Broadview
www.broadview.com.cn

CDA数据分析师系列丛书

SPSS Modeler+Weka
数据挖掘从入门到实战

经管之家 主编　李御玺 唐绍祖 马伯 曾珂 编著

面向商业数据挖掘建模分析人员的教材
从具体的商业数据分析案例入手，轻松掌握数据挖掘的目的、方法、工具。

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

SPSS Modeler+Weka 数据挖掘从入门到实战

经管之家 主编　李御玺 唐绍祖 马伯 曾珂 编著

中国科学技术大学 图书馆藏

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

内 容 简 介

本书是一本面向商业数据挖掘建模分析人员的教材，从具体的商业数据分析案例入手，帮助读者掌握数据挖掘的目的、方法、工具与分析步骤。本书所采用的分析工具为目前颇受好评的IBM SPSS Modeler及开源软件Weka。IBM SPSS Modeler有很好的用户接口，也有不错的分析功能，但不具有前沿技术的分析模块，以及很难与现有的信息系统结合，而Weka恰能弥补其不足。同时，这两个软件都不需要编程，适合初学者。本书具体内容由四位活跃在数据挖掘教学和项目开发领域的人员完成，内容侧重软件的实际操作。本书力图以浅显的方式解释复杂的技术原理，尽量避免涉及过多的数学内容。

未经许可，不得以任何方式复制或抄袭本书之部分或全部内容。
版权所有，侵权必究。

图书在版编目（CIP）数据

SPSS Modeler+Weka 数据挖掘从入门到实战 / 经管之家主编；李御玺等编著. —北京：电子工业出版社，2019.5
（CDA 数据分析师系列丛书）
ISBN 978-7-121-31911-2

Ⅰ. ①S… Ⅱ. ①经… ②李… Ⅲ. ①统计分析—应用软件 Ⅳ. ①C819

中国版本图书馆 CIP 数据核字（2017）第 137107 号

策划编辑：张慧敏
责任编辑：王　静
印　　刷：北京京师印务有限公司
装　　订：北京京师印务有限公司
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：787×980　1/16　印张：17.5　字数：398 千字
版　　次：2019 年 5 月第 1 版
印　　次：2019 年 5 月第 1 次印刷
定　　价：69.00 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。

前　言

感谢您选择《SPSS Modeler+Wake 数据挖掘从入门到实战》。本书内容源于李御玺教授的数据挖掘相关课程讲义，讲义历经多次修改，逐渐适合作为数据挖掘实用教材，并在获得学员们的高度评价后再被编辑成书。本书的另一位作者常国珍也长期活跃在数据挖掘的项目实施和培训中，2014 年其与李教授相识，并与李教授对出版本书之事一拍即合。

读者对象

本书的撰写采取了算法与软件实操双向并行的策略。在理论上，本书尽量用例子来说明数据挖掘算法背后的理论及意义，避免艰涩的数学公式，以求读者能用最简单的方式理解理论的精髓。在软件实操上，本书以各领域的实用案例为基础，逐步地将软件的功能引出，以求读者能了解软件功能的使用场景。有了坚实的理论基础及软件操作能力，再辅之以众多的实用案例，本书的读者就能逐步进入多姿多彩的数据挖掘世界。本书是以读者第一次接触数据挖掘为前提来撰写的。读者若有数据库、统计及计算机基础，则学习起来会较为轻松。

工具介绍

IBM SPSS Modeler 可谓商业数据挖掘领域的“重型武器”，其功能全面、算法安全可靠、追求执行效率与操作上的简单易用，并被广泛运用于许多企业中。其缺点是缺乏前沿的分析模块及很难与现有的信息系统结合，而开源软件 Weka 恰能弥补其不足。Weka 简单好用，拥有许多前沿的分析模块并易于与现有的信息系统整合。其缺点是在数据预处理部分，便利性不如 IBM SPSS Modeler 简单、易用。这两个软件对初入数据分析领域的读者而言是很好的入门工具。

阅读指南

本书分为 15 章。第 1 章介绍数据挖掘的起源及应用。同时说明如何建立一个 SPSS Modeler 及 Weka 的项目。第 2 章介绍数据挖掘的方法论 CRIPS-DM。同时说明如何将数据汇入 SPSS Modeler 及 Weka 的项目中，并做初步的数据探索。第 3 章介绍基本的数据挖掘技术。同时说明如何利用 SPSS

