决策树算法在智能导学系统中的应用

邱 涛,李 雯
(江西理工大学 信息工程学院,江西 赣州 341000)

摘 要:决策树算法是一个经典的数据挖掘分类算法,如今已经被广泛应用到各个领域,并且取得了很好的效果,此外,对决策树算法的改进也在不断的进行中。将决策树算法应用在智能导学系统中,其目的是为了使智能导学系统能更好对学习者进行分类。采用的方式是应用决策树算法对学习者输入的资料对其进行分类,并对不同类型的学习者应用不同的教学计划。结果表明应用决策树算法分类能明确的把握学习者的特性,提高系统的分类效率。由此得出结论,将决策树算法应用在智能导学系统中是十分可行的。

关键词:决策树算法;智能导学系统;分类
中图分类号:TP18 **文献标识码**:A **文章编号**:1673-629X(2009)12-0189-04

The Application of Decision Tree Algorithm in Intelligence Teaching System

QIU Tao, LI Wen
(School of Information Engineering, Institute of Technology of Jiangxi, Ganzhou 341000, China)

Abstract: Decision tree algorithm is a classical data mining classification algorithm, it's widely used in every field and has many great effect nowadays. Further more, the amelioration to decision tree algorithm is keeping on. The purpose to apply decision tree algorithm in intelligence teaching system is to lets the system do better classification. Use decision tree algorithm to classcify the learner by their input information, and apply different learning program to them. The result shows the application of decision tree algorithm lets system get the learner's trait clearly and improve the classification efficiency. There is conclusion that apply decision tree algorithm to intelligence teaching system is feasible.

Key words: decision tree algorithm; intelligence teaching system; classification

0 引 言

智能导学系统是在 Internet 的远程教育系统的基础上,加入人工智能技术,使系统具有智能性,能更好地适应学习者的要求。基于 Internet 的远程教育具有开放性、灵活性、学习终身性和资源共享性等优点。充分利用这些优点不仅可以满足学习者个性化学习需要,而且可以在很大程度上提高学习效率。但据初步调查表明,大多数学习者在网络学习时,会遇到不同程度上的困难,主要原因是目前的网络学习系统中存在网络课程机械化、学习资源混杂化、缺少反馈、缺少有效的导学机制[1]等问题。利用软件工程、数据挖掘和人工智能等技术,分析设计智能导学系统,可以很好地解决上述问题,体现出系统的智能性。

收稿日期:2009-03-23;**修回日期**:2009-06-27
基金项目:江西省科学技术研究项目(GJJ08285)
作者简介:邱 涛(1986-),男,江西赣州人,硕士研究生,研究方向为数据挖掘,人工智能;李 雯,教授,硕士生导师,研究方向为基于网络的数据库应用技术。

1 数据挖掘中的分类技术

1.1 分类技术

数据挖掘的主要任务有分类分析、聚类分析、关联分析、序列模式分析等,其中的分类分析由于其特殊地位,一直是数据挖掘研究的热点之一,至今已经提出很多算法。数据挖掘分类就是分析输入数据,通过在训练集中的数据表现出来的特性,为每一个类找到一种准确的描述或者模型[2]。

分类就是把一些新的数据项映射到给定类别的中的某一个类别,例如一篇文章在发表时,可以自动地把这篇文章划分到某一个文章类别,一般的过程是根据样本数据利用一定的分类算法得到分类规则,新的数据过来就依据该规则进行类别的划分。分类在数据挖掘中是一项非常重要的任务,有很多用途,比如预测,即从历史的样本数据推算出未来数据的趋向。还有分

析用户行为,通过这种分类,可以得知某一商品的用户群,对销售来说有很大的帮助。分类器的构造方法有机器学习方法、神经网络方法等。常见的统计方法有KNN算法,基于事例的学习方法。机器学习方法包括决策树法和归纳法[3]。

1.2 分类技术在智能导学系统中的应用

分类技术在智能导学系统中也起到关键的作用,例如,需要根据用户进入系统时注册的信息,对学生进行分类,根据各个类别学生的特点,对学生实行不同的教学方案,并根据他们一段时间学习效果的反馈,不断更新学习的计划,充分体现了系统的智能性。学习者在进入智能导学系统时候,根据系统提示进行注册,由于不同的学习者具有不同的特性,不能将每个人都应用同一套教学计划,而预先拟定不同的教学计划。针对不同类型的学习者,因材施教提高他们的学习效率,而根据学习者资料对其行进的分类也更能体现出系统的智能性。

