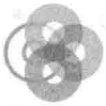
大数据技术与应用专业规划教材

数据挖掘实用案例分析

◎ 赵卫东 董亮 著

清华大学出版社



大数据技术与应用专业规划教材

数据挖掘实用案例分析

◎ 赵卫东 董亮 著

清華大學出版社
北京

内 容 简 介

数据挖掘已经广泛应用于各行各业，并催生了数据分析师的兴起。本书结合项目实践，首先对数据挖掘的核心问题进行了总结，并以保险推荐为例说明数据挖掘过程中每个步骤需要关注之处；然后，结合香水销售分析，讨论可视化图形的基本应用。为增强本书的实用性，提高读者的动手能力，后续章节详细地分析了数据挖掘在银行信用卡、餐饮、商务酒店、制造业、公安等领域的应用。此外，本书还介绍了卷积神经网络在音频数据处理方面的实际应用。

本书内容深入浅出，案例生动形象，可以作为高校相关专业“数据挖掘”“机器学习”“商务数据分析”等课程的实验教材，也可以供学习数据分析的社会人士参考。

图书在版编目(CIP)数据

数据挖掘实用案例分析/赵卫东，董亮著. —北京：清华大学出版社，2018
(大数据技术与应用专业规划教材)
ISBN 978-7-302-49049-4

Ⅰ. ①数…　Ⅱ. ①赵… ②董…　Ⅲ. ①数据采集－案例　Ⅳ. ①TP274

中国版本图书馆 CIP 数据核字(2017)第 295509 号

责任编辑：闫红梅　常建丽
封面设计：刘　键
责任校对：焦丽丽
责任印制：杨　艳

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
课件下载：http://www.tup.com.cn，010-62795954

印 装 者：北京国马印刷厂
经　　销：全国新华书店
开　　本：185mm×260mm　　**印　　张**：16.5　　**字　　数**：400 千字
版　　次：2018 年 2 月第 1 版　　**印　　次**：2018 年 2 月第 1 次印刷
印　　数：1～1500
定　　价：49.00 元

产品编号：075293-01

FOREWORD 前言

目前，高校的数据分析类课程（如数据挖掘、机器学习、大数据分析等）教学方式大多以“知识点”为核心组织教学，学生主要以学习知识为主，工程应用实践机会较少。教师将所要教授的知识点在课堂上讲述，课后再以作业练习、课程实验、课程设计等形式帮助学生深入理解课堂上所学的知识。尽管为提高教学效果，目前许多高校尝试了大型开放式网络课程（Massive Open Online Course，MOOC）、翻转课堂、移动课堂、同伴学习和小规模限制性在线课程（Small Private Online Course，SPOC）等教学方法的改革，但总体上来说，对于应用性较强的课程教学，还存在改进的空间，尤其是对学生的动手实践能力要求较高的数据分析类课程。现有的教学方法在传授理论知识时，缺少实际应用环节的支持，学生缺少在实际应用的背景下充分理解所学知识的机会，难以培养学生应用专业知识分析解决问题的技能和创新思维能力。

数据分析的方法是科学，但这些方法的选择和应用过程因问题而异，带有很强的艺术性。在现有专业课程教学模式下，学生仅仅了解需要学习基本的理论知识，缺少实践动手经历，难以获得这些知识的应用技巧，很少接触与企业实际项目相关的内容，因此学生的应用能力较弱，与企业实际的需求脱节。例如，在“数据分析”课程中，一般的教学方式是教师将具体数据分析的方法教授给学生，学生能够理解算法或方法的内容，但难以解决实际项目中应用具体算法碰到的问题。目前亟待克服数据分析类课程教学脱离企业所需能力的培养痛点，在课程学习的知识基础上，解决实际问题，引导学生解决数据分析实际问题的必要技能和思维方法。

实际上，数据分析绝大部分的教材和书籍还基本停留在基本理论和方法的介绍，实验部分的内容比较简单或者缺失，实际应用的内容不足。还有些实战性的书籍没有按照教材的方式编写，案例也比较粗略，数据分析过程中的一些技能解释肤浅。有关实际项目中数据分析过程思路的分析以及难点解析对教学，尤其是对实验或案例教学非常重要。最近几年，作者与多家企业合作，在数据分析领域辛苦耕耘，亲自参与了多个实际数据分析项目，熟悉数据分析过程的酸甜苦辣，希望通过本教材弥补国内数据分析实用教材的不足，也希望本教材的出版能改善国内数据分析类课程教学资料短缺的情况。

