JIYU DASHUJU DE
GAOFENGXIAN
XUESHENG
YUCE YANJIU

基于大数据的高风险学生预测研究

余小高 /著

厦门大学出版社 XIAMEN UNIVERSITY PRESS
国家一级出版社
全国百佳图书出版单位

本书为全国教育科学“十三五”规划2016年度教育部重点课题“基于大数据的高风险学生预测研究”(课题批准号：DCA160263)课题成果，并获得该课题资助出版。

基于大数据的高风险学生预测研究

余小高 著

厦门大学出版社 XIAMEN UNIVERSITY PRESS
国家一级出版社
全国百佳图书出版单位

图书在版编目(CIP)数据

基于大数据的高风险学生预测研究/余小高著.—厦门:厦门大学出版社,2019.5
ISBN 978-7-5615-7230-6

Ⅰ.①基… Ⅱ.①余… Ⅲ.①数据处理—应用—学生工作—研究 Ⅳ.①G455—39

中国版本图书馆 CIP 数据核字(2018)第 268593 号

出 版 人 郑文礼
责任编辑 陈进才

出版发行 厦门大学出版社
社　　址 厦门市软件园二期望海路 39 号
邮政编码 361008
总 编 办 0592-2182177　0592-2181406(传真)
营销中心 0592-2184458　0592-2181365
网　　址 http://www.xmupress.com
邮　　箱 xmup@xmupress.com
印　　刷 虎彩印艺股份有限公司

开本 787 mm×1 092 mm　1/16
印张 12.5
字数 305 千字
版次 2019 年 5 月第 1 版
印次 2019 年 5 月第 1 次印刷
定价 39.00 元

厦门大学出版社
微信二维码

厦门大学出版社
微博二维码

目 录

第一章　绪论

从课程学习角度，高风险学生是指期末考试多门课程成绩不合格导致留级以及退学的学生。通过对高风险学生的预测，每学期的前半段将能了解期末考试有问题的学生，及时掌握这些学生的学习情况，分析对他们学习产生影响的潜在因素，找出他们在学习中存在的问题，可以提醒教师、辅导员等及时采取合适的措施对这些学生进行干预和帮扶，督促、引导他们顺利通过期末课程考试，正常完成学业，减少留级甚至退学的风险。为了提高人才培养质量，降低学生留级或退学风险，本书在分析国内外研究现状的基础上，首先设计教育大数据服务模式，搭建教育大数据集成平台，对校园一卡通系统、学生管理信息系统、教务管理系统、网络日志等学生数据进行采集、清洗、存储和分析；然后研究学生特征提取和选择的方法，从而对学生特征进行恰当且准确的提取和选择；最后研究和选择合适的预测方法，构建满足需求的高风险学生预测模型并进行评估，设计并实现原型系统，以期能为我国教育大数据的开发和利用提供借鉴，为提高人才培养质量提供参考。

第一节　研究背景

一、问题提出

大数据技术可以从海量的数据中挖掘出隐藏的模式和有价值的知识[1]，目前已在电子商务、金融和生物等领域得到成功应用。近年来，大数据环境下的教育大数据挖掘(Educational Big Data Mining，EBDM)得力于远程教育、在线教育和教育信息化等应用的推动，已迅速发展成一个新兴的研究领域并成为人们关注的热点[2]。

教育大数据挖掘是综合运用了教育学、心理学、社会学、统计学、数学和计算机等多学科知识并快速发展的一门新兴学科[3]。计算机科学与教育学的组合形成了计算机教育，统计学与计算机科学的相互结合产生了数据挖掘，而教育学与统计学的交叉融合则催生了学习分析，等等。将教育学、统计学和计算机科学这三门学科交叉融合则产生了教育大数据挖掘学科。教育大数据挖掘利用教育学知识和大数据技术，分析和处理对学生教育过程中形成的海量数据，从这些大数据中挖掘出其潜在的价值，找出教育中的新知识、新规律，为学习者和教育工作者服务，确保教育部门和有关机构制定的教育目标能够更加高效地完成。

教育大数据挖掘在个性化教学、学生成绩预测、高风险学生预测、知识推荐与自主学习等领域有着非常广泛的应用，学生成绩预测是根据学习者在学习过程中形成的各种数据，采用合适的预测方法来预测、评估或判断其学习效果。在教育数据挖掘领域，学生成绩预测是最早的研究方向之一，也是最重要的研究方向[4-6]。学生成绩预测是高风险学生预测的基础，当前有关学生成绩预测的研究成果采用的数据量较小、数据源单一、数据结构简单，对数

