MineSet™ 3.0 企业版
参考指南

文档号码：007-3558-001CHS

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, Clif f B runk, Eben Haber, Ara Jerahian, Andy Kar, Ed Karrels, Eser Kandogan, Alex Kozlov, Alan Norton, Peter Rathmann, Mario Schkolnick, Dan Sommerfield, Pete rWelch 和 Brett Zane-Ulman。

MineSet™ 3.0 企业版 参考指南
文档号码：007-3558-001CHS

目录

图表清单

表格清单

关于本指南

*《MineSet 3.0 企业版参考指南》*用于说明 MineSet 系统数据挖掘和可视化工具的技术特征和高级功能。 MineSet 产品的当前信息也可以在下列网址中找到 http://www.sgi.com/software/mineset。

关于读者

该指南要求读者已经从 “工具管理器”中熟悉 MineSet 的操作。该指南的目的是向读者阐述背景细节并说明 MineSet 是如何工作的。Windows 用户将会发现文档中给定的方法和路径名是熟悉的。对于 IRIX 的用户应熟悉 UNIX 命令。

寻找信息

该指南的大部分内容解决的是从命令行和配置文件中安装和运行 MineSet 的有关问题。在有些样例中对可编程接口进行了描述。每章的小结可在 “文档的结构”中找到。

关于如何使用 MineSet 工具的信息，参考 *《MineSet for Windows 3.0 企业版用户指南》*。

文档的结构

第 1 章，概念和解释
本章包括 MineSet 工具和概念的描述和解释。

附录 A，配置和数据文件样例
本章描述了 MineSet 所带的数据和配置文件样例。

指南中的插图

该指南的大多数插图来自于 MineSet 3.0 for Windows。在一些样例中， IRIX 和 Windows 的操作有本质的不同，本指南对两种版本都进行了说明。

录入约定

指南中将使用下列类型的约定和符号：

斜体 可执行名称、文件名、程序变量、工具、实用工具、可变命令行参数、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
屏幕上的命令行文字和提示

黑体固定宽度类型
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

概念和解释

添加列

如果在将数据集中的数据列送入分类工具或可视化工具之前，想对它进行添加、删除或排序操作，则可在 “工具管理器”的 “数据转换”面板中实现。下面的 “添加列”按钮条目描述了添加列的过程。“删除列”以及 “按列名排序”将在各自的条目中讨论。

添加列按钮

您可以用*添加列*按钮创建一个新列，它的值来自于数学表达式的计算结果。例如，您可以添加一个新列，其值为两个已存在列的比值。单击*添加列*可得到一个对话框，让您指定新列的名字和表达式 （图 1-1）。

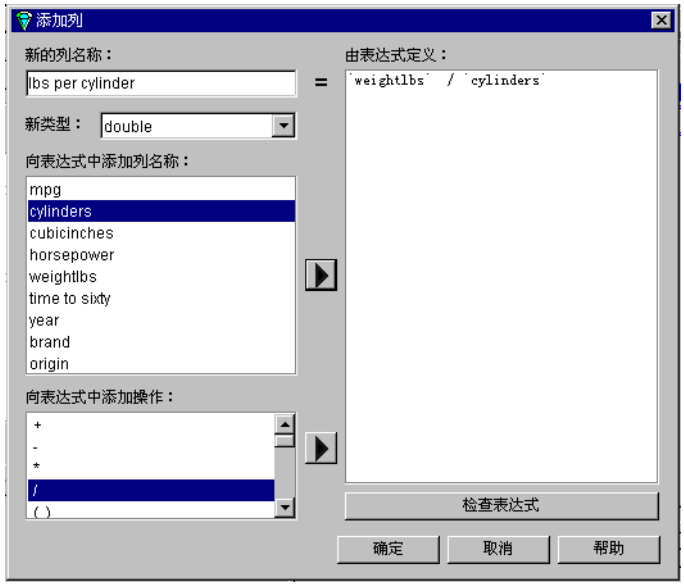

图 1-1 添加列对话框

对话框左边有一个字段可输入新列的名字，弹出式菜单可指定列的类型 （整型、字符型、浮点型等）。

对话框的右边是大的文本输入区域，可用来输入定义表达式。作为输入列名和操作符的快捷方式，对话框左下方的滚动列表框中列出了当前表中的所有列和所有可能的操作符。要在表达式中插入一个列名或操作符，可在滚动列表框中双击它，或先选中它然后单击滚动列表框右边的箭头按钮。

在 “筛选”和 “添加列”面板中使用的表达式语言与 C、 C++ 和 Java 中使用的类似。基本操作符是一样的 :

+	加
-	减
*****	乘
/	除
( )	括号
%	求余 （除法操作后所剩的余数）
!	逻辑非
~	逻辑非
&&	逻辑与
||	逻辑或
^	逻辑异或
==	等于
!=	不等于
<=	小于等于
<	小于
>=	大于等于
>	大于
&	按位与
|	按位或

表达式语言同样提供下列操作：

isNull( )	判断括号中的值是否为空。
if( )then( )else( )	如果第一对括号中的值为真，那么该表达式返回值就为第二对括号中的值。否则，返回值就是最后一对括号中的值。
(x) : (y) : (z)	相当于 C 的语法 if/then/else
divide(x,y,z)	返回 x 除以 y 的商，如果 y 等于 0，则返回 0。

strlen(x)	表达式返回值等于字符串中字符的个数。
substring(x , yz)	返回值是字符串 x 的子串，该子串始于 y，长度为 z。遵从 C、C++ 和 Java 中的约定，字符串中第一个字符的索引号为 0。

要检查您所创建的表达式，单击*检查表达式*按钮。如果存在错误，就会出现一个对话框，指出是什么错误及其位置。直接单击*确定*，该表达式就会自动被检查，并且错误信息直到表达式被修改正确后才会消失。

“添加列”对话框可以检查类型相容性：如果您赋给一个字符型列以数值表达式，就会出现警告消息，并且新列的类型会被自动改正 （反之亦然）。

您还可以添加自己的用户定义函数。参考 *《MineSet 3.0 企业版接口指南》*中的 “内置函数”。

组合

使用 “工具管理器”中的*组合* 按钮需要对数组及分布有基本的了解，参见 “数组”条目，该条目提供了在组合特征中所涉及概念的基本介绍。

您可利用 “工具管理器”中的 *组合*按钮创建简单的组合过程，产生数组或分布列。单击该按钮可产生 “组合”对话框 （图 1-2）。对话框中有三个列表，在中间列表中显示当前表中的列。

要进行组合操作，请选择列的名字，然后单击左边和中间列表之间的左向箭头按钮。您可利用下面的弹出式菜单指定索引值 （如果结果是数组）或一个分布列 （如果结果是被分布开的）。

要进行组合操作时，您可以利用底部的五个选项：求和、平均、最小最大值或计数。如果要对数值型列进行组合，您可以选择这些选项的任意搭配。对于其它类型，只允许计数。如果您选择了不止一个的选项，就会得到多个结果。例如，选择了平均和最大值运算，您将会得到一个平均值结果和另一个最大值结果。复选框 “在组合中包括空值”允许您在计算中忽略或包含空值。

图 1-2 “组合”对话框

下面给出列名的三个列表 :

- *要组合的列*。
- *分组索引列*（默认）； 这样可以保证在整个操作过程中该列表中的列不发生变化。对于分组索引列中具有相同组合值的每套记录，在结果表中只能有一个输出结果，其值出现在求和、最小值、最大值或计数组合列中 （取决于面板底部的复选框）。
- *要删除的列*。

在您处理完附加组合准则对话框以后，“表处理”窗口的 “当前列”文本框内显示了应用这些准则之后所产生的新列的名字。

对话框右下部的菜单用于创建分布列或数组列。您可利用这些索引和分布菜单指定一个索引数组或分布列。参见第 6 页 “数组”和第 7 页 “分布列”。

数组

数组是某一类型变量的集合，例如：浮点型、整型、字符型等 （参见第 37 页 “更改列类型或名称”中的可能类型的清单）。数组总有一个索引，该索引必须是已分组的列。数组可以是一维、二维、三维等等。一维数组可看以作为列表。二维数组可以看作为表单。数组维数越高就越难将其可视化。

数组对于 “树可视化工具”很有用处；如果您想定制在 “散点可视化工具”、“平伸可视化工具”以及 “地图可视化工具”中使用的滑动条，数组是必须的。

例如，假如您的数据集代表按照年龄组和州所划分的消费额。为了限制单元格数目，您可以将参与者的年龄分为三组：0-20、 21-40、41-60。结果表显示在表 1-1。

表 1-1 按照年龄组和州组合的消费模式

州	年龄组	总支出 （$）
CA	0-20	$50
CA	21-40	$454
CA	41-60	$693
NY	0-20	$35
NY	21-40	$541
NY	41-60	$628

表 1-1 显示了六行，按州排序。相同的含义也可以用一维数组列表示，正如表 1-2 所示。在这种表示过程中，*总支出 （$）*成为一个数组，并以分组列 *年龄组*为索引。

表 1-2 数组以 “年龄组”为索引

州	总支出 （$）[年龄组]
CA	[$50, $454, $693]
NY	[$35, $541, $628]

分布列

分布列与数组相似，但在几个重要方面存在不同。不产生包含许多值的一个新数组列，而按每一个索引值产生一个新列。例如，如果在表 1-2 中的数据没有成为一个数组，取而代之的是按照*年龄组*进行分布，结果将在表 1-3 中显示。

表 1-3 按照 “年龄组” 分布的列

州	总支出 （$）0-20	总支出 （$）21-40	总支出 （$）41-60
CA	$50	$454	$693
NY	$35	$541	$628

该分布的例子扩展了表 1-1，增加了列数而减少了行数。

如果您有超过一个的分组列 （例如，*年龄组*和*工资组*），您可以生成一个按照*年龄组*和*工资组*组合索引的二维数组。表 1-4 通过加入*工资组*列的区别特征来改进以前的样例（只显示了代表 加利福尼亚 的行）。通过这些您可以查看特定区域以及特定收入档次中的消费总数。利用这些技术可以在数据集中挖掘更精练的信息。

表 1-4 按照年龄组和工资组分布的列

州	年龄组	工资组	总支出 （$）
CA	0-20	$0-$25,000	$30
CA	0-20	$25,001-$50,000	$15
CA	0-20	超过 $50,000	$5
CA	21-40	$0-$25,000	$120
CA	21-40	$25,001-$50,000	$234
CA	21-40	超过 $50,000	$100
CA	41-60	$0-$25,000	$101
CA	41-60	$25,001-$50,000	$290
CA	41-60	超过 $50,000	$302

如果您将总支出 （$）变成了数组，在表 1-5 中显示了相同的数据产生的结果：

表 1-5 总支出 （$）中包括工资组时的结果

州	工资组	总支出 （$）[年龄组]
CA	$0-$25,000	[$30, $120, $101]
CA	$25,001-$50,000	[$15, $234, $290]
CA	超过 $50,000	[$5, $100, $302]

如果您想利用*年龄组*和*工资组*产生数组，结果将显示在表 1-6：

表 1-6 利用年龄组和工资组产生数组的结果

州	总支出 （$）[年龄组] [工资组]
CA	[$30, $120, $101, $15, $234, $290, $5, $100, $302]

最后，如果您利用*工资组*进行分布而以*年龄组*作为索引，结果显示在表 1-7：

表 1-7 利用*工资组*进行分布而以*年龄组*作为索引的结果

州	总支出 （$）[年龄组]，工资 $0-25,000	总支出 （$）[年龄组]，工资 $25,001-$50,000	总支出 （$）[年龄组]，工资超过 $50,000
CA	[$30, $120, $101]	[$15, $234, $290]	[$5, $100, $302]

对于每个数组成员上面的样例只有一个相关值，并且分布操作仅仅对存在的数据进行了重组织。 MineSet 为那些输出数组成员含有多个值的数据集提供了几个组合选项。最普遍的选项就是求和。例如，当累加消费进行预算的时候这是很有用的。您也可以计算整个数据的最小、最大和平均值以及对其计数。

当对一个数据集进行上述操作的时候，对于一个特殊的组很有可能没有合适的取值。在这种情况下，对于*最小*、*最大*、*平均*、和*求和*组合操作，数据移动器都以空值处理。对于 *计数* 操作， DataMove 的值为 0。

动画

只要被使用的数据集中至少有一个列被映射到滑动条成员，则 “散点可视化工具”、“平伸可视化工具”以及 “地图可视化工具”提供动画功能。这就意味着您可以利用动画显示一维以上的数据集的变化，例如：时间。典型情况下，独立的属性，例如，时间或年龄是最好的滑动条搭配，但是任何分组的列都可以被使用。

动画控制面板

主可视化工具窗口的右边的动画控制面板包含一个最多带有两个滑动条的汇总窗口、一个信息字段、动画按钮、路径滑动条、速度滑动条、数据点切换按钮、同步滑动条按钮和对于 “散点可视化工具”来说的管道轨迹菜单。

独立维控制滑动条

在汇总窗口附近出现的滑动条数目依赖于在可视化工具主窗口中所显示的数据集。数据集可以具有一、二维滑动条或者无滑动条。

具有两个独立维的数据集

如果数据集具有二维独立可变数据 （例如 *company.scatterviz*），则主窗口右边的控件将变为可见的 （参见图 1-3）。

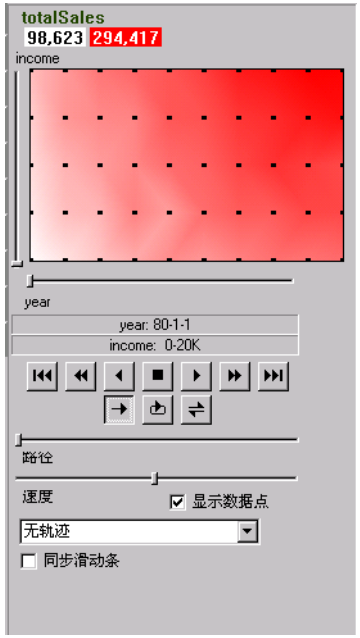

图 1-3 带有汇总窗口和两个滑动条控件的动画控制面板

在主窗口的右边是汇总窗口和滑动条控件。汇总窗口下面的水平滑动条用来选择第一独立维的数据点，而在其左面的垂直滑动条用来选择第二独立维的数据点。水平滑动条所代表的维由其下面的标签标识，垂直滑动条所代表的维由其上面的标签标识。

滑动条在工具管理器中由以下一种方法指定：

- 对于地图、散点或平伸可视化工具，将列名映射到“数据目标”窗格中的“滑动条 1”或“滑动条 2”元素，或者
- 利用“组合面板”创建一或两个数组列。这些数组列自动被应用到可视化工具的动画滑动条。

动画汇总窗口

汇总窗口显示了全部的列值，这些列被映射到动画滑动条所有可能设置下的汇总可视化元素。也就是说，它显示了汇总属性如何随滑动条变化。汇总窗口中越白的部分表示主窗口中实体所代表的汇总值越低。颜色越深，代表的值越高。

汇总窗口还显示了黑点，在一维或二维范围上均匀分布。这些点指示离散数据点的精确位置。您可利用 “显示数据点”复选框关闭这些黑色的点。

例如，当 *company.scatterviz* 文件被第一次打开时，二维汇总窗口显示从白 （在左边）到红 （在右边）的颜色范围。白色对应于低销售量；红色代表高组合销售量。在这个样例中，红色越深代表日常生活、汽车和家庭保险的销售总值越高。

*动画路径*是指汇总窗口中代表动画数据的黑点之间的路径。在汇总窗口中有三种方法创建动画路径。

打开文件，可采用这几种方法之一创建动画路径：

- 在汇总窗口中，挑选一个黑点并定义起始点。单击按住鼠标左键并在窗口中拖动光标。通过放开鼠标左键结束路径。
- 在汇总窗口中，通过单击鼠标左键定义起始点。然后，通过移动光标至窗口的其它位置后单击鼠标中键来定义终止点。在两点的中间出现一条直线。要加入更多的线段，继续重复单击鼠标中键。只有用三键鼠标才能完成这一操作。
- 在汇总窗口中，通过单击鼠标左键定义起始点。然后拖动独立维滑动条中的一个，沿着这维中画一条直线。如果有两个滑动条，利用第二个滑动条并在其控制下沿着坐标轴画一条直线。

动画按钮和滑动条

您可以利用在二维汇总窗口下面的类似 VCR 的按钮和滑动条 （路径和速度）控制动画。

动画按钮

一旦在汇总窗口中绘制了路径 （参见第 12 页 “动画汇总窗口”），您可以利用类似 VCR 按钮沿着路径控制动画。中间的*停止*按钮以蓝色加亮，指示初始状态。使用相邻的*播放*按钮 （在*停止*的右边）或*反向播放* （在左边）在所画路径中按向前或相反方向开始简单运动。（是*播放*还是*反向播放*，是由画路径的顺序定义的，而不是由自左向右或自右向左的运动定义的。）

要停止和再启动动画，单击*停止*按钮，然后再按*播放*或*反向播放*按钮。注意当您停止时，动画继续到最近的离散数据点。

在 *播放*按钮旁边是*单步*按钮，以及*单步播放*和*反向单步播放*按钮。单击以上按钮之一将改变当前路径点位置到下一个离散数据点。

最外边的按钮是*快进*和*快退*按钮。在*停止*状态时单击这些按钮之一将使路径点位置变到路径的终止点 （对于*快进*）或起始点 （对于*快退*）。在*播放*状态下单击*快进*按钮将加快动画速度。

动画流

在动画按钮的下面是三个动画流按钮。

单次播放（默认） — 动画向前或反向运动直到到达路径的终点，然后停止。

循环播放 — 当动画到达路径的终点后 , 它自动回到起始点然后重新开始。

往返播放 — 当动画到达路径的终点时，它反向沿着原路径向另一端点运动；到达端点后动画再次反向，循环也再次开始。

动画滑动条

通过分组和各种组合，可以自动或手动创建滑动条。该操作并没有在 “工具管理器” 的当前历史中显示，但它们确实出现在工具的配置文件中。

映射为 “滑动条 1” 和 “滑动条 2” 的列最终形成在动画中用到的列的索引值 （例如，颜色和大小）。这些列或者是数字型的 （整型、浮点型、双精度型），或者被分组。如果映射为滑动条的列已经被分组了，该列就不需要再自动分组了，并且该列已被当作滑动条的索引使用。然而，如果该列未被分组，利用自动分组选项就可以创建一个分组的列。（参见第 34 页 “分组”，以及个别工具条目可找到更多的信息。）

当动画停止时，您可以移动 “路径滑动条” 沿着路径重新设置位置点。注意，当您使用 “路径滑动条” 时，汇总窗口中的光标沿着画好的路径移动，而且绘制区域下面的和左边的滑动条将会随光标的位置相应地移动。然后利用*播放*或*反向播放* 按钮从新指定的点重新开始动画过程。您可以将 “路径滑动条” 拖到离散数据点之间的任意位置，然而当您放开滑动条时，路径位置就会移动到最近的离散数据点上。

使用 “速度” 滑动条调节动画沿着路径播放的速度。

数据点和内插

当动画播放时，在 “散点”、“平伸” 或地图 “可视化” 工具中映射成大小、颜色或坐标轴 （位置）的组合变量在平滑的变化。然而，在选中点消息框和以及在*指针位子*字段中所显示的信息只是位置最近的离散数据点的数据值，没有显示内插的数据值。

动画按照下列的方式产生：如果您有了 1991 和 1992 年的数据，并且它们对应于 “散点可视化” 工具中一个实体的大小。进一步假定 1991 年的大小为 20，1992 年的大小为 40，当您将年滑动条从 1991 移向 1992 时，实体大小随着 20 到 40 之间内插的结果而均匀变化。例如， 1991 和 1992 的中点，其大小为 30，当您靠近 1992 时，大小就接近 40。然而，您无法终止离散数据点之间的动画，而您也不能将路径滑动条拖到离散数据点之间的静态位置。

汇总窗口的数据点代表与数据文件中实际数据相对应的位置。例如，大小 20 和 40 代表实际数据，而 30 则不能。在这个样例中，在与每年相对应的滑动条位置上汇总窗口中都有数据点。

注意：并非所有的变量必须随滑动条的变化而变化。如果存在两个滑动条，一些变量会随着其中一个滑动条的变化而变化，而另外的变量将随两个滑动条的变化而变化。

运动轨迹

在 “散点可视化” 工具中，轨迹菜单可以显示动画过程中运动点的轨迹。您可以选择 “线形轨迹”、“淡出轨迹”、“管状轨迹” 或无轨迹。需要更多的信息可参见 *《MineSet 3.0 for Windows 企业版用户指南》* 中 “在散点可视化工具和平伸可视化工具中创建动画”。

应用模型

在 “工具管理器” 的 “数据转换面板” 中的*应用模型*按钮可以让您：

- 选用以前创建的分类工具并将它应用于新数据。
- 在当前数据表上测试以前创建的分类工具。
- 将当前数据拟合到以前创建的分类工具的结构。

在对话框的左上部 （图 1-4）是当前服务器上可用的所有的分类工具的列表。如果您选择了一个分类工具，右边列出该分类工具所需要的列名字和类型。如果这些必要条件与当前表匹配，底部的消息将会给出提示。如果当前表没有所选分类工具所需的所有列，底部的消息也有相应提示，所丢失的列会在右边的列表中被选中，并且底部的按钮变为无效。

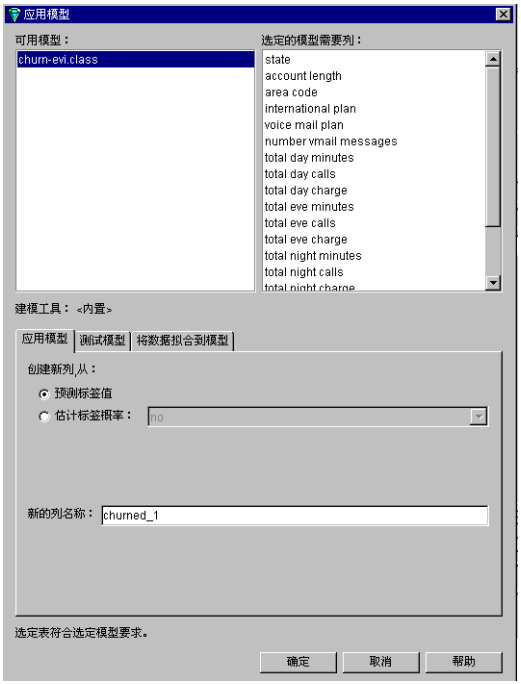

图 1-4 应用模型对话框：选择分类工具

应用模型面板

“应用模型”面板将以前创建的分类工具应用于当前表，正如图 1-5 所示。对于分类工具有两种应用模式：

- 为当前表中的记录*预测标签值*。例如，如果您创建了一个分类工具来测定客户波动，您可以利用该选项添加一列来标记每个客户可能波动或不可能波动。
- 为标签值*估计概率值*。用分类工具估计每个记录含有指定标签值的概率 （例如，客户波动 = 是），而不是用它来预测每一记录的标签值。假设用创建的分类工具来测定客户波动 , 您可以利用该选项来添加一个列，该列指示了每个客户可能波动的概率。

*新列名*文本字段可让您指定新列的名字。

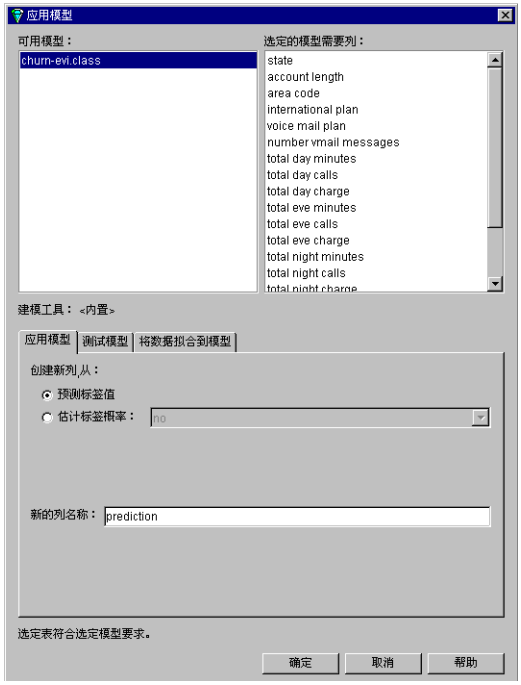

图 1-5 应用模型面板

测试模式面板

“测试模型”面板用当前数据表测试先前创建的分类工具。该表必须包括选中的分类工具所需的一定名称和类型的列。与 “应用模型”不同，“测试模型”也要求表包含一个标签列，该列与创建分类工具时所用的标签列拥有相同的名字和类型。

“测试模型”面板所拥有的选项可以使您：

- 显示表记录上应用分类工具的混淆矩阵。
- 显示指定标签值应用分类工具的上升曲线。
- 显示指定标签值应用分类工具的 ROI 曲线。

- 显示分类工具的可视化过程，并将数据表作为测试集，（这仅仅与 “决策树”和 “选项树”分类工具有关）。
- 选择用于记录权重的属性。

“测试模型”面板下部的文本字段显示这一结果。

将数据拟合到模型

“将数据拟合到模型”模式用于按照以前创建的分类工具格式化当前表中的数据。这将会产生一个新的分类工具，并拥有与原先的分类工具一样的结构；然而，新的分类工具利用表中的数据来更新概率估计 （见第 33 页 “修正”）。因为表中的所有数据用来拟合分类工具的结构，所以没有误差估计。“依照模型拟合数据”不能在利用推进建立的分类工具上使用。当您想要在独立的测试集上评估新分类工具的性能时 （完全从拟合数据中分离），可以使用 “测试模型”。

“将数据拟合到模型”面板所具有的选项可以使您

- 显示新分类工具的可视化过程
- 为新分类工具指定名称
- 选择一个属性用作记录权重

应用模型

在建立预测模型之后，您可将它应用于记录来预测其标签。例如，如果您建立了分类工具 （离散标签的一种预测模型），对于预测蝴蝶花类型，您可以将分类工具应用于只包含描述性属性的记录，然后添加带有预测蝴蝶花类型的新列。

为了确保数据质量 , 在建立分类工具以后，您可将它应用于训练集，以求找出被分类工具误标的记录。这些记录可能会确保更进一步的调查。也许它们是 “噪声”，或者会产生特别的直观结果。

例如，假如采用了*只用于分类器*模式导入蝴蝶花数据集的 “决策树”。如果您将此分类工具应用于数据集，您将得到一个包含预测标签的新列 （iris type_1）。然后您可以添加新列，以表达式 （iris type != iris type_1）定义为 **int** 类型。当分类工具分类错误时，新列为 1，当正确分类时，新列为 0。图 1-6 显示了 “散点可视化工具”中的数据图，其中新列以绿色为 0 （正确），红色为 1 （错误）的设置映射为颜色。观察这张图，可以确定哪里出错。

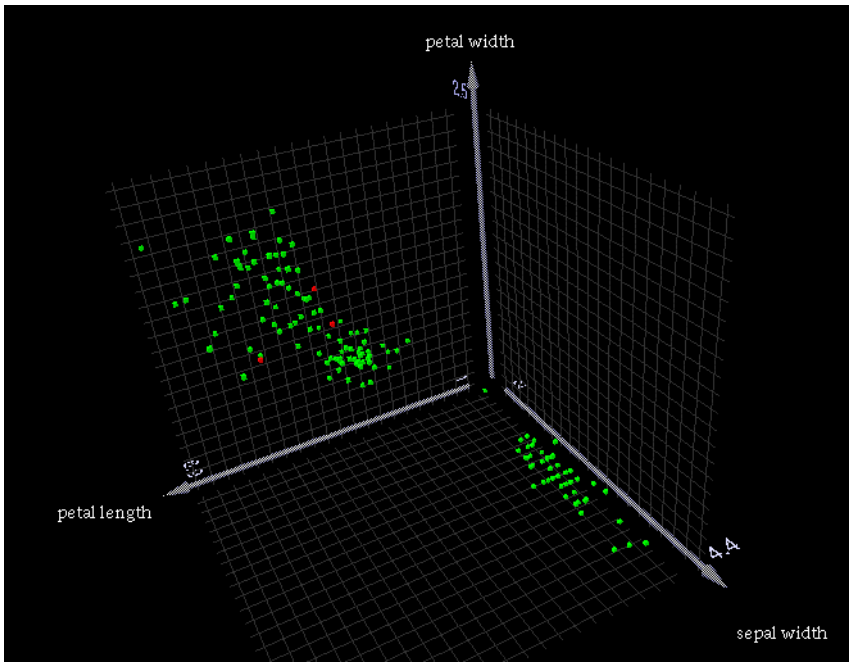

图 1-6 蝴蝶花数据集错误分类，例 1

另一个种方法是将新列用表达式 （iris type != iris type_1）+ 0.01 定义为浮点型。在原始标签被映射为颜色的条件下，可以使用 “散点可视化工具”，并且该新列映射为大小。不正确的预测显示为大立方体； 正确的预测显示为小立方体 （参见图 1-7）。

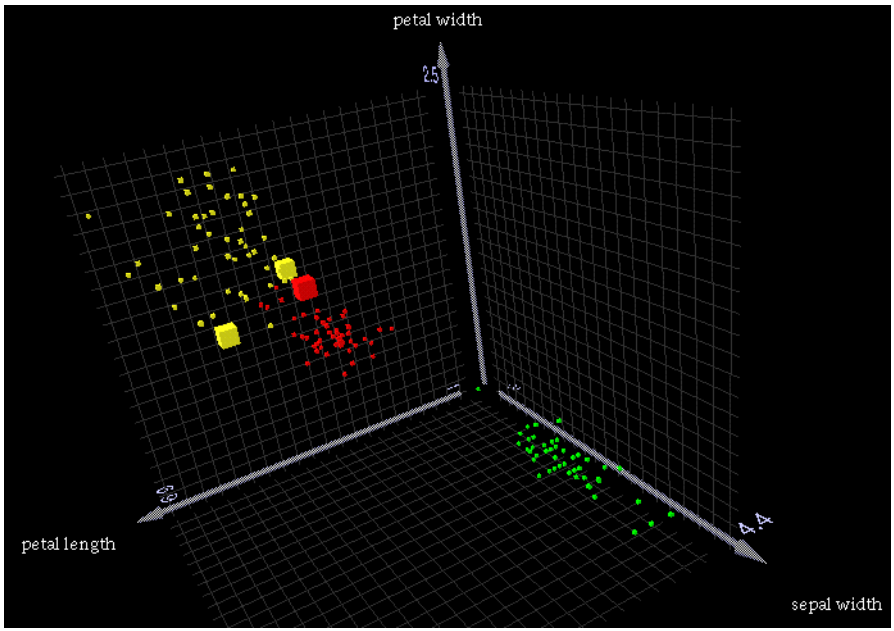

图 1-7 蝴蝶花数据集错误分类，例 2

要得到更多的关于误差和错误划分的信息可参见第 80 页 “错误估计”。

关联规则

在大型数据集中，可以通过用关联规则来建造、证实和图形化表达模式模型，进行数据挖掘。这些模式被关联规则表达，这些规则指示数据集中列值相伴出现的频率。关联规则的典型应用是市场供求分析，它将辨认哪些商品可能相伴出现在消费者的 “购物筐” 中。

发现并图形化显示关联规则与许多行业有关。关联规则能够产生有用关联的例子是超级市场存货计划、销售计划以及直销中的附加邮寄。

在利用关联规则工作时有两个步骤：

1. 产生规则：数据文件由 “关联规则产生器” 处理，它可以创建可被可视化工具利用的文件。
2. 规则可视化：该操作显示产生的关联规则。

“关联规则” 可产生简单 （一对一）和多路关联规则。这部分描述简单关联规则。关于多路规则的描述，参见第 25 页中的 “多路规则”。

简单关联规则可以这样表达：如果 X 为真，那么 Y 也存在为真的可能性。 MineSet 将 X 视为规则的左项 （LHS）而 Y 为规则的右项 （RHS）。

应用关联规则的一个例子就是得到消费者购买模式的市场供求数据。这里，市场供求就是消费者到商店一次所购买的商品。在此情形下的样例规则可能是 : “既买尿布又买婴儿粉的人占 80%”，这个百分数被视为规则的*置信度*。

在该例中，“尿布” 是规则左项 （LHS），而 “婴儿粉” 是规则右项 （RHS）。

这些规则的一些应用有：

- 如果商品 A 出现在 RHS，而 LHS 可以帮助我们确定商店应该怎样做才能促进其销售。
- 如果商品 B 出现在 LHS，而 RHS 可以帮助我们确定如果商店不再经营商品 B，哪种商品将会受到影响。

“关联规则产生器” 处理一个输入文件，然后产生一个由规则组成的输出文件。例如，如果 X 和 Y 是记录中的属性，那么规则：

X=>*Y*

说明只要 *X* 在一个记录中出现，则 *Y* 也很有可能一起出现。

关联的强度由四个数字量化 :

- 规则*支持度*描述整个数据集中规则出现的频繁度，或者说 *X* 和 *Y* 作为整个记录的一部分在数据集中一起出现的频率。例如，如果支持为 1%，则 *X* 和 *Y* 在所有记录中的 1% 中出现。

  您可以为产生的规则指定最小支持度阈值。默认的最小支持阈值为 1%。最小支持越低，产生的规则越多，工具的性能可能越慢。

  满足*最小支持度阈值*的规则很重要，这是因为两个原因:

  - 只有当有相当的一部分记录支持规则的时候，规则才会有商业价值。例如，如果每个买鱼子酱的人同样也买伏特加酒，规则鱼子酱 => 伏特加酒就有 100% 的置信度。但是，如果只有少数的人买鱼子酱 （也就是说，支持很低），那么对于零售商来说规则就有局限性。
  - 如果只有很少一部分的记录支持规则，那么该规则就不具有统计显著性。规则可能是因为偶然性，因此不要轻易根据规则做出判断。

- 规则的置信度定量为规则两边一起出现的记录数目除以 LHS 规则出现的记录数的商。例如，如果置信度为 50%，那么 *Y* 在 *X* 出现的记录中有 50% 的出现概率。这样，如果知道 *X* 在一个记录中出现了，那么 *Y* 也在同一记录中出现的概率为 50%。您也可以为产生的规则指定最小置信度阈值。最小置信度阈值的默认值为 50%。

- *期望置信度*可以衡量那些左项和右项看似没有关系的规则的置信度。它是根据数据集中出现 RHS 项的记录数目来计算的。所以期望置信度与置信度之间的差别衡量了根据 LHS 项的出现而进行预测的能力的变化。

- *上升度*代表了置信度和期望置信度之间的比率。值越大代表规则越不可预测。这就告诉您，当试图确定 RHS 是否在某一记录中出现时，规则的 LHS 可以提供多少附加信息。

关联规则生成器不报告置信度小于期望置信度的规则。换句话说，如果在 A 和 B 一起出现的频率小于 B 单独出现频率时，规则 A=>B 就不会列出。

注意： 如果只给出 Y 和 X=>Y 形式的规则，则对于 X 什么也推断不出来。规则只指定从 LHS 到 RHS 之间的关系。

关联选项卡

“关联”选项卡位于 “工具管理器”的 “数据目标”窗格中的 “挖掘工具”选项下，从中您可以产生关联规则。这些规则，通常称为市场需求分析，可以帮助您确定客观现象的一般分类。

文件需求

“关联规则”需要下列文件，这些文件由 “工具管理器”创建来产生规则可视化进程：

- 运行 “关联规则产生器”产生的规则文件，命名为 *.rules.data* 文件
- 带有后缀 *.rules.schema* 的方案文件，由此您可以在 “记录查看器”中查看规则
- *.rules.scatterviz* 文件

.rules 中缀并不是必须的，但是当这些文件由 “工具管理器”产生时就会被用到。

可视化关联规则

有几种方法可以启动规则可视化工具：

- 利用 “工具管理器”来配置和启动 “关联规则”工具 （参见第 24 页 “关联规则配置”）。一旦规则生成，“工具管理器”就自动启动 “散点可视化工具”。

- 如果您知道要用到的配置文件 , 可双击文件的图标 （在文件管理器窗口中）。这样就启动了 “散点可视化工具”并自动装入您所指定的配置文件。这只有当配置文件名以 *.scatterviz* 结尾时才起作用 （对于用 “工具管理器”创建的 “散点可视化工具”配置文件总是这样）。
- 输入这个命令：
  - 在 DOS 命令行提示下：

    `CD` *文件目录*

    `Viz` [*filename.scatterviz*]
  - 在 UNIX shell 命令行提示符下：

    `scatterviz` [*filename.scatterviz*]

当启动 “散点可视化工具”时，您必须指定配置文件，而不是数据文件。

关联规则配置

您可以利用 “工具管理器”配置关联规则的部分，这样可以大大简化任务。要在 “工具管理器数据目标”窗格中配置 “关联规则”文件，选择 “挖掘工具”，然后单击 “关联”选项卡。根据数据源，您可以分组或删除列以简化您的可视化过程。

关联规则选项

关联规则对话可让您指定几个选项。

置信度

除以 LHS 规则出现的记录数后，该选项可定量描述规则两边一起出现的记录数。例如，如果置信度为 50%，那么 *Y* 在 *X* 出现的记录中有 50% 的出现概率。最小置信度阈值的默认值为 50%。

支持度

该选项定量描述了整个数据集中规则的流行程度，或者说 *X* 和 *Y* 作为整个记录数目的一部分在数据集中一起出现的频率。例如，如果支持为 1%，则 *X* 和 *Y* 一起出现的频率占在整个记录数的 1%。您可以为产生的规则指定最小支持度阈值。默认的最小支持度阈值为 1%。

加权

在下列情况下关联规则允许记录加权，即指定某些记录比其它记录更重要或者因不均匀采样而进行弥补时。如果 “使用权重” 复选框没有被选中，每个记录的权重为 1。当复选框被选中时，在下拉菜单中可让您选择包含每个记录权重的列。*权重保留为属性*复选框，如果被选中，由 “关联规则产生器” 所建立的规则中将包含权重列。如果复选框未被选中，权重列将被排除在 “产生器” 建立的任何规则之外。参见第 138 页 “记录加权” 可找到进一步的解释。

多路规则

如果您选择了 “多路规则” 按钮，“关联规则产生器” 产生所有满足最小支持和置信度阈值的关联规则，包括那些在 LHS 和 RHS 中拥有很多项的规则。这个规则的一个例子是啤酒加空心粉暗示着薯条、色拉和白酒。

多路关联规则产生了 LHS 和 / 或 RHS 上拥有许多项的规则。图 1-8 说明了为多路规则的产生而配置的 “工具管理器关联” 面板。版本显示来自于 MineSet 3.0 for Windows。IRIX 版本是类似的。

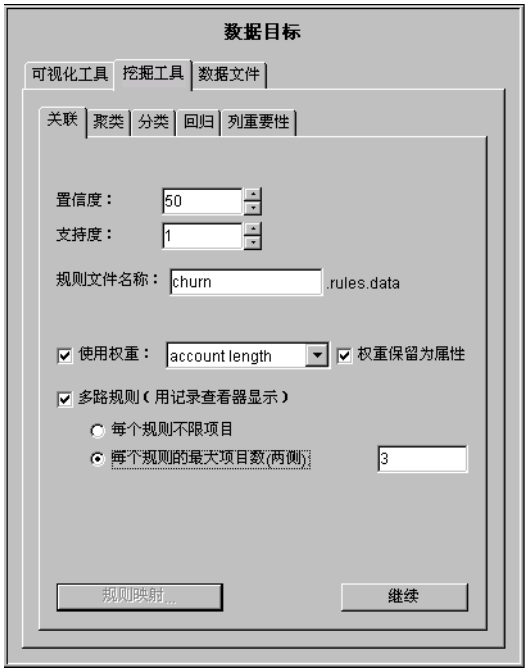

图 1-8 初始的工具管理器窗口显示多路关联的产生。

与其用 “散点可视化工具”还不如用 “记录查看器”显示多路规则。其中每行显示一条规则。表的前两列包含了 LHS 和 RHS 中项的数目。接下来的四列包含了支持度、置信度、期望置信度和上升度值。最后两列包含 LHS 和 RHS 项。在 LHS 和 RHS 列中，项之间由 and 分隔。在上面的样例规则中，LHS 包含两项并以啤酒和空心粉代表。RHS 包含三项，分别以薯条、色拉和白酒代表。

您可以通过在 “每个规则中最大总项目数”字段中输入数据来限制所产生规则的大小。该数字指示在任何规则中允许拥有的最大项数。规则中的项数是 LHS 和 RHS 中项数的总和。上述例子规则中有五个项，简单规则又有两个项。

注意： 产生多路规则要花费较多的时间。观察状态窗口可发现每次迭代产生的规则数目。如果产生了太多的规则，请删除操作并增加最小支持度或置信度阈值，或者降低在每个规则中最大可允许项数。

关联规则映射按钮

关联规则可让您将规则属性映射为显示内容中的可视成分。特殊的映射可以检验对于更进一步理解可视化过程的设想。在 “工具管理器” 中的 “数据目标” 窗格中，单击 *规则映射* 按钮打开在图 1-9 中显示的 “关联规则映射” 面板。每个成员只显示了在弹出式菜单中可用的选择。

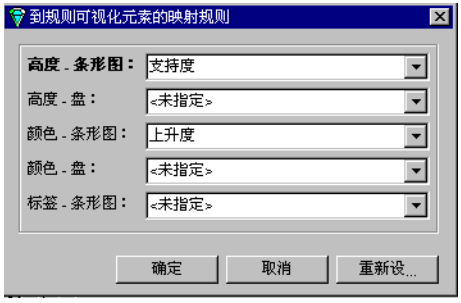

图 1-9 关联规则映射面板

您可以将从 *.rules.data* 文件中自动产生的列 （支持度、置信度、期望置信度和上升度）映射为该面板中的可视成员；默认值显示在表 1-8。参照第 24 页 “关联规则选项”可得到这些值的解释 :

表 1-8 关联规则映射

可视成员	描述
高度 - 条形图	可以指定条形图高度所代表的含义 （默认情况下映射为支持度）。
高度 - 盘	可指定盘高度所代表的含义。
彩色 - 条形图	可指定条形图颜色代表的含义 （默认情况下，映射为上升度）。
颜色 - 圆盘	可以指定圆盘颜色所代表的含义。
标签 - 条形图	标签 - 条形图可使您指定条形图标签所代表的含义。

关联规则可视化

图形显示关联规则允许您开发和比较产生的规则。利用 “散点可视化”工具在栅格场景中显示规则。左项 （LHS）项在一个坐标轴上，而右项 （RHS）项在另一个坐标轴形图上。规则的属性和特征在它的 LHS 和 RHS 项的结合处被显示出来。显示的内容包括条、盘和标签。

如果显示的视图太小，项标签不出现在坐标轴上。可以用 “伸缩”轮放大视图直到项标签出现。您可以查看特殊规则的标签，方法是当鼠标处在选择模式时，将鼠标光标放在个体条上。特殊规则的所有细节都会在视图左上方的区域中显示出来。

主窗口的底部显示了指示所示属性 （例如，条高度和颜色）和与基本规则 （例如，置信度和支持度）相关的值之间映射关系的图例。

条高度对应于支持度，而条颜色对应于上升度值。当这一画面放得足够大时， LHS 和 RHS 轴以项的名字标记，除非该项在配置文件中被关掉。

通过使用 “工具管理器”（见表 1-8）或使用编辑器改变配置文件，您可以修改标记以及条形图和盘的高度和颜色所代表的内容。当一个变量被映射为圆盘或条颜色时，颜色映射会自动的产生。如果您希望改变这些默认颜色映射，可以编辑配置文件。

如果将鼠标放在一个代表 “关联规则”对象的条形图上后，单击鼠标左键，信息就会出现在 “选择窗口”中。单击的同时按下 Shift 键就可以选择多条规则。

追溯是一个术语，用于指代在可视化过程中观察基本数据的行为。追溯表达式对选定规则进行逻辑 “与”的操作来指定。如果原始表中的列与 *.rules.data* 文件中的列不匹配，那么在执行了追溯操作时，规则产生器产生一个特殊列来帮助创建筛选表达式。这就意味着对追溯特性面板的修改无效，因为一个特殊的字符列已经映射为对 *.rules.scatterviz* 文件的追溯。

当您在一个或多个规则中追溯时， MineSet 显示所有满足规则的记录。从可视化工具 “选项”下拉式菜单中选择 “显示原始数据”。参见第 78 页 “追溯”条目可得到关于追溯的更多信息。

关联规则的样例文件

MineSet 软件包含了样例文件以展示关联规则的特征和能力。参见附录 A，配置和数据文件样例

将文件从 .ruleviz 转换到 .scatterviz

MineSet 2.6 版以前，关联规则在它们自己的可视化工具中显示，并且配置文件在格式上略有不同。

如果存在 *.ruleviz* 文件，您可以将它们转换成 *.scatterviz* 格式，方法是编辑 *.ruleviz* 文件并另存为 *.scatterviz* 文件。样例 1-1 和 样例 1-2 显示了 *.ruleviz* 和 *.scatterviz* 格式的差异。样例 1-2 包含了注解有助于您了解它们之间的变化。两种配置文件使用相同的数据文件。

注意： 在旧的 ruleviz 文件格式中，大小被称为高度，置信度称为可预测度，支持称作流行度。

样例 1-1 group.ruleviz

```
MineSet 2.5
input
{
    file "group.rules";
}

expressions
{
    double `pred/expected` = predictability/expected;
}
view
{
    height predictability;
    height max 10;
    height legend on;

    disk height expected;
    disk height legend label "disk height: expected predictability";

    color prevalence;
    color colors "white" "purple";
    color scale 0 10;
    color legend "0%" "10%";

    message "%s  implies  %s\npredictability: %.2f predictability/expected:
        %.2f  prevalence: %.2f", LHS, RHS, predictability, `pred/expected`,
        prevalence;

    options grid size 3;
    options hide disk distance 600;
    options hide item distance 600;
}
```

样例 1-2 group.rules.scatterviz

```
MineSet 3.0
input
{
# 将 group.rules 重命名为 group.rules.data:
    file "group.rules.data";

# rules.data 文件的架构总是 .
# 这样的。请添加下列行:
    int nlhs;
```

```
    int nrhs;
    float support;
    float confidence;
    float `expected confidence`;
    string LHS;
    string RHS;
}
expressions {
    float lift  = confidence / `expected confidence`;
}


view
{
# 替换高度可预测度:
    size confidence, scale 1.;
    size legend label "Bar Height: confidence";

# 替换期望的盘高度 :
    disk height `expected confidence`, scale 1.;
    disk height legend label "Disk Height: expected confidence";

# 替换颜色流行度:
    color support;
    color colors "white" "purple", legend label "Color: support";
    color scale 0 9;
    color legend "0%" "9%";

# 添加两个轴映射 ( 在旧的文件中没有 ):
    axis RHS, max 100, orderby alpha;
    axis LHS, max 100, orderby alpha;

# 确保形状为条形图 :
    options entity shape bar;
    options axis label size 20;

    message "%s implies %s\n support=%2.2f%%, confidence=%2.2f%%,
        expected confidence %2.2f%%, lift=%2.2f",LHS, RHS, support, confidence,
        `expected confidence`, lift;

    options grid color "#202020";

    options hide disk distance 600;
    options hide entity label distance 600;
}
```

自动分组

工具管理器列分组面板提供选择自动分组的选项。如果选择了您想要分组的列，在 Windows 中，选择 “自动分组”，或在 IRIX 中，从 “自动阈值”选项卡中 “自动选择分组的数目”，MineSet 将会为您分组。当您想用程序的机器学习功能为分组提供参考时，这是很有用的。见第 34 页 “分组”中的分组选项的完整描述。表 1-9 列出了可用的自动分组选项：

表 1-9 自动分组选项。

选项	Windows 位置	IRIX 位置	描述
自动分组	基本的	N/A	自动分组所选的列。
自动选择分组数	高级的	Auto。阈值	告诉 MineSet 选择分组的数目。
分成 _ 分组	高级的	Auto。阈值	可让您选择分组数目。
使用方法	高级的	Auto。阈值	可在自动、统一区间和统一权重中选择。
离散标签	高级的	N/A	可为分类选择标签。如果选择了自动方式，您必须从菜单中选择标签。
每个组的最小权重	高级的	N/A	为了减少组数，您可输入每组的最小权重 （或在未设置权重的情况下计算）。

在自动分组的过程中，选择阈值是为了使不同组中标签的分配尽可能的不同。这种方式持续产生阈值来分割区间直到没有明显的多余间隔。值的差别越大，选择组数越多 （为对数关系）。

单击自动选择框可通知 MineSet 根据实例中的总权重来自动确定最小权重：总权重越大，每组的最小权重就越高 （它们之间成对数关系）。

修正

修正可使您从一个训练集中建立结构，然后对大数据集进行修正以改进概率估计。修正的任务就是使模型概率估计与基本数据一致。修正比导入大模型结构的速度快。当利用预留误差估计时，您抽出了一部分数据用于检验。当您通过模型结构修正所有数据时，最终模型的误差减小了，因为计数，权重和概率反映了整个数据集。您可以在任何导入工具 “高级选项”面板中找到修正。

导入工具建立的模型拥有两个部分：

- 决策树和选项树的结构，结构就是树的形状。为了明显起见，结构是每个属性的分组的数目，而如果属性是数值型，那么结构就是阈值。
- 概率估计，结构的每个特殊部分可用来估计每个类出现的概率。这些估计通常是建立在结构中不同点上训练记录的计数的基础之上。对于决策树，概率是通过叶节点上记录的权重来确定的。对于证据分类工具，概率是通过每个属性值或区间的条件概率来确定的。条件概率是通过 “证据可视化工具”左窗口中的矩形图来显示的，该窗口显示了在给定 （在 .. 条件之上）每个标签值的情况下每个属性值的相对概率。

用一套记录来修正模型并不改变模型的结构，但会改变基于新数据的概率估计。修正很有用，原因在于：

1. 可以从一个小的训练集中建立结构，然后用大的数据集进行修正以改进结构中的概率估计。相对导入模型的结构，修正是一个更快的过程。
2. 当使用了预留误差估计，则抽出一部分数据进行检验。一旦导入了模型结构、估计了误差，则可以通过该结构对所有的数据进行修正，这样可以减小最后模型的误差。当计数、权重和概率出现在模型的结构中时，它们反映了所有数据，而不仅仅是训练集部分的特征。

当使用可视化工具中的追溯功能时，可以看到对应于权重的数据被显示，它反映了整个数据集的特征。如果没有使用修正功能，那么所显示的权重仅代表训练集。

对于所有的导入工具，您可以通过 “数据目标”面板的 “高级选项”按钮来访问 “修正检验集”（IRIX 的深层选项）。在 “数据目标”窗格中单击 “挖掘工具”并选择 “分类”或 “回归”选项卡，然后为所有导入工具选择 “分类工具”（或 “回归工具”）和错误估计模式。当 “推进”为可用时，修正复选标记被禁用。

分组

分组可以把一个或多个列中的信息分组，处理过程中创建一个新列 （例如，关于年龄区间的一列 0-18， 19-25， 26-35 等等）。“工具管理器”中的分组可以节约计算时间并简化可视化过程。如何分组的细节，可参见 *《MineSet 3.0 for Windows 企业版用户指南》*中 “为列改变或创建新分组”。

分组选项

对任一选择列， Windows 分组选项对话框初始给出三向选择：“不分组”、“自动分组”和 “自定义阈值”。 IRIX 版本给出两种选择，“自动阈值”和 “用户指定阈值”。

分组记号

MineSet 对组名使用了修改的间隔记号：

（ *较低的阈值* . . . *较高的阈值*]

在较低阈值旁边的括号 “）”指示该阈值不包含在区间之内。方括号，“]”指示阈值包含在区间之内。例如，（10.5 . ..12.6]指示值区间大于 10.5 并且包括 12.6。如果省略了较低的阈值，那么区间就为小于并包含较高边界的所有值。例如， (... 10.5] 表明值的区间小于等于 10.5。如果较高的阈值被省略，区间就为大于较低边界的所有值。例如， (12. 6...] 就是指值的区间大于 12.6。

分组方式

将数据按类分组有三种解决方法：

- *熵*—需要您选择离散标签。阈值的选择是为了使不同组中的标签分布尽可能的不同。这种解决方法持续产生阈值来分割区间直到没有明显的多余间隔。

  在每组最小权重文本字段中，可以指定任何组中的最小权重； 这就防止了以比指定值更小的权重来创建组。如果两个结果子区间中无一包含指定的最小权重，则不会对区间进行分割。默认情况下，每个记录只有一个权重。在这种情况下，指定每组中 “最小权重”与指定每组实例最小数目一样。

  利用算法自动地设置值较为每组指定最小权重的做法更好。根据记录权重的总和，*使用权重*复选框可利用该算法为每组计算最小权重：总权重越大，每组的最小权重就越高 （它们之间成对数关系）。

- *统一区间*—算法将值区间分解成指定数目的相同大小子区间。极限区间的较低和较高边界包括数据观察范围之外的所有值。例如，如果属性的值在 3-8 之间，并且您指定了四个组，那么所找出的阈值为 4.25、 5.5 和 6.75，与区间相一致：

  - 小于 4.25
  - 大于 4.25 小于包括 5.5
  - 大于 5.5 小于包括 6.75
  - 大于 6.75

  极限区间的较低和较高边界包括数据观察范围之外的所有值。MineSet 对这些组名的记号为：

  - (... 4.25]
  - (4.25 ... 5.5]
  - (5.5 ... 6.75],
  - (6.75...]

