

基于 C4.5 决策树算法的数据挖掘技术在《计算机程序设计》课程混合式教学的应用研究

李俊杰

(陆军军医大学基础医学院计算机教研室, 重庆 400030)

摘要: 随着大数据、人工智能技术与在线教学的逐步深度融合, 使得在线教学平台数据的精准分析和挖掘成为可能。本文利用交互式教学平台 PYTHON123, 记录了学习者的各项学习行为数据, 并建立 C4.5 决策树分类模型, 对学生学习行为数据进行科学评判和预测评估, 从而更加全面、有效、科学的指导学习者行为和效果。

关键词: C4.5; 决策树; 数据挖掘; 混合式教学

近年来, 混合式教学由于其自身具有的实时交互、非固定时空、个性化指导、海量资源的优势, 受到了越来越多教育者的拥护。计算机程序设计课程由于实践性强、学生知识基础参差不齐、亟需老师给予更多的针对性指导的特点, 完美契合了混合式教学的使用优势。本文利用 PYTHON 慕课平台 (www.python123.io), 从 2019 年 -2021 年进行了针对大二年级临床医学专业学生的《计算机程序设计》课程, 并采用线上线下混合式教学。线下方面主要是翻转课堂教学改革和编程实操, 线上方面开展了自主学习、碎片化学习、互动式学习平台。由于课程特点, 课程采用了终结性考核与形成性考核相结合的方式。其中形成性考核评定需要综合量化评定。如何科学客观的评价学习效果, 指引混合式教学更好的实施, 成为现阶段教育研究的重要课题。基于此, 本文在分析学生在线学习行为的基础上, 构建 C4.5 决策树模型, 对学习结果进行科学评判, 进而反向推判在线学习行为与学习效果之间相关性影响因素, 旨在调整优化教学模式, 改善教学方法, 指导教学实施, 提高混合式教学效果。

一、理论基础

(一) 数据挖掘技术

数据挖掘是指从大量的数据中通过算法搜索隐藏于其中信息的过程。数据挖掘可以分析大量数据中的关系和趋势, 即将原始数据转换为有用的信息和知识。数据挖掘一般基于大量的样本, 同时推测出相关性、演化趋势等。技术层面通过采用统计分析、分类、聚类、关联规则、预测、估计等方法。数据挖掘的一般过程包括: 定义目标、熟悉数据、清洗数据、数据挖掘、算法实现、分析应用。

(二) C4.5 决策树算法的概述

决策树算法是数据挖掘中常用的算法之一, 本质上以一棵树呈现结构, 并以将相同属性的数据进行归类。树包括根节点和叶子节点, 每一个节点代表一个类别的属性, 决策树可以从节点判断具体数据属性和层级。最常用的决策树算法有 ID3、C4.5、CART、ASSISTANT 算法等。

C4.5 算法在继承了 ID3 算法特点的基础上, 采用信息增益率作为选取属性的标准。

(1) 对于集合 D 分为 C 个类别, D 的熵值为

$$entropy(D) = - \sum_{i=1}^{|C|} Pr(c_i) \log_2 Pr(c_i)$$

(2) 信息增益值, A 是 D 的属性特征, 并划分为 V 个值, 即 D 划分成为不同 V 个子集, 则 D 的信息熵为:

$$entropy(A, D) = \sum_{j=1}^V \frac{|D_j|}{|D|} \times entropy(D_j)$$

信息增益值计算为 $Gain(D, A) = entropy(D) - entropy(A, D)$

信息增益率则将熵值进行归一化处理, 除以类属性, 计算成为信息增益率, 为

$$GainRation(D, A) = \frac{Gain(D, A)}{- \sum_{j=1}^V (\frac{|D_j|}{|D|} \times \log_2 \frac{|D_j|}{|D|})}$$

