基于决策树算法的学生成绩的预测分析

刘 志 妩
(沈阳理工大学信息科学与工程学院 辽宁 沈阳 110159)

摘 要 针对学生成绩问题,用决策树算法建立分类规则和分析预测模型,描述如何根据分析预测结果发现影响学生成绩的因素,进而有益于教师改进教学方法,提高学生的学习效果。通过实例验证,建立决策树,得出分类规则,结果表明该算法能够对学生数据进行正确分类,得到有价值的结论。

关键词 数据挖掘 决策树 学生成绩 预测和分析

中图分类号 TP391 **文献标识码** A

DOI:10.3969/j.issn.1000-386x.2012.11.081

FORECAST AND ANALYSIS OF STUDENTS' MARKS BASED ON DECISION TREE ALGORITHM

Liu Zhiwu
(*School of Information Science and Engineering , Shenyang Ligong University, Shenyang 110159, Liaoning, China*)

Abstract Decision tree algorithm is used to establish classification rule and analysis-forecasting model for students' marks. The paper addresses how the analysis-forecasting result is used to find out the factors affecting students' marks, and this in turn benefits teachers in improving their teaching methodology and students in raising their study results. Through example validation, the decision tree is established, the classification rules are derived as well. Result indicates that this algorithm can correctly classify data of students' marks and obtain valuable conclusion.

Keywords Data mining Decision tree Students' marks Forecast and analysis

0 引 言

高等教育的重点是提高教育质量,为社会培养具有综合素质的复合型人才。而提高学生成绩是衡量教学质量的主要依据之一,也是评价学生对知识的掌握程度的重要标志之一,因此,通过对学生成绩进行预测分析,可以为教学管理者深化教学改革,合理安排教学计划,提高教学质量提供重要依据。

数据库系统虽然可以高效地实现数据的录入、查询和统计的功能,但却无法发现海量数据中隐藏的关系和规则。而采用数据挖掘技术,可以从海量数据中发现隐藏的知识和规律。

各学校多年来都积累了大量的学生成绩数据,将数据挖掘技术应用于成绩预测分析,可以对其进行全面分析,得到潜在的影响学生成绩的因素,使教学管理者可以得到许多有价值的信息和知识,并利用其提高教学质量和教学管理水平。因此,利用现代化技术预测分析学生成绩是目前教育界的高度重视的问题之一。

本文主要介绍用数据挖掘中的决策树 C4.5 算法,建立学生成绩的预测分析模型及分类规则,并用实例进行验证。

1 决策树 C4.5 算法

1.1 决策树方法

决策树方法是利用信息论中的互信息(信息增益)寻找数据库中具有最大信息量的属性字段,建立决策树的一个结点,再根据该属性字段的不同取值建立树的分支;每个分支子集重复建立树的下层结点和分支的过程。采用决策树,可以将数据规则可视化,也不需要长时间的构造过程,实际应用中的决策树可能很复杂,但每一条从根结点到叶子结点的路径的含义仍然是可以理解的。决策树的这种易于理解性,对于数据挖掘的使用者来说是一个显著的优点,因此决策树方法在知识发现系统中应用较广泛。

决策树是通过一系列规则对数据进行分类。该方法的思路就是从训练集数据中,自动地构造决策树,从而可以根据这个决策树对任意实例进行判定。决策树可分为分类树和回归树两种,分类树对离散变量作决策树,而回归树是对连续变量作决策树。决策树算法的核心是确定分支准则,即如何从众多的属性中选择一个最佳的分支属性。

最早的决策树算法是由 Hunt 等人于 1966 年提出的概念学习系统 CLS,之后是 Quinlan 于 1986 年提出的 ID3 算法和 1993 年提出的能处理连续属性的 C4.5 算法。ID3 只能处理离散型描述属性,C4.5 算法是 ID3 的改进算法,不仅可以处理离散型描述属性,还能处理连续型描述属性。

本文根据需求确定分析目标,采用 C4.5 算法建立决策树

收稿日期:2012 - 03 - 29。**刘志妩**,副教授,主研领域:数据仓库,OLAP,数据挖掘技术。

分析模型和分类规则,很好地对学生成绩进行预测分析。

1.2　决策树 C4.5 算法

决策树 C4.5 算法用信息增益作为选择根结点和各内部结点中分支属性的评价标准,克服了 ID3 算法使用信息增益选择属性时偏向于取值较多的属性之不足。其处理数据的过程如下。

