# 基于 Windows 的
# MineSet 3.0 企业版教程 ®

文档编号 007-4006-001CHS

出品人

作者：Helen Vanderberg， Pam Sogard， Sandra Motroni
插图：Dany Galgani
制作：Linda Rae Sande
设计：Barry Becker， Amit Bleiweiss， Jeff Brainerd， Cliff Brunk，Eben Haber，
Ara Jerahian，Eser Kandogan，Andy Kar，Ed Karrels，Alex Kozlov，Brian
Lovrin， Alan Norton， Peter Rathman， Gerald Rousselle，
Mario Schkolnick，Dan Sommerfield，Peter Welch，and Brett Zane-Ulman.



基于 Windows 的 MineSet 3.0 企业版教程 ®
文档编号 007-4006-001CHS

# 目录

# 关于本教程

*《针对 Windows 的 MineSet 3.0 企业版教程》*介绍了 Windows 环境中的 MineSet。MineSet 是数据挖掘和可视化处理工具的集成软件包，它提供了有关数据挖掘概念和过程的快速概览。本教程描述了几个基本任务，可以帮助您立即使用 MineSet。一旦您熟悉了界面，可以参考 *《针对 MineSet Windows 3.0 企业版用户指南》*以获得对其他 MineSet 特性的全面描述。用户指南将作为 MineSet 的一部分联机提供。有关详细信息可访问 http://mineset.sgi.com 。

## 本教程的适用对象

本教程适用于最终用户。不需要编程方面的经验，也不需要任何统计方面的预备知识（尽管这些知识很有帮助）。但需要一点 Windows 的基本知识。

## 本教程的使用前提

为了使用该教程，应该在您的系统中安装 MineSet，或者您可以访问这样的系统，其中的示例依赖于它。MineSet 的安装说明可以在 *《针对 MineSet Windows 3.0 企业版用户指南》*和 MineSet 的 网页 http://mineset.sgi.co 上获得，另外在 MineSet 的网页上还可以下载 MineSet 的测试版。

您不必为本教程访问数据库。在 MineSet 发行的版本中包括所有需要的数据。

## 本教程的结构

第 1 章，“数据挖掘基本原理”，介绍数据挖掘的概念，并解释如何用数据挖掘来解决问题。常用的数据挖掘任务与 MineSet 工具相互联系；每个工具的详细信息在随后章节中阐述。

第 2 章，“数据挖掘过程”，描述与数据挖掘过程有关的任务。提供了一个使用 MineSet 进行数据挖掘的研究案例。

第 3 章，“客户波动教程”，为使用 MineSet 进行的数据挖掘过程提供了一个详细教程。它从初始屏幕开始，逐屏讲解如何通过 MineSet 工具使用*客户波动数据集*（随 MineSet 发行版提供的一个数据集）。

第 4 章，“进一步的探索”，继续讲述 MineSet 数据挖掘中更为复杂的变化。

附录 A，“在 MineSet 可视化工具中漫游”，讲述了移动和操作可视化工具窗口的各种方法。

## 印刷约定

本教程使用以下几种字体约定：

| | |
|---|---|
| *斜体* | 斜体用于命令、文件名、变量和用户界面按钮名称。 |
| `Courier` 或宋体 | Courier 或宋体用于系统输出的示例和文件内容。 |
| **`Courier`粗体或宋体粗体** | Courier 粗体或宋体粗体用于逐字键入的命令和其他文本。 |

*第 1 章*

# 数据挖掘基本原理

本章综述了数据挖掘的方法，模型的建立和评测，以及与这些主题相关的 MineSet 功能：



## 关于数据挖掘

数据挖掘的目的是为了发现数据中的模式，从而将这种认识应用到解决问题之中。结合了强大的可视化工具的分析数据挖掘将为知识发现开辟新的道路。数据挖掘系统可以自动查找并显示新模式，而这些模式能够带来新的认识。具体的例子如：确定属性之间的关系，区分不同特征数据的子集，以及从历史数据中推测将来事件发生的概率。

在普通数据库查询或联机分析处理 （OLAP）中，您必须直接指定数据元素之间的所有关系。而数据挖掘却可以发现对您来说可能是未知或未见的关系。

当进行一个商业或科学过程 （例如：从用户帐单、药物测试或销售点交易中获取数据）时，开始通常要检索被分析或被挖掘的数据。检索的数据量可能会很大以至于无法用数据挖掘以外的手段进行分析。一旦对这种数据进行了适当的转换，就可以将其存储在数据仓库中。有关详细信息，请参阅 第 10 页的 “准备数据”。

## 本教程中使用的术语

可以将 MineSet 使用的数据文件看成是巨大的表。行是单独的记录，列则是每个记录的*属性*。分类任务中的*标签*是为分类选择的特定属性值。例如：在本教程通篇使用的范例文件 （*客户波动*）中，任务将记录分类为已离开公司的用户 （客户波动的）和没有离开公司的用户。标签属性 （列）为 “客户波动的”，可能的标签值为 “是”和 “否”。

*离散*标签是一个只能拥有有限值的标签，如性别、薪水范围 （例如：少于 $40,000 ； $40,000 到 $80,00 ； 超过 $80,000），以及年龄范围 （例如：不到 21， 21 到 35， 36 到 50， 50 以上）。连续标签可以拥有一个大范围内的所有值，例如：年薪、年销售额、以及每加仑英里数。

## 数据挖掘方法

数据挖掘将假设检验与数据驱动式挖掘结合起来。在假设检验中，调查者将检验针对数据体的某个设想，以确认或否认其有效性。在某些情况下，数据本身就可以促进挖掘过程。在挖掘中，调查者从数据中得出结论，并允许数据本身暗含着结论。通常，数据挖掘问题可以通过两种方法的结合来解决。例如：结论可以产生新的假设，而这些假设又被继续进行检验、确认或否认。数据挖掘是统计学与机器学习的交叉点。

MineSet 工具包可以对数据进行分析、挖掘和图形化显示，从而使您可以看见、研究并最终理解这些数据。您可以用不同方式组织和检验数据。挖掘工具会自动查找模式，并建立能够使用可视化工具查看的模型。当您将可视化工具直接应用于数据上时，就会获得对数据更深刻、更根本的认识，通常是发现隐藏在更深层的模式以及重要的趋势走向。

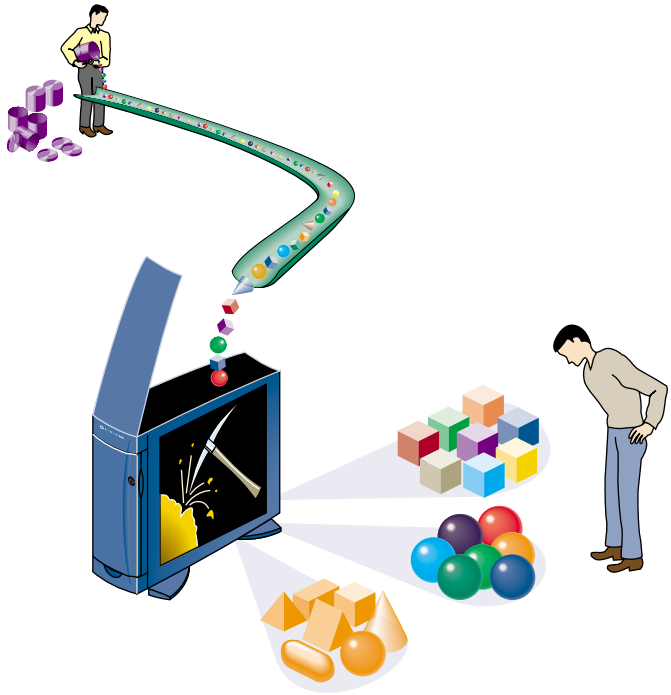


图 1-1　　分析式数据挖掘发现数据中的模式

MineSet 中典型的分析式数据挖掘操作的结果将同时包括一个描述数据的模型和对该模型的可视化处理。这个可视化处理是三维界面的，它使您可以交互地处理对象，并进行动画演示。可视化处理将使您理解模型，并可对复杂的数据模式进行综述以便在决策时获得极其宝贵的认识。 MineSet 是一个集成系统，其中的分析式运算法则可以产生可视化处理，您可以为进一步的挖掘选择可视化处理元素。

## 分析式数据挖掘运算法则

分析式数据挖掘运算法则可以从数据中自动建立模型。通常使用两个系列的建模运算法则：监督和非监督。预测建模任务被称为*监督*任务，其目的是基于其他列的值来预测某列的值。这类任务类似于一个老师的督导，他告诉你问题的正确答案来指导你。

描述建模的目标是发现数据中的隐含模式和段落。它们是*非监督*任务。既没有正确答案的任何概念，也没有明确公认的性能量度。非监督任务展示行为类似的模式和段落，从整体的角度提供对数据的认识。

## 监督建模

在监督建模中，您希望预测的是一个被称为 “标签” 的特殊属性。通过对标签和其他属性之间关系的编码，模型可以对新的、未标注的数据进行预测。另外，通过观察模型本身，您可以认识到标签与其他属性之间的关系。例如：如果一个用户已经离开了您的公司 （通常被称为损失或客户波动），您就可以建立一个模型，既可以预测哪一个用户可能波动，又可以帮助您了解导致这种行为的原因和模式。

分类和回归是两种最普通的监督建模任务。如果标签是离散的 （即：包含一个固定集合的值），则该任务叫做分类；如果标签是连续的 （即：可以取一个连续范围中的任何值 - 例如：收入或股票价格），则该任务叫做回归。

### 分类

分类任务将离散的标签值分配给未标注的记录 （关于这些术语的定义，请参阅第 2 页的 “本教程中使用的术语”）。在这种情况下，记录将被划分到预定义的组中。例如：一个简单的分组可能把用户的帐单记录归为以下指定的两类：在 60 天内支付帐单的人和超过 60 天支付帐单的人。其他的数据分类任务可能会按性别或收入来对用户进行分组。分类器还会预测特定记录采取指定标签值的概率。例如：假如给定了用户记录的其他属性值， MineSet 可以计算某个用户在 60 天内支付帐单的概率。

分类器是一种模型，假如给定了数据集的其他属性，它就可以预测另外一个属性。MineSet 可以从训练集 （数据集的一个子集）中自动导入 （建立）一个分类器。当导入分类器时， MineSet 还会生成一个模型的可视化处理，它可以帮助您了解分类器的工作方式，从而提供宝贵的认识。一旦生成了一个分类器，它就可以被用来对未标注记录 （即：对那些遗失了标签属性的记录）进行分类或预测。这个概念将在第 3 章中作进一步讲解。

MineSet 有四种分类模型导入工具：决策树、选项树、证据 （简单 Bayes），以及决策表分类器。可以使用三维可视化工具查看每个模型。

### 回归

除了标签不是离散的以外，回归是类似于分类的一种监督建模任务。例如：预测薪水或股票的价格就是一种回归，而预测薪水是否在给定的范围内或股票价格将会上升还是下降是一个分类任务。

### 评测模型的准确性

预测模型很少是最优的，因此评测它们的准确性是数据挖掘过程中的一个重要部分。用来评测准确性的工具取决于模型的类型。通常根据分类器的误差率来评测它们。这种评测中最常用的是错误分类，或被错误分类记录的比例。当评测模型的准确性时，用没有用于建立模型的数据进行测试是非常重要的。 MineSet 为评测误差提供了大量方法。关于详细信息，请参阅第 4 章，“进一步的探索”。

## 非监督建模

非监督建模的目的是发现行为相似的数据段及其规则 （聚类）。非监督建模是一种描述性任务，而不是一种预测性任务。模型无法直接用于预测，因此无需保存部分数据以作为验证分类器的测试集。 MineSet 为非监督建模提供了两种方法 - 关联和聚类。

### 关联

为了生成关联，任务将要确定数据属性 A 和 B 之间的隐含规则，例如 A 隐含 B。关联通常用于寻找具有密切关系的分组，这些分组将揭示哪些物品经常与其他物品一起被购买。典型的密切关系分组是市场购物篮分析，用以预测特定物品被同时购买的频率。例如：购买婴儿商品隐含着该用户购买低焦油含量香烟的概率要比购买普通香烟的概率高，这个发现可能会帮助商店有差别地安排货架。

### 聚类

聚类运算法则将数据分割为具有相似特征的记录组或聚类组。例如：健康保险公司可能会发现以下特性可以定义一个段：20 到 45 岁、技术员、孩子少于两个、科幻电视迷、可支配收入为每年 10,000 到 20,000 美元。

然后，在新的科幻电视节目中插播电视广告，非常适合这些人的有关健康保险内容的就会更加有效和有针对性。

## 可视化数据挖掘

分析式数据挖掘运算法则可以由数据可视化技术进行补充，这些技术利用了人类大脑惊人的模式识别能力。以下是可用的 MineSet 可视化工具：

- 地图可视化工具 - 数据被显示在一张图上，通常是一张地图
- 散点可视化工具 - 数据点的显示是一维、二维或三维的。附加属性可以被映射为颜色、大小和形状。最后，还有两个附加属性可以被映射到滑动条，允许动画和闪现，共八维。 MineSet 中的列重要性操作可以帮助您识别指定任务要映射的重要属性。
- 平伸可视化工具 - 类似于散点可视化工具，其区别在于数据密度用颜色的通透性显示，看起来象一片模糊的半透明云彩。结果与单独处理每个数据点的效果近似。
- 树可视化工具 - 数据被映射到节点上，以便观察数据的分层结构。

