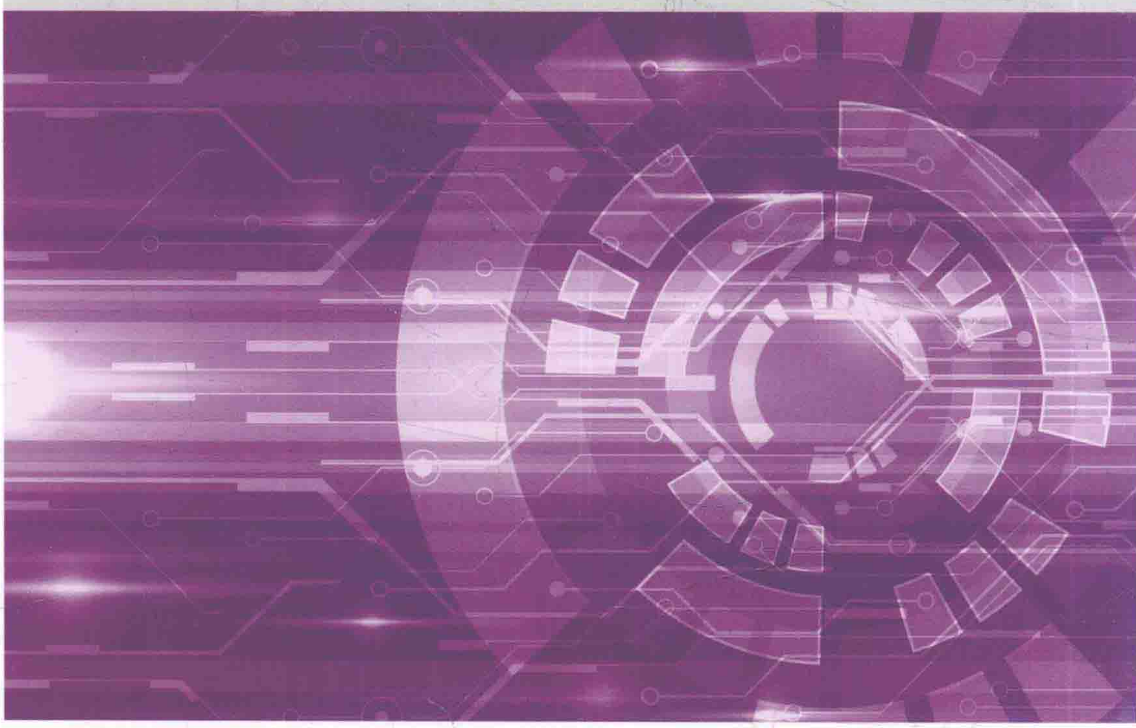
· 大数据应用人才培养系列教材 ·

数据挖掘基础

■总主编◎刘　鹏　张　燕　■主编◎陶建辉　■副主编◎姜才康



清华大学出版社

大数据应用人才培养系列教材

数据挖掘基础

总主编　刘　鹏　张　燕

主　编　陶建辉

副主编　姜才康

清华大学出版社

北　京

内容简介

本书介绍了数据挖掘的基本概念，包括数据挖掘的常用算法、常用工具、用途和应用场景及应用状况，讲述了常用数据挖掘方法，如分类、聚类、关联规则的概念、思想、典型算法、应用场景等。此外，本书还从实际应用出发，讲解了基于日志的大数据挖掘技术的原理、工具、应用场景和成功案例。日志挖掘技术现在已得到了广泛的运用。

通过以上内容的学习，读者将了解数据挖掘的基本概念、思想和算法，并掌握其应用要领。本书可以作为培养应用型人才的课程教材，也可作为相关开发人员的自学教材和参考手册。

图书在版编目（CIP）数据

数据挖掘基础/陶建辉主编．—北京：清华大学出版社，2018
（大数据应用人才培养系列教材）
ISBN 978-7-302-50219-7

Ⅰ．①数…　Ⅱ．①陶…　Ⅲ．①数据采集-教材　Ⅳ．①TP274

中国版本图书馆 CIP 数据核字（2018）第 114763 号

责任编辑：贾小红
封面设计：刘　超
版式设计：魏　远
责任校对：马子杰
责任印制：刘海龙

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
印 装 者：北京密云胶印厂
经　　销：全国新华书店
开　　本：185mm×260mm　　印　　张：11.5　　字　　数：266 千字
版　　次：2018 年 6 月第 1 版　　印　　次：2018 年 6 月第 1 次印刷
印　　数：1～2500
定　　价：48.00 元

