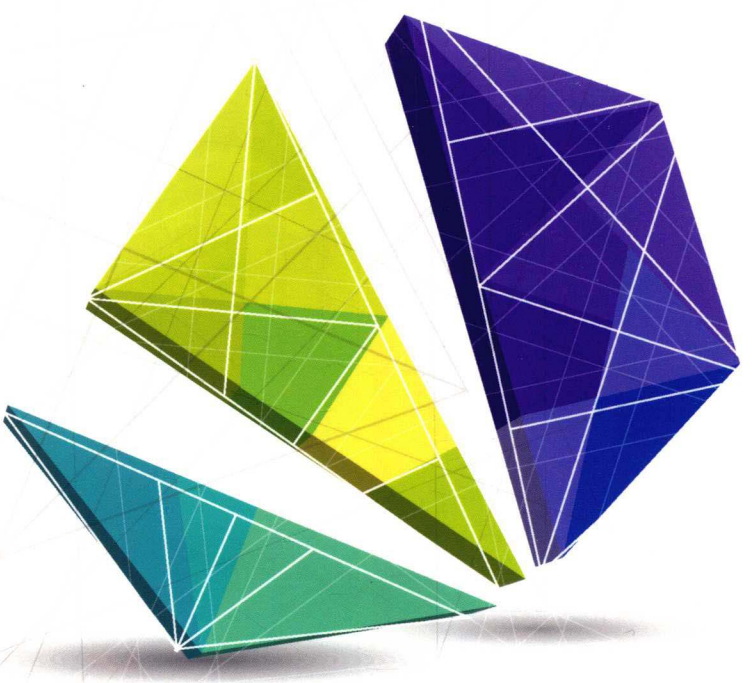
高等学校大数据技术与应用规划教材

数据挖掘

SHUJU WAJUE

宋万清　杨寿渊　陈剑雪　高永彬　编著

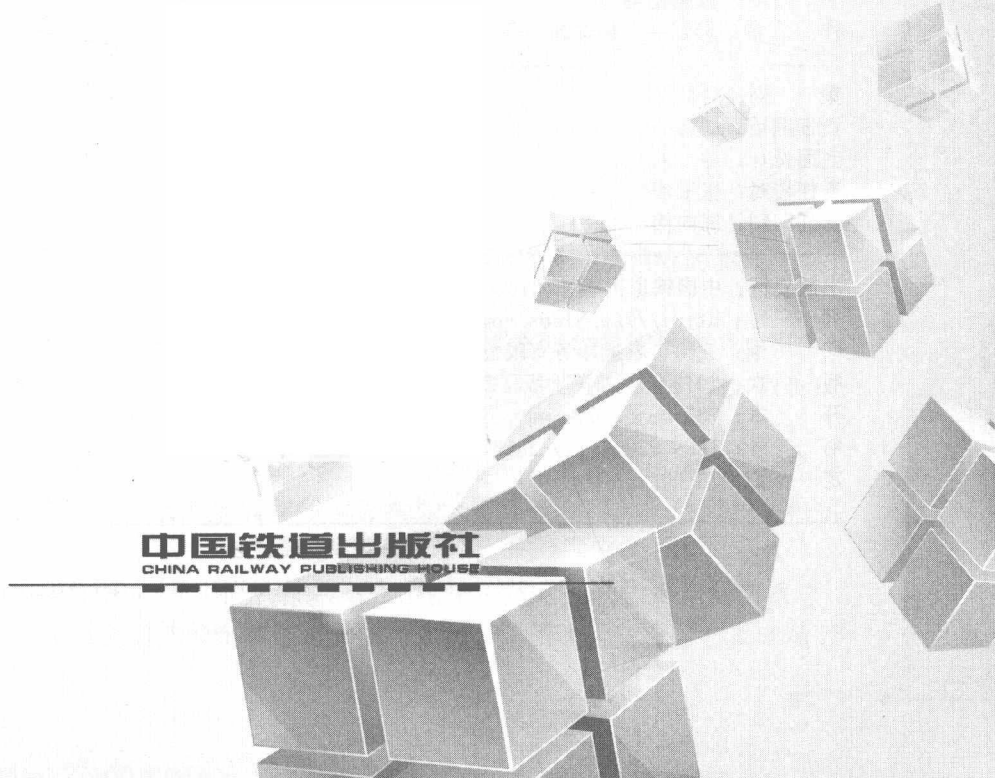


中国铁道出版社
CHINA RAILWAY PUBLISHING HOUSE

术与应用规划教材

数据挖掘

宋万清　杨寿渊　陈剑雪　高永彬　编著
方志军　钱亮宏　主审



中国铁道出版社
CHINA RAILWAY PUBLISHING HOUSE

内 容 简 介

本书着力于介绍数据挖掘基础知识、基本原理、常用算法，主要内容包括数据挖掘概述、数据的描述与可视化、数据的采集和预处理、数据的归约、关联规则挖掘、分类与预测、 非线性预测模型、聚类分析、深度学习简介、使用 Weka 进行数据挖掘。本书通俗易懂，注重基础知识、基本原理和基本方法，注重启发和引申，以培养学生独立思考和独立发现的能力。

