基于粗糙集的影响大学生心理健康的研究

徐 怡[1,2],余 浩[3],刘 刚[3],倪治伟[3]
(1. 安徽大学 计算智能与信号处理教育部重点实验室,安徽 合肥 230039;
2. 安徽大学 计算机科学与技术学院,安徽 合肥 230601;
3. 安徽大学 互联网学院,安徽 合肥 230039)

摘 要:大学生心理健康是影响大学生未来发展的最主要因素之一。由于影响大学生心理健康的因素复杂而且难以预估,导致高校在改善大学生心理健康的过程中存在盲目性。为了及时准确地帮助心理健康存在隐患的学生,首先通过向本校本科生分发调查问卷,然后利用粗糙集理论中基于信息熵的属性约简算法找出影响因子,利用粗糙集理论中基于决策树的规则提取算法提取出具有支持度、置信度高的规则集,最后利用评估规则集准确度的一般方法验证了规则集的有效性。该研究成果可以指导高校制定出具有针对性的改善大学生心理健康的政策,从而及时准确地帮助存在心理健康隐患的大学生。

关键词:粗糙集;信息熵;决策树;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)05-0121-04
doi:10.3969/j.issn.1673-629X.2020.05.023

Research on College Students' Mental Health Based on Rough Set Theory

XU Yi[1,2], YU Hao[3], LIU Gang[3], NI Zhi-wei[3]
(1. Key Laboratory of Intelligent Computing & Signal Processing of Ministry of Education, Anhui University, Hefei 230039, China;
2. School of Computer Science and Technology, Anhui University, Hefei 230601, China;
3. School of Internet, Anhui University, Hefei 230039, China)

Abstract: College students' mental health is one of the main factors affecting their future development. Due to the complex and difficult factors affecting college students' mental health, there is blindness in the process of improving college students' mental health. In order to timely and accurately help students with mental health problems, we firstly distribute questionnaires to the undergraduates, and then find out the affecting factors by using the attribute reduction algorithm based on information entropy in rough set theory and extract the rule set with high degree of support and confidence by using rule extraction algorithm based on decision tree in rough set theory. Finally we adopt the general method of accuracy evaluation rule sets to verify the effectiveness of the rule set. The research results can guide universities to formulate targeted policies to improve the mental health of college students and help to improve their mental health accurately and timely.

Key words: rough set; information entropy; decision tree; attribute reduction; rule extraction

0 引 言

在如今的社会环境中,大学生在入学前承受压力大,课业繁重,部分学生存在或多或少的心理健康问题。但是心理健康的影响因素复杂,比如是否为单亲家庭,性格类型,参加课外活动的情况等等,各因素的重要程度也有差别,同时其内在联系也模糊不清,对高校的策略制定提出了严峻的挑战。因此,利用科学的方法找出影响因素,挖掘出有指导意义的依赖规则就变得十分重要。

波兰学者 Z. Pawlak 在 1982 年提出的粗糙集理论

收稿日期:2019-04-15 **修回日期**:2019-08-16 **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);安徽大学国家级大学生创新创业训练计划(201810357072)
作者简介:徐 怡(1981-),女,博士,副教授,研究方向为智能信息处理、粒计算、三支决策。
网络出版地址:http://kns.cnki.net/kcms/detail/61.1450.TP.20191218.1112.026.html

是一种能够定量分析处理不完整、不一致、不精确性信息与知识和不确定性的数学工具[1]。粗糙集理论与其他理论在处理不确定和不精确的问题的区别是,它不需要提供数据集合以外的任何先验信息处理这个问题,所以问题的不确定性的描述或处理可以更加客观[2]。基于粗糙集理论的应用研究主要集中在属性约简、规则获取等方面,基于粗糙集的理论发展为数据挖掘提供了许多有效的方法[3]。决策集的一种树结构,决策树方法具有速度快、易于转换为简单易懂的分类规则等优点。

对于决策树,数据的准备往往是简单或者是不必要的,而且能够同时处理数据型和常规型属性,在相对短的时间内能够对大型数据源做出可行且效果良好的结果。

文中首先设计了高校心理健康状态调查问卷,面向本校大一到大四的学生分发调查问卷收集数据,经过离散化处理后,利用粗糙集理论中基于信息熵的属性约简算法找出影响大学生心理健康的关键因素,最后利用基于属性重要度的决策树规则提取算法挖掘出影响因素与心理健康程度的依赖关系,得到支持度、置信度高的规则集。

