ID3算法在教学质量评价中的应用研究

冯 菁,姚宏亮
(合肥工业大学 计算机与信息学院,安徽 合肥 230009)

摘 要:决策树是数据挖掘中的一种分类算法,它是一种以实例为基础的归纳学习算法,来发现数据模式和规则。介绍了数据挖掘的定义及分类,详细介绍了决策树ID3算法。又根据ID3算法,对院校中收集的大量教学评价数据样本进行分析,获得不同属性上的信息增益,生成最终决策树,可将此树转换成一个if—then规则的集合。生成规则和决策树,然后对新数据进行分析和预测。通过数据建模以发现规律和模式,从而提取有价值的信息,避免目前教学质量评价中的不合理性,实例验证和分析的结果表示该方法的有效性。为教学质量评价提供合理、科学的决策支持,从而提高教学质量,改进教学成果。

关键词:数据挖掘;教学质量评价;ID3算法;决策树
中图分类号:G434 **文献标识码**:A **文章编号**:1673-629X(2011)06-0250-04

Application of ID3 Algorithm in Teaching Quality Evaluation

FENG Jing, YAO Hong-liang
(School of Computer and Information, Hefei University of Technology, Hefei 230009, China)

Abstract: Decision tree algorithm is a classical data mining classification algorithm, it's one of inductive learning based on instances. And it can be used for classification and prediction tasks. First, introduces the definition as well as the main technologies of data mining. Then, it describes the ID3 algorithm in details, and optimizes the ID3 algorithm. The basic ideas and implementation methods of ID3 algorithm is discussed. A kind of date mining algorithm ID3 has been used in the processing of the teaching quality evaluation date, and some rules are established. Meanwhile, advices on how to improve the defects are put forward. Evaluation of teaching quality is a very important part of teaching process, this decision information for the teaching manager to improve the quality of teaching, so as to continuously improve the quality of education.

Key words: data mining; teaching quality evaluation; ID3 algorithm; decision tree

0 引 言

我国高等职业技术教育培养的是高素质技能型人才,高职院校教师教学质量的评价是依据教师教学工作特点、社会需要、人才培养规格和特定教学对象等因素来进行的。这样才能更全面、更准确、更合理地评价教师教学的各项工作。

而每一所院校每一学期、每一学年都会保留大量的数据材料,如何从积累的数据中获取有效的信息,对教师做出聘任、晋升,或增加奖金等奖、惩方式,为此决策提供有说服力的依据,做到客观公正,通过合理的奖惩调动教师的工作积极性有着重要的意义,有助于学校管理者对教师进行适度的监督和控制,从而提高教育教学质量。

收稿日期:2010-11-21;**修回日期**:2011-02-21
基金项目:国家自然科学基金项目(61070131)
作者简介:冯 菁(1974-),女,硕士,高级工程师,研究方向为数据挖掘;姚宏亮,博士,副教授,主要研究方向为数据挖掘与机器学习。

1 数据挖掘

1.1 数据挖掘(Data Mining)的定义

随着科技的发展,计算机、网络、数据库等技术广泛应用于日常管理中,各行各业积累了大量的信息数据,对数据库的存取与查询操作,已远远不能满足要求。人们需要从海量数据中获得这些数据背后的更重要信息,如数据的整体特征描述,企图发现事件间的相互关联,以及发展趋势进行预测[1]。

数据挖掘,从数据中挖掘知识,就是从大量的、不完全的、有噪声的、模糊的、随机的数据中,提取隐藏在其中的、人们事先不知道的、但又是潜在有用的信息和知识的过程[1-4]。与数据挖掘相近的术语有:从数据库发现知识(KDD)、数据分析、知识抽取、模式分析、信息收割、数据融合以及决策支持等。人工智能领域称为知识发现,而数据库领域则称为数据挖掘[1]。