## 用于数据挖掘任务的 MineSet 工具

如果您有关于分类、回归和聚类的数据挖掘问题，可以使用以下 MineSet 工具：

- 决策树导入和分类工具 - 导入一个产生决策树可视化处理的分类器。
- 选项树导入和分类工具 - 导入一个类似于决策树导入和分类工具的分类器。但是，它还建立备用选项并在分类过程中均衡它们，通常会提高准确性。
- 证据导入和分类工具 - 创建其自己的分类器并产生一个可视化处理，以显示基于所提供数据的证据。
- 决策表导入和分类工具 - 创建一个分层的可视化处理，显示每一层的各维数对组合。在保留前后位置的同时，您可以迅速概化上寻和细化下寻 （察看下一层细节，概化上寻相反）。
- 聚类运算法则 - 根据特征的相似性对数据进行分组，然后将其显示为一系列的棒状图和直方图，类似于统计可视化工具。默认情况下，聚类运算法则使用聚类可视化工具显示结果，但是也可以使用其他的可视工具。
- 回归树 - 导入一个预测真实值的回归工具，注意：结果是带有不同等级的值，而不是预先确定的区间。
- 列重要性 - 在区分两个标签值时，确定指定列的重要性。用于观察变量改变带来的各种影响，或提出映射到散点和平伸可视化工具坐标轴的列。

MineSet 中还包含补充工具以辅助知识发现过程

- 统计可视化工具 - 数据以棒状图和直方图的形式显示，每列一个图。连续列显示为棒状图，离散列显示为直方图。
- 直方图可视化工具 - 数据以直方图形式显示，连续列被分组 （被分解为不同的范围）。
- 记录查看器 - 原始数据显示在电子表格中。

下一章，第 2 章，将描述典型的数据挖掘过程以及如何使用这些工具。

# 数据挖掘过程

本章介绍有关知识发现过程的特定任务。知识发现过程是一个反复迭代的过程 （如图 2-1 所示），您一旦发现了新的模式并提高了对数据的认识后，就又回到早期的阶段。

该过程的通常步骤是：

1. 识别数据源 - 请参阅 第 9 页的 “识别数据”。
2. 准备数据 - 请参阅 第 10 页的 “准备数据”。
3. 建立模型 - 请参阅 第 12 页的 “建立模型”。
4. 评测模型 - 请参阅 第 12 页的 “评测模型”。
5. 使用模型 - 请参阅 第 12 页的 “使用模型”。

## 识别数据

识别数据任务将从决定需要哪些数据来解决问题开始。例如：关于用户行为的预测通常是根据具体问题来重新定义的必要目标。在定义问题时，调查者必须识别用于解决该问题的数据并探索其他可能的数据源。

数据可能位于一个较为困难的位置，或是以一种模糊的形式出现。有时存在多个互不兼容的初始数据库。而且，如果数据很少或不完整，就会需要更多的数据。新数据的形式将根据现有数据的形式被收集起来。MineSet 支持多种商业数据库（Oracle、Informix、SQL）的本地接口、 ODBC 接口，并且可以读取不同文件格式的数据 （制表符分隔的单一结构文件、 MineSet 二进制文件、 Excel、 SPSS、 Mutable 等等。）


图 2-1　　数据挖掘过程

## 准备数据

在加载到 MineSet 之前，数据可能需要修改 （该步骤通常叫做清洗。）特别指出的是，以下问题是常见的：

- 数据格式可能与 MineSet 的表示法不兼容 （例如：来自老式大型计算机的二进制、编码、或 EBCDIC 字符串）。
- 数据可能有拼写错误或本身是错的，也可能不完整或取值错误。
- 字段描述可能不清楚或混乱，或者可能对于不同的源含义不同。例如：订购日期可能指的是订单被发送、盖戳、收到或键入的日期。
- 数据可能会过期；例如：用户可能会搬家，变换主人或改变消费模式。

即使是清楚的数据也需要在适用于挖掘和可视化处理之前进行转换。

## 转换数据

转换可以极大地提高模型的性能。例如：如果您正在分析电话公司的数据，可能会发现：比较单独给定的任何一个元素，长途电话的费率 （销售额除以通话使用的总分钟数）是一个更好的用户行为预测器。数据转换是开发一个合理模型的核心部分，随着分析进程不断地推进，您甚至可能会返回并以另外的方式转换数据。您可以通过下列方法来转换数据：

- 添加列，通常是对已有的数据应用一个数学公式来创建一个新字段。
- 删除无关、多余或包含明显无用预测器的列。
- 筛选可视化处理。例如：您可能希望只看到最强的规则或最有利可图的用户段。
- 对数据进行分组 - 将一个连续范围内的数据分解为几个离散的段 （例如：[1-10]，[11-20] 等）。
- 组合数据 - 将记录集合在一起，然后查找总和、最大值、最小值和平均值。
- 对数据进行采样以获得数据的一个随机子集 （通过百分比或计数）。
- 应用先前创建的分类器，对带有类标签的新记录进行标注，或估计给定标签值的概率。

在 MineSet 中，大部分的转换是使用 “工具管理器”中的 “数据转换”面板来完成的。

## 建立模型

知识发现过程的核心是模型的建立，该任务由数据挖掘运算法则自动完成。这部分将在第 3 章中阐述。

## 评测模型

评测模型的准确性可以进一步提炼您对该模型及其用法的认识，您可以通过筛选数据、删除列、创建新列等方法来改进模型。

MineSet 采用四种模型评测方法：误差估计、混淆矩阵、上升曲线和 ROI （投资回报）曲线。

## 使用模型

可以通过将模型应用到新数据上来使用它。新数据可能会产生新问题，从而需要进一步改进。

在 第 3 章的电信范例中，可以通过创建模型来确定哪些用户可能离开他们的电话公司。然后，利用模型来评测用户记录，以识别最可能离开的特定用户。可以给这些用户一定的刺激使他们留下。

## 将过程应用到特定数据库上

下面两章将逐步向您介绍关于客户波动数据集的知识发现过程 - 一个准备好的电信用户数据集。当您完成该范例时，请思考一下这里提出的过程以及您的操作是如何前进和返回的 （如 图 2-1 所示）。

# 客户波动教程

本章向您逐步介绍使用 MineSet 提供的*客户波动*数据集的一个知识发现过程的范例。假设您使用的系统中已经安装了 MineSet 和所有的样本文件。每一步都将进行详细讲解，除非特别注明，否则每一步都是建立在以前步骤之上的。这些步骤如下：



## 关于原始数据

*客户波动*数据集用于管理电信用户 - 定期使用电话的人。用户可以选择为他们提供电话服务的公司，当这些用户改变了原先的选择，就被称为 “客户波动”，这会导致以前服务公司的收入损失。电信公司可能拥有电话记录数据库，其中包含电话信息（来源、目标、日期、持续时间）、帐单数据库、用户数据库和用户服务数据库。用户的相关信息将出现在所有这些数据库中。当这些信息被结合在一起时，就会产生一套用户特征。MineSet 提供的客户波动数据集就是这样一个集合；识别数据并将用户特征创建到记录中的步骤已经完成了。该数据集将用于本章的其余部分，其中为每个用户开创了一个记录。

## 运行 MineSet

1. 选择 “开始” > “程序” > “MineSet 3.0 企业版” > “MineSet”，或双击桌面上的 MineSet 图标以运行 MineSet。
2. 如果 “登录到服务器” 对话框 ( 图 3-1) 没有以默认的方式出现，请选择 “文件” > “连接到服务器”。如果您希望将当前系统同时作为用户机和服务器来使用，可以在产生的对话框中，单击 “将这台机器作为当前用户”。如果您希望使用另外一个系统作为服务器，请键入服务器名称、登录名和口令 （如果有）。


**图 3-1** 工具管理登录窗口

3. 单击*确定*。如果您曾经登录到 MineSet，系统就会显示一个恢复的界面。为使用本教程，您必须打开一个新文件 （第 4 步）。
4. 在 “工具管理器” 窗口中，选择 “文件” > “打开新的数据文件”。如果生成的对话框没有显示 data 目录，可以转到 MineSet 的安装目录，默认情况下为 MineSet 3.0 > data。
5. 选择 *churn.schema*，右边 “预览列” 面板中将出现如 图 3-2 所示的一系列条目。
6. 单击*打开*。

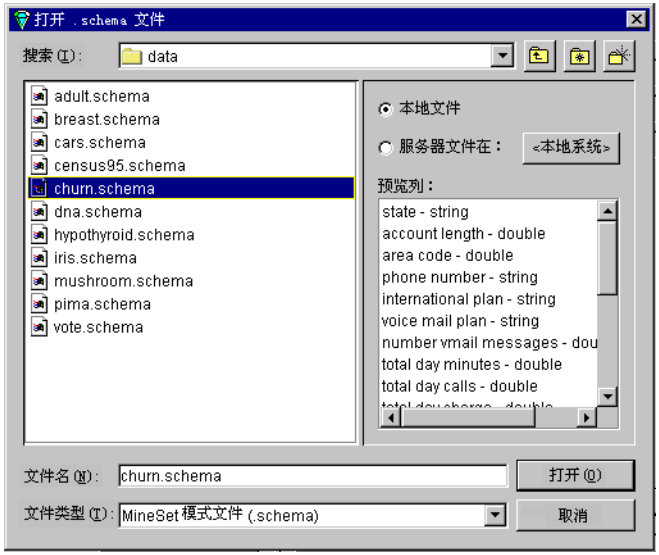



**图 3-2** “打开新的数据文件”窗口

这样您就可以访问电信用户数据集了。下一次当您运行 MineSet 时，系统将自动把您带到该位置 （或是最后一次 MineSet 运行时您所在的任何地方），并且您所做的一切选项选择都将被保存起来。

## 查看记录

在 “MineSet 工具管理器”中，您可以以电子表格的形式查看记录，步骤如下：

1. 在 “工具管理器”的 “数据目标”面板的上面一行选项卡中，单击 “可视化 工具”选项
   卡；然后在下面一行选项卡中，单击 *记录* 以进入 “记录查看器”选项卡。

   用于本章其余部分的 *客户波动* 数据集中为每个用户建立了一个记录。“工具管理器”左侧的 “数据转换”面板上列出了不同类型的列：州 （字符串），帐号时间长度（双精度）等等。如果列是数字，就被定义为双精度或浮点数；或者如果列是由字符组成的，则被定义为字符串。

2. 单击右下方的 *调用工具*。

数据显示为电子表格。列及其含义如 表 3-1 所示。

**表 3-1** “MineSet 记录查看器”显示的客户波动数据集中列的详细信息

| 列名称 | 值 |
|---|---|
| 州 | 用户居住的美国州名的两个字母缩写 |
| 帐号时间长度 | 一个数值，表示用户拥有长途通话服务的月数 |
| 地区代码 | 三位数字的电话公司代码 |
| 电话号码 | 3+4 位电话公司代码 |
| 国际方案 | 国际电话的特殊定价程序，以是 / 否值表示 |
| 声音邮件方案 | 对带有服务商提供的声音邮件的用户的特殊定价程序，以是 / 否值表示 |
| 声音邮件的数目 | 每天声音邮件的平均数 |
| 白天的分钟总数<br>晚上的分钟总数<br>夜间的分钟总数<br>国际的分钟总数 | 按白天、晚上、夜间或国际价格收取费用的平均通话分钟数 |
| 白天的呼叫总数<br>晚上的呼叫总数<br>夜晚的呼叫总数<br>国际的呼叫总数 | 在白天、晚上、夜间或国际间进行的平均呼叫数 |
| 白天的费用总数<br>晚上的费用总数<br>夜晚的费用总数<br>国际的费用总数 | 按白天、晚上、夜间或国际价格收取的平均通话费用 |
| 用户服务电话的数目 | 该用户在过去半年中打电话寻求服务商用户支持的呼叫次数。 |
| 客户波动的 | 在过去的半年中，该用户是否改变过长途载波，以是 / 否值表示 |

3. 关闭 “记录查看器”窗口。您应该再次看到 “工具管理器”窗口，仍然使用*客户波动*数据源。
4. 在 “工具管理器”窗口的 “数据目标”面板中，应该仍然显示 “可视化工具”选项卡；在选项卡的下面一行中单击 “统计”选项卡。
5. 单击*调用工具*。

   “统计可视化工具”显示画面中包含许多直方图和盒形图。直方图显示的是离散变量值的分布，盒形图显示的是连续变量的汇总统计。

   每个盒形图 （在 图 3-3 的右侧）显示出关于单列数据的统计，包括最小值、最大值、平均值 （红色）、中值、以及四等分中的两个值 （25% 和 75%）。这些值被标记为线，标准偏差 （红色）显示在 +/- 符号后面。

   平均值是将数据添加到一列中，然后再除以记录数得到的。中值是给定列中的数字以大小顺序排列时位于中间的数。标准偏差是列中数据离散性的量度。

   直方图中包括特定的离散值：州名称或是 / 否值。下滚以在画面中查找客户波动的直方图 （请参阅 图 3-3 左侧）。显示出 5000 名用户当中的 707 名已经离开了该服务商。在本教程中，客户波动列一直是很重要的。

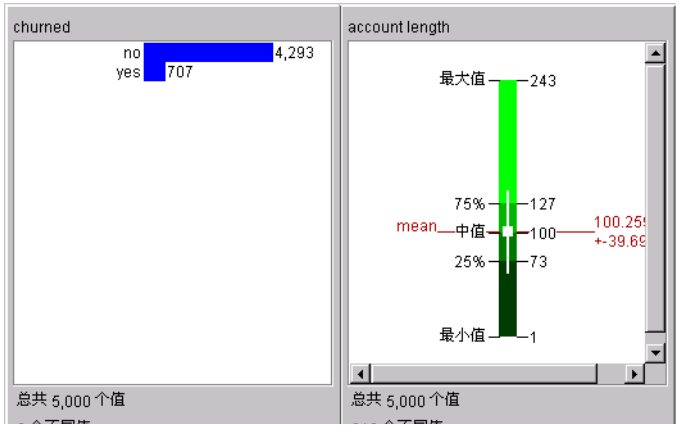



**图 3-3** “统计可视化工具”生成的典型直方图和盒形图

6. 关闭 “可视化工具”窗口，返回 “工具管理器”窗口。

## 建立一个证据分类器

现在您可以进行分析数据挖掘了。请确认 MineSet 已连接到适当的服务器上，并且数据源是 *churn.schema*。如果您在会话之间激活了 MineSet，历史文件将自动返回到您离开时的位置。

