第36卷第5期 中国科学技术大学学报 Vol. 36, No. 5
2006年5月 JOURNAL OF UNIVERSITY OF SCIENCE AND TECHNOLOGY OF CHINA May 2006

文章编号:0253-2778(2006)05-0546-04

基于非线性拟合方程的多变量决策树算法*

吴 强,李金龙,杨振宇,王煦法

(中国科学技术大学计算机科学技术系,安徽合肥 230027)

摘要:根据数据属性间存在的线性相关和非线性相关影响决策树性能的特点,提出了一种用拟合回归建立决策树的算法,并利用这种相关性来提高分类能力.该算法选择了一个较优的属性子集,对此子集中的属性进行加权组合,用于构造决策树的节点,采用二次多项式来拟合两个属性间可能存在的相关性,从而构造出分类能力更强的决策树.研究中用 UCI 标准数据集对各种算法进行测试及比较,实验结果及分析表明此决策树算法具有良好性能.

关键词:决策树;相关性;属性组合;多变量决策树

中图分类号:TP181 **文献标识码**:A

A multi-variable decision tree algorithm based on nonlinear fitting equation

WU Qiang, LI Jin-long, YANG Zheng-yu, WANG Xu-fa

(*Department of Computer Science and Technology, USTC, Hefei* 230027, *China*)

Abstract: Linear and nonlinear correlations between features affect the ability of the decision tree. To exploit those correlations and improve classification ability, a new method was proposed to construct decision trees in which a quadratic fitted model was used. The algorithm selected approximatively optimal feature sets, and these features were weighted and assembled so as to construct a node of the decision tree. At the same time to model those possible linear and nonlinear correlations between features, features were fitted using quadratic fitted equations each time, then the datasets were partitioned by fitted results, and the subdataset is handled by recursive process to build a decision tree with better performance of classification. UCI standard data sets were tested in the research, and classification results of different algorithms show that the proposed decision tree algorithm has better performance than other tested decision tree algorithms.

Key words: decision tree; correlation; combining features; mult-ivariate decision tree

0 引言

决策树分类方法以其易理解性、需要信息量少、效率及准确率较高等优点在分类挖掘算法中占有重要地位.由于应用的需求,到目前为止,已经涌现出大量的决策树分类算法及其改进算法,年代较远的如 CART 算法[1]、ID3 算法[2]、C4.5 算法[3]等,较新的如多变量决策树[4]、分层归纳[5,6]、功能树[7,8]、基于神经网路的混合型决策树[9]等. CART、ID3、C4.5 这些算法的基本思想是每次贪心选择一个分类能力

* **收稿日期**:2005-08-16;**修回日期**:2006-03-09

基金项目:国家自然科学基金委海外青年学者联合会研究基金(60428202)资助.

作者简介:吴强,男,1964 年生,博士生/高工.研究方向:知识发现,决策理论,数学模型. E-mail:qiangw@mail.ustc.edu.cn

通讯作者:王煦法,教授. E-mail:xfwang@ustc.edu.cn

最强的属性作为当前节点的测试属性,然后用测试属性将数据集划分成若干子集,对各子集再进行相似处理,它们之间的最大区别是对属性分类能力的评价标准不同. 后来的研究表明 CART、ID3、C4. 5 算法没有考虑到属性间的相关性,而这种相关性在提高决策树的分类能力及效率上存在很大潜力. 于是,研究者们提出了两种改进策略:一是选择与分类目标最相关,具有最大分类能力的属性子集来构造节点,这是个 NP 完全难题,人们采用了求近似最优子集的方法[10,11];二是通过线性分类器组合加权属性的决策树算法,比较有代表性的如分层归纳[5,6]、功能树[7,8]等,它们的设计思想是先用若干个线性分类器处理数据,产生样例的类别概率分布,然后把得到的类别概率分布作为新属性添加到原始数据集中,最后用一种分类方法根据新数据集建立决策树,这些算法由于考虑到了属性间的相关性,其分类能力有所提高.

1　预处理知识

1.1　传统决策树分类算法

决策树算法主要思想是通过构造决策树来发现数据中蕴涵的分类规则,如何构造精度高、规模小的决策树是决策树算法的核心内容.

