文章编号：2095-2163(2020)08-0127-04 中图分类号：TP391.1 文献标志码：A

# 基于决策树的数据挖掘学生行为分析

陈馨瑶
（浙江农林大学 信息工程学院，杭州 311300）

**摘 要**：在大数据飞速发展的时代，如何合理利用高校大数据对学生进行标准化以及个性化服务管理，成为当前高校管理者的重点关注问题。通过对决策树分类算法的研究，采用决策树 C4.5 算法，从学校数据中心读取数据后进行预处理，除去异常数据后，基于一卡通消费数据以及高校图书馆进出数据进行数据挖掘与分析，通过消费数据以及图书馆进出数据与学生成绩绩点的对比分析得出结果，为高校管理服务提供更全面化的支持。

**关键词**：高校大数据；决策树算法；数据挖掘与分析

## Analysis ofstudent behavior in data mining based on decision tree

CHEN Xinyao
（School of Information Engineering, Zhejiang Agriculture and Forestry University, Hangzhou 311300, China）

**【Abstract】** In the era of the rapid development of big data, the data of colleges and universities are also increasing rapidly. How to make rational use of big data to standardize students and individualized service management has become the focus of attention of university managers at present. According to the research of decision tree classification algorithm, the decision tree C4.5 algorithm is used to read the data from the school data center and preprocess the data. After removing the abnormal data, the data mining and analysis are carried out based on the consumption data of card and the data of university library.

**【Key words】** university big data; decision tree algorithm; data mining and analysis

## 0 引 言

随着高校信息化环境中累积数据的不断增加，形成了一个较全面的大数据环境。如何对高校大数据进行有效共享以及交换管理，并通过大数据挖掘分析思想与方法，改进校园师生管理模式，同时利用大数据分析结果为校园生活提供更加清楚详细的服务，是现如今高校服务体系建设所不能忽视的问题。

众多高校利用校园大数据开发了许多针对校园管理和师生服务的应用，依赖大数据挖掘方法对学生个人行为进行分析，得到学习行为特征，创建个性化学习库。利用校园大数据开发了许多针对校园管理和师生服务的应用，依赖大数据挖掘方法来支撑高校教育的校园管理政策以及学生个人行为的分析，受到普遍关注。从学生画像系统到数据统一管理平台的建设都代表着中国高校信息化建设在教育领域中的迅猛发展。

## 1 基本概念及算法

决策树（Decision Tree）的概念是指在已知各种情况发生的概率，构造决策树来求出净现值的期望值大于等于零的概率，以此来判断分析方法的可行性[1]。

决策树的基本算法：

（1）ID3 算法。ID3 的核心是在决策树各个节点上利用信息增益准则选择特征，递归地构建决策树[2]。

①从根节点开始计算节点所有特征的信息增益，由该特征的不同取值建立子节点；

②对子节点递归地调用以上方法；

③直至所有特征的信息增益很小或无特征为止，得到决策树

（2）C4.5 算法。C4.5 算法继承了 ID3 算法的优点，利用信息增益率替代信息增益来选择属性，克服了 ID3 算法中属性选择偏取值多的属性问题。同时在树的构造过程中进行剪枝，且能够完成对连续属性的离散化处理和不完整数据的处理[3]。

