数据挖掘技术在数字化校园中的应用研究

Application Study of Data Mining Technology in Digital Campus

张治斌　王艳萍　（河南理工大学 计算机科学与技术学院 河南 焦作 454000）

摘要：在讨论了数据挖掘技术的基本概念、决策树方法的基础上，提出了决策树算法在数字化校园中的应用，以高校学生等级的划分为例介绍了该算法的实施过程，并对结果进行了分析，得出供高校管理者决策的结论。

关键词：数据挖掘 决策树算法 数字化校园 学生等级

1 引言

数字化校园是以数字化信息为依托，利用计算机技术、网络技术、通讯技术支持学校教学和管理信息流，实现教育、教学、科研、管理、技术服务等信息收集、处理、整合、存储、传输、应用，使教学资源得到充分优化利用的一种虚拟教育环境[1]。数字化校园建设已经成为现代高校建设的重要组成部分，如何更好地利用数字化校园信息，提高高校教学效率，从而为社会培养出更多高素质人才，是一个值得研究的问题。数字化校园是面向教师和学生的，并为教师和学生服务。利用数据挖掘技术，在了解学生的各个方面信息的基础上，通过决策树算法得到学生学习成绩的总体发展趋势，为高校教学提供决策支持作用。

2 数据挖掘技术

2.1 数据挖掘的基本概念

数据挖掘（Data Mining，DM）是从大量的、不完全的、有噪声的、模糊的、随机的实际数据中，提取隐含在其中的、人们不知道的，但又是潜在有用的信息和知识的过程[2]。目的是发现未知的关系和以数据拥有者可以理解并对其有价值的新颖方式来总结数据，进而预测未来可能发生的行为，从而为决策行为提供有利的支持。

2.2 决策树方法

决策树方法是数据挖掘的核心技术算法之一，他通过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，常用于预测模型中。目前，国际上最有影响力的决策树方法是 ID3 决策树生成算法，C4.5 算法是 ID3 算法的改进，该算法主要采用信息增益比来确定被测试的属性[3]。

决策树（decision tree）是一个类似于流程图的树结构，其中每个内部节点表示在一个属性上的测试，每个分支代表一个测试输出，而每个树叶节点代表类或类分布。树的最顶层节点是根节点。通常情况下，采用自顶向下递归的各个击破的方式构造决策树，在此过程中，选择合适的属性作为测试属性；采用剪枝方法控制生成的决策树的大小；是两个关键的问题。

决策树的基本算法是贪心算法，它以自顶向下递归的各个击破方式构造决策树，算法 Generate_decision_tree 生成一棵决策树的基本步骤[2]：

输入：训练样本 samples，由决策属性表示，侯选属性的集合 attribute_list。

输出：一棵决策树。

（1）创建节点 N。

（2）if samples 都在同一个类 C then

（3）返回 N 作为叶节点，以类 C 标记；

（4）if attribute_list 为空，以类 C 标记。

（5）返回 N 作为叶节点，标记为 samples 中最普通的类；//多数表决

（6）选择 attribute_list 中具有最高信息增益的属性 test_attribute；

（7）标记节点 N 为 test_attribute；

(8) for each test_attribute 中的已知值 ai //划分 samples

(9) 由节点 N 长出一个条件为 test_attribute = ai 的分支;

(10) 设 si 是 samples 中 test_attribute = ai 的样本的集合;//一个划分

(11) if s_i 为空 then

(12) 加上一个树叶,标记为 samples 中最普通的类;

(13) else 加上一个由 Generate_decision_tree 返回的节点;

以上递归步骤当下列条件成立时停止:

(1) 给定节点的所有样本属于同一类。

(2) 没有剩余属性可以用来进一步划分样本。在此情况下,使用多数表决。

(3) 分支 test_attribute = ai 没有样本。在这种情况下,以 samples 中的多数类创建一个树叶。

3 数字化校园整体框架

基于当前高等院校校园网的基本设施和已有的各种应用服务,一个基于通用的统一身份认证和统一信息展示的数字化校园解决方案的总体框架。这个框架能够集成各种校园网中的应用[1]。各个子系统在数字化校园中的位置如图 1 所示。