传统决策树的构造过程可分为以下几个步骤:

(Ⅰ)规范原始数据,即从原始数据中抽象出属性集和类别属性;

(Ⅱ)在候选属性集中选择最有分类能力的属性作为当前节点的测试属性;

(Ⅲ)根据当前节点测试属性取值的不同,将训练数据集划分为若干子集;

(Ⅳ)针对上一步中得到的每一个子集,重复进行上述的(Ⅱ)、(Ⅲ)两个步骤,直到最后子集已能给出类别标识或者测试属性已无有助分类的信息;

(Ⅴ)生成叶子节点,即给出类别标识.

通过上述步骤,我们就得到了对数据集进行分类的决策树. 在决策树算法的建树过程中,测试属性的选择是一个关键的步骤. 必须对属性的“分类能力”给出合理的评价标准,这样才能每次选择“最有分类能力”的一个属性作为当前节点的测试属性. 根据评价标准的不同发展出了很多决策树分类算法,典型的评价标准有信息熵理论、基尼系数(gini index)等. 基于信息熵理论的算法有 ID3[2]、C4. 5[2]等,而基于基尼系数的算法有 CART[1]等. 除评价标准不同外,这些算法结构和特性相似,都没有考虑属性间可能存在的相关性.

1.2　多分类器组合决策树分类算法

基于多分类器组合的决策树分类算法是新兴的决策树分类算法,也是当前的研究活跃区域. 这类算法的基本设计思想是希望能够利用多个不同特性的分类器联合建立决策树. 其基本过程是先选择一个分类器对数据集分类,产生原始数据集的类别概率分布,将这些类别概率分布加入到原始数据集中作为新属性,从而构成了新数据集;然后使用其他分类器对新数据集作相似的处理,直至最终由最后一个分类器给出类别判定结果.

基于多分类器组合的典型算法有分层归纳(cascade generalization)[5]、约束分层归纳(constrained cascade generalization)[6]、功能树(functional tree)[7,8]等. 分层归纳方法可以分为两类,称之为松耦合和紧耦合[5]. 松耦合是先用若干分类器来分类数据,分类的结果是数据的类别概率分布,然后把得到的概率分布作为新属性添加到原始数据集中,最后用一种分类方法根据新数据集来建立决策树;紧耦合又称为局部分层归纳,它在建树过程中生成的中间节点处运用分类器来生成新属性,这些新属性被下层节点用来划分数据集和分类.

约束的分层归纳方法思想是来自未扩展分层归纳方法. 松耦合只在决策树的根节点处用其他分类器归纳一次,以产生新属性,而紧耦合则是在决策树的每一层都用分类器来产生新属性. 约束归纳引进一个代表归纳层次的参数,通过该参数可以很方便地控制归纳层次,而不仅仅只是松耦合或紧耦合两种情况. 调整约束参数,可以得到各个归纳层次的决策树,从而可以选择性能最好的一棵作为预测器.

功能树是另外一种基于多分类器合并的决策树算法,它是沿袭紧耦合分层归纳方法的算法思想来建立决策树,二者的最大区别是功能树剪枝可能产生一种称为功能叶的叶节点. 功能叶不直接给出类别信息,而是根据建树过程中生成的构造函数来计算样例的类别.

合理利用数据集属性间的相关性将有助于提高决策树的分类能力,但是对于一个给定的数据集,其属性间是否有相关性,有什么类型的相关性,我们只能从样本来推测. 上述若干种多分类器组合的策略仅是属性间可能存在的线性相关,对于属性间可能存在的非线性相关,目前的研究还很少. 本文提出一

种新的决策树算法,该算法结合了上述两种策略,即选择一个较优的属性子集,对此子集中的属性进行加权组合,并用于构造树的节点,同时考虑属性间可能存在的线性和非线性相关的情形,采用一次和二次多项式来拟合两个属性间可能存在的相关性,从而构造出分类能力更强的决策树.

2 改进算法及实验分析

2.1 算法思想及过程

为能更好地利用属性间隐含的关系,先考虑根据候选属性做拟合,然后对拟合出的连续结果根据分类问题的类别信息离散化. 结合决策树算法的贪心策略,每次选择 n 个属性进行一次拟合,并根据拟合结果划分数据集,对子数据集进行相应的过程,这样就可以建立一个基于拟合的决策树. 为弥补未选择全部属性拟合的局限性,可考虑把每次拟合的结果作为新的属性加入到原数据集中,构造出更复杂的拟合方程. 这样虽然可能会造成过拟合,但实验统计结果显示,生成的决策树结点数比 C4.5 算法结果来说明显变少,降低了决策树的复杂性.

