文章编号: 1003- 207(2004) 04- 0144- 05

利用上凸函数对决策树算法的改进

高学东, 尹阿东, 张 健, 宫 雨, 武 森

(北京科技大学管理学院, 北京 100083)

摘 要: 针对决策树分类方法的计算效率进行深入研究, 根据信息增益计算的特点, 引入了上凸函数的概念, 用于提高决策树分类过程中信息增益的计算效率。利用我们所提出的“一致性定理”和“特殊一致性定理”, 从理论上证明了利用上凸函数对信息增益计算进行改进后, 构造的决策树与原决策树具有相同的分类准确率。同时我们通过对大数据集的实验, 发现在相同规模的数据集下, 改进后的决策树算法比原算法有更高的计算效率, 并且这种计算效率的提高有随着数据集规模的增加而增加的趋势。

关键词: 决策树; ID3 算法; 上凸函数; 信息熵

中图分类号: TP18 文献标识码: A

1 引言

目前在数据挖掘领域中, 存在许多解决分类问题的模型, 诸如神经网络[1]、遗传算法[2]、模糊数学[3]、贝叶斯分类[4]、粗糙集[5]等方法。分类知识发现领域中这些方法的使用都已经取得了令人满意的效果[6,7,8], 但使用最为广泛的分类模型仍然是决策树算法[9,10,11]。决策树分类模型之所以被广泛使用, 主要有以下几个方面的原因:

(1) 与神经网络或贝叶斯发类等其他分类模型相比, 决策树的分类原理简单易懂, 很容易被使用人员理解和接受。

(2) 在决策树分类过程中, 不需要人为设定任何参数, 更适合于知识发现的要求。

(3) 决策树分类方法不需要任何除训练数据集和测试数据集以外的附加信息, 保证决策树与其它分类方法相比具有更高的分在速度。

(4) 决策树分类方法与其它分类模型相比, 具有非常好的分类准确率。

2 决策树算法的改进

2.1 ID3 决策树改进算法的原理

ID3 决策树算法的核心思想是利用信息熵原理, 选择信息增益最大的属性作为分类属性, 递归地拓展决策树的分枝, 完成决策树的构造。

信息增益计算公式的基本形式为 $gain(A) = I(P) - E(A)$, 其中 $E(A)$ 为属性 A 作为分类属性后的信息熵, 其表达形式为 $E(A) = \sum_{i=1}^{v} Pi \times I(p_j)$, v 为分类属性的取值种类, $I(p_j)$ 称为信息量。信息量 $I(p_j)$ 计算公式的基本形式为 $I(p_j) = -\sum_{i=1}^{n} p_j \times \log p_j$, 其中 n 为类主属性的取值种类, p_j 为某一类主属性所覆盖的记录数占该分类属性上总的记录数的百分比。

由公式 $I(p_i) = -\sum_{i=1}^{n} p_j \times \log p_j$, 其中 $\sum_{i=1}^{n} p_j = 1$, $n \in N$ 的特点, 我们提出针对信息量计算的改进方法, 用以简化信息量计算的复杂度。

信息量计算公式中, 用 $\log_2 p$ 函数作为信息量的计算单位(以下简写为 $\log p$), 根据 $\log p$ 函数的基本性质, 推出信息量计算的改进方法。

定量 **1**: 设 $f(x)$ 在 $[a, b]$ 上连续, 在 (a, b) 内具有一阶和二阶导数, 那么

(1) 若在 (a, b) 内, $f''(x) > 0$, 则 $f(x)$ 在 $[a, b]$ 上的图形是上凹(Concave) 的;

(2) 若在 (a, b) 内, $f''(x) < 0$, 则 $f(x)$ 在 $[a, b]$ 上的图形是上凸(Convex) 的。

信息量计算公式中所用的 $\log p$ 函数中, p 代表某一类记录数占总记录数的百分比, 其定义域为 $[0, 1]$, 并且当 $[0, 1]$ 上任意两点 p_1, p_2, 满足 $p_1 - p_2 = \Delta p \to \alpha(o)$ 时, $\log p$ 函数在 $[0, 1]$ 区间上连续, 所以

收稿日期: 2004- 03- 24

作者简介: 高学东(1963-), 男(汉族), 北京科技大学管理学院, 教授, 博士生导师, 研究方向: 管理过程优化.