产品编号：075033-01

编写委员会

总主编　刘　鹏　张　燕
主　编　陶建辉
副主编　姜才康
参　编　徐　昉　袁　华　梁英杰
　　　　王海洋　朱　辉

总　序

短短几年间，大数据就以一日千里的发展速度，快速实现了从概念到落地，直接带动了相关产业的井喷式发展。数据采集、数据存储、数据挖掘、数据分析等大数据技术在越来越多的行业中得到应用，随之而来的就是大数据人才缺口问题的凸显。根据《人民日报》的报道，未来3~5年，中国需要180万大数据人才，但目前只有约30万人，人才缺口达到150万之多。

大数据是一门实践性很强的学科，在其呈现金字塔型的人才资源模型中，数据科学家居于塔尖位置，然而该领域对于经验丰富的数据科学家需求相对有限，反而是对大数据底层设计、数据清洗、数据挖掘及大数据安全等相关人才的需求急剧上升，可以说占据了大数据人才需求的80%以上。比如数据清洗、数据挖掘等相关职位，需要大量的专业人才。

迫切的人才需求直接催热了相应的大数据应用专业。2018年1月18日，教育部公布了“大数据技术与应用”专业备案和审批结果，已有270所高职院校申报开设“大数据技术与应用”专业，其中共有208所职业院校获批“大数据技术与应用”专业。随着大数据的深入发展，未来几年申请与获批该专业的职业院校数量仍将持续走高。同时，对于国家教育部正式设立的“数据科学与大数据技术”本科新专业，除已获批的35所大学之外，2017年申请院校也高达263所。

即使如此，就目前而言，在大数据人才培养和大数据课程建设方面，大部分专科院校仍然处于起步阶段，需要探索的问题还有很多。首先，大数据是个新生事物，懂大数据的老师少之又少，院校缺“人”；其次，院校尚未形成完善的大数据人才培养和课程体系，缺乏“机制”；再次，大数据实验需要为每位学生提供集群计算机，院校缺“机器”；最后，院校没有海量数据，开展大数据教学实验工作缺少“原材料”。

对于注重实操的大数据技术与应用专业专科建设而言，需要重点面向网络爬虫、大数据分析、大数据开发、大数据可视化、大数据运维工程师的工作岗位，帮助学生掌握大数据技术与应用专业必备知识，使其具备大数据采集、存储、清洗、分析、开发及系统维护的专业能力和技能，成为能够服务区域经济的发展型、创新型或复合型技术技能人才。

无论是缺“人”、缺“机制”、缺“机器”，还是缺少“原材料”，最终都难以培养出合格的大数据人才。

其实，早在网格计算和云计算兴起时，我国科技工作者就曾遇到过类似的挑战，我有幸参与了这些问题的解决过程。为了解决网格计算问题，我在清华大学读博期间，于 2001 年创办了中国网格信息中转站网站，每天花几个小时收集和分享有价值的资料给学术界，此后我也多次筹办和主持全国性的网格计算学术会议，进行信息传递与知识分享。2002 年，我与其他专家合作的《网格计算》教材正式面世。

2008 年，当云计算开始萌芽之时，我创办了中国云计算网站（chinacloud.cn）（在各大搜索引擎“云计算”关键词中排名第一），2010 年出版了《云计算（第 1 版）》，2011 年出版了《云计算（第 2 版）》，2015 年出版了《云计算（第 3 版）》，每一版都花费了大量成本制作并免费分享对应的几十个教学 PPT。目前，这些 PPT 的下载总量达到了几百万次之多。同时，《云计算》一书也成为了国内高校的优秀教材，在中国知网公布的高被引图书名单中，《云计算》在自动化和计算机领域排名全国第一。

