信息科学与技术丛书

姚志勇　编著

SAS编程与数据挖掘商业案例

- 从 PDV 角度详尽剖析 Base SAS 常用语句代码及应用
- 数据挖掘理论和商业应用紧密结合
- 原创朴素贝叶斯文本分类和 EM 迭代算法代码
- 三个典型的数据挖掘商业案例分析

机械工业出版社
CHINA MACHINE PRESS

信息科学与技术丛书

SAS 编程与数据挖掘商业案例

姚志勇　编著

机械工业出版社

本书是作者多年来在企业实践工作中的经验总结，详细讲解了使用SAS进行商业数据挖掘的方法。其中包含了目前公开出版的诸多SAS教材没有的大量实战内容。

本书内容全面、新颖独创、综合性强，适合企业人员使用，也可作为数学、统计学、金融、电子商务、医药等专业的本科生、硕士生学习SAS编程和数据挖掘的参考资料。

读者可在http://www.cmpbook.com下载书中的SAS程序及相关测试数据集。

图书在版编目（CIP）数据

SAS编程与数据挖掘商业案例/姚志勇编著 .—北京：机械工业出版社，2010.5

（信息科学与技术丛书）

ISBN 978-7-111-30535-4

Ⅰ.①S… Ⅱ.①姚… Ⅲ.①商业统计－统计分析－应用软件，SAS Ⅳ.①F712.3

中国版本图书馆CIP数据核字（2010）第078039号

机械工业出版社（北京市百万庄大街22号 邮政编码100037）

策划编辑：车 忱

责任编辑：车 忱

责任印制：乔 宇

北京机工印刷厂印刷（北京振兴源印务有限公司装订）

2010年5月第1版·第1次印刷

184mm×260mm·22.25印张·551千字

0 001—4 000册

标准书号：ISBN 978-7-111-30535-4

定价：42.00元

凡购本书，如有缺页、倒页、脱页，由本社发行部调换

电话服务

社服务中心：（010）88361066

销 售 一 部：（010）68326294

销 售 二 部：（010）88379649

读者服务部：（010）68993821

网络服务

门户网：http://www.cmpbook.com

教材网：http://www.cmpedu.com

封面无防伪标均为盗版

前　言

当前国内的诸多数据挖掘书籍几乎都是基于理论说明，很少深入介绍数据挖掘实践，涉及 SAS 开发的更是少见。因此，从商业应用出发，基于实践而不是基于理论的数据挖掘书籍呼之欲出。本书作者从商业需求出发，以商业人士的眼光来看待企业数据挖掘，并给出大量的商业实践案例。把主流的数据挖掘技术用真实案例来实现是本书出版的初衷，同时为了满足初学者需求，作者也给出了数据挖掘必备的基础编程知识模块。

全书共分两部分。

第一部分是 SAS 编程：

第 1 章和第 2 章主要介绍 SAS 系统和编程基础，同时介绍 SAS 数据处理最核心的内容——数据指针和 PDV 流程。该核心内容贯穿第一部分，是已出版的其他 SAS 图书没有的。

第 3～9 章主要介绍 SAS 的数据处理技术，也是第一部分的主要内容，包括数据集处理、变量处理和观测处理等多种数据处理技术，同时也介绍了循环控制等稍难的内容，重要的是给出了诸多实际案例及商业应用。尽管第 3～9 章从表面上看和诸多已经出版的 SAS 图书没有什么大的不同，但是这些章节最大的亮点是作者对每一个示例和案例从数据指针和 PDV 流程的角度给予了最详细的程序解读，让读者真正读懂程序，而不是停留在程序的表面。

第 10 章是第一部分的难点。作者还是站在商业实践的角度逐一介绍宏最常用的部分，同时也给出了非常详细的程序解读。

第 11 章介绍 SQL 过程。有关内容在国内同类书中都出现过，但是作者独辟蹊径，融合了项目实践中诸多真正有用的语句，同时也给出了诸多开发建议和应注意的问题。

第 12 章介绍数据处理实践。该章共包括四个方面的内容，几乎都是目前国内没有出现过的，如 HASH 对象及商业应用、正则表达式等。随机抽样也是数据处理经常面临的问题，这里作者开发了在 SAS 系统中如何处理分层不等比例抽样的代码，这也是目前国内其他 SAS 图书没有介绍过的。

