MineSet™ 3.0 Enterprise Edition Reference Guide

ドキュメント番号 007-3558-001JPN

編集協力者

執筆：Sandra Motroni and Helen Vanderberg
イラスト：Dany Galgiani
制作：Linda Rae Sande
技術協力：Barry Becker、Amit Bleiweiss、Jeff Brainerd、Cliff Brunk、Eben Haber、Ara Jerahian、Andy Kar、Ed Karrels、Eser Kandogan、Alex Kozlov、Alan Norton、Peter Rathmann、Mario Schkolnick、Dan Sommerfield、Peter Welch、Brett Zane-Ulman

MineSet™ 3.0 Enterprise Edition Reference Guide
ドキュメント番号 007-3558-001JPN

目次

図目次　xv

表目次　xvii

はじめに　xix
対象読者　xix
このマニュアルの内容　xix
このマニュアルの構成　xx
マニュアル内の図表について　xx
表記上の決まり　xx

1. MineSet **の概念と機能説明　1**
項目の追加　1
「項目の追加」ボタン　1
集計処理　4
配列　6
分割型項目　7
アニメーション　10
アニメーション・コントロール・パネル　11
独立次元を制御するスライダ　11
アニメーション・サマリウィンドウ　13
アニメーション用のボタンとスライダ　14
アニメーション制御ボタン　14
アニメーション用のスライダ　15
データポイントと補間　16
動作試験　16

モデルの適用　17
「モデルの適用」パネル　18
「モデルのテスト」パネル　19
「モデルへのデータ適合」パネル　20
モデルの適用　21
相関規則　23
「相関規則分析」タブ　26
必要なファイル　26
相関規則の可視化　27
相関規則の設定　27
相関規則のオプション　28
相関規則の対応付け用のボタン　30
スキャタ・ビジュアライザによる相関規則の分析　31
相関規則のサンプルファイル　32
.ruleviz から .scatterviz へのファイル変換　33
階級自動生成　36
バックフィッティング　37
階級生成　39
階級生成のオプション　39
ブースティング　42
項目の型または名前の変更　42
選択ポイント　43
クラシファイア　43
クラシファイアの名前　44
「クラシファイア」タブ　45
クラスタリング　45
単一 k-means クラスタリング法　46
反復型 k-means クラスタリング法　47
クラスタリングのオプション　49
属性の重み　49
「クラスタオプション」ダイアログ・ボックス　50

クラスタ・ビジュアライザ　52
　必要なファイル　52
　クラスタ・ビジュアライザの起動　52
色の選択　53
　カラーブラウザによる色の選択（Windows システム）　53
　カラーブラウザによる色の選択（IRIX システム）　55
「重要項目」タブ　58
重要項目　58
　重要項目の検出　58
　重要項目と他の分析との相違　61
項目　62
コマンド行の操作　63
設定ファイル　63
混同マトリックス　64
生成コストの考慮　65
相互検証　66
データ・クリーニング　66
「データの可視化 / マイニング」パネル　67
「データファイル」タブ　67
データのインポート　68
「データ変換」パネル　68
デシジョン・テーブル　69
　デシジョン・テーブルモデルの生成　70
　デシジョン・テーブル・ビジュアライザの起動　70
　離散型ラベル　71
　軸に対する項目の対応付けによるデータの解析　71
　デシジョン・テーブルの解釈　72
　デシジョン・テーブルのオプション　74
　プルダウン・メニュー　75
　「表示」メニュー　76
　「属性値の順序付け」メニュー　77

決定木　77
　決定木の作成　78
　IRIX システム上での並列化処理　79
　詳細分析オプション　79
　決定木分析のオプション　79
　検索パネルとフィルタパネル　83
「離散型ラベル」メニュー　84
ドリルスルー　85
ドリルダウンとドリルアップ　86
誤差推定　87
エビデンス・モデル　91
エビデンス分析　91
　エビデンス・クラシファイアの作成　92
　エビデンス・ビジュアライザの起動　94
　エビデンス分析のオプション　95
　エビデンス・ビジュアライザのメニュー　98
「ファイル」メニュー　100
　Windows システム　100
　IRIX システム　101
必要なファイル　102
「フィルタ」ボタン　103
「フィルタ」パネル　103
増加比率　105
IRIX システムの「ヘルプ」メニュー　105
Windows システムの「ヘルプ」メニュー　106
ヒストグラム・ビジュアライザ　106
履歴の表示　106
予備法　107
分析　107
Tool Manager で分析を実行するときのモード　109
　分析 における誤差の取扱い　109
　分析の詳細オプション　110
分析のステータス・ウィンドウ　112

国際化　114
IRIX システム上でのロケールの設定　114
他の言語とエンコーディングに対する拡張機能（IRIX 専用）　114
反復型 k-means　117
ラプラス補正　117
学習曲線　118
改善曲線　120
損失マトリックス　121
マップ・ビジュアライザ　125
マップ・ビジュアライザに必要なファイル　127
マップ・ビジュアライザの起動　128
Tool Manager によるマップ・ビジュアライザ の設定　129
.gfx ファイルと .hierarchy ファイルの作成　129
スライダとアニメーションの作成　131
マップ・ビジュアライザのオプション　131
マップ・ビジュアライザの設定情報が保存されるファイル　134
「マイニングツール」タブ　134
複数のオブジェクトの選択　135
相互情報量　135
Naive-Bayes　135
ツリー・ビジュアライザ以外のビジュアライザで使用できるナビゲーション・コントロール　136
ツリー・ビジュアライザで使用できるナビゲーション・コントロール　139
「属性値の順序付け」メニュー　140
正規化相互情報量　141
NULL 値　141
選択式決定木　143
選択式決定木モデルの作成　144
必要なファイル　144
選択式決定木クラシファイアの作成　144
IRIX システム上での並列化処理　144
選択式決定木分析のオプション　145
IRIX システム上での並列化処理　146

予測値　148
普及率　148
枝刈り　148
ランダムシード　149
レコードビューワ　149
　レコードビューワの起動　149
　行番号の振り直し　150
　レコードビューワでの検索　150
　データの保存　150
レコードの重み付け　151
「回帰」タブ　151
回帰ツリー　152
　回帰ツリーの作成　152
　「連続型のラベル」メニュー　153
　回帰ツリーのオプション　153
　回帰モデルにおける誤差推定　156
回帰モデル名　156
「項目の削除 」ボタン　157
投資利益率 (ROI) 曲線　157
ファイルの保存　159
サンプルファイルのディレクトリ　159
スキャタ・ビジュアライザ　159
　必要なファイル　160
　スキャタ・ビジュアライザの起動　161
　スキャタ・ビジュアライザの設定　161
　スキャタ・ビジュアライザ用のスライダの作成　162
　スキャタ・ビジュアライザのオプション　162
　アニメーション・コントロール・パネル　166
　スキャタ・ビジュアライザにおける NULL 値の取扱い　167
　設定ファイルとデータファイルのサンプルファイル　167
「選択」メニュー　167
マップ・ビジュアライザ、スキャタ・ビジュアライザ、スプラット・ビジュアライザのスライダの作成　169

項目名のソート　169
スプラット・ビジュアライザ　169
　スプラット・ビジュアライザの不透明度　171
　必要なファイル　173
　スプラット・ビジュアライザの起動　174
　「スプラット」オプション　175
　設定情報が保存されるファイル　177
　スプラット・ビジュアライザにおける NULL 値の取扱い　177
　スプラット・ビジュアライザ用のスライダの作成　178
　アニメーション・コントロール・パネル　178
　スプラット・ビジュアライザのプルダウン・メニュー　181
　「形状」メニュー　181
　設定ファイルとデータファイルのサンプルファイル　182
分割の下限値　182
分割の基準　183
統計量ビジュアライザ　184
　統計量ビジュアライザの使用方法　184
　統計量ビジュアライザのプルダウン・メニュー　186
　統計量ビジュアライザの「表示」メニュー　186
「テーブルの履歴 」ボタン　187
　「現在の履歴表示は」フィールド　187
　「前処理 :」フィールドと「次処理 :」フィールド　188
Tool Manager　191
Tool Manager のオプション　192
訓練事例　192
ツリー・ビジュアライザ　193
　必要なファイル　194
　ツリー・ビジュアライザの起動　194
　ツリー・ビジュアライザのオプション　195
　設定情報が保存されるファイル　203

ツリー・ビジュアライザのプルダウン・メニュー　203
「表示」メニュー　204
「選択」メニュー　210
ツリー・ビジュアライザの表示メニュー　211
ツリー・ビジュアライザの移動メニュー　211
「ヘルプ」メニュー　213
ツリー・ビジュアライザにおける NULL 値の取扱い　213
ツリー・ビジュアライザに関する制限事項　215
設定ファイルとデータファイルのサンプルファイル　215
端数の切捨て　215
均一な範囲　216
均一な重み付け　216
「表示」メニュー　216
可視化ツール　217
警告オプション　219
Web 公開　220
重み付け　220
西暦 2000 年問題への対応　220

A. **設定ファイルとデータファイルのサンプルファイル　223**
相関規則ビジュアライザ用のサンプルファイル　225
クラスタリング用のサンプルファイル　226
重要項目用のサンプルファイル　226

決定木分析用のサンプルファイル　228
解約 (Churn)　228
車の原産国　229
性別の判断　230
給料を決める要因　232
アヤメ (iris) の分析モデル　233
キノコ (Mushroom) の分析モデル　234
政党への帰属　235
乳癌の診断　235
甲状腺機能低下症 (Hypothyroid) の診断　236
ピマ族における糖尿病の診断　237
DNA 境界　237
デシジョン・テーブル用のサンプルファイル　238
解約 (Churn)　239
車の原産国　241
性別の判断　242
給料を決める要因　243
アヤメの分析モデル　248
キノコの分析モデル　249
政党への帰属　250
乳癌の診断　251
甲状腺機能低下症 (Hypothyroid) の診断　252
ピマ族における糖尿病の診断　253
DNA 境界　254

エビデンス・ビジュアライザ用のサンプルファイル　255
解約 (Churn)　255
車の原産国　256
性別の判断　257
給料を決める要因　259
アヤメの分析モデル　261
キノコの分析モデル　261
政党への帰属　263
乳癌の診断　263
甲状腺機能低下症 (Hypothyroid) の診断　264
ピマ族における糖尿病の診断　265
DNA 境界　265
マップ・ビジュアライザ用のサンプルファイル　266
選択式決定木用のサンプルファイル　269
解約 (Churn)　270
車の原産国　270
アヤメの分析モデル　270
キノコの分析モデル　271
政党への帰属　271
乳癌の診断　272
甲状腺機能低下症 (Hypothyroid) の診断　272
DNA 境界　272
回帰ツリー分析用のサンプルファイル　272
解約（Churn）　273
車の燃費効率　273
給料を決める要因　274
アヤメ (Iris) の属性値　276
ピマ族における糖尿病の診断　276
スキャタ・ビジュアライザ用のサンプルファイル　277
スプラット・ビジュアライザ用のサンプルファイル　280
ツリー・ビジュアライザ用のサンプルファイル　282

図目次

図 1-1 「項目の追加」ダイアログ・ボックス　2
図 1-2 「集計処理」ダイアログ・ボックス　5
図 1-3 サマリウィンドウと 2 つのスライダを持つアニメーション・コントロール・パネル　12
図 1-4 「モデルの適用」ダイアログ・ボックス：クラシファイヤの選択　18
図 1-5 「モデルの適用」パネル　19
図 1-6 アヤメ (Iris) データセットの誤った判別モデル（例 1）　22
図 1-7 アヤメ (Iris) データセットの誤った判別モデル（例 2）　23
図 1-8 Tool Manager の「相関規則分析」タブにおける多重規則の設定例　29
図 1-9 「相関規則の対応付け」パネル　30
図 1-10 複数のカラーボックス　53
図 1-11 カラーブラウザダイアログ・ボックス（Windows システム）　54
図 1-12 カラーブラウザの HSB タブ（Windows システム）　55
図 1-13 カラーブラウザの RGB タブ（Windows システム）　55
図 1-14 複数のカラーボックス　56
図 1-15 カラーブラウザ（IRIX システム）　57
図 1-16 アヤメ (Iris) データセットに関する混同マトリックス　64
図 1-17 予備法によるクラシファイヤの誤差推定　88
図 1-18 相互検証のクラシファイヤ (k=3)　90
図 1-19 分析を適用するときのツール実行環境　108
図 1-20 学習曲線　119
図 1-21 改善曲線　121
図 1-22 キノコ (mushroom) データセットに関する混同マトリックス（デフォルト設定）　122
図 1-23 キノコ (Mushroom) データセットに関する混同マトリックス（損失マトリックスを使用）　123
図 1-24 キノコ (Mushroom) データセットに関する混同マトリックス（NULL 値を許可する損失マトリックスを使用）　124

図 1-25 1990 年の米国の人口を表示したマップ・ビジュアライザの例 126
図 1-26 NULL 値を高さに割当てた場合（中央上のオブジェクト）と色に割当てた場合（右下のオブジェクト） 143
図 1-27 投資利益率 (ROI) 曲線 158
図 1-28 *u* の値が大きい場合と小さい場合の不透明度関数 171
図 1-29 *u* = 5.3 と u = 30 の場合のグラフ 172
図 1-30 統計量ビジュアライザによる数値型項目の表示 185
図 1-31 統計量ビジュアライザによる離散型項目の表示 186
図 1-32 「テーブルの履歴」ボタン 187
図 1-33 「履歴の表示」ダイアログ・ボックス (Windows) 189
図 1-34 「履歴の表示 (View History」ダイアログ・ボックス (IRIX) 190
図 1-35 訓練事例内のサンプルレコード 193
図 1-36 ツリー・ビジュアライザの「設定オプション」ダイアログ・ボックス (Windows システム) 196
図 1-37 ツリー・ビジュアライザの「設定オプション (Configuration Options)」ダイアログ・ボックス （IRIX システム） 197
図 1-38 ツリー・ビジュアライザの「検索」ダイアログ・ボックス (Windows) 204
図 1-39 ツリー・ビジュアライザの「検索 (Search)」ダイアログ・ボックス (IRIX) 206
図 1-40 ツリー・ビジュアライザでの検索結果の例 208
図 1-41 高さ、色、ディスク、ラベルに NULL 値を割当てた場合の表示例 214
図 A-1 「解約 (Churn)」データセットに対するドリルダウン 240
図 A-2 デシジョン・テーブル・ビジュアライザによる「成人 (adult)」データセットの解析（IRXI システム) 245
図 A-3 「成人 (adult)」データセットの詳しい解析 246

表目次

表 1-1 州別・年齢別に集計した消費パターン 7

表 1-2 階級生成された年齢をインデックスとする配列 7

表 1-3 階級生成された年齢を基準として分割された項目 8

表 1-4 年齢階級別・給与階級別に集計した消費額 8

表 1-5 年齢階級をインデックスとして消費額を1次元配列にしたテーブル 9

表 1-6 年齢階級と給与階級をインデックスとして消費額を 2 次元配列にしたテーブル 9

表 1-7 年齢階級をインデックスとする配列（消費額）を給与階級に基づいて分割したテーブル 10

表 1-8 相関規則の対応付け 31

表 1-9 階級自動生成のオプション 36

表 1-10 デフォルトのファイル拡張子 102

表 1-11 韓国語のリソースファイルの編集例 116

表 1-12 ツリー・ビジュアライザ以外のビジュアライザのナビゲーション・ボタン 136

表 1-13 ツリー・ビジュアライザ以外のビジュアライザの調整スライダとダイヤル 137

表 1-14 ツリー・ビジュアライザ以外のビジュアライザ内のグラフに対する操作 138

表 1-15 ツリー・ビジュアライザのナビゲーション・ボタン 139

表 1-16 ツリー・ビジュアライザの調整スライダとダイヤル 140

表 1-17 systune パラメータ 147

表 1-18 40-50 歳代のテーブル 179

表 1-19 50-60 歳代のテーブル 179

表 1-20 40-50 歳代のテーブルと 50-60 歳代のテーブルの補間結果 180

はじめに

このマニュアルでは、データマイニング・ツールと可視化ツールを組合せたソフトウェア・パッケージである MineSet の技術的な特徴と高度な機能を説明します。MineSet の最新情報については、World Wide Web の http://www.sgi.com/software/mineset.（日本語版 http://mineset.sgi.co.jp/）でも参照することができます。

対象読者

このマニュアルでは Tool Manager による MineSet の操作方法に慣れたユーザを対象として、MineSet の機能や仕組みを分かりやすく解説しています。Windows 版 MineSet の操作手順やパス名は、Windows 環境で一般的に使用されるものです。IRIX 版 MineSet を使用する場合は、UNIX コマンドに関するある程度の知識が必要になります。

このマニュアルの内容

このマニュアルでは主に、コマンド行または設定ファイルを通じて MineSet を操作する方法について説明しています。また、必要に応じてプログラミング・インタフェースについても説明しています。各章の要約は「このマニュアルの構成」に記載しています。

MineSet ツールの使用方法については、『*MineSet 3.0 Enterprise Edition User's Guide for Windows*』を参照してください。

このマニュアルの構成

第 1 章 「MineSet の概念と機能説明」
MineSet の概念や用語、MineSet の各種機能や可視化ツール（ビジュアライザ）についてトピックス別に詳しく説明しています。

付録 A 「設定ファイルとデータファイルのサンプルファイル」
MineSet パッケージにサンプルとして付属している設定ファイルとデータファイル（ビジュアライザの実行時に必要となるファイル）の内容と使用方法について説明しています。

マニュアル内の図表について

このマニュアルに記載している図表は、主に MineSet 3.0 for Windows のものです。IRIX 版と Windows 版の図表が大きく異なる場合は、両方の図表を掲載しています。

表記上の決まり

このマニュアルでは、次の表記法と記号を使用しています。

『』	ほかのマニュアルのタイトルを表します。
「」	本書のほかの章や節のタイトルを表します。また、メニュー名やボタン名などの UI（User Interface）を表します。
->	プルダウン・メニューの階層構造を表します。
< >	キーボードのジェネリック・キー（Ctrl、Shift、Alt など）を表します。キーの操作方法として、次に例を示します。
< Enter >	< Enter > キーを押します。
< Alt >-h	< Alt > キーを押しながら h キーを押します。
< Alt >-h c	< Alt > キーを押しながら h キーを押した後、すぐに c キーのみを押します。

< Shift >-< Ctrl >-n
: < Shift> キーを押しながら < Ctrl> キーと n キーを同時に押します。

< Ctrl >-x< Ctrl >-c
: < Ctrl > キーを押しながら x キーを押した後、すぐに < Ctrl > キーを押しながら c キーを押します。

イタリック
: コマンド名、ファイル名、プログラム変数、およびメニュー名やボタン名などの UI（User Interface）を表します。

クーリエ書体
: システム出力結果例や、ファイルの内容を表します。

クーリエ書体のボールド体
: ユーザが文字通り入力するコマンドや、他のテキストを表します。

第 1 章

MineSet の概念と機能説明

項目の追加

データセットのデータ項目を追加、削除、分類（ソート）した上でデータをクラシファイア (Classifier) やビジュアライザに送る場合は、Tool Manager の「データ変換」パネルを使用します。項目の追加処理については、下記の「項目の追加 ボタン」を参照してください。項目の削除については「項目の削除」ボタン」、項目名のソートについては「項目名のソート 」をそれぞれ参照してください。

「項目の追加」ボタン

新しい項目を作成するには、「 項目の追加 」ボタンを使用します。新しい項目の値は、数式に基づいて計算されます。たとえば、2 つの既存の項目値の比率を表す新しい項目を追加するときに、このボタンを使用します。このボタンをクリックすると、ダイアログ・ボックスが開き、新しい項目の名前と表現（式）を指定することができます (図 1-1)。

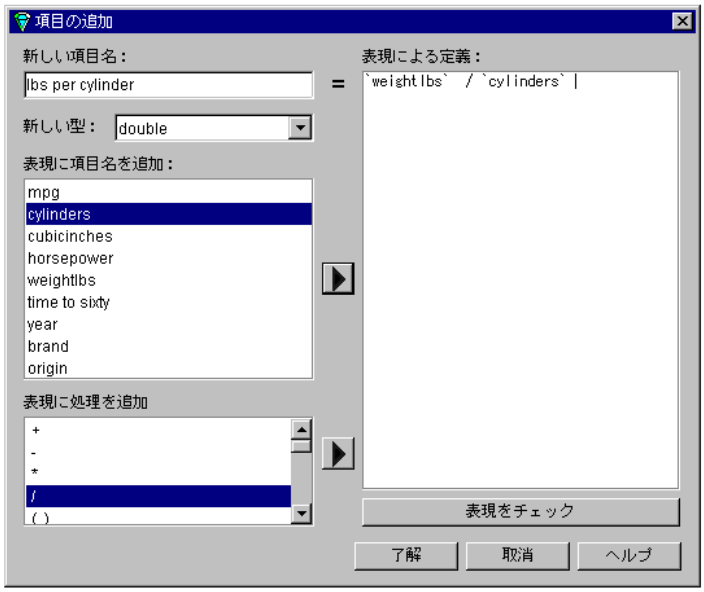

図 1-1 「項目の追加」ダイアログ・ボックス

このダイアログ・ボックスの左側には新しい項目の名前を入力するフィールドがあり、ポップアップ・メニューを使用して項目のデータ型（整数、文字列、浮動小数点など）を指定できるようになっています。

ダイアログ・ボックスの右側にある大きいテキスト入力領域には、項目値の表現（式）の定義を入力します。項目名と演算子が入力しやすいように、ダイアログ・ボックスの左下にあるスクロールリストに、現在選択されているテーブルのすべての項目と、使用可能なすべての演算子が表示されます。項目名または演算子を表現に挿入するには、スクロールしたリストに表示される該当の項目をダブルクリックするか、その項目をクリックしてからリストの右側にある矢印ボタンをクリックします。

「項目の追加」ダイアログ・ボックスと「フィルタ」パネルで使用される式の構文は C、C++、Java で使用される式と似ています。下記のように、基本的な演算子は同じものが使用されます。