据分析与处理的方式方法也较为简单，不能满足实际应用的需要。

而做好高风险学生预测，需要学生的性格爱好、心理状况、个人基本数据、各阶段各科目的成绩、受教育程度、文化和社会背景、家庭情况、人际关系、上网情况等方方面面的复杂数据[7-11]，这些数据来源丰富，数据结构复杂，数据量巨大。因此，本书采用大数据技术，搭建教育大数据集成平台，采集、清洗、存储、分析和处理海量的学生数据，在每学期的中前期有效地预测课程考试有可能不及格的高风险学生，及时对这些学生采取合适的措施进行帮扶，引导他们顺利完成课程学习。

二、本书的研究意义

高风险学生预测可以尽早地让这些学生发现学习中存在的问题，帮助他们认识问题所在，改进学习方法，提高学习成绩；也能够及时帮助辅导员、教师等有关人员了解学生情况，及时对高风险学生进行干预并采取帮扶措施帮助这些学生进步。因此，本书的完成具有重要的实用价值，可以有效地挖掘长期积累的教育大数据，教育教学管理人员根据挖掘的结果，能够更准确而详细地了解校情和学情，提高教育教学质量，减少学生留级或辍学的风险。

教育大数据挖掘是一门跨越多学科的新兴交叉学科，互联网＋、大数据、移动计算、物联网、云计算、云存储等推动了教育大数据挖掘学科的迅速发展。然而，当前教育大数据挖掘的研究还存在采用的数据量较少、数据处理方法简单，研究成果适用范围较窄等研究不够深入的问题。本书采用大数据技术，通过研究，选择合适的数据处理方法对多源海量异构的学生数据进行分析处理，提取和选择满足需求的高风险学生预测模型特征，构建较通用的、符合期望的高风险学生预测模型。因此，本书的研究具有较强的理论价值，拓宽了教育大数据挖掘的适用范围，丰富了其研究方法。

基于大数据的高风险学生预测是一个崭新的研究课题，本书的研究能够利于学校的教育教学工作，强化学风建设，提高学生自主学习的积极性，达到提高人才培养质量的目的；同时，本书的研究能够对教育大数据挖掘的科学研究发挥积极作用，促进该学科的快速发展。

第二节　研究现状与分析

一、研究现状简述

美国在大数据的研究与应用等方面处于领先地位，根据实际应用的需求，一些美国学校和机构研发并推出了不同的学生成绩预测系统，这些系统在实际应用中取得了一定的成效[12]。

（一）功能角度

根据学生成绩预测功能的实现形式不同，当前的学生成绩预测系统主要有以下四种类型。

1. 独立预测

由学校或企业机构主导开发并独立运行的学生成绩预测系统，实现对学生在线学习进行预测的相关功能，比如美国普渡大学研发并推出的课程信号系统和 Desire2Learn 机构推出的学生成功系统。美国普渡大学研发的用于监测学生学习状态的课程信号系统是属于在

线学习预测系统，该系统根据一种新的算法能够预测出哪些学生处于学业风险中，并对这些风险学生进行预警和干预。根据学生在学习过程中的不同学习状态，系统能够设定对应的“警示信号”进行预警，教师根据这些“警示信号”对学习者进行针对性的干预，帮助这些风险学生健康成长，顺利完成学业。学生成功系统由美国 Desire2Learn 机构研发并推广，该系统能够提供一系列有关学习的服务推送给教师、学生等用户，比如了解学生的学习情况、预测学生在完成学业中存在的风险、对问题学生进行预警和干预、生成学生的学习分析报告等。该系统能够分析学生在学习过程中有哪些因素困扰其学习，这些影响因素包括学生课堂学习的出勤率、老师布置的学习任务是否完成、课程学习的参与程度等，根据这些影响因素生成高风险学生预测模型，能够准确地预测出学业存在高风险的学生，并采取适当的干预措施和帮扶方法，引导这些学生完成学业[13]。

2. 预测过程可视化

为了增强风险学生预测结果的可解释性和可读性，将可视化工具应用到学习管理系统中，实现预测过程可视化功能，能够直观地了解学生的学习情况，比如可汗学院推出的学习仪表盘，采用的是动态图形化界面，简单直观。可汗学院研发的学生学习状态的信息跟踪和镜像技术的学习仪表盘，能够对学生在线学习的各种行为进行动态跟踪，采用可视化的图形界面，直观地分析和显示其学习情况，并对风险学生进行预警。该系统记录和整合了大量的有关学生学习的信息，并根据用户提出的需求采用统计和分析方法对学生学习数据进行处理，以数字、文本、图表等多种形式可视化和直观地展示出来，帮助相关人员进行学习分析[14]。