Modeler 及 Weka 建立 KNN 模型并进行分类预测。第 4 章介绍数据挖掘的进阶技术、数据挖掘技术的绩效增益及两个重要的数据挖掘网站。第 5 章详细介绍数据预处理技术，同时说明如何利用 SPSS Modeler，针对银行的信用风险评估数据，进行数据预处理。第 6 章介绍如何有效地挖掘对项目有帮助的关键变量。同时说明如何利用 SPSS Modeler 及 Weka，挖掘有效变量。第 7 至 15 章则为数据挖掘模型的介绍。这些模型均为热门且应用最为广泛的模型。对于每个模型的介绍，先以实例说明其理论，随后以实用的案例介绍如何在 SPSS Modeler 及 Weka 中操作，让每个读者充分了解每个模型的实际运用效果。

如果时间允许，则读者可以采取通读本书内容并按照示例进行操作的方式，但是这样效率可能不高。更高效的方法是结合工作中遇到的问题，先集中精力把书上的示例操练好，然后带入工作中的实际数据实现同样的算法，最后修改部分设置，以满足工作中的特定需求。

本书特点

本书作为市场上为数不多的理论与软件实操相结合并面向商业数据挖掘的书籍，和其他统计软件图书有很大的不同，本书结构新颖，案例贴近实际，讲解深入透彻。

- **场景式设置**

本书从银行、电信、零售、医疗等行业中精心归纳、提炼出各类数据挖掘案例，方便读者搜寻与实际工作相似的问题。

- **启发式描述**

本书注重培养读者解决问题的思路，以最朴实的思维方式结合启发式的描述，帮助读者发现规律、总结规律和运用规律，从而启发读者快速找出问题的解决方法。

售后服务

尽管作者们对书中的案例精益求精，但疏漏之处在所难免，如果发现书中的错误或某个案例有更好的解决方案，则敬请与本书作者联系，作者邮箱为 leeys@mail.mcu.edu.tw。

学习方法

只有对数据分析的流程熟悉了，才能实现从模仿到灵活运用。在产品质量管理方面，对流程的掌控是成功的关键，在数据挖掘项目中，流程同样是重中之重。数据挖掘是一个先后衔接的过程，一个步骤的失误会带来完全错误的结果。数据挖掘的流程大致包括抽样、数据清洗、数据转换、建模和模型评估这几个步骤。如果在抽样中的取数逻辑不正确，就有可能使因果关系倒置，得到完全

此为试读，需要完整PDF请访问：www.ertongbook.com

相反的结论。数据转换方法如果选择不正确，模型就难以得到预期的结果。而且，数据挖掘是一个反复试错的过程，每一步都要求有详细的记录和操作说明，否则分析人员很可能迷失方向。

学习数据挖掘最好的方法就是动手做一遍。本书语言通俗，但高度凝练，很少涉及公式，这会让读者大意，如果读者不动手做一遍，则很难体会到书中表述的思想。本书提供了相应的演练数据，也同时给出了相关方面的参考资料，供学员学习。

致谢

本丛书从策划到出版，张慧敏主编倾注了大量心血，经管之家的董事长赵坚毅先生提供了多方面的支持，特在此表示衷心的感谢！

为保证丛书的质量，使其更贴近读者，我们邀请了北京大学的殷子涵进行试读和修改完善。感谢各位预读员的辛勤、耐心与细致，使得本书能以更加完善的面目与各位读者见面。还要感谢刘莎莎参与本出的编写工作。

再次感谢您的支持！

作　者

读者服务

轻松注册成为博文视点社区用户（www.broadview.com.cn），扫码直达本书页面。

- **提交勘误**：您对书中内容的修改意见可在 提交勘误 处提交，若被采纳，将获赠博文视点社区积分（在您购买电子书时，积分可用来抵扣相应金额）。
- **交流互动**：在页面下方 读者评论 处留下您的疑问或观点，与我们和其他读者一同学习交流。