学习数据分析的最好方法就是做中学，使用实际数据解决实际问题，而不是单纯学习技术。实际上，有效的数据分析需要对业务进行深入理解，在此基础上形成有效的分析思路，并通过实验反复比较，才能真正解决客户的问题。在数据时代，现实应用中往往不乏数据。从生活中的小数据、简单问题开始，做各种假设，探索其中的规律。不断尝试常用的分析语言、工具和技术，在应用中不断学习新的知识，弥补课堂教学的不足，尤其是体会数据分析过

程中书本上难得看到的分析技巧，并在应用中举一反三。如此反复，随着分析问题的深入，不断提高分析能力，体会数据分析的艰辛和解决客户问题的快乐。

本教材不局限于数据分析基本理论和基本方法的介绍，而是立足实际应用，突出实际数据分析项目中的思路，以及数据分析中的难点。但希望读者具有一定的统计学、机器学习（数据挖掘）、数据科学，以及必要的相关专业知识。也不追求过多的案例堆积，希望读者能理解数据分析的思路，举一反三。这些内容是作者多年项目实践和教学成果的总结，其中的分析思路只有参与实际的项目，才能体验到数据分析的难点和艺术性，这是目前教学过程中培养学生工程性思维的重要问题，也是真正提高学生创新能力和动手能力的手段。这些内容是数据分析的基础，也是从事大数据分析必须掌握的知识和技能。有关数据挖掘常用算法的介绍，读者可以参阅作者已经出版的教材《商务智能（第4版）》（清华大学出版社，2016年）或其他专业书籍。

全书分为11章，具体的内容简介如下：

第1章从数据分析的流程出发，讨论了在数据分析各个阶段需要做的工作以及经常遇到的主要问题，尤其是数据挖掘算法使用时容易遇到的难题。数据挖掘过程有一定的标准，但是针对具体的业务需求，如何设计合理、有效的数据分析流程，需要有一定的经验和技巧，数据的预处理、算法的选择等主要步骤都充分体现了数据挖掘的艺术性。

第2章以保险产品推荐项目为例，突出了数据挖掘选择合适的算法并非很简单的事情，需要在理解分析问题以及对多种算法熟悉的基础上，通过实验对初选的几种算法进行比较、调优，才能选择对解决问题效果比较好的算法。

第3章介绍了多维分析常用的可视化图形，这是数据分析的基本功。这些图形可以帮助数据分析师探索数据，找出数据中存在的问题以及基本规律。

第4章介绍了IBM SPSS Modeler 18数据挖掘工具的常用组件。在学习数据分析的不同阶段，根据学习者的基础、问题的分析难度等，可以选择不同的工具或平台。尽管分析工具并不是数据挖掘最重要的事情，但学习成本低、功能强大的分析工具对于问题的解决也是不可少的。对于编程基础有限的数据分析师，可以选择类似IBM SPSS Modeler 18的挖掘工具或TensorFlow等开源工具。尽管如此，对于有一定数据分析基础的读者，推荐学习Python、R等针对数据分析的语言，这些语言比较灵活，功能也十分强大。

第5章对香水的销售数据进行分析，讨论受欢迎的香水以及特点，并找出影响香水销售的主要因素，为香水的营销提供依据。

第6章对银行的客户信用记录、申请客户信息、拖欠历史记录、消费历史记录等人口属性、交易数据进行综合分析，讨论用户银行信用卡拖欠和欺诈行为特征，为银行推广信用卡以及风险管理提供依据。

第7章从大众点评网抓取火锅店海底捞的菜品介绍以及客户评论数据，以客户为中心，分析客户对火锅的偏好，为火锅店的选址、菜品的选择和设计，以及火锅店的竞争力都提供了参考。

第8章以携程网上某商务宾馆的客户评分、评论数据为基础，通过情感分析，分析了客户对商务宾馆的偏好，并了解客户的消费行为，比较多家商务宾馆的竞争优劣势，为商务宾馆改进经营提供了参考。

第9章在某耐热导线工厂最近2年的质量管理数据的基础上，分析了这些数据存在的

问题，探索耐热导线的加工流程中几个工序之间半成品或成品质量指标的关系，提高最终产品的合格率。

第10章利用公安人口数据和违法犯罪人员行为特点的数据，建立风险评分模型，实现对高危人群的特征分析，识别具有违法、犯罪、可疑或可能的高危人员。

第11章讨论深度学习在音频处理领域的应用，介绍了常用的深度神经网络模型，重点分析卷积神经网络在音频质量评价领域的应用。

数据挖掘是一个多学科交叉的领域，本书通过少数实际的具体案例，阐述数据分析项目的过程以及一些要点，可作为普通高等学校“数据挖掘”“商务数据分析”“商务智能”等课程的案例和实验指导材料，也可供有志于数据分析师的读者参考。配套实验数据、源代码、软件等可以从清华大学出版社网站下载。由于作者水平有限，书中难免有错误之处，希望读者不吝指出。

