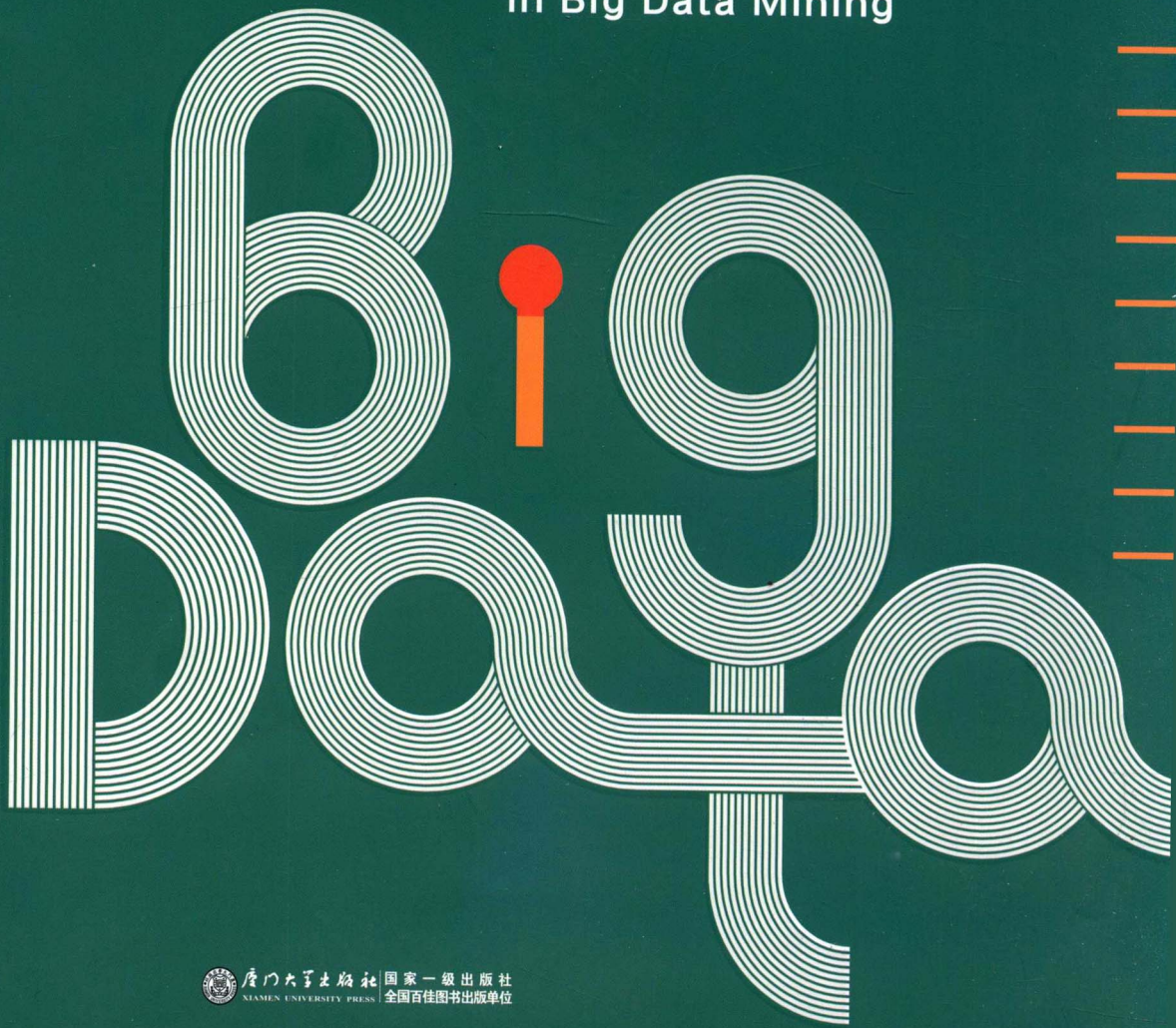
谢邦昌 朱建平 王小燕 著

厦门大学数据挖掘研究中心
厦门大学管理学院 MBA 中心
大数据丛书

2

Excel
在大数据挖掘中的应用

The Application of Excel
in Big Data Mining



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

王小燕 著

Excel
在大数据挖掘中的应用

The Application of Excel in Big Data Mining

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

图书在版编目(CIP)数据

Excel 在大数据挖掘中的应用/谢邦昌，朱建平，王小燕著. —厦门：厦门大学出版社，2016.5

ISBN 978-7-5615-6012-9

Ⅰ.①E… Ⅱ.①谢…②朱…③王… Ⅲ.①数据处理-表处理软件 Ⅳ.①TP274

中国版本图书馆 CIP 数据核字(2016)第 065182 号

出 版 人 蒋东明
责任编辑 吴兴友
封面设计 王 琳
美术编辑 李夏凌
责任印制 吴晓平

出版发行 厦门大学出版社
社 址 厦门市软件园二期望海路 39 号
邮政编码 361008
总 编 办 0592-2182177 0592-2181253(传真)
营销中心 0592-2184458 0592-2181365
网 址 http://www.xmupress.com
邮 箱 xmupress@126.com
印 刷 厦门市万美兴印刷设计有限公司

开本 787mm×1092mm 1/16
印张 21.25
插页 2
字数 500 千字
印数 1～3 000 册
版次 2016 年 5 月第 1 版
印次 2016 年 5 月第 1 次印刷
定价 55.00 元

厦门大学出版社
微信二维码

厦门大学出版社
微博二维码

前 言

21世纪是大数据的时代，又是“互联网+”的时代，借助于计算机存储和分布式计算带来的便捷，信息呈爆炸式增长，各行各业时刻都在生成、收集并处理大量的数据，依赖大数据分析技术的时代悄悄来临。从谷歌推出的流感预测，到百度的市场物价预测和春运的迁徙动态图，能察觉到大数据分析已深入我们的生活。数据挖掘作为大数据分析的核心技术，面对这些繁杂又冗余的大数据，其性能、速度、可视化程度都受到了时代的挑战。

