分类号_____ 密级_____

UDC _____

昆明理工大学 专业硕士学位论文

数据挖掘在学生在线测试与预测中的应用研究

研 究 生 姓 名 卫明

指导教师姓名、职称 袁梅宇 副教授

学 科 专 业 计算机技术

研 究 方 向 教育数据挖掘

论 文 工 作

起 止 日 期 2015 年 7 月~2018 年 3 月

论 文 提 交 日 期 2018 年 3 月

http://www.ixueshu.com

学 位 论 文 出 版 授 权 书

我同意将本人学位论文著作权中的数字化复制权、发行权、汇编权和信息网络传播权的专有使用权在全世界范围内授予中国学术期刊（光盘版）电子杂志社（下简称“杂志社”），同意其在《中国优秀博硕士学位论文全文数据库》和 CNKI 系列数据库中出版，未经杂志社书面许可，我不再授权他人以数字化形式出版本文。我同意《中国优秀博硕士学位论文全文数据库出版章程》规定享受相关权益。

如有任何第三方未经杂志社许可使用本人论文，杂志社应追究其法律责任，诉讼的全部费用由杂志社承担。胜诉后，由杂志社与本人按 5：5 的比例分配所获赔偿金。

作者签名：卫明

2018 年 05 月 31 日

http://www.ixueshu.com

学位论文作者信息

论文题目	数据挖掘在学生在线学习测试与预测中的应用研究				
姓　　名	卫明	学号	2014704132	答辩日期	2018 年 05 月 29 日
论文级别	博士□　　硕士☑				
院 / 系 / 所	信息工程与自动化学院		专　　业	计算机技术	
联系电话			E_mail		
通信地址(邮编).					
备注：					

☑公开 □保密（___年___月至___年___月）(保密的学位论文在解密后应遵守此协议)

联系电话：010-62791951 62793176 62790693　　传真：010-62791814

通信地址：北京清华大学邮局 84-48 信箱　采编中心　　邮编：100084

学位论文使用授权书

本论文作者完全了解学校关于保存、使用学位论文的管理办法及规定，即学校有权保留并向国家有关部门或机构送交论文的复印件和电子版，允许论文被查阅和借阅。本人授权昆明理工大学可以将本学位论文的全部或部分内容编入学校有关数据库和收录到《中国博士/优秀硕士学位论文全文数据库》进行信息服务，也可以采用影印、缩印或扫描等复制手段保存或汇编本学位论文。

注：保密学位论文，在解密后适用于本授权书。

作者签名：王明　　导师签名：袁梅宇

2018年05月31日　　2018年05月31日

http://www.ikueshu.com

学院：信息工程与自动化学院

学号：2014704132

专业：计算机技术

（一式三份，交研究生院学位工作处）

一　遵守学术行为规范承诺

本人已熟知并愿意自觉遵守《昆明理工大学研究生学术规范实施细则（试行）》的所有内容，承诺所提交的毕业和学位论文是终稿，不存在学术不端行为，且论文的纸质版与电子版内容完全一致。

二　独创性声明

本人声明所提交的论文是我个人在导师指导下进行的研究工作及取得的研究成果。尽我所知，除了文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得昆明理工大学或其他教育机构的学位或证书而使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示了谢意。本人完全意识到本声明的法律结果由本人承担。

三　关于论文使用授权的说明

本人完全了解昆明理工大学有关保留使用学位论文的规定，即：学校有权保留送交论文的复印件，允许论文被查阅和借阅；学校可以公布论文的全部或部分内容，可以采用影印、缩印或其他复制手段保存论文。（保密的论文在解密后应遵守此规定）

本学位论文属于（必须在以下相应方框内打“√”，否则一律按“非保密论文”处理）：

1、保密论文：　□本学位论文属于保密。

2、非保密论文：　□本学位论文属于内部论文，网上延后公开。

□√本学位论文不属于保密范围，适用本授权书。

是否同意授权以下单位（必须在以下相应方框内打“√”，否则一律按“同意授权”处理）：

□√同意授权　　□不同意授权

http://www.lxueshu.com

将本人学位论文著作权中的数字化复制权、发行权、汇编权和信息网络传播权的专有使用权在全世界范围内授予中国学术期刊（光盘版）电子杂志社，并在《中国优秀博硕士学位论文全文数据库》和 CNKI 系列数据库中出版。

研究生本人签名：王明　　签字日期：2018 年 05 月 31 日

研究生导师签名：[illegible]　　签字日期：2018 年 5 月 31 日

摘要

教育是对中华民族伟大复兴具有决定性意义的事业。教育数据挖掘是一个新兴的多学科的研究领域，探索获取各种教育信息系统中的数据信息的方法和技术，教育数据挖掘得到了机器学习、人工智能、教育学、认知学等领域研究人员的广泛认同，国内外都在积极投入研究。本论文研究学习者的知识掌握状态的学习者知识模型，是教育数据挖掘中的核心研究领域，具有很好的研究价值和实际意义。

本文主要完成以下三个方面的工作。

第一，研究如何构建学习者知识模型，通过挖掘智能教学系统的日志数据，了解并掌握学生对知识的学习过程和掌握程度。

第二，比较模型的预测与真实结果，得到评价指标，评估学习者知识模型与实际的符合程度。

第三，将研究成果应用到学校在校教育系统获取到的实际数据中，证明本文方法得到的模型有一定的准确性。

具体地，论文以 KDD CUP 2010 竞赛的公开数据集为研究对象，应用多种预处理方法，通过繁琐而极费时间的处理，将 5.29G 大小、2 千万个样本的数据缩减为个人电脑可以挖掘的数据集，然后进行建模和模型评估。本文模型的准确率达到 88.9495%，均方根误差 RMSE 为 0.2848，比竞赛冠军的 RMSE 0.271157 稍差，冠军是由台湾大学著名教授林智仁领衔的 25 人豪华团队。并将本文模型与逻辑回归、SVM 算法、贝叶斯网络和 BP 神经网络算法所构建的模型比较，发现本文模型效果最好。本文还对模型的泛化能力进行评估，证明在极端情况下模型也具备很好的泛化能力。最后，挖掘学校在线教育平台中的学生实际数据，得到准确率为 82.3301%。

本文的研究遵从机器学习中著名的奥卡姆剃刀原则：对数据最简单的解释也就是最好的解释。将公开数据集里的 20 个特征通过预处理转换为 3 个特征，极大地简化的问题，取得了良好的效果。

关键词：教育数据挖掘；知识模型；KDD CUP 2010

http://www.ixueshu.com

Abstract

Education is a decisive cause for the great rejuvenation of the Chinese nation. Educational data mining is a new and multi-disciplinary research field, exploring the methods and techniques to obtain data from various educational information systems. Educational data mining has been widely recognized by the researchers in the fields of machine learning, artificial intelligence, education, cognition and so on. This paper studies the learner's knowledge model of the knowledge mastery state. It is the core research field in the education data mining. It has good research value and practical significance.

This article mainly completed the following three aspects of the work.

Firstly, it studies how to construct the learner's knowledge model, and to understand and master the students' learning process and mastery of knowledge by digging the log data of the intelligent teaching system.

Second, it compares the prediction and real results of the model, gets the evaluation index, and evaluates the degree of conformity between the learner's knowledge model and the actual situation.

Third, it applies the research results to the actual data obtained from the school education system, proving that the model obtained by this method has certain accuracy.

Specifically, the paper takes the open data set of the KDD CUP 2010 competition as the research object, uses a variety of preprocessing methods, and reduces the data of the size of the 5.29G and the 20 million samples into the data set that the personal computer can excavate through the tedious and extremely time-consuming processing, and then carries out modeling and model evaluation. The accuracy of this model is 88.9495%, the root mean square error RMSE is 0.2848, which is a little worse than the RMSE 0.271157 of the competition champion. The champion is the 25

http://www.ixueshu.com

Deluxe team, led by Lin Zhiren, a famous professor at National Taiwan University. Compared with the models constructed by logistic regression, SVM algorithm, Bayesian network and BP neural network, we find that the model is the best. This paper also evaluates the generalization ability of the model, which proves that the model has good generalization ability in extreme cases. Finally, the actual data of the students in the online education platform are excavated, and the accuracy rate is 82.3301%.

This study follows the famous Occam razor principle in machine learning: the simplest explanation of data is the best explanation. The 20 features in the open data set are converted to 3 features by preprocessing, which greatly simplifies the problem and achieves good results.

Key words: Educational Data Mining; Knowledge Model; KDD CUP 2010

http://www.ixueshu.com

目录

http://www.lwchu.com

http://www.lwlieshu.com

http://www.ixueshu.com

第一章 绪论

随着机器学习、数据挖掘和教育信息化的发展，研究教学过程中学生自主学习的日志数据逐渐成为较为热门的领域，将数据挖掘技术应用到挖掘学生的学习日志，以及应用到学生学习过程的方方面面，逐渐成为一个称为教育数据挖掘的一门新兴技术。本文以 KDD CUP 2010 竞赛提供的公开数据集为研究对象，采用特征工程技术对原数据进行预处理，然后划分训练集和测试集，使用训练集来训练分类器，使用测试集来评估分类模型，对学生学习数学的日志记录进行挖掘，获取到学生的水平和题目难度信息，从而预测学生能否做对下一道题。这样，可以评估学生对所学知识点的掌握情况，以助于合理安排教学进度，提高学习效率和教学质量。

1.1 选题的背景与研究意义

数据挖掘技术起源于 20 世纪 80 年，到现在已经取得了非常广泛及重大的进展。教育数据挖掘综合了多个学科的理论和技术来解决教育教学实践中遇到的问题。对教育相关的数据进行分析预测，找到一条可行的可以解决教育教学中遇到的问题的道路。

http://www.kaoshu.com

1.1.1 选题背景

如今我国高等院校基本上都已使用在线学习系统、学生成绩管理系统、学籍管理系统等系统管理数据[1]。这些系统收集了多年的教学信息，数据量十分庞大，教育界慢慢把目光聚焦在如何有效利用已经这些大量的教育数据信息，从中发现有利于提高学生学习效率和提高教师教学质量的有用信息。随着在线教育 E-Learning 的推广，学生可以随时随地自主学习，创造了跨时空的生活、工作和学习方式，使知识获取的方式发生了根本变化。教和学可以不受时间、空间和地点条件的限制，知识获取渠道灵活与多样化。随着时间的推移，在线教育系统中相应地积累了数量庞大的数据，手工处理这些数据几乎不可能，研究如何挖掘学生相

关数据势在必行。通过分析学生学习的日志数据，找出它们内在的联系和规律，预测学生的表现，为提高教学质量和优化教育服务。教育数据挖掘技术是信息技术与教育相结合的产物，为解决教育信息化产生诸多问题的一门新兴技术。

美国教育部在发布的《通过教育数据挖掘和学习分析技术来提高教与学：问题简述》（Enhancing Teaching and Learning through Educational Data Mining and Learning Analytics: An Issue Brief）报告中，主张通过教育数据挖掘、学习分析和可视化数据分析来改进自适应学习系统，实现个性化学习。

1.1.2 研究意义

教育数据挖掘能够帮助教师和学生改进教学质量。首先，在大数据背景下，数据挖掘技术可以有效帮助教师改进教学。例如，教师可以查看学生在一道题上的停留时间，判别学生在答错一道题以后有没有进行复习。大数据可以统计学生在网上提问的次数以及参与讨论的多少，然后在对学生的学习行为进行引导；通过记录学生在学习过程中的鼠标点击量，可以用于研究学生学习活动的轨迹，研究学生对知识点的反应情况，学习知识点的用时，从而了解哪些知识点需要重复或强调，以及哪些陈述方式或学习工具最为有效。

http://www.lxueshu.com

大数据还可以帮助教师全面、正确地评价学生的学习能力。传统的学生评价方式往往依靠感觉、直觉和考试。人的感觉和直觉不完全可靠，考试存在局限，学生可能会做错已经懂的题，也可能碰巧做对不懂的题，因此，传统的评价方式不那么可靠。而大数据凭借日常点点滴滴的信息采集，运用科学的逻辑推理，能客观地展现学生的完整形象，可以较为客观地对学生进行审视与评估。

可见，应用教育数据挖掘的数据分析结果，教师可以更好地了解学生，观察和了解学生的学习行为，找到最合适的教学方法和教学顺序。也可以针对不同特点的学生采用不同的教学方法与教学策略，及时发现问题并进行有效干预，从而显著提高教学的质量与效率。

本文试图利用先进的机器学习和数据挖掘技术，通过对学生学习日志数据进行挖掘，建立学生学习模型，对学生掌握各个知识点的情况进行建模。

由于我们无法进入到学生头脑中，查看学习模型是否符合实际，因此，合理的思路是将学习模型视为一个黑箱，查看黑箱预测结果是否与学生的实际做题结果相符，如果符合则可认为学习模型性能好。例如，本文的学习模型预测准确率到达 80%以上，则认为学习模型的置信度较高。

获取到置信度高的学习模型有很好的研究意义。首先，模型可以预测学生在未来的重要考试中大致的得分，这对学生本人准备重要考试至关重要。其次，可以直接判断某个学生对某个知识点的掌握程度，只需要从题库中获取该知识点的给定数目的题目，使用学习模型来预测该学生的得分即可进行判断。如果没有学习模型，就大概只能依靠把学生叫来做题才能做到。

正是因为学习模型重要，由 ACM（Association for Computing Machinery，计算机协会）的 SIGKDD（Special Interest Group on Knowledge Discovery and Data Mining，ACM 知识发现和数据挖掘国际会议）组织的年度竞赛 KDD CUP 在 2010 年才将教育数据挖掘作为竞赛的主题，邀请全世界的数据挖掘团队来解决这一世界难题。本文的研究对象就是 KDD CUP 2010 公开数据集。

http://www.kuleshu.com

1.2 国内外教育数据挖掘研究动态

近年来，随着教育信息化数字化校园建设，现代远程教育的开展和 Web 2.0 等广泛应用的促进下，大量的研究者开始着手教育数据挖掘(Educational Data Mining，简称 EDM)方面的研究[2]。在数字化教育建设的近几年中，教育数据领域迎来了巨大的变化。在线学习系统、智能手机应用和社交网络为 EDM 研究提供了大量的应用和数据。截止至 2016 年底，全球超过 6000 万名师生[3]使用在线学习系统 MOODLE[4]进行学习和交流。到 2017 年底，全球超过 23 亿人在使用智能手机，社交媒体 Facebook 的活跃用户数超过 16 亿人[5]。近两年兴起的新型教学

模式大规模公开在线课程(Massive Open Online Courses，简称 MOOCs[6])，截止至 2016 年底，在 MOOCs 网站 Coursera 上注册的用户人数已超过 1500 万[7]。

