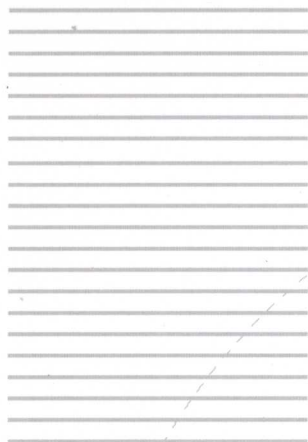




光华管理学院
Guanghua School of Management



张俊妮 ● 编著

北京大学光华管理学院教材

Business Statistics

商务统计系列

数据挖掘与应用

DATA MINING
AND ITS
APPLICATION



北京大学出版社
PEKING UNIVERSITY PRESS

商务统计系列



数据挖掘与应用



北京大学出版社
PEKING UNIVERSITY PRESS

图书在版编目(CIP)数据

数据挖掘与应用/张俊妮编著. —北京:北京大学出版社, 2009. 6
(北京大学光华管理学院教材·商务统计系列)
ISBN 978-7-301-15239-3

I. 数… II. 张… III. 数据采集-高等学校-教材 IV. TP274

中国版本图书馆 CIP 数据核字(2009)第 088538 号

书 名: 数据挖掘与应用

著作责任者: 张俊妮 编著

责任编辑: 朱启兵

标准书号: ISBN 978-7-301-15239-3/F·2196

出版发行: 北京大学出版社

地 址: 北京市海淀区成府路 205 号 100871

网 址: <http://www.pup.cn>

电 话: 邮购部 62752015 发行部 62750672 编辑部 62752926

出版部 62754962

电子邮箱: em@pup.pku.edu.cn

印刷者: 河北滦县鑫华书刊印刷厂

经 销 者: 新华书店

730 毫米×980 毫米 16 开本 12.5 印张 220 千字

2009 年 6 月第 1 版 2009 年 6 月第 1 次印刷

印 数: 0001—4000 册

定 价: 22.00 元

未经许可,不得以任何方式复制或抄袭本书之部分或全部内容。

版权所有,侵权必究

举报电话:010-62752024 电子邮箱:fd@pup.pku.edu.cn

丛书总序

教材建设是大学人才培养和知识传授的重要组成部分。对管理教育而言,教材建设尤为重要,一流的商学院不仅要有一流的师资力量、一流的生源、一流的教学管理水平,而且必须使用一流的教科书。一流的管理类教科书必须满足以下标准:第一,能把所在领域的基础知识以全面、系统的方式和与读者友好的语言呈献给读者;第二,必须有时代感,能把学科前沿的研究成果囊括进去;第三,必须做到理论和实务(包括案例分析)相结合,有很强的实用性;第四,能够启发学生思考现实的管理问题,培养他们分析问题和解决问题的能力;第五,可以作为研究人员和管理人士的工具书。

中国的管理教育是伴随改革开放而产生的。真正意义上的管理教育在中国不过十多年的历史,但巨大的市场需求使得管理教育成为中国高等教育各学科中发展最快的领域,管理类教科书市场异常繁荣。但总体而言,目前国内市场上管理类教科书的水平仍不能令人满意。国内教科书作者大多数在所涉及领域并没有真正的原创性研究和学术贡献,所撰写的教科书普遍停留在对国外教科书的内容进行中国式排列组合的水平上;国外引进的原版教科书虽然具有学术上的先进性,但由于其写作背景是外国的管理实践和制度安排,案例也都是取自于西方发达国家,对中国读者而言,总有一种隔靴搔痒的感觉。如何写出一流的中国版的管理类教材,是中国管理教育发展面临的重要任务。

北京大学光华管理学院一直重视教材建设工作。1999年夏,我们曾与经济科学出版社签约,以每本20万元的稿酬,向全国征集MBA教科书作者。这个计划公布之后,我们收到了十几本教科书的写作方案。遗憾的是,经专家委员会评审,没有一本可以达到我们所期望的水平。究其原因,主要

