工程设计与分析系列

SAS 统计分析与数据挖掘

谢龙汉　尚　涛　编著

视频教学

★ SAS——全球最为著名的统计分析软件

★ SAS——金融、科研、数据挖掘领域必备之软件

★ 基础知识—实训实例—工程实例

★ 实例操作视频教学，轻松学习

電子工業出版社
PUBLISHING HOUSE OF ELECTRONICS INDUSTRY
http://www.phei.com.cn

工程设计与分析系列

SAS 统计分析与数据挖掘

谢龙汉　尚　涛　编著

電子工業出版社

内 容 简 介

本书基于 SAS 9.2 版本编写，从 SAS 编程出发，用案例形式介绍 SAS 数据挖掘在各领域的广泛应用。全书分为 SAS 基础篇、提高篇、应用篇，每章均给出大量分析案例，具体内容为 SAS 软件与数据挖掘简介，SAS 编程基础，图形与报表制作，描述性分析，假设检验，回归分析，方差分析与因子分析，相关分析与对应分析，判别分析，聚类分析，生存分析，时间序列分析，以及 SAS 在具体数据挖掘项目中的应用等。

本书最大特点是抛弃了其他同类书籍中只说理论、缺少案例分析的弊病，全书给出大量数据挖掘分析案例，为读者展示 SAS 在数据整合、数据挖掘、商业智能、金融数据分析、金融风险管理等项目中的强大应用技术。

本书读者对象为高等院校统计、信息等相关专业的本科生、研究生，科研单位的科技人员和企事业单位计算机工作者，以及数据分析、商业咨询、金融工程、商业智能工作者。

图书在版编目（CIP）数据

SAS 统计分析与数据挖掘/谢龙汉，尚涛编著. —北京：电子工业出版社，2012.1
（工程设计与分析系列）
ISBN 978-7-121-14888-0

Ⅰ. ①S… Ⅱ. ①谢… ②尚… Ⅲ. ①统计分析—应用软件，SAS Ⅳ. ①C819

中国版本图书馆 CIP 数据核字（2011）第 217233 号

策划编辑：许存权
责任编辑：陈韦凯　　特约编辑：刘丽丽
印　　刷：涿州市京南印刷厂
装　　订：涿州市桃园装订有限公司
出版发行：电子工业出版社
　　北京市海淀区万寿路 173 信箱　邮编　100036
开　　本：787×1 092　1/16　印张：29.25　　字数：749 千字
印　　次：2012 年 1 月第 1 次印刷
册　　数：4 000 册　　定价：59.00 元（含 DVD 光盘 1 张）

前　言

SAS 是当今国际最著名的数据分析软件。本书从 SAS 编程出发，用案例的形式介绍 SAS 数据分析在各个领域的广泛应用。

本书是在结合作者多年的使用经验和工作经验的基础上编写的，在编写过程中，突出了以下特点：

（1）直观易懂性。全书以图解实例的形式介绍基础知识和实例操作，所有的知识块和案例分析都尽可能详细，直观易懂，使读者能够在最短的时间内获取最多的知识。

（2）先进性。以最新的 SAS 9.2 版为蓝本进行讲解，广泛吸收国内外优秀教材的成果进行内容设计，在系统介绍基本理论和基本方法的同时，注意介绍新的成熟内容，以及统计学在实际问题中的应用。

（3）实用性。全书采用了基础知识介绍和实例操作相结合的方法，互相补充，本书的实例大多来源于经济生活之中，使读者在学完本书后能够快速地将知识应用于实践。

（4）结构清晰，讲解详尽。全书采用基础知识－程序实现－综合实例分析的讲解方法，循序渐进地提高读者的 SAS 编程知识，而且每个知识点和实例都做了尽可能详细的讲解，使读者学习起来轻松自如。

（5）配有全部的案例数据、程序与多媒体示范。本书的配套光盘中提供了所有实例的数据、SAS 程序、视频操作  ，读者可以在观看录像中增强对知识点的理解。

本书共分为 21 章，依次介绍了 SAS 9.2 的基本编程知识、基本统计分析、高级统计分析、SAS 数据挖掘、SAS 在金融中的应用，以及各章节中的案例分析等内容。

