国外计算机科学经典教材

# DATA MINING
## Concepts, Models, Methods, and Algorithms

# 数据挖掘
## ——概念、模型、方法和算法

(美) Mehmed Kantardzic 著
闪四清 陈茵 程雁 等译

清华大学出版社

# 数据挖掘

## ——— 概念、模型、方法和算法

(美) Mehmed Kantardzic 著

闪四清 陈茵 程雁 等译

清华大学出版社

北 京

## 内 容 简 介

作为一本教科书，本书全面讲述了数据挖掘的概念、模型、方法和算法。本书共包括 13 章和 2 个附录，全面、详细地讲述了从数据挖掘的基本概念到数据挖掘的整个过程，以及数据挖掘工具及其典型应用领域。

本书编写严谨、内容权威、结构合理、科学规范、语言流畅，特别适合作为高等院校数据挖掘课程的教科书，还适合作为数据挖掘研究人员必备的参考书。

EISBN:04-71-22852-4



**北京市版权局著作权合同登记号　图字：01-2002-3730**

**图书在版编目(CIP)数据**

数据挖掘：概念、模型、方法和算法/(美)康塔尼克著；闪四清等译. — 北京：清华大学出版社，2003

书名原文：Data Mining Concepts, Models, Methods, and Algorithms

ISBN 7-302-06777-5

I. 数… II. ①康… ②闪… III. 数据采集 - 教材 IV. TP274

中国版本图书馆 CIP 数据核字(2003)第 047114 号

**出 版 者**：清华大学出版社　　**地　　址**：北京清华大学学研大厦

http://www.tup.com.cn　　**邮　　编**：100084

**社 总 机**：010-62770175　　**客户服务**：010-62776969

**组稿编辑**：曹康

**文稿编辑**：陈宗斌

**封面设计**：康博

**版式设计**：康博

**印 刷 者**：北京牛山世兴印刷厂

**发 行 者**：新华书店总店北京发行所

**开　　本**：185×260　**印张**：20.25　**字数**：518 千字

**版　　次**：2003 年 8 月第 1 版　2003 年 8 月第 1 次印刷

**书　　号**：ISBN 7-302-06777-5/TP・5042

**印　　数**：1～4000

**定　　价**：42.00 元

# 前　言

从传统上讲，分析人员已经完成了从已记录数据中提取有用信息的任务。但是，现代商业和科学领域的数据增长需要应用基于计算机的方法。随着数据集规模的不断扩大和复杂度的增加，从内行的分析人员直接分析到使用更加复杂和尖端的技术来进行间接的、自动化的分析，这种转换是不可避免的。现代化的计算机技术、网络技术和传感器技术使数据的搜集和组织成为一项毫不费力的任务。但是，所获得的数据需要从已记录的数据转换成有用的信息和知识。数据挖掘就是应用基于计算机的方法论的整个过程，包括新的知识发现技术。

现代世界是一个数据驱动的世界。我们被数据所包围着，这些数据是数值型数据或其他类型，它们都必须经过分析和处理，转换成通知、指导、回答或帮助理解和作出决策的信息。现在是互联网、内联网，数据仓库和数据集市的时代，改变经典数据分析的基本范型的时机成熟了。极大的数据集(有时有数亿条个人记录)被存储在中心数据仓库中，允许分析人员使用更为综合、更为强大的数据挖掘方法。同时数据量很大且还在增加，数据源是无限的，所覆盖的领域是广泛的：包括工业、商业、金融和科学等。

近些年来，从原始数据中发现新知识的方法爆炸性地增长。有鉴于此，一个新的数据挖掘学科被专门研发出来，以便从这些巨型数据集中提取有价值的信息。如果低成本计算机(针对软件实现)，低成本传感器、通信、数据库技术(用于搜集和存储数据)，以及那些能使用计算机并能提出“有趣的”和“有用的”应用问题的应用专家系统都得到迅速的扩展，这丝毫不让人感到惊奇。

数据挖掘技术近来变成了决策者们的热门话题，因为它从大量的历史数据中提供有价值的、隐藏的商业和科学“情报”。但是，数据挖掘实质上不能算是一门新技术。从已记录数据中提取信息和知识是一个在科学和医学研究中已经建立完善的概念。新的内容是几门学科及相应的技术的聚合，这些技术为数据挖掘在科学和企业界发展创造了独一无二的机会。