本书适合作为数据科学与大数据、信息管理、统计等专业的本科层次基础课教材，也可作为相关专业研究生层次的参考用书。

图书在版编目（CIP）数据

数据挖掘/宋万清等编著. —北京：中国铁道出版社，2018.12

高等学校大数据技术与应用规划教材

ISBN 978-7-113-25167-3

Ⅰ.①数… Ⅱ.①宋… Ⅲ.①数据采集-高等学校-教材 Ⅳ. ①TP274

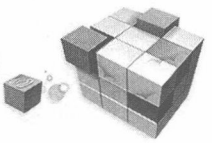
中国版本图书馆 CIP 数据核字（2018）第 261647 号

书　　名：数据挖掘
作　　者：宋万清　杨寿渊　陈剑雪　高永彬　编著

策　　划：曹莉群　　**读者热线：**（010）63550836
责任编辑：周海燕　冯彩茹
封面设计：穆　丽
责任校对：张玉华
责任印制：郭向伟

出版发行：中国铁道出版社（100054，北京市西城区右安门西街 8 号）
网　　址：http://www.51eds.com
印　　刷：三河市宏盛印务有限公司
版　　次：2019 年 1 月第 1 版　2019 年 1 月第 1 次印刷
开　　本：787 mm×1 092 mm　1/16　**印张：**11.75　**字数：**254 千
书　　号：ISBN 978-7-113-25167-3
定　　价：38.00 元

前　言



随着信息技术的普及和应用，各行各业产生了大量的数据，人们持续不断地探索处理这些数据的方法，以期最大程度地从中挖掘有用信息，面对如潮水般不断增加的数据，人们不再满足于数据的查询和统计分析，而是期望从数据中提取信息或者知识为决策服务。数据挖掘技术突破数据分析技术的种种局限，结合统计学、数据库、机器学习等技术解决从数据中发现新的信息并辅助决策这一难题，是正在飞速发展的前沿学科。近年来，随着教育部“新工科”建设的不断推进，大数据技术受到广泛的关注，数据挖掘作为大数据技术的重要实现手段，能够挖掘数据的关联规则，实现数据的分类、聚类、异常检测和时间序列分析等，解决商务管理、生产控制、市场分析、工程设计和科学探索等各行各业中的数据分析与信息挖掘问题。

截至 2018 年本书出版，共有 283 所高校获批“数据科学与大数据技术”专业，其中 985 及 211 高校占比为 13%。目前国内数据人才缺口更是达到百万级。数据科学是一门交叉学科，除了计算机相关知识，还需要统计和数学基础，以及业务应用能力。目前，数据科学与大数据逐渐成为高校信息类、管理类和数学统计类专业的必修课程，同时，作为面向各专业的通识课也广受欢迎。

本书作为立足于应用型本科数据科学与大数据教学的入门级教材，具有如下特色：

（1）内容安排合理且全面，从数据的预处理到常用数据挖掘算法的描述，循序渐进，深入浅出。

（2）难度适中，适用于本科中低年级的入门级教材，零基础要求，对编程及数学知识不作要求。

（3）融入了大量本领域的前沿知识与方法，如包括基于 GAN 网络的深度学习的最新进展。

（4）理论与案例相结合，理论与实践相结合，包含了 Weka 工具的使用。特别地在第 10 章还给出了完整的数据挖掘应用案例，使读者能够在数据挖掘平台上感受完整的数据分析过程。

本书全面介绍了数据挖掘的基础知识、基本原理、常用算法以及相应的实践工具，主要内容分为以下四块内容：

（1）数据挖掘基本知识。第 1 章为数据挖掘概述，主要介绍数据挖掘的基本概念、基本流程及算法等。第 2 章介绍数据的描述与可视化，包括数据按属性分类、数据的基本统计描述、数据的相似性度量方法及数据的可视化技术等。

（2）数据预处理。第 3 章介绍数据的采集和预处理，包括数据的采集、数据预处理的目的和任务、数据清洗、数据集成和数据变换等。第 4 章介绍数据的归约，包括线性回归和主成分分析。

