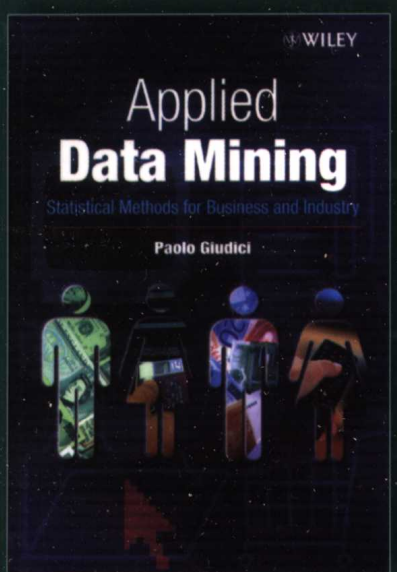
国外计算机科学教材系列

实用数据挖掘

Applied Data Mining

Statistical Methods for Business and Industry



[意] Paolo Giudici 著
袁方 王煜 王丽娟 等译
袁方 审校

WILEY
PHEI
電子工業出版社
Publishing House of Electronics Industry
http://www.phei.com.cn

国外计算机科学教材系列

实用数据挖掘

Applied Data Mining

Statistical Methods for Business and Industry

[意] Paolo Giudici 著

袁 方 王 煜 王丽娟 等译

袁 方 审校

電子工業出版社

Publishing House of Electronics Industry

北京 · BEIJING

内容简介

本书对面向应用的数据挖掘方法进行了清晰的阐述，包括经典的多元统计方法、贝叶斯多元统计方法、基于机器学习的数据挖掘方法和基于计算的数据挖掘方法等。介绍了数据挖掘领域中许多最新的研究成果，如关联规则、序列规则、图示马尔可夫模型、基于存储的推理、信用风险和Web挖掘等。并详细介绍了选自实际工业项目的6个应用实例，强调了数据挖掘方法的实用性。

本书主要面向计算机科学、信息管理、应用统计学和经济学等专业的高年级本科生和研究生。对实际从事海量数据分析和处理的技术人员也有很好的指导作用和参考价值。

Paolo Giudici: **Applied Data Mining: Statistical Methods for Business and Industry.**
ISBN 0-470-84678-X

版权贸易合同登记号　图字：01-2003-6403

图书在版编目（CIP）数据

实用数据挖掘/（意）朱迪茨（Giudici, P.）著；袁方等译. -北京：电子工业出版社，2004.6
（国外计算机科学教材系列）
书名原文：Applied Data Mining: Statistical Methods for Business and Industry
ISBN 7-120-00012-8

Ⅰ.实… Ⅱ.①朱… ②袁… Ⅲ.数据采集-教材 Ⅳ.TP274

中国版本图书馆CIP数据核字（2004）第043558号

责任编辑：谭海平　许菊芳
印　　刷：北京兴华印刷厂
出版发行：电子工业出版社
　　北京市海淀区万寿路173信箱　邮编：100036
经　　销：各地新华书店
开　　本：787 × 980　1/16　　印张：18.5　字数：408千字
印　　次：2004年6月第1次印刷
定　　价：28.00元

出 版 说 明

21世纪初的5至10年是我国国民经济和社会发展的重要时期，也是信息产业快速发展的关键时期。在我国加入WTO后的今天，培养一支适应国际化竞争的一流IT人才队伍是我国高等教育的重要任务之一。信息科学和技术方面人才的优劣与多寡，是我国面对国际竞争时成败的关键因素。

当前，正值我国高等教育特别是信息科学领域的教育调整、变革的重大时期，为使我国教育体制与国际化接轨，有条件的高等院校正在为某些信息学科和技术课程使用国外优秀教材和优秀原版教材，以使我国在计算机教学上尽快赶上国际先进水平。

电子工业出版社秉承多年来引进国外优秀图书的经验，翻译出版了“国外计算机科学教材系列”丛书，这套教材覆盖学科范围广、领域宽、层次多，既有本科专业课程教材，也有研究生课程教材，以适应不同院系、不同专业、不同层次的师生对教材的需求，广大师生可自由选择和自由组合使用。这些教材涉及的学科方向包括网络与通信、操作系统、计算机组织与结构、算法与数据结构、数据库与信息处理、编程语言、图形图像与多媒体、软件工程等。同时，我们也适当引进了一些优秀英文原版教材，本着翻译版本和英文原版并重的原则，对重点图书既提供英文原版又提供相应的翻译版本。