最初，本书打算作为指导学生的介绍性资料。但是我很快发现，来自不同背景和职业，需要搞清楚大量原始数据的人们，显然也会欣赏一本包含一些最重要的数据挖掘方法、工具和算法的书籍。因此，本书的读者面很广：从希望学习数据挖掘的基本过程和技术的学生，到直接参加跨学科的选择数据挖掘工具小组的分析人员。本书回顾了一些最先进的技术，分析了高维数据空间中的大量原始数据，以提取有助于决策的新信息。书中涉及到的大部分技术定义、分类和解释都已获得广泛认可，在书末的参考书目中它们都曾经出现过。本书重点关注一种对数据挖掘过程的所有阶段来说都是

系统的、平衡的方法，并用充足的示例来展现这些方法。我希望这些经过精心准备的例子能给读者自己的数据挖掘技术和工具的选择以及构造提供额外的论据和指南。要更好地理解大部分已介绍技术的实现细节对读者来说也是个很大的挑战，他们要开发自己的工具或改进他们所用的方法和技术。

要讲授数据挖掘，必须强调所用方法的概念和属性，而不是机械地应用不同的数据挖掘工具。不论所有基于计算机的工具怎样吹嘘，它们也不能代替那些决定过程怎样设计及采用什么工具的实践者。对方法、模型以及它们怎样运转及其运转原理的深入理解是有效和成功运用数据挖掘技术的先决条件。任何在数据挖掘领域的研究者和实践者都要意识到这些问题，以便成功地应用一种特定的方法，理解一种方法的局限性，或者开发新技术。本书提出和讨论了这些问题和理论，然后描述了起源于统计学、机器学习、计算机图形学、数据库、信息检索、神经网络、模糊逻辑和进化计算的具有代表性的和流行的方法。讨论了那些在揭示大型数据集的重要模式、趋向和模型的方法中，已经被证明具有关键性的方法。

虽然关注技术是容易的，当您阅读本书时记住，仅凭技术是不能提供整个解决方案的。写作本书的目标之一是把与数据挖掘有关的那种夸大减到最小，而不是对合理的期望作出错误的允诺。我努力采用更为客观的方法，描述了在数据挖掘应用中得出可靠、有用结果所必需的过程和算法。

我不提倡使用任何特殊的产品或技术优于使用其他产品或技术。数据挖掘过程的设计者必须要有足够的素养，以便选择适当的方法和软件工具。我希望当一个读者读完本书时，可以成功和有效地开始并完成数据挖掘过程的所有阶段的基本活动。

Mehmed Kantardzic
Louisville, KY
2002 年 8 月

# 作 者 介 绍

Mehmed Kantardzic 在波斯尼亚的 Sarajevo 大学获得学士、硕士和博士学位，1994 年以前，他一直在 Sarajevo 大学电子工程学院任副教授。1995 年，他进入 Louisville 大学，从 2001 年起，担任 Louisville 大学计算机工程和计算机科学系的副教授和数据挖掘实验室主任。他的研究兴趣包括：数据挖掘和知识发现、软计算、神经网络概括的可视化、互联网上的多媒体技术以及分布式智能系统。Kantardzic 博士近年来集中研究数据挖掘技术在生物医学研究中的应用。他编写过 5 本书，启动和领导超过 30 个研究和开发项目。在仲裁期刊和会议学报中发表了 120 多篇文章。同时，Kantardzic 博士也是 IEEE、ISCA、SPIA 的会员，是丹佛 ISCA'99 会议的规划主席，也是 2001 年弗吉尼亚阿林顿智能系统国际会议的主席。

# 目　　录

# 第1章　数据挖掘的概念

**本章目标**

- 理解对大型的、复杂的和信息丰富的数据集进行分析的必要性。
- 明确数据挖掘过程的目标和首要任务。
- 描述数据挖掘技术的起源。
- 认识数据挖掘过程所具有的迭代特点，说明数据挖掘的基本步骤。
- 解释数据的质量对数据挖掘过程的影响。
- 建立数据仓库和数据挖掘之间的联系。

## 1.1　概述

现代科学和工程建立在用“首要原则模型(first-principle models)”来描述物理、生物和社会系统的基础上。这种方法从基础的科学模型入手，如牛顿运动定律或麦克斯韦的电磁公式，然后基于模型来建立机械工程或电子工程方面的各种应用。在这种方法中，用实验数据来验证基本的“首要原则模型”，以及对一些难以直接测量或者根本不可能直接测量的参数进行评估。但是在许多领域，基本的“首要原则模型”往往是未知的，或者研究的系统太复杂而难以进行数学定型，随着计算机的广泛应用，像这样的复杂系统生成了大量的数据。在没有“首要原则模型”时候，可以利用这些易得的可用数据，通过对系统变量之间可以利用的关系(即未知的输入输出相关性)进行评估来导出模型。这样，传统的建模及基于“首要原则模型”进行分析的方法与开发模型及直接对数据进行相应分析的方法之间普遍存在着范型变换。