（3）数据挖掘算法详解。第 5 章介绍关联规则挖掘，包括关联规则挖掘的概念、关联规则挖掘算法及应用实例。第 6 章介绍分类与预测，包括决策树模型、贝叶斯分

类模型、线性判别模型、逻辑回归模型以及模型的评估与选择方法。第7章介绍非线性预测模型，包括支持向量机和神经网络。第8章介绍聚类分析，包括聚类分析概述、k-均值聚类、k-中心聚类以及聚类评估。第9章介绍深度学习，包括深度学习的来由、深度学习网络的基本结构、卷积神经网络及一个应用实例。

（4）数据挖掘实践。第10章为使用Weka进行数据挖掘，包括Weka的基本操作、如何使用Weka进行关联规则挖掘、分类、回归和聚类等。

另外，附录还介绍了拉格朗日乘子法在支持向量机中的优化算法。

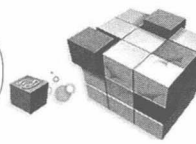
本书由宋万清、杨寿渊、陈剑雪、高永彬编著。具体分工如下：上海工程技术大学宋万清编写第2、5、6、8、10章和附录，上海工程技术大学陈剑雪编写第3、7章，上海工程技术大学高永彬编写第9章，江西财经大学杨寿渊编写第1、4章。全书由上海工程技术大学方志军、上海交通大学钱亮宏主审。同时，本书部分内容借鉴了许多学者的研究成果，在此深表谢意！

由于编者水平有限，加之时间仓促，书中难免存在疏漏和不足之处，敬请读者批评指正。

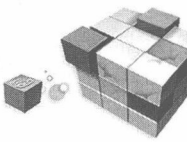
编　者

2018年8月

目　录



第1章 数据挖掘概述 «



1.1 什么是数据挖掘

先来看一个在数据挖掘界流传甚广的故事：

全球最大的零售商沃尔玛通过对海量的原始交易记录分析发现了一个有趣的现象：与尿不湿一起购买最多的商品竟然是啤酒！这两种毫不相干的商品的销售数据为何会具有如此高的相关度？这着实令人费解。为了一探究竟，沃尔玛派出市场调查人员深入调查分析，终于弄清楚了产生这种现象的原因。原来美国的太太们常叮嘱她们的丈夫不要忘了下班后为小孩买尿不湿，而丈夫们在买尿不湿时常常习惯性地捎上几罐啤酒以犒劳自己。这是一个了不起的发现，沃尔玛以此为依据尝试将这两种商品并排摆放，结果啤酒和尿不湿的销量双双增长。

以上故事是数据挖掘的典型案例，通过对大量数据进行处理分析，从中发现有价值的知识和规律。类似的案例还有许多，如淘宝、亚马逊（Amazon）等电商通过海量的顾客网购记录分析顾客的消费习惯，并以此为依据向顾客推荐商品；谷歌（Google）通过对检索词频分析成功预知了2009年冬季流感的到来；芝麻信用通过海量的网络交易数据分析对用户进行信用评估和风险测控；腾讯通过对海量游戏数据进行分析实现游戏产品的市场预测和精准营销等。

1.1.1 数据、信息和知识

数据（Data）产生于对客观事物的观察与测量，我们把被研究的客观事物称为实体（Entity）。实体可以通过各种可观测的属性（或称为特征）来描述，例如，人作为一个实体，有年龄、性别、身高、体重等属性，这些属性有时也称为变量（Variable），而这些变量的取值就是数据。在计算机科学中，数据是一个非常广泛的概念，它泛指所有能够被计算机处理的单一符号、符号组合，甚至模拟信号。

信息（Information）是数据的内涵，要得到信息需对数据进行解释或加工处理。信息与数据既有区别又有联系，数据是信息的载体，是具体的数字或符号，是具体的；信息是数据的内在含义，是抽象的。

对信息进行再加工，进一步抽象和概括，就得到了知识（Knowledge）。知识通常

表现为模式或规律，它是对信息之间的逻辑联系的抽象概括，具有简单、可重复、可推广的特点，例如表 1.1 所示的关系数据库。

表 1.1　学生成绩表

姓名	学号	综合成绩	等级
李明	001	82	B
刘艳	002	91	A
张凯	003	95	A
杨林	004	97	A
王二小	005	85	B
钱晓兰	006	87	B
刘丽	007	99	A

比如“李明的综合成绩是 82”“李明的等级是 B”“刘艳的综合成绩是 91”“刘艳的等级是 A”是信息，而“如果综合成绩>90 那么等级为 A”“如果综合成绩>80 且综合成绩<90 那么等级为 B”则是知识，它们是对多条信息的抽象概括，提取其规律。

