文章编号:1000 - 5862(2019)05 - 0469 - 04

# 融合 GINI 指数的 C4.5 算法的分类研究

聂 斌,李 欢,罗计根,杜建强,周 丽,黄 强
(江西中医药大学计算机学院,江西 南昌 330004)

**摘要:** 信息增益率倾向于取值数较少的属性和产生不平衡的划分,GINI 指数偏向于取值数较多的属性且区间趋于平衡的划分. 基于此,该文提出融合 GINI 指数的 C4.5 改进算法,首先计算候选属性的信息增益率和 GINI 指数,其次计算信息增益率和 GINI 指数的比值,最后筛选出比值最大的属性作为划分结点,改进了 C4.5 算法的不足. 以 10 次 10 折交叉验证准确率和运行时间为评价指标,通过 5 组 UCI 数据测试改进算法性能,并与 ID3、C4.5 和 CART 算法对比实验. 实验结果表明: 融合 GINI 指数的 C4.5 算法减轻了属性取值多少对划分结点选择的影响,并且缓和了划分区间的不平衡,提高了分类准确率和运行效率,算法更加稳定,可行有效.

**关键词:** C4.5 算法; GINI 指数; 决策树; 中医药信息
**中图分类号:** TP 301.6 **文献标志码:** A **DOI:** 10.16357/j.cnki.issn1000-5862.2019.05.05

## 0 引言

分类是人工智能领域中应用较为广泛的技术之一,常用的分类方法有决策树[1]、支持向量机[2]、神经网络[3]和贝叶斯算法[4]等,决策树算法相比于其他算法易理解,可解释性强. C4.5[5]算法是决策树生成算法之一,且广泛运用于网络流量、天气预警和车位识别等分类研究. 针对 C4.5 算法的不足,夏修臣等[6]运用余弦相似度改进 C4.5 算法,有效地降低了分裂属性的维度,缩减了决策树的规模,降低了算法的复杂度; 曾繁慧等[7]采用云的方法将数值型属性离散化,将分辨数作为分裂属性的选取标准,降低了算法的时间复杂度,提高了算法的准确率; Peng Hai等[8]利用 CPU-FPGA 异构平台和 OpenCL 设计流程,改进 C4.5 算法的训练过程,使训练速度提高了 3 倍; M. Z. F. Nasution 等[9]运用主成分分析进行特征约简,再运用 C4.5 算法分类,模型的鲁棒性和准确率得到了提高. 黄秀霞等[10]运用泰勒级数和等价无穷小简化信息增益率计算公式,并引入其他非类属性对该属性的 GINI 指数均值,减少非类属性间的冗余误差.

C4.5 算法运用信息增益率为属性选择度量标准,但信息增益率偏向于可取值数较少的属性[11],因此,C4.5 算法采用了一个启发式策略: 先取出信息增益大于平均水平的候选属性,再从中选择信息增益率最高的属性为划分结点,但该方法没有改进信息增益率的不平衡划分[12]. CART 算法[13]是另一种决策树生成算法,其采用 GINI 指数为属性选择度量标准,GINI 指数偏向于可取值数较多的属性[14]且区间趋于平衡的划分. 基于此,针对信息增益率偏向于可取值数较少的属性和产生不平衡划分的问题,本文提出融合 GINI 指数的 C4.5 算法,该算法将信息增益率与 GINI 指数的比值作为属性选择度量标准,改进信息增益率的不足,提高 C4.5 算法的准确率.

## 1 C4.5 算法

设 $D$ 为标记类元组的训练集,类标号属性可取 $n$ 个不同的值,则该集合中有 $n$ 个不同的类 $C_i(i = 1,2,\cdots,n)$. $C_{i,D}$ 是 $D$ 中 $C_i$ 类元组的集合,$|D|$ 和 $|C_{i,D}|$ 分别是 $D$ 和 $C_{i,D}$ 元组的个数.

假设 $C_i$ 类样本所占集合 $D$ 的比例为 $p_i$,记为 $|C_{i,D}|/|D|$. 属性 $A$ 有 $v$ 个值分别为$\{a_1,a_2,\cdots,a_v\}$,其中取值为 $a_v$ 的元组记为集合 $D_j$,若用 $A$ 来对

收稿日期: 2019-01-03
基金项目: 国家自然科学基金(61562045),江西省卫生计生委中医药科研计划(普通)(2017A282)和江西省科技厅重点研发计划(20171ACE50021)资助项目.
作者简介: 聂 斌(1972-),男,江西峡江人,副教授,主要从事数据挖掘、机器学习、人工智能和中医药信息学的研究. E-mail: ncunb@163.com

集合 $D$ 进行划分,则会产生 $v$ 个分支结点.

