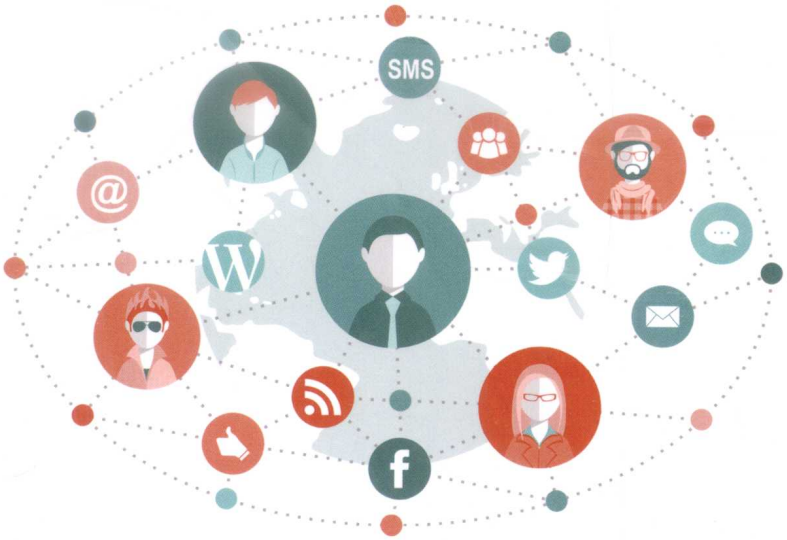
大数据时代，让你的数据发挥最大的价值！



数据挖掘

你必须知道的32个经典案例

任昱衡　李倩星　米晓飞　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

大数据时代，让你的数据发挥最大的价值！

数据挖掘

你必须知道的32个经典案例

任昱衡　李倩星　米晓飞　著

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

内 容 简 介

本书是为广大数据分析师量身定制的入门读物，它旨在帮助读者站在大数据时代的制高点。数据分析处于统计学、计算机信息科学、运筹学、数据库等多个领域的交叉地带，大数据时代的到来大大丰富了数据分析的内涵，数据分析师的职责与以往相比发生了巨大的改变。

本书全面介绍了经典数据分析、模式识别、机器学习、深度学习、数据挖掘、商务智能等多个领域的数据分析算法，将大数据时代的数据分析热点技术一网打尽。本书为每个数据分析算法都搭配了一个经典案例，并按照由易到难的原则构建知识框架，充分照顾了不同水平读者的阅读习惯。

通过阅读本书，读者将对大数据时代下的数据分析有一个全面的认识。无论是入门级的数据分析员还是有一定基础的数据分析师，都能通过本书完善、加深对数据分析的认识。

图书在版编目（CIP）数据

数据挖掘：你必须知道的 32 个经典案例 / 任昱衡，李倩星，米晓飞著. —北京：电子工业出版社，2016.1
ISBN 978-7-121-27579-1

Ⅰ. ①数… Ⅱ. ①任… ②李… ③米… Ⅲ. ①商业信息—数据采集—案例
Ⅳ. ①F713.51

中国版本图书馆 CIP 数据核字（2015）第 272252 号

策划编辑：李 冰 责任编辑：王丽萍
印 刷：三河市兴达印务有限公司
装 订：三河市兴达印务有限公司
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：720×1 000 1/16 印张：16.75 字数：255 千字 彩插：2
版 次：2016 年 1 月第 1 版
印 次：2016 年 1 月第 1 次印刷
定 价：59.00 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。

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。

服务热线：（010）88258888。

前　言

自 2015 年以来，“大数据时代”已成为最热门的名词之一。大数据在企业决策中扮演着越来越重要的角色，各个行业都不约而同地提出了大数据的口号，与大数据相关的新名词、新产品不断涌现，“统计分析和数据挖掘”跻身最受欢迎的求职技能行列，数据分析师的薪酬待遇也远远超过平均水平。与数据分析技能之火爆相对应的是数据分析人才的缺失。麦肯锡公司的研究报告表明，截至 2018 年，全球将面临 150 万数据分析人才方面的缺口。这意味着，有 150 万的其他行业从业人员将有可能把握住机会，转型为令人艳羡的数据分析师。

全面的数据改革迫在眉睫，但如何真正落实大数据，仍然是一个富于争议的话题。数据分析行业内部面临着相同的困境，在大数据时代，新的数据分析方法层出不穷，原有的数据分析方法也在不断完善，这些都导致数据分析师这一工作发生了令人措手不及的变化。为了帮助新的从业人员尽快熟悉数据分析这一工作，以及帮助原有的数据分析师尽快完成转型，

