内容全面、讲解透彻、突出实战

IBM SPSS Modeler 18.0

数据挖掘权威指南

张浩彬 周伟珠 编著

联袂推荐

暨南大学教授、博士生导师刘建平
暨南大学研究生院副院长、经济学院统计学系
副主任、教授、博士生导师陈光慧
天善智能创始人梁勇
IBM技术专家刘咏梅
IBM资深数据科学家钟云飞
广东省环保厅环境咨询专家委员会专家
广东柯内特环境科技有限公司总经理朱斌

中国工信出版集团
人民邮电出版社
POSTS & TELECOM PRESS

IBM SPSS Modeler 18.0

数据挖掘权威指南

张浩彬 周伟珠 编著

人民邮电出版社
北京

图书在版编目（CIP）数据

IBM SPSS Modeler 18.0数据挖掘权威指南 / 张浩彬，周伟珠编著. -- 北京 : 人民邮电出版社，2019.4
ISBN 978-7-115-50759-4

Ⅰ. ①I… Ⅱ. ①张… ②周… Ⅲ. ①统计分析－软件包 Ⅳ. ①C819

中国版本图书馆CIP数据核字(2019)第022470号

内 容 提 要

本书是一本以数据挖掘应用为主导，以 SPSS Modeler 为实践框架的应用指南，内容涵盖数据挖掘方法论、数据读取、数据处理、数据可视化、统计分析与检验、数据挖掘算法、自动建模、集成与扩展、模型部署以及性能优化等，力求帮助读者全面掌握数据挖掘项目的主要内容以及实践细节。除了操作层面，本书也尽可能地把专业晦涩的数据挖掘知识及商业应用内容以通俗易懂的方式传递给读者，同时所有场景会结合 IBM SPSS 工具进行实现并提供样例学习，方便读者在学习的同时加深巩固和理解。如果你是在校学生、刚刚从事数据分析的大学毕业生、数据分析爱好者、市场营销人员、产品运营人员或者数据分析师，如果你希望提升自己的数据挖掘技术，那么就适合阅读本书。

◆ 编　　著　张浩彬　周伟珠
责任编辑　王峰松
责任印制　焦志炜

◆ 人民邮电出版社出版发行　　北京市丰台区成寿寺路 11 号
邮编　100164　　电子邮件　315@ptpress.com.cn
网址　http://www.ptpress.com.cn
固安县铭成印刷有限公司印刷

◆ 开本：787×1092　1/16
印张：29.25
字数：697 千字　　2019 年 4 月第 1 版
印数：1－2 000 册　　2019 年 4 月河北第 1 次印刷

定价：108.00 元

读者服务热线：(010)81055410　印装质量热线：(010)81055316
反盗版热线：(010)81055315
广告经营许可证：京东工商广登字 20170147 号

前　言

当人民邮电出版社编辑王峰松找到我和伟珠时，问我们要不要写一本关于 SPSS Modeler 的图书，我当时还是有一些迟疑的。第一点迟疑是，我之前已经写了一本关于数据挖掘算法的图书，虽然关于 SPSS Modeler 的操作只占了其中 100 页左右的篇幅，但作为一本彩色印刷的算法图书，这真的不少了。第二点迟疑是，正是因为写过一本书，所以才知道写作的艰辛，尤其是这类和数学及工具应用密不可分的书籍。我问王编辑，为什么还要再写一本书？王编辑反问了我一句:“你第一本书是主要讨论算法的，那你觉得你在 IBM 作为 SPSS 工程师的这段时间，关于 SPSS Modeler 的话已经说完了吗？”真是一个让人难以拒绝的反问，我内心有一个声音简直要脱口而出：“当然没有。”就这样，我和伟珠两个人就开始了这本书的写作，我们希望写一本“纯粹的 SPSS Modeler 工具书”。

