统计分析及关联挖掘在大学生心理健康中的应用①

亓文娟, 黄书城

(武夷学院 数学与计算机学院, 武夷山 354300)

摘 要: 为深入了解影响大学生心理健康的主要因素以及心理症状之间的关系, 以某高校 2011 级的学生心理测试数据为基础, 采用统计分析和关联规则挖掘两种方法, 从性别、学生干部、独生子女、来源地、家庭结构、家庭月收入等方面进行了分析研究, 根据研究结果为高校开展大学生心理健康教育的规划、决策提供依据.

关键词: 统计分析; 关联规则; Clementine; 心理健康

Statistical Analysis and Association Rule Mining of Application in College Students' Mental Health

QI Wen-Juan, HUANG Shu-Cheng

(Mathematics and Computer Science department, Wuyi University, Wuyishan 354300, China)

Abstract: To better understand the relationship between the main factors affecting the mental health of college students as well as psychological symptoms between a university's 2011' students' psychological test data, the research uses statistical analysis and association rule mining two species method. From gender, only-child or not, native place, student cadre or not, family structure, family's monthly income to analysis research. According to the research results will help educators to get a deeper understanding of students' mental health problems and provide a basis for them to make plans and decisions about college studnets' psychological educaiton.

Key words: statistical analysis; association rules; clementine; mental health

大学生心理健康不是单纯的精神障碍, 心理健康能提升个人的能力, 使他们更易实现人生的目标. 目前中国大学生存在的心理问题呈增多趋势, 由于心理问题造成的大学生行为偏差的个案不断增多, 直接影响到学生的健康成长和校园稳定. 加强大学生心理健康教育工作是新形式下全面实施素质教育的重要举措, 是高等学校德育工作的重要组成部分. 目前各高校均成立了大学生心理咨询中心, 每年均对刚入校的学生进行心理测试, 在测试的过程中已积累了大量的数据. 为深入了解影响大学生心理健康的主要因素以及心理症状之间的关系, 本文采用统计分析和关联规则挖掘两种方法来分析了性别、学生干部、独生子女、来源地、家庭结构、家庭月收入和大学生心理症状间的关联关系, 根据挖掘结果可以为大学生心理健康教育的规划、决策提供依据, 进一步探索学生心理障碍的早期预防、干预的新办法, 防患于未然, 塑造大学生健康的人格, 使高校的心理辅导工作更有针对性, 提高心理辅导工作的水平与效率.

1 相关概念和理论

1.1 关联规则概述

关联规则挖掘(Association Rule Mining)是数据挖掘领域成果颇丰而且比较活跃的研究分支, 是用于发现隐藏在大型数据集中令人感兴趣的联系[1].

(1) 设 I=[i1,i2,,i3,… in]项的集合, 数据集 D 是事务的集合, 其中每个事务 T 是项的集合, 使得 T⊆I. 每一个事务有一个表示符, 称作 TID. 事务 T 包含一个项目集 A 当且仅当 A⊆T, 一个关联规则就是形如 A→B

① 基金项目:福建省"大学生创新创业训练计划"项目(201310397022)

收稿时间:2014-02-27;收到修改稿时间:2014-03-31

的逻辑蕴涵式[2], 其中 A⊂I, B⊂I, 并且 A∩B=∅.

(2) 支持度 Support(A→B)=P(A∪B) = Support(A∪B)=S 即项集 A 和项集 B 的并集 A∪B 在所有事务 D 中出现的概率. 支持度的度量反映了关联规则是否具有普遍性.

(3) 置信度 Confidence(A→B) =P(B|A)=Support(A∪B)/Support(A)=C 即在出现了项集 A 的事务 D 中, 项集 B 也同时出现的概率. 置信度的度量反映了关联规则的可靠性.

(4) 提升度 Lift(A→B)= P(B|A)/ P(B)即置信度与期望置信度的比值, 其值大于 1 才是有用的关联规则.