1. 在 “工具管理器”的 “数据转换”面板中，删除以下列。单击一列，然后按住 Ctrl 键以选中其余的列。然后单击*删除列*按钮。

   电话号码
   白天的分钟总数
   白天的呼叫总数
   晚上的分钟总数
   晚上的呼叫总数
   夜间的分钟总数
   夜间的呼叫总数
   国际的分钟总数
   国际的呼叫总数

   电话号码列没有预测值。分钟总数和呼叫总数列与收费总数列相关，只能增加极少额外信息，删除这些列可以缩短导入模型的处理时间，并且能够生成一个更加简洁的可视化处理。

2. 在 “工具管理器”窗口的 “数据目标”面板中选项卡的上面一行，单击 “挖掘工具”选项卡。

3. 在 “选项卡”的下面一行单击 “分类”选项卡，然后从以下下拉菜单中进行选择：

   *模式*：分类器和误差

   *导入工具*：证据

   *离散标签*：churned

   您即将导入一个证据分类器，以帮助勾画可能会客户波动用户的特征。默认情况下，分类器和错误对数据使用一种预留的方法，从三分之二的数据中导入分类器，并将剩余部分作为估计误差率的测试集。

4. 单击*继续*

   当导入工具读取数据时，系统可能会警告您：“州”列将被删除，因为它具有太多的唯一值。如果是这样，请转到 “工具管理器文件” > “首选项”菜单，将默认最大属性值更改为 100。

“工具管理器”底部的 “状态”窗口将显示导入过程的进程和汇总信息，包括误差估计 11.40% 正负 .78%。当导入步骤完成后，系统将自动调用 “证据可视化工具”，显示可视化模型 （图 3-4）。

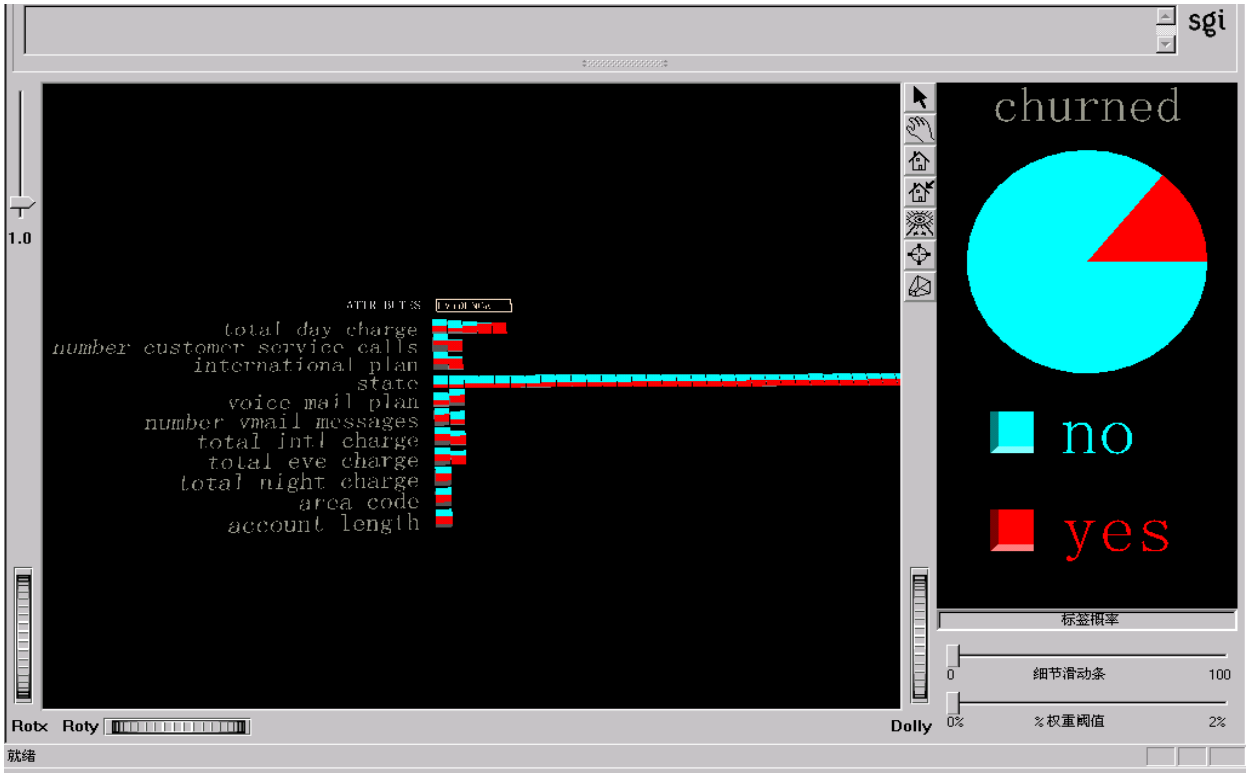



**图 3-4** “证据可视化工具”窗口

“证据可视化工具”是按照区分标签 “客户波动”的作用大小对列进行排序的，从顶部开始。要快速调整视图，可以使用伸缩滚轮。在本教程中，左边面板中的方块饼状图和右边面板中的圆饼图都被称为 “图”。

“标签概率”面板 （在 图 3-4 的右侧）将显示一个代表*先验概率*的圆饼图。先验概率是指具有 “是”（红色楔形）或 “否”（蓝色楔形）客户波动值的随机记录的概率，不考虑任何属性值。从数学角度来说，这个数是带有类标签的记录数除以记录总数得到的。

在 “证据”面板 （在 图 3-4）中，数据集中的列由图来表示，每个属性带有一个值或一个取值范围。将光标模式从 “抓取”（手形）改变为 “选取”（箭头），并单击标题为 “证据”的框。可以切换为饼图显示的是*条件概率分布*，即：由圆饼图表示的具有特定属性值的用户 （例如：“声音邮件方案”列中的值 “是”）具有客户波动值 “是”的概率。

单击一个图以更新右边的面板，显示根据模型得到的期望概率。

用屏幕边框上的滚轮来漫游，或者使用鼠标按钮和 Ctrl 键的不同组合来漫游。关于漫游控制的详细信息，请参阅 附录 A，“在 MineSet 可视化工具中漫游”。

您可以看到影响客户波动的因素，因为在 图 3-4 的第一和第二行中表示客户波动的薄片从左向右逐渐增加，所以，明显存在一个严重的问题。使用公司服务最多的用户也更可能频繁地发生客户波动。公司将失去它最宝贵的用户。

要找到一个类标签 （例如：客户波动值为 “是”），可以在右边的 “标签概率”面板中选择一个值。在 “右边”面板中单击标签 “是”的按钮，证据将会显示为条形图。用鼠标指向条形图将会显示出估计的概率。

这里显示的具有差别的属性也可以用来为散点图形可视化工具选择坐标轴。“州”属性在属性列表中显示得相对较高，暗示着可能有地理联系。

尽管证据模型独立地使用并显示各属性，但是在许多数据库中属性都不是独立的，而且，将一个属性集结合起来进行考虑更有利于标签的确定。状态窗口中的误差估计显示分类器的误差率期望值大概为 12%，稍后我们将生成一个更加准确的决策树。在转而使用 “平伸可视化工具”之前，请关闭 “证据可视化工具”。

## 利用平伸可视化工具查看概率

“平伸可视化工具”需要映射到颜色的列必须具有一个数值。客户波动的列是一个字符串，它必须被转换为一个数字 （p_churned 表示客户波动的概率），再映射到 “平伸可视化工具”中：

1. 在 “数据目标”面板上面一行选项卡中单击 “可视化工具”选项卡，然后在下面一行选项卡中，单击 “平伸”选项卡以进入 “平伸可视化工具”。
2. 在 “数据转换”面板中，单击*添加列*。

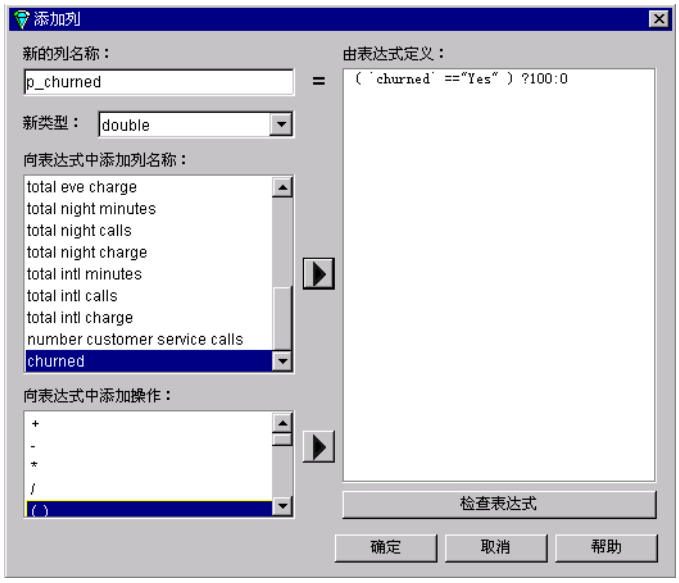



**图 3-5** 添加一个新列

3. 在 “添加列” 对话框 ( 图 3-5) 的 “新列名称” 文本字段中，输入新的名称，**p_churned**。目的是为了根据客户波动列建立一个数字列。

   在 “由表达式定义” 文本字段中，创建表达式 **(‘churned‘ == ”yes”) ? 100 : 0**。您可以从左边的两个滚动列表中创建该表达式：“向表达式中添加列名称” 和 “向表达式中添加操作”，或者您可以将其直接键入。该表达式可翻译成 “如果客户波动列中的值为 “是”，给 p_churned 赋值 100，否则，给它赋值 0。” 该表达式的目的是将一个字符串 （是或否）翻译转换为一个数值。请确认 “新类型” 文本字段被设置为双精度。

   单击 *检查表达式* 以确保没有语法错误。单击 *确定* 以关闭对话框并再单击 *确定* 添加列。

4. 在 “数据转换” 面板的 “平伸” 选项卡下，选择每个元素旁边下拉菜单中的列，将列映射到可视元素。在本教程中，下拉菜单为：

   对于坐标轴 1，选择 “白天的费用总数”。

   对于坐标轴 2，选择 “用户服务呼叫数”。

   对于坐标轴 3，选择 “国际方案”。

   对于颜色，选择 “p_churned”，结果将类似于 图 3-6。

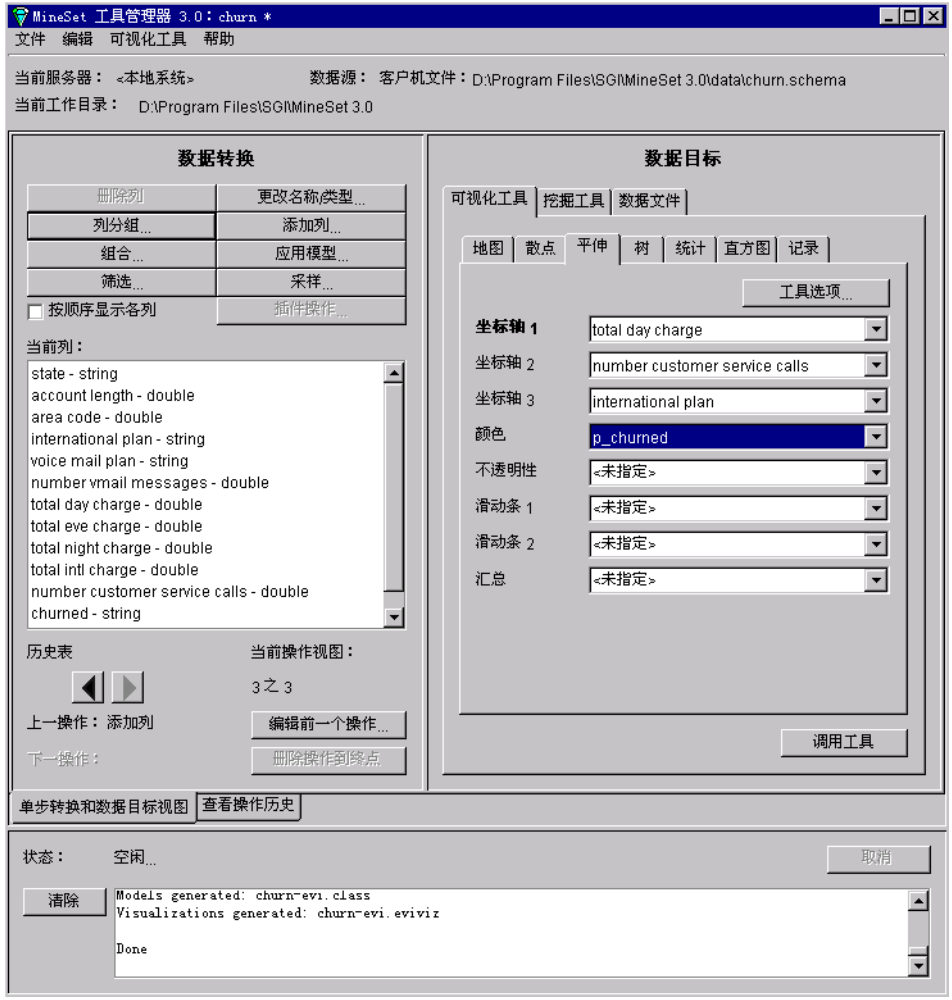



**图 3-6** 对于平伸可视化工具将列映射到元素

5. 单击 *调用工具*。

   如图 图 3-7 所示，数据被绘制在 “平伸可视化工具” 窗口中。左上方滑动条可以控制色彩浓度变化。关于窗口操作的帮助，请参阅 附录 A，“在 MineSet 可视化工具中漫游”。要在场景中漫游并查看不同的区域，可以同时按住鼠标的左右按钮，并在场景中移动光标。“平伸可视化工具” 可以使您通过查看复杂数据在多维空间内的不断变化，来帮助进行分析。