不开玩笑地说，这本书在最开始写作的时候，名字就暂定为《IBM SPSS Modeler 最强工具书》。当然，这么“土”的名字通过性也不大，倒是因为这个初衷，我和伟珠两个人吃了不少苦头，因为我们是真的想把这些年关于 Modeler 的使用经验和使用感悟都写下来，告诉每一个喜欢和使用这个工具的读者。还记得，本书第一稿写完的时候一共包括了 23 章，页数一度接近 900 页。

本书现在的名字叫《IBM SPSS Modeler 18.0 数据挖掘权威指南》，是一本以数据挖掘应用为主导、以 SPSS Modeler 为实践框架的应用指南，内容涵盖数据挖掘方法论、数据读取、数据处理、数据可视化、统计分析与检验、数据挖掘算法、自动建模、集成与扩展、模型部署以及性能优化等，力求帮助读者全面掌握数据挖掘项目的主要内容以及实践细节。除了操作层面，本书也尽可能地把专业晦涩的数据挖掘知识及商业应用内容以通俗易懂的方式传递给读者，同时所有场景会结合 IBM SPSS 工具进行实现并提供样例学习，方便读者在学习的同时加深巩固和理解。简单来说，如果你是在校学生、刚刚从事数据分析的大学毕业生、数据分析爱好者、市场营销人员、产品运营人员或者数据分析师，如果你希望提升自己的数据挖掘技术，就适合阅读本书。

本书特色

本书第一个特色是“全”。作为一本“SPSS 字典”，从本书结构来看，23 章的内容中涵盖了数据挖掘方法论、数据读取、数据处理、数据可视化、统计分析与检验、数据挖掘算法、自动建模、集成与扩展、模型部署以及性能优化等，全面覆盖在数据挖掘项目中用户可能遇到的内容。

本书的第二个特色是“透”。本书的每一章内容，既包括理论的讲解，又涵盖应用的实践，而且在工具介绍上尽可能涵盖每一个选项的内容和应用形式，力求让读者“吃透”该

章节的内容。

本书的第三个特色是“重实践”。从本书的内容上看，作为数据挖掘算法与工具操作相结合的图书，实践是少不了的。更进一步的是，本书每章都附有对应的实战技巧，每个实战技巧都是由我和伟珠两个人多年的应用经验总结而成。

致谢

感谢“探数寻理”的读者关注与支持。感谢IBM大中华区分析事业部刘咏梅、龙力辉、钟云飞、秦思等多位领导及同事的帮助和建议，是你们的大力支持让本书变得更加完善。感谢刘建平教授、陈光慧教授、朱斌董事长、梁勇总经理等多位书评作者，感谢你们能够在百忙之中抽出时间阅读书稿，并提出宝贵的意见和建议。感谢人民邮电出版社编辑王峰松的大力支持和辛勤工作，让本书能够顺利出版。

联系方式和电子资源

由于作者水平有限，本书难免会出现纰漏和不足之处，恳请各位读者批评指正。如果读者有任何意见和建议，欢迎在微信中搜索“wetalkdata”并关注“探数寻理”公众号，与本书作者进行互动和沟通。