本书精心选择了 32 个流行的数据分析算法，并佐以案例，供大家了解大数据时代下数据分析行业的最新动态。

本书特色

1．条理清晰，内容翔实，全面介绍了大数据时代下的数据分析算法体系

数据分析处于统计学、人工智能、模式识别、机器学习等多个领域的交叉处。本书分为 8 章，每一章都有独立的主题，涵盖了数据分析所涉及的大部分学科。同时，本书不同章之间存在紧密的关联，揭示了不同学科之间的异同，以及它们是如何丰富数据分析的内涵，并影响数据分析学科发展的轨迹的。通过阅读本书，读者将站在制高点，一览无余地看到大数据时代下不同数据分析算法是如何彼此关联，构成完整体系的。

2．语言通俗易懂，内容由易到难，适合各层次读者学习

本书语言诙谐有趣，对每个数学公式都不厌其烦地举例讲解，即使毫无高等数学基础的读者也能够读懂本书所讲解的内容。同时，本书严格按照由易到难的学习规律编写，从较为简单的经典数据分析方法写起，逐渐过渡到较为晦涩的模式识别、机器学习等内容，通过阅读本书，读者将从一个“数据分析小白”迅速进阶为基础扎实、能独当一面的数据分析师。本书的内容涵盖了每个算法的原理、公式和适用场景、算法的优缺点。无论是数据分析菜鸟，还是有一定经验的数据分析师，本书都能够帮助你拓展、加深对数据分析的认识。

3．案例丰富精彩，应用贴近实际，满足读者的多重需求

本书为每一个算法都配备了一个精心选择的商业案例，这些案例横跨十几个行业，将大数据时代为人称道的经典案例一网打尽，满足了读者的三大需求：首先，结合案例学习算法能将算法讲解得更加清楚，加深读者

对算法的认识；其次，这些案例展示了数据分析在各行各业的最新应用，读者能够通过它们切实感受到数据分析的魅力，激发读者学习数据分析的热情；最后，这些案例涉及多个领域，不仅能够迎合从事不同领域工作的读者的口味，也能够使读者了解到数据分析在不同领域的现状，从而帮助读者选择进一步深入学习的方向。

本书内容及体系结构

第 1 章　经典的探索性数据分析案例

本章介绍了 4 种最基本的数据分析方法，分别是数据收集、数据可视化、异常值分析和对比分析。通过学习这 4 种数据分析方法，读者将对数据分析师的工作内容有一个初步的了解，使读者能够完成初级的数据分析任务。

第 2 章　经典的相关分析与回归分析案例

本章的主题是相关分析和回归分析。这两种分析方法经典、古老而有效，至今仍被广泛应用。其中，相关分析能够为回归分析做准备，回归分析又从侧面验证了相关分析结果的正确性。本章涉及了 1 种最常见的相关分析方法和 3 种最常见的回归分析方法，通过阅读本章，读者将获得解决小数据样本下的一大类数据分析问题的能力。

第 3 章　经典的降维数据分析案例

本章介绍了粗糙集算法、因子分析、最优尺度分析、PCA 降维算法等 4 种降维算法。本章是小数据分析和大数据分析交界的一章，这 4 种降维算法既可以为小数据分析服务，也可以为大数据分析服务。本章展示了降维分析与相关分析、回归分析的关联，加深了读者对小数据分析的理解，并为读者打开了大数据分析的大门。

第 4 章　经典的模式识别案例

本章感兴趣的问题是模式识别问题。模式识别算法研究的是如何让机器像人一样认识世界，它运用了较多的数学知识，并借助编程方法来实现。图像分析、遗传算法、决策树、K 均值是本章关心的主题，本章选取了与数据分析关系最密切的案例，旨在使读者了解模式识别与数据分析的区别与联系。

第 5 章　经典的机器学习案例

本章关心的内容是机器学习，机器学习学科致力于让机器拥有和人类一样的思考能力。通过阅读本章的语义搜索、顺序分析、文本分析、协同过滤 4 个算法，读者将发现机器学习更多的是从机器的角度来思考问题，这要求读者拥有更深入的编程思维方式，以便于更好地实现机器学习算法。

第 6 章　经典的深度学习案例

本章是对上一章的延伸，介绍了支持向量机、两种神经网络和 RBM 算法。深度学习是一个很大的命题，本章仅选取了与数据分析最相关的部分。另外，除了向读者介绍 4 种深度学习算法以外，本章还向读者指明了机器学习未来的发展方向，这将同样影响到数据分析未来的发展。