   您可以保存 “工具管理器” 的当前状态，包括特定的选项。具体步骤为：选择 “文件” > “将当前会话保存为”，然后指定 *churn1.mineset*。

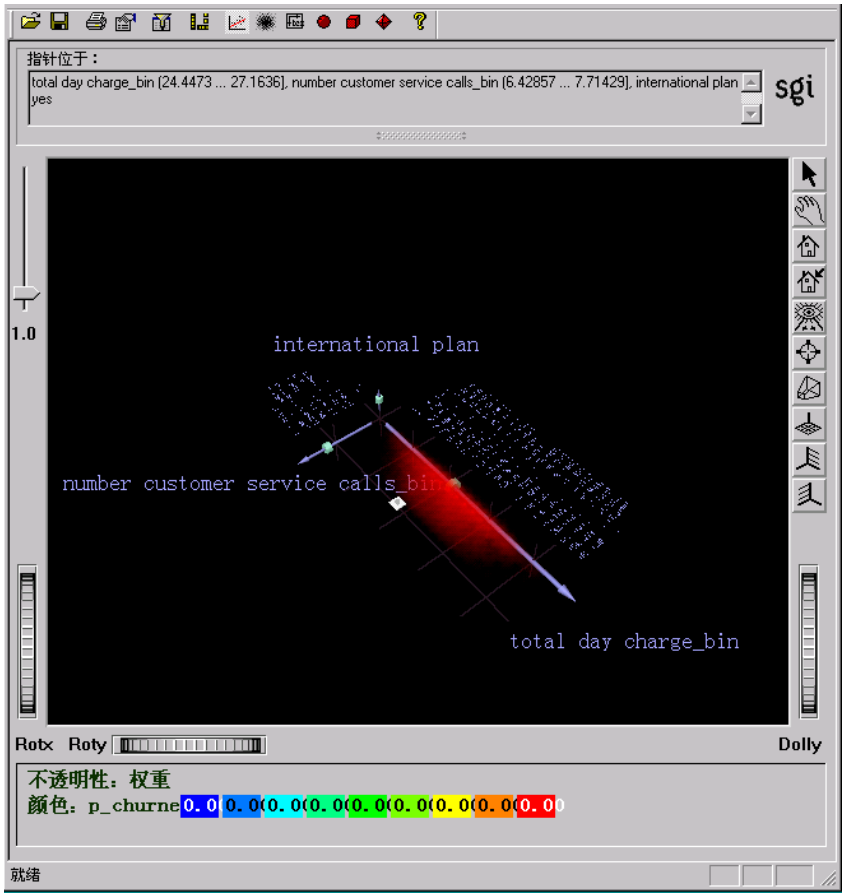



**图 3-7** “平伸可视化工具” 窗口

在 图 3-7，所示的可视化处理中，有两处客户波动的概率最高：白天的费用总数高的用户在黄色到红色区域，显示在该图的底部；而白天的费用总数低的用户服务呼叫数高 （在该图左上部附近）。拨打很多用户服务电话的低付费用户离开了，您可能并不希望保留这些用户，因为他们用了您的钱却回报很少。图形底部的高付费用户是更好的对象。

6. 关闭 “平伸可视化工具”，返回 “工具管理器”窗口。

## 可视化地理分布

如 第 19 页的 图 3-4，所示，证据模型表明 “州”是一个良好的区分属性。如果您的模型没有此项显示，原因可能是您没有如 第 18 页的 “建立一个证据分类器”，中第 4 步所述改变最大值。本部分建立在以前的计算之上，按地理方位显示数据以说明客户波动是如何随着 “州”变化的。

您已经添加了由数据集原有列生成的列 （p_churned）。现在您可以将数据转换到一个较小的数据集中，其中包含每个州的平均客户波动值。这种转换叫做组合。

1. 在 “工具管理器”窗口的 “数据转换”面板中，单击*组合*。

   在 “组合”对话框中将 p_churned 移到左列中 （选中它，然后单击左箭头）。在新窗口中突出显示 p_churned，选中 “平均值”和 “计数”，并确保没有选中 “总和”、“最小值”和 “最大值”。将州留在中间列中，然后将其余所有项目移到右列。（按住 Ctrl 键以选中多个列。）请确保您的屏幕如 图 3-8 所示。单击*确定*以应用这些选项。

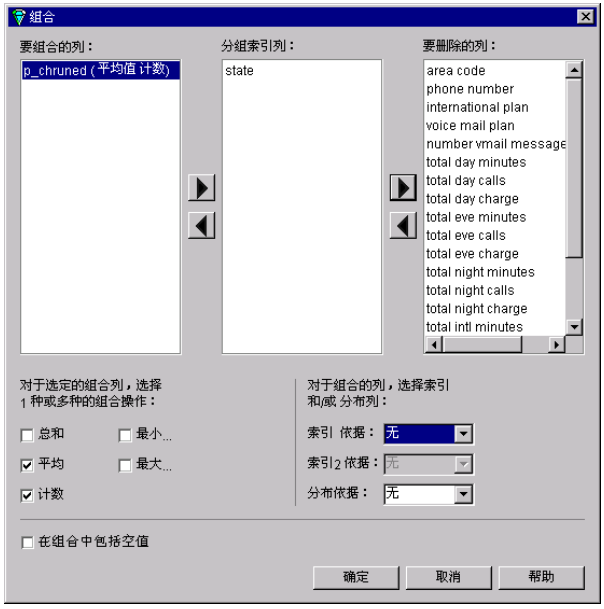



**图 3-8** “聚合”对话框

2. 在“数据目标”面板的上面一行选项卡中，单击“可视化工具”选项卡，然后再单击“记录”选项卡。单击*调用工具*以查看每个州的记录，其中包括平均客户波动的用户数和每个州客户波动用户的总数。
3. 关闭“记录查看器”窗口，返回到“工具管理器”窗口。您现在将把该数据链接到美国地图。
4. 从“工具管理器数据面板”面板的下面一行选项卡中，单击“地图可视化工具”选项卡。然后单击*工具选项*按钮。将出现“地图可视化工具选项”对话框 ( 图 3-9)。
5. 单击“实体文件”文本字段右侧的按钮，从此处进入 MineSet 的安装目录，并选择 *config* > *mapviz* > *gfx_files*>*usa.state.hierarchy.*


**图 3-9** 地图可视化工具选项对话框

6. 单击*打开*以选定该文件，然后单击*确定*关闭“地图可视化工具选项”面板。下一步是将可视元素与列相链接。

7. 将当前“数据转换”面板中的列映射到“数据目标”面板中的元素，具体步骤如下（请参阅 图 3-10）。从可视元素旁边的下拉菜单中：

   实体条选择州
   高度条选择 count_p_churned
   颜色条选择 avg_p_churned.

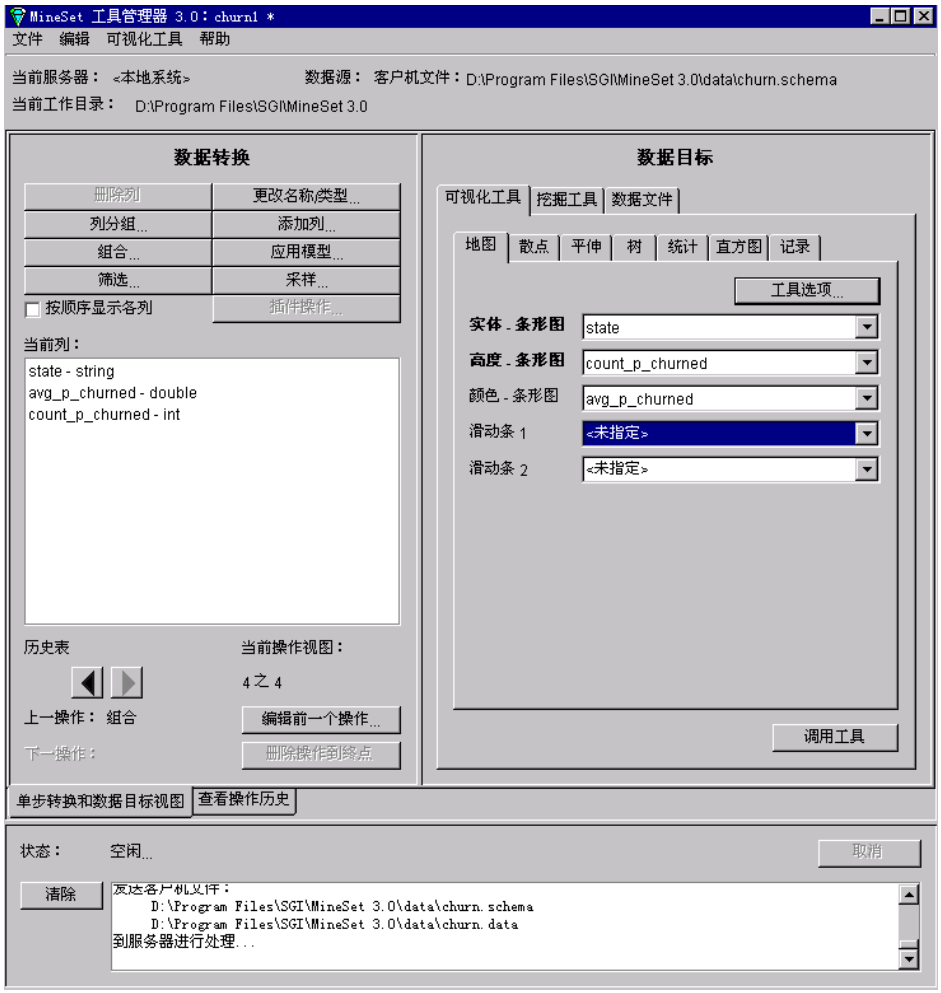



**图 3-10** 可视化工具将列映射到可视实体

8. 单击 *调用工具*以查看客户波动用户按照州的分布 ( 图 3-11)。

可视化处理显示客户波动用户在美国的分布情况。对于每个州，颜色指示客户波动的概率，高度指示该州的用户数目。例如，在 图 3-11 中，选定了缅因州，虽然显示出平均客户波动率为 18.4466%，但是客户波动计数为 103。即：该平均值只是基于 103 名用户的。西弗吉尼亚州显示的最高，客户波动率是基于 158 名用户的。颜色显得最清楚、明亮的州的平均客户波动率都超过了 21%（得克萨斯、蒙大拿、华盛顿、加利弗尼亚和新泽西）。这个可视化处理表明在客户波动和地理位置之间没有明显的关系，尽管不同的州的客户波动率确实不同。

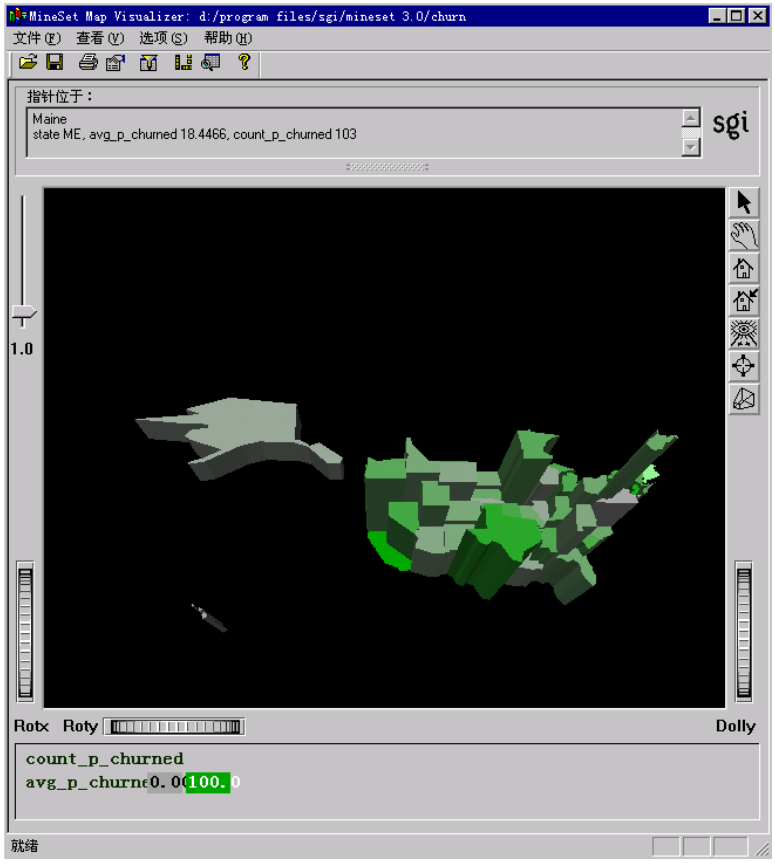



**图 3-11** 带有平均客户波动分布的 “可视化工具”窗口

使用 “文件” > “退出” 来关闭 “地图可视化工具”。下一个例子研究 “决策树” 分类器，使用同一个数据集生成不同的可视化处理。

## 创建决策树分类器

与 “证据” 分类器不同，“决策树” 分类器可以显示属性的相互作用，即：影响分类的属性值的组合。本部分使用历史模式开始，按照以下步骤建立一个决策树分类器并将其可视化。

1. 单击 “工具管理器数据转换” 面板底部的 “操作的历史视图” 选项卡，切换到历史模式。
2. 用鼠标右键单击 “添加列” 和 “聚合” 操作并选择删除，将它们从历史中删除。
3. 切换回 “单步转换和数据目标视图”。现在您应该在 “当前视图为：2 的 2”
4. 在 “数据目标” 面板的上面一行选项卡中，单击 “挖掘工具” 选项卡。
5. 在下面一行选项卡中，单击 “分类” 选项卡，然后从下拉菜单中选择以下选项：

   *模式：*分类器和误差

   *导入工具：*决策树

   *离散标签：*churned
6. 单击*继续*

   MineSet 将分类并创建 “决策树” 模型，如 图 3-12 所示。比起 图 3-4 中 “证据可视化工具” 显示的估计错误率有明显改进 （6.36% + / -0.60%），证明了早期的假设，即：属性之间的相互作用是显著的。在 图 3-12，决策树中的每个节点上都有两个条状物，每条对应一个标签值。鼠标指向条将显示记录计数和标签值的百分比。每个节点的基准指示到达该节点的记录数、颜色和每个子树的估计误差率 （请参见可视化处理底部的图例）。

