

MineSet for Windows™ 3.0

企业版用户指南®

文档编号 007-4005-001CHS

制作群体

撰稿: Sandra Motroni 和 Helen Vanderberg

插图: Dany Galgiani

制作: Linda Rae Sande

工程: Barry Becker、Amit Bleiweiss、Jeff Brainerd、Cliff Brunk、Eben Haber、Ara Jerahian、Andy Kar、Ed Karrels、Eser Kandogan、Alex Kozlov、Alan Norton、Peter Rathmann、Mario Schkolnick、Dan Sommerfield、Peter Welch 和 Brett Zane-Ulman。

© 2000, Silicon Graphics, Inc.—版权所有

未经 Silicon Graphics, Inc. 书面许可, 不得以任何方式复制或抄袭本文档的全部或部分内容。

有限的和限制性权利说明

政府部门对于该产品的使用、复制或公开受到以下条款的限制: FAR 52.227-14 中的“数据权利”条款和 / 或类似条款, 或 FAR、DOD、DOE 或 NASA FAR 附录中的后续条款。非出版发行的权利依据美国的版权法保留。合约商 / 制造商是 Silicon Graphics, Inc., 1600 Amphitheatre Pkwy, Mountain View, CA 94043-1351。

Silicon Graphics 是注册商标, SGI、MineSet 和 Silicon Graphics 徽标都是 Silicon Graphics, Inc. 的商标。Oracle 是 Oracle 公司的注册商标。Excel、Windows 和 Windows NT 都是 MicroSoft 公司的注册商标。MATLAB 是 The Matchworks, Inc. 的商标。SPSS 是 SPSS, Inc. 的注册商标。DBMS/COPY 是 Conceptual Software, Inc. 的商标。

“树可视化工具”的美国专利号为 No. 5,528,735 ; 5,555,354 ; 5,671,381 ; 和 5,861,885。“平伸可视化工具”的美国专利号为 No. 5,861,891。“地图可视化工具”、“散点可视化工具”和“平伸可视化工具”中的二维滑动条的专利正在申请之中。“证据可视化工具”、“决策表”和“散点动画”的专利正在申请之中。

MineSet for Windows™ 3.0 企业版用户指南®
文档编号 007-4005-001CHS

目录

图表清单 ix

表格清单 xiii

关于本指南 xv

本指南的适用对象 xv

查找 MineSet 信息 xv

本文档的结构 xvi

本指南中的插图 xvii

相关读物 xviii

格式约定 xviii

1. 数据挖掘和 MineSet 工具概述 1

数据挖掘技术 1

关于数据挖掘和数据挖掘方法 2

分析数据挖掘 3

 监督建模 3

 非监督建模 5

可视化数据挖掘 6

MineSet 的数据挖掘工具 7

数据挖掘过程概述 8

 识别数据 9

 准备数据 9

 建立模型 10

 评测模型 11

 配置模型 11

2. 利用 MineSet 访问数据 13

在何处可以获得 MineSet 13

MineSet 软件是如何工作的 14

- MineSet 与应用程序 15
 - 运行 MineSet 15
 - 使用 “工具管理器” 窗口 18
 - 用 “记录查看器” 查看原始记录 19
 - 改变 “记录查看器” 列 20
 - 筛选 “记录查看器” 中的数据 21
 - 保存 “记录查看器” 中的数据 22
 - 用 “统计可视化工具” 查看记录统计 22
 - 了解棒状图 22
 - 理解直方图 23
 - 运行 “统计可视化工具” 24
 - 使用样例数据文件工作 25
 - 运行 “直方图可视化工具” 25
 - 在 MineSet 三维可视化工具中漫游 27
 - 在树可视化工具中漫游 27
 - 在 “非树可视化工具” 中漫游 29
 - 获得帮助 31
- 3. 加工数据 33
 - 为什么要加工数据？ 33
 - 用 “工具管理器” 转换数据 34
 - 删除和添加列 35
 - 为列改变或创建新的分组 37
 - 通过组合创建新列 40
 - 通过筛选限制列内容 42
 - 更改列类型或名称 43
 - 应用模型 46
 - 数据采样 46
 - 用 “历史表” 回溯历史操作 47
 - 加权记录 49
 - 查找重要的列 50

- 4. **用散点和平伸可视化工具检查数据** 53
 - 散点和平伸可视化工具概述 53
 - 散点可视化工具概述 53
 - 平伸可视化工具概述 55
 - 为散点和平伸可视化工具转换数据 57
 - 平伸可视化工具的处理技术 59
 - 运行散点和平伸可视化工具 60
 - 映射散点可视化工具可视元素 60
 - 映射平伸可视化工具元素 61
 - 查看散点和平伸可视化工具中的结果 64
 - 查看模式 64
 - 为散点可视化工具创建滑动条 68
 - 在“散点”和“平伸”可视化工具中创建动画 69
 - 使用可视化工具汇总窗口演示动画 69
 - 在散点可视化工具中显示动画轨迹 70
 - 在平伸可视化工具中演示动画 72
 - 平伸可视化工具中汇总窗口的说明 73
 - 处理散点和平伸可视化结果 75
 - 更改画面 75
 - 在散点可视化工具中选择和追溯 77
 - 用“形状”菜单更改散点可视化工具画面 79
 - 用“形状”菜单更改平伸可视化工具画面 80
- 5. **用树可视化工具检查数据** 81
 - 树可视化工具概述 81
 - 运行树可视化工具 82
 - 映射树可视化工具可视元素 84
 - 用树可视化工具查看结果 86
 - 树可视化工具示例 87
 - 在树可视化工具中更仔细地观察数据 90
 - 用全局观察窗口观察整个图片 92
 - 用“搜索”面板查找指定的对象 93
 - 用“标号”面板标记重要的位置 94
 - 用“筛选”面板筛选数据 97
 - 调整树可视化工具画面 97

- 6. 用地图可视化工具检查数据 99
 - 地图可视化工具概述 100
 - 运行地图可视化工具 102
 - 准备数据 103
 - 组合数据 103
 - 选择地图形状 104
 - 链接地图可视化工具元素 104
 - 查看地图可视化工具 108
 - 查看模式 109
 - 在地图可视化工具中创建动画 110
 - 处理地图可视化工具结果 110
 - 更改地图可视化工具画面 110
 - 选项和追溯 111
- 7. 理解预测建模 113
 - 预测建模概述 113
 - 生成模型 114
 - 证据模型 114
 - 决策树模型 116
 - 选项树模型 117
 - 决策表模型 118
 - 回归树模型 120
 - 评测预测模型 121
 - 使用所有数据进行分类 122
 - 预留误差估计 124
 - 交叉验证误差估计 125
 - 创建学习曲线 126
 - 应用预测模型 129
 - 选择模型 129
 - 应用模型 130
 - 进入后续内容 131

8.	用决策、选项和回归树建模和预测	133
	决策、选项和回归树概述	134
	决策树	134
	选项树	135
	回归树	138
	运行决策、选项和回归树	139
	用决策树可视化工具查看结果	140
	使用决策树主窗口对记录分类	141
	其它有用的选项	143
	用选项树可视化工具查看结果	143
	用回归树可视化工具查看结果	144
	使用回归树主窗口预测数值	145
	回归树的误差估计	145
	其它有用的选项	145
	用决策、选项和回归树预测	146
9.	用决策表分类器和可视化工具建模和预测	147
	决策表分类器概述	147
	运行决策表	149
	用决策表可视化工具查看结果	151
	查看“决策表”窗格	152
	“标签概率”窗格	153
	“决策表”示例	153
	用决策表预测	156
10.	用证据分类器和可视化工具建模和预测	157
	证据分类器和可视化工具概述	157
	“证据可视化工具”窗口	158
	运行证据工具	159
	用证据可视化工具查看结果	161
	证据视图	161
	概率视图	165
	条形视图	166
	改变证据可视化工具视图	168
	用证据分类器预测	169

- 11. 改进预测建模 171
 - 确保模型的准确性 171
 - 测试模型 171
 - 将数据拟合到模型 173
 - 在误差估计中修正 175
 - 用推进提高准确性 175
 - 用混淆矩阵和损失矩阵调整模型 176
 - 使用混淆矩阵来研究错误 176
 - 显示混淆矩阵 178
 - 定义损失矩阵 181
 - 用上升曲线和 ROI 曲线评测模型 183
 - 用上升曲线检查预测 183
 - 使用投资回报曲线寻找销售利润 185
- 12. 用聚类划分数据 189
 - 聚类概述 189
 - 用“工具管理器”启动聚类 190
 - 使用聚类样例文件工作 192
 - 理解“聚类可视化工具”主窗口 193
 - 聚类的其它可视化处理 194
- 13. 用关联规则分析数据 195
 - 关联规则生成和可视化处理概述 195
 - 关联规则生成 196
 - 规则可视化处理 197
 - 启动关联规则 199
 - 建立关联 199
 - 记录加权 200
 - 将规则属性映射到可视元素 200
 - 说明“散点可视化工具”中的关联规则 201
 - 追溯 204
 - 多路关联规则 204
 - MineSet 用户词汇表 207
 - 索引 217

图表清单

图 1-1	数据表样例	3
图 1-2	数据挖掘过程	8
图 2-1	工具管理器登录	16
图 2-2	“打开数据文件”窗口	17
图 2-3	“工具管理器”窗口	18
图 2-4	“记录查看器”画面	19
图 2-5	记录查看器筛选面板	21
图 2-6	“统计可视化工具”显示的数值列	23
图 2-7	“统计可视化工具”显示的离散列	24
图 2-8	选定“统计可视化工具”的“数据目标”窗格	25
图 2-9	带有“直方图可视化工具”的“数据目标”窗格	26
图 3-1	“数据转换”面板	34
图 3-2	“添加列”对话框	36
图 3-3	“列分组”对话框	37
图 3-4	“高级分组选项”窗格	38
图 3-5	“组合”对话框	41
图 3-6	“筛选”对话框	43
图 3-7	“类型”弹出列表	44
图 3-8	“采样”对话框	46
图 3-9	“查看历史”对话框	48
图 3-10	“列重要性”选项卡	50
图 3-11	列重要性的高级模式	51
图 4-1	散点可视化工具屏幕样例	54
图 4-2	具有一维汇总滑动条的平伸可视化工具样例	55
图 4-3	选定“散点可视化工具”的“数据目标”面板	60
图 4-4	将列映射到平伸可视化工具的可视元素	62
图 4-5	为映射添加一个整数类型的列	63

图 4-6	显示选定实体的信息	65
图 4-7	操作数据上方的选取拖动工具	67
图 4-8	散点可视化工具管状运动轨迹示例	71
图 4-9	具有汇总窗口和滑动条控制的平伸可视化工具动画控制面板	72
图 4-10	滑动条移动后改变的可视化处理（请与图 4-2 比较）	74
图 4-11	散点和平伸可视化工具的查看菜单	75
图 4-12	散点和平伸可视化工具筛选面板	76
图 4-13	散点可视化工具选项菜单	77
图 4-14	平伸可视化工具选项菜单	78
图 5-1	树可视化工具主窗口的画面示例	82
图 5-2	选定树可视化工具的工具管理器数据目标窗格	84
图 5-3	“商店”数据集的树可视化工具初始视图	88
图 5-4	突出显示对象及其基本信息	90
图 5-5	选定（被聚光的）对象的示例	92
图 5-6	树可视化工具的全局观察窗口	93
图 5-7	树可视化工具中的搜索结果示例	94
图 5-8	树可视化工具“标号面板”	95
图 5-9	“选择标号”对话框	95
图 5-10	带有标号标记的主窗口	96
图 6-1	使用地理形状的地图可视化工具样例	100
图 6-2	带有地理轮廓上相关群体条形图的地图可视化工具	101
图 6-3	显示带有指定端点的美国地图的可视化工具样例	102
图 6-4	在地图可视化工具中组合	103
图 6-5	将列映射到地图可视化工具的可视元素	105
图 6-6	“地图可视化工具选项”对话框	107
图 6-7	年份滑动条位于 1990 年的 Population.usa.mapviz 示例	108
图 7-1	“证据导入工具”为客户波动数据集生成的证据可视化处理	115

图 7-2	“决策树导入工具”为客户波动数据集生成的决策树	116
图 7-3	“选项树导入工具”为汽车数据集生成的选项树	117
图 7-4	“决策表导入工具”为蘑菇数据集生成的决策表	119
图 7-5	回归器根据成人数据集生成的回归树	120
图 7-6	蝴蝶花错误分类示例	123
图 7-7	预留的错误估计选项	124
图 7-8	交叉验证的误差估计选项	125
图 7-9	带有客户波动标签集的客户波动数据集的学习曲线	127
图 7-10	学习曲线选项	128
图 7-11	“测试和应用模型”窗口：选择分类器	130
图 7-12	应用模型面板	131
图 8-1	汽车数据集的决策树	135
图 8-2	汽车数据集的选项树	137
图 8-3	成人数据集的回归树	138
图 8-4	工具管理器数据目标窗格，分类选项卡	139
图 8-5	蝴蝶花数据集的决策树	142
图 9-1	‘蘑菇’数据集的决策表	148
图 9-2	显示分类器的工具管理器数据目标面板	149
图 9-3	蘑菇数据集的决策表可视化处理	151
图 9-4	在蘑菇数据集中细化下寻的决策表	154
图 9-5	细化下寻决策表的特写	155
图 10-1	应用到蝴蝶花数据集的证据可视化工具	158
图 10-2	显示概率的证据可视化工具	159
图 10-3	“工具管理器数据目标”窗格	160
图 10-4	用于客户波动数据集的证据可视化工具块状图	162
图 10-5	证据可视化工具中的蘑菇数据集	164
图 10-6	证据可视化工具圆饼图	166
图 10-7	汽车数据集标签值“日本”被选定的	167
图 11-1	测试模型面板	172
图 11-2	“将数据拟合到模型”面板	174

图 11-3	蘑菇数据集的混淆矩阵	177
图 11-4	分类器选项窗格显示选中的混淆矩阵	178
图 11-5	显示蘑菇数据集错误分类的混淆矩阵	180
图 11-6	“损失矩阵编辑”窗格	181
图 11-7	带有损失矩阵蘑菇数据集的混淆矩阵	182
图 11-8	上升曲线	184
图 11-9	投资回报曲线	187
图 12-1	“聚类”选项卡	190
图 12-2	使用迭代 k- 均值算法聚类	191
图 12-3	聚类可视化工具主窗口	192
图 13-1	关联规则可视化处理主窗口的详细视图	198
图 13-2	关联生成的初始工具管理器窗口	199
图 13-3	关联规则映射	201
图 13-4	在指定 brand.rules.scatterviz 样例时的初始关联规则视图	202
图 13-5	代表规则的条形图上的光标	203
图 13-6	建立多路关联规则的初始工具管理器窗口	205

表格清单

表 2-1	处理记录查看器	20
表 2-2	“树可视化工具”中的漫游图标	27
表 2-3	操纵 “树可视化工具” 场景	28
表 2-4	“非树可视化工具” 中的漫游图标	29
表 2-5	操纵 “树可视化工具” 场景	30
表 3-1	“工具管理器数据转换” 面板上的按钮功能	35
表 3-2	高级分组选项	39
表 3-3	列类型含义	45
表 3-4	历史表含义	47
表 4-1	平伸可视化工具中允许映射的列类型	58
表 4-2	平伸可视化工具中允许映射的列类型	58
表 4-3	映射散点可视化工具中的可视元素	61
表 4-4	映射平伸可视化工具中的可视元素	62
表 4-5	散点和平伸可视化工具的查看菜单选项	75
表 5-1	树可视化工具可视元素	85
表 5-2	“商店数据” 的元素映射	88
表 5-3	树可视化工具显示参数	97
表 6-1	追溯操作	109
表 6-2	地图可视化工具的查看菜单选择	110
表 6-3	地图可视化工具的选项菜单选项	111
表 7-1	对学习曲线结果的操作	129
表 10-1	操作 “证据可视化工具视图”	168
表 11-1	测试模型面板选项	173
表 11-2	将数据拟合到模型选项	174
表 12-1	聚类方法选项	190
表 13-1	关联规则组件	197
表 13-2	将关联规则映射到可视元素	201

关于本指南

*MineSet 3.0 企业版用户指南*描述了 MineSet 挖掘和可视化处理工具的特性和功能。关于 MineSet 产品的最新信息，可以在万维网上访问 <http://mineset.sgi.com>。

本指南的适用对象

您不必是数据挖掘方面的专家就可以使用本指南或 MineSet，但是对数据及其代表含义的理解会有助于您更加容易解释结果。如果您有数据挖掘技术的经验，本指南仍然可以帮助您了解 MineSet 运算法则的工作方式，以及它们是如何被应用和可视化的。

如果要使用“工具管理器”以便从数据库中将数据导入 MineSet 工具，就应该主要参考 *《MineSet 3.0 企业版接口指南》* 以执行相关任务。一旦数据被加载到各种可视化工具中，就不需要数据库或编程背景了。

查找 MineSet 信息

本指南解决有关运行 MineSet 的任务。书中的大部分内容都是在讲述如何运行各种工具。在“[本文档的结构](#)”中有逐章的摘要。

关于技术细节和更完全的讲解，请参阅 *《MineSet 3.0 企业版参考指南》*。关于特定的 MineSet 运算法则方面非常专业的信息，请参阅 http://www.sgi.com/software/mineset/mineset_data.html 中的白皮书列表。

关于以下方面的信息，如：将 MineSet 导出到其他位置或应用程序、用命令行操作或系统管理，请参阅 *《MineSet 3.0 企业版接口指南》*。

MineSet 还为第三方产品提供了插入应用程序的方法，如 AcPro。

本文档的结构

本指南的前三章将介绍数据挖掘和 MineSet 产品。以后的章节将集中介绍所示的特定工具和过程：

第 1 章，“数据挖掘和 MineSet 工具概述”

本章简短概述了数据挖掘的原理，解释了术语，并介绍了利用 MineSet 工具包进行分析和可视数据挖掘的过程。

第 2 章，“利用 MineSet 访问数据”

本章描述了如何运行 MineSet 和如何使用一些基本的工具查看数据。

第 3 章，“加工数据”

本章描述了使用“工具管理器”转换原始数据的要求和过程。讲解了研究数据未知特征的“列重要性”和“聚类”工具。

第 4 章，“用散点和平伸可视化工具检查数据”

本章介绍了“散点和平伸可视化工具”的界面。无论是静态地还是通过动画演示，这些工具对于可视化多维数据都是很有价值的。

第 5 章，“用树可视化工具检查数据”

本章描述了“树可视化工具”的界面。该工具对于可视化分层数据很有价值。

第 6 章，“用地图可视化工具检查数据”

本章描述了“地图可视化工具”的界面。该工具对于具有地理或空间背景的数据很有用。

第 7 章，“理解预测建模”

本章介绍了与描述建模相反的预测建模，以及 MineSet 工具提供的各种分类器。

第 8 章，“用决策、选项和回归树建模和预测”

本章描述了如何生成并使用“决策”、“选项”和“回归”工具。这些工具对于按照属性集合并根据这些属性进行一系列的决策来分类数据很有价值。选项树可以同时显示拆分对多个属性产生的影响。回归树对于预测基于连续值的属性（例如真实生活中的事件）很有用。

第 9 章，“用决策表分类器和可视化工具建模和预测”

本章完整地描述了“决策表”的界面。该工具对于可视化分类数据中作出的决策和创建分类器很有价值。

第 10 章，“用证据分类器和可视化工具建模和预测”

本章描述了如何生成并使用“证据分类器”。该工具对于检查指定结果在给定属性下出现的概率来分类数据很有价值。

第 11 章，“改进预测建模”

本章描述了“混淆矩阵”、“损失矩阵”以及“ROI 和上升曲线”的用途及用法，将预测建模与真实生活联系起来。

第 12 章，“用聚类划分数据”

本章描述了“聚类可视化工具”的界面。该工具对于检查数据和确定聚类模式很有价值。

第 13 章，“用关联规则分析数据”

本章描述了“关联规则可视化工具”。该工具对于挖掘大的数据集和发现其中数据的相互关系很有价值。

MineSet 用户词汇表

词汇表解释了本指南中频繁使用的技术词汇和词组。

本指南中的插图

本指南中的硬拷贝显示了屏幕画面的黑白插图。联机版本中的这些插图都是彩色的。如果发现在硬拷贝中某个图形或屏幕画面看不清，可以参考联机版本中的页面，以获得更加清晰的插图。

相关读物

关于数据挖掘及相关技术的一般信息，可以参阅以下读物：

- *Data Mining Techniques for Marketing, Sales, and Customer Support*, Michael Berry 和 Gordon Linoff 合著， John Wiley & Sons 出版。
- *Data Mining Solution: Methods and Tools for Solving Real World Problems*, Christopher Westphal 和 Teresa Braxton 合著， John Wiley & Sons 出版。

如果希望了解关于数据挖掘更加专业的信息，《*MineSet 3.0 企业版接口指南*》中的附录 A 列出了更多的读物。

格式约定

本指南使用以下的类型约定和符号：

斜体 可执行名称、文件名、程序变量、工具、实用工具、变量命令行参数、以及示例、代码和语法语句中由用户提供的变量

粗体 关键字

`固定宽度类型` 屏幕上的命令行文本和提示

粗体固定宽度类型 用户的输入，包括键盘按键（打印和非打印）；以及示例、代码和语法语句中由用户提供的文字

数据挖掘和 MineSet 工具概述

本章介绍数据挖掘的有关主题以及 MineSet 工具是如何在这种环境下工作的。如果对 MineSet 已经有所了解，可以跳过本章。有关主题包括：

- 第 1 页的 “数据挖掘技术”
- 第 2 页的 “关于数据挖掘和数据挖掘方法”
- 第 3 页的 “分析数据挖掘”
- 第 6 页的 “可视化数据挖掘”
- 第 7 页的 “MineSet 的数据挖掘工具”
- 第 8 页的 “数据挖掘过程概述”

MineSet 是一个进行数据分析的集成工具包。市场和营销专家、金融分析家、保险商以及任何需要分析数据的人都可以使用 MineSet 的图形用户界面。对于希望为商业和科学目的编写决策支持应用程序的程序开发员， MineSet 也提供了接口。

数据挖掘技术

在这里和本书最后的词汇表中概述了基本数据挖掘术语的用法。例如：在这种情况下，数据是进行一个商业或科学过程（例如：客户帐单、药物测试或销售点交易）时最初被收集的记录集合。这些记录被组织为带有行列的表，以便 MineSet 使用。可以赋予表一个描述性名称，如“用户”，代表某个特定企业所有关于用户的记录。在数据挖掘环境中，通常将列称为属性，将行称为实例或记录。

在 MineSet 使用的运算法则和数学公式中，有一些被称为导入工具，因为它们可以导入要建立的模型。模型本身被称为分类器，因为它们对数据的行和列进行分类。因此，可以获得行和列的记录内容，并以多种方式对它们进行排序或分类。

关于数据挖掘和数据挖掘方法

数据挖掘的目的是为了发现数据中的规律，从而将这种认识应用到问题解决之中。数据挖掘可以解决的典型问题包括：

- 欺诈侦测
- 客户波动分析
- 电话模式分析
- 市场目标
- 确定市场划分
- 改善操作程序
- 改善医疗服务
- 市场供求分析

当使用查询或联机分析处理（OLAP）收集数据库中的数据信息时，必须直接指定数据元素之间的所有关系。例如：要按区域查询所有销售额。这种方法预先假设您认为销售额随区域变化，因此要用数据检验这个假设以证明或否认它的有效性。

这就是 OLAP 与基于发现的数据挖掘的结合点，在数据挖掘中可以发现未知的关系。这种方法允许数据本身向调查者提出结论。就是这种发现原来未知关系的能力将数据挖掘与 OLAP 及其他方法（如查询）区分开来。

可视化数据挖掘以可视的形式展现数据，因此只要通过可视化处理就可以了解总体趋势。这种可视化和研究复杂数据模式的能力在决策中被证明是极其宝贵的。这种可视数据挖掘是描述性的，即：它对已经被测量或量化的现有数据进行描述。

分析数据挖掘使用运算法则自动从数据中建立模型。然后这些运算法则将根据选择的算法以各种方法对数据进行处理。可以将数据分析结果以可视形式显示，用未经试验的数据检验它们，或是将模型适用到全新的数据集上。分析数据挖掘使用模型预测下一部分存在数据的特征。

分析数据挖掘可以生成关于数据的模型。它用于工作的运算法则可分为两大类 - 监督和非监督建模方法。

在 MineSet 中，可视化处理与挖掘模型的结合与单独应用任何一种技术相比，能够为用户提供更加深入的认识。

分析数据挖掘

在该论述中，数据出现在由行列组成的表中。数据纪录由行代表：数据属性由表的列代表（请参阅图 1-1）。

说明性属性					标签
	年龄	婚姻状况	性别	受雇年限	信用风险
记录 1	35	D	M	3	> 60
记录 2	45	M	F	19	< 60
记录 3	29	S	M	45	> 60

图 1-1 数据表样例

监督建模

监督建模任务的目标是根据记录中其他列（属性）的值，预测同一个记录中某一列（属性）的内容或值。一般倾向于预测一指定的属性或特征。该属性叫做标签。通过对标签和其他属性之间关系的编码，可以使用监督建模运算法则建立模型，对新的、未标注的数据进行预测。

然后，就可以使模型可视化，获得对标签和其他属性之间关系的认识。例如：如果用户已经离开了公司（通常被称为缩减或客户波动），就可以建立一个模型，既可以预测哪些客户可能客户波动，又可以帮助了解导致这种行为的原因和模式。

两种最普通的监督建模任务是分类和 *回归*。如果标签是离散的（即：包含一个固定集合的值 - 例如：“是”或“否”），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：收入和股票价格），则该任务叫做回归。

分类

分类任务是将分类标签值分配给未标注的记录。在这种情况下，记录将被划分到预定义中的组中。例如：一个简单的分组可能把客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。另外的数据分类可能会按性别或收入来对客户进行分组。分类过程还可以预测标签采取指定值的概率。例如：可以计算用户在 60 天内支付帐单的概率。

在创建这些分类过程中，MineSet 利用运算法则，根据所给的数据导入（建立）模型。MineSet 可以自动从训练集中导入分类器。训练集是原始数据的随机子集。例如：可能决定只使用 40% 的数据来建立模型。模型一旦生成，就可以被用来对标签值未知的记录进行分类或预测类概率。如：当新记录添加到表中时，标签值可能是未知的。这时，模型就可以预测标签列的值。如果后来发现标签值与预测的相同，就说模型是正确的。原始数据集中没有被用于训练的部分可以用来检验导入模型的准确性。

MineSet 可以导入四种分类模型，每一种都可以用可视化工具查看。根据运算法则在排序和分类数据时作出决策的方式，这些模型是不同的。例如：决策树可能根据用户交谈是否超过 256 分钟将电话记录分为两个组。而选项树可能提供三种方式以搜索可能波动的用户。关于这些区别的详细信息，请参阅第 7 页的“MineSet 的数据挖掘工具”，有关示例将在以后的章节中进行介绍。

回归

除了标签取值是连续而不是离散的以外，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，尽管预测薪水是否在给定的范围内（例如：少于 \$60K，\$60K 到 \$120K 之间，多于 \$120K），或股票价格将会上升还是下降是一个分类任务。

MineSet 还可以从训练集或数据样本中自动导入回归器。回归器一旦生成，就可以被用来预测问题中连续属性的值，例如：薪水为 \$60,420。MineSet 当前只有一个回归模型：回归树。和 MineSet 建立的所有其他模型一样，回归树模型也可以用 MineSet 可视化处理进行可视化。

评测模型的准确性

估计监督模型的准确性是数据挖掘过程中的重要部分。通常按照分类器的错误率或分类错误比例对其进行评测。而回归器评测则是根据它们预测误差中标准偏差的绝对误差进行的。当评测模型的准确性时，用非建立模型所用的数据对它进行测试是非常重要的。

非监督建模

非监督建模发现行为相似的数据段及其规则。既没有正确答案的任何理念，也没有明确公认的性能量度。非监督建模展示相似的模式和段，从整体的角度提供对数据的认识。模型无法直接用于进行预测；因此无需将部分数据保留下来作为建立分类器的训练集。MineSet 为两种最普通的非监督建模提供了运算法则：关联和聚类。

关联

在生成关联时，主要任务是描述 A 是否凭借某种规则暗含着 B。经典的关联分组是市场供求分析，用以预测特定物品被同时购买的频率。例如：购买婴儿商品暗示着该客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群集组。例如：健康保险公司可能会发现以下这些特征可以定义一个段：20 到 45 岁，技术员，孩子少于两个，科幻电视迷，并且可支配收入为每年 5,000 到 10,000 美元。然后，通过在新的科幻电视中插播电视广告，就可以利用对这些人非常适合的健康保险节目更加有效地针对他们。

可视化数据挖掘

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。以下的 MineSet 可视化工具为工作示例：

- 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。额外的属性可以被映射为颜色和大小。还有两个附加的属性可以被映射到滑动条，允许动画和闪现，第八个属性可以被映射为颜色代码，通过潜在有趣的动画变量值组合引导动画。MineSet 中的列重要性操作可以帮助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变化的不透明性显示。得到的结果与单独处理每个数据点所得的结果近似，并且对于数据点太多以至于无法在散点可视化工具中显示的数据集尤其有用。
- 树可视化工具 - 数据被映射到节点上以便观察数据的分层分解。决策树、选项树和回归树都将数据显示在各种分支的树状可视化处理中。

MineSet 的数据挖掘工具

如果有需要分类、回归和聚类的数据集，可以使用以下这些 MineSet 工具：

- 决策树导入和可视化工具 - 导入一个产生分支格式决策树的可视化处理的分类器。
- 选项树导入和可视化工具 - 导入一个类似于决策树导入和分类工具的分类器。但是，它还建立另外的树并在分类过程中均衡它们，通常会提高准确性。
- 证据导入和可视化工具 - 创建其自己的分类器并进行可视化处理以显示基于所提供数据的证据。
- 决策表导入和可视化工具 - 创建一个分层的可视化处理，显示每一级上的维数对。在保留上下文的同时，可以迅速地概化上寻获得概览，细化下寻查看细节。
- 回归树导入和可视化工具 - 导入一个预测实值属性的回归器，即：结果带有不同等级的值，而不是带有预先确定的区间。可以使用树可视化工具查看回归器的结构。
- 关联规则 - 对控制关联（其中 A 暗示着 B）的规则进行编码，通常称为市场供求分析。使用散点可视化工具显示规则，通常用条形和盘形。
- 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。
- 列重要性分析 - 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 中还包含额外的工具来辅助知识发现过程：

- 统计可视化工具 - 以棒状图和直方图形式显示数据，每一列一个图。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 - 以直方图形式显示数据，每一列数据一个直方图，如果必要，可对连续列进行分组。
- 记录查看器 - 将原始数据显示为电子表格。

数据挖掘过程概述

本部分介绍有关知识发现过程的特定任务。该过程是一个反复迭代的过程，一旦发现了新的模式并提高了对数据的认识以后，通常就又回到早期的阶段，如[图 1-2](#)所示。



图 1-2 数据挖掘过程

该过程包括以下步骤：

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 9 页](#)的“准备数据”。
3. 建立模型 — 请参阅[第 10 页](#)的“建立模型”。
4. 评测模型 — 请参阅[第 11 页](#)的“评测模型”。
5. 配置模型 — 请参阅[第 11 页](#)的“配置模型”。

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：关于客户行为的可预测性通常是一个必要目标。在重新定义问题时，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偏僻的位置，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、MATLAB 等等。）一些工具（如：Conceptual Software, Inc. 的 DBMS/COPY）可以将数据从 100 多种格式中转换为 MineSet 可以使用的格式，并且 MineSet 的导入数据能力（在“工具管理器文件”菜单中）也可以进行类似的转换（关于详细信息，请参阅 *MineSet 3.0 企业版参考指南* 和 *《MineSet 3.0 企业版接口指南》*）。

准备数据

在加载到 MineSet 之前，数据可能需要修改（该步骤叫做清洗。）特别地，常见问题包括：

- 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是错误的，可能不完整，也可能取值有误。
- 字段描述可能不清楚，或者可能对于不同的来源有不同的含义。例如：订购日期可能指的是订单被发送、盖戳、收到或输入的日期。
- 数据可能会过期；例如：客户可能会搬家、变换成员或改变消费模式。

即使是清楚易懂的数据也需要在应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：如果正在分析电话公司的数据，可能会发现：相对单独给定的任何一个元素，长途电话的费率（按通话使用的分钟总数划分的销售额）是一个更好的用户行为预测器。数据转换是开发合理模型的核心部分。随着不断地前进，您甚至可能会返回来以另外的方式转换数据。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新的字段。
- 删除无关、冗余、或是包含明显无用预测器的列。
- 根据列值的布尔表达式筛选数据，以影响模型或可视化处理过程。例如：可能希望只看到最重要的规则或最有利可图的客户段。
- 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）。
- 应用一个先前创建的分类器、回归器或聚类模型，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。插件模块可用于创建转换或特定种类模型。

建立模型

知识发现过程的核心是模型的建立，该任务由分析型数据挖掘运算法则自动完成。可以使用所有的数据建立模型，也可以保留部分数据用来测试模型的准确性。建立模型可以提供多个导入模型的方式。这些选择不仅会影响可视化处理显示的方式，而且还会影响运算法则在建立分类器时的决策。在[第 7 章，“理解预测建模”](#)中将进一步讨论这些选项。

评测模型

评测模型的准确性可以进一步提炼对该模型及其用法的认识。一些模型，特别是“决策树”分类器和“选项树”分类器，可以直接通过可视化处理评测模型的不同部分并显示它们。

MineSet 提供了四种模型评测方法：误差估计、混淆矩阵、上升曲线和 ROI（投资回报）曲线。在[第 11 章，“改进预测建模”](#)中将对这些方法进行讲解。

配置模型

可以通过将模型应用到新数据上来配置它。新数据可能会产生新的问题，从而需要进一步的改进。

在《*MineSet 3.0 企业版教程*》的电信示例中，创建了一个模型来确定哪些用户可能波动（即：离开他们的电话公司）。然后，客户记录将通过模型进行评测以识别最可能波动的特定客户。可以给这些客户一定的刺激使他们留下。

可以继续了解关于客户波动数据集 -（MineSet 软件中一个准备好的电信用户数据集） - 的知识发现过程。当完成该指南中的范例时，请思考一下您的商业或科学活动是如何向前发展并反复的，从而建立起一个经过充分测试的分析数据挖掘模型。

在下一章，[第 2 章，“利用 MineSet 访问数据”](#)，您就可以利用 MineSet 及提供的样例文件直接开始工作了。

利用 MineSet 访问数据

本章描述如何开始运行 MineSet，以及如何在“工具管理器”图形用户界面中使用程序查看数据。讨论的主题包括：

- 第 13 页的“在何处可以获得 MineSet”
- 第 14 页的“MineSet 软件是如何工作的”
- 第 15 页的“MineSet 与应用程序”
- 第 15 页的“运行 MineSet”
- 第 18 页的“使用“工具管理器”窗口”
- 第 19 页的“用“记录查看器”查看原始记录”
- 第 22 页的“用“统计可视化工具”查看记录统计”
- 第 24 页的“运行“统计可视化工具””
- 第 25 页的“运行“直方图可视化工具””
- 第 27 页的“在 MineSet 三维可视化工具中漫游”
- 第 31 页的“获得帮助”

另见《*MineSet 3.0 企业版教程*》以练习使用各种工具。当希望使用自己数据库中的数据时，请参阅《*MineSet 3.0 企业版接口指南*》中的“设置 MineSet”。

在何处可以获得 MineSet

可以从 CD 中或从 Internet 上安装 MineSet。一旦安装了 MineSet，通常在您自己的系统上就可以立刻开始使用它，并利用 MineSet 发行版提供的格式化数据文件（位于 MineSet 安装目录中的 *data* 目录下）进行练习。下一个部分将对 MineSet 操作进行简短讲解。如果希望立即运行，请参阅第 15 页的“运行 MineSet”。

MineSet 软件是如何工作的

MineSet 是以客户机 / 服务器的方式进行工作的，服务器进程可以与客户机运行在同一个系统中，也可以在另一个通常是更强大的系统中。服务器必须安装在客户机可以访问的地方。“数据移动器”是一个运行在服务器上的进程，用于访问数据库文件、执行数据转换、运行挖掘操作并生成可视化处理文件。这些可视化处理文件随后将被传送到客户机。

“工具管理器”是一个客户机进程，可以通过它提供的图形用户界面（GUI）完成与 MineSet 的大部分交互任务。这些交互任务是本章的主要部分。可以使用“工具管理器”指定数据源、应用于数据的转换集、使用的挖掘和可视化处理工具、以及如何保存工作的结果。

一旦完成这些规范，就可以指挥 MineSet 访问“工具管理器”中指定的所需数据，并执行其中指定的操作。然后，“工具管理器”将该信息发送到服务器“数据移动器”进程。如果要被分析的数据存在于单一结构文件中，MineSet 将读取一个配置（.schema）文件，其中包含表的图元信息以及真正数据所在的 .data 文件的名称。如果要分析的数据在 Oracle、Sybase 或 Informix_ 数据库中，“数据移动器”将对数据库执行查询以获取需要的数据集。

.schema 文件描述了输入数据文件的格式。在《*MineSet 3.0 企业版接口指南*》中对这些文件进行了详细说明。

注意：在使用 MineSet 工具之前，请先按照 MineSet 发行版注意事项中的安装和许可说明进行。系统管理员可能需要设置“数据移动器”配置文件。《*MineSet 3.0 企业版接口指南*》中介绍了详细的设置信息。

MineSet 与应用程序

MineSet 具有一个应用程序接口（API），因此可以将 MineSet 的可视化工具和数据挖掘引擎集成到应用程序中。关于详细信息，请参阅 [《MineSet 3.0 企业版接口指南》](#)，或访问 MineSet 网站。

运行 MineSet

运行 MineSet 的最简单方法是在桌面上创建 *mineset.exe* 快捷方式。要创建快捷方式，可以从 MineSet 的安装目录中，进入 *\bin* 目录并单击 *mineset.exe* 文件。从“文件”菜单中选择“创建快捷方式”。

也可以从桌面上的“开始”>“程序”>“MineSet 3.0 企业版”菜单进入 MineSet。可以用鼠标左键选择运行它，或者按住鼠标右键将其拖到桌面上，然后再松开鼠标右键。

1. 单击 MineSet 快捷方式，将出现“MineSet 工具管理器”。
2. 如果默认情况下，“登录到服务器”对话框 (图 2-1) 没有出现，可以选择“文件”>“连接到服务器”。如果希望将当前系统同时作为客户机和服务器来使用，可以在生成的对话框中，单击“将这台机器作为当前用户”。如果希望使用另一个系统作为服务器，请键入服务器名称、登录名称和口令（如果有的话）。

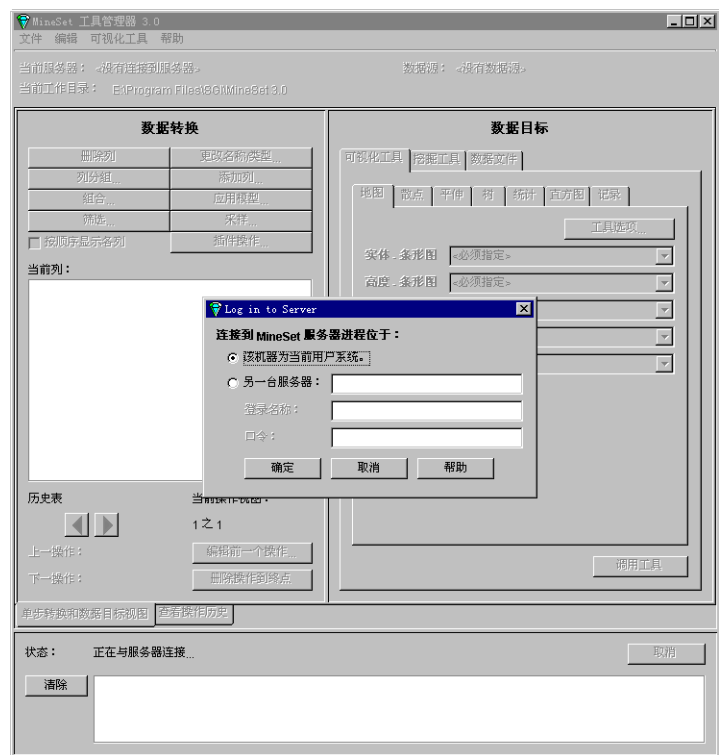


图 2-1 工具管理器登录

3. 在“工具管理器”窗口中，选择“文件”>“打开新的数据文件”。在出现的“打开数据文件”窗口中，输入使用数据的路径。如果希望用 MineSet 附带的样例文件进行练习，可以在 *MineSet 3.0\data* 目录下找到它们，它位于 MineSet 的初始安装位置。
- 要使用样例文件工作，可在“打开数据文件”窗口中选择 *churn.schema*。该文件是 MineSet 提供的一个数据样表。右边“预览列”面板中的项目是数据列，如图 2-2 所示。单击确定。

churn.schema 文件赋予访问电信用户数据集的权利。下一次运行 MineSet 时，系统将自动把您带到最后一次 MineSet 被激活时所处的状态，并且您所做的一切选项选择都将被恢复。

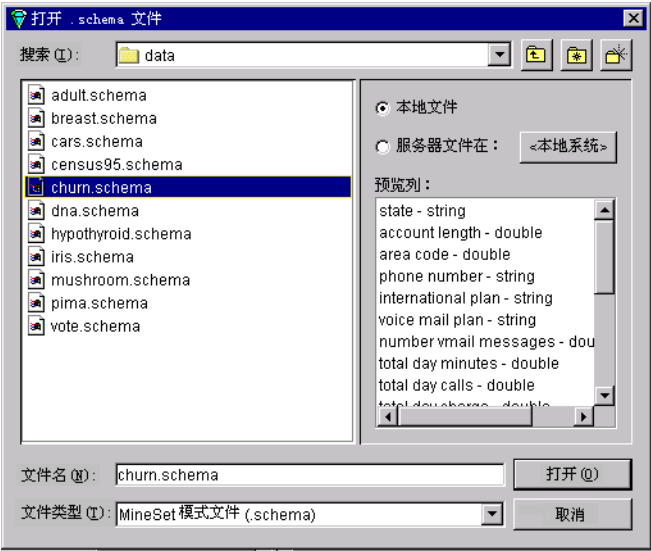


图 2-2 “打开数据文件”窗口

《MineSet 3.0 企业版参考指南》中包括一个可用来练习的样例文件列表；关于每个文件的内容，请参阅“配置和数据文件样例”。

使用 “工具管理器” 窗口

“工具管理器” 是执行所有操作的主要起点，因此对该部分进行一个快速浏览是很有帮助的。

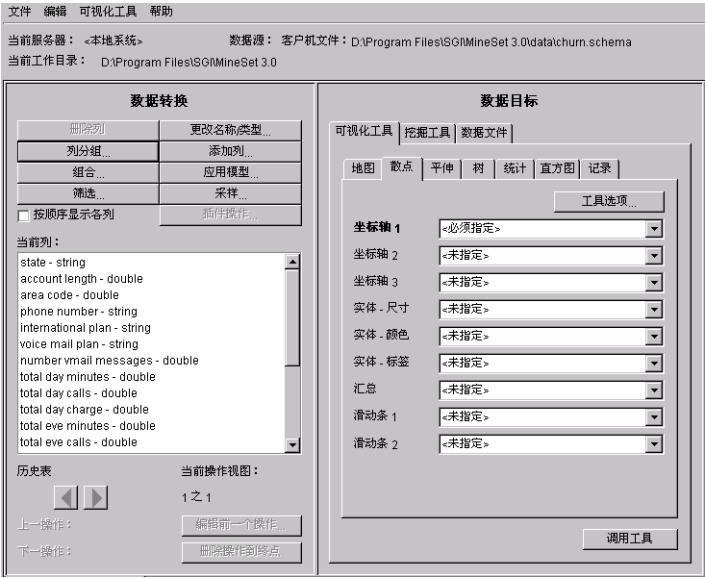


图 2-3 “工具管理器” 窗口

左边的“数据转换”面板是转换所有数据的地方。操作中使用顶部按钮进行的操作会影响以下的“当前列”。可以使用下部的按钮在转换的历史前后浏览；底部的“状态”窗口显示当前数据集进行的操作。可以使用右边的“数据目标”面板在“可视化工具”和“挖掘工具”之间切换，或将当前状态保存在“数据文件”中。以下将详细讲述其中的每个功能。

用“记录查看器”查看原始记录

熟悉选定的数据集有一个简单方法，就是使用“记录查看器”查看数据库记录和列中的数据值。记录将以电子表格的形式出现。

打开“工具管理器”，如“运行 MineSet”所示，并执行以下步骤：

- 1. 在“工具管理器数据目标”面板中，单击“可视化工具”选项卡。
- 2. 在上面一行选项卡中，单击“记录”以进入“记录查看器”。
- 3. 单击面板右下角的调用工具，将出现“记录查看器”（图 2-4）。

文件 查看							帮助
							5,000 行, 21 列
行 #	state	account length	area code	phone number	international plan	voice mail plan	numb
1	AR	116	510	409-5519	no	no	
2	WI	48	510	419-5480	no	no	
3	NE	75	408	343-1965	yes	no	
4	NC	85	510	404-2871	no	no	
5	NH	178	510	373-2387	no	no	
6	OH	43	510	342-5249	no	yes	
7	WI	90	415	420-8308	no	no	
8	DE	125	408	359-9794	no	no	
9	IL	53	415	402-7954	no	no	
10	NV	111	415	396-8198	no	yes	
11	IN	94	408	402-1251	no	no	
12	DE	129	510	332-6181	no	no	
13	MT	119	510	374-5301	no	no	
14	TN	25	415	337-3699	no	no	
15	VT	80	415	342-7514	no	no	
16	VA	115	415	367-3971	no	no	

图 2-4 “记录查看器”画面

如果希望以不同的方式查看结果，“记录查看器”可以进行各种排序、重编号以及调整大小（请参阅下面的表 2-1。）

改变 “记录查看器” 列

可以使用一些方法改变 “记录查看器” 的外观，如表 2-1 所示：

表 2-1 处理记录查看器

目的:	方法:
调整列大小	单击右边的列分界线，并向需要的方向拖动。
重排列	单击并将标题单元拖动到需要的位置。
隐藏列	在 “视图” 菜单中选择隐藏 / 显示列，取消对要隐藏的列的选择。
按列值排列记录顺序	单击列标题单元。
按多列排列记录顺序	按住 Ctrl 键并单击每个希望用于排序的列标题，单击的顺序就是希望排列它们的顺序。
将排序反转	再次单击列标题单元。
返回到原始顺序	单击行 # 列标题。
对行进行重编号	从 “视图” 菜单中选择 “重编号行”（无法撤消该功能）。
搜索一个数值	从 “视图” 下拉菜单中选择 “搜索” 窗格，在 “查找” 字段中输入一个值，然后突出显示希望搜索的列，单击 <i>查找下一个</i> 或 <i>查找上一个</i> 。要选择多列，可以使用 Shift- 单击；若要选择非连续的列，可以使用 Ctrl- 单击。

筛选“记录查看器”中的数据

可以对数据进行筛选，这样就可以只看到一定范围内的值。

1. 从“记录查看器”窗口中，选择“查看” > “筛选工具”窗格。
2. 选择一个列或表达式，从中建立筛选表达式。例如：在汽车数据集中，可以建立表达式 ‘cubicinches’ >400 来筛选那些容量大于 400 ci 的记录。（请参阅图 2-5）。
可以选择左边的元素并单击向右的箭头来建立表达式；或是简单地将表达式直接键入到文本字段中。
3. 单击应用，将立即开始筛选。窗格将一直保留，直到关闭它。