可以根据定理 1, 检验 $\log p$ 函数的凹凸性。

因为 $(\log' p) = \frac{1}{p \times \ln 2}$, $(\log'' p) = -\frac{1}{p^2 \times \ln 2} < 0$, 所以根据定理 1 可知, $\log p$ 函数在定义域 $[0, 1]$ 上是上凸函数。

由上凸函数的基本性质, 可得如下结论:

性质 **1**: 若对于任意的 $\alpha, \beta \in [a, b]$, $t \in (0, 1)$, 若 $f(x)$ 在 $[a, b]$ 上为上凸函数, 则有

$$tf(\alpha) + (1-t)f(\beta) \leqslant f[t\alpha + (1-t)\beta] \quad (\text{公式 2.1})$$

恒成立。

由于前面已经证明 $\log p$ 在 $[0, 1]$ 上是上凸函数, 所以对于二值类主属性的信息量计算公式 $I(p_1, p_2) = -(p_1 \log p_1 + p_2 \log p_2)$, 其中 $p_1 + p_2 = 1$, 根据上凸函数的性质, 可以简化为 $I(p_1, p_2) = -(p_1 \log p_1 + p_2 \log p_2) \geqslant -\log(p_1^2 + p_2^2)$。对于信息量计算公式这样处理的好处是, 在利用 ID3 算法, 计算分类属性的信息增益的过程中, 凡遇到与信息量计算相关的内容都按上述方法进行简化处理, 可以减小信息增益的计算费用。众所周知, 计算机对 $\log p$ 函数的取值计算是需要花费大量时间的, 上述二值类主属性的信息量计算中, 若按原 ID3 算法, 则需要对 $\log p$ 函数计算两次, 而利用改进的 ID3 算法则只需要对 $\log p$ 函数计算一次, 就可以得出信息量的近似值, 所以上述这种 ID3 改进算法的使用会提高决策树构造的效率。

以上是针对凸函数的基本性质, 处理二值类主属性的信息量计算的简化问题。当类主属性取多值时, 信息量计算公式的形式为 $I(p_j) = -\sum_{i=1}^{n} p_j \times \log p_j$ 时, 其中 $n \in N$, 应如何做相应的简化处理呢? 这时就需要用到凸函数基本性质的推广定理。

性质 **2**: 设 $f(x)$ 为凸函数, S 是非空凸集, $x_i \in S$, $\lambda_i \geqslant 0$, $\sum_{i=1}^{m} \lambda_i = 1$, $m \in N$, 则

$$\sum_{i=1}^{m} \lambda_i f(x_i) \leqslant f\left(\sum_{i=1}^{m} \lambda_i x_i\right) \quad (\text{公式 2.2})$$

恒成立。

对于多值类主属性的信息量计算公式, 根据性质 2 可简化为 $-\sum_{j=1}^{n} p_j \log p_j \geqslant -\log\left(\sum_{j=1}^{n} p_j^2\right)$。

这种对多值类主属性的信息量计算的简化, 与原 ID3 算法计算 n 个 log 函数值相比, 只需要计算一次 log 函数值, 即可求出信息量的近似值, 减少了 $(n-1)$ 个 log 函数值的计算过程, 大大地提高了决策树构造过程中信息量的计算效率。

2.2 ID3 决策树改进算法的理论证明

ID3 决策树改进算法利用凸函数的性质, 简化了信息量的计算复杂度, 但是采用改进算法选择出的分类属性与原 ID3 算法选择出的分类属性是否完全一致呢? 下面分别以二值类主属性和三值类主属性为例, 从理论上说明 ID3 改进算法与原 ID3 算法可以获得完全相同的分类结果。

(1) 二值类主属性的情况

定理 2(一致性定理) 在 $(0, 1)$ 区间上, 任意一点 p_1', 有 $I(p_1, 1-p_1) \geqslant I(p_1', 1-p_1')$ 存在时, 则 $-\log[p_1^2 + (1-p_1)^2] \geqslant -\log[p_1'^2 + (1-p_1')^2]$ 恒成立。

证明: 信息量计算公式为 $I(p_1, p_2) = -(p_1 \log p_1 + p_2 \log p_2)$, 其中 $(p_1 + p_2) = 1$, 所以上述公式可转化为 $I(p_1, 1-p_1) = -[p_1 \log p_1 + (1-p_1)\log(1-p_1)]$。对上述信息量计算公式求一阶导数, 得 $I'(p_1, 1-p_1) = \log\frac{1-p_1}{p_1}$。

当 $p_1 = \frac{1}{2}$ 时, $I'(p_1, 1-p_1) = \log\frac{1-p_1}{p_1} = 0$;

