医学数据挖掘

案例与实践

华琳　李林
主编

夏翊　郑卫英　安立
刘薇　信中
副主编

清华大学出版社

医学数据挖掘
案例与实践

华琳　李林
主编

夏翊　郑卫英　安立
刘薇　信中
副主编

清华大学出版社
北京

内 容 简 介

基于大数据时代生物医学数据的爆炸式增长，本书从医学科研中的实际问题出发，以案例的形式深入浅出地介绍了近年来崭新的医学数据挖掘技术，包括决策树模型、支持向量机、随机森林分类、关联规则、贝叶斯网络构建等，并详细介绍了数据挖掘软件(SPSS、SAS、R 等)的操作步骤，重点突出实用性和可操作性，以期提高读者对医学科研数据的深层次处理与分析的能力。

本书主要取材于编者近年来从事生物医学数据深度挖掘方面的研究与教学工作内容，既适用于医学院校本科生及研究生、医学基础及临床科研工作者和相关技术人员作为教材，也可作为科学研究的参考用书。

图书在版编目（CIP）数据

医学数据挖掘案例与实践 /华琳等主编. —北京：清华大学出版社，2016
ISBN 978-7-302-44188-5

Ⅰ. ①医… Ⅱ. ①华… Ⅲ. ①数据处理－计算机应用－医学 Ⅳ. ①R-39

中国版本图书馆 CIP 数据核字（2016）第 148620 号

责任编辑：龙启铭
封面设计：傅瑞学
责任校对：梁　毅
责任印制：宋　林

出版发行：清华大学出版社
网　　址：http://www.tup.com.cn, http://www.wqbook.com
地　　址：北京清华大学学研大厦 A 座　　邮　　编：100084
社 总 机：010-62770175　　邮　　购：010-62786544
投稿与读者服务：010-62776969，c-service@tup.tsinghua.edu.cn
质量反馈：010-62772015，zhiliang@tup.tsinghua.edu.cn
印 装 者：北京国马印刷厂
经　　销：全国新华书店
开　　本：185mm×230mm　　**印　　张**：12.5　　**字　　数**：273 千字
版　　次：2016 年 9 月第 1 版　　**印　　次**：2016 年 9 月第 1 次印刷
印　　数：1～2000
定　　价：30.00 元

产品编号：069958-01

编写人员名单

主编：

华　琳	首都医科大学
李　林	首都医科大学

副主编（按拼音先后顺序）：

安　立	北京朝阳医院
刘　薇	北京同仁医院
夏　翊	首都医科大学
信　中	北京同仁医院
郑卫英	首都医科大学

序

众所周知，大数据时代已经到来，且已涉及到了各行各业。大数据时代对生物医学领域也产生了巨大的影响。随着信息技术的飞速发展，公共卫生信息系统、远程医疗、移动医疗的广泛建立与使用、以及个性化医疗设备、健康可穿戴设备的开发，生物医学领域进入了海量数据时代。生物医学大数据广泛涉及人类健康相关的各个领域：临床医疗、公共卫生、医药研发、医疗市场与费用、个体行为与情绪、人类遗传学与组织学、社会人口学、环境、健康网络与媒体数据。

生物医学技术的发展离不开大数据，大数据也成为生物医学领域的最新驱动力。在临床实践及科学研究的过程中，大数据的利用、开发和整理，可以为临床及科研提取更多更有价值的信息，甚至可以颠覆以往的一些研究结果。生物医学大数据将改变医学实践模式，改善医药卫生服务质量，最终有利于实现个体化治疗和群体性预防。但是，生物医学大数据区别于其他领域的大数据特征，它具有多维性（文本、数字、影像、基因等多种数据）、异质性和时空动态性等特征，这些都为大数据分析工作带来了一定的难度。因此，大数据分析技术和方法对大数据在生物医学领域的应用具有重要意义。

本书主编华琳副教授多年从事生物医学统计与数据深度挖掘方面的研究与教学工作，具有丰富的理论知识和实践经验。她擅长统计研究设计、各类型数据统计分析、数据挖掘及生物信息学分析，近5年内在相关领域以第一作者和通讯作者发表论文30余篇，其中SCI论文近20篇。她为人朴实、待人诚恳、学风正派、实事求是。我很赞赏本书以“知识应用化”为主线、以丰富的实际案例作解释的组织结构、重点突出实用性和可操作性。

本书从医学临床及科研中的实际问题出发，以案例的形式深入浅出地介绍了近年来崭新的医学数据挖掘技术。相信本书既会是高等学校相关专业的一本好教材，也可作为医学基础及临床科研工作者和相关技术人员的一部好参考书，能提高读者对医学科研数据的深层次处理与分析的能力。

