大数据时代网络教育学习成绩预测的研究与实现

——以本科公共课程统考英语为例

孙 力 程玉霞

(江南大学 人文学院,江苏无锡 214122)

[摘要] 合适的数据分析技术能使我们借助网络学历教育学生在学习和管理系统中产生的数据和信息,发现相关规律,进而为网络学历教育教学和管理流程的优化提供有益的决策依据。本文采用数据挖掘中数据分类 C5.0 决策树方法,通过分析网络学历教育本科学生英语学习及相关信息,实现了对其英语统考成绩的预测。在分析英语统考前景预测的目标特性后,在 SPSS 的 Clementine 12.0 数据挖掘环境中,历经数据提取、数据预处理、决策树构建和决策树优化等步骤,本研究构建了网络教育本科英语统考成绩的预测模型,并提出了模型实现方法;同时对模型相关属性的重要性进行了分析,提出了提高网络教育本科学生英语学习水平和统考通过率的相应策略。

[关键词] 网络教育;数据挖掘;决策树方法;英语统考;预测模型

[中图分类号] G434 [文献标识码] A [文章编号] 1007-2179(2015)03-0074-07

一、引言

近年来,大数据的概念逐渐兴起,人们用它来描述和定义信息爆炸时代产生的海量数据及与之相关的技术发展与创新(黄荷,2012)。大数据带来的机遇是可以利用数据分析技术预测未来。数据挖掘是从大量不完全、有噪声、模糊、随机的数据中,提取隐含在其中、人们事先不知道但又潜在有用的信息和知识的过程(洪建峰,2013)。它是一种深层次的数据分析方法,主要依靠人工智能、机器学习和统计学技术,对数据进行归纳推理,从中挖掘出潜在的模式,预测未来趋势,为决策提供支持。

我国网络高等学历教育经过十五年的发展,各试点高校对网络学习相关系统进行了完善,尤其是管理、学习、评价和监控系统的运用,产生了大量学生学习过程的相关数据。例如,学生个人信息,课程考试成绩,在线学习次数、时间点、学习兴趣点和学习时长,作业完成情况,参与讨论情况,过程性评价等。目前这些数据仅存在于各类网络学习和管理系统中,未真正发挥应有的作用。如能运用数据挖掘技术和学习分析技术,建立相关系统对其进行提取和标准化处理,可以为网络学习流程和管理流程的优化设计,提供相关决策依据(魏顺平,2013);同时,可以了解学生学习的效率、意愿、耐心度和专心度、相关兴趣点等个性化信息,为学生提供网络学习的个性化提醒和指导,以及学习资源的个性化推送服务等(傅钢善等,2014;陈益均等,2013);最终通过建立以学习者数据为核心的学习支持系统,构建智能化网络学习环境。

网络教育的部分公共课程统考是本科层次学生毕业电子注册的必要条件。其中英语是网络教育学生的难点。各试点高校都在尝试采用各种方式提高英语统考的通过率。本文运用数据挖掘技术,以江南大学网络教育本科学生为研究对象,通过对学习平台中学生个人相关信息、入学测试成绩、入学后英语类课程及其他课程学习情况的数据进行分析,预

[收稿日期]2015-03-10 [修回日期]2015-04-26 [DOI 编码]10.13966/j.cnki.kfjyyj.2015.03.009

[作者简介]孙力,博士,江南大学人文学院教授,继续教育与网络教育学院副院长,研究方向:网络教育系统的构建及开发(lisun@jiangnan.edu.cn);程玉霞,江南大学人文学院硕士研究生。

测其网络统考英语课程的考试成绩。

二、数据挖掘技术理论描述

数据挖掘是利用模式识别、统计和数学的技术,从大量数据中筛选发现新的有意义的关系、模式、变化和主要结构的方法(陈文伟等,2004)。随着大数据时代的到来,它被越来越多地应用到人工智能、机器学习、市场分析、商务管理和决策支持等领域。数据挖掘由三个步骤组成:数据预处理阶段、模型设计阶段和数据分析阶段。分类和聚类技术是其中最有应用价值的两大技术。

