文章编号:1000-5862(2018)04-0384-05

调和平均优化选择划分属性的决策树改进算法

王 卓[1],聂 斌[2*],罗计根[2],杜建强[2],陈 爱[1],周 丽[2]
(1. 南昌大学软件学院,江西 南昌 330047;2. 江西中医药大学计算机学院,江西 南昌 330004)

摘要:针对信息增益和信息增益率对属性取值数的偏好,提出了一种调和平均优化选择划分属性的决策树改进算法. 首先计算候选划分属性的信息增益,找出信息增益高于平均水平的属性,然后分别计算这些属性的信息增益率和信息增益的调和平均值,从中筛选调和平均值最大的属性,建立分支决策,并用递归方法建立决策树. 通过 4 份不同规模数据实验,利用信息增益、信息增益率、GINI 指数以及该文提出的方法作为属性划分的标准,分别考察其准确性在训练集、测试集、10 次 10 折交叉验证(或 5 次 5 折交叉验证),以及其平均值. 实验结果表明:该方法准确性较好、运行时间较短,具有一定程度的优越性.

关键词:决策树;信息增益率;调和平均;中医药信息
中图分类号:TP 391 **文献标志码**:A **DOI**:10.16357/j.cnki.issn1000-5862.2018.04.11

0 引言

信息增益准则对可取值数目较多的属性有所偏好[1-2],为减少这种偏好可能带来的不利影响,C4.5 算法[3]不直接使用信息增益,而是使用“信息增益率”选择最优划分属性. 然而,信息增益率准则对可取值数目较少的属性有所偏好[1]. 因此,C4.5 算法先是在候选划分属性中找出信息增益高于平均水平的属性,再从中选择信息增益率最高的. 随后有 CART、PUBLIC、SPRINT 算法等对该问题进行研究,另外,也有学者从剪枝的角度优化决策树,还有一些研究者从其它方面改进决策树算法[4-16],取得了一定的效果. 本文研究发现,为了调和所选划分属性受属性取值多少的影响,提出一种调和平均优化选择划分属性的算法,先计算信息增益率和信息增益的调和平均值,从中筛选平均值最大的属性,作为划分属性.

1 决策树选择划分属性的方法

1.1 信息熵、信息增益与信息增益比

在信息论与概率统计中,熵(entropy)[17]是表示随机变量不确定性的度量. 设 X 是一个取有限个值的离散随机变量,其概率分布为 $P(X=x_i)=p_i, i=1,2,\cdots,n$,则随机变量 X 的熵定义为 $H(X)=-\sum_{i=1}^{n} p_i \log p_i$,其中,若 $p_i=0$,则定义 $0\log 0=0$. 对数若以 2 为底,则熵的单位为比特,若以自然对数为底,熵的单位为纳特. 由定义可知,熵只依赖于 X 的分布,而与 X 的取值无关,所以熵 $H(X)$ 或 $H(p)$ 可表示为 $H(p)=-\sum_{i=1}^{n} p_i \log p_i$,熵越大,随机变量的不确定性就越大.

信息增益[17]表示已知属性 X 的信息使得类 Y 的信息的不确定性减少的程度. 设属性 A 对训练数据集 D 的信息增益 $G_{ain}(D,A)$,定义为集合 D 的经验熵 $H(D)$ 与属性 A 给定条件下 D 的经验条件熵 $H(D|A)$ 之差:

$$G_{ain}(D,A)=H(D)-H(D|A). \quad (1)$$

以信息增益作为划分训练数据集的属性,存在偏向于选择取值较多的属性的问题,从而引入信息增益比对此问题进行校正.