第二部分是数据挖掘商业案例：

第 13 章主要介绍数据挖掘概念和流程。数据挖掘流程尤其是商业流程是本章的重点。该流程告诉读者一个真正的商业数据挖掘流程在商业环境中是如何实施的。

第 14 章重点介绍响应模型。响应模型是商业实践中最常用的预测模型，基于第 13 章的流程规范给出了一个具体的商业案例研究。

第 15 章是客户行为分析。该章有目前全球最流行的行为分析，包括“行为年龄”和“行为性别”（注意完全不同于具有自然属性特征的“真实年龄”和“真实性别”），作者运用 Naïve Bayesian 技术开发出一整套模型，并对该模型拥有完全自主知识产权。

第 16 章介绍文本挖掘。该章首先介绍了文本挖掘的流程，然后开发出基于 Naïve Bayesian 文本分类算法和 EM 迭代思想的大型代码，并成功应用于商业实践。

本书特色如下：

（1）国内少有的由商界人士编写的 SAS 数据挖掘图书。书中有近一半的内容是目前国

出 版 说 明

随着信息科学与技术的迅速发展，人类每时每刻都会面对层出不穷的新技术和新概念。毫无疑问，在节奏越来越快的工作和生活中，人们需要通过阅读和学习大量信息丰富、具备实践指导意义的图书来获取新知识和新技能，从而不断提高自身素质，紧跟信息化时代发展的步伐。

众所周知，在计算机硬件方面，高性价比的解决方案和新型技术的应用一直备受青睐；在软件技术方面，随着计算机软件的规模和复杂性与日俱增，软件技术不断地受到挑战，人们一直在为寻求更先进的软件技术而奋斗不止。目前，计算机在社会生活中日益普及，随着Internet延伸到人类世界的方方面面，掌握计算机网络技术和理论已成为大众的文化需求。由于信息科学与技术在电工、电子、通信、工业控制、智能建筑、工业产品设计与制造等专业领域中已经得到充分、广泛的应用。所以这些专业领域中的研究人员和工程技术人员越来越迫切需要汲取自身领域信息化所带来的新理念和新方法。

针对人们了解和掌握新知识、新技能的热切期待，以及由此促成的人们对语言简洁、内容充实、融合实践经验的图书迫切需要的现状，机械工业出版社适时推出了“信息科学与技术丛书”。这套丛书涉及计算机软件、硬件、网络和工程应用等内容，注重理论与实践的结合，内容实用、层次分明、语言流畅，是信息科学与技术领域专业人员不可或缺的参考书。

目前，信息科学与技术的发展可谓一日千里，机械工业出版社欢迎从事信息技术方面工作的科研人员、工程技术人员积极参与我们的工作，为推进我国的信息化建设作出贡献。

机械工业出版社

内没有出现过的。同时克服了诸多国内 SAS 书籍只翻译（如 SAS 帮助文档）而不予解释，或者即便解释也只是停留在程序表面的严重弊端，从程序的内在运行流程并基于作者多年的商业实践给予程序代码更详细到位的解析。

（2）国内首次用 SAS 编程开发 Naïve Bayesian 分类算法和 EM 迭代（原创代码）。

（3）国内首次把数据挖掘理论和商业应用相结合，给出三个非常典型的数据挖掘项目案例，并配有大型的算法代码。

（4）国内首次介绍非常实用的抽样技术。

（5）国内首次介绍 SAS 的一个重要组件——HASH 对象，并给出商业实践案例。

本书按照惯例，SAS 的关键字在代码中使用小写，而在正文中使用大写。

本书适合多层次多专业的人士阅读，如数学、统计学、经济学、保险和商业管理等专业的本科生、研究生及相关从业人员。希望这是一本让 SAS 用户和数据挖掘工作者都非常喜欢的书。