拟合变量个数 n 可以作为算法的参数,选择使用不同的参数 n 可实现不同的算法策略,至于是否将每次的拟合结果作为新属性加入原数据集中,需根据实验结果比较和不同策略的差异、特点等来决定. 具体算法流程如下所示:

算法中使用符号定义为:

samples:训练数据集,包含 N 个样本; alist:候选属性集合;F:拟合方程;n:拟合变量个数;Node:表示一个节点;$\boldsymbol{Y}=(c_1,c_2,\cdots,c_N)$:$N$ 个样本的类别属性;$\boldsymbol{R}=(r_1,\ r_2,\cdots,r_N,)$:拟合结果,为一连续浮点数向量;$\boldsymbol{D}=(d_1,\ d_2,\cdots,d_N,)$:拟合结果 $\boldsymbol{R}$ 离散化后的结果,也是一个向量;error:拟合误差;算法:Regress 由给定的训练数据集和拟合方程生成一棵拟合回归决策树;输入:samples,alist,F,n;输出:一棵决策树.

算法:

(Ⅰ)Node=Create();//创建节点 Node;

(Ⅱ)从 alist 中选取满足拟合误差最小的 n 个属性,用拟合方程 F 做一次拟合得到连续结果 $\boldsymbol{R}$;

(Ⅲ)把(Ⅱ)中选择的 n 个属性记录在节点 Node 中,把 $\boldsymbol{R}$ 作为一个新属性加入到训练数据集中,即 alist=alist$\cup\boldsymbol{R}$;

(Ⅳ)对 $\boldsymbol{R}$ 进行离散化得到 $\boldsymbol{D}$,即令 $d_i=c_j$;其中,c_j 在 Y 的所有分量中离 r_i 最近;

(Ⅴ)根据 $\boldsymbol{D}$ 来计算拟合误差,error=count_none_zero($\boldsymbol{Y}-\boldsymbol{D}$);//count_none_zero($\boldsymbol{Y}-\boldsymbol{D}$),求向量 $\boldsymbol{Y}-\boldsymbol{D}$ 中非零分量个数;

(Ⅵ)if (error=0) then

Node 标记为叶子,返回节点 Node,转向(Ⅷ);

else Node 标记为中间节点;

(Ⅶ)for each (d_i in $\boldsymbol{D}$),由节点 Node 生成一个拟合结果为 d_i 的分枝;设 S_i 是训练数据集 samples 中所有拟合结果为 d_i 的所有样本的集合;加上一个由 Regress(S_i,alist,F,n)返回的节点;//递归调用 Regress;

(Ⅷ)end

通过改变拟合方程 F 和拟合变量个数 n 可以得到各种基于拟合回归的决策树算法,本文讨论最简单的情况,即使用两个拟合变量,取拟合方程为

$$aX_1^2+bX_2^2+cX_1X_2+dX_1+eX_2+f \quad (1)$$

其中,a、b、c、d、e、f 是拟合系数,这样我们就得到了建立拟合回归决策树的具体算法 Regress 1;为预防把数据集划分为太多分枝而产生的数据碎片问题,可以考虑将多分类问题预先转化为二分类问题[11],然后依据 Regress 1 再相似处理,最终生成一棵二叉树,我们把这种算法取名为 Regress 2.

2.2 实验分析

我们从 UCI 标准数据集中选出 15 组数据对各种算法进行测试,同时给出传统决策树算法 C4.5[3] 的实验结果,以利于对比分析. 所选数据集的数据特征如表 1 所示.

实验采用 10 倍交叉验证的方法,表 2 给出了各种算法的实验结果. 其中,数字代表测试集的分类错误百分比,数字前面的(+)号代表此结果优于 C4.5 算法的结果,(−)号代表此结果不如 C4.5 的结果,(=)号代表此结果与 C4.5 算法的结果相当.

首先看算法 Regress 1 的结果,与 C4.5 相比,Regress 1 具有一定优势,除多类别数据集 glass、led7、led24、waveform-21 的结果比 C4.5 差外,其余 5 个数据集的结果与 C4.5 相当,有 6 个数据集的结果明显比 C4.5 要好;再看算法 Regress 2 的结果,由于 Regress 2 只是 Regress 1 在多类问题上的改进算法,所以两者在二类别数据集上的结果相同,而在多类别数据集上 Regress 2 的结果除有 4 个与 C4.5 相当外,其余结果都比 C4.5 的好. 因此,算法

Regress 2 的结果比 C4.5 要好.