北京大学泌尿外科研究所副所长

前言

背景：随着现代医学及技术的迅速发展，生物医学海量数据日益激增，医学工作者将面临如何对来自临床实践及基础实验研究中的数据进行正确统计分析并提取特征信息，从而用于临床研究、诊断和治疗。当前，对于大部分医学生及医学工作者，基础统计分析及SPSS软件实现已经普及，对于从临床实践及基础实验研究中所获取的数据进行基本统计分析的过程也已经掌握。但是，基础统计分析提取的信息较为有限，距研究者的期望相去甚远。许多研究实践证实采用高级统计分析方法以及高维数据的数据挖掘技术可以提取更多更有价值的信息，相应的研究结果被业界广泛认同。但是，当前医学生及医学工作者对这些数据分析知识缺乏了解，这一方面造成了大量宝贵的基础实验数据、临床研究数据处于闲置无用的状态，另一方面也无助于医学知识的有效积累、无助于提升临床诊治的精准度。近年来，作者一直致力于生物医学数据深度挖掘方面的研究，同时为众多医学研究生、青年医师、医学工作者提供高级统计分析服务。在研究工作与医学研究生及医学工作者的接触过程中，我们深深体会到目前临床与基础医学研究中的海量数据存在着进一步深度挖掘的空间，因此我们一直怀有将近二十年来的研究与教学工作的经验与大家分享的激情。

目的：为了提高广大医学生和医学工作者对于数据深度分析及数据挖掘的能力，本书介绍一些高级统计分析方法，如多元回归分析、主成分分析、生存分析、Meta分析，以及针对高维数据的数据挖掘技术，如决策树模型、支持向量机、随机森林分类、关联规则等；除了重点介绍近年来崭新的医学数据分析方法及数据挖掘技术外，还重点描述、示范这些方法的软件实现的具体内容。提升读者数据处理与分析能力是我们的最终目的。

特色：

（1）本书内容覆盖面广：内容涉及常见的复杂医学数据分析及数据挖掘，如主成分分析、生存分析、倾向性匹配、决策树模型、支持向量机和关联规则分析等，从而宽展了读者的群体面；

（2）案例典型，完备翔实：以疾病或系统为主线，采用案例式编排，深入浅出地帮助读者熟悉复杂数据分析处理的各种方法，增强可读性与针对性；

（3）软件实现具体细致：展示分析解决的流程及相应的软件操作实现，避免大量的公式及烦琐的计算，提高实用性与可操作性；

（4）作者研究经验结晶：作者长期从事医学数据处理方面的相关工作，积累了丰富的基础临床医学数据处理经验。

读者对象：本书适合于医学院校本科生及研究生、医学基础科研及临床科研工作者及相关技术人员作为教材或作为科学研究的参考资料使用。

本书主要取材于编者近年来从事生物医学数据深度挖掘方面的研究与教学工作内容，部分数据资料在与临床医生探讨后进行了改编。在此我们向曾经与我们分享研究过程的医生朋友们表示由衷的感谢，向参与过讨论的研究生表示感谢。我们还要感谢北京市自然科学基金项目（No. 7142015）的支持。

本书的编写之初和编写过程中得到了北京同仁医院刘薇主任医师、信中副主任医师和北京朝阳医院安立副主任医师的鼓励和全力支持，借此机会我们对他们表示深深的敬意和感谢。同时也感谢首都医科大学生物医学工程学院领导的大力支持。

本书虽然由编写动机到成书历经时间不短，但全新取材工作量巨大，具体编写时间有限，加之编者水平和所涉猎范围所限，书中不足和缺陷在所难免。希望得到专家、同行和读者的批评指正，使本书得到不断完善。

编者

2016 年 6 月

目录

第1章

数据预处理

在进行数据分析工作之前，一般需要对数据作必要的处理，这称之为数据预处理。数据预处理工作在多数情况下是十分必要的。在数据整理的过程中，数据中的异常值和缺失值比较常见，虽然这是数据采集人员和统计工作人员最不愿意见到的，但又是无法完全避免的情况。

异常值一方面可以根据专业知识判别，比如血压值接近于零是明显的不合理数据。此外，数值过度偏离均值可能就是异常值，如果不做处理会对最终结果造成影响。

缺失值产生的原因有很多，包括主观原因（如主观失误，历史局限等），以及客观原因（如失访、实验仪器失效等）。当缺失值总数较小时，大多数统计方法都会采取将缺失值直接删除的做法，此时对最终的分析结果影响不大。但是，当缺失值数量较大时，简单地删除缺失值会丢失大量的数据信息，基于此种做法有可能会得到错误的结论。本章简单介绍处理数据中异常值和缺失值的常用方法。