- *统一权重*—算法将值区间分为指定数量的等权重组。与 “统一区间” 在值域中阈值将值划分为等大小间隔的区间不一样，而 “统一权重” 确定了将记录分成等权重子集组的阈值。默认情况下，每个记录只有一个权重。在这种情况下，“统一权重” 解决方法产生了指定数目的组，每组包含了大约相等的实例数。

统一区间和统一权重都可以确定修剪因子，修剪因子指定了在分组之前排除在值区间最外的极值部分。默认的修剪因子为 0.05。这就排除了记录中 5% 带有极值的记录 （其中 2.5% 为区间中的最低值， 2.5% 为区间中的最高值）。修剪的目的就是减小在阈值产生过程中异常值的影响。

所有方法都可让您决定是人工指定组数还是让算法自动选择数目。对于 “统一区间” 和 “统一权重” 解决方法，自动分组过程是以相异值的数目为基础的：相异值越多，选择的组就越多 （为对数关系）。

典型情况下，在确定阈值时所有可用的实例都将被使用。当被分组的属性后来被用于导入模型时，误差估计指标将为最优。这是因为检验集中的分布信息被用于确定阈值。

*仅使用训练集*防止了分组解决方法在确定阈值时看到实验集中的记录。这就给出了分类工具错误率的更实际的估计。*只使用训练集*需要用户指定相同的 “预留率” 和 “随机子” （参见第 101 页 “导入工具误差选项”）用于为估计分类工具误差而创建预留集。

“使用权重” 菜单可以让您用任何数值型属性对实例进行加权。改变记录加权会影响 “自动” 或 “统一权重” 方法，但是不会影响 “统一区间” 方法。

如果您单击*应用*，“工具管理器” 会挑选组阈值并将之显示在 “所选列的阈值是” 的文本字段中。“分组列” 窗口底部的文本字段显示了分组算法的进程和任何出现的错误。

推进 （boosting）

推进是一种算法，该算法创建几个不同模型并利用加权投票法合并它们的预测过程。推进模型通过在相比其它数据而言很难进行模型化的数据中将归纳过程集中在样例上的方式来不断改进准确性。在所有使用 “推进” (无可视化) 复选框的导入工具模型中，“深层导入工具”选项对话框中的 “推进”是可用的。

在一些情况下，在创建模型过程中最重要的问题是错误率。例如，假设您已经分析了数据集的客户波动预计，到达了较为满意的一点，并准备创建用于预计用户中谁最有可能波动的模型。在这种情况下，您不再对模型的可视化感兴趣 , 因为您对相关的因素有了相当好的理解。您也想尽可能得到最好分类准确度。在这种情况下，您也许想要推进。

推进并不总是增加准确度，但通常是这样的。推进模型不能被可视化，虽然可以看到其混淆矩阵、上升曲线、学习曲线和 ROI 曲线。推进是密集的计算过程，通常要比相应没有推进的导入工具的计算时间长 25 倍。修正并不与推进分类工具一起工作 , 这是因为推进对多重模型和用于训练它们的记录进行了特殊的加权。

您可对拥有任何可取值数目的标签来使用推进。推进并不总能改进归纳模型的错误率；这也被强调为多值标签带来的问题。

更改列类型或名称

该 “工具管理器”选项可使您改变列的类型或名字。参见 *《MineSet 3.0 for Windows 企业版用户指南》* 中的 “更改列类型或名字”可得到更多信息。

选择点

选择点是当您从 “工具管理器挖掘工具” 的 “聚类” 选项卡中单击了迭代 k- 均值时，显示一个高级选项。选择点是聚类参考数目和 0 之间的一个点，较高的选择点建议了较大的类别数而较低的选择点则建议了较小的。选择点为 1.0 时总会找出上边界。在聚类过程中，如果您的界限是 1 到 5 个类，那么选择点为 0.4 时将选出两个类，而当选择点为 0.8 时将选出四个类。选择点为 1.0 时总选出五个。

分类工具

*分类工具*在假设其它属性已知的情况下，预测一套数据的某一个属性。分类工具是一种模型。属性是数据集中的继承特征。例如，如果有了通讯公司的用户数据，假设已知用户是否有声音邮件、出国计划以及使用电话的时间，就可利用分类模型来预测用户是否会波动。被预计的属性被称为标签，用于预计的属性被称为描述性属性。

MineSet 可以从一个训练集中自动建立分类工具。训练集由提供了标签的数据中的记录组成，（也可参见 第 176 页 “训练集”）。例如，您提供了一个数据库表，每一列对应于一个描述性属性 （例如，声音邮件计划的出现，每天打电话的平均分钟数），其中一列为标签 （客户波动或不波动）。从训练集中自动创建分类工具的算法被称为导入工具。

当产生了分类工具， MineSet 也产生可视化过程来帮助理解分类工具是怎样操作的。可视化过程也可以为数据本身提供有价值的直观结果。一旦分类工具产生了，它就可对未知标签值的记录分类。该值由分类工具来预测。

分类工具有两个部分：结构和概率估计

- 结构—对于 “决策树” 和 “选项树”，结构就是树的形状。对于 “证据”，结构是每个属性分组的数目，如果属性是数值型的那么就是阈值。“决策表” 的数学结构与 “决策树” 的相同，但其显示与 “证据” 可视化过程相似。
- 概率估计—结构的每一部分都估计了每个类的概率。这些估计通常根据结构中不同点上训练的记录数目来进行。对于 “决策树”，概率是由叶节点上面的记录的权重来决定的。对于 “证据” 分类工具，概率是通过每个属性值或区间的条件概率来确定的。当忽略其它所有的属性时，先验概率意味着在训练数据集里随机选择的记录中发现特殊类标签的概率。例如，*糖尿病*的概率，换句话说，带有类标签的记录总数，除以数据集中记录的总数。条件概率是，当您选择了*糖尿病*标签时，落入例如 *age_60+* 类中特殊记录的概率 （即建立在选择标签的条件之上）。

注意： 参见 *《MineSet 3.0 企业版接口指南》* 附录 A 中的关于分类工具的进一步阅读材料的清单以及有关在 MineSet 样例文件里使用的数据集声明。

分类工具名字

正如 “工具管理器” 中所确定的那样，产生的分类工具以阶段文件名作为前缀，至于合适的后缀，例如， *-dtable.class* 是用于 “决策表” 的分类工具， *-dt.class* 是用于 “决策树” 分类工具。默认情况下，所有的分类工具都存储在 服务器上的 *file_cache* 目录中，默认为 *mineset_files*。这些分类工具将来可用于给未标记的记录进行分类；也就是说，它们可以用来对未标记的数据集预计标签 （参见第 15 页 “应用模型” 和 第 33 页 “修正”）。

分类选项卡

要访问 MineSet 分类工具，在 “工具管理器” 的 “数据目标” 窗格中单击 “挖掘工具” 选项卡可找到 “分类” 选项卡。您可在四个模式中选择一个：“分类器和错误估计”、“仅用于分类工具”、“估计误差” 和 “学习曲线”。这些模式中的每一个都可以与任何不同的导入工具进行组合：“决策树”、“选项树”、“证据” 和 “决策表”。要想知道每个导入工具的具体信息参见指南中相应的条目。有关这些分类工具的用法，参见 *《MineSet 3.0 for Windows 企业版用户指南》*。

分类

聚类选项卡可使您访问 MineSet 的聚类功能，对于开发不熟悉的数据集来说是非常有用的挖掘工具。因为聚类是一种与关联规则发现相似的描述性挖掘工作，所以您不需要将特殊的列设计为标签。除此而外，数据集从不分为训练和检验集；分类模型总是从整个数据集中创建并进行评估。

聚类模型存储为一套原形记录，每类一个。它们代表了聚类中所有数据的加权平均并被作为 *聚类中心* 或 *重心*：不象标准数据库记录，聚类中心对于数值型列以汇总统计的形式，对于类目列以直方图的形式为每一列保存了分布。

从 “工具管理器” 的 “数据目标” 窗格中的 “挖掘工具” 选项卡中可运行聚类。聚类的目的是确定数据集中有哪些相似的特征。然后，您可以看到带有不同参数的聚类和实验结果。

MineSet 中的所有聚类都在 k- 均值方法的基础上使用了合并算法；算法将相似的记录分组形成了类，其目标是使每组中整体的相似性最大。

要使用聚类工具，在 “工具管理器” 主窗口中的 “数据目标” 窗格中选择 “挖掘工具” 选项卡。从随后产生的选项卡中选择聚类。这样，主聚类面板就会显露出来。

您可以从主 “聚类”面板中单击*运行*来启动聚类过程；并没有必需的选项。默认情况下，您将利用单步 k- 均值聚类方法来发现数据中的三个类。一旦聚类完成后，您将看到聚类的评估结果，然后 “聚类可视化工具”将随之出现。

聚类面板提供了下列选项：

- 方法—允许在单步 k- 均值和迭代 k- 均值方法之间进行选择。默认情况下是单步 k- 均值。两种方法都在下面进行了详细的描述。
- 聚类的数目—对于简单 k- 均值法，您必须指定要寻找的聚类数目。默认值为 3。
- 聚类数目范围—对于迭代 k- 均值方法，您必须指定聚类可能数目的上限和下限。默认为 1 ... 10.
- 选择点—对于迭代 k- 均值方法，您必须指定*一个选择点*。该值在 0 和 1 之间，并有助于选择聚类的最终数目。参见第 42 页 “迭代 k- 均值方法聚类”。默认值为 0.5。

注意： 聚类是一个计算密集过程，要在较大的数据集上完成需花费一些时间，尤其在运行迭代 k- 均值模式下。如果数据集超过了 10,000 条记录，最好对数据的样本进行聚类。

用单步 k- 均值方法聚类。

术语 *k- 均值*是指一种方法，该功能根据记录间的相似性来确定尽可能好的聚类方案。这种方法是 MineSet 中聚类的最简单的形式。您指定了所需的类数目，而该算法将数据中的记录分组使每个类中的整体离散度最小。离散度是指类的凝聚性；离散度越高，每个记录离类中心越远。从技术角度讲，离散度可以通过记录到所指定的类中心之间距离的均方根的大小来度量。

算法本身是迭代性的，其处理过程如下：

1. 例如，可选择寻找五个类。
2. 五个类的中心在记录空间中被初始化为随机位置。随机子参数的选择会导致不同的开始位置。

3. 数据中的每个记录被分配给中心离它最近的类中。然后在每类新数据的基础上重新计算类中心位置。
4. 如果记录与其它中心间的距离比其与所属类中心之间的距离还要近，那么记录将移至最近的类中。然后，类中心将根据新数据重新计算。这一步骤将一直重复直到没有新的改进。

这一算法确保在有限的迭代中完成计算。

步骤 4 决定了聚类的运行时间。因此，过程窗口在每运行一次步骤 4 时出现一个进度条，并伴有注释指明到目前为止已进行了多少次迭代。您可以设置在计算停止前所允许的最大迭代次数限制 （运行步骤 4）。默认值为 20。

类的名字从 1 开始编号，虽然类名由数字代表，但对于类来说并不代表顺序。

迭代 k- 均值方法聚类

迭代 k- 均值聚类是单步 k- 均值聚类方法的一个更复杂的分支，在 “工具管理器”的 “聚类”选项卡上作为一种方法列出。不象单步 k- 均值方法，它并不需要指定所需创建类的确切数目，但是需要下限和上限以及一个选择点。该算法将在上下限之间的某处挑选适合于数据集的一些类。选择点是 0 和聚类参考数目 之间的一个点，较高的值代表了较大量的类而较低值则代表了较小量的类。选择点为 1.0 时总会选为上边界。例如，如果界限为一和五个类，那么选择点为 0.4 时将选出两个类，而当选择点为 0.8 时将选出四个类。选择为 1.0 时总选出五个类。

要运行迭代 k- 均值方法，您要选择三个参数：类数目的下限 （默认为 1），类数目的上限 （默认为 10），以及一个选择点 （默认为 0.5）。在给定这些参数的条件下，算法进行如下：

1. 利用聚类最小数目来运行单步 k- 均值算法。这将产生初始聚类结果。
2. 找出具有最大离散度的类，将之一分为二，创建两个新类。初始类的一半记录分配到新的类中，而另一半则分配给另一新类中。在所包含数据的基础上重新计算新类的中心。
3. 如果记录与其它中心间的距离比其与现属类中心之间的距离还要近，那么记录将移至最近的类中。然后在每类新数据的基础上重新计算类中心位置。这与单步 k- 均值方法的步骤 4 一样。作为单步 k- 均值算法，该步骤一直被重复直到没有记录移动或运行了所允许的最大的次数为止，该数值由最大迭代次数参数指定 （默认为 20）。

   步骤 2 和 3 一直被重复直到产生最大类数目为止：

该过程可以在最大和最小聚类数目之间产生一系列聚类结果。

最终呈现出聚类结果由选择点参数来确定 （默认为 0.5），过程如下：

通过衡量每个类的平均聚类离散度而对每一聚类结果进行了评估。随着类数目的增加，平均离散度总是下降。然而，它并不是均匀的下降。选择点选择了所需的离散程度，用最小类数聚类离散度到最大类数聚类离散度之间的比例衡量。所测的聚类离散度与该值最接近的聚类是最终聚类结果。如果选择点为 1.0，则总是选出最大类数的聚类，如果选择点为 0.0 则总是选出最小的。

之所以称为聚类是因为是在分割过程中的不断分离结合。最初的聚类 （在类数最小的基础上进行）用序数来命名，就好象单步 k- 均值一样。类每次被分割时，两个新类赋予了分割类的名字，但加上了 "A" 或 "B"。例如，类名字为 "2-B-A" 就是由初始聚类中类 2 派生而来的，并且分割了两次。

当创建了聚类模型，您可以在状态窗口中显示出统计结果：iris 数据集中的样例：

```
聚类结果：
----------------------------------------
记录到重心的整体均方根距离：0.216 +- 0.0928 RMS 从记录到每个重心的距离：
    类 1： 0.2306 +- 0.09484
    类 2： 0.1921 +- 0.09932
    类 3： 0.2247 +- 0.08058
模型另存为 iris.cluster
```

数据是每个聚类离散度以及整个聚类优化拟和度的度量。根据不同数据集、数据或不同的聚类数所得到的聚类离散度是不可比较的。

聚类选项

k- 均值聚类算法依赖于记录到类中心之间距离的计算。数据中的每一列被看作为所有记录多维空间中的独立维。默认情况下，数据中的每一列对最后的距离起着相同的作用。然而，您可以改变每一列或属性的影响力 , 可以通过在 “聚类高级选项”对话框中指定*属性权重*来实现。

属性权重

列的属性权重是一个大于或等于 0 的值，可在聚类算法的距离计算中确定列的影响力。属性权重设为 1 代表平均影响力。设为 2 代表影响力恰好为同样列的两倍。设为 0 致使该列对聚类不产生作用 （就好象在进行聚类前将该列从数据中删除一样）。属性权重不必为整数，也可以小于 l。

单击主聚类面板中*高级选项*或*深层选项*（IRIX）按钮可设置属性权重。“高级选项”对话框的上部显示了当前属性权重，所有值都默认为 1。

要改变一个或多个权重，选择您期望改变其权重的列（按住 Shift 并单击可选择或撤消选择区间，按住 Control 并单击可同时选择或取消选择多个值）。现在，在选定的权重字段中敲入新权重。

最后，单击 *设置* 按钮已选择的权重将改变成新值。您可以利用 *选择全部选定* 按钮同时选择和设置所有的权重。

您需要不断地调节属性权重以有助于发现更好理解的聚类结果。加权准则为：

- 含有大量值的字符或枚举 （枚举数组）类列应设低权重 （通常为 0）。如果用了该算法这些列将导致聚类结果的严重偏差，它们以后有助于解释聚类结果。
- 如果发现有几列具有较强的相关性，调节权重使这些列的 “权重总和”为 1，否则，这些列会过分的影响聚类结果。
- 类目型 （字符或枚举）列之间的距离趋向更大，通常将它们的权重设得比实值列还要低。

聚类选项对话框

您可以从 “工具管理器”上的 “挖掘工具”选项卡中访问 “聚类选项”对话框。选择 “聚类”选项卡，然后选择*高级选项*按钮来得到 “聚类选项”对话框。所表现出选项取决于您利用了单步还是迭代 k- 均值方法。

- *属性权重*

  聚类提供了对数据集中每个属性值进行不同方式加权的机会。参见第 200 页 “加权”可得到关于该选项的更多信息。

- *距离度量*

  这种度量确定了记录与聚类中心之间距离的测量方法。默认的选择是欧氏距离，在数据中的每一列为一维的条件下，它度量了多维空间中沿直线的距离。弹出式菜单提供了可替换的选项：曼哈顿距离。曼哈顿距离通过将沿每维坐标轴上的距离相加得到。这种度量方法因曼哈顿中横向的街道街区而得名：不可能从点 A 直接到点 B，因为只能沿着平行于坐标轴 （街道）的路来行进。

- *最大迭代次数*

  该选项设置了通过聚类算法所使用数据集数目的限制。更特别的是，它在 k- 均值算法中限制了步骤 4 运行的次数。参见第 41 页 “用单步 k- 均值方法聚类。”。

- *随机子*

  不同的随机子将导致初始聚类中心的不同起始点。参见第 41 页 “用单步 k- 均值方法聚类。”

- *使用权重*

  与 MineSet 中大多数挖掘工具一样，聚类支持记录加权。这一选项允许您指定一列（必须为数值型），该列将指定数据中每个记录的权重。

- *权重保留为属性*

  如果该框被选择了，则指定的加权列也会被聚类算法用做为正常的属性。如果未选择该框，则指定加权列将不被用做属性； 它将会从属性加权部分中消失 （固有值为 0）。

聚类可视化工具

“聚类可视化工具”展示了一系列棒状图和直方图。一旦执行了聚类操作，您可以利用“聚类可视化工具”直接查看类中心。

作为选择 , 您可以在训练数据上直接利用应用模型特征 （见第 15 页 “应用模型”）为每个记录设定类。您可以利用 MineSet 中的许多其它工具来开发聚类结果。

文件需求

“聚类可视化工具”需要以下文件：

- 数据文件由制表符分隔的字段行组成。利用“工具管理器”可以很容易的创建这个文件。您可以通过从数据源（例如数据库）中抽取数据来产生数据文件，然后将其格式化以便于聚类可视化工具的使用。数据文件有用户定义的扩展名（由聚类可视化工具提供的样例文件具有 *.clusterviz.data* 扩展名）。
- 配置文件描述了输入数据的格式以及如何进行显示。工具管理器可以创建这个文件，或者您自己可以利用任何编辑器（例如， jot、 vi、 Emacs 或者您喜欢的文本编辑器）来产生这个文件。

  配置文件必须具有 *.clusterviz* 扩展名。当启动“聚类可视化工具”，或当打开文件时，您必须指定配置文件，而不是数据文件。

启动聚类可视化工具

有几种方法启动“聚类可视化工具”：

- 利用“工具管理器”来配置并启动“聚类可视化工具”。（参见第 17 页“工具管理器”来浏览“工具管理器”的功能，这些功能对所有 MineSet 工具都一样；参见*《MineSet 3.0 for Windows 企业版用户指南》*的“在工具管理器”中利用“聚类可视化工具”样例。）
- 从“工具管理器”的“可视化工具”下拉式菜单中选择“聚类可视化工具”。通过选择“文件”>“打开”来打开配置文件。
- 如果您知道要利用哪个配置文件，双击配置文件的图标。这样就启动了“聚类可视化工具”并自动装入了指定的配置文件。只有当配置文件以 *.clusterviz* 结尾时这种方法才起作用（在利用“工具管理器”为“聚类可视化工具”创建的配置文件的情况下，总是这样）。
- 从 UNIX 命令行中输入：

  `clusterviz [` *`配置文件`* `]`

  *配置文件*是可选的，它指定了要使用的配置文件的名字。如果未指定配置文件，那么您必须用“文件”>“打开”来指定一个。

颜色选择

工具选项对话框中有许多设置都有选择颜色选项。当使用 “工具管理器” 时，单击*高级选项*按钮 （Windows）或*深层选项* （IRIX）按钮来得到工具选项对话框。

用颜色选择器 来 选 择 颜 （Windows）

要显示 “颜色选择器”，单击颜色开关或 “工具选项” 中颜色列表里的 + 符号或者 “高级选项” 面板。如果选中了颜色开关列表，该选项列表如图 1-10 中所示 （最初是空的）。

图 1-10 颜色开关列表

当单击新颜色开关时，您将看到该颜色出现在 “颜色选择器” 的 “近期” 和 “预览区域” 内 （图 1-11）。如果您改变了主意，可以返回到 “近期栅格” 中显示的以前颜色开关上。一旦您选中了颜色，单击*确定*，然后该颜色就加入到您的颜色列表中。不用取消 “颜色选择器”，您就可以添加几种颜色。

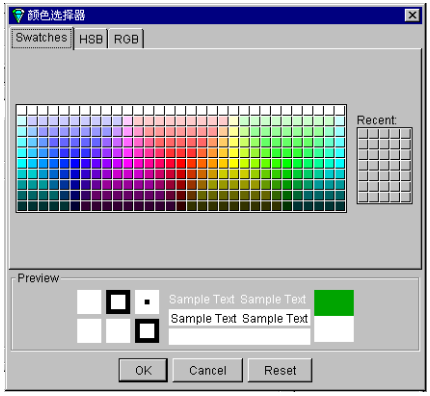

图 1-11 颜色选择器对话框

您可以在颜色选择器窗格中用 HSB （色度、饱和度和亮度），或以 RGB （红、绿和蓝）方式来选择颜色。在 HSB 窗格中，单击大的颜色方块。白色的圆圈就会出现。

当选中 H 单选钮时，您可以通过在方形区域中拖动圆圈来调节饱和度和亮度。要调节色度，移动彩虹条旁边的滑动条。

当选中 S 单选钮时，您可以通过在方形区域中拖动圆圈来调节色度和明亮度。要调节饱和度，移动彩色条旁边的滑动条。

当选中 B 单选钮时，您可以通过在方形区域中拖动圆圈来调节色度和饱和度。要调节明亮度，移动彩色条旁边的滑动条。

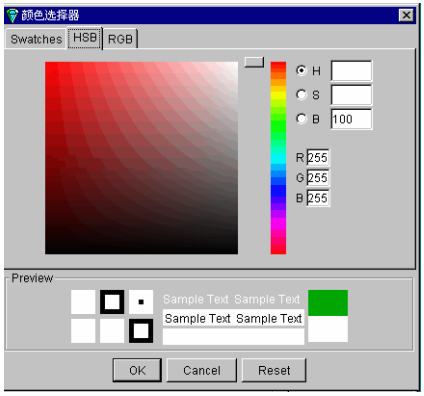

图 1-12 “颜色选择器”对话框的 HSB 窗格

在 RGB 窗格中，您可以利用滑动条来指定红、绿和蓝色值的量从而选择颜色。

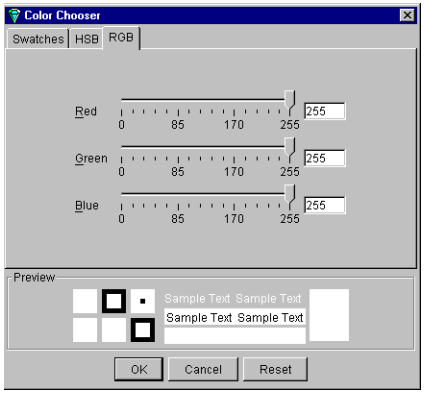

图 1-13 “颜色选择”对话框中的 RGB 窗格

用 “颜色浏览器”来选 择 颜 （IRIX）

如果 “颜色浏览器”是可用的，它将会出现在工具选项面板上。MineSet 拥有一个使用颜色开关的颜色列表选择器。该部分描述了如何为 MineSet 可视化工具选择、应用和改变颜色选项。

如果只有一个颜色被选择 （例如，一个栅格颜色），则会出现单一颜色开关。

单击开关可产生颜色浏览器，您可以改变开关的颜色 （图 1-15）。

如果选择了颜色列表开关，就会出现一列开关 （它们最初是空的），显示如图 1-14。

图 1-14 颜色开关列表

要编辑颜色，用鼠标左键单击开关。这也就同样选择了用按钮改变颜色的开关。如果用鼠标中键单击开关，则开关被选中，但还不能出现颜色选择器。

在开关列表的旁边是四个按钮。第一个是标有加号 （+）的按钮，用它可以在列表的末尾添加新的颜色。添加了开关后，颜色选择器出现了，您可以选择开关的颜色。如果列表中已经有了最大数目的颜色，那么 “添加”按钮被禁用。

下一个按钮标有减号 (-)。该按钮删除选中的颜色。如果没有选择开关，或者列表中只有最小数目的颜色，该按钮被禁用。

在这个 “删除”按钮的旁边是另外两个按钮来向左和向右转换所选的颜色。如果没有选择开关或这些开关已经在列表的底部，这些按钮则被禁用。

如果颜色数量超过了显示面积所能容纳的范围，则滚动箭头就会添加至列表的每一端。如果硬件用完了颜色，颜色开关被文本标签替换，将颜色以十六进制计数来显示。

在可视化 “配置选项”面板中的 “颜色”面板上，当您单击颜色开关或添加按钮时 “颜色浏览器”（图 1-15）就会出现。

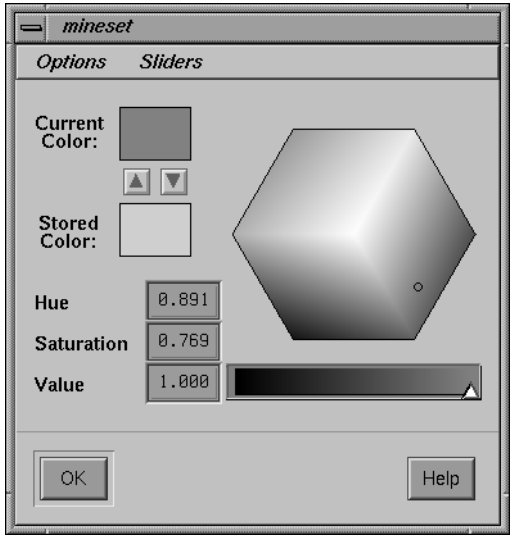

图 1-15 颜色浏览器 （IRIX）

用 “颜色浏览器” 选择颜色：

1. 颜色六边形中，您可以在小白色圆圈的上面移动鼠标。
2. 按鼠标左键按钮，然后在六边形内移动鼠标。小圆圈下面的颜色就会出现在在*当前颜色*标签旁边的矩形中。当您选择颜色时，这个矩形可作为调色板。
3. 当小圆圈在您所想要颜色的上部时，松开鼠标按钮。所选开关立即呈现出选中的颜色。
   您可在不取消 “颜色浏览器” 的情况下编辑几个颜色； 单击选项面板中的任意颜色开关就可以在已经出现的 “颜色浏览器” 中编辑该颜色。

当您对颜色作出选择后，单击 *确定* 按钮。“颜色浏览器” 窗口关闭。

列重要性选项卡

通过工具管理器的列重要性选项卡可以访问 “列重要性” 工具。见下面的 “列重要性”。

列重要性

在 “工具管理器” 的 “挖掘工具” 面板上，*列重要性*可从标有 “列重要性” 的选项卡上运行。在对所选标签列区别不同值的过程中，它有助于发现数据集中哪些是最重要的列。

列重要性和其它数据挖掘工具，例如，聚类 （在第 40 页 “分类” 中讨论）之间的差异是：利用列重要性 , 可以决定使用哪个标签来确列重要性。在聚类中，数据本身显示了那些是区分因素。因为分类模型表述方法之间的差异，不同属性对于不同模型就可能显得更为重要。 MineSet 提供的样例文件，显示了在某种情况下列重要性非常有用，可见附录 A，配置和数据文件样例。

寻找重要列

按照*列重要性*，例如，要寻找三个最好的列用于发现*适当信贷风险*标签，然后将它们选出来并映射为“散点可视化工具”的坐标轴。当您选择了标签并单击了*运行*，就会出现一个弹出式窗口，窗口中带有三列，这三列就是最好的鉴别工具。一种叫做“纯度”的测量（从 0 到 100）会告诉您用列区分不同标签的优劣程度。添加更多的列只能够增加纯度。

纯度可作为标签值分布斜率的一种度量。累计纯度测量是分割数据的纯度的度量。使用列分割数据，该列以决策树中分割数据的同样方式被发现为重要的。分割中的每个集都有其本身的纯度度量，并且分割过程中的纯度度量是这些个别度量的组合。对于已知的分割集，如果每个类具有同等代表性，其纯度为 0；如果每个记录为同一类，则纯度为 100。同样地，如果分割中的每个集具有同等类代表性，则累计纯度为 0；如果每个分割集中包含的记录都属相同的类，则累计纯度为 100。

“列重要性”有两种模式：

- 简单模式

  要调用简单模式，从弹出式菜单中选择离散标签，然后指定您想要看到的列数，然后单击*运行*。

- 高级模式

  利用“高级”模式可以控制列的选择。要输入高级模式，单击“列重要性”面板中的*高级模式*。出现一个对话框后，您可以选择加权属性并决定对于确定重要性它是否表现为普通的属性。对话框包含两个列名列表：左边的列表包含可用的属性，而右边的列表包含了被选为重要的属性（可通过人工或重要性算法来确定）。

高级模式可以以两种不同的方式工作：寻找几个新的重要属性或者对可用属性进行分级。

- 寻找几个重要属性

  要输入这个子模式，单击对话框中部两个单选钮的第一个 （*.. 寻找 [数字] 个附加重要属性*）。如果在没有再改变其它设置的情况下单击了*运行*，其效果与简单模式中的一样，寻找指定数目的重要列并自动移入到右边的列中。每个列的旁边，都给出了累计纯度 （也就是说，向上所有列的累计纯度）。

  作为选择，通过将列名从左列表移向右列表，可以事先指定想要包含的列并让系统添加更多的列。例如，要选择*柱面*列并让系统找到超过三个的列，单击*柱面*列名，然后单击列表之间的向右箭头。

  单击*运行*可以看到每个列的累计纯度，还有列表以前的纯度。纯度为 100 意味着，利用已知的列，您可以在数据集中很好的区别不同的标签值。

- 对可用的属性分级

  利用高级模式您可以计算每个列要加入到已被标为重要的列中所产生的纯度变化，例如，您可以将*柱面*移到右边的列表中，然后请求系统计算左列表中剩下的每列所产生的纯度增加改进。对于右边 （已经标为重要）的列计算累计纯度。

  为进入这种子模式，单击对话框底部两个单选钮中的第二个 （*... 对左栏各列计算改良纯度，对右栏各列计算累计纯度）*。该子模式允许对过程进行优化控制。如果两个级别很接近，但您可能更喜欢其中之一 （例如，因为它收集成本低、更实用或更容易理解）。

列的重要性取决于该列以前是否被标为重要。例如，*网收入*单独来看可能是一条好列，但它可能没有*工资*重要，因为它们看起来是高度相关的。三个最好的列并不是一定由那些单独来看级别很高的列组成。如果两列分别以美元或另外一种货币形式来表示收入，它们的级别是一样的；但是，一旦其中之一被选中，另外一个不能再提高区分能力。

对于为 “散点可视化工具”和 “平伸可视化工具”寻找三个最佳坐标轴的过程来说，列选择是很有用的。将标签选为在 “树可视化工具”中使用的关键字时，它对于为 “树可视化工具”寻找好的区分等级 （即分开不同标签值的等级）也是很有用的。

利用自动离散化过程将所有浮点型的值 （**双精度** 或 **浮点型**）先离散化。如果左边列表中的列没有赋值，算法将不予考虑；这要么是因为它具有单一值 （例如，当它在被离散为单一个区间时），要么是因为它将分割的记录数并不具有统计意义。

分类工具中列重要性的不同

这一部分描述 “列重要性”由 “证据”和 “决策表导入工具”所选择的重要性分级以及利用 “决策树”导入工具所选择的分割之间的差异。因为 “列重要性”使用全部数据，因此以下过程假定您在 “分类工具”模式中运行了导入工具 , 从而导入工具也就使用了全部数据。

离散过程

“列重要性”算法和 “证据导入工具”利用自动离散化算法 （同样的算法已应用于 “工具管理器”中的自动分组）将所有连续属性 （列）离散化。“决策树算法”事先并不产生属性离散化，而是在建成树时寻找阈值。

进行自动离散化的优势是它将连续的区间同时离散成几个间隔，而 “决策树”只能进行三叉分割。

“决策树”算法的主要优点在于使子集产生离散的数据块 （当检验完成时，符合特殊的节点的那些）。这样，与 “全局”离散化相对，该离散化过程对于那些记录来说是 “局部”的过程。

重要性方程

与排斥多向分割的 “决策树”，“选项树”和 “决策表导入工具”形成对照，“证据导入工具”和 “列重要性”算法根据作为纯度度量的 “交互信息”来对属性进行分级。这样如果您允许 “决策树”找出重要列，那么结果将倾向于具有较少值的属性而不是具有较多值的属性。“决策树”在默认情况下将变为 “交互信息”。

依照其它属性

证据导入工具独立为每个属性分级。如果几个属性高度相关，则它们有相同的级别。如果从 “列重要性”中使用 “高级选项”，并且在没有任何属性被选为重要的情况下 “... 计算改良纯度”（也就是说，移到右边的列表中），所显示的属性分级与由证据导入工具选择的分类次序相匹配。

“列重要性”算法、“决策树导入工具”以及 “决策表导入工具”都提供了比 “证据导入工具”更有力的重要性能力。而所有这些在选择一个重要分级时要考虑到其它列。

在列重要性之中，列被鉴别为重要与右边列表中的列有关系。如果两个列高度相关，并且其中一列已被选中，另外一列将有可能永远不被选中； 列之所以被选中是因为它可以提供比现有信息更多的有关标签的信息。

在 “决策表导入工具”中，建议按钮提供了与 “列重要性”算法类似的重要性工具。列被决定是否重要与映射框中已经出现的列有关。但是，在 “决策表导入工具”建议模式和 “列重要性”算法之间主要有三个不同。第一，“决策表导入工具”排斥多向分割。第二，“决策表导入工具”能够进行彻底地查询以找到更好的列顺序。最后，“决策表导入工具”并不列出任何仅仅用于对列分级的纯度。

“决策树导入工具”提供了更灵活的重要性分级方法，因为不同的列在不同的子树上可以被选中。例如，一列可以被选为根的左子节点而另一个为根的右子节点。以上这些适用于“决策树”，而对于选择一小部分列在“散点可视化工具”或“平伸可视化工具”中显示的情况则是不合适的。对于这些情况，“列重要性”算法是优越的，这是因为它建立了易忘记的“决策树”，在其中，任何层次的树在每个节点都检验相同的列。在“列重要性”条件下，为了以前所选列的组合必须选择某一列。

列

参见第 1 页 “添加列”，第 143 页 “删除列”，和第 15 页 “列名排序”。

命令行操作

每个可视化工具以及 MineSet 本身可以从命令行中启动，在提示符下输入工具名字，例如：

Windows 系统类型：

```
Viz [ 配置文件 ]
```

Irix 系统类型 （例如）：

```
scatterviz [ 配置文件 ]
```

指定配置文件 （*配置文件*）是可选的，但如果不指定，工具启动时仅仅带有可操作的“文件”和“帮助”菜单。您必须利用“文件”>“打开”来指定配置文件。

在 IRIX 中，不同的工具使用命令：clusterviz、eviviz、scatterviz、splatviz、statviz、treeviz 和 mapviz。关联规则使用第 20 页 “关联规则”中的两部分命令。

配置文件

对于每个工具， MineSet 需要有两个文件：

- 数据文件由制表符分隔的字段行组成，并具有 *.data* 扩展名。
- 配置文件描述了输入数据的格式以及如何显示它们。文件将工具的缩写作为扩展名，例如， eviviz、 scatterviz。

在利用工具管理器来指定那些您计划使用的分类工具或可视化工具的选项或参数时，您可以自动的创建配置文件。作为选择，您可以利用 ascii 文本编辑器为每个工具创建配置文件。例如，打开 Word Pad 或类似的文本编辑器并选择 “文件类型 : 所有文档 (*.*)” 您可以用这样的方式打开任何 *.schema* 和 *.data* 文件在 *《MineSet 3.0 企业版接口指南》* 可以找到更多的细节和样例。

混淆矩阵

混淆矩阵可以更加细致的描述由分类工具产生的误差。混淆矩阵给出了所产生的误差的*类型*，而不是简单的分析预测正误的数目。在 “工具管理器” 的 “挖掘工具” 选项卡环境下，对于所有的分类工具，“显示混淆矩阵” 是一个 “高级” 选项 （选择 “分类” 选项卡）。图 1-16 显示了在数据集上导入的 “决策树” 的混淆矩阵。

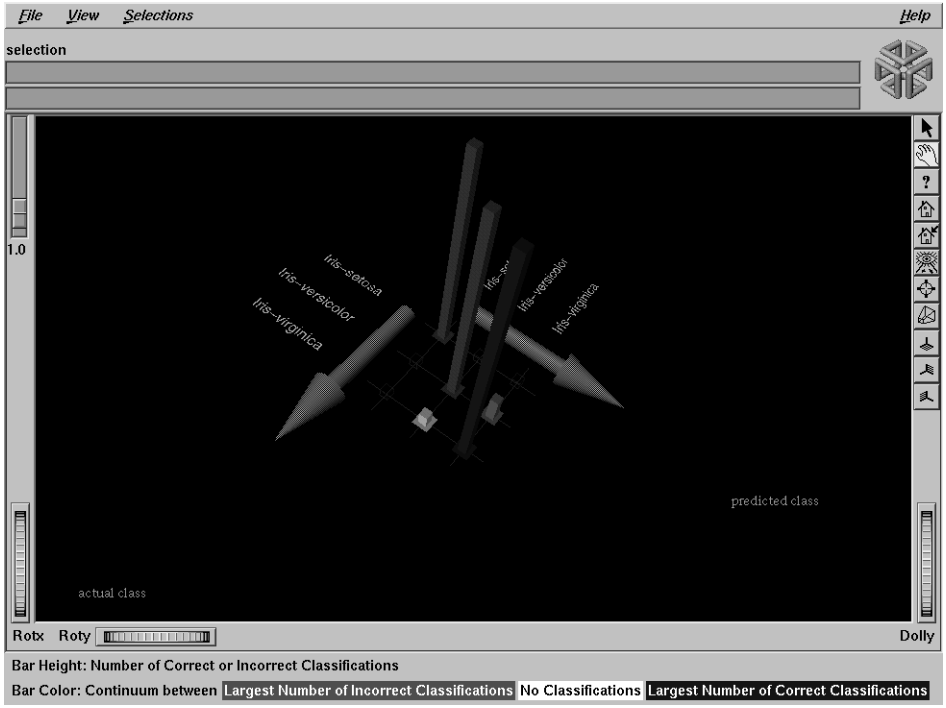

图 1-16 Iris 数据集的混淆矩阵

两个坐标轴代表：

- 由分类工具预测的类值，以及
- 在实验设置中给出的实际类值 （预留设置）。

对角线上的条目是正确的预测。非对角线条目指示不正确的预测。这种代表方式显示 *iris-versicolor* 和 *iris-virginica* 是经常会混淆的，然而 *iris-setosa* 的预测总是正确的。

当产生不同错误类型的代价不均匀时，使用损失矩阵影响产生的错误差通常是有很用的（见第 111 页 “损失矩阵”）。

注意： 混淆矩阵显示了测试集产生的误差；这样，如果数据的基本分布并没有明显地改变，它就代表了在实际情况下误差期望的真实分布。 MineSet 中的混淆矩阵在修正之前就进行计算，并且在是否应用修正的情况下都是一样的。（也可参见 第 33 页 “修正”。）

代价复杂性

代价复杂性是 CART （分类和回归树）中发展起来的高级的修剪方法选项。要寻找该修剪方法，从 “工具管理器” 中选择 “分类挖掘工具” 选项卡并选择 “决策树”。单击*高级选项*按钮可得到显示修剪选项的对话框。（您也可以选择 “回归挖掘工具” 选项卡并按照同样的方法操作。）

代价复杂度修剪通过互换树的误差率 （它的代价）和树上叶子的数目 （它的复杂度）试图产生形状优化的树。在代价复杂度计算过程中，修剪训练集就分割为学习集和修剪集。学习集被用于长成一个修剪树。该树被修剪后产生出一系列复杂性递减的树。修剪集然后被用于在序列中找出代价最小的树。最小代价树的大小被标出。学习和修剪集被合并并长成一棵树。该树然后被修剪为最小代价树的大小。

代价复杂度修剪参数允许您选择比最小代价树更小的树。该参数指出了标准差的值比可以接受的最小代价树所花费的代价更大。将该参数设为 0 以选择最小代价树；设置参数为 0.5 以选择最小的树，其误差率不大于 0.5 倍标准差，比最小代价树差。高值表示更多的修剪。如果数据包含了噪音 （误差和异常），增加值来创建更小的树。如果树被修剪成为一个节点，降低值以降低修剪量并更多地显示树的结构。

修剪过程比限制树高度或增加分割下限过程的速度要慢，这是因为先要创建整个树然后再进行修剪。然而，修剪过程是有选择地进行的，所以具有较低的误差率。

交叉验证

交叉验证是估计分类工具误差的一种方法，其过程为将数据集分割成 一定数量 （*k*）的子集 （通常为 10），然后建立同样数量的分类工具。这一过程可以重复多次来增加估计的可信度。导入工具被训练和检验了 *k* 次；每次都用所有的数据减去一个不同的子集进行训练，然后在预留的子集上进行检验。

“交叉验证”是任何导入工具都可用的 “估计误差选项”。对话框允许您在交叉验证中设置子集数以及过程重复次数。您也可以将随机子设置为不同的数，或保持为相同的数以确保总可以从相同点切入数据。

数据清洗

数据很可能表现为含糊的形式，或与其它存在的数据不相容。字段很可能遗漏或不再有效。将这样的复杂性清理为一种有序状况的过程被称为数据清洗，并且在很多情况下是数据挖掘操作的预处理过程。第三方提取工具可以支持该过程。

将数据从其它格式转化为 MineSet 格式，并做一些基本的数据清洗操作，可从 “工具管理器”菜单条中选择 “文件” > “导入数据”。参见 *《MineSet 3.0 企业版接口指南》* “导入数据文件”部分。

数据目标窗格

“工具管理器”中的 “数据目标”窗格提供了这些主要选项：“可视化工具”、“挖掘工具”和 “数据文件”，所有这些都是顶层选项卡。“可视化工具”提供了现有数据的直接可视化处理方法并显示相应的直观结果。“挖掘工具”创建数据模型 （伴随可视化）可以用于预测数据特征。“数据文件”用于将操作过的数据保存到文件。

数据文件选项卡

在 “工具管理器” 的 “数据目标” 窗格中利用 “数据文件” 选项卡可以将操作过的数据保存为将来可以在客户机和服务器上使用的数据文件。如果您单击了 “数据文件” 选项卡，就会出现带有两个切换按钮的面板，用它可以指定文件是保存在服务器上还是客户机上。选择的客户文件名出现在 “客户” 复选框的旁边。如果选择了 “客户”，*选择客户文件*按钮将带出一个对话框，用它可选定要保存的文件名称。当您选择 “服务器”，您可以将文件名直接输入临近 “服务器” 复选框的文本字段。

注意： 对于服务器文件，不允许有路径名；所有的文件保存在 “DataMove” 缓冲目录中。

数据导入

MineSet 提供数据导入实用工具，可从 “工具管理器文件” 菜单中实现 （“导入数据”）。“导入数据” 面板允许您选择想要导入的文件以及一些导入选项。参见 *《MineSet 3.0 企业版接口指南》* 来了解这些选项的描述以及所支持的数据格式列表。

数据转换面板

利用主 “工具管理器” 窗口左边的 “数据转换” 面板，您可以对将要工作的数据表格进行操作。在您利用 “文件” 菜单打开了一个表之后，表的列名出现在 “数据转换” 面板的*当前列*窗口中。所显示的选项的功能为：

- *删除列*可允许您删除与当前可视化处理或挖掘过程无关的一个或多个列。
- *分组列*可使您将每个记录分配给在一定列值区间内的组。例如，年龄列可以分为不同的区间：(0...18]，(19...25]，(26...35]，等等。"(" 指示了下限没有包含在区间之内。"]" 指示上限包含在区间内 （也可见 “分组”）。

- *组合*—创建了新列，它可以代表所选列的求和、平均、计数、最大或最小值。组合也能够从以其它分组列为索引的列中产生数组。应用组合操作产生的结果表通常比原始表的少很多行，因为新表中的每一行是原始表中几个行的组合 （参见第 4 页 “组合”）。
- *筛选*—可以让您根据包含列值的表达式来选择数据的子集，例如，留下那些年龄小于 20 的记录 （见 “筛选”）。
- *更改名称 / 类型*—可以使您改变列的名称以及类型 （参见 “更改名称 / 类型”）。
- *添加列*—可以让您根据数学表达式来添加一个新列 （参见 “添加列”）。例如，利用表达式根据 “年龄” 列来添加 “未成年人” 列，如果年龄小于或等于 18，那么未成年人列的值为 1； 否则未成年人列的值为 0。
- *应用模型*—可以让您利用以前创建的分类工具来标记新记录，估计标签值的概率，在新数据上测试分类工具或对存在的分类工具进行数据拟合 （见 “应用模型”）。
- *采样*—允许您选择数据的随机子集。对于很大的数据集这是有用的 （参见 “采样”）。
- *按顺序显示各列*—允许您按照字母顺序对列进行排序。在 Windows 系统中，该功能表现为复选框。
- *插件 操作*—只有当插件 API 可用时才出现。这些功能使用户可以象执行 MineSet 本身操作一样来访问插件操作。参见 MineSet 主页 （http://mineset.sgi.com）以及《MineSet 3.0 企业版接口指南》中的 “MineSet 插件功能”，可以找到更多的信息。

决策表

与决策树相似，决策表的结构具有等级之分，但却是利用一对属性而不是单一属性在每一层上将数据分解。“决策表导入工具”为数据分类指定了最重要的属性 （列），然后可视化工具将图形显示为块状的套合图。可视化结果中的每个块状图依次被分为更小的块状图以代表下一个最重要的属性。每个可视化结果都可以包含几个层，这些层代表重要程度递减的属性。更多的信息可参见 *《MineSet 3.0 for Windows 企业版用户指南》*中“利用决策表可视化工具”。

决策表中使用的分类方法与决策树的一样。通过选取样例所在区域的最主要的类来执行分类操作。如果被分类的记录落入无训练数据出现的区域，分类过程可以通过选择表等级中高一层的优势类来执行。

导入决策表

“决策表”分类工具可以从数据中被导入或自动产生。数据由记录以及与每个记录都相关的标签组成。（参见第 99 页 “导入工具”。）

“决策表”分类工具的自动导入是这样一个过程，其记录计数 （或更平常的情况下为记录权重）被用于计算表等级中每个节点上的概率。在可视化过程中，每个节点上记录的分布用一个块状图来表示。

所有连续属性被分为离散的组，分组原则是使得这些区间内的类分布尽可能的不同。区间的数目可以被自动确定。利用 “工具管理器”进行明确的分组，您可以更改任何属性自动分组。

沿坐标轴 （属性）上任意一行中块的数目显示了由导入工具产生的离散区间数目。如果只有一个区间，那就意味着该属性本身在预计标签时不起作用。初始时，标签的先验概率在 “标签概率”窗格中显示。先验概率是训练集中的各个类的比例。

在决策表等级中的每一层上有三种方法可以将属性映射到 X 和 Y 轴：手工、自动以及以特征搜索的方式生成。

启动决策表可视化工具

有几种方法可以启动 “决策表可视化” 工具：

1. 在 “工具管理器” 的 “分类” 选项卡中运行 “决策表导入” 工具。在导入工具建立分类工具之后，会自动调用 “决策表可视化工具”。
2. 利用 “工具管理器” 从 “可视化工具” 菜单中启动三维可视化工具，然后打开 *.dtableviz* 文件 （有关信息参见第 175 页 “工具管理器”中的 “工具管理器” 功能，该功能在 MineSet 工具中是通用的）。
3. 如果您知道要使用的配置文件，双击配置文件的图标。这样就启动了 “三维可视化工具” 并自动加载指定的配置文件。只有当配置文件名以 *.dtableviz 结束时才有效*。（所有由 “工具管理器” 为 “决策表可视化工具” 创建的配置文件都带有 *.dtableviz* 后缀。）
4. 在提示符下，通过输入下列命令可以从 IRIX 命令行中启动 “决策表可视化工具”：

   `dtableviz` [*filename.dtableviz*]

   其中文件名是可选的。如果您又指定配置文件，那么您必须使用 “文件” > “打开” 来选择一个文件。

离散标签

“离散标签” 菜单提供列的清单，这些列包含了数据集中的离散值 （单击*离散标签*旁边的箭头）。离散属性 （分组值、字符串值或整数）只有有限的值。您应该选择带有很少值的列作为 （最好是两到三个）标签属性 （参见训练集）。如果没有离散属性，菜单显示 “没有离散标签”，并且*运行*按钮也禁用。然后您必须利用 “工具管理器” 的 “数据转换” 面板加通过分组或添加列功能创建一个离散属性。

通过 “将列映射到坐标轴” 来 “查看数据”

您可以通过将列映射到 X 和 Y 轴来发现数据中属性之间的关系。在 “工具管理器”“数据转换” 窗格中的当前列面板中，选择一个列名，然后：

- 在 Windows 系统中，单击 X 列表或 Y 列表中的单元格，下拉式菜单会产生一个可用选择的列表。
- 在 IRIX 系统中，在 X 坐标轴或 Y 坐标轴窗口中选择一项。映射的头两列显示在最高层，接下来的映射按层次顺序依次下降。

如果您映射了奇数个属性，最后一层中显示了与奇数属性值相对应的单独的一列块状图。如果两个属性之间的作用并不确定，则将一个属性映射为 X 轴，而另外一个属性映射为 Y 轴。

可视化过程可能会交换某一层上的 X 和 Y 的映射关系，以求更好地保持整个块状图方阵的纵横比，但是它不会将属性移到不同的层上，除非属性已经被映射到该层。

如果您并不知道将哪个属性映射到哪个轴上，那么单击*建议*和*继续*按钮。

“建议” 有两种不同方法：使用或不使用特征子集查询。在深层导入工具选项面板中：

- 如果 “建议使用特征查询” 复选框不被选中，则 MineSet 运行 “列重要性” 进行搜寻。“列重要性” 在第 52 页 “列重要性”中做进一步描述。
- 如果 “建议使用特征查询” 复选框被选中， MineSet 先启动 “列重要性”，然后从 “列重要性” 推荐的顺序开始遍历所有可能的顺序。在每一阶段都会运行误差估计，并且对每个选项都进行了查看。可以预测，这将是一个冗长的过程，并且您可以随时单击*停止*终止该过程。特征查询持续的时间越长，分类工具的结果越准确。

解释决策表

每个类标签的*先验概率*被描述为“标签概率窗格”中的饼图，并出现在屏幕的右边。类标签的先验概率是，在忽略所有属性值的情况下，随机抽取记录时在数据中标签的概率。从数学上讲，该数为具有该类标签的记录数除以记录总数得到的比值。

左边“主窗口”中每个矩形块或块状图的*概率分布*显示了，在每个类中，属性值具有特殊组合的记录的比例。对于给定的数据这些概率是精确的。

默认情况下，沿坐标轴的标称属性值按照它们预测类的重要性进行了排序，这有助于找出重要的属性值。如果标签是被分组的属性，那么就会使用代表最高值的类。如果标签是标称性的，那么在先验概率饼图中占有较大份额的类就可用于确定顺序。当您选择了一个特定的类，并请求按照标签概率进行排序（通过选择“标称顺序”>“标签概率”），该类确定了标称值的顺序。通常，标称属性值可以按照字母顺序或权重顺序来排序，分组属性值总按它们的自然顺序显示。

当您细化下寻到一个矩形块中（利用鼠标右键），原块状图代表的数据重新显示在新的块状图方阵中，其中分配到细化层的属性被用作坐标轴。块状图放置在灰色基准块的顶部。如果基准块完全被块状图覆盖，这一层中的任意数值组合都有相应数据。通常情况下，基准块的大部分不会被覆盖。这就显示了没有数据存在的区域。在稀疏数据集中，有大量的区域是空的。点取或选择基准块与在细节层中逐层上寻的效果是一样的。在基准块上细化下寻致使每个在其顶部的块状图细化到下一层次上。

如果属性具有未知（空）值，那么该未知值由问号（?）标注。如果存在空值并且没有参与其余属性值的排序，则空值总显示为第一个值。利用“查看”>显示“空值”可以切换“空值”的显示。

“漫游”可以根据运行的平台是 Windows 还是 IRIX，以及您用的两键鼠标还是三键鼠标而有所不同，在这里描述了 Windows 平台下的两键类型。您可以从所在平台上的“特性选项”面板中改变按钮模式。

如果在左边窗格中选中 （单击鼠标左键）了一个块状图，那么在右边窗格中的概率饼图将显示相应的经验概率分布，该分布正好与块状图的显示相匹配。当选择了左边的多个块状图 （按下 Ctrl- 键的同时单击鼠标左键），那么右边的饼图将显示一个记录集的概率分布 （与标签有关），该记录集由所选块状图定义。也可以通过选择块状图下的基准块，选择整一组块状图。这与在层次结构中逐层上寻的效果相同。

类在右边饼图的下面列出，并按照份额大小为排序。概率最大的类在最上面。当选择了左边的值，这一顺序也响应变化的概率饼图。在当前选项设置下要预测的类显示在顶部，该类将被估计。如果标签是已被分组的属性，根据份额而定的顺序并没有变化。除此而外，如果标签是已分组的属性，颜色就按照连续谱系列分配：值最高的组是红色的；否则，就用随机颜色来代表。

要想看到更多细节信息，移动右边 “标签概率” 窗格下面的 “细节滑动条”。

决策表选项

选择 “高级选项”（Windows）或 “深层导入工具选项”（IRIX）弹出 “高级导入工具选项” 对话框。该对话框由四个面板组成：

- 顶部的面板指示了在 “工具管理器” 的 “数据目标面板” 中所做的选择。
- 从上边数第二个面板可让您设置损失矩阵和加权属性。参见第 111 页 “损失矩阵” 和第 20 页 “加权”。
- 左下角的面板可让您指定深层 “导入工具选项”，在后面的章节中描述。
- 右下角的面板可使您指定 “错误估计” 选项 （除非在 “数据目标” 面板中选择的是 “只用于分类工具” 模式，在这种情况下该面板是空的）。面板中显示的选项取决于您所选择的 “错误估计” 类型 （参见第 18 页 “应用模型” 和 第 80 页 “错误估计”）。

要微调 “决策表”导入算法，您可以改变下列 “决策表”导入工具选项。

- 最大尺寸

  当拥有大型数据集时，这一项特别有用。默认情况下，限制是 10,000 个节点 （这些节点与可视化过程中的块状图对应）。通过单击复选框以及敲入限制值可以限制大小。限制节点数目可以加快导入过程并节省存储空间。虽然限制规模会降低运行时间，但会增加错误率。如果最大尺寸小于显示所有数值组合的必需值，可视化过程显示的属性数目可能比已被映射的属性数目少。

- 最大属性数

  该选项确定了在查询时，可以允许多少不同的列加入到 “决策表”中。限制属性会简化结果并加快过程。这一限制仅仅影响由*建议*模式添加的列，手工添加的列并不受这一限制的影响。

- 每组的最小权重

  “决策表”导入工具将所有的连续属性离散化。该选项可使您定义每组中实例的最小数目。自动设置根据数据集的大小自动计算了这一数目 : 数据集越大，组也就越大。如果数据集非常大，您可以得到过多的离散区间，要想减少组数，就需要增加该值。

细化下寻和概化上寻

在可视化工具中，您可以在 “决策表”块状图中细化下寻或概化上寻看到更多或更少的细节 （参见第 80 页 “细化下寻和概化上寻”）。在块状图上单击鼠标右键就可以进行细化下寻。这样将以更细的块状图组来显示原来的数据，这些数据按下一层的一对属性进行了分解。如果光标放在基准块上 , 单击鼠标右键对基准上所有的块状图上细化下寻。如果光标放在背景上，单击将细化下寻或概化上寻全部所有的块状图。要进行概化上寻，在块状图、基准块或背景上按住 Ctrl 键并单击鼠标右键 （或使用鼠标中键）。

下拉式菜单

五个下拉式菜单可以让您访问 “决策表可视化工具” 的附加功能 : “文件”、“查看”、“名称排序”、“选项” 以及 “帮助”。如果您没有指定配置文件就启动了 “决策表可视化工具”，那么只有 “文件” 和 “帮助” 菜单是可用的。参见 “文件”、“帮助” 以及 “选项” 条目可以得到有关菜单的细节。

查看菜单

“查看” 菜单可使您控制 “主窗口” 中的一些显示的一些内容包括下面的设置：

- “筛选面板” 打开筛选面板根据所选的标准筛选数据的面板 （参见第 95 页 “筛选面板”）。
- “设置背景颜色” 可打开 “颜色选择器” 并让您选择一个新的背景颜色 （参见第 48 页 “颜色选择”）。
- “窗口控件” 可以隐藏或显示主窗口周围的外部控件。
- “空位置” 可切换空值所对应的显示内容。如果存在空值，则显示为第一个值，并与其它非空值略有偏移。
- “使用场景查看器”（或 “使用检查查看器”）可转换成可选的三维漫游模式。既然鼠标中键已用在 “场景” 查看器中的漫游功能上，那么要使用上寻功能必须在单击鼠标中键的同时按下 Ctrl 键。该选项在 Windows 并不支持。
- “显示为证据”（IRIX 上的证据模式）显示了每个块状图的条件概率，而不是最初根据记录权重得到的分布图。如果有一个或多个类很小的时候，这是很有用的。
- “工具条” 和 “状态条” 允许您显示或隐藏 “工具” 和 “状态条”。在 IRIX 上该选项不可用。

名称排序菜单

“名称排序”菜单可控制标称属性值如何排序，并提供了下列选项 :

- “按字母排序”使具有标称值的属性从左到右 （或从上到下）按字母顺序进行排序。
- “依据权重大小”使值从左到右进行排序，那些具有最大权重的记录在左面。
- “依据标签概率”（默认情况下）使标称属性值按照对应某一类的份额大小进行排序。如果标签是已分组的属性，默认情况下则使用最高的组。如果标签是标称的，那么在先验概率饼图中占有最大份额的类被视为默认类值。如果选择了特定类 , 并且随后就要求按照标签概率进行排序，那么被选中的类用于确定次序。在所有情况下，如果存在 “空值”，它就为第一个值。