2 智能导学系统

2.1 智能导学系统简介

智能导学系统是利用各种成熟的数据挖掘技术对远程教育系统(E-Learning)进行优化,使系统更具有智能性,也更能符合学生的要求。随着 Internet 的普及和在教育领域中的应用,远程教育(E-Learning)得到了高速发展。基于 Internet 的网络教学,实现了全球信息资源的共享,有效地突破了时空局限,扩大了教学规模,使更多的人有了学习的机会。然而目前的远程教育存在资源利用率低,学生与学生之间,学生与教师之间,以及教师与课程安排人员之间相互独立,没有有效的信息反馈,没有考虑到学生的差异性,只是一些静态的网络教育资源的罗列。这造成 E-Learning 没有发挥出它自己应有的特色,而与传统教学一样,所有的学生都进行同样模式的教学,教学资源也没有得到有效共享。现代远程教育的全过程基本上都是通过浏览网站的形式进行的,学生在 Web 上的行为会产生大量的信息,这些信息在远程教育的全过程中十分宝贵。如何充分挖掘这些信息及其背后潜在的信息并反馈来指导远程教育中的各个环节,以此来扩大影响、吸引招生和为学生提供个性化的服务内容,从而增强远程教育的竞争力,已逐渐成为革新 E-Learning 技术中的一个研究热点。

2.2 智能导学系统的构建

目前国内外的研究者利用传统的数据挖掘算法,其中包括聚类、分类、关联规则、模式挖掘和文本挖掘来打造个性化远程教育系统。利用聚类方法将学生分组,然后基于同组中成功的相似学习者来给学生推荐好的学习方法和学习资源。利用数据挖掘工具为教育者获得更多的教学反馈,评估课程内容的结构和学习过程的有效性,从而能自适应调整教学计划和组织教学内容,提高学习者的学习效率。面向负责人和管理者,挖掘出知识来指导网站建设,从而提高网站效率以适应学生的学习行为(优化服务大小、网络传输分配等)。为解决当前的远程教育系统存在形式单一和被动教学等问题,提出了一个基于学习者个性因素的智能导学模型,从而满足学习者主动学习的要求。构建了一个个性化协作学习系统,并提出了一种新颖的打分或者交换的用户动态聚类算法,从学生的资源请求中发现学生兴趣,并有效地将具有相同兴趣的学生自动组成学习社区。

2.3 智能导学系统的关键问题

随着智能导学系统的不断发展,人工智能、数据挖掘等各种技术、理论越来越多的被应用在远程教育系统中,以体现系统的智能性。目前在智能导学系统中还有许多需解决的关键问题[1]:

(1)规则的产生和教学策略的制定。如何从大量的信息中提取有效的规则,然后根据规则制定与之相应的、行之有效的教学策略,这需要丰富的教学经验和反复、大量的实验。并且能够在教学的过程中,根据学习者的实际情况,动态地调整教学计划。

(2) 智能导学系统和学习工具、练习工具、作业工具、测试工具、答疑工具、交流工具等系统接口参数的确定。工具系统通过智能导学系统传递的参数,合理调用规则库里的规则。

(3)条件可信度、加权因子和规则可信度的确定。由于主观因素,领域专家确定的初始值因人而异,与实际情况存在误差。需要根据教学的实际效果,调整各初始值,使之更适应学生的实际情况。

(4)规则绩效的测评。建立怎样的一个测评系统,能有效地评估学习者在根据规则形成的个性化学习环境中取得的学习效果,给规则一个合理的评价,这是反馈系统的核心部分。

(5)智能导学系统个性化分析参数的调整和规则的自适应修改。根据规则绩效测评的结果,适当地修改个性化分析参数和规则可信度,使个性化分析出的规则更符合学习者的学习规律。

3 决策树算法在学生分类中的应用

3.1 决策树算法简介

决策树是一个类似于流程图的树结构,是一棵有向、无环树。树中的每一个结点代表数据集中的一个

属性,从根结点起,除叶结点以外每个结点都是对所代表属性的一次判断。根据判断的结果进入该结点的不同分枝,叶结点代表的是分类的结果。它是一种逼近离散值函数的方法,在这种方法中析取到的函数被表示为一棵决策树[4]。得到的决策树最顶层节点作为根节点,将其每一个子节点作为测试属性,每个分枝代表一个测试输出,而每个叶子结点代表类或类的分布。根据测试结果,选择某个分枝,为了分类一个特定数据项目,从根结点开始,一直向下判定,直到达到一个叶子结点为止。这样,一个决策树就形成了[4]。

决策树是一个预测模型,决策树算法的基础是自顶向下分裂的贪心算法[5]。相对与其他的多类分类方法,二叉决策树算法的生成过程中产生的节点数大大减少[6]。但在实际的智能导学系统应用中,面临的是多分类的问题,所以将决策树分类方向扩展为多分类。

