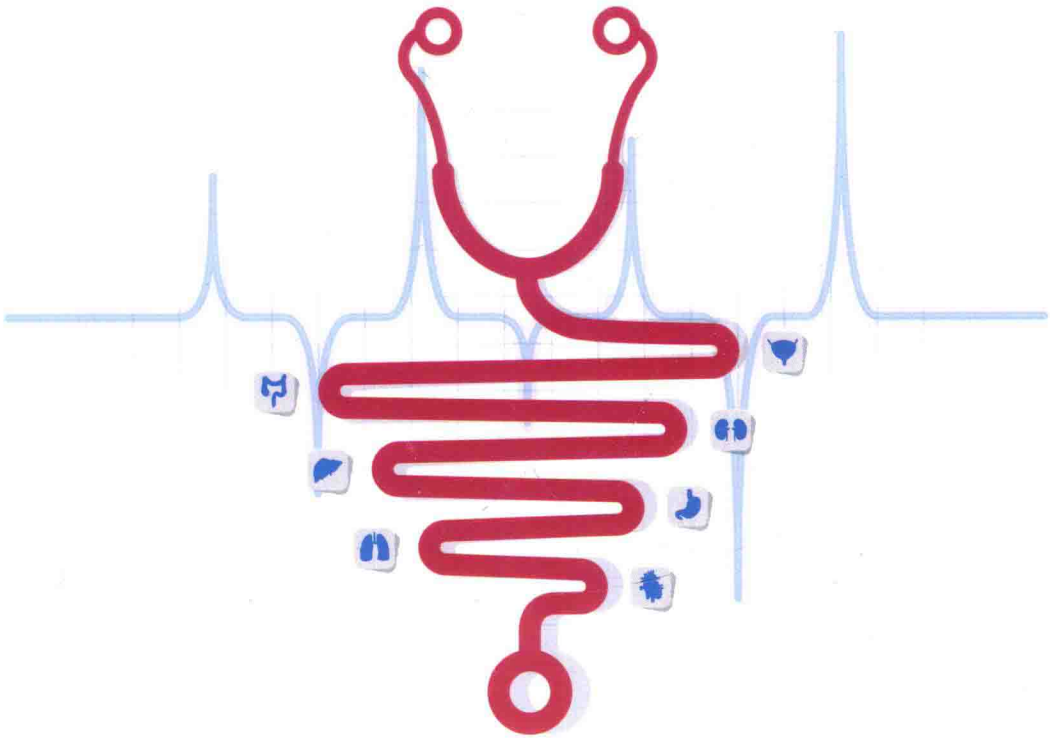
CDA数据分析师系列丛书

Broadview
www.broadview.com.cn

采用大量案例与实证
重点探讨数据挖掘技术如何与临床医学深度融合



医疗革命

医学数据挖掘的理论与实践

邵学杰 / 著

中国工信出版集团

電子工業出版社
PUBLISHING HOUSE OF ELECTRONICS INDUSTRY
http://www.phei.com.cn

CDA数据分析师系列丛书

医疗革命

医学数据挖掘的理论与实践

邵学杰 / 著

電子工業出版社
Publishing House of Electronics Industry
北京•BEIJING

内 容 简 介

本书以数据挖掘与模式识别的七大原理在临床医学中的运用案例为切入点，系统而全面地介绍了医学数据挖掘的基本方法与原理，对数据分析的常用算法进行了通俗易懂的讲解。本书最大的特色是采用了案例分析与实证的方法，每一个原理、算法都在案例讲解中生动地体现出来。更重要的是，本书对临床医学的数据挖掘与模式识别技术进行了开创性、系统性的讨论，用案例展现了数据挖掘技术如何与临床医学相结合，为广大的医生、医学数据挖掘工作者提供了很实用的技术示范、理念导入、系统思考。

本书所有概念的讲解基本结构为原理讲解与案例实操的二元结构，兼顾初学者与专业人士的需要。本书重点探讨了数据挖掘技术如何与临床医学深度融合，如何运用现代的数据挖掘理念、模式识别与机器学习的基本方法解决临床科研中的应用问题，为广大的科研型临床医生提供助力，为广大的数据分析人员找到行业应用的范例，为广大初学者提供努力学习的方向；更重要的是在这个大数据时代，我们可以亲自见证数据技术是如何改变并深刻影响着临床医学的科研与教学的。

图书在版编目（CIP）数据

医疗革命：医学数据挖掘的理论与实践 / 邵学杰著. —北京：电子工业出版社，2016.9
（CDA 数据分析师系列丛书）
ISBN 978-7-121-29867-7

