分类号： 密级：

UDC： 学号： 406629512016

南 昌 大 学 硕士 研 究 生

学 位 论 文

**移动健康管理系统中**

**心血管病患者的数据挖掘及其应用**

## Cardiovascular Disease Data Ming and Application of

## Mobile Health Management System

秦元博

培养单位（院、系）：软件学院 软件工程

指导教师姓名、职称：芦东昕 教授

指导教师姓名、职称：

申请学位的学科门类：工学

学科专业名称：软件工程

论文答辩日期：2015年5月20日

答辩委员会主席：舒坚

评阅人：校级盲审

校级盲审

2015年5月20日

**一、学位论文独创性声明**

本人声明所呈交的学位论文是本人在导师指导下进行的研究工作及取得的研究成果。据我所知，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，也不包含为获得南昌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谢意。

学位论文作者签名（手写）： 签字日期： 年 月 日

**二、学位论文版权使用授权书**

本学位论文作者完全了解南昌大学有关保留、使用学位论文的规定，同意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。本人授权南昌大学可以将学位论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存、汇编本学位论文。同时授权北京万方数据股份有限公司和中国学术期刊（光盘版）电子杂志社将本学位论文收录到《中国学位论文全文数据库》和《中国优秀博硕士学位论文全文数据库》中全文发表，并通过网络向社会公众提供信息服务，同意按“章程”规定享受相关权益。

学位论文作者签名（手写）： 导师签名（手写）：

签字日期： 年 月 日 签字日期： 年 月 日

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 论文题目 | 移动健康管理系统中心血管病患者的数据挖掘及其应用 | | | | |
| 姓 名 | 秦元博 | 学号 | 406629512016 | 论文级别 | 博士□ 硕士 |
| 院/系/所 | 软件学院 | | 专业 | 软件工程 | |
| E\_mail | 262366210@qq.com | | | | |
| 备注： | | | | | |

公开 □保密（向校学位办申请获批准为“保密”， 年 月后公开）

**摘 要**

近几十年来，国内外高血压的覆盖率、致死率仍呈上升趋势，截至2013,年，中国有2.7亿高血压患者。万幸地是，多年实践表明，可通过防控手段提升高血压的检测、治疗与控制效果。

高血压作为一种严重的心血管病，既是众多严重慢性疾病的致病因素，也是多因素综合作用的结果，针对发病、并发症事件的研究具有重要意义。国内外关于高血压发病方面的研究较为丰富，但大多是非政府主导的、范围有限的研究；关于高血压并发症方面的研究则较少，且大多基于统计学方法。

本文试图在深圳医疗大数据产业积累的基础上，以高血压的发病、患严重并发症事件的相关数据为研究对象，采用数据挖掘、数据库、统计学等方法，做了以下工作：给出1450万份居民健康档案的探查方案与结果；建立高血压数据库并提高存取性能；生成发病、并发症事件的数据集；给门诊数据做分词处理；预处理数据；建立风险模型，并评估、比较效果；用PHP实现K-M算法绘制生存曲线；探讨高血压并发症的疾病谱图。

研究重点在于建立、评估、比较风险模型。基于R平台，采用逻辑斯蒂回归、朴素贝叶斯、随机森林分别建立发病、并发症事件的风险模型，通过30次10折交叉验，计算AUC评估、比较模型。针对逻辑斯蒂回归模型，深入分析主要影响因素，得出风险公式，并回到数据集做单变量分布与分析。

**关键词：**高血压及其并发症；风险建模；生存曲线；疾病谱图；分类算法

**ABSTRACT**

In recent decades, the coverage and mortality of hypertension still shows an obviously upward trend both domestic and overseas. As of 2013, China had nearly 270 million patients with hypertension. Fortunately, years of practice has shown that detection, treatment and control can be enhanced by prevention.

As a serious cardiovascular disease, hypertension is not only risk factor of many serious chronic diseases, but also result of the combined effects of multiple factors. The study on hypertension incidences and complication is significant. There are abundant studies on hypertension incidences both domestic and overseas, while most of them are non-government-led and of limited range. What’s more, studies on hypertension complications are of less quantity and are based on statistical methods mostly.

This paper attempts to study hypertension and its complications based on Shenzhen large medical data. The following work has been done. Give explore scheme and result of 14.5 million health records of residents. Establish hypertension database and improve access performance. Generate specialized datasets. Segament outpatient texts and preprocess data. Establish, evaluate and compare risk models of hypertension incidences and serious complicaitons. Realize K-M algorithm with PHP and draw survival curves. Discuss the spectrum of hypertension and its complications.

This study focused on the establishment, evaluation and comparison of risk models. Based on R platform, it established models using logistic regression, Naive Bayes and random forest respectively, evaluated and compared models using 10-fold cross-validation by 30 times. It also analysed main factors, derived risk formula and do univariate distribution and analysis on datasets.

**Key Words:** hypertension and complications; risk model; survival curve; hypertension spectrum; classification algorithm

**目 录**

[第1章 引言 1](#_Toc421047928)

[1.1 研究背景 1](#_Toc421047929)

[1.2 移动健康管理系统 3](#_Toc421047930)

[1.3 国内外研究现状 4](#_Toc421047931)

[1.3.1 健康管理 4](#_Toc421047932)

[1.3.2 心血管病患者的数据挖掘及其应用 6](#_Toc421047933)

[1.4 研究内容和创新 8](#_Toc421047934)

[第2章 基础知识和理论 10](#_Toc421047935)

[2.1 高血压及其并发症 10](#_Toc421047936)

[2.2 分类算法 11](#_Toc421047937)

[2.2.1 逻辑斯蒂回归 11](#_Toc421047938)

[2.2.2 朴素贝叶斯 12](#_Toc421047939)

[2.2.3 随机森林 14](#_Toc421047940)

[2.3 模型评估方法 15](#_Toc421047941)

[2.3.1 10折交叉验证 15](#_Toc421047942)

[2.3.2 AUC与ROC 15](#_Toc421047943)

[第3章 数据库建立与优化 18](#_Toc421047944)

[3.1 数据库建立 18](#_Toc421047945)

[3.1.1 数据库需求和特征评分标准 18](#_Toc421047946)

[3.1.2 数据类型定义规则 19](#_Toc421047947)

[3.1.3 建立高血压数据库 20](#_Toc421047948)

[3.2 Merge引擎分表 21](#_Toc421047949)

[第4章 数据处理 23](#_Toc421047950)

[4.1 数据存在性与分布情况 23](#_Toc421047951)

[4.2 研究对象 25](#_Toc421047952)

[4.3 数据集生成 26](#_Toc421047953)

[4.3.1 分组规则 26](#_Toc421047954)

[4.3.2 数据集生成规则 28](#_Toc421047955)

[4.4 门诊数据的分词 30](#_Toc421047956)

[4.5 数据预处理 31](#_Toc421047957)

[4.5.1 数据清洗 31](#_Toc421047958)

[4.5.2 数据转换 33](#_Toc421047959)

[4.5.3 数据规约 35](#_Toc421047960)

[第5章 风险建模及评估 36](#_Toc421047961)

[5.1 数据集介绍 36](#_Toc421047962)

[5.2 风险建模 37](#_Toc421047963)

[5.3 模型评估 38](#_Toc421047964)

[5.4 模型比较 42](#_Toc421047965)

[5.5 结果与分析 46](#_Toc421047966)

[5.5.1 高血压发病事件 46](#_Toc421047967)

[5.5.2 高血压患严重并发症事件 52](#_Toc421047968)

[第6章 生存分析和疾病谱图 59](#_Toc421047969)

[6.1 生存分析 59](#_Toc421047970)

[6.1.1 K-M算法及实现 59](#_Toc421047971)

[6.1.2 结果与分析 61](#_Toc421047972)

[6.2 疾病谱图 63](#_Toc421047973)

[6.2.1 数据计算 63](#_Toc421047974)

[6.2.2 结果与分析 64](#_Toc421047975)

[第7章 结论与展望 66](#_Toc421047976)

[7.1 结论 66](#_Toc421047977)

[7.2 进一步工作的方向 68](#_Toc421047978)

[致 谢 69](#_Toc421047979)

[参考文献 70](#_Toc421047980)

**第1章 引言**

**1.1 研究背景**

高血压是全球普遍存在的重大公共卫生问题，西欧的覆盖率为44%，北美的覆盖率为28%。到2025年，在世界人口范围内，75%的高血压化将发生在发展中国家[[[1]](#endnote-2)]。高血压是心脑血管病、肾功能衰竭等常见慢性疾病的常见致病因素，约半数的脑卒中和心肌梗死与它有关[[[2]](#endnote-3)]。而在我国，每年有200万以上的人死于高血压，具有高致残率、高致死率, 直接医疗费用366亿元/年[[[3]](#endnote-4)]。

至2012，全国有2.9亿心血管病患者，包括2.7亿高血压、700万脑卒中、450万心力衰竭，250万心肌梗死。全国心血管病死亡率呈上升趋势，2012年心血管病死亡率为255万，平均10秒钟就有1人死于心血管病病[[[4]](#endnote-5)]。包括高血压在内的心血管病死亡占农村38.7%，城市41.1%，加强政府主导下的心血管病防治刻不容缓，心血管病死亡率如图1.1所示：

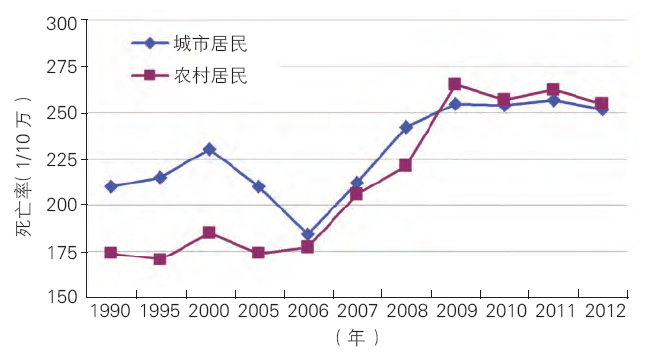


图1.1 1990 ~2012年中国心血管病死亡率

我国15岁及以上人口高血压患病率高达24%，如图1.2所示。至2012年，我国将拥有2.66亿高血压患者[]。

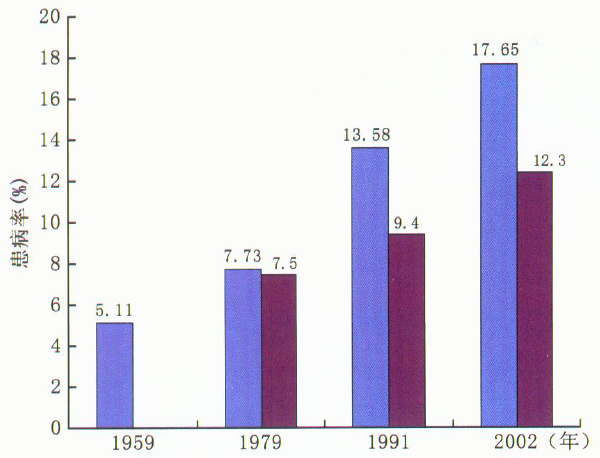


图1.2 我国成年人高血压患病率

国内外多年实践证明，防控手段能有效维持血压，继而减少因高血压而导致的中风、心力衰竭、缺血性心脏病、肾功能衰竭等慢性疾病。至2002为止的十年间，我国高血压患者的检测率、治疗率、控制率分别上升了3.9%、12.6%、3.3%[[[5]](#endnote-6)]。

近年来，医疗领域信息化建设持续推进，产生大量医学健康数据，不仅包含居民健康档案、电子病历等个人医疗健康信息，还包括多种涉及国家安全的公共卫生事件管理信息。基于医学健康数据的统计和分析能够为居民提供个性化医疗健康管理方案，同时为区域卫生管理者提供有关疾病防控、医疗行为监督的决策支持。

科技部将“数字化医疗技术研究”纳入重点发展方向，设立了新的医改阶段目标“2020年，覆盖城乡居民的基本医疗卫生基本制度基本建立”，提出了“强化医学科技信息的加工处理、深入分析和信息产品推送能力，为推动医学科技的创新发展提供信息支撑”等方针[[[6]](#endnote-7)]。

区域医疗系统涵盖了上百万居民人口和上百家医疗机构的医疗数据，数据量随时间推移快速增长。以深圳市为例，2011年的统计数据表明，深圳市全市常住总人口为1046.74万，全市有卫生医疗机构2465家[[[7]](#endnote-8)]，其中医院110家，社康中心611家，疾病控制中心8家，妇幼保健院7家，诊疗8878.0万人次，其中急诊683.5万人次，收治住院病人96.2万人次，病床使用率89.7%。每个病人的电子病历、影响归档和生理信号等健康数据每天都在增加。

目前深圳市医疗信息系统的存量医疗数据已超过100TB，仅电子影像数据便以每年数10TB的速度飞速增长。作为公共医疗信息建设较完善的地区，对医疗及健康大数据分析与应用的需求更为迫切。

国内外的探索经验已经表明，对医疗资源的深入应用可广泛提升区域医疗卫生事业。我国目前医疗健康信息化建设集中存在以下不足：

1、信息利用率不高：我国医疗卫生领域的信息化建设重点仍然放在电子病历和医疗档案的录入、存储、检索以及简单统计上，缺乏对医疗数据资源的深入价值挖掘。

2、应用途径单一：现有医疗信息化系统的用途大部分只限于单向的收集数据和统计结果，缺乏从医疗健康大数据中总结规律、获取知识及情报的手段，更缺乏根据所获得的信息及时进行主动管理、反馈及干预的手段。

通过建立各种健康预测、管理和治疗模型，可为慢病与亚健康的人群提供风险评估、康复管理技术及服务，提高信息利用率，增加应用途径。

为了实现这一目标，如何高效处理、存储、分析、整合蕴含大量应用价值的医疗和健康大数据，从中挖掘获取具有广泛医学意义的知识和结论，并形成对经济及社会具有广泛影响的应用，成为亟待解决的问题。

**1.2 移动健康管理系统**

本文为面向区域医疗和公共卫生的健康大数据处理分析研究及示范应用课题下一部分，数据来源于深圳市医学信息中心，课题主要研究面向区域医疗应用的健康大数据挖掘与建模。

深圳市医学信息中心是深圳市卫计委直属事业单位，2005年获得深圳市政府投资项目立项，制定“深圳医疗卫生139工程”。作为深圳市区域卫生信息化平台的所有者，提供项目实施所需的区域卫生医疗数据及管理服务需求方面的指导。数据覆盖了全市1450万人的居民健康档案（历史1600万份，超30亿条诊疗记录），接入了全市68家公立医院（含所有三甲医院）及近600家社康中心。

金证股份已经成功开发了便携式全科医生工作站系统，集全科医疗检查和信息化于一体，方便医务人员出诊、入户随访和开展公共卫生服务，同时在全国10多个省、4000多个村卫生室使用，如图1.3所示：



图1.3 全科医生工作站

移动健康管理系统建立在在全科医生工作站和区域医疗健康信息系统的基础上。依托该平台，针对慢病高危因素挖掘及预后评估问题，研究如何从覆盖1450万居民的健康档案中发现慢病的高危因素及对应诊疗保健方案，研究如何以直接观测到症状体征作为观察变量，将病人的健康状态进行标量化，从健康大数据中得到健康状态，并发现有用的个性化诊疗保健信息。

由于不同基层医护人员的行医经验差别较大，且慢病等区域主要疾病存在一定的个体差异性，课题基于不同患者基本情况及病情发展做出不同的研究，本文以高血压患者的有关数据为研究对象，具体研究高血压发病、并发症的有关情况，并给出包括风险预测、生存曲线、疾病谱图、保健指导等方面的分析。

高血压及其并发症的有关研究成果一方面可通过全科医生工作站服务于基层和社区医院，另一方面可通过区域医疗健康信息系统应用到本地众多公立医院，保障更多居民的健康。

**1.3 国内外研究现状**

**1.3.1 健康管理**

一、国内外发展情况

现代医疗以防治为主，保健为辅，这与健康管理的理念相符。健康管理在发达国家发展较早、服务覆盖面广。1929年美国成立世界上第一个健康管理组织[[[8]](#endnote-9)]，有约70%的人群每年体检一次，并做相应的健康管理服务，有效降低慢病发病和死亡。在欧洲，大多数国家已经建立国家主导的全民医疗卫生服务系统。英国已建立了覆盖每位居民的健康档案，平均每位医生覆盖1500位居民。

在中国，健康管理起步较晚，2001年国内第一家健康管理公司注册成立，2008年上半年。但至今，有关的健康管理机构已有五千多家。健康管理在中国市场有巨大发展潜力，已成为新兴朝阳产业。

二、国内外研究现状

当今健康管理的主题是慢病管理，也是许多研究的重点所在。慢病通常病程长且发展缓慢，影响范围大且易导致死亡，严重影响患者及其家庭的生活质量，耗费大量医疗费用，所带来的经济负担已达到所有疾病的46%[[[9]](#endnote-10)][[[10]](#endnote-11)]。同时，全球人口老龄化问题不断加剧，预计2050年我国超过60岁的老人比例将高达22.7%[[[11]](#endnote-12)]，随之引发的则是慢性疾病的发生及覆盖率的不断增长，以及对健康监护的巨大需求。

在常见慢病中，高血压不仅多种其他慢病的致病因素，也是多种因素共同作用的结果，不仅需在发病时进行治疗，更需要长期跟踪监护。

然而，传统的医院就诊方式无法长期有效的检测高血压的变化及发展过程，并不适合高血压的诊断与治疗。医疗信息服务商可以提供一些相关的服务，例如，可通过对高血压患者的数据进行深入挖掘，得出糖尿病的主要影响因素。同时，也可通过对病人前期的各种病例数据以及其他患者的康复数据进行数据挖掘，分析与评估出最优化的康复保健方案。

在国外，Intel与IBM公司都曾尝试用数据挖掘方法进行糖尿病人的病情管理及医师推荐[[[12]](#endnote-13)][[[13]](#endnote-14)]，并发表在数据挖掘顶级会议KDD上[]。近年来健康管理系统在心衰[[[14]](#endnote-15)][[[15]](#endnote-16)][[[16]](#endnote-17)]、高血压[[[17]](#endnote-18)]、糖尿病[][[[18]](#endnote-19)][[[19]](#endnote-20)]、心血管疾病[[[20]](#endnote-21)][[[21]](#endnote-22)][[[22]](#endnote-23)]等多种慢性疾病的管理中得到较多的应用。但是此类研究仅在小规模人群上进行了跟踪，在医学上的可靠性有待检验，远未实现针对普遍人群的健康管理与疾病预警。

在我国，健康管理中有超过90%还以体检为主，无法提供更深入的健康评估报告、跟踪情况、干预治疗等更进一步的服务。

随着以数据挖掘为代表的各项研究与技术的成熟，健康管理行业的发展呈现三大趋势——信息化、标准化、个性化。信息化以电子健康档案、电子病历为代表，重点在于信息系统的建立与管理；标准化以分布式数据库、数据融合等为代表，重点在于标准化数据存取与流程；个性化以大大小小新兴的健康管理软件为代表，如掌上春雨医生等，提供适合个人的、个性化的服务。

**1.3.2 心血管病患者的数据挖掘及其应用**

一、 医疗数据挖掘

2013年是世界大数据元年，如今的大数据表现出大量、多样、快速、有价值[[[23]](#endnote-24)]这四大趋势。数据挖掘技术的进步为医疗领域带来巨大的变革动力。

医疗数据挖掘在公共卫生领域的应用是当前的一个研究热点。近年来，国际上许多国家的政府或专业卫生组织都基于数据挖掘技术建立了多个公共卫生监控系统[[[24]](#endnote-25)]，较为流行的包括如下系统：

1、 NEDSS[[[25]](#endnote-26)] (National Electronic Disease Surveillance System）

2、 ProMED-mail[[[26]](#endnote-27)](http://www.promedmail.org)

3、 GPHIN[[[27]](#endnote-28)] (Global Public Health Intelligence Network)

4、 HealthMap[[[28]](#endnote-29)] (http://www.healthmap.org)

5、 MediSys[[[29]](#endnote-30)] (http://www.medusa.jrc.it)

6、 EpiSPIDER[[[30]](#endnote-31)] (www.epispider.org)

7、 BioCaster[[[31]](#endnote-32)] (http://biocaster.nii.ac.jp)

8、 NBII Wildlife Disease Information Node[]

此外，Google、Facebook、Twitter等互联网巨头都通过对社交网络数据进行深入挖掘分析，实现对流感等传染性疾病进行预警与跟踪。

例如，Google的流感爆发预测软件Google Flu[[[32]](#endnote-33)]，比较了2003~2008年季节性流感的传播时期的数据和美国人最常检索的5000万词条，通过分析搜索记录预测流感爆发情况。基于Google地图的禽流感全球跟踪插件[[[33]](#endnote-34)]就在Nature上发表并在学术界引起较大反响。然而，今年的跟踪验证结果表明[[[34]](#endnote-35)]，由于社交网络的数据不稳定性，上述技术在预测效果上都存在较大缺陷，某些情况下与真实数据相距甚远。