第 1 章　数据挖掘概述。介绍数据挖掘的涵义，为什么要进行数据挖掘，数据挖掘的用途、过程等，最后介绍 SAS 软件系统在数据挖掘中的地位。

第 2 章　SAS 模块概述。详细介绍 SAS 软件系统和 SAS 各个常用的模块。

第 3 章　SAS 程序设计基础。详细介绍 SAS 编程中的各个步骤，包括 SAS 安装过程，SAS 数据步和过程步，程序调试及 SAS 函数，掌握 SAS 编程基础。

第 4 章　数据预处理。内容包括数据输入、数据整理、数据步变量控制、数据修改与选择，最后介绍 SAS 与 SPSS 软件之间的数据转换。

第 5 章　数据汇总与报表制作。介绍最基本的数据报表生成过程，PROC PRINT 和 PROC TABULATE 过程。

第 6 章　SAS 绘图。统计图是统计描述的重要工具，本章介绍 SAS 绘制图形的 GPLOT、GCHART、G3D 过程。

第 7 章　数据描述。介绍利用统计图、统计量和数据分布进行数据描述的过程。

第 8 章　描述性统计分析。叙述描述性统计分析中的平均数、中位数、众数等度量集中趋势，用极差、标准差、变异系数等度量。

第 9 章　ANALYST 模块。主要介绍利用 ANALYST 模块进行数据管理。

第 10 章　参数估计和假设检验。介绍对于定量资料的统计描述和简单推断，主要有 UNIVARIATE、MEANS、TTEST 过程。

第 11 章　方差分析与协方差分析。介绍方差分析与协方差分析的基本原理及其 SAS 过程。

第 12 章　回归分析。介绍线性回归、REG 过程、多项式回归、逐步回归 LOGISTIC 过程及非线性回归。

第 13 章　主成分分析与因子分析。介绍主成分分析与因子分析的原理、数学模型及 SAS 过程，然后应用于案例分析。

第 14 章　相关分析和对应分析。介绍相关分析和对应分析的原理、数学模型及案例分析。

第 15 章　判别分析。介绍判别分析的最大似然法、Fisher 判别分析法、Bayes 判别分析法、逐步判别分析法等，并应用于案例分析。

第 16 章　聚类分析。介绍聚类分析的基本原理，以及 CLUSTER、FASTCLUS、VARCLUS、TREE 过程。

第 17 章　生存分析。介绍 SAS 对生存分析的应用，并进行研究探讨。

第 18 章　时间序列分析。介绍时间序列的数学模型，SAS 的 ARIMA 过程及案例分析。

第 19 章　SAS 数据挖掘应用。主要介绍 SAS 数据挖掘方法论——SEMMA，SAS 的企业数据挖掘套件 SAS/EM 及其案例分析。

第 20 章　SAS 在数据预测中的应用。介绍利用 SAS 软件系统对有关数据预测的一些案例进行分析研究，为读者展示数据预测的相关技术和方法。

第 21 章　SAS 在金融数据分析中的应用。介绍利用 SAS 软件对部分金融学中数学模型进行分析求解，展示 SAS 软件在金融数据挖掘中的应用。

本书主要由尚涛完成，香港中文大学谢龙汉博士提供技术支持和指导，并对本书进行了校对和完善。参加本书编写和光盘开发的还有林伟、魏艳光、林木议、王悦阳、林伟洁、林树财、郑晓、吴苗、李翔、莫衍、朱小远、唐培培、耿煜、邓奕、张桂东、鲁力、刘文超、刘新东等，同时也非常感谢腾龙工作室其他成员的帮助和支持。

由于时间仓促，书中难免有疏漏之处，请读者谅解。读者可通过电子邮件reader.toptech@gmail.com或者 reader.toptech @163.com 与我们交流。

编著者

目　录

第1章　数据挖掘概述

每个企业每天都会产生大量的数据，这些数据来自于不同的数据源。数据挖掘是对大量的原始数据进行选择、分析和建模，从中发现以前没有发现的趋势和模式，通过数据和文本挖掘得到的信息对企业战略决策有很大帮助。因此，数据挖掘使得人们从原始数据到更加明智的业务决策。本章将详细叙述数据挖掘的过程、工具、用法，以及 SAS 软件系统在数据挖掘中的位置等，为利用 SAS 软件系统进行数据挖掘打下基础。