除了资料分享，在 2010 年，我们在南京组织了全国高校云计算师资培训班，培养了国内第一批云计算老师，并通过与华为、中兴、360 等知名企业合作，输出云计算技术，培养云计算研发人才。这些工作获得了大家的认可与好评，此后我接连担任了工信部云计算研究中心专家、中国云计算专家委员会云存储组组长、中国大数据应用联盟人工智能专家委员会主任等。

近几年，面对日益突出的大数据发展难题，我们也正在尝试使用此前类似的办法去应对这些挑战。为了解决大数据技术资料缺乏和交流不够通透的问题，我们于 2013 年创办了中国大数据网站（thebigdata.cn），投入大量的人力进行日常维护，该网站目前已经在各大搜索引擎的“大数据”关键词排名中位居第一；为了解决大数据师资匮乏的问题，我们面向全国院校陆续举办多期大数据师资培训班，致力于解决“缺人”的问题。

2016 年年末至今，我们已在南京多次举办全国高校/高职/中职大数据免费培训班，基于《大数据》《大数据实验手册》以及云创大数据提供的大数据实验平台，帮助到场老师们跑通了 Hadoop、Spark 等多个大数据实验，使他们跨过了“从理论到实践，从知道到用过”的门槛。

其中，为了解决大数据实验难问题而开发的大数据实验平台，正在

为越来越多的高校教学科研带去方便，帮助解决“缺机器”与“缺原材料”的问题。2016 年，我带领云创大数据（www.cstor.cn，股票代码：835305）的科研人员，应用 Docker 容器技术，成功开发了 BDRack 大数据实验一体机，它打破了虚拟化技术的性能瓶颈，可以为每一位参加实验的人员虚拟出 Hadoop 集群、Spark 集群、Storm 集群等，自带实验所需数据，并准备了详细的实验手册（包含 42 个大数据实验）、PPT 和实验过程视频，可以开展大数据管理、大数据挖掘等各类实验，并可进行精确营销、信用分析等多种实战演练。

目前，大数据实验平台已经在郑州大学、成都理工大学、金陵科技学院、天津农学院、西京学院、郑州升达经贸管理学院、信阳师范学院、镇江高等职业技术学校等多所院校部署应用，并广受校方好评。该平台也可以云服务的方式在线提供（大数据实验平台，https://bd.cstor.cn），实验更是增至 85 个，师生通过自学，可用一个月时间成为大数据实验动手的高手。此外，面对席卷而来的人工智能浪潮，我们团队推出的 AIRack 人工智能实验平台、DeepRack 深度学习一体机以及 dServer 人工智能服务器等系列应用，一举解决了人工智能实验环境搭建困难、缺乏实验指导与实验数据等问题，目前已经在清华大学、南京大学、南京农业大学、西安科技大学等高校投入使用。

在大数据教学中，本科院校的实践教学应更加系统性，偏向新技术的应用，且对工程实践能力要求更高。而高职、高专院校则更偏向于技术性和技能训练，理论以够用为主，学生将主要从事数据清洗和运维方面的工作。基于此，我们联合多家高职院校专家准备了《云计算导论》《大数据导论》《数据挖掘基础》《R 语言》《数据清洗》《大数据系统运维》《大数据实践》系列教材，帮助解决“机制”欠缺的问题。

此外，我们也将继续在中国大数据（thebigdata.cn）和中国云计算（chinacloud.cn）等网站免费提供配套 PPT 和其他资料。同时，持续开放大数据实验平台（https://bd.cstor.cn）、免费的物联网大数据托管平台万物云（wanwuyun.com）和环境大数据免费分享平台环境云（envicloud.cn），使资源与数据随手可得，让大数据学习变得更加轻松。

在此，特别感谢我的硕士导师谢希仁教授和博士导师李三立院士。谢希仁教授所著的《计算机网络》已经更新到第 7 版，与时俱进日臻完美，时时提醒学生要以这样的标准来写书。李三立院士是留苏博士，为我国计算机事业做出了杰出贡献，曾任国家攀登计划项目首席科学家。他的严谨治学带出了一大批杰出的学生。

