华南农业大学学报(社会科学版)
2006年第3期 JOURNAL OF SOUTH CHINA AGRICULTURAL UNIVERSITY No. 3 2006
(第5卷) (SOCIAL SCIENCE EDITION) (Vol. 5)

模糊数据挖掘在高校学生管理中的应用

蔡秀娟[1], 蒋艳荣[2]

(1. 华南农业大学 食品学院,广东 广州 510642; 2. 华南理工大学 计算机科学与工程学院,广东 广州 510640)

摘 要: 分析了目前高校学生管理工作中存在的问题,探讨了模糊数据挖掘理论在高校学生管理中的应用。研究设计了一种数据挖掘方法,能够对高校学生进行聚类分析,从而发现高校内部学生的类型,并可判断某学生属于哪一种类型,对高校学生管理工作具有理论和现实意义。

关键词: 模糊数据挖掘;聚类分析;高校学生管理

中图分类号: G40-054 **文献标识码**: A **文章编号**:1672-0202(2006)03-0143-05

数据挖掘是从大型数据库或数据仓库中发现并提取隐藏在其中的信息或知识的过程,其目的是帮助分析人员寻找数据间潜在的关联,发现被忽略的要素,而这些信息对预测趋势和决策行为是十分有用的。从事务库中发现关联规则是数据挖掘中常用的方法[1,2]。近年来,因为模糊集理论的简洁性和与人类推理的相似性,它越来越多的用于智能系统中。集成了模糊理论的数据挖掘能够从数据中发现更通用和更重要的知识和规则。因此,模糊数据挖掘应用于越来越多的领域,并取得了良好的效果。

随着我国教育体制改革的深入,当代大学生的学习、生活、就业方式等发生了根本性的变化,接受信息的渠道多样化、复杂化。他们又生活在信息时代,易于接受新事物新思想,有较强的平等意识、公民意识、求实精神和批判精神。因此,在新形式下如何针对这些新情况对大学生进行客观、公平、有效的管理,是高校学生工作面临的新问题。同时,随着素质教育的全面推进和网络在校园的迅速发展,要求学生管理工作应该突出现代性,要求学生管理者具有更广泛的管理知识和现代化的管理手段[3]。

一般来说,高校学生管理工作主要包括学生综合测评、奖学金评选、助学金评选、就业管理、学生干部选拔、日常事务管理(如宿舍管理、违纪行为管理)等方面的管理工作。这些工作直接牵涉到学生的切身利益,也影响到学生今后的前途和发展,因而,在管理工作中如何做到客观和公平是非常重要的,而且,只有做到客观、公平、有效的管理,才能充分调动学生的学习积极性,促进学生健康发展。

尽管有一些学者试图将一些新的方法引入学生管理工作[4-6],目前的高校学生管理工作中的主观因素仍然太重、缺乏科学分析和根据,主要体现在:一方面,管理者是根据各种文件规定来管理学生的,而文件只是规定了各种条件程序或名额限制等,如果满足条件的学生超出名额限制时,如何客观合理的从中选出真正优秀的学生管理文件中没有给出具体的操作方法,这时候依赖于管理者的主观判断;另一方面,对学生的评价没有科学的理论根据,目前的管理方法一般由管理者根据有关文件对学生的德、智、体等方面进行量化打分,然后求出学生的评价值,在对学生量化打分时特别是德育分,没有一种科学的计算方法;另外,在选拔和冲突事务管理中,一般没有明确的文

收稿日期:2006-03-02

作者简介:蔡秀娟(1965-),女,广东郁南县人,华南农业大学食品学院助理研究员,主要研究方向为思想政治教育和教育管理.

