巍巍交大　百年书香
www.jiaodapress.com.cn
bookinfo@sjtu.edu.cn

策划编辑　高　锐
责任编辑　胡思佳　柳卫清
封面设计　黄燕美

Python 数据挖掘技术

Python SHUJU WAJUE JISHU

扫描二维码
关注上海交通大学出版社
官方微信

ISBN 978-7-313-29113-4
9 787313 291134 >
定价：49.80元

大数据人才培养精品系列教材

Python数据挖掘技术

主编　孙玉荣　张　佳

上海交通大学出版社
SHANGHAI JIAO TONG UNIVERSITY PRESS

大数据人才培养精品系列教材

Python 数据挖掘技术

Python SHUJU WAJUE JISHU

主编　孙玉荣　张　佳

本书以构建完整的数据挖掘技能体系为目标，按照CRISP-DM标准组织教材内容，从技术角度着重讲述了求解实际问题涉及的关联规则、分类、聚类等常用挖掘算法的原理，并设计了相应案例，提供了算法实现的Python代码，以强化读者对各知识点的理解与掌握。

上海交通大学出版社
SHANGHAI JIAO TONG UNIVERSITY PRESS

大数据人才培养精品系列教材

Python 数据挖掘技术

Python SHUJU WAJUE JISHU

主编　孙玉荣　张　佳

上海交通大学出版社
SHANGHAI JIAO TONG UNIVERSITY PRESS

内容提要

本书主要介绍数据挖掘的基本技术和应用，全书共分为11章，内容包括数据挖掘概述、Python数据挖掘基础、数据获取与预处理技术、数据可视化、关联规则、决策树算法、朴素贝叶斯分类算法、逻辑回归算法、KNN算法、聚类分析算法、主成分分析。

本书适合作为普通高等教育计算机类、工商管理类相关专业数据挖掘课程的教材，也可作为数据分析与挖掘技术人员的参考用书。

图书在版编目(CIP)数据

Python数据挖掘技术 / 孙玉荣，张佳主编. —上海：上海交通大学出版社，2023.10

ISBN 978-7-313-29113-4

Ⅰ.①P… Ⅱ.①孙… ②张… Ⅲ.①软件工具—程序设计 Ⅳ.①TP311.561

中国国家版本馆CIP数据核字(2023)第198155号

Python数据挖掘技术

Python SHUJU WAJUE JISHU

主　　编：孙玉荣　张　佳

出版发行：上海交通大学出版社　　地　　址：上海市番禺路951号

邮政编码：200030　　电　　话：021-64071208

印　　制：三河市骏杰印刷有限公司　　经　　销：全国新华书店

开　　本：850 mm×1 168 mm　1/16　　印　　张：15.25

字　　数：358千字

版　　次：2023年10月第1版　　印　　次：2023年10月第1次印刷

书　　号：ISBN 978-7-313-29113-4

定　　价：49.80元

前　言

Preface

随着信息技术的飞速发展，海量数据成为一种重要资源。这是一种无形的资源，你可能痛苦于信息资源的泛滥，但当你带着像开盲盒一样好奇、希翼的心情去探索这些资源时，它就是宝藏之源。信息化时代不仅仅需要速度，还需要智慧，因此数据挖掘诞生了。挖掘的字面解释是“挖；发掘”，有让看不见的东西浮现之义，现常引申为深入地开发与探求。因此数据挖掘一词可以形象、生动地描述出人类对信息资源的探索过程。

在当今时代，专业交叉融合会为不同的行业带来创新与活力，数据挖掘无疑是融合的桥梁之一。编者编写本书的目的是引导数据行业新人入门。就讨论的主题而言，本书内容针对的是解决生活中常见的需求任务，包括事物客观情况描述需求、分类与聚类任务需求及预测需求。就数据的挖掘过程而言，本书以 CRISP-DM 行业标准流程为指导，包括业务理解、数据理解、数据挖掘模型构建、模型评估与部署。基于以上考虑，本书分 11 章进行介绍。

第 1 章和第 2 章为本书的基础内容。第 1 章介绍了数据挖掘的必然性，对比了数据挖掘与机器学习的内涵，概述了数据挖掘的方法，讲解了挖掘过程遵循的规则，罗列了部分数据挖掘工具，讨论了模型构建中的关键问题。“工欲善其事，必先利其器”，挖掘工具的使用方法介绍很有必要。第 2 章主要介绍了 Python 的相关知识，但仅靠本书介绍的知识远远不足以让读者掌握该工具的使用方法，编者期望读者以给出的信息为引导，有目的地、高效地选择相关资料去补充学习。

第 3 章和第 4 章涉及业务理解和数据理解的相关知识。第 3 章介绍了数据预处理的原因、方法，还需读者从本书外补足统计类基础知识，诸如常用的统计量的使用、假设检验等。第 4 章介绍了数据可视化的相关知识，数据之美能加深人们对业务的理解，激发人们的创作热情，于养眼养心之时高质高效地完成求解的任务，而可视化工具就是呈现数据之美的有效途径。

第 5 章介绍了关联规则的相关知识，包括基本概念、寻找规则的理论基础及选择规则的准则等知识。作为规律性描述的方法，关联规则是我们认识新生事物的工具，利用该工具可以了解事件发生的因果关系，也可以利用事件的相关性去处理事务。

第 6 章至第 11 章介绍了各类算法的原理、适用范围及其优缺点、模型的评价方法等内容。其中，第 6 章至第 9 章介绍了求解分类任务的系列方法，指导读者进一步理解熵和概率的相关知识，理解分类的目标不仅仅是类别区别，还可以预测未来。第 6 章着重介绍了 ID3 算法的模型构建过程，涉及熵知识的应用。第 7 章和第 8 章与概率应用相关，包括条件概率、先验概率、后验概率等知识。在这里读者将更深入地理解条件假设，并对其意义有新的见解。第 9 章则介绍了 KD 树的构建过程，涉及距离度量知识的应用。第 10 章介绍了聚类的相关知识，着重介绍了 K 均值聚类分析算法。通过学习本章，读者可以更好地区分监督和无监督两个概念，在

遇到问题且没有指导的情况下也有应对的策略。第 11 章介绍了主成分的相关知识。主成分分析解决的主要问题是多维度海量数据的分析问题，读者可以利用该方法降维度以降低分析任务的难度，也可以利用该方法进行综合评价研究。本章还介绍了距离度量的相关知识与度量方式的选择。第 9 章的距离度量方法和本章一样，但考虑到第 9 章主要讨论的是分类任务，故放到分类系列方法中介绍，编者认为这样的安排使全书结构更清晰。

本书从第 5 章到第 11 章，在内容的组织上都基于同样的架构，介绍相关概念、关键问题讨论、任务求解过程、评价方法及优缺点的讨论，并适时给出了相应的代码复现及实例验证。本书的最大特色就是基于每一章的内容指引读者学习的方向，以使读者能构建数据挖掘的框架和问题求解体系。书中介绍的每一种方法，其出现都有经典的应用需求来源，但每一种方法并不仅限于解决与经典相关的任务，当对这些方法及其相关概念有了深度理解之后，创新性的应用就为时不远了。

本书由中南林业科技大学孙玉荣和张佳主编。孙玉荣负责第 1 章、第 2 章、第 3 章、第 4 章、第 5 章、第 6 章、第 9 章、第 11 章的编写，全书的审核及代码实现与调试工作，张佳负责第 7 章、第 8 章、第 10 章的编写。本书在编写过程中，还得到了黄慧华、王昱皓、朱颖芳、龚志伟、叶萍、李奕宏、邹昕莹等同仁的支持和帮助，在此表示感谢。

我们面对的不再是求解一道有标准答案的数学题，因为生活中要解决的大多数问题是多解的，甚至是无解的，希望本书能为读者建立起新的求解体系。本书编写内容多基于工作实践，书中存在的欠妥和偏颇之处，请读者及时指出，我们会不断改进，与读者共同成长。