本章内容

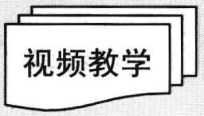
- 数据挖掘简介
- 数据挖掘用途
- 数据挖掘过程
- SAS——数据挖掘领域的领导者
- SAS 在各种商业解决方案中应用

1.1　数据挖掘简介

数据挖掘，简言之就是从大量繁杂的数据中获取隐含其中的信息，如对顾客分类、聚类、欺诈甄别、潜在顾客识别等，现在应用领域很广，如设计、零售、金融、银行、医疗、政府决策、企业财务、商业决策。下面将从数据挖掘的概念、起源等说起。

1.1.1　数据挖掘的含义

数据挖掘（Data Mining）就是从大量数据中获取有效的、新颖的、潜在有用的、最终可理解的模式的非平凡过程。美国 SAS 软件研究所将数据挖掘定义为“数据挖掘是按照既定的业务目标，对大量的企业数据进行探索、揭示隐藏其中的规律性并进一步模型化的先进、有效的方法。”数据挖掘的广义观点：数据挖掘就是从存放在数据库，数据仓库或其他信息库中的大量的数据中“挖掘”有趣知识的过程。数据挖掘又称数据库中知识发现（Knowledge Discovery in Database, KDD），也有人把数据挖掘视为数据库中知识发现过程的一个基本步骤。知识发现过程由以下步骤组成：



- 数据清理；
- 数据集成；
- 数据选择；
- 数据变换；
- 数据挖掘；
- 模式评估；
- 知识表示。

数据挖掘可以与用户或知识库交互。

并非所有的信息发现任务都被视为数据挖掘。例如，使用数据库管理系统查找个别的记录，或通过互联网的搜索引擎查找特定的 Web 页面，则是信息检索（Information Retrieval）领域的任务。虽然这些任务是重要的，可能涉及使用复杂的算法和数据结构，但是它们主要依赖传统的计算机科学技术和数据的明显特征来创建索引结构，从而有效地组织和检索信息。尽管如此，数据挖掘技术也已用来增强信息检索系统的能力。

1.1.2 数据挖掘的起源

需要是发明之母。近年来，数据挖掘引起了信息产业界的极大关注，其主要原因是存在大量数据，可以广泛使用，并且迫切需要将这些数据转换成有用的信息和知识。获取的信息和知识可以广泛用于各种应用，包括商务管理、生产控制、市场分析、工程设计和科学探索等。

在国内，数据挖掘应用最多的行业也是数据积累最快的行业，如电信、金融、保险、互联网、零售等行业。

数据挖掘利用了来自如下一些领域的思想：

- 来自统计学的抽样、估计和假设检验；
- 人工智能、模式识别和机器学习的搜索算法、建模技术和学习理论。

数据挖掘也迅速地接纳了来自其他领域的思想，这些领域包括最优化、进化计算、信息论、信号处理、可视化和信息检索，一些其他领域也起到重要的支撑作用。特别的是，需要数据库系统提供有效的存储、索引和查询处理支持。源于高性能（并行）计算的技术在处理海量数据集方面常常是重要的。分布式技术也能帮助处理海量数据，并且当数据不能集中到一起处理时更是至关重要。

1.1.3 统计学与数据挖掘

统计学和数据挖掘有着共同的目标：发现数据中的结构。事实上，由于它们的目标相似，有些人（尤其是统计学家）认为数据挖掘是统计学的分支。这是一个不切合实际的看法。因为数据挖掘还应用了其他领域的思想、工具和方法，尤其是计算机学科，例如，数据库技术和机器学习，而且它所关注的某些领域和统计学家所关注的有很大不同。

