基于粗糙集的高校学生实践能力因素研究

徐 怡,汤天贺,张 屹,刘埠远,张添翼
(安徽大学 计算机科学与技术学院,安徽 合肥 230601)

摘 要:学生实践能力是衡量一个高校学生综合能力的重要指标,但是由于影响学生实践能力的因素很多,没有统一的标准去衡量一个高校学生实践能力的好坏。因此,高校师生也无法采取针对性措施提高高校学生的实践能力。为了准确分析影响高校学生实践能力的因素,设计了影响高校学生实践能力的问卷,向大一至大四学生发放,采集数据,然后采用粗糙集理论中基于信息熵的启发式属性约简算法计算各个属性的属性重要度,找出影响高校学生学习实践能力的关键因素。再采用粗糙集理论中基于分辨矩阵的属性值约简算法,挖掘出影响高校学生实践能力的关键因素,导出规则集。通过实验验证了该规则集的有效性。研究成果可以对高校教学工作的开展提供参考,继而提高学生的实践能力。

关键词:高校学生;实践能力;粗糙集理论;属性约简;规则提取
中图分类号:TP301 **文献标识码**:A **文章编号**:1673-629X(2020)04-0031-05
doi:10.3969/j.issn.1673-629X.2020.04.006

Research on Factors Affecting College Students' Practical Ability Based on Rough Set Theory

XU Yi,TANG Tian-he,ZHANG Yi,LIU Bu-yuan,ZHANG Tian-yi
(Department of Computer Science and Technology,Anhui University,Hefei 230601,China)

Abstract:Students' practical ability is an important indicator to measure the comprehensive ability of a university student. However, there are many factors that affect students' practical ability,so there is no unified standard to measure the practical ability of a college student. Therefore,college teachers and students can not take targeted measures to improve the practical ability of college students. In order to accurately analyze the factors affecting the students' practical ability in colleges and universities,we design a questionnaire which affects the practical ability of college students. We distribute the questionnaire to the freshman to senior students for data collection,and then use the information entropy-based heuristic attribute reduction algorithm in rough set theory to calculate the importance of each attribute and find out the key factors that affect the learning ability of college students. Then we adopt the attribute value reduction algorithm based on the discernibility matrix in rough set theory to find out the key factors affecting the students' practical ability and derive the rule set. The validity of the rule set is verified by experiments. The research results can provide reference for the development of teaching work in universities,and then improve students' practical ability.

Key words:college students;practical ability;rough set theory;attribute reduction;rule extraction

0 引 言

目前对于学生实践能力的评估还是通过某些课程的分数或者人为观察,得到的结果很可能与实际情况不符。这对于高校或教师衡量学生综合能力具有很大影响。标准的可以客观准确评估学生实践能力的规则体系对现阶段教学是非常有必要的。文中通过对高校学生的真实数据进行处理,得到其中隐含的规则。具体做法是向高校大一至大四的学生发放调查问卷收集数据,然后利用基于粗糙集的属性约简算法和规则提取算法对数据进行处理,最后通过实验验证导出规则的科学性。

文中运用的数据分析方法是基于粗糙集的属性约简和规则提取[1]。粗糙集是波兰科学家 Pawlak 在 1982 年提出的一种处理模糊和不确定知识的数学工

收稿日期:2019-04-10 **修回日期**:2019-08-13 **网络出版时间**:2019-12-18
基金项目:国家自然科学基金(61402005);安徽省自然科学基金(1308085QF114);安徽省高等学校省级自然科学基金(KJ2013A015);安徽大学计算智能与信号处理教育部重点实验室课题项目资助(2014);国家级大学生创新训练项目(201810357071)
作者简介:徐 怡(1981-),女,博士,副教授,研究方向为智能信息处理和粗糙集理论;汤天贺(1998-),男,研究方向为粗糙集理论。
网络出版地址:http://kns.cnki.net/kcms/detail/61.1450.TP.20191218.1110.002.html