Ⅰ. ①医… Ⅱ. ①邵… Ⅲ. ①医学－数据采集－研究 Ⅳ. ①R-39

中国版本图书馆 CIP 数据核字(2016)第 211991 号

策划编辑：石　倩
责任编辑：石　倩
印　　刷：三河市双峰印刷装订有限公司
装　　订：三河市双峰印刷装订有限公司
出版发行：电子工业出版社
　　北京市海淀区万寿路 173 信箱　　邮编：100036
开　　本：787×980　1/16　印张：12　字数：308 千字
版　　次：2016 年 9 月第 1 版
印　　次：2016 年 9 月第 1 次印刷
定　　价：49.00 元

凡所购买电子工业出版社图书有缺损问题，请向购买书店调换。若书店售缺，请与本社发行部联系，联系及邮购电话：（010）88254888，88258888。

质量投诉请发邮件至 zlts@phei.com.cn，盗版侵权举报请发邮件至 dbqq@phei.com.cn。

本书咨询联系方式：010-51260888-819　faq@phei.com.cn。

序

Big Data（大数据）在这几年突然火红于日常生活的各项领域中，连临床医疗也不例外，其实早期就存在许多通过数据来佐证或者分析预测结果的例子，但是当时在大多数的情况之下，统计运算不够快速成为了资料分析的一大限制，因此大多数资料是被临床研究工作者们放在一边而从未思考该如何运用的。伴随着信息科技的进步以及发达，能为我们所分析的数据将呈现爆炸性的成长，因此人们能从数据中学习的知识会更加丰富。和其他科学领域相比，需要透过大量临床试验的医学领域算是进步较缓慢的学科。管仲曾说过："不明于数欲举大事，如舟之无楫而欲行于大海也。"意思是说在不清楚相关数据的情况下想做大事，无疑是没有桨的船想航行于汪洋大海中一样。也就是说，在医疗大数据的时代下，"dry lab"的医疗数据研究将会是协助医学领域快速进步的一大重要关键。本书通过大量临床医学的实例，由浅入深地介绍各项数据分析以及数据挖掘的方法和工具，将大量的临床医学数据化繁为简。相信无论是在校的学生或是临床研究者，本书都将会是学习或科研路上不可或缺的好伙伴。

谢邦昌
台北医学大学管理学院及大数据研究中心　院长/主任
中华市场研究协会理事长
中华资料采矿协会荣誉理事长

前　言

在医学大数据时代，数据技术带来了临床医学科研的革命性进步。本书通过对医疗数据挖掘的基本理论的阐述，将现代统计学与数据挖掘技术有机结合，讲述了大量的医学数据挖掘的案例，提供了大量的医学数据挖掘的实操方法。医学数据模式识别的七大原理与案例讲解是本书具有独创性的对医学数据技术的全面概括与总结，七大原理的首次提出也是医学数据挖掘技术上升到系统理论的重要实践与创新。无论是预测性建模、解释性建模、知识性建模与描述性建模，抑或是序列模式建模、依赖关系建模、异常模式建模，模式识别的类型规律跃然纸上，为专业人士或初学者厘清了数据挖掘与模式识别的基本类型特征。

不仅如此，本书选取的大量的医学数据挖掘案例为本书的实用性增加了学以致用的特色，凡认真阅读本书的读者都会从理论与实操两个层面全面、系统、实用地了解医学数据挖掘的原理与方法。本书以胰腺癌与二型糖尿病的关联规则、乳腺癌图片智能识别的挖掘算法、心电信号大数据的人工智能识别、低位前切保肛术的荟萃分析、贝叶斯网络预测高血压患者心血管风险、基线静息心率评估心血管事件、老年肺癌研究的荟萃分析等实用数据技术为切入点，使初学者能够掌握医学数据挖掘的基本理论与方法，因此是一本很好的入门级教科书。

对于资深的临床医生、医学博士、论文写作者而言，本书也是一本很好的案例参考书。特别是对于医学科研课题而言，本书提供了强大的实际操作技术培训与案例讲解，从顶级的国际期刊《自然》、《细胞》、《柳叶刀》等杂志选取经典的数据分析案例，用生动的方法让读者可以学到医学论文中数据、图表、算法的实际使用方法；因而对于专业人员而言，本书又是一本很好的资深级别的专业用书。