编　者

目　录

Contents

第 1 章　数据挖掘概述 …… 1

1.1　数据治理 …… 2

1.1.1　数据储量 …… 2

1.1.2　各国数据治理的战略地位 …… 3

1.2　数据挖掘与机器学习 …… 4

1.2.1　两者概念区分 …… 4

1.2.2　两者间的联系 …… 5

1.3　数据挖掘技术 …… 6

1.3.1　数据挖掘本质 …… 6

1.3.2　数据挖掘任务 …… 6

1.4　数据挖掘过程模型 …… 8

1.4.1　9 步模型 …… 8

1.4.2　CRISP-DM 模型 …… 9

1.5　数据挖掘工具 …… 13

1.6　模型构建中的几个关键问题 …… 14

1.6.1　业务理解 …… 14

1.6.2　数据理解与预处理 …… 14

1.6.3　挖掘算法的选用 …… 14

1.6.4　建模用数据集的选用 …… 14

第 2 章　Python 数据挖掘基础 …… 16

2.1　搭建 Python 开发环境 …… 17

2.1.1　Python 第三方库介绍 …… 17

2.1.2　安装 Anaconda …… 18

2.2　Python 数据类型 …… 21

2.2.1　数字类型 …… 21

2.2.2　序列容器 …… 22

2.2.3　非序列容器 …… 27

2.2.4　数据类型的嵌套 …… 29

2.3　Python 程序控制结构 …… 29

2.4 NumPy 科学计算包 …… 30
2.4.1 NumPy 数据类型、视图和副本 …… 30
2.4.2 NumPy 数组基础 …… 31
2.4.3 NumPy 数组操作介绍 …… 34
2.5 pandas 数据分析包 …… 37
2.5.1 pandas 核心数据结构——Series …… 37
2.5.2 pandas 核心数据结构——DataFrame …… 38
2.5.3 数据分析操作基础 …… 39
第 3 章 数据获取与预处理技术 …… 43
3.1 数据 …… 44
3.1.1 数据定义 …… 44
3.1.2 数据分类 …… 44
3.2 数据源 …… 45
3.2.1 数据库数据 …… 45
3.2.2 数据仓库数据 …… 46
3.2.3 事务数据 …… 47
3.2.4 数据矩阵 …… 47
3.2.5 图状结构数据 …… 48
3.2.6 时序数据 …… 49
3.2.7 其他类型数据 …… 49
3.3 数据收集 …… 49
3.3.1 构造数据仓库 …… 50
3.3.2 网络爬虫技术 …… 50
3.3.3 数据集网站 …… 50
3.4 数据质量问题 …… 51
3.4.1 数据完整性问题 …… 51
3.4.2 异常数据 …… 52
3.4.3 数据的不一致 …… 53
3.4.4 多维度数据处理 …… 53
3.4.5 数据量太少 …… 53
3.4.6 数据量过多 …… 53
3.5 数据预处理 …… 54
3.5.1 数据清洗 …… 54
3.5.2 数据集成 …… 57
3.5.3 数据变换 …… 58
3.5.4 数据归约 …… 59

3.6　数据安全 …… 60
3.6.1　数据安全的战略地位 …… 60
3.6.2　数据霸权 …… 60
3.6.3　基础数据界定 …… 61
3.6.4　责任和义务 …… 61
第 4 章　数据可视化 …… 63
4.1　数据可视化定义 …… 64
4.2　常用的可视化工具 …… 65
4.3　常见的可视化图形 …… 66
4.3.1　散点图 …… 66
4.3.2　箱形图 …… 67
4.3.3　热力图 …… 69
4.3.4　直方图 …… 70
4.3.5　聚类谱系图 …… 70
4.3.6　词云图 …… 71
4.4　数据可视化与数据挖掘 …… 72
4.5　Python 数据可视化简介 …… 72
4.5.1　Python 绘图环境搭建 …… 72
4.5.2　Matplotlib 绘图操作方式 …… 73
4.5.3　matplotlib. pyplot 模块的绘图程序设计方式 …… 73
4.5.4　Matplotlib 图层结构 …… 76
4.5.5　图形绘制流程 …… 78
第 5 章　关联规则 …… 81
5.1　关联规则基础知识 …… 82
5.1.1　基本概念 …… 82
5.1.2　数据的离散化 …… 83
5.2　Apriori 算法原理 …… 85
5.2.1　关联规则的评价参数 …… 85
5.2.2　规则的分类 …… 87
5.2.3　Apriori 算法中的两个关键问题 …… 88
5.2.4　Apriori 算法描述及其执行流程 …… 91
5.3　关联规则应用案例 …… 93
5.3.1　代码实现 …… 93
5.3.2　生成的规则 …… 98
5.3.3　关联规则的价值衡量 …… 98
5.4　关联规则应用讨论 …… 99

第 6 章 决策树算法 …… 101

6.1 决策树算法基础知识 …… 102

6.1.1 基本概念 …… 102

6.1.2 构造决策树的关键问题 …… 105

6.2 ID3 算法原理 …… 106

6.2.1 信息增益与属性选择 …… 106

6.2.2 ID3 算法描述 …… 108

6.2.3 ID3 算法的优缺点 …… 108

6.3 决策树的优化 …… 109

6.3.1 拟合能力和泛化能力 …… 109

6.3.2 剪枝策略介绍 …… 110

6.4 决策树模型性能评价 …… 113

6.4.1 混淆矩阵 …… 113

6.4.2 几个常用评估指标的计算 …… 114

6.5 ID3 算法应用案例及 Python 代码实现 …… 117

6.5.1 应用案例 …… 117

6.5.2 代码实现 …… 118

第 7 章 朴素贝叶斯分类算法 …… 125

7.1 贝叶斯算法基础知识 …… 126

7.1.1 贝叶斯决策理论 …… 126

7.1.2 先验概率和后验概率 …… 127

7.1.3 条件概率 …… 127

7.1.4 使用条件概率分类 …… 128

7.2 朴素贝叶斯分类算法基础知识 …… 128

7.2.1 朴素的由来 …… 128

7.2.2 关键问题 …… 129

7.2.3 算法原理 …… 129

7.2.4 朴素贝叶斯分类算法的三种类型 …… 132

7.3 朴素贝叶斯分类算法的优化 …… 135

7.3.1 不完全数据集 …… 135

7.3.2 连续型数值型属性 …… 136

7.3.3 属性之间的独立性 …… 137

7.4 朴素贝叶斯分类算法 Python 代码实现 …… 137

7.4.1 准备工作 …… 138

7.4.2 先验概率估计 …… 138

7.4.3 求出类条件概率并计算可能性 …… 139

7.4.4　构建分类器并进行检验 …… 141
7.5　朴素贝叶斯分类算法应用案例 …… 142
7.5.1　代码实现 …… 142
7.5.2　可视化效果展示 …… 144
7.6　朴素贝叶斯分类算法的优点和缺点 …… 144
7.6.1　优点 …… 144
7.6.2　缺点 …… 145
第 8 章　逻辑回归算法 …… 146
8.1　回归基础知识 …… 147
8.1.1　回归分类 …… 147
8.1.2　线性回归和逻辑回归 …… 147
8.1.3　线性回归 …… 147
8.1.4　二分类 …… 149
8.1.5　多分类 …… 150
8.2　逻辑回归原理 …… 151
8.2.1　逻辑回归的关键问题 …… 152
8.2.2　算法核心内容 …… 152
8.3　逻辑回归算法的优化 …… 155
8.3.1　正则化策略 …… 155
8.3.2　多类别逻辑回归算法 …… 156
8.4　逻辑回归算法的性能评价 …… 157
8.5　逻辑回归算法 Python 代码实现 …… 158
8.6　逻辑回归算法应用案例 …… 160
8.6.1　代码实现 …… 160
8.6.2　可视化效果展示 …… 162
8.7　逻辑回归算法应用场景探讨 …… 163
8.8　逻辑回归算法的优点和缺点 …… 165
第 9 章　KNN 算法 …… 167
9.1　KNN 算法基础知识 …… 168
9.1.1　基本概念 …… 168
9.1.2　KNN 算法的关键问题 …… 169
9.2　KNN 算法原理 …… 171
9.2.1　核心思想 …… 171
9.2.2　算法流程 …… 172
9.2.3　KD 树求解分类过程 …… 172
9.3　基于 KD 树的近邻算法 …… 175
9.3.1　问题实例 …… 175

