21世纪统计学系列教材

R语言数据挖掘

（第2版）

薛 薇 编著

Data Mining with R

(Second Edition)

中国人民大学出版社

计学系列教材

R语言数据挖掘

（第2版）

薛　薇　编著

Data Mining with R

(Second Edition)

中国人民大学出版社
· 北京 ·

图书在版编目（CIP）数据

R语言数据挖掘/薛薇编著. —2版. —北京：中国人民大学出版社，2018.7
21世纪统计学系列教材
ISBN 978-7-300-25825-6

Ⅰ.①R… Ⅱ.①薛… Ⅲ.①程序语言-程序设计-教材②数据采集-教材 Ⅳ.①TP312②TP274

中国版本图书馆CIP数据核字（2018）第111710号

21世纪统计学系列教材
R语言数据挖掘（第2版）
薛 薇 编著
R Yuyan Shuju Wajue

出版发行	中国人民大学出版社		
社　　址	北京中关村大街31号	**邮政编码**	100080
电　　话	010－62511242（总编室）	010－62511770（质管部）	
	010－82501766（邮购部）	010－62514148（门市部）	
	010－62515195（发行公司）	010－62515275（盗版举报）	
网　　址	http://www.crup.com.cn		
	http://www.ttrnet.com（人大教研网）		
经　　销	新华书店		
印　　刷	北京密兴印刷有限公司	**版　　次**	2016年4月第1版
规　　格	185 mm×260 mm　16开本		2018年7月第2版
印　　张	26.25插页1	**印　　次**	2018年7月第1次印刷
字　　数	664 000	**定　　价**	55.00元

第 2 版前言

《R 语言数据挖掘》（第 2 版）仍然坚持第 1 版原理讲解深入浅出，案例代码可反复实现，兼顾数据挖掘原理和实践，手把手教会读者数据挖掘的风格。同时，在第 1 版的基础上进行了以下修订：

第一，在保持内容框架不变的基础上，对各章节文字进行了全面梳理和规范，使得全书通篇文字表述更加一致、完整和严谨。

第二，为使读者更易理解原理，对第 4 章加权 K-近邻法的距离和相似度变换过程进行了更加详尽的讲解，调整了旁置法的表述；调整了第 5 章分类树生长过程异质性下降测度公式的书写形式，更加详尽地论述了对交叉验证的剪枝过程，增加了对随机森林输出结果的说明；统一了神经网络中权值、误差的数学书写；对 EM 聚类过程进行了更加详尽的讲解。

第三，增加了 RStudio 简介。

本书可作为高等院校数据科学和大数据技术本科专业，以及其他相关专业本科生和研究生的学习教材，同时也可作为商业企业、科研机构、政府管理部门等相关数据分析人员的阅读参考书。请读者到中国人民大学经管图书在线（http://www.rdjg.com.cn）网站，下载本书案例数据和 R 程序代码。

特别感谢中国人民大学出版社对本书出版的大力支持，感谢王珏、刘茜、王艳红、周天旺、要卓、陈笑语等同学对本书的贡献。书中不妥和错误之处，望读者不吝指正。

薛 薇

于中国人民大学应用统计科学研究中心、中国人民大学统计学院

第 1 版前言

我们已经步入一个大数据时代。大数据时代不仅仅意味着数据的积累与存储，更意味着对数据的建模与分析。

近年来，数据挖掘不断汲取并集成机器学习、统计学和可视化等学科领域的研究成果，在众多行业获得了可观的应用案例，造就了卓有成效的发展。这一切使得大数据分析不再是一种漂浮在云端、飞翔在风口的奢望，大数据分析已日益成为许多个人、企业和组织进行科学决策的重要方法工具。

由于采取彻底的开放性策略，R 语言已成为近年来出类拔萃的数据挖掘工具之一。其特点主要是：开源性，即可以免费下载并升级；全面性，即数据挖掘方法丰富，覆盖面广；操作简便性，即直接采用函数调用相关算法，通过简单编程即可完成复杂的数据处理和方法拓展；可扩展性，即 R 语言通过网络社区平台吸引越来越多的专家学者和应用人员成为开发者，为 R 语言不断增添更有效、更前沿的数据挖掘方法。所以，R 语言是一款应用前景广阔的数据挖掘工具。