读者可以通过关注公众号，回复“指南案例数据”获取本书所有章节对应的数据文件以及数据模型文件。

资源与支持

本书由异步社区出品，社区（https://www.epubit.com/）将为您提供相关资源和后续服务。

配套资源

本书提供如下资源：

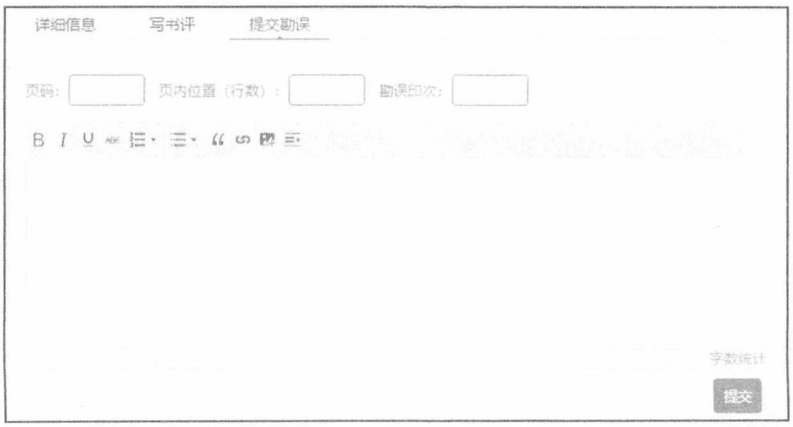
- 本书彩图；
- 数据文件以及数据模型文件。

要获得以上配套资源，请在异步社区本书页面中单击 配套资源，跳转到下载界面，按提示进行操作即可。注意：为保证购书读者的权益，该操作会给出相关提示，要求输入提取码进行验证。

提交勘误

作者和编辑尽最大努力来确保书中内容的准确性，但难免会存在疏漏。欢迎您将发现的问题反馈给我们，帮助我们提升图书的质量。

当您发现错误时，请登录异步社区，按书名搜索，进入本书页面，单击“提交勘误”标签，输入勘误信息，单击“提交”按钮即可。本书的作者和编辑会对您提交的勘误进行审核，确认并接受后，您将获赠异步社区的 100 积分。积分可用于在异步社区兑换优惠券、样书或奖品。



扫码关注本书

扫描下方二维码，您将会在异步社区微信服务号中看到本书信息及相关的服务提示。

与我们联系

我们的联系邮箱是 contact@epubit.com.cn。

如果您对本书有任何疑问或建议，请您发邮件给我们，并请在邮件标题中注明本书书名，以便我们更高效地做出反馈。

如果您有兴趣出版图书、录制教学视频，或者参与图书翻译、技术审校等工作，可以发邮件给我们；有意出版图书的作者也可以到异步社区在线提交投稿（直接访问 www.epubit.com/selfpublish/submission 即可）。

如果您是学校、培训机构或企业，想批量购买本书或异步社区出版的其他图书，也可以发邮件给我们。

如果您在网上发现有针对异步社区出品图书的各种形式的盗版行为，包括对图书全部或部分内容的非授权传播，请您将怀疑有侵权行为的链接发邮件给我们。您的这一举动是对作者权益的保护，也是我们持续为您提供有价值的内容的动力之源。

关于异步社区和异步图书

“异步社区”是人民邮电出版社旗下 IT 专业图书社区，致力于出版精品 IT 技术图书和相关学习产品，为作译者提供优质出版服务。异步社区创办于 2015 年 8 月，提供大量精品 IT 技术图书和电子书，以及高品质技术文章和视频课程。更多详情请访问异步社区官网 https://www.epubit.com。

“异步图书”是由异步社区编辑团队策划出版的精品 IT 专业图书的品牌，依托于人民邮电出版社近 30 年的计算机图书出版积累和专业编辑团队，相关图书在封面上印有异步图书的 LOGO。异步图书的出版领域包括软件开发、大数据、人工智能、软件测试、前端、网络技术等。

异步社区

微信服务号

目　录

第 1 章　IBM SPSS Modeler 基本介绍

IBM SPSS Modeler（以下简称 SPSS Modeler）是一款强大且易用的数据挖掘软件。它的设计遵循 CRISP-DM 方法论，在功能上能够覆盖整个数据挖掘生命周期的使用，不但内置丰富稳健的数据挖掘算法，而且提供了各种不同的数据处理方式以及多种生动的图形展现方式。

1.1　SPSS 简介

SPSS 最初称为“Statistical Package for the Social Sciences”，即社会科学统计软件包。1968 年，SPSS 由斯坦福 3 个学生所开发，它是世界上最早的统计分析软件。在 1984 年，SPSS 公司推出全球第一个统计分析软件微机版本（SPSS/PC+），并使其能很快地应用于自然科学、技术科学、社会科学的各个领域。之后在 1999 年，SPSS 公司收购了 Clementine 产品线，并将其改名为 SPSS Modeler。随着这次收购，以及 SPSS 产品能力和服务范围的扩展，SPSS 公司将英文全称改为 Statistical Product and Service Solutions，即统计产品与服务解决方案，这也标志着公司战略方向和产品设计的重大转型。在 2009 年，SPSS 公司被 IBM 收购，而直到现在，IBM SPSS 产品线下最主要的两款产品依然为 IBM SPSS Statistics 以及 IBM SPSS Modeler，前者定位于为统计分析工具，后者则定位为数据挖掘工具。到 2018 年为止，IBM 已经发布了 SPSS Statistics 25.0 以及 SPSS Modeler 18.1.1。