在众多公共卫生监管系统中，比较具有代表性的是美国的国家疾病监测信息系统NEDSS。该系统首先创建标准以保证全国范围手机的数据格式统一，从而使得大型机构通过电子数据交换就可实现疾病报告的自动化。从收集数据的层面统一标准是其成功与特色之处。

但是，该系统侧重信息的收集、交换与汇报，在数据挖掘及分析预警方面的工作相对滞后。

在我国，区域医疗目前建设的重点在于各级医疗机构数据的共享、信息的互联互通。随着医院信息系统(HIS)、电子病历(EMR)、影像归档和通信系统(PACS)、实验室信息管理系统(LIS)等信息化建设的日趋成熟，海量医疗和健康数据汇集到区域医学信息中心，且医疗健康数据的完整性有了很大提高。这就为基于医疗数据的数据挖掘提供了新的机遇和挑战。

二、心血管病患者的数据挖掘

国际上最著名的从事高血压等心血管病的数据挖掘机构当属在NHLBI (National Heart, Lung and Blood Institute)的指导下，于1948年成立Framingham Heart Study[[[35]](#endnote-36)]。通过持续跟踪参与实验的心血管病患者，获取高质量的定制数据并在高血压的数据挖掘领域取得瞩目成就，首次提出肥胖是高血压发病的主要风险因素之一[[[36]](#endnote-37)]。

我国已拥有海量的有关心血管病诊疗的临床数据，在管理数据、统计分析上已具备一定的基础，但运用数据挖掘技术进行心血管病患者的数据挖掘及其应用的水平还比较薄弱，大都还停留在基于统计学方法的阶段，研究内容也多停留在疾病的发病事件上，基于医疗数据挖掘的研究相对比较少，其中对并发症及其相互关系的研究就更少。

数据挖掘已经广泛运用到金融和零售等行业[[[37]](#endnote-38)]，技术相对成熟，将其应用到心血管病数据的分析与挖掘具备技术可行性。

同时，我们也注意到医疗数据挖掘有别于其他领域的数据挖掘[[[38]](#endnote-39)]，具有数据量大、来源复杂、信息载体多样化、缺失情况严重、部分带有时序性、隐私性强、数据库冗余等特性，处理时要注意数据预处理，尤其是数据清洗、过滤及去隐私化等。

数据挖掘方法通常取决于兴趣方向和数据类型：对于预测型数据，通常采用分类、回归、时间序列分析等；对于描述型数据，通常采用聚类、关联规则分析、序列分析等[[[39]](#endnote-40)]。最常使用的算法为ID3/CART、Apriori系列、模糊算法等，及近年引入的医疗领域的遗传算法、神经网络、粗糙集等。

本文研究的高血压发病及严重并发症事件属于疾病诊断问题，归属于预测型数据的分类问题。在预测型数据的分类问题的研究上，Pasolli等根据MIT-BIH心律失常数据库，采用SVM、KNN、ANN这三种分类器分别对ECG波形进行分类，最高达到了86%的准确率[[[40]](#endnote-41)]。

**1.4 研究内容和创新**

针对上述需求，本文在深圳市居民健康档案、全科医生工作站、区域医疗健康信息系统的基础上，使用数据库技术、数据挖掘技术、统计分析、医学知识等，探查、分析已获取的居民健康档案，建立高效、专用的高血压数据库；通过数据预处理方法生成符合研究需求的高质量数据集；以高血压人群的居民健康档案为主要研究对象，对高血压患者的诊疗数据进行分词；建立高血压的发病、患严重并发症的风险模型，实现高血压发病事件和患严重并发症事件的预警，采用30次10折交叉验证，计算AUC并绘制ROC曲线，评估模型，并重点分析逻辑回归模型；绘制生存曲线，研究患有不同程度高血压的人群的主要风险因素随时间变化的生存率；探索疾病谱图，研究高血压人群确诊后患上其他病症的情况，初步探究病症，尤其是高血压严重并发症之间的关系。整合多个模型，研究成果将作为移动健康管理系统的一部分，保障更多人的健康。

本文研究的数据由深圳市医学信息中心和深圳市金证科技股份有限公司共同提供，由深圳市68家公立医院（含所有三甲医院）及近600家社康中心负责采集，覆盖了全市1450万人的居民健康档案，历史数据共1600万份，超过30亿条诊疗数据。

研究的第一步工作是探查居民健康档案，生成高血压数据探查方案，并作为糖尿病、冠心病等后期研究时探查数据的模板。分析数据探查结果，核实数据的可用性，明确具体变量的可用样本量、变量的取值分布和时间分布、表间对应情况等。

提取居民健康档案中高血压患者的有关数据，辅以健康人、糖尿病、肿瘤等人群的基本信息，采用MySQL建立专用的高血压数据库，通过添加索引、Merge表等数据库方法，提高数据库存取效率。

在专用数据库的基础上，按照研究需要进行详细的数据探查，在探查结果的基础上从数据库中分别筛选高血压发病事件、严重并发症事件的实验组、对照组，生成发病数据集、严重并发症数据集。

数据库中除了传统数值型数据，还存有文本、影像等。其中，门诊诊疗数据为文本型，使用ICTCLAS2015工具做分词处理，得出确诊率最高的病症，以危险性、致死率为标准找出其中常见的严重并发症，为后续研究做铺垫。

高血压发病、严重并发症这两个事件都属于二分类问题，本文对二者进行风险建模的方法类似，都是采用如下三种分类算法进行建模与研究：逻辑斯蒂回归、朴素贝叶斯和随机森林。

为了得到可靠稳定的模型，采用十折交叉验证，计算AUC并绘制ROC曲线评估建模结果，并回到数据集，做单变量分布与分析。

研究居民发病、患严重并发症随随变化的情况，采用K-M算法，分析生存率并绘制生存曲线。

研究同时患有以并发症为重点的多种疾病的高血压人群所患疾病的时间先后关系，建立疾病谱图，探究高血压、并发症、其他疾病之间的多元关系。

综上所述，本文以与高血压发病、患严重并发症这两类事件有关的数据为研究对象，经过建立专用慢病数据库、生成相应数据集、数据处理、风险建模及评估、计算生存曲线、绘制疾病谱图等步骤，最终获得可靠稳定的发病风险模型、并发症风险模型、相应生存曲线和高血压疾病谱图，高血压发病事件、患严重并发症事件的主要影响因素及分析结果。

**第2章 基础知识和理论**

**2.1 高血压及其并发症**

一、高血压

血压的常规测量项目为收缩压(SBP, Systolic Blood Pressure)、舒张压(DBP, Diastolic Blood Pressure)。目前我国实行《2000年中国高血压治疗指南标准》，根据该标准，高血压分级如表2.1所示：

表2.1 高血压分级

|  |  |  |
| --- | --- | --- |
| 类别 | 收缩压/mmHg | 舒张压/mHg |
| 正常血压 | <120 | <80 |
| 正常高值 | 120~139 | 80~89 |
| 高血压 | ≥140 | ≥90 |
| 1级高血压 | 140~159 | 90~99 |
| 2级高血压 | 160~179 | 100~109 |
| 3级高血压 | ≥180 | ≥110 |
| 单纯收缩期高血压 | ≥140 | ≥90 |

以收缩压、舒张压中较高的级别为准。本文划分健康人群和高血压人群时，采用严格的高血压定义，即正常高值不算健康人群。

二、高血压并发症

高血压在早期属于可控、不可愈、对生活影响小的疾病，但随着时间推移，它会逐渐破坏人体的正常循环，危害身体机能损害重要器官，继而造成冠心病、糖尿病、中风等严重并发症，严重时甚至威胁到生命安全。

通常，根据受损器官不同，高血压并发症主要有为心脏病、脑血管病、肾病、大小动脉病、眼病这几类。且高血压经常同时破坏多个靶器官，并发多种疾病，严重威胁人类健康。

研究时，考虑到血压、血糖代谢、血脂代谢、胰岛分泌对人体的相互作用，本文参考糖尿病、冠心病的并发症，将其列入病症范围加以考虑。三种慢病及其并发症关系如图2.1所示：



图2.1 高血压、糖尿病、冠心病及其并发症关系

其中，高血压主要并发症如下：

1、 心脏病：心力衰竭、心肌梗死、左心室肥厚

2、 脑血管病：脑卒中

3、 肾病：肾功能衰竭、肾硬化

4、 大小动脉病：动脉硬化、主动脉夹层

5、 眼病：视网膜动脉硬化

**2.2 分类算法**

本文采用逻辑斯蒂回归、朴素贝叶斯和随机森林三种分类算法探究高血压的发病事件、严重并发症的建模效果。

**2.2.1 逻辑斯蒂回归**

回归方法通常用于描述响应变量与一个或多个解释变量之间的关系。逻辑回归模型和线性模型的最大区别在于，逻辑回归模型的结果变量为二项型[[[41]](#endnote-42)]。

二项逻辑斯蒂回归模型条件分布如下[[[42]](#endnote-43)]：

(2.1)



(2.2)



这里，是输入，是输出，和是参数，称为权值向量，b称为偏置，为和的内积。



给定训练的数据集，由逻辑斯蒂回归模型我们得到的对数似然函数为



(2.3)



本文中使用牛顿方法来求解，下面简要介绍牛顿方法。

我们在数值分析方法中知道，求解时，如果f可导，那么可以通过迭代公式



(2.4)



来求解。



当我们求解极大似然函数的最大值，就转换成了求极值的问题，进而也就是求解的问题，那么求解的迭代公式为



(2.5)



当为向量时，牛顿法可表示为



(2.6)



其中是的Hessian矩阵。



**2.2.2 朴素贝叶斯**

朴素贝叶斯算法是一种简单常用的分类器，它基于贝叶斯定理和特征之间条件独立假设。对于输入的样本集，首先基于特征之间条件独立条件算出输入和输出之间的联合概率分布，然后对于输入的样本，算出其最大后验概率分布的[]。



我们给定样本,其中为维度为的向量集合，另外我们定义类别集合分别为。我们定义是的联合概率分布。加入我们有如下的训练集：



(2.7)



朴素贝叶斯方法根据训练集来学习上面提到的联合概率分布。首先我们需要得到先验概率分布：

(2.8)



另外我们得到基于的的条件概率分布:



(2.9)



根据条件概率分布公司，我们可以得到联合概率分布。朴素贝叶斯分类器基于特征独立假设，基于的的条件概率分布可以写成如下形式：



(2.10)



朴素贝叶斯基于特征独立这一假设来获得X,Y之间的联合概率分布，实际上在生活中我们知道绝大多数特征之间实际上是有关系的，特征独立会让朴素贝叶斯模型变得比较简单，使用效率大大提高。

当我们需要使用朴素贝叶斯进行分类的时候，对于输入，我们需要计算后验概率分布，根据贝叶斯定理可以得到：



(2.11)



带入基于条件独立假设的的的条件概率分布，得到：



(2.12)



我们可以得到的后验概率分布，于是得到朴素贝叶斯的基本可以表示为：



(2.13)



我们可以看到上式中的分母实际上是相同的，所以可以得到：

(2.14)



**2.2.3 随机森林**

在介绍随机森林之前，首先介绍一下决策树的概念。决策树是一种树形的分类器，内部节点代表一个属性，可测试节点属性值，不同的内部节点分支输出不同的属性值，叶节点代表输出结果。

比如图2.2中的决策树，判断是否出门。首先判断天气属性，如果属性值为晴朗，则继续判断潮湿属性，如果湿度高则不出门，否则出门。由决策树可以推出随机森林的概念。

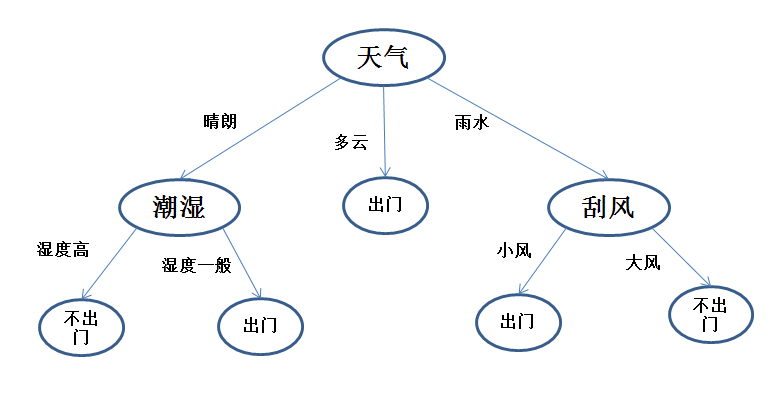


图2.2 判断是否出门的决策树例子

在L.Breiman给出了明确的定义[[[43]](#endnote-44)]：随机森林是一个分类器，由多个树结构分类器组成，其中是互不关联且同一分布的随机向量.每棵树为输入X投票，以选出最受欢迎的类。图2.3给出了一个随机森林算法的示意图：

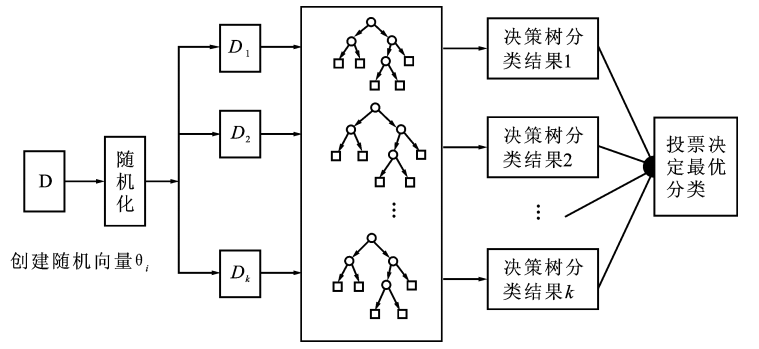


图2.3 随机森林算法示意图

随着森林中树的数目的增加，森林的泛化误差会收缩至一个临界值。森林的泛化误差由森林的强度以及森林的相关系数决定。使用随机特征选择分裂节点的错误率与Adaboost相近，但是面对噪声它更健壮。内部评估监测错误，强度，关联度等，这些参数用来对分裂时增加特征数时作出反应。内部评估也用来估量变量的重要性。

随机森林在运行的速度方面有着显著的优势，且不用担心过拟合问题。你可以根据实际情况的需要来生成任意多棵树。随机森林在当前的分类算法的一个研究热点。

**2.3 模型评估方法**

**2.3.1 10折交叉验证**

为了得到稳定可靠的模型，通常选用某种评价标准对候选集进行评估。

给定样本量为N的训练数据，若候选集记为，有效模型记为，从中选出指定该指标最小的模型[[[44]](#endnote-45)]即



(2.15)



是评价标准，评价标准是是有效模型的概率。常用模型选择方法有交叉验证、holdout验证、留一验证。



交叉验证（Cross-Validation），也称为交叉比对，是完整的样本空间中按一定该比例划分训练集和测试集，对测试集进行预测并求出误差，记录平方和。重复操作直至全部样本均有且仅有一次预测。样本的预测误差平方和记为PRESS(Predicted Error Sum of Sqares)。

通常采用5折、10折或20折交叉验证，效果较为明显[[[45]](#endnote-46)]。本文采用30次10折交叉验证，每次将数据集分成10份，轮流将其中9份做训练，1份做测试，并最终覆盖全部数据，10次结果的均值作为对算法精度的估计。

**2.3.2 AUC与ROC**

AUC（Area Under the Curve）来源于ROC曲线。ROC是接受操作特征曲线的缩写（Receiver Operating Characteristic curve）。ROC一般用于对二元分类器进行评判，并且分类器输出的是一个连续的值，通过定义阙值的方式实现样本分类。

比如需要预测人是否患有某种疾病，输出的是概率，我们可以定义概率大于0.5为有某种疾病，小于0.5是健康。

在详细了解ROC曲线是我们需要了解一下混淆矩阵的概念，如表2.2所示：

表2.2 混淆矩阵

|  |  |  |
| --- | --- | --- |
| 预测输出/真实值 | p | n |
| p' | 真阳性(TP) | 假阳性(FP) |
| n' | 假阴性(FN) | 真阴性(TN) |

假如我们要预测一个人是否有高血压，可能有下面四种情况：

1、真阳性（TP）：预测患有高血压，实际患有高血压。

2、假阳性（FP）：预测患有高血压，实际上没有高血压。

3、真阴性（TN）：预测没有患有高血压，实际上也没有高血压。

4、假阴性（FN）：预测没有患有高血压，实际上患有高血压。

我们将这四种情况放入到表中，就得到了混淆矩阵，帮助我们对这四种情况进行分析，还可得到假阳性比率（FPR）和真阳性比率（TPR）。

(2.16)



(2.17)



在ROC曲线中，定义FPR为X轴，TPR为Y轴。由于阙值是连续的，不同的阙值可以得到相应的(x,y)坐标，画出相应的ROC曲线。在对ROC曲线进行分析时，往往会和（0，0），（1，1）连成的直线进行比较，如图2.4所示：

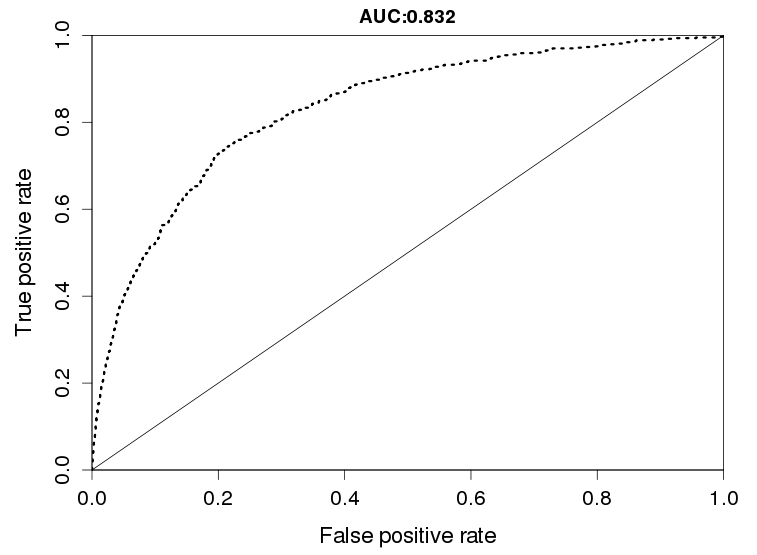


图2.4 一个简单的ROC曲线

当点在直线上方时，表示该预测战胜了随机猜测，具备一定的预测效果，否则是没有战胜随机预测，不具备预测效果[[[46]](#endnote-47)]。

另外一个常用的概念是AUC，即ROC曲线下方的面积，AUC的值在0到1之间，其表示的意义是随机的从样本中抽取一个阳性样本和一个阴性样本，分类器进行分类时，阳性样本输出的值比阴性样本输出的值大的机率即为AUC的值。AUC能够克服由于样本不平衡造成的评价指标偏倚问题。比如单纯计算准确率，阳性样本占比99%，阴性样本占比1%，假设一个分类器预测所有的样本为阳性，即使准确率有99%，显然这样的分类器是没有预测效果的。

**第3章 数据库建立与优化**

**3.1 数据库建立**

**3.1.1 数据库需求和特征评分标准**

本文研究的数据来自深圳市68家公立医院（含所有三甲医院）及近600家社康中心，覆盖了全市1450万人的居民健康档案，历史数据共1600万份，总计超过30亿条诊疗数据，建立数据库前需要定义数据库需求和特征评分标准。

数据库需求包括目标样本（横向）及目标特征（纵向）。特征评分标准综合考虑了项目需要和文献调研结果。

一、 数据库需求

1、目标样本

本文以高血压人群、健康人群作为数据库的目标人群，按照一名患者对应一份样本的原则，以档案标识符为线索查找其他数据。

其中，高血压人群的档案标识符由高血压随访表得出，该表包含了所有高血压患者。健康人群则要复杂些，需从整个居民健康档案中剔除患有各类慢病、重大疾病的人群，再剔除各类体检中血压值不正常的人群，剩余的便是健康人群。

2、目标特征

居民健康档案主要包括九大数据模块，以其中七个数据模块中的部分数据表为对象，按照一定评分标准提取有关特征。

数据表概况如表3.1所示：

表3.1 数据表概况

|  |  |
| --- | --- |
| 模块名称 | 需研究的数据表 |
| 基本信息 | 个人生活方式、既往史、市民个人信息、家族遗传史、吸烟史、现病史、现在健康主要问题、饮酒史、职业病危害因素接触史、中医体质辨识 |
| 健康体检 | 查体体检结果、[健康体检信息表](#表结构!B78)、[特殊人群辅助检查体检结果](#表结构!B121)、[一般人群辅助检查体检结果](#表结构!B172)、[一般情况体检结果](#表结构!B251)、[脏器功能体检结果](#表结构!B273)、[中医体质辨识结果](#表结构!B289) |
| 门诊诊疗 | [门诊诊疗基本信息](file:///H:\qin\硕士毕设\20140717数据字典表\3.%20门诊诊疗.xlsx#表结构!B1)、[健康档案总索引表](file:///H:\qin\硕士毕设\20140717数据字典表\3.%20门诊诊疗.xlsx#表结构!B18)、[症状（主诉、病史）](file:///H:\qin\硕士毕设\20140717数据字典表\3.%20门诊诊疗.xlsx#表结构!B49)、[疾病诊断表](file:///H:\qin\硕士毕设\20140717数据字典表\3.%20门诊诊疗.xlsx#表结构!B171) |

