关键词
数据挖掘
Excel软件应用

高效办公 Express to Office Efficiency
"职"通车

Excel 2007 数据挖掘完全手册

谢邦昌 朱建平 来升强 编著

系统性 详细叙述数据挖掘的一般概念、通行规范、方法技术以及软件应用等，使读者获得一个较为清晰和正确的数据挖掘观念

具体性 围绕Excel 2007的数据挖掘模块，对Excel 2007强大的表格工具详加讲解，有助于读者在工作表中完成种种复杂的数据分析任务

实用性 提供一些大型应用案例，通过详细的操作讲解和结果解释，可令读者获得实际的数据挖掘经验，从而能迅速加以应用

清华大学出版社

高效办公“职”通车

Excel 2007 数据挖掘完全手册

谢邦昌　朱建平　来升强　编著

清华大学出版社
北　京

内 容 简 介

本书围绕 Excel 2007 的数据挖掘模块，通过大量操作示范，介绍了主流的数据挖掘方法。全书包括数据挖掘算法介绍、Excel 2007 数据挖掘模块介绍、其他分析工具介绍、数据挖掘范例 4 篇，共 26 章。除了给出有关的理论和原理阐述之外，还提供了一些大型应用案例。通过详细的操作讲解和结果分析，读者可以获得实际的数据挖掘经验，并能迅速地在自己所处的领域中加以应用。

利用 Excel 2007 的数据挖掘模块，读者无须经过专业培训，就能完成多种数据挖掘任务。本书适用于学习数据挖掘和相关课程的学生、运用 Excel 2007 进行复杂大型数据分析的职场人士及咨询公司从业人员等。

图书在版编目（CIP）数据

Excel 2007 数据挖掘完全手册/谢邦昌，朱建平，来升强编著. —北京：清华大学出版社，2008.7
（高效办公“职”通车）

ISBN 978-7-302-17474-5

Ⅰ. E…　Ⅱ. ①谢…　②朱…　③来…　Ⅲ. 电子表格系统，Excel 2007-手册　Ⅳ. TP391.13-62

中国版本图书馆 CIP 数据核字（2008）第 057618 号

责任编辑：吴颖华　孙　斌
封面设计：张　岩
版式设计：牛瑞瑞
责任校对：马军令
责任印制：王秀菊
出版发行：清华大学出版社　　**地　　址**：北京清华大学学研大厦 A 座
http://www.tup.com.cn　　**邮　　编**：100084
社　总　机：010-62770175　　**邮　购**：010-62786544
投稿与读者服务：010-62776969，c-service@tup.tsinghua.edu.cn
质　量　反　馈：010-62772015，zhiliang@tup.tsinghua.edu.cn
印 刷 者：清华大学印刷厂
装 订 者：三河市溧源装订厂
经　　销：全国新华书店
开　　本：185×260　**印　张**：19.25　**字　数**：426 千字
版　　次：2008 年 7 月第 1 版　　**印　次**：2008 年 7 月第1 次印刷
印　　数：1～5000
定　　价：32.00 元

丛 书 序

随着计算机技术的全面迅速发展，人们将在行政办公、财务管理、会计、统计、审计等众多领域面对计算机的应用和管理。掌握计算机在这些领域的应用，一方面可以极大地提高工作效率，另一方面也可以提高业务水平。信息时代，许多行业都要求工作者有很强的计算机操作技能，做到运用自如，熟练而且深入地掌握软件的应用。而要做到这一点，必须从软件的实际应用入手。

正是在这一大背景下，我们策划了本套丛书，精选了应用领域较广泛、较常用的一些软件，如 Excel、SPSS、用友财务软件、金蝶财务软件等，旨在帮助广大办公人员、财务人员、统计分析人员、审计人员及相关专业的学生快速掌握这些软件的应用，用以解决实际工作或学习中的问题，提高自身的应用水平。