页面入口：http://www.broadview.com.cn/31911

目　录

第 1 篇　理论篇

第 2 篇　准备篇

此为试读，需要完整PDF请访问：www.ertongbook.com

第 3 篇 关系篇

第1篇　理论篇

第1章

数据挖掘简介

近年来，信息产业高速发展，人们越来越关注如何将信息转换成有用的、直观的知识。因此，在1991年，William Frawley和Gregory Piatetsky Shapiro提出了数据挖掘的概念，即从现有的大量数据中，撷取不明显的、之前未知的、可能有用的知识的过程。数据挖掘的目标是建立一个决策模型，根据过往的行动来预测未来的行为。例如，分析一家公司的不同客户对公司产品的购买情况，进而分析出哪一类客户会对公司的产品有兴趣。在讲究实时、竞争激烈的网络时代，若能事先破解消费者的行为模式，将是公司获利的关键因素之一。

1.1 数据挖掘的起源、定义及目标

数据挖掘可以从数据中撷取不明显的、之前未知的信息，举一个著名的例子：美国的沃尔玛超市为了能够准确了解顾客在其门店的购买习惯，对其顾客的购物行为进行数据挖掘。一个意外的发现是：跟尿布一起销售最多的商品竟是啤酒！在美国，一些年轻的父亲下班后经常要到超市去买婴儿尿布，而他们中有 30%～40%的人同时也为自己买一些啤酒。产生这一现象的原因是：美国的太太们常叮嘱她们的丈夫在下班后为小孩买尿布，而丈夫们在买尿布后又随手带回了他们喜欢的啤酒。既然尿布与啤酒被一起购买的机会很多，于是，沃尔玛就在其一个个门店中将尿布与啤酒并排摆放在一起，结果是尿布与啤酒的销售量双双增长。按照常规思维，尿布与啤酒风马牛不相及，若不是借助数据挖掘技术对大量交易数据进行分析，沃尔玛是不可能发现这个有价值的规律的。

与此同时，数据挖掘能够发现信息之间的联系，但不一定是因果关系，在数据挖掘的过程中能发现有用的信息。比如，美国的分析员做过这样一项研究：在冰激凌热卖时，溺水的人数会增加。而冰激凌与溺水的人数之间存在的联系并不是因果关系，所以，减少冰激凌的销售量不一定会降低溺水的人数。而二者之所以会存在联系，是因为冰激凌在夏天热卖，而夏天去游泳的人数增加，溺水的人数也会增加。所以，通过这个例子可以看出，数据挖掘要在大量的数据中找到有用的知识，不能发现关联之后就随意下定结论。

1.2 数据挖掘的发展历程

数据挖掘虽然是计算机应用领域的新名词，但也经历了几十年的发展历程。

- 第一阶段：1960 年以后，各种新兴的数据收集模式开始出现，例如磁带、软盘、硬盘等，人们开始掌握了收集数据的基本方法。
- 第二阶段：1980 年以后，随着收集的数据量的增多，人们开始需要数据库，并逐渐建立起了数据库，但是此时还不能查询数据。
- 第三阶段：1990 年以后，数据统计的概念出现，人们可以进入数据仓库完成简单的数据统计，但并不能做太精细的决策。
- 第四阶段：2000 年以后，随着数据库和计算机网络的广泛应用，加上使用先进的自动数据生成和采集工具，人们所拥有的数据量急剧增大。针对大规模数据的分析处理方法——数据挖掘出现了。

