信息管理与信息系统创新应用系列教材

商务智能与数据挖掘实验教程

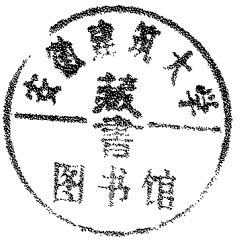
朱慧云 曹玲 编著

科学出版社

信息管理与信息系统创新应用系列教材

商务智能与数据挖掘实验教程

朱慧云　曹　玲　编著



科学出版社

内 容 简 介

本书综合经济管理专业知识和商务智能、数据挖掘模型开发于一体，结合商业背景设计若干实践项目，全面阐述使用 IBM SPSS Modeler、Weka、RapidMiner 等软件进行数据分析与挖掘的原理、方法和步骤，介绍社会网络分析软件 UCINET 与文献可视化分析软件 CiteSpace 的使用，紧密结合理论教学，使学生在有限的实验课时中，加深对所学知识的理解和掌握。目前国内商务智能与数据挖掘实验指导教程的相关书籍不多，结合商业背景的更是稀少，本书强调数据挖掘在商业决策领域中的应用，弥补大多数同类书籍商业应用不足的缺点。

本书可作为经管类专业本科生、研究生的实验教材，也可在 MBA、EMBA 教学和企业培训中使用，还可供从事商务智能与数据挖掘相关工作的专业人员参考。

图书在版编目（CIP）数据

商务智能与数据挖掘实验教程/朱慧云，曹玲编著. —北京：科学出版社，2017

（信息管理与信息系统创新应用系列教材）

ISBN 978-7-03-055281-5

Ⅰ. ①商… Ⅱ. ①朱…②曹… Ⅲ. ①数据处理-教材 Ⅳ. ①TP274

中国版本图书馆 CIP 数据核字（2017）第 278054 号

责任编辑：惠 雪 曾佳佳/责任校对：彭 涛

责任印制：张克忠/封面设计：许 瑞

科学出版社出版

北京东黄城根北街16号

邮政编码：100717

http://www.sciencep.com

三河市书文印刷有限公司印刷

科学出版社发行 各地新华书店经销

*

2017 年 11 月第 一 版 开本：720×1000 B5

2017 年 12 月第二次印刷 印张：11 3/4

字数：237 000

定价：59.00 元

（如有印装质量问题，我社负责调换）

前　言

随着计算能力和互联网技术的迅猛发展，不断递增的海量数据集使得传统数据分析工具变得力不从心。针对如何面对海量数据信息的挑战，商务智能与数据挖掘技术得到了快速的发展。它能够有效地组织这些海量数据，洞悉数据的蛛丝马迹，发现数据的潜在价值，预测数据的发展趋势，从而帮助预测分析和制定决策。因此，加强商务智能与数据挖掘领域的理论与实践学习，现已成为经管类专业特别是信息管理专业学生的必修内容。

本实验教程通过大量的实例，从广为人知的商业软件 IBM SPSS Modeler，到开源软件 Weka 和 RapidMiner，再到社会网络分析软件 UCINET、文献可视化分析软件 CiteSpace，循序渐进地引导学生做好各章的实验。第一部分着重介绍 IBM SPSS Modeler 数据挖掘软件的基本操作和使用方法，购物篮分析、客户细分和客户分类三个数据挖掘的经典实例，使学生熟悉 IBM SPSS Modeler 中关联分析、聚类分析、分类分析等功能。第二部分介绍两种功能强大并较为常用的数据挖掘开源软件，Weka 和 RapidMiner。通过学习与应用 Apriori 算法、决策树算法、划分方法中 K 均值算法分别对数据集进行关联规则挖掘、分类、聚类分析来熟悉这两个软件。第三部分介绍社会网络分析软件 UCINET 和文献可视化分析软件 CiteSpace，使学生了解 UCINET、CiteSpace 软件的基本操作和应用环境，进行科研合作网络特征的社会网络分析、文献信息可视化分析、绘制知识图谱的基本流程。此外，还给出了 IBM SPSS Modeler、Weka、RapidMiner 等软件的数据和文件的下载链接(请访问网址：http://www.ecsponline.com/，选择“网上书店”，检索图书书名，在图书详情页面“页源下载”栏目中获取)，以便于读者学习和使用。