是当时中国管理学院的教授、学者大多数并没有真正从事有关中国商业实践、管理实践的理论性和实证性研究。我们得出的结论是:没有一流的学者,没有一流的学术研究成果,就不可能写出一流的教科书。国外有大量优秀的教科书,这些教科书都是成千上万的优秀学者在对每一个具体的管理问题进行出类拔萃的研究的基础上写成的,是学术研究的结晶。国内学者如果没有研究的积累,要写出包含中国管理实践的好的教科书是不可能的。所以,我们果断地中断了这个计划。

自1999年以来,师资队伍的建设成为光华管理学院工作的重中之重,除了通过出国培训、合作研究等方式提升原有教师的水平外,我们还从国内外引进了六十多位优秀的新教师,使得光华管理学院成为真正与国际接轨的研究型商学院。我们的绝大多数教师不仅受过良好的科学研究训练,具有很好的理论素养,而且潜心于中国管理实践的研究和教学。不少教师已在国际一流的学术刊物上发表论文,受到国际同行的关注。

今天,我们有充分的信心向社会呈献一套由光华管理学院教师撰写的优秀的管理类系列教科书。本系列教材包括“应用经济学”、“金融学”、“会计学”、“市场营销学”、“战略管理”、“组织管理”、“管理科学与工程”和“信息系统管理”等子系列,涵盖管理学院教学的各个层面。这些教科书都是作者在光华管理学院多次授课的讲义的基础上反复修改写成的,经受过课堂实践的考验。

当然,我们深知,优秀教材的建设是一项长期的任务,不可能一蹴而就,也不是一个学院所能独立完成的,需要所有管理学院的通力合作。我们欢迎兄弟院校的师生和广大读者对教材提出批评和建议,以便我们不断改进。

让众多的没有机会进入光华管理学院的读者分享我们的课堂内容,是我们的荣幸,更是我们的责任。我们将以开放的姿态和与时俱进的精神为管理学教育的发展而努力,将更多更高水平的教材奉献给广大读者!

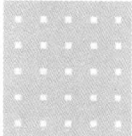
张维迎

北京大学光华管理学院

2005年7月11日



序



近年来,数据挖掘在诸多领域得到了广泛应用。不论是海关、税务等政府部门,还是金融、电信、医疗等行业,将数据管理起来、活用起来,都能获取巨大的直接回报和间接效益。因此,数据挖掘的知识值得进行广泛传播。

2004年春季我第一次开设“数据挖掘与应用”课程时,搜寻了关于数据挖掘的各种专著和教材,却找不到一本我理想中的教材。在我看来,有些书介绍了数据挖掘的某些应用,却没有深入探讨背后的方法;而另一些书又过于技术化,讲了很多方法的细节,却没有和应用有机地结合起来。我面对的学生有一定量化背景,但理论上不要求特别深。我希望给他们讲授实际能够运用的数据挖掘的体系、对各种数据挖掘方法的直觉理解以及它们的有机应用;同时,我又希望保持一定的严谨性,为学生对这些方法的理解和运用提供坚实的基础。此外,我还希望与某种主要软件结合起来讲授,使得学生能够边学边用,加强学习效果。由于一些机缘,当时选择了 SAS 软件。

经过数年的摸索、与学生的互动以及不断改进,终于完成了这本书,满足了我所设想的多重目的,在课堂上试用时也达到较好的教学效果。我希望这本书能为更多老师和学生服务,也能为更多对数据挖掘感兴趣的人服务。

在此,我要感谢所有选修了北京大学光华管理学院“数据挖掘与应用”课程的同学,他们的学习热情和宝贵的反馈意见使本书得以逐渐成型。我还要特别感谢孙鹏、李娜和魏巍同学,他们帮助我进行了重要的资料整理工作;也感谢罗荣华同学帮助我校对了部分书稿。