本丛书是集体智慧的结晶，在此谨向付出辛勤劳动的各位作者致敬！书中难免会有不当之处，请读者不吝赐教。我的邮箱：gloud@126.com，微信公众号：刘鹏看未来（lpoutlook）。

刘 鹏
于南京大数据研究院
2018 年 5 月

前　言

数据挖掘是知识发现不可缺少的部分，是将未加工的数据转换为有用信息的过程。为了贯彻国家大数据战略，尽快帮助应用型院校学生学习和掌握数据挖掘的基本知识以及基本应用技能，我们以通俗、简明并结合实际应用的方式编写了《数据挖掘基础》教材。

《数据挖掘基础》教材讲述了数据挖掘概念、数据挖掘的常用方法，包括分类方法、聚类方法和关联规则方法。此外，本教材还从实际应用出发，讲解了日志的挖掘与应用方法。

分类是数据挖掘中的一种重要方法，在给定数据基础上构建分类函数或分类模型，该函数或模型能够将数据归类为给定类别中的某一类别，就是分类。一般通过构建分类器实现具体分类，分类器是对样本进行分类方法的统称。本教材将对分类的基本概念及知识，如决策树、分类器、贝叶斯分类器、支持向量机等内容进行讲解和研究。

聚类的过程，就是将相似数据归并到一类的过程，形成同类对象具有共同特征，不同类对象之间有显著区别。聚类的目的是通过数据间的相似性将数据归类，并根据数据的概念描述来制定对应的策略。本教材将对聚类基本概念及常用算法进行讲解，着重研究了聚合分析方法，并介绍了聚类方法应用场景。此外，还详细讲解了聚类方法的实现例子。

关于关联规则，我们从营销界流传的“啤酒与尿布”经典案例入手，介绍关联规则的概念、定义和分类，并分析了关联规则的挖掘过程，包括频繁项集产生、强关联规则和关联规则评价标准，重点介绍了关联规则最为经典的算法——Apriori 算法，并分析了关联规则挖掘技术在国内外的应用现状，以及关联规则挖掘实例。

日志分析挖掘的综合实战章节讲述了日志概念、日志处理、日志分析原理及工具、日志挖掘应用，以及日志分析挖掘实例。

我们衷心希望本教材可以帮助读者学习到数据挖掘的基础知识，掌握数据挖掘的基本方法，以及体会到数据挖掘在实际应用中的精妙之处。

感谢编写组的全体老师，他们相互鼓励、相互学习、相互促进，为《数据挖掘基础》教材的编写付出了辛勤的劳动！本书的问世也要感谢清华大学出版社王莉编辑给予的宝贵意见和指导。

《数据挖掘基础》编写组

2018 年 5 月

目　录

第 1 章　数据挖掘概念

1.1　数据挖掘概述……1
1.1.1　什么是数据挖掘……2
1.1.2　数据挖掘常用算法概述……2
1.1.3　数据挖掘常用工具概述……4
1.2　数据探索……5
1.2.1　数据概述……5
1.2.2　数据质量……7
1.2.3　数据预处理……10
1.3　数据挖掘的应用……11
1.3.1　数据挖掘现状及发展趋势……11
1.3.2　数据挖掘需要解决的问题……12
1.3.3　数据挖掘的应用场景……14
1.4　作业与练习……18
参考文献……18

第 2 章　分类

2.1　分类概述……19
2.1.1　分类的基本概念……19
2.1.2　解决分类问题的一般方法……20
2.1.3　决策树……21
案例：Web 机器人检测……23
2.1.4　模型的过分拟合……24
2.2　贝叶斯决策与分类器……25
2.2.1　规则分类器……25
2.2.2　分类中贝叶斯定理的应用……26
2.2.3　分类中朴素贝叶斯的应用……27
2.3　支持向量机……28
2.3.1　最大边缘超平面……29
2.3.2　线性支持向量机 SVM……30
2.3.3　非线性支持向量机 SVM……33