本书以数据挖掘概念和 R 语言入门开篇，目的是使读者能够快速总览数据挖掘的理论轮廓，厘清相关概念，掌握 R 语言入门和深入学习的路线。后续，本书以数据挖掘过程为线索，以应用实例为辅助，详细讨论 R 语言数据挖掘的数据组织和整理、可视化图形、主流数据挖掘方法原理和算法步骤以及应用实现等内容。其间，为使读者快速入门 R 语言，起步数据挖掘的实践应用，本书首先系统介绍了 R 语言的数据对象、常用系统函数、流程控制等服务于数据组织和整理的程序设计基础知识，以及 R 的各种主流可视化图形。然后，围绕数据预测、揭示数据内在结构、揭示数据关联性、诊断异常数据等数据挖掘核心目标，依次讨论了诸多主流数据挖掘方法和 R 的实现过程，涉及近邻分析、决策树、人工神经网络、支持向量机、聚类算法、关联规则、模式甄别、网络分析等众多经典模型和算法。覆盖内容之广泛，R 实现步骤之详尽，数据应用之经典，都是国内外同类书籍中不多见的。这是本书的特点之一。

同时，R 语言数据挖掘中的数据挖掘方法是核心，R 语言实现是形式，两者是“道”与“术”的关系。我们认为“道”和“术”的结合，无论对数据挖掘的初学者还是应用实践者都是必要的。“道”是原理，此原理不是数学公式的简单罗列，而是给出直观透彻的方法认知。“术”是操作，此操作不是函数命令的简单呈现，而是算法实现和应用的通用模板，是帮助读者实现数据挖掘实践的有效工具。本书力图阐述“道”，利用 R 语言充分展现“道”，通过有代表性的数据案例，画龙点睛地阐明“术”。每章都配有案例数据和 R

程序代码，使读者不但知其然，更知其所以然。这是本书的特点之二。

进一步，目前R语言包的数量已多达7 000多个，而且还在快速增长。R的开放性决定了可能有诸多包都可以实现相同的数据挖掘算法。对此，本书选择R中主流且被有效验证和广泛使用的包，既保证经典性，也兼顾有效性，同时解决了初学者因陷于众多R的“包”围中而无从下手的问题。这是本书的特点之三。

最后，对R语言数据挖掘的初学者，建议按照本书章节结构，循序渐进地学习，并参照书中示例，边学边做，以加深概念理解和提升R语言熟练度。对有一定R语言基础或数据挖掘应用经验的学习者，因本书各章节具有相对独立性，所以采用“以数据为导向”和“以问题为导向”的有针对性的R语言数据挖掘学习策略均是可行的。

本书努力迎合广大R语言数据挖掘读者的主流需求，适合高等院校相关专业的本科生和研究生学习使用，以及商业企业、科研机构、政府管理部门等相关人员阅读参考。请读者到中国人民大学经管图书在线（http://www.rdjg.com.cn）下载本书案例数据和R程序代码。

特别感谢中国人民大学出版社对本书出版的大力支持，感谢王珏、刘茜、王艳红、周天旺、要卓、陈笑语等同学对本书的贡献。书中不妥和错误之处，望读者不吝指正。

薛 薇

目　录

第1章 Chapter 1 数据挖掘与R语言概述

蓬勃发展的互联网（移动互联网）技术、物联网技术和云计算技术，不但将人类社会与物理世界有效地连接起来，更创造性地建立了一个数字化的网络体系。运行其中的搜索引擎服务、大型电子商务、互联网金融、社交网络平台等，不断改变着人们生活与生产的方式。同时，参与其中的个人、企业和组织每时每刻都在释放巨大的比特数字流，从而造就了一个崭新的大数据时代。

人类的数据生产能力达到空前。2009年IBM的一项早期研究结果显示，人类文明诞生以来其数据总量的90%是在之前两年内产生的。2020年全世界所产生的数据规模预计将达到2016年的45倍。其规模已远远超出了传统的G或T的量级，而达到以P（1 000T），E（100万T）或Z（10亿T）为单位的水准。

