高校贫困学生辅助评价的研究①

林志兴

(三明学院 现代教育技术中心, 三明 365004)

摘 要: 高校对贫困学生的认定存在认定成本高、可信度不强以及标准不一致等问题. 本文通过分析高校一卡通的消费行为, 刻画贫困学生的消费特征, 应用马尔科夫模型对贫困学生消费行为进行建模, 并提出了相似指标的概念和计算方法. 通过对学生消费行为与贫困学生消费行为模型进行相似指标计算, 对贫困学生进行认定. 该方法具有计算效率高, 速度快, 计算成本低廉以及数据获取容易, 在同一所学校中评价标准一致, 对贫困学生的平均识别率达到 90%以上等特点, 可以作为高校在复评贫困学生的一个有力的辅助工具.

关键词: 一卡通; 马尔科夫模型; 状态概率转换矩阵; 贫困学生认定

Research on Assistant Identification of Poor Students in Colleges

LIN Zhi-Xing

(Modern Educational Technology Center, Sanming College, Sanming 365004, China)

Abstract: There are some problems in the identification of poor students in colleges and universities, such as high cost, the lack of credibility and the inconsistency of standards. Through anglicizing the consumption behavior of campus one card solution, we can depict the consumption characterization of poor students. And then we build Markov model for consumption behavior of poor students, and put forward the concept and calculation method of line index. Based on calculating the similarity index for both the consumer behavior of student model and the consumption behavior of poor student model, we can identify the poor students. This method has the characteristics of high computational efficiency, fast speed, low cost and easy data acquisition, which has consistent evaluation criteria in the same school. And the average recognition rate for poor students is than 90%. It can be used as a powerful auxiliary tool for the identification of poor students in colleges and universities.

Key words: campus card system; Markov model; state probability transition matrix; identification of poor students

高校对贫困学生的认定, 主要是依据学生家庭的人均收入以及民政部门的证明, 但是这种认定方法由于没有确定具体的判定模型, 不一定能够真正准确掌握学生家庭的实际经济状况, 不能做到客观鉴定. 另外, 考虑到部分贫困学生的特殊心理, 贫困学生有可能不会主动提供相应材料, 对学校评定工作的开展会有较大影响, 可能导致贫困资助并没有落实到真正需要资助的学生身上. 虽然高校贫困学生认定政策总体上看是合理的, 但是在实际工作开展过程中由于各方面的客观原因, 还是会存在一定的问题.

本文通过分析高校使用的一卡通系统中学生消费行为, 应用数据挖掘技术构造一个贫困学生的评分模型, 为学校贫困生评定提供依据. 该方法特别适用于识别一些冒领贫困生助学金的行为, 以及一些贫困学生因心理压力和个人隐私问题不能主动申请的现象.

本文接下来从四个部分进行介绍, 第一部分介绍贫困学生认定工作以及校园一卡通的应用, 第二部分介绍本文提出的方法, 第三部分介绍开展的实验以及相应的分析结构, 第四部分对本文进行了总结和展望.

1 相关研究

1.1 贫困生认定现状

① 基金项目:福建省教育厅中青年教师科研项目高校教育信息化专项资助[JA15463]
收稿时间:2016-10-08;收到修改稿时间:2016-11-10 [doi: 10.15888/j.cnki.csa.005816]

现在高校主要采用两种方法进行贫困学生的认定[1], 第一种方式是通过生源地政府出具的证明材料, 这是高校进行贫困学生认定的主要依据. 第二种方式是通过评选认定, 该方式由学生主动申请, 贫困生认定工作组根据申请材料以及证明材料等进行审核, 并考察其日常支出等情况, 并进行民主评议, 最后确定贫困学生.

以上两种方法可以比较有效地认定贫困学生, 但是还是存在一些问题[2]. 首先是需要耗费较多的资源进行认定, 其次是仅仅根据书面材料进行认定, 其可信度较低. 由于在生源地政府开具证明受到人为影响较大, 特别是开具虚假证明追溯成本比较高, 没有高校会逐一去生源地确认证明是否虚假, 因此, 该认定方式可信性较差. 第三是没有统一的界定标准. 由于各个地区经济发展不均衡, 各地的贫困标准和程度都是相对于当地的情况, 导致贫困学生的认定结果地区差异性很大.