+	加算
-	減算
*	乗算
/	除算
( )	式をグループ化する括弧
%	剰余（除算した後の余り）
!	論理否定 (NOT)
~	論理否定 (NOT)
&&	論理積 (AND)
||	論理和 (OR)
^	排他的論理和 (exclusive OR)
==	等しい
!=	等しくない
<=	より小さいか等しい
<	より小さい
>=	より大きいか等しい
>	より大きい
&	ビット積 (bitwise AND)
|	ビット和 (bitwise OR)

上記の演算子のほかに、次の関数や構文が用意されています。

isNull( )	括弧内の値が Null かどうかを判定します。
if ( ) then ( ) else ( )	最初（if の後）の括弧内の値が true である場合は 2 番目（then の後）の括弧内の値を返し、それ以外の場合は 3 番目（else の後）の括弧内の値を返します。
(x) ? (y) : (z)	C 言語の if/then/else 構文と同様に、(x) が true である場合は (y) を返し、(x) が false である場合は (z) を返します。

divide(x, y, z)	y がゼロでない場合は x を y で割り、y がゼロである場合は z を返します。この関数は、y==0 ? z : x/y という C の構文と同じです。
strlen(x)	文字列 x の長さ（半角文字数）を返します。
substring(x, y z)	文字列 x の部分文字列（y の位置から始まる長さ z の文字列）を返します。C、C++、Java の構文規則と同様に、文字列を構成する最初の文字のインデックスは（1 ではなく）0 です。

作成した表現（式）の構文をチェックするには、「表現をチェック」ボタンをクリックします。エラーが検出されると、ダイアログ・ボックスにエラーの種類と発生箇所が表示されます。「了解」ボタンをクリックすると、式が自動的にチェックされ、エラーが訂正されるまでダイアログ・ボックスは閉じません。

「項目の追加」ダイアログ・ボックスではデータ型の整合性がチェックされます。たとえば、文字列型の項目に数値式（または、数値型の項目に文字列）を代入しようとすると、警告メッセージが表示され、新しい項目のデータ型が自動的に訂正されます。

式の中でユーザが定義した関数を使用することもできます。詳細については、『*MineSet 3.0 Enterprise Edition Interface Guide*』の「Plug-in Functions」を参照してください。

集計処理

Tool Manager の「集計処理」ボタンを使用するには、配列と分布の基本概念を理解しておく必要があります。集計処理機能に関連する基本概念の説明については、「配列（Array）」を参照してください。

「集計処理」ボタンを使用すると、単純な集計処理、配列の作成、項目の分割を行うことができます。このボタンをクリックすると、「集計処理」ダイアログ・ボックスが表示されます（図 1-2）。このダイアログ・ボックスには 3 つのリストがあり、中央のリストには現在選択されているテーブルの全項目が表示されます。

集計処理を実行するには、中央のリスト内で項目の名前を選択し、左側のリストと中央のリスト間にある左向き矢印をクリックします（選択された項目が左側のリストに移動します）。右下のポップアップ・メニューでは、配列型項目のインデックスまたは分割型項目の分割基準を指定します。

項目値の集計処理のタイプ（ 合計、平均、最小、最大 、カウント）を指定するには、ダイアログ・ボックスの左下にある 5 つのトグルボタンを使用します。数値型の項目を集計処理するときは、これらのオプション（集計処理タイプ）を任意に組み合わせることができます。数値以外の項目については、「カウント」だけを選択できます。複数のオプション（集計処理タイプ）を選択すると、複数の項目が作成されます。たとえば、「平均 」と「最大 」を選択すると、平均値を表す 1 つの項目と最大値を表すもう 1 つの項目が作成されます。「集計処理に null を含める」チェックボックスを使用すると、集計処理時に Null 値を計算対象に入れるか除外するかを指定することができます。

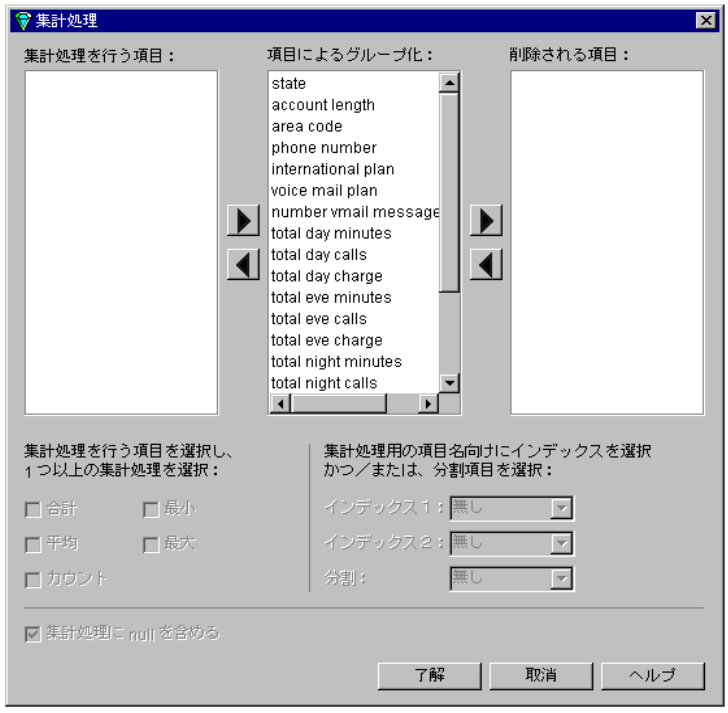

図 1-2　「集計処理」ダイアログ・ボックス

「集計処理 」ダイアログ・ボックスには、次の 3 つの項目名リストがあります。

- 「集計処理を行う項目」
- 「項目によるグループ化」（デフォルト）
  このリスト内の項目は集計処理中に変更されません（集計処理時にグルーピングの基準としてのみ使用されます）。これらの各項目の値の組合わせと一致する複数のレコードが集計処理されて単一のレコードが作成され、新しい項目としてテーブルに追加されます。新しい（集計処理された）項目の値は、ダイアログ・ボックスの左下にあるトグルボタンの設定に応じて、合計、平均、最小、最大、カウントいずれかになります。
- 「削除される項目」

ダイアログ・ボックス内で集計処理条件を設定して「了解」ボタンをクリックすると、集計処理条件に基づいて作成された新しい項目名が「 テーブル操作」ウィンドウの「現在のデータセットの項目名」テキストボックスに表示されます。

集計処理を通じて配列型項目または分割型項目を作成するときは、ダイアログ・ボックスの右下のポップアップ・メニューを使用して、配列型項目のインデックスまたは分割型項目の分割基準を指定します。詳細については、「配列（Arrays）」（6 ページ）と「分割型項目」（7 ページ）を参照してください。

配列（Arrays）

配列とは、特定のデータ型（浮動小数点、整数、文字列など）の変数からなる集合体です（使用可能なデータ型の一覧については、「項目の型または名前の変更」（42 ページ）を参照してください）。配列には常にインデックス（添字）が付けられます。配列のインデックスは、階級生成された項目でなければなりません。1 次元配列、2 次元配列、3 次元配列、さらにそれ以上の多次元配列を作成することができます。1 次元配列は一種のリスト、2 次元配列はテーブル（表）とみなすことができます。配列の次元が増えるに従って、可視化が難しくなります。

ツリー・ビジュアライザ (Tree Visualizer) では、配列を使用すると便利です。スキャタ・ビジュアライザ (Scatter Visualizer)、スプラット・ビジュアライザ (Splat Visualizer)、マップ・ビジュアライザ (Map Visualizer) のスライダをカスタマイズするときは、配列を使用する必要があります。

たとえば、州別・年齢別に消費額（$ 単位）を表すデータセットがあると想定します。データセットの行数を減らすには、対象者の年齢を 0-20、21-40、41-60 の 3 つのグループに階級生成して集計します。その結果、表 1-1 に示す表が作成されます。

表 1-1 州別・年齢別に集計した消費パターン

State（州）	Age_bin（年齢階級）	Total $ Spent（消費額）
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 には州別・年齢階級別に集計された 6 行のデータが表示されています。表 1-2 に示すように、同じ州の消費額を 1 つの項目内の 1 次元配列で表すこともできます。表 1-2 では、「*Total $ Spent*（消費額）」項目が配列になり、階級生成された年齢（「*Age_bin* (年齢階級)」項目）がその配列のインデックスとして使用されています。

表 1-2 階級生成された年齢をインデックスとする配列

State（州）	Total $ Spent [Age_bin]（消費額 [年齢階級]）
CA	[$50, $454, $693]
NY	[$35, $541, $628]

分割型項目

分割型項目は配列型項目と似ていますが、大きな相違がいくつかあります。分割型項目の場合は、複数の値を持つ単一の項目（配列）が作成されるのではなく、インデックスの個々の値ごとに 1 つの項目が作成されます。たとえば、表 1-2 のデータから配列を作成しないで、階級生成された年齢（「Age_bin (年齢階級)」項目）に基づいて消費額を分割すると、表 1-3 に示すようなテーブルが作成されます。

表 1-3 階級生成された年齢を基準として分割された項目

State **(州)**	Total $ Spent 0-20 **(年齢 0-20 の消費額)**	Total $ Spent 21-40 **(年齢 21-40 の消費額)**	Total $ Spent 41-60 **(年齢 41-60 の消費額)**
CA	$50	$454	$693
NY	$35	$541	$628

この例では、表 1-1 と比べて項目数が増えていますが、行数は少なくなっています。

階級生成された項目が複数存在する場合は、それらの項目に基づいてレコード値をクロス集計することができます。たとえば、Age_bin（年齢階級）および Salary_bin（給与階級）という項目が存在する場合は、年齢階級別・給与階級別に消費額を集計することができます。「Salary_bin (給与階級)」項目の個別値を追加して表 1-3 のテーブルを細分したデータを表 1-4 に示します（ここではカリフォルニア州 (CA) のデータだけを示しています）。このようにテーブルを細分すると、年齢グループ別に加えて所得水準別に消費動向を分析することができます。データセットの特性を詳しく分析する場合は、このような手法を利用してください。

表 1-4 年齢階級別・給与階級別に集計した消費額

State **(州)**	Age_bin **(年齢階級)**	Salary_bin **(給与階級)**	Total $ Spent **(消費額)**
CA	0-20	$0-$25,000	$30
CA	0-20	$25,001-$50,000	$15
CA	0-20	Over $50,000	$5
CA	21-40	$0-$25,000	$120
CA	21-40	$25,001-$50,000	$234
CA	21-40	Over $50,000	$100
CA	41-60	$0-$25,000	$101

表 1-4（続き）年齢階級別・給与階級別に集計した消費額

State **（州）**	Age_bin **（年齢階級）**	Salary_bin **（給与階級）**	Total $ Spent **（消費額）**
CA	41-60	$25,001-$50,000	$290
CA	41-60	Over $50,000	$302

Age_bin（年齢階級）をインデックスとして「Total $ Spent（消費額）」項目を 1 次元配列に変換すると、同じデータによって表 1-5 に示すテーブルが作成されます。

表 1-5 年齢階級をインデックスとして消費額を 1 次元配列にしたテーブル

State **（州）**	Salary_bin **（給与階級）**	Total $ Spent [Age_bin] **（消費額 [年齢階級]）**
CA	$0-$25,000	[$30, $120, $101]
CA	$25,001-$50,000	[$15, $234, $290]
CA	Over $50,000	[$5, $100, $302]

Age_bin（年齢階級）と Salary_bin（給与階級）をインデックスとして「Total $ Spent（消費額）」項目を 2 次元配列に変換すると、表 1-6 に示すテーブルが作成されます。

表 1-6 年齢階級と給与階級をインデックスとして消費額を 2 次元配列にしたテーブル

State**（州）**	Total $ Spent [Age_bin] [Salary_bin]**（消費額 [年齢階級][給与階級]）**
CA	[$30, $120, $101, $15, $234, $290, $5, $100, $302]

最後に、Age_bin（年齢階級）をインデックスとして消費額を配列に変換した上で、Salary_bin（給与階級）を基準として配列の各要素を分割すると、表 1-7 に示すテーブルが作成されます。

表 1-7 年齢階級をインデックスとする配列（消費額）を給与階級に基づいて分割したテーブル

State（州）	Total $ Spent [Age_bin], Salary $0-25,000（消費額 [年齢階級], 給与 $0-25,000）	Total $ Spent [Age_bin], Salary $25,001-$50,000（消費額 [年齢階級], 給与 $25,001-$50,000）	Total $ Spent [Age_bin], Salary Over $50,000（消費額 [年齢階級], 給与 $50,000 超）
CA	[$30, $120, $101]	[$15, $234, $290]	[$5, $100, $302]

上記の例では、配列の各要素に対応する値が 1 つだけ存在し、分割によって既存のデータ値が再編成されているだけですが、MineSet では複数のデータ値を単一の配列要素に分割するための集計処理オプションがいくつか用意されています。最もよく使用されるのは、データ値を合計するオプション（合計）です。このオプションは各種の経費を合計して予算を計算する場合などに便利です。合計の他に、最小、最大、平均、カウントの各オプションがあります。

データセットの値を分割する場合、特定の階級に該当する値が存在しない可能性もあります。そのような場合、最小、最大、平均、合計の集計処理オプションでは、データムーバ (DataMover) によって NULL 値が割当てられます。カウントオプションでは、0 が割当てられます。

アニメーション

スキャタ・ビジュアライザ (Scatter Visualizer)、スプラット・ビジュアライザ (Splat Visualizer)、マップ・ビジュアライザ (Map Visualizer) では、使用中のデータセットの少なくとも 1 つのスライダ要素が項目に割当てられている場合に、アニメーションを実行することができます。アニメーションを使用すると、時間などの次元に沿ってデータセットの変化を表示することができます。通常、時間や年齢などの独立属性がスライダ次元として最も適していますが、階級生成された項目を使用することもできます。

アニメーション・コントロール・パネル

アニメーション・コントロール・パネルは、ビジュアライザのメインウィンドウの右側に表示されます。アニメーション・コントロール・パネルの内部には、サマリウィンドウ（およびそれに隣接する最高 2 つのスライダ・コントロール）、情報フィールド、アニメーション・ボタン、パススライダ、スピードスライダ、データポイントの表示 / 非表示を切替えるチェックボックス、同期スライダ、（さらにスキャタ・ビジュアライザの場合は）「軌跡」メニューがあります。

独立次元を制御するスライダ

サマリウィンドウの左側と下側に表示されるスライダの数は、ビジュアライザのメインウィンドウに表示されるデータセットに応じて異なります。1 つまたは 2 つのスライダ次元（独立次元）を持つデータセットもあれば、独立次元を持たないデータセットもあります。

2 つの独立次元を持つデータセット

相互に独立して値が変化する 2 つの次元を持つデータセット（たとえば *company.scatterviz* など）の場合は、2 つのスライダ・コントロールを持つアニメーション・コントロール・パネルがメインウィンドウの右側に表示されます（図 1-3 参照）。

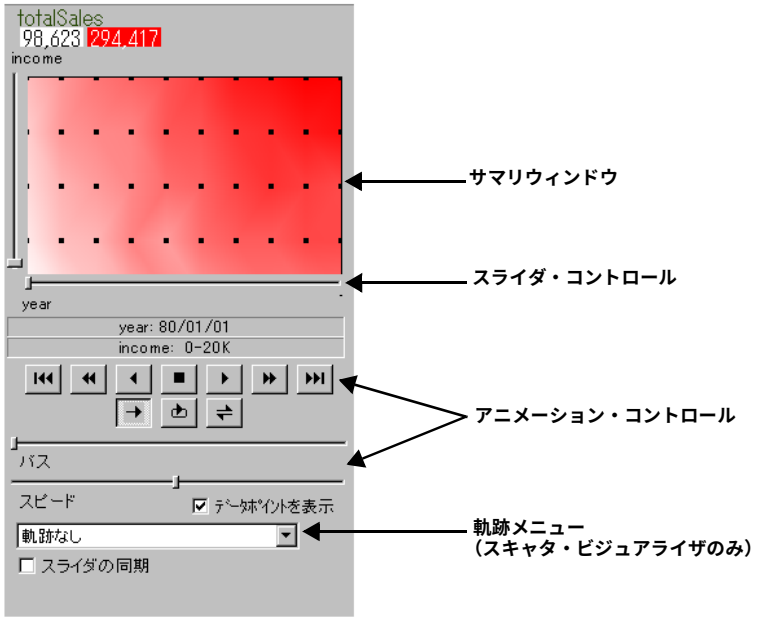

図 1-3 サマリウィンドウと 2 つのスライダを持つアニメーション・コントロール・パネル

アニメーション・コントロール・パネルの上部には、サマリウィンドウとスライダ・コントロールが表示されます。サマリウィンドウの下側の水平スライダは、最初の独立次元のデータポイントを選択するときに使用します。左側の垂直スライダは、2 番目の独立次元のデータポイントを選択するときに使用します。水平スライダの次元（属性）の名前はスライダの下にあるラベルで示され、垂直スライダの次元（属性）の名前はスライダの上にあるラベルで示されます。

Tool Manager を使用してスライダを設定するには、次のような方法があります。

- マップ・ビジュアライザ、スキャタ・ビジュアライザ、またはスプラット・ビジュアライザを使用する場合は、Tool Manager の「データの可視化 / マイニング」パネル上で、項目名を「スライダ 1」と「スライダ 2」に対応付けます。
- 「集計処理」ダイアログ・ボックスを使用して 1 つまたは 2 つの配列型項目を作成します。これらの配列型項目はビジュアライザのアニメーション・スライダで自動的に使用されます。

アニメーション・サマリウィンドウ

アニメーション・サマリウィンドウには、アニメーション・スライダの各種設定に基づき、サマリ可視化要素に対応付けされた項目の合計値が表示されます。すなわち、サマリウィンドウでは、スライダ次元に沿ってサマリ属性がどのように変化するかが示されます。サマリウィンドウ内の領域のカラー濃度が薄い（色が白い）ほど、メインウィンドウ内の要素によって表される合計値は小さくなり、カラー濃度が強いほど、合計値は大きくなります。

また、サマリウィンドウには、データの 1 次元または 2 次元に沿って均一の間隔で黒いドットが表示されます。これらの黒いドットは、離散データポイントの正確な位置を示しています。「データポイントの表示」チェックボックスを使用すると、これらの黒いドットを非表示にすることができます。

たとえば、*company.scatterviz* ファイルを最初に開くと、白（左側）から赤（右側）の範囲の色が 2D サマリウィンドウに表示されます。白は契約高が少ないことを表し、赤は契約高が多いことを表しています。この例では、赤の濃度が強くなるほど、生命保険、自動車保険、住宅保険の契約高が多くなります。

サマリウィンドウの黒いドット間を結ぶ経路は、アニメーション・パス (*animation path*) と呼ばれます。データのアニメーションは、このアニメーション・パスに沿って再生されます。

サマリウィンドウ内でアニメーション・パスを作成するには、最初に可視化ファイルを開いた後、次のいずれかの手順を実行します。

- マウスの左ボタンで黒いドットをクリックしてパスの始点を設定し、マウスの左ボタンを押したままパスの終点までカーソルをドラッグします。マウスの左ボタンを放すと、パスが終結します。
- マウスの左ボタンで黒いドットをクリックしてパスの始点を設定した後、カーソルを別の場所（パスの終点）に移動し、マウスの中ボタンをクリックして終点を設定します。こうすると、パスの始点と終点の間に直線が表示されます。マウスの中ボタンを繰返しクリックすると、別の直線（パス）を追加することができます。この方法は 3 ボタンのマウスでしか使用できません。
- マウスの左ボタンで黒いドットをクリックしてパスの始点を設定した後、独立次元のスライダの 1 つをドラッグし、その次元の軸に沿った直線を引きます。スライダが 2 つある場合は、2 番目のスライダをドラッグし、その次元の軸に沿った別の直線を引きます。

アニメーション用のボタンとスライダ

アニメーションを制御するには、2D サマリウィンドウの下にあるビデオに似たボタンとスライダ（「パス」と「スピード」）を使用します。

アニメーション制御ボタン

サマリウィンドウ内でパスを作成すると（「アニメーション・サマリウィンドウ」（13 ページ）を参照）、サマリウィンドウの下にあるビデオ様式のボタンを使用して、パスに沿ったアニメーションを制御することができます。最初は中央の「 停止」ボタンが青色で強調表示されています。「停止」ボタンの右側にある「 再生」ボタンをクリックすると、パスに沿って順方向にアニメーションの再生が開始されます。「停止」ボタンの左側にある「逆再生」ボタンをクリックすると、パスに沿って逆方向にアニメーションの再生が開始されます。「再生」ボタンと「逆再生」ボタンは、左から右または右から左への動きではなく、アニメーション・パスが描かれた順序に基づく動きを制御します。

アニメーションを一度停止してから再開するには、「停止」ボタンをクリックした後、「再生」ボタンまたは「逆再生」ボタンをクリックします。ただし、「停止」ボタンをクリックしても、一番近い離散データポイントに達するまでアニメーションは停止しません。

「再生」ボタンと「逆再生」ボタンの隣りには、「最速の先送り」ボタンと「最速の巻戻し」ボタンがあります。停止状態のときにこれらのボタンをクリックすると、パスの位置が終点（「最速の先送り」ボタンを押したとき）または始点（「最速の巻戻し」ボタンを押したとき）まで移動します。アニメーション再生中にこれらのボタンをクリックすると、再生が高速になります。

一番外側には順方向と逆方向の「 一つのステップずつ先送り」ボタンがあります。これらのボタンをクリックすると、現在のパスの位置が前後の一番近い離散データポイントまで移動します。

アニメーション の動作フロー用ボタン

アニメーション制御ボタンの下には、次の 3 つのアニメーション の動作用ボタンがあります。

「一回だけアニメーションを実行」ボタン（デフォルト）— 再生中にパスの終点に達するか、逆再生中にパスの始点に達すると、アニメーションは停止します。

「ループ」ボタン — パスの終点または始点に到達したら、パスの始点または終点に戻って、アニメーションを繰返し再生します。

「スイング」ボタン — パスの終点または始点に到達したら、再生方向を逆転してアニメーションを繰返し再生します。

アニメーション用のスライダ

階級生成処理や集計処理と組み合わせて使用されるアニメーション用のスライダは、自動的に作成するか手作業で作成することができます。この操作は Tool Manager の現行履歴には表示されませんが、ツールの設定ファイルに保存されます。