设属性 A 对训练数据集 D 的信息增益比 $G_{ainRatio}(D,A)$,定义为其信息增益 $G_{ain}(D,A)$ 与训练数据集 D 关于属性 A 的值的熵 $H_A(D)$ 之比

收稿日期:2017-08-06
基金项目:国家自然科学基金(61562045,61363042),江西省自然科学基金重大项目(20152AXCB20007),江西省高校科技落地计划(LD12038),江西省教育科学“十二五”规划一般课题(15YB005)和江西中医药大学自然科学基金(2013ZR0068)资助项目.
通信作者:聂 斌(1972-),男,江西峡江人,副教授,主要从事中医信息学、数据挖掘和人工智能方面的研究. E-mail:864860723@qq.com

$$G_{ainRatio}(D,A) = G_{ain}(D,A)/H_A(D). \quad (2)$$

1.2　平均信息增益

需要注意的是,信息增益率准则对可取值数目较少的属性有所偏好,因此 C4.5 算法并不是直接选择信息增益率最大的候选划分属性,而是使用了一个启发式算法,即先从候选划分属性中找出信息增益高于平均水平的属性,再从中选择信息增益率最高的属性[1-2].

计算属性的全部未用属性平均信息增益方法为

$$A_{VERGain(D,A)} = \frac{1}{v}\sum_{i=1}^{v}(I(p,n) - E(A)). \quad (3)$$

2　信息增益和信息增益率的调和平均值

信息增益和信息增益率的调和平均值(Harmonic Average of Information Gain and Information Gain Rate,HAIGIGR)为

$$H_{AIGIGR}(A) = 2G_{ain}(D,A)G_{ainRatio}(D,A)/(G_{ain}(D,A) + G_{ainRatio}(D,A)). \quad (4)$$

3　调和平均优化选择划分属性的决策树改进算法

为了尽量减少信息增益和信息增益率对属性取值数的偏好,提出一种调和平均优化信息增益与信息增益率选择属性的决策树改进算法. 该算法的基本思想为:(i) 计算候选划分属性的信息增益,找出信息增益高于平均水平的属性;(ii) 分别计算这些属性的信息增益率和信息增益的调和平均值,从中筛选平均值最大的属性,建立分支决策;(iii) 用递归方法建立决策树.

3.1　算法描述

调和平均优化选择划分属性的决策树改进算法描述为

输入:训练数据集 D,属性集 A,阈值 ε;

输出:决策树 T.

1) 若 D 中所有实例属于同一类 C_j,则置 T 为单结点树,并将 C_j 作为该结点的类,返回 T;

2) 若 $A = \varnothing$,则置 T 为单结点树,并将 D 中实例最大的类 C_j 作为该结点的类,返回 T;

3) 否则,按(1) 式计算 A 中各属性对 D 的信息增益,并从大到小排序;

4) 按(3) 式计算 A 中各属性对 D 的信息增益的算术平均值;

5) 结合第 3)、第 4) 步,删除小于信息增益平均值的属性信息增益;

6) 按(1) ~ (4) 式求剩下属性的信息增益和信息增益率的调和平均值 $H_{AIGIGR}(A)$,选择该值最大的属性 A_g;

7) 若 A_g 的值小于阈值 ε,则置 T 为单结点树,并将 D 中实例数最大的类 C_j 作为该结点的类,返回 T;

8) 否则,对 A_g 中的每一可能值 a_i,依 $A_g = a_i$ 将 D 分割为子集,即若干非空 D_i,将 D_i 中实例数最大的类作为标记,构建子结点,由结点及其子结点构成树 T,返回 T;

9) 对结点 i,以 D_i 为训练集,以 $A - \{A_g\}$ 为属性集,递归地调用第 1) 步 ~ 第 8) 步,得到子树 T_i,返回 T_i.

3.2　算法的理论分析

调和平均优化信息增益与信息增益率选择属性生成决策树,减少了对属性取值多与取值少的偏好,从而尽可能地减少过拟合现象,在一定程度上可以提高分类准确性和提升生成决策树的效率.

4　实验验证

为了验证调和平均优化选择划分属性的决策树改进算法的有效性,选取了 ID3、C4.5,以及 CART 3 个经典算法进行比较.

