

# **S+Miner 入门手册**

北京宏能畅然数据应用有限公司编译

**2007 年 4 月**

**Proprietary Notice** Insightful Corporation® owns both this software program and its documentation. Both the program and documentation are copyrighted with all rights reserved by Insightful Corporation. The correct bibliographic reference for this document is as follows: *Insightful Miner™8 Getting Started Guide*, Insightful Corporation, Seattle, WA.  
Printed in the United States.

**Copyright Notice** Copyright © 2006, Insightful Corporation. All rights reserved.  
Insightful Corporation  
1700 Westlake Avenue N, Suite 500  
Seattle, WA 98109-3044  
USA

**Trademarks** Insightful, Insightful Corporation, the Insightful logo, S, S+, S-PLUS, S+FinMetrics, S+SeqTrial, S+SpatialStats, S+ArrayAnalyzer, S+EnvironmentalStats, S+Wavelets, S-PLUS Graphlets, Graphlet, Trellis, and Trellis Graphics are either trademarks or registered trademarks of Insightful Corporation in the United States and/or other countries. Intel and Pentium are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries. Microsoft, Windows, MS-DOS and Windows NT are either registered trademarks or trademarks of Microsoft Corporation in the United States and/or other countries. Sun, Java and Solaris are trademarks or registered trademarks of Sun Microsystems, Inc. in the United States or other countries. UNIX is a registered trademark of The Open Group. All product names mentioned herein may be trademarks or registered trademarks of their respective companies.

1.概览.....	4
1.1 绪论.....	4
1.2 S+Miner 界面的概述.....	4
1.3 一个数据分析问题.....	5
1.4 数据访问.....	7
1.5 数据探索.....	9
1.6 创建模型.....	17
1.7 概要.....	25
2 扩展教程.....	26
2.1 绪论.....	26
2.1.1 Insightful 数据挖掘方法 .....	26
2.2 目标定义.....	28
2.3 数据访问.....	28
2.4 数据探索.....	32
2.4.1 准备数据.....	35
2.4.2 保存工作簿.....	38
2.5 创建模型.....	38
2.5.1 导入训练数据.....	39
2.5.2 作图.....	41
2.5.3 训练模型.....	43
2.5.4 查看模型.....	44
2.5.5 选择模型.....	47
2.5.6 输出模型.....	51
2.6 发布模型.....	51
2.6.1 输入打分数据.....	53
2.6.2 模型的输入.....	54
2.6.3 预测.....	54
2.7 S-PLUS 库的探索 .....	54
2.7.1 使用 S-PLUS 图 .....	55
2.7.2 用 S-PLUS 脚本节点进行建模和预测.....	57
2.8 概要.....	61
2.9 参考文献.....	61

# 1.概览

## 1.1 绪论

**S+Miner** (Insightful Miner 的简称) 是一个企业级的数据挖掘工具。S+Miner 设计目标之一是, 使得它可以与你已经在用的软件无缝衔接。 你可以对很多数据源进行导入及导出, 包括如:

- ✘ Excel 和 Lotus 等电子表格、
- ✘ Access 类的数据库、
- ✘ SAS 和 SPSS 类的统计软件
- ✘ Oracle/DB2/SQLServer 等流行的关系型数据库。

一旦访问到了数据, 你就可以进行以下操作:

- ✘ 可以使用各种图、交叉表以及描述性统计等手段探索你的数据;
- ✘ 使用 S+Miner 数据进行清洗和操纵管理工具, 为分析模型准备数据;
- ✘ 拟合各样概率模型, 包括线形回归, 逻辑回归, 和分类决策树。
- ✘ 用标准工具评估你的模型的效果, 例如提升图。

概览部分简要地向你介绍了 S+Miner 挖掘网络的概念, 并探索一些简单的流程网络实例向你展示如何使用 S+Miner 去解决现实中的数据挖掘问题。

## 1.2 S+Miner 界面的概述

S+Miner界面包含一系列用于进行数据挖掘的节点面板, 外加一个用来设计一个可视化挖掘网络的工作簿。当你载入一个新的工作簿开始数据挖掘时, 界面如图 1.1

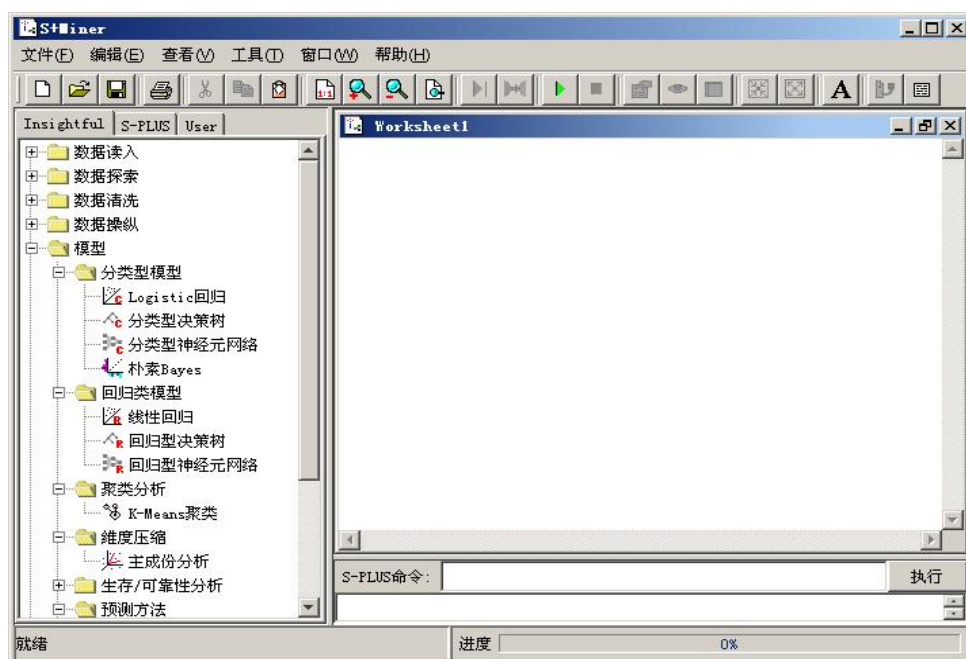


图 1.1: S+Miner 图形用户界面

通过从在左边的探索器框中拖拽组件到右边桌面框的工作簿来创建挖掘网络节点。在节点间建立连接并设置节点的属性。

在桌面框下面是信息框，显示节点运行时的状态。**S+Miner** 的警告和错误提示在这个框中显示。当你运行网络时，**S+Miner** 通过**S+Miner**管道架构传递数据、进行节点计算；它是一个节点接着一个节点来处理数据。为了加快处理进程，**S+Miner**可以通过临时文件对每个节点以二进制形式建立缓存。在默认情况下，数据一次通过管道10000行，但是也可以用全局的或单个节点调整这个数据值。

在下面章节部分，用一个简单的挖掘网络来了解一下它的本质特征和这些特征在医学领域是如何结合起来解决数据挖掘问题的。

### 1.3 一个数据分析问题

这个例子用到的数据是来自于杜克大学心血管病数据库，它有3504个病人由6个变量组成。这些病人由于胸腔疼痛到杜克大学医疗中心咨询。这个习题的目的很简单：

预测一个病人有显著冠状疾病发病的概率，在至少一个重要的冠状动脉中有大于或等于75%的直径都变窄的情况就称为显著的冠状疾病。

这个数据集中的六个变量如下：

sex	0=男性，1=女性
age	患者的年龄，以年计
cad. dur	冠状动脉病兆持续时间，以月计
cholesterol	病人的胆固醇水平
Si gdz	显著冠心病发病标识：Presence (or absence)
Tvdl m	严重冠心病标识：Presence (or absence)
	这也被叫做三血管疾病(three-vessel)或左主冠动脉疾病

这个分析显著冠心病发病标识来做因变量（sigdz）。为了进行这个分析，创建一个**S+Miner**挖掘网络来评估两个结果模型并确定哪个模型对显著冠状疾病发病的概率预测效果更好。

启动**S+Miner**，以便开始分析这个例子。**S+Miner**屏幕出现，随即出现如图1.2的对话框。



图1.2：启动文件选择对话框。

按以下操作打开这个问题的工作表：

1. 点击 **浏览** 按钮显示 **打开** 对话框
2. 在打开文件选择对话框的底部，点击 **例子** 文件夹图标。（点击这个图标复制所有在安装 **Examples** 文件夹中的文件到

/username/iminer\_work\_8\_0/examples 目录并保存原始工作表和数据集在安装例子目录下)

3. 查看 打开 对话框并显示新的 username/iminer\_work\_8\_0/examples 文件夹。

双击dukestudy文件夹，选择文件dukecath.imw，并点击打开。

**注：**使用Solaris操作系统的用户，如果你是在Solaris操作系统里运行S+Miner，例子文件夹图标在打开对话框的右边。

除此之外，安装例子直接复制到/username/iminer\_work/examples目录下。

4. 在 启动选择文件 对话框点击 确定。

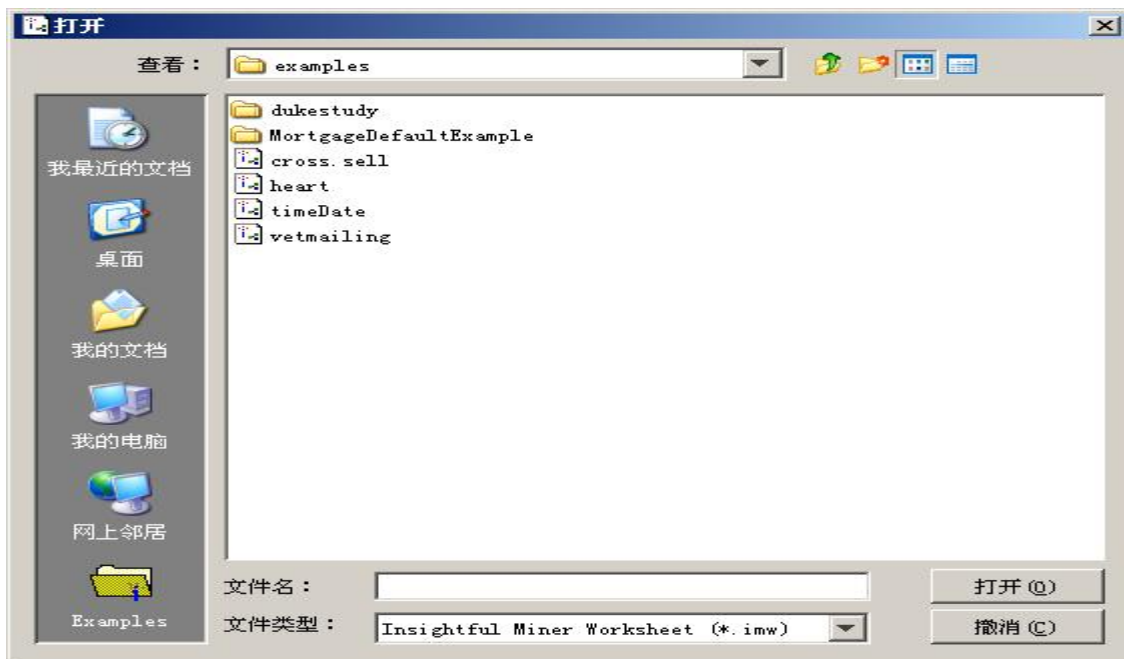


图1.3 点击例子文件夹图标（左下）把安装例子目录中的文件直接复制到

username / iminer\_work\_8\_0/examples目录下

这个打开的工作表包括一个挖掘网络的例子，如图1.4。注意到所有的网络节点都出现一个红色的状态指示器，它表示节点被连接但是数据是缺失的。当你读入数据后，节点变成黄色，表示你可以运行。当你完成节点对话框且运行网络时，所有的状态指示器变成绿色表示节点已经成功完成。

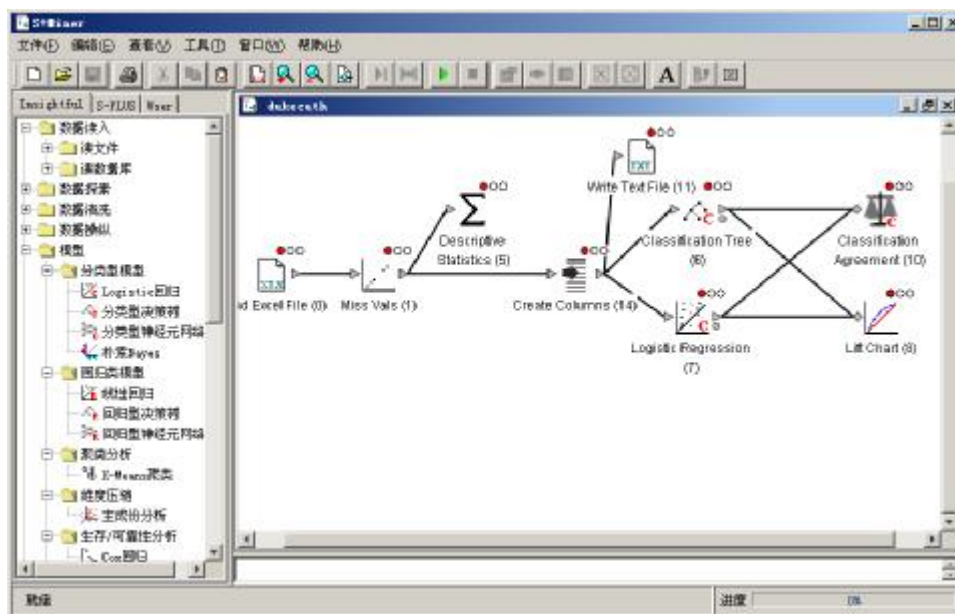


图1.4：状态指示器显示的红色的时候表示节点暂不可用或数据缺失。

图1.4的网络展示了一些数据挖掘的步骤。接下来，检查网络中每个节点确定它们在做什么。

## 1.4 数据访问

图1.4展示的挖掘网络例子起始于一个 **读Excel文件(Read Excel File (0))** 节点：一种S+Miner读入数据的方法。你可以用 **数据读入** 组件读入数据，包括**读文本文件**、**读固定格式文本文件**、**读SAS文件**、**读其他类型文件**，或者是用 **读数据库** 组件。

1. 双击 **读Excel文件** 节点打开它的属性对话框。对话框如图 1.5

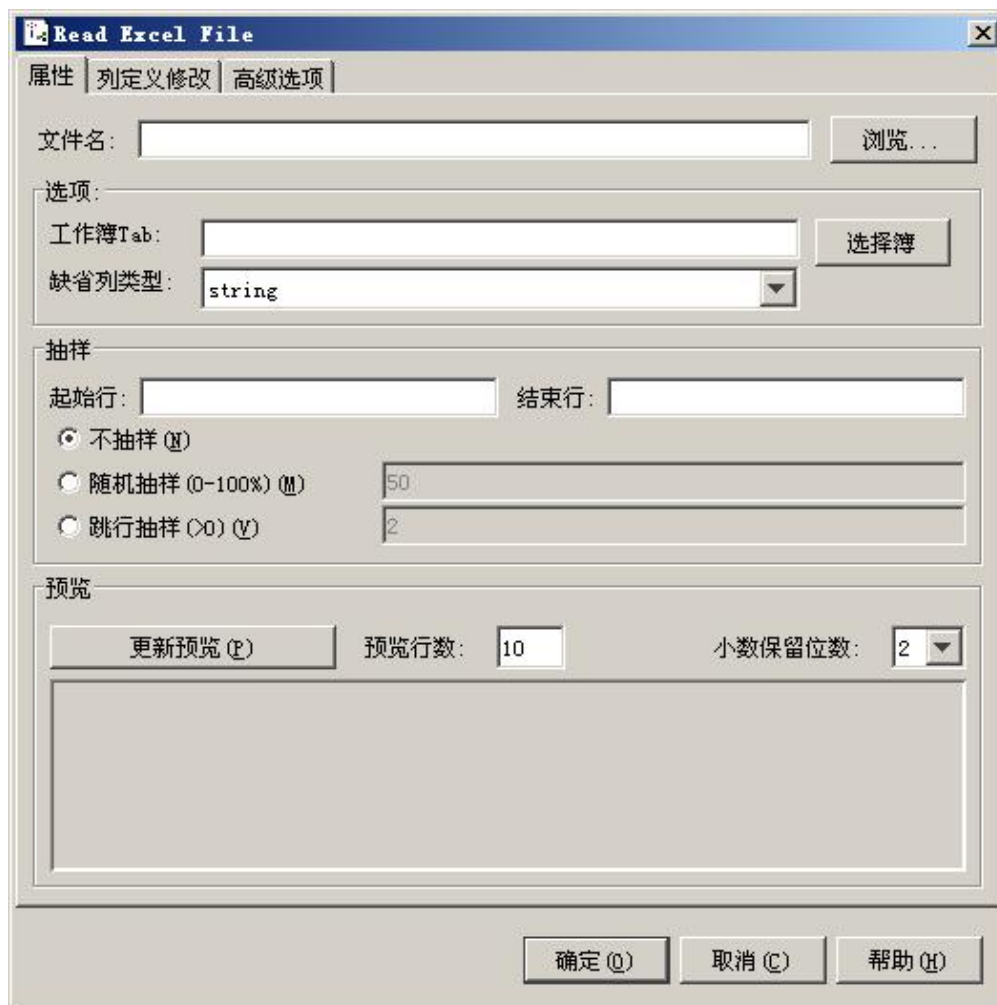


图1.5： 读Excel文件组件的属性对话框

2. 点击 **浏览** 按钮显示 **打开** 对话框
  3. 因为曾经点击过 **例子** 文件夹图标（在对话框的左下角），**打开** 对话框会显示 **username/iminer\_work\_8\_0/examples/dukestudy** 文件夹。
  4. 在 **dukestrdy** 文件夹中，选择数据文件 **acath.xls**，并点击 **打开** 按钮。（如果你在Microsoft Windows 下操作，且你的选项设定成隐藏文件扩展名，那文件名就显示成acath。）
- 在 **预览** 选项组中，点击 **更新预览** 来显示前十行数据（默认状态）

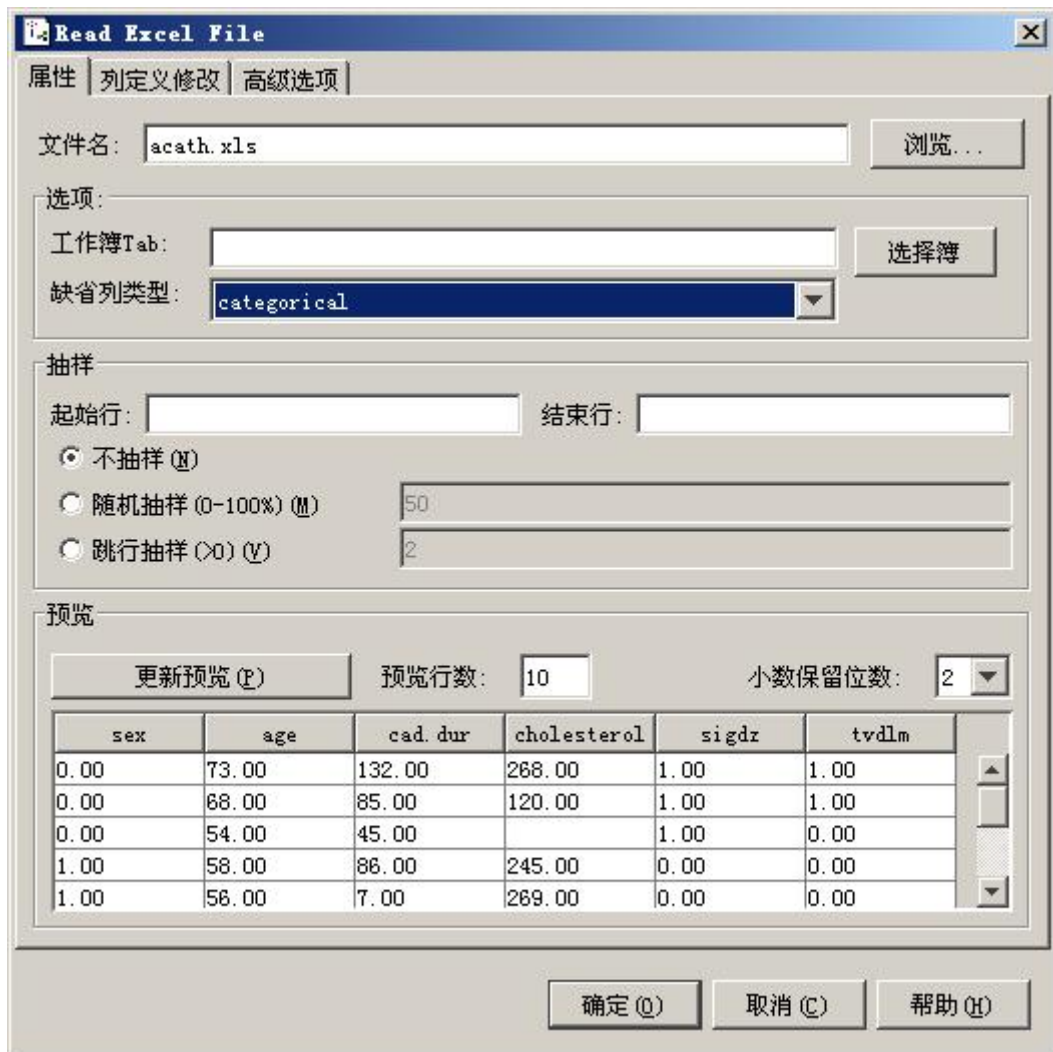


图1.6:完整的读Excel文件节点

因为这个例子的目标是预测显著冠状疾病发病的概率（sigdz），如果想建立一个以sigdz作为因变量的模型，这要求把因变量设定成分类型变量。但它是连续型或数值型变量形式输入的。为了把sigdz变成分类型变量，点击 **列定义修改** 选项，再进行以下操作（图1.7）：

5. 在下拉列表中选择 sigdz 行。
6. 在 **设置类型** 选项组中，点击 **分类型**。
7. 点击 **确定** 按钮关闭对话框。
8. 在S+Miner 工具栏中，点击 **运行至此** 按钮，运行网络到此。


**读Excel文件** 节点的状态指示器这时变成绿色，表示它成功完成了数据的读入。来看一下例子数据集的帮助文件，从S+Miner菜单选项，点击**帮助** ► **帮助索引**，选择**acath.xls**数据集。





图1.7: 把sigdz从连续型变量变成分类型。


## 1.5 数据探索

1. 打开 **读 Excel 文件(Read Excel File (0))** 节点的查看器检查你导入的数据。点击 **读 Excel 文件** 节点，然后点击 S+Miner 工具栏上的查看器按钮 .