第 7 章　经典的数据挖掘案例

本章介绍了判别分析、购物篮分析、马尔可夫链、AdaBoost 元 4 种算法，实质上是对以上 6 章的查漏补缺。大数据时代加速了各个学科的融合，数据科学家借鉴了不同学科知识后创造出的数据分析算法也就具有了多种学科的特质。本章将这些“混血”算法集合起来，向读者展示了数据分析最多变的一面。

第 8 章　经典的商业智能分析案例

本章是对数据分析的升华和总结，在真正的数据分析项目中，数据分析师总是会运用多种数据分析方法来构建模型，本章所介绍的案例就是这样运用多种方法构建模型的例子。同时，本章还进一步辨析了数据分析和数据挖掘的异同，并隐含了作者对所有读者的寄语，读完本章后，读者就能对大数据时代下的数据分析有一个全面深入的认识了。

本书读者对象

- 刚刚入行的数据分析员
- 统计学、管理学、金融学、计算机技术与科学等专业的学生
- 想要提高数据分析能力的数据分析师
- 希望转行做数据分析的从业人员
- 想要增加对数据分析了解的主管人员
- 其他对数据分析感兴趣的读者

目 录

第 1 章 经典的探索性数据分析案例 …… 1
1.1 探索性数据分析综述 …… 2
1.2 数据巧收集——红牛的大数据营销案例 …… 4
1.2.1 状况百出的红牛企业 …… 4
1.2.2 红牛企业巧妙收集消费者数据 …… 6
1.2.3 数据收集小结 …… 8
1.3 数据可视化——数据新闻促使英军撤军 …… 9
1.3.1 维基解密带来的海量数据 …… 9
1.3.2 百花齐放的数据新闻 …… 11
1.3.3 数据可视化小结 …… 15
1.4 异常值分析——Facebook 消灭钓鱼链接 …… 16
1.4.1 Facebook 和广告商之间的拉锯战 …… 17
1.4.2 异常值分析指导排名算法工作 …… 18

1.4.3 异常值分析小结 …… 22
1.5 对比分析——TrueCar 指导购物者寻找最合算的车价 …… 24
1.5.1 火中取栗的 TrueCar 网站 …… 24
1.5.2 数据对比赢得消费者信赖 …… 26
1.5.3 对比分析小结 …… 29

第 2 章 经典的相关分析与回归分析案例 …… 31

2.1 相关回归综述 …… 32
2.2 皮尔逊相关值——纽约市政府利用相关分析监控违法建筑 …… 34
2.2.1 简约而不简单的消防检测系统 …… 34
2.2.2 使用相关分析洞察 60 个变量的关系 …… 36
2.2.3 相关分析小结 …… 39
2.3 时间序列分析——人寿保险的可提费用预测 …… 41
2.3.1 人寿保险公司和可提费用 …… 41
2.3.2 使用四种时间序列回归预测模型解决问题 …… 43
2.3.3 时间序列分析小结 …… 46
2.4 线性回归分析——梅西百货公司的十二项大数据策略 …… 48
2.4.1 从“一亿豪赌”说起的零售商困境 …… 48
2.4.2 SAS 公司帮助梅西百货构建模型 …… 50
2.4.3 线性回归分析小结 …… 53
2.5 Logistic 回归分析——大面积流感爆发的预测分析 …… 56
2.5.1 究竟谁才是流感预测算法之王 …… 56
2.5.2 向 Logistic 算法中引入更多变量 …… 58
2.5.3 Logistic 回归分析小结 …… 61

第 3 章 经典的降维数据分析案例 …… 63

3.1 降维分析算法综述 …… 64
3.2 粗糙集算法——协助希腊工业发展银行制定信贷政策 …… 66
3.2.1 银行信贷政策的制定原则 …… 66
3.2.2 粗糙集算法原理和应用 …… 67
3.2.3 粗糙集算法小结 …… 71

3.3 因子分析——基于李克特量表的应聘评价法 …… 73
3.3.1 源于智力测试的因子分析 …… 73
3.3.2 使用因子分析解构问卷 …… 75
3.3.3 因子分析小结 …… 78
3.4 最优尺度分析——直观评估消费者倾向的分析方法 …… 80
3.4.1 市场调查问题催生的最优尺度分析 …… 80
3.4.2 六种经典的最优尺度分析解读方法 …… 82
3.4.3 最优尺度分析小结 …… 86
3.5 PCA 降维算法——智能人脸识别的应用与拓展 …… 88
3.5.1 刷脸的时代来了 …… 88
3.5.2 使用 PCA 算法完成降维工作 …… 90
3.5.3 PCA 算法小结 …… 93