在图书选题上，我们大都选择国外著名出版公司出版的高校教材，如Pearson Education培生教育出版集团、麦格劳－希尔教育出版集团、麻省理工学院出版社、剑桥大学出版社等。撰写教材的许多作者都是蜚声世界的教授、学者，如道格拉斯·科默（Douglas E. Comer）、威廉·斯托林斯（William Stallings）、哈维·戴特尔（Harvey M. Deitel）、尤利斯·布莱克（Uyless Black）等。

为确保教材的选题质量和翻译质量，我们约请了清华大学、北京大学、北京航空航天大学、复旦大学、上海交通大学、南京大学、浙江大学、哈尔滨工业大学、华中科技大学、西安交通大学、国防科学技术大学、解放军理工大学等著名高校的教授和骨干教师参与了本系列教材的选题、翻译和审校工作。他们中既有讲授同类教材的骨干教师、博士，也有积累了几十年教学经验的老教授和博士生导师。

在该系列教材的选题、翻译和编辑加工过程中，为提高教材质量，我们做了大量细致的工作，包括对所选教材进行全面论证；选择编辑时力求达到专业对口；对排版、印制质量进行严格把关。对于英文教材中出现的错误，我们通过与作者联络和网上下载勘误表等方式，逐一进行了修订。

此外，我们还将与国外著名出版公司合作，提供一些教材的教学支持资料，希望能为授课老师提供帮助。今后，我们将继续加强与各高校教师的密切联系，为广大师生引进更多的国外优秀教材和参考书，为我国计算机科学教学体系与国际教学体系的接轨做出努力。

电子工业出版社

教材出版委员会

译 者 序

随着计算机技术，特别是数据库技术的快速发展和广泛应用，各行各业积累的数据量越来越大，传统的数据处理方式已很难充分利用蕴藏在这些数据中的有用知识，于是数据挖掘技术应运而生。数据挖掘是为了发现事先未知的规则和联系而对大量数据进行选择、探索和建模的过程，目的在于得到对数据库的拥有者来说清晰而有用的结果。

近几年陆续有几本关于数据挖掘的书籍（影印版或翻译版）问世。但本书的特点在于其理论介绍与应用实例的统一，无论是从事数据挖掘的理论研究，还是把数据挖掘技术应用于实际工作都具有很好的参考价值。本书适合于计算机科学、商业管理和统计学专业的高年级本科生和研究生阅读，也适合于数据挖掘的应用技术人员参考。

本书的第一部分主要介绍数据组织及数据挖掘方法。其中包括数据仓库、Web仓库、数据集市、数据矩阵、一元分析、多元分析、聚类分析、决策树和神经网络等。本书的第二部分是应用实例，对市场购物篮分析、Web点击流分析、网站用户分析、客户关系管理、信用评分、电视观众预测等实际应用案例进行了详细的分析。

本书的作者Paolo Giudici博士是意大利Pavia大学的统计学副教授，数据挖掘实验室负责人，欧洲工商业统计网数据挖掘工作组的协调人，在数据挖掘领域有很多高水平的著作和论文出版，本书就是作者研究工作和实际应用的总结。

本书译者为袁方（前言、第1章、第2章、第5章）、王煜（第7章~第10章）、王丽娟（第3章、第4章）、孟增辉（第6章）、李弛（第11章）、侯亚丽和李宁译（第12章）。在各章译稿的基础上，全书由袁方审校、定稿。郁亚辉和张明也参与了部分章节的翻译工作。李天柱教授仔细审校了部分译稿，并提出了许多宝贵的修改意见。

由于本书的专业性很强，加上时间紧迫及我们的水平有限，因此译文中难免有错误和不当之处，敬请各位读者批评指正。

前　言

在当今的信息社会中，日益增加的数据量需要有效的工具来对这些数据进行建模和分析。数据挖掘和实用统计方法是从这些数据中抽取知识的合适工具。数据挖掘可以定义成为发现事先未知的模式而对大型数据库进行选择、探查和建模的过程。它和实用统计方法的主要区别在于其范围，实用统计方法关心的是统计方法在数据上的应用，而数据挖掘则是旨在生成决策规则（用于说明商业目标）的数据抽取和分析的全过程。

