第17卷 第3期
2007年3月

计算机技术与发展
COMPUTER TECHNOLOGY AND DEVELOPMENT

Vol.17 No.3
Mar. 2007

基于遗传算法的决策树优化模型

华文立,胡学刚

(合肥工业大学 计算机与信息学院,安徽 合肥 230009)

摘 要:在分析C4.5算法原理的基础上,进一步讨论了C4.5算法在决策树的规模控制、属性选择、滤躁和去除不相关属性等方面的不足,讨论了决策树挖掘中对训练数据进行属性约简的必要性。从实用的角度提出了一种利用遗传算法进行寻优的、基于属性约简的决策树构建模型,并为此模型设计了一个适应度函数。该模型具有自适应的特点,通过调整适应度函数的参数,可以约束遗传算法的寻优方向,实现对决策树的优化。实验表明,决策树寻优后,在所用训练集属性减少的同时,分类精度却有一定程度的提高,而分类规则的规模却降低了,因此,该模型具有一定的实用价值。

关键词:决策树;属性约简;遗传算法;适应度函数

中图分类号:TP18 **文献标识码**:A **文章编号**:1673-629X(2007)03-0116-03

The Model of Decision Tree Based on Genetic Algorithm

HUA Wen-li, HU Xue-gang

(School of Computer and Information, Hefei University of Technology, Hefei 230009, China)

Abstract: Based on the analysis of C4.5 algorithm, presents the defects of the scale control of decision tree and attribute selection, and in eliminating noise and irrelevant attributes. The paper also discusses the necessity of conducting attribute reduction for the training data in the course of decision tree mining. In addition, for the practical demands, the paper, based on attribute reduction, proposes a model for decision tree to optimize it by adopting genetic algorithm. Then a fitness function is designed for the model. The model maintains the characteristic of self-adjustment, can control the optimization direction of genetic algorithm, and optimize the decision tree by adjusting the parameters of fitness function. An experiment is conducted and the findings of the experiment show that after the optimization of the decision tree, the attributes of training data will be reduced, the classification accuracy will be improved and the scale of the classification rules will be made smaller. Therefore, the model is of great practical value.

Key words: decision tree; attribute reduction; genetic algorithm; fitness function

0 引 言

分类是数据挖掘领域中重要的研究课题之一,分类规则是在已知训练样本的特征和分类结果的基础上,为每一种类型找到一个合理的描述或模型,然后再用这些分类的描述或模型对未知的新数据进行分类[1]。目前已有多种分类理论,如粗糙集理论、神经网络、统计模型、贝叶斯分类器、支持向量机和决策树等,其中,决策树是较为常用的方法之一。其中最著名的决策树算法有ID3,C4.5,CART等,文中将针对目前较流行的C4.5算法进行讨论。

决策树学习是以实例为基础的有监督的归纳学习算法,通过一组无次序、无规则的实例推理出决策树表示形式的分类规则。对决策树的评价有一些量化的评价标准,除去分类的正确性应当放在第一位考虑,决策树的复杂程度是另外一个需要考虑的重要因素。从实用的角度看,如果决策树构造得过于复杂,那么对于用户来说这个决策树是难以理解的,因此,权衡决策树分类精度和决策树规模是有必要的。

1 构建决策树时进行属性约简的必要性

C4.5算法[2]是自上而下构造决策树的,选择分裂属性的依据是信息增益比率,采用后修剪策略,相对于其他分类算法来说,分类准确率是比较高的。但C4.5也有以下缺陷:

(1)C4.5采用分而治之的策略,因此获得的决策树可能是局部最优的,但不一定是全局最优的。

(2)C4.5算法采用先构建树、后修剪策略,而构建树时,可能会由于过早引入躁声或不相关属性,造成树的先天结构不好,单靠修剪,并不能从根本上改变树的

收稿日期:2006-05-19

作者简介:华文立(1969-),男,安徽涡阳人,安徽电子信息职业技术学院讲师,硕士研究生,研究方向为软件工程、数据挖掘;胡学刚,教授,硕士生导师,研究方向为数据挖掘与人工智能。