谨以此书献给我的父母,他们给了我生命,也不断地给予我关怀和鼓励。

张俊妮

2009年4月7日

于北大燕园

CONTENTS

目录

第一章 数据挖掘概述 /1

- 1.1 什么是数据挖掘 /2
- 1.2 数据挖掘的应用 /2
- 1.3 数据挖掘方法论 /8

第二章 数据理解和数据准备 /12

- 2.1 数据理解 /14
- 2.2 数据准备 /16
- 2.3 使用 SAS 进行数据理解和数据准备：
FNBA 信用卡数据 /29

第三章 关联规则挖掘 /39

- 3.1 关联规则的实际意义 /40
- 3.2 关联规则的基本概念及 Apriori
算法 /41
- 3.3 负关联规则 /45
- 3.4 序列关联规则 /47
- 3.5 使用 SAS 进行关联规则挖掘 /48

第四章 多元统计中的降维方法 /52

- 4.1 主成分分析 /53

CONTENTS

目 录

4.2 探索性因子分析 /59

4.3 多维标度分析 /65

第五章 聚类分析 /70

5.1 距离与相似度的度量 /72

5.2 k 均值聚类法 /76

5.3 层次聚类法 /79

第六章 预测性建模的一些基本方法 /85

6.1 判别分析 /86

6.2 朴素贝叶斯分类算法 /88

6.3 k 近邻法 /89

6.4 线性模型与广义线性模型 /91

第七章 神经网络 /99

7.1 神经网络架构及基本组成 /100

7.2 误差函数 /106

7.3 神经网络训练算法 /110

7.4 提高神经网络模型的可推广性 /114

7.5 数据预处理 /116

7.6 使用 SAS 建立神经网络模型 /117

7.7 自组织图 /129

CONTENTS

目录

第八章 决策树 /136

- 8.1 决策树简介 /137
- 8.2 决策树的生长与修剪 /139
- 8.3 对缺失数据的处理 /145
- 8.4 变量选择 /146
- 8.5 决策树的优缺点 /147

第九章 模型评估 /158

- 9.1 因变量为二分变量的情形 /159
- 9.2 因变量为多分变量的情形 /169
- 9.3 因变量为连续变量的情形 /170
- 9.4 使用 SAS 评估模型 /171

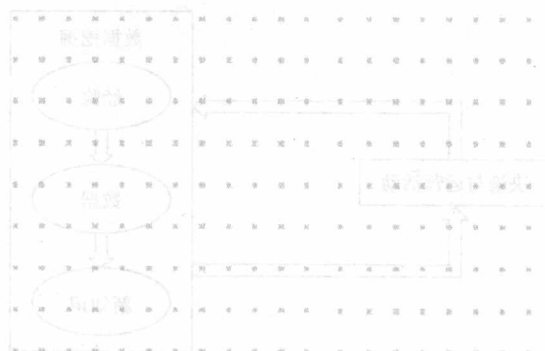
第十章 模型组合与两阶段模型 /175

- 10.1 模型组合 /176
- 10.2 随机森林 /180
- 10.3 两阶段模型 /182

参考文献 /184

第一章 数据挖掘概述

1



■ 什么是数据挖掘

■ 数据挖掘的应用

■ 数据挖掘方法论

1.1 什么是数据挖掘

任何一个组织(政府部门、企业、学校等)在决策与运营活动中都会积累丰富的经验,同时也面临着在不断变化的环境下做出快速而正确决策的挑战。数据挖掘方法首先把组织所积累的经验转换为可度量的数据,对其进行分析后,提炼出对运营管理有指导意义的新知识,进一步改进决策、改善运营活动(见图1.1)。这是一个持续改进的过程,决策经营活动不断积累新的经验、新的数据,使用数据挖掘方法分析新的数据后不断产生新的知识,不断地促进决策与经营活动的改进和完善。Berry 和 Linoff(2000)将数据挖掘定义为对大量数据进行探索和分析、以便发现有意义的模式和规则的过程。