通过学习并应用实验教程中的内容，学生能够更深刻地掌握相关的商务智能与数据挖掘基础理论知识，熟悉社会网络分析与可视化软件的使用，同时也提高

了利用理论知识解决实际问题的能力，使得单一的知识传授型教学转变为素质教育为主的实践式教学。与此同时，每章末还留有复习思考题，为学生留下巩固复习、启发思考的空间。

本书由朱慧云、曹玲编写，其中朱慧云负责第1～8章的编写，曹玲负责第9、10章的编写。编者总结多年教学过程中的实践经验，对传统的课程进行改革，整合成本书。本书获得江苏省高校品牌专业建设工程项目(编号：1181181601002)资助，在此表示感谢。

由于编者水平有限，书中难免存在不妥之处，在此恳切地希望广大读者进行批评指正。

作　者

2017年9月

目　录

第二部分　开源数据挖掘软件使用篇

第三部分　社会网络分析与可视化软件使用篇

第一部分

IBM SPSS Modeler 软件使用篇

第 1 章　IBM SPSS Modeler 软件使用基础

1.1　实 验 目 的

(1) 了解 IBM SPSS Modeler 数据挖掘软件的基本操作和环境。

(2) 初步掌握使用 IBM SPSS Modeler 的不同节点导入不同格式存储的数据。

(3) 熟悉 IBM SPSS Modeler 提供的图形节点，通过对数据的可视化展示了解数据类型和数据分布。

1.2　背 景 知 识

1) IBM SPSS Modeler

SPSS Modeler 是一款商业数据挖掘软件，能够为个人、团队、系统和企业做决策提供预测性智能。它可提供各种高级算法和技术 (包括文本分析、实体分析、决策管理与优化)，快速建立预测性模型，并将其应用于商业活动，从而改进决策过程[1]。

借助 SPSS Modeler，您可以使用各种分析技术访问数据源，如数据仓库、数据库、Hadoop 分布或平面文件，以便从您的数据中发现隐含的模式。这些统计技术使用历史数据来预测当前状况或未来事件。这些统计技术还包括数据访问、数据准备、数据建模和交互可视化功能。

SPSS Modeler 在提供大量强大且稳健的数据挖掘模型供分析人员使用的同时保持非常友好的易用性，提供图形化的操作环境，使用鼠标即可完成数据挖掘全过程，降低了入门要求，减少了学习时间。

2) *数据挖掘方法论*——CRISP-DM

SPSS Modeler 根据 CRISP-DM(cross-industry standard process for data mining) 即“跨行业数据挖掘标准流程”设计[2]。CRISP-DM 模型将一个数据挖掘流程分为六个不同的，但顺序并非完全不变的阶段。这六个阶段分别是：

(1) 商业理解 (business understanding)。从商业的角度了解项目的要求和最终目的是什么，并将这些目的与数据挖掘的定义以及结果结合起来。

(2) 数据理解 (data understanding)。数据理解阶段开始于数据的收集工作。

接下来就是熟悉数据的工作，具体如：检测数据的量，对数据有初步的理解，探测数据中比较有趣的数据子集，进而形成对潜在信息的假设。收集原始数据，对数据进行装载，描绘数据，并且探索数据特征，进行简单的特征统计，检验数据的质量，包括数据的完整性和正确性，缺失值的填补等。

(3) 数据准备(data preparation)。数据准备阶段涵盖了从原始粗糙数据中构建最终数据集 (将作为建模工具的分析对象) 的全部工作。数据准备工作有可能被实施多次，而且其实施顺序并不是预先规定好的。这一阶段的任务主要包括：制表，记录，数据变量的选择和转换，以及为适应建模工具而进行的数据清理等。

(4) 建模 (modeling)。在这一阶段，各种各样的建模方法将被加以选择和使用，通过建造、评估模型将其参数校准为最理想的值。比较典型的是，对于同一个数据挖掘的问题类型，可以有多种方法选择使用。如果有多重技术要使用，那么在这一任务中，对于每一个要使用的技术要分别对待。一些建模方法对数据的形式有具体的要求，因此，在这一阶段，重新回到数据准备阶段执行某些任务有时是非常必要的。

