决策树算法在大学生心理健康测评系统中的应用①

亓文娟, 晏 杰

(武夷学院 数学与计算机学院, 武夷山 354300)

摘 要: 决策树是非常流行的数据挖掘方法. 介绍了决策树的理论, 分析了决策树的构造, 讨论了 C5.0 算法的思想及其优缺点, 同时为深入了解影响大学生心理健康的主要心理症状及因素, 将 C5.0 算法应用于大学生心理健康测评数据, 根据挖掘结果可以更深入的了解学生心理健康问题, 为高校开展好大学生心理健康教育有着现实的意义.

关键词: 数据挖掘; 决策树; C5.0 算法; 心理健康

Applying Decision Tree Algrithm in College Students' Mental Health Assessment System

QI Wen-Juan, YAN Jie

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

Abstract: Decision tree is a very popular method of data mining. This paper introduces the theory of decision tree to analyze the structure of the decision tree, C5.0 algorithm discussed the idea of their advantages and disadvantages. in-depth understanding of the main factors of psychological symptoms and mental health of college students, the C5.0 algorithm is applied mental Health assessment data mining results based on a better understanding of students' mental health problems, to carry out a mental health education colleges and universities have practical significance.

Key words: data mining; decision tree; C5.0 algorithm; mental health

随着生活节奏的不断加快和社会竞争的日益激烈, 大学生面临的学习、生活、情感和就业压力明显增大, 由此产生的心理问题日益突出, 直接影响到学生的健康成长和校园稳定. 大学生心理健康状况引起了全社会的普遍关注, 对大学生心理健康状况的深入研究及对大学生心理干预模式的探索是国内外学者普遍关注的热点. 本研究的目的就是为了弥补对大学生心理问题传统分析统计的不足, 利用数据挖掘技术对大学生心理问题进行研究, 从而发现影响大学生心理健康的主要心理症状及因素, 为学校心理辅导方面提供一个更为科学的决策基础, 为有心理健康问题的学生提供治疗方案, 为大学生的心理健康提供早期预防、干预的新方法, 使学校的心理健康教育工作更具合理性.

1 决策树的概念

决策树是一个类似于流程图的树状结构, 是一种用来表示人们为了作出某一个决策而进行的一系列判断过程的树形图, 这种方法用于表示"在什么条件下会得到什么结果"之类的规则[1].

决策树由决策结点、分支和叶子组成. 决策树最上面的结点称为根节点, 是整棵决策树的开始. 每个分支是一个新的决策结点或者是树的叶子结点. 每一个决策结点代表一个问题或决策, 通常对应于待分类对象的属性. 每个叶子结点代表一种可能的分类结果. 在沿着决策树从上到下遍历的过程中, 每个结点都会遇到一个测试, 对每个结点上问题的不同测试结果导致不同的分支, 最后到达一个叶子结点.

2 决策树的构造

决策树算法发现分类规则主要是利用数据集通过构建一个决策树来完成的. 决策树算法的核心内容是如何构造精度高、规模小的决策树, 一般分为两个步骤[2], 即第一步生成决策树: 由在数据预处理时划分好

① 基金项目:武夷学院校科研基金资助项目(XL201307);福建省"大学生创新训练计划"项目(201310397022)
收稿时间:2015-02-27;收到修改稿时间:2015-04-15

的数据集(即训练样本集)生成决策树的过程; 第二步决策树剪枝技术: 检验、校正和修改生成的决策树的过程, 主要目的是将那些影响预衡准确性的分枝剪除, 在数据预处理划分的测试数据集中的数据这时就发挥了它存在的价值, 对训练集数据构建决策树过程中产生的初步规则进行校验. 一般来说, 决策树规模越小越好, 因为它具备预测能力, 树越小越强, 所以要尽可能的往规模小的树构建.

2.1 决策树的生成

决策树的生成过程由以下几个步骤实现[3]:

(1)对训练样本数据进行进一步的处理, 是根据实际需求以及所处理数据的特性来决定的, 最关键的是选择合适的决策属性, 并且是最能体现样本特殊性的, 并分别确定每个样本决策属性值.

(2)选择决策树的当前决策节点的依据是在决策属性集中最有分类标识能力的属性, 给定“指标”在训练样本集上最佳的属性就是最有分类标识能力的属性.