(1) 元数据预处理

通过 ETL 将所有的元数据转换成数据仓库,如果元数据是连续型,则应离散化处理。

(2) 计算每个属性的信息增益和信息增益率

计算过程如下:

(a) 计算每个训练集分类信息的期望值

设训练数据集为 T,在 T 中类别标识属性有 m 个独立的取值,即定义了 m 个分类 $C_i, i=1,2,\cdots,m$,R_i 为数据集 T 中属于 C_i 类的子集,r_i 是 R_i 中元组的数量,则 T 在分类中的期望信息量可由式(1)计算。

$$I(r_1,r_2,\cdots,r_m) = -\sum_{i=1}^{m} P_i \times \log_2 P_i \tag{1}$$

式中,$P_i = \frac{r_i}{|T|}$,P_i 表示任意样本属于 C_i 类的概率,$|T|$ 表示训练样本数据集中的元组数。

期望信息量 I 用来衡量将 T 分为 C_i 类的不确定性。数值越大意味着不确定性越大,反之亦然。

(b) 计算属性 A 的信息熵

假设属性 A 具有 n 个不同的取值$\{a_1,a_2,\cdots,a_n\}$,则通过属性 A 的取值将数据集 T 划分为 n 个子集,其中 T_j 表示在数据集 T 中属性 A 的取值为 $a_j(j=1,2,\cdots,n)$ 的子集,如果 A 被选为决策属性,则这些子集将对应该结点的不同分支。

用 T_{ij} 表示 T_j 子集中属于 $C_i(i=1,2,\cdots,m)$ 类的元组数,则属性 A 对于分类 C_i 的熵可由式(2)计算:

$$E(A) = \sum_{j=1}^{n} \left(\frac{T_{1j}+T_{2j}+\cdots+T_{mj}}{|T|} I(T_{1j},T_{2j},\cdots,T_{mj}) \right) \tag{2}$$

属性 A 的每个取值对分类 C_i 的期望信息量计算如下:

$$I(T_{1j},T_{2j},\cdots,T_{mj}) = -\sum_{i=1}^{m} P_{ij}\log_2(P_{ij}) \tag{3}$$

其中,$P_{ij} = \frac{T_{ij}}{T_j}$ 表示 T_{ij} 在 T_j 中的比重。

(c) 计算属性 A 的信息增益

属性 A 为分类提供的信息量就是属性 A 的信息增益,由式(4)计算:

$$G(A) = I(r_1,r_2,\cdots,r_m) - E(A) \tag{4}$$

(d) 计算信息增益率

信息增益率定义如下:

$$GR(A) = \frac{G(A)}{I(A)} \tag{5}$$

必须用式(5)对每个属性$(A,B,C,\cdots)$计算增益率。

(3) 构造决策树

信息增益率是选择决策树分裂属性的基础,拥有最大增益率的属性将被选择作为决策树的分支属性。我们将要构建决策树的训练集 T,按照计算的增益率划分成 n 个子集。如果第 i 个子集 T_i 中所有的元组类别相同,该节点将成为决策树的叶结点,并停止分裂。训练集 T 中不符合上述条件的其他子集将继续递归分割构造树的分支,直到所有的子集中的元组属于同一类别。生成决策树后,我们可以从树中提取规则,用于对新的数据集进行分类。

2　实例分析

2.1　学生成绩的元数据

我们以学校的工业电气自动化专业的学生的一些课程成绩数据为例,通过数据挖掘分析,找到各科成绩的内在联系,从而有的放矢,提高学生的整体学习质量。学生成绩数据库包含学生序号(ONS),和某些主要课程的分数。例如:电工基础(记为 FEE)、电机与拖动(记为 EMD)、自动控制原理(记为 ACP),自动控制系统(记为 ACS)和高等数学(记为 HM),部分数据列于表 1 中。

表 1　学生成绩表

ONS	HM	EMD	ACP	ACS	FEE
1	80	75	65	70	68
2	74	67	62	74	65
3	78	81	54	63	62
4	65	68	58	65	68
5	88	82	78	80	82
6	91	85	80	83	85
7	85	82	78	84	78
…	…	…	…	…	…
210	78	72	56	62	64

2.2　数据预处理

为了便于进行数据挖掘,对表 1 中的数据进行规范化,将小于 60 分的成绩用 0 表示,大于等于 60 分的成绩用 1 表示,结果将表 1 转换为一个数据只有 0 和 1 的表。