   在这个例子中，决策树的根部是白天的费用总数，表明这是最重要的唯一因素 - 用户白天打电话的费用，划分的阈值为 44.96。

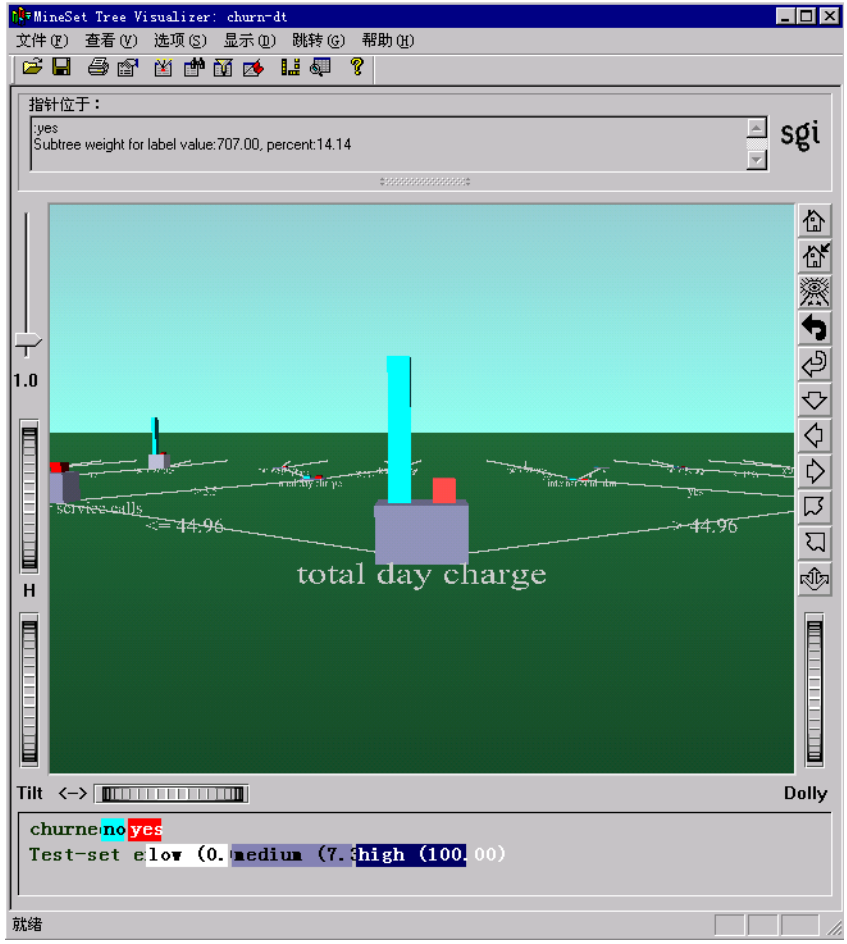



**图 3-12** “树可视化工具”窗口

您可以使用伸缩轮或鼠标按钮的组合在 “树可视化工具” 中横向浏览。（关于漫游的详细信息，请参阅 附录 A，“在 MineSet 可视化工具中漫游”）。通过选择根部的红色条（客户波动的：是），在右边您可以看到客户波动的用户为 14.14%。单击右边的线 （白天的费用总数 >$44.96）以转到子节点，其中包含在白天电话花费最多的用户。这些用户中有 59.31% 客户波动。再次单击任一条支线以转到下一个节点，它将显示在白天电话花费最多的用户，以及可能或不可能拥有声音邮件方案的用户。在这些用户中，拥有声音邮件的用户客户波动率较低，为 9.33%。可能向用户提供声音邮件可以帮助减少客户波动。

要理解这个树是自动从数据中导出的，这一点很重要，在导入过程中决定了节点的属性和阈值。

要细节追溯和查看原始数据，可以选择一个节点基准成一个条，然后选择 “选项” > “显示原始数据”。“记录查看器” 将显示与选定节点相匹配的记录。

如果您希望进一步探索 MineSet，并了解有关应用分类器的详细信息，请阅读下一章：第 4 章，“进一步的探索”。

# 进一步的探索

本章继续探讨 MineSet 工具。假设您已经阅读了 第 3 章，“客户波动教程”，并已准备使用 MineSet 的其他部分：



## 探索数据聚类

当您遇到不熟悉的数据集时，利用聚类运算法会发现一些有趣的属性或特征。这个非预测性运算法则将记录划分为在一些方面类似的聚类。对于这个示例，返回到 “工具管理器”窗口，重新打开 *churn.schema* 文件，开始一个新的历程。

1. 在 “数据目标”面板的上面一行选项卡中，单击 “挖掘工具”选项卡。
2. 在下面一行选项卡中，单击 “聚类”选项卡，进行以下选择：

   *方法：*单步 k- 均值

   *聚类数：*3

3. 在 “工具管理器” 的 “数据转换” 面板中，选中以下列，然后单击*删除列*（请参阅图 4-1），将其从 “当前列” 面板中删除：

   state
   account length
   area code
   phone number
   international plan （因为它与国际的费用总数相关）
   voice mail plan （因为它与声音邮件消息数相关）
   可以使用 Ctrl 键进行多项选择

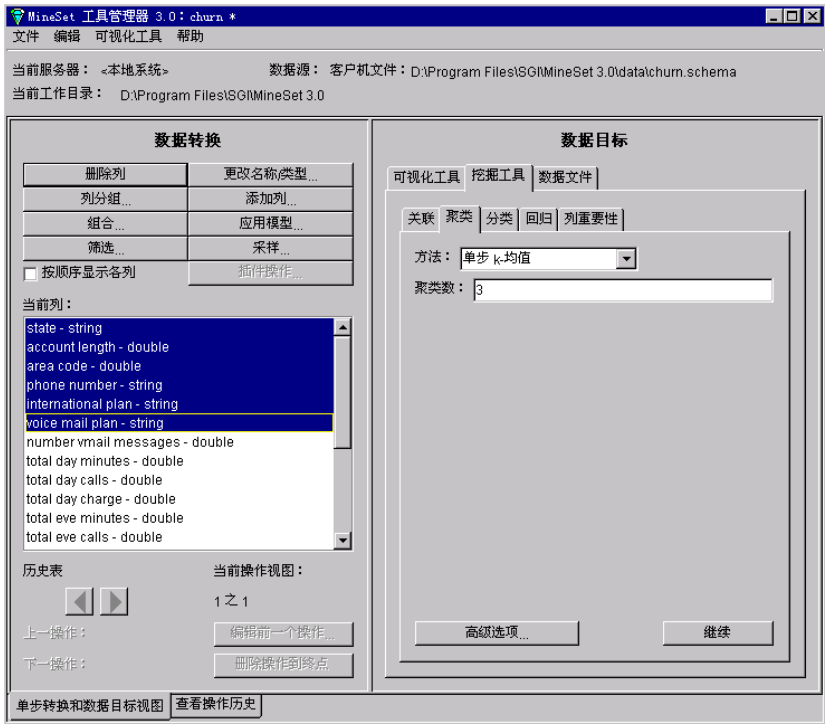



**图 4-1** 删除列以准备聚类

被删除的列是那些不太可能影响聚类结果的列。“客户波动” 列将被保留以帮助解释结果。当您进一步研究数据集时，可以试验删除其它的列。

4. 单击*高级选项*以设置属性的权重。

   默认情况下，每个列的权重都被设置为 1，即：每个列的重要性都相等。在该例中，客户波动列被设置为 0，用来查看当数据集被聚合时，该属性是否能够自发生成。单击*设置*，然后再单击*确定*。

5. 在 “工具管理器”窗口的右侧单击*继续*。

   当运算法则选择用于划分记录的显著特征时，“工具管理器”底部的 “状态”窗口会显示聚类操作的进程。模型被保存为 *churn.cluster*。

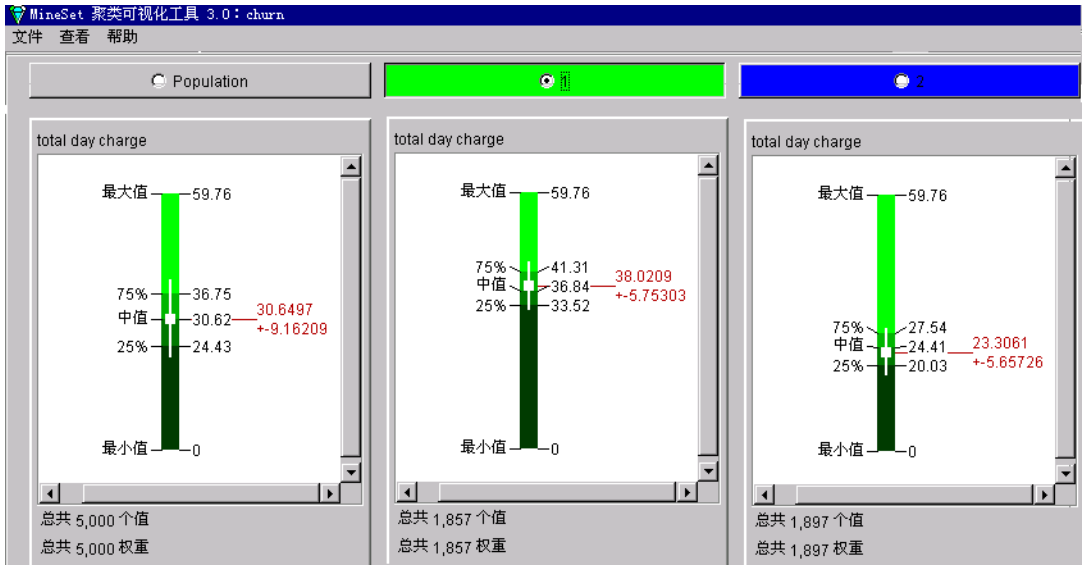



**图 4-2** “聚类可视化工具”产生的盒形图

图 4-2 所示为 “聚类可视化工具”结果的部分视图。调整窗口以查看所有聚类。按照列的重要性对其进行排序以区分聚类之间的差别。显然，声音邮件消息、白天的时间总数、以及白天的费用总数是最重要的列。顶部列中聚类之间的颜色和平均值差别很大，但是随着画面的下滚，这种差别会逐渐变小。

6. 在画面的顶部，单击聚类编号旁边的圆圈。这将会改变属性次序，即用于区分该聚类与其他类的重要属性会改变次序。
7. 在 “聚类可视化工具”窗口中选择 “文件”> “退出”，以关闭窗口并返回到 “工具管理器”窗口。

## 将模型中的列与坐标轴联系起来

利用 “聚类可视化工具”，您可以查看数据集中独立的属性，检查最显著的属性，并了解它们之间的差别。但是，如果要了解聚类之间属性的相互关系，就需要使用 “散点可视化工具”，它可以提供一个更清晰的视图。要将聚类模型应用于 “散点可视化工具”，需要确定哪些列被映射到各种不同的坐标轴。

1. 在 “工具管理器”的 “数据转换”面板中，单击*应用模型*并从可用模型列表中选择 *churn.cluster* 。单击*确定*。

   虽然 “聚类可视化工具”显示最主要的有三列，但是每个聚类的重要性次序是独立的，在属性之间没有相互作用的迹象。在这点上，“列重要性”工具很有用。
2. 在 “数据目标”面板的上面一行选项卡中，单击 “数据文件”选项卡，然后再单击 “服务器”复选框。在文本字段中，键入文件名 **churn-crop** ，然后单击*创建文件*。这将保存客户波动数据集的简短版本，以备本教程以后使用。


**图 4-3** 将数据文件保存到服务器

## 在聚类模型中查找重要的列

1. 在 “工具管理器” 窗口的 “数据目标” 面板中，单击 “挖掘工具” 选项卡，然后再单击 “列重要性” 选项卡。默认情况下，工具将按照重要性顺序选择顶部的三列，离散的标签是 “cluster”。


**图 4-4** 聚类的列重要性选择

2. 单击*继续*。

   出现的面板显示：

   1. number vmail messages
   2.total day minutes
   3.total eve minutes

   状态窗口显示白天花费在电话上的时间是一个关键，所有其他列显示出相关性。下一步是将这些列映射到 “散点可视化工具” 中的坐标轴。

## 映射到 “散点可视化工具”

1. 在 “数据目标”面板上面一行选项卡中单击 “可视化工具”选项卡，然后在下面一行选项卡中，单击 “散点”选项卡以进入 “散点可视化工具”。

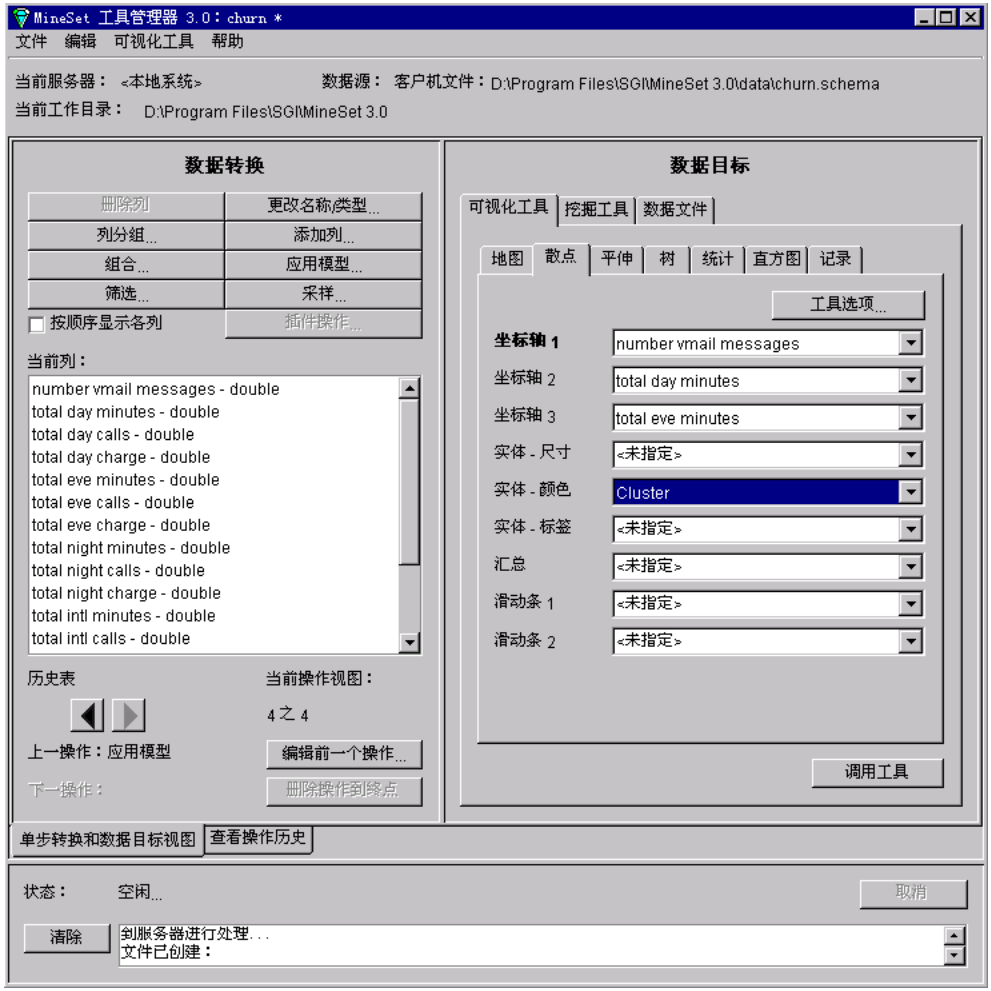



**图 4-5** 将映射到 “散点可视化工具”中的坐标轴

2. 在 “数据目标” 面板中，使用下拉菜单将这些元素映射到以下的列 （请参阅 图 4-5):