数据挖掘在各行各业中都有应用，比如其最早应用于银行、通信业，现在也在零售业、保险业及政府中有所应用（见图 1-1）。

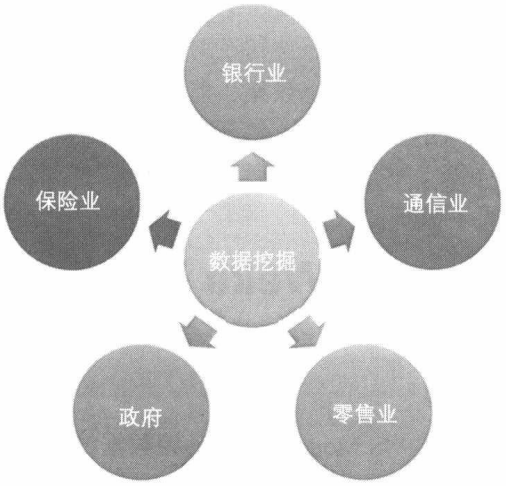

图 1-1

1．银行

金融事务需要收集和处理大量数据，由于银行在金融领域的地位、工作性质及业务特点，市场竞争激烈程度决定了它对信息化、电子化的需求比其他领域更迫切。利用数据挖掘技术可以帮助银行产品开发部门描述客户以往的需求并预测未来。例如，汇丰银行对不断增长的客户群进行分类，为每种产品找到最有价值的客户，这样其产品才能推销得好，而且比盲目推销产品节省了 30%的销售费用。再例如，银行通过数据挖掘发现，有盗刷信用卡行为的人，其使用信用卡的第一笔消费往往小于 10 元，所以，银行根据这条规律，冻结了那些第一笔消费小于 10 元的客户的账号，减少了客户盗刷信用卡所带来的经济损失。

2．零售业

在过去，零售商依靠供应链软件、内部分析软件甚至直觉来预测库存需求。随着竞争压力一天天地增大，很多零售商都开始致力于找到更准确的方法来预测其连锁商店应保有的库存。通过数据挖掘可以为产品存储决策提供准确、及时的信息。

3．保险业

对受险人员进行分类有助于确定适当的保险金额度。通过数据挖掘可以得到不同行业、不同年龄段、不同社会层次的人的信息，从而可以评估他们的保险金。另外，还可进行保险金种类关联分析，分析购买了某种保险的人是否又同时购买了另一种保险，也可预测什么样的顾客会购买新险种。总而言之，数据挖掘在保险业中有广泛的应用。

4．政府

数据挖掘被广泛应用于电子政务中的综合查询、经济分析、宏观预测、应急预警、风险分析及预警、质量监督管理及监测、决策支持等系统，它为公众提供了一个智能化、高效的网上政府。例

如，几乎每个政府网站都有类似“公众意见调查”的栏目，这是了解公众需求的一个很好的途径，但是从网站公布的调查结果看，其结论大多还停留在对单个问题求总数、求比例等简单分析上。利用数据挖掘技术可以在网上建立一个能有效地收集、监测和分析公众数据的系统，提炼出实用、有效的信息，建成以公众需求为主导的电子政务。将数据挖掘技术引入电子政务，可大大提高整个电子政务系统的智能化水平。

1.3 SPSS Modeler 和 Weka 基础操作

1.3.1 SPSS Modeler 软件简介

IBM SPSS Modeler（以下简称 SPSS Modeler）是一组数据挖掘工具，通过它可以快速建立预测模型，并将其应用于商业活动中，从而改进决策过程。

使用 SPSS Modeler 处理数据主要分为 3 个步骤。

- 首先，将数据读入 SPSS Modeler 中。
- 然后，通过一系列操作运行数据。
- 最后，将数据发送到目标位置。

这个操作过程被称为数据流，因为数据以一条条记录的形式，依次经过各种操作，最终到达目标位置（模型或某种数据输出）。

数据流工作区是 SPSS Modeler 窗口中最大的区域，也是构建和操作数据流的区域，如图 1-2 所示。

SPSS Modeler 中的大部分数据和建模工具位于节点选项卡中，该选项卡位于数据流工作区（简称工作区）的底部（见图 1-3）。要将节点添加到工作区中，在节点选项卡中双击节点对应的图标或将其拖曳到工作区中即可。随后可将各个图标连接以创建一个数据流。每个选项卡中均包含一组不同的数据流操作阶段中使用的相关节点，例如：

- 源（Source）节点。此类节点可将数据引入 SPSS Modeler 中。
- 记录选项（Record Ops）节点。此类节点可对数据记录执行操作，例如选择、合并和追加等。
- 字段选项（Field Ops）节点。此类节点可对数据字段执行操作，例如过滤、导出新字段和确定给定字段的测量级别等。
- 图形（Graphs）节点。此类节点可在建模前后以图表形式显示数据。图表形式包括散点图、直方图、网络节点和评估图表。
- 建模（Modeling）节点。此类节点可使用 SPSS Modeler 中提供的建模算法，例如神经网络、决策树、聚类算法和数据排序等。
- 输出（Output）节点。此类节点可生成能在 SPSS Modeler 中查看的数据、图表和模型等多种输出结果。

此为试读，需要完整PDF请访问：www.ertongbook.com