我们都逐渐习惯面对这样的一个事实——超量的数据充斥着我们的电脑、网络和生活，政府机构、科研机构和企业都投入大量的资源去收集和存储数据。实际上，这些数据中只有一小部分将会被用到，因为在很多情况下，要么数据量简直太大了，难于管理，要么就是数据结构太复杂，不能进行有效的分析。这种情况是怎么发生的呢？根本的原因是人们创建一个数据集时往往把精力都集中在如数据的存储效率的问题上，而没有去考虑数据最终是怎样使用和分析的。

对大型的、复杂的、信息丰富的数据集的理解实际上是所有的商业、科学、工程领域的共同需要，在商务领域，公司和顾客的数据逐渐被认为是一种战略资产。在当今的竞争世界中，吸取隐藏在这些数据后面的有用知识并利用这些知识的能力变得愈加

重要。运用基于计算机的方法，包括新技术，从而在数据中获得有用知识的整个过程，就叫做数据挖掘。

数据挖掘是一个反复迭代的过程，在这个过程中，所取得的进步用“发现”来定义，而这种发现是通过自动或手工方法取得的。在对什么将会构成一个“有趣的”结果没有预定概念的初步探测性分析方案中，数据挖掘非常重要。它从大量的数据中搜寻有价值的、非同寻常的新信息，是人和计算机合力的结果；它在人类描述问题和目标的知识与计算机的搜索能力之间寻求平衡，以求获得最好的效果。

在实践中，数据挖掘的两个基本目标往往是预测和描述。预测涉及到使用数据集中的一些变量或域来预测其他我们所关心变量的未知或未来的值；另一方面，描述关注的则是找出描述可由人类解释的数据模式。因此，可以把数据挖掘活动分成下述两类。

1) 预测性数据挖掘：生成已知数据集所描述的系统模型。

2) 描述性数据挖掘：在可用数据集的基础上生成新的、非同寻常的信息。

在预测领域的后期，数据挖掘的目标是得出一种模型，以可执行码来表示。这种可执行码可以用于执行分类、预测、评估或者其他相似的任务。而描述性领域的后期，数据挖掘的目标是利用大型数据集中的未知模式和关系获得对所分析系统的理解。对特定的数据挖掘的应用，预测和描述的相对意义有相当大的变化。预测和描述的目标都是通过数据挖掘技术来实现的，本书将在后面介绍这些技术。数据挖掘的基本任务如下：

1. 分类——预测学习功能的发现，此功能将一个数据项分到几个预定义类中的一类。

2. 回归——预测学习功能的发现，此功能将一个数据项映射到一个真实值预测变量。

3. 聚类——一种普遍的描述性任务，寻求以确定有限的一组类别或类来描述数据。

4. 总结概括——一项附加的描述任务，寻找对数据集或子集的简单描述方法。

5. 关联建模——发现描述变量之间或者数据集或其一部分的特征值之间的重要的相关性的本地模型。

6. 变化和偏差检测——发现数据集中最重要的变化。

针对复杂的和大型的数据集的数据挖掘任务，第 4 章给出了更加正式的带有图形化解释和说明性示例的方法。这里给出了当前介绍性的分类和定义，只是让读者对可使用数据挖掘技术来解决的问题和任务的广阔领域有一个初步感受。

数据挖掘成功地达到预定目标，很大程度上依赖于设计者投入的精力、知识和创造力。从本质上讲，数据挖掘就像是解题：从问题的个别方面来看，结构并不复杂。但把它作为一个整体时，它们就能组成一个详尽的系统。当你试着去拆分这个系统时，你可能会遭遇失败，开始把各部分组合在一起又往往会为整个过程而苦恼。但是，一旦你知道怎么从部分着手，你就会发现其实问题并没有开始那么困难。同样的道理可以类推到数据挖掘中，开始的时候，数据挖掘过程的设计者可能对数据源知道的不多。

如果他们知道很多，就很可能对完成数据挖掘失去兴趣。从个别来看，数据似乎是简单、完整和可解释的。但是从整体的角度看时，它们完全是另外一个面貌——具有威胁性、难以理解，就像是一道难题。因此，要想在数据挖掘过程中成为一个分析者和设计者，除了要具备非常专业的知识外，还要有创造性的思维以及从不同角度看问题的主动性。

