先导思想力
教材领航

DATA MINING AND ITS APPLICATION WITH SAS AND R
(2ND EDITION)

数据挖掘与应用

以SAS和R为工具

(第二版)

张俊妮◎著

北京大学出版社
PEKING UNIVERSITY PRESS

数据挖掘与应用

——以SAS和R为工具

（第二版）

张俊妮◎著

北京大学出版社
PEKING UNIVERSITY PRESS

图书在版编目(CIP)数据

数据挖掘与应用：以 SAS 和 R 为工具/张俊妮著. —2 版. —北京：北京大学出版社，2018. 10
(光华思想力书系·教材领航)
ISBN 978-7-301-29909-8

Ⅰ.①数… Ⅱ.①张… Ⅲ. ①数据采集-高等学校-教材 Ⅳ. ①TP274

中国版本图书馆 CIP 数据核字（2018）第 214457 号

书　　名 数据挖掘与应用：以 SAS 和 R 为工具（第二版）
SHUJU WAJUE YU YINGYONG (DI-ER BAN)
著作责任者 张俊妮　著
责 任 编 辑 裴蕾
标 准 书 号 ISBN 978-7-301-29909-8
出 版 发 行 北京大学出版社
地　　址 北京市海淀区成府路 205 号　100871
网　　址 http://www.pup.cn
电 子 信 箱 em@pup.cn　　QQ:552063295
微信公众号 北京大学经管书苑 (pupembook)
电　　话 邮购部 010-62752015　发行部 010-62750672　编辑部 010-62752021
印 刷 者 涿州市星河印刷有限公司
经 销 者 新华书店
720 毫米×1020 毫米　16 开本　22.25 印张　528 千字
2009 年第 1 版
2018 年 10 月第 2 版　2018 年 10 月第 1 次印刷
定　　价 58.00 元

丛书编委会

丛书序言一

很高兴看到“光华思想力书系”的出版问世，这将成为外界更加全面了解北京大学光华管理学院的一个重要窗口。北京大学光华管理学院从1985年北京大学经济管理系成立，以“创造管理知识，培养商界领袖，推动社会进步”为使命，到现在已经有三十余年了。这三十余年来，光华文化、光华精神一直体现在学院的方方面面，而这套“光华思想力书系”则是学院各方面工作的集中展示，同时也是北京大学光华管理学院的智库平台，旨在立足新时代，贡献中国方案。

作为经济管理学科的研究机构，北京大学光华管理学院的科研实力一直在国内处于领先位置。光华管理学院有一支优秀的教师队伍，这支队伍的学术影响在国内首屈一指，在国际上也发挥着越来越重要的作用，它推动着中国经济管理学科在国际前沿的研究和探索。与此同时，学院一直都在积极努力地将科研力量转变为推动社会进步的动力。从当年股份制的探索、证券市场的设计、《证券法》的起草，到现在贵州毕节实验区的扶贫开发和生态建设、教育经费在国民收入中的合理比例、自然资源定价体系、国家高新技术开发区的规划，等等，都体现着光华管理学院的教师团队对中国经济改革与发展的贡献。

多年来，北京大学光华管理学院始终处于中国经济改革研究与企业管理研究的前沿，致力于促进中国乃至全球管理研究的发展，培养与国际接轨的优秀学生和研究人员，帮助国有企业实现管理国际化，帮助民营企业实现管理现代化，同时，

为跨国公司管理本地化提供咨询服务，从而做到“创造管理知识，培养商界领袖，推动社会进步”。北京大学光华管理学院的几届领导人都把这看作自己的使命。

作为人才培养的重地，多年来，北京大学光华管理学院培养了相当多的优秀学生，他们在各自的岗位上作出贡献，是光华管理学院最宝贵的财富。光华管理学院这个平台的最大优势，也正是能够吸引一届又一届优秀的人才的到来。世界一流商学院的发展很重要的一点就是靠它们强大的校友资源，这一点，也与北京大学光华管理学院的努力目标完全一致。

今天，“光华思想力书系”的出版正是北京大学光华管理学院全体师生和全体校友共同努力的成果。希望这套丛书能够向社会展示光华文化和精神的全貌，并为中国管理学教育的发展提供宝贵的经验。