当 $p_1 < \frac{1}{2}$ 时, $I'(p_1, 1-p_1) = \log\frac{1-p_1}{p_1} > 0$, 此时函数 $I(p_1, 1-p_1)$ 单调递增;

当 $p_1 > \frac{1}{2}$ 时, $I'(p_1, 1-p_1) = \log\frac{1-p_1}{p_1} < 0$, 此时函数 $I(p_1, 1-p_1)$ 单调递减。

又由于函数 $I(p_1, p_2) = -(p_1 \log p_1 + p_2 \log p_2)$ 是关于 $p_1 = \frac{1}{2}$ 左右对称的, 所以只讨论 $p_1 < \frac{1}{2}$ 或 $p_1 > \frac{1}{2}$ 其中的一种情况即可, 接下来, 我们选择 $p_1 < \frac{1}{2}$ 的情况进行讨论。

当 $p_1 < \frac{1}{2}$ 时, 由 $I'(p_1, 1-p_1) > 0$ 可知, 在 $p_1 \in \left(0, \frac{1}{2}\right)$ 区间上, 函数 $I(p_1, 1-p_1)$ 单调递增。又根据凸函数的性质, 有 $I(p_1, 1-p_1) = -[p_1 \log p_1 + (1-p_1)\log(1-p_1)] \geqslant -\log[p_1^2 + (1-p_1)^2]$ 存在, 其中函数 $(-\log[p_1^2 + (1-p_1)^2])' = -2\left[\frac{1}{p_1^2 + (1-p_1)^2} \times (2p_1 - 1)\right] > 0$, 所以函数 $-\log[p_1^2 + (1-p_1)^2]$ 在 $\left(0, \frac{1}{2}\right)$ 区间上也是单调递增

的。

若在$\left(0, \frac{1}{2}\right)$区间上, 任意一点$p'_1$, 有$I(p_1, 1-p_1) \geqslant I(p'_1, 1-p'_1)$存在时, 则有$\frac{1}{2} > p_1 > p'_1$恒成立; 又由于函数$-\log[p_1^2 + (1-p_1)^2]$在$\left(0, \frac{1}{2}\right)$区间上也是单调递增的, 进而推出$-\log[p_1^2 + (1-p_1)2] \geqslant -\log[p'^2_1 + (1-p'_1)^2]$恒成立。同理可证, 在$\left(0, \frac{1}{2}, 1\right)$区间上, 上述证明过程仍然成立, 定理 2 得证。

(2) 三值类主属性的情况

定理**3**(特殊一致性定理) 在$(0, 1)$区间上, 任意两点p'_1, p'_2, 有$I(p_1, p_2, 1-p_1-p_2) \geqslant I(p'_1, p'_2, 1-p'_1-p'_2)$存在, 则$Con(p_1, p_2, 1-p_1-p_2) \geqslant Con(p'_1, p'_2, 1-p'_1-p'_2)$恒成立。其中$Con(p_1) = -\log\left[(p_1)^2 + \left(\frac{v-\delta_1 p_1}{\delta_2}\right)^2 + (1-p_1 - \frac{V-\delta_1 p_1}{\delta_2})^2\right]$

证明: 信息量计算公式为$I(p_1, p_2, p_3) = -(p_1\log p_1 + p_2\log p_2 + p_3\log p_3)$, 由于满足$(p_1 + p_2 + p_3) = 1$, 所以信息量计算公式还可以表示为$I(p_1, p_2, 1-p_1-p_2) = -[p_1\log p_1 + p_2\log p_2 + (1-p_1-p_2)\log(1-p_1-p_2)]$, 信息量计算公式$I(p_1, p_2, 1-p_1-p_2)$的空间形式表示为三维曲面。由于三维曲面的空间形态难以表述, 我们利用$V = \delta_1 p_1 + \delta_2 p_2$, $(\delta_1 \neq 0, \delta_2 \neq 0)$的平面截三维曲面, 研究三维曲面在平面$V = \delta_1 p_1 + \delta_2 p_2$上的投影所形成的曲线的形状, 对比分析 *ID*3 算法和 *ID*3 改进算法的差异程度。

根据上述信息量计算公式可知, 平面$V = \delta_1 p_1 + \delta_2 p_2$截曲面$I(p_1, p_2, 1-p_1-p_2)$获得曲线的函数形式如下: (为简化, 以下用$I(p_1)$替代$I(p_1, p_2, 1-p_1-p_2)$表示信息量)

$$I(p_1) = -\left[p_1\log p_1 + \frac{V-\delta_1 p_1}{\delta_2}\log\frac{V-\delta_1 p_1}{\delta_2} + (1-p_1\frac{V-\delta_1 p_1}{\delta_2})\log(1-p_1-\frac{V-\delta_1 p_1}{\delta_2})\right] \quad (\text{公式 2.3})$$