「スライダ 1」と「スライダ 2」に対応付けられる項目は、アニメーションで使用される項目（たとえば、色やサイズ）のインデックスになります。これらの項目は数値型（整数、浮動小数点、倍精度浮動小数点）であるか、または階級生成されていなければなりません。スライダに対応付けされた項目が既に階級生成されている場合、階級自動生成は実行されず、その項目がスライダのインデックスとして使用されます。一方、項目が階級生成されていない場合は、階級自動生成が実行されます（「階級生成」（39 ページ）を参照）。

アニメーションの停止中は、「パス」スライダを使用して、パス上の位置を変更することができます。「パス」スライダをドラッグすると、サマリウィンドウ内のカーソルがパスに沿って動き、（描画領域の下と左側にある）1D スライダもカーソルに従って動きます。「再生」ボタンまたは「逆再生」ボタンをクリックすると、新しく指定した位置からアニメーションが再開されます。「パス」スライダは、パス上の離散データポイント間にある任意の位置にドラッグすることができます。ただし、マウスのボタンを放すと、パスの位置は一番近い離散データポイントに移動します。

パスに沿ったアニメーションの再生速度を調整するには、「スピード」スライダを使用します。

データポイントと補間

スキャタ・ビジュアライザ、スプラット・ビジュアライザ、またはマップ・ビジュアライザでは、アニメーションの進行に伴って、サイズ、色、軸（位置）に対応付けされた項目（属性）の集計値が滑らかに変化します。ただし、「選択」メッセージ・ボックスと「ポインタを通過」フィールドには、一番近い離散データポイントの値だけが表示され、各データポイント間で補間されたデータ値は表示されません。

ここでアニメーションが再生される過程を簡単に説明します。たとえば、1991 年と 1992 年のデータ値が存在し、これらのデータ値がスキャタ・ビジュアライザの特定要素のサイズに対応付けされていると仮定します。また、1991 年のサイズは 20、1992 年のサイズは 40 と仮定します。年度のスライダを 1991 年から 1992 年に移動すると、要素のサイズは 20 から 40 の間で均等に補間されて変化します。たとえば、1991 年と 1992 年の中間点における要素のサイズは 30 です。1992 年に近づくにつれ、要素のサイズは 40 に近づきます。ただし、アニメーションを 2 つの離散データポイントの中間で停止することはできません。また、「パス」スライダを離散データポイントの中間位置にドラッグして停止することもできません。

サマリウィンドウ内のデータポイントは、データファイル内にある実際のデータに対応するスライダ位置を示します。たとえば、20 および 40 というサイズは実際のデータに対応していますが、30 は実際のデータには対応していません。この例のサマリウィンドウでは、各年度に対応するスライダ位置にデータポイントが存在します。

スライダの動きに伴って、すべての項目値を変化させる必要はありません。たとえば、スライダが 2 つある場合、一部の項目はどちらか一方のスライダに応じて変化させ、残りの項目は両方のスライダに応じて変化させることができます。

動作試験

スキャタ・ビジュアライザでは、「軌跡」メニューを使用して、アニメーション中に移動するポイントの軌跡を表示することができます。このメニューでは、「ライン軌跡」、「フェードアウト軌跡」、「チューブ軌跡」、または「軌跡なし」を選択することができます。詳細については、『*MineSet 3.0 Enterprise Edition User's Guide for Windows*』のスキャタ・ビジュアライザとスプラット・ビジュアライザにおけるアニメーションの作成に関する説明を参照してください。

モデルの適用

Tool Manager の「データ変換」パネルにある 「モデルの適用」ボタンを使用すると、次の処理を実行することができます。

- 以前に作成したクラシファイア (Classifier) を新しいデータに適用する。
- 以前に作成したクラシファイアを現在のテーブルに適用して、クラシファイアの精度をテストする。
- 以前に作成したクラシファイアの構造に現在のテーブル内のデータを適合（フィット）させる。

次のダイアログ・ボックス（図 1-4）の左上には、サーバ上で利用できる クラシファイアのリストが表示されます。クラシファイアを選択すると、そのクラシファイアに必要な項目の名前とデータ型が右側のリストに表示されます。クラシファイアに必要なすべての項目が現在のテーブルに含まれている場合は、ウィンドウの下部にその旨のメッセージが表示されます。選択したクラシファイアに必要なすべての項目が現在のテーブルに含まれていない場合は、ウィンドウの下部にその旨のメッセージが表示され、欠けている項目が右側のリストに強調表示され、下にある 3 つのボタンは使用できない状態になります。

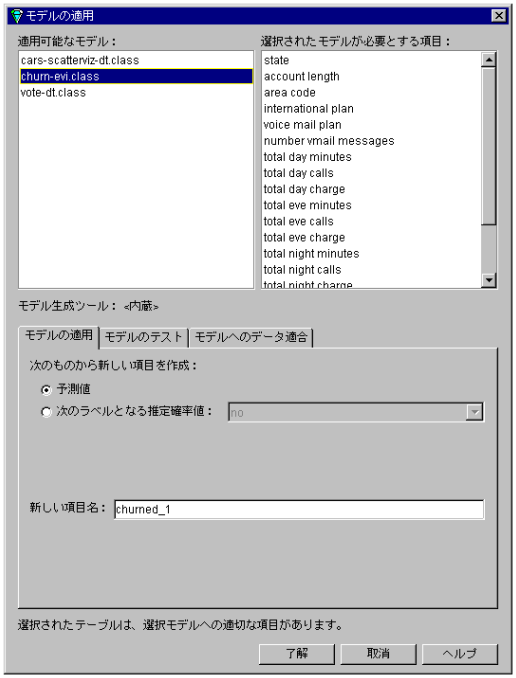

図 1-4 「モデルの適用」ダイアログ・ボックス：クラシファイアの選択

「モデルの適用」パネル

以前に作成したクラシファイアを現在のテーブルに適用するには、「モデルの適用」パネル（図 1-5）を使用します。クラシファイアを適用するときは、次の 2 つのオプションを選択することができます。

- テーブル内の各レコードの予測値
  たとえば、解約する顧客を判断するクラシファイアを作成した場合は、このオプションを使用して、個々の顧客の解約傾向の有無を示す項目を追加することができます。
- 次のラベルとなる推定確率値
  クラシファイアを使用して各レコードのラベル値を予測する代わりに、各レコードが次のラベル（churn = yes など）となる確率を予想します。たとえば、解約する顧客を判断するクラシファイアを作成した場合は、このオプションを使用して、個々の顧客が解約する確率を示す項目を追加することができます。

新しい項目の名前を指定するには、「新しい項目名」テキスト・フィールドを使用します。

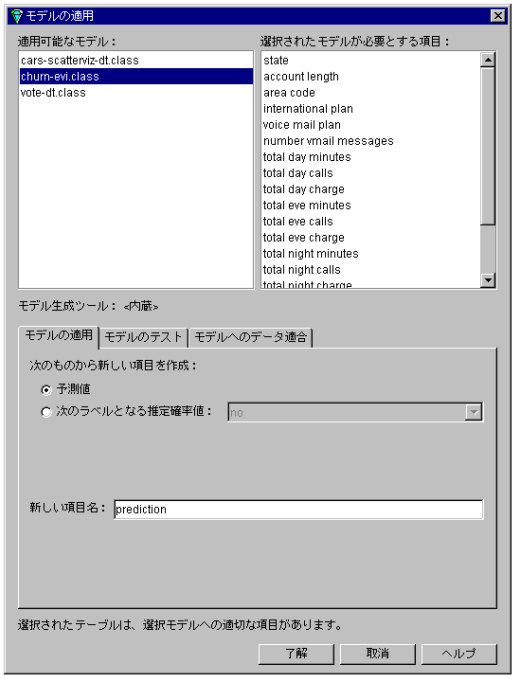

図 1-5 「モデルの適用」パネル

「モデルのテスト」パネル

以前に作成したクラシファイアを現在のテーブルに適用して、クラシファイアの精度をテスト（検定）するには、「モデルのテスト」パネルを使用します。クラシファイアをテストするには、そのクラシファイアに必要なすべての項目（名前とデータ型が一致する項目）がテーブルに含まれていなければなりません。また、モデルを適用する場合と違って、モデルをテストするときは、クラシファイアの構築時に使用したラベルと同じ名前とデータ型のラベルがテーブルに含まれていなければなりません。

「モデルのテスト」パネルには、次のような機能を持つオプションがあります。

- テーブルのレコードに関する混同マトリックスを表示する。
- 特定のラベル値に関する 改善曲線を表示する。
- 特定のラベル値に関する ROI 曲線を表示する。
- テストセット (test set) として使用されるレコードと一緒にクラシファイアを可視化する（決定木 (Decision Tree) または選択式決定木 (Option Tree) の場合のみ）。
- レコードの重みとして使用する属性を選択する。

モデルのテスト結果は、パネルの下部にあるテキスト・フィールドに表示されます。

「モデルへのデータ適合」パネル

以前に作成したクラシファイアの構造に現在のテーブル内のデータを適合させるには、「モデルへのデータ適合」パネルを使用します。クラシファイアにデータを適合すると、元のモデルと同じ構造を持つ新しい分析モデルが生成されます。ただし、新しいクラシファイアでは、現在のテーブル内のデータに基づいて確率評価が更新されます(「バックフィッティング」（37 ページ）を参照)。現在のテーブル内の全データがクラシファイアの構造に適合されるため、誤差推定は行われません。ブースティングによって構築されたクラシファイアの場合は、「モデルへのデータ適合」パネルを使用できません。別個のテストセット（適合されるデータとは全く別のレコードセット）に基づいて新しいクラシファイアの精度を評価する場合は、「モデルのテスト」パネルを使用してください。

「モデルへのデータ適合」パネルには、次のような機能を持つオプションがあります。

- 新しいクラシファイアを可視化する。
- 新しいクラシファイアの名前を指定する。
- レコードの重みとして使用する属性を選択する。

モデルの適用

構築された 予測モデル をレコードに適用すると、レコードのラベルを予測することができます。たとえば、アヤメの種類を予測するクラシファイア（離散型ラベルを予測するモデル）を構築した場合、記述属性だけを含むレコードにそのクラシファイアを適用すると、予測されたラベル（アヤメの種類）を示す新しい項目が追加されます。

クラシファイアの精度を確認するには、クラシファイアを構築した後で、そのクラシファイアを訓練事例に適用し、誤ったラベルが予測されたレコードの有無を調べます。そのようなレコードは「ノイズ」であるか、特別な意味を持つレコードであるため、詳しい調査が必要になります。

たとえば、「クラシファイアのみ」モードを使用してアヤメ (Iris) データセットの決定木を作成した場合、そのクラシファイアをデータセットに適用すると、予測されたラベルを示す新しい項目 (iris type_1) が作成されます。次に、(iris type != iris type_1) という表現（式）を使用して定義した整数 (int) 型の項目を追加します。この判定用の項目の値は、クラシファイアによって誤ったラベルが割当てられた場合は 1 になり、正しいラベルが割当てられた場合は 0 になります。判定用の項目の値が 0 (OK) の場合は緑、1（誤り）の場合は赤でレコードが描画されるようにカラーセットの色に項目を対応付けし、各レコードをプロットしたスキャタ・ビジュアライザを図 1-6 に示します。このグラフを見ると、誤ったラベルが割当てられた箇所が分かります。

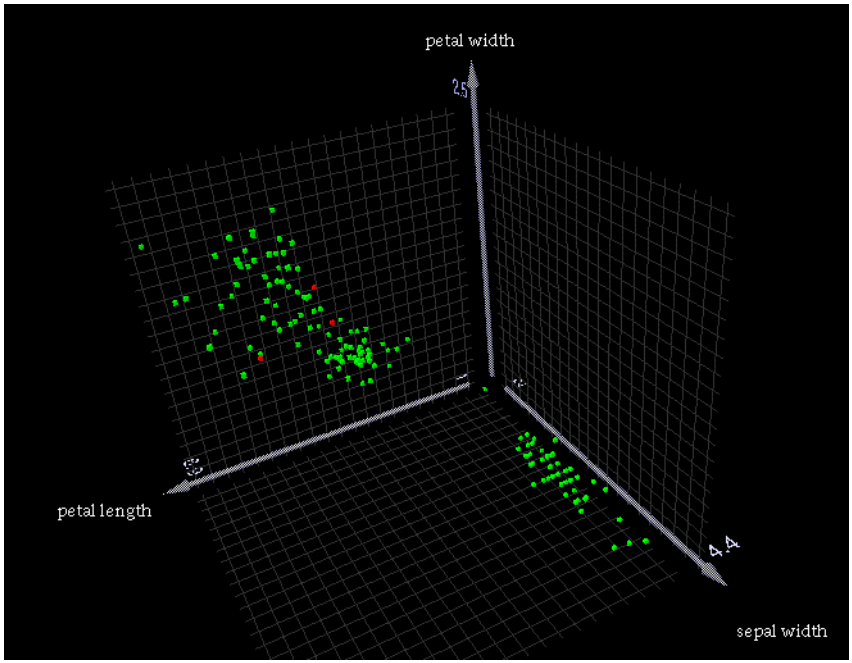

図 1-6 アヤメ (Iris) データセットの誤ったクラス判別（例 1）

または、(iris type != iris type_1) + 0.01 という表現（式）を使用して単精度浮動小数点数 (float) 型の新しい項目を追加しても、誤ったクラス判別を見つけることができます。元のラベルを色に対応付けし、この新しい項目をサイズに対応付けしてスキャタ・ビジュアライザを実行すると、誤ったクラス判別が生成されたレコードが大きい立方体で表示され、正しいクラス判別が生成されたレコードは小さい立方体で表示されます（図 1-7 を参照）。

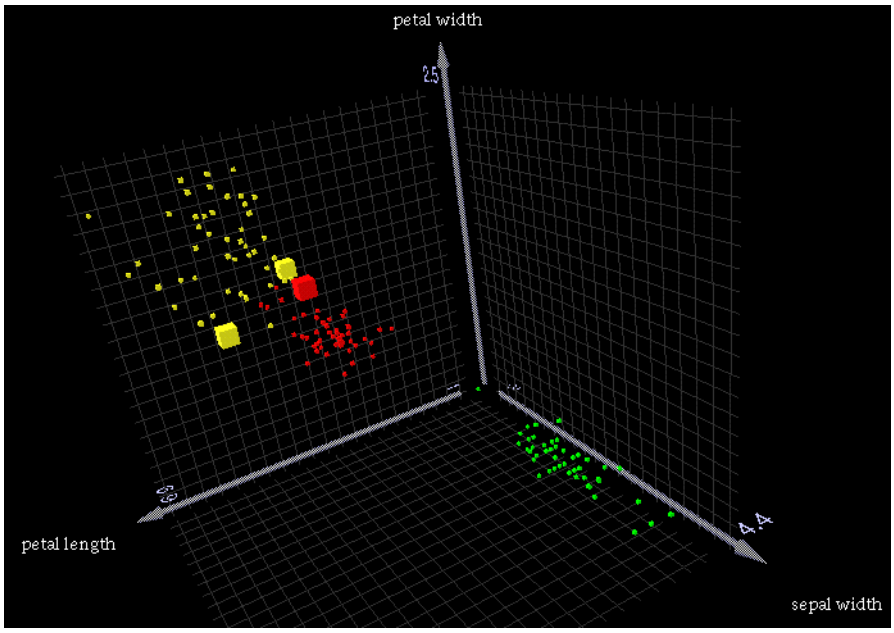

図 1-7 アヤメ (Iris) データセットの誤ったクラス判別（例 2）

予測値の誤差とクラス判別の誤りの詳細については、「誤差推定」（87 ページ）を参照してください。

相関規則

MineSet の相関規則分析を使用すると、膨大なデータベース内に潜む特定のデータパターンの検出、検証、可視化を通じて、データを発掘することができます。検出されたパターンは相関規則によって表され、データセット内でどのような組み合わせの属性値が一緒に発生するかを示します。相関規則の代表的な適用例に「マーケット・バスケット分析」があります。マーケット・バスケット分析では、ショッピングの「買物かご」に一緒に入れられる傾向のある商品を検出します。

相関規則を検出してグラフィック表示するという分析手法は、スーパーマーケットの在庫管理、商品棚の構成、ダイレクトマーケティングのような数多くの業務に役立ちます。

相関規則を適用する操作は、次の 2 つの段階に分けられます。

1. 相関規則の生成 — データファイルを相関規則分析によって処理し、スキャタ・ビジュアライザで取り扱えるファイルを生成します。
2. 相関規則の可視化 — 生成された相関規則をグラフィカルに表示します。

相関規則分析では、単純な相関規則（左辺と右辺が 1 対 1 に対応する規則）または多重の相関規則（左辺と右辺に複数の事象が含まれる規則）を生成することができます。ここでは、単純な相関規則について説明します。多重の相関規則の説明については、「多重規則」（28 ページ）を参照してください。

単純な相関規則では、X が “ 真 ” であると仮定した場合、一定の確率で Y も “ 真 ” になるという関係が成立します。MineSet では規則の左辺 (LHS) を X で表し、右辺 (RHS) を Y で表します。

相関規則の適用例として、「マーケット・バスケット」のデータに基づく顧客の購買パターンの分析が挙げられます。ここで言うマーケット・バスケットとは、顧客が 1 回の来店で購入する商品のリストです。たとえば、「おむつを購入する顧客の 80% はベビーパウダーも一緒に購入する」という相関規則がこの分析から導出されます。80% という確率値は、相関規則の信頼度と呼ばれます（下記参照）。

この例では、「おむつ」が規則の左辺 (LHS) の事象 (X)、「ベビーパウダー」が規則の右辺 (RHS) の事象 (Y) です。

このような相関規則の適用例をいくつか紹介します。

- 商品 A が右辺 (RHS) にある場合、左辺 (LHS) は商品 A の売上げを伸ばす対策を決定するのに役立ちます。
- 商品 B が左辺 (LHS) にある場合、右辺 (RHS) は商品 B の販売を中止した場合にどの製品が影響を受けるかを判断するのに役立ちます。

相関規則分析は入力ファイルを処理し、生成された相関規則を格納した出力ファイルを作成します。レコード内に X および Y という属性（事象）がある場合、次の形式の規則が生成されます。

X=>*Y*

この相関規則は、レコード内に事象 X（左辺）が発生した場合、一定の頻度（確率）で事象 Y（右辺）も発生すると予想されることを示します。

相関規則の精度（説明力）は次の 4 つの指標値によって定量化されます。

- 支持度はデータセット全体において特定の相関規則が観察される頻度です。すなわち、支持度は左辺 (LHS) の事象 (*X*) と右辺 (RHS) の事象 (*Y*) が一緒に発生する度数をレコードの総数で割った値です。たとえば、支持度が 1% である場合は、レコード総数の 1% において事象 *X* と事象 *Y* が一緒に発生します。

  生成される相関規則について支持度の最小区間を指定することができます。デフォルトの区間は 1% です。支持度が小さいほど多くの規則が生成されるため、ツールのパフォーマンス（実行速度）は低下します。

  支持度の最小区間を満たす相関規則が重要なのは、次の 2 つの理由によります。

  – 特定の規則に実務上の価値が生じるのは、かなりの割合のレコードにその規則を適用できる場合に限られます。たとえば、キャビアを購入する人が全員ウォッカも購入するならば、規則 Caviar => Vodka の信頼度は 100% となります。ただし、キャビアを買う人がほんの少数しかいない（すなわち、支持度が極めて低い）場合、この規則は小売業者にとって利用価値の低いものになります。

  – 規則を適用できるレコードの数が極めて少ない場合、その規則は統計学的に有意とは言えません。そのような規則は偶然の産物である可能性が高いため、その規則に基づいて意思決定を下すのは賢明でありません。

- 信頼度は左辺 (LHS) の事象 (*X*) と右辺 (RHS) の事象 (*Y*) が一緒に発生するレコード数を左辺 (LHS) の事象 (*X*) が発生するレコード数で割った値です。たとえば、事象 *X* と事象 *Y* の相関規則の信頼度が 50% である場合は、事象 *X* が発生する全レコードの 50% において、事象 *Y* が同時に発生することが期待されます。したがって、特定のレコードで事象 *X* が発生することが分かれば、そのレコードで事象 *Y* が同時に発生する確率は 50% ということになります。生成される相関規則について信頼度の最小区間を指定することができます。デフォルトの最小区間は 50% です。

- 期待される信頼度は右辺 (RHS) の事象 (Y) が発生するレコード数をデータセット全体のレコード総数で割った値です。すなわち、左辺 (LHS) の事象と右辺 (RHS) の事象が完全に独立である（無相関である）場合に、規則の信頼度がどのような値になるかを示します。したがって、信頼度を期待される信頼度で割った比率（下記の改善率 (Lift)）は、左辺 (LHS) 事象によって予測能力がどの程度改善されるかを示す尺度になります。

- 改善率は、信頼度を期待される信頼度で割った比率です。この比率が大きいほど、相関規則の予測能力が高いことになります。すなわち、特定のレコードで右辺 (RHS) の事象 (Y) が発生するかどうかを判断するときに、左辺 (LHS) の事象 (X) によってどの程度の情報が得られるかを示します。

相関規則分析では、信頼度が期待される信頼度を下回る（改善率が 1.0 未満である）相関規則は報告されません。すなわち、事象 Y だけが発生する度数よりも事象 X と Y が一緒に発生する度数のほうが少ない場合、X => Y という規則は報告されません。

注記 : 事象 Y だけが発生することが分かっている場合、X => Y という形式の相関規則は、事象 X の発生に関する情報を何も提供しません。相関規則 X => Y は、左辺から右辺への片方向だけの関係を示します。

「相関規則分析」タブ

相関規則を生成するときは、Tool Manager の「データの可視化 / マイニング」パネルにある「マイニングツール」タブをクリックし、表示されるタブから「相関規則分析」を選択します。相関規則は、データセット内でどのような組み合わせの属性（項目）値が一緒に発生するかを示します。相関規則の代表的な適用例に「マーケット・バスケット分析」があります。

必要なファイル

スキャタ・ビジュアライザを使用して相関規則をグラフィカルに表示するには、次のファイルが必要です。これらのファイルは Tool Manager を通じて自動的に作成されます。

