基于 C4.5 算法在水利水电建筑工程专业成绩分析中的应用*

胡　勇[1,2]，胡　玲[3]

（1. 重庆水利电力职业技术学院　信息工程系，重庆　402160；2. 重庆师范大学　数学与计算机科学学院，重庆　400047
3. 重庆大学　社会科学研究处，重庆　400044）

［摘　要］ 针对学生成绩问题，给出了学生成绩数据挖掘模型。决策树方法是数据挖掘中非常有效的分类方法。根据学生成绩数据特点，采用了 C4.5 决策树算法。C4.5 算法是决策树核心算法 ID3 的改进算法，它构造简单，速度较快，容易实现。选取决策属性，挖掘结果表明，该算法能够正确将学生成绩数据分类，并得到若干有价值的结论，供决策分析。

［关键词］ 数据挖掘；C4.5 算法；决策树
［中图分类号］TV；G642　　［文献标识码］A　　［文章编号］1005-2909(2006)04-0108-04

随着高校毕业生的增多，就业的压力越来越大，如何提高学生成绩是每一所高校的目标。影响学生成绩的因素很多，传统的统计分析方法已不适应深入分析的需要。笔者对学生成绩数据库应用数据挖掘中的 C4.5 算法进行了分析，得出了影响学生某科成绩的真实原因，教育管理人员可以此为依据，制定相应措施，提高教学效果。

一、决策树算法原理

决策树方法是数据挖掘的核心技术算法之一，它通过将大量数据有目的地分类，从中找出一些潜在的、对决策有价值的信息，用于预测模型中。国际上最早和最有影响的决策树方法是由 Quinlan 研制的 ID3 决策树生成算法。C4.5 算法是 ID3 算法的改进，该算法的基本工作流与 ID3 算法相同。决策树方法的基本思想是采用信息论中的概念，用信息增益作为决策属性分类判别能力的度量，进行决策节点属性的选择。C4.5 算法采用信息增益率作为属性选择的度量标准，理论和实验表明，采用信息增益率比采用信息增益更好。

C4.5 算法中，决策属性信息增益的计算方法如下：

设 S 是训练样本数据集，S 中类别标识属性有 m 个独立的取值，也就是说定义了 m 个类 $C_i, i=1,2,\cdots,m$；R_i 为数据集 S 中属于 C_i 类的子集，用 R_i 表示子集 R_i 中元组的数量。

集合 S 在分类中的期望信息量可以由以下公式给出，$I(r_1,r_2,\cdots,r_m)=-\sum_{i=1}^{m} p_i-\log_2(p_i)$ 式中：p_i 表示任意样本属于 C_i 类的概率；$p_i=r_i/|s|$，$|s|$ 为训练样本数据集中的元组数量。假设属性 A 共有 v 个不同的取值 $\{a_1,a_2,\cdots,a_v\}$，则通过属性 A 的取值可将数据集 S 划分为 v 个子集，其中 S_j 表示在数据集 S 中属性 A 的取值为 a_j 的子集，$j=1,2,\cdots,v$，如果 A 被选为决策属性，则这些子集将对应该节点的不同分枝。

如果 s_{ij} 表示 s_j 子集中属于 C_i 类的元组的数量，则属性 A 对于分类 $C_i(i=1,2,\cdots,m)$ 的熵可由下式计算：$E(A)=\sum_{j=1}^{v}\frac{S_{1j}+\cdots\cdots+S_{mj}}{|S|}I(S_{1j},\cdots\cdots,S_{mj}C)$

属性 A 的每个取值对分类 C_i 的期望信息量 $I(S_{1j},\cdots\cdots,S_{mj})$，可由下式给出）$I(S_{1j},\cdots\cdots,S_{mj})=-\sum_{i=1}^{m} p_{ij}\log_2(P_{ij})$ 式中：$P_{ij}=S_{ij}/|S_j|$，它表示在 S_j 子集中属于 C_i 类的比重。

由此可得到对属性 A 作为决策分类属性的度量值（称为信息增益）为：

* ［收稿日期］2006-09-21
［作者简介］胡　勇（1969-），男，重庆人，重庆水利电力职业技术学院讲师，重庆大学硕士研究生，从事数据挖掘和人工智能研究。

$$Gain(A) = I(r_1, r_2, \cdots, r_m) - E(A),$$

信息增益率为 Ratio(A):Gain(A)/E(A)。

该算法需要计算每个决策属性的信息增益率,具有最大信息增益率的属性被作为给定数据集 S 的决策属性节点,并通过属性的每一个取值建立由节点引出的分枝。

二、实例分析

下面通过一个实例来看一下数据挖掘的步骤及决策树方法的实现。

(一)确定成绩对象

下面以某高校的水利水电建筑工程专业的学生成绩数据为例,希望从学生成绩中发现各科成绩的关联,从而提高整体学习成绩。为此选定一个数据模型:学生成绩数据库,含学号、水利工程施工技术、水工制图、工程力学、水工建筑物、水工概预算这些字段,具体见表1。