因此, 需要找到一种方法来回避以上的缺陷, 其中, 考察学生的日常支出是一种有效的方法, 但是, 该方法如果采用人工调查方式, 也将存在耗费成本高, 可信度低的问题. 在这里, 我们提出了一种基于一卡通消费行为的数据挖掘方法, 对已经入学使用一卡通消费的学生进行评分, 作为高校贫困学生认定的一个辅助工具.

1.2 一卡通在贫困生评价中应用

随着网络信息化的发展, 校园一卡通应用越来越多, 其发展也越来越被重视, 已经成为高校信息管理的核心部分. 为了便捷管理, 一卡通已经实现食堂就餐[3]、校内超市购物[4,5]、计算机机房上网、门禁管理、考勤管理、校医院看病买药等校园“一站式”的服务. 在使用校园一卡通的过程中, 广大学生和教职工产生大量消费数据.

对校园一卡通使用者个体产生的数据, 可以应用数据挖掘技术, 分析挖掘使用者个体的消费习惯. 在校园一卡通中, 消费的主体是学生群体, 通过分析学生的行为习惯[3], 比如消费习惯[4,5]、学习习惯、生活习惯等, 可以刻画学生行为特征, 而根据这些特征进行多维分析或者因果分析, 可以对校园建设[6]、校园管理、学生评估[7-10]等作出更合理的决策.

已经有一部分研究开展应用一卡通消费数据进行贫困学生认定的工作, 费小丹等人[7]应用聚类技术, 提取贫困学生的特征, 并提出一个可以直观反映学生贫困程度的贫困指数. 夏冉等人[8]构建了一个校园信息化平台, 其中应用数据挖掘技术分析学生行为, 作为学生资助工作的重要信息来源. 张冬冬等人[9]运用 C4.5 决策树分类算法, 分析某高校贫困学生信息, 并根据决策树规则, 建立了一套比较完善的贫困生认定规则体系, 为高校贫困生认定工作提供决策支持. 张建明[10]应用决策树生成了完整的决策树及其规则. 在此基础上,将决策树挖掘算法应用于贫困生评定工作中. 单菊芬[11]通过建立一卡通消费记录的数据仓库、以及 OLAP 数据模型, 然后应用 OLAP 模型通过多维展示, 解决系部、专业、生源地、家庭收入状况等指标的多维分析问题, 并应用关联分析对贫困学生的特征进行发现, 根据知识发现的结果, 为贫困生评价机制提出改进的建议. 秦微微[12]应用统计方法对一卡通数据进行处理、整合、描述、建模以及推断. 通过数据挖掘方法, 确定贫困学生与普通学生在消费方面具有显著性差异的指标, 如: 大消费额比例、相对就餐消费得分等, 应用统计检验, 证明这些指标可以作为高校贫困生合理评判的辅助依据. 何倩[13]基于层次分析法理论提出了一个多层次的贫困生认定指标体系结构, 在这个体系结构中, 包含十个具体的指标的认定体系, 可以通过不同层次权重的调整, 直接应用管理决策分析技术. 王德才[14]应用支持向量机分类算法,对一卡通消费记录进行非线性分类, 确定不同类别的特征. 其次, 在应用关联规则算法对支持向量机分类结果进行分析, 提取频繁模式,并针对关联规则效率不高的问题, 提出了改进思路, 利用矩阵的运算特性和增量方式, 有效的提高了关联规则的计算效率. 该方法针对部分贫困生消费行为模式进行对比分析, 挖掘贫困生的消费频繁模式, 同时采用统计角度进行消费总额、消费频度、消费平均值分析, 在贫困学生的识别应用上得到较好的效果.

2 方法

本文首先通过研究贫困生使用一卡通行为, 刻画贫困生的一卡通使用特征. 根据使用特征, 构建一个马尔科夫模型, 并以此模型对学生使用行为进行评分, 该评分作为认定或识别贫困生的一个参考.