尽管数据挖掘是一个非常重要和快速发展的领域，但是相关的文献资料还不全面，特别是在统计的角度相关文献更少。大多数数据挖掘的书籍或者过分强调学术性（面向计算机科学），或者过分强调实用性（市场驱动），本书试图在数据挖掘方法和工商业领域的应用之间建立一个桥梁，这个桥梁是通过采用一致的和严格的统计建模方法实现的。

本书不仅介绍来自于机器学习和统计领域的数据挖掘方法，还针对欲达到的商业目的对它们进行介绍，因此在书名中加上了“实用”这个词。本书的第二部分是一组案例研究，依据性能和可用性与第一部分中介绍的方法进行了对照。第一部分从比较宽的范围介绍了目前用于数据挖掘的所有方法，并把它们集成到了一个功能框架中，这些方法从本质上被分成基于计算的方法（如关联规则、决策树和神经网络等）和基于统计的方法（如回归模型、广义线性模型和图模型等）。

第二部分的案例研究给出了在工业领域从事大数据量项目的专业指导，如客户关系管理、Web分析、风险管理及更广范围的市场营销和金融管理。介绍了一些必要的形式化工具和数学工具。想了解更多内容的读者可以参看参考文献，第2章～第6章后面给出了阅读指导。

本书是我从1989年开始学习过程的总结，当时我是美国明尼苏达大学的统计学专业研究生。从那时起，我的研究工作一直集中在计算和多元统计之间的相互影响上，1998年我开始建立一个由统计人员组成的数据挖掘小组，现已发展成意大利Pavia大学的数据挖掘实验室。

感谢Wiley出版社提出的建议和对这项工作的鼓励，特别是统计学和数学编辑及助理编辑 Sian Jones 和 Rob Calver。还要感谢 Greg Ridgeway，他审阅了最后的手稿并提出了一些改进建议。最后，最衷心地感谢我的妻子 Angela，她长期鼓励我在这个领域进行研究，本书献给她和我的儿子 Tommaso，他于2002年5月24日出生，当时我正在修改本书的手稿。

我希望人们喜欢这本书并最终应用于实际工作，我非常高兴看到对本书的意见和建议（我的E-mail地址是 giudici@unipv.it），在后续版本中我愿意吸收读者的意见和建议。

Paolo Giudici

2003年于意大利 Pavia 大学

目　录

第一部分　方法

第二部分 商业应用

第1章 绪 论

目前，每一个个人和组织（企事业单位及家庭）都能得到大量的关于自身和生存环境的数据与信息。这些数据具有预测外部环境变化趋势的潜力，但到目前为止，这种潜力没有得到充分开发利用，特别是在作为本书主题的商业领域。这有两个主要原因。一是数据分散在相互无关的不同档案系统中，数据缺乏良好的组织结构；二是缺少对统计工具及其数据潜力的深刻认识，这导致了对相关数据进行有效分析综合的产生。

两个方面的成就有助于克服这些问题。一是软硬件产品的低价格、高性能趋势得到继续，这允许各个组织去收集数据并组织成结构化的数据，以便于访问和转换；二是方法的研究，特别是在计算和统计领域，方法的研究导致了灵活的和可扩展的算法能够用于分析大的存储数据。这两个方面的进展意味着数据挖掘作为重要的支持决策的智能工具已经扩展到了很多商业领域。

本章介绍数据挖掘的本质思想，给出数据挖掘的定义并和统计学、计算机科学中的相关主题进行比较，描述数据挖掘的过程并对数据挖掘软件给予简要介绍。本章的最后一部分介绍本书的内容组织并给出进一步阅读的建议。

1.1 什么是数据挖掘

为了理解数据挖掘（data mining）的含义，有必要看一下它的字面意思。英文中挖掘（mine）就是抽取（extract），这个词通常是指从地下抽取隐藏的贵重资源的挖掘操作。该词和数据之间的联系是：对数据进行深入的研究，目的在于从大量数据中去发现事先未注意到的额外信息。从科学研究的角度看，数据挖掘是一门相对较新的学科，主要基于计算科学、市场营销学和统计学等学科的研究。用于数据挖掘的很多方法来源于两个研究分支，一个是机器学习，另一个是统计学，特别是多元的计算统计学。