件规定如何处理,到底如何处理,依赖于管理者所掌握的管理知识和管理手段。

对学生各方面表现的评估和评价是高校学生管理工作的基础,不论是综合测评、三好学生评选,还是学生干部选拔,都与对学生在相关方面表现的评估有关。而目前存在的评估方法都有一定的主观性、片面性,缺乏科学根据。因此,可以利用模糊数据挖掘理论中的方法来指导学生管理工作,克服目前学生管理工作中存在的问题。

一、数据挖掘的一般过程

数据挖掘的一般过程可用图1表示:

(1) 数据清理和集成:消除噪声或不一致的数据,并将多种数据源组合在一起。

(2)数据选择和变换:从数据库中检索与分析任务相关的数据,并将数据进行一定的变换,转换或者统一为适合挖掘的形式。如通过汇总或聚集操作。

(3)数据挖掘:为基本步骤,使用智能方式提取数据模式。

(4)模式评估:根据某种兴趣度度量,识别表示知识的真正有趣的模式。

(5)知识表示:使用可视化和知识表示技术,向用户提供挖掘的知识。

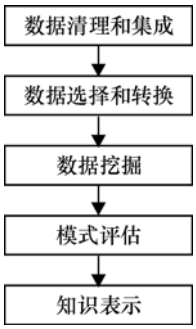


图1 数据挖掘的基本过程

二、模糊挖掘算法

在数据挖掘处理过程中,数据挖掘算法是最为关键的,利用模糊挖掘算法和从数据仓库中提取的数据,我们可以发现目前组织中存在的个体类型,另外还可以判断一个个体属于这些类型中的哪一类。

1. 模式发现

①在数据仓库的全体数据记录上,建立待分类的样本集U。

把要分类的对象称为样本,如 $u_1, u_2, \cdots u_n$,$U=\{u_1, u_2, \cdots, u_n\}$ 为样本集。应该将具体属性进行数量化,数量化的属性称为样本指标,设有m项指标,这可用m维向量描述样本,即:

$$u_i = \{u_{i1}, u_{i2}, \cdots, u_{im}\} \quad i = 1,2,\cdots,m \tag{1}$$

由于在实际数据中,采集到的数据往往不是[0,1]闭区间的数,所以应将这些原始数据标准化,按下式求出各数据的标准化值:

$$u'_{ik} = \frac{u_{ik} - \bar{u}_k}{S_k} \tag{2}$$

$\bar{u}_k$ 为其平均值,设样本的某一项指标k就可取到n个数据 $u_{1k}, u_{2k}, \cdots, u_{nk}$,则它们的平均值按下式计算:

$$\bar{u}_k = \frac{1}{n}\sum_{i=1}^{n} u_{ik} \tag{3}$$

在这里,i=1,2,…,n;k=1,2,…,m

然后按下式求出这些原始数据的标准差

$$S_k = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(u_{ik} - \bar{u}_k)^2} \tag{4}$$

这时得到的标准化数据还不一定都在[0,1]闭区间之内,还须采用下面的极值标准化公式:

$$u''_{ik} = \frac{u'_{ik} - \min\limits_{1\le i\le n}\{u'_{ik}\}}{\max\limits_{1\le i\le n}\{u'_{ik}\} - \min\limits_{1\le i\le n}\{u'_{ik}\}} \tag{5}$$

②建立模数相似关系 R

R 可表示为相似矩阵,一般的形式如下:

$$R = \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1n} \\ r_{21} & r_{22} & \cdots & r_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ r_{n1} & r_{n2} & \cdots & r_{nn} \end{bmatrix} \tag{6}$$

在这里,$0 \le r_{ij} \le 1; i = 1,2,\cdots,n; j = 1,2,\cdots,n$

计算 r_{ij} 的方法很多,这里采用最大最小法,即:

$$r_{ij} = \frac{\sum\limits_{k=1}^{m} min(u_{ik},u_{jk})}{\sum\limits_{k=1}^{m} max(u_{ik},u_{jk})}, (i,j = 1,2,\cdots,n) \tag{7}$$