1.1.3.1 统计学的性质

试图为统计学下一个太宽泛的定义是没有意义的。尽管可能做到，但会引来很多异

议。相反，更要关注统计学不同于数据挖掘的特性。

差异之一同 1.1.2 节中最后一段提到的相关，即统计学是一门比较保守的学科，目前有一种趋势是越来越精确。当然，这本身并不是坏事，只有越精确，才能避免错误，发现真理。但是，如果过度的话，则是有害的。这个保守的观点源于统计学是数学的分支这样一个看法，尽管统计学确实以数学为基础（正如物理和工程也以数学为基础，但没有被认为是数学的分支），但它同其他学科还有紧密的联系。

数学背景和追求的确加强了这样一个趋势：在采用一个方法之前先要证明，而不是像计算机科学和机器学习那样注重经验。这就意味着有时候和统计学家关注同一问题的其他领域的研究者提出一个很明显有用的方法，但它却不能被证明（或还不能被证明）。统计杂志倾向于发表经过数学证明的方法而不是一些特殊方法。数据挖掘作为几门学科的综合，已经从机器学习那里继承了实验的态度。这并不意味着数据挖掘工作者不注重精确，而只是说明如果方法不能产生结果的话就会被放弃。

正是统计文献显示了统计的数学精确性。同时还显示了其对推理的侧重。尽管统计学的一些分支也侧重于描述，但是浏览一下统计论文的话，就会发现这些文献的核心问题就是在观察了样本的情况下如何去推断总体。当然这常常也是数据挖掘所关注的。下面会提到数据挖掘的一个特定属性，就是要处理的是一个大数据集。这就意味着，由于可行性的原因，常常得到的只是一个样本，但是需要描述样本取自的那个大数据集。然而，数据挖掘问题常常可以得到数据总体，例如，关于一个公司的所有职工数据，数据库中的所有客户资料，去年的所有业务。在这种情形下，推断就没有价值了（如年度业务的平均值），因为观测到的值也就是估计参数。这就意味着，建立的统计模型可能会利用一系列概率表述（如一些参数接近于 0，则会从模型中剔除掉），但当总体数据可以获得的话，在数据挖掘中则变得毫无意义。在这里，可以很方便地应用评估函数：针对数据的足够的表述。事实是，常常所关注的是，模型是否合适而不是它的可行性，在很多情形下，使得模型的发现很容易。例如，在寻找规则时常常会利用吻合度的单纯特性（如应用分支定理），但当应用概率陈述时则不会得到这些特性。

统计学和数据挖掘部分交迭的第三个特性是在现代统计学中起核心作用的“模型”。“模型”这个术语更多的含义是变化的。一方面，统计学模型是基于分析变量间的联系，但另一方面这些模型关于数据的总体描述确实没有道理。关于信用卡业务的回归模型可能会把收入作为一个独立的变量，因为一般认为高收入会导致大的业务。这可能是一个理论模型（尽管基于一个不牢靠的理论）。与此相反，只需在一些可能具有解释意义的变量基础上进行逐步的搜索，从而获得一个有很大预测价值的模型，尽管不能做出合理的解释（通过数据挖掘去发现一个模型的时候，常常关注的就是后者）。

还有其他方法可以区分统计模型，但在这里将不做探讨。这里关注的是，现代统计学是以模型为主的。而计算、模型选择条件是次要的，只是如何建立一个好的模型。但在数据挖掘中，却不完全是如此。在数据挖掘中，准则起了核心的作用。在很多情形下，模型的选择并不都是显而易见的，选择一个合适的模型是不可能的，最合适的计算方法也是不可行的。在这种情形下，从另外一个角度出发，应用设计的一系列技术来回答 MVA 问题，暂不考虑模型和最优判别的选择。

相对于统计学而言，准则在数据挖掘中起着更为核心的作用并不奇怪，数据挖掘所继承的学科如计算机科学及相关学科。数据集的规模常常意味着传统的统计学准则不适合数据挖掘问题，不得不重新设计。尽管一些统计学的准则已经得到发展，但更多的应用是机器学习。