(3)训练样本数据集根据当前决策节点属性取值的不同划分为若干子集, 属性有几个取值则形成几个子集, 因为每个取值都形成了一个子集.

(4)针对步骤 3 得到的每一个子集, 重复步骤 2 和 3, 直到最后的子集符合子集中所有元组都属于相同类别、该子集是遍历了所有决策属性得到的、子集中的所有剩余测定属性存在取值完全相同, 而分类属性并不相同时, 停止根据这些决策属性进一步进行子集划分三个条件之一.

(5)生成叶子节点.

2.2 决策树的剪枝

在决策树中要克服噪声的基本技术是使用修剪技术, 它的目的是简化决策树而变得容易理解. 由于决策树的生成过程采用自上而下、分而治之的策略, 样本数随着迭代深度的增加也随之减少, 虽然降低了算法的时间复杂度, 但是忽略了样本的整体分布情况, 引起了对噪声的敏感. 出现这种情况是因为在样本划分算法更深层次里, 统计特性集中在训练样本的一个子集. 由于失去了一般代表性而无法用于对新数据的分类和预测, 就出现了过分匹配现象. 删除由于噪声数据而引起的分枝就是决策树剪枝目的, 从而避免决策树的过分匹配.

决策树有两种常用的修剪方法[4]:

(1)预剪枝(Pre-Pruning): 事先指定决策树生长的最大深度是预剪枝最直接的方法, 它使决策树不能充分的增长. 另一种方法是采用检验技术对树节点进行检验, 决定是否允许决策树的相应分枝继续生长, 可事先指定一个最小的允许值.

(2)后剪枝(Post-Pruning): 该技术基于决策树充分生长的条件下, 根据一定的规则, 剪去决策树中冗余的叶节点或毫无意义的分枝. 边修剪边检验是后剪枝的特点, 它的思想是: 在决策树不断的剪枝过程中, 通常采取训练样本集中的数据或测试样本集的数据进行检验, 并计算出目标变量的准确率和错误率. 如果有一个叶子节点剪枝后, 测试集上的准确度或其他测度依然不降低, 则剪掉该叶子节点.

3 决策树算法C5.0

最早的决策树算法是亨特 CLS(Concept Learning System)提出[5], 后来由 J R Quinlan 在 1979 年提出了著名的 ID3 算法, 主要针对离散型属性数据, C4.5 是 ID3 的改进算法, 增加了对连续属性的离散化, C5.0 是 C4.5 应用于大数据集上的分类算法, 主要在执行效率和内存使用方面进行了改进.

3.1 算法优缺点

C5.0 算法可以处理多种数据类型,如: date, times, timestamps 等, 数据处理速度更快内存占用方面的性能大大提高,由于采用了提升(Boosting) 方法, 产生的决策树是较小的, 拥有更高的分类精度. 其优点主要表现为: 在面对数据遗漏和输入字段很多的问题时非常稳健; 通常不需要很长的训练次数进行估计; C5.0 模型比一些其他类型的模型易于理解, 模型推出的规则有非常直观的解释; 提供强大技术以提高分类的精度. 但 C5.0 算法对连续性的字段比较难预测.

3.2 C5.0 算法选择决策树分支的标准

C5.0 决策树算法采用属性的信息增益来确定决策树分支的标准, 寻找最佳分组变量和分割点[6]. 设 S 是一个样本集合, 目标变量 C 有 K 个分类, freq(Ci, S)表示属于 Ci 类的样本数, |S|表示样本几何 S 的样本数. 则集合 S 的信息熵定义为:

$$\text{Info(S)}=-\sum_{i=1}^{k}((freq(C_i,S)/|S|)\times\log_2(freq(C_i,S)/|S|))$$

如果某属性变量 T,有 N 个分类, 则属性变量 T 引入后的条件熵定义为:

$$Info(T)=-\sum_{i=1}^{n}((|T_i|/|T|)\times Info(T_i))$$

属性变量 T 带来的信息增益为:

$Gain(T)= Info(S) -Info(T)$

3.3 C5.0 算法思想[7]

设 R 是非标称属性集;C 是标称属性;S 是训练集;tree()是决策树生成的函数:

函数 tree(R,C,S)//函数返回值类型为决策树

{

/**********相关定义**********

*{dj|j=1,2,… ,m}为属性 D 的值;

*{Sj|j=1,2,…,m}为 S 的子集,分别包含属性 D 的不同值 dj ;

**********/

If(S 为空)then 返回单一失败节点;

If(S 包含的记录的标称属性值均相同)then 返回具有该标称属性值的单一节点;

If(R 为空)then 返回用 S 的最常见值赋值的单一节点;/*此时为出错, 记录没有被适当分类*/

在 R 中找寻具有最大信息增益的属性 D;

生成一棵以 D 为根的树, 分支为 $d_1,d_2,…,d_m$;

递归调用函数 tree(R-{D},C,S_1); tree(R-{D},C,S_2); … ;tree(R-{D},C,Sm);

}

4 C5.0在大学生心理健康测评系统中的应用

4.1 数据预处理

本文的数据来源于福建省某高校共 5065 名 2012 级学生在入校后所做的大学生心理健康量表, 其中男生 2263 人, 女生 2802 人. 旨在通过学生对心理症状 104 个预设问题的回答, 判断大学生的心理状况. 大学生心理健康共有 9 个维度, 分别为躯体化、强迫、人际关系敏感、焦虑、抑郁、敌对、恐怖、偏执和精神病.

数据预处理是数据挖掘过程中一个非常重要的环节. 数据挖掘所处理的数据集通常不仅具有海量数据, 而且可能存在大量的噪声数据、冗余数据或不完整数据等, 比如学生测试时漏填或填写不规范等异常都会导致数据库产生大量的噪声数据, 所以对数据进行预处理是很有必要的, 一般需要用掉挖掘过程中 70%的工作量.

(1) 数据抽取

数据抽取也称为数据取样, 通过它可以使数据的规律性和潜在特性更加明显. 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, 同时被测试的学生都是 2012 级并且 97.8%的学生都是汉族, 对挖掘结果不产生影响, 所以将这些属性删除.

(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. 《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

将测试数据 SCL-90 总分根据中国常模结果, 160 分以下为“健康(A)”、161 分与 200 分之间为“进一步检查(B)”、201 分与 250 分之间为“很明显(C)”、 250 分以上为“比较严重(D)”四个等级作为大学生心理健康与否的标准; 对于“家庭月收入”等属于连续型数据的属性进行离散化, 本文按 2000 元以下为“低”, 2000-5000 元之间为“中”, 5000 元以上为“高”划分为三个区间.

4.2 决策树挖掘模型的建立

比较著名的商用数据挖掘软件主要有 SPSS Clementine、SAS Enterprise Miner、IBM Intelligent Miner、SQL Server Data Mining、Oracle DM 等, 选择 SPSS Clementine12.0 作为本模型建立和分析的平台[8]. Clementine 中有 4 种决策树算法: C5.0、CART、QUEST 和 CHAID 算法. 其中 CART、QUEST 是二分支决策树, C5.0、CHAID 是多分支决策树. CART、CHAID 算法的目标变量可以是连续的, 也可以是离散的. C5.0、QUEST 算法的目标变量只能是离散的. 选择 C5.0 构建决策树模型[9], 如图 1 所示.

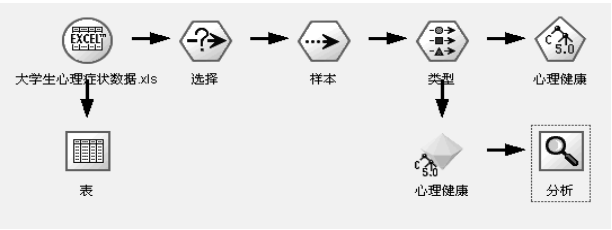

图 1　决策树 C5.0 挖掘数据流

4.3 挖掘结果

通过添加选择节点, 可以对数据源进行条件选择, 包含或丢弃满足某个条件的数据, 通过样本节点对数据源进行抽样, 本文采用随机抽取 70%的数据作为训练样本集, 挖掘分析结果如图 2-6 所示.