由于写作期间，Modeler 还没更新到 18.1.1，而且目前来看 18.1.1 属于 18.0 的小版本更新，因此全书的操作实现均基于 SPSS Modeler 18.0 版本。

1.2　SPSS Modeler 的特点

SPSS Modeler 作为一款数据挖掘利器，它的优势（见图 1-2-1）主要表现在 4 方面：专业性、易用性、扩展性以及高性能。

1. 专业性

（1）覆盖整个数据挖掘生命周期：SPSS Modeler 提供数据处理、分析探索、模型创建、评估及部署整个数据挖掘流程功能。

（2）高效的数据处理：SPSS Modeler 提供一系列数据处理功能，包括数据合并、导出、抽样、筛选和汇总等。当用户的连接数据源为数据库的时候，数据处理过程中生成的 SQL 可直接通过 SQL Pushback 技术将生成

图 1-2-1　SPSS Modeler 的优势

的 SQL 回推到数据库端运行，减少数据 I/O 处理时间，从而提高运行速率。

（3）丰富、稳健的数据挖掘模型：SPSS Modeler 提供一系列高级数据挖掘技术，专为满足各种数据挖掘应用程序所需而设计，包括 40 种常用的分类算法、聚类算法和关联规则，其中有 12 种支持 Spark 基于内存计算。

2. 易用性

（1）图形化操作界面：如图 1-2-2 所示，SPSS Modeler 支持图形化界面、菜单驱动和拖拉式的操作。SPSS Modeler 提供了数据源、记录处理、字段处理、图形、模型、输出和导出 7 大类节点，在数据挖掘过程中，只需要把相关节点通过鼠标拖拉的方式连接在一起即可完成整个过程，而无须任何编程操作。