对公式 2.3 求一阶导数, 得

$$I'(p_1) = -\log p_1 + \frac{\delta_1}{\delta_2}\log\frac{V-\delta_1 p_1}{\delta_2} - \left(\frac{\delta_1}{\delta_2} - 1\right)\log\left(1-p_1-\frac{V-\delta_1 p_1}{\delta_2}\right) \quad (\text{公式 2.4})$$

公式 2.4 中, δ_1, δ_2 取不同的值, 表示不同的平面截曲面$I(p_1, p_2, 1-p_1-p_2)$获得不同的曲线, 当$\delta_1 = \delta_2$时, 公式2.4的形式转化为$I'(p_1) = \log\frac{V-\delta_1 p_1}{\delta_1 p_1}$。

当$p_1 = \frac{V}{2\delta_1}$时, $I'(p_1) = \log\frac{V-\delta_1 p_1}{\delta_1 p_1} = 0$;

当$p_1 < \frac{V}{2\delta_1}$时, $I'(p_1) = \log\frac{V-\delta_1 p_1}{\delta_1 p_1} > 0$, 此时函数$I(p_1)$单调递增;

当$p_1 > \frac{V}{2\delta_1}$时, $I'(p_1) = \log\frac{V-\delta_1 p_1}{\delta_1 p_1} < 0$, 此时函数$I(p_1)$单调递减。

又由于用$V = \delta_1 p_1 + \delta_2 p_2$, $(\delta_1 = \delta_2)$平面截得的曲线$I(p_1)$是关于$p_1 = \frac{V}{2\delta_1}$左右对称, 所以只讨论$p_1 < \frac{V}{2\delta_1}$或$p_1 > \frac{V}{2\delta_1}$其中的一种情况即可, 接下来, 我们选择$p_1 < \frac{V}{2\delta_1}$的情况进行讨论。

当$p_1 < \frac{V}{2\delta_1}$时, 由$I'(p_1) > 0$可知, 在$p_1 \in \left(0, \frac{V}{2\delta_1}\right)$区间上, 函数$I(p_1)$单调递增。又根据凸函数的性质, 有$I(p_1) \geqslant -\log\left[(p_1)^2 + \left(\frac{V-\delta_1 p_1}{\delta_2}\right)^2 + (1-p_1-\frac{V-\delta_1 p_1}{\delta_2})^2\right]$, 设

$$Con(p_1) = -\log\left[(p_1)^2 + \left(\frac{V-\delta_1 p_1}{\delta_2}\right)^2 + (1-p_1-\frac{V-\delta_1 p_1}{\delta_2})^2\right]$$

当$\delta_1 = \delta_2$时, 有$Con'(p_1) = -\frac{2}{\left[(p_1)^2 + \left(\frac{V-\delta_1 p_1}{\delta_1}\right)^2 + (1-p_1-\frac{V-\delta_1 p_1}{\delta_1})^2\right]} \times \frac{2\delta_1 p_1 - V}{\delta_1} > 0$, 所以函数$Con(p_1) = -\log\left[(p_1)^2 + \left(\frac{V-\delta_1 p_1}{\delta_2}\right)^2 + (1-p_1-\frac{V-\delta_1 p_1}{\delta_2})^2\right]$在$\left(0, \frac{V}{2\delta_1}\right)$区间上也是单调递增的。

若在$\left(0, \frac{V}{2\delta_1}\right)$区间上, 任意一点$p'_1$, 有$I(p_1) \geqslant I(p'_1)$存在时, 则有$\frac{V}{2\delta_1} > p_1 > p'_1$恒成立; 又由于函数$Con(p_1) = -\log\left[(p_1)^2 + \left(\frac{V-\delta_1 p_1}{\delta_2}\right)^2 +\right.$

$(1- p_1- \frac{V- \delta_1 p_1}{\delta_2})^2]$ 在 $\left(0, \frac{V}{2\delta_1}\right)$ 区间上也是单调递增的, 进而推出 $Con(p_1) \geqslant Con(p_1')$ 恒成立。同理可证, 在 $\left(\frac{V}{2\delta_1}, 1\right)$ 区间上, 上述证明过程仍然成立, 定理 3 得证。