$D$ 的信息熵为

$$I_{nfo}(D) = -\sum_{i=1}^{n} p_i \log_2 p_i. \quad (1)$$

属性 $A$ 的信息熵为

$$I_{nfo_A}(D) = \sum_{j=1}^{v} (D_j/D) I_{nfo}(D_j). \quad (2)$$

属性 $A$ 对训练集 $D$ 进行划分所获得的信息增益为

$$G_{ain}(A) = I_{nfo}(D) - I_{nfo_A}(D). \quad (3)$$

属性 $A$ 的分裂信息为

$$S_{plitInfo_A}(D) = -\sum_{j=1}^{v} (|D_j|/|D|) \log(|D_j|/|D|). \quad (4)$$

属性 $A$ 对训练集 $D$ 进行划分所获得的信息增益率为

$$G_{ainRate_A}(D) = G_{ain}(A)/S_{plitInfo_A}(D). \quad (5)$$

重复上述过程依次计算每个属性的信息增益率. 由于信息增益率倾向于可取值数较少的属性,因此,C4.5 算法采用了一个启发式策略: 先取出信息增益大于平均水平的候选属性,再从中选择信息增益率最高的属性为划分结点,结点的分支数为属性的取值个数.

## 2 改进的 C4.5 算法

### 2.1 改进算法理论分析

在信息增益率计算过程中,由于

$$\log_2 a_1 + \log_2 a_2 + \cdots + \log_2 a_m < \log_2(a_1 + a_2 + \cdots + a_m)\ (a_1, a_2, \cdots, a_m < 1), \quad (6)$$

所以,(4) 式属性的分裂信息值通常会随属性的可取值数目减少而减少,进而(5) 式属性的信息增益率会增大. 因此,信息增益率会倾向于可取值数较少的属性,C4.5 使用了一个启发式改进了此处的不足,但其以每个结点的划分子集为属性的不同属性值集合,即结点的分支数等于属性的可取值数. 当属性的可取值数量过大时,产生的分支数和叶子结点数也会增大,生成的决策树复杂度高; 当属性的各属性值集合不平衡时,每个结点的划分子集也会不平衡.

CART 是一种二叉树生成算法,其采用 GINI 指数为属性选择标准,GINI 指数越小,则数据集的纯度越高,选择以划分后 GINI 指数最小的属性为最优划分结点,GINI 指数偏向于取值数较多的属性,且划分平衡. 鉴于此,采用比值法将 GINI 指数和信息增益率融合,理论上不仅改进了信息增益率对取值数较少属性的偏好和划分不平衡问题,还使每个结点将按属性取值划分改为 2 元划分,消除了分支之间的冗余,降低了决策树的复杂度.

### 2.2 改进算法模型建立

C4.5 算法根据(5) 式选择信息增益率最大的属性为划分结点,结点的分支数为划分属性的取值数. 本文提出的算法是选择信息增益率与 GINI 指数比值最大的属性为划分结点,每个结点采用 2 元划分,从而减少信息增益率对可取值数较少属性的偏向,以及减少结点的分支数,降低决策树的复杂度,提高分类精度,算法思想步骤如下:

**输入** 数据集 $D$,候选划分属性列表 attributeList;

**输出** 一棵二叉树;

**Step 1** 创建一个根节点 node;

**Step 2** 若数据表为空,则 node 为叶节点,return;

**Step 3** 若类属性都是同一个值 $C$,则 node 为叶节点,其类别标记为 $C$,return;

**Step 4** 若数据表中自变量属性列表为空,则 node 为叶节点,其类别标记为类属性最多的类,return;

**Step 5** 根据(1) 式计算数据集 $D$ 的信息熵,计算数据集 $D$ 的 $G_{ini}(D)$ 为

$$G_{ini}(D) = 1 - \sum_{i=1}^{n} p_i^2; \quad (7)$$

**Step 6** 对于数值型输入变量,将数据按升序排序后,从小到大依次以相邻数值的中间值为组限,将数据集 $D$ 划分为 2 组; 对于多分类型输入变量,将多类别合并成 2 个“超类”;

**Step 7** 根据(2) ~(5) 式计算按某一分割点划分属性 $A$ 的信息增益率 $G_{ainRate_{(A,a)}}(D)$,此时公式中 $v$ 的取值为 2;

**Step 8** 计算按某一分割点 $a$ 划分属性 $A$ 的基尼指数为

$$G_{ini_{(A,a)}}(D) = (|D_1| G_{ini}(D_1) + |D_2| G_{ini}(D_2))/|D|; \quad (8)$$