(一)数据分类技术

数据分类指分析数据库中的一组对象,找出其共同属性,然后根据分类规则,把它们划分为预先设定好的不同类别。数据分类过程一般分两个部分:先是确定分类规则,也称为学习或训练过程,即先将训练样本数据集作为输入,依据数据集特征为每一类别建立分类规则或描述;然后通过更大量的测试数据集测试这些分类规则,以生成更恰当的分类规则,并依据最终的分类规则形成数据分类。目前分类方法包括基于决策树的分类,如 ID3 算法和 C4.5 算法;基于统计的分类,如贝叶斯分类算法;基于神经网络的分类,如后向传播算法;源自关联规则挖掘概念的分类和遗传算法等。

(二)数据聚类技术

数据聚类是根据在数据中发现的描述对象及其关系的信息,将数据对象分组而形成数据簇。最终目标是:簇内的对象相互之间相关,不同簇的对象之间不相关。簇内相似度越大,同时簇间差别越大,数据聚类效果越好。已有的分类方法包括划分方法、层次方法、基于密度的方法、基于网格的方法和基于模型的方法等。

总之,分类是按照某种标准给数据贴“标签”,再根据标签区分归类。聚类是事先没有“标签”而通过分析找出数据之间存在聚集性原因的过程(Kantardzic,2011)。分类适合类别或分类体系已经确定的场合;聚类适合不存在分类体系、类别数不确定的场合,一般作为某些应用的前端。

三、数据分析技术的选用

我们通过采集和分析本科生个人相关信息及在网络学习平台留下的学习信息,预测学生在网络统考英语课程的前景。由于统考合格是网络教育本科生毕业电子注册的必备条件,我们将预测的结果仅定义为“合格”与“不合格”。这属于数据挖掘的数据分类问题,本研究将采用基于决策树的分类方法。

决策树方法起源于概念学习系统,就是使用树的结构对数据进行分类,每个条件下的记录集就像一棵树的叶节点。根据字段数据取值的不同,可以对决策树进行分支,在决策树各个分支的子集中再重复建立分支和决策树各下层节点,形成一棵决策树。目前最有影响的决策树算法是 ID3 和 C4.5。

ID3 主要是选择运用信息最大属性的增益值来进行样本训练划分,其目的是能够使熵在分裂系统时达到最小,以此提高决策树算法的精确度和运算速度。ID3 算法的缺陷在于运用信息增益作为分裂属性的标准,在取值时会不自然地偏向于取值较多的属性,然而大部分情况下,这种属性不能提供更多有价值的信息。C4.5 是改进 ID3 形成的新算法,它能够同时处理连续值和离散值的属性。C4.5 选择测试的标准主要采用信息增益比,这在很大程度上弥补了 ID3 的不足。

C5.0 算法是 C4.5 算法的修订版(商业版),适用于处理大数据集,计算速度快,占用内存资源较少。C5.0 算法根据能够提供最大信息增益的字段划分样本,对第一次划分出来的子样本递归划分,直到不能再分为止,最后重新检查最底层的划分,去掉贡献不大的分支,得到最终模型。C5.0 可以产生两种模型:决策树和规则集。决策树由算法划分样本直接产生,每个叶子节点表示一个特定的训练数据子集,训练数据集中的每个样本只属一个叶子节点。也就是说,任何一个给定的样本通过决策树只能得到一个预测结果(Zhu et al.,2009)。

C5.0 决策树分类主要分为两个过程。首先是学习过程,就是通过对大量的训练数据集学习来构造决策树。第二步是利用构造的决策树进行分类,先利用测试数据集评估决策树分类的准确率,如果准确率可以接受,则将生成的决策树用于新的数据分类。本研究采用 C5.0 为数据挖掘的内核算法。