第 4 章 经典的模式识别案例 …… 95

4.1 模式识别综述 …… 96
4.2 图像分析——谷歌的超前自动驾驶技术 …… 98
4.2.1 以安全的名义呼吁自动驾驶技术 …… 98
4.2.2 快速成熟的无人驾驶技术 …… 100
4.2.3 图像分析小结 …… 103
4.3 遗传算法——经典的人力资源优化问题 …… 105
4.3.1 使用有限资源实现利益最大化 …… 105
4.3.2 遗传算法的计算过程 …… 107
4.3.3 遗传算法小结 …… 110
4.4 决策树分析——“沸腾时刻”准确判断用户健康水平 …… 111
4.4.1 打造我国最大健身平台 …… 111
4.4.2 信息增益和决策树 …… 113
4.4.3 决策树小结 …… 116
4.5 K 均值聚类分析——HSE24 通过为客户分类降低退货率 …… 118
4.5.1 在电子商务市场快速扩张的 HSE24 …… 119
4.5.2 使用 K 均值聚类为客户分类 …… 120

4.5.3 K均值聚类小结 …… 123

第5章 经典的机器学习案例 …… 127

5.1 机器学习综述 …… 128
5.2 语义搜索——沃尔玛搜索引擎提升15%销售额 …… 130
5.2.1 注重用户体验的沃尔玛公司 …… 130
5.2.2 语义搜索引擎的底层技术和原理 …… 132
5.2.3 语义搜索技术小结 …… 135
5.3 顺序分析——搜狗输入法的智能纠错系统 …… 137
5.3.1 搜狗输入法的王牌词库和智能算法 …… 137
5.3.2 频繁树模式和顺序分析算法 …… 140
5.3.3 顺序分析小结 …… 143
5.4 文本分析——经典的垃圾邮件过滤系统 …… 144
5.4.1 大数据时代需要文本分析工作 …… 145
5.4.2 垃圾邮件过滤中的分词技术和词集模型 …… 146
5.4.3 文本分析小结 …… 149
5.5 协同过滤——构建个性化推荐系统的经典算法 …… 151
5.5.1 协同过滤算法为什么这么流行 …… 151
5.5.2 基于用户和基于产品的协同过滤 …… 153
5.5.3 协同过滤算法小结 …… 155

第6章 经典的深度学习案例 …… 159

6.1 深度学习综述 …… 160
6.2 支持向量机——乔布斯利用大数据对抗癌症 …… 162
6.2.1 乔布斯和胰腺癌的八年抗战 …… 162
6.2.2 医学统计学和支持向量机 …… 164
6.2.3 支持向量机小结 …… 168
6.3 感知器神经网络——最佳的房产价格预测算法 …… 169
6.3.1 如何在我国预测房价 …… 170
6.3.2 多层感知器和误差曲面 …… 171
6.3.3 感知器神经网络小结 …… 175

6.4 自组织神经网络——如何又快又好地解决旅行商问题 …… 177
6.4.1 最优路径问题的典型模式和解决方法 …… 177
6.4.2 自组织神经网络的拓扑结构和权值调整 …… 178
6.4.3 自组织神经网络小结 …… 182
6.5 RBM 算法——为新闻报道智能分类 …… 183
6.5.1 新闻报道智能分类的难与易 …… 183
6.5.2 RBM 算法的学习目标和学习方法 …… 185
6.5.3 RBM 算法小结 …… 188

第 7 章 经典的数据挖掘案例 …… 191

7.1 数据挖掘综述 …… 192
7.2 判别分析——美国运通构建客户流失预测模型 …… 194
7.2.1 美国运通公司的旧日辉煌 …… 194
7.2.2 判别分析的假设条件和判别函数 …… 196
7.2.3 判别分析小结 …… 200
7.3 购物篮分析——找出零售业的最佳商品组合 …… 201
7.3.1 名动天下的“啤酒和尿布”案例 …… 202
7.3.2 购物篮分析的频繁模式 …… 203
7.3.3 购物篮分析小结 …… 207
7.4 马尔可夫链——准确预测客运市场占有率 …… 208
7.4.1 复杂的客运市场系统 …… 209
7.4.2 概率转移矩阵的求解方法 …… 210
7.4.3 马尔可夫链小结 …… 213
7.5 AdaBoost 元算法——有效侦测欺诈交易的复合算法 …… 215
7.5.1 弱分类器和强分类器之争 …… 215
7.5.2 AdaBoost 元算法的分类器构建方法 …… 217
7.5.3 AdaBoost 元算法小结 …… 220