从大量的知识中总结出原理和法则，就得到了智慧（Wisdom），它是更高层次的抽象。从数据到信息，再到知识，再到智慧，是一个不断抽象概括的加工过程，可以用图 1.1 来表示。

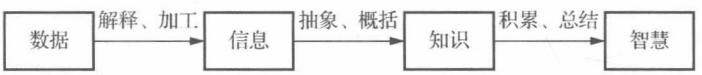


图 1.1　从数据到知识、智慧的过程

1.1.2　数据挖掘的定义

通过上一小节的学习已经知道数据、信息和知识是不同的，有数据并不等于有信息和知识，必须对数据进行解释、加工、抽象和概括才能得到信息和知识。真正有价值、可供人类利用的是信息和知识，而不是数据，因此将数据转化为信息和知识至关重要。计算机的早期时代是通过人工的方式实现由数据到知识的转化，但随着计算机和互联网技术的发展，数据规模以指数方式爆炸式增长，人工处理方式已不可行。从大量的数据中以自动或半自动的方式抽取信息、发现新的知识就是数据挖掘（Data Mining，DM）的基本任务。

关于数据挖掘，一个比较公认的定义是：

数据挖掘是利用人工智能、机器学习、统计学等方法从海量的数据中提取有用的、事先不为人所知的模式或知识的计算过程。

形象地说，数据挖掘就是挖矿，从矿山（数据库）中发掘有用的矿藏（知识）。如果从数学角度来看，数据挖掘就是一个变换，它将输入的数据变换为有用的模式或知识。

在数据库领域研究者常常使用“数据库中的知识发现（Knowledge Discovery in

Database, KDD)”，这个术语最早由 Piatetsky 和 Shapiro 提出，其含义是指从原始数据出发，经过数据清洗、集成、选择、变换、提取、评估得到有价值的信息和知识的整个过程，如图 1.2 所示。

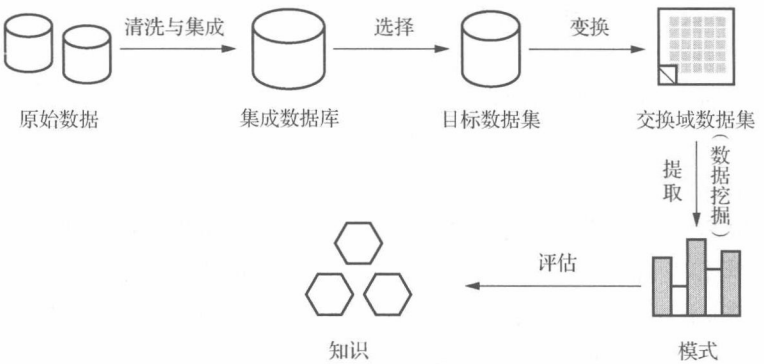


图 1.2 知识发现的过程

在这个过程中，从数据中提取事先不为人所知的、有价值的模式是最为关键的一步，数据库研究者把这一步称为数据挖掘，但现在学术界通常把数据挖掘与数据库中的知识发现视为等同的概念。

在图 1.2 中，把数据的清洗、集成、选择和变换等步骤称为数据的预处理。之所以需要这些预处理步骤是因为现实数据具有数量大、维数高、高度冗余、含噪声、有缺失等特点，如果不对其进行预处理，一方面计算量太大，另一方面由于无关数据的干扰导致挖掘的结果不可靠。数据预处理的过程需要大量使用统计学、信号与图像处理等领域的方法。

数据经过预处理之后，便可对其进行挖掘，从中提取事先不为人所知的、有价值的知识，这些知识包括关联规则、分类模型、预测模型、聚类模型、时间序列模型和异类检测模型，将在 1.2.2 节对这些概念做出解释。为了从数据中挖掘出关联规则、分类、预测、聚类和异类知识，需要大量使用概率统计、模糊数学、模式识别、人工智能、机器学习、专家系统甚至神经科学等领域的方法。此外，为构建一个高效的数据挖掘系统，还需要知识表示、高性能计算等领域的知识，因此数据挖掘是一门综合性的交叉学科。

1.1.3 数据挖掘的发展简史

在计算机诞生之初，数据是零散存储的，随着数据量的增大和复杂性的提高，数据的管理和有效利用问题提上了日程。20 世纪 60 年代产生了数据库和数据库管理系统，其中著名的有美国通用电气公司（General Electric Company, GE）开发的 IDS（Integrated Data Store）和 IMS（Information Management system）。1970 年国际商用机器公司（International Business Machines Corporation, IBM）的 E.F.Codd 博士提出了关