在写作的过程中，胡远文、于召鑫、黄黎明、蒲实、朱荣斌等在资料收集方面做了一些工作，在此表示感谢。

赵卫东
2017年8月
复旦大学

CONTENTS 目录

第 1 章　数据分析过程的主要问题 …… 1

1.1　业务理解 …… 1
1.2　数据理解 …… 2
1.3　数据质量问题与预处理 …… 3
1.4　数据分析常见陷阱 …… 9
1.5　数据分析方法的选择 …… 10
1.5.1　分类算法 …… 11
1.5.2　聚类算法 …… 15
1.5.3　关联分析 …… 16
1.5.4　回归分析 …… 17
1.5.5　深度学习 …… 19
1.5.6　统计方法 …… 19
1.6　数据分析结果的评价 …… 19
1.6.1　分类算法的评价 …… 20
1.6.2　聚类结果的评价 …… 21
1.6.3　关联分析的评价 …… 22
1.6.4　回归分析结果的评价 …… 22
1.6.5　深度学习的评价 …… 23
1.7　数据分析团队的组建 …… 24
1.7.1　项目经理 …… 24
1.7.2　业务专家 …… 24
1.7.3　数据工程师 …… 25
1.7.4　数据建模人员 …… 25
1.7.5　可视化人员 …… 25
1.7.6　评估人员 …… 25
1.8　数据分析人才培养的难题 …… 26
1.8.1　数理要求高 …… 26
1.8.2　跨学科综合能力 …… 26

1.8.3 国内技术资料少 …… 26
1.8.4 实践机会少 …… 27

第 2 章 数据挖掘算法的选择——保险产品推荐 …… 28

2.1 业务理解 …… 28
2.2 数据分析目标 …… 29
2.3 数据探索 …… 30
2.3.1 数据质量评估 …… 30
2.3.2 探索数据统计特性 …… 32
2.3.3 数据降维 …… 34
2.4 模型选择过程 …… 36
2.4.1 算法初选 …… 37
2.4.2 算法验证 …… 40
2.4.3 算法优化 …… 43
2.4.4 平衡数据集 …… 43
2.4.5 修改模型参数 …… 46
2.5 总结 …… 48

第 3 章 常用可视化的多维分析 …… 50

3.1 箱图 …… 51
3.2 雷达图 …… 53
3.3 标签云 …… 55
3.4 气泡图 …… 56
3.5 树图 …… 57
3.6 地图 …… 58
3.7 高低图 …… 59
3.8 双轴图 …… 60
3.9 关系图 …… 61
3.10 热图 …… 63

第 4 章 SPSS Modeler 建模组件介绍 …… 65

4.1 数据预处理组件 …… 65
4.1.1 数据清理组件 …… 65
4.1.2 数据集成组件 …… 66
4.1.3 数据选择组件 …… 67
4.1.4 数据变换组件 …… 67
4.2 数据挖掘建模组件 …… 68
4.2.1 模型筛选 …… 68
4.2.2 自动建模 …… 68

4.2.3 决策树模型 …… 69
4.2.4 贝叶斯网络模型 …… 70
4.2.5 神经网络模型 …… 70
4.2.6 支持向量机模型 …… 71
4.2.7 时间序列模型 …… 71
4.2.8 统计模型 …… 71
4.2.9 聚类模型 …… 73
4.2.10 关联分析 …… 73
4.2.11 KNN 模型 …… 74
4.2.12 数据挖掘模式评估 …… 74
4.3 知识表示 …… 74
4.3.1 图形节点 …… 75
4.3.2 数据输出 …… 75
4.3.3 数据导出 …… 76

第 5 章 香水销售分析 …… 77

5.1 香水销售数据预处理 …… 77
5.2 香水销售数据统计分析 …… 79
5.3 影响香水销量的因素分析 …… 84
5.4 香水适用场所关联分析 …… 87
5.5 香水聚类分析 …… 89
5.6 香水营销建议 …… 92

第 6 章 银行信用卡欺诈与拖欠行为分析 …… 93

6.1 客户信用等级影响因素 …… 94
6.1.1 客户信用卡申请数据预处理 …… 94
6.1.2 信用卡申请成功影响因素 …… 96
6.2 信用卡客户信用等级影响因素 …… 102
6.3 基于消费的信用等级影响因素 …… 104
6.4 信用卡欺诈判断模型 …… 105
6.4.1 基于 Apriori 算法的欺诈模型 …… 106
6.4.2 基于判别的欺诈模型 …… 109
6.4.3 基于分类算法的欺诈模型 …… 110
6.5 欺诈人口属性分析 …… 114
6.5.1 欺诈人口属性统计分析 …… 115
6.5.2 基于逻辑回归的欺诈人口属性分析 …… 116
6.5.3 逾期还款的客户特征 …… 119
6.5.4 基于决策树分析逾期客户特征 …… 120
6.5.5 基于回归分析逾期客户特征 …… 123