决策树

“决策树”是预测模型，通过利用独立的或已知的属性值来帮助确定标签值或未知属性的值。预测标称值 （通常为字符串，例如 "Yes" 和 "No"）或仅可取较少值的属性值的任务是指分类过程。通过预测每个记录的标签，决策树可以对数据进行分类。用于分类的基本结构是决策树。一旦 “决策树导入工具”已经对数据进行了分类，则 “三维可视化工具”就会显示它的结构。

“决策树”方法显示了各种属性之间的相互作用，也就是说，属性值的组合是如何影响预测标签的。预测标签是指给定记录中未知的特征。在 “决策树”结构已知的条件下，数据在后序节点中的分布取决于在前一节点所做的决策。

在完整的可视化过程中的，决策树中每个节点上的条形图代表了各个标签值。您可以将光标放在条上来显示记录数目 （或权重）以及标签值的百分数。在每个节点的基准块处显示了到达它的记录数目 （或权重）。

创建决策树

“决策树”分类工具可从数据中自动导入 （或产生）。通过登录到服务器并以平常的方式选取一个数据集后就可以开始了。

从 “工具管理器”中选择*分类*选项卡，并且从*导入工具*弹出式菜单中选择 “决策树”。除非您希望，否则您不必做更多的限定。只需单击*继续*按钮。就可利用 “树可视化工具”来显示生成的决策树。在可视化过程结果中：

- 决策节点上的标签指定了在该节点上被检验的属性。
- 决策树中的叶节点指定了一个类。
- 基准块的颜色指示了子树的误差估计。
- 每个节点顶上的垂直条形图显示了节点上类的分布。

鼠标指到节点会显示以下信息 :

- *子树权重*—所指节点之下子树中的训练集记录的权重。该值映射为基准块的高度。
- *检验集误差 / 损失*—子树误差的估计 （或如果给出了损失矩阵，则为损失）。 +/- 后面的数是估计的标准差。标准差越高，误差估计越不准确。误差 / 损失估计和标准差在叶节点只包含很少记录或当检验集误差接近 0% 或 100% 的时候可信度较低。
- *检验集权重*—到达该节点的检验集中的记录权重 （如果没有设定权重则为记录的数目）。
- *纯度*—是介于 0 到 100 之间的数，指示了节点上的标签值分配的不均衡度。如果节点的记录来自于单一类，则纯度为 100。如果标签值具有相同的权重，纯度为 0。纯度是在修正之后进行计算的。

IRIX 中的并行过程

如果您已经安装了 MineSet 的多处理器版本，当分支包含了超过 1000 条以上的记录，在“决策树”中可以并行处理基于树的算法。（参见第 134 页 “在 IRIX 系统中的 “并行计算””。）您可以通过改变 “工具管理器”中 “特性”面板里的并行线程的数目来控制并行模式 （参见第 92 页 “文件菜单”）。只有在 IRIX 系统中该选项才可用。

深层导入工具选项

当修正可用并且在训练集 “显示为盘形图”的选项被选中时，垂直条形图显示了以盘形式出现的训练集的分布。盘的高度与条的高度比例一致。您应该期望盘形图的高度与“深层导入工具选项”中用于 “支持”比率的条高度值一样。

决策树选项

从 “工具管理器决策树”窗格中，选择*高级选项* （或*深层选项* IRIX）弹出 “分类工具高级选项”对话框。该对话框由四个面板组成：

- 顶部的面板指示了在 “工具管理器”的 “数据目标面板”中所做的选择。
- 从上边数第二个面板可让您设置损失矩阵和加权属性。参见第 111 页 “损失矩阵”和第 20 页 “加权”。
- 用左下边的面板可以指定深层 “导入工具”选项。
- 右下角的面板可使您指定 “误差估计”选项 （除非在 “数据目标”面板中选择的是“仅用于分类器”模式，在这种情况下区域是空的）。该面板中显示的选项取决于您所选择的 “误差估计”类型 （参见第 18 页 “应用模型”）。

要微调 “决策树”导入算法 , 您可以改变一些 “决策树”导入工具选项。“分类工具”选项对话框的 “导入工具选项”部分提供以下功能 :

- 限制树高

  默认情况下，在 “决策树”中的高度 （层的数目）并无限制。通过单击复选框并输入上限值以限制高度。限制层的数目可以加快导入过程，改善并行过程的负载平衡度并且不需要分解太多的节点有助于研究 “决策树”。但限制大小会增加误差率，设置该选项并不影响最大层之前各层属性的选择和分割过程。

- 拆分准则

  该选项提供了五个拆分准则供选择。下面是技术性定义。对于一个给定的问题，很难说哪种标准是最好的。可以全部进行试验，选择误差估计最低的或生成的 “决策树”最容易理解的标准。

  *公共信息*是指父节点和加权平均的子节点纯度之间的变化 （也就是说，*平均信息量的变化*）。加权平均是根据每个子节点拥有的记录数目来进行的。

  标准化公共信息 （默认情况下）是公共信息除以以 2 为底的子节点数目的对数。

  *增量比*是指在忽略标签值的条件下，公共信息除以分割的*熵*。“标准化公共信息”和“增量比”将优先权赋给了只带很少值的属性。

  *平方检测*应用 x^2 检测的统计独立性检验所有候选分割。然后选择导致标签值分支独立性最小的分割方法。

  Gini 是 CART （“分类和回归树”）中使用的拆分标准。与 “公共信息”相似，Gini 测量了父节点和加权平均子节点纯度之间的变化。与 “公共信息”不同的是，Gini 计算的节点纯度，减去了该节点处标签概率平方的和。

- 拆分下限

  这是一个权重下限，如果权重没有设定，就为记录的数目，而且必须至少有两个子节点达到该下限。该选项的默认值为 2。例如，如果节点中有一个三向分割，三个子节点中至少有两个其权重必须为 2 或更大（如果没有设置权重则为两条记录或更多）。这就提供了另外一种限制“决策树”大小以及提高运行速度的方法。

  提高拆分下限是为了增加概率估计的可靠度，因为每个叶节点处的记录数是巨大的。如果您预计数据包含了大量的噪音（误差或异常），或假如您利用树来估计概率（参见标题 2: 应用模型），将拆分下限增至 5 或更大。如果数据集很小（< 100 条记录），您可以减小该数到 1。

- 修剪

  “决策树”是在“树高”限制和“拆分”标准的约束之下建立的经过。统计检验一些子树还没有一个叶子节点更有意义，在这种情况下那些子树会被修剪掉。有三种修剪选项可以让您控制决策树的修剪操作：“置信度”、“代价复杂性”和“无”。

  修剪过程比限制树高度或增加分割下限过程的速度要慢，这是因为要先创建整个树然后再进行修剪。然而，有选择的进行修剪将形成更准确的分类工具。

  *置信度*是默认的修剪方法。高值指示修剪掉更多的部分；而低值会减少修剪。默认的修剪置信度 0.7 指示了应用于“决策树”的推荐修剪数量。如果数据包含了噪音（误差或异常），可增加该值来创建更小的树。最低的值为 0（不修剪）；没有上限。修剪参数为 0 的情况下，只有当单个节点的误差率至少与子树的一样低时，子树才被修剪。

*代价复杂性*将树的误差率 （它的代价）和树中叶节点数 （它的复杂度）进行交换，并允许您选择比最小代价树更小的树。将该参数设为 0 得到最小代价树； 设置参数为 0.5 得到最小的树，其误差率不大于 0.5，标准差比最小代价树差。默认的设置为 0，选择最小的树，代价也最小。高值指示更多的修剪。如果数据包含了许多噪音（误差和异常），增加该值来创建更小的树。如果树被修剪成为一个节点，降低该值以降低修剪量并显示更多的树的结构。参见第 60 页 “代价复杂性”可得到更多的信息。

*无*不执行修剪。虽然这会产生对于训练数据来说较为冗余的树，导致试验数据较高的分类误差，但它却允许您调查决策树的完整结构。

- 允许一次性拆分

  单击复选框可以使导入工具在拥有超过两个值的标称属性上产生两向分割。通常情况下，标称属性上的分割产生的分支与它们拥有的值一样多 （参见第 72 页 “创建决策树”）。例如，属性 “颜色”上的分割会产生红色、绿色、黄色以及蓝色的分支。如果一次性拆分功能可用，那么导入工具也可以产生只有两个分支的分割，例如红色和非红色。一次性拆分正好孤立了列拥有值中的一个。在数据中包含具有很多可能值的属性，而其中一些值可以很好的区分标签时，这种类型的分割是很有用的。

- 推进

  “推进”用于提高分类准确性，虽然它是一个耗时的过程。在推进过程中不执行可视化过程。

  当执行过程继续时，状态窗口显示了估计误差。估计误差是该算法重复将新的权重分布在训练集上以及在重新加权的数据集上导入分类工具的结果。参见第 37 页 “推进 （boosting）”可得到更多的信息。

搜索和筛选面板

在 “查看”菜单中选择 “搜索面板”和 “筛选面板”（IRIX 为 “显示”菜单）会产生一个对话框，您可指定搜索 / 筛选对象的标准。对于决策树该项选择总是一样的。这些在下面描述。

搜索或筛选可以被限制在特殊标签上进行，或者通过选择类列表中的值 （在顶层窗格中）或通过使用类项 （在底部的窗格中），这些都提供更强大的比较操作 （例如匹配）。下滑滑动条可以得到其它标准。下面描述了其它项：

- *子树权重*可以限制对子树中那些带有给定权重 （如果没有设置权重则为记录数）的条形图或基准块的查询或筛选操作 （取决于是选择了条形图还是基准块单选钮）。您就可以限制对权重不低于 50 的条形图进行查询操作。
- *检验属性*可以将搜索或筛选操作限制在正在检验的具有给定标签值的节点上。注意，决策节点标签代表了检验属性，而叶节点标签表示了预测标签。例如，如果选择了*包含年龄的检验属性*，就只考虑那些检验了年龄值的节点。
- *检验值*可以将搜索或筛选操作限制在带有指定收入线的节点上。
- *百分数* 可以将搜索或筛选操作限制在权重占整个节点权重达到指定百分数的条形图上。例如，您想找到所有对于一个指定类所占权重大于 80% 的节点。那么单击该类标签并选择 *百分比 >80*。如果您选择了基准块而不是条形图 （基准块的值为 0）则设置这一项是毫无意义的。
- *纯度* 可以将搜索或筛选操作限制在给定纯度范围的节点上。例如，如果想要看纯度节点 （有一个类占优势），您可以选择*纯度 >90*。
- *检验集的子树权重* 可以通过给定检验集权重限制对子树的查询或筛选操作 （如果没有设置权重则为检验集记录的数目）。
- *检验集的误差 / 损失*可以通过给定估计误差或损失范围限制对节点进行的查询或筛选操作。
- *均值误差 / 损失标准偏差* 可以通过限定标准差范围限制对节点进行的查询或筛选操作。
- *层* 可以限制对指定层或层范围进行查询或筛选操作。例如，只查询前五个层。

对于决策树，以下选项较少用到。

- *等级* 找出所有的节点和线，而这些节点和线与从根生出的路径上的已知值相匹配。然后标记出这些节点的子节点。
- *将空值看作 0* 在决策树导入工具中不使用，这是因为不会对决策树产生空值。

一旦搜索操作结束，黄色聚光灯就会加亮与搜索标准匹配的对象。要显示黄色聚光灯下的对象信息，将箭头移到聚光灯上，则信息会显示在左上角，标签 “指针位于” 之下；在黄聚光灯下选择并缩放对象，用鼠标左键单击聚光灯即可；如果您在单击的同时按下了 Shift 键，就不会进行缩放操作。

一旦完成了筛选操作，则画面只会显示与筛选标准匹配的节点。

离散标签菜单

“工具管理器” 中 “分类” 选项卡上的 “离散标签” 菜单提供了可能存在的离散标签的列表。离散属性 （分组值、字符串值或几个整数）只有有限的值。选择只带有很少值（例如，两或三个）的标签是很明智的。如果没有离散属性，菜单显示没有离散标签，并且*继续*按钮也禁用。然后您必须，利用 “工具管理器” 的 “数据转换” 面板并借助分组或添加列功能创建一个离散属性。

追溯

在当前可视化过程中观察与选定内容所对应的原始数据通常是非常有用的。这就是 *追溯*。通过在可视化过程中选择对象，并将下层的数据送入工具管理器，您就可以查看原始数据或利用其它工具来进行分析。 MineSet 中的所有工具都具有这样的功能，并且某些工具还具有该条目结尾所说明的附加特征。

在可视化过程 “选项”下拉式菜单中有两项可以执行追溯：

- 发送到工具管理器

  当选择了该选项，产生可视化过程的操作历史被置入 “工具管理器”。与可视化工具中用户选择相对应的筛选操作被加入到历史中，该筛选操作被尽早的插入操作历史中，下面给出了具体的限制。

- 显示原始数据

  与 “发送到工具管理器”过程相似，该选项从产生可视化过程的操作历史开始，并将筛选操作尽早添加到历史中，所有非筛选操作在历史中的筛选操作被删除以后才进入历史。新历史用于生成在 “记录查看器”中显示的表。如果筛选操作能够被放在历史的开始位置，则显示的数据是原始记录； 否则，就会产生警告信息，指示数据不完全是原始数据。工具管理器的状态不会改变 （除非当前没有运行）。

在另一种情况下，“工具管理器”执行该操作。如果 “工具管理器”没有运行 , 则它自动启动。

只有用 “工具管理器”产生的可视化过程才可用于追溯操作。这些可视化过程的每个 *.schema* 文件都包含了*历史*部分并用来通知 “工具管理器”文件是怎样生成的。几个特殊目的的挖掘可视化过程，例如 “学习曲线”、“混淆矩阵”以及 “上升曲线”并不包括在历史中，并且也不支持追溯。

当选择了用于追溯的对象，您就在可视化表的基础上隐含地指定了筛选语句。如果在可视化之前表已经转换，“工具管理器”也许不得不改变筛选操作将它更早放入历史中。例如，如果该筛选操作是建立在已分组列的基础之上，“工具管理器”必须将其改变来引用分组之前的列。

因为 “工具管理器”不能为所有的操作调节筛选过程，因此筛选过程就不能总放到历史的开始。特殊情况下，如果筛选过程引用了由*添加列*，*组合*或 *应用模型*创建的列，则筛选过程在历史中的位置不会比创建列的操作更早。更重要的是，“工具管理器”不能将筛选操作过程移到存在的采样操作之前，因为那样会戏剧性地改变采样结果。

“显示原始数据”模式力图削减历史以便在尽可能早的阶段显示记录。然而，要防止该模式过多的显示并未选择的记录，削减操作就不能删除任何其它存在的筛选过程 （或者是由用户创建或者是以前追溯产生）。

细化下寻和概化上寻

细化下寻或概化上寻是指那些在实体上单击来观察更多或更少细节的操作。只有在 “地图可视化工具”和 “决策表可视化工具”中才可用。

细化下寻是用鼠标右键单击对象。概化上寻操作是在按下 Ctrl 键的同时用鼠标右键 （或用鼠标中键）单击对象。在背景上进行的概化上寻或细化下寻操作会在全局范围内执行。

错误估计

当创建了分类工具，如果能够知道该工具将来的执行情况是非常有用的 （也就是分类工具误差率代表的意义）。影响分类误差率的因素包括：

- 训练集中的可用记录数目

  因为导入工具必须从训练集中学习，所以训练集越大，分类工具的可靠程度越高；但是训练集越大，导入工具建立分类工具所花的时间越长。随着训练集的规模增加，误差率的改进也随之下降 （这就是一种回报效应的减少）。

- 属性数目

  更多的属性意味着导入工具要计算更多的组合、产生更困难的问题以及花费更多的时间。注意，有时随机相关会使导入工具误入歧途，从而可能建立准确性很低的分类工具 （从技术上说这被称为冗余）。如果属性与任务并不相关，则将它从训练集中删除 （可以用 “工具管理器”实现）。

- 属性中的信息

  有时属性中并没有足够的信息来预测标签并保证较低的误差率 （例如，根据眼睛的颜色试图确定某人的工资）。添加其它属性 （例如，专业，每星期工作的小时数以及年龄）也许能够降低误差率。

- 未来并未标记记录的分布

  如果未来的记录产生的分布基础与训练集中的不同，误差率很可能会很高。例如，如果您从包含家庭轿车的训练集中建立了分类工具，在试图对包含许多赛车的记录进行分类时该分类工具也许并不适用，因为属性值的分布可能差异性很大。

下面描述了分类工具中用于估计误差率的两种普通方法。两种方法都假定未来的记录将从具有相同分布的训练集中抽取得来。

- 预留

  一部分记录 （通常为三分之二）用作训练集，而剩下的记录为*检验集*。导入工具只显示了数据的三分之二并建立分类工具。然后，导入的分类工具对检验集进行分类，并且检验集中的误差率或损失就是估计误差率或估计损失。图 1-17 显示了这种误差估计方法。

  当修正被启用并且 “用盘显示训练集”的选项被选中时，垂直条形图显示了以盘形式出现的训练集的分布。盘的高度与条的高度比例一致。您应该期望盘在条形图上的高度与 “深层导入工具”选项中用于 “预留”模式的比率的值一致的高度。

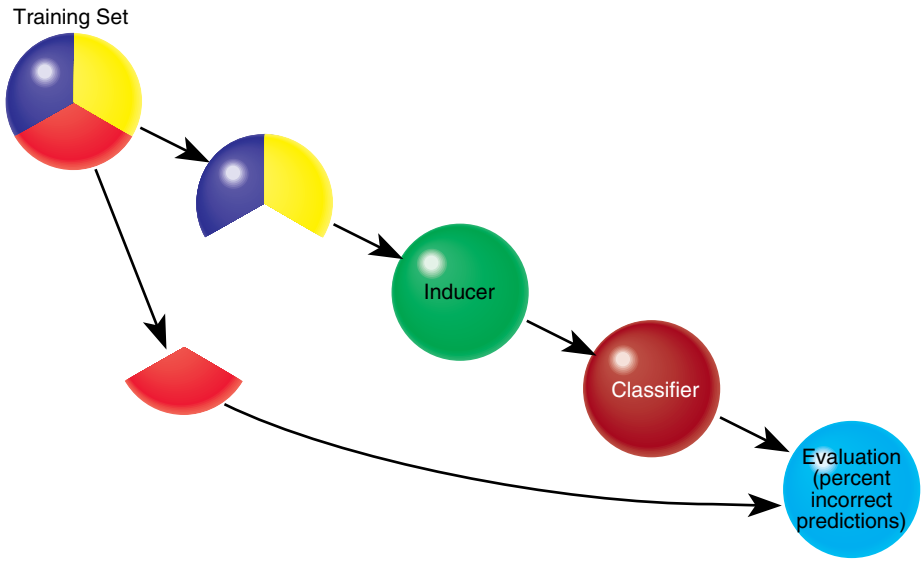

图 1-17 估计分类工具的准确性

这种方法虽然很快，但是因为它仅仅使用了数据的三分之二来创建分类工具，因此就没有充分地利用数据来进行学习。如果利用了所有的数据，那么就有可能建立一个更为准确的分类工具。

- 交叉检验

  数据被分为 *k* 个互不重叠的大小相近的子集。导入工具训练和检验了 *k* 次；每次都用所有的数据减去一个不同的子集进行训练，然后在预留子集上进行检验。估计误差率则是所得到误差的平均。图 1-18 显示了 *k*=3 的交叉检验 （注意默认值为 *k*=10）。

  交叉检验可以重复很多次 （*t*）。如果进行了 *t* 次，那么 *k*- 子集交叉检验就会有 *k***t* 个分类工具产生并被评估。这也就意味着交叉检验的次数是 *k***t* 次。默认情况下，*k*=10 并且 *t*=1，所以交叉检验的时间是创建单独分类工具 10 倍左右。

  增加重复的次数 （*t*）也就增加了运行的时间，但是改善了误差的估计和相应的置信区间。

您可以增加或减少 k。减少到 3 或 5 就会缩短运行时间；但是估计就会因为训练集规模的变小而产生偏差。可以增加 k，但只推荐用于较小的数据集。

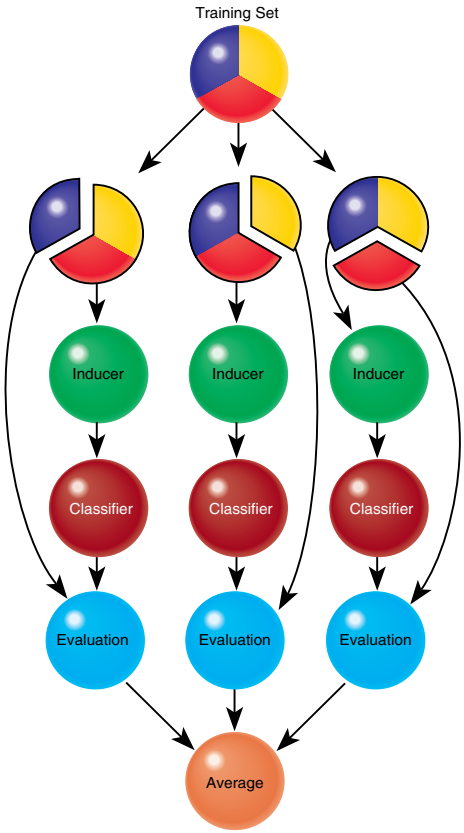

图 1-18 分类工具的交叉检验 （k=3）

通常情况下，预留估计应该在开发阶段使用，并且数据集应有超过 5000 以上的记录。交叉检验应该用于分类工具的最后建立阶段，并且数据集的规模较小。

修正是将整个数据集应用于只利用训练集中的数据所建立起来的分类工具或模型。当利用预留误差估计时，您剩下了一部分数据用于检验。当您通过分类工具结构修正所有数据时，您可以降低最终分类工具的误差，因为计数、权重和概率反映的是整个数据集的信息。详细讨论，参见第 33 页 “修正”。

证据模型

“证据分类工具”模型将数据集中的每条记录分配给一类。“证据可视化工具”显示了证据模型的结构。可视化工具有助于理解分类过程中特定属性值的重要性。同样，它也可直观的表现于分类过程，以及回答 “What if”问题。参见 *《MineSet 3.0 for Windows 企业版用户指南》* 可以得到如何使用和解释 “证据可视化工具”的详细信息。

证据导入工具

证据导入工具建立了一个模型，该模型假定对分类来讲各属性值的概率是独立的。例如， *iris* 数据集假定四个属性 （萼片长度、萼片宽度、花瓣长度以及花瓣宽度）对于每个 iris 类（iris-setosa、iris-versicolor 和 iris-virginica）来说是独立的。当然这个简化的模型并不太真实，但这个模型用于最初的数据开发却非常有效，并且它的分类预测性能在实际应用过程中也很有用。

每个属性值或值的区间 （对于分组的连续属性）定义在一张图表上，并依次给出了每个类标签的条件概率。对已知的记录进行分类，一种做法是计算每个类的概率，具体方法是将其先验概率乘以矩阵中每一行的相应条件概率。最终的乘积给出了每个类的相对概率，并且得出最高的值就对应预计类。如果属性的值未知，则它被忽略。这些 “空”值由那些位置略有偏差的图所代表，并且当您从 MineSet 提供的样例中调出 *iris.schema* 文件时这些值会出现在左边。参见附录 A，配置和数据文件样例可得到路径名，或参照 *《MineSet for windows 3.0 企业版用户指南》* 中提供的信息。

当您用 “证据导入工具”显示了 iris 数据集时，每个类标签的*先验概率*由屏幕右侧 “标签概率”窗格中的圆饼图描述。类标签的先验概率是，在忽略所有属性值的情况下，随机抽取记录时在数据中见到该标签的概率。从数学上讲，该数为具有类标签的记录除以记录总数的比值。在 “主窗口”中 “证据”（块状图）和 “概率”（饼图）之间的切换可以通过单击 “证据”标签或通过在 “查看”菜单中切换 “证据模式”。

条件概率，由屏幕左侧 “主窗口”中的块状图来表示，显示了每个标签值条件下的每个属性值之间的相对概率。块状图中份额的大小指示了，在考虑了记录中的已知属性值之后分类工具添加到先验概率中的证据数量。如果份额的大小一样，那么值就是不相关的，并且分类工具向所有的类中添加了相同数量的证据。

从技术上看，图中的份额代表了在给定类标签 L 的条件上，属性值 A 的标准化条件概率。条件概率 P(A|L)，是指从带有标签 L 的记录中随机抽取的记录 采取 A 值的 概率。在默认设置下，根据记录权重可计算概率。例如， P （0.75 < 花瓣宽度 < 1.65 | iris-versicolor）为 91.6，因为带有 iris-versicolor 标签的记录有 36 个，其中 33 个的花瓣宽度在这个区间内。

条形图 / 证据模式中基准块的高度显示了带有所选标签的证据。整个支持的证据值通过 -log[1-P(L)/sum(P(Li))] 计算得出，而反对证据值由 -log[P(L)/sum(P(Li))] 计算得到。对最小证据的删减影响了基准块并同样影响了条形图。

*重要性*是关于标签的一种对预测能力的度量。“主窗口”提供了有价值的直观结果，这不仅在于每个属性值的重要性会影响类值，而且还在于特定属性值的重要性。

导入证据分类工具

“证据”分类工具的自动导入过程中用记数 （或权重）来计算概率。“证据”分类工具自动从数据中导入 （产生）。由记录和与每个记录相关的标签所组成的数据集被称为*训练集*。

用以下方法可以得到概率 :

1. 所有连续属性被离散化 （分组），分组原则是这些组内的类分布尽可能的不同。如果系统中安装了多处理器版本的 MineSet，那么区间的数目被自动确定，这样就可以调用并行过程，进行并行分组。任何属性的自动分组过程可以通过执行 “工具管理器”中的分组操作而被覆盖。
2. 先验概率是训练集中每个类比例。
3. 条件概率是指，在训练集中每个类标签出现的条件下每个属性值的概率。（该图显示的属性值已进行了标准化。）

每行中块图的数目就是由导入工具产生的离散区间的数目。如果只有一个区间，那就意味着该属性本身在预计标签时不起作用。初始时，标签的先验概率在 “标签概率”窗格中显示。

每个属性值或值的区间 （对于分组的连续属性），正好定义了一个块图，依次给出了每个类标签的条件概率。

- 拉普拉斯校正

  这将会使概率向均值靠拢，从而避免了极端数 （例如 0 和 1）。如果在 “证据导入工具”对话框的*高级选项*中 （或*深层导入工具选项*IRIX 系统）选择了拉普拉斯校正选项，并且将因子设为空或 0 后，就可以应用因子为 "1/t 训练集权重 " 的启发式过程自动进行拉普拉斯校正。也可参见第 10 页 “拉普拉斯校正”。

- 每组的最小权重

  “证据导入工具”将所有的连续属性离散化。该选项可使您定义每组中实例的最小数目。自动设置是一个可根据数据集规模来设置该数值的启发式过程 : 数据集越大，从总体上说，每个分组中所允许的最小记录数目就越大，并且分组的宽度就越小。如果数据集非常大，您可能得到过多的离散区间，要想减少组数，就应增加该值。

- 自动选择列

  该选项提供一个进程，只选择那些最有助于预测的列。因为有一些列可能降低证据分类工具的估计准确性，所以该过程可自动查询有用的列子集。只有那些被该过程判为有用的列被使用。这一过程要花费很长的时间，尤其是在有很多列的情况下。删除那些可能降低准确性的高度相关列是很有用的。自动选择列和 “推进”不能同时使用，否则将会消耗很长的时间来完成。

自动选择列执行了一个查询过程，该过程用于查询可降低分类工具误差的列的最优集。这些列的选择过程是这样实现的，通过利用包装器靠近过程来估计不同属性集的误差。每个属性子集通过利用交叉检验估计分类工具误差的手段来进行评估。添加或删除列的根据是，利用 “最好 - 第一”查询机制得到的误差估计。默认的模式下，查询从空的属性集开始。通过选择*向后*选项，查询从全套设定的选项开始，这是比较慢的，因为更大的模型得从头建立 （参见 *《MineSet 3.0 企业版接口指南》*中的 “可视化工具的数据文件格式”）。

启动证据可视化工具

有几种方法可以启动 “证据可视化工具”：

- 从 “工具管理器”中的 “分类”选项卡上运行 “证据导入工具”。在导入工具建立分类工具之后，会自动调用 “证据可视化工具”。下面的内容是使用与 “证据可视化工具”结合在一起的 “工具管理器”。
- 用 “工具管理器”从 “可视化工具”菜单中启动 “三维可视化工具”。
- 如果您知道要使用的配置文件，双击配置文件的菱形图标。这样就启动了 “三维可视化工具”并自动加载指定的配置文件。只有当配置文件以 *.eviviz* 结尾时才有效（用 “工具管理器”为 “证据可视化工具”创建的配置文件总是这样）。
- 从 IRIX 命令行中启动 “证据可视化工具”，可以输入：

  `eviviz` [*configFile*]

  *配置文件*是可选的，它指定了所使用的配置文件的名称。如果未指定配置文件，那么您必须用 “文件” > “打开”来指定一个。

调用证据可视化工具的 IRIX 选项

-quiet 选项删除弹出的以指示进程的对话框。通过在您的 *.Xdefaults* 文件中加入一行代码可以使该选项永久可用

```
*minesetQuiet:TRUE
```

也可参见第 19 页 “警告选项”。

在 Windows 上，该选项在 “三维可视化工具特性”面板中是可用的。

证据导入工具选项

选择 “高级选项”（IRIX 上为选择 “深层导入工具选项”）可生成 “导入工具选项”对话框。该对话框由三个面板组成 :

- 顶部的面板显示了在 “工具管理器”的 “数据目标”面板中所做的选择。“误差估计”的类型由该模型限定。
- 左下角的面板可让您指定 “深层导入工具”选项 （描述如下）。
- 右下角的面板可使您指定 “误差估计”选项 （除非在 “数据目标”面板中选择的是 “仅用于分类器”模式，在这种情况下该区域内是空的）。该面板中显示的选项取决于您所选择的 “误差估计”类型 （参见第 80 页 “错误估计”）。

如果在 “高级导入工具”选项中指定了 “损失矩阵”，标记为 *使用损失矩阵*的按钮就会出现在概率饼图的右下角。该项选定后 （默认情况下）“损失矩阵”用于调节概率的显示。最大的薄片所显示的是考虑了 “损失矩阵”的类。要想看到不使用 “损失矩阵”时概率的状态，取消选择 *使用损失矩阵*按钮。灰薄片指示了存在一个经过编辑用于预计空值的列。如果灰薄片是最大的薄片，则分类工具预测为空值。要得到更多的信息，参见第 111 页 “损失矩阵”。

在 “证据可视化工具主窗口”中 “选择对象”

在选择模式中，光标表现为箭头。您可以加亮对象 （饼图或条），方法是将光标移动到该对象上。关于该对象的信息就会在 “主窗口”的上面显示。只要光标停留在对象上，信息就一直显示。

- 如果对象是一个饼图，那么消息将采用这样的格式 :

  < 属性名字 >: < 值或区间 >
  权重 = < 权重 >

  这里，*权重*是落入区间或具有该属性值的所有数据点的权重和。图高与该数成比例。除非使用了记录加权，否则权重就代表了记录计数。

- 如果对象是条形图，那么消息将采用这样的格式 :
  (< 属性 > = < 值 >) ==> Prob (< 选择的标签 >) = x% [low%-high%]
  证据 =z
  < 选择的标签 > ==> Prob (< 属性 > = < 值 >) = y% [low%-high%]
  权重 = < 权重 >

  这里， *x* 是在记录具有加亮属性值的条件之上具有所选标签值的概率。方括号的范围， [low%-high%] 给出了 95% 的置信区间。同样地， y% 是在记录具有加亮属性值的条件之上具有所选标签值的概率。条的高度代表了证据而非概率。证据的数量 z 与条的高度直接相关。为了确定哪个类被估测了，证据可以进行求和计算 （不象概率，必须相乘）。*权重* 是具有该值的数据点的总权重。

  从技术角度讲，*支持证据*被定义为

$$-\log\left[1-\frac{P(A\mid L)}{\sum_{i=1}^{N}P(A\mid L_i)}\right]$$

  而*反对证据*被定义为

$$-\log\left[\frac{P(A\mid L)}{\sum_{i=1}^{N}P(A\mid L_i)}\right]$$

A 是属性值， L 是所选的标签值，而 N 是标签值的数目。当计算条高度时，上述括号中的表达式中加上了一个很小的数以阻止条无限制的增高。“主窗口”中的“支持”或“反对”的周围有一个框说明可以进行单击操作。在框上单击可以切换所代表的意义。

灰色矩形基准块的高度 （上面有直立着的条形图）代表由先验概率贡献的证据的数目。例如，如果标签是轿车的汽缸，很少有三汽缸的轿车，所以当*支持证据*显示时基准块很低，而当*反对证据*显示时，基准块很高。您可以将顶部的单独条形图的高度加入这一高度。

*支持证据*在确定预测特定标记值的过程中什么值最有效的时候比较用处。

证据的数量 （条形图高度）不是直接从高亮度显示的概率中派生出来的。相反，证据取决于相对于其它根据上述等式得到的其它标签值概率的条件概率。

您也可用以下的方式从属性行中选择任意数量的值，即当光标位于属性值之一的上方时单击鼠标左键。右边 “标签概率窗格”中的大饼图改变为反映您所选的项；它现在显示的是在刚选中属性值的条件下的后验概率。类保持了顺序排列，所以与最大薄片对应的一个类出现在右边列表中的顶部。通过一起乘以每个属性的条件概率，“证据可视化工具”形成了新的后验概率分布，然后将该结果乘以先验概率并标准化为 1。

这种乘法操作与条件独立性假设对应。当违反了该假设，并且出现几个属性实例时，预测类的概率就很可能是错误的 （即使最终的分类是正确的）。当您运行导入工具时，显示在 “状态”窗口中的估计误差有助于您判定这一假设的合理性。如果误差率 / 损失较低，则该假设表现出适当的鲁棒性。

当在 “标签概率”窗格中选择了特定标签 , 对于属性的每个值 “主窗口”显示了条而不是块状图或饼图。条上的标题是 “支持证据”，“支持”周围的框指示了它可以被选择。

单击 “支持证据”中的 “支持”，标题切换显示为 “反对”。其结果是，条形图高度改变成显示反对标签的证据。

选择条形图而对右边 “标签概率窗格” 中的大概率饼图所起的作用与选择块状图或饼图所起的一样。条形图的高度指示了由所选值贡献下的支持或反对选定标签的证据数目。因为概率的对数用于代表证据，所以条形图高度就被添加到累计证据中 （反之，就必须乘以概率）。

证据可视化工具菜单

五个下拉式菜单可以让您使用 “证据可视化工具” 的附加功能 : “文件”、“查看”、“选择”、“名称排序” 以及 “帮助”。如果没有指定配置文件就启动了 “证据可视化工具”，那么只有 “文件” 和 “帮助” 菜单可用。参见 “文件”，“帮助” 条目可以得到有关菜单的细节。

查看菜单

利用 “查看” 菜单可以控制 “证据可视化” 窗格中的显示设置 : 普通菜单项的信息参见第 185 页 “查看菜单”。根据系统平台，菜单也包含了一些附加选项：

- *显示为证据*切换 “证据” 的显示。如果选择了的项，画面左侧显示证据。如果没有选取该项，画面显示概率。
- *按重要性排序*展示了属性的排序，这种排序是根据涉及所选标签的分类中属性的用处来进行的。如果该选项被关闭，那么属性将会显示出与 “工具管理器” “当前列” 中相同的顺序。
- *减去最小证据*当选择了标签并且显示了条形图的时候才应用。该选项选定时 （默认情况下），减去了整个标签值中最小值对应的条形图的高度。每个属性中各个值不同数量也不同，但是对于已知的属性值，所减去的数量对应标签值是固定的。激活该选项会因为对所有标签值都减去了相同的量而放大一些小的差别。
- *使用拉普拉斯校正*切换到使用 “拉普拉斯校正”。如果在深层导入工具选项对话框中指定了拉普拉斯校正，则就利用了该值，否则提供默认值。
- *利用景观查看器*（或*利用检查查看器*）可转换成三维漫游的可选模式。

名称排序按钮

“名称排序”菜单可控制标称属性值如何排序。有关选择的信息可参见第 129 页 “标称排序菜单”。

选择菜单

“选择”菜单可以对基本数据进行追溯。要进行追溯，首先选择值和 / 或类的某种组合，然后在两种追溯方法中选择一个对基本记录进行追溯。对于几个工具该菜单是相似的，参见第 153 页 ““选项”菜单”可得到菜单项的解释。

文件菜单

绝大多数可视化工具的 “文件”菜单是相似的，并包含一些选项 :

Windows 系统

- *打开* 装载并打开一个配置文件，并在主窗口中显示。以前显示的数据被清除。使用 *打开* 来查看一个新的数据集，或在改变配置之后查看相同的数据集。
- *重新打开* 重新打开当前已打开的文件。在配置文件或数据文件更新之后可以使用该功能。
- *保存图象*将当前可视化工具窗口中的内容保存到图象文件内，并可以选择是保存整个窗口，包括所有的符号，或者只是保存主要的图形对象画面 （默认情况下是整个窗口）。
- *打印图象*将当前窗口中的内容输出给打印机。您可以利用 “打印”对话面板 （默认情况下为系统的默认打印机）来指定输出打印机，并且可以象 *另存为* 对话框一样，选择是打印全部窗口还是只打印主要画面窗口。
- *打印预览*显示了潜在的打印输出。
- *打印设置* 打开一个对话面板来设置打印输出 （默认情况下是系统的默认打印机）。
- *网上发布* 将当前文件打印到网上兼容文件中去。

- *特性*启动 “特性”对话框来指定鼠标映射并在命令执行过程中切换警告信息。在这里还可以指定声音效果、动画播放以及默认的可视化工具字体大小等。
- *启动工具管理器*启动 “工具管理器”（如果还没有运行），并恢复其在调用 “树可视化工具”时的状态。
- *近期文件*打开任意可视化工具使用过的近期文件。菜单列出了最近打开的四个文件。
- *退出*关闭所有的窗口并退出应用程序。

IRIX 系统

- *打开*为并打开并 装载一个配置文件，并在主窗口中显示。以前显示的数据被清除。使用*打开* 来查看一个新的数据集，或在改变配置之后查看同一数据集。
- *重新打开* 重新打开当前已打开的文件。在配置文件或数据文件更新之后可以使用该功能。
- *另存为* 将当前可视化工具窗口中的内容保存到图象文件中去。用户指定了文件名称（对于 “树可视化工具”，默认为 *treeviz.rgb*），格式 （默认为 *rgb*），并可以选择是保存整个窗口、包括所有的图例，或者只是保存主要的图形对象画面 （默认情况下是整个窗口）。
- *打印图象*将当前 “可视化工具”窗口中的内容输出给打印机。您可以利用 “打印”对话面板 （默认情况下为系统的默认打印机）来指定输出打印机，并且可以象*另存为* 对话框一样，选择是打印全部窗口还是只打印主要画面窗口。
- *网上发布*将可视化工具保存为 *.mtr* 文件，该文件适合于网上发布。
- *启动工具管理器*启动 “工具管理器”（如果还没有运行），并恢复其在调用 “树可视化工具”时的状态。
- *退出*关闭所有的窗口并退出应用程序。

文件需求

大多数 MineSet 可视化工具除配置文件以外至少需要两个文件 :.*data* 文件和 *.schema* 文件，当您利用 “工具管理器” 调用工具时，上述文件自动产生。

.data 文件由制表符分隔的字段行组成。您可以通过从数据源 （例如，Oracle，INFORMIX 或 Sybase 数据库）中抽取数据来生成数据文件并用您熟悉的文本编辑器（例如， WordPad、jot、vi 或 Emacs）对其进行格式转换。参见 *《MineSet 3.0 企业版接口指南》* 可以找到所需似的文件格式。表 1-10 给出了使用 “工具管理器” 为可视化工具创建的文件扩展名类型。

表 1-10 默认样例文件扩展名。

工具	数据文件扩展名	方案文件扩展名	配置文件扩展名
关联规则	*.rules.data*	*.rules.schema*	*.rules.scatterviz*
聚类可视化工具	无	无	*.clusterviz*
决策表可视化工具	*.dtableviz.data*	无	*.dtableviz*
证据导入工具	无	无	*.eviviz*
地图可视化工具	*.mapviz.data*	*.mapviz.schema*	*.mapviz*
记录查看器	*.data*	*.schema*	无
散点可视化工具	*.scatterviz.data*	*.scatterviz.schema*	*.scatterviz*
平伸可视化工具	*.splatviz.data*	*.splatviz.schema*	*.splatviz*
统计可视化工具	无	无	*.statviz*
树可视化工具	*.treeviz.data*	*.treeviz.schema*	*.treeviz*

当启动可视化工具或打开文件时，您必须指定配置文件，而不是 *.data* 文件或 *.schema* 文件。“记录查看器” 可以打开任何 *.schema* 文件。

数据文件可以拥有用户定义的扩展名 （所提供的样例有 *.data* 扩展名）。

筛选按钮

工具管理器中 “数据转换”面板上的按钮可以使您利用数学表达式来筛选数据。结果表只包括表达式为真时的记录 （或者，如果记录为数值型的，则为非 0 值）。当单击*筛选*，就会产生 “筛选”对话框。

该对话框可以让您在左边选择列名和操作符，在右边建立表达式。要了解表达式定义语言的完整描述，参见第 1 页 “添加列”。

筛选面板

从 “查看”下拉式菜单中可以访问 “筛选面板”，以下部分描述这个选项。“散点”，“平伸”，“映射”和 “决策表可视化工具”使用该面板。“树可视化工具”拥有一个类似的面板。“证据可视化工具”有其自己的相应功能部分。

筛选面板 带出一个面板，利用该面板您可以根据一个或多个标准，限制在主查看区域中显示的实体的数目。您可以使用筛选面板来微调显示设置，强调特定信息或缩减显示信息的数量。

筛选范围（仅存在于 “散点可视化工具”中）可让您指定图形对象的高度是在整个数据集的范围内还是在筛选过的数据中进行比例变化。

筛选面板有两个部分。上面的部分可让您在字符型变量的基础上进行筛选操作。要选择变量的所有值，单击*设置全部*。要清除当前选择，单击*清除*。要选择一个值就单击它。要撤消对一个值的选择，只需再单击它一下。

下面的部分可让您在字符型和数值型变量的基础上进行筛选操作。当您操纵滑动条时，只有那些值保持不变的变量才能用于筛选。滚动窗格可显示不同等级的内容 （“包含”、“相等”、“匹配”和 “为空”）。

要对数值型值进行筛选，就输入一个值并选择关系操作 （=， !=，>，<， >=， <=）。要对字母数字值进行筛选，就输入字符串。您可以使用三种字符比较操作类型中的任一种 :

- 包含指示了包含了适当的字符串。例如， California 包含了字符串 Cal 和 forn。
- 相等则需要正好匹配的字符串。
- 匹配允许通配符 :
  - 星号 （*）代表了任何数量的字符。
  - 问号 （?）代表了一个字符。
  - 方括号 （[ ]）包含了要匹配的字符列表。

例如， California 与 Cal*、 Cal?fornia 和 Cal[a-z]fornia 匹配。

在一些情况下 （通常与 “工具管理器”中的分组有关），出现了值选项菜单，而不是文本字段。要忽略那个变量，选择 “选项”菜单中的*忽略*。您可以使用带有这些选项的关系操作符 （例如 >=）。这就意味着选择了包括后续值的指定值。

作为数值型和字符型的比较操作的补充，您还可指定 `is null` 操作，当值为空时操作返回真。

每个字段的右边是附加选项菜单，用该菜单可以指定 “与”或 “或”操作选项。例如，您可以指定 “销售 >20 和 <40”。对于已知变量可以有任意多个 “与”或 “或”从句，但是不能将 “与”和 “或”用到单一变量中去。

单击*应用*按钮来启动筛选操作。如果该面板为当前活动面板时，您按下了*回车*键，就会自动启动筛选操作。

增益比

“增益比”是由除以熵 （纯度的变化）的 “公共信息”形成的 “决策树”分割标准，该熵是忽略了标签值时所做分割产生的熵。可参见第 71 页 “决策树”可得到关于 “决策树”分割标准的讨论。

帮助 （IRIX）

除了 F1 或 Shift-F1 键组合以外，每个工具中的 “帮助”菜单可访问五个帮助功能：

- *单击得到帮助*将光标变为问号。将光标放在可视化工具主窗口中的对象上并单击鼠标会使帮助屏幕出现；该屏幕中包含了那个对象的信息。关闭帮助窗口后，鼠标恢复到原来的形状并撤消了帮助功能。该功能的键盘快捷方式为 Shift+F1。（注意，在有些工具中也可以将箭头状的光标放在对象上并且按下 F1 功能键来访问对象的帮助屏幕。
- *概述*提供了工具主要功能的简单总结，包括如何打开文件以及如何交互操作结果视图。
- *索引*提供了完整帮助系统的索引。该选项当前被禁用。
- *键与快捷方式*提供了所有可利用加速键的可视化工具功能的快捷方式。
- *产品信息* 产生了带有可视化工具版本号码和版权注意事项的屏幕。

帮助 （Windows）

除了 F1 键，每个工具中的 “帮助”菜单提供了对所有 “MineSet 3.0 企业版”手册的访问。*产品信息* 产生带有可视化工具版本号码和版权注意事项的屏幕。

直方图可视化工具

“直方图可视化”工具将数据中的连续列自动分组并将结果送入统计 “可视化工具”中。参见第 16 页 “统计可视化工具”。您可以设置下列 “直方图可视化工具”选项：

- 您可以挑选组的数目或者让 MineSet 自动为您做这件事。
- 您可以设置修剪因子。修剪因子说明了部分极值排除在产生分组的范围以外。默认的修剪因子为 0.05。这就排除了 5% 具极值的实例 （2.5% 具有区间内的最低值而 2.5% 具有最高值）。修剪的目的就是减小在阈值产生过程中数值的影响。

历史

您可以通过使用工具管理器中的*历史表*按钮来得到当前阶段的以前操作历史。参见第 171 页 “历史表按钮”。您也可以利用 “文件”下拉菜单来保存的您的工作。您也可以在完成转换操作之后单击 “工具管理器”的 “数据目标”窗格中的 “数据文件”选项卡来保存数据文件。当您恢复工作时，加载您想要工作的文件或合适的阶段，而操作历史也保存在里面。

预留

预留是指在数据集的部分样本上训练模型的过程，以便于该模型使用剩下的部分进行检验。该过程可在没有用于训练模型的数据上估算模型误差率。用于训练的数据的比例在 MineSet 中被称为 “预留比”。

导入工具

导入工具是一种从*训练集*中建立预测模型的算法，需要带有标签的记录。训练集是导入工具用来“学习”如何建造模型的数据集中的一部分。一旦建立了模型，它的结构就能被可视化或用于没有标记记录的分类。运行导入工具是 CPU 和 I/O 密集过程。正是由于这个原因， MineSet 导入工具在 MineSet 服务器上运行，而不是在 MineSet 客户机上运行 （参见图 1-19）。

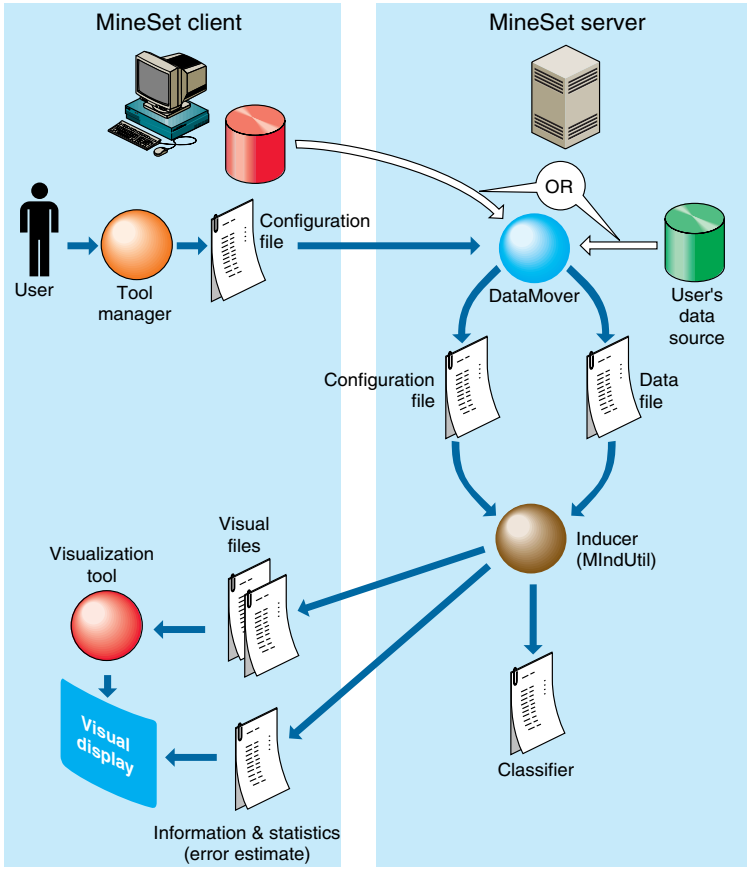

图 1-19 模型的工具执行过程

导入工具需要训练集，该训练集是一个带有属性或特征的表，属性之一被指定为标签。一旦建立了模型，就可以对新记录预计标签。这些新记录必须存在于拥有模型所用全部属性的表中，并具有它们在训练集中的名字和类型。该表不需要包含标签属性。如果存在，在预计的过程中会忽略掉。

工具管理器中的导入工具模式

有四种模式运行导入工具。

- 分类器和错误估计 （或回归器和错误估计）
- 仅用于分类器 （或仅用于回归器）
- 估计误差
- 学习曲线

*分类器和错误估计*模式 （或*回归器和错误估计*模式）使用了预留方法来建立模型 : 随机抽取部分数据来进行训练 （通常为三分之二）而剩下的用于检验。预留比例可以在 Windows 版本的*高级选项*中设置，或者在 IRIX 版本中的*深层导入工具选项*中设置 （参见第 80 页 “错误估计”）。该方法是默认方法并被推荐为初始开发进程。这种方法速度较快并提供了误差估计。

仅用于分类器 （或*仅用于回归器工具*）选项使用了所有的数据建立模型。该方法没有误差估计。当只有很少的数据或当您建立了最终的模型时，可以利用该模式。

*估计误差*模式估算了使用全部数据建立的模型的误差 （使用了*仅用于分类器*或*仅用于回归器工具*模式）。“估计误差”利用了交叉检验，由此导致了较长的运行时间。（参见第 61 页 “交叉验证”。）当只有少量数据时使用该方法。该导入模型与由*仅用于分类器*或*仅用于回归器工具*模式导入的模型一样。

*学习曲线*模式估算了训练集规模对分类工具误差率的影响 （参见第 10 页 “学习曲线”）。

导入工具误差选项

下列选项可用于微调导入工具的误差估计。该 “误差选项”的可用性取决于您选择的导入工具模式。

在*分类器和错误估计*（或*回归器和错误估计*）以及*误差估计*选项中，您可以设置随机子用以确定数据是如何分为训练和检验集的。改变随机子将造成训练集和检验集的不同划分。如果误差估计略有变化，那么导入过程就会不稳定。

在*分类器和错误估计*（或*回归器和错误估计*）选项中您可以设置记录的 “支持比率”以保留训练集。默认为 0.666667 （三分之二）。剩下的记录用于估算误差。

在*估计误差*选项中，您可以设置交叉检验倍数以及重复进程的次数。参见第 61 页 “交叉验证”。

高级导入工具选项

MineSet 支持所有导入工具的几个高级选项。您可以考虑产生某种错误的代价，或对非均匀采样过程的补偿 （真实总量中某些部分比其它部分的采样密度高很多）。另一个选项可使您创建更复杂的模型，该模型以增加计算时间的代价换取更好的准确性。

- **修正** —*修正检验集设置*选项是一个复选框，在使用分类器和错误估计模式时适用于所有导入工具的 Windows 中版本的*高级选项*，或 IRIX 版本中的*深层导入工具选项*中可以找到。当推进可用时，修正对话框就被禁用。参见第 33 页 “修正”。
- **混淆矩阵** —*显示混淆矩阵*选项是一个复选框，使用 “分类器和错误估计”模式时，可以适用于所有导入工具的 Windows 版本中的*高级选项*或 IRIX 版本中的*深层导入工具选项*中找到。参见第 58 页 “混淆矩阵”。
- **投资回报曲线—**投资回报 （ROI）曲线与 “上升曲线”相似，但它以损耗的形式而不是以误差的形式显示准确度； 并将所使用的损失矩阵考虑在内。*显示 ROI 曲线*选项是适用于所有分类过程导入工具的一个复选框，可以在 Windows 版本的*高级选项*或 IRIX 版本中的*深层导入工具选项*中找到。 ROI 曲线需要已选定的标签值。然后为该标签值产生并显示了 ROI 曲线。参见第 143 页 “投资回报曲线”。

- **上升曲线—**上升曲线是一个图形，该图形显示了记录的随机排序与在描述特殊标签值过程中由分类工具所创建的排序之间的差异。参见第 110 页 “上升曲线”。
- **损失矩阵—** “损失矩阵”可让您改变产生某些误差的代价或对其重新加权。损失矩阵与混淆矩阵一起可帮助减少误差的代价。*使用损失矩阵*选项是一个适用于所有导入工具的复选框，它出现在 Windows 版本中的*高级选项*或 IRIX 版本中的*深层导入工具选项*下的对话框的顶部。参见第 111 页 “损失矩阵”。
- **加权设置—**记录加权可给每个记录赋予权重；被采样两次的记录通常得到的权重为 0.5，而剩下的部分其权重被赋予 1。*使用加权*选项是一个适用于所有导入工具的复选框，它出现在 Windows 版本中的 *高级选项* 或 IRIX 版本中的 *深层导入工具 选项* 下对话框的顶部。为权重选择列。*权重保留为属性*选项确定导入工具为了模型化的目的是否可以利用该属性。在某种情况下，即当权重是作为实验设计一部分的分层样例 （stratified sample）的结果时，该模型就不提供对权重列的访问，因为它不是真实世界中实体的性质。
- **学习曲线—**学习曲线作为由一定量记录所创建的模型功能，是一个显示由导入工具产生的模型误差的图形。典型情况下，产生模型所用的记录越多，误差越低。参见第 108 页 “学习曲线”。

设置特殊选项

- 总是忽略类型数组属性。
- 日期被认为是字符串。除非有少量的日期，否则因为离散属性的限制，这样的属性通常被忽略。在运行导入工具之前，您应该将日期分组。

导入工具状态窗口

在您按下*继续*之后，在 “数据目标”面板中，“工具管理器”主窗口下部的 “状态窗口”显示了算法进程并展示了导入模型的特别信息。例如，对于 “决策树”它显示了节点数目，叶子数目以及 “决策树”的深度。该信息自动保存在您的工作站上带有 *.out* 扩展名的阶段文件中。

- 对于 “分类器和错误估计”（或 “回归器和错误估计”），开始一系列点代表了读入文件，然后显示了有关分类工具创建的进程的信息，再后显示了检验集分类的进程。
- 对于 “仅用于分类器”（或 “仅用于回归器”）模式，没有检验集分类阶段。
- 对于 “估计误差”，显示了次数和倍数。
- 对于 “学习曲线”， x 轴上的每个平均点将在线上描述出来，并且平均点的每次运行都将由一个点代表。

分类器和错误估计模式状态窗口细节

当您选择 “分类器”和 “错误估计模式”（或 “仅用于回归器”），“状态”窗口显示：

- 用于将数据分为训练和检验集的随机子。
- 用于产生模型的记录数目。
- 用于评估结果模型的记录数目。
- 所做的正确和不正确预计的数目。
- 对于分类工具，平均标准化均方差代表了概率估计的准确性。对于每个检验记录，均方差是指 1 减去正确标签值概率估计的平方，再加上其它标签值 （不正确）概率估计平方和的结果。标准化均方差是方差半均方根，是一个介于 0 和 1 之间的数。平均标准化均方根是在实验集中利用适当的权重进行平均的标准化均方差 （加权平均）。
- 对于分类工具，分类误差就是错误预计的百分数。
- 对于分类工具，平均均方差和分类误差显示了均值的标准差以及均值的置信区间。如果数据来自于相同的分布，这就是一个可以从分类工具得到的范围。对于误差估计 （不是损失），就会使用比通常的双标准差规则更准确的公式。