计算机科学、教育学和统计学是与 EDM 联系最紧密的学科，它们之间的关系如图 1.1 所示[8]。

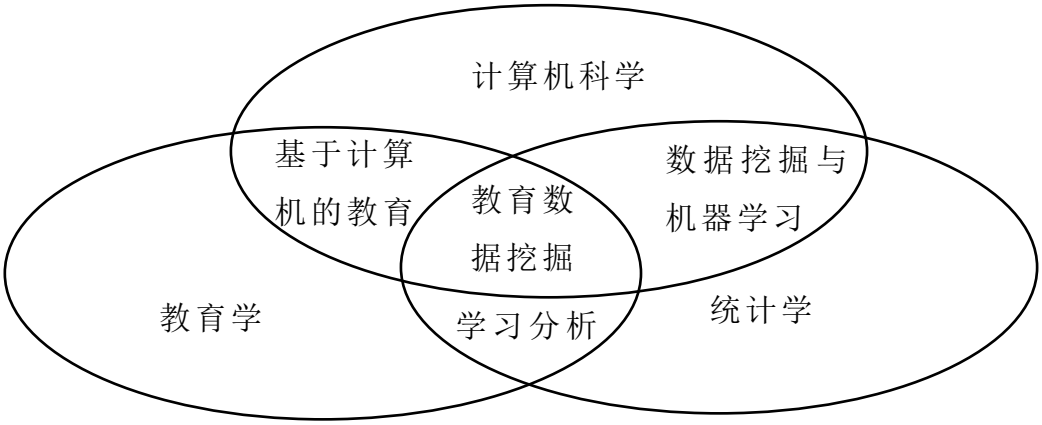

图 1.1 EDM 涉及的主要学科

从图中可以看到，这三大学科两两交叉又衍生了数据挖掘与机器学习(Data Mining and Machine Learning，简称 DM&ML)、基于计算机的教育(Computer-Based Education，简称 CBE)及学习分析(Learning Analytics，简称 LA)。通过与这 3 个领域的对比可以看出 EDM 的特点[9]。

EDM 与一般的 DM&ML 研究的不同之处在于其数据具有教育学科的性质，主要表现在以下几个方面：

- 多学科

EDM 数据与多个学科的概念和技术都有关系。数据如学习目的、教学方式、教学评估、兴趣、人际关系和家庭背景等方面涉及到与社会学、教育学和心理学等相关的概念和技术。研究者面对这一类数据，首先要知道其中的概念，同时还要熟知与它们相关的测量和评价技术。

- 多层次

EDM 数据的多层次特性源于教育机构构成的多层性，如学生一般按学校、学区、院系、专业和班级进行划分；加上教学材料的内容结构多层性，如课本内容可按单元、章节、知识点进行组织。

- 多精度

EDM 数据的多精度来自于教育研究有的可能要持续几年甚至几十年，有的教学研究需要精确到秒甚至是毫秒。因此研究者可以根据不同的需求按时间精度进行分类。

- 多情景

EDM 数据的多情景特性来自于教与学的活动的本身特点。学生获得知识的多少、好坏与教师水平、学习环境有关；也与学生自身的学习能力、自身的学习状态和学生人际关系都有关系。上面任何一项的改变都会影响学生获得知识的效率。

EDM 与一般的基于计算机的教育（CBE）研究的主要不同在于应用目的的不同，基于计算机的教育（CBE）研究的目标是辅助或替代传统的教学过程，而 EDM 则主要解决传统教学缺少或难以实现的功能，如：学生可以了解自己的学习效率、学习效果等；老师可以了解教学的效率，改进教学材料，了解学生的个体和总体情况及预测学生的学习成绩等。

EDM 与一般 LA 研究的不同之处在于两者采用了不同的方法：EDM 用机器学习和数据挖掘，而 LA 采用统计；从另一角度来看，LA 侧重于描述已发生的事件或其结果，而 EDM 侧重于发现新知识与新模型[10]。

http://www.ixueshu.com

1.2.1 EDM 的发展历程

EDM 的发展大致可分为两个时期：

第 1 个时期是从上世纪 80 年代初到上世纪末，研究者开始将简单的数据挖掘技术应用于教育领域，但取得的成果很少。这一时期采用的数据挖掘方法主要是关联规则算法和统计分析[11]。

第 2 个时期则是从本世纪初至今，EDM 进入快速发展阶段。随着计算机的普及和高速发展，EDM 进入了高速发展时期。EDM 数据来源更加多样，特别是在线学习系统广泛应用，为 EDM 研究提供了丰富的数据，同时采用的数据挖掘技术更加多样化[12]。和国外 EDM 研究相比，国内 EDM 研究相对起步晚，研究深度和广度均落后于国外[13]。

在 EDM 发展的各个时期，均有相应的综述性论文发表[8,14-20]。2010 年 Romero 等人编写的第一本关于 EDM 技术的专业书《Handbook of

Educational Data Mining》[21]，共有 36 章，对 EDM 的概念做了详细阐述了，介绍了主要技术和与之对应的典型案例。2014 年 Baker 在 MOOCs 网站 Coursera(https://www.coursera.org) 上开设课程“Big Data in Education”，讲授 EDM 的基础知识和技术。这两本书详细介绍了 EDM 技术。

1.2.2 EDM 的最新研究进展

教育领域中基于不同终端的数字化学习日渐普及先进学习平台记录学习行为数据的多样性与海量性，使得教育大数据的应用类型更加多样化，教育数据价值日益凸显。

主要的 EDM 应用类型如下：

可视化(Visualization，简称 VS)将信息或知识通过现代技术化作静态或动态图像形象地展示在人们面前。在 EDM 中，人们通过可视化技术更加直观地理解教育数据，如用户在线论坛数据[22]、在线评估过程中产生的数据[23]、教师和学生之间的互动[24]、考试成绩[25]或者学生团体活动的相关数据[26]等。

学生建模(Student Modeling，简称 SM)通过对学生的行为、动机和学习策略等方面建立模型来揭示其学习特征。例如学习者的应答数据，包括正确的、部分正确的和错误的应答、应答时长、所需提示、误答次数等等；学习者的技能练习数据（内容及持续时间）。在 EDM 中，采用了贝叶斯网[27-31]、序列模式挖掘[32-34]、关联规则[35,36]和逻辑回归[37]等方法对学生特点和学习行为进行自动建模[37]。本文根据学生在线学习系统进行数学学习的交互日志，对学生的应答数据进行建模。包括学生答题次数，每道题被回答的次数和答对次数，每道题目所涉及的知识点，答对某道题需要的提示次数等。通过对这些数据的分析，得出学生的答题能力，问题难度，知识点的难易程度等特征。

学生表现预测(Predicting Student Performance，简称 PSP)通过对现有数据分析处理，预测学生下一阶段的学习表现。学生表现预测是 EDM 最早也是最流行的应用之一[38]，通过对教育数据的挖掘，提取出有用的信息，对学生数据进行建模，运用相关算法对学生表现进行预测。例如

http://www.ixueshu.com

根据学生在线学习记录预测学生的最终分数[39]或者预测学生下一阶段的表现[40]等等。本文基于学生在在线学习系统进行数学学习的交互日志所提供的挑战数据集建立的模型，分别使用决策树 C4.5 算法和逻辑回归算法对挑战数据集中的训练部分进行学习训练，预测学生在未来题目上的表现。

推荐系统(Recommender System，简称 RS)通过对学生数据的分析，总结学生的特点、状态，合理的向学生推荐相应的学习内容，包括课程、资料和方法[41]。

自适应系统(Adaptive System，简称 AS)可以根据学生建模的结果做自适应变化的学习系统[42]。

随着开展 EDM 研究的计算机技术专家越来越多，未来专用的教育数据挖掘技术很有可能会逐渐形成体系。与传统教育研究相比，EDM 具有无可比拟的优势。EDM 借助大数据处理技术，可以快速处理包含数万甚至数十万学生信息的数据，完成数据建模、预测、可视化等一系列复杂的操作。

http://www.ixdaohu.com

1.2.3 现有研究的不足及发展趋势

认清 EDM 研究目前存在的不足，使我们对其研究现状有了更加清晰的认识，同时也让我们更加清楚 EDM 未来的发展趋势。

首先是研究选题的不足。在 EDM 众多的研究类型中，学生表现预测和推荐系统对教育的影响最大。它们不仅加深了我们对教育理念的认识，更促进了我们更进一步探索教育发展；同时改变了传统教学方式，提高了教与学的质量。在 EDM 研究中，选题时要注意教育与数据挖掘两者之间的关系。数据挖掘是其方法，而教育是其目的。因此，要利用先进的数据挖掘技术解决教学过程中遇到的问题。

其次是研究方法的不足。在 EDM 的研究中，对数据预处理技术的研究较少。现有的 EDM 文献中处理的数据一般是意义清晰的最终数据集，很少对数据预处理工作进行详细描述。然而，EDM 具有多情景、多语义、存在大量噪声和数据缺失等特征。数据预处理方法对于 EDM 研究的重要性不亚于数据挖掘算法，在有的情况下甚至超过后者。因此，

研究者应特别重视数据预处理方法的研究和论述，特别是那些具有推广价值的预处理技术。本文在预测学生未来题目表现的过程中，着重介绍对学生数据预处理的过程，包括训练集和测试集的划分、属性选择、属性变换和缺失值处理等。

1.3 论文结构

本文共分为六个章节。用到的数据挖掘工具为 Weka[43]。数据集为 KDD CUP 2010[44]提供的公开数据即根据学生使用的在线学习系统进行数学学习的交互日志。

第一章：对研究的背景以及意义做了相关的介绍，论述了课题研究的意义——即预测学生能否正确完成下一道题的科研价值，并总结了本文的主要工作。

第二章：教育数据挖掘概述。本章介绍教育数据挖掘的特点及挖掘方法；阐述教育数据挖掘工作过程，包括预处理、模型训练和模型评估等；介绍本文用到算法。为下一章数据挖掘过程做准备。

第三章：数据挖掘过程。本章主要讲述所使用的数据集、数据预处理、模型训练和模型评估，并将本文方法与竞赛团队 Y10、NTU 结果做了对比。

第四章：改进的挖掘过程。本章主要讲述如何对挖掘过程进行改进。采用更接近实际的划分训练集与测试集方法，重新对数据集进行预处理，发现预测结果稍差，但更接近实际，并与其它分类器结果比较，发现 C4.5 决策树分类器效果最好。本章还通过实验评估模型的泛化能力，证明论文所构建模型的有很好的泛化能力。

第五章：实际应用。本章通过对学校“教育在线”网络教学平台中实际存在的学生数据进行处理，主要是集中在对学生做题等相关教学行为的教育数据分析，进而对学生在测试做题的过程中进行预测，验证本文上一章中所提出的更接近实际的挖掘过程的可行性和正确性。

第六章：总结与展望。总结了本文的主要工作和最终结果，并对论文的创新与不足进行分析，提出下一步工作。

http://www.ixueshu.com

第二章 教育数据挖掘概述

教育数据挖掘需要将数据挖掘、学习分析、人工智能等先进技术结合起来，涉及多学科知识。教育数据挖掘的工作包括数据预处理、预测、评估等过程。数据的预处理需要经过特征选择、特征变换、缺失值插补、训练集与测试的划分等过程。

2.1 教育大数据

教育大数据的定义最早从产生教育大数据的主体出发，将教育大数据分为广义的和狭义的两类[52]：广义的教育大数据泛指所有来源于日常教育活动中人类的行为数据；狭义的教育大数据是指学习者行为数据，主要来自于学生管理系统在线学习平台和课程管理[49]。也有研究指出教育大数据指整个教育活动过程中所产生的以及根据教育需要采集到的，一切用于教育发展并可创造巨大潜在价值的数据集合。

基于以上研究，可以认为教育大数据的定义包含三层含义：第一个含义，教育大数据是教育领域的大数据，是面向特定教育主题的多类型、多维度、多形态的数据集合；第二个含义，教育大数据是面向教育全过程的数据，通过数据挖掘和学习分析支持教育决策和个性化学习；第三个含义，教育大数据是一种分布式计算架构方式，通过数据共享的各种支持技术达到共建共享的思想[53]。也就是说，我们把教育大数据定义为：面向教育全过程时空的多种类型的全样本的数据集合。教育大数据不仅仅是建设教育大数据中心，不仅仅是分析全过程学习数据，更多的是一种共享的生态思想[54]。

2.1.1 教育大数据的特点

教育大数据有更强的实时性、连续性、综合性和自然性，并使用不同的应用程序来分析和处理不同复杂度和深度的数据。以“大数据教育”为特征的数据时代的教育，教育大数据主要来自于教学活动过程，在课

http://www.kuaishu.com

堂教学、考试评价和互动网络中直接产生。教育数据挖掘和学习分析在教育领域中的应用模式详情见下表 2.1。

表 2.1 教育数据挖掘和学习分析在教育领域中的应用模式

学习者知识建模	学习者掌握的知识（概念、技能、过程性知识和高级思维技能等）	1．学习者的应答数据，包括正确的、部分正确的和错误的应答、应答时长、所需提示、误答次数等等 2．学习者的技能练习数据（内容及持续时间） 3.学习者的测试结果数据（形成性和总结性）
学习者行为建模	学习者不同的学习行为范式与其学习结果之间的关系	1．学习者的应答数据，包括正确的、部分正确的和错误的应答、应答时长、所需提示、误答次数等等 2．在课堂/学校环境下的学习行为变动数据
学习者体验建模	学习者对自己学习体验的满意程度	1．满意度问卷调查数据 2．学习者对后续学习单元或课程采取的行为和表现的数据
学习者建档	学习者聚类分组	学习者的应答数据，包括正确的、部分正确的和错误的应答、应答时长、所需提示、误答次数等等
领域建模	对主题模块的划分和排序	1．学习者的应答数据，包括正确的、部分正确的和错误的应答、应答时长、所需提示、误答次数等等 2．领域模块分类数据 3．技能与问题之间、问题与问题之间的关联性数据
学习组件分析和教学策略分析	促进有效学习的学习组件、有效的在线教学策略、在线课程的整体效果	1.学习者的应答数据，包括正确的、部分正确的和错误的应答、应答时长、所需提示、误答次数等等 2．领域模块分类数据. 3.技能与问题之间、问题与问题之间的关联性数据
趋势分析	未来趋势的容及原因	１．选取三个以上的数据点，用于纵向庭势识别 2．数年内的入学记录、学位、生源等学生基本信息数据
适应性和个性化	对学习者的学习建议、学习体验对后续学习的促进化用、学习体验的实时改善	1.学生的历史数据等 2.学生的学业成绩数据

http://www.ixueshu.com

哈佛大学、斯坦福大学和其他世界知名大学还推出了一个教育数据相关的研究方案[45]；此外，美国学校管理者协会，全球信息技术研究和咨询公司 Gartner 公司联合启动“封闭缺口：将数据转化为行动”项目，旨在促进学生信息系统和数据管理系统的使用学习[46]。在 2012 年 3 月，奥巴马政府投资了 2 亿美元的资金，对海量数据的进行开发和研究，提高采集、储存、保留、管理、分析和共享的效率[47]。为了促进“大数据”教育，美国的大学和中小学在“大数据”在教育中的应用提供有效的指导，2012 年 10 月，美国教育部颁发的《通过教学教育数据挖掘和分析学习与学习》的报告[48]。