2.4 分类在实际场景中的应用案例……36
案例一：如何解决文章主题关键字与搜索引擎关键字带来的检索结果差异……36
案例二：甄别新金融交易方式的欺诈行为……36
案例三：在线广告推荐中的分类……37
2.5 作业与练习……40
参考文献……41

第 3 章 聚类

3.1 聚类概述……42
3.1.1 聚类的基本概念……42
3.1.2 聚类算法……45
3.2 聚合分析方法……48
3.2.1 欧氏距离……48
3.2.2 聚合过程……49
3.2.3 聚类树……51
3.2.4 聚合分析方法应用例子……52
3.3 聚类在实际场景中的应用案例……53
3.4 聚类的实现例子……54
3.5 作业与练习……61
参考文献……61

第 4 章 关联规则

4.1 关联规则概述……63
4.1.1 经典案例导入……63
4.1.2 关联规则的基本概念和定义……64
4.1.3 关联规则的分类……67
4.2 关联规则的挖掘过程……68
4.2.1 知识回顾……68
4.2.2 频繁项集产生……69
4.2.3 强关联规则……71
4.2.4 关联规则评价标准……71
4.3 关联规则的 Apriori 算法……73
4.3.1 知识回顾……73
4.3.2 Apriori 算法的核心思想……74
4.3.3 Apriori 算法描述……74
4.3.4 Apriori 算法评价……76

4.3.5 Apriori 算法改进 …… 77
4.4 关联规则的 FP-growth 算法 …… 78
4.4.1 构建 FP 树 …… 79
4.4.2 从 FP 树中挖掘频繁项集 …… 82
4.5 实战：关联规则挖掘实例 …… 83
4.5.1 关联规则挖掘技术在国内外的应用现状 …… 83
4.5.2 关联规则应用实例 …… 83
4.5.3 关联规则在大型超市中应用的步骤 …… 86
4.6 作业与练习 …… 88
参考文献 …… 88
第 5 章 综合实战——日志的挖掘与应用
5.1 日志概念 …… 90
5.1.1 日志是什么 …… 91
5.1.2 日志能做什么 …… 91
5.2 日志处理 …… 93
5.2.1 产生日志 …… 93
5.2.2 传输日志 …… 93
5.2.3 存储日志 …… 96
5.2.4 分析日志 …… 100
5.2.5 日志规范与标准 …… 111
5.3 日志分析原理及工具 …… 113
5.3.1 日志分析原理 …… 114
5.3.2 日志分析工具 …… 120
5.3.3 日志分析系统规划建设 …… 123
5.4 日志挖掘应用 …… 127
5.4.1 安全运维 …… 127
5.4.2 系统健康分析 …… 128
5.4.3 用户行为分析 …… 129
5.4.4 业务分析设计 …… 130
5.5 日志分析挖掘实例 …… 131
5.6 作业与练习 …… 133
参考文献 …… 133
第 6 章 数据挖掘应用案例
6.1 电力行业采用聚类方法进行主变油温分析 …… 134

6.1.1 需求背景及采用的大数据分析方法 …… 134
6.1.2 大数据分析方法的实现过程 …… 135
6.1.3 大数据分析方法的实现结果 …… 137
6.2 银行信贷评价 …… 138
6.2.1 简介 …… 138
6.2.2 神经网络模型 …… 138
6.2.3 实证检验 …… 139
6.3 指数预测 …… 140
6.3.1 金融时间序列概况 …… 140
6.3.2 小波消噪 …… 141
6.3.3 向量机 …… 142
6.3.4 指数预测 …… 143
6.4 客户分群的精准智能营销 …… 143
6.4.1 挖掘目标 …… 143
6.4.2 分析方法和过程 …… 144
6.4.3 建模仿真 …… 148
6.5 使用 WEKA 进行房屋定价 …… 150
6.6 作业与练习 …… 154
参考文献 …… 155