和计算机科学、人工智能相关的机器学习用于发现数据中的关系和规则，这些关系和规则可以表示成一般规律。机器学习的目的是再现数据生成的过程，允许分析者从已知数据归纳出新的未知的案例。1962年，Rosenblatt提出了称为感知器的第一个机器学习模型，接着神经网络在20世纪80年代后半期得到发展。与此同时，一些研究者完善了主要用于分类问题的决策树理论，统计学一直用于建立分析数据的模型并且目前可以用计算机来完成。从20世纪80年代后半期开始，作为统计分析基础的计算方法的重要性日益增加，统计方法用于实际的多元统计应用得到同步发展。从20世纪90年代开始，统计学家也对机器学习方法表现出了兴趣，这导致了方法学的重要发展。

到20世纪80年代末，机器学习方法的应用已超出计算和人工智能领域。特别是在数据库市场的应用，数据库用来提高市场竞争力，数据库中的知识发现（KDD）用于描述所有从已知数据中发现关系和规则的方法。逐渐地，KDD扩展成描述从数据库中推断信息的整个过程，从初始商业目标的确定到决策规则的使用。数据挖掘用于描述KDD中的一个组成部分，在KDD中把学习算法应用于数据。

1995年在加拿大蒙特利尔召开的第一届知识发现和数据挖掘国际会议上，“数据挖掘”概念第一次由Usama Fayaad提出，这次会议一直被认为是该领域的主要会议之一。KDD是一种分成若干阶段的集成分析技术，目的在于从大量的已知数据中推断事先未知的、看起来没有任何明显的规则或重要联系的知识。随着数据挖掘概念的建立，逐渐变成了整个推断知识过程的同义词。先前的定义忽略了一个重要的方面——数据挖掘的根本目的，数据挖掘的目的是得到根据相关性可以测量的结果——商业利益。这里给出数据挖掘更完整的定义：

> 数据挖掘是为了发现事先未知的规则和联系而对大量数据进行选择、探索和建模的过程，目的在于得到对数据库的拥有者来说清晰而有用的结果。

在商业领域中，结果的效用变成了商业结果本身。因此，数据挖掘与统计分析的区别不是数据量的大小，也不是使用的分析方法的不同，而是对已知的数据库、分析工具和商业知识的集成。应用数据挖掘方法意味着遵循一个集成的过程，包括：把商业需求转换成要分析的问题，检索用于分析的数据库，应用基于统计技术的计算机算法得到对战略决策有用的重要结果。战略决策本身又提出新的测量需求和新的商业需求，形成由数据挖掘引起的知识的良性循环（Berry 和 Linoff, 1997）。

数据挖掘不只是计算机算法和统计技术的应用，它是一个商业智能过程，可以和信息技术一起支持商业决策。

1.1.1 数据挖掘与计算

数据挖掘的出现与计算机技术的发展，特别是与数据库的快速发展密切相关。在这一小节我们将阐明一些概念。

简单易用的查询和报告工具帮助我们在不同层次中探索商业数据。查询工具检索信息，报告工具则清晰地表现信息，分析结果可以通过客户－服务器网络、内联网甚至互联网进行传送。网络允许共享，所以数据分析可以在最合适的平台上进行。这可以充分发挥远程服务器的数据分析潜力，并在本地PC上得到分析报告。一个客户－服务器网络必须能够适应所有类型的远程请求，从简单的数据排序到使用SQL语言对数据库中的数据进行抽取概括的特殊请求。

数据检索和数据挖掘的相同点是，从档案文件或数据库中抽取感兴趣的数据和信息。区别在于数据检索对信息的抽取规则是事先定义好的，抽取的是外在信息。一个典型的例

子是，某公司市场部要查询所有至少同时按顺序购买过一次产品A和产品B的客户的个人信息。该请求可能基于一个未得到证实的想法：这样的购买行为之间存在某种联系。查询得到的人群可能是未来商业广告的目标，这样得到的成功率（即实际购买推销产品的客户数与作为推销对象的总用户数的比值）比以其他方式得到的成功率要高许多。更进一步说，没有对数据进行初步的统计分析就很难预测成功率，而且对于得到更好的客户特性信息以改善商业竞争效果是不可能的。

