文章编号: 1006-2475(2013) 03-0082-03

基于数据挖掘的高校学生学习成绩分析应用研究

樊同科,孙姜燕

(西安外事学院现代教育技术中心,陕西 西安 710077)

摘要: 高校的学生成绩管理是各高校教务管理工作的核心和基础。大多数高校的学生成绩是以多种形式保存,一般只限于对成绩的查询及简单的统计上面,没有对这些积累的海量数据背后的有用信息进行挖掘分析。针对这些海量数据构建数据仓库,利用数据挖掘技术的分类预测算法对学生成绩进行挖掘分析,表明学生成绩的高低是与学生本身的特质、生源地、教师学历等多因素有联系的。通过分析得出的这些联系可以为学校的决策和管理部门提供分析和管理的依据。反过来也可以指导和促进教学,以提高整体的教学质量。

关键词: 数据挖掘; 决策树; 学生成绩分析

中图分类号: TP311 **文献标识码:** A **doi**: 10.3969/j.issn.1006-2475.2013.03.021

Analysis and Application of College Students' Academic Record Based on Data Mining

FAN Tong-ke, SUN Jiang-yan

(Modern Education Technology Center of Xi'an International University, Xi'an 710077, China)

Abstract: Management of college students' academic record is an important part of work in educational administration. Most colleges store students' records in various ways, mainly limited to searching and simple statistics, and useful information behind such mass data is not analyzed. A data base is constructed based on such data and students' academic record is analyzed by using classification and prediction algorithm, showing that students' academic performance is related to students' particularity, origin and teachers' education background. Such pertinence can provide evidence for decision-making and administration departments of schools for analysis and administration. In turn, it can instruct education and improve overall educational quality.

Key words: data mining; decision tree; analysis of student's academic record

0 引 言

高等院校要生存,要发展,教育质量是根本,民办院校更是如此。衡量教育质量的一个指标就是学生学习成绩。学习成绩的好坏是受众多因素影响的,它是教师和教学管理部门进行教改和提高教学质量分析与研究的重要依据。由于时间的推移,学校积累了大量的学生成绩数据,这些数据的保存形式可能是多样的,彼此是分散的,不一致的。通过建立数据仓库[1],利用数据挖掘工具,挖掘分析这些数据间潜在的联系信息,对教学工作趋势的预测以及提高高等院校综合数据利用的能力,加强管理决策的合理性和科学性都有推进作用。

1 数据挖掘

数据挖掘是指从海量的、不完整的、模糊的实际应用数据中,提取隐含在其中的人们事先不知道的但又可能有用的信息和知识的过程[2]。通俗地讲,数据挖掘就是利用各种分析方法和工具,对数据库中积累的大量繁杂的历史数据进行分析、归纳与整合的工作,以发现数据内部的信息和关系的过程,提供企业管理层在进行决策时的参考依据[3]。

数据挖掘的模式按功能可分为两大类: 描述性和预测性[4]。描述性模式主要用于数据挖掘的初期,目的是了解某个系统数据存在的问题及特征,是为预测做准备; 预测性模式是在描述分析得到结论的基础上,通过预测性分析能得到大量数据间潜在的联系信

收稿日期: 2012-12-10

基金项目: 陕西省教育科学"十二五"规划 2012 年课题(SGH12534) ; 陕西省 2012 年度自然科学基础研究计划项目(2012JM8045)

作者简介: 樊同科(1979-) ,男,陕西扶风人,西安外事学院现代教育技术中心讲师,硕士,研究方向: 数据仓库,数据挖掘。

息,能够为决策者提供直接的决策依据[5-6]。预测型模式常用的数据模型包括: 决策树模型、规则推理模型和神经网络模型。对于回归型问题,可以选择神经网络; 对于分类型问题,可以用神经网络来解决,也可以用决策树做分类。

2　分类预测

分类和预测是两种数据分析形式,它们可用于抽取能够描述重要数据集合或预测未来数据趋势的模型。分类方法用于预测数据对象的离散类别; 预测方法则用于预测数据对象的连续取值[7]。

2.1　C4.5 决策树分类

决策树采用基于实例的归纳学习算法,旨在从一组无次序、无规则的实例中推理出决策树形式的分类规则,采用自顶向下的递归方式,在决策树的内部节点进行属性值的比较并根据不同属性判断从该节点向下的分支,在决策树的叶结点得到结论。内部节点用矩形表示,而叶节点用椭圆表示。跟踪一条由根到叶节点的路径,该叶节点就存放该元组的类预测[8-9]。