附录 A 大数据和人工智能实验环境

附录 B Hadoop 环境要求

附录 C 名词解释

第1章

数据挖掘概念

数据挖掘是什么？与现有的统计学、概率学、信息学等学科有什么不一样？作为一门新兴的学科，数据挖掘有两个特点：一是数据的广泛性、多样性；二是数据研究的共性。数据的类型多种多样，既有传统的结构化数据，也有网页、文本、图像、视频、语音等非结构化数据。数据挖掘主要包括两个方面：一方面用数据挖掘算法、工具来研究数据；另一方面将获得的知识应用到各个领域中。在数据采集中并非所有的信息发现任务都被视为数据挖掘，如通过 Inter 搜索引擎查找特定的 Web 页面属于信息检索领域。数据挖掘是将未加工的数据通过相应的算法、工具转换为有价值的信息的过程。

数据挖掘的应用对现代社会的影响是多方面的，如对社会学研究有着巨大的影响，一是社交网络、网络科学的研究形成新的研究层面，同时提供了新的研究方向、新的实用价值，如广告精准投放、热点及舆情分析等；二是新的数据来源和数据挖掘算法、工具使它的研究进一步量化、去经验化等。

1.1 数据挖掘概述

数据挖掘知识体系涉及内容广泛，本节将主要介绍一些基本的概念、算法和工具。

1.1.1 什么是数据挖掘

数据挖掘（Data Mining）就是从大量的、不完全的、有噪声的、模糊的、随机的数据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。数据挖掘的数据源包括数据库、数据仓库、Web 或其他数据存储库。

世界幸福报告（The World Happiness Report）是一个具有标志性意义的调查报告（官方网址：http://worldhappiness.report/），旨在对世界各国幸福状态进行了解和研究。世界幸福报告在各国政府、组织、民间广泛使用，它被用来指导其政策趋向及其政策开展的效果，获得了几乎全球的认可。2016 年的世界幸福报告显示，几乎全球区域以及整个世界人口，各个国家之间的幸福不平等程度显著增加。2016 年度的世界幸福报告（数据来源官方网址：https://www.kaggle.com/ unsdsn/world-happiness）涵盖了全球的 157 个国家。

国家统计、经济学、心理学、调查分析、健康、公共政策等各个领域的专家们通过这些指数来研究评估一个国家的发展状况。数据挖掘在这个研究评估中起到了关键作用，其中涉及的如数据集、数据分类、层次聚类、数据可视化等。

并非所有的信息发现任务都可称作数据挖掘，例如，在数据库管理系统中检索一条记录，属于信息检索（Information Retrieval）领域的任务，尽管也是从大量数据中检索有用信息，但并不满足数据挖掘的概念。

1.1.2 数据挖掘常用算法概述

在面对海量数据时，需要使用一定的算法，才能从中挖掘出有用的信息，下面介绍数据挖掘中常用的算法。

1. 分类算法

（1）决策树算法。决策树算法是一种典型的分类算法，首先利用已知分类的数据构造决策树，然后利用测试数据集对决策树进行剪枝，每个决策树的叶子都是一种分类，最后利用形成的决策树对数据进行分类。决策树的典型算法有 ID3、C4.5、CART 等。决策树算法的基本步骤如下。

① 生成决策树。遍历训练集数据，根据数据中具有分类能力的属性作为决策树的节点，不断展开，直到该节点属于一个叶子节点，表明已经找到分类结果。

② 决策树剪枝。决策树剪枝是对步骤①中生成的决策树进行校验和修正，使用测试数据集对分类过程中产生的规则进行校验，剪掉影响分类结果准确性的分枝。

（2）贝叶斯分类算法。贝叶斯分类算法是统计学的一种方法，其中朴素贝叶斯算法在许多情况下可以与决策树和神经网络算法相媲美，而且方法简单，准确度高，速度快。贝叶斯算法是基于贝叶斯定理的，而贝叶斯定理假设一个属性值对给定类的影响独立于其他属性值，但这种假设在很多情况下是不成立的，因此为了降低这个假设的影响，产生了很多改进算法，如 TAN（Tree Augmented Bayes Network）算法。