2.1.2 教育数据的挖掘方法

教育数据挖掘主要通过分类分析、聚类分析、回归分析和关联规则等算法和技术来发现隐藏在教育数据库中有用的信息[50]-[51]。

(一)分类分析

教育中最常用的数据挖掘技术就是分类，即决策树和神经网络训练一组预先分类的数据，建立一个模型来分类的其他数据。该过程包括两个阶段：学习和分类。在学习阶段，分类算法使用训练数据训练分类模型。在分类阶段，测试数据用于评估分类器的性能。如果性能满足要求，可以用分类模型来预测新的数据集。神经网络具有预测的作用，对实验数据做训练，通过神经网络的权值来对预测数据，神经网络可以对有效的数据做提取的能力，因此说神经网络对于在线学习的发展法相是一个大趋势。决策树是一种分类模型和预测模型，树结构表示一个群决策，每个树代表一个自变量和因变量。它由三部分组成，包括决策点、状态节点和结果节点，具体的决策树方法包括分类回归树和卡方自动交互检测。

(二)聚类分析

聚类分析主要是对数据进行聚类，也就是分组，这点与分类分析类似。但是，聚类的训练数据和测试数据没有事先做过类别标签，完全由

聚类算法根据数据特征进行聚类，因此聚类属于无监督学习。有两种聚类算法，分别是层次聚类分析和非层次聚类分析。

(三)回归分析

回归分析是自变量与因变量的关系。对于教育大数据，研究中自变量用自己的信息表示，例如学习时间、班级信息、出勤率等可变参数，因变量可以是考生成绩等变量。

(四)关联规则

关联和相关性经常用于发现变量之间的关系。通过对教育数据集进行相关分析，我们发现，教学中的学习内容及其重要，有益的数据可以起到良好的效果，通过关联规则来获取课堂中学生感兴趣且有用的内容，对提高学生的综合水平有较好的作用。

2.2 教育数据挖掘的工作过程

EDM 的工作过程包含数据分析、特征提取、生成模型和模型评估[22]，如图 2.1 所示。从教育的角度来看，EDM 模型是从数据中挖掘用户信息的过程，再反馈有价值的信息给用户的过程。

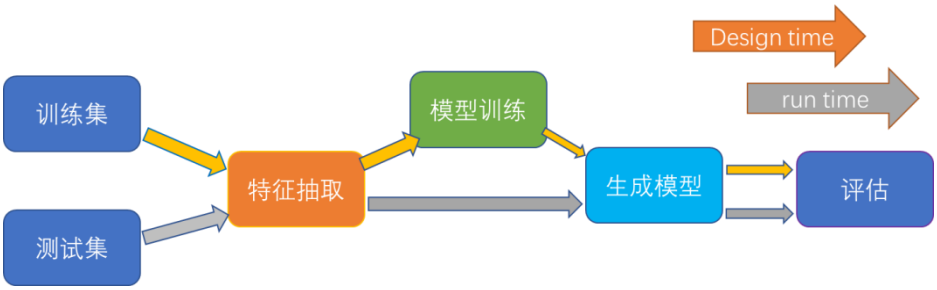

图 2.1 EDM 工作过程

除应用于教育领域外，EDM 的工作过程与通常的数据挖掘应用完全相同。

本文的挖掘过程包括学生学习记录数据预处理，训练模型，用处理好的数据进行预测，对模型性能进行评估。其中，数据预处理主要对收集到的数据进行划分训练集与测试集、特征选择、二次抽样、特征转换、缺失值插补等数据处理操作，得到与数据挖掘算法相匹配的数据并存入数据库中；选择分类算法来训练模型，可以选择常用的 C4.5 决策树算

法和逻辑回归算法；然后使用训练好的模型对未知标签的数据进行预测；模型评估主要对模型进行评价，检验训练所得模型的各种性能参数和模型的泛化能力。

2.2.1 预处理

数据是数据挖掘的基础，要得到好的挖掘效果就需要好的数据。因此，数据预处理（数据准备）是构建挖掘模型的重要步骤。数据预处理发生在已经理解数据之间的关联及内容以后，包括数据清理、数据转换、整合等。据统计，预处理过程一般要占数据挖掘项目的 40%~70% 的时间，在一些复杂项目中甚至占总项目时间的 80%。因此，数据预处理是非常重要的挖掘环节，其处理好坏将直接影响整个项目的效果。

挖掘过程最耗时的就是数据预处理，本文的数据预处理时间非常长。该过程如此耗时的主要原因是：存储在数据仓库中的数据并不一定适合模型的构建及应用。虽然建立数据仓库时，数据转换过程将不同数据源里的数据整合起来并完成必要的清理和格式化，但是用于挖掘的数据可能分布在多个表中，仓库里的数据还可能存在缺失值、无效值和不完整值等，这些都不利于挖掘模型的构建。因此，在构建挖掘模型前，还需要对已有数据进行一系列的转换过程，例如，填补缺失值、替换无效值、整合表数据、计算时间序列、以及行列置换等。

特征选择是指从原始特征中选择最少的特征，使所选特征与类别之间具有最大相关度，特征与特征之间具有最小相关度。

初步进行特征选择的方式有 4 种：(1)用映射或变换的方法把原始特征变换为较少的新特征；(2)从原始特征中挑选出一些最具代表性的特征；(3)根据专家的知识挑选最有影响的特征；(4)用数学的方法进行选取，找出最具分类信息的特征。

本文所用数据量特别大，用作实验的数据有 5.29G，包含 20 个属性，结合训练集和测试集所含属性，很容易排除一些无用属性，选出一些最具代表行的属性。因此最开始进行特征选择的时候，我们采用上述方式（2）、（3）进行初步筛选，再快速筛选无关变量。

2.2.2 模型训练

经过预处理后的数据可以直接用于训练分类模型。通常可以将实验数据分为三个部分，即，一是训练集，用以训练模型；二是验证集，用以选出最优模型；三是测试集，用以评价模型的性能。

划分好训练集、验证集和测试集之后，使用训练集来训练分类模型。本文直接使用 Weka 提供的成熟可视化工具，因此模型训练是比较简单的事。

2.2.3 模型评估

模型评估就是评价训练出来的模型在多大程度上符合实际，一般用评估度量来定量表示。

常见的评估分类器性能的度量通常有：准确率、召回率、精度。

准确率(Accuracy) 的定义是：对于给定的测试数据集，分类器正确分类的样本数与总样本数之比。

精度是针对我们预测结果而言的，它表示的是预测为正的样本中有多少是真正的正样本。那么预测为正就有两种可能了，一种就是把正类预测为正类(TP)，另一种就是把负类预测为正类(FP)[55]，也就是

$$P=TP/(TP+FP) \quad (2.1)$$

召回率是针对我们原来的样本而言的，它表示的是样本中的正例有多少被预测正确了。那也有两种可能，一种是把原来的正类预测成正类(TP)，另一种就是把原来的正类预测为负类(FN)[56]。

$$R=TP/(TP+FN) \quad (2.2)$$

均方根误差（RMSE，Root Mean Square Error）亦称标准误差，是观测值与真值偏差的平方和与观测次数比值的平方根，常用于衡量观测值同真值之间的偏差。

除了这些评估指标之外，还有一些其他指标，如 F-score，Kappa，AUC[57]等。

2.3 分类算法介绍

本文使用 C4.5 决策树算法、逻辑回归、支持向量机 SVM 算法、贝叶斯网络算法和 BP 神经网络算法这几种简单有效的分类算法。决策树算法的容易理解和解释，能够同时处理数据型和标称型特征，易于实现；逻辑回归算法速度快，比较适合本文的二元分类问题，逻辑回归模型简单易于理解，可以直接看到各个特征的权重；支持向量机 SVM 算法，它起源于逻辑回归，是一种二类分类模型，其基本模型定义为特征空间上的间隔最大的线性分类器，其学习策略便是间隔最大化，最终可转化为一个凸二次规划问题的求解，比较适合本文的二元分类问题；贝叶斯网络就是在信息不完备的情况下，通过可以观察变量推断不可观察的变量，与本文预测情况类似，适合本文预测使用；BP 神经网络算法，能学习和存贮大量的输入--输出模式映射关系，输入学习样本，使用反向传播算法对网络的权值和偏差进行反复的调整训练，本文涉及到多个属性间的互相关系，通过 BP 网络算法反复训练可以得到比较理想的模型，然后进行预测。

2.3.1 C4.5 决策树算法

C4.5 算法使用信息增益率作为分裂规则（需要用信息增益除以该属性本身的熵），连续属性的分裂只能二分裂，离散属性的分裂可以多分裂，比较分裂前后信息增益率，选取信息增益率最大的。

本文预测模型以决策树算法 C4.5 和逻辑回归算法对学生做对下一道题进行预测，下面先介绍 C4.5 算法。

C4.5 主要是在 ID3 算法的基础上改进得出的，主要可以理解成信息增益改成了信息增益率，具体从如下几个方面做了改进：

（1）能够处理连续型属性和离散型属性的数据。

（2）能够处理有缺失值的数据。

（3）以信息增益率做为属性选择的指标。

（4）对树做剪枝处理，防止决策树过拟合。

信息增益率将分裂信息作为分母，属性取值数目越大，分裂信息值越大，从而部分抵消了属性取值数目所带来的影响。

2.3.2 逻辑回归

逻辑回归（Logistic Regression）算法的原理可以简单的描述为这样的过程：

(1) 构造预测函数

虽然名字里带“回归”，但是逻辑回归实际上是一种分类方法，用于解决二元分类问题（即只有两种类别标签）。

构造预测函数为：

$$h_\theta(x) = g(\theta^T x) = \frac{1}{1+e^{-\theta^T x}} \quad (2.3)$$

$h_\theta(x)$函数的值有特殊的含义，它表示结果取 1 的概率，因此对于输入 x 分类结果为类别 1 和类别 0 的概率分别为：

$$\mathrm{P}(\mathrm{y}=1|\mathrm{x};\theta) = h_\theta(x)$$

$$\mathrm{P}(\mathrm{y}=0|\mathrm{x};\theta) = 1 - h_\theta(x) \quad (2.4)$$

(2) 构造代价函数：

$$\mathrm{Cost}(h_\theta(x),\mathrm{y}) = \begin{cases} -\log(h_\theta(x)) & if\ y=1 \\ -\log(1-h_\theta(x)) & if\ y=0 \end{cases} \quad (2.5)$$

$$\mathrm{J}(\theta) = \frac{1}{m}\sum_{i=1}^{m} Cost(h_\theta(x^i), y^i) \quad (2.6)$$

(3) 梯度下降法求 J(θ)的最小值

求 J(θ)的最小值可以使用梯度下降法，根据梯度下降法可得 θ 的更新过程：

$$\theta_j := \theta_j - \alpha\sum_{i=1}^{m}(h_\theta(x^i) - y^i)x_j^i, (j = 0, \cdots, n) \quad (2.7)$$

(4) 梯度下降过程向量化， vectorization 后 **θ** 更新写成：

$$\theta := \theta - \alpha \cdot (\frac{1}{m}) \cdot x^T \cdot (g(x \cdot \theta) - y) \quad (2.8)$$

对于逻辑回归的损失函数构成的模型，可能会有些权重很大，有些权重很小，导致过拟合（就是过分拟合了训练数据），使得模型的复杂度提高，对未知数据的预测能力即泛化能力较差。常用正则化方法来解决过拟合问题。

2.3.3 支持向量机 SVM 算法

支持向量机(Support Vector Machine，SVM)是 Corinna Cortes 和 Vapnik 等于 1995 年提出的[58]，它在解决小样本、非线性及高维模式识别中表现出许多特有的优势，并能够推广应用到函数拟合等其他机器学习问题中。支持向量机包括核技巧，使它成为实质上的非线性分类器。支持向量机的学习策略就是间隔最大化，形式的化为求解凸二次规划问题，也等价于正则化的和业务损失函数的最小化问题。

2.3.4 贝叶斯网络算法

贝叶斯分类算法是统计学的一种分类方法，它是一类利用概率统计知识进行分类的算法。贝叶斯网络主要用于概率推理及决策，具体来说，就是在信息不完备的情况下通过可以观察随机变量推断不可观察的随机变量，并且不可观察随机变量可以多于以一个，一般初期将不可观察变量置为随机值，然后进行概率推理一个贝叶斯网络定义包括一个有向无环图（DAG）和一个条件概率表集合[59]。DAG 中每一个节点表示一个随机变量，可以是可直接观测变量或隐藏变量，而有向边表示随机变量间的条件依赖;条件概率表中的每一个元素对应 DAG 中唯一的节点，存储此节点对于其所有直接前驱节点的联合条件概率。

2.3.5 BP 神经网络算法

BP（Back Propagation）网络是 1986 年由 Rumelhart 和 McCelland 为首的科学家小组提出[60]，是一种按误差逆传播算法训练的多层前馈网络，是目前应用最广泛的神经网络模型之一。BP 网络能学习和存贮大量的输入--输出模式映射关系，而无需事前揭示描述这种映射关系的数学方程。它的学习规则是使用最速下降法，通过反向传播来不断调整网络的权值和阈值，使网络的误差平方和最小。BP 神经网络模型拓扑结构包括输入层（input）、隐层(hide layer)和输出层(output layer)[61]。

2.4 本章小结

本章主要介绍了教育数据挖掘的相关理论，包括教育数据挖掘的概念、教育数据挖掘的挖掘方法和工作过程，并对数据预处理、模型训练和评估作了介绍，还介绍了本文使用的几种分类模型。

第三章 数据挖掘过程

教育数据挖掘过程包括数据的预处理、预测、评估几个部分。本章的数据挖掘方法借鉴伊朗学生 Yasser Tabandeh 和 Ashkan Sami 的论文《Classificationof Tutor System Logswith High Categorical Features》，这两个学生组成的团队 Y10 使用有限的计算机设备取得了优异的成绩。其数据预处理过程是先对学习记录进行特征选择，然后对选取的特征进行数值化，最后再对所得数据进行二次抽样得到最终数据。本文做法是先对学习记录进行特征选择，然后进行二次抽样，最后再进行数值化得到最终数据。这样把二次抽样放在特征选择之后，好处是减少计算量与处理时间，最终模型结果稍强于 Y10 团队，比开发设计 LIBSVM 模式识别与回归软件包的台湾大学林智仁(Lin Chih-Jen)教授带领的 25 人冠军团队 NTU 稍差。

3.1 数据集描述

本文数据集使用 KDD CUP 2010 提供的公开数据集，该数据集由美国学生使用在线学习系统进行数学学习的交互日志构建而得。

KDD CUP 2010 数据集由一个开发数据集（Development Data Sets）和一个挑战数据集（Challenge Data Sets）组成。每个数据集都分为训练部分（Training Portion）和测试部分（Test Portion）。其中开发数据集中学生答对与否的标签（Performance Lables）是公开的，而挑战数据集已将此标签隐藏起来。KDD CUP 2010 提供所示的三个开发数据集和所示的两个挑战数据集，分别如表 3. 1 和表 3. 2 所示。

表 3. 1 开发数据集

数据集	学生数	步骤数	文件名
Algebra I2005-2006	575	813361	algebra_2005_2006.zip
Algebra I2006-2007	1840	2289726	algebra_2006_2007.zip
Bridge to Algebra 2006-2007	1146	365871	bridge_to_algebra_2006_2007.zip

