基于自主学习环境下使用 C4.5 算法的专业预测①

The Degree Forcast using C4.5 Algorithm based Independent Studying

刘 敏 滕 华 毛嘉莉 董 文 (西华师范大学 计算机学院 四川南充 637002)

摘 要: 网络环境下的自主学习模式存在学习者所修的课程杂乱无章、没有明显的专业信息、学习盲目等诸多问题,人为地对自主学习行为进行指导和规范并不现实。利用面向属性归纳的决策树 C4.5 算法对自主学习信息进行分析,事先预测出自主学习者的专业取向,不仅能够帮助自主学习者有效地学习,而且能够实现网络教育中对自主学习者的有效管理,为学位决策提供可靠的依据。

关键词: 学位决策 自主学习模式 智能导学 数据挖掘 C4.5 算法 面向属性归纳

近几年来,随着我国高校本科教学改革的全面深入,先进的信息技术和在全国各高等院校中广泛实行的学分制,以及本科教学质量与教学改革工程[1]的进一步实施为自主学习模式的形成提供了孕育的摇篮,使得为学习者提供随时随地的个性化教育和终身教育的自主学习模式成为了网络教育发展的主要方向。在这种基于网络环境的自主学习模式下,学习者可以根据自己的兴趣广泛学习自己喜欢的课程,但是这样会造成学习者所修的课程杂乱无章、没有明显的专业信息、学习盲目等诸多问题。为确保学习者完整系统地学习,从而顺利完成学业,也为了实现有效的网络教育管理,为决策者提供可信的决策依据,智能导学[2]势在必行。

智能导学就是运用专家系统和数据挖掘技术为自主学习者提供实时的学习指导,其中预测出自主学习者的专业取向、规范其学习方向是解决问题的关键[3]。为了得到自主学习者与专业取向之间的映射关系,需要总结出一系列由自主学习者到专业取向的映射规则。由于学习者人数众多,相关的学习信息庞大繁多,依靠人工找到相关规则和进行分类是不可能的,因此需要用数据挖掘的方法对上述映射规则进行提取并利用这些映射规则对正在学习的学习者进行分类。

1 数据来源、预处理与实验环境

本文实测数据是西华师范大学计算机学院 2002、2003 和 2004 级计算机专业与通信专业学生的学习数据,数据来源包括 SQL Server 和 Access 数据库、Excel 表格和为学位预测模块设计的调查问卷表格。这些数据来源于不同的数据库系统中,主要用于日校生的学绩管理,往往不适合直接挖掘,需要做数据预处理的工作。数据预处理工作准备是否充分,对于挖掘算法的效率乃至正确性都有关键性的影响。数据预处理方法一般包括数据的选择(选择相关的数据)、净化(消除冗余数据)、转换、归约等。

在上述原始数据中,存在很多的属性,如学号、班级、姓名、性别、课程学分、每门课及成绩、授课时间、专业名称等等,数据准备时必须去掉那些不相关或弱相关属性,可以通过采用面向属性的归纳(AOI)等方法去掉这些不必要的属性。

1.1 属性删除

将有大量不同取值且无概化操作符的属性或者可用其它属性来代替它的较高层概念的那些属性删除。比如学生姓名、授课时间等,它们的取值太多且无法在该取值域内找到概化操作符,所以应将其删除。

1.2 相关分析

相关分析的好处在于:一方面能够减少输入变量之间的冗余度,从而保证计算的效率和输出的简捷;另一方面为了避免与输出变量无关的输入延误甚至误导挖掘进程,必须要保证输入变量与输出变量之间有一

①科研启动项目:决策树算法在学位决策系统中的应用研究(06B012)

定的相关度。对属性进行相关分析的方法有相关系数法、统计检验法等。对于有些属性，可以根据逻辑上直观地判断决定取舍，例如学号、性别、班级等属性对学位专业无关，而课程成绩与课程学分两个属性高度相关(相关系数为 0.9524)，因此去掉学号、性别、班级、课程学分等属性。

1.3 属性概化

用属性概化阈值控制技术沿属性概念分层上卷或下钻进行概化。如：按 100 分为满分计，就有 100 种分值；而按照该门课程是否结业，以 60 分为界则分为及格和不及格两类。考虑到有的学生修了该门课程，有的学生又没有修，于是把该课程成绩不及格的作为没有修该门课程的对待，把该课程成绩及格的作为修了该门课程的对待，这样就把课程作为了判断属性，每门课程对应的课程成绩就概化为“修”和“未修”，作为了属性取值。

