Introduction to Data Mining

数据挖掘教程

李保坤　张丽娟　编著

西南财经大学出版社

Introduction to Data Mining

数据挖掘教程

李保坤　张丽娟　编著

西南财经大学出版社

图书在版编目(CIP)数据

数据挖掘教程/李保坤,张丽娟编著. —成都:西南财经大学出版社,2009.7

ISBN 978-7-81138-440-6

Ⅰ.数… Ⅱ.①李…②张… Ⅲ.数据采集—教材 Ⅳ.TP274

中国版本图书馆 CIP 数据核字(2009)第 120169 号

数据挖掘教程

李保坤 张丽娟 编著

责任编辑:于海生 黄慧英

封面设计:杨红鹰

责任印制:封俊川

出版发行:	西南财经大学出版社(四川省成都市光华村街 55 号)
网　　址:	http://www.bookcj.com
电子邮件:	bookcj@foxmail.com
邮政编码:	610074
电　　话:	028-87353785　87352368
印　　刷:	郫县犀浦印刷厂
成品尺寸:	170mm×240mm
印　　张:	9
字　　数:	165 千字
版　　次:	2009 年 7 月第 1 版
印　　次:	2009 年 7 月第 1 次印刷
印　　数:	1—2000 册
书　　号:	ISBN 978-7-81138-440-6
定　　价:	18.00 元

作者简介

李保坤，美国新墨西哥州立大学博士，西南财经大学统计学院副教授，应用统计研究所副所长。

前 言

几年前，全球信息技术领域的领军人物比尔·盖茨在预言未来技术发展时把数据挖掘列为最重要的方向之一。因此，无论你是正在高校攻读的学生，还是在社会上打拼的专业技术人员，只要有一颗学习上进的心，你就不得不关注数据挖掘这一新领域。

数据挖掘是统计学与计算机科学的交叉学科，解释数据挖掘原理及算法术语有的来自统计学领域，有的来自计算机科学领域，对许多学习者来说，数据挖掘的书籍总有一些不适应的感觉。此书将尽量避免很专业化的术语，使用通俗易懂的语言，让具有基础统计学、线性代数以及计算机基本知识的读者就可以理解和学习。为了方便读者能够验证所学习的算法，本书还配有一套数据挖掘软件——西南财经大学数据挖掘系统。

笔者在美国新墨西哥州立大学留学期间开始系统学习数据挖掘。作为学习过多门计算机科学研究生课程的数理统计学专业博士研究生，理解数据挖掘的原理应是非常容易的，但苦恼的是很难找到合适的软件来验证这些原理的实际应用效果。著名数据挖掘商用软件为了显示自己的“强大”或“专业”，有的把一些研究人员已经验证不合理的方法硬塞了进来以充数；有的把数据粗略化以便加快速度；有的为了单一挖掘目的而牺牲其他功能；等等。而且功能全面且界面友好的免费软件因为编程困难且费时费力，普通的数据挖掘研究人员难以制作，导致在网络上难以找到。为了消化所学习过的数据挖掘算法并应用，为数据挖掘的后学者提供帮助，笔者自己动手使用 VBA 编写了一套在微软 Word 上运行的数据挖掘插件 EASY MINER。现已经较好地用于西南财经大学数据挖掘课程的教学之中。

EASY MINER 受到学生和老师的普遍欢迎。同学们不仅在上数据挖掘课时通过这个软件理解数据挖掘原理，而且用它在课后作科研论文所需的数据分析和参加建模竞赛的练习。为了让更多的人能够轻松学习和理解数据挖掘原理并应用，我们把 EASY MINER 的功能模块重新用 VS. NET 实现出来。在此过

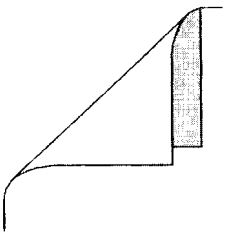
程中，北京工业大学继续教育学院张丽娟老师（中国地质大学博士）补充了一些数据挖掘算法，对已有的算法进行了优化处理并增加了图形显示模块。另外，该软件还结合了SAS EM、微软SQL Server挖掘和马克威数据分析系统中挖掘功能的一些优点，使改进后的EASY MINER与以前相比有了新的飞跃。我们把它重新命名为西南财经大学数据挖掘系统。