- 相関規則分析によって作成される規則ファイル（接尾辞は *.rules.data*）
- 相関規則をレコードビューワに表示するためのスキーマ・ファイル（接尾辞は *.rules.schema*）
- スキャタ・ビジュアライザ用の設定ファイル（接尾辞は *.rules.scatterviz*）

各ファイルの中間接尾辞 *.rules* は必須ではありませんが、Tool Manager を使用して設定ファイルを自動作成すると、この中間接尾辞がファイル名に付きます。

相関規則の可視化

スキャタ・ビジュアライザ (Scatter Visualizer) を起動して相関規則をグラフィカルに表示するには、いくつかの方法があります。

- Tool Manager を使用して、相関規則分析を設定して起動します（「相関規則の設定」（27 ページ）を参照）。相関規則が生成されると、Tool Manager によってスキャタ・ビジュアライザが自動的に起動されます。
- 使用する設定ファイルが分かっている場合は、（ファイル管理ウィンドウ内で）その設定ファイルのアイコンをダブルクリックします。こうするとスキャタ・ビジュアライザが起動され、選択した設定ファイルが自動的に読み込まれます。この起動方法を利用できるのは、設定ファイルの拡張子が .scatterviz である場合に限られます（Tool Manager を使用してスキャタ・ビジュアライザ用の設定ファイルを作成すると、ファイル拡張子が常に .scatterviz になります）。
- コマンド行に次のコマンドを直接入力します。
  - Windows システム上では、MS-DOS プロンプトのウィンドウに次のコマンドを入力します。

    `CD` *file-directory*

    `Viz` [*filename.scatterviz*]
  - IRIX システム上では、UNIX シェルのコマンド行プロンプトに次のコマンドを入力します。

    `scatterviz` [*filename.scatterviz*]

filename.scatterviz はスキャタ・ビジュアライザ用の設定ファイルの名前です。ビジュアライザを起動するときは、データファイルではなく設定ファイルを指定してください。

相関規則の設定

Tool Manager を使用して相関規則のオプションを設定すると、分析作業を大幅に簡素化することができます。相関規則のオプションを設定するには、Tool Manager の「データの可視化 / マイニング」パネルにある「マイニングツール」タブをクリックし、表示されるタブから「相関」を選択します。分かりやすいビジュアルを作成するには、データソースに応じて、項目値の階級生成や一部の項目の削除が必要になる場合があります。

相関規則のオプション

「相関」タブでは、次のようなオプションを指定することができます。

信頼度

生成される相関規則について信頼度の最小区間を指定します（デフォルトは 50%）。信頼度は左辺 (LHS) の事象 (*X*) と右辺 (RHS) の事象 (*Y*) が一緒に発生するレコード数を左辺 (LHS) の事象 (*X*) が発生するレコード数で割った値です。たとえば、事象 *X* と事象 *Y* の相関規則の信頼度が 50% である場合は、事象 X が発生する全レコードの 50% において、事象 *Y* が同時に発生することが期待されます。

支持度

生成される相関規則について支持度の最小区間を指定します（デフォルトは 1%）。支持度はデータセット全体において特定の相関規則が観察される頻度です。すなわち、支持度は左辺 (LHS) の事象 (*X*) と右辺 (RHS) の事象 (*Y*) が一緒に発生する度数をレコードの総数で割った値です。たとえば、支持度が 1% である場合は、レコード総数の 1% において事象 *X* と事象 *Y* が一緒に発生します。

重み

相関規則を生成するときは、特定のレコードが他のレコードよりも重要であると指定したり、標本抽出の不均一性を補正したりするために、レコードの重み付けを行うことができます。「重み付けとして使用」チェックボックスをオフにすると、各レコードの重みは 1 になります。このチェックボックスをオンに設定すると、各レコードの重み値が格納された項目（重み項目）を右側のポップアップ・メニューから選択することができます。相関規則分析によって生成される規則の中で重み項目を通常の属性として使用する場合は、「属性としても使用」チェックボックスをオンに設定します。このチェックボックスをオフにした場合、相関規則分析によって生成される規則の中で重み項目は使用されません（各レコードの重み付け専用に使用されます）。詳細については、「レコードの重み付け」（151 ページ）を参照してください。

多重規則

「多重規則」チェックボックスをオンに設定すると、左辺 (LHS) と右辺 (RHS) が 1 対 1 に対応する単純な規則に加えて、左辺 / 右辺に複数の事象が含まれる多重規則が相関規則分析によって生成されます。生成されるすべての規則は、信頼度と支持度に関する最小区間の制限（上記参照）を満たします。多重規則とは、"beer and linguini implies potato chips and salsa and wine."（ビールとパスタ料理を飲食する人は、ポテトチップス

とサルサとワインも同時に飲食する）のように、左辺または右辺に 2 つ以上の事象が含まれる規則です。

Tool Manager の「相関」タブで多重規則を生成するときの設定例を図 1-8 に示します。この図は Windows 版のウィンドウですが、IRIX 版のウィンドウもほぼ同じ構成です。

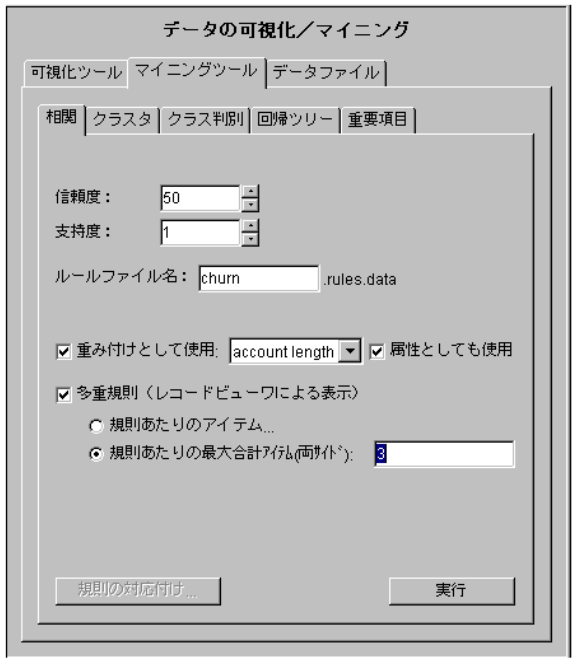

図 1-8 Tool Manager の「相関」タブにおける多重規則の設定例

多重規則を表示するときは、スキャタ・ビジュアライザではなくレコードビューワを使用します。レコードビューワのテーブルでは、1 行に 1 つの規則が表示されます。最初の 2 つの項目は、左辺 (LHS) の事象数と右辺 (RHS) の事象数を示します。次の 4 つの項目はそれぞれ、支持度、信頼度、期待される信頼度、改善率の値を示します。最後の 2 つの項目には、LHS と RHS の事象が表示されます。LHS と RHS の各事象は、単語 "and" で区切られます。前述の多重規則の例では、LHS の 2 つの事象が "beer and linguini"(ビールとリングィーニ "）と表示され、RHS の 3 つの事象が "potato chips and salsa and wine"（ポテトチップスとサルサとワイン）と表示されます。

生成される多重規則のサイズ（左辺と右辺の事象数の合計）を制限するには、「規則あたりの最大合計アイテム」フィールドに数値を入力します。この数値は、1 つの規則で使用可能な最大の事象数 (左辺と右辺の事象数の合計) を示します。前述の多重規則の例では、5 つの事象（左辺 = 2、右辺 = 3）が使用されています。左辺と右辺が 1 対 1 に対応する単純な規則の事象数は 2 つです。

注記 : 多重規則の生成には時間がかかります。相関規則分析による反復処理ごとに、生成される規則の数をステータス・ウィンドウで確認してください。生成される規則の数が多すぎる場合は操作を中止し、信頼度と支持度に関する最小区間を大きくするか、または「規則あたりの最大合計アイテム」を少なくしてください。

相関規則の対応付け用のボタン

相関規則は、スキャタ・ビジュアライザの可視化要素の属性に対応付けることができます。適切な対応付けを行って仮説を検定すると、相関規則の可視化をより細かく分析することができます。このような対応付けを行うには、Tool Manager の「データの可視化／マイニング」パネルの「ルールビジュアライザの対応付け」ボタンをクリックし、「相関規則の対応付け」パネル（図 1-9）を開きます。各可視化要素の右側にあるポップアップ・メニューには、その可視化要素で選択できる項目だけが表示されます。

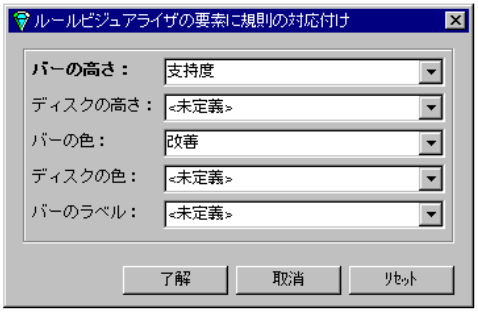

図 1-9 「相関規則の対応付け」パネル

このパネルでは、Tool Manager によって自動作成された規則ファイル (.rules.data) 内の各項目（支持度、信頼度、期待される信頼度、改善率など）を可視化要素に対応付けることができます。表 1-8 に、対応付け先の可視化要素とそのデフォルト設定を示します。規則の構成要素の説明については、「相関規則のオプション」（28 ページ）を参照してください。

表 1-8 相関規則の対応付け

可視化要素	指定する内容
バーの高さ	バーの高さに対応付ける項目（デフォルトは支持度 (Support)）
ディスクの高さ	ディスクの高さに対応付ける項目
バーの色	バーの色に対応付ける項目（デフォルトは改善率 (Lift)）
ディスクの色	ディスクの色に対応付ける項目
バーのラベル	バーのラベルに対応付ける項目

相関規則の可視化

相関規則をグラフィカルに表示すると、生成された複数の規則を詳しく分析・比較することができます。相関規則はスキャタ・ビジュアライザのグリッド・ランドスケープ上に表示されます。左辺 (LHS) の項目が 1 つの軸に表示され、右辺 (RHS) の項目がもう 1 つの軸に表示されます。個々の規則は左辺 (LHS) の項目と右辺 (RHS) の項目の交点に表示され、各規則の構成要素と特性はバー、ディスク、ラベルによって表されます。

表示されたビューが小さすぎる場合は、各軸の横に項目のラベルが表示されません。その場合は、項目のラベルが表示されるまで、Dolly ダイヤルを使用してビューをズームインしてください。また、特定の規則のラベルを表示するには、マウスを選択モードにして、目的のバーの上にマウスポインタを移動する方法もあります。こうすると、その規則の詳細情報がビュー領域の左上隅に表示されます。

メインウィンドウの下部には、スキャタ・ビジュアライザの可視化要素（バーの色や高さなど）と相関規則の構成要素（信頼度、支持度など）の対応関係を示す説明（凡例）が表示されます。

デフォルトでは、バーの高さが支持度を表し、バーの色が改善率を表します。グラフを十分にズームインすると、（設定ファイルで非表示と指定していない限り）左辺 (LHS) と右辺 (RHS) の各軸に項目名のラベルが表示されます。

Tool Manager またはテキストエディタを使用して設定ファイルを編集すると、スキャタ・ビジュアライザの可視化要素と相関規則の構成要素間の対応関係（バーとディスクの高さや色が表す構成要素、ラベルの値など）を変更することができます（表 1-8

を参照)。相関規則の構成要素を可視化要素に対応付けると、カラーマップが自動的に作成されます。このデフォルトのカラーマップを変更する場合は、設定ファイルを編集してください。

相関規則を表すバーの上にカーソルを置いて、マウスの左ボタンをクリックすると、「選択」ウィンドウに情報が表示されます。多重規則を選択する場合は、<Shift> キーを押しながらマウスの左ボタンをクリックします。

ドリルスルーとは、現在選択されているオブジェクトの元データ（データソースから取得したデータ）を表示して細かく分析する操作を指します。ドリルスルー表現は、選択された規則の論理積によって作成されます。オリジナルのテーブル中の項目は .rules.data ファイル内の項目と一致していないため、規則分析によって特殊な文字列型項目が作成され、ドリルスルーの実行時にその項目を使用してフィルタ表現が作成されます。*.rules.scatterviz* ファイル内ではこの特殊な文字列型項目がすでにドリルスルー表現に組込まれているため、ドリルスルーの「設定」パネルで各種オプションを変更しても、フィルタ表現は影響を受けません。

1 つまたは複数の規則のドリルスルーを実行すると、それらの規則を満たす全レコードが抽出されます。それらのレコードを表示するには、ビジュアライザの「選択」プルダウン・メニューから「オリジナルデータを表示」を選択します。ドリルスルーの詳細については、「ドリルスルー」（85 ページ）を参照してください。

相関規則のサンプルファイル

MineSet ソフトウェアでは、相関規則分析の機能や特長を紹介するためにサンプルファイルが用意されています。詳細については、付録 A「設定ファイルとデータファイルのサンプルファイル」を参照してください。

.ruleviz から .scatterviz へのファイル変換

MineSet 2.6 とそれ以前のバージョンでは、相関規則が独自の規則ビジュアライザで表示されており、多少異なる形式の設定ファイル (.ruleviz) が使用されていました。

規則ビジュアライザ用の設定ファイル (*.ruleviz*) を使用している場合は、その .ruleviz ファイルを編集してスキャタ・ビジュアライザ用の設定ファイル (*.scatterviz*) に変換し、拡張子を *.scatterviz* として保存してください。例 1-1 と 例 1-2 に、.ruleviz ファイルと *.scatterviz* ファイルの違いを示します。変更箇所が容易に分かるように、例 1-2 の設定ファイル (*.scatterviz*) にはコメント行が含まれています。これら 2 つの設定ファイルでは、同じデータファイルを使用しています。

注記： 従来の .ruleviz ファイルでは、サイズ が高さ、信頼度が予測値、支持度が普及率という用語で記述されていました。

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;
}
```

例 1-2 group.rules.scatterviz

```
MineSet 3.1
input
{
# Rename group.rules to group.rules.data:
    file "group.rules.data";

# The schema for the rules.data file is always
# the following. Add these lines:
    int nlhs;
    int nrhs;
    float support;
    float confidence;
    float `expected confidence`;
    string LHS;
    string RHS;
}
expressions {
    float lift  = confidence / `expected confidence`;
}


view
{
# This replaces height predictability:
    size confidence, scale 1.;
    size legend label "Bar Height: confidence";

# This replaces disk height expected:
    disk height `expected confidence`, scale 1.;
    disk height legend label "Disk Height: expected confidence";

# This replaces color prevalence:
    color support;
    color colors "white" "purple", legend label "Color: support";
    color scale 0 9;
    color legend "0%" "9%";

# Add these two axis mappings (not present in old file):
    axis RHS, max 100, orderby alpha;
    axis LHS, max 100, orderby alpha;