2.1 贫困生使用一卡通行为特征研究

首先, 从全校抽取三个班级的学生作为样本, 提

取他们 2016 年 3 月 1 日至 2016 年 5 月 31 日共 92 天的一卡通使用记录.

其次, 标识学生类别, 分为普通学生和贫困学生.

第三, 计算两种类别学生在这 92 天内, 早中晚三餐的就餐次数以及三餐平均金额和方差. 表 1 列出了两种类别学生的就餐情况.

表 1 贫困学生与普通学生三餐就餐情况

	贫困学生			普通学生		
	平均就餐次数	每餐金额均值	每餐金额方差	平均就餐次数	每餐金额均值	每餐金额方差
早餐	63.4	3.67	1.23	62.5	3.89	1.87
午餐	81.7	6.23	1.12	78.2	7.17	2.03
晚餐	80.2	5.76	1.08	73.8	8.06	2.45

第四, 分析两种类别学生就餐情况, 确定两类学生在哪些方面具有显著差别, 也就是确定区分两类学生的特征.

从表 1 可以看出, 贫困学生与普通学生在早餐行为上差别不显著, 但是在午餐和晚餐方面, 差别比较明显, 无论从平均就餐次数、金额方面都有显著不同. 而且可以看出, 贫困学生的午餐晚餐金额方差也显著低于普通学生, 这说明贫困学生午餐晚餐行为比较有规律, 基本都在学校食堂就餐, 而且每餐消费能力显著低于普通学生.

因此, 可以看出贫困学生在午餐晚餐的平均就餐次数和每餐平均金额上与普通学生显著不同. 我们可以将这几个指标作为区分两类学生的特征, 并作为下一步分析的基础. 在此, 我们定义贫困学生就餐次数为 C, 每餐平均金额为 E, 午餐用下标 L 表示, 晚餐用下标 D 表示. 即: 贫困学生午餐平均就餐次数为 C_L 和午餐每餐平均金额 E_L, 贫困学生晚餐的平均就餐次数为 C_D 和晚餐每餐平均金额 E_D.

2.2 马尔科夫链以及状态转移概率矩阵

马尔科夫模型是研究无后效性随机过程的一个主要模型, 其经常应用在消费行为分析以及异常行为分析中[16], 其基础是马尔科夫链. 马尔科夫链(Markov Chain), 描述了一种状态序列 $X_1, X_2, X_3, \ldots\ldots$, 其每个状态值取决于前面有限个(一个或几个)状态. 这些状态所有可能取值的集合, 被称为“状态空间”, 而 X_n 的值则是在时间 n 的状态. 如果 X_{n+1} 对于过去状态的条件概率分布仅仅由前一个状态决定, 则可以认为 X_{n+1} 对于过去状态的条件概率分布是 X_n 的一个函数,

因此:

$$P\left(X_{n+1}=x\middle|X_1=x_1,X_2=x_2,\ldots,X_n=x_n\right)$$
$$=P\left(X_{n+1}=x\middle|X_n=x_n\right)$$

马尔科夫链在实际中得到广泛的应用[15].

在马尔科夫模型中, 经常会使用到状态转移概率矩阵. 状态转移概率就是在当前状态 X_n 等于 i 的条件下, 使下一个状态 X_{n+1} 等于某一状态 j 的条件概率. 状态转移概率矩阵记录了当前时间的各个状态到下一时间各个状态的转换概率. 其数学公式为:

$$Pij=P\left(x_{n+1}=j\middle|x_n=i\right)$$

其中 i, j 分别表示当前时间和下一时间的某一状态, 可以看出, 状态转移概率矩阵是一个 $m*m$ 的矩阵, 其中 m 为该马尔科夫链的状态空间的大小.

以一卡通行为分析为例, l 表示在学校用午餐, $\bar{l}$ 表示没在学校用午餐, d 代表在学校用晚餐, $\bar{d}$ 代表没有在学校用晚餐, 那么 P_{ld} 表示某位学生在一段时间内, 在学校用午餐而且也在学校用晚餐的概率, $P_{l\bar{d}}$ 表示某位学生在学校用午餐但是没有在学校用晚餐的概率.