在这个数字化校园框架中,利用 PKI 体系结构作为统一身份认证系统的基础,以 LDAP 目录作为校园网内各种身份和信息数据的存储每体,从而实现 Portal 信息展示平台,为校园网内各种应用服务的集成与展现提供了途径。

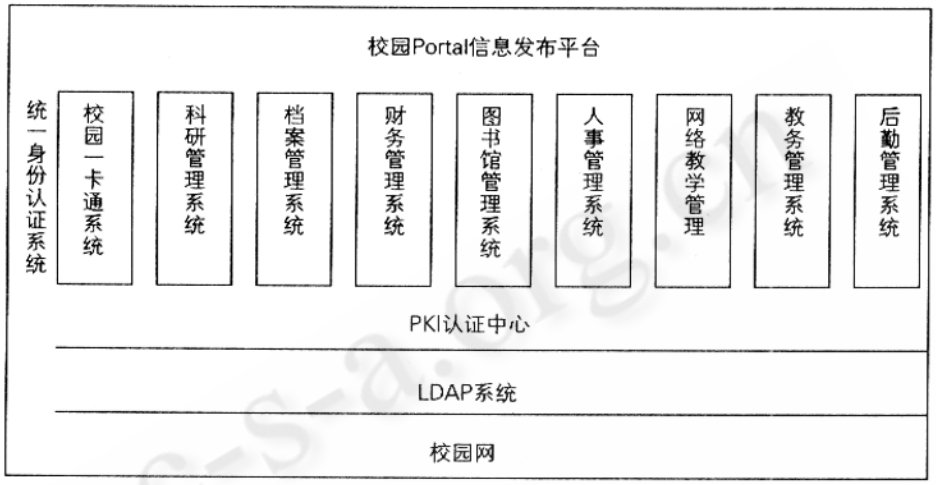

图 1　数字化校园的整体框架

4 数据挖掘技术在数字化校园中应用

数据挖掘过程主要经历以下主要阶段:确定数据挖掘对象、数据准备等。下面将结合数字化校园介绍数据挖掘关键过程的应用。

4.1 确定数据挖掘对象

定义清晰的挖掘对象,认清数据挖掘的目标是数据挖掘的第一步。在数字化校园信息库中,主要的信息就是教师和学生,如何更好地协调教师和学生的关系,更好地促进教育事业的发展,本文先从本科生着手,来研究本科生在校的基本情况,从而确定以学生为主体(以下简称学生)。

4.2 数据准备

收集和描述数据是整个数据挖掘工作中相当重要的一部分。数据准备一般包括两个步骤:数据的选择和数据的预处理。在这里主要是在校本科生的家庭出身情况、学习情况、每月消费情况、每月借书情况、社会工作情况。例如从校园一卡通系统中可以找到某个学生的这个月的消费情况。下面的挖掘方法并未对学生

信息的各个子库中所有数据进行直接挖掘，而是以学生的数字化校园中的基本信息作为基础信息，通过对学校的各个子库的个人信息进行加工处理，运用简单的统计方法对每个子库信息进行聚合[5]，从而得到进行数据挖掘的基本信息。

把从各个子库中得到我们想要的数据必须经过处理才能应用到数据挖掘技术中去。例如我们把学生通过文字所表现的不同属性进行量化，以便于算法分析，我们把学生分为：A、B、C、D、E 五个等级，即各个方面都表现优秀的学生为 A、中等靠上但次于优秀的为 B、中等生为 C、中等靠下为 D、各个方面都很差的为 E[6]。

