Moodle二次开发研究之数据挖掘*

霍　静

（天水师范学院物理与信息科学学院　　甘肃天水，741001）

摘　要：Moodle　开源教学平台在我国教育领域得到广泛应用，为推动Moodle平台更深层次应用，以Moodle平台Feedback板块为例，结合学生C语言考试成绩，探讨Moodle平台数据的挖掘需求以及相应的挖掘方法。通过建立决策树、将数据规则可视化，探讨影响学生计算机成绩优良的因素，为优化学习、教学提供了科学参考。

关键词：数据挖掘；Moodle；c4.5

Abstract: Moodle, an open source Elearning platform, has found broad application in China's education sector. To bring the application of Moodle platform to a higher level, the paper discusses Moodle platform's data mining needs and relevant data mining methods by taking the Feedback module of Moodle platform for example and combining students' scores of C Programming Language examination. It also discusses factors influencing students' grades on computer courses through the creation of decision tree and the visualization of data rules, which provides a scientific guidance for optimizing learning and teaching.

Key words: Data mining ; Moodle; C4.5

中图分类号：TP368.1　　**文献标识码：**A　　**文章编号：**1001-9227（2014）01-0125-03

0　引　言

*Moodle*平台是基于建构主义教育理论而开发的课程管理系统，是一个免费的开放源代码的软件[1]。截止*2013*年*9*月，笔者在*CNKI*中以“*Moodle*”作为检索词查找发表期刊，返回记录*1032*条。检索结果几乎全是介绍*Moodle*平台使用经验、或基于*Moodle*平台的教学应用研究报告，检索界面如图*1*所示：

图1

笔者再次以“*Moodle*+挖掘”作为检索词查找发表期刊，返回记录仅*2*条，检索界面如图*2*所示。

图2

收稿日期：2013-11-05

*基金项目：天水师范学院校级项目“基于Moodle的泛在学习环境的二次开发（TSB1114）阶段性成果

作者简介：霍静（1978-），女，讲师，硕士，主要研究方向为计算机网络。

可见国内已有的众多*Moodle*平台教学应用研究缺少对*Moodle*平台记录的各种信息的深入挖掘。因此本文以*Moodle*平台中*Feedback*板块为例，结合学生*C*语言考试成绩，探讨*Moodle*平台数据的挖掘需求以及相应的挖掘方法，推动*Moodle*平台深层次的应用。

1　开发环境搭建

Moodle 1.9（支持*Windows　OS*）+*EXCEL2003*+*Feedback*模块。

1.1　挖掘算法

数据挖掘就是从大量的随机实际应用数据中，提取隐含在其中的、有用的信息和知识的过程[2]。分类问题是数据挖掘领域的一个分支,通过分类,可以确定对象属于哪个预定义的目标类[3]。分类算法中决策树算法是解决分类问题的最有效方法[4]。决策树算法中*C4.5*算法应用较多[5]。

设*S*是训练样本集,有*m*类样本集并且有*s*个训练样本，则对一个给定的样本分类所需期望信息：

$$I(S_1,S_2,\ldots,S_m)=-\sum_{i=1}^{m}\frac{S_i}{S}\log_2\frac{S_i}{S} \qquad (1)$$

属性*A*的值$\{a_1,a_2,\ldots,a_v\}$的将*S*划分为子集$\{S_1,S_2,\ldots,S_v\}$,设S_j包含类C_i的S_{ij}个对象。属性*A*划分的熵：

$$E(A)=\sum_{j=1}^{v}\frac{S_{1j}+\ldots+S_{mj}}{S}I(S_{1j}+\ldots+S_{mj}) \qquad (2)$$

属性*A*划分的信息增益为:

$$Gain(A)=I(S_1,S_2,\ldots,S_m)-E(A) \qquad (3)$$

属性*A*分割训练集*S*的信息量为:

$$SplitInfo(A,S)=I\left(\frac{|S_1|}{|S|},\frac{|S_2|}{|S|},\ldots,\frac{|S_y|}{|S|}\right) \qquad (4)$$

其中{S1,S2,...,Sy}是根据属性A的取值分割S后产生的所有子集。

属性A的信息增益率为:

$$GationRation(A,S)=\frac{Gain(A)}{SplitInfo(A,S)} \quad (5)$$

由上可知，C4.5算法对给定的训练集S中所有非类别属性计算信息增益率,选择信息增益率最大的属性作为测试属性[6]。

1.2 分析数据准备

(1)数据收集

在feedback模块中设计调查问卷，学生填写提交后将数据导出为Excel文件。调查表有学生兴趣、教师教法、课堂效果、课外学习时间及调查学生学号五项。

(2)数据合成

Feedback导出的调查明细默认是将调查问卷中的每个问题做为一条记录,不符合按行(某学生)列(所有问题项集)来对问卷做数据分析的要求，因此将导出数据先做行列转置变换[7]，得到206条记录。此外，结合该门课程成绩表，以学号为关键字，将两个数据表合并成一个表。