3. 个性化预测

将个性化工具应用到学生的学习系统中，实现对不同学生进行个性化预测的功能，比如电子顾问采用的就是这种方法。美国亚利桑那州立大学研发并推广的电子顾问，让不同学生分别完成一定的图像游戏任务，据此探索出这些学生的职业偏好，从而制定符合不同学生特点的学习路径，进行个性化帮扶。该系统要求学生每学期学习完规定的课程任务，并取得相应的学分[15]。学生在学习过程中，该系统能够为这些学生提供个性化的点播功能，并引导和支持这些学生完成规定的学习任务，对学习中存在高风险的学生进行预测，帮助其顺利完成学业。

4. 模块组件预测

将风险学生成绩预测系统作为一个功能模块应用到学生在线学习平台中，比如海星预警系统，该系统具有对风险学生进行预测的功能，能够预测出风险学生并对其进行干预，帮助这些学生顺利完成学业。海星企业成功平台中实现的海星预警系统能够帮助不同学生制定符合其特点的学习任务，引导这些自主学习，顺利完成学业。该系统以大数据技术、分布式存储和运算技术、学习分析技术等为基础，有效地分析不同学生的学习努力程度，以尽快地把握这些学生的学习情况，了解不同学生的兴趣、爱好等个性化特点，并对有风险的学生进行干预，降低这些学生的退学率[16]。

上述从功能角度，分析了四种当前学生成绩预测系统的主要类型，介绍了五个代表性的学生成绩预测系统，即课程信号系统、学生成功系统、学习仪表盘、电子顾问和海星预警系统，表 1.1 从这五个系统的实现预测形式、风险学生名单发布方式与内容、实现的技术及其成效与不足等方面进行了分析。

表 1.1 五个典型学生成绩预测系统比较

	实现形式	内容	发布方式	技术	优点	缺点
课程信号系统	学校自主研发	成绩/努力程度/辍学	电子邮件/短信/消息	数据挖掘/预测算法/分析工具	减少辍学率/节省时间和经费	个性化弱/干预过多/帮扶不足
学生成功系统	企业机构研发	学业风险/辍学	可视化/电子邮件	预测模型/学习分析/可视化	预测/干预/减少辍学率	普适性不高
学习仪表盘	可视化预测	知识点掌握程度	电子邮件/仪表盘	信息跟踪/镜像技术/学习分析	个性化学习	应用范围窄/预警单一
电子顾问	个性化预测	学习路径预警	电子邮件	个性化点播工具/学习分析	课程推荐/学习个性化/帮扶	预警形式单一/无法及时准确
海星预警系统	功能模块	课程成绩/努力程度	电子邮件/红旗/短信	分布式运算/大数据/自适应	掌握学生/减少辍学率	预警单一

(二)教学环境角度

学生成绩预测一直是教育科学研究的重点,以下根据教学环境的不同,分别从三个方面介绍该领域的研究现状。

1. 封闭式教学

封闭式教学系统主要是指单机学习系统和基于客户/服务结构的管理信息系统,这种类型的系统主要由学校内部学生、教师和教学管理等人员使用,系统不能提供学生之间的互动和交流等功能。这种类型系统的用户量少,用户之间无法互动和交流,因此与系统有关的数据量很小。例如,Natek 利用决策树技术对高校信息系统中的数据进行分析,发现了对学生课程及格率产生影响的关键因素,并成功地对学生课程考试的最后分数进行了预测[17];根据蒙特卡罗理论,Caro 等人模拟了最近 10 年来的学生数据,并给高校教育教学管理者提供学生成绩的分析结果,帮助他们进行决策支持[18]。

2. 开放式教学

当今信息技术的快速发展,促进了网络技术在教育领域的广泛应用,有力地推动了网络教育课程的发展,并取得了较好的效果。这些网络教育课程一般是以 Web 技术为基础,采用一定程度的人工智能技术对学习数据进行分析,帮助学生提供学习成绩。开放式教学环境相比于封闭式教学,提供了学生之间互动和交流的学习功能。智能导学系统(Intelligent tutoring system,ITS)是开放式教学最典型的代表。ITS 采用了人工智能技术,是一种开放式的智能学习系统,给学生的学习提供了互动和交流的功能,允许老师利用系统对学生进行管理,并记录不同学生的学习情况。同时,ITS 的数据涉及学生登陆系统的日志文件、论坛讨论和发言、作业完成情况和教学资源等,数据十分丰富。由于 ITS 记录的这些数据与学生的学习行为息息相关,得到了许多研究者的重视。例如,Lara 等人利用 ITS 记录大量的学生数据,构建了学生成绩预测模型,根据该模型来预测学生能否顺利完成课程的学习任务并通过考试[19];Romero 等人分析了学生在学习论坛中的交流和发言记录,采用聚类和分类等技术,成功地对学生的课程最终成绩进行了预测[7];Hackey 等人逻辑回归技术处理和分