如图 1.8 所示，**读 Excel 文件** 节点的查看器是一般的节点查看器，它是 S+Miner 中许多节点的查看器，包括所有的输入输出节点和数据操纵节点。节点查看器由六个选项页组成：

- ¥ 第一项显示所有的数据集。
- ¥ 第二项到第五项显示四种不同数据类型的数据包括（连续型，分类型，字符串，和日期）
- ¥ 第六项显示其他类型的数据。

节点查看器每页的底部都总结了节点输出数据的概况：5 个连续型列数（或变量）和 3, 504 个观测样本。图 1.8 显示了默认的 5 个连续型变量。

2. 点击 **连续型** 按钮检查这些变量。这个图表显示了数据的一些特征：**缺失值个数** 列显示了 cholesterol 变量缺失 1246 个值，tvdlm 变量缺失 3 个值。
3. 点击 **分类型** 按钮显示了唯一一个分类型变量 sigdz。点击这一行的任意地方，可以看见它的不同水平。
4. 检查完这些数据，点击窗口右上角的  按钮关闭节点查看器。

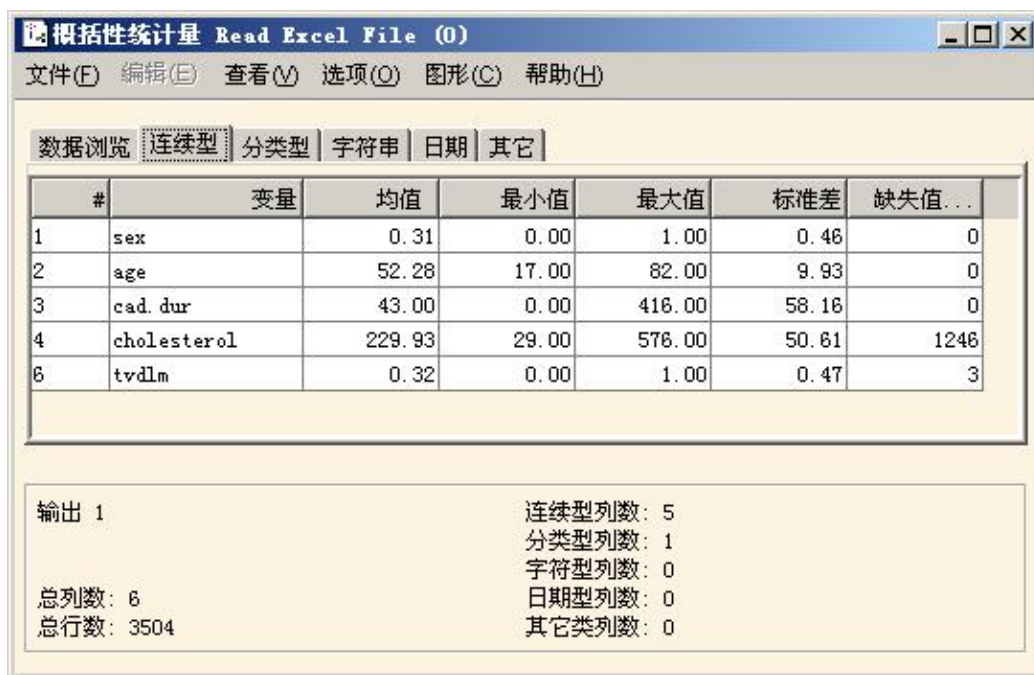


图 1.8: 读 Excel 文件节点的查看器

## 清洗数据

接下来用 **缺失值处理(Miss Vals (1))** 节点去除缺失的行，因为它们对分析没有意义。

1. 右键点击工作表中的 **缺失值处理** 节点，选择**属性**（图1.9）。



图 1.9: 选择去除行作为处理缺失值的方法。

2. 点击cholesterol， 按住CTRL点击tvdlm选择两列。在 **选择方法** 的下拉列表框中选择**去除行**，点

击 设置方法。

3. 点击确定，点击 **运行至此** (运行到此)。
4. 右键点击 **缺失值处理** 节点选择 **查看器**。
5. 点击 **连续型** 按钮观察对话框底部数据概况。如图 1.10，这个数据集仅有 2258 行。

	sex	age	cad. dur	cholesterol	sigdz
	continuous	continuous	continuous	continuous	categorical
1	0.00	73.00	132.00	268.00	1
2	0.00	68.00	85.00	120.00	1
3	1.00	58.00	86.00	245.00	0
4	1.00	56.00	7.00	269.00	0
5	0.00	41.00	15.00	247.00	1
6	0.00	35.00	44.00	257.00	0
7	0.00	58.00	7.00	168.00	1
8	0.00	81.00	2.00	246.00	1
9	0.00	58.00	79.00	221.00	1
10	0.00	47.00	6.00	272.00	1
11	0.00	66.00	8.00	257.00	1
12	0.00	48.00	69.00	236.00	1

输出 1	连续型列数: 5
	分类型列数: 1
	字符型列数: 0
	日期型列数: 0
	其它类列数: 0
总列数: 6	
总行数: 2258	

图 1.10: 运行 缺失值处理节点

为了得到数据的形象表述，可以对每个连续型变量进行做图：

6. 点击第一行的任何地方来选择列表格的第一行。
7. 按住 **SHIFT** 点击列表框中最后一行，来选择数据集中所有连续变量。
8. 从节点查看器窗口顶部菜单中选择 **图形** ► **概括图**，如图 1.11。

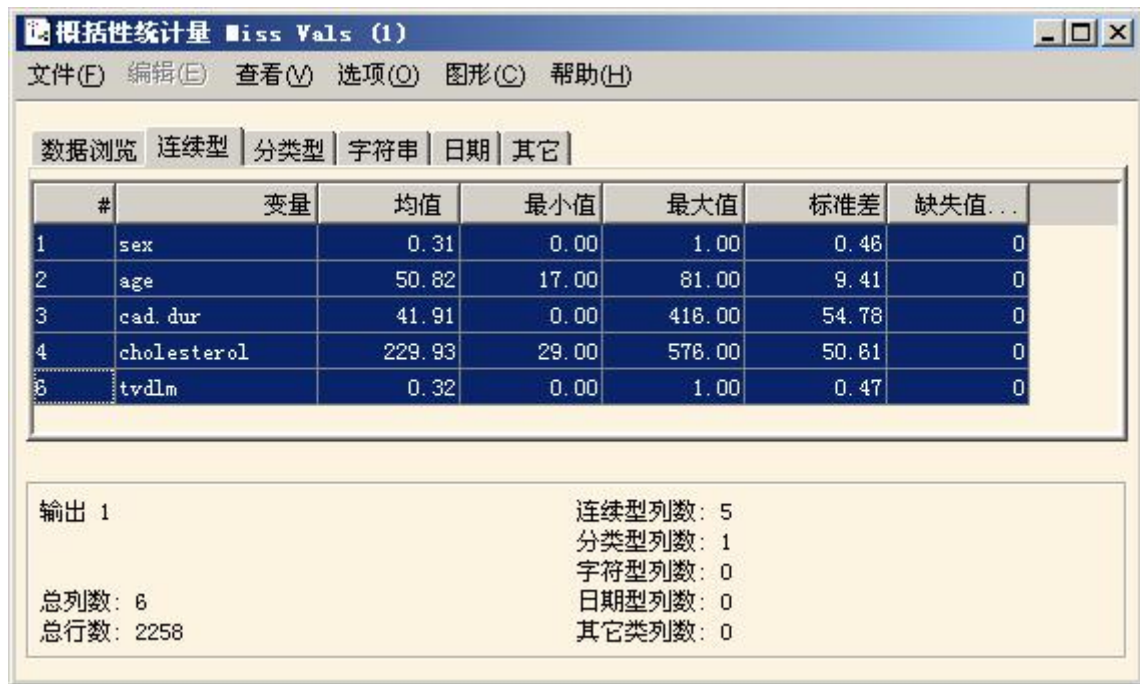


图 1.11: 在节点查看器中创建单变量的图

图 1.12 的图表查看器显示了数据的概况和每个被选变量的图形:

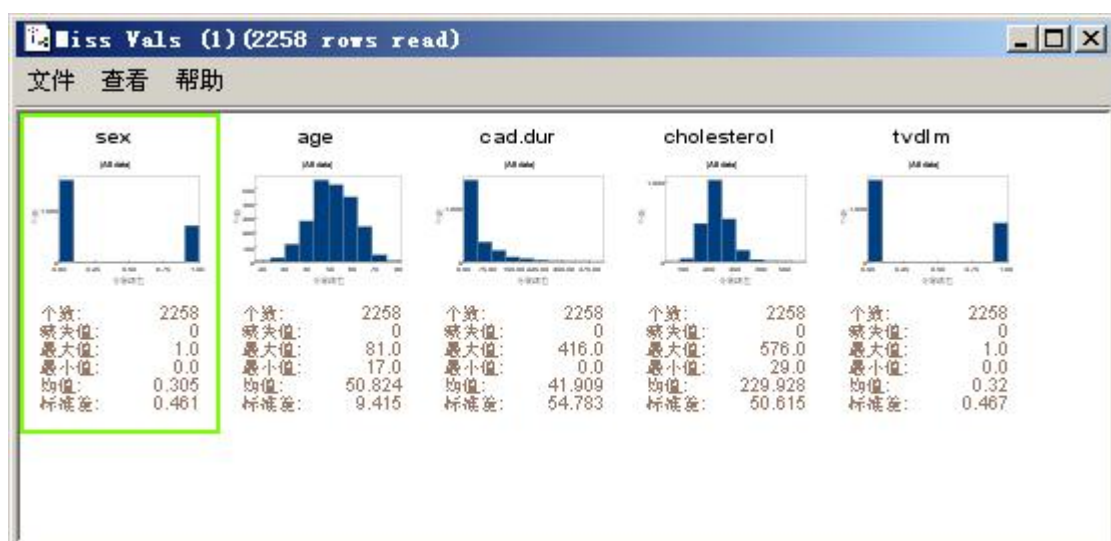


图 1.12: 数据概况和数据集中每个连续变量的图。对连续型数据做出直方图。

9. 回到查看器窗口, 点击 **分类型** 按钮, 对数据中变量 sigdz 重复以上做法:

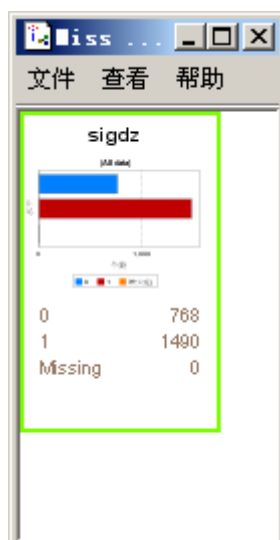


图 1.13: 数据概况和每个分类型变量的图。对于分类型变量，用条形图来显示。

为了扩大分类型变量的图，双击 **sigdz** 图。如图 1.14 所示，一个 **选择的图** 窗口打开，显示数据概况和 **sigdz** 变量的条形图。

分类型变量图显示了患有显著性冠状疾病的人数（患病比大约 2：1），通过图下的水平的计数值来显示。下一部分给出这些观测。

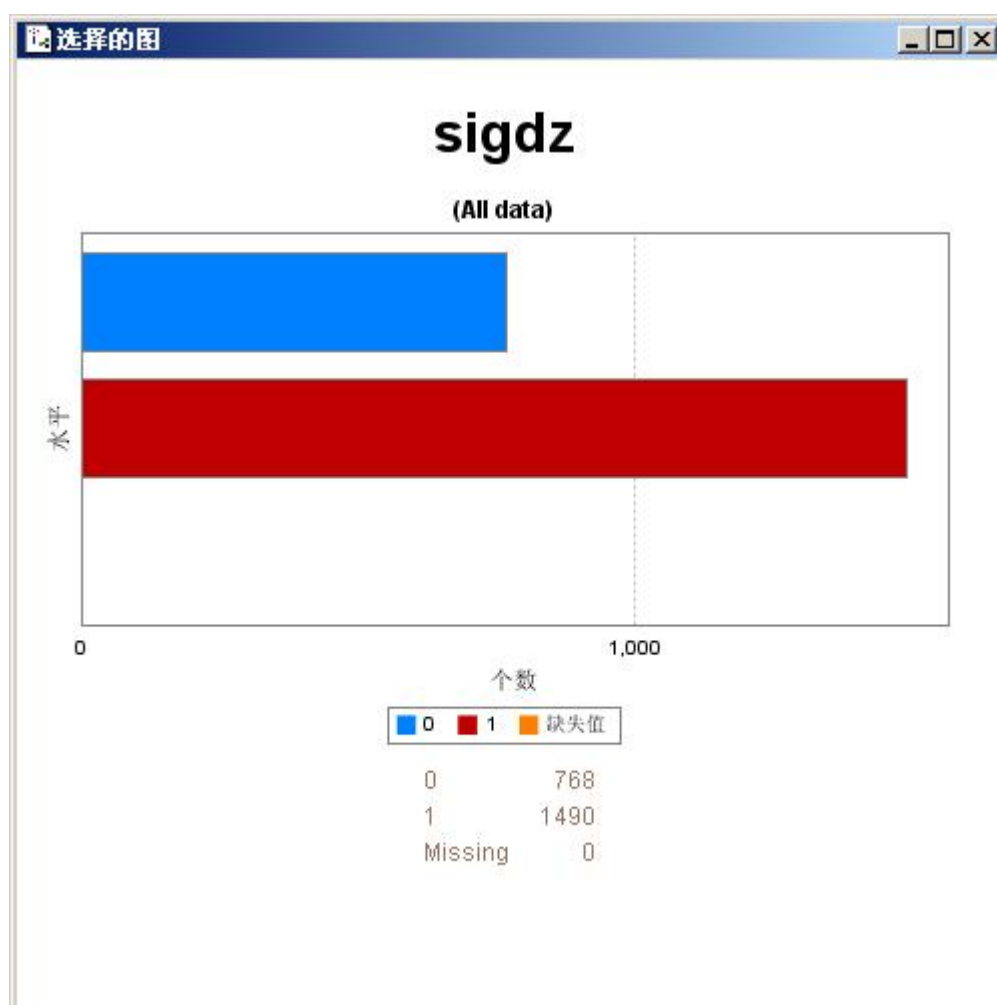



图 1.14: **sigdz** 的放大图，显示了有无患有显著性冠状疾病的数量比是 2：1。

10. 关闭 **选择的图** 窗口。
11. 查看完数据后，关闭节点查看器和图形窗口。

## 进一步的数据探索

通过查看数据的描述性统计，你可以对数据有更好的理解，你可以删除缺失值并修改列。通过运行 **描述性统计** 节点你可以得到均值、标准差和数据极值。

设定 **描述性统计** 节点属性如下：

1. 右键点击 **描述性统计** 节点选择属性。
2. 在 **可用列** 列表框中选择所有变量，点击右侧的双箭头按钮  把变量移到 **显示** 列表框中，


如图 1.5。点击 **确定** 按钮，再点击工具栏上的 **运行至此** 按钮 .



图 1.15: 用**描述性统计**节点来选择那些你希望计算概率的变量，如均值，标准差和极值。

3. 右键点击 **描述性统计** 节点选择 **查看器**。变量的概率如图 1.16 所示。可以看出 Cad.dur 的直方图显示了较高的峰度。



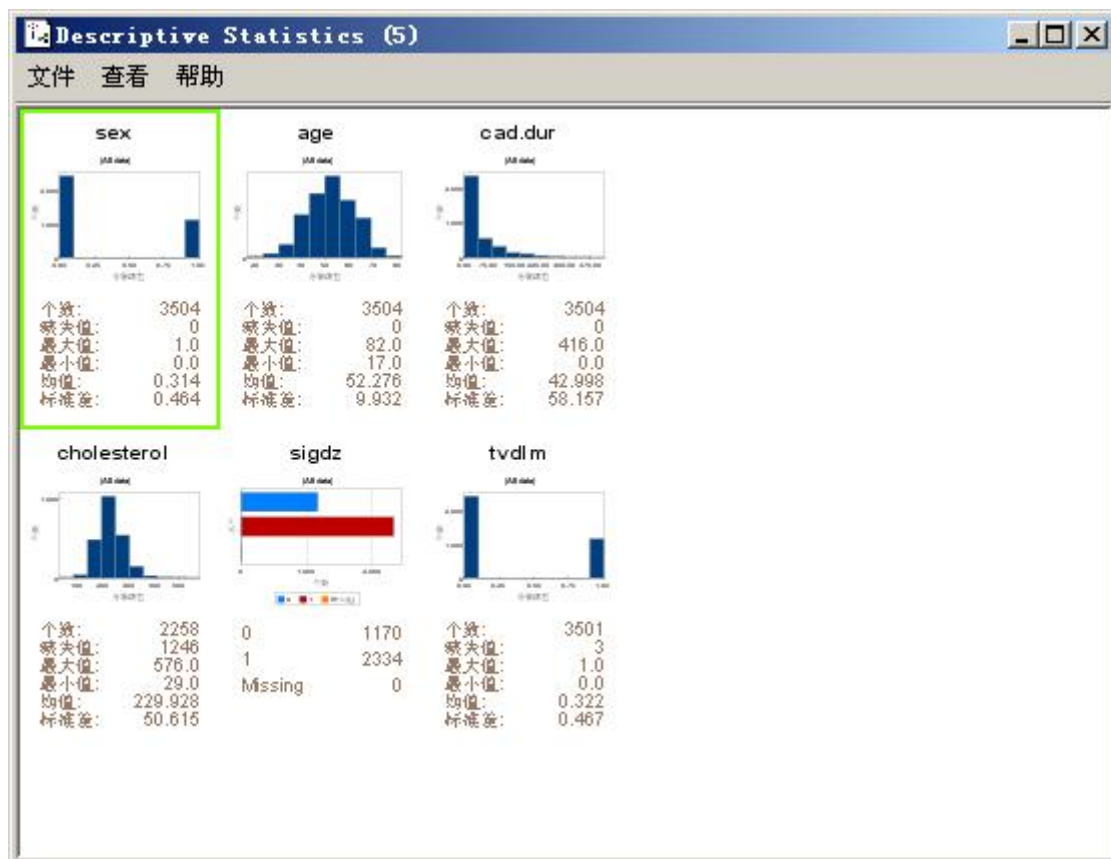


图 1.16: 描述性统计节点的输出，显示变量概率

在下节数据操纵部分，你可以通过创建一个变量的 log 变换来描述 cad.dur 的峰度；为了演示，还创建一个新的变量 age\*cholesterol，且可以把它引入在模型中。在后面的创建模型部分，你可以用 Logistic 回归检验和分类树检验对 sigdz 的预测。