我们相信，无论您是初学者还是资深的专业人士，本书都将为您提供极大的可读性、趣味性和科学性。

目　录

第 1 章　数据分析与数据挖掘的力量 1

1.1　葡萄牙医生解决世界新生儿出生缺陷的故事 2

1.2　医学数据挖掘的主要定义 5

1.2.1　数据挖掘的定义 5

1.2.2　医学数据挖掘的故事 5

1.3　医学数据模式识别的七大原理与案例讲解 6

1.3.1　什么是模式识别 6

1.3.2　7 个小故事 7

1.4　临床医学领域的机器学习与人工智能 12

1.5　神经元网络的基本原理 13

第 2 章　临床医学的数据挖掘 20

2.1　房颤与肾功能关联现象的故事 21

2.2　支持向量机的算法原理与应用 30

2.2.1　一个故事的开场白 30

2.2.2　支持向量机的主要特点 31

2.2.3　支持向量机的应用案例 39

2.3　疾病规律与统计学革命 43

2.3.1　肝胆外科的统计学故事 43

2.3.2　双盲实验的诞生 44

2.3.3　几则很有趣的医学统计学故事 47

2.4　老年肺癌研究 50

2.4.1　数据的抓取与来源 50

2.4.2　癌症与老龄化的相关性分析 51

2.4.3　老年人肺癌手术适用性评估关键词频率 53

2.4.4　老年肺肿瘤的数据分析 54

2.4.5　英国肺癌患者 38 年来死亡率研究 59

2.4.6　老龄肺癌死亡率数据的三维分析 59

2.5 临床医学与数据挖掘的边缘学科 62
2.5.1 几个实例 62
2.5.2 医学统计学与医学数据挖掘的区别 69
2.5.3 有关数据挖掘是边缘学科的几个实例 72
2.5.4 一个医学数据挖掘的案例 74

第 3 章 临床医学与数据技术的深度融合 90

3.1 二型糖尿病与胰腺癌的故事 91
3.2 Cox 回归的基本原理与应用 94
3.2.1 Cox 回归的基本原理 94
3.2.2 晚期肺癌伴脑转移患者的预后多因素 Cox 回归 95
3.2.3 本案例的几点启示 100
3.3 医学数据分析中的故事 101
3.4 聚类的临床医学意义 103
3.4.1 聚类算法的基本定义 103
3.4.2 临床医学数据挖掘中聚类的意义 104
3.4.3 案例 112
3.5 贝叶斯算法的应用案例 113
3.5.1 一个流传甚广的故事 113
3.5.2 一个贝叶斯算法的医学案例 114

第 4 章 临床医学的模式识别 126

4.1 模式识别是什么 127
4.1.1 定义 127
4.1.2 临床医学模式识别的故事 127
4.2 基线静息心率的故事 130
4.3 决策树算法 132
4.4 最大期望（EM）算法 135
4.5 算法的规律与临床医学的本质 140
4.5.1 算法的本质是什么 140
4.5.2 数据挖掘中医学的本质 141

第 5 章 医学数据挖掘的常用工具 146

5.1 SAS 挖掘软件运用案例 147
5.2 Weka 软件介绍 150
5.3 Matlab 案例 152

5.4 R 语言案例 162
5.5 临床医生如何用好挖掘工具 164

第 6 章 专业级医学 SCI 论文中的统计工具 169

6.1 医学数据中的 *T* 值与 *P* 值故事 170
6.2 K 线图的故事 172
6.3 国际顶级期刊上的数据技术 174
6.4 SCI 荟萃分析中的统计学工具 180
6.4.1 研究对象及入选标准 181
6.4.2 统计学处理 181

第1章

数据分析与数据挖掘的力量

- 葡萄牙医生解决世界新生儿出生缺陷的故事
- 医学数据挖掘的主要定义
- 医学数据模式识别的七大原理与案例讲解
- 临床医学领域的机器学习与人工智能
- 多层神经元网络算法的基本原理

1.1 葡萄牙医生解决世界新生儿出生缺陷的故事

每年，全球大约有数以百万计的新生儿缺陷患者，原因包括遗传的、环境的、病毒性的，其中有高达 25%以上的新生儿先天缺陷找不到明确的原因。虽然超声医学、分子遗传检测技术已经有长足的进步，但依然有 8%左右的新生儿先天缺陷在世界某些地区找不到原因，葡萄牙医生用数据挖掘方法的解决方案对我们很有启发。

葡萄牙医生首先以全球各诊所新生儿出生记录数据为基础，包括出生年月日、性别、家庭住址三项基础统计数据，然后用空间地理信息做匹配关联分析，就是分析出生婴儿与空间地理位置的关联性，结果如图 1-1 和图 1-2 所示。