数据挖掘是计算机行业中发展最快的领域之一，以前数据挖掘只是结合了计算机科学和统计学而产生的一个让人感兴趣的小领域，如今，它已经迅速扩大成为一个独立的领域。数据挖掘的强大力量之一在于它具有广泛的方法和技术，以应用于大量的问题集。既然数据挖掘是一个在大型数据集上进行的自然行为，其最大的目标市场应该是整个数据仓库、数据集市和决策支持业界。包括诸如零售、制造、通信、医疗、保险、运输等行业的专业人士。在商业界，数据挖掘可用于发现新的购买倾向、设计投资战略和在会计系统中探测未经认可的开支，增加销售业务。其结果可用于向顾客提供更集中的支持和关注。数据挖掘技术也能应用于解决商业过程重构问题，其目标是了解商业操作和组织之间的相互作用和关系。

对一些法律的执行部门和专门的调查机构来说，它们的任务是识别欺诈行为和发现犯罪倾向。这些单位也成功地运用了数据挖掘技术。例如：这些方法能辅助分析人员识别麻醉品组织的相互交流作用中的犯罪行为模式、洗黑钱活动、内部贸易操作、连环杀手的行动以及越境走私犯的目标。数据挖掘技术也被情报部门的人员使用，他们把维持大型的数据源作为与国家安全问题相关活动的一部分。本书附录 B 对当今数据挖掘技术的典型商业应用作了一个简洁的纵览。

## 1.2　数据挖掘的起源

看看作者们对数据挖掘的描述有多大不同！显然我们在数据挖掘的定义上还远没有达成一致，甚至没有制定出到底什么是数据挖掘，数据挖掘是使用学习方法将统计学强化后的一种形式，它是一个全新的革命性的概念吗？从我们的观点看，大部分数据挖掘问题和相应的解决方法都起源于传统的数据分析。数据挖掘起源于多种学科，其中最重要的两门是统计学和机器学习，统计学起源于数学，因此，它强调数学上的精确。在实践测试之前，在理论基础上建立一些东西的要求是明智的，相比之下，机器学习更多地起源于计算机实践。这就导致了实践的倾向，自觉地对一些东西进行检验来查看它表现的好坏，而不是去等待有效性的正式证据。

如果说数据挖掘的统计学方法与机器学习方法之间的主要区别之一是数学和形式化被给予的地位的话，另一个区别就在于模型和算法规则之间侧重点不同。现代统计学几乎完全是由模型概念驱动的，是一个假定的结构，或者说是一个结构的近似，这

个结构能够产生数据。统计学强调模型，而机器学习倾向于强调算法。这不会让人感到吃惊，“学习”这个词包括了过程的概念，即一种含蓄的算法。

数据挖掘中的基本模型法则也起源于控制理论，控制理论主要应用于工程系统和工业过程。通过观察一个未知系统(也被称为目标系统)的输入输出信息，以决定其数学模型的问题通常被叫做系统识别。系统识别的目标是多样化的，并且是从数据挖掘的立场出发的。最重要的是预测系统的行为，并解释系统变量之间的相互作用和关系。

系统识别通常包括两个组织严密的步骤：

(1) 结构识别——在这一步骤中，我们要应用到关于目标系统的先验知识来决定一类模型，在这类模型中搜寻将要导出的最适合的模型。通常这类模型都由一个参数函数 y=f($u$,$t$)来表示，y 表示模型的输出，u 是一个输入向量，t 是一个参数向量，函数 f 的测定是依赖于问题的，函数基于设计者的经验、直觉和控制目标系统的自然法则。

(2) 参数识别——在第二步中，当模型结构已知时，我们要做的就是应用优化技术来测定参数矢量 t 以便结果模型 y*=f($u$,$t$*)能恰如其分地描述目标系统。