内容安排

本丛书强调软件与职业应用相结合，以实例为载体，着重介绍常用软件的操作功能和实践应用技巧。本套丛书包括：

《用友 ERP-U8 财务软件应用实务》
《金蝶 K/3 财务软件应用实务》
《SPSS 在统计分析中的应用》
《Excel 2007 在统计分析中的应用》
《Excel 2007 在会计工作中的应用》
《Excel 2007 在财务管理中的应用》
《Excel 2007 在审计分析中的应用》
《Excel 2007 函数、公式范例应用》
《Excel 2007 数据图表范例应用》
《Excel 2007 VBA 办公范例应用》
《Excel 2007 数据挖掘完全手册》

丛书特色

1．软件与职业应用相结合，实用性强。深入浅出地讲述了行政办公、财务管理、会计、统计、数据挖掘、审计各职业领域的关键知识，系统介绍了相应的软件应用方法及技巧，对实际工作有极大的帮助和指导意义。

2．内容丰富，案例典型。本套书每章都有实践案例，读者可以根据自己的情况进行取舍，直接应用于具体的工作之中。

3．结构合理，逻辑清晰。从全新的实例角度出发，按照“基本知识点讲解——实践应用——解决问题”的逻辑结构编写，全面介绍了这些软件在日常工作中的应用。符合读者的学习思路，可以使广大读者在最短的时间内学习并利用应用软件的各种强大功能，少走

弯路，迅速提升专业技能和提高工作效率。

4．光盘特色。本套丛书大部分都配有光盘，汇集了书中所用的应用软件、实例素材，及应用实例的视频，极大地方便了读者的学习。

读者定位

1．适合作为行政办公、财务管理、会计、统计、审计等领域在职工作人员提高自身业务水平的参考用书，还适合于非统计类的研究生及从事相关数据分析人员学习。

2．适合作为高校财务管理、会计、统计、审计、市场营销、电子商务、信息管理等相关专业的教材或学习用书。

3．适合作为各相关领域应用培训或职业培训的教学用书。

售后服务

如果读者在阅读图书的过程中有什么问题或需要帮助，可以登录本丛书的信息支持网站 http://www.thjd.com.cn 或通过 zzfangcn@vip.163.com（010-62788951-269）联系，也可以在 http://www.thjd.com.cn 的读者留言栏目留言，我们将尽快给您提供帮助与支持。

前　言

目前，各行各业都开始利用计算机及相应的信息技术进行管理和决策，这使得各企事业单位生成、收集、存储和处理数据的能力大大提高。数据量与日俱增，大量复杂信息层出不穷，人们将面临着复杂数据的处理问题。Excel 是当前使用最普遍的电子表格软件，它能容易地完成图表的制作、统计、分析以及数据处理，不但功能强大，而且简单易用。最新版本的 Microsoft Office Excel 2007 支持超过 104 万笔记录的单张数据工作表，并可以同时存储 1.6 万列的数据。为能有效提升 Excel 2007 用户数据处理和分析的能力，微软公司提供了一个免费的数据挖掘模块。通过调用该模块，Excel 2007 用户可以方便快速地完成以往只有使用专业数据挖掘软件才能完成的任务。因此，我们编写了《Excel 2007 数据挖掘完全手册》这本书，其目的是使具有一定 Excel 基础的读者，能够在了解相关统计思想与方法的基础上，运用该软件对复杂数据和海量数据进行处理、分析。

本书的编写力求以统计思想为主线，以数据挖掘技术应用为目的。基本内容和特点具体体现为：第 1 篇详细叙述数据挖掘的一般概念、通行规范、方法技术以及软件应用等，使读者获得一个较为清晰和正确的数据挖掘观念。第 2 篇围绕 Excel 2007 的数据挖掘模块，通过大量操作示范，详细讲述了 Excel 2007 数据挖掘模块的九大模型的使用。这些模型包括决策树、贝叶斯概率分类、关联规则、聚类分析、时序聚类、线性回归、Logistic 回归、类神经网络和时间序列分析，基本涵盖了主要的数据挖掘技术和方法。第 3 篇介绍了 Excel 2007 的其他分析工具，结合数据挖掘技术和方法，使用改进的 Excel 表格工具，可以很方便地进行图形化的分析。第 4 篇是数据挖掘的案例分析，包括投资决策、信用评级，以及市场销售和客户细分等领域的数据挖掘模型。通过详细的操作讲解和结果解释，读者可以获得实际的数据挖掘经验，并能迅速在自己所处的领域中加以应用。