要删除筛选，可以清除表达式窗口中的表达式，然后单击应用，或从“记录查看器”窗口中选择“查看” > “删除筛选器”。

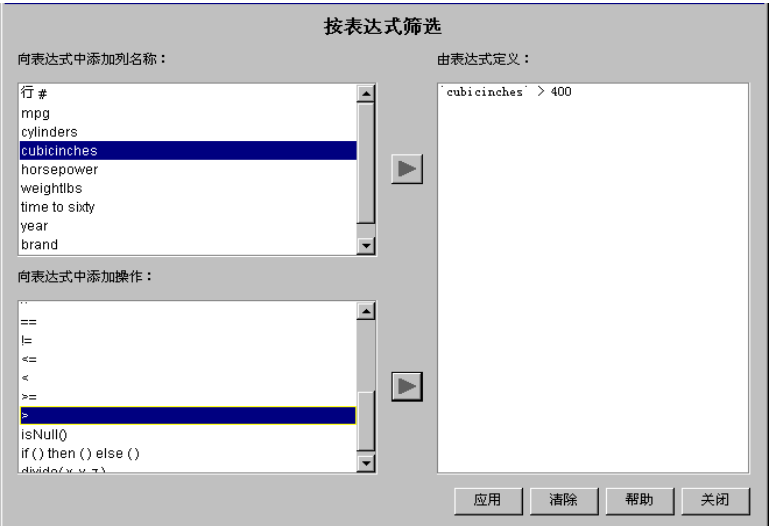


图 2-5 记录查看器筛选面板

可以打开任意多个筛选面板。要一次应用多个筛选器，应该首先应用一个，对行进行重新编号，然后再应用下一个。

注意：对行重新编号不能被撤消。要返回到原始数据，必须重新打开文件。

保存“记录查看器”中的数据

可以保存“记录查看器”中的文件，使用“文件” > “保存”或“文件” > “另存为”。

利用“另存为”，可以将数据保存为以下四种格式：二进制文件、ASCII 文件、HTML 文件或文本文件。当以二进制或 ASCII 格式保存时，将保存数据文件（表）和模式文件（向“工具管理器”和“数据移动器”描述数据文件的内容）。HTML 格式将文件保存为 HTML 表。文本格式将文件保存为制表符分隔的形式，其中第一行为列标题。

用“统计可视化工具”查看记录统计

用“统计可视化工具”可以发现数据集中记录的详细信息。可以计算某些基于数据集中记录数的统计。根据列中的数据类型是数值型或离散型，分别把统计结果显示为棒状图和直方图。

了解棒状图

棒状图将列中数值的最小值、最大值、平均值、中值以及两个四分位数（25% 和 75%），显示为穿过垂直彩色条形图的线。数据集总体标准偏差以 +/- 值表示。当不一致的值少于 50,000 个时，将显示四分位数（请参阅图 2-6）。如果在列中不同的值多于 50,000 个时，统计结果将显示为灰色垂直条形图。关于术语平均值、中值和标准偏差，请参考词汇表。

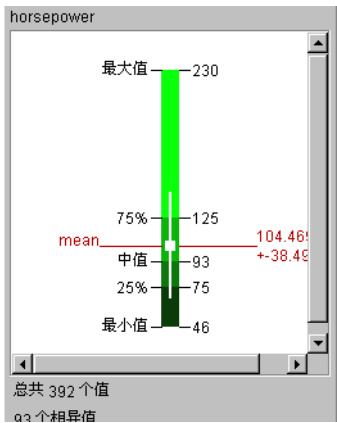


图 2-6 “统计可视化工具”显示的数值列

理解直方图

直方图将显示非数值数据（如字符串或分组）列的结果。（请参阅图 2-7。）这表示数据表中的列可以包括字符串（如“是”或“spore_color，”）或分组的值（如“10-70”）。可以有多达 100 个不同的项目。默认情况下，字符串标称所取的属性值是按照计数降序排列的，但是可以使用“查看”下拉菜单来选择另一种排序方法。字符串的两种排序方式分别为：按照计数排序（或当选定了权重时按照权重排序），或者按照名称的字母顺序排序。如果不同的类别少于等于 100 个，列面板还将包含不同值的计数。如果值已经被分组，那么分组的次序与排序方法无关。

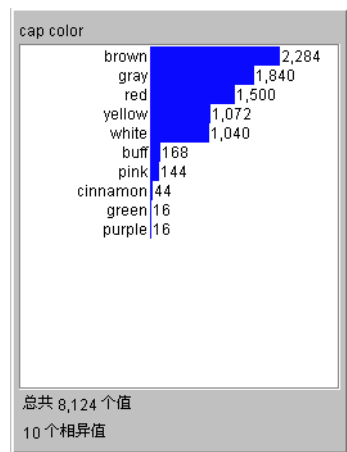


图 2-7 “统计可视化工具” 显示的离散列

创建了数据的可视化处理以后，当将鼠标在一个字段上掠过而不单击时，可以在直方图中看到被缩截的原文信息。

运行 “统计可视化工具”

在选择了一个数据集后，运行 “统计可视化工具” 最简单的方法就是：

- 1. 在 “工具管理器数据目标” 面板中，单击 “可视化工具” 选项卡。
- 2. 在选项卡的下面一行中，单击 “统计” 以进入 “统计可视化工具” (图 2-8)。
- 3. 单击面板右下角的 调用工具 。

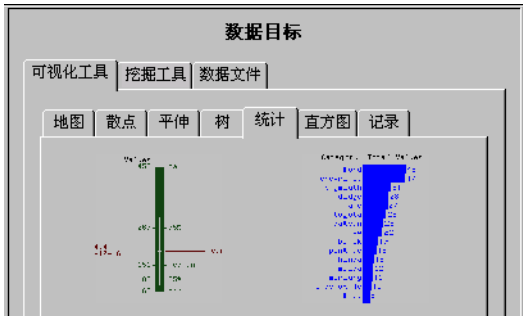


图 2-8 选定 “统计可视化工具” 的 “数据目标” 窗格

关于运行工具的其他方法，请参阅 《MineSet 3.0 企业版参考指南》中的 “统计可视化工具”。

使用样例数据文件工作

可以使用一些样例数据文件进行工作，以熟悉 “统计可视化工具” 的特性和功能。这些文件在 MineSet 3.0 下的 \examples 目录中，位于 MineSet 的初始安装目录。

从 “工具管理器” 菜单条上的 “可视工具” 运行 “统计可视化工具”，然后使用 “文件” > “打开” 下拉菜单 > 打开所有 .statviz 文件。

运行 “直方图可视化工具”

“直方图可视化工具” 自动将所有包括连续 （数值型）值的列分组，并将结果发送到 “统计可视化工具” 进行显示。关于分组，请参阅词汇表。

运行 “直方图可视化工具” 的最简单的方法是：

- 1. 在 “工具管理器数据目标” 面板中，单击 “可视化工具” 选项卡。
- 2. 从工具选择卡中，选择 “直方图”（图 2-9）。

3. 可以设置以下选项或直接转到第 4 步：
- 选取分组数，或允许 MineSet 进行选择。

■ 设置修剪因子，它指示分组产生前被剔除极值的分数。默认值为 0.05。这将排除 5% 带有极值的实例（低端 2.5%，高端 2.5%）。修剪会减小例外对阈值生成的影响。例外是位于数据主体外部的单独实例。



图 2-9 带有“直方图可视化工具”的“数据目标”窗格

4. 单击面板右下角的调用工具。
- “工具管理器”窗口底部的“状态”窗口将显示运算法则处理数据的进程。然后当前的数据会以直方图的形式显示，从中可以对数据回落的位置有一个快速的总体认识。请参阅第 23 页的“理解直方图”。

在 MineSet 三维可视化工具中漫游

本部分描述了各种 MineSet 可视化工具中的漫游控件。树可视化工具拥有的一组控件将在以下的“在树可视化工具中漫游”中进行介绍。非树可视化工具拥有的另一组控件，在第 29 页的“在“非树可视化工具”中漫游”中进行了介绍。

在树可视化工具中漫游

“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（观察点）。本部分包括两个表，可作为“树”、“决策树”、“选项树”和“回归树”可视化工具控件的快速参考。表 2-2 描述了漫游图标。

表 2-2 “树可视化工具”中的漫游图标

图标	操作
	将图返回到为首页视图设定的大小和位置。默认情况下，首页视图的大小和位置是第一次调用可视化工具时所用的。可以使用下一个图标改变首页位置。
	为图设置新的首页视图。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）。
	将一个节点移向根节点。
	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 2-3 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 2-3 操纵 “树可视化工具” 场景

操作	滑动条或滑轮	等同的鼠标操作
在场景表面掠过	N/A	同时按住左右鼠标按钮（或鼠标中键）并移动鼠标
提高或降低条形高度以突出差别	高度滑动条（左上）	N/A
上下移动观察点	H（水平）滑轮	按住鼠标右键，然后上下移动鼠标。
左右移动观察点	左右移动滑轮（<-->）	同时按住左右鼠标按钮（或鼠标中键）并左右移动鼠标。
前后移动观察点	Dolly（伸缩）滑轮	同时按住左右鼠标按钮（或鼠标中键）并上下移动鼠标。
改变照相机的上下倾斜角度	Tilt（倾斜）滑轮	N/A
沿指示的方向移动	N/A	同时按住 Alt 键和左右鼠标按钮（或鼠标中键）并移动鼠标。当向前移动时，观察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，观察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（或使用展开漫游图标）。

在“非树可视化工具”中漫游

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 2-4](#) 描述了漫游图标。

表 2-4 “非树可视化工具”中的漫游图标

图标	名称	操作
	选取	改变程序到选取模式（箭头）。在选取模式（也称为选择模式）中，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，可以在窗口中移动图： 一要在窗口中移动图，可以按住鼠标右键并移动鼠标。 一要旋转图，可以按住鼠标左键并移动鼠标。 一要对图进行缩放操作，可以同时按住鼠标左右键（或鼠标中键）并移动鼠标。
	主视图	将图返回到为首页视图设计的大小和位置。默认情况下，首页视图的大小和位置是第一次调用可视化工具时所用的。可以使用设置主视图图标来改变主视图位置。
	设置主视图	为图设置新的首页视图。当希望保存某个视图或位置时，可以使用它。
	全景视图	将图移到中央位置，使整个图在窗口中都能被看到。
	搜索对象	将选择的点移动到面板中间，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。
	三维透视	切换三维透视模式。
	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 2-5 描述了非树可视化工具的调整滑动条和滑轮。

表 2-5 操纵 “树可视化工具” 场景		
操作	滑动条或滑轮	鼠标或键盘操作
在 “选取” 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。将光标沿着希望图移动的方向移动。
提高或降低块、圆饼或条形图的高度以突出差别	高度滑动条 （左上）	N/A
绕坐标轴 X 旋转场景	X 旋转滑轮	在抓取模式中，单击并按住鼠标左键。将光标沿着希望图旋转的方向移动。
绕坐标轴 Y 旋转场景	Y 旋转滑轮	在抓取模式中，单击并按住鼠标左键。将光标沿着希望图旋转的方向移动。
将场景放大和缩小	Dolly （伸缩）滑轮	在抓取模式中，单击并按住左右鼠标按钮 （或鼠标中键）。向下移动鼠标放大场景，向上移动鼠标缩小场景。
在细节级别中细化下寻 （仅适用于 “决策表” 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。
在细节等级中概化上寻 （仅适用于 “决策表” 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 并单击鼠标右键 （或单击鼠标中键）。

获得帮助

除了在第 xv 页的“查找 MineSet 信息”中列出的资源以外，其他有关 MineSet 的信息资源包括：

- 从《*MineSet 3.0 企业版教程*》中可以迅速了解到各种工具的用法。
- 每个工具中都有上下文相关的帮助：在感兴趣的区域按 F1 键。
- <http://mineset.engr.sgi.com/movies>. 上的联机演示影片概述了一些工具的用法。

加工数据

本章介绍使用 MineSet 工具加工数据的概念。这些内容是创建分类器工作的预备知识，后者将在下章讨论。本章包括以下几个部分：

- 第 33 页的 “为什么要加工数据？”
- 第 34 页的 “用 “工具管理器” 转换数据”
- 第 47 页的 “用 “历史表” 回溯历史操作”
- 第 49 页的 “加权记录”
- 第 50 页的 “查找重要的列”

为什么要加工数据？

面对一个陌生的数据集，可能首先用 “记录查看器” 和 “统计可视化工具” 对其进行研究。（请参阅第 2 章，“利用 MineSet 访问数据”）本章中，将通过加工或转换原始数据继续该过程。例如：可能发现信息多得无法了解它们的意义。简化可视化处理最简单的方法就是对列进行添加、删除或合并。该操作不会改变基本数据，改变的只是当前工作的会话文件。应用的转换顺序将被称为是该 “工具管理器” 会话的历史。

“工具管理器” 是一个图形用户界面（GUI），可以使用它完成与 MineSet 组件的大部分交互任务。可以用 “工具管理器” 选择现有的数据源，转换或分析该数据，并使用任何一个 MineSet 工具可视化结果。

关于如何与 MineSet 可视化工具进行交互的详细信息，请参阅 *《MineSet 3.0 企业版参考指南》* 指定工具项。

注意： “工具管理器” 可能需要一些手工操作以使非 MineSet 创建的数据文件能相容；请参阅 *《MineSet 3.0 企业版接口指南》*。

用 “工具管理器” 转换数据

按钮的功能（如图 3-1）列于表 3-1 中。要转换数据，可以从 “当前列” 面板中选择列，然后单击适当的按钮。

- 1. 运行 Mineset （请参阅第 2 章中的 “运行 MineSet”）。
- 2. 使用 “文件” > “打开新的数据文件” 下拉菜单，向 “工具管理器” 窗口中选择一个数据集。在 “打开数据文件” 对话框中单击文件，在 “数据转换” 面板的当前列窗口中将出现数据集的列标题 （图 3-1）。

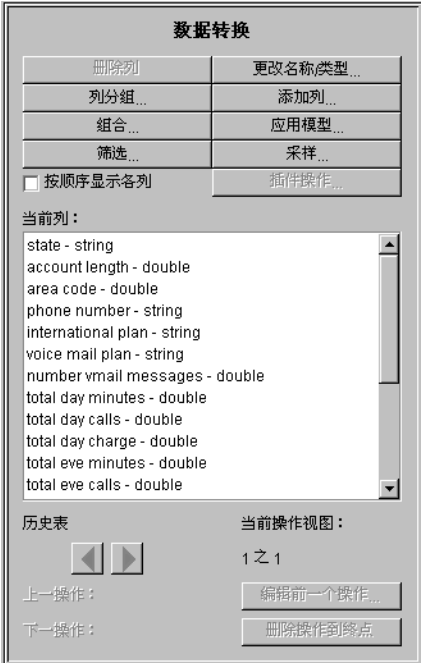


图 3-1 “数据转换” 面板

表 3-1 “工具管理器数据转换”面板上的按钮功能

按钮名称:	操作:
删除列	删除当前所有选定的列。
分组列	将每个记录分配到一定的范围（分组）。
组合	在分组的同时，执行一些列的总和、平均值、最小值、最大值、或计数组合。还可以排列索引组合的列。
筛选	根据布尔表达式创建数据的子集。只保留表达式判定为真的记录。
更改名称 / 类型	可以更改列的名称和类型，如：从浮点改为整型。（请参阅表 3-3 中的说明。）
添加列	根据数学表达式添加新列，表达式中可以包含其他列。
应用模型	用现有的模型标注新记录或修正数据。
采样	在大数据集中选择数据的随机子集。
按顺序显示各列	按字母顺序排列数据集的列。

删除和添加列

要删除或添加列，可以在“工具管理器数据转换”窗格中进行以下操作：

删除列—选择与可视化处理或挖掘无关的列，然后单击“删除列”按钮。删除无关的列可以节省计算时间。要选择多个非连续的列以同时删除，可以在选择附加的列时按住 Ctrl 键。对于连续列使用 Shift 键。

添加列—单击添加列按钮打开对话框，在其中指定新列的名称和数学表达式（图 3-2）。例如：可以基于类“age”添加一个叫做“minor_age”的新列，使用表达式：“如果年龄小于等于 18，那么 minor_age 为真；否则 minor_age 为假。”这样的表达式写作：“if ('age'<= 18) then ('minor_age'= 1) else ('minor_age'= 0)。”

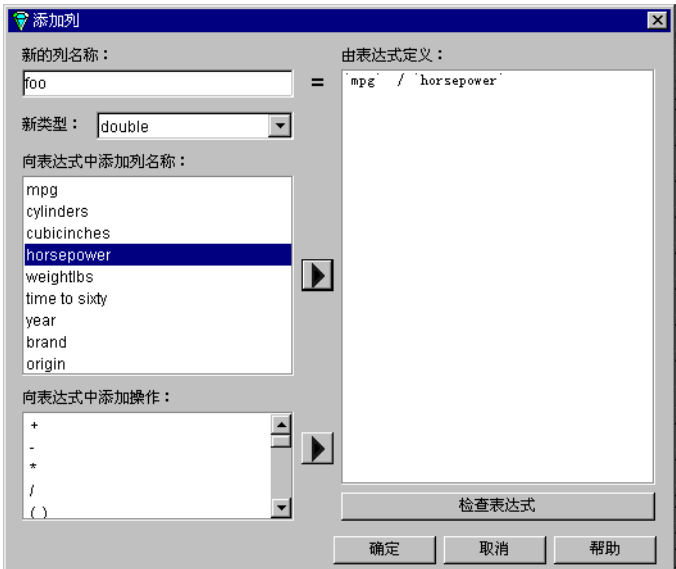


图 3-2 “添加列”对话框

要完成该操作，单击“添加列”按钮，将出现一个对话框（图 3-2）。

1. 在对话框中输入新列的名称，所有使用下面的弹出菜单指定列类型（整型、字符串、浮点型等等）。列名称中不能包括空格。
2. 在大的文本项区域中，填入表达式的定义：
 - 使用左下的滚动列表从可用的列和操作符中快速选择。
 - 要在表达式中插入列名称和操作符，可以在滚动列表中双击它，或者先选定它然后单击列表的向右箭头。
3. 单击检查表达式按钮，检查表达式用法是否正确。单击“确定”，关闭状态框。再次单击“确定”，关闭“添加列”对话框。

“添加列”对话框将检查类型兼容性：如果为字符串列分配了数值表达式（反之亦然），就会出现一个警告消息，新列的类型将被自动更正。

为列改变或创建新的分组

要组织模型或可视化处理，通常必须对记录进行分组，尤其要创建动画需要的滑动条。可以更改分组的边界，用不同的列数目和范围进行尝试，以简化可视化处理。

例如：如果希望将年龄范围分为：0-18, 19-30, 31-50, 51-60, 61+。这些范围表示为：(...18], (18 ...30], (30 ... 50], (50 ...60], (60 ...]。数字旁边的圆括号表示范围中不包括该数字。数字旁边的方括号表示范围中包括该数字。请参阅词汇表中的词条分组，以及《*MineSet 3.0 企业版参考指南*》中的“分组”。

在“工具管理器数据转换”窗格中，单击列分组 以打开一个对话框，指定分组选项（图 3-3）。



图 3-3 “列分组”对话框

1. 在对话框的顶部窗格中，单击要被分组的列。默认值为“不分组”。选择分组方法，并在“新列名称”中选择分组列的名称，或者直接键入新名称。如果选择多个列进行分组，“新列名称”将处于非活动状态。

要删除原来的列，请确保标注为“删除原始列”的框被选中。

2. 选择 “自动分组”、“自定义阈值” 或 “均匀分隔”（即平均分隔的分组）。
- 自动分组让 MineSet 通过机器学习建议分组阈值。
 - 可以使用自定义阈值在 “阈值” 文本字段中指定分组阈值。
 - 均匀分隔可以指定范围的起点、终点，以及每个分组的大小。
- 单击 “显示高级选项” 箭头，将出现 “高级分组控制” 窗格（图 3-4）。

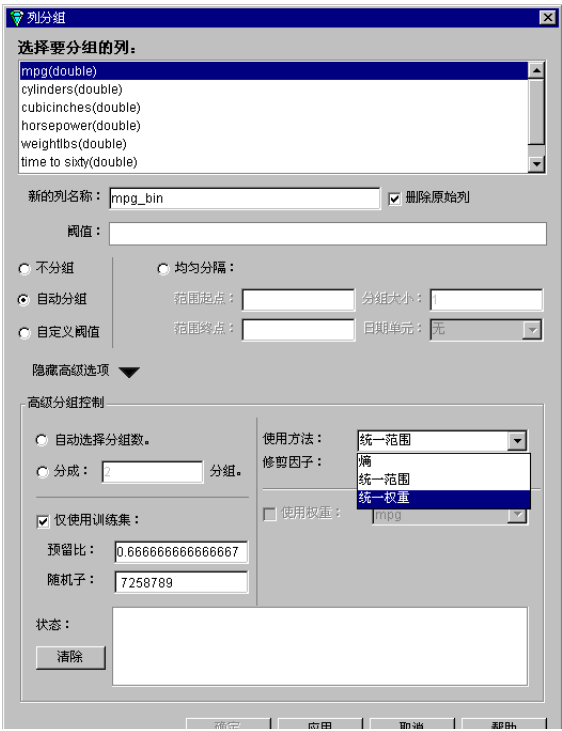


图 3-4 “高级分组选项” 窗格

3. 在“高级分组选项”窗格中，可以进行以下选择：

表 3-2 高级分组选项

选择	产生的操作
自动选择 分组数	MineSet 自动确定最佳的分组数。
分成 ____ 分组	将数据集划分为指定的组数。
使用方法	在该菜单中可以选择三种创建分组的方法。
熵	选取分组，使每个分组中的标签值相似，从而最小化熵或分组标签的混乱度。
统一范围	将数据集在范围内进行平均分布的分组。
统一权重	将数据集划分为具有相等权重（实例数）数据的组。
修剪因子	（仅适用于统一范围和统一权重）。删除分组阈值以外的数据。外层是位于数据主体外部的单独实例。
使用训练集	将数据集划分为子集，并且仅使用占总体支持比，比例大小的子集创建分组。
预留比	指定用于创建分组的数据集的百分比。
随机子	指定开始划分数据集的起点。随机子数值随时间改变。如果希望下次运行获得完全相同的数据，必须指定相同的随机子。

当选择 *自定义阈值* 设置您自己的分组阈值时，还可以输入：

- **阈值**
在文本字段中，输入由逗号隔开的阈值范围，如果范围中不包括该数字，则在其旁边用圆括号；如果范围中包括该数字，则在其旁边用方括号。例如：所有 18 岁以上（不包括 18 岁），30 岁以下（包括 30 岁）的年龄表示为：(18 ... 30]。
使用自定义阈值字段中的日期项要求使用双引号。如果输入的日期不带有双引号，系统将自动添加上。
“1/1/99”、“2/1/99”、“3/1/99”、“4/1/99”、“5/1/99”、“6/1/99”
还可以使用 *mm/dd/yyyy* 格式（例如：“5/22/1999”。）
- **均匀分隔**
当单击该按钮时，下面的文本字段将被激活。在适当的字段中输入分组范围的起点、终点，以及分组大小。

对均匀分隔阈值使用的日期不要加双引号。在该字段中可以使用两种日期格式（1/1/96 或 01/01/1999）。

在 *《MineSet 3.0 企业版参考指南》* 中的“分组”中可以获得详细信息。

通过组合创建新列

可以通过组合创建现有列内容以外的新列。例如：使用客户波动数据集可以在列中创建一个数据的子集，其中包含每个州的平均客户波动数。还可以使用各种准则分布列中的内容。因为组合重新对数据进行了加工，因此参考 *《MineSet 3.0 企业版参考指南》* 是很重要的，从而可以了解数组和分布的背景，获得对该特性更全面的认识。请参阅组合在词汇表中的定义。

组合过程将原始数据集中的行组合成新的组合行，从而减少数据集中行的总数。

要创建组合，请按照以下步骤进行：

1. 在“工具管理器数据转换”窗格中，单击组合，打开一个对话框以创建简单组合、生成数组或分布列（图 3-5）。当前数据集的列最初会出现在中间的“分组索引列”文本字段中。

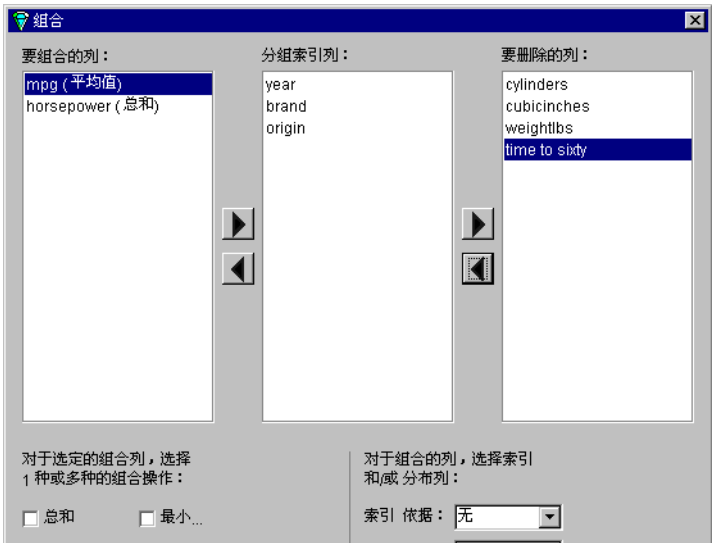


图 3-5 “组合”对话框

2. 选择要组合的列，然后单击左箭头按钮，将列移到左边。将“分组索引列”项留在中间（希望索引的列），然后将所有其余的列移到右列中。（在选择的同时按住 **Ctrl** 键以选择多个列。）

要组合的列—将列移到此处以创建组合。组合列中的值将被求和、平均、给出最小值或最大值、或计数，取决于面板底部被选中的复选框。

分组索引列—保留在此处的列在操作中不会改变。对于分组索引列中每个具有相同组合值的记录集合，生成表中的输出只有一个记录。

注意： 如果在分组列表中有许多列和数值列，那么在结果中得到的行可能要比期望的多。通常，在分组列表中使用的是具有很少值的几列。

要删除的列—删除的列不用于组合。

3. 单击该列下面的复选框，指定操作中如何对值进行组合：总和、平均、最小值或最大值、或计数。很少使用最小值和最大值。

对于数值列可以选择任何组合。对于其他类型，只允许计数。每个选项都会给出单独的结果；选择平均和最大值将给出一个带有平均值的列，以及另一个带有最大值的列。默认列名称职能识别用于创建它的过程 `count_state`、`avg_weightlbs` 等等。

4. 使用右下的弹出菜单指定索引项（如果结果是数组）和列分布（如果结果可以被分布）。

数组对于创建滑动条很有用，而几种可视化工具中的动画都需要滑动条。为了显示一个特征是如何按照（索引于）另一个特征而变化的，事先必须将数组中被索引的列分组。关于数组和分布的详细介绍，请参阅 *《MineSet 3.0 企业版参考指南》* 中的“组合”。

5. 单击“确定”。
6. 请确保“数据转换”窗口的“当前列”文本框中显示的新列名称是应用准则得到的。

在 *《MineSet 3.0 企业版教程》* 中详细描述了一个使用组合的示例，通过州来区分平均客户波动。

通过筛选限制列内容

可以用基于包含列值的表达式选择数据的子集，例如：可以只保留那些年龄小于 20 岁的记录，或英里 / 加仑小于 25 的记录。一旦使用“工具管理器”加载了数据集，就可以开始按以下步骤进行筛选：

1. 在“工具管理器数据目标”面板中，单击筛选，打开一个对话框，使用数学表达式筛选数据（图 3-6）。



图 3-6 “筛选”对话框

2. 从左边选择一个列名称和操作符，并在右边建立一个表达式，或者直接在“由表达式定义”文本字段中将其键入。产生的表中只包括表达式判定为真（或者，如果是数值型，为非零）的记录。请参阅《*MineSet 3.0 企业版参考指南*》中的“筛选面板”。关于表达式定义语言的全面介绍，请参阅《*MineSet 3.0 企业版接口指南*》。

更改列类型或名称

一些数据库将数值型值存储为字符串。Oracle 数据库将所有数字（整数和实数）以单一格式存储，默认值为“工具管理器”中的双精度数据类型。使用更改类型按钮以确保这些值被正确处理。使用同一个按钮更改列名称。

更改列类型

- 1. 在 “工具管理器数据转换” 窗格中，单击更改名称 / 类型。将出现一个对话框，显示列标题及其类型（图 3-7）。
- 2. 选择可以转换的列，然后单击新类型菜单。列表中将显示可能的列类型（无效的类型显示为灰色）。对选择的解释详见表 3-3。

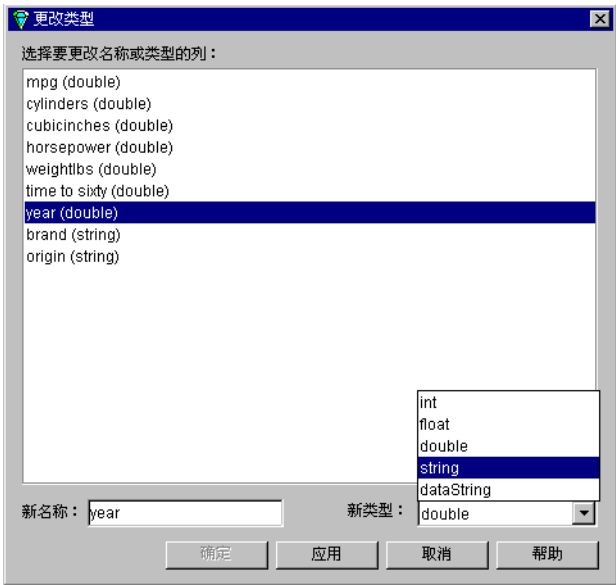


图 3-7 “类型” 弹出列表

表 3-3 列类型含义

类型	含义
整数	32 位带符号整数。
浮点	单精度浮点数，小数点为可选的。
双精度	双精度浮点数，小数点为可选的。
数据字符串	数据字符串 的值不是保存在普通字符串表中，因此每个数据字符串值都是独立存储的，即使多个值都是相同的。很少使用。
字符串	存储在普通字符串表中的字符串，可以节省内存。如果无法确定是使用字符串 还是 数据字符串，请使用字符串。
日期	指定为数据库中的日期类型。
分组	由分组操作创建的列。

- 3. 选择了新类型后，单击应用，使更改生效。
- 4. 单击关闭，退出对话框。

如果试图将不合适的字段（如名称）转换为数字，产生的值将都为零。

更改列名称

更改列名称可以在与更改列类型相同的对话框中进行。

- 1. 从“更改类型”对话框中，选择希望重命名的列，在“新名称”文本字段中键入新名称，单击应用。
- 2. 然后单击关闭，退出对话框。

应用模型

如果事先创建了一个模型，可以使用应用模型按钮标注当前表中的新记录、估计标签值的概率、测试模型在当前表上的性能、或在现有的模型上修正当前表。关于有关示例，请参阅第 7 章，“理解预测建模”。

数据采样

如果数据集太大，以至于不能有效工作，可能希望使用一个随机采样的子集。加载了数据文件以后，从“工具管理器数据转换”窗格中：

- 1. 单击采样按钮，打开“采样”对话框（图 3-8）。
- 2. 选择一个选项：
 - 单击“百分比”复选框，然后在文本字段中键入希望采集的当前数据集的百分比。该百分比是近似值。
 - 单击“计数”复选框，建立样本中需要的记录数。



图 3-8 “采样”对话框

- 3. 单击“互补样本”复选框，以获得除随机样本以外的所有记录。例如：如果在需要10%的样本时“互补样本”复选框没有被选中，那么当选中该框时，就可以获得其余90%的数据。单击“确定”。
- 4. （可选）在“随机子”字段输入数值，由此生成随机样本。如果没有指定数值，根据当前时间生成的随机数就被作为随机子。如果希望不同的随机样本，必须指定不同的随机子。对不同的数据集数据挖掘过程使用同样的随机子可以使您每次都使用同样的随机样本工作。当需要测试新的发现在数据集中的支持程度时，请改变随机子。请参阅词汇表中的词条*随机子*。

用“历史表”回溯历史操作

要查看以前操作的历史，以及在改变主意或发生错误时需要返回，可以使用“数据转换”窗格底部的两个*历史表*按钮。单击左箭头按钮，显示以前步骤中的表。单击右箭头按钮，返回到当前状态的表。关于按钮含义的解释，请参阅[表 3-4](#)。

表 3-4 历史表含义

字段	含义
当前视图是	计数将发生变化，指示正在查看的步骤。
上一操作：和下一操作：	显示邻近的操作，有助于跟踪。
编辑上一个操作	打开对话框，编辑显示在 <i>上一个：</i> 字段中的操作。请注意以后操作中使用的删除操作。
删除操作到终点	删除当前步骤以后的所有操作。如果还不在操作历史的结尾，“数据目标”面板将在单击按钮后被再次激活。
“查看操作历史”选项卡	打开“数据转换”表的一个完整历史（ 图 3-9 ）。
“单步转换和数据目标视图”选项卡	将您带到历史中一个单独的点，用“工具管理器”显示事件的状态。

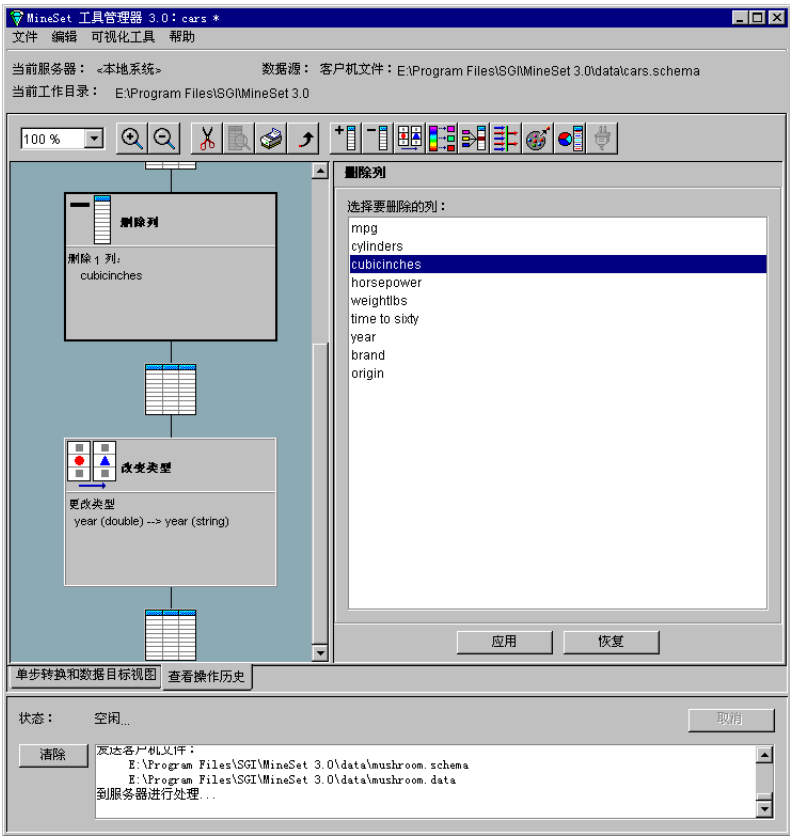


图 3-9 “查看历史”对话框

同编辑上一个操作一样，更改一个操作通常会影响（有时使无效）历史中的后续操作。可以选择一个指定的操作进行编辑、添加或查看。当更改影响历史时，操作历史视图会发出警告，并显示更改后的新历史。单击图表，返回到以前的操作，或单击操作之间的表以到达操作之间的表状态。

要返回到“工具管理器”中的上一个位置，可以单击“单步转换和数据目标视图”选项卡。

加权记录

在其他转换数据的方法中，如果需要，加权记录可以使一些记录比其他记录更重要，或更不重要。例如：电话公司将所有欺诈电话存储在数据集中，同时只存储了非欺诈电话的一小部分。通过使用记录加权，可以使每个记录充分反映其对总体的真实作用。

一些数据集已经被组合过，这些记录具有一个与它们自身相关的自然“计数”（例如：关于美国城市的统计通常具有一个相关的人口计数）。该计数属性可以被映射到权重，它相当于将每个记录按照计数次数进行复制。

记录加权的语义学就是：记录权重为 2 相当于两个记录权重为 1 的记录。允许浮点型权重。

记录的加权可以通过所有导入工具中的高级选项对话框完成（请参阅导入工具在词汇表中的定义）。加载了数据集后，请按照以下步骤加权记录：

1. 在“工具管理器数据目标”窗格中，选择“挖掘工具”选项卡，然后单击“分类”选项卡。
2. 从“模式”弹出菜单中，选择导入工具。
3. 单击高级选项按钮，然后在出现的对话框中，单击“使用权重”复选框。这表示在第 4 步中选定的属性将对记录进行加权。
4. 为权重选择列。
5. 如果第 4 步中使用的属性也作为普通属性出现，请单击“权重保留为属性”复选框。如果该框没有被选中，权重属性将不会影响导入过程。
6. 单击确定，接受这些选项，然后单击继续，运行选定的导入工具。

查找重要的列

列重要性可以帮助发现在预测选定标签列的不同值的过程中最重要的列。列重要性和聚类的区别（请参阅第 12 章，“用聚类划分数据”）在于：利用列重要性可以决定使用哪个标签来确定列的重要性。而在聚类中，虽然数据本身就会显示区分因素，但却不提供标签。

例如：使用列重要性查找“散点可视化工具”中映射到坐标轴的最好的三列。只需简单选择标签并运行工具，然后会出现一个窗口，其中的三列就是最好的区分器。一个被称为“纯度”（0 到 100 的一个数）的量度会告诉您用这些列区分不同标签的效果。添加更多的列会提高纯度。



图 3-10 “列重要性”选项卡

- 1. 在“工具管理器数据目标”窗格中，单击“挖掘工具”选项卡，然后再单击“列重要性”选项卡，如图 3-10 所示。
- 2. 若要以简单模式运行，可以在弹出菜单中选择离散标签，在“要查找的列数”字段中指定一个数字，然后单击继续。状态窗口的底部将给出结果。

若要控制列的选择，可以单击高级模式按钮，这时会出现“高级列重要性”对话框，（请参阅图 3-10）。

- 3. 在此对话框中，可以单击“使用权重”复选框，并从“关于列的重要性”弹出菜单选择一个列。这将有助于在确定重要性时，决定该列是否表现为常规属性。请参阅第 49 页的“加权记录”。关于工具操作的详细信息，请参阅《MineSet 3.0 企业版参考指南》中的“列重要性”。
- 4. 对话框中包含两个列名称列表：左边列表中包含可用的属性，右边列表中包含（被您或“列重要性”运算法则）选择为重要的属性。单击右向箭头，可以将选定的列从左边移到右边。

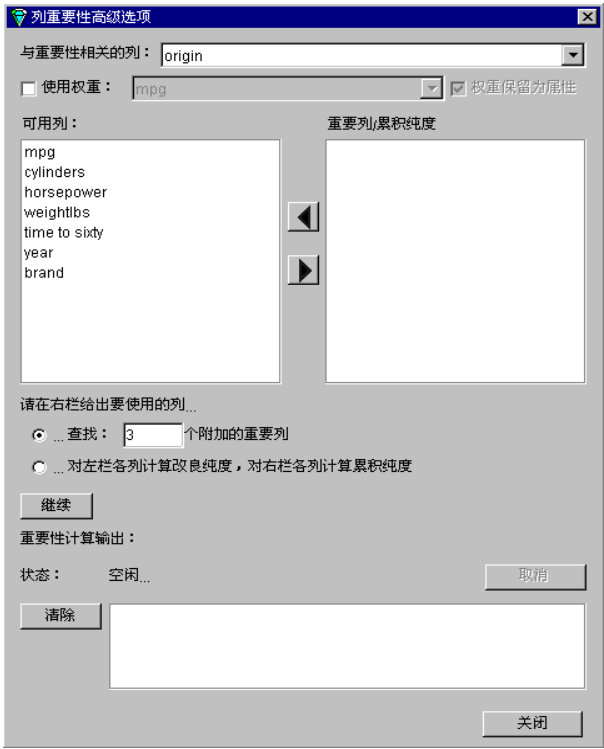


图 3-11 列重要性的高级模式

5. 在对话框的中间部分有两个高级模式选项：

- 查找多个重要属性

单击选项 ... 查找 [数字] 个附加的重要列上的选择钮。如果直接单击继续而没有进一步的更改，那么效果将与简单模式一样。指定数目的重要列被找到并被移动到右边的列中，同时给出了累积纯度。

但是，如果单击右向箭头，将列名称从左边列表移到右边列表，那么就可以指定这些被包括的列，并让系统添加更多的列。

单击继续，查看每个列的累积纯度，以及列表中以前的纯度。纯度为 100 表示：使用给定的列，可以完全区分开数据集中不同的标签值。

- 对可用属性排序

可以检验每一列添加到右边列表中时纯度的增幅。如果使用“汽车”数据集，如图 3-11 所示，可以将列 `cylinders` 移到右边列表中，然后请求系统计算保留在左边列表中的每一列能够产生的纯度增幅。为右边的列（已经标记为重要的）计算累积纯度。结果将显示在右边的列中。

单击 ... 对左边各列计算改良纯度，对右边各列计算累积纯度上的选择钮。该子模式允许对过程进行精密控制。如果两个列排得非常近，可能希望使一个替换另一个（可能因为它收集更便宜、更可靠、或者更易于理解）。

“散点和平伸可视化工具”查找最佳的三个坐标轴时，列重要性很有用。应被选择为“树可视化工具”查找好的区分层（区分开不同标签值的分层，此时该标签为“树可视化工具”的关键字）时，它也很有用。

关于使用“列重要性”样例文件的详细信息，请参阅《*MineSet 3.0 企业版参考指南*》中“列重要性”。

用散点和平伸可视化工具检查数据

“散点”和“平伸”可视化工具比较类似，其区别在于“散点可视化工具”将数据显示为一系列单独的实体，而“平伸可视化工具”则显示数据点的组合，看上去像是三维景观中不透明的、彩色的云。要了解它们各自优点的详细信息，请参阅 [《MineSet 3.0 企业版参考指南》](#)。本章主要包括以下几个主题：

- 第 53 页的“散点和平伸可视化工具概述”
- 第 57 页的“为散点和平伸可视化工具转换数据”
- 第 60 页的“运行散点和平伸可视化工具”
- 第 64 页的“查看散点和平伸可视化工具中的结果”
- 第 69 页的“在“散点”和“平伸”可视化工具中创建动画”
- 第 75 页的“处理散点和平伸可视化结果”

提供了一些样本配置和数据文件，用于演示“散点”和“平伸”可视化工具的特性和功能。这些文件在 MineSet 3.0 下的 \examples 目录中，位于 MineSet 的初始安装目录。

散点和平伸可视化工具概述

本部分讲解“散点”和“平伸”可视化工具的用法，以及选择使用时的一些原则。

散点可视化工具概述

“散点可视化工具”的单个数据点对应于数据文件中的行。当数据点少于 50,000，或对数据采取了一些处理使其成为小的组合集合时，这种可视化处理的效果很好。“散点可视化工具”生成的散点图可以被制成动画以更清楚地显示关系（[图 4-1](#)）。

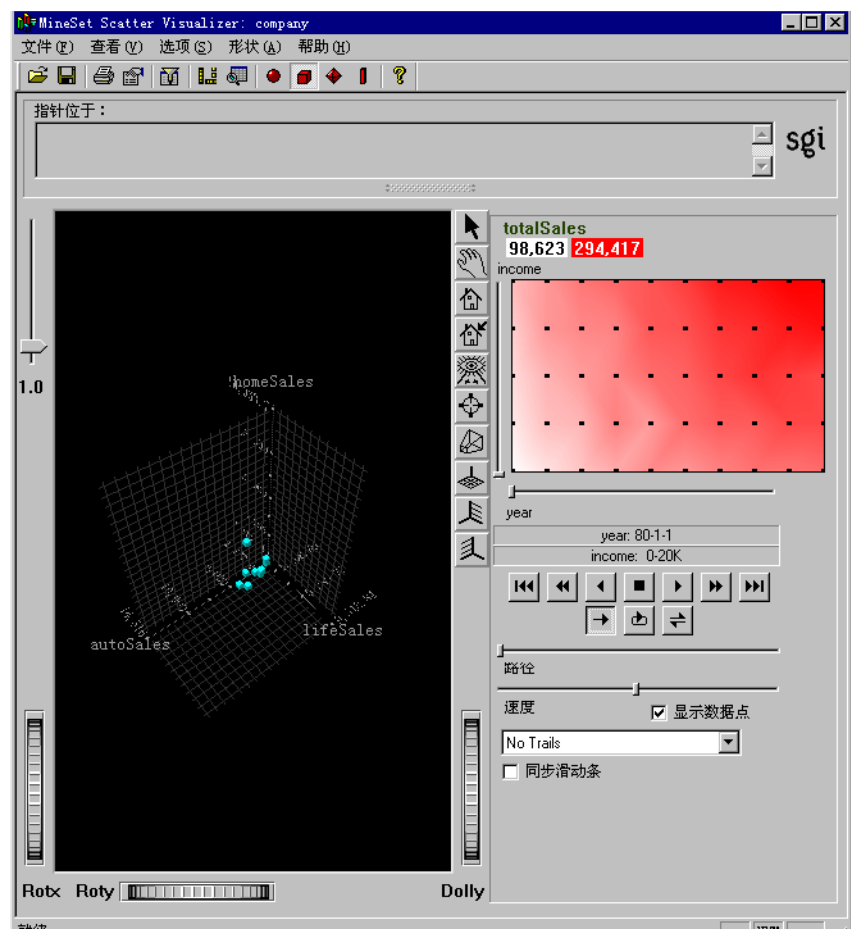


图 4-1 散点可视化工具屏幕样例

“散点可视化工具”显示的是三维景观，其中的数据列被映射到实体和元素（如坐标轴、大小和颜色）。如果将一个或两个数值型变量映射到滑动条，就可以将大小、颜色或实体位置制成动画。在图 4-1 所示的示例中，数据代表一些公司随时间变化的销售额。如果将时间变量映射到一个滑动条，将销售额映射到大小，那么实体随时间滑动条的增长或缩短就被制成了动画。该示例为 MineSet 3.0\examples\company.scatterviz。

当回放动画路径时，可以看到大小、颜色变化和数据点运动的趋势或异常。可以在三维景观中浏览以寻找感兴趣的观察点，或对变量值进行缩放使其更加突出。要清除场景，可以对画面进行过滤，仅显示符合特定准则的实体。

平伸可视化工具概述

利用“平伸可视化工具”，可以可视化地分析多个变量之间的关系，当使用动画特性时，可以更清楚地看到一些关系。“平伸可视化工具”使用图形对象，称为平伸，代表数据点的组合。平伸对象的颜色和不透明度在动画过程中可以更改，而位置却不能。