续表3.1 数据表概况

|  |  |
| --- | --- |
| 模块名称 | 需研究的数据表 |
| 住院诊疗 | [健康档案总索引表](file:///H:\qin\硕士毕设\20140717数据字典表\4.%20住院诊疗.xlsx#表结构!B1)、[症状（主诉、病史）](file:///H:\qin\硕士毕设\20140717数据字典表\4.%20住院诊疗.xlsx#表结构!B30)、[检验](file:///H:\qin\硕士毕设\20140717数据字典表\4.%20住院诊疗.xlsx#表结构!B48)、[手术/操作记录表](file:///H:\qin\硕士毕设\20140717数据字典表\4.%20住院诊疗.xlsx#表结构!B86)、[疾病诊断表](file:///H:\qin\硕士毕设\20140717数据字典表\4.%20住院诊疗.xlsx#表结构!B124)、[检查结果记录表](file:///H:\qin\硕士毕设\20140717数据字典表\4.%20住院诊疗.xlsx#表结构!B144)、住院诊疗基本信息表 |
| 住院病案 | [健康档案总索引表](file:///H:\qin\硕士毕设\20140717数据字典表\5.%20住院病案.xlsx#表结构!B9)、[疾病诊断表](file:///H:\qin\硕士毕设\20140717数据字典表\5.%20住院病案.xlsx#表结构!B83)、[住院病案信息表](file:///H:\qin\硕士毕设\20140717数据字典表\5.%20住院病案.xlsx#表结构!B102) |
| 疾病管理 | [高血压病人档案表](file:///H:\qin\硕士毕设\20140717数据字典表\6.%20疾病管理.xlsx#表结构!B1)、[高血压患者随访记录表](file:///H:\qin\硕士毕设\20140717数据字典表\6.%20疾病管理.xlsx#表结构!B24)、[老年人健康随访记录表](file:///H:\qin\硕士毕设\20140717数据字典表\6.%20疾病管理.xlsx#表结构!B103)、[市民个人信息](file:///H:\qin\硕士毕设\20140717数据字典表\6.%20疾病管理.xlsx#表结构!B137)、[糖尿病患者随访记录表](file:///H:\qin\硕士毕设\20140717数据字典表\6.%20疾病管理.xlsx#表结构!B193)、[肿瘤患者随访记录表](file:///H:\qin\硕士毕设\20140717数据字典表\6.%20疾病管理.xlsx#表结构!B265)、[重性精神疾病患者随访记录表](file:///H:\qin\硕士毕设\20140717数据字典表\6.%20疾病管理.xlsx#表结构!B330) |
| 疾病控制 | [职业健康检查信息表](file:///H:\qin\硕士毕设\20140717数据字典表\7.%20疾病控制.xlsx#表结构!B381)、[职业性放射性疾病报告卡](file:///H:\qin\硕士毕设\20140717数据字典表\7.%20疾病控制.xlsx#表结构!B420) |

其中，基本信息模块的个人生活方式表已拆分为三个子表，分别为吸烟史、饮酒史、运动史，且仅保留最近更新的数据。

生活方式相关的历史数据的缺乏给以生活方式为重心的高血压患者的数据挖掘带来非常大的负面影响，在划分发病事件、严重并发症事件的实验组和对照组时，要注意考查生活方式相关特征的数据量。

二、特征提取评分标准

根据项目需要和文献调研结果，将有潜力的特征进行分级，如表3.2所示：

表3.2 分级标准

|  |  |  |  |
| --- | --- | --- | --- |
| 来源 | 1级 | 2级 | 3级 |
| 项目需要 | 必备 | 需要 | 参考 |
| 文献调研 | 价值高 | 一定价值 | 参考价值 |

特征评分标准的部分内容如表3.3所示：

表3.3 特征评分标准中一般情况体检结果部分

|  |  |  |
| --- | --- | --- |
| 字段名 | 文献分级 | 项目分级 |
| 身高(cm) | 1 | 1 |
| 体温(℃) | 2 | 2 |
| 体重（kg） | 1 | 1 |
| 体重指数 | 1 | 1 |
| 臀围(cm) | 1 | 1 |
| 腰围(cm) | 1 | 1 |
| 收缩压(mmHg) | 1 | 1 |
| 舒张压(mmHg) | 1 | 1 |
| 脉率（次/分钟） | 1 | 1 |
| 呼吸频率（次/分钟） | 2 | 2 |

**3.1.2 数据类型定义规则**

根据上文所述的数据库需求及特征提取评分标准，数据类型定义规则如下：

1、包括居民健康档案识别号、病历号在内的ID号的类型为bigint(20)，且大多不能为null；

2、包括身高、收缩压在内的测量数据的类型为decimal(18)或decimal(18, 2)；

3、包括性别、婚姻状况在内的类别数据的类型为char，长度根据具体特征而定；

4、包括运动方式、文化程度在内的文本数据的类型为varchar，长度根据具体特征而定；

5、包括出生时间、建档时间在内的时间数据的类型为date。

**3.1.3 建立高血压数据库**

在上述工作的基础上，本文分四步建立了专用的高血压数据库：搭建环境、统一字符集、数据传输和修改数据类型。

一、搭建环境

出于提高数据库存取速度、便于系统实现等方面的考虑，在内网的服务器上搭建LAMP环境。本文LAMP的组合方式为CentOS、Apache、MySQL、PHP。

在场内给服务器安装64位CentOS 6.0，配置内网接口，服务器信息如图3.1所示：



图3.1 服务器信息

配置IP地址等，实现内网上的IP访问。通过yum安装Apache、MySQL、PHP及支撑插件。

本地安装Navicat Premium，以SSH方式分别连接外网Oracle和内网MySQL。安装putty远程操作服务器，安装WinSCP传输文件。

二、统一字符集

原始数据含有中文，必须统一各平台的中文编码。原始数据使用utf-8，所以服务器上的MySQL编码、数据传输编码、本地MySQL编码向其靠拢，字符集设为“urf8 – UTF-8 Unicode”、排序规则设为“utf8\_general\_ci”。

修改内网服务器的my.cnf文件，并在MySQL的命令行界面下用如下命令统一字符集，直到所有项设置完毕：

alter database [数据库名] character set utf-8;

show variables like ‘collation\_%’;

show variables like ‘character\_set\_%’;

set [需修改的项] = utf8;

三、传输数据

按照数据库需求及特征提取标准，将数据迁移至位于内网服务器的MySQL数据库中。同种数据库间的数据通过Navicat传输，异种数据库间的数据通过Spoon.bat传输。

数据库进行插入操作时，会通过锁表保证数据安全性与一致性，导致运行缓慢。本文使用MyISAM引擎，行格式设置为Fixed，加快存取操作的速度。以常规体检表为例，采用InnoDB需要2~3天，但MyISAM引擎只需2小时。

四、修改数据类型

由于异种数据库上数据类型的定义不同，根据数据字典重新设置数据类型。

**3.2 Merge引擎分表**

考虑到包括常规体检表在内的多张表数据量过大，为了提高存取效率，在已添加表索引的基础上，通过分表实现数据库优化。

分表的常规方法是用程序把不同查询分配给不同子表，但该方法对查询不透明，不易使用。

MySQL提供了两个解决方案：Partition和MySQL Merge存储引擎[[[47]](#endnote-48)]。

其中，MySQL Merge表有如下五个优点：分离静态和动态的数据；利用数据结构的相似性优化查询；查询时可以访问更少的数据；适合大数据集；通过alter或.mrg文件可修改Merge表，并FLUSH TABLES刷新表缓存，动态增/减子表。

本文采用Merge存储引擎对数据量最大的体检记录表、常规体检表、辅助体检表进行了分表操作。

分表后，数据库存取效率有了大幅度提升。同一条查询语句的查询时间由原本的数十分钟缩短到数秒钟。以体检记录表为例，共有2190648条记录，分表后查询时间仅为8.400s，如图3.2所示：

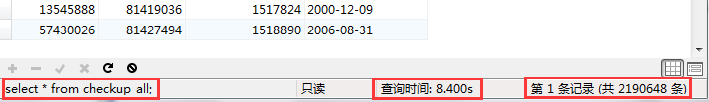


图3.2 Merge后的查询效率

分表的过程中需注意：子表间的结构需要一致，主表和子表的结构也要一致；Merge表保存了所有子表的索引，故不能在Merge表中根据索引进行唯一性检索；以MyISAM作为子表的引擎；REPLACE不会在Merge表中工作；AUTO\_INCREMENT不会按照预期方式工作。

以常规体检表为例，具体步骤如下：

1、 给常规体检表添加索引：

alter table checkupresult\_general add index idx\_mpi(MPI\_ID);

2、 建立相应子表，具体数量依据需求变化，常规体检表分为5张：

create table general\_1 like chekcupresult\_general;

3、 建立Merge表：

create table general\_all like checkupresult\_general;

4、 修改子表引擎为MyISAM：

alter table general\_1 ENGINE = MyISAM;

5、 给总表添加Merge引擎并关联子表：

alter talbe general\_all ENGINE = MERGE;

union( general\_1, general\_2, general\_3, general\_4, general\_5) INSERT\_METHOD = LAST;

6、 往各子表中写入数据，各50万条：

insert into general\_1 select \* from checkupresult\_general limit 0, 500000;

7、查询新表：

select \* from general\_all;

**第4章 数据处理**

**4.1 数据存在性与分布情况**

研究的第一步工作是进行数据探查，核实数据的可用性。不同居民拥有不同的看病历史，档案中的数据也有所不同，有些数据可能只有很少人有，有些数据可能都是无意义的值。以高血压人群、健康人群为主，糖尿病人群、癌症人群为辅，探查数据的存在性与分布情况，核实特征的可用性。

一、数据存在性

以目标表的目标特征为探查对象，以MySQL为探查工具，探查数据的存在性。探查的目的在于确定该项特征是否存在，标记并排除所有值都为空、所有值都为0的特征，并标记出文本类型的属性，根据研究需要单独做分词处理。同时记录下各表中有效的特征的记录数，需探查的数据表如表4.1所示：

表4.1 需探查的数据表

|  |  |
| --- | --- |
| 中文名 | 主要内容 |
| 常规体检 | 体质指数、腰围、收缩压、舒张压、抑郁评分等 |
| 市民信息 | 性别、年龄、文化程度、婚姻状况等 |
| 高血压随访 | 血压、生活方式、遵医行为等 |
| 糖尿病随访 | 血糖、生活方式、遵医行为等 |
| 饮酒史 | 饮酒量、频率、品种等 |
| 吸烟史 | 是否吸烟、吸烟量、是否戒烟等 |
| 市民重大健康问题 | 重要靶器官上的疾病等 |
| 疾病诊断 | 门诊、住院、症状等 |
| 住院病案信息 | 住院、症状等 |

以常规体检表为例，它的数据存在性探查结果如表4.2所示：

表4.2 常规体检表的数据存在性探查

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 待探查的字段 | 1次 | 2次 | 3次 | 4次 | 5次 | 5次或以上 |
| 身高 | 2792 | 9744 | 15962 | 20959 | 27762 | 143716 |
| 体温 | 0 | 0 | 0 | 0 | 0 | 0 |
| 体重 | 2792 | 9748 | 15952 | 20964 | 27771 | 143709 |
| 体质指数 | 66450 | 43309 | 7576 | 9275 | 635 | 3907 |
| 臀围 | 44584 | 21364 | 1679 | 3511 | 74 | 898 |
| 腰围 | 2888 | 10184 | 16681 | 21744 | 28188 | 141224 |

续表4.2 常规体检表的数据存在性探查

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 待探查的字段 | 1次 | 2次 | 3次 | 4次 | 5次 | 5次或以上 |
| 收缩压 | 2850 | 9766 | 16010 | 21124 | 27730 | 143453 |
| 舒张压 | 2853 | 9762 | 15994 | 21146 | 27731 | 143447 |
| 脉率 | 10276 | 22993 | 30537 | 32089 | 22947 | 94986 |
| 呼吸频率 | 65291 | 41407 | 6924 | 8536 | 455 | 3472 |

表4.2中的“次数”指数据库中，拥有多次该项记录的样本数量。其中，除了体温特征为“不存在”，其余特征都为“存在”，且有效的记录数为2792。其余情况依此例分析，探查结果体现在研究设计上，篇幅所限，不赘述。

二、数据分布情况

进一步探查“存在”的特征的数据分布情况。对于连续型数据，如年龄，按照相关定义划分年龄段分别探查；对于离散型数据，如性别，按照类别分别探查，部分结果如表4.3所示：

表4.3 分布情况探查（部分结果）

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 饮酒频率 | 1（从不） | 2（偶尔） | 3（经常） | 4（每天） | —— |
|  | 30401 | 6496 | 725 | 120 | —— |
| 开始饮酒年龄 | <18（未成年） | 18-28（青春期） | 29-40（成熟期） | 41-65（中年期） | >=65（老年期） |
|  | 46936 | 3626 | 624 | 47 | 6 |
| 日饮酒量 | 0-10 | 11-50 | 51-100 | >100 | —— |
|  | 49683 | 965 | 366 | 178 | —— |
| 饮酒种类 | 1(白酒) | 2（啤酒） | 3（红酒） | 4（黄酒） | >5（其他） |
|  | 2968 | 1300 | 0 | 968 | 329 |
| 戒酒标志 | 1（未戒酒） | 2（已戒酒） | —— | | |
|  | 10261 | 26666 | —— | | |
| 戒酒年龄 | <18  （未成年） | 18-28  （青春期） | 29-40  （成熟期） | 41-65  （中年期） | >=65  （老年期） |
|  | 16507 | 1317 | 209 | 11 | 0 |
| 醉酒标志 | 1 | 2 | —— | | |
|  | 1877 | 34363 | —— | —— | —— |
| 更新日期 | 2011年 | 2012年 | 2013年 | 2014年 | —— |
|  | 98 | 9434 | 33633 | 29926 | —— |

表4.3描述了饮酒史的数据分布探查情况。其中，开始饮酒年龄为数值型数据，按照健康学定义将其划分为“未成年”、“青春期”、“成熟期”、“中年期”、“老年期”；饮酒频率为类型数据，有“从不”、“偶尔”、“经常”、“每天”这4种，以字符方式存放在数据库中。其余情况依此例分析，篇幅所限，不赘述。



**4.2 研究对象**

通常，采集居民健康档案数据的情景如图4.1所示：

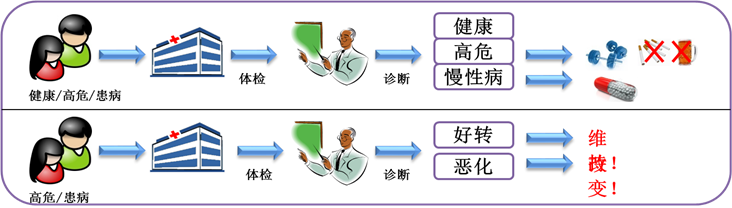


图4.1 数据采集情景

上半部分模拟高血压发病事件的数据采集情景，下半部分模拟高血压患严重并发症事件的数据采集情景。

一般，在确诊高血压或者严重并发症后，居民会在一定程度上改变生活方式，严重时甚至直接入院长期看护。因此，患病状态发生改变的患者在确诊后，与生活方式有关的数据参考价值不大，筛选特征时要特别注意。

探查结果显示，具有5次或以上常规体检记录，且其中至少一次以上收缩压或舒张压不在正常范围内，且至少有一次生活方式记录的人数为12334，其中具有辅助查体体检记录的人数为768。

其中，符合数据量要求的个人信息、常规体检信息、高血压随访表中的记录也较为完整。遗憾的是，同时拥有辅助体检记录的人群非常少，只可做小范围的研究。

具体如图4.2所示：

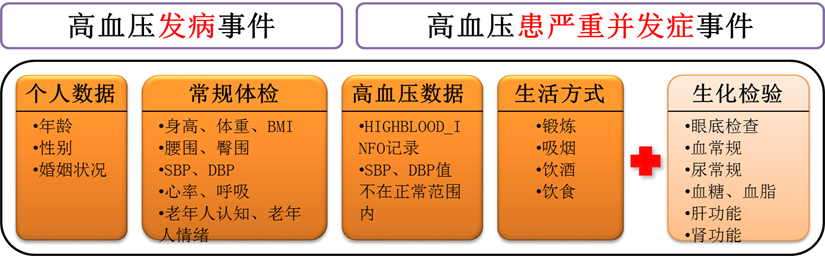


图4.2 研究对象示意图

1、发病事件

常规体检次数次的居民中，挑选符合记录较为完整的高血压患者作为实验组，挑选不曾患有任何严重慢病的健康居民作为对照组。以两组人都健康时的数据为研究对象，生成发病数据集。



2、并发症事件

在常规体检次数次的高血压患者中，挑选患有严重并发症的患者作为实验组，挑选病情维持良好的患者作为对照组。以两组人刚被确诊为高血压时的有关数据为研究对象，生成并发症数据集。



**4.3 数据集生成**

**4.3.1 分组规则**

首先对高血压数据库中的样本人群进行分组，分组情况如表4.4所示：

表4.4 实验分组情况

|  |  |  |
| --- | --- | --- |
| 类型 | 实验组 | 对照组 |
| 发病数据集 | 高血压患者 | 无重大慢病记录的患者 |
| 并发症数据集 | 有严重并发症的患者 | 无严重并发症的患者 |

一、发病事件

以健康时（特指未患高血压时）的有关数据为对象，研究实验组、对照组高血压发病风险，分组情况如图4.3所示：



图4.3 发病数据集的分组

实验组、对照组均有5次或以上常规体检记录。实验组要求5次常规体检至少一次测出临界值或高血压，且其中健康时的；对照组要求5次常规体检血压、血糖、血脂均在正常范围。

具体规则及样本数如表4.5所示：

表4.5 发病数据及分组规则及样本数

|  |  |  |
| --- | --- | --- |
| 组名 | 实验组（高血压患者） | 对照组（健康组） |
| 分组规则 | 1、常规体检次数≥5；  2、有患病前的记录，且SBP(90, 120)、DBP(60, 80)；  3、健康记录的时间比高血压随访记录的时间早6个月以上。 | 1、常规体检次数≥5；  2、所有记录的SBP(90, 120)、DBP(60, 80)；  3、不在高血压、糖尿病、肿瘤病人、重性精神病等重大疾病的登记表中；  4、现在主要健康问题表中所有疾病的字段的值为1（未患病）；  5、疾病诊断表中疾病名称要排除有心、脑、肾、糖尿、高血压等关键词的人群。 |
| 样本人数 | 33934名 | 58111名 |

二、并发症事件

以初次患有高血压时的有关数据为对象，研究实验组、对照组患严重并发症的风险，分组情况如图4.4所示：

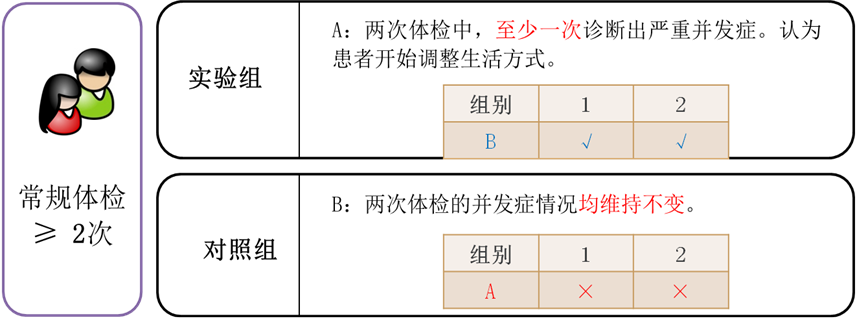


图4.4 并发症数据集分组

实验组、对照组在确诊高血压后，均有2次或以上常规体检记录。对于实验组，要求门诊、住院的诊断记录至少有一次诊断出严重并发症；对于对照组，要求所有记录都无严重并发症记录。其中，诊断记录的提取见后文中文本数据的分词部分。

具体规则及样本人数如表4.6所示：

表4.6 发病数据及分组规则及样本人数

|  |  |  |
| --- | --- | --- |
| 组名 | 实验组（患有严重并发症） | 对照组（预后稳定） |
| 分组规则 | 1、常规体检次数；  2、确诊高血压后，要有2次常规体检记录；  3、第一次常规体检记录与最后一次常规体检记录，二者相差至少2年；  4、该患者有高血压随访记录；  5、确诊高血压后，门诊、住院的诊断记录中有患严重并发症的记录。 | 1、常规体检次数；  2、确诊高血压后，要有2次常规体检记录；  3、第一次常规体检记录与最后一次常规体检记录，二者相差至少2年；  4、该患者有高血压随访记录；  5、确诊高血压后，门诊、住院的诊断记录没有患严重并发症的记录。 |
| 样本人数 | 647名 | 36404名 |