一般而言，系统识别不是一个一次通过的过程，结构和参数识别都要重复进行直到找到满意的模型为止，图 1-1 图形化地描述了迭代的过程。每次迭代中的典型步骤如下：

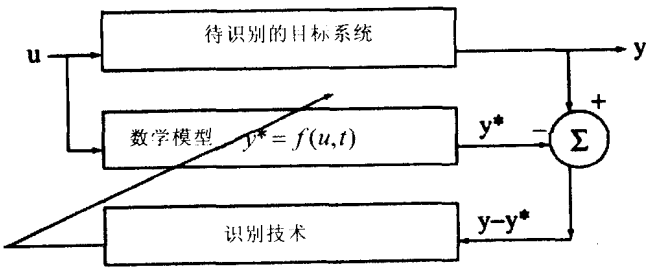


图 1-1　参数识别结构图

(1) 详细说明一类公式化(数学化)的模型并对它们进行参数化，y* = f(u,t)代表需识别的系统。

(2) 选择最适合可用数据集的参数(差值 y - y*最小)，完成参数识别。

(3) 导入正确性检验来检测识别出来的模型是否能正确响应没见过的数据集(通常称为检验，确认，或核查数据集)。

(4) 一旦正确性检验的结果满足要求就停止这一过程。

如果我们对目标系统一无所知，结构识别就会变得困难，我们必须在通过试验和在有误差的情况下选择结构。我们对大多数工程系统和工业过程了解较多，而在大多数应用数据挖掘技术的目标系统中，这些结构完全是未知的，或者这些结构太复杂而难以得到一个适当的数学模型。因此，用于参数识别的新技术已经被开发出来了，而且这些技术是当今一系列数据挖掘技术的一部分。

最后，我们要区别在数据挖掘中“模型”和“模式”是怎么解释的？“模型”是一个“大型”的结构，或许是对多数(有时是全部)案例的关系的总结。反之，“模式”是

一个局部结构，满足于少数的案例或者很小的数据空间区域。值得注意的是，“模式”这个词用于模式识别时的意义和它用于数据挖掘时的意义有很大的区别。在模式识别中，它是指将一个特定对象特征化的一个度量向量，是多维数据空间里的一个点。在数据挖掘中，模式仅仅是一个局部模型。本书把 N 维数据向量作为样本。

## 1.3　数据挖掘过程

数据挖掘作为一门学科，我们没有试图去涵盖关于它的所有可能的方法和所有不同的观点，而是从一个可能的、十分广泛的数据挖掘的定义开始。

定义：数据挖掘是一个从已知数据集合中发现各种模型、概要和导出值的过程。

这里，“过程”一词相当重要。即使是在一些专业环境中，也有这样的一种观点：数据挖掘只是采摘和应用基于计算机的工具来匹配出现的问题并自动获取解决方案。这是一种对世界人为的理想化所形成的误解，为什么这是错的呢？有几种原因，一个原因是：数据挖掘不只是一些独立工具的一个集合，它们彼此完全不同，并且等待着去匹配问题。第二个原因在于把一个问题和一种技术视为等同的观念。在极少数情况下，研究问题可以充分、精确地陈述出来，使得方法的单独和简单的应用将会满足。实际上，现实中所发生的是：数据挖掘变成了一个反复的过程。一个人对数据进行研究，利用一些分析工具对数据进行检查，决定从另外一个角度来看它，可能会对数据进行修改，然后又回到开始，应用别的数据分析工具，得到一个更好的或不同的结果。这个过程可能循环许多次，每一种技术都被用到，以便查明数据的细微的不同的方面——询问一个数据的细微不同的问题。在这里不得不描述的是令现代数据挖掘激动人心的发展史。尽管如此，数据挖掘仍然不是统计学、机器学习以及其他方法和工具的随意应用，它不是在分析技术空间里面乱闯，而是一个精心策划和深思熟虑过的，决定什么才是最有用的、最有前景的和最有启迪作用的一个过程。

认识到这一点很重要：从数据中发现或估计其相关性，或从中完整地挖掘出新数据，只是人们所采用的一般实验性程序中的一部分，这些人包括科学家，工程师和其他应用标准步骤从数据中得出结论的人。适合数据挖掘问题的一般实验性程序包括以下步骤。

### 1.3.1　陈述问题和阐明假设

大多数基于数据的模型研究都是在一个特定的应用领域里完成的。因此，为了提出一个有意义的问题的陈述，拥有领域内详尽的知识和经验是必不可少的。不幸的是，许多应用研究往往以牺牲对问题的清晰描述为代价而集中在数据挖掘技术上，在这一

步中，模型建立者通常会为未知的相关性指定一组变量，如果可能，还会指定此相关性的一个大体形式作为初始假设。对当前问题可能会有几个阐明的假设。这一步要求将应用领域的专门技术和数据挖掘模型相结合，实际上，这往往意味着数据挖掘专家和应用专家之间密切地相互协作。在成功的数据挖掘应用中，这种协作并没有停止在初始阶段，而是持续了数据挖掘的整个过程。