估计误差模式状态窗口细节

当您已经选择了 “估计误差”模式，该 “状态”窗口将包含有关分类工具下列信息：

- 交叉验证的次数和倍数。
- 随机子。
- 标准差的估计准确度。
- 估计准确度的 95% 置信区间。

对于回归工具，显示了下列信息：

- 估计均方差和估计平均绝对误差。
- 有关上述准确度矩阵的 95% 置信区间。

设置特殊选项

- 总是忽略类型数组属性。
- 日期被认为是字符串。除非有少量的日期，否则因为离散属性的限制，这样的属性通常被忽略。在运行导入工具之前，您应该将日期分组。

国际化

在 Windows 系统中，可以通过桌面 “控制面板”的 “区域设置”部件来设置地区。该部分用于在 IRIX 系统中设置地区。

MineSet 支持国际化数据集。图形界面中的文本标签为英语，但是您可以查看用于数据编码相对应的语言表示的多字节列名称和数据值。只要安装了相应的语言产品，MineSet 就自动支持日语、中文和韩语的 EUC 编码。对于其它语言和编码，参见第 105 页 “扩展到其它的语言和编码 （仅限 IRIX）”。

在 IRIX 系统中设置地区

客户和服务器系统中必须有您正在使用语言的地区和字体，以及任何用于远程显示的系统。要查看系统中安装的地区列表，在 shell 提示符下输入下列命令：

```
locale -a
```

要设置地区，将环境变量 LANG 设置为合适的地区，而地区是从上述命令产生的列表中选择出来的。例如，要将地区设置为 Japanese， EUC 编码，则利用 csh，输入下列命令：

```
setenv LANG ja_JP.EUC
```

然后从同样的 shell 中调用 MineSet。要想为所有的应用程序永久地设置地区，参考 IRIX 文档。

扩展到其它的 语 言 和 编 （仅限 IRIX）

要让 MineSet 在安装过程所包含的地区之外运行，需将资源文件拷贝到合适的目录中去并进行修改。 MineSet 可视化工具以二维和三维字体使用 “打开发明器”。为了使文本能够正常的显示，您必须安装了类型 II （通常称为 CID 轮廓）字体。

对于下列地区安装过程中需包含资源文件：

- ja_JP.EUC
- ko_KR.euc
- zh_CN.ugb
- zh_TW.ucns

在地区 *locale_name* 中运行 MineSet （参见第 10 页 “在 IRIX 系统中设置地区”中怎样列出安装的地区）：

1. 正常安装 MineSet。
2. 以 root 登录。

3. 从 */usr/lib/X11/app-defaults* 拷贝下列资源文件至 */usr/lib/X11/locale_name/app-defaults* 中：
   - *Clusterviz*
   - *Dtableviz*
   - *Eviviz*
   - *Mapviz*
   - *Mineset*
   - *Scatterviz*
   - *Splatviz*
   - *Statviz*
   - *Treeviz*
4. 在 */usr/lib/X11/locale_name/app-defaults* 中编辑资源文件。您需要知道资源名字以及您想用的字体的特别要求 （参见表 1-11 中的样例）。
5. 在 *locale_name* 设置地区并调用 MineSet。

样例 1-3 用于 “韩语”的资源文件变化

为 “韩语”所需做的变化在表 1-11 中给出。列出的字体来自于下列文件的列表中：

- */usr/lib/X11/fonts/ps2xlfd_map.korean*
- */usr/lib/X11/fonts/ps2xlfd_map.korean.outline °£*

表 1-11 “韩语”字体资源

文件	“英语”资源 （一些行会自动换行）	“韩语”资源 （一些行会自动换行）
Clusterviz Statviz	titleFont:screen12	titleFont:screen12， -ksg-mj-medium-r-normal--14-130-75-75-c-140-ksc5601.1987-0
Clusterviz Statviz	gradationsFont:screen11	gradationsFont:screen11， -ksg-mj-medium-r-normal--12-110-75-75-c-120-ksc5601.1987-0
Clusterviz Statviz	balloonFont:screen11	balloonFont:screen11， -ksg-mj-medium-r-normal--12-110-75-75-c-120-ksc5601.1987-0
Clusterviz Statviz	xFontEncoding:ISO8859-1	xFontEncoding:ksc5601.1987-0

表 1-11 “韩语”字体资源

文件	“英语”资源 （一些行会自动换行）	“韩语”资源 （一些行会自动换行）
Dtableviz，Eviviz，Mapviz，Scatterviz，Splatviz，Treeviz	myDefaultFont:Helvetica-Narrow	myDefaultFont:Helvetica-Narrow； Gungso-Regular--KSC-H
Mineset	zoom2*fontList:-*-*-medium-r-*-*-6-*-*-*-*-*-*-*	zoom2*fontList:-*-*-medium-r-*-*-6-*-*-*-*-*- * -；-ksg-*-medium-*--12-*:
	zoom3*fontList:-*-*-medium-r-*-*-8-*-*-*-*-*-*-*	zoom3*fontList:-*-*-medium-r-*-*-8-*-*-*-*-*- * -；-ksg-*-medium-*--12-*:
	zoom4*fontList:-*-*-medium-r-*-*-10-*-*-*-*-*-*-*	zoom4*fontList:-*-*-medium-r-*-*-10-*-*-*-*-*-* - ；-ksg-*-medium-*--14-*:
	zoom5*fontList:-*-*-medium-r-*-*-12-*-*-*-*-*-*-*	zoom5*fontList:-*-*-medium-r-*-*-12-*-*-*-*-*-* - ；-ksg-*-medium-*--14-*:
	zoom6*fontList:-*-*-medium-r-*-*-14-*-*-*-*-*-*-*	zoom6*fontList:-*-*-medium-r-*-*-14-*-*-*-*-*-* - ；-ksg-*-medium-*--18-*:
	zoom7*fontList:-*-*-medium-r-*-*-16-*-*-*-*-*-*-*	zoom7*fontList:-*-*-medium-r-*-*-16-*-*-*-*-*-* - ；-ksg-*-medium-*--24-*:
	zoom8*fontList:-*-*-medium-r-*-*-24-*-*-*-*-*-*-*	zoom8*fontList:-*-*-medium-r-*-*-24-*-*-*-*-*-* - ；-ksg-*-medium-*--24-*:

迭代 k- 均值

迭代 k- 均值是一种由 MineSet 聚类工具使用的方法。该方法将在用户选择的范围内自动的选择聚类的数目；聚类的数目不限。该算法通过产生几个侯选聚类结果并利用离散度矩阵组合和用户可选选择点来选择其中最好的一个方式，以确定聚类的数目。

拉普拉斯校正

“证据可视化工具”利用一个导入过程，记数 （或权重）在此被用于计算概率。拉普拉斯校正避免了极端概率 （0 和 1）。

要寻找拉普拉斯校正功能，从 “工具管理器”的 “挖掘工具”中单击 “分类”选项卡并选择 “导入工具：”>“证据”。当您单击了 “高级选项”按钮，在产生的对话框中显示出 “证据选项”下的拉普拉斯复选框。

在一个样例中使用了拉普拉斯校正，在该例中我们并不将概率 1 分配给下述的事件，即 “病人被测出 AIDS 为阳性就得了绝症”，而是概率接近于 1，允许存在误差或未被代表的样例。拉普拉斯校正会使 “证据可视化工具”概率偏向平均值，这样就避免了极端值 0 和 1。

这意味着 “证据主窗口”中的每个图对于每个类来说都有非 0 的切片。组中的记录越少，偏向均值的变化就越大。如果选择了拉普拉斯校正，并且因子为空或设为 0 ，那么通过应用因子 “1/ 训练集权重”的启发式过程，就可以进行自动拉普拉斯校正。

学习曲线

作为由大量记录所创建的模型的一个功能，学习曲线是用于显示导入工具产生的模型误差的一个图形。典型情况下，产生模型所用的记录越多 , 误差越低。

学习曲线是通过为曲线上的每个点产生指定数目的模型来创建的。每个模型用随机采样的记录产生，而剩下的记录 （那些没有用于训练的）用来估计误差。

产生学习曲线将会耗费大量的 CPU 时间。如果 t_i 是在训练集 i 上训练导入工具的次数（i 的范围为 1 到所有点的数目），并且对每个点运行 k 次，总次数为

$$k*\sum_i t_i$$

每点运行次数的增加会成比例地增加运行时间，但会改善对均值的估计。运行次数的默认值为 3。

“散点可视化工具”筛选面板可以用于筛选要显示的数据类型 （平均点、置信区间、插值点或实际轨迹）。例如，您也许想删除轨迹和置信区间数据点，只保留平均和插值点。

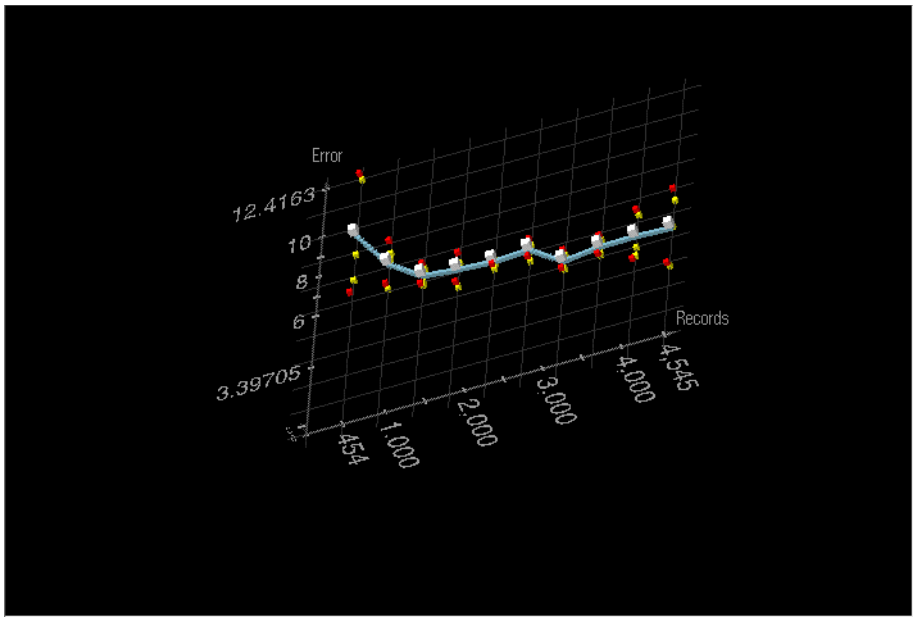

图 1-20 学习曲线

“学习曲线”是 “挖掘工具”选项卡 “分类”菜单 （或 “回归”菜单）中的一种模式。任何导入工具都可以使用它。当选中了 “学习曲线”模式，*高级选项* 对话框 （或*深层选项* 对话框）允许您在对话框的右侧指定*学习曲线 选项*，包括：

- 学习曲线中的点数，
- 每个点的运行次数，
- 在起始点和结束点所用记录的数量。

每个中间点所用的记录数目将自动计算。

必须指定学习曲线中点的数目；它也必须大于或等于 1。可以指定起始和结束点的记录数目以允许为指定范围的训练集产生学习曲线。如果这些选项为空，它们将根据学习曲线上的点数目和训练集中的记录数目自动计算。这将默认覆盖整个范围的训练集。例如，假定文件包含了 80,000 条记录。如果您在学习曲线上指定了 3 点，算法在 20,000，40,000 和 60,000 个记录上产生点。通常“缩小”到一个较小的范围很有用。例如，可以在 1000 到 10,000 个记录范围内产生学习曲线。

上升曲线

上升曲线画出了记录的随机排序与在描述特殊标签值时由分类工具所创建的顺序之间的差异。例如，您创建了一个模型来预计什么样的用户将要 “客户波动”，而现在就要在他们波动之*前*将目标对准他们。上升曲线可以帮助完成这一目标。

上升曲线是这样的图形，X 轴显示了记录的数量从 0 到 100%，而 Y 轴显示了与那些拥有给定标签值的用户记录的数目 （在本例中 `Churn=yes`）。在图 1-21 中显示了两条曲线。

较低的曲线 （红）显示了在记录随机排序的情况下，波动用户数量的期望。上面的曲线（白色）显示了，根据每个记录的分类工具评价 （概率估计）顺序排列时，波动用户的数量。那些代表分类工具找出的最有可能波动的用户的记录最先出现；那些波动可能较小的用户最后出现。通过分类工具曲线和随机曲线之间的差异可以看出分类工具排序所表现出的优势。

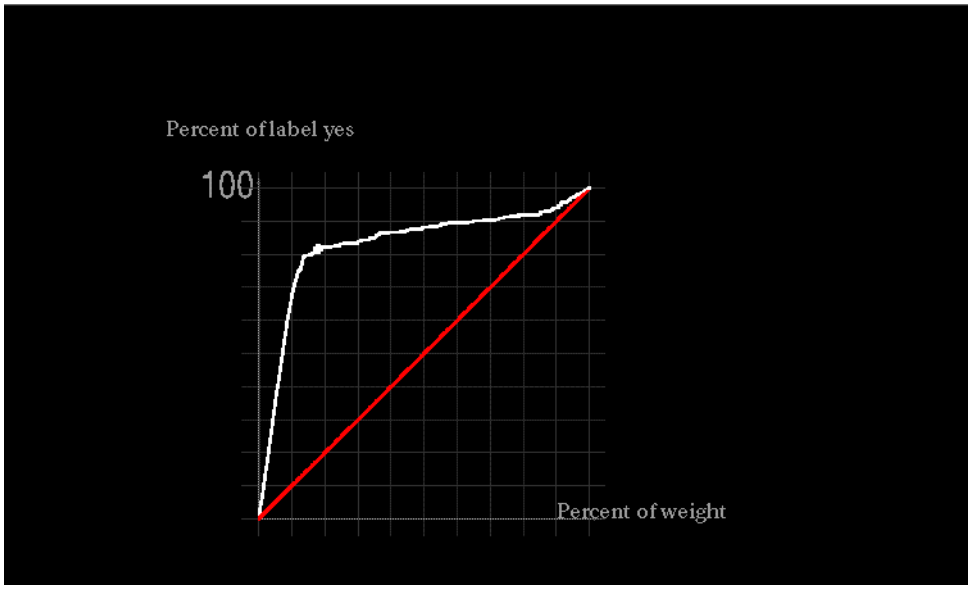

图 1-21 上升曲线

在建立该上升曲线的过程中，将选择的分类工具应用于该检验集。数据集的指定部分用于训练，然后，导入分类工具在剩下的数据集上运行。当使用 “分类器和错误估计”模式时，*显示上升曲线*选项是*高级选项*（或*深层选项*）下的一个复选框，它适用于所有导入工具。“上升曲线”需要一个选定的标签值。上升曲线产生并展示了 ' 个标签值。

损失矩阵

损失矩阵的目的就是控制分类工具可能产生的错误类型。在许多情况下，一些错误比另一些更致命。一个典型的例子来自于蘑菇数据集，其中蘑菇被分为可以食用的和有毒的。将可食用的蘑菇划分为有毒的会损失一些钱 （未吃蘑菇的价格）。而将有毒的蘑菇划分为可以食用的会损失上千倍的费用 （住院的费用）。分类工具利用损失矩阵来避免这些代价昂贵的错误。

损失矩阵和混淆矩阵结合使用最有效果。混淆矩阵指出了分类工具产生的错误的类型和数量。如果混淆矩阵指示了分类工具产生了大量代价昂贵的错误，那么在损失矩阵中赋给这种错误较高的权重也许是改进分类工具表现的最佳途径。可用 “高级选项”（ IRIX 系统下的深层选项）下的 “使用损失矩阵”复选框来激活损失矩阵。对于所有的分类导入工具都可用。

在下面的例子中说明的该过程：图 1-22 显示了在只有 0.1 （10%）的比例被用作训练集时，利用 “决策树导入工具”建立的*蘑菇*数据集的一个混淆矩阵。

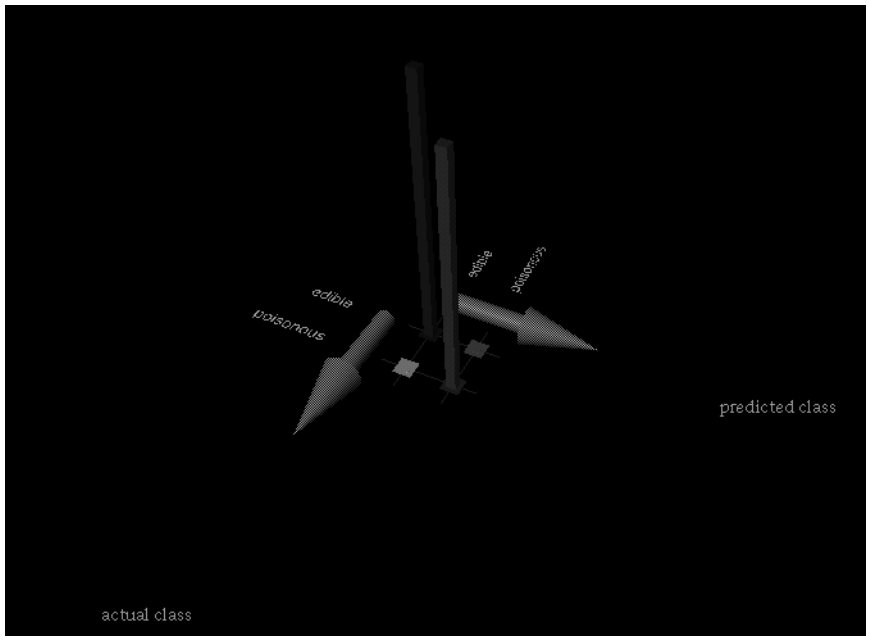

图 1-22 利用默认设置为蘑菇数据集建立的混淆矩阵

代表有毒蘑菇的 8 条记录被划分为可食用的（0.1%）；代表可食用蘑菇的 15 条记录被划分为有毒的 （0.2%）； 3793 可食用的蘑菇和 3496 有毒蘑菇被正确的区分开。当分类工具的错误率仅为 0.31 （小于百分之一）时，我们的估计损失将为 $10000 * 8 + $2 * 15 = $80,030 或更糟。

图 1-23 显示了同样数据集的混淆矩阵，但这是在使用损失矩阵代表上述费用的条件下运行了 “决策树导入工具”之后生成的。新分类工具非常保守，在将有毒蘑菇划分为可食用蘑菇方面没有犯错误；但是它却将 1558 （1543+8）个可食用蘑菇划分为有毒蘑菇。估计的损耗总和是这样：$10000 * 0 + $2 * 1 5 58 = $3116只有前述费用的 3%。

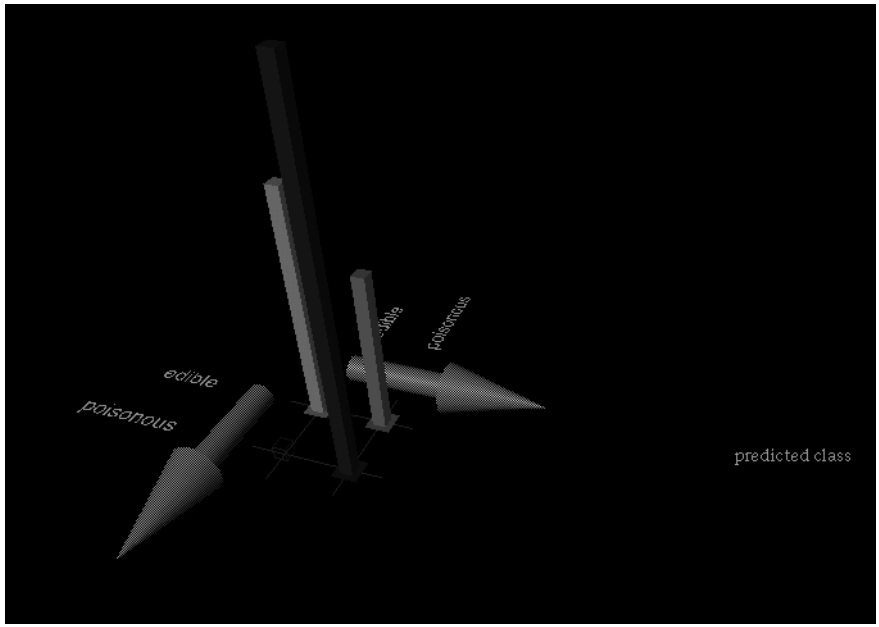

图 1-23 带有损失矩阵的蘑菇数据集混淆矩阵

损失矩阵也可有助于预计未知 （空值）值，该值以问号显示 （?）。例如，设想如果请教外面专家判断蘑菇是有毒的还是可食用需花费 $1。那样的话，一些分类过程将导致未知预计。运行 “决策树导入工具”产生的混淆矩阵显示在图 1-24 中，其中有 1551 个为未知，只有 15 个可食用的蘑菇被划分为有毒的。整个的费用为 $10000*0 + $1 * 1551 + $2 * 15 = $1581。

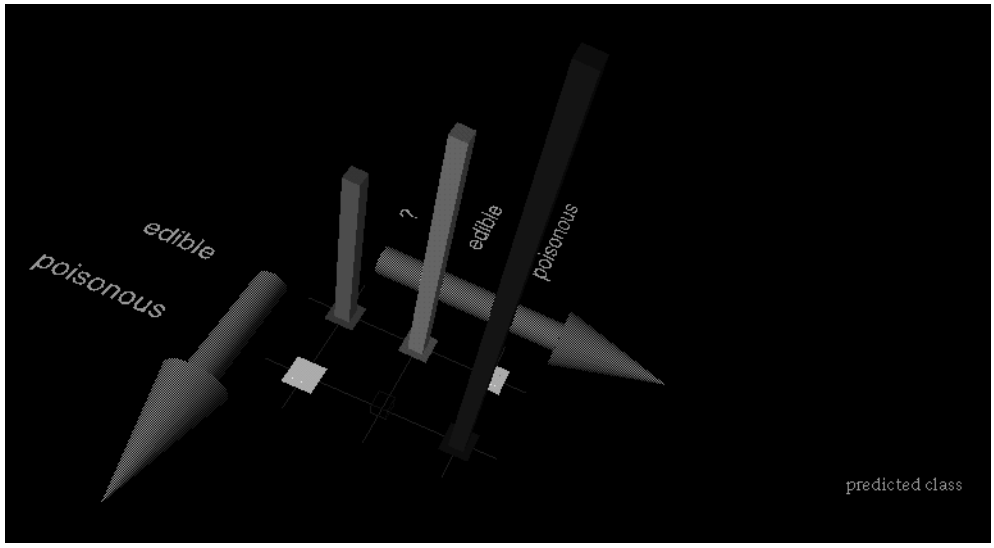

图 1-24 带有损失矩阵的蘑菇数据集混淆矩阵混淆矩阵允许未知预计

对于决策树，损失矩阵的基础是在树叶处所做的概率估计。对于可靠估计：

1. 利用 “决策树”和 “选项树”中的 “深层选项”将 “分割下限”从默认值增加到更高的值 （例如：5)。总体上说，训练集越大并且噪声越多该值就应越高。
2. 使用大训练集。您也许需要大训练集从而得到更可靠的估计。
3. 使用 “选项树”。它们通常提供更好的概率估计并试图降低损失。例如，运行上述蘑菇样例，将 $10000 改变为 $100 并且不允许出现未知值，则 “决策树”产生的估计损耗为 $1464，而 “选项树”产生的估计损耗为 $662。

*使用损失矩阵*选项是*高级选项*（或*深层选项*）下的复选框，它适用于所有分类过程的导入工具。*编辑矩阵*按钮可用于定义损失矩阵。为避免预计产生未知值，在未知预计列中填入矩阵中的最大值。

如果在 “证据可视化工具”中指定了 “损失矩阵”，那么就会在*高级选项*（或 IRIX 系统中的*深层导入工具选项*）中的概率圆饼图的右下方出现一个标有*使用损失矩阵*的按钮。选择 （默认情况下）“损失矩阵”用于调节概率的显示。最大的切片所显示的是考虑了 “损失矩阵”的类。要想看到不使用 “损失矩阵”时概率的状态，取消选择*使用损失矩阵*按钮。当使用了 “损失矩阵”，就会出现一条灰色的切片，这是因为您编辑损失矩阵时生成了一个用于预计空值的列。如果灰切片是最大的切片，则分类工具预计为空值。

地图可视化工具

“地图可视化工具”向您展示了三维景观中以条形图表现的数据。该工具在创建与空间相关的可视化过程中有用，而在数据具有地理意义时尤其有用。

数据项与可视景观中的条形图对象相关。然而，这些对象具有可识别的形状和位置。景观由一些这样的对象组成，每个都具有独立的高度和颜色 （参见图 1-25）。除了可以在景观中动态的进行漫游之外，您还可以利用概化上寻和细化下寻来进行概括或观察更高的精细度，还能利用动画来观察数据沿一维或二维参变量的变化。

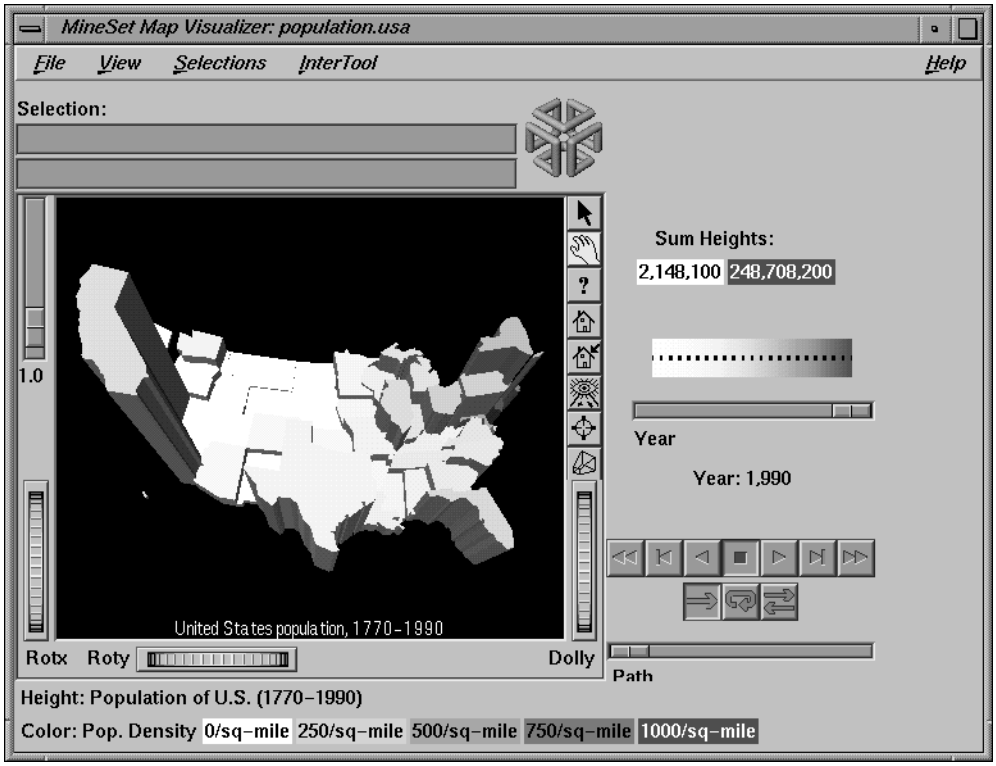

图 1-25 地图可视化工具样例显示了 1990 年美国的人口分布

景观也可以由这些地理对象轮廓的光滑平面组成，并在指定的位置生成 “条形图”。

另一种景观图是由一些端点位于指定位置的线组成，这些线都具有独立的宽度和颜色。线是具有宽度和颜色属性，而不具有高度和颜色属性的任意形状的对象和条形实体。

地图可视化工具的文件需求

“地图可视化工具”需要数据、 gfx、等级和配置文件：

- 数据文件由制表符分隔的字段行组成。典型情况下，“工具管理器”创建该文件。您可以不使用 “工具管理器”产生该文件 （对于所必需的文件格式，参见 *《MineSet 3.0 企业版接口指南》*，“为地图可视化工具创建数据、配置、等级和 GFX 文件”）。

  数据文件是从数据源 （例如， Oracle、 INFORMIX 或 Sybase 数据库）中提取数据的结果，然后将之转化为 “地图可视化工具”能够使用的格式。数据文件有用户定义的扩展名 （“地图可视化工具”所提供的样例文件具有 *.data* 扩展名）。

- gfx 文件由形状描述和将要显示的 1、 2 或 3 维对象的位置组成。

  Gfx 必须有一个 *.gfx* 扩展名。 MineSet 包含不同的 *.gfx* 文件，包括按照州、电话区域编码和邮政编码等不同标准来划分的美国地图以及按省来划分的加拿大地图。您也可以手工产生所需要的 *.gfx* 文件 （参见 *MineSet 3.0 企业版接口指南*，“为地图可视化工具创建所需格式的数据、配置、等级和 GFX 文件”）。

- 等级文件由下列描述组成

  - 所显示的不同图形对象的列名
  - 描述图形对象形状和位置的 *.gfx* 文件的文件名
  - 一种对图形对象等级关系的可选描述，可以让您看到更多或更少的细节，被称为细化下寻和概化上寻功能。

  等级文件可以定义州和包含多个州的区域之间的关系，并按照州或区域的层次来显示人口数。例如， *gfx_files/usa.state.gfx* 文件描述了美国 50 州的形状； *gfx_files/usa.state.hierarchy* 文件描述了将州划分为区域，再将区域划分为东西部以及将东西部合成完整的美国这种等级关系。

  要得到更多的信息，参见 *《MineSet 3.0 企业版接口指南》*，“为地图可视化工具创建数据、配置、等级以及 GFX 文件”。

- 配置文件描述了输入数据的格式以及如何显示它们。一般情况下，利用 “工具管理器”创建该文件。您也可以利用自己喜欢的文本编辑器（例如， WordPad， jot， vi 或 Emacs）而不是 “工具管理器”来产生该文件 （参见 *《MineSet 3.0 企业版接口指南》* 中的 “为地图可视化工具创建数据、配置、等级以及 GFX 文件”。

  若想从 “文件”下拉菜单中选择 “打开”时能够被列出，配置文件应具有 *.mapviz* 扩展名。当启动 “地图可视化工具”，或打开一个文件时，需指定配置文件而不是数据文件。

启动地图可视化工具

有几种方法可以启动 “地图可视化工具”：

- 利用 “工具管理器”来配置和启动 “地图可视化工具”。参见 *《MineSet 3.0 for Windows 企业版用户指南》* 有关 “工具管理器”的详细信息，这些信息对所有的 MineSet 工具都通用。
- 如果您知道要利用的配置文件，双击配置文件的图标。这样就启动了 “地图可视化工具”并自动加载了指定的配置文件。这只有当配置文件以 *.mapviz* 结尾时才有效（对于利用 “工具管理器”为 “地图可视化工具”创建的配置文件来说总是这样）。
- 从 IRIX 命令行中启动 “地图可视化工具”可以通过输入

  `mapviz` [*configFile*]

  configFil 是可选的，它指定了所使用的配置文件的名字。如果未指定配置文件，那么您必须用 “文件” > “打开”来指定一个。

调用地图可视化工具的选项

在 IRIX 系统上，您可以利用 **`-quiet`** 选项来删除指示进程的弹出式对话框。通过加入一行

```
*minesetQuiet:TRUE
```

到您的 *.Xdefaults* 文件可以使该选项永久可用。

您也可以设置一个警告执行语句，参见第 199 页 “警告选项”条目。 Windows 用户可以从 “文件” > “特性”菜单中设置这些选项。

利用工具管理器来配置地图可视化工具

该部分描述了如何利用 “工具管理器”来配置 “地图可视化工具”。虽然 “工具管理器”大大简化了配置 “地图可视化工具”的任务，但您仍然可以利用文本编辑器为该工具建立一个配置文件 （参见 *《MineSet 3.0 企业版接口指南》*中 “为地图可视化工具创建数据、配置、等级以及 GFX 文件”）。

产生 .gfx 和 .hierarchy 文件

要使用 “地图可视化工具”，您必须为应用程序提供定义显示图形对象的两个文件：

- 一个或多个 *.gfx* 文件，该文件定义了所要显示的图形对象的形状。
- *.hierarchy* 文件，描述了多个互有联系的地图 （*.gfx*）文件之间的关系。

这些文件不是由 “工具管理器”创建的； 它们必须以 MineSet 的一部分存在或由用户创建。 Windows 用户可以在 MineSet 所安装的 *\mapviz.gfx_files 文件目录下找到它们。* IRIX 用户可以在 */usr/lib/MineSet/mapviz/gfx_files* 目录中找到它们。如果您决定自己来创建，可以在 *《MineSet 3.0 企业版接口指南》*中的 “为地图可视化工具创建数据、配置、等级和 GFX 文件”找到说明。

.gfx 和 *.hierarchy* 文件是 MineSet 包的一部分，它们包括

- 美国各州
- 美国各县
- 美国 5 位邮政编码所代表的区域
- 美国电话区域码所代表的区域
- 加拿大各省及领土
- 墨西哥各州
- 澳大利亚各州及领土
- 欧洲中西部各国家
- 法国和荷兰的地区

“地图可视化工具”需要的数据文件具有

- 指示地理对象的一列 （例如：州）。列中的每一行必须指定唯一的地理对象 （在样例中，意味着每一个州对应一行）。
- 至少有一个数字型值的列映射为 （利用数学表达式）每个地理条的高度和 / 或颜色。这些列可以是标量、一维数组或二维数组。如果该列为数组，必须使用滑动条来指定数据点以映射到高度和颜色。

如果高度和颜色映射为一维或二维数组，这些数组必须具有相同的索引 （参见 *《MineSet 3.0 企业版接口指南》* 中的 “为地图可视化工具创建数据、配置、等级以及 GFX 文件”）。

创建滑动条和动画

参见第 10 页 “动画”。

地图可视化工具选项

单击*工具选项*按钮将显示一个对话框，该对话框可允许您对 “地图可视化工具”中一些选项默认值进行改变。

下面的部分描述 “地图可视化工具”、“选项”对话框中的按钮和字段。

地理布局

实体文件 指定了在 “地图可视化工具”主窗口中用于代表地理 “实体”对象的 *.hierarchy* 文件。

*轮廓文件*指定了所要画的对象的轮廓，该轮廓显示为光滑平面，其上将放置 3-D 实体对象。

*寻找文件*按钮可浏览文件，并寻找所要使用的 *.hierarchy* 文件。

*实体文件*和*轮廓文件*字段是可选的。如果不提供 “实体文件”，那么 “地图可视化工具”所创建的图形实体对象将由画面中任意形状和随意放置的简单矩形组成。

高度

该部分指定了初始高度*比例*值 （默认为 1.0）以及是否在 “地图可视化工具”窗口的底部显示高度图例。

颜色

要使用这些 “颜色”选项，您必须将列映射为 “数据目标”面板所需的 * *颜色 – 条形图*。要得到如何选择和改变颜色的更详细的解释，参照第 48 页 “颜色选择”。

颜色列表 — 您可以利用颜色列表标签旁边的 + 按钮来指定颜色列表。这就会产生一个颜色编辑器，您可以利用它指定要加入列表的颜色。

映射 — 您可以指定在图形显示中所表现的颜色变化是*连续的*还是*离散的*。如果您选择了*连续的*，作为在 “映射”字段中被映射为颜色值的一个函数，颜色值在 “颜色列表”字段中所输入的颜色之间逐渐变化。

弹出式按钮右边的字段可让您输入颜色所映射的指定值。在该字段中值的数目必须与从“使用的颜色列表”字段中输入的颜色数目相同。

您可以按照需要在该字段中输入多种颜色以用于显示。如果列中映射为 * 颜色 - 条形图的颜色数目超过了您所选择的颜色的数目，“地图可视化工具”将实时加入适量的随机抽取的颜色。要得到更多的关于选择颜色的信息，参见第 48 页 “颜色选择”。

图例开关 — 可让您选择显示或隐藏颜色图例。

标准化开关 — 可让您确定 “地图可视化工具”是否在颜色列的最大和最小值之间自动按比例变化 （这被称为颜色标准化），正好与您手工指定阈值相反。当*标准化开关*可用时，阈值必须在 0 到 100 之间，代表了颜色列中最大值到最小值之间范围的百分数。

滑动条

参见第 154 页中的 “为地图可视化、散点可视化、平伸可视化创建的滑动条”。

消息字段

该选项允许您指定当实体被选中时所显示的消息。要列出和描述用于输入字段的格式类型，参见 *《MineSet 3.0 企业版接口指南》* 里的 “为地图可视化工具创建数据、配置、等级和 GFX 文件”中 “消息语句”部分。

可让您指定显示在地图可视化工具主窗口底部的字符串。该字符串必须包含在双引号内。

执行字段

该选项可让您输入在双击一个实体时运行的命令。格式与消息语句相似。如果不出现执行语句，双击也不起作用。

要得到 “执行”字段更详细描述，参见 *《MineSet 3.0 企业版接口指南》* 中的 “为地图可视化工具创建数据、配置、等级和 GFX 文件”。

重设工具选项

在对 “工具选项”对话框作出修改以后，如果您想要将所有的设置重设为默认值，单击 *重设* 按钮。

接受工具选项

一旦您完成了对 “工具选项”对话框的修改，单击*确定*将返回 “工具管理器”主窗口。

地图可视化工具文件设置

“工具管理器”将 “地图可视化工具”的信息存储在几个文件中，并具有相同的前缀：

- < *前缀* >.*mapviz.data* 包含数据。
- < *前缀* >.*mapviz.schema* 描述了数据文件。
- < *前缀* >.*mapviz* 包含了地图可视化工具所需的信息。
- < *前缀* >.*mineset* 包含了创建其它文件所需的所有信息。

要指定前缀，使用 “工具管理器”主窗口中 “文件”菜单的菜单选项 *保存 ...*。如果您不指定前缀，则它将根据数据源的类型自行确定。

当您使用了 *调用工具*按钮，根据需要 *.data*， *.schema*，和 *.mapviz* 文件将被更新。

挖掘工具选项卡

从 “工具管理器”中的 “数据目标”窗格中单击 “挖掘工具”选项卡可得到下列选项卡 :

- 关联，关联规则
- 聚类，聚类
- 按照以下模式分类：仅用于分类器、分类器和错误估计、估计误差以及学习曲线。这些选项卡中可用的导入工具是 : 决策树、选项树、证据和决策表

- 按照回归模式进行回归 : 回归器和错误估计，仅用于回归器，估计误差以及学习曲线
- 列重要性，列重要性
- 任何插件，例如 ACpro

多重选择

在大多数工具中，利用 Shift- 单击鼠标左键就可以完成选择。如果在对象上单击鼠标左键的同时没有按下 Shift 键，则在选择了光标下面对象的同时也撤消了以前所有选择的对象。单击鼠标左键时按下 Shift 键可以切换该对象的选择而不影响其它选择。（“平伸可视化”工具具有不同的界面，可在 “平伸可视化工具”条目中找到其描述。）

当您选择了一项，一条描述该项的信息会出现在工具的主窗口中，在默认情况下，可视化工具只显示最后选择的对象的信息。分离的 “记录查看器”窗口显示了一个表，该表显示了所有选项的值。在 “树可视化工具”和 “地图可视化工具”中，选择 “选项” > “显示值”来查看该表。

如果已经为特殊的工具设置了消息，该消息也出现在表中。可通过拖动列之间的分隔符来重新调整表中列的大小。您也可以单击一个值并在表的顶部显示该值的完整文本。

交互信息

“交互信息”是一个 “决策树”分割标准，该标准控制着分类工具进行决策的路线。“交互信息”是指父节点和子节点纯度加权平均之间纯度的变化 （也就是*熵*）。加权平均是根据每个子节点上记录的数目来进行的。参见第 71 页 “决策树”可得到其它分割标准讨论。

简单 -Baye 方法

“证据导入工具”有时被称为单纯 Bay 或简单 Bay。它建立了一个模型，该模型假定在已知类的条件下，属性值概率独立。例如，对于 *iris* 系统数据集，四个属性 （萼片长度、萼片宽度、花瓣长度以及花瓣宽度）对于每个 iris 类 （ iris-setosa， iris-versicolor 和 iris-virginica）来说是独立的。当然这个简化的模型并不很真实，但这个模型用于最初的数据开发却非常有效，并且它的分类预计表现在实际应用过程中也很好。

用窗口控件在非树可视化工具中漫游

该部分由三个表组成，并可用于 “证据”、“决策表”、“地图”、“散点”以及 “平伸可视化工具”漫游控件的快速查询。表 1-12 描述了漫游按钮。

表 1-12 在非树可视化工具中的漫游按钮

按钮	名字	动作
	选取	改变程序至选取模式 （箭头）。在选取模式中，您可以加亮 （掠过）或选择 （单击）图中的成员。
	抓取	改变程序至抓取模式 （手形）。抓取模式中，您可以在窗口中移动图象： 要旋转图，按下鼠标左键并移动鼠标。 要在窗口中移动图，按下鼠标的左键和右键 （如果系统配置了三键鼠标，可以使用鼠标中键）并移动鼠标。
	主视图	将图返回至设计为主视图的位置。默认情况下，当可视化工具被第一次调用时，这是起始位置。您可以通过设置主视图按扭来改变主视图位置。
	设置主视图	为图设置新的主视图。当您想保存某个视图或位置时使用。
	全景视图	将图移至窗口的中心位置并且使它的全部可见。
	搜索对象	将您选择的点移至窗格的中央并对其缩放。当鼠标的光标变为瞄准形状时，将它移到您想要看地更清楚的一点，然后按住鼠标左键 （或利用鼠标鼠标中键，如果您的系统配置了三键鼠标）。
	三维透视模式	切换三维透视模式。
	顶视图	将图变为顶视图 （只适用于 “散点”和 “平伸可视化工具”）。
	前视图	改变图为前视图 （只适用于 “散点”和 “平伸可视化工具”）。
	侧视图	改变图为侧视图 （只适用于 “散点”和 “平伸可视化工具”）。

表 1-13 描述了在非树可视化工具中调节滑动条和滑动轮。

表 1-13 在非树可视化工具中调节滑动条和滑动轮

滑动条或滑动轮	动作
高度滑动条 （左上）	升高或降低块、饼或条形图的高度来强调差异。
细节滑动条 （只适用于 “证据”和 “决策表可视化工具”）	筛选出不重要的属性。
% 权重阈值滑动条 （只适用于 “证据”和 “决策表可视化工具”）	筛选出记录权重小于数据集中记录权重总和指定百分数的属性，最多为 2%。
Rotx 轮	围绕 X 轴旋转图。
Roty 轮	围绕 Y 轴旋转图。
缩放轮	将图放大或缩小。

表 1-14 列出了在非树可视化工具中，对图进行的几种操作。

表 1-14 控制非图可视化工具场景

动作	滑动条或滑动轮	鼠标或键盘动作
在选取和抓取模式中切换	N/A	按下 Esc 键或漫游按钮。
移动画面	N/A	在抓取模式中，单击并按住鼠标右键。向您想要移动图的方向移动光标。
升高或降低块，饼或条形图的高度来强调差异	高度滑动条 （左上）	N/A
将画面围绕 X 轴旋转	Rotx 轮	在抓取模式中，单击并按住鼠标左键。向您想要旋转图的方向移动光标。
将画面围绕 Y 轴旋转	Roty 轮	在抓取模式中，单击并按住鼠标左键。向您想要移动图的方向移动光标。
将画面放大或缩小	调节轮	在抓取模式中，单击并按住鼠标左键 （或鼠标中键）。鼠标向下移用于放大，向上移用于缩小。
在细节层中细化下寻（只适用于 “决策表”和 “地图可视化工具”）	N/A	将鼠标箭头放在指定图上 （或所有图的背景上）并单击鼠标右键。
在细节层中概化上寻（只适用于 “决策表”和 “地图可视化工具”）	N/A	将鼠标箭头放在指定图上 （或所有图的背景上）并 Ctrl- 单击鼠标右键 （或单击鼠标中键）。

用窗口控件在树可视化工具中漫游

“树可视化工具”显示是很好的方法，它好象是通过照相机来查看画面。要改变视图，您只需改变照相机的位置 （视点）。该部分由两个表组成，并可用于 “树”、“决策树”、“选项树”以及 “回归树可视化工具”控制的快速查询。表 1-15 描述了漫游按钮。

表 1-15 树可视化工具中的漫游图标

图标	动作
	将图返回至指定为主视图的位置。默认情况下，当可视化工具被第一次调用时，这是图的起始位置。您可以使用下一个图标来改变主视图位置。
	为图设置新的主视图。用该功能来保存某个视图或位置。
	将图移至窗口的中心位置并且它的全部可见。
	恢复以前的移动 （象 “网上浏览器”中 “后退”按钮一样）。
	重做已恢复的移动 （象 “网上浏览器”中的 “前进”按钮一样）。
	把节点移向树根。
	从一个节点或条移向左边。
	从一个节点或条移向右边。
	沿着节点左路径向下移动。
	沿着节点右路径向下移动。
	从当前模式中弹出一个可选路径的菜单。

表 1-16 描述了在树可视化工具中调节滑动条和滑动轮。

表 1-16 在树可视化工具中调节滑动条和滑动轮

滑动条或滑动轮	动作
高度滑动条 （左上）	升高或降低条高度来强调差异
H 轮	上下移动视点
倾斜轮	改变照相机的上下倾角
左右移动轮 （<-->）	将视点左右移动
调节轮	将照相机前后移动

标称排序菜单

“证据可视化工具”和 “决策表”上的 “标称排序”菜单可控制标称 （命名的）属性值如何排序，并提供了以下选项：

- 按字母排序使具有标称值的属性从左到右 （或从上到下）按字母顺序进行排序。
- 依据权重大小从左到右进行排序，那些具有最大权重的记录在左面。
- 依据标签概率 （默认情况下）使标称属性值按照份额的大小进行排序，该份额对应于类中的一个。如果标签是分组的属性，则默认情况下使用最高的组。如果标签是标称的，那么在先验概率饼图中占有最大份额的类用做默认。如果选择了特殊类，并且随后就必须按照标签概率进行排序，那么被选中的类用于确定次序。在所有情况下，如果存在 “空值”，它就保持为第一个值。

标准化公共信息

“决策树”工具使用了 “标准化公共信息”作为默认分割标准。从技术上说，这是 “公共信息”除以子节点数目的以 2 为底的对数。可参考第 71 页 “决策树”以获得关于 “决策树”分割标准的讨论。

空

一些可视化工具，例如，“平伸可视化工具”以及“地图可视化工具”，当含有未知数据值或空值的字段映射为可视属性时，使用了特殊的代表方式。（关于空值的讨论，参见 *《MineSet 3.0 企业版接口指南》*的 “MineSet 中的空值”一章）。当组中的每个记录为映射为颜色的列而具有空值的时候，该平伸过程的结果颜色为灰色。如果组合中的一个或多个记录为映射为颜色的列而具有非空值的时候，那么那个（或那些）值就用于计算颜色。一个值与空值的和为空值时，则一个值和空值的平均就为该值（也就是说，value + Null = Null； avg（val， Null） = val）。

当空值显示在不同可视化工具显示窗口区域时：“选取窗口”、“选择窗口”或“指针位于”区域上时，则它显示为问号（?）。

对于包含映射到坐标轴上的空值的数字型列，在坐标轴定义的范围之下有一个特殊的空值位置。这有助于显示该空值与其它值不连续。利用“查看”菜单中的“显示空位置”选项，可以关闭数字坐标轴的空值位置。对于映射到坐标轴的字符串值型列，空值（由 ? 代表）被看作是另一个值。

在“地图可视化工具”中，当以下为真时会出现空值：

- 数据库或数据文件包含空值时。
- “工具管理器”根据分组产生数组，并且没有数据落入特定的分组中。例如， 30-40 岁的人口没有数据，该组为空。
- “工具管理器”用于产生数组并且指定了空值枚举选项。在该例中，对于组值为空的情况，创建一个数组成员代表所有值的组合。“工具管理器”给特殊的组赋予问号（?）字符。要查看该组的值，移动相应的滑动条至最左边位置。如果空组中没有数据，与它相关的值也为空，并且“地图可视化工具”也将相应的图形对象表示为“空”对象。
- 空值的表达式和组合可以产生空值（参见“ MineSet 中的空值”）。
- 当空值映射为可视属性时，“地图可视化工具”使用特殊的代表方式。空高度导致 0 高度的深灰对象；空颜色导致了具有适当高度（由映射为高度的值来定义）的对象，但具有深灰的颜色。

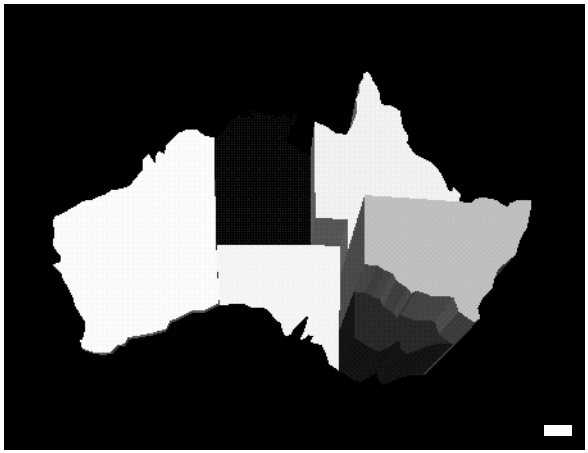

图 1-26 空值映射为高度 （中上部的对象）和颜色 （右下部的对象）的方式

当选择具有空值的对象时，问号 （?）则显示在选择字段中。在 “地图可视化工具” 中映射的空值显示在图 1-26 里，其中空值映射为 “高度” 和 “颜色”。

选项树

“选项树” 是预测模型。它通过利用独立的或已知的属性值来帮助确定标签值或未知属性的方法来进行预测。“选项树” 通过将每个记录分配给类的方式来开始数据分类过程。用于分类的基本结构是决策树，在第 71 页 “决策树” 中有描述。“选项树” 以*选项节点*扩展了普通 “决策树” 模型。“选项节点” 显示了树中决策节点处可供选择的几个选项。“选项树” 一般比普通的决策树更大、更复杂。

导入选项树

从数据中可自动导入 （产生）“选项树” 模型。由记录和一个与每个记录相关的标签所组成的数据称为*训练集*。

文件需求

“选项树”导入工具需要训练集。从数据源 （例如， MineSet ASCII 文件、二进制文件、Oracle、 INFORMIX 或 Sybase 数据库中的 表）中抽取数据可以产生文件。要应用产生的分类器，您应该拥有一个具有分类器所需属性的记录数据集，即除了标签属性外具有所有其他训练集中属性的记录数据集。

创建选项树导入工具

从数据中自动导入 （产生）“选项树分类”工具。这个过程从通过登录到服务器并以平常的方式选择一个数据集开始。

从 “工具管理器”中选择*分类*选项卡，并且从*导入工具*弹出式菜单中选择 “选项树”。除非您希望，否则您不必做更多的限定。只需单击*继续*按钮。“选项树”使用 “树可视化工具”为其显示。

IRIX 中的并行过程

如果您已经安装了 MineSet 的多处理器版本，当分支包含了超过 1000 条以上的记录，在“决策树”中可以并行处理基于树的算法。（参见第 134 页 “在 IRIX 系统中的 “并行计算””。）您可以通过改变 “工具管理器”中 “特性面板”的并行模式来控制线程的数目 （参见 “文件菜单”）。该选项在 IRIX 系统中才可用。

选项树选项

选择 “高级选项”（IRIX 上为 “深层导入工具选项”）可生成导入工具选项对话框。该对话框由四个面板组成：

- 顶部的面板指示了在 “工具管理器”的 “数据目标面板”中所做的选择。
- 从上边数第二个面板可让您设置损失矩阵和加权属性。参见第 111 页 “损失矩阵”和第 13 页 “记录加权”。

- 左下角的面板可让您指定 “高级导入工具” 选项 （描述如下）。
- 右下角的面板可使您指定 “误差估计选项”（除非在 “数据目标面板” 中选择的是仅用于分类器模式，在这种情况下此区域内是空的）。该面板中显示的选项取决于您所选择的 “误差估计” 类型 （参见第 80 页 “错误估计”）。

要微调 “选项树” 导入工具算法，您可以改变在 “决策树” 中的任何选项，这些选项在第 71 页 “决策树”中描述。除此之外，还提供了下面的选项。

- 最大根选项数

  该整数默认为 5，限制了在根上创建的最大选项数目。该导入工具不允许出现全部选项，因为其他的属性可能级别更低。

- 减少值

  该整数默认为 2，定义了在每一层中最大选项数目降低的量。最大根选项数的默认值为 5，这就说明决策树节点的第二层至多有三个选项 （5-2=3）。而决策树节点的第三层将被限制为只有一个选项 （3-2=1）。再向下的层也一样被限制为一个选项。

- 最小拟合比

  该比率确定何时将属性作为选项排除。当导入工具给每个属性一个拟合得分时，则导入工具选择最好的属性以及其它良好的并可作为选项的属性。拟合比率确定了哪些选项应达到什么样的程度才能被选中。比率值 *f* 意味着考虑一个选项，属性评分必须至少达到 （*1-f*） **b*，其中 *b* 为最好属性的得分。拟合比率为 1 则选择所有的属性（所以，如果在该属性上有分割产生，就说明该属性满足了上面描述的限制选项）。拟合比率为 0，则就会形成普通的 “决策树”（无选项节点）。默认值为 0.9。

导入 “选项树” 所需的时间与所创建的选项节点数目有密切关系。因为选项节点一般都在顶部附近创建 （对于可理解性和限制误差来说最有效果），对导入 “选项树” 所需时间的准确估计是所创建没有子选项的选项数目乘以建立 “决策树” 的时间。在默认设置下，根节点最多可以拥有五个的选项，并且每个子节点可以拥有最多三个的选项。则选项总数可达 15 （ 3 乘以 5）。如果最大根选项数增至 6，则选项数目的限制为 48 （6*4*2）； 如果增至 7，则选项数目的限制为 105 （7*5*3）。保持最大根选项数为 5，但将减少量改为 1，则选项的限制为 120 （5*4*3*2）。这样，上例中所期望的导入时间将比普通 “决策树” 所需的时间长两个数量级。降低最小拟合比选项通常导致了比限制因子少的选项，从而减少了导入时间。

在 IRIX 系统 中 “并行计算”

MineSet 的 IRIX 多处理器版本提供了并行处理。它允许将计算密集的任务以并行的方式完成。要运行 64 位版本，需安装并行服务器并从 “工具管理器特性” 菜单中选择并行版本。

“DataMove” 是运行在服务器上的一个进程，提供了对存储在文件中的数据库和数据进行访问的功能，并为挖掘和可视化工具转换数据。

IRIX 6.4 及以上的版本支持大内存 （64 位）模式。如果您有 IRIX 6.2，那么您仍然可以使用 32 位数据挖掘实用程序，但是您必须更新为 IRIX 6.5 以确保获得 64 位支持和 p 线程功能。要想获得 64 位寻址的优势，根据您的系统配置，您也需要改变 **systune** 资源参数。

systune 参数确定了可用系统资源的默认限制。表 1-17 列出了 IRIX 所推荐的 **systune** 参数值 （要得到更多的信息，参见 systune （1M）参考页）。

表 1-17 systune 参数

参数	定义	推荐值
rlimit_pthread_cur	线程数目的当前限制	1024
rlimit_rss_cur	内存使用的当前限制	机器上的物理内存数量
rlimit_vmem_cur	虚拟内存使用的当前限制	您机器上的逻辑交换空间或大约两倍的物理内存
rlimit_nofile_cur	打开文件数目的当前限制	1024 或线程数目的限制