实验组样本人数明显减少主要是因为严重并发症的定义较为严格，以死亡率为评估标准，因此符合条件的患者较少。

**4.3.2 数据集生成规则**

一、发病事件

1、常规体检表：提取半年至一年半内，最早的健康记录。

身高、体重、体质指数、臀围、腰围、收缩压、舒张压、脉率、呼吸频率、老年人抑郁评分

2、健康体检信息表：按档案标识符、体检流水号唯一对应。

检测日期（该时间将作为基准时间，用于筛选生活方式等表中的数据）

3、市民个人信息：按档案标识符提取最接近基准时间的数据。

性别、文化程度、出生日期、婚姻状况、职业

4、高血压患者随访记录表：提取随访日期与基准时间差3个月的最早数据。

随访类型、随访日期、症状、收缩压、体重、心率、日吸烟量、日饮酒量、周运动次数、运动时间、摄盐情况、心理情况调整代码、遵医行为、辅助检查结果描述。

5、饮酒史：提取更新时间与基准时间差3个月的最早的数据。

饮酒频率、开始饮酒年龄、日饮酒量、饮酒种类、戒酒标志、戒酒年龄、更新时间。

6、吸烟史：提取更新时间与基准时间差3个月的最早的数据。

开始吸烟年龄、日吸烟量、戒烟标志、戒烟年龄、更新时间

7、运动史：提取更新时间与基准时间差3个月的最早的数据。

运动频率、运动次数、运动时间、运动方式、更新时间。

二、并发症事件

1、常规体检表：提取确诊高血压后的最早的数据。

身高、体重、体质指数、臀围、腰围、收缩压、舒张压、脉率、呼吸频率、老年人抑郁评分。

2、健康体检信息：按档案标识符、体检流水号唯一对应。

体检日期（该时间将作为基准时间，用于筛选生活方式等表中的数据）

3、市民个人信息：按档案标识符唯一对应。

性别、文化程度、出生日期、婚姻状况、职业。

4、高血压患者随访记录：提取随访日期与基准时间差3个月且时间最接近的数据

随访类型、随访日期、症状、收缩压、体重、心率、日吸烟量、日饮酒量、周运动次数、运动时间、摄盐情况、心理情况调整代码、遵医行为、辅助检查结果描述。

5、饮酒史：提取更新时间与基准时间差3个月且时间最接近的数据。

饮酒频率、开始饮酒年龄、日饮酒量、饮酒种类、戒酒标志、戒酒年龄、更新时间。

6、吸烟史：提取更新时间与基准时间差3个月且时间最接近的数据。

开始吸烟年龄、日吸烟量、戒烟标志、戒烟年龄、更新时间

7、运动史：提取更新时间与基准时间差3个月的最早的数据。

运动频率、运动次数、运动时间、运动方式、更新时间。

8、现在健康主要问题：用记录序号提取更新时间与基准时间相差2年以内的数据。

脑血管疾病情况代码、肾脏疾病情况代码、心脏疾病情况代码、血管疾病情况代码、眼部疾病情况代码、神经系统疾病情况代码、其他系统疾病情况代码、更新日期。

9、疾病诊断表：提取诊断日期高血压确诊日期相差2年以内的数据。

高血压确诊日期、病症1、时间1、病症2、时间2、……、病症24、时间24。该部分涉及门诊中的文本数据，具体情况见后文中门诊数据的分词结果。

**4.4 门诊数据的分词**

门诊数据大多以文本方式存储，生成数据集前需要对其进行分词：首先统计出现次数最多的病症，并以病症的死亡率、危险性为参考，得出主要的病症集合；再以病症集合为标准，返回去搜索原始数据，求得每个样本的患病情况，以“病症名 + 患病与否”的方式添加到数据集中。

本文的分词对象是疾病诊断表的病症名称，该特征是医院为病人诊断时录入的症状描述，数据内容如图4.5所示：



图4.5 病症特征的数据内容

观察数据发现：同一种诊断会有多种别名，比如高血压、高血压病、高压等；同一份记录记录了多种病。

ICTCLAS2015作为分词工具，可以进行常用语的简单分词，但对专业词汇必须自己添加用户词典。基于医学知识，建立用户词典。通过逐条读取并分析病症数据，得到病症名及对应的发病次数，主要结果如表4.7所示：

表4.7 分词结果

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 病症 | 频次 | 病症 | 频次 | 病症 | 频次 |
| 高血压 | 36626 | 糖尿病 | 11346 | 冠心病 | 6224 |
| 高脂血 | 3736 | 急性上呼吸道感染 | 3073 | 咽炎 | 2925 |
| 眩晕 | 1599 | 上呼吸道感染 | 1567 | 慢性胃炎 | 1219 |
| 2型糖尿病 | 1215 | 支气管炎 | 1145 | 白内障 | 1093 |
| 头晕 | 1089 | 结膜炎 | 1086 | 急性支气管炎 | 1055 |
| 脑梗塞 | 993 | 失眠 | 959 | 原发性高血压 | 755 |
| 感冒 | 762 | 贫血 | 626 | 头痛 | 620 |
| 咳嗽 | 614 | 脑动脉硬化 | 577 | 尿毒症 | 497 |
| 痛风 | 401 | 干眼症 | 381 | 心律失常 | 361 |
| 慢性肾衰竭 | 360 | 扁桃体炎 | 341 | 睡眠障碍 | 325 |

选择排名靠前的并发症病症作为新特征，共挑选出24种：糖尿病、冠心病、脑梗塞、脑动脉硬化、心律失常、动脉硬化、中风、慢性肾功能衰竭、心肌缺血、心绞痛、脑卒中、脑出血、心力衰竭、心肌梗死、腔梗、心肌梗塞、肾功能衰竭、脑血栓、短暂性脑缺血发作、慢性心力衰竭、视网膜动脉硬化、陈旧性心肌梗塞、脑栓塞、高血压脑出血。

以上述新特征为基础，返回数据库，以并发症实验组为目标对象，以确诊高血压后新患上的病症作为研究对象，结果如表4.8所示：

表4.8 分词结果示例

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| MPI | HYPER\_DATE | DISEASE1 | DATE1 | DISEASE24 | DATE24 |
| \*\*\*\* | 2013-03-01 | 0 | null | 1 | 2014-09-01 |
| \*\*\*\* | 2013-03-01 | 1 | 2013-05-03 | 1 | 2014-09-01 |

MPI为居民标识符，HYPER\_DATE为高血压确诊时间，DISEASE1 ~ DISEASE24分别对应上述24种常见病症，DATE1 ~ DATE24为确诊时间。

**4.5 数据预处理**

现实世界中的数据通常是不完整并含有许多噪声的[37]。数据预处理可以提高数据质量、提升处理效率，这一步骤及其前期准备步骤通常占了整个数据挖掘过程70%以上的时间。

居民健康档案中的脏数据通常以如下方式呈现：缺失值、错误值、不同的缩写、自由格式的文本串、值与字段不匹配、字段间不对应、词移位、两条记录对应同一实体、互相矛盾的记录、同一实体的某个属性有多个不同的值、错误的引用等[[[48]](#endnote-49)]。

**4.5.1 数据清洗**

一、缺失数据处理

通常，缺失数据的呈现方式有两种——空值和零值。然而，并非所有零值都是无意义的。因此在填补缺失数据前，需要人工筛出“零值无意义”的特征，将零值转换为空值，后期统一填补。

1、零值转空值

特征：心率、身高、体重、BMI、臀围、腰围、收缩压、舒张压、呼吸、开始吸烟年龄、戒烟年龄、开始饮酒年龄、戒酒年龄、文化程度。

标准SQL语句为：

update [表] set [特征] = null where [特征] = [零值];

2、填补空值

1）需人工填写的特征：教育程度、婚姻状况、锻炼方式。

2）填补全局常量的特征：职业、戒酒标志、戒酒年龄、症状。

3）填补均值的特征：身高、体重、腰围、收缩压、舒张压、呼吸、老年人抑郁评分、日饮酒量、日吸烟量、运动频率、周运动次数、坚持锻炼时间、食盐摄入、精神状态调整。

4）填补出现频率最高的值的特征有：性别、教育程度、婚姻状况、饮酒频率、醉酒标志、锻炼方式、遵医行为。

标准SQL语句为：

update [表] set [特征] = [值] where [特征] is null;

二、修改错误值

原始数据基本由人工录入，失误难以避免，错误值的呈现方式通常为偏离合理取值范围。

1、数值型

该类错误值通常扩大/缩小10的倍数（十进制），可参考正常取值范围修正。

特征：身高、体重、腰围、收缩压、舒张压、呼吸、老年人抑郁评分、日饮酒量、戒酒年龄、日吸烟量、运动频率、周运动次数、饮食状况、食盐摄入。

标准SQL语句为：

update [表] set [特征] = [特征] / 10 where [特征] > [值界限];

2、离散型

该类错误值千奇百怪，只能通过其他方式清理数据。

特征：性别、教育程度、婚姻状况、职业、饮酒频率、戒酒标志、醉酒标志、坚持锻炼时间、锻炼方式、症状、精神状态调整、遵医行为。

标准SQL语句为：

update [表] set [特征] = null where [特征] != [值];

update [表] set [特征] = [值] where [特征] is null;

三、值的补集

医疗机构采集居民健康数据时，统一特征有多种表述形式。如，婚姻状况特别以字符、文本加以描述，字符型用“1”、“2”等值，文本用 “未婚”、“已婚”等值。

另一种情况是，同一特征的数据出现在不同表中。如，舒张压的数据可能出现在常规体检表或高血压随访表中。

高血压随访记录中的体重、收缩压、日饮酒量、日吸烟量、运动时间(分钟/次)可以分别补充到常规体检中的体重、收缩压，饮酒史的日饮酒量，吸烟史的日吸烟量，生活方式的运动时间中。

标准SQL语句为：

update [表] set [特征1] = [特征2] where [特征1] is null and [特征2] is not null;

**4.5.2 数据转换**

一、时间数据

居民个人信息只记录居民出生日期，不记录年龄，不能直接使用，需把出生日期转换为年龄，标准SQL语句为：

update [表名] set AGE = round(DATEDIFF(SYSDATE(), DATE\_OF\_BIRTH) / 365);

高血压随访表只记录随访时间，需取第一次随访时间作为高血压确诊时间，并转换为患病年龄，标准SQL语句为：

Update[表名] set HYPER\_AGE = round(DATEDIFF(HYPER\_DATE, DATE

\_OF\_BIRTH) / 365);

二、可推导的数据

部分特征可由其他特征计算得来，且在线性模型中具有更好的效果。例如，常规体检表中的BMI，可由该表中的求得，标准SQL语句为：



update [表] set BMI = round((WEIGHT / (HEIGHT \* HEIGHT)) \* 10000, 2) where WEIGHT is not null and HEIGHT is not null;

三、值的替换

部分特征虽然为分类特征，但却以文本方式存储，如性别，值为男、女，但实际上可替换为字符1、2，以减少内存消耗并提高处理速度，特征值替换结果如表4.9所示：

表4.9 特征值替换结果

|  |  |
| --- | --- |
| 特征名称 | 值的说明 |
| 性别 | 1男 2女 |
| 文化程度 | 0其他 1初中及以下 2中专及中技 3高中  4技校 5大专 6大学本科 7研究生 |
| 婚姻状况 | 1未婚 2已婚 3丧偶 4离婚 5其他 |
| 饮酒频率 | 1从不 2偶尔 3经常 4每天 |
| 主要饮酒品种 | 1白酒 2啤酒 3红酒 4黄酒 5其他 |
| 是否戒酒 | 1未戒酒 2已戒酒 |
| 吸烟状况 | 1从不吸烟 2已戒烟 3吸烟 |
| 摄盐情况 | 1轻 2中 3重 |
| 心理调整情况 | 1良好 2一般 3差 |
| 遵医行为 | 1良好 2一般 3差 |
| 随访类型 | 1控制满意 2控制不满意 3不良反应 4并发症 |

四、类型标记

类型标记对照如表4.10所示：

表4.10 类型标记对照

|  |  |
| --- | --- |
| 类型标记 | 类型名称 |
| 0 | 健康 |
| 1 | 健康 -> 高血压 |
| 3 | 高血压 -> 预后稳定 |
| 4 | 高血压 -> 严重并发症 |

类标为“0”、“1”的，根据实验的分组结果进行添加；类标为“3”、“4”的，根据3.4章门诊数据中高血压并发症的分词结果，按照医学上病症的死亡率、危险性进行排序，然后进行添加类标。

添加“0”、“1”类标的标准SQL语句为：

update [发病表] set CLASSIFICATION = 0;

update [发病表], [发病实验组表] set CLASSIFICATION = 1 where MPI\_ID in (select MPI\_ID from [发病实验组表]);

添加“3”、“4”类标的标准SQL语句为：

update [并发症表] set CLASSIFICATION = 3;

update [并发症表], [并发症分词表] set CLASSIFICATION = 4 where [严重并发症1] = 1 or [严重并发症2];

根据分词结果，在所有24种常见病症有，有12中属于高血压严重并发症范畴，详见3.4章对严重并发症的定义，在此不赘述。

**4.5.3 数据规约**

有些特征的值范围很大，但值又主要集中在较小的区域，如果直接进行挖掘，模型对这种值不敏感。为了消除量纲的影响，常用的两种做法是数据中心化、数据标准化。

数据中心化，即各项数据减去特征均值。例如，年龄数据18、20、22、24、26，其均值为22，那么中心化后的数据集：18-22，20-22，22-22，24-22，26-22，即：-4，-2，0，2，4。

数据的标准化，即数据中心化后的值再除以标准差。例如，上述年龄数据集的标准差为（2根号2），那么标准化后的数据集为：，，，，，即：，，0，，。



(4.1)



(4.2)



本文选择对数据进行标准化处理，具体过程于建模时在R平台上完成。

**第5章 风险建模及评估**

虽然发病、并发症这两类事件的数据、处理方式、结果意义有所不同，但风险建模的过程类似，故本文合并到一处讲述。

首先，分组介绍发病数据集、并发症数据集的样本及特征情况；然后，采用逻辑斯蒂回归、朴素贝叶斯、随机森林这三种分类器分别对两类事件进行风险建模；再然后，采用30次10折交叉验证评估模型，绘制ROC并计算AUC，分别比较三种分类器在两类事件共六份数据集上的建模结果；最后，以逻辑斯蒂回归得出的风险模型为重点研究对象，详述主要影响因素及权重。

整个建模、评估过程在R平台上完成，R是用于统计计算和图形的系统，除了常规功能外，还提供编程语言、高级图形、与其他语言的接口、调式设备等[[[49]](#endnote-50)]。建模前，从MySQL中导出.csv格式的数据集，并统一R平台和数据集的中文编码为utf - 8，用WinSCP把数据集存至R所在的服务器。

**5.1 数据集介绍**

为了更好地建模，本文针对发病、并发症事件分别生成了3份数据集。生成原则分别为：最大限度保留样本、一定程度保留样本和特征、最大限度保留特征，具体如表5.1所示：

表5.1 数据集详情

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 事件 | 生成原则 | 实验组样本量 | 对照组样本量 | 特征数 |
| 发病 | 最大限度保留样本 | 30209 | 57120 | 9 |
| 一定程度保留样本和特征 | 21707 | 17869 | 9 |
| 最大限度保留特征 | 2203 | 842 | 15 |
| 患严重并发症 | 最大限度保留样本 | 647 | 36404 | 17 |
| 一定程度保留样本和特征 | 637 | 23760 | 16 |
| 最大限度保留特征 | 351 | 7485 | 21 |

对应的数据集名称从上往下分别为：datarisk、hyperrisk.sample、hyperrisk.attribute、databfz、hyperbfz.sample、databfz.attribute。

根据上表，我们观察到：最大限度保留特征时，样本数量减少一个数量级，而特征的数量仅增加了6个。这主要是因为与生活方式有关的特征，如，锻炼、饮酒、吸烟等，并非医疗机构数据采集的重点，所以该类特征本就具有有效样本量少的特点。其次，由于本文采用的是高血压患者尚未患病时的相关数据，有严格的时间限制，而居民健康档案仅保留最近更新的一份与生活方式有关的特征。在这两个因素的双重作用下，增加特征数将势必大幅度地减少样本量。

不难发现，类似情况也出现在高血压并发症数据集上。

**5.2 风险建模**

本文采用逻辑斯蒂回归、朴素贝叶斯、随机森林这三种分类方法分别对高血压的发病、患严重并发症事件的共六份数据集进行风险建模。

首先导入所需的包，如图5.1所示：

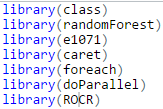


图5.1 建模及评估所需的R包

class中有分类常用的函数，如batchSOM、condense、knn、knn.cv、lvq等；randomForest中有随机森林常用的函数，如combine、getTree、grow、na.roughfix、outlier、plot.randomForest、predict.randomForest、randomForest、treesize等；e1071提供统计相关的函数支持，如朴素贝叶斯、模糊集、SVM等；caret，是分类和回归训练（Classification And REgression Traning）的缩写，有建立分类和回归模型常用的函数，如calibration、confusionMatrix、cox2、histogram.train、lattice.rfe等；foreach支持foreach语法；doParallel提供并行处理相关的函数支持；ROCR提供绘制ROC曲线、敏感度特异性曲线等的函数支持。

由于六份数据集上的风险建模过程类似，以发病事件的“最大限度保留样本”数据集，即datarisk为例详述建模过程。

从服务器上把数据集导入内存，并填充其中残余的空值。把类标从数据集中分出来，标记类别型数据，包括性别、文化程度、婚姻状况等，对数据进行标准化处理，主要代码如下：

datarisk <- read.table("~/workspace/data\_qin/datarisk.txt", header = T)

datarisk <- na.roughfix(datarisk)

y <- datarisk$CLASSIFICATION

datarisk$CLASSIFICATION <- NULL

x <- datarisk

factor.features <- c("SEX\_CODE", "EDU\_NAME", "MARITAL\_STATUS")

for (feature in factor.features){

x[,feature] <- as.factor(x[,feature])

}

other.features <- setdiff(colnames(x), factor.features)

x[,other.features] <- scale(x[, other.features])

逻辑斯蒂回归建模的关键代码如下：

model.log <- glm(label ~ ., data = data, family = "binomial")

朴素贝叶斯建模的关键代码如下：

model.nb <- naiveBayes(label ~ ., data = data)

随机森林建模的关键代码如下：

model.rf <- randomForest(label ~ ., data = data)

重复上述过程，分别处理发病事件的三份数据集，得出模型。类似地，按上述流程处理并发症事件的三份数据集。发病事件得出九个模型，并发症事件也得出九个模型，具体建模过程在此不赘述。

**5.3 模型评估**

采用30次10折交叉验证评估模型，计算AUX，最后绘制ROC曲线，评估建模结果。

首先，按照8:2的比例从数据集中抽选训练集和测试集，即80%做训练集，20%做测试集。随后，再将训练集平均分成10份，把其中9份做训练集，留出1份做测试集。然后，再将训练集平均分成10份，测试集平均分成10份，分别使用逻辑回归、朴素贝叶斯、随机森林对其进行训练及测试。最后，求出AUC的均值作为1次10折交叉验证的结果，并保存起来。

评估过程如图5.2所示：



图5.2 评估过程示意图

重复上述过程30次，完成30次10折交叉验证，再次求AUC均值作为对模型准确性的最终评估结果。为节省时间，本文用30个进程并行处理。

首先，定义10折交叉验证的函数，关键代码如下：

CV <- function(y, fold = 10){

flds <- createFolds(y, k = fold, list = TRUE, returnTrain = FALSE)

cv.test <- list()

for(i in 1:fold){

cv.test[[i]] <- flds[[i]]

}

cv.train <- list()

for(i in 1:fold){

cv.train[[i]] <- setdiff(1:length(y), cv.test[[i]])

}

cv <- list()

cv$test <- cv.test

cv$train <- cv.train

return(cv)

}

然后，定义函数，分别采用三种分类器处理每份数据，每轮结束后计算auc，关键代码如下：

KFoldClassifiers <- function(X, Y, fold){

require(e1071)

require(randomForest)

pred.log <- numeric(length(Y))

pred.nb <- numeric(length(Y))

pred.rf <- numeric(length(Y))

cv <- CV(Y, fold = fold)

for(i in 1:fold){

x <- X[cv$train[[i]],]

y <- Y[cv$train[[i]]]

train.data <- as.data.frame(cbind(x,y))

names(train.data) <- paste0("V", 1:ncol(train.data))

names(train.data)[ncol(train.data)] <- "label"

train.data$label <- as.factor(train.data$label)

test.data <- as.data.frame(X[cv$test[[i]],])

names(test.data) <- paste0("V", 1:ncol(test.data))

model <- glm(label ~ ., data = train.data, family = binomial(logit))

pred <- predict(model, test.data, type = "response")

pred.log[cv$test[[i]]] <- pred

model <- naiveBayes(label ~ ., data = train.data)

pred <- predict(model, test.data, type = "raw")

pred.nb[cv$test[[i]]] <- pred[,2]

model <- randomForest(label ~ ., data = train.data)

pred <- predict(model, test.data, type = "prob")

pred.rf[cv$test[[i]]] <- pred[,2]

