系统的一个有效的途径,同时在专家系统的发展中被视为"瓶颈". 随着决策树归纳学习的深入研究,已经产生了许多构造决策树的方法[1],尽管由这些方法产生的决策树在构造基于知识的专家系统中是有用的,但是常常经受与人的思维与感知密切相联的认识上的不确定性的困扰,因而具有精确描述特征的决策树归纳学习已经不能适应一个系统中不精确知识自动获取的要求,为了在不确定(模糊)环境下达到不精确知识自动获取的需要,对模糊决策树归纳学习的研究已经成为当前的一个研究热点. 目前已出现了一些模糊决策树归纳学习方法,并开始应用于实际. 本文将针对决策树与模糊决策树及它们各自的代表算法进行分析比较.

1 决策树与模糊决策树的描述

决策树归纳学习与模糊决策树归纳学习分别是归纳学习与模糊归纳学习的一个重要分支, ID3 是决策树学习算法的代表,模糊 ID3 和 Min - U 是模糊决策树学习算法的代表.

1.1 决策树的一般概念描述[2]

设 E 是一组分类例子集合; A 为一组描述例子特征(属性)的集合; $T(E)$ 是一个决策树归纳的停止准则, $IDM(A_i, E)$ 是一个评价函数, $A_i \in A$;决策树的描述过程如下:

如果 E 满足停止准则 $T(E)$,
那么 返回该树的叶结点,标记 E 中的最多数目的例子类别;
否则 选择一个特征 $A_{best} \in A$ 使得 $IDM(A_{best}, E)$ 的值最大;
对特征 A_{best} 的每一个取值 A_j,递归生成子树 $T(E_j)$,(其中 E_j 由 E 中特征 A_{best} 取值为 V_j 的例子组成)返回一个非叶结点,标记分类特征为 A_{best};
结束.

从决策树的描述发现,建立决策树的核心在于评价函数 $IDM(A_i, E)$ 和停止准则 $T(E)$ 如何确定.

收稿日期: 1999 - 12 - 10
基金项目: 河北省自然科学基金资助项目(698139); 河北省科委博士基金资助项目(97543306)
作者简介: 黄冬梅(1963—),女,河北保定人,河北农业大学副教授,主要从事模糊数学方面的研究.

1.2 模糊决策树的描述[3]

模糊决策树是决策树的一种推广.

设$\Omega_i \subseteq F(U)$　$(1 \leq i \leq m)$是给定的 m 族模糊子集且满足 $M(\Omega_i) > 1$. 如果一个有向树满足:

1) 树中的每个结点属于 $F(U)$,

2) 对树中每一个非叶结点 N,它的所有子结点将是 $F(U)$ 的一个子集族,记为 Γ,则存在 $i(1 \leq i \leq m)$,使得 $\Gamma = \Omega_i \cap N$,

3) 每一叶结点对应一个或多个分类决策值.

则称其为一个模糊决策树, 每一组模糊子集Ω_i 对应于一个属性而Ω_i 中的每一个模糊子集对应于该属性的一个值.

2 决策树与模糊决策树的比较分析

下面主要对决策树与模糊决策树的异同点进行比较.

1) 属性值及分类. 在决策树中每一示例的属性值及分类都分别取互斥的属性值及分类中的一个,属性值及分类均是明确的, 它们是示例空间上的集合, 代表示例的属性及分类是否属于该属性值及要学习的概念. 在模糊决策树中,属性值或分类反映了与人的思维、认识过程和理解中密切相关的不确定性(模糊性), 所以将它们看成示例空间上的模糊集合, 代表着示例的属性及分类隶属于该属性值及要学习的概念的程度. 特别地情况, 属性值为示例空间上的模糊数. 模糊决策树的归纳学习与决策树归纳学习相比较, 由于合理地处理了不确定信息、噪音数据等问题, 从而有较强的分类能力及稳健性, 使知识表示的方式更为自然,为决策者提供了丰富的决策信息.

2) 扩展结点. 在决策树归纳学习中,因为属性值是互斥的,因此在对某一结点进行扩展(分枝)时,按扩展属性的属性值可将该结点上的例子集合 S 划分成 v 个不相交的子集合 $S_1, S_2, \cdots S_v$. 这些子集合满足 $\bigcup_{i=1}^{v} S_i = S, S_i \cap S_j = \quad (i \neq j)$. 其中 S_i 表示扩展属性的值 V_j 所对应的例子集合 $(i = 1, 2, \cdots v)$, v 是扩展属性的值的个数. 在模糊决策树归纳学习中,因为属性值为示例空间上的模糊集合,根据模糊集合的特点,当对某一结点按其扩展属性所对应的属性值进行扩展(分枝)时,$S_1, S_2, \cdots, S_v$ 构成 S 的一个模糊分割,即每一 S_i 为一模糊集合且 $\bigcup_{i=1}^{v} S_i \subseteq S$, $S_i \subseteq S(i = 1, 2, \cdots, v)$. 模糊决策树在知识的归纳过程中并入了认识上的不确定性, 使归纳出的知识在容许不精确或冲突的信息方面更稳健.

