

数据挖掘在高中生综合素质评价中的应用

刘慧婷¹, 刘军², 朱永斌¹

(1. 安徽大学 计算机科学与技术学院, 安徽 合肥 230601;

2. 阜阳市第十中学, 安徽 阜阳 236000)

摘要:自普通高中生综合素质评价工作开展以来,在高等院校选拔人才时起到了辅助作用。文中引入数据挖掘技术,使其与综合素质评价工作有机地结合起来,符合时代潮流的发展趋势,具有一定的研究价值。文中把改进的基于0-1矩阵向量内积法运用到普通高中生综合素质评价工作中,这种方法与经典Apriori算法相比,由于只需要对事物数据库进行一次扫描,所以效率比经典Apriori算法提高很多。实验结果证明用这种算法来处理学生综合素质评价数据较为合理。

关键词:数据挖掘;综合素质评价;关联规则

中图分类号:TP391

文献标识码:A

文章编号:1673-629X(2014)03-0246-04

doi:10.3969/j.issn.1673-629X.2014.03.061

Application of Data Mining in Comprehensive Quality Evaluation of Senior High School Students

LIU Hui-ting¹, LIU Jun², ZHU Yong-bin¹

(1. School of Computer Science and Technology, Anhui University, Hefei 230601, China;

2. The Tenth Middle School of Fuyang City, Fuyang 236000, China)

Abstract:Comprehensive quality evaluation of senior high school students is useful when advanced education institutions choose talents. It introduces data mining technology to combine with the comprehensive quality evaluation, which conforms to the development trend of times and has certain research values. Apply the improved vector inner product method based on 0-1 matrix to comprehensive quality evaluation of senior high school students. Compared with the classical Apriori algorithm, the improved method is more effective than the former, because it needs to scan the transaction database only once. Experimental results show it is reasonable to apply the improved vector inner product method to deal with the evaluation data of students' comprehensive quality.

Key words: data mining; comprehensive quality evaluation; association rule

0 引言

自安徽省统一高考实行自主命题选拔人才以来,综合素质评价工作与其紧密配合,已经在高校进行鉴别和选择人才、公平客观的反映学生综合素质等各项工作中发挥了不可或缺的作用^[1-2]。经过几年的积累,我校的综合素质评价数据库中存在大量的评价数据。为了使学生的心理、身体等各方面素质得到均衡发展而又不失个性,为了向高校输送更多符合时代发展要求的高素质人才,综合素质评价数据所反映的实质性内容愈发显得重要^[3]。文中运用数据挖掘技术研究学生综合素质评价数据与学生各项身心发展指标两者之间的联系,并期望所分析的结果对学校今后在教育

教学工作中起到一定的积极指导作用。

1 综合素质评价中挖掘工作步骤

利用数据挖掘技术解决实际问题一般包括如下几个步骤^[4]:数据收集、数据预处理、数据转化、进行数据挖掘得出规则及模式、对模式的后评价^[5]。

1) 明确目的,选择、收集挖掘数据。

2) 数据的转化。这是保证挖掘工作所需数据集质量的重要环节,如果这项工作出现较多的纰漏,挖掘出来的结果很可能是垃圾。文中在学业水平测试成绩的转化处理时,使用了离散化的方法^[6-7]。

3) 关联规则的挖掘。对优化后的数据集采用

收稿日期:2013-05-07

修回日期:2013-09-06

网络出版时间:2014-01-07

基金项目:国家自然科学基金资助项目(61202227)

作者简介:刘慧婷(1978-),女,副教授,博士,硕士生导师,研究方向为数据挖掘。

网络出版地址: <http://www.cnki.net/kcms/detail/61.1450.TP.20140107.1525.021.html>

合适的数据挖掘算法对其进行挖掘,获得有价值的关联规则^[8]。文中采用了基于0-1矩阵的向量内积法,对学生综合素质评价活动中出现的各项数据进行关联规则的挖掘。

4)评估与应用挖掘的结果。对所挖掘得到的关联规则的实际价值进行分析说明,将挖掘结果应用到实际人才培养中去。

2 基于0-1矩阵的向量内积算法的基本步骤

传统的Apriori算法以及后来提出的一些改进或扩展的算法实现较复杂,尤其当项目集维数较大时,挖掘效率很低^[9-10],所以,一种将原始数据库中的数据按特定规则转换成0-1矩阵,通过列向量作内积运算来快速生成频繁项目集的算法便出现了。该算法只需扫描一次数据库,预算简单,大量减少I/O次数,降低了算法实现的难度^[11-12]。算法思想及步骤如下所述:

1)布尔矩阵表示。设事务数据库*T*中由*j*个不同项目组成,即 $I = \{I_1, I_2, \dots, I_j\}$ 。*P*为一组事务集。 $T = \{P_1, P_2, \dots, P_i\}$,每个事务*P*都有唯一标识*Tid*,项目集的长度为*k*,称为*k*-项集,则有如下操作:

定义*i* × *j* 矩阵 $R_{i \times j}, R_{i \times j} = \begin{cases} 1, I_j \subset P_i \\ 0, I_j \not\subset P_i \end{cases}$

即完成数据库*T*向0-1矩阵 $P_{i \times j}$ 转化。

2)增加布尔矩阵的第一行标志位,0表示候选频

繁项目集,1代表非频繁项集,初始化该矩阵时,初值设为全0。

3)求出各个项目的支持度,并依据事先设定的最小支持度(min_sup),判断求得频繁1-项集,并将所有非频繁项集的标志位设为1,并将其删除。

4)将第3步求得的频繁1-项集两两作内积运算。即对*D*中向量 $\alpha_m (m = 1, 2, \dots, i - 1)$,依次与其后面的向量 $\alpha_n (n = m + 1, m + 2, \dots, i)$ 两两内积,然后,判断 $\alpha_m \wedge \alpha_n$ 是否大于或等于最小支持度,若满足条件,将其加入频繁2-项集,并且记录第*m*列与第*n*列的列号即(*m*,*n*的值)。

5)将对所有频繁*k*-项集按(4)步方法求向量内积,然后判断所得各项频度是否大于或等于最小支持度,产生频繁*k* + 1项集,并将非频繁项集的标志位改为1,并将其删除。

6)循环以上步骤,直到没有更大频繁项集产生。

7)求频繁项集的各非空子集的可信度,判断*P*是否大于或等于最小可信度,若为真,则生成强关联规则。

3 综合素质评价中数据的清理与转化

3.1 评价数据的清理

从数据库中随机抽取学生综合素质评价数据如表1所示。

表1 部分学生基本信息和评价数据表

学业水平测试号	政治面貌	班干部	性别	语文	数学	外语	科学	人文	技术	实证材料	公民道德素养
092224610920009	团员	否	女	71	75	84	125	110	50		合格
092224610920038	团员	否	女	67	80	63	112	106	43		合格
092224610910133	团员	否	男	57	89	93	93	89	35		合格
092224610910560	团员	否	男	62	79	63	82	80	43		合格
092224610920483	团员	否	女	83	87	68	130	116	45		合格
092224610920612	非团员	否	女	82	67	64	86	83	54		合格
092224610920742	团员	否	女	78	92	93	122	119	51	公民道德素养:2011-2012年获校级“三好学生”	优秀

在综合素质评价的各项指标中选取政治面貌、班干部任职情况、性别、学业水平测试各项成绩、实证材料的记录情况和综合素质评价相应项目成绩6项标准作为数据挖掘源对象,学业水平测试号作为表的主键,用以区分各项记录。

该项挖掘工作的数据预处理主要包括以下几点:

1)合适补充遗漏的数据值。在获得数据和整理数据的同时,发现数据中含有一些属性值的空缺和丢失,这就需要对其进行修护和填补。针对这种情况,可以通过使用另一种方法来获取学生的资料。如:学生性别和班干部任职情况存在丢失的数据情况,可以经过实地调查或询问班主任等方法重新获取数据,将学

生的信息完善起来。

2)数据信息出现异常值。在归纳学生的学业水平测试成绩的过程中,发现了有一些学生的成绩全为0分,经过分析调查,此种情况绝大多数是社会考生在我校报名参加了普通高中学业水平测试之后,由于种种原因,没有参加测试,因此,各学科的成绩均为0分。这种情况产生的数据应当作为异常值处理,删除此极端记录。

3)除此之外,由于疾病、家庭原因或学生自身主观原因造成学生中途辍学,虽有学生信息记录,但是没有参与综合素质评价工作,此记录也应当被列为无效记录,应该予以删除。

3.2 评价数据的转化

数据的清理工作完成后,进入到数据的转换工作。根据文中所挖掘数据的特点,需要将学业水平测试成绩进行离散化,性别、政治面貌、班干部任职等属性值也需要用离散化的数值表示。学业水平测试成绩经过分析,可以将其分为四组,结果如表 2 所示。