9.3.2 算法实现 …… 175
9.4 KNN 算法的优缺点及其改进 …… 184
9.4.1 优点 …… 184
9.4.2 缺点 …… 184
9.4.3 改进 …… 184
9.5 KNN 算法的应用场景 …… 184
第 10 章 聚类分析算法 …… 187
10.1 聚类分析算法基础知识 …… 188
10.1.1 分类与聚类 …… 188
10.1.2 聚类分析概述 …… 188
10.1.3 聚类分析的两种类型 …… 189
10.1.4 聚类分析的关键问题 …… 189
10.1.5 聚类算法类型 …… 189
10.2 相似性度量 …… 191
10.2.1 距离度量相似性 …… 191
10.2.2 相关系数度量相似性 …… 193
10.3 原型聚类算法介绍 …… 193
10.3.1 K 均值聚类分析算法介绍 …… 193
10.3.2 其他原型聚类算法介绍 …… 196
10.4 K 均值聚类分析算法的优化 …… 198
10.4.1 后处理 …… 198
10.4.2 二分 K 均值聚类分析算法 …… 198
10.4.3 K 均值++聚类分析算法 …… 198
10.5 K 均值聚类分析算法的代码复现 …… 199
10.5.1 K 均值聚类分析算法流程 …… 199
10.5.2 Python 代码 …… 200
10.6 聚类分析算法实例应用 …… 202
10.6.1 鸢尾花数据集聚类代码 …… 202
10.6.2 效果展示 …… 204
10.7 聚类性能度量 …… 204
10.7.1 外部指标 …… 205
10.7.2 内部指标 …… 206
10.8 K 均值聚类分析算法的优点和缺点 …… 206
第 11 章 主成分分析 …… 208
11.1 主成分分析基础知识 …… 209
11.1.1 主成分分析相关概念 …… 209
11.1.2 与主成分分析相关的数学概念 …… 210

11.1.3　主成分分析的关键问题 …… 212
11.2　主成分分析的基本原理 …… 214
11.2.1　主成分获取的理论基础 …… 214
11.2.2　主成分的线性组合 …… 215
11.2.3　主成分的求解过程 …… 215
11.2.4　主成分的算法描述 …… 219
11.3　主成分的作用与用途 …… 219
11.3.1　主成分的作用 …… 219
11.3.2　主成分的用途 …… 220
11.4　主成分分析应用举例 …… 220
11.4.1　降维处理 …… 220
11.4.2　相关系数矩阵与协方差矩阵结果对比 …… 224
11.5　主成分分析的 Python 代码实现 …… 225
11.6　主成分分析的优点和缺点 …… 230
11.6.1　优点 …… 230
11.6.2　缺点 …… 230
参考文献 …… 232

第1章

数据挖掘概述

本章主要介绍数据挖掘的相关概念、数据挖掘中运用的统计分析方法、数据挖掘任务、数据挖掘过程模型，帮助读者构建数据挖掘学科的整体知识框架。

要求：了解数据治理的战略地位、数据挖掘的概念、常用统计分析方法、本书教学选用的数据挖掘工具，掌握数据挖掘过程。

重点：了解数据挖掘与机器学习的异同、数据挖掘技术、数据挖掘过程、影响建模效果的关键问题。

难点：数据挖掘过程的理解。

1.1 数据治理

20 世纪 90 年代之前，信息技术主要用于数据的保存与处理，以支持各行业的业务运作。20 世纪 90 年代之后，以美国银行业零售业务为代表的金融业，将信息技术的应用扩充至面向顾客的服务型信息处理，数据的价值逐渐被各行业的高层管理者所关注。进入 21 世纪，伴随着各行业之间的激烈竞争与行业信息量激增，数据成为各行业重要的资产和发展的新动力，各行业都在通过加强数据的利用寻找新业务增长点。信息时代里的数据不再仅仅是因为支持行业业务动作而存在，更多的是为了各行业提升自身业务附加价值而存在的，可以这么说，在信息时代，谁抓住数据信息带来的红利，谁将在竞争中取得优势，在这样的背景下，数据治理的需求逐渐浮出水面。

1.1.1 数据储量

提起数据治理，就不得不提数据储量。21 世纪，随着通信技术的发展，各行业信息系统采集、处理和积累的数据量越来越多，全球大数据储量呈爆炸式增长。

图 1-1 描述了 2014—2019 年全球大数据储量（数据来源于 IDC 前瞻产业研究院）并展示了其增长趋势。

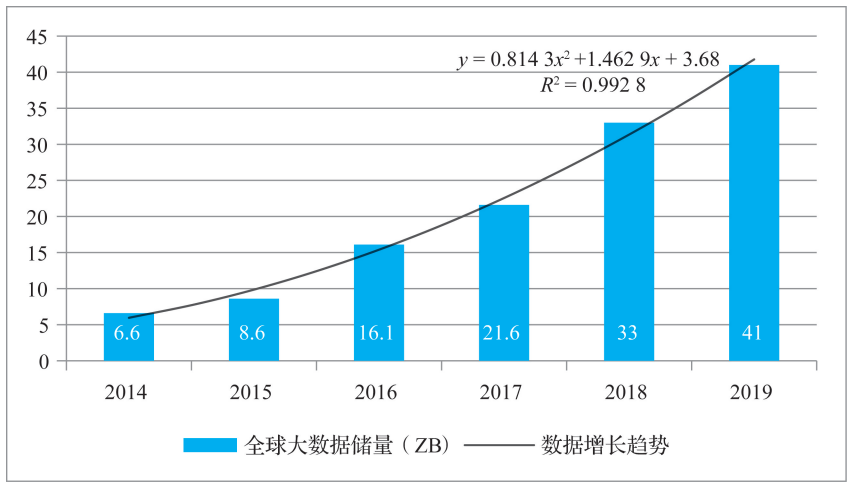


图 1-1 2014—2019 年全球大数据储量及其增长趋势

从图 1-1 可以看出，全球大数据储量从 2014 年的 6.6 ZB 增长到了 2019 年的 41 ZB（1 ZB 的数据储量约等于 10.74 亿个 1 TB 容量的移动硬盘总存储量），每年的大数据储量都处于不断增长状态。全球大数据储量的增长趋势可以描述为：

$$y=0.814\,3x^2+1.462\,9x+3.68(R^2=0.992\,8)$$

其中，x 表示年份对应的序号，$x\geqslant 1$；y 为相应年份对应的数据储量。

这就意味着过去的每一天、每一个小时、每一分钟、每一秒，人类都制造了大量的数据，可以预知，未来随着技术的进步和产业的变迁，数字洪流将席卷世界的每一个角落。

为定量表示同种量的大小而约定的定义和采用的特定量称为计量单位。数据量的计量单位用存储单位表示。图 1-2 直观地展示了数据存储单位的大小。

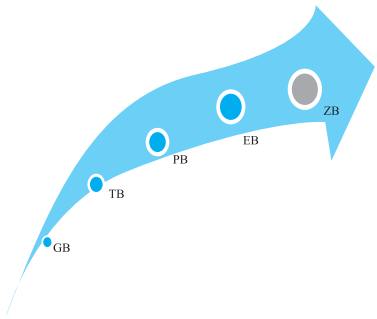


图 1-2　数据存储单位的大小

在计算机中，数据的最小单位是二进制的一个数位，记作位(bit)，也称比特，人们规定 8 位二进制数为 1 个字节(Byte，简称 B)，1 B=8 bit，字节是计算机存储的基本单位。常用的数据存储单位还有千字节(KB)、兆字节(MB)、吉字节(GB)、太字节(TB)、拍字节(PB)、艾字节(EB)、泽字节(ZB，又称皆字节)、尧字节(YB，又称佑字节)，这些存储单位之间的进率是 2^{10}，据此可推出以下换算关系，见表 1-1。