决策树最早起源于 Hunt 等人提出的概念学习系统,然后发展到 Quinlan 的 ID3 算法,最后演化为能处理连续值属性的 C4.5 算法[10]。

2.2　C4.5 决策树分类算法描述

决策树的构造采用自上而下、分而治之的递归方式[11]。初始时,根节点包含数据集中的所有样本。若一个节点包含的样本均为同一类别,则该节点成为叶节点并标记为该类别; 否则,采用信息增益的度量选择合适的分类属性,将数据集划分为若干个子集。该属性称为相应节点的测试属性。对测试属性的每个已知值都创建一个分支,同时也包含一个被划分的子集。递归地对所获得的每个划分形成一棵决策树。一旦一个属性出现在某个节点上,则不能再出现在该节点之后所产生的子树节点上。当一个节点包含的样本均为同一类别或没有样本满足测试属性值,则算法终止[12]。

在进行属性度量(即分类规则)[13-14]时,采用信息增益和信息增益率来衡量,具有最好度量得分的属性被选作给定元组的分裂属性。

(1) 信息增益[13,15]。

设训练样本集 D 有 s 个样本,类别属性有 k 个不同的取值,定义 k 个不同的类 D_j,$j \in \{1,2,3,\cdots,k\}$,则计算元组分类时所需的期望信息公式为:

$$\mathrm{Info}(D) = -\sum_{j=1}^{k} P(D_j)\log_2(P(D_j)) \tag{1}$$

其中 $P(D_j)$ 表示训练集 D 中的类别 D_j 在整个分布中出现的概率,因为采用二进制编码,所以对数函数以 2 为底。

计算类别期望信息度量的公式为:

$$\mathrm{Info}(D)_A = -\sum_{j}^{v} P(V_j) \sum_{i}^{k} P(D_j \mid V_i)\log_2 P(D_j \mid V_i) \tag{2}$$

其中,D 表示训练样本集,A 表示某个属性,属性 A 将 D 划分为 v 个子集,D_j 表示子集 V 中的样本个数。

信息增益的计算公式为:

$$\mathrm{Gain}(A) = \mathrm{Info}(D) - \mathrm{Info}(D)_A \tag{3}$$

(2) 信息增益率[13,15]。

信息增益率的度量偏向于选择具有大量值的属性。分裂信息的计算公式为:

$$\mathrm{SplitInfo}(D)_A = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} \times \log_2\left\{\frac{|D_j|}{|D|}\right\} \tag{4}$$

信息增益率的公式为:

$$\mathrm{gain\_ratio}(A) = \frac{\mathrm{Gain}(A)}{\mathrm{SplitInfo}(A)} \tag{5}$$

3　应　用

3.1　数据准备

在学校的教务管理系统中保存了学校历年学生所有课程的成绩数据,希望从这些数据中挖掘出与提高学生学习成绩有关系的因素。从建立的数据仓库中,对我校全校的公共课《计算机应用基础》抽取如表 1 所示字段的数据模型,包含记录 20 条,作为决策树算法的训练样本数据。其中评价结果分为“优”、“良”、“中”、“差”4 个级别,“优”级 = [80,100],“良”级 = [70,79],“中”级 = [60,69],“差”级 = [0,59]。在抽取的样本数据中,男生共 11 人,其中“优”级 4 人,“良”级 2 人,“中”级 3 人,“差”级 2 人; 女生 9 人,其中“优”级 4 人,“良”级 2 人,“中”级 1 人,“差”级 2 人。样本数据中成绩字段为连续属性,经过分类转化为离散属性为: S1 = [85,100],S2 = [70,85],S3 = [60,70],S4 = [0,60]。将成绩属性作为归纳属性集,将性别、院系、生源地、教师学历、教师职称联合作为决策属性来划分样本类别。

表 1　训练集字段

学生姓名	性别	院系	教师学历	教师职称	生源地	成绩	评价结果

3.2　构造决策树

在抽取的 20 条元组中,有“优”级元组 8 条,“良”级元组 4 条,“中”级元组 4 条,“差”级元组 4 条。通过计算每个属性的信息增益,可以找出训练集样本 D 的分裂规则,如样本中学生成绩评价结果的期望信息 Info(D) = 1.9217(位),对 D 中元组按照性

别进行分类所需的期望信息为 $\mathrm{Info}(D)_{性别}=1.8913$（位）。因此，按照性别属性划分的信息增益为 Gain（性别）$=\mathrm{Info}(D)-\mathrm{Info}(D)_{性别}=0.0304$（位），对于性别属性的分类信息为：$\mathrm{SplitInfo}(D)_{性别}=0.9927$，性别属性的增益率为：Gain_ratio（性别）$=0.0307$。

按照以上的计算方法分别计算教师职称、院系、教师学历、生源、成绩的增益率分别为 0.2244、0.2271、0.1152、0.3083、0.7945。由此可以看出，增益率最大的为成绩属性，则其可作为决策树的根节点，可把决策树分成 4 棵子树。每棵子树的属性字段均为序号、姓名、性别、院系、教师学历、教师职称、生源、评价结果。若同一棵子树的评价结果完全相同，则其归为一类，作为该决策树的叶子节点。若同一棵子树的评价结果不同，则继续计算各属性的增益率，再划分类别。