限于作者的水平，并且由于数据挖掘涉及领域非常广泛，所以本书只是把目前主流的数据挖掘付诸实践，今后如有机会，将不断更新和改进。

姚志勇

目　录

第1章 SAS系统简介

1.1 系统简介

SAS 系统是一个面向对象的、跨平台的、模块化和组合化的应用软件系统，具有完备的数据存取、数据管理、数据分析和数据展现功能，特别是在数据处理和统计分析领域，SAS 系统至今仍然被誉为全球标准软件系统。

1.1.1 SAS 系统与商务智能系统

随着商务智能的日益普及，传统的 SAS 系统也已经进化到与商务智能和数据挖掘保持同步，基于商务智能的商务解决方案已成为 SAS 产品的重要组成部分。传统的 SAS 系统已经演化成服务端的一个模块基础，SAS BI 通过包含 SAS 基础在内的所有服务模块，并以 Web 等作为中间媒介，最终提供给客户包括联机分析系统 OLAP、客户端分析软件 EG 和客户端报表工具 Portal 等在内的一揽子商务解决方案，如图 1-1 所示。

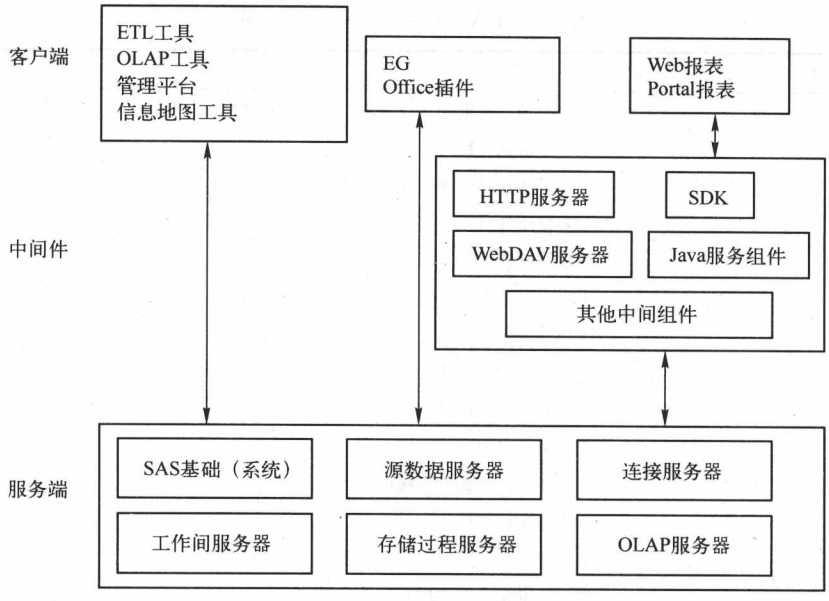

图 1-1 SAS 智能商务平台

本书内容主要集中在 SAS 基础（系统）模块，并未涉及其他任何模块。

1.1.2 SAS 系统与其他数据库的数据交换

SAS 系统提供了多种与现行数据库（如主流数据库 TERADATA、DB2、Oracle）的接

口，以及其他应用类软件（如 SPSS 等），如图 1-2 所示。所以在数据获取上除了自身产生的数据外，还可以通过与现行数据库的接口实现从其他数据库获取数据。

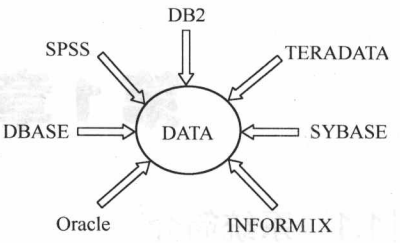

图 1-2　数据库交换系统

在商业实践中，从外部数据库获取数据的频率要远远大于 SAS 本身产生数据的频率，主要原因是商业数据往往都是海量的，一般情况下，需要处理的观测数都会在百万级以上。而 SAS 本身虽然具有建立逻辑数据库的机制，但是与真正意义上的数据库还存在一定差距。比如，对数据用户登录机制、查询优化机制、索引机制、数据存储机制、安全机制等都没有主流数据库（如 DB2、TERADATA 等）高效。因此，提供与这些主流数据库的接口是 SAS 系统处理大型数据的必然选择。