二、数据的预处理

(一) 数据的采集

在混合式教学所使用的 PYTHON123 平台里, 包含了学生各项学习行为数据, 如学生的年级、院系、专业、答疑发帖数、章节练习完成率、登录时长、视频浏览时长、摸底考试成绩等, 通过 csv 格式按年度和班级批量导出, 然后对数据进行合并、填充、转换、清洗、集成等。

(二) 数据的建模和泛化

对前期的数据, 提取期中期末成绩、章节作业完成率、登录时长、视频浏览时长、考前测试成绩作为一个向量 {D1, D2, D3, D4, D5}, 属性值分别取 {A、B、C、D}, 分别代表优, 良, 中, 差。划分范围根据历史成绩和当前成绩平均值取得。

(三) 数据的处理

对数据处理过程的关键是如何对学习者的学习行为进行离散化分级, 分成 A、B、C、D。主要包括时间数据处理和分数数据处理。

时间数据处理: 在以时间为单位的各项行为中, 如登陆总时长和视频浏览时间进行分级。例如在视频浏览时长方面, 每次课时长 10 分钟, 45 个学时, 根据模型估算和以往出现频率, 则划分等级 B=[430, 380], 等级 C 为 [380, 330], 等级 D 为 [330, 280], 等级 A 为 [280, 230]。这样就将连续的数值划入不同离散区间。

分数数据处理: 分数基于百分制, 完成率是分数形式, 这些数据需要经过定量计算。分数和完成率根据以往估算获取一个平均分数, 划分成 A, B, C, D 的形式, 对应阈值换算, 这样就完成了数据分级。

三、决策树建立和 C4.5 算法运用

利用 C4.5 算法核心思想, 以信息增益率为评分标准, 建立学习评价指标体系。

表 1 部分原始数据

序号	登录时长 (分钟)	视频浏览时 长 (分钟)	章节练习 完成率	摸底考试 成绩 (分)	答疑发帖 数 (个)	等级
1	3984	425	0.94	75	65	A
2	2820	408	0.75	64	50	B
3	3124	438	0.80	80	79	A
4	2975	377	0.67	69	48	B

数值进行量化得到表 2 中的数据:

表2 量化处理后的数据

序号	登录时长 (分钟)	视频浏览时 长(分钟)	章节练习 完成率	摸底考试 成绩(分)	答疑发帖 数(个)	等级
1	A	B	A	A	B	A
2	C	B	B	B	C	B
3	A	A	A	A	A	A
4	C	C	C	B	B	B

建立数据集：选取分类模型的训练集和测试数据集进行决策树的生成。训练数据集和测试数据集样本各为480例，按照表中规则，得到属性A样本96例，属性B样本182例，属性C样本170例，属性D样本32例，则熵值为：

$$entropy(s) = -\sum_{i=1}^4 Pr(c_i) \log_2 Pr(c_i) = 1.754$$

然后计算机每个决策分类的熵值，再得到信息增益率。根据规则计算每个属性的信息增益率，得到决策树如图1所示：

图1的决策树由IF—THEN划分规则构成，假设汇总一个数据需要进行分类，从决策树的根结点出发，沿着树枝一直进行下去，最终会得到数据分类的结果。

规则类似于：IF（“摸底考试成绩”=A）AND（“章节完成率”=A）AND（“视频浏览时长”=B）AND（“登录时长”=B）AND（“答疑发帖数”=B）THEN “学生预测成绩”=A；……通过诸如此类一系列上述规则进行类别的判定。

