文章编号: 1006-2467(2010)07-0883-04+0891

ID3 算法的改进和简化

朱 颢 东[1,2,3]

(1. 郑州轻工业学院 计算机与通信工程学院, 郑州 450002;
2. 中国科学院 成都计算机应用研究所, 成都 610041; 3. 中国科学院 研究生院, 北京 100039)

摘 要: 针对ID3 算法倾向于选择取值较多的属性的缺点, 引进属性重要性来改进ID3 算法, 并根据改进的 ID3 算法中信息增益的计算特点, 利用凸函数的性质来简化该算法. 实验表明, 优化的 ID3 算法与原 ID3 算法相比, 在构造决策树时具有较高的准确率和更快的计算速度, 并且构造的决策树还具有较少的平均叶子数.

关键词: 决策树; ID3 算法; 属性重要性; 信息增益; 凸函数
中图分类号: TP 301 文献标志码: A

Research on Improvement and Simplification of ID3 Algorithm

ZHU Hao-dong[1,2,3]

(1. School of Computer and Communication Engineering, Zhengzhou University of Light Industry, Zhengzhou 450002, China;
2. Chengdu Institute of Computer Application, Chinese Academy of Sciences, Chengdu 610041, China;
3. The Graduate School of the Chinese Academy of Sciences, Beijing 100039, China)

Abstract: For the shortcoming that ID3 algorithm tends to choose attribute which has many values, attribute importance was introduced to improve ID3 algorithm. Next, according to the character of information gain, the improved ID3 algorithm was simplified to reduce the complexity of computing information gain by the convex function. Through experiment testing, the optimized ID3 algorithm can spend much less time to construct the high accurate decision tree and this decision tree has less average leaves.

Key words: decision tree; ID3 algorithm; attribute importance; information gain; convex function

决策树分类是一种十分重要的数据挖掘方法, 在众多决策树构造方法中, ID3 算法最具影响力[1], 人们对其进行了深入研究[2-10]. 在此基础上, 本文从解决 ID3 算法倾向于选择取值较多的属性的缺点以及简化 ID3 算法的角度进行了研究.

1 ID3 算法的不足及其改进

ID3 算法的核心计算公式为[1]

$$\mathrm{Gain}(A) = I(s_1, s_2, \cdots, s_m) - E(A) \qquad (1)$$

研究表明[2-10]: 式(1)倾向于选择分支值较多的属性作为分支节点, 而分支值较多的属性却并不总是最好的属性. 例如: Bratko 研究小组在研究判断病情的各种因素时, 用 ID3 确定"病人的年龄(有 9 种值)"为应该首先选择的属性, 但在实际情况中, 医学专家却认为这个属性在判断病情时没那么重要[3]. 文献[3-7]中在 ID3 算法里引入一个用户兴趣度, 取值为[0, 1], 具体由用户根据经验给出, 虽然能在一定程度上克服该算法的缺点, 但在使用用户兴

收稿日期: 2009-09-24
基金项目: 四川省科技计划项目(2008GZ0003); 四川省科技厅科技攻关项目(07GG006-019)
作者简介: 朱颢东(1980-), 男, 河南虞城人, 博士生, 主要研究方向为软件过程技术与方法、文本挖掘、智能信息处理.

趣度时需注意:

(1) 对用户感兴趣的属性,可根据先验知识或领域知识进行测试,选择符合实际情况的用户兴趣度.

(2) 当大多数属性数据量较大、个别属性数据量较小,且人们对这些属性重要性认识不足时,选择这些属性的用户兴趣度,使其不会出现大数据掩盖小数据的现象.

(3) 决策树中的属性如果许多有先验知识或领域知识,可根据实际情况选择用户兴趣度,但不宜做太多选择,可以逐步进行,否则会因人为因素影响决策效果.

由上可知,兴趣度的取值全靠用户经验,很难反映事实,并且对那些不熟悉的用户来说,更不能较好地选择这个兴趣度,从而使得这些改进算法不能较好地被使用.本文在上述研究的基础上,把粗集理论用于 ID3 算法,用客观属性重要度来代替全靠经验确定的主观用户兴趣度,从而使得该参数的确定更具说服力,以克服 ID3 算法的上述缺点.