数据挖掘与数据检索的不同在于，数据挖掘寻找现象之间事先未知的关系和关联，能够基于数据判断决策的有效性对目标数据做出合理的评价。不要把数据挖掘和用来建立多维报告的工具（如OLAP，在线分析处理）相混淆，OLAP是用来揭示对应二维报告的变量之间关系的图形工具。与OLAP不同，数据挖掘对所有可用变量以不同的方式进行组合，也就是说，我们可以超越OLAP中对概括结果的可视化表示，为商业领域建立有用的模型。数据挖掘不仅仅对数据进行分析，还是一个相当复杂的过程，数据分析只是其中的一个方面。

OLAP是一种重要的商业智能工具。查询和报告工具描述数据库（广义上包括数据仓库）中包含的内容，而OLAP解释为什么存在某些特定的关系。用户对变量之间可能存在的关系做出假设，通过观察数据来证实自己的判断。假设用户想弄清楚为什么一些债务没有偿还，首先他猜想这些人的收入低而且许多债务是高风险的，然后对该假设进行验证。OLAP提供一个描述收入、负债和无力偿还债务之间的经验关系的图形表示（称为多维超立方体），对该图的分析能够证实他的假设。

因此，OLAP也可以用来抽取对商业数据库有用的信息。与数据挖掘不同，假设要由用户提出并能从数据中发现。此外，推断过程是纯计算过程，没有使用模型工具或统计学提供的结论。对于数据库，OLAP可以使用较少的变量来提供有用信息，但如果要处理成十上百的变量时，就会出现问题。找到一个好的假设并通过分析数据库来证实或否认它，OLAP工具日益变得困难和费时。

OLAP不是数据挖掘的替代物，这两个技术是互补的，一起使用会产生很好的效果。OLAP可以用在数据挖掘的预处理阶段。由于可以将重点放在最重要的数据、识别特例及寻找主要内部联系上，所以OLAP使得对数据的理解变得很容易。用特定概要变量表示的数据挖掘的最后结果可以方便地用OLAP超立方体描述。

我们可以把用于从数据库中推断知识的商业智能工具的发展过程总结成如下的简单序列：

查询和报告 → 数据检索 → OLAP → 数据挖掘

查询和报告有最小的信息容量，而数据挖掘的信息容量最大；前者最容易实现，而后者最难实现。这就需要在信息量和实现的难易程度上有一个折中。对工具的选择还必须考虑到特殊的商业需求和公司信息系统的特性。信息缺乏是数据挖掘的最大障碍之一，由

于建立的数据库常常与数据挖掘无关，所以重要的信息可能会丢失。不准确的信息同样是个问题。

数据仓库的建立可以解决上述大部分问题，数据在数据仓库中的有效组织与高效可扩展的数据挖掘方法相结合，使数据能正确、有效地用于公司决策。

1.1.2 数据挖掘和统计学

统计学一直为数据分析提供方法。统计的方法和机器学习方法的区别在于，统计方法的发展不仅和分析数据有关，而且还要依照一个概念参考规范。虽然这样的统计方法一致而严格，却限制了它们快速适应在信息技术和机器学习应用中产生的新方法。统计学家现在对数据挖掘产生了很大兴趣，这将对数据挖掘的发展大有帮助。

很长时间以来统计学家把数据挖掘看做是“数据垂钓”、“数据打捞”或“数据调查”的同义语，所有这些都是对数据挖掘含义的片面理解，这主要来自两方面的批评。首先，数据挖掘没有惟一的理论上的参考模型，而是许多模型相互竞争，模型的选择依据所要检验的数据，而且不管数据多么复杂，总可以找到一个模型能很好地适应数据；其次，巨量的可用数据可能导致找到在数据中根本不存在的关系。

虽然这些批评值得考虑，但我们应当注意到现代数据挖掘技术对泛化能力给予了很大关注，这意味着当选择一个模型时，要考虑其预测性能，复杂的模型将受到惩罚。不能忽视这样的事实，很多重要的发现是事先未知的且不能用做研究假设，特别是对于大型数据库。

最后一个可以把数据挖掘和统计学方法区分的特征是，当统计学传统上只关注为检查特定假设收集的主要分析数据时，数据挖掘还可以分析为其他目标收集的次要数据。例如，当分析存放在数据仓库中的公司数据时，数据挖掘可进行多方面的分析。此外，统计数据可以是试验数据，但在数据挖掘中使用的数据是典型的观测数据。