系数据模型，使得关系数据库逐渐成为数据库系统的主流。传统数据库仅解决了数据的查询、操纵、定义和控制等底层功能，但企业需要的不是数据，而是能够辅助决策的高层信息和知识，这就需要将多源异构的数据进行集成、统一处理并实时地生成报表，以供决策者参考。为了解决这个问题，Bill Inmon 于 1990 年提出了数据仓库（Data Warehouse）和联机分析处理（On Line Analytical Processing, OLAP）等思想，使得数据库管理系统更加智能化，能够自动生成报表，实时提供一些辅助决策的知识。

尽管数据仓库和联机分析处理具有较高的自动化水平，但它们仍然只能做事先设定的处理和操作，而无法自动发现新知识。20 世纪 90 年代中期，随着互联网技术的发展，信息量急剧增长，且形式多样化，更新换代速度加快，为了提高数据转化为知识的效率，研究者开始将人工智能和机器学习的方法引入数据库管理系统，使系统具有从数据中探索和发现新知识的能力，于是数据挖掘作为一门新学科正式形成。

近十几年来，随着多媒体信息技术、移动网络技术、物联网和云计算技术的发展，不仅数据量爆炸式增长，种类越来越多样化，而且对数据处理的时效性要求也越来越高，大数据的概念由此形成。大数据给数据挖掘带来了新的发展机遇，使得数据挖掘的应用越来越广泛和普及，同时也给数据挖掘带来了新的挑战，这主要表现在如下 4 个方面：①数据量大大超过了传统计算机硬件和软件能够处理的范围；②数据的质量低，存在大量缺失和错误；③数据高度冗余，价值密度低；④实时性要求高。为了应对这些挑战，高性能计算、云计算以及近年发展起来的深度学习成为大数据挖掘必不可少的工具。

1.2 数据挖掘的基本流程及方法概述

1.2.1 数据挖掘的基本流程

在 1.1.1 节扼要介绍了知识发现的基本过程，将这一过程稍作完善，即得到数据挖掘的基本流程，如图 1.3 所示。

数据挖掘大致分为数据预处理、数据挖掘、模式评估 3 个阶段。其中预处理大致包括清洗、集成、选择、变换等步骤。由于原始数据中含有噪声、错误、缺失等，因此预处理的第一步是对数据进行清洗，消除数据中的噪声和无关数据，修复错误、填补缺失数据等。接下来是对来自不同数据源中的数据进行集成，将有关的数据组合在一起构建数据仓库。数据仓库中的数据并非都与挖掘主题有关，必须从中选出与挖掘主题密切相关的数据，这样一方面可以减少计算量，另一方面还可以消除无关数据的干扰。选择完数据之后，还需要对目标数据进行变换，如线性回归分析（Linear Regression Analysis）、主成分分析（Principal Component Analysis）、多维标度分析（Multidimensional Scaling）、傅里叶变换（Fourier Transform）、离散余弦变换（Discrete Cosine Transform）、小波变换（Wavelet Transform）等，变换的主要目的是消除冗余，简化数据，这一步骤也称数据归约（Data Reduction）。