2.3 建立贫困学生评分系统

根据 2.1 中的分析, 可以看出午餐晚餐的就餐次数和每餐平均金额可以作为区分两类学生的特征. 结合马尔科夫模型, 我们进行评分系统的模型构建.

首先, 在原来抽取的样本数据中, 进一步计算贫困学生午餐与晚餐的马尔科夫状态转移概率矩阵, 如表 2 所示.

表 2 贫困学生马尔科夫状态概率转移矩阵

	晚上就餐		晚上没有就餐	
	概率	每餐平均金额	概率	每餐平均金额
中午就餐	P_{LD}	E_{LD}	$P_{L\bar{D}}$	$E_{L\bar{D}}$
中午没有就餐	$P_{\bar{L}D}$	$E_{\bar{L}D}$	$P_{\bar{L}\bar{D}}$	$E_{\bar{L}\bar{D}}$

其中, P_{LD} 表示贫困学生中午就餐而且晚上也就餐的概率, E_{LD} 代表这两餐的平均金额;

$P_{L\bar{D}}$ 表示贫困学生中午就餐而且晚上没有就餐的概率, $E_{L\bar{D}}$ 代表这两餐中午餐的平均金额;

$P_{\bar{L}D}$ 表示贫困学生中午没有就餐而晚上就餐的概率, $E_{\bar{L}D}$ 代表这两餐中晚餐的平均金额;

表示中午没有就餐而且晚上没有就餐的概率, $E_{\bar{L}\bar{D}}$ 代表这两餐的平均金额. 很明显, 平均金额为 0.

其次, 定义相似指标 S, 将公式 1 以及贫困学生午餐晚餐平均就餐次数和每餐平均金额作为评判贫困学生的标准, 与这些特征相似的, 说明是贫困生的可能性就越

大, 如果不相似, 说明不是贫困生的可能性很大.

同时需要计算每个学生使用一卡通行为特征, 在此, 我们定义如下:

给定一个学生,

① 计算该学生午餐晚餐相应统计值: 午餐平均就餐次数为 C_l, 午餐每餐平均金额为 E_l, 晚餐餐平均就餐次数为 C_d, 晚餐每餐平均金额为 E_d.

② 计算该学生午餐晚餐马尔科夫状态转移概率矩阵以及相应统计值:

其中每个学生中午就餐而且晚上也就餐的概率为 P_{ld}, E_{ld} 代表这两餐的平均金额;

中午就餐而且晚上没有就餐的概率 $P_{l\bar{d}}$, $E_{l\bar{d}}$ 代表这两餐中午餐的平均金额;

学生中午没有就餐而晚上就餐的概率 $P_{\bar{l}d}$, $E_{\bar{l}d}$ 代表这两餐中晚餐的平均金额;

中午没有就餐而且晚上没有就餐的概率 $P_{\bar{l}\bar{d}}$, $E_{\bar{l}\bar{d}}$ 代表这两餐的平均金额,很明显, 平均金额为 0.

由每个学生使用一卡通行为特征以及贫困学生使用一卡通行为特征, 我们可以定义以下相似指标:

首先, 我们定义公式 1. 公式 1 中 W_1 为待测学生与贫困学生在午餐晚餐之间的差距, 可以看出, 当 W_1 的值越大, 待测学生被判别为贫困学生的可能性就越低. 其中 *days* 中为计算的天数.

$$W_1=\left|\frac{C_L-C_l}{days}\times(E_L-E_l)\right|+\left|\frac{C_D-C_d}{days}\times(E_D-E_d)\right| \quad (1)$$

其次, 我们定义公式 2 中 W_2 为待测学生与贫困学生在马尔科夫状态转换概率矩阵之中的差距, 可以看出, 当 W_2 的值越大, 待测学生被判别为贫困学生的可能性就越低.

$$\begin{aligned}W_2=&\left|(P_{LD}-P_{ld})*(E_{LD}-E_{ld})\right|+\left|(P_{L\bar{D}}-P_{l\bar{d}})*(E_{L\bar{D}}-E_{l\bar{d}})\right|\\&+\left|(P_{\bar{L}D}-P_{\bar{l}d})*(E_{\bar{L}D}-E_{\bar{l}d})\right|+\left|(P_{\bar{L}\bar{D}}-P_{\bar{l}\bar{d}})*(E_{\bar{L}\bar{D}}-E_{\bar{l}\bar{d}})\right|\end{aligned} \quad (2)$$

最后, 我们从公式 3 中定义了相似度 *S*, 其中, *S* 值越大, 说明被判别为贫困学生的可能性就越高.

$$S=\frac{1}{2^{(W_1+W_2)}} \quad (3)$$

可以看出, 相似指标 S 越大, 就说明与贫困学生的特征越相似, 具有更大概率被判别为贫困学生.