析在线课程中学生的学习经验和他们已有的学习数据，得到了在线课程的学习经验是学生能否完成后续课程学习任务的关键因素[20]。

3. 新型教学环境

近年来，随着信息技术的飞速发展和在教育领域的广泛应用，大量的新型教育教学环境如雨后春笋一样地产生。例如，移动智能设备[21]、各种类型的游戏[22,23]、丰富多彩的社交网络[24-26]和增强现实技术[27]在教育教学中大量和深入的应用，以及大规模开放在线课程(Massive Open Online Course，MOOC)[28]的推广与普及。众多研究者对这些新型的教育教学环境进行了研究，例如文献[22]根据学生参加计算机游戏的数据来分析和预测这些学生的学习类型，预测结果的准确率大于85%。

以上是国外的大数据环境下的教育数据挖掘的研究现状与分析，而国内在这方面研究的广度和深度上与国外的研究有较大的差距，而且研究起步较晚[29]。最近十年来，国内学者对EBDM的研究已取得了一些进展[28,30-31]，但在总体上主要还存在以下三个方面的不足：(1)研究成果的创新性较弱，大多数研究成果仅仅是对国外研究成果进行述评、跟踪和改进；(2)研究技术深度较低，大多数研究成果仅仅在教育类期刊上发表；(3)研究范围较小，大多数研究成果集中在智能导学系统[32]和学生的个性化学习两个领域[33]，几乎没有关于学业高风险学生的预测研究。由于许多国外高校学生一般由多个公司提供在校的日常消费和网络等服务，不由学校负责提供服务，很难获得各种学生数据；而国内教育体制和人们社会习惯的差异，国内高校在教育数据获取和积累方面比国外占有优势，利用这一优势，可以为我国高校的学生、教师、教育教学管理者提供更好的服务。

二、高风险学生预测系统概述

(一)系统架构

系统架构主要包含四个方面：预测目的、预测内容、预测方式和预测结果，其架构如图1.1所示。

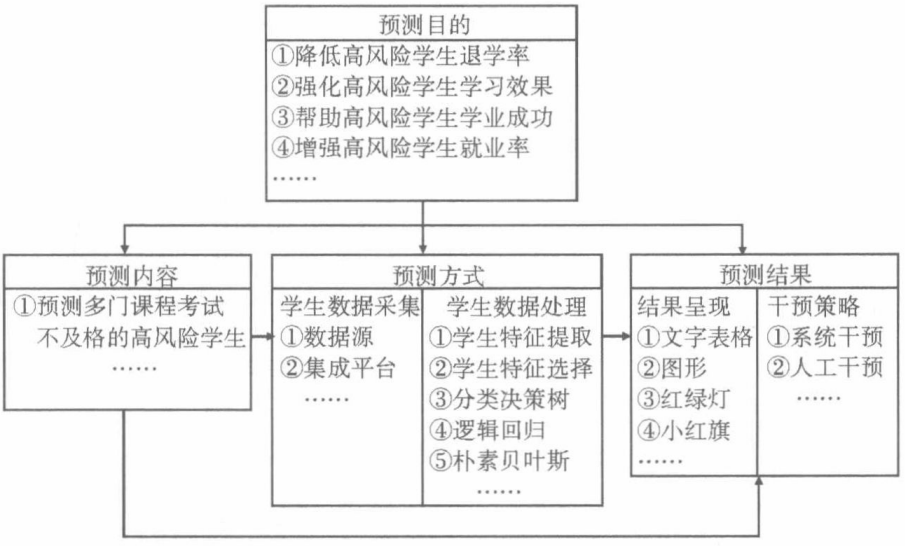

图1.1　预测系统架构

1. 预测目的

明确高风险学生预测目的是为预测系统的建设奠定了基础，指明了方向，主要有如下四种预测目的：①提高学生学习效果，②帮助学生学业成功，③降低学生退学率，④提升学生就业率。预测目的是整个预测过程的根本，直接影响教育大数据的采集，对预测内容、预测方式和预测结果产生影响。本书研究的主要目的是降低学生退学率，提高学生学习效果，从而达到学生学业成功和提高学生就业率。

2. 预测内容

当前在教育领域，对高风险学生的预测内容主要包括预测多门课程考试不及格的高风险学生等，根据预测目的确定具体的预测内容，预测方式依据预测内容的不同进行相应的改变。本书研究的主要内容是根据学生管理系统、教务系统、学生的基本情况、网络日志、一卡通等数据来预测多门课程不及格的高风险学生，预防这些学生留级或退学的风险。

3. 预测方式

高风险学生预测方式主要包括教育大数据采集、处理和预测算法，明确使用采集的技术，确定采集的学生数据源、数据类型，搭建数据集成平台。根据采集的学生数据确定数据处理技术和预测算法。大数据环境下，本书利用教育大数据集成平台，采集学生全方位数据，包括学生管理系统的数据、教务系统的数据、学生一卡通数据、学生基本数据和网络日志等，根据预测目的的提取和选择学生特征，组合逻辑回归、贝叶斯、决策树等预测方法构建预测模型，提高预测效果。