（3）CART 算法。CART 算法由以下两步组成：

①决策树生成：基于训练数据集生成决策树，生成的决策树要尽量大；

②决策树剪枝：用验证集对已生成的树进行剪枝并选择最优子树，用损失函数最小作为剪枝的标准。

## 2 决策树算法的实际应用

### 2.1 学生日常行为与成绩分析

通过对中心数据库中数据的实时监控，抽取需

作者简介：陈馨瑶（1995-），女，硕士研究生，主要研究方向：数据共享与交换、数据挖掘、数据分析。

收稿日期：2020-06-22

要的数据分类存入数据仓库中,利用工具和算法对数据进行分析,实现可视化监控。以此保障高校管理人员对学生异常行为的监管以及预警。

针对实际情况,对数据库中学生行为数据进行预处理后,通过 C4.5 算法对学生行为数据进行挖掘分析。如:$m$ 个样本的连续特征 $A$ 有 $m$ 个,从小到大排列为 $a_1,a_2,\ldots,a_m$,则 C4.5 取相邻两样本值的平均数,共取得 $m-1$ 个划分点。其中第 $i$ 个划分点 $T_i$ 表示为:$T_i=\frac{a_i+a_{i+1}}{2}$。分别计算以该点作为二元分类点时的信息增益。选择信息增益最大的点作为该连续特征的二元离散分类点。如取到的增益最大的点为 $a_t$,,则小于 $a_t$ 的值为类别 1,大于 $a_t$ 的值为类别 2,这样就做到了连续特征的离散化。要注意的是,与离散属性不同的是,如果当前节点为连续属性,则该属性后面还可以参与子节点的产生选择过程。表 1、表 2 分别给出了与学生相关数据,具体计算步骤如下。

**表 1 学生去图书馆次数以及平均绩点**

**Tab.1 Number of student visiting the library and GPA**

| 学号 | 性别 | 天数 | 平均绩点 |
|---|---|---|---|
| 201701310229 | 男 | 48 | 4.25 |
| 201701310128 | 男 | 112 | 4.2 |
| 201701310331 | 男 | 94 | 3.01 |
| 201701310301 | 女 | 111 | 3.87 |
| 201701310210 | 女 | 106 | 3.83 |
| 201701310118 | 女 | 132 | 3.82 |
| 201701310101 | 女 | 122 | 3.78 |
| 201701310213 | 女 | 119 | 3.77 |
| 201701310201 | 女 | 41 | 3.76 |
| 201701310311 | 女 | 104 | 3.73 |
| 201701310115 | 女 | 125 | 3.71 |

**表 2 学生日常行为与成绩**

**Tab. 2 Students´ daily behavior and performance**

| 学号 | 性别 | 是否图书馆 | 平均成绩 |
|---|---|---|---|
| 1 | 男 | 是 | 85 |
| 2 | 男 | 是 | 76 |
| 3 | 女 | 否 | 87 |
| 4 | 男 | 否 | 69 |
| 5 | 女 | 是 | 92 |
| 6 | 女 | 否 | 56 |
| 7 | 女 | 否 | 61 |
| n | 男 | 是 | a |

①对数据源进行预处理。

②计算每个属性的信息增益和信息增益率。

③在根节点属性中,每个可能的取值都对应了一个子集,对样本子集递归地执行第二步,直至规划的每个子集中的观测数据在分类属性上取值都一致,以此来生成决策树。

④根据构造的决策树提取分类规则,对新的数据集进行分类。

特征数越多的特征对应的特征熵越大。其作为分母,可以校正信息增益容易偏向于取值较多的特征的问题。C4.5 的思路是将数据分成二部分,对每个样本设置一个权重(初始可以都为 1),然后划分数据。一部分是有特征值 A 的数据 D1,另一部分是没有特征 A 的数据 D2.。然后对于没有缺失特征 A 的数据集 D1 与对应的 A 特征的各个特征值,一起计算加权重后的信息增益比,最后乘上一个系数。这个系数是无特征 A 缺失的样本加权后所占加权总样本的比例。

$$entopy(\text{平均成绩})=-\frac{a_1}{a}\log_2\frac{a_1}{a}-\frac{a_2}{a}\log_2\frac{a_2}{a}=x_1. \tag{1}$$

$$entopy(\text{男})=-\frac{b_1}{n}\log_2\frac{b_1}{n}-\frac{b_2}{n}\log_2\frac{b_2}{n}=x_2. \tag{2}$$

$$entopy(\text{女})=-\frac{b_2}{n}\log_2\frac{b_2}{n}-\frac{b_1}{n}\log_2\frac{b_1}{n}=x_3. \tag{3}$$

$$entopy(\text{性别})=\frac{b_1}{n}*x_2-\frac{b_2}{n}*x_3=x_4. \tag{4}$$

信息增益 :信息熵表示的是不确定度。均匀分布时,不确定度最大,此时熵就最大。当选择某个特征对数据集进行分类时,分类后的数据集信息熵会比分类前的小,其差值表示为信息增益。信息增益可以衡量某个特征对分类结果的影响大小。

$$Gain(\text{性别})=x_1-x_4=x_5. \tag{5}$$

分裂信息:

$$Split\_Info(\text{性别})=-\frac{b_1}{n}\log_2\frac{b_1}{n}\log_2\frac{b_2}{n}=x_6. \tag{6}$$