通过研究式(1)发现,对于

$$I(s_1, s_2, \cdots, s_m) = -\sum_{i=1}^{m} P_i \log P_i \tag{2}$$

只要给定了训练样本和类别数,它在整个决策树构造中始终为定值,为了减少计算量,改善算法的时间复杂度,可以把它从式(1)中摘除,此时只保留$E(A)$即可.

在$E(A)$中,

$$I(s_{1j}, s_{2j}, \cdots, s_{mj}) = -\sum_{i=1}^{m} P_{ij} \log P_{ij}$$

的计算十分耗时,因为每次都要计算多个 $\log P_{ij}$,而计算 $\log P_{ij}$ 很耗时.因此,本文研究了 $\log P_{ij}$ 函数的基本性质,证明了 $\log P_{ij}$ 是一种上凸函数,利用上凸函数的特性,对信息量计算公式进行简化,从而达到优化的目的.

2 ID3 算法的改进和简化

2.1 ID3 算法的改进

定义 1 属性依赖度[9-10].给定一个决策表 $S=\langle U, C, D, V, f\rangle$,存在一属性集 R, $R\subseteq C$,决策属性 D 对于属性集 R 的依赖度定义为

$$\gamma(R, D) = \frac{\mathrm{Card}(\mathrm{POS}_R(D))}{\mathrm{Card}(U)} \tag{3}$$

式中:U 为数据对象集;C 为条件属性集;D 为决策属性集;V 为全部属性的值集;f 为一个映射函数;$\mathrm{Card}(U)$为模糊集中的标准函数,代表 U 的对象个数;$\mathrm{POS}_R(D)$为粗糙集中的标准正域函数.

显然有 $0\leqslant\gamma(R, D)\leqslant 1$, $\gamma(R, D)$给出了决策属性 D 与属性集 R 之间相依性的一种测度.它反映了属性集 R 对于决策 D 的重要程度.在已知 R 的前提下,一个属性 $a\in R$ 对于决策 D 的重要性定义为

$$\mathrm{SGF}(a) = \gamma(R, D) - \gamma(R-\{a\}, D) \tag{4}$$

$\mathrm{SGF}(a)$表明了把属性 a 从 R 中去掉后对分类决策的影响程度,即不能被正确分类样本所占的比例.将式(4)引入式(1),可得改进公式:

$$\mathrm{Cain}(A) = I(s_1, s_2, \cdots, s_m) - \sum_{j=1}^{n} \frac{s_{1j}+s_{2j}+\cdots+s_{mj}}{s} \mathrm{SGF}(A) \sum_{i=1}^{m} P_{ij} \log P_{ij} \tag{5}$$

式(5)在解决大数据量覆盖小数据量方面具有一定优势,而且还降低了属性值较多又不是很重要属性的地位,解决了 ID3 算法偏向于选择分支值较多属性的缺陷.

2.2 ID3 算法的简化

在式(5)中,根据前述理由可以把式(2)从式(5)中摘除,此时式(5)只保留后一项,而不影响最终结果.可以进一步改进为

$$E^{*}(A) = -\sum_{j=1}^{n} \frac{s_{1j}+s_{2j}+\cdots+s_{mj}}{s} \times \mathrm{SGF}(A) \sum_{i=1}^{m} P_{ij} \log P_{ij} \tag{6}$$

式(6)中函数 $\log P_{ij}$ 是一个类似于 $\log x$ 的对数函数,其中 $P_{ij}\in[0, 1]$ 是 s_j 中样本属于类 C_i 的概率.对于 $\forall x_1, x_2\in[0, 1]$,若满足 $x_1 - x_2 = \Delta x \to \alpha(0)$时,$\log x$ 函数在$[0, 1]$上连续,根据文献[11]可以证明该函数的凹凸性,相应地,也就证明了函数 $\log P_{ij}$的凹凸性.

性质 1[11] $\log x$ 函数在$[0, 1]$上是凸函数.