表1　学生成绩

学号	水利工程施工技术	水工制图	水工概预算	水工建筑物	工程力学
1	76	71	68	71	81
2	71	65	63	72	74
3	60	36	67	28	80
4	67	84	71	61	78
5	64	58	72	54	72
6	73	80	66	58	67
7	62	81	78	52	79
8	74	78	47	60	56
9	62	48	63	52	67
10	72	73	73	49	60
……	……	……	……	……	……
161	65	53	89	50	50

(二)数据准备

第一步:将上表中的数据规范化,用0表示成绩小于60分,1表示成绩大于或等于60分,得到一个新表:以方便下一步数据挖掘的工作。

第二步:选取训练实例集。

从所有学生中进行抽样,将抽样数据作为训练集,共计有161条记录。经统计,在这161条记录的训练集中单科成绩及格人数和不及格人数如表2所示:

表2　成绩抽样统计

	水利工程施工技术	水工制图	水工概预算	水工建筑物	工程力学
及　格	82	57	34	32	39
不及格	79	104	127	129	122

三、依据C4.5算法构造决策树

(一)选取训练样本数据集,取属性"水工建筑物"作为类别标识属性,属性"水利工程施工技术""水工制图"、"水工概预算"、"水工建筑物"、"工程力学"作为决策属性集。训练样本数据集中,共

有161个元组,其中水工建筑物类所对应的子集中元组个数,分别为:水工建筑物及格人数 $P=32$,不及格人数 $N=129$,为了计算每一个决策属性的信息增益,首先计算课程"水工建筑物"所含有的期望信息量:

$$I(P,N)=I(32,129)=-\frac{32}{161}\log_2\frac{32}{161}-\frac{129}{161}\log_2\frac{129}{161}=0.7195$$

然后计算水利工程施工技术决策属性的期望信息量(即熵值)。当水利工程施工技术成绩=1和=0,且水工建筑物及格和不及格时,计算出水利工程施工技术所包含的总信息量。经统计,水利工程施工技术和水工建筑物的统计数据如表3所示。

表3　成绩搭配表

成绩搭配	人数
水利工程施工技术成绩=1且水工建筑物成绩=1	28
水利工程施工技术成绩=1且水工建筑物成绩=0	54
水利工程施工技术成绩=0且水工建筑物成绩=1	4
水利工程施工技术成绩=0且水工建筑物成绩=0	75

可得到:

E(水利工程施工技术)=(82/161)I(28,54)+(79/161)I(4,75)=0.6136

因此属性"水利工程施工技术"的信息增益为:

Gain(水利工程施工技术)=I(P,N)-E(水利工程施工技术)=0.7195-0.6136=0.1059

属性"水利工程施工技术"的信息增益率为:

Ratio(水利工程施工技术)=Gain(水利工程施工技术)/E(水利工程施工技术)=0.1726

同理可得其他课程的信息增益和信息增益率,结果如表4所示:

表4　课程信息增益率

	水工制图	水工概预算	工程力学	水利工程施工技术
Gain	0.2136	0.095	0.1701	0.1059
Ratio	0.4222	0.1521	0.3096	0.1726

由于属性"水工制图"具有最大信息增益率,故而选择该属性作为该决策树的根节点,可以看出所有课程当中水工制图是最能区别训练集中决定水工建筑物成绩与否的课程。

(二)创建一个树结点,并创建该结点的子链,每个子链代表所选属性的一个唯一值。使用子链的值进一步细化子类。当出现以下两种情形之一时可以停止分类:1.一个结点上的数据都是属于同一类别;2.没有属性可以再对属性进行分割。根据各个课程的信息增益率,应该选择水工制图作为所建决策树的根结点。由于水工制图的属性值只有两个:1(及格)和0(不及格),所以在水工制图下可以建立两个分支。经统计,水工制图不及格且水工建筑物不及格的人数为100,其准确率为100/104=96.2%。因此对水工制图不及格这个分支停止分割。又经统计,水工制图及格的57人中有26人水工建筑物及格,31人水工建筑物不及格,所以应对水工制图及格这个分支进行分割。从上表可知,应该选取工程力学作为分割结点进行细化。分割后经统计显示,水工制图和工程力学都及格的学生中,有26人水工建筑物及格,6人水工建筑物不及格,准确率为26/32=81.3%;水工制图及格但工程力学不及格的学生中,有22人水工建筑物不及格,3人水工建筑物及格,准确率为22/25=88%。由此可构建出数据的决策树,如图1所示.