增益率:

$$Gain\_Ratio(\text{性别})=\frac{x_5}{x_6}=x_6. \tag{7}$$

在上述计算中,$a$ 表示平均成绩;$b$ 表示性别;$n$ 表示总人数。 通过对增益值和增益率的计算可以推出,女生相较男生更愿意去图书馆,并且女生成绩较男生成绩更好。

同理计算出去图书馆对平均成绩的影响,选择最大的增益率作为决策树节点,以此类推,最后具体展示各个因素对学生成绩的影响以及影响权重,方便学校对学生学习的把握。

## 2.2 实验结果与分析

本节主要通过对 2018 年学生使用一卡通在食

堂消费以及出入图书馆情况与学生的期末绩点进行对比分析，得出这些行为与学生成绩的基本关系。

以 2016-2018 级本部本科生为例分析，其中 2018 年 3-6 月、9-12 月 2016-2018 级本部本科生在食堂的消费总人数为 7 870 人，3-6 月总消费记录数为 1 781 557 条，9-12 月与总消费记录数为 2 631 610 条，合计共 4 413 167 条。

本文对以上数据进行分段累加与计数，并进行统计分析。具体建模方法如下：

诸多数据库表字段中选取需要使用的部分，如消费流水表，原表有 50 多个字段，只抽取其中相关的几个字段，主要包含刷卡消费的关键信息：用户学工号、消费金额、消费天数、消费时间、年级、学院。

按时间段分析：2018 年 3-6 月、2018 年 9-12 月大部分学生在校期间。

按年级段：本次仅对 2016-2018 级校本部本科生进行分析。

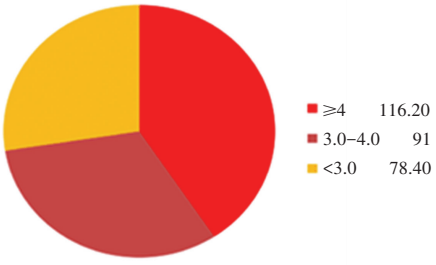


**图 1　学生 8 点前消费次数与绩点相关情况分析**

**Fig. 1　Analysis of the correlation between consumption and GPA before 8 o´clock in 2018**

绩点的高低顺序与消费次数的高低次序相一致，绩点较高的学生平均消费次数也较多。可见八点前平均消费次数的情况可以反映学生的学习情况，且八点前有过消费记录的同学均无绩点 2.8 以下情况。

总体而言九点前平均消费次数高的人群挂科门数少，但是九点前平均消费次数最少的人群挂科门数并不是最多的。九点前平均消费次数为 70 的占人数的大部分，消费次数与人数呈正相关关系。

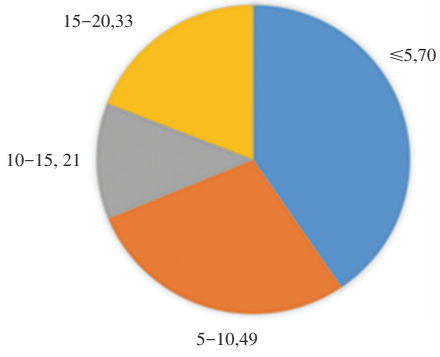


**图 2　9 点前一卡通消费次数与挂科门数的分析统计**

**Fig. 2　Analysis and statistics on the number of card consumption and the number of door hangings before 9 o´clock 2018**