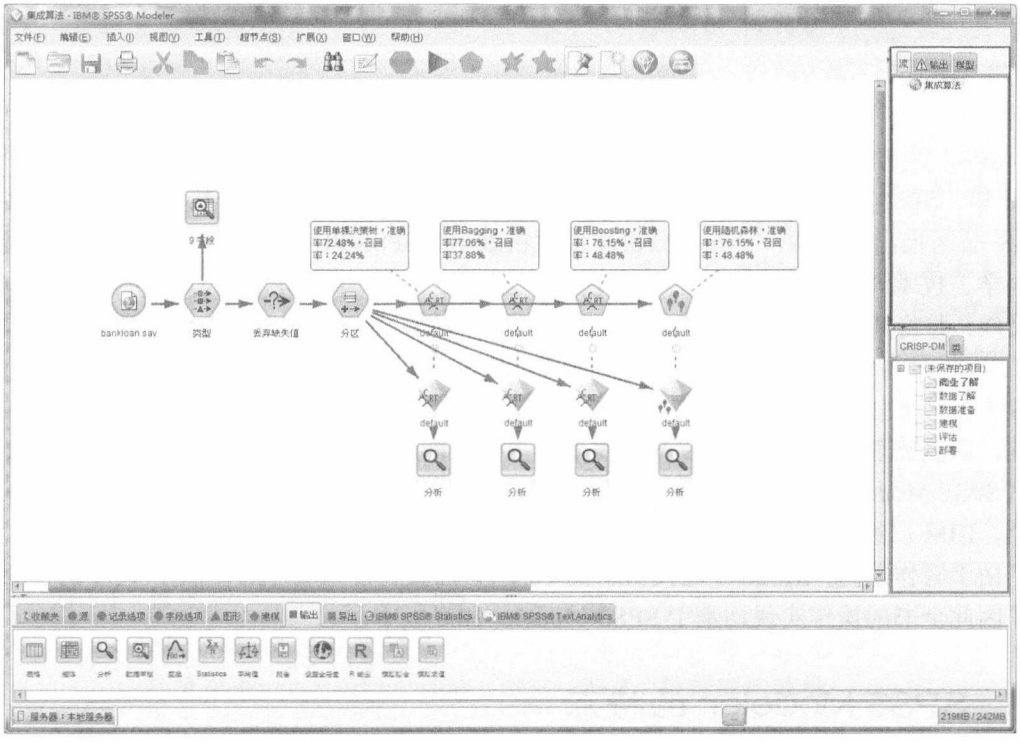


图 1-2-2　SPSS Modeler 图形化操作界面

（2）自动建模算法：如图 1-2-3 所示，SPSS 提供了自动分类、自动数值等自动建模节点帮助用户快捷便利地进行模型选择。例如，自动分类节点能够自动运行所有分类模型，按准确率和运行时间等指标排序，选择最优模型，同时也支持混合多个模型进行组合投票。

（3）便捷的参数调整：各个建模节点中都带有默认模式和专家模式。一般情况下，默认模式能够帮助初学者快速开始数据挖掘过程。而进一步的专家模式，则使得用户能够根据建模目标及实际业务数据特征等进行参数调整，如图 1-2-4 所示。

（4）丰富清晰的中文帮助文档：帮助文档提供了从数据挖掘整个过程点对点的详细说明及应用举例。同时帮助文档提供关键词搜索，可以获取需要的各种问题解答说明。

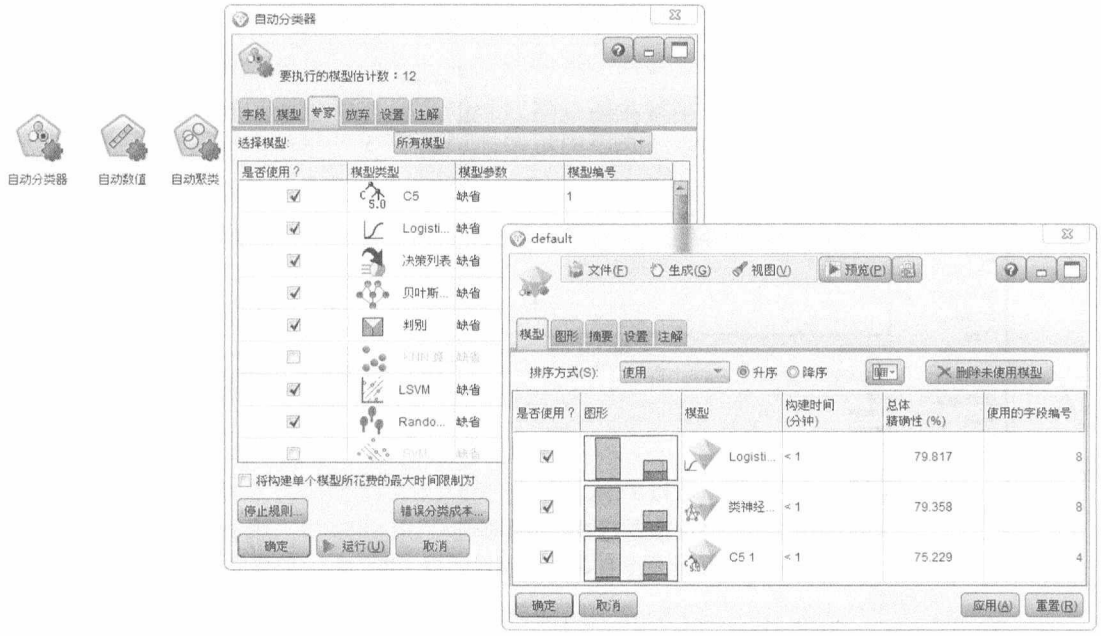


图 1-2-3 SPSS Modeler 自动建模算法

3. 扩展性

（1）多种数据源的集成：SPSS Modeler 支持与多种不同数据源的连接，支持传统关系型数据库（如 Oracle、DB2 和 SQL Server 等）、Hadoop 分布式架构数据、分隔符隔开的文本文件和固定宽度的文本文件、SPSS Statistics 文件、Excel 等各种数据源。