上述原始数据的预处理工作均在计算机中进行。通过进一步对原始数据存在着不完整的、有噪声的和不一致的问题采用原始数据预处理，从中筛选出 713 条有效数据，建立了符合挖掘需要的数据库，其中每条数据记录包含了学生几年的修课情况和毕业时的学位专业名称。限于篇幅，这里只列出所有数据集中有代表性的 15 个数据，并且每个数据只是有代表性的选择了 10 个属性(表 1)。

表 1　预处理后的学习信息表(部分)

概率统计	数字电路	电磁场与波	汇编	数字信号	现代交换原理	离散数学	组成原理	操作系统	数据库原理	专业
修	修	未修	修	修	未修	修	未修	修	修	计算机
修	未修	修	修	未修	未修	修	修	修	修	计算机
修	修	修	修	修	修	未修	修	未修	未修	通信
修	未修	修	修	修	修	未修	未修	未修	未修	通信
修	修	修	未修	修	修	未修	未修	未修	修	通信
未修	修	未修	修	未修	未修	未修	修	修	修	计算机
修	修	未修	未修	未修	未修	修	修	修	修	计算机
修	修	未修	修	未修	未修	修	修	修	修	计算机
修	未修	修	修	修	修	未修	未修	修	未修	通信
修	修	修	修	修	修	修	未修	未修	未修	通信
修	修	未修	修	未修	未修	修	修	修	修	计算机
修	修	修	修	修	修	修	修	修	修	计算机
修	未修	未修	修	修	修	修	未修	未修	修	通信
未修	修	未修	修	修	未修	修	修	修	修	计算机
修	修	未修	修	未修	未修	修	修	修	修	计算机

显然，前期的数据预处理工作提高了数据的质量，进而提高了挖掘结果的质量。

本文的实验环境为实验室局域网，系统采用 J2EE 技术，Web 服务器基于 Windows 2003 Server，数据库服务器用 SQL Server 2000 实现。系统的测试工作在一台工作站(神舟承运 L720T)上完成，其操作系统为 Windows XP2。

2 算法选择

决策树分类方法(Decision Tree)是利用最广泛的分类技术。决策树，又称判定树，是一种类似二叉树或多叉树的树结构，树中的每个非叶节点(包括根节点)对应于训练样本集中一个属性上的测试，非叶节点的每一个分枝对应属性的一个测试结果，每个叶子节点则代表划分的一个类或类分布。最顶端节点为根节点，从根节点到叶子节点的一条路径形成一条分类规则。决策树可以很方便地转化为分类规则，是一种非常直观的分类模式表示形式。

要进行决策树分类，首先必须构造一棵决策树。决策树构造的 C4.5 算法是 J. R. Quinlan 于 1993 年提出的一种对 ID3 的改进算法。C4.5 算法克服了 ID3 在应用中的一些不足，表现在如下四个方面：(1) C4.5 算法引入了增益率(Gain Ratio)来克服 ID3 算法中的最大增益(Gain)偏向于多值属性的特点；(2) C4.5 算法使用一种悲观估计来补偿由于 ID3 算法对分类器的准确率的乐观估计造成的偏差，即 C4.5 算法使用一组独立于训练样本的测试样本来评估准确性，而不像 ID3 算法使用训练样本估计每个规则的准确率；(3) C4.5 算法在 ID3 的基础上加进了对连续型属性，属性值空缺情况的处理；(4) 对树剪枝也有了较成熟的方法，在决策树的构造过程中或者构造完成之后对决策树进行剪枝[4]。

由于本文的数据具有以下特点：首先，具有分类预知性的特点，已知每一组数据对应的分类是某个学位专业名称；其次，本文的主要任务是学习数据中的分类规律；再次，根据挖掘任务的分类性，应该选取受记录字段影响最小、模型易理解性、模型易训练性、模型易实施性、通用性、有用性最高的决策树算法进行挖掘，而 C4.5 算法是一种很有效的决策树算法，它能够对不完整的数据进行处理。因此，C4.5 算法作为了本工作数据挖掘的首选算法，实现了智能导学中对学生几年来所有已结业课程的专业信息的归纳分析，从中提取出

最可能的授位专业信息,从而进行学位预测的功能。

3　用 C4.5 算法生成学位预测决策树

在数据预处理后用确定的决策树算法生成决策树,C4.5 算法的主要处理过程如下:

(1) 计算学生学位分类所需的总信息熵

设 S 为 s 个样本的训练样本集,共有 m 类样本 C_i $(i=1,\cdots,m)$,s_i 为类 C_i 中的样本数,计算公式为:

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

其中 p_i 是任意样本属于 C_i 的概率,可用 s_i/s 来估计。在本例中,计算机学院学生的专业分为"计算机"和"通信"两类,故 m=2。

(2) 计算每个属性的信息熵

设属性 X 具有 v 个不同的取值 $\{x_1,x_2,\cdots,x_v\}$,将 S 划分为 v 个子集 $\{S_1,S_2,\cdots,S_v\}$,其中 S_i 包含 S 中这样一些样本,它们在 X 上具有值 $x_i(i=1,2,\cdots v)$。如果选择 X 作为测试属性,则这些子集就是从代表样本集 S 的节点生长出来的新的叶节点。设 s_{ij} 是子集 S_i 中类别为 C_i 的样本数,则根据 X 划分样本的信息熵值为:

$$E(X)=\sum_{i=1}^{v}\frac{s_{1j}+\cdots+s_{mj}}{s}I(s_{1j},\cdots,s_{mj})$$

其中 $I(s_{1j},s_{2j},\cdots,s_{mj})=-\sum_{i=1}^{m}P_{ij}\log_2(P_{ij})$;$P_{ij}=\frac{s_{ij}}{|S_i|}$ 是 S_i 中类为 C_i 的样本的概率。本文的数据中,课程是分类属性,每门课程即每个属性的取值均为"修"和"未修",因此 v=2。

(3) 计算该属性的信息增益和信息增益率

用属性 X 划分样本集 S 后所得的信息增益值为:

$$Gain(X)=I(s_1,s_2,\cdots,s_m)-E(X)$$

信息增益函数对于那些可能产生多分枝的测试倾向于生产大的函数值,但是输出分枝多,并不表示该测试对未知的对象具有更好的预测效果,信息增益率函数可以弥补这个缺陷。

"信息增益率"是 C4.5 算法为了去除多分枝属性的影响而对 ID3 算法中信息增益的一种重要改进。使用"信息增益率函数",即考虑了每一次划分所产生的子结点的个数,又考虑了每个子结点的大小(包含的数据实例的个数),从而就使得考虑的对象主要是一个个地划分,而不再考虑分类所蕴涵的信息量。理论和实验表明,采用信息增益率比采用信息增益更好,有效地克服了 ID3 方法选择偏向取值多的属性。属性的信息增益率函数为:

$$A(X)=\frac{Gain(X)}{I(s_1,s_2,\cdots,s_v)}$$

其中 v 为该节点的分枝数,s_i 为第 i 个分枝下的记录个数。

(4) 归纳决策树

依次计算每个属性的信息增益 Gain(X) 以及信息增益率 A(X),选取信息增益率最大的,但同时获取的信息增益又不低于所有属性平均值的属性作为测试属性,以该属性作为结点属性的每一个分布引出一个分枝,据此划分样本。要是节点中所有样本都在同一个类,则该节点成为树叶,以该学位专业标记该树叶。如此类推,直到子集中的数据记录在主属性上取值都相同,或没有属性可再供划分使用,递归地形成初始决策树。另外,在节点处记下符合条件的统计数据:该分枝总数、计算机专业个数和通信专业个数。

之所以选取信息增益率大而信息增益不低于平均值的属性,是因为高信息增益率保证了高分枝属性不会被选取,从而决策树的树型不会因某节点分枝太多而过于松散。过多的分枝会使得决策树过分地依赖某一属性,而信息增益不低于平均值保证了该属性的信息量,使得有利于分类的属性更早地出现。

4　结果分析与实验结论

应用中算法通过两次处理,一次是训练过程,导出分类算法,一次是测试过程,即得出正确识别率。C4.5 算法从经过预处理的 713 条有效数据中随机选取 2/3 左右即 475 条作为训练数据,其余 238 条数据作为测试数据。通过运用 C4.5 算法能够得到专业分类决策树,Quinlan 的 C4.5 算法以文本形式输出,也可用文献[5]介绍的方法以树型图示方法展示出来。限于篇幅,在此不列出完整的决策树模型。