**Step 9** 计算按某一分割点 $a$ 划分属性 $A$ 的信息增益率 $G_{ainRate_{(A,a)}}(D)$ 与基尼指数 $G_{ini_{(A,a)}}(D)$ 的比值 $R_{atio_{(A,a)}}(D)$ 为

$$R_{atio_{(A,a)}}(D) = G_{ainRate_{(A,a)}}(D)/G_{ini_{(A,a)}}(D); \quad (9)$$

**Step 10** 循环 Step 4 ~ Step 9,计算其他属性的信息增益率与基尼指数的比值;

**Step 11** 选择信息增益率与基尼指数的比值最高的属性,若该属性为 $A$,则把结点 node 标记为属性 $A$;

**Step 12** 删除属性 $A$;

**Step 13** 找到使 $R_{atio_A}(D)$ 最大的分割点,按属性 $A$ 的最佳分割点从 node 结点中生长出 2 个分支;

**Step 14** 循环 Step 3 ~ Step13;

**Step 15** 终止.

根据上述算法思想,构建融合 GINI 指数的 C4.5算法流程图,如图 1 所示.

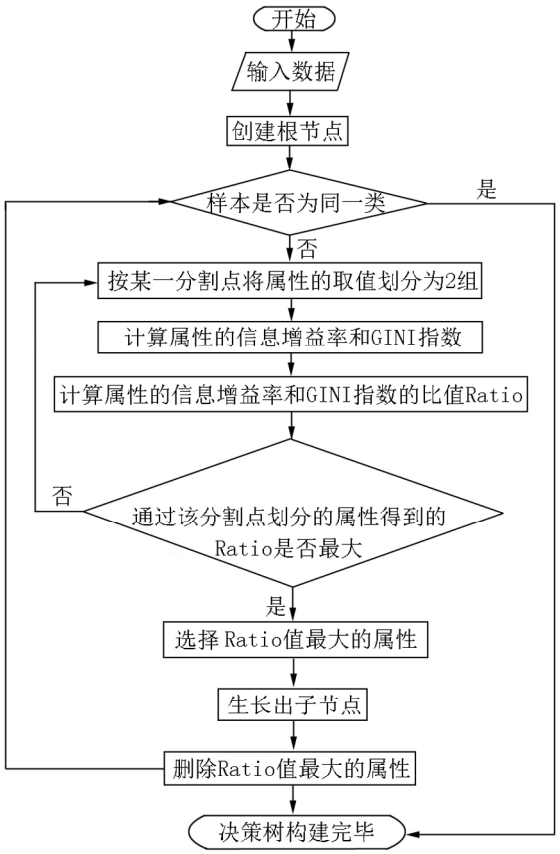


**图 1　融合 GINI 指数的 C4.5 算法流程**

# 3　实验结果及分析

## 3.1　实验数据说明

为了验证改进算法的有效性,本文选取了 UCI 数据库中的 5 组分类数据集,表 1 对 5 组数据集的属性数、样本数、类别数做了简要描述.

**表 1　UCI 数据集**

| 数据集 | 属性数 | 样本数 | 类别数 |
|---|---|---|---|
| post-operative-patient | 8 | 87 | 3 |
| vote | 16 | 435 | 3 |
| hamberman | 4 | 306 | 2 |
| balance-scale | 4 | 625 | 3 |
| CAR | 6 | 1 729 | 4 |

## 3.2　实验结果分析

将 5 组数据集在实验环境为 Window10 操作系统(64 位)、Intel(R) Core(TM) i5-3470 CPU、8G 的 RAM 以及 eclipse 开发平台上展开实验. 在具体实验过程中,叶子结点的最大容许个数为 3,最大容许误差为 0.001. 通过比较 ID3、C4.5、CART 和改进 C4.5 算法的 10 次 10 折交叉验证准确率和运行时间,分析 3 种算法的性能,实验结果见表 2. 为了更加直观地比较实验结果,分别绘制了图 2 和图 3 来表示 3 种算法的 10 次 10 折交叉验证准确率和运行时间.