北京大学光华管理学院名誉院长

丛书序言二

“因思想而光华。”正如改革开放走过的 40 年，得益于思想解放所释放出的动人心魄的力量，我们经历了波澜壮阔的伟大变迁。中国经济的崛起深刻地影响着世界经济重心与产业格局的改变；作为重要的新兴经济体之一，中国也越来越多地承担起国际责任，在重塑开放型世界经济、推动全球治理改革等方面发挥着重要作用。作为北京大学商学教育的主体，光华管理学院过去三十余年的发展几乎与中国改革开放同步，积极为国家政策制定与社会经济研究源源不断地贡献着思想与智慧，并以此反哺商学教育，培养出一大批在各自领域取得卓越成就的杰出人才，引领时代不断向上前行。

以打造中国的世界级商学院为目标，光华管理学院历来倡导以科学的理性精神治学，锐意创新，去解构时代赋予我们的新问题；我们胸怀使命，顽强地去拓展知识的边界，探索推动人类进化的源动力。2017 年，学院推出“光华思想力”研究平台，旨在立足新时代的中国，遵循规范的学术标准与前沿的科学方法，做世界水平的中国学问。“光华思想力”扎根中国大地，紧紧围绕中国经济和商业实践开展研究；凭借学科与人才优势，提供具有指导性、战略性、针对性和可操作性的战略思路、政策建议，服务经济社会发展；研究市场规律和趋势，服务企业前沿实践；讲好中国故事，提升商学教育，支撑中国实践，贡献中国方案。

为了有效传播这些高质量的学术成果，使更多人因阅读而受益，2018 年年初，

在和北京大学出版社的同志讨论后，我们决定推出“光华思想力书系”。通过整合原有“光华书系”所涵盖的理论研究、教学实践、学术交流等内容，融合光华未来的研究与教学成果，以类别多样的出版物形式，打造更具品质与更为多元的学术传播平台。我们希望通过此平台将“光华学派”所创造的一系列具有国际水准的立足中国、辐射世界的学术成果分享到更广的范围，以理性、科学的研究去开启智慧，启迪读者对事物本质更为深刻的理解，从而构建对世界的认知。正如光华管理学院所倡导的“因学术而思想，因思想而光华”，在中国经济迈向高质量发展的新阶段，在中华民族实现伟大复兴的道路上，“光华思想力”将充分发挥其智库作用，利用独创的思想与知识产品在人才培养、学术传播与政策建言等方面作出贡献，并以此致敬这个不凡的时代与时代中的每一份变革力量。

北京大学光华管理学院院长

前　言

本书的第一版于 2009 年 4 月出版。当时，“大数据” 这一词尚未流行，数据挖掘仅在一些行业崭露头角。近十年过去了，大数据已经成为时代特征，数据分析也成为几乎所有组织都积极获求的能力。本书的改版希望有助于回应这样的需求。

在第一版的基础上，本书增加了缺失数据、回归模型中的规则化和变量选择、卷积神经网络、支持向量机、协同过滤这五章内容。在已有各章内，本书亦增加了新的内容和示例。近些年来，R 因为其自由、免费、开源，已经发展为数据分析领域最强大的软件之一。因此，本书除了继续展示 SAS 程序，还增加了 R 程序。

我要特别感谢首都师范大学的田旭平同学，她为本书新增的 SAS 程序和大部分 R 程序写了初稿，并录制了所有程序操作教程的视频。她充满责任心，有很强的动手能力。她美妙的声音也让视频令人愉悦。我也要感谢本书的编辑，北京大学出版社的裴蕾女士，她认真负责的工作提升了本书的质量。

我还要感谢北京大学光华管理学院的同学王菲菲（现已就职）、杨兵（现已就职）、万雅婷、任图南、林颖倩、张轩瑜，他们重新整理了本书第一版使用的数据，并为本书新增的小部分 R 程序写了初稿。我也要感谢所有在本书第一版出版之后修过北京大学光华管理学院“数据挖掘与应用”课程的同学们，他们的反馈为本书新增内容的选题提供了基础。