表 3.2 挑战数据集

数据集	学生数	步骤数	文件名
AlgebraI2008-2009	3310	9426966	algebra_2008_2009.zip
Bridge to Algebra 2008-2009	6043	20768884	bridge_to_algebra_2008_2009.zip

上述数据集都是学生在计算机辅助导学系统的学习记录。学生学习并解答导学系统中的题目，每次人机交互以日志的方式记录下来。挑战数据集分为 AlgebraI2008-2009 和 BridgetoAlgebra 2008-2009 两种，论文简称 A89 和 B89，A89 的属性有 22 个，B89 有 20 个，但是 B89 的数据集更大，更具有挑战性，因此论文选用 B89 作为实验对象。

bridge_to_algebra_2008_2009.zip 压缩文件的大小为 439M，包含三个 txt 文件，其中，bridge_to_algebra_2008_2009_train.txt 文件是训练集，大小为 5.29G，共有 20012498 条记录；bridge_to_algebra_2008-_2009_test.txt 文件是测试集，大小为 131M，共有 756386 条记录；bridge_to_algebra_2008_2009_submission.txt 文件是提交文件，大小为 7.28M，共有 756386 条数据，只包括 Row 和 CorrectFirstAttempt 两个属性。

B89 训练集和测试集都有 20 个属性和一个目标属性，用制表符（Tab）分割各个属性，它们的含义如表 3.3 所示。

表 3.3 B89 各属性含义说明

序号	属性名称	含义说明
1	Row（行标号）	对于挑战数据集，行标号是对每个文件中的行进行重新编号，并不是直接从原始数据集中得到。
2	Anon Student Id（学生匿名 ID）	学生匿名唯一标识符。

续表 3.3

序号	属性名称	含义说明
3	Problem Hierarchy（问题层次）	该问题的所属的课程层级。由单元名（Unit）和章节名（Section）组成，用逗号分隔开两个部分。
4	Problem Name（问题的名称）	问题的唯一标识符。
5	Problem View（问题遇到的次数）	到现在时间为止，学生遇到该问题的总次数。
6	Step Name（步骤名称）	代表所有问题中各个步骤的称谓。每个问题都可以分解为一个或多个步骤。
7	Step Start Time（步骤开始时间）	该步骤起始时间，可为空（Null）。
8	First Transaction Time（第一次事务时间）	步骤首次处理时间。
9	Correct Transaction Time（正确的事务时间）	步骤正确作答时间，可为空（Null）。
10	Step End Time（步骤结束时间）	步骤最后一个事务结束的时间。
11	Step Duration(sec)（步骤持续时间）	一个步骤的持续时间，由一个步骤所内含的所有事务的持续时间的和构成。若开始时间为空，可为空。单位为秒。
12	Correct Step Duration(sec)（正确步骤持续时间）	正确步骤持续时间，单位为秒。
13	Error Step Duration(sec)（错误步骤持续时间）	如果第一次尝试错误（不正确的尝试或请求提示都记为错误），步骤持续的时间。单位为秒。

续表 3.3

序号	属性名称	含义说明
14	Correct First Attempt（第一次尝试正确）	在导学系统中，学生在一个步骤中首次尝试是否正确，正确记为 1，错误为 0。
15	Incorrects（错误总数）	一个步骤中学生做出的错误尝试的总数。
16	Hints（提示总数）	学生请求提示的总次数。
17	Corrects（正确总数）	一个步骤中学生做出的正确尝试的总数（同一个问题遇到多次才会增加）。
18	KC(SubSkills)（KC 子技能）	用于问题的特定技能，若存在的话。解题中识别到的有效知识点。一个步骤可以有不同知识点，用“~~”隔开。
19	Opportunity(SubSkills)（子技能机会）	学生遇到 KC 项所列出的知识点次数，每次遇到该知识点增加一次。若有多个知识点，则数据以“~~”隔开。
20	KC(KTracedSkills)（KC 知识追踪技能）	该知识组件用于在线学习系统，格式与 KC(SubSkills)保持一致。
21	Opportunity(KTracedSkills)（知识追踪技能机会）	格式与 Opportunity(SubSkills)一致，只针对 KC(KTracedSkills)知识组件计数。

挑战数据集的训练部分提供所有的值，但测试部分没有提供如下属性的值：Step Start Time、First Transaction Time、Correct Transaction Time、Step End Time、Step Duration(sec)、Correct Step Duration(sec)、Error Step Duration(sec)、Correct First Attempt、Incorrects、Hints、Corrects 和 Corrects。究其原因是因为，我们要预测的答案是 Correct First Attempt，所以不会提供；如果提供其他上述值，几乎可以直接得到答案。例如，若 Hints 不为 0，说明该学生已经请求提示，Correct First Attempt 必为 0。因此没有提供这些值。

3.2 数据预处理

B89 的数据文件太大，训练文件有 5.29GB，共有 20012498 条记录，因此在一般配置的计算机中，由于内存的原因，几乎不能一次加载完成。即使能够加载，分类器也不可能直接对其进行学习和预测。因此，首先要进行预处理，通过特征选择、二次抽样，并进行特征变换，以抽取出合适的特征，降低数据规模，使这些数据能够被普通计算机处理。

数据预处理历来是数据挖掘的重点，最终的挖掘效果大部分都和预处理密切有关。预处理有一些特别的技术诀窍，论文采用的方式是利用数据库系统数据结构化且统一管理，查询迅速、准确的优势，编写非过程化语言的 SQL 语句进行预处理。

3.2.1 特征选择

特征选择就是要删除一些对预测作用不大的属性。

首先对数据集进行分析，经过对照开发数据集和挑战数据集，可以看到有一些属性没有在测试集中出现。具体有 Step Start Time、First Transaction Time、Correct Transaction Time、Step End Time、Step Duration(sec)、Correct Step Duration(sec)、Error Step Duration(sec)、Correct First Attempt、Incorrects、Hints、Corrects 和 Corrects 十个属性。因此，在第一次特征选择的时候，将这些属性直接删除。另外，还需要删除 ProblemHierarchy 属性，因为该属性包含单元名称和章节信息，完全依赖于 ProblemName 属性。最后，将 ProblemName 和 StepName 两个属性合并成为一个 ProblemStep 的属性，以增加建模的准确性和速度。

第一次特征选择后，剩下的属性有 Anon Student Id、ProblemStep、ProblemView、KC(SubSkills)、Opportunity(SubSkills)、KC（KTracedSkills）、Opportunity(KTracedSkills)和 Correct First Attempt，一共 8 个属性。

使用 Weka 工具检查上述属性与目标属性的相关性，删除一些属性对预测影响不大的属性，最终选择出如下 4 个属性：Anon Student Id、ProblemStep、KC(KTracedSkills)和 Correct First Attempt。实际还包含属性 Row，行号属性有助于将来进一步预处理。

3.2.2 二次抽样

经特征选择后的 B89 数据集还是相当大，约有 1.49G。一般电脑难以处理这样的大数据，需要进一步缩小数据规模。

二次抽样出约 1/7 的数据，进一步降低数据规模。抽样后数据量也相当大，约有 2858928 条数据，200M 大小。

二次抽样是随机抽取部分数据。按照常识，从时间上越接近测试样本的学生表现越能帮助预测学生的未来表现，因此，抽取与测试样本相隔较近的样本显然比随机抽取优越，下一章讨论这个问题。

3.2.3 特征变换

由上述过程得到的三个重要属性（StudentId、ProblemStep、KC）都是标称型，取值数量都非常大。大多数包括决策树在内的分类器都不容易用这类数据来训练模型。在有限的时间和硬件资源上，很难将决策树算法运用到这样的数据上。另外，逻辑回归算法更容易处理数值型属性。因此，有必要通过特征变换算法将标称型属性转换为数值型属性。

经过多方比较，最终采纳的特征变换算法如算法 3.1 所示。

算法 3.1 特征变换算法
对于属性集里的每一个标称型属性 Fc
增加一个新的数值型属性 Fn 到属性集
对于 Fc 中的每一个取值 v
N=包括 v 的所有样本数量
Np=包括 v 且为正例的样本数量
A=Np/N（v 中正例的百分比）
把 A 填入 Fn
在属性集里移除 Fc

采用算法 4.1 进行特征变换，创建如下三个属性以取代原来的对应属性。

（1） StudentChance：由 StudentId 属性转换而来，表示一个学生解答问题的能力，即，该学生可能正确回答问题的概率。

（2） PSChance：由 ProblemStep 属性转换而来，表示一个问题步骤的难易程度，即，该步骤被正确解答的概率。

（3） KCChance：由 KC 属性转换而来，表示某个知识点的难易程度，即，包含该知识点的步骤被正确解答的概率。

上述三个属性非常容易理解，学生答题是否正确就是由学生水平（StudentChance）、问题步骤的难度（PSChance）和知识点难度（KCChance）这三个因素决定。

3.3 模型训练

使用 Weka 工具对预处理后的训练集进行训练，训练后的决策树模型如图 3.1 所示。从图中可以看到，学生水平（StudentChance）最为重要，根据其值是否小于等于 0.4051 分到左子树和右子树，然后再判断知识点难度（KCChance），根据学生水平和知识点难度可判断学生能否做对题目。

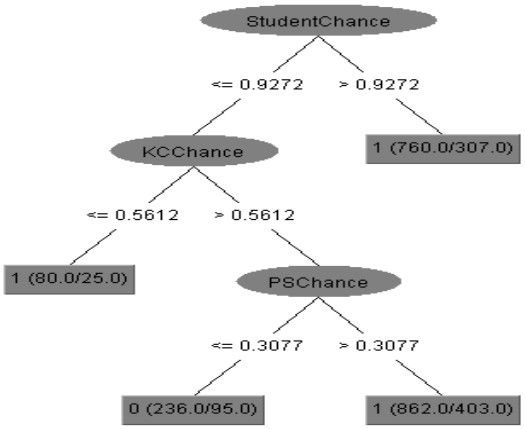

图 3.1 部分决策树

决策树对未知样的分类过程是，自决策树根节点开始，自上向下沿某个分支向下搜索，直到到达叶节点，叶节点的类别标签就是该未知样本的类别。

按照决策树的分类过程，自根节点 StudentChance 开始，向下搜索，如果遇到类别标签为 1 的叶节点，可推断未知样本的类别标签是 1。

3.4 模型评估

模型评估可评价分类模型的性能。常用十折交叉验证方法进行模型评估。

逻辑回归（Logistic）分类器和 C4.5 分类器结果对比如表 3.4 所示。

表 3.4 Logistic 分类器和 C4.5 分类器结果对比

分类器	分类准确率	RMSE
Logistic	88.8591%	0.2856
C4.5	88.9495%	0.2848

从表中可以看出，C4.5 决策树在分类准确率上比 Logistic 稍好，RMSE 也是 C4.5 效果稍小，C4.5 的效果更好。

本文方法与竞赛团队 Y10 和冠军团队 NTU 对比如表 3.5 所示。可以看出，本文方法略好于 Y10 的结果，比冠军团队 NTU 稍差。

表 3.5 本文方法与竞赛团队 Y10、NTU 结果对比

	Y10	NTU	本文方法
C4.5 RMSE	0.2921	0.2711	0.2848
LogisticRMSE	0.2933	0.2729	0.2856

鉴于本方法仅使用三个属性和有限的计算设备，RMSE 0.2848 的成绩与冠军团队 RMSE 0.2711 差距很小，成绩已经非常优异。

3.5 本章小结

本章主要讲述所使用的数据集、数据预处理、模型训练和模型评估，并将本文方法与竞赛团队 Y10，NTU 结果做了对比，预测效果很好。

受限于个人电脑的处理能力，仅仅使用了 20 个属性中的三个属性和一个目标属性，分类的准确率就能达到 88.9495%，如果再对分类器参数进行优化，或者采用集成学习算法，估计还有提升空间。

第四章 改进的挖掘过程

上一章主要实现了团队 Y10 的挖掘方法。本章通过讨论分析采用更接近实际的划分训练集与测试集方法，重新对数据集进行预处理，再进行预测，并比较不同算法模型的优劣。发现 C4.5 算法预测结果稍差，但更接近实际，而且效果最好。本章还对模型的泛化能力进行评估，证明模型有很好的泛化能力。

4.1 改进的思路

上一章的模型评估采用十折交叉验证方法，也就是把全部数据集划分为大致相等的十份，其中一份用作测试集，其余九份用作训练集。把目光集中在用作测试集的一份样本上，在预处理时，采用特征变换算法已经隐含了测试集中的预测答案，即算法中计算 A=Np/N，N 包含的是全部十份的样本总数，Np 也是全部十份中为正例的样本数量，Np 的计算已经将测试集中的样本分布计算到预处理后的属性中。这时，测试集已经变成提前知道答案的测试集，其预测结果肯定会好于真实结果。但是在真正对挑战测试集中的测试部分进行预测时，由于答案已经隐藏，无法预先看到答案，其预测准确率肯定会差一些。

为了保证对分类预测性能评估的真实性，本章对上述方法进行了两点改进。第一，严格按照 KDD CUP 2010 竞赛组织者抽取测试集的方式来抽取测试集，将测试集中的答案隐藏，保证对分类器评估的有效性和合理性；第二，不采用完全随机的数据抽样，而是考虑数据的时效性，只抽样离测试样本最近的问题步骤。因为学生做题是有先后时间顺序的，对于学生的答题能力肯定是随着时间的推移越来越强，因此只抽取离测试样本最近的问题步骤，最能反映当前学生的状态，能够更准确预测学生的答题是否正确。

4.2 更接近实际的挖掘过程

首先研究 KDD Cup 2010 竞赛组织者抽取测试集的方式，严格按照该方式从数据集中抽取训练集和测试集；然后用训练集训练分类模型，再用测试集来评估学习到的模型。这样的评估结果更接近实际。

4.2.1 划分训练集和测试集的思路

KDD CUP 2010 的训练文件和测试文件是按照图 4. 1 所示的方法来划分的。图中的每一根水平实线代表学生的每一条答题记录，遵循时间的先后顺序。竞赛的测试文件是由每个学生每个单元的最后一个问题组成。