从上述的证明过程中, 可以看出 ID3 改进算法与原 ID3 算法在选择分类属性时, 具有相同的分类准确率。

2. 3　ID3 决策树改进算法实例及对比分析

我们以十四条记录[12]为基础, 随机生成多条记录组成多个数据集, 仍然在 Celeron900, RAM256 的计算机上, 利用 PB8. 0 前端开发工具, 实现 ID3 算法和 ID3 改进算法构造决策树的过程。然后, 利用 ID3 算法和 ID3 改进算法构造的决策树, 分别从以下几个方面进行对比分析:

(1) 生成决策树的形态

根据 ID3 改进算法的基本原理以及 ID3 算法和 ID3 改进算法对比分析, 可以发现, ID3 改进算法在构造决策树的过程中, 计算出的信息量与信息量的实际数值非常接近。同时根据我们提出并证明的两个"一致性定理", 可以保证决策树节点分类属性选择的一致性, 所以经过实验分析, 我们发现在不同规模的数据集中, 利用 ID3 算法和 ID3 改进算法构造的决策树, 在"节点数量","叶节点数量","树的深度","规则数量" 等几个方面完全相同, 说明 ID3 改进算法构造的决策树具有与 ID3 算法构造的决策树完全相同的分类准确率。

(2) 生成决策树节省的时间率

ID3 改进算法在计算信息量时, 花费的计算时间在绝大多数情况下, 比 ID3 算法花费的时间少许多, 这就决定了利用 ID3 改进算法构造决策树花费的时间也应该比 ID3 算法花费的时间少。为了证明上述结论, 我们分别以不同规模的数据集, 利用 ID3 算法和 ID3 改进算法构造决策树。在每一个数据集中, 分别利用上述两种方法对决策树逐级拓展过程中, 计算决策树每个节点上的信息增益所花费的计算时间进行测试, 取十次计算时间的平均值作为算法确定信息增益花费的计算时间。然后通过上述实验数据, 对比分析在构造决策树的过程中, 随着数据集规模的变化, ID3 改进算法比 ID3 算法在计算信息增益过程中节省时间率的变化情况。

表 **1**　***ID3*** 算法和 ***ID3*** 改进算法计算信息增益所用的计算时间　(单位: 毫秒)

记录数量 n	10^2	$5* 10^2$	10^3	$5* 10^3$	10^4	$5* 10^4$	10^5	$1.3* 10^5$	$5* 10^5$	10^6
ID3 所用时间- 平	3. 74	3. 74	3. 72	3. 67	4. 76	3. 60	3. 55	3. 45	3. 82	3. 90
ID3 改进所用时间- 平	3. 45	3. 48	3. 45	3. 40	4. 54	3. 34	3. 28	2. 85	3. 14	3. 34
节省时间	0. 29	0. 26	0. 26	0. 27	0. 22	0. 26	0. 27	0. 60	0. 68	0. 56
节省时间率(%)	7. 73	6. 83	6. 88	7. 36	4. 52	6. 85	7. 21	17. 14	17. 78	14. 40

从表 1 中, 可以看出, 在 100- 10 万条记录的数据集规模下构造决策树时, 改进的 ID3 算法计算信息增益的效率比原算法平均提高 8% 左右; 在 10 万- 100 万条记录的数据集规模下构造决策树时, 改进的 ID3 算法计算信息增益的效率比原算法平均提高 15% 左右, 这充分说明使用 ID3 改进算法能够以更高的效率构造决策树。

根据表 1 中的实验数据, 得出 ID3 改进算法节省时间率随数据集规模的变化趋势, 如图 1 所示:

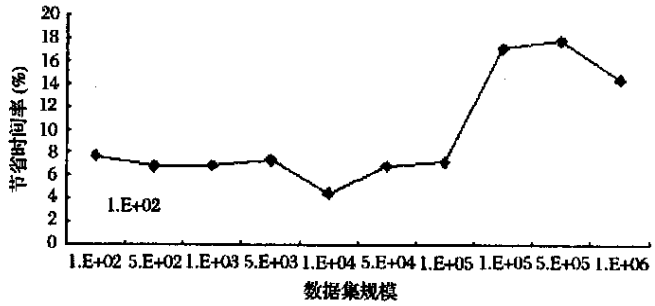

图 1　ID3 改进算法节省时间率随数据集变化趋势

从图 1 中, 可以看出, 在构造决策树的过程中, 随着数据集规模的增大, ID3 改进算法与 ID3 算法相比, 所节省的计算时间有增多的趋势, 这说明在处理大规模数据集的决策树构造过程中, ID3 改进算