图 2 和图 3 中横坐标代表数据集,实线是 CART,虚线是 C4.5,线-点交叉线是 ID3,点线是改进的 C4.5,图 2 的纵坐标代表 10 次 10 折交叉验证准确率,图 3 的纵坐标代表运行时间. 从图 2 可看出,在数据集 post-operative-patient、vote、haberrman 和 balance-scale 上,明显的看出点线在最上方,这表明在这 4 组数据集上,3 种算法中改进的 C4.5 算法 10 次 10 折交叉验证准确率最高. 从图 3 可看出,在 5 组数据集上,点线位于实线的下方,虚线和线-点交叉线的上方,这说明改进算法的运行速度比 CART 快,但比 C4.5 和 ID3 慢.

**表 2　C4.5、CART 和改进 C4.5 算法的 10 次 10 折交叉验证准确率和运行时间**

| | ID3 | | CART | | C4.5 | | 改进 C4.5 算法 | |
|---|---|---|---|---|---|---|---|---|
| | 准确率 | $t$/ms | 准确率 | $t$/ms | 准确率 | $t$/ms | 准确率 | $t$/ms |
| post-operative-patient | 0.581 7 | 1 | 0.556 7 | 194 | 0.600 0 | 26 | 0.701 7 | 50 |
| vote | 0.929 0 | 4 | 0.923 1 | 10 | 0.911 2 | 2 | 0.933 9 | 10 |
| hamberman | 0.715 6 | 39 | 0.720 9 | 86 | 0.704 4 | 17 | 0.742 8 | 75 |
| balance-scale | 0.707 8 | 24 | 0.720 9 | 200 | 0.704 4 | 9 | 0.742 8 | 183 |
| CAR | 0.971 1 | 16 | 0.974 8 | 139 | 0.973 8 | 140 | 0.985 9 | 127 |

通过表 2 中的数据详细比较 3 种算法的 10 次 10 折交叉验证准确率,在 postoperative-patient 数据集上,改进算法的准确率为 0.701 7,比 C4.5 算法高 0.101 7,比 ID3 算法高 0.120 0,比 CART 算法高 0.145 0;在 vote 数据集上,改进算法的准确率为 0.933 9,比 C4.5 算法高 0.022 7,比 ID3 算法高 0.004 9,比 CART 算法高 0.010 8;在 haberrman 数据集上,改进算法的准确率为 0.742 8,比 C4.5 算法高 0.038 4,比 ID3 算法高 0.027 2,比 CART 算法高出 0.021 9;在 balance-scale 数据集上,改进算法的准确率为 0.742 8,比 C4.5 算法高 0.038 4,比 ID3 算法高 0.027 2,比 CART 算法高 0.021 9;在 CAR 数据集上,改进算法的准确率为 0.985 9,比 C4.5 算法高 0.012 1,比 ID3 算法高 0.014 8,比 CART 算法高 0.011 1.

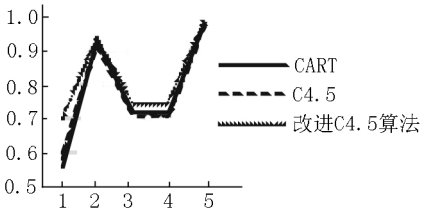


**图 2　3 种算法的 10 次 10 折交叉验证准确率**

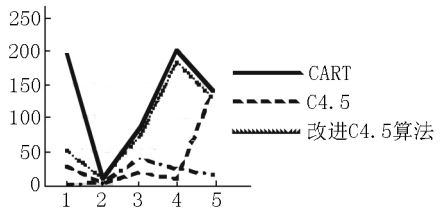


图3 3种算法的运行时间

综上所述,虽然改进算法的运行速度比 CART 快、比 ID3 和 C4.5 慢,但其 10 次 10 折交叉验证准确率比 ID3、C4.5 和 CART 都高. 因此,融合 GINI 指数的 C4.5 算法的准确率得到了一定的提高.