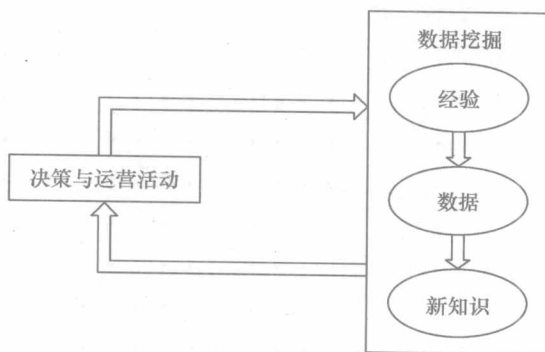


图 1.1 数据挖掘与决策和运营活动

1.2 数据挖掘的应用

数据挖掘在很多领域都得到了广泛应用,以客户为导向的应用有市场篮分析、获取客户、客户细分、客户保持、交叉销售、向上销售、客户终身价值分析等,以运营为导向的应用有盈利分析、定价、欺诈发现、风险评估、雇员流失分析、生产效率分析等。下面我们从几个具体的案例来看数据挖掘的应用。

案例一：银行业 [摘自 Kudyba(2004)]

[应用一] 企业贷款信用风险

在给企业贷款时,银行不可避免地面临着信用风险,这种风险可以通过两类指标来刻画:一是企业贷款违约的概率,二是一旦企业违约所带来的损失。如果能够很好地预测信用风险,银行能够根据信用风险的大小,基于自身的风险偏好选择客户群体,为不同的客户提供不同的贷款产品或不同的贷款利率。

对违约事件进行预测可能存在两类错误,第一类错误将实际会违约的企业判断为不违约者,这会产生大量的信用损失(贷款的本金、利息等);第二类错误将实际不会违约的企业判断为违约者,这会导致银行失去潜在的业务和盈利机会。最大限度地减少这两类错误,将会为银行带来可观的收益。

在 20 世纪 90 年代早期美国经历经济衰退之前,大多数美国银行及穆迪、标准普尔等风险分析仲裁机构的决策依赖于信贷人员、信用调查分析人员等的专家意见,很少使用基于统计方法的风险分析。在这次危机之后,美国银行开始重视如何更加一致地诠释并管理风险。他们的解决办法是使用数据仓库和数据挖掘技术,对大量数据进行收集、存储和维护,应用高级建模方法对信用风险进行建模,并对所使用的模型进行经常性的监测和修正。巴塞尔协议就特别强调银行内部的信用风险管理,因此,很多银行都使用内部的历史数据和现代统计技术,建立了内部评级模型。

以花旗银行为例。它收集了反映企业财务状况的年度财务报表、企业所处行业的总体情况、企业的市场地位、企业管理质量、企业管理层的风险偏好、审计报告的质量、企业开业时间、企业作为花旗银行客户的时间等信息。在收集这些信息时,重点关注违约行为的确认,以及有违约行为的企业在违约之前的各种信息。

数据中存在的自相矛盾和错误会导致任何建模努力付诸东流,所以花旗银行对所收集的数据进行了大量的清理工作。因为数据发生的频率比较低,例如财务报表多为年度数据,所以每一个数据点都是有价值的,需要尽一切努力保证数据的准确性,并尽可能少地丢弃数据。数据清理需要自动过程与手动过程的有机结合,例如,需要详细检查资产负债表是否平衡、违约日期能否确定等。根据这些数据,花旗银行建立了分地区、分行业的一系列模型来预测风险类别,每一个风险类别都与一定范围的违约概率相联系。

模型建立后,还需要验证模型的预测是否准确。对信用风险模型最重要的验证是通过收集实际违约企业的数据实现的,需要查验模型是否将实际信用水

平低的企业归入风险比较高的类别,模型预测为高风险的企业中是否实际违约率更高。随着时间的推移,由于行业环境变化等因素,模型验证时会发现现有信用风险模型的性能逐步下降,所以还需要及时对模型进行更新。这时,可以将新的企业的数据加入建模数据集,同时将时间过长的数据从建模数据集中去除,根据新的数据集更新模型。