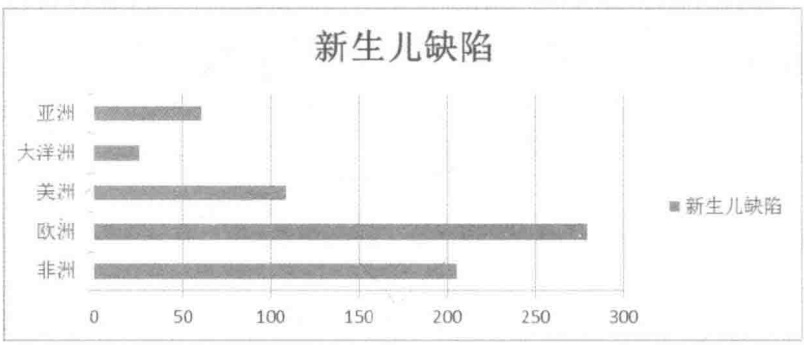


图 1-1　各大洲新生儿缺陷抽样分布数

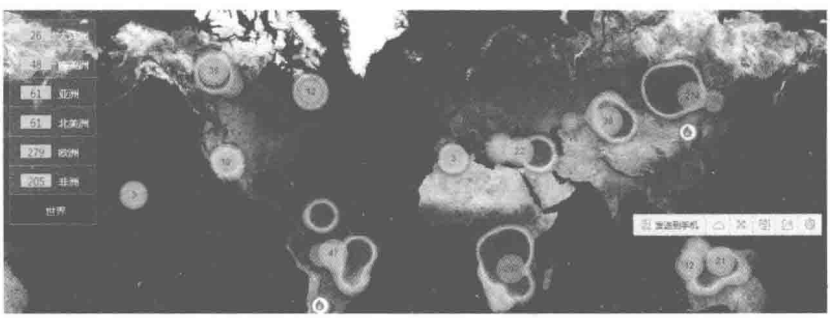


图 1-2　各大洲新生儿缺陷分布示意图

资料来源：葡萄牙医生 2014 年全球各大洲新生儿缺陷抽样调查报告

这项研究采用了最简单的单变量模型：变量是缺陷新生儿的出生地址，方法是采用全球抽样的均匀分布法，确保抽样数据的代表性。

抽样方法的正确性是指抽样的代表性和随机性，代表性反映样本与批质量的接近程度，而随机性反映检查批中单位产品被抽入样本纯属偶然，即由随机因素所决定。在对总体质量状况一无所知

的情况下，显然不能以主观的限制条件去提高抽样的代表性，抽样应当是完全随机的，这时采用简单随机抽样最为合理。在对总体质量构成有所了解的情况下，可以采用分层随机或系统随机抽样来提高抽样的代表性。在采用简单随机抽样有困难的情况下，可以采用代表性和随机性较差的分段随机抽样或整群随机抽样。这些抽样方法除简单随机抽样外，都是带有主观限制条件的随机抽样法。通常只要不是有意识地抽取质量好或坏的产品，尽量从批的各部分抽样，都可以近似地认为是随机抽样。

1. 单纯随机抽样（simple random sampling）

将调查总体全部观察单位编号，再用抽签法或随机数字表随机抽取部分观察单位组成样本。

优点：操作简单，均数、率及相应的标准误计算简单。

缺点：总体较大时，难以一一编号。

2. 系统抽样（systematic sampling）

该方法又称机械抽样、等距抽样，即先将总体的观察单位按某一顺序号分成 n 个部分，再从第一部分随机抽取第 k 号观察单位，依次用相等间距，从每一部分各抽取一个观察单位组成样本。

优点：易于理解、简便易行。

缺点：总体有周期或增减趋势时，易产生偏性。

3. 整群抽样（cluster sampling）

总体分群，再随机抽取几个群组成样本，群内全部调查。

优点：便于组织、节省经费。

缺点：抽样误差大于单纯随机抽样。

4. 分层抽样（stratified sampling）

先按对观察指标影响较大的某种特征，将总体分为若干类别；再从每一层内随机抽取一定数量的观察单位，合起来组成样本。有按比例分配和最优分配两种方案。

优点：样本代表性好，抽样误差减少。

以上四种基本抽样方法都属单阶段抽样，实际应用中常根据实际情况将整个抽样过程分为若干阶段来进行，称为多阶段抽样。

葡萄牙医生在本故事中采用了分群与分层抽样调查相结合的方法，按五大洲分群抽取，每个洲又按历史高发地区分层抽取。整群的聚类（cluster）是数据挖掘技术上一个很重要的概念，把某维度属性相近的实例聚类是数据技术最基础的方法；聚类后，距离太远的数据就是异常值。对数据处理的常规方法第一步就是聚类，把某些属性相近似的数据聚类后就可以进一步分析它们之间的关系，数据的聚类可以做回归（预测），数据的离散可以做预警（异常值）。