结构。

(3)C4.5 算法依据属性的信息增益比率[3]来选择分裂属性,所构建的决策树深度仍然有可能很大,因此可能会造成导出的规则过多、规则长度过大,而这些在实际应用中都会直接影响决策树的预测速度。

(4)C4.5 算法主要从属性在分类能力上度量属性的重要性,而检验属性分类能力是单个进行的,忽略了属性的相关性,对属性的优化组合缺乏重视。事实上,在构建树时,每个属性的重要性并不相同。

基于以上分析,笔者认为应该在构建决策树时对属性进行约简,通过对训练样本进行属性约简,抽出重要属性用于构建决策树,尽可能去除训练样本中的躁声和不相关属性,降低躁声和不相关属性对构建树的影响,优化属性组合。同时,结合实际需求,应该控制由决策树导出的规则数量和规则长度,提高所构造决策树的分类速度和效率。因此,在构建树时应该对树的规模(决策树的规模指规则数量和规则长度)有所控制,在可以接受的分类精确度范围内,控制决策树的规模。

虽然属性选择对构建决策树很重要,但如何选择始终是个难题。显然,穷举每一种可能属性组合的方案是不可取的,因为当训练集属性很多时,可能的数据组合方式是个天文数字。例如文中实验中所用样本数据的属性共有 41 个,这样,可能的属性组合是 2^{41}种,不可能用穷举的方式去寻求最优方案。一种可行的方案是依靠专家经验给出重要属性的组成模式,但这样做仅依赖于专家的个人经验,具有一定的风险。为此,提出一种将遗传算法和 C4.5 算法结合起来的寻优模型,利用遗传算法强大的寻优能力进行属性约简,优化属性组合。

2　基于属性约简的决策树寻优模型

从目前技术看,求属性的最小约简是个 NP 难题,因此,如何求出近似最优属性约简就变的很有实际意义了。文中提出一种逐渐逼近的寻优模型,模型中将使用遗传算法[4]进行属性组合的寻优。模型如图 1 所示。

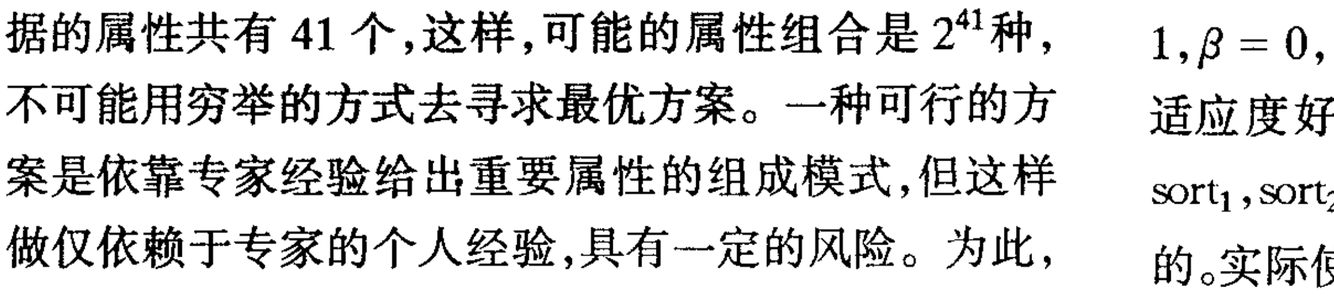

图 1　决策树寻优模型

模型的工作原理如下:首先随机产生初始种群,种群中每个个体都代表一种属性组合;第二,按照种群中每个个体所代表的属性组合准备训练数据,并将训练数据送往"决策树生成模块"生成决策树;第三,由"模式评估模块"计算每个决策树适应度,适应度使用文中定义的适应度函数(后面将详细说明)进行计算;第四,"遗传算法繁殖模块"将根据第三步得出的评价结果选择优良父辈并繁殖下一代,将新一代作为种群,并回到第二步。循环停止条件是达到规定的循环次数或产生的决策树满足规定的分类准确率要求。