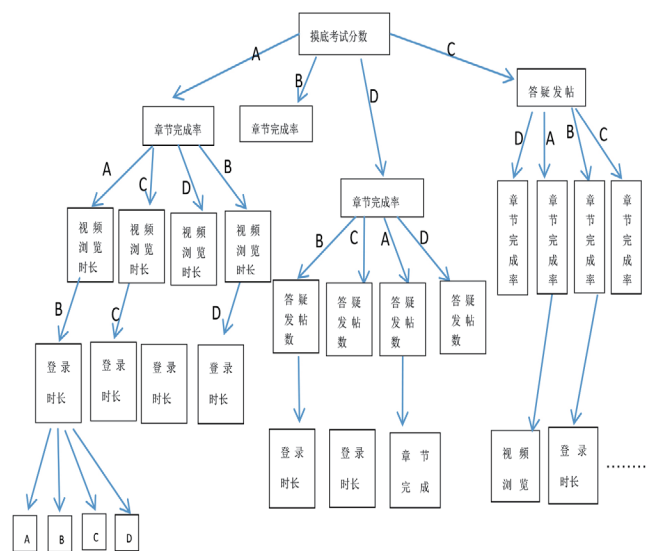


图1 决策树的构建和属性分类

四、模型准确性评估

生成的分类模型对测试数据集中的未知数据，是否能够准确判断，还要进行验证性评估。下表是对测试数据集进行准确性评估的结果。

表3 测试数据集预测结果

测试样本被分类别	分类样本数	预测结果为A的 样本数	预测结果为B的 样本数	预测结果为C的 样本数	预测结果为D的 样本数	准确率	错误率
A	98	88	5	4	1	89.80%	10.20%
B	172	8	156	5	3	90.70%	9.30%
C	152	3	7	137	5	90.13%	9.87%
D	58	0	3	4	51	86.21%	13.79%
合计	480	99	171	150	60	88.20%	11.8%

从表3中数据看到，对测试样本集进行分类总体正确率在88.20%左右，错误率在11.8%。这个结果基本达到我们预期。同时我们还可以进一步调整等级阈值范围，可有效提升准确率。检测结果表明，利用决策树C4.5算法得到的教学模型对未来测试样本的正确率能够达到预先要求，也达到了我们正确评估学习效果的初衷。

指导建议：

（1）根据以上结论，如何有效安排在线学习。例如，某学生希望取得“良”的形成性评估成绩，那么他进行在线学习时，登录总时长、视频浏览时长、章节练习完成率、摸底考试成绩以及参与答疑发帖和互动讨论均应达到相应要求。

（2）教师要提供教学反馈信息，及时改进教学过程，提高学生分数。根据翔实的统计数据，教师可以掌握每个学生学习的情况和整体趋势，以及各章节模块学习的详细情况。从而对学生难点，进行针对性的指导。

五、小结

本文首先介绍了数据挖掘和C4.5算法决策树的相关原理，接着对所授课程在线学习过程构建评判系统，即通过对原始数据量化处理，分类属性类别，最终构建学习行为投入与评判结果的决策树模型。经测试和分析，该决策树构建预测了学生学习行为数据和学习评判结果之间的关系，并反向推测试验了模型的正确性。本文力图解决传统决策树算法只能对对称型数据处理的弊端，尝

试对连续型数据量化和数据处理，同时也在算法精度和拟合度进一步提升，消除了以往过度匹配问题。本文的研究为评估学生在线学习合理投入，提供了较为科学的方法和手段。由于受研究样本数量的局限，导致对学习行为的分析还存在不够全面的问题，需要后期更加深入的研究，进一步发挥决策树算法模型优势，更好地指导课程教学改革实践。

参考文献：

- [1] 杨露. 基于数据挖掘技术的就业信息管理系统设计与实现[J]. 中小企业管理与科技, 2015(29): 202.
 - [2] 黄荣喜. 高职院校就业信息平台中决策树应用研究[J]. 广西教育, 2016(39): 179-181.
 - [3] 史达伟, 耿焕同, 吉辰, 黄超. 基于C4.5决策树算法的道路结冰预报模型构建及应用[J]. 气象科学, 2015(02): 204-209.
- 基金支持项目：陆军军医大学2019年教育研究课题（2019B02）：数据挖掘技术在《计算机程序设计》网络教学平台中的应用研究；陆军军医大学基础医学院2019年教育研究课题（2019JCXYB12）：基于大数据技术的混合式教学实践研究

作者简介：李俊杰（1980—），男，硕士，副教授，计算机科学与技术。