Excel是当前使用最普遍、入门极快的电子表格软件，它能轻易地完成大量数据的统计分析、处理与制图，不仅功能强大，还操作简单。最新版本Microsoft Office Excel 2013是Excel发展历程中一个里程碑级的产品，对比Excel 2010，它非常适用于“大数据”和“互联网+”时代下的数据挖掘工作。它提供了数项让人眼前一亮的新功能，包括Power Query，Power View，Power Map和PowerPivot。

为了提升Excel的数据挖掘和统计分析能力，微软公司提供了一个免费的SQL Server 2012数据挖掘外接程序。通过该程序，Excel 2013能够快速地完成许多专业数据挖掘软件才能完成的任务。在大数据处理非常频繁和迫切的时代，许多学者在学习或工作中面临数据挖掘模型如何实现的难题，因此，我们编写了本书，旨在用最简单的Excel 2013软件工具，轻松、快速地实现复杂的数据处理、分析和挖掘工作。

本书的编写以统计原理为主线，以数据挖掘模型的应用为目的。全书共19章，基本内容和特点体现为：第1篇包括第一章至第四章，概述数据挖掘在大数据时代的重要性，详细总结常用的数据挖掘理论和技术的一般概念、方法技术，并总结了常用的数据挖掘软件和工具的特点和适用性，使读者获得清晰的数据挖掘观

念。第2篇从第五章至第十七章，详细地讲述了数据挖掘加载项的配置要求和安装过程。接着从理论、操作、实例展示三个方面重点讲述了数据挖掘加载项的9种数据挖掘模型的使用，包括决策树、贝叶斯概率分类、关联规则、聚类分析、时序聚类、线性回归、Logistic回归、神经网络聚类、时间序列分析。内容全面，涵盖了目前学术界和实业界主要的数据挖掘技术和方法。同时，通过实例操作介绍多种数据挖掘辅助技术，包括关键影响因素、类别检测、从示例填充、预测、异常值检测、应用场景分析。第3篇为实例篇，包含第十八、十九章，以台湾房屋信用贷款数据和台湾健康食品行业数据为例，应用第2篇中多种数据挖掘方法，在Excel 2013数据加载项上展开实例分析，通过详细的操作演示和结果解释，读者可获得实际的数据挖掘经验，并能迅速在自己的学习和工作中加以应用。最后为附录部分，通过详细的操作示范，讲述了Excel 2013的四大特色数据分析功能，包括Power Query，Power View，Power Map和PowerPivot。