4. 预测结果

高风险学生预测的结果主要有预测结果的信息呈现和干预策略两个方面的内容，即根据预测目的、预测内容和预测方式等确定预测结果以文字表格、图形、红绿灯及小红旗等形式呈现，提供对高风险学生进行系统干预或人工干预等的干预策略。预测结果的信息呈现方式是风险预测系统的直观展示，干预策略是给高风险学生提供合适的建议或反馈。

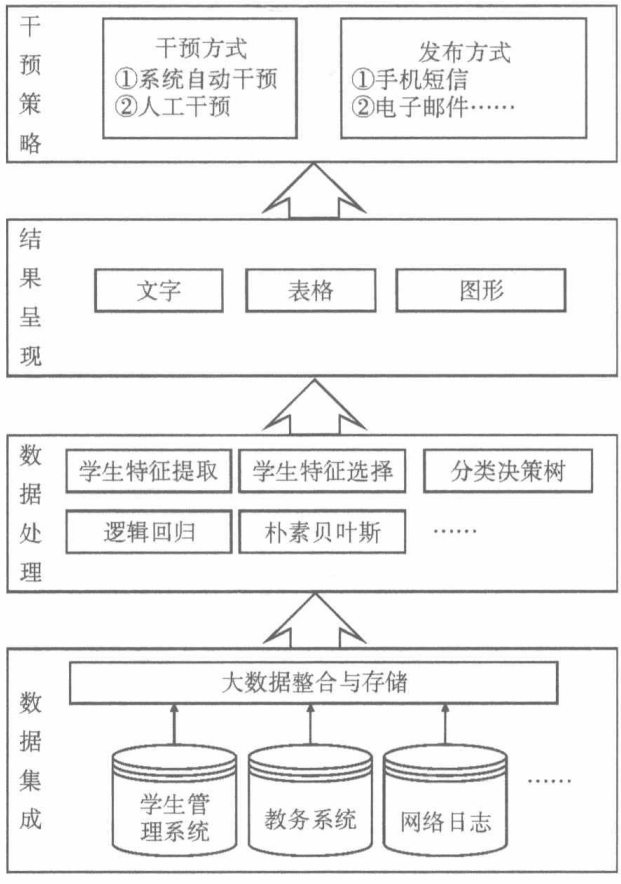

图 1.2　风险预测系统功能

(二)系统功能及过程

1. 风险预测功能

风险预测功能分为数据集成、数据处理、结果呈现和干预策略四个功能模块，如图 1.2 所示。

(1)数据集成

建立教育大数据集成平台，将学生管理系统的数据、教务系统的数据、学生基本数据和网络日志等多种复杂数据进行集成。其中，学生基本数据主要包括学生档案资料、家庭情况、学习风格、习惯、态度、

情感、社会关系等数据。

(2)数据处理

数据处理技术主要从两个方面展开:①学生特征提取和选择,包括基于时间轴的学生基本特征提取及选择、基于校园一卡通的学生特征提取及选择、基于网络日志的学生特征提取及选择、基于内容的分析、学生情感识别、社会网络分析、话语分析[34]、学生性格分析和语境分析等,其中学生情感识别技术主要从面部表情识别、语音情感识别和肢体语言识别这三个方面展开研究,利用内容挖掘和智能分析等技术,可以发现学生的语音、文本、绘图、录像、视频中包含的有关情绪信息。②预测方法的选择及改进,主要有分类决策树、逻辑回归、朴素贝叶斯等。通过这些数据分析技术对学生数据进行分析,可以在每学期的前半段预测期末考试存在多门课程不及格的高风险学生。

(3)结果呈现

结果呈现是对预测的期末考试多门课程不及格的高风险学生进行直观呈现,通过对教育大数据的分析,准确预测出高风险学生名单,采用文字、表格、图形等方式直观地呈现出来。

(4)干预策略

预测为高风险的学生,需要对这些学生进行干预,帮助他们学习进步。预测系统有完善的干预策略库,存放解决各种问题的具体干预策略,教育管理者或系统能够通过手机短信或电子邮件等向学生提供合适的干预。