如图 1-3 所示，数据之间的关系可以从图形上表示出来，因此数据挖掘完全可以可视化地表现出来。就是说数据之间是有空间分布关系与距离的，用空间分布关系来表示数与数之间的关系，是现代数学的重要特征。

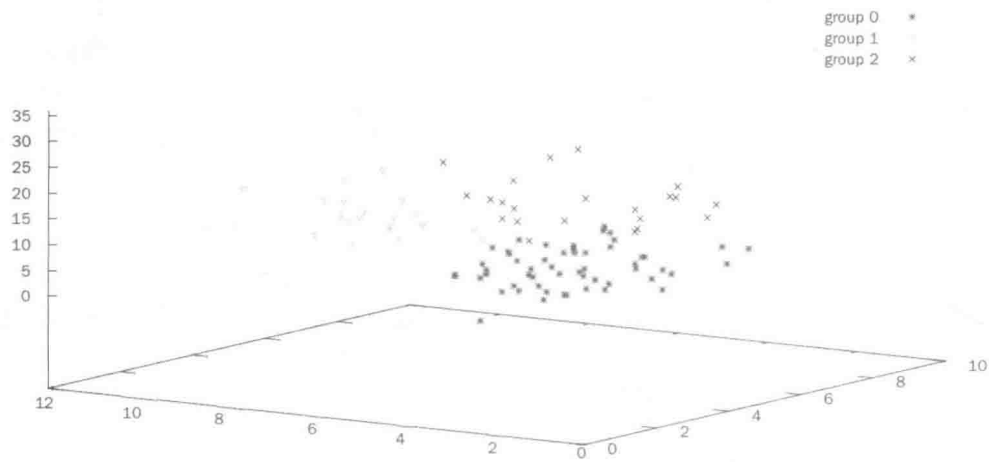


图 1-3　数据聚类图示效果

本故事中，葡萄牙医生的重要发现是：

① 欧洲大量新移民聚类中产生了新生儿缺陷高发的现象，这一数据甚至超过了传统落后地区非洲的新生儿出生缺陷率。

② 伊拉克战争、叙利亚战争、也门内战导致的难民大量涌入欧洲，人口的大规模迁徙改变了欧洲的新生儿人口健康状况。

就这样，葡萄牙医生用了一个简单的变量（婴儿出生地），代入了一个简单的分析框架——空间地理坐标与新生儿缺陷的关联性，用抽样方法获取数据，最后导出了近年来欧洲新生儿缺陷增加的主要原因：大规模移民难民潮。其中一个典型调查发现西班牙边境地区一个废弃的化学工厂是外来移民长期居住后新生儿缺陷发生的重要原因。

这是一个用数据进行知识发现（Knowledge-Discovery in Databases，KDD）的故事也是一个典型的流行病监测模型。数据库知识发现是数据挖掘最核心的意义。计算机时代，大量的数据被存放在数据库中，而不管是关系型数据库还是非关系型数据库，大量数据存储的成本都非常高昂；尤其在中国的三甲医院中，每天都有大量的门诊与住院数据产生，其中 80%是图像数据。一个普通的三甲医院每年产生大约 15TB～20TB 的新数据，这些数据中包含着许许多多疾病的新规律与知识发现，而用传统的统计学方法，用传统的手工或计算机方法已经无法处理或者无法准确地处理。这就是现代大数据技术产生的背景，包括传统的统计学、计算机技术、最优化分析技术、机器学习与人工智能、在线分析与检索技术。

1.2 医学数据挖掘的主要定义

1.2.1 数据挖掘的定义

数据挖掘(Data mining),又译为资料探勘、数据采矿。它是数据库知识发现(Knowledge-Discovery in Databases，KDD) 中的一个步骤。数据挖掘一般是指从大量的数据中通过算法搜索隐藏于其中信息的过程。数据挖掘通常与计算机科学有关，并通过统计、在线分析处理、情报检索、机器学习、专家系统 (依靠过去的经验法则) 和模式识别等诸多方法来实现上述目标。