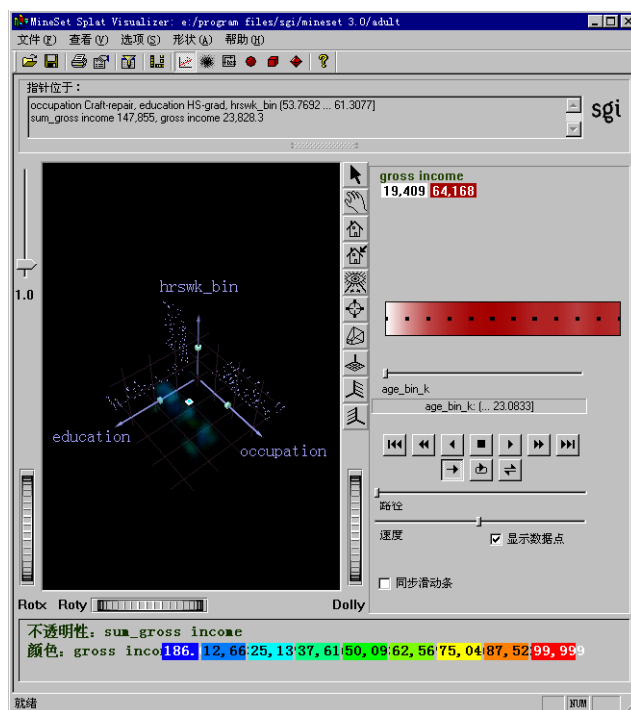


图 4-2 具有一维汇总滑动条的平伸可视化工具样例

图 4-2 显示的是一个三维景观的“平伸可视化工具”视图，*adult94* 样本数据集的列被映射到其中的坐标轴、滑动条、颜色和不透明度。它类似于散点图，只不过散点图是单独地画出每个数据点，而“平伸可视化工具”将彼此接近的数据点（在同一个分组内）组合起来并把它们画为一个单独的平伸对象。结果与你为每个单独的点画一幅散点得到的画面很近似。所得图象可以被认为是三维的彩色直方图。

从“平伸可视化工具”的结果中，可以：

- 使用动画面板可以了解数据中的全局性移动及趋势。变化的颜色和不透明度展示出实际的移动情况（关于图形的颜色方案，请参阅本书的联机版）。
- 在三维景观上掠过可以突出特定的维或观察点。
- 使用高度滑动条（在主窗口的左上）增强能见度，可以降低或提高平伸不透明度。数据密的区域颜色可能变化很小，因为颜色是取决于许多值的平均值的。
- 对画面进行过滤，仅显示符合特定准则的平伸。可以对相应于坐标轴、滑动条、权重、和颜色的列进行过滤。
- 选取卷中单个平伸的纹理信息。
- 用框选择器定义一个选定的区域，以追溯原始数据或发送到“工具管理器”。

例如：图 4-2 中的左边坐标轴显示的是在轴上按平均收入排列的每个职业。行政管理职业列在坐标轴的末端，平均收入最高，并为该值提供了一个自然级数。另一方面，*education*（图 4-2 中右边的坐标轴）值的顺序通常从低到高，但是在少数情况下，顺序上也会不规则。这种意料之外的顺序可能很有意思，因为它指出了数据与期望不相符的地方。

为散点和平伸可视化工具转换数据

在转换决策中要对数据有相当的计划和了解。组成数据集的列在映射到坐标轴、颜色或滑动条之前，通常需要经过一定方式的处理或转换。在查看了一个结果之后，可能经常需要返回来以另外的方式转换数据。

在 MineSet 中，大部分的转换是使用“工具管理器”中的“数据转换”窗格来完成的，（请参阅第 34 页的“用“工具管理器”转换数据”）。以下是一些可能需要的转换：

- 添加列，通常是对已有的数据应用一个数学公式。
- 删除无关、冗余、或是包含明显无用预测因素的列。可以减小数据集的大小及加速处理。
- 筛选数据。例如：可能只希望使用含有在一定范围内取值的特定属性的记录。
- 更改列名称或类型。
- 分组数据—将连续范围的数据划分到离散的组中。
- 适用于“平伸可视化工具”的组合数据—如果不用“工具管理器”手工组合数据，该工具将自动执行此任务。但是，最好在“工具管理器”中组合数据，以便在服务器上完成任务，而不用将组合的数据集送到客户机系统（请参阅第 59 页的“平伸可视化工具的处理技术”）。
- 适用于“散点可视化工具”的组合数据—组合可以将大的数据集压缩为小的组合集合。组合后，表中的每一行都将对应于散点图中的一个实体。在组合过程中可以为动画创建数组列。

“平伸可视化工具”中的交互性与表现出来的数据点数目无关；它只取决于坐标轴维中的分组数。如果数据集非常大，请在“工具管理器”中进行显式组合。这可以是服务器使用流操作执行进程，而比将整个数据集发送到客户机并在那里进行组合要好得多（请参阅第 59 页的“平伸可视化工具的处理技术”）。

表 4-1 所列为可以在“平伸可视化工具”中映射到可视实体的列类型。

表 4-1 平伸可视化工具中允许映射的列类型

实体	列类型	要求
坐标轴	数值型	需要进行分组，手工或由“工具管理器”自动执行。
	分组型	可以被直接映射。
	字符串	可以被直接映射。
颜色	数值型	可以被直接映射。
不透明度	数值型	如果没有映射，默认为记录计数。
滑动条	数值型	必须进行分组，手工或由“工具管理器”自动执行。
汇总	数值型	可以被直接映射。

表 4-2 所列为可以在“散点可视化工具”中映射到可视实体的列类型。

表 4-2 平伸可视化工具中允许映射的列类型

实体	列类型	要求
坐标轴	任意，包括数组	可以被直接映射。
实体 - 大小	数值型或数组型	可以被直接映射。
实体 - 颜色	任意，包括数组	可以被直接映射。
实体 - 标签	任意，包括数组	可以被直接映射。
滑动条	分组型	可以在组合过程中被隐式映射。
汇总	数值型	可以在组合过程中被隐式映射。

如果字符串列被映射到坐标轴，那么分组将被定义为该列标识值。这些值按顺序排列，从而与映射到颜色的属性关联（如果没有属性被映射到颜色，那么它们就与不透明度关联）。沿着字符串值的坐标轴查看颜色的变化可以了解该列与映射到颜色的列的关联程度。

平伸可视化工具的处理技术

在“工具管理器”中进行组合可以避免在客户机上执行该任务。这意味着大部分的工作将由服务器完成。通常，服务器比客户机更加强大，并且执行流操作，不需将整个数据集加载入内存。而且，在“工具管理器”中组合可以避免将大的数据集传输到客户机。完成以上操作的步骤为：

1. 使用“工具管理器”对用于坐标轴和滑动条的数值列进行分组。
2. 在按照坐标轴和滑动条列分组的同时，通过计数和平均组合映射到颜色的列。
3. 将产生的计数组合映射到不透明度。
4. 将产生的平均组合映射到颜色。

以 *adult94* 数据集（提供分布状态）为例，以上所示的具体过程步骤列在了下面。

1. 分组 `age` 和 `hours_per_week`。您将在需要分组值的滑动条上使用前者，而将后者用作坐标轴上分组数值的示例。
2. 组合 `gross_income` 使用“计数”和“平均值”。保留“分组索引列”窗格中的 `education`，`occupation`，`age_bin` 和 `hours_per_week_bin`，并删除其他所有的列。这将提供一个可视化处理，显示如何 `gross_income` 按照其他三个准则变化的方式。
3. 将 `education`，`occupation` 和 `hours_per_week_bin` 映射到坐标轴。
4. 将 `avg_gross_income` 映射到颜色，`count_gross_income` 映射到不透明度，`age_bin` 映射到滑动条。根据映射坐标轴的准则，该映射使颜色和不透明度按年龄显示收入（平均和计数）的变化。

当调用工具时，请注意所有过程都是在服务器上完成的，并且数据文件 *adult94.splatviz.data* 中包含的行是原始数中行的组合。生成的可视化处理如图 4-2 所示。

运行散点和平伸可视化工具

运行“散点”和“平伸”可视化工具最简单的方法是从“工具管理器”中：

- 1. 打开“工具管理器”，选择服务器和数据源。具体介绍请参阅第 15 页的“运行 MineSet”。在第一个示例中，使用的是客户波动数据集。
- 2. 在“工具管理器”的“数据目标”窗格上，选择“可视化工具”选项卡；然后在下面一行选项卡中，选择“散点可视化工具”或“平伸可视化工具”，（图 4-3）。

映射散点可视化工具可视元素

- 1. 在“工具管理器”的“数据转换”窗格滚动菜单中选择，将当前列映射到可视元素。例如：对于“散点可视化工具”，如果使用客户波动数据集，则为坐标轴 1 选择 number of vmail message，为坐标轴 2 选择 total day charge，为坐标轴 3 选择 total eve charge。

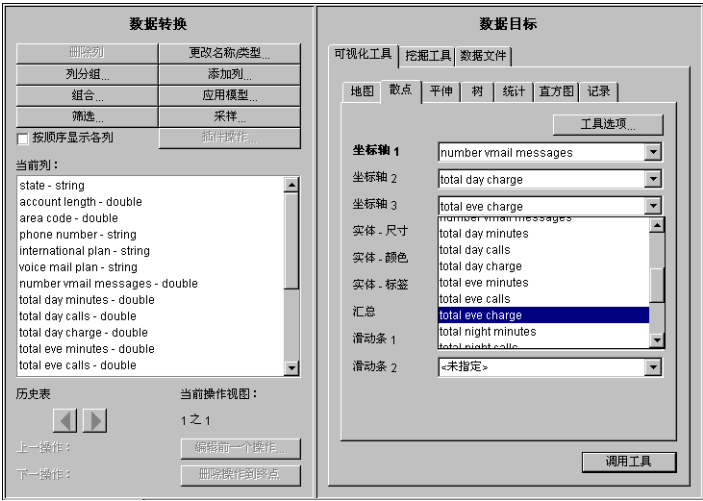


图 4-3 选定“散点可视化工具”的“数据目标”面板

表 4-3 标识了散点可视化工具的可视元素的映射效果。

表 4-3 映射散点可视化工具中的可视元素	
可视元素	操作:
坐标轴 1	将数据只分配给第一个坐标轴通常没有用。将数据分配给所有三个坐标轴可生成一个三维图。
坐标轴 2	
坐标轴 3	
实体 - 大小	将大小、颜色和标签分配给实体。用 “散点可视化选项” 对话框为这些映射指定选项，可以单击 “散点可视化工具选项” 按钮打开对话框。
实体 - 颜色	
实体 - 标签	
汇总	确定汇总滑动条背景的颜色
滑动条 1	将列直接映射到一个或两个动画滑动条
滑动条 2	

2. 要撤消映射，可以选择元素，然后再从可用列列表中选择 <未指定>。如果使用自己的数据集，可以考虑运行 “列重要性” 工具以帮助确定可能的映射（请参阅 [“查找重要的列”](#)。）
3. 在 “数据目标” 面板的右下角单击 *调用工具*。

映射平伸可视化工具元素

与 “散点可视化工具” 类似，“平伸可视化工具” 也要求在运行之前将列链接到可视元素。在适合可视元素准则之前，一些列可能需要处理。本示例使用系统提供的例子中的蘑菇数据集，主要显示如何对列进行处理以满足映射的要求。按第 60 页的 [“运行散点和平伸可视化工具”](#) 所述，从 “工具管理器” 开始，选择蘑菇数据集，并选择 “平伸可视化工具”。

1. 在“工具管理器”的“数据转换”窗格滚动菜单中，选择将“当前列”中的项目映射到“可视元素”中的项目。



图 4-4 将列映射到平伸可视化工具的可视元素

例如：对于坐标轴 1 选择 **habitat**，对于坐标轴 2 选择 **odor**，对于坐标轴 3 选择 **spore print color**。

表 4-4 标识“平伸可视化工具”可视元素的映射效果。

表 4-4 映射平伸可视化工具中的可视元素

可视元素	操作:
坐标轴 1 坐标轴 2 坐标轴 3	将数据只分配给第一个坐标轴通常没有用。将数据分配给所有三个坐标轴可生成一个三维图。
颜色	需要一个数值列以确定平伸的颜色。如果没有属性映射到颜色，效果就会是单色的。
不透明度	默认情况下的记录计数。如果在“工具管理器”中进行了计数组合（或加权列的总和），或者数据集中的列已经是基于计数或权重的，那么可以为此元素使用该列。

表 4-4 映射平伸可视化工具中的可视元素

可视元素	操作:
滑动条 1	将列直接映射到一个或两个滑动条。列必须是数值型或分组的。
滑动条 2	
汇总	确定汇总滑动条背景的值。如果没有汇总列被映射，默认情况下为计数。如果映射了汇总列，那么在汇总中会显示该列的加权平均值。

在映射列中，如果需要数值列，可以使用如下的表达式将两值的字符串列变成新创建的数值: ('stringCol'==" value1")? 1:0。关于详细信息，请参阅第 3 章中的“删除和添加列”。

本例中，添加了一个整数列，代表可食性的概率，以便将该特征映射到颜色。蘑菇数据集只包含字符串列。非连续属性（如字符串）不可能被映射到颜色，因为不可能对这样的属性求平均。

2.
- 在“工具管理器数据转换”窗格中单击添加列按钮。在生成的对话框中，创建一个新列，名称为 p_edible，类型为 int。表达式为 (edibility == "xedible")?100:0。请注意引号的位置。



图 4-5 为映射添加一个整数类型的列

3. 将 `p_edible` 映射到颜色，从而根据显示的颜色定义蘑菇是可食用的还是不可食用的。
要撤消映射，可以从下拉菜单中选择 < 未指定 >。
4. 在“数据目标”窗格的右下角单击 *调用工具*。生成的可视化处理提供了一些对蘑菇的有趣认识。

以下部分介绍 MineSet 产品提供的另一种访问数据的方法。

查看散点和平伸可视化工具中的结果

与所有可视化工具一样，如果在没有指定配置文件（即：名称扩展名为 `.scatterviz` 或 `.splatviz` 的文件）的情况下运行“散点”或“平伸”可视化工具，那么就只能使用“文件”和“帮助”下拉菜单。要显示可视化工具主窗口中的所有菜单和控件，必须打开配置文件。使用“文件”>“打开”以查看配置文件列表。这些文件在 `MineSet 3.0/examples` 目录中，位于 MineSet 的初始安装目录。

当打开有效的配置文件时，将会看见三维景观。图 4-1 所示的是 `company-total.scatterviz`，它显示不同收入等级的人寿保险、汽车保险和家庭保险销售额随时间的变化。

查看模式

“散点”和“平伸”可视化工具的两种查看模式为 *抓取*和*选取*。要在这两种模式之间切换，可以将光标移到主窗口中，然后按 `Esc` 键。单击可视化工具窗口主面板边框上的箭头或手形，也可以从一种模式改变为另一种模式。

散点可视化工具中的选择模式

当处于“散点可视化工具”的选择模式时：

- 要显示“散点可视化工具”上部文本字段中实体的有关信息—将鼠标移到对象上。
- 要选择一个实体，在对象上按鼠标左键。要选择多个实体，单击鼠标左键的同时按住 Ctrl 键。
- 要清除选择—单击黑色背景。关于漫游的详细信息，请参阅第 27 页的“在 MineSet 三维可视化工具中漫游”。

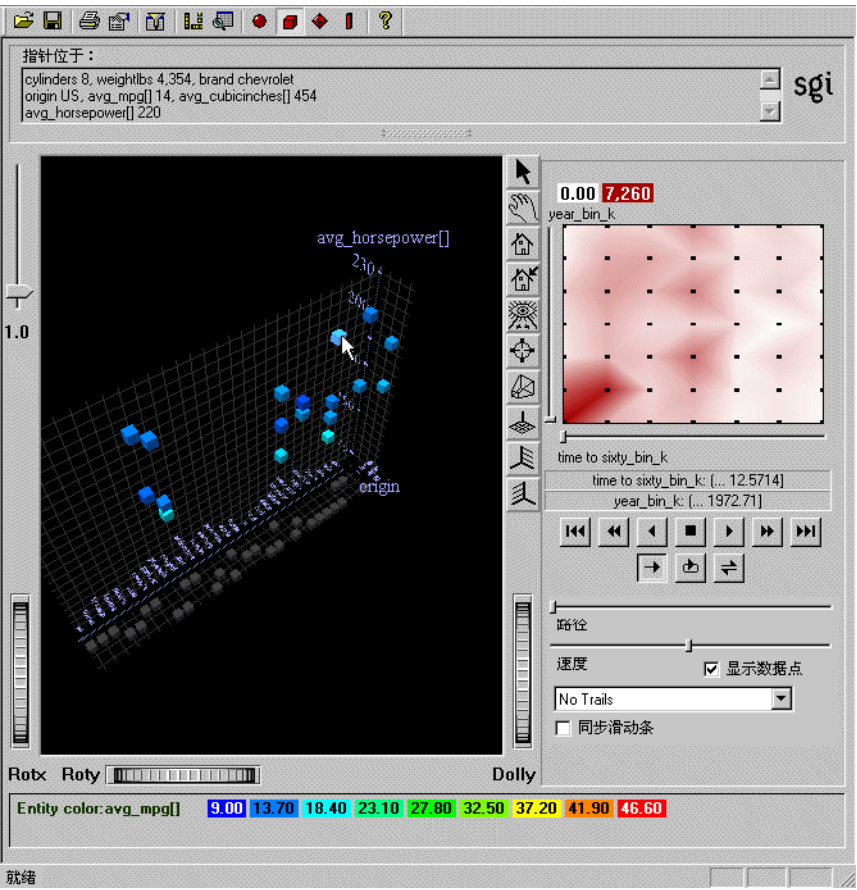


图 4-6 显示选定实体的信息

关于从命令行指定 “散点” 和 “平伸” 可视化工具的信息，请分别参阅它们在
《*MineSet 3.0 企业版参考指南*》中各自的条目。

平伸可视化工具中的选择模式

在 “平伸可视化工具” 的选取模式中，可以移动三维选取拖动工具穿过数据点的密云，查找场景中区域的有关信息。该选取拖动工具由一个圆柱体和一个正方形组成。

- 要沿圆柱体的坐标轴平行移动，可以单击圆柱体并将鼠标指针沿希望的方向拖动。
- 对于受到正方形选定的平面约束的移动，可以单击正方形并拖动。使用 **Shift** 键将运动约束在平面内的一个坐标轴上。
- 要在平行与约束模式之间切换，可以在光标位于拖动工具上时按下 **Ctrl** 键。（不需要按鼠标按钮。）
- 另外，每个坐标轴都有一个盘，与选取拖动工具的位置排成一行。在坐标轴上移动盘会使拖动工具移动，反之亦然。

当选取拖动工具位于数据上时，圆柱体的颜色改变为数据下面云团状 “平伸” 的颜色，并且有关该区域的信息会显示在视图区域顶部（图 4-7）。如果没有数据，圆柱体将保持淡灰色，有关该位置的信息会显示在区域顶部以帮助漫游。

当完成拖动松开鼠标按钮时，当前所在 “平伸” 图形的信息会显示在顶部的选择窗口中。如果移动了动画滑动条，该选取信息就会被更新。使用鼠标可以将该选定信息剪切并粘贴到其他应用程序中，例如：报表或数据库。

取消选定 “选项” > “显示选取拖动工具”，可以将选取拖动工具从场景中删除。

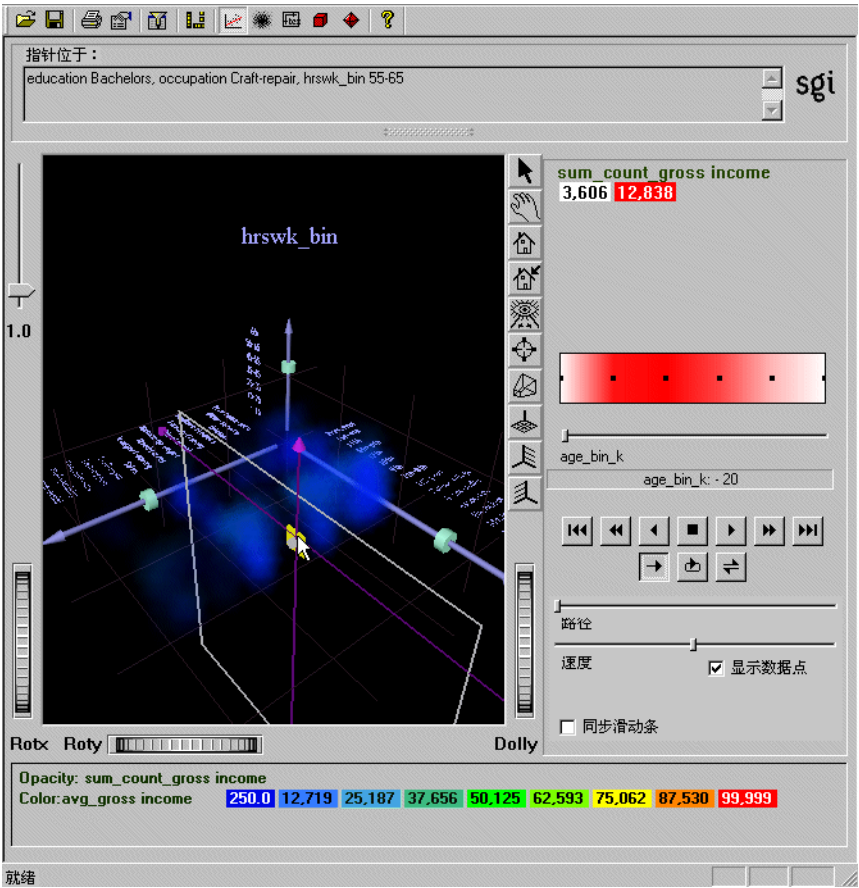


图 4-7 操作数据上方的选取拖动工具

在第 27 页的“在 MineSet 三维可视化工具中漫游”中可以找到窗口图标的说明。请参阅《MineSet 3.0 企业版参考指南》中的“散点可视化工具”和“平伸可视化工具”以获得可视化工具选项的补充描述，以及“颜色浏览器”以获得改变颜色的详细信息。

为散点可视化工具创建滑动条

当觉得一个值随某一列的值变化时，可以将该列映射到滑动条。如果列是（整数、浮点、数据点类型）数值型的或分组的，就可以被映射到滑动条。如果列已经被分组，那么它的名称后就会带有 `_bin`。列类型注释在“当前列”列表中的列名称后，例如：白天的电话总数 - 双精度。在大多数情况下，“工具管理器”中从列到滑动条的简单映射将自动创建滑动条。

自动的滑动条创建

如果在组合步骤中没有隐式地指定滑动条，“工具管理器”将通过分组和组合自动创建滑动条。这些自动操作发生在单击调用工具之后。没有删除或映射到可视实体的列将用于确定特定实体的数目（即：它们将称为自动组合过程中被分组的列）。单击“工具管理器”中的工具选项，可以在“工具管理器选项”对话框中找到当前滑动条的索引。

要组合每一个被映射到可视元素的数值列。可以在“工具选项”面板中选择组合的类型。如果希望对不同的实体进行不同的组合，必须手工创建滑动条。

手工滑动条创建

要保证预期的结果，最好用“工具管理器”显式分组与组合。为了完成该操作，首先要对计划映射到可视实体的列进行组合（使用平均、计数或总和），然后按其他的字符串和分组列进行分组，最后按希望成为动画滑动条的分组列“排列索引”。对于所有按滑动条变量索引的组合列，结果将具有数组列。当滑动条以这种方式被指定时，它们将无法直接映射到滑动条元素。关于组合的详细信息，请参阅《*MineSet 3.0 企业版参考指南*》中的“组合”。

如果在当前数据表中有一个数组列，“散点可视化工具”中将有一个对应于该列索引的滑动条。对于一个索引，将有一个一维的 X 滑动条；对于两个索引，将同时创建 X 和 Y 滑动条。当前表中的所有数组列都必须具有相同的索引；否则，将不会创建滑动条。关于创建数组列的详细信息，请参阅第 3 章中的“通过组合创建新列”和《*MineSet 3.0 企业版参考指南*》。

在“散点”和“平伸”可视化工具中创建动画

使用可视化工具主窗口右边的动画控制面板创建动画。只有映射了滑动条时，才会出现动画窗口。例如：`company.scatterviz` 就提供了这样一个数据集。请参阅《[MineSet 3.0 企业版参考指南](#)》中的“动画”。当觉得一个值根据特定准则变化（如人口密度随时间变化）时，可以将该列映射到滑动条。`adultjobs.splatviz` 是一个使用动画滑动条的“平伸可视化工具”示例。

最有效的方式是为滑动条或坐标轴选择独立的属性，为坐标轴、大小和颜色选择从属的属性。如果只有一个滑动条被映射，（例如 `adultjobs.splatviz` 或 `adult94.scatterviz`），那么汇总窗口将被压缩。该滑动条的维由它下面的标签标识。如果数据集没有映射到滑动条的列（如 `brand.scatterviz`），那么将不会出现滑动条控制。

使用可视化工具汇总窗口演示动画

可以在“散点”或“平伸”可视化工具动画控制面板的汇总窗口部分创建动画。一维或二维中的汇总窗口将根据主窗口中数据的组合显示一个汇总值的数组。在汇总滑动条窗口中，每个黑点有一个汇总值。汇总颜色被指定了高值，指定给白色的是低值（内插在这两种颜色之间的是中间值）。如果对于特定位置上的汇总变量没有数据，或只有空值，那么它将显示为灰色以表示未知。使用该滑动条左边图中所有数据的指定组组合算汇总值。可以使用动画面板上的“显示数据点”复选框关闭这些黑点。

该示例介绍如何使用准备好的 `company.scatterviz` 文件创建动画。

1. 在“工具管理器”菜单条中选择“可视工具”，再选择“三维可视化工具”。
2. 在菜单中，选择“文件” > “打开”菜单。
3. 打开 `company.scatterviz` 文件。该文件在 `MineSet 3.0\examples` 目录中，位于 MineSet 的初始安装目录。

4. 在汇总窗口中，选择一个黑点作为动画路径起始点。在该点上单击并按住鼠标左键，然后在窗口上拖动光标。松开鼠标左键以终止路径。
 - 还可以在一个点上单击鼠标左键定义起始点。然后拖动一个独立维的滑动条，并沿该维画一条直线。如果有两个滑动条，那么也可以使用第二个滑动条沿第二个坐标轴画一条直线。
 - 定义路径的另一个方法是：在一个点上单击鼠标左键，然后在另一个点上单击鼠标中键，作出连续的路径段。该选项仅适用于三键鼠标。
5. 使用汇总窗口下的 VCR 状按钮控制动画。单击第二行的路径按钮之前，请先单击第一行的播放按钮。关于控制的详细信息，请参阅《*MineSet 3.0 企业版参考指南*》中的“动画”。

在散点可视化工具中显示动画轨迹

可以用“散点可视化工具”显示运动轨迹，以演示实体不断改变的动画路径。当创建动画时，轨迹会以选定的形式显示在每个选定的实体后面。可以从“散点可视化工具动画控制”面板中的运动选项菜单（图 4-8）中选择：

- 没有轨迹—默认选项
- 线型轨迹—细的彩色线
- 淡出轨迹—类似于线轨迹的透明彩色线，最近的位置也最不透明
- 管状轨迹—三维管状轨迹，其厚度随实体在动画路径中运动时大小的变化而变化。管状轨迹太多会明显降低动画速度。

所有轨迹都根据最初的实体进行了颜色编码。当实体改变了颜色，如从红色变为蓝色，那么随着汇总窗口下滑动条的移动，对应的轨迹也将会在两个位置之间逐渐改变颜色。轨迹建立在未映射属性在整个路径上保持不变的点之间。作为示例，可以显示收入随年份的改变，并留下可视轨迹以显示两点之间的变化速度。

按照少数列分组的组合数据可能非常适合用于显示运动轨迹。最初，将显示散点图中受路径影响的所有点的运动轨迹。单击任意实体以选中它，从而只为选定的点显示轨迹。这可以用来减少视觉混乱。具有空位置的实体在轨迹中表现为中断。

图 4-8 所示为具有管状运动轨迹的“散点可视化工具”示例。

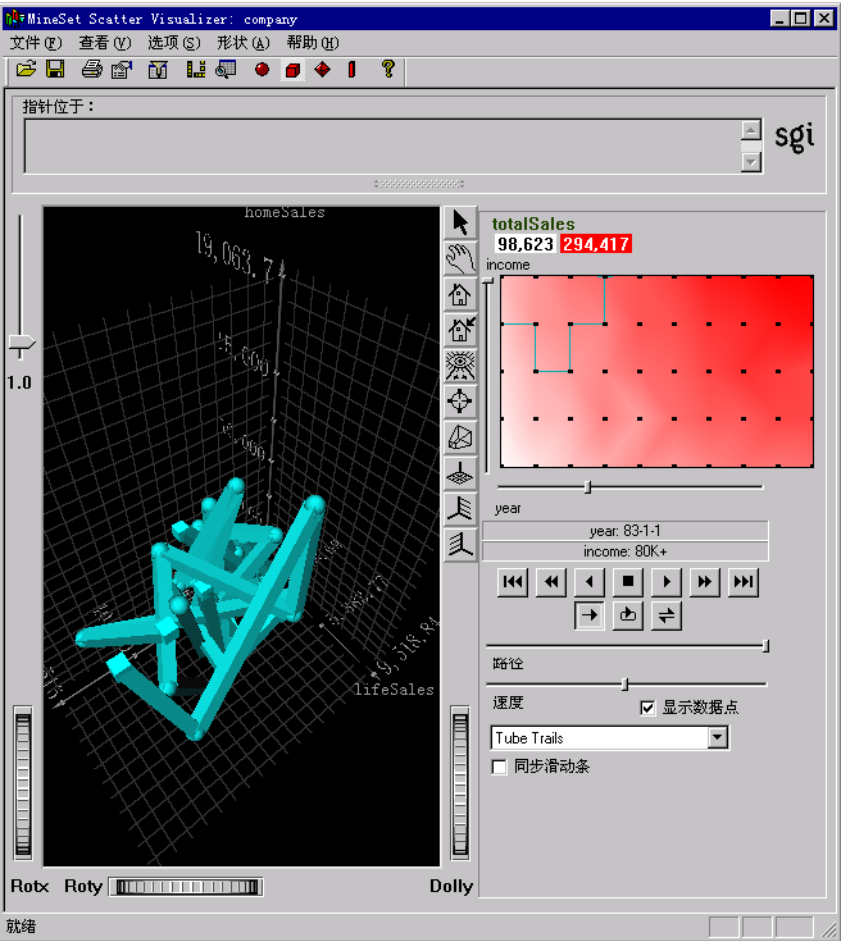


图 4-8 散点可视化工具管状运动轨迹示例

在平伸可视化工具中演示动画

从本质上看，在“平伸可视化工具”中演示动画与在“散点可视化工具”中是一样的，除了“平伸可视化工具”没有管状轨迹选项以外。关于 *adultJobs.splatviz* 的示例显示在图 4-9 中。

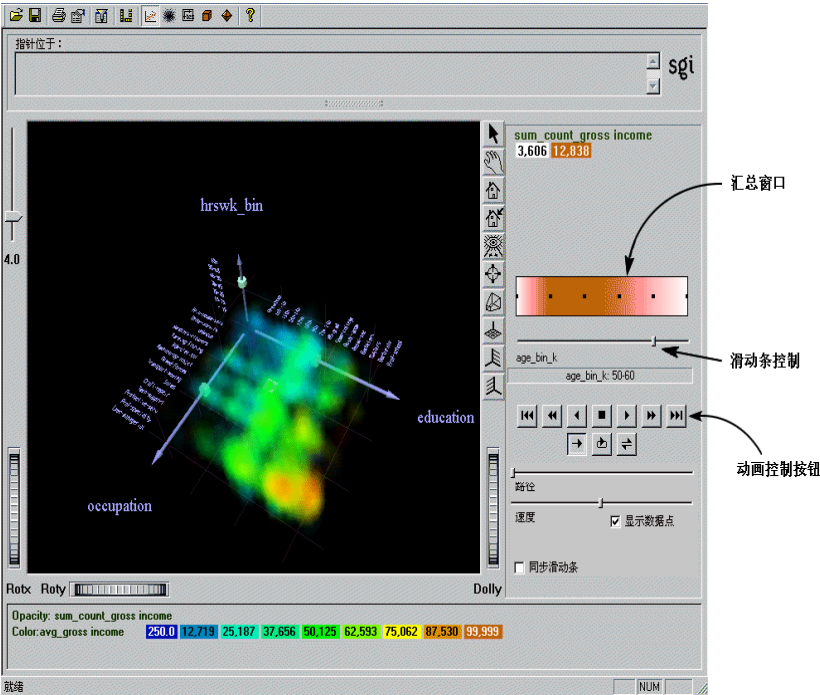


图 4-9 具有汇总窗口和滑动条控制的平伸可视化工具动画控制面板

如果配置文件没有指定滑动条映射，那么将不会出现滑动条控件。

平伸可视化工具中汇总窗口的说明

汇总窗口（请参阅图 4-9）中显示滑动条维中汇总属性的组合。主窗口中的云团状“平伸”代表汇总值的颜色密度。如果没有列被显式映射到汇总，那么计数将被用于显示滑动条上代表数据最多的位置。

在图 4-9 所示的例子中，汇总滑动条白色区域中的黑色圆点表明：当滑动条处于该位置时，左边的平伸图是由 3,606 个记录组成的。滑动条上最红点处的图拥有 12,838 个记录。

平均间隔的黑色圆点指示离散数据点的精确位置。可以取消对动画面板底部“显示数据点”框的选择以隐藏这些黑点。这些位置之间的滑动条位置使用基本数据的内插来产生图象。

汇总窗口中的颜色密度

使用 *adultJobs.splatviz* 文件可以解释颜色密度对值的反映方式。汇总窗口显示的颜色分为从白色（左）到红色（中）再到白色（右）。红色代表数据较多，而白色代表数据较少。在本例中，滑动条中间较高密度的红色表明人口最密的范围是 20-50 岁。

在汇总窗口中创建路径

创建路径的方式与在“散点可视化工具”中使用的完全一样，并且控件也是一样的。但是，当动画演示时，云团的大小和颜色平滑地变化是毫无意义的。状态字段中的信息将显示内插的数据值。当滑动条停止运动时，滑动条位置将继续向前，直到最近的、没有使用内插数据值的离散数据点。

汇总滑动条上的每个分组位置都有一个表。表的每一行（原始数据的组合）定义场景中的一个云团状“平伸”。汇总上对应于邻近分组的表的行数不必相同，因为位置之间的数据分布是不同的。例如：如果图 4-2 上的可视化处理通过将滑动条缺口移到右边（请参阅图 4-10）而从显示 40-50 岁变为显示 50-60 岁，那么一些位置就会显示以前未出现的云团，反之亦然。

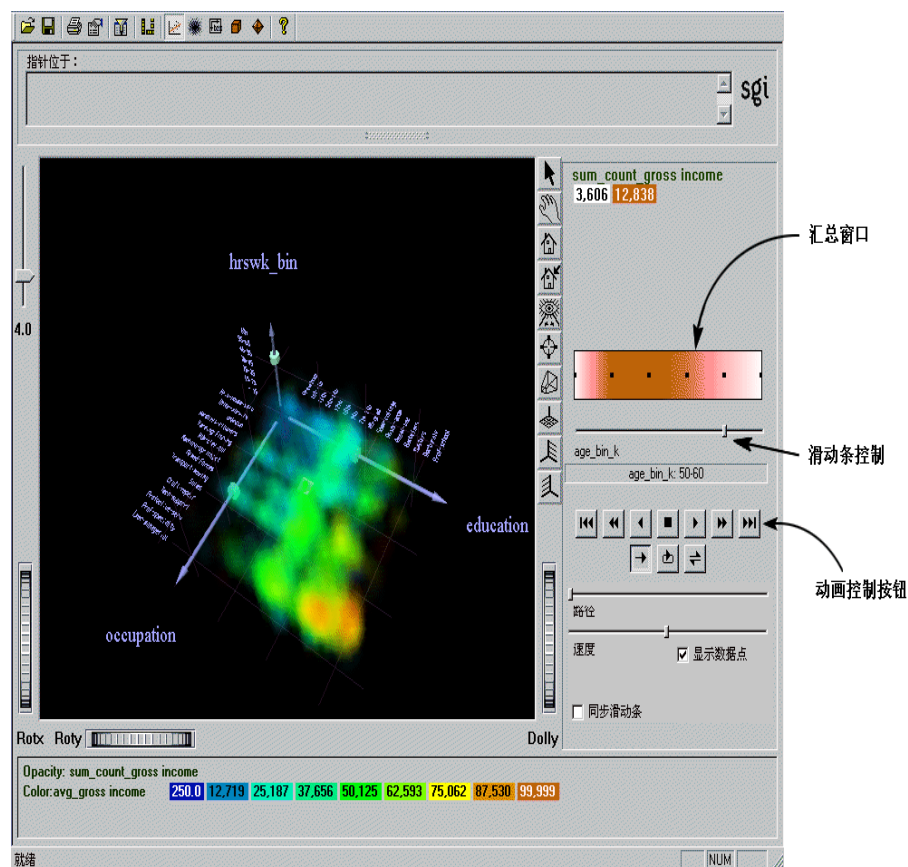


图 4-10 滑动条移动后改变的可视化处理（请与图 4-2 比较）

对于一维滑动条上的内插，将会合并两个邻近的表，然后使用空间列作为唯一的关键字对其进行组合。权重是通过简单内插得到的（如果表中缺少特定的行，那么将假定权重为 0）。用于颜色的平均值也是内插得到的，都是由权重进行加权。

关于内插技术细节的示例，请参阅《MineSet 3.0 企业版参考指南》中的“平伸可视化工具”。

处理散点和平伸可视化结果

可以通过五个下拉菜单获得 “散点” 和 “平伸” 可视化工具的功能。它们是 “文件”、“查看”、“选项”、“形状”、和 “帮助”。如果在没有指定配置文件的情况下运行可视化工具，那么只有 “文件” 和 “帮助” 菜单可以使用。一旦打开文件，就可以使用其他菜单。“文件” 和 “帮助” 菜单是标准菜单，关于详细信息，请参阅 [《MineSet 3.0 企业版参考指南》](#)。

更改画面

通过 “查看” 菜单可以控制可视化工具窗口中显示的特定部分。图 4-11 所示为 “散点” 和 “平伸” 可视化工具的 “视图” 菜单。



图 4-11 散点和平伸可视化工具的查看菜单

下表中对这些选项进行了说明：

表 4-5 散点和平伸可视化工具的查看菜单选项

选项	描述
筛选面板	打开 “筛选” 面板（图 4-12）以处理画面内容
设置背景颜色	打开颜色选择器以指定新的画面背景颜色
显示窗口控件	显示或隐藏主窗口周围的外部控件
动画面板	显示或隐藏所有动画控件面板
空位置	显示或隐藏坐标轴上具有空位置或未知位置的实体



图 4-12 散点和平伸可视化工具筛选面板

如果希望根据一定的准则筛选显示的数据量，可以使用“筛选”对话框，如图 4-12 所示。在“视图”菜单中选择“筛选”面板。可以使用上面的窗格根据字符串列进行筛选。要选择一个值，可以单击它。要取消对一个值的选择，只需再次单击它。要选择列中所有的值，可以单击全部选中按钮。要清除当前所有选择，可以单击清除按钮。

可以使用底部的窗格同时根据字符串列和数值列的值进行筛选。关于详细信息，请参阅《*MineSet 3.0 企业版参考指南*》中的“筛选面板”。

利用下方的“仅显示筛选集”，可以指定主窗口中的景观是包括整个数据集的范围，还是只包括筛选后的数据。

单击应用按钮，开始筛选。如果在面板处于活动状态时按下回车键，那么筛选将自动开始。

在散点可视化工具中选择和追溯

可以使用“散点可视化工具选项”菜单追溯基本数据。

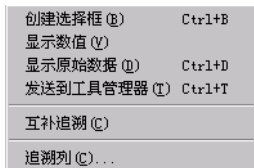


图 4-13 散点可视化工具选项菜单

- *创建框选择器*创建一个三维的框选择器，可以拉伸并移动它以选择体积区域。当它处于活动状态时，“记录查看器”表将打开，其中显示内部实体组合数据的有关信息。请确保光标处于“选择”模式。要移动选择框，可以用鼠标左键单击其中一面并沿需要的方向拖动。拖动时按住 **Shift** 键可以将运动限制在离拖动最近的坐标轴上。要更改选择框的大小，可以沿需要的方向拖动灰色尺度卡。不能在体积界限之外调整大小或移动。灰色尺度卡将不断调整大小以保持固定的屏幕大小。如果觉得它们显得太大，可以将其缩小，这样它们会减小相对于框的大小。
- *显示值*可以显示“记录查看器”表中选定实体的值。
- *显示原始数据*可以检索并显示对应于选定实体的记录。生成的记录显示在“记录查看器”表中。如果没有选择任何实体，那么该菜单项目为灰色。
- *发送到工具管理器*将根据当前选择在“工具管理器”历史表的开始处插入一个筛选操作。用于追溯的实际表达式由当前选项确定。如果没有选择任何实体，那么该菜单项目为灰色。
- 使用 *互补追溯*可以使显示原始数据和发送到工具管理器选项读取所有没有被选择的数据。

- **追溯列**将打开一个面板，可以用来选择用于追溯的列。与其他可视工具不同，数据中没有特定的列被指定为关键数据。“散点可视化工具”不可能确定用户在追溯表达式中需要的列。例如：数据可能是汽车的商标、型号和重量。可能需要追溯原始数据，指定应该考虑商标和型号，而不考虑重量。默认情况下，映射到图形要求的所有列在追溯中都被认为是有意义的。其他的列则不是这样，但是可以在“特性”对话框中突出显示它们来达到此目的。

关于详细信息，请参阅 *《MineSet 3.0 企业版参考指南》* 中的“追溯”。

在平伸可视化工具中选择和追溯

“平伸可视化工具选项”菜单与“散点可视化工具”在用于追溯基本数据的选项方面稍有不同。

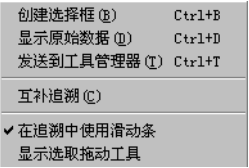


图 4-14 平伸可视化工具选项菜单

- **创建框选择器**创建一个三维的框选择器，可以拉伸并移动它以选择体积区域。当它处于活动状态时，“记录查看器”表将打开，其中显示内部实体代表的所有组合数据的有关信息。关闭该窗口将清除当前所有选择。所有选择框内的实体或使用 **shift-** 单击选择的实体将显示在表窗口中。要移动选择框，可以用鼠标左键单击其中一面并沿需要的方向拖动。拖动时按住 **Shift** 键可以将运动限制在离拖动最近的坐标轴上。要更改选择框的大小，可以沿需要的方向拖动灰色尺度卡。不允许在体积界限之外调整大小或移动。灰色尺度卡将不断调整大小以保持固定的屏幕大小。如果觉得它们显得太大，可以将其缩小，这样它们会减小相对于框的大小。

- 显示原始数据可以检索并显示对应于“框选择器”选定的记录。生成的记录显示在“记录查看器”表中。
- 发送到工具管理器将根据当前框选择在“工具管理器”历史的开始处插入一个筛选操作。用于追溯的实际表达式由当前框选择的范围确定。如果没有选择任何实体，那么将出现一个警告。
- 使用互补追溯可以使显示原始数据和发送到工具管理器选择读取所有没有被选择的数据。
- 在追溯中使用滑动条在追溯时使用滑动条。
- 显示选取拖动工具可以开关选取拖动工具的可见性（默认情况下为开）。当框选择开始时，选取拖动工具被删除，但是它可以在框选择处于活动状态的同时被激活。

关于详细信息，请参阅 [《MineSet 3.0 企业版参考指南》](#) 中的“追溯”。

用“形状”菜单更改散点可视化工具画面

可以使用“形状”菜单更改“散点可视化工具”中绘制散点的方法。以下的不透明图元可以用于代表散点。

- 立方体将绘制一个不透明的立方体，它的体积与映射到大小可视元素的属性成正比。
- 菱形将绘制一个线框三角形。它的方向随其颜色变化，并且它的大小与映射到大小可视元素的属性成正比。
- 球形将绘制一个不透明的球体，它的体积与映射到大小可视元素的属性值成正比。
- 条形将绘制拉长的直棒，它的高度由映射到大小可视元素的属性值决定。

用“形状”菜单更改平伸可视化工具画面

可以使用“平伸可视化工具形状”菜单更改绘制平伸的方法。可以选择交互的交换准确性。在所有近似方法中，纹理是理想 高斯型 密度分布最准确的表示法。因为大部分计算机对硬件辅助结构映射都支持的很好，因此纹理通常是最好的选择。三种平伸图元类型是：

- 线性形将绘制与高斯型平伸线性近似的一小组三角形。
- 高斯形将绘制与高斯型平伸近似的一大组三角形。
- 纹理形使用纹理映射三角形绘制最准确的表示法。在不支持硬件辅助结构映射的系统上，该过程可能会非常慢。

另外还可以选择以下的不透明图元代表平伸图元：

- 立方体将绘制一个不透明的立方体，它的体积与映射到不透明度可视元素的属性（如果有的话）或计数成正比。
- 菱形将绘制一个线框三角形。它的方向随其颜色变化，并且它的大小与映射到不透明可视元素的属性（如果有的话）或计数成正比。
- 球形将绘制一个不透明的球体，它的体积与映射到不透明度可视元素的属性（如果有的话）或计数成正比。

用树可视化工具检查数据

本章包括以下几个部分：

- 第 81 页的 “树可视化工具概述”
- 第 82 页的 “运行树可视化工具”
- 第 86 页的 “用树可视化工具查看结果”
- 第 97 页的 “调整树可视化工具画面”

提供了一些样本配置和数据文件，用于演示 “树可视化工具” 的特性和功能。这些文件在 *MineSet 3.0* 下的 *examples* 目录中，位于 *MineSet* 的初始安装目录。

树可视化工具概述

“树可视化工具” 是一个图形接口，它将数据显示为三维景观。它将数据以树的形式分层显示。树的每一层根据不同属性的值进行分支。每个树节点都显示一个代表它是所有子树数据的图形。图形由基准块组成，基准块的高度和颜色取决于数据属性。每个基准块上条形和 / 或盘的编号、标签、高度和颜色也由指定的属性确定。

如图 5-1 所示，“树可视化工具” 将数据的定量特征和相关特征显示为分层连接的节点。每个节点由条形和盘组成，它们的高度和颜色对应于数据值的组合（通常为总和、平均和计数）。连接节点的线称为边，它显示数据集与其子集的关系。

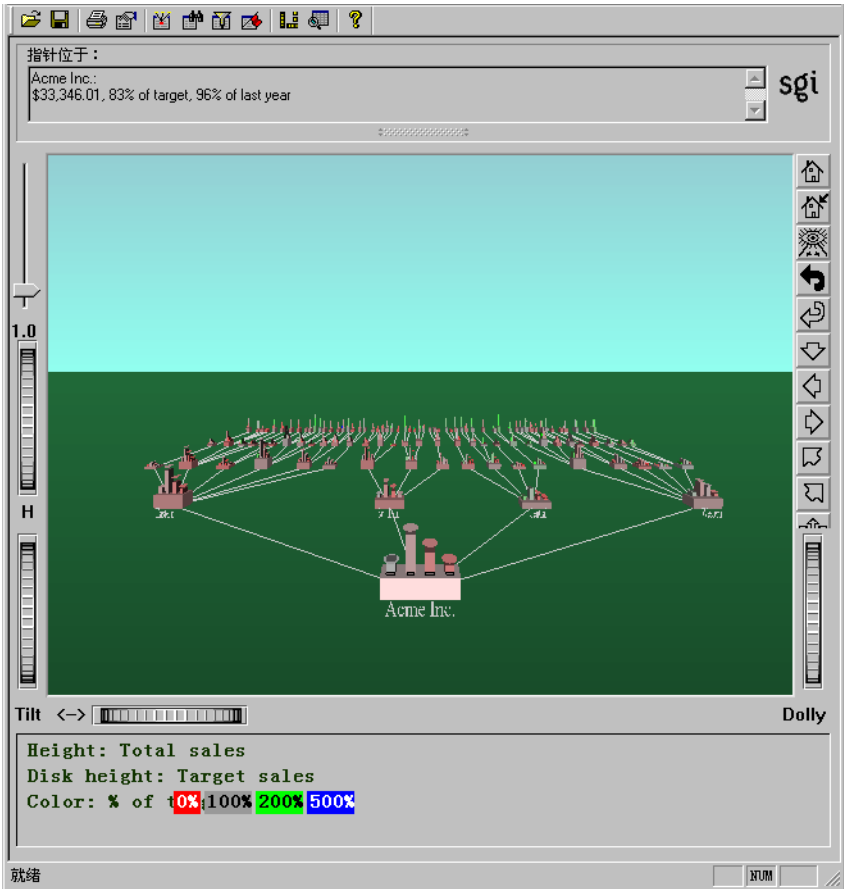


图 5-1 树可视化工具主窗口的画面示例

运行树可视化工具

本部分描述如何使用“工具管理器”配置“树可视化工具”。虽然“工具管理器”大大简化了“树可视化工具”的配置任务，但是您也可以使用文本编辑器手工建立配置文件。还有其他多种运行“树可视化工具”的方法，关于详细信息，请参阅 [《MineSet 3.0 企业版参考指南》](#)。

