一种新型决策树属性选择标准

倪春鹏，王正欧

（天津大学系统工程研究所，天津，300072）

摘要：讨论传统决策树算法中三种常用的基于熵的属性选择标准，提出一种基于属性重要性排序的建立决策树的新方法。该方法在决策树的每个内结点首先依据属性重要性将属性进行排序，然后选择最重要的属性作为分类属性生成决策树，并抽取出规则。与传统的决策树数据分类方法相比，此方法可有效地选择出对于分类最重要的分类属性，增强决策树的抗干扰能力，并提高规则的预测精度。

关键词：决策树；重要性排序；数据分类

中图分类号：TP181 文献标识码：A 文章编号：1672－3090（2004）04－0437－04

在数据挖掘技术中，决策树分类算法以其抽取规则简便，规则易于理解等优点得到了广泛的应用。1986年 Quinlan首先提出 ID3算法，并在1993年提出后继的 C4.5算法，后者已成为广泛流行的决策树算法。决策树算法一般采用自顶向下的贪婪算法，在每个内结点选择分类效果最好的属性进行下一步的分类，直到这棵树能准确地分类训练样本，或所有的属性都已被使用过。影响决策树分类算法分类效果的主要问题是，在每个内结点如何选取要测试的属性以及剪枝技术。传统的属性选择标准中应用最为广泛的是，基于熵理论的各种标准，如信息增益（Information Gain）、信息增益率（Gain Ratio）等[1~3]。本文讨论基于熵理论的三种主要分类标准的特性，并针对这类标准的缺点提出了一种基于属性重要性排序的新的属性选择标准，依此建立的决策树，避免了传统决策树抗干扰属性能力不强的缺点，提高了决策树分类预测精度。

1 决策树算法简介

建立一棵决策树首先需要选择一个属性作为根节点，然后把该属性的每一个可能值作为一个分支。各分支将上一节点所含数据分成若干子集，在每个分支中再从所剩余的属性中找出一个属性作为该分支的下一个节点，如此往复，直到各分支数据同属一类，或所有属性均被选用为止。如 ID3算法采用信息增益作为选择节点属性的评价标准。假设 S为样例集，目标属性具有 n个不同的值，则 S的分类熵定义为

$$Entropy(S) = \sum_{i=1}^{n} -p_i \log_2 p_i \qquad (1)$$

式中：p_i——S中属于类别 i的比率。

假设一个实例集 S中的实例分属于 k个类别（d_1，d_2，…，d_k），其概率分别为 p_1，p_2，…，p_k。

假设属性 A将 S分为不相交的 n个子集 S_1，S_2，…，S_n，|S|为样例总数，$|S_i|$为属性 A取值为 i的样例数，则 S被 A所划分得到的熵 Entropy(A，S)定义为

$$Entropy(A,S) = \sum_{i \in Value(A)}^{n} \frac{|S_i|}{|S|} Entropy(S_i) \qquad (2)$$

式中：Value(A)——属性 A的取值集合；i——Value(A)中的某个值。

一个属性 A相对于样例集合 S的信息增益（IG）定义为

$$IG(A,S) = Entropy(S) - Entropy(A,S) \qquad (3)$$

在选择根节点及各内节点测试属性时，以信息增益值大的作为测试属性。

由于训练例子集中的噪音、错误项或干扰属性的影响，以此训练集生成的决策树常常包含了这些错误的信息，它能够正确分类训练集中的数据，但在分类测试例子集时精度不高，而且据此生成的决策树规模较大。这种现象被称为过拟合（Overfitting）。为了改善决策树的的性能，减小过拟合度，还需对决策树进行剪枝。剪枝分为前剪枝和后剪枝。前剪枝算法是在树的生长过程完成前就进行剪枝，如限制最小节点大小法，限制节点不纯度法等；后剪枝算法是当决策树的生长过程完成后再进行剪枝。如错误率降低，修剪方法中

收稿日期：2003－06－20

基金项目：国家自然科学基金资助项目（60275020）.

作者简介：倪春鹏（1976－），男，天津大学系统工程研究所，硕士生.