}

ROCRpred = prediction(pred.log, Y)

auc.log <- as.numeric(performance(ROCRpred, "auc")@y.values)

ROCRpred = prediction(pred.nb, Y)

auc.nb <- as.numeric(performance(ROCRpred, "auc")@y.values)

ROCRpred = prediction(pred.rf, Y)

auc.rf <- as.numeric(performance(ROCRpred, "auc")@y.values)

return(c(auc.log, auc.nb, auc.rf))

}

定义按一定比例划分训练集、测试集的函数，关键代码如下：

gen.train.test <- function(X, Y, ratio){

num <- 1:length(Y)

train.index <- sample(num, length(Y)\*ratio)

test.index <- setdiff(num, train.index)

x <- list()

y <- list()

x[["train"]] <- X[train.index,]

x[["test"]] <- X[test.index,]

y[["train"]] <- Y[train.index]

y[["test"]] <- Y[test.index]

data <- list(x, y)

return(data)

}

定义绘制ROC曲线的函数，关键代码如下：

get.roc <- function(train.x, train.y, test.x, test.y){

train.data <- cbind(train.y, train.x)

colnames(train.data)[1] <- "label"

train.data$label <- as.factor(train.data$label)

model <- glm(label ~ ., data = train.data, family = binomial(logit))

pred.log <- predict(model, test.x, type = "response")

model <- naiveBayes(label ~ ., data = train.data)

pred.nb <- predict(model, test.x, type = "raw")

model <- randomForest(label ~ ., data = train.data)

pred.rf <- predict(model, test.x, type = "prob")

ROC.log = prediction(pred.log, test.y)

ROC.nb = prediction(pred.nb[,2], test.y)

ROC.rf = prediction(pred.rf[,2], test.y)

return(list(log = ROC.log, nb = ROC.nb, rf = ROC.rf))

}

调用30个进程并行处理30次10折交叉验证，保存结果，关键代码如下：

cl <- makeCluster(30, type = "SOCK")

registerDoParallel(cl)

k.fold.rep <- foreach(i = 1:30, .combine = "rbind", .export = c("CV"), .packages = c("class","e1071", "randomForest", "caret", "ROCR")) %dopar%{

KFoldClassifiers(x, y, 10)

}

stopCluster(cl)

result[["datarisk"]] <- k.fold.rep

最后，绘制ROC曲线，关键代码如下：

roc[["datarisk"]] <- get.roc(x, y, train.test[[1]][["test"]], train.test[[2]][["test"]])

以上过程为高血压发病事件的中datarisk数据集（最大限度保留样本）的处理过程，类似地，处理发病事件、严重并发症事件的所有数据集，详细过程在此不赘述。

**5.4 模型比较**

各数据集上，分别采用逻辑斯蒂回归、朴素贝叶斯、随机森林进行建模的效果用30次10折交叉验证评估如下：

一、发病事件

1）datarisk数据集上的结果如表5.2所示：

表5.2 datarisk的评估结果

|  |  |  |  |
| --- | --- | --- | --- |
| 序号 | Logistic | NaiveBayes | RandomForest |
| 平均值 | 0.950900547722674 | 0.945441670770052 | 0.963448550751506 |
| 标准差 | 0.0000143082288063568 | 0.00000897467251479031 | 0.000143078057972043 |

更直观的ROC曲线如图5.3所示：

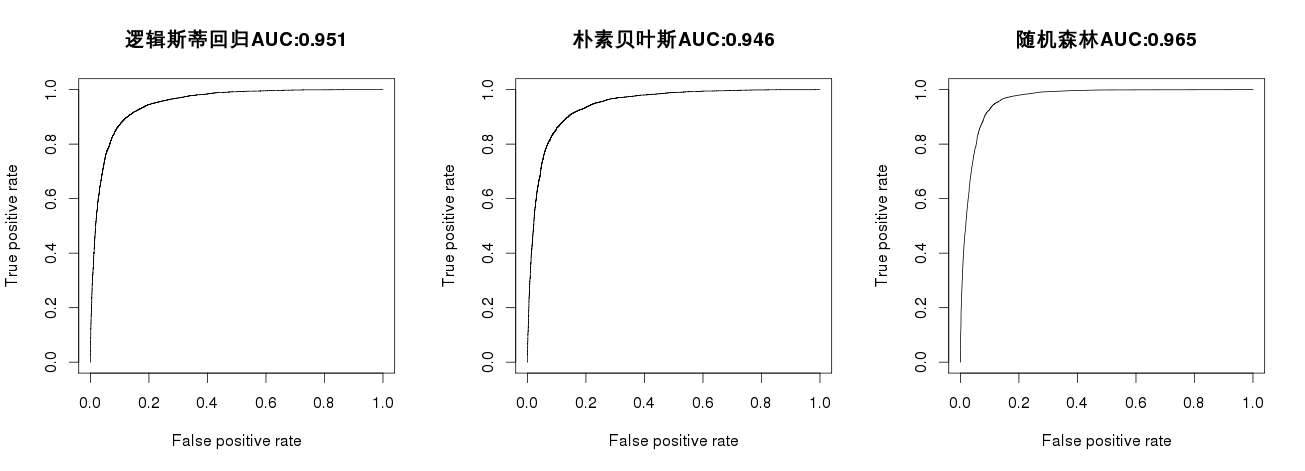


图5.3 datarisk的ROC曲线

我们发现，在datarisk数据集上，随机森林的建模效果最好，AUC = 0.9634，朴素贝叶斯的建模效果最稳定，标准差 = 0.0000089。

2）hyperrisk.sample数据集上的评估结果如表5.3所示：

表5.3 hyperrisk.sample的评估结果

|  |  |  |  |
| --- | --- | --- | --- |
| 序号 | Logistic | NaiveBayes | RandomForest |
| 平均值 | 0.930929849493694 | 0.925238553774099 | 0.941149551886023 |
| 标准差 | 0.0000289502081528086 | 0.0000160111299511268 | 0.000273015951058807 |

更直观的ROC曲线如图5.4所示：

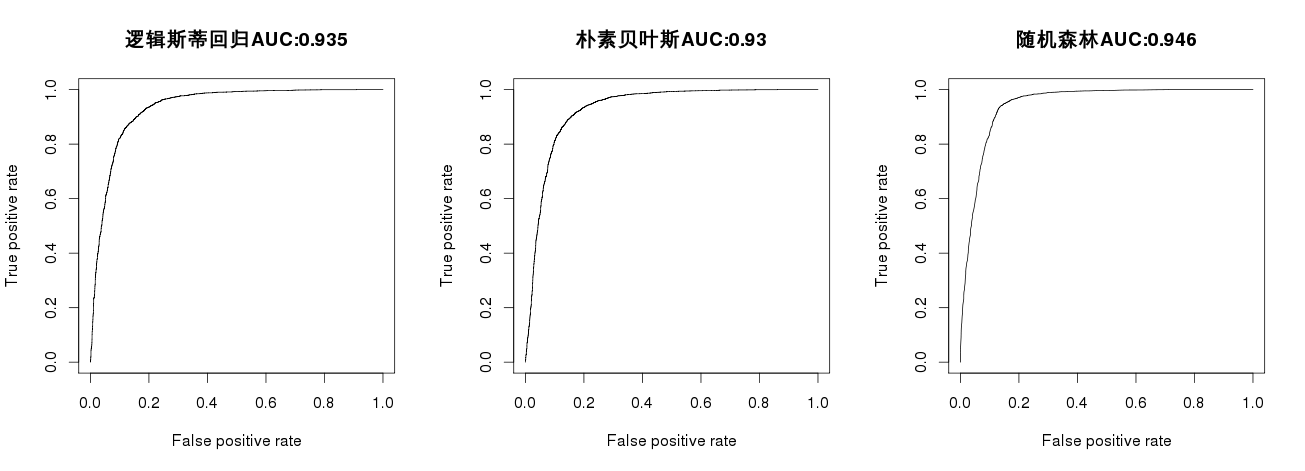


图5.4 hyperrisk.sample的ROC曲线

我们发现，在hyperrisk.sample数据集上，随机森林的建模效果最好，AUC = 0.9411，朴素贝叶斯的建模效果最稳定，标准差 = 0.000016。

3）hyperrisk.attribute数据集上的评估结果如表5.4所示：

表5.4 hyperrisk.attribute的评估结果

|  |  |  |  |
| --- | --- | --- | --- |
| 序号 | Logistic | NaiveBayes | RandomForest |
| 平均值 | 0.796110437960621 | 0.791060528660028 | 0.823175520589257 |
| 标准差 | 0.00100993233088368 | 0.00111704209677349 | 0.00281394069698871 |

更直观的ROC曲线如图5.5所示：

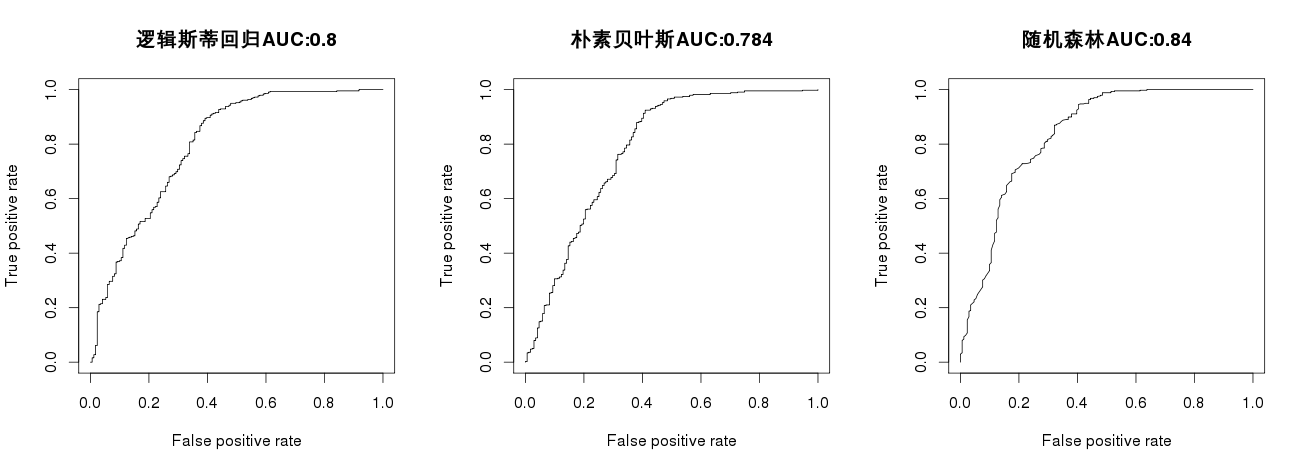


图5.5 hyperrisk.attribute的ROC曲线

我们发现，在hyperrisk.sample数据集上，随机森林的建模效果最好，AUC = 0.8231，逻辑斯蒂的建模效果最稳定，标准差 = 0.001。

综上所述，在高血压发病事件的三份数据集中，最大限度保留样本的数据集建模效果最好；三种分类器中，随机森林的建模效果最好，AUC = 0.9634；朴素贝叶斯的模型最稳定，标准差相对较小。

二、并发症事件

1）databfz数据集上的评估结果如表5.5所示：

表5.5 databfz的评估结果

|  |  |  |  |
| --- | --- | --- | --- |
| 序号 | Logistic | NaiveBayes | RandomForest |
| 平均值 | 0.811219655277616 | 0.788735963076074 | 0.772905447596963 |
| 标准差 | 0.00125835321289736 | 0.0024988561336113 | 0.00391688048773276 |

更直观的ROC曲线如图5.6所示：

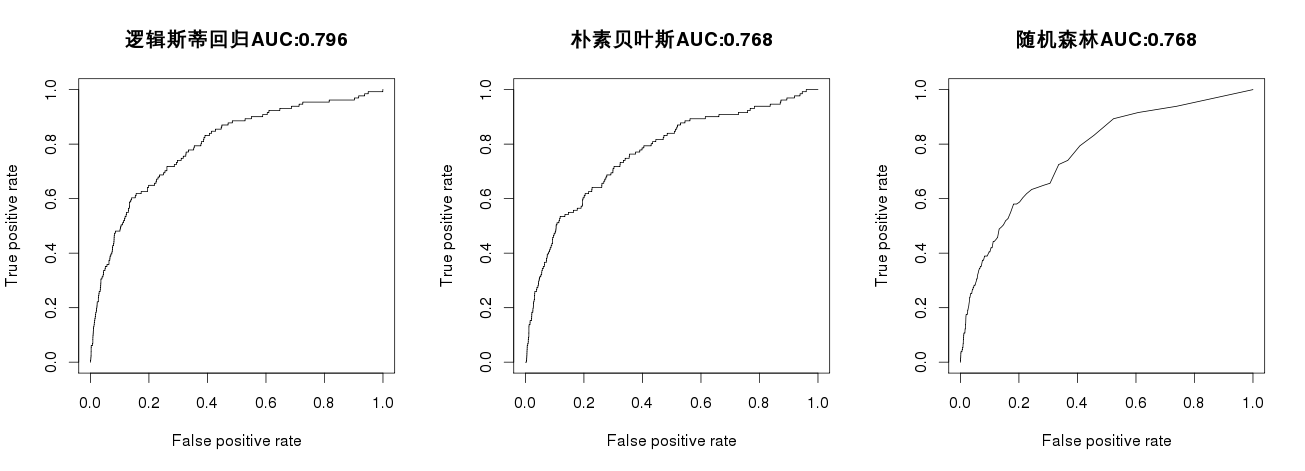


图5.6 databfz上的ROC曲线

我们发现，在databfz数据集上，逻辑斯蒂回归的建模效果最好，AUC = 0.8112，同时它的建模效果也最稳定，标准差 = 0.001258。

2）hyperbfz.sample数据集上的评估结果如表5.6所示：

表5.6 hyperbfz.sample的评估结果

|  |  |  |  |
| --- | --- | --- | --- |
| 序号 | Logistic | NaiveBayes | RandomForest |
| 平均值 | 0.830613852766591 | 0.791987511432050 | 0.791878194994586 |
| 标准差 | 0.00080427755820522 | 0.000811199194839366 | 0.00258287052943688 |

更直观的ROC曲线如图5.7所示：

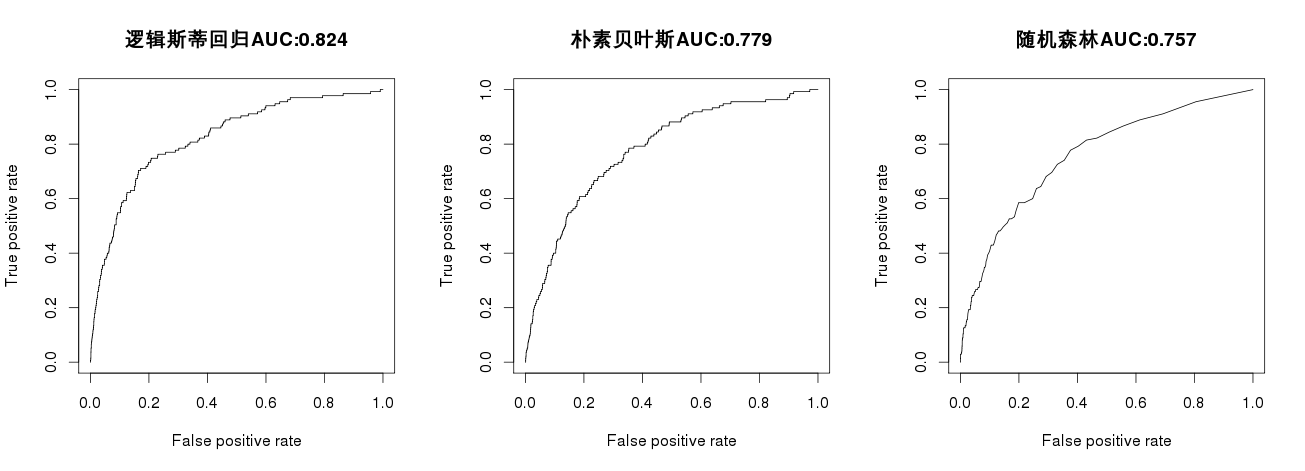


图5.7 hyperbfz.sample上的ROC曲线

我们发现，在hyperbfz.sample数据集上，逻辑斯蒂回归的建模效果最好，AUC = 0.8306，同时它的建模效果也最稳定，标准差 = 0.0008。

3）hyperbfz.attribute数据集上的评估结果如表5.7所示：

表5.7 hyperbfz.attribute上30次10折交叉验证结果

|  |  |  |  |
| --- | --- | --- | --- |
| 序号 | Logistic | NaiveBayes | RandomForest |
| 平均值 | 0.777703288366386 | 0.754563553691525 | 0.772673707053134 |
| 标准差 | 0.00345129575737908 | 0.00257990248868838 | 0.0035245915162859 |

更直观的ROC曲线如图5.8所示：

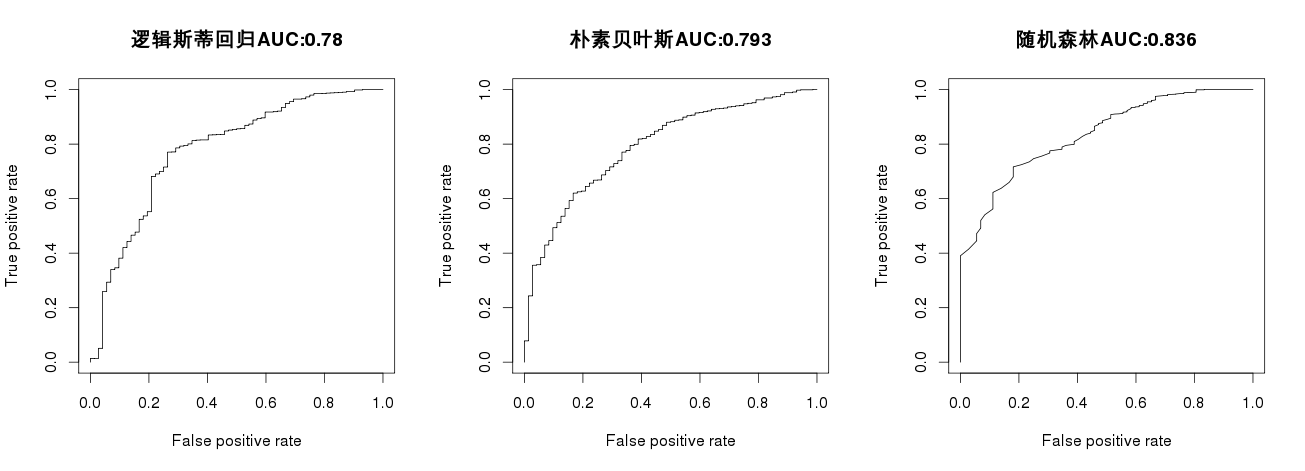


图5.8 hyperbfz.attribute上的ROC曲线

我们发现，在hyperrisk.sample数据集上，逻辑斯蒂回归的建模效果最好，AUC = 0.7777，朴素贝叶斯的建模效果最稳定，标准差 = 0.0025。

综上所述，在三份数据集中，一定程度保留样本和特征的那份数据集建模效果最好；三种分类器中，逻辑斯蒂回归的建模效果最好，AUC = 0.8306；朴素贝叶斯的模型最稳定，标准差相对较小。

**5.5 结果与分析**

在三种分类算法中，只有逻辑斯蒂回归得出的模型能用于分析主要影响因素。参考上文模型比较的结果，本文选定datarisk、hyperbfz.sample数据集上逻辑斯蒂回归算法的建模结果做具体分析，研究发病、患严重并发症事件的主要影响因素、风险公式、单变量分布。

**5.5.1 高血压发病事件**

发病事件以高血压人群为实验组，健康人群为对照组，研究他们患高血压之前的有关数据。

一、主要影响因素

三份数据集上，逻辑斯蒂回归风险模型的AUC = 0.95/0.93/0.80，效果最好的模型基于datarisk数据集。只保留显著性不大于0.05且权重大于0.01的作为主要影响因素，在三个发病风险模型中出现过的有：身高、体重、BMI、腰围、收缩压、舒张压、性别、文化程度、婚姻状况、年龄。