(5) 评估 (evaluation)。从数据分析的角度考虑，在这一阶段中，已经建立了一个或多个高质量的模型。但在进行最终的模型部署之前，需要更加彻底地评估模型，回顾在构建模型过程中所执行的每一个步骤，是非常重要的，这样可以确保这些模型达到企业的目标。一个关键的评价指标就是，是否仍然有一些重要的企业问题还没有被充分地加以注意和考虑。在这一阶段结束之时，有关数据挖掘结果的使用应达成一致的决定。

(6) 部署 (deployment)。部署，即将其发现的结果以及过程组织成为可读文本形式。模型的创建并不是项目的最终目的。尽管建模是为了增加更多有关于数据的信息，但这些信息仍然需要以一种客户能够使用的方式被组织和呈现。这经常涉及一个组织在处理某些决策过程中，如在决定有关网页的实时人员或者营销数据库的重复得分时，拥有一个“活”的模型。

根据需求的不同，部署阶段可以是仅仅像写一份报告那样简单，也可以像在企业中进行可重复的数据挖掘程序那样复杂。在许多案例中，往往是客户而不是数据分析师来执行部署阶段。然而，尽管数据分析师不需要处理部署阶段的工作，对于客户而言，预先了解需要执行的活动，从而正确地使用已构建的模型是非常重要的。

3) *数据流*

使用 SPSS Modeler 处理数据有三个步骤。首先，将数据读入 SPSS Modeler，

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。这一操作序列称为数据流，因为数据以一条条记录的形式，从数据源开始，依次经过各种操作，最终到达目标 (模型或某种数据输出)(图 1-1)[3]。

图 1-1　一个简单数据流

1.3　实 验 内 容

(1) 初步认识 SPSS Modeler 软件，了解软件的主窗口，学习对节点的基本操作、构建数据流等。

(2) 使用 SPSS Modeler 的数据库源节点、变量文件节点等导入数据。

(3) 使用 SPSS Modeler 提供的图形节点，对数据进行可视化展示。

1.4　实 验 步 骤

1.4.1　初步认识 IBM SPSS Modeler 软件

1) IBM SPSS Modeler 主窗口

依次单击开始→所有程序→IBM SPSS Modeler 18.0→IBM SPSS Modeler 18.0，启动程序，显示 IBM SPSS Modeler 主窗口(图 1-2)。

SPSS Modeler 主窗口由菜单栏、工具栏、数据流工作区、节点选用板、管理器和项目窗口组成。

菜单栏。菜单栏位于 SPSS Modeler 主窗口顶部，包含软件的绝大多数命令。

工具栏。SPSS Modeler 主窗口顶部有一个图标工具栏，其中包含许多有用功能，如创建新流、打开现有流、运行当前流等。

数据流工作区。数据流工作区是 SPSS Modeler 窗口的最大区域，也是构建和操作数据流的位置。通过在界面的主工作区中绘制与业务相关的数据操作图表来创建流。每个操作都用一个图标或节点表示，这些节点通过流连接在一起，流表示数据在各个操作之间的流动。在 SPSS Modeler 中，可以在同一流工作区或