通过实验、评估验证了规则集的有效性。研究成果可以指导高校制定出具有针对性的改善大学生心理健康的政策,从而准确及时地帮助存在心理健康隐患的大学生。

1 粗糙集理论

为引出粗糙集的属性约简算法,下面介绍文中涉及到的粗糙集的基本概念[4-6]:

定义 1:完整的信息系统 S,可以用四元组表示为 $S=\{U,R,V,f\}$,简记为 $S=\{U,R\}$。$U=\{x_1,x_2,\cdots,x_n\}$ 是一个由有限个对象构成的论域,$A=C\cup D$ 表示域属性集合,其中 $C=\{a_1,a_2,\cdots,a_n\}$ 是条件属性,$D=\{d\}$ 为决策属性;V 表示属性值域;f 是信息函数从 $U\times R$ 到 V 的信息函数,即 $f:U\times R\rightarrow V$,用于表示记录 x 在属性 $a\in A$ 上的取值。

定义 2:任取非空属性子集 $B\subseteq R$,如果对 $x_i,x_j\in U,\forall r\in B,f(x_i,r)=f(x_j,r)$ 均成立,则 B 为不可分辨关系,记为 $\mathrm{Ind}(B)$。$\mathrm{Ind}(B)$ 即可把论域 $\cup$ 中分为若干个等价类,等价类的集合记为 $\cup=\mathrm{Ind}(B)$,为基本集。另外如果不存在集合 X 表示成某些基本集的并时,称 X 为 B 粗糙集。

定义 3:任取子集 $X\subseteq U$,则 X 关于知识 R 的上近似和下近似是:

$$R^-(X)=\{x\in U,[x]_R\cap X\neq\varnothing\}$$

$$R^-(X)=\{x\in U,[x]_R\subseteq X\}$$

其中,$[x]_R$ 表示元素 x 的 R 等价类。确定域 $\mathrm{Pos}(X)$ 表示 U 中在 R 下能确定归入集合 X 的元素的集合,否定域表示为 $\mathrm{Neg}(X)$。

定义 4:在数据规约中,利用两个属性集合 $P,R\subseteq Q$ 之间的相互依赖程度可以确定一个属性 a 的重要度。属性 P 对 R 的依赖程度用 $\gamma_R(p)$ 表示。

$$\gamma_R(p)=\frac{\mathrm{card}(\mathrm{pos}R(P))}{\mathrm{card}(U)}$$

$$\mathrm{pos}R(P)=\{x\mid x\in \mathrm{Upr}R(X)\,\&\,X\in\frac{U}{P}\}$$

属性 a 加入 R,对于分类 U/P 的重要程度定义如下:

$$\mathrm{SGF}(a,R,P)=\gamma_R(p)-\gamma_{R-|a}(p)$$

定义 5:属性集合 P 的信息熵 $H(P)$:

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

定义 6:设 U 是一个论域,P 是 U 的一个条件属性集集合,d 为决策属性,$r\in P$ 是核属性的充分必要条件为:

$$H(\{d\}\mid P)<H(\{d\}\mid P-\{r\})$$

下面介绍基于信息熵的粗糙集属性约简算法。

2 基于粗糙集理论的属性约简算法

如果一个集合有无一个属性对于它对决策表的条件信息熵的大小不造成任何改变,表明这个属性就可以被约简[7-9]。

输入:决策表 $\mathrm{DT}=(U,C\cup D)$。

主要步骤:

Step1:令决策属性集合为 D,条件属性集合为 C,计算 D 的信息熵 $H(D)$。

Step2:计算决策表中属性集 C 对决策属性 D 的互信息量 $I(C,D)$。

Step3:求核属性 core。

(1)初始化 core 为空集;

(2) $\forall a\in C$,若有 $f(x,C-a)=f(y,C-a)$,a 就是核属性。

Step4:令 core 为 R,计算 R 对决策属性 D 的互信息量 $I(R,D)$。

Step5:对 $\forall a\in C-R$,计算其对 D 的互信息量最大的属性,$R=R\cup a$。

Step6:计算此属性集 R 对决策属性 D 的互信息量 I。当属性集 R 的 I 和全部 C 的 I 相等时,则结束;否则转向 Step5。

得到的约简集需通过规则提取才能得到有指导意义的规则集。下面介绍基于决策树的规则提取算法。

3　基于决策树的规则提取算法

要理解好决策树,首先说明一些基本概念及其决策树的使用过程 0-0。再以概念为基础介绍基于决策树的规则提取算法,支持度和置信度的概念[10-13]。

3.1　决策树概念

决策树是知识表示的一种形式。决策树具有树结构,树结构由多个节点和分支组成。决策树的第一个节点称为根节点,根节点是应用决策树时的唯一入口点。下面的根节点和内部节点选择一个属性组,换句话说,它们会在内部问一个问题,并将连接节点分支,和树枝将有答案的可能值,称为叶节点和终端节点决定节点用于确定预计值或对于一个给定的类别分类。