[应用二] 信用卡

信用卡行业中有大量的数据。持有信用卡的客户每次购买都会留下购买时间、金额、商店的类型、地点等信息;银行有关于客户其他银行账户活动的各种信息;银行还有可能获取来自外部的另外一些信息,如客户人口信息、生活方式等。数据仓库和数据挖掘对信用卡行业而言至关重要,数据分析和营销活动不间断地在各方面相互影响。

- **客户拓展。**目标群体既可以是目前还未拥有该银行信用卡的潜在客户,也可以是现有客户(提供第二张或第三张信用卡)。统计模型能够预测目标群体中接受概率比较高的人,从而能够帮助策划成功的客户拓展活动。

- **客户保留。**保留现有客户是最廉价的拓展方式,当市场趋于饱和时,这种方式尤其凸显其重要性。数据挖掘方法能够分析曾经注销过信用卡的客户的特征,在下次客户拓展活动中,可以事先确认具有类似特征的客户,针对他们设计特定的项目来保留他们。

- **交叉销售。**将产品出售给现有客户要比出售给新客户更有可能盈利,因为关于现有客户已经收集到高质量的信息,而且现有客户已经展现了一定的忠诚度。交叉销售可能涉及银行的各类金融产品,银行也有可能销售其他金融机构(如保险公司)或非金融机构(如旅行社)的产品。数据挖掘方法能够帮助分析应该通过哪些销售方式联系哪些现有客户,从而产生丰厚的交叉销售的利润。交叉销售活动及其结果又能给消费者数据库增添更多的信息,为下一次营销活动提供更有价值的帮助。

- **增加长期价值。**可以将关于一个客户的所有账户(存储账户、投资账户、信用卡账户等)的数据整合起来,建立统计模型来测算客户的长期价值。

- **违约管理。**管理信用卡客户群体的风险就像管理企业贷款的风险一样重要。通过收集和全面分析违约的信用卡账户,建立统计模型,能够比较准确地预测违约行为的发生。

- **发现欺诈行为。**信用卡被盗用的威胁越来越严重。银行可以使用各种统计方法及时发现不正常的购买模式,并立即采取行动。

案例二：中国海关[摘自易万达(2006)]

2004 年我国进出口贸易突破 1 万亿元人民币,成为世界第三大对外贸易国;2005 年我国进出口贸易总额突破 1.4 万亿元人民币。海关作为对外贸易的直接窗口,是连接国内外市场的桥梁,其重要作用也日益凸现。同时,进出口环节的违规和走私活动更加频繁,海关面临的形势更加复杂,所承担的打击走私、征收关税、货物监管的任务也更加艰巨。海关执法评估系统(Enforcement Assessment System, EAS)利用统计数据增强了海关管理和分析信息的能力。

海关统计是全口径统计,所有进出口货物都需按《海关法》规定如实申报。海关数据有几大来源:

- 原始凭证,包括《中华人民共和国海关进(出)口货物报关单》以及经海关核发的其他申报单证。
- 外贸企业基本情况,包括企业名称、企业资产、行业信息等外部数据。
- 来自银行、国税部门、港口等其他部委和行业部门的信息。外贸企业可能在某一个方面提供虚假信息,但是在整体上伪造信息的可能性非常低。通过掌握全面的数据,海关极大提高了对进出口贸易管理和监督工作的效率。例如,真实的商品信息能够帮助查证外贸企业是否在通关过程中伪报瞒报商品的数量或品名;来自银行的企业资金流变化的信息真实反映该企业真正的运营情况,能够帮助查证该企业是否伪报瞒报商品的价格。
- 国际来源,主要是香港地区、韩国、日本等周边贸易大国或地区的海关数据。

海关总署使用 EAS 系统在宏观、中观、微观等多个层面加强了管理监督能力,下面试举几例进行说明。

[应用一] 宏观:总体税收的预测

海关税收约占中央财政收入的 1/3 左右,海关税收的稳定性对国民经济健康稳定的发展起着至关重要的作用。及时准确地预测和把握海关的总体税收情况,对海关整体计划、组织、管理、控制等工作起着关键的指导作用。