#### 4. 关闭 描述性统计 节点查看器。

### 数据操纵

为了创建 cad.dur 的 log 变换，需要用一个表达式创建一个新的变量叫做 lcad。同样，用一个表达式创建变量 age.chol。你可以用 **创建新列** 节点创建两个新的变量。

为了创建这些列（图 1.17）：



1. 右键点击 **创建新列** 节点选择 **属性**；
2. 从 **选择类型** 下拉列表中点击 **连续型**；
3. 点击 **加入**；
4. 在 **名称** 下写入 lcad，在 **创建新列表达式** 下写入  $\log(\text{cad.dur}+1)$ ；
5. 再次点击 **加入**；
6. 在 **名称** 下写入 age.chol，在 **创建新列表达式** 下写入  $\text{age}*\text{cholesterol}$ ；
7. 点击 **确定**；
8. 从工具栏中点击 **运行至此** (  ) 按钮来运行 **创建新列** 节点，点击查看器 (  ) 来查看新增两列的数据。



图 1.17: 用创建新列节点通过操纵已存在的变量创建新列 (lcad 和 age.chol)

9. 点击 **连续型** 来查看新增变量, 如图 1.18。




图 1.18: 运行创建新列节点, 两个新列 lcad 和 age.chol 被创建



10. 关闭节点查看器。

随着新变量的增加，你已具备了创建模型和对因变量 **sigdz** 进行预测的所有数据。在做这些以前，保存更改的数据集把它写入文本文件中。这样做，你以后还可以对数据进行检索。

1. 在 **创建新列** 节点上方，双击 **写文本文件** 节点。
2. 写入 **acath\_modified.txt** 作为 **文件名**，并在 **分割符** 下拉列表中选择 **single space delimited**。

点击确定。然后点击 **运行至此** (  )。文件被保存到 `username/Iminer_work_8_0/examples` 目录下。如果 **acath\_modified.txt** 已经存在，可以覆盖并保存它。

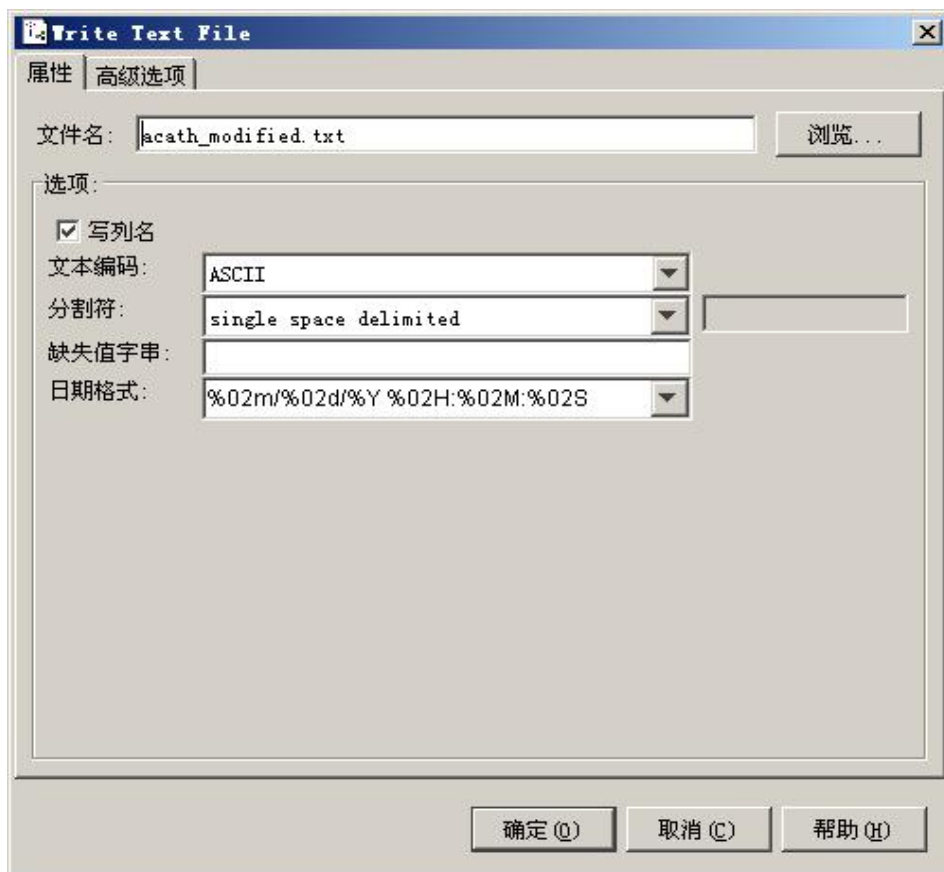


图1.19：用**写文本文件**节点来输出已修改的Excel数据**acath.xls**到文本文件**acath\_modified.txt**中。你可以用这些数据做其他分析。

## 1.6 创建模型

S+Miner 为依据自变量预测因变量提供工具。这个例子用来比较分类决策树和 Logistic 回归两种方法何种方法对于预测 **sigdz** 更有效（图 1.20）。

**Sigdz** 变量是二元变量；它表示一个由于胸痛来医院治疗病人是否患有显著性冠状动脉疾病。二元因变量数据是分类决策树和 Logistic 回归两种方法中最普遍的模型。这个挖掘网络例子使用了两种模型。两种模型都是用同一个因变量(**sigdz**)和自变量(**sex**, **lcad**, **age**, **cholesterol**, 和 **age.chol**)来预测响应变量。

这个例子没有用 cad.dur 和 tvdlm 做因子：而是用 lcad(cad.dur 的 log 变换)来代替因子 cad.dur，tvdlm 包含的信息在实际预测 sigdz 时并没有用到。

对于这个模型，指定分类变量 sigdz 为响应变量或是因变量，其余变量为预测因子或是自变量。

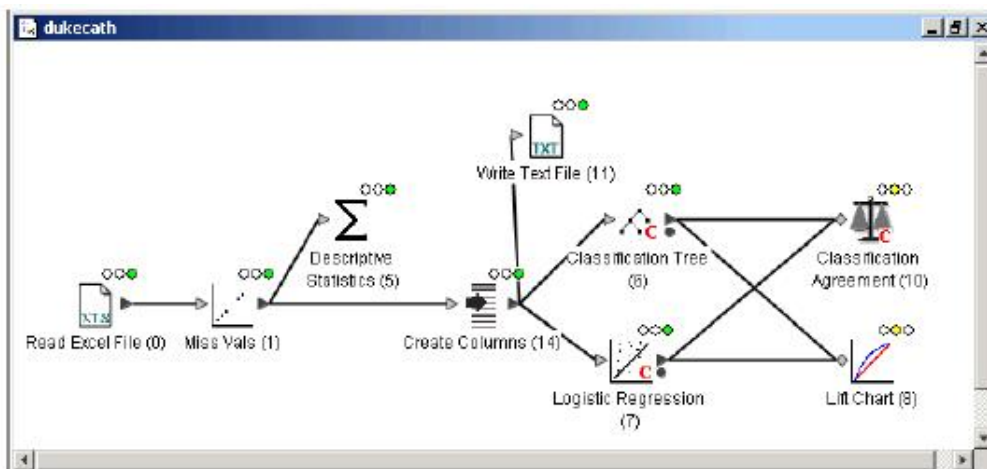


图 1.20: 网络的后一部分集中来比较用分类决策树和 Logistic 回归两个节点进行预测的效果。

运行完这些后，用分类吻合度和提升图节点来评估两节点的性能。

## 创建分类决策树



1. 右键点击 分类决策树 节点从快捷菜单中选择 属性。这部分完整的对话框如图 1.21。



图 1.21: 分类决策树的属性页。

分类决策树节点允许你选择因变量和自变量用于预测。

这里你感兴趣于预测显著性冠状疾病(sigdz)，它是两个预测模型的因变量。

2. 在 **可用列** 列表框中，选择变量 sigdz。
3. 点击因变量列表框左侧的  按钮
4. 在 **可用列** 列表框中，CTRL-点击 age, cholesterol, age.chol, lcad, 和 sex。
5. 点击 **自变量列** 表框左侧的  按钮。
6. 在页面底部的 **方法** 功能组中，选择 **集成树**。集成树是树的集成。树模型的预测是根据集成树的平均。
7. 点击 **集成树**。

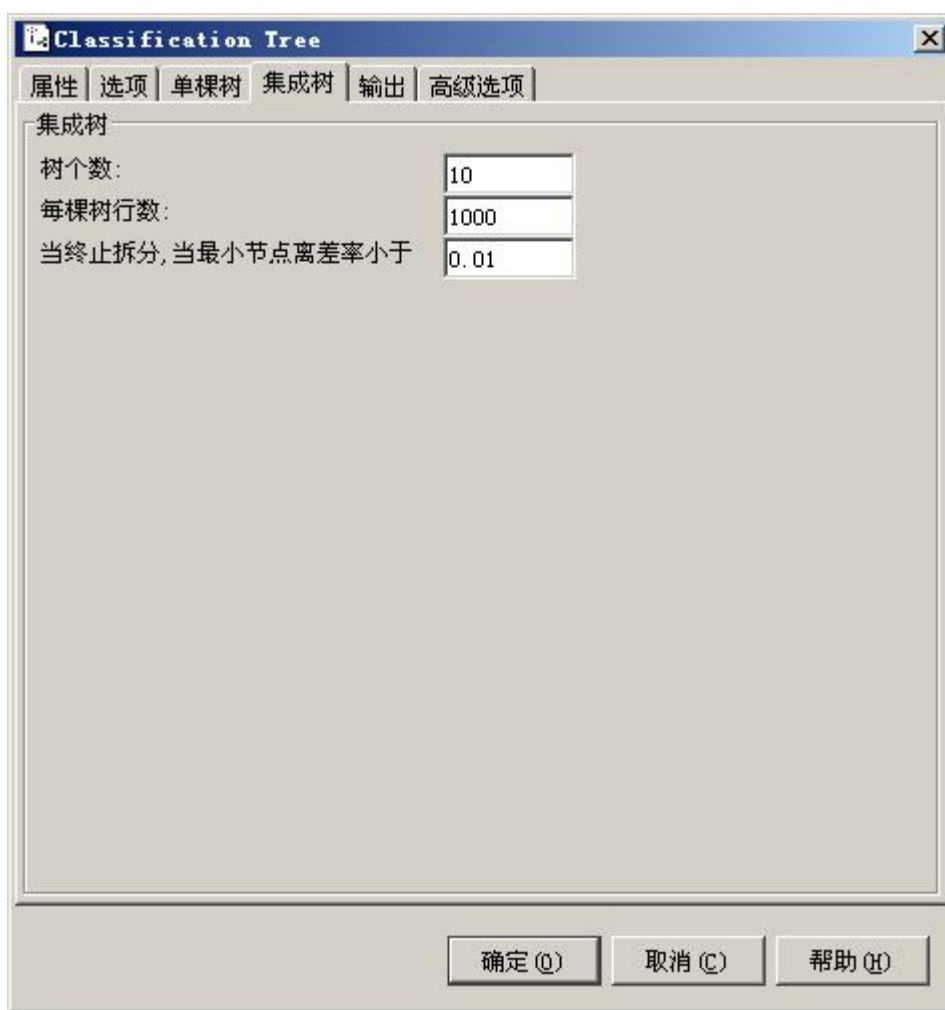


图 1.22: 分类决策树对话框中的集成树页面

8. 在每棵树行数写入 1000(acath.xls 数据集有 3504 行，这个值一定要小于数据集的总行数)。图 1.22 显示了这部分完全的对话框页面。
9. 点击 **输出** 选项。



图 1.23: 分类决策树对话框的输出页面

10. 在 **新列** 选项组下面的 **概率** 组中，点击 **指定类别**，从下拉列表里选择 **1**。

如果是默认状态，S+Miner 只能返回因变量最终水平的概率值。如果想显示 **1** 水平的概率，你必须通过明确的选择这个选项来选择变量值 **1**。图 1.23 显示了这部分完全的对话框页面。

11. 点击 **确定** 关闭对话框。

## 创建 Logistic 回归检验

接下来，具体介绍 **Logistic 回归** 节点的属性。

1. 重复分类决策树的 1-5 的步和 9-10 的步用于 **Logistic 回归** 节点。

图 1.24 显示了 Logistic 回归节点属性页完整的对话框



图 1.24: 为 **Logistic 回归** 节点选择变量的对话框。

2. 点击 **确定** 来接受改变。

3. 点击  按钮来运行网络。

看两组模型：

4. 右键点击 **分类决策树** 节点，在快捷菜单中选择查看器。打开 **分类决策树** 查看器，如图 1.25

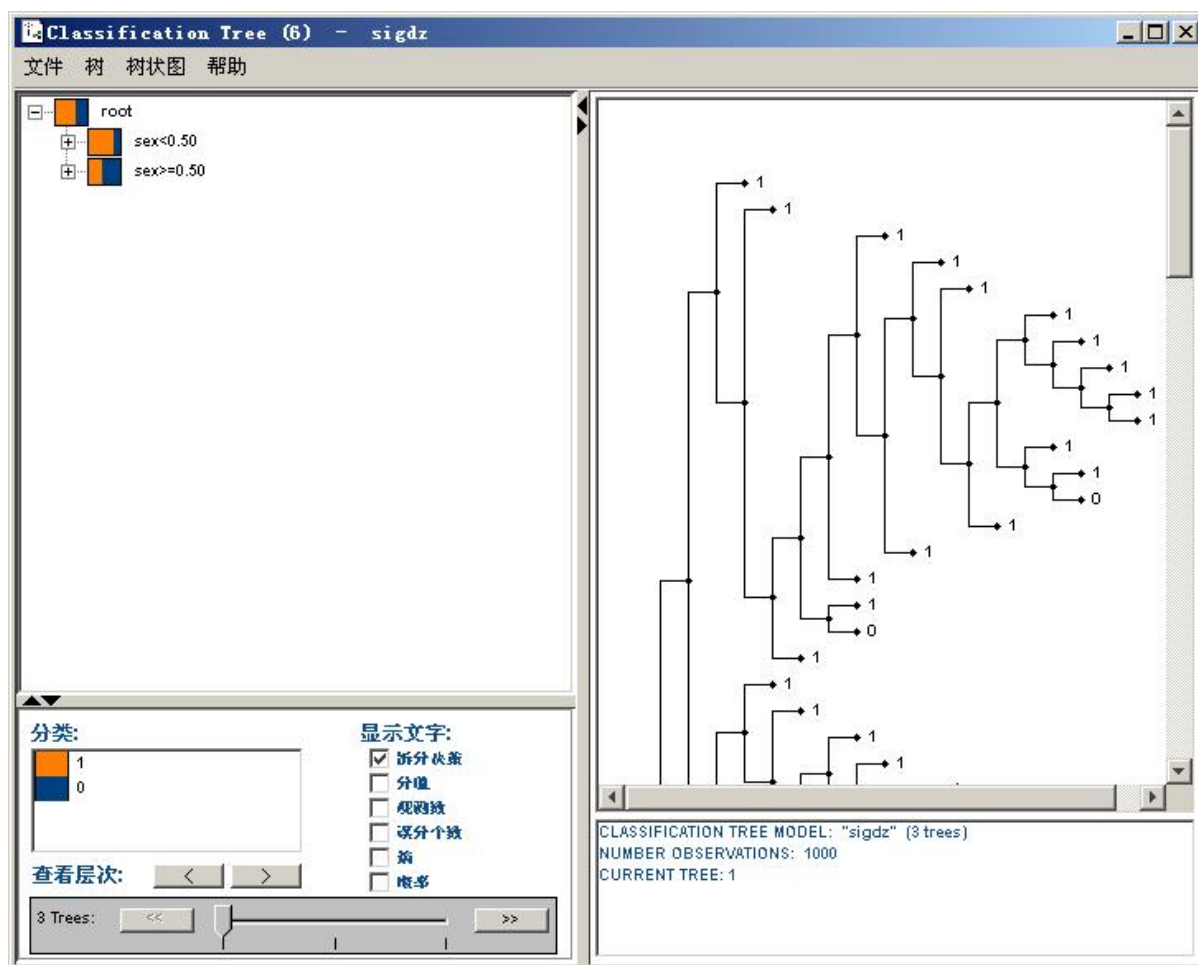




图 1.25: 分类决策树节点的查看器

分类决策树算法为每个通过管道式架构输送的数据块分别拟合一棵树。对于这个数据，有 2258 行，你选择的数据块大小（1000 行）产生了 3 棵树。这 3 棵树的集合叫做集成树。你可以通过点击查看器底部左边灰色栏里的双箭头按钮（和）来翻看这三棵树。这个集成树所做的预测是计算这三棵树模型的平均来实现的。这就是常说的平均块模型。

5. 打开 **Logistic 回归** 节点的查看器。一个网络窗口浏览器打开，显示了模型的系数表，如图 1.26。（最大化浏览器来查看结果。）

## Logistic Regression (7)

DEPENDENT VARIABLE: SIGDZ

Coefficient Estimates				
Variable	Estimate	Std.Err.	t-Statistic	Pr(> t )
(Intercept)	-8.60	1.37	-6.29	3.92E-10
sex	-2.06	0.11	-18.13	1.11E-68
age.chol	-0.00	1.1E-4	-3.43	6.2E-4
age	0.16	0.03	5.98	2.66E-9
cholesterol	0.03	0.01	4.86	1.25E-6
lcad	-0.01	0.04	-0.13	0.90

Analysis of Deviance		
Source	DF	Deviance
Regression	5	559.90
Error	2,252	2,335.38
Null	2,257	2,895.29

Correlated Coefficients	
Coefficients	Correlation
age.chol and age	-0.97
age.chol and cholesterol	-0.98
age and cholesterol	0.96

图 1.26: Logistic 回归节点的查看器

6. 当你查看完节点，关闭查看器。

### 比较模型

在上节，为了比较分类决策树和 Logistic 回归模型，使用了两个不同的节点：**分类吻合度** 和 **提升图** 节点。

**分类吻合度** 节点用于比较多个分类模型的精度。这个问题中，是通过比较 **分类决策树** 节点和 **Logistic 回归** 节点来产生。它利用模型产生的预测值生成一个误分矩阵。它揭示了模型所作的正确分类的个数和比例。如图 1.27



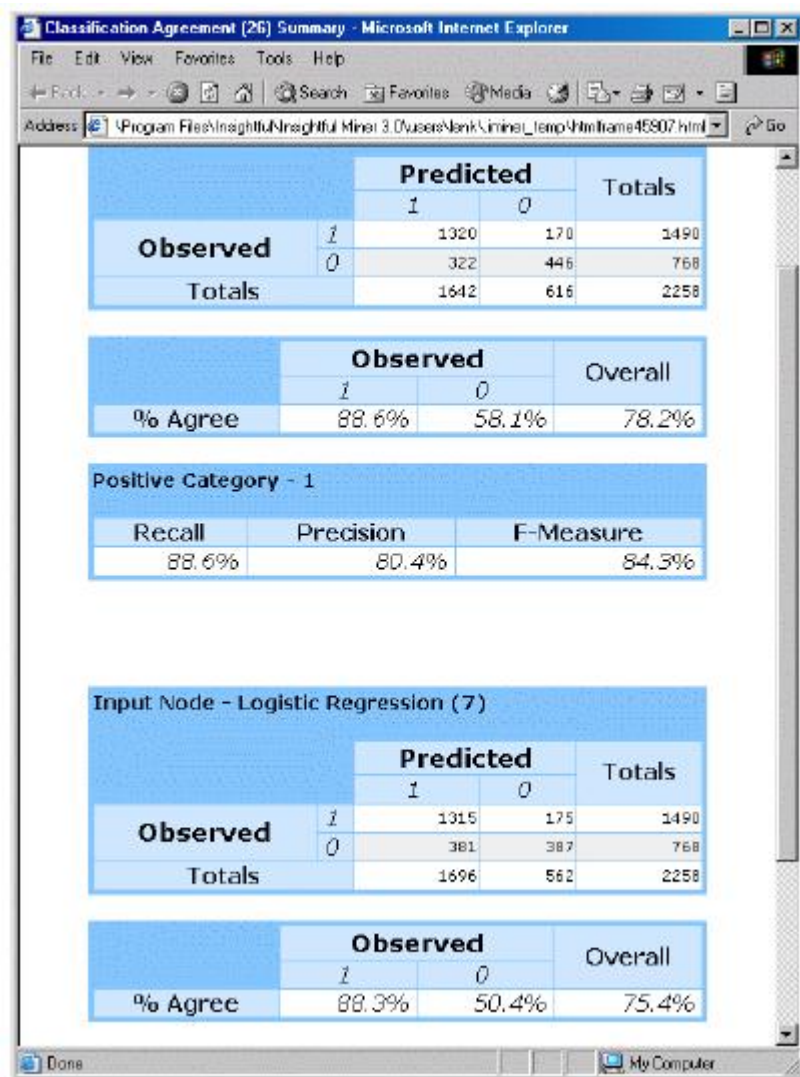


图 1.27: 分类吻合度节点的输出, 比较了分类决策树节点和 Logistic 回归节点的输出结果。

可以观察到 分类决策树 节点整体精确度 78.2%, 而 Logistic 回归节点的是 75.4%。因此 分类决策树 节点整体的预测结果稍好一些, 这主要是由于它对于非显著性冠状疾病的预测效果更好一些 (58.1%对 50.4%), 所以从整体来看, 这个模型更好。