要从“工具管理器”中运行“树可视化工具”：

1. 选择“文件”>“连接到服务器”，然后登录到服务器。
2. 选择“文件”>“打开新的数据文件”，并选择或键入所需的文件名。
3. 在“数据目标”面板中，单击“可视化工具”选项卡（图 5-2）。
4. 从工具选项卡中，选择“树可视化工具”选项卡。
5. 用可视元素旁边的下拉菜单将指定属性从数据集映射到“数据目标”中适当的可视元素。关于映射的详细信息，请参阅第 84 页的“映射树可视化工具可视元素”。
6. 单击“工具选项”按钮，在“树可视化工具配置选项”面板中指定“高度和颜色组合”。这一步对于获得有意义的可视化处理至关重要。

默认情况下，父节点条形的高度（或颜色）代表所有子节点条形值的总和；但是，对于某些数据集，平均值、最大值、最小值或计数可能是更好的选择。组合可以被指定为条形、基准块或盘的高度或颜色。关于组合和其他内容的详细信息以及详细的高级选项，请参阅《*MineSet 3.0 企业版参考指南*》。

7. 单击调用工具，运行可视化工具。

“工具管理器”窗口底部的“状态”框将显示进程与统计结果。

映射树可视化工具可视元素

当“树可视化工具”被选定时，“数据目标”面板将显示一些可视元素，可以将数据中的属性映射到这些可视元素（图 5-2）。从每个可视元素旁边的下拉菜单中选择列。
表 5-1 列出可视元素以及其用法的说明。通常，使用“树可视化工具”的最好方法是用元素映射来进行试验，直到找到属性和可视元素的最好组合。

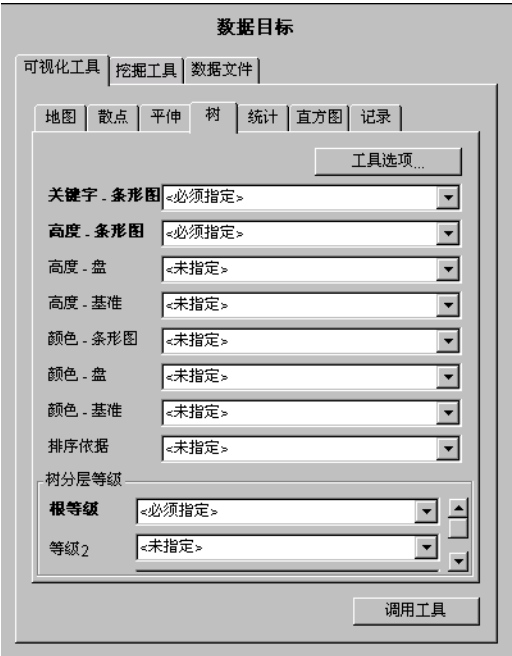


图 5-2 选定树可视化工具的工具管理器数据目标窗格

表 5-1 列出了可视元素。

表 5-1 树可视化工具可视元素

可视元素	说明
关键字 - 条形	用它来指定条形代表的值。如果列是数值型的，通常最好对列进行分组。否则， MineSet 将为每个属性值创建一个条形，这将使可视化处理难于使用。
高度 - 条形	用它来指定条形高度代表的值。通常，条形越高，代表的值就越大。查看工具选项面板中关于每级组合属性的选项。
高度 - 盘	用它来指定盘高度代表的值。可选的盘与条形放在同一个位置。如果盘属性的量度单位与条形高度属性的一样（例如：今年的销售额和去年的销售额），那么盘最有用。如果没有指定映射，则不会显示任何盘。
高度 - 基准	用它来指定基准块高度代表的值。如果没有指定映射，缺省使用条形高度映射。查看 “工具选项” 面板中关于每级组合属性的选项。
颜色 - 条形	用它来指定条形颜色代表的值。可以通过 “工具选项” 面板分配指定的颜色（请参阅 《MineSet 3.0 企业版参考指南》中的 “颜色选择”），或允许 MineSet 自动分配颜色。查看工具选项面板中关于每级组合属性的选项。
颜色 - 盘	用它来指定盘颜色代表的值。可以通过 “工具选项” 面板分配指定的颜色（请参阅 《MineSet 3.0 企业版参考指南》中的 “颜色选择”），或允许 MineSet 自动分配颜色。 只有当指定了盘高度时，该选项才有效。

表 5-1 树可视化工具可视元素	
可视元素	说明
颜色 - 基准	用它来指定基准块颜色代表的值。可以通过 “工具选项” 面板分配指定的颜色（请参阅《MineSet 3.0 企业版参考指南》中的“颜色选择”），或允许 MineSet 自动分配颜色。查看工具选项面板中关于每级组合属性的选项。 如果没有指定映射，就使用条形图颜色映射。
按 ... 排序	用它来按照选定属性的值对节点的分布分类。默认情况下，排序顺序为从左到右的升序。
树分层根等级	用它来指定如何将来自数据源的表转换为分层结构。树的第二级具有的节点数与选定属性值的数目一样多。 如果属性的可能值有许多，树将很大并且难于分析。对列进行分组可以减轻这种情况。
树分层 2, 3, ... 级	用这些层次来将数据划分到更多的分层结构中。默认情况下，“可视元素” 列表具有三个等级。如果指定了第三个等级，“树可视化工具” 会自动添加第四个等级。每指定一个额外等级的同时会添加另一个。可以根据需要指定任意多的等级。 如果属性的可能值有许多，树将很大并且难于分析。对列进行分组可以减轻这种情况。

用树可视化工具查看结果

当 “树可视化工具” 第一次运行时，分层的根节点位于场景的前面，在主窗口底部附近。根节点的背后是它的子代。每个节点都由带有条形的基准组成。可以通过 “工具管理器” 更改条形高度和颜色代表的值，或者手工更改 `.treeviz` 配置文件（请参阅《MineSet 3.0 企业版接口指南》中的 “树可视化工具”）；通常基准块代表所有条形的组合。基准块用边连接起来，这些边代表节点与其子代的连接。

子集中的值可以自动在相邻的高等级中求和与显示。条形下的基准块可以提供有关所有条形组合值的信息。代表负值的条形显示在基准块的顶部下面。可以通过禁用基准高度面来更清晰地查看负值条形（请参阅第 97 页的“调整树可视化工具画面”，或《MineSet 3.0 企业版参考指南》中的“树可视化工具”）。

树可视化工具示例

图 5-3 显示的是打开 stores.treenviz 文件的场景。该文件在 MineSet 3.0\examples 目录中，位于 MineSet 的初始安装目录下。要想看到它，请从工具管理器“可视工具”菜单中，打开“三维可视化工具”；然后在树可视化工具“文件”菜单中，选择“打开”。数据文件中的每个记录都包含以下列：地区（东部、南部、中部或西部）、州、城市、商店数目、商品类型（用具、服装、电器或家具）、今年的销售额、去年的销售额、以及目标销售额。还有两个新添加的列：今年达到的销售额占目标销售额百分比，和今年达到的销售额占去年销售额的百分比。表 5-2 显示的是列到“树可视化工具”可视元素的映射。

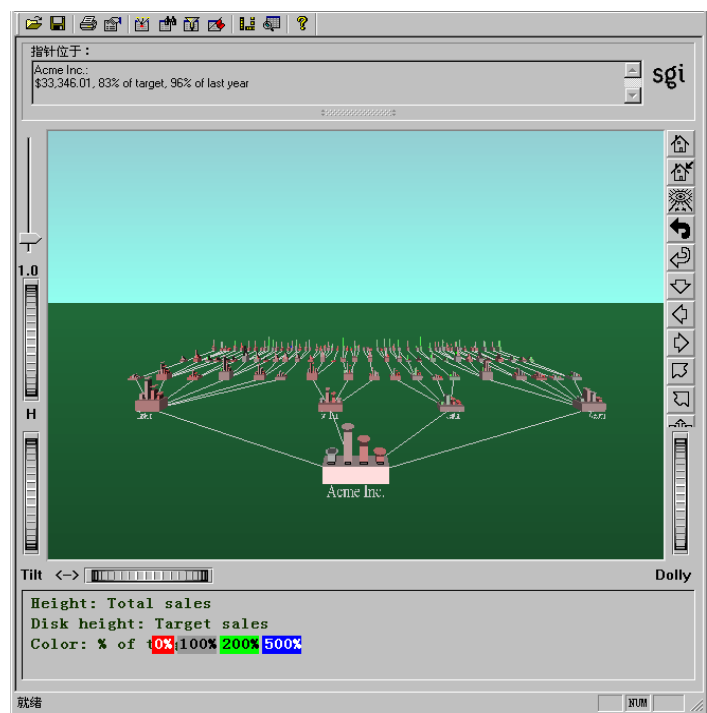


图 5-3 “商店”数据集的树可视化工具初始视图

树顶部节点中的条形代表每种产品的总销售额。盘代表目标销售额。下一级将销售额按地区分开，再下一级按州分开，等等。直到单个的商店。

在检查“树可视化工具”场景时，可以很容易地看到每个地区、州、城市或商店对于每种商品目标销售额的符合程度。还可以按照地区、州、商品类型等等来比较销售额。

表 5-2 显示的是该示例中列到可视元素的映射。

表 5-2 “商店数据”的元素映射

列	组合	映射
地区	无	层次 1

表 5-2 “商店数据” 的元素映射

列	组合	映射
州	无	层次 2
城市	无	层次 3
商店标识号	无	层次 4
商品类型	无	关键字 - 条形
今年的销售额	求和	条形高度
去年的销售额	求和	无
目标销售额	求和	盘高度
目标百分比	无	条形颜色
去年的销售额百分比	无	无

映射是获得有意义的可视化处理的关键。在熟悉该工具以前， 试验是寻找最佳映射模式的最好方法。

在树可视化工具中更仔细地观察数据

有两种方法可以查看每个节点的细节：

- 突出显示节点或对象（条形或基准块），可以查看基本数据。（请参阅“突出显示对象或节点”）
- 选择节点或对象，将观察点缩放到该位置，可以查看基本数据。（请参阅“选择对象”）可以一次选择多个位置。选择一个节点或对象也可以突出显示它。聚光灯可以保持对该位置的跟踪，即使它在场景的远处。

突出显示对象或节点

要突出显示一个对象（条形或基准块），可以将鼠标放在该对象上面。关于该对象的信息将出现在视图区域的左上部，“指针位于”标签的下面。（图 5-4）。要突出显示一个节点并获取关于该节点的信息，可以将指针放在通向该节点的线上。

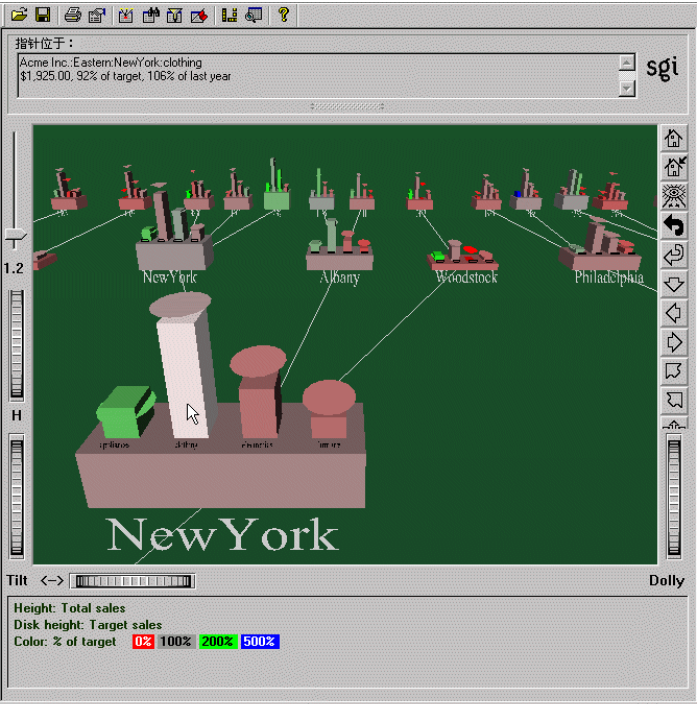


图 5-4 突出显示对象及其基本信息

选择对象

要选择一个对象并将镜头移动到它的位置，可以单击该对象。单击的同时按住 **Shift** 键则选择对象而不直接聚焦到它。只要选定了对象，就会显示信息。

如果单击对象的同时按住 **Ctrl** 键，就可以锁定对该对象的选择。如果对象当前没有被选定，单击选择它，反之亦然。使用该技术可以同时选择多个对象。“指针位于”字段中只显示最后一个所选对象的信息，可以选择“选项” > “显示数值”或追寻到所选对象背后的原始数据，从而看到所有被选的值。

当选择对象时，在它的上面会出现一个白色的聚光灯（图 5-5）。在搜索时，会出现一个黄色的聚光灯（请参阅第 93 页的“用“搜索”面板查找指定的对象”）。聚光灯是可见的，即使选定的对象是远处背景中的一个子节点。

聚光灯的边是对象的代用品：当指针移到聚光灯的边上时，相关对象就被突出显示，有关信息会出现在视图的左上角。单击聚光灯的边，选择相关的对象并（如果没有按下 **Shift** 键）直接聚焦到它。聚光灯只有在边上实线部分是活动的，而在中央透明部分是非活动的。在透明区域单击可以选择聚光灯后面的对象。

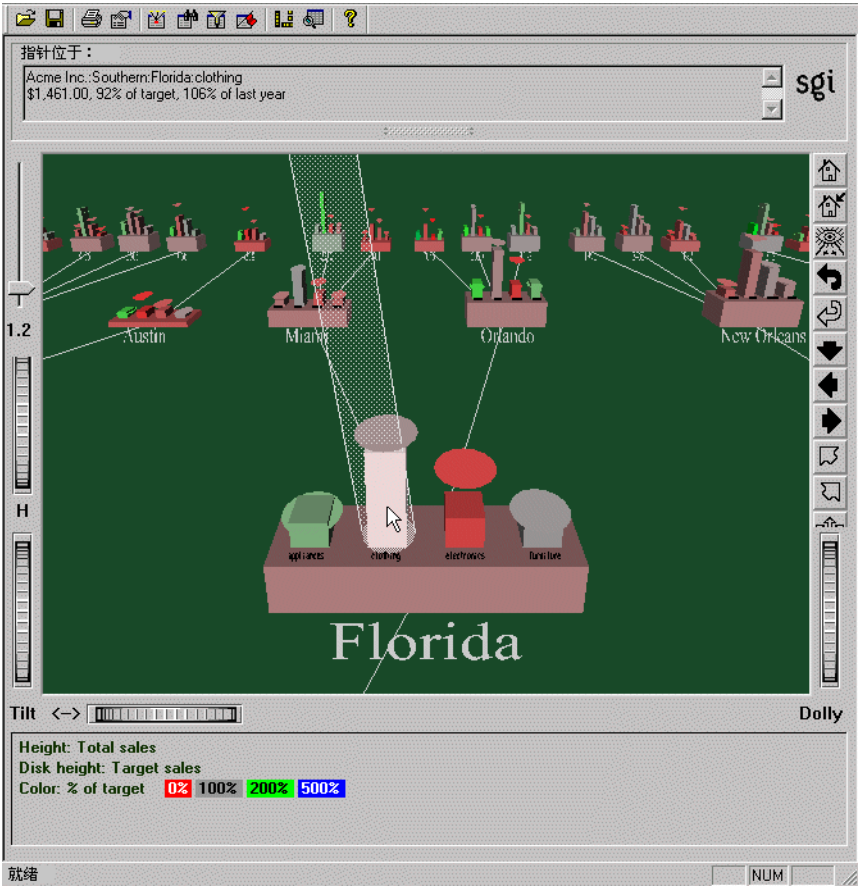


图 5-5 选定（被聚光的）对象的示例

用全局观察窗口观察整个图片

要打开带有整个分层的俯视图窗口（图 5-6），可以选择“视图”>“全局观察”。如果需要在每次查看场景时显示“全局视图”，请设置配置文件中的“overview”选项（请参阅《MineSet 3.0 企业版接口指南》中的“全局视图”一节）。

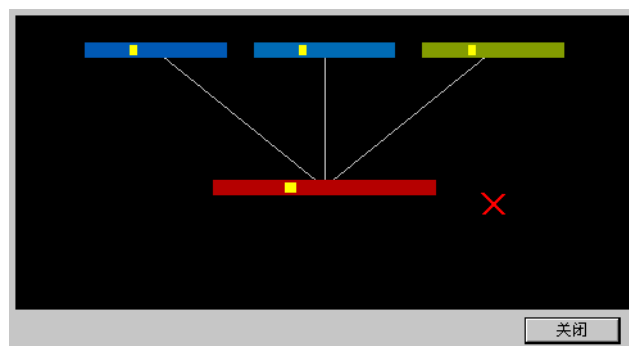


图 5-6 树可视化工具的全局观察窗口

全局观察窗口中的“X”表示当前的位置。全局视图可以帮助跟踪您在整个场景中的位置和观察点。它还有助于迅速到达指定的位置：

- 要在全局视图中选择节点、将主视图直接聚焦到它并突出显示它，可以单击该节点。这类似于在主窗口中单击节点。
- 要将视图直接聚焦到新的位置，即使在该位置没有节点存在，也可以在该位置同时单击鼠标左右键（或鼠标中键）。

用“搜索”面板查找指定的对象

要在“树可视化工具”窗口中搜索指定的对象，可以选择“查看”>“搜索面板”。填入搜索规则，然后单击“搜索”。关于“搜索”面板的详细信息，请参阅 [《MineSet 3.0 企业版参考指南》](#)。

一旦搜索完成，黄色聚光灯将突出显示符合搜索规则的对象（请参阅图 5-7）。要显示黄色聚光灯下面对象的有关信息，可以将指针移到该聚光灯上；信息将出现在左上角，“指针位于”标签下。要选择并直接聚焦到黄色聚光灯下的对象，可以单击聚光灯；如果单击的同时按住了 Shift 键，将不会进行聚焦。

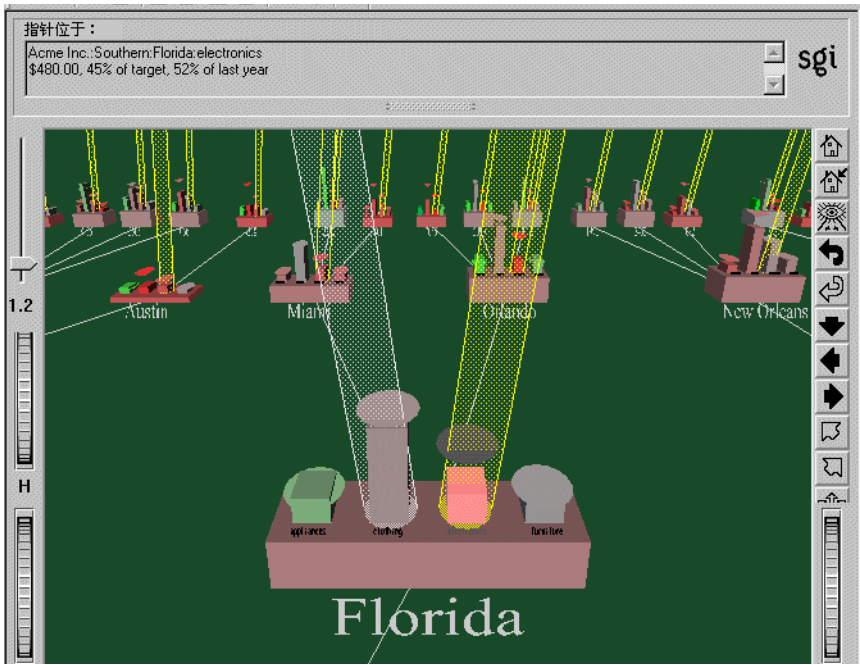


图 5-7 树可视化工具中的搜索结果示例

用“标号”面板标记重要的位置

“标号”面板可以命名并保存重要的位置（或观察点）以备将来参考。所有标号都可以用主视图中的彩色标记指示。如果标号代表选定的对象，就将标记放在该对象上。如果它代表观察点的位置，就将标记放在该位置。要转到标号，可以单击标记。可以通过选择“显示”菜单中的“标号标记”打开和关闭所有标记（请参阅标号标记，在[第 97 页](#)的“调整树可视化工具画面”小节内）。

要保存位置并用标号标记它（可选）：

1. 在树可视化工具“视图”菜单中，选择“标号面板”。将打开一个窗口，如[图 5-8](#)所示。如果还没有放置任何标号，列表将是空的。



图 5-8 树可视化工具 “标号面板”

- 2. 单击标号按钮，标记主窗口中的当前位置。将显示一个对话框，如图 5-9 所示。
- 3. 输入希望用于标号的名称，并选择一种颜色。默认名称为当前选定对象的名称。选择的颜色将控制出现在主窗口中标记的颜色，并且代表该标号。



图 5-9 “选择标号” 对话框

图 5-10 所示为一个主窗口样例，其中的标记代表创建的标号。

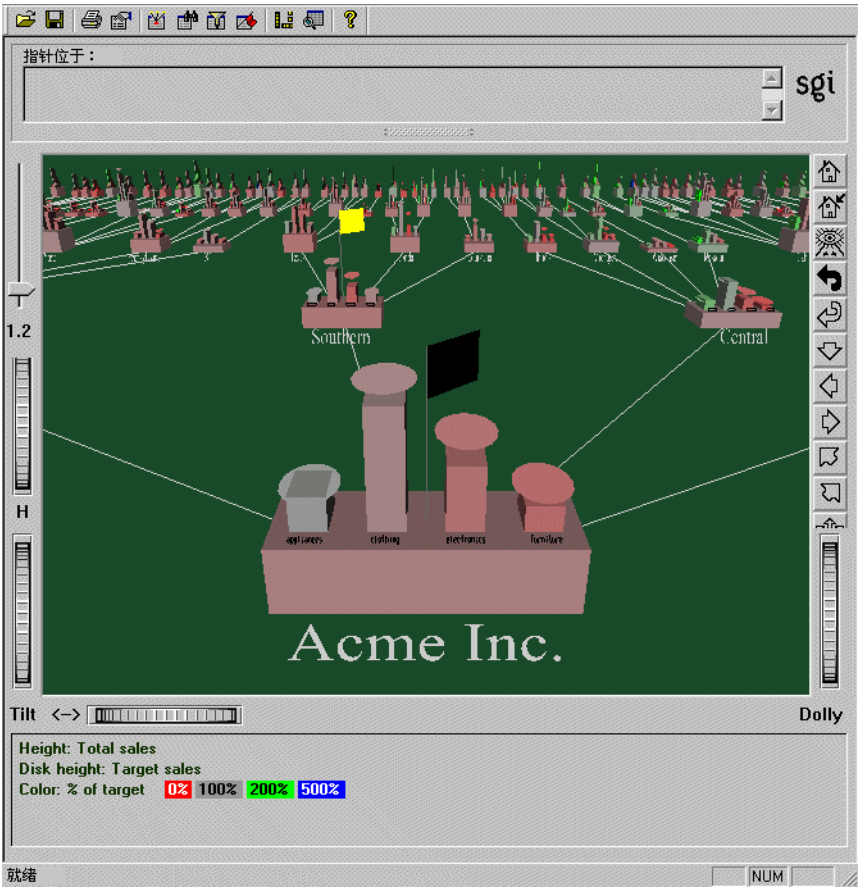


图 5-10 带有标号标记的主窗口

“标号”窗口边上的按钮可以执行以下任务：

- 要转到面板中选定标号代表的位置，可以单击“标号”面板中的跳转到按钮。双击面板中的标号可以产生同样的效果。如果标号选定的对象已不存在（它可能已被筛选掉，或自标号创建以来数据发生过更改），那么显示的位置将在原来对象位置的附近。
- 要删除面板中选定的标号，可以单击删除按钮。

- 要更改面板中选定标号的名称或颜色，可以单击修改按钮。
- 要将面板中选定的标号按列表顺序向上移动，可以单击向上按钮。
- 要将面板中选定的标号按列表顺序向下移动，可以单击向下按钮。
- 要退出“标号”面板，可以单击关闭按钮。

存储标号信息的文件与配置文件具有相同的名称，后缀为 *.marks*。只要标号发生更改，就会被保存到该文件中。如果所有标号都被删除了，*.marks* 文件也将被删除。如果标号更改无法保存（例如：由于权限错误），就会出现一个警告；此后再进行的标号更改尝试时，该警告将不再重复出现。

用“筛选”面板筛选数据

“筛选”面板将删除选定的信息，因此可以调整显示层次。您可以使用“筛选”面板强调指定的信息，或压缩数据量以获得更好的性能。要访问“筛选”面板，可以从“树可视化工具查看”菜单中选择它。关于使用“筛选”面板的详细信息，请参阅 *《MineSet 3.0 企业版参考指南》* 中的“树可视化工具”。

调整树可视化工具画面

表 5-3 描述了用于调整“树可视化工具”画面的多个选项。这些树可视化工具“画面”菜单中的选项可以控制一些显示参数。

表 5-3 树可视化工具显示参数

选项	描述
基准高度（开关）	打开或关闭基准高度。要查看负数，或想使条形高度互相比较变得更容易，可以关闭该选项。打开该选项可以提供所有条形的汇总信息。该开关的初始值可以用配置文件中的语句“BASE HEIGHT”来更改。
标号标记（开关）	打开和关闭代表标号的标记（请参阅“用“标号”面板标记重要的位置”）。

表 5-3 树可视化工具显示参数

选项	描述
零位置（子菜单）	控制零高度对象的显示方式。默认情况下，它们的显示与其他对象类似：高度为零的实心立方体（平面）。子菜单可以指定将它们显示为轮廓（空心的正方形），或完全隐藏起来（不画）。该开关的初始值可以用配置文件中的选项“Zero”来更改。（请参阅《MineSet 3.0 企业版参考指南》）。
空位置（子菜单）	控制空高度对象的显示方式。它的选项与零菜单中的一样；但是空选项的默认值是将对象显示为一个轮廓。该开关的初始值可以用配置文件中的选项“null”来更改。（请参阅《MineSet 3.0 企业版参考指南》）。

用地图可视化工具检查数据

地图可视化工具在三维景观上用条形图或可识别的地理形状来显示数据。当数据具有地理背景或依附一些特定的地形布局时，该工具很有用。本章主要包括以下几个主题：

- 第 100 页的 “地图可视化工具概述”
- 第 102 页的 “运行地图可视化工具”
- 第 103 页的 “准备数据”
- 第 108 页的 “查看地图可视化工具”
- 第 110 页的 “在地图可视化工具中创建动画”
- 第 110 页的 “处理地图可视化工具结果”

我们提供了一些配置文件和数据文件样例，用于示范 “地图可视化工具” 的特性和功能。这些文件保存在 *MineSet 3.0\examples* 目录中，位于 **MineSet** 的初始安装目录。

地图可视化工具概述

“地图可视化工具”可以用来观察与空间相关的数据。您可以使用 “地图可视化工具” 在三维场景中漫游，可以对数据进行概化上寻和细化下寻，可以全局观察数据或进一步查看局部细节，还可以使用动画来了解数据在一维或二维独立空间中的变化方式。这里，所谓独立的维度是指任意像年龄或年份这样与其他无关的变量。当数据集中包含有映射到滑动条的独立变量时，动画面板会出现在主窗口右侧。（如 图 6-1 所示）

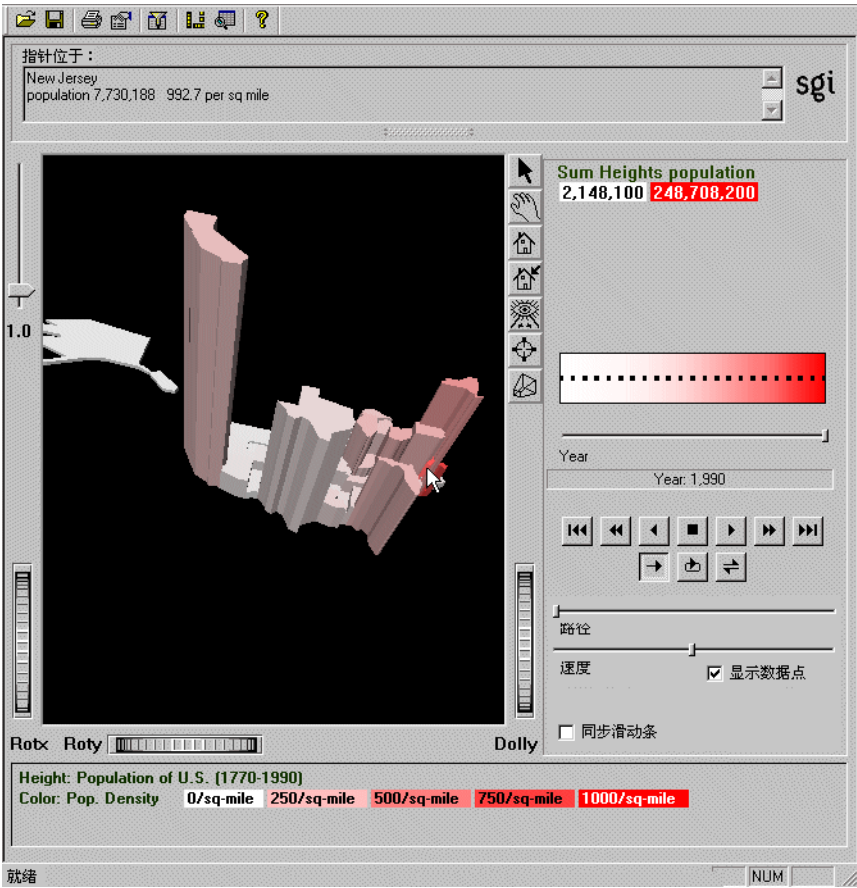


图 6-1 使用地理形状的地图可视化工具样例

场景还可以由描绘成简单轮廓的地理对象组成，在特定的位置放有“条形图”圆柱体。（请参阅图 6-2）。

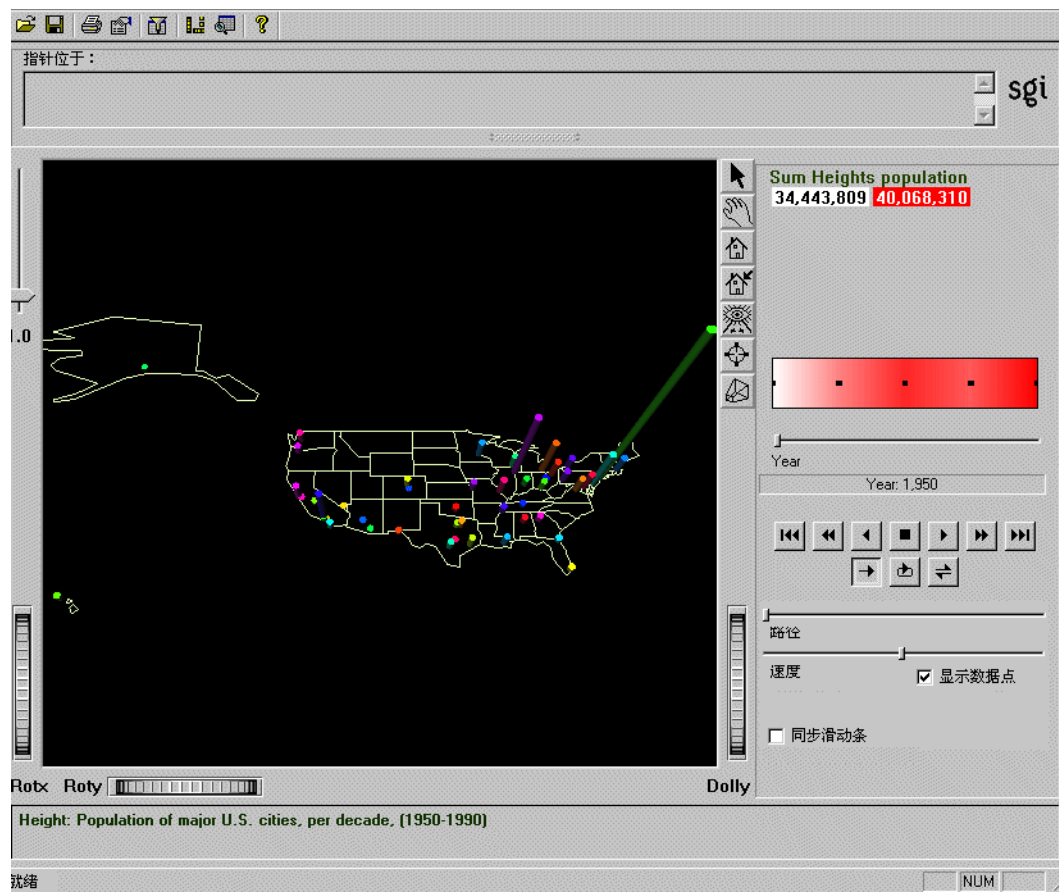


图 6-2 带有地理轮廓上相关群体条形图的地图可视化工具

另一个可能的场景会在地图上显示连接着特定位置的线，每条线具有各自的宽度和颜色（请参阅图 6-3）。这些线的特征对应于普通对象和圆柱体的高度和颜色属性。



图 6-3 显示带有指定端点的美国地图的可视化工具样例

运行地图可视化工具

以下是运行“地图可视化工具”最常用的方法：

1. 打开“工具管理器”，选择服务器和数据源。如果需要帮助，请参阅第 2 章，“[利用 MineSet 访问数据](#)”。本示例使用“客户波动”数据集。
2. 在“工具管理器数据目标”面板中单击“可视化工具”选项卡；然后在下面一行选项卡中，单击“地图可视化工具”。

准备数据

本部分讲述转换地理数据的步骤，以便“地图可视化工具”进行显示。

组合数据

在地图上描绘数据之前，必须先将数据转换到较小的数据集。这种转换叫做组合。

1. 在“工具管理器数据转换”面板中单击组合。



图 6-4 在地图可视化工具中组合

2. 将感兴趣的列移到“要组合的列”列表中，然后使用底部的复选框标识需要进行的数学操作。将空间分组保留在中间的“分组索引列”字段中，然后将其他所有的列从可视化处理中移除。关于操作的说明，请参阅[第 41 页](#)上的[图 3-5](#)。

例如：使用 *客户波动* 数据集，在“组合”对话框中，单击列，然后再单击向左箭头，将其国际电话总数（**total intl calls**）移到左边列表中。在底部复选框中选中 *平均和计数*，关闭 *总和*。将州（**STATE**）保留在中间列表中，然后将其余所有项目移到右边列表。（按住 **Ctrl** 键可以选中多个列。）单击 *确定* 以应用您的选择。

下一步是查找适当的空间形状。

选择地图形状

1. 在“数据目标”面板中，单击“可视化工具”选项卡上的 *工具选项* 按钮，访问地理文件。“工具管理器”将显示“地图可视化工具选项”对话框。
2. 单击“实体”文本字段右侧的按钮，选择一个地理实体文件。这些文件在 **MineSet** 的初始安装目录中，位于 *MineSet 3.0\config\mapviz\gfx_files* 下。本示例使用 *usa.state.hierarchy*。
3. 单击 *确定*，找到选定的文件，然后再次单击 *确定*，关闭“地图可视化工具选项”面板。关于 *.gfx* 文件的详细信息，请参阅《*MineSet 3.0 企业版参考指南*》中的“可视化工具”一节。

下一步是将可视元素与列相链接。

链接地图可视化工具元素

将地图元素链接到可视实体时，一些列可能需要提前进行处理才能适合可视元素的要求：

- 如果列中拥有太多的离散值，请参阅第 3 章中的“为列改变或创建新的分组”。
- 如果列的类型不正确，请参阅第 3 章中的“更改列类型或名称”。



图 6-5 将列映射到地图可视化工具的可视元素

1. 将工具管理器“数据转换”窗格中的列映射到“数据目标”窗格中的实体。在每个实体旁边的弹出菜单中进行选择以确定映射项目。
对于该示例，为实体 - 条形图选择 `state`，为高度 - 条形图选择 `count_total intl calls`。
关于映射可行性的详细信息，请参阅[“映射到滑动条”](#)和[“映射到地图”](#)。
2. 单击 *调用工具*，查看数据的地理分布。

映射到滑动条

当您觉得一个值根据特定准则变化（如收入随年龄变化）时，可以将该列映射到滑动条，而将其他列映射到别的可视元素。例如：[图 6-1](#) 中有一个映射到列“年”的滑动条。这样可以通过移动滑动条观察数据的逐年变化。

如果列是（整数、浮点数、双精度类型）数值型的或被分组过的，就可以被映射到滑动条。如果列已经被分组，那么它的名称后就会带有 `_bin` 字样。列类型注释位于“当前列”字段中的列名称后，例如：`total day calls - double`。

可以自动或手工创建滑动条。当把列映射到滑动条时，“工具管理器”将自动生成滑动条，滑动条通过自动分组和组合来创建。请参阅词汇表中的词条：[分组](#)和[组合](#)。这些自动操作发生在单击[调用工具](#)之后。

如果在当前数据表中有一个数组列，“工具管理器”可以创建对应于该数组列下标的滑动条。“工具选项”对话框中指出了当前滑动条下标。要显示它们，可以在“工具管理器数据目标”窗格中单击“可视化工具”选项卡，然后再单击[工具选项按钮](#)。当前表中的所有数组列都必须具有相同的下标；否则，将不会创建滑动条。

关于创建数组列的详细信息，请参阅第 3 章中的[“通过组合创建新列”](#)和《[MineSet 3.0 企业版参考指南](#)》中的[“组合”](#)。

关于自动和手工建立滑动条的详细信息，请参阅《[MineSet 3.0 企业版参考指南](#)》中的[“地图、散点、平伸可视化工具中的滑动条创建”](#)。

映射到地图

MineSet 提供了多种可识别地形和其他空间实体组成的 `.gfx` 和 `.hierarchy` 文件。除非将实体条映射到一个图形元素，否则无法得到可识别的地图。如果希望显示其他可视实体或地图，可用以下步骤取代第 104 页的[“选择地图形状”](#)：

1. 在“工具管理器可视化工具”选项卡中，单击[工具选项按钮](#)。
2. 在对话框中选择所需的选项。“地图可视化工具选项”对话框，如[图 6-6](#) 所示。
 - 对于地图形状的实体，可以单击“实体文件”文本字段右边的[查找文件按钮](#)，然后选择一个在“地图可视化工具”主窗口中用于代表对象的 `.hierarchy` 地理实体文件。可以在 `/config/mapviz/gfx_files` 中找到这些示例。

- 对于矩形，可以单击“轮廓文件”文本字段右边的查找文件按钮，指定要描绘的轮廓对象。数据将表现为一个放置有三维实体的平面。

“实体文件”和“轮廓文件”字段是可选的。如果没有提供“实体文件”，可以看到实体对象由任意摆放在场景中的矩形组成。

关于其他选项的描述，请参阅《*MineSet 3.0 企业版参考指南*》中的“可视化工具”。



图 6-6 “地图可视化工具选项”对话框

3. 完成了对“工具选项”对话框的更改后，请单击确定，关闭“工具管理器”窗口。然后可以单击调用工具。

查看地图可视化工具

与所有可视化工具一样，如果在没有指定配置文件的情况下运行了“地图可视化工具”，那么可视化工具主窗口将是空的，并且只能使用“文件”、“查看”和“帮助”下拉菜单。要显示所有菜单和控制，请使用“文件”>“打开”，以查看配置文件列表。这些文件在 *MineSet 3.0\examples* 目录中，位于 MineSet 的初始安装目录。

如果已经指定了有效的配置文件，可视化工具将会显示地理景观。例如：图 6-7 所示的是指定 *population.usa.mapviz* 并将年份滑动条移到最右边时得到的结果。

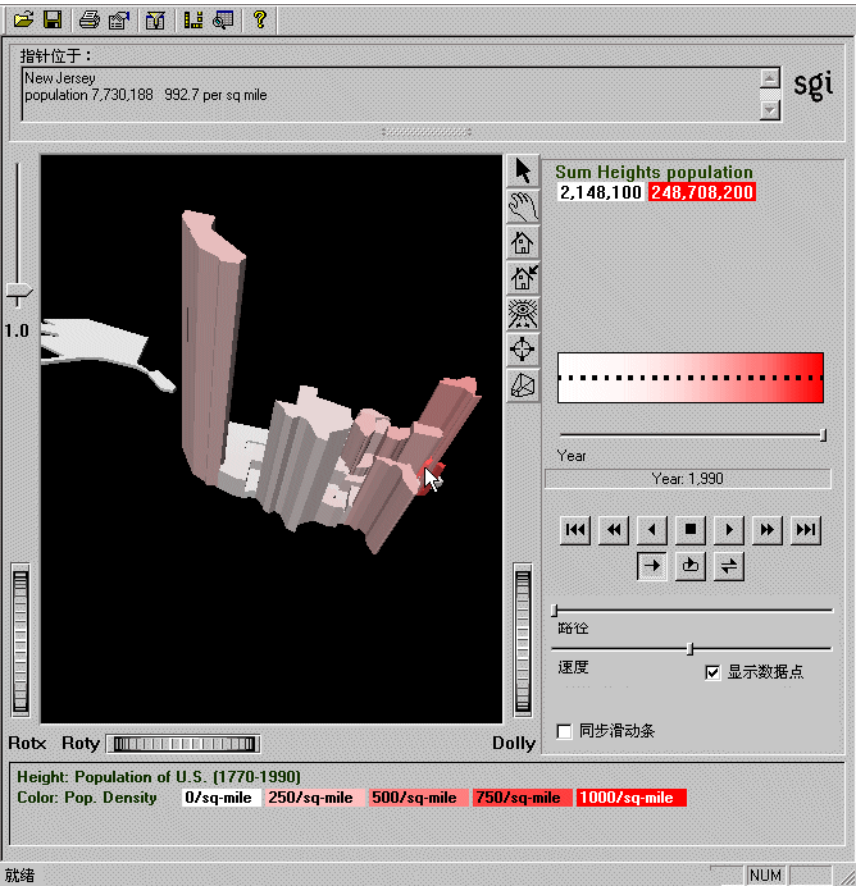


图 6-7 年份滑动条位于 1990 年的 Population.usa.mapviz 示例

该画面显示了美国各州的人口和人口密度。各州的人口由州图形的高度代表。高度在主窗口右边区域的整个动画控制范围中相互关联。

查看模式

两种查看模式为抓取和选取。要在这两种模式之间切换，可以将光标移到主窗口中，然后按 Esc 键。想从一种模式改变为另一种模式，您也可以单击窗口边框上的箭头或手形图标。关于详细信息，请参阅第 64 页。

上寻和下寻

只有当 .data 和 .hierarchy 文件支持概化上寻和细化下寻时，您才可以在选择模式中细化下寻以获得详细信息，或概化上寻以查看更全面的视图。“地图可视化工具”样本文件支持下寻和上寻操作。

表 6-1 追溯操作

目的:	方法:
细化下寻对象	在对象上单击鼠标右键。
细化下寻所有对象	在背景上单击鼠标右键。
概化上寻对象	在对象上按住 Ctrl 键并单击鼠标右键（或鼠标中键）。
概化上寻所有对象	在背景上按住 Ctrl 键并单击鼠标右键（或鼠标中键）。

可以重复细化下寻到数据支持的最细微级别的粒度。对象的数目越大，“地图可视化工具”在追溯操作后用于重建场景的时间就越长，并且动画控制的执行也越慢。

默认情况下，“地图可视化工具”一开始将显示最细微级别的对象细节，因此，这时只有概化上寻（更大的粒度）是处于活动状态的。

可视化空数据

当在特定位置缺少数据或数据为空时，将显示一个问号。在“地图可视化工具”中出现空值，可能由于数据库或数据文件包含空，也可能由于其他特定原因引起（请参阅《*MineSet 3.0 企业版参考指南*》中的“空”一节）。

在地图可视化工具中创建动画

可以使用主窗口右边的动画控制面板创建动画。只有映射了一个或两个滑动条时，才会出现动画窗口。请参阅《*MineSet 3.0 企业版参考指南*》中的“动画”。

处理地图可视化工具结果

使用下拉菜单“查看”和“选项”，可以获得“地图可视化工具”的扩充功能。

更改地图可视化工具画面

“查看”菜单提供的选项汇总在表 6-2 中。关于详细信息，请参阅《*MineSet 3.0 企业版参考指南*》中的“查看”菜单。

表 6-2 地图可视化工具的查看菜单选择

选项	描述
筛选面板	打开“筛选”面板，根据选定准则对主窗口中显示的实体数进行筛选。请参阅《 <i>MineSet 3.0 企业版参考指南</i> 》中的“筛选”一节。
显示窗口控件	显示或隐藏主窗口周围的外部控件
动画面板	显示或隐藏主窗口右边所有的动画控制面板
数据点	使二维汇总窗口中出现或隐藏由黑色圆点组成、用于指示精确数据值的栅格
使用随机颜色	使配置文件中的颜色映射规定激活或失效。
显示 X-Y 坐标	用于形成和改进 .gfx 文件，而不是用于数据分析。关于详细信息，请参阅《 <i>MineSet 3.0 企业版参考指南</i> 》中的“地图可视化工具”。

选项和追溯

“选项”菜单可以对“地图可视化工具”中选定的数据执行表 6-3 中所列的操作。关于详细信息，请参阅《MineSet 3.0 企业版参考指南》中的“地图可视化工具”。

表 6-3 地图可视化工具的选项菜单选项

选项	描述
显示值	显示选定对象的值的“记录查看器”表
显示原始数据	检索并将记录显示为“记录查看器”表
发送到工具管理器	根据当前选择在“工具管理器”历史的开始处插入一个筛选操作。
在追溯中使用滑动条	确定在追溯表达式创建过程中是否使用了滑动条位置。默认情况下，将追溯限制在由滑动条位置定义的记录中。
互补追溯	获取全部没有被显示原始数据和发送到工具管理器选定的数据。

关于追溯的详细信息，请参阅《MineSet 3.0 企业版参考指南》中的“追溯”。

关于“地图可视化工具”附带样例文件的说明，请参阅《MineSet 3.0 企业版参考指南》中的“样例文件”。

理解预测建模

本章介绍了预测建模并讲述了如何使用 MineSet 从数据集中生成预测模型。由 MineSet 生成的可视化处理可以帮助了解模型是如何操作的。可以使用追溯技术检查基本数据。讨论的主题包括：

- 第 113 页的 “预测建模概述”
- 第 114 页的 “生成模型”
- 第 121 页的 “评测预测模型”
- 第 129 页的 “应用预测模型”
- 第 131 页的 “进入后续内容”