数据挖掘利用了来自如下一些领域的思想：① 来自统计学的抽样、估计和假设检验；② 人工智能、模式识别和机器学习的搜索算法、建模技术和学习理论。数据挖掘也迅速地接纳了来自其他领域的思想，这些领域包括最优化、进化计算、信息论、信号处理、可视化和信息检索。一些其他领域也起到重要的支撑作用，尤其需要数据库系统提供有效的存储、索引和查询处理支持。源于高性能 (并行) 计算的技术在处理海量数据集方面常常是重要的。分布式技术也能帮助处理海量数据，并且当数据不能集中到一起处理时更是至关重要。

1.2.2 医学数据挖掘的故事

医学数据挖掘一般是指从大量的医学数据中通过算法搜索来认识隐藏于其中疾病新规律的过程。

今天这里要讲述一个关于肠道菌群与心血管疾病关联性的故事。在微生物学诞生后不久，人们就发现，在动物的消化道中存在有不少微生物。例如在牛、羊、兔等食草动物的胃或盲肠中，就存在大量以细菌为主的微生物群体。由于食草动物摄入的植食性饲料中，纤维素、半纤维素等多糖难以依靠动物体自身分泌的酶液消化，而微生物群体中包含的纤维素消化菌、半纤维素消化菌等可以较好地将多糖转化为低聚糖和寡糖，从而促进对这些营养物质的吸收。

随着医学的发展，人们也注意到，在人类的肠道，尤其是结肠 (也就是平常所说的大肠) 中，也存在着大量微生物。这些以细菌为主的微生物种类极多，数量极大。肠道菌群并非是生来就有的，它们实际上是“外来户”。在母体子宫内，胎儿所处的是一个几乎无菌的环境，因此胎儿肠道内也是无菌的。当胎儿出生之后的几天内，细菌通过分娩时阴道物质摄入、哺乳时的口腔摄入以及空气吸入等途径进入新生儿体内，并在肠道内定植，形成新生儿最初的肠道菌群。随着婴儿的成长，肠道菌群的种类结构逐渐趋于稳定，最终形成成熟的肠道菌群。这些微小的生物群体就这样不知不觉地定居到人体之内，悄无声息地与主人相随一生。

近期的多项研究表明，肠道菌群和人体的代谢疾病具有重要关系。肠道菌群失衡可能是造成肥胖、糖尿病等多种代谢异常的重要原因之一。造成代谢异常的主要原因，是失衡的肠道菌群产生的脂多糖等内毒素进入人体，被免疫细胞识别后产生多种炎症因子，使得机体进入低度炎症状态，从而产生代谢异常。例如，若长期进食高脂、高糖食物，可造成肠道菌群中条件致病菌比例增加，而

共生菌比例下降，从而使得食物中摄取的能量更容易转化为脂肪累积于皮下，造成肥胖。此外，低度炎症还能促使机体对胰岛素响应程度下降，造成胰岛素抵抗，进而发展为糖尿病。

这些医学观察的结论完全得益于数据挖掘技术的进步，医生们从医治经验中发现患有肠道疾病的人往往也同时患有心血管疾病。一开始医生们并没有注意到这个现象，当越来越多的病例记录了同一现象时，医生们开始怀疑两者之间的关联性。但是怀疑代替不了科学结论，需要定量化的数据支持，越来越多的病例数据汇总后经过关联规则算法最终找到了大量的支持病例，最终现代医学解开了这个秘密。肠道菌群与中风，原本风马牛不相及的两个病种终于确立了因果关系。

有意思的是，最新的医学数据挖掘表明，肠道菌群的数量分布居然与抑郁症有关联，医学科学家正在试图解开这个秘密。

这个故事生动地表达了医学数据挖掘的魅力与能量。利用大量的临床医学数据发现新的医学疾病规律正是数据挖掘在医学，特别是临床医学领域的巨大意义。

1.3 医学数据模式识别的七大原理与案例讲解

1.3.1 什么是模式识别

模式识别是指对表征事物或现象的各种形式的（数值的、文字的和逻辑关系的）信息进行处理和分析，以对事物或现象进行描述、辨认、分类和解释的过程，是信息科学和人工智能的重要组成部分。数据挖掘的本质就是模式识别。医学数据的七种模式识别方法分别是：

① 解释性数据建模；
② 描述性建模；
③ 预测性建模；
④ 知识性建模；
⑤ 序列模式建模；
⑥ 依赖关系的建模；
⑦ 异常与趋势建模。

建模就是建立模型，就是为了理解事物而对事物做出的一种抽象，是对事物的一种无歧义的书面描述。建立描述过程的性能的数学模型也称为建模，系统建模主要用于三个方面。①分析和设计实际系统。②预测或预报实际系统的某些状态的未来发展趋势，预测或预报基于事物发展过程的连贯性。③对系统实行最优控制。