# Make sure the shape type is bar:
    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;
}
```

階級自動生成（Automatic Binning）

Tool Manager の「階級の生成」ダイアログ・ボックスには、階級自動生成を設定するためのオプションがあります。階級生成したい項目を指定した後、Windows 版の場合は「階級自動生成」、IRIX 版の場合は「区間自動生成 (Automatically chose number of bin)」タブの「自動的に階級を選択します」をクリックすると、MineSet が自動的に階級生成を実行します。この機能は機械学習を利用したプログラムを使用して階級を指摘する場合に便利です。階級生成のオプションの詳細については、「階級生成」（39 ページ）を参照してください。表 1-9 に、階級自動生成に関するオプションの一覧を示します。

表 1-9 階級自動生成のオプション

オプション	Windows 版	IRIX 版	機能説明
階級自動生成	基本オプション	N/A	選択された項目を自動的に階級生成する。
自動的に階級数を選択します	詳細 オプション	「区間自動生成」タブ (Automatically chose number of bin)	階級数をアルゴリズム内部で自動的に選択する。
グループ化 : ___ 階級生成	詳細 オプション	「区間自動生成」タブ (Automatically chose number of bin)	ユーザが階級数を指定する。

表 1-9（続き）階級自動生成のオプション

オプション	Windows 版	IRIX 版	機能説明
次のようなアプローチ	詳細 オプション	「区間自動生成」タブ (Automatically chose number of bin)	階級生成アルゴリズムとして、ユーザが「自動」、「均一な範囲」、「均一な重み付け」のいずれかを選択する。
離散型ラベル	詳細 オプション	N/A	ユーザが分析モデルのラベルを選択する。「自動」を選択した場合は、分析モデルの対象とする離散型ラベルをメニューから選択する。
階級での重み付け下限値	詳細 オプション	N/A	ユーザが 1 つの階級あたりの最小重み（重みが設定されていない場合はレコード数）を指定して、階級数を制限する。

「エントロピー」アルゴリズムを選択して階級の自動生成を行うとき、上限と下限の区間は、個々の階級内部でのラベルの分布ができる限り異なるように設定されます。新しい階級生成の範囲が有意と判定されなくなるまで、区間が計算されて範囲が分割されます。値が異なれば、それだけ多くの階級が選択されます（両者の関係は対数的です）。

「自動」チェックボックスにチェックマークを付けると、全レコードの重み合計に基づいて階級あたりの最小重みが自動的に計算されます。重み合計が大きければ、階級あたりの最小重みも大きくなります（両者の関係は対数的です）。

バックフィッティング

バックフィッティングでは、訓練事例だけに基づいてモデルを作成した後、データセット全体をそのモデルにバックフィッティング（再適用）して予測精度を改善します。バックフィッティングの目的は、確率評価と元のデータセット間の整合性を高めることです。バックフィッティングは大きいモデルの構造を帰納的に作成するよりも高速です。予備 (Hold-out) 法で誤差を評価するときは、データセットの一定の割合がテスト用に残されます。モデルが作成された後、そのモデルの構造にデータセット全体をバックフィットすると、モデルの構造内部のレコード数（カウント）、重み、確率などがデータセット全体を反映した値になるため、最終的な誤差が低減されます。バックフィッティングに関するオプションは、各分析の「詳細設定」ダイアログ・ボックスにあります。

分析ツールによって構築されるモデルは、次の 2 つの部分からなります。

- 構造 ― 決定木と選択式決定木については、ツリーの形状が分析モデルの構造です。エビデンス・クラシファイアについては、各属性の階級数と、（属性が数値型である場合は）区間がモデルの構造です。
- 確率推定 ― モデルの構造の各部分によって各クラスの確率が推定されます。通常、これらの確率推定は、構造内の個々のポイントにある訓練事例のレコード数に基づいて計算されます。決定木については、リーフ部分のレコードの重みに基づいて確率が計算されます。エビデンス・クラシファイアについては、個々の属性値の条件付き確率または一定範囲の属性値の条件付き確率に基づいて確率が計算されます。条件付き確率は特定のラベルの値が現れるという条件のもとで特定の属性値が現れる相対的な確率であり、エビデンス・ビジュアライザの左側のウィンドウ内では矩形グラフによって表されます。

レコードセットに対してモデルをバックフィットしてもモデルの構造は変更されませんが、実際のデータに基づいて確率推定が更新されます。バックフィッティングにはいくつかの利点があります。

1. 小さい訓練事例に基づいてモデルの構造を作成した後、そのモデルの構造に大きいデータセットをバックフィットすると、確率推定の精度を改善することができます。バックフィッティングは、モデルの構造を帰納的に作成する処理よりも高速です。
2. 予備法で誤差を推定するときは、データセットの一定の割合がテスト用に残されます。モデルが作成されて誤差推定が完了した後、そのモデルの構造にデータセット全体をバックフィットすると、最終的な誤差を低減することができます。その理由は、モデルの構造内のレコード数（カウント）、重み、確率などが、訓練事例だけではなくデータセット全体を反映した値になるためです。

ビジュアライザのドリルスルー機能を使用するときは、データの重みが表示されますが、この重みはデータセット全体を反映しています。バックフィッティングを行っていない場合は、訓練事例だけを反映した重みが表示されます。

テストセットをバックフィッティングするときは、「データの可視化 / マイニング」パネルにある「詳細設定」ボタン（Windows システム）または「詳細オプション (Futher options)」ボタン（IRIX システム）をクリックして、「テストセットのバックフィット」チェックボックスにチェックマークを付けます。「データの可視化 / マイニング」パネルで「マイニングツール」タブをクリックし、「クラス判別」タブまたは「回帰」タブを選択します。次に、各分析の実行モードとして「クラシファイアとエラー」または「回帰ツリーとエラー」を選択し、ブースティングが有効になっているときは、バックフィッティングを使用できません。

階級生成

項目の値を階級生成すると、1 つまたは複数の項目の情報をグループにまとめて新しい項目（たとえば 0-18、19-25、26-35 などの年齢グループの値を表す項目）を作成することができます。Tool Manager を使用して項目を階級生成すると、計算時間が短縮され、可視化に適した簡略なデータが作成されます。項目の階級生成の詳細については、『*MineSet 3.0 Enterprise Edition User's Guide for Windows*』の項目の新しい階級の変更または作成に関する説明を参照してください。

階級生成のオプション

Windows 版の「階級生成のオプション」ダイアログ・ボックスには、項目の階級生成方法として、「階級生成を行いません」、「階級自動生成」、「区間の指定」の 3 つのオプションがあります。IRIX 版には、「区間自動生成 (Automatic Thresholds)」および「区間のユーザ指定 (User Specilied Thresholds)」という 2 つのオプションがあります。

階級生成の表記

MineSet 3.1 では、階級生成の範囲について次のような新しい表記が使用されます。

(下限区間 ... 上限区間] (*lower-threshold* ... *upper-threshold*]

下限区間の左側の丸カッコ "(" は、下限区間が範囲に含まれないことを示します。上限区間の右側の角カッコ "]" は、上限区間が範囲に含まれることを示します。たとえば、(10.5 ... 12.6] は 10.5 を越えて（10.5 は含まれない）12.6 以下の範囲を示します。下限区間を省略すると、上限区間以下のすべての値が範囲に含まれます。たとえば、(... 10.5] は 10.5 以下の範囲を示します。上限区間を省略すると、下限区間を越えるすべての値（下限は含まれない）が範囲に含まれます。たとえば、(12.6 ...] は 12.6 を越える範囲を示します。

階級自動生成アルゴリズム

項目を自動的に階級生成するために、次の 3 つのアルゴリズムが用意されています。これらのアルゴリズムは「次のようなアプローチ」オプションで選択することができます。

- エントロピー― このアルゴリズムを使用するときは、分析モデルの対象とする離散型ラベルを「離散型ラベル」オプションで選択する必要があります。上限と下限の区間は、個々の階級内部でのラベルの分割ができる限り異なるように設定されます。新しい階級生成の範囲が有意と判定されなくなるまで、区間が計算されて範囲が分割されます。

  「階級での重み付け下限値」テキストフィールドでは、1 つの階級あたりの最小重みを指定します。ここに指定した値より小さい重みの階級は作成されません。分割によって作成される範囲に含まれるレコードの重み合計が指定の最小値に達しない場合、分割は行われません。デフォルトでは、各レコードの重みは 1 です。その場合、階級あたりの最小重みは、階級あたりの最小レコード数と同じになります。

  ユーザが階級あたりの最小重みを指定する代わりに、最小重みを自動的に算出するような機能も用意されています。「自動」チェックボックスにチェックマークを付けると、全レコードの重み合計に基づいて階級あたりの最小重みが自動的に計算されます。重み合計が大きければ、階級あたりの最小重みも大きくなります（両者の関係は対数的です）。

- 「均一な範囲」― 個々の範囲のサイズが均一になるように区間を設定して、指定された数の階級を作成します。両端（下限と上限）の範囲には、データセット内に実際に存在しない値も含まれます。たとえば、ある属性の値が 3-8 の範囲にあり、4 つの階級を指定した場合、区間は 4.25、5.5、6.75 となります。

  - 4.25 未満
  - 4.25 ～ 5.5
  - 5.5 ～ 6.75
  - 6.75 超

  両端（下限値と上限値）の範囲には、データセット内に実際に存在しない値も含まれることに注意してください。MineSet では、これらの階級名の表記は次のようになります。

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

  - (6.75...]
- 「均一な重み付け」— 個々の範囲の重みが均一になるように区間を設定して、指定された数の階級を作成します。範囲のサイズを均一にする区間が選択される「均一な範囲」とは異なり、「均一な重み付け」ではデータセットを同じ重みのサブセットに分割するような区間が計算されます。デフォルトでは、各レコードの重みは 1 です。その場合、「均一な重み付け」では、ほぼ同数のレコードが含まれる階級が指定された数だけ作成されます。

「均一な範囲」と「均一な重み付け」では、「端数の切り捨て」を指定することができます。このオプションを指定すると、階級が生成される前に、あらかじめ極端な値が排除されます。デフォルトの端数切り捨ては 0.05 です。すなわち、すべての値のうち 5% の極端な値（下限の 2.5% と上限の 2.5%）が排除されます。このオプションは、区間を決めるときに、外れ値の影響を低減する効果があります。

どの階級生成アルゴリズムでも、ユーザが階級の数を明示的に指定するか、アルゴリズム内部で階級の数を自動的に算出することができます。後者の場合、「均一な範囲」と「均一な重み付け」では、個別値の数に基づいて階級の数が決定されます。個別値の数が多いほど、階級の数も多くなります（両者の関係は対数的です）。

デフォルトでは、区間を設定するときに、データセット全体が使用されます。すなわち、区間が決定されるときに（訓練事例だけではなく）テストセットの分布情報も使用されるため、階級生成された属性を後で使用してモデルを作成するときに、モデルの誤差が甘く評価される傾向があります。

「訓練事例のみを使用」オプションを選択すると、区間を決定するときにテストセットの分布情報が除外されるため、クラシファイアの誤差率がより現実的に評価されるようになります。「訓練事例のみを使用」をオンに設定する場合、「予備比率」と「ランダムシード」は、誤差推定用のテストセットの作成時に指定した値と同じにする必要があります（「分析における誤差の取扱い」（109 ページ）を参照）。

「重み付けとして使用」オプションでは、数値型の属性を使用して各レコードの重みを設定することができます。各レコードの重みを変更すると、「エントロピー」と「均一な重み付け」のアルゴリズムが影響を受けますが、「均一な範囲」は影響を受けません。

「適用」ボタンをクリックすると、階級生成の区間が計算されて「選択された項目の区間」フィールドに表示されます。「階級の生成」ウィンドウの一番下にあるテキストフィールドには、階級生成アルゴリズムの進行状況とエラーメッセージが表示されます。

ブースティング

ブースティングは、複数の異なるモデルを作成し、直前の処理で誤ったモデルが生成されたレコードの重みを調整しながら、各モデルを繰返し適用して予測精度を上げるアルゴリズムです。ブースティングを行うと、予測が困難なサンプルが集中的に作成されるため、通常はモデルの精度が改善されます。ブースティングを有効にするには、Tool Manager の「詳細設定」ダイアログ・ボックスにある「ブースト (可視化なし)」チェックボックスにチェックマークを付けます。

モデルを構築するときに、その精度 （低い誤差率）が何よりも重要視される場合があります。たとえば、顧客データを解析して他社に乗換える顧客の特性を調べるときに、解約する可能性が最も高い顧客を正確に予測することだけに集中し、各種要因の可視化は不要であるといったケースが考えられます。このように、クラス判別の正確さが重要視される場合は、ブースティングを使用してください。

ブースティングが適用されたモデルは可視化できませんが、混同マトリックス、改善曲線、学習曲線、ROI 曲線は作成できます。ブースティングは膨大な計算量を要する処理であるため、通常の 25 倍くらいの時間がかかる場合があります。ブースティングでは複数のモデルや訓練事例が特殊な方法で重み付けされるため、ブースティングされたクラシファイアにはバックフィットを適用できません。

ブースティングは任意の数のラベル値を持つデータセットに適用することができます。ブースティングを適用しても、分析モデルの誤差率が常に改善されるとは限りません（特に、3 つ以上のラベル値を持つデータセットではこの傾向が強くなります）。

項目の型または名前の変更

Tool Manager には、項目の型または名前を変更するためのオプションがあります。詳細については、『*MineSet 3.0 Enterprise Edition User's Guide for Windows*』の項目の型または名前の変更に関する説明を参照してください。

選択ポイント（Choice Point）

選択ポイントは、Tool Manager の「マイニングツール」の「クラスタ 」タブから「反復型 k-means 」をクリックすると表示される詳細オプションのひとつです。選択ポイントは、たとえばクラスタ数を決める基準を示唆する値（範囲は 0.0 ～ 1.0）です。選択ポイントに大きい値を指定すると、クラスタ数が多くなる傾向になり、小さい値を指定すると、クラスタ数が少なくなる傾向になります。選択ポイントに 1.0 を指定すると、常に上限値が採用されます。たとえば、クラスタ数の下限が 1、上限が 5 である場合に、選択ポイントを 0.4 と指定すると 2 つのクラスタが選択され、選択ポイントを 0.8 と指定すると 4 つのクラスタが選択される可能性が高くなります。選択ポイントを 1.0 と指定すると、常に 5 つのクラスタが選択されます。

クラシファイア（Classifier）

クラシファイア (*Classifier*) では、データセットに含まれる他の複数の属性に基づいて特定の属性を予測します。クラシファイアはひとつのモデルで、属性はデータセットに含まれる固有の特性です。たとえば、電話会社の顧客に関するデータがある場合は、ボイスメールや国際電話プランの利用状況、電話の利用時間などの情報（属性）をクラス判別モデルによって分析し、特定の顧客が解約する（他社に乗換える）かどうかを予測することができます。予測される属性はラベル (Label) と呼ばれ、予測に使用される属性は記述属性と呼ばれます。

MineSet は訓練事例からクラシファイアを自動的に生成することができます。訓練事例はデータセット内のレコードのうち、ラベルの値が判明しているレコードから構成されます（「訓練事例」（192 ページ）を参照）。たとえば、個々の記述属性（ボイスメールの利用有無、1 日の平均通話時間など）ごとに 1 つの項目と、ラベル（解約の有無）を表す 1 つの項目を持つデータベースのテーブルを訓練事例として使用することができます。訓練事例に基づいてクラシファイアを自動的に構築するアルゴリズムを、分析と呼んでいます。

MineSet ではクラシファイアが生成されると、クラシファイアの構造を視覚的に分かりやすく表現するビジュアルが同時に作成されます。このビジュアルを見ると、データ自身の特性を細かく分析することもできます。クラシファイアが作成されると、そのモデルに基づいて、ラベルの属性が判明していないレコードのクラシファイアを判別することができます。属性値はクラシファイアによって予測されます。

クラシファイアは、次の 2 つの部分からなります。

- 構造 — 決定木と選択式決定木については、ツリーの形状が分析モデルの構造です。エビデンスについては、各属性の階級の数と、（属性が数値型である場合は）区間がクラシファイアの構造です。デシジョン・テーブルの数学的構造は決定木と似ていますが、ビジュアル表示はエビデンス・ビジュアライザと似ています。
- 確率推定 — モデルの構造の各部分によって各クラスの確率が推定されます。通常、これらの確率推定 は、構造内の個々のポイントにある訓練事例のレコード数に基づいて計算されます。決定木については、リーフ部分のレコードの重みに基づいて確率が計算されます。エビデンス・クラシファイアについては、個々の属性値または一定範囲の属性値の条件付き確率に基づいて確率が計算されます。事前確率は、訓練事例からレコードをランダムに抽出した場合に、クラスのラベル（たとえば、「糖尿病」などの特定のクラスのラベル）が観察される確率です（他の属性は無視します）。数学的に、この値は特定のラベルを持つレコードの数をデータセット内のレコードの総数で除した値です。条件付き確率は、特定のラベル（たとえば、「糖尿病」）が現れるという条件のもとで特定の属性値（たとえば、「60 歳以上」）が現れる相対的な確率を示します（すなわち、ラベルの条件付き選択の確率です）。

注記 : クラシファイアに関する詳しい参考文献の一覧や、MineSet のサンプルファイルで使用されているデータセットの謝辞については、『*MineSet 3.0 Enterprise Edition Interface Guide*』の付録 A を参照してください。

クラシファイアの名前（Classifier Name）

生成されるクラシファイアの名前の接頭語は、Tool Manager で指定されるセッション・ファイル名になり、接尾語はモデルのタイプを表す適切な文字列（デシジョン・テーブルモデルの場合は *-dtable.class*、決定木モデルの場合は *-dt.class* など）になります。デフォルトでは、すべてのクラシファイアがサーバ上の *file_cache* ディレクトリ（デフォルトは *mineset_files*）に保存されます。これらのクラシファイアを使用すると、ラベルが判明していないデータセットのラベルを予測することができます（「モデルの適用」（17 ページ））と「バックフィッティング」（37 ページ）を参照）。

「クラシファイア 」タブ

MineSet のさまざまなクラシファイアを使用するには、Tool Manager の「データの可視化 / マイニング」パネルにある「マイニングツール」タブをクリックし、「クラス判別」タブを選択します。「クラス判別」タブでは、各分析の実行モードとして、クラス判別とエラー、クラシファイアのみ、誤差推定、学習曲線の 4 種類を選択することができます。これらの各モードは任意の分析（決定木、選択式決定木、エビデンス、デシジョン・テーブル）と組合わせて使用することができます。各分析の詳細については、このマニュアルの該当のトピックスを参照してください。クラシファイアの使用方法については、『*MineSet 3.0 Enterprise Edition User's Guide for Windows*』を参照してください。

クラスタリング

「マイニングツール」の「クラスタ」タブでは、MineSet のクラスタリング機能を利用することができます。クラスタリングは相関規則の探索と似た説明的なマイニング処理であるため 、特定の項目（属性）をラベルとして指定する必要はありません。また、データセットは訓練事例とテストセット (test set) には分割されず、常にデータセット全体が評価されてクラスタリング・モデルが構築されます。

クラスタリング・モデルは、クラスタごとに 1 つのプロトタイプ・レコード (prototypical record) として保存されます。これらのレコードはクラスタ内の全データの加重平均を表すもので、クラスタ中心（またはクラスタ重心）と呼ばれます。標準的なデータベース・レコードとは違って、クラスタ中心では、各項目の分布が数値型項目の場合はサマリ統計の形式、離散（カテゴリ型）項目の場合はヒストグラムの形式で保持されます。

クラスタリング・ツールを使用するには、Tool Manager の「データの可視化 / マイニング」パネルにある「マイニングツール」タブをクリックし、「クラスタ」を選択します。クラスタリングの目的は、データセット内の類似した特性を決定することです。作成された各クラスタは、さまざまなパラメータを使用して検査することができます。

MineSet のすべてのクラスタリングでは、k-means 目的関数に基づく組合わせアルゴリズムが使用されます。このアルゴリズムでは、類似したレコードをまとめることによってクラスタが形成されます。その際、各クラスタ内部での類似性が最大になるような配慮がなされます。

クラスタリング・ツールを使用するには、Tool Manager のメインウィンドウの「データの可視化 / マイニング」パネルにある「マイニングツール」タブを選択し、表示されるタブから「クラスタ」を選択します。クラスタリングのメインウィンドウが表示されます。

クラスタリングのメインウィンドウにある「実行」ボタンをクリックすると、クラスタリング処理が開始されます。デフォルト設定を使用する場合は、オプションを指定する必要はありません。デフォルト設定では、単一 k-means クラスタリング法に基づいて 3 つのクラスタが作成されます。クラスタリング処理が終了すると、クラスタリングに関する評価情報が表示された後、クラスタ・ビジュアライザが自動的に起動されます。

クラスタリングのメインウィンドウには次のオプションがあります。

- 「メソッド」— クラスタリング法として「単一 k-means」または「反復型 k-means」を選択します（各方法の詳細については下記の説明を参照）。
- 「クラスタ数」— 「単一 k-means」を選択した場合のみ、クラスタの数を指定します。デフォルトは 3 です。
- 「クラスタ数の範囲」— 「反復型 k-means」を選択した場合のみ、クラスタ数の下限と上限を指定します。デフォルトは 1 ... 10 です。
- 「ポイントの選択」— 「反復型 k-means」を選択した場合のみ、選択ポイントを指定します。選択ポイントは最終的なクラスタ数を決めるときの基準を示唆する値（範囲は 0.0 〜 1.0）です（「反復型 k-means クラスタリング法」（47 ページ）を参照）。デフォルトは 0.5 です。

注記 : クラスタリングは計算負荷の高い処理であるため、大量なデータを処理する場合（特に反復型 k-means 法を指定した場合）は、処理時間が長くなる可能性があります。データセットのレコード数が 10,000 件を越える場合は、標本抽出したデータをクラスタリングの操作対象とすることをお勧めします。

単一 k-means クラスタリング法

k-means という用語は、レコード間の類似性に基づいて適正なクラスタリングを決定する目的関数を意味します。単一 k-means 法は MineSet の最も単純なクラスタリング法です。ユーザが任意の数のクラスタを指定すると、各クラスタ内部での散らばりが最小になるように（類似性が最大になるように）データセット内のレコードがグループ化されます。散らばり (Dispersion) はクラスタの凝集性を表す値です。散らばりが大きいほど、各レコードはクラスタ中心から離れてバラバラに分布することになります。数

学的には、散らばりはクラスタ中心と各レコード間の距離の平方二乗平均 (RMS: Root Mean Square) です。

単一 k-means 法のアルゴリズム自体は反復的であり、次のような処理が実行されます。

1. ユーザが任意の数（たとえば 5 つ）のクラスタを指定します。
2. 5 つのクラスタ中心がレコード空間内のランダムな位置に初期設定されます。「ランダムシード」パラメータの値を変更すると、クラスタ中心の初期位置が変わります。
3. データセット内のレコードとクラスタ中心との距離が計算され、その距離が最も小さいクラスタに各レコードが割当てられます。次に、各クラスタ内の新しいレコードに基づいてクラスタ中心が再計算されます。
4. 現在属しているクラスタの中心よりも別のクラスタの中心に近いレコードが存在する場合は、それらのレコードが近い方のクラスタに移動されます。次に、各クラスタ内の新しいレコードに基づいてクラスタ中心が再計算されます。このステップは、改善が見られなくなるまで反復的に実行されます。

このアルゴリズムは有限の反復回数で収束（終了）することが保証されています。

クラスタリングの実行時間は主にステップ 4 によって決まります。したがって、進行状況ウィンドウにはステップ 4 の反復ごとに 1 つの進行バーが表示され、それまでの反復回数が明示されます。アルゴリズムの実行を停止するまでの最大の反復回数（ステップ 4 の実行回数）を指定することができます。デフォルトは 20 です。

各クラスタには 1 から始まる連続番号の名前が付けられます。クラスタ名は数値で表されますが、クラスタ自体に特定の順序付けはありません。

反復型 k-means クラスタリング法

反復型 k-means クラスタリング法は単一 k-means 法を拡張したより複雑なアルゴリズムです。単一 k-means 法とは違って、反復型 K 平均法では正確なクラスタ数を事前に指定する必要はありませんが、クラスタ数の下限、クラスタ数の上限、選択ポイントという 3 つのパラメータを指定する必要があります。下限から上限の間でデータセットに適したクラスタ数が選択されます。選択ポイントは最終的なクラスタ数を決めるときの基準を示唆する値（範囲は 0.0 ～ 1.0）です。選択ポイントに大きい値を指定すると、クラスタ数が多くなる傾向になり、小さい値を指定すると、クラスタ数が少なくなる傾向になります。選択ポイントに 1.0 を指定すると、常に上限値が採用されます。たとえば、クラスタ数の下限が 1、上限が 5 である場合に、選択ポイントを 0.4 と指定すると 2 つのクラスタが選択され、選択ポイントを 0.8 と指定すると 4 つのクラスタが

選択される可能性が高くなります。選択ポイントを 1.0 と指定すると、常に 5 つのクラスタが選択されます。

反復型 k-means 法のクラスタリングを実行するときは、クラスタ数の下限 (デフォルトは 1)、クラスタ数の上限（デフォルトは 10)、選択ポイント（デフォルトは 0.）という 3 つのパラメータを指定します。これらのパラメータに基づいて、次のような処理が実行されます。

1. クラスタ数の下限をクラスタ数とみなして、単一 k-means アルゴリズムが実行されます。これによって初期クラスタリングが作成されます。
2. 散らばりが最も大きいクラスタが検出され、そのクラスタを半分に分割して 2 つの新しいクラスタが作成されます。元のクラスタ内にあったレコードは半分ずつに分けられ、2 つのクラスタに均等に分配されます。次に、2 つの新しいクラスタの中心が計算されます。
3. 現在属しているクラスタの中心よりも別のクラスタの中心に近いレコードが存在する場合は、それらのレコードが近い方のクラスタに移動されます。次に、各クラスタ内の新しいレコードに基づいてクラスタ中心が再計算されます。これは単一 k-means 法のステップ 4 と同じです。単一 k-means 法の場合と同様に、レコードを移動する必要がなくなるか、「最大 # 反復数」パラメータで指定された反復回数（デフォルトは 20 回）に達するまで、このステップが繰り返されます。

   上記のステップ 2 と 3 は、クラスタ数の上限に到達するまで反復的に実行されます。

上記の処理を実行すると、最小クラスタ数から最大クラスタ数の範囲のクラスタリングが作成されます。

ユーザに提示される最終的なクラスタリングは、「ポイントの選択」パラメータ（デフォルトは 0.5）に基づいて次のように決定されます。

各クラスタの散らばりの平均値を計算して、個々のクラスタリングが評価されます。クラスタの数が増えると、平均散らばりは必ず減少しますが、一様には減少しません。適正な散らばりは、「ポイントの選択」パラメータによって決定されます。適正な散らばりとは、次の式によって計算される値です。

[最小クラスタ数のときの散らばり] - [(最小クラスタ数のときの散らばり - 最大クラスタ数のときの散らばり) * 選択ポイント]

この値に最も近い散らばりを持つクラスタリングが最終的なクラスタリングとして選択されます。選択ポイントを 1.0 と指定すると常に最大クラスタ数が採用され、選択ポイントを 0.0 と指定すると常に最小クラスタ数が採用されることに注意してください。

反復型 k-means 法のクラスタ名は、分割過程での起源に基づいて命名されます。最小クラスタ数による最初のクラスタリングでは、単一 k-means 法の場合と同様に連続番号を使用して各クラスタが命名されます。クラスタが分割されるたびに、2 つの新しいクラスタに元のクラスタの名前（連続番号）が付けられますが、その後に "A" または "B" が付加されます。たとえば、"2-B-A" という名前のクラスタは、初期クラスタリングのクラスタ 2 がその後、2 回分割されたことを意味します。

クラスタリング・モデルが構築されると、次のような統計情報がステータス・ウィンドウに表示されます。この例では、「アヤメ (iris)」データセットを使用しています。