另一个用来比较的节点是经典的 提升图, 它主要用来比较两个模型的累计收益和提升度。提升程度与初始状态相比: 在提升图里用一条参考直线来表示初始状态。

1. 打开 提升图 节点的查看器。
2. 在 图类 下方, 选择 累计收益图, 如图 1.28



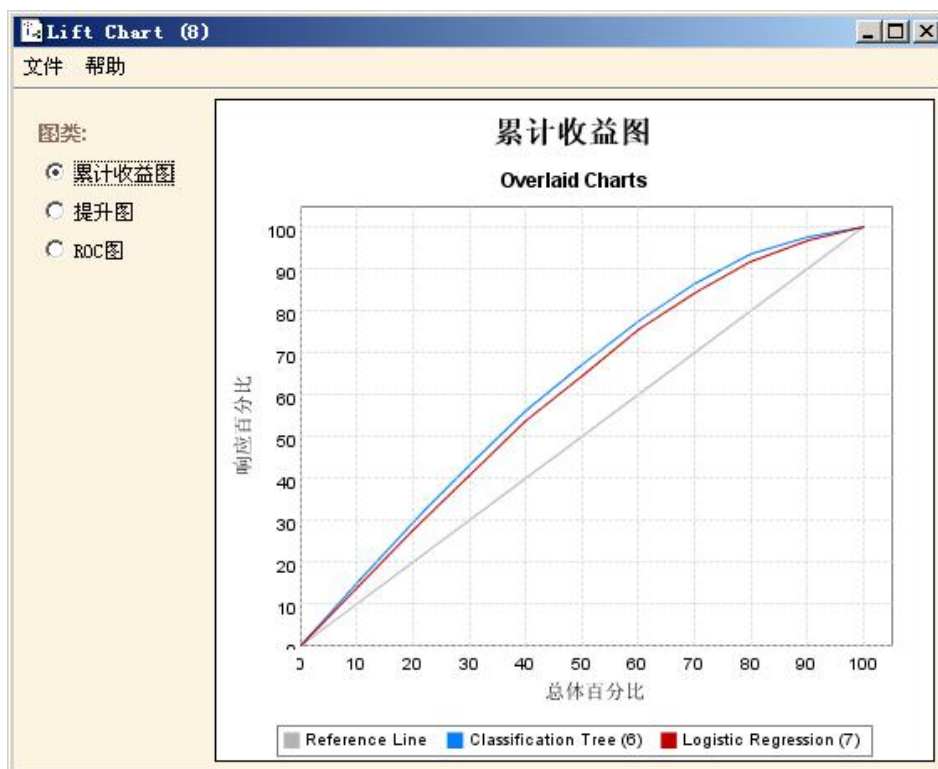


图 1.28: 提升图节点的查看器

可以看出分类决策树用上方的蓝线来表示, 表明它比 Logistic 回归的红线提升程度稍高一些, 所以, 首选分类决策树模型来预测 sigdz。

## 1.7 概要

从这个数据例子中你至少可以得出以下两条结论:

1. 分类决策树模型预测显著冠状疾病发病概率要比 Logistic 回归模型更精确。
2. 如果你用分类决策树模型进行预测, 可以可靠地预测出一个病人在当时患有显著性冠状动脉疾病的概率是 88.6%。

可以发现这个模型没有使用表示患有严重冠状动脉疾病 tvdlm 变量。由此你可以实现另一个研究, 预测一个患有显著性冠状疾病的人患有严重冠状疾病的概率。你可以使用这个数据集的子集来实现这个分析, 或是将患显著性冠状疾病的情况作为发展成严重性冠状动脉疾病的标识。

为了简便, 很重要的一点是这个模型只运行一个确定的被叫做训练数据的集合。理想情况, 你可以使用 **分割** 节点把一部分数据用来训练(运行模型), 另一部分用来测试(预测模型), 最后用一个新的数据集检验(确认模型)。

第二章的例子, 是一个扩展教程, 描述了一个更复杂的例子, 用训练和测试数据创建一个模型来预测消费者家庭抵押贷款拖欠的情况。这个模型使用一个新的数据集。随着你对这个工具逐渐的熟悉, 你就可以知道怎样应用 S+Miner 的能力来进行你的数据挖掘。

## 2 扩展教程

### 2.1 绪论

在这个扩展教程中，将使用 S+Miner 的建模功能来预测金融数据。在这个例子中，建立一个模型来预测哪个消费者在家庭抵押贷款中会拖欠。例如，想象一下如果你在一个私营的抵押公司工作，你想从另一个抵押公司贷款，但是你决定慎重对待所要面临的风险。你想买的家庭抵押贷款不拖欠的可能性高于 0.98。

家庭抵押贷款是美国很大的一项业务。不仅是因为它为家庭融资与再融资提供了主要的市场，也因为它是各种放款业务交易的第二大活跃市场。贷款通过拖欠和不拖欠的风险来评估。一个关键的问题是建立一个模型，基于已知的消费者属性和消费者资源的结合（如信用度积分、贷款历史、房屋价值等等）来预测他是否会拖欠贷款。不仅在需要精确评估贷款的第二大市场，在需要建立成功的贷款渠道策略的第一大市场，这些模型都是有价值的。

统计建模在这个领域发挥了巨大的作用。响应变量和预测因子的关系（例如，给出有过消费者的拖欠史）是很能说明问题的。S+Miner 非常适合这类问题，因为它提供了先进的半参数和非参数的方法，这种方法在因变量和预测间不用假定具体的参变量（例如，线性）。

在这个扩展教程中，你可以开发各种模型来拟合训练数据集，用新数据（测试数据）比较模型，然后选择最好的模型。使用你认为最好的模型，预测或是给一些潜在的消费者评分。对潜在的贷款客户列表进行预测或打分。根据你愿意承担的风险过滤你的客户列表，并决定如何放贷。

#### 2.1.1 Insightful 数据挖掘方法

Insightful 根据开发和应用现实数据挖掘中获取的经验制定了一套数据挖掘的流程。图2.1高度概括了这个过程。以下是这些步骤。

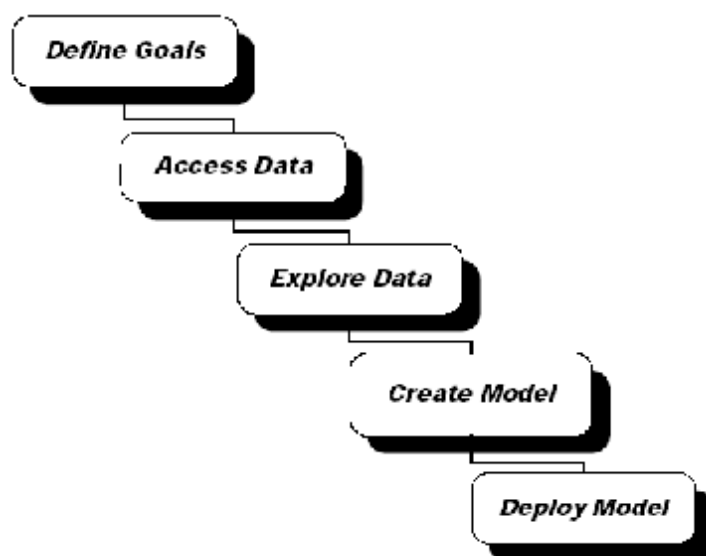


图 2.1 数据挖掘过程：目标定义，数据访问，数据探索，创建模型，和应用模型。

以下的部分，你可以进一步分解这些步骤，以便了解怎样利用S+Miner的先进模型和方析能力，把高度概括的视图转化成解决现实的数据挖掘问题的方法。

下面的 S+Miner 工作簿显示了图 2.1 中各个步骤概况的例子。网络流程图 2.2 的上半部表示访问，探索和建模，下部的表示使用 S-PLUS 模型进行评估。

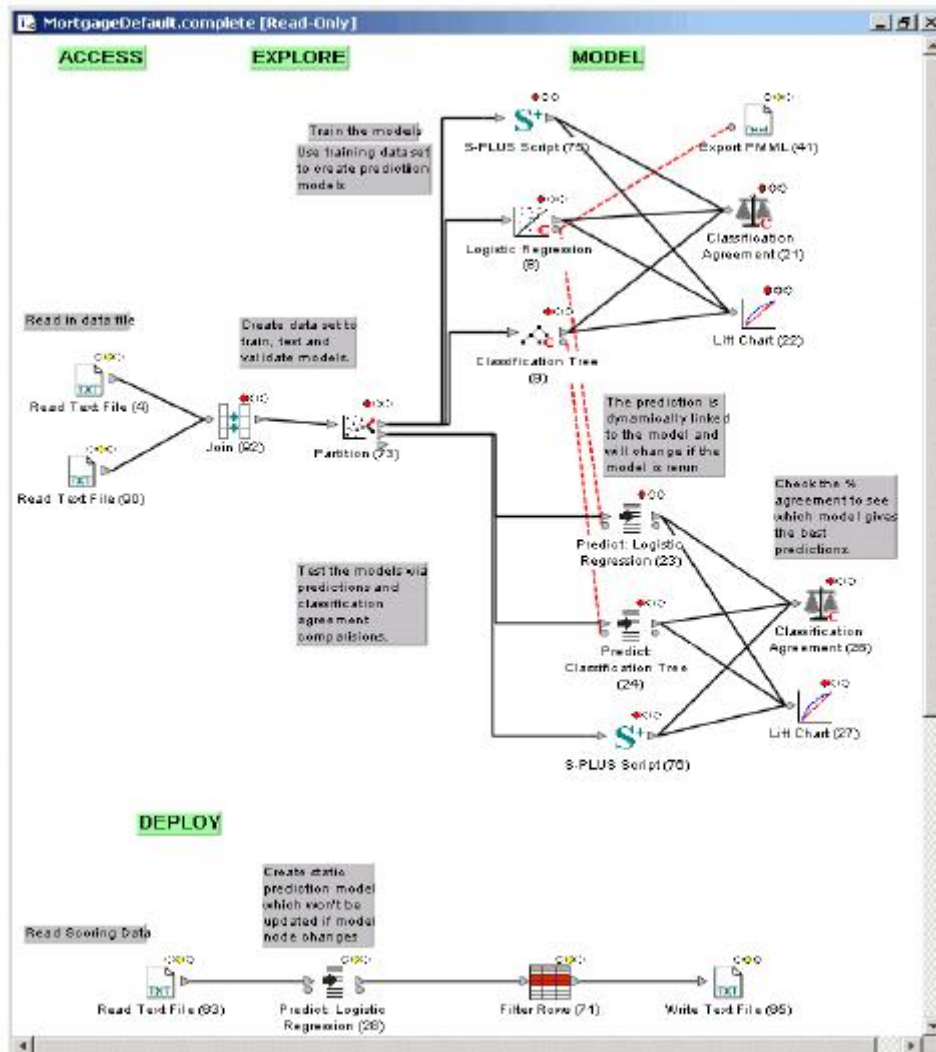


图2.2: 在文件名为MortgageDefault.complete.imw的工作簿中的完整建模流程网络。它预测消费者家庭抵押贷款中无拖欠的概率。

在这章扩展教程中的examples/MortgageDefaultExample文件夹里你可以找到例子和数据。你可以在一个工作簿中解决全部问题，如图2.2；而这个例子是通过一系列的工作簿来解决的。

## 2.2 目标定义

如上面绪论所讨论的，问题是根据已有的消费者数据预测获得无拖欠贷款的概率。最终的结果将是一份揭示消费者拖欠的最小可能即 ( $\text{Pr}(\text{NoDefault}) > 0.98$ ) (无拖欠的概率大于0.98) 的文件。

这个例子创建的模型包含两个数据集。第一个数据集的变量在图表2.1中。

图表 2.1: **mortdef.txt** 数据文件的变量

变量	描述
ID	消费者身份证号
Status	分类变量：拖欠或不拖欠
Delinquency	不良行为积分
PerPastDue	过期的应付比例，包括一部分本金和利息
MonthsPastDue	过期的月份数
CurrentLTV	当前的贷款值
paymentDiff	还款差额

第二个数据集是独立的信用报告机构给出的信用积分。第二个数据文件的变量如下：

图表 2.2: **mortdef.creditscore.txt** 数据文件的变量

变量	描述
ID	消费者身份证号
CreditScore	信用积分

这个例子的数据是从真实的家庭抵押贷款数据中修改得来的。在实际中，在训练和测试数据集中拖欠与无拖欠贷款的比率是要低的多。

随着目标的确立，你可以开始创建 S+Miner 挖掘网络来解决问题了。第一步是加载数据。

## 2.3 数据访问

数据来自于三个文本文件，如图表 2.3 所述。

图表 2.3: 抵押贷款拖欠模型和预测中应用到的数据文件。

文件名	描述
examples/MortgageDefaultExample/Mortdef.txt	消费者名单，包括贷款，还款历史及他们的贷款状态
examples/MortgageDefaultExample/Mortdef.creditscore.txt	mortdef 文本文件列出了每个消费者的信用积分
examples/MortgageDefaultExample/Mortdef.score.txt	用来预测哪个消费者会拖欠贷款的数据

这个例子的第一个阶段，你只需要第一个数据文件**mortdef.txt**。如果你不知道如何下手，先关闭所有的工作簿和窗口，然后仍然像概览里所述的那样打开文件。

1. 从主菜单中，选择 **文件** ► **新建** 来打开一个新的工作簿。

2. 在探索器窗口中，点击 **读文本文件** 组件，用鼠标拖拽到桌面框的工作簿中，释放鼠标按钮。  
初始时，**读文本文件** 节点状态指示器是红色的，表示它尚未具备运行条件。在你运行它之前，你必须设定它的属性。
3. 双击 **读文本文件** 节点来打开 **属性** 页对话框。



图2.3:读文本文件对话框的属性页

4. 点击 **浏览** 按钮，点击 **例子** 文件夹图标。（在窗口浏览器的左下角，Solaris操作系统在右侧。）  
复制例子文件夹中的内容从安装目录到默认的被操作系统定义的用户目录下的例子文件夹中，然后保留原始的需要再进入的例子文件夹。
5. 双击 **MortgageDefaultExample** 文件夹，然后从这个文件夹中选择数据文件 **mortdef.txt**。
6. 点击 **打开**。

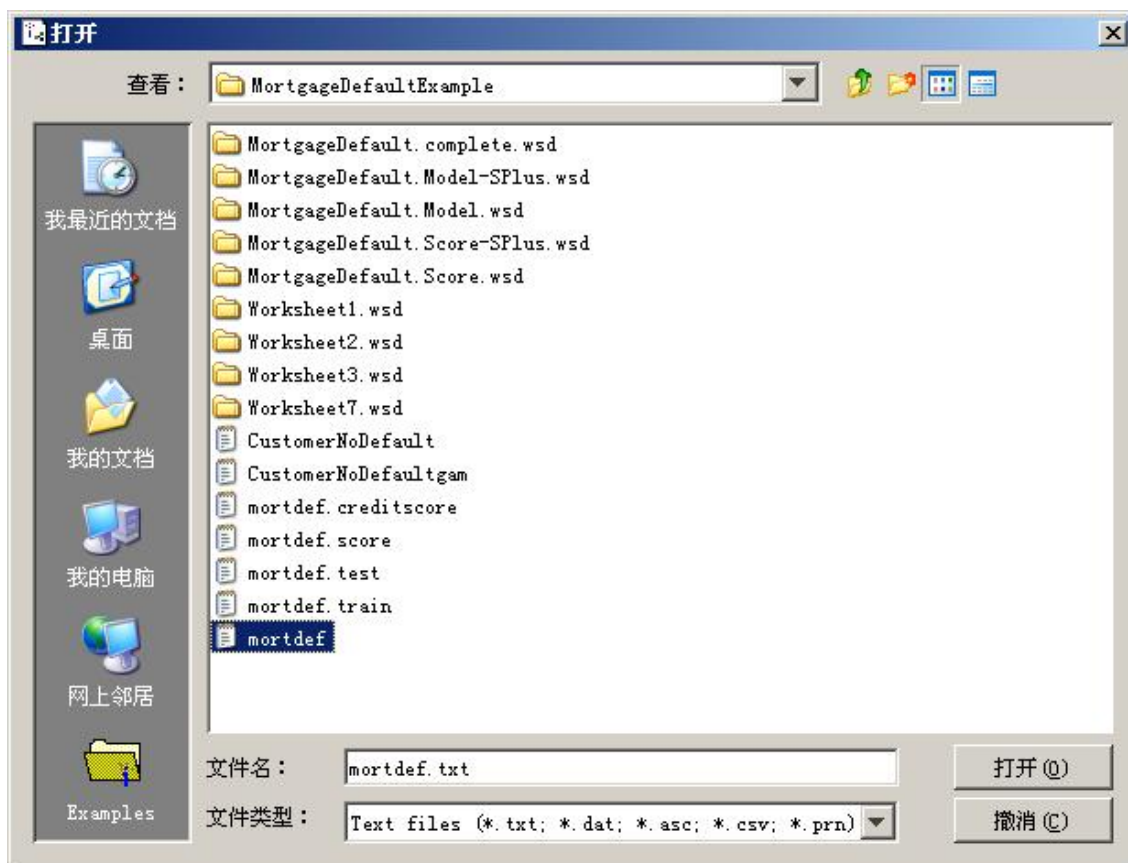


图2.4:通过浏览对话框打开例子文件夹。

**注：** 如果不想通过浏览找到文件，你可以在文件名文本框中键入文件名。如果你没有指定完整的路径，S+Miner在相同的文件夹里找文件作为工作簿。

7. 想要预览数据文件的前十行数据，点击 **预览** 选项组中的 **更新预览**。（这个部分完整的对话框如图2.3）。在默认状态下，每列数据都是以连续型的形式读入的，非数字性的字符列以字符串形式读入。字符串列一般用来储存鉴别信息且它的每一行都是不同的并不在模型中使用。为了更加了解每一列，检查查看器中值来得到更多的有关每一列包括的值的有关信息。可以以字符串形式来读入列，然后通过检查来确定它们适当的类型。

这个例子，以分类型形式读入status列，字符型形式读入ID列。数据中其他所有列都是数值型的，以连续型形式读入。用 **读文本文件** 对话框中的 **列定义修改** 页来改变这些变量的列的类型。

8. 点击 **列定义修改** 选项页。（这部分的完整对话框如图2.5）





图2.5: 读文本文件对话框中的列定义修改页面

**注:** 如果列不够宽, 不足以显示出全部的列名, 你可以通过移动鼠标扩展它们, 用鼠标拖拽两列之间的垂直线, 向左或向右拖, 直到这列足够宽, 释放鼠标。

根据你扩展的第一列的宽度, 你必须通过右侧的滚动轴来查看新列。

9. 点击包含变量**status**行的任意地方来选择它。

10. 在对话框右下方的 **设置类型** 选项组中点击 **分类型**。

注意到在 **新类型** 列中出现了提示, **status**的数据类型从字符串 (S) 变成了分类型 (C)。接下来也可以用这种方法来定义因变量。

11. 在 **设置角色** 组件中, 点击 **因变量**。

可以注意到在 **新角色** 列里出现 (D), 表示**status**被设定成因变量。同时这个信息也被加载到流程网络中。

12. 点击ID行的任意地方, 从 **设置类型** 组中选择**string**。

13. 点击 **确定** 按钮关闭对话框。


注意到读文本文件节点状态指示器这时变成黄色，表示现在已具备运行条件。但是，还需要读入第二个数据文件。

14. 右键点击工作表的空白区域，选择 **创建新节点**。

出现一个探索浏览框，你可以选择一个节点添加到当前工作簿中。

15. 选择 **读文本文件** 节点，点击**确定**。

16. 点击节点并把它拖拽到第一个节点的下方。

17. 从工具条中选择**属性**工具（）。

18. 点击 **浏览** 选择mortdef.creditscore.txt。

19. 点击 **打开**。

20. 点击 **列定义修改** 页。

21. 点击ID行的任意地方，在 **设置类型** 选项组中选择 **字符串**。

22. 点击 **确定** 关闭对话框。注意到 **读文本文件** 节点状态指示器变成黄色，表示它可以运行。

23. 点击S+Miner工具条上的**运行**按钮（）。

当两个节点都可以运行了，查看一下工作表底部的消息栏，有关于运行时间缓存大小的信息，和一些警告错误消息。当节点成功完成后，状态指示器变成绿色。图2.6显示了两个节点运行后的工作表。