表 1-1　数据存储单位及其换算关系

中文单位	中文简称	英文简称	英文标记	进率
千字节		KiloByte	KB	2^{10}
兆字节	兆	MegaByte	MB	2^{20}
吉字节	吉	GigaByte	GB	2^{30}
太字节	太	TeraByte	TB	2^{40}
拍字节	拍	PetaByte	PB	2^{50}
艾字节	艾	ExaByte	EB	2^{60}
泽字节	泽	ZettaByte	ZB	2^{70}
尧字节	尧	YottaByte	YB	2^{80}

1.1.2　各国数据治理的战略地位

伴随着海量数据的产生，数据资源逐步上升为国家发展的一种战略资源。2009 年，美国总统奥巴马签署了开放和透明政府备忘录，旨在建立更加开放透明的政府、参与合作的政府，体现了美国政府对开放数据的重视。2012 年，美国政府投资“大数据研究与发展先导计划”，开启了美国的大数据时代。与此同时，美国很多行业开始通过大数据创新思维，围绕大数据的采集和利用能力开展各种工作，如气候变化、交通模式、健康和疾病数据、购买行为以及通过媒体互动表现出的社会行为。

2010 年，欧洲联盟(以下简称欧盟)委员会公布了“2020 战略”，这是继里斯本战略之后欧盟的第二个十年经济发展规划。欧盟认为，为挖掘创新潜力，应尽可能地以最好的方式使用资源，这些资源就是数据，开放数据将成为新的就业和经济增长的重要工具。2010 年 11 月，欧盟通信委员会向欧洲议会提交了“开放数据：创新、增长和透明治理的引擎”的报告。该报告以开放数据

为核心，制定了应对大数据挑战的战略。2011 年 11 月，欧盟数字议程采纳了该报告，并于 12 月 12 日正式推进这一战略。2013 年，法国政府发布了数字化路线图，将大数据列为国家创新战略重点实施领域之一。

2013 年，日本政府正式公布了新 IT 战略——“创建最尖端 IT 国家宣言”。该宣言全面阐述了 2013—2020 年期间以发展开放公共数据和大数据为核心的日本新信息技术国家战略，提出要把日本建设成为一个“具有世界最高水准的广泛运用信息产业技术的社会”。

在中国，2015 年，中共十八届五中全会首次提出“国家大数据战略”，发布了《促进大数据发展行动纲要》；2016 年，《政务信息资源共享管理暂行办法》出台；2017 年，《大数据产业发展规划(2016—2020 年)》(以下简称《规划》)指出“十三五”时期是中国全面建成小康社会的决胜阶段，是实施国家大数据战略的起步期，是大数据产业崛起的重要窗口期，必须抓住机遇加快发展，实现从数据大国向数据强国转变。《规划》明确了“十三五”时期大数据产业的发展思路、原则和目标，将引导大数据产业持续健康发展，有力支撑制造强国和网络强国建设。2020 年的统计数据(来源于前瞻产业研究院)表明，中国数据产量约占全球数据产量的 23%。

综上所述，数据资产成为提高国际、国内竞争力的核心资产，数据给各国发展带来了重大的机遇，同时也带来了巨大的挑战，以数据为原材料的数据治理成为国际社会的一项具有战略意义、势在必行的工作。

国际数据管理协会(DAMA)认为数据治理是对数据资产管理行使权力和控制的活动集合。目前，对数据治理存在以下三个层面的理解：宏观上，数据治理是以国家、国际组织、多利益攸关方等为主体，对数据权利、流通、管理等方面所进行的全球化治理；中观上，数据治理是指国家或地区对其主权范围内的数据质量、权属、流动机制等方面进行的宏观管理；微观上，数据治理则是指组织或企业个体对其数据的实用性、可用性、完整性和安全性所做的整体管理。

数据治理是一种体系，在该体系的运作下，通过对数据的获取、处理、使用等方面进行监管来保障数据质量。本书关注的正是微观层面下个体对数据质量采取的相关措施，促使零散数据变为统一数据，改变数据的混乱状况，从而为挖掘数据中隐藏的有价值的知识提供可靠的数据基础。

1.2 数据挖掘与机器学习

1.2.1 两者概念区分

仅利用数据库系统的录入、查询、统计等功能，往往难以发现数据中存在的关系和规则，也难以根据现有数据预测未来的发展趋势，更缺乏挖掘数据中隐藏的有价值的信息的手段。在信息时代，大量的数据背后隐藏着许多重要的信息，人们希望能够对其进行更高层次的分析，以便更好地利用这些数据。

从技术角度看，数据挖掘(data mining，DM)是知识发现(knowledge discovery in database，KDD)的一个步骤。数据挖掘是指从大量的、不完全的、有噪声的、模糊的、随机的数据集中提取隐含在其中的、人们事先不知道但又是潜在有用的信息和知识的过程。这个定义包含以下四层含义：

(1)数据源必须是真实的、大量的、含噪声的,发现的知识是用户感兴趣的知识。

(2)发现的知识要可接受、可理解、可运用。

(3)并不要求发现放之四海而皆准的知识,仅支持特定问题的发现。

(4)这些知识的表现形式可能是数据共性的描述、数据间依存的条件规则,也可能是数据间存在的某种模型关系。

简单而言,数据挖掘就是通过对大量数据进行某种操作,从中发现有用的知识。数据挖掘过程涉及机器学习、数理统计、神经网络、数据库、模式识别、粗糙集、模糊数学等学科的相关技术。数据挖掘过程的三个重要步骤是数据准备、规律寻找和规律应用,如图 1-3 所示。

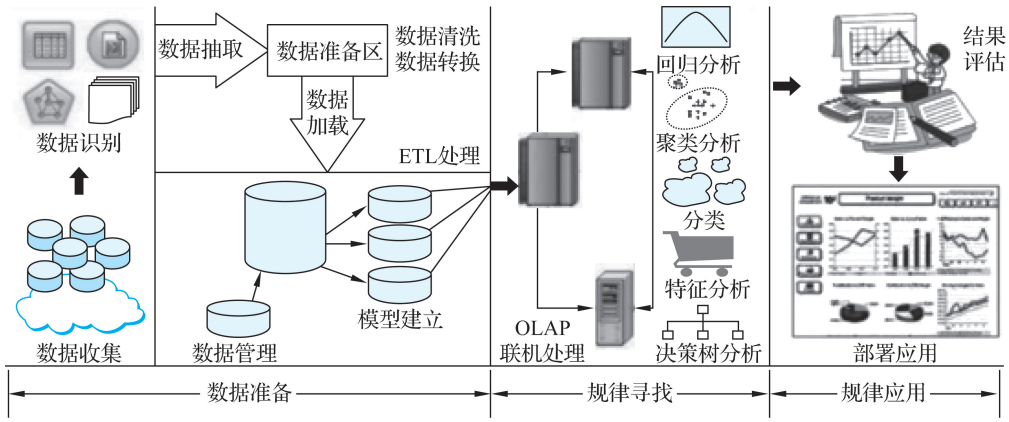


图 1-3 数据挖掘过程的三个重要步骤

从信息系统的角度看,孕育于商务智能需求的数据挖掘是一种决策支持过程,通过高度自动化地分析企业的数据,做出归纳性的推理,从中挖掘出隐含的、并有潜在价值的信息,帮助决策者调整市场策略,提高效益,减少风险。营销界的经典案例"尿片和啤酒"生动地说明了数据挖掘技术出色的领域知识发现和对决策的支持效果。目前,数据挖掘技术广泛应用于商务管理、生产控制、市场分析、工程设计和科学探索等方面。

机器学习(machine learning,ML)是一门多领域交叉学科,涉及概率论、统计学、逼近论、凸分析、算法复杂度理论等多门学科。它是人工智能的核心,是使计算机具有智能的途径之一,其核心任务是专门研究计算机模拟人类学习行为的方法,从而使计算机通过获取新的知识或技能,重新组织已有的知识结构,不断改善自身的性能。机器学习研究的关键问题是改进和创新知识发现方法,克服现存算法的计算性能瓶颈。