以决策树算法 ID3 为例,其基本思想是采用信息论中的互信息(或称信息增益)作为决策属性分类判别能的度量,进行决策节点属性的选择。在 ID3 算法中,决策节点属性的选择应用了信息论中熵概念来完成,通过信息增益最大(或最大熵压缩)的属性建立决策树,这样选择的节点属性保证了决策树具有最小的分枝数量和最小的冗余度[7]。

$H(U) = -\sum_i P(u_i)\log P(u_i)$。其中 $|S|$ 表示例子集 S 的总数,$|u_i|$ 表示类别 u_i 的例子数,类别 u_i 出现的概率为:$P(u_i) = \frac{|u_i|}{|S|}$。

3.2　决策树算法的优点

总体而言,运算速度快,精度高。

(1)决策树方法不需要假设先验概率分布,这种非参数化的特点使其具有更好的灵活性和鲁棒性。

(2)决策树方法不仅可以利用连续实数或离散的数值样本,而且可以利用"语义数据",比如离散的语义数据:东,南,西,北等。

(3)决策树方法产生的决策树或产生式规则集具有结构简单直观,容易理解,以及计算效率高的特点。

(4)决策树方法能够有效地抑制训练样本噪音和解决属性缺失问题,因此可以解决由于训练样本存在噪声而使得分类精度降低的问题。

3.3　决策树算法的形式过程

决策树,一个树性的结构内部节点上选用一个属性进行分割,每个分叉都是分割的一个部分。叶子节点表示一个分布决策树生成算法分成两个步骤:1. 树的生成,开始数据都在根节点递归进行数据分片;2. 树的修剪,去掉一些可能是噪音或者异常的数据。决策树使用:对未知数据进行分割,按照决策树上采用的分割属性逐层往下,直到一个叶子节点。

(1)基本算法。

自上而下分而治之的方法开始时,所有的数据都在根节点属性都是种类字段 (如果是连续的,将其离散化)所有记录用所选属性递归的进行分割属性的选择是基于一个启发式规则或者一个统计的度量。

(2)停止分割的条件。

一个节点上的数据都是属于同一个类别没有属性可以再用于对数据进行分割。

(3)伪代码:

```
Procedure BuildTree(S)
    用数据集 S 初始化根节点 R
    用根结点 R 初始化队列 Q
    While Q is not Empty do {
      取出队列 Q 中的第一个节点 N
      if N 不纯 (Pure) {
        for 每一个属性 A
        估计该节点在 A 上的信息增益
        选出最佳的属性,将 N 分裂为 N1、N2
        }
    }
```

3.4　智能导学系统数据的预处理

为了体现远程教育系统具有智能性,必须先根据学生在进入系统时注册的信息中抽取出学生的特性,将学生进行分类,为不同类别的学生制定不同的学习方案,并根据学生在学习方案中的学习效果的反馈,动态地修改学生方案,使学习的过程尽量适合学生的特性,提高学生的学习效率。而在学习过后,系统还将根据学习的结果动态调整学习计划。

为了抽取学生的特性,必须根据学生的注册信息提取出有用的项,作为学生的属性 k_i,而学生的特性可以由多维向量 $K = \{k_1, k_2, \cdots, k_n\}$ 来表示。而根据已有学生的选课记录和学习效果,将学生预分为:经济管理类、信息技术类、机械建筑类、文学法律类、体育艺术类、未定型类(这里以系统认为的学生适合的专业作为类别名称、未定型类用于决策树算法后没有明确分类的个体)。此分类并不能完全概括所有学生的分类,如果有新的教材进入系统,会产生新的类别,这种情况可以利用其他数据挖掘算法解决系统中的矛盾。文中的分类方法只在已有的类别上讨论学生的适应程度。

根据学生注册时的信息,将学生的分类属性定为:年龄、性别、爱好、特殊天赋、所学专业、身体条件。其中为了简便起见,将爱好和特长首先预定几个值,所学专业则根据具体专业情况划分入以上几个大类,不在类中的以其他作为预定值。提取出的属性集合及其分类情况如表 1 所示(该导学系统主要针对对象为在校

大学生和研究生)。

表 1 分类属性

属性名	年龄	性别	爱好	特殊天赋	所学专业	身体条件
属性取值	1~30 岁 30~100 岁 其他	男 女	电器装配 体育运动 棋牌对弈 书籍阅读 其他、无	运动能力 逻辑思维 记忆能力 歌唱舞蹈 其他、无	经济管理 机械建筑 文学法律 信息技术 体育艺术 其他	体格强健 一般 体质偏差

根据表 1 的分类,具体学生情况可以细分为 4000 多种不同情况,从数据库中读出记录如图 1 所示。

1437	21	男	电脑游戏	无	信息技术
1438	26	女	书籍阅读	记忆能力	体育艺术
1439	33	男	棋牌对弈	运动能力	机械建筑
1440	45	男	体育干家	逻辑思维	文学法律