1.1 异常值的常见处理方法

先通过简单的数据分析，了解预先要处理的数据分布特性，从而发现数据的异常情况。反映数据的集中趋势指标主要有均值（mean）、中位数（median）和众数（mode）等。反映数据离散趋势的指标主要有方差（variance）、标准差（standard deviation）、极差（range）和四分位数间距（quartile interval，即第 75 位百分位数与第 25 位百分位数的差值）等。通常也常用一些图形反映数据的分布状况，从而发现异常的数据。

例如，一组数据有 15 个数值：100，110，110，120，131，132，130，121，124，132，132，135，192，122，125。可以通过 SPSS 软件中的描述性统计，获得数据的分布特性（如计算平均值、中位数、标准差、方差等），结果如图 1.1 所示。

一般来说，如果一个数据大于 Q3+1.5IQR（其中 Q3 表示第三四分位数，即第 75 位百分位数，IQR（Inter-Quartile Range）表示四分位数间距）或者小于 Q1-1.5IQR（其中 Q1 为

Statistics

VAR00001

N	Valid	15
	Missing	0
Mean		127.7333
Median		125.0000
Mode		132.00
Std. Deviation		20.37318
Variance		415.067
Range		92.00
Minimum		100.00
Maximum		192.00
Percentiles	25	120.0000
	50	125.0000
	75	132.0000

图 1.1 数据的分布特性

第一四分位数，即第 25 位百分位数），就认为该数据为异常值。可以通过做箱式图（box plot）发现数据中的异常值。

根据上面的数据做出的箱式图，如图 1.2 所示，其中粗线表示的是中位数，箱子的高度表示的是四分位数间距，圆圈和小星星表示的是异常值。该数据中有两个异常值，分别为第 1 个值 100 和第 13 个值 192，其中 100<Q1−1.5IQR=120−1.5×(132−120)=102，192> Q3+1.5IQR=132+1.5×(132−120)=150，这两个异常值可以考虑在后期的统计中进行剔除。

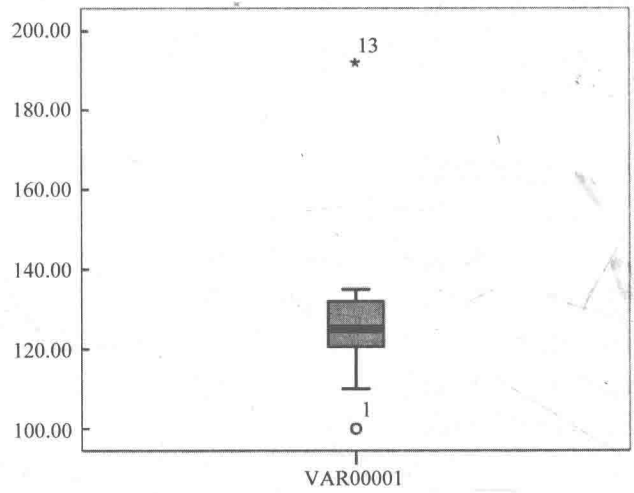

图 1.2 数据的箱式图

此外，对于多变量统计数据的异常值识别，常用的检验思路是观察各样本点到样本中心的距离。如果某些样本点到样本中心的距离太大，就可以判断为异常值。这里距离的度

量一般使用马氏距离（Mahalanobis Distance）。因为马氏距离不受量纲的影响，而且在多元条件下，马氏距离还考虑了变量之间的相关性，这使得它优于欧氏距离。考虑到由于个别异常值会导致均值向量和协方差矩阵出现巨大偏差，这样计算出来的马氏距离起不到检测异常值的作用，从而导致传统的马氏距离检测方法不稳定，因此需要利用迭代的思想构造一个稳健的均值和协方差矩阵估计量，然后计算稳健马氏距离（Robust Mahalanobis Distance），从而使得异常值能够正确地被识别出来。

在 R 软件的 mvoutlier 软件包中提供了基于稳健马氏距离的异常值检验方法。在介绍多变量数据的异常值判断前，先简单介绍 R 软件及其软件包。

R 软件是一款免费的共享统计软件，它提供了若干的统计程序包，以及集成的统计工具和各种数学计算。R 软件有 Linux、MacOS X 和 Windows 三种版本。用户可以在 R 网站上免费下载。例如，想下载 R 软件的 Windows 版本，可以从 http://cran.r-project.org/bin/windows/base/网址中下载。当前 R 软件的 Windows 最高版本为 R-3.2.3。不同的版本所携带的软件包也不同。R 软件大概每 3 个月更新一次版本。

相比于其他软件环境，R 软件有很多优点，如 R 软件的源代码公开，任何人都可以查看 R 的分析过程，最新的统计方法都是在 R 中实现，具有让人眩目的数据可视化功能，是唯一在所有平台上均可运行的数据分析和挖掘环境等。

