SCIENCE & CULTURE
科文图书

全球顶尖商学院MBA课程精华

高管商学院

数据挖掘

[美] 迈克尔·贝里(Michael J. A. Berry)
戈登·利诺夫(Gordon S. Linoff) 著
中国人民大学副校长 袁卫 等译

MASTERING DATA MINING

针对企业高管面临的棘手问题
提供简洁实用的MBA解决方案

CEO、CFO、CTO、CIO……必备的
案头管理工具书

中国劳动社会保障出版社

全球顶尖商学院MBA课程精华

高管商学院

数据挖掘

MASTERING DATA MINING

[美] 迈克尔·贝里(Michael J. A. Berry)
戈登·利诺夫(Gordon S. Linoff) 著

中国人民大学副校长 袁卫 等译

中国劳动社会保障出版社

著作权合同登记号：图字 01－2003－4151 号
图书在版编目（CIP）数据

数据挖掘/（美）贝里（Berry，M.J.A.），（美）利诺夫（Linoff，G.S.）著；袁卫等译．—北京：中国劳动社会保障出版社，2004.10
高管商学院
书名原文：Mastering Data Mining
ISBN 7-5045-4709-3

Ⅰ.数… Ⅱ.①贝… ②利… ③袁… Ⅲ.信息技术－应用－企业管理：供销管理 Ⅳ.F274

中国版本图书馆 CIP 数据核字（2004）第 095397 号

中国劳动社会保障出版社出版发行
（北京市惠新东街 1 号 邮政编码：100029）
出版人：张梦欣

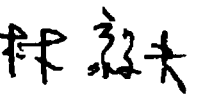
*

北京科文剑桥图书有限公司承销
（北京安定门外大街 208 号三利大厦 邮政编码：100011）
购书热线：010－64203023

*

北京民族印刷厂印刷装订 新华书店经销
680 毫米×980 毫米 16 开本 27.25 印张 454 千字
2004 年 10 月第 1 版 2004 年 10 月第 1 次印刷
定价：48.00 元
读者服务部电话：010－64929211 发行部电话：010－64911190
出版社网址：http：//www.class.com.cn

林毅夫推荐序



我对“高管商学院”这套丛书的出版感到十分高兴。改革开放以来，市场的开放和不断增加的经济自由度已使我国疾步驶入经济发展的快车道。大好的市场机会和不断加深的国内、国外市场的竞争，对企业高层管理人员的战略把握能力提出了越来越高的要求。许多高层经理人员实战经验丰富，但是理论知识相对薄弱，如果能在实践基础上学习系统的工商管理知识，他们将具有更高的技能、思路和眼光，将可以提高决策质量以迎接挑战，把握我国改革开放给企业发展带来的机遇。“终身学习和发展”已成为越来越多的企业家和高管人员的共识！

这套丛书既不是工商管理的入门读物，也非艰深的理论文献。全套丛书的理论体系完整，内容新颖实用，以清新的表达方式和大量的信息介绍了企业管理的方方面面。每本书的作者或在著名大学的工商管理学院任教，或从事着与实践密切相关的培训工作。多位作者还担任管理方面重要学术期刊的编委，《税收与企业战略》一书的作者更是1997年的诺贝尔经济学奖的获得者。而且，最难能可贵的是，这些作者能够随着管理理论的不断发展而推陈出新，有的书在国外已经修订过十几次，这套“高管商学院”丛书呈献的则都是最新版本。

这套丛书既包括了企业高层管理者应该具备的基础经济学知识，又涵盖了企业具体管理的核心内容。比如，《经济学》一书以简明易懂的语言，介绍了经济学的基本原理，使读者具备准确地判断经济事件的能力，并对经济学产生持久的兴趣。《管理思潮》是近年来管理思想革新的简洁实用的总结，书中对各种新的管理理念作了详尽的介绍，并对如何将之运用到复杂的日常组织中作了深入浅出的阐述。在企业核心运作过程中可能遇到的困难，也可以在本套丛书中找到答案。比如，《财务管理》全面介绍了财务管理的各项指标和操作技术，被誉为“财务小圣经”。丛书中对全新的管理方法也有系统介绍，比如数据挖掘就是信息领域发展最快的技术。通过《数据挖掘》一书，读者不仅可以掌握