6.5.6 根据消费历史分析客户特征 …… 128
6.5.7 基于聚类分析客户特征 …… 128
6.5.8 基于客户细分的聚类分析 …… 134

第 7 章 海底捞火锅运营分析 …… 138

7.1 火锅相关数据抓取 …… 139
7.2 数据预处理 …… 140
7.3 数据分析 …… 145
7.3.1 海底捞运营分析 …… 145
7.3.2 店铺选址分析 …… 148
7.4 菜品关联分析 …… 153
7.5 用户评论与评分的关联分析 …… 160
7.6 顾客情感分析 …… 168

第 8 章 商务宾馆竞争分析 …… 172

8.1 目前经济型酒店行业竞争态势 …… 172
8.2 用户相关数据准备 …… 174
8.3 通过 Python 编程抓取评论 …… 180
8.4 数据预处理 …… 183
8.5 商务宾馆客户数据分析 …… 184
8.5.1 酒店评分影响因素 …… 184
8.5.2 酒店评分与酒店业绩关系 …… 187
8.5.3 酒店评分分析 …… 189
8.5.4 客户情感分析 …… 198
8.5.5 竞争分析 …… 205
8.6 建议 …… 214

第 9 章 耐热导线工厂质量管理数据分析 …… 215

9.1 项目概述 …… 215
9.2 耐热导线生产质量数据预处理 …… 216
9.3 耐热铝线质量检测数据分析 …… 218

第 10 章 基于逻辑回归模型的高危人员分析 …… 225

10.1 高危人员分析需求 …… 226
10.2 高危人群相关数据收集与预处理 …… 226
10.3 建立模型 …… 229

第 11 章 卷积神经网络在音频质量评价领域的应用 …… 236

11.1 深度学习基础 …… 236

11.1.1 深度学习的发展过程 …… 237
11.1.2 深度学习常用技术框架 …… 237
11.1.3 常用的深度学习算法 …… 239
11.2 音频质量评价 …… 241
11.2.1 音频样本及特征预处理 …… 242
11.2.2 音频特征选择 …… 244
11.2.3 卷积神经网络模型训练 …… 245
11.2.4 模型参数调优 …… 248
11.3 性能验证 …… 249
参考文献 …… 251

第1章

数据分析过程的主要问题

数据分析是一种入门容易但要精通却很难的学科。做好数据分析并非依赖于某一种技术或方法，其关键是分析思路，通过对业务进行调研，思考过程具有逻辑，并引入一定的创新理念，最后形成可行性建议。数据分析人员为了完成分析任务，获得较好的分析结果，不仅要懂得行业知识，对业务流程有一定的了解，还要理解数据背后的隐含信息，能够对数据进行合理的解读，而且要从变化的角度和时间维度对需求进行把握，确定用哪些数据来解决行业问题，这是数据分析的基础。

数据分析的主要流程是：明确分析目标、数据收集、数据预处理、建模分析、结果评估、结论整理及建议，通过对现状、原因等分析最终实现预测分析，确保数据分析维度的充分性和结论的合理有效性。

1.1 业务理解

数据分析过程中需要理解需求和分析目标，深入理解与分析目标相关联的业务背景，包括行业知识、领域知识及业务流程等，若数据分析人员对业务背景不熟悉，其分析方法和过程就难以贴合实际需求。业内专业人员往往以数据分析人员分析的结论为常识。

为了从数据中挖掘出有价值的结果，与领域专家进行充分交流，要亲临一线去了解业务实际情况，切忌“数据空想”，对业务知识理解其逻辑和原理，不仅有助于在数据预处理过程中对异常数据进行甄别和剔除，而且有助于分析过程中数据探索和挖掘方法的选择，对于结果是否符合预期，也可直观得出结论，否则容易出现模型的准确率虽然很高，经过业务专家评价时发现模型的某一自变量为目标变量的特征表现，最终模型毫无价值。

对数据分析目标的理解，包括定性分析和定量分析，前者给出与目标变量关联的自变量列表或目标变量的性质预测等，后者除了列举相关自变量，还要对其权重等进行定量分析，

在实际数据分析过程中，需要依据不同的业务目标设计分析方案。

在业务理解中，要以方法论的层面进行流程梳理，以实现快速确认分析目标相关联的影响因素，将分析过程以结构化的方式展现，利于理顺思路，而且不局限于某一行业应用，只要变换行业影响因素，即可应用于其他行业。例如，在企业经营活动的分析中，可以应用图 1.1 所示的分析框架，其中主要包括产业基础、企业运营分析、企业财务分析、竞争分析、营销分析、客户分析，此分析框架基本涵盖了大部分的企业经营活动，具体分析中可以适当进行增减和完善，并且可以按照不同的行业进行细化，形成行业分析框架。