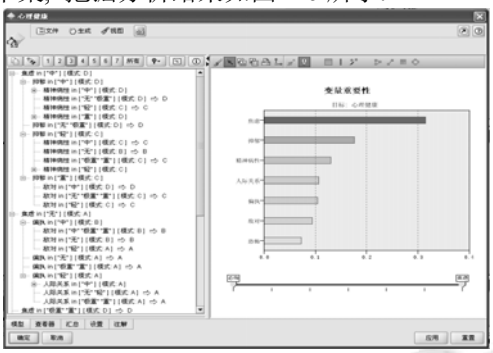

图 2 心理健康症状决策树 C5.0 挖掘结果

图 3 强迫症状与属性间的挖掘结果

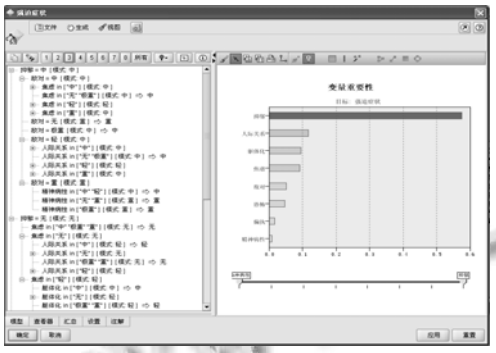

图 4 强迫症状与其他心理症状间的挖掘结果

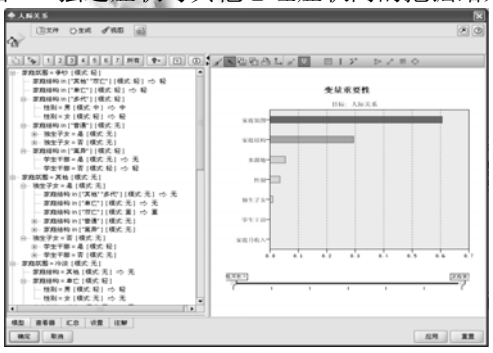

图 5 人际关系敏感症状与属性间的挖掘结果

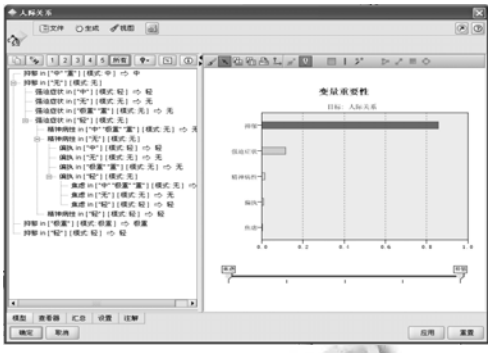

图 6 人际关系敏感症状与其他心理症状间的挖掘结果

4.4 评估及建议

采用 30%的数据作为测试集, 对所生成的分类规则进行验证, 结果如表 1 所示.

表 1 C5.0 训练集和测试集的准确率(%)

	心理健康	强迫	焦虑	人际关系
训练集	97.3	92.82	92.74	91.48
测试集	96.48	92.91	91.68	91.21

从各个角度分析来看, 整体而言, 大学生的心理素质是健康的. 从图 2 中可以看出影响心理健康的最主要因素为焦虑, 焦虑无则心理健康 A; 焦虑重或极重则心理健康 D; 根据统计分析得知强迫症和人际关系敏感症的学生比例较高, 从图 4 和图 6 中可以看出与这两种心理疾病最为密切的心理症状为抑郁症. 由于就业压力、经济压力、感情原因、精神压力过大、不良的校园文化、教育问题以及自身原因等诸多原因很容易导致大学生焦虑症的出现. 由于当今社会就业压力的严峻, 许多大学生急于求成, 力不从心, 并未找到理想的工作; 对于家庭贫困的学生, 经常为经济问题犯愁, 担心生活费、学费等, 虽然有助学金、助学贷款等措施, 但不是每个贫困生都能享受到; 而相当一部分学生由于工作不好找, 读完本科读研究生, 读完研究生读博士, 精神压力过大, 很容易患上焦虑症; 伴随着心理和生理的发育, 大学生在恋爱方面存在着严重的心理障碍, 有些人因为感情破裂, 产生报复心理, 有人因为毕业与恋人分手, 这段感情始终放不下, 对于生活阅历浅的大学生很容易钻死胡同; 也许是由于扩招的原因, 目前很多高校虽然做到了教书育人, 但侧重于教书, 并未做好学生的思想工作; 受社会大环境、学校素质教育、家庭的影响, 养成了大学生缺乏承受挫折的能力, 一旦遇到问题直接撒手, 没有主张, 学习上有目标但动力不足, 有信心但不能持之以