1.2 数据挖掘的分类

数据挖掘分类方法有多种,根据挖掘任务可分为

分类和预测模型发现、数据总结、聚类、关联规则发现、序列模式发现、依赖关系或依赖模型发现、异常和趋势发现等[2]。根据挖掘方法,可分为机器学习方法、统计方法、神经网络方法和数据库方法。机器学习包含归纳学习方法、基于案例的学习、遗传算法等。统计方法包含回归分析、判别分析、聚类分析、探索性分析等。神经网络方法包含前向神经网络、自组织神经网络等。数据库分析方法包含多维数据分析方法、面向属性的归纳方法等[2]。

1.3 决策树方法

大部分数据挖掘方法采用规则发现技术或决策树分类技术来发现数据模式和规则。其核心是某种归纳算法,这类方法通常先对数据库的数据进行挖掘,生成规则和决策树,然后对新数据进行分析和预测。这类方法的主要优点是规则和决策树都是可读的[1]。

决策树是数据挖掘中的一种分类算法,它是一种以实例为基础的归纳学习算法[5,6],同时也是一种着眼于一组无次序、无规则的事例,推理出决策树表示形成的分类规则[5]。该算法是利用树形结构来表示决策集合,这些决策集合通过对数据样本集的分类产生规则。树的每一个非叶子结点表示对一个属性的测试,其分枝代表测试的每个结果,每个叶结点代表一个类别。在建树的过程中,需要使用剪枝来剪去数据中的噪声和局外者,从而提高在未知数据上分类的可靠性[7]。决策树分为分类树和回归树两种,分类树对离散变量做决策树,回归树对连续变量做决策树。常用的决策树算法有ID3、C4.5、CART等[2,5,8]。

2 ID3算法

1986年Quinlan提出了著名的ID3算法。ID3算法是基于信息熵的决策树分类算法,该算法的核心是在决策树中各级结点上选择属性,用最高信息增益作为结点的测试属性[9],使得在每一个非叶子结点进行测试时,能获得关于被测试样本集最大的类别信息,使用该属性将样本集分成子集后,系统的熵值最小。期望该非叶子结点到达各后代叶结点的平均路径最短,使生成的决策树平均深度较小,从而提高分类速度和准确率[10]。

树的生成算法(ID3):

设S是s个数据样本的集合,假定决策属性,具有m个不同的值,定义m个不同类C_i,($i=\{1,\cdots,m\}$),s_i是类C_i中的样本数。对一个样本分类的期望信息可由下面公式(1)给出:

$$I(s_1,s_2,\cdots,s_m)=-\sum_{i=1}^{m}\frac{s_i}{s}\log_2\frac{s_i}{s} \tag{1}$$

如果以属性A作为决策树的根,属性A具有v个不同值$\{a_1,a_2,a_3,\cdots,a_v\}$,它将$S$分成$v$个子集$\{S_1,S_2,\cdots,S_v\}$,其中$S_j$包含$S$中这样一些样本,它们在$A$上具有值$a_j$,则这些子集对应于由包含集合$S$的结点生长出来的分枝。对于给定的子集$S_j$,有公式(2):

$$I(s_{1j},s_{2j},\cdots,s_{mj})=-\sum_{i=1}^{m}\frac{s_{ij}}{s}\log_2\frac{s_{ij}}{s} \tag{2}$$

根据A划分成的子集的熵由公式(3)给出:

$$H(A)=\sum_{j=1}^{v}\frac{s_{1j}+\cdots+s_{mj}}{s}I(s_{1j}+\cdots+s_{mj}) \tag{3}$$

熵值越小,子集划分的纯度越高。在属性A上分枝将获得的信息增益为公式(4):

$$\mathrm{Gain}(A)=I(s_1,s_2,\cdots,s_m)-H(A) \tag{4}$$

ID3选择Gain(A)最大的属性A作为样本集的根结点[6,11],各分枝的样本子集递归使用ID3方法,建立决策树结点和分枝,直到样本子集属于同一类。这种方法使生成的决策树平均深度最小,有较快的速度,这样就生成了一棵决策树。