具[2-3]。它能有效地分析不精确、不一致性、不完整等各种不完备信息,还可以通过对数据的分析和推理,从中发现隐含的知识,揭示潜在的规则。目前,粗糙集已经广泛应用于各个领域,例如人工智能领域中的机器学习、知识获取、分析决策等;也可以与其他软计算方法结合,设计出更智能更有效的混合系统[4]。

利用粗糙集理论进行数据处理,提取数据中隐含的规则,最重要的一点就是对于属性约简和规则提取算法的研究。属性约简可以有效降低知识系统的维数,而规则提取则在此基础上从中得出有效的规则[5]。对于数量庞大的数据集来说,数据之间的联系性通常代表特殊的现实意义。但由于数据库的庞大,人工处理几乎不可能,而引入粗糙集方法可以将大量数据精炼成规则形式描述的知识,便于分析[6-7]。

文中的粗糙集在规则提取方法中运用信息熵的概念,用来描述数据之间的相关性[8]。先运用粗糙集对收集的 176 份有效问卷进行属性约简,然后对约简后的信息系统提取规则,最终得到影响学生实践能力的具体因素。

1 粗糙集基本知识

文中涉及的相关概念如下[9-10]:

定义 1:一个知识表达系统(或信息表) S 可以表示成有序四元组 $S=\{U,A,V,F\}$; U 是论域,是全体样本的集合; A 代表属性集合, $A=C\cup D$, C 是条件属性集,反映的是对象的特征, D 是决策属性集,反映的是对象的类别; V 是属性值的集合, V_r 表示属性 r 的取值范围; F 为信息函数,用于确定 U 中每一个对象 x 的属性值; $F:U\times A\rightarrow V$, 即任一 $x_i\in U, r\in A$, 则 $F(x_i,r)=V_r$。

设 U 为一个论域,对于属于条件属性集合 C 中的任一属性集,都可以导出相应的等价划分。

定义 2: U/R 表示 U 上由 R 导出的所有等价类。$[x]_R$ 表示包含元素 $x\in U$ 的 R 等价类,由同一属性集导出的等价类中的对象在属性集上是不可分辨的。例如, P 是 U 上的一个等价类簇,如果 $Q\in P$ 且 $Q\neq\varnothing$,则 Q 的所有等价类的交也是一个等价关系,记作 $\mathrm{IND}(Q)$。

定义 3: P 为等价关系簇, $R\in P$, 如果有 $\mathrm{IND}(P)=\mathrm{IND}(P-\{R\})$,则称 R 是 P 中不必要的;否则 R 为 P 中必要的。如果每一个关系 $R\in P$ 都是必要的,则 P 是独立的,否则 P 为依赖的。

定义 4: U 为论域, Q 和 P 是 U 上的两个等价关系簇,且 P 包含 Q ,若 Q 是独立的,且两者划分的等价关系相同,则 Q 是 P 的一个约简,记作: $\mathrm{RED}(P)$ 。 P 中所有绝对必要关系的集合称为等价关系簇 P 的核,记作: $\mathrm{CORE}(P)$ 。

文中采用的属性约简算法以核作为基础,逐渐扩充必要的属性。为了对决策表中属性重要度做有效度量,引入信息熵的概念。

定义 5[11]:信息 P 的熵 $H(P)$ 定义为:

$$H(P)=-\sum_{i=1}^{n}p(X_i)\log(p(X_i))$$

文中用到的是条件属性对于决策属性的影响,进一步运用信息熵,得出关于决策属性的条件信息熵。

定义 6[12]:决策信息系统 $S=\{U,A=C\cup D,V,F\}$, C 、 D 为 U 上的一个等价关系集合, C 、 D 在 U 上导出的划分分别为:

$$U/\mathrm{IND}(C)=\{X_1,X_2,\cdots,X_n\}$$

$$U/\mathrm{IND}(D)=\{Y_1,Y_2,\cdots,Y_n\}$$

则 D 相对于 C 的条件信息熵 $H(D\mid C)$ 为:

$$H(D\mid C)=-\sum_{i=1}^{|U/C|}p(X_i)\sum_{j=1}^{|U/D|}p(Y_j\mid X_i)\log p(Y_j\mid X_i)$$