这个模型的关键是确定适应度函数和选择算子。在遗传程序设计算法中,评价函数用来区分群体中个体(问题的解)的好坏,是种群中个体优劣的一种量化反映,它的构造直接影响问题求解的效率。适应度函数反映了对个体的评价倾向,遗传算法正是基于适应值对个体进行选择,以保证适应性好的个体基因有更多的机会遗传到下一代个体中,适应度函数的构造一般根据具体问题的不同而不同。

考虑到实际使用的需要,适应度函数应该综合考虑分类器的分类准确率、导出的规则数目及规则平均长度。基于以上考虑,定义适应度函数为:

$$\text{fitness}(T) = \alpha\text{sort}_1 + \beta\text{sort}_2 + \gamma\text{sort}_3$$

其中,sort_1,sort_2,sort_3 分别代表种群中某个决策树 T 在分类误报率、规则数量、规则平均长度上的排名,数值低者名次靠前。α,β,γ 分别代表 sort_1,sort_2,sort_3 的权重,且 $\alpha+\beta+\gamma=1$,$0\leq\alpha,\beta,\gamma\leq1$。例如,当 $\alpha=1$,$\beta=0$,$\gamma=0$,分类误报率的高低就成为判断决策树适应度好坏的唯一标准;当 $\alpha=\beta=\gamma$ 时,则表示 sort_1,sort_2,sort_3 对决策树适应度好坏来说是同等重要的。实际使用中可以根据使用者的需要,选择 α,β 和 γ 的值。选择算子的确定应该考虑在搜索空间范围和整个算法的收敛速度上取得平衡,为此,将采用基于适应度大小顺序的选择算子[5]。

3　实验及分析

实验所用训练和测试数据均来自 KDDCUP'99[6],共有 290 439 条数据,由 DOS 类入侵数据和 Normal 数据组成,每个样本数据由 41 个属性组成。从所有数据中随机抽取约一半数据 145 443 条作为训练集数据,剩余的 144 996 条数据做为测试数据。每个种群个体基因由 41 个 0 或 1 组成,1 代表对应属性被选中,0 代表对应属性未选中。使用 C4.5 Release 8[7]生成决策树。实验目的是建立针对 DOS 入侵的误用分类器。

实验分四组进行,然后将结果进行对照分析:第一次实验,作为后面三次实验的对照试验,没有进行属性约简和属性组合寻优,直接由训练集生成决策树,训练

数据包含全部 41 个属性;第二、三、四次实验,使用寻优模型进行属性约简和优化组合,具体实验参数如表 1 所示。

表 1 实验参数表

	种群规模	交叉概率	变异概率	选择压力	随机种子	适应度函数参数 (α,β,γ)	迭代代数(代)
第二次实验	30	0.8	0.05	5	0.05	(1.0,0.0,0.0)	100
第三次实验	30	0.8	0.05	5	0.05	(0.6,0.2,0.2)	100
第四次实验	30	0.8	0.05	5	0.05	(0.3333,0.3333,0.3333)	100

实验及参数说明:

(1)遗传算法迭代次数共 100 代,也可以根据情况增加迭代次数。

(2)为避免遗传算法过早局部收敛,将种群规模设置为 30 个个体,交叉概率为 0.8,变异概率为 0.05。

(3)实验均选择第 100 代中适应度排名前 10 名的个体分别建立决策树,并使用同一测试集测试决策树的性能。

(4)适应度函数的 3 个参数的不同设置,表示对种群中个体适应度好坏的评价倾向。例如,在第二次实验中,设置 $\alpha=1,\beta=0,\gamma=0$,表示个体好坏仅以分类准确度为准,不考虑规则数量和规则长度,因此这样的寻优属于不控制决策树规模的寻优。而第三、第四次实验中,逐渐降低准确度在适应度函数中的权重,提高规则数量和规则长度的权重,这种做法,有利于在不过分损失分类准确度的前提下,选择规则数量和规则长度理想的决策树,实际应用价值更高。