3 ID3算法在教师教学评价中的应用

数据挖掘的决策树方法在淮北职业技术学院计算机系中教师业务档案信息管理中得到成功应用。这里只讨论授课教师课堂教学的数据挖掘,不涉及教师科研与职称情况。

3.1 数据预处理

大致分为三个步骤:数据选取集成、数据预处理、数据变换[2,7]。具体参考文献[3,6]。

本系任课教师教学质量评估指标体系表共三份,一份是教师同行评议,一份是系领导评议,还有一份是学生评议,三份表的内容如表1所示,格式类似,共两级指标,每项分值10分。

表1　淮北职业技术学院教师教学质量评估表(二)

(系领导、教师评估用表)

系别:　　课程:　　考评人员:　　计分:　　表1

一级指标	二级指标		分值分配	得分
教学内容	正确性	讲课内容正确,概念清楚……	4	
	符合教学大纲	符合大纲要求,重点突出,深入浅出……	2	
	先进性	本学科学术动态与新进展……	1	
	理论联系实际	实践教学……	3	
教学态度	备课情况	教学准备充分、教案规范、条理清楚、重点突出	3	
	教书育人	责任心强,爱岗敬业……	3	
	遵守教学纪律	按授课计划授课,不迟到……	2	
	组织教学		2	
教学方法	略	略	10	
教学效果	略	略	10	
评价等级				

收集三份调查打分表,汇总计算出各位教师的日常教学质量,并算出平均数。现对四十九名教师某年

度考核的教学质量评议表进行讨论，为了把问题简单化，把上述表中四项一级指标作为数据库单表（jx 表名）中的字段（属性），分别为 A1，A2，A3，A4。计分参照分配到系里评优指标，分为优（9-10），良（7-8.9），中（其他）三个等级。最终结果（评价等级）作为单表的第五个字段（A5 属性）。A1，A2，A3，A4 中的等级划分为：1、优 2、良 3、中，按此要求进行数据转换。得出本系教师日常教学质量数据（见表2）。

ID	教学内容A1	教学态度A2	教学方法A3	教学效果A4	
1	良	良	良	优	良
2	良	良	良	良	良
3	良	良	良	中	良
4	优	优	优	优	优
5	中	良	中	中	中
6	中	中	中	良	中
7	中	中	良	中	中
8	良	优	优	良	良
9	良	良	中	良	良
10	中	良	良	中	良
11	优	良	良	优	优
12	中	优	良	良	良
13	良	良	优	中	良
14	良	良	中	中	良
15	中	良	中	中	中
16	优	良	良	优	优
17	良	良	中	中	中
18	优	良	优	良	优
19	中	良	优	良	良
20	中	优	优	良	良
21	良	中	中	良	良
22	中	良	中	良	中
23	良	优	中	良	良
24	良	优	优	优	优
25	良	优	中	良	良
26	良	中	中	中	中

3.2 ID3 算法建立决策树

对所有属性进行信息增益计算，先计算该样本对于 A5 类别属性的期望信息（信息熵）。

分析表中数据，其中的优、良、中人数各为 11，26，12。则有：

I(11,26,12)= 1.46604

下面计算每个属性的条件信息熵和信息增益。

（1）对于属性 A1，属性值为“优”共 13 人，其中类别属性 A5 为优，良，中人数分布为 10，3，0。s11=10，s21=3，s31=0；属性值为“良”，共 19 人，其中类别属性 A5 优，良，中人数分布为 1，15，3。s12=1，s22=15，s32=3；属性值为“中”，共 17 人，其中类别属性 A5 优，良，中人数分布为 0，8，9。s13=0，s23=8，s33=9。

查询 SQL 语句：Select Jx. ID from Jx where (((jx. A5="中") and ((jx. A1)="优"))显示 A1 为优，A5 结果为中的那些人。