4.1　实验数据及说明

实验采用了 4 份数据,包括 Iris 数据集(http://archive.ics.uci.edu/ml/datasets/Iris),Breast Cancer Wisconsin 数据集(http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Original%29),天气数据集(http://blog.csdn.net/czp11210/article/details/51161531),QSAR biodegradation Data Set(http://archive.ics.uci.edu/ml/datasets/QSAR+biodegradation).

Iris 数据集:条件属性(自变量)4 个,分类属性(因变量)1 个,样本总数为 150 条.

Breast Cancer Wisconsin 数据集:条件属性有肿块厚度、细胞大小的均匀性、细胞形状的均匀性、边际附着力等 9 个,分类属性 1 个,样本总数为 699 条.

天气集数据:4 个自变量,并且全部为离散型自变量,因变量 1 个,样本总数 13 条.

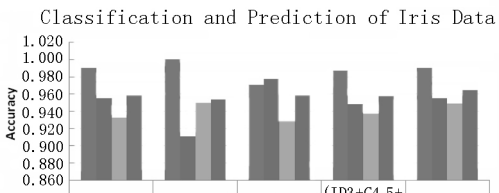
QSAR biodegradation Data Set:41 个自变量(属性),样本总数 1 055 条.

以上数据集实验验证时,按 7∶3 比例,随机将样本分成训练集和测试集建模,叶子节点最大容许个数为 3;叶子节点最大容许误差为 0.001. 另外,天气数据集样本较少,用 5 折交叉验证的方式进行建模,其它数据集均采用 10 折交叉验证的方式进行建模.

4.2 实验结果

计算设备配置为:AMD A10-8700P Radeon R6, 10Compute Cores 4C + 6G 1.80 GHz,RAM 8.00GB, 64 位操作系统. 对以上 4 份实验数据,利用信息增益和信息增益率的调和平均值(Harmonic Average of Information Gain and Information Gain Rate,HAIGIGR)、信息增益(Information Gain, IG)、信息增益率(Information Gain Rate, IGR)以及 GINI 指数作为属性划分的标准,建立决策树模型. 分别就不同数据集的分类准确性、计算时间分析等方面进行比较分析,比较结果如图 1 ~ 图 7 所示,其中 Training set 为训练集,Test set 为测试集,10 times 10-fold cross validation 为 10 次 10 折交叉验证,5 times 5-fold cross validation 为 5 次 5 折交叉验证,(ID3 + C4.5 + CART) / 3 为 ID3、C4.5、CART 3 种方法结果的平均数,The above average 为训练集、测试集、10 次 10 折交叉验证 3 种测试结果的平均数,computing time(ms)为计算时间.

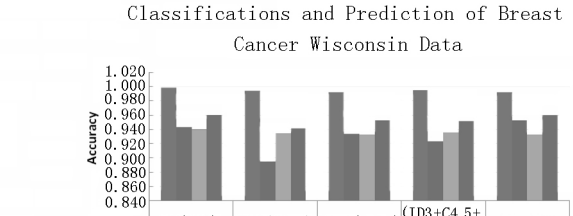
4.2.1 分类准确性分析 Iris 数据集的实验结果如图 1 所示,实验结果表明:(i)在训练集、测试集及 10 次 10 折交叉验证单方面,4 种属性划分的准确性均较高,稍各有所长;(ii)在训练集、测试集及 10 次 10 折交叉验证的平均数方面,HAIGIGR 在 3 种测试方法中的准确性平均值最高;(iii)对信息增益、信息增益率以及 GINI 指数 3 者结果的平均值而言,HAIGIGR 的准确性最高.