海关对税收的预测公式为:

$$\text{海关总体} \quad \text{应征税商品} \quad \text{应征税商品} \\ \text{税收预测值} = \frac{\text{进口值预测值}}{\text{综合税率预测值}}$$

以 2005 年第一季度后预测 2005 年全年税收为例,海关根据历史数据使用时间序列方法预测 2005 年 4 月至 12 月应征税商品进口值,与前三个月的数值相加

得到 2005 年全年应征税商品进口值的预测值;同时海关使用简单回归方法预测应征税商品综合税率。若国际市场和国内经济稳定、相关外贸政策变动不大,2005 年税收的预测值为 5261 亿元人民币。2005 年海关实际税收额为 5278 亿元人民币,可见预测值非常准确。

[应用二] 中观:直属海关层面的分析

海关总署共有 41 个直属海关。由于区域经济、地理环境、港口特性等因素的影响,进出不同直属海关的商品和数量不尽相同。例如,广州海关、深圳海关进口的商品主要集中于电子类产品和仪器仪表,进出口商品总量不大,但相对税收却可能很高;而青岛海关、宁波海关进口的商品则主要集中于大宗散货类的产品,总量很大,但税收会相对较少。如果强调单一的评价标准,并不能够反映各海关真实的业绩。

以税率为横坐标,各税率区间的商品进口金额百分比为纵坐标作图(图 1.2),可以清晰地看出不同海关的税率结构之间存在差异。根据税率结构对直属海关进行聚类分析,可将直属海关分为 10 个不同的组,在同一组中可对比不同海关绩效的优劣。例如,可使用应征综合税率和实际征收税率的差异作为考核指标,如果同一组中大部分直属海关实际征收税率都与应征综合税率相差不大,但有一个直属海关实际征收税率显著高于应征综合税率,那么这个直属海

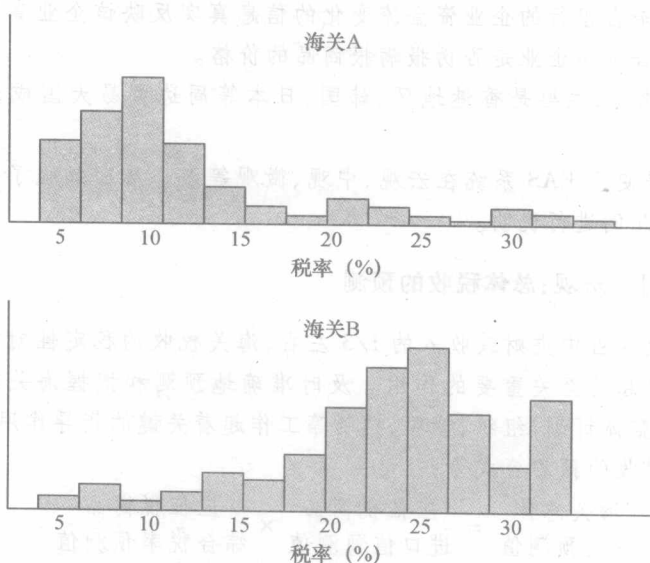


图 1.2 不同海关的税率结构

关的管理绩效比较高。再如,在同一组中比较各海关的通关时间,如果出现差别很大的情况,可以结合通关流程分析,及时发现一些海关在管理上的差距,整体提升海关的管理水平。

【应用三】 微观:进出口货物的异常波动

进出口商品的价格分布看似无规则,但如果按照来源地、尺寸、材料、交易时间等因素分类,可以看出同类商品的价格基本符合正态分布。因此,可以把商品细分到各类,然后把该类中价格处于最低 2.5% 和最高 2.5% 的商品作为重点查验的对象。这项举措对提高查验准确率起了积极的作用。

案例三:意大利信息系统联盟 [摘自 <http://www.sas.com> 所列的成功案例]

在意大利,信息系统联盟(CSI)在信息技术和远程通信领域代表了公共管理机构。CSI 为公众健康服务设计了决策支持系统,该系统回答了下列一些问题:

- 与去年相比,医疗开销增长了多少?哪类病人主要导致了这一增长?
- 在不同地区,各类疾病的医疗开销是怎样分布的?
- 不同年龄段患者的平均开销是怎样变化的?
- 哪类疾病需要患者支付的医疗开销最大?