本书适合多层次多专业的读者，如数据、统计、经济金融、工商管理等专业的本科生、研究生学习，还适合于非统计类从事相关数据分析的人员阅读。

本书由台北医学大学谢邦昌教授设计整体框架，并同厦门大学朱建平教授、湖南大学王小燕助理教授共同撰写。在本书编写过程中，厦门大学研究生范新妍、黄晓雯、张悦涵为本书第2篇部分章节和附录的资料收集和录入，以及部分案例的整理做了大量的工作。谢邦昌和朱建平教授对本书进行了修改和总纂，台北医学大学邓光甫教授审阅了全书。

本书在编写和出版过程中，得到了厦门大学数据挖掘研究中心、台北医学大学大数据研究中心、厦门大学管理学院MBA中心、湖南大学金融与统计学院和厦门大学出版社的支持，陈丽贞和吴兴友同志为本书的组稿、编辑做了大量的工作，在此表示衷心感谢！编写一本好的书并不容易，尽管我们努力想奉献给读者一本满意的书，但仍有达不到读者各方面的要求的地方。书中难免有疏漏或错误之处，恳请读者多提宝贵意见，以便今后进一步修改与完善。

本书的编写和出版得到了国家社会科学基金重大项目“大数据与统计学理论的发展研究”(13&2D148)的资助。

作者

2016年1月

CONTENT

BIG DATA

第一篇 基本知识

第二篇 Excel 2013 数据挖掘模型

附 录

第一篇

基本知识

BIG DATA

第一章
大数据与数据挖掘

随着云计算、电子商务、网络技术的飞速发展，数据种类和规模在金融服务、零售、科研、医疗保健、能源、交通等诸多行业飞速增长。2006 年全球新产生的数据量大约为 180 EB，2011 年达到 1.8 ZB，有研究机构预测到 2020 年这个数字将增长到 35.2 ZB，大数据时代伴随而来，它成为继“云端”以后在 IT 界深入人心的流行词汇。

1.1 大数据的定义

“大数据”一词最早出现在 1997 年迈克尔·考克斯的论文中，直到 2008 年 9 月，《科学》杂志发表论文“Big Data：Science in the Petabyte Era”，大数据一词开始广泛传播。如同“云端”一样，在出现之初并没有给它明确的定义。大数据到底是什么，它极其复杂，至今还没有一个得到普遍认同的官方定义。

大数据不仅仅等于数据量大，亚马逊大数据科学家 John Rauser 将其定义为“任何超过一台计算机处理能力的庞大数据量”；Informatica 中国区首席产品顾问但彬认为“大数据＝海量数据＋复杂类型的数据”，其规模和复杂程度超过了常用技术按照合理的成本和时限捕捉、管理以及处理这些数据集的能力；维基百科将其描述为“任何由于规模庞大且高度复杂而难以通过现有数据库管理工具或者传统数据处理应用进行处理的数据集”。

通俗地说，大数据是现有的一般技术难以管理的海量数据。“现有的一般技术难以管理”，一方面是由于大数据的结构极其复杂，传统的关系数据无法管理；另一方面是大数据在量方面的急剧增加，导致查询数据的反应时间超过容许范围。

1.2 大数据的 4V 特征

顾名思义，大数据具有数据量大的特征，但这不是它的全部，一般而言，它具有以下四个特征：

大量性(volume):是指数据量的庞大性以及规模的完整性,主要体现在存储和计算两方面,其单位从TB上升到PB级别,在互联网时代,社交网络、电子商务等把人类带入了"PB"新时代。这与数据存储和网络技术的发展密切相关。据统计,互联网一天产生的全部内容足以制作1.68亿张DVD,一天发出2 940亿封邮件以及200万个帖子。

多样性(variety):它不仅仅具有量的剧增性,同时也有数据复杂性的提升,主要是指大数据包含的数据类型繁多。电商和物联网的发展,使得它不再局限于传统的结构化数据,还包括半结构以及非结构化数据,如文本、音频、图片、搜索记录等,后者所占比例高达85%。

高速性(velocity):主要体现在大数据的增长速度快,以及传输和处理速度快。大数据通过云计算,可以在短短几十分钟内,将传统模式下十几天才能完成的数据存储完毕。

价值性(value):这主要体现了大数据的应用价值。但是它的价值性具有稀疏性、不确定性和多样性。数据量大并不意味着它们都是有效的,这其中大部分数据都是噪音,价值密度很低。数据量增长的同时,隐藏的有价值信息并没有相应比例增长,从而获取有用信息的难度加大。