本书适合多层次多专业人士如数学、统计、经济金融、管理类等专业的本科生、专科生学习，还适合于非统计类的研究生及从事相关数据分析的人员阅读。

本书在编写及出版的过程中，得到了厦门大学经济学院计划统计系、台湾辅仁大学统计资讯学系和清华大学出版社的大力支持，在此一并表示衷心感谢！编写一本好书并不容易，尽管我们努力想奉献给读者一本满意的书，但仍有一些内容达不到读者各方面的要求。书中难免有疏漏之处，恳请读者多提宝贵意见，以便今后进一步修改与完善。

为了方便读者高效、便捷地使用本书，特免费提供本书所有实例的原始数据、源文件，请登录清华大学出版社网站（www.tup.tsinghua.edu.cn）下载。

本书的编写得到了厦门大学讲座教授基金和国家教育部“新世纪优秀人才支持计划”（Program for New Century Excellent Talents in University，NCET）的资助。

编　者

2008 年 3 月

目　录

第 1 篇　数据挖掘算法介绍

第 1 章　数据挖掘简介 3
1.1　数据挖掘的定义 3
1.2　数据挖掘的重要性 3
1.3　数据挖掘的功能 3
1.4　数据挖掘的步骤 4
1.5　数据挖掘建模的标准 CRISP-DM 5
第 2 章　数据挖掘运用的理论和技术 7
2.1　回归分析 7
2.1.1　简单线性回归分析 7
2.1.2　多元回归分析 7
2.1.3　岭回归分析 8
2.1.4　Logistic 回归分析 9
2.2　关联规则 9
2.3　聚类分析 10
2.4　判别分析 11
2.5　类神经网络分析 12
2.6　决策树分析 13
2.7　其他分析方法 15
第 3 章　数据挖掘与相关领域的关系 17
3.1　数据挖掘与统计分析的不同 17
3.2　数据挖掘与数据仓储的关系 17
3.3　知识发现与数据挖掘的关系 18
3.4　OLAP 与数据挖掘的关系 19
3.5　数据挖掘与机器学习的关系 19
3.6　网络挖掘与数据挖掘的关系 20
第 4 章　数据挖掘商业软件产品及其应用现状 21
4.1　数据挖掘商业软件的分类 21
4.2　主要软件的介绍 21

4.3 顾客关系管理 22
4.4 数据挖掘的行业应用 23

第 2 篇 Excel 2007 数据挖掘模块介绍

第 5 章 安装与设定 Excel 2007 数据挖掘加载项 27
5.1 系统需求 27
5.2 开始安装 27
5.3 完成安装验证 30
5.4 组件设定 30
5.5 配置完成检查 35
第 6 章 Excel 2007 数据挖掘入门 37
6.1 Excel 2007 数据挖掘功能介绍 37
6.2 数据挖掘使用说明 37
6.2.1 目录查询 37
6.2.2 开始功能 38
6.2.3 视频和教学 39
6.3 数据挖掘连接配置 39
6.3.1 设定目前的连接 39
6.3.2 跟踪 41
6.4 数据准备 41
6.4.1 浏览数据 41
6.4.2 清除数据 44
6.4.3 分割数据 46
6.5 数据建模 50
6.6 精确度和验证 51
6.6.1 准确性图表 51
6.6.2 分类矩阵 52
6.6.3 利润图 53
6.7 模型用法 53
6.7.1 浏览功能 53
6.7.2 查询功能 56
6.8 模型管理 57
6.8.1 重新命名挖掘模型 57
6.8.2 删除挖掘结构 57
6.8.3 清除挖掘结构 58
6.8.4 用原始数据处理挖掘结构 58