Berry 和 Linoff（1997）给出的两种数据挖掘的分析方法，区分了自顶向下方法（确认的）和自底向上方法（探索的）。自顶向下分析的目的在于承认或拒绝假设，扩展我们只是部分理解的知识，主要依靠传统的统计方法达到这一目标。自底向上分析在于用户寻找先前没有注意到的信息，在数据中寻找把它们联系起来建立假设的方法。自底向上分析是典型的数据挖掘方法。现实中这两种方法是互补的。实际上，自底向上分析中得到的信息可以指出重要的关系和趋势，但不能解释为什么这些发现是有用的，在多大程度上是有效的。自顶向下分析是确定性工具，可以用来确定发现的结果，并可基于这些结果来评价决策的质量。

至少有三个方面的因素可以把统计数据分析和数据挖掘区别开。首先，数据挖掘分析海量数据，这对统计学分析提出了新的问题。在许多应用中由于计算效率的问题，不可能分析甚至访问整个数据库，对数据库中的数据进行抽样成为必需。抽样必须考虑数据挖掘的目标，所以不能使用传统的统计学理论。第二，许多数据库都是不适合统计学分析需要

的标准形式，如从互联网上得到的数据。这就需要从统计学领域之外寻找合适的分析方法。第三，数据挖掘的结果必须有持续性，这意味着必须不断关注由数据分析模型得到的商业结果。

总之，从统计学的角度来看，数据挖掘并没有什么新内容，但是由于它们具有各自的特性，仍有理由认为统计学方法应该能够研究和形式化数据挖掘中用到的方法，这意味着一方面要从统计学和实用的观点来看待数据挖掘提出的问题，另一方面要发展概念化的规范，以便由统计学家把数据挖掘引导到一个广泛且一致的分析框架中。

1.2 数据挖掘过程

数据挖掘是从明确目标到评价结果的一系列活动，包括七个阶段：

1. 明确数据分析的目标。
2. 对数据进行选择、组织和预处理。
3. 探索性分析数据及转换。
4. 确定在分析阶段使用的统计方法。
5. 用选定的方法分析数据。
6. 评价和比较使用的方法，选择最后的分析模型。
7. 解释最终模型和它在决策过程中的应用。

1.2.1 过程的七个阶段

明确目标

明确目标就是定义分析的目的。要弄清所分析的现象并不总是容易的。实际上，公司的目标通常是清楚的，但是潜在的问题很难转化为分析需要的具体目标。对问题和目标的明确描述是正确建立分析的先决条件。这显然是最困难的部分，因为此时确定的目标决定随后的方法如何组织，因此必须明确无疑。

组织数据

当目标确定后就该为分析选择数据了。首先要确定数据源，通常使用容易获得且可靠的内部资源，这种数据还有一个优点，就是它产生于公司自己的经验和经历。公司的数据仓库是理想的数据源，存储所有不会改变的历史数据，从中我们可以容易地抽取主题数据库、数据集市或感兴趣的数据。如果没有数据仓库，则可以组合该公司的不同数据源创建数据集市。

一般地，数据集市的建立为随后的数据分析提供基本的输入，通常用列表形式表示，称为数据矩阵，这基于分析的需求和前面所建立的目标。当一个数据矩阵建立后，要进行初

步的数据清洗，换句话说，就是对数据进行质量控制。这是一个用于找出现存的不适于分析的变量的处理过程，同时起到检查变量内容、判断是否丢失数据或有不正确数据的作用。如果有重要信息丢失，则需要检查这个阶段并找到根源。

最后，对数据子集或抽样的分析也是有用的。因为对于整个数据集市进行完全分析得到的信息质量并不总是比研究抽样得到的更好。实际上，数据挖掘面对的数据库都是很大的，使用抽样可以节省分析时间。可以对比抽样之外的数据来检验模型的正确性，还可以减小统计模型因受噪声影响失去泛化和预测能力的危险。

数据的探索性分析

对数据的初步探索分析与OLAP技术非常相似。初期对数据重要性的评价有助于原始变量的转换、更好地理解现象或者导出基于满足特定初始假设的统计模型。探索分析可以找到反常数据——与其他项不同的数据项，这些数据项不能删除，因为它们可能含有对于达到分析目标来说很重要的信息。对数据的探索性分析是必须的，它可以使分析者预测哪一种统计方法最适合下一阶段的分析，而这一选择必须考虑在前一阶段得到的数据的质量。探索分析可能会因为收集到的数据不够充分而要求重新抽取。用于数据挖掘的主要探索方法将在第3章讨论。