其中, $p(Y_j\mid X_i)=\dfrac{|Y_j\cap X_i|}{X_i}$ 。

定义 7:在决策信息系统 $S=\{U,A=C\cup D,V,F\}$ 中,若 $\forall B\subseteq C$, $H(D\mid B)=H(D\mid C)$ 且 B 相对于 D 是独立的,则称 B 是 C 关于 D 的属性约简。

定义 8:设 U 是一个论域, P 是 U 的一个条件属性集合, D 为决策属性,则 $r\in P$ 是核属性的充分必要条件为 $H(D\mid P)<H(D\mid P-\{r\})$ 。

定义 9:设 $S=\{U,A=C\cup D,V,F\}$ 是一个决策信息系统,其中 C 是条件属性集合, D 是决策属性集合,且 $R\subset C$,则对于任意属性 $a\in C-R$ 的重要度 $\mathrm{SGF}(a,R,D)$ 的定义为: $\mathrm{SGF}(a,R,D)=H(D\mid R)-H(D\mid R\cup\{a\})$ 。

当 a 添加进入 C 中,信息熵变化越大, a 关于 D 越重要。

定义 10[13]: $U=\{x_1,x_2,\cdots,x_n\}$, $c\in C_0$, $c(x)$ 是对象 x 在属性 C 上的值, $D(x)$ 是对象 x 在决策 D 上的值,则分辨矩阵记为 $\boldsymbol{M}(S)=[c_{ij}]_{m*n}$,其 i 行 j 列处元素为:

$$\text{if } c(x_i)\neq c(x_j), D(x_i)\neq D(x_j) \textit{ then } c\in C \textit{ else } 0$$

定义 11:支持度,关联规则 $X\Rightarrow Y$ 在 D 中的支持度,是 D 中事务包含 $X\cap Y$ 的百分比,即 $\mathrm{Sup}(R_x)=\dfrac{[x]_C\cap[x]_D}{U}$,衡量决策规则的强度。

定义 12:置信度,是包含 X 的事务中同时包含 Y 的百分比,即 $\mathrm{Cer}(R_x)=\dfrac{[x]_C\cap[x]_D}{[x]_C}$,反映了决策规则的可信性。

定义 13:覆盖度,即 $\mathrm{Cov}(R_x)=\frac{[x]_C \cap [x]_D}{[x]_D}$,用来评估决策规则的质量,反映了决策规则条件类对决策类的覆盖程度。

2　属性约简算法

属性约简对于一个信息系统来说非常重要,它可以减少信息系统的规模。即使用一部分属性和数据就可以达到与之前相同的决策效果。通常信息系统中并不是所有属性都一样重要,而去掉冗余属性的步骤被为称属性约简,文中选用了基于信息熵的属性约简算法[14],即利用信息熵来区分属性的重要程度。算法描述如下:

输入:信息系统 $S=\{U,A=C\cup D,V,F\}$ 即属性集合, C 是条件集合, D 是决策集合;

输出:信息系统的核与最小约简 P。

Step1:决策属性集 D 相对条件属性集合 C 的条件熵 $H(D\mid C)$。

Step2:计算 C 中的核属性集 Core。

Step2.1: Core = $\varnothing$;

Step2.2:对于每个 $a\in C$,IF $H(D\mid C)<H(D\mid C-\{a\})$,则 Core = Core + $\{a\}$;

Step2.3:输出 Core。

Step3:约简。

Step3.1: P = Core, $B=C-$ Core;

Step3.2:计算条件信息熵 $H(D\mid P)$,IF $H(D\mid P)=H(D\mid C)$ 转 Step4,否则继续执行;

Step3.3:对于每个 $r\in B$,计算条件信息熵 $H(D\mid P\cup\{r\})$,求 $\mathrm{SGF}(\{r\})=H(D\mid P)-H(D\mid P\cup\{r\})$;