6.8.5 用新数据处理挖掘结构 58
6.8.6 导出挖掘结构 59
6.8.7 导入挖掘结构 60
第 7 章 决策树 61
7.1 基本概念 61
7.2 决策树模块的建立 61
7.3 决策树与判别函数比较 61
7.4 计算方法 62
7.4.1 确定预测精度的标准 62
7.4.2 选择分裂（分层）技术 62
7.4.3 定义停止分裂（分层）的时间点 62
7.4.4 选择适当大小的决策树 63
7.5 Excel 2007 决策树算法 63
第 8 章 贝叶斯概率分类 71
8.1 基本概念 71
8.2 Excel 2007 贝叶斯概率分类 73
第 9 章 关联规则 84
9.1 基本概念 84
9.2 关联规则的种类 85
9.3 关联规则的算法：Apriori 算法 85
9.4 Excel 2007 关联规则 86
第 10 章 聚类分析 96
10.1 基本概念 96
10.2 层次聚类分析 96
10.3 聚类分析原理 97
10.4 Excel 2007 聚类分析 100
第 11 章 时序聚类 116
11.1 基本概念 116
11.2 相关研究和算法 116
11.3 Excel 2007 时序聚类 117
第 12 章 线性回归 126
12.1 基本概念 126
12.2 简单回归分析 127
12.3 多元回归分析 130

12.4 Excel 2007 线性回归 133
第 13 章 Logistic 回归 142
13.1 基本概念 142
13.2 logit 变换 142
13.3 Logistic 分布 143
13.4 列联表的 Logistic 回归模型 144
13.5 Excel 2007 Logistic 回归 145
第 14 章 类神经网络 161
14.1 基本概念 161
14.2 类神经网络的架构与训练算法 163
14.3 类神经网络的特性 163
14.4 类神经网络应用 163
14.5 类神经网络优缺点 164
14.6 Excel 2007 类神经网络 165
第 15 章 时间序列分析 175
15.1 基本概念 175
15.2 时间序列的成分 177
15.3 时间序列数据的图形介绍 178
15.4 利用平滑法预测 182
15.5 用趋势方程预测时间序列 186
15.6 预测含趋势与季节成分的时间序列 187
15.7 利用回归模型预测时间序列 188
15.8 其他预测模型 189
15.9 单变量时间序列预测模型 189
15.10 时间趋势预测模型 192
15.11 Excel 2007 时间序列 193
第 16 章 DMX 介绍 198
16.1 DMX 介绍 198
16.2 DMX 函数介绍 199
16.2.1 模型建立 200
16.2.2 模型训练 201
16.2.3 模型使用（预测） 201
16.2.4 其他函数语法 202
16.3 DMX 数据挖掘语法 205
16.3.1 决策树 206

16.3.2 贝叶斯概率分类.....207
16.3.3 关联规则.....207
16.3.4 聚类分析.....208
16.3.5 时序聚类.....209
16.3.6 线性回归.....210
16.3.7 Logistic 回归.....211
16.3.8 类神经网络.....212
16.3.9 时间序列.....213
16.4 DMX 应用范例.....214
16.4.1 分类.....215
16.4.2 估计.....216
16.4.3 预测.....217
16.4.4 关联分组.....217
16.4.5 聚类.....218

第 3 篇 其他分析工具介绍

第 17 章 分析关键影响因素.....223
第 18 章 检测类别.....228
第 19 章 从示例填充.....231
第 20 章 预测.....233
第 21 章 突出显示异常值.....235
第 22 章 应用场景分析.....238
22.1 目标查找.....238
22.2 假设.....240
第 23 章 Visio 2007 数据透视分析.....243

第 4 篇 数据挖掘范例

第 24 章 上市公司投资价值分析的挖掘模型.....251
24.1 研究动机与目的.....251
24.2 挖掘模型的构建.....251
24.3 变量筛选.....252
24.4 决策树模型.....253
24.5 贝叶斯概率模型.....255
24.6 Logistic 回归模型.....255
24.7 预测准确度比较.....256