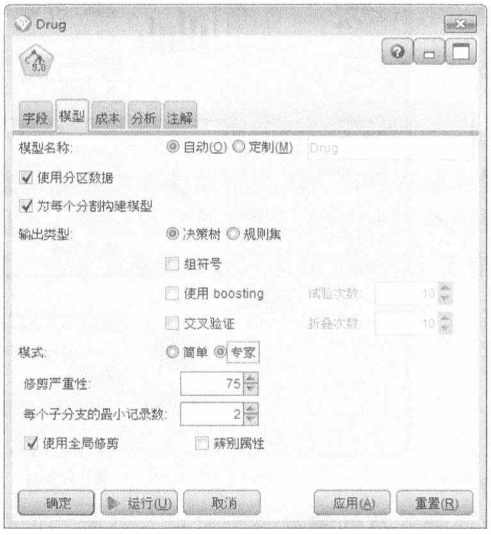


图 1-2-4 SPSS Modeler C5.0 算法参数调整界面

（2）在开源工具上更好的扩展和支持：SPSS Modeler 在开源技术上一直有很好的支持。SPSS Modeler 15 版本开始集成 R 语言，SPSS Modeler 16 版本开始集成 Python，SPSS Modeler 17 版本集成 Spark。来到 SPSS Modeler 18 版本后，SPSS Modeler 在集成上更进一步，以往在集成 Python 及 Spark 上需要 SPSS Modeler Analytics Server 组件的支持，现在能够直接在 SPSS Modeler 的客户端上集成 Python，并且能够把相关的 R 语言代码/Python 代码直接集成为一个建模节点，如图 1-2-5 所示。

（3）全新的扩展中心：除了通过在 SPSS Modeler 中嵌入相关的 R/Python 代码定制相关节点外，IBM 公司也开发了更多的功能在 GitHub 上，现在用户可以直接在 SPSS Modeler 上下载应用相关的功能节点。新的扩展功能包括天气数据获取、GIS 集成和地理空间应用等。SPSS Modeler 的扩展中心如图 1-2-6 所示。