以后的章节中将深入介绍可视化处理和模型。关于详细的技术信息，请参阅 *《MineSet 3.0 企业版参考指南》*。

预测建模概述

MineSet 包含许多用于生成预测模型的数据分析挖掘运算法则。如果给定了记录中的其他几个属性值，预测模型就可以预测被称为标签的属性值。如果在属性中有足够的信息，模型就可以对标签进行准确的预测。被预测的标签指的是给定记录中未知的特征，例如：在信用卡历史数据集中，如果知道了 “年龄”、“性别” 和 “职业” 等属性的值，预测模型的任务可能就是预测 “信用风险” 属性的值。

如果数据中含有未知的标签，请使用预测建模，但是如果需要可视化含有未知标签的数据，使用描述建模更为合适。MineSet 的两类预测建模分别为分类和回归。分类预测离散值，回归预测连续范围内的值。下面的章节将对它们分别加以详细介绍。本章主要讨论所有预测模型的一般过程。

生成模型

MineSet 用于生成预测模型的运算法则取决于将要预测的属性类型。对于分类任务，MineSet 的运算法则将生成“证据”模型、“决策树”、“选项树”或“决策表”。对于回归任务，MineSet 将生成“回归树”。

以下五个小节对 MineSet 支持的各种类型预测模型进行了简短的介绍，并且利用图例说明它们是如何生成的。可以使用“工具管理器数据目标”面板中的高级选项按钮制订模型生成运算法则工作的方式。关于这些选项是如何影响模型生成器的详细技术描述，请参阅《*MineSet 3.0 企业版参考指南*》中的“工具管理器中的导入工具模式”。

关于单独的导入工具列表，请参阅《*MineSet 3.0 企业版接口指南*》中的附录 A。

证据模型

本部分讲述如何生成叫做“证据分类器”的预测模型，有关详细信息请参阅第10章，“用证据分类器和可视化工具建模和预测”。

要建立预测模型，先使用“工具管理器”选择数据集。关于详细信息，请参阅第2章中的“运行 MineSet”该示例使用客户波动数据集建立模型，用来预测哪些用户可能改变他们的电信提供商。

1. 在“工具管理器数据目标”窗格中，单击“挖掘工具”选项卡。
2. 单击“分类”选项卡，然后从弹出菜单中进行下列选择：

模式：仅用于分类器

导入工具：证据

离散标签：churned（举例）

即将导入证据分类器以帮助特征化可能会波动的用户。在该示例中，将使用所有的数据来建立模型，而不为准确性测试保留任何数据。

3. 单击继续。

“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。当导入步骤完成后，系统将自动调用“证据可视化工具”，显示模型的可视化处理，如图 7-1 所示。

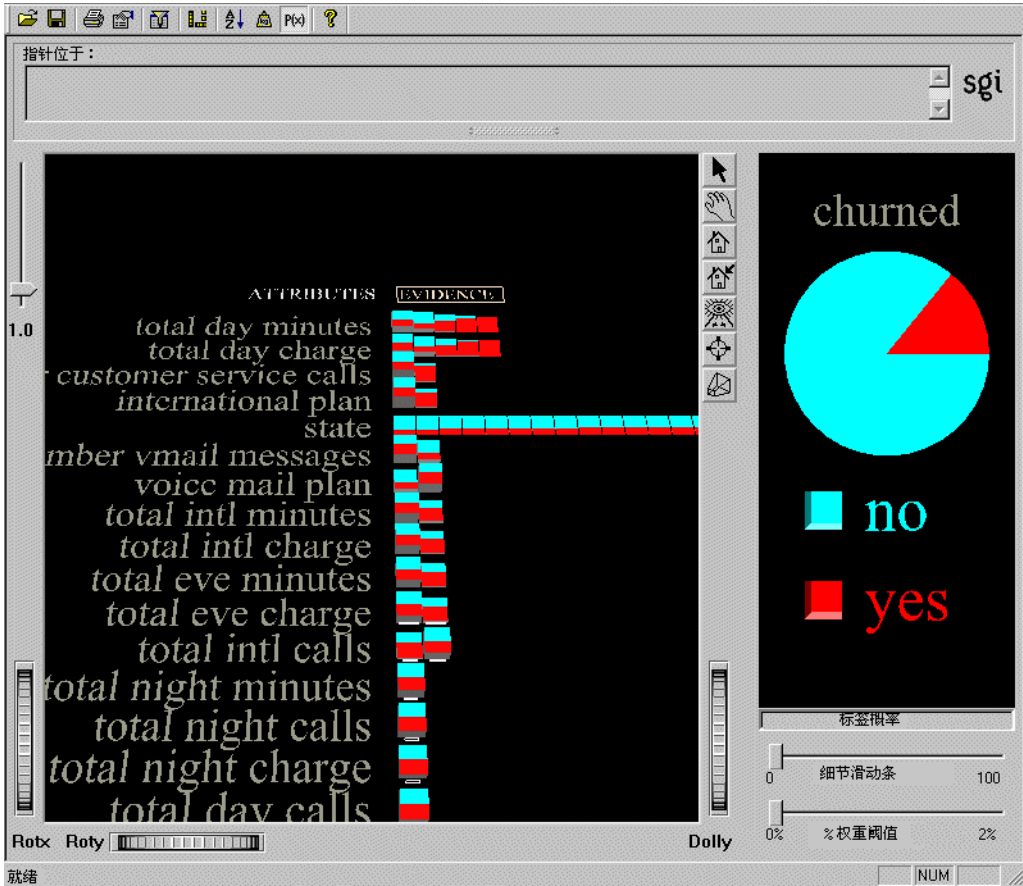


图 7-1 “证据导入工具”为客户波动数据集生成的证据可视化处理

该模型显示每个属性预测客户波动值的准确性。关于如何根据预测行为使用模型，请参阅第 10 章，“用证据分类器和可视化工具建模和预测”。

决策树模型

在上一个示例中，使用了“证据”方法导出并可视化分类器。如果选择了“决策树”方法，可以显示各种属性是如何相互作用的，即：属性值的组合是如何影响被预测标签的。有了“决策树”，数据如何在以后的节点（决策点）上分布将取决于上一个节点作出的决策。在图 7-2 所示的例子中，根决策点的分支为白天的分钟总数属性是否大于或小于 264.45。这会影响后续分支中值的分布；显然，较高的条形图出现在可视化处理的左边。

要生成“决策树模型”，其步骤与生成“证据模型”的步骤一样，除了在“分类”选项卡上选择“导入工具”时，需要选择“决策树”。在第 8 章，“用决策、选项和回归树建模和预测”中将全面讲解该任务。该方法的结果如图 7-2 所示。从根扩展出去子树中类的分布是不同的，这表明属性之间的相互作用是非常显著的。

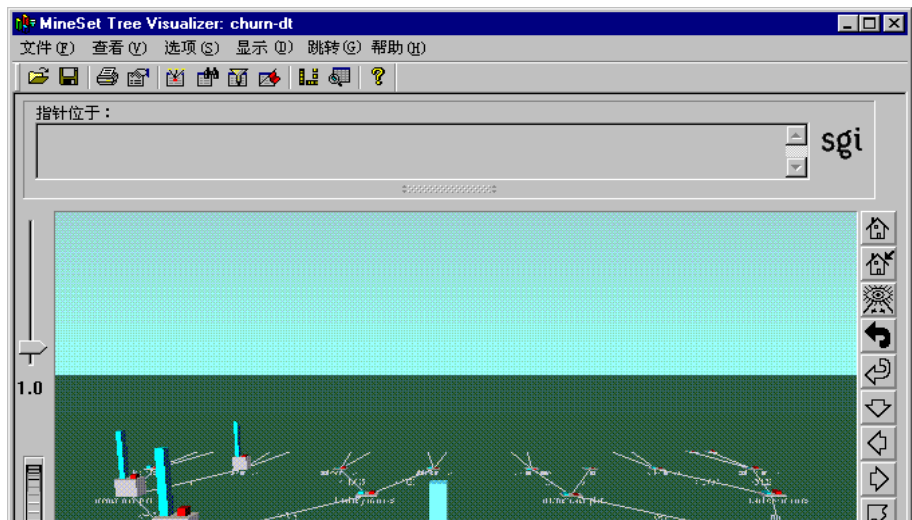


图 7-2 “决策树导入工具”为客户波动数据集生成的决策树

该决策树每个节点上的条形图代表标签值（类）的分布。当光标放在条形图上时，画面上的状态区域中会显示该标签的记录权重和百分比。每个节点的基准高度代表记录的权重（如果没有使用权重，就用记录计数）。

在这个示例中，代表决策树根的小块被标签“白天的分钟总数”所标记，表明这是预测客户波动中一个最重要的因素——这些用户的交谈时间，划分的阈值为 264.45 分钟。在 第 8 章，“用决策、选项和回归树建模和预测”中将对“决策树”进行详细讲解。

选项树模型

“决策树导入工具”为每个子树选择一个“最好”的属性；然而可能有多个用以拆分的好属性。这时，“选项树”可以创建选项节点。“选项树”的样子与“决策树”类似，但它是由“选项树导入工具”生成的。

顶节点是一个“选项节点”，显示有多个好属性可以在根部被选择。可以设置每个节点生成的选项数，表示该点最好的属性。然后将为每个选项生成一个分支。图 7-3 所示的例子使用了汽车数据集。在第 8 章，“用决策、选项和回归树建模和预测”中将对“选项树”进行详细讲解。

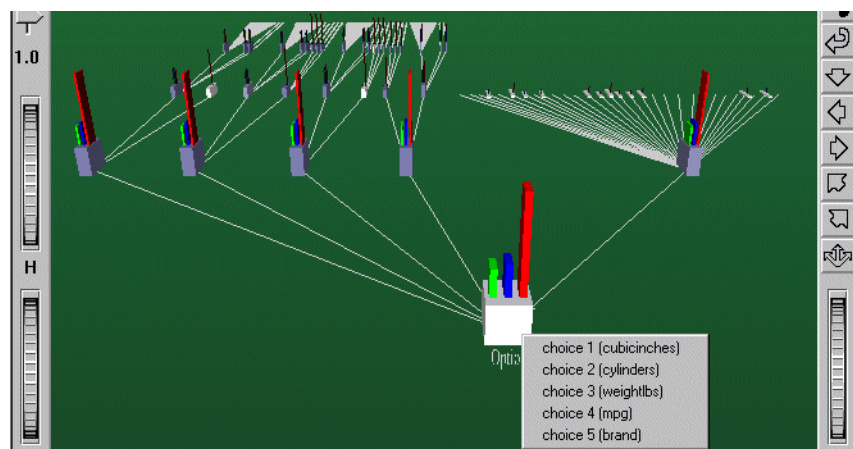


图 7-3 “选项树导入工具”为汽车数据集生成的选项树

在示例数据集（图 7-3），要预测汽车的产地是在欧洲、日本还是美国。“决策树导入工具”为根选取了体积属性。相对而言，“选项树导入工具”选择了多个属性作为根的其他不错的选项，其中包括：体积、汽缸、重量、英里 / 加仑和商标。在根上单击 **Ctrl** 键和鼠标右键可以看到这些选项。

选项节点还可以出现在根以外的其他地方。但在默认情况下，它们只出现在根或紧邻根的下一级。

相对于“决策树”，“选项树”通常需要 10 到 15 倍的时间用来建立，并且规模要大得多，但是它们具有两个显著的优点：

- 灵活性—可以使用选项节点看到多个合适的选项。可以使用选项节点从多个而不只是一个属性中进行选项设置。当掠过树时，可以跟踪一个易于理解或适合用户对该问题背景知识的选项。
- 准确性—通常，“选项树”的错误率低于“决策树”。“选项树”的分类是使每个选项给每个标签值“投票”，然后将选票平均。这好比拥有一个专家小组，每个人都试图根据不同的标准来预测标签。选项节点将所有专家的选票平均，产生一个更加稳定、风险更小的分类器。

决策表模型

“决策表”导入的模型表示分层结构中属性对之间的关联。导入工具进行决策的方式与“决策树”的一样，但是属性的评测遍及树的所有层次，而不是某个特定的子树。然后，结果的表现形式是分层表而不是树。在可视化处理中的小块图上单击将显示其详细组成信息，如图 7-4 所示。

回归树模型

“回归树导入工具”生成一个回归器，它是一个类似于分类器的预测模型。其区别在于回归器预测的是具有连续值标签，例如个人的年薪。而分类器只预测离散值，例如：“是”或“否”，或一个指定的范围（如 20-100）。

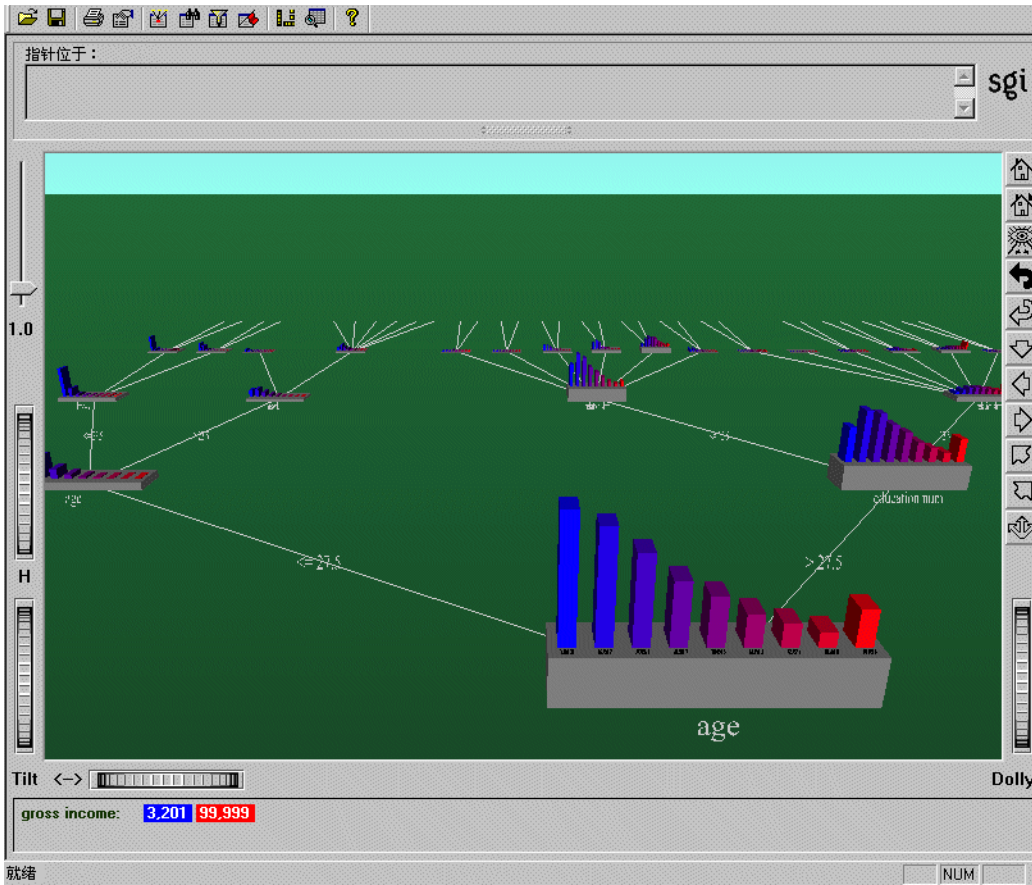


图 7-5 回归器根据成人数据集生成的回归树

下面的示例讲述如何生成回归器，在第 8 章，“用决策、选项和回归树建模和预测”中将介绍这个任务。

要建立预测回归器，先使用“工具管理器”选择一个数据集。关于详细信息，请参阅第 2 章，“利用 MineSet 访问数据”。该示例使用成人数据集。

1. 在“工具管理器数据目标”窗格中，单击“挖掘工具”选项卡。
2. 单击“回归”选项卡，然后从弹出菜单中进行下列选择：

模式：仅用于回归器

导入工具：回归树

连续标签：总收入（例如）

即将导入一个回归器，以帮助特征化数据集中记录代表的成人的总收入。在该示例中，将使用所有的数据来建立模型。

3. 单击继续。

“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，也可以在对话框中观看进程指示器。当导入步骤完成后，系统将自动调用“回归树可视化工具”，显示一个模型的可视化处理，如图 7-5 所示。某些数据集生成的时间较长。

可以为拥有连续属性（即：属性值在一个连续范围内变化）的数据集生成一个回归器，如果没有这样的列，可以从“工具管理器数据转换”窗格中添加一个新的连续列，请参阅第 3 章中的“删除和添加列”。

第 126 页的“创建学习曲线”中将讨论学习曲线分类器模式。

评测预测模型

预测模型的目的是为了进行预测，因此模型的有效性显然取决于其预测的准确性。模型的准确性由错误率衡量。本部分讨论 MineSet 建模模式的范围是：“仅用于分类器”使用所有的数据建模，不进行错误估计；估计错误率的方法通常有两种—预留和交叉验证。

在估计错误的预留方法中，数据集的一部分（通常为三分之二）被用来生成该模型。导入工具使用该训练数据集中的标签构造模型，其余的数据用于测试模型和估计错误率。

交叉验证是估计分类器错误的另一种方法，它将数据集拆分为一定数量的数据夹或子集（一般是 10 个），然后建立相同数量的分类器。这个过程可以被重复多次以提高估计的可靠性。

预留与交叉验证方法都是期望将来采集的记录与训练集的数据分布是相同的。《*MineSet 3.0 企业版参考指南*》的“错误估计”和“交叉验证”对此进行更加详细的讨论。

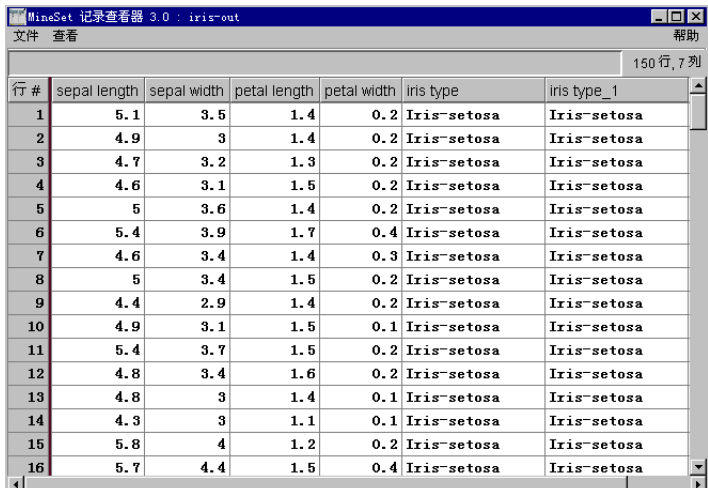
使用所有数据进行分类

建立模型的仅用于分类器模式将使用所有可用的数据来建立分类器。当不要考虑估计预测模型的错误率时（例如：只是希望查看可视化处理），这是很有用的。

假设希望得到用仅用于分类器模式导入的蝴蝶花数据集的“决策树”。这个例子讲述了建立并应用模型的整个程序。按照以下步骤，可以发现模型的错误分类：

1. 使用“文件” > “打开新的数据文件”下拉菜单，从“工具管理器”窗口中选择一个数据集。选择 *iris.schema* 数据集。
2. 在“工具管理器数据目标”窗格中，单击“挖掘工具”选项卡，然后单击“分类”选项卡并选择
 - 模式：仅用于分类器
 - 导入工具：决策树
 - 离散标签：蝴蝶花类型将使用所有的数据建立“决策树”模型，确定用以区分不同类型蝴蝶花的属性。
3. 单击继续，运行“决策树”。检查，然后关闭可视化处理。创建好的模型使用了所有的数据。以下的步骤将讲述在这种情况下，如何评测正确和不正确的预测。

- 4. 在“工具管理器数据转换”面板中单击应用模型。选择 *iris-dt.class*（蝴蝶花决策表分类器），然后单击确定接受默认的新列名称 *iris type_1*，其中含有预测的标签。
- 5. 单击添加列按钮，创建另一个标识错误的新列（称之为 *iris_fault*）。然后，可以添加一个列，定义为带有表达式（'*iris type*'!= '*iris type_1*'）的 **int** 类型。
要建立表达式，可以从左边的列表中选择，然后按向右箭头，或直接在文本字段中键入。
- 6. 单击检查表达式，然后单击“确定”。这将添加一个基于 *iris type_1* 的新列 *iris_fault*。当分类器分类错误时，该列中有一个 1；当分类正确时，该列中有一个 0。
- 7. 验证结果。在“工具管理器数据目标”窗格中，单击“可视化工具”，从“工具”弹出菜单中选择“记录查看器”，然后单击调用工具。图 7-6 所示为结果的一个示例。



行 #	sepal length	sepal width	petal length	petal width	iris type	iris type_1
1	5.1	3.5	1.4	0.2	Iris-setosa	Iris-setosa
2	4.9	3	1.4	0.2	Iris-setosa	Iris-setosa
3	4.7	3.2	1.3	0.2	Iris-setosa	Iris-setosa
4	4.6	3.1	1.5	0.2	Iris-setosa	Iris-setosa
5	5	3.6	1.4	0.2	Iris-setosa	Iris-setosa
6	5.4	3.9	1.7	0.4	Iris-setosa	Iris-setosa
7	4.6	3.4	1.4	0.3	Iris-setosa	Iris-setosa
8	5	3.4	1.5	0.2	Iris-setosa	Iris-setosa
9	4.4	2.9	1.4	0.2	Iris-setosa	Iris-setosa
10	4.9	3.1	1.5	0.1	Iris-setosa	Iris-setosa
11	5.4	3.7	1.5	0.2	Iris-setosa	Iris-setosa
12	4.8	3.4	1.6	0.2	Iris-setosa	Iris-setosa
13	4.8	3	1.4	0.1	Iris-setosa	Iris-setosa
14	4.3	3	1.1	0.1	Iris-setosa	Iris-setosa
15	5.8	4	1.2	0.2	Iris-setosa	Iris-setosa
16	5.7	4.4	1.5	0.4	Iris-setosa	Iris-setosa

图 7-6 蝴蝶花错误分类示例

然后，可以将结果应用到一个散点图，以使新列映射到颜色，颜色集中绿色是 0（没有错误），红色是 1（错误）。关于“散点可视化工具”的示例，请参阅第 4 章，“用散点和平伸可视化工具检查数据”。

预留误差估计

不是使用所有的数据建立模型，而是仅使用部分数据作为训练集来导入分类器。“分类器和误差”模式（MineSet 的默认模式）将自动将数据集划分为训练和测试两个子集。可以改变用作训练集的数据比例以满足要求。

要了解预留错误估计是如何应用到选定的数据集的，可以按照下列步骤执行：

- 1. 从 *iris.schema* 数据集开始（请参阅第 122 页）。
- 2. 在“工具管理器数据目标”面板中，单击“挖掘工具”选项卡，然后单击“分类”选项卡并选择：
 - 模式：分类器和错误估计
 - 导入工具：决策树
 - 离散标签：蝴蝶花类型将蝴蝶花数据集划分为两个子集。训练集中包含三分之二（默认的支持记录比例）的记录用于建立“决策树”模型。剩余的三分之一记录用于估计模型的错误率。
- 3. 单击高级选项按钮，使用文本面板改变训练集的支持比例，如图 7-7 所示。还可以改变随机子以获得不同训练集和测试集。请参阅词汇表中的词条随机子。单击确定，关闭面板。

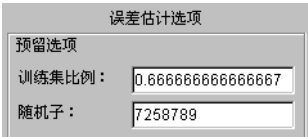


图 7-7 预留的错误估计选项

- 4. 单击继续，运行“决策树”导入工具。

主窗口底部的状态区域中显示用测试集（剩余的三分之一数据）估计的导入工具错误率。可以向下滚动，查看所有信息。

交叉验证误差估计

为建立最终的分类器或对小数据集可以使用交叉验证的方法。使用的过程应该可以说明其原因。交叉验证是获得对误差更加精确的估计的方法。在 n -fold 交叉验证（其中 n 代表赋予名称一个数字， $fold$ 是希望将数据划分成的子集）中，数据集被划分为 n 个独立的子集。每个子集将依次被留下，其余的 $n-1$ 个子集（比最初指定的数目少 1）组成为一个训练集。生成的模型用留下的子集进行评测。然后，对这 n 个独立的估计进行平均，并将得到的数据组合起来建立最终的模型。N-fold 交叉验证使用的时间大约是“分类器和误差”或“仅用于分类器”方法的 $n+1$ 倍。

将交叉验证应用到一个选定数据集并设置倍数：

1. 在“工具管理器数据目标”窗格中，选择“挖掘工具”选项卡，然后单击“分类”
2. 如果使用了所有数据（如在“仅用于分类器”模式中），在“模式”弹出菜单中选择误差估计以访问误差。由于使用了交叉验证，运行时间可能较长。当数据很少时，可以使用误差估计。导入的分类器与仅用于分类器模式导入的分类器完全一样。
3. 单击高级选项按钮，改变交叉验证的运作方式。在“误差估计选项”（请参阅图 7-8）中，可以设置 Fold 和 Time 的值以重复误差估计过程。单击确定接受所有更改。

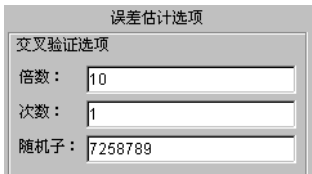


图 7-8 交叉验证的误差估计选项

4. 单击继续，运行算法。

关于“误差估计”更加详细的讨论，请参阅 [《MineSet 3.0 企业版参考指南》](#)。

创建学习曲线

学习曲线显示导入工具生成的分类器的误差，与用于创建分类器的记录数成反比。通常，用于生成分类器的记录越多，误差就越小。

通过指出希望调查的训练集的大小范围，和对于每个训练集大小希望建立导入工具的次数，可以创建学习曲线。要指出希望调查的训练集的大小范围，可以输入记录的起始数（最小值）和终止数（最大值）以及要检查的点数。每个分类器都是使用记录的随机样本生成的，并使用剩余的记录（没有用于训练的记录）估计其误差。

1. 从第116页使用的 *churn.schema* 数据集开始。
2. 在“工具管理器数据目标”面板中，单击“挖掘工具”，然后单击“分类”选项卡并选择：

模式：学习曲线

导入工具：决策树

离散标签：churned（举例）

系统将提示不计算电话号码列，因为其中包含了太多不同的值。可以删除这个列，或是单击“确定”。

3. 单击继续。

该过程可能需要较长的时间，因为导入工具正在为曲线上的每个点都生成一个分类器。状态窗口中会显示训练范围、已经完成的训练次数、和每次的平均误差。

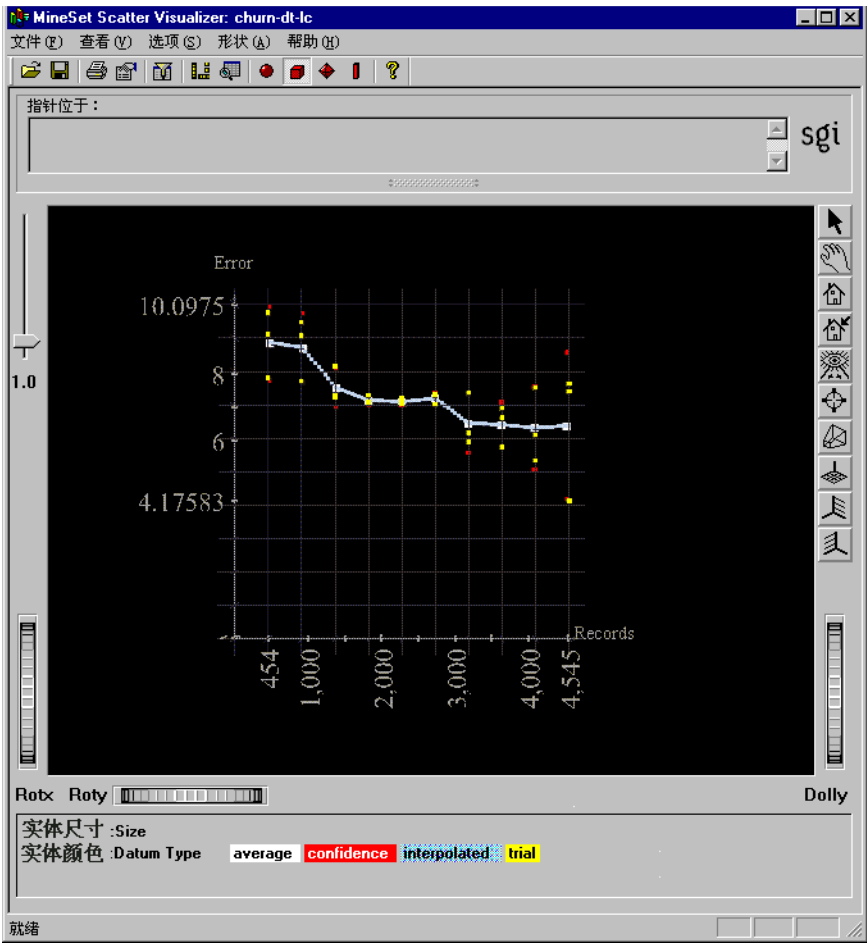


图 7-9 带有客户波动标签集的客户波动数据集的学习曲线

图 7-9 指出了客户波动数据集中给定记录数的错误率。如果训练了 454 个记录，错误率为 9.41%；如果训练了几乎所有的记录，那么错误率为 5.27%。通过找出错误率不再降低的点，可以确定生成给定错误率的分类器所需的训练记录数。在这个例子中，对 3,000 个记录进行训练就足够了。可以根据准确性适当的学习曲线上的点选取一个样本大小。

图中显示了四种类型的点：

- 黄色的点是运行多次（运算法则运行的次数）后得到的实际误差估计。
- 白色的点是平均值。
- 蓝色的点（有时象一条蓝色的线）内插在白色的点之间。
- 红色的点显示的是关于平均值的一个 95% 的置信区间，该平均值基于每次运行的实际误差估计。

默认情况下，运算法则使用所有的记录建立学习曲线，但是可以指定其他模式：

1. 使用前面的“工具管理器数据目标”窗格中的“学习曲线”、“决策树”和客户波动选项，单击高级选项。
2. 可以在高级选项对话框中指定这些“学习曲线选项”（图 7-10）
 - 学习曲线中的点数，必须大于 1
 - 每一点运行次数
 - 起始和终止点使用的记录数。

系统将自动计算中间每个点使用的记录数。

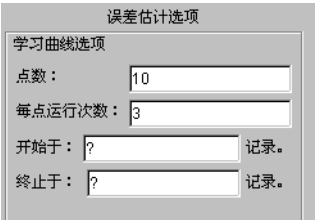


图 7-10 学习曲线选项

如果范围和点数都留成空白，系统将自动根据学习曲线中的点数和训练集中的记录总数计算。该默认值覆盖了训练集的整个范围。例如：假设有一个文件含有 80,000 个记录。如果在学习曲线中指定了 3 个点，运算法则就在 20,000、40,000 和 60,000 个记录处生成点。通常，在较小的范围上进行“放大”这是很有用的。例如：可以仅用 1,000 到 10,000 个记录生成学习曲线。

增加每一点运行的次数会相应地增加运行时间，但是也会改善对误差的估计。运行次数的默认值是 3。

可以为所有的导入工具生成学习曲线。可以按照默认值运行，也可以确定曲线中的点数、运行的次数或记录的范围。表 7-1 所示是特定选择对学习曲线的影响。

表 7-1 对学习曲线结果的操作

要获得该结果：	方法：
减小置信区间 （更高的准确性）	增加运行次数。 增加测试集的大小。
加速处理	使用小样本 （请参阅第 46 页的 “数据采样”）

关于导入工具操作的详细信息，请参阅 《MineSet 3.0 企业版参考指南》 中的 “导入工具”。

应用预测模型

建立好预测模型后，可以将其应用到其他记录，预测它们的标签。例如：如果为预测蝴蝶花 类型建立了分类器，可以将这个分类器应用到仅包含描述属性的记录，就会添加一个带有预测蝴蝶花类型的新列。

选择模型

MineSet 生成的模型被保存到指定的服务器目录中。例如： *churn-dt.class* 是从客户波动数据集中生成的决策树模型。可以将这个分类器应用指定到另一个数据表。使用 “决策树” 为该示例建立了模型后，请继续进行以下步骤：

- 1. 在 “工具管理器数据转换” 窗格中单击应用模型按钮。
- 2. 在 “测试和应用模型” 对话框（图 7-11）的 “可用模型” 面板中选择一个模型。接着，右边的面板会列出该分类器所需的列名称和类型。

如果当前内存加载的表符合这些要求，那么每个标签底部的按钮（应用模型，测试模型，或拟合数据到模型）将被激活。如果不符合要求，右边列表中遗漏的列就会被自动选择，底部的按钮将无效。

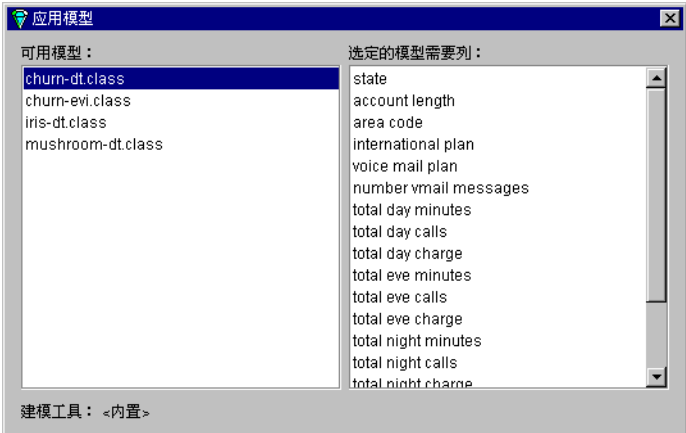


图 7-11 “测试和应用模型” 窗口：选择分类器

应用模型

选定了模型后，希望应用：

- 在“测试和应用模型”对话框中，单击“应用模型”选项卡（图 7-12）并从下面选择一个：

预测的标签值：添加要预测的列，例如：是否某个用户要波动。可以通过新列看到模型使用新数据时的预测效果。

标签估计概率：估计每个记录具有给定标签值的概率。从弹出菜单中选择可用的值。这将添加一个新列，其中的数值表示估计的正确概率。
- 在文本字段中输入新列的名称，例如：p_churned，代表用户波动的概率。

- 3. 单击 “确定”。
- 4. 要将模型图形化地显示为一个列表，从 “工具管理器数据目标” 窗格中单击 “可视化工具”，并在 “工具” 弹出菜单中选择 “记录查看器”，然后单击调用工具。

将显示标注的列 churn 和 p_churned，可以比较预测和结果。

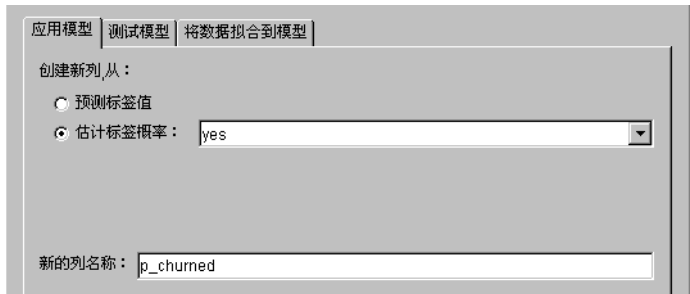


图 7-12 应用模型面板

模型应用可能会在销售活动中进行，其中模型是从某城市进行的商业活动中生成的，并根据那个城市中的反应生成记录（标签）列。模型可能被应用，然后商业活动邮件可以只被发送给另一个城市中的那些被分类器标注为可能会作出回应的人，从而节省了邮寄成本。

进入后续内容

以后的几章将介绍查看数据的不同方式，并指出了每种可视化工具特有的优势。因此，您将发现自己从用 “记录查看器” 查看原始未转换的数据开始，进而认识了 “统计” 和 “直方图” 可视化工具，然后又使用 “列重要性” 工具搜索重要的列，以准备利用 “散点”、“平伸” 或 “地图” 可视化工具获得已知数据的一个有用的可视化处理。

如果没有想过从何开始，也许会希望从聚类开始，并让这种方法引导知识的发现。

如果准备测试并应用一个模型，以及确定模型的准确性，可以在第 11 章，[“改进预测建模”](#) 中找到几种方法。

用决策、选项和回归树建模和预测

决策树、选项树和回归树看起来很相似，尽管它们的运算法则非常不同。本章介绍这三种基于树的模型用法的类似和差别。要了解导入工具所使用的运算法则和可能选项的详细信息，请参阅 *《MineSet 3.0 企业版参考指南》*。

本章包括以下几个部分：

- 第 134 页的 “决策、选项和回归树概述”
- 第 139 页的 “运行决策、选项和回归树”
- 第 140 页的 “用决策树可视化工具查看结果”
- 第 143 页的 “用选项树可视化工具查看结果”
- 第 144 页的 “用回归树可视化工具查看结果”
- 第 146 页的 “用决策、选项和回归树预测”

提供了一些配置和数据文件样例，用于演示 “树可视化工具” 的特性和功能。这些文件在 MineSet 3.0 下的 \examples 目录中，位于 MineSet 的安装目录。

决策、选项和回归树概述

本部分讲解三种类型树的用法，以及选择使用时的一些原则。每种树工具都由建立模型的导入工具和显示结果树的可视化工具组成。导入工具可以是分类器（决策树和选项树），也可以是回归器（回归树）。

决策树

决策树是预测模型。它使用相关的、或已知的属性值进行预测，以帮助确定标签或未知属性的值。对按标称值（通常是字符串，如“是”或“否”），或只在一个小范围内取值的属性的值进行预测的任务称为分类。决策树通过预测每个记录的标签值来对数据分类。用于分类的基本结构就是决策树，如[图 8-1](#)所示。一旦“决策树导入工具”建立了数据分类器，“树可视化工具”就可以将它的结构显示出来。

树中的第一个节点是根节点，代表所有的数据。从此处，树发出两个或更多的分支，每一个都代表具有不同指定属性（列）值的数据。例如：[图 8-1](#)是汽车数据集的决策树可视化处理，其中的记录是按体积分类的。第一个分支处是 169.5 立方英寸，比它小的在右边的枝上，比它大的在左边的枝上。树可以在同一个属性上拆分成多个节点。节点也可以拆分成多个分支。例如：对于蘑菇数据集，记录被分为可食用的和非可食用的，而第一个分支又拆分成九个分支，对应于九种气味类型。

对象将在分支末端（叶节点）到达节点，此处所有的记录，或是说几乎所有的记录都具有相同的类（标签）。在[图 8-1](#)中，类是“决策树”窗口底部列出的体积值的不同范围。

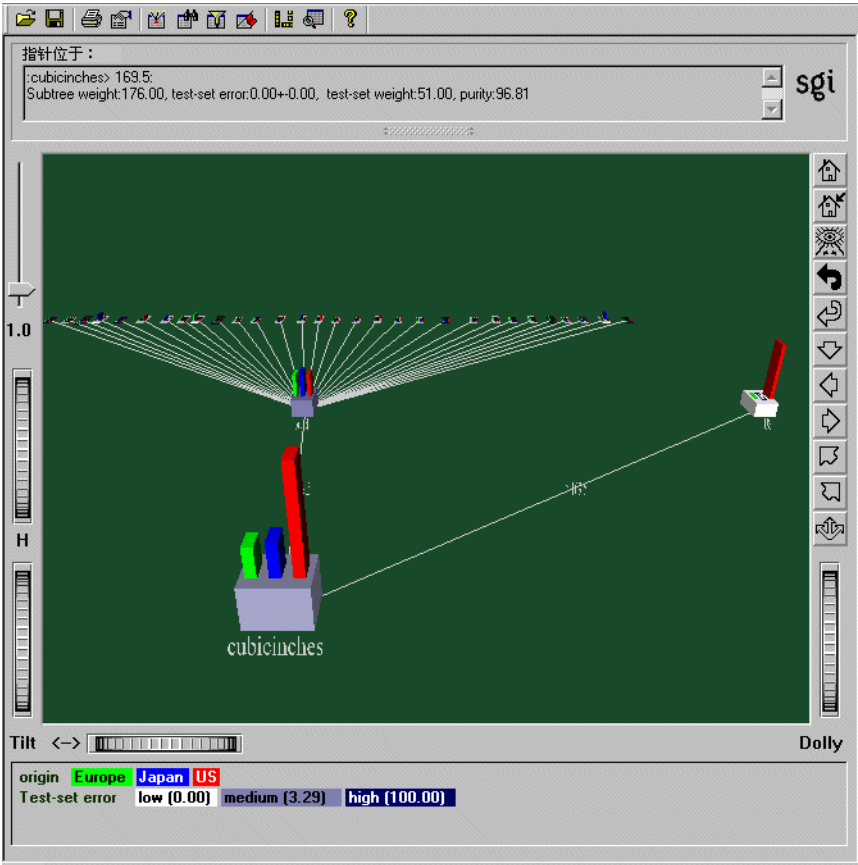


图 8-1 汽车数据集的决策树

选项树

与“决策树”分类器类似，“选项树”也为每个记录分配一个类。用于分类的基本结构就是决策树，如前面部分所述。一旦“选项树导入工具”建立了分类器，“树可视化工具”就可以将它的结构显示出来。一个选项树实际上由多个决策树组成。它不是选取一个属性来为根节点拆分，而是选取多个，并为每个属性都生成一个决策树。

图 8-2 所示为汽车数据集的选项树，其中的目标是预测汽车的产地（美国、日本或欧洲）。“选项树”允许选项节点，从而扩展了“决策树分类器”。“选项节点”提供了多个选项，可以在树中的决策节点进行选择。例如：在图 8-2 中，根就是具有五个选项的选项节点：体积、汽缸数、重量（磅）、英里 / 加仑和商标。

选项节点有两个作用：

- 通过显示可以选择的多个选项，提高影响类标签确定因素的内容广泛性。选项节点在节点处不是使用单独一个属性，而是提供多个选项。当查看树时，可以根据以前的经验，选择一个更易于理解或是更好预测的选项，或是根据误差估计选择选项。

在所示的汽车数据集中，可以选择汽缸子树，因为它的值很少，或是选择重量子树，因为它的估计误差较小（1.53）。误差估计仅仅是估计；通常，如果两个选项之间的误差差别小于它们平均标准偏差的两倍，那么从统计上来看，误差就是相同的。

- 它们通过平均每个选项子树的选票，降低了创建分类器不准确的风险。每个选项生成的子树都可以被看成是一个“专家”。选项节点平均这些专家的选票。这样的平均可以生成错误率较低、更好的分类器。

如图 8-2 所示，在汽车数据集中，根节点的估计错误率为 0.76%，比所有的子节点都低。有时，明显的选项不一定就是最好用的。例如：尽管商标似乎是这个任务的一个明显属性，但是训练集中可能不包含所有的商标（实际上，它也没有包含所有商标）。对于未知的商标，“决策树”猜测为大多数类（美国），犯了两个错误。但是，当存在其他选项时，它们被平均，从而减少了错误。

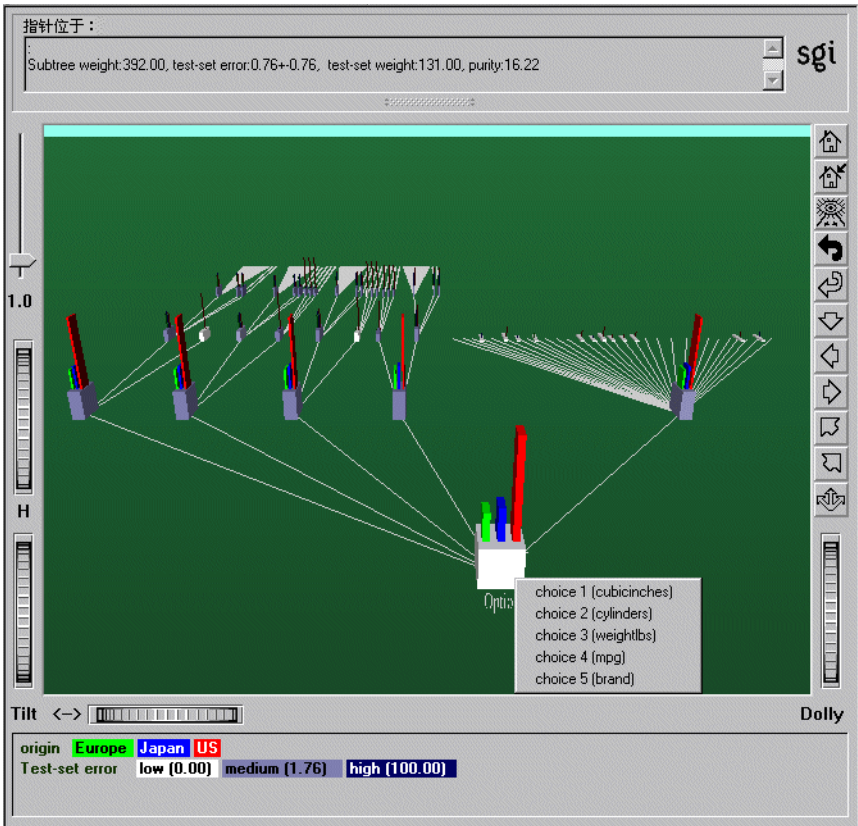


图 8-2 汽车数据集的选项树

但是，选项树有两个缺点：

- 默认设置下，建立选项树需要的时间大约比建立决策树所需时间长 10 到 15 倍。
- 创建的“树可视化工具”文件非常大，包含的节点是正常决策树的 10 到 15 倍。

在数据集上运行“选项树”导入工具，确定在综合性和错误率方面的优点是否能够平衡较长导入时间的缺点，同时可以获得额外的认识，例如：在建立决策树时要删除或使用哪个属性。

回归树

回归树是在给定一组描述性属性的情况下，预测连续标签值的任务。回归和分类是类似的，除了在分类中，被预测标签的取值只能是一小部分离散值以外，例如：蝴蝶花按类型可以被分为蓝色、杂色、纯白色蝴蝶花。在回归中，被预测模型可以在连续范围内任意取值，例如：一个人的年薪可以是任何大于零的数。

当生成了一个回归器时，“选项树可视化工具”会显示相应的回归树。该可视化处理可以帮助理解回归器及它是如何进行预测的。另外，它可提供对数据本身的宝贵认识。回归器一旦生成，就可以用于预测未标注记录的标签值。