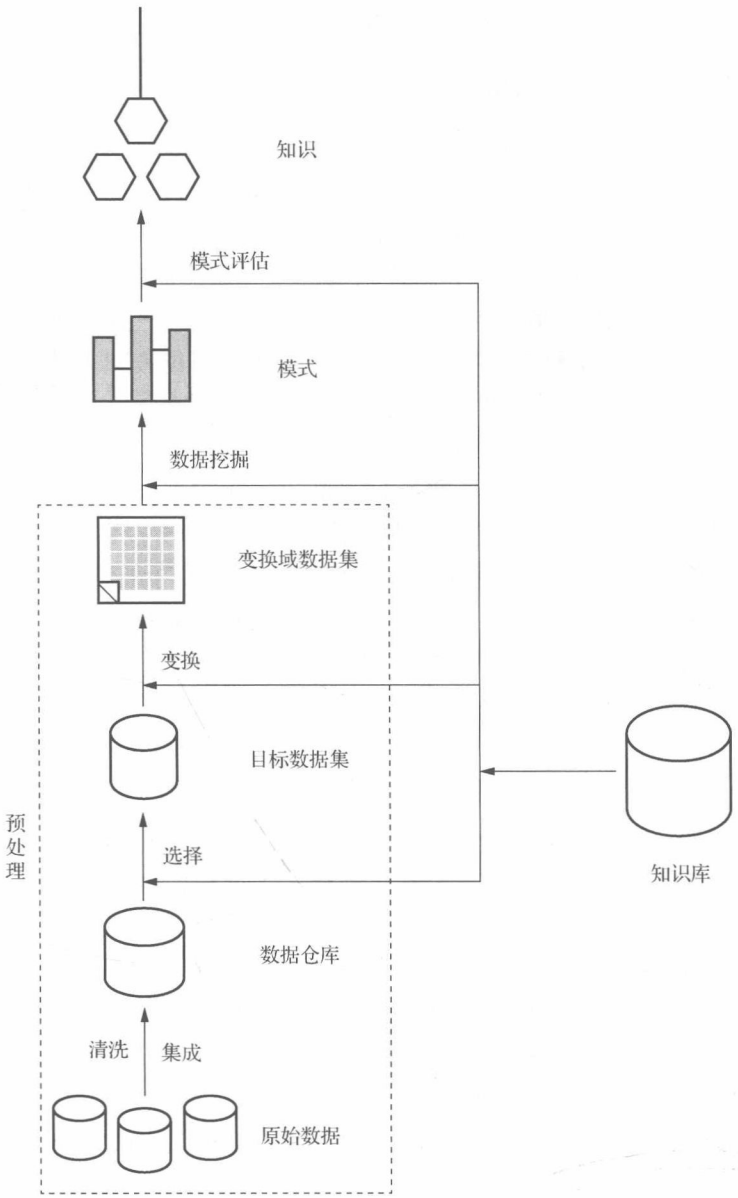


图 1.3 数据挖掘流程图

数据挖掘阶段的主要任务是从变换后的数据集中挖掘出事先不为人所知的模式或知识，这一阶段将在 1.2.2 节中介绍。挖掘出新的模式和知识后，还需要对其进行评估，按照一定的标准，如挖到的知识的新奇性、有效程度和应用价值等，从中筛选出新奇的、有效的、有价值的模式和知识。此外还需要用适当的方式将这些知识表示或展现出来，因此，知识表示和可视化技术也是数据挖掘的重要研究内容。

在数据选择、数据变换（归约）、数据挖掘、模式评估等步骤都需要专业领域知

识的参与，如对业务的理解、对数据的理解、对模式的评估标准等，因此需要有一个专业领域知识库来指导和支持数据挖掘的整个过程。

1.2.2 数据挖掘的任务和方法概述

数据挖掘所要挖掘的模式和知识包括以下内容：

（1）对数据集的概要总结。通过对数据集中的数据进行统计分析得出数据集的总体特征，对数据集进行简明、准确的描述，或对两个数据集进行对比，给出两个数据集的差异的概要性描述。例如，从某校教职工数据库中选择讲师数据进行挖掘分析，可得到讲师的概要性描述："65%（age<30）and（age>24）"，这就表示该校的讲师中有65%的人年龄介于24岁和30岁之间；又比如，抽取该校教职工数据库中的讲师数据和副教授数据进行对比分析，可以得到如下概要性描述：

"讲师：70%（papers<3）and（teaching course<2）"

"副教授：65%（papers>=3）and（teaching course>=2）"

这就表示该校讲师中有70%的人发表的论文数量小于3且所授课程小于2门；而该校的副教授中有65%的人发表了至少3篇论文且讲授至少2门课程。

（2）数据的关联规则。关联规则是描述数据之间潜在联系的一种方式，通常用形如 $A-B$ 的蕴含式来表示。例如，从某零售店的原始销售记录中挖掘出如下关联规则：

$$\text{contains}(X,\text{'bread'}) \rightarrow \text{contains}(X,\text{'milk'})[\text{support}=10\%,\text{confidence}=60\%]$$

这就表示所有顾客中有10%的人同时购买了面包和牛奶两样商品，而在购买了面包的顾客中有60%的顾客同时购买了牛奶。其中前一个百分比称为支持度，其大小反映了关联规则的普遍程度，支持度越大表示该关联规则覆盖的范围越大；后一个百分比称为置信度，是一个条件概率，其值越大则表示购买了面包的顾客同时购买牛奶的概率越大。

又如，某房地产销售公司从历史销售记录中挖掘出如下关联规则：

(年龄>30)∧(年龄<50)∧(年收入>20万元)→（是否成交='yes'）

[support=20%,confidence=85%],

这就表示该公司的客户中年龄介于30岁与50岁之间、年收入大于20万元的客户占20%，而在年龄介于30岁与50岁之间、年收入大于20万元的客户中有85%的最终成交了。