I(10,3,0)= 0.77935

I(1,15,3)= 0.91328

I(0,8,9)= 0.9975

H(A1)= 13/49 * I(10,3,0) +19/49 * I(1,15,3) +17/49 * I(0,8,9)= 0.906967

（2）对于属性 A2，属性值为“优”共 10 人，其中类别属性优，良，中人数分布为 2，7，1。属性值为“良”，共 26 人，其中类别属性优，良，中人数分布为 7，13，6。属性值为“中”，共 13 人，其中类别属性优，良，中人数分布为 2，6，5。

H(A2)= 1.41833

（3）对于属性 A3，H(A3)= 1.16766。

（4）对于属性 A4，H(A4)= 1.207373。

它们的信息增益分别为：

Gain(A1)= 0.55907

Gain(A2)= 0.04771

Gain(A3)= 0.29838

Gain(A4)= 0.258667

根据 ID3 算法取 Gain(A1) 为根结点。因为 A1 的信息增益最大。先按 A1 为属性分类，得到如下决策树。再继续用上述算法递归。

生成的决策树如图 1 所示。

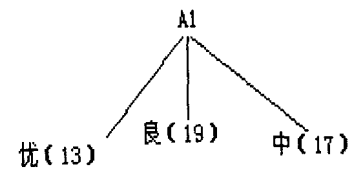

图 1 按“教学内容”生成的决策树

3.3 I 分枝递归 ID3 算法

A1 属性值为“优”子树的划分，在此 13 条记录中，所需的期望信息 I(10,3,0)= 0.77935。

（1）计算 A2 值为“优”，“良”，“中”，类别属性为“优”，“良”，“中”的条件熵的计算为：

I(1,0,0) = 0；I(7,1,0)= 0.54356；I(2,2,0)= 1

H(A2)= 0.64219。

（2）计算 A3 属性条件熵，A3 值为“优”，类别属性为“优”，“良”，“中”的人数分别为 3，0，0。A3“良”，类别属性“优”，“良”，“中”的人数为 7，0，0。A3 为“中”，类别属性“优”，“良”，“中”的人数为 0，0，3。

H(A3)= 0。

（3）计算 A4 属性条件熵，H(A4)= 0.67667。

Gain(A2)= 0.137159

Gain(A3)= 0.77935，

Gain(A4)= 0.102678

选 A3 作为子树的根结点，对此分支继续 ID3 算法递归，得到决策树如图 2 所示。最终得到的决策树如图 3 所示。

为了增加决策树的可读性及可理解性，需要对决策树进行修剪。但是对于测试数据来说，修剪决策树必然会导致误差率的增加，我们设计一个允许最大误差率，得到一棵经过剪枝后的决策树。本例用先剪枝

方法，通过提前停止树的构造（例如，通过决定在给定的结点上不再分裂或划分训练样本的子集）而对树“剪枝”。

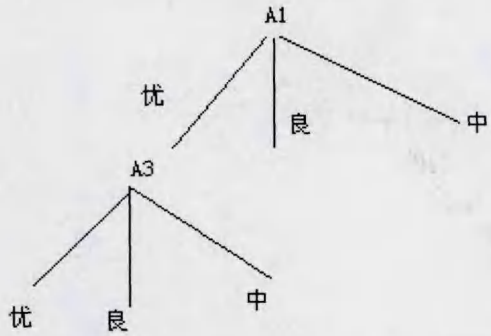

图2　按“教学方法”第二次生成的决策树

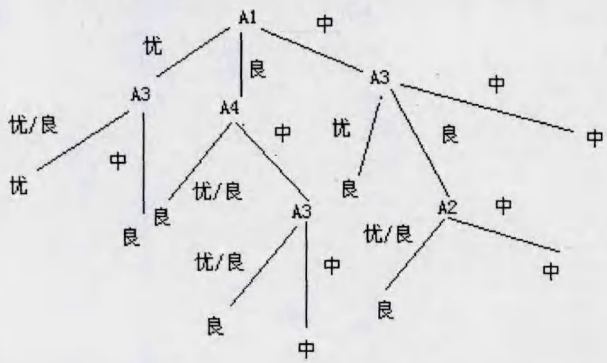

图3　ID3 算法最终生成的决策树

4　生成的分类规则及分析

根据 ID3 算法生成的最终决策树，可将从此树转换成一个 if—then 规则的集合，从根结点，分支，到叶子结点，每个条路径对应一组属性测试的合取，决策树代表这些合取式的析取。决策树最大的优点就是可以直接提取分类规则。