1.3 大数据的预测魅力

大数据的预测目前主要体现在商业价值上,已有相关的公司利用大数据技术进行研究。国内最有代表性的有三家公司,分别是百度、阿里巴巴以及腾讯。具体来说,大数据应用到了如下方面:

(1)大数据预测世界杯赛事

在世界杯紧张进行的同时,百度、谷歌、微软以及高盛等科技公司都推出了赛事预测平台,百度预测结果尤为亮眼。在小组赛阶段,百度以58.33%的准确率居首,在淘汰赛阶段,百度和微软全程预测16进8及8进4的比赛,准确率100%;谷歌预测了12场比赛仅一场错误,准确率为91.67%。百度在这场预测大赛中领跑,其大数据研究院利用百度大数据搜索过去5年内全球987支球队的3.7万场比赛数据,并与乐彩网等建立数据战略合作伙伴关系,将博彩数据融入预测模型,同时考虑了团队实力、主场优势、最近表现、世界杯整体表现和博彩公司赔率等五个因素。这意味着大数据已取代章鱼保罗来掌控赛事预测。

(2)百度迁徙

百度作为技术型公司,已诞生了多个大数据背景下的产品。百度迁徙是百度在2014年春运期间基于大数据技术推出的一个品牌项目。它基于LBS大数据进行全样本的数据处理、分析和挖掘,在业界首次实现了全程、动态、即时和直观地展示人口迁徙的轨迹与特征,并实现可视化。

(3)疾病预测

基于用户的搜索数据以及购物行为,并结合天气、环境和人口流动等因素,百度与中

国疾病预防和控制中心合作，就流感、肺炎、肺结核、性病这四种疾病，对全国每一个省份以及大多数市区的活跃度和趋势图等进行了全面的监控。这是继世界杯预测后百度大数据预测的又一款产品。用户可以查看过去 30 天以内的数据以及未来 7 天的预测趋势。

(4)市场物价预测

CPI 表征物价的浮动情况，但是目前的 CPI 存在偏差性，仅仅编制一个物价总指数显然不足以准确反映不同收入阶层的实际价格水平。同时，CPI 存在较长的时滞性，会弱化它的信息反馈功能。大数据时代数据的获取不再是问题，数据更新速度更快，这为统计调查提供了海量的原始资料。大数据可以帮助人们了解未来物价走向，提前预知通货膨胀或经济危机。阿里巴巴团队通过阿里 B2B 大数据提前预知亚洲金融危机，这是市场物价预测的典型案例。

此外，阿里巴巴在业务驱动下，已通过大数据着手分析各种用户消费行为，“云端＋大数据”已成为它的重要战略。阿里巴巴会记录淘宝和天猫的每一笔消费记录，交易及信用数据成为它的一手资料，从而建立其数据库。而云平台阿里云是阿里巴巴支撑大数据的主要组成部分，在网购高峰期如 2013 年的“双十一”，75％的交易都是在阿里云平台上运行。

大数据的预测应用还有景点预测、城市旅游预测、高考考研预测、金融预测、票房预测、选举预测、奥斯卡获奖预测等等，大数据的预测魅力将在各个领域得到体现。

1.4 数据挖掘定义

“Data mining is the process of seeking interesting or valuable information in large data bases.”

数据挖掘(data mining)是近年来数据库应用领域中相当热门的话题。数据挖掘指在数据库中，利用各种分析方法对累积的大量历史数据进行分析、归纳与整合等工作，以提取出有用的信息，找出有意义且用户有兴趣的模式(interesting patterns)，供企业管理层进行决策。

数据挖掘指找寻隐藏在数据中的信息，如趋势(trend)、模式(pattern)及相关性(relationship)的过程，也就是从数据中发掘信息或知识(也称 knowledge discovery in databases，KDD)，也有人称为数据考古学(data archaeology)、数据模式分析(data pattern analysis)或功能相依分析(functional dependency analysis)，目前已被许多研究人员视为结合数据库系统与机器学习技术的重要领域，许多产业界人士也认为此领域是一项增加各企业潜能的重要指标。

事实上，数据挖掘并不只是一种技术或一套软件，而是一种结合多项专业知识的应用。但我们对数据挖掘应有一个正确的认识：数据挖掘工具是从数据中发掘出各种可能的假设(hypothesis)，但并不帮你证实这些假设，也不帮你判断这些假设对你的价值。