数据挖掘的技术和理论，更重要的是还可以了解数据挖掘在各个业务领域的可能应用方式，从而为今后从事实际的数据挖掘工作提供参考。

总之，这是一套非常适合高管人员阅读的丛书，能够帮助高管人员成为具有国际视野和战略眼光，既能把握宏观走向、社会脉搏，又熟悉企业微观操作的新时代企业界的领军人物。

2004 年 10 月

序 言

迈克尔和我的第一本书《市场营销、销售及客户服务的数据挖掘技术》(Data Mining Techniques for Marketing, Sales, and Customer Support, John Wiley & Sons 出版,1997,以下简称《数据挖掘技术》)发行以来,读者反应强烈、好评如潮,令我们深受感动。我们编写那本书的初衷是对数据挖掘作个综合且又易于理解的介绍,看来我们的目的是达到了。

自从那本《数据挖掘技术》问世后,世界发生了很大变化。我们办了自己的公司——数据挖掘者(Data Miners),专门致力于数据挖掘工作。数据挖掘者的使命是像强调模型的结果一样强调对方法的理解,像重视数据挖掘技术一样重视数据挖掘的过程。

对于毫无经验的创业者来说,放着定期领取工资的稳定工作不做,而去独立闯荡肯定是一件既刺激又冒险的行动。但它确实给我们提供了接触企业并向实践学习的机会,使我们学会了如何面对客户、如何分配资金以及如何选择客户等。它还为我们创造了与数据挖掘主要客户以及该领域顶尖人物合作的机会。

我们的朋友和家人不止一次地问过我们,为什么还要花费那么多时间写第二本书?简单地说就是《数据挖掘——客户关系管理的科学与艺术》一书需要写。数据挖掘的领域最近几年发展得非常迅速,商界和技术领域的实践需要我们写,希望这本书能够满足读者的要求。

要了解数据挖掘市场发展得有多快,我们只需看一看商业专栏的记载就够了。在本书写作的过程中,我们见证了由于数据挖掘在客户关系管理和电子商务领域的杰出作用使得大公司发生并购或者扩展的许多范例。

■ 世界最大的私人控股软件公司——SAS 公司,推出了公司历史上最成功的产品——数据挖掘软件包。我们习惯开玩笑地说:"在 SAS 系统中,统计就是一切。那么数据挖掘就只能是这一切之外的了。"若按照

这一定义，数据挖掘就没有存在的必要了！

■ 在分析软件市场上，SAS 最强的竞争对手 SPSS 通过开发研制国际领先的数据挖掘软件，使得其竞争力得到提高，而在几年前 SPSS 还无力与 SAS 竞争。

■ DoubleClick，这一互联网最大的广告销售商购买了 Abacus，一个善于用目录数据建立预测模型的公司。

种种迹象表明，数据挖掘已经成为各种问题的解决方法之一。如果做得好的话，数据挖掘确实可以解决许多难题，但这需要理解数据挖掘的整个过程。这一理解来自于过去所做项目成功与失败两个方面的经验和教训。过去的每一个项目都积累了宝贵的经验。在这本书中，我们将带领你们去重温这些项目，以分享我们的经验。

《数据挖掘——客户关系管理的科学与艺术》一书将对数据挖掘作全面的介绍。第一部分从商业领域开始。本部分的四章将回答如下的问题：数据挖掘为什么重要？数据挖掘有哪些诀窍？第 4 章将特别介绍客户及客户关系管理。尽管数据挖掘在许多领域都有成功的应用，但在客户关系管理领域引起了人们格外的兴趣和重视。

第二部分从技术层面上介绍数据挖掘。第 5 章回顾数据挖掘技术（这些技术和方法在我们的第一本书里有详细的介绍），第 6 章讨论数据，接下来的第 7 章研究如何建立好模型。这一章十分重要，因为它包含了我们多年的经验和教训。

第三部分是本书最长的部分，也是最重要的部分。这部分是数据挖掘的案例研究。尽管这些案例都是商业领域的，但涉及的范围还是很广的，从对数十万兆字节数据的探索（hundreds of giga - bytes of data）到为网络银行客户服务预报下一个标题广告，从而改进印刷过程。所有这些案例讨论的都是实际的问题，同时都给出了其中所用的方法、数据、所得到的结果以及经验和教训。