对结果进行整理与去除数据后得到整棵的决策树如图 1 所示。

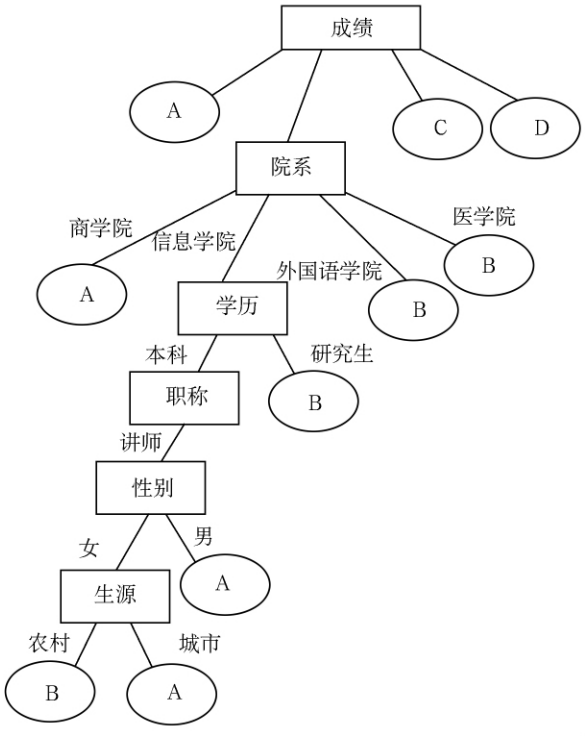

图 1 计算机基础课程学生成绩完整的决策树

3.3 结论

从以上构造的某校计算机应用基础课程学生成绩的分类预测模型中，可以分析得知影响学生该成绩的最重要的因素是学生所在的院系，这实际上是与学生的入学成绩和中学文理科差别有关，信息学院学生在高中时大部分为理科，且入学成绩较高。其次是教师学历、教师职称、学生性别和生源地，这符合该校的现实状况。从对该树的分析可知，该校要继续加强对学生的教育教学管理，提高教学质量，招收更多的高素质的学生入校，不断提升教学层次和办学水平。同时要加强教师队伍建设，鼓励教师提升学历，丰富教学经验，提高教学水平等。

4 结束语

数据挖掘作为一种新兴的数据分析技术，其研究取得了令人瞩目的成就，已经成功地应用到了许多领域。数据挖掘的方法有很多，本文只用了决策树分类预测方法对学生成绩进行分析研究。在实际应用时，可以使用其它方法或多种方法结合起来分析研究，以发现其中隐藏的有用的信息。

参考文献：

[1] Inmon W H. 数据仓库 [M]. 北京：机械工业出版社，2003.
[2] 陈文伟. 数据仓库与数据挖掘教程 [M]. 北京：清华大学出版社，2006.
[3] 张云涛，龚玲. 数据挖掘原理与技术 [M]. 北京：电子工业出版社，2004.
[4] 王海著，闫宏印. 数据仓库和数据挖掘在信用社客户关系管理中的应用 [J]. 计算机与现代化，2010(1)：162-165.
[5] 王远庆. 基于数据挖掘技术的个人信用评估模型研究 [D]. 成都：西南财经大学，2009.
[6] 施蕾，唐艳琴，张欣星. 数据挖掘中决策树方法的研究 [J]. 计算机与现代化，2009(10)：29-32.
[7] 朱明. 数据挖掘导论 [M]. 合肥：中国科学技术大学出版社，2012.
[8] 郑岩. 数据仓库与数据挖掘原理及应用 [M]. 北京：清华大学出版社，2011.
[9] 林向阳. 数据挖掘中的决策树算法比较研究 [J]. 中国科技信息，2010(2)：94-95.
[10] 李强. 创建决策树算法的比较研究——ID3，C4.5，C5.0 算法的比较 [J]. 甘肃科学学报，2006，18(4)：84-87.
[11] 朱娟，杨丰华. 改进的决策树算法在教务管理数据挖掘系统中的应用 [J]. 软件导刊，2010，9(4)：78-79.
[12] 韩慧，王建新，孙俏，等. 数据仓库与数据挖掘 [M]. 北京：清华大学出版社，2009.
[13] Jiawei Han，Micheline Kamber. 数据挖掘概念与技术 [M]. 范明，孟小峰译. 北京：机械工业出版社，2007.
[14] 李慧慧，万武族. 决策树分类算法 C4.5 中连续属性过程处理的改进 [J]. 计算机与现代化，2010(8)：8-10.
[15] 陈文伟. 数据仓库与数据挖掘教程 [M]. 北京：清华大学出版社，2006.