在意大利,公众健康服务是由本地公众健康服务权威机构来管理的。在皮德蒙特高原就有 22 家独立的本地权威机构。CSI 开发了商业智能系统,将 22 家权威机构的信息集中在一个数据仓库中。该系统能够检查数据并为医生提供图形化的报告,这些报告向医生显示了针对不同性别、不同年龄段等患者的治疗方案的分布,据此医生可以比较自己的治疗方案和其他医生的治疗方案。该系统的另一个优势是让医生知道他们的信息正在被监视。药品公司经常主动向医生推广自己的新药,但药品价格不总是合适的,医生知道他们的信息被监视,就会更加谨慎地开药。

CSI 还能帮助回答诸如世界卫生组织(WHO)建议的高血压治疗方案是否被遵循的问题。CSI 分析了所有患者 6 个月以来的高血压特效药的使用情况。结果表明高血压的治疗占总开销的 30% 以上,这一比例是相当高的。通过对 4 000 名医生的聚类分析,CSI 发现了一组没有遵守 WHO 指导方针的医生。皮德蒙特高原地区决定安排相关的培训活动,向这些医生解释 WHO 的指导方针,找出他们没有遵守这些指导方针的原因。

案例四：零售业 [摘自 <http://www.sas.com> 所列的成功案例]

Staples 是美国一家经营办公用品的连锁零售商店,拥有 1 100 多家分店,年销售额近 110 亿美元。它使用数据挖掘方法成功地管理各家分店。Staples 采用的数据包括历史销售数据、客户(包括商户和家庭)的统计数据、分店所处的地段特征及该地段的竞争水平等。

Staples 每周收集并分析 800 多万笔交易数据,为 1 100 多家分店进行每周和每日的销售预测。这些数据还被用于公司的其他事务,如季度销售预测、用人计划、存货管理、年度预算等。Staples 还使用数据挖掘为新的分店选址,通过预测近 500 家预选店址未来三年的销售额,选择最优的店址。因为关闭一家分店的成本大约在 50 万到 100 万美金之间,通过数据挖掘避免错误的选址决策已经为公司节省了几百万美元。

总结:数据挖掘带来的收益

随着信息技术的发展,各类经营管理活动产生的数据量越来越大,不少组织的数据量以兆字节(Terabyte)为单位计算,数据分析能力往往成为一个组织的核心竞争力。数据挖掘能够帮助满足客户需求、降低风险、最大化收益、简化管理流程、优化资源配置等等。使用数据挖掘技术获取超过 10 倍的直接投资回报司空见惯。数据挖掘还往往带来很多无法直接度量的效益,如信息流动的通畅、管理监督能力的提升等。

1.3 数据挖掘方法论

数据挖掘活动主要分为无监督和有监督两大类。在无监督数据挖掘中,我们对各个变量不区别对待,而是考察它们之间的关系。这类方法有描述和可视化、关联规则分析、聚类分析、主成分分析等。在有监督数据挖掘中,我们希望建立根据一些变量来预测另一些变量的模型,前者被称为自变量,后者被称为因变量。有监督数据挖掘能从数据中获取深度细致的信息,应用非常广泛(如第 1.2 节的案例中针对贷款企业违约率的预测、针对信用卡客户对营销活动的响应情况的预测、零售商店的销售预测等)。我们将要介绍的两种数据挖掘方法论主要针对的是有监督数据挖掘,它们为数据挖掘项目提供了设计和实施的