	IG(ID3)	IGR(C4.5)	Gini(CART)	(ID3+C4.5+CART)/3	HAIGIGR
Training set	0.991	1.000	0.971	0.987	0.991
Test set	0.956	0.911	0.978	0.948	0.956
10 times 10-fold cross validation	0.933	0.950	0.929	0.937	0.949
The above average	0.960	0.954	0.960	0.958	0.965

图 1 鸢尾花数据集的分类预测结果比较

Breast Cancer Wisconsin 数据集的实验结果如图 2 所示,实验结果表明:(i)在训练集、测试集及 10 次 10 折交叉验证单方面,4 种属性划分的准确性都较高,且各有所长;(ii)在训练集、测试集及 10 次 10 折交叉验证的平均数方面,HAIGIGR 在 3 种测试方法中,测试集及 10 次 10 折交叉验证的平均值最高;(iii)对信息增益、信息增益率以及 GINI 指数 3 者结果的平均值而言,HAIGIGR 的仅次于信息增益.



	IG(ID3)	IGR(C4.5)	Gini(CART)	(ID3+C4.5+CART)/3	HAIGIGR
Training set	0.998	0.994	0.992	0.995	0.992
Test set	0.943	0.895	0.933	0.924	0.952
10 times 10-fold cross validation	0.939	0.934	0.933	0.935	0.933
The above average	0.960	0.941	0.953	0.951	0.959

图 2 乳腺癌数据集的分类预测结果比较

天气集数据集的实验结果如图 3 所示,结果表明:(i)在训练集、测试集及 10 次 10 折交叉验证单方面,4 种属性划分的准确性都不稳定,在训练集上较高,在测试集波动较大,HAIGIGR 在 5 次 5 折交叉验证方面最好;(ii)在训练集、测试集及 5 次 5 折交叉验证的平均数方面,HAIGIGR 方法准确性偏低;(iii)对信息增益、信息增益率以及 GINI 指数 3 者结果的平均值而言,HAIGIGR 的准确性在训练集持平,在测试集较低,在 5 次 5 折交叉验证最高.

QSAR biodegradation Data Set 的实验结果如图 4,结果表明:(i)在训练集、测试集及 10 次 10 折交叉验证单方面,HAIGIGR 在训练集上准确性仅次于信息增益,在测试集及 10 次 10 折交叉验证方面较高;(ii)在训练集、测试集及 10 次 10 折交叉验证的平均数方面,HAIGIGR 在 3 种测试方法中的准确性最高;(iii)在信息增益、信息增益率以及 GINI 指数三者结果的平均值而言,HAIGIGR 的准确性最高.

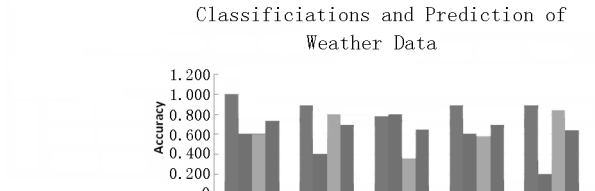
4.2.2 计算时间分析 Iris 数据集. 采用 4 种属性划分标准建立决策树、计算其准确性、生成决策树图等,HAIGIGR 方法的运行时间仅为 31 ms,是 4 种方法中时间最少的,且优势明显. 实验结果如图 5 所示.

Breast Cancer Wisconsin 数据集. 采用 4 种属性划分标准建立决策树、计算其准确性、生成决策树图等,HAIGIGR 方法的运行时间仅为 47 ms,仅高于信息增益方法,优于其它 3 种方法,且优势明显. 实验结果如图 6 所示.

天气集数据集. 采用 4 种属性划分标准建立决策树、计算其准确性、生成决策树图等,4 种方法使

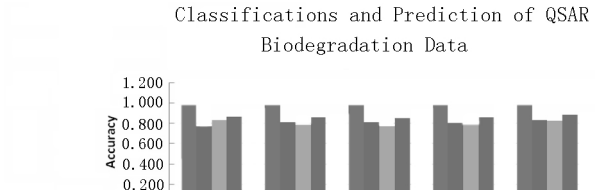
用时间都较短,接近于 0.