张俊妮

2018 年 7 月

于北大燕园

目 录
CONTENTS

第1章

数据挖掘概述

1.1 什么是数据挖掘

任何一个组织（政府部门、企业、学校等）在决策与运营活动中都会积累丰富的经验，同时也面临着在不断变化的环境下做出快速而正确决策的挑战。数据挖掘方法首先根据组织所积累的经验收集可度量的数据（包括内部数据和外部数据），对这些数据进行分析后，提炼出对运营管理有指导意义的新知识，进一步改进决策、改善运营活动（见图 1.1）。这是一个持续改进的过程，决策运营活动不断积累新的经验，新的数据不断被收集，使用数据挖掘方法分析新的数据后不断产生新的知识，不断地促进决策与运营活动的改进和完善。

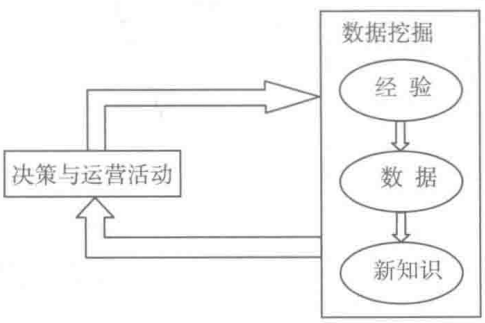

图 1.1 数据挖掘与决策和运营活动

Berry and Linoff（2000）将数据挖掘定义为：对大量数据进行探索和分析，以便发现有意义的模式和规则的过程。数据挖掘活动主要分为无监督和有监督两大类。在无监督数据挖掘中，我们对各个变量不区别对待，而是考察它们之间的关系。这类方法包括：描述和可视化、关联规则分析、主成分分析、聚类分析等。在有监督数据挖掘中，我们希望建立根据一些变量来预测另一些变量的模型，前者被称为自变量，后者被称为因变量。这类方法包括：线性及广义线性回归、神经网络、决策树、随机森林、支持向量机等。有监督数据挖掘能从数据中获取深度细致的信息，应用非常广泛。

在大数据时代，大量数据的收集和存储成为常态，对数据进行探索和分析也成为常态。应该收集什么数据、如何从数据中发现有意义的模式和规则才是真正的挑战，这里的“有意义”指的是根据具体需要用数据分析来回答和解决问题。

1.2 统计思想在数据挖掘中的重要性

在数据挖掘项目中，即使数据量很大、算法非常先进，也需要统计思想的指引。我们通过一些案例来说明这一点。

案例：谷歌流感趋势

2009 年 2 月，谷歌的一个研究小组在《自然》杂志上发表论文（Ginsberg et al., 2009），介绍了“谷歌流感趋势”。研究小组以“咳嗽”“发烧”等与流感相关的关键词的搜索频率为自变量，以疾病预防控制中心的数据为因变量，根据历史数据建立了统计模型。他们发现，可以通过监测相关关键词的变化趋势，追踪美国境内流感的传播趋势，而这一结果不依赖于任何医疗检查。谷歌的追踪结果很及时，而疾病控制中心则需要汇总大量医师的诊断结果才能得到一张流感传播趋势图，延时为一至两周。这一结果令人激动，谷歌流感趋势也成为引爆“大数据”这个名词的著作《大数据时代：生活、工作与思维的大变革》（维克托 · 迈尔 · 舍恩伯格，2012）的开篇案例。