四、英语统考成绩预测的实现

依据上述讨论,本研究运用数据挖掘的数据分

类技术实现网络教育本科生英语统考成绩的预测,历经数据提取、数据预处理、决策树构建、决策树优化和预测实现等步骤(见图1)。

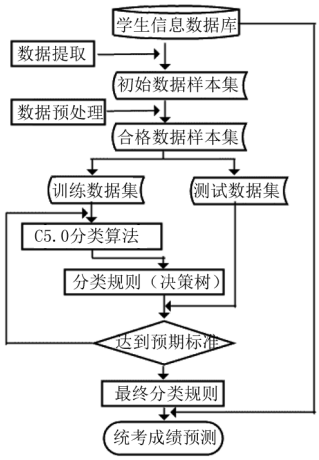

图1 英语统考成绩预测实现流程

(一)分类规则构建环境

Clementine 是 ISL(Integral Solutions Limited)公司开发的数据挖掘工具平台,SPSS 公司收购 ISL 后,对 Clementine 产品进行了重新整合和开发。目前 Clementine 已经成为世界上最常用的数据挖掘工具。SPSS 和一个从事数据挖掘研究的全球性企业联盟制定了关于数据挖掘技术的行业标准:CRISP-DM (Cross-Industry Standard Process for Data Mining)。与以往仅仅局限在技术层面的数据挖掘方法论不同,CRISP-DM 将数据挖掘技术与具体商业目标相结合,使数据挖掘成为商业过程,并将具体的商业目标映射为数据挖掘目标(Zhu et al.,2009)。目前世界上50%以上的数据挖掘工具均采用 CRISP-DM 的数据挖掘流程。

CRISP-DM 的数据挖掘流程包含商业理解、数据理解、数据准备、建模、模型评估和结果部署六个步骤(SPSS White Paper,2004)。Clementine 完全支持 CRISP-DM 标准,其智能预测模型有助于快速解决出现的问题。由于其对商业目标的深入理解,Clementine 最后得到的数据挖掘结果的配置非常有效(刘世平等,2003)。本研究选用的分类规则,即分类决策树的形成环境是 SPSS 的 Clementine 12.0。

(二)分类规则的构建及优化

在 Clementine 12.0 中用 C5.0 算法构建英语统考成绩预测的分类规则(即决策树的形成及优化),可分以下七个步骤。

1. 数据的提取和预处理

数据预处理是数据挖掘前的数据准备工作,目的是去除与挖掘目标不相关的数据属性和内容,为数据挖掘提供干净、准确、更有针对性的数据,减少挖掘算法的数据处理量,提高挖掘效率和最终结果的准确度。数据预处理的方法包括数据选取、数据清理、数据属性取值一致化、数据集成、数据转换和数据简化等。我们按学生基本数据、学习过程和成绩数据从数据库中提取已有英语统考成绩学生的相关数据。由于英语统考学生在学习期间可多次参加考试,为准确起见,本次只提取首次成绩,又考虑到与英语统考结果关联度较大的相关数据属性,我们制定了如下数据预处理规则:

1)由于学号和学生姓名一一对应,学生基本信息保留“学号”“性别”“入学年龄”“生源地”“所属专业”5 个属性。40 岁及以上入学者可免英语统考,故去除入学年龄>=40 学生的所有记录。

2)学生成绩数据保留“入学测试英语”“入学测试计算机”“入学测试高等数学”“入学测试大学语文”“大学英语二”“大学英语三”“所学课程平均”“学位英语”八个属性。由于入学测试时高等数学和大学语文分别是理工类和文史类专业的测试课程,除这两个属性外,去除其他成绩数据空缺的记录。

3)学生学习过程数据合并为“在线学习情况”属性,并根据平台的形成性评价系统给出的总成绩按标准化规则,从高到低以150分制赋值。

4)“统考大学英语”仅分为“合格”和“不合格”两个取值。

我们从江南大学网络教育平台数据库中提取本科在籍学生的数据后,按前面所述的数据预处理规则进行相应处理,保留了7000条相关数据,以 Excel 数据表格形式保存为“江大网络.xls”。

2. 建立 Clementine 数据源

启动 Clementine 并新建流文件后,选择界面下部“源”子菜单内的“Excel”,将其拖入面板中。双击面板中的“Excel”图标,在弹出编辑界面中选择“导入文件”,选择文件“江大网络.xls”并导入,面板中图标名称变为“江大网络.xls”。

3. 关联数据

选择界面下部“字段选项”子菜单内的“类型”,

将其拖入面板中。选择“江大网络. xls”图标,单击鼠标右键,选择“连接”,并连接到“类型”。双击“类型”图标,在弹出“类型”对话框中点击“清除所有值”后,点击“读取值”,并在“方向”列表中进行属性方向调整。其中,由于“学号”属性对于分类无作用,方向为“无”;“统考大学英语”为目标分类属性,方向为“输出”;其他属性方向均为“输入”。