数据挖掘中的建模，其本质就是模式识别的方法，包括数学定量描述与归因分析定性描述。医学模式识别就是利用临床医学大数据来建模，找到疾病之间的相互关系，无论是依赖关系，关联关系还是序列模式等关系都可以在数据中找到真相。

下面我们分别简述七个小故事来深入浅出地讲解这七个原理。

1.3.2 7 个小故事

1. 解释性数据建模

第一个小故事发生在朝鲜战争时期，第一、二次战役后以美国为首的联合国军遭遇了志愿军的重大打击，特别是长津湖一战，志愿军九兵团全线包围美骑一师，经过拼命抵抗美军好不容易才冲出重围。更有甚者，美军第八集团军司令沃克中将在给儿子的授勋途中车祸死亡，不得不由李奇微将军来接任。李奇微上任后的第一件事就是阅读美军的作战日志，他发现志愿军的每一次攻势都只有七天左右，他称之为“礼拜攻势”；经过深思熟虑，李奇微发现这是由于志愿军的补给只能维持七天所致。于是乎，美军依靠战场日志的“回放”发现了志愿军的弱点，美军制定了对付志愿军“礼拜攻势”的“磁性战术”，志愿军进攻时美军节节撤退避其锋芒，第七天开始发动反攻。这种战术给志愿军带来了很大的威胁，李奇微也成为志愿军很难对付的敌人。

这是一个典型的数据挖掘与模式识别案例。这里的数据记录就是美军的“战场日志”。通过对战场日志的梳理，李奇微发现了志愿军的攻击规律是七天一个周期，为了搞清楚原因，美军将领李奇微运用他的军事经验对“礼拜攻势”进行了很好的归因分析——志愿军是由于补给问题而导致的。这也是一个解释性建模的典型案例。

解释性建模是数据挖掘中的一个重要的模式识别方法。解释性建模的实质是模糊建模，模糊建模的概念由 Zadeh 提出后，在数据挖掘、模式识别、故障诊断、预测、监督与控制等方面得到了迅速的发展和应用，成为模糊理论与应用中重要的研究方向。模糊模型的特点在于它用模糊规则对知识进行表达，而且可以解决一些复杂的、非线性的、用传统的数学方法难以解决的问题。早期的模糊建模主要针对简单系统，采用总结专家经验的方式进行。

故事中的李奇微将军的战术就是典型的专家经验的合理运用。但是对于复杂系统，由于难以获得完备的专家经验，而数据相对容易获得，因此近年来基于数据的模糊建模成为研究的热点；但目前大多数研究将模糊模型作为一种函数逼近器，追求模糊模型对实际系统的拟合程度，即以模型的精确性为建模目标。

一个好的数学模型具备以下三点：① 描述性；② 预测性；③ 说明性。具体地说就是，一个好的数学模型能描述建模基于的系统，并且对其做出预测，同时能解释为什么这么建模以及建模得出的结论。针对以上三点，我们来看看数据和模型的区别。首先数据可以说是具有描述性，但仅是局部描述性，除非给出的数据能遍历每一种情况，而数学模型则具有全局描述性。其次，数据的预测性表现在可以通过数据建立模型，来给出预测结果。最后，好的数学模型能明确解释数据的走向，而光看数据你只能知道数据是怎么变化的，但不知道为什么这么变。建模和数据是相辅相成的，针对一个问题，建模是将其抽象到纯数学层面以寻求普适的解决方法与结论，数据是用来验证建模的结论，或者是辅助求解模型的（比如有些固定参数需要通过具体的实验或者观测数据来确定）。当然，只有用在好模型上，数据才会显得有意义。最后，如果数学建模真的因为大数据而没用了，也不会有那么多应用数学家还在探讨关于数学建模的问题。

2. 描述性建模

第二个小故事发生在微软公司的面试题中：如何在已知身高、体重、性别、年龄四个指标，但无法知道 18 个学生中任何一个人的照片及影像资料的条件下分析出这 18 个学生的身材数据?