强规则即置信度和支持度均大于给定阈值(最小置信度阈值和最小支持度阈值)的关联规则, 否则称为弱规则. 给定一个事务集 D, 挖掘关联规则问题就是产生强规则的问题.

1.2 经典算法 Aprior 算法

Apriori 算法是布尔关联规则挖掘频繁项集的原创性算法[2]. 该算法利用逐层搜索的迭代方法找出数据库中项集的关系, 以形成规则, 其过程由连接与剪枝组成. Apriori 算法寻找最大项目集的基本思想是: 第一步, 统计所有含一个元素项目集出现的频率, 并找出不小于 minsup 的一维最大项目集. 从第 k 步(k≥2)开始根据第 k-1 步生成的(k-1)维最大项目集产生 k 维侯选项目集, 搜索数据库得到侯选项目集的项集支持度, 与 minsup 比较, 直到没有候选项集为止, 最终找到 k 维最大项目集[3].

Apriori 算法利用 Apriori 性质来提高频繁模式逐层产生的效率, 也减小了搜索的空间. (性质 1: 频繁项集的所有非空子集也都是频繁项目集. 性质 2: 非频繁项目集的所有超集也都是非频繁项目集.)在数据量较小的情况下大大压缩了候选频繁项集的大小, 取得了很好的性能. 但当频繁项集数据量很大的时候, 它有两个方面的开销可能是巨大的: (1)产生大量的候选项集(2)重复扫描事务数据库, 数据要在外存与内存之间转换处理, 开销很大, 同时运行效率也较低, 在频繁项长度变大的情况下, 运算时间显著增加[4]. 为了提高 Apriori 算法的效率, 国内外的许多文献提出了 Apriori 算法的优化, 如划分、事务压缩、散列技术、抽样、动态项集计数等[5].

2 大学生心理健康数据预处理及统计分析

本文的数据来源于某高校 2011 级共 4320 名学生在入校后所做的大学生心理健康量表, 采用教育部《中国大学生心理健康测评系统》课题组编写的症状自评量表 SCL_90 进行测试.

2.1 数据预处理

数据预处理是数据挖掘过程中一个非常重要的环节, 为了提高关联规则挖掘的准确性、有效性和可伸缩性, 在关联规则挖掘之前, 需要对关联规则所用的数据进行数据抽取、数据清洗、数据规范, 然后将数据生成后面挖掘模块所需要的数据格式.

(1) 数据抽取

在测试获取的数据中, 由于学号、姓名、各题答案、测试日期等属性值都是唯一性的, 挖掘这些属性没有任何意义, 同时统计得知其中汉族的学生占 97.2%, 同时年级都是 2011 级, 对挖掘结果不产生影响, 所以将这些属性删除[6].

(2) 数据清洗

数据清洗包括缺失值处理、异常数据处理、噪声数据处理、重复数据检查以及数据的有效性验证等[7]. 《中国大学生心理健康测评系统》对部分属性缺失值已经做了处理, 但学生的独生子女、学生干部、来源地、家庭结构等属性的缺失值未做处理, 由于空缺值较少, 本文采用人工填充的方法, 利用多数属性值填充该空缺.

(3) 数据规范

根据连续性数据离散化, 离散型数据类别化的基本原则[8], 将测评数据根据测评的结果分为轻、中、重、极重四个区间, 但由于挖掘出来的支持度和置信度均不是很高, 本文将各心理症状值分为有、无两个区间; 将连续数据 “家庭月收入”进行离散化, 按 2000 元以下, 2000-5000 元之间, 5000 元以上划分为低、中、高三个区间; 将离散数据 “来源地”进行转化, 例如将边远农村概化为高层概念农村, 经过概化后来源地分为大中城市、小城镇、农村.