2. 预测过程

预测过程如图 1.3 所示。

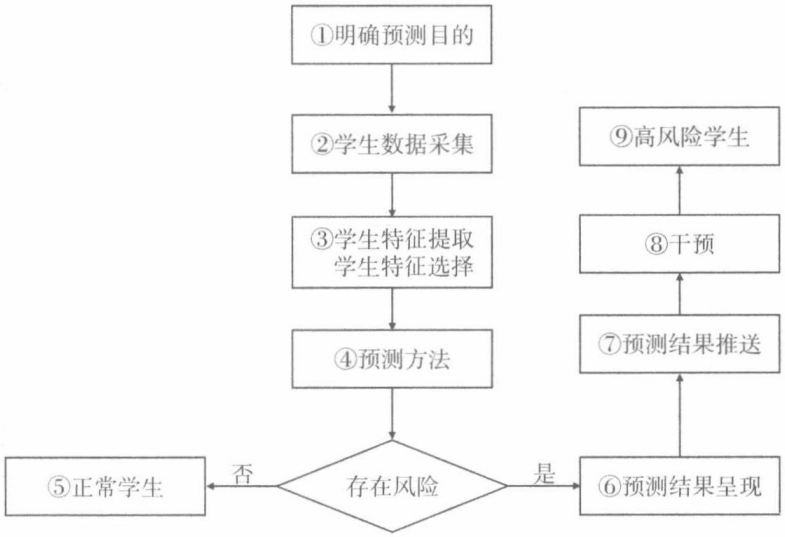

图 1.3　预测过程

一般而言,高风险学生的预测过程可分为如下九个步骤。

(1)明确高风险学生预测目的。

(2)学生数据采集。根据第一步的预测目的,确定采集技术,明确需要采集的学生数据,搭建大数据集成平台。

(3)学生特征提取和选择。将采集的数据进行清洗后,采用合适的数据分析技术提取和

选择学生特征，比如基于时间轴的学生基本特征提取、基于校园一卡通的学生特征提取和基于网络日志的学生特征提取等。

(4)预测方法。根据预测目的和预测内容，选择预测方法，比如分类决策树算法、逻辑回归算法和朴素贝叶斯算法等等，并将这些算法组合预测，以求取得最佳预测效果。

(5)学习情况呈现。根据第三步判断，若不是高风险学生，则呈现当前学生为正常学生。

(6)预测结果呈现。根据第三步判断，若是高风险学生，则呈现预测结果。

(7)预测结果推送。可以分别向教师、教育教学管理者和学生推送各自合适的预测结果。

(8)干预。向高风险学生提出个性化建议及推荐合适的资源。教师和教育教学管理者根据预测结果，人工对这些高风险学生进行干预，提出适合学生的个性化建议，并推荐适合学生的优质资源；根据预测结果，系统可以对高风险学生进行干扰，自动向学生推荐合适的建议和资源。

(9)个性化建议和优质资源推荐给相应的高风险学生。

三、当前研究存在的问题

近年来，EBDM 的研究得到迅速发展，越来越多的学者对风险学生成绩预测这一课题的研究表现出浓厚的兴趣，但是当前的研究还处于基础阶段，体现在以下四个方面。

(一)数据源比较单一

当前研究的数据仅来自某个信息系统或某门课程的数据[35-37]；数据来源过于简单，采集的数据仅仅是简单的数据集，无法全面和系统化地对学生进行预测。

(二)数据量比较小

研究的数据集涵盖的学生不多，一般只有几十到几百个学生，占全体学生比例较少，数据量过少[38-40]，不具有典型性和代表性。

(三)数据处理比较简单

根据检索已发表的研究成果，其研究的数据在表达的内容和数据结构上都较为简单，这些数据含有较少的噪声，只需要简单的数据清洗工作就可以进行分类和预测[35,41]。

(四)应用范围过窄

当前的研究工作基本是采集某一特定课程的学生数据，根据这些数据构建预测模型，预测学习该门课程的学生考试能否及格，无法对多门课程进行组合预测，这使研究成果的应用范围过于狭窄[4,7]。

第三节　研究内容

本书采用调研、数据采集、数据分析、干预实验、模型构建和原型系统开发等方法对学生基本档案、学生成绩数据、校园一卡通消费数据、学生宿舍门禁数据、图书馆门禁、借阅数据和网络日志等学生数据进行集成和清理，并有针对性地进行特征提取和特征选择，选择合适的预测方法，选择有效的预测模型，开发高风险学生预测原型系统，在每学期的前半段预测