描述性建模也是数学中常见的建模方法，其基本原理是从特殊到一般，即从分析事物的具体情况出发，经过数学语言的构建而得到一个可以具体描述事物特征的方法。描述性数学模型反映了从特殊到一般的认识过程，它是从分析客观事物的具体特征入手，经过逐步抽象而得到的。把客观事物中的关系概括于一个数学结构之中，是描述性数学模型的主要特征，也是解决问题的重要手段。

例如：乘法的交换律，如 3×4=4×3 和 15×3.5=3.5×15，我们可以用 a×b=b×a 来抽象表达，这就是一种描述性建模的思路。从特殊到一般的归纳与演绎，从具体特征到抽象表达，数学化的架构过程就是描述性建模的主要方法。

确立数与数之间的关系，可以采用代数方法或是几何方法，也可以采用代数加几何的解析几何方法。下面我们用坐标系的方法来直观地观察数与数的集合关系。数据如表 1-1 所示。

表 1-1　cop lot（height-weight-sex）不同性别下身高对于体重的散点

num	name	sex	age	height	weight
1	1 alice	f	13	56.5	84.0
2	2 becka	f	13	65.3	98.0
3	3 gail	f	14	64.3	90.0
4	4 karen	f	12	56.3	77.0
5	5 kathy	f	12	59.8	84.5
6	6 mary	f	15	66.5	112.0

attach(st)

如图 1-4 与图 1-5 所描述的那样，先以纵坐标为体重、横坐标为身高看看两者之间的关系。加入“性别”这个分组变量后，转换纵坐标为身高，横坐标为体重再次分析 18 个学生的数据。利用纵坐标、横坐标的变换来观察数据之间的关系是描述性建模的重要分析方法。描述性建模的主要方法就是从具体的特例中利用数据语言抽象概括出事物的特征：18 个同学中除去两个同学外（这两个同学一个瘦小，一个肥胖）其余同学身材匀称（身高与体重正相关）。

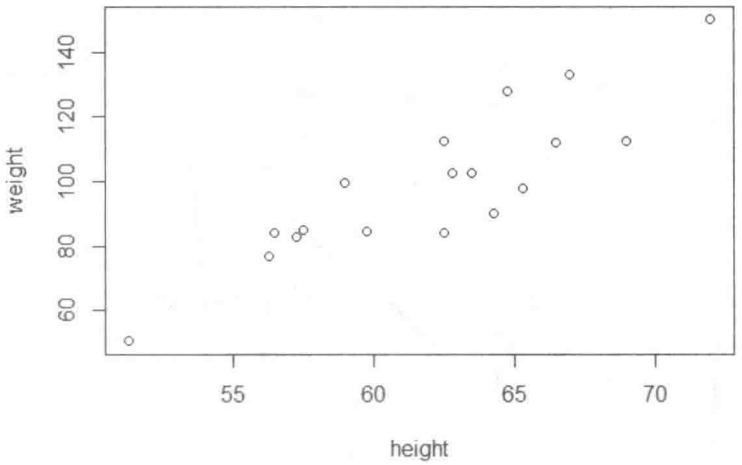


图 1-4　plot(height,weight)#身高对于体重的散点图

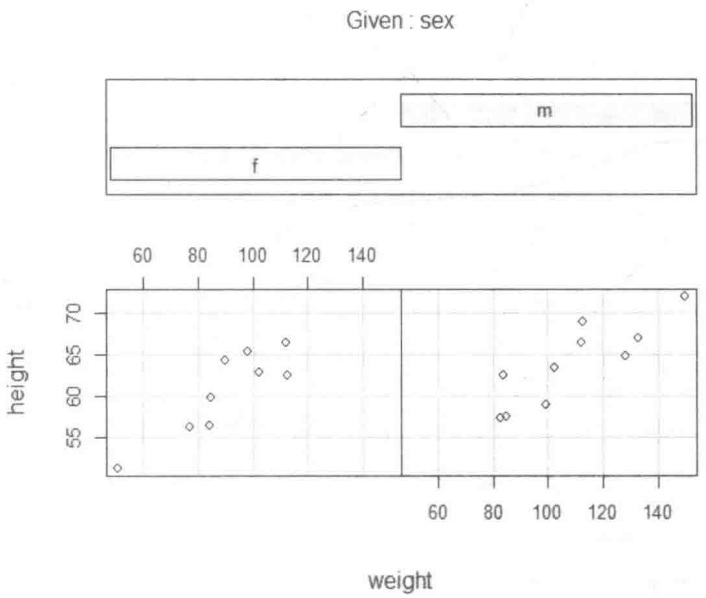


图 1-5　cop lot(height~weight|sex)#不同性别下身高对于体重的散点图

3. 预测性建模

第三个小故事是关于谷歌的大数据预测建模故事。如图 1-6 所示，谷歌公司依据全球用户对流感药物的在线查询情况已经可以提前六个月预测流感的爆发日期与流行路径。

每天都有成千上万的人通过 Google 来搜索信息，从旅途需要花费多长时间到怎样治疗他们孩子的病，各式各样的信息都有，这无疑极大地方便了人们的生活。