以AUC=0.95的模型为研究对象，详情如表5.8所示：

表5.8 高血压发病风险模型

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 参数 | 因素 | 权重 | 相关性 | P值 | 平均值 | 最小值 | 最大值 | 单位 |
| x0 | 调整项 | -4.80913 | 负相关 | 5.58E-06 | —— | | | —— |
| x1 | 身高 | -0.23797 | 负相关 | 8.24E-09 | 163.68 | 100 | 200 | cm |
| x2 | 体重 | 0.48528 | 正相关 | < 2E-16 | 58.62 | 30 | 173 | kg |
| x3 | 收缩压 | 0.72107 | 正相关 | < 2E-16 | 112.81 | 21 | 200 | mmHg |
| x4 | 舒张压 | 0.79118 | 正相关 | < 2E-16 | 70.80 | 50 | 150 | mmHg |
| x5 | 性别1 | 3.41190 | 正相关 | 1.26E-03 | 1男  2女 | | | 0/1 |
| x6 | 性别2 | 3.41855 | 正相关 | 1.23E-03 | 0/1 |
| x7 | 文化程度1 | 1.54730 | 正相关 | < 2E-16 | 1初中及以下  2中专及中技  3高中  4技校  5大专 | | | 0/1 |
| x8 | 文化程度2 | 2.37116 | 正相关 | < 2E-16 | 0/1 |
| x9 | 文化程度3 | 1.30299 | 正相关 | < 2E-16 | 0/1 |
| x10 | 文化程度4 | 3.30570 | 正相关 | 4.89E-16 | 0/1 |
| x11 | 文化程度5 | 1.87094 | 正相关 | < 2E-16 | 0/1 |
| x12 | 婚姻状况2 | 0.93748 | 正相关 | < 2E-16 | 2已婚  3丧偶  5离婚 | | | 0/1 |
| x13 | 婚姻状况3 | 2.05518 | 正相关 | 7.43E-03 | 0/1 |
| x14 | 婚姻状况5 | -0.90172 | 负相关 | < 2E-16 | 0/1 |
| x15 | 年龄 | 1.69205 | 正相关 | < 2E-16 | 47.06 | 18 | 103 | year |

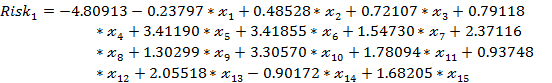
R上的逻辑回归会把离散型数据按值划分为不同的特征，值为0/1。性别、文化程度、婚姻状况都是离散型，数据说明见上表。

正相关表示随着值的增加，发病风险也会增加，该类因素包括：体重、收缩压、舒张压、性别（男）、性别（女）、文化程度（初中及以下）、文化程度（中专及中技）、文化程度（高中）、文化程度（技校）、文化程度（大专）、婚姻状况（已婚）、婚姻状况（丧偶）、年龄。其中，性别为女，文化程度为技校、中专及中技，婚姻状况为丧偶，比其它值更能增加风险。

负相关表示随着值的增加，发病风险反而降低，该类因素包括：身高、婚姻状况（离婚）。但这些因素的权重都不超过1。

二、发病风险公式

根据逻辑斯蒂回归模型，得出如下发病风险公式：



公式中，为发病的风险概率，对应的因素见表5.8。



三、单变量分布与分析

1、身高（权重：-0.23797）

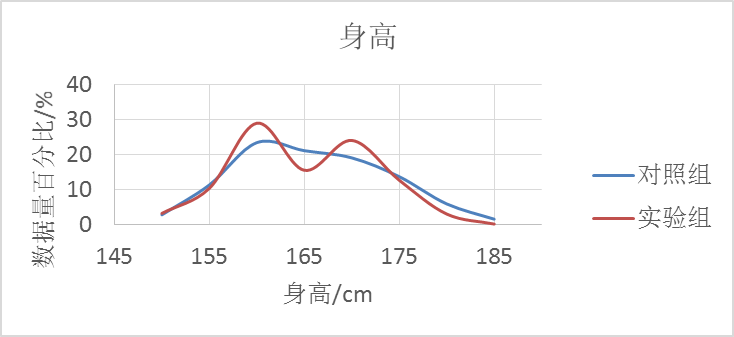


图5.9 身高分布对比

考虑到男女身高存在差异性，计算身高和性别的相关性系数后发现为-0.6948416（0.3~0.8表示具有中度相关性），具有中度相关性。女性身高通常在[155, 175]区间，男性身高通常在[165, 185]区间。



从图5.9可以看出，对照组身高有两个不明显的波峰，且较均匀分布在[160, 175]区间；而实验组身高有两个明显波峰，主要集中在160cm、170cm附近。

数据分布显示，高血压人群中，女性的身高主要集中在160cm附近，男性的身高主要集中在170cm附近，属于较低范围；而健康人群主要落在在[160， 175]区间，属于较高范围。身高与高血压发病率呈负相关，这也与建模结果相符。且身高的P值为5.58E-6，十分显著。



由此得出，身高越低，高血压发病率就越高。

2、体重（权重：048528）

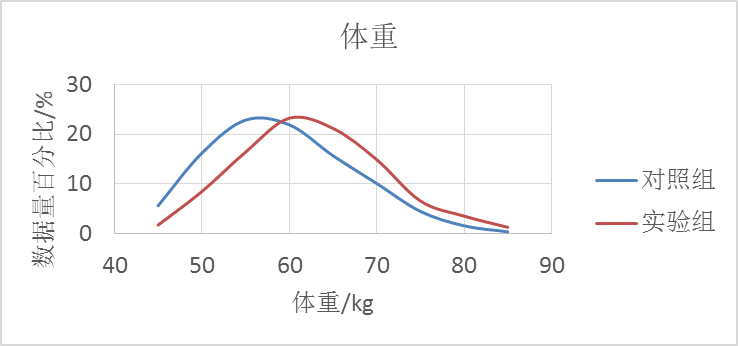


图5.10 体重分布对比

文献调研显示，BMI经常出现在主要影响因素中，但本文的模型中BMI不是显著的因素，计算体重和BMI的相关性系数后发现为 0.734319，具有中度相关性。可见在本文的模型中，BMI的影响很大程度体现在体重上，导致BMI本身不显著。

从图5.10可以看出，两组的分布曲线都呈正态分布，且幅度类似。但实验组峰值为60kg，对照组峰值为55kg。

数据分布显示，高血压人群体重偏高，BMI也偏高；健康人群体重偏低，BMI也偏低。体重与高血压发病率呈正相关，这也与与建模结果相符。且体重的P值小于2E-16，十分显著。

由此得出，体重越高，BMI越高，高血压发病率就越高。

3、收缩压（权重：0.72107）

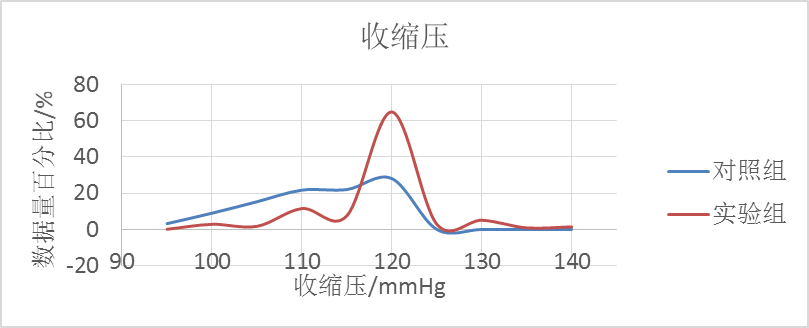


图5.11 收缩压分布对比

医学上，收缩压120mmHg是健康血压和正常高值的分水岭，而正常高值也被称为临界高血压值，处在正常高值的人是高血压高危人群。

从图5.11可以看出，实验组的收缩压大多集中在120mmHg附近，而对照组的收缩压大多低于120mmHg。另外，实验组分布曲线较对照组陡峭，尤其在波峰120mmHg处，实验组有65%左右的数据落在附近，而对照组只有30%的左右落在该处。最后，两个曲线于115mmHg处交叉，在区间，实验组人数明显多于对照组；而在区间，实验组人数明显少于对照组。



即，高血压人群的大多在正常高值范围内，或者接近正常高值；而健康人群的则大多在正常范围内。

数据分布显示，高血压人群的大多在正常高值范围内，或者接近正常高值；而健康人群的则大多在正常范围内。收缩压越偏离正常范围，其高血压发病率也就越高，呈正相关，这也与建模结果相符。且收缩压的P值小于2E-16，十分显著。

由此得出，收缩压越偏离正常范围，高血压发病率就越高。

4、舒张压（权重：0.79118）

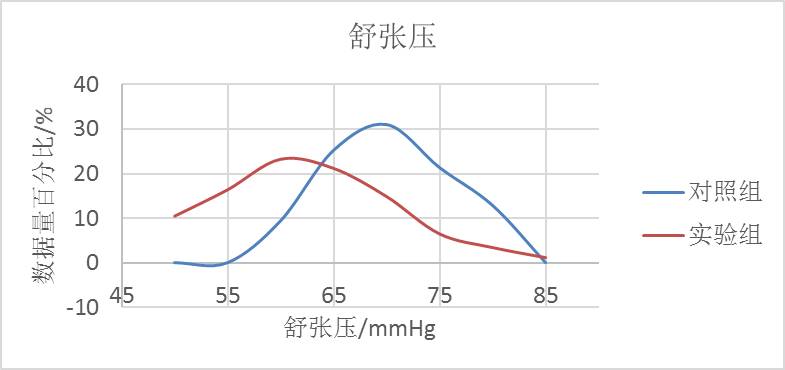


图5.12 舒张压分布对比

医学上，舒张压60mmHg是低血压和正常血压的分水岭，低血压也是血压不正常的表现之一。

从图5.12可以看出，两组的分布曲线具有相似的形状。实验组有40%的数据处于区间；而对照组只有10%的数据落在该区间。而且，两条曲线于64mmHg处交叉，在区间，实验组人数明显多于对照组；而在区间，实验组人数明显少于对照组。



即，高血压人群大多处于低血压范围内，或接近低血压范围；而健康人群则处于正常血压范围内。

数据分布显示，高血压人群大多处于低血压范围内，或接近低血压范围；而健康人群则处于正常血压范围内。舒张压越偏离正常范围，其高血压发病率也就越高，呈正相关，这也与建模结果相符。且收缩压的P值小于2E-16，十分显著。

由此得出，舒张压越偏离正常范围，高血压发病率就越高。

5、性别（权重：男/3.41190 女/3.41855）

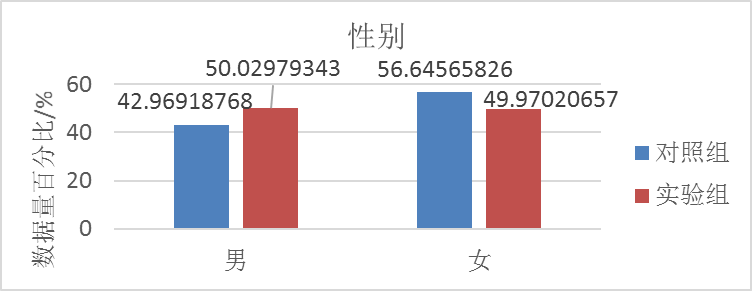


图5.13 性别分布对比

从图5.13可以看出，两组性别分布相对平均，但实验组男性更多，对照组女性更多，但二者差距并不十分明显。

数据分布显示，高血压人群中有更多男性，健康人群中有更多女性，这与模型中男性权重低于女性权重相反。同时，男性的P值为1.26E-03，女性的P值为1.23E-03，比较显著。

考虑到性别与身高具有中度相关性，该现象可能是由身高带来，性别本身可能并非主要影响因素。

6、文化程度（权重：初中及以下/1.54730 中专及中技/2.37116 高中/1.30299 技校/3.30570 大专/1.87094）

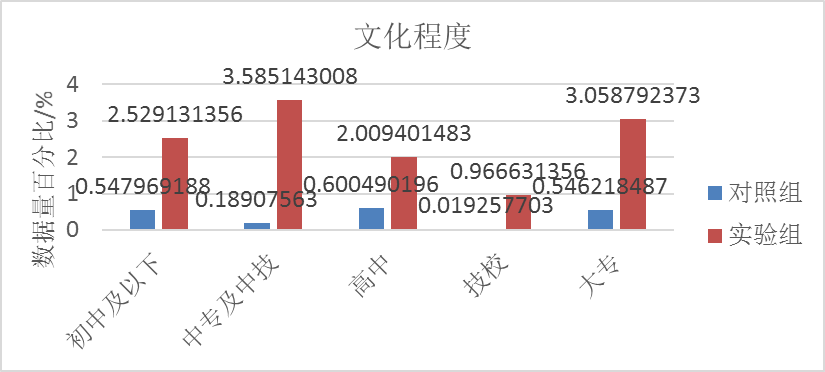


图5.14 文化程度分布对比

从图5.14可以看出，文化程度特征中的值——初中及以下、中专及中技、高中、技校、大专——所含数据量占总数据量的比例不足20%，无法从数据分布上得出支持或反对模型的结论。

7、婚姻状况（权重：已婚/0.93748 离婚/-0.90172）

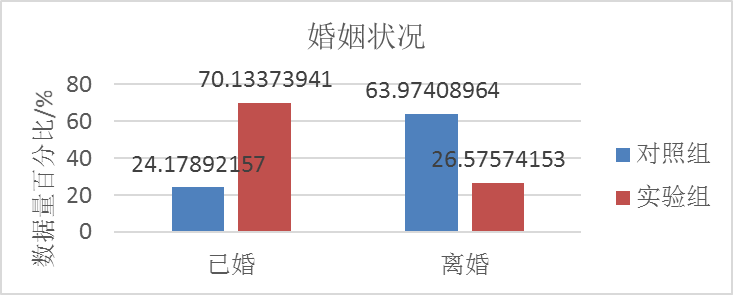


图5.15 婚姻状况分布对比

从图5.15可以看出，实验组较对照组有更多的已婚人群，数据量高达70%；对照组组有更多的离婚人群，数据量高达64%。

数据分布显示，高血压人群中有更多人已婚，健康人群有更多人离婚，这也与建模结果相符。且二者的P值都小于2E-16，十分显著。

由此得出，已婚人群高血压发病率高，离婚人群高血压发病率低。

8、年龄（权重：1.69205）

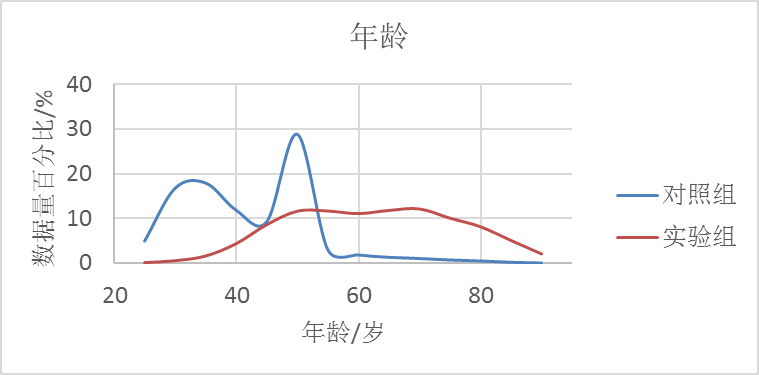


图5.16 年龄分布对比

从图5.16可以看出，实验组较为平缓，有不明显的两个波峰50岁、70岁，对照组较为陡峭，有十分明显的两个波峰30岁、50岁。而且，实验组大多集中在区间，对照组大多集中在区间，即发病人群较健康人群年长。最后，两组曲线于50岁处交叉，实验组80%的数据分布在50岁以上，对照组80%的数据分布在50岁以下。



数据分布显示，高血压人群中年长者较多，且大多在50岁以上；健康人群中年轻者较多，且大多在50岁以下。高血压与年龄呈正相关，这也与建模结果相符。且年龄的P值小于2E-16，十分显著。

由此得出，年龄越大，高血压发病率越高，尤其在50岁以后，要重视体检。

**5.5.2 高血压患严重并发症事件**

并发症事件以患严重并发症的人群为实验组，预后稳定的人群为对照组，两组人都为高血压患者，研究两者高血压早期的有关数据。

一、主要影响因素

三份数据集上，逻辑斯蒂回归建模的AUC = 0.81/0.83/0.78，效果最好的模型基于hyperbfz.sample数据集。只保留显著性不大于0.05且权重大于0.01的作为主要影响因素，在三个并发症风险模型中出现过的有：腰围、收缩压、舒张压、性别、婚姻状况、症状、遵医行为、年龄、患病年龄。

以AUC=0.83的模型为研究对象，结果如表5.9所示：

表5.9 高血压患严重并发症风险建模详情

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 参数 | 因素 | 权重 | 相关性 | P值 | 平均值 | 最小值 | 最大值 | 单位 |
| x0 | 调整项 | -3.45822 | 负相关 | < 2E-16 | —— | | | —— |
| x1 | 腰围 | -0.23900 | 负相关 | 1.50E-07 | 81.06 | 50 | 128 | cm |
| x2 | 收缩压 | -0.44546 | 负相关 | 1.09E-15 | 132.16 | 78 | 220 | mmHg |
| x3 | 舒张压 | -0.21958 | 负相关 | 0.000104 | 82.04 | 40 | 150 | mmHg |
| x4 | 性别2 | -0.33567 | 负相关 | 0.007517 | 2女 | | | 0/1 |
| x5 | 婚姻状况9 | 1.41673 | 正相关 | 1.86E-05 | 9其他 | | | 0/1 |
| x6 | 症状1 | -0.81707 | 负相关 | 4.03E-05 | 1头痛头晕  9无症状 | | | 0/1 |
| x7 | 症状9 | -0.73302 | 负相关 | 3.81E-06 | 0/1 |
| x8 | 遵医行为2 | -0.42265 | 负相关 | 0.003135 | 2一般 | | | 0/1 |
| x9 | 年龄 | 0.72990 | 正相关 | < 2E-16 | 57.20 | 18 | 102 | year |

R上的逻辑回归会把离散型数据按值划分为不同的特征，值为0/1。性别、婚姻状况、症状、遵医行为都是离散型，数据说明见上表。

正相关表示随着值的增加，并发症的风险也会增加，该类因素包括：婚姻状况（其他）、年龄。

负相关表示随着值的增加，并发症风险反而降低，该类因素包括：腰围、收缩压、舒张压、性别（女）、症状（头痛头晕、无症状）、遵医行为（一般）、患病年龄。

二、患严重并发症风险公式：

根据并发症的逻辑斯蒂回归模型，得出如下并发症风险公式：



公式中，为患严重并发症的风险概率，x0~x10的对应如下主要影响因素规约化后的值：调整项、腰围、收缩压、舒张压、女、婚姻状况（其他）、症状（头痛头晕）、症状（无症状）、遵医行为（一般）、年龄、患病年龄。



三、单变量分布与分析

1、腰围（权重（-0.23900））

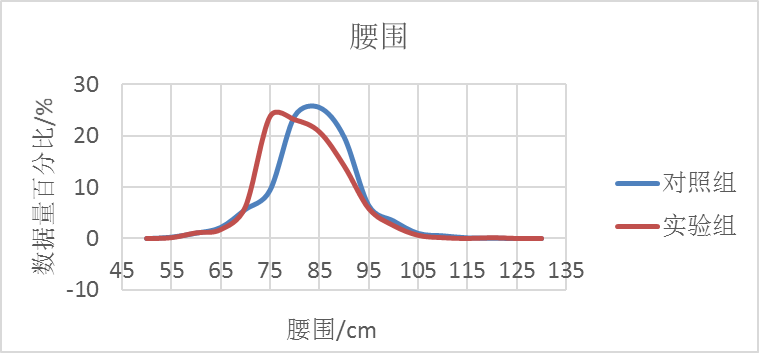


图5.17 腰围分布对比

从图5.17可以看出，两组分布曲线类似，均呈正态分布。其中，实验组波峰在75cm附近，对照组波峰在78cm附近，且实验组腰围整体较对照组腰围偏小。即患严重并发症的人群腰围更小，预后稳定的人群腰围更高。

但是，腰围作为肥胖的一个重要指标，通常认为腰围越大患严重并发症的风险也越高，这与数据分布相反。针对这点，抽样查看实验组身高变动，发现患者在确诊高血压前腰围减小，表明患者当时身体状况已经不好。

由此得出，对于高血压患者而言，并非腰围越低越健康，更可能腰围越低身体状况越不佳，发病率就越高。

2、收缩压（权重：-0.44546）

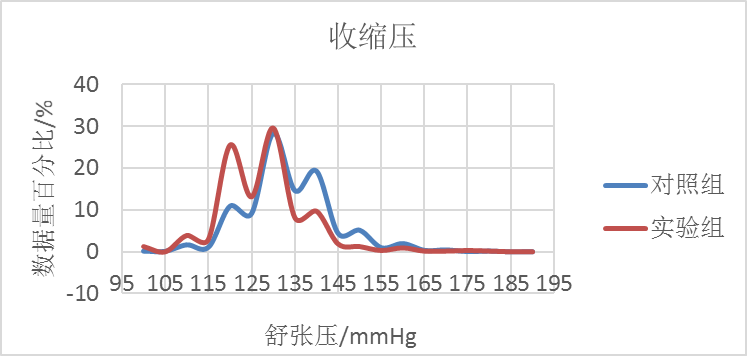


图5.18 收缩压分布对比

计算相关性系数后发现收缩压与其他特征仅具有低相关度，即收缩压与其它特征无线性关系。另外，两组人群均为高血压患者，都需服用降压药让收缩压控制在正常高值范围内。



从图5.18可以看出，两组曲线形状类似，都有6个波峰，且波峰几乎都落在110mmHg、120mmHg、130mmHg、140mmHg、150mmHg、160mmHg处，而且两组均有80%的数据处于区间，但实验组120mmHg的人更多，对照组140mmHg的人更多。此外，两条曲线于135mmHg处交叉，实验组比对照组有更多人的收缩压低于135mmHg。