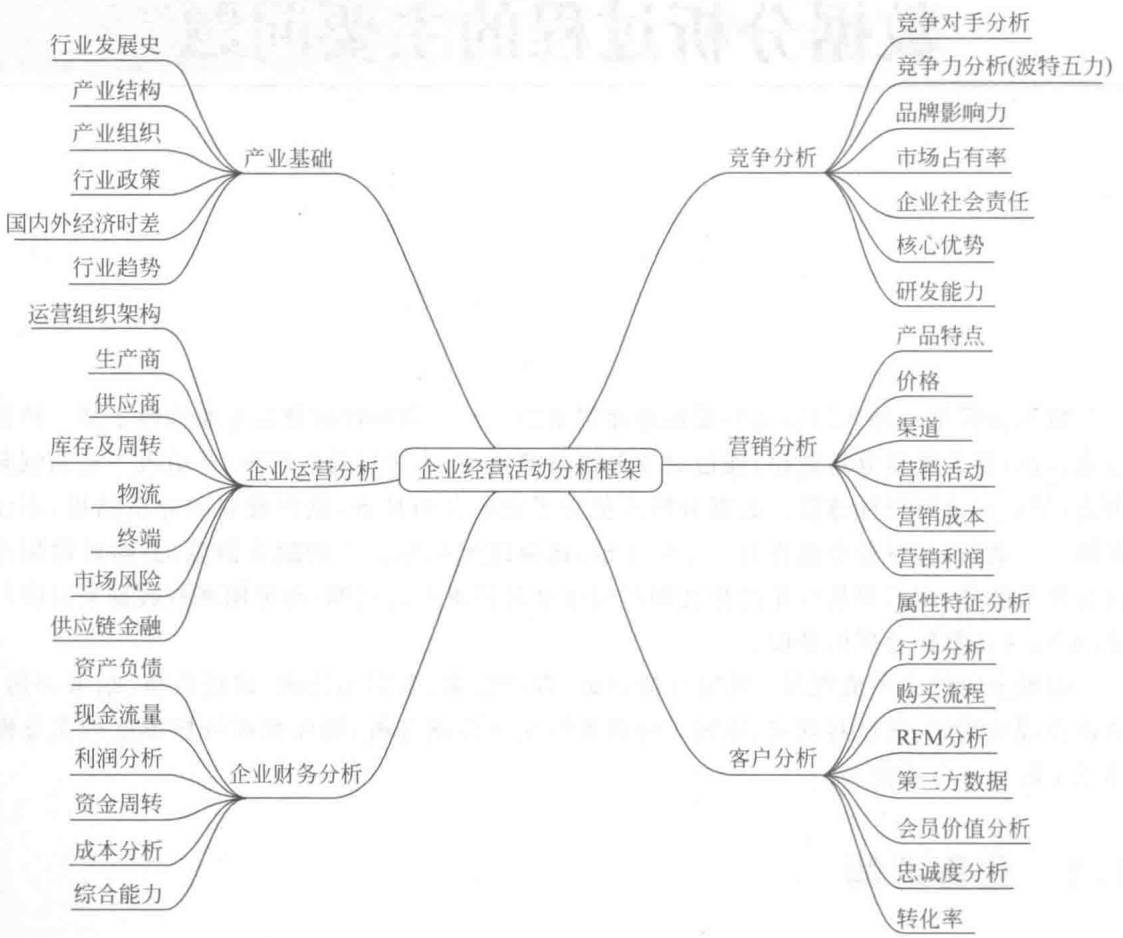


图 1.1 企业经营活动分析框架

对业务理解的分析框架中，主要从宏观的角度结构化、模块化指导数据分析，把问题分解成各个相关联的子模块，为后续数据分析进行规划，起到提纲挈领的作用。

1.2 数据理解

数据分析从字面上看是由数据和分析两部分组成的，其中数据是基础和根本，没有数据样本作为支撑，再好的结论也是无本之木，对现有数据理解到位有助于建立合理的分析框架。分析目标相关联的自变量数据往往可遇不可求，多数情况下，数据资料与分析的目标没

有直接相关性，需要对数据本身进行探索，查看其数据特性或样本特征，结合这些特征来挖掘其与分析目标之间的关系。

为了提高数据分析的准确性，需要多维的源数据，数据量较大可能会产生更多的冗余数据，处理过程较麻烦，经过预处理和降维后，可以得到更多样的支持数据，在初创型企业的数据量较少的情况下，可通过爬虫抓取非结构化数据，并转化为结构化数据作为补充。