从学生基本信息库中抽取数据,从综合素质评价系统中获取评价数据,从学业水平测试成绩库中查找

学生的成绩,然后合成一张表,并进行相应的处理,作为数据挖掘的对象,如表 3 所示。

表 2 学业水平测试成绩分类表

取值范围	成绩类别
528 ~ 660 分	优秀
462 ~ 527 分	一般
396 ~ 461 分	及格
396 以下	不及格

表 3 学生信息合成处理表

学业水平测试号	政治面貌	班干部	性 别	总 分	实证材料	公民道德素养
092224610920009	团员	否	女	515		合格
092224610920038	团员	否	女	471		合格
092224610910133	团员	否	男	456		合格
092224610910560	团员	否	男	409		合格
092224610920483	团员	否	女	529		合格
092224610920612	非团员	否	女	436		合格
092224610920742	团员	否	女	555	公民道德素养:2011-2012 年获校级“三好学生”。	优秀

根据 Apriori 算法的特点,需要把表 3 加以转化,按照一个事务对应一条记录的准则将表 3 转换成事务库,记录中的属性值按特定的代码转换。以“公民道德素养”评价模块为例。

将表 3 中前 5 位学生的基本信息资料和综合素质评价数据按表 4 中的规则转换成事务数据库,如表 5 所示。

根据基于向量积的求解方法,将事务数据库化为 0-1 矩阵,如下:

1	0	0	1	0	1	0	1	0	0	0	1	0	1	0
1	0	0	1	0	1	0	1	0	0	0	1	0	1	0
1	0	0	1	1	0	0	0	1	0	0	1	0	1	0
1	0	0	1	1	0	0	0	1	0	0	1	0	1	0
1	0	0	1	0	1	1	0	0	0	0	1	0	1	0

4 挖掘算法的实现

以公民道德素养一项指标为例,具体步骤如下所述:

- 1) 选择 vs2008 和 sql2005 为开发平台。
连接数据库,导入数据,并创建 dataGridView 控件,指定数据源。
- 2) 将事务数据库中的数据转化为 0-1 矩阵。
(1) 创建一维字符串数组,长度为 15。
(2) 创建二维数组 X ,用于存放转换后的 0-1 矩阵,转换时加标志位,初始化为全“0”,“0”表示频繁项集,“1”表示非频繁项集。
- (3) 创建锯齿数组 T ,存放事务数据库中的各属性值。

表 4 转换代码表

项目名	属性值	转换代码
团员情况	团员	I_1
	非团员	I_2
班干部任职情况	班干部	I_3
	非班干部	I_4
性别	男	I_5
	女	I_6
测试总成绩	优秀	I_7
	一般	I_8
	及格	I_9
	不及格	I_{10}
有无实证材料	有	I_{11}
	无	I_{12}
公民道德素养	优秀	I_{13}
	合格	I_{14}
	不合格	I_{15}

表 5 转换后数据库中部分事务集

TID	项目表
T_1	$I_1, I_4, I_6, I_8, I_{12}, I_{14}$
T_2	$I_1, I_4, I_6, I_8, I_{12}, I_{14}$
T_3	$I_1, I_4, I_5, I_9, I_{12}, I_{14}$
T_4	$I_1, I_4, I_5, I_9, I_{12}, I_{14}$
T_5	$I_1, I_4, I_6, I_7, I_{12}, I_{14}$

- (4) 假定事务数据库中的记录数为 P ,所有列的可能属性值为 q 。
- 3) 求频繁项集。
- 4) 生成规则。

- (1) 根据所得频繁项集 L , 计算其所有非空子集 S 。
- (2) 对每个非空子集计算它的可信度。 $\text{Sup_count}(L)/\text{Sup_count}(S) = P$ 。
- (3) 若 $P \geq \text{min_conf}$, 则输出强关联规则: $S \Rightarrow (L - S)$ 。

5 测试结果

在学校的学生基本资料信息库中和综合素质评价数据库中分别抽取 2010 届、2011 届和 2012 届普通高中高三学生的数据, 经过对数据的整理, 得到共计约 1 600 余条记录, 对政治面貌、是否是班干部、性别、学业水平测试汇总成绩等级、有无实证材料与综合素质评价六个指标(即“公民道德素养”、“交流与合作”、“审美与表现”、“实践与创新”、“学习态度与能力”和“运动与健康”)进行关联规则的挖掘。获取的关联规则如下:

设置最小支持度为 22%, 最小可信度为 60%。

(1) 关联规则 [政治面貌=团员] \wedge [实证材料=无] \rightarrow [公民道德素养=合格]

支持度: 0.82 可信度: 1

(2) 关联规则 [政治面貌=团员] \wedge [是否班干部=否] \rightarrow [公民道德素养=合格]

支持度: 0.85 可信度: 0.91

(3) 关联规则 [政治面貌=团员] \wedge [是否班干部=否] \rightarrow [实证材料=无]