```
クラスタリング結果：
----------------------------------------
レコードから重心までの平方二乗平均距離： 0.216 +- 0.0928 RMS 距離：
   Cluster 1: 0.2306 +- 0.09484
   Cluster 2: 0.1921 +- 0.09932
   Cluster 3: 0.2247 +- 0.08058
iris.cluster としてモデルを保存
```

上記の統計値は、個々のクラスタの散らばり (RMS) と、クラスタリング全体のフィット率を表しています。クラスタの数やデータセットの内容が異なる複数のクラスタリングの散らばりを単純に比較することはできません。

クラスタリングのオプション

k-means クラスタリング・アルゴリズムの基本概念は、クラスタ中心と各レコード間の距離を計算することです。レコードの各項目（属性）は、全レコードからなる多次元空間内の別々の次元として取り扱われます。デフォルトでは、レコードの各属性が最終的な距離に与える影響度（重み）は同じです。ただし、「クラスタオプション」ダイアログ・ボックスで「属性の重み」を指定すると、各項目（属性）の影響度を調整することができます。

属性の重み

項目の属性の重みはゼロ以上の値であり、クラスタリング・アルゴリズムの距離計算における各項目（属性）の影響度を左右します。属性の重みを 1 に設定すると平均的な影響度となり、2 に設定すると同じ項目が 2 つあるのと同じ効果が得られます。属性の重みを 0 に設定すると、その項目はクラスタリングに影響を与えません（項目を

データから除去した上でクラスタリングを実行したとみなされます)。属性の重みは整数値である必要はなく、1 未満の重みも許されます。

属性の重みを設定するには、最初にクラスタリングのメインウィンドウにある「詳細設定」ボタン（Windows システム）または「詳細オプション」ボタン（IRIX システム）をクリックします。「クラスタオプション」ダイアログ・ボックスの最上部には現在の属性の重み（デフォルトはすべて 1）が表示されます。

1 つまたは複数の重み値を変更するには、目的の項目を選択して（複数の連続した項目を選択または選択解除する場合は、< Shift > キーを押しながら範囲の最初と最後の項目をクリックします。複数の連続していない項目を選択または選択解除する場合は、< Ctrl > キーを押しながら各項目をクリックします。)、新しい重み値を「選択された重み」フィールドに入力します。

最後に「設定」ボタンをクリックすると、選択した項目の重みが新しい値に変わります。「全て選択」ボタンを使用すると、すべての項目を選択して、それらの重み値を一度に変更することができます。

より分かりやすいクラスタリングを作成するには、属性の重みを調整しなければならない場合があります。属性の重みの設定に関するガイドラインは次の通りです。

- 一般的に、値の個数が多い文字列 (String) 型または列挙 (enum) 型の項目の重みは低い値（必要に応じてゼロ）に設定します。このような項目はクラスタリングを大きく歪ませる可能性があります。
- 高い相関のある複数の項目が見つかった場合は、それらの項目の重み値の合計が 1.0 になるように属性の重みを調整します。そうしないと、相関の高い項目によってクラスタリングが過大な影響を受ける可能性があります。
- 離散型（文字列または列挙型）項目間の距離は大きくなる傾向があるため、通常は実数型項目よりも小さい重み値を離散型項目に割当てるのが適切です。

「クラスタオプション」ダイアログ・ボックス

クラスタリングに関するさまざまなオプションを設定する「クラスタオプション」ダイアログ・ボックスを表示するには、Tool Manager の「マイニングツール」タブをクリックし、表示されるタブから「クラスタ」を選択した後、「詳細設定」ボタン（Windows システム）または「詳細オプション (Futher options)」ボタン（IRIX システム）をクリックします。設定するオプションは、選択したクラスタリング法（単一 k-means 法または反復型 k-means 法）に応じて異なります。

- 属性の重み

  クラスタリングでは、データセット内の個々の属性（項目）に異なる重みを割当てることができます。このオプションの詳細については、「属性の重み」（44 ページ）と「重み付け」（200 ページ）を参照してください。

- 距離の測定法

  クラスタ中心とレコード間の距離を測定する方法を指定します。デフォルトのユークリッド距離 (Euclidean) では、レコードの個々の項目が全レコードからなる多次元空間内の 1 つの次元とみなされ、その多次元空間内のベクトルノルム（各次元に沿った距離の二乗和の平方根）として距離が計算されます。ポップアップ・メニューには別の計算方法としてマンハッタン距離 (Manhattan) が表示されます。マンハッタン距離は、各次元の軸に沿った距離の和として計算されます。マンハッタン距離という名前は、マンハッタンの街路を人が歩く様子に由来しています。各次元の軸（街路）に平行のパスに沿ってしか移動できないため、A 地点から B 地点に直線的に進むことはできません。

- 最大 # 反復数

  クラスタリング・アルゴリズムによってデータセットを走査する回数を制限します。より厳密には、単一 k-means クラスタリング法のステップ 4 を実行する回数を制限します。詳細については、「単一 k-means クラスタリング法」（46 ページ）を参照してください。

- ランダムシード

  ランダムシードを変更すると、クラスタ中心の初期位置が変わります。詳細については、「単一 k-means クラスタリング法」（46 ページ）を参照してください。

- 重み付けとして使用

  MineSet のほとんどのマイニングツールと同様に、クラスタリングではレコードの重み付け（重みの設定）がサポートされます。このオプションでは、データの各レコードの重みを指定する項目（数値型でなければなりません）を選択できます。

- 属性としても使用

  このチェックボックスをオンに設定した場合は、クラスタリング・アルゴリズムにおいて、重みの項目が通常のデータ属性としても使用されます。チェックマークを付けない場合は、重みの項目は通常の属性とはみなされず、ダイアログ・ボックスの「属性の重み」セクションにも表示されません（暗黙的にゼロの重み値が対応付けされます）。

クラスタ・ビジュアライザ

クラスタ・ビジュアライザでは、クラスタリングの分析結果が複数のボックスグラフとヒストグラムで表示されます。クラスタリング処理を一度実行すると、クラスタ・ビジュアライザを使用してクラスタ中心を直接表示することができます。

また、「モデルの適用」機能（「モデルの適用」（17 ページ）参照）を訓練事例に直接適用すると、各レコードにクラスタを割当てることができます。こうすると、MineSet のさまざまなマイニングツールを利用してクラスタを解析できるようになります。

必要なファイル

クラスタ・ビジュアライザには次のファイルが必要です。

- タブで区切られたフィールド行からなるデータファイル

  データファイルは Tool Manager を使用して簡単に作成することができます。データファイルを作成するには、データベースなどのソースからデータを抽出し、クラスタ・ビジュアライザ専用の形式に変換します。データファイルの拡張子はユーザが自由に指定します（なお、クラスタ・ビジュアライザ用のサンプルファイルの拡張子は *.clusterviz.data* になっています）。

- 入力データの形式と表示方法を指定する設定ファイル

  設定ファイルは、Tool Manager を使用して自動作成するか、任意のテキスト・エディタ（jot、vi、Emacs など）を使用して手作業で作成します。

  設定ファイルの拡張子は *.clusterviz* でなければなりません。クラスタ・ビジュアライザを起動するとき、またはファイルを開くときは、データファイルではなく設定ファイルを指定してください。

クラスタ・ビジュアライザの起動

クラスタ・ビジュアライザを起動するには複数の方法があります。

- Tool Manager を使用し、クラスタ・ビジュアライザを設定して実行します。（すべての MineSet ツールで共通に使用される Tool Manager の機能については、「Tool Manager」（191 ページ）を参照してください。Tool Manager を通じてクラスタ・ビジュアライザを操作する方法については、『*MineSet 3.0 Enterprise Edition User's Guide for Windows*』を参照してください。）

- Tool Manager の「可視化ツール」メニューから「クラスタ・ビジュアライザ」を選択し、「ファイル」メニューの「開く」オプションを使用して設定ファイルを開きます。
- 使用する設定ファイルが分かっている場合は、その設定ファイルのアイコンをダブルクリックします。こうするとクラスタ・ビジュアライザが起動され、選択した設定ファイルが自動的に読込まれます。この起動方法を利用できるのは、設定ファイルの拡張子が *.clusterviz* である場合に限られます（Tool Manager を使用してクラスタ・ビジュアライザ用の設定ファイルを作成すると、ファイル拡張子は常に .clusterviz になります）。
- 次のコマンドをプロンプトに入力して、IRIX シェルのコマンド行からクラスタ・ビジュアライザを起動します。

  `clusterviz [`*`configFile`*`]`

  configFile は任意指定の引数で、設定ファイルの名前を表します。コマンド行で設定ファイルを指定しなかった場合は、「ファイル」->「開く」オプションを使用してファイル名を指定する必要があります。

色の選択

MineSet ツールのほとんどのオプション・ウィンドウには、色を選択するためのオプションがあります。各ツールのオプション・ウィンドウを開くには、Tool Manager 内で「詳細設定」ボタン（Windows システム）または「詳細オプション (Futher options)」ボタン（IRIX システム）をクリックします。

カラーブラウザによる色の選択 （Windows システム）

カラーブラウザ (Color Chooser) を表示するには、「ツールオプション」ウィンドウ内または「詳細設定」ウィンドウ内で、カラーリスト内のカラーボックスまたは + 符号をクリックします。複数のカラーボックスから色を選択できる場合は、カラーボックスのリスト（図 1-10 参照）が表示されます（初期状態では空の場合もあります）。

図 1-10 複数のカラーボックス

色のリスト内の + 符号をクリックすると、カラーブラウザ（図 1-11）が表示され、カラーリストに追加する色を選択することができます。「直前」エリアには 1 つ前の色が表示され、「プレビュー」エリアには最新の色が表示されます。気が変わった場合は、「直前」エリアに表示された色に戻すことができます。使用する色を決めて「了解」ボタンをクリックすると、その色が色のリストに追加されます。複数の色を追加するときに、いちいちカラーブラウザを終了する必要はありません。

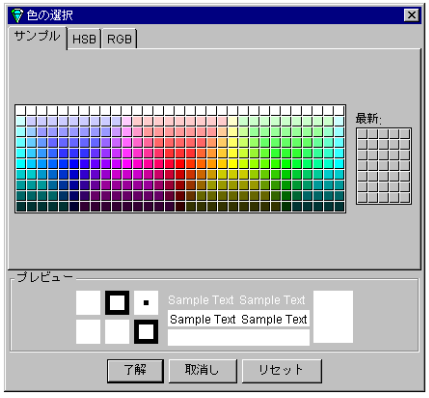

図 1-11 カラーブラウザ (Color Chooser) ダイアログ・ボックス （Windows システム）

カラーブラウザの HSB (Hue, Saturation, Brightness) タブまたは RGB (Red, Green, Blue) タブを使用して、色を選択することもできます。HSB タブで大きい色付きボックスをクリックすると、白いサークルが表示されます。

H ラジオボタンを選択した場合は、ボックス内でサークルをドラッグして、彩度と輝度を調整することができます。色調 (Hue) を調整するには、虹色のバーの隣りにあるスライダを動かします。

S ラジオボタンを選択した場合は、ボックス内でサークルをドラッグして、色調と輝度を調整することができます。彩度 (Saturation) を調整するには、虹色のバーの隣りにあるスライダを動かします。

B ラジオボタンを選択した場合は、ボックス内でサークルをドラッグして、色調と彩度を調整することができます。輝度 (Brightness) を調整するには、虹色のバーの隣りにあるスライダを動かします。

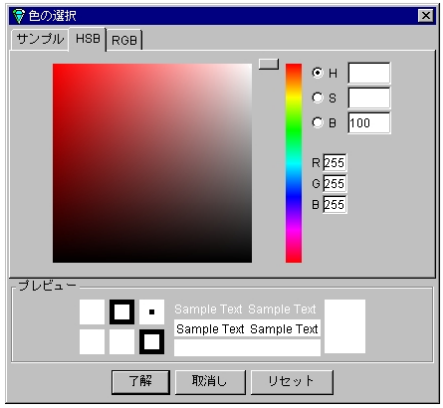

図 1-12 カラーブラウザの HSB タブ（Windows システム）

RGB タブではスライダを使用して、色の 3 要素（R：赤、G：緑、B：黒）の値を個別に調整することができます。

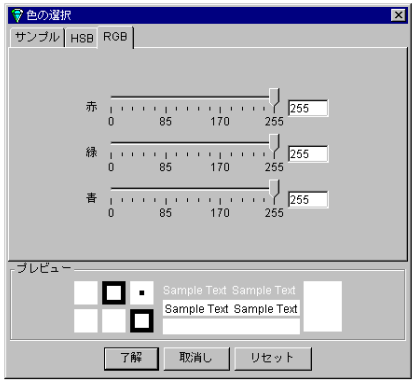

図 1-13 カラーブラウザの RGB タブ（Windows システム）

カラーブラウザによる色の選択 （IRIX システム）

カラーブラウザが使用できるツールについては、そのオプション・ウィンドウにカラーオプションが表示されます。MineSet では、カラーボックスを使用した色のリストチューザが用意されています。ここでは、MineSet 可視化ツール（ビジュアライザ）のウィンドウで使用される色の選択や編集を行う方法を説明します。

選択できる色が 1 つだけの場合は（たとえばグリッドの色）、1 つのカラーボックスが表示されます。

カラーボックスをマウスの左ボタンでクリックすると、カラーブラウザが表示され、そのボックスの色を編集することができます（図 1-15 参照）。

複数のカラーボックスから色を選択できる場合は、カラーボックスのリスト（図 1-14 参照）が表示されます（初期状態では空の場合もあります）。

図 1-14 複数のカラーボックス

色を編集するには、カラーボックスをマウスの左ボタンでクリックします。こうすると、カラーブラウザが起動され、ボタンを押して色を変更できるようになります。カラーボックスをマウスの中ボタンでクリックすると、そのカラーボックスが選択されますが、カラーブラウザは起動されません。

カラーボックスのリストの右側には 4 つのボタンがあります。最初のボタンはプラス符号 (+) が付いた「追加 (+)」ボタンであり、リストの最後に新しい色を追加するときに使用します。このボタンをクリックすると、新しいカラーボックスがリストの最後に追加され、カラーブラウザが表示されて、そのカラーボックスの色を選択できるようになります。リストに表示されている色が最大数に達している場合、「追加 (+)」ボタンは使用できません。

「追加 (+)」ボタンの隣りには、マイナス符号 (-) が付いた「削除 (-)」ボタンがあります。このボタンは選択した色を削除するときに使用します。カラーボックスが選択されていないとき、またはリストに最小限の色しか表示されていないとき、このボタンは使用できません。

「削除 (-) 」ボタンの隣りには、選択した色を左右に移動する 2 つのボタンがあります。カラーボックスが選択されていないとき、または選択したボックスがリストの端にあって移動できないとき、これらのボタンは使用できません。

色のリストに多くの色が配置されて一部が見えないときは、リストの両端にスクロール用の矢印が表示されます。ハードウェアの制限で色を表示できないときは、カラーボックスの代わりに、色を 16 進表記で示すテキストラベルが表示されます。

カラーボックスをマウスの左ボタンでクリックするか、またはビジュアライザの「設定オプション (Configuration Options)」ウィンドウの「色 (Colors)」パネルにある「追加 (Add)」ボタンをクリックすると、カラーブラウザが表示されます（図 1-15）。

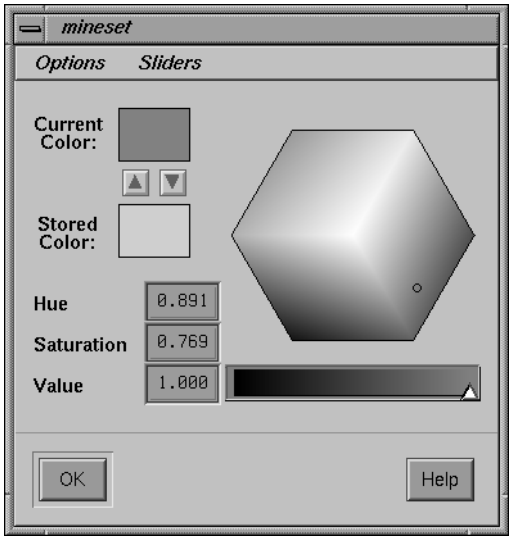

図 1-15　　カラーブラウザ（IRIX システム）

カラーブラウザを使用して色を選択する手順を次に示します。

1. 色付きの六角形の中にある白くて小さい円の上にマウスのカーソルを移動します。
2. マウスの左ボタンを押して、六角形の中でマウスのカーソルを動かします。「現在の色 (*Current Color*)」ラベルの右側にある長方形の領域に、小さい円の下にある色が表示されます。色を選択するときは、この長方形の領域がカラーパレットの役割を果たします。
3. 小さい円が希望の色の上に来たらマウスのボタンを離します。ボタンを離すとすぐにその色がカラーボックスに表示されます。

   複数の色を選択する場合も、いちいちカラーブラウザを閉じる必要はありません。別のカラーボックスをクリックすれば、すでに開いているカラーブラウザでその色を編集することができます。

色を選択したら、「了解 (*OK*)」ボタンをクリックし、カラーブラウザのウィンドウを閉じます。

「重要項目」タブ

Tool Manager の「重要項目」タブを選択すると、重要項目機能を使用することができます。下記の「重要項目」を参照してください。

重要項目

重要項目を使用するには、Tool Manager の「データの可視化 / マイニング」パネルにある「マイニングツール」タブをクリックし、表示されるタブから「重要項目」を選択します。このツールを使用すると、ラベルの値の推定に関して最も重要と考えられる項目をデータセットから簡単に検出することができます。

クラスタリング（「クラスタリング」（45 ページ）参照）などの他のデータマイニング・ツールと重要項目の大きな違いとして、重要項目機能では、各項目の寄与率の判定に使用するラベルをユーザ自身が指定することが挙げられます。これに対して、クラスタリングでは、データ自身を解析することによって、データの分析モデルに有効な属性が示されます。MineSet では個々のクラス判別モデルごとに別々のアルゴリズムが採用されているため、重要と判定される項目（属性）も異なる場合があります。重要項目機能の適用が有効と思われる事例とそのサンプルファイルの説明については、付録 A「設定ファイルとデータファイルのサンプルファイル」を参照してください。

重要項目の検出

たとえば、「良好な信用リスク (*good credit risk*)」というラベル値を推定するうえで重要な 3 つの項目を 重要項目によって検出し、スキャタ・ビジュアライザの 3 つの軸に対応付ける場合を考えてみます。ユーザがこのラベルを指定して「実行」ボタンをクリックすると、ポップアップ・ウィンドウが表示され、3 つの重要な項目が示されます。指定したラベルに対する各項目の重要度は、「寄与率」（0 から 100 までの数値）という尺度によって表されます。項目の数を増やすと、寄与率が単調に増加します。

寄与率はラベル値の分布の偏りを表す尺度です。累積寄与率はデータを分割するときの寄与率を表す尺度です。決定木の分割方法と同様に、データは重要と判断された項目に基づいて複数のパーティションに分割されます。パーティション内の各セットには独自の寄与率尺度があり、各パーティションの寄与率尺度は各セットの寄与率尺度の組合わせになります。パーティション内にある特定のセットの寄与率は、すべてのレコードが別々のクラスに属していれば 0 になり、各レコードが同じクラスに属していれば 100 になります。同様に、パーティションの累積寄与率は、パーティション内

の各セットの全レコードが別々のクラスに属していれば 0 になり、パーティション内の各セットのレコードがすべて同じクラスに属していれば 100 になります。

重要項目には次の 2 つのモードがあります。

- 標準モード

  標準モードを開始するには、ポップアップ・メニューから離散型ラベルを選択し、表示する項目の数を指定して、「実行」ボタンをクリックします。

- 詳細モード

  詳細モードでは、項目の選択過程を細かく制御することができます。詳細モードを開始するには、「重要項目」ウィンドウで「詳細モード」をクリックします。表示されるダイアログ・ボックスでは、重みとして使用する項目（属性）を選択し、その項目を通常のデータ属性として分析モデルで使用するかどうかを指定します。ダイアログ・ボックスには項目名のリストが 2 つ表示されています。左側のリストには選択可能な属性が表示され、右側のリストには選択された属性（ユーザが選択した属性または主成分抽出アルゴリズムによって選択された属性）が表示されます。

  詳細モードには 2 つの機能があります。1 つは新しい重要属性を検出することであり、もう 1 つは属性をランキングすることです。

  – 複数の重要属性の検出

    このサブモードを開始するには、ダイアログ・ボックスの中央にある 2 つのラジオボタンのうち最初のボタン（「... 検索 : ［数字］重要項目を追加」）をクリックします。そのまま「実行 (*Go*)」ボタンをクリックすると、標準モードを選択した場合と同じ結果になります（すなわち、指定した数の重要項目が検出されて、それらの項目が右側のリストに自動的に移動されます）。各項目の近くには、累積寄与率（検出された項目を含む全項目の寄与率）が表示されます。

    ユーザが特定の項目を事前に選択した上で、残りの項目をツールによって自動選択する場合は、左側のリストから右側のリストに項目を移動しておきます。たとえば、「シリンダー (*cylinders*)」項目を事前に選択した上で、さらに 3 つの項目をツールによって自動選択する場合は、「シリンダー (*cylinders*)」項目をクリックした後、左右のリストの間にある右向き矢印をクリックします。

    「実行」ボタンをクリックすると、リスト内の以前の寄与率とともに、各項目の累積寄与率が表示されます。寄与率が 100 である場合は、選択された項目を使用してデータセット内のラベルの値を完全に推定することができます。

– 属性のランキング

詳細モードでは、重要項目として検出された項目（右側のリストに入っている項目）の寄与率に対して、左側の各項目が与える向上度を評価することもできます。たとえば、右側のリストに「シリンダー (*cylinders*)」項目を移動した後、左側に残っている各項目によって右側の項目の寄与率がどれくらい改善されるかを計算することができます。累積寄与率は右側の項目（すでに重要項目として検出された項目）について計算されます。

このサブモードを開始するには、ダイアログ・ボックスの下にある 2 つのラジオボタンのうち 2 番目のボタン（「... 左側項目リストに寄与率向上度、右側項目リストに累積寄与率を表示」）を選択します。このサブモードでは、重要項目の選択過程を細かく調整することができます。たとえば、ランキングがほとんど同じ項目が 2 つ存在する場合、（収集にコストがかからない、信頼度が高い、分かりやすいなどを基準にして）どちらか一方の項目だけを選択することができます。

項目の寄与率は、他の重要な項目と併用すると変化する場合があります。たとえば、「純収入 (*net-income*)」はそれ自身として重要な項目ですが、「給与 (*salary*)」と併用するとそれほど重要でなくなります。これは両者の相関関係が強いためです。ランキングの高い順に 3 つの項目を組合わせても最適の属性セットが形成されるとは限りません。ドル建ての収入と別の通貨建ての収入のランキングは同じですが、どちらか一方がすでに選択されている場合、他方を選択しても全体としての説明力は増加しません。

重要項目機能は、スキャタ・ビジュアライザまたはスプラット・ビジュアライザで使用する 3 つの軸を見つけるときに役立ちます。また、ツリー・ビジュアライザでキーとして使用するラベルを選択するときに、説明変数に適した項目階層（ラベル値を分割する階層）を見つけるときにも役立ちます。

単精度 (float) または倍精度 (double) の 浮動小数点値は、自動離散化アルゴリズムを通じて事前に離散化されます。値を持たない項目が左側のリストに表示される場合、その項目は離散化アルゴリズムから除外されたものです。このような項目は（たとえば、1 つの範囲に階級生成されて）単一の値になってしまったか、分割後のレコード数が統計的に有意な値に達しなかったものです。

重要項目と他のクラシファイアとの相違

ここでは、重要項目による項目の選択、エビデンス分析とデシジョン・テーブル分析による寄与率のランキング、およびデシジョン・テーブル分析による分割の相違について説明します。重要項目ではデータセット内の全データが使用されるため、下記の説明では各分析が「クラシファイアのみ」モードで実行され、すべてのデータが使用されることを前提としています。

離散化プロセス

重要項目とエビデンス分析では、自動離散化アルゴリズム（Tool Manager の階級自動生成で使用されるのと同じアルゴリズム）に基づいてすべての連続属性（項目）が離散属性に変換されます。決定木のアルゴリズムでは属性（項目）が事前に離散化されず、ツリーの構築時に区間が検出されます。

自動離散化の主な利点は、連続範囲が複数の間隔に同時に離散化されることです。これに対して、決定木では二者択一の分割しか行われません。

決定木のアルゴリズムの主な利点は、データのサブセットが複数の離散的なチャンク（検定が実行される位置にある特定のノードに到達したレコード集合）に分割されることです。すなわち、これらのチャンク（レコード集合）には、「グローバル」な離散化ではなく「ローカル」な離散化が適用されます。

寄与率のランキング

エビデンス分析と重要項目では、相互情報量を 寄与率の尺度として属性がランキングされます。一方、決定木分析、選択式決定木分析、デシジョン・テーブル 分析では、正規化された相互情報量がデフォルトの寄与率の尺度に設定されているため、多岐の分割にはペナルティが課せられます。したがって、決定木分析では、値の数が多い属性よりも値の数が少ない属性のほうが重要と判断される傾向があります。決定木のデフォルト設定は「相互情報」に変更することができます。

他の属性との依存関係

エビデンス分析では、個々の属性が相互に独立にランキングされます。複数の属性の相関が強い場合、各属性のランキングはほぼ同じになります。重要項目機能の「詳細」モードで、重要な属性を選択しないで（属性を右側のリストに移動しないで）、「... 左側項目リストに寄与率向上度、右側項目リストに累積寄与率を表示」オプションを使用すると、属性のランキングがエビデンス分析によるソート順序と同じになります。

重要項目、決定木分析、デシジョン・テーブル分析には、エビデンス分析よりも強力な重要項目機能があります。これらの分析機能では、他の項目（属性）との依存関係に基づいて寄与率のランキングが行われます。

重要項目では、右側のリストに表示された一連の項目と比較して各項目の寄与率が判断されます。相関関係が非常に強い 2 つの項目が存在する場合、どちらか一方が選択されると、他方は選択されなくなります。ラベルの推定に関して、すでに選択されている項目よりも多くの情報を提供できる項目だけが選択されます。

デシジョン・テーブル分析では、「提唱」ボタンが重要項目機能と似た機能を備えています。このボタンをクリックすると、すでにマッピング・ボックスに入っている一連の項目と比較して各項目の寄与率が判断されます。ただし、デシジョン・テーブル分析の提唱モードと重要項目アルゴリズムとの間には重要な相違が 3 つあります。第一に、デシジョン・テーブル分析では多岐の分割にペナルティが課されます。第二に、デシジョン・テーブル分析では、より徹底した探索を実行して、項目の寄与率のランキングが決定されます。第三に、デシジョン・テーブル分析では寄与率の値が表示されず、項目のランキングだけが行われます。

決定木分析の重要項目機能はさらに柔軟であり、個々のサブツリーごとに異なる項目を選択することができます。たとえば、ルートの左側の子ノードと右側の子ノードで別々の項目を選択することができます。この機能は決定木には適していますが、スキャタ・ビジュアライザやスプラット・ビジュアライザに表示する少数の項目を選択する場合には適していません。後者の場合には、重要項目機能の方が適しています。その理由は、ツリーの各レベルで全ノードの同じ項目が検定される " 自明な " 決定木が作成されるためです。重要項目では、以前に選択された項目のすべての組合わせを評価するときに、単一の項目を選択する必要があります。

項目

「項目の追加」（1 ページ）、「「項目の削除」ボタン」（157 ページ）、「項目名のソート」（169 ページ）を参照してください。

コマンド行の操作

MineSet ソフトウェア本体と各ビジュアライザは、そのツール名をプロンプトに入力して、コマンド行から起動することができます。

Windows システム上での操作 :