这本书的文字内容主要参考了美国麻省理工学院的数据挖掘开放讲义、国外许多大学老师关于数据挖掘课程的教学资料以及网络上对有关算法的介绍材料。书中使用的数据均来自统计学教材或数据挖掘教材中使用的标准数据，数据分析结果和图形展示由我们自己制作的西南财经大学数据挖掘系统软件生成。如果读者希望更深入地了解数据挖掘知识，或下载有关资料，请访问作者本人的数据挖掘学习网站：http://bali.cai.swufe.edu.cn/dm.htm。在此希望读者通过这本书和软件而大大增进学习效率。

随着数据挖掘算法的不断发展，我们将适时根据其发展变化修订本教材以及和本教材配套的西南财经大学数据挖掘系统。由于作者才疏学浅，本书的错误疏漏之处肯定不少。因此恳请各位统计学、计算机科学、数学等有关领域的读者不吝赐教。同时也恳请学习者将使用本教材的建议和意见及时反馈给我，对此我表示衷心的感谢。

李保坤 博士
2009年6月18日
于西南财经大学

目 录



第一章 介绍

1.1 这本书的读者对象

数据挖掘通常要涉及统计和机器学习(或者叫做人工智能)方面的算法。如果作者的目的只是让读者掌握数据挖掘的技术和工具的话,这类书籍因为缺乏详细的解释,因此对读者的指导作用就不会太强。另外也有许多关于数据挖掘算法比较专业的书籍,它们的对象是统计研究人员或者高年级的研究生,里面没有具体的商业案例分析,因此一般的读者会觉得太涩。有鉴于此,我们在写作此书时内容上主要突出了以下两个特色:

(1)介绍分类、预测、数据精简等数据挖掘核心技术的基础理论和算法;

(2)采用商业案例说明这些算法的使用。

另外,这本书在形式上和普通的书籍有一个显著的区别:它配备了一套演示各种算法的软件——西南财经大学数据挖掘系统,供读者理解数据挖掘思想、算法以及进行数据挖掘练习。本书使用的数据全部放在商业分析人士所熟悉的 EXCEL 电子表格里(此外,软件还可处理文本文件、sas 的.dat 格式文件、ACCESS 数据库文件),引入的案例浅显易懂,所介绍的数据挖掘算法均可使用软件实现。本书的对象是学习数据挖掘技术的商学院学生和在商业部门的实际操作人员。虽然写作这本书的目的是为了满足学生学习数据挖掘思想和技术的需要,那些在实际工作中使用数据挖掘的分析人员和咨询人员也会发现这是一本很好的入门教材。

1.2 什么是数据挖掘

数据挖掘是一个相当新的领域,目前还处于演变之中。最早的关于数据挖

掘的国际会议于1995年举行。关于数据挖掘有许多定义。Hand等人2001年给出的一个简明定义:“数据挖掘就是从大型数据集合里挖掘出有用的信息”,可说是抓住了数据挖掘的本质。Berry和Linoff于2000年给出了一个稍微长一点的定义:“数据挖掘是通过自动或者半自动的方法对大量的数据进行处理和分析以便发现有意义的模式或者规律”。还有一个定义来源于一家信息技术研究公司——加特那集团:“数据挖掘是从大量的存储数据里进行筛选,采用模式识别技术以及统计和数学技巧,发现有意义的新的相互关系、模式以及趋势的过程。”

1.3 数据挖掘的用途

数据挖掘可用于许多领域。部队可使用数据挖掘技术了解各种因素对炸弹落点的准确度影响;情报机构可以利用它决定在大量被截获的情报中哪些是有价值的;网络安全专家可以使用它决定一段网络代码是否有破坏性;医疗研究人员可以使用数据挖掘方法预测癌症复发的可能性。尽管数据挖掘的方法和工具具有广泛的适用性,但是考虑到这本书的许多读者是商学院学生,我们在选择案例时倾向于商业领域的应用。人们应用数据挖掘方法所希望解决的一些普通的商业问题有:

(1)大量的潜在顾客里面哪些人最可能成为买家

我们使用分类技术(如Logistic回归、分类树以及其他方法)可以挑选出其特征数据和目前的最佳顾客以及最为近似的顾客。同时我们还可以用预测技术预测他们会花多少钱。

(2)哪些客户将来最可能搞欺诈活动(或者已经搞了欺诈)

我们可以利用分类技术识别出哪些更可能涉嫌欺骗医疗补助申请,并对这些申请多加关注。

(3)哪些贷款申请人可能搞欺骗

我们可以用一些分类技术识别出有欺骗倾向的贷款申请人或者用Logistic回归的方法为申请人算出一个“欺骗概率”值,即对每一个申请人我们都可以计算出其搞欺骗的可能性。

(4)哪些客户更可能会放弃订购服务(如电话、杂志等)

对此我们也可以用分类技术识别出放弃订购服务的客户来,或者用Logistic回归的方法为客户算出一个“流失概率”值。这样的话,一些对客户的鼓励措施,如折扣等就会被用到最需要的地方。

1.4 数据挖掘的起源

数据挖掘处于统计学和机器学习(也被称为人工智能)领域的交叉点上。处理数据和建模的许多方法在统计学领域早就存在——例如线性回归、Logistic回归、判别分析以及主成分分析。但是,经典统计学的两个核心难点——计算复杂并且数据稀少——在数据挖掘里就不存在,因为在数据挖掘的应用中数据量大,并且目前的计算机计算能力超强。

正因为此,Daryl Pregibon 把数据挖掘描述为"建立在规模和速度上的统计学"。有人把这一说法进行了推广:数据挖掘是"建立在规模、速度和简单化上的统计学"。在此简单并不完全是指算法上的简单,更多的是指推理逻辑上的简单。在经典统计学环境里因为数据少,同一个样品既被用于作估计,也被用来决定估计值的可靠程度。这使得许多人对置信区间和假设检验的思想感到不易理解,关于它们的限制条件也不好领会。与此对照,数据挖掘用一个样本进行模型拟合,用另一个样本进行拟合评估的做法就较容易被人理解。

计算机科学为我们提供了机器学习技术,例如分类树和神经网络。它们依赖超强的计算力而不需按照经典统计模型的方式求解。另外,数据库管理功能的扩张也是数据挖掘诞生的一个因素。

经典统计学理论里强调的东西(例如:判定一个模式或者一个感兴趣的结果是否是随机的)在数据挖掘中就不是问题。和统计学作个比较,数据挖掘处理的是以无限制方式存放的大量数据的数据集合,不可能对需要解决的问题施加统计理论所需的严格限制。

因为计算能力的过于强大,数据挖掘的一些方法容易产生"过度拟合"危险。过度拟合指的是现有样本跟一个模型拟合太过,以至于模型不仅描述了数据的根本特性,而且也描述了其随机特性。按工程上的术语指这个模型不光是拟合信号,还拟合噪声。

1.5 术语和注释

数据挖掘是几个领域结合的产物,作挖掘的人通常用好几个术语来指示同一个东西。例如,在机器学习领域,被预测的变量是输出变量或者目标变量。对于统计人员,被预测变量是因变量。下面是一些常用的数据挖掘术语:

"算法"指的是用于实现某一数据挖掘技术——如分类树、辨识分析等的特

定程序。

“属性”也被称为“特性”、“变量”、或者从数据库的观点，是一个“域”。

“个体”是关于一个单元的测量值的集合——例如一个人的身高、体重、年龄，等等；它也被称作“记录”或者“排”（每一排通常代表一个记录，每一列代表一个变量）。

“置信度”在形如“如果买了 A 和 B，就要买 C”的关联法则里有特定的含义。置信度是已经买了 A 和 B，还要买 C 的条件概率。在统计学里，关于估计值的误差大小，置信度有更广泛的含义。