确定统计方法

有多种统计方法可以使用，也有许多算法，所以对于现存方法的分类是非常重要的。方法的选择依赖于所研究的问题或数据变量的类型。数据挖掘过程由应用引导，因此方法可以根据分析的目的分类。可以区分出三大类方法：

- **描述性方法：**旨在更加简要地描述数据类，它们也称为对称的、无监督的或间接的方法。观测数据被划分为若干未知的类（聚类分析法、Kohonen图法），变量可能根据未知的联系互相关联（关联方法、对数线性模型、图模型）。在这种方式中，变量在同一层次处理，没有因果关系的假设。第4章和第5章将给出这些方法的示例。
- **预测性方法：**旨在描述一个或多个同其他所有变量有关系的变量，它们也称为不对称的、有监督的或直接的方法。这需要寻找分类规则或基于数据的预测，这些规则帮助我们预测或分类未来的与解释变量或输入变量有关的一个或多个应答变量或目标变量的结果。这种类型的主要方法是从机器学习领域发展起来的，例如神经网络（多层感知器）和决策树；也有经典的统计模型，如线性和对数回归模型。第4章和第5章将阐述这些方法。
- **局部方法：**旨在识别数据库中子集的特征，描述性方法和预测性方法都是全局的而非局部的。局部方法的例子有分析相关数据的关联规则、识别反常值（异常值）等，这些都将在第4章介绍。

从功能的观点看，这个分类是彻底的，进一步的区别将在有关文献中讨论。每一个方法可以单独使用，也在多阶段分析中应用于其中一个阶段。

数据分析

当统计方法确定后必须转化成可以计算的适当算法，以便从数据库中合成出需要的结果。市场上有大量为数据挖掘所设计的专业或非专业软件。对于大多数的标准应用而言，我们并不需要开发特定算法，这些软件提供的算法已经足够了。然而，控制数据挖掘过程的算法必须对不同的方法和软件方案有彻底的理解，这样它们才能适应公司的特殊需求，在获得决策时能正确地解释结果。

对统计方法的评价

从可用的统计模型中找到最好的数据分析模型，对于产生最终决策是必需的，因此模型和决策规则的选择依赖于使用不同方法得到的结果之间的比较，这是对于将要用于待处理数据的特定统计方法正确性的重要诊断检测。可能没有一个方法能够满足目标集，这时要为分析重新寻找更适合的新方法。

在评价一个特定方法的性能及对统计类进行诊断测量时，必须考虑其他因素，例如时间限制、资源限制、数据质量和数据的有效性。在数据挖掘中仅使用一种统计方法分析数据是不可取的，不同的方法强调不同的方面，否则某一方面有可能被忽略。

选择最好的最终模型，需要快速简单地应用和比较不同方法，比较产生的结果，然后对得到的不同规则给予商业评价。

方法的实现

数据挖掘不仅对数据进行分析，还要把结果集成到公司决策过程中。商业知识、抽取规则和它们在决策过程中的参与，使我们从分析阶段转移到决策引擎的产生阶段。一旦模型被选择并在数据集上测试，分类规则可被用于整个空间。例如可以预测哪些客户更有可能使我们获利，可以对不同的消费群体使用不同的商业策略，从而增加公司利润。

数据挖掘有这么多好处，因此正确地实现其过程，发挥它的全部潜能是很重要的。公司内部的数据挖掘内容必须渐进地开展，建立现实的目标，一直观察结果的推进，最终目标是把数据挖掘与其他活动完全集成，以支持公司决策。

集成过程可以分为4个阶段：

- **策略阶段**：这个阶段需要研究用到的商业过程，以便辨别数据挖掘可以带来哪些益处，此阶段得到的结果将确定数据挖掘工程的商业目标和评价准则。
- **训练阶段**：在此阶段更加细致地评价数据挖掘活动，建立一个指导性计划，用上一阶段确定的目标和准则评价结果。计划的选择是一个基础因素，必须简单易用，且必须对产生利益足够重要。如果这个指导性计划是积极的，则有两种可能的结果：对不同数据挖掘技术有用性的初步评价和对一个数据挖掘系统原型的定义。