《数据挖掘技术》

图书基本信息

书名：《数据挖掘技术》

13位ISBN编号：9787302310143

10位ISBN编号：7302310149

出版时间：2013-3

出版社：清华大学出版社

作者：Gordon S.Linoff,Michael J.A. Berry

页数：620

译者：巢文涵,张小明,王芳

《数据挖掘技术》

内容概要

《数据挖掘技术:应用于市场营销、销售与客户关系管理(第3版)》内容简介：谁将是忠实的客户？谁将不是呢？哪些消息对哪些客户细分最有效？如何最大化客户的价值？如何将客户的价值最大化？《数据挖掘技术:应用于市场营销、销售与客户关系管理(第3版)》提供了强大的工具，可以从上述和其他重要商业问题所在的公司数据库中提取它们的答案。自《数据挖掘技术:应用于市场营销、销售与客户关系管理(第3版)》第1版问世以来，数据挖掘已经日益成为现代商业不可缺少的工具。在这个最新版本中，作者对每个章节都进行了大量的更新和修订，并且添加了几个新的章节。《数据挖掘技术:应用于市场营销、销售与客户关系管理(第3版)》保留了早期版本的重点，指导市场分析师、业务经理和数据挖掘专家利用数据挖掘方法和技术来解决重要的商业问题。在不牺牲准确度的前提下，为了简单起见，即使是复杂的主题，作者也进行了简洁明了的介绍，并尽量减少对技术术语或数学公式的使用。每个技术主题都通过案例研究和源自作者经验的真实案例进行说明，每章都包含了针对从业者的宝贵提示。书中介绍的新技术和更为深入的技术包括：线性和逻辑回归模型、增量响应（提升）建模、朴素贝叶斯模型、表查询模型、相似度模型、径向基函数网络、期望值最大化（EM）聚类和群体智慧。新的章节专门讨论了数据准备、派生变量、主成分分析和其他变量减少技术，以及文本挖掘。在建立了全面的数据挖掘应用业务环境，并介绍了所有数据挖掘项目通用的数据挖掘方法论的各个方面之后，《数据挖掘技术:应用于市场营销、销售与客户关系管理(第3版)》详细介绍了每个重要的数据挖掘技术。

《数据挖掘技术》

作者简介

Gordon S.Linoff和Michael J.A.Berry在数据挖掘领域的知名度众所周知。他们是Data Miners公司——一家从事数据挖掘的咨询公司——的创始人，而且他们已经共同撰写了一些在该领域有影响力和得到广泛阅读的书籍。他们共同撰写的第一本书是Data Mining Techniques的第一个版本，于1997年出版。自那时起，他们就一直积极地挖掘各种行业的数据。持续的实践分析工作使得两位作者能够紧跟数据挖掘、预测以及预测分析领域的快速发展。Gordon和Michael严格地独立于供应商。通过其咨询工作，作者接触了所有主要软件供应商（以及一些小的供应商）的数据分析软件。他们相信好的结果不在于是采用专用的还是开源的软件，命令行的还是点击的软件，而是在于创新思维和健全的方法。
Gordon和Michael专注于数据挖掘在营销和客户关系管理方面的应用——例如，为交叉销售和向上销售改进推荐，预测未来的用户级别，建模客户生存期价值，根据用户行为对客户进行划分，为访问网站的客户选择最佳登录页面，确定适合列入营销活动的候选者，以及预测哪些客户处于停止使用软件包、服务或药物治疗的风险中。Gordon和Michael致力于分享他们的知识、技能以及对这个主题的热情。当他们自己不挖掘数据时，他们非常喜欢通过课程、讲座、文章、现场课堂，当然还有你要读的这本书来教其他人。经常可以发现他们在会议上发言和在课堂上授课。作者还在blog.data-miners.com维护了一个数据挖掘的博客。
Gordon生活在曼哈顿。在本书之前，他最近的一本书是Data Analysis Using SQL and　Excel，已经由Wiley于2008年出版。
Michael生活在马萨诸塞州剑桥市。他除了在Data Miners从事咨询工作之外，还在波　士顿大学卡罗尔管理学院讲授市场营销分析（Marketing Analytics）课程。