③聚类分析

采用最大树方法进行聚类分析,即构造一个以所有被分类的对象为顶点的图,当 $r_{ij} \neq 0$ 时,顶点 i 与顶点 j 就可以连一条边。其方法是先画出顶点集中的某一个 i,然后按 r_{ij} 从大到小的顺序依次连边,要求不产生回路,直到所有的顶点都被连通为止,这样就得到一棵最大树。树的每一条边都能赋以某一权数,即 r_{ij}。但是由于具体的连法不同,这种最大树不可能是唯一的。

然后对最大数取入截集,即去掉那些权数 $r_{ij} < \lambda$ 的边,$\lambda \in [0,1]$。这样就把一棵树截成互不连通的几棵子树。虽然最大树并不唯一,但取了截集之后,所得的子树是一样的。

这些子树即为数据仓库中归纳发现的模式。

2. 预测

①求各模式平均指标

对所得的每一模式,按下式求其平均指标:

$$M_{ij} = \sum u_{kj}/P \quad (i = 1,2,\cdots,s; j = 1,2,\cdots,m) \tag{8}$$

其中 s 表示总共的模式数,k 表示该模式(即第 i 种模式)由仓库中的哪几条记录推出,P 表示推出该模式的记录总数。

②预测

将待预测的样本 X 与数据仓库中已分类的模式作比较,求出它们之间的贴近度就可以判定该样本接近哪一模式,从这一模式的整体情况预测其结果。

三、模糊数据挖掘算法在高校学生评估管理中的应用

有许多标准可以用于对高校学生的评估。本文选定“心理素质”、“工作能力”、“学习成绩”、“诚信” 和“道德价值”五个指标要素。我们采用程度系数来量化每个指标元素,例如“道德价值”分为较好、一般、差,分别用数值 1,2,3 来代表。从数据仓库中选择 10 个学生数据,见表格 1。

表1 高校学生信息

高校学生	心里素质	工作能力	学习成绩	诚信	道德价值
学生1	4	0.92	3.68	2	3
学生2	5	0.93	3.72	1	2
学生3	4	0.94	3.76	2	2
学生4	3	0.93	3.72	1	1
学生5	6	0.97	3.88	2	3
学生6	3	0.93	3.72	1	2
学生7	2	0.95	3.8	1	1
学生8	5	0.96	3.84	2	2
学生9	1	0.96	3.84	0	1
学生10	3	0.97	3.88	2	2

1、通过计算,得到标准化矩阵和归一化矩阵:

$$\begin{bmatrix} 0.33 & 1.62 & 1.00 & 2.04 \\ 1.17 & 1.00 & 0.67 & 0.19 \\ 0.33 & 0.38 & 1.00 & 0.19 \\ 0.50 & 1.00 & 0.67 & 1.67 \\ 2.00 & 1.50 & 1.00 & 2.04 \\ 0.50 & 1.00 & 0.67 & 0.19 \\ 1.33 & 0.25 & 0.67 & 1.67 \\ 1.17 & 0.88 & 1.00 & 0.19 \\ 2.17 & 0.88 & 2.33 & 1.67 \\ 0.50 & 1.50 & 1.00 & 0.19 \end{bmatrix} \quad \begin{bmatrix} 0.00 & 1.00 & 0.20 & 1.00 \\ 0.45 & 0.55 & 0.00 & 0.00 \\ 0.00 & 0.09 & 0.20 & 0.00 \\ 0.09 & 0.55 & 0.00 & 0.80 \\ 0.91 & 0.91 & 0.20 & 1.00 \\ 0.09 & 0.55 & 0.00 & 0.00 \\ 0.55 & 0.00 & 0.00 & 0.80 \\ 0.45 & 0.46 & 0.20 & 0.00 \\ 1.00 & 0.46 & 1.00 & 0.80 \\ 0.09 & 0.91 & 0.20 & 0.00 \end{bmatrix}$$