注意： 在设置新参数后，您必须重新启动您的机器。

如果您已经安装了 MineSet 的多处理器版本，当分支包含了超过 1000 条以上的记录，可以并行处理基于树的算法。树上的每个节点可以估计在相应层上最有可能的分割，这些任务可以进行并行处理。默认情况下，一个程序所能容纳线程的最大数量被自动确定。您可以通过改变“工具管理器”中“特性面板”的并行模式来控制线程的数目。并行过程会造成内存碎片，这就会使可用于并行计算的最大数据集的规模小于在单处理器中计算的最大数据集的规模。

预测度

当用关联规则进行工作时，如果您用左侧 （LHS）和右侧 （RHS）作为两个坐标轴将栅格可视化，那么预测度描述了在所有满足 LHS 条件的记录中同时满足 RHS 条件的记录所占的比率，或者是流行度除以 LHS 项出现的频率，（参见第 28 页 “关联规则可视化”可得到 RHS 和 LHS 的解释）。它指出了 X 和 Y 一起出现的频率，是所有出现 X 的记录的一部分。例如，如果预测度为 50%，那么在所有 X 出现的记录中， X 和 Y 一起出现的概率为 50%。因此，如果您知道 X 出现在一个记录中，您可以预计在该记录中 Y 有 50% 的机会出现。例如，如果记录显示了在人们买婴儿尿布和香烟之间相关性中的可预测度为 50%，那么在买婴儿尿布的人中，将会有 50% 的人也买香烟。

流行度

当用关联规则进行工作时，如果您用左侧 （LHS）和右侧 （RHS）作为两个坐标轴将栅格可视化，那么流行度描述了在 RHS 和 LHS 中一起出现的项的频率。（参见第 28 页 “关联规则可视化”可得到 RHS 和 LHS 的解释）。流行度定量描述了全部记录当中，文件中 X 和 Y 一起出现的部分。例如，如果流行度为 1%，X 和 Y 一起出现的记录占总数的 1%。

修剪

修剪是 “决策树” 导入选项，其改变将影响树的高度和 “分割下限”。默认的修剪因子为 0.7。高值指示更多的修剪。参见第 71 页 “决策树”可得到全面的讨论。

随机子

在 MineSet 中有几个选项对话框可以指定随机子。它是指随机数产生算法的起始点。然后，产生的随机数用于数据集的采样中。例如，通过每次采样都使用相同的随机子，您可以确保每次改善模型都使用了相同的数据选择结果。如果您改变了随机子，那么您将从基本数据集中得到不同的选择结果。

记录查看器

“记录查看器” 允许您直接查看您的数据。这就给您一个熟悉它们中的数据值和列的机会。“记录查看器” 也允许您：

- 通过重置大小、重新调整或隐藏它们来进行操作
- 在任意给定列中根据值对数据进行排序或筛选
- 通过多列进行排序 （只适用于 Windows）
- 在排序和筛选之后对行进行重编号
- 查询给定的值
- 将您的操作文件以一定的格式保存

启动记录查看器

有几种方法可以启动记录查看器：

- 从“工具管理器”中：
  - 在“数据目标”面板中单击“可视化工具”选项卡。
  - 单击“记录”选项卡（Windows）或从“工具”菜单（IRIX）中选择“记录查看器”。
- 从“工具管理器可视化工具”菜单中，选择“记录查看器”。然后，使用“记录查看器文件”菜单来打开文件。
- 双击 .schema 文件图标（只适用于 Windows）。
- IRIX 用户可以在外壳命令行提示符下输入该命令：

  `recordview` [*filename*]

对行重编号

“记录查看器”允许您在任意一点对行进行重编号。如果您在排序和筛选之后执行了该操作，那么重编号操作就不能恢复。要回到原始数据，您必须重新打开文件。

要进行重编号，选择“查看”>“重新对行编号”。

在“记录查看器”中查找

“记录查看器”允许您在数据中查找一个值。打开“搜索”面板，选择“查看”>“搜索”面板。要想进行查找，输入值，加亮您想要进行查询的列，然后单击*查找下一个*或*查找上一个*。

保存数据

“记录查看器”允许您保存数据，包括您对数据所做的任何改动。您可以利用“文件”>“保存”或“文件”>“另存为”来保存您的文件。

如果您用了 “保存”，文件将在原始的名字和格式下进行保存。如果这是您第一次保存文件，则它以 MineSet 二进制格式进行保存。“另存为”将带出 “保存数据”窗口，您可以输入所需的文件名以及所希望的格式类型。

利用 “另存为”功能，您可以将文件按照四种格式进行保存 : 二进制、 ASCII、 HTML 或文本格式。当您以二进制或 ASCII 格式进行保存时，同时保存了数据文件和方案文件。 HTML 格式将文件保存为 HTML 表。文本格式将文件以制表符分隔的形式进行保存，列的名字为第一行的内容。

记录加权

在某种实验设计中，真实人口部分比其它部分的采样更频繁。有时，您可能想要得到人口总数的 1% 作为样本，那么采样总数 0.1% 的这一小部分只会产生 0.001% 的样本，这就显得太少 （例如，您可能只得到两个人）。记录加权让您给每个记录一个权重； 这样，被两次采样的人得到的权重为 0.5，而剩下的人被赋予的权重为 1。

另一个例子，电话公司将欺诈电话保存在一个数据集中，而只存储了非欺诈电话的一小部分。通过利用记录加权，可以将总量的真实部分赋给每个记录。

最终，一些数据集已经进行了组合，并且记录具有与它们相关的自然记数 （例如，美国城市的统计数据通常拥有人口的相关记数）。该记数属性可被映射为权重，并与根据记数量来复制每个记录价值。

记录加权的意义是权重为 2 的记录与两个权重为 1 的记录等价。并且允许浮点型的权重。也可参见第 20 页 “加权”。

回归选项卡

回归选项卡允许访问 MineSet 回归工具。在 “工具管理器”中的 “数据目标”面板中，单击 “挖掘工具”选项卡来显示 “回归”选项卡。

回归树

回归工具是一个执行回归操作的预测模型。在已知一套描述性属性的条件下，回归过程的任务是预测连续标签值。回归过程和分类过程相似，其区别是在分类过程中，预计的标签只具有很少量的离散值。

当回归工具产生时， MineSet 也产生可视化过程。该可视化过程有助于您理解回归工具以及它是如何进行预测的。除此之外，它还能够提供数据本身有价值的直观结果。一旦产生，回归工具就可用于为未标记的记录预测标签值。

导入 “回归树”

使用了称为导入工具的算法后，就可以从数据中自动的导入 （产生）“回归树”回归工具。如果使用了除 “仅用于回归器”以外的任何选择，数据都会被分割形成训练集。训练集由数据库中的记录组成，而数据库中的连续标签也已知。例如，您可以提供带有列（例如，年龄、教育、职业、每周工作小时数等等）的表，而且其中一列包含描述性的属性 （收入总量）。通过登录到服务器并以平常的方式选择一个数据集后就可以开始了。

Windows 用户：

“工具管理器”中，从 “挖掘工具”下选择*回归*选项卡。

IRIX 用户：

从 “工具管理器”中选择*回归*选项卡， *导入工具*菜单自动显示了 “回归树”。

除非您希望，否则您不必做更多的限定。除了标签值是连续的而不是离散的以外，该选择与 “决策树”的完全一样。只需单击 *调用工具*或*运行*按钮。“回归树”使用 “树可视化工具”为其显示。

连续标签

“连续标签”菜单提供了连续标签的列表。该列表包括了表现为数值型值的属性。选择您想要进行模型化操作的标签属性。例如，要为预测总收入产生回归工具，选择“总量收入”。如果没有连续属性，菜单显示*没有连续标签*，并且*继续*按钮被禁用。只有连续属性才能产生回归工具。如果数据集中没有包含连续属性，您可以利用“工具管理器数据转换”面板添加一个新的连续列。

回归树选项

选择*高级选项*（IRIX 中的*深层导入工具选项*）可生成“回归选项”对话框。该对话框由四个面板组成：

- 顶部的面板指示了在“工具管理器”的“数据目标面板”中所做的选择。
- 从上面数第二个面板可指定“损失矩阵”，参见第 111 页 “损失矩阵”，并设置加权属性，参见第 200 页 “加权”。
- 左下角的面板可让您指定其它“导入工具”选项（描述如下）。
- 右下角的面板可使您指定“误差估计选项”（除非在“数据目标面板”中选择的是“仅用于分类器”模式，在这种情况下此区域是空的）。该面板中显示的选项取决于您所选择的误差估计类型（也可参见第 80 页 “错误估计”）。

要微调“回归树”导入算法，您可以改变下列“回归树”导入工具选项。

- 限制树高度

  默认情况下，“回归树”中的高度（层的数目）并无限制。通过单击复选框并输入上限值以限制高度。限制层的数目可以加快导入过程，并且不需要分解太多的节点就可以研究“回归树”。虽然限制规模会减少运行时间，但会增加错误率。设置该选项并不影响在最大层之前的层上所选的属性。

- 拆分标准

  在树导入过程中，该选项允许您指定用于在竞争属性分割中选择的标准。对于回归树， MineSet 支持四种拆分标准：

  - 方差

    这就造成 “回归树导入工具”选择那些能使树中每点处的节点内方差最小的分割。当产生叶子时，在叶子处的预测是符合叶子的记录标签值的平均。方差是数据集中每个值之间的平方差，除以成员或值的总数。该模式是统计中最常用的。

  - 绝对偏差

    这就造成导入工具选择那些能使树中每点处的节点内绝对偏差最小的分割。当产生叶子时，在叶子处的预测是符合叶子的记录标签值的中值。

  - 标准化方差

    与方差类似，这是一个分割标准，但对于多向分割会产生偏差。

  - 标准化绝对偏差

    就象绝对偏差，这是一个分割标准，但对于多向分割会产生偏差。

  如果存在问题，很难说哪种标准是最好的。对每个都进行试验，并选择最低误差估计或最容易理解的那个 “回归树”。

- 拆分下限

  这就是一个权重的下限 （或者，如果权重没有设定，就为记录的数目），而且必须至少存在于两个子节点中。该选项的默认值为 5。例如，如果节点中有一个三向分割，三个子节点中至少有两个其权重必须为 5 或更多。这就提供了限制 “回归树”规模的另一种方法。

  提高拆分下限是为了增加概率估计的可靠度，因为每个叶子的记录数是巨大的。这也就将创建更小的树并降低导入时间。如果您估计数据中包含了噪音 （错误或异常），可提高分割下限。如果数据集很小 （ < 100 条记录），您可以降低分割下限。

- 代价复杂度 - 修剪

  代价复杂度修剪通过互换树的误差率 （它的代价）和树上叶子的数目 （它的复杂度）试图产生形状优化的树。在代价复杂度计算过程中中，训练集被分割为学习集和修剪集。学习集被用于长成一个修剪树。该树被修剪后产生出一系列具有递减复杂性的树。修剪集然后被用于在序列中找出代价最小的树。最小代价树的大小被标出。学习和修剪集被合并并长成一棵树。该树然后被修剪为最小代价树的大小。

  代价复杂度修剪参数允许您选择比最小代价树更小的树。该参数指出了标准差的值比可以接受的最小代价树所花费的代价更大。将该参数设为 0 以选择最小代价树；设置参数为 0.5 以选择最小的树，其误差率不大于 0.5，标准差比最小代价树差。默认设置为 0，选择最小规模的树 , 其误差率不大于标准误差率但又差于最小代价树。高值指示更多的修剪。如果数据可能包含了噪音 （误差和异常），增加该值来创建更小的树。如果树被修剪成为一个节点，降低值以降低修剪量并更多地显示树的结构。

  修剪过程比限制树高度或增加分割下限过程的速度要慢，这是因为要创建整个树然后再进行修剪。然而，有选择的进行修剪将形成更准确的回归器。

回归工具中的误差估计

当评估分类工具时，自然矩阵是错误的 （分类工具预测了错误标签样例的数目）。当提供了损失矩阵，不同类型的分类错误会有不同的相关代价。在这种情况下，损耗是一个自然度量。

对于回归过程，其任务是预测一个实数值，该值没有自然评估矩阵。经常使用的两种度量是*均方差*以及*平均绝对误差*。在 “均方差”中，使用了预测标签值和实际标签值之间的平均方差。在平均绝对误差中，使用了预测标签值和实际标签值之间差的绝对值的平均。

回归器名称

产生的回归器以阶段文件 （在 “工具管理器”中确定）的前缀命名，而后缀为 *-rt.regress*，例如 *churn-rt.regress*。默认情况下，所有的回归工具存储在服务器的 *file_cache* 目录中，在 Windows 中默认为 *MineSet 文件*，在 IRIX 中默认为 *mineset_files*。Windows 系统在当前工作目录中将可视化过程存储为三维条形图图标。这些文件有 *-rt.treeviz* 后缀。

这些回归工具可用于为未做标记的数据集预测标签。要应用已存贮的回归器，您需要根据回归器所使用的属性选择记录数据集。该回归器将为每个记录预测新的连续标签值。参见第 15 页 “应用模型”和第 33 页 “修正”。

删除列

当您需要简化可视化过程或模型时，您可以从当前工作中删除列，而不将它们从数据集中删除。在 “工具管理器”的 “数据转换窗格”中，从 “当前”列文本窗格中选择一列并单击*删除列*按钮。这样就从您的计算中，但并未从数据集中删除该列。

投资回报曲线

投资回报 （ROI）曲线与上升曲线相似，但它以损失的形式而不是以误差的形式显示准确度；并将所使用的损失矩阵考虑在内。

可以从任何导入工具的 “高级选项”面板的 “工具管理器”对 ROI 曲线进行配置，只需使用 “误差估计选项”部分下的复选框。对于回归工具此功能不可用。

ROI 曲线中的每个点按照每个记录的期望损失进行排序，也就是那些由所选标签值标记的点。很相似，曲线中每个点的高度指示了累计利润 （与损失相反），而不是到该点为止的所有记录的累计准确度 （与误差相反）。

在所选的标签列条件下 （参见第 111 页 “损失矩阵”）“损失矩阵”中的成员乘以分配给由分类工具产生的类的概率可以计算期望损失。因此，如果分类工具对于其预测很确定，那么期望损失就较低，并且记录将出现在 ROI 曲线的左边。

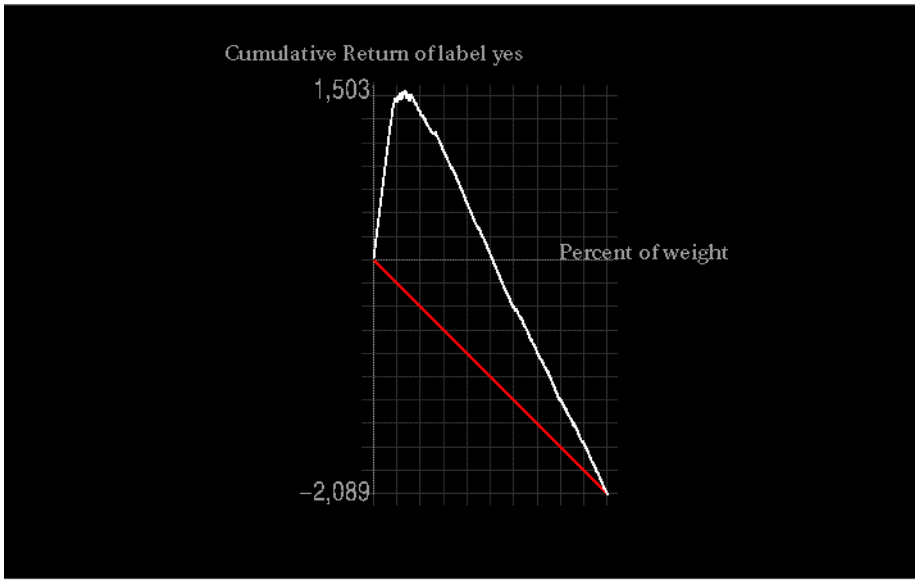

图 1-27 投资回报曲线

ROI 曲线意义在于，用户对于数据集中的每个记录都应采取措施。该措施与所选择的标签值相关。例如，在客户波动数据集中，与标签 “是”相关的措施，也许会向人们提供一些市场资料。这也许会阻止人们波动；但是如果不加区别的进行该措施的代价是昂贵的。如果分类工具用于预测是否向特殊的人发送信函，那么 ROI 曲线的顶点就显示在检验集上必须保留大约多少钱。

当填入损失矩阵并与 ROI 曲线一起使用时，应特别小心。在一定预测标签下的列确定了标签值的结果 ROI 曲线。在所有可能的类上，该列中的成员需要代表采取与该标签值相关的措施所得到的期望增益和损耗。例如，客户波动过程中，在列 “预计是”以及在 “实际值为不”行下的成员可能包含一个等于 2 的值来指示邮寄给将要客户波动的人小册子 （该措施与 “是”有关）的费用为 $2。另一方面，在列 “是”和行 “是”下的成员可能有一个为 -10 的值来指示，如果防止了用户的客户波动将为公司节省用于邮寄的费用 $10。

保存文件

您可以保存您正在工作的阶段部分，方法是使用 “工具管理器文件”下拉式菜单，或通过单击 “工具管理器”的 “数据目标”窗格中的 “数据文件”选项卡来保存特殊转换。当您想恢复工作时，从 “工具管理器”“文件”下拉式菜单中装入该阶段或文件，则文件和历史表都在那里恢复。

样例文件目录

MineSet 提供一系列样例文件供用户使用。在前面的部分中进行了详细的说明。平台不同目录也不同。

Windows 用户可以在安装 MineSet 的目录的 *\examples* 下找到样例文件。

IRIX 用户可以在 */usr/lib/MineSet/examples* 中找到文件。

散点可视化工具

利用 “散点可视化工具”可直观的分析几个变量的关系从统计上或者通过动画过程。在没有大量记录 （少于 50,000）的条件下，对于观察单个数据点或者将大量的记录合并为不同组合结果的小集合时，它是很有用的。如果数据集有大量的记录，则可考虑使用 “平伸可视化工具”。在 “散点可视化工具”中进行分析可使用 :

- 三维景观
- 包含一个二维滑动条的动画控制面板
- 图形对象被称为*实体*，在三维景观中可以进行动画过程。在动画过程中，实体的位置，颜色和大小都可以改变。

“散点可视化工具”通过将数据集中的每个记录，或行映射为三维景观中的实体的方式来对数据进行可视化。数据中的变量可以映射为实体的大小，颜色以及位置。同样，您也可以将一或两个数值型变量映射到动画控制面板中的滑动条。如果映射为实体的大小，颜色和位置的变量取决于映射为滑动条的变量，则用该滑动条可执行动画过程。例如，数据可代表几个公司一段时间的销售。如果时间映射为滑动条而销售变量映射为大小，那么就可以实现实体随时间滑动条增大或缩小的动画。

在创建了数据的可视化过程后，“散点可视化工具”可用不同的方法分析数据。您可利用动画控制面板对动画路径进行一维或二维的跟踪。通过回放所创建的动画路径，您可以观察到实体大小、颜色和运动的趋势或异常变化。在三维景观中，您可以对显示结果定向来强调特殊的维或角度。“散点可视化工具”可以对变量值进行放大以对其进行特别强调。同样，您也可以筛选显示结果，只显示那些符合标准的实体。

文件需求

“散点可视化工具”需要下列文件 :

- 数据文件由制表符分隔的字段行组成。利用 “工具管理器”可以很容易的创建这个文件。如果您正在产生该文件，参见 *《MineSet 3.0 企业版接口指南》*中按所需格式“为散点可视化工具创建的数据和配置文件”。

  您可通过从数据源 （例如数据库）中抽取数据来产生数据文件，并将它按照 “散点可视化工具”的需要进行格式转化。数据文件拥有用户定义的扩展名 （“散点可视化工具”中提供的文件具有 *.data* 扩展名）。

- 配置文件描述了输入数据的格式以及如何进行显示。“工具管理器”可以创建该文件，或者您可利用自己喜欢的文本编辑器 （例如，WordPad、jot、 vi 或 Emacs）来产生该文件 （参见 *《MineSet 3.0 企业版接口指南》*中的 “为散点可视化工具创建数据和配置文件”）。

  配置文件必须有 *.scatterviz* 扩展名。当启动 “散点可视化工具”，或当打开文件时，您必须指定配置文件，而不是数据文件。

启动散点可视化工具

有几种方法可启动 “散点可视化工具”：

- 在 “工具管理器”中的 “可视化工具”选项卡上运行 “散点可视化工具”。参见第 147 页 “配置散点可视化工具”可得到与 “散点可视化工具”配合使用 “工具管理器”的更多信息。
- 利用 “工具管理器”从 “可视化工具”菜单中启动 “三维可视化工具”。（参见本书的 “工具管理器”条目以及 *《MineSet 3.0 for Windows 企业版用户指南》* 中有关 “工具管理器”功能的部分，该部分对所有的 MineSet 工具都是公用的。）
- 如果您知道要使用哪个配置文件，双击配置文件的图标。这样就启动了 “散点可视化工具”并自动装入您所指定的配置文件。这只有当配置文件名以 *.scatterviz* 结尾时才起作用 （对于 “用工具管理器”为 “散点可视化工具”创建的配置文件总是如此）。
- 从 IRIX 命令行中启动 “散点可视化工具”，可以输入：

  `scatterviz` [*configFile*]

  *配置文件*是可选的，它指定了所使用的配置文件的名字。如果未指定配置文件，那么您必须用 “文件” > “打开”来指定一个。

配置散点可视化工具

- 在 Windows 系统中使用工具管理器：

  在 “工具管理器”的 “数据目标”窗格中，单击 “可视化工具”选项卡并选择 “散点”。“散点可视化工具”面板显示了您可以映射为列的各种成员。在每个文本字段右边的弹出式菜单中选择一列。每个弹出式菜单中的可用选择被限制为那些合适类型的列。

- 在 IRIX 系统中使用工具管理器：

  在 “工具管理器”的 “数据目标”窗格中，单击 “可视化工具”选项卡；在 “工具”弹出式菜单中选择 “散点可视化工具”。显示星号的可视成员需要映射。从 “当前列”中选择一列并映射为右边窗格中的成员。

- 用 “文本编辑器” 创建一个 “配置文件”：

  虽然 “工具管理器” 大大简化了配置 “散点可视化工具” 的任务，但您也可以利用文本编辑器为该工具手工创建配置文件。参见 *《MineSet 3.0 企业版接口指南》* 中 “为散点可视化工具创建数据和配置文件”。

为 “散点可视化工具” 创建滑动条

当数据集中的列独立变化时，例如，当数据在一个时段中变化时，为 “散点可视化工具” 创建滑动条是有用的。滑动条是创建动画的先决条件，可以手工或自动创建。参见 *《MineSet 3.0 for Windows 企业版用户指南》* 中的 “滑动条创建” 部分可得到更多的信息。

散点可视化工具选项

要在 “散点可视化工具” 中改变不同的选项，返回 “工具管理器” 的 “数据目标面板”，并选择 “可视化工具” 选项卡，然后选择 “散点可视化工具”。单击 *工具选项* 按钮可以生成一个新对话框。在这里您可以改变 “散点可视化工具” 中某些选项的默认值。

“散点可视化工具” 对话框有四个基本选项块：

- 实体
- 滑动条
- 坐标轴
- 其它

实体选项

该选项组可让您指定控制 “散点可视化工具”图形显示的实体的外观特征：

- *实体图例开关* — 可让您确定是显示还是隐藏实体图例。
- *实体尺寸* — 可让您将实体按比例放大到最大尺寸，比例尺寸或者默认尺寸 （不调节）。您可以指定实体大小的图例是显示还是隐藏。
- *实体颜色* — 可让您控制实体所显示的颜色。您可以：
  - 指定使用的颜色列表
  - 指定映射类型
  - 将颜色列表映射为值列表
  - 指定颜色图例是显示还是隐藏
  - 映射颜色到实体
- *实体形状* —可让您为实体选择一个可视代表物：立方体、条形、球形或四面体。
- *实体标签颜色* — 您可以通过在上面单击来修改标签的颜色。这样就出现了 “颜色选择”对话框，利用它可以完成颜色的改变。
- *实体标签尺寸* —控制了实体标签的大小。较小的数字会减小规模，而较大的数据会增大规模。

要使用 “颜色”选项，您必须将一列映射为 “数据目标”面板中的实体颜色。要得到如何选择和改变颜色的更详细的解释，参见第 48 页 “颜色选择”。

*颜色列表*可让您利用颜色列表标签旁边的 + 按钮来指定颜色列表。这就会产生一个颜色编辑器，您可以利用它指定要加入列表的颜色。

*颜色映射*可让您指定在图形显示中颜色的变化是*连续的*还是*离散的*。如果您选择了*连续*，作为在*颜色映射*字段中被映射为颜色的值的一个功能，颜色值在*颜色列表*字段中输入的颜色之间逐渐的变化。

弹出式按纽右边的字段可让您指定映射为该颜色的值。如果您不指定任何映射值，那么就会使用颜色变量值的范围。参见第 48 页 “颜色选择”可得到关于选择颜色的更多信息。

汇总选项

汇总滑动条允许在一或两个附加变量上进行动画过程。滑动条上的每个位置都有一个颜色，该颜色与映射为总和的变量组合值相对应。“汇总”选项可让您指定在 “汇总”窗口中显示的变量应使用什么颜色。您也可以指定是显示还是隐藏汇总图例，该汇总图例用于指示这些值是什么。要得到动画的更多信息，参见第 10 页 “动画”。

如果您有一个数组，您可指定 X 或 Y 滑动条。这些选项旁边的弹出式按钮提供了可用关键字的列表，您可以指定将哪个用做滑动条。

滑动条选项

“滑动条”选项控制了滑动条映射的解释过程。要得到更多的信息可参见第 148 页 “为 “散点可视化工具”创建滑动条”。

坐标轴选项

坐标轴选项可让您为每个坐标轴指定：

- 标签 （如果该框为空，“散点可视化工具”就默认地使用每个坐标轴的列名）。
- 颜色
- 每个坐标轴的尺寸类型 （可以为*最大尺寸*，*比例尺寸*，或*不调节*）。
  - *最大尺寸*可指定对于一个指定的尺寸坐标轴可以独立地进行缩放。如果一个坐标轴具有最大尺寸，两倍于另一个坐标轴，则不管数据值的大小它都是另一个轴的两倍长。当使用不同的坐标单位的坐标轴进行比较时，该选项最为有用 （例如，将收入和年龄比较）。对于非数值型的数据该选项不起作用。

  - *比例尺寸*可在坐标轴最大值的基础上对其进行缩放。如果两个坐标轴具有相同的“比例规模”，但是其中一个具有最大值并且是另一个的两倍，则前一个是后一个的两倍长。当使用相同的坐标单位坐标轴比较时，该选项是非常有用的（例如，收入与支出）。该选项对于非数据型坐标轴起作用。
  - *不调节*与*比例尺寸*为 1.0 等价。
- 尺寸值
- 该坐标轴是否应延伸以包括 0 值。

其它选项

在对话框底部的“其他选项”，包含了下列字段：

- *消息*允许您指定当实体被选中时所显示的消息。要列出和描述用于输入字段的格式型式，参见 *《MineSet 3.0 企业版接口指南》*里的为“散点可视化工具创建数据和配置文件”中的“消息语句”部分。
- *执行*该选项可让您输入在双击一个实体时执行的 UNIX 命令。格式与消息语句类似。如果不出现执行语句，双击也不起作用。要得到执行字段更详细描述，参见 *《MineSet 3.0 企业版接口指南》*里的为“散点可视化工具创建数据和配置文件”中的“消息语句”部分。
- *隐藏标签距离*控制了实体标签变为不可见的距离。较小的距离可以改善性能，但标签很快就消失。数据越高，标签隐藏的距离越大。
- *坐标轴标签尺寸*控制了坐标轴标签的大小。较小的数字会减小尺寸，而较大的数据会增大尺寸。
- *栅格（X，Y，Z）尺寸*可指定各自坐标轴栅格线之间的间隔。较小的数字会减小尺寸，而较大的数据会增大尺寸。如果指定了 0，那么就不画出栅格线。
- *栅格颜色*通过在颜色块上面单击可改变栅格的颜色。这时将出现“颜色选择器”对话框，利用它可以完成颜色的改变。

重设工具选项

如果想要将所有的选项恢复成默认值，单击*重设*选项。

保存新工具选项

一旦您完成了对 “工具选项”对话框的修改，单击*确定*将返回 “工具管理器”主窗口。

动画控制面板

动画控制面板，将出现在主窗口的右边，由最多带有两个滑动条的汇总窗口，信息字段，动画按钮和动画滑动条组成。参见第 10 页 “动画”。

在散点可视化工具中的空值处理

当具有未知数据值或空值的字段映射为可视化属性时，“散点可视化工具”使用特殊的表现方式。（对于空值的讨论，参见 *《MineSet 3.0 企业版接口指南》*的 “MineSet 中的空值”一章。）当空值映射为实体尺寸时，该实体被显示成立方体的轮廓。当空值映射为实体的颜色时，实体显示为深灰色。当空值显示在 “选择窗口”或 “指针位于”的区域时，它显示为问号 （?）。

如果空值映射为实体的位置 *x*， *y*，或 *z* 时，则结果取决于 “查看”菜单中的 “空位置”（参见第 197 页 “查看菜单”）。如果设置了该选项，则坐标轴上带有空位置的实体显示在对应坐标轴范围内。如果没有设置该选项，那么就不显示带有空位置的实体。

样例配置和数据文件

我们提供样例数据和配置文件来展示 “散点可视化工具”的特征和能力。这些文件的详细描述可参见附录 A，配置和数据文件样例。

“选项”菜单

“选项”菜单对所有工具来说是相似的；不适合某种特殊工具的功能不在菜单中显示。“选项”菜单可以对基本数据进行追溯。要执行追溯，首先选择一个或多个实体或平伸，然后在两种追溯方法中选择一种对选定的记录进行追溯。

- *创建选择框*创建三维选择框，它可以伸展或转换为选择区域的体积范围。当选择框被激活时，一个“记录查看器”格式的表格将被打开，用以显示所有选中实体所代表的组合数据的信息。关闭窗口就删除了选择框，而不取消它的选择。框中的任何实体以及按下 Shift 键选择的实体将在表窗口中显示。要改变选择框，用鼠标左键单击它的一面，并将它朝需要的方向拖动。拖动时按住 Shift 键会限制向离运动方向最近的坐标轴运动。要改变选择框的范围，可朝需要的方向拖动灰色标记。在体积界限以外调整大小或进行移动是不允许的。灰色标记会自动调整大小来保持恒定的屏幕尺寸。如果有时它们显得太大，您可以靠近选择框，它们将根据框的大小减小它们的尺寸。
- *显示数值*弹出一个窗口显示被选中的实体。
- *显示原始数据*检索并显示了与已被选中实体相关的记录。结果记录显示在一个表查看器中。如果什么也没被选中，该选项被禁用 （变灰）。
- *送到工具管理器*在当前 （框）选择的基础上，在“工具管理器”历史的起点插入一个筛选操作。用于追溯的精确表达式由当前在主窗口中的选择来确定。如果什么也没被选中，该项被禁用 （变灰）。
- *互补追溯*可使用*显示原始数据*和*发送到工具管理器*选择，取出所有没被选中的数据。

- *追溯列*带出了一个面板，您可以选择用于追溯的列。不象其它可视化工具，在数据中没有特殊列，被指定为数据的关键字。可视化工具不可能确定在追溯表达式中用户想要的列。例如，您可能拥有有关轿车的品牌，型号以及重量的数据。您可能想追溯这些原始数据，并确定需要考虑品牌和型号，而不考虑重量。默认情况下，所有被映射为图形必备成分的列在追溯中都被认为是有意义的。其它的列则不是如此，但您可以通过在 “追溯列”对话框中将它们加亮来达到这样的效果。

要得到追溯的更多信息，参见第 78 页 “追溯”。

为地图可视化、散点可视化、平伸可视化创建的滑动条

您可以将一个列映射为滑动条来显示它根据指定的标准是如何变化的。如果列是数字型的 （整型、浮点型和双精度型）或分组的它就可以被映射为滑动条。如果列已经被分组，那么它在名字后将带有 `_bin`。在 “当前列”字段中列的名字之后标注出了列的类型，例如：日通话的总数 - 双精度型。

从 “工具管理器”的 “数据转换”窗格里的*当前列*中，通过将数字型列映射为一个或两个滑动条，您可以为 “地图”，“散点”以及 “平伸”可视化工具自动地创建滑动条。参见第 13 页 “动画按钮和滑动条”。

列名排序

您可以不改变数据集，而按照字母顺序对列名进行排序以便于参考。从 “工具管理器”中的 “数据转换”窗格中，单击*按顺序显示各列*复选框 （Windows），或 *排序列名字*按钮 （IRIX）。

平伸可视化工具

利用 “平伸可视化工具”可以从统计学角度或者通过动画过程直观的分析几个变量的关系。对于具有大量记录的数据集，它是特别合适的。对于量不太大的记录，如果您想观察各个数据点，可以选择 “散点可视化工具”。要进行数据分析，可使用

- 三维景观
- 包含一个二维滑动条的动画控制面板
- 被称为 *平伸*的图形对象，代表了数据点的组合。在动画过程中可以改变平伸的颜色和透明度 （而不是位置或大小）。

“平伸可视化工具”可以通过将列映射为坐标轴、滑动条、颜色和透明度来将数据可视化。所得到的三维景观可以被认为每个数据点都被分散画出的散点图的近似。它并不是真正的散点图，因为聚在一起的数据点 （属于相同的组）被合并，并以单一的云团状平伸画出。

每个被映射为坐标轴或滑动条的数字型列首先必须进行分组。如果跳过了该分组步骤，“工具管理器”会用自动统一分组实现该步骤 （参见第 34 页 “分组”）。字符型列可直接映射到坐标轴上。任何数字型列可以映射成颜色。对于属于同一组的所有数据点，通过取映射到颜色的列的平均值，可生出平伸的颜色。平伸的透明度是根据同一组数据点数目的加权而定的。如果没有数据点映射为透明度，那么记录记数就用于确定透明度。可视化结果的交互性独立于数据点数目所代表的意义；它只取决于轴维度上分组的数目。

如果您的数据集很大，在工具管理器中进行组合。服务器会处理该过程，而不用将所有的数据送入客户端并在那里进行组合。参见第 4 页 “组合”。

在动画控制面板中，最多能有两个数字型列可以映射为滑动条。当动画面板中的滑动条沿着其路径从一点移向另一点时，平伸在动画过程中改变了其颜色和透明度。不象 “散点可视化工具”，平伸的位置和大小都不发生变化；它们在固定的，统一间隔的位置上。只有它们的颜色和透明度在变化，造成实际运动的幻觉。

如果字符串列映射到坐标轴上，那么分组被定义为该列的相异值。通过对映射为颜色的列的值进行平均，沿字符串坐标轴的值的顺序可通过对这些平均值的排序来自动确定。沿字符串值坐标轴观察颜色的变化，可看出该列与映射为颜色的列之间相关性。如果没有颜色，那么透明度变量会被用于确定顺序。

平伸可视化工具透明度

映射为透明度的列应该是记录记数或用于对记录进行加权的列。平伸透明度， α 根据下列关系建立：

$$\alpha = 1 - e^{-u \cdot weight}$$

*权重*是映射为透明度的列 （或者如果没有映射为透明度的列，则为记录记数）。当*权重*的值变大时，该函数的形状为透明度渐进于 1 （完全透明）。当您调节主窗口左边的透明度比例滑动条时， *u* 是用于比例变化的。图 1-28 显示了该函数当 *u* 为高值或低值时形状。图 1-29 显示了 u 为低值或高值时相同的可视化过程。

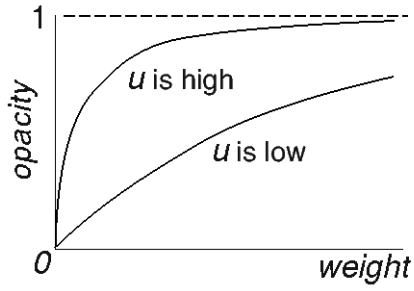

图 1-28 当 *u* 为高值或低值时透明度函数的形状

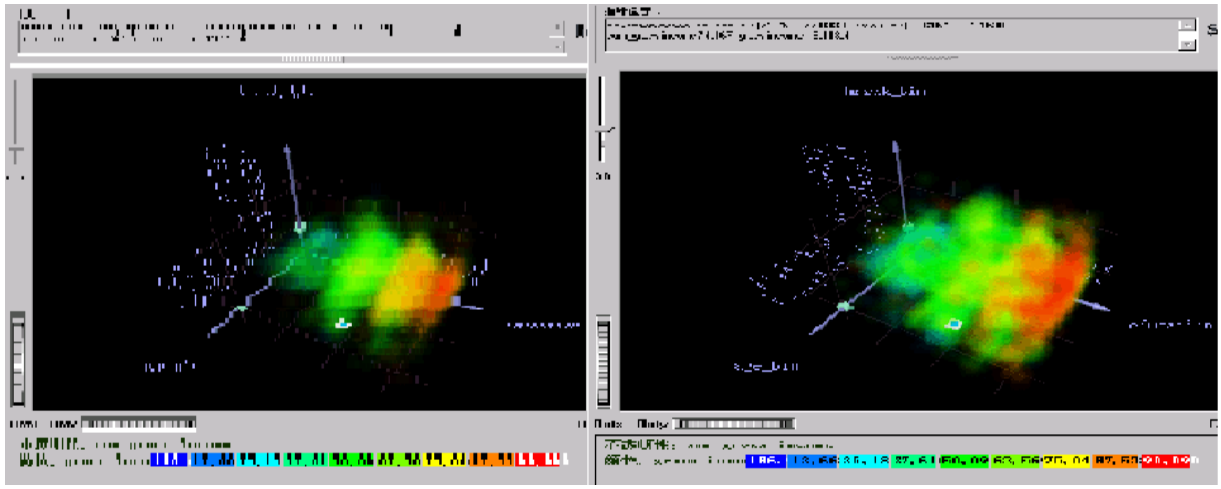

图 1-29 当 $u = 5.3$，和 $u = 30$ 时的图像

如果没有记录映射为透明度，“散点可视化工具”产生了一列在组合时产生记录记数的列。这就意味着所有的记录都进行相等的加权。在该列中进行了组合过程，当坐标轴和滑动条列进行分组时，在映射为颜色的列上进行了平均组合过程。所有其它列都是不必要的并且将被删除。您不需要将任何东西映射为透明度，除非您想对每个记录进行加权，而且其权重不为 1。

关于在“工具管理器”而不是在客户端进行处理的信息，参见 *《MineSet 3.0 for Windows 企业版用户指南》* 中的 “平伸可视化工具”的 “处理技术”部分。

当您调用了该工具，所有的过程将在服务器上进行，而数据文件， *adult94.splatviz.data*，包含了原始数据中进行了行组合的行。

在一些情况下，您可能想要一个对记录进行加权的列。例如，如果您有一个数据集其中一列为人口，另一列为平均工资 （您想映射为颜色），您可以将人口映射为透明度，平均工资映射为颜色；然后用 “平伸可视化工具”进行组合。该组合过程通过坐标轴和滑动条列进行分组，因此，它将求透明度列 （在这种情况下为人口）的总和。新列称为 sum_population。 average_salary 列被修订了，它仍然是平均工资，但经过了每行中人口值的加权。这样，平均工资列仍然显示了它所代表的所有人的平均工资。

作为选择，由于数据集规模的原因，如果您想避免客户端的处理和存储，您可以按照下面所示在 “工具管理器”中执行同样的组合过程：

1. 创建一个新列，定义 temp = population * avg_income。
2. 执行组合过程：按照坐标轴和滑动条列进行分组，求组合人口的总和并且求组合 temp 的总和。
3. 创建一个新列，定义为
   avg_salary = sum_temp/sum_population
   这就会产生加权平均。
4. 现在您可以将 sum_population 映射为透明度，将 avg_salary 映射为颜色。

注意，如果您没有在 “工具管理器中显式地做这些工作，那么，这些步骤就由 “平伸可视化工具”自动进行的。然而，如果您在 “工具管理器”中执行了它们，会比在服务器上进行要更有效，因为这样会形成非常小的能够被客户端检索的文件。

平伸可视化工具文件需求

“平伸可视化工具”需要下列文件 :

- 数据文件由制表符分隔的字段行组成。利用 “工具管理器”可以很容易的创建这个文件。如果您正在产生该文件，参见 *《MineSet 3.0 企业版接口指南》*中的为 “平伸可视化工具创建数据和配置文件”获取所需的文件格式。

  您可通过从数据源 （例如数据库）中抽取数据来产生数据文件，并将它按照 “平伸可视化工具”的需要进行格式转化。数据文件拥有用户定义的扩展名 （“散点可视化工具”提供的样本文件具有 *.data* 扩展名）。

- 配置文件描述了输入文件的格式以及如何进行显示。“工具管理器”可以创建该文件，或可用您所喜欢的文本编辑器来产生该文件（参见 *《MineSet 3.0 企业版接口指南》* 中的为 “平伸可视化工具”创建 “数据”和 “配置文件”文件一章）。

  配置文件必须有 *.splatviz* 扩展名。当启动 “平伸可视化工具”，或当打开文件时，您必须指定配置文件，而不是数据文件。

启动平伸可视化工具

有几种方法可以启动 “平伸可视化工具”：

- 利用 “工具管理器”来配置和启动三维可视化工具。在 *《MineSet for Windows 3.0 企业版用户指南》* 中有详细的描述。
- 从 “工具管理器”的 “可视化工具”下拉式菜单中选择 “平伸可视化工具”。通过选择 “文件” > “打开”来打开配置文件。
- 如果您知道要利用的配置文件，双击配置文件的图标。这样就启动了三维可视化工具并自动加载指定的配置文件。只有当配置文件以 *.splatviz* 结尾时才有效（对于用 “工具管理器”为 “平伸可视化工具”创建的配置文件总是这样）。
- 在 UNIX 命令行提示下，敲入：

  splatviz [*configFile*]

  *配置文件*是可选的，它指定了所使用的配置文件的名字。如果未指定配置文件，那么您必须用 “文件” > “打开”来指定一个。

调用 “平伸可视化工具”的 IRIX 选项

-quiet 选项删除弹出以指示过程的对话框。要使该选项永久可用，可通过加入下列行：

```
*minesetQuiet:TRUE
```

到您的 *.Xdefaults* 文件。

Windows 用户可以通过 “文件” > “特性”菜单达到同样的效果。

平伸可视化工具形状选项

“平伸”选项可以为 “平伸”指定一定量的特征，然后 “平伸可视化工具”将进行图形显示。

- *平伸颜色*—可控制用于平伸的颜色。您可以
  - 指定所用颜色列表
  - 指定映射类型
  - 将颜色列表映射为值列表
- *平伸形状*—可选择用于画出平伸的方法之一：线性形、高斯形、纹理形、球状形、立方体或四面体。参见第 16 页 “形状菜单”可得到更深入的解释。

要使用 “颜色”选项，您必须将一列映射为 “数据目标”面板中的颜色。如果在颜色列表中什么也没有输入，则使用默认的颜色图。默认的颜色图是一个从蓝 （最低值）到红（最高值）的连续谱。要得到如何选择和改变颜色的更详细的解释，参照第 48 页 “颜色选择”。

颜色列表您可以利用颜色列表标签旁边的 + 按钮来指定颜色列表。这就会产生一个颜色编辑器，您可以利用它指定要加入列表的颜色。

颜色映射您可以指定在图形显示中所表现的颜色变化是*连续*或 *离散的*:。如果您选择了“连续”，作为在*颜色映射*字段中被映射为颜色的值的一个功能，颜色值在*颜色列表*字段中所输入的颜色之间逐渐的变化。

弹出式按钮右边的字段可让您输入映射为该颜色的指定值。如果您不指定任何映射值，那么就会使用颜色列中值的范围。参见第 48 页 “颜色选择”可得到关于选择颜色的更多信息。

汇总选项

汇总选项可让您指定在 “汇总”窗口中使用什么颜色。只有当您将列映射为汇总时才可用。

其它选项

在对话框底部的“其他选项”，包含了下列字段：

- *隐藏标签距离*控制了坐标标签（对于字符串值坐标轴）变为不可见时的距离。增加该数值可以让标签在更远的距离上出现。数据越高，标签隐藏的距离越大。
- *坐标轴标签尺寸*控制了坐标轴标签的大小。较小的数字会减小尺寸，而较大的数据会增大尺寸。
- *栅格颜色*通过在上面单击可改变栅格的颜色。这样就出现了“颜色选择器”对话框，利用它可以完成颜色的改变。
- *栅格（X，Y，Z）尺寸*可指定各自坐标轴栅格线之间的间隔。较小的数字会减小尺寸，而较大的数据会增大尺寸。如果规模设置为 0，则在该维中就没有栅格线。

重设工具选项

单击*重设选项*将所有选项的值设为默认值。

保存平伸可视化工具设置

当您按下*调用工具*，“工具管理器”在几个文件里保存了“平伸可视化工具”的信息，所有文件都具有相同的前缀：

- < *前缀* >.*splatviz.data* 包含了数据。
- < *前缀* >.*splatviz.schema* 描述了数据文件。
- < *前缀* >.*splatviz* 包含“平伸可视化工具”所需的信息。

与当前工具选项一起保存整个对话，从“工具管理器文件”菜单中使用下列菜单选项之一：

- *保存当前对话 ...* 默认前缀建立在数据源的基础上
- *另存当前对话 ...* 指定您自己的前缀

保存文件的前缀为 <*prefix*>.*mineset*，并且包含了 MineSet 返回当前状态所需的所有信息。

当您使用了*调用工具*后，如果需要， *.data*、*.schema* 和 *.splatviz* 文件将被更新。

在平伸可视化工具中的空值处理

当具有未知数据值或空值的字段映射为可视化属性时，“平伸可视化工具”使用了特殊的表示方式。（关于空值的讨论，参见 *《MineSet 3.0 企业版接口指南》* 的 “MineSet 中的空值”一章。）当组中的每个记录映射为颜色的列具有空值的时候，该平伸过程的结果为灰色。如果组合中的一个或多个记录为映射颜色的列具有非空值的时候，那么那个（或那些）值将被用于计算颜色。一个值与空值的和为空值，一个值和空值的平均就为该值 （也就是说， value + Null = Nul ； avg （val， Null） = val）。

当空值在正文处出现时，它显示为问号 （?）。（“选择窗口”与 “指针位于”区域在它们自己的部分中讨论。）

对于包含映射到坐标轴上的空值的数字型列，在坐标轴定义的范围之下有一个特殊的空值位置。这有助于显示该空值与其它值不连续。利用 “查看”菜单中的 “空位置”选项，可以关闭数字坐标轴的空值位置。对于映射到坐标轴的字符串值型列，空值 （由？代表）被看作是一个值。

为平伸可视化工具创建的滑动条

在主窗口旁边的汇总窗口附近出现的滑动条数目依赖于在配置文件中指定的滑动条映射。任何带有鉴别标签的滑动条的出现，取决于数据集是有两个、一个还是滑动条映射。

映射为滑动条 1 和滑动条 2 上的列最终形成滑动条的索引值。这些列要么是数字型的（整型、浮点型、双精度型），要么被分组。如果映射为滑动条的列已经被分组了，该列就不需要自动分组了，并且该列已被当作滑动条的索引使用。然而，如果该列未被分组，利用自动统一分组过程就可以创建一个分组的列。（参见第 34 页 “分组”可得到更多的信息。）在形成自动分组中用到的列从当前表中被删去。

动画控制面板

动画控制面板，出现在主窗口的右边，由最多带有两个滑动条的汇总窗口，信息字段，动画按钮和动画滑动条组成。参见第 10 页 “动画”。

在每个单独的滑动条位置上 （由汇总窗口的黑点指示）画面对应于内存中数据表。要在一维滑动条上进行插值，两个相邻表被合并，然后将空间列作为特定键进行组合。当滑动条从一个分组位置移向另一个时，则对每个平伸的权重 （映射为透明度）进行插值（如果表缺少特定行，则假定权重为 0）。用于平伸的平均值也进行了插值，但只是按照权重进行加权。

样例 1-4 插值过程

该样例描述了插值过程的技术细节。假设我们想要显示一个图象，该图象代表了外部滑动条之上在 40-50 岁表和 50-60 岁表之间的插值。表 1-18 和表 1-19 分别为年龄 =40-50 和年龄 =50-60 的表，是两个滑动条位置。

表 1-18 年龄在 40 到 50 之间

教育程度	职业	工作时间	收入	加权
HS-grad	Exec-Man.	15-25	25000	2
HS-grad	Mach-op	15-25	30000	1
硕士	技师	25-35	35000	3

表 1-19 年龄在 50 到 60 之间

教育程度	职业	工作时间	收入	加权
HS-grad	Exec-Man.	15-25	70000	1
职业的	Mach-op	35-45	40000	2

下面就是 “平伸可视化工具” 进行插值的过程。对于表 1-18，添加了一个等于 *（1-t）weight* 的新列以及一个等于 *(1-t) (weight)(valu e* 的权重新列。对于表 2，添加了一个等于 *（t）（weight）的新列，*以及一个等于 *（t）（weight）（value）*的权重新列。两个表被合并。

通过以空间坐标轴为键，将两个表合并，并汇总了两个新列的组合结果。这就确保了对于所有的空间坐标轴，不存在具有相同分组值的两行。最后，将总和值除以总和权重来得到内插值。在这种情况下，该内差值是相对 *收入*而言的。如果 *t*=.5，结果表将为表 1-20。

表 1-20 在表 1 和表 2 之间插值的之间过程。

教育程度	职业	工作时间	收入	加权
HS-grad	Exec-Man.	15-25	40000	1.5
HS-grad	Mach-op	15-25	30000	.5
硕士	技师	25-35	35000	1.5
职业的	Mach-op	35-45	40000	1

如果外部查询滑动条是两维的，则使用双线性插值。

该普查数据集包含了将近 150,000 行。外部滑动条的目的是在其中漫游，并显示数据中附加维的汇总信息。红色的区域代表了总和值高的地区；白色区域显示了值较低的地区。当滑动条停在黑点上时，则图象显示未插值的数据。可以跟踪滑动条的轨迹，并且利用滑动条下的 VCR 控制面板显示动画。

要显示动画是如何形成的，假定您具有 8 年的数据， 1990-1997 （也就是说，在汇总窗口中有 8 个数据点）。可以通过检查平伸如何随滑动条从一年移向另一年而进行变化的方式开始。假定 1990 年中，在给定的位置平伸的值为 20 （被映射为颜色）并且权重为 2，这就意味着它代表了 2 条记录。进一步假定 1991 年 中，同样的平伸的值为 40 权重为 200。

1991 年 中的平伸比 1990 中的更加透明，这是因为它代表了更多记录的组合结果 （或者具有大得多的权重的记录）。当您将年滑动条从 1990 移到 1991 时，权重则按照 2 到 200 的线性插值结果进行变化。

通过计算经记录权重加权所得两个值的平均可计算出该值。例如，在 1990 和 1991 之间，权重为 101，并且该值为 ((1-.5)*2*20+.5*200*40)/((1-.5)*2+.5*200) = 39.8。当您接近 1992，其大小也接近 40。

在离散点之间您不可能停止动画过程，并且您不能将路径滑动条拖到离散数据点之间的静止位置。汇总窗口中的数据点代表了与数据文件中实际数据相对应的滑动条位置。例如，值 20 和 40 代表了实际值的组合，而 39.8 则不是。

在平伸可视化工具中的下拉式菜单

五个下拉式菜单可以让您访问 “平伸可视化工具”的附加功能：它们为 “文件”、“查看”、“选项”、“形状”以及 “帮助”。它们的描述可参照第 92 页 “文件菜单”，第 197 页 “查看菜单”，第 15 页 ““选项”菜单”，以及第 97 页 “帮助 （IRIX）”等条目。

形状菜单

在该工具中使用的平伸是用来对小云团进行模型化的 （参见 *Proceedings of SIGGRAPH ，90*， Vol.24， No.4，第 367-376 页 *Lee Westover 著的* “Footprint Evaluation for Volume Rendering”）。

“形状”菜单可改变绘制平伸的方法。您可以选择交互式地交换准确度。在每次逼近估计中，纹理平伸是理想高斯型密度的最准确表达。因为大多数计算机可较好地支持硬件辅助纹理化，因此纹理平伸通常是最好的选择。在 SGI 平台中，只有 Indy 或更早的系统受到软件执行慢的速度限制。三种平伸类型是：

- *线性形* 绘制了少量的三角对高斯型平伸进行线性估计。
- *高斯形* 绘制了大量的三角估计高斯型平伸。
- *纹理形*利用映射为矩形的纹理给出最准确的近似。在不支持硬件辅助纹理映射的机器上该过程是很慢的。

有选择的，下列透明原始的方法是允许的。

- *球形*绘制了透明球，其直径随
  权重的立方根 （或权重）而变化。
- *立方体*绘制了一个立方体，其宽度随记数 （权重）的立方根而变化。
- *四面体*绘制了一个三角框架，其大小随记数 （权重）的立方根而变化。

配置和数据文件样例

有一些由 MineSet 提供的样例数据和配置文件，它们用来展示 “平伸可视化工具” 的特征和功能。每个文件的细节描述参见附录 A，配置和数据文件样例。

拆分下限

拆分下限是一个用于改进 “决策树导入工具” 和 “回归树导入工具” 的选项。增加它会产生更小的树，但却会降低准确度。

拆分下限是权重上的下限 （通常情况下，如果权重没有设定，就为记录的数目），必须至少存在于两个子节点中。该选项的默认值为 2。例如，如果节点中有一个三向分割，三个子节点中至少有两个其权重必须为 2 或更多 （如果没有设置权重则为两条记录或更多）。这就提供了限制 “决策树” 规模的一种方法。

因为每个叶子的记录数是巨大的，所以提高分割下限是为了增加概率估计的可靠度。这也将创建更小的树并降低导入时间。如果您认为数据包含了噪音 （错误或异常），或者在估计概率时使用了树 （参见第 15 页 “应用模型”），则将分割下限增至 5 或更多。如果您的数据集很小 （<100 条记录），您可以减小该数到 1。参见第 71 页 “决策树”。

拆分标准

该选项提供了 “决策树”中的三个拆分标准。下面是技术性定义。如果存在问题，很难说哪种标准是最好的。可以全部进行试验，然后选择产生最低误差估计或最容易理解的那个 “决策树”。

*公共信息*是指父节点和子节点纯度加权平均之间纯度的变化 （也就是*熵*）。加权平均是根据每个子节点上记录的数目来进行的。

标准化公共信息 （默认情况下）是公共信息除以子节点数目的以 2 为底的对数。

*增益比*是指在忽略标签值的条件下，交互信息除以拆分的*熵*。

*标准化公共*信息和*增益比*将优先权赋给了带有少量值的属性。

提供给 “回归工具”的选项确定了在回归树中列的选择。

*方差*确定了在回归树中列是怎样进行选择的。方差选择的列可以产生使节点内方差最小化的拆分。选择方差将产生的回归树，其叶节点处具有平均预测。

*绝对偏差*选择的列可以产生使节点内绝对偏差最小化的拆分。“选择绝对偏差”将产生的回归树，其叶节点处具有中值预测器。

标准化方差 （默认情况下）为方差除以以 2 为底的子节点数目的以 2 为底的对数。

*标准化绝对偏差*为绝对偏差除以以 2 为底的子节点数目的以 2 为底的对数。

统计可视化工具

“统计可视化工具”，可以从 “工具管理器”的 “数据目标”窗格中的 “可视化工具选项卡”上进行访问，它显示了带有一系列小面板的窗口，每个在 “工具管理器”窗格的*当前列*中列出的列中分配一个。只有有限个列面板可以同时显示，使用边滚动条，或者平行或垂直拉伸 “统计可视化工具”窗口，可以看到更多列面板。

在将数据集中一定数量的记录输入 “工具管理器”的基础上，利用 “统计可视化工具”您可以看到某些统计结果。列面板的格式根据列类型以及该列内存在的相异值的数目而变化。列通常分为两种类型：*数值型*和*离散型*，分别显示为棒状图以及直方图。