性质 2[11] 设 $f(x)$为凸函数,X 为非空集合,$x_i\in X$, $\exists\lambda_i\geqslant 0$, $\sum_{i=1}^{m}\lambda_i = 1$,则

$$\sum_{i=1}^{m}\lambda_i f(x_i) \leqslant f\left(\sum_{i=1}^{m}\lambda_i x_i\right) \tag{7}$$

根据性质 1 可知,$\log P_{ij}$ 在$[0, 1]$上为凸函数,因此,对于式(6)中的 $-\sum_{i=1}^{m} P_{ij}\log P_{ij}$ 部分,根据性质 2,可以转换为

$$-\sum_{i=1}^{m} P_{ij} \log P_{ij} \geqslant -\log\left(\sum_{i=1}^{m} P_{ij}^2\right)$$

此时,式(6)可用下式代替:

$$E(A)^{*} = -\sum_{j=1}^{n} \frac{s_{1j}+s_{2j}+\cdots+s_{mj}}{s} \times \mathrm{SGF}(A) \log\left(\sum_{i=1}^{m} P_{ij}^2\right) \tag{8}$$

式(8)优点:按原 ID3 算法需要对 $\log P_{ij}$ 计算 m 次,而利用简化的 ID3 算法只需要对 $\log P_{ij}$ 计算 1 次,就可以得出信息量的近似值,减少了 $m-1$ 个 $\log P_{ij}$ 函数值的计算过程,极大提高了决策树构造的计算效率.

由于 $P_{ij}=\frac{s_{ij}}{s_j}$, $\sum_{i=1}^{m}s_{ij}=s_j$, 则

$$\sum_{i=1}^{m}P_{ij}^2=\sum_{i=1}^{m}\left(\frac{s_{ij}}{s_j}\right)^2=\frac{1}{s_j^2}\sum_{i=1}^{m}s_{ij}^2=\left(\frac{1}{s_j^2}\left[\sum_{i=1}^{m}s_{ij}\right)^2-\sum_{i=1}^{m}\sum_{l=1\wedge l\neq i}^{m}(s_{ij}s_{lj})\right]=\left(\frac{1}{s_j}\sum_{i=1}^{m}s_{ij}\right)^2-\sum_{i=1}^{m}\sum_{l=1\wedge l\neq i}^{m}\left(\frac{s_{ij}}{s_j}\frac{s_{lj}}{s_j}\right)=1-\sum_{i=1}^{m}\sum_{l=1\wedge l\neq i}^{m}(P_{ij}P_{lj})$$

因此,有

$$\log\left(\sum_{i=1}^{m}P_{ij}^2\right)=\log\left[1-\sum_{i=1}^{m}\sum_{l=1\wedge l\neq i}^{m}(P_{ij}P_{lj})\right]=\frac{\ln\left[1-\sum_{i=1}^{m}\sum_{l=1\wedge l\neq i}^{m}(P_{ij}P_{lj})\right]}{\ln 2}$$

又因为当 $x\to 0$ 时,函数 $\ln(1+x)\to x$,而 $-\sum_{i=1}^{m}\sum_{l=1\wedge l\neq i}^{m}(P_{ij}P_{lj})\to 0$,所以有

$$\frac{\ln\left[1-\sum_{i=1}^{m}\sum_{l=1\wedge l\neq i}^{m}(P_{ij}P_{lj})\right]}{\ln 2}\to-\frac{\sum_{i=1}^{m}\sum_{l=1\wedge l\neq i}^{m}(P_{ij}P_{lj})}{\ln 2}$$

因此,式(8)变为

$$E^{*}(A)=\sum_{j=1}^{n}\frac{(s_{1j}+s_{2j}+\cdots+s_{mj})}{s}\times\mathrm{SGF}(A)\frac{\sum_{i=1}^{m}\sum_{l=1\wedge l\neq i}^{m}(P_{ij}P_{lj})}{\ln 2}\qquad(9)$$

由于 ln 2 为常数,故可以从式(9)中去掉而不影响最终结果,此时,式(9)可以变为

$$E(A)^{*}=\sum_{j=1}^{n}\frac{(s_{1j}+s_{2j}+\cdots+s_{mj})}{s}\times\mathrm{SGF}(A)\sum_{i=1}^{m}\sum_{j=1\wedge j\neq i}^{m}P_{ij}^2\qquad(10)$$

同式(8)相比,式(10)完全消除了比较耗时的 $\log P_{ij}$ 函数计算,仅由加、减、乘、除 4 种简单计算方式构成,大大减少了计算机的运行时间,提高了判别速度.采用式(10)计算每个属性的平均熵,从中选取熵值最小的属性作为结点属性,由此可获得一个优化的 ID3 算法.