删除某结点为根结点的子树,使它成为叶子结点。该结点下训练样例分属 n个不同分类 $A_1, \cdots, A_n$,若分类 A_i为包含训练样例最多的分类,则把分类 A_i定义为该结点的分类。仅当修剪后的树对于验证集的分类性能不比原来的树差时才删除该结点。反复地修剪结点,每次总是选取那些删除后可以最大提高决策树分类精度的结点,继续修剪,直到进一步的修剪是有害时为止[4]。前剪枝效率高,但可能减小搜索视野;后剪枝比前剪枝的预测精度高,不会减小搜索视野,但由于允许先生成过拟合的决策树再反复回溯进行裁减,故大大增加了运算量。

2 三个基于熵的属性选择标准

信息熵是信息理论中用于分析不确定程度的一种重要度量,它从统计学角度得到描述一个给定问题所需的最小信息量,从而以所需信息量的多少来衡量不确定性的程度。信息熵的定义见式(1)。信息增益(IG),增益律(GR)和正规增益(NG)是分类问题中用于评价属性重要性的三个常用的基于熵的度量[5],分别定义为式(3)、式(4)、式(5)。

$$GR(A,S)=IG(A,S)/\left(\sum_{i=1}^{n}\frac{|S_i|}{|S|}\log\frac{|S|}{|S_i|}\right) \tag{4}$$

$$NG(A,S)=\frac{IG(A,S)}{\log_2 n} \tag{5}$$

式中:n——划分的区间数。

2.1 对 $y=p\log_2 p$对的讨论

$p\log_2 p$是三个标准中的基本元素,设 $y=p\log_2 p$,首先讨论它的一些性质。

当 $p=1$时,$y=0$;$p=0$为函数 $y=-p\log_2 p$的第一类间断点,$y=-p\log_2 p=\log_2(1/p)/(1/p)$,$p\to 0^+\Rightarrow 1/p\to\infty$,此时 $1/p$比 $\log_2 1/p$趋进于正无穷大的速度大,因此,此时 $y\to 0$。p是概率值,所以 $p\in[0,1]$,$dy/dp=-(\log_2 p+p/p)=-(\log_2 p+1)=0\Rightarrow p=1/2$,所以当 $p<1/2$时,$dy/dp>0$;$p>1/2$时,$dy/dp<0$;$p=1/2$时,$dy/dp=0$。$p=1/2$时,取最大值;$p\in(0,1/2)$时,y为增函数;$p\in(1/2,1)$时,y为减函数。

综上可得:当 $p=0$时,$y=0$;$p=1$时,$y\to 0$。0为 y的最小值。

2.2 对增益律 $IG(A,S)$的讨论

若 S共有 m类,某属性 A将 S分为不相交的 n个子集 $S_1, S_2, \cdots, S_n$,其中 $S_1, S_2, \cdots, S_n$只含有一个类别,即在 $S_i(1\leqslant i\leqslant n)$中,$p_1AS_j=0, \cdots, p_{i-1}AS_j=0$,$p_iAS_j=1$,$p_{i+1}AS_j=0, \cdots, p_mAS_j=0(1\leqslant i\leqslant m)$,其中,$p_iAS_j$表示区间 S_j属于第 i类的概率,则由2.1得 $Entropy(A,S)=0$。$IG(A,S)=Entropy(S)$,设划分 S的区间数为 n,S的例子数为 k,只要 $m\leqslant n\leqslant k$就可能得到上述的划分,一个极端的情况是将每个例子划分为一类,显然这种划分对数据分类不具有任何意义,但用信息增益标准衡量却是最好的划分方法之一,因此信息增益标准有一种偏好较细划分的倾向。

2.3 对信息增益 $GR(A,S)$的讨论

当 $|S_i|\approx|S|$时,$|S|/|S_i|\to 1\Rightarrow\log_2|S|/|S_i|\to 0$,即此时 $GR(A,S)$趋于无穷大,因此增益律标准有一种偏好不均匀划分的倾向。

2.4 对正规增益 $NG(A,S)$的讨论

正规增益 $NG(A,S)$是在信息增益 $IG(A,S)$的基础上除以 $\log_2 n$,即将信息增益 $IG(A,S)$中的以2为底的对数换为以 n为底的对数,这样制约了 n增大时全式的值,它有效区分了2.2中极端状况($m<n$)与有效划分($m=n$)。在大多数属性离散化和属性选择的实验中,NG比 IG和 GR得到的结果要好[5]。

注意到 NG表达式(5)中,当划分数 n增大时,NG减小,当划分出的各区间 S_i纯度(属于某类别得概率)增大时,NG增大,两种相反的作用使 NG标准在划分不均匀时可能不能正确地区分最优的划分。下面举一反例:

某数据库如表1所示。

表1 样例数据库

	数据值									
属性 A	0.2	0.2	0.2	0.2	0.2	1.15	1.15	1.15	1.15	1.15
属性 B	0.75	0.75	0.75	0.75	0.75	1.56	1.56	1.56	1.56	1.93
类别	1	1	1	1	1	2	2	2	2	3

显然,属性 B比属性 A的分类效果更好,分别计算 A,B的 NG值。

$$Entropy(S)=-\frac{1}{2}\log_2\frac{1}{2}-\frac{2}{5}\log_2\frac{2}{5}-\frac{1}{10}\log_2\frac{1}{10}=1.361$$

$$Entropy(A,S)=\frac{1}{2}\left(-\frac{4}{5}\log_2\frac{4}{5}-\frac{1}{5}\log_2\frac{1}{5}\right)=0.361$$

$$Entropy(B,S)=0$$

$$NG(A,S)=1.361-0.361=1$$

$$NG(B,S)=\frac{1.361-0}{\log_2 3}=0.8587$$

$$NG(A,S)>NG(B,S)$$

选属性 A_i与事实矛盾。

3　基于属性重要性的属性选择标准

属性重要性的定义方法很多, 文献 [6] 提出的输入输出关联法 (IOC)用样本值的变化而引起的输出变化的累加值作为衡量数据属性重要性的标准。对于某个属性 k,如果样本值变化而引起的输出变化越大, 则说明该属性越重要, 反之则说明该属性对于数据分类的意义不大。该方法可用下式表示:

$$C(k)=\sum |x(i,k)-x(j,k)|\cdot \operatorname{sign}|y(i)-y(j)| \quad i\neq j \tag{6}$$

式中: $C(k)$——第 k个属性的输入输出关联值; $x(i,k)$, $x(j,k)$——第 i, j个样本的第 k个条件属性值; $y(i)$, $y(j)$——第 i, j个样本的决策属性值; $\operatorname{sign}(x)$——符号函数。

当 $x>0$时返回 1,当 $x=0$时返回 0。为消除量纲不同所造成的影响, 对该方法进行归一化处理:

$$x'=\frac{x-\min(A)}{\max(A)-\min(A)} \tag{7}$$

式 (6)变为

$$C(k)=\sum |x'(i,k)-x'(j,k)|\cdot \operatorname{sign}|y(i)-y(j)| \quad i\neq j \tag{8}$$

在决策树的生长过程中, 本文提出可在每个结点均选择输入输出关联值最大的属性作为测试属性建立决策树。对于定性描述的属性, 为使输入输出关联值更好地反映输入输出变化的关系, 本文将式 (8)修改为式 (9)(此时由于取信号函数值, 故不需归一化处理)。

$$C(k)=\sum \operatorname{sign}|x(i,k)-x(j,k)|\cdot \operatorname{sign}|y(i)-y(j)| \quad i\neq j \tag{9}$$

这样, 输入输出关联值作为属性选择标准避免了计算属性对实例集合的划分、及划分中实例属于某类别的概率, 因而避免了第二部分讨论中基于熵的各标准的问题; 而输入输出关联值本身的定义又可使其能区分出与类别变化一致度最高的属性。因此, 输入输出关联值可作为建立决策树中一种较理想的属性选择标准, 它减少了基于熵的各标准中以对数计算为累加计算的计算量, 从而提高了属性选择的效率。

4　仿真试验

这里采用加利福尼亚大学机器学习公共数据库[7]中的 3个数据库进行仿真试验, 并采用 C4.5作为规则生成器。其结果见表 2、表 3。

表 2　数据库相关信息

数据库	例子总数	条件属性数	类别数	训练集所占比率/%
Zoo	101	16	7	75
Breast cancer	554	9	6	75
Lymphography	148	18	4	75

表 3　用四种标准建立决策树预测精度

数据库	IG/%	GR/%	NG/%	C(k)/%
Zoo	90	90	95	98.5
Breast cancer	71.93	72.63	82.5	86.5
Lymphography	65.52	68.7	68.75	68.75

三种基于熵的属性选择标准在很多情况下可以较好地选择出测试属性, 因此得到广泛的应用。而输入输出关联法选择测试属性克服了三种标准的缺点, 经仿真试验表明, 可较好地选择出分类效果好的属性建立决策树。

5　结论

本文分析了建立决策树中三种基于熵的属性选择标准, 指出三种标准中存在的缺点, 提出以输入输出关联法作为属性选择标准建立决策树, 取得了较好的效果。

参　考　文　献

[1] Buntine W, Niblett T. A Further Comparion of Splitting Rules for Decision-tree Induction[J]. Machine Learning, 1992, 8(1): 75—85.