**表 3　2018 年 9 点前一卡通消费次数与挂科门数的分析统计**

**Tab. 3　Analysis and statistics on the number of card consumption and the number of door hangings before 9 o´clock 2018**

| 挂科门数 | 9 点前平均消费次数 | 人数 |
|---|---|---|
| ≤5 | 70 | 3167 |
| 5-10 | 49 | 181 |
| 10-15 | 21 | 29 |
| 15-20 | 33 | 9 |

在进行图书馆进馆天数与学习成绩关联性分析之前，首先对数据进行了异常处理。排除了空白数据以及平均绩点在 1.0 以下的数据，以保证统计分析的准确性与可靠性，同时针对男生与女生进行了区分统计，分开对比。

表 4 为全校学生 2018 年度的各项努力程度指标的聚类结果。

**表 4　学生努力程度聚类结果**

**Tab. 4　Student effort clustering results**

| 类编号 | 学生占比/% | 课堂考勤 | 图书借阅 | 入馆次数 | 课程通过率 | 学习时长 | 绩点 |
|---|---|---|---|---|---|---|---|
| 0 | 8.2 | 0.98 | 23.6 | 107.3 | 1 | 317.9 | 3.42 |
| 1 | 5.7 | 0.90 | 3.7 | 31.5 | 1 | 57.3 | 3.32 |
| 2 | 6.8 | 0.99 | 13.8 | 139.7 | 0.95 | 387.2 | 2.69 |
| 3 | 29.5 | 0.95 | 4.2 | 37.5 | 0.97 | 73.5 | 2.57 |
| 4 | 17.2 | 0.72 | 2.1 | 25.9 | 0.93 | 33.7 | 2.27 |
| 5 | 15.2 | 0.67 | 3.7 | 26.3 | 0.90 | 37.8 | 2.09 |
| 6 | 17.4 | 0.98 | 1.3 | 11.8 | 0.96 | 23.5 | 2.31 |

2018 年本科生平均入馆天数多的人平均绩点较高。在每个平均绩点段男生入馆天数均少于女生，可见女生比男生自主学习的积极性更高。

同时，针对 2018 早上 8 点前有一卡通消费记录

后,又在图书馆有入馆记录的本科生,在数据库中将这些记录导出并计算次数,与其 2018 年的平均绩点进行对比,通过综合平均次数进行对比绩点。在统计时,同样做了异常处理,去除了绩点在 1.0 以下的数据以保证数据的准确性与可靠性。分析结果如图 4 所示。

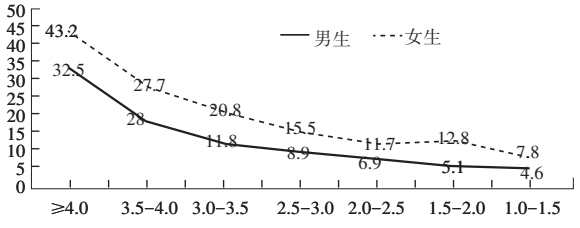


**图 3 2018 年本科生平均进馆天数与平均绩点相关分布情况分析**

**Fig. 3 Analysis on thedistribution of average entry days and GPA of undergraduate students in 2018**

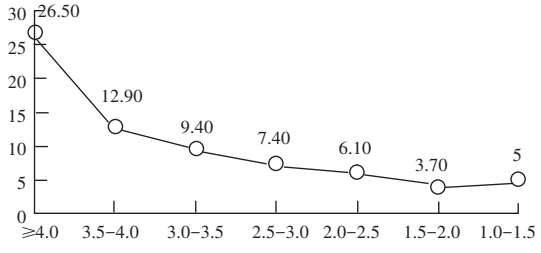


**图 4 平均综合次数与平均绩点对比分析趋势**

**Fig. 4 Trends in comparative analysis of average number of combinations and GPA**

排除平均绩点在 1.0-1.5 的异常结果,平均综合次数与平均绩点总体呈正相关关系,平均综合次数高平均绩点也相对较高。可见在早上 8 点前有一卡通消费记录后又在图书馆有入馆记录的本科生自我管理能力较强且成绩较理想。