本书以 R-2.12.2 Windows 版本为例，下载完成后双击 R-2.12.2-win.exe 按步骤进行安装，安装完成后双击桌面上的图标 R.12.2，就会出现如图 1.3 所示的界面。

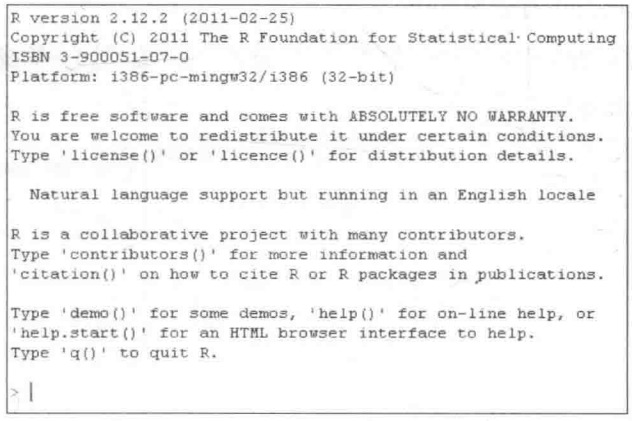

图 1.3　R 软件界面图

在该界面中出现的>符号是软件自动出现的，用户不用输入这个符号，只需要直接在后面输入命令。R 软件中附带了很多软件包，这些软件包是函数、数据集和文件的组合。所有软件包都可以通过 CRAN 的网站（https://cran.r-project.org/）下载获得。用户也可以自动联网安装这些软件包。

下面来安装判断多变量数据异常值的 mvoutlier 软件包。选择菜单 Packages→Install package（s），此时要求选择 CRAN 镜像，这里选择 China（Beijing 2），如图 1.4 所示。

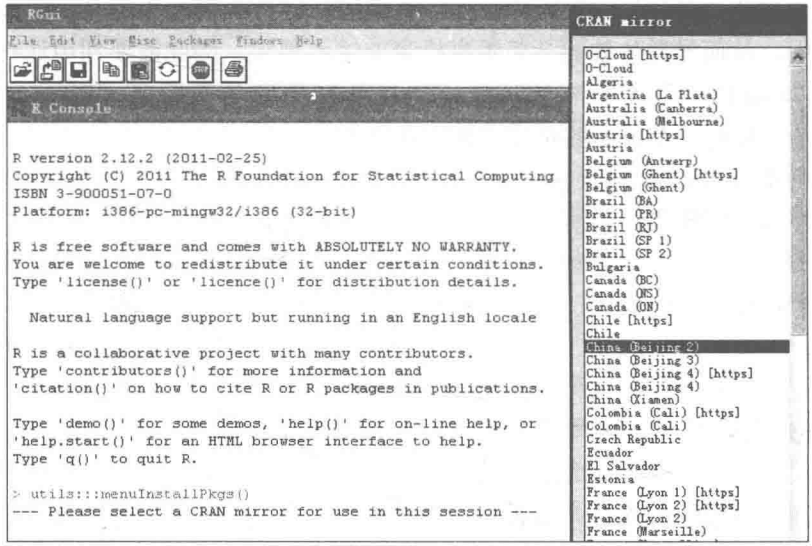

图 1.4 安装 R 软件包菜单

在弹出的 R 自带的软件包中选择 mvoutlier，然后进行安装，如图 1.5 所示，单击 Packages 对话框中的【OK】按钮进行安装即可。

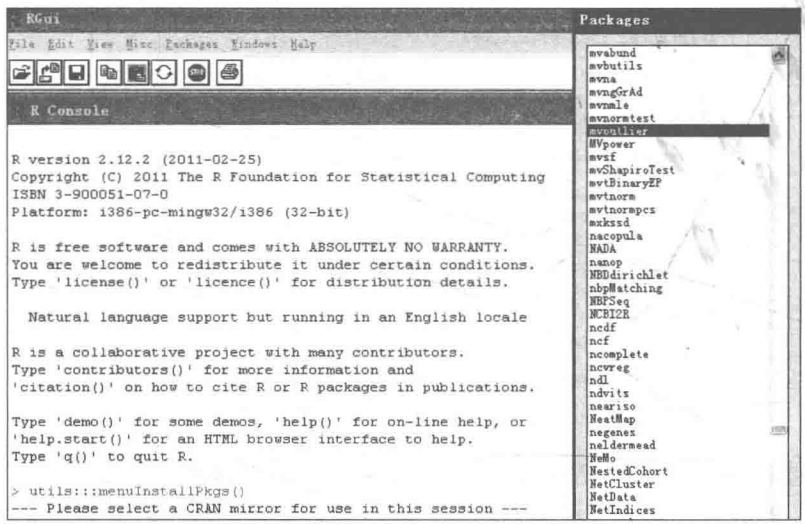

图 1.5 安装 mvoutlier 软件包界面