3　实例验证

3.1　准确度与叶子数目的对比实验

实验中以 UCI 提供的部分数据集为实验数据,采用文献[12]算法对实验数据进行离散化,使用 10 层交叉法测试决策树的平均分类精度.同时,为了比较优化 ID3 算法和传统 ID3 算法的复杂性,本文将每个数据集分成 10 份形成 10 个子集,每个子集中的数据随机抽取,在每一个数据子集上分别用两算法构造决策树,然后统计 10 次构造决策树中平均叶子结点数.表 1 是各数据集的分布情况,表 2 所示为实验结果.

表 1　数据集

Tab. 1　The data sets

数据集名称	样本数	D	C	数据集名称	样本数	D	C
Breast	817	2	9	Lymph	189	4	18
Diabetes	798	2	8	Bupa	428	2	6
Iris	209	3	4	Segmentation	2 932	7	19

表 2　实验结果

Tab. 2　The experimental results

数据集名称	传统 ID3 算法		优化 ID3 算法	
	平均准确度/%	叶子数	平均准确度/%	叶子数
Breast	85.8	8.0	90.5	6.2
Diabetes	71.5	16.5	79.2	10.8
Iris	73.1	7.0	78.7	5.0
Lymph	73.2	8.5	77.8	7.2
Bupa	82.8	10.0	88.9	7.8
Segmentation	73.2	10.8	84.1	8.1
平均值	78.8	10.1	83.2	7.5

由表 2 可见,与传统 ID3 算法相比,优化 ID3 方法有较好的平均分类精确度,同时构造的决策树叶子数较少,从而有较低的复杂性.

3.2　生成决策树时间的对比实验

以文献[1]中“天气表”的 14 条记录为基础,首先随机生成多条记录以组成多个数据集,然后在每一个数据集中,分别对传统 ID3 算法和优化 ID3 算法进行 20 次计算时间的测试,取其平均值作为算法构造决策树花费的计算时间.其实验结果如图 1 所示.由图可见,在不同规模的数据集中,优化 ID3 算法构造决策树所用计算时间比传统 ID3 算法构造决策树所用的时间少,这说明使用优化 ID3 算法能够以更高的效率构造决策树.

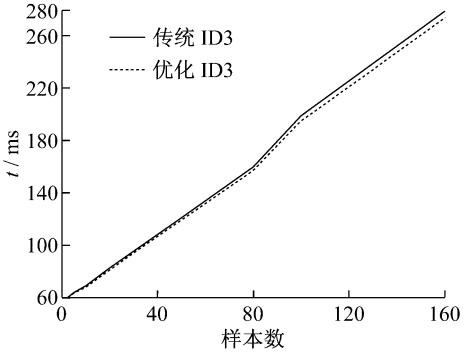


图 1 两算法执行时间对比

Fig. 1 Comparison of the two algorithms execution time

根据图 1, 可获得优化 ID3 算法相比于传统 ID3 算法所节省的时间 Δt 以及相应的时间节省率 η, 如图 2 所示. 由图可见, 在构造决策树的过程中, 随着数据集规模的增大, 优化 ID3 算法与传统 ID3 算法相比节省的计算时间增多, 优化 ID3 算法的高效性越明显.

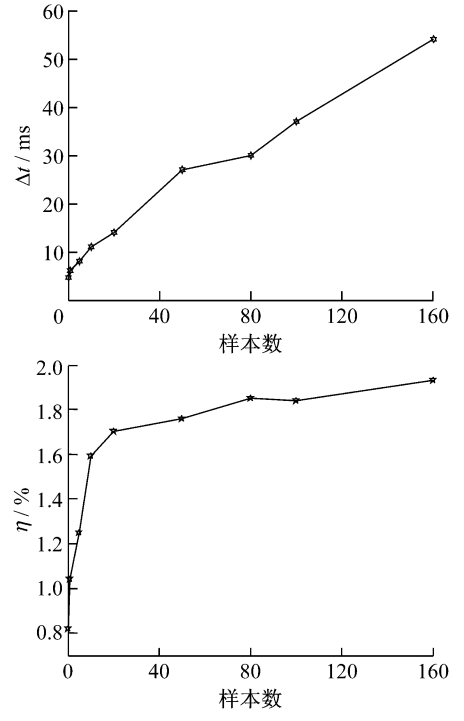


图 2 优化的 ID3 算法节省时间和时间节省率随数据集变化趋势

Fig. 2 The saving time trend and saving time rate trend of optimized ID3 with data sets