《数据挖掘技术》

书籍目录

《数据挖掘技术》

章节摘录

版权页： 插图： 1.每个业务都是服务业务 处于服务行业的公司，信息将赋予其竞争优势。这就是为什么连锁饭店会记录你首选无烟的房间，而租车公司会记录你喜欢的车的类型。此外，传统上认为自身不是服务提供者的公司也开始从不同的角度来思考。汽车经销商是出售汽车还是运输工具？如果是后者，那么每当你自己的车在商店里时，经销商就为你提供一辆替代车是合理的，许多经销商现在就是这么做的。 即使是日用商品也可以通过服务得到加强。一家家庭供热石油公司如果能够监视你的使用情况，并在你需要更多的石油时向你提供石油，那么相比一家公司期望你在油箱枯竭和管道冻结前记得打电话来安排你的订单，它销售的产品更好。对于信用卡公司、长途运输公司、航空公司以及所有类型的零售商而言，服务竞争通常会与价格竞争一样多或更多。 2.信息即产品 许多公司发现他们拥有的客户信息不仅对自己有价值，而且对其他人同样有价值。一家具有忠诚卡方案的超市有一些消费者包装食品行业会喜欢的信息——关于谁在购买哪些产品的知识。信用卡公司有一些航空公司想要了解的信息——谁在买大量的机票。超市和信用卡公司都处于知识经纪人的位置。超市可以通过打印优惠券向消费者包装食品公司索取更高的收费，此时超市会承诺通过向适当的购物者打印适当的优惠券获得更高的回报率。信用卡公司可以向航空公司收费，其目标是为经常旅行、但乘坐其他航空公司航班的人提供频繁的飞行积分。 Google了解人们正在Web上寻找什么。它在出售赞助商链接（以及其他事物）时利用这种知识。保险公司会为确保某人在搜索“汽车保险”时，为其提供它们站点的链接而支付相应的费用。金融企业将支付赞助商链接，从而当有人搜索诸如“抵押贷款再融资”之类的短语时显示其链接。

《数据挖掘技术》

编辑推荐

《数据挖掘技术:应用于市场营销、销售与客户关系管理(第3版)》的主题包括：如何创建稳定、持久的预测模型；数据准备和变量选择；用诸如回归、决策树、神经网络、基于记忆的推理之类的有指导技术来建模特定目标；用诸如聚类、关联规则和链接分析之类的无指导技术来发现模式；建模业务的事件发生时间问题，如下一次购买时间和预期的剩余生存期等；挖掘非结构化文本。

《数据挖掘技术》

精彩短评

1、尼玛，花了很长时间才啃完的一本书
2、大而全，基本涵盖了数据挖掘的方方面面，较基础的一本书。
3、门外汉入门数据挖掘，必读书
4、很想买第二版来着，但是貌似只能买二手和定制印刷的了。正好第三版出来了，毫不犹豫的入了。由于亚马逊在我心中的美好印象，我选择了它。可是你是故意让我失望的吗，在阅读前言的时候发现印刷的错误。作者写道：丝毫没有考虑到撰写这本书给我们的个人生活带来的影响。这次买的书，华丽丽的在'生'和'活'字之间插入了ai ok de这六个英文字母，于是就变成“丝毫没有考虑到撰写这本书给我们的个人生ai ok de活带来的影响”.百思不得其解啊，Z.CN!
5、翻译实在太烂，什么叫“古代水手学会了如何避免为保护西西里和意大利大陆之间狭窄海峡的锡拉岩礁岩石和卡律布迪斯漩涡”，这本书是清华大学体育学院翻译的？
6、CRM啊。。。和marketing沾边也是一点点有没有
7、系统详细，值得入手
8、内容全面，侧重营销领域，有较好的应用例。
9、广告、销售、运营感觉都是广义的客户关系管理的一部分，在技术上手段也有很多相似之处
10、理论很全，希望是结合软件操作，或代码呈现的方式。可能是很多细节需要在理解理论的基础上多尝试。
11、粗略的翻到前50页，已经发现了几处明显的错别字，感觉中文版根本没有认真校对。另外，翻译质量实在说不过去，太烂。我觉得不是译者的水平问题，而显然是责任心问题。我一般不写书评，这次实在无法忍受。
12、这本介绍的没有重点，如果偏实践业务，应该对多些案例，如果偏技术，那就应该更深入，并且翻译的实在是太差了，各种拗口不知所云。
13、这本书为什么打分这么高，可能每个人偏重不一样吧。以后有时间再看一遍。
14、大部头，CRM相关业务比较多，数据挖掘的算法和实现没有涉及。适合业务分析师看的书。
15、RT.....
16、帮朋友买的,朋友说很有用.
17、趁着有整块时间终于读完了。。基本覆盖了data mining能用到的各种方法，也只是粗浅的介绍，具体推导过程肯定不会一个一个详述，技术本身不重要，怎么与业务结合才是需要重点培养sense的。最后，翻译嘛，有的地方还是各种拗口。
18、讲了很多理论，实例只有文字叙述，而且不够详细，只能参考一下