该学期期末课程考试不及格的高风险学生。本书研究内容的总体框架如图 1.4 所示。

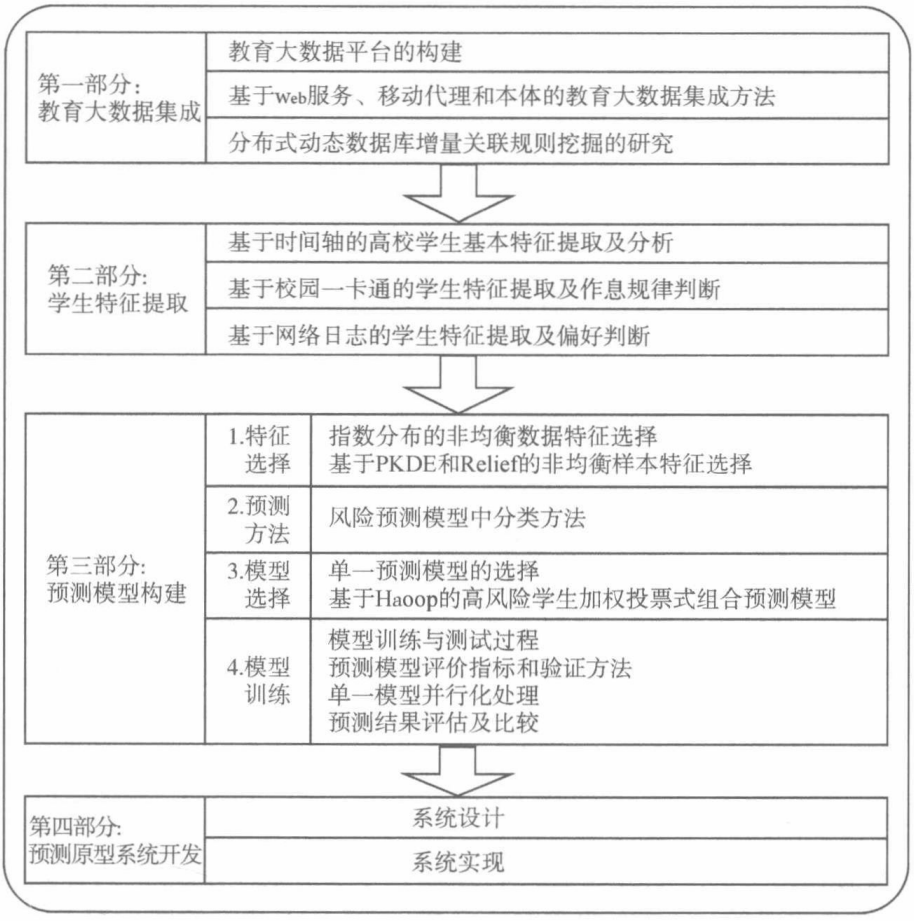

图 1.4 研究内容的总体框架

根据图 1.4，本书研究内容主要包括教育大数据集成、学生特征提取、预测模型构建和预测原型系统开发四个部分。学生特征提取是在对各种教育、教学、学生学习和学生管理等的数据源进行集成和分析的基础上，从这些数据中提取和选择出能够识别学生类型的特征，这步工作是整个研究的基础；高风险学生预测模型的构建是以提取和选择出合适的学生特征为基础，设计出有效的高风险学生预测模型；原型系统开发是用编程工具完成高风险学生预测模型的软件开发。下面对各个研究部分进行介绍。

一、教育大数据集成

以大数据技术为基础，根据教育大数据的含义及其构成，搭建教育大数据平台，进行需求分析，构建教育大数据的服务模型，研究基于 Web 服务、移动代理和本体的教育大数据集成方法和分布式动态教育大数据增量关联规则挖掘，为高风险学生预测所需的数据集成、特征提取、特征选择和预测模型的实现提供数据集成和分析平台。

二、学生特征提取

学生特征提取是通过采集并处理有关学生的各种数据，提供高风险学生预测模型能准确识别高风险学生而必需的学生特征。根据学生数据来源的不同，分别提出基于时间轴的

高校学生基本特征提取及分析、基于校园一卡通的学生特征提取及作息规律判断、基于网络日志的学生特征提取及其偏好判断，从学生基本特征提取、学生作息特征提取和网络日志特征提取三个方面完成学生特征提取。

（一）基本特征提取

学生基本特征是指学生独有的基本象征和标志，根据其内容分为三大类：个人基础特征、学习成绩和社会关系；个人基础特征主要有：姓名、性别、年龄、身高、相貌（以登记照为主）、签名（个人字迹）、民族、身体健康、家庭情况（单亲、父母双亡、兄弟姐妹人数、经济状况等）、生源地等；学习成绩主要指高考成绩、每学期期中期末各科成绩等；社会关系主要有：同学关系、师生关系以及与社会人员的关系等。这些数据主要来自于学生信息系统、教务管理系统等。本部分首先主要采用常规处理方法，对学生的基本数据进行采集、分析、处理、转换、统计和计算，然后对有部分缺失的数据和相互矛盾的数据进行单独处理。

（二）作息特征提取

学生作息特征用于判断学生作息规律情况，以此来预测学生的学习成绩好坏。该部分的数据来源主要是学生的校园一卡通系统，包括学生的宿舍门禁、图书馆门禁、图书借阅数据、早餐消费记录、午餐消费记录、晚餐消费记录、校内营业点消费记录、校医院看病记录等等。为了更好地对这些数据进行解释，弄清楚其实际意义，需要对校园一卡通系统进行研究，并对多个不同学校、专业、年级的学生进行访谈和调研，研究和设计对应的算法来判别和提取学生的作息特征，并判断出学生的作息规律情况。