图2.6: 前两个节点运行后的工作簿

## 2.4 数据探索

随着数据的读入，你可以更详细的检查它并为问题模型的建立阶段作准备。

进入第一个 **读文本文件** 节点的查看器。



1. 点击 **读文本文件 (0)**，然后再点击S+Miner 工具条中的查看器按钮。（图2.7显示了查看器对话框中 **连续型** 的页面）

图2.7: 节点查看器的**连续型**页面

在节点查看器的底部，显示了数据文件有7列和4979行。每种类型变量的个数也在右下角显示出来。查看器中的每个按钮都总结了不同类型的数据。第一个按钮显示的是全部的数据集。

节点查看器的 **连续型** 页面显示了数据集中每个变量的最大值，最小值，均值，标准差。最后一列显示了缺失值的个数。（如图2.7）。

你可以在概括性统计量页面中根据任意单个列对数据进行排序。

2. 通过点击它的列名，可以根据StDev列对数据进行降序排列，如图2.8所示。（再次点击就会根据StDev列对数据进行升序排列）

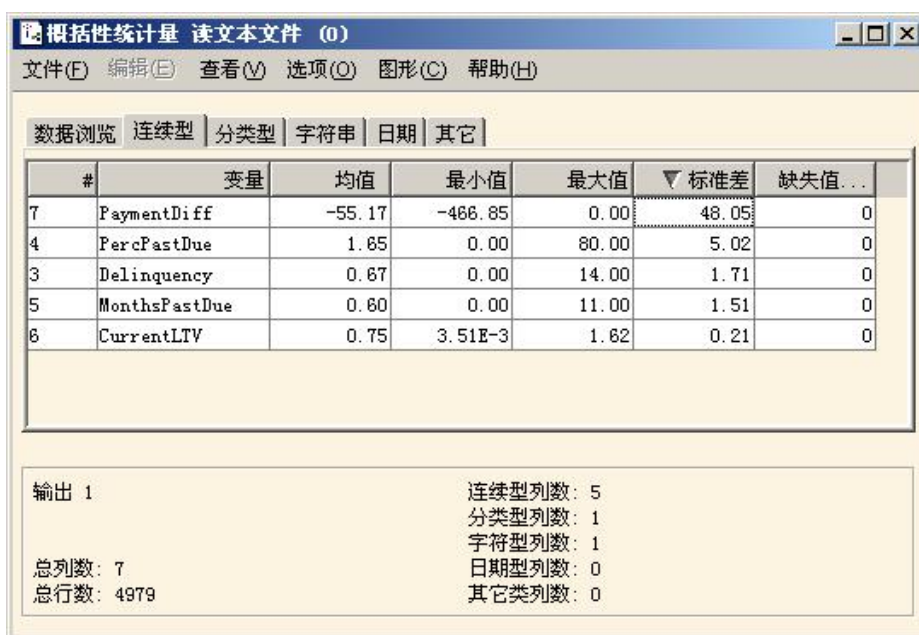


图2.8: 通过点击列名在节点查看器中对数据进行排序

可以注意到在变量名旁边出现了一个倒三角表示当前数据的排列状态。（如果三角形尖朝上表示数据按升序排列。）

3. 通过点击顶部菜单项你可以实现对查看器的操纵。例如：

¥ 从 **查看器** 菜单中，你可以以 **HTML** 形式查看浏览概括信息。

¥ 从 **编辑** 菜单中，你可以复制数据到剪贴板里。

¥ 从 **选项** 菜单中，你可以选择数的精度。

¥ 从 **图形** 菜单中，你可以选择不同形式为数据作图。

参看使用S-SPLU作图向导或是S+Miner用户手册来查找用S-PLUS作图并把它们作为新节点添加到工作簿中的更多信息。

4. 点击查看器节点的 **分类型**。你只有一个类型变量status，如图2.9所示。

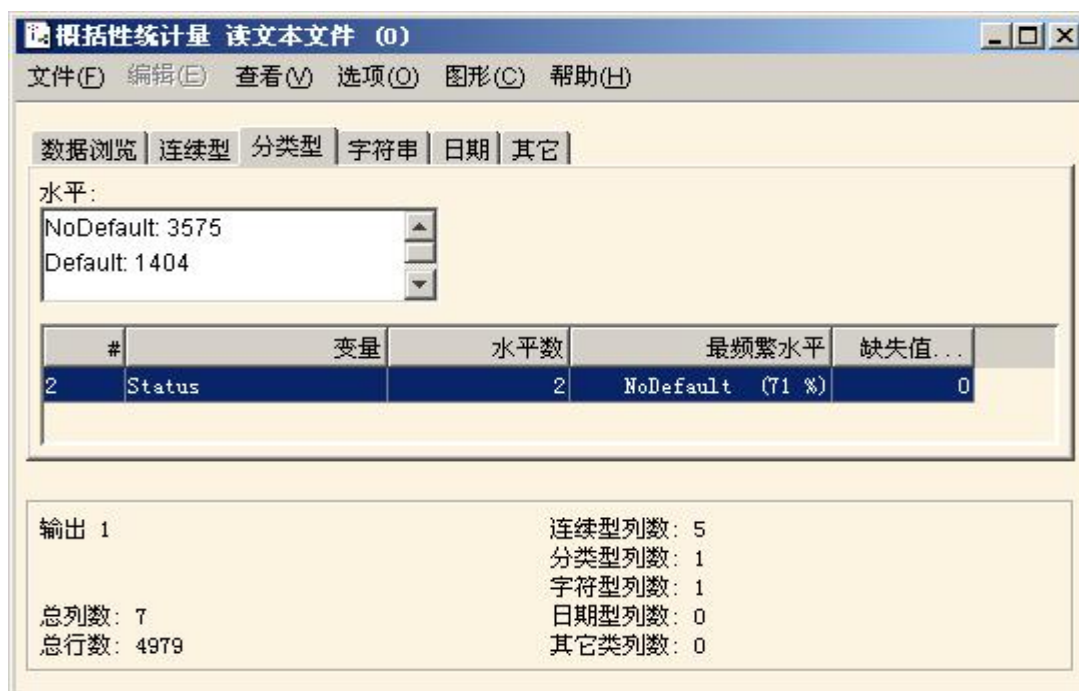


图2.9: 节点查看器的**分类型**页面

总结分类型变量中每个变量的水平个数，最频繁水平和缺失值个数。

5. 点击status变量所在行的任意地方。在页面的左上角的水平栏里显示了变量在不同水平上的个数。

可以看到在NoDefault水平上有3575个观测值，在Default水平上有1404个观测值。

6. 点击 **数据浏览** 来显示全部数据集，如图2.10你可以通过下部和右侧的滚动条来查看全部数据。

	ID	Status	Delinquency	PercPastDue	MonthsPastDue
	string	categorical	continuous	continuous	continuous
1	"1"	NoDefault	0.00	0.00	0.00
2	"2"	Default	1.00	5.00	0.00
3	"3"	NoDefault	0.00	0.00	0.00
4	"4"	NoDefault	4.00	0.00	0.00
5	"5"	NoDefault	0.00	0.00	0.00
6	"6"	NoDefault	0.00	0.00	0.00
7	"7"	NoDefault	0.00	0.00	0.00

输出 1

连续型列数: 5  
分类型列数: 1  
字符型列数: 1  
日期型列数: 0  
其它类列数: 0

总列数: 7  
总行数: 4979

图2.10: 节点查看器中数据浏览页面图

### 2.4.1 准备数据

一个消费者的信用度在预测这个消费者是否会拖欠贷款是很有意义的，所以需要通过合并两个数据集载入这个信息。

为了找出数据最好的模型，试图比较一些模型。

这个过程要分以下三个阶段来完成。

- Ø 训练模型
- Ø 测试模型
- Ø 检验模型

合并完数据文件后，把新的数据集分割成训练数据和测试数据。通过分割数据并用不同的数据测试模型，你可以对模型误差有个很好的估计。如果你希望最后所选择的模型是无偏的，你可以创建三个数据集：一个用来训练，一个用来测试，一个用来检验。为了方便，一般把数据分成两个子集：一个用于训练模型一个用来测试模型。

注意到 **分割** 节点右侧有三个黑色的三角形，表示有三个输出端口。你可以使用这些输出端口来输出为了不同操作而分割的数据，如写文件。（如图2.11的例子。）

如果你创建完两个数据集，完整的流程网络如图2.11。在examples/MortgageDefaultExample/MortgageDefault.Explore.imw 文件中提供了完整的工作薄的副本。通过在当前的工作簿上创建流程网络来继续这个例子。

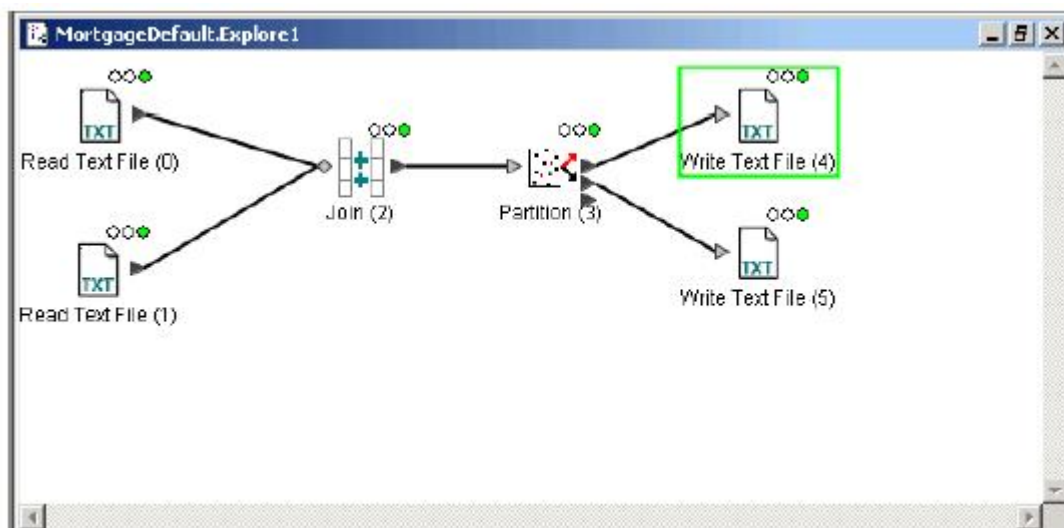



图2.11:读数据，联接数据，分割数据成两个数据集然后写入一个文本文件，整个过程的网络流程图。

## 合并数据


在这个练习中，首先把两个文件联接起来，然后分割数据。

为了把两个文件联接成一个文件，使用 **联接** 节点。两个文件都有ID列，你可以很容易匹配数据行。也可能有更复杂的联接，在S+Miner用户手册的第六章有关于 **联接** 节点的信息。

7. 双击 **数据操纵/列操纵** 文件夹下的 **联接** 节点。一个新的联接节点出现在工作簿中。拖拽这个节点到 **读文本文件** 节点右侧。
  8. 左击 **读文本文件(0)** 的输出端口拖住鼠标移动至 **联接** 节点上的输入端口，释放鼠标按键。一个联接出现在两个节点间。
  9. 重复以上步骤，连接 **读文本文件(1)** 节点到 **联接** 节点下面的输入端口。
- 现在联接节点的输入已经确定了，设定节点属性。在这个数据文件中，每个消费者的ID都是一对一的，所以你不需担心没有匹配上的行。完整的属性页如图2.12。
10. 右击 **联接** 节点图标然后从菜单中选择 **属性**。
  11. 在 **为所有输入设置属性** 选项组中，选 **键** 的下拉列表框里的值为1。在 **键** 值右侧的下拉列表框中选择ID。点击 **设置所有输入**。
  12. 点击 **确定** 关闭对话框。

13. 点击工具栏中的 **运行至此** 按钮()。

为了确保数据正确的合并，打开查看器并核对列名和类型。

14. 点击S+Miner工具栏中的查看器按钮()。打开 **数据浏览** 页，通过翻看滚动条可以看出最后一列现在是CreditScore。

至此，你已经创建了训练数据集和测试数据集。




图2.12:联接节点完整的属性页

## 分割数据

15. 在探索框中 **数据操纵/行操纵** 文件夹下，双击 **分割** 节点。把分割节点放在联接节点的右侧。
16. 左键点击 **联接** 节点的输出端口拖拽鼠标与 **分割** 节点相连。
17. 右键点击 **分割** 节点，并选择 **属性**。
18. **训练** 文本框中键入70，**测试** 文本框中键入30。

为了可重复测试这个例子，在 **分割** 节点中设定一个随机种子。

19. 点击 **高级选项** 选项页，点击 **输入随机种子**。使用默认值5。点击 **确定**。



20. 在工具栏中，点击**运行**().


在信息栏里可以注意到，只有 **分割** 节点运行了。节点指示器是绿色的就不再运行了。**分割** 节点上端的输出端口输出了70%的随机样本，剩下的30%从输出端口下端的节点输出。接下来，把两个数据集导入文件里为后续工作使用。

## 写文本文件



21. 拉动 **探索器** 框滚动条至下部的 **数据输出/写文件** 文件夹。找到文件夹下的 **写文本文件** 组件。
22. 添加两个 **写文本文件** 节点到你的工作簿，把它放在 **分割** 节点后边，如图2.11。
23. 联接 **分割** 节点到 **写文本文件** 节点。

**注：**如果想删除节点连线，可以右击连线然后选择删除连线。也可以通过右键点击连线去掉直连线前面的对号来改变连线形状，从直连线到直角连线。同样，你可以把所有的连线都改成直角连线，通过点击编辑  全选，然后点击查看  切换连接方式。

24. 双击上面的 **写文本文件** 节点来打开 **属性** 页。
25. 点击 **浏览**，然后点击 **例子** 图标。
26. 打开MortgageDefaultExample文件夹，然后在 **文件名** 文本框中键入 **Mymortdef.train.txt** 点击**打开**。
27. 把 **分割符** 文本框里改成single space delimited，然后点击 **确定**。
28. 双击下面的 **写文本文件** 节点打开 **属性** 页。
29. 点击 **浏览**，然后打开MortgageDefaultExample文件夹。在 **文件名** 文本框中键入 **Mymortdef.test.txt**。点击**打开**。如果它没有存在就创建了一个新的文件，或是覆盖一个已存在的文件。(如果你不想创建一个新的文件或是覆盖一个已经存在的文件那么可以换一个文件名。)
30. 点击 **确定**。
31. 点击工具条上的 **运行** 按钮()。

当你只提供一个文件名时，默认路径是你的工作簿地址。如果你还没有保存工作簿，在这种情形，默认路径是你默认的用户名路径，具体取决于你的系统。例如：


**c:/Documents and Settings/username/iminer\_work\_8\_0/examples。**

训练数据和测试数据都在这个路径下。在例子目录下

(examples/MortgageDefaultExample/MortgageDefault.Explore.imw) 完整的工作簿表明你可以联接或收集 **读文本文件(1)**，**联接** 和 **分割** 节点创建一个聚合节点，这个节点联接数据文件再分割数据文件成两个数据集。你可以增加这个节点到你的 **用户库** 里，以备将来使用。在 **用户手册** 里提到了如何创建和使用聚合节点，并把它们加入 **用户库** 里。

## 2.4.2 保存工作簿

如下保存工作簿::

32. 从主菜单中选择，**文件**  **保存** 并选择一个位置保存文件。
  33. 在**文件名**文本框中键入文件名并点击 **保存**。你键入的文件名被自动附上扩展名**.imw**。
- 接下来，进行建模。

## 2.5 创建模型

现在所说的建模都是建立预测模型。用包含已知目标变量的训练集，来产生一个预测模型。你可以使用这个模型来进行预测，叫做为目标变量打分，在这个例子中，你将要预测消费者拖欠贷款

的概率。

首先，用数据产生训练模型，然后使用各模型来预测并与观测到的数据进行比较。你可以通过添加以前的工作簿和流程网络来继续这个例子，如图2.2 (**examples/MortgageDefaultExample/MortgageDefault.complete.imw**)。

取而代之的，你也可以开始一个新的工作簿来读取你前一部分输出的训练集和测试集文件，探索数据。最后的挖掘网络如图2.13。完整的工作簿在 **examples/MortgageDefaultExample/MortgageDefault.Model.imw** 目录下。

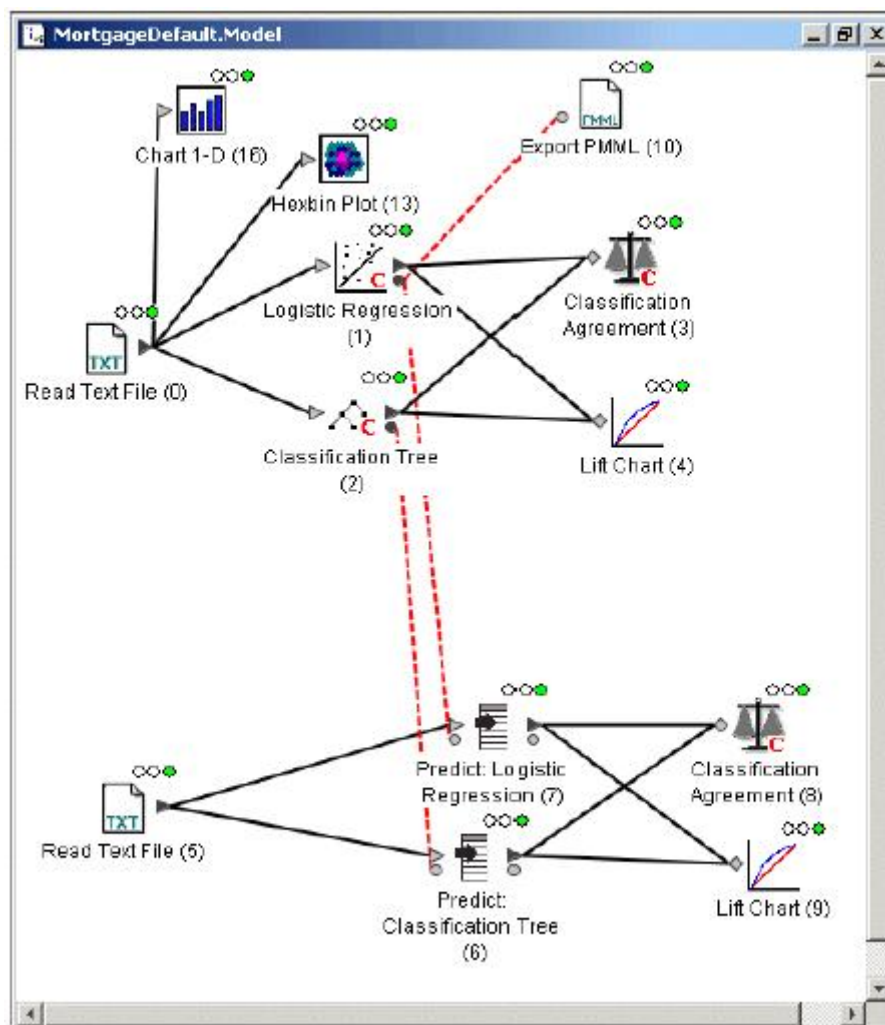


图2.13:模型阶段完整的工作簿，MortgageDefault.Model.imw。

### 2.5.1 导入训练数据

首先，打开一个新的工作簿然后读取训练数据，从探索数据部分里通过分割合并数据集得到的数据。

1. 在主菜单中，点击 **文件** ► **新建**。
2. 在探索器窗口中在 **数据读入/读文本文件** 文件夹下点击 **读文本文件** 组件添加这个节点到工作簿中。用鼠标拖拽节点到离左框大约1寸的位置放下。
3. 双击 **读文本文件** 节点打开 **属性** 页。
4. 点击 **浏览**，然后打开文件夹 **examples/MortgageDefaultExample**。选择 **mortdef.train.txt** 文件，点