通常人们总结大数据有“4V”的特点，即大量（volume）、高速（velocity）、多样（variety）、价值（value）。那么如何采用有效的方法快速分析这些大量和多样化的数据，并挖掘出其内在的价值呢？我们说大数据分析一般需要四个核心要素：基于云计算的基础设施、分布式的大数据体系、数据分析方法与算法、行业应用知识与经验。

沿着这个思路，大数据分析的一名初学者应如何寻找合适的突破口，并通过渐进的学习，成为理想中的数据分析师或数据科学家呢？我们认为从数据挖掘方法入手，无疑是最佳选择。这个学习方案一方面可保证初学者在一开始就可以持续进行一般的数据分析，并通过增加数据量、引进新方法提高自己的分析能力，从而逐渐成为一名方法应用与算法研究的专家。另一方面，在达到一定水平之后，向下可以进一步研究大数据的分布式计算环境与计算方法，并深入学习云计算的基础知识，成为大数据系统建设的高手；向上也可以结合自己所从事行业的实际问题，通过具体实践积累应用经验，成为该领域大数据分析的翘楚。

R语言正是目前应用最为广泛的数据挖掘与分析工具。其突出特点表现为：第一，共享性。使用者可以到相应的网站免费下载和使用。第二，分析方法丰富。R不仅包括众多经典通用的统计和数据挖掘方法，还拥有大量面向不同应用领域问题的前沿和专用的模型算法。第三，操作的简便性和灵活性。R支持计算机编程。用户可以通过编程实现数据整理的自动化和批量化，可以通过调用R的现成模型和算法解决一般性的数据挖掘问题，还可以自行编写程序解决特殊的数据挖掘问题。第四，成长性。R语言通过开放的网络社区

平台，不断吸引更多的专家学者和应用人员成为R的开发者，更多更有效、更前沿的方法正不断融入R中。

1.1 什么是数据挖掘

大数据对于数据挖掘，是挑战，更是机遇。褪去了发展初期的浮躁与喧哗，数据挖掘在理论方法与软件工具上都有了长足的发展，并在诸多领域积累了成熟的应用案例，取得了扎实的应用成果。人们曾经将数据挖掘形象地比喻为从数据“矿石”中开采知识“黄金”的过程，如今面对数据的“矿石”，数据挖掘充分汲取机器学习、统计学、分布式和云计算等技术养分，在方法研究、算法效率、软件工具集成环境和创新应用等方面不断开拓，正将昔日的数据“矿锤”升级为现代化的数据“挖掘机”，成为大数据时代最有效的数据分析利器。所以，数据挖掘具有多学科综合性、方法性与工具性的特征。对此，初学者应具有较强的数据操作能力和学习领会能力，以举一反三、触类旁通。

数据挖掘的发展过程是一个兼容并蓄的成长过程。如图1-1所示，一般来说，数据挖掘经历了三个主要发展阶段，从初期局限于数据库中的知识发现（knowledge discovery in database，KDD），发展到中期内涵不断丰富完善以及多学科的融合发展，乃至今天成为大数据时代的关键分析技术，数据挖掘已经取得了实质性的跨越。

图1-1 数据挖掘发展历程示意图

目前，对数据挖掘的理解已达成如下共识：

首先，数据挖掘是一个利用各种方法从海量的有噪声的各类数据中提取潜在的、可理解的、有价值的信息的过程。这里，信息可进一步划分为两大类：一类是用于数据预测的信息；一类是用于揭示数据内在结构的信息。

其次，数据挖掘是一项涉及多任务、多学科的庞大的系统工程，涉及数据源的建立和管理、从数据源提取数据、数据预处理、数据可视化、建立模型并评价以及应用模型评估等诸多环节，如图1-2所示。

针对复杂问题且涉及海量数据的数据挖掘任务往往是一项大规模的系统工程。为更加规范地开展数据挖掘工作，NCR，SPSS和戴姆勒-奔驰（Daimler-Benz）三家公司联合制定了跨行业数据挖掘标准CRISP-DM（cross industry standard process of data mining），SAS公司也发布了相关数据挖掘标准SEMMA（sample，explore，modify，model，assess）。这些标准希望通过对数据挖掘过程中各处理步骤的目标、内容、方法以及应注意