## 3 结束语

校园数据主要以教职工和学生为主,除了学生行为画像,针对师生在校园内的行为可以通过数据分析提供更多帮助。本文通过决策树算法对学生消费行为和学习行为进行分析,并通过对比成绩绩点数据,挖掘了消费行为和学习行为与成绩之间的关系。这些数据资料在数据分析、数据挖掘领域具有较高的应用价值。可以针对师生行为的数据分析与挖掘,为师生提供更好的帮助与服务,为高校管理者提供更多的管理层面的数据支持,保障高校建设与管理的高效运行。

## 参考文献

[1] 刘继玺. 基于云服务模式的智慧农作移动平台构建[D]. 南京农业大学,2016.
[2] 廖琳. 决策树在学生信息管理系统中的应用研究[D]. 广西大学,2014.
[3] 田欣.决策树算法的研究综述[J]. 现代营销(下旬刊),2017(1):36.
[4] 孙昊. 学习行为分析与学业预警系统研究与设计[D]. 苏州大学,2017.

(上接第 126 页)

**表 2 不同算法去噪后的均方误差**

**Tab. 2 Mean square error of different algorithms after denoising**

| 脉冲噪声+高斯噪声 | 0.01+0.01 | 0.1+0.02 | 0.03+0.1 | 0.05+0.05 |
|---|---|---|---|---|
| 噪声图像 | 0.012 6 | 0.048 | 0.075 0 | 0.053 3 |
| 硬阈值函数 | 0.002 0 | 0.004 9 | 0.005 8 | 0.004 2 |
| 软阈值函数 | 0.001 1 | 0.002 9 | 0.004 7 | 0.003 |
| 中值滤波去噪 | 0.001 8 | 0.004 5 | 0.016 5 | 0.009 2 |
| 本文算法 | 0.000 94 | 0.001 5 | 0.003 1 | 0.002 1 |

## 6 结束语

针对图像信号中含有的混合噪声用单一的方法无法得到好的去噪效果问题,本文结合中值滤波和小波变换中的阈值去噪各自的优点,对小波阈值函数进行改进,对含有混合噪声的图像进行处理。经过仿真实验并比较信噪比和均方误差,结果表明这种综合去噪方法优于单一的去噪方法,具有实用性和有效性,能够改善含有混合噪声的图像质量,从而为后期的图像处理奠定了基础。

## 参考文献

[1] 王蓓,张根耀,李智,等. 基于新阈值函数的小波阈值去噪算法[J]. 计算机应用,2014,34(5):1499-1502.
[2] 关雪梅. 一种基于中值滤波和小波变换的图像去噪处理算法研究[J]. 中州大学学报,2020,37(1):121-124.
[3] 方挺,任文文. 中值滤波和小波变换相结合的图像去噪研究(英文)[J]. 无线互联科技,2016(14):119-120.
[4] 李智,张根耀,王蓓,等. 基于中值滤波和小波变换的图像去噪[J]. 现代电子技术,2014,37(13):72-74.
[5] 倪培峰,胡雄. 一种基于改进阈值函数的小波阈值降噪算法[J]. 电子技术应用,2016,42(8):98-100,104.
[6] 周峡,徐善顶. 一种改进小波阈值函数的图像去噪方法研究[J]. 南京工程学院学报(自然科学版),2019,17(4):44-49.
[7] 刘永平,郭小波. 基于新阈值函数和小波分析的数字图像去噪方法[J]. 电脑与信息技术,2020,28(2):5-7,20.
[8] 杨立. 基于改进小波阈值函数的图像去噪[J]. 重庆理工大学学报(自然科学),2013,27(2):93-95,125.
[9] 代少升,崔俊杰,张德洲,等. 基于中值滤波和小波变换的红外图像去噪方法[J]. 半导体光电,2017,38(2):299-303.
[10] 张绘娟,张达敏,闫威,等. 基于改进阈值函数的小波变换图像去噪算法[J]. 计算机应用研究,2020,37(5):1545-1548,1552.