依据以上量化标准，我们把统计得到用于数据样本的一个 6 维向量进行初步量化。

(1) 学生每月消费：超过 500 元的为高、500—300 元的为中、低于 300 的低。

(2) 图书馆平均每月借书(每月按图书馆开放 25 天计算，以下简称图书馆借书)：每月光顾图书馆 4 次以上为优，2—4 次为良，少于 2 次的为中。

(3) 专业课平均成绩：高于 85 分的为优，75—85 之间的为良，60—75 之间的为中。

(4) 参加社会活动情况：1 表示经常参加社会活动，0.5 表示参加社会活动适度，0 表示基本上不参加社会活动。

(5) 家庭出身：农表示出身农民，工表示出身工人，干表示出身干部。

(6) 学生等级：各个方面都表现优秀的学生为 A，中等靠上但次于优秀的为 B，中等生为 C，中等靠下为 D，各个方面都很差的为 E。

(下面介绍一个训练样本，该数据样本选自 2003 级计算机专业某个班学号的前 15 名。)

4.3 构造决策树

根据选取训练样本数据集，取属性“学生等级”作为类别标识属性，属性“家庭出身”、“每月平均消费水平”、“专业课平均成绩”、“图书观借书”、“参加社会活动”作为属性集[4]。训练样本集类 A、B、C、D、E 所对应的样本个数记为 s_1、s_2、s_3、s_4、s_5。其中 $s_1=2, s_2=4, s_3=4, s_4=3, s_5=2$。

首先，对给定的样本分类所需的期望信息：

$I(s_1,s_2,s_3,s_4,s_5)=I(2,4,4,3,2)=-2/15\ \log_2(2/15)-4/15\ \log_2(4/15)-4/15\ \log_2(4/15)-3/15\ \log_2(3/15)-2/15\ \log_2(2/15)=1.9311$

下一步，计算每个决策属性的熵，从属性“家庭出身”开始，对每个分布计算期望信息。

对于家庭出身=“干”：$s_{11}=1, s_{21}=1, s_{31}=2, s_{41}=0, s_{51}=0$

$I(s_{11},s_{21},s_{31},s_{41},s_{51})=-1/4\ \log_2(1/4)-1/4\ \log_2(1/4)-2/4\ \log_2(2/4)=2.0068$

表 1　学生等级训练样本集

学号	家庭出身	每月平均消费水平	专业课平均成绩	图书馆借书	参加社会活动	学生等级
03020501	干	高	中	中	1	C
03020502	干	高	良	中	0.5	C
03020503	干	低	良	良	0.5	B
03020504	农	低	优	优	0.5	A
03020505	农	低	良	良	0	B
03020506	工	高	良	中	0.5	C
03020507	农	低	中	中	0.5	D
03020508	工	高	中	中	0	E
03020509	工	中	中	中	0.5	D
03020510	农	低	良	中	0.5	C
03020511	农	低	良	中	1	B
03020512	农	中	中	中	0	E
03020513	干	中	优	良	1	A
03020514	工	高	中	中	0	D
03020515	农	低	良	良	1	B

对于家庭出身=“工”：$s_{12}=0, s_{22}=0, s_{32}=1, s_{42}=2, s_{52}=1$

$I(s_{12},s_{22},s_{32},s_{42},s_{52})=-1/4\ \log_2(1/4)-1/4\ \log_2(1/4)-2/4\ \log_2(2/4)=2.0068$

对于家庭出身=“农”：$s_{13}=1, s_{23}=3, s_{33}=1, s_{43}=1, s_{53}=1$

$I(s_{13},s_{23},s_{33},s_{43},s_{53})=-1/7\ \log_2(1/7)-3/7\ \log_2(3/7)-[1/7\ \log_2(1/7)]\times 3=1.6697$

如果样本按家庭出身划分，对这个给定的样本分类所需的期望信息为：

$E(\text{家庭出身})=4/15\ I(s_{11},s_{21},s_{31},s_{41},s_{51})+4/15\ I(s_{12},s_{22},s_{32},s_{42},s_{52})+7/15\ I(s_{13},s_{23},s_{33},s_{43},s_{53})=1.8494$