从所有学生数据中抽样作为数据训练集,共有 210 条记录。其中各科及格人数和不及格人数统计如表 2 所示。

表 2　各门课成绩统计(人数)

课程	HM	EMD	ACP	ACS	FEE
及格(1)	162	158	148	137	125
不及格(0)	48	52	62	73	85

2.3　用 C4.5 算法构造决策树

表 2 显示了样本训练集中,含有基于课程的五个分类,在每个类别中,根据成绩的及格与否将学生人数分为两个子集。

课程 ACS 被选为类别标识属性,其余课程作为决策属性集。构造决策树的目的是发现 ACS 课程与其它课程的内在联系。

训练数据集中包含 210 个元组,其中 ACS 类所对应的子集中的元组数为:及格人数 $r_1=137$, 不及格人数 $r_2=73$。

为了计算每个决策属性的信息增益,首先要计算课程 ACS(标识属性)的期望信息量如下:

$$I(r_1,r_2) = I(137,73) = -\frac{137}{210}\log_2\frac{137}{210} - \frac{73}{210}\log_2\frac{73}{210} = 0.932$$

进一步统计,其他作为决策属性的任一课程与标识属性课程 ACS 的成绩搭配情况,例如课程 HM 成绩及格(为 1)且课程 ACS 成绩也及格(为 1)的人数为 110 人,HM 成绩及格(为 1)且 ACS 成绩不及格(为 0)的人数为 52 人,HM 成绩不及格(为 0)

且 ACS 成绩及格(为 1)的人数为 27 人,HM 成绩不及格(为 0)且 ACS 成绩也不及格(为 0)的人数为 21 人。其它课程成绩与 ACS 课程的成绩搭配情况,列于表 3 中。而其它任两门课程成绩与标识属性课程 ACS 的成绩搭配情况列于表 4 中。其中只列出 EMD 和 FEE 两门课成绩与 ACS 成绩的搭配情况。

表 3 两门课程的成绩搭配情况

搭配 1	HM	ACS	学生人数
	1	1	110
	1	0	52
	0	1	27
	0	0	21
搭配 2	EMD	ACS	学生人数
	1	1	118
	1	0	40
	0	1	19
	0	0	33
搭配 3	ACP	ACS	学生人数
	1	1	105
	1	0	43
	0	1	32
	0	0	30
搭配 4	FEE	ACS	学生人数
	1	1	108
	1	0	17
	0	1	25
	0	0	60

表 4 三门课程的成绩搭配情况

EMD	FEE	ACS	学生人数
1	1	1	53
1	1	0	41
1	0	1	52
1	0	0	12
0	1	1	29
0	1	0	2
0	0	1	3
0	0	0	18

然后参照表 2 的搭配 1,按式(2)计算决策属性课程 HM 的期望信息量(即墒)如下。

$$E(HM) = \frac{162}{210} \times I(110,52) + \frac{48}{210} \times I(27,21)$$

其中:

$$I(110,52) = -\frac{110}{162} \times \log_2 \frac{110}{162} - \frac{52}{162}\log_2 \frac{52}{162} = 0.9053$$

$$I(27,21) = -\frac{27}{48} \times \log_2 \frac{27}{48} - \frac{21}{48} \times \log_2 \frac{21}{48} = 0.9888$$

$$\therefore\ E(HM) = 0.7714 \times 0.9053 + 0.2268 \times 0.9888 = 0.9244$$

按式(4),得决策属性课程 HM 的信息增益为:

$$G(HM) = I(r_1, r_2) - E(HM) = 0.932 - 0.9244 = 0.0076$$

按式(5),可得决策属性课程 HM 的信息增益率为:

$$GR(HM) = G(HM)/E(HM) = 0.0076/0.9244 = 0.0082$$

用同样的方法,可以对其它决策属性进行信息增益和信息增益率的计算。计算结果列于表 5 中。

表 5 各门课程的信息增益和信息增益率

	HM	EMD	ACP	FEE
增益	0.0076	0.0711	0.0241	0.2369
增益率	0.0082	0.0826	0.0265	0.341

由表 5 结果可知,决策属性 FEE(电工基础课)的信息增益率最大,因此将该属性选作决策树的根结点,并且因为 FEE 属性只有两种取值:0(不及格)和 1(及格),所以,从该结点可以分裂出两个分支:一支为不及格(记为:分支 0),另一支为及格的(记为:分支 1)。由表 3 搭配 4 的数据,可见,FEE 和 ACS 都及格的人数为 108 人,它占 FEE 及格人数(125 人,参见表 2)的比例为:

$$108/125 = 0.864$$

它表示分支 1 的估计准确率为 86.4%,满足我们设置的 80% 的标准,因此分支 1 可以停止分裂。

在分支 0 中,FEE 不及格人数为 85 人(见表 2),FEE 和 ACS 都不及格的人数为 60 人,准确率为 70.59%,不满足我们的要求,因此需要进一步分裂。

为确定下一个分支结点,我们用上述方法计算除根结点之外的另三个属性的信息增益率,结果显示,属性 EMD 具有最大的信息增益率,因此它被选择为根结点的分支 0 的下一个分支结点。

同样属性 EMD 也有两个取值 0 和 1,所以也分裂为分支 1 和分支 0。由表 4 可以看到,在 FEE 和 EMD 都不及格的学生中,有 18 人 ACS 成绩不及格,有 3 人 ACS 成绩及格,所以在 EMD 的分支 0 上,ACS 不及格的估计准确率为 18/21 = 85.7%。分支 0 满足预先设定的标准,可以停止分裂。

另外,在表 3 中也可以看到,在 FEE 不及格且 EMD 及格的学生中,有 52 人 ACS 成绩及格,12 人不及格,因此在 EMD 结点的分支 1 上,ACS 及格的估计准确率为: 52/64 = 81.3%,分支 1 满足预先设定的标准,也可以停止分裂。则我们所构造的决策树如图 1 所示。

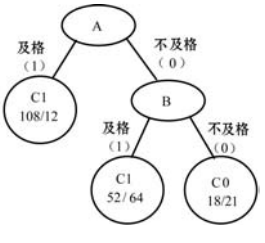

图 1 估计学生成绩决策树

图 1 中,结点 A——电工基础(FEE),结点 B——电机与拖动(EMD),结点 C1——自动控制系统(ACS)及格,结点 C0——自动控制系统(ACS)不及格。

2.4 分类规则描述

决策树算法的主要优势就是可以用来直接抽取分类规则。对于决策树的从根结点到每个叶结点的路径用 IF…THEN 的形式描述分类规则。这里仅以 ACS 属性提取的分类规则描述如下:

IF 电工基础成绩及格,THEN 自动控制系统成绩通常也及

(下转第 330 页)

2.3 接收用户文字输入功能的实现

在用户提交了虚拟会议室会场布置的结果后，系统会弹出2个文本输入框。用户可以手动敲入用户名、会务资源编号来确定会场布置。主要使用的BB模块如下：Creat System Font、Set Font Properties、2D Text、Input String、Identity、Create String。实现脚本如图7所示。

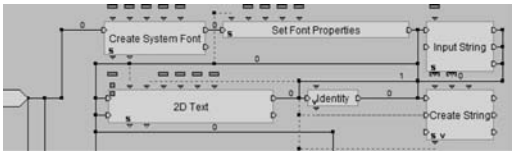

图7 文本输入的脚本

3 系统前台效果验证

原型系统在开发制作阶段是以CMO格式保存的，会务申请人员客户端需要同时安装Virtools DEV、Virtools Server、SQL Server客户端才能正常运行该CMO文件。因此，本系统单机版需要发布成exe文件供用户下载，在客户端单机运行。运行初始前台效果及自定义会场布置操作效果如图8和图9所示。

图8 系统初始化效果图 图9 自定义会场布置效果图

4 结 语

本文采用Virtools实现了将虚拟现实技术应用于会务服务的目的，设计了用户自定义会场布置系统。该系统仿真实用效果强，易于推广至其他领域。当前虽然关于Virtools的虚拟仿真文献和资料较多，但涉及Virtools与数据库关联的资料则相当少。此外，本文原型系统的后台仿真数据库设计也不够完善，仍需从用户需求出发，逐步完善。