QSAR biodegradation Data Set. 采用 4 种属性划分标准建立决策树、计算其准确性、生成决策树图等,HAIGIGR 方法的运行时间仅略高于信息增益,比其它方法时间要短,优势明显. 实验结果如图 7 所示.



	IG(ID3)	IGR(C4.5)	Gini (CART)	(ID3+C4.5 +CART)/3	HAIGIGR
Training set	1.000	0.889	0.778	0.889	0.889
Test set	0.600	0.400	0.800	0.600	0.200
5 times 5-fold cross validation	0.600	0.800	0.360	0.587	0.840
The above average	0.733	0.696	0.646	0.692	0.643

图 3　天气数据集的分类预测结果比较

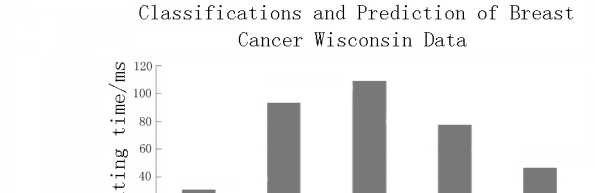


	IG(ID3)	IGR(C4.5)	Gini(CART)	(ID3+C4.5+ CART)/3	HAIGIGR
Training set	0.992	0.984	0.981	0.986	0.989
Test set	0.773	0.811	0.811	0.798	0.833
10 times 10-fold cross validation	0.828	0.782	0.768	0.793	0.824
The above average	0.864	0.859	0.853	0.859	0.882

图 4　QSAR biodegradation 数据集的分类预测结果比较

	IG(ID3)	IGR(C4.5)	Gini(CART)	(ID3+C4.5+ CART)/3	HAIGIGR
computing time(ms)	31	78	62	57	31

图 5　鸢尾花数据集的计算时间结果比较

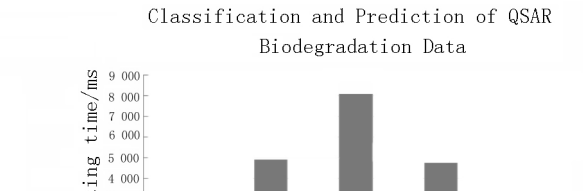


	IG(ID3)	IGR(C4.5)	Gini(CART)	(ID3+C4.5+ CART)/3	HAIGIGR
computing time(ms)	31	93	109	78	47

图 6　乳腺癌数据集的计算时间结果比较

5　结论

通过4份不同规模数据实验结果归结为,利用



	IG(ID3)	IGR(C4.5)	Gini(CART)	(ID3+C4.5+ CART)/3	HAIGIGR
computing time(ms)	1 238	4 921	8 054	4 738	1 929

图 7　QSAR biodegradation 数据集的计算时间结果比较

HAIGIGR、IG、IGR、GINI 指数作为属性划分的标准时:(i)在训练集、测试集及 10 次 10 折交叉验证单方面,4 种属性划分的准确性都较高,偶尔在上数据集上波动较大;(ii)在训练集、测试集及 10 次 10 折交叉验证的平均数方面,HAIGIGR 方法最好;(iii)对信息增益、信息增益率以及 GINI 指数 3 者结果的平均值而言,HAIGIGR 方法准确性较好;(iv)在运行时间方面,HAIGIGR 方法使用时间较短,有明显优势.

实验表明,利用 HAIGIGR、IG、IGR、GINI 指数作为属性划分的标准的准确率均较高,而 HAIGIGR 能较好地解决属性值倾向的问题,稳定性好,可行有效.

在研究和实验中发现,每次随机将样本分成训练集和测试集建模时的结果会有一定的差异,本文只取了一次实验的结果. 如何使得实验结果保持相对稳定,将是下一步工作的研究重点.

6　参考文献

[1] 周志华. 机器学习 [M]. 北京:清华大学出版社,2016.

[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 [EB/OL]. [2017-03-17]. http://ishare.iask.sina.com.cn/f/12391571.html.

