文章编号：1000-5013(2016)01-0071-03 **doi**：10.11830/ISSN.1000-5013.2016.01.0071

一种改进 ID3 型决策树挖掘算法

潘大胜，屈迟文

（百色学院 信息工程学院，广西 百色 533000）

摘要： 分析经典 ID3 型决策树挖掘算法中存在的问题，对其熵值计算过程进行改进，构建一种改进的 ID3 型决策树挖掘算法. 重新设计决策树构建中的熵值计算过程，以获得具有全局最优的挖掘结果，并针对 UCI 数据集中的 6 类数据集展开挖掘实验. 结果表明：改进后的挖掘算法在决策树构建的简洁程度和挖掘精度上，都明显优于 ID3 型决策树挖掘算法.

关键词： 数据挖掘；ID3 型决策树；熵值计算；UCI 数据集

中图分类号： TP 301.6 **文献标志码**： A

数据仓库技术的出现，成功地解决了大数据的存储和管理问题，并配套数据挖掘技术从海量数据中提取最有价值的信息[1-2]. 数据挖掘技术是从已有数据信息中提取有价值信息，甚至形成知识. 目前，已经出现的数据挖掘技术包括基于机器学习的挖掘技术、基于聚类分析的挖掘技术、基于关联规则的挖掘技术等[3-5]. 在各种基于机器学习的挖掘技术中，决策树挖掘算法获得了广泛的应用[6-12]. ID3 型决策树挖掘法是最经典的决策树挖掘算法之一. 在决策树的各级节点上，采用信息增益构建属性选择依据，最高的信息增益属性最终会被确定为节点的判别标准，这样就可以达成训练样本的信息最小化，提升决策树的构建速度，并简化决策树的结构. 本文在经典 ID3 型决策树算法的基础上，构建一种改进型的决策树挖掘算法.

1 改进的决策树挖掘算法

经典的 ID3 型决策树挖掘算法具有搜索空间完整、测试次数少、分类挖掘速度快、构建的决策树节点少、适合于噪声数据和离散数据的挖掘等优点. 然而，其最大的局限性在于它的建树原则是依据信息熵理论计算出的属性增益. 这种算法选取出的最佳属性是一种多值属性，并不一定是最优的，导致其挖掘结果往往只具有局部最优的特性，而无法达到全局最优. 文中对经典的 ID3 型决策树算法进行改进，构建一种能够达到全局最优的决策树.

信息熵值的计算是 ID3 型决策树算法挖掘过程实现的关键. 文中主要对熵值的计算进行改进，进而在此基础上重新设计执行流程. 在经典 ID3 型决策树算法中，属性 A 在计算过程中会出现多值，选取这些值的个数作为权重系数，融入属性 A 的熵值计算中.

假设属性 A 存在 m 个属性，这些属性出现的概率用 $p_1, p_2, \cdots, p_m$ 表示，属性 A 作为节点对应的 m 个子节点可以用属性值集合表示为 $\{\theta_1, \theta_2, \cdots, \theta_m\}$，这些属性值对应的信息熵 $G(\theta_1), G(\theta_2), \cdots, G(\theta_m)$. 改进算法最终设计的用于计算属性 A 的熵，其表达式为

$$\overline{G}(A) = m\sum_{i=1}^{m} p_i \times G(\theta_i). \tag{1}$$

改进决策树算法主要有以下 5 个执行步骤.

步骤 1 对于任意一个属性 A_i，假设它有 m_i 个属性值，可以用 $\{\theta_1, \theta_2, \cdots, \theta_{m_i}\}$ 表示. 这些属性值对

收稿日期： 2015-11-13

通信作者： 潘大胜(1975-)，男，副教授，主要从事数据挖掘技术的研究. E-mail：bspandsh@163.com.

基金项目： 广西自然科学基金青年基金资助项目(2014GXNSFBA118283)

应的概率分别为 $p_1, p_2, \cdots, p_{m_i}$,其中,每个属性值对应的信息熵计算式为

$$G(\theta_i) = \sum_{i=1}^{n} \frac{2t_i f_i}{t_i + f_i}. \tag{2}$$

步骤 2 借助式(1)计算属性 A_i 对应的熵.

步骤 3 按照步骤 1,2 的方法,继续计算属性 $A_{i+1}, A_{i+2}, \cdots$ 对应的熵值,并从所有的熵值中,选择最小的熵值所对应的属性为节点.

步骤 4 按照步骤 1~3 的方法,继续各个后继节点.

步骤 5 当新一轮计算结果确定的各个节点为叶子节点时,决策树构建完毕;否则,继续执行以上各步骤.

2 验证性实验

2.1 实验所用数据集

为了验证文中算法的有效性,选择 6 种加利福尼亚大学的 UCI 测试数据集(Wine, Spect Heart, Balance Scale, Vehicle Silhouettes, Hill Valley, Yeast)执行数据挖掘实验.

Wine 数据集是用于判断白酒种类的化学成分数据,这些数据主要来自产于意大利的白酒,这些化学成分涉及酒精、果酸、黄酮等. Wine 数据集中含有 178 个样本数据,13 个整数属性,3 个分类类别.

Spect Heart 数据集是根据 CT 图像判断病患心脏是否正常的数据集,这些数据是根据质子发射 CT 扫描结果得出的. Spect Heart 数据集中含有 267 个样本数据,22 个属性特征,每个属性只有"-1"和"1"两种表达可能,2 个分类类别,即正常和不正常.