击打开。

设定属性。

5. 通过点击 **更新预览** 查看数据文件的前十行数据。
6. 点击 **列定义修改** 选项页。
7. 点击Status变量所在行的任意地方。
8. 在 **设置角色** 组中，点击 **因变量**。在 **设置类型** 组中，点击 **分类型**。
- 在建模过程中去除ID行。
9. 点击列名ID。在 **选择列** 组中点击 **排除**。
10. 点击 **确定** 关闭对话框。

现在，读入数据文件时没有读入ID列。关于输入数据的详细介绍参看S+Miner用户手册的39页。  
完整的对话框如图2.14。点击 **确定** 接受改变。



图2.14:读训练数据页面中完整的列定义修改页。




## 2.5.2 作图

如果用这个数据集来评估财物贷款来确定拖欠的风险，可以用一个预测模型来检查数据模式。这个模型使用贷款信息(过期时间，当前贷款值，支付差额，置信度和逾期借款)来预测拖欠的概率。

为了得到初始的观点，首先使用一维图形节点来做出数据直方图，确定哪个列能给你提供足够的信息来确定拖欠与无拖欠的概率。

11. 拖拽 **一维图形** 节点到工作簿。

12. 双击 **一维图形** 节点打开 **属性** 对话框。

13. 在 **属性** 页中，在 **可用列** 选项组中点击变量status通过点击(  )把它添加到 **分组字段** 列表框中。

14. 选择 **可用列** 列表框中其余的变量通过点击按钮(  )把它们添加到**显示组**中

15. 点击 **确定** 关闭对话框，右键点击选择 **运行至此**，来运行挖掘网络。右键点击选择 **查看器** 显示直方图如图2.15。

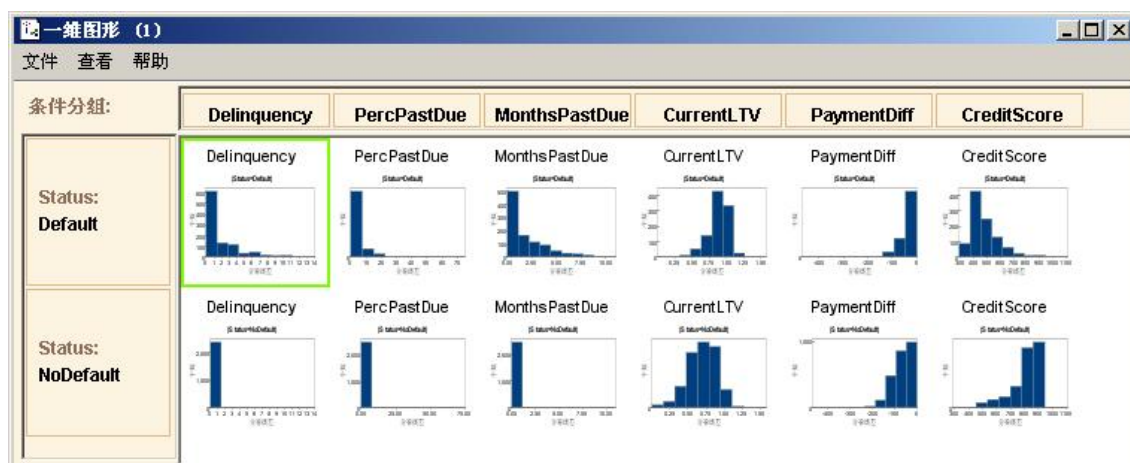


图2.15:信用打分和当前贷款值变量的直方图

这个直方图展示了置信度和当前贷款值变量，表现了一些我们感兴趣的变量status值特别是 Creditsore 和CurrentLTV。

接下来，使用六边分箱图来进一步估计这两列。因为六边分箱图不是对每行数据的独立点进行作图，可以用于大型数据集并展现出易懂的图。你也可以使用六边分箱图具体处理 **所有行**，实现大型数据Trellis图的特征。

在这个例子中，设定六边分箱图的x轴和y轴分别为Creditscore和currentLTV，设变量status为条件变量。

### 设定六边分箱图节点的属性

16. 在探索器窗口中点击 **S-PLUS** 页

17. 在 **二维图形-连续型** 文件夹中，双击 **六边分箱图**，添加一个 **六边分箱图** 节点到工作簿中。

18. 把 **六边分箱图** 节点放到 **读文本文件** 节点的右上方，然后连接节点。

19. 双击 **六边分箱图** 节点显示它的 **属性** 对话框。
20. 在 **属性** 页中，在 **x轴** 文本框中选择 **currentltv**。在 **y轴** 文本框中选择 **creditscore**。
21. 在 **条件分组** 文本框中选择 **status**。
22. 在 **行处理** 选项组中，选择 **所有行**。(使用大型数据库的Trellis函数时选择所有行。)
23. 你可以查看对话框其他的选项，可以接受默认选项。 更多有关使用 **六边分箱图** 节点和其他 S-PLUS图节点工具的信息，请参看S+Miner用户手册中的S-PLUS库章节。
24. 在 **数据** 页面中，点击 **应用** 显示 **六边分箱图**。图形使用标准色，如图2.16。

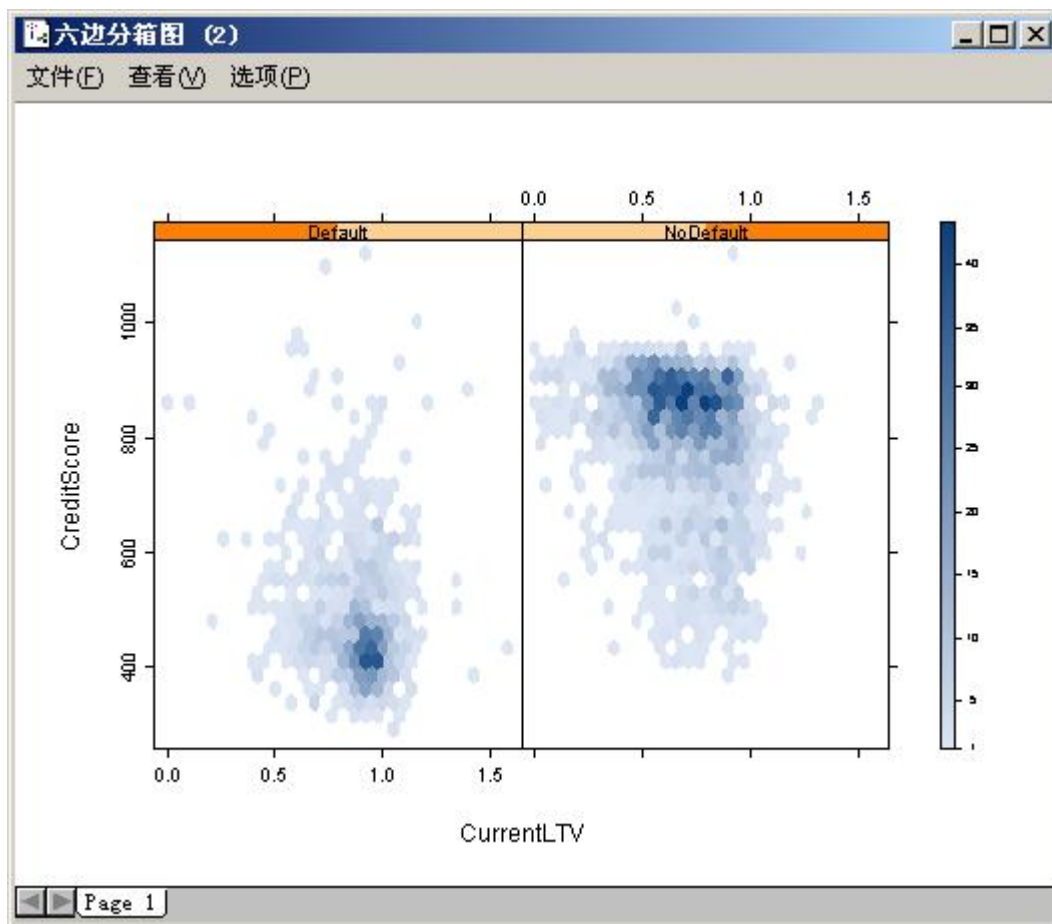


图2.16:抵押拖欠贷款例子的六边分箱图

注意到这个例子的数据，消费者有较低的置信度和相对高的当前贷款等级值，这意味着冒着较大的风险拖欠贷款。

## 改变展示图的颜色

如果你不想让展示图显示背景颜色，你可以改变六边分箱图的颜色。通过把**Default** 变成**Standard**，如图2.16:

25. 在图窗口的主菜单下，点击 **选项** ► **设置颜色**。
26. 点击 **标准**。

同样你也可以通过点击 **设置颜色** 对话框中的 **编辑颜色**，来选择一种新的主题。改变成单个的



颜色。或是混合的颜色，参看S+Miner用户手册查看更多信息。

### 2.5.3 训练模型

因变量是一个分类型变量，它需要进行分类型检验。在这个例子中，使用Logistic回归和分类决策树两种模型。这两个模型对于预测分类型变量的数据都是有效的。

27. 在 **Insightful** 探索器中 **模型/分类型模型** 文件夹中，双击 **Logistic回归** 组件添加节点到工作簿中。把它与 **读文本文件** 节点相联。


28. 重复以上步骤来创建 **分类决策树** 节点。把这个节点放在 **Logistic回归** 节点下方，如图2.13。

29. 双击 **Logistic回归** 节点来打开属性页。注意到status变量有两个标志  和  因为当你读入数据的时候，你已经定义status变量为因变量。

30. 点击 **自动** 按钮，把status变量移动到 **因变量列** 列表框中。

因为当你读入数据时你没有设定 **自变量**，所以你必须手动移动它们。

31. 点击选择**Delinquency**，然后按住SHIFT键点击最后一个变量**Creditscore**。整个列就都被选择了。

点击右侧的双箭头按钮  把这些变量导入 **自变量列** 列表框中。

**注：** 你可以在自变量列表框中选择有交互作用的项并点击交互键把交互项也添加到自变量列表框中，通过选择两个以上变量然后点击交互按钮来添加交互项。你也可以通过点击左侧的双箭头按钮把它们移出列表框。

核查确保模型的其他属性都正确设定。默认情况S+Miner返回值为因变量最后水平的概率结果。改变默认设定来得到你希望的预测结果。

32. 点击 **输出**。在 **新列** 组中，选择 **指定类别**，然后在下拉列表框中选择**Default**。

33. 为了在随后作图过程中含有 **自变量**，可以在 **复制输入列** 选项组中，选择 **自变量**。

34. 完整的对话框如图2.17所示。点击 **确定** 完成属性页的设置。



图2.17: 完整的**Logistic 回归**节点**属性**对话框中**输出**的页面。


把 **分类型决策树** 节点的属性值设定成与 **Logistic回归** 节点相同的。

35. 双击 **分类决策树** 节点打开它的 **属性** 对话框。

点击每个按键来查看它的默认值。对于这个例子来说大多数的设定都是很好的；然而，你必须设定自变量并输出保存它们以备以后之用。

36. 重复以上的30-34步骤，然后点击 **确定** 按钮完成属性值的设定。

## 运行模型节点

37. 点击工具栏上的 **运行** 按钮  来运行挖掘网络。

## 2.5.4 查看模型

你现在已经有两个数据模型。分别检查它们来理解它们之间的差别。

38. 右击 **Logistic回归** 节点选择 **查看器**。这个节点的查看器是以HTML形式显示的，如图2.18

**注：**在窗口中，S+Miner通过连接打开。html形式的文件（例如因特网）。在Solaris操作系统中，默认连接到Netscape浏览器中。

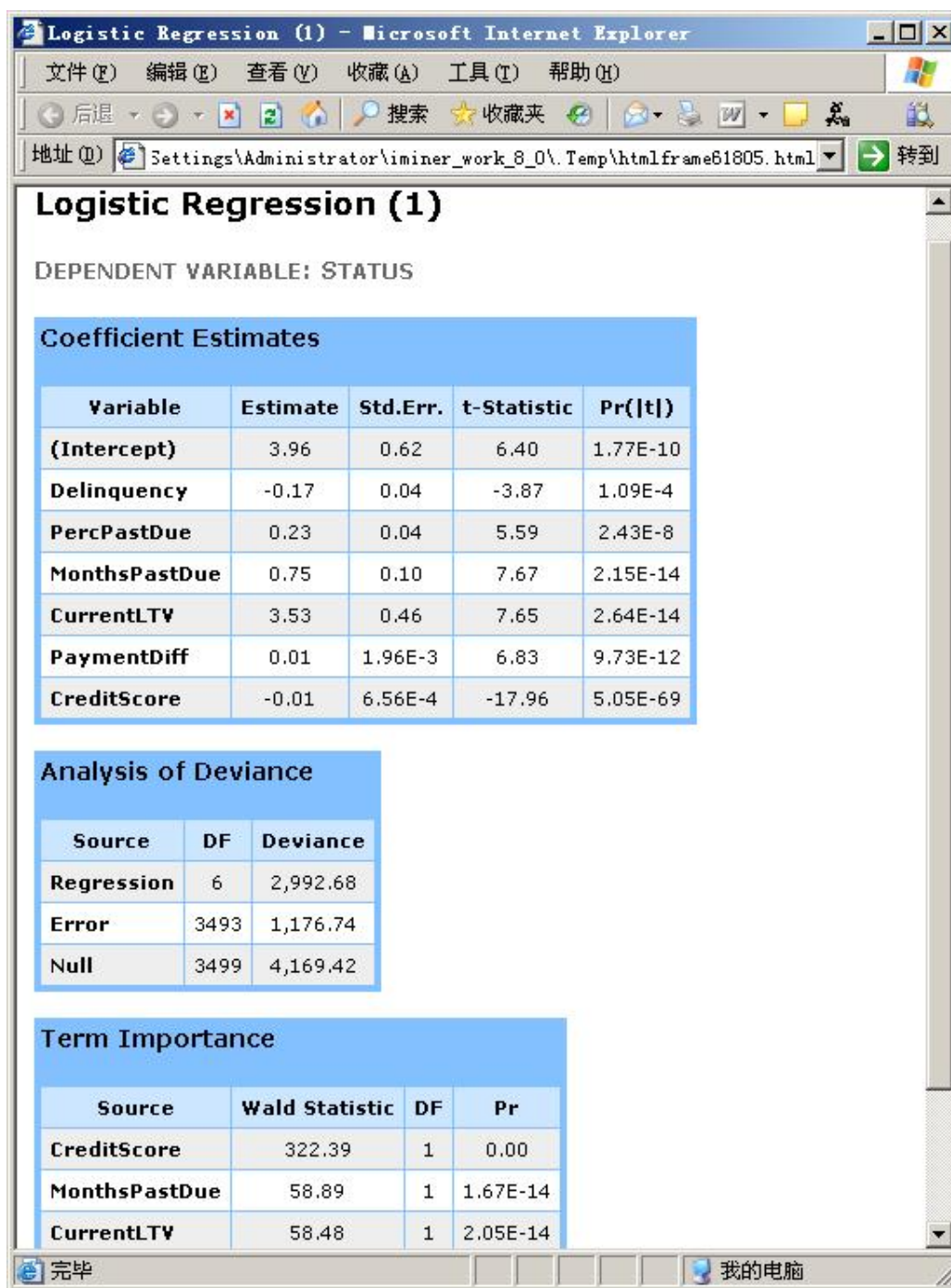


图2.18: Logistic回归节点的查看器。

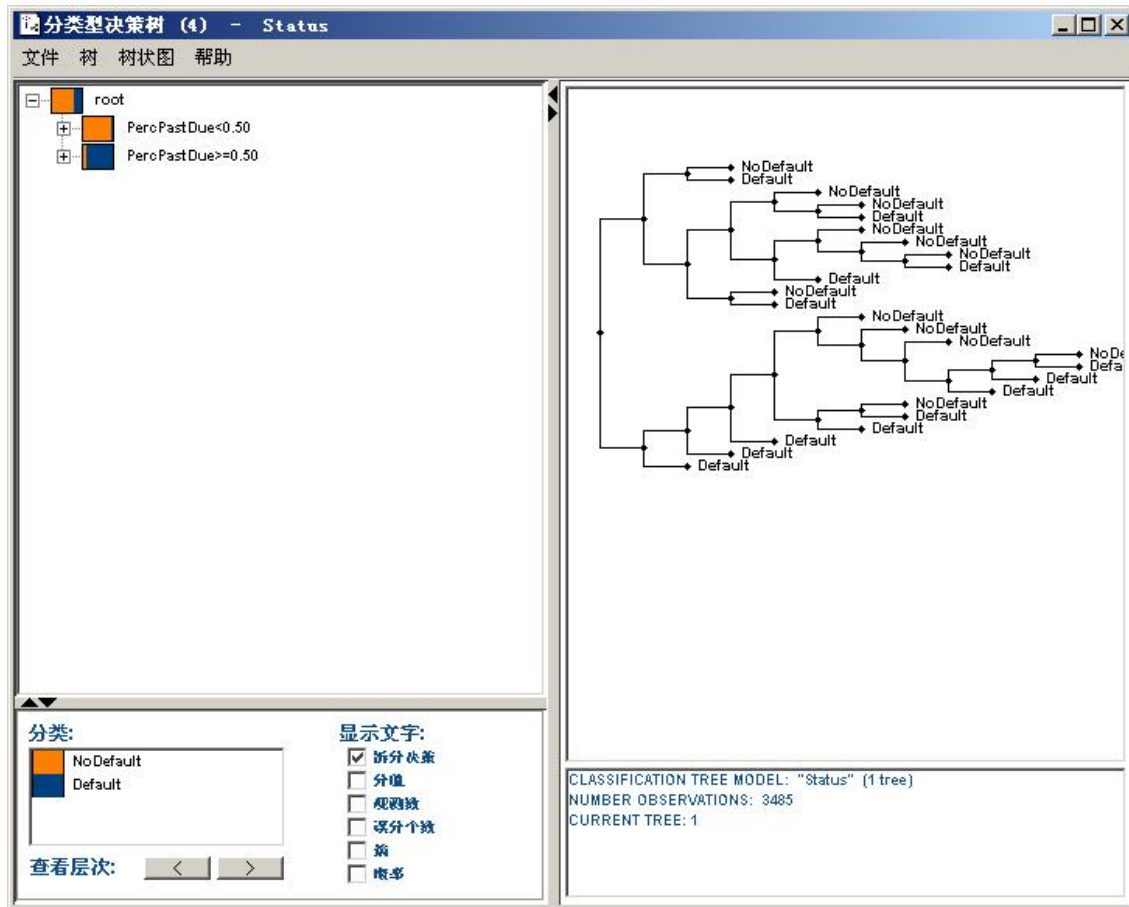
**Coefficient Estimates**表和**Term Importance**表表明这个模型的自变量都是有显著的。这个节点顶部的三个变量是CreditScore, MonthsPastDue, 和CurrentLTV。注意到, 当CreditScore和PaymentDiff是显著的, 但它们的系数就非常小。你可以通过添加交互项来改进模型; 但是这个例子不包含这部分练习。

S+Miner中的 **分类决策树** 节点储存了关于树中所有变量的信息, 包括每个拆分的重要性。打开分类决策树节点的浏览器 (如图2.19) 查看**Relative Term Importance**图。

39. 右击 **分类决策树** 节点, 点击 **查看器**。

40. 通过点击在窗口左上部PercPastDue<0.50前面的加号标志来展开节点。

你也可以点击窗口右侧文本框的节点来展开对树的浏览。



图

2.18, 展开层次探索框第一层的分类决策树节点的查看器


41. 点击 **树** ► **查看列重要性** 来显示相对的列重要性的箱图

图2.20中的条形图显示了模型中每个列在离差上的改变。对于每个分拆，由于分拆离差的改变，你可以知道列（变量）分拆。在模型中这些列在离差上大的改变是很重要的，并在图形的上端显示。最重要的预测变量是那些在各自列中值大于零的。你可以用这些信息来过滤出数据集的列，这些数据的离差接近零。





图2.20: 分类决策树模型的变量对照图

42. 点击每个窗口右上角的关闭按钮 () 关闭每个查看器。

### 2.5.5 选择模型

你希望了解用这个模型来预测个体拖欠贷款问题的有效性。**分类吻合度** 组件提供的交叉矩阵 (一个模型有一个交叉矩阵) 揭示了这些信息。除此之外 **提升图** 组件也提供了一个参照图来帮助你评估模型。