4. 选择训练数据

1)选择统考大学英语成绩为“合格”的数据:选择界面下部“记录选项”子菜单内的“选择”,将其拖入面板中,并与“类型”图标连接。双击“选择”图标,在弹出“选择”对话框中,构建模式为“包含”,条件为“统考大学英语 = ‘合格’”。为了平衡训练数据中“合格”比例过大的情况,选择“记录选项”子菜单的“样本”,拖入面板中,并连接到“选择”图标。双击“样本”图标,设置采样方法为“简单”“样本”“n 中取 1”为 5(默认为 2),图标名称变为“取 15”。

2)选择统考大学英语成绩为“不合格”的数据:选择界面下部“记录选项”子菜单内的“选择”,将其拖入面板中,并与“类型”图标连接;双击“选择”图标,在弹出的“选择”对话框中,构建模式为“包含”,条件为“统考大学英语 = ‘不合格’”。

3)数据合并:选择界面下部“记录选项”子菜单内的“合并”,将其拖入面板中,并与“取 15”和统考成绩不合格的“选择”图标同时连接;双击“合并”图标,在弹出的对话框中,设置“合并方法”为“关键字”,将所有属性放入合并关键字中,并选择“包含匹配和不匹配的记录”。

5. 选择算法并建模

选择界面下部“字段选项”子菜单内的“类型”,将其拖入面板中,并与“合并”图标连接。同时,参照“关联数据”步骤,进行数据清除、读取和各属性的方向调整。选择界面下部“建模”子菜单内的“C5. 0”,将其拖入面板中,并连接到刚建立的“类型”图标。双击“C5. 0”图标,设置模型名称为“统考英语预测 1”,其余设置均为系统默认值。鼠标右击“C5. 0”图标,选择“执行”,面板右侧分栏出现“统考英语预测 1”图标,建模完成。将“统考英语预测 1”图标拖入面板中,双击该图标可查看建模结果。

6. 模型分析

将“统考英语预测 1”图标与连接“江大网络. xls”的“类型”图标连接,再选择界面下部“输出”子菜单内的“分析”,将其拖入面板中,与“统考英语预测 1”图标连接,右击“分析”图标,选择“运行”,可以看到模型输出结果与实际数值的比较,即模型的准确度。

7. 模型优化

通过前述步骤,得到的决策树由于训练数据集包含 14 个属性(学号和统考大学英语除外),显得较为复杂,程序实现较为困难。鉴于 14 个属性与统考英语成绩关联度有大小之分,在不显著影响模型准确度的前提下,我们尝试逐步去除一些关联度较小的属性,力求获得决策树复杂度和模型准确度之间的一个最佳平衡点。做法为:每次减少属性后,获得新的 Excel 数据源,重复步骤(1)-(6),得到简化模型及其准确度,并与前面的模型比较。历次属性去除情况及所得模型的决策树复杂度和预测准确度见表一。

表一 模型(决策树)优化情况

序号	模型名称	包含属性情况	决策树子节点数	预测准确度(%)
1	统考英语预测 1	包含所有 15 个属性	54	79. 18
2	统考英语预测 2	删除生源地	50	81. 85
3	统考英语预测 3	删除生源地、入学考试高等数学和入学考试大学语文	47	82. 12
4	统考英语预测 4	删除生源地、入学考试高等数学、入学考试大学语文和所属专业	44	82. 92
5	统考英语预测 5	删除生源地、入学考试高等数学、入学考试大学语文、所属专业和性别	22	80. 84
6	统考英语预测 6	删除生源地、入学考试高等数学、入学考试大学语文、所属专业、性别和入学测试计算机	10	77. 52

从表一可以看出,减少训练数据集包含的属性数量,所得到的决策树子节点数量相应减少,即决策树的复杂度相应降低。从减少属性数量对预测准确度的影响看,高等数学和大学语文由于不是所有学生都参加,应与统考英语成绩无关,属于无关属性,去除后预测准确度反而上升。同样,生源地和所属专业也属于无关属性。性别和入学测试计算机属相关属性,去除后预测准确度下降。从影响程度看,入学测试计算机属性更大。因此,综合考虑模型复杂度和预测准确度,本研究选择模型“统考英语预测 5”为最终结果。该模型在 Clementine 中的分类模型流程和所形成的决策树分别见图 2 和图 3。