(a) 标准化矩阵 (b)归一化矩阵

图2 标准化和归一化矩阵

2、用最大最小法,建立模糊相似矩阵:

$$\begin{bmatrix} 1.00 & & & & & & & & & \\ 0.21 & 1.00 & & & & & & & & \\ 0.13 & 0.08 & 1.00 & & & & & & & \\ 0.59 & 0.36 & 0.20 & 1.00 & & & & & & \\ 0.68 & 0.33 & 0.10 & 0.48 & 1.00 & & & & & \\ 0.24 & 0.64 & 0.11 & 0.44 & 0.21 & 1.00 & & & & \\ 0.29 & 0.24 & 0.00 & 0.47 & 0.45 & 1.00 & & & & \\ 0.25 & 0.76 & 0.26 & 0.28 & 0.37 & 0.46 & 0.22 & 1.00 & & \\ 0.37 & 0.27 & 0.09 & 0.40 & 0.61 & 0.16 & 0.41 & 0.34 & 1.00 & \\ 0.48 & 0.41 & 0.24 & 0.32 & 0.40 & 0.53 & 0.04 & 0.48 & 0.20 & 1.00 \end{bmatrix}$$

图3 相似矩阵

3、采用最大树法进行聚类分析,得到最大树:

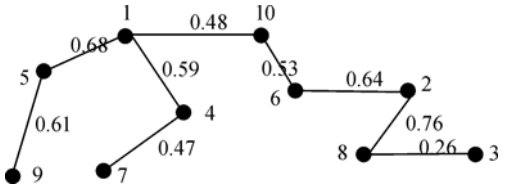


图 4　最大树

设 λ = 0.53,可以得到 4 类:{学生 3},{学生 7},{学生 1,学生 4,学生 5,学生 9},{学生 2,学生 6,学生 8,学生 10}。

最后,根据新的学生的材料,计算相似度,预测该学生与哪一种类型接近,从而更好的评估该学生的表现。

四、结论

数据挖掘能够发现组织内部的知识模式和规律,本文探讨了模糊数据挖掘在高校学生管理中的应用,能够对高校学生进行聚类分析,从而发现高校内部的学生的类型,这对高校学生管理工作具有理论和现实指导意义。作为高校学生工作的教育管理者,应该在科学评估和分析的基础上,根据实际情况,对学生做到客观、公平、有效的管理,只有这样,才能充分调动学生的学习积极性,促进学生健康发展。

参考文献:

[1] R Agrawal, T Imielinksi, A Swami. Mining association rules between sets of items in large database, The 1993 ACM SIGMOD Conf[DB/OL]. (1993). http://acm. lip. tsinghua. edu. cn.

[2] R Agrawal, T Imielinksi, A Swami. Database mining: a performance perspective[J]. IEEE Trans. Knowledge Data Eng, 1993,5(6): 914 - 925.

[3] 邬补科,沈晓敏. 新时期高校学生管理工作的现状及创新[J]. 财经论丛, 2003,(9):99 - 101.

[4] 林勇娟,倪水雄. 大学生管理工作中的博弈思想[J]. 高教论坛, 2004,(1):148 - 150.

[5] 董荣凤, 等. 大学生素质评测专家系统的开发和应用[J]. 中国管理科学,2000, 8(11):435 - 440.

[6] 林勇娟, 罗　艺. 决策理论在高校学生管理中的应用[J]. 高教论坛, 2005, 10(5): 166 - 169.

Application of Fuzzy Data Mining in College Students Management

CAI Xiu-juan, JIANG Yan-rong

(1. *College of Food. South China Agriculture University*, *Guangzhou* 510642;
2. *School of Computer Science and Engineering*, *South China University of Technology*, *Guangzhou* 510640, *China*)

Abstract: In this paper,the problems existing in college student management are analyzed, and the application of fuzzy data mining in college student management is discussed. A method of data mining is designed to make clustering analysis for college students, which can find the knowledge models inside the college students and do a prediction for new students. This research owns a great theoretic and practical guiding potential for the management of college students.

Key Words: fuzzy data mining; clustering analysis; college students management