```
Viz [configFile]
```

IRIX システム上での操作 :

```
scatterviz [configFile]
```

設定ファイル (*configFile*) は指定しなくてもかまいません。ただし、設定ファイルを指定しなかった場合は、ツールの起動後に「ファイル」メニューと「ヘルプ」メニューしか使用できません。その場合は、「ファイル」->「開く」オプションを選択して設定ファイル名を指定してください。

IRIX システム上では、さまざまなツールでコマンド（clusterviz, eviviz, scatterviz, splatviz, statviz, treeviz, mapviz）が使用されます。相関規則分析では、2 つの部分からなるコマンドが使用されます（「相関規則」（23 ページ）を参照）。

設定ファイル

MineSet の各ツールを使用するには、次の 2 つのファイルが必要です。

- タブで区切られたフィールド行からなるデータファイル（拡張子は *.data*）
- 入力データの形式と表示方法を指定する設定ファイル

  設定ファイルの拡張子は、そのツール名の略称（.eviviz、.scatterviz など）になります。

Tool Manager を使用して分析モデルのオプションやパラメータを指定するか、使用するビジュアライザを設定すると、各ツールに適した設定ファイルが自動的に作成されます。また、テキストエディタを使用して、各ツールの設定ファイルを手作業で作成することもできます。たとえば、ワードパッド (Word Pad) などのテキスト・エディタを開き、「ファイルの種類」として「すべてのファイル (*.*)」を選択すると、スキーマファイル (*.schema*) やデータファイル (*.data*) などの任意のファイルを編集することができます。各ファイルの詳細と作成例については、『*MineSet 3.0 Enterprise Edition Interface Guide*』を参照してください。

混同マトリックス

混同マトリックスを使用すると、クラシファイア (Classifier) を通じて発生する誤差の原因を詳しく解析することができます。混同マトリックスでは、正しい予測と誤った予測の数が単に表示されるだけではなく、発生した誤差のタイプが細かく分析されます。混同マトリックスは、すべてのクラシファイアで使用できる詳細オプションです。混同マトリックスを使用するときは、Tool Manager の「データの可視化 / マイニング」パネルにある「マイニングツール」タブをクリックし、表示されるタブから「クラス判別」を選択します。図 1-16 に、アヤメ (Iris) データセットから帰納された決定木に関する混同マトリックスを示します。

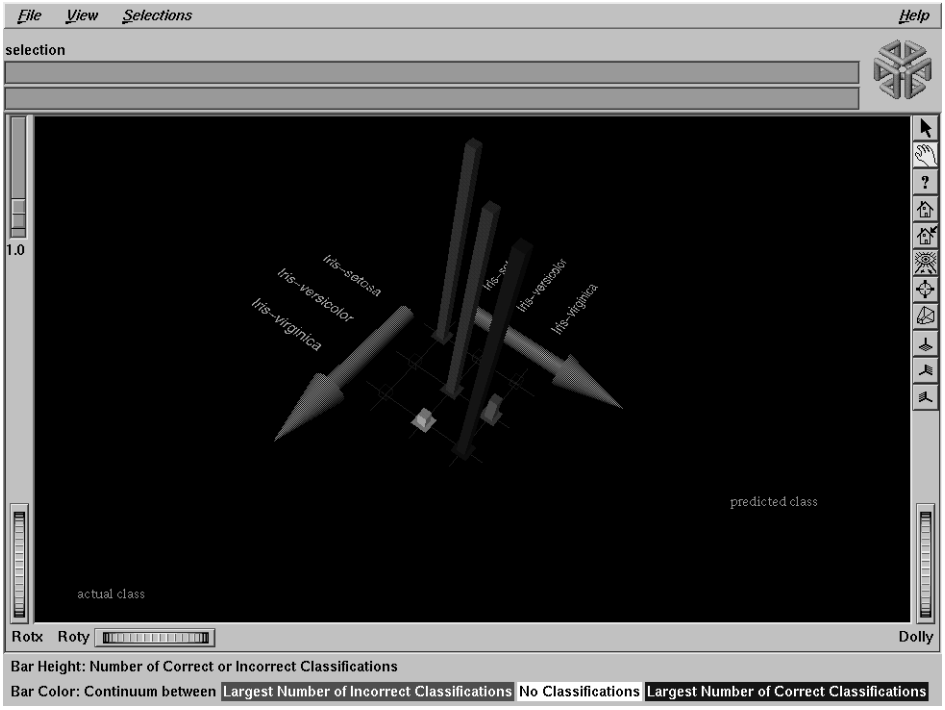

図 1-16 アヤメ (Iris) データセットに関する混同マトリックス

混同マトリックスの 2 つの軸は次の値を表します。

- クラシファイアによって予測されたクラス値
- テストセット（予備セット）内に存在する実際のクラス値

対角線上の値は正しい予測を表し、対角線以外の値は誤った予測を表します。この例では、*iris-versicolor* と *iris-virginica* が頻繁に誤った予測が生成されている一方、*iris-setosa* は常に正しい予測が生成されていることが分かります。

誤差のタイプごとにコスト（損失）が異なる場合は、損失マトリックスを使用して、各タイプの誤差の影響を推定してください（「損失マトリックス」（121 ページ）を参照）。

注記： 混同マトリックスではテストセットに対する誤差が分析されるため、元のデータの分布が大きく変わらなければ、実際のデータにおける誤差の分布を的確に予想することができます。MineSet の混同マトリックスはバックフィッティングの実行前に計算されるため、バックフィッティングの適用前でも適用後でも結果は変わりません（「バックフィッティング」（37 ページ）も参照してください）。

生成コストの考慮

生成コストの考慮は、CART (Classification And Regression Trees) で開発された高度なツリー枝刈り法です。この枝刈り法を使用するには、Tool Manager の「マイニングツール」タブをクリックし、「クラス判別」（または「回帰」）タブから「決定木」を選択した後、「詳細設定」ボタンをクリックし、表示されるダイアログ・ボックス内で「枝刈りオプション」を設定します。

生成コストの考慮による枝刈り法では、ツリーの誤差率（コスト）とツリー内の枝の数（複雑性）のトレードオフ関係を調整することによって、最適なサイズのツリーを生成します。コストと複雑性のバランスを調整するために、訓練事例は学習事例と枝刈りセットに分割されます。最初に、学習事例に基づいてツリー全体が構築された後、そのツリーが枝刈りされて、それほど複雑でない複数のツリーが作成されます。次に、枝刈りセットに基づいて、複数のツリーの中から最もコストの低いツリーが選ばれ、そのツリーのサイズが記録されます。最後に、学習事例と枝刈りセットを再結合されてツリーが構築され、そのツリーが最低コストのツリーのサイズに切り詰められます。

この枝刈り法のパラメータ（枝刈り係数）を指定すると、最低コストの（誤差率が最も低い）ツリーよりもサイズの小さいツリーを選択することができます。この枝刈り係数は、一定範囲の標準誤差（コスト = 誤差率の増加）を許容して、ツリーの複雑性を抑えるものです。枝刈り係数をゼロに設定すると、最低コストのツリーがそのまま選択されます。枝刈り係数を 0.5 に設定すると、誤差率が（最低コストのツリーの誤差率 + 標準誤差 * 0.5）未満であるツリーのうち、最もサイズの小さいツリーが選択されます。枝刈り係数を 1.0（デフォルト値）に設定すると、誤差率が（最低コストのツリーの誤差率 + 標準誤差 * 1.0）未満であるツリーのうち、最も小さいサイズのツリー

が選択されます。枝刈り係数を大きくすると、除去されるサブツリーの数が増えツリーのサイズは小さくなります。データにノイズ（誤差または異常値）が含まれている場合は、枝刈り係数を増加させて小さいツリーを作成してください。ツリーが切り詰められて 1 つのノードになってしまう場合は、枝刈り係数を減少させて大きいツリーを作成してください。

枝刈りによるツリーの簡素化では、ツリー全体が構築されてからサブツリーが除去されるため、最大レベル数または分割の下限値による簡素化よりも実行時間は長くなります。ただし、サブツリーが選択的に除去されるため、モデルの誤差率が減少します。

相互検証

相互検証は、クラシファイアの誤差を評価する 1 つの方法です。この方法では、データセットを任意の数 (*k*) の層（通常 K=10）に分割し、同じ数のクラシファイア (Classifier) を作成します。このプロセスは誤差率の評価を何回も繰返し、推定値の信頼度を向上させます。分析は訓練事例として使用され、*k* 回繰り返されます。データセット全体から 1 つの別の層を差し引いて訓練事例を適用し、それを予備層で検定します。

相互検証は、誤差推定オプションの 1 つとして任意の分析で使用することができます。相互検証用オプションのダイアログ・ボックスでは、相互検証の層数と繰返し回数を設定することができます。また、ランダム・シードを別の数に設定したり、同じ数に維持して同じポイントからのデータを常にカットすることもできます。

データ・クリーニング（Data Cleaning）

データはさまざまな場所に散在していたり、不明瞭な形式や既存のデータと互換性のない形式で保存されたりしている場合があります。また、一部のフィールド（項目）が欠けていたり、無効な値が格納されている可能性もあります。このような複雑性を整理・整頓する作業は「データ・クリーニング (Data Cleaning)」と呼ばれ、通常、データマイニングを始める前に必須の準備作業になります。データ・クリーニングにはサードパーティが提供する各種のデータ抽出ツールを利用することができます。

他のデータ形式から MineSet のデータ形式にデータを変換し、基本的なデータ・クリーニング作業を実行するには、Tool Manager の「ファイル」メニューから「データのインポート」を選択します。詳細については、『*MineSet 3.0 Enterprise Edition Interface Guide*』の「Importing Data Files」を参照してください。

「データの可視化 / マイニング」パネル

Tool Manager の「データの可視化 / マイニング」パネルには、最上位レベルのタブとして、「可視化ツール」、「マイニングツール」、「データファイル」という 3 つのタブがあります。「可視化ツール」タブでは、データをグラフィカルに表示して相互関係を分析するための各種ビジュアライザを選択することができます。「マイニングツール」タブでは、データに基づくモデルとビジュアルを作成して、データの詳細解析や予測を行うことができます。「データファイル」タブには、解析したデータをファイルとして保存するためのオプションがあります。

「データファイル」タブ

Tool Manager の「データの可視化 / マイニング」パネルにある「データファイル」タブを使用すると、操作したデータをクライアント上またはサーバ上のデータファイルとして保存し、将来の分析で再利用することができます。「データファイル」タブをクリックすると、データファイルの保存先（サーバまたはクライアント）を指定するためのトグルボタンを持つウィンドウが表示されます。「クライアント」チェックボックスの隣りには、指定したクライアント・ファイルの名前が表示されます。「クライアント」を選択して「新しいクライアント ファイルを選択」ボタンをクリックすると、クライアント・ファイルの名前を選択するダイアログ・ボックスが表示されます。「サーバ」を選択した場合には、サーバのファイル名を隣りのテキストフィールドに直接入力することができます。

注記： サーバのファイルについてはパス名を指定できません。サーバのファイルはすべてデータムーバのキャッシュ・ディレクトリに保存されます。

データのインポート

MineSet にはデータのインポート機能が用意されています。インポート機能を利用するには、Tool Manager の「ファイル」メニューから「データのインポート」を選択します。表示される「データのインポート 」ウィンドウでは、インポートするファイル名のほかに、インポートに関するさまざまなオプションを指定することができます。各オプションの説明やインポートできるファイル・フォーマットの一覧については、『*MineSet 3.0 Enterprise Edition Interface Guide*』を参照してください。

「データ変換」パネル

Tool Manager のメインウィンドウの左側にある「データ変換」パネルを使用すると、データセット内のテーブルに対してさまざまな操作を実行することができます。「ファイル」メニューを使用してテーブルを選択すると、「データ変換」パネルの 「現在のデータセットの項目名」ウィンドウに、選択したテーブルの項目の見出しが表示されます。このパネルには、次のような機能を持つオプションボタンがあります。

- 「項目の削除」— 可視化やマイニングに関係ない項目を削除します。
- 「階級の生成」— 各レコードを一定の範囲（階級）の項目値からなるグループに割当てます。たとえば、年齢の項目を (0...18]、(19...25]、(26...35] などのグループに階級生成することができます。"(" は下限が範囲に含まれないことを示し、"]" は上限が範囲に含まれることを示します。「階級生成も参照してください。
- 「集計処理」— 選択した項目の合計値、平均値、レコード数、最大値、最小値を表す新しい項目を作成します。また、他の階級生成済み項目をインデックスとして、特定項目の値を要素とする配列を作成することもできます。通常、集計処理を実行すると、元のテーブルにある複数の行が集計処理されて新しいテーブルの行が作成されるため、集計処理後のテーブルは元のテーブルよりも行数が少なくなります（「集計処理」（4 ページ）を参照）。
- 「フィルタ」— 項目値を含む条件式に基づいてデータのサブセットを選択します。たとえば、年齢が 20 歳未満のレコードだけを抽出することができます（「フィルタリング」を参照）。
- 「データ型の変更」— 項目の名前とデータ型を変更します（「データ型の変更」を参照）。
- 「項目の追加」— 数学式に基づいて新しい項目を追加します（「項目の追加」を参照）。たとえば、「年齢 (age)」項目を使用し、"if age is less than or equal to 18 then minor becomes 1; else minor becomes 0"（年齢が 18 歳以下の場合は minor = 1、それ

以外の場合は minor = 0）という表現を使用して「未成年 (minor)」という項目を追加することができます。

- 「モデルの適用」— 作成済みのクラシファイアを使用して、新しいレコードのラベル値の予測、ラベル値の確率の評価、新しいデータを使用したクラシファイアのテスト、既存のクラシファイアに対するデータのバックフィットを行います（「モデルの適用」を参照）。
- 「標本」— データのサブセットをランダムに選択します。膨大なデータセットを操作する場合に便利です（「標本抽出」を参照）。
- 「項目のソート」— 項目をアルファベット順にソートします。このオプションは、Windows システム上ではチェックボックスとして表示されます。
- 「プラグインの操作」— このオプションはプラグイン API が利用できるときだけ表示されます。このオプションを使用すると、MineSet の固有機能と同じ方法でプラグイン API を操作することができます。詳細については、MineSet のホームページ (http://mineset.sgi.com) と『*MineSet 3.0 Enterprise Edition Interface Guide*』の「MineSet Plug-In Capability」を参照してください。

デシジョン・テーブル（Decision Table）

デシジョン・テーブルは決定木と似た階層構造のモデルですが、各レベルのデータを分割するときに単一の属性ではなく 2 つの属性が使用されます。デシジョン・テーブル分析では、データのクラス判別にとって最も重要な属性（項目）が確定された後、デシジョン・テーブル・ビジュアライザによってそれらの項目がグラフィカルに表示されます。各項目はネスト（入れ子）構造の表形式になったケーキグラフとして表示され、次に重要な属性（項目）を表す小さいケーキに分割されてケーキグラフとして表示されます。すなわち、階層を順番に展開するに従って、表示される各レベルの属性の寄与率が減少していきます。デシジョン・テーブル・ビジュアライザの詳細については、『*MineSet 3.0 Enterprise Edition User's Guide for Windows*』を参照してください。

デシジョン・テーブルで使用されるクラス判別法は、決定木と似ています。すなわち、クラス判別法を行うときは、標本（レコード）が抽出される領域において大多数を占める（優勢な）クラスがラベルとして選択されます。学習データが存在しない領域にレコードが属する場合は、テーブルの階層内で 1 つ上のレベルにある優勢なクラスがラベルとしてクラス判別されます。

デシジョン・テーブルモデルの生成

デシジョン・テーブル・クラシファイアはデータに基づいて自動的に生成されます。データは複数のレコードと各レコードに対応するラベルから構成されます（「分析」（107 ページ）を参照）。

デシジョン・テーブル・クラシファイアを自動的に生成するときは、レコードの数（より一般的にはレコードの重み）に基づき、次の手順に従ってテーブル階層内の各ノードの確率が計算されます。ビジュアルでは、これらの各ノードにおけるレコード分布がケーキグラフによって表されます。

連続的な属性は、複数の離散的な範囲に階級生成されます。その際、各範囲でのクラス分布ができるだけ異なるように考慮されます。範囲の数は自動的に決定されます。Tool Manager を使用して手動で属性を階級生成すれば、自動的な階級生成を変更することができます。

軸（属性）のペアの 1 つに沿って 1 行に表示されるケーキの数は、分析 (Inducer) によって生成される離散範囲の数に一致します。範囲が 1 つしか存在しない場合、その属性はラベルの予測に効果がなかったことを意味します。最初は、ラベルの事前確率が右側のラベル確率ウィンドウに表示されます。

デシジョン・テーブル階層内の各レベルにおいて X 軸と Y 軸に属性を割り当てるには、手動、自動、項目の自動選択の 3 つの方法があります。

デシジョン・テーブル・ビジュアライザの起動

デシジョン・テーブル・ビジュアライザを起動するにはいくつかの方法があります。

1. Tool Manager の「クラス判別」タブからデシジョン・テーブル分析を実行します。デシジョン・テーブル分析でクラシファイアが構築されると、デシジョン・テーブル・ビジュアライザが自動的に起動されます。
2. Tool Manager を使用して、「可視化ツール」メニューからデシジョン・テーブル・ビジュアライザを起動し、*.dtableviz* ファイルを開きます。すべての MineSet ツールで共通に使用される Tool Manager の機能については、「Tool Manager」（191 ページ）を参照してください。

3. 使用する設定ファイルが分かっている場合は、その設定ファイルのアイコンをダブルクリックします。こうするとデシジョン・テーブル・ビジュアライザが起動され、選択した設定ファイルが自動的に読込まれます。この起動方法を利用できるのは、設定ファイルの拡張子が .dtableviz である場合に限られます（Tool Manager を使用してデシジョン・テーブル・ビジュアライザ用の設定ファイルを作成すると、ファイル拡張子は常に *.dtableviz* になります）。
4. 次のコマンドをプロンプトに入力して、IRIX シェルのコマンド行からデシジョン・テーブル・ビジュアライザを起動します。

   dtableviz ［*filename.dtableviz*］

   configFile は任意指定の引数であり、設定ファイルの名前を表します。コマンド行で設定ファイルを指定しなかった場合は、「ファイル」->「開く」オプションを選択してファイル名を指定する必要があります。

離散型ラベル

「離散型ラベル」メニューには、データセット内で離散的な値を取る項目のリストが表示されます（「離散型ラベル」の隣りにある矢印をクリックしてください）。離散的な属性（階級生成値、文字列、数個の整数など）は限られた数の値しか取りません。ラベルとして使用する属性については、できる限り少ない数（理想的には 2 ～ 3 個）の値を取る属性を選択してください（「訓練事例を参照）。離散的な属性が存在しない場合は、このメニューに「離散型のラベルがありません」と表示され、「実行」ボタンは使用できない状態になります。その場合は、Tool Manager の「データ変換」パネルを使用して新しい項目を追加するか既存の項目を階級生成して、離散的な属性を新たに作成する必要があります。

軸に対する項目の対応付けによるデータの解析

項目（属性）を X 軸または Y 軸に対応付けると、データの属性間の関係を解析することができます。Tool Manager の「データ変換」ウィンドウにある「現在のデータセットの項目名」パネルで項目の名前を選択した後、次の手順を実行してください。

- Windows システム上では、「X リスト」内または「Y リスト」内のセルをクリックし、表示されるプルダウン・メニューから適切な項目を選択します。
- IRIX システム上では、「X 軸」ウィンドウ内または「Y 軸 (Y-axis)」ウィンドウ内で項目を選択します。対応付けられる最初の 2 つの項目は最上位レベルに表示され、3 番目以降の項目は 2 番目以降のレベルに表示されます。

奇数個の項目（属性）を対応付けると、一番下のレベルには最後の（奇数番目の）属性値を表すケーキが 1 つだけ表示されます。2 つの属性間に相関関係があると思われる場合は、1 つの属性を X 軸に対応付けし、もう 1 つの属性を Y 軸に対応付けます。

ビジュアルでは、ケーキグラフ全体の縦横比を適正に維持するために、特定のレベル上で X 軸と Y 軸の対応付けが交換される場合がありますが、各属性が対応付けされたレベルと異なるレベルに属性が移動されることはありません。

どの属性をどの軸に対応付ければよいか分からない場合は、「提唱」チェックボックスをオンにして「実行」ボタンをクリックしてください。

「詳細分析オプション」の「検出機能で使用することを提唱」チェックボックスのオン / オフに応じて、次のような 2 通りの推測処理が実行されます。

- チェックボックスがオフの場合
  重要項目 (Column Importance) 機能によって項目（属性）が検出されます。重要項目 の詳細については、「重要項目」（58 ページ）を参照してください。
- チェックボックスがオンの場合