   坐标轴 1 选择 number vmail messages
   坐标轴 2 选择 total day minutes
   坐标轴 3 选择 total eve minutes
   实体颜色选择 “cluster”（在应用模型时创建）

3. 单击 *调用工具*。

   图 4-6 中的 “散点可视化工具” 窗口清楚地显示出聚类颜色上的区别。蓝色散点立方体代表聚类 2，扁平饼形平均地被红色和绿色分开 - 聚类 1 和聚类 3。这个饼形表明声音邮件的数目很低。显然，白天的分钟总数和晚上的分钟总数是相互依赖的。如果您单击了一个感兴趣的可视点，就会显示相关数据。关闭 “散点可视化工具” 窗口，返回 “工具管理器”，进行下一步。

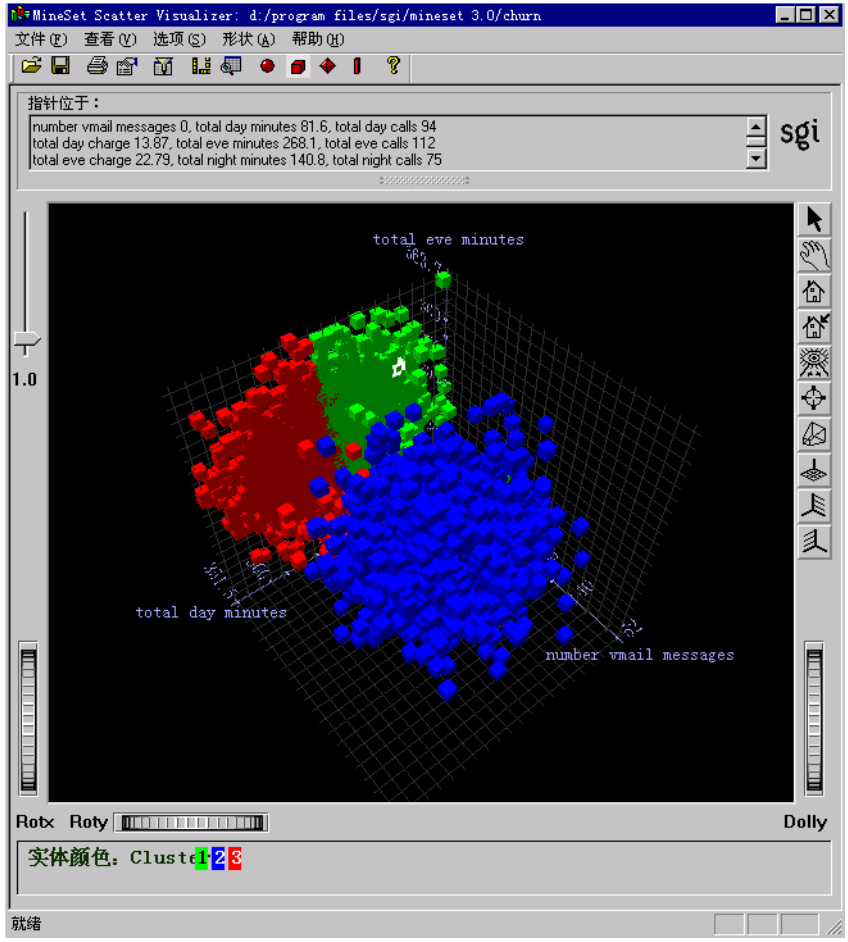



**图 4-6** 根据聚类绘制的“散点可视化处理”

## 调用决策表

在这个例子中，不是使用 “散点可视化工具” 来查看聚类数据，而是可以将这些数据可视化为一个 “决策表”。

1. 在 “工具管理器” 窗口中，选择 “文件” > “打开新的数据文件”。
2. 在 “打开 schema 文件” 窗口中，单击*服务器文件*按钮，然后选择 *churn-crop.schema* 这是以前保存的文件。如果您在会话之间激活了 MineSet，您将自动被带回到离开的地方。这可以通过选择 “工具管理器文件” > “首选项” 菜单来设置。
3. 单击文件，然后再单击*打开*。
4. 在 “数据目标” 面板的上面一行选项卡中，单击 “挖掘工具” 选项卡。
5. 在下面一行选项卡中，单击 “分类” 选项卡，然后从下拉菜单中选择以下选项：

   *模式*：分类器和误差估计

   *导入工具*：决策表

   *离散标签*：churned

   请确保您的离散标签正确。您将要导入一个决策表并允许运算法则建议哪些列对于映射到坐标轴 X 和 Y 是最重要的。
6. 请验证*建议*复选框被选中，然后单击*继续*。

当工具确定了合适的映射后，您就可以看到被映射到坐标轴的列了。“工具管理器” 底部的 “状态” 窗口将显示导入过程的进程和汇总信息，包括分类误差率。当导入步骤完成以后，系统将自动调用 “决策表可视化工具”，显示模型。操纵伸缩滚轮以调整视图，或者用鼠标按钮进行漫游 - 请参阅 附录 A，“在 MineSet 可视化工具中漫游”。

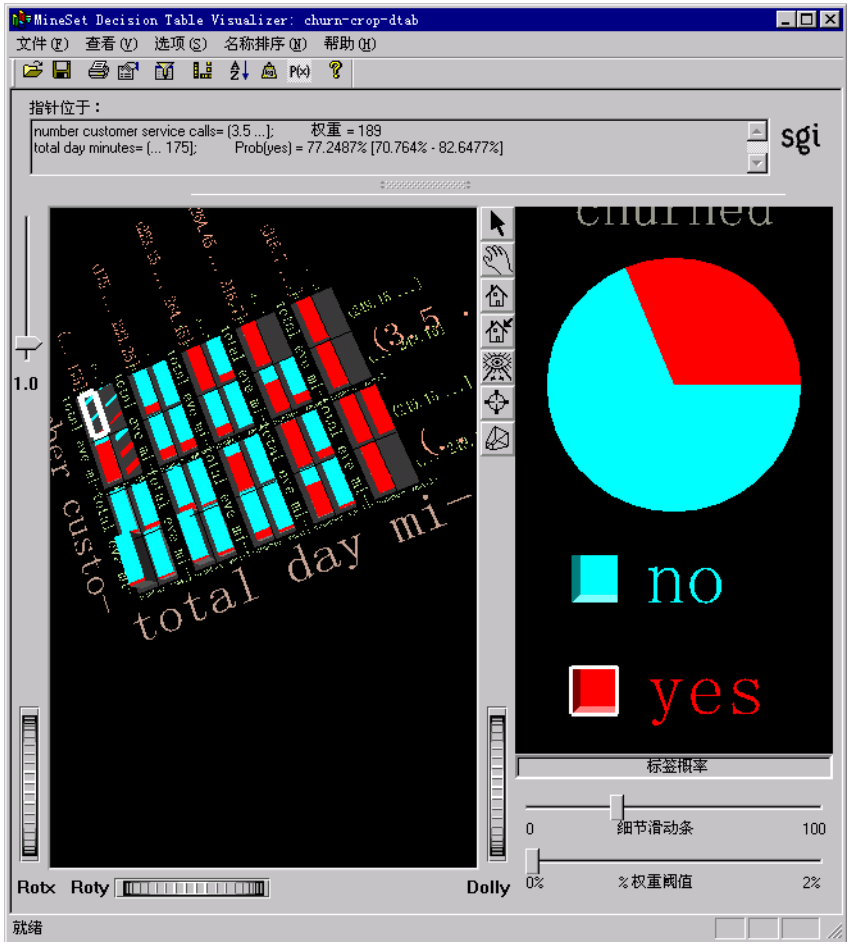



**图 4-7** 显示聚类客户波动结果的决策表

图 4-7 在 “标签概率”面板中显示了一个圆饼图，类似于 “证据可视化工具”，指示出客户波动的总百分比。在左侧的面板中，数据被显示为块状图，它表明在数据的某一子集中存在多少客户波动。显然，白天分钟总数高的用户总是波动。

“决策表”将显示不同等级细节的数据，开始只考虑几个列，随着您进一步的查看将增加更多的细节。将光标模式从“抓取”改变为“选取”，并用光标掠过场景，相关数据显示在窗口之上。请注意落在预期模式以外的部分 - 白天的分钟总数少于 175，并且用户服务电话超过 3.5。使用鼠标按钮进行概化上寻和细化下寻，
请参阅 附录 A，“在 MineSet 可视化工具中漫游”，完成对 “决策表”的检查后，请关闭画面并返回到 “工具管理器”。

## 针对用户使用模型

您曾经创建了一个模型来预测哪些用户可能客户波动。既然您已经有了一个这样的模型，就可能希望在可能客户波动的用户客户波动*之前*针对他们。上升曲线可以帮助您完成这个任务。

在上升曲线图中，坐标轴 X 显示从 0 到 100% 的记录数，坐标轴 Y 显示具有给定标签值的用户的记录数 （在这里， *churn= 是* ）。在 图 4-10 中显示了两条曲线。下面的曲线或直线 （红色）显示的是在给定一个记录随机次序的情况下预期客户波动的用户百分比。上面的曲线 （白色）显示的是当按照分类器对每个记录的评分 （概率估计）安排次序时客户波动用户的百分比，被模型识别为最可能客户波动的用户记录将首先出现；那些概率较小的记录将出现在后面。模型次序的优点在于可以发现模型曲线与随机曲线之间的差别。

建立该上升曲线时，在测试集中应用了一个选定的模型。在下面的例子中，用一个指定的数据集段来进行训练，然后将导入的模型在数据集的剩余部分上运行。虽然可以通过在分类器的 “高级选项”中选择 “上升曲线”来生成上升曲线，但是在本教程中将介绍一个更为复杂的示例，一个涉及到采样和将模型应用到数据集的示例。

## 创建一个训练样本

进行对于这个示例，请先返回到 “工具管理器”基本窗口，然后使用 “文件”> “打开新的数据文件窗口”并返回到本地文件 *churn.schema* 以开始一个新的历程。

1. 在 “数据转换”面板中，单击*采样*。在 “采样”对话框键入采样的百分比 *40*，然后单击*确定*。

   该选择仅简单地在整个导入模型的数据集中随机采集 40% 的样本。


**图 4-8** 选择一个样本进行测试

2. 在 “数据目标”面板上面一行选项卡中单击 “挖掘工具”选项卡；然后在下面一行选项卡中，单击 “分类”并从下拉菜单中进行以下选择：

   *模式：*仅用于分类器

   *导入工具：*决策树

   *离散标签：*churned

   您正在导入一个基于 40% 随机样本的决策树分类器，选择 “仅分类器”，是因为这只是训练集。测试集就是数据集的剩余部分 （除了 40% 已采集的记录以外）。

3. 单击*继续*

   产生的决策树将对模型进行演示，这在下一个阶段是必需的。根权重被相当程度地减少，因为样本的数量小于整个数据集，并且在每个节点的底部没有显示任何颜色，表明没有可用的误差估计。

您可以在状态字段看到，分类器被自动保存在名称 *churn-dt.class* 下。下一步是在客户波动数据集的剩余部分上使用该分类器。

## 应用模型

关闭 “决策树”窗口，返回到 “工具管理器”窗口。因为您已经使用了数据集的 40% 来建立模型，所以还剩下 60% 作为测试集。

1. 在 “数据转换”面板中，单击*编辑上一个操作*。“采样”对话框会再次出现。
2. 在 “采样”对话框中，再次在 “百分比”文本字段中输入 40，但是这一次单击 “互补样本”框，表明您需要样本的其他部分。
3. 单击*确定*。
4. 单击 “数据转换”面板中的*应用模型*按钮。
5. 从可用模型列表中选择 *churn-dt.class*。这是建立在客户波动数据集之上的 “决策树”模型。
6. 在面板的下半部分单击 “测试模型”选项卡；打开*显示上升曲线*，并将 *ROI/ 上升标签*下拉菜单设置为 “是”。