恒; 由于大学生在人际交往与沟通中存在着诸多问题, 出现了较多的宅男宅女寻求在网络的虚拟世界里找到交际的满足, 有的甚至染上网瘾, 迷恋于网络世界, 自我封闭, 不愿与人面对面交往, 久而久之, 影响了大学生的认知、情感和心理定位.

从图 3 和图 5 中, 我们可以看出家庭氛围是影响大学生心理健康的主要因素. 随着当今社会离婚率的提高, 很多父母忽略了家庭教育对孩子的影响, 殊不知父母是孩子最好的老师, 大学生最早接受的教育便是家庭教育, 形成了最初的道德观、价值观. 大学生的性格、心理特点、心理品质和行为习惯的家庭氛围有着直接的关系, 在和谐的家庭中长大的孩子, 身心愉悦, 而家庭中充满了争吵, 冷淡、溺爱、暴力, 孩子的身心必然受创, 导致孩子心理健康问题也越来越多, 越来越严重. 家庭结构也是影响大学生心理健康的第二因素. 健康家庭的孩子对生活充满希望, 对自己的感情生活也非常有信心, 而父母双亡的大学生由于父母关爱的缺失, 缺乏安全感, 神经敏感, 感情脆弱, 做事情总是畏首畏尾, 其心理问题极为显著, 单亲或父母离异的不健全家庭, 总会不同程度、不同层面的对子女的心理健康有所伤害.

对于大学生而言, 在大学期间提高心理素质, 引导他们顺利度过大学时光是非常有意义的. 针对大学生心理存在的各种问题, 高校要充分做好心理教育、预防与咨询的各项工作, 利用网络、广播、校报等媒介进行宣传, 开展形式多样的心理健康宣传教育活动或邀请心理方面的专家开展心理健康教育专题讲座; 进一步加强心理健康教育工作的师资队伍, 通过开展心理教师业务研讨与培训, 提高心理教师的水平; 始终保持心理健康工作要与辅导员工作相结合, 使辅导员能够给予问题学生适当关怀, 通过沟通交流, 消除困惑, 同时学校要与有心理问题的大学生家长进行及时沟通, 使家长充分认识到家庭教育的重要性, 及时对大学生进行教育引导, 使其健康成长.

5 总结

本文分析了决策树的构造, 讨论了 C5.0 算法的思想及其优缺点, 并采用 clementine 进行了决策树 C5.0 挖掘模型的构造, 对大学生心理健康数据进行数据挖掘, 根据挖掘结果分析了影响大学生心理健康的主要心理症状及因素, 并给出了相关建议, 对指导心理健康的相关部门及人员制定正确的辅导计划, 辅助决策有很好的帮助, 为学生的身心健康发展铺路搭桥.

参考文献

1 张晓帆.基于 SQL Server 2005 中药指纹图谱数据挖掘方法研究[硕士学位论文].沈阳:沈阳药科大学,2009.
2 刘文.基于聚类算法和支持向量机算法的文本分类算法研究[硕士学位论文].镇江:江苏科技大学,2012.
3 张婧.基于数据挖掘的汽车售后服务业客户消费行为分析研究[硕士学位论文].武汉理工大学,2009.
4 周怡,王世伟.医学数据挖掘－SQL Server 2005 案例分析.北京:中国铁道出版社,2008.
5 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.
6 高玉蓉.基于决策树的土地利用现状信息提前研究[硕士学位论文].杭州:浙江大学,2006.
7 熊蜀峰.聂黎明.基于 C5.0 算法的学生成绩分析决策树构造.科技信息,2010(8):24–25.
8 亓文娟,晏杰,郭磊,卢荣辉,黄书城.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.
9 吴小刚,周萍,彭文惠.决策树算法在大学生心理健康测评中的应用.计算机应用与软件,2011,(10):240–244.