实验结果如表 2 所示。从实验结果可以看到:

(1)从分类精度分析。第一次实验中,未进行寻优,此时利用 C4.5 R8 依据训练集构建的决策树,对测试集预测分类的误报数为 71。而后三次实验均为寻优后、依据同一训练集构建决策树,对同一测试集的预测分类误报分别为 28.5,34.5 和 39.2,误报分别降低了 59.9%、51.4% 和 44.8%,而后三次实验优化后的属性数分别为 22.7,24.3 和 19.6。实验表明,训练样本集的属性个数并不是越多越好,经过寻优后,构建树所需的属性大大减少了,同时分类准确性却提高了。

(2)从分类规则规模分析。第三、第四次实验导出的规则数量和规则长度均低于第一次实验,尤其是规则数量大幅度减少了,这说明树的规模得到了有效控制。并且可以看到,树的规模减小的同时,分类的准确性依然远远高于第一次试验。这因为在后两次实验中,适应度函数的评价倾向既考虑分类精度,又控制树的规模所致。这也恰恰是第二次实验在分类精度提高很多的同时,所产生的树的规模却没有得到改善的根本原因所在。

表 2 实验结果数据

	属性总数	训练数据总数	测试集数据总数	训练集误报	测试集误报	修剪后的导出规则数	修剪后的平均规则长度
第一次实验	41	145443	144996	19	71	159	6.57
第二次实验	22.7	145443	144996	6.5	28.5	162.1	7.37
第三次实验	24.3	145443	144996	8.2	34.5	46.7	6.34
第四次实验	19.6	145443	144996	12.3	39.2	33.1	5.48

注:除第一次实验外,表中其它实验的结果数据均为第 100 代前 10 名分类器相应指标数据的平均值。规则长度指规则前件包含的属性个数。

4 总 结

文中所提出的决策树寻优模型是个自适应模型,通过对属性的约简和优化组合,能实现对决策树的优化,降低了躁声和不相关属性对构造决策树的影响,提高了决策树的分类准确性。另外,模型中提出的适应度函数,既考虑了分类器的准确率,又考虑了规则的规模(规则数量和规则平均长度),有利于提高分类器的分类速度和效率,因此更具实用价值。实验表明,所述的寻优模型对决策树的优化是有效的。

以上实验所用程序使用 VC++6.0 进行编码并调试通过,实验所用操作系统为 Windows XP。

未来的工作是:进一步研究适应度函数三个构成要素的内在联系;研究将文中所提出的决策树寻优模型运用于构造入侵检测系统任务中;改进程序算法,利用遗传算法天然的易于并行化的特点,提高学习效率,缩短学习时间。

参考文献:

[1] Han Jiawei, Kamber M. 数据挖掘概念与技术[M]. 范 明等译. 北京:机械工业出版社,2001.

[2] Quinlan J R. C4.5: Programs for Machine Learning[M]. [s. l.]: Morgan Kaufmann Publishers, 1993.

[3] Dunham M H. 数据挖掘教程[M]. 郭崇慧译. 北京:清华大学出版社.2005.

[4] 周 明,孙树栋. 遗传算法原理及应用[M]. 北京:国防工业出版社,1999.

[5] 任庆生,叶中行,曾 进,等. 对常用选择算子的分析[J]. 上海交通大学学报,2000,34(4):564-566.

[6] KDDCUP'99 data[J/OL]. 1999. http://kdd.ics.uci.edu/databases/kddcup99/.

[7] Quinlan J R. C4.5 Release 8[EB/OL]. 1992. http://www.rulequest.com/Personal/.