4 结 语

与传统的 ID3 算法相比, 优化的 ID3 方法有较好的平均分类精确度, 同时构造的决策树叶子数较少, 从而有较低的复杂性. 在相同规模的数据集中, 优化 ID3 算法构造决策树所用的计算时间比传统 ID3 算法构造决策树所用的计算时间少, 充分说明优化 ID3 算法提高了决策树构造的效率. 特别是随着数据规模的增大, 算法的效率和性能就越好, 算法的优越性越明显.

参考文献:

[1] Quinlan J R. Induction of decision tree[J]. **Machine Learning**, 1986, 4(2): 81-106.

[2] 刘小虎, 李 生. 决策树的优化算法[J]. 软件学报, 1998, 10(9): 797-800.
LIU Xiao-hu, LI Sheng. The decision tree optimization algorithm[J]. **Journal of Software**, 1998, 10(9): 797-800.

[3] 张桂杰, 王 帅. 决策树分类 ID3 算法研究[J]. 吉林师范大学学报(自然科学版), 2008, 29(3): 135-137.
ZHANG Gui-jie, WANG Shuai. Decision tree classification[J]. **Jilin Normal University Journal (Natural Science Edition)**, 2008, 29(3): 135-137.

[4] 孙爱东, 朱梅阶, 涂淑琴. 基于属性值的 ID3 算法改进[J]. 计算机工程与设计, 2008, 29(12): 3011-3012.
SUN Ai-dong, ZHU Mei-jie, TU Shu-qin. Improved ID3 algorithm based on attribute values[J]. **Computer Engineering and Design**, 2008, 29(12): 3011-3012.

[5] 鲁 为, 王 枞. 决策树算法的优化与比较[J]. 计算机工程, 2007, 33(16): 189-190.
LU Wei, WANG Cong. Optimization and comparison of decision tree algorithm[J]. **Computer Engineering**, 2007, 33(16): 189-190.

[6] 乔 梅, 韩文秀. Rough 集中属性分类贡献能力综合测度研究[J]. 系统工程学报, 2006, 21(5): 508-614.
QIAO Mei, HAN Wen-xiu. Research on syntheses measure of attribute classification contribution ability in rough set[J]. **Journal of Systems Engineering**, 2006, 21(5): 508-614.

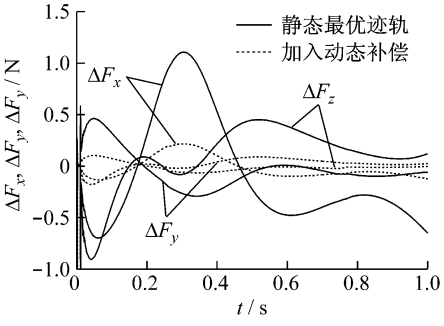
[7] 王静红, 王熙照, 邵艳华, 等. 决策树算法的研究及优化[J]. 微机发展, 2004, 14(9): 30-32.
WANG Jing-hong, WANG Xi-zhao, SHAO Yan-hua, *et al*. Research and optimization of decision tree algorithm[J]. **Microcomputer Development**, 2004, 14(9): 30-32.

[8] 张 彦, 刘瞰东, 李茂青. 基于信息论的决策树算法探讨[J]. 自动化技术与应用, 2006, 25(1): 4-7.
ZHANG Yan, LIU Tun-dong, LI Mao-qing. Decision tree algorithm based on the information theory [J]. **Techniques of Automation and Applications**, 2006, 25(1): 4-7.

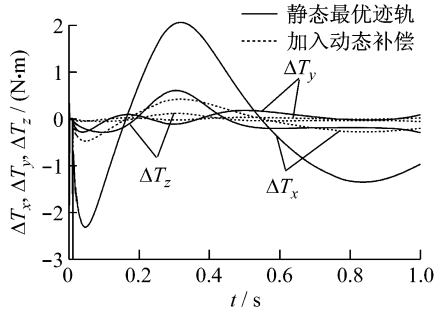
[9] 蒋 芸, 李战怀, 张 强, 等. 一种基于粗糙集构造决策树的新方法[J]. 计算机应用, 2004, 24(8): 21-23.