在朴素贝叶斯算法中，将每个样本数据分为 n 个属性，计算每个属性属于分类 C_i 的概率，根据假设可知，所有属性属于分类 C_i 之积即为该属性属于分类 C_i 的概率，找出概率的最大值，就计算出该属性最有可能属于哪一个类别。

（3）支持向量机。支持向量机（Support Vector Machine，SVM）是建立在统计学理论的 VC 维理论和结构风险最小原理基础上的，它在解决小样本、非线性及高维模式识别中表现出许多特有的优势，并能够推广应用到函数拟合等其他机器学习问题中。支持向量机算法将在后面章节做详细介绍。

2. 聚类算法

聚类算法不同于分类算法，不会考虑类标号，这是因为在很多情况下，开始并不存在类标号。聚类算法可以根据最大化类内相似性、最小化类间相似性的原则进行聚类或分组，这样就形成了对象的簇，同一个簇内的数据具有较高的相似性，不同簇之间的数据具有较低的相似性。常见的分类算法有 K-MEANS 算法、K-MEDOIDS 算法等。在对 2016 世界幸福报告分析时，首先对不同国家根据幸福指数进行聚类，再用不同降维方法在二维空间对数据进行展示，最后分析国家区域与幸福指数的关系。

3. 关联规则

关联规则是形如 $X \rightarrow Y$ 的蕴涵式，X 和 Y 分别称为关联规则的先导和后继。我们先从一个例子中感受一下关联规则的重要性。

这里有一则沃尔玛超市的趣闻。沃尔玛曾经对数据仓库中一年多的原始交易数据进行了详细的分析，发现与尿布一起被购买最多的商品竟然是啤酒。借助数据仓库和关联规则，发现了这个隐藏在背后的事实：美国的妇女经常会嘱咐丈夫下班后为孩子买尿布，而 30%~40%的丈夫

在买完尿布之后又要顺便购买自己爱喝的啤酒。根据这个发现，沃尔玛调整了货架的位置，把尿布和啤酒放在一起销售，大大增加了销量。

从这个例子中我们感受到了关联规则分析的重要性，而关联规则一般分为两个阶段，第一阶段必须先从数据中找出所有高频项目组合，第二阶段再由这些高频项目组中产生关联规则。常用的关联规则算法有Apriori 算法、FP-树频集算法等。

1.1.3 数据挖掘常用工具概述

下面介绍几种常用的数据挖掘工具。

1. Weka 软件

Weka（Waikato Environment for Knowledge Analysis）的全名是怀卡托智能分析环境，是一款免费且非商业化的数据挖掘软件，也是基于Java 环境下开源的机器学习与数据挖掘软件。Weka 的源代码可在其官方网站下载。它集成了大量数据挖掘算法，包括数据预处理、分类、聚类、关联分析等。用户既可以使用可视化界面进行操作，也可以使用Weka 提供的接口，实现自己的数据挖掘算法。图形用户界面包括 Weka Knowledge Flow Environment 和 Weka Explorer。用户也可以使用 Java 语言调用 Weka 提供的类库实现数据挖掘算法，这些类库存在于 weka.jar 中。

2. Clementine（SPSS）软件

Clementine 是 SPSS 所发行的一种资料探勘工具，集成了分类、聚类和关联规则等算法，Clementine 提供了可视化工具，方便用户操作。其通过一系列节点来执行挖掘过程，这一过程被称作一个数据流，数据流上面的节点代表了要执行的操作。Clementine 的资料可视化能力包含散布图、平面图及 Web 分析。

3. KNIME 软件

KNIME（Konstanz InformationMiner）是基于 Eclipse 开发环境来精心开发的数据挖掘工具，可以扩展使用 Weka 中的数据挖掘算法。与Clementine 类似，KNIME 使用类似数据流的方式实现数据挖掘过程，挖掘流程由一系列功能节点组成，每个节点有输入、输出端口，用于接收数据或模型以及导出结果。

4. RapidMiner 软件

RapidMiner 在 2015 年 KDnuggets 举办的第 16 届国际数据挖掘暨