建立好基于随机样本的分类器后，就可以将其应用到客户波动数据集的剩余部分了。


**图 4-9** 准备对整个数据集测试分类器

7. 单击*运行测试*。该过程可能需要花费一些时间。图 4-10 所示为生成的上升曲线，选定点的详细信息显示在上部。

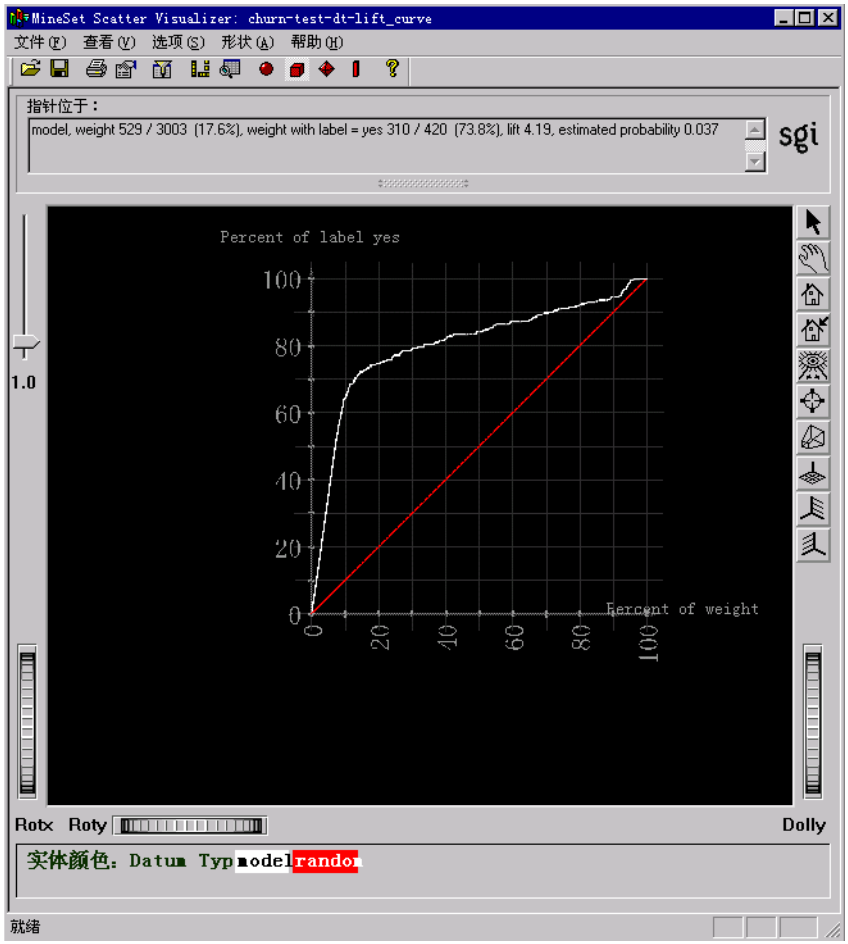



**图 4-10** 上升曲线

沿着白色 （模型）线移动鼠标，在不同的点上单击，查看上升曲线以及 *churn = 是*的用户的百分比。寻找曲线的拐点，在该例子中是分类器的估计概率为 0.054 的地方。

在该点，向可能客户波动的用户发送刺激的投资回报迅速减少。下一步是将分类器应用到整个数据集。

8. 返回 “应用模型” 对话框；单击 “应用模型” 选项卡，选择 *churn-dt.class* 并进行以下选择：

   *标签概率估计值*是

   *新列名称*：**p_churned**（您必须将其键入。）

   当您单击*标签概率估计值*时，要选择 “是” 来与 “测试模型” 步骤中的相应选择相匹配。该过程将添加一个新列 （p_churned），它代表某些即将客户波动用户的概率。单击*确定*。

9. 在 “工具管理器” 的 “数据转换” 面板上单击*筛选*；在 “由表达式定义” 文本字段中 创建表达式 **p_churned > 0.054**。检查表达式，然后单击*确定*。

   图 4-10 中所示是由第 8 步检索而得的估计概率图。目的是为了只选择那些最可能客户波动的用户。在现实生活情况下，这一步针对未标注的数据进行，以预测哪些现有的用户可能客户波动。

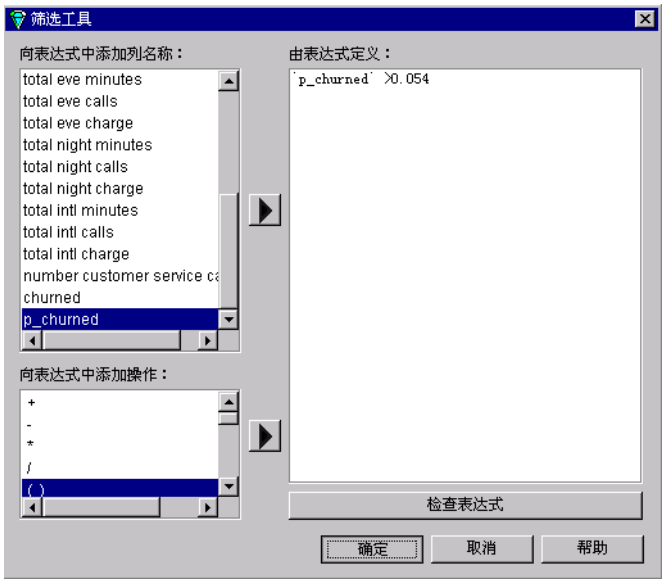


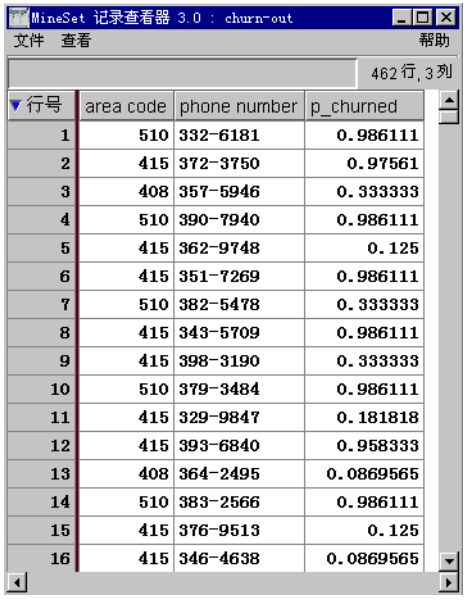

**图 4-11** 为客户波动概率进行筛选

最后，您可以查看 “记录查看器” 中的结果，删除早期引用中不必要的列，具体细节如下。

10. 在 “工具管理器” 的 “数据目标” 面板中，单击 “可视化工具” 选项卡，然后单击 “记录” 选项卡。在 “数据转换” 面板中，选择除地区代码、电话号码和 p_churned 以外的所有列，然后单击*删除列*按钮。要选择多列，可以按住 Shift 键进行连续选择，或按住 Ctrl 键进行特定选择。

11. 单击 “调用工具”。

    如 图 4-12 所示，结果是一个很有用的电话列表，列出的是根据该模型得出的最有可能客户波动的用户的电话。




| 行号 | area code | phone number | p_churned |
|---|---|---|---|
| 1 | 510 | 332-6181 | 0.986111 |
| 2 | 415 | 372-3750 | 0.97561 |
| 3 | 408 | 357-5946 | 0.333333 |
| 4 | 510 | 390-7940 | 0.986111 |
| 5 | 415 | 362-9748 | 0.125 |
| 6 | 415 | 351-7269 | 0.986111 |
| 7 | 510 | 382-5478 | 0.333333 |
| 8 | 415 | 343-5709 | 0.986111 |
| 9 | 415 | 398-3190 | 0.333333 |
| 10 | 510 | 379-3484 | 0.986111 |
| 11 | 415 | 329-9847 | 0.181818 |
| 12 | 415 | 393-6840 | 0.958333 |
| 13 | 408 | 364-2495 | 0.0869565 |
| 14 | 510 | 383-2566 | 0.986111 |
| 15 | 415 | 376-9513 | 0.125 |
| 16 | 415 | 346-4638 | 0.0869565 |

**图 4-12** 记录查看器结果

在 “记录查看器” 中，对于每个记录都有一个数字，估计用户客户波动的概率。筛选保留了最高数字的那些用户。它提供的只是可能客户波动的用户列表，您应该对他们发送刺激 （例如：通过电话招揽、发送信件等等）。关闭 “记录查看器”，继续探索客户波动数据集。

# 减少分类错误代价

在 MineSet 中，您可以使用以下三种工具来减少模型建立中的错误：混淆矩阵可以给出错误和不正确预测的详细图示；损失矩阵可考察比较严重的错误；而投资回报曲线将显示在何时投入更多的时间和金钱会收获甚微。

## 显示一个混淆矩阵

返回到 “工具管理器”窗口，重新打开 *churn.schema*。

1. 在 “数据目标”面板中单击 “挖掘工具”面板；然后单击 “分类”并从下拉菜单中进行以下选择：

   *模式*：分类器和误差估计

   *导入工具*：决策树

   *离散标签*：churned

2. 单击*高级选项*，将出现 图 4-13 所示的 “分类器”选项面板。

   在该过程中会出现一条消息：属性 “phone_number”被删除，因为它具有 100 个以上不一致的值。单击 “确定”，进行处理。

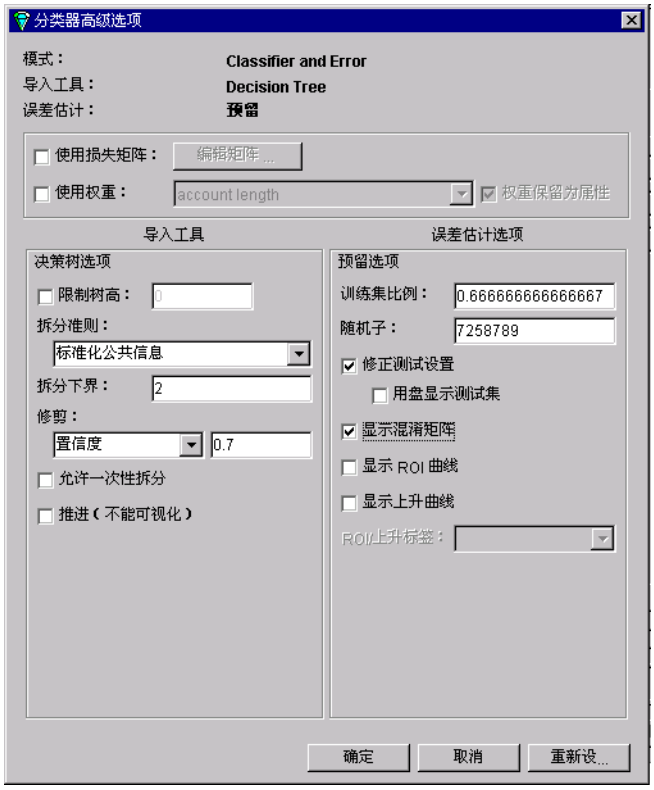



**图 4-13** “分类器高级选项”面板

3. 打开*显示混淆矩阵*和*修正测试集*。请确保关闭了*显示上升曲线*和*显示 ROI 曲线*，然后单击*确定*。

4. 在 “工具管理器数据目标”面板的 “分类”面板中单击 “继续”。

“混淆矩阵”将显示分类器在分类中出现错误的地方。关闭 “树可视化工具”，检查 “混淆矩阵”。从这里，您可以根据您对数据的了解建立一个 “损失矩阵”，使对某些错误的容错性降低。

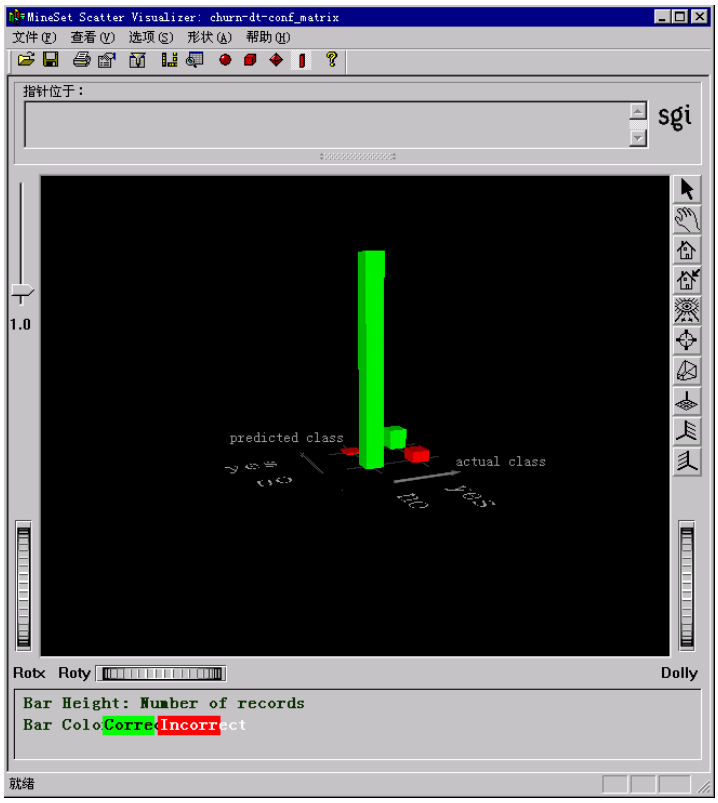



**图 4-14** “混淆矩阵”显示正确和错误的分类

在 图 4-14 所示的窗口中，两个绿色条形 （一长一短）代表正确的分类。两个红色条形代表错误分类。在分类中发生的重要错误分类由两个红色条形中的较大一条代表 - “被预测的类：否，实际的类：是。”这些用户预计不会客户波动，但是实际上却客户波动了，代价高的错误分类达到了 4.7% （要查看百分比，可以使用 “选项”> “显示值”）。您可以尝试在对数据的现有了解上建立一个 “损失矩阵”，来减少该错误，为这个红色条形代表的错误提高加权。

## 定义损失矩阵

建立 “损失矩阵” 的目的是控制哪些错误分类器需要多加注意，以及哪些错误分类器需要避免。

1. 通过 “文件” > “退出” 关闭 “混淆矩阵” 画面，返回到 “工具管理器” 窗口。
2. 单击*高级选项*，返回到 “分类器” 选项面板。
3. 在 “分类器高级选项” 对话框中，打开*使用损失矩阵*。
4. 单击*编辑矩阵*，加权出现错误的代价。将出现一个 “损失矩阵” 面板，如 图 4-15 所示。