用于回归的基本结构就是回归树，如图 8-3 所示。

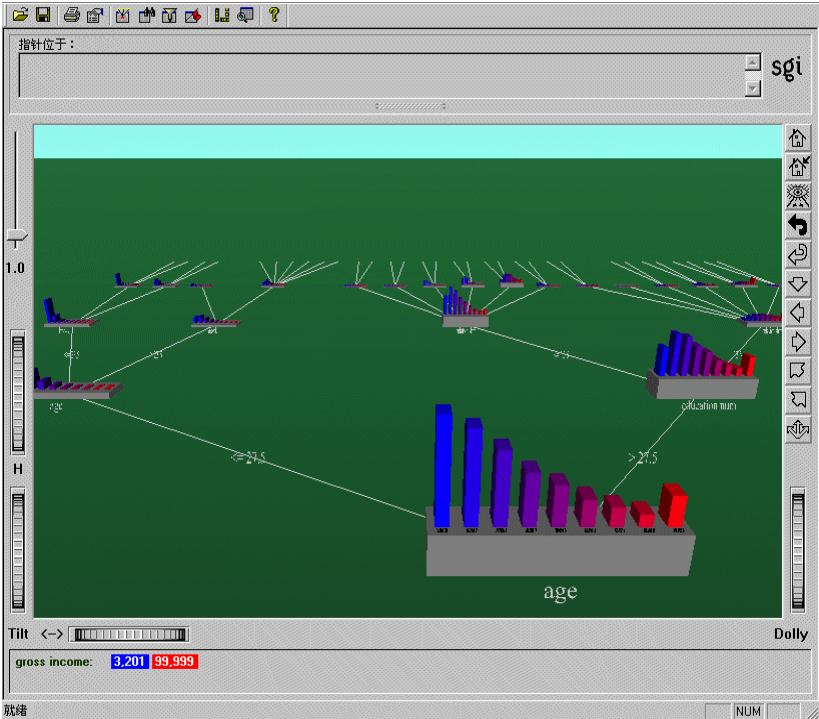


图 8-3 成人数据集的回归树

运行决策、选项和回归树

运行树工具最简单的方法是从“工具管理器”中：

- 1. 在“工具管理器文件”菜单中，连接并登录到服务器；再从同一个菜单中选择“打开新的数据文件”，并选择或键入需要的文件名。
- 2. 在“数据目标”窗格中，单击“挖掘工具”（图 8-4），然后在下面一行选项卡中单击“分类”或“回归”：
 - 对于“决策树”和“选项树”，单击“分类”选项卡。
 - 对于“回归树”，单击“回归”选项卡。
- 3. 从弹出的“模式”菜单中选择一个模式。关于四种可用模式的详细信息，请参阅《MineSet 3.0 企业版参考指南》中的“误差估计”条目。
- 4. 从弹出的“导入工具”菜单中，选择导入工具。
- 5. 从“离散标签”（对于“回归树”是“连续标签”）弹出菜单中，选择希望用作标签的属性（列）。例如：对于蘑菇数据集，可食性是一个合理选择。对于客户波动数据集，已波动客户的是一个合理选择。

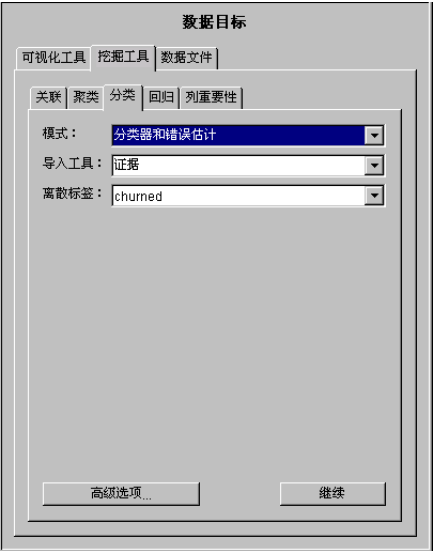


图 8-4 工具管理器数据目标窗格，分类选项卡

6. 单击继续，运行导入工具。“工具管理器”窗口底部的“状态”窗格将显示进程与统计结果。一旦 MineSet 完成计算，就会出现可视化工具窗口。

用决策树可视化工具查看结果

决策树和选项树可视化处理是类似的。它们都是由边（线）连接的两类节点组成：决策节点和叶节点。

决策节点上的标签将指定在该节点被测试的属性，用于测试属性的值（或值的范围）将显示在边上。每一个可能的属性值都刚好符合一条边。例如：图 8-1 中决策树的根测试属性重量磅；从节点生出的两条边指定了该属性值的范围（小于等于 3018，和大于 3018），以便每个可能值都能与右边或左边的分支相匹配。

决策树中的叶节点指定类。在图 8-1 中，如果跟随右边的分支从根到标注的叶 (... 16.5]，就会看到“决策树”分类记录的分界是：当达到 16.5 英里 / 加仑或更少时，重量大于 3018 磅、马力大于 141、以及立方英寸大于 311.5。

基准块的颜色指示子树的误差估计：靛青显示高误差，灰色显示中等误差，白色显示低误差。如果没有测试集记录到达节点（表示不存在误差估计），基准的颜色为黑色。

每个节点顶上的垂直条形图显示节点处类的分布。节点的基准块具有高度和颜色。高度对应于到达该节点训练集记录的权重（如果没有设置权重，就是记录的数目）。通常，权重越高，每个节点处的类分布就越可靠（关于加权记录的详细信息，请参阅第 3 章中的“加权记录”）。

将鼠标箭头放在节点上面，会显示以下信息：

- 子树权重—被指向节点下方的子树中训练集记录的权重。该值被映射到基的高度。
- 测试集错误 / 损失—子树错误（如果给定了损失矩阵，就是损失）的估计。+/- 后面的数字是估计的标准偏差。标准偏差越高，误差估计就越不准确。对于带有很少记录的叶，或者当测试集错误接近于 0% 或 100% 时，错误 / 损失估计和标准偏差较不可靠。

- **测试集权重**—到达节点的测试集记录权重（如果没有设置权重，就是记录的数目）。
- **纯度**——一个从 0 到 100 的数，指示节点处标签值分布的倾斜度（请参阅词汇表中的词条**倾斜度**）。如果节点的记录都来自于单独一个类，那么纯度为 100。如果标签值的权重相同，那么纯度为 0。在修正之后计算纯度（请参阅词汇表中的词条**修正**）。

只有“分类和错误估计”模式，生成测试集误差 / 损失和权重。可以使用“测试分类器”选项生成基于已有分类器和测试集的可视化处理。

使用决策树主窗口对记录分类

要分类具有未知标签的记录，可以从树根开始，然后跟随该记录属性值指示的分支，根据记录属性值选择适当的边，就可以到达一个叶节点。与叶节点相关联的标签或类就是预测的记录分类。

有一些决策完成得很快，并且路径较短。其他决策的路径可能较长。通常，每个叶节点对应于一个规则，它是决策节点处的所有测试与从根通向它的沿线上所有值（或值的范围）的合取。

为蝴蝶花数据集（图 8-5）产生的决策树非常简单，很适于演示。该文件可以从“工具管理器可视工具”中打开；选择“三维可视化工具”，然后选择 `\examples\iris-dt.treeviz`。在这个树的根部，误差率是 6%，标准偏差是 3.39%。标准偏差高是因为文件小，测试集中只有 50 个记录。纯度为 0.0，表明分布是同一的。

根的左边子节点有 0 个测试集错误，且纯度为 100，因为所有花瓣长度小于等于 2.6 英寸的记录都是蓝色蝴蝶花类；因此对于花瓣长度小于等于 2.6 英寸的所有记录，蓝色蝴蝶花的预测很可能非常准确。根的右边子节点估计误差为 8.57%。在该子节点中，符合的记录都是花瓣长度大于 2.6 英寸的，没有任何记录属于蓝色蝴蝶花类；因此这个类很可能是杂色蝴蝶花或纯白色蝴蝶花。因为在该节点处只有两种可能性，因此纯度比根部高（36.91）。

决策树叶节点将数据划分为共享同一个分类规则（到达每个叶的路径）的群集。通过查看叶节点，就可能看到共享相同属性集的群集。

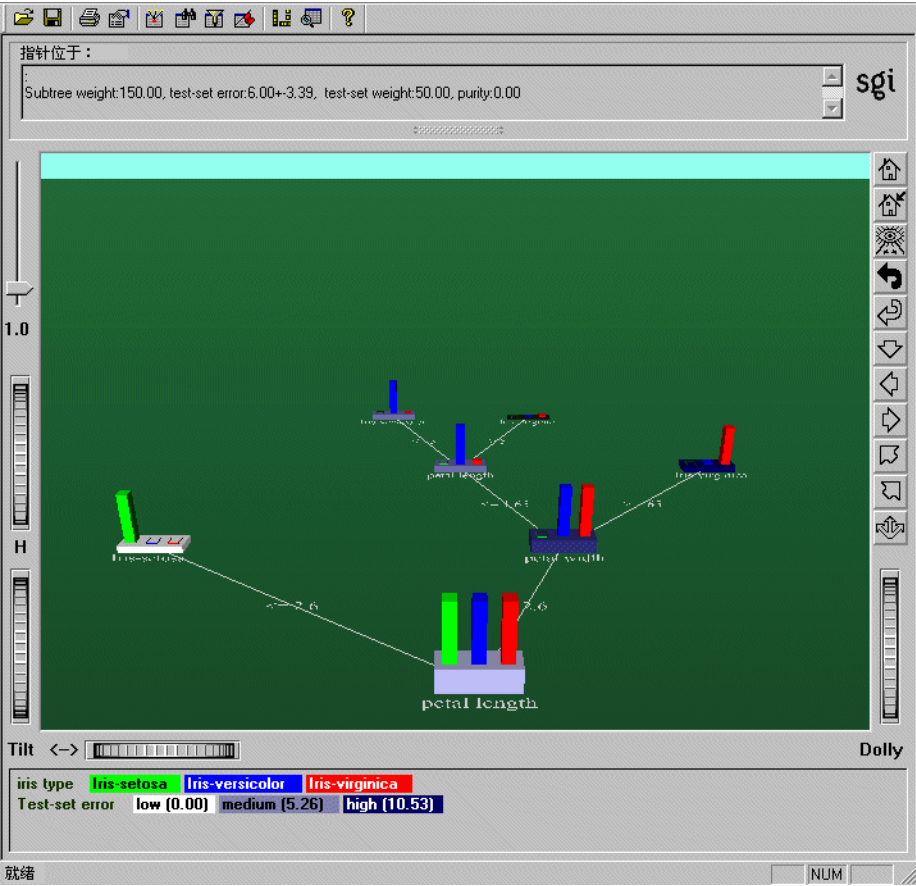


图 8-5 蝴蝶花数据集的决策树

其它有用的选项

“决策树可视化工具”使用“树可视化工具”进行显示。“树可视化工具”具有多个有用的设备，例如：搜索面板、筛选面板、全局观察窗口，它可以从“工具管理器”的“查看”菜单中进入。以上所述详见第5章，“用树可视化工具检查数据”和《MineSet 3.0 企业版参考指南》中的“树可视化工具”。

决策树也有各种选项，用于调整导入运算法则、剪枝和评价量度。以上所述详见《MineSet 3.0 企业版参考指南》中的“决策树”。

用选项树可视化工具查看结果

决策树和选项树可视化处理从根本上是类似的。但是，两种类型的树之间还是有一些区别：

- 选项树最左边的选项是决策树导入工具的唯一选项。当转到右边时，选项是按适合度得分降序排列的。适合度得分不必与显示的测试集错误相匹配。这是在意料之中的，因为导入工具使用的是非理想评分功能。测试集估计也有天然的可变性：测试集越大，估计就越准确。
- 选项节点的错误率可以和它每个子节点的错误率不同。因为选项节点平均了子节点的预测，因此它的错误率可能不同。在某些情况下，它的错误要比所有子节点的错误都要低，表现出平均的好处。
- 选项节点每个子节点上的示例（条形中显示的）分布与选项节点本身的示例分布完全一样。这是因为选项节点没有决策：选项全被显示为子节点。
- “选项树”可视化处理有一个很有用的在树中漫游的特性，即：在选项节点上单击鼠标右键。将显示子节点选项的列表。

在 MineSet 3.0\examples\cars-odt.treeviz 中可以发现简单的“选项树”价值检验。“选项树可视化工具”使用“树可视化工具”进行显示。“树可视化工具”具有多个有用的设备，例如：搜索面板、筛选面板、全局观察窗口。以上所述详见第5章，“用树可视化工具检查数据”。

选项树也有各种选项，用于调整导入运算法则、修剪和评价量度。以上所述详见《*MineSet 3.0 企业版参考指南*》中的“决策树”。

用回归树可视化工具查看结果

回归树和决策树可视化处理是类似的。它们都是由边连接的两类节点组成：决策节点和叶节点。

决策节点指定在该节点被测试的属性。用于测试属性的值（或值的范围）将显示在边上。每一个可能的属性值都刚好符合一条边。例如：图 8-3 中回归树的根测试属性年龄；从节点生出的两条边划分了该属性的值（小于 27.5，和大于或等于 27.5），以便每个可能值都能与右边或左边的分支相符合。如果值未知并且没有用问号标注在边上，那么就会预测当前节点的平均值或中值标签。

回归树中的叶节点预测一个值。在 `\examples\adult-rt.treeviz` 中可以找到此处讨论的有用的示例。在图 8-3 中，如果跟随最左边的分支从根到标注的叶 4002.300，就会发现回归树预测每周工作时间不足 35.5 小时的 19.5 岁以下的人平均总收入为 \$4002.30。

“回归树”中节点上的每个条形对应于连续标签值的一个子区间。每个节点覆盖的连续标签值范围可以不同。每个节点上的条形构成一个直方图指示记录权重（数）在该范围内分布情况条形数目由根处的记录权重决定。最左边的条形通常对应于最低的值。因此，每个节点的大小和中点可能不同。条形颜色指示条形覆盖的子区间的中点。最大范围有左边的蓝色和右边的红色指示。仅包括有限标签值范围的记录的节点的直方图范围不是从蓝色到红色。

节点的基准块具有高度。高度对应于到达该节点训练集记录的权重（如果没有设置权重，就是记录的数目）。通常权重越高，节点处的分布就越可靠。

将鼠标箭头放在节点上面，会显示以下信息（请参阅词汇表中术语的定义）：

- **子树权重**—被指向节点下方的子树中训练集记录的权重。该值被映射到基的高度。
- **平均值**: 连续标签的平均值。
- **标准偏差**: 连续标签的标准偏差。标准偏差越高，模型就越不可靠。
- **中值**: 连续标签的中值。
- **绝对偏差**: 连续标签的绝对偏差。绝对偏差越高，模型就越不可靠。

使用回归树主窗口预测数值

要预测记录的值，可以从树根开始，然后跟随该记录属性值指示的分支。按记录属性值跟随适当的边，就可以到达一个叶节点。与叶节点相关联的预测就是预测的记录值。

回归树的误差估计

评测分类器时，固有量度是错误的（分类器预测错误标签的示例数）。当提供了损失矩阵时，不同类型的分类错误可能具有不同的相关代价。此时，损失就是固有量度。关于详细信息，请参阅第 11 章中的“[定义损失矩阵](#)”。

对于预测真实值的回归任务，没有单独的固有评价量度。频繁使用的两种量度是**标准偏差**和**绝对偏差**。标准偏差是均方误差的平方根。绝对偏差是预测标签值和实际标签值差的绝对值的平均值。

其它有用的选项

“回归树可视化工具”使用“树可视化工具”进行显示。“树可视化工具”具有多个有用的设备，例如：搜索面板、筛选面板、全局观察窗口。以上所述详见第 5 章，“[用树可视化工具检查数据](#)”和《[MineSet 3.0 企业版参考指南](#)》中的“树可视化工具”。

回归树也有各种选项，用于调整导入运算法则、修剪、和评价量度。以上所述详见《*MineSet 3.0 企业版参考指南*》中的“回归树”。

用决策、选项和回归树预测

到此为止，本章已经讨论了利用已知的结果（例如：客户波动的或未波动的，和可食的或有毒的）对数据进行分类，但是对于预测未知或将来的结果，决策、选项、和回归树可能是最有用的。可以根据其中的类已知的数据建立模型，然后使用它对其中的类未知的新数据进行分类。关于建立和应用模型的详细信息，请参阅第 7 章，“理解预测建模”和第 11 章，“改进预测建模”。

用决策表分类器和可视化工具建模和预测

本章讨论“决策表”的特性和功能，主要包括以下几个部分：

- 第 147 页的“决策表分类器概述”
- 第 149 页的“运行决策表”
- 第 151 页的“用决策表可视化工具查看结果”
- 第 156 页的“用决策表预测”

我们提供了一些配置和数据文件样例，用于演示“决策表”的特性和功能。这些文件安装在 *MineSet 3.0* 下的 *\examples* 目录中，位于 *MineSet* 的初始安装目录。

决策表分类器概述

决策表是执行分类的预测建模工具（关于分类器和预测建模的详细信息，请参阅第 1 章，“数据挖掘和 *MineSet* 工具概述”和第 7 章，“理解预测建模”）。它结合了一个导入工具（生成决策表模型的一个运算法则）和一个可视化工具。与证据模型不同，“决策树”模型不假设各属性是相互独立的。

决策表是数据的层次化分类，在每个层次上应用两个属性。“决策表”导入工具识别出对于分类数据最重要的属性（列），然后，可视化工具以一系列块状图图形化显示出生成的模型。可视化处理中的每个块图可以进一步被划分为更小的块，代表下一层最重要属性。每个可视化处理可以包含多个等级，代表重要性递减的属性。图 9-1 所示为‘蘑菇’数据集“决策表”可视化处理的顶级，其中确定可食性的两个最重要属性为气味和孢子印的颜色（spore print color）。

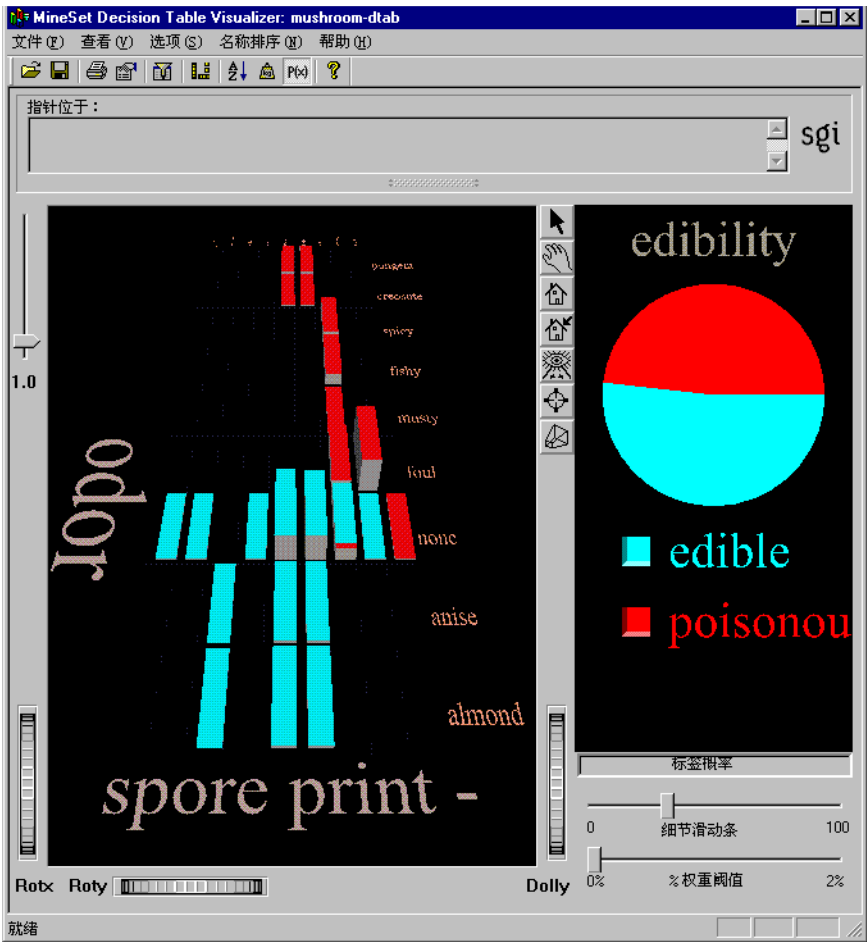


图 9-1 ‘蘑菇’数据集的决策表

运行决策表

- 创建“决策表”分类器最简单的方法是：在“工具管理器”（关于详细信息，请参阅《*MineSet 3.0 企业版参考指南*》）中选择：
- 1. 从“工具管理器文件”菜单中，连接并登录到服务器；
 - 2. 从同一个菜单中选择“打开新的数据文件”，并选择或键入需要的文件名。
 - 3. 在“数据目标”窗格中，单击“挖掘工具”选项卡（图 9-2），然后从下面一行选项卡中选择“分类”。

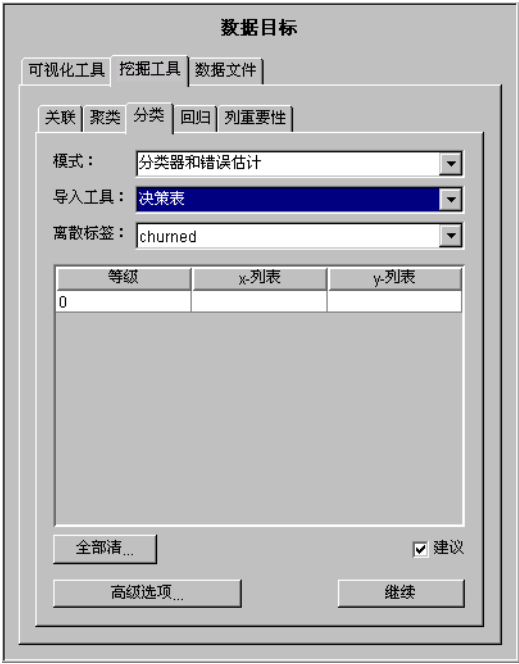


图 9-2 显示分类器的工具管理器数据目标面板

- 4. 从弹出的“模式”菜单中选择一个模式。关于四种模式的详细信息，请参阅第 7 章，“理解预测建模”。
- 5. 从弹出的“导入工具”菜单中，选择“决策表导入工具”。

6. 从弹出的“离散标签”菜单中，选择希望为标签使用的属性。
7. 在 x- 列表和 y- 列表下拉菜单（请参阅图 9-2）中，指定希望查看的属性。可以用以下三种方法完成该操作：
 - 选中“建议”框，可以让 MineSet 自动选择属性。MineSet 将识别对于预测类最有用的属性，并显示它们。
 - 也可以从“数据目标”面板的下拉菜单中手工选择属性。如果需要一个起点，可以使用“列重要性”查找最重要的属性（请参阅第 3 章中的“查找重要的列”）。
 - 还可以选择一部分属性，然后让 MineSet 选择其余的属性。按照以上所述方法选择属性，然后选中“建议”框。MineSet 将映射其余的有关属性。
8. 单击继续以运行“导入工具”。

“工具管理器”窗口底部的“状态”面板将显示进程与统计结果。还可以单击进程对话框中的取消或立即显示 可视化 来中断自动属性建议过程。单击立即显示 可视化 将停止当前服务器的计算，并使用中间的结果为已经映射的列构建决策表。

注意：关于其他“决策表”导入工具选项的详细信息，请参阅《MineSet 3.0 企业版参考指南》中的“决策表”。

用决策表可视化工具查看结果

“决策表”可视化工具有两个窗格，左边的“决策表”面板和右边的“标签概率”面板。图 9-3 所示为蘑菇数据集的“决策表”可视化处理，其中显示了最高两层的细节。

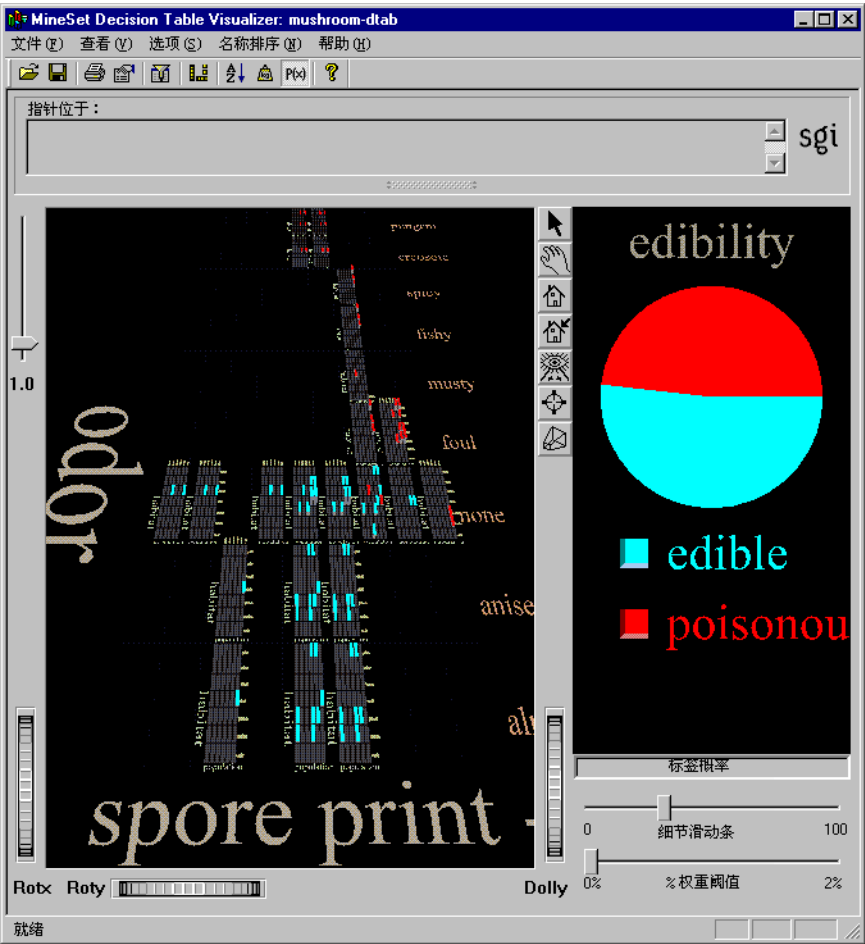


图 9-3 蘑菇数据集的决策表可视化处理

查看“决策表”窗格

左边的“决策表”面板由块状图组成，块状图就是带有彩色切片的方块图，那些切片代表着具有特定属性值的记录的标签概率。标签概率代表具有这些指定属性值的记录属于某个特定类的概率。例如：在图 9-3 中，“决策表”运行在蘑菇数据集上。产生的图显示具有白色孢子印和腥臭气味的蘑菇有毒的概率是 100%。然而，如果具有白色孢子印，却没有气味，那么它有毒的概率只有 7.69%，可以食用的概率为 92.31%。要了解这些百分比值，可以选择左边的一个块，然后将鼠标箭头放在右边标签（可食的或有毒的）旁边的彩色框上面。百分比值将显示在菜单条和主窗口之间的区域内。

用鼠标右键单击，可以将“决策表”面板中的元素进一步细分为越来越小的块状图，此过程叫做细化下寻。要更仔细地检查块状图：

- 要了解当前细节等级上两个属性的值，可以将鼠标箭头（在选择模式下）放在所需要的块状图上。属性值和记录权重将显示在菜单条和主窗口之间的区域内。块状图的高度与权重成正比（请参阅词汇表中的词条权重）。
- 要细化下寻到下一个细节等级，可以将鼠标箭头放在需要的块状图上，然后单击鼠标右键，或者单击背景在所有块上进行全局细化下寻。图 9-5 所示为对‘蘑菇’数据集细化块的特写视图。
- 要从较低层次中概化上寻回去，可以在单击鼠标右键的同时按住 **Ctrl** 键（或使用鼠标中键）。通过在背景上按住 **Ctrl** 键并同时单击鼠标右键（或使用鼠标中键）可以进行全局概化上寻。
- 要了解更高一层图块代表的属性值和记录权重，可以将鼠标箭头置于基准块（块状图下面的灰色图块）上。
- 要了解某图块的定义值，可以先选中最粗细节级别的基准块，然后逐层选中相应层次的基准块，直到当前层次的上一层为止。相关的一对属性值将出现在菜单条下面的选择面板之中。

在每个细节等级上，一个属性的名称显示在块状图组的左边，它的值显示在右边；另一个属性的名称显示在块状图组的底部，它的值显示在顶部。如果属性的总数为奇数，那么最低的等级将只有一个属性。

“标签概率”窗格

“标签概率”窗格（在“工具管理器”窗口的右侧）将显示一个表示整个数据集标签概率的圆饼图。在圆饼图下显示有所有类标签的列表。

如果希望更仔细地检查标签概率，可以按照以下步骤进行：

- 要了解指定属性集合的标签概率，可以单击所需的块状图。“标签概率”窗格中的圆饼图显示了属性集合记录的标签概率，这个属性集合由块状图代表。
- 要了解属性组合的标签概率，可以使用 **Ctrl-** 单击所需的块状图。通过 **Ctrl-** 单击，可以选择许多场景中不同细节等级上的块。可视化工具窗口右边的“标签概率”窗格将显示选定记录集的标签概率。
- 要了解对每个标签指定的百分比和置信度，可以将鼠标箭头放在需要标签旁边的彩色框上。数字将显示在菜单条和“决策表”窗格之间的区域内。

“决策表”示例

在这个介绍如何使用“决策表”的示例中，假设希望指出哪些蘑菇可以被安全地食用，哪些蘑菇是有毒的。开始时，需要从蘑菇数据集中创建一个决策表（请参阅图 9-1）。左边的面板显示出决策表的最高细节等级。最高等级选择了气味和孢子印属性来进行分类，因为它们是由“列重要性”运算法则确定的两个最重要属性。只有一个顶级块具有多个类，其属性值气味等于无，孢子印颜色等于白色。当细化下寻到下一个等级时（在图上单击鼠标右键），块图将表示属性：生长地和总数（请参阅图 9-4）。该细化下寻区域的特写如图 9-5 所示。在这个等级上，不存在混合的块。因此，如果发现具有白色孢子印并且没有气味的蘑菇与其他一些蘑菇生长在树林中，就可以确定它是有毒的。

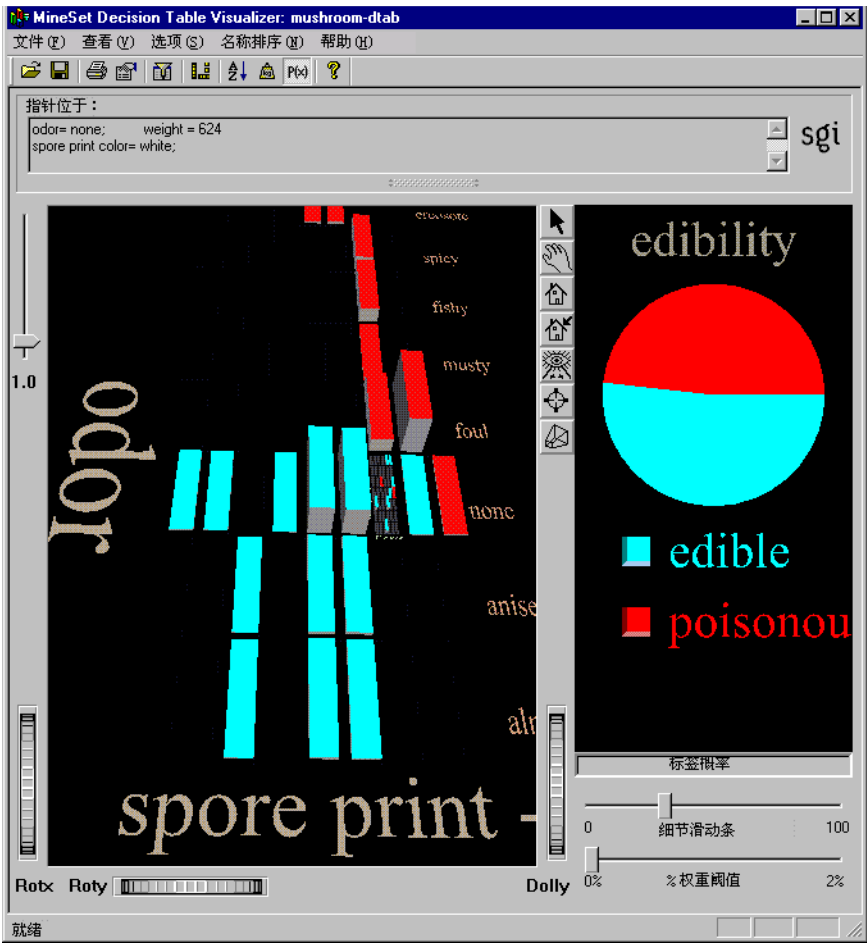


图 9-4 在蘑菇数据集中细化下寻的决策表

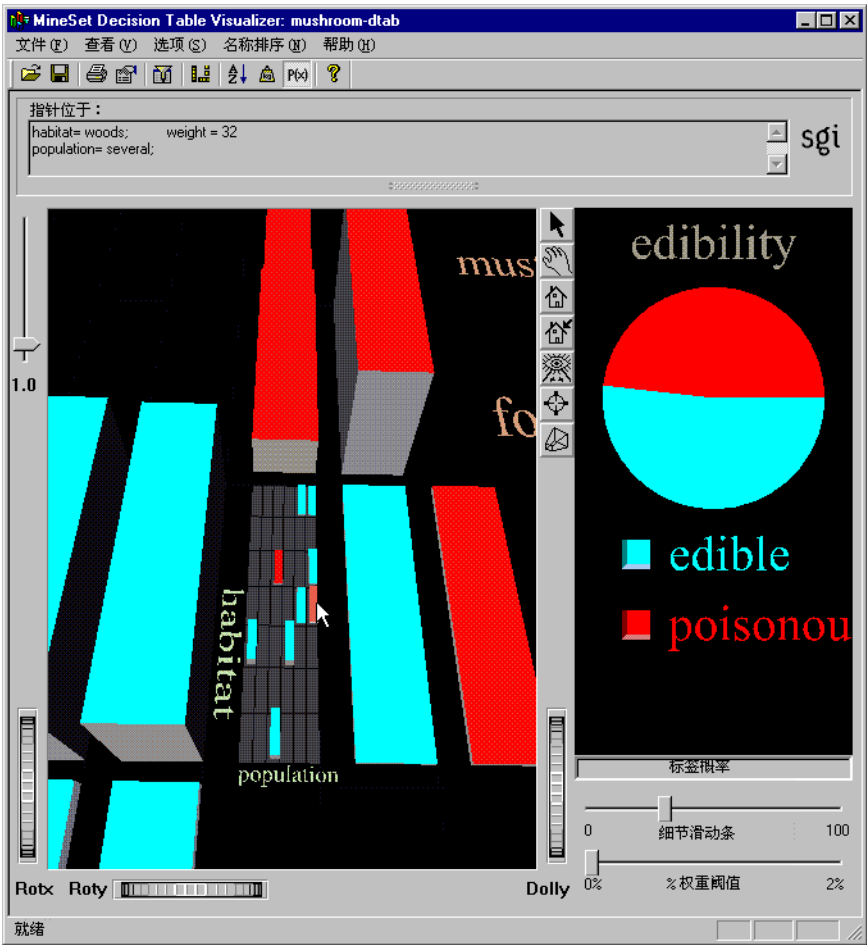


图 9-5 细化下寻决策表的特写

用决策表预测

到此为止，本章已经讨论了利用已知的结果（例如：可食的或有毒的）对数据进行分类，但是对于预测未知的或将来发生的结果，“决策表”可能是最有用的。可以根据具有已知类信息数据建立模型，然后使用它对未知类信息的新数据进行分类。关于建立和应用模型的详细信息，请参阅第 7 章，“理解预测建模”和第 11 章，“改进预测建模”

关于菜单的详细信息，请参阅 《*MineSet 3.0 企业版参考指南*》中的“决策表”。

用证据分类器和可视化工具建模和预测

本章讨论“证据分类器”和“证据可视化工具”的特性和功能，主要包括以下几个部分：

- 第 157 页的“证据分类器和可视化工具概述”
- 第 159 页的“运行证据工具”
- 第 161 页的“用证据可视化工具查看结果”
- 第 169 页的“用证据分类器预测”

提供了一些配置和数据文件样例，用于演示“证据可视化工具”的特性和功能。这些文件在 *MineSet 3.0* 下的 *\examples* 目录中，位于 *MineSet* 的初始安装目录。

证据分类器和可视化工具概述

证据工具是执行分类的预测建模工具（关于分类器和预测建模的详细信息，请参阅第 1 章，“数据挖掘和 *MineSet* 工具概述”和第 7 章，“理解预测建模”）。它结合了一个导入工具（生成 Naive Bayes 模型的运算法则）和一个可视化工具。与“决策树”模型不同，“证据”模型假设属性是独立的，它甚至在属性不是独立时仍然能产生合理的结果。

“证据可视化工具”显示来自分类器的信息，如图 10-1 所示。可视化工具可以帮助理解用于分类的指定属性（列）的重要性。

“证据可视化工具”窗口

一旦建立了分类器，结果就会显示在“证据可视化工具”窗口中。最初，左边窗格包含分类器使用的每个属性的块状图行（图 10-1）。块状图类似于其中显示比例的圆饼图，除了它是带有矩形薄片的方块。可以在块状图（代表证据）、圆饼图（代表概率）和条形图（提供选定标签的详细信息）之间进行切换（关于详细信息，请参阅第 161 页的“用证据可视化工具查看结果”）。

图 10-1 所示为证据视图，图 10-2 所示为概率视图。

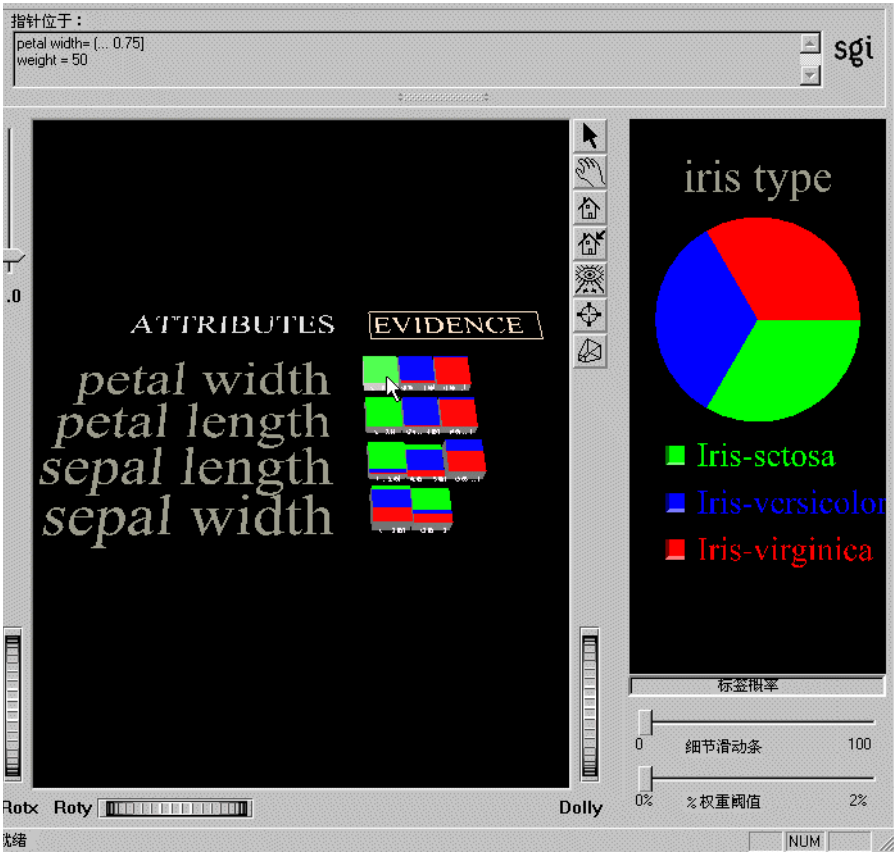


图 10-1 应用到蝴蝶花数据集的证据可视化工具

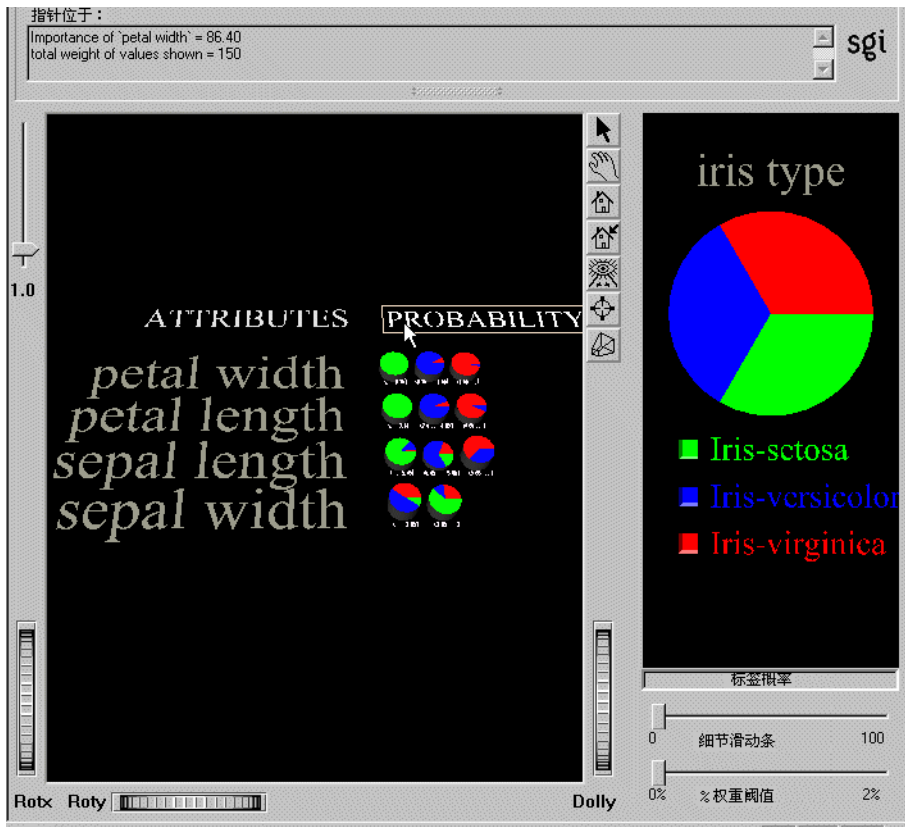


图 10-2 显示概率的证据可视化工具

运行证据工具

运行“证据”工具最简单的方法是：在“工具管理器”（关于详细选项，请参阅《*MineSet 3.0 企业版参考指南*》）中：

1. 从“工具管理器文件”菜单中，连接并登录到服务器。
2. 从同一个菜单中选择“打开新的数据文件”，并选择或键入需要的文件名。
3. 在“数据目标”窗格中，单击“挖掘工具”选项卡（图 10-3），然后从下面一行选项卡中选择“分类”。

- 4. 从弹出的“模式”菜单中选择一个模式。关于四种模式的详细信息，请参阅第 7 章，“理解预测建模”。
- 5. 从弹出的“导入工具”菜单中，选择“证据”。
- 6. 从弹出的“离散标签”菜单中，选择希望为标签使用的属性。可食性对于蘑菇数据集是一个很好的属性。
- 7. 单击继续以运行“导入工具”。

“工具管理器”窗口底部的“状态”窗格将显示进程与统计结果。

注意： 关于其他“证据”导入工具选项的详细信息，请参阅 《MineSet 3.0 企业版参考指南》中的“证据导入工具选项”。



图 10-3 “工具管理器数据目标”窗格

用证据可视化工具查看结果

一旦运行了“证据可视化工具”，就可以用多种方法查看结果。在可视化工具窗口的左边窗格中，可以在证据视图、概率视图和条形视图之间切换：

- 证据视图显示块状图，代表证据（请参阅第 162 页的“证据视图”）。
- 概率视图显示圆饼图，代表概率（请参阅第 165 页的“概率视图”）。
- 条形视图显示条形图，代表对特定标签有利和不利的证据（请参阅第 166 页的“条形视图”）。

“标签概率”窗格（可视化工具窗口的右边）显示一个圆饼图，代表整个数据集中标签值的分布（请参阅[“查看证据视图中的概率”](#)）。该窗格是为左边窗格中的所有三种视图（证据、概率和条形）显示的。

证据视图

当“证据可视化工具”第一次运行时，主面板将显示证据视图（[图 10-4](#)）。

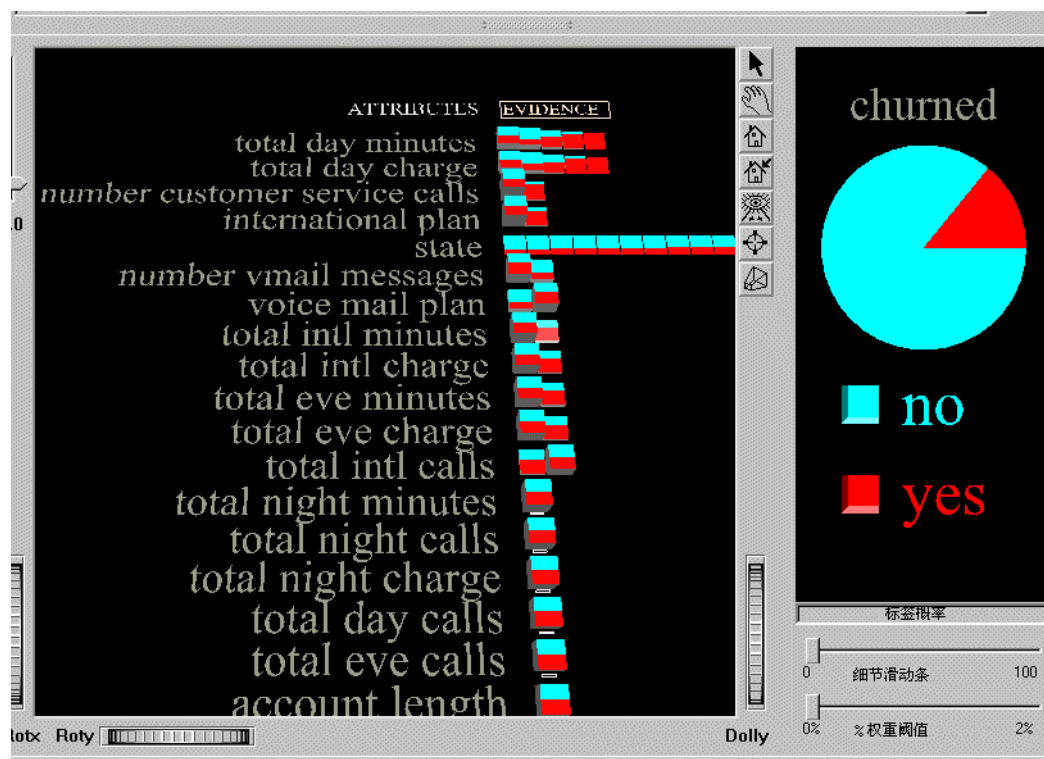


图 10-4 用于客户波动数据集的证据可视化工具块状图

解释证据视图

“证据视图”为数据集中的每个属性或列显示一行块状图。每一行中对于每个离散属性值都有一个方块。如果属性值不是离散的，而是连续的， MineSet 自动将值分为几个范围和组（例如：0 到 5、5 到 10，等等），划分方式是使相邻块之间的差别最大化。块的高度与具有该属性值的记录权重成正比（请参阅词汇表中的词条权重）。如果没有设置权重属性，那么高度就代表记录的数目。

在块状图中，薄片的尺寸代表条件概率分布（证据），而不是直接概率。例如：在客户波动数据集中，如果一个块的一半为“是”，一半为“否”，就表示该块代表的用户特定属性值与总体的波动概率相同。换言之，该属性值没有预测值。相反，如果一个块的90%为“是”，10%为“否”，就表示具有该特定属性值的用户比总体更可能客户波动。但是，它并不表示：用户有90%客户波动的概率，因为整个概率取决于许多变量，而不只是这一个。关于更详细的讲解，请参阅《[MineSet 3.0 企业版参考指南](#)》中的“证据导入工具”。

行的次序指示预测标签时属性的相对重要性。每个属性都被分配了一个介于1到100之间的重要性值。将鼠标箭头放在场景中属性的名称上面，就可以看到这个值，该信息显示在菜单条和可视化工具面板之间的区域内。例如：在图10-5所示的蘑菇数据集中，蘑菇的气味是可食性的极端标志，因此气味属性被排在了列表的首位，重要性值为90.69。相反，茎的形状根本无法指出可食性，因此茎的形状被排在了列表的最后，重要性值为.75。

菌盖的颜色属性也是一个有趣的例子。它的重要性被排在了后面，因为大部分蘑菇的菌盖是白色的。菌盖为白色的蘑菇几乎被平均分成了可食的和有毒的，非常接近先验概率（即：数据集中任何蘑菇有毒的一般概率）。因此，从统计学角度来看，该属性没有为任何一类增加证据，因此重要性被排在了后面。但是，对于棕色、橙色或黄色的菌盖颜色（菌盖颜色行中的第一、二、四块），图象就完全不一样了。棕色或橙色菌盖的蘑菇100%可以食用。黄色菌盖的蘑菇100%是有毒的。尽管这听起来很有说服力，但是菌盖不是白色的蘑菇非常少，因此菌盖颜色不能成为好的区别特征。