了解业务流程中数据产生过程，明确数据代表的意义，并对数据的结构和各字段之间的关系进行分析，在分析过程中需要结合业务逻辑，对数据的理解是整个数据分析过程的基础，如果这一过程出现问题，将影响最终分析结果的正确性。

从历史的角度，数据的产生过程本身是变化的，在时间的维度上，不仅要关心数据是如何产生的及产生的频度，还要关心用户的动作数据，这些都将产生趋势特征，在数据分析过程中，需要关注业务变化导致的数据变化。

同时，由于需求会发生变化，新的数据会加入进来，数据分析方案也要具有一定的扩展性，以应对企业发展的变化和原始数据变化带来的影响，能够在设计模型后对其进行修正和动态改进。

1.3 数据质量问题与预处理

数据质量要求数据是完整的和真实的，并且具有一致性和可靠性。在数据分析过程中，高质量的数据更容易具有较高的区分度。相反，在数据分析领域，有一个著名的“垃圾进，垃圾出”结论，如果数据具有较多缺失值、异常值和无效记录，那么依此数据建立的模型在实际应用中将无法保证其结果真实和有效，数据预处理占用整个数据挖掘项目 60%的工作量，目标就是保证输入模型的数据是符合业务实际情况的，基于正确的数据，才可以谈模型的选择和应用。

1. 数据量较少

数据挖掘需要有一定的数据量作为支撑，随着数据量的增多，其中的规律越发明显，也更容易发现其中分析目标相关的因素，特别是在神经网络或深度学习等算法中，其前提条件就要求有大量的训练数据，否则就容易引起模型过拟合的问题。

数据分析过程中一般要将样本划分为训练集、验证集、测试集，如果数据量较少，可以只需要训练集和测试集，其中训练集的数据量一般为 50%～80%。在某些数据质量较高、区分度较明显的业务场景中，数据量可以更少，一般来说，数据量是自变量数量的 10～20 倍为佳。

在数据的数量足够多的情况下，还要关注数据的质量，如果给定的数据虽然较多，但其中样本的覆盖范围较少，与分析目标相关维度的数据数量才是关键的，否则最终分析得到的结论可能会有较大的局限，不能完全反映数据的本质。

2. 数据量过多

数据集中数量过多时，对全部数据集进行分析要耗费更多的计算资源，要求硬件配置较高，并且由于数据中各类数据的比例往往是不平衡的，例如，两家公司的产品销售的开始时间点并不一致，其销量相差悬殊，如果直接应用到模型中进行竞争分析，则可能出现较大的

结果误差，这种情况可以应用数据采样技术随机提取样本子集。

在面对海量的同质化数据时，如商品交易数据，可以通过聚集技术按照时间、空间等属性进行平均值等汇总，减少数据数量，由于采用了统计汇总后的数据，结果的可视化层次更高，也更加稳定，缺点是可能存在细节丢失的情况。

另外一种情况是在小概率事件的处理中需要关心数据集的不平衡问题。例如，在车辆运行异常检测时，车辆正常运行的时间远超过出现故障的时间，所以正常的数据量占了绝大多数，异常数据量极少，或者是在广告点击事件、地震检测、入侵检测、垃圾邮件过滤等这类稀有事件的分析中，要对数据集应用采样技术，或对异常数据进行复制，提高其占比。

3. 维度灾难

当数据中的自变量较多时，会出现维度灾难问题，特别是在矩阵数据中，其中冗余变量占比较高时，可用数据变成稀疏矩阵，在分类算法处理时就没办法可靠地进行类别划分，在聚类算法中则容易使聚类质量下降，为了从中获得稳定的分析结果，需要耗费大量的运算时间，分析过程低效，为了应对此问题，可以采用线性代数的相关方法将数据从高维空间影射到低维空间中，其中主成分分析(PCA)、奇异值分解(SVD)等方法比较常用。

下面通过对信用卡消费行为与是否存在欺诈进行分析，来展示 PCA 的主要用法。信用卡用户消费统计记录如图 1.2 所示，其中包括了卡类别、日均消费金额、日均次数等消费行为统计后的结果值，还包括用户的属性信息，如性别、年龄、职业等，排除目标字段，共有 19 个输入变量可供选择。