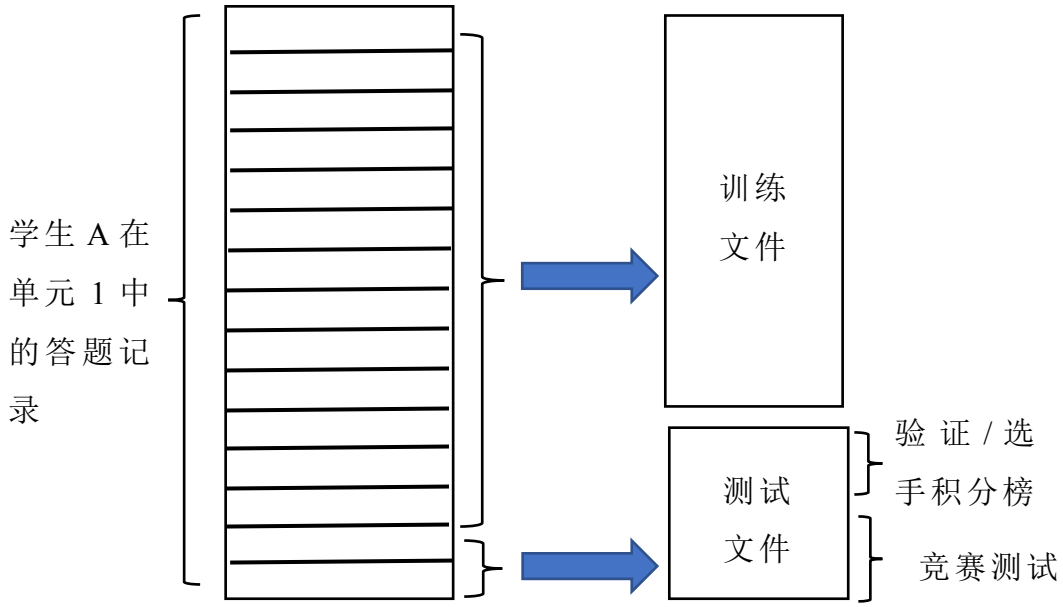

图 4. 1 KDD CUP 2010 划分训练文件和测试文件的方法

由于 KDD CUP 2010 的测试文件没有提供 Correct First Attempt 值，无法通过该测试文件来评估分类模型的性能。合理的思路是从训练文件中划分出单独的训练集和测试集，训练集用于训练模型，测试集用于评估模型。

按照这种思路划分，将原训练文件中每个学生在每个单元练习的最后一个问题抽取出来，组成用于验证模型的测试集。此时，如果将剩下

的所有问题都作为训练集，显然训练集过大，一般计算机无法处理。直接能想到的解决方案有两种：一是随机抽样，随机抽取剩下的部分问题作为训练集；二是考虑时间顺序，只抽取每个单元最后几个问题组成的训练集，扔掉单元内离测试问题较远的问题。从常识上来判断，第二种方式更有优势，因为知识学习是一个循序渐进的过程，时间越近的多个问题的知识相关程度越高。

综上所述，最终采纳的方案是将每个学生每个单元的最后一个问题抽取出来组成测试集，将剩余数据中每个学生每个单元的最后三个问题抽取处理组成训练集。

4.2.2 新的预处理过程

新的训练集和测试集划分方式要考虑单元，按照时间先后次序分别将每个单元的最后几个问题划分到训练集和测试集，因此要全部重做预处理。

总体来说，训练集和测试集的划分逻辑较为复杂，很难用 SQL 语句实现，更好的办法是用 Java 语言编程实现。本文使用 Java 语言实现训练集和测试集的划分。其流程图如图 4.2 所示。

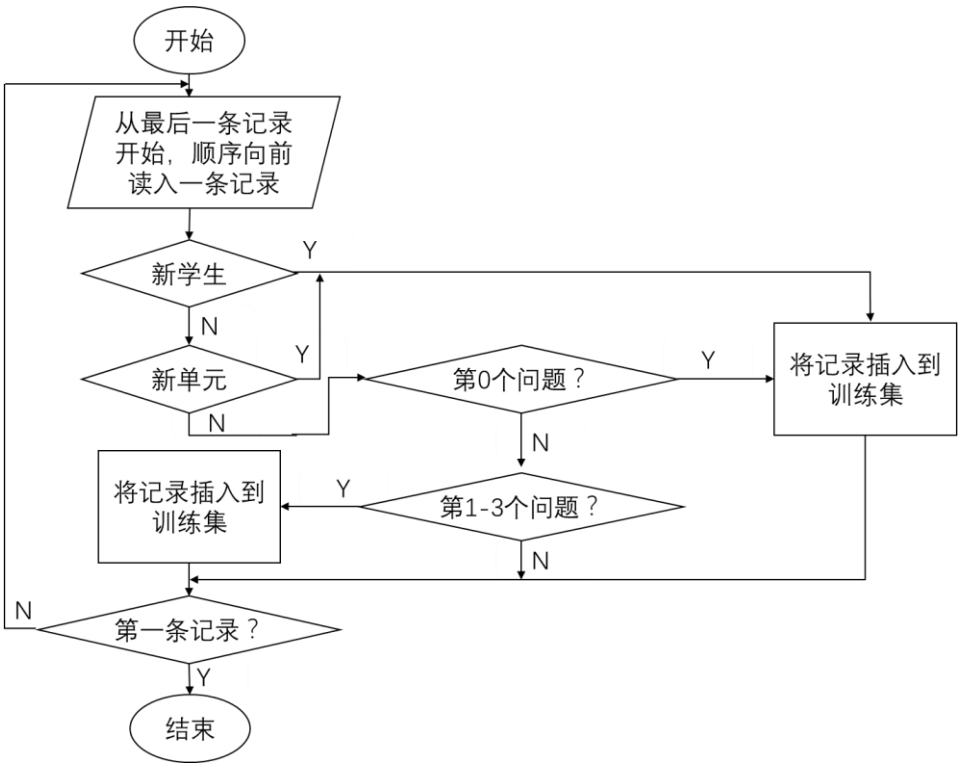

图 4.2 划分训练集和测试集的流程图

原训练文件按照时间顺序排列，从后向前容易实现数据拆分。从流程图可以看到，总体的目标是将第 0 个问题（逆序的第 0 个问题就是顺序后的第一个问题）的步骤划分至测试集，将第 1 至第 3 个问题的步骤划分至训练集，丢弃其他问题的步骤。当然，该流程图省略了很多细节，例如，必须记录学生姓名，才能判断新的记录是否是新学生的学习日志；必须对问题有一个计数器，才能判断到底将记录插入到测试集，或是训练集，或是丢弃。

最终实现的 Java 代码参见附录 B。

由于数据量非常大，本人计算机需要十几个小时才能完成处理。

最终的训练集和测试集的样本数分别为 2246653 和 774378，大约占原数据的 15.19%。

4.2.3 缺失值插补

经过前面的训练集测试集划分和预处理之后，将 KDD CUP 2010 训练集属性转换为数值型，这时会发现一个问题：测试集有些数据没有在训练集中出现过，无法通过特征变换转换为数值型。

经过检查，测试集中一共有 774378 个样本，其中，有 3367（约 0.43%）的学生没有在训练集中出现；有 32361（约 4.2%）的问题步骤没有在训练集中出现；有 28（约 0.0036%）的知识组件没有在训练集中出现。

这些测试集中的缺失值对分类器评估无疑会带来负面影响。如果 StudentChance 为空，说明这些学生根本就没有在训练集中出现过。试想一下，要对一个一无所知学生成绩进行预测，只能靠猜，依靠猜测的测试样本准确率肯定很低。再看 PSChance，该值为空带来的影响更大，因为缺失比例更高。换句话说，通过训练，我们能够了解某个步骤能被正确解答的概率，在一定程度上表示该步骤的难度。如果对某个步骤的难度一无所知，要预测也只能靠猜。KCChance 也类似。

那么扩大抽样范围能否解决这个问题？首先，哪怕将全部除测试集以外的剩余步骤都划分到训练集，还是有 3367 个步骤的 StudentChance 为空，因为那些学生就只完成了单元的最后一个也是唯一一个问题，这个问题的步骤已经全部划分到测试集，没法在训练集中重复使用。经分

析后，认定扩大抽样范围并不能解决这个问题。但是，扩大抽样范围对PSChance 和 KCChance 肯定有益，例如，将原算法划分到训练集的最后三个问题扩大至十个问题，也许有一些问题步骤就会在训练集中出现。扩大抽样范围带来性能提升好处的同时又会增加计算的复杂度，到底是好是坏只能通过实验才能说明。

最简单的办法就是直接将这些缺失值的测试样本删除。但是相对于简单删除不完全的样本，用最可能的值插补缺失值丢失的信息比较少。缺失值插补的方法有多种，最简单的是将缺失值替换为均值，称为均值插补。

具体做法是，将测试集中为空的值替换为训练集中相应 Chance 的均值。最后完成全部预处理工作。

4.2.4 挖掘过程和结果

启动 Weka 工具，使用逻辑回归 Logistic 分类器对预处理后的数据集进行挖掘，得到的分类准确率为 87.8009%，均方根误差 RMSE 为 0.3095。效果比前面稍差，但更符合实际。

使用 C4.5 分类器对预处理后的数据集进行挖掘，得到的分类准确率为 88.2308%，均方根误差 RMSE 为 0.3083，比逻辑回归稍好。

同时分别使用贝叶斯 BayesNet 分类器、BP 神经网络分类器 Back Propagation 和支持向量机 SVM 分类器对预处理后的数据集进行挖掘，与上述结果进行对比。

使用贝叶斯 BayesNet 分类器对预处理后的数据集进行挖掘，得到的分类准确率为 86.5974%，均方根误差 RMSE 为 0.3127。效果比逻辑回归和决策树 C4.5 效果都差。

使用 BP 神经网络 Back Propagation 分类器对预处理后的数据集进行挖掘，得到的分类准确率为 87.4799 %，均方根误差 RMSE 为 0.3127。效果比贝叶斯好，但比逻辑回归和决策树 C4.5 效果都差。

使用支持向量机 SVM 分类器对预处理后的数据集进行挖掘，得到的分类准确率为 87.6352%，均方根误差 RMSE 为 0.3109。效果比 BP 神经网络稍好，但比逻辑回归和决策树 C4.5 效果都差。

采用新方法预测结果如表 4. 1 所示。

表 4. 1 采用新方法预测结果对比

算法	分类准确率	RMSE
BayesNet	86.5974%	0.3210
Back Propagation	87.4799 %	0.3127
SVM	87.6352%	0.3109
Logistic	87.8009%	0.3905
C4.5	88.2308%	0.3083

从表中可以看出 C4.5 预测效果最好，虽然效果比前面稍差，但更符合实际，而且它是所有模型中预测效果最好的。

4.3 检验模型泛化能力

机器学习的目的是使学习到的模型不仅对已知数据有很好的预测能力，而且对未知数据也能产生正确的输出。通过训练集训练得到的模型在多大程度上能够对新实例预测正确输出称为泛化。

在目前实验中，测试集都是训练集中没有的新实例，预测准确率实际评估的就是模型的泛化能力。现在，让我们再考虑一种极端的情形，只预测那些在训练集中没见过的学生，或是没见过的题，或是没见过的知识组件实例，看看我们的模型到底怎么样。前文描述过，这类情形纯粹依靠猜测，如果在这样的极端情况下还能达到一定的准确率，我们大致可认为模型的泛化能力很好。

由于目标实例的一个或多个属性缺失，只能使用训练集的均值进行插补。具体方法使用 4.2.3 节所述的缺失值插补方法，得到极端的测试集。使用模型对极端的测试集进行预测，将预测结果与真实结果比较，得到的分类准确率低于目前结果，符合预期。实验结果如表 4.2 所示。

表 4.2 模型评估结果

算法	分类准确率
Logistic	82.7147%
C4.5	82.7091%

综上所述，论文所构建模型的有很好的泛化能力，即对于训练集中没有见过的学生、没有见过的问题步骤，或者没见过的知识组件的极端测试样本，模型的预测准确率仍然能够超过 82%。

4.4 小结

本章主要讲述如何对挖掘过程进行改进。采用更接近实际的划分训练集与测试集方法，重新对数据集进行预处理，发现预测结果稍差，但更接近实际，并与其它分类器结果比较，发现 C4.5 决策树分类器效果最好。本章还通过实验评估模型的泛化能力，证明论文所构建模型的有很好的泛化能力。

第五章 实际应用

本章通过对学校“教育在线”网络教学平台中实际存在的学生数据进行处理，主要是集中在对学生做题等相关教学行为的教育数据分析，进而对学生在测试做题的过程中进行预测，验证本文上一章中所提出的更接近实际的挖掘过程的可行性和正确性。

5.1 教育在线数据集描述

“教育在线”网络教学平台是由优慕课在线教育科技（北京）有限责任公司（前身为清华发现教育技术研究所）所研发的一套在线教育教学软件，用于支持基于传统课堂的网络辅助教学模式，本章测试所用到的数据为“教育在线”系统后台数据库中原始数据，且进行相关必要的处理以满足程序的需求。本章节中测试实验的数据为一个班级中 28 名学生所有做题的数据，包含 346 个题目及 18 个知识点，最后形成 1000 条答题数据。

5.1.1 StudentChance 数据集

StudentChance 数据集反应的是学生解答问题的能力，即该学生可能正确解答问题的概率，通过对“教育在线”系统的调研分析，其中能表明 StudentChance 数据集主要涉及到如表 5.1 的相关数据表。

表 5.1 StudentChance **数据集涉及到的相关原始数据情况**

数据表	描述	主要字段	描述
THEOL_USER	用户表	USERNAME	用户编号
EOL_TEST_ANSWER	测试题目表	ANSWERID	题目编号
EOL_TEST_ANSWER_QUESTION	测试题目答题表	mark	得分

编程实现对原始数据的抽取，得到本文所需的 StudentChance 数据集，相关编码简写如下。

```
//统计学生正确答题总数
SELECT count(*) FROM 测试题目答题表 where ANSWERID
in (测试题目表)
where USERID = (用户表)
and mark > 0";
```

5.1.2 PSChance 数据集

PSChance 数据集反应的是一个问题步骤的难易程度，即该步骤被正确解答的概率，通过对“教育在线”系统的调研分析，其中能表明 PSChance 数据集主要涉及到如表 5.2 相关数据表。

表 5.2 PSChance 数据集涉及到的相关原始数据情况

数据表	描述	主要字段	描述
EOL_LESSON	课程表	numb	课程编号
EOL_TEST	测试表	CATEID	测试编号
EOL_TEST_ANSWER	测试题目表	ANSWERID	题目编号
EOL_TEST_ANSWER_QUESTION	测试题目答题表	mark	得分

编程实现对原始数据的抽取，得到本文所需的 StudentChance 数据集，相关编码简写如下。

```
//统计单个题目被学生做对正确的次数
SELECT count(*) FROM 测试题目答题表 where ANSWERID
in (测试题目表  where TESTID
in (测试表 where CATEID = (课程表  where numb = “该课程编号”) )
and mark > 0";
```

5.1.3 KCChance 数据集

KCChance 数据集反应的是某个知识点的难易程度，即包含该知识点的步骤能被正确解答的概率，通过对“教育在线”系统的调研分析，其中能表明 KCChance 数据集主要涉及到如表 5.3 的相关数据表。

表 5. 3 KCChance 数据集涉及到的相关原始数据情况

数据表	描述	主要字段	描述
EOL_TEST_QUESTION	测试题目选项表	QUESTIONID	题目编号
EOL_TEST_ANSWER_QUESTION	测试题目答题表	QUESTIONID、mark	编号、得分
EOL_QUESTIONBANK	知识点表	KNOWLEDGEKEY	知识点编号

编程实现对原始数据的抽取，得到本文所需的 KCChance 数据集，相关编码简写如下。

```
//统计单个知识点被做对的正确次数
SELECT count(*) FROM 测试题目答题表 where QUESTIONID
in (测试题目选项表  where QUESTIONID )
in (知识点表 where KNOWLEDGEKEY = "知识点编号";
```

5.2 在线预测的设计实现过程

根据上面数据集的收集情况后，需要进行相关处理将数据存储到 mysql 数据库中为程序所用，并进行均值计算和存储，然后将相关数据回填到训练集和测试集中，最后进行预测操作。