(3)数据清洗

去掉表中填写不符合要求的作废调查记录，得到193条有效记录。

(4)数据转换

调查表中课外学习时间、成绩是连续值属性数据，由于采用C4.5分类算法，这里需将连续值属性数据离散化，具体标准为：课外学习时间小于一小时（差）、介于一小时至两小时（中）、大于两小时（优）；成绩大于等于60（通过）、成绩小于60（未通过）。

(5)数据减化[8]

在本例中，学号属性主要用来将调查表和成绩表数据合并，对成绩分析没有任何影响，将该属性从数据中去掉。更进一步，为了便于后期数据属性表示公式化，将问卷列属性学生兴趣、教师教法、课堂效果、课外学习时间、成绩分别用E1至E5表示；属性值优、中、差分别用A、B、C表示；课程成绩只有通过、未通过，分别用Y、N表示。

经过上述处理，得到如表1所示训练数据集（部分）。

表1

E1	E2	E3	E4	E5
B	A	B	B	Y
A	A	B	A	Y
B	B	A	B	N
B	B	B	C	N
B	A	A	B	Y
A	B	B	A	Y
B	A	A	A	Y
.....	.....	.....	.....	.....

1.3 建立决策树

给定样本分类的熵：

$I(s_1,s_2)=I(64,129)=-\log_2 47/193-\log_2 129/193=0.916297$

每个属性的信息增益：

属性E_1的熵:

$E_1=A$，$S_{11}=14$，$S_{21}=35$，$I(S_{11},S_{21})=0.896574$

$E_1=B$，$S_{12}=40$，$S_{22}=94$，$I(S_{12},S_{22})=0.961574$

$E_1=C$，$S_{13}=0$，$S_{23}=1$，$I(S_{13},S_{22})=0$

E1作为决策树第一个划分结点的熵:

$$E(E_1)=\frac{64}{193}I(S_{11},S_{21})+\frac{128}{193}I(S_{12},S_{22})+\frac{1}{193}I(S_{13},S_{23})=0.904657$$

划分后的信息增益是:

$Gain(E1)=I(S1,S2)-E(E1)=0.916297-0.904657=0.011640$

E1分割训练集对应的信息量是:

$$SplitInfo(E_1)=-\frac{64}{193}\log_2\frac{64}{193}-\frac{128}{193}\log_2\frac{128}{193}-\frac{1}{193}\log_2\frac{1}{193}=0.897657$$

E1分割训练集的信息增益率是:

$$GationRation(E_1,S)=\frac{Gain(E_1)}{SplitInfo(E_1)}=0.012967$$

同上，属性E2、E3、E4的信息增益率计算结果如图3所示。

$$GationRation(E_2,S)=\frac{Gain(E_2)}{SplitInfo(E_2)}=0.011577$$

$$GationRation(E_3,S)=\frac{Gain(E_3)}{SplitInfo(E_3)}=0.101299$$

$$GationRation(E_4,S)=\frac{Gain(E_4)}{SplitInfo(E_4)}=0.013356$$

通过比较四个计算结果值，其中E3的信息增益率最大,因此先将E3作为决策树根节点,根据属性的三个不同值建立三个分支,然后对每个分支的子集再次采用C4.5算法作进一步划分,直到分支划分结束，最后决策树如图3。

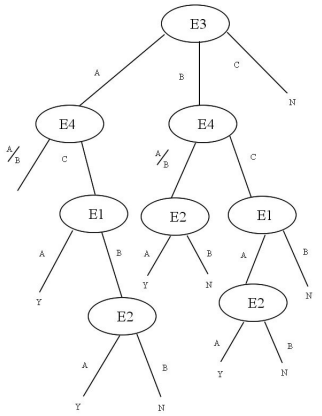

图3

1.4 规则

$If(E_3=A\wedge E_4=B)then E_5=$优

$If(E_3=A\wedge E_4=C\wedge E_1=A)then E_5=$优

$If(E_3=A\wedge E_4=C\wedge E_1=B\wedge E_2=A)then E_5=$优

$If(E_3=B\wedge E_4=B\wedge E_2=A)then E_5=$优

$If(E_3=B\wedge E_4=C\wedge E_1=A\wedge E_2=A)then E_5=$优

从规则中可以得出影响考试成绩优良与否的关键因素是课堂效果、课外学习时间及学生兴趣。说明了C语言课程一方面要重视理论学习同时课外学生自主学习也很重要，还要注意培养学生的学习兴趣。该结论为指导教师进一步提高学生考试成绩提供了理论依据和明确的指导方向。