[4] Chen Kunhuang,Wang Kungjeng,Wang Kungmin,et al. Applying particle swarm optimization-based decision tree classifierfor cancer classification on gene expression data [J]. Applied Soft Computing,2014:24(C):773-780.

[5] Chen Cuihua,He Binbin,Zeng Ze. A method for mineral prospectivity mapping integrating C4.5 decision tree,weights-of-evidence and m-branch smoothing techniques:a case study in the eastern Kunlun Mountains,China [J]. Earth Science Informatics,2014,7(1):13-24.

[6] Huang Aihui. C4.5 algorithm of decision tree improvement and application [J]. Science Technology and Engineer-

ing,2009(1):34-36,42.
[7] Jia Ping,Dai Jianhua,Pan Yunhe,et al. Novel algorithm for attribute reduction based on Mutual-information gain ratio [J]. Journal of Zhejiang University:Engineering Science,2006,40(6):1041-1044,1070.
[8] 王靖,王兴伟,赵悦. 基于变精度粗糙集决策树垃圾邮件过滤 [J]. 系统仿真学报,2016,28(3):705-710.
[9] 张楼,曹健. 面向大数据分析的决策树算法 [J]. 计算机科学,2016(S1):374-379,383.
[10] 于菲,张敏灵. 基于决策树集成的偏标记学习算法 [J]. 模式识别与人工智能,2016,29(4):367-375.
[11] 王杰,蔡良健,高瑜. 一种基于决策树的多示例学习算法 [J]. 郑州大学学报:理学版,2016,48(1):81-84.
[12] 王忠民,张琮,衡霞. CNN 与决策树结合的新型人体行为识别方法研究 [J]. 计算机应用研究,2017 (12):1-2.
[13] 王世东,刘毅,王新闯,等. 基于改进决策树模型的矿区土地复垦适宜性评价 [J]. 中国水土保持科学,2016,14(6):35-43.
[14] 李瑞红,李智,童玲. 蚁群路径优化决策树在慢性肾病分期诊断中的应用 [J]. 软件导刊,2017,16(2):135-138.
[15] 谢振平,孙桃. 自组织决策树的联想记忆在线学习模型 [J]. 模式识别与人工智能,2017,30(1):21-31.
[16] 张巍,聂进,滕少华. 基于互信息的模糊决策树及其增量学习 [J]. 江西师范大学学报:自然科学版,2014,38(1):89-94.
[17] 李航. 统计学习方法 [M]. 北京:清华大学出版社,2012.

The Improvement Decision Tree Algorithm for Harmonic Mean Optimization on Selection Attributes

WANG Zhuo[1],NIE Bin[2*],LUO Jigen[2],DU Jianqiang[2],CHEN Ai[1],ZHOU Li[2]
(1. School of Software,Nanchang University,Nanchang Jiangxi 330047,China;
2. School of Computer,Jiangxi University of Traditional Chinese Medicine,Nanchang Jiangxi 330004,China)

Abstract: Aiming at the preference of information gain and information gain rate for the number of attribute values, an improved decision tree algorithm is proposed to adjust the attribute of optimal selection. The basic idea of the algorithm is as follows. Firstly, the information gain of the candidate partitioning attribute is calculated to find out the attribute of the information gain higher than the average level. Then, the harmonic average of the information gain and information gain of these attributes are calculated respectively, value of the largest attribute, the establishment of branch decision. Lastly, the use of recursive method to establish decision tree. Through four experiments of different scale data, the information gain, information gain rate, GINI index and the method proposed in the paper are used as the criteria of attribute classification to examine the accuracy of the method in the training set, the test set, ten times the ten-fold cross validation (or five times the five-fold cross validation), and the three aspects of the average. The results show that the proposed method is of good accuracy and low running time, and has certain advantages.

Key words: decision tree; information gain ratio; harmonic mean; information of traditional Chinese medicine

(责任编辑: 冉小晓)