5.2.1 数据计算及转存

（1）StudentChance 能力值存储及计算过程中需要考虑学生是否为第一次参与答题，相关编码简写如下。

```
//计算学生 StudentChance 能力值
questions_num = getStudentTotalNumberOfQuestions(studentNum); //取得学生答题总数
if( questions_num == 0 ){   //判断学生是否第一次参与答题
    //学生是第一次答题，取得所有学生的平均答题能力值
} else {
    //取得学生正确答题总数
    //除法计算得到结果
}
```

将上述方法得到的数据转存到数据库中，StudentChance 表中应有的数据字段简要如表 5.4。

表 5. 4 StudentChance **数据表**

字段名	描述
student_num	学生编号
questions_num	学生答题数量
questions_correct_num	学生正确答题数量

（2）PSChance 能力值存储及计算过程中需要考虑该题目是否为第一次参与测试中，相关编码简写如下。

```
 //计算单个题目难易度
questions_num = getTotalNumberOfQuestions (numb,question_id); //取得单个题目出现总次数
if( questions_num == 0 ){    //题目是否第一次参与测试中国答题
    //题目是第一次参与测试中，取得所有题目的平均难易程度
} else {
    //取得单个题目被正确答题的总次数
    //除法计算得到结果
}
```

将上述方法得到的数据转存到数据库中，PSChance 表中应有的数据字段简要如表 5.5。

表 5. 5 PSChance **数据表**

字段名	描述
question_id	题目编号
questions_num	题目被答题总次数
questions_correct_num	题目被正确答题总次数

（3）KCChance 能力值存储及计算过程中需要考虑知识点是否为第一次参与测试中答题，相关编码简写如下。

```
//计算单个知识点难易度
points_number = getTotalNumberOfPoints(studentNum); //取得单个知识点出现总次数
if(points _num == 0 ){    //判断知识点是否第一次参与测试中答题
    //知识点是第一次参与测试中答题，取得所有知识点的平均难易程度
} else {
    //取得单个知识点被正确答题的总次数
    //除法计算得到结果
}
```

将上述方法得到的数据转存到数据库中，KCChance 表中应有的数据字段简要如表 5.6。

表 5. 6 KCChance **数据表**

字段名	描述
point_id	知识点编号
qoints_num	知识点被答题总数
qoints _correct_num	知识点被正确答题总数

5.2.2 均值计算及转存

上一节中所用到的三个均值，考虑到测试中可能有新的学生、题目和知识点的加入，当数据表为空时程序执行上可能会报错误，因而就使用平均值来赋予新的对象，且这样的处理可能更接近实际。

均值计算及转存通过代码实现完成，其相关数据表组成如表 5.7。

表 5. 7 average_value **数据表**

字段名	描述
edit_time	修改时间
student_chance	学生能力平均值
ps_chance	题目难易平均值
kc_chance	知识点难易平均值

5.2.3 数据回填

数据回填就是对训练集和测试集的数据进行特征变换，用前序章节得到的数值型数据对训练集和测试集中的文本数据进行变换，得到数值型属性，进而进行算法的预测。

5.3 实际数据中的模型性能

在上一小节的设计实现过程中，训练集和测试集中都进行的数据的特别变换处理，本小节中使用训练集构建 C4.5 决策树分类器，预测测试集的类别，并输出实际的和预测的类别标签以及分布，最后评价模式的实验性能。

从 mysql 取数据，构建出训练集和测试集后，进行预测处理，相关编码简写如下。

```
//创建 Instances 对象
Instances train_data = new Instances("train_data", train_atts, 0);


//指定分类标签为最后一个属性，即答题正确与否
train_data.setClassIndex(test_data.numAttributes() - 1);
//从 mysql 取数据，构建训练集和测试集
.....
//训练分类器
C4.5 classifier = new C4.5();
classifier.buildClassifier(train_data);


//输出预测
for(....){
    double pred = classifier.classifyInstance(test_data.instance(i)); //得到预测值
    System.out.print(pred); //输出打印预测值
}
```

通过上述操作，使用 C4.5 决策树算法进行模型训练和预测，最后输出打印的结果如图 5.3。

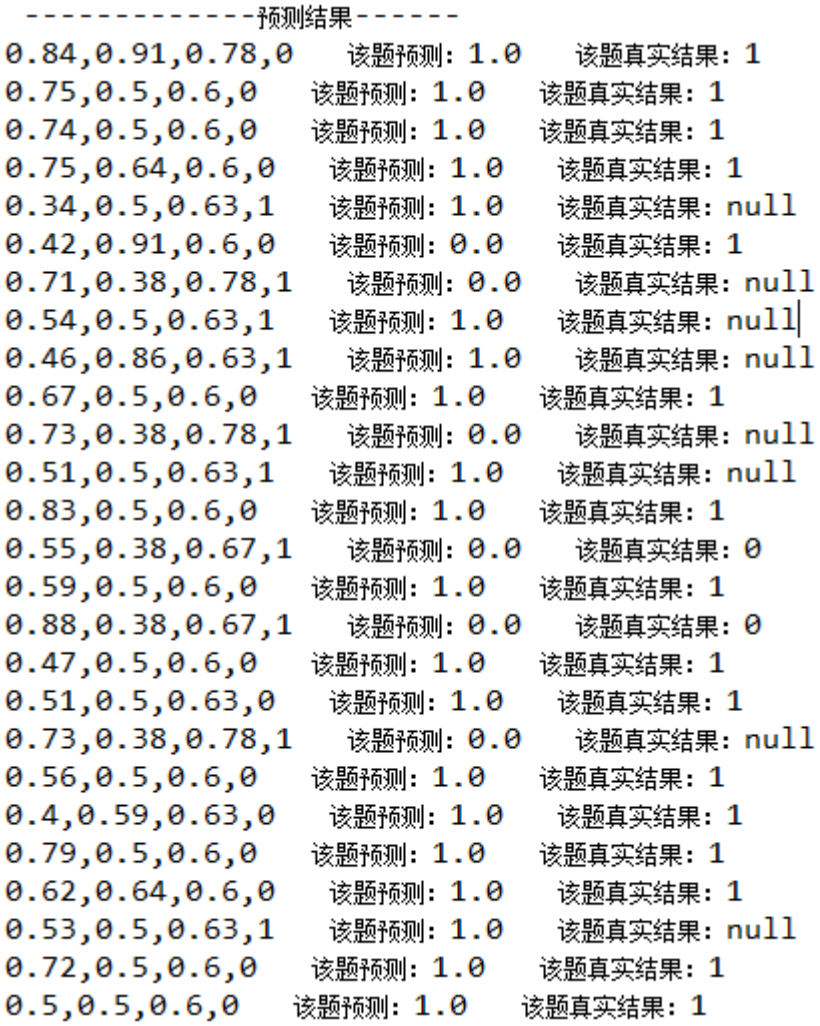

图 5.3 “教育在线”平台测试预测实验结果图

通过比较最终的预测结果，上述实验得到的分类准确率为 82.3301%，由于本文中使用的“教育在线”平台中的数据为一个班级的 28 名同学产生的答题数据，训练范围不是很大，得到上述测试预测实验结果是符合预期的。

5.4 小结

本章通过对学校“教育在线”网络教学平台中实际存在的学生数据进行处理，主要是集中在对学生做题等相关教学行为的教育数据分析，

进而对学生在测试做题的过程中进行预测，得到分类准确率为 82.3301%，验证本文上一章中所提出的更接近实际的挖掘过程的可行性和正确性。

第六章 总结与展望

随着高校在线教育的普及，大量与学生有关数据存在于在线系统中。而随着时间的推移，这个数据量会越加庞大，如何处理这些数据，使这些数据有利于教育教学，是我们迫切要解决的问题。因此研究挖掘学生相关数据势在必行。本章主要对本文的整个研究工作进行了总结，并对下一步工作进行展望。

6.1 总结

本文首先对教育数据挖掘的应用和发展做了简短的介绍，并分析了教育数据挖掘的不足及发展趋势。引出研究学生在线学习相关数据的必要性：通过对学生在线学习的日志数据的分析，得出学生学习能力和知识点难度等相关数据，预测学生的表现，为提高教学质量和优化教育服务。文中详细阐述了数据预处理的过程和自己改进的地方。

本文以 KDD CUP 2010 竞赛提供的公开数据集为研究对象，详细阐述对竞赛数据进行的预处理工作，并对预处理方法进行改进。以更接近实际的获取到学生水平和题目难度信息，从而预测学生能否做对下一道数学题，预测的准确率达到 88.2308%，具有很好的预测效果。这样，可以评估学生对所学知识点的掌握情况，以改进教学与管理方法，提高学生学习效率和教学质量。

综上所述，本文主要完成了以下工作：

（1）对数据挖掘的过程：数据预处理、模型训练和模型评估进行详细介绍。本文重点讲述数据预处理的过程：特征选择、二次抽样，划分训练集与测试集、特征变换和缺失值插补。并将本文方法结果与竞赛团队 Y10 结果做了对比，证明本文方法效果更好。

（2）针对竞赛团队 Y10 的方法，提出一种更符合实际的训练集与测试集划分方法，并用编程实现。并对缺失的值进行均值插补，实现对学生下一题能否做对的预测，并与其它分类器结果比较，发现 C4.5 决策

树分类器效果最好。最后通过实验评估模型的泛化能力，证明论文所构建模型的有很好的泛化能力。

结合团队 Y10 的方法进行实验时，发现在数据预处理过程中，采用的特征变换算法已经隐含了测试集中的预测答案，导致预测结果优于实际结果。为了保证对分类预测性能评估的真实性，本文对上述方法进行了两点改进：第一，严格按照 KDD CUP 2010 竞赛组织者抽取测试集的方式来抽取测试集，将测试集中的答案隐藏，保证对分类器评估的有效性和合理性；第二，不采用完全随机的数据抽样，而是考虑数据的时效性，只抽样离测试样本最近的问题步骤。因为学生的知识点掌握程度是随着时间而增强的，离测试样本越近，越能反映出学生当时的能力，预测也就越准确。重新对数据集进行预处理，并利用编程实现对训练集与测试集的划分，建立模型，实现对学生能否做对下一题的预测，预测的准确率达到 88.2308%，有很好的效果。

（3）根据文中提出的改进方法，用实际数据去验证它的可行性和正确性。以学校“教育在线”网络教学平台中实际存在的学生学习数据为原始数据进行数据挖掘，应用上述改进方法对学生的做题情况进行预测，准确率达到 82.3301%，效果很好，验证了本文所提出的更接近实际的挖掘方法的可行性和正确性。

6.2 展望

通过本文的研究，初步实现了数据挖掘技术在学生在线学习中对学生行为预测的应用。但是，仍然存在着许多问题需进一步研究：

（1）在第五章中采用的样本数较少，只有 1000 条数据，后面继续研究本方法在扩大样本容量后的预测效果。

（2）文中实验受限于个人电脑的处理能力，采用了 20 个属性中的三个属性和一个目标属性。下一步对分类器参数进行优化，扩大选择属性范围，采用集成算法进行处理，研究其对性能提升的影响。

致谢

首先，非常感谢研究生期间我的导师袁梅宇副教授。在刚读研究生的时候，第一次见到袁老师，就觉得他是一个治学严谨，有很高学术水平的老师。在研究生学习期间，随着深入的接触，越来越佩服袁老师治学和为人。袁老师不光治学严谨，学术水平高，而且还教会了我很多为人处世的道理。袁老师因材施教，从来不限制学生学术研究的空间，并在学习上悉心指导，耐心的解答学生在学习中遇到的各种问题。从论文的开题到中期答辩和预答辩，直到最后论文审核定稿，袁老师都严格把关，耐心指导。在这里非常感谢袁老师。我会永远铭记您的教诲，您的教诲会伴随着我一路走下去。

其次，要感谢我的师兄何佳，师兄孟卓，师姐刘翠翠。特别是何佳师兄，在平常生活中给予关怀，在论文写作过程中给予耐心指导，不抛弃，不放弃。对师兄的帮助我会铭记在心，在这里非常感谢师兄。平时大部分时间都是在实验室和宿舍度过的，接触最多的就是师兄师姐们，不仅在学习上给我提供了很多帮助和建议，而且在平时生活中也给我了很多鼓励和帮助。非常感你们。愿我们的友谊永远持续下去。

参考文献

[1] 王迎云.基于决策树算法的学生成绩挖掘与分析[D].安徽大学,2012.

[2] Anjewierden A, Kolloffel B, Hulshof C. Towards educational data mining: Using data mining methods for automated chat analysisto understand and support inquiry learning processes. In: Proc.of the Int'l Workshop on Applyi-ng Data Mining in e-Learning (ADML 2007). 2007.

[3] Adams BS, Cummins, Davis, etal. NMC Horizon Report:2017 Higher Educ-ation Edition[J]. Journal of Open Learning,2017.

[4] Cole J, Foster H. Using Moodle: Teaching with the Popular Open Source Course Management System.2nded. O'Reilly Media, Inc.2007.

[5] Phuntusil N, Limpiyakorn Y. Predicting Engaging Content for Increasing Organic Reach on Facebook[M]// Information Science and Applications 2017. 2017.

[6] Bendezu-Quispe G, Torres-Roman JS, Salinas-Ochoa B, etal. Utility of massive open online courses (MOOCs) concerning outbreaks of emerging and reemerging diseases[J]. F1000research, 2017, 6:1699.

[7] Coursera. https://www.coursera.org/.

[8] Romero C, Ventura S. Data mining in education. Wiley Interdisciplinary Reviews-Data Mining and Knowledge Discovery, 2013,3(1):12 27. [doi: 10.1002/widm.1075].

[9] 周庆,牟超,杨丹.教育数据挖掘研究进展综述[J]. 软件学报, 2015, 26(11):3026-3042.

[10] Li T, Fu GS. An overall view of the educational data mining domain. Modern Educational Technology, 201 0,20(10):21-25 (inChinese with Engli-sh abstract). [doi: 10.3969/j.issn.1009-8097.2010.10.004]

[11] Wang YG, Zhang Q. MOOC: Characteristics and learning mechanism. Education Research, 2014, (9):112-120,133(in Chinese with English abstra-ct).

[12] Meng WJ. Essence of network-based education: individualized and self-regulated learning supported by interactive systems with emotional comm-unication. Education Research, 2002, (4):52-57 (in Chinese).

[13] 牟智佳,俞显,武法提.国际教育数据挖掘研究现状的可视化分析:热点与趋势[J].电化教育研究, 2017(4):108-114.

[14] Wu YW, Li S, Tian QH. Research and Implementation of mashup intelligent question-answering system. Computer Engineering,2013,39(7):2-33-236, 241 (in Chinese with English abstract). [doi: 10.3969/j.issn.1000-3428.2013.07.052]

[15] Jiang YR, Han JH, Wu WM. Adaptive approach to personlized learning sequence generation. Computer Science, 2013,40(8):204-209 (in Chinese with English abstract). [doi: 10.3969/j.issn.1002-137X.2013.08.043]

[16] Peña-Ayala A. Educational data mining: A survey and a data mining-based analysis of recent works. Expert Systems with Applications, 2014,41:1432-1462. [doi: 10.1016/j.eswa.2013.08.042]

[17] Mohamad SK, Tasir Z. Educational data mining: A review. Procedia — Social and Behavioral Sciences, 2013,97:320-324. [doi:10. 1016/j.sbspr-o.2013.10.240]

[18] Baker RS, Yacef K. The state of educational data mining in 2009: A review and future visions. Journal of Educational Data Mining,2009,1(1):3-17.