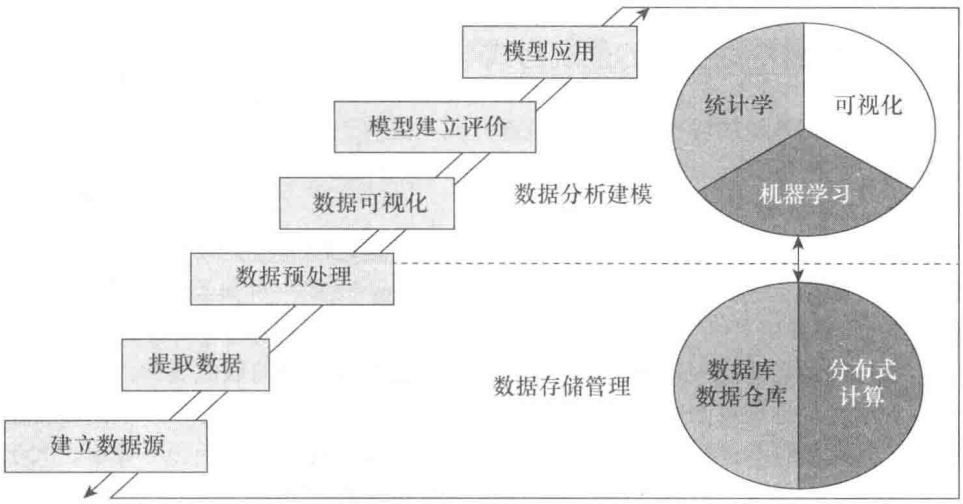


图 1-2　数据挖掘过程示意图

的问题等提出可操作性的建议，帮助学习者从方法论的高度深入理解并掌握数据挖掘的一般规律。

进一步，数据挖掘的诸多环节本质上可归纳为两个具有内在联系的阶段：数据存储管理阶段和数据分析建模阶段，涉及计算机科学和统计学等众多交叉学科领域。

当前，数据挖掘的对象是大数据系统。大数据往往来自不同的采集渠道以及不同的数据源，数据量庞大且杂乱有噪声。高效合理地存储数据，有效地保障数据的一致性等，在数据挖掘中尤为重要，也始终是数据挖掘的难点，涉及计算机学科中的数据库和数据仓库计算、分布式计算、并行处理等多个研究领域。大数据的存储管理有两个层面：一是基础设施层面，包括对存储设备、操作系统、数据库、数据仓库、分布式计算等方面的整体评估，需求的客观理解，系统架构、技术和产品的选择，稳定、高效的数据基础设施体系的建立等一系列问题；另一个是数据管理工具层面，包括数据的抽取检索、集成清洗以及其他预处理的软件、技术和管理等诸多方面。数据的存储管理是数据分析的基础和保障，也在某种程度上为选择数据分析方法提供依据。

数据挖掘中的数据预处理、数据可视化、建立和评价模型等环节，其核心目标是发现数据中隐藏的规律性，这是统计学和从属计算机科学的机器学习（machine learning）以及具有跨学科（统计和计算机）特点的可视化研究的主要任务，也是本书讨论的重点。事实上，从统计学视角看数据挖掘会发现，数据挖掘与统计学有高度一致的目标：数据分析。正因如此，数据挖掘对统计学而言似乎并不陌生。然而，尽管目标一致，但仍提出数据挖掘概念的重要原因是：数据分析对象是大数据。大数据特征决定了数据处理需要多学科的共同参与，数据分析需要一种集中体现多学科方法和算法优势的理论和工具，这就是数据挖掘。

1.2　数据挖掘的结果

如果将数据挖掘视为一个系统，那么这个系统的输入是数据，系统的输出就是数据挖掘结果。从数据挖掘系统的输出入手，讨论数据挖掘结果的呈现方式和基本特征，是一种快速总览数据挖掘特点的有效途径，也是打开深入理解数据挖掘内涵之门的钥匙。

1.2.1 数据挖掘结果的呈现方式

数据挖掘系统的输出，其一般呈现方式主要有三类：第一，数学模型；第二，推理规则；第三，图形。

1. 数学模型

数学模型即通过数学函数的形式定量反映变量之间的相关数量关系。如最常见的一般线性回归模型 $y=\beta_0+\beta_1 x_1+\beta_2 x_2+\beta_3 x_3+\cdots+\beta_k x_k+\varepsilon$ 是一种典型的数学模型。