2 结束语

本文以Moodle平台中Feedback板块为例，使用C4.5算

（下转第128页）

3 下位机 Flash 的写入流程

Flash 存储器修改数据是一个复杂的过程，所有的 *Flash* 存储器数据操作（读操作除外）都需要使用命令序列，命令序列由向 *Flash* 存储器地址范围些某些数据的操作组成。所有目标为该地址范围的数据转移操作都被解释为命令序列。

（*1*）*Flash* 擦除序列：

首先向 *0xc000AA* 地址写入 *0x80* 数据，然后向 *0xC00054* 地址写入 *0xAA* 数据，最后将目标 *Flash* 页面地址写入 *0x03* 数据，即可擦除一页 *Flash*，或者写入 *0x33* 则可以擦除一个扇区。擦除该页的 *Flash* 之后即可通过 *Flash* 编程将新数据写入到 *Flash* 中，周而复始的进行每一次的程序更新操作。

（*2*）*Flash* 编程序列*:*

Flash 页面模式：通过向 *0xC000AA* 的伪地址写入 *0x0050* 数据之后立即往目标 *Flash* 地址写入 *0x00AA* 可以让目标地址的那一页 *Flash* 进入页面模式。

Flash 写入缓冲区：*Flash* 内容是不可以直接写入的，我们只能将数据写入 *Flash* 写入缓冲区中，然后启动写命令，*Flash* 内部管理器会将缓冲区的数据通过一个内部加电序列写入到真正的 *Flash* 存储器中。

启动 *Flash* 写：通过向 *0xC000AA* 地址写入 *0Xa0* 数据然后向 *0xC0005A* 地址写入 *0xAA* 数据启动 *Flash* 写入。

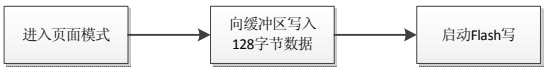

图3 流程图

4 下位机数据接收流程

通过 *RS485* 接口，上位机将编译好的程序目标代码下载到单片机中，下位机通信服务程序将上位机发送的程序代码通过校验无误之后，对代码包头部的命令字段进行分析，如果是数据命令则将程序代码写入到目标 *Flash* 区域中，并循环接收完整个新程序的所有代码。如果是修改启动地址命令则修改当前单片机启动地址，当所有代码都接收完毕之后，修改单片机当前运行的 *Flash* 区域地址，并重新启动单片机，完成程序的更新流程，由于所有的程序更新操作是基于单片机自身的，所以与选用的通讯格式或者协议都没有关系，也可以选择使用工业以太网或者是 *CAN* 网络等。

在整个程序更新的过程中所有的命令、地址、及数据都是由上位机发出，所以在整个写入流程具有很高的灵活性和可定制性。

5 总 结

通过在系统工作过程中接收程序升级数据，充分利用了 *XC2267* 的内部 *FLASH* 空间及现场具备的现场总线技术，对系统的实时工作影响小，仅需要在工作间隙进行数据发送和接收工作，如果在整个程序的工作过程中由于种种原因而造成数据错误而中断整个流程，但是由于没有进入到修改程序启动地址的最后一步，那么先前所有写入的数据无法产生实际作用，也不会对当前的仪表工作产生任何不利的影响，那怕是升级过程中断电也不会产生任何不利后果，将升级过程的风险降到了最低。

参考文献

[1] 胡汉才.单片机原理及其接口技术(第3版)[M].北京：清华大学出版社,2010,05.

[2] 肖 看,李群芳.单片机原理、接口及应用:嵌入式系统技术基础(第2版)[M].北京：清华大学出版社,2010,09.

[3] XC2267 16位单片机微控制器用户手册.英飞凌公司,2004.

（上接第126页）

法，建立成绩分析决策树模型，探讨了影响学生成绩优良的因素，为教师侧重改进哪些教学环节提供理论依据，也为针对 *Moodle* 平台的数据挖掘提供了一个范例。但是，文中调查方案设计仅有五个属性，是否还有其它没有考虑的因素是今后需要考虑的问题。此外，*C4.5* 算法的不足乃至算法的改进也是今后应用要研究解决的课题。

参考文献

[1] http://docs.moodle.org/dev/.

[2] http://wiki.mbalib.com/wiki/数据挖掘.

[3] 孙 娟.王熙照.规则简化与模糊决策树剪枝的比较[J].计算机工程,2006,6:210.

[4] Wang Xiao Hua.An Automatic Fuzzy Text Classification Based on Statistical Word[C].The 6th International Conference for Young Computer Scientists, HangZhou,2001.

[5] 李一平,姚宏亮.C4.5算法在成绩分析中的应用[J].计算机工程应用技术, 2011,6:51.

[6] 吕瑞雪.基于决策树的中学生成绩挖掘与分析[D].呼和浩特:内蒙古大学,2010.

[7] 宋 赟.基于数据挖掘技术的学生成绩分析[D].重庆:重庆师范大学,2009.

[8] 李一平,姚宏亮.C4.5算法在成绩分析中的应用[J].计算机工程应用技术,2011,6:51.