数据分布显示，在高血压早期，那些后期患并发症的人群更严格地控制了收缩压，那些预后良好的人群相对较宽松的控制了收缩压。后期患并发症的人群更多地将收缩压控制在正常血压的范围内，而预后稳定的人群则大多控制在正常高值范围内。且收缩压越低越接近正常值，患并发症的人就越多。收缩压与高血压发病率呈负相关，这也与建模结果相符。且收缩压的P值为1.09E-15，极为显著。



由此得出，收缩压的控制并非越严格越好，需要考虑到身体状况。收缩压控制在正常高值即可，若强制控制在正常收缩压反而会增加患严重并发症的风险。收缩压越低，严重并发症的发病率就越高。



3、舒张压（权重-0.21958）

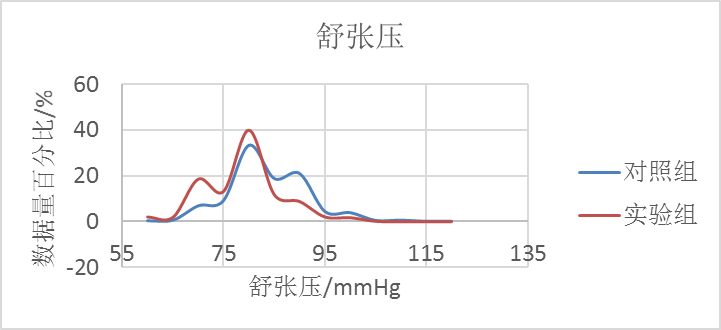


图5.19 舒张压分布对比

计算相关性系数后发现舒张压与其他特征仅具有地相关度，即舒张压与其他特征无线性关系。另外，两组人群均为高血压患者，都需服用降压药让舒张压控制在正常高值范围内。



从图5.19可以看出，舒张压的情况与收缩压的情况类似，差别只在于舒张压只有4个波峰——70mmHg、80mmHg、90mmHg，且两条曲线于80mmHg处交叉，这也是正常血压与正常高值的分水岭。

数据分布显示，后期患并发症的人群更多地将舒张压控制在范围内，而预后稳定的人群则大多控制在正常高值范围内。且舒张压越低岳接近正常值，患并发症的人就越多。舒张压与高血压发病率呈负相关可这也与建模结果以及收缩压的情况相符。且舒张压的P值为0.000104，十分显著。



由此得出，舒张压的空值并非越严格越好，需要考虑到身体状况。舒张压控制在正常高值即可，若强制控制在正常舒张压反而会增加患严重并发症的风险。舒张压越低，严重并发症的发病率就越高。



4、性别（权重：女/-0.33567）

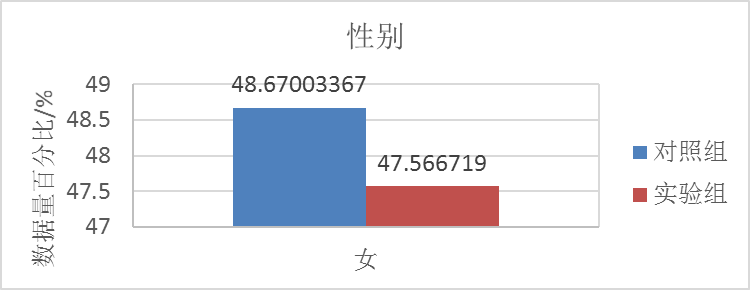


图5.20 性别分布对比

从图5.20可以看出，两组性别的数据量百分比相差不大，实验组47%，对照组48%，实验组女性更少，但二者差距并不明显。

数据分布显示，后期患严重并发症的人群中有更少女性，呈负相关，这也与建模结果相符。且女性的P值为0.007517，较为显著。

由此得出，性别为女，严重并发症发病率就越低。

5、婚姻状况（权重：其他/1.41673）

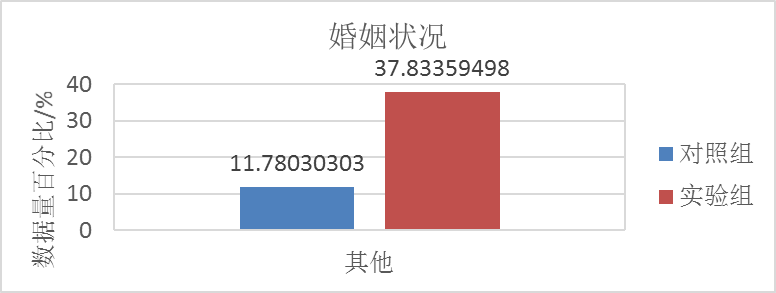


图5.21 婚姻状况分布对比

婚姻状况特征有未婚、已婚、离婚、丧偶、其他这几个值。

从图5.21可以看出，实验组有38%的婚姻状况为“其他”的人群，远高于对照组的12%，二者差距十分明显。

数据分布显示，后期患严重并发症的人群中有更多人婚姻状况为“其他”，呈正相关，这也与建模结果相符。且婚姻状况为“其他”的P值为1.86E-5，极为显著。

由此得出，与未婚、已婚、离婚、丧偶相比，婚姻状况为“其他”，严重并发症的发病率就越高。

6、症状（权重：头痛头晕/-0.81707 无症状/-0.73302）

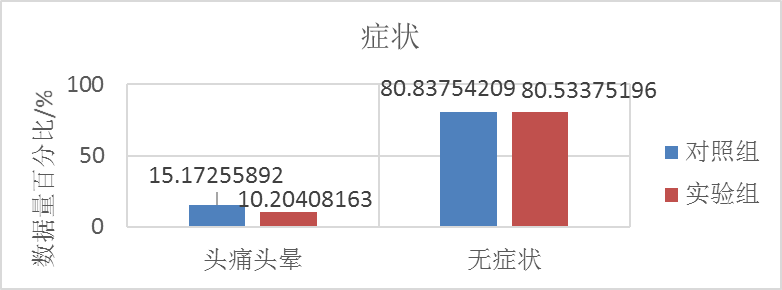


图5.22 症状分布对比

考虑到症状特征另外几个值，四肢发麻、心悸胸闷、其他，为更加严重的症状，而症状本身在判定病情时具有重要参考价值。

从图5.22可以看出，两组值为“头痛头晕”、“无症状”的数据分布情况类似，实验组的人数都比对照组少，但二者差距不大。

数据分布显示，后期患严重并发症的人群中有更少的人症状为头痛头晕或无症状，即患严重并发症的人群有轻症状的人少，呈负相关，这也与建模结果相符。且头痛头晕的P值为4.03E-5，无症状的P值为3.81E-6，极为显著。

由此得出，与四肢发麻、心悸胸闷个、其他相比，症状为头痛头晕或无症状，患严重并发症的概率就越低。

7、遵医行为（权重：一般/-0.42265）

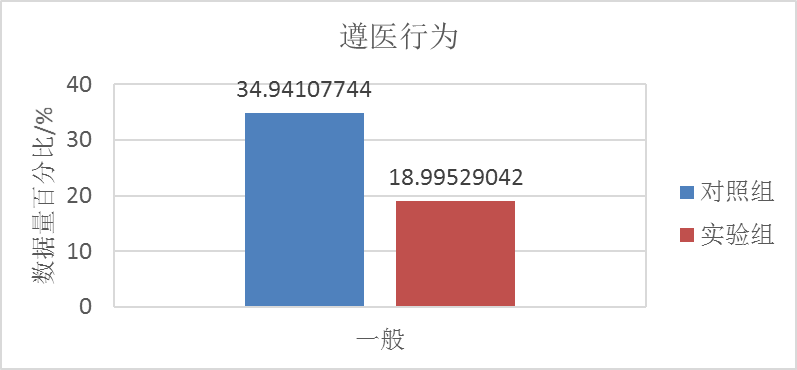


图5.23 遵医行为分布对比

从图5.23可以看出，遵医行为“一般”的情况中，实验组只有19%，对照组却有35%，二者差距明显。

数据分布显示，患严重并发症的人群中有更少的人遵医行为“一般”，呈负相关，这也与建模结果相符。同时，遵医行为“一般”的P值为0.003135，较为显著。

由此得出，遵医行为“一般”的人，即比较遵守医嘱的人，严重并发症的发病率越低。

8、年龄（权重：0.72990）

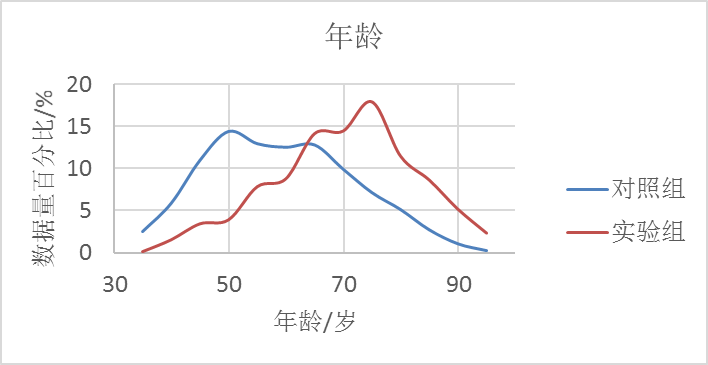


图5.24 年龄分布对比

从图5.24可以看出，两组波形差别较明显，实验组大多分布在在区间，对照组大多分布在区间。而且，两组曲线于67岁处交叉，对照组有75%的数据落在67岁以上，实验组有65%的数据落在67岁以下。



数据分布显示，在高血压前期，那些后期患严重并发症的人群更为年长，且大多在67岁以上，那些预后良好的人群更为年轻，且大多在67岁以下。并发症与年龄呈正相关，这也与建模结果相符。且年龄的P值小于2E-16，极为显著。

由此得出，年龄越大，严重并发症的发病率就越高，尤其在67岁以后，更要重视身体健康。

**第6章 生存分析和疾病谱图**

在医学研究中，常用到生存分析方法。所谓生存期，是指从某个标准时刻，如发病、确诊、开始治疗或进行手术的时间算起，至死亡或复发为止的时间[[[50]](#endnote-51)]。

本文在门诊数据分词结果和并发症数据集的基础上，对于高血压的预后考核，用死亡率来表示并发症的发病，生存率来表示预后良好的情况。大致步骤为：以高血压人群确诊时间为时间起点，计算患者各种并发症的发病相对时间，分析高血压人群2年内的并发症发展情况。

根据上文中风险分析的结果，选择收缩压、年龄、性别3个显著风险因子对高血压人群进行分组，以月为单位，统计糖尿病、冠心病、脑梗死3种并发症的各自发病情况。作生存分析时，将糖尿病、冠心病、脑梗死视为同一事件,采用K-M算法求生存率。

**6.1 生存分析**

**6.1.1 K-M算法及实现**

K-M（Kaplan - Meier）分析方法是一种求生存函数的非参数方法，其基本思想是将生存时间由小到大依次排列，在每个死亡点上，计算其期初人数、死亡人数、死亡概率、生存概率和生存率[[[51]](#endnote-52)]。不考虑删失，生存率公式如下：



其中，n为总数，di为t时间内总死亡数。

用PHP代码统计出各因素的生存率并存放在表中。以收缩压为例，新建表SBP\_survive并设置所需字段，然后运行PHP代码即可，所得结果存入数据库。

代码如下所示：

$conn=odbc\_connect("\*\*\*","\*\*\*","\*\*\*",SQL\_CUR\_USE\_ODBC);

if($conn) echo "connect success!";

else echo "failed to connect";

$sql = "select \* from copy\_hyperbfz where SBP<130";

$exec=odbc\_exec($conn,$sql);

$arr\_date=array(1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23,24);

$arr = array();

$num=0;

for($i=0;$i<count($arr\_date);$i++) $arr[$i]=0;

while($row=odbc\_fetch\_array($exec)){

$date\_cha = array();

$num++;

for($i=1;$i<25;$i++){

if($row["DISEASE".$i]==1&&$row["DATE".$i]!=null){

$date\_cha[]=abs(strtotime($row["DATE".$i]) - strtotime($row["HYPER\_DATE"]))/60/60/24/30;

}

}

if(!empty($date\_cha)){

$pos = array\_search(min($date\_cha), $date\_cha);

$min = $date\_cha[$pos];

for($i=0;$i<count($arr\_date);$i++){

if($min<=$arr\_date[$i]) $arr[$i]++;

else continue;

}

$arr\_survive = array();

for($i=0;$i<count($arr\_date);$i++) $arr\_survive[$i]=($num-$arr[$i])/$num;

echo "<br>";

for($i=0;$i<count($arr\_date);$i++){

echo $arr\_date[$i]." ";

echo $arr\_survive[$i];

echo "<br>";

}

**6.1.2 结果与分析**

一、SBP

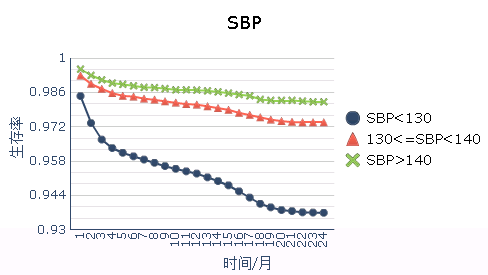


图6.1 收缩压生存曲线

SBP的生存曲线如图6.1所示。SBP(Systolic Blood Pressure)，即收缩压，按照SBP的取值范围将高血压人群分为三组：。 可以看出，的分组生存率更低，两个分组的生存率反而更高。



通常情况下，SBP的正常范围为，为，高血压范围为或以上。对于高血压人群而言，需服药将收缩压控制在的范围内，收缩压越低，表明控制效果越好。理论上，控制效果好的分组应该有更高的生存率，这与图6.1的观察结果相反。



考查收缩压的数据分布，发现在的人群中，的人生存率并不低，但的人生存率很低，即上图中的结果主要是由的人群带来的。而在医学上，120是健康值与正常高值的分水岭。



上述观察表明，血压控制并非越严格越好，将其控制在正常高值范围内即可，不必强行控制在健康血压范围，不然会给身体带来更大负担增加并发症的发病风险，与并发症风险模型的分析结果相符。

二、年龄

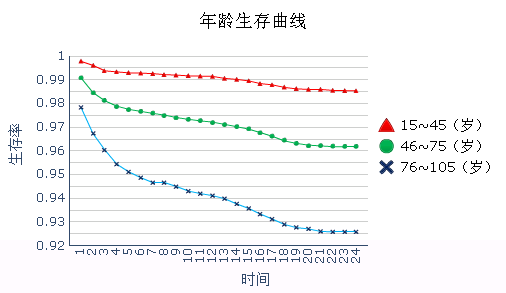


图6.2 年龄生存曲线

年龄生存曲线如图6.2所示。按照年龄取值范围将高血压人群分为三组：。从图中可以看出，年龄越高的人群生存率越低，与理论结果一致。



三、性别

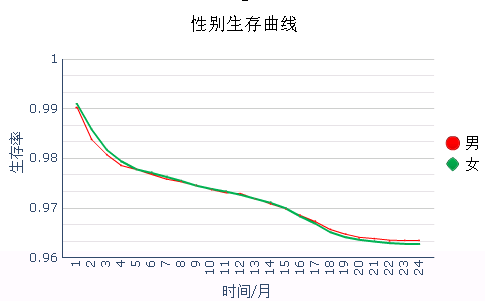


图6.3 性别生存曲线

按性别不同将高血压人群分为男女两组，生存曲线如图6.3所示。可以看出，男、女生存率变化基本一致，可见性别不是影响生存率的重要因素。

**6.2 疾病谱图**

**6.2.1 数据计算**

数据来源于病患的门诊数据，在分词结果的基础上，研究高血压人群确诊高血压后的常见病症，得到病症随时间变化的情况，按时间顺序绘制疾病谱图。

疾病谱图由点、有向线段组成，点代表病症，有向线段代表患病先后顺序，生成病症数据、病症关系数据的流程如图6.4所示：



图6.4 疾病谱图生成流程

一、生成病症数据

获得疾病谱图数据后，再次进行过滤，选出部分病症作为疾病谱图的点数据，合并相似病症，红色部分为已被合并的病症，具体如图6.5所示：



图6.5 疾病谱图中的疾病数据

二、生成病症关系数据

以高血压确诊时间为时间起点，分别获取3个月、6个月、1年、2年、3年、3年以上这6份数据作为点间的有向线段，部分结果如表6.1所示：

表6.1 疾病谱图中的关系数据示例

|  |  |  |
| --- | --- | --- |
| 起点 | 终点 | 数量 |
| 高血压 | 糖尿病 | 31 |
| 高血压 | 冠心病 | 29 |
| 高血压 | 高脂血症 | 29 |
| 糖尿病 | 高脂血症 | 23 |
| 高血压 | 脑梗塞 | 23 |
| 糖尿病 | 冠心病 | 23 |

**6.2.2 结果与分析**

将处理后的数据导入绘图工具Cytoscape中，点的大小、有向线段的粗细代表发病人数，点越大或线越粗代表发病人数愈多，疾病谱图如图6.6所示：

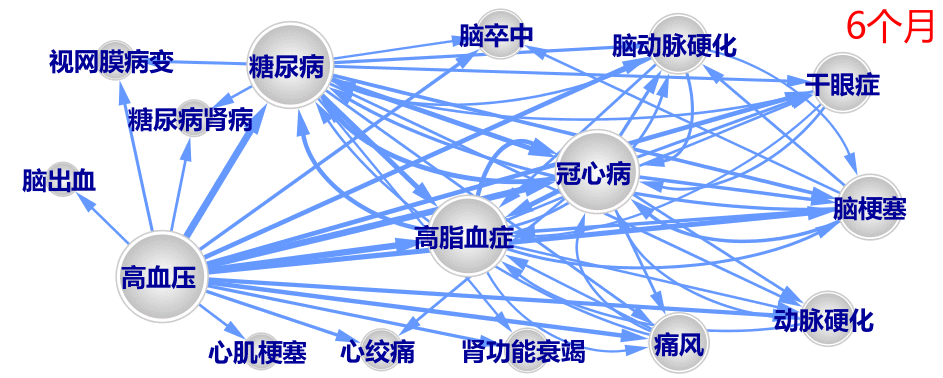
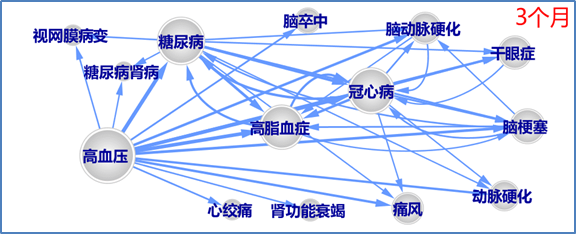
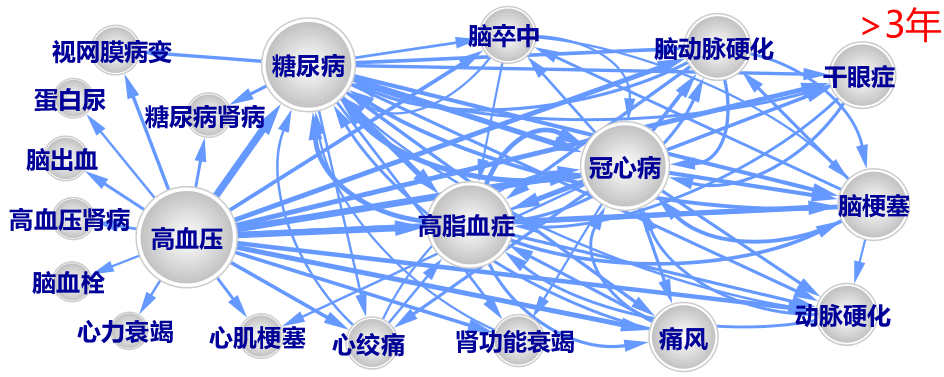
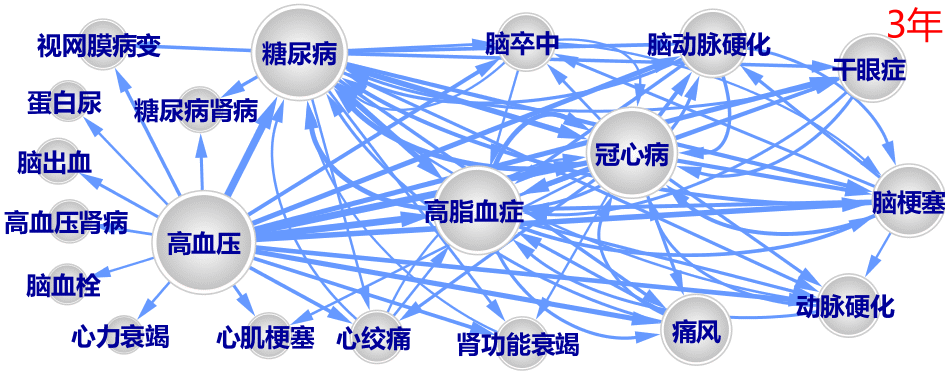
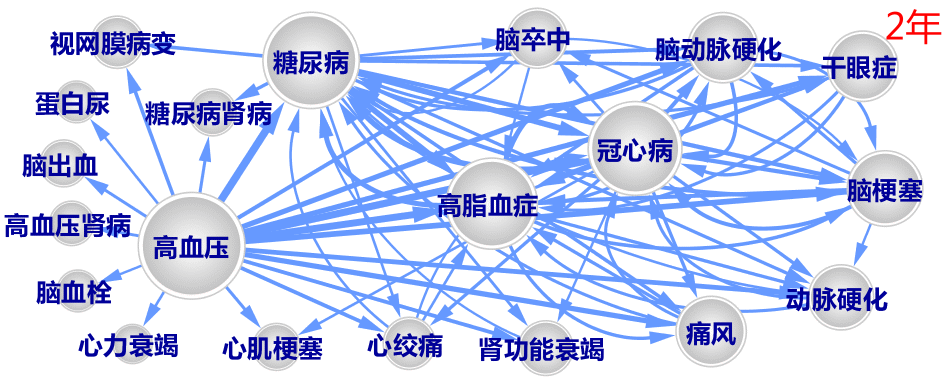
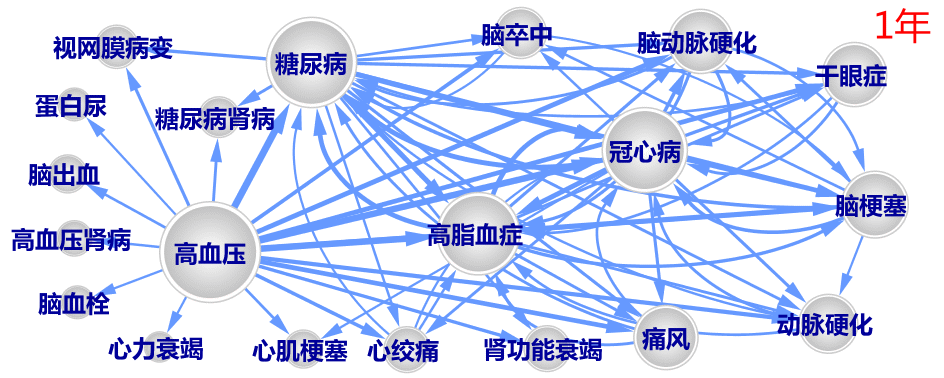