2. 推理规则

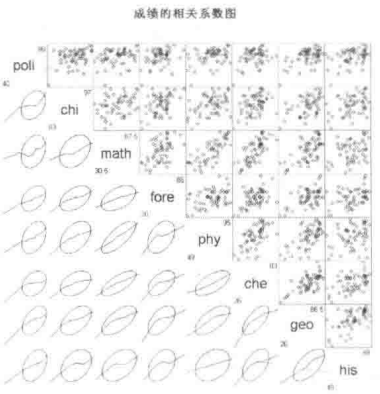
推理规则即通过一种逻辑表达式的形式反映变量之间的取值规律。规则集是多个推理规则的集合。推理规则由条件（IF）和结论（THEN）两部分组成。条件是变量、变量值以及关系运算符和逻辑运算符组成的式子。关系运算符包括等于、不等于、大于、大于等于、小于、小于等于，逻辑运算符包括“并且”、“或者”和“非”。结论是目标变量取值。

例如，IF（收入＝3）并且（年龄小于44）THEN 购买行为＝购买，就是一个常见的推理规则。

推理规则是基于逻辑表述的，直观且容易理解。

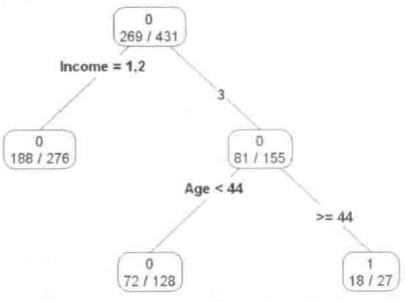
3. 图形

图形也是一种直观呈现数据挖掘结果的主要方式。它既可用于直观展示变量间相关性的特征（见图1-3（a））、数据分布的特征（见图1-3（b）），也可以是上述推理规则的图形表达（见图1-3（c）），抑或是无法以数学模型形式表达的其他复杂分析模型（见图1-3（d））。

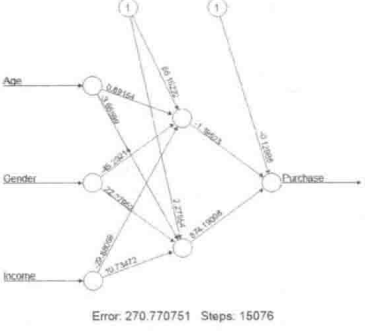


(a)相关系数矩阵示意图

(b)文本词云图示意图



(c)决策树结果示意图



(d) 神经网络模型结果示意图

图1-3 示意图

图 1-3（a）展现了多个变量简单的相关性方向和强弱；图 1-3（b）展示了一批文本中各词汇出现的频率；图 1-3（c）为决策树分析结果，是一个推理规则集的图形表示；图 1-3（d）为一个神经网络模型结果。

1.2.2　数据挖掘结果的基本特征

数据挖掘是一个从大数据中挖掘出有用信息的过程。如上所述，数据挖掘结果具有不同的呈现方式。此外，数据挖掘结果（即有用信息）还具有以下三个重要特征：潜在性、可理解性和有价值性。具体如图 1-4 所示。