2.2 对大学生心理健康数据的统计分析

被测试的 4320 名学生中, 男生 1999 人, 女生 2321 人, 学生干部 913 人, 非学生干部 3407 人, 独生子女 955 人, 非独生子女 3365 人, 农村 2682 人, 小城镇 870 人, 大中城市 768 人, 单亲家庭 224 人, 非单亲家庭 4096 人, 高收入家庭 712 人, 低收入家庭 2118 人, 中

等收入家庭 1490 人. 其中有躯体化症占 17.85%(771 人), 有强迫症占 61.02%(2636 人), 有人际关系敏感症占 42.43%(1833 人), 有抑郁症占 29.26%(1264 人), 有焦虑症占 31.57%(1364 人), 有敌对症占 32.75%(1415 人), 有恐怖症占 23.94%(1034 人), 有偏执症占 39.47%(1705 人), 有精神病占 31.64%(1367 人). 如图 1 所示. 本文以比例高的强迫症和人际关系敏感症两种心理疾病为例, 从性别、独生子女、来源地、学生干部、家庭结构、家庭月收入进行统计分析, 如图 2-7 所示.

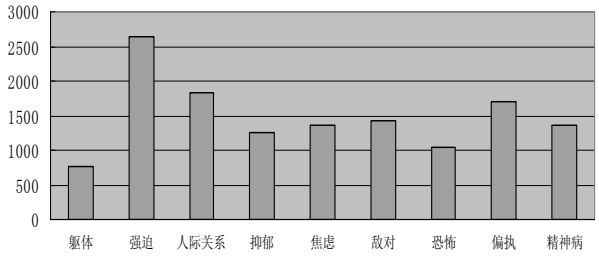

图 1　九维心理症状统计图

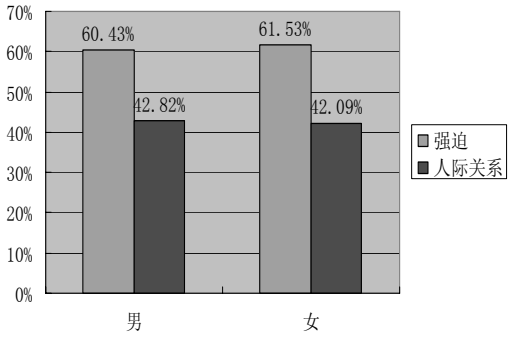

图 2　性别与心理症状的关系图

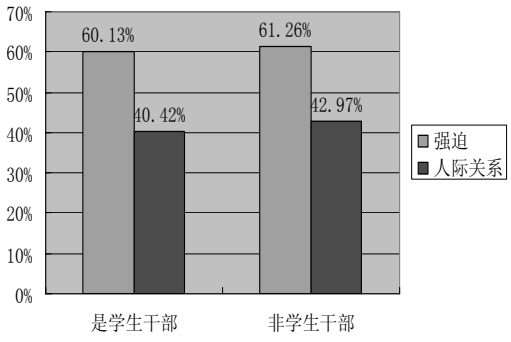

图 3　学生干部与心理症状的关系图

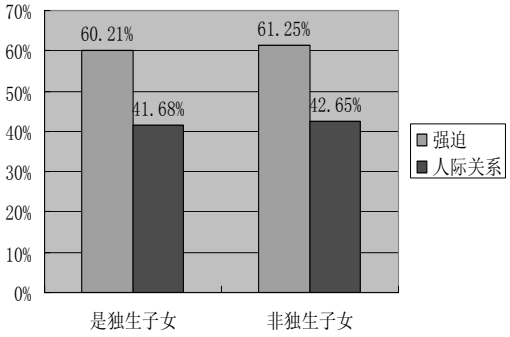

图 4　独生子女与心理症状的关系图

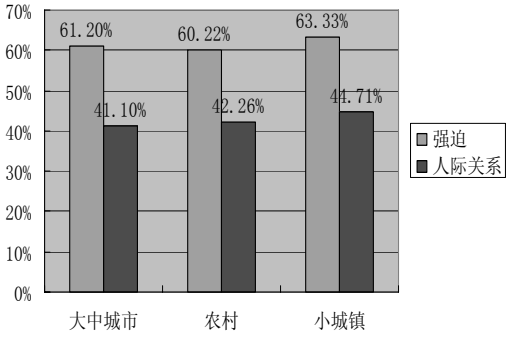

图 5　来源地与心理症状的关系图

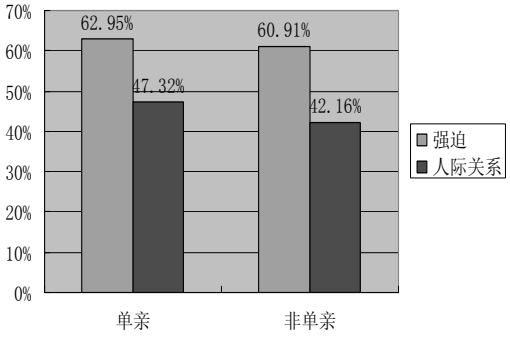

图 6　是否单亲与心理症状的关系图