如何读取统计可视化工具

棒状图根据*数值型*列画出，每个数值型列由整型、浮点型、双精度型或日期型值组成。每个棒状图面板显示了单一列数据的统计结果，包括这些数值型值的最小值、最大值、平均值、中值以及两个四分之一值 （25% 和 75%）。这些值显示为横穿具有绿色底纹的垂直棒形的线，并且数量的标准差显示为 +/- 值。只有少于 50,000 个值的时候，才会显示四分之一值。（参见图 1-30）。如果列中的相异值超过 50,000 条 ，则统计结果显示为灰色的条。

平均值为列数据的和除以记录的总数。中值为在给定列中按照大小顺序排列的数的中间数。标准差是列中数据的离散度的度量。

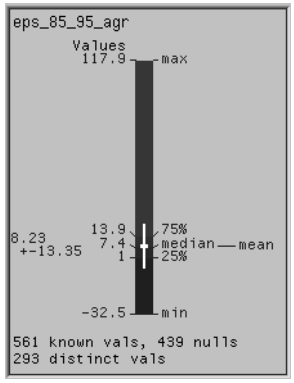

图 1-30 统计可视化工具显示的数字型列

直方图用于绘制*离散型*（或*标称型*）列的，具有非数值型（字符、分组或枚举型）值（参见图 1-31）。离散型列面板最多显示 100 个相异值以及每个相异值对应的实例数目的直方图。离散行的默认顺序是降序排列，但是您可以利用“查看”下拉式菜单来选择另一种排序方式。如果只有 100 个或更少的列目，列面板可以包含相异值的记数。

只要是显示离散值就可以使用直方图，例如，yes/no 值或州名。不论棒状图或直方图，将数据集中记录的数目，以“总数”表示，而该特殊盒子中的所代表相异记录的数目以“相异值”表示。

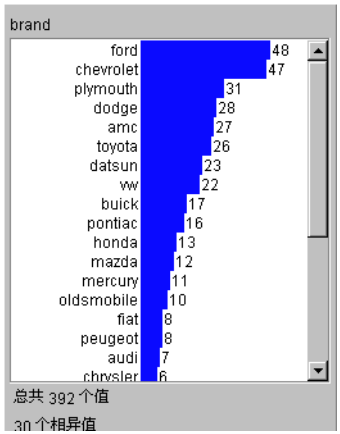

图 1-31 统计可视化工具显示的离散列

如果您从图标中启动了 “统计可视化工具”，则在主窗口中只有 “文件”和 “帮助”菜单能够使用。要想让主窗口显示所有的菜单和控制，请打开一个 *.statviz* 文件。利用 “文件” > “打开”可以看到可用配置文件的列表。

统计可视化工具下拉式菜单

三个下拉式菜单可以让您访问 “统计可视化工具”的附加功能：它们标记为 “文件”、“查看”和 “帮助”。如果没有指定配置文件就启动了 “统计可视化工具”，那么只有 “文件”和 “帮助”菜单可用。参见 “文件”和 “帮助”条目。

统计可视化工具的查看菜单

“统计可视化工具”中的 “查看”下拉式菜单对直方图和棒状图进行分类：

- *按记数值排列标称*指定，标称 （离散）列显示为依照每值实例记数进行降序排列的直方图。
- *按字母顺序排列标称*指定按照字母顺序进行排列。

历史表按钮

MineSet 允许您使用一系列操作来转换数据表。这一系列转换被记录下来，并且可以通过 “工具管理器”“数据转换”面板下部的*历史表*按钮来回溯特定的转换。利用这些按钮您可以看到转换过程中的每一步，并且当您出错时可以返回。当您单击了向左箭头按钮，列窗口显示在前一步骤中出现的表。单击向右箭头按钮则返回到当前表状态。

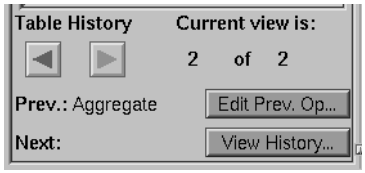

图 1-32 “历史表按钮”的 “当前操作视图为”字段

“当前视图为”字段

在历史表按钮的右边是信息字段*当前视图为*，它记录了您所做的转换次数并且指示了您正在查看的步骤。该字段中的两个整数指示了您正在观察转换序列中的哪一步，以及现存总步骤数。例如，如果您做了两次转换，您可以查看原始表 （总数 3 中的第 1 个），在第一次变化之后的表（总数 3 中的第 2 个），或第二次变化之后的表（总数 3 中的第 3 个）。

“上一个”和 “下一个”按钮

当您利用*历史表*按钮向前或向后查看以前的转换时，*上一个*和 *下一个*字段 （在箭头按钮之下）有助于您定位在历史表中的位置。对于您所查看的任何一个表，*上一个：*字段告诉您前一个转换是什么，而 *下一个：*字段告诉您下一个转换。

“编辑前一个操作”按钮

*编辑前一个操作*按钮允许您编辑在*前一个*字段中显示的操作。（在*当前视图为：某数中的第一个时*，该按钮变灰，因为这是指原始表，以前没有发生过转换。）当您单击 *“编辑前一个操作”*按钮时，则弹出有关以前操作的对话框，您可以对该转换进行改变。例如，如果前面的转换为列分组，当您单击*编辑前一个操作*时，出现*列分组*对话框。

通过改变以前的转换，您可以影响对当前表建立的后续转换操作。例如，如果您删除了在后续分组操作中用到的列，则分组操作就变为无效。*编辑历史*按钮有助于您避免这样的问题。

删除操作 至 终 点 按 （只适用于 Windows）

当您在历史中做备份时，“工具管理器”的 “数据目标”面板 （在主窗口的右边）被禁用 （变灰），这是因为您不在历史的末尾。为使 “数据目标”面板变为有效，您必须向前一直到历史的末端，或者单击*删除操作至终点*按钮删除当前视图后所做的全部操作。

操作历史选项卡

当您单击了*操作历史*选项卡 *（在 IRIX 上的查看历史*按钮*）*，显示当前列和数据目标的面板被一个显示*数据转换*表全部历史的面板所代替 （图 1-33。）表现为小的矩形表的每个表格都包含了列的列表，并通过一个大的矩形图表 （指示了在表上执行的操作）与下个表格相连。

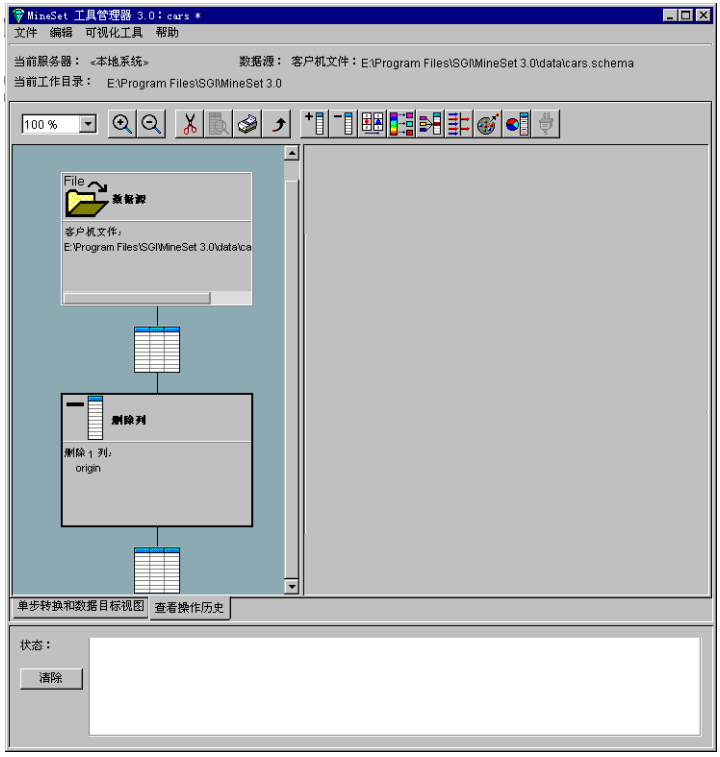

图 1-33 查看历史对话框 （Windows）

Windows 版本中的工具条上的图标可帮助您用 “工具管理器”对数据集进行检查和执行不同的操作。单击左边窗格的一个操作，然后单击*单步转换和数据目标查看*选项卡（*IRIX 上的查看单步*按钮）会呈现出处于选定操作阶段的 “工具管理器”（参见图 1-33 和
图 1-34。）单击*操作视图历史*选项卡 （在 IRIX 上为*视图历史*按钮）将返回以前的视图。

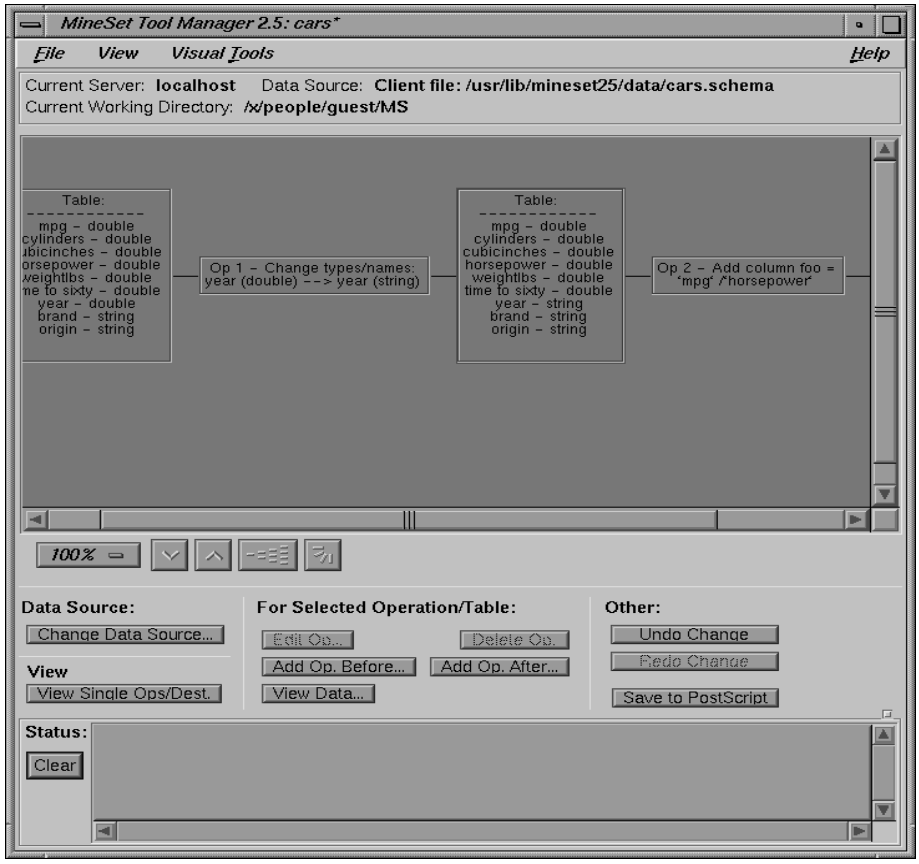

图 1-34 查看历史对话框 （IRIX）

利用*编辑前一个操作*改变一个操作通常影响了 （有时无影响）历史中的后继操作。您可以选择一个特定的操作来编辑、添加或查看。当编辑操作影响了历史，“视图历史”对话框将警告您，并显示新的历史。

在 IRIX 版本下 “视图历史”面板中的图表窗口下面的一排按钮允许您改变图表的方向和大小。

工具管理器

“工具管理器”是用户用于指定配置文件，数据文件和使用工具的图形界面。在该部分中讨论了 “工具管理器”的一般操作。关于如何具体使用，可参考 *《MineSet 3.0 for Windows 企业版用户指南》*。

在 MineSet 客户端可运行工具管理器。该过程遵从下列路径：

1. “工具管理器”打开一个 DataMover 的连接，该连接运行在 MineSet 服务器上。在一些情况下，该服务器可以是您客户端的工作站，而在其它情况下则是不同的机器。
2. “工具管理器”可以指定
   - 数据库和表或在客户端或服务器上包含数据的二进制或 ASCII 文件
   - 应用挖掘或可视化工具
   - 通过工具选项，设定数据如何显示
   - 阶段文件保存了您工作的历史

   经 “DataMover”返回的信息用于指导相互作用。作为结果，“工具管理器”产生了配置文件。该文件包含了用于执行后序步骤的用户定义参数。
3. “工具管理器”将步骤 2 的配置文件的拷贝转发给了 “DataMover”。
   “DataMover”通过以下途径处理文件
   - 通过访问数据库或文件
   - 执行指定的数据转换
   - 当需要时运行挖掘工具
   - 当需要时生成可视化处理文件

   这些可视化处理文件包含 MineSet 工具可读入的以指定格式保存的数据。然后，这些可视化处理文件的一份拷贝就转给了 MineSet 客户端。
4. “工具管理器”调用合适的 MineSet 可视化工具。
5. 该工具访问可视化处理文件并显示了数据。
6. 如果您产生了一个模型，该模型可应用于新的数据。

工具管理器特性

“工具管理器特性”对话框允许您设置下列选项：

- *启动时自动恢复*当您登录 MineSet 时，允许您返回上次使用的最后阶段。 MineSet 保存历史，并以您上次离开时相同的状态打开了文件。
- *使用二进制数据*文件告诉 MineSet 使用二进制文件，这样可以减少处理时间。
- *最大属性值数*可让您设置数据集中属性值的截断数。在计算中将不使用具有更多单一值的任何列。
- *并行过程*可让您为 IRIX 的并行过程设置选项。

训练集

训练集是一个包含属性的表，其中之一被指定为类标签。该标签是您将试图预测的属性。

该样例的目标是在蝴蝶花属性萼片长度、萼片宽度、花瓣长度以及花瓣宽度已知的条件下预测花 （iris-setosa， iris-versicolor 或 iris-virginica）的类型。图 1-35 显示了样例训练集中的几个记录。

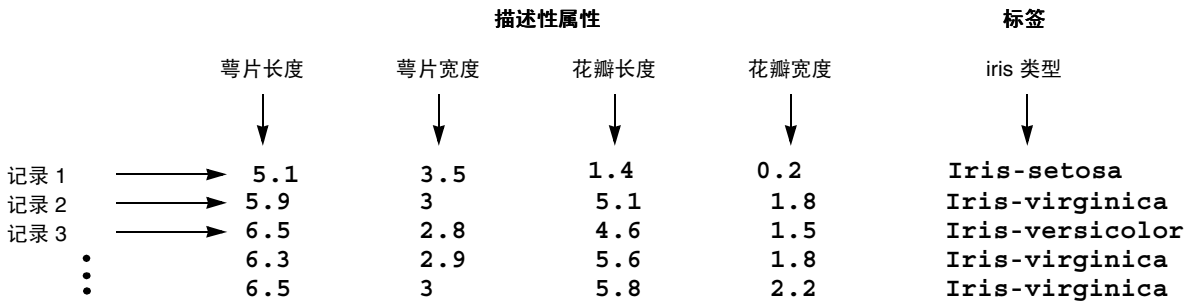

图 1-35 训练集中的记录样例

一旦建立了模型，就可以为新记录预测标签值。这些新记录必须存在于拥有模型所用的全部属性的表中，并具有与这些属性在训练集中相同的名字和类型。该表不需要包含标签属性。如果该属性存在，那么在预测的过程中就忽略掉。

树可视化工具

“树可视化工具”是用来在三维场景中显示数据的图形界面，它将您的数据表现为等级块（节点）和伴随圆盘物的条形图，您可以动态的漫游、查看数据集的部分或全体。如何使用“树可视化工具”在 *《MineSet for Windows 3.0 企业版用户指南》* 中有详细的讨论。

“树可视化工具”通过将数据表现为以等级形式相互连接的节点来展现数据定量和相关的特征。每个节点都包含了其高度、颜色和圆盘位置与数据值组合相对应的条形图。与节点相连的边（显示为线）显示了各部分数据与其子集之间的关系。

在次一级组中的值可以进行汇总并在相邻的更高的层中自动地显示。条形图下面的基准块可以提供所有条形图的组合信息。代表负值的条形图显示在基准块顶部之下，通过使基准块高度无效，您可以更清楚的观察负值条形图（参见第 19 页“树可视化工具显示菜单”，或 *《MineSet 3.0 企业版接口指南》* 中为“树可视化工具”创建“数据”和“配置文件”里的“基准块高度说明”部分。

文件需求

“树可视化工具”需要下列文件：

- *数据*文件由制表符分隔的字段行组成。利用“工具管理器”可轻易地创建该文件，参见 *《MineSet for Windows 3.0 企业版用户指南》* 中的描述。如果您要自己生成该文件，参见 *《MineSet 3.0 企业版接口指南》* 的以所需的文件格式为“树可视化工具”创建“数据”和“配置文件”。

  数据文件拥有用户定义的扩展名（“树可视化工具”中提供的样例文件具有 *data* 扩展名）。

- *配置*文件描述了输入数据的格式以及它们如何转化为等级结构。利用“工具管理器”可轻易地创建该文件。参见 *《MineSet for Windows 3.0 企业版用户指南》* 中的描述。您也可利用自己喜欢的文本编辑器（例如，WordPad、jot、vi 或 Emacs）来产生该文件，参见 *《MineSet 3.0 企业版接口指南》* 中的“为树可视化工具创建数据和配置文件”。

配置文件必须具有 *.treeviz* 扩展名。当启动 “树可视化工具”，或打开一个文件时，指定的是配置文件而不是数据文件。

启动树可视化工具

有几种方法启动 “树可视化工具”：

- 使用工具管理器，参见 *《MineSet for Windows 3.0 企业版用户指南》*。
- 从 “工具管理器” 的 “可视化工具” 下拉式菜单中选择 “树可视化工具”。通过选择 “文件” > “打开” 来打开配置文件。
- 如果您知道要利用的配置文件，双击配置文件的图标。这样就启动了三维可视化工具并自动加载指定的文件。只有当文件名以 *.treeviz* 结尾才有效 （对于用 “工具管理器” 为 “树可视化工具” 创建的配置文件总是这样）。
- 从 IRIX 命令行中输入：

  `treeviz` [*configFile*]

  *配置文件*是可选的，它指定了所使用的配置文件的名字。如果未指定配置文件，那么您必须用 “文件” > “打开” 来指定一个。

您可以使用一个警告系统并在调用工具时禁用对话框，参见第 199 页 “警告选项”。

树可视化工具选项

单击*工具选项*按钮可以生成新对话框。（图 1-36 和图 1-37）。这就允许您改变 “树可视化工具” 中一些选项的默认值。

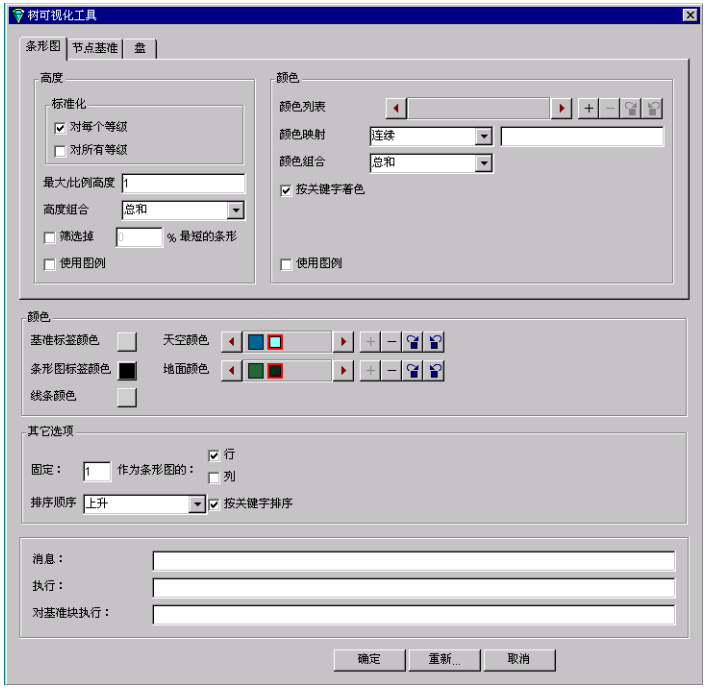

图 1-36 树可视化工具配置选项对话框 （Windows）

在 Windows 版本中，对话框的顶部显示了三个选项卡：*条形图节点基准块*以及*圆盘*。每个选项卡展示了一系列选项，您可以对可视化过程进行细致的配置。要得到关于选择颜色的更多信息，参考第 48 页 “颜色选择”。要指定每个对话框的 “高度” 部分，参见第 180 页 “标准化高度”。

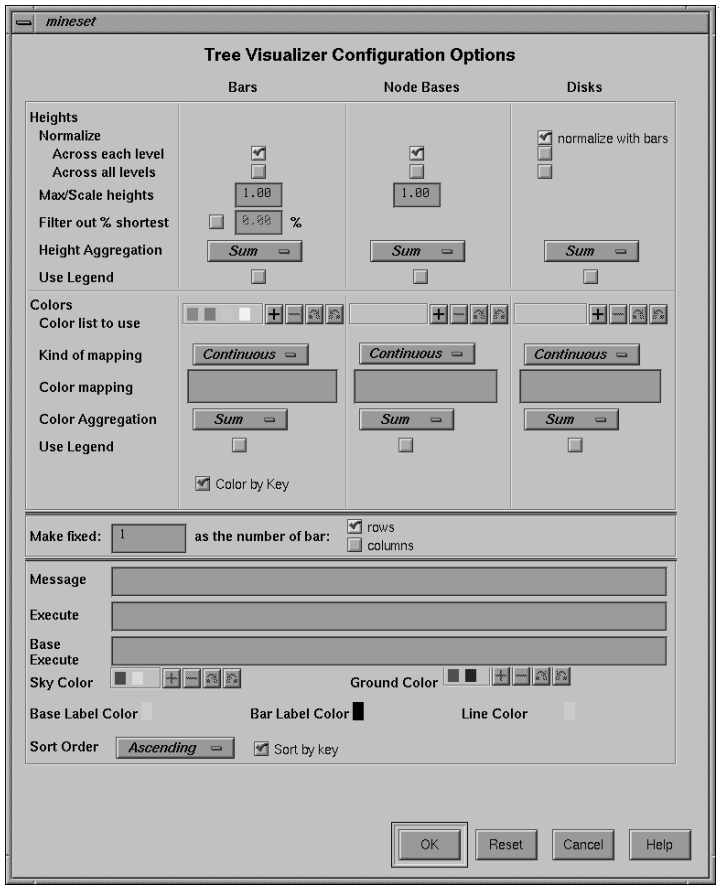

图 1-37 树可视化工具配置选项对话框 （IRIX）

在 IRIX 版本中，对话框的顶部显示了三个选项卡：*条形图节点基准块*和*圆盘图*。在这个对话框中可以对所有的选项进行访问。

标准化高度

该选项可使您对所有层次中的每个层上条形图、节点基准块以及圆盘图的高度进行标准化。标准化高度确定了高度变量的最大值；它标准化了与该高度相关的所有值。这样，如果最大值为 30.0，并且最大条高度设为 1.0 （任意单位），那么大小为 15.0 的值将映射为值 0.5。

层内标准化独立地标准化了等级的每一层。如果数据是按等级汇总的，该选项最为有用，并且防止了等级最顶层将底层的项目过度压缩。在所有层上进行地标准化将所有的层一起标准化，而不管在等级中的位置。如果对条形图没有进行任何选择，也就不会有标准化过程。

节点基准块的标准化独立于 “条形图”。如果未加特定选择，用于条形图的标准化方法也用于节点基准块，虽然其值的标准化独立进行。

如果存在薄圆盘并且选择了*与条形图一起标准化*，则圆盘与条形图结合在一起进行标准化：代表了相同值的圆盘和条形图具有相同的高度。如果在 “圆盘”列中选中了其它标准化选项，则圆盘的标准化独立于条形图：不管最高的圆盘和最高的条形图所代表的实际值，它们都具有相同的高度。

最大 / 比例高度

该选项可让您指定最高条形图和节点基准块的高度。默认值为 1.0 （任意单位）。如果在看到视图以后，您觉得高度太低或太高，使用该字段来进行调节。例如，在字段中输入 2 可以使所有条形图的高度变为原来高度的两倍；输入 0.5 使所有条形图变为原来的一半。

如果指定了标准化过程，该值代表了最高条形图或基准块的高度。如果没有指定标准化过程，则所有的值都要用该均值进行比例缩放。当对两个不同的数据集视图进行比较时，后者可能会有用。

筛选掉 最短的 %

该选项可让您筛选掉只包含短条形图的节点。首先，计算了画面中最高的条形图 （如果高度按照层进行了标准化，那么就为每一层中最高的条形图）。那么只显示那些至少包含一个条形图的节点，该条形图高度相当于最高高度的给定百分比。例如，如果您在字段中输入 5%，那么就会显示至少包含一个高度不低于最高高度 5% 的条形图的节点 。（这些条形图的祖辈节点也显示）。这一选项可以粗略的筛选掉小的、不感兴趣的节点。这并不是找出带有一定值的特殊节点的精确原理。使用该选项可以加快缓慢而复杂的画面的再现，或减少由许多接近 0 高度的条形图所产生的杂乱。

虽然筛选出了小的节点，但在等级上的任何聚合中它们仍然要进行记数。

高度组合

默认情况下，父节点条形图的高度是所有子节点条形图高度的汇总；但是这些高度可以为均值、最大值、最小值、计数或任何出现的值。这些值可用于条形图高度、基准块高度以及圆盘高度值。

颜色

利用这一套选项可以

- 指定所用颜色列表
- 指定映射类型
- 将颜色映射为条形图、节点基准块以及圆盘图

要使用这些 “颜色”选项，您必须已经将一个属性映射到 “数据目标”面板所需的 “颜色 - 条形图”、“颜色 - 圆盘”或 “颜色 - 基准块”。要得到如何选择和改变颜色的更详细的解释，参照第 48 页 “颜色选择”。

*使用的颜色列表*可让您利用颜色列表标签旁边的 + 按钮来指定颜色列表。这就会产生一个颜色编辑器，您可以利用它指定要加入列表的颜色。

*映射类型*可让您指定在图形显示中颜色的变化是*连续的*还是*离散的*。如果您选择了 “连续”，颜色值 （条形图、节点基准块或圆盘的）在 “颜色列表”中所输入的颜色之间逐渐的变化。如果您选择了 “离散”，颜色只在指定界限内变化。

颜色映射 让您指定颜色所映射的值。

颜色组合

默认情况下，父节点条形图的颜色是所有子节点条形图值的汇总；但是这些颜色可以为均值、最大值、最小值或任何出现的值。该组合可用于条形图颜色、基准块节点颜色以及圆盘颜色。

根据关键值来确定颜色

该选项可让您通过它们的关键值来自动地对条形图上色。如果指定了另一个上色方法，则忽略该选项。如果您未指定颜色列表，或指定的颜色不充足，则随机抽取一些附加颜色；如果指定了额外的颜色，则它们被忽略。

行列固定

默认情况下，在一行上放置了所有的条形图。该选项允许改变列或行的数目。如果既没有选中行又没选中列或者该数置为 0，那么无论行或列都不固定，并且显示了最为接近正方形的近似行列数。

消息

该选项可让您敲入您想要的任何消息。消息语句指定了当指针在一个对象上移动或当一个对象被选中时所显示的消息。默认情况下，对于基准块使用与条形图相同的消息。如果未指定任何消息，则使用包含了所有列名称和值的缺省消息。

消息的格式必须与正在使用的数据类型相匹配：

- 字符串必须使用 %s。
- Ints 必须使用整型格式 （象 %d）。
- Floats 和 doubles 必须使用浮点格式 （象 %f）。

要得到消息字段的详细描述，参见 *《MineSet 3.0 企业版接口指南》* 中为 “树可视化工具”创建 “数据”和 “配置文件”一章里的 “消息声明”。

双击执行和基准块双击执行

这些选项可让您敲入一个当在条形图或基准块上双击时执行的命令。如果只填写 “执行”字段，则它既应用于条形图也应用于基准块。如果字段都被填写，则 “执行”应用于条形图，而 “基准块执行”应用于基准块。格式与消息声明一样。如果不出现执行声明，双击也不起作用。

要得到 “执行” 字段的详细描述，参见 *《MineSet 3.0 企业版接口指南》* 中为 “树可视化工具” 创建 “数据” 和 “配置文件” 一章里的 “执行声明”。

天空颜色

您可以指定一种或两种颜色。如果只指定了一种颜色，则天空为纯色的。如果指定了两种颜色，则天空在两种颜色之间渐变。当指定了两种颜色，则第一种为天空顶部的颜色，而第二个为底部的颜色。

地面颜色

您可以指定一种或两种颜色。如果只指定了一种颜色，则地面为纯色的。如果指定了两种颜色，则地面在两种颜色之间渐变。对于地面，第一种为远处水平线的颜色，而第二种为近处地面的颜色。

基准块标签颜色

您可以指定基准块前面的标签颜色。

条形图标签颜色

您可以指定条形图前面的标签颜色。

线条颜色

您可以指定与基准块相连的线的颜色。

排序顺序

如果您选择了 *按照关键值排序* 复选框，所显示的节点按照即定顺序排列。在复选框旁边的菜单可指定按照升序排列还是按照降序排列。

重设工具选项

在对 “工具选项” 对话框作出修改以后，如果您想要将所有的设置重设为默认值，单击 *重设选项* 按钮。

保存新工具选项

一旦您完成了对 “工具选项”对话框的修改，单击*确定*将返回 “工具管理器”主窗口。

保存树可视化工具设置

“工具管理器”将 “树可视化工具”的信息存储在几个文件中，并具有相同的前缀：

- *<prefix>.treeviz.data* 包含数据。
- *<prefix>.treeviz.schema* 描述数据文件。
- *<prefix>.treeviz* 包含了 “树可视化工具”所需的信息。
- *<prefix >.mineset* 包含了创建其它文件所需的所有信息。

要指定前缀，则用 “工具管理器”主窗口中 “文件”菜单的*另存当前阶段为 ..* 菜单选项。如果您不指定前缀，则它将根据数据源的类型而定。

当您使用了*调用工具*按钮，如果需要，*.data*、*.schema* 和 *.treeviz* 文件将被更新。

树可视化工具下拉式菜单

利用六个标为 “文件”、“查看”、“选项”、“显示”、“跳转”和 “帮助”的下拉式菜单，您可以访问 “树可视化工具”的所有功能。该 “文件”菜单对大多数 MineSet 工具是一样的，参见第 92 页 “文件菜单”。

查看菜单

“查看”菜单 （在 IRIX 系统上为 “显示”菜单）包含四个选项：“全局观察”、“搜索面板”、“筛选面板”以及 “标记面板”。为了与数据进行交互，每一个选项都产生一个对话框。

搜索面板

在 “查看”（或 “显示”）菜单中选择*搜索*，则产生一个对话框可让您指定查找对象的标准 （图 1-38 和图 1-39）。

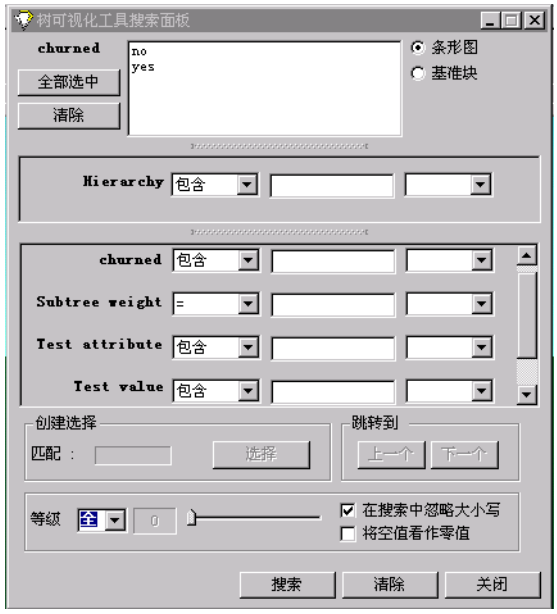

图 1-38 树可视化工具的查找对话框 （Windows）

您可以指定要查询的等级。默认情况下，则对整个等级进行查询。要限制所查询的层，从选项菜单中选择关系操作符 （例如 :<=），再指定层的操作数，使用 “层”滑动条来选择要进行查询的层。层 0 是根的等级，层 1 是其下一层，依次类推。例如，要选择根与其下面的两层，可选择 <= 2。

您也可以选择是查询条形图还是基准块。

“等级”字段可让您指定要搜索的节点。“等级”字段下面是让您为单独列指定查询标准的字段 （在 “当前列”中定义：“工具管理器表处理”窗格中的窗口。

要指定查询是否区分大小写，单击“搜索”面板中的*在搜索中忽略大小写*复选框。例如，如果切换开关打开（在该按钮上有一个复选标记），则字符串“Hello”与“hellO”相同。

标有*将空值看作 0* 的复选框默认为关闭，在这种情况下，在查询中带有空值的比较不能返回 TRUE。如果它是打开的，则空被视为 0。

当对条形图进行查询时，默认为在所有条形图中进行查询。仅查询指定的条形图，您必须从列表中选择它们。*设置全部*按钮选中所有的条形图；如果要对大多数条形图进行查询，这是很有用的，您只需再取消选择少量的条形图即可。*清除*按钮清除对所有的条形图的选择。如果没有选择条形图，则忽略条形图列表，并对所有的条形图进行查询。

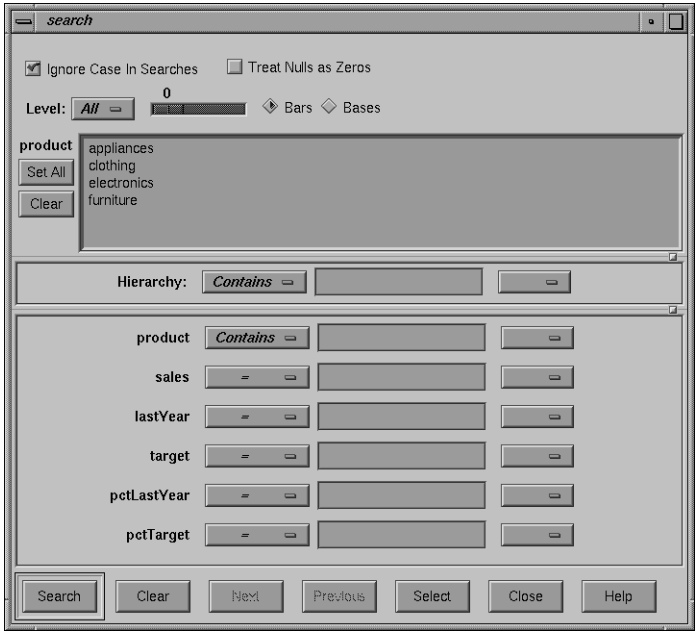

图 1-39 树可视化工具的查找对话框（IRIX）

要查询数字型值，输入该值，并选择关系操作（=、!=、>、<、>=、<=）。要查询字母型值，输入您要查询的字符。您可以使用三种字符比较类型中的一种 :

- “包含”指示包含了适当的字符串。例如， California 包含了字符串 Cal 和 forn。
- “相等”则需要正好匹配的字符串。
- “匹配”允许通配符：
  - 星号 （*）代表了任何数量的字符。
  - 问号 （?）代表了一个字符。
  - 方括号 （[ ]）包含要匹配的字符列表。

  例如， California 与 Cal*， Cal?fornia 和 Cal[a-z]fornia 匹配。

在一些情况下 （通常与 “工具管理器”中的分组有关），出现了值选项菜单，而不是文本字段。要忽略一个变量 , 选择 “选项”菜单中的 “忽略”。您可以使用选项中的关系操作符 （例如 >=）。这意味着被选值以及其后序值都被选定了。

除了数值型和字符型的比较操作以外，您还可指定*为空*操作，当值为空时操作返回真。

每个字段的右边是附加选项菜单，用该菜单可以指定 “与”或 “或”操作选项。例如，您可以指定 “销售 >20 与 < 40”。对于已知列可以有多个 “与”或 “或”从句，但是不能将 “与”和 “或”混用到单一列中去 .

如果等级的不同层次有不同数据类型的关键字，例如，顶层通过字符串进行选择，而第二层通过整数进行选择），那么 “等级”查询字段以字符型来对待并且提供了字符操作而不是数值型操作。

如果选中*在搜索中忽略大小写*复选框，则所有的字符查询比较都不区分大小写。

“搜索”面板底部的按钮包括了部分或全部下列功能：

- *搜索*启动查询过程。当面板在激活状态下并且按下 “回车”键后，该按钮自动被激活。
- *清除*关闭了所有查询聚光灯，并且清除了查询字段中的值。
- *下一个*按照从左到右的顺序选择并缩放聚焦到下一个匹配的对象。在选择了最后一个匹配的对象之后，单击 *下一个*则将视图返回到 “起始”主视图位置。只有在找到有匹配的对象之后， *下一个*才有效。
- *上一个*在与 *下一个*按钮相反的顺序上进行选择和缩放。
- *选择*选择了所有与查询标准相匹配的对象。然后，可通过 “选择”菜单与这些对象进行交互。
- *关闭*关闭了搜索窗口并关掉了聚光灯。如果 “搜索”面板重新打开，则它所处的状态与最近一次*关闭*之前的相同；再次单击*搜索*则重复上次的查询。

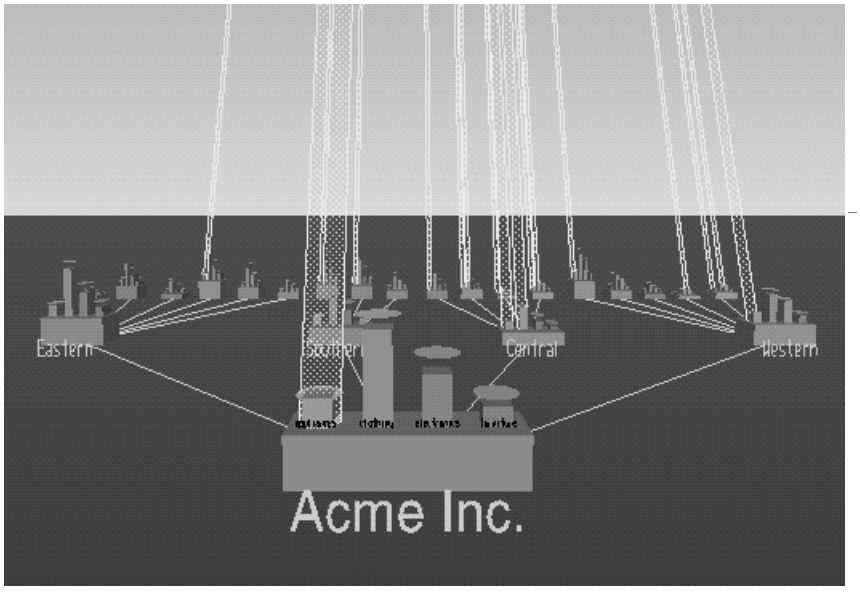

图 1-40 在 “树可视化工具”中查询的 “结果样例”

一旦搜索操作结束了，黄聚光灯就会加亮与查找标准匹配的对象。（参见图 1-40）。要显示黄聚光灯下的对象信息，将鼠标移到聚光灯上，则信息会显示在左上角，在标签*指针位于：*之下。要选择并缩放黄聚光灯下的对象，用鼠标左键单击聚光灯；如果您在单击时按下了 Control 键，则不进行缩放。

筛选面板

“筛选”面板筛选出选择的信息，这样可以微调显示的等级。您可以使用 “筛选”面板来强调特定信息，或者减少数据量以求更好的显示效果。对于大多数的 MineSet 工具，“筛选”面板是相似的，参见第 95 页 “筛选面板”。

“树可视化工具筛选”对话框中的字段遵循在 “搜索”对话框中使用的相同约定。如果选中*在筛选中忽略大小写*复选框，则所有的字符筛选比较都不区分大小写。

如果某节点不满足筛选标准，没有条形图满足标准，或没有子节点满足标准，则该节点不显示。然而，会出现特殊的对象满足筛选标准，但是它在树结构上的前辈不是这样。同样，在相同节点中的其它条形图也许不满足该标准。因为在解释上下关系时这个位置是重要的，所以，删除这些条形图也许并不好。因此，您可以在控制这些对象应该如何绘制的三个单选钮中选取一个：*填色*、*轮廓*以及*隐藏*。然而，请注意，如果对象由于“显示 0”或 “显示空”菜单而以不填色的形式被绘制，则它们都遵循原有设置。例如，如果 “空”被隐藏，则它们总是隐藏，而不管筛选中选择什么。

当对指定条形图进行筛选时是一个例外。在这种情况下，其它的条形图被删除并且不占用空间，而不管单选钮的设置是什么。

该 “高度滑动条”可让您筛选出只包含短条形图的节点。该值的大小表示为最大高度的百分数。首先，计算了画面中最高的条形图 （如果高度按照层进行了标准化，那么就为每一层中最高的条形图）。那么只显示那些至少包含一个高度不小于最大高度的指定百分比的条形图的节点。

例如，如果您在字段中输入 5%，那么就会显示包含至少一个条形图的节点，而该条形图高度不低于最高条形高度的 5%。（这些条形图显示为星号）。这一选项可以粗略的筛选出小的，不感兴趣的节点。这并不是找出带有一定值的特殊节点的精确方法；要实现该目的请使用查询面板。使用该选项可以加快复杂画面的显示或限制 0 高度附近的条形图所带来的杂乱。您也可以通过使用 “高度筛选”命令在配置文件中设置筛选选项。

虽然筛选出了小的节点，但在层级上的任何累计中仍然要对它们进行记数。

在 “高度筛选”滑动条下面的 “深度”滑动条可让您显示层级，于是在指定时间内只会各显示一定数量的层。当您位于根等级，您只能看到被滑动条指定的层目。行中的节点进行了调整来优化它们的可视性。当移动到等级中较低层的节点时，附加行自动变为可见的。已有的节点自动调节它们的位置来容纳新加入的节点；这样，一些节点也许就要移动。注意，轮廓显示了等级中的所有节点；并不仅仅是顶部节点；这样，轮廓的布局也许就与主视图的布局不匹配。轮廓中的 X 近似于主视图中的对应位置；在两个布局当中不存在精确的映射。

要启动筛选面板，单击*筛选*按钮。如果当面板活动时按下了*回车*键，就会启动自动筛选操作。要关闭面板，单击*关闭*按钮。

树可视化工具选项菜单

“选项”菜单可以对基本数据进行追溯。该菜单有五个项。

- *显示数值*显示了一个所有被选对象值的表 （记录查看器格式）。
- *显示原始数据*检索并显示与已被选中实体对应的记录。结果记录显示在表查看器中。
- *发送到工具管理器*根据当前的选择框，在 “工具管理器”历史的起点插入一个筛选操作。用于追溯的实际表达式由当前的框的范围来确定。如果什么也没选中，则出现一个警告信息。

- *互补追溯*影响*显示原始数据*和*发送到工具管理器*选择，当其被使用时，取出所有*未*被选中的数据。
- *标准化子树*确定了子树中成员的最大高度，并且对与高度相关的所有值进行标准化。

要得到追溯的更多信息，参见第 78 页 “追溯”。

树可视化工具显示菜单

“树可视化工具显示”菜单可让您控制几个显示参数。

- *基准块高度*是一个复选框，可打开和关闭基准块高度。要观察负值，或使之易于与条形图高度比较，可将该选项关闭。将它打开则提供了有关所有条形图的汇总信息。用配置文件中的 “基准块高度”语句可改变该选项的初始值。
- *标号标志*是一个切换选项，可让您打开或关闭代表标号的标志 （参见 *《MineSet for Windows 3.0 企业版用户指南》*中的 “标记面板”部分）。
- *零值* 是一个子菜单，可以控制高度为 0 的对象怎样显示。默认情况下，它们象其它对象一样显示：高度为 0 的立方体 （平面）。子菜单可让它们显示为轮廓 （显示为中空的正方形），或完全隐藏 （不画出）。利用配置文件中的 “ 0”选项可以改变其初始值。
- *空值*是一个子菜单，可以控制高度为空的对象是怎样显示的。它与 0 菜单有相同的选项；但是空选项的默认值是将对象显示为轮廓。利用配置文件中的 “空”选项可以改变初始值。

树可视化工具跳转菜单

“跳转”菜单重复了主窗口右手上边的按钮功能。它也指定了一些功能的键盘快捷键。

- *主视图*带您到一个指定的位置。默认情况下，该位置为画面的初始视图点。初始情况下，该位置是在调用“树可视化工具”和指定配置文件之后所显示的第一个视点。如果您已经利用“树可视化工具”进行工作并且单击了*设置主视图*菜单项，然后单击*主视图*，那么将返回您最后单击*设置主始视图*时的视点位置。该功能的键盘快捷键为 Control+H。
- *设置为主视图*，将主视图位置变为您的当前位置。单击*主视图*菜单项，将返回您最后一次单击*设置主视图*时的视点位置。
- *全景视图*显示了整个结构，保持画面的倾斜角度。要得到该画面的俯视图，倾斜画面垂直向下，然后单击*全景视图*菜单项。
- *向后*可让您返回到前一个位置。如果您刚启动“树可视化工具”并未从初始位置移动，则该菜单项变灰。该功能的键盘快捷键为 Control+B。
- *向前*可让您向前到您单击*向后*菜单项时的位置。如果您没有单击过*向后运行*菜单项，则*向前运行*菜单项变灰。该功能的键盘快捷键为 Control+R。
- *父节点*只有在选择了对象时才激活。如果选择了条形，单击该菜单项则选择了该条形图所在的基准块。如果选择了基准块，单击该菜单项则移到父节点上。一旦到达了根节点（等级的最高层），*父节点*菜单就变灰。该功能的键盘快捷键为 Control+U。
- *左移*可让您向左选择相邻兄弟节点。如果选择了条形，则选择它左边的条形图。如果选择了基准块，那么如果父节点在左边还有另一个子节点，则该子节点被选中。如果没有选择，或者当前选择左边没有兄弟节点，则该按钮变灰。
- *右移*可让您向右选择兄弟节点。如果选择了条形图，则选择它右边的条形。如果选择了基准块，那么如果父节点在右边还有一个子节点，则该子节点被选中。如果没有选择，或者当前选择在右边没有兄弟节点，则该按钮变灰。

- *第一个子项*可让您选择当前节点的第一个子节点。如果没有选择，或选择了条形，或当前的选择没有子节点，则该按钮变灰。
- *最后的子项*可让您选择当前节点的最后的子节点。如果没有选择，或选择了条形，或当前的选择没有子节点，则该按钮变灰。

帮助菜单

对于大多数 MineSet 工具帮助菜单是一样的，参见第 97 页 “帮助 （IRIX）”。

在树可视化工具中的空值处理

空代表未知值 （参见 *《MineSet 3.0 企业版界面指南》* 的 “MineSet 中的空”）。

在 “树可视化工具”中，在下列情况下出现空：

- 数据库或数据文件包含空值时。
- 在配置文件中不出现 “跳过遗漏”选项 （参见 *《MineSet 3.0 企业版接口指南》* 的为 “树可视化工具”创建 “数据”和 “配置文件”中的 “跳过遗漏”）并且数据表示等级中一个节点中的键值，而对于另一个却没有提供。例如，在代表州财政的过程中，如果没有关于 Texas 州收入税的记录，则 Texas 的收入税为空。这与一个记录显示 Texas 的收入税为 0 不同，因为在这种情况下它显示的税为 0。
- 当使用 “工具管理器”在分组的基础上来产生一个数组，并且没有数据落入指定分组中时，该分组对应的值为空。例如， 30-40 岁的年龄段没有数据，则该分组为空。
- 当在 “工具管理器”中产生数组并且指定了空枚举选项时，会创建与每个条形图中的第一个条对应的特殊数组条目，用它来代表组值为空的组中所有值的组合。（该条标以问号 （?），代表空。如果空组中没有数据，则与其相关的值也为空。

  注意： 如果整个数据中与空组相关的所有值都为空，则 “树可视化工具”忽略空组并不显示它。
- 空值的表达式和组合可以产生空值。

当空值映射为可视化属性时，在 “树可视化工具”中使用特殊代表物。如果空映射为高度，则该对象以轮廓模式进行绘制 （然而可通过 “显示”菜单或配置文件对其进行配置（参见第 192 页 “树可视化工具显示菜单”））（*《MineSet 3.0 企业版接口指南》* 的为 “树可视化工具”创建 “数据”和 “配置文件”中的 “空”）。对于一个条形或一个基准块，看上去就象空正方形。（它看上去并不象立方体，因为它没有高度。）对于盘，它看上去象一个圆圈。如果空值映射为颜色，则它被绘制为深灰色 （参见图 1-41）。

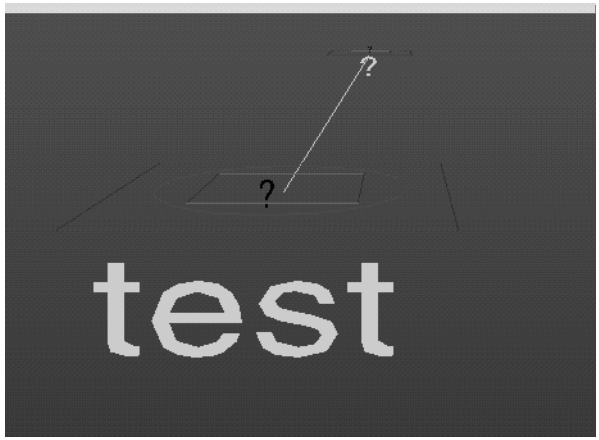

图 1-41 “空值”的代表物映射为高度、颜色、盘以及标签

当选择具有空值的对象时，问号 （?）显示在选择字段中。

树可视化工具限制

对于 “树可视化工具”，“工具管理器”不支持下列内容：

- 数据未经组合就直接显示的非组合等级结构。
- 实时监控。
- 一些不常用的选项 （跳过遗漏、轮廓、收缩、根标签、速度、爬升速度、叶空白、根叶空白、叶边缘空白、初始位置、初始角度、条标签大小、基准块标签大小以及 lod）。

- 变长数组。
- 在创建等级之后计算的表达式。例如，如果您想计算一个百分数，该百分数必须在等级组合过程之后才能计算，因为不可能对百分数进行组合。

样例配置和数据文件

所提供的配置文件和数据文件展示了 “树可视化工具” 的特征和功能。使用 “树可视化工具” 将 “决策”、“选项” 或 “回归树” 可视化的样例可参见附录 A，配置和数据文件样例。

修剪因子

修剪因子是 “高级分组” 操作，它允许您在启动分组操作之前将极端值 （称作离群值） 从数据集中排除。默认的修剪因子为 0.05。这就排除了 5% 具极值的实例 （2.5% 具有区间内的最低值而 2.5% 具有最高值）。修剪的目的就是减小在阈值产生过程中离群值的影响。

统一范围

统一范围是在自动数据分组中使用的一个选择，在数据分组过程中值区间被分为统一大小的子区间。参见第 34 页 “分组”，可得到如何用 “工具管理器” 的*列分组*按钮进行应用的详细讨论。

统一权重

统一权重是在自动数据分组中使用的一个选择，在数据分组过程中的值区间被分为一定数量的等权重组，因此每个组包含了相同数量的实例，如果每个实例的权重为 1，那么每组的权重就相等。参见第 34 页 “分组”，可得到如何用 “工具管理器” 的*列分组*按钮进行应用的详细讨论。

查看菜单

“查看”菜单可控制不同的显示选项，对于大多数可视化工具来说它是相似的。根据不同的平台，其菜单包含下列一些或所有选项：

- *筛选面板*根据一个或多个标准，减少在主查看区域中显示的实体的数目。您可以使用筛选面板来微调显示设置，强调特别信息或缩减显示信息的数量。在筛选面板右下部的 “调整比例”复选框，可以指定主窗口中的场景是覆盖了整个数据集还是仅包括筛选的数据。参见第 95 页 “筛选面板”来填写该对话框，或在字段上移动鼠标并按下 Shift F1 获得帮助。
- *设置背景色*产生一个颜色选择器，您可以指定一个新的背景颜色
- *显示窗口控件*可让您隐藏或显示在显示窗口周围的外部控件。
- *空位置*切换空值的显示。
- *动画面板*可让您显示或隐藏动画控制面板。对于不具有独立维的数据集，该菜单项被禁用。

可视化工具

该部分提供了 MineSet “工具管理器”可访问的可视化过程工具的概述，从而您可以使用不同的可视化查看方式来查看您的数据：

- **聚类可视化工具**显示了类或组的统计结果，并将统计结果与整个数据集的聚类结果放在一起，这样您可以看到是什么特征将类与其它类区分开来。
- **决策树可视化工具**可让您从分级的列中观察数据。例如，可以根据产品类、地理、销售提高以及销售补偿代表计划来检查商业利润。数据每次被分配了两个属性，所以您可以细化下寻深入到每层的属性对中。

- **证据可视化工具**可让您了解特殊属性值是如何有助于预测给定标签概率的。例如，如果在蝴蝶花数据集中，您看到属性 “萼片长度为 5.45...5.85”，那么您可以看到标签为 iris-setosa 的概率为 86.54%。当 “证据导入工具”创建模型时，“证据可视化工具”显示了不同属性是如何对所产生的决策提供贡献，并且允许进行 “What if” 分析。
- **直方图可视化工具**将数据中所有的连续型列自动分组，并将结果送入 “统计可视化工具”中以直方图的形式显示。
- **地图可视化工具**可让您可视化存在地理相关性的数据。例如，您可以将不同国家区域可视化，显示市场计划的有关效应。“地图可视化工具”的细化下寻功能可让您将焦点集中在指定区域，并且在更小的地理单元中进行更细致的分析。
- **记录查看器**可让您在类似于电子表格的行和列中查看数据。
- **散点可视化工具**在 1、 2 或 3 维空间中显示数据点。附加属性可以映射为尺寸、颜色以及形状。最后，两个属性可以映射为滑动条，在总共 8 维空间里，允许动画和闪现。当您移动滑动条，显示的内容也跟着变化来反映独立变量中的变化。“散点可视化工具”也用于显示 “关联规则”。
- **平伸可视化工具**与 “散点可视化工具”在很多方面有相似之处，区别只是在显示数据密度的时候使用了透明度。当需要表达大量的数据而又不需要画出每个点时，该工具是非常有用的。
- **统计可视化工具**计算并显示当前数据集的汇总信息 （最大值、最小值、中值、标准差、相异值以及四分之一值）。
- **树可视化工具**在以等级关系分析数据的过程中很有用。交互式的 “闪现”进展允许您在不同的等级层上检查数据关系。例如，“树可视化工具”可用于检查公司的生产线，图形显示每个产品对公司收入的贡献。等级的每个分支在更细的子层上显示信息，根据生产线并最终根据个体产品显示收入状况。

  如果每个决策由树中分离的节点来代表，那么 “树可视化工具”也用于查看 “决策树”、“选项树分类工具”以及 “回归树回归工具”的结果模型。

警告选项

在 IRIX 上，当在命令行上进行操作时，有两个 MineSet 选项将影响任何工具的调用：

- **`-warnexecute`** 指示了如果您试图执行一个在执行语句中指定的命令，则会出现一个警告，而您可以选择执行还是不执行命令。这是着眼于系统的安全而设计的，例如，从网上得到的文件，并且当利用 mtr 文件来执行命令时自动启用。（参见 *《MineSet 3.0 企业版接口指南》* 可得到关于 mtr 文件的解释。）

  您可永久启用该选项

  - 在 IRIX 系统上通过添加：

    `*minesetWarnExecute:TRUE`

    到您的 *.Xdefaults* 文件或

  - 在 Windows 系统上通过改变注册项中的值：

    `MINESET_WARN_EXECUTE`

    您可以使用 “文件 - 特性 “对话框来完成这些。

  `-quiet` 取消弹出以指示过程的对话框。可以永久启用该选项，方法是将下行

- **`*minesetQuiet:TRUE`**

  添入 *.Xdefaults* 文件。

在 Windows 上，这些选项在 “三维可视化工具” 的 “特性对话框” 中是可用的 （在 “文件” 菜单中可用）。

网上发布

MineSet 网上扩展允许在网上将 MineSet 软件产生的文件和数据可视化。MineSet 的 *.mtr* 扩展名可让您将 MineSet 配置、方案以及数据文件放入档案文件中，并作为 html 标签镶嵌在网页中。当该页被加载到 Internet Explorer 中时，在浏览器窗口中可以启动 MineSet 可视化工具。运行浏览器的机器必须安装了 MineSet 客户端软件。