“因变量”在有约束学习里是那个被预测的变量；也被称作“输出变量”、“目标变量”或者“结果变量”。

“估计”指的是预测一个连续型输出变量的值，也被称作“预测”。

“特征”也被称作“属性”、“变量”或者从数据库的观点，称为“域”。

“输入变量”是在有约束学习里作预测的变量，也被称作“自变量”、“预测变量”。

“模型”通常指的是一个数学公式，包括为它设置的参数（许多模型具有用户可以调节的参数）。

“结果变量”在有约束学习里是那个被预测的变量，也被称作“因变量”“输出变量”、“目标变量”或者“输出变量”。

“P(A|B)”读作“已知 B 已经发生，A 将发生的概率”。

“预测”指的是预测一个连续输出变量的值，也被称作“估计”。

“记录”是关于一个单元的测量值的集合——例如一个人的身高、体重、年龄，等等；它也被称作“个体”或者“排”（每一排通常代表一个记录，每一列代表一个变量）。

“分数”指的是一个估计的值或者类。

“给新数据打分”的意思是利用训练数据得出的模型预测新数据里的输出值。

“有约束学习”指的是用已有记录得到算法（逻辑回归、回归树等）的过程。在这些记录里人们感兴趣的输出变量是已知的，这个算法“学习”如何预测新记录里输出变量的值，这些值在新纪录里是没有的。

“测试数据”指的是只在模型建立和选择的过程的末期，用于评价最终模型对新数据的处理效果的那部分数据。

“训练数据”指的是用于拟合模型的那部分数据。

“验证数据”指的是用于评价模型拟合状况、调整模型、选择最佳模型的那部分数据。

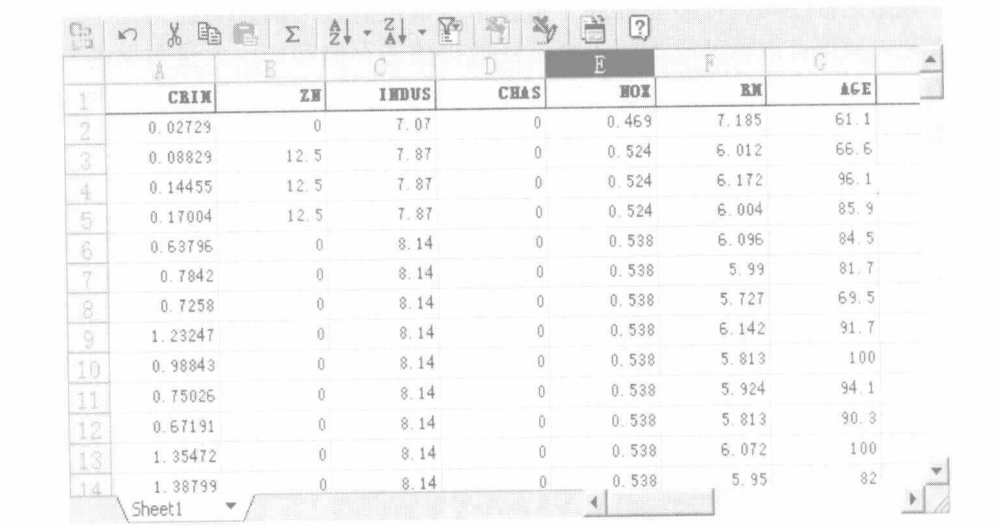
“无约束学习”指的是人们试图从数据中了解一些东西的分析，而不是预

测感兴趣的输出值(例如输出结果是否属于某个聚类)。

“变量”也被称为“特性”、“属性”或者从数据库的观点,是一个“域”。

1.6 数据集合的组织

数据集合几乎总是以变量为列,记录为行的形式组织和显示的。下面的例子(波士顿住房数据,见图1.1)记录了一些人口统计区的14个变量。每一行代表一个人口统计区。例如第一区的人口平均犯罪率(CRIM)是0.027 29,房屋面积超过25 000平方英尺(1平方英尺≈0.09平方米,下同)以上的家庭数(ZN)是0,等等。在做有约束学习的时候,这14个变量中有一个变量是结果变量,通常列在最后或者最开头(在本例中,结果变量是列在最后的中间值)。