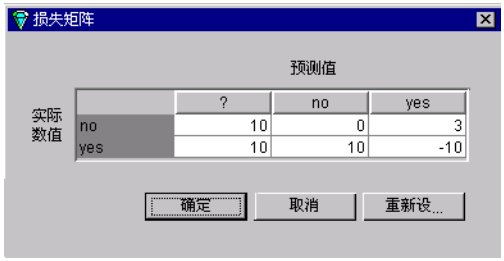



**图 4-15** “损失矩阵” 显示权重

5. 在 “损失矩阵” 的行中从左到右设置以下值：

   *实际值*：否：10—0—3

   *实际值*：是 : 10—10—(-10)

   问号下列中的值应该高一些，以免分类器预测为 “未知”。

使用这些值，如果您预测一个用户不会客户波动并且预测正确，就既无收益也无损失（以零代表）。如果您预测一个用户不会客户波动 （因此就没有向他们发送任何刺激），而他们却客户波动了，就会导致 10 的损失 （以 +10 代表，该数字表示损失）。如果您错误地预测一个用户要客户波动，而他们却没有，就会损失 3，代表发送一个不必要邮件的代价。如果您的邮寄计划起到了作用，保留了一些原本要客户波动的用户，您就可以得到 10 （由 -10 代表）。下一步应该是调查您的投资回报。

## 查看投资回报曲线

投资回报曲线使您可以看到出现某种错误的代价，并向您指出继续采取行动也不会再有效的那一点。

1. 请确保打开了 “分类器高级选项”面板中的*修正测试集*、*显示混淆矩阵*、以及*使用损失矩阵*。
2. 请确保*显示 ROI 曲线*也被选中。
3. 请确保 *ROI/ 上升标签*被设置为 “是”，然后单击*确定*。
4. 在 “工具管理器”窗口的 “分类”面板中单击*继续*。

将出现三个画面窗口：决策树、混淆矩阵和 ROI 曲线。“混淆矩阵”显示出分类器在进行 churn=no 预测时更加保守，因此降低了假负性。一边的错误增加了，但是另一边的错误却减少了。关闭 “混淆矩阵”和 “决策树”画面，查看 “ROI 曲线”窗口。

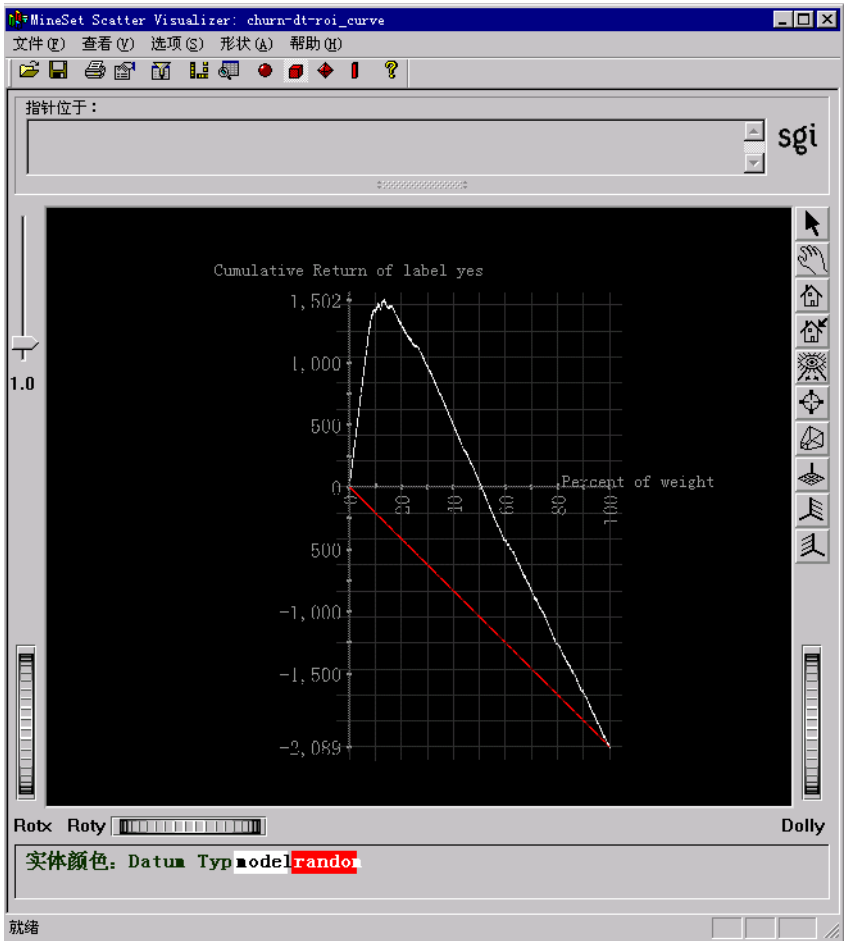



**图 4-16** 投资回报曲线

图 4-16 所示的 ROI 曲线与上升曲线有类似之处。横穿中间的水平线代表零收益和损失。红线代表采用随机总体样本并向他们发送邮件产生的预期情况，如果您向每个人都发送了邮件，由于邮寄的代价就会带来损失。但是，存在一个投资回报的最优点，即曲线的拐点，为 1488 或总体的 12.6%。

## MineSet 的进一步探索

参见 *MineSet 3.0 Enterprise Edition User's Guide*， *MineSet 3.0 Enterprise Edition Reference Guide*，和 *《针对 Windows 的 MineSet 3.0 企业版教程》*中对这些工具的描述及分析数据挖掘运算法则 中的内容。可以选择 “帮助”> “MineSet 用户指南”来运行联机手册。

该教程只是 MineSet 工具包的简单介绍。其他有关方面的内容可以在 *《MineSet 用户指南》*中找到：

- 散点可视化工具。
- 分层的树可视化工具。
- 选项树导入工具和分类器。
- 关联规则生成器和可视化工具。
- 回归，允许预测连续值而不是离散值。
- 转换，包括分组、分布和数组的索引。
- 记录加权，允许向不同的记录分配不同的权重，因为这些记录比其他记录更加重要（例如：可高度获利的用户）。
- 学习曲线，可以帮助您确定对数据集的采样是否可以结束，在不过多降低导入分类器准确性的前提下加速知识发现过程。
- 许多工具选项，包括颜色处理、消息框。
- 可视工具的动画滑动条。
- 批处理。程序 mineset_batch 可用于相互无影响地执行操作。当作业需要定期 （例如：每晚一次）运行时，这是很有用的。
- 使用高级技术的误差估计，例如交叉验证。

在 *《MineSet 用户指南》*中还描述了文件和数据处理的操作细节。

**注意：** 数据挖掘运算法则会发现一些非因果的关联。一个著名的发现就是鞋子尺寸与阅读能力之间的强烈相关性：鞋子尺寸越大，阅读能力就越好。这个关系虽然是真的，但却不是因果关联的；鞋子尺寸与阅读能力都会随着年龄而增加（随着儿童不断长大，他们的鞋子尺寸和阅读能力都会增加。）警告您不要把发现的关系归为因果关系，穿大号的鞋子不可能提高您的阅读能力。

# 在 MineSet 可视化工具中漫游

## 在树可视化工具中导航

“树可视化工具”画面就好像是正在通过一个照相机观看场景。要改变视图，可以改变照相机的位置 （视点）。本部分包括两个表，可作为 “树”、“决策树”、“选项树”和“回归树”可视化工具控制的快速参考。表 A-1 描述了导航按钮。

**表 A-1** “树可视化工具”中的导航图标

| 图标 | 操作 |
|---|---|
|  | 将图返回到为原始视图设计的大小和位置。默认情况下，该图的大小和位置是第一次调用可视化工具时所用的。您可以使用下一个图标改变原始位置。 |
|  | 为图设置新的原始视图。使用它来保存特定的视图和位置。 |
|  | 将图移到中央位置，使整个图在窗口中都能被看到。 |
|  | 撤消上一次移动 （类似网页浏览器上的 “向后”按钮）。 |
|  | 重复已被撤消的移动 （类似网页浏览器上的 “向前”按钮） |
|  | 将一个节点移向树根。 |
|  | 将一个节点或条移向左边。 |
|  | 将一个节点或条移向右边。 |
|  | 将一个节点沿左边路径向树的下方移动。 |
|  | 将一个节点沿右边路径向树的下方移动。 |
|  | 弹出一个菜单，其中包括发自当前节点的可能路径。 |

表 A-2 列出的几个操作可以用于树可视化工具中的场景。大多数操作都可以用可视化工具上的一个控制或鼠标动作来完成。

**表 A-2** 操作 “树可视化工具”场景

| 操作 | 滑动条或滑轮 | 对应的鼠标操作 |
|---|---|---|
| 在场景表面掠过 | N / A | 同时按住左右鼠标按钮 （或鼠标中间的按钮）并移动鼠标。 |
| 提高或降低条形高度以强调差别 | 高度滑动条 （左上） | N / A |
| 上 / 下移动视点 | 水平滑轮 | 按住鼠标右键，然后上 / 下移动鼠标。 |
| 从左 到 右移动视点 | 左 到 右滑轮 (<-->) | 同时按住左右鼠标按钮 （或鼠标中间的按钮）并从左到右移动鼠标。 |
| 前 / 后移动视点 | 伸缩滑轮 | 同时按住左右鼠标按钮 （或鼠标中间的按钮）并上 / 下移动鼠标。 |
| 改变照相机的上 / 下倾斜角度 | 倾斜滑轮 | N / A |
| 沿着您指的方向移动 | N / A | 同时按住 Alt 键和左右鼠标按钮 （或鼠标中间的按钮）并移动鼠标。当向前移动时，视点也会按照当前的倾斜角度向下移动。类似地，当向后移动时，视点会按照当前的倾斜角度向上移动。 |
| 选择一个节点的子节点 | N / A | 在父节点上按住 Ctrl 键并单击鼠标右键，然后在子节点上单击，移动到子节点 （或使用展开导航图标）。 |

## 在非树可视化工具中导航

本部分包括两个表，可作为 “证据”、“地图”、“散点” 和 “平伸” 可视化工具控件的快速参考。表 A-3 描述了导航按钮。

**表 A-3** “非树可视化工具” 中的导航按钮

| 按钮 | 名称 | 操作 |
|---|---|---|
|  | 选取 | 改变程序到选取模式 （箭头）。在选取模式中，您可以突出显示 （刷过）或选择 （单击）图中的元素。 |
|  | 抓取 | 改变程序到抓取模式 （手形）。在抓取模式中，您可以在窗口中移动图：<br>— 要在窗口中移动图，可以按住鼠标右键并移动鼠标。<br>— 要旋转图，可以按住鼠标左键并移动鼠标。<br>— 要对图进行伸缩操作，可以同时按住鼠标左右按钮 （或使用鼠标中间的按钮）并移动鼠标。 |
|  | 首页 | 将图返回到为原始视图设计的大小和位置。默认情况下，该图的大小和位置是第一次调用可视化工具时所用的。您可以使用设置首页图标来改变首页位置。 |
|  | 设置首页 | 为图设置新的原始视图。当您希望保存某个视图或位置时，可以使用它。 |
|  | 查看全部 | 将图移到中央位置，使整个图在窗口中都能被看到。 |
|  | 缩放 | 将选择的移动到面板中间，并缩放它。当鼠标光标成为瞄准器形状时，将它移动到希望看得更清楚的点，然后单击鼠标左键。 |
|  | 三维 | 切换到三维视角。 |
|  | 顶视图 | 将图改变为顶视图 （仅适用于 “散点” 和 “平伸” 可视化工具）。 |
|  | 前视图 | 将图改变为前视图 （仅适用于 “散点” 和 “平伸” 可视化工具）。 |
|  | 侧视图 | 将图改变为侧视图 （仅适用于 “散点” 和 “平伸” 可视化工具）。 |

表 A-4 描述了非树可视化工具中的调整滑动条和滑轮。

**表 A-4** 操作 “非树可视化工具”场景

| 操作 | 滑动条或滑轮 | 鼠标或键盘操作 |
|---|---|---|
| 在 “选择”和 “抓取”模式间切换 | N/A | 按下 Esc 键或导航按钮。 |
| 移动场景 | N/A | 单击并按住鼠标右键。将光标沿着您希望图移动的方向移动。 |
| 提高或降低饼形、圆饼形或条形的高度以突出差别 | 高度滑动条 （左上） | N/A |
| 绕坐标轴 X 旋转场景 | Rotx 滑轮 | 单击并按住鼠标左键。将光标沿着您希望图旋转的方向移动。 |
| 绕坐标轴 Y 旋转场景 | Roty 滑轮 | 单击并按住鼠标左键。将光标沿着您希望图旋转的方向移动。 |
| 将场景放大和缩小 | 伸缩滑轮 | 单击并同时按住左右鼠标按钮 （或鼠标中间的按钮）。向下移动鼠标放大场景，向上移动鼠标缩小场景。 |
| 筛选出不太重要的属性 | 细节滑动条 ( 仅适用于 “证据”和 “决策表”可视化工具 ) | N/A |
| 筛选出记录权重小于指定占数据集中全部记录权重百分比的属性值，最大为 2% | % 权重阈值滑动条 ( 仅适用于 “证据”和 “决策表”可视化工具 ) | N/A |

**表 A-4** 操作 “非树可视化工具”场景

| 操作 | 滑动条或滑轮 | 鼠标或键盘操作 |
| --- | --- | --- |
| 在细节等级中下寻 （仅适用于 “表”和 “地图”可视化工具） | N / A | 将鼠标箭头放在指定的图 （或所有图的背景）上，然后单击鼠标右键。 |
| 在细节等级中上寻 （仅适用于 “表”和 “地图”可视化工具 | N / A | 将鼠标箭头放在指定的图 （或所有图的背景）上，然后按住 Ctrl 键并单击鼠标右键 （或单击鼠标中间的按钮）。 |