## 4 结语

对于信息增益率倾向于取值数较少属性和产生不平衡的划分,本文提出融合 GINI 指数的 C4.5 算法,该算法运用 GINI 指数偏向于取值数较多的属性且区间趋于平衡的划分的特点,减轻了属性取值多少对划分结点选择的影响. 经过 5 组 UCI 数据集实验,并与 ID3、C4.5 和 CART 对比,改进算法表现出更高的 10 次 10 折交叉验证准确率. 在实验过程中发现,本文提出的算法对类别不平衡数据的分类准确率不高,因此,后续将对该方面做进一步的研究.

## 5 参考文献

[1] 陈亚慧,叶继华. 基于决策树分类的个性化农产品移动信息服务系统 [J]. 江西师范大学学报: 自然科学版,2016,40(2):145-148.

[2] 冷强奎,刘福德,秦玉平. 一种基于混合二叉树结构的多类支持向量机分类算法 [J]. 计算机科学,2018,45(5):220-223,237.

[3] 杨国亮,王志元,张雨. 一种改进的深度卷积神经网络的精细图像分类 [J]. 江西师范大学学报: 自然科学版,2017,41(5):476-483.

[4] Tang Bo,He Haibo,Baggenstoss P M,et al. A Bayesian classification approach using class-specific features for text categorization [J]. IEEE Transactions on Knowledge and Data Engineering,2016,28(6):1602-1606.

[5] Quinlan J R. C4.5: programs for machine learning [M]. San Francisco: Morgan Kaufmann Publisher,1993.

[6] 夏修臣,王秀英. 基于余弦相似度的改进 C4.5 决策树算法 [J]. 计算机工程与设计,2018,39(1):120-125.

[7] 曾繁慧,李艺. 因素空间理论的决策树 C4.5 算法改进 [J]. 辽宁工程技术大学学报: 自然科学版,2017,36(1):109-112.

[8] Peng Hai,Zhang Xiaofan,Huang Letian. An energy efficient approach for C4.5 algorithm using OpenCL design flow [EB/OL]. [2018-11-19]. https://ieeexplore.ieee.org/document/8280132.

[9] Nasution M Z F,Sitompul O S,Ramli M. PCA based feature reduction to improve the accuracy of decision tree c4.5 classification [EB/OL]. [2018-11-19]. http://iopscience.iop.org/article/10.1088/1742-6596/978/1/012058/pdf.

[10] 黄秀霞,孙力. C4.5 算法的优化 [J]. 计算机工程与设计,2016,37(5):1265-1270,1361.

[11] 周志华. 机器学习 [M]. 北京: 清华大学出版社,2016.

[12] Han Jiawei,Micheline Kamber,Pei Jian. 数据挖掘: 概念与技术 [M]. 3 版. 范明,孟小峰,译. 北京: 机械工业出版社,2012.

[13] Breiman L I,Friedman J H,Olshen R A,et al. Classification and regression trees(CART) . wadsworth [J]. Encyclopedia of Ecology,1984,40(3):582-588.

[14] 孙喜洲. 数据挖掘分类技术在健身会所管理系统中的应用研究 [D]. 青岛: 中国海洋大学,2011.

[15] UCI. Machine learning repository [EB/OL]. [2019-01-17]. http://archive.ics.uci.edu/ml/index.php.

## The Study on Classification of C4.5 Algorithms with GINI Index

NIE Bin,LI Huan,LUO Jigen,DU Jianqiang,ZHOU Li,HUANG Qiang
(School of Computer Science,Jiangxi University of Traditional Chinese Medicine,Nanchang Jiangxi 330004,China)

**Abstract:** The information gain rate tends to take fewer attributes and produce an imbalance partition. The GINI index tends to take more attributes and produce the balanced partition. Based on this,an improve C4.5 algorithm combining GINI index is proposed. The algorithm first calculates the information gain rate and GINI index of candidate attributes,and then calculates the ratio of information gain rate to GINI index. Finally,the attribute with the largest ratio is selected as the segmentation node,which improves the shortcomings of the C4.5 algorithm. Taking ten times and ten fold cross-validation accuracy and running time as evaluation index,the improved algorithm performance is tested through five UCI data sets and compared with ID3,C4.5 and CART algorithms. The results show that the C4.5 algorithm combining GINI index reduces the influence of attribute value on the selection of partition nodes,and alleviates the imbalance of partition interval,which improves the classification accuracy and operation efficiency. The algorithm is more stable and feasible.

**Key words:** C4.5 algorithm; GINI index; decision tree; information of Chinese medicine (责任编辑: 冉小晓)