Step3.4:选择 SGF($\{r\}$) 最大的属性 r, $B=B-\{r\}$, $P=P+\{r\}$,同时把 SGF 为零的属性值去掉;

Step3.5:转 Step3.2。

Step4:输出约简 P。

算法的核心思想就是从核属性集开始,对剩下的属性计算条件信息熵。条件信息熵的值为 0,表示此属性对于信息系统是不必要的。以此作为判断条件,进行多轮计算,直到现有属性集可以替代原来的属性集。

3　规则提取算法

属性约简缩小了数据的规模。但是约简后的数据仍然有冗余。规则提取就是进一步地去掉不必要的信息,删除每个样本中的多余属性值。用少量的数据值就可以区分一条样本。由于处理的数据量较少,所以文中选用的是基于分辨矩阵的规则提取算法,该算法较为简单,也容易实现。中间加入了一些启发式信息,来提高效率。

算法描述如下[15-16]:

输入:信息系统 $S=\{U,A=C\cup D,V,F\}$ 即属性集合, C 是条件集合, D 是决策集合;

输出:规则集 R。

Step1:根据 S 构造分辨矩阵 $\boldsymbol{M}$。

Step2:计算每行的核属性集 C。

Step2.1: $C=\varnothing$;

Step2.2:对于 $\boldsymbol{M}$ 的一行来说,它的核属性集 C 是一行中所有属性个数为一的元素的集合。

Step3:更改 $\boldsymbol{M}$。

Step3.1: i 为 $\boldsymbol{M}$ 的行数, j 为 $\boldsymbol{M}$ 的列数;

Step3.2:IF $M[i][j]\cap C\neq\varnothing$ THEN $M[i][j]=\varnothing$;

Step3.3:对于 $\boldsymbol{M}$ 的一行来说,IF 所有元素的并集为空,转 Step4,否则继续执行;

Step3.4:挑选 $\boldsymbol{M}$ 的一行中出现次数最多的一个属性 r, $C=C+\{r\}$,转 Step3.2。

Step4:输出规则。

Step4.1:对于 $\boldsymbol{M}$ 的第 i 行,此时得到规则:核属性集 C 中的属性对于第 i 行决策表中的描述→决策表第 i 行的决策值;

Step4.2:对 $\boldsymbol{M}$ 的每一行进行处理。

算法首先构建差别矩阵,其中的一行代表此条知识与其他知识的区分情况。对每一行求出所对应的核属性,然后用非核属性填充到核属性集合,直到核属性集合可以唯一区分此条知识。得到的核属性集合及其值构成了规则,下面通过实验将对规则进行分析处理,包括置信度、覆盖度、支持度的计算,以及验证。

4　实验分析

4.1　数据收集

为了得到大量数据研究影响大学生实践能力的关键因素,设计出一份有关大学生实践能力调查的调查问卷,问卷内容分别从个人信息、个人生活与学习习惯以及对实践的态度三个主要方面入手。本问卷共设置 20 个问项,其中总成绩排名、小学期实习成绩、暑期实践活动完成的情况、评价自己实践能力四条,根据不同的选项分别设置分数为 0、1、2、3 分,分数相加所得的结果作为实践能力的决策属性。总分在区间 0-7 分者实践能力弱,在区间 9-12 分者实践能力强。为方便数据处理,所有问项设置为单项选择,每个答案相互独立。具体的调查问卷见表 1。