### 1.3.2 数据收集

这一步是关于数据是怎样产生和收集的。通常有两种截然不同的可能。第一种是当数据产生过程在专家(建模者)的控制之下时：这种方法被认为是“设计实验”。第二种情况是专家不能影响数据产生过程时：这种方法被认为是“观察法”。观察设置，也就是数据随机产生，在大多数数据挖掘应用中都被采用。具有代表性的是，数据收集完成后取样的分布也是完全未知的，或者说其分布是在数据搜集过程中部分或者不明确地给出的。但是，我们要理解数据搜集是怎样影响它的理论分布的，这一点相当重要。这样的先验知识对以前的建模以及后来的对结果的最终解释都是相当重要的。同样，对于用于评估模型的数据以及后面用于测试和应用于模型的数据，要确定它们来自同样的未知的样本分布也是很重要的。如果分布不同，那么评估的模型就不能在最终的结果应用中成功地使用。

### 1.3.3 数据预处理

在观察设置中，数据常常采集于已存在数据库、数据仓库和数据集市中。数据预处理通常包括至少两个常见任务：

1. 异常点的检测(和去除)——异常点是与众不同的数值，这些数值和大多数观察值不一致。一般来讲，异常点是由测量误差、编码和记录误差产生的，有时也来自于自然的异常值。这种不具备代表性的样本以后会严重影响模型的产生。对异常点有两种处理办法：

a) 把检测并最终去除异常点作为预处理阶段的一部分。

b) 寻找不受异常点影响的健壮性建模方法。

2. 比例缩放、编码和选择特征——数据预处理过程包括几个步骤，如各种比例缩放和不同类型的编码。例如，一个取值范围为[0，1]的特征和一个取值范围为[-100，1000]的特征，它们在应用技术中的加权是不一样的，对最终的数据挖掘结果的影响也不尽相同。因此，推荐对它们进行比例缩放并使它们加权相同以进行进一步的分析。同样，通过为后来的数据建模提供较少量资料丰富的特征，详细应用的编码方法通常可以完

成维度归约。

这两类预处理任务只是在数据挖掘过程中大量预处理活动的说明性的例证。数据预处理步骤不应该与数据挖掘的其他阶段完全独立起来考虑，在数据挖掘过程的每一次迭代中，所有的活动加在一起都能为后面的迭代定义新的和改进的数据集。通常，通过把先验知识合并为具体应用比例缩放和编码的形式，一种好的预处理方法能为数据挖掘技术提供最佳的陈述。更多关于这些技术和预处理阶段的内容大体上将会在第 2 章和第 3 章中给出。在第 2 章和第 3 章中，我们把预处理和相应的技术功能性地划分为两个子阶段：数据准备和数据维度归约。

### 1.3.4　模型评估

选择并实现适当的数据挖掘技术是这一阶段的主要任务。这个过程往往并不是直截了当的，实际上，实现是建立在几个模型的基础上的，从中选择最好的模型是额外的任务，从数据中学习和发掘的基本原则将会在本书的第 4 章介绍，随后，第 5~13 章解释和分析一些特殊的技术，应用这些技术可以从数据中成功地学习，也可以应用这些技术找到适当的模型。

### 1.3.5　解释模型和得出结论

在大多数情况下，数据挖掘模型应该有助于决策。因此，要对这种模型进行说明以使模型有用，因为人们不会在复杂的“黑箱模型”的基础上作决策。注意，模型准确性的目标和模型说明的准确性的目标有点互相矛盾。一般来说，简单的模型容易说明，但是其准确性就差一些。现代的数据挖掘方法寄望于使用高维度的模型来获得高精度的结果。用特定的技术验证这些结果对这些模型进行解释说明被看作是一项独立的任务，同时也是非常重要的。用户不会想要一个数百页的数值结果，这样的结果难以理解，不能总结、解释，也不能用这样的结果来进行成功的决策。

尽管本书将重点放在数据挖掘过程中的第 3 步和第 4 步，我们还是必须了解它们只不过是一个更为复杂的过程中的两个步骤而已，不管是个别地来看数据挖掘的各个阶段，还是整个的数据挖掘过程，都是高度反复的，如图 1-2 所示，对整个过程的良好理解对任何成功的应用都是重要的。如果没有恰当地收集和预处理数据，或者没有对问题进行有意义的明确表述，不管第 4 步中所使用的数据挖掘方法有多强大，最终模型都将是无效的。