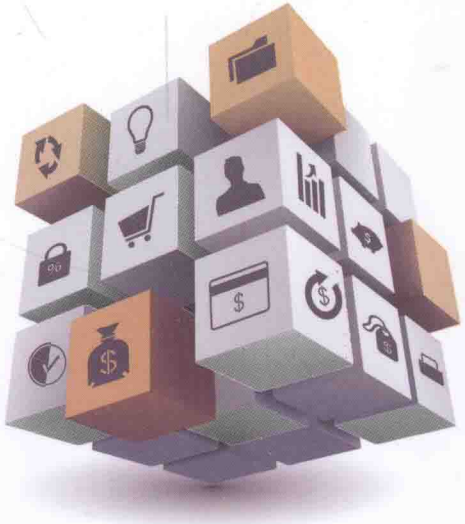
15个行业案例经验分享

IBM SPSS Modeler

数据与文本挖掘实战

王国平 郭伟宸 汪若君 编著

- SPSS Modeler操作+数据挖掘原理+数据挖掘案例
- 理论结合实践，解决数据挖掘在商业中的各种问题



提供本书资源文件下载

清华大学出版社

IBM SPSS Modeler

数据与文本挖掘实战

王国平 郭伟宸 汪若君 编著

清華大學出版社
北京

内 容 简 介

本书主要包括两部分内容：在数据挖掘部分，重点介绍了各种数据挖掘方法的基本原理及应用，包括回归分析、时间序列分析、因子分析、决策树分析、判别分析、聚类分析、人工神经网络、贝叶斯网络以及社交网络分析等；在文本挖掘部分，重点介绍了文本挖掘的节点，以及具体的实现过程。每一章都详细介绍了数据和文本挖掘的基本原理和分析过程，同时在实例中也介绍了 SPSS Modeler 中大部分节点的使用方法及应用步骤。

本书与同类书籍相比，安排了较多的实例，使读者能够边学边练，在短时间内就可以有一个较大的提高，方便读者熟悉 SPSS Modeler 的基本操作，并通过系统的案例使读者掌握应用技巧。

本书对于高校理工学科、经济金融学科及数量分析方面的学生，以及数据挖掘和分析方面的研究人员和从业人员等，具有很强的可读性、可操作性与可使用性，尤其适合商业销售、经济管理、社会研究和人文教育等行业的相关人员阅读。

图书在版编目（CIP）数据

IBM SPSS Modeler 数据与文本挖掘实战/王国平，郭伟宸，汪若君编著. —北京：清华大学出版社，2014
ISBN 978-7-302-37212-7

Ⅰ. ①I… Ⅱ. ①王… ②郭… ③汪… Ⅲ. ①统计分析－软件包 Ⅳ. ①C819

中国版本图书馆 CIP 数据核字（2014）第 152149 号

责任编辑：王金柱
封面设计：王　翔
责任校对：闫秀华
责任印制：刘海龙

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：三河市中晟雅豪印务有限公司
经　　销：全国新华书店
开　　本：180mm×230mm　　**印　　张**：19.25　　**字　　数**：493 千字
版　　次：2014 年 11 月第 1 版　　**印　　次**：2014 年 11 月第1 次印刷
印　　数：1～3000
定　　价：55.00 元

产品编号：056342-01

前　言

数据挖掘是一个逐步演进的过程，在电子数据处理的初期，人们就试图通过某些方法来实现自动决策支持，当时机器学习正成为人们关注的焦点，机器学习的过程就是将一些已知的并已被成功解决的问题作为范例输入计算机，相应的软件通过学习这些范例总结并生成相应的规则，通常这些规则具有通用性，使用它们可以快速解决某一类的实际问题。随着神经网络技术的形成和发展，人们将注意力转向知识工程，知识工程不同于机器学习（向计算机输入范例，让它生成规则），而是直接给计算机输入已被代码化的规则，计算机通过使用这些规则来解决某些问题。

上个世纪 80 年代人们又在新的神经网络理论的指导下，重新将注意力转回到机器学习的方法上，并将其成果应用于处理大型商业数据库。随着新术语——知识发现（简称 KDD，即 Knowledge Discovery in Database）逐渐被人们所接受，并用 KDD 来描述整个数据挖掘的过程，包括最开始的制定业务目标到最终的结果分析，利用数据挖掘（Data Mining）来描述使用挖掘算法进行数据挖掘的子过程，在这一过程中，数据挖掘工具的选择变得越来越重要。

IBM SPSS Modeler 强大的数据挖掘功能将复杂的统计方法和机器学习技术应用到数据当中，帮助客户揭示了隐藏在交易系统、企业资源计划、结构数据库和普通文件中的模式和趋势，让客户始终站在行业发展的前端，IBM 公司于 2009 年收购了 SPSS 数据分析软件公司，并将其和 Clementine 数据挖掘软件进行整合，且将 Clementine 更名为 IBM SPSS Modeler，再次推向全球市场，本书介绍的是 15.0 版本，也是目前的最新版本。