1.1.3　SAS 语言与 SAS 系统

SAS 语言是 SAS 系统的基础，是用户和系统直接对话的语言，其特点是用户不必告诉 SAS“怎么做”，只要告诉它“做什么”就行了。

SAS 语言结构主要由 DATA 和 PROC 两个基本步骤任意组合而成。其中，DATA 步完成对数据的获取、加工和处理；PROC 步用于数据分析和输出报告。

SAS 系统包含很多软件模块，见表 1-1。

表 1-1　SAS 主要软件模块及功能

软件模块	功　能
BASE	数据管理、基础统计、报表生成和图形显示
STAT	统计分析软件包
ETS	计量经济和时间序列分析软件包
OR	运筹学软件包
QC	质量控制软件包
IML	矩阵语言运算软件包
GRAPH	图形软件包
FSP	数据处理的交互式菜单系统
AF	应用开发软件包
ASSIST	菜单驱动界面
ACCESS	与其他数据库接口的接口集
EIS	应用开发软件包
INSIGHT	可视化数据探索软件包
SHARE	进行数据库并发式控制的软件包
CONNECT	分布处理不同操作平台上的 SAS 系统

SAS 商务智能所有的解决方案及其产品都是以上面所列软件模块作为基础和核心的。而其中的 BASE 又是 SAS 基础（系统）的核心，主要功能是数据管理和数据处理，并有报表输出和基本统计量输出功能。STAT 是 SAS 的统计分析软件算法包，包括方差分析、回归分析、多变量分析、非参数分析等共 50 多个 PROC 步，每个 PROC 步提供多种不同的算法选

择，从而组成一个完整的统计分析方法集。历经 40 多年，至今该软件仍然是全球统计分析领域的领头羊。

BASE 模块和 STAT 模块是本书重点介绍的内容。

1.1.4 SAS 9 浏览窗口简介

SAS 9 是一个集成了几乎所有 SAS 软件模块的集成软件。在 Windows 操作系统环境下，如果安装了 SAS 软件，就可以在开始程序菜单里面看到快捷方式。单击快捷方式后，打开的主界面包括浏览器窗口、日志窗口、编程窗口以及通用的菜单栏和工具条，如图 1-3 所示（SAS 9.2）。

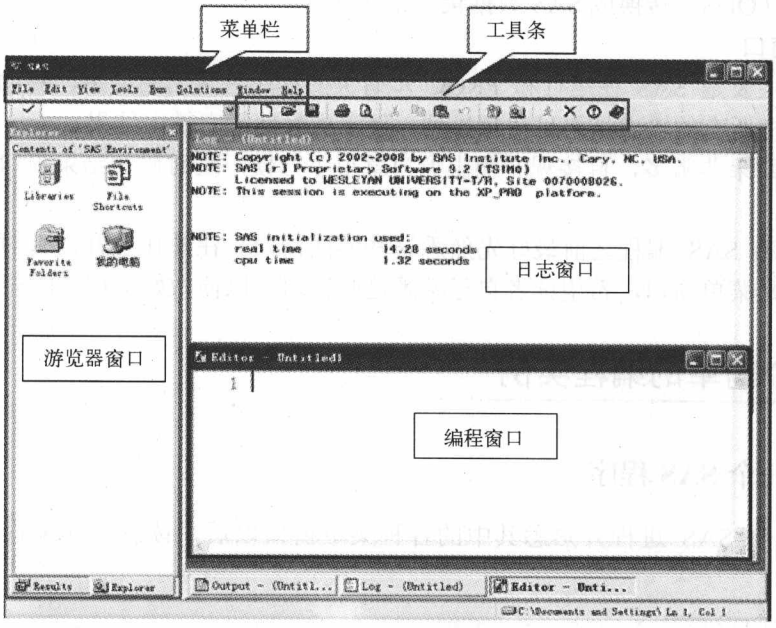

图 1-3 SAS 操作界面

1. 菜单栏

菜单栏由多个子菜单组成，不同的子菜单实现不同的功能。所以，激活不同的子菜单，窗口菜单的内容也会不同。与菜单栏功能相同的是弹出菜单，只要在当前子窗口右击鼠标即可。

2. 工具条

工具条包含了若干按钮，每一个按钮代表一个常用的命令，如保存程序、运行程序等。用户可以在菜单栏 Tools 子菜单中的 Customize 中设定工具条，非常简单。