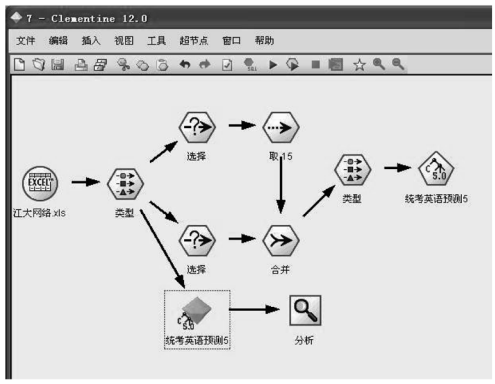

图 2 分类规则模型流程图

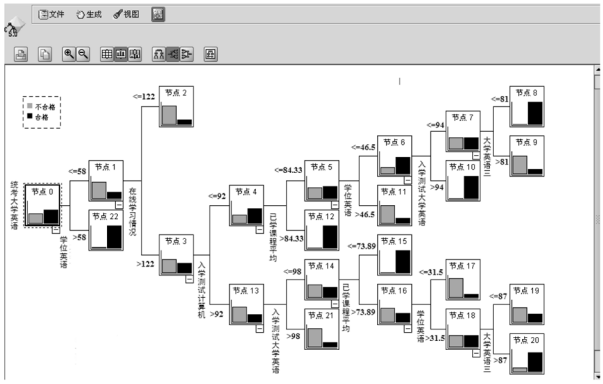

图 3 分类规则内容—决策树

(三)英语统考成绩预测的实现方法

图 3 所示的统考英语预测模型,即是所形成决策树的展开。本研究通过将 Clementine 12.0 构建的基于 C5.0 算法的分类规则,即决策树代码,转换成可执行的程序代码(其实就是 if-else 的嵌套组合),通过 PHP 中的类方法实现(Adhatrao et al. , 2013)。

江南大学网络教育的教学教务管理和学生学习平台采用 SQL Server 为后台数据库,所有学生的相关信息均存储在该数据库中。对于学生而言,学号是其在管理系统中的唯一标识,可以通过读取网页输入的学号作为查询条件,通过 SQL Query 编写的 SQL 查询接口,从数据库中读出该学生“入学年龄”“入学测试英语”“入学测试计算机”“大学英语二”“大学英语三”和“学位英语”的成绩,并读取其所有已学课程的成绩,作平均值处理后形成“所学课程平均”属性的数据;读取其形成性评价系统形成的网上学习总成绩并作 150 分制标准转换后形成“在线学习情况”属性的数据。通过在服务器端运行 PHP 语言实现的分类规则脚本程序,我们可以得到该学生本科英语统考的预测结果。

五、结果及分析

本研究采用数据挖掘分类方法中的 C5.0 决策树方法,以江南大学已参加网络教育英语统考学生的相关数据为训练数据,在 SPSS 的 Clementine 12.0 数据挖掘环境中,通过对相关属性的不断精简,最终构建了网络教育本科英语统考成绩的预测模型。该模型以“入学测试英语”“入学测试计算机”“大学英语二”“大学英语三”“已学课程平均”“在线学习情况”“学位英语”七个属性为决策树的形成因素,构建的决策树包含 22 个子节点,深度为 7,预测的准确度为 80.84%。

Clementine 环境中形成决策树时获得各属性的重要性可以通过双击图 2 中的“统考英语预测 5”图标得到(见图 4)。“学位英语”对统考英语成绩预测的重要性最大。究其原因,主要是两者的考试要求较为接近,考试的题型和题量基本相同。学士学位的获得比毕业要求更高。单纯从考试难度而言,学位英语的难度要略高于统考英语。而从考试的形式而论,英语统考是完全基于在线题库的全机考模式,学位英语是传统的试卷笔试模式,并且有一定的考试范围。对于成人学生而言,更加适应传统的笔试模式。因此综合相比,两者的考试难度相当。如果达到了学位英语考试的要求,说明学生的英语总体水平上了一个台阶,英语统考通过的可能性自然增加。