因此，这种划分的信息增益是：

$\text{Gain}(\text{家庭出身})=I(s_1,s_2,s_3,s_4,s_5)-E(\text{家庭出身})=0.0817$

类似地，我们可以计算 Gain(每月平均消费水平)=0.4076，Gain(专业课平均成绩)=1.2668，Gain(图书馆借书)=0.6963，Gain(参加社会活动)=0.266，由于专业课平均成绩在 属性中具有最高信息增益，它被选作测试属性。创建一个节点，用专业课平均成绩标记，并对于每个属性值，引出一个分支。样本据此划分，重复上述步骤，最后返回的最终判定树如图 2 所示。

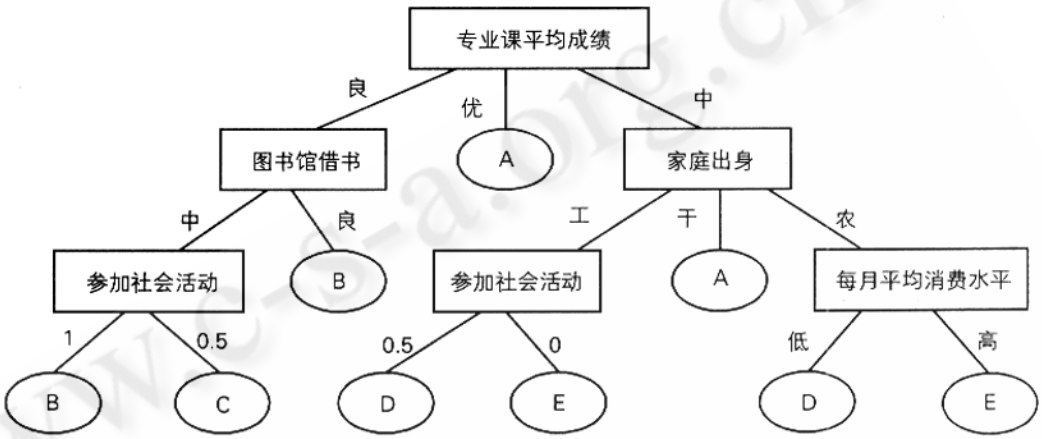

图 2 学生等级决策树

4.4 结果分析

比较以上 5 个属性的信息增益，得到学生等级决策树，我们可以得到以下结论：

(1) 可以看出专业课水平的高低是决定学生等级的关键因素。

(2) 图书馆借书次数较多，可以看出学生比较重视学习。

(3) 参加社会活动积极的学生，也是相对较好的学生。

(4) 每月消费较高的部分同学比较侧重于学习之外的别的方面，所以这些学生是较差的。

(5) 并不是来自家庭贫困的学生都是好学生，也不是来自家庭富裕的学生都是差学生，虽说大学生关键是靠个人的努力，学校的管理和督促对那一部分消费比较高的学生来说还是能起到一定的作用的。

5 结论

在目前数字化校园建设日益兴起的环境下，本文根据数字化校园系统中所存储的学生信息，利用数据挖掘技术的决策树方法分析了影响学生等级的重要因素，这只是数据挖掘技术在数字化校园系统中一个简单的应用。如何充分地利用高校资源，把数据挖掘技术和数字化校园更好地结合起来是当前高校面临的一个很重要的现实问题，从而达到提高教学质量和提高大学生素质的目的。

参考文献

1 陆炯，数字化校园的总体框架与若干关键技术的研究[D]，南京大学：南京大学，2004.

2 Jiawei Han, Micheline Kamber，数据挖掘：概念与技术[M]，北京：机械工业出版社，2001.188－194.

3 陈文伟、黄金才，数据仓库与数据挖掘[M]，北京：人民邮电出版社，2004.13－123.

4 雷松泽、郝燕，基于决策树的就业数据挖掘[J]，西安工业学院学报，2005，25(5)：429－432.

5 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.

6 丁智斌、袁方、董贺伟，数据挖掘在高校学生成绩分析中的应用[J]，计算机工程与设计，2006，27(4)：590－592.