第 8 章 经典的商业智能分析案例 …… 223

8.1 商业智能分析综述 …… 224
8.2 KXEN 分析软件——构建欧洲博彩业下注预测平台 …… 226

8.2.1 现代博彩业背后的黑手 …… 226
8.2.2 集体智慧和庄家赔率的联系 …… 228
8.2.3 KXEN 软件小结 …… 231
8.3 数据废气再利用——物流公司数据成功用于评估客户信用 233
8.3.1 数据废气和黑暗数据的异同 …… 234
8.3.2 论如何充分利用物流公司数据 …… 235
8.3.3 数据废气再利用小结 …… 239
8.4 必应预测——使用往期信息预测自然灾害 …… 240
8.4.1 预测自然灾害的必要性 …… 241
8.4.2 微软大数据预测的优与劣 …… 242
8.4.3 必应预测小结 …… 245
8.5 点球成金——助力 NBA 大数据分析的多种神秘软件 …… 246
8.5.1 NBA 的有效球员数据 …… 247
8.5.2 有关点球成金的靠谱方法 …… 249
8.5.3 点球成金小结 …… 251

第1章

经典的探索性数据分析案例

探索性数据分析综述

数据巧收集——红牛的大数据营销案例

数据可视化——数据新闻促使英军撤军

异常值分析——Facebook 消灭钓鱼链接

对比分析——TrueCar 指导购物者寻找最合算的车价

数据的筛选、重组、结构化、预处理等都属于探索性数据分析的范畴，探索性数据分析是帮助数据分析师掌握数据结构的重要工具，也是奠定后续工作之成功的基石。本章选取了 4 个经典的案例，从 4 个角度展示了探索性数据分析是怎样工作的。通过对本章的学习，读者将掌握基本的数据分析方法。

1.1 探索性数据分析综述

正如学习英语要从 A、B、C、D 开始一样，我们学习数据分析也要从数据的收集和预处理开始。有一些急于求成的人认为数据的收集和预处理很简单，没有什么技术含量，马马虎虎看一下就好了，重点还是应该放在后续的数据分析算法上。这种看法是非常肤浅的，在数据分析项目中，数据的收集和预处理往往占据整个项目工作量的十之八九。正是这些简单的工作决定了整个项目的成败。

举个例子，当我们对一份数据作回归分析，发现吸烟越多得癌症的概率就越小时，我想正常人的反应都是返回检查数据是否出错了，而不是欣喜若狂地赶快去医学杂志发表论文。再比如，当我们分析一份数据时，发现甲二苯甘酸越多，胞嘧啶就越少时，请问大家能一眼判断出这个结果是否具有发表论文的价值吗？对于欠缺相关专业知识的人来说，只有首先进行探索性数据分析，才能准确理解分析结果的意义。这就是为什么我们需要探索性数据分析的原因。

从广泛意义上讲，探索性数据分析主要包括数据的预处理和数据

的探索性分析。其中数据的预处理是指对数据进行清洗、转化、重组和筛选，而数据的探索性分析则包括基本的五数总括、数据分布等，简单的相关分析和方差分析等，也都属于数据的探索性分析的范畴。通常情况下，我们不认为探索性数据分析包含数据的收集，但篇幅所限，本书将数据的收集也归入探索性数据分析这一章。

数据的收集是整个数据分析项目的原点，没有收集来的数据，什么数据分析技术都是纸上谈兵。对于一小部分较为常见的问题，比如预测市政府的财政收入，预测未来某一时间的天气数据等问题，都可以从相关的公开网站下载相关的数据包。但对于大部分商家根据自身情况提出的特定问题来说，则需要专门设计收集数据的方法，一个巧妙的方法可以节约成千上万的资金，本章要介绍的红牛营销案例就是一个经典的例子。

在真实生活中，收集的数据往往是不能直接用来进行高级数据分析的，这是因为原始数据中会包含许多残缺值和错误值，比如将一个人的身高记录为 17.5 米，这显然是一个录入出错的值，数据预处理就是要将原始数据中的残缺值和错误值一一剔除，只留下有意义的数据。除此之外，也并不是所有的数据变量都适合进行数据分析，因此数据预处理还要承担起挑选有价值的数据变量的任务。

简单的探索性数据分析主要用于研究数据的分布结构。研究一个数据变量的极大值、极小值、中位数各是什么，数据呈正态分布还是偏态分布是十分重要的。比如回归分析就要求分析变量具有正态分布的特征。探索性数据分析可以使数据分析师直观掌握数据的各项特征，这一点将帮助数据分析师在后续分析中选择更合适的数据分析技术。