图6.6 不同考察期的疾病谱图



续图6.6 不同考察期的疾病谱图

可以看出，随着时间推移，并发症种类增多，由3个月时的13种发展到3年后的19种，增加了蛋白尿、脑出血、高血压肾病、脑血栓、心力衰竭、心肌梗塞6种病症。而糖尿病、冠心病、高血脂症、脑梗塞是最常见的几种高血压并发症。

**第7章 结论与展望**

**7.1 结论**

本文在全科医生工作站和区域医疗健康信息系统的基础上，运用数据挖掘技术、数据库技术、统计分析方法、医学知识等，以高血压发病事件、高血压患严重并发症事件的相关数据为研究对象，给出了完整的研究方案和相应成果。

本文的工作主要包括：

1、探查深圳市1450万份居民健康档案，包括30亿条诊疗数据，确定数据基本对应关系，确定了数据库需求和特征评分标准。

2、建立高血压数据库，采用Merge引擎进行分表，优化数据库存取性能。

3、通过数据存在性与分布性探查、数据清洗、数据转换、数据规约等步骤，生成两大事件共计6份数据集。

4、运用逻辑斯蒂回归、朴素贝叶斯、随机森林三种分类算法处理数据集，得出两大事件共计18个风险模型。

5、采用30次10折交叉验证，计算AUC并绘制ROC曲线，比较建模效果，发病事件效果最好的模型AUC = 0.96，并发症事件效果最好的模型AUC = 0.83。

6、根据逻辑斯蒂回归模型，分别得出发病事件、并发症事件的主要影响因素，给出风险公式，分析建模结果。

7、分别按收缩压、年龄、性别划分高血压人群，采用K-M算法计算生存率，用PHP绘制生存曲线，比较生存曲线并进行分析。

8、通过计算疾病数据和疾病关系数据，绘制疾病谱图，给出简单分析。

针对高血压发病、严重并发症事件，分别得出如下结论：

1、在逻辑斯蒂回归、朴素贝叶斯、随机森林三种分类算法中，随机森林对发病数据集的建模效果最好，AUC=0.96；逻辑斯蒂回归对并发症数据集的建模效果最好，AUC=0.83。

2、对高血压发病事件进行逻辑斯蒂回归建模，得出如下结论：身高越低、体重越高、BMI越大、血压值越偏离正常范围、文化程度为初中及以下/中专及中技/高中/技校/大专、婚姻状况为已婚、年龄越大的人群，高血压发病率偏高，尤其是在50岁左右，要重视体检。

3、对高血压严重并发症事件进行逻辑斯蒂风险建模，得出如下结论：对于高血压患者而言，并非腰围越低越健康，更可能腰围越低身体状况越不佳，并发症发病率越高；血压控制在正常高值范围内即可，不必强行控制在健康血压范围内，控制得越严格给身体带来的负担越重，并发症发病率越高；女性的并发症发病率更低，但并不十分明显；与未婚、已婚、离婚、丧偶相比，婚姻状况为“其他”的人群并发症发病率更高；门诊诊疗出头痛头晕、无症状等较轻的症状人，并发症发病率更低；比较遵守医嘱的人，并发症发病率更低。



4、分别按照收缩压、年龄、性别，将并发症事件的样本划分为不同的组，分别绘制生存曲线。生存曲线下降越快，表明患严重并发症的人越多；反之，表明患严重并发症的人越少。收缩压的生存曲线表明，血压控制并非越严格越好，支持了风险建模中关于血压的结论；年龄的生存曲线表明，年龄越高，生存率下降越快，支持了风险建模中关于年龄的结论；性别的生存曲线表明，不同性别生存率下降趋势一致，不存在较大差别，支持了风险建模中关于性别的结论。

5、通过分词得出疾病数据和疾病发生的先后关系数据，绘制高血压疾病谱图，发现随着时间推移，并发症种类增多，由3个月的13种严重并发症发展到3年的19种。最早发生的疾病为糖尿病、冠心病、高血脂症、脑栓塞，后期继发的为蛋白尿、脑出血、高血压肾病、脑血栓、心力衰竭、心肌梗塞。

对于高血压发病事件，随机森林算法得出的模型AUC=0.96，具有非常高的准确性，对于预测居民患高血压的风险具有重要实用价值。对主要影响因素及有关数据的分析，一方面支持了年龄、肥胖、偏离正常范围的血压、身高是高血压发病事件的高危因素，另一方面定量地确定了文化程度、婚姻状况对发病的影响。

对于高血压严重并发症事件，逻辑斯蒂回归得出的模型AUC=0.83，具有很高的准确性，对于高血压患者的预后评估与指导具有重要参考价值。对主要影响因素及有关数据的分析，一方面支持了症状、遵医行为对高血压预后的积极作用，另一方面也明确了预后期间血压值应该控制的范围，纠正了人们认为腰围越小患严重并发症的可能性越低的错误认识。而对高血压患者的生存曲线进行的研究支持了本文的观点。

**7.2 进一步工作的方向**

本文基于高血压患者的数据进行了一些研究并取得了初步成功，但仍旧有许多工作可以进行，主要包括：

1、针对高血压的发病、并发症事件，本文是以常规体检次数大于或等于5次的人群为研究对象，但这些患者中拥有血检数据的人数较少，血检数据包括了许多重要的生化指标，如血压、血糖、胆固醇含量、蛋白酶含量等。后续研究可把研究重点放在血检数据上，在保证一定样本量的前提下，以此为标准生成数据集，进一步研究高血压的发病、并发症事件。

2、本文把严重的高血压并发症视为一个事件，对其进行风险建模，后期可针对具体的某个并发症，比如糖尿病、冠心病、慢性肾衰竭、心力衰竭等进行风险建模、生存分析，专注研究高血压的一个并发症的相关情况。

3、由于深圳是一座国际化都市，人口成分复杂，流动性强，再加上作为中国和世界的交通枢纽，一旦重大传染病在此地爆发，后果不堪设想。健康档案覆盖了深圳市1450万居民，记录了包括呼吸道传染病、艾滋病等传染病的有关，对这方面的数据进行挖掘具有重大意义。

4、深圳市卫计委统计显示，呼吸道疾病是深圳市发病率最高的疾病之一，而呼吸道疾病又通常与当地天气、气候条件紧密关联。随着环境污染，尤其是空气污染的加剧，挖掘呼吸道疾病的发病与环境的关系对于指定卫生医疗管理策略具有重要意义，也对防治该类疾病具有重大作用。

希望在医疗健康的数据挖掘上，我国能有比Google Flu等国际一流的成果更好的研究，为世界的医疗卫生行业发展做出应有贡献。

**致 谢**

写到这里已经是论文的结篇之词，几次抬手总敲不出字来，总结与回首总是令人感慨万千。硕士时光不比大学，大多都在校外度过，上课、交换、实习构成了我硕士时光的主旋律，但曲调的生命力往往由副歌造就，家人、朋友、同学、老师，还有遇到过的可爱人们，那些不期而遇的温暖，不禁令人思及莞尔。

首先要感谢导师和中心的老师。导师带我入行并给我宝贵的实习机会，到更广阔的平台去学习、研究。非常感谢老师在长达一年的时间里给我的指导，从最初的收集资料，到探查数据、学习统计与医学知识、学习数据挖掘基础及工具、跟踪项目进度、带学弟学妹，给我宝贵的锻炼机会，也十分理解找工作期间的忙碌，最重要的是，老师在研究上的睿智、细致和坚持给我很大启发，感谢有机会参与863项目，在此向他们表示最真挚的感谢。

同时，感谢交换期间带过我的老师、师母和实验室的同学们带给我的启发和温暖，感谢所有任课老师在学业上给我的指导，感谢中心的老师们给我机会和指导，感谢中心的同事，感谢学弟学妹给过的支持与帮助。感谢班上的同学们，风雨同舟，一起解决学习上、生活上遇到的问题，教会我学习和做人的道理，共同进步。

最后，感谢家人和朋友，在生活和学习上给与的支持和帮助，这些温馨与幸福是最宝贵的回忆。

秦元博

2015年4月1日

**参考文献**

1. [] Okechukwu S Ogah, Ikechi Okpechi, Innocent I Chukwuonye, et al. Blood pressure, prevalence of hypertension and hypertension related complications in Nigerian Africans: A review[J]. World Journal of Cardiology, 2012, 12: 327-340. [↑](#endnote-ref-2)
2. [] 王文, 朱曼璐, 王拥军, 吴兆苏, 等. 《中国心血管病报告2012》概要[J]. 中国循环杂志, 2013, 06: 408-412. [↑](#endnote-ref-3)
3. [] 王增武, 黄艳红. 中国高血压患病率与知晓率和控制率现状[J]. 中华健康管理学杂志, 2013, 7(2): 139-140. [↑](#endnote-ref-4)
4. [] 陈伟伟, 高润霖, 刘力生, 等. 中国心血管病报告2013概要[J]. 中国循环杂志, 2014, 7: 487-491. [↑](#endnote-ref-5)
5. [] 刘力生. 中国高血压防治指南2010[J]. 中华高血压杂志, 2011, 08: 701-743. [↑](#endnote-ref-6)
6. [] 温家宝. 关于发展社会事业和改善民生的几个问题[J]. 求是, 2010, 7(1): 19 [↑](#endnote-ref-7)
7. [] 范明宽. 深圳市社区卫生服务现状研究[D]. 华中科技大学, 2013. [↑](#endnote-ref-8)
8. [] 张晓燕, 唐世琪, 梁倩君. 美国健康管理模式对我国健康管理的启示[J]. 中华健康管理学杂志, 2010, 4(5): 315-317. [↑](#endnote-ref-9)
9. [] Berk M L, Monheit A C. The concentration of health care expenditures, revisited[J]. Health Affairs, 2001, 20(2):9-18. [↑](#endnote-ref-10)
10. [] World Health Organization. Reproductive Health, World Health Organization. Chronic Diseases, Health Promotion. Comprehensive cervical cancer control: a guide to essential practice[M]. World Health Organization, 2006. [↑](#endnote-ref-11)
11. [] Gavrilov L A, Heuveline P. Aging of population[J]. The encyclopedia of population, 2003, 1:32-37. [↑](#endnote-ref-12)
12. [] Nachman L, Baxi A, Bhattacharya S, et al. Jog falls: a pervasive healthcare platform for diabetes management[M]//Pervasive Computing. Springer Berlin Heidelberg, 2010: 94-111. [↑](#endnote-ref-13)
13. [] Neuvirth H, Ozery-Flato M, Hu J, et al. Toward personalized care management of patients at risk: the diabetes case study[C]//Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2011: 395-403. [↑](#endnote-ref-14)
14. [] Windham B G, Bennett R G, Gottlieb S. Care management interventions for older patients with congestive heart failure[J]. The American journal of managed care, 2003, 9(6): 447-59; quiz 460-1. [↑](#endnote-ref-15)
15. [] Phillips C O, Wright S M, Kern D E, et al. Comprehensive discharge planning with postdischarge support for older patients with congestive heart failure: a meta-analysis{j}. Jama, 2004, 291(11): 1358-1367. [↑](#endnote-ref-16)
16. [] Ekman I, Swedberg K. Home-based management of patients with chronic heart failure - focus on content not just form![J]. European heart journal, 2002, 23(17): 1323-1325. [↑](#endnote-ref-17)
17. [] Green B B, Cook A J, Ralston J D, et al. Effectiveness of home blood pressure monitoring, Web communication, and pharmacist care on hypertension control: a randomized controlled trial[J]. Jama, 2008, 299(24): 2857-2867. [↑](#endnote-ref-18)
18. [] Tanguay S, Heywood P. MyFoodPhone: the start of a mobile health revolution[J]. Mobile persuasion, 2007, 20: 21-27. [↑](#endnote-ref-19)
19. [] World Health Organization. How to investigate drug use in health facilities: selected frug use indicators[J]. 1993. [↑](#endnote-ref-20)
20. [] 代少升, 张跃, 张和君, 等. 便携式远程心电实时监护仪的研制[J]. 医疗卫生装备, 2006, 27(9): 1-2. [↑](#endnote-ref-21)
21. [] 童基均, 郭希山, 潘敏, 等. 基于无线技术的远程病人监护系统[J]. 国际生物医学工程杂志, 2007, 30(2): 128-128 [↑](#endnote-ref-22)
22. [] 冯娟. 家用便携式远程医疗监护系统[J]. 中国科技信息, 2009(4): 200-201. [↑](#endnote-ref-23)
23. [] Sagiroglu S, Sinanc D. Big data: A review[C]//Collaboration Technologies and Systems (CTS), 2013 International Conference on. IEEE, 2013: 42-47. [↑](#endnote-ref-24)
24. [] Brownstein J S, Freifeld C C, Madoff L C. Digital disease detection—harnessing the Web for public health surveillance[J]. New England Journal of Medicine, 2009, 360(21): 2153-2157. [↑](#endnote-ref-25)
25. [] National Electronic Disease Surveillance System Working Group. National Electronic Disease Surveillance System (NEDSS): a standards-based approach to connect public health and clinical medicine[J]. Journal of public health management and practice: JPHMP, 2001, 7(6): 43. [↑](#endnote-ref-26)
26. [] Victor L Y, Madoff L C. ProMED-mail: an early warning system for emerging diseases[J]. Clinical infectious diseases, 2004, 39(2): 227-232. [↑](#endnote-ref-27)
27. [] Blench M. Global public health intelligence network (GPHIN)[C]//Proceedings of the conference of the American machine translation association (AMTA). Waikiki, Hawai’i: AMTA. 2008. [↑](#endnote-ref-28)
28. [] Freifeld C C, Mandl K D, Reis B Y, et al. HealthMap: global infectious disease monitoring through automated classification and visualization of Internet media reports[J]. Journal of the American Medical Informatics Association, 2008, 15(2): 150-157. [↑](#endnote-ref-29)
29. [] Linge J P, Belyaeva J, Steinberger R, et al. Medisys: Medical information system[J]. Advanced ICTs for Disaster Management and Threat Detection: Collaborative and Distributed Frameworks, 2010: 131-142. [↑](#endnote-ref-30)
30. [] Keller M, Blench M, Tolentino H, et al. Use of unstructured event-based reports for global infectious disease surveillance[J]. Emerging infectious diseases, 2009, 15(5): 689. [↑](#endnote-ref-31)
31. [] Collier N, Doan S, Kawazoe A, et al. BioCaster: detecting public health rumors with a Web-based text mining system[J]. Bioinformatics, 2008, 24(24): 2940-2941. [↑](#endnote-ref-32)
32. [] Butler D. Mashups mix data into global service[J]. Nature, 2006, 439(7072): 6-7. [↑](#endnote-ref-33)
33. [] Ginsberg J, Mohebbi M H, Patel R S, et al. Detecting influenza epidemics using search engine query data[J]. Nature, 2009, 457(7232): 1012-1014. [↑](#endnote-ref-34)
34. [] Lazer D M, Kennedy R, King G, et al. The parable of Google Flu: traps in big data analysis[J]. 2014. [↑](#endnote-ref-35)
35. [] Kannel W B, McGee D L. Diabetes and cardiovascular disease: the Framingham study[J]. Jama, 1979, 241(19): 2035-2038. [↑](#endnote-ref-36)
36. [] Hubert H B, Feinleib M, McNamara P M, et al. Obesity as an independent risk factor for cardiovascular disease: a 26-year follow-up of participants in the Framingham Heart Study[J]. Circulation, 1983, 67(5): 968-977. [↑](#endnote-ref-37)
37. [] 数据挖掘: 概念与技术[M]. 机械工业出版社, 2007. [↑](#endnote-ref-38)
38. [] 朱凌云, 吴宝明, 曹长修. 医学数据挖掘的技术, 方法及应用[J]. 生物医学工程学杂志, 2003, 20(3): 559-562. [↑](#endnote-ref-39)
39. [] 龚著琳, 陈瑛, 苏懿. 数据挖掘在生物医学数据分析中的应用[J]. 上海交通大学学报 (医学版), 2010, 30(11): 1420. [↑](#endnote-ref-40)
40. [] Pasolli E, Melgani F. Active learning methods for electrocardiographic signal classification[J]. Information Technology in Biomedicine, IEEE Transactions on, 2010, 14(6): 1405-1416. [↑](#endnote-ref-41)
41. [] Hosmer Jr D W, Lemeshow S. Applied logistic regression[M]. John Wiley & Sons, 2004. [↑](#endnote-ref-42)
42. [] 李航. 统计学习方法[J]. 2012. [↑](#endnote-ref-43)
43. [] Breiman L. Random forests[J]. Machine learning, 2001, 45(1): 5-32. [↑](#endnote-ref-44)
44. [] 家会臣, 靳竹萱, 李济洪. Logistic 模型选择中三种交叉验证策略的比较[J]. 太原师范学院学报: 自然科学版, 2012, 11(1): 87-90. [↑](#endnote-ref-45)
45. [] Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection[C]//Ijcai. 1995, 14(2): 1137-1145. [↑](#endnote-ref-46)
46. [] Fawcett T. An introduction to ROC analysis[J]. Pattern recognition letters, 2006, 27(8): 861-874. [↑](#endnote-ref-47)
47. [] Zaitsev P, Tkachenko V. High Performance MySQL: Optimization, Backups, and Replication[M]. O'Reilly Media, 2012. [↑](#endnote-ref-48)
48. [] 郭志懋, 周傲英. 数据质量和数据清洗研究综述[J]. 软件学报, 2002, 13(11): 2076-2082. [↑](#endnote-ref-49)
49. [] Team R C. R Language Definition[J]. 2000. [↑](#endnote-ref-50)
50. [] Kleinbaum D G, Klein M. Survival analysis[M]. New York: Springer-Verlag, 1996. [↑](#endnote-ref-51)
51. [] Kaplan E L, Meier P. Nonparametric estimation from incomplete observations[J]. Journal of the American statistical association, 1958, 53(282): 457-481. [↑](#endnote-ref-52)