参 考 文 献

[1] 黄炜. 基于Virtools的工厂虚拟漫游系统的设计与实现[J]. 软件设计开发，2011，7(3)：565－567.
[2] 肖灵君，刘紫薇. 基于3ds Max和Virtools的虚拟校园的开发与设计[J]. 科技信息，2009(12)：199－200.
[3] 杨琳，赵建民，朱信忠，等. 虚拟校园三维全景漫游技术研究[J]. 计算机工程与科学，2007，29(10).
[4] 肖杰华，陈立平，等. 面向生产制造过程的虚拟仿真系统建模研究[J]. 计算机工程，2002，28(6)：44－46.
[5] 张荣华. 几何建模技术在虚拟校园漫游系统开发中的应用[J]. 计算机工程与设计，2008，29(23).
[6] 赵玲，梁宏宝，等. 基于虚拟现实的水处理车间仿真的研究[J]. 计算机仿真，2006，23(6)：286－288.
[7] 刘明坤. VT游戏创作秘笈[M]. 北京：中国青年出版社，2010.
[8] 赵震，张兰成. 产品设计中交互效果图表现技法[M]. 北京：机械工业出版社，2011.
[9] 罗建勤，张明. Animation交互式漫游动画——Virtools＋3dsMax虚拟技术整合[M]. 北京：中国科学技术出版社，2010.
[10] 王丹东，徐英欣，胥林. 三维游戏设计师宝典Virtools行为模块词典大全[M]. 重庆：电脑报电子音像出版社，2009.

(上接第314页)

格，准确率为86.4%，学生人数的覆盖率是：125/210＝59.5%。

IF 电工基础成绩不及格 并且电机与拖动成绩也不及格，THEN 自动控制系统成绩通常不及格，准确率为85.7%，学生人数的覆盖率是：21/210＝10%。

IF 电工基础成绩不及格 但电机与拖动成绩及格，THEN 自动控制系统成绩一般及格，准确率为81.25%，学生人数的覆盖率是：64/210＝30.5%。

因此，我们可以得出结论：学生的电工基础课程学习的情况会严重影响到专业课自动控制系统的学习效果。学生的电机与拖动课程学习的情况也会影响到专业课自动控制系统的学习效果。因此，我们要教学生学好专业课自动控制系统，必须重视基础课电工基础课程教学效果。那些电工基础课不及格的学生，必须重视电机与拖动课程的学习，才能在自动控制系统课程上取得好成绩。

3 结 语

本文通过对学生成绩的数据分析，提出了提高学生的自动控制系统课程成绩的数据挖掘模型，采用决策树C4.5算法，进行分析，实验表明，应用该算法，构造的决策树分类正确。

数据挖掘技术近年来广泛应用于金融、保险、医药等行业，取得了一些重要成果。然而，在教学管理中使用的数据挖掘技术的成功案例较少。本文是使用数据仓库和数据挖掘技术的教学管理领域的一个尝试，但仍有一些问题需要进一步研究和探索。本文所涉及的内容，在将数据仓库和数据挖掘技术应用于教育领域探索出了一个切实可行的方法，可为进一步研究教育和教学管理决策支持系统打下一定基础。

参 考 文 献

[1] 刘红岩，等. 数据挖掘中的数据分类综述[J]. 清华大学学报：自然科学版，2002，42(6)：727－730.
[2] Gehrke J，Ramakrishnan R，Ganti V Rainforest. A framework for Fast Decision Tree Construction of Large Datasets[C]//Proceeding 1998 International Conference Very Large Data Bases(VLDB'98). New York，Aug 1998：416－427.
[3] Polat K，Gunes S. A Novel Hybrid Intelligent Method Based on C4.5 Decision Tree Classifier and One against all Approach for Multi Class Classification Problems[J]. Expert System with Applictions，2008.
[4] 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，2010：160－164.
[5] 陈志泊，等. 数据仓库与数据挖掘[M]. 北京：清华大学出版社，2009：111－125.
[6] 刘向峰，张洪伟，牟锐，等. 数据挖掘在销售管理系统中的应用[J]. 计算机应用研究，2004(6)：189－191.
[7] Quinlan JR. C4.5 Programs for Machine Learning[EB]. 1993.
[8] 王倩. 决策树在信息检索中的性能研究[J]. 微计算机信息，2008(1－3)：201－208.
[9] 钟晓，马少平，张钹，等. 数据挖掘综述[J]. 模式识别与人工智能，2001，14(1)：48－53.
[10] 张海笑，徐小明. 数据挖掘中分类方法研究[J]. 山西电子技术，2005(2)：20－21.
[11] 廖开际，等. 数据仓库与数据挖掘[M]. 北京：北京大学出版社，2008：167－173.