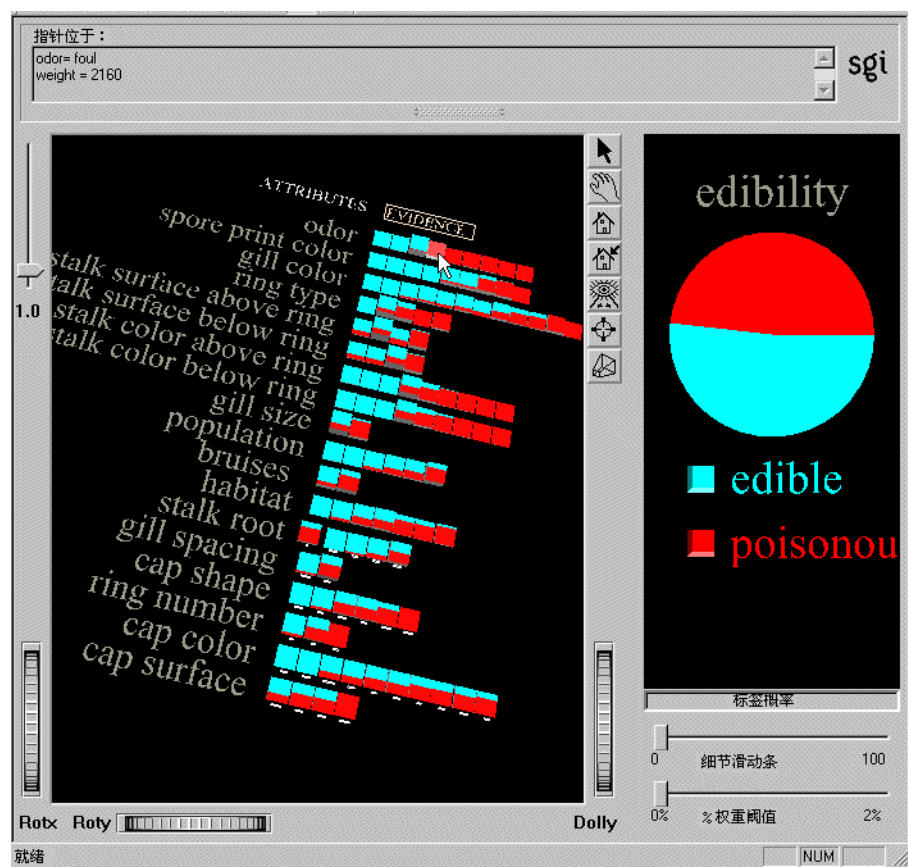


图 10-5 证据可视化工具中的蘑菇数据集

查看证据视图中的概率

当“证据可视化工具”第一次打开时，“标签概率”窗格将显示整个数据集的标签概率。例如：在图 10-5 中，使用客户波动数据集，随机记录具有“是”客户波动标签的概率为 14.14%。如果将鼠标箭头放在标签旁边的彩色框上，该标签的概率和权重将被显示在窗口的左上部。

如果希望查看具有某个属性特定值的记录的概率圆饼图，可以单击代表该值的块（鼠标必须处于选择模式以选择属性）。“概率”面板将只显示符合选定值的记录的概率圆饼图。要取消对图的选定，可以单击主窗格背景的任何地方。

要查看具有特定属性组合的记录的概率，可以通过 **Ctrl-** 单击它们来选择多个块。“概率”面板将显示具有这种属性组合的记录的概率圆饼图。因为每个属性的概率都要被计算，就好像它们之间相互独立一样，因此组合概率可能不完全准确。会出现一个警告提醒您这一点，将它们用作估计。只有当属性真正独立的时候，它们才是完全准确的，但这样的情况很少。

有时在选择属性组合时，可能选择一些相互排斥的属性。在这种情况下，“标签概率”窗格将显示一个灰色的圆饼。这表明具有该特定组合值的记录很少或者根本就没有。

要重新设置“标签概率”以显示先验概率，可以在主窗格的背景上单击任何位置。

概率视图

概率视图（图 10-6）显示代表概率的圆饼图，而不是块状图。要从证据视图进入概率视图，可以单击图顶部白色轮廓框中的“证据”一词，或在可视化工具“查看”菜单中切换“显示为证据”复选框。要返回到证据视图，单击“概率”一词或再次切换复选框。

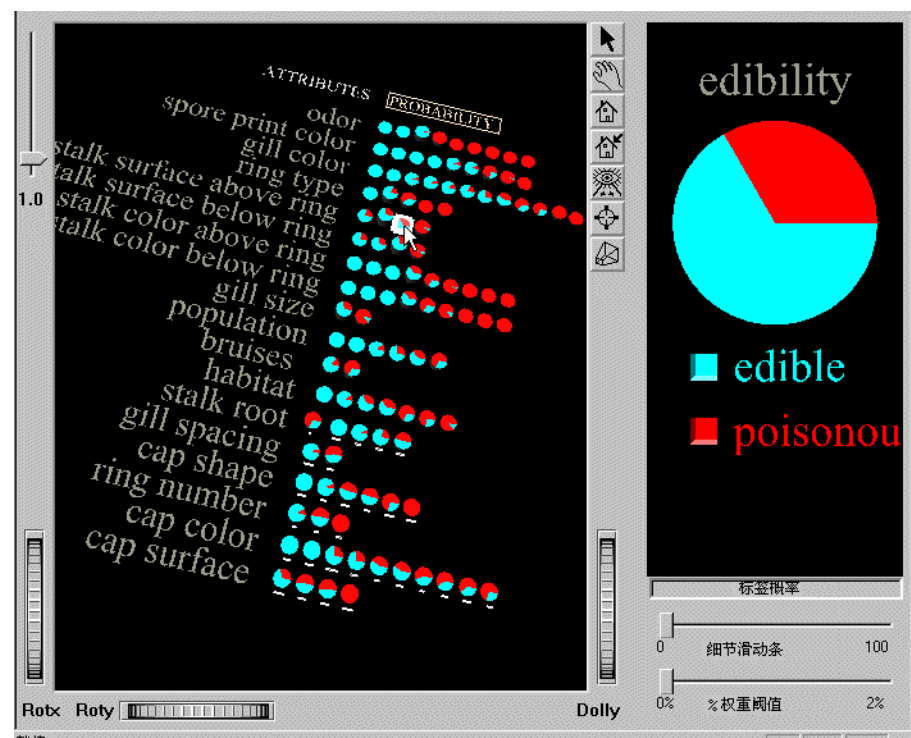


图 10-6 证据可视化工具圆饼图

条形视图

条形视图显示有利于（或不利于）选定标签值的证据。要进入条形视图，可以在“标签概率”面板中单击所需标签旁边的彩色框。接着，主窗口将显示条形图（图 10-7）。每个条形的高度代表有利于选定标签值的证据。要退出条形视图，可以再次单击彩色框。

单击条形视图顶部的白色轮廓框中的“有利（for）”一词，可以切换到“不利证据”模式。要返回到“有利证据”模式，可以单击“不利（against）”一词。

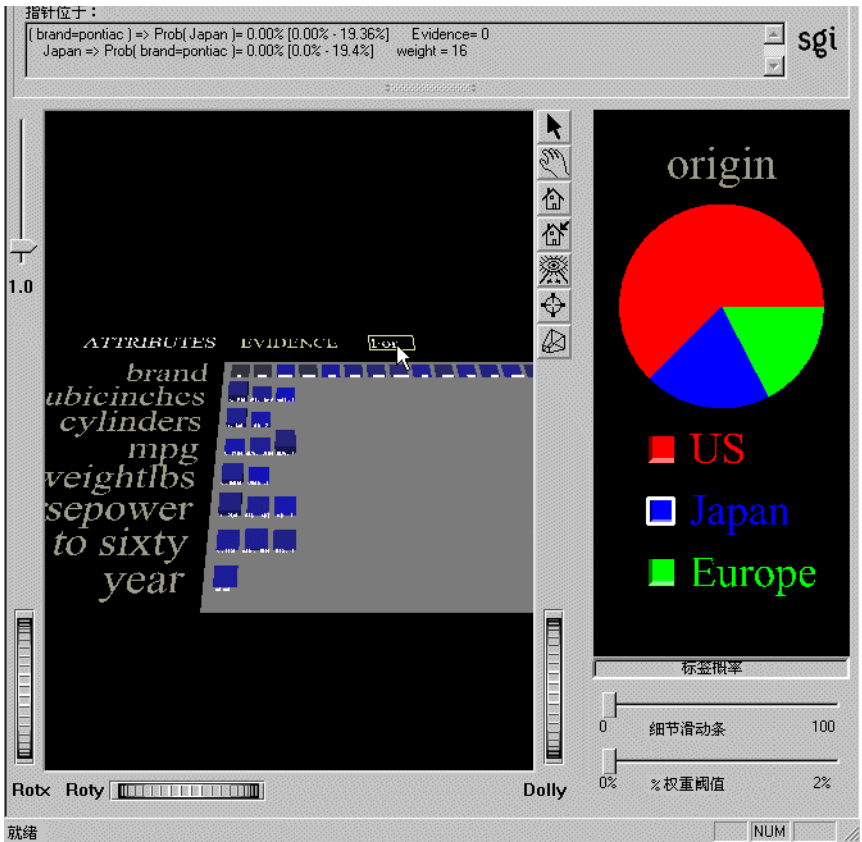


图 10-7 汽车数据集标签值“日本”被选定的

当确定哪些值对预测特定标签值最有帮助时，“有利证据”会很有用。相反情况下，即：当确定哪些值对预测标签不会出现最有帮助时，“不利证据”会很有用。关于这些值的计算方法详细的技术说明，请参阅《MineSet 3.0 企业版参考指南》中的“证据导入工具”。

条形的颜色可以从完全饱和的颜色到完全灰色。条形的灰度以 95% 的置信区间为准。而这又取决于该值的权重。因此，近似于灰色的条形具有低的权重和大的置信区间。这些条形的高度不可能准确。相反，完全饱和条形的高度和相应的证据值是可以信赖的，因为它是基于大权重的，代表许多记录。当条形被突出显示时，确切的记录数（权重）反映在文本输出行中。

改变证据可视化工具视图

可以使用[表 10-1](#) 所示的各种操作更好地检查不同视图中的结果。

表 10-1 操作 “证据可视化工具视图”

目的:	操作:
在块状图（证据）和圆饼图（概率）之间切换	单击白色轮廓的“证据”或“概率”一词，或切换“视图”>“显示为证据”。
在块状图和圆饼图或条形图之间切换	在“标签概率”窗格中单击所需标签旁边的彩色方块。
在条形图和块状图或圆饼图之间切换	在“标签概率”窗格中单击所需标签旁边的彩色按钮。
在条形视图中的“有利”和“不利”之间切换	单击“有利”或“不利”旁边的白色轮廓框。
显示特定块或圆饼的属性值和权重	将鼠标箭头放在块或圆饼上。
显示标签概率	将鼠标箭头放在“标签概率”窗格中标签旁边的框上。
为具有特定属性值的新记录（在右边窗格中）显示标签的预期分布	单击代表该值的块。

表 10-1 操作 “证据可视化工具视图”

目的:	操作:
显示属性重要性和总权重	将鼠标箭头放在左边的属性名称上。
为带有特定属性值组合的记录获得估计的标签概率分布	通过 Ctrl- 单击选择两个或多个块，并检查“标签概率”窗格中的圆饼图。
清除选择	单击左边窗格背景区域的任何地方。
查看具有选定属性值的所有记录	单击代表需要的属性值组合的块，然后从“选择”菜单中选择“显示原始数据”。
查看除具有选定属性值记录以外的所有记录	在“选择”菜单中选“互补细节追溯”，然后从同一个菜单中选择“显示原始数据”。
在图之间放大或减小高度差别	使用主窗口左侧的高度滑动条。
减少（或增加）显示的属性行数，从最少（或最多）的预测属性开始	使用细节滑动条。删除重要性小于滑动条值的属性。
筛选出低计数或低权重的属性值	使用权重阈值滑动条。删除权重小于滑动条指示的百分比的值。
为一些属性值已知的新记录预测类标签	单击对应于这些属性值的图。右侧圆饼图中最大的部分对应于预测的类。

用证据分类器预测

迄今为止，讨论主要集中于使用“证据分类器”和“证据可视化工具”检查标签已知的数据。实际上，这个工具在预测标签未知的数据中的标签时最有用。在这个例子中，使用已知的数据创建了模型（请参阅《MineSet 3.0 企业版参考指南》中的“训练集”）并保存了它，然后又将其应用到标签未知的数据上。请参阅《MineSet 3.0 企业版参考指南》中的“应用模型”。

关于可视化工具菜单的详细信息，请参阅《MineSet 3.0 企业版参考指南》中的“证据可视化工具菜单”。

改进预测建模

本章建立在前面章节发展出的概念上，使用“工具管理器”检查分类中出现错误的代价，并将数据挖掘操作按实际需要进行调整。MineSet 导入工具的高级选项可以提高准确性，补偿错误的代价，并且确定从何时起，追加投资将不会再带来任何收益。本章包括以下部分：

- [第 171 页的“确保模型的准确性”](#)
- [第 176 页的“用混淆矩阵和损失矩阵调整模型”](#)
- [第 183 页的“用上升曲线和 ROI 曲线评测模型”](#)

在本指南的联机版本（可从“帮助”菜单中进入）中可以看到彩色的屏幕插图。

确保模型的准确性

可以用不同的方法来检查已建立的模型的准确性。在[第 121 页的“评测预测模型”](#)已经讨论了几种方法。本部分继续讨论使用不同数据源验证模型的方法。如果还没有模型，请回到[第 7 章，“理解预测建模”](#)，获得创建模型的帮助。

测试模型

如果已经创建了模型，MineSet 允许使用“测试模型”工具在当前使用的数据上测试它。数据集中列的名称和类型必须与创建模型时使用的相同。

与“应用模型”（详见第 7 章中的“应用预测模型”）不同，“测试模型”要求表包含的标签列的名称和类型也要与建立模型时使用的相同。原因是测试必须从已具有正确答案的数据开始工作。测试文件将被分配一个默认名称，后缀为 *-test*，如 *iris-test*。

下面的示例从蝴蝶花数据集中创建一个测试文件。一旦登录到服务器上，就可以选择数据集（示例文件是 *MineSet 3.0\data\iris.schema*，位于 MineSet 最初安装的目录中）。

要在当前数据集上测试模型，请按以下步骤进行：

- 1. 在“工具管理器数据转换”窗格中单击应用模型按钮。
- 2. 在“测试和应用模型”对话框中，单击“测试模型”选项卡。
- 3. 从对话框顶部出现的可用模型列表中选择模型的名称（如：*iris-dt.class*），然后会出现默认测试文件名。（图 11-1）。

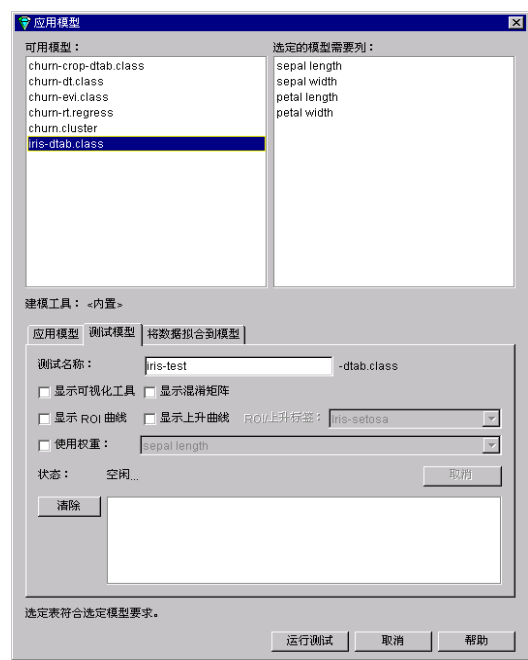


图 11-1 测试模型面板

如果当前表中遗漏了一些列，那么它们的名称就会突出显示在右边“选定模型需要列”列表中。请选择其他模型或恢复被遗漏的列。“测试模型”面板将提供以下选项：

表 11-1 测试模型面板选项

选择	操作
测试名称	显示测试文件的名称，可以更改。
显示 可视化 工具	显示具有测试集表的分类器可视化处理。
显示混淆矩阵	显示基于表记录的模型混淆矩阵。请参阅第 178 页的“显示混淆矩阵”。
显示 ROI 曲线	显示使用指定标签值的模型 ROI 曲线。请参阅第 185 页的“使用投资回报曲线寻找销售利润”。
显示上升曲线	显示使用指定标签值的模型上升曲线。请参阅第 183 页的“用上升曲线和 ROI 曲线评测模型”。
ROI/ 上升标签	显示可用标签值的弹出菜单。
使用权重	从加权记录的弹出菜单中选择可用的属性。请参阅第 3 章中的“加权记录”。

4. 单击运行测试。“测试模型”面板底部的文本字段中将显示出结果，同时显示选定的可视化工具选项。

在这个例子中，整个 *iris.schema* 数据集运行在事先创建好的模型（*iris-dt.class*）上，在 150 个测试实例中发现了 4 个错误。

将数据拟合到模型

当有新数据时，可以使用“将数据拟合到模型”面板（图 11-2）将当前表中的数据拟合到事先创建好的模型中。虽然，这样做生成了一个与原来模型结构相同的新模型；但是，新模型使用新数据更新了概率估计（请参阅“在误差估计中修正”）。因为表中的所有数据都被拟合到新模型的结构中，因此不存在误差估计。“将数据拟合到模型”不能被用于使用推进建立的模型。请参阅下面的“用推进提高准确性”。如果需要评测建立在独立测试集（与拟合数据不相交）上的新模型的性能，可以使用“测试模型”工具。

要将数据拟合到现有的模型，请按以下步骤进行：

- 1. 在“工具管理器数据转换”窗格中单击应用模型按钮。
- 2. 在“测试和应用模型”对话框中单击“将数据拟合到模型”选项卡。（图 11-2）在表 11-2 中会出现拟合到模型选项。



图 11-2 “将数据拟合到模型”面板

“将数据拟合到模型”面板（图 11-2）中有以下选项：

表 11-2 将数据拟合到模型选项

选择	操作
显示 可视化工具	显示新模型的可视化处理（仅适用于基于树的模型）。
用盘显示训练集	设置盘用以指示可视化处理中训练集数据所'的比例。
使用权重	从弹出菜单中选择用作记录权重的属性。

在误差估计中修正

当使用一个记录集对模型修正时，就会更新概率估计，但不会改变模型的结构。修正类似于将数据拟合到模型，这种模式默认情况下使用修正。修正在以下情况中很有用：

- 从小训练集中建立了一个模型结构，然后用大数据集对其修正。通常这比导入更大的模型结构要快。
- 需要更高的准确性。当计数、权重和概率表现在模型的结构中时，它们反映了所有的数据，而不仅仅是训练集内的部分数据。

启用修正：

1. 在“工具管理器”窗口的“数据目标”窗格中，单击“挖掘工具”选项卡。
2. 单击“分类”选项卡。
3. 选择“分类器和错误估计”模式，然后单击高级选项按钮，打开“高级选项”对话框。单击“误差估计选项”下的“修正”测试集复选框。当“推进”被启用时，复选标记被禁用。

用推进提高准确性

在一些情况下，错误率在选择模型时是最重要的准则，因此需要启用推进。推进是一种案算法，它创建多个不同的模型并使用加权投票方式将模型的预测结合起来。推进将导入过程集中于相对模型更为困难的数据示例上，从而提高模型的准确性，但是，‘被推进’的模型不能可视化。

要在选定模型上激活推进：

1. 在“工具管理器数据目标”窗格上，单击“挖掘工具”。
2. 单击“分类”选项卡。
3. 在任一模式中单击高级选项按钮，打开“高级选项”对话框。在“导入工具”选项部分单击“推进”（无可视化）复选框。

推进经常但不是总能提高准确性。推进模型无法可视化，但是仍然可以查看混淆矩阵、上升曲线、学习曲线和 ROI 曲线。推进是计算性很强的过程，其运行时间通常是相应不带推进的导入工具运行时间的 25 倍。推进模型不能使用修正。

注意： 虽然推进使用的理论没有推广超出双类的问题，MineSet 还是可以让您对具有任意多个取值的标签进行推进。此时，推进不太可能改进错误率。

推进反复地为训练集指定新的权重，并为被重新设置权重的训练集导入模型。请参阅《*MineSet 3.0 企业版参考指南*》。以获得关于推进的详细信息。

用混淆矩阵和损失矩阵调整模型

当在分类中发现错误时，知道分类器发生混乱的方式以及这种混乱造成的损失是很有帮助的。MineSet 提供了两种互补的矩阵来解决这个问题：混淆矩阵和损失矩阵。

使用混淆矩阵来研究错误

假设要将蘑菇分类成有毒的和可食用的。将蘑菇分类为有毒的可能需要花费 2 美元（蘑菇的成本）。但是，将一个有毒的蘑菇分类为可食用的，可能会使您住在医院中花去 10,000 美元（乐观估计）。MineSet 的混淆矩阵和损失矩阵的功能可以处理这种情况，并且可以给出错误和不正确预测的详细图片。

图 11-3 所示为蘑菇数据集的混淆矩阵，其决策树仅使用了 10% 的数据进行训练，远少于建议的训练集。关于详细信息，请参阅第 7 章，“理解预测建模”。

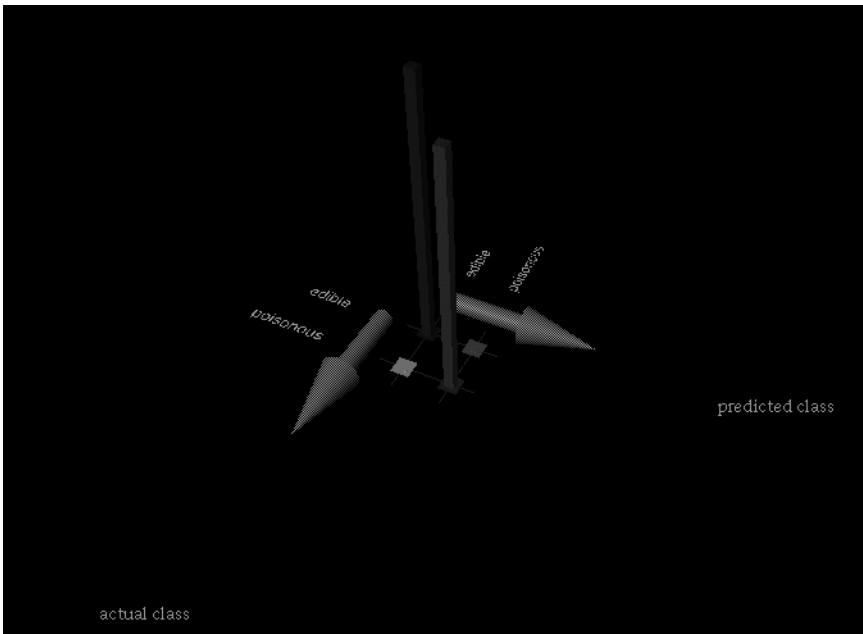


图 11-3 蘑菇数据集的混淆矩阵

代表有毒蘑菇的 8 个记录被分类为可食用的（0.1%）；代表可食用蘑菇的 15 个记录被分类为有毒的（0.2%）。其余的 3793 个可食用蘑菇和 3496 个有毒蘑菇的分类正确。尽管模型错误率仅为 0.31%（小于 1%），但是按照上一段示例中的估计损失为 $\$10000 \times 8 + \$2 \times 15 = \$80,030$ 。

混淆矩阵用一个对错误类型的可视化处理，给出关于模型出错的详细图片。两个坐标轴分别代表测试集（支持集）中预测类值和实际类值。混淆矩阵的将先于修正计算。下面介绍在上述情景中如何使用混淆矩阵。

显示混淆矩阵

在本例中，使用“工具管理器”窗口打开数据文件 *mushroom.schema*，然后研究蘑菇可食用和有毒分类中的混乱情况。该文件在 *MineSet 3.0\data* 目录中，位于 MineSet 的初始安装目录。

- 1. 在“工具管理器数据目标”窗格中单击“挖掘工具”选项卡；然后单击“分类”选项卡并从以下弹出菜单中进行选择：

模式：分类器和错误估计

导入工具：决策树

离散标签：可食用性
- 2. 单击高级选项，将出现“分类器高级选项”窗格，如图 11-4 所示。

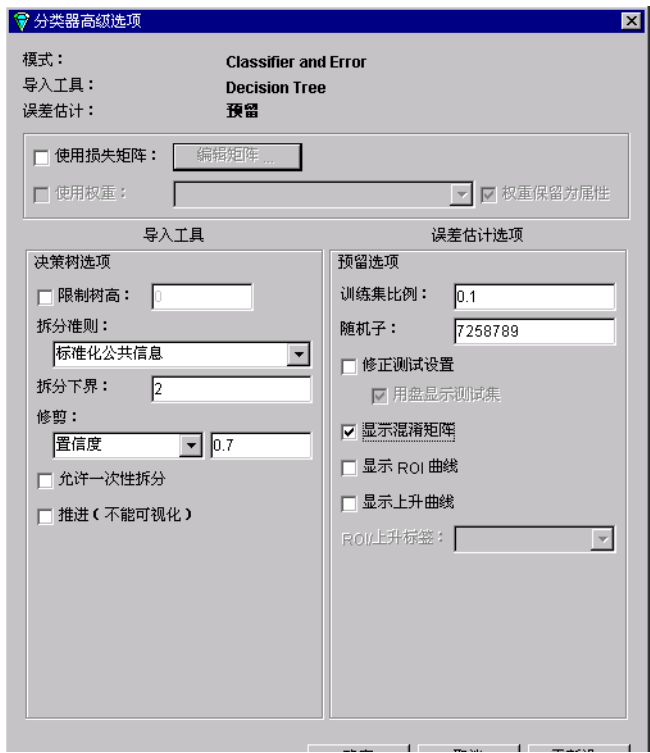


图 11-4 分类器选项窗格显示选中的混淆矩阵

3. 在“误差估计选项”窗格中，将预留比更改为 0.1，关闭“修正”测试集，单击“显示混淆矩阵”复选框，然后单击“确定”。

减小训练集的大小，为演示之便使分类器产生错误。典型的训练集要大的多。

4. 在“工具管理器数据目标”窗格中单击继续。

当决策树可视化处理出现时，请关闭它并检查图 11-5 所示的混淆矩阵。混淆矩阵显示分类器在分类中出现错误的地方。从这里，可以根据对数据的了解建立一个“损失矩阵”，使对某些种类错误的容错性降低。

图 11-5 所示的画面表明使用这个被削减的训练集，有 8 个有毒蘑菇被错误分类为可食用的，占总数的 0.1%。在可视化处理中的选择对象，然后从“选项”菜单中选择“显示”，查看“选择列表”。

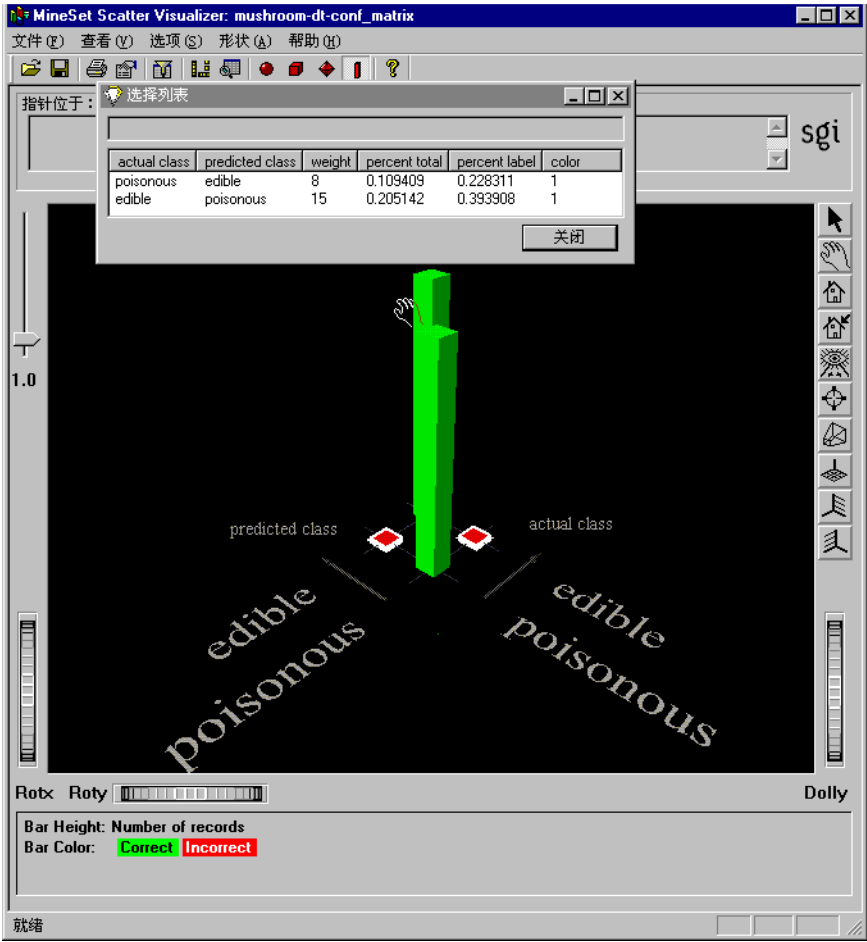


图 11-5 显示蘑菇数据集错误分类的混淆矩阵

定义损失矩阵

本部分讲解如何使用“损失矩阵”减少代价高的分类错误类型。示例将继续上一部分操作，使用蘑菇数据集、决策树导入工具、对可食性标签分类。

1. 用“文件”>“退出”来关闭“混淆矩阵”画面，返回到“工具管理器”窗口。
2. 在“数据目标”窗格的“挖掘工具”中，单击高级选项，打开“分类器”选项窗格。
3. 在窗格的第二部分，单击使用损失矩阵；然后稍等。
4. 单击编辑矩阵，加权发生错误的代价。“损失矩阵”窗口类似于图 11-6 所示。

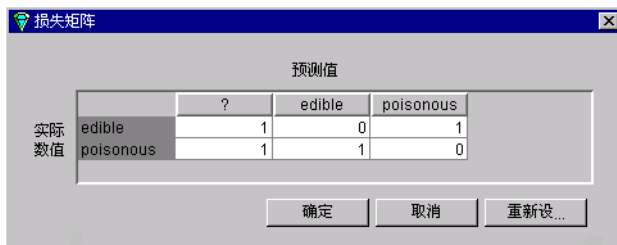


图 11-6 “损失矩阵编辑”窗格

5. 在“损失矩阵”的“实际值”行中设置这些值，如图 11-6 所示从左到右读取：
可食用的：1—0—1
有毒的：1—50—0
把有毒蘑菇错分为可食用蘑菇的代价被加权为 50 倍于反向错误分类的代价。
6. 单击确定以设置“损失矩阵”值，然后继续运行分类器。
新的分类器非常保守；不会发生将有毒蘑菇分类为可食用蘑菇的错误。相反，将可食用蘑菇分类为有毒蘑菇的错误却有 1558 个。新的估计损失为 $\$10,000 \times 0 + \$2 \times 1,558 = \$3,116$ ，仅为没有考虑损失的分类器成本的 3%。图 11-7 所示为产生的“混淆矩阵”。

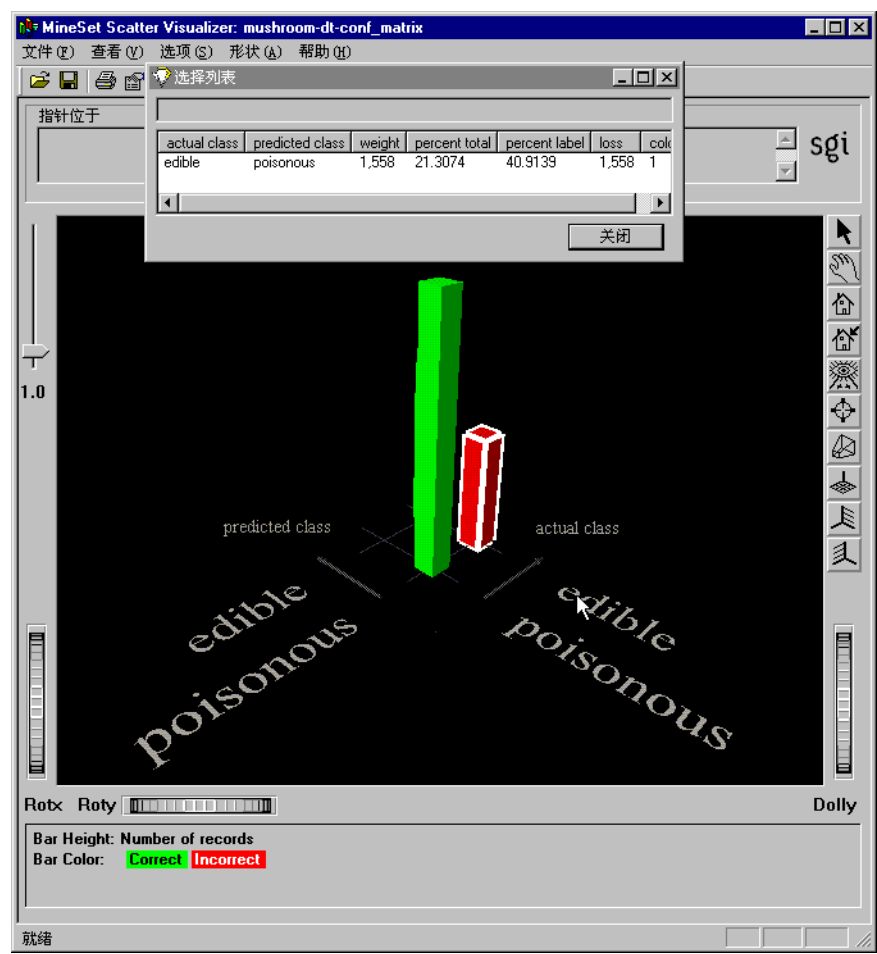


图 11-7 带有损失矩阵蘑菇数据集的混淆矩阵

关于“误差估计”的详细信息，请参阅 [《MineSet 3.0 企业版参考指南》](#)。

用上升曲线和 ROI 曲线评测模型

当商家从事一项长期项目时，能在下一阶段更改计划或放弃项目之前预测收益或损失将会大有益处。MineSet 可以判断问题的后果，并在事件产生显著影响之前进行更正。可以利用“上升曲线”和“投资回报曲线”（ROI）两种方法来进行这项工作。

用上升曲线检查预测

上升曲线显示预测特定标签值中随机记录顺序与模型生成的顺序之间的差别。例如：当创建了一个预测哪些用户可能波动的模型时，就可能希望在可能波动的用户客户波动之前瞄准它们。上升曲线能够帮助实现这个目标。本示例使用客户波动数据集。

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 显示的记录数对应于具有给定标签值的用户（在这里，Churn=yes）。

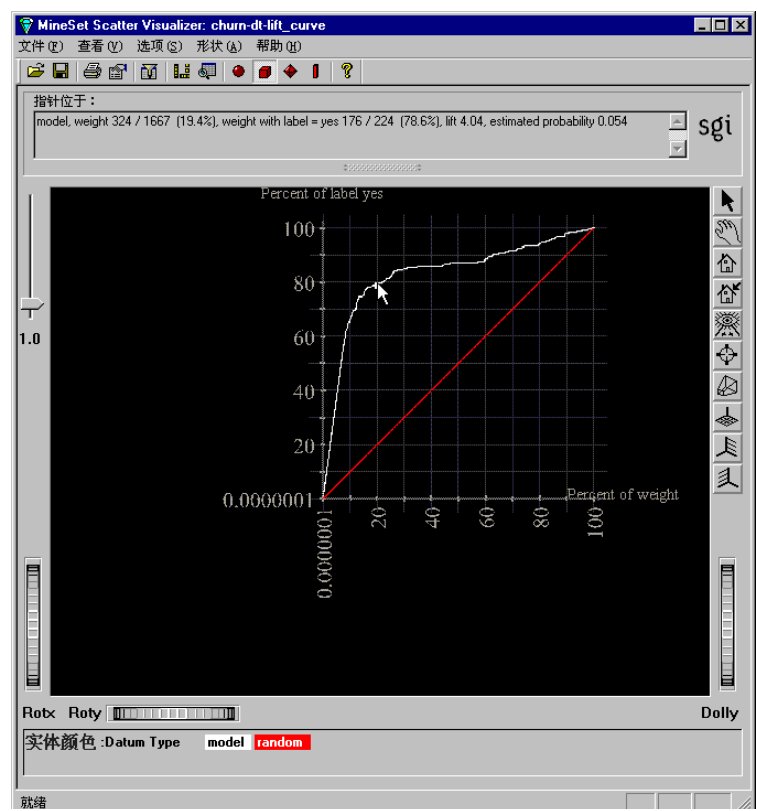


图 11-8 上升曲线

图中显示了两条曲线，如图 11-8 所示。下面的曲线（红色）显示在随机顺序下放置记录预期客户波动用户数曲线。上面的曲线（白色）显示按照模型对每个记录的评分（概率估计）顺序放置记录时客户波动用户数的百分比曲线。此时被模型识别为最可能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。模型排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。

要在选定模型上生成上升曲线：

1. 在“工具管理器数据目标”窗格中，单击“挖掘工具”选项卡。
2. 单击“分类”选项卡。
3. 选择“分类器和错误估计”模式以及“决策树”导入工具。
4. 单击高级选项，然后在“分类器”选项弹出菜单中，单击“误差估计选项”下的显示上升曲线复选框。在 ROI/ 上升标签弹出菜单中选择标签值（yes）。
5. 单击确定接受导入工具选项，然后关闭对话框。
6. 单击继续。

该过程可能需要花费一些时间。关闭“决策树”。产生的上升曲线将显示任务栏下选择窗口中所有选定点的详细信息。将指针沿白色（模型）线移动，然后单击各个点，查看 churn=yes 的用户百分比和升幅。查找曲线的拐点，在本例中，拐点在分类器估计概率为 0.056 的地方。在这一点，向可能波动的用户追加刺激能带来的投资回报将迅速减少。下一步是将分类器应用到整个数据集（请参阅第 7 章中的“使用所有数据进行分类”）。

使用投资回报曲线寻找销售利润

“投资回报曲线”（ROI）类似于“上升曲线”，但是它显示损失而不是错误方面的准确性，它取决于如何对“损失矩阵”加权（请参阅第 181 页的“定义损失矩阵”）。如果模型对其预测结束很肯定，那么期望损失将较小，记录靠近于 ROI 曲线的左边。

ROI 曲线希望您对数据集中的每个单独记录的结果都进行操作。该操作与选定的标签值相关联。例如：在客户波动数据集中，与标签“yes”相关联的操作可能是向某个人发送销售材料。这可能阻止他波动，但是如果不加区别地发送材料，成本就会非常昂贵。如果模型是用于预测是否向某人发送邮件的，那么 ROI 曲线最高点显示的就是在测试集中大约能够节省的钱。

可以从“投资回报”曲线中看到出现某种错误的代价，并指出继续采取行动也不会再有效的那一点。要使用“分类器和错误”模式在客户波动数据集上生成 ROI 曲线，可以按照以下步骤进行：

1. 在“工具管理器数据目标”窗格中，单击“挖掘工具”选项卡。
2. 单击“分类”选项卡，然后单击高级选项按钮。
3. 选中“修正”测试集、“显示混淆矩阵”，然后在“分类器”选项窗格中单击“使用损失矩阵”。
4. 在上面窗格中单击编辑矩阵以加权发生错误的成本，如第 181 页的“定义损失矩阵”。在本例中，用以下方法填充行：

实际值：否： 1 0 2

实际值：是： 1 0 -10

单击确定以设置“损失矩阵”。

5. 将显示 ROI 曲线设置为“打开”。
6. 在 ROI/ 上升标签下拉菜单中选择“yes”（对于客户波动示例）。
7. 单击确定。
8. 在“工具管理器”窗口中单击继续。

将出现以下三个画面窗口：“决策树”、“混淆矩阵混淆矩阵”、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，因此加权为负数。在考虑损失的情况下，“混淆矩阵”显示权重为 4%。一边的错误增加了，但另一边的错误却减少了。使用“文件”>“退出”以关闭“混淆矩阵”和“决策树”画面，然后查看“ROI 曲线”窗口（图 11-9）。

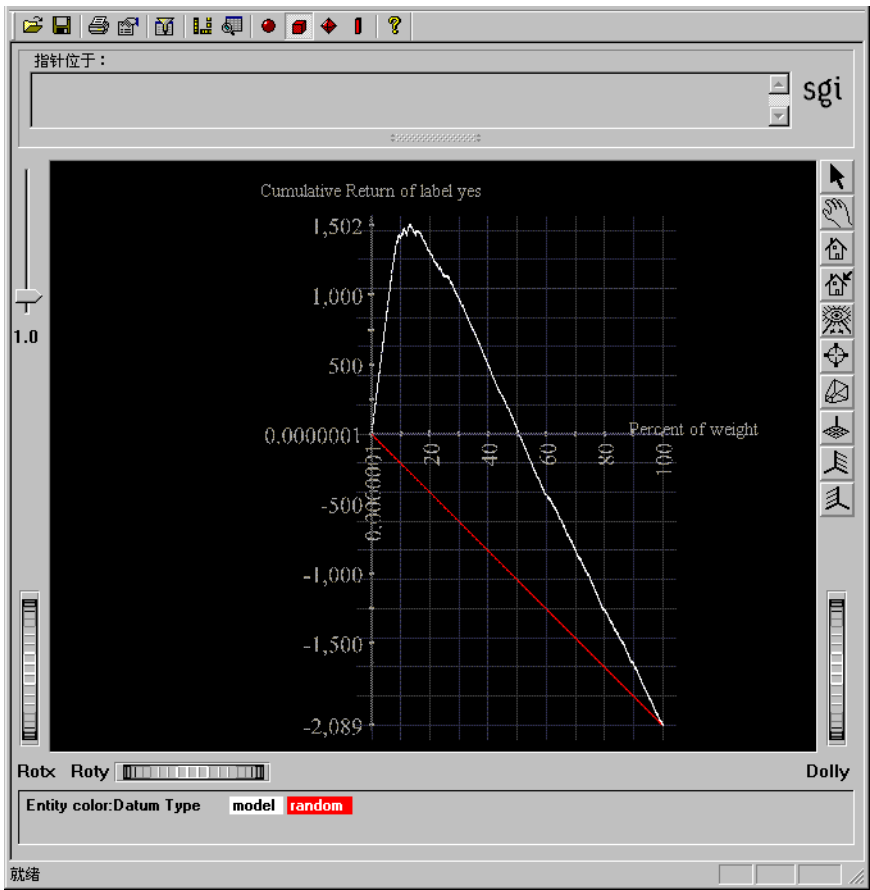


图 11-9 投资回报曲线

ROI 曲线类似上升曲线。中间的水平线代表零收益和零损失。红色线代表采用随机样本并向他们发送邮件时的期望效果。如果向每个人都发送邮件，将会有所损失，因为邮寄成本太高。但是，有一个最优的投资回报点，即曲线的顶点 1448，或总体的 15.2%。

在使用 ROI 曲线之前填充“损失矩阵”时，请务必小心。特定预测标签下的列将决定为该标签值生成的 ROI 曲线。如果要为使用标签 **churned** 的客户波动数据集填充“损失矩阵”，则该列中的条目需要表示出：在各类数据之上，与该标签值相关的操作能够带来的期望收益或损失。例如：客户波动中列“预测是”、行“实际值否”下的条目包含值 2 表明向某个不会波动的人邮寄小册子的成本（与“是”相关联的操作）是 2 美元。另一方面，列“是”、行“是”下的条目的值是 -10，表明阻止了用户的客户波动，除邮寄成本外为将公司节省 10 美元。这些成本难于估计，很小的变化就会显著影响生成的 ROI 曲线顶点位置。

进一步的详细信息，请参阅 *《MineSet 3.0 企业版参考指南》*，以及访问 MineSet 的网站 <http://mineset.sgi.com>，获得最近的更新。

用聚类划分数据

聚类的目的是为了确定数据集中哪些元素或特征是类似的。这对于研究新的数据集是很有用的。例如：蝴蝶花数据集显示出三个明显不同的蝴蝶花聚类 - 蓝色蝴蝶花，杂色蝴蝶花，和 纯白色蝴蝶花。本章包括以下主题：

- “聚类概述”
- “用 “工具管理器” 启动聚类”
- “使用聚类样例文件工作”
- “理解 “聚类可视化工具” 主窗口”
- “聚类的其它可视化处理”

聚类概述

聚类是描述性数据挖掘任务，所以不用将列指定为标签。因此，聚类被归为非监督学习运算法则。聚类模型通常是建立于整个数据集之上并在整个数据集上进行评测的。

MineSet 将类似的记录分在一起组成聚类，目的是为了使每个组中的相似性最大化。MineSet 提供两种不同的聚类模式：单步的 k 均值和迭代的 k- 均值。在 [《MineSet 3.0 企业版参考指南》](#) 的 “聚类” 中对两种模式进行深入讲解。

一旦聚类被运行，就可以直接使用 “聚类可视化工具” 查看聚类中心，它将显示数据集中独立的属性。可以检查最显著的属性，并且了解它们之间的区别。但是，要了解聚类之间属性的相互关系，只有 “散点可视化工具” 和 “决策表” 才能提供更清楚的视图。

要使用 “散点可视化工具” 查看聚类模型，需要确定应该使用 “列重要性” 工具将哪些列映射到不同的坐标轴。首先使用 “应用模型” 工具（请参阅[第 129 页的 “应用预测模型”](#)），然后从可用模型列表中选择（数据集名称）.cluster。这将提供新的 “聚类” 列。在 “散点可视化工具” 中将该列映射到颜色。

用“工具管理器”启动聚类

聚类没有先决条件，也没有必需的选项。默认情况下， MineSet 使用单步 k- 均值聚类方法挖掘数据中的三个聚类。聚类一旦完成，就可以看到聚类的评测，然后将出现“聚类可视化工具”。

- 1. 在“工具管理器数据目标”窗格中，单击“挖掘工具”，然后再单击“聚类”选项卡，如图 12-1 所示。



图 12-1 “聚类”选项卡

- 2. 要控制聚类方法的选择，可以从“方法”弹出菜单中选择单步或迭代 k- 均值，如表 12-1 所示。

表 12-1 聚类方法选项

方法	描述
单步 k- 均值	默认方法。指定寻找的群集数（默认值为 3）。
迭代 k- 均值	需要上界和下界以及选择点（0 到 1 之间的一个数—默认值为 0.5）。

关于选择点，请参阅词汇表定义。

- 3. 单击继续。

注意： 聚类计算是密集的操作，在处理较大的数据集时可能需要一定的时间，尤其是以迭代 k- 均值模式运行时。如果数据集中的记录数多于 10,000，请尝试聚类一个数据抽样。

群集被分配了从 1 开始的连续数值名称。虽然聚类名称用数字表示，但是对聚类来说，不存在次序。



图 12-2 使用迭代 k- 均值算法聚类

在迭代 k- 均值中，根据群集在拆分过程中的出处命名它们。使用连续的数字命名最初的聚类（基于聚类的最小数），就像在单步 k- 均值中一样。当拆分一个聚类时，两个新的聚类将按被拆分的聚类命名，附加一个“A”或“B”。例如，命名为“2-B-A”的聚类是在最初聚类中从聚类 2 中派生出来的，并且被拆分了两次。