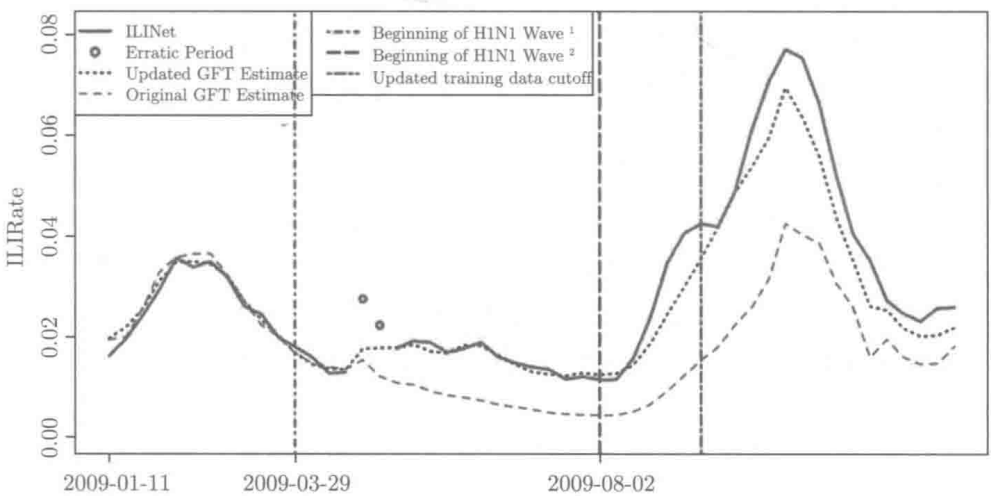

图 1.2　2009 年谷歌流感趋势拟合

然而，在 2009 年甲型 H1N1 流感流行的时候，谷歌流感趋势模型严重低估了流感的数量（见图 1.2[1]）。谷歌工程师们认为，出现这种现象的原因是之前谷歌流感趋势模型使用的数据都是关于季节性流感的，而 2009 年的 H1N1 流感是病毒性流感，人们搜索的关键词发生了很大变化，所以模型不再适用（Cook et al., 2011）。他们对模型使用的关键词包进行了大幅度修改。原始模型含大约 40 个关键词，修改后的模型含大约 160 个关键词，两个模型只有 11 个共同的关键词。不仅如此，关键词的类别也发生了很大的变化。图 1.3[2] 展现了修改前和修改后关键词在各个

1. 资料来源：Cook et al. (2011), https://doi.org/10.1371/journal.pone.0023610.g001。这里只取了原图的第一部分。

2. 资料来源：Cook et al. (2011), https://doi.org/10.1371/journal.pone.0023610.t001。

类别的分布状况。在原始模型中,“肺炎”(pneumonia)等与流感复杂性相关的关键词占全部关键词的 42%,而在修改后的模型中,这一比例仅为 6%。修改后的模型对 2009 年疾病控制中心的数据达到了高度拟合(见图 1.2)。

Query Category	Sample Query	Original Model Relative Category Volume	Updated Model Relative Category Volume
Symptoms of an influenza complication	[symptoms of bronchitis]	6%	11%
Influenza complication	[pnumonia]*	42%	6%
Specific influenza symptom	[fever]	6%	39%
General influenza symptoms	[early signs of the flu]	2%	30%
Cold/flu remedy	[robitussin]	12%	4%
Term for influenza	[influenza a]	<1%	3%
Antibiotic medication	[amoxicillin]	12%	0%
Related disease	[strep throat]	16%	<1%

*Search users often misspell the word *pneumonia*.

图 1.3 谷歌流感趋势模型中关键词包的修改

好景不长。2013 年 1 月,美国流感发生率达到峰值,谷歌流感趋势的估计比实际数据高两倍。这一不精确的估计再次引起了媒体的关注。

英国《金融时报》专栏作家蒂姆·哈福德发表了《大数据,还是大错误》一文(Hartford, 2014),指出在大数据时代,数据的规模很大,相对容易采集,而且可以实时地更新。大数据的鼓吹者们提出了四个令人兴奋的论断,每一个都能从谷歌流感趋势的“成功”中得到“印证”:

(1) 数据分析可以生成准确率惊人的结果;

(2) 因为每一个数据点都可以被捕捉到,所以可以彻底淘汰过去抽样统计的方法;

(3) 不用再寻找现象背后的原因,我们只需要知道两者之间有统计相关性就行了;

(4) 不再需要科学的或者统计的模型,理论被终结了,数据已经大到可以自己说出结论。

谷歌流感趋势的教训之一是抽样偏差。我们感兴趣的是推断整个人群中的流感患病率,但在谷歌上进行搜索的群体可能代表不了整个人群。例如,在谷歌上进行搜索的群体中年龄比较大的人和儿童所占的比例可能低于其在整个人群中的比例。

这一教训不是新近才有的。关于抽样偏差的一个著名案例是 1936 年的美国总统选举。参加竞选的候选人是民主党的罗斯福和共和党的兰登。《文学摘要》杂志给有电话或汽车,或者阅读该杂志的人发出 1 000 万份调查问卷,回收的问卷中有 1 293 669 份支持兰登,972 897 份支持罗斯福。所以该杂志预测兰登会胜出。而当时并不出名的盖洛普公司使用统计抽样的方式进行民意调查,调查了几千人,得到罗斯福会胜出的结论。最终结果是罗斯福获胜。1936 年,整个人群中只有少量