表 1 问卷表

编号	问题	选项				
1	性别	男	女			
2	年级	大一	大二	大三	大四	
3	专业类别	理学	文学	工学	经管	其他
4	成绩排名	前 10%	前 10% ~30%	前 30% ~60%	后 40%	
5	做事喜欢按照计划还是随遇而安	按照计划	随遇而安			
6	暑期实践活动完成情况	很好	较好	一般	不好	
7	对所学专业知识掌握如何	很好	较好	一般	不好	
8	遇到问题时更倾向于	自己独立思考	寻求他人帮忙	自己手动实践		
9	小学期成绩	前 10%	前 10% ~30%	前 30% ~60%	后 40%	
10	自己在做一件事情的时候结果如何	总是很好	偶尔很好	一般	不太好	
11	是否重视学科迁移	很关注	偶尔关注	不关注		
12	参加社团数量	1 个及以下	2 个	3 个及以上		
13	是否经常参加公益活动或者志愿活动	经常参加	偶尔参加	很少参加	从未参加	
14	是否喜欢实践活动	非常喜欢	比较喜欢	一般喜欢	不喜欢	
15	是否经常主动与人交流	经常	偶尔	极少	从不	
16	课余时间花费在什么地方	网上娱乐	自主学习	兼职	参加各种活动	
17	评价自己的实践能力	很好	较好	一般	不好	
18	更注重学习成绩还是实践能力	学习成绩	实践能力			
19	对于将来的打算	自己创业	继续深造	参加工作	继承家产	
20	评价自己的性格	内向	外向			

问卷的发放对象主要为在校大二到大四学生。总共收到 200 份问卷,去除随意填写以及填写不完整的问卷,真实有效的问卷数量为 176 份。因大一新生参与实践的机会和相关评价指标较少,为使得到的结果更为准确,更加具有普遍性,在所得 176 份有效问卷中,大一新生所占比例较少,仅占比约 11%。

4.2 数据处理

根据第 2 节中所提到的基于信息熵的属性约简算法对收集到的 176 份数据进行属性约简,约简结果为以下 10 条属性:性格,做事风格,重视学科迁移,积极主动和他人交流,课余时间,遇到问题,几个社团,公益或志愿者,学习成绩还是实践能力,自己动手做事。

基于第 3 节中所提到的分辨矩阵方法将属性约简的结果进行进一步处理,得到 10 条覆盖度 0.5 以上,支持度 0.4 以上,置信度 0.8 以上有效规则,如表 2 所示。

表 2 规则表

序号	规则的条件	
1	自己动手做事情=总是很好学习成绩还是实践能力=学习成绩	能力=强
2	自己动手做事情=总是很好课余时间=自主学习	能力=强
3	自己动手做事情=总是很好积极主动和他人交流=偶尔	能力=强
4	自己动手做事情=总是很好	能力=强
5	遇到问题=寻求他人帮忙学习成绩还是实践能力=实践能力	能力=弱
6	课余时间=网上娱乐 学习成绩还是实践能力=实践能力	能力=弱
7	学习成绩还是实践能力=实践能力	能力=弱
8	几个社团=1 个及以下 学习成绩还是实践能力=实践能力	能力=弱
9	自己动手做事情=偶尔很好 几个社团=1 个及以下	能力=弱
10	公益或志愿者=偶尔参加 课余时间=自主学习	能力=强

为验证所得到规则的准确性,从所有的 176 份数据中分别随机抽取 60% ,70% ,80% ,90% 的数据作为训练数据,其他数据作为测试数据进行交叉测试。每组测试 100 次,将测试结果取平均值,所得结果统计如表 3 所示。从表 3 可以看出当抽取的数据比例在 50% ~90% 之间时,分类精度可达到 0.6 以上,并且随着训练集的不断增加,所得到的分类精度也趋于稳定,说明实验所采用的基于信息熵的属性约简算法与规则提取算法得出了可靠性较高的结果。

表 3　测试抽取比例表

抽取数据比例	训练集	测试集	分类精度
60%	100	67	0.65
70%	117	50	0.69
80%	134	33	0.74
90%	150	17	0.73