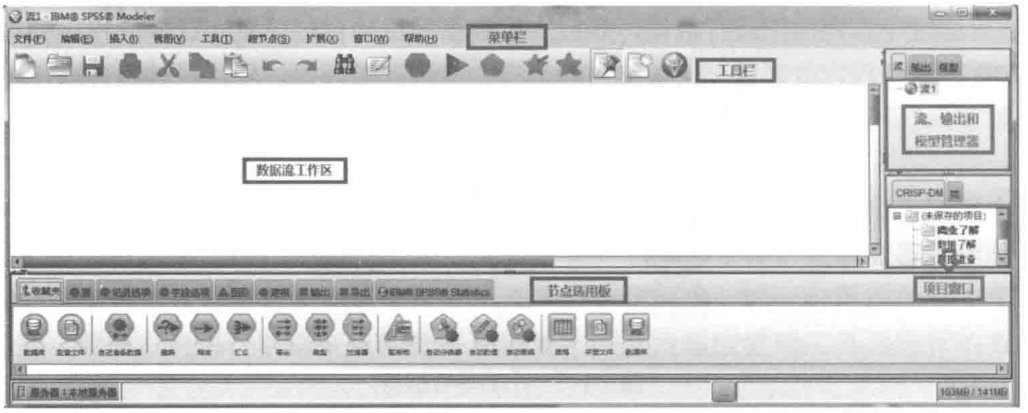


图 1-2　IBM SPSS Modeler 主窗口

通过打开新的流工作区来一次处理多个流。会话期间，流存储在 SPSS Modeler 窗口右上角的“流”管理器中。

节点选用板。SPSS Modeler 中，每个操作都用一个节点表示。SPSS Modeler 中的大部分数据和建模工具位于节点选用板中，节点选用板位于流工作区下方窗口的底部。节点选用板包括多个选项卡，每个选项卡均包含一组不同的流操作阶段中使用的相关节点。

流、输出和模型管理器。管理流、输出和模型，包括三个选项卡。可以使用“流”选项卡打开、重命名、保存和删除在会话中创建的流。“输出”选项卡中包含由 SPSS Modeler 中的流操作生成的各类文件，如图形和表格，可以显示、保存、重命名和关闭此选项上列出的表格、图形和报告。“模型”选项卡是管理器选项卡中功能最强大的选项卡，该选项卡中包含所有模型块，这些模型块是针对当前会话在 SPSS Modeler 中生成的模型。这些模型可以直接从“模型”选项卡上浏览或将其添加到工作区的流中。

项目窗口。窗口右侧底部是项目窗口，用于创建和管理数据挖掘项目。“CRISP-DM”选项卡提供了一种组织项目的方式。“类”选项卡提供了一种在 SPSS Modeler 中按类别 (即按照所创建对象的类别) 组织工作的方式。

2) 节点

源节点。使用源节点能够导入以多种格式存储的数据，这些格式包括平面文件、IBM SPSS Statistics (.sav)、SAS、Microsoft Excel 和 ODBC 兼容关系数据库，也可以使用用户输入节点生成综合数据。

记录选项节点。此类节点对数据记录执行操作，如选择记录、合并记录等。

字段选项节点。此类节点对数据字段执行操作，如过滤字段、导出新字段等。

图形节点。此类节点在建模前后以图表形式显示数据。

建模节点。SPSS Modeler 提供了多种借助机器学习、人工智能和统计学的建模算法。建模节点提供了这些算法，使用这些算法可以根据数据生成新的信息以及开发预测模型。每种算法各有所长，同时适用于解决特定类型的问题。

输出节点。此类节点生成可在 SPSS Modeler 中查看的数据、图表和模型等多种输出结果。

导出节点。此类节点生成可在外部应用程序中查看的多种输出结果。

IBM SPSS Statistics 节点。此类节点从 IBM SPSS Statistics 中导入数据或将数据导出到其中，并用于运行 IBM SPSS Statistics 过程。

3) 对节点的操作

(1) 增加节点。可以采用以下三种方式在数据流区域增加一个节点：

在节点选用板上双击节点，自动放置节点到数据流区域。注意：它会自动地连接到“中心”节点。

将节点从节点选用板拖放到数据流区域中。

单击节点选用板中的节点，然后单击流工作区。

(2) 编辑节点。在节点上右击，展开一个节点，点击 “编辑”，就可以在弹出的对话框中对节点进行设置(图 1-3)。在数据流工作区双击节点也可以编辑节点。

(3) 连接节点。连接节点可以采用以下方式：

双击节点选用板上的节点，可以自动把新节点连接到数据流区域中的“中心”节点上。

选中某个节点，然后单击右键打开上下文菜单。在菜单中选择连接。此时，开始节点和光标处将同时显示连接图标(图 1-4)。单击工作区中的第二个节点以连接这两个节点。

在流工作区中，可以使用鼠标中键单击某个节点并将其拖到另一个节点。(如果鼠标没有中键，可以通过按住 Alt 键的同时使用鼠标从一个节点拖到另一个节点来模拟此操作)