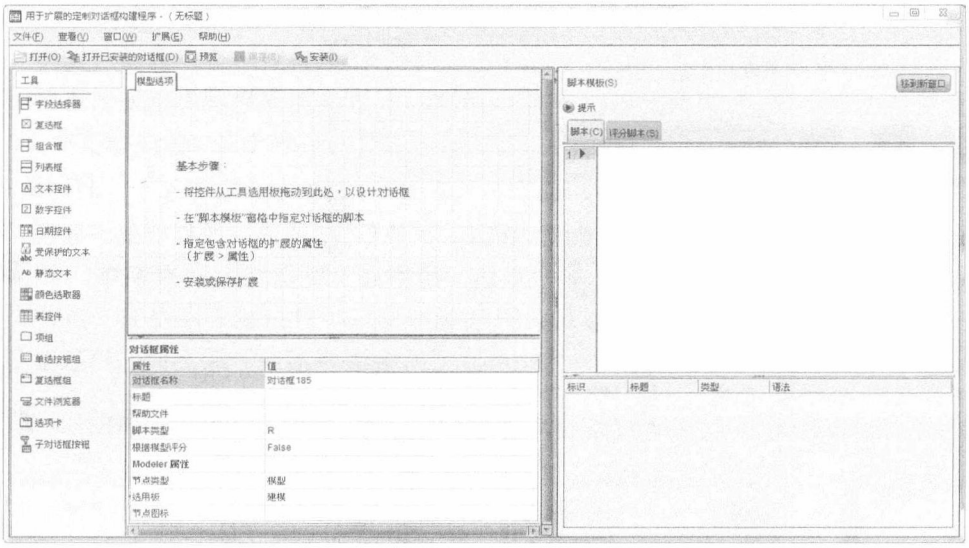


图 1-2-5 Modeler 用于扩展的定制对话框（R/Python 的集成）

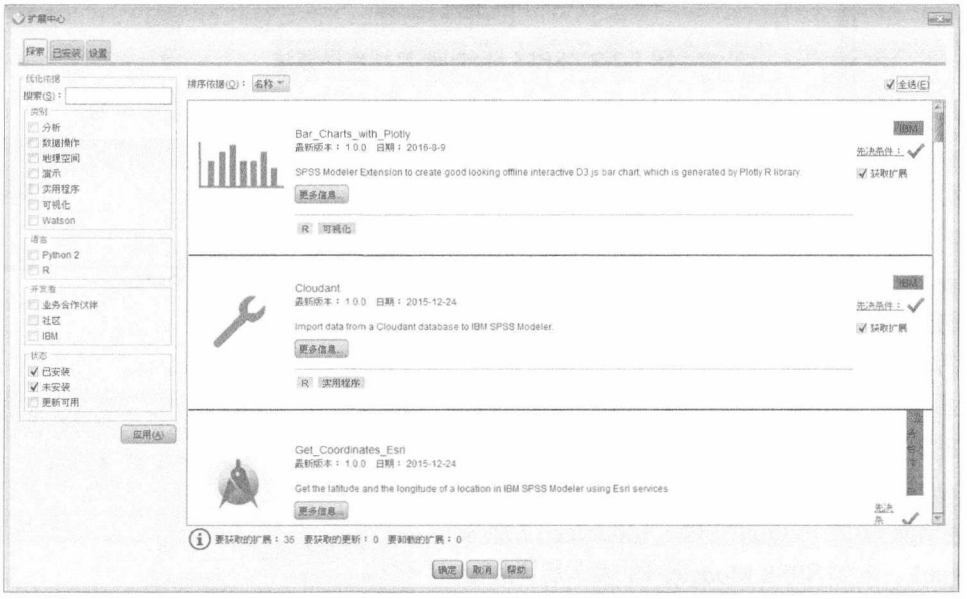


图 1-2-6 SPSS Modeler 扩展中心

4. 高性能

作为一款功能强大的商业化软件，为满足用户对企业化生产环境及分布式计算的需求，SPSS Modeler 提供了对应的服务器版本以及 Analytics Server，通过 Analytics Server 可以直接连接 Hadoop 分布式数据源。同时，提供了多种可以在 Hadoop 上运行的算法，直接转换成 Map Reduce 或 Spark 在 Hadoop 上运行，从而大大提升计算性能。

1.3 CRISP-DM 方法论

在实际的商业挖掘项目中，数据挖掘绝不仅仅是拿到一份数据后建立模型这么简单，

要做好一个数据挖掘项目，需要将丰富的业务知识、高质量的数据以及科学的算法理论进行结合。更具体地说，它是一个从商业问题中来、到商业应用中去的过程，而在一个典型的商业数据挖掘项目过程中如何界定商业问题、怎么获取高质量数据、怎么完成对数据的清洗、如何建立合适的模型、怎么把模型结果应用到商业领域当中都是这个过程的核心要素。因此，为了能够在整个数据挖掘项目过程中更加专业化及标准化，SPSS Modeler 遵循跨行业数据挖掘标准流程（Cross Industry Standard Process for Data Mining，CRISP-DM）方法论进行设计。