	卡类别	日均消费金额	日均次数	单笔消费最低	单笔消费最高	年收入	是否存在欺诈	性别	年龄	婚姻	户籍	教育	居住类型	职业	工作年限	保险缴纳	车辆情况	总评分	信用等级	额度
1	普卡	764	6	45.300	1127.200	54000	0	男	32	已婚	河南	本科	租房	个体户	9	有	无	60	D-风险客户	10000
2	普卡	797	2	48.000	1303.000	56000	0	女	34	已婚	河北	本科	租房	个体户	11	有	无	60	D-风险客户	10000
3	普卡	106	4	6.600	129.900	22297	0	女	60	已婚	北京	本科	自购房	私营企业	37	有	有	60	D-风险客户	10000
4	普卡	800	2	48.000	1308.000	56000	0	男	34	已婚	北京	本科	租房	私营企业	11	有	无	60	D-风险客户	10000
5	普卡	968	4	58.700	2411.900	67400	0	男	31	未婚	浙江	大专	租房	私营企业	11	有	无	60	D-风险客户	10000
6	普卡	101	3	5.000	102.000	19360	0	男	32	已婚	湖南	硕士及以上	自购房	私营企业	7	有	有	60	D-风险客户	10000
7	普卡	106	4	6.700	130.200	22599	0	女	45	已婚	上海	本科	自购房	私营企业	22	有	有	60	D-风险客户	10000
8	普卡	800	2	48.100	1315.300	56000	0	女	33	已婚	黑龙江	本科	租房	个体户	10	有	无	60	D-风险客户	10000
9	普卡	366	3	26.700	634.700	39198	0	女	34	已婚	海南	本科	租房	私营企业	11	有	无	60	D-风险客户	10000
10	普卡	107	2	6.900	131.500	22975	0	男	29	未婚	天津	硕士及以上	自购房	私营企业	4	有	有	60	D-风险客户	10000

图 1.2 信用卡用户消费统计记录

在 SPSS Modeler 中应用主成分分析/因子节点对数据进行降维，选择日均消费金额等 9 个字段作为输入，以 70%训练集、30%测试集的比例进行分区，选择“专家”模式，参数为默认值，运行后的主要结果如图 1.3 所示。

在总方差解释表中，前 4 个变量的初始特征值大于 1，分别为日均消费金额、日均次数、单笔消费最低、单笔消费最高，这 4 项累积占全部变量的 84.507%，也符合主成分的 80%以上的标准，说明这 4 项作为输入变量比较合理。

降低维度的另一种方法是通过特征子集选择的方式，将那些不相关的特征，如身份证号、姓名等剔除，只选择与目标变量紧密相关的特征。除了剔除属性，还可以使用特征加权技术，结合领域知识人为赋予某些特征更大的影响力权重。

在深度学习领域，常用特征提取和特征创建的技术将原始数据中的特征进行重构，以获得模型需要的特征，并且在重构过程中加以格式转换和数据变换。常用的技术包括傅里叶变换和小波变换，前者将时域信号转化为频域信号，后者主要处理时间序列等类型。

4. 数据不完整

除了数据量要多，还要求数据的种类要多。例如，要对企业产品的销售情况进行分析或预测，除了需要有企业产品相关的市场、销售情况等信息外，还需要有客户相关资料、竞品的

公因子方差

	初始	提取
日均消费金额	1.000	0.914
日均次数	1.000	0.697
单笔消费最低	1.000	0.922
单笔消费最高	1.000	0.785
年收入	1.000	0.553
年龄	1.000	0.927
工作年限	1.000	0.928
总评分	1.000	0.938
额度	1.000	0.942

提取方法：主成分分析法。

总方差解释

成分	初始特征值			提取载荷平方和		
	总计	方差百分比/%	累积/%	总计	方差百分比/%	累积/%
1	3.351	37.234	37.234	3.351	37.234	37.234
2	1.901	21.123	58.357	1.901	21.123	58.357
3	1.280	14.217	72.575	1.280	14.217	72.575
4	1.074	11.932	84.507	1.074	11.932	84.507
5	0.783	8.704	93.211			
6	0.304	3.379	96.590			
7	0.145	1.611	98.201			
8	0.112	1.243	99.444			
9	0.050	0.556	100.000			

提取方法：主成分分析法。

图 1.3 PCA 主成分分析结果示例

销售情况、市场数据、财务数据等，甚至要有交通物流、CPI 等宏观数据支持，但是现实情况中，很多数据缺失，要么这些数据并没有进行记录，要么它们在竞争对手的系统中，无法获得，这种情况将直接影响数据挖掘方法的选择，此时可以通过编写程序，来抓取外部数据作为补充。

数据缺失也是数据不完整的一种表现，可能是空白值或空值，也可能是存在大量的无效值，例如，所有记录的某一字段值均相同，或者某一字段中超过一半的记录为空或无效，在出现数据缺失时，分析人员要查找缺失原因，是原信息录入系统缺陷，还是人为操作失误，或者字段为选填等业务原因，并按照不同的原因进行数据预处理。例如，由于系统 Bug 导致的，则需要修复 Bug 并重新计算，如果当前字段中的数值是随时间逐渐生成的，则为业务原因，需要结合实际业务进行处理。