作为一个数据挖掘平台，Modeler 结合商业技术可以快速建立预测性模型，进而应用到商业活动中，帮助人们改进决策过程。同那些仅仅着重于模型的外在表现而忽略了数据挖掘在整个业务流程中应用价值的其他数据挖掘工具相比，Modeler 功能强大的数据挖掘算法，使数据挖掘贯穿于业务流程的始终，在缩短投资回报周期的同时极大地提高了投资回报率。

编写特点

目前，市场上大多数的 SPSS Modeler 图书基本上还是按照较早版本的 Clementine 编写，而且大多是理论性的介绍，没有结合具体的案例进行深入分析。

本书与同类书籍相比，安排了较多的实例，并具有以下优势。

- 理论：解决案例所涉及的理论知识和算法，SPSS Modeler 作为数据挖掘的工具毕竟

不是智能化的，需要了解工具的内在理论和逻辑，才能更有效地进行数据挖掘。

- 实例分析：使用数据挖掘理论对案例进行分析，找出解决问题的技术路线，帮助读者从解决问题的角度进行思考。

面向读者

本书由数据挖掘与分析研究人员编写，书中实例都具有很高的参考价值。本书对于高校理工学科、经济金融学科及数量分析方面的学生，以及数据挖掘和分析方面的研究人员和从业人员等，具有很强的可读性、可操作性与可使用性，尤其适合商业销售、经济管理、社会研究和人文教育等行业的相关人员阅读。

资源下载

提供本书资源文件下载，下载地址： http://pan.baidu.com/s/1eQEedKA 。

致谢

本书是编者近年来使用 SPSS Modeler 的经验汇总与提炼，在写作过程中，得到了编者领导、同事、老师、同学以及朋友的帮助，借本书出版之际，向他们表示诚挚的感谢！

最后还要特别感谢清华大学出版社的支持，以及各位编辑热情细致的工作。由于作者水平有限，书中难免会出现不足和错误，敬请广大读者批评与指正。

作 者

2014 年 9 月于上海

目 录

第 1 部分 数据挖掘篇

第 2 部分 文本挖掘篇

第 1 部分

数据挖掘篇

数据挖掘（Data Mining）是近年来数据库应用技术中相当热门的议题，看似神奇，实际上也不是什么新概念，因其所用之诸如预测模型、数据分割、链接分析（Link Analysis）、偏差侦测（Deviation Detection）等，美国早在第二次世界大战前就已运用在人口普查及军事等方面。

近年来，数据挖掘在各领域的应用非常广泛，只要该产业拥有极具分析价值与需求的数据仓储或数据库，皆可利用 Mining 工具进行有目的的挖掘与分析。一般较常见的应用案例大多发生在零售业、制造业、金融业、保险业、通信业以及医疗服务业等。

在销售数据中挖掘顾客的消费习惯，由交易记录找出顾客偏好的产品组合，包括找出流失顾客的特征与推出新产品的时机点等都是零售业常见的实例；直效营销强调的分众概念与数据库营销方式在导入数据挖掘的技术后，使直效营销的发展性更为强大，例如利用数据挖掘分析顾客群的消费行为与交易记录，结合基本数据，并依其对品牌价值等级的高低来区分顾客，进而达到差异化营销的目的；制造业对数据挖掘的需求大多运用在品质控管方面，在制造过程中找出影响产品品质最重要的因素，以期提高作业流程的效率。

近年来电信公司、信用卡公司、保险公司以及股票交易商对于诈欺行为的侦测都很有兴趣，这些行业每年因为诈欺行为而造成的损失都非常可观，数据挖掘可以从一些信用不良的客户数据中找出相似特征并预测可能的欺诈交易，达到减少损失的目的。金融业可以利用数据挖掘来分析市场动向，并预测个别公司的营运以及股价走向。数据挖掘的另一个独特的用途是在医疗业，用来预测手术、用药、诊断或流程控制的效率。

尤其是近年来数据挖掘在 CRM 中的应用，CRM（Customer Relationship Management，客户关系管理）是近来引起热烈讨论与高度关切的议题，尤其是在直效营销的崛起与网络快速发展的带动下，若跟不上 CRM 的脚步就如同跟不上时代。CRM 不是设一个客服专线就算了，更不只是把一堆客户的基本数据输入计算机就行，完整的 CRM 运行机制在相关的硬软件系统能健全地提供支持之前，有太多的数据准备工作与分析需要推动。企业通过数据挖掘可以分别针对策略、目标定位、操作效能与测量评估等 4 个方面的相关问题，有效地从市场与顾客所搜集、累积的大量数据中挖掘出对消费者而言最关键、最重要的答案，并赖以建立真正由客户需求点出发的客户关系管理。