43. 从探索器窗口中的 **评估/分类型** 文件夹, 添加 **分类吻合度** 和 **提升图** 节点到工作簿中。把这些节点连接到模型的输出端口中。

注意到这些新节点的输入端口是个菱形块而不是三角形。这表明这个节点可以接受不止一个输入。

44. 运行挖掘网络。

45. 打开 **分类吻合度** 节点的查看器 (如图2.21) 通过滚动条浏览窗口。

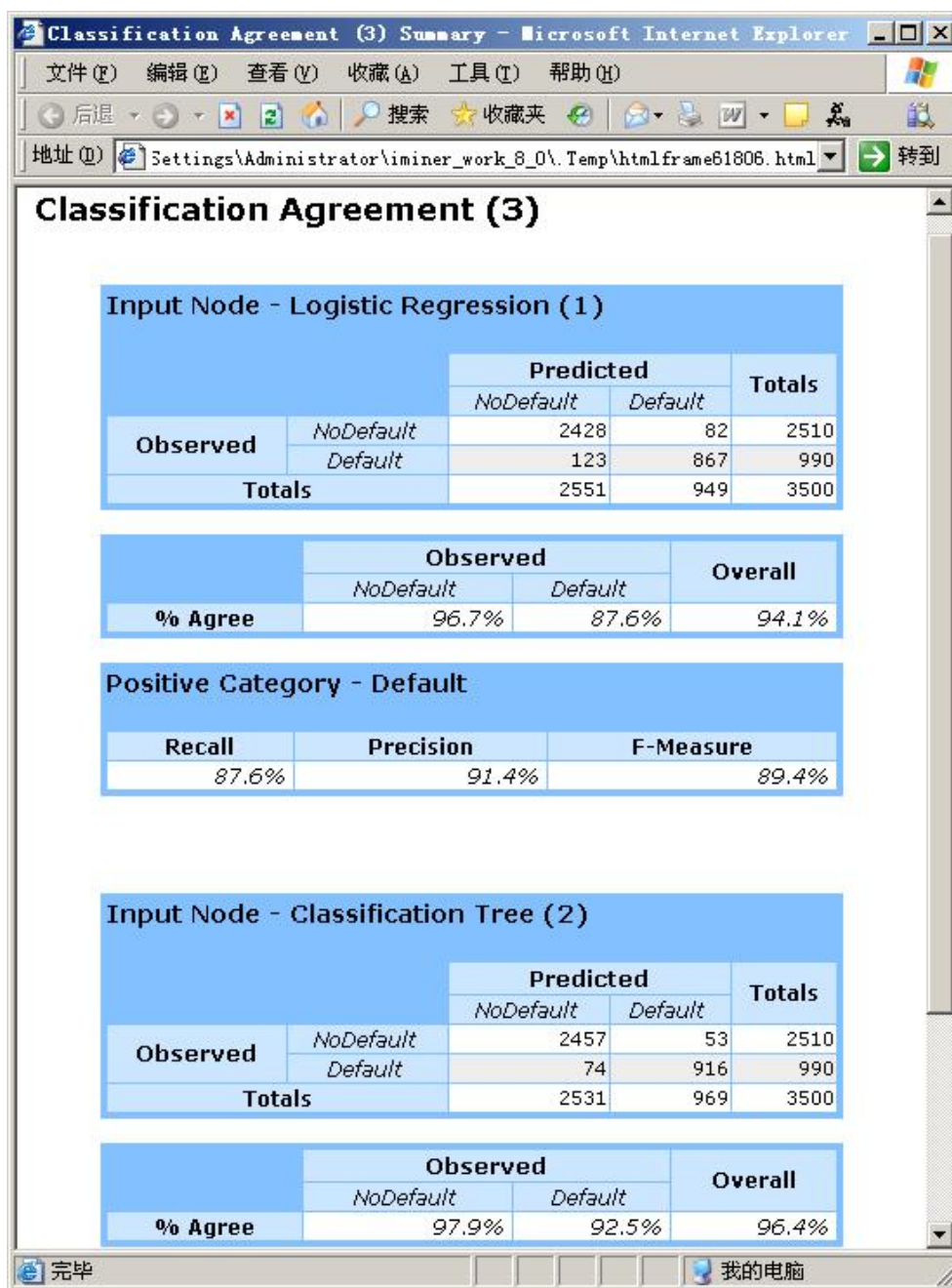


图2.21：分类吻合度节点的查看器。

**分类吻合度** 组件产生的交叉矩阵揭示了已被模型分类的观测值的数量和比例。

分类决策树有很高的正确预测度（最大的预测度百分比）可以达到96.4%。Logistic回归模型的预测比率是94.1%

46. 关闭 **分类吻合度** 节点的查看器。

47. 打开 **提升图** 节点的查看器（如图2.22）

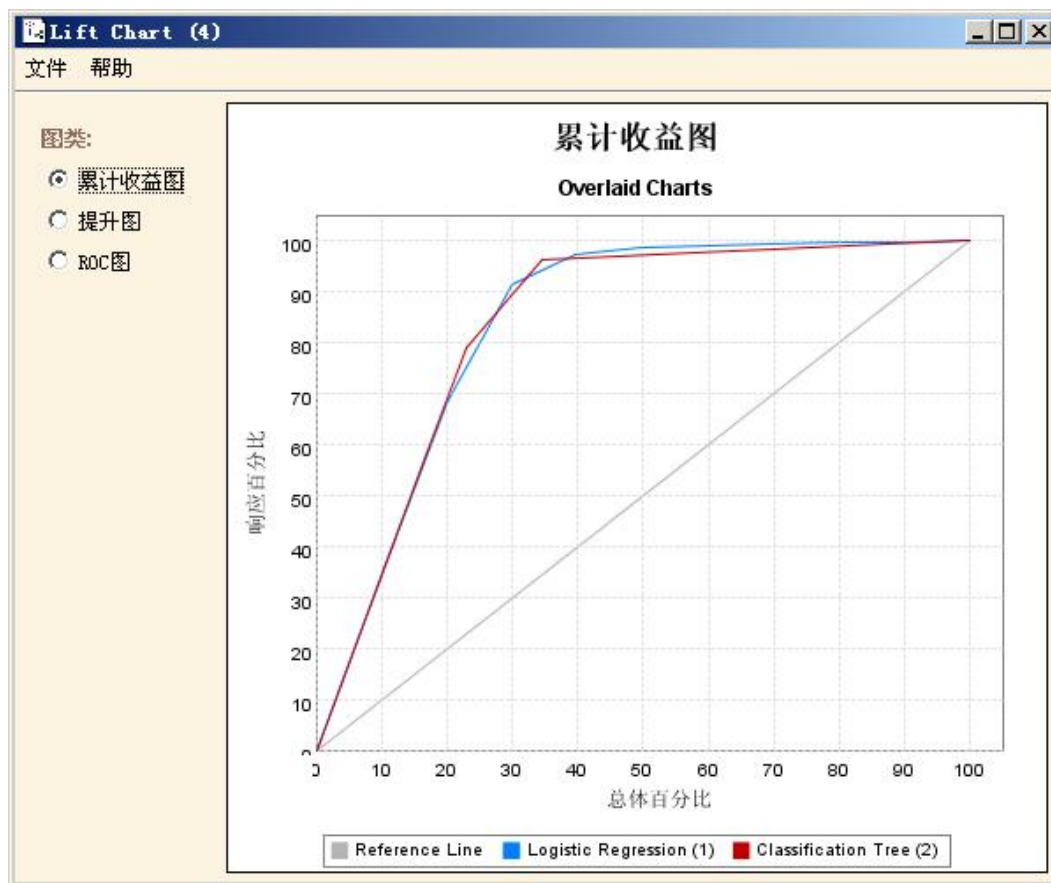


图2.22: 提升图节点的查看器

提升图节点查看器提供了这个模型的比例图。这个组件产生了三个不同的图：提升图，累计收益图，和ROC图。图形由提升曲线和基准线组成。提升曲线和基准线间的面积越大，模型越好。正如你看到的图2.22，分类决策树模型提供了最大的累计收益。

48. 关闭 **提升图** 节点查看器。

## 模型测试

现在你已近创建了模型并做了初步的对照，你可以用一个新的数据集测试模型。首先，读入你在探索工作簿中创建的测试数据。然后在两个模型里创建预测节点，然后把这个预测应用到测试数据中。

### 读入测试数据

49. 添加一个 **读文本文件** 节点到工作簿中，放在已存在的 **读文本文件** 节点的下方。
50. 打开 **属性** 页对话框，点击 **浏览**，选择 **mortdef.test.txt**。
51. 点击 **列定义修改**。
52. 点击变量名为 **status** 所在行的任意地方。
53. 在 **设置角色** 选项组中，点击 **因变量**，并在 **设置类型** 选项组中点击 **分类型**。
54. 点击列名ID，在 **选择列** 选项组中点击 **排除**。
55. 点击 **确定** 关闭对话框。

## 创建预测因子节点

56. 右击 **分类决策树** 节点，在弹出的对话框里选择 **创建预测节点**。

57. 注意到 **预测：分类决策树** 节点出现在在工作簿中由红色虚线连接到模型中。如果需要，调整新节点的位置，连接到**读文本文件**节点的输出中。

**注：**模型的输出端口带一个小圆圈区别与其他是三角形或菱形的输入输出端口，删除模型连接即可创建一个静态预测节点，意味着即使模型有所改变预测模型也不会改变。

58. 打开 **预测：分类决策树** 节点的属性对话框。完整的对话框页面如图2.23：



图2.23：预测：分类决策树节点对话框的属性页。

59. 为了增加自变量到输出数据，在 **复制输出列** 选项组中，选择 **自变量**。

60. 点击 **确定** 关闭对话框。

61. 对 **Logistic回归模型** 节点重复步骤56-60。

62. 运行挖掘网络。

## 比较模型

接下来，看一下模型在测试数据上的预测结果。好的变量选择意味着训练数据正确分类的百分比与测试数据的百分比是相近的。一般来说测试数据的正确分类的百分比略低。同时，并不希望累计收益或是提升图有很大的区别。如果又很大区别，就应该调整模型。

63. 添加 **分类吻合度** 节点和 **提升图** 节点到工作簿。

64. 连接这些节点到 **预测** 节点的输出端口。

65. 运行网络。

66. 打开两个评估节点的查看器。

这两个模型的预测比率很接近，分类决策树的94.2%与Logistic回归的94.8%。用这个结果来决定选择哪个模型，感觉有点牵强。为了演示目的，这里选择Logistic回归模型并以PMML模型文件形式输出。通过输出模型，你今后可以在另一个工作簿里使用它来评估数据。

## 2.5.6 输出模型

67. 从探索器窗口中的 **模型/文件** 文件夹，创建一个 **输出PMML** 节点把它放在 **分类吻合度** 节点上方，如图2.13。

68. 连接 **Logistic回归** 模型节点右边的端口到 **输出PMML** 模型节点的输入端口。

69. 打开**输出PMML**节点的属性对话框并设定PMML文件名为**logisticRegModelMortgage.xml**。

70. 点击 **确定** 关闭对话框。

71. 在工具条上点击**运行至此**按钮（）。

通过这个例子可以看出，输出PMML节点的查看器和Logistic回归节点的查看器相似；然而，不总是这种情况。通常，建模过程的最后一步是用检验数据来评估最后所选模型产生的误差。因为你使用训练数据来构建模型并用测试数据来选择模型，根据这些数据集测量的误差是偏高的，高于使用新数据测量你所希望看到的。考察新数据集（检验数据）能提供无偏的误差估计。这个练习不演示这一步骤。

## 2.6 发布模型

Insightful方法对数据进行挖掘的最后步骤是打分/发布模型。在这步里，使用一个新的数据集，来预测每个消费者无拖欠的概率。

为了模拟真实的情况，创建一个新的工作簿，然后输入新数据到模型中。预测不会拖欠贷款概率>0.98的消费者，并把结果写入文本文件。完整的工作簿如图2.24

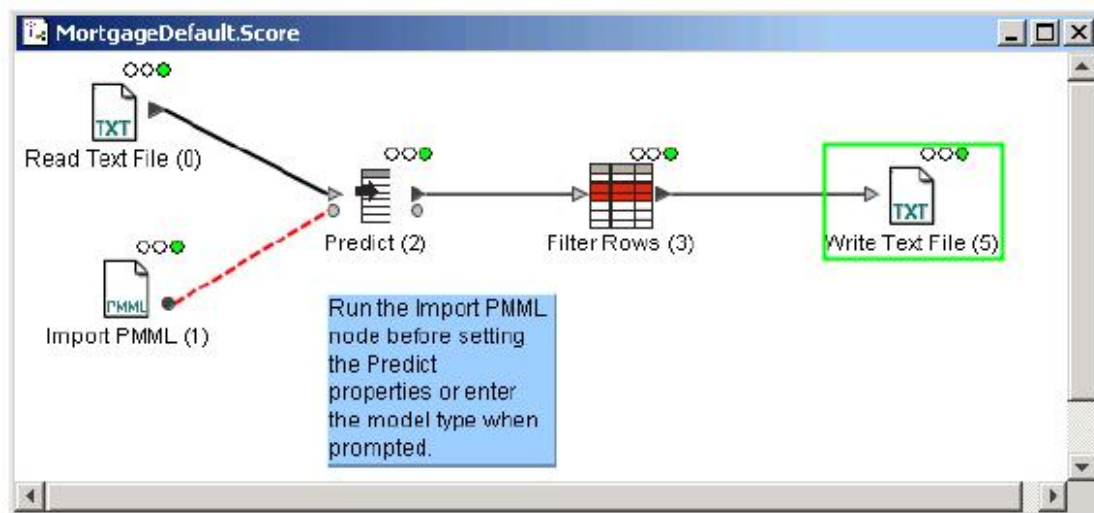


图2.24：使用一个输入模型评估新数据的完整工作簿。

添加前三个网络节点到工作簿，然后为每个节点设定属性。**读文本文件** 节点的完整的属性对话框如图2.25。

图2.25：打分数据集的**读文本文件**节点的完整对话框



## 2.6.1 输入打分数据


1. 从探索器窗口下的 **数据读入/读文件** 文件夹下双击 **读文本文件** 组件，添加一个 **读文本文件** 节点到工作簿。
2. 从探索器窗口下的 **模型/文件** 文件夹下双击 **读入PMML** 组件并添加 **读入PMML** 节点到工作簿。放在 **读文本文件** 节点的下方。
3. 从探索器窗口下的 **模型/预测方法** 文件夹下双击 **预测** 组件，添加 **预测** 节点到工作簿。放到其余两个节点的右侧。
4. 连接 **读文本文件** 节点的输出端口到 **预测** 节点的输入端口。
5. 连接 **读入PMML** 节点的输出模型端口到 **预测** 节点的输入模型端口。
6. 双击 **读文本文件** 节点打开属性对话框。
7. 点击 **浏览**，选择mortdef.score.txt然后点击 **确定**。
8. 点击 **列定义修改**。（完整的对话框页面如图2.26）



图2.26: 输入打分数据的 读文本文件 节点的 列定义修改 页面的完整对话框。

9. 点击ID变量行的任意地方。
10. 在 设置类型 选项组中, 点击 字符串。
11. 点击 确定 关闭对话框。

## 2.6.2 模型的输入

12. 双击 读入PMML 节点打开属性对话框。
13. 键入logisticRegModelMortgage.xml到PMML文件名文本框或是从浏览中选择这个文件。
14. 当节点仍被选择的时候通过点击 运行至此  按钮运行节点。

## 2.6.3 预测

15. 双击 预测 节点打开属性对话框。在预测节点的 属性 对话框中, 在 复制输入列 选项中清除 因变量 文本框然后选择 自变量 和 其他。点击 确定。  
象前面的第15步描述的那样为了阻止选择变量status, 用预测数据输出消费者ID列, 在运行网络之前, 增加一个 过滤行 节点来过滤那些非拖欠概率大于0.98的消费者。然后你可以输出这个列表到文本文件中进行发送。
16. 从探索器窗口下的 数据操纵/行操纵 文件夹下, 双击 过滤行 组件, 并把它添加到工作簿中。  
把新节点放到 预测 节点的右侧, 然后连接 预测 节点。
17. 双击节点打开它的属性对话框。
18. 在 属性 页面里, 限定条件 文本框中, 键入PREDICT.prob>0.98。点击 确定 关闭对话框。
19. 在探索器窗口下的 数据输出/写文件 文件夹下, 双击 写文本文件 组件, 添加 写文本文件 节点。把新节点放到过 滤行 节点的右侧, 连接 过滤行 节点。
20. 双击节点打开 属性 对话框。
21. 在 属性 页面, 点击 浏览 键入examples/MortgageDefaultExample文件夹。
22. 在 文件名 文本框中, 键入CustomerNoDefault.txt, 点击打开按钮。在 分隔符 列表框中选择 tab delimited。点击 确定。
23. 运行网络。

你已经创建了一个tab-delimited形式的文本文件包含哪个消费者最有可能拖欠贷款的信息。你现在可以发送这个文件到银行贷款部门用它进行风险评估。

## 2.7 S-PLUS 库的探索

你可以用S-PLUS的图表功能来增加你的探索和预测的能力, 你也可以用 **S-PLUS脚本** 节点创建复杂模型。

S-PLUS软件中的S语言是S+Miner系统的基础且不需要明确安装。在探索器窗口中有S-PLUS页面。这本书里我们不介绍S语言的细节。参看即将出版或已经出版的S-PLUS书目查看更多详细信息。

**注：**S+Miner不仅包括S-PLUS库还包括S语言的运行环境，你无须为S+Miner安装额外的S-PLUS版本。如果你想更广泛的使用S-PLUS或是S语言，可以考虑使用S-PLUS企业版。

S-PLUS企业版提供了S+Miner所没有的性质，如S-PLUS GUI，S-PLUS集成的开发环境，S-PLUS控制台应用，以及嵌入应用和其他接口(包括OLE，DDE，和COM)。

## 2.7.1 使用 S-PLUS 图

在探索器窗口下点击**S-PLUS**页显示**S-PLUS**的节点。**S-PLUS**提供了很多做探索图的工具，这些图能为数据提供更多的信息。例如，看一个因变量是条件变量status的直方图的例子。

1. 打开提供的例子工作簿，**examples/MortgageDefault.Model.imw**。
2. 运行流程网络。
3. 从S-PLUS页中，打开 **数据探索/一维图形-连续性** 文件夹，拖拽 **直方图** 组件，把它放到 **预测：Logistic回归** 节点旁。

完整的工作簿如图2.27 (**MortgageDefault.Model-splus**)

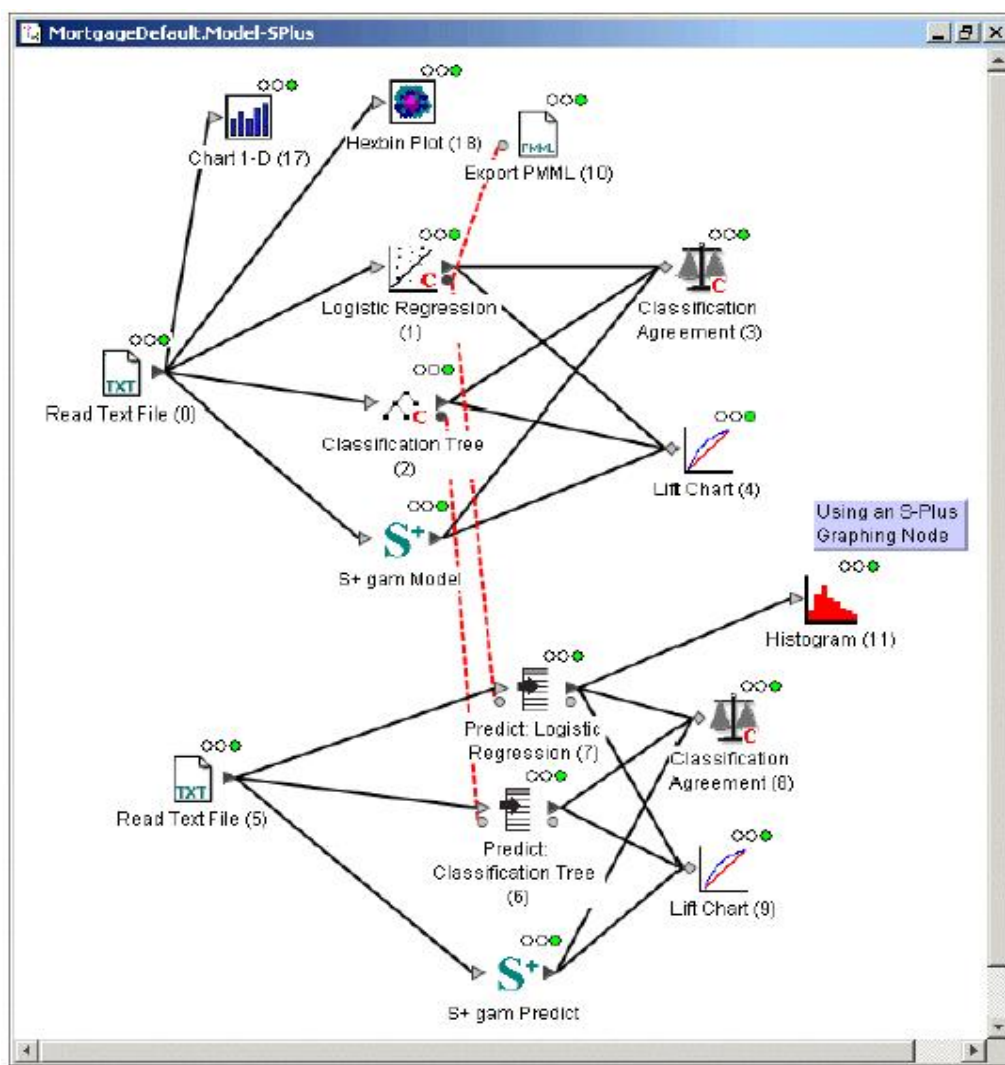


图2.27:用S-PLUS脚本节点对拖欠贷款数据作图并对照的模型，包含S-PLUS模型的

完整工作簿(MortgageDefault.Model-SPlus.imw)