对缺失值可以采用众数、中位数、均值、最近距离等方法对缺失值进行人为补充，或者也可以通过回归或贝叶斯定理等预测缺失值。为了提高数据的纯度，也可以删除含有缺失值的记录，但如果缺失值的记录数较多时，删除操作可能会丢失样本特征，此时可以删除对应的字段，对于缺失值超过 30%的字段，可不作为模型输入变量。

5. 异常数据

在数据收集阶段由于人为或系统处理等原因，会导致产生异于常规的数据。其中异常数据分为两类：一类是错误的数据；另一类为小概率事件，或称为稀有事件。在系统预处

理阶段要视情况对数据进行探索，并结合行业内的业务知识对其进行识别，一旦发现错误数据，则将其剔除或修正。对于稀有事件，如信用卡欺诈行为、垃圾邮件等，这类正常数据不但不能修正和删除，反而要重点分析其特征。

通过查看散点图或箱图的方式查看离群点信息，如图 1.4 所示，可以看到方框中的年收入达到 21 亿元，已经超过绝大多数人的收入范围，极有可能为异常数据。还可基于距离或统计模型等进行检测，如应用线性回归、主成分分析等方式来区分异常数据，除此之外，还可应用深度学习（如 RNN 方法）来检测。

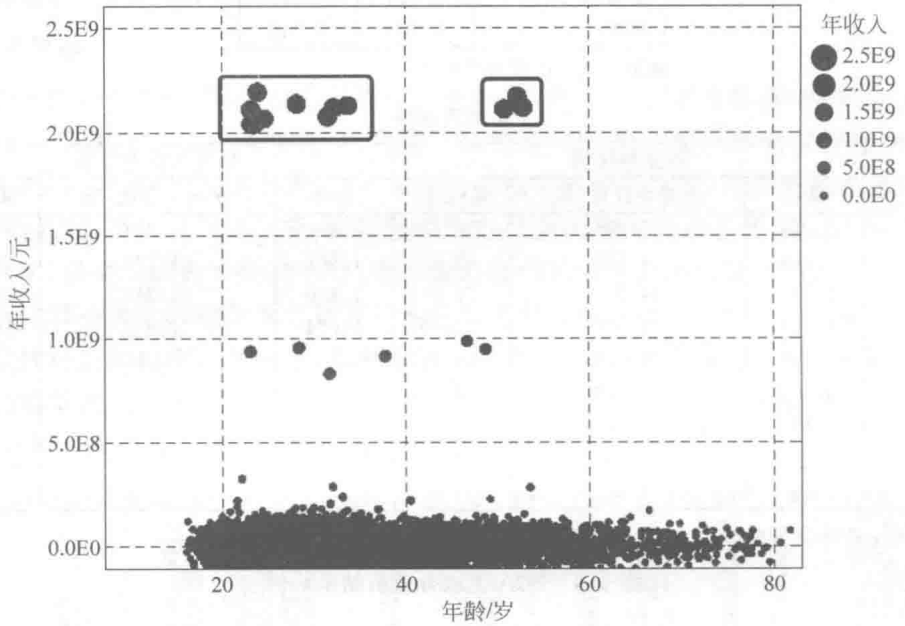


图 1.4　通过散点图查看离群点

当异常数据并非在离群点时，没有显著异常，可能是由于人为输入错误或系统误差导致的，虽然这些数值是不正确的，但是由于其与真实值之间区分较少，所以较难发现这类噪声数据。可以通过抽样的方式进行人工检测，或者对比不同数据源系统中的数据，进行一致性检测。

6. 重复数据

在数据分析中如果出现较多的重复数据，将对模型的结果产生误差，在数据处理过程中可以使用 SQL 或 Excel 中的去重复方法将重复数据滤除。有时候在记录中所有字段都是非重复数据，但选择其中部分字段时则容易产生重复样本，即样本子集中含有重复数据，特别是手动选取某几个字段作为模型输入时，容易忽略这一细节，所以，在将其应用到模型之前，需要进行过滤，将重复数据滤除。在 SPSS 中可以使用“区分”节点，对选择的自变量进行去重。利用“区分”节点去重复如图 1.5 所示。

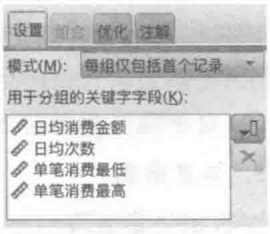


图 1.5　利用“区分”节点去重复

在模式中选择“每组仅包括首个记录”，其他重复的记录将滤除，用于分组的字段即为流向下一节点的变量，只有日均消费金额等 4 个字段中的值均为重复时，才会被滤除。