对于样例和安装指令，参考 *《MineSet 3.0 企业版接口指南》*。

加权

通常权重代表了实际记录的记数，但是当数据集进行了不均匀采样时，一些记录通过加权，也就是通过增加了的重要性，可能被赋予了更多的重要性。可以在 “工具管理器” 的 “数据目标” 面板中通过选择 “挖掘工具” 来加权记录。对于任何一个导入工具，单击*深层选项*（或*高级选项*）按钮，可得到带有 “使用加权” 复选框的对话框。第二加权的意义是权重为 2 的记录与两个权重为 1 的记录等价。并且允许浮点型的权重。也可参见第 13 页 “记录加权”。

适应 2000 年问题

MineSet 支持 Y2K 日期。对于美国地区，日期是输入形式可以是 MM/DD/YY 或 MM/DD/YYYY。MineSet 对于 2 数位年遵循 X/Open 标准：大于 68 的 2 数位年可定为 1969 年到 1999 年，而小于或等于 68 的 2 数位年被定为 2000 年到 2068 年。

对于欧洲语言体系，日期的输入形式可以为 DD/MM/YY 或 DD/MM/YYYY，对于 2 位数年的处理方法同上。

在其它语言体系中，如果要输入 2 位数的年份，则在显示过程中自动扩展为 4 位数的年份。

配置和数据文件样例

MineSet 中的几个配置和数据文件样例可以展示不同工具的功能。本节将详细描述和解释每个工具的样例文件，具体条目是按工具的字母顺序排序。

- Windows 用户可在安装 MineSet 的目录 *\examples* 中找到样例文件。
- IRIX 用户可以在 */usr/lib/MineSet/examples* 中找到相应的文件。

文件描述如下：

关联规则样例文件

下面提供的数据和配置文件样例将在已备数据集的基础上对 “关联规则”进行可视化处理。其中一些文件与等级数据集相对应。规则文件包含了通过运行 “关联规则生成器”得到的规则。按照惯例，包含规则的文件应该具有 *.rules.data* 扩展名。每个配置文件指定了对应规则文件是如何显示的，配置文件必须有 *.scatterviz* 扩展名。在本部分中提到的文件在 MineSet 的 *Windows* 目录下安装的 *\examples* 目录中，对于 IRIX 用户，则在 */usr/lib/MineSet/scatterviz/examples* 中。

- *group.rules.data* 和 *group.rules.scatterviz*

  这些文件为每个产品组，例如，面包和薰烤食品、牛奶以及碳酸饮料等提供了生成的规则和配置规范。

- *category.rules.data* 和 *category.rules.scatterviz*

  这些文件为产品组中的产品类别，例如，冷藏和未冷藏牛奶，提供了生成的规则和配置规范。

- *adult94.rules.data* 和 *adult94.rules.scatterviz*

  这些文件为调查数据库提供了生成的规则和配置规范，显示了婚姻状况、受教育水平、年龄、收入以及其它变量之间的关联。

- *germanCredit.rules.data* 和 *germanCredit.rules.scatterviz*

  这些文件为德国的信贷数据集提供了生成的规则和配置规范，显示了信贷历史、职业、存款以及与其它变量之间的关联。

- *cars.rules.data* 和 *cars.rules.scatterviz*

  这些文件为轿车数据集提供了生成的规则和配置规范，显示了不同属性之间的关联。

聚类样例文件

下面的样例是聚类发挥作用的例子。该样例与 MineSet 所提供的样例数据集相关。它显示了如何利用 “聚类”挖掘工具，并解释了不同的结果和选项。

轿车数据集相对简单，主要涉及了一些熟悉的概念，如马力、自重以及到达 60 mph 所需的时间。

当“聚类可视化工具”第一次出现时，属性根据它们在区分类的过程中作用的大小从上到下进行排列。

如果您选择了类 1，则该类控制了由条形图和直方图所代表的属性的排序优先权。在其它类中的属性顺序仍然以类 1 为基础。例如，如果您单击了类 1，属性序列为汽缸、重量，然后是每加仑英里数。排序中的变化将仅出现在可视化过程中，而不是在基础数据集中。您可以比较其它类的相同行，观察属性在不同类之间的差异。当您选择了类 2，您可以在更低的层上观察到一个不同的属性排序。在这种情况下，产地是最重要的，然后是汽缸、马力、每加仑英里数。

列重要性样例文件

下面是 “列重要性”发挥作用的样例。该样例与 MineSet 所提供的样例数据集相关。它显示了如何利用 “列重要性”挖掘工具工作，并解释了不同的结果和选项。

当用户更改他们的电话业务服务商时，这个术语被称为 “客户波动”，这是通讯产业中共同的问题。文件 *churn.schema* 和 *churn.data* 即为这个样例。 Windows 用户可以在 MineSet 安装的目录下的 *\data* 目录中找到，该样例 IRIX 用户可以在 */usr/lib/MineSet/data* 中找到该样例。

简单运行 “列重要性”模式产生了下列三个属性：

- 白天总费用。
- 用户服务电话数目。
- 州。

通过从高级模式中运行 “计算改良纯度”，您可以观察到 “白天总费用”和 “白天总分钟数”具有相同的纯度级别 （48.67）。将其中之一移到右边栏中 （例如，白天总分钟数）并且重新运行 “计算改良的纯度”，您可以观察到另外一个 （白天总费用）没有数值。这两个属性高度相关。

当 “白天总分钟数”在右边一栏时观察属性值，我们可以看到下列内容：

- 国际计划 （4.1）
- 用户服务电话数目 （8.1）
- 州 （4.7）

您可以选择将 “国际计划”移向右边，因为该信息易于使用并且容易测量。

另外两个属性 （用户服务电话数和州）仍然非常重要 （实际上，它们的重要性在增加），所以它们显然与 “国际计划”不相关。

通过这种方式观察属性的重要性，您可以确定哪一个属性可以被同样好 （或几乎同样好，但是更容易测量或理解的其他属性代替。通过观察纯度，您可以确定附加属性能起多大作用。例如，在上述的方案中，州明显地提高了纯度。在 *iris* 数据集中，所选择的第三个属性 （*萼片长度*）只是略微提高了纯度，简单的二维散点图将给出与三维散点图近乎一样多的信息。

决策树样例文件

以下是 “决策树”导入工具发挥作用的样例。每一个样例都与 MineSet 所提供的样例数据文件相关。通过运行导入工具，您可以生成如下面描述的 *-dt.treeviz* 文件。

可以通过打开保存在 *data* 目录中的 *.schema* 文件而将数据文件加载到 MineSet 中（例如，*churn.schema*）可以从 *examples* 目录中打开具有 *-dt.treeviz* 扩展名的分类工具可视化文件。

- Windows 用户可以在安装 MineSet 目录的数据和样例目录中找到这些文件。
- IRIX 用户可以在 *data* 和 *examples* 目录中找到这些文件，而上述两个目录在 */usr/lib/MineSet/treeviz/examples* 中。

客户波动

当用户将他们的电话业务从一个通讯公司转向另一个时，这个术语被称为 “客户波动”，这是通讯产业中共同的问题。在 *examples* 目录中的文件， *churn-dt.treeviz*，显示了一个为解决该问题而导入的 “决策树”分类工具。通过在 *data* 目录中的文件 *churn.schema* 上运行导入工具可以产生该文件，注意其标签设为客户波动 （是、否）。该文件是虚拟的，但却是以从实际数据中建立的模式为基础的。

在该树中的根分割是以白天用户交谈的时间量 （白天总分钟数）为基础进行的。每天交谈超过 264 分钟的用户的客户波动率比没有超过的用户明显地高很多 （60% 比 11%）。这些很可能是最有利润潜力的用户。

左边的子树代表了每天交谈少于 264 分钟的用户。他们的客户波动率为 11% ；但是如果他们打了三个以上的用户服务电话，则客户波动率增至 49%。

右边的子树代表了每天交谈多于 264 分钟的用户。他们的客户波动率为 59% ；但是如果他们有声音邮件计划，则客户波动率降为 9.3%。如果他们没有声音邮件计划，则客户波动率几乎为 75%。

轿车的产地

*轿车*数据集包含了从 70 年代到 80 年代不同型号轿车的信息。属性包括重量、加速度以及每加仑英里数（mpg）。*examples* 目录中的文件， *cars-dt.treeviz*，显示了一个为解决该问题而导入的 “决策树”分类工具。通过在 *data* 目录中的 *cars.schem* 文件上运行导入工具可以产生该文件，注意其标签设为产地 （日本、美国、欧洲）。如果您有一个关于轿车属性的数据集，您可能想知道不同产地轿车的特征。

- Windows 用户可以在安装 MineSet 的目录中 *examples\cars-dt.treeviz* 和 *\data\cars.schema* 下找到这些文件。
- IRIX 用户可以在 */usr/lib/MineSet/treeviz/examples/cars-dt.treeviz* 和 */usr/lib/MineSet/data/cars.schema* 中找到这些文件。

注意，树中左边的分割是关于品牌的。根部的分割不是关于品牌的，这是因为 “决策树”导入工具不支持多向分割，而在引擎体积上的分割被视作更好的区分器。您可以使用 “工具管理器删除列”功能来隐藏品牌，这样使得问题更加有趣。

在 “决策树”中，您可以看到对于美国产轿车来说引擎体积是一个非常好的区分器。带有大引擎 （>169.5 立方英寸）的轿车全是美国造的，但是较小的轿车却产自各地。通过选择 “选项” > “显示原始数据”，您可以观察到不是美国产的大引擎轿车是 Mercedes。请注意，在树中的根节点 （也就是，整个训练数据集）有更多的美国轿车（62.50%），然而在体积属性上进行一次分割后，更难预测具有小引擎轿车的产地。根的纯度为 16.2 表明只有一个类占优势 （在这种情况下为美国）。右节点 （> 169.5 立方英寸）的纯度为 96.81，表明我们已经找出了非常纯的子集 （几乎所有具有大引擎的轿车都产自美国）。实际上，右子树的误差率估计为 0% （绿色基准）。根左边的节点的纯度为 0.23 并且具有更高的误差率 31.25% （橙色基准）。该子问题比最初的一个更难解决：每个类的记录数目几乎一样。

预测性别

*成人*数据集包含了有工作的成年人的信息。该数据集是从美国调查局中抽调的，它包含了年龄超过 16 岁的人的有关信息，他们每周工作至少一小时，每年总收入超过 $100。您也许想知道怎样将男性和女性特征化，文件 *adult-sex-dt.treeviz* 显示了为解决该问题而导入的 “决策树”分类工具，该文件通过在 *adult.schema* 上运行导入工具而产生，并且其标签设为*性别*。该数据集包含了几乎 50,000 条记录；所以当您在工作站上运行 “决策树导入工具”时将要花费几分钟的时间。

- Windows 用户可以在安装 MineSet 的目录 *\examples\adult-sex-dt.treeviz* 和 *\data\adult.schema* 下找到这些文件。
- IRIX 用户可以在 */usr/lib/MineSet/treeviz/examples/adult-sex-dt.treeviz* 和 */usr/lib/MineSet/data/adult.schema* 下找到这些文件。

这些可视化过程结果提供了以下直观结果：

- 婚姻关系是最重要的属性。丈夫通常是男性。（有趣的是，有一位丈夫是女性，这表明调查局存在数据质量问题，并没有意识到这是同性之间的婚姻。）同样地，如果是妻子，那么她通常是女性，除了三个记录例外。

为了使问题更加有趣，删除关系属性并且产生一个新 “决策树”。在这种情况下：

- 最重要的属性就是婚姻状况。
- 从基准块的高度来看，大多数人要么已离婚，要么与普通人结婚或从未结婚。很少有人分居、与军人结婚、或丧偶。
- 根节点上的分布显示该数据集中男性居多。（该数据集包含了有关有工作的成年人的信息而并不代表整个人口。）
- 最左边的节点包含了离异有工作的成年人。我们可以看到这里的分布比根部更加平衡 （60% 为女性， 40% 为男性）。第二个节点包含了已婚有工作的成年人。我们可看到 89% 是男性。第三个节点包含了从未结婚的有工作的成年人。他们的数量大致与离婚组的相同，男性略多一点。最右边的节点包含了丧偶的有工作的成年人，81% 为女性 （也许是因为她们较高的生活期望值）。 " 丧偶 " 一词表示失去配偶的任何一方。

如果您想将注意点集中在职业女性上，以发现新的结果，那么您可以使用查询面板来找出含有大量女性人口的部分。您可以选择

- 性别与女性匹配 （单击窗口顶部的女性）
- 子树权重 >1000
- 百分数为 > 80

三个黄聚光灯显示了匹配节点。因为两个节点在同一条路径上，所以观察离根最近的节点 （在右边）。该路径翻译为规则

```
婚姻状况 = " 丧偶 " 显示其中 81.23% 为女性

婚姻状况 = 离婚和职业 =
      管理文书显示 87.67% 为女性
```

在该训练集中，在根部的 16,192 个记录中，有 1233 （丧偶）和 1045 （离婚和在职）个女性满足这些规则。该单一部分包含了数据集中超过 14　的职业女性的信息。

工资因子

如果您有一个有工作的成年人的数据集，您可能想知道什么因素影响着工资水平。您可以将记录分为两类：每年的收入在 $50,000 以下或以上的成年人。每个记录就有了一个属性，该属性具有两个值中的一个。" 50,000 ー " 和 " 50,000+ " 您可以运行 MineSet 分类工具来帮助确定是什么因素影响了工资。样例文件 *adult-salary-dt.treeviz* 显示了为解决该问题而导入的 “决策树” 分类工具。通过在数据文件 *adult.schema* 上运行导入工具，以及在用户定义阈值 50000 的基础上对 *gross_income* 进行分组可以得到该文件，注意标签被设为 *gross_income_bin* 。

- Windows 用户可以在安装 MineSet 的目录 *examples**adult-salary-dt.treeviz* 和 *data**adult.schema* 下找到这些文件。
- IRIX 用户可以在 */usr/lib/MineSet/treeviz/examples/adult-salary-dt.treeviz* 和 */usr/lib/MineSet/data/adult.schema* 中找到这些文件。

这些可视化过程提供了以下直观结果：

- 代表整个训练集的根节点显示了 76.07% 的有工作的成年人年收入在 $50,000 以下。
- 年龄是最重要的因素。27 岁以下的人中只有 3.07% 的部分收入超过 $50,000。基准颜色为绿色，表明规则很准确 （大约 3% 的误差率）。
- 对于预测年龄超过 27 岁的人的工资，教育水平是很重要的因素。调查局对每个人都分配了教育层次指标。“决策树” 在 12.5 处发生分割；层次 13 正好相对于学士学位。具有学士或更高学位的人中大约有 55% 的收入在 $50,000 以上。
- 对于年龄大于 27 并且受到良好的教育的部分，婚姻关系是重要的工资预测因素。对于那些已经结婚的人，收入达到 $50,000 以上的机率，对于丈夫为 73%，对于妻子为 75%。（但是，包含妻子的节点只有一个较小的基准块，表明几乎没有女性符合该规则。）如果在这组中的人没有结婚，收入达到 $50,000 以上的机率，对于男性会降为 27%，对于女性为 25%。

蝴蝶花分类

在该数据集中，每个记录描述了蝴蝶花的四个特征：萼片宽度、萼片长度、花瓣宽度、花瓣长度。每个蝴蝶花 被进一步划分为 iris-setosa、 iris-versicolor 或 iris-virginica 类型。该目标就是理解如何按不同特征来区分类型。

在运行分类器之前，单击 “工具管理器分类工具” 选项卡中的 “列重要性” 选项卡；然后单击*继续*。您得到了特征重要性的排序。petal_width，petal_length 和 sepal_length。在 “散点可视化工具” 中，您可以将这些属性映射到坐标轴上， iris_type 映射为颜色，并观察聚类。

文件 *iris-dt.treeviz* 显示了为解决该问题而导入的 “决策树” 分类工具。通过在 *iris.schema* 上运行导入工具可产生该文件。

- Windows 用户可以在安装 MineSet 的目录 *\examples\iris-dt.treeviz* 和 *\data\iris.schema* 中找到这些文件。
- IRIX 用户可以在 */usr/lib/MineSet/treeviz/examples/iris-dt.treeviz* 和 */usr/lib/MineSet/data/iris.schema* 下找到这些文件。

运行 “树可视化工具”，您可以看到根节点有 6% 的误差率，即使纯度很低 （0）。纯度测量了分布的斜度，在根部的分布非常统一：每个标签值有 50 个记录。左边的分支（petal-length<=2.6 英寸）到达了只包含 iris-setosas 类型的绿色的节点 （0 误差）。另外一个分支利用了 petal_width 上的检验同样也能够很快地划分类别。路径 petal-length > 2.6 和 petal-width <= 1.65 和 petal-length > 5 到达包含 4 个记录的不纯叶节点。iris-virginica 类型有三个记录，而 iris-versicolor 只有一个。“决策树” 并没有分割该节点，因为它被视为为没有意义 （默认情况下，每个分割必须包含两个子节点其权重至少为 2。节点颜色也为黑色，说明了没有实例符合该节点，所以也没有关于它的估计误差率。

总结如下：萼片长度 <= 2.6 英寸的花被预测为 iris-setosa，那些萼片长度为 > 2.6 英寸且 <=5 英寸以及萼片宽度 <= 1.65 英寸的被预测为 iris-versicolor，那些萼片长度为 >2.6 英寸并且萼片宽度为 > 1.65 或萼片长度 为 > 5 英寸并且萼片宽度为 <= 1.65 的被预测为 iris-virginica。

当使用 “列重要性” 来区分数据时，因为 “决策树” 在连续属性上产生了两向分割，因此树的根分割与列重要性中的第一个属性不同 （参见第 52 页的 “列重要性”可得到更多的细节）。

蘑菇分类

文件 *mushroom-dt.treeviz* 显示了为蘑菇分类而导入的 “决策树”分类工具。通过在 *mushroom.schema* 上运行导入工具可以产生该文件。

- Windows 用户可以在安装 MineSet 目录的 *examples**mushroom-dt.treeviz* 和 *data**mushroom.schema* 中找到这些文件。
- IRIX 用户可以在 */usr/lib/MineSet/treeviz/examples/mushroom-dt.treeviz* 和 */usr/lib/MineSet/data/mushroom.schema* 中找到这些文件。

分类目的是在给定的数据集上分辨哪些蘑菇可以食用而哪些是有毒的。在该数据集中有 8000 多条记录，运行该导入工具也许会花几秒种。

每个蘑菇有很多特征，包括菌帽颜色、裂纹和气味。如果您建立了 “决策树”分类工具，您可以看到只用气味属性就可以对区分蘑菇是否有毒情况中的 50% 作出判断。如果蘑菇没有气味，则有毒的概率为 3.4%。下一个属性是茎的形状，如果呈锥形，则蘑菇可以食用，但是如果形状为越来越粗，则有 11.6% 的可能性是有毒的。有 1032 个蘑菇记录到达该节点。您可以沿着更深的节点来观察如何考虑其它属性的作用。

党派隶属

该数据集包含投票记录。目的是从关键投票数据的信息中下找出议员所属党派。数据集包括了由*议会季度年鉴*（*CQA*）所鉴定的美国众议院每个议员在 16 次关键表决中的投票。 CQA 列出了九种类型的投票：投票赞成、约定赞成以及宣布赞成 （这三个简称为是）；投票反对、约定反对以及宣布反对 （这三个简称为否）；投票表明避免利益冲突以及不投票或以其他方式表明立场 （这三个简称为未知处理）。

在运行分类工具之前，观察 16 个投票结果看哪个特征最重要。然后运行 “决策树”分类工具。

文件 *vote-dt.treeviz* 显示了为解决该问题而导入的“决策树”分类工具。在 *vote.schema* 上运行导入工具可产生该文件。

- Windows 用户可以在 安装 MineSet 目录的 *\examples\vote-dt.treeviz* 和 *\data\vote.schema* 中找到该文件。
- IRIX 用户可以在 */usr/lib/MineSet/treeviz/examples/vote-dt.treeviz* 和 */usr/lib/MineSet/data/vote.schema* 下找到该文件。

乳腺癌诊断

乳腺癌数据集包含了有关妇女接受乳腺癌诊断的信息。每个记录包含了病人细胞大小、肿块厚度以及边缘附着力等属性。最终的属性是诊断为恶性的还是良性的。文件 *breast-dt.treeviz* 显示了为解决该问题而导入的“决策树”分类工具。通过在 *breast.schema* 上运行导入工具可产生该文件。

- Windows 用户可以在安装 MineSet 目录 *\examples\breast-dt.treeviz* 和 *\data\breast.schema* 中找到这些文件。
- IRIX 用户可以在 */usr/lib/MineSet/treeviz/examples/breast-dt.treeviz* 和 */usr/lib/MineSet/data/breast.schema* 下找到这些文件。

该“决策树”表明细胞大小一致性是非常有力的区分属性。根分布为 65% 对 35%（纯度为 7.07），根的两个子节点更加倾斜，其中左节点的误差率仅为 1.29%。根本身是极好的区分器：如果您将树高度限制为单一层，则误差率为 7.3%。

甲状腺机能减退诊断

甲状腺机能减退疾病数据集与乳腺癌的相似，只是我们试图预测的是甲状腺机能减退而不是癌症。文件 *hypothyroid-dt.treeviz* 显示了为解决该问题而导入的“决策树”分类工具，通过在 *hypothyroid.schema* 上运行导入工具可产生该文件。

- Windows 用户可以在安装 MineSet 目录的 *\examples\hypothyroid-dt.treeviz* 和 *\data\hypothyroid.schema* 下找到这些文件。
- IRIX 用户可以在 */usr/lib/MineSet/treeviz/examples/hypothyroid-dt.treeviz* 和 */usr/lib/MineSet/data/hypothyroid.schema* 下找到这些文件。

在数据集中有 3163 条记录，其中的绝大多数没有甲状腺机能减退症状 （95.23%）。这就意味着预测否定的判断在多数时间里是对的。但是，我们所担心的正是那些模型预测为健康而实际具有甲状腺机能减退的人们。这种假否定是非常严重的。通过从 “深层导入工具”选项中选择混淆矩阵，您会看到有五个具有甲状腺机能减退的病人被错误地分类。

观察 “决策树”，您可以看到根节点为绿色 （高准确度）。根部的 fti 单一属性显示了要给出许多否定的诊断是相对容易的。具有高 fti 的人 99.7% 是否定的，而那些带有未知值的人也是否定的 （医生不测量这个属性也许是因为其它的一些属性很明显），但是剩下的 （218 个人）却是很难诊断的病例。我们从 3163 条记录入手，但是真正 “有趣的”并值得挖掘的只有 218 个，这是因为要确定大多数病例的分类很容易。在该样例中，大多数数据并不令人感兴趣，您想把注意力很快集中在一小部分人上。对于这 218 个人，您可以看到有 66% 是肯定的而 34% 是否定的。

当您沿着树向下，可以增加高度比例值 （可视化工具的左上部的滑动条）查看不同的高度。在其中，大多数人具有甲状腺机能减退的节点满足条件 “fti<=64.5 并且 tsh>5.95”。它包含 151 条甲状腺机能减退记录其中的 140 条。

Pima 糖尿病诊断

该数据集是关于利用统计学原理对糖尿病进行诊断的问题的，它是从 Arizona， Phoenix 的美州部落中收集而来的。该目的是根据一些医学特征 （例如，血压、体重、葡萄糖水平以及年龄）来确定一个病人是否患有糖尿病。

文件 *pima-dt.treeviz* 显示了为解决该问题而导入的 “决策树”分类工具。通过在 *pima.schema* 上运行导入工具可产生该文件。

- Windows 用户可以在安装 MineSet 目录 *\examples\pima-dt.treeviz* 和 *\data\pima.schema* 中找到这些文件。
- IRIX 用户可以在 */usr/lib/MineSet/treeviz/examples/pima-dt.treeviz* 和 */usr/lib/MineSet/data/pima.schema* 下找到这些文件。

DNA 边界

在 DNA 数据集中有 3,186 条记录。 DNA 边界是从分子生物学领域的角度来界定的。结合连接处是 DNA 链上的点，在其上 “多余的” DNA 在蛋白质产生过程中已经被删除。我们的任务是识别 exon/intron 之间的界限 （被称为 EI 位置）；以及 intron/exon 之间的界限 （被称为 IE 位置）；或两者都不是。IE 边界被作为 “接受者” 而 EI 边界为 “捐献者”。这些记录最初从 GenBank 64. （*genbank.bio.net*) 中得到，其属性提供了 60 个核苷酸的观察窗口。分类点处于窗口的中点，这样就保证了联结处的每一边为 30 个核苷酸。

在该例中，“决策树” 的根节点显示了这三个类的分布。通过指向条形图，您可以看到其组成为约有 24% 的 exon/intron，24% 为 intron/exon，而 52% 则为 “不是”。根节点前面的 “left_01” 标示了这是一个重要的需要首先查看的属性。“left_01” 记号指示我们正在讨论的处于连结结合处左边的第一个核苷酸。第一个核苷酸 （以及所有核苷酸）的属性值选择为 “A”、“G”、“T” 以及 “C” 核苷酸。如果 “left_01” 核苷酸是一个 “G”，那么就沿着 “G” 分支并一直到下一个节点，其分布现在显示了这样的核苷酸比在根上的更象 “exon/intron” 或 “intron/exon”：该分布对于 “exon/intron” 为 32%，而对于 “entron/exon” 为 43%，而对于 “没有” 为 24%。如果 “left_01” 核苷酸为 “A” “T” 或 “C”，那么就采用相应的 “A”，“T” 或 “C” 分支并且在三种情况下，“没有” 的概率急剧增加 （分别为 87%、 87% 和 95%）。该检验和分支过程就一直重复直到到达带有预测类 （“exon/intron” “entron/exon” 或 “没有”）的最后节点。

对于该数据集，由于该领域的概率自然性，“证据分类工具” 可能比 “决策树分类工具” 更加合适。这可以通过比较估计的误差率来得到证实。

决策表样例文件

下面的样例显示了 “决策表” 如何起作用，这些样例均与 MineSet 所提供的样例数据文件相关。通过运行 “决策表” 导入工具 （在 “深层导入工具选项” 中打开 “建议使用特征查询”），您可以得到将在下面描述的 *-dtab.dtableviz* 和 *-dtab.dtableviz.data* 文件。

注意：数据 （*.data*）以及附带的配置 （*.schema*）文件在客户端工作站的 *data* 目录中。分类可视化工具文件，带有 *-dtab.dtableviz* 扩展名，在客户端工作站的 *examples* 目录中。要将数据文件装入 MineSet，打开 *.schema* 文件。

- Windows 用户可以在安装 MineSet 目录的 *\examples* 和 *\data* 下找到这些文件。
- IRIX 用户可以在 */usr/lib/MineSet/treeviz/examples* 和 */usr/lib/MineSet/data* 目录中找到这些文件。

客户波动

*客户波动*是指一个电话用户离开一个公司而转投另一个公司。该样例显示了是什么因素使用户重选电话公司。文件 *churn.schema 和 churn.data* 用于产生这个样例。 Windows 用户可以在 *Program Files\SGI\MineSet 3.0\data 目录下找到它们。* IRIX 用户可以在 */usr/lib/MineSet/data* 中找到这些文件。

文件 *churn-dtab.dtableviz* 显示了将客户波动的属性作为标签而导入的分类工具结构，该分类工具的误差率为 5.5%。 14.3% 的记录 代表了波动的用户。为第一细节层而选择的两个属性是用户服务电话次数和白天总费用。通过观察在这两个属性上的分布，您可以看到客户波动随着白天总费用的增加而增加，而只有在白天总费用少于 29.75 的情况下例外。如果服务电话的次数超过 3，那么客户波动率就高。大约 3/4 的记录白天总费用小于 38 并且用户服务电话等于或小于 3。

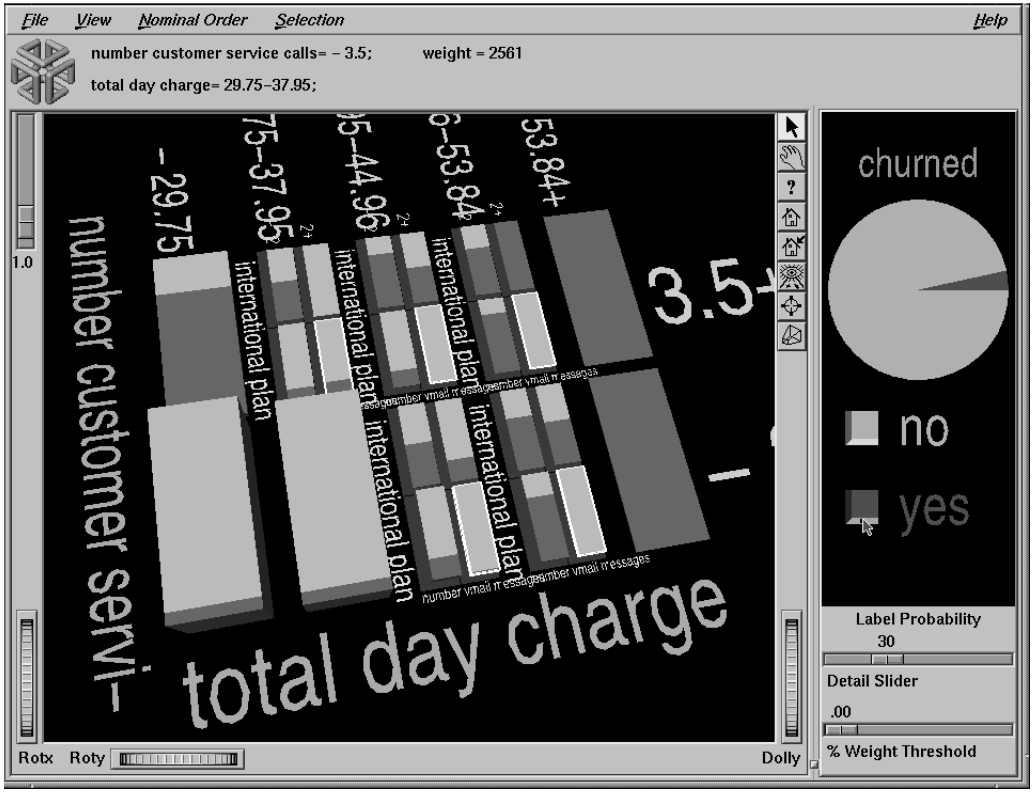

图 A-1 在客户波动数据集中细化下寻

当查看用户客户波动时，从并不清楚的区域开始进行细化下寻。图 A-1 显示了对所有块状图进行的细节下寻，其中没有一个明显的优势类。下一个要考虑的属性是国际计划和声音邮件的数目。在那些日费用很高的用户中，很明显，具有国际计划以及几乎没有声音邮件消息与用户波动率具有很好的相关性。选择每个细化下寻区域中右下角的块，然后用鼠标刷过右边概率窗格中 “churned=yes”旁边的盒子，则显示了在所选的区域中只有 3.4% 的用户客户波动。

现在，在每个先前的细化下寻区域左中上部的块状图上进一步细化下寻。这样做时可以得到非常类似的分布。一致的模式表明：高的夜晚总费用以及高的国际总费用与客户波动之间呈较好的相关性。您可以甚至细化下寻到另一层，来观察国际通话总数的效果，但是该层有助于得出结论的记录太少，对在该样例基础上得到的预测结果也许不会太可信。如果您对国际通话如何影响客户波动以及它与其它变量如何相关感兴趣，可以返回“工具管理器”将国际通话总数映射为等级中更高的层次，并且再运行决策表。

虽然“州”与客户波动有很好的相关性，但并不选择它，这是因为算法对于取值较少的变量有内置优先权。这就防止了算法选择类似于社会保障数目一样的属性，该数目唯一指示着每一条记录，这样，产生的训练集高度准确，但对于划分将来未标记的数据没有任何用处。

轿车的产地

*轿车*数据集包含了从 70 年代到 80 年代不同型号轿车的信息。属性包括重量、加速能力以及每加仑英里数 （mpg）。文件 *cars-dtab.dtableviz* 显示了对于该数据集而导入的“决策表分类工具”的结构。将标签设为“产地”（日本、美国、欧洲），并通过在 *cars.schema* 上运行导入工具可以产生该文件。

- Windows 用户在安装 MineSet 目录 *\examples\cars-dtab.dtableviz* 和 *\data\cars.schema* 中可以找到这些文件。
- IRIX 用户可以在 */usr/lib/MineSet/examples/dtableviz/cars-dtab.dtableviz* 和 */usr/lib/MineSet/data/cars.schema* 下找到这些文件。

因为品牌属性唯一地确定了产地，所以分类工具的结构就极为简单。仅显示的两个属性是品牌和汽缸型号。查看各个品牌对汽缸数目的趋向比较有趣。例如，Mazda 共有 21 种不同的型号， Honda 有 18 种型号，但是它们都只有 5 个或更少的汽缸。相反地，Cadillac 只生产带有 6 个或更多汽缸的轿车。

如果首先删除品牌属性，那么该样例会变得更加有趣。另外一个有用的转换可以是将汽缸转化为字符串型，这样可以显示每个汽缸值，而不是以一个组形式出现。作为选择，通过确定的映射，可以在品牌和汽缸之外创建附加的细节层。

预测性别

*成人*数据集包含了有关从业成年人的信息。该数据集是从美国调查局中抽调的。它包含了年龄超过 16 岁的，每周工作至少一小时，每年收入超过 $100 的人的信息。您也许想知道怎样将男性和女性特征化。

文件 *adult-sex-dtab.dtableviz* 显示了为解决该问题而导入的 “决策表分类工具”的结构。在删除婚姻关系列 （使得分类工具更加琐细）之后，将标签设为性别，并通过在 *adult.schema* 之上运行导入工具可以产生该文件。对于每个值的组合来方便地查看记录分布，您可以使用左边比例滑动条来缩放块状图的高度。

- Windows 用户可以在安装 MineSet 目录 *examples**adult-sex-dtab.dtableviz* 和 *data**adult-sex.schema* 中找到这些文件。
- IRIX 用户可以在 */usr/lib/MineSet/examples/dtableviz/adult-sex-dtab.dtableviz* 和 */usr/lib/MineSet/data/adult-sex.schema* 下找到这些文件。

在 “决策表可视化工具”中，“标签概率窗格”显示了从业 ' 为男性的先验概率要高于职业女性。“证据可视化工具”显示了婚姻状况和职业对于确定性别是非常重要的属性，但是，它并没有显示这两个属性之间的依赖性。最高层显示了几个相互关系。例如，职业为手工业维修的人大多数与普通人结婚 （更为特殊地是，他们之中 98.6% 是丈夫），而职业为 “Other-Service”的人大多数却是 “ Never-married”（ 48% 为男性）。

起初您也许会觉得奇怪，“Marital_status = Married-civilian-spouse”的大多数人为男性，但是一旦您考虑到该数据是从纳税申报表中收集而来，您就会觉得这很合理，因为这些男性的妻子们不工作，只是与她们的丈夫一起被收入档案材料。

“Occupation = Admin-Clerical”中的离婚率最高。那些职业为 “Other-Service”的人其离婚率也很高，但是他们更倾向于分居，因为他们中 “分居”的数目甚至要比 “Admin-Clerical”的还高。

假设您想在 “寡居”并且 “Occupation = Admin-Clerical”的条件下找出是女性的概率。在证据可视化工具中，通过在这两个属性值上单击就可以得到近似的答案 （女性为 94.7%）。这里我们通过在这两个属性的交点处的块上单击鼠标左键就可以得到确切的结果 （女性为 95.2%）。

在 “已婚城市夫妇”和 “职业 = 未知”的对应块上细化下寻至最下面一层，有一个比其它任何职业和婚姻状况组合更明显地模式：更年轻的成员趋向于女性，更年长的成员趋向于男性。

工资因素

对于一个从业的成年人的数据集，您也许想找出是什么因素影响工资。首先，将总收入分为两组，少于 50,000 的为一组，超过 50,000 的为一组。然后，您可以运行 MineSet 分类工具来帮助确定是什么因素影响着工资。文件 *adult-salary-dtab.dtableviz* 显示了为解决该问题而导入的 “决策表”分类工具。在 *adult.schema* 上运行导入工具可以产生该文件，并利用用户定义的阈值将 *gross income* 分为五组。

- Windows 用户可以在安装 MineSet 目录的 *\examples\adult-salary-dtab.dtableviz* 和 *\data\adult-salary.schema* 中找到这些文件。
- IRIX 用户可以在 */usr/lib/MineSet/examples/dtableviz/adult-salary-dtab.dtableviz* 和 */usr/lib/MineSet/data/adult-salary.schema* 下找到这些文件。

因为标签是数值型的，所以可以用连续的色谱将颜色分配给每个类。同样在右边概率窗格中的类并不是通过份额大小来进行排序的，这是因为它们具有数值顺序。红色被分配给最高的组 （50,000+）。

在顶层所选择的两个属性是婚姻关系和 “Education_ num”。因为属性 “education num”仅仅是可能的不同教育程度的列举，并不是您所想象的是教育的年数，所以并无特殊作用。但是，它们却存在着一个近似的相关性。如果您更喜欢看到实际的字符串值，可用教育程度来代替该列。如果您只是删除该列， education_num，并重新运行使用特征查询，那么该算法可能不会在等级的顶部选取教育程度，因为它的记录太多。

该模型中属性的顺序是自动选中增加准确性的。应用领域知识来进行映射的模型通常可以提供更有用的可视化过程。 *adult-salary3-dtab.dtableviz* 提供了这样的模型，并显示在图 A-2 中。在这里工资首先被分为 3 个区间 （20,000 和 60,000 是阈值）。在第一层映射的属性是：关系和性别；第二层为教育程度和职业；第三层为 hrswk （每周工作小时数）和年龄。

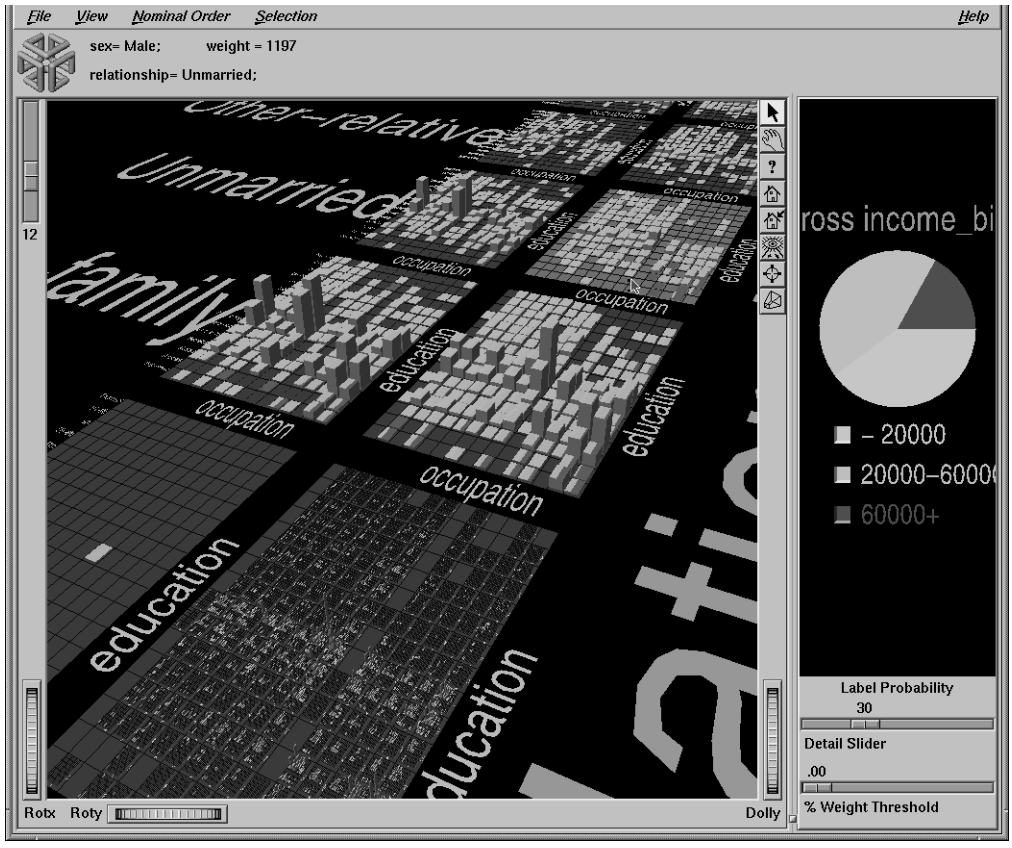

图 A-2 决策表可视化工具使用成人数据集

我们知道，在顶层 “婚姻关系”和 “性别”有很强的相关性。我们当然期望所有的丈夫都是男性，而所有的妻子都是女性，但是我们可以立即看到情况并不是这样。通过单击对应男性妻子的块状图，我们可以看到其中有三个他们的工资都落入 20,000-60,000 区间以内。可以在这些块状图中细化下寻显示出有关这些不正常记录的更多信息。您也许希望选择这些块 （使用鼠标左键）追溯其基本数据，由此您可以发现所有其它属性的值。

在背景上单击鼠标右键，这样会全局地细化下寻到下一个层。现在，对于教育程度和职业的每个组合，每个块状图都被一个方阵所替代。对于每个方阵其排序是一样的，并且整个的排序都是按照与收入之间相关性来进行的。如果您选择下拉式菜单中的 “名称顺序” > “按权重排序”，那么整个排序将按照记录权重的大小来进行。最流行的职业和教育程度层将出现在每个方阵的左下角。“高中毕业”是比例最多的教育程度，并且 “专业技能型”是最普通的职业，但是职业为专业技能型的高中毕业生却不多。

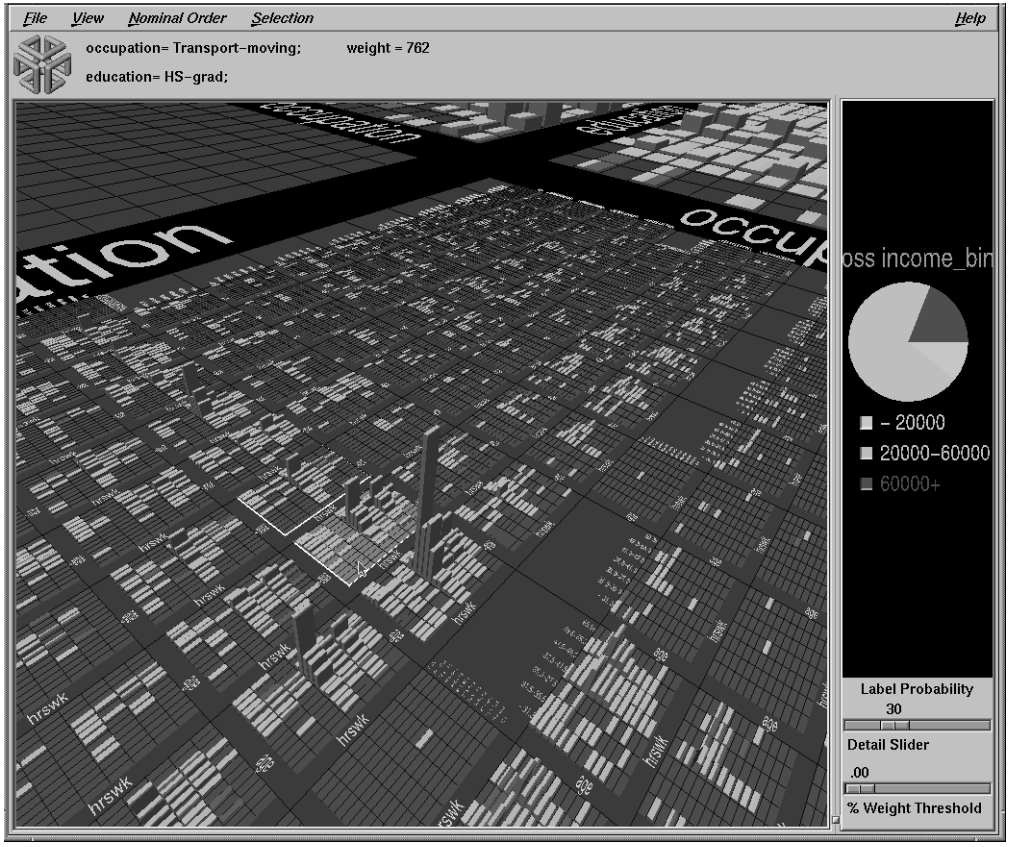

图 A-3 成人数据集的近景查看

返回按照收入的排序，您可以看到对于性别和婚姻关系每个组合的不同分布。在不属于同一家庭关系的男性和女性之间并无很大的不同，但是在未婚男性和女性之间的差异却十分明显。（参见图 A-3）。在男性方阵左下角中有很特别的红块聚类，而在女性方阵中却没有。通过稍微放大高度比例，您将会注意到女性方阵中在 “职业 = 主管办事员”和“其它服务”处有一个尖峰，而在对应的男性方阵中却没有这样的尖峰。

在人口最多的块 （男性丈夫）上单击鼠标右键。该操作可能会花掉一分钟，这是因为可视化工具需要为该块的下一层建造全部几何形状。该几何形状是根据需要创建的，因为将它们从头开始全部创建所需的时间过长，并且过于浪费，而用户很少对许多过于细节的区域进行查看。如果您在背景上误击了右键，而下一层细节量又特别大，那么就得等很长的时间，这是因为计算量大的原因。如果细化下寻将花费了很长的时间，那么就会出现一个带有取消按钮的进度条。

考虑一下通过为职业和教育程度的每个组合显示 hrswk 分布而创建的许多年龄组。一个使人惊奇的事实是，不管显示了多少块图，在占所有丈夫 2.5% 的地方有一个尖峰。（参见图 A-3）如果您必须为典型的丈夫挑选一个特征，那么应该有理由说他高中毕业，做手艺维修，年龄在 41 到 59 之间，每周工作时间在 38 到 41 小时之间。

比较从事销售且高中毕业的丈夫们与职业为管理经理且高中毕业的丈夫们之间的工资分布，虽然在年龄分布和工作小时方面是相似的，但是对于经理们的这一组，属于收入高于 60,000 类的概率为 34%，相比之下销售人员的概率为 27%。要看到在顶部显示的这些概率，首先单击标签概率窗格右边 60000+ 收入类旁边的按钮，然后挑选左边的块状图。

蝴蝶花分类

在该数据集中，每个记录描述了 蝴蝶花 花朵的四个特征：萼片宽度、萼片长度、花瓣宽度、花瓣长度。每个 蝴蝶花 被进一步划分为 iris-setosa， iris-versicolor 或 iris-virginica 类型。目标就是把握每个蝴蝶花 类型的特征。

在运行分类器之前，单击 “工具管理器的分类工具”选项卡中的 “列重要性”选项卡；然后单击 “建议”，再单击*继续*。您得到了特征重要性的排序：萼片宽度、萼片长度和花瓣长度。在 “散点可视化工具”中，您可以将这些映射到坐标轴上，将 蝴蝶花 类型映射为颜色，并观察所得到的聚类。

文件 *iris-dtab.dtableviz* 显示了为解决该问题而导入的 “决策表分类工具”结构，通过在 *iris.schema* 上运行导入工具可产生该文件。

- Windows 用户可以在安装 MineSet 目录的 *examples**iris-dtab.dtableviz* 和 *data**iris.schema* 中找到这些文件。
- IRIX 用户可以在 */usr/lib/MineSet/examples/dtableviz/iris-dtab.dtableviz* 和 */usr/lib/MineSet/data/iris.schema* 下找到这些文件。

在 “决策表可视化工具”中，我们可以看到 “萼片宽度”是非常好的区分属性。当添加了花瓣宽度后，您可以看到对于那些萼片宽度在 0.75 到 1.65 之间的记录，所有 iris versicolor 的实例都出现在 “花瓣宽度 < 3.05”的组中。

在纯度不为 100% 的三个块中细化下寻。顶部的两个块中都包含了唯一的 iris-versicolor 实例，使它们无法纯净。对于 “花瓣宽度 <3.05”的块，很难孤立异常的 iris-versicolor 类型，但是，对于那个特殊的块，通过使用萼片长度可以孤立 iris versicolor。

蘑菇分类

文件 *mushroom-dtab.dtableviz* 显示了为解决该问题而导入的 “决策表分类工具”结构。通过在 *mushroom.schema* 上运行导入工具可以产生该文件。

- Windows 用户可以在 MineSet 安装目录中的 *examples**mushroom-dtab.dtableviz* 的 *data**mushroom.schema* 下找到这些文件。
- IRIX 用户可以在 */usr/lib/MineSet/examples/dtableviz/mushroom-dtab.dtableviz* 和 */usr/lib/MineSet/data/mushroom.schema* 下找到这些文件。

目的是在给定的数据集上预测那些蘑菇可以食用而那些是有毒的。在该数据集中有 8000 多条记录，运行该导入工具也许会花几分钟。注意，在预留三分之一进行准确性估计的默认模式下，将保留三分之一的记录用于测试模型。

每个蘑菇有很多特征，包括菌帽颜色、裂纹和气味。在“决策表可视化工具”中，气味和茎形状出现在顶层上。注意，气味可以很好地区分蘑菇是否可以食用。只有在没有气味并且茎形状为“向上扩大”时，才有些不好区分。所以我们自然地在这个单独的块上细化下寻。现在，我们看到只有那些具有这两个值的记录通过裂纹和菌褶大小的值被分解，注意裂纹和菌褶大小之间的作用，该作用很难用任何其它分类工具发现。

既然该数据集中的所有记录都是标称的，则这些值就通过它们在预测可食性过程中所起的作用进行排序。您也许想将这些值按照字母顺序或根据权重进行排序。要实现这一点，从名称顺序菜单中选取合适的方法。如果您单独考虑裂纹或菌褶大小，那么您将不能预测完全可食用或有毒蘑菇的大类，但是将它们一起考虑，我们可以看到，如果没有裂纹并且菌褶较大，那么该类型中的 814 个蘑菇可以食用。相反地，如果有分裂而且均褶较窄，那么该类型中的所有 11 个蘑菇都有毒。要想区分另外两种情况，我们将必须进一步细化下寻。

在“决策表可视化工具”中，向右移动“% 权重阈值”滑动条。最终，那些带有霉味的蘑菇对应块状图将从画面中删除。原因是只有不到 1% 的记录标记为“气味 = 霉味”。

党派隶属

该数据集包含投票记录。目的是在有关关键投票数据的条件下找出议员所属党派。数据集包括了由*议会季度年鉴*（*CQA*）所鉴定的美国众议院每个议员在 16 次关键表决中的投票。 CQA 列出了九中类型的投票：投票赞成、约定赞成以及宣布赞成 （这三个简称为是）；投票反对、约定反对以及宣布反对 （这三个简称为否）；投票表明避免利益冲突以及不投票或以其他方式表明立场 （这三个简称为未知处理）。

在运行分类工具之前，观察 16 个投票结果看是否能确认哪个特征最重要，然后运行“决策表可视化工具”。对于该数据集，您也许想要按照字母顺序对值进行排列，所以“是”类型的投票都出现右上角，而“不”类型的投票则在左下角。

在顶层，我们可以看到 “合成燃料公司缩减”和 “医生费用冻结”。在这些变量中存在着一个令人惊奇的关系，用任何其它模型都几乎不可能得出。除三个以外其它所有投票反对医生费用冻结的都是民主党，几乎每个民主党成员既投票反对费用冻结同时又投票反对合成燃料缩减 （206 个中只有 3 个不符合该模式）。奇怪地是，除 5 个以外所有投票赞成医生费用冻结的共和党成员，都投票反对合成燃料缩减。在这些风马牛不相及的问题中存在的奇怪关系暗示着这些结果以某种方式结合着，而这需要更深入的调查。

顶层中大多数块图除了中间的一个 （只包含 6 个记录）以及在两个问题上都投赞成票的一个代表以外几乎都是纯净的。在这里我们可以在另一层上细化下寻来区分该组中代表的政治倾向。利用 “反卫星实验禁令”和 “采纳财政议案”对该组的 55 个代表进一步区分。

文件 *vote-dtab.dtableviz* 显示了为解决该问题而导入的 “决策表分类工具”。在 *vote.schema* 上运行导入工具可产生该文件。

- Windows 用户可以在安装 MineSet 目录的 *\examples\vote-dtab.dtableviz* 和 *\data\vote.schema* 中找到这些文件。
- IRIX 用户可以在 */usr/lib/MineSet/examples/dtableviz/vote-dtab.dtableviz* 和 */usr/lib/MineSet/data/vote.schema* 下找到这些文件。

乳腺癌诊断

乳腺癌数据集包含了有关妇女接受乳腺癌诊断的信息。每个记录包含病人的细胞大小、肿块厚度以及边缘附着力等属性的记录。最终的属性是诊断为恶性的还是良性的。该文件 *breast-dtab.dtableviz* 显示了为解决该问题而导入的 “决策表分类工具结构”，通过在 *breast.schema* 上运行导入工具可以产生该文件。

- Windows 用户可以在安装 MineSet 目录的 *\examples\breast-dtab.dtableviz* 和 *\data\breast.schema* 下找到这些文件。
- IRIX 用户可以在 */usr/lib/MineSet/examples/dtableviz/breast-dtab.dtableviz* 和 */usr/lib/MineSet/data/breast.schema* 下找到这些文件。

在 “决策表可视化工具”中，等级的顶部显示了细胞形状的分割和均匀性的形状。如果两个属性同时都为低值，那么所给的样例有 99.2% 为良性；相反，两个值都比较高的训练记录 100% 为恶性。

在不纯的四个块图上细化下寻。现在，边缘附着力和裸露核也参与区分过程。在该层的每个块图中记录更少；其结果是噪声更多，并且更难掌握趋势。边缘附着力和裸露核都高的值更倾向于恶性肿瘤，但不确定。注意，裸露核的第一个值为空。这些空值块图的分布比其它的块图更值得怀疑，所以您也许希望通过取消选择 “查看 > 显示空值”将它们隐藏。

如果您总体细化下寻两个层以上时，您可以注意到一些有趣的特征。这些块图非常小，有很大的多维空间区域是空的。只有一些小区域，其中很多记录都进行了聚类。当所有的值都低时，就会出现一个巨大尖峰 （100% 良性）。尖峰可占数据的 20%。

甲状腺机能减退诊断

甲状腺机能减退数据集与乳腺癌的一样。文件 *hypothyroid-dtab.dtableviz* 显示了为解决该问题而导入的 “决策表分类工具”结构。通过在 *hypothyroid.schema* 上运行导入工具可产生该文件。

- Windows 用户可以安装在 MineSet 目录的 *\examples\hypothyroid-dtab.dtableviz* 和 *\data\hypothyroid.schema* 中找到这些文件。
- IRIX 用户可以在 */usr/lib/MineSet/examples/dtableviz/hypothyroid-dtab.dtableviz* 和 */usr/lib/MineSet/data/hypothyroid.schema* 中找到这些文件。

在数据集中有 3，163 条记录，其中的绝大多数并不具有甲状腺机能减退 （95.45%）。这就意味着预测否定的判断，这在多数时间里是对的。但是，我们所担心的正是那些模型预测为健康而实际具有甲状腺机能减退的人们。这种假否定是非常严重的。

为了避免假否定，在这种情况下，您也许想调节损失矩阵使后验概率向预测甲状腺机能减退方面倾斜。将本来患有疾病的人诊断为健康要付出很高的代价，而将实际健康的人诊断为有病只意味着不得不进行更准确的检查或对他们进行不必要的治疗。

在该数据集中使用 “决策表导入工具”，我们注意到：

- 如果 fti 为空，那么 tbg 就不为空并且 t3 几乎总为空 （再细化下寻一层）。
- 除了两个例外，当 t4u 为空时 fti 也为空。
- 当 fti 为低值时甲状腺机能减退更易发生。

Pima 糖尿病诊断

该数据集是关于利用统计学对糖尿病进行诊断的问题的，它是从 Arizona， Phoenix 的美州部落中收集而来的。其目的是根据一些医学特征 （例如，血压、体重、葡萄糖水平以及年龄）来确定一个病人是否患有糖尿病。