4. 连接 **直方图** 节点到 **预测Logistic回归** 节点，然后双击 **直方图** 节点打开它的属性对话框。完整的对话框如图2.28。

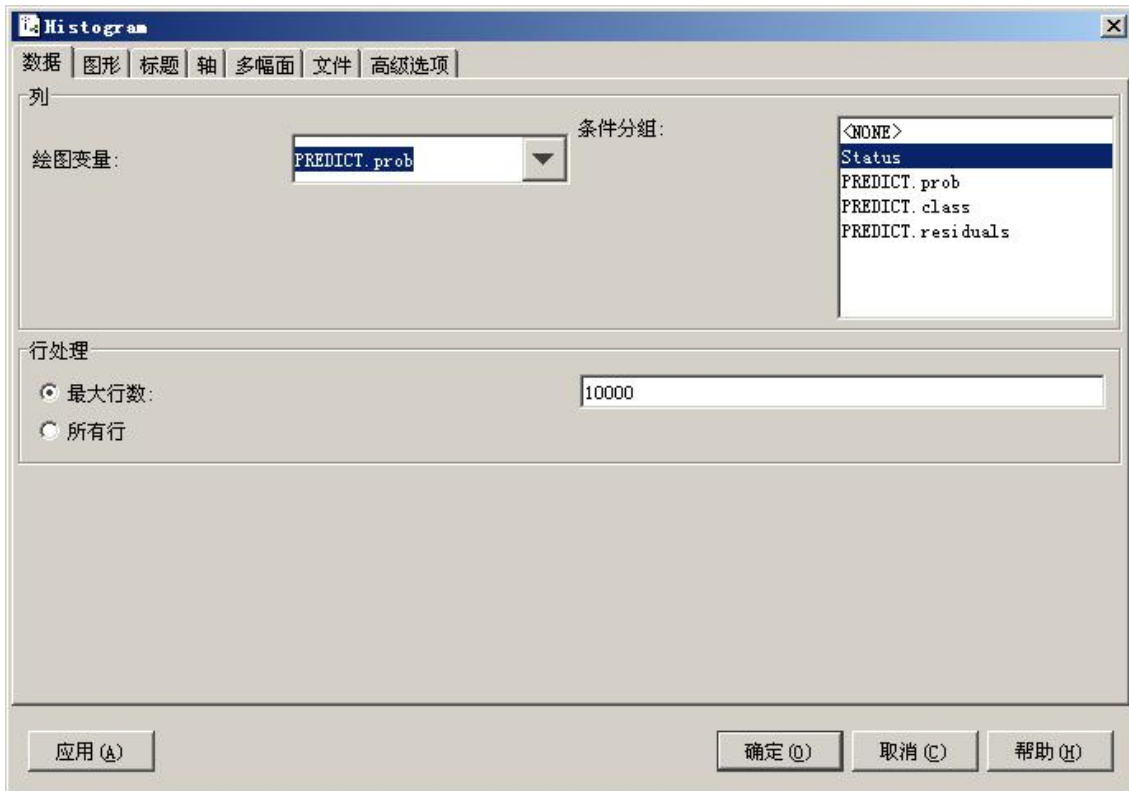


图2.28: 完整的直方图节点的属性框。

5. 在 **列** 选项组的 **绘图变量** 文本框中，选择**PREDICT\_prob**。在 **条件分组** 中选择变量status。
6. 点击 **应用** 来运行节点，并展示Percent of Total vs PREDICT\_prob的直方图。如图（2.29）

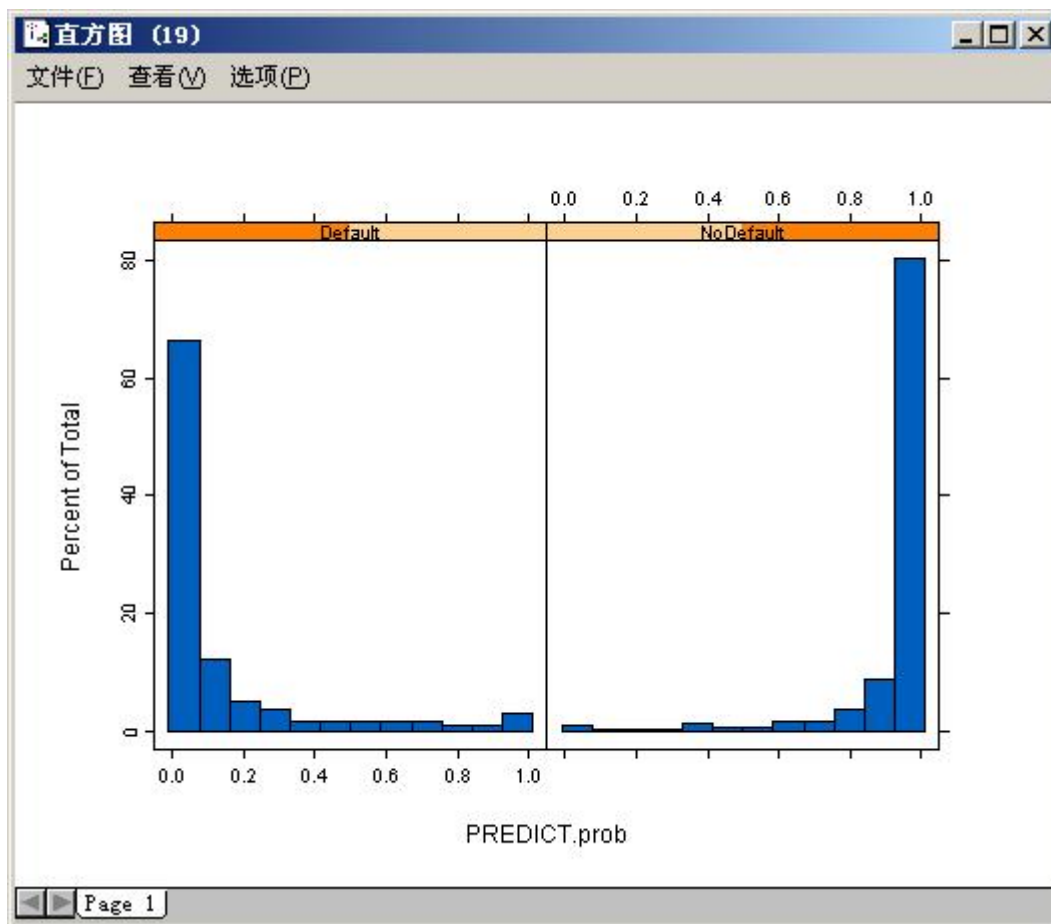


图2.29: logistic回归模型中Percent of Total 对PREDICT.prob的直方图

这个图显示：这个例子为预测概率作出了很好的工作，但是Logistic回归模型不能用在预测变量和因变量间是非线性回归的关系中。对于一个非线性模型用哪种预测更好呢？一个备选模型应该是GAM模型(广义添加模型)。一个二元的GAM模型与Logistic回归模型相似，但是GAM模型的预测变量和因变量不一定是线性关系，它允许任意的平滑函数关系。对于更多的细节参看Hastie和Tibshirani(1990)，或是参看S-PLUS用户指南第一章。你可以用S-PLUS脚本节点创建模型，具体在下部分讲解。

7. 如果你不想修改初始的工作簿，使用不同的文件名保存这个工作簿。

## 2.7.2 用 S-PLUS 脚本节点进行建模和预测

比较S-PLUS的GAM模型和前面的两个模型作：分类决策树和Logistic回归。完整的工作簿如图2.27。工作簿文件是 **examples/MortgageDefaultExample/MortgageDefault.Model-Splus.Imw**。你可以在S+Miner用户手册里找到有关创建和使用S-PLUS脚本节点的相关信息。

### 用S-PLUS脚本节点创建GAM模型

1. 使用浏览方式，打开**examples/MortgageDefaultExample**文件夹下的**MortgageDefault.Model.imw**。
2. 从 **工具** 文件夹中添加 **S-PLUS脚本** 节点到工作簿里。连接这个节点到 **读文本文件(0)** 节点。
3. 双击 **S-PLUS脚本** 节点打开属性对话框。

4. 在 **脚本** 页面中，点击 **加载**。选择文件

**examples/MortgageDefaultExample/MortgageDefault.gamModel.ssc。**

点击 **打开**。

你可以在S-PLUS代码或是属性对话框中的 **选项** 页设置属性。

在这个例子中用对话框来设定选项。这是S-PLUS脚本节点默认的状态，如选项页面中，在 **输入输出信息设定** 选项组，选择 **此处指定**。对于GAM模型，你希望以此使用所有的数据；在 **行处理** 组中，指定 **单个数据块**（这是默认设定的）。

5. 在的对话框里的 **输出列** 选项组中选择 **预先设定** 和 **新列**。

接下来，通过输出新变量来完成 **新列**。完整的 **选项** 页如图2.30。



图2.30: S-PLUS 脚本GAM模型节点的完整的选项页面。

6. 在第一行，**名称** 列里，键入**Status**。从 **类型** 列表框里选择 **分类型** 并从 **角色** 列表框中选择 **因变量**

7. 在第二行，**名称** 列中，键入**PREDICT.prob**。从 **类型** 列表框里选择 **连续型** 从 **角色** 列表框里选择**预测**。

8. 在第三行，**名称** 列中，键入**PREDICT.class**。从 **类型** 列表框里选择 **分类型** 并在 **角色** 列表框里选择 **信息**。

9. 点击 **确定** 关闭 **属性** 对话框。

10. 给节点重命名，右击节点选择 **更名**，或是左击节点名然后键入**S+gam模型**。

11. 连接 **S+gam模型** 节点到 **分类吻合度** 和 **提升图** 节点。

12. 运行网络。



GAM模型输出了几个图形。通过节点查看器查看默认属性输出的这些图。你可以像设定节点属性一样设定查看器的选项。接下来比较新模型和另一个模型：

13. 打开训练数据的 **分类吻合度** 节点查看器

GAM模型整个的吻合度在Logistic回归和分类决策树的吻合度之间。

## 用S-PLUS脚本节点预测GAM模型

你可以使用另一个S-PLUS节点从GAM模型来创建一个预测节点。在**S+gam Model**节点的代码中你可以写入输出模型的信息。这些信息可以通过另一个节点访问并用来预测。

14. 添加 **S-PLUS脚本** 节点到另一个预测节点下，并连接它到 **读文本文件** 节点来测试数据。

15. 打开属性对话框到脚本页，键入

**examples/MortgageDefaultExample/MortgageDefault.gamPredict.ssc。**

为了预测，特别是大数据集，可以使用 **多个数据块** 选项。这里在此通过对话框指定选项，而不是在S-PLUS脚本中设定。完全整的对话框如图2.31。

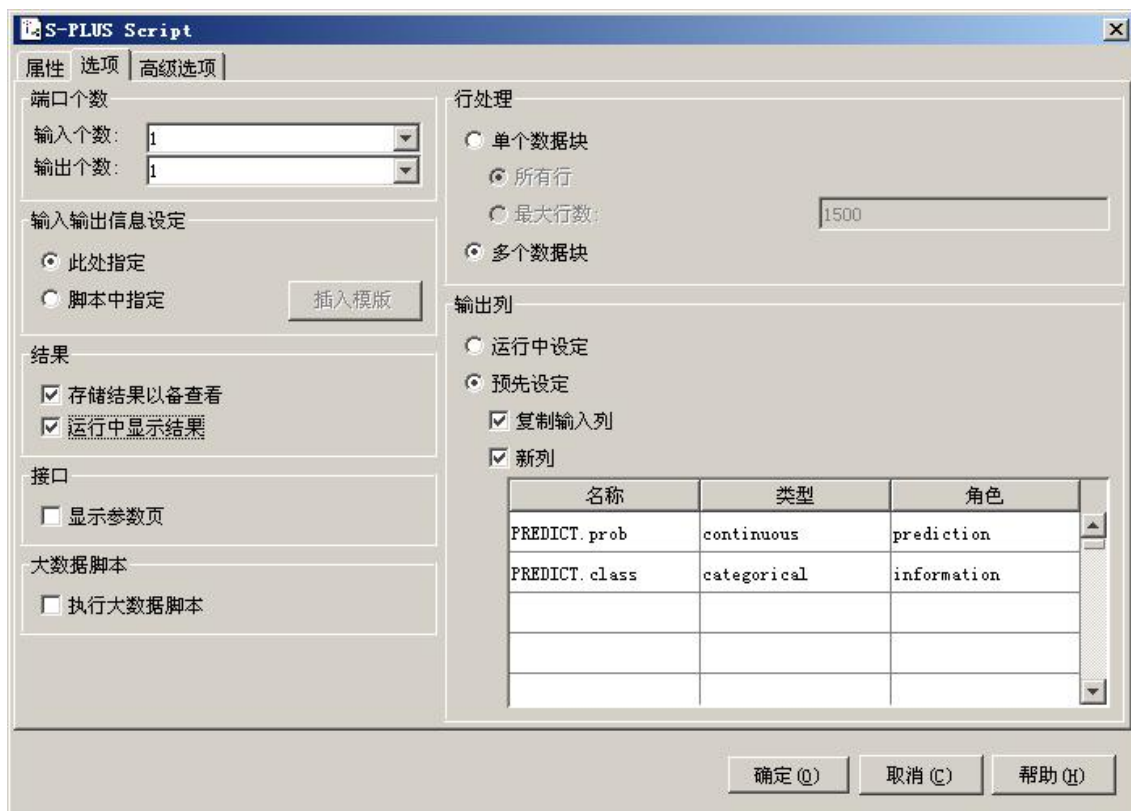


图2.31: S-PLUS脚本GAM预测节点完整的选项属性页。

16. 点击打开 **选项** 页面。在行处理组中，选择 **多个数据块**。

17. 在 **输出列** 选择 **预先设定**，选择 **复制输入列** 和 **新列**。

18. 完成新列表的第一行，键入**PREDICT.prob**，**continous**，和**prediction**。

19. 完成表的第二行，键入**PREDICT.class**，**categorical**，和**information**。点击 **确定**。

20. 重新命名节点为**S+gam预测**。

21. 连接 **S+gam预测** 节点到 **分类吻合度** 和 **提升图** 节点。

22. 运行挖掘网络。

复制这个预测节点到打分工作簿中为数据打分，就像你在logistic 回归模型中所作的一样。完整的工作簿是examples/MortgageDefaultExample/MortgageDefault.Score-Splus.imw。

使用 **S+gam预测** 节点添加一个打分网络到已有的打分工作簿中。完整的工作簿如图2.32。

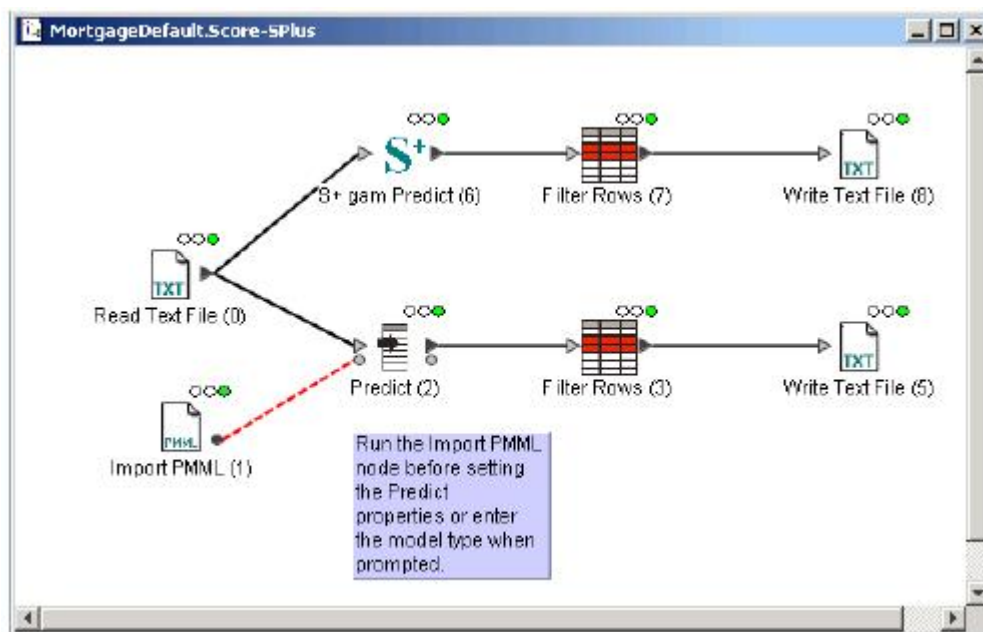


图2.32：用logistic回归和S-PLUS GAM模型来评估抵押贷款得分数据的完整的工作簿  
(MortgageDefault.Score-Splus.imw)

23. 在模型工作簿中选择 **S+gam预测** 节点并按CTRL-C来复制节点。
24. 打开打分工作簿 **examples/MortgageDefaultExample/MortgageDefault.Score.imw**，并按CTRL-V 粘贴预测节点到工作簿中。
25. 拖拽节点到已存在的预测节点，然后连接它到 **读文本文件** 节点。
26. 添加 **过滤行** 节点和 **写文本文件** 节点如图2.32，并连接节点。
27. 双击 **过滤行** 节点打开它的属性框。在 **限定条件** 文本框中键入PREDICT.prob>0.98，并点击 **确定**。
28. 双击 **写文本文件** 节点打开属性页面。在 **文件名** 文本框中，键入 **CustomerNoDefaultgam.txt**
29. 运行挖掘网络。

你现在有一列贷款项表单，上面有拖欠贷款概率的最低标准。

在这个例子中，你创建的所有模型对于预测拖欠贷款概率作出了很好的工作。使用S-PLUS模块，可以用新方法探索数据并为数据创建更复杂的非线性模型。

## 2.8 概要

在这个例子中，你：

- ¥ 创建了一个模型来预测顾客拖欠家庭贷款的概率；
- ¥ 使用所有可用的消费者的数据，通过合并数据文件和分割数据，来训练和测试模型；
- ¥ 通过使用Logistic回归和分类决策树模型来预测消费者贷款状态，创建一个精度在94%以上的模型（在建模中，仅仅使用了预测因子；为了改善模型，你还可以使用预测因子的交叉项。）；
- ¥ 比较了标准S+Miner模型和一个由S-PLUS提供的附加的GAM模型。这个模型用法与其他模型一样（再者，如果你使用更多的时间来探索数据和交互项，你可能开发出更好的模型。）；
- ¥ 最后，你创建了一个消费者的列表，上面记录着在可接受的范围内他们拖欠贷款的风险。最后一列是预测概率，你也可以考虑不同的风险情况。并利用这些信息在进行哪项贷款这个问题上做出最后的决定。

## 2.9 参考文献

Hastie, T. and Tibshirani, R. (1990) *Generalized Additive Models*. Chapman and Hall, London.  
*S-PLUS Guide to Statistics, Volume 1*, Insightful Corp., Seattle, WA.