2.4 评价标准

正确率是广泛用于信息检索和统计学分类领域的度量值, 用来评价结果的质量. 正确率是检索出相关文档数与检索出的文档总数的比率, 衡量的是检索系统的查准率.

正确率 = 提取出的正确信息条数 / 提取出的信息条数

本实验中, 我们采用准确率来衡量已经定义的相似指标 *S* 的准确程度,

准确率= 预测正确的贫困学生数量 / 贫困学生总数量 (4)

根据以上公式, 根据高校情况, 确定贫困学生的比例 *t*%, 因此贫困学生总数量为总学生数量的 *t*%. 同时我们选取 S 值为前 *t*%的学生作为预测为贫困学生, 并根据已经确定的贫困学生标记, 确认预测的正确性.

3 实验

3.1 实验数据

实验数据来源是某高校 2016 年 3 月至 5 月的一卡通食堂就餐流水记录, 共 243 万条, 总人数为 12563 人. 其中教职工人数为 783 人, 学生人数为 11753 人.

数据进行以下预处理:

① 确定早餐, 午餐, 晚餐时段

设置 6:00:00-10:00:00 为早餐时间, 该时段发生的刷卡流水记录作为早餐行为; 设置 10:00:01-15:00:00 置为午餐时间, 设置 15:00:01-20:00:00 置为晚餐时间.

② 合并同一卡号相同日期的同一时段数据

将同一个卡号的同一天同一时段的流水合并为一条记作一次就餐行为, 合计多条流水金额为一条流水金额(如同一天的多次午餐刷卡行为合并为一次午餐行为, 并且将多次流水的金额合计成为本次午餐消费金额).

3.2 模型建立与评分

步骤一, 将样本三个班级学生流水数据作为训练数据, 根据模型计算

① 贫困学生午餐平均就餐次数 C_L 和午餐每餐平均金额 E_L, 贫困学生晚餐平均就餐次数为 C_D 和晚餐每餐平均金额 E_D. 该特征已经由表 1 计算所得.

② 贫困生的马尔科夫状态转移概率矩阵数据, 如表 3 所示.

表 3　贫困学生的马尔科夫状态转移概率矩阵值

	晚上就餐		晚上不就餐	
	概率	平均金额	概率	平均金额
中午就餐	0.803	5.88	0.197	6.11
中午没有就餐	0.921	5.56	0.079	0

步骤二, 随机取一个班级作为测试数据, 该班级没有出现在样本班级中, 并登记该班级中每一位学生的类别属性, 确定为贫困学生或普通学生, 其中参数 *days* 为 92, 表示测试时间区段为 2016 年 3 月 1 日至 5 月 31 日, 共 92 天.

步骤三, 计算该班级每个学生的午餐晚餐就餐次数为 C_l, C_d, 以及每餐平均金额 E_l, E_d, 以及学生午餐与晚餐的马尔科夫状态转移概率矩阵的值. 如表 4 所示.

表 4 学生马尔科夫状态概率转移矩阵

	晚上就餐概率	晚上就餐每餐平均金额	晚上没有就餐概率	晚上没有就餐每餐平均金额
中午就餐	P_{ld}	E_{ld}	$P_{l\bar{d}}$	$E_{l\bar{d}}$
中午没有就餐	$P_{\bar{l}d}$	$E_{\bar{l}d}$	$P_{\bar{l}\bar{d}}$	$E_{\bar{l}\bar{d}}$

步骤四, 应用公式(1), (2), (3)计算该班级每个学生的贫困生评分, 并记录评分.