关联规则挖掘是数据挖掘的重要内容，将在第5章详细介绍关联规则挖掘的算法。

（3）分类与预测。所谓分类，就是按照一定的规则将样本数据划分成不同的类，分类的关键在于选择合适的分类规则，这些规则通常是从样本数据中学习而获得。所谓预测，就是利用某个函数模型来估计样本的某些属性的值，所利用的函数模型可以是线性的也可以是非线性的，可以是参数模型也可以是非参数模型，这些模型和参数通常需要通过从训练数据中学习得到。分类和预测是紧密相关的，分类可以看作预测的特殊情形，即因变量只能取有限的离散值的情形。

例如，商业银行可以根据信用卡申请人的年龄、职业、收入水平、财产状况等对

信用卡申请人进行分类，将信用卡申请人分为低、中、高风险三类，分类方法可以是决策树模型、支持向量机或神经网络模型等，这些模型将在第 6 章和第 7 章详细介绍。

（4）聚类。所谓聚类，就是依据数据内在的相似性将其划分为若干类，使得同类数据之间的相似度尽可能大，并且不同类数据之间的相似度尽可能小。聚类与分类不同，其区别在于事先并不知道样本数据有哪些类，是探索性的。聚类的关键在于选择合适的相似性度量。

例如，手机销售公司可依据消费者的年龄、性别、职业、收入水平、居住地等属性对消费者进行聚类分析，探索各类消费者的特点，以促进营销。将在第 8 章详细介绍聚类算法。

（5）异类检测。异类（Outlier）也称异常点，是指那些不符合大多数数据对象所构成的规律（模型）的数据对象，如分类模型中的反常实例、聚类模型中的离群点等。传统的数据挖掘算法为了提高模型的拟合优度常常将异类当作噪声除去，但在某些应用中异类往往是重要的，如诈骗识别、异常行为检测、网络异常检测等，对于这些应用异类检测尤其重要。研究者提出了许多异类检测算法，如基于数据对象的概率分布的算法、机器学习算法等。

（6）时间序列模型。像股票价格这样的数据是随时间不断演化的，人们关心的是其演化规律，即数据在时间维度上的相关性，如趋势、周期性、自回归模式等，这就是时间序列模式，这些模式可以用各种各样的时间序列模型来描述。

在挖掘关联规则、分类、预测、聚类、异类和时间序列模式等知识时，人们需要用到各种机器学习算法，如贝叶斯网络、支持向量机、人工神经网络、深度学习等。所谓机器学习，就是用计算机程序模拟人类的学习过程，是一个从训练数据中获取经验并不断改进系统自身性能的有反馈的信息处理与控制过程。例如，神经网络就是一个典型的学习系统，它由多个神经元连接而成，每一个神经元的功能实际上是一个简单的非线性函数，其结构可以用图 1.4 表示。

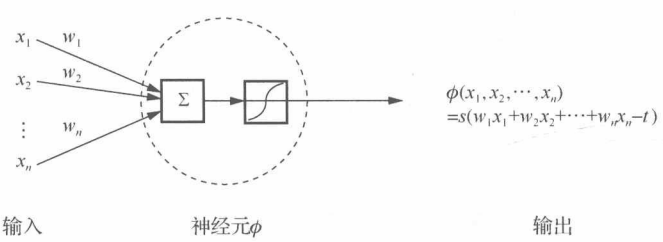


图 1.4　单个神经元结构示意图

图中 $w_1,w_2,\cdots,w_n$ 称为连接权值，$s(x)$ 称为激活函数，t 是激活阈值。激活函数 $s(x)$ 通常取如下 Logistic 函数：

$$s(x)=\frac{1}{1+e^{-kx}} \quad x\in\mathbf{R}$$

其图像是S形。通过适当地设置连接权值和激活阈值，单个神经元具有一定的分类和预测能力，但毕竟模型过于简单，无法胜任复杂数据的分类和预测。如果将多个神经元按照适当的方式连接起来，就得到了一个人工神经网络（简称神经网络）。图 1.5 所示的三层神经网络就是一个典型的分类器。

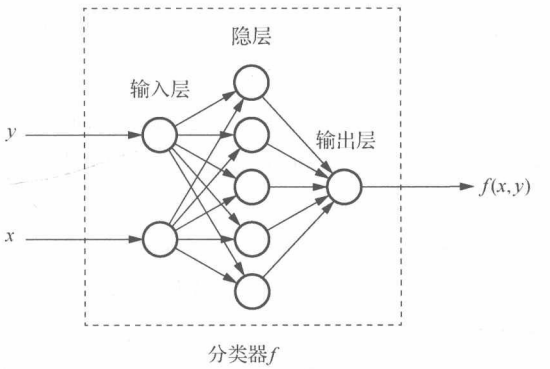


图 1.5　三层神经网络构成的分类器

这个分类器是由多个神经元线性组合和嵌套构成的复杂的非线性函数 $f(x,y)$，当然它依赖于连接权值和激活阈值的设定，用 $\boldsymbol{w}$ 表示所有连接权值和激活阈值所构成的向量，则分类器的完整表达式为 $f(x,y,\boldsymbol{w})$。如何训练这个分类器使它“学会”对样本数据进行分类呢？不妨假设待分类的样本数据是坐标平面上的一些点：