[19] Romero C, Ventura S. Educational data mining: A survey from 1995 to 2005. Expert Systems with Applications, 2007,33(1):135-146. [doi: 10.101-6/j.eswa.2006.04.005]

[20] Borrego M, Foster MJ, Froyd JE. Systematic literature reviews in engineering education and other developing interdisciplinary fields. Journal of Engineering Education, 2014,103(1):45-76. [doi: 10.1002/jee.20038]

[21] Romero C, Ventura S, Pechenizkiy M, Baker RS. Handbook of Educatio-nal Data Mining. CRC Press, 2011.

[22] Burr L, Spennemann DH. Patterns of user behaviour in university online forums. Int'l Journal of Instructional Technology and Distance Learning, 2004,1(10):11-28.

[23] Pechenizkiy M, Trcka N, Vasilyeva E, van der Aalst W, De Bra P. Proce-ss mining online assessment data. In: Proc. of the Int'l Working Group on Educational Data Mining. 2009. 279-288.

[24] Mostow J, Beck J, Cen H, Cuneo A, Gouvea E, Heiner C. An educational data mining tool to browse tutor-student interactions: Time will tell. In: Proc. of the Workshop on Educational Data Mining, National Conf. on Artificial Intelligence. 2005. 15-22.

[25] Juan AA, Daradoumis T, Faulin J, Xhafa F. SAMOS: A model for monit-oring students' and groups' activities in collaborative e-learning. Int'l Journal of Learning Technology,2009,4(1):53-72. [doi:10.1504/IJLT.200-9.024716]

[26] Baker RS, Corbett AT, Aleven V. Improving contextual models of guessing and slipping with a truncated training set. In: Proc. Of the Educa-tional Data Mining 2008. 2008.67-76.

[27] García P, Amandi A, Schiaffino S, Campo M. Evaluating Bayesian netw-orks' precision for detecting students' learning styles. Computers& Education, 2007,49(3):794-808. [doi: 10.1016/j.compedu.2005.11.017]

[28] Jonsson A, Johns J, Mehranian H, Arroyo I, Woolf B, Barto A, Fisher D, Mahadevan S. Evaluating the feasibility of learning student models from data. In: Proc. of the Educational Data Mining: Papers from the AAAI Workshop. 2005. 1-6.

[29] Chang KM, Beck J, Mostow J, Corbett A. A Bayes net toolkit for student modeling in intelligent tutoring systems. In: Ikeda M, Ashley KD, Chan TW, eds. Proc. of the 8th Intelligent Tutoring Systems. Springer-Verlag, 2006. 104-113. [doi: 10.1007/11774303_11]

[30] Arroyo I, Murray T, Woolf BP, Beal C. Inferring unobservable learning variables from students' help seeking behavior. In: Lester JC, Vicari RM,

Paraguacu F, eds. Proc. of the Intelligent Tutoring Systems. Springer-Verlag, 2004. 782-784. [doi: 10.1007/978-3-540-30139-4_74]

[31] Antunes C. Acquiring background knowledge for intelligent tutoring systems. In: Proc. of the EDM. 2008. 18-27.

[32] Andrejko A, Barla M, Bieliková M, Tvarozek M. User characteristics acquisition from logs with semantics. In: Proc. of the Int'l Conf. on Information System Implementation and Modeling. 2007. 103-110.

[33] Robinet V, Bisson G, Gordon M, Lemaire B. Searching for student intermediate mental steps. In: Proc. of the 11th Int'l Conf. on User Modeling. 2007. 35-39.

[34] Huang J, Zhu A, Luo Q. Personality mining method in Web based education system using data mining. In: Proc. of the IEEE Int'l Conf. on Grey Systems and Intelligent Services 2007 (GSIS 2007). IEEE, 2007. 155-158. [doi: 10.1109/GSIS.2007.4443256]

[35] Matsuda N, Cohen WW, Sewall J, Lacerda G, Koedinger KR. Predicting students' performance with simstudent: learning cognitive skills from observation. In: Luckin R, Koedinger KR, Greer J, eds. Proc. of the 2007 Conf. on Artificial Intelligence in Education:Building Technology Rich Learning Contexts That Work. Amsterdam: IOS Press, 2007. 467-476.

[36] Feng M, Beck J. Back to the future: A non-automated method of constructing transfer models. In: Barnes T, Desmarais M, Romero C, Ventura S, eds. Proc. of the Int'l Working Group on Educational Data Mining, Spain, 2009. 240-248.

[37] Frias-Martinez E, Chen SY, Liu X. Survey of data mining approaches to user modeling for adaptive hypermedia. IEEE Trans. On Systems, Man, and Cybernetics, Part C: Applications and Reviews, 2006,36(6):734-749. [doi: 10.1109/TSMCC.2006.879391]

[38]莫增文.基于时间序列分析技术的预测模型设计与应用[D].中国科学院大学, 2014.

[39]范洁.基于数据挖掘的在线学习行为评估系统设计与实现[D].国防科学技术大学, 2005.

[40]骆毅.基于不确定性系统研究方法的高校学生学习成绩分析与预测[D].哈尔滨工程大学, 2011.

[41]郑清雅.云计算环境下基于学习风格的教学资源推荐系统设计与实现[D].沈阳师范大学, 2016.

[42] 姜思璐,刘建国.大数据下协作学习的个性化自适应学习系统设计研究[J].长春师范大学学报, 2016(10):72-76.

[43] 李德有,李凌霞,郭瑞波.基于Weka平台的机器学习方法探究[J].电脑知识与技术, 2012, 08(10):2334-2337.

[44] Yu H, Lo H, Hsieh H. Feature engineering and classifier ensemble for KDD cup 2010[C]// Jmlr Workshop & Conference. 2010.

[45]强海燕.世界一流大学人文课程之比较——以哈佛大学、斯坦福大学、多伦多大学为例[J].比较教育研究, 2012(11):22-26+40.

[46] Gartner.迎接大数据和数据专家[J]. 网络运维与管理, 2014(11):84-85.

[47] Wang Y J. The Age of Big Data[J]. Chinese Journal of Stroke, 2013.

[48]徐鹏,王以宁,刘艳华等.大数据视角分析学习变革——美国《通过教育数据挖掘和学习分析促进教与学》报告解读及启示[J].远程教育杂志, 2013(6):11-17.

[49]李婷,傅钢善.国内外教育数据挖掘研究现状及趋势分析[J].现代教育技术, 2010, 20(10):21-25.

[50]杨为民.在线学习的现状与发展研究[D].西北师范大学, 2007.

[51]徐海波.浅析面向在线教育的大数据应用[J].数字技术与应用, 2015(12):85-86.

[51] Shen R, Yang F, Han P. Data analysis center based on e-learning platform. In: Hommel G, Huanye S, eds. Proc. of the InternetChallenge: Technology and Applications. Springer-Verlag, 2002. 19-28. [doi: 10.100-7/978-94-010-0494-7_3]

[52]杜婧敏，方海光，李维杨，等。教育大数据研究综述[J]. 中国教育信息化，2016(19):1-4.

[53]孙钰林，大数据在职业教育中的应用[J]. 中国高教研究，2017(4):107-110.

[54]冯媛.大数据在职业教育中的应用与前景展望[J]. 科学大众(科学教育)，2015(8).

[55]蒋传进. 组合预测中的预测精度与预测风险分析[J]. 统计与决策, 2015(3):78-80.

[56]富震. 基于 SVM 主动学习技术的 PU 文本分类[J]. 计算技术与自动化, 2014(1):127-131.

[57] Han J, Kamber M, Pei J. Data Mining (Third Edition)[M]. 2011.

[58] Cortes C, Vapnik V. Support-vector networks[C]// Machine Learning. 1995:273-297.

[59]黄解军. 贝叶斯网络结构学习及其在数据挖掘中的应用研究[D]. 武汉大学, 2005.

[60] Yao Z, Fei M, Li K, et al. Recognition of blue-green algae in lakes using distributive genetic algorithm-based neural networks[J]. Neurocomputing, 2007, 70(4):641-647.

[61]杨永雷. BP 神经网络在坐标转换方面的应用[J]. 建筑工程技术与设计, 2015(15).

附录 A 攻读学位其间发表论文目录

[1] 卫明，袁梅宇，张秋明，王赟. 基于射电天文图像恢复的改进方法[J]. 传感器与微系统,2018(04):54-57.

附录 B 划分训练集和测试集的程序清单

划分训练集和测试集的 Java 源程序 Preprocessing.java

```
package kdd2010.b89;
import java.sql.Connection;
import java.sql.DriverManager;
import java.sql.PreparedStatement;
import java.sql.ResultSet;
import java.sql.Statement;
/***
 * 预处理 KDD Cup 2010 数据
 * 抽样原数据集，将原训练数据拆分为训练集和测试集。
 * 训练集由每个学生每个单元的 TRAINSETPROBLEMNUMBER 个问题组成，测试集由每个学生每个单元的最后一个问题组成。
 *
 */
public class Preprocessing {
    static final int TRAINSETPROBLEMNUMBER = 3; // 抽取的训练集中每个学生一个单元的问题数
    static final long POPULATION = 20012498; // 样本容量
    static final long BATCH = 1000; // 批次处理的样本数

    public static void main(String[] args) {
        Connection con = null;
        Statement stmt = null;
        PreparedStatement pstmtInsertTrain = null;
        PreparedStatement pstmtInsertTest = null;
        PreparedStatement pstmtSelect = null;
        ResultSet rs = null;
```

```
            try {
                // 加载数据库驱动
                Class.forName("com.mysql.jdbc.Driver");
                // 创建数据库连接
                con =
DriverManager.getConnection("jdbc:mysql://localhost:3306/weka", "weka",
"weka");
                // 创建 Statement 对象
                stmt = con.createStatement();
                String sqlTrain = "CREATE TABLE B89Train(Row int NOT NULL,
StudentId char(255), ProblemStep char(255), KC char(255), CFA tinyint, PRIMARY
KEY (Row))";
                String sqlTest = "CREATE TABLE B89Test(Row int NOT NULL,
StudentId char(255), ProblemStep char(255), KC char(255), CFA tinyint, PRIMARY
KEY (Row))";
                // 创建训练集和测试集表
                stmt.executeUpdate(sqlTrain);
                stmt.executeUpdate(sqlTest);
                // 因为记录数超大,企图一条 SQL 语句加载全部数据无法运行。
只能分批查询
                String sqlSelect = "SELECT * FROM B89 WHERE Row
BETWEEN ?AND ? ORDER BY Row DESC";
                pstmtSelect = con.prepareStatement(sqlSelect);
                long idx = POPULATION - BATCH; // 指针，idx 所指向的行是处
理过数据的上界
                pstmtSelect.setLong(1, idx);
                pstmtSelect.setLong(2, POPULATION);
                rs = pstmtSelect.executeQuery();
                String sqlInsertTrain = "INSERT INTO B89Train VALUES
(?, ?, ?, ?, ?)";
```

```
pstmtInsertTrain = con.prepareStatement(sqlInsertTrain);
String sqlInsertTest = "INSERT INTO B89Test VALUES (?, ?, ?, ?, ?)";
pstmtInsertTest = con.prepareStatement(sqlInsertTest);
long finishedSteps = 0;
int problemsPerUnit = 0;
String oldStudent = "";
String oldUnit = "";
String oldProblem = "";
String unit = "";
while (true) {
    if (!rs.next()) {
        rs.close(); // 先关闭数据集，才能重新使用
        idx--; // 避免重复使用一条记录
        if (idx < 1)
            idx = 1; // 检查边界
        pstmtSelect.setLong(2, idx);
        idx = idx - BATCH;
        if (idx < 1)
            idx = 1; // 检查边界
        pstmtSelect.setLong(1, idx);
        rs = pstmtSelect.executeQuery();
        rs.next(); // 保证跳到第一条记录
    }
    finishedSteps++;
    // 打印进度
    if (finishedSteps % 100 == 0)
        System.out.print(finishedSteps + " 条记录已处理\n");
    // 分隔出单元信息
    unit = rs.getString("Problem Hierarchy").trim();
```

```
                    unit = unit.split(",")[0];
                    if (oldStudent.equals(rs.getString("Anon Student Id").trim())
&& oldUnit.equals(unit)) {
                        // 老学生且老单元
                        if (oldProblem.equals(rs.getString("Problem
Name").trim())) {
                            // 老问题
                            // 只处理指定范围的问题
                            if (problemsPerUnit <=
TRAINSETPROBLEMNUMBER) {
                                if (problemsPerUnit == 0) {
                                    // 最开始的一个问题
                                    // 插入到测试集
                                    pstmtInsertTest.setInt(1, rs.getInt(1));
                                    pstmtInsertTest.setString(2,
rs.getString(2).trim());
                                    pstmtInsertTest.setString(3,
rs.getString(4).trim() + rs.getString(5).trim());
                                    pstmtInsertTest.setString(4, rs.getString(7));
                                    pstmtInsertTest.setInt(5, rs.getInt(6));
                                    pstmtInsertTest.execute();
                                } else {
                                // 后面的 TRAINSETPROBLEMNUMBER 个问题
                                // 插入到训练集
                                    pstmtInsertTrain.setInt(1, rs.getInt(1));
                                    pstmtInsertTrain.setString(2,
rs.getString(2).trim());
                                    pstmtInsertTrain.setString(3,
rs.getString(4).trim() + rs.getString(5).trim());
                                    pstmtInsertTrain.setString(4, rs.getString(7));
```

```
                                pstmtInsertTrain.setInt(5, rs.getInt(6));

                                pstmtInsertTrain.execute();
                            }
                        }

                    } else {
                        // 新问题
                        oldProblem = rs.getString("Problem Name").trim();
                        problemsPerUnit++;

                        // 插入到训练集
                        if (problemsPerUnit <=
TRAINSETPROBLEMNUMBER) {
                            pstmtInsertTrain.setInt(1, rs.getInt(1));
                            pstmtInsertTrain.setString(2,
rs.getString(2).trim());
                            pstmtInsertTrain.setString(3, rs.getString(4).trim()
+ rs.getString(5).trim());
                            pstmtInsertTrain.setString(4, rs.getString(7));
                            pstmtInsertTrain.setInt(5, rs.getInt(6));

                            pstmtInsertTrain.execute();
                        }
                    }

                } else {
                    // 新学生或者新单元
                    // 保存学生、单元和问题，以便比对
                    oldStudent = rs.getString("Anon Student Id").trim();
```

```
                    oldUnit = unit;
                    oldProblem = rs.getString("Problem Name").trim();

                    problemsPerUnit = 0;

                    // 插入到测试集
                    pstmtInsertTest.setInt(1, rs.getInt(1));
                    pstmtInsertTest.setString(2, rs.getString(2).trim());
                    pstmtInsertTest.setString(3,  rs.getString(4).trim()  +
rs.getString(5).trim());
                    pstmtInsertTest.setString(4, rs.getString(7));
                    pstmtInsertTest.setInt(5, rs.getInt(6));

                    pstmtInsertTest.execute();
                }

                if (rs.getInt(1) == 1) {
                    break; // 到了第一条记录，该退出了
                }
            }

        } catch (Exception e) {
            e.printStackTrace();
        } finally {
            // 关闭数据库连接
            if (con != null) {
                try {
                    con.close();
                } catch (Exception e) {
                    // 不处理了
```

```
                    }
                }
            }

        }

    }
```