1.1.3.2 数据挖掘的性质

由于统计学基础的建立在计算机的发明和发展之前，所以，常用的统计学工具包含很多可以手工实现的方法。因此，对于很多统计学家来说，1000 个数据就已经是很大的了。但这个“大”对于英国大的信用卡公司每年 3.5 亿笔业务或 AT&T 每天 2 亿个长途呼叫来说相差太远了。很明显，面对这么多的数据，则需要设计不同于那些“原则上可以用手工实现”的方法。这意味着计算机（正是计算机使得大数据可能实现）对于数据的分析和处理是关键的，分析者直接处理数据将变得不可行。相反，计算机在分析者和数据之间起到了必要的过滤的作用，这也是数据挖掘特别注重准则的另一原因。尽管有必要，把分析者和数据分离开很明显导致了一些关联任务。这里就有一个真正的危险：非预期的模式可能会误导分析者，这一点下面会讨论。

在现代统计中计算机是一个重要的工具，它们确实是，并不是因为数据的规模。对数据的精确分析方法（如 Bootstrap 方法、随机测试、迭代估计方法）及比较适合的复杂的模型，正是有了计算机才是可能的。计算机已经使得传统统计模型的视野大大的扩展了，还促进了新工具的飞速发展。

统计学很少会关注实时分析，然而数据挖掘问题常常需要这些。例如，银行事务每天都会发生，没有人能等 3 个月得到一个可能的欺诈的分析。类似的问题发生在总体随时间变化的情形。我的研究组有明确的例子，显示银行债务的申请随时间、竞争环境、经济波动而变化。

1.1.3.3 讨论

主要有几个问题：

① 数据挖掘有时候是一次性的实验，这是一个误解。它更应该被看作是一个不断的过程（尽管数据集是确定的）。从一个角度检查数据可以解释结果，以相关的观点检查可能会更接近等。关键是，除了极少的情形下，很少知道哪一类模式是有意义的。数据挖掘的本质是发现非预期的模式，同样非预期的模式要以非预期的方法来发现。

② 与把数据挖掘作为一个过程的观点相关联的是认识到结果的新颖性，许多数据挖掘的结果是所期望的。然而，可以解释这个事实并不能否定挖掘出它们的价值。没有这些实验，可能根本不会想到这些。实际上，只有那些可以依据过去经验形成的合理的解释的结构才会是有价值的。

③ 在大数据集中发现模式的可能性当然存在，然而，也不应就此掩盖危险：所有真正的数据集（即使那些是以完全自动方式搜集的数据）都有产生错误的可能。关于人的数据集（如事务和行为数据）尤其有这种可能。

1.1.4 数据挖掘相关的一些问题

1.1.4.1 数据挖掘和统计分析

非要去区分数据挖掘和统计的差异其实是没有太大意义的。一般将其定义为数据挖掘技术的 CART（Classification and Regression Trees）、CHAID（Chi-Square Automatic Interaction Detector）或模糊计算等理论方法，也都是由统计学者根据统计理论所发展衍生。换另一个角度看，数据挖掘有相当大的比重是由高等统计学中的多变量分析所支撑。但是为什么数据挖掘的出现会引发各领域的广泛关注呢？主要原因在于相对传统统计分析而言，数据挖掘有下列几项特性：

① 处理大量实际数据更强势，且无须专业的统计背景去使用数据挖掘的工具；

② 数据分析趋势为从大型数据库获取所需数据并使用专属计算机分析软件，数据挖掘的工具更符合企业需求；

③ 就纯理论的基础点来看，数据挖掘和统计分析有应用上的差别，毕竟数据挖掘目的是方便企业终端用户使用而并非给统计学家检测用的。

1.1.4.2 数据挖掘和数据仓库

若将数据仓库（Data Warehousing）比作“矿坑”，数据挖掘就是深入矿坑采矿的工作。毕竟数据挖掘不是一种无中生有的魔术，也不是点石成金的炼金术。若没有足够丰富完整的数据，是很难期待数据挖掘能挖掘出什么有意义的信息的。

要将庞大的数据转换成为有用的信息，必须先有效率地收集信息。随着科技的进步，功能完善的数据库系统就成了最好的收集数据的工具。数据仓库，简单地说，就是收集来自其他系统的有用数据，存放在一个整合的储存区内。所以，其实就是一个经过处理整合，且容量特别大的关系型数据库，用于储存决策支持系统（Design Support System，DSS）所需的数据，供决策支持或数据分析使用。从信息技术的角度来看，数据仓库的目标是在组织中，在正确的时间，将正确的数据交给正确的人。