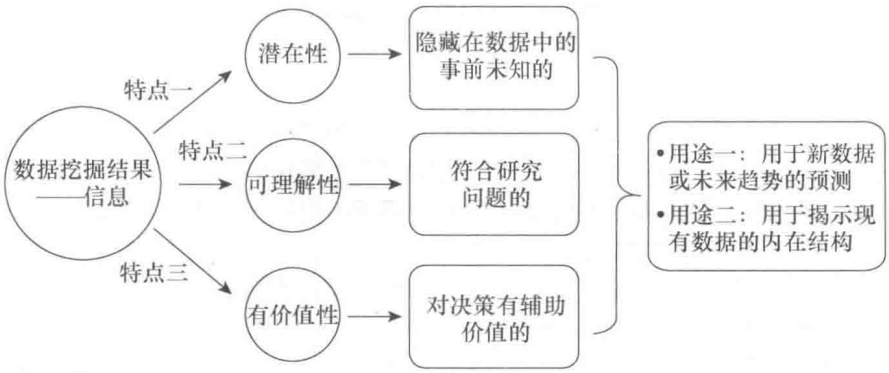


图 1-4　数据挖掘结果示意图

1. 潜在性

发现大量数据中隐含的变量相关性、数据内在结构特征等是数据挖掘的重要任务，也是数据挖掘的核心成果。研究变量相关性以及数据内在结构特征是统计学擅长的，其传统分析思路是：基于对研究问题的充分理解，依据经验或历史数据首先预设数据中存在某种相关性假定，然后验证这种假定是否显著存在于当前数据当中。这是一种典型的验证式分析思路。然而，大数据分析中的数据量庞大，变量个数多且类型复杂，以传统方式预设假定将非常困难，甚至不可能。所以数据挖掘通常会倾向于采用一种归纳式的分析思路，即事先不对数据中是否存在某种关系做任何假定，而是通过“机械式”的反复搜索和优化计算归纳出所有存在于数据中的规律。

这样的分析思路有优势，但也存在问题。优势在于它既可能找到隐藏于数据中的人们事先知道的规律性，也可能发现那些人们事先未知的规律。存在的问题是由此得到的分析结果，一方面可能是类似于传说中“尿布和啤酒”① 的典型案例，另一方面也可能是令人无法理解和没有价值的。

2. 可理解性

数据挖掘结果的可理解性是指分析结论具有符合研究问题的可解释性。例如，在消费者行为偏好的数据挖掘中，若分析结果是一段时间内顾客的消费金额与其身高有密切关

① 位于美国阿肯色州的著名连锁超市沃尔玛（Wal-Mart）通过分析顾客消费数据库发现，啤酒和尿布同时购买的可能性很大。这个结论让超市的管理者很惊讶。究其原因发现：住在该超市周边的顾客大多为年轻夫妇，通常妻子总是嘱托丈夫在下班时给孩子买尿布，而年轻的爸爸们在给孩子买尿布的同时，也会买些啤酒犒劳自己。超市根据这个分析结论重新调整了货品的摆放位置，以减少爸爸们在超市里来回拿取商品所花费的时间。

系，那么这样的结论一般就不具有可解释性。事实上，数据挖掘揭示出的不可理解的相关性，可能是一种虚假相关，也可能是因其他相关因素传递而产生的表象。

3. 有价值性

数据挖掘结果是否有价值体现在对决策是否有指导意义。对决策没有指导意义的结果是没有价值的。例如，在居民健康管理的数据挖掘中，若分析结论是90%的居民每日就餐次数是3次，且三餐的平均就餐时间是早上7点、中午12点、晚上7点。这种分析结论的价值很低，因为这是一般常识。

谁是导致数据挖掘结果有可能无法理解和没有价值的“元凶”呢？答案是：海量大数据。事实上，发现海量大数据中隐藏的可理解的、有价值的信息，难度要远大于小数据集，因为会出现分析小数据集时不曾出现的诸多新问题。其中的一个主要问题就是“机械式的挖掘”给出的“信息”很可能只是数据的某种“表象”而非“本质”。用统计术语讲就是，很可能并不是数据真实分布或关系的反映，而是海量数据自身的某种无意义的随机性的代表。

为此，人们试图借助统计学方法对“表象”和“本质”加以区分。作为数据挖掘成员中的一分子，统计学确实在区分挖掘出的信息是系统性的本质还是随机性的表象上具有重要作用。通常的做法是以分析数据是随机样本为前提，采用统计推断式的假设检验。统计推断以随机样本为研究对象，通过找到样本的某些特征并计算这些特征在原假设成立下出现的概率，判断它们是否具有统计上的显著性，即这些特征是系统性的还是由样本的随机性所致。事实上，数据挖掘发展初期也确实采纳了这种方式，所以某些数据挖掘方法貌似统计方法也很正常。但问题在于随着大数据的出现以及数据挖掘应用的不断拓展，这样的思路出现了以下主要问题。