文件 *pima-dtab.dtableviz* 显示了为解决该问题而导入的 “决策表分类工具” 结构。通过在 *pima.schema* 上运行导入工具可产生该文件。

- Windows 用户可以在安装 MineSet 目录的 *examples**pima-dtab.dtableviz* 和 *data**pima.schema* 中找到这些文件。
- IRIX 用户可以在 */usr/lib/MineSet/examples/dtableviz/pima-dtab.dtableviz* 和 和 */usr/lib/MineSet/data/pima.schema* 下找到这些文件。

使用 “决策表可视化工具” 后我们注意到：

- 在细节分支的第二层上您可以看到小于 28 岁的女性几乎没有人怀过六次孕的 （只有 4 人）。您可能希望从这些记录中得到关于这四人的有关信息。
- 血糖浓度高、体重过大就指示了发生糖尿病的可能性很大。

DNA 边界

文件 *dna-dtab.dtableviz* 显示了为解决该问题而导入的 “决策表分类工具”结构。通过在 *dna.schema* 上运行导入工具可产生该文件。

- Windows 用户可以在安装 MineSet 目录的 *\examples\dna-dtab.dtableviz* 和 *\data\dna.schema* 中找到这些文件。
- IRIX 用户可以在 */usr/lib/MineSet/examples/dtableviz/dna-dtab.dtableviz* 和 */usr/lib/MineSet/data/dna.schema* 中找到这些文件。

在 DNA 数据集中有 3,186 条记录。 DNA 边界是从分子生物学领域的角度来界定的。结合连接处是 DNA 链上的点，在其上 “多余的” DNA 在蛋白质产生过程中已经被删除。我们的任务是识别 exon/intron 之间的界限 （被称为 EI 位置）；以及 intron/exon 之间的界限 （被称为 IE 位置）；或两者都不是。 IE 边界被称为 “接受者”而 EI 边界为 “捐献者”。这些记录最初从 GenBank 64. （*genbank.bio.net*) 中得到，其属性提供了 60 个核苷酸的观察窗口。分类点位于窗口的中点，这样就保证了联结处的每一边为 30 个核苷酸。

从 “决策表可视化工具”中，您可以看到一个另人惊奇的模式，而在任何其它分类工具模型中都不很明显。顶层中在 left_01 和 right_02 之间存在着很明显的相互作用。Exon/intron 只在 right_02 = T 时才存在。而 Intron/extron 只在 left_01=G 时才存在。对于 left_01 和 right_01 的其它值只有很少的份额结合。

证据可视化工具样例文件

下面的样例显示了分类工具如何发挥作用。各样例都与 MineSet 所提供的样例数据集相关。通过运行导入工具，您可以产生下面将要描述的 *.eviviz* 文件。

数据文件可以通过打开 *data* 目录中相应的 *. shema* 文件而被装入 *Minset*，（例如，*churn.schema*）。分类工具可视化文件，具有 *.eviviz* 扩展名，可以从 *examples* 目录中找到并打开。

客户波动

*客户波动*是指一个电话用户放弃一个公司而转投另一个公司。该样例显示了是什么使用户变更电话公司。

文件 *churn.schema* 和 *churn.data* 用于产生这个样例。要将数据文件装入 MineSet，打开 *.schema* 文件。

- Windows 用户可以在安装 MineSet 目录 *examples**churn.data* 和 *data**churn.schema* 中找到这些文件。
- IRIX 用户可以在 */usr/lib/MineSet/examples/eviviz/churn.data* 和 */usr/lib/MineSet/data/churn.schema* 中找到这些文件。

文件 *churn.eviviz* 显示了将客户波动的属性作为标签而导入的分类工具结构。该分类工具的误差率为 12%。 14.1% 的记录代表了波动的用户。两个最重要的属性，白天总分钟数和白天总费用，明显地相关。如果从 “深层导入工具” 选项中选择 “自动特征选择” 之后运行了导入工具，只利用四个属性 （白天总费用、服务电话数目、声音邮件业务以及声音邮件消息数目）就可将误差率降到 10.5%。日总费用超过 53.78 的所有 29 个用户都客户波动了。

大量的用户服务电话也是客户波动的预测工具。大量用户服务电话可能指示了在使用复杂设备或接收不可靠服务时出现了障碍。具有国际业务的用户也更有可能客户波动。有些州的用户比其它州的更有可能客户波动， California 和 New Jersey 有最多的客户波动率，而 Virginia 为最少。要观察拥有记录数超过记录总数的 2% 的州，则将 “% 权重阈值” 滑动条滑到右边，这样将从显示中删除了绝大部分州的值。如果您也选择了 “名称顺序 > 加权”，那么具有最多记录的州 West Virginia （WV）出现在最左面。在区分客户波动的过程中，许多属性 （列表的低部）是没有用的。注意，白天费用是很好的预测工具，而夜间费用则不是。

轿车的产地

*轿车*数据集包含了从 70 年代到 80 年代初不同型号轿车的信息。属性包括*重量*、*加速能力*、以及每加仑英里数（*mpg*）。文件 *cars.eviviz* 显示了为解决该问题而导入的“证据分类工具”结构。通过在 *cars.schema* 运行导入工具可以产生该文件， *cars.schema* 的标签设为产地 （日本、美国、欧洲）并且将汽缸列转变为字符类型。汽缸列转换为字符类型是为了观察到所有的值并且避免自动离散化。

- Windows 用户可以在安装 MineSet 目录 *\examples\cars-eviviz* 和 *\data\cars.schema* 中找到这些文件。
- IRIX 用户可以在 */usr/lib/MineSet/examples/eviviz/cars.eviviz* 和 */usr/lib/MineSet/data/cars.schema* 中找到这些文件。

如果您有一个关于轿车属性的数据集，您可能想知道不同产地轿车的特征。从右边饼图中标签值的分布可以看出，数据集中的多数轿车都是美国制造的 （62.5%）而一小部分产自日本 （20.2%）和欧洲 （17.3%）。因为每个品牌都只与产地国家相关，因此很明显，品牌是产地最好的预测工具。因为这个原因，它具有最高的重要性并且在列表的顶部。通过观察圆饼图的高度，可以看出许多轿车有四个汽缸，大多数重量小于 3000 lbs 并且大多数可以在 13 至 20 秒的范围内将速度提到每小时 60 英里。

观察个别属性值份额的分布。如果轿车的引擎大小为 >169 立方英寸，那么可以几乎肯定它产自美国；并且肯定不会产自日本。另一个饼图显示了美国轿车通常有 6 或 8 个汽缸，每加仑的英里数较低，大马力 （超过 134），大重量 （超过 2981 lbs）并且具有较强的加速能力。日本汽车具有更好的汽油英里数，3 或 4 个汽缸（一小部分为 6 汽缸）并且具有较小的引擎。如果您在 “标签概率窗格”中单击了 “欧洲”，您可以看到代表欧洲轿车的证据的条形图。例如，如果轿车有五个汽缸则说明肯定产自欧洲。但是，相应饼图的高度显示在数据中只有三个轿车有五个汽缸。如果轿车有较好的英里数，则很明显它产自欧洲。如果轿车的英里数小于 41，那么它产自欧洲的机会就为 83%。如果一个轿车是欧洲产的，那么其公里数好于 41 mpg 的机会为 10.4%。但是没有美国轿车和只有 2% 的日本轿车其 mpg 在这个范围内 ，所以这是产地为欧洲最强的证据。

假如您只知道一个轿车有 40mpg 并且重量为 3000lbs，而又想预测轿车的产地。那么选择合适的饼图 （或块图）：mpg=30.95-41.15 并且 weightlbs=2981.5+。右边的概率结果显示为美国 84%，欧洲 16%。因为在训练集中没有重量为 lbs>2981.5 的日本轿车，所以不可能是日本轿车。如果您选中拉普拉斯校正选项 （值为 0.5）并再次运行导入工具，您将得到不同的结果：欧洲的概率为 16%，美国的概率为 82%，而日本的概率为 2%。这是因为拉普拉斯校正可以防止块图中的任何份额完全变为 0。可以肯定的是，对于日本为什么不能制造重量超过 2981 lbs 的轿车没有一个很基本的原因，因此，当概率 （饼图）相乘时，并没有删除预测为日本轿车的概率。

预测性别

*成人*数据集包含了有关从业成年人的信息。该数据集是从美国调查局中抽调的。它包含了年龄超过 16 岁的人的有关信息，他每周工作至少一小时，每年收入超过 $100 的人员信息。您也许想知道怎样将男性和女性特征化。文件 *adult-sex.eviviz* 显示了为解决该问题而导入的 “证据分类工具”结构。在删除婚姻关系列 （使得分类工具更加琐细）之后，将标签设为性别，并通过在 *adult.schema* 之上运行导入工具可以产生该文件。

- Windows 用户可以在安装 MineSet 目录 *\examples\adult-sex.eviviz* 和 *\data\adult-sex.schema* 中找到这些文件。
- IRIX 用户可以在 */usr/lib/MineSet/examples/eviviz/adult-sex.eviviz* 和 */usr/lib/MineSet/data/adult.schema* 中找到这些文件。

在 “证据可视化工具”中，“标签概率窗格”显示了从业者为男性的先验概率要高于职业女性。

- 婚姻状况对于性别来说是最重要的预测工具。如果一个工人是已婚市民，那么是男性的概率很大。但是，如果一个工人丧偶并且正在工作，那么很有可能为女性。
- 所列出的第二个属性显示了职业。研究这些可以知道在特定的性别中流行什么职业。从左到右按照男性所占比例的降低列出了不同的职业：军队 （100%），手工维修 （95%），交通运输 （95%）以及种植渔业 （94%）。而女性则代以房屋内务 （94%）以及主管办事员 （67%）。通过单击 “标签概率窗格”中 “女性”旁边的按钮，然后在 “职业 = 主管办事员”上移动鼠标，您可以看到 23% 的女性其工作为主管办事员。相反地，假设一个人的工作为主管办事员，那么其性别为女性的概率为 67%。

  假设您要在 “丧偶”并且 “职业 = 主管办事员”的条件下找出是女性的概率。（处于选取模式时）当在 “女性”旁边的盒子之上移动鼠标时，通过单击该属性值并且阅读顶部 的文本可以达到上面所说的目的。

- 如果一个人的工作种类是 “私人企业”或 “非私人企业”，那么其为男性的概率就比较高。相反地，如果一个人的工作种类为 “州政府工作人员”，那么其为女性的条件概率就比较高，但是后验概率 （在考虑先验概率之后）并不高 （单击它并且观察右边的后验概率）。在为州政府工作的条件上，女性所占的份额增加了，但是并没达到由此预测其为女性的那种程度。

  旋转该视图后，您可以通过观察图的高度得出大多数人在私有企业中工作。
- 通过观察总收入属性，您可以看出收入范围越高，为男性的概率越大。
- 教育程度并没有提供多少信息，除了博士学位，因为在其中男性更多一些。
- 不同的职业对于男性或女性有不同的分布。
- 种族属性显示，在条件概率中非裔美国女性工作的百分比要高于其他种族的百分率。单击该值可以看到，在男性和女性之间后验概率大致相等。
- 在该数据集中男性每周工作的小时数要多于女性。

工资因素

如果您有一个从业成年人的数据集，您可能想知道什么因素影响着工资。首先，将 gross_income 分为五组，阈值为 10,000、 20,000、 30,000 以及 60,000。每个记录的属性为这五组值中的一个。您可以运行 MineSet 分类工具来帮助确定是什么因素影响了工资。文件 *adult-salary.eviviz* 显示了为解决该问题而导入的 “证据分类工具”结构。在 *adult.schema* 上运行导入工具可以产生该文件，并利用用户定义的阈值将 gross_income 分为五组。

- Windows 用户可以在安装 MineSet 目录的 *examples**adult-salary.eviviz* 和 *data**adult.schema* 中找到这些文件。
- IRIX 用户可以在 */usr/lib/MineSet/examples/eviviz/adult-salary.eviviz* 和 */usr/lib/MineSet/data/adult.schema* 中找到这些文件。

“证据导入工具”中的属性按照重要性划分等级；这样，可以认为婚姻关系、婚姻状况、年龄、职业、教育程度、每周小时数以及性别是重要的。因为标签是数值型的，所以可以用连续的色谱将颜色分配给每个类。红色分配给最高的组 （60,000+）。根据份额的大小，在 “标签概率窗格”中列出了类标签。当您单击 “主窗口”中的值时，类标签值的顺序发生改变以保持该标签在顶部为最大预测类。

- *婚姻关系*显示了丈夫和妻子比未婚工人或无家庭的工人所挣的钱多。妻子的收入比丈夫的收入略高。
- *婚姻状况*显示出大多数人结过婚 （左数第二个图很高）。已婚工人比未婚的所挣的工资高。
- *年龄*显示了年龄是一个关键的因素。直到 61 岁以前，这时许多人都已退休，挣 $50,000 以上的概率随年龄的增长而增加。
- 不同的职业对应不同的概率。管理型和专业型的工作为每年收入超过 $60,000 提供了证据。
- *教育程度*是重要的因素。当只考虑教育程度时，收入超过 $60,000 的证据大都来自教育程度为硕士、博士或从专业学校进入大学的那些工人。
- 每周小时数显示了工作的小时数越多，挣得更多钱的证据越高。
- *性别*显示了女性会为每年收入少于 $60,00 提供更多的证据。
- 调节*权重百分数*滑动条来删除 native_country 的值，权重较低的教育程度以及职业值也被删除。

蝴蝶花分类

在该数据集中，每个记录描述了蝴蝶花花朵的四个特征：萼片宽度，萼片长度，花瓣宽度，花瓣长度。每个蝴蝶花 被进一步划分为 iris-setosa， iris-versicolor 或 iris-virginica 类型。我们的目标就是把握每个蝴蝶花 类型的特征。

在运行分类工具之前，单击 “工具管理器的分类工具”选项卡中的 “列重要性”选项卡；然后单击*继续*，您得到这些特征重要性的等级顺序：萼片宽度、萼片长度和花瓣长度。在 “散点可视化工具”中，您可以将这些映射为坐标轴， iris_type 映射为颜色，并观察自然聚类结果。

文件 *iris.eviviz* 显示了为解决该问题而导入的 “证据分类工具”结构。通过在 *iris.schema* 上运行导入工具可产生该文件。

- Windows 用户可以在安装 MineSet 目录的 *\examples\airis.eviviz* 和 *\data\iris.schema* 中找到这些文件。
- IRIX 用户可以在 */usr/lib/MineSet/examples/eviviz/iris.eviviz* 和 */usr/lib/MineSet/data/iris.schema* 中找到这些文件。

在 “证据可视化工具”中，我们可以看到萼片长度和萼片宽度是极好的区分属性，而花瓣长度和花瓣宽度略差。向右移动重要性阈值滑动条可以看到以花瓣为基础的属性首先消失。

蘑菇分类

文件 *mushroom.eviviz* 显示了为解决该问题而导入的 “证据分类工具”结构。通过在 *mushroom.schema* 上运行导入工具可以产生该文件。

- Windows 用户可以在安装 MineSet 目录中的 *\examples\mushroom.eviviz* 和 *\data\mushroom.schema* 中找到这些文件。
- IRIX 用户可以在 */usr/lib/MineSet/examples/eviviz/mushroom.eviviz* 和 */usr/lib/MineSet/data/mushroom.schema* 中找到这些文件。

我们的目的是在已知的数据集上预测哪些蘑菇可以食用而哪些是有毒的。在该数据集中有 8000 多条记录，运行该导入工具也许会花几秒种。注意，预留三分之一做准确性估计的默认模式下，将保留三分之一的记录用于测试。

每个蘑菇有很多特征，包括菌帽颜色、裂纹和气味。默认情况下，“证据导入工具”通过重要性 （也就是在预测标签过程中的作用）将属性排序。气味和孢子颜色出现在列表的顶部，这是因为，这些属性，块图中的分布随着值的不同而有很大的不同。既然在该数据集中的所有记录都是标称的，那么这些值就通过它们在预测可食性过程中所起的作用从左到右进行排序。您也许想将这些值按照字母顺序或根据权重进行排序。要达到这一点，从名称顺序菜单中选择合适的方法。通过将鼠标改为箭头 （单击主画面右上角的箭头图标或按下 Esc 键），然后单击右边窗格中有关该类的按钮，您就可以看到有毒蘑菇的特征。较高的条形图与指示蘑菇有毒的值相关。

在 “证据可视化工具”中，向右移动 “细节”滑动条，可以从从画面中删除具有最低重要性的属性。目前所知，最重要的属性是气味，它的重要性为 92 ；而所有其它属性的重要性都小于 48。几乎所有的值都是很好的区别器，但是如果没有气味 （无），那么就会出现两类的混淆。“证据可视化工具”可让您观察也许是关键的特定值，即使属性本身并不总是重要。例如， stalk_color_below_ring 不是理想的区分属性，这是因为在大多数时间里它的值为白色。因为颜色为白色的可食性蘑菇和有毒蘑菇数量相等，因此白色没有预测力。当 stalk_color_below_ring 为灰色或浅黄色，那么可以进行极准确的区分，但是很少有蘑菇带有这种颜色。

党派隶属

该数据集包含投票记录。我们的目的是在有关关键投票数据的条件下找出议员所属党派。数据集包括了由*议会季度年历*（*CQA*）所鉴定的美国众议院每个议员在 16 次关键表决中的投票。 CQA 列出了九中类型的投票：投票赞成、约定赞成以及宣布赞成 （这三个简称为是）；投票反对、约定反对以及宣布反对 （这三个简称为否）；投票表决避免利益冲突以及不投票或以其它方式表明立场 （这三个简称为未知处理）。

在运行分类工具之前，观察 16 个投票结果看能否区分哪个特征最重要，然后运行 “证据可视化工具”。对于该数据集，您也许想按照字母顺序将值进行排列，以使投票反对的在最左边，未决定的在中间，而代表是的在右边。

一些结果清楚地指示了一个人的党派倾向。民主党投票赞成医生费用冻结以及对 El Salvador 的援助，而共和党投票赞成采纳财政议案以及对 Nicaragua 中的 Contras 进行援助。

在党派中移民问题并没有引起分歧，毫无疑问，政客们在这上面有明确的立场，因为 235 人中只有 7 人对该问题无所适从。

文件 *vote.eviviz* 显示了为解决该问题而导入的 “证据分类工具”结构。在 *vote.schema* 上运行导入工具可产生该文件。

- Windows 用户可以在安装 MineSet 目录的 *examples**vote.eviviz* 和 *data**vote.schema* 中找到这些文件。
- IRIX 用户可以在 */usr/lib/MineSet/examples/eviviz/vote.eviviz* 和 */usr/lib/MineSet/data/vote.schema* 中找到这些文件。

乳腺癌诊断

乳腺癌数据集包含了有关妇女接受乳腺癌诊断的信息。每个记录包括了病人细胞大小、 z 肿块厚度以及边缘附着力等属性，最终的属性是诊断为恶性的还是良性的。文件 *breast.eviviz* 显示了为解决该问题而导入的 “证据分类工具”结构。通过在 *breast.schema* 上运行导入工具可以产生该文件。

- Windows 用户可以在安装 MineSet 目录中 *examples**breast.eviviz* 和 *data**breast.schema* 中找到这些文件。
- IRIX 用户可以在 */usr/lib/MineSet/examples/eviviz/breast.eviviz* 和 */usr/lib/MineSet/data/breast.schema* 中找到这些文件。

在 “证据可视化工具”中，您可以看到 sample_code_number 处在一个被均分为两部分的区间里，也就意味着它并不能指示乳腺癌是良性的还是恶性的。

甲状腺机能减退诊断

甲状腺机能减退数据集与乳腺癌的一样。文件 *hypothyroid.eviviz* 显示了为解决该问题而导入的 “证据分类工具”结构。通过在 *hypothyroid.schema* 上运行导入工具可产生该文件。

- Windows 用户可以在安装 MineSet 的目录的 *\examples\hypothyroid.eviviz* 和 *\data\hypothyroid.schema* 下找到这些文件。
- IRIX 用户可以在 */usr/lib/MineSet/examples/eviviz/hypothyroid.eviviz* 和 */usr/lib/MineSet/data/hypothyroid.schema* 下找到这些文件。

在数据集中有 3163 条记录，其中的决大多数没有甲状腺机能减退症状 （95.45%）。这就意味着预测否定的判断在多数时间里是对的。但是，我们所担心的是那些模型预测为健康而实际具有甲状腺机能减退的人们。这种假否定是非常严重的。

观察一下 tsh 介于 6.35 和 27.5 之间的块图。它显示了许多关于甲状腺机能减退的证据。然而，单击它时，在右边的后验概率饼仍然预测为 “否定”，这是因为 “否定”的先验概率太大。

为了避免假否定，在这种情况下，您也许想调节损失矩阵使后验概率向预测甲状腺机能减退方面倾斜。将本来患有疾病的人诊断为健康就要付出很高的代价，而将实际健康的人诊断为有病只意味着进行更准确的检查或对他们进行不必要的治疗。

在 “证据可视化工具”中，您可以看到 fti 是很重要的。，头两个区间提供了很多关于甲状腺机能减退的证据。

Pima 糖尿病诊断

该数据集是关于利用统计学对糖尿病进行诊断的问题，它是从 Arizona， Phoenix 的美州部落中收集而来的。其目的是根据一些医学特征 （例如，血压、体重、葡萄糖水平以及年龄）来确定一个病人是否患有糖尿病。

文件 *pima.eviviz* 显示了为解决该问题而导入的 “证据分类工具” 结构。通过在 *pima.schema* 上运行导入工具可产生该文件。

- Windows 用户可以在安装 MineSet 目录中 *\examples\pima.eviviz* 和 *\data\pima.schema* 中找到这些文件。
- IRIX 用户可以在 */usr/lib/MineSet/examples/eviviz/pima.eviviz* 和 */usr/lib/MineSet/data/pima.schema* 中找到这些文件。

在 “证据可视化工具” 中，您可以看到许多无关的属性。当 plasma_glucose 增加时，患糖尿病的概率也随着增加。当年龄＞ 27 时，如果怀孕的次数较高 （超过 6），也将成为很好的指示器。

DNA 边界

文件 *dna.eviviz* 显示了为解决该问题而导入的 “证据分类工具” 结构。通过在 *dna.schema* 上运行导入工具可产生该文件。

- Windows 用户可以在安装 MineSet 目录中 *\ examples\dna.eviviz* 和 *\data\dna.schema* 中找到这些文件。
- IRIX 用户可以在 */usr/lib/MineSet/examples/eviviz/dna.eviviz* 和 */usr/lib/MineSet/data/dna.schema* 中找到这些文件。

在 DNA 数据集中有 3,186 条记录。 DNA 边界是从分子生物学领域的角度来界定的。结合连接处是 DNA 链上的点，在其上 “多余的” DNA 在蛋白质产生过程中已经被删除。我们的任务是识别 exon/intron 之间的界限 （被称为 EI 位置）；以及 intron/exon 之间的界限 （被称为 IE 位置）；或两者都不是。IE 边界被称为 “接受者” 而 EI 边界为 “捐献者”。这些记录最初从 GenBank 64. （*genbank.bio.net*) 中得到，其属性提供了 60 个核苷酸的观察窗口。分类点处于窗口的中点，这样就保证了联结处的每一边为 30 个核苷酸。

从 “证据可视化工具” 中，您可以看到位于中心附近的属性都被选为很重要，而远离结合处的属性就不是很重要。

如果您单击并选择了左边与 “left_01:G” 和 “left_02:A” 对应的窗格中的图表，那么右边标签概率窗格中的饼图将改变以显示由 “证据分类工具” 所预测的每个类的概率分布。在已知这两个值的条件下，圆饼图显示了建立的证据模型将最高的概率分配给了 “intron/exon”，其后是 “exon/intron” 和 “none”。

如果您调用了自动特征选择，虽然运行时间增加了很多 （有时以小时记），但准确性却略有改进。在这种情况下，运行自动特征选择时应有所考虑。

地图可视化工具样例文件

所提供的配置和数据文件样例展示了 “地图可视化工具”的特征和功能。

- Windows 用户可以在安装 MineSet 目录的 *\examples\mapviz* 下面找到这些文件。*.gfx* 和 *.hierarchy* 文件可以在 *\config\mapviz* 中找到。
- IRIX 用户可以在 */usr/lib/MineSet/examples/mapviz 中找到这些文件。.hierarchy* 和 *.gfx* 文件可以在 */usr/lib/MineSet/mapviz/gfx_files* 中找到。
  - *blocks.mapviz*、 *blocks.data*、 *blocks.gfx*、和 *blocks.hierarchy*
    这个简单的样例显示了四个相邻的方块。每个方块的高度和颜色随着 *blocks.data* 中基本数据的不同而不同。您可以使用鼠标中键来概化上寻查看上部或下部方块对的组合值，接着继续概化上寻查看概化成一个方块图的组合值。您可以利用鼠标右键细化下寻来观察重新出现的更细粒度对象。
  - *population.australia.mapviz*， *population.australia.data*，*australia.states.gfx*，和 *australia.states.hierarchy*
    该数据文件为澳大利亚每个州及地区包含一条记录。每一行包含了三个以制表符分隔的项：有关州或地区、人口值以及地区面积。

    该样例图形显示了 1991 年澳大利亚各州和地区的人口和人口密度。图形对象的高度代表了相关的人口；颜色代表了相关的人口密度。在显示结果下部的图例描述了颜色范围及其相关值。
  - *population.canada.mapviz*、 *population.canada.data*、 *canada.provinces.gfx*、和 *canada.provinces.hierarchy*
    该数据文件为加拿大的每个省和地区包含一条记录。在该例中，每行包含了 13 个以空格分隔的值 （每个值代表 1871 至 1991 年之间的每个十年）。

该样例图形显示了从 1871 至 1991 年间，以十年为间隔加拿大各省和地区的人口和人口密度。动画控制面板可让您动态查看经历一段时间后的数据集的变化。动画操作在第 10 页的 “动画控制面板”中有解释。

- *population.europe.mapviz*、 *population.europe.data*、 *europe.countries.hierarchy* 和 *europe.countries.gfx*
  当进行图形展示时，所显示的是 1992 年中西部欧洲的人口与人口密度。
- *population.usa.mapviz*，*population.usa.data*， *usa.state.gfx*，和 *usa.state.hierarchy*
  当进行图形显示时，所显示的是从 1770 至 1990 年间美国的人口与人口密度。该动画控制可让您动态查看人口和密度随时间的变化。
- *population.usa.city.mapviz*， *population.usa.city.data*， *usa.state .gfx*、*usa.state.hierarchy*、 *usa.city.gfx*、*usa.city.hierarchy*
  和 *usa.state.gfx* 文件指定美国，显示为背景。 *usa.city.gfx* 文件确定了在该背景上的城市位置。 *.data* 文件说明了每个城市的人口。

  该样例图形显示了 1950 至 1990 年间美国最大的 48 个城市的人口。没有数据映射为颜色。动画控制可允许您动态查看人口随时间的变化。
- *perhouse.perage.mapviz*、*perhouse.perage.data*、 *usa.state.gfx* 和 *usa.state.hierarchy*
  该样例图形显示了从 1988 年 7-8 月到 1991 年 5-6 月间用户的家庭消费数据。颜色映射为消费 ' 家庭成员的性别；高度代表一定时间段和年龄组中每个家庭平均花费的美元数量。该数据有两个独立维：时间和年龄。汇总窗口中颜色浓度最高的区域指示了最高的消费，称为 “1989 年 5-6 月 （年龄：30-39）” 以及 “1990 年 5-6 月 （年龄：30-39）。
- *telecom.mapviz*、*telecom.data*、 *usa.city.lines.gfx*、 *usa.city.lines.hierarchy*、*usa.state.gfx* 和 *usa.state.hierarchy*
  该样例图形显示了带有弧线的平面地图。这些线连接两个端点。这些线具有不同的数值、宽度和颜色。在该例中，宽度和颜色是随机的；但是它们与端点之间连接的持久度是相关的。

- *fasta.m.data*，*fasta.m.mapviz*，*fasta.m.gfx*，和 *fasta.m.hierarchy*

  该样例数据文件包含了两个完整染色体组（经 Tom Flores 博士允许发表，欧洲生物信息研究所）间序列比较的部分结果。当进行图形显示时，科学家可以很快找出并确定两组基因之间相似的部分。如上所示的在可视化数据开采中显示大量信息的能力可以被扩展以用来包含更多有关个体基因的信息。科学家可以更容易的开采数据，因而可以更好地理解相似基因序列的目的和功能。

  在该样例中，“图”是生物有机体的圆形基因，被称为*生殖道枝原体*（MG）。MG 基因分为 500 个相等段，每段代表了基因中的 1000- 核苷酸序列。滑动条选择了第二个基因的一段，被称作*流感嗜血菌*（HI），用于两个基因之间的交叉比较。“动画控制面板”中的“汇总窗口”指示了哪一段显示出了最大的相似性，并且您可以移动滑动条来仔细研究那些有趣的特殊段。因此“图”上块的高度和颜色指示了每个 MG 段和每个 HI 段的相对相似性，而较高的块对应于较高的相似性度量。相似性通过“倒数值”来衡量，其范围从 0.0 到 1.0。

选项树样例文件

下面的样例显示了“选项树导入工具”可能起作用的情况。每个样例与 MineSet 所提供的样例数据文件相关。通过运行导入工具，您可以产生下面描述的 *-odt.treeviz* 文件。文本述的每个任务，其方案和目标在“树可视化工具样例文件”中都进行了描述。这里对于几个样例数据集我们指出了“选项树”的优缺点。

注意： 拥有 *.schema* 扩展名的分类器可视化数据文件，位于客户工作站的 *data* 目录中。带有 *-odt.treeviz* 扩展名的可视化文件，位于客户工作站的 *examples* 目录中。要将数据文件装入 MineSet，请打开 *.schema* 文件。

- Windows 用户可以在安装 MineSet 目录的 *\examples* 和 *\data* 子目录下找到这些文件。
- IRIX 用户可以在 */usr/lib/MineSet/examples/treeviz* 和 */usr/lib/MineSet/data* 下找到这些文件。

客户波动

该数据集的 “选项树”显示出日总费用、日总分钟数以及用户服务电话对于根来说是好属性：它们几乎具有相同的估计错误率。根据优先权和对数据的理解，您可以选择下降到某个子树。注意，当右子树开始于用户服务电话时，第二个检验将在日总费用或日总分钟数 （在根的左选项上）上进行的。但是，由于已经在一个属性上出现了拆分，所以阈值将不会相同。

轿车的产地

该数据集的 “选项树”显示了几个对于根的好属性。它们包括：体积、汽缸、重量（lbs），速度 （mpg）及品牌。注意，根的估计误差率比其任何一个子节点都要低。

蝴蝶花分类

这是一个 “选项树”的应用效果比 “决策树”差的样例。“决策树”的根显示了 6% 的错误，而 “选项树”的根具有 8% 的错误，“选项树”的效果似乎差一些。但是：

- 误差估计的标准偏差相当高：3.88% 和 3.39%。统计中的经验估计法认为如果偏差小于两倍标准差，那么在 95% 的置信水平下，此偏差不是统计显著的。此样例中，两种分类模型间 2% 的偏差甚至不大于一倍标准差；因此，在 95% 的置信水平下，分类误差率很可能不具统计差异。

- 对于小文件 （ 蝴蝶花只有 150 条记录），不同随机种子得出不同结果。例如，将随机子变为 3 时，在不改变 “决策树分类工具”误差率的情况下，将会把 “选项树分类工具”的误差从 8% 降至 4% （记着要重置随机子）。这并*不*意味着已经产生了更准确的分类器，因为其误差估计并不稳定。由于只有 50 条记录用于检验，每一个错误就要占 2% 的错误率。 4% 和 8% 之间的差别只是多犯了两个或更多的错误。
- 对于小文件 （蝴蝶花有 150 条记录），可使用 MineSet 中的 “误差估计”选项。它在更窄的置信区间内会产生更好的估计。当您运行该模式，状态窗口显示了 “决策树”分类工具的误差率为 4.67% +/- 1.73%，而 “选项树”分类工具的误差率为 4.00% +/- 1.61%。在该样例中，差异也并不明显，但是 “选项树”效果略优。
- 即使 “选项树”的误差率较高，在分配概率估计方面它们也许 （通常也是）更好。对于该数据集，“决策树”的估计均方差为 3.94 ； “选项树”为 3.67 （虽然在 95% 的置信水平下差异并不明显）。

蘑菇分类

该数据集的 “选项树”显示了在根节点选择的五个选项其总估计误差率为 0。观察该结果，您也许更倾向于使用左选项 （菌斑），因为虽然准确性相似，但它比气味更容易检验 （导入的 “决策树”的根检验）。您也许想删除气味和菌褶大小两个属性，然后只是为了准确性 （使估计误差率为 0%）再建立正规的 “决策树”。

然而请注意，删除根选项而代以由 “决策树”选择的兄弟选项不能必然产生 “选项树”所显示的具有相同准确率的分类器。被删除的属性也许是树的下层要用到的属性。例如，从轿车数据集中删除品牌属性将明显地增加误差率，即使是在根部五个选项中有四个根本没有用到它。

党派隶属

该数据集与 蝴蝶花 数据集的表现十分相似。“选项树”与 “决策树”有同样的误差率。在 “估计误差”模式下，交叉检验估计显示在误差率和均方差方面，选项树比 “决策树”略好一点 （但在 95% 的置信水平下并不显著）。

乳腺癌诊断

对于*分类工具和误差估计*以及*估计误差模式*，“选项树”的误差率比 “决策树”略低，但是差异并不明显 （在 95% 置信水平下）。

甲状腺机能减退诊断

该数据集的误差率非常低 （小于 1%），但是这是因为被检查是否患有甲状腺机能减退的大多数人 （95%）并没有得病。如果我们利用损失矩阵来避免错误否定 （对实际患上了甲状腺机能减退但被诊断否定的情况给予 100 的惩罚），我们可以看到 “选项树”的损失将明显低于 “决策树”：182 比 523 （全部），或 0.17 比 0.5 （每条记录）。在 95% 的置信度水平下，差异很明显。

DNA 边界

对于该数据集，“选项树”比 “决策树”略微准确一些；但是通过观察根选项，您会注意到它会选择*左 1*、*2*，和*右 1*、*2*、*5*。如果已知背景知识，即与边界更为接近的属性可能会更重要一些，您也许想在*右 5* 上排除选项分割。在您将根选项的最大数目更新为 4 （从 5）之后，误差率从 5.65% 升至 6.59%。如果根不再将*右 5* 作为选项来使用，这也许很奇怪；而将根选项数目从 5 改为 4 的另一个结果就是也减少了在树的更深层中所出现的选项数目 （因为降低了参数）。这样使得其它 4 个子树的个别误差率增加了。但是，选项树的误差率 （在 95% 置信水平下）仍然明显好于 “决策树”的误差率 7.06% +/- 0.79%。

回归树样例文件

下述样例描述了可以使用回归树的情况，并且突出了 “回归树导入工具”的一些功能。每一个样例与 MineSet 所提供的样例数据文件相关。通过运行导入工具，您可以产生将在下面描述的 *-rt.regress files* 文件。

注意： 拥有 *.schema* 扩展名的数据文件，位于客户工作站的 *data* 目录中；带有 *-rt.treeviz* 扩展名的回归可视化文件位于客户工作站的 *examples* 目录中。要将数据文件装入 MineSet，请打开 *.schema* 文件。

- Windows 用户可以在安装 MineSet 目录的 *\examples* 和 *\data* 子目录下找到这些文件。
- IRIX 用户可以在 */usr/lib/MineSet/examples/treeviz* 和 */usr/lib/MineSet/data* 下找到这些文件。

客户波动

*客户波动*数据集包含了有关电信公司用户的通话模式的信息。在分类样例中，该数据集用于确定是什么因素致使用户波动或离开该公司转到竞争对手那边去。在回归样例中，我们将试图确定是什么因素影响了每天公司从用户手中索取的费用。

文件 *churn-rt.treeviz* 显示了在该数据集上产生并用于预测日总费用的 “回归树”。有趣的是，整个树只在一个属性 （日总分钟数）上产生分支；它不断深入地分割该属性并逐步缩小范围。这是因为日总分钟数正比于日总费用 -- 用户只为使用该系统的时间支付费用。“回归树”能够反映这个事实。

轿车耗油量

*轿车*数据集包含了从 70 年代到 80 年代初不同型号轿车的信息。该数据集中的属性包括重量、提速能力以及每加仑英里数 （mpg）。文件 *cars-rt.treeviz* 显示了在该数据集上导入的 “回归树”回归工具，并将每加仑英里数作为连续标签。

通过单击顶部节点，我们可以看到数据集中轿车的平均 mpg 在 23.5 左右。在数据集的 “回归树”中的第一个分割显示了对轿车英里数影响最大的因素是轿车本身的重量。“决策树”揭示了众所周知的事实，就是轿车的自重越大其英里数越低。通过观察基准节点的两个子节点，我们注意到右边的子节点比左边的更蓝一些，这就是说它的每加仑英里数更少。通过加亮该节点，我们可以看到轿车的重量小于 3018lbs. 时，其 mpg 大约为 28.3，而当重量大于 3018lbs 时大约为 16.6mpg。

进一步观察较重的轿车，我们注意到下一个分割出现在轿车的马力上，并且马力越大的轿车其英里数越低。下一层的分割出现在轿车制造的年份上，可以看出越新的轿车英里数越大。现在，让我们看一个不寻常的轿车。通过使用筛选面板，我们试图发现一个节点其平均 mpg <24 而最大值 >30。在筛选过程中，我们很快将树减少至一个节点，而轿车重量小于 3018 lbs，功率大于 77 马力，生产于 1980 以前。在这个范围里只存在一辆不寻常的轿车；通过选择节点上最右边的条并通过选项 > 显示原始数据菜单项进行追溯，我们找到了这辆轿车，它是 1978 产的 Dodge，重量约为 2000 lbs，83 hp，它可以达到每加仑 33.5 英里。

工资因素

*成人*数据集是从美国统计局获得的，其中包含有工作的成年人的信息。它包含了年龄超过 16 岁，每周工作至少一小时，每年收入超过 $100 的成人的信息。我们可以利用 “回归树导入工具”来确定是什么因素在影响一个人的工资；并在已知其它信息的条件下，对其应得的工资水平进行粗略的估计。

文件 *adult-rt.treeviz* 显示了在成人数据集上导入的 “回归树回归工具”，并将总收入作为连续标签。注意，该数据集非常大 （大约有 50000 条记录），因此在该数据集上导入回归工具要花费几分钟时间。

顶部节点的条形图是表示统计局数据中工资值的直方图。注意，可用数据量随工资层的降低而逐渐减少。我们有许多关于每年收入 $3,000 的人的数据，但是随着图的深入它们同样在减少。在直方图的最后部分该趋势发生逆转，该直方图指出存在相当数量的每年收入大约为 $100,000 的人的数据。这个差别也许是数据本身存在的趋势，也许是因为采样存在偏差。

“回归树”中第一个分割产生在年龄属性上。正如所预认的，年轻人所赚的钱通常比年长者少。从顶节点及其两个子节点，我们可以看到有关这三组人群的汇总统计特征。我们注意到在该项研究中，每个人的平均工资在 $33,500 左右，而年龄小于 27 岁的人其平均工资为 $14,300 左右；年龄大于 27 岁的人其平均工资在 $40,000 左右。

在年龄小于 27 岁的人群的下两次分割中，我们看到回归树先将他们分成两个部分：23 岁及以下的人群和 23 岁以上的人群。有意思的是，这两个分支上的后续分割都产生在每周工作小时属性上，这显示出对于这两个年龄段来说，人们工作的时间越多，总收入也就越高。

现在将注意力放在年龄超过 27 岁的人群上，我们发现在所受教育程度上迅速产生了三个分割。受过 13 年以上教育的人 （与学士学位相对应）会挣更多的钱。通过观察被教育程度分割的两个子节点，我们可以看到每年收入的 $90,000 左右的大多数人都至少接受了某种高等教育。

我们可以使用筛选面板，迅速找出每年平均收入超过 $50,000 的人群的位置。在筛选面板中，选择平均 >50000。顶层节点在该筛选过程中消失，这是因为挣那么多的钱并不是常见的事。具有学士学位并且年龄超过 27 岁的人落入该范围以内。沿着第一个分割的左分支一直到末端，我们在该范围内发现了另一组人群：年龄超过 36 岁的已婚男性，每周工作超过 35 小时，并且受到过良好的教育 （10 年或更多）。

如果我们再次使用筛选工具，并且寻找绝对偏差大于 $25,000 的节点，我们可以发现经济条件变化范围最大的人群。在该筛选过程中第一个被保留的节点是那些年龄超过 27 岁并且具有学士学位的人群。该节点上的直方图显示了以均值为中心的分布，不过年收入达到 $100,000 左右的人群具有不寻常的数目。

蝴蝶花

该数据集中的每个记录描述了蝴蝶花的五个特征：萼片宽度、萼片长度、花瓣宽度、花瓣长度以及花类型。该回归过程的目的是根据其它特征来预测萼片宽度。文件 *iris-rt.treeviz* 显示了为预测萼片宽度而在该数据集上运行 “回归树导入工具”的结果。

观察顶部节点，我们可以看到在萼片宽度值中存在一个间隙，其中没有花存在。“回归树导入工具”首先使用萼片长度变量在该数据集上产生分割。萼片长度小于 2.6 时，可能的萼片宽度子集很有限。另一方面，萼片长度值大于 2.6 指出了一个更大的萼片长度的均匀分布。那些萼片长度小于 2.6 的蝴蝶花的平均萼片宽度为 0.24，而对应那些萼片长度大于 2.6 的蝴蝶花其萼片平均宽度为 1.68。对于具有大萼片长度的蝴蝶花，我们可以看到在萼片长度上树再次产生分割，这一次阈值为 4.85。在相同变量上两个连续分割说明在这两个变量中可能存在着某种受限制的功能性关系。

回到萼片宽度小于 2.6 的那些蝴蝶花，我们可以看到下面的分割是在花瓣宽度属性上进行的。有意思地是，这部分树中的值好象被分隔开了，那些花瓣宽度小于 3.25 的蝴蝶花呈现出的值处于三个狭窄而且被分开的范围内。

糖尿病诊断

该数据集与利用统计学对糖尿病进行诊断的问题有关，它是从亚利桑那州，凤凰城的美州部落中收集而来的。文件 *pima-rt.treeviz* 显示了在该数据集上运行 “回归树导入工具” 的结果，并将血糖水平作为待预测的连续变量。

该树的第一个分割出现在糖尿病指标上，显示出患有糖尿病的人其血糖水平比正常人高（141 比 110）。下一个分割是按 2 小时血清胰岛素属性进行的，当其值高于 125 时将导致高血糖。

“回归树” 从顶部节点开始，遵循每个节点处的决策进行预测，对新记录进行检查。例如，一个糖尿病病人其 2 小时血清胰岛素为 110，可以预测得其血糖水平为 105。

散点可视化工具样例文件

这里提供的样例数据文件和配置文件展示了 “散点可视化工具” 的特征和功能。下面的 *.data* 和 *.scatterviz* 文件位于 *examples* 目录中。要将数据文件装入 MineSet，请打开 *.schema* 文件。

- Windows 用户可以在安装 MineSet 目录的 *\examples* 子目录下找到这些文件。
- IRIX 用户可以在 */usr/lib/MineSet/examples/scatterviz* 中找到这些文件。

“散点可视化工具”样例文件如下所示：

- *company.data*
  该文件包含了几个保险公司对三个险种的假想销售数据：人寿保险、汽车保险和家庭保险。该数据跨度为 10 年 （增量为一年），包括了五个收入范围 （用户的年收入）。

- *company.scatterviz*
  该文件指定年份到一个滑动条维，收入范围到另一个滑动条。人寿保险、汽车保险和家庭保险的销售额成为 “散点可视化工具”场景中的三个维。滑动条汇总窗口中的颜色浓度代表了所有公司在全部保险业务的总销售值。

- *company-total.scatterviz*
  该文件包含了与 *company.scatterviz* 中所之指定的内容相同的内容，所不同的是由公司在所有保险业务范围内的销售总额确定每个公司的规模。

- *ompany-life.scatterviz*
  该文件包含了与 *company.scatterviz* 中所指定的内容相同的内容，所不同的是每个对象的颜色指示了作为整个销售额一部分的人寿保险销售额。

- *store-type.data* 和 *store-type.scatterviz*
  这两个文件显示了不同类型的商店在不同产品上的三年期销售额。由滑动条所代表的独立变量是时间。每个实体代表一种商店类型 （例如，食品商店、药店、服务站等等）。对于每个商店类型，数据文件包含了几个产品组的销售总额，例如，酒精饮料、麦片等等。数据的时间跨度为 36 个月，按月增长。

  该配置文件用月作为滑动条维。一个坐标轴是酒精饮料的销售额，另一个为烟草制品的销售额。未使用第三个坐标轴。

- *brand.data* 和 *brand.scatterviz*
  这些文件显示了不同类型商店中几种软饮料品牌的销售额。在该数据集中，实体的品牌和商店的类型映射到坐标轴。销售总额被映射为每个品牌实体的大小。颜色映射是随机的。由于没有独立的变量，也就没有用滑动条。

- *cars.data* 和 *cars.scatterviz*
  这些文件显示了几个轿车型号的重量、马力、型号年代以及提速能力。坐标轴为立方英寸， mpg 以及提速到 60 迈的时间。重量映射为实体大小。

- *people.data* 和 *people.scatterviz*
  这些文件显示了一个假想的人口样例中的身高、体重、人口密度以及胆固醇水平。

- *nl.births.data* 和 *nl.births.scatterviz*
  这些文件显示了荷兰的人口出生模式。对于每个区域，显示了人口密度、出生率以及人口总数。动画滑动条映射为母亲的年龄和年份。

- *adult94.data* 和 *adult94.scatterviz*
  这些文件显示了使用散点可视化工具应用于 *adult.data* 的复杂样例。可视化过程中的三个坐标轴为 avg_hrswk （也就是每周工作的小时数）、 avg_gross_income 和 avg_education_num。不幸地是 “教育数”与受教育的年数并不完全对应，但很接近。右边的滑动条可在不同的年龄区间内进行动画过程。通过按照职业、种族和性别的分组创建每个组合结果。这就意味着对于这三个属性存在一个对应每个取值组合的实体。颜色正如图例所示，代表了不同的职业。每个实体的大小对应于记录计数。汇总滑动条同样按照数据密度进行上色。要想知道该可视化过程是如何创建的，您可以从 “文件”菜单中选择 “启动工具管理器”。这将打开用于创建实例的 “工具管理器”。

  初始时，画面显示了年龄小于 20 岁的人群的有关信息。注意，平均工作小时数 （约为 14）和平均收入 （约为 $4000 ）比较低。如果您利用滑动条在年龄属性上进行动画过程，并且从三个垂直的视角观察画面 （使用主窗口右部的三个按钮），您将注意到有不同的趋势出现。例如，如果您定向画面只观察每周小时数与收入之间关系，您可以发现人们工作的时间随着年龄的增长而延长，一直到 25 岁，然后每周工作的时间很少超过 49 小时，直到退休。然而，收入将一直增长，直到 50 岁，然后处于平稳状态，然后再次降低。实际的趋势多少与职业的选择和其它的因素有关。

假设您对工艺维修和特殊行业两种职业间的比较感兴趣。可以打开 “筛选”面板（查看 > 显示筛选面板）并且从职业列表中选择 “工艺维修”和 “特殊行业”。现在，当您进行动画时，您可以看到从事 “特殊行业”的人在实际中往往开始于低收入，但是随着年龄的增长很快赶上从事 “工艺维修”的人。“特殊行业”在教育坐标轴上比 “手艺维修”高很多。您也许希望通过只显示女性或者特定的种族来进一步限制筛选过程。或者在动画过程中尝试选择一些不同运动轨迹。

- *census.data* 和 *census.scatterviz*

  这些文件也显示了组合统计数据的图形。原始数据集包含了大约 150,000 行。在组合之后，对于教育、性别、产业以及职业 （它们作为分组依据列）的每个组合会出现一个立方体 （一个组合结果）。

平伸可视化工具样例文件

这里提供的配置文件和数据文件展示了 “平伸可视化工具”的特征和功能。这些文件在 *examples* 目录中。

Windows 用户可以在安装 MineSet 目录的 *\examples* 子目录下找到这些文件。

IRIX 用户可以在 */usr/lib/MineSet/examples/splatviz* 中找到这些文件。

- *蘑菇*
  mushroom.data 文件包含了超过 5,000 只蘑菇的预组合数据。分组依据列为：odor、gill_color 和 cap_color。对于原始数据中这三个列的每个组合，都有一个计数和一个平均可食性指标，当其为 0 时表示可以食用，为 1 时则有毒。因为蘑菇不可能是部分有毒，因此在 0 到 1 之间的平均可食性指标意味着在组合结果中，有些蘑菇可以食用，而另一些有毒。

  该可视化过程将每个列的特定值都根据平均可食性沿着坐标轴进行排序。很明显，气味是可食性最好的判定因素。可以注意到，大多数平伸要么为 0，要么为 1，这就意味着这三个列在对两类蘑菇进行划分时都有用处。实际上，列重要性功能选择了映射为坐标轴的列。向下移动透明度滑动条可以确定哪个平伸拥有最高的计数。最透明的平伸代表了 288 只具有相同 odor、 gill_color 以及 cap_color 值的蘑菇。要进行确认，可以用 sum_count_poison>280 进行筛选并选中剩下的平伸以观察它们的计数。注意，所有 gill_color=buff 的蘑菇都是有毒的。

- *adultJobs*
  adultJobs.data 文件从 *adult94* 中派生出来，*adult94* 是一个发行版中提供的数据集。它使用依据教育、职业、每周工作小时数 （分组的）以及年龄 （分组的）分组的组合结果创建而成。 gross_income 列按照计数和平均进行组合。要使用 “平伸可视化工具”进行显示，将 age_bin 映射为滑动条，将其它分组列映射为坐标轴。count_gross_income 列映射为透明度，而 avg_gross_income 映射为颜色。

  当滑动条处在最左边的位置时，图的颜色几乎完全是蓝色的。这就意味着不论职业，教育程度或工作的小时数，大多数小于 20 岁的年轻人收入较低。向右移动滑动条，可以注意到收入随更高的教育程度和不同的职业是怎样朝着坐标轴的末端快速升高的。通过透明度的变化，您可以看到最为普遍的教育程度类型是 HS，即大专或大本学位。

  移动 “汇总”滑动条可以显示收入的分布是如何随年龄坐标轴列变化的。

- *adultJobs2*
  adultJobs2 也是在 *adult94* 数据集的基础上得到的。这里，坐标轴表示工作类别、教育程度以及职业。两个映射为滑动条的列为年龄 （已分组）和每周工作小时数 （已分组）。收入再次被进行计数和平均组合并分别映射到颜色和透明度。因为在二维滑动条上有更多的位置，因此几乎没有由每个位置代表的记录。这就会造成颜色和透明度的更大变化。“汇总”滑动条 hrs_per_week 维中部的红色区域指出几乎所有的人每周工作的小时数在 35 和 45 之间。注意，一些职业与特定的工作类别有固定的联系。例如，军队中的每个人的工作类别都为 Fed-Government。

- *censusIncome*
  这个例子建立在一个与 adult94 相似的数据集之上，不过由于其规模较小，这个数据集没有包括数据的分布信息。要试图理解毛收入 （gross income）和总收入 （total income）之间的差别，必须将 gross_income、 total_income 和 hrs_per_week 映射为坐标轴。颜色显示了年龄。通过研究图象，我们可以知道有很多记录 total_income=gross_income，但是也有很大一部分记录具有较高的 total_income，而 gross_income 为 0。奇怪的是，在很多情况下 gross_income 大于 total_income。

注意不同年龄的人群所集中的地方。许多老人（黄色）在 hrs_per_wk=0 的平面内。他们应该是退休人员。许多子节点和年轻的成年人（蓝色）在 gross_income=total_income=0 的线上。注意，非常透明的平伸位于该范围的外边缘附近。这些位置包括了所有落入坐标轴的最大组中的点。例如， total_income 最高的组为 70,300+。高于 70,300 的任何点都进入了该组。

要想更好地观察变化的密度，请调节透明度滑动条。在低透明度等级中，对角线表示出大多数人 gross_income=total_income，或他们只有 total_income 而没有 gross_income。当您提高等级时，您可以看到几乎所有的范围都包含点。该数据集包含了 150,000 条记录。

- *客户波动*
  客户波动是指一个用户离开一个公司转向另一个公司。该样例显示了电话公司用户的波动。用于产生该样例的数据是 *churn.schema*。

  使用列重要性，我们可以发现 total_day_charge、 number_customer_service_calls 和 international_plan 是重要的区别因素。这些列映射为坐标轴。然后我们创建一个新数值型列， churn，其等价于 churned==Yes，将之映射为颜色。

  在可视化结果中，红色区域指示了较高的波动可能。该区域表示打过三个以上服务电话但 total_day_charge 很低的用户对应于高 churn 值。您也许想对高消费用户进行加权，施以比其它用户大得多的权重。要达到这一目标，创建一个新列 total_charge，它等价于

  ```
  `total_day_charge`+`total_eve_charge`+`total_night_charge`
  ```

  或该总和的幂。然后将 total_charge 列映射为透明度。这就意味着每个记录按照 total_charge 进行了加权。现在，可视化过程显示了 total_day_charge 坐标轴高值端附近的令人感兴趣的附加区域。

树可视化工具样例文件

这里提供的配置文件和数据文件展示了 “树可视化工具”的特征和功能。这些文件在 *examples* 目录中。

- Windows 用户可以在安装 MineSet 目录的 *\examples* 子目录下找到这些文件。
- IRIX 用户可以在 */usr/lib/MineSet/examples/treeviz* 和 */usr/lib/MineSet/data* 中找到这些文件。

“树可视化工具”样例文件如下所示：

- *store.data* 和 *store.treeviz*
  当进行图形显示时，这些文件显示了有关连锁店的假想销售数据。树中的等级包括所有商店、区域、州、城市以及个别商店。对于等级中的每一层显示了四个产品。在该配置中，高度代表销售额；颜色代表完成目标的百分数。

- *stateRevenue.data* 和 *stateRevenue.treeviz*
  当进行可视化显示时，这些文件显示了 1992 年度每个州的财政收入成分，该数据可从美国统计局中得到。（可访问 http://www.census.gov/govs/state/stfin92.dat）。高度代表了税收额。背景中下降的节点显示了各种税收对显示在根节点中的整体收入的贡献。

- *beer.data* 和 *beer2.data*、*beer.treeviz* 和 *beer2.treeviz*
  当进行图形显示时，这些文件显示了建立在啤酒消费者调查基础上的假想数据。树包含三个层：

  1. 第一层是关于种类的 （例如，啤酒或淡啤酒）。
  2. 第二层是品牌编码 （随机分配）。
  3. 第三层是个别产品编码；例如：12 包装与 6 包装 （随机分配）。

  每个图包含了七个条形，代表了七个年龄组。条形高度代表了该年龄组的消费额。颜色代表被男性和女性消费的百分比。品牌、产品以及在这些文件中使用的数据只是样例。

  beer.treeviz 和 *beer2.treeviz* 产生了相同的图形结果，但是它们按照不同的方式构建而成。在 *beer.treeviz* 中，每种类型的啤酒由一个记录代表，每个记录包括男性和女性的消费量；这些值存贮在一个枚举数组中。

  在 *beer2.treeviz* 中，每种啤酒有七条记录，每个记录代表一个年龄组。注意，在 *beer* 文件中，年龄组包含在配置文件中；在 *beer2* 文件中，它们包含在数据文件中。

  beer 文件比 *beer2* 文件所需的存贮空间少；但是配置文件略微复杂一些。在一些情况下，以 *beer2* 文件所使用的格式产生数据可能更容易一些。

索引

C

D

E

F

G

H

I

J

K

L

M

N

O

P

Q

R

S

X

Y

Z