法在计算效率上比 ID3 算法有更大的优越性。

从上述"生成决策树的形态"和"生成决策树的节省时间率"两个实验得到的结果可以得出如下结论:

(1) ID3 改进算法构造的决策树与 ID3 算法构造的决策树有完全相同的形态和分类准确率。

(2) 在相同规模的数据集中, ID3 改进算法构造决策树所用的计算时间比 ID3 算法构造决策树所用的计算时间少, 充分说明 ID3 改进算法提高了决策树构造的效率。

3 结论

本文根据 ID3 算法中的信息增益计算原理的特点, 利用上凸函数的性质简化信息增益的计算, 进而提高 ID3 算法中信息增益的计算效率。同时从实验和理论两方面证明, 改进的 ID3 算法与原 ID3 算法相比, 具有相同的准确率和更高的计算速度。

在实际应用方面, 我们以武钢 1995- 2002 年的销售数据为基础, 经过对原始数据集的预处理, 以"钢材利润率"作为类主属性, 选择"销售地区"、"销售产品"、"客户行业"、"销售渠道"等属性作为候选属性, 利用改进的 ID3 算法对武钢销售数据进行分类知识发现。得出诸如"产品大类= 线材 ∧ 客户行业= 电力 ∧ 销售地区= 华南 钢材利润率= 低"此类的规则, 为武钢销售公司对钢材销售市场的预测分析提供有力的决策支持。

参考文献:

[1] 邵华, 赵宏. 一种与神经元网络杂交的决策树算法[J]. 小型微型计算机系统, 2001, 22(8): 964- 966.

[2] 肖勇, 陈意云. 用遗传算法构造决策树[J]. 计算机研究与发展, 1998, 35(1): 49- 52

[3] Wang X Z, Chen B, Qian G L, et al. On the Optimization of Fuzzy Decision Trees[J]. Fuzzy Sets and Systems, 2000, 112: 117- 125.

[4] Chen M S, Yu P S, Liu B. A Method to Boost Naive Bayesian Classifiers[C]. In: Proceedings of The Sixth Pacific- Asia Conference on Knowledge Discovery and Data Mining, 2002: 115- 122.

[5] 赵卫东, 盛昭瀚, 何建敏. 粗糙集在决策树生成中的应用[J]. 东南大学学报(自然科学版), 2000, 30(4): 132- 137.

[6] Ling C X, Zhang H. Toward Bayesian Classifiers with Accurate Probabilities[C]. In Proceedings of The Sixth Pacific- Asia Conference on Knowledge Discovery and Data Mining, 2002: 123- 134.

[7] Provost F, Domingos P. Tree Induction for Probability- Based Ranking[J]. Machine Learning, September 2003, 52(3): 199- 215.

[8] Bredensteiner E J, Bennett K P. Feature Minimization within Decision Trees[J]. Computational Optimizations and Applications, 1998, 10: 111- 126.

[9] Elouedi Z, Mellouli K, Smets P. Belief Decision Trees: Theoretical Foundations[J]. International Journal of Approximate Reasoning, 2001, 28: 91- 124.

[10] Carmela C, Francesco M, Roberta S. A Statistical Approach to Growing A Reliable Honest Tree[J]. Computational Statistics and Data Analysisi, 2002, 38: 285- 299.

[11] Bartlett P L, Mendelson S. Rademacher and Gaussian complexities: Risk Bounds and Structural Results[J]. Journal of Machine Learning Research, 2002, 3: 463- 482.

[12] Quinlan J R, Induction of Decision Tree[J]. Machine Learning, 1986, 1(1): 81- 106.

An Improved Algorithm of Decision Trees by Using the Convex Function

GAO Xue- dong, YIN A- dong, ZHANG Jian, GONG Yu, WU Sen

(School of Management, Beijing University of Science and Technology, Beijing 100083, China)

Abstract: In this paper, we research deeply the theory of decision trees induction. According to the character of expected information and the quality of convex function, we propose a new algorithm to raise the efficiency of calculating expected information in the process of inducing the decision trees. By using the theory of consistency and special consistency, we also prove that the accuracy of decision trees constructed by the improved algorithm is equal to the one of ID3 algorithm. At the same time, through the experiment of testing the large datasets, we find that the new algorithm has higher calculative efficiency than the old one in the same datasets. Moreover with the larger scale of datasets, the calculation of expected information has more rapid efficiency.

Key words: decision tress; ID3 Algorithm; convex function; expected information