第 25 章 信用卡用户信用评测的挖掘模型 259
25.1 研究背景 259
25.2 研究动机 260
25.3 研究目的 260
25.4 Excel 2007 构建数据挖掘模型 260
25.4.1 决策树分析 260
25.4.2 聚类分析 263
25.4.3 Logistic 回归 269
第 26 章 市场营销与客户细分的挖掘模型 271
26.1 研究动机与目的 271
26.2 研究方法与限制 271
26.3 数据分析 271
26.4 挖掘建模 273
26.4.1 决策树 273
26.4.2 单纯贝叶斯分类 280
26.4.3 聚类分析 282
26.4.4 决策树 286
26.4.5 Logistic 回归 288
26.4.6 关联分析 292
26.5 结论 295

第1篇

数据挖掘算法介绍

- 数据挖掘简介
- 数据挖掘运用的理论和技术
- 数据挖掘与相关领域的关系
- 数据挖掘商业软件产品及其应用现状

第1章

数据挖掘算法介绍

第 1 章　数据挖掘简介

1.1　数据挖掘的定义

Data mining is the process of seeking interesting or valuable information in large database.

数据挖掘（data mining）是近年来数据库应用领域中相当热门的话题。数据挖掘一般是指在数据库或数据仓库中，利用各种分析方法与技术，对过去累积的大量繁杂数据进行分析、归纳与整合等工作，提取出有用的信息，例如趋势（trend）、模式（pattern）及相关性（relationship）等，并将其中有价值的信息作为决策参考提供给决策者。通俗地说，数据挖掘就是从数据中发掘信息或知识，有人称为知识发现（knowledge discovery in database，KDD），也有人称为数据考古学（data archeology）、数据模式分析（data pattern analysis）或功能相依分析（functional dependency analysis）。目前，数据挖掘已经成为数据库系统、机器学习、统计方法等多个学科相互交叉的重要领域，而在实务界，越来越多的企业开始认识到，实施数据挖掘可以为企业带来更多潜在的商业机会。

但我们对数据挖掘应有一个正确的认知：数据挖掘不是一个无所不能的魔法。数据挖掘的种种工具都是从数据中发掘出各种可能成立的“预言”，并对其潜在价值加以“估计”，但数据挖掘本身并不能在实际中查证和确认这些假设，也不能判断这些假设的实际价值。

1.2　数据挖掘的重要性

现代企业经常会搜集大量的数据，这些数据涵盖了市场、客户、供货商，及其竞争对手等重要信息，但是由于信息超载与无结构化，企业的决策者无法充分利用这些庞大的数据资源，仅能使用其中的一小部分，这可能导致决策失误，甚至出现决策错误。而借助数据挖掘技术，企业完全有能力从浩瀚的数据海洋中，挖掘出全面而又有价值的信息和知识，并作为决策支持之用，进而形成企业独有的竞争优势。

1.3　数据挖掘的功能

一般而言，数据挖掘包括下列五项功能，这些功能大多为成熟的计量和统计分析方法。

1．分类（classification）

按照分析个体的属性状态分别加以区分，并建立类组（class）。例如，将信用申请者的风险等级分为高风险、中风险和低风险三类。使用的方法有决策树（decision tree）、判别分

析（discriminant analysis）、类神经网络（artificial neural network），以及记忆基础推理（memory-based reasoning）等。

2．估计（estimation）

根据已有的数值型变量和相关的分类变量，以获得某一属性的估计值或预测值。例如，根据信用卡申请者的教育程度和从事职业来设定其信用额度。使用的方法有相关分析、Logistic 回归及类神经网络等。

3．预测（prediction）

根据个体属性的已有观测值来估计该个体在某一属性上的预测值。例如，由顾客过去刷卡消费额预测其未来的刷卡消费额。使用的方法有回归分析、时间序列分析及类神经网络等。

4．关联分组（affinity grouping）

从所有对象决定哪些相关对象应该放在一起。例如，超市中相关的洗漱用品（牙刷、牙膏、牙线）放在同一货架上。在客户营销系统上，这类分析可以用来发现潜在的交叉销售（cross-selling）商品聚类，进而设计出有价值的组合商品集合。