1.2.2 两者间的联系

大数据技术的蓬勃兴起和云计算技术的崛起使数据处理、分析和计算成本低廉,机器学习已成为数据挖掘领域中的一个新兴分支与细分领域,也成为当今信息界火热的关键词。数据挖掘可以认为是数据库技术与机器学习的交叉,它利用数据库技术来管理海量的数据,并利用机器学习和统计分析来进行数据分析。数据挖掘集统计学、计算机、数学等学科的相关知识于一体,是多学科交叉的典范,具有极其广阔的应用前景。数据挖掘和机器学习两大领域之间交叉渗透,才会使知识发现更有价值,极大地提高领域知识的应用效果。

1.3 数据挖掘技术

1.3.1 数据挖掘本质

统计分析方法有成熟的数学基础，在数据挖掘过程中有着大量的运用。经典的统计分析方法主要有：

(1)描述统计分析。描述统计分析是对数据的分布状态、数字特征和随机变量之间的关系进行估计和描述的方法，常用于数据的集中度分析、相关分析。很多统计分析方法在应用中要求数据服从或近似服从一定的分布规律，如正态分布，需要对数据的分布进行检验以保证分析结果的可靠性和稳定性。

(2)回归分析。回归分析研究的是因变量(目标)和自变量(预测器)之间的关系，常用于预测分析、时间序列模型以及发现变量之间的因果关系。例如，用回归分析研究司机的鲁莽驾驶与道路交通事故数量之间的关系。

(3)聚类分析。聚类分析是一种探索性分析，聚类与分类的不同在于，聚类所要求划分的类是未知的。在分类的过程中，人们不必事先给出一个分类的标准，聚类分析能够从样本数据出发自动进行分类。聚类分析使用不同的方法，常常会得出不同的结论。不同研究者对同一组数据进行聚类分析，由于设定的参数不同，所得到的聚类数未必一致。

(4)判别分析。判别分析是根据已知的分类目标，对样品数据建立判别函数，进而依据此函数判断给定的新样品归属于哪个类别。

(5)主成分分析(principal component analysis，PCA)。通过正交变换将一组可能存在相关性的变量转换为一组线性不相关的变量，转换后的这组变量称为主成分。主成分分析常用于研究目标的变量降维处理、综合评价等问题。

(6)因子分析。因子分析是一种旨在寻找隐藏在多变量数据中，无法直接观察到却影响或支配可测变量的潜在因子，并估计潜在因子对可测变量的影响程度以及潜在因子之间的相关性的多元统计分析方法，常用于减少变量个数、对原始变量进行分类等。

(7)时间序列分析。时间序列分析是对动态数据进行处理的统计方法，它研究随机数据序列所遵从的统计规律，用于解决实际问题。时间序列通常由四种要素组成：趋势、季节变动、循环波动和不规则波动。

数据挖掘的本质就是从大量数据中自动地抽取模式、关联、变化、异常和有意义的结构，以对数据进行解释、分析和预测，基于统计学，综合数据库技术和人工智能技术完成知识发现。

1.3.2 数据挖掘任务

从大量的数据出发，找出其中蕴含的领域知识，为制定合适的决策提供依据是数据挖掘的社会需求。这些数据来源于金融、经济、政府、人口统计、生命周期、商业等领域，通过数据挖掘相关技术找出其中存在的模式、趋势、事实、关系、模型、关联规则、序列等知识，以指导相关领域负责人制定诸如产品的目标市场、资金分配、贸易选择、广告投放、销售的地理位置等决策。

根据客户需求目标，可以将数据挖掘的任务分为预测类任务（应用历史数据构造模型，预测未来数据的走向）和描述类任务（数据中潜在的规律）两大类，在实践过程中，需要根据任务选择合理的数据挖掘方法。常见的数据挖掘任务如下：

（1）分类分析：可用于完成预测类任务和描述类任务。分类是一种有监督的机器学习方法，根据训练数据集和类标号属性（监督员），构建规则或者决策树模型来分类现有数据或新数据，从而实现对数据的描述或预测。分类分析常应用于风险管理、广告投放、信誉证实、性能预测、目标市场、医疗诊断等目标需求任务。

（2）聚类分析：描述类任务首选。聚类是一种无监督的机器学习方法，把数据按照相似性归纳成若干个类别，同一类中的数据彼此相似，不同类中的数据相异。根据聚类结果，可以从宏观角度了解数据的分布模式以及可能的数据属性之间的相互关系。聚类分析常用于模式识别、空间数据分析、文档分类、市场销售、土地使用、城市规划等目标需求任务。

（3）关联规则分析：数据关联是数据库中存在的一类重要的、可被发现的知识。在交易数据、关系数据或其他信息载体中，查找存在于项目集合或对象集合之间的频繁模式、关联、相关性或因果结构，即关联规则。关联规则分析常用于购物篮分析、交叉销售、产品目录设计等目标需求任务。

（4）序列分析及时间序列分析：序列是指数据元素的有序排列，该序列中的每个元素由不同项目集合组成，在数据挖掘中被称为事务，事务具有时间或空间属性。对于给定的数据挖掘目标，该类分析的关键技术问题之一就是找出事务之间的时间或空间属性，从而明确分析问题的项目集合，该类分析的关键技术问题之二是如何构建序列。该类分析常用于解决客户购买行为模式预测、Web 访问模式预测、疾病诊断、自然灾害预测、DNA 序列分析及工业控制等问题。

（5）孤立点分析：异常对象被称为孤立点或离群点，异常检测也称为偏差检测或例外挖掘。孤立点是一个明显偏离其他数据点的对象，它就像是由一个完全不同的机制生成的数据点一样。孤立点分析是数据挖掘中的一个重要部分，它的任务是发现与大部分对象显著不同的对象。大部分数据挖掘方法都会将这种差异信息视为噪声而丢弃，然而在一些应用中，这种罕见的数据可能蕴含着更大的实用价值。孤立点分析目前被广泛应用于电信和信用卡的诈骗检测、贷款审批、电子商务客户分类、网络入侵检测、故障检测与诊断、药物研究、气象预报、体育竞技中的运动员状态检测等领域。

目前，随着“互联网＋”应用的深入，数据挖掘技术成为各行业抓住机遇发展、迎接挑战的关键所在。

图 1-4 给出了某通信公司数据挖掘任务提取的一个例子。通信运营商的产品通常存在较高的同质化，导致各大公司间的竞争加剧。为了最大化盈利能力，各公司需要制定合适的战略。根据数据获取的来源，客户档案数据可以分为两大类，一类为客户个人信息及其购买产品信息，记作 A；另一类为通信行为数据，记作 B。购买产品信息即客户套餐信息，通信行为数据则包含客户网上行为信息。根据 A 中的数据信息可以构建相关模型对客户价值进行识别。例如，针对客户的购买产品信息将客户分为 VIP、普通客户等；找出客户购买产品与个人信息中的关联关系；对客户满意度进行分析，预测流失可能性等。根据 B 中的数据信息可以构建客户行为特征体系，

并应用聚类算法对客户进行分群(若干个不同的社会群体)。最后根据客户价值模型与客户行为特征进行业务目标群体分类,应用不同的策略对各群体进行业务推广。

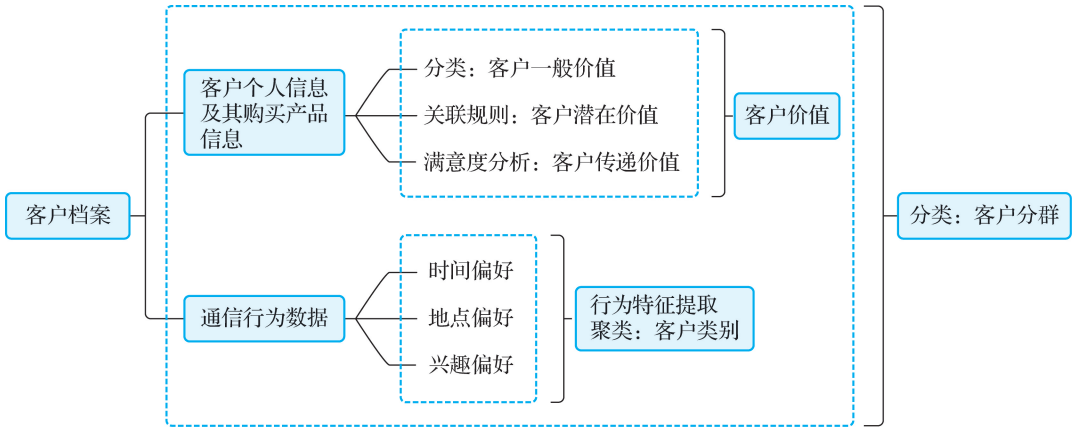


图 1-4　数据挖掘任务提取

1.4　数据挖掘过程模型

数据挖掘过程也称为数据挖掘工作流程,指以数据为原材料,按一定顺序对数据进行提取、整理,进而发现领域知识的全过程,这一过程通常包含了一系列复杂的步骤。数据挖掘过程模型是指对这些步骤进行组合和描述而形成的共同遵守的准则和依据。数据挖掘过程模型主要分为学术模型和工业模型两大类。其中,学术模型以 9 步模型(nine-steps model)为主流,工业模型则以跨行业数据挖掘标准流程(cross-industry standard process for data mining,CRISP-DM)模型为主流。