许多人对于数据仓库和数据挖掘时常混淆，不知如何分辨。其实，数据仓库是数据库技术的一个新主题，利用计算机系统帮助操作、计算和思考，让作业方式改变，决策方式也跟着改变。

数据仓库本身是一个非常大的数据库，它储存着由组织作业数据库中整合而来的数据，特别是指 OLTP（On-Line Transactional Processing，事务处理系统）所得来的数据。将这些整合过的数据置放于数据仓库中，而公司的决策者则利用这些数据作决策；但是，这个转换及整合数据的过程，是建立一个数据仓库最大的挑战。因为将作业中的数据转换成有用的策略性信息是整个数据仓库的重点。

综上所述，数据仓库应该具有这些数据：整合性数据（Integrated Data）、详细和汇总性的数据（Detailed and Summarized Data）、历史数据、解释数据的数据。从数据仓库挖掘出对决策有用的信息与知识，是建立数据仓库与使用数据挖掘的最大目的，两者的本质与过程是两回事。换句话说，数据仓库应先行建立完成，数据发掘才能有效率的进行，因为数据仓库本身所含数据是干净（不会有错误的数据参杂其中）、完备，且经过整合的。因此，两者之间的关系或许可解读为数据挖掘是从巨大数据仓库中找出有用信息的一种过程与技术。

1.1.4.3 数据挖掘和 OLAP

OLAP（Online Analytical Process）意指由数据库所连接出来的在线分析处理程序。有些人会说：已经有 OLAP 的工具了，所以，不需要数据挖掘。事实上两者间是截然不同的，主要差异在于数据挖掘用在产生假设，OLAP 则用于查证假设。简单来说，OLAP 是由使用者所主导，使用者先有一些假设，然后利用 OLAP 来查证假设是否成立；而数据挖掘则是用来帮助使用者产生假设。所以，在使用 OLAP 或其他 Query 的工具时，使用者是自己在做探索（Exploration），但数据挖掘是用工具在帮助做探索。

例如，市场分析师在为超市规划货品架柜摆设时，可能会先假设婴儿尿布和婴儿奶粉会是常被一起购买的产品，接着便可利用 OLAP 的工具去验证此假设是否为真，又成立的证据有多明显；但数据挖掘则不然，执行数据挖掘的人将庞大的结账数据整理后，并不需要假设或期待可能的结果。透过数据挖掘技术可找出存在于数据中的潜在规则，于是可能得到（如尿布和啤酒常被同时购买的）意料之外的发现，这是 OLAP 所做不到的。

数据挖掘常能挖掘出超越归纳范围的关系，但 OLAP 仅能利用人工查询及可视化的报表来确认某些关系，是以数据挖掘此种自动找出甚至不会被怀疑过的数据模型与关系的特性。事实上已超越了经验、教育、想象力的限制，OLAP 可以和数据挖掘互补，但这项特性是数据挖掘无法被 OLAP 取代的。

1.1.4.4 数据挖掘步骤

以下提供一个数据挖掘的进行步骤作为参考：

- 理解业务与理解数据；
- 获取相关技术与知识；
- 整合与查询数据；
- 去除错误或不一致及不完整的数据；
- 由数据选取样本先行试验；
- 建立数据模型；
- 实际数据挖掘的分析工作；
- 测试与检验；
- 找出假设并提出解释；
- 持续应用于企业流程中。

由上述步骤看出，数据挖掘牵涉了大量的准备工作与规划过程，事实上许多专家皆认为整套数据挖掘的进行有 80%的时间精力是花费在数据前置作业阶段，其中包含数据的净化与格式转换甚至表格的连接。由此可知，数据挖掘只是信息挖掘过程中的一个步骤而已，在进行此步骤前还有许多的工作要先完成。

1.1.4.5 数据挖掘理论

数据挖掘是近年来数据库应用技术中相当热门的议题，看似神奇、听来时髦，实际上却也不是什么新东西。因其所用如预测模型、数据分割、连结分析（Link Analysis）、偏差侦测（Deviation Detection）等，美国早在第二次世界大战前就已应用，运用在人口普查及军事等方面。