《数据挖掘技术》

精彩书评

1、书本身非常好，但翻译很差。outbound在营销中明明是“外呼”，也就是外呼营销，打电话营销的意思，书里居然翻译成“出站”，相当无语。。客户关系生命周期里，“潜在客户”、“新客户”、“已建立的客户”、“前客户”这几个名词翻译的也不敢恭维，应该叫“潜在客户”、“新客户”、“老客户（或者活跃客户等等）”、“流失客户”更合适吧。
2、内容4分，翻译-1分翻译实在太差，不如直接看英文版，清华大学出版社的烂名声果然不是盖的“古代水手学会了如何避免为保护西西里和意大利大陆之间狭窄海峡的锡拉岩礁岩石和卡律布迪斯漩涡”这一看就知道不是人翻译出来的。
3、这本书不是纯讲数据挖掘理论的书，从本书的副标题你大概也能猜得到。对于像我这样数据挖掘领域的门外汉，读起这本书也没有多大的困难。这本书不是纯讲技术的书，但是其对技术理论的理解还是很有帮助。作者无论介绍数据挖掘的概念和技术，都是通过实例娓娓道来，之后引导读者进入数据挖掘的世界里面。书中的示例都是作者数十年工作中遇到的数据挖掘案例，碰到问题以及解决方案的总结，而不是凭空捏造的案例，从中可以看到作者在数据挖掘领域的造诣。作者对于数据挖掘的算法和原理虽有介绍，但是不多，所以对于初入数据挖掘领域的人阅读本书还是没有什么困难的，对于数据挖掘的专业人士，如果想了解具体的算法还是看专门的书籍吧。对于数据挖掘软件的操作，本书也没有介绍。本书重点是对技术的解释和数据挖掘的实际应用。对于各种主题，作者都有介绍。即使对于复杂的主题，作者也进行了简明的介绍，和其他数据挖掘书目不同的是，作者尽量减少了技术术语和数学公式的使用。每个章节都有作者宝贵经验的提示。本书内容分为四个部分。第一部分讨论了数据挖掘的业务上下文。作者为了讲解DM的方法，还铺垫统计学的一些基础知识和典故。第二部分介绍了有指导的数据挖掘技术，包括决策树，人工神经网络，最近邻方法：基于存储的推理和协同过滤，购物篮分析和关联规则。第三部分介绍了无指导学习的数据挖掘技术，如聚类分析等。最后一部分对数据挖掘术语中的数据进行了介绍，对于了现在的大数据等热点概念。电信运营商如何通过呼叫数据发现群体中的领袖？谁正在家里使用传真机?Google如何成为世界的统治者？Facebook上面你的朋友暴露了你什么信息？想要了解这些问题的答案吗？本书的第16章，链接分析（LinkAnalysis）会告诉你问题的答案。除了这些外，很多引人入胜的问题你可以在本书找到答案。这里不一一列举。阅读本书中，发现一个问题。可能由于纸张篇幅的限制，本书结束的时候没有参考文献和术语表索引，对于查询术语等有点不便。总之，正如本书作者所说，本书是这样一本书，当你开始自己的数据挖掘生涯时，也许想要阅读它。

《数据挖掘技术》

章节试读

1、《数据挖掘技术》的笔记-第407页

关联分析-无指导DM
1、关联规则类别
可操作的规则
平凡规则
费解的
2、度量关联规则的好坏
支持度p(condition and result)
置信度=支持度／p(condition)提升度=支持度／p(condition)p(result)卡方值

2、《数据挖掘技术》的笔记-第398页

凝聚聚类的局限性
计算复杂度高
难以可视化
对离群点很敏感

实际应用
不适合大量数据，适合以客户为中心的应用
1、基于用户偏好的产品聚类
第一步：定义距离或相似度：简化地计算两种产品同时被采购的比例
修正：用两种产品都买的客户百分比除以两种产品都没有买的客户的百分比
第二步：计算距离矩阵
2、根据用户响应对直销活动聚类

3、《数据挖掘技术》的笔记-第32页

Match the Profile 不知道为什么会被译者“ 天才 ”的译作“ 匹配剖析 ”， 是在用中国话翻译么？

这一节讲述的方法有误，用贝叶斯公式可以证明，计算分数时，应该用读者特征分布与全国人口分布比较后得到的index做连乘而不是累加。
或者index需要取对数。

4、《数据挖掘技术》的笔记-第53页

数据挖掘人员的挑战在于要找出哪些模式有益，哪些无益。考虑以下模式，所有这些模式都曾在大众媒体文章中被引用过，就像它们有预测价值：

- 在野党(总统竞选失败的政党)在非大选年选举中会获得国会席位。
- 当美国职业棒球联盟赢得世界职业棒球大赛时，共和党会继续把持着白宫。
- 华盛顿红人队赢得最后一场主场比赛时，执政党继续执掌白宫。
- 在美国总统竞选中，高个子的男人常常会赢。<原文开始></原文结束>