本例中所生成的分类规则如下：

(1) if A1 = 优 and A3 = 优 thenA5 = 优

(2) ifA1 = 优 and A3 = 良 then A5 = 优

(3) if(A1 = 优 andA3 = 中) thenA5 = 良

(4) if(A1 = 良 andA4 = 优) thenA5 = 良

(5) if(A1 = 良 andA4 = 良) thenA5 = 良

(6) if(A1 = 良 and A4 = 中 and A3 = 优) then A5 = 良

(7) if(A1 = 良 and A4 = 中 and A3 = 良) then A5 = 良

(8) if(A1 = 良 and A4 = 中 and A3 = 中) then A5 = 中

(9) if(A1 = 中 andA3 = 优) thenA5 = 良

(10) if(A1 = 中 andA3 = 中) thenA5 = 中

(11) if(A1 = 中 andA3 = 良 andA2 = 优) thenA5 = 良

(12) if(A1 = 中 andA3 = 良 andA2 = 中) thenA5 = 中

从上面分析可得出，A1 教学内容属性在此评价体系中最重要。从这里可以得出较公正的客观评价。其次教学方法也是比较重要的指标。学校也可依照此结果，改进教学质量，提升教师在某一方面的改进工作。教师在增进教学内容基础上，改进教学方法，端正教学态度，提升教学效果。

5　结束语

目前在数据挖掘领域中，存在许多解决分类问题的模型，最为广泛的分类模型仍然是决策树算法。决策树方法在分类过程中，不需要人为设定任何参数，更适合于知识发现的要求；决策树分类方法不需要任何除测试数据集以外的附加信息，保证决策树与其它分类方法相比具有更高的分类速度，具有非常好的分类准确率。

ID3 算法是决策树最有影响的算法，把决策树技术 ID3 算法应用于高等职业教育教学质量评价、学生成绩预测、教师师资管理等方面，给予公平、公正、客观的评价标准，学校利用这些隐藏的信息发现有用的价值，为教学管理部门提供决策支持依据，指导日常教学与管理工作，提升管理水平，更好地开展教学工作，以合理的理念、科学的态度引导高职院校的发展，提升整个学校的办学水平。

参考文献：

[1] 张友生，徐　峰. 系统分析师技术指南[M]. 北京：清华大学出版社，2004.

[2] 邵峰晶，于忠清. 数据挖掘原理与算法[M]. 北京：中国水利水电出版社，2003.

[3] 肖志明. 决策树算法在高校教学评价中的应用研究[J]. 广西轻工业，2008，24(11)：164-167.

[4] 覃宝灵. 决策树技术在教学质量评价中的应用研究[J]. 电脑知识与技术，2007，3(13)：191-192.

[5] 郭亮山. 浅淡数据挖掘技术在公安领域中的应用[J]. 福建警察学院学报，2008(4)：32-36.

[6] 杨　静，张楠男，李　建，等. 决策树算法的研究与应用[J]. 计算机技术与发展，2010，20(2)：114-116.

[7] Han JiaWei, Kamber M. Data Mining: Concepts and Technique[M]. 北京：高等教育出版社，2001.

[8] 袁　燕. 决策树算法在高校教学评价系统中的应用[J]. 浙江海洋学院学报，2006，25(4)：440-444.

[9] 桂维魁，陈　涛，柳　洋. 基于 ID3 算法的考试成绩分析决策树的构造[J]. 天津城市建设学院学报，2008，14(2)：139-141.

[10] 李　霞. ID3 分类算法在银行客户流失中的应用研究[J]. 计算机技术与发展，2009，19(3)：158-160.

[11] Quinlan J R. Induction of decision tress[J]. Machine learning，1986(8)：81-106.