在本部分，我们首先介绍数据挖掘的基本知识，随后重点介绍各种数据挖掘模型的基本理论和 SPSS Modeler 的具体应用实例，包括回归分析、时间序列分析、因子分析、决策树分析、判别分析、聚类分析、人工神经网络、贝叶斯网络以及社交网络分析等模型，其中实例是本部分的重点，也是本书的重点。

第 1 章

数据挖掘概述

学习要点

- 了解数据挖掘的基本概念。
- 理解数据挖掘与传统技术的关系。
- 了解几种常用的数据挖掘软件。
- 熟悉使用软件进行数据挖掘的基本步骤。

数据挖掘对于大多数人来说都是一个陌生的事物，让读者在较短的时间内快速熟悉它就是本书第 1 章的主要任务，我们将会从最基础的知识开始讲起，由浅入深，逐步介绍数据挖掘的知识，最后结合实例进行详细阐述。

本章首先介绍数据挖掘的一些基本知识，包括数据挖掘定义、发展历史、技术特征，以及与传统技术的关系比较，同时还介绍了三种主要的数据挖掘软件。随后重点介绍了使用 SPSS Modeler 软件对电信行业的目标客户进行深入研究的基本步骤，主要是为了让读者了解什么是数据挖掘以及如何进行数据挖掘等问题。

1.1 什么是数据挖掘

数据挖掘是一种通过数理模式来分析大量资料，以找出不同的客户或市场划分，分析出消费者喜好和行为的方法。数据挖掘可以描述为：是按企业既定业务目标，对大量的企业数据进行探索和分析，揭示隐藏的、未知的或验证已知的规律性，并进一步将其模型化的先进、

有效的方法。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中的一个步骤。数据挖掘一般是指从大量的数据中自动搜索隐藏于其中的有着特殊关系性信息的过程。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（Data Mining）在技术上的定义就是从大量的、不完全的、有噪声的、模糊的、随机的实际应用数据中，提取隐含在其中的、人们事先不知道的，但又是潜在有用信息和知识的过程。与数据挖掘相近的同义词有数据融合、数据分析和决策支持等。这个定义包括多层含义：数据源必须是真实的、大量的、含噪声的；发现的是用户感兴趣的知识；发现的知识要可接受、可理解、可运用；并不要求发现放之四海皆准的知识，仅支持特定的问题。数据挖掘在商业角度的定义是一种新的商业信息处理技术，其主要特点是对商业数据库中的大量业务数据进行抽取、转换、分析和其他模型化处理，从中提取辅助商业决策的关键性数据。由此可见，数据挖掘是一种深层次的数据分析方法。

1.1.1　数据挖掘的定义

数据挖掘又称数据库中的知识发现（KDD），是目前人工智能和数据库领域研究的热点问题，数据挖掘是一种决策支持过程，它主要基于人工智能、机器学习、模式识别、统计学、数据库、可视化技术等，高度自动化地分析企业的数据，做出归纳性的推理，从中挖掘出潜在的模式，帮助决策者调整市场策略，减少风险，做出正确的决策。

数据挖掘是通过分析每个数据，从大量数据中寻找其规律的技术，主要有数据准备、规律寻找和规律表示 3 个步骤。

- 数据准备是从相关的数据源中选取所需的数据并整合成用于数据挖掘的数据集。
- 规律寻找是用某种方法将数据集所含的规律找出来。
- 规律表示是尽可能以用户可理解的方式（如可视化）将找出的规律表示出来。

并非所有的信息发现任务都被视为数据挖掘。例如，使用数据库管理系统查找个别的记录，或通过因特网的搜索引擎查找特定的 Web 页面，则是信息检索领域的任务。虽然这些任务是重要的，可能涉及复杂的算法和数据结构，但是它们主要依赖传统的计算机科学技术和数据的明显特征来创建索引结构，从而有效地组织和检索信息。尽管如此，数据挖掘技术也已用来增强信息检索系统的能力。

当前，DMKD（数据挖掘和知识发现）研究方兴未艾，其研究与开发的总体水平相当于数据库技术在上个世纪 70 年代所处的地位，迫切需要类似于关系模式、DBMS 系统和 SQL 查询语言等理论和方法的指导，才能使 DMKD 的应用得以普遍推广。预计在本世纪，DMKD 的研究还会形成更大的高潮。研究焦点可能会集中到以下几个方面：发现语言的形式化描述；寻求数据挖掘过程中的可视化方法；研究在网络环境下的数据挖掘技术（Web Mining）；加