Balance Scale 数据集是根据心理学实验判断天平是否平衡的数据集,这些数据根据天平 2 个托盘的质量、距离和测试者的心理判断天平是否会发生倾斜. Balance Scale 数据集中含有 625 个样本数据,每个样本数据含有 4 个属性特征,这些属性需要在"1,2,3,4,5"中取值,3 个分类类别(左侧倾斜、右侧倾斜、平衡).

Vehicle Silhouettes 数据集是根据图像的二维特征信息判断汽车类别所形成的数据集合. Vehicle Silhouettes 数据集中含有 846 个样本数据,每个样本数据含有 18 个属性特征,这些属性需要在整数空间上取值,4 个分类类别.

Hill Valley 数据集是用于判断多数据点连接曲线后凹凸形状的数据集,它先将 100 个数据点连接成曲线,再判断曲线的凹凸形状. Hill Valley 数据集中含有 1 212 个样本数据,每个样本数据含有 100 个属性,这些属性在实数空间上取值,2 个分类类别.

Yeast 数据集是利用据蛋白质成分推测细胞位置的数据集. Yeast 数据集含有 1 484 个样本数据,每个样本数据含有 8 个属性,这些属性在实数空间上取值,10 个分类类别.

2.2 ID3 算法与改进算法的实验对比

分别采用 ID3 型算法和改进算法执行挖掘分类实验,结果如表 1 所示. ID3 算法与改进算法在决策树构建的节点数和挖掘精度上的对比结果,如表 1 所示.

表 1 ID3 算法与改进算法的性能对比结果

Tab. 1 Comparison results of the performance of the ID3 algorithm and the improved algorithm

数据集	节点数		挖掘精度	
	ID3 算法	改进算法	ID3 算法	改进算法
Wine	340	252	0.81	0.93
Spect Heart	273	221	0.91	0.97
Balance Scale	166	142	0.76	0.85
Vehicle Silhouettes	132	118	0.77	0.88
Hill Valley	96	67	0.72	0.84
Yeast	55	32	0.69	0.82

由表 1 可知:改进算法相比于经典的 ID3 算法,其构建的决策树的节点数量更少,表明改进算法构建的决策树更加精简,且在挖掘精度方面也有明显的提高.

3 结束语

针对 ID3 型决策树挖掘算法中存在的数据挖掘问题,重新设计 ID3 型决策树构建中的熵值计算过程,构建一种改进的 ID3 型决策树挖掘算法. 通过 UCI 数据集中的 6 种数据集展开实验,结果表明改进

算法具有更精简的决策树结构、更高的挖掘精度，优于 ID3 型决策树挖掘算法.

参考文献：

[1] SETTOUTI N，AOURAG H. A comparative study of the physical and mechanical properties of hydrogen using data mining research techniques[J]. JOM：the Journal of the Minerals，Metals & Materials Society，2015，67(9)：2145-2153.

[2] 沈伟. 基于数据挖掘技术的高职院校招生决策仓库设计与实现[J]. 网络安全技术与应用，2015(3)：165-167.

[3] 钱昭勇，陈梅. 数据挖掘技术在突发事件应急系统中的应用与研究[J]. 电子技术与软件工程，2014，19(8)：217.

[4] PALACIOS A，MARTINEZ A，SANCHEZ L，et al. Sequential pattern mining applied to aeroengine condition monitoring with uncertain health data[J]. Engineering Applications of Artifical Intelligence，2015，44：10-24.

[5] 罗来曦，朱渔. 以 XML 为基础的 Web 数据挖掘技术系统的框架设计与实现[J]. 电子技术与软件工程，2014，15(7)：201.

[6] BANAEE H，LOUTFI A. Data driven rule mining and representation of temporal patterns in physiological sensor data[J]. IEEE Journal of Biomedical and Health Informatics，2015，19(5)：1557-1566.

[7] PORRO-MUNOZ D，OLIVETTI E，SHARMIN N，et al. Tractome：A visual data mining tool for brain connectivity analysis[J]. Data Mining and Knowledge Discovery，2015，29(5)：1248-1279.

[8] 吴春琼，胡国柱，徐静. 高职院校项目驱动模式下基于数据挖掘决策树分类的教学效果分析[J]. 吕梁教育学院学报，2015，32(10)：18-21.

[9] 李楠，段隆振，陈萌. 决策树 C4.5 算法在数据挖掘中的分析及其应用[J]. 计算机与现代化，2008，160(12)：160-163.

[10] 高媛媛. 基于数据挖掘的财务舞弊识别研究：决策树-神经网络组合模型的构建[J]. 科技经济市场，2014(11)：93-95.

[11] ELYASIGOMARI V，MIRJAFARI M S，SCREEN H R C，et al. Cancer classification using a novel gene selection approach by means of shuffling based on data clustering with optimization[J]. Applied Soft Computing，2015，35：43-51.

[12] 李翔，刘韶涛. FP-Growth 的并行加权关联规则挖掘算法[J]. 华侨大学学报(自然科学版)，2014，35(5)：523-527.

An Improved ID3 Decision Tree Mining Algorithm

PAN Dasheng，QU Chiwen

(School of Information Engineering，Baise University，Baise 533000，China)

Abstract：By analyzing the problem of ID3 decision tree mining algorithm，the entropy calculation process is improved，and a kind of improved ID3 decision tree mining algorithm is built. Entropy calculation process of decision tree is redesigned in order to obtain global optimal mining results. The mining experiments are carried out on the UCI data category 6 data set. Experimental results show that the improved mining algorithm is much better than the ID3 type decision tree mining algorithm in the compact degree and the accuracy of the decision tree construction.

Keywords：data mining；ID3 decision tree；entropy calculation；UCI data set

(责任编辑：黄晓楠　　英文审校：吴逢铁)