表 1　实验选用数据集的数据特征

Tab. 1　Features of the selected data set for experiment

数据集名称	属性个数	离散属性个数	连续属性个数	数类别个数
corral	6	6	0	2
diabetes	8	0	8	2
glass	9	0	9	7
iris	4	0	4	3
led7	7	7	0	10
led24	24	7	0	10
mofn-3-7-10	10	10	0	2
monks1	6	6	0	2
monks2	6	6	0	2
monks3	6	6	0	2
parity5+5	10	10	0	2
pima	8	0	8	2
tic-tac-toe	9	9	0	2
vehicle	18	0	18	4
waveform-21	21	0	21	3

表 2　各种算法的测试错误率

Tab. 2　Error rate of test for the different methods

数据集名称	C4.5 算法	Regress 1	Regress 2
corral	18.8	(+)0.00	(+)0.00
diabetes	30.9	(+)25.39	(+)25.39
Glass	37.5	(−)52.78	(+)19.44
iris	8.0	(=)8.00	(=)8.00
led7	32.5	(−)49.6	(+)27.93
led24	37.1	(−)44.73	(+)21.83
mofn-3-7-10	14.5	(+)10.16	(+)10.16
monks1	23.4	(+)5.56	(+)5.56
monks2	34.7	(+)28.47	(+)28.47
monks3	2.8	(+)2.78	(+)2.78
parity5+5	50.0	(+)34.38	(+)34.38
pima	23.8	(+)22.66	(+)22.66
tic-tac-toe	16.9	(−)18.44	(−)18.44
vehicle	31.6	(+)30.50	(+)23.40
waveform-21	29.3	(−)33.43	(+)21.32

从实验结果及分析可以看出,由于考虑了属性间可能存在的线性和非线性相关性,选用二元二次多项式建立的拟合回归决策树取得了良好的实际效果.

3　结论

本文借助拟合回归的思想,提出一种建立拟合回归决策树的方法,通过用二元二次方程作为拟合方程,拟合出了更平滑的分类界面,同时考虑了属性间可能存在的线性和非线性相关性. 在 UCI 数据集上的实验结果及分析表明这种算法相对于经典的 C4.5 算法在性能上有了较大提高. 但是,同样由于采用了非线性的拟合方程,使得分类器的可理解性较传统的决策树分类器有所降低. 如何均衡分类器的分类精度和可理解性这一对矛盾仍是一个需要进一步研究的问题. 其他待研究的工作还包括如何利用属性间相关性来辅助分类的方式和方法,如何从数据集中发现更多类型的相关性辅助分类等.

参考文献(References)

[1] Olshen R A, Breiman L, Friedman J H, et al. Classification and Regression Trees[M]. Boca Raton, Florida: Chapman & Hall, 1984.

[2] Quinlan J R. Induction of decision trees[J]. Machine Learning, 1986, 1(1): 81-106.

[3] Quinlan J R. C4.5: Programs for Machine Learning [M]. San Francisco: Morgan Kaufmann Publishers, 1993.

[4] Brodley C E, Utgoff P E. Multivariate decision trees [J]. Machine Learning, 1995, 19(1): 45-77.

[5] Gama J, Brazdil P. Cascade generalization [J]. Machine Learning, 2000, 41(3): 315-343.

[6] ZHAO H M, Ram S. Constrained cascade generalization of decision trees[J]. IEEE Transactions on Knowledge and Engineering, 2004, 16(6): 727-739.

[7] Gama J. Functional trees for classification[C]// 2001 IEEE International Conference on Data Mining. IEEE Computer Society, 2001: 147-154.

[8] Gama J. Functional trees [J]. Machine Learning, 2004, 55(3): 219-250.

[9] ZHOU Z H, CHEN Z Q. Hybrid decision tree[J]. Knowledge-Based Systems, 2002, 15(8): 515-528.

[10] LIU Huan, Setiono R. Feature transformation and multivariate decision tree induction [C] // Discovery Science. Berlin/Heidelberg: Springer-Verlag, 1998: 279-291.

[11] SHA Hui-xin, YE Dong-yi. A knowledge roughness based approach to multivariate decision tree construction[J]. Journal of Fuzhou University(Natural Science), 2004, 32(2): 138-141.
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, 2004, 32(2): 138-141.

[12] Schapire R E, Singer Y. Improved boosting algorithms using confidence-rated predictions [J]. Machine Learning, 1999, 37(3): 297-336.