附录 C 决策树文字表示

```
=== Run information ===

Scheme:       weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     B89
Instances:    2841774
Attributes:   4
              StudentChance
              PSChance
              KCChance
              CFA
Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===

J48 pruned tree
------------------

PSChance <= 0.8611
|   PSChance <= 0.56
|   |   PSChance <= 0.3846
|   |   |   StudentChance <= 0.8708: 0 (56457.0/7660.0)
|   |   |   StudentChance > 0.8708
|   |   |   |   PSChance <= 0.2581: 0 (19448.0/2164.0)
|   |   |   |   PSChance > 0.2581
|   |   |   |   |   StudentChance <= 0.9106: 0 (12316.0/4686.0)
|   |   |   |   |   StudentChance > 0.9106
|   |   |   |   |   |   StudentChance <= 0.9649
|   |   |   |   |   |   |   KCChance <= 0.5349
```

```
|   |   |   |   |   |   |   |   PSChance <= 0.3696: 0 (1089.0/457.0)
|   |   |   |   |   |   |   |   PSChance > 0.3696: 1 (188.0/88.0)
|   |   |   |   |   |   |   KCChance > 0.5349
|   |   |   |   |   |   |   |   KCChance <= 0.7191
|   |   |   |   |   |   |   |   |   StudentChance <= 0.9272
|   |   |   |   |   |   |   |   |   |   KCChance <= 0.5612: 1 (80.0/25.0)
|   |   |   |   |   |   |   |   |   |   KCChance > 0.5612
|   |   |   |   |   |   |   |   |   |   |   PSChance <= 0.3077: 0 (236.0/95.0)
|   |   |   |   |   |   |   |   |   |   |   PSChance > 0.3077: 1 (862.0/403.0)
|   |   |   |   |   |   |   |   |   StudentChance > 0.9272: 1 (760.0/307.0)
|   |   |   |   |   |   |   |   KCChance > 0.7191
|   |   |   |   |   |   |   |   |   StudentChance <= 0.9368: 0 (1672.0/765.0)
|   |   |   |   |   |   |   |   |   StudentChance > 0.9368: 1 (302.0/128.0)
|   |   |   |   |   |   StudentChance > 0.9649: 1 (76.0/14.0)
|   |   PSChance > 0.3846
|   |   |   StudentChance <= 0.8548
|   |   |   |   StudentChance <= 0.8125: 0 (28336.0/9079.0)
|   |   |   |   StudentChance > 0.8125
|   |   |   |   |   PSChance <= 0.5172: 0 (22857.0/9369.0)
|   |   |   |   |   PSChance > 0.5172
|   |   |   |   |   |   KCChance <= 0.8963: 0 (6652.0/3207.0)
|   |   |   |   |   |   KCChance > 0.8963: 1 (1188.0/551.0)
|   |   |   StudentChance > 0.8548
|   |   |   |   StudentChance <= 0.9012
|   |   |   |   |   PSChance <= 0.4643: 0 (15273.0/7075.0)
|   |   |   |   |   PSChance > 0.4643
|   |   |   |   |   |   PSChance <= 0.5227
|   |   |   |   |   |   |   StudentChance <= 0.8719: 0 (8438.0/4169.0)
|   |   |   |   |   |   |   StudentChance > 0.8719: 1 (15544.0/6841.0)
|   |   |   |   |   |   PSChance > 0.5227: 1 (12437.0/5146.0)
```

```
|   |   |   |   StudentChance > 0.9012: 1 (27886.0/9890.0)
|   PSChance > 0.56
|   |   StudentChance <= 0.8479
|   |   |   PSChance <= 0.7474
|   |   |   |   StudentChance <= 0.7788
|   |   |   |   |   StudentChance <= 0.6823: 0 (8911.0/3223.0)
|   |   |   |   |   StudentChance > 0.6823
|   |   |   |   |   |   PSChance <= 0.6374: 0 (8550.0/3471.0)
|   |   |   |   |   |   PSChance > 0.6374
|   |   |   |   |   |   |   StudentChance <= 0.7391
|   |   |   |   |   |   |   |   StudentChance <= 0.6863: 1 (335.0/122.0)
|   |   |   |   |   |   |   |   StudentChance > 0.6863
|   |   |   |   |   |   |   |   |   PSChance <= 0.6887: 0 (3291.0/1462.0)
|   |   |   |   |   |   |   |   |   PSChance > 0.6887: 1 (3448.0/1691.0)
|   |   |   |   |   |   |   StudentChance > 0.7391: 1 (10374.0/4673.0)
|   |   |   |   StudentChance > 0.7788
|   |   |   |   |   PSChance <= 0.6417
|   |   |   |   |   |   PSChance <= 0.6118
|   |   |   |   |   |   |   StudentChance <= 0.8096
|   |   |   |   |   |   |   |   PSChance <= 0.5789
|   |   |   |   |   |   |   |   |   PSChance <= 0.5664
|   |   |   |   |   |   |   |   |   |   KCChance <= 0.7568
|   |   |   |   |   |   |   |   |   |   |   PSChance <= 0.5647: 0 (134.0/53.0)
|   |   |   |   |   |   |   |   |   |   |   PSChance > 0.5647
|   |   |   |   |   |   |   |   |   |   |   |   KCChance <= 0.5541: 0 (4.0)
|   |   |   |   |   |   |   |   |   |   |   |   KCChance > 0.5541: 1
(45.0/18.0)
|   |   |   |   |   |   |   |   |   |   KCChance > 0.7568: 1 (130.0/49.0)
|   |   |   |   |   |   |   |   |   PSChance > 0.5664: 0 (1321.0/562.0)
|   |   |   |   |   |   |   |   PSChance > 0.5789: 0 (3372.0/1656.0)
```

```
|   |   |   |   |   |   StudentChance > 0.8096: 1 (11960.0/5726.0)
|   |   |   |   |   PSChance > 0.6118
|   |   |   |   |   |   StudentChance <= 0.7989
|   |   |   |   |   |   |   StudentChance <= 0.7971: 1 (1473.0/720.0)
|   |   |   |   |   |   |   StudentChance > 0.7971
|   |   |   |   |   |   |   |   PSChance <= 0.6163: 1 (28.0/10.0)
|   |   |   |   |   |   |   |   PSChance > 0.6163: 0 (191.0/68.0)
|   |   |   |   |   |   StudentChance > 0.7989: 1 (8219.0/3547.0)
|   |   |   |   PSChance > 0.6417: 1 (53467.0/20136.0)
|   |   |   PSChance > 0.7474
|   |   |   |   StudentChance <= 0.7391
|   |   |   |   |   StudentChance <= 0.5547: 0 (1313.0/494.0)
|   |   |   |   |   StudentChance > 0.5547
|   |   |   |   |   |   PSChance <= 0.8219
|   |   |   |   |   |   |   KCChance <= 0.8983
|   |   |   |   |   |   |   |   PSChance <= 0.7609
|   |   |   |   |   |   |   |   |   KCChance <= 0.7005
|   |   |   |   |   |   |   |   |   |   KCChance <= 0.6537
|   |   |   |   |   |   |   |   |   |   |   KCChance <= 0.6097: 0 (132.0/57.0)
|   |   |   |   |   |   |   |   |   |   |   KCChance > 0.6097: 1 (56.0/18.0)
|   |   |   |   |   |   |   |   |   |   KCChance > 0.6537: 0 (247.0/84.0)
|   |   |   |   |   |   |   |   |   KCChance > 0.7005: 1 (1397.0/693.0)
|   |   |   |   |   |   |   |   PSChance > 0.7609
|   |   |   |   |   |   |   |   |   StudentChance <= 0.6823
|   |   |   |   |   |   |   |   |   |   StudentChance <= 0.6159: 0
(684.0/314.0)
|   |   |   |   |   |   |   |   |   |   StudentChance > 0.6159: 1
(1687.0/782.0)
|   |   |   |   |   |   |   |   |   StudentChance > 0.6823
|   |   |   |   |   |   |   |   |   |   KCChance <= 0.8976: 1 (3490.0/1408.0)
```

```
|   |   |   |   |   |   |   |   |   |   KCChance > 0.8976
|   |   |   |   |   |   |   |   |   |   |   StudentChance <= 0.7352: 0
(60.0/16.0)
|   |   |   |   |   |   |   |   |   |   |   StudentChance > 0.7352: 1 (6.0/1.0)
|   |   |   |   |   |   |   KCChance > 0.8983
|   |   |   |   |   |   |   |   PSChance <= 0.7558
|   |   |   |   |   |   |   |   |   StudentChance <= 0.6842: 0 (350.0/168.0)
|   |   |   |   |   |   |   |   |   StudentChance > 0.6842: 1 (671.0/259.0)
|   |   |   |   |   |   |   |   PSChance > 0.7558: 1 (3457.0/1216.0)
|   |   |   |   |   |   PSChance > 0.8219
|   |   |   |   |   |   |   StudentChance <= 0.6489
|   |   |   |   |   |   |   |   KCChance <= 0.8476
|   |   |   |   |   |   |   |   |   KCChance <= 0.8238
|   |   |   |   |   |   |   |   |   |   KCChance <= 0.6983
|   |   |   |   |   |   |   |   |   |   |   KCChance <= 0.5588: 1 (4.0)
|   |   |   |   |   |   |   |   |   |   |   KCChance > 0.5588: 0 (65.0/24.0)
|   |   |   |   |   |   |   |   |   |   KCChance > 0.6983: 1 (299.0/116.0)
|   |   |   |   |   |   |   |   |   KCChance > 0.8238
|   |   |   |   |   |   |   |   |   |   KCChance <= 0.8277
|   |   |   |   |   |   |   |   |   |   |   StudentChance <= 0.5771: 1 (2.0)
|   |   |   |   |   |   |   |   |   |   |   StudentChance > 0.5771: 0
(21.0/2.0)
|   |   |   |   |   |   |   |   |   |   KCChance > 0.8277
|   |   |   |   |   |   |   |   |   |   |   StudentChance <= 0.5625: 1 (5.0)
|   |   |   |   |   |   |   |   |   |   |   StudentChance > 0.5625
|   |   |   |   |   |   |   |   |   |   |   |   StudentChance <= 0.6055: 0
(36.0/7.0)
|   |   |   |   |   |   |   |   |   |   |   StudentChance > 0.6055
|   |   |   |   |   |   |   |   |   |   |   |   StudentChance <= 0.6259
|   |   |   |   |   |   |   |   |   |   |   |   |   PSChance <= 0.8478
```

| | | | | | | | | | | | | | | KCChance <= 0.8438: 1 (15.0/1.0)

| | | | | | | | | | | | | | | KCChance > 0.8438

| | | | | | | | | | | | | | | StudentChance <= 0.6118: 0 (2.0)

| | | | | | | | | | | | | | | StudentChance > 0.6118: 1 (6.0/1.0)

| | | | | | | | | | | | | | PSChance > 0.8478

| | | | | | | | | | | | | | | KCChance <= 0.8438: 0 (6.0/1.0)

| | | | | | | | | | | | | | | KCChance > 0.8438: 1 (2.0)

| | | | | | | | | | | | | | StudentChance > 0.6259: 0 (19.0/6.0)

| | | | | | | | | KCChance > 0.8476: 1 (1054.0/358.0)

| | | | | | | | StudentChance > 0.6489: 1 (6310.0/1939.0)

| | | | StudentChance > 0.7391: 1 (119022.0/30036.0)

| | StudentChance > 0.8479: 1 (491368.0/100456.0)

PSChance > 0.8611

| PSChance <= 0.9847

| | StudentChance <= 0.8194

| | | StudentChance <= 0.7102

| | | | StudentChance <= 0.4667

| | | | | StudentChance <= 0.1818

| | | | | | KCChance <= 0.901

| | | | | | | KCChance <= 0.7941: 1 (6.0/1.0)

| | | | | | | KCChance > 0.7941: 0 (14.0/4.0)

| | | | | | KCChance > 0.901: 0 (155.0/2.0)

| | | | | StudentChance > 0.1818

```
|   |   |   |   |   |   StudentChance <= 0.3841
|   |   |   |   |   |   |   PSChance <= 0.8672: 1 (7.0/1.0)
|   |   |   |   |   |   |   PSChance > 0.8672: 0 (209.0/80.0)
|   |   |   |   |   |   StudentChance > 0.3841
|   |   |   |   |   |   |   StudentChance <= 0.4051
|   |   |   |   |   |   |   |   KCChance <= 0.9317: 1 (132.0/32.0)
|   |   |   |   |   |   |   |   KCChance > 0.9317: 0 (6.0)
|   |   |   |   |   |   |   StudentChance > 0.4051: 1 (201.0/96.0)
|   |   |   |   StudentChance > 0.4667: 1 (21886.0/3986.0)
|   |   |   StudentChance > 0.7102: 1 (115518.0/12632.0)
|   |   StudentChance > 0.8194: 1 (629788.0/37378.0)
|   PSChance > 0.9847: 1 (1094716.0/679.0)

Number of Leaves  : 81

Size of the tree :  161

Time taken to build model: 55.18 seconds
```

免费论文查重：http://www.paperyy.com
3亿免费文献下载：http://www.ixueshu.com
超值论文自动降重：http://www.paperyy.com/reduce_repetition
PPT免费模版下载：http://ppt.ixueshu.com

阅读此文的还阅读了：

26. 离群数据挖掘在电力负荷预测中的应用研究

27. 数据挖掘算法在水质评价预测中的应用

28. 基于在线商店流量数据的销售预测研究

29. 数据挖掘技术的应用研究

30. 基于数据挖掘的软件测试应用研究

31. 数据挖掘技术在公安预测预警中的应用

32. 数据挖掘技术的应用研究

33. 基于数据挖掘的疾病预测

34. 探析数据挖掘在智能在线答疑系统中的应用

35. 基于数据挖掘的软件缺陷预测技术研究

36. 基于数据挖掘技术的医疗设备绩效预测方法的应用研究

37. 数据挖掘技术的应用研究

38. 在线学习系统中数据挖掘的应用

39. 基于数据挖掘的证券指数预测研究

40. 数据挖掘模型在股市预测中的应用综述

41. 数据挖掘技术的应用研究

42. 数据挖掘在股票预测中的应用

43. 基于数据挖掘的证券指数预测研究

44. 销售预测中数据挖掘的应用

45. 数据挖掘技术在在线开放课程中的应用

46. 数据挖掘技术在在线考试系统中的应用研究

47. 数据挖掘技术于在线考试系统的应用

48. 数据挖掘技术的应用研究

49. 数据挖掘技术在股票预测中的应用探讨

50. 电气设备在线数据预测的研究