步骤五, 重复步骤二, 一共测试十个班级.

3.3 实验结果分析

实验测试了十个班级, 根据该校贫困生评定情况, 我们设置 t=15, 即选取前 15%的学生预测为贫困生, 并采用公式 4, 计算识别准确率, 结果如表 5.

表 5 中, 第二列为十个班级的人数, 第三列为每个班级的贫困学生数量 n, 第四列为根据公式计算出相似指标 S 在该班级前 n 名中, 实际为贫困学生的数量, 第五列是根据公式计算出的正确识别率. 可以从表 5 看出, 应用公式计算出来的相似指标 S 可以比较好的表示出贫困学生的特征. 实验证明用 S 进行识别贫困学生的可信度比较高, 可以达到 90%以上.

表 5 实验结果以及正确识别率

	班级学生	贫困学生	识别正确	识别率(%)
1	30	5	4	80.00
2	28	5	5	100.00
3	31	5	5	100.00
4	29	5	4	80.00
5	33	5	4	80.00
6	32	5	5	100.00
7	35	5	5	83.33
8	31	5	4	80.00
9	33	5	5	100.00
10	31	5	5	100.00
平均值				90.33

4 总结

本文通过分析一卡通的流水数据, 刻画贫困学生在使用一卡通的特征. 根据使用特征, 构建一个马尔科夫模型, 并以此模型对学生使用行为进行评分, 并将此该评分作为认定或识别贫困生的一个参考. 通过实验证明, 该方法有效地通过一卡通行为识别贫困学生, 其平均识别率达到 90%以上, 能够作为高校在复评贫困学生的一个有力的辅助工具. 该方法特别适用于识别一些冒领贫困生助学金的行为, 以及一些贫困学生因心理压力和个人隐私问题不能主动申请的现象. 该方法计算简单, 识别率高, 可以作为高校贫困学生管理系统的一部分.

参考文献

1 王平,龚文涛.基于 SOA 的高校贫困生认定体系的研究.微型电脑应用,2015,(10):55–56.

2 郑晓涛.高校贫困生认定工作的现实困境与建设路径.淮海工学院学报(人文社会科学版),2014,12(1):124–126.

3 张硕.基于 WEKA 的校园一卡通数据挖掘与分析[学位论文].武汉:华中师范大学,2014.

4 田丽萍.基于数据挖掘的超市客户消费数据研究[学位论文].天津:河北工业大学,2012.

5 陈锋.基于校园一卡通系统的高校用户就餐消费行为分析与数据挖掘.中国教育信息化,2014,(9):47–49.

6 曾秋凤.数据挖掘在校园一卡通数据库中的应用研究[硕士学位论文].南昌:江西师范大学,2008.

7 费小丹,董新科,张晖.基于校园一卡通消费数据的高校贫困生分析.电脑知识与技术,2014,(20):4934–4936.

8 夏冉,张小莉,姚建民.基于一卡通系统的数据挖掘技术与校园信息化关系的研究.科技风,2008,(4):33.

9 张冬冬,李玉龙,王玉鑫.数据挖掘技术在高校贫困生认定中的应用.西安文理学院学报:自然科学版,2013,16(4):90–94.

10 张建明.基于数据挖掘的高校贫困生认定系统设计和分析[硕士学位论文].南京:东南大学,2015.

11 单菊芬.基于数据挖掘技术的高校贫困生管理系统设计和分析[硕士学位论文].南京:南京邮电大学,2012.

12 秦微微.基于数据挖掘技术的高校贫困生评判指标的选取[硕士学位论文].长春:东北师范大学,2015.

13 何倩.基于层次分析法对高校贫困生认定指标体系的研究.黑龙江教育学院学报,2011,30(3):21–23.

14 王德才.数据挖掘在校园卡消费行为分析中的研究与应用[硕士学位论文].哈尔滨:哈尔滨工程大学,2010.

15 Ching WK, Ng MK. Markov chains: Models, algorithms and applications.Springer Berlin, 2012, 83(483):xiv.

16 韩忠明,张晨,李斌.基于 Markov 模型的异常用户检测.计算机仿真,2014,31(6):316–320.