3. 日志窗口

日志窗口可显示程序运行过程中的所有信息。例如，图 1-3 中日志窗口的文字就是在 SAS 启动后看到的日志信息。日志窗口对程序的调试起着非常重要的作用，因为程序员在开发程序的过程中，很少有一次就能开发所有的代码并运行成功的，一般都需要调试若干次才获得通过，这时所有调试出现的信息包括程序出错的信息都出现在日志窗口中。通过日志窗口，一般就能知道程序出错的原因，并找出相应的 Bug。

4．浏览器窗口

单击图 1-3 中左下角的 Explorer 按钮，就可以显示对应的浏览器窗口，包含四个对象，其中最常用的是 Library 逻辑库对象。

5．编程窗口

这是最重要的人机交互界面。一般来说，用户所有的开发代码都是在这个窗口完成和调试的。

6．输出窗口

输出窗口主要显示 SAS 在运行和 PROC 步有关的程序时输出的结果，如包含一些统计量的输出，一些报表的输出等。后面将专门提到，输出窗口的输出结果还可以通过 SAS 的文件输出系统（ODS）转换成 SAS 数据集等格式文件。

7．结果窗口

结果窗口主要是 SAS 在运行和 PROC 步有关的程序时输出的结果，和输出窗口不同的是，结果窗口把所有的输出结果呈树状排列，便于更好地检索输出结果。

如果输出结果非常多，直接观看输出窗口就很不方便，这时候在结果窗口中检索输出会更加便捷。

读者在熟悉 SAS 编程之前最好先熟悉这些界面窗口。在此作者并没有对这些窗口做详细介绍，特别是菜单窗口，希望读者自行熟悉这些窗口，以便做好编程前的充分准备工作。

1.2 一个简单的编程实例

1.2.1 编写一个 SAS 程序

在打开一个 SAS 进程并熟悉其中的各种菜单窗口以后，读者可在编程窗口输入如下代码：

```
data test;
input x y z ;
cards;
1 2 3
4 5 6
;
run;
```

这段程序仅包含七行代码：

第 1 行：产生数据集的 DATA 步语句。后面的 test 是要产生的 SAS 数据集。

第 2 行：输入变量的 INPUT 语句。该语句表明 test 数据集包含三个变量 x、y 和 z。

第 3 行：输入数据的 CARDS 语句。其后接具体的输入数据。

第 4，5 行：数据块。是要输入的具体数据。

第 6 行：分号。分号是 SAS 不同语句之间间隔的标志。

第 7 行：RUN 语句。表明到此为止就可以提交上面六行语句。

1.2.2 提交一个 SAS 程序

有四种方式可以运行该语句：

第 1 种：单击工具条里面的小人图标。

第 2 种：单击菜单栏 RUN 下面的 SUBMIT。

第 3 种：按热键〈F3〉。

第 4 种：保存上面的代码到某个盘符目录下，然后右键单击文件名，可以通过批处理方式提交，此时选择 Batch Submit With SAS 9.2；当然也可以交互式提交，此时选择 Submit With SAS 9.2。批处理提交和直接提交的区别在于：前者是全部提交所有的 SAS 代码，而后者是逐行提交，如果没有遇到 RUN 语句，进程会一直停留在运行状态，并等待用户的下一个命令。所以，一般情况下，除非在所有代码都调试成功后才用批处理命令，否则交互式命令是常用的选择。

按〈F3〉键运行上面的代码后，会发现日志窗口有如下显示：

```
1     data test;
2     input x y z;
3     cards;
NOTE: The data set work.test has 2 observations and 3 variables.
NOTE: data statement used (Total process time):
      real time           0.01 seconds
      cpu time            0.01 seconds
5     ;
6     run;
```