1.4.1　9 步模型

1996 年的数据挖掘应用处于萌芽状态,Fayyad 提出了数据挖掘的 9 步模型,如图 1-5 所示。

图 1-5　数据挖掘的 9 步模型

该模型以确定数据挖掘目标为起点,以应用知识为终点,将数据挖掘过程细分为 9 步:

(1)定义、理解数据挖掘目标。数据挖掘总是开始于一个业务目标,而数据挖掘人员的首要工作就是很好地理解这一目标,只有这样才能顺利开展数据挖掘工作。因此,数据挖掘项目团队包括客户业务需求管理人员需要通过组织讨论会进行良好的沟通,确定数据挖掘的目标。

(2)选择、创建数据集。在定义目标之后,就需要找出可用的数据、获取额外的辅助数据,根据数据的属性将这些数据整合到一个数据集中。

(3)准备和清洗数据。在现实世界中,数据往往是“脏”的,由于人为操作、测量工具及其他方面的原因,已有部分数据呈现出不正确、不完整、不准确或不相关的特征。这些特征通常以缺失

值、噪声数据、异常值等形式出现，会对挖掘出的知识价值造成困扰，所以需要通过一些方法增强数据的可靠性，尽量提高数据的质量。这一过程称为数据清洗，采用的方法可能是复杂的统计方法，也可能是特殊的数据挖掘算法。例如，获取的某数据空值过多，此时可以建立回归方程，通过计算得到该数据的值。

(4)转换数据。不同的数据挖掘算法有特定的数据属性要求，为方便数据的使用或使挖掘的结果有意义，在这个步骤中，将原始数据转换为可以理解的格式或符合挖掘要求的格式。例如，属性转换、不同数据源的同类数据格式统一或者数据降维减少有效变量的数目。在不同的应用领域，数据转换的方式也是千差万别的。

(5)选择合适的数据挖掘方法。根据不同的数据挖掘目标来制定不同的数据挖掘方法。例如，为了获得更多客户，可以首先学习驱动已有客户的盈利因素，然后获取具有合适特征的新客户；降低信贷风险意味着那些目前信誉良好的客户可能会变质，同时提前减少他们的信用额度；提高客户保留可能会聚焦于改进现有客户的体验，或者获取持续期预期会很长的新客户。

(6)选择数据挖掘算法。算法是数据挖掘中必不可少的一项技术，要结合数据本身的特点与其用途创建并优化算法，才能更加精准地获得想要的数据。例如，挖掘目标是选取广告的最佳位置，由于获取的是人口统计数据，所以可选用相似度模型；挖掘目标是确定新商店的最佳位置，由于获取的是位置范围内的人口数量、已有商店的数量等数据，所以可选用逻辑回归、决策树模型等。

(7)执行数据挖掘算法。根据选取模型的运算结果反复调整模型参数，直到得到满意的精度要求为止，模型构建完毕。

(8)评价结果。模型构建完毕后，还需对模型的性能进行进一步的分析，理论上可以根据分类、聚类、关联等算法的性能分析指标评价算法的性能，对模型的灵敏性、特效性进行度量，实践中要关注模型的有用性和可理解性，以取得更好的应用效果，最后对模型进行归纳与总结。

(9)应用发现的知识。数据挖掘的最终目的是将挖掘到的新知识应用于生产实践。当然，知识数据一定要使用恰当，否则数据挖掘再精妙也是徒劳的。因此，这一步的成败也就决定了整个数据挖掘过程的成败。

9步模型结构更倾向于理论、学术上的探讨，概念清楚，理论完善，但在实践中有时很难严格地区分过细的步骤，如(2)、(3)、(4)步体现的都是数据预处理，影响实践效率，有时并不实用，因此，在实践中人们常以此模型结构为指导。

1.4.2 CRISP-DM模型

1996年，以建立“数据挖掘方法和过程”标准为目标，SPSS、戴姆勒-克莱斯勒和NCR公司共同成立了一个兴趣小组(community-of-interest)，时称SIG组织。该组织于1999年提出了CRISP-DM模型，并于2000年正式推出CRISP-DM 1.0版。SIG组织在伦敦、纽约、布鲁塞尔等地都拥有工作组，成员发展迅速，因此该模型发出即受到广泛关注。如今CRISP-DM模型成为数据挖掘行业标准。

视频
CRISP-DM模型介绍

1. CRISP-DM方法论

CRISP-DM方法论是一种层次模型，采用自顶向下的分析方法对求解问题域过程进行抽象，理

论上讲包括四级层次(从一般到具体)的抽象:阶段、通用任务、特定任务和流程实例,如图 1-6 所示。

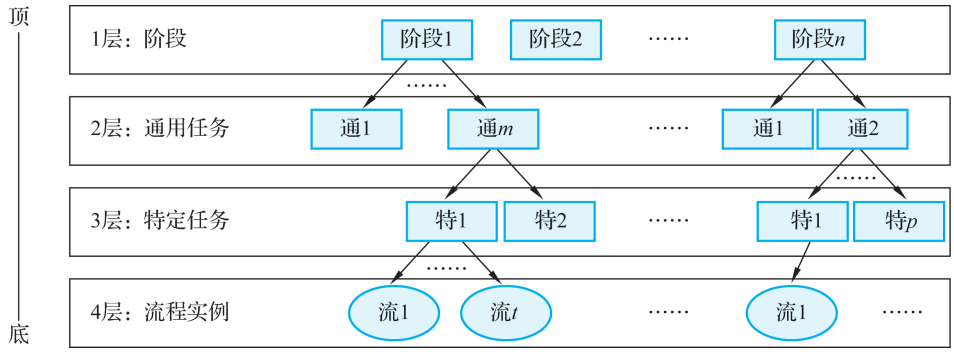


图 1-6　CRISP-DM 方法论的四级层次分类

从图 1-6 中可以看出,数据挖掘过程由四级层次构成,第 1 层抽象出数据挖掘过程的各个阶段,该层的任务是将数据挖掘过程分为 n 段,分别记为阶段 1,阶段 2,……,阶段 n。针对每一阶段 $i(i=1,2,\cdots,n)$,第 2 层列出了其可能包含的通用任务,假设某一阶段包含 m 个通用任务,分别记为通 1,通 2,……,通 m。通用任务设定旨在尽可能地完整和稳定数据挖掘流程,完整是指尽可能地覆盖数据挖掘情形的整个过程和所有可能的数据挖掘应用;稳定则是指对于不可预见的发展(如新的建模技术产生),模型应该仍然有效。第 3 层是对应于第 2 层中的某个通用任务的,就其具体情况做出实施行动。第 4 层则给出特定任务实施的结果,包括一次活动、决策和结果的记录。它表示一个特定项目中发生的实际情况,而不是一般情况。

从形式上看,该模型具有两个特征:将每一个阶段的任务都描述成了一个个简单的事件序列集合,阶段间相互独立(不考虑任务间的联系),从而简化了数据挖掘流程模型。该模型自顶向下,从简单粗略到复杂精细逐步完成数据挖掘任务。