每隔四年，略高于半数的美国选民都很兴奋地投票选举总统候选人。几个月后，候选人接管而失

望就此开始——政客们根本不能兑现选民所期望的所有承诺。两年后的国会选举会出现反弹，通常是由感到失望的支持者不投票导致的结果。因为该模式有一个基本的解释，似乎很可能会持续到将来，表明它有预测价值。

而后两个所谓的预言——涉及体育赛事的那两个预言，看上去明显没有预测价值。无论共和党人和美国联盟在过去可能已经共享过多少次胜利(作者没有研究过这一点)，但是没有理由可以预测在将来会继续关联。

那候选人的高度呢？自1948年杜鲁门(其身材矮小，但比杜威高)当选之后，只有卡特击败福特和布什打败克里是较为矮小的候选人赢得普选的两次选举。在2000年的选举中，如果我们假设模式是与赢得普选而非选举人票相关，那么戈尔的6英尺1英寸对布什总统的6英尺还是符合该模式。在2008年，打篮球的奥巴马击败了较为矮小的麦凯恩。高度看上去与当总统这份工作毫不相关。然而，我们的语言展示了“身高歧视”：我们把仰视看成是表示尊敬的姿态，而俯视表示蔑视。身高与更好的童年营养相关，其反过来会提高智商以及其他社会成功的指标。

5、《数据挖掘技术》的笔记-第17页

引言：本文为数据挖掘技术（第三版）第一章读书笔记

数据挖掘的良性循环包含了四个步骤

识别业务机会；
挖掘数据将其转换为可操作的信息；
根据信息采取行动；
度量结果；
根据以上步骤的提示，成功的关键是把数据挖掘合并到业务流程，并能促进数据挖掘人员与使用结果的业务用户之间的通信。

1.识别业务机会

数据挖掘的良性循环首先在于识别合适的业务机会。为了避免浪费分析工作，受限应愿意针对结果采取行动。数据挖掘很适用于以下几类业务。
规划新的产品介绍
计划直接营销活动
理解客户的流失/波动
评价营销测试的结果
分配营销预算以吸引最有利可图的客户

度量过去的努力和有关业务的特设问题同时启示数据挖掘的机会：

什么类型的客户会影响过去的活动？
最好的客户住在哪里？
在自动柜员机前漫长的等待是客户流失的原因嘛？
有利可图的客户会使用客户支持嘛？
哪些产品应该使用Clorox漂白来升级？
2.将数据转换为信息

数据挖掘的重点是讲数据转换为可操作的结果。但是许多缺陷降低了使用数据挖掘的能力：
坏的数据格式，例如客户地址中不包含邮政编码
混乱的数据字段，例如交货日期在一个系统定义为“计划交付日期”，在另一个系统定义为“实际交付日期”

功能缺陷，例如呼叫中心系统不允许对每个客户进行备注
法律影响
组织因素，因为业务部门不愿意改变他们的行动，特别是没有奖励的情况下
不及时，因为结果可能来得太晚而不再适合采取行动
3.根据信息采取行动

采取行动是数据挖掘良性循环的目的，数据挖掘可以使业务更加明智，随着时间的推进，更明智的决定会促成更佳的结果。
当客户在线时，把结果合并到推荐系统
通过直接邮寄，电子邮件，电话营销等向客户和潜在客户发送消息；使用数据挖掘，不同的信息应该发送给不同的人
优先客户服务
调整库存水平
数据挖掘的结果必须交给可以接触到客户或影响客户关系的核心业务流程

4.度量结果

需要使用可度量的结果来验证行动的结果，分为以下四个组进行验证。

目标组：接受处置，且具有指示响应的模型评分
控制组：接受处置，且随机或基于较低的模型评分进行选择
对照组：不接受处置，且随机或基于较低的模型评分进行选择
模型化对照组：不接受处置，且具有指示响应的模型评分

以活动为例，多去思考下以下几个问题。

该活动是否抵达并带来有利可图的客户？
得分更高的模型评分会表明更高的响应度嘛？
这些客户是被保留还是可预期？
本活动所出大的最忠实客户的特征是什么？
新获得的顾客会购买额外的产品吗？
一些信息或优惠比其他的更有效嘛？
活动所触达的客户可通过其他备用渠道触达嘛？

http://8jo.cn/2015-02-05

《数据挖掘技术》

版权说明