用该模型生成的规则来预测测试数据集中的未知数据属于哪一分类,例如在学位决策系统中针对某同学前几年的修课情况用该模型预测其专业方向如图 1。

实验中使用测试样本 238 条,其中计算机专业样本 151 条,通信专业样本 87 条,测试结果的正确识别率为 89.253%。从实验结果来看,决策树模型虽然显示了一个很不平衡的结构,但得出了很容易理解的决策

规则。通过分析决策规则我们发现，公共选修课都不会起到决策作用，这和人的主观判断完全一致。同时，道路。

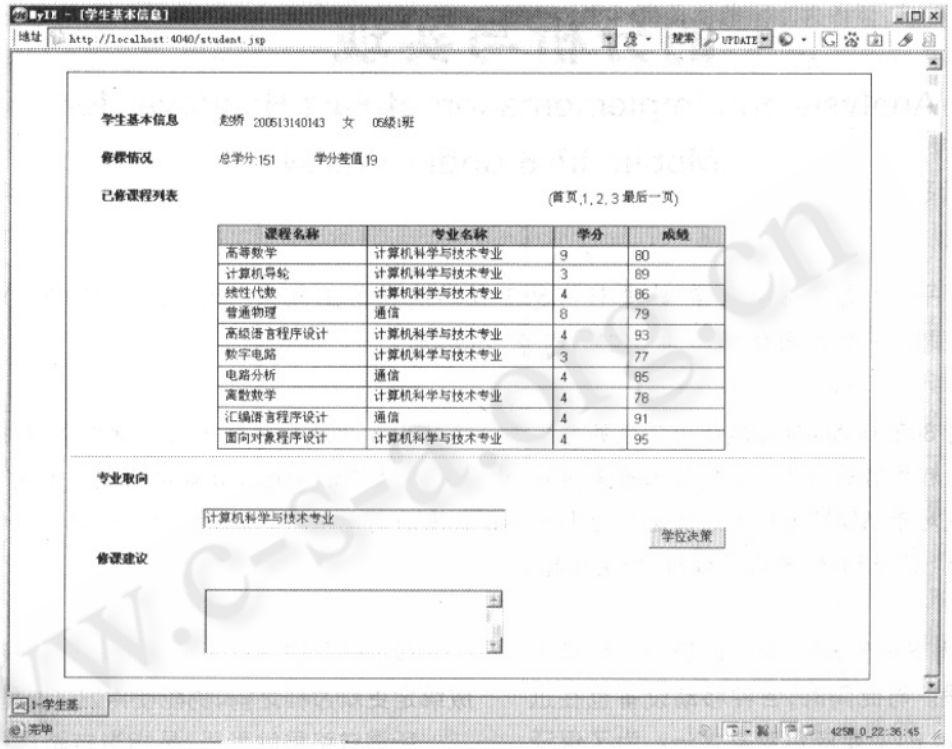

图 1　学位决策系统中的专业预测结果

实验结果还较好的体现了各门专业课在本专业课程体系中的权重。由此可见，该实验说明在智能导学中用 C4.5 算法进行专业取向的预测是有效可行的，其结果能够规范自主学习者的学习方向，并作为管理决策者的决策依据。

5　结束语

基于网络环境下的自主学习模式是网络教育发展的新阶段，对自主学习进行指导的智能导学涉及到了人工智能、数据挖掘和教育理论等多方面的知识，使用数据挖掘技术对自主学习的专业取向进行预测只是智能导学付诸实践的一个大胆尝试。就目前而言，把数据挖掘技术用于网络教育的工作才刚刚开始，相信沿着这条多学科研究道路走下去，最终将深化对自主学习者学习规律的认识，为自主学习者提供良好的学习指导、规范其学习行为，提高学校的管理及决策能力，为形成和谐的基于下一代互联网的自主学习模式铺平

参考文献

1　“质量工程”全面启动，中国大学教学 2007,2:4.
2　荆永君等，基于 Internet 的智能导学系统设计，中国教育网络，总第 4 期，2005,1 ~ 2:44 - 47.
3　刘敏、滕华、董文、谯石，基于 IPv6 校园网的智能选课系统的 J2EE 架构的设计与实现，数据通信，2006,3:56 - 58.
4　Quinlan, Ross. J. C4.5: Programs for machine learning [M], San Mateo, CA: Morgan Kaufmann Publishers, 1993.
5　姜欣、徐六通、张雷，C4.5 决策树展示算法的设计[J]，计算机工程与应用，2003,28(4):93 - 95.