(下转第 **891** 页)

优轨迹时误差值被控制在目标广义力的 0.090‰以内. 虽然迭代过程精度为0.003‰, 但由于存在惯性力、哥氏力等因素, 其误差值较大. 在加入动态补偿轨迹后, 其误差降低至 0.02%以内. 可见: 补偿曲线能够很好地抵消机构动力学的影响; 同时, 说明对模型的简化合理有效. 另外, 离散点轨迹求解的平均时间为4 ms, 完全能够满足实时控制的要求.



(a) 力误差



(b) 力矩误差

图 6　广义力误差轨迹

Fig. 6　Error track of wrench

4　结　语

本文建立了主从式双并联十二自由度广义负载仿真系统的动力学模型, 并以广义力为优化目标进行了轨迹规划. 将优化轨迹分解为静态最优轨迹和动态补偿轨迹. 静态最优轨迹建立了静态下目标广义力和位姿的映射关系, 并结合牛顿迭代法和静力学给出了迭代流程; 补偿轨迹则将最优轨迹离散化, 并在离散点领域内根据线性化动力学模型而进行轨迹补偿. 仿真结果表明, 该方法具有较好的合理性、有效性及实时性.

参考文献:

[1] 焦宗夏, 华　清, 王晓东. 负载模拟器的评价指标体系[J]. 机械工程学报, 2002, 38(11): 26-30.
JIAO Zong-xia, HUA Qing, WANG Xiao-dong. Estimation for performance of load simulator[J]. **Chinese Journal of Mechanical Engineering**, 2002, 38(11): 26-30.

[2] 王益群, 王燕山. 气液联合多通道同步加载比例复合控制系统[J]. 机械工程学报, 2005, 41(10), 180-184.
WANG Yi-qun, WANG Yan-shan. Pneumatic-hydraulic multi-channel synchronized loading proportional compound control system[J]. **Chinese Journal of Mechanical Engineering**, 2005, 41(10): 180-184.

[3] Raath A D, Van Waveren C C. A time domain approach to load reconstruction for durability testing[J]. **Engineering Failure Analysis**, 1998, 5(2): 113-119.

[4] Zhang S Y, Han J W. Application of μ theory in compliant force control[J]. **Chinese Journal of Aeronautics**, 2006, 19(1): 89-96.

[5] Huang Q T, Jiang H Z. Spacecraft docking simulation using hardware in the loop simulator with Stewart platform[J]. **Chinese Journal of Mechanical Engineering**, 2005, 18(3): 415-418.

[6] 王宣银, 吴　剑, 吴乐彬. 基于并联六自由度电液伺服机构的力控制方法[J]. 上海交通大学学报, 2007, 41(1): 111-115.
WANG Xuan-yin, WU Jian, WU Le-bin. The force control method based on 6-DOF parallel electro-hydraulic servo mechanism[J]. **Journal of Shanghai Jiaotong University**, 2007, 41(1): 111-115.

[7] 吴江宁. 并联式六自由度平台及其控制研究[D]. 杭州: 浙江大学机电系, 1996.

(上接第 **886** 页)

JIANG Yun, LI Zhan-huai, ZHANG Qiang, *et al*. New method for constructing decision tree based on rough sets theory[J]. **Computer Applications**, 2004, 24(8): 21-23.

[10] 倪春鹏, 王正欧. 一种新型决策树属性选择标准[J]. 武汉科技大学学报(自然科学版), 2004, 27(4): 437-441.
NI Chun-peng, WANG Zheng-ou. A new attribute selection criterion of decision tree[J]. **Journal of Wuhan University of Science and Technology (Natural Science Edition)**, 2004, 27(4): 437-441.

[11] 同济大学应用数学系. 高等数学: 上册[M]. 第 5 版. 北京: 高等教育出版社, 2004.

[12] Hu X, Cercone N. Data mining via generalization, discrimination and rough set feature selection[J]. **Knowledge and Information System: An International Journal**, 1999, 1(1): 21-27.