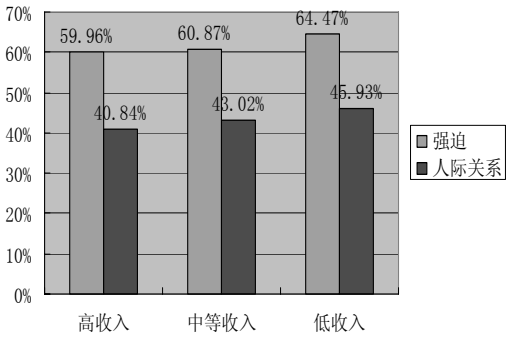

图 7　家庭收入与心理症状的关系图

分析及建议: 女生有强迫症的比例高于男生, 而男生的人际关系敏感的比例高于女生, 非学生干部的学生有强迫症和人际关系敏感的比例都高于学生干部中有强迫症和人际关系敏感的的比例, 非独生子女有强迫症的比例高于独生子女, 来自农村的学生有强迫症的比例低于大中城市和小城镇的学生, 而小城镇的人际关系敏感比例较高, 单亲家庭的学生有强迫症和人际关系敏感的比例高于非单亲家庭的学生, 低收入家庭的学生有强迫症和人际关系敏感的比例高于中、高收入家庭的学生. 作为学生干部的学生, 在日常生活的各种琐事中了提高了协调性和与师生的沟通能力, 有人际关系敏感的比例自然比非学生干部有人际关系敏感的比例低; 独生子女由于缺少同兄弟姐妹交往, 往往更多地在家庭外与同学、朋友交往, 反而在社会交往方面学了许多, 这点否定了以往在社会流行的独生子女“孤僻”、“不合群”、“处处以个人为中心”、“难以与人交往”的看法; 农村的孩子由于受经济压力和就业压力等的影响, 心理健康状况不及城市孩子; 单亲家庭的孩子由于家庭的不完整, 无法同时享受父母的爱, 对待问题的看法上会有些偏激, 心理健康状况不容忽视; 高收入家庭的孩子不用为经济而烦恼, 心理健康状况比低、中等收入的学生要好. 女生、非学生干部、非独生子女、小城镇、单亲、低收入家庭的学生更应该引起相关部门的重视及心理疏导.

3 对大学生心理健康数据的关联规则挖掘

本文采用商用数据挖掘软件 SPSS Clementine12.0 作为挖掘模型建立和分析的平台. Clementine 中关联挖掘有“GRI 模型”、“Carma 模型”、“Apriori 模型”三种[9], 算法可处理 Transactional 和 Tabular 两种数据格式. 选择经典的“Apriori”算法建立模型, 在类型节点中选择方向为“两者”, 在过滤节点中过滤掉与分析无关的属性, 构建关联挖掘数据流[10], 如图 8 所示:

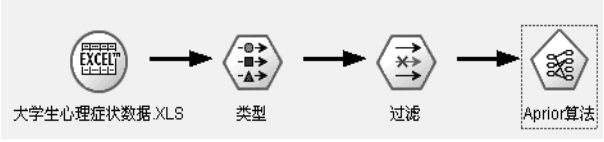

图 8 多维关联挖掘数据流

根据统计的不同属性和不同心理症状情况比例的结果, 作为支持度和置信度阈值的参考依据, 在挖掘时不断调整支持度和置信度阈值, 分别获得性别、学生干部、独生子女、来源地、家庭结构、家庭月收入六个属性和大学生心理症状间的关联关系. 本文以比例高的强迫症和人际关系敏感症两种心理疾病为例进行挖掘分析, 挖掘结果如表 1- 3 所示.

表 1 九维心理症状维间关联规则(部分)

序号	前件	后件	支持度/%	置信度/%	提升
1	有抑郁, 有焦虑	有强迫	22.824	98.479	1.614
2	有恐怖, 有人际关系	有抑郁	20.162	80.712	2.759
3	有焦虑, 有敌对	有强迫	21.852	98.623	1.616
4	有焦虑, 有敌对	有精神病	21.852	80.508	2.544
5	无强迫, 无人际关系	无躯体化	35.972	98.327	1.197
6	无强迫, 无偏执	无人际关系	35.972	94.466	1.641
7	无强迫, 无偏执	无躯体化	35.972	98.391	1.198
8	有抑郁, 有焦虑	有人际关系	22.824	93.002	2.192
9	有抑郁, 有敌对	有焦虑	20.394	85.358	2.703
10	有焦虑, 有精神病	有抑郁	22.986	82.377	2.815

表 2 属性—强迫维间的关联规则(部分)

序号	前件	后件	支持度/%	置信度/%	提升
1	单亲家庭	有	5.185	62.946	1.032
2	大中城市, 低收入	无	8.634	42.359	1.087
3	独生子女, 农村	无	7.616	45.289	1.162
4	学生干部, 女	无	9.537	42.233	1.083
5	非单亲家庭, 高收入	有	15.255	65.706	1.077
6	学生干部, 低收入	无	10.741	42.241	1.084
7	小城镇, 独生子女	有	5.463	67.797	1.111
8	高收入, 女	有	8.194	64.972	1.065
9	独生子女, 女	有	8.462	63.462	1.04
10	学生干部, 独生子女	无	5.208	42.667	1.095

表 3 属性—人际关系维间的关联规则(部分)

序号	前件	后件	支持度/%	置信度/%	提升
1	非独生子女, 高收入	有	12.292	47.269	1.114
2	农村, 高收入	有	10.185	45.227	1.066
3	大中城市, 独生子女	无	9.028	60.769	1.056
4	非学生干部, 高收入	有	13.079	45.664	1.076

5	大中城市, 低收入	无	8.634	64.879	1.127
6	学生干部, 低收入	无	10.741	61.853	1.074
7	高收入	有	16.481	45.927	1.082
8	非单亲家庭, 高收入	有	15.255	46.889	1.105
9	单亲家庭	有	5.185	47.321	1.115
10	非单亲家庭, 大中城市	无	16.852	60.852	1.057