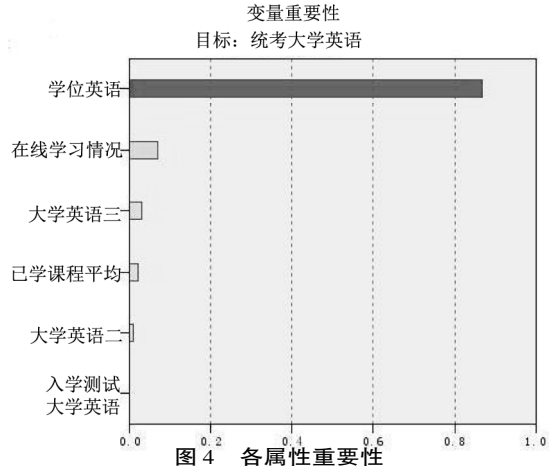

图 4 各属性重要性

网络教育学生主要通过在线学习的形式完成课程学习,达到学习目标。学生的在线学习情况直接反映了学生平时学习的状态和态度。“在线学习情况”成绩好,说明学生平时学习态度比较认真,坚持

网络学习,各项学习任务能按时按要求完成,英语统考通过的可能性就高。因此,在线学习情况对英语统考成绩预测的重要性占第二位是合理的。

“大学英语三”和“大学英语二”是网络教育本科学生的两门英语课程。其中,“大学英语三”的课程要求与英语统考的考试大纲要求更为接近,两者对英语单词、语法、听力、翻译和写作等分项的具体要求类似;而“大学英语二”的课程要求要低于英语统考。英语课程的学习是学生提高英语水平和通过英语统考的先决基础条件,因而“大学英语三”和“大学英语二”对英语统考成绩预测的重要性排在第三位和第五位是可以理解的。通过一系列恰当的措施有效提高“大学英语三”和“大学英语二”的学习效率和效果,尤其是前者对学生提高英语水平和英语统考的通过率意义深远。

在所有关联属性中,“已学课程平均成绩”的重要性排在第四位。该属性反映了学生网络学习的最终效果。学生平时学习坚持得好,投入的时间和精力多,课程的平均成绩自然就好。与其相对应,学生投入英语学习的时间也就相应增多。这直接关系到英语的学习效果,最终影响英语统考的成绩。

“入学测试大学英语”是试点高校针对就读学历教育学生入学组织的英语基础水平测试,其成绩反映了学生入学前的英语基础水平,是后续英语学习的基础,对于预测学生入学后的英语学习成绩有一定的重要性,但不是关键因素,因为它与后续学习情况关联度更高。

“入学测试计算机”的成绩直接反应了学生入学前的计算机应用水平。因为网络学习主要是通过计算机网络进行课件学习、完成各类学习任务、参与学习讨论和疑难问题解决等学习主要环节;英语统考的完成也需具备一定的计算机应用能力。因此,计算机应用水平对日常学习和英语统考的通过率具有一定影响,这是该属性对英语统考预测具有一定关联度的原因所在。

综上所述,英语统考成绩是对网络教育本科学生英语学习效果的最终考核。我们所选取用于生成预测结果的七个属性中,“入学测试英语”和“入学测试计算机”分别代表了学生的学习基础;“在线学习情况”和“所学课程平均成绩”分别代表了学生的学习状态和整体学习效果;“大学英语二”和“大学英语三”是学生英语学习的阶段性结果;“学位英语”是与英语统考同等重要的英语学习最终考核;将这七个属性作为英语统考的预测依据是合理的。

除了两个代表学生基础的属性外,另外五个属性都是通过具体的学习过程形成的。通过最终的英语统考预测结果可以分析出学生在整个英语学习过程中的薄弱阶段,从而进一步分析决定学生英语学习各阶段学习效果的学习行为,如网上学习的参与度、网络学习次数、网络学习时长、网上提交作业情况、网上交互讨论情况、学生前期考试行为等。通过对这些学习行为的统计、干预和预警提醒,同时,采取一系列有效的措施,特别是提高和完善对学生的教学管理、学习指导和支持服务,激发学生网络学习的自觉性,可以提高在线学习效率和效果,提升学生各阶段性结果的成绩,最终提高英语统考通过率。而这也正是我们后续研究的重点。