$$E=\{(x_1,y_1),(x_2,y_2),(x_3,y_3),\cdots\}$$

这些样本数据分属于两个不同的类，即Ⅰ和Ⅱ。要让分类器学习就必须提供“学习材料”，即训练样本，这里准备的训练样本是从所有样本数据中随机地抽取一部分，并用人工方式标注所抽取的每一个样本数据所属的类别，样本数据（x_i,y_i）所属的类别用可用一个标签 μ 表示，即

$$\mu_i=\begin{cases}1,\text{如果}x_i,y_i\text{属于第Ⅰ类}\\-1,\text{如果}x_i,y_i\text{属于第Ⅱ类}\end{cases}$$

这些训练样本连同其标签就构成了训练数据集 D：

$$D=\{(x_1,y_1,\mu_1),(x_2,y_2,\mu_2),(x_3,y_3,\mu_3),\cdots,(x_n,y_n,\mu_n)\}$$

分类器学习的过程就是不断地调节向量 $\boldsymbol{w}$，使得式（1.1）平方误差函数

$$E=\sum_{i=1}^{n}[f(x_i,y_i,\boldsymbol{w})-\mu_i]^2 \qquad (1.1)$$

最小化的过程，具体的计算原理将在第 7 章学习。分类器经过学习后可以达到非常好的分类效果，如图 1.6 所示，其中“+”表示Ⅰ类样本，“o”表示Ⅱ类样本。

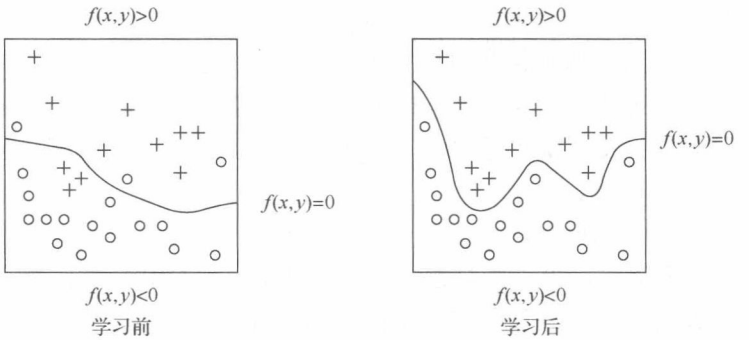


图 1.6 分类器学习的效果

机器学习算法大致可以分为监督学习（Supervised Learning）和非监督学习（Unsupervised Learning）两类。所谓监督学习，就是有教师指导的学习，训练数据必须是已经标注的样本数据，或者说训练目标由人指定，如回归、分类等；所谓非监督学习，就是无教师指导的学习，训练数据是无标签数据，学习算法只能从数据本身提取模式和规律，如聚类、自编码学习等。关于机器学习的更多知识将在第 6 章、第 7 章和第 9 章介绍。

1.3 数据挖掘的应用

本章开篇举了啤酒和尿不湿的例子，这是数据挖掘的一个典型应用。事实上，数据挖掘技术从一开始就是面向应用的，应用领域非常广泛，包括商务、银行、保险、医疗、电信、科研、教育、电子出版、娱乐、社交媒体、智能电网等。下面仅就商务领域、医疗和医学领域、银行和保险领域、社交媒体领域举几个典型的应用实例。

1.3.1 数据挖掘在商务领域的应用

数据挖掘在商务领域的应用包括：库存及物流管理、数据库营销、客户群体划分、背景分析、交叉销售、客户流失性分析等。美国运通公司（American Express）有一个用于记录信用卡业务的数据库，数据量达到 54 亿字符，并仍在随着业务进展不断更新。运通公司对这些数据进行挖掘，在此基础上制定了“关联结算（Relation Ship Billing）优惠”的促销策略，即如果一个顾客在一个商店用运通卡购买一套时装，那么在同一个商店再买一双鞋，就可以得到比较大的折扣。这种策略取得了极大的成功，实现了商店销售量和运通卡使用率的双双增长。

农夫山泉通过定期采集饮用水的生产、运输、销售、财务等环节的场景数据，每月收到约 3 TB 的数据，其中不乏图像、视频、音频等非结构化数据，通过对这些数