关于聚类运算法则操作的详细信息，请参阅 *《MineSet 3.0 企业版参考指南》* 中的“聚类”。

使用聚类样例文件工作

MInceSet 提供的一个汽车样例数据集比较简单，处理的是一些熟悉的概念，如：马力、车辆重量和达到 60 英里 / 小时所需的时间。

使用 “工具管理器文件” > “打开新的数据文件” 下拉菜单，选择 *cars.schema* 文件。在 MinceSet 安装目录的 *\data* 目录下查找。

- 1. 在 “工具管理器数据目标” 面板中，单击 “挖掘工具” 选项卡。
- 2. 单击 “聚类” 选项卡，将 “方法” 设置为单步 k- 均值，群集数设置为 3，然后单击右下方的继续。



图 12-3 聚类可视化工具主窗口

3. 在产生的“聚类可视化工具”窗口（图 12-3）中，单击显示聚类 1 的绿色条，它控制棒状图和直方图所代表属性的优先次序。当在群集 1 上单击时，属性重要性（即：直方图出现的次序，从上到下读）为汽缸、体积、马力、重量磅，然后是英里 / 加仑。这个重要性次序仅由可视化处理的这个聚类控制。
4. 在其他的聚类中比较同一列，以了解聚类之间属性的区别。
5. 选择聚类 2，并且注意属性次序的不同。在本例中，产地更为重要，其次是商标、体积、汽缸、重量磅，再其次是马力，最后是英里 / 加仑。

理解“聚类可视化工具”主窗口

“聚类可视化工具”显示按行列安排的棒状图。每一行对应数据中的一个属性，每一列代表顶部窗格中命名的一个聚类。总体列显示的是对于整个数据集的统计。

按照第 190 页的“用“工具管理器”启动聚类”中的步骤，可以得到一个类似于图 12-3 所示的画面，其中聚类按次序编号。运算法则无法检测聚类代表的内容，因此需要人工翻译解释。属性在“聚类可视化工具”窗口中显示的次序是很重要的；默认情况下，属性是按用于聚类向彼此区分的重要性次序显示的。

改变列的次序：

1. 单击含有聚类名称的窗格，按区分该聚类与其他所有聚类的重要性次序重新显示属性。
2. 单击含有“总体”的窗格以恢复默认排序。

聚类的其它可视化处理

“聚类可视化工具”提供了一种易用的聚类基本可视化处理。但是，这不是可视化聚类的唯一方式。《*MineSet 3.0 企业版教程*》中介绍了一个使用“散点可视化工具”将聚类结果转换为画面的示例。主要想法是应用刚建立的聚类模型创建一个新列，然后使用选择的可视化工具。对于查看聚类结果，“散点”、“证据”和“决策表”可视化工具特别有用。

用关联规则分析数据

关联规则挖掘数据的一致性，即：哪些数据实例可能与其他指定实例同时发生。这就是市场供求分析，它来自于一个比喻：哪些产品可能在同一个购物筐中被一起购买。本章主要包括以下几个主题：

- 第 195 页的 “关联规则生成和可视化处理概述”
- 第 199 页的 “启动关联规则”
- 第 201 页的 “说明 “散点可视化工具” 中的关联规则”
- 第 204 页的 “追溯”
- 第 204 页的 “多路关联规则”

提供了一些配置和数据文件样例，用于演示 “关联规则” 的特性和功能。这些文件在 *MineSet 3.0* 下的 *\examples* 目录中，位于 *MineSet* 的初始安装目录。

关联规则生成和可视化处理概述

如果希望知道在超级市场的冷冻食品展示柜中，为什么将鸡蛋替代品与华夫饼干放在一起，那么以下您将看到在其中发挥作用的关联规则。关联规则可以发现数据中的关系和隐含模式—古典数据驱动式的知识发现过程中不是很明显的相互关系在此变得非常清晰，并且可以用可视化形式显示。关联规则对于发现大数据集中的隐含模式尤其有用，可以用关联规则表明数据集中某些项目同时出现的频率。这是一个非监督学习运算法则，因为它不是集中于任何一个特定的属性上。而是同等对待所有的属性，对于值得注意的规则进行全局搜索。

关联规则的发展有以下两个阶段：首先 “关联规则生成器” 将对数据文件进行处理，创建一个可用于可视化工具的文件。然后可视化工具显示出该文件，在本例中为 “散点可视化工具”。

关联规则生成

关联规则生成器可以生成简单的（一对一）和多路的规则。本部分描述简单关联规则。关于多路规则的描述，请参阅第 204 页的“多路关联规则”。

简单关联规则规定：如果 X 为真，那么在一定的概率下 Y 也为真。MineSet 称 X 为规则的左侧（LHS），称 Y 为规则的右侧（RHS）。

考虑顾客在一次进商店时所购买的一组物品。在这种情况下，规则可能是：“80% 买尿布的人同时会买婴儿奶粉。”整个百分比被称作规则的置信度。

在本例中，“尿布”是规则左侧（LHS）的项目，“婴儿奶粉”是规则右侧（RHS）的项目。在这种情况下：

- 如果婴儿奶粉出现在 RHS，那么 LHS 就可以帮助确定什么措施可以增加该项目的销售。
- 如果尿布出现在 LHS，那么 RHS 就可以帮助确定如果商店中停止出售尿布，什么商品将会受到影响。

“关联规则生成器”首先对输入文件进行处理，然后生成一个由规则组成的输出文件。如果 X 和 Y 为记录中的项目，那么如下的规则：

$X \Rightarrow Y$

表明：只要 X 出现在记录中，那么预计 Y 将会以一定频率出现。

关联的强度有四个数字来量化，汇总于表 13-1

表 13-1 关联规则组件		
量度	描述	统计描述
支持度	LHS 和 RHS 同时发生的频率	$P(LHS \cap RHS)$
置信度	在所有 LHS 发生的情况中，RHS 也发生的分数，或者支持度除以 LHS 项目发生的频率。	$P(RHS \mid LHS)$
期望置信度	RHS 项目发生的频率	$P(RHS)$
上升度	置信度与期望置信度的比值	$P(RHS \mid LHS) : P(RHS)$

请参阅以下词汇在词汇表中的定义：支持度，置信度，期望置信度，和上升度。关于详细信息，请参阅《MineSet 3.0 企业版参考指南》中的“关联规则”。

注意： 如果只知道 Y 和形如 $X \Rightarrow Y$ 的规则，那么对于 X 将无所知。规则指定蕴涵只能是从 LHS 到 RHS。

规则可视化处理

当关联规则显示在栅格图上时，可以看到关联规则的图形化形式。左侧 (LHS) 项目在一个坐标轴上，右侧 (RHS) 项目在另一个坐标轴上。如图 13-1 所示，规则的属性显示在其 LHS 和 RHS 项目的交叉点上。该画面可以包括条形图、盘和标签。

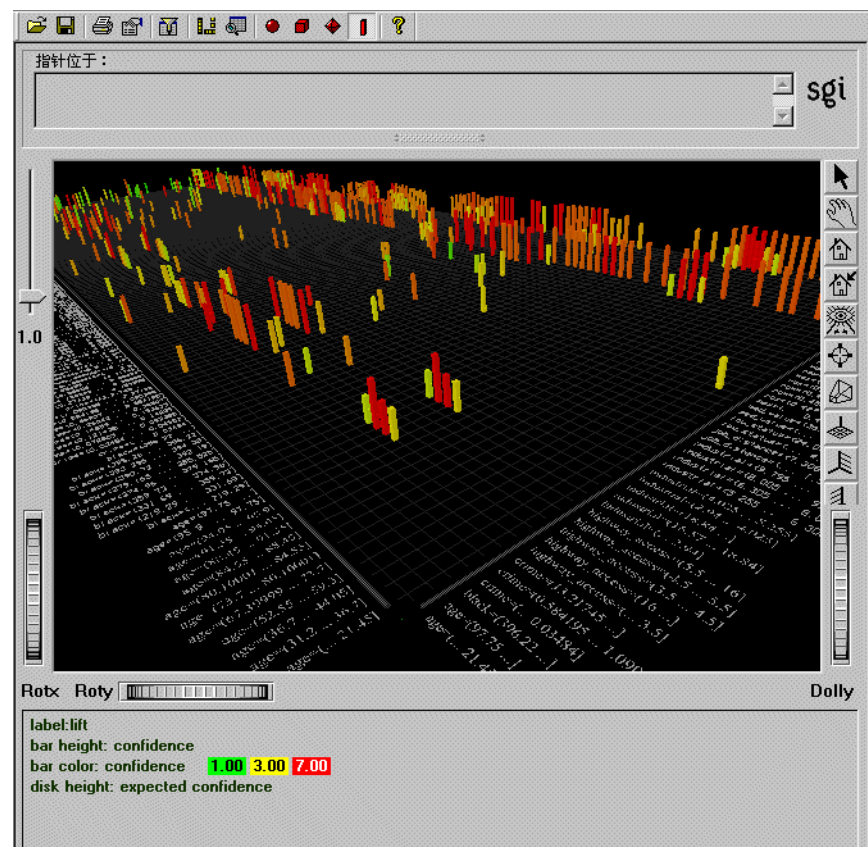


图 13-1 关联规则可视化处理主窗口的详细视图

主窗口的底部将显示一个图例，指示显示属性（例如条形高度和颜色）与基本规则关联值（例如置信度和支持度）之间的映射。

启动关联规则

本部分描述如何使用“工具管理器”简化配置关联规则工具的任务。

建立关联

本例中使用汽车数据集建立简单关联。您可能会发现以下属性之间的关联：英里 / 加仑、马力、加速度、重量、引擎大小、产地、品牌以及汽车的生产日期。例如：英里数是否随时间提高？引擎功率是否减小？

以下步骤是建立关联最简单的方法：

1. 打开“工具管理器”，选择服务器，并选择 *cars.schema* 作为数据源（请参阅第 15 页的“运行 MineSet”）。
2. 在“工具管理器”的“数据目标”面板中，单击“挖掘工具”选项卡；在下面一行选项卡中，选择 关联（对应于关联规则）。将显示图 13-2。



图 13-2 关联生成的初始工具管理器窗口

3. （可选步骤）关联规则的运算法则只适用于离散值；因此连续类型的列会被自动分组。如果希望其它的分组，可以使用“数据转换”窗格上的 *列分组* 按钮。（关于详细信息，请参阅 [第 37 页](#) 的“*为列改变或创建新的分组*”）。
4. 单击 *继续*，以运行关联规则生成器。或者选择以下设置：

置信度—指定规则的最小置信度阈值。置信度小于该值的规则将不会生成。默认值为 50%。可能的取值为 1-100。

支持度—指定作为记录总数百分比的最小支持度阈值。小于该值的关联规则将不会生成。默认值为 1%。可能的取值为 1-100。
5. （可选项）一旦完成关联规则选项的选择，可以单击 *规则映射* 按钮，将规则属性（如支持度、置信度、期望置信度以及上升度）映射到可视元素。请参阅下面的“*将规则属性映射到可视元素*”。

记录加权

当希望将某些记录指定为更重要的记录，或补偿不均衡的采样时，关联规则允许记录加权。如果没有选中 *按列加权* 框，那么每个记录的权重都是 1。只有当数据集中包含可以加权的列时，*按列加权* 框才会处于活动状态。

要启用记录加权，可以单击 *按列加权* 复选框。当选中该框时，会出现一个弹出菜单让您选择包含每个记录权重的列。如果选中了 *权重属性* 框，其中将包括“关联规则生成器”发现的规则中的权重列。如果没有选中该框，那么所有的记录将被赋予同样的权重，以确定“生成器”发现的规则。关于记录加权的进一步讲解，请参阅 [第 3 章](#) 中的“*加权记录*”。

将规则属性映射到可视元素

“关联规则”工具可以将规则属性映射到画面的可视元素。单击“规则映射”按钮，打开“将规则映射到可视化元素”面板，如 [图 13-3](#) 所示。

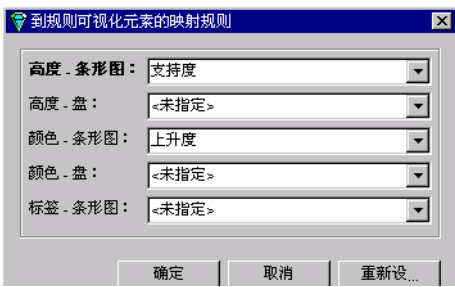


图 13-3 关联规则映射

可以被映射的可视元素在下面列出；前面有星号的项目是可选的：

表 13-2 将关联规则映射到可视元素

可视元素	映射
高度 - 条形图	指定条形高度代表的属性。默认值为 “支持度”
* 高度 - 盘	指定盘高度代表的属性。
* 颜色 - 条形图	指定条形颜色代表的属性。默认值为 “上升度”
* 颜色 - 盘	指定盘颜色代表的属性。
* 标签 - 条形图	指定条形标签代表的属性。

说明 “散点可视化工具” 中的关联规则

关联规则使用有效配置文件的说明显示 “散点可视化工具” 中规则文件的数据。要查看样例文件，可以在 “MineSet 三维可视化工具” 中使用 “文件” > “打开”，查看配置文件列表。这些文件在 *MineSet 3.0\examples* 目录中，位于 MineSet 的初始安装目录。例如：指定图片中 *brand.rules.scatterviz* 的结果，如图 13-4 所示。关联规则文件可以通过其名称结尾的 *.rules.scatterviz* 来识别。

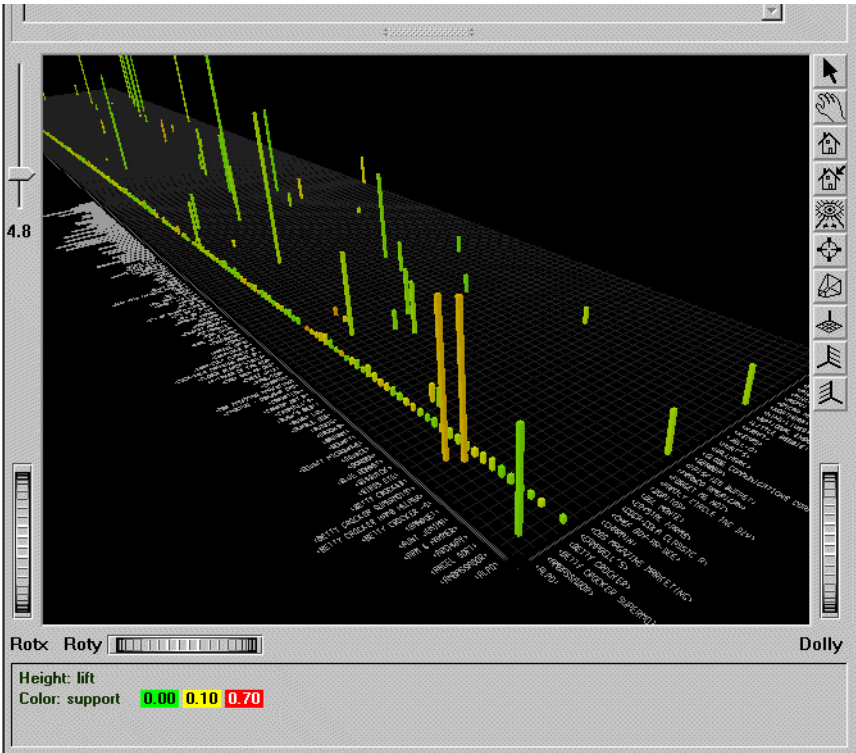


图 13-4 在指定 brand.rules.scatterviz 样例 时的初始关联规则视图

规则显示在“散点可视化工具”中的栅格图上。左侧 (LHS) 项目在一个坐标轴上，右侧 (RHS) 项目在另一个坐标轴上。如图 13-4 所示，规则的属性显示在其 LHS 和 RHS 项目的交叉点上。例如：条形高度对应于支持度，条形颜色对应于上升度。如果显示的视图太小，项目标签不会出现在坐标轴的两侧。可以使用窗口右下方的伸缩滑轮来放大视图，直到项目标签出现。当鼠标为选取模式时，还可以将鼠标指针放在单独的条形上查看特定规则的标签（请参阅图 13-5）。关于该特定规则所有详细信息都将显示在视图区域的左上角。

例如：在图 13-4 中，条形高度对应于支持度，条形颜色对应于上升度。

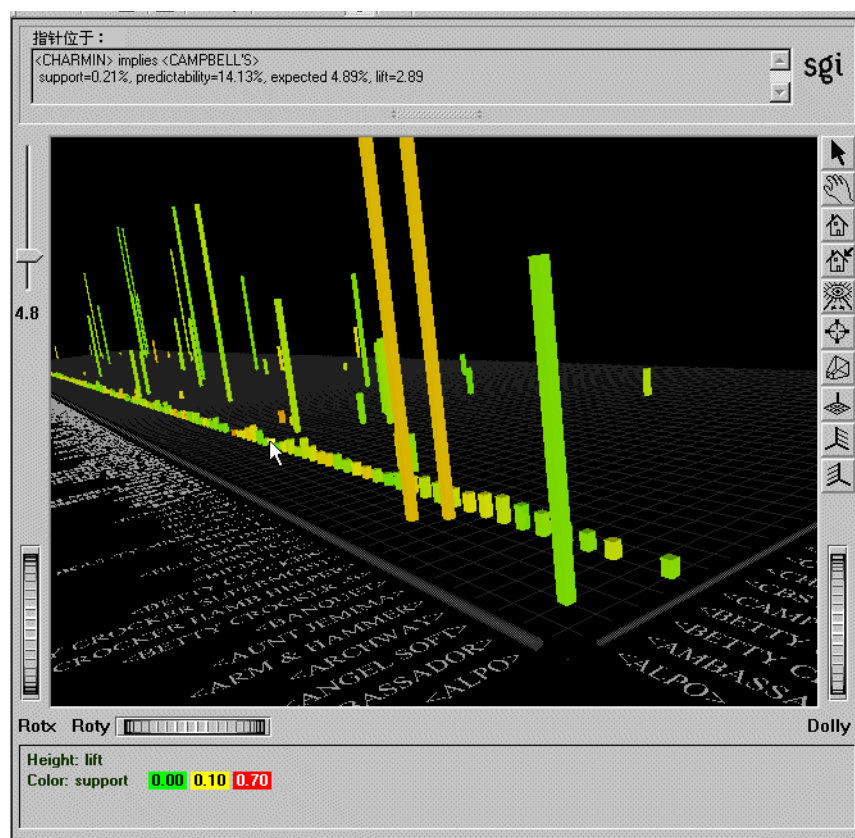


图 13-5 代表规则的条形图上的光标

如图 13-5 所示，将鼠标指针放在关联规则对象上，会显示出该对象的信息。只要指针停留在对象上面，信息就会显示，但是有一定的显示时间。如果将指针放在对象上面并单击鼠标左键，那么在主窗口上方也会显示同样的信息，该信息将一直保留到指针离开该对象。另外，可以按住 **ctrl** 键并单击选择多个规则。

使用鼠标可以将选择窗口中的文本剪切并粘贴到其他应用程序中，例如：报表或数据库。

追溯

通过对选定的规则执行逻辑 “与” 来确定追溯表达式。因为原始表中的列与 *.rules.data* 文件中的列不匹配，规则生成器将产生一个专门的列以帮助构建执行追溯时的筛选表达式。这意味着对追溯设置选项面板的更改将不会有效，因为已经将一个专门的字符串值列映射到 *.rules.scatterviz* 文件中用于追溯。

追溯规则时，MineSet 将显示满足规则的所有记录。

多路关联规则

在一些情况下，更加复杂的规则可能会很有用，这些规则的 LHS 和 / 或 RHS 上有多个项目。这就是多路关联规则。图 13-6 所示的是为多路规则进行配置的 “工具管理器关联” 面板。

如果选中了 “多路规则” 复选框，关联规则生成器将会生成所有满足最小支持度和置信度阈值的规则，包括那些在 LHS 和 RHS 中有多个项目的规则。这样的规则可能是：“啤酒和空心粉暗示着土豆片、甜面酱和葡萄酒。”

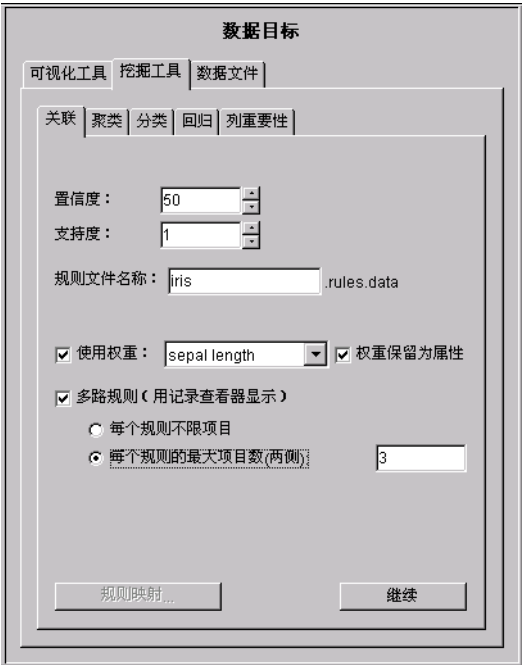


图 13-6 建立多路关联规则的初始工具管理器窗口

使用“记录查看器”可以显示多路规则，因为在“散点可视化工具”中没有明显的方法来显示更复杂的规则。它们被显示为一行一个规则。表中头两列包含 LHS 和 RHS 中的项目数。以下的四列包含支持度、置信度、期望置信度和上升度值。最后两列包含 LHS 和 RHS 项目。在 LHS 和 RHS 列中，项目由“和”分隔开。上面的规则示例中，LHS 包含两个项目“啤酒和空心粉”，RHS 包含三个项目“土豆片、甜面酱和葡萄酒。”

可以在“每个规则的最大项目总数”字段中输入一个数字，限制生成规则的大小。它指明规则中允许的项目数最大值。规则中的项目数是 LHS 和 RHS 中项目数的总和。上面的规则示例中有三个项目；简单规则有两个项目。

注意： 生成多路规则需要的时间较长。查看 “工具管理器” 状态窗口，其中指明每次迭代生成的规则数。如果生成的规则太多，可以取消操作，增加最小支持度或置信度阈值，或减少每个规则允许的最大项目数。

MineSet 用户词汇表

组合

在组合（如下）操作中数据以各种方式组合起来。

组合

在一些指定的列集合上，使用一个或多个操作汇总多行数据的过程。组合操作可以生成汇总统计，转置表，甚至可以生成基于原始数据的数组。

运算法则

完成某些任务的正式过程。例如：导入工具运算法则是一组机器指令，它告诉计算机如何建立分类器或回归器。

数组

能够存储多个项目的数据类型。数组类型由组合产生，可视化工具用它来存储随时间变化的动态数据。MineSet 支持一维数组（也称为向量）和二维数组（也称为矩阵）。

属性

预测模型进行输入的列。在执行预测建模时，数据中的列被划分为属性（提供输入）和标签（代表模型的输出）。

自动分组

这个运算法则可以选择分组阈值（请参阅分组），从而使不同分组中的标签分布尽可能不同。形式上，这个运算法则将使每个分组内的熵（请参阅熵）最小化。

修正

修正将整个数据集应用到由较小样本建立的模型。产生的模型保留了原始结构。修正的目的是为了在保持模型误差估计的同时，使模型的内部分布准确地反映所有数据。

分组

分组将连续的数据划分到离散的组中，将实际数据转换为分类数据。例如：一个连续的年龄范围可以被分组为：0-18、19-25、26-35 等等。这些组或分组被定义为不会重叠的连续区域。

块状图

在显示数据分布的可视化处理中展示的分段小块。类似于圆饼图，但是是方块。

选择点

选择点是介于 0 和 1 之间的值，它起着引导的作用，例如：引导聚类数的选择；较高的选择点意味着较大的数，而较低的选择点意味着较小的聚类数。选择点为 1.0 通常选取上界。在聚类中，如果界限是一个和五个类，那么值为 0.4 的选择点将选取两个类，而值为 0.8 的选择点将选取四个类。值为 1.0 的选择点将选取五个。

分类器

试图描述关于其他列（属性）的列（标签）的预测模型。从标签已知的数据中构建的分类器，将用于预测标签未知的新数据的标签值。从内部来看，分类器是为每个输入行预测离散值的运算法则或数学公式。例如：从蝴蝶花数据集中建立的分类器可以预测给定花瓣和雄蕊的长度与宽度的蝴蝶花类型。分类器还会估计每个标签值的概率。例如：从汽车数据集中建立的分类器可以预测指定汽车在美国制造的概率。分类器仅适用于离散标签。回归器类似于分类器，但用于连续标签。

聚类

聚类确定数据集中的哪些元素是相似的。它按照一个运算法则或数学公式对记录进行分组，这个运算法则或数学公式试图找到相似记录同时倾向的质心或中心。该过程将数据集划分为相互排斥的子集，不依赖于任何预定义的类。

条件概率

某个事件（A）在其他某个事件（B）发生的情况下发生的概率。写作 $P(A|B)$ ，读作“给定 B 时 A 的概率”。条件概率以矩形图显示在证据可视化工具左边的窗口中，这些矩形图显示了在给定（有条件的）每个标签值的情况下每个属性值的相对概率。条件概率可以被看成是给定标签值的证据。

具有连续值的属性

可以在整个连续值范围内取值的属性。在 MineSet 中，整型、双精度型和浮点型可以被看成是连续的。

置信度

关联规则（ $X \rightarrow Y$ ）的置信度，将 X 和 Y 同时出现的频率量化为 X 出现的记录数的分数。例如：如果置信度为 50%，在 X 出现的记录中，X 和 Y 同时出现的概率为 50%。因此，知道了 X 在记录中出现，那么 Y 也在该记录中出现的概率为 50%。

配置文件

识别数据如何从 .data 文件映射到满足可视化工具要求的文件。MineSet 配置文件的有以下几种后缀：.treeviz, .scatterviz, .splatviz, .mapviz, .eviviz, .dtableviz, .statviz, 和 .clusterviz

交叉验证

估计预测错误的方法。交叉验证将数据集拆分为 k（一般为 10）个同样大小、被称为数据夹的部分，建立 k 个预测模型，然后在经过其余数据夹的训练之后，在不同的数据夹上测试每个模型。这个过程可以被重复多次以提高估计的可靠性。

数据清洗

数据清洗是通过修改数据的形式或内容改进数据质量的过程，例如：删除或改正错误的数值。

数据文件

MineSet 的数据文件是用制表符分隔的原始数据单一结构文件，根据它来创建模型或可视化处理。MineSet 的数据文件使用的后缀是 .data。它们可能是二进制（更小并且更快）或 ASCII（可以人工读取）格式。

数据集

数据集是具有数据行和列的表。其中的行有时被称为记录或实例。MineSet CD 中包括多种样本数据集，例如：有关人们电信习惯的客户波动数据集，大众感兴趣的汽车数据集，基于美国人口调查局筛选后的资料的成人数据集。还可以将数据集描述为一个模式及符合该模式的实例集合。通常假设这些实例没有次序。

决策树

基于分支系列测试的预测模型。每个测试都要检查数据集中的单独一列，并用它确定下一个被应用的测试。所有测试的结果将确定要预测的标签。

离散属性

具有有限数目的不同值的列。在 MineSet 中，字符串、分组值和日期被认为是离散的。如果被用于分类器的标签，整数也被认为是离散的。

追溯

检索基于可视化处理的部分原始数据的操作。

熵

信息论中的一种量度，指的是数据的无序程度。数据集的熵越高，其中的值就越不同，越混杂。MineSet 中的许多挖掘运算法则都通过划分数据来使熵最小化。

错误率

模型对数据集所做预测的正确率或错误率。错误率通常由独立的测试集确定，在学习过程没有使用过该测试集。有时会使用更复杂的准确性估计技术，例如交叉验证，尤其是对于只包含少量实例的数据集。

证据

概率估计的准确程度将影响最终的预测。证据分类器根据每个属性的证据总和预测其标签值。在形式上，证据是正态化条件概率的负对数。

期望置信度

期望置信度是数据集中 RHS 项目出现的频率。因此，期望置信度和置信度之间的区别在于：因为 LHS 项目的存在，预测能量中量度的改变。期望置信度指出项目之间不存在任何关系时的置信度。

筛选工具

在 MineSet 中，筛选工具通常是一个布尔表达式，用来选择或操作列内容。例如：“m.p.g. < 25”表示只选择显示小于 25 英里 / 加仑的项目（即：对于该表达式为真的记录）。筛选工具是存储起来的经过选择的准则集合，这些准则指定数据集中的记录子集。

预留

用于训练模型的数据集的比例（通常为三分之二），其余部分将用于测试。预留方法是 MineSet 支持的最简单的误差估计方法。

假设检验

假设是有效性可以被检验的假定解释。假设检验是一种自顶向下的方法，它试图证实或证伪事先的想法。在建立假设前，要为注意到的行为提出可能的解释。经过扩展，假设将规定被选择来进行分析的数据。

导入工具

从训练集中建立模型运算法则。MineSet 支持四种分类导入工具（决策树、选项树、证据、决策表）以及一种回归导入工具（回归树）。

实例

用于记录或行的术语。

知识发现

知识发现是寻找数据中新的、值得注意并且有用的模式的过程。数据挖掘是知识发现的子集。它允许数据建议提出新假设并用于检验。

标签

预测模型试图预测的单独列。用户必须预先选择标签。例如：“蝴蝶花类型”是蝴蝶花数据集中一个很好的候选标签。

上升度

置信度与关联规则生成器所创建规则的期望置信度的比值。通常，这个数字越大，规则就越值得注意。

上升曲线

评价模型预测准确性的方法。上升曲线可视地显示出分类器随机提供确定标签的的优越之处。

平均值

平均值可被认为是样本中数据值的总和除以数据分组数所得到的数。平均值通常由 $\bar{x}$ 表示，一般被称为“平均数。”（例如：1、2、3 的平均值 $(1+2+3) / 3$ 为 2。）

中值

当数字以大小顺序排列时中间的数。

模型

基于观察到的现象的描述，通常省略了某些细节。模型暗含着预测。例如：如果形成了直接邮件模型，就意味着“这就是我们认为的直接邮件用户的样子。”学习者可以更改值并观察更改对系统操作的影响。请参阅“分类器”。

可预测性

“置信度”的另一个术语。

先验概率

类标签的先验概率是在忽略所有属性值的情况下，在随机选择记录的数据中发现该标签的概率。从数学的角度来说，这个数是带有类标签的记录数除以记录总数得到的。（请参阅“条件概率”。

纯度

“列重要性”相关内容中的纯度是标签值分布正确性的量度。累积纯度是数据关于标签值被划分的好坏程度的量度。使用列划分的数据在“决策树”中以同样方式进行划分时被认为是重要的数据。分组中的每个集合都有其自己的纯度量度，并且分组内的纯度量度是这些单独量度的组合。对于分组中给定的集合，如果每个类都有相同的表示法，那么纯度为 0，如果每个记录都是同一类的，那么纯度为 100。类似的，如果分组中每个集合的类表示法相同，那么累积纯度为 0；如果分组中的每个集合包含的记录都有相同的类，那么累积纯度为 100。在 MineSet 中，纯度取决于“熵”。

随机子

随机种子是为选择随机样本数据而选择特定起始点的方法。当需要不同的随机样本时，可以指定另外的随机子。对不同的数据集数据挖掘过程使用同样的随机子可以使您每次都使用同样的随机样本工作。当需要测试研究中发现的特定模式的稳定性时，可以改变随机子。

范围

列中最大和最小可能值的差。

回归器

一种预测模型，其中标签的取值为连续值。回归器与分类器很相似。

关系型数据库

关系型数据库是数据仓库的核心。数据及其之间的关系被组织到表中—每个项目中都包含相同字段的记录集合。一些字段被指定为关键字，以便按照指定值关键字段索引的搜索可以快速检索数据。如果在某个字段中具有相同的值，不同表中的记录可能被链接在一起。具体例子如：Sybase、Informix、OLEDB、SQL server 和 Oracle。

规则

规则被规定为用于处理数据的行为模式，例如：关联规则或分类规则。它也是决策树中从根到叶分类数据的唯一路径。基于规则的系统按照给定的程序，将一组“if-then”规则应用到一组事实上以进行推断。

投资回报（ROI）

一个金融术语，它通过测量投资的利润（回报）增值来衡量项目的价值。也被称为 ROI。

行

关系表中的记录。

模式

数据集中所有列的描述。模式包括每一列的名称和类型信息。模式可以被存储在后缀为 .schema 的文件中。

偏度

频率分布的不对称程度。

标准偏差

数据的分散量度。被定义为方差的平方根。

支持度

如果给定了关联规则 $X \rightarrow Y$ （ X 蕴涵 Y ），支持将把 X 与 Y 同时在文件中出现的频率量化为全部记录数的分数。例如，如果支持为 1%，那么 X 和 Y 将在全部记录数的 1% 中同时出现。

表

标准化的关系数据集。

测试集

测试集由数据集中被保留用于测试导入的分类器错误率的记录组成。请参阅“训练集”。

训练集

MineSet 中的训练集是数据的子集，在主要数据操作之前被划分出来用于建立分类器或模型。它由数据库中确定标签的记录组成，派生于描述性属性。导入工具用它来学习如何构造分类器。请参阅“测试集”。

修剪因子

修剪因子在分组前指定被排除在取值范围外的数据集极值分数。默认的裁剪因子为 0.05。它排除了 5% 具有极值的实例（2.5% 的最小值和 2.5% 的最大值）。裁减会减小外层对阈值生成的影响。

统一范围

统一范围是数据自动分组中使用的选择，在这些数据中取值范围被分为统一大小的子区间。

统一权重

统一权重是数据自动分组中使用的选择，在这些数据中取值范围被分为指定的分组数，以便每个分组中的记录数目相同。如果启用了记录加权，将对范围进行划分以便每个分组包含的总权重相同。

权重

MineSet 支持两种权重类型：

记录权重是应用到数据中每一行的数值。对权重为 k 的行，挖掘工具将当作似乎存在 k 个这样的行来处理。也支持非整型权重。权重值由数据中用户选择的列提供。

属性权重是用户指定的列重要性量度。聚类运算法则使用属性权重影响距离量度。

A

API, 15

按钮, 工具管理器, 34

B

“标号”按钮, 95

“标号”面板, 94-97

 获取当前位置, 96

“标号”命令, 94

帮助, 31

棒状图, 理解, 22

保存工具选项

 地图可视化工具, 107

保存数据

 记录查看器, 22

历史表, 47-48

C

“常量”命令, 79, 80

采样, 46-47

菜单

 分散可视化工具, 75-79

 平伸可视化工具, 75-80

参数

 显示选项, 97

查看历史, 47-48

查看模式

 地图可视化工具, 109

 分散可视化工具, 64

查找指定值, 93

查找重要的列, 50-52

纯度, 141

纯度的量度, 141

存储

 数据位置, 94

错误率（选项树）, 143

错误 / 损失估计, 140

D

decision trees(决策树)

 error/loss estimates(错误 / 损失 估计), 140

 measure of purity(计算纯度), 141

“地图可视化工具选项”对话框, 107

地图可视化工具

 查看模式, 109

 动画控制面板, 110

 空值与, 110

 显示数据, 101, 102

 选项

 保存, 107

 选择对象, 111

 主窗口, 108-110

迭代 k- 均值, 190, 191

动画, 100

动画控制面板（地图可视化工具）, 110

- 动画控制面板（平伸可视化工具）
 - 汇总窗口, 73
- 动画控制面板（散点可视化工具）
 - 汇总窗口, 69
- 端点, 102
- 对象
 - 搜索, 93
 - 选择
 - 地图可视化工具, 111
 - 平伸可视化工具, 66
 - 树可视化工具, 90, 91, 93
- 多个对象, 选择, 91
- 多路规则, 204
 - “工具管理器”及其, 205
 - 显示, 205

E

- error/loss estimate(错误 / 损失 估计), 140
- 二维组合, 69, 73

F

- F1 帮助, 31
- “发送到工具管理器”命令, 77, 79
- “分类器和错误”模式, 124
- 非监督建模, 5-6
- 分层, 81
- 分类, 4
- 分类规则, 142
- 分类器
 - 损失矩阵, 181-182
 - 混淆矩阵, 176-180
 - 应用到记录, 175
- 分类器模式, 121
 - 分类器和错误, 124
 - 估计错误, 125

- 仅分类器, 122-123
- 学习曲线, 126-129
- 分散可视化工具
 - 查看模式, 64
 - 动画控制面板
 - 汇总窗口, 69
 - 概述, 53
 - 显示数据, 69
 - 选项, 67
 - 选择对象, 77, 78
 - 运动轨迹, 70
 - 演示, 71
- 分析
 - 关系, 53, 81
 - 模式和趋势, 56
- 分析数据挖掘, 2-6
- 分组
 - 平均的阈值, 39
 - 自定义, 39
- 分组列, 37
- 负值, 87

G

- Gaussian 命令, 80
- “估计错误”模式, 125
- “关闭”按钮
 - “筛选”对话框, 97
- “关联规则映射”面板, 201
- 根节点, 86
- 更改标号, 97
- 更改列类型, 43-45
- 工具管理器, 18
 - “数据转换”面板, 34
 - 按钮, 34
 - 采集数据, 46-47
 - 多路规则
 - 图表, 205

- 分组列, 37-40
- 更改列类型, 43-45
- 更改列名称, 43-45
- 工具选项
 - 分散可视化工具, 67
- 关联规则生成器窗口, 199
- 配置选项
 - 关联规则, 199-201
- 筛选数据, 42-43
- 添加列, 35-36
- 应用模型, 46
- 转换数据, 34-47
- 估计错误, 121-125
- 关联, 5
- 关联规则, 195-206
 - 多路, 204
 - 显示, 205
 - 记录加权, 200
 - 建立关联, 199-200
 - 将数据映射到, 200
- 配置
 - “工具管理器”及其, 199-201
- 期望置信度, 197
- 上升, 197
- 生成, 196
 - 设置选项, 199-200
- 市场供求, 196
- 显示, 197, 201-203
 - 图表, 202, 203
- 支持, 197, 200
 - 最小阈值, 200
- 置信度, 196, 197, 200
 - 期望的, 197
 - 最小阈值, 200
- 转换数据, 199-200
- 细节追溯, 204
- 关联规则可视化工具
 - 图表, 198
 - 主窗口, 198
- 关联规则生成器, 195-206

- “工具管理器”及其, 199
- 设置选项, 199-200
- 输出, 196
- 显示图例, 198
- 关系, 分析, 53, 81
- 规则可视化工具
 - 显示规则, 201-203

H

- “互补细节追溯”命令, 77, 79
- 滑动条
 - 创建, 68
- 回归, 5
- 回归树
 - 概述, 120-121
- 回归树导入工具
 - 概述, 138
 - 误差估计, 145
- 汇总窗口（平伸可视化工具）, 73
- 汇总窗口（散点可视化工具）, 69
- 汇总值, 87
- 获得帮助, 31
- 混淆矩阵, 176-180

J

- “结构”命令, 80
- “仅分类器”模式, 122-123
- 基准, 86, 87
 - 决策树和, 140
- 记录
 - 分类, 141-142
 - 分类器和, 175
- 记录查看器, 19-22
 - “筛选”面板, 21

- 保存数据, 22
- 使用列, 20
- 运行, 19
- 记录加权, 49
 - 关联规则和, 200
- 加工数据
 - 工具管理器, 33-47
 - 为什么要加工数据?, 33
- 加权记录, 49
- 监督建模, 3-5
 - 评测准确性, 5
- 简单 k- 均值, 190
- 剪切选择信息, 66
- 建立关联规则, 199-200
- 建立模型, 10
- 建模
 - 非监督的, 5-6
 - 概述, 113
 - 监督的, 3-5
- 节点, 86
 - 决策树
 - 查看信息, 140
 - 选项树, 143
- 景观, 81
- 聚光灯, 91, 93
- 聚类, 6, 189-193
 - 方法, 190
 - 可视化工具, 193
 - 运行, 190
- 聚类可视化工具, 189-193
- 决策表
 - 概述, 118-119
- 决策树
 - 纯度的量度, 141
 - 错误 / 损失估计, 140
 - 分类记录, 141-142
 - 概述, 116-117
 - 节点

- 查看信息, 140
- 显示, 143
- 决策树导入工具
 - 查看节点信息, 140
 - 分类规则, 142
- 均方误差, 145

K

- k- 均值
 - 迭代, 190
 - 简单, 190
- 可视化工具
 - 漫游在, 27-31
- 可视数据挖掘, 2, 6

L

- 类
 - 指定, 140
- 类型列, 45
- 联机帮助, 31
- 列
 - 分组, 37
 - 更改类型, 43
 - 更改名称, 45
 - 类型, 定义, 45
 - 添加, 35
 - 重要性, 50-52
- 录入约定, xviii
- 轮廓, 98
- 轮廓文件字段, 107

M

- measure of purity(计算纯度), 141

.marks 文件名扩展名, 97

MineSet

运行, 15

MineSet API, 15

MineSet 工具, 7-8

命名

观察点, 94

蘑菇分类数据集

... 的混淆矩阵, 177, 182

模式, 56

模型

建立, 10

配置, 11

评测, 5, 11, 121

评测准确性, 5, 11

生成, 114

选择, 129

应用, 11, 129-131

O

OLAP 和数据挖掘, 2

P

purity(纯度), 141

“平伸类型”菜单, 79, 80

配置关联规则

“工具管理器”及其, 199-201

配置模型, 11

平面, 101

平伸

绘制选项, 80

已定义, 55

平伸可视化工具, 53-80

菜单, 75-80

动画控制面板

汇总窗口, 73

分析数据, 56

概述, 55

组合数据点, 55

获取信息, 66

显示数据, 73, 79, 80

选择对象, 66

主窗口, 64-67

评测模型, 121

评价模型, 5, 11

Q

期望置信度, 197

趋势, 56

全局观察窗口（树可视化工具）, 92

R

ROI 曲线, 185

S

Subtree weight option(子树权重), 145

“筛选”按钮, 76

“筛选”面板

分散可视化工具, 76

记录查看器, 21

“删除”按钮, 96

“视图”按钮

地图可视化工具, 110

“数据目标”面板

聚类, 191

“数据转换”面板, 34

损失矩阵, 181-182

三维组合, 69, 73

三维景观, 81

- 三维图, 61, 62
 - 散点可视化工具, 53-80
 - 菜单, 75-79
 - 主窗口, 64-65
 - 散点图, 56
 - 筛选, 42-43
 - 筛选数据
 - 记录查看器, 21
 - 删除标号, 96, 97
 - 上升, 197
 - 上升曲线, 183-185
 - 生成关联规则, 196
 - 生成模型, 114
 - 识别数据, 9
 - 使用学习曲线, 126-129
 - 市场供求分析, 196
 - 视图
 - 地图可视化工具, 100, 101, 102
 - 显示选项, 107
 - 树可视化工具
 - 突出显示信息, 91, 93
 - 书面文档
 - 录入约定, xviii
 - 树可视化工具
 - 标记观察点, 94-97
 - 概述, 81
 - 获取信息, 90, 91, 93
 - 决策树和, 143
 - 搜索对象, 93
 - 突出显示信息, 91, 93
 - 选择对象, 90, 91, 93
 - 主窗口, 82
 - 数据
 - 识别, 9
 - 准备, 9
 - 数据点
 - 组合, 55
 - 注释, 94
 - 数据挖掘, 2
 - 方法, 2
 - 分析, 2-6
 - 和 OLAP, 2
 - 可视的, 2, 6
 - 综述, 8-11
 - 数据集
 - 查找指定值, 93
 - 显示数据, 101, 102
 - 动画控制面板, 69-74
 - 三维景观, 81
 - 数据转换, 34-47
 - 数量, 81
 - 搜索, 93
 - 搜索聚光灯, 93
- T**
- 添加列, 35
 - 条形
 - 负值与, 87
 - 决策树和, 140
 - 统计可视化工具, 22-25
 - 棒状图, 理解, 22
 - 运行, 24
 - 直方图, 理解, 23
 - 投资回报曲线, 185
 - 突出显示的对象
 - 地图可视化工具, 111
 - 分散可视化工具, 77, 78
 - 平伸可视化工具, 66
 - 树可视化工具, 90
 - 图例
 - 关联规则, 198
 - 推进, 175
- W**
- 位置

- 标记, 94-97
- 文档
 - 录入约定, xviii
- 误差估计, 121-125
 - 回归树导入工具, 145
- 误差估计选项
 - 均方误差, 145
 - 平均绝对误差, 145

X

- “细节追溯列”命令, 78
- “显示选取拖动工具”命令, 79
- “显示原始数据”命令, 77, 79
- “显示值”命令, 77, 78
- “线性”命令, 80
- “向上”按钮, 97
- “向下”按钮, 97
- “修改”按钮, 97
- “选项树”导入工具
 - 错误率, 143
 - 决策树对, 137
 - 显示选项树, 143
- “学习曲线”模式, 126-129
- 显示
 - 决策树, 143
 - 决策树节点, 140
 - 数据, 81, 101, 102
 - 动画控制面板, 69-74
 - 选项树, 143
- 显示参数, 97
- 显示关联规则, 197, 201-203
- 显示选项
 - 地图可视化工具, 107
- 修改标号, 97
- 修正, 175
- 选取拖动工具, 66, 79

- 选项节点, 136
 - 排序, 143
 - 已定义, 136
- 选项树
 - 概述, 117-118
- 选项树导入工具
 - 概述, 134-137
- 选择菜单
 - 地图可视化工具, 111
 - 分散可视化工具, 77
- 选择对象
 - 地图可视化工具, 111
 - 分散可视化工具, 77, 78
 - 平伸可视化工具, 66
 - 树可视化工具, 90, 91, 93
 - 全局观察窗口, 93
- 选择模式
 - 平伸可视化工具, 66
- 选择模型, 129
- 学习曲线
 - 使用, 126-129

Y

- 颜色
 - 决策树, 140
- 叶节点, 140
- 应用程序编程接口, 15
- 应用模型, 11, 46, 129-131
- 映射
 - 关联规则和, 200
 - 字符串, 59
- 预测建模
 - 概述, 113
- 运动轨迹
 - 散点可视化工具和, 70
 - 演示, 71
- 运行

- MineSet, 15
- 记录查看器, 19
- 聚类, 190
- 统计可视化工具, 24
- 直方图可视化工具, 25
- 阈值
 - 分组, 平均的, 39
 - 自定义分组, 39
- Z
- “转到”按钮, 96
- 在可视化工具中漫游, 27-31
- 粘贴选择信息, 66
- 证据
 - 概述, ??-115
 - 概述 <\$startrange, 114
- 证据导入工具
 - 运行, 139, 149
- 支持, 197, 200
 - 最小阈值, 200
- 支持误差估计, 124
- 直方图, 理解, 23
- 直方图可视化工具, 25-26
 - 运行, 25
- 置信度, 196, 197, 200
 - 最小阈值, 200
- 主窗口
 - 地图可视化工具, 108-110
 - 分散可视化工具, 64-65
 - 回归树, 144-145
 - 平伸可视化工具, 64-67
 - 树可视化工具, 82
- 注释数据点, 94
- 转换数据
 - 工具管理器, 34-47
 - 为什么要转换数据?, 33
- 细节追溯
 - 关联规则和, 204
- 准备数据, 9
- 准确性
 - 评测, 5, 11
- 子代节点, 86
- 子节点
 - 选项树和, 143
- 字符串
 - 映射, 59
- 最小支持阈值, 200
- 最小置信度阈值, 200
- 组合, 40-42, 87
 - 二维, 69, 73
 - 数据点, 55