第一，大数据的海量特性极大地限制了上述分析思路的可行性。若认为数据挖掘的数据对象是一个样本，那么这个样本通常是大样本。对于以小规模数据集为研究对象发展起来的统计推断而言，样本表现出的某些特征如果确实是由随机性导致的“表象”，那么在统计推断过程中能够得到原假设成立下出现的概率很小而被正确地确认为随机性的结论。这种分析思路在小数据集上是可行的，但在数据挖掘中的海量大样本集上就不再奏效。因为任何统计不显著的随机性都可能因样本量大，而被有倾向性地误判为显著，即误判为系统性的、有意义的，即使是“表象”，也会被误判为“本质”。

第二，数据挖掘的研究对象往往是总体而非随机样本。数据挖掘的对象一般是现有数据集，它们通常就是人们关注的总体而不是样本。从这个角度讲，统计推断不再必要。当然，数据挖掘并不否认统计推断的重要作用。若将现有数据放到一个更大的时空中去，那么目前这个数据总体也可以视为更大时空中的一个样本。但问题是需确保样本是随机样本，否则统计推断还会因丧失原本的理论基础而不再适用。

另外，有些数据挖掘应用问题只能基于总体而不能基于样本来研究。例如，在信用卡欺诈甄别研究中，若确实存在极少数的恶意透支行为，这些交易数据会因数量很小而不易或无法进入随机样本。若以样本为研究对象，样本中的欺诈特征会变得不再明显甚至消失，从而得到不存在欺诈行为的分析结论。

基于上述原因，数据挖掘不再以统计推断方式验证数据挖掘的结果是否有意义，而是采用一种“退而求其次”的做法，即要求行业专家深度参与数据挖掘过程，并由行业专家判断数据挖掘结果的意义和价值。例如，“所有前列腺癌患者都是男性”，“加油站的信用卡刷卡金额通常在个位为零上出现峰值”，这些结论是否可理解和有价值，需由行业专家去评估。

1.3　数据挖掘能做什么

通常，数据挖掘可以解决四大方面的问题：第一，数据预测；第二，发现数据的内在结构；第三，发现关联性；第四，模式诊断。

1.3.1　数据预测

顾名思义，数据预测就是基于对历史数据的分析，预测新数据的特征，或预测数据的未来发展趋势等。例如，一份关于顾客特征及其近 12 个月消费记录的数据包含诸如顾客的性别、年龄、职业、年收入等属性特征，以及顾客购买商品的种类、金额等消费行为数据。现希望依据该份数据，找到如下问题的答案：

- 具有某种特征（如已知年龄和年收入）的新顾客是否会购买某种商品？
- 具有某种特征（如已知年龄）和消费行为（购买或不购买）的顾客，其平均年收入是多少？
- 某种商品在未来 3 个月将有怎样的销售量？

上述问题均属数据预测的范畴，并有各自不同的应用特点。

第一个问题的答案无非是买或者不买。若将买或不买视为消费行为的两个类别（图 1-5（a）中的圆圈和三角形分别代表“买”和“不买”），则解决该问题的思路是基于已有数据，研究顾客的属性特征与其消费行为间的规律，并借助某种数学模型或者推理规则等定量反映这一规律。进一步，依据该规律对新顾客（图 1-5（a）中的菱形点）的消费行为（菱形点应归为圆圈还是三角形）进行预测。数据挖掘将这类对数据所属类别进行预测的问题统称为分类预测问题。分类的目标是找到某些可将两类或多类分开的数学模型或者推理规则，它们几何上对应一条或若干条直线（或平面或超平面），如图 1-5（a）所示的虚线。进一步，依据新数据与直线（或平面或超平面）的位置关系，预测新数据所属的类别。