由以上研究数据可得出影响大学生实践能力的关键因素以及推测大致导致其有相关性的原因:自己动手做事情的积极性、自主学习的积极性与实践能力呈正相关,而遇到问题时倾向请求他人帮忙与实践能力呈负相关。具体表现为学生对实践本身的内心接纳程度越高,学习与接受能力越强,实践能力也越强;另外从结果中得出的一条规则得出,在课余时间进行网上娱乐的学生的实践能力相比在课余时间进行自主学习的学生实践能力更弱,从而印证了上述观点。实验中得出的其中一条结论引人注目:认为实践能力相对于学习成绩更为重要的学生实践能力更弱,而认为学习成绩更重要的学生往往实践能力也很好。这说明实践能力的提高需要以理论知识为基础,实践也即为所学理论知识的验证。二者相辅相成,不可分而论之。而大学社团对于学生的实践能力也存在一定的影响,社团所提供的工作与交流环境将更有助于培养学生的实践能力。

根据以上研究成果所得出的结论以及原因的推测可以帮助高校制定出台相关的规则制度,也可帮助教师抓住如何提升学生实践能力的关键所在,并改进实践授课方法,从而使学生更愿意动手实践,采取正确的实践方法,进而提升学生的实践能力。

5　结束语

为了准确了解影响高校学生实践能力的关键因素,以提高学生的实践能力,通过对文中设计的调查问卷进行分析,利用粗糙集理论的属性约简和规则提取算法,从决策表中提取影响高校学生实践能力的关键因素,导出规则集。通过实验验证了该规则集的有效性。研究成果可以对高校教学工作的开展提供参考,继而提高学生的实践能力。

参考文献:

[1] 王国胤,姚一豫,于　洪. 粗糙集理论与应用研究综述[J]. 计算机学报,2009,32(7):1229-1246.

[2] PAWLAK Z, SKOWRON A. Rough sets: some extensions [J]. Information Sciences,2007,177(1):28-40.

[3] PAWLAK Z,GRZYMALA-BUSSE J W,SLOWINSKI R, et al. Rough sets[J]. Communications of the ACM,1995,38 (11):88-95.

[4] 邬阳阳,汤建国. 大数据背景下粗糙集属性约简研究进展[J]. 计算机工程与应用,2019,55(6):31-38.

[5] 刘　明. 基于粗糙集的属性约简方法研究[D]. 成都:电子科技大学,2016.

[6] PAWLAK Z. Rough set approach to knowledge-based decision support[J]. European Journal of Operational Research, 1997,99(1):48-57.

[7] PAWLAK Z. Rough sets,decision algorithms and Bayes´ theorem[J]. European Journal of Operational Research,2002, 136(1):181-189.

[8] 钱文彬,杨炳儒,徐章艳,等. 基于信息熵的核属性增量式高效更新算法[J]. 模式识别与人工智能,2013,26(1):42-49.

[9] 张文修,吴伟志,梁吉业,等. 粗糙集理论与方法[M]. 北京:科学出版社,2001.

[10] WANG Xiaoyan,YANG Sichun,JI Bin. Research on attribute reduction based on variable precision rough sets model [C]//Proceedings of 2011 IEEE international conference on computer science and automation engineering(CSAE 2011) VOL01. Beijing:IEEE,2011:4.

[11] BAO Zhongkui,YANG Shanlin. Attribute reduction for set-valued ordered fuzzy decision system[C]//2014 sixth international conference on intelligent human-machine systems and cybernetics. Hangzhou:IEEE,2014:96-99.

[12] YANG Xibei,LIANG Shaochen,YU Hualong,et al. Pseudo-label neighborhood rough set: measures and attribute reductions[J]. International Journal of Approximate Reasoning, 2019,105(1):112-129.

[13] 王治和,崔晓慧. 改进的差别矩阵启发式属性约简算法[J]. 计算机工程与设计,2016,37(4):1032-1036.

[14] 何国建,陶宏才. 一种基于粗集理论的属性约简改进算法[J]. 计算机应用,2004,24(11):75-76.

[15] 蔡兴雨,徐　怡,程智炜. 基于粗糙集理论的影响高校学生成绩因素研究[J]. 计算机技术与发展,2016,26(11):200-204.

[16] 姚　晟,徐　凤,赵　鹏,等. 邻域粗糙集模型的规则提取方法研究[J]. 小型微型计算机系统,2018,39(6):1323-1327.