以某项数据挖掘任务为例,将以上四级层次描述为表 1-2 所示的形式。

表 1-2　某项数据挖掘任务各级层次分类

阶　　段	沟通交流		数据理解				数据挖掘		结果表达与解释
通用任务	商业理解	技术理解	收集数据	数据清洗	数据转换	……	算法选择	算法性能	商务分析
特定任务	确定目标		数据源 1; 数据源 2; ……	缺失值处理; 异常值处理; ……	数据格式转换; 降维处理; ……	……	关联关系? 分类? 聚类? 预测?	模型参数调整; 模型性能分析; ……	可视化结果呈现; ……
流程实例	列出目标		数据仓库	数据集	数据集		选择的模型	挖掘的模型	商业知识表达; 决策建议; ……

从表 1-2 中可以看出,第 1 级将本次数据挖掘任务分成了 4 个阶段,第 2 级根据各阶段的性

质制定了通用任务，这两级设置体现了所有数据挖掘的共性问题，第 2 级通用任务尽可能地覆盖了数据挖掘过程的所有任务，并易于扩充新的任务（添加新任务并不影响现有任务的流程）。第 3 级特定任务则与数据挖掘实例紧密关联，如缺失值处理、异常值处理等是数据清洗要求完成的具体任务。第 4 级流程实例则得到相关数据挖掘任务的阶段性成果，最后汇聚到结果表达与解释阶段，进行商业知识表达、提供决策建议，完成数据挖掘任务。

以此方法论为基础可以看出，数据挖掘不单是数据的组织或呈现，也不仅是数据分析和统计建模，而是一个理解业务需求—寻求解决方案—接受实践检验—理解业务需求的完整过程。

2. CRISP-DM 模型的生命周期

图 1-7 所示层次结构模型提供了一个数据挖掘项目生命周期的大概描述，数据挖掘生命周期由项目的各阶段以及这些阶段各自的任务和任务之间的关系等组成。CRISP-DM 模型认为数据挖掘生命周期经历 6 个阶段。

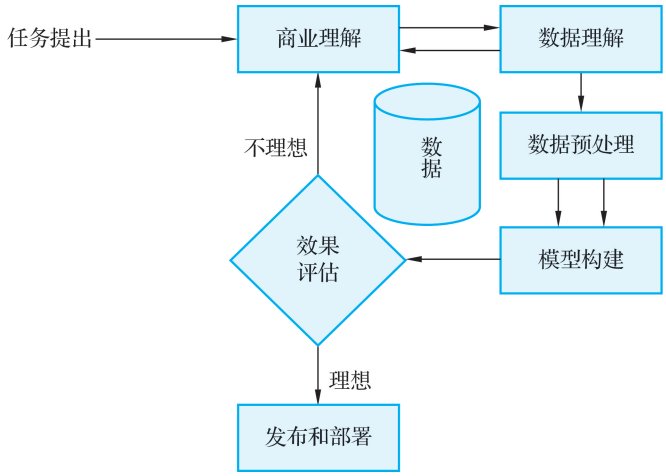


图 1-7　CRISP-DM 模型的生命周期

图 1-7 中的箭头指明了各阶段最重要的关联关系，该模型为数据挖掘项目提供了一个完整的过程描述。从项目商业理解（business understanding）到项目发布和部署（deployment），整个过程由 6 个阶段组成，各阶段的顺序并不是一成不变的，有时需要在不同阶段间进行向前或向后移动，这完全取决于每一个阶段的处理结果或下一个阶段的具体任务，如果该结果是下一个阶段所需要的，则进入数据挖掘流程的下一阶段。

1）商业理解

第一阶段从商业角度全面理解客户真正要达到的目标，明确客户要求，将商业问题转化为数据分析问题，确定项目挖掘目标，同时生成数据挖掘问题定义和完成目标的初步计划。在问题定义中，要说明工作开展的背景、问题的业务价值和工程结果的评价标准。在为达到数据挖掘目标而确定的计划中要写明牵头和配合部门，各个角色的主要任务，以及整个工程各阶段的时间安排。

商业目标是以商业术语描述的，而数据挖掘目标是以技术术语描述的。数据挖掘目标：描述项目的预计输出，该输出使得商业目标得以实现。

2)数据理解

数据理解阶段从收集原始数据开始,首先根据第一阶段项目目标寻找可用的数据源(如某些系统、工具,或者某些部门的存档文件),梳理出需要的数据,对多数据源还需要进行整合,在此基础上,了解数据的含义及数据之间的关系,对多数据源,还要弄清多数据源间的关联关系。最后对数据的质量问题进行判别,如数据是否完整,是否存在缺失值,并对出现的质量问题列出可能的处理方法。在数据理解阶段,可能会发现更有价值的信息或新的困难,因此,需要返回商业理解阶段对目标和计划进行调整。

3)数据预处理

数据预处理阶段对选用的原始数据进行选择、清洗、构造、整合以及格式转换等操作,使数据达到建模需求。选择数据是指找出与数据挖掘目标相关的属性数据;清洗数据包括数据一致性处理、无效值处理与缺失值填补等,是提升数据质量、保证分析结果正确的关键步骤。通过属性派生、全新记录生成等方式构造数据;对多表或多条记录进行合并或聚合操作实现数据的整合;受数据挖掘算法限制,格式转换可能涉及数据属性的顺序变更、数据记录的顺序变更或者不改变数据含义的数据符号形式的改变等。

数据预处理阶段与模型构建阶段关系十分紧密,在建模过程中预处理数据的质量会对建模结果造成深远的影响。由于数据质量问题引起的建模结果问题需要返回数据预处理阶段以重新调整数据,因此该阶段工作有可能重复执行多次,这样就会造成数据预处理过程与建模过程循环迭代,导致数据挖掘过程顺序不唯一。

4)模型构建

在这个阶段,可以选择和应用不同的数据挖掘技术建立模型,借助数据挖掘工具,经过多次执行,通过查全率、查准率、可信度、处理性能等指标调整模型参数得到最优的模型,以获取最佳的数值。这里指实际选用的数据挖掘技术,如 C4.5 决策树模型。许多建模技术对数据都做了明确的假定,如所有属性同分布、不允许有缺失值、分类属性必须是符号等。因此,在实际操作过程中需要经常跳回到数据预处理阶段对数据进行调整。如果是非数据质量问题导致挖掘结果不理想,则需要换一种数据挖掘技术,相应地根据算法重新准备数据。

5)效果评估

模型构建阶段所做的评估仅仅是从建模技术层面考察模型的准确度和通用性因素,而本阶段的评估则是评价所建模型多大程度地满足第一阶段制定的商业目标,并努力寻求商业理由以解释模型的欠缺。评价的方法就是把生成的结果与在项目开始定义的评价准则进行比较。CRISP-DM 模型给出了评价用等式:

$$RESULTS = MODELS + FINDINGS$$

其中,RESULTS 指模型的所有输出结果,MODELS 指模型本身计算的结果,FINDINGS 是除模型本身计算结果之外新发现的信息(如没有被数据挖掘项目所覆盖的数据质量问题),该信息不一定与当前项目商业目标有关,但对项目的发起者却非常重要。

模型评估的结果不外乎成功与失败两种可能。如果成功了,那么就进入发布和部署阶段,将数据挖掘获取的知识并入企业产品的业务目标中。如果失败了,则需返回到前面的步骤,或者重新设定业务目标,或者重新收集数据进行建模。

6)发布和部署

建模的目的就是从数据中发现知识,而获得的知识需要以便于用户使用的方式重新组织和展现,最后生成项目总结报告提交给相关人员,这就是结果部署阶段的工作。根据业务目标的不同,制定模型部署的方案与计划也不同。简单的部署方案可以只提交一份数据挖掘报告,复杂的部署方案则需要将模型集成到企业的核心运营系统当中,通过与企业核心运营系统对接和联调,使模型上线发布。由于算法模型一般是基于历史数据构建的,所以在模型部署并运行一段时间之后,业务场景可能已经发生了变化,导致原有的模型可能已经无法满足当前的业务需要。因此在模型部署上线的同时,还需同步上线模型的监督和维护系统,以持续跟踪模型的运行状况并及时地进行调整。