5．同质分组（clustering）

将异质总体分成为同质性的类别（clusters），即聚类。其目的是识别出总体中所包含的混合类别的组间差异，并根据每个类别的特征对所有个体进行归类。同质分组相当于营销术语中的细分（segmentation）。应该注意的是：聚类分析根据数据自动产生各个类别，事先是不知道或无须知道总体中潜在的类别信息。使用的方法有 k-means 等动态聚类法及 agglomeration 等层次聚类法。

1.4 数据挖掘的步骤

数据挖掘的步骤会随不同领域的应用而有所变化，每一种数据挖掘技术也会有各自的特性和使用步骤，针对不同问题和需求所制定的数据挖掘过程也会存在差异。此外，数据的完整程度、专业人员支持的程度等都会对建立数据挖掘过程有所影响（蔡维欣，2003）。这些因素造成了数据挖掘在各不同领域中的运用、规划，以及流程的差异性，即使同一产业，也会因为分析技术和专业知识的涉入程度不同而不同，因此对于数据挖掘过程的系统化、标准化就显得格外重要。如此一来，不仅可以较容易地跨领域应用，也可以结合不同的专业知识，发挥数据挖掘的真正精神。

数据挖掘完整的步骤如下：

① 理解数据和数据的来源（understanding）。

② 获取相关知识与技术（acquisition）。

③ 整合与检查数据（integration and checking）。

④ 去除错误或不一致的数据（data cleaning）。

⑤ 建立模型和假设（model and hypothesis development）。

⑥ 实际数据挖掘工作（data mining）。

⑦ 测试和验证挖掘结果（testing and verfication）。

⑧ 解释和应用（interpretation and use）。

由上述步骤可看出，数据挖掘牵涉了大量的准备工作与规划工作，事实上许多专家都认为整套数据挖掘的过程中，有80%的时间和精力是花费在数据预处理阶段，其中包括数据的净化、数据格式转换、变量整合，以及数据表的链接。可见，在进行数据挖掘技术的分析之前，还有许多准备工作要完成。

1.5 数据挖掘建模的标准 CRISP-DM

CRISP-DM 是 Cross-Industry Standard Process for Data Mining 的简称，中文翻译为“数据挖掘的跨行业标准过程”。CRISP-DM 是由欧洲几家在数据挖掘应用上有经验的公司共同筹划组织的一个特别小组所提出的。该组织的成员包括数据仓储供货商 NCR、德国汽车航天公司 Daimler-Chrysler、统计分析软件供货商 SPSS 和荷兰的银行保险公司 OHRA，除了 NCR 与 SPSS 等是专注于数据挖掘软件开发的成员之外，也有其他众多厂商参与实验，通过实际操作过程，整体规划设计，并在 2000 年推出了 CRISP-DM 1.0 模型，把数据挖掘过程中必要的步骤都加以标准化。CRISP-DM 模型强调完整的数据挖掘过程，不能只针对数据整理、数据显示、数据分析以及构建模型，而应该将对企业的需求问题的理解，以及后期对模型的评价与模型的延伸应用都纳入到数据挖掘过程中。因此，CRISP-DM 从方法学的角度强调了实施数据挖掘项目的方法和步骤，同时独立于每种具体数据挖掘算法和数据挖掘系统。

CRISP-DM 分为六个阶段（phase）和四个层次（level），分别简介如下。

六个阶段如下。

1. 定义商业问题（business understanding）

本阶段的主要工作是要针对企业问题以及企业需求进行了解确认，针对不同的需求做深入的了解，将其转换成数据挖掘的问题，并拟定初步构想。在此阶段中，需要与企业各层次进行讨论，只有对要解决的问题有了非常清楚而全面的了解，才能正确地针对问题拟定分析过程。

2. 数据理解（data understanding）

此阶段包括建立数据库与分析数据。在这个阶段必须先收集数据，了解数据的含义与特性，并过滤出所有可能有用的数据，然后进行数据整理并评估数据的质量，必要时再将分属不同数据库的数据加以合并或整合。数据库建立完成后再进行数据分析，并找出影响最大的数据，进而判断是否有必要进一步收集更为详细的数据。

3. 数据预处理（data preparation）

此阶段和数据理解阶段为数据准备阶段的核心，这是建立模型前的最后一步数据准备