在编写这本书时，我们特别强调数据挖掘的实际应用。我们希望这本书能够帮助读者掌握这门艺术。

译者序

数据挖掘是信息领域发展最快的技术，很多不同领域的专家，比如统计学家、数据库专家等，都从中获得了发展的空间，使得数据挖掘日益成为企业界讨论的热门话题。随着信息技术的发展，人们采集数据的手段日益丰富与高明，由此积累的数据日益膨胀，数据量达到 GB 甚至 TB 级，而且高维数据也日益成为主流。这些海量数据及其高维特征使得传统的数据分析手段相形见绌。计算机性能的日益更新，使得人们能够期望计算机帮助我们分析与理解数据，帮助我们以丰富的数据为基础做出正确决策。于是数据挖掘这一融合多种分析手段，从大量数据中发现有用知识的方法就应运而生，并在使用中得以蓬勃发展。

数据挖掘是一个多学科交叉的领域。一方面，数据挖掘以计算技术的发展为首要条件，没有数据的有效组织，从一堆数据垃圾中发现有用知识是痴人说梦；没有大量计算算法的支持，即使是简单的查询也会耗时巨大，更不用说发现有用模式。另一方面，即使数据得到了有效组织，计算算法足够先进，要想发现海量数据中隐藏的有用信息，还必须综合利用统计学、模式识别、人工智能、机器学习、神经网络等学科的专业知识。比如数据挖掘使用的分析方法，有相当大比重是靠高等统计学中的多元分析来支撑的，一般将之定义为数据挖掘技术的 CART、CHAID 或模糊计算等理论方法，也都是由统计理论发展衍生的。当然，所有这些学科的发展必然会从不同角度关注数据分析技术的进展，数据挖掘也为这些学科的发展提供了新的机遇与挑战。

正因为数据挖掘的强大功能，它引起了学术界与实务界的广泛关注，吸引了众多研究与开发人员。目前国内外很多大学都开设有数据挖掘的课程，而且出版发行了若干数据挖掘方面的专业书籍。然而，长期以来，数据挖掘方面的专著往往侧重于机器学习领域，对案例的讲解所作的研究不够，而且多数专著过于理论化，与数据挖掘的本来目的——方便企业

末端使用者的使用这一目标相违背，因此，选择一本合适的数据挖掘教材困难较大。

《数据挖掘——客户关系管理的科学与艺术》一书的英文原著，一直被作为我们中国人民大学统计学系数据挖掘中心数据挖掘领域理论与应用研究的经典教材。在长期的探索和研究中，我们体会到从案例中学习是快速领会数据挖掘概念的最佳途径。本书通过丰富的来自于实际的案例以及大量宝贵的数据分析经验，阐述了数据挖掘的本质。它是掌握数据挖掘理论的宝典。事实上，这本书已得到了世界范围内的广泛认可。从著名的网上书店——亚马逊书店关于数据挖掘书籍的销售统计看，本书英文版长期雄踞销量第一的宝座。本书的主要特点在于其清晰的理论架构、系统且易于理解的理论解说及丰富的实际案例。通过对本书的学习，读者不仅可以从中掌握数据挖掘的核心技术和理念等，而且更主要的是可以了解数据挖掘在各个业务领域的应用，从而为今后从事实际的数据挖掘工作提供攻玉之石。

本书分为三个部分共 14 章。全书的翻译由袁卫和王星组织完成。参加本书初稿翻译的成员 包括：袁卫（前言，第 1 章）、王星（第 2、3、4、5、6 章）、伍业锋（第 7、8 章）、匡红波（第 9、10、11、12 章）、戴稳胜（第 13、14 章）。张尧庭教授和谢邦昌教授从学术角度对本书进行了全面审校。

由于本书涉及领域较宽，知识面广，许多领域译者并不熟悉而且有少量的专业术语目前并没有固定的译法，翻译难度很大，因此，译文中的错误和不妥之处在所难免。我们欢迎读者及数据挖掘爱好者给我们提出批评与建议。

我们的联系方式是：

地址：中国人民大学科研楼 A 座 901

电话：86 - 10 - 62511709，13683215697，13693043707

网站：http://www.dmchina.org

E - mail：dmcstat@sina.com

中国人民大学副校长
中国统计学会副会长　袁卫

2003 年 7 月

目 录

第一部分 本书焦点

第二部分 数据挖掘的三大支柱