	A	B	C	D	E	F	G
1	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE
2	0.02729	0	7.07	0	0.469	7.185	61.1
3	0.08829	12.5	7.87	0	0.524	6.012	66.6
4	0.14455	12.5	7.87	0	0.524	6.172	96.1
5	0.17004	12.5	7.87	0	0.524	6.004	85.9
6	0.63796	0	8.14	0	0.538	6.096	84.5
7	0.7842	0	8.14	0	0.538	5.99	81.7
8	0.7258	0	8.14	0	0.538	5.727	69.5
9	1.23247	0	8.14	0	0.538	6.142	91.7
10	0.98843	0	8.14	0	0.538	5.813	100
11	0.75026	0	8.14	0	0.538	5.924	94.1
12	0.67191	0	8.14	0	0.538	5.813	90.3
13	1.35472	0	8.14	0	0.538	6.072	100
14	1.38799	0	8.14	0	0.538	5.95	82

Sheet1

图1.1 波士顿住房数据

1.7 数据挖掘迅速发展的因素

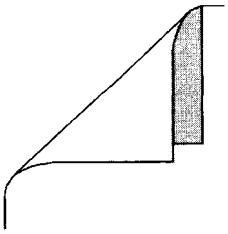
或许推动数据挖掘发展的最重要的因素是数据的增长。2003年零售业巨无霸沃尔玛在一个10兆兆位(Terabyte)的数据库里显示出每天达成2千万笔交易。而在1950年,最大的几家公司的数据以电子表格的形式也只占用几十兆位(1 Terabyte = 1 000 000 Megabyte)。

数据本身的增长并不是简单地由经济和知识库的扩张驱动的,而是由数据

自动存取装置的大量增加以及这些装置的成本降低造成的。不仅更多的事件被记录,而且每一个事件有更多的信息被捕捉下来。可扫描的条形码、销售点标记、鼠标点击历史和全球定位卫星的数据都是这方面的例子。互联网的增长制造了许多新的信息源。目前人们可以在互联网上搜索图书、定购货品并且其详细情况均可记录下来。在营销领域,从注重商品和服务到注重客户及其需要的转变产生了对客户详细信息的需求。

用于记录客户交易以辅助日常商业活动的实用数据库可以处理简单的查询,但却不足以进行较复杂的综合分析。这些可操作数据库的数据因此被挖掘、转换并被送到一个数据仓库(或者称为数据加工厂)——一个把企业的决策系统结合在一起的大型综合数据存储系统。该系统可以连接一些完成单一任务的较小的数据店(Data Mart),这些数据店可以存储外部信息资源的数据(如信用等级数据)。

数据挖掘使用的许多处理和分析技术目前没有强大的计算能力是不可能实现的。数据存取器件成本的不断下降使得建造存储和产生大量数据的装置成为可能。总而言之,计算能力方面的持续迅速的改进是数据挖掘发展的一个基本动力。



第二章 数据挖掘过程概览

2.1 数据挖掘的核心思想

2.1.1 分类

分类或许是数据分析最基本的形式。一份工作录用通知的接收者可能接收也可能拒绝这份工作;一项贷款的申请人可能会把钱准时还上,也可能过期还上或是宣告破产;一笔信用卡交易可能是正常的也可能是欺诈性的;在网络上下载的一段代码可能携带病毒也可能不携带病毒;一种疾病的患者可能康复,也可能继续在病中或者病死。数据挖掘的一项基本任务就是用类别已知的数据找出规则,然后把这些规则用在未进行分类的数据上。

2.1.2 预测

预测和分类相似,差别在于我们是预测一个变量的数值,而不是一个类别(比如购买者或者非购买者)。当然,在分类时我们试图去预测一个类别,而"预测"这个术语在这本书里指的是预测一个连续变量的数值。(在数据挖掘书籍中,"估计"有时被用来表示预测一个连续变量的值,"预测"既可以用于连续变量的数据,也可以用于类型变量的数据。)

2.1.3 关联分析

有了储存客户交易信息的大型数据库自然就产生了对购买物品进行的关联分析(哪种物品和哪种物品是搭配着买的)。通过关联分析得到的"关联法则"然后以多种方式被利用。例如,百货商店可以利用关联法则在扫描了一个顾客的采购单后印制优惠券,优惠券上打折扣的商品是由通过分析大量顾客的采购单得到的关联法则决定的。