3.2　决策树使用过程

决策树的使用是通过决策树变换的分类规则来确定未知类别数据对象的分类。

首先根据所建立的决策树生成 if-then 格式的分类规则,然后在分类规则的前提下对所判断数据对象的属性值进行比较。如果与规则的前提一致,则该规则的分类就是数据对象的类。

3.3　算　法

以建立决策规则树为目的构造决策树算法。

Step1:在约简属性集中选择 AS(attribute significance)大的属性作为节点,如果各个属性的 AS 相等,选择复合程度最小的属性,如果复合值再一致,则选择序号较小的属性。

Step2:依据所选的属性进行分类,然后对每个类重复上述操作,直到所有的类别中的决策属性相等,或属性集合为空,或者属性选择不能再继续分类,从而产生相应的叶子节点。选择属性后,将其从 reduce 属性集中删除,以确保所选属性不会重复用于每个分支。

Step3:读树。每个叶子节点是一类,就是一个规则。

3.4　支持度和置信度

Support(支持度):表示同时包含 A 和 B 的属性占所有属性的比例。如果用 $P(A)$ 表示使用 A 属性的比例,那么 Support = $P(A\&B)$。支持度是该规则在决策表中的所占比例。

Confidence(置信度):表示使用包含 A 的属性中同时包含 B 属性的比例,即同时包含 A 和 B 的属性占包含 A 属性的比例。公式为:Confidence = $P(A\&B)/P(A)$。置信度计算方法是该规则和该规则有关的全部不相容规则比例。举例如表 1 所示,其中 β 表示规则在表的个数,D 为决策属性,假设决策表有 100 个元组。

表 1　举　例

序号	属性 1	属性 2	属性 3	D	β
1	1	0	1	3	1
2	2	2	2	4	4

则有:

(1) CD = 1/(1+4+1) ×100% = 16.7%

SD = 1/100×100% = 1%

(2) CD = 4/(1+4+1) ×100% = 66.7%

SD = 4/100×100% = 25%

4　实验分析

文中选择了通过面向本校大一到大四的学生分发问卷的形式收集数据,通过抽样调查得到的数据经过筛选后具有一定的普遍性和可靠性。

对处理后的数据使用上文所介绍的算法得到规则集,然后通过支持度和置信度的计算验证了规则集的最简性,再通过交叉测试验证分类精度的方法验证了规则集的有效性。

4.1　设计问卷

在进行调查问卷之前,考虑到大学生的隐私问题和后期处理的方便,对调查问卷进行了几次修改。

本次实验面向本校大一到大四的学生随机分发了 300 份调查问卷。剔除掉 29 份无效问卷后得到 271 份有效问卷。

4.2　处理过程及结果

首先建立大学生的心理健康特征决策表。在构建所有大学生的心理健康特征决策表时,将所有学生调查问卷视为论域 U。将大学生心理健康的影响因素构成条件属性集 C,心理健康总体评价作为决策属性集 D,得到决策表[14]。

利用第 2 节描述的基于信息熵的属性约简算法对决策表进行处理,得到约简后的属性集:$\{a_3, a_5, a_{10}, a_{13}\}$,分别表示性格类型、单亲家庭、课外活动、人际关系。

在属性约简的基础上,利用第 3 节描述的规则提取算法对约简后的决策表进行处理,可以得出以下 5 条规则。

(1) If $a_5 = 0$ and $a_{13} = 1$, then $d = 0$

(2) If $a_5 = 0$ and $a_3 = 0$, then $d = 0$

(3) If $a_3 = 0$ and $a_{13} = 0$, then $d = 0$

(4) If $a_3 = 1$ and $a_{13} = 1$, then $d = 1$

(5) If $a_{10} = 1$ and $a_{13} = 1$, then $d = 1$

4.3　规则集有效性评估

为验证所得规则为最简规则,分别计算了 5 条规则的支持度和置信度,如表 2 所示[15]。

表 2 CD 和 SD 计算 %

序号	SD	CD
1	16.73	100
2	26.62	100
3	19.26	100
4	14.58	100
5	20.00	100

由结果可知所有规则的 CD 即置信度均为 1,即可推出不能去掉任一规则,即为最简规则。如果在 5 条规则中加一条,例如:

If $a_{10}=0$ and $a_{13}=1$, then $d=0$;则规则 1 的 CD 会变化为 66.7%,SD 也会下降,同理对规则 3 和规则 4 也会有同样的影响。

验证了最简性后,为了验证所得 5 条规则的有效性,从 279 分数据随机抽取部分数据作为训练数据,另一部分作为测试数据,按照不同比率抽取,进行三组交叉测试,每组 100 次,取 100 次的分类精度平均值作为最终的分类精度,结果如表 3 所示。

表 3 分类精度测试结果

测试集	训练集	次数	分类精度
50%	50%	100	0.631 2
70%	30%	100	0.695 7
90%	10%	100	0.716 4

从结果中可以得到分类精度在 60% 以上,证明了规则集的有效性。同时训练数据与分类精度正相关,进一步证明了算法的可靠性和规则集的有效性。

5 结束语

经过获取数据、处理数据和数据挖掘之后得到的结果中可以认识到影响大学生心理健康的主要因素为是否为单亲家庭、学习情况和人际关系,有少许影响的为性格类型。在本研究中,了解到为单亲家庭的同学更容易有心理问题。同时学习情况较差并且人际关系较差的同学也存在着心理健康风险,性格类型对大学生心理健康的影响存在但并不显著。与人们的认知相同,性格外向学习情况好的同学普遍心理健康状况好,这在收集问卷的过程中也有所体会。为了准确地挖掘出影响大学生心理健康的因素,利用粗糙集的知识构建了一种数据挖掘模型,并通过实验验证了其可靠性,可以协助高校有针对性地帮助可能存在心理健康问题的大学生,对提高大学生整体心理健康具有一定的指导价值。

参考文献:

[1] PAWLAK Z. Rough sets[J]. International Journal of Computer and Information Sciences,1982,11:341-356.

[2] PAWLAK Z. Rough sets - theoretical aspects of reasoning aboutdata[M]. Dordrecht: Kluwer Academic Publishers, 1991:33-34.

[3] SŁOWIŃSKI R. Handbook of applications and advances of the rough set theory[M]. Dordrecht:Kluwer Academic Publishers,1992:49-60.

[4] ZIARKO W. Rough sets,fuzzy sets and knowledge discovery [C]//Proceedings of the International workshop on rough sets and knowledge discovery (RSKD'93). Banff, Alberta, Canada:Springer-Verlag,1993:24-31.

[5] 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.

[6] 匡乐红,徐林荣,刘宝琛,等. 基于粗糙集原理的泥石流危险度区划指标选取方法[J]. 地质力学学报,2006,12(2):236-242.

[7] 王国胤,于 洪,杨大春. 基于条件信息熵的决策表约简[J]. 计算机学报,2002,25(7):759-766.

[8] 范翔宇,王红卫,索中英,等. 基于粗糙集-信息熵的辐射源威胁评估方法[J]. 北京航空航天大学学报,2016,42(8):1755-1761.

[9] 高 爽,冬 雷,高 阳,等. 基于粗糙集理论的中长期风速预测[J]. 中国电机工程学报,2012,32(1):32-37.

[10] 黄解军,潘和平,万幼川. 数据挖掘技术的应用研究[J]. 计算机工程与应用,2003,39(2):45-48.

[11] 朱绍文,胡宏银,王泉德,等. 决策树采掘技术及发展趋势[J]. 计算机工程,2000,26(10):1-3.

[12] 苗夺谦,王 珏. 基于粗糙集的多变量决策树构造方法[J]. 软件学报,1997,8(6):425-431.

[13] 管红波,田大钢. 基于属性重要性的决策树规则提取算法[J]. 系统工程与电子技术,2004,26(3):334-337.

[14] 侯利娟,王国胤,聂 能,等. 粗糙集理论中的离散化问题[J]. 计算机科学,2000,27(12):89-94.

[15] 谢 宏,程浩忠,牛东晓. 基于信息熵的粗糙集连续属性离散化算法[J]. 计算机学报,2005,28(9):1570-1574.