3) 匹配规则及推理. 在决策树中,每一条从根到叶的路径对应一条规则,该规则的真实度为 1,而且测试例子仅与一条规则相匹配,推理是基于清晰逻辑的;在模糊决策树中,每一条从根到叶的路径对应一条模糊规则,该规则的真实度小于等于 1,而且测试例子可与多条模糊规则相匹配,并且带有一定的隶属度,推理是基于模糊逻辑的. 由此可看出,模糊决策树与决策树相比较,模糊决策树更贴近于自然,更符合人类的思维. 从决策的角度看,它提供的知识更为合理.

4) 叶结点及规则的真实度. 在决策树中,叶结点仅包含一类的例子,相应的规则的真实度为 1;在模糊决策树中,叶结点包含不止一类的例子,相应规则的真实度小于等于 1,从而为构造模糊专家系统提供了更有效的途径.

5) 训练示例. 对训练示例,决策树的测试精度为 100 %;而模糊决策树的测试精度非 100 %.

3 ID3 算法与模糊 ID3 算法的比较分析

决策树归纳学习算法以 ID3 为代表, ID3 学习算法采用分治策略,在决策树的递归构造过程中,在树的结点上利用特征的信息增益大小作为分枝属性选择的启发式函数, 选择信息增益最大的特征作为分枝属

性,ID3 具有描述简单,分类速度快的优点,适合于大规模数据的处理,被广泛应用于模式识别,专家系统等领域. 下面是 ID3 的算法描述.

设 $E = D_1 \times D_2 \times \cdots \times D_n$ 是 n 维有穷向量空间,其中 D_j 是有穷离散符号集, E 中的元素 $e = < v_1, v_2 \cdots, v_n >$ 叫做例子,其中 $v_j \in D_j, j = 1, 2, \cdots, n$. 设 PE 和 NE 是 E 的两个例子集,分别做正例集和反例集,假设向量空间 E 中的正例集 PE 和反例集 NE 的大小分别为 p 和 n,ID3 基于下列两个假设:

1) 在向量空间 E 上的一棵正确决策树对任意例子的分类概率同 E 中正、反例的概率一致;

2) 一棵决策树能对一例子作出正确类别判断所需的期望信息比特为:

$$I(p, n) = -\frac{p}{p+n}\log_2\frac{p}{p+n} - \frac{n}{p+n}\log_2\frac{n}{p+n}.$$

如果以属性 A 作为决策树的根,A 具有 v 个值 $\{v_1, v_2 \cdots v_v\}$,它将 E 分成 v 个子集 $\{E_1, E_2 \cdots E_v\}$,假设 E_i 中含 p_i 个正例和 n_i 个反例,那么子集 E_i 所需的期望信息为 $I(p_i, n_i)$,以属性 A 为根所需要的期望信息为:

$$E(A) = \sum_{i=1}^{v}\frac{p_i + n_i}{p+n} I(p_i, n_i).$$

因此,以 A 为根的信息增益是 $\mathrm{Gain}(A) = I(p, n) - E(A)$.

ID3 选择使 $\mathrm{Gain}(A)$ 达最大的属性 A 作为扩展(分枝)属性. 容易将 ID3 扩展到多类样本的分类问题. ID3 的一个优点是它的归纳时间是和所给任务的困难度(如例子个数,用来描述对象的属性个数,研究概念的复杂度即决策树的结点数)成线性递增关系[4].

模糊 ID3 算法是决策树归纳学习算法 ID3 的一种推广. 模糊 ID3 算法亦采用分治策略,在决策树的递归构造过程中,在树的结点上利用属性的分割模糊熵作为扩展(分枝)属性的启发式函数,选择分割模糊熵最小的属性作为扩展属性,它具有分类速度快的优点,适合处理模糊数据,可应用于模糊专家系统等领域. 下面是模糊 ID3 算法的描述.

设训练例子集为 $E = \{1, 2, \cdots, N\}$,共有 m 个属性 $A_1, A_2 \cdots, A_m$,每个属性 A_i 的值域分别为 $\mathrm{Range}(A_i) = \{T_{i1}, T_{i2}, \cdots, T_{ik_i}\}$ $(i = 1, 2, \cdots, m)$. 第 j 个例子关于第 i 个属性的取值为定义在 $\mathrm{Range}(A_i)$ 上的一个模糊子集(模糊向量),每一个 $T_{ij}(1 \leq i \leq m, 1 \leq j \leq k_i)$ 是定义于 E 上的一个模糊子集. 设有两类 $\{P, N\}$ (P 类和 N 类集分别记作 P 和 N),则每一例子可表示为一个 $\sum_{j=1}^{m} k_i + 1$ 维的向量:

$$(r_{11}, \cdots, r_{1k_1}, r_{21}, \cdots, r_{2k_2}, \cdots, r_{m1}, \cdots, r_{mk_m}, P \text{ 或 } N).$$

给定一个判别叶子的标准 $\delta(0 < \delta < 1)$,递归地考虑非空结点 D.