如图 1-3-1 所示，在 CRISP-DM 方法论中，它把一个数据挖掘项目划分为 6 个阶段：商业理解、数据理解、数据准备、建立模型、模型评估及结果部署。

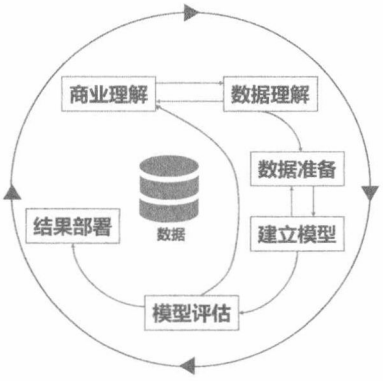


图 1-3-1 CRISP-DM 方法论

1. 商业理解

在数据收集及建立模型之前，应该先完成对商业目标的界定。在这个阶段，需要与相关业务及技术人员对数据挖掘目标的达成、对现有资源的评估及对计划的制定进行充分讨论。商业理解阶段是整个数据挖掘过程路线图的基础所在。在商业理解阶段，需要完成以下工作：

- 确定业务目标；
- 评估情况；
- 确定数据挖掘目标；
- 制定项目计划。

2. 数据理解

在数据理解阶段，需要深入理解可用于数据挖掘项目的相关数据资源。只有完成对数据资源的充分掌握，才能避免在下一阶段（数据准备）中发生意外问题，因此可以利用表格、图形或统计指标对数据进行进一步的数据探索。在数据理解阶段，需要完成以下工作:

- 收集初始数据；
- 描述数据；
- 探索数据；
- 验证数据质量。

3. 数据准备

在数据准备阶段，需要花费大量的时间对数据进行清洗，以保证在建模时具备高质量的数据基础。在实际的数据挖掘项目中，数据准备阶段的工作往往占整个项目工作的 50%～70%。值得高兴的是，假如用户在商业理解及数据理解阶段投了足够多的精力，将能有效地减少在此阶段不必要的返工。在数据准备阶段，需要完成以下工作：

- 选择数据；
- 清理数据；
- 构建新数据；
- 集成数据；
- 格式化数据。

4. 建立模型

建立模型是整个数据挖掘项目中的核心阶段，通过前面的数据准备，用户已经获得可用于数据建模的高质量数据，接下来就是通过构建合适的模型从数据中获得真正的洞察。在建立模型阶段，可能会需要进行多次迭代，以找到一个能够圆满解决商业问题的模型。在建立模型阶段，需要完成以下工作：

- 选择建模技术；
- 生成测试设计；
- 构建模型；
- 评估模型。

5. 模型评估

模型评估是验证用户的工作是否获得成功的关键。在此阶段，除了需要对算法模型进行技术上的评估外，还需要根据在业务理解阶段设定的目标进行业务评估，以确保项目成果能满足实际的业务需求。在模型评估阶段，需要完成以下工作：

- 评估结果；
- 审核过程；
- 确定后续步骤。

6. 结果部署

结果部署是最终结果的运用过程。在此阶段，需要把在数据中获得的洞察应用到具体业务中，以求实现最终的商业价值。在结果部署阶段，需要完成以下工作：

- 指定部署计划；
- 计划监视和维护；
- 生成最终报告；
- 执行最终项目审核。

1.4 SPSS Modeler 下载与安装

SPSS Modeler 的客户端支持 Windows 及 Mac OS 操作系统，SPSS Modeler 服务器端支持 Windows 及 Linux 操作系统。考虑到数据挖掘过程中需要消耗大量的资源，IBM 官方建议对应的系统配置内存应大于或等于 4GB，并且至少有 20GB 的硬盘空间。

1. SPSS Modeler 试用下载

步骤 1：首先登录 IBM SPSS 官方网站。在该官网上提供了 SPSS Modeler 的下载链接，并支持 30 天试用。链接地址为：

https://www.ibm.com/analytics/cn/zh/technology/spss/

步骤 2：在单击图 1-4-1 所示的“SPSS 最新版本下载”按钮后，将弹出如图 1-4-2 所示的对话框。在此处，选择“SPSS Modeler 免费试用”选项。如果用户此前没有注册过 IBMid（IBM 账号），那么在下载前会要求用户注册并登录。