[2] Kononenko I, Se J H. Attribute Selection for Modeling[J]. Future Generation Computer Systems, 1997, 13(2—3): 181—195.

[3] Shih Y S. Families of Splitting Criteria for Classification trees[J]. Statistics and Computing, 1999, 9(4): 309—315.

[4] Tom Mitchell. Machine Learning[M]. 北京: China Machinery Press, McGraw-Hill Education (Asia), 2003. 42—47.

[5] Hong S J. Use of Contextual Information for Feature Ranking and Discretization[J]. IEEE Transactions on Knowledge and Data Engineering, 1997, 9(5): 718—730.

[6] 文专, 王正欧. 一种高效的基于排序的 RBF神经网络属性选择方法[J]. 计算机应用, 2003, (8): 34—36.

[7] P M Murphy, C J Merz. UCI Reposiory of Machine Learning Databases[EB/OL]. http://www.ics.uci.edu/mlearn/MLRepository.html, 1998.

A New Attribute Selection Criterion of Decision Tree

NI Chun-peng, WANG Zheng-ou

(Institute of Systems Engineering, Tianjin University, Tianjin 300072, China)

Abstract: This paper discusses three common entropy-based attribute selection criteria of the traditional decision tree arithmetic and presents a new decision tree building method based on attribute importance ranking. The method ranks attributes based on the importance of the attributes in every decision tree interior nodes, and then selects the most important attribute as the ranking attribute to build a decision tree and extracts rules. Compared with the traditional data classification methods used in decision tree, the proposed method can find out the most important attribute efficiently, raises the anti-jamming capacity of decision tree and improves the prediction precision of rules produced.

Keywords: decision tree; importance ranking; data classification

[责任编辑 彭金旺]

(上接第430页)

定法[J]. 生物化学与生物物理进展, 1986, 13(4): 64.

[3] 向荣, 王鼎年. 过氧化脂质硫代比妥酸分光光度法的改进[J]. 生物化学与生物物理进展, 1990, 17(3): 241.

[4] 夏奕明, 朱连珍. 血和组织中谷胱苷肽过氧化酶活力测定[J]. 卫生研究, 1987, 16(4): 29.

[5] Naayanan. Differential Alterations in ATP-supported Calcium Transport Actives of Sarcoplasmic Reticulum and Sarcolemma of Aging Myocardium[J]. Biophasmic Acta, 1981, 678(8): 442.

[6] 邵洪, 汪代良, 尤忠义, 等. 氧自由基与蛋白质代谢[J]. 国外医学分子生物学分册, 1990, 12(1): 42.

[7] Han Qing-bong, Shu Hu-ying, Wang Jian, et al. Study on the Effects of Foshousany plus Danshen in Preventing IUGR Rats with Passive Smoking from Peroxidation in Erythrocyte Lipid[J]. J of Tongji Med Univ, 1995, 15(2): 120.

Protective Effects of Ligustrazine on Digestive System of Buurned Rats

ZHU Han-rong, HE Li-ya, ZHU Lei, CHENG Cai-lian,
CHEN Yong, ZHAO Xue-hua, WU Mian-yun, XIA Rui-yun

(College of Medical Science, Wuhan University of Science and Technology, Wuhan 430080, China)

Abstract: To observe the effects of ligustrazine on the contents of ATPase, GSH-Px, MDA in stomach, intestine and liver of severely burned rats, the rats were randomly divided into three groups. Tissues of stomach, intestine and liver were gathered and homogenized respectively, and contents of ATPase, GSH-Px, MDA were assayed. It was found that the contents of ATPase, GSH-Px in stomach, intestine and liver decreased significantly while MDA were obviously enhanced in comparison with that of normal group. After the treatment with ligustrazine, the contents of ATPase, GSH-Px were significantly higher but MDA lower than that of non-treated burned group. So it can be concluded that ligustrazine has protective effects on stomach, intestine and liver in burned rats. This might be because ligustrazine can reduce free radicals and has antioxidant effects on stomach, intestine and liver.

Keywords: ligustrazine; lipid peroxide; burn

[责任编辑 徐前进]