支持度: 0.84 可信度: 0.9

(4) 关联规则 [总分=一般] \wedge [是否班干部=否] \rightarrow [审美与表现=C]

支持度: 0.29 可信度: 0.54

(5) 关联规则 [是否班干部=否] \wedge [实证材料=无] \rightarrow [学习态度与能力=C]

支持度: 0.49 可信度: 0.54

(6) 关联规则 [总分=一般] \wedge [实证材料=无] \rightarrow [是否班干部=否]

支持度: 0.53 可信度: 0.91

上述关联规则的含义如下:

(1) 在学生信息数据库中, 政治面貌为团员且无实证材料的记录为 82%, 这些记录的学生在公民道德素养一项中全被评为合格。

(2) 在学生信息数据库中, 政治面貌为团员且不是班干部的记录为 85%, 在这些记录中有 91% 的学生在公民道德素养一项中成绩为合格。

(3) 在学生信息数据库中, 政治面貌为团员且不是班干部的记录为 84%, 在这些记录中有 90% 的学生无实证材料。

(4) 在学生信息数据库中, 总分为一般且不是班干部的记录为 29%, 在这些记录中有 54% 的学生在审美与表现一项中成绩为 C。

(5) 在学生信息数据库中, 不是班干部且无实证材料的记录为 49%, 在这些记录中有 54% 的学生在学习态度与能力一项中成绩为 C。

(6) 在学生信息数据库中, 总分为一般且无实证材料的记录为 53%, 在这些记录中有 91% 的学生不是班干部。

6 结束语

由以上研究, 可以发现, 学校应该注重学生各方面素质的综合发展, 不仅要稳抓文化课教学, 还要培养学生独立的思考, 探索知识的能力, 以及社会实践活动能力, 对广大一线教师来说就是要对学生因材施教, 不搞“一刀切”, 在保持稳抓文化课教育的同时, 开展学生各方面能力的培养。

参考文献:

- [1] 张志远. 考试命题要符合素质教育的要求[J]. 宁夏教育, 1998(6): 13-14.
- [2] 罗祖兵, 邱月. 高中综合素质评价中的关键表现及其作用[J]. 教育科学研究, 2012(11): 44-48.
- [3] 梁宝华. 基于数据挖掘的大学生综合素质评估系统的设计与实现[D]. 南宁: 广西师范大学, 2007.
- [4] Han Jiawei, Kamber M. 数据挖掘: 概念与技术[M]. 第3版. 范明, 孟小峰, 译. 北京: 机械工业出版社, 2012.
- [5] 孔丽英. 数据挖掘在计算机等级考试中的应用[J]. 计算机教育, 2010(2): 38-41.
- [6] Coenen F. Data mining: Past, present and future[J]. Knowledge engineering review, 2011, 26(1): 25-29.
- [7] Liao Shu-Hsien, Chu Peihui, Hsiao Pei-Yuan. Data mining techniques and applications - A decade review from 2000 to 2011[J]. Expert systems with applications, 2012, 39: 11303-11311.
- [8] Hullermeier E. Fuzzy machine learning and data mining[J]. Wiley interdisciplinary reviews - Data mining and knowledge discovery, 2011(1): 269-283.
- [9] 刘以安, 刘强, 邹晓华, 等. 基于向量内积的关联规则挖掘算法研究[J]. 计算机工程与应用, 2006, 42(21): 172-174.
- [10] 梁宝华, 张步群, 陆军, 等. 基于排序向量内积的关联规则挖掘算法[J]. 计算机工程与应用, 2008, 44(26): 126-128.
- [11] 袁万莲, 郑诚, 翟明清. 一种改进的 Apriori 算法[J]. 计算机技术与发展, 2008, 18(5): 51-53.
- [12] 赵艳芹. 关联规则数据挖掘算法的研究[D]. 哈尔滨: 哈尔滨工程大学, 2006.

数据挖掘在高中生综合素质评价中的应用

作者：

刘慧婷, 刘军, 朱永斌, [LIU Hui-ting](#), [LIU Jun](#), [ZHU Yong-bin](#)

作者单位：

刘慧婷, 朱永斌, [LIU Hui-ting](#), [ZHU Yong-bin](#)(安徽大学 计算机科学与技术学院, 安徽 合肥, 230601), 刘军, [LIU Jun](#)(阜阳市第十中学, 安徽 阜阳, 236000)

刊名：

计算机技术与发展

英文刊名：

Computer Technology and Development

年, 卷(期):

2014(3)

本文链接: http://d.wanfangdata.com.cn/Periodical_wjfz201403060.aspx