分析及建议: 表 1 列举了各心理症状之间的关联关系, 通过挖掘结果可以看出人际关系敏感与抑郁、焦虑、精神病、恐怖等症状有着较高的相关性, 强迫症与焦虑、抑郁、恐怖、精神病、敌对、偏执、人际关系敏感等症状有着较高的相关性, 而偏执与精神病、抑郁、焦虑有着较高的相关性. 表 2 列举了部分属性与强迫症之间的关联程度, 例如规则 1 表示单亲家庭中有强迫症的学生占所调查学生的比例为 5.185%, 而所有单亲家庭中有强迫症的比例为 62.946%. 表 3 列举了部分属性与人际关系敏感症之间的关联程度. 挖掘结果可以看出独生子女及单亲家庭子女这两种特殊群体有强迫症和人际关系敏感症的学生虽然支持度不高, 但置信度较高, 是绝不能忽视的群体, 他们的存在心理问题的机率大于多子女家庭和非单亲家庭. 由于社会大环境的影响, 学校素质教育的缺乏, 家长对独生子女也是倍加呵护, 养成孩子自私的心理, 在人际交往与沟通中存在着以自我中心、自我封闭、社会功利、猜疑嫉妒、江湖义气等类型. 而单亲家庭子女由于亲子关系的失调, 监护者教养方式的失当, 社会评价压力以及自身心理调试能力不强, 产生不安全感, 自卑感而自闭、孤僻甚至逆反. 担任过学生干部或生活在大中城市的学生社会交际面广, 社会阅历相对比较丰富, 人际关系处理的较好, 而农村孩子受生活环境, 物质条件和见闻等的影响, 心理压力过大. 生活在小城镇的学生有强迫症的比例高于生活在农村的学生, 小城镇的独生子女居多, 而非独生子女多来自农村. 女性有强迫症较男性置信度较高, 而男性有人际关系敏感症较女性置信度较高, 这与重男轻女的世俗观念是分不开的.

针对大学生心理存在的各种问题, 高校要充分做好大学生心理健康教育与咨询的各项工作, 促进学生健康成长. 比如可以通过开展形式多样的校园心理健康宣传教育活动, 利用广播、校报、校园网、班级会议进行宣传, 同时要充分发挥课堂教学在大学生心理健康教育中的主渠道作用, 开设心理学和健康教育系列的校级选修课或邀请心理方面的专家开展心理健康教育专题讲座. 在注重心理健康宣传工作的同时也要心理健康教育软硬件建设, 进一步完善大学生心理健康教育的各项规章制度, 加强心理健康教育工作队伍建设, 组织开展心理教师业务研讨与培训, 成立大学生心理健康协会等. 心理健康教育工作始终要与辅导员、班主任及任课教师的工作相结合, 通过心理测试平台与各位老师的沟通交流, 及时发现易感人群, 同时针对特殊群体给予适当的关怀, 使学校心理健康教育工作更有效, 使学生的心理健康水平得到提高.

4 结语

本文探讨了关联规则经典算法 Apriori 算法, 针对大学生心理健康测评数据, 采用统计分析和关联规则挖掘两种方法, 分析了性别、学生干部、独生子女、来源地、家庭结构、家庭月收入和大学生心理症状间的关联关系, 根据挖掘结果可以更深入了解学生心理问题, 同时针对如何加强和改进大学生心理危机干预工作提出了一点建议, 对于大学生心理咨询教师及辅导员能够更好的开展学生心理疏导工作, 不断推动大学生心理健康教育工作科学化建设有着重要的意义.

参考文献

1 Han JW, Kamber M.范明,孟小峰译.数据挖掘概念与技术.北京:机械工业出版社,2006.
2 王璇.改进的 Aprior 算法在大学生心理数据分析中的应用.中原工学院学报,2011,1:35–38.
3 何广东.数据挖掘在大学生心理问题中的应用.无线互联科技,2013,2:196–197.
4 晏杰,亓文娟.基于 Apriori&FP-growth 算法的研究.计算机系统应用,2013,23(5):122–125.
5 蒋盛益,李霞,郑琪.数据挖掘原理与实践.北京:电子工业出版社,2011.
6 任丽君.大学生心理问题数据挖掘系统的设计与实现.东莞理工学院学报,2008,10:55–60.
7 姜淑芳,常勇.数据挖掘的应用研究.科技经济市场,2011,10:18–21.
8 元昌安.数据挖掘原理与 SPSS Clementine 应用宝典.北京:电子工业出版社,2009.
9 王家胜,牟肖光.读者借阅多维关联规则挖掘模型的建立与分析.计算机应用,2011,11:3084–3086.
10 亓文娟,晏杰,黄书城,郭磊,卢荣辉.关联规则挖掘在大学生心理健康测评系统中的应用研究.湖南工业大学学报,2013,11:94–99.