CRISP-DM 模型来源于实践,并用于指导人们的数据挖掘实践。该模型通过 6 个步骤对企业数据挖掘的过程进行了描述,目前是数据挖掘领域事实上的行业标准。本书会以此标准为指导,讲解各实例数据挖掘任务。

1.5 数据挖掘工具

随着科技的发展和应用的深入,信息的表现形式变得更加多样化,不再是单一的结构化数据,更多的是非结构化数据,数据量也呈爆发式增长,对数据挖掘的技术要求也在不断地提高。数据挖掘工具是指使用数据挖掘技术从大型数据集中发现并识别模式的计算机软件。目前有大量的数据挖掘工具可供选择,下面列出几种常用的数据挖掘工具:

(1)R 语言。R 语言是一种简单而强大的编程语言,提供了完整的数据处理、计算和制图软件系统。其功能包括:数据存储和处理系统;数组运算(其在向量、矩阵运算方面的功能尤其强大);完整连贯的统计分析工具;优秀的统计制图功能;可操纵数据的输入和输出,可实现分支、循环,用户可自定义功能。

(2)RapidMiner。RapidMiner 是世界领先的数据挖掘解决方案,它提供了丰富的数据挖掘算法,用户也可以通过建模选项板开发预测模型。RapidMiner 常用于解决诸如营销响应率、客户细分、客户忠诚度及终身价值、资产维护、资源规划、预测性维修、质量管理、社交媒体监测和情感分析等典型商业关键问题,覆盖了汽车、银行、保险、生命科学、制造业、石油和天然气、零售业、通信业,以及公用事业等各大行业事务处理,极大地简化了数据挖掘过程的设计和评价。

(3)IBM SPSS Modeler。IBM SPSS Modeler 是世界领先的可视化数据科学和机器学习解决方案,具有可视化界面,用户无须编程即可实现数据分析、地理空间分析、文本分析和社交网络分析等建模任务,有效提高工作效率。

(4)Kaggle。Kaggle 是全球最大的数据科学社区。2010 年,联合创始人兼首席执行官安东尼·高德布卢姆(Anthony Goldbloom)在澳大利亚墨尔本创立了 Kaggle,主要是为开发商和数据科学家提供举办机器学习竞赛、托管数据库、编写和分享代码的平台,受到了许多科学家和开发者的关注。公司和研究人员在这里张贴他们的数据,来自世界各地的统计人员和数据挖掘者竞相制作模型。

(5)SAS Data Mining。SAS Data Mining 是一款商业软件,它为描述性和预测性建模提供了更好的理解数据的方法。

(6)Python。于 1991 年诞生的高级程序设计语言 Python 是当前世界上最受欢迎的编程语言之一。Python 提供了丰富的标准程序包和第三方程序包,在数据挖掘领域得到广泛应用。利用其提供的 NumPy、pandas、Matplotlib、SciPy、scikit-learn 可以完成数据挖掘任务,本书内容的学习就由 Python 实现。

1.6 模型构建中的几个关键问题

借助于数据挖掘的相关工具,完成分析任务看似一项简单的工作,但分析结果、质量却因人而异。造成这一问题的关键是分析思路与数据处理。从提出任务到完成任务的整个过程看,模型构建过程中存在几个关键的问题值得关注。

1.6.1 业务理解

业务理解是挖掘任务成功的最关键的一步。大多数情况下,业务理解是理解需求和制定分析目标的基石。因此,进行数据挖掘需了解行业知识、领域知识并亲临一线去了解业务实际情况,与领域专家进行充分交流,在此基础上把整个任务模块分解成各个相关联的子模块,从而制定出合理的总目标与实现总目标的各个子目标。

1.6.2 数据理解与预处理

了解业务处理过程中数据的产生过程,明确数据代表的意义,对数据的结构和字段间的关系进行分析,是选取特征变量以及构建挖掘模型的基础。要进一步关注数据质量,数据质量直接关系到挖掘出的知识价值,防止“垃圾进,垃圾出”的有效手段就是获得高质量的数据,其中预处理技术是关键。因此,需要加深对数据预处理知识的理解,对预处理方法进行深入了解,充分理解方法的缺陷并尝试进行改善。

1.6.3 挖掘算法的选用

当前人们提出的各种数据挖掘算法不下百种,但没有任何一种数据挖掘算法是万能的。选用算法时应注意:

(1)不同的算法应用于具体数据的含义和能力不同。

(2)一个问题可能存在多种算法可以求解,但挖掘质量可能有差异。

(3)有些算法可以用于多种数据类型,有些算法则对某些数据类型不适用。

(4)有些算法的参数选择依赖于经验。

(5)有些算法对数据有特殊要求,需要进行某些转换、过滤之类的操作。

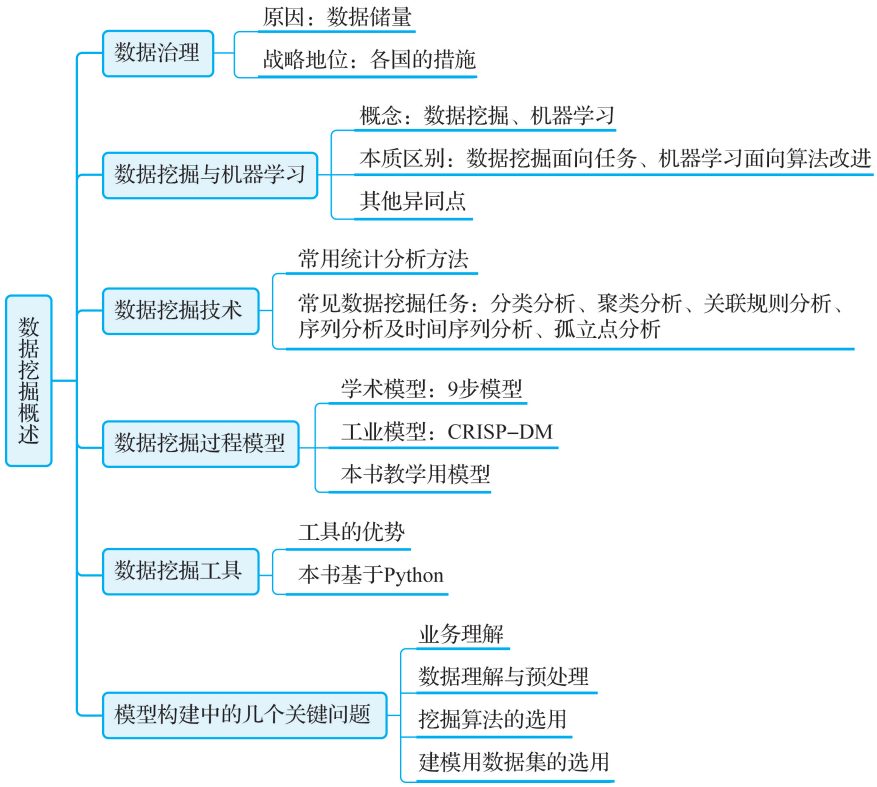
(6)通过历史数据所建立的模型,其分析和挖掘出的结果与当前实际客体的行为不一定完全相同,因而在应用挖掘出的知识进行决策时存在非系统性风险。

1.6.4 建模用数据集的选用

一般情况下,算法中的参数设定与给定的数据集有关,因此,数据集的选用是算法执行效果

的关键。例如，在构建预测模型中，首先需要将数据集划分成训练用数据集和验证用数据集。实验过程中采用 N 折交叉验证的方法，可以提高算法的泛化能力，因此，数据集的构建是建模的一个关键问题。

本章小结



本章习题

1. 简述数据挖掘与数据治理的关系。

2. 简述数据挖掘与机器学习的异同点。

3. 结合自己的经历和认识，描述数据挖掘给某个行业带来的变化，你认为数据挖掘可能需要完成的任务有哪些？

4. 数据挖掘过程模型是一成不变的吗？谈谈你对数据挖掘过程的理解。

5. 有哪些影响建模效果的关键问题？