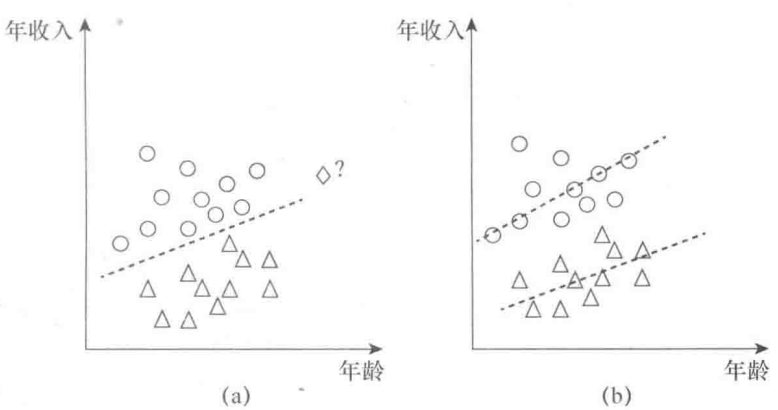


图 1-5　数据预测示意图

第二个问题是对顾客的平均年收入进行预测。目标是找到不同类别客户（购买或不购买）其年收入与年龄间的相关关系，并借助某种数学模型定量反映这种关系。进一步，依

据这种关系对新顾客的平均年收入进行预测。该问题的研究思路与第一个问题基本类似，不同点在于该问题的答案是个数值。数据挖掘将这类数值预测问题统称为回归预测问题。回归预测的目标是找到可反映某个数值型变量与其他诸多变量间相关关系的数学模型，它们几何上对应一条直线（或平面），称为回归直线（或回归平面），如图1-5（b）所示的两条虚线。

对于第三个问题，可以该商品近12个月的销售量为研究对象，分析销售量随时间推移所呈现的变化趋势并进行预测。这类问题是单个时间序列的预测问题，属统计学研究的范畴，这里不做讨论。

总之，这里讨论的数据预测主要包括数据的分类预测和回归预测。但需要注意的是，数据预测是在上述数学模型或推理规则仍适用于新数据的假设下进行的。

事实上，完全可以采用统计学的建模思路解决这两个问题，而且统计学对此也早有极为成熟的分析逻辑，例如Logistic回归和多元线性回归等。但正如前面提及的，传统统计学以随机样本为研究对象、以某种线性关系假设为前提的验证式分析思路，并不能很好地适应大数据背景下的数据挖掘。所以，按照一定策略的“机械式”归纳搜索是数据挖掘解决数据分类和回归问题的主要方式。由此需关注如下两个问题：

第一，用于预测的数学模型或推理规则，是否正确地反映了变量间的总体相关性，是不是数据取值的主体且重要规律的反映。

数据挖掘的对象是海量大数据，大数据量是一把“双刃剑”。它既为探索事物规律、发现变量间相关性提供了数据支撑，也最大程度地掩盖了数据中最重要的最一般化的规律和相关性。因数据量大导致数据挖掘发现的规律或相关性很可能仅仅是大数据中的某个数据子集的局部特征，而数据预测则要求预测依据是数据中一般性和全局性规律的抽象和体现，因为只有这样才有预测的普适性。

为此，需探索规律或相关性是全局性的还是局部性的。一种常见的方法是视已有数据为总体，通过随机抽样大幅减少数据量得到一个小的随机样本，并探索其中的规律和相关性。若总体中的规律仍然存在于小样本中，则有理由认为这个规律是全局性的，因为随机小样本中的规律和相关性通常不会是海量大数据总体中的局部特征，这是随机抽样所决定的。可见，统计学的随机抽样在数据挖掘中仍有非常重要的意义。

第二，用于预测的数学模型或推理规则，是否具有较高的预测性能。

衡量模型是否具有较高的预测性能，通常要看它对新数据的预测结果是否准确，在新数据集上的预测误差是否较小。所以，一般以预测误差作为模型预测性能的测度。预测误差越小，模型的预测性能越好。由于新数据预测结果是否准确在建模时无从得知，计算模型对已有数据集的预测误差测度。预测误差可以是数据分类中的错判率，或者回归分析中的残差方差等，这都是数据挖掘可以直接借鉴的。

但问题是由此计算出的预测误差很可能因数值偏低而放大模型对新数据预测的准确性，无法客观测度模型的预测性能。原因在于，无论统计方法还是数据挖掘方法，都是以最小化当前预测误差为前提的，在最大化拟合已有数据的基础上估计（或搜索）预测模型的参数（或推理规则）。当预测模型（或推理规则）基于已有数据全体时，它在数据全体上的预测误差一定是最小的，但却无法得知它在其他数据集上是否仍有理想的表现，是否会因预测误差增加过大而无法用于新数据的预测。所以，找到对新数据预测误差的估计方法，是数据预测中的重要问题。我们将在后续章节集中讨论这一问题。