(4) 删除节点之间的连接。在连接箭头上按住鼠标右键选择“删除连接”，可以删除这个连接(图 1-5)。

在节点上右击，展开一个节点，点击 “断开连接”，可以删除此节点上的全部连接。

选中节点，然后按 F3 键删除所有连接。

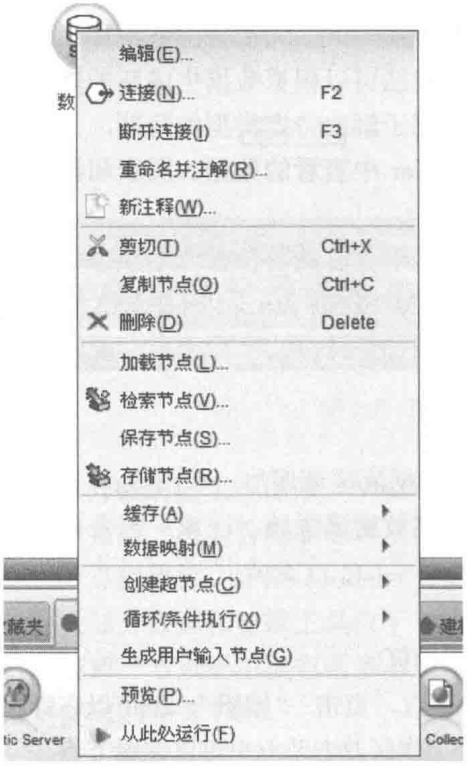


图 1-3　编辑节点

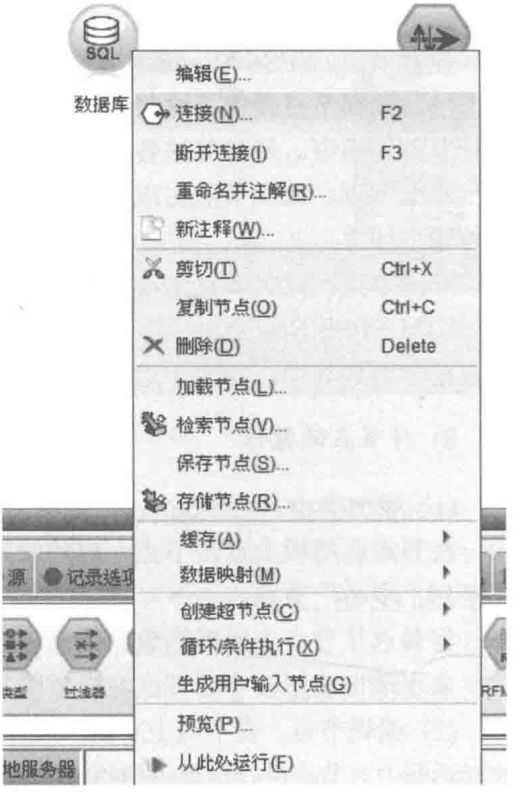


图 1-4　连接节点

图 1-5　删除节点之间的连接

(5) 删除节点。在节点上右击，展开一个节点，点击 “删除”，可以删除此节点(图 1-6)。在数据流工作区中选中节点，按下键盘上的 Delete 键，也可以删除节点。

4) *构建数据流*

使用 SPSS Modeler 进行的数据挖掘重点关注通过一系列节点运行数据的过程，这一过程被称为流。这一系列节点代表要对数据执行的操作，而节点之间的连接指示数据流的方向。通常，可以使用数据流将数据读入 SPSS Modeler，通过

图 1-6　删除节点

一系列操作运行数据，然后将其发送至某个目的地。

采用下列步骤构建数据流：

(1) 将节点添加到流工作区。

(2) 连接节点以形成流。

(3) 指定任意节点或流选项。

(4) 执行流。

1.4.2　数据导入

SPSS Modeler 包括数据库、变量文件、Excel、SAS 文件等源节点，分别用于不同格式存储的数据的导入。

1) *数据库源节点*

数据库源节点可用于使用 ODBC(开放数据库连接) 从多种其他数据包中导入数据，这些数据包包括 Microsoft SQL Server、DB2、Oracle 等。因此，要读