（三）网络日志特征提取

网络日志特征用于识别学生访问各种网站的内容及花费的时间。从学校网络中心服务器、路由器和防火墙等设备下载的网络日志来看，网络日志涵盖的网址数量很庞大，并含有大量的噪声数据，当前还没有成熟的研究方法来进行大规模的网络日志特征提取。网络日志特征提取的研究从以下三个方面展开：第一，对网址进行分类。结合已有的网址分类表、人工网址分类和计算机辅助网址分类等方法，将网络日志中数量庞大的网址分为主类、大类和小类三层，逐层细化和展开。第二，计算学生浏览特殊网站的时间。比如，教育学习类网站、游戏类网站和视频类网站等等，学生浏览这些网站对其成绩影响很大，提取这些特征能大大增强高风险学生预测模型的性能，因此估算学生浏览这些特殊网址时间的精度要求很高。第三，估算学习浏览其他网站的时间。学生除了浏览特殊网站外，还浏览了数量巨大的其他网站，但是这些网站对高风险学生预测模型的性能影响比较小，可采用较为粗略的估算浏览网站时间算法，同时对那些异常网站进行单独处理。

三、预测模型构建

构建高风险学生预测模型是指选择合适的学生特征和预测分类器，利用已有的学生数据对预测分类器进行训练，建立具有良好性能的高风险学生预测模型，对高风险学生进行准确的预测。该部分从以下三个方面进行研究：特征选择、模型选择和模型训练。

（一）特征选择

通过对学生数据的分析和处理，提取出学生特征，但提取的学生特征较多，同时高风险学生占整个学生的比例很小，属于典型的非均衡数据的特征提取与选择问题。为了避免过

拟合和更好地解释学生的行为与结果之间的关系，以利于指导实践，在高风险学生预测模型中不能完全采用这些提取的特征，需要选择能够准确识别出高风险学生的关键特征。

(二)模型选择

选取高风险学生预测模型是指选取合适的预测分类器来预测高风险学生。目前常用的分类器有决策树、支持向量机、神经网络、朴素贝叶斯、逻辑回归等。该部分将选取合适的候选模型，并研究组合预测模型，以便进行预测模型训练和评估，选择性能最优的高风险学生预测模型。

(三)模型训练

该部分是用已有的学生数据对各候选模型和组合模型进行训练，对它们的性能进行评估，最终选取效果最优的预测模型。要完成这部分的研究，需要采用大数据技术，解决海量数据存储和处理、模型训练与样本选择等问题。

四、预测原型系统开发

该部分是将选取的高风险学生预测模型用编程语言开发出一个可运行的软件系统，对模型进行检验和改进，以便进一步为学生、老师和教育教学管理者提供服务。该部分涉及大数据技术、软件技术选型、系统架构设计、系统功能设计、数据接口设计、界面设计和项目管理等。系统功能主要包括学生数据采集、高风险学生预测模型实现、各类用户管理、预测结果的审核发布与通知等部分。

第四节 主要创新

通过研究，本书取得了如下研究成果及创新：

1. 搭建了教育大数据平台，集成教育领域中多种数据来源

构建教育大数据平台的目的是对教育大数据进行集成、存储、运算、分析和挖掘，为大数据环境下的教育应用系统的开发和运行提供基础支撑。针对教育大数据集成中存在数据源多、安全性低、数据结构多样和数据利用率低等问题，本书提出了一个多源异构的教育大数据集成方法。引入 Web 服务、移动代理和本体技术，探讨了教育大数据服务模型。基于 Web 服务的技术优势，结合移动代理的安全特征，利用本体对多数据源的包装和全局的中介，给出了满足教育大数据集成的方法，并研发了原型系统。在实际的应用中，体现了该方法的合理性和科学性，为教育大数据的应用奠定了基础。

2. 研究了基于时间轴的高校学生基本特征提取与分析方法

为了发现高风险学生，干预并帮助这些学生顺利完成学业，以时间轴为主线，分析学生基本特征值的变化，找出高风险学生，为教育管理者提供参考。首先分析了高校学生基本特征，讨论了时间轴的特点，指出了特征值的来源，论述了基本特征值变化的分析方法及特征值的分布式存储；然后设计了高校学生基本特征值变化分析的 4 层体系架构，并介绍了其工作原理和工作流程；最后通过实例证明了该方法的有效性和正确性。该研究成果对教育工作者发现高风险学生具有应用价值。

3. 设计了基于校园一卡通的学生特征提取及作息规律判断方法