1) 计算 $f_P(D) = M(D \cap P)/M(D)$ 和 $f_N(D) = M(D \cap N)/M(D)$,如果有一个超过 δ,则表示为叶子.

2) 在结点 D 上计算每一属性的分割模糊熵 $FE(D, A_i)$.

$$FE(D, A_i) = \sum_{j=1}^{k_i}\frac{M(D \cap T_{ij})}{m^*} E(D \cap T_{ij}),$$

其中, $m^* = \sum_j M(D \cap T_{ij})$, $E(D \cap T_{ij}), = -\frac{a}{a+b}\log_2\frac{a}{a+b} - \frac{b}{a+b}\log_2\frac{b}{a+b}$,

而 $a = M(P \cap D \cap T_{ij})$, $b = M(N \cap D \cap T_{ij})$,

筛选出最小者对应的属性作为该结点上的扩展属性,根据其属性值进行分枝扩展.

从模糊 ID3 算法的描述发现,模糊 ID3 算法的核心仍是评价函数和停止准则的选取,模糊 ID3 算法选取分割模糊熵为最小的属性作为扩展属性. 容易将模糊 ID3 扩展到多类样本的分类学习问题.

下面给出决策树归纳学习算法 ID3 与模糊决策树归纳学习算法模糊 ID3 的比较.

1) 在决策树归纳学习算法 ID3 中,所使用的评价函数为信息增益 $\mathrm{Gain}(*)$ (选取使 $\mathrm{Gain}(A)$ 达最大的属性 A 为扩展属性),这种方法使生成的决策树平均深度较小,从而有较快的分类速度;而在模糊决策树归纳

学习算法模糊 ID3 中,所使用的评价函数为分割的模糊熵 FE(*) (使 FE(A) 达最小者所对应的属性 A 为扩展属性),这种方法同样是使生成的模糊决策树的平均深度较小,有较快的分类速度.

2) 设 d 表示属性的个数, b 表示一个属性的最大的可能值的个数, N 表示训练例子的个数, 对 ID3 算法,构造决策树时根据 Gain(*) 计算的数目选取属性的费用,称为构造决策树的最坏情况的费用[5],为

$$\sum_{i=2}^{d} i \cdot b^{d-i} = \frac{2b^{d} - b^{d-1} - d + 1}{(b-1)^{2}} b + d = O(b^{d})$$

对模糊 ID3 算法,构造模糊决策树的最坏的费用亦为 $O(b^{d})$,这说明两种算法复杂性相当.

3) 决策树归纳学习算法 ID3 与模糊决策树归纳学习算法模糊 ID3 都采用分治策略. 模糊 ID3 算法是 ID3 算法在模糊环境下的直接推广,当算法中的隶属度仅取 0 或 1 时,模糊 ID3 算法退化为 ID3 算法.

参 考 文 献:

[1] SAFAVIAN S R, LANDGREBE D. A survey of decision tree classifier methodology[J]. IEEE Trans On Systems Man and Cybernetics, 1991, 21(3): 660 - 674.

[2] 钱国良. 归纳学习算法研究与应用[D]. 哈尔滨:哈尔滨工业大学, 1998.

[3] 王熙照. 模糊示例学习研究[D]. 哈尔滨:哈尔滨工业大学, 1998.

[4] 何钦铭, 王申康. 机器学习与知识获取[M]. 杭州:浙江大学出版社, 1997.

[5] UTGPFF P E. Incremental induction of decision trees[J]. Machine Learning, 1989(4): 161 - 186.

Comparision between Decision Tree and Fuzzy Decision Tree

HUANG Dong-mei[1], HA Ming-hu[2], WANG Xi-zhao[2]

(1. Department of Basic Science, Hebei Agricultural University, Baoding 071001, China;
2. Department of Mathematics, Hebei University, Baoding 071002, China)

Abstract: The difference and similarities between decision tree and fuzzy decision tree are analyzed. Fuzzy decision tree is the generalization of decision tree in fuzzy environment, and the knowledge represented by fuzzy decision tree is more natural to the way of human thinking.

Key words: decision tree; fuzzy decision tree; ID3 algorithm; fuzzy ID3 algorithm

(责任编辑: 傅爱民)