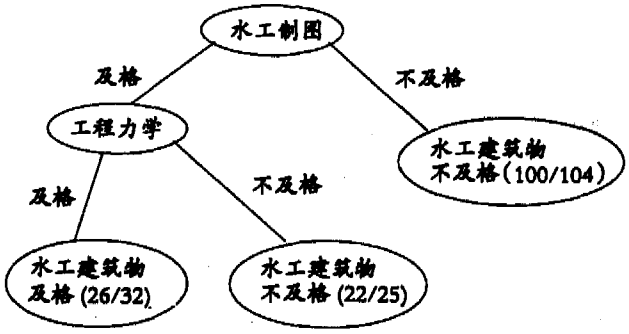

图1　C4.5生成的决策树

四、规则知识描述

从决策树中只提取属性"水工建筑物"类的规则。分类规则如下:

由该决策树可以得出下列规则:

(1)IF 学生的水工制图成绩不及格

THEN 其水工建筑物成绩通常也不及格。

准确度=(104-4)/104=96.2%

覆盖率=104/161=64.6%

(2)IF 学生的水工制图及格且工程力学成绩不及格,

THEN 水工建筑物成绩不及格。

准确度=(32-6)/32=81.3%

覆盖率=32/161=20%

(3)IF 学生的水工制图成绩及格且工程力学成绩及格

THEN 其水工建筑物成绩及格

准确度=(25-3)/25=88%

覆盖率=25/161=16%

由以上规则可以看出,学生水工制图的学习效果将直接影响其对水工建筑物的学习效果。工程力学的学习对水工建筑物的学习也有一定的影响。因此在进行水工建筑物教学时应考虑学生的水工制图基础。水工制图程度较好而水工建筑物程度一般的学生应更重视工程力学的学习。

五、结论

通过对学生成绩数据库的分析,提出了提高水工建筑物成绩的数据挖掘模型,采用数据挖掘中的核心算法C4.5决策树算法来进行分析,实验表明应用该算法使数据挖掘构造简单、分类正确、速度较快。

〔参考文献〕

〔1〕史忠植.高级人工智能[M].北京:科学出版社,1998.

〔2〕谭旭,王丽珍,卓明.利用决策树发掘分类规则的算法研究.云南大学学报,2000,22(6):45-49.

〔3〕唐华松,姚耀文.数据挖掘中决策树算法的探讨[J].计算机应用研究2001,(8):21-23.

〔4〕敖南.基于决策树的分类方法研究[D].南京:南京师范大学,2003

〔5〕李雄飞,李军.数据挖掘与知识发现[M].北京:高等教育出版社,2003

〔6〕刘向锋,张洪伟,牟锐,等.数据挖掘在销售管理系统中的设计和实现[J].计算机应用研究,2004(6):189-191.

〔7〕谷琼,朱莉,蔡之华,袁红星.基于决策树技术的高校研究生信息库数据挖掘研究[J].《电子技术应用》2006,(1):20-21.

The water conservancy water electricity construction engineering professional result analysis of application Base on C4.5 algorithm

HU Yong[1,2], HU Ling[3]

(1. *Information Engineering Department, Chongqing College of Water Resources and Electric Engineering, Chongqing* 402160, *China*;
2. *Mathematics and Calculator Science College, Chongqing Normal University, Chongqing* 400047, *China*;
3. *Office of Social Sciences, Chongqing University, Chongqing* 400045, *China*)

Abstract: Aim at the student the result problem, give student the result data scoops out the model. the decision tree method is a very valid classification method, in the data that scoop out. according to student the result data characteristics, adopted the C4.5 decision tree algorithm. C4.5 algorithm are the improvement algorithm of the decision trees core algorithm ID3, it construct in brief, the speed compare quickly, easy realization. selection decision belong to sex, scoop out the result enunciation, that algorithm can be right to get student the result data classification, and some worthy conclusion, provide the decision the analysis.

Key words: data mining; C4.5 algorithm; decision tree