前 3 行是代码，第 4 行的 NOTE 说明有一个数据集 Work.test，在什么地方呢？有两种方式可以找到该数据集：

第 1 种：双击浏览窗口里面的 Libraries，进去以后会发现有四个逻辑库（关于什么是逻辑库，在第 2 章有详细说明，读者可暂时把它理解为类似于一个 Windows 系统下的目录），其中有个叫 Work 的逻辑库正是要寻找的对象，双击该逻辑库，就会发现里面藏有 test 文件集。双击该文件集，结果会显示该文件集有三个变量 x、y 和 z，对应值正是刚才输入的三个变量值。

第 2 种：单击工具条中的 SAS Explorer 图标，在浏览器窗口中会呈现一个树状的结构，如图 1-4 所示。

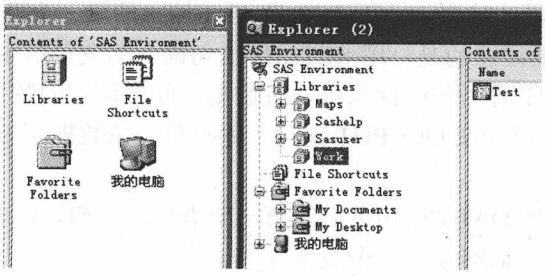

图 1-4 浏览器窗口

读者注意到，其实这个浏览器窗口就是图 1-3 中左边的浏览器窗口，只不过呈树状结构，看起来更方便。双击其中的 Work 逻辑库，在最右边也能看到 Test 数据集。

接下来的一个 NOTE 观测了该程序运行的时间。一般情况下，在项目实施过程中，会把这些时间观测输出，以便调试和优化现有的程序。

1.2.3 保存和打开一个 SAS 程序

刚才编写的这个程序在编程窗口中的名字仍然是 Unititled1，如果需要保存这个程序，有两种方式：

第 1 种：单击工具条的保存按钮，保存到指定目录。

第 2 种：单击菜单栏 File 菜单，选择 Save，保存到指定目录。

需要注意的是，SAS 程序文件的扩展名是 SAS，但可以用任何文本文件名作为扩展名，如 TXT，因为 SAS 程序文件实际上是一个文本文件。

如果要打开一个已经保存的 SAS 程序文件，只要选择 File 菜单下的 Open 选项，打开指定的文件即可。

1.3 DATA 步的数据指针和 PDV 流程

1.3.1 数据指针和 PDV 流程

SAS 语言是先编译后执行的语言，每一个 DATA 步开头标志着编译的开始，接下来的所有语句都被称为“程序”，SAS 总是把这些语句转换成计算机可识别的语言，然后通过逐行语句编译、转换，最后输出成 SAS 数据集。在理解一个 DATA 步所需要走过的正常流程之前，读者需要理解两个概念：

第一个概念是数据指针，为简化起见，读者可以把它理解成在当前的内存缓存区，输入数据所在的位置。

第二个概念是 PDV，它是 Program Data Vector 的缩写，也就是在 DATA 步中所有涉及的变量被变成当前向量的一部分。来自输入数据每一个原始数据或数据集的这些变量，和其他程序语句建立的新变量一起，都放在 PDV 里面。

下面以刚才产生的 Test 数据集为例讲解上面的流程。

第 1 步：编译程序，开辟内存空间。在 Test 数据集中，七行程序首先被编译，并在内存中开辟一块变量空间给三个变量 x、y 和 z。

第 2 步：初始化变量为缺失值。用 INPUT 语句输入的这三个变量在 DATA 步执行之前被置为缺失值。但是需要注意的是，用 RETAIN 语句读入的变量值则被保留。

第 3 步：执行 INPUT 等语句。读入第一条数据，并把读入的数据放入 PDV。

第 4 步：执行 RUN 语句或 OUTPUT 语句。输出第一条数据到被创建的 DATA 步后面的数据集中。

第 5 步：SAS 返回到 DATA 语句之后的第一个语句，并初始化 PDV 中的非 RETAIN 变量 x、y 和 z 为缺失值，然后继续下一个数据读入。

第 6 步：当 SAS 读完并处理完来自 INPUT 语句后面的 1;2;3 和 4;5;6 后，程序到了输入