一般
体质偏差
体格强健
一般

图 1 数据库各学生的具体数据

3.5 决策树算法在学生分类中的应用

根据这些属性集合 K,并根据学习者所选的课程所属类别,进行决策树分类算法。取前 4000 条数据,图 2 为决策树的生成过程。

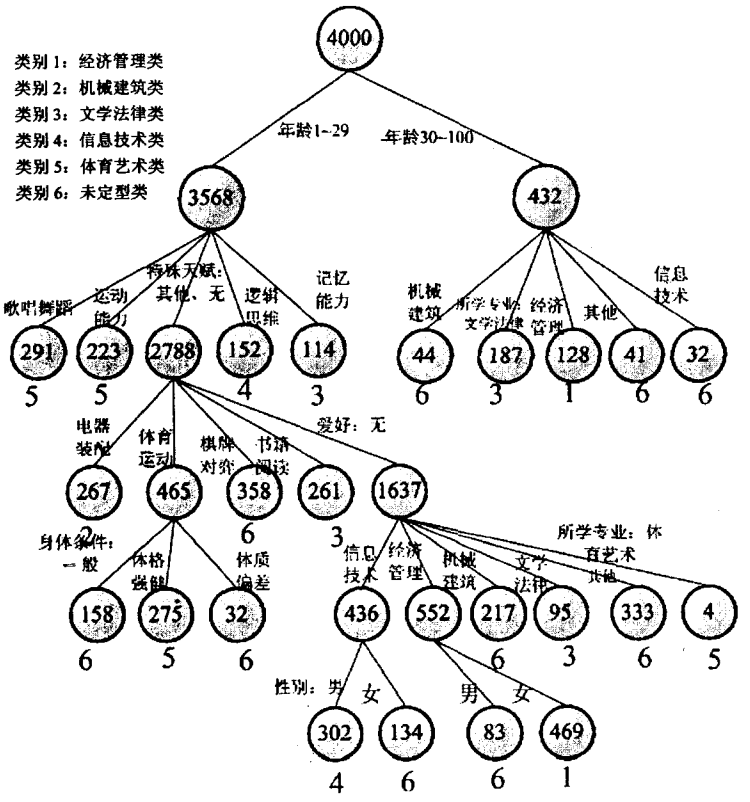

图 2 生成的决策树

步骤 1:将根节点按照各属性分裂得到熵值,计算机公式上述内容中已给出。

步骤 2:取能得到最大的熵值的属性作为决策树的分裂属性,并计算是否有类型超过阈值(阈值预设为 80%),如果有,则可以确定其类型。

步骤 3:依次对各未确定类型的节点进行属性分裂,直到所有属性分裂完毕。

步骤 4:对于没有某种类型超过阈值的节点,暂时归为未确定类型节点。

如上所述,决策树算法可以将新注册的学习者归入到 6 大类当中,对于前 5 类,智能导学系统将安排相关类型的教材及学习方案,并推荐给学生;对于未定型类,导学系统将给学生人数最多的几套方案,任学习者自己选取。

4 结束语

决策树分类算法运算速度快、精度高,决策树分类算法在智能导学系统中的应用是导学系统体现其智能性的第一步,也是关键的一步。决策树分类算法在智能导学系统中应用的优点主要有两个方面:(1)能够对原有的数据做出分析,以便产生精确的分类;(2)决策树方法产生的决策树或产生式规则集具有结构简单直观,容易理解,以及计算效率高的特点。由于属性和其取值情况众多,学习的人数也非常多,所以在计算时也会产生很大的数据量。但据其分类效果以及智能导学系统之后的过程而言,它还是非常高效的,而且作为智能导学系统的第一步,决策树算法的应用也体现出了智能导学系统的智能性。此外现在也已出现多种对决策树算法的改进,能进一步缩小计算量,提高效果。综上所述,将决策树算法应用在智能导学系统中是十分可行的。

参考文献:

[1] 荆永君,钟绍春,程晓春,等.智能导学系统设计[J].广西师范大学学报:自然科学版,2004,22(3):19-23.

[2] 董 贺,荣光怡.数据挖掘中数据分类算法的比较分析[J].吉林师范大学学报:自然科学版,2008,29(4):107-108.

[3] Mitchell T M.机器学习[M].张银奎,曾华军,译.北京:机械工业出版社,2006.

[4] 麦 青.浅谈数据挖掘中的决策树算法[J].福建电脑,2008,24(11):58-59.

[5] Shen Shi-kai, Hong Wang Wu, Sun-yan. A study of the employment of higher institutions based on the decision tree model[J]. 通讯和计算机,2008,5(10):28-32.

[6] 方 勇,戚飞虎. A new decision tree learning algorithm[J]. 哈尔滨工业大学学报:英文版,2005,12(6):684-689.

[7] 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):164-165.