随着我国网络教育的深入发展,个性化学习支持服务正越来越被重视和研究,各类应用系统也正在逐步进入开发和应用阶段(顾小清等,2012;吴永和等,2013)。个性化学习支持服务即运用数据挖掘和数据分析技术,依据网络学习系统已有的大量数据,关注学习者的学习背景、学习习惯、学习兴趣和学习关注度等个性化因素与其学习效果的关联度,对学习者网络学习过程的主要环节进行个性化提醒、学习指导、学习资源和学习方法推荐等学习支持服务。本研究所形成的学生网络教育英语统考成绩的预测结果可以作为学生英语学习和统考辅导的个性化服务依据。

[参考文献]

[1] Adhatrao, K., Gaykar, A., Dhawan, A., Jha, R., & Honrao, V. (2013). Predicting student’ performance using ID3 and C4. 5 classification algorithms [J]. International Journal of Data Mining & Knowledge Management Process, (3):5.

[2]陈文伟,黄金才(2004). 数据仓库与数据挖掘[M]. 北京:人民邮电出版社(第2版):4.

[3]陈益均,殷莉(2013). 基于数据挖掘的学生成绩影响模型的研究[J]. 现代教育技术, (1): 94-96.

[4]傅钢善,王改花(2014). 基于数据挖掘的网络学习行为与学习效果研究[J]. 电化教育研究, (35): 53-57.

[5]顾小清,张进良,蔡慧英(2012). 学习分析:正在浮现中的数据技术[J]. 远程教育杂志,(1):18-25.

[6]黄荷(2012). 今日谈:大数据时代降临[J]. 半月谈, (17):

49-50.

[7]洪建峰(2013). 数据挖掘技术在远程教育中的应用研究[J]. 微型电脑应用,(8):43-45.

[8] Kantardzic, M. (2011). Data mining: Concepts, models, methods and algorithms[M], Wiley-IEEE Press.

[9]刘世平,姚玉辉(2003). 数据挖掘工具的评判[J]. 数字财富,(6): 74-76.

[10]SPSS White Paper(2004). Working with telecommunications churning in the telecommunications industry[R]. SPSS White Paper.

[11]魏顺平(2013). 学习分析技术:挖掘大数据时代下教育数据的价值[J]. 现代教育技术,(2):5-11.

[12]吴永和,陈丹,马晓玲. (2013). 学习分析: 教育信息化的新浪潮[J]. 远程教育杂志,(4):11-19.

[13]Zhu Xiaoliang, Wang Jian, Yang Hongcan, & Wu Shangzhuo (2009). Research and Application of the improved Algorithm C4.5 on Decision Tree[A]. International Conference on Test and Measurement (ICTM)[C]. (2):184-187.

(编辑:李学书)

Research and Implementation of Network Education Results Prediction in Big Data Era: A Case of English Unified Exanination for Online Undergraduates

SUN Li & CHENG Yuxia

(*School of Humanities, Jiangnan University, Wuxi* 214000, *China*)

Abstract: *With the development of information society, information storm brought by big data is changing our life, working and thinking style. More and more students are participating in online diploma education. They learn through accessing and using online resources, online homework, interactive discussions and examinations. Such participation has left or generated a giant useful data and information in various types of course management and learning systems. Using appropriate data analysis techniques, these data and information can help us obtain useful knowledge, find relevant disciplines, and provide useful basis for decision making, programming all aspects of online learning, optimizing processes for teaching management, and improving the quality of teaching and changing the design of educational software. All of these are inevitable demands for sustainable development of online education. We analyzed English learning and other relevant information of undergraduate students in online diploma education using data classification technology in data mining, and forecasted the prospects when they took the English unified examination, which was necessary for their graduation. After briefly describing the concepts and relevant theories of data mining, we compared the differences between classification and clustering techniques in data mining and analyzed the characteristics of achieving the goal for forecast. We determined to use C5.0 decision tree classification in data mining. Using the relevant data of students who took online education in Jiangnan University and had taken the English unified examination as the training data, our research went through four steps: data retrieving, data preprocessing, decision tree structuring and optimization. Then, in Clementine* 12.0 *data mining environment of SPSS, we built a forecasting model of the undergraduate English unified examination. This article also proposes the implementation method of the model, discusses and analyzes the constructed forecasting model and the importance of the related properties, and proposes appropriate policies about how to improve the level of undergraduate English online education and the throughput rate of the English unified examination.*

Key words: *online education; data mining; decision tree; the English unified examination; forecasting model*