  初めに、重要項目 (Column Importance) 機能が実行された後、それによって検出された属性の組合わせを開始点として、すべての組合わせがテストされます。各段階で誤差が評価され、すべてのオプションが試行されます。予想される通り、処理に非常に長い時間がかかります。「停止」ボタンをクリックすると、この処理をいつでも中止することができます。属性検出処理を長く実行するほど、クラシファイアの精度が高くなります。

デシジョン・テーブルの解釈

デシジョン・テーブル・ビジュアライザでは、各クラスラベルの事前確率 (*Prior probability*) が、画面の右側にある「ラベル確率」ウィンドウ内の円グラフによって表されます。クラスラベルの事前確率 (Prior probability) とは、属性値に関係なくレコードをランダムに選択した場合、データセット内でそのラベルが観察される確率です。数学的に、この値は特定のクラスラベルを持つレコードの数をデータセット内のレコードの総数で除算した値になります。

左側にあるメインウィンドウ内の矩形ブロック（ケーキ）の確率分布 (*Probability distribution*) は、特定の組合わせの属性値を持つレコードが各クラス内で現れる比率を示します。この確率分布は、特定のデータセットの特性を正確に表しています。

デフォルトでは、離散属性の個々の値はクラスラベルの予測に寄与する度合い（ラベル確率）に基づいてソートされます。したがって、重要な属性値を簡単に識別するこ

とができます。ラベルが階級生成された（元々は連続的な）属性である場合は、最大値の階級のクラスがソート基準として使用されます。ラベルが離散的な属性である場合は、事前確率の円グラフでスライスが最大であるクラスがソート基準として使用されます。特定のクラスを選択して、ラベル確率によるソートを要求すると（「属性値の順序付け」メニューの「ラベル確率順」オプションをクリックすると）、選択したクラスがソート基準として使用されます。なお、離散型属性の個々の値はアルファベット順または重み順にソートすることもできます。階級生成された属性の値は常に自然な順序で表示されます。

グラフ上でマウスの右ボタンをクリックして矩形ケーキの次のレベルにドリルダウンすると、そのケーキによって表されるデータがケーキの新しいマトリックス内に再表示されます（このマトリックスでは、現在のレベルに割り当てられた属性が軸として使用されます）。ケーキグラフは灰色のベースの上に表示されます。ベースがケーキによって完全に覆われている場合は、そのレベルにあるすべての組合わせの属性値がデータによって表されていることを意味します。通常、ベースの大部分はケーキによって覆われていません（データが存在しない領域があることを意味します）。非常に希薄なデータセットでは、データが存在しない領域が多くの割合を占めています。灰色のベースを選択することは、1 つ上のレベルのケーキを選択するのと同じ効果があります。ベースをドリルダウンすると、ベースの上にある各ケーキが展開されて次のレベルが表示されます。

属性が NULL 値である場合は、その NULL 値が疑問符 (?) で表されます。NULL 値は常に最初（一番左側）の値として表示され、他の値と一緒にソートされることはありません。NULL 値の表示 / 非表示を切替えるには、「表示 」メニューの「NULL 値の表示 」オプションを使用します。

ナビゲーション方法は、プラットフォームの種類（Windows または IRIX）とマウスのタイプ（2 ボタンまたは 3 ボタン）に応じて異なります。ここでは、Windows の 2 ボタン・マウスを想定して説明します。マウスボタンのモードは、デスクトップの「設定」パネルで変更することができます。

左側のウィンドウ内でマウスの左ボタンをクリックして 1 つのケーキを選択すると、そのケーキによって示される確率分布とまったく同じ事後確率分布が右側の円グラフに表示されます。<Ctrl> キーを押しながらマウスの左ボタンをクリックして複数のケーキを選択すると、選択されたケーキによって定義されるレコード集合の（ラベルに関する）確率分布が右側の円グラフに表示されます。灰色のベースをクリックして、ケーキのグループ全体を選択することもできます。これは階層内で 1 つ上のレベルのケーキを選択するのと同じ効果があります。

右側の円グラフの下に表示されるクラスが数値型でない場合は、各クラスがスライスサイズの順番に並べられます（確率が最大であるクラスが一番上に表示されます）。左

側の属性値を選択すると、円グラフ（確率分布）の変化を反映するように各クラスの表示順序が変わります。ラベルが階級生成された属性である場合、各クラスの表示順序は変更されず、連続的なスペクトルに基づく色が各クラスに対応付けされます（最大値の階級が赤色になり、それ以外の階級はランダムな色になります）。

「ラベル確率 (Label Propability)」ウィンドウの下にある「詳細」スライダを右側に移動すると、より詳細な情報が表示されます。

デシジョン・テーブルのオプション

「詳細設定」（Windows システム）または「詳細分析オプション (Further inducer options)」（IRIX システム）を選択すると、「分析オプション」ダイアログ・ボックスが表示されます。このダイアログ・ボックスには次の 4 つのパネルがあります。

- 一番上のパネルには、Tool Manager の「データの可視化 / マイニング」パネルで設定した項目が表示されます。
- 一番上から 2 番目のパネルでは、損失マトリックスと重み属性を設定することができます。詳細については、「損失マトリックス」（121 ページ）と「重み付け」（220 ページ）を参照してください。
- 左下のパネルでは、詳細な「分析オプション」を設定することができます。
- 右下のパネルでは、「誤差推定のオプション」を設定することができます。ただし、「データの可視化 / マイニング」パネルで「クラシファイアのみ」モードを選択した場合、このパネルには何も表示されません。このパネルに表示されるオプションは、選択した誤差推定法に応じて異なります（「モデルの適用」（21 ページ）と「誤差推定」（87 ページ）を参照）。

デシジョン・テーブル分析の各種オプションを設定すると、デシジョン・テーブル分析のアルゴリズムを微調整することができます。

- 最大サイズ

  データセットが大きい場合は、このオプションが特に便利です。デフォルトの最大サイズは 10,000 ノードです（この値はビジュアルのケーキ数に一致しています）。最大サイズを変更するには、フィールドをクリックして新しい制限値を入力します。最大サイズを制限すると、分析の実行速度が高速化され、ディスク領域が節約されますが、誤差率が増加する場合があります。値のすべての組み合わせを表示するのに必要なサイズよりも最大サイズが小さい場合は、対応付けされた属性の一部がビジュアルに表示されません。

- 最大の属性値

このオプションは、検出時にデシジョン・テーブルで使用できる項目（属性）の数を制限します。属性数を制限すると、ビジュアルが簡略になり検出時間が短縮されます。この制限値は「提唱」モードによって自動検出される項目だけに適用され、手動で追加される項目には影響を与えません。

- 階級での重み付け下限値

  デシジョン・テーブルでは、連続的な属性が離散的な範囲に階級生成されます。このオプションでは、1 つの階級に入れる事象の最小値を設定することができます。自動設定の場合は、データセットのサイズに基づいて事象の最小値が設定されます。一般的に、データセットが大きくなるほど、階級も大きくなります。データセットが非常に大きい場合は、必要な数を越える階級が作成される場合があります。階級の数を減らすには、重み付け下限値を高くしてください。

ドリルダウンとドリルアップ

デシジョン・テーブル・ビジュアライザ内では、ケーキグラフのドリルダウンまたはドリルアップを行って、詳細レベルを増減することができます（「ドリルダウンとドリルアップ」（86 ページ）を参照）。ドリルダウンを行うには、1 つのケーキをマウスの右ボタンでクリックします。こうすると、そのケーキが小さい複数のケーキで置き換えられ、1 つ下のレベルにある属性ペアによって分割されたデータが表示されます。カーソルが灰色のベース上にあるときにマウスの右ボタンをクリックすると、そのベース上の全ケーキに対してドリルダウンが行われます。カーソルが背景上にあるときにマウスの右ボタンをクリックすると、すべてのケーキに対してグローバルなドリルダウンが行われます。ドリルアップを行うには、1 つのケーキ、灰色のベース、または背景上で < Ctrl > キーを押しながらマウスの右ボタンをクリックします（3 ボタン・マウスを使用する場合は、マウスの中ボタンでオブジェクトをクリックします）。

プルダウン・メニュー

デシジョン・テーブル・ビジュアライザでは、5 種類のプルダウン・メニュー（「ファイル」、「表示」、「属性値の順序付け」、「選択」、「ヘルプ」）を通じて、さまざまな機能を利用することができます。設定ファイルの名前を指定しないでデシジョン・テーブル・ビジュアライザを起動した場合は、「ファイル」メニューと「ヘルプ」メニューしか使用できません。「表示」と「属性値の順序付け」以外のメニューの説明については、「ファイル」メニュー、「ヘルプ」メニュー、「選択」メニューを参照してください。

「表示」メニュー

「表示」メニューの下記のオプションを使用すると、デシジョン・テーブル・ビジュアライザのメインウィンドウに表示される情報を制御することができます。

- 「フィルタパネル」
  ユーザが指定する条件に従ってデータをフィルタリングできるパネルが表示されます（「「フィルタ」パネル」（103 ページ）を参照）。
- 「色の選択」
  ダイアログ・ボックスが表示され、新しい背景色を指定することができます（「色の選択」（53 ページ）を参照）。
- 「拡張コントローラーの表示」
  メインウィンドウ内にある外部コントロールの表示 / 非表示を切替えます。
- 「NULL 位置の表示」
  NULL 値の表示 / 非表示を切替えます。NULL 値は他の非 NULL 値から少し離れて、最初の値として表示されます。
- 「ランドスケープ ビューワの使用」
  別の 3D ナビゲーション・モードに切替えます。ランドスケープ・ビューワでのナビゲーションではマウスの中ボタンが使用されるため、ドリルアップを行うときは、< Ctrl > キーを押しながらマウスの中ボタンをクリックする必要があります。このオプションは Windows システム上では使用できません。
- 「エビデンスとして表示」、IRIX システム上では「エビデンスモード」
  レコードの重みに基づく分布（最初に表示される情報）ではなく各ケーキの条件付き確率を表示します。このオプションは、一部のクラスが小さいときに便利です。
- 「ツールバーとステータスバー」
  ツールバーとステータスバーの表示 / 非表示を切替えます。このオプションは Windows システム上だけで使用できます（IRIX システム上では使用できません）。

「属性値の順序付け」メニュー

「属性値の順序付け」メニューでは、離散属性の値をどのようにソートするかを制御することができます。このメニューには、次のオプションがあります。

- 「アルファベット順」
  離散属性値を左から右に（または下から上に）アルファベット順でソートします。
- 「重み順」
  レコードの重みが大きい属性が左側に並ぶように、属性値を左から右にソートします。
- 「ラベル確率順」（デフォルト）
  クラスの 1 つを表すスライスのサイズに基づいて、属性値を左から右にソートします。ラベルが階級生成された属性である場合は、最大値の階級のクラスがソート基準となります。ラベルが離散的な属性である場合は、事前確率の円グラフでスライスが最大であるクラスがソート基準となります。特定のクラスを選択して、ラベル確率によるソートを要求すると、選択したクラスがソート基準となります。いずれの場合も、NULL 値は一番左側に最初の値として配置されます。

決定木

決定木は、従属属性（既知の属性）の値に基づいてラベル属性（未知の属性）の値を推定する予測モデルです。離散属性の値（通常は "yes"、"no" などの文字列）または数個の値しか取らない属性の値を予測する処理は「判別モデル」と呼ばれます。決定木分析は、決定木の構造に基づき、各レコードのラベルを予測することによってデータを判別モデルとして生成します。決定木分析によって分析されたモデルのデータの構造は、ツリー・ビジュアライザによってグラフィカルに表示されます。

決定木分析では、さまざまな属性の相互作用（すなわち、属性値の組合わせがラベルの予測に与える影響）が示されます。予測されるラベルは、個々のレコード内にある未知の属性です。決定木分析では、以前のノードにおける分析モデル結果によってそれ以降のノードにおけるデータの分布が決まります。

ツリー・ビジュアライザによるグラフでは、決定木の各ノードにあるバーが個々のラベル値を表します。バーの上にカーソルを移動すると、そのラベル値に関連するレコード数（重み）とパーセント値が表示されます。各ノードのベースには、そのノードに到達したレコードの数（重み）が表示されます。

決定木の作成

決定木クラシファイアはデータに基づいて自動的に導出（生成）されます。最初にサーバにログインして、通常の手順に従ってデータセットを選択します。

Tool Manager の「クラス判別」タブで、「分析」ポップアップ・メニューから「決定木」を選択してください。必要な場合を除き、特別なオプションを設定する必要はありません。単に「実行」ボタンをクリックすると、作成されます。決定木では、ツリー・ビジュアライザを使用して分析結果が表示されます。ビジュアルには、次の情報が表示されます。

- 決定木ノード上のラベルは、そのノードでテストされる属性を表します。
- 決定木内のリーフノードは、特定のクラスを表します。
- ベースの色は、サブツリーの誤差推定を表します。
- 各ノードの一番上にある縦のバーは、そのノードにおけるクラスの分布を表します。

ノードをマウスでポイントすると、次の情報が表示されます

- サブツリーの重み — サブツリー内部の訓練事例レコードのうち、ポイントされたノードの下位にあるレコードの重み。この値はベースの高さに反映されます。
- テストセットの誤差 / 損失 — サブツリーの誤差推定（損失マトリックスが指定された場合は予想損失）。+/- の後の数値は、予想値の標準偏差です。標準偏差が大きくなるほど、予想値の精度は低くなります。リーフのレコード数が少ない場合、またはテストセットの誤差率が 0% または 100% に近い場合は、予想値と標準偏差の信頼度が低くなります。
- テストセットの重み — ポイントされたノードに到達したテストセット・レコードの重み（重みが設定されていない場合はレコード数）。
- 寄与率 — ノードにおけるラベル値の分布の歪みを表す値（範囲は 0 ～ 100）。ノードの全レコードが単一のクラスに属している場合、寄与率は 100 になります。ラベル値の重みがすべて同じである場合、寄与率は 0 になります。寄与率はバックフィットの後で計算されます。

IRIX システム上での並列化処理

マルチプロセッサ版の MineSet をインストールした場合は、決定木の枝に 1,000 個を越えるレコードが含まれているときに、ツリー・アルゴリズムを並列的に計算することができます（「IRIX システム上での並列化処理」（146 ページ）を参照）。スレッドの数を調整するには、Tool Manager の「設定」パネルで「並列化処理」モードを変更します（「「ファイル」メニュー」（100 ページ）を参照）。この並列化オプションは、IRIX システム上だけで使用することができます。

詳細分析オプション

バックフィットがオンに設定されていて、「訓練事例をディスクとして表示」オプションをオンに設定した場合は、訓練事例の分布がディスクとして表示されます。ディスクの高さはバーの高さに対応します。すなわち、ディスクの高さは「詳細分析オプション」で設定された予備比率を表します。

決定木分析 (Decision Tree) のオプション

「詳細設定」（Windows システム）または 「詳細分析オプション (*Further Inducer Options*)」（IRIX システム）を選択すると、「詳細分析オプション」ダイアログ・ボックスが表示されます。このダイアログ・ボックスには次の 4 つのパネルがあります。

- 一番上のパネルには、Tool Manager の「データの可視化 / マイニング」パネルで設定した項目が表示されます。
- 一番上から 2 番目のパネルでは、損失マトリックスと重み属性を設定することができます。詳細については、「損失マトリックス」（121 ページ）と「重み付け」（220 ページ）を参照してください。
- 左下のパネルでは、詳細な「分析オプション」を設定することができます。
- 右下のパネルでは、「誤差推定のオプション」を設定することができます。ただし、「データの可視化 / マイニング」パネルで「クラシファイアのみ」モードを選択した場合、このパネルには何も表示されません。このパネルに表示されるオプションは、選択した誤差推定法に応じて異なります（「モデルの適用」（21 ページ）を参照）。

ダイアログ・ボックスの「分析オプション」セクションにある各種オプションを設定すると、決定木分析のアルゴリズムを微調整することができます。

- 「ツリーの高さ制限」

  デフォルトでは、決定木のレベル数（高さ）に制限はありません。レベル数を制限するときは、このチェックボックスをクリックして最大レベル数を入力します。決定木のレベル数を制限すると分析の実行が高速になり、数多くのノードに気を取られないで決定木を調査することができます。また、並列処理の負荷バランスが改善されます。ただし、レベル数を制限すると、誤差率が増加する可能性があります。このオプションを設定しても、最大レベル数よりも前のレベルで選択された属性は影響を受けません。

- 「分割の基準」

  このオプションでは 5 種類の分割基準を選択することができます。下記の定義は技術的な説明であり、個別の問題に最も適した分割基準を判断するのは困難です。すべての分割基準を試して、誤差推定が最も小さくなる基準を選択するか、最も分かりやすい決定木が生成される基準を選択してください。

  「相互情報量」は、子ノードの加重平均寄与率と親ノード間の寄与率の変化（エントロピー）です。加重平均寄与率は、個々の子ノードに存在するレコードの数に基づいて計算されます。

  「正規化された相互情報量」（デフォルト）は、「相互情報量」を子ノードの数の対数（底は 2）で除算した値です。

  「増加比率」は、「相互情報量」を分割のエントロピーで除算した値です（ラベル値は無視します）。「正規化された相互情報」と「増加比率」は、数個の値しか取らない属性に適しています。

  「カイ二乗」は、分割のすべての候補に対してカイ二乗統計独立性検定を適用した後、ラベル値の独立性が最小になる分割を選択します。

  「ジニ集中係数」は CART (Classification And Regression Trees) で使用される分割基準です。「相互情報量」と同様に、「ジニ集中係数」でも子ノードの加重平均寄与率と親ノード間の寄与率の変化（エントロピー）が測定されます。ただし、「相互情報量」とは違って、ジニ集中係数のノードの寄与率は「 1 - (ノードにおけるラベル確率の二乗和)」という式で計算されます。

- 「分割の下限値」

  このオプションの値は、ノードの子ノードのうち、少なくとも 2 つのノードで設定する必要のある重み（重みが設定されていない場合はレコード数）の下限です。このオプションのデフォルト値は 2 です。たとえば、ノードを 3 方向に分割する場合は、3 つのうち少なくとも 2 つの子ノードに 2 以上の重み（重みが設定されていない場合は 2 つ以上のレコード）を割当てる必要があります。これは決定木のサイズを制限して実行時間を短縮する代替手段になります。

  分割の下限値を増加させると、各枝上のレコード数（重み）が増えるため、確率推定の精度が改善される傾向があります。データにノイズ（誤差または異常値）が含まれていると思われる場合、またはツリーを使用して確率を予測する場合は(「モデルの適用」を参照)、分割の下限値を 5 以上に増加させてください。データセットが非常に小さい（レコード数が 100 未満）場合は、この下限値を減少させてもかまいません。

- 「枝刈り」

  決定木は、「ツリーの高さ制限」と「分割の下限値」による制限に基づいて構築されます。その後、統計テストが実施されて、単一のリーフノードよりも有意でないサブツリーが検出され、そのようなサブツリーが枝刈りされます。決定木内のサブツリーの枝刈りを制御するために、「信頼度」、「生成コストも考慮」、「無し」という 3 つのオプションが用意されています。

  枝刈りによるツリーの簡素化では、ツリー全体が構築されてからサブツリーが除去されるため、最大レベル数または分割の下限値による簡素化よりも実行時間は長くなります。ただし、サブツリーが選択的に除去されるため、より正確なクラシファイアが生成されます。

  「信頼度」はデフォルトの枝刈りオプションです。大きい枝刈り係数を指定すると除去されるサブツリーの数が増加し、小さい枝刈り係数を指定すると除去されるサブツリーの数が減少します。デフォルトの枝刈り係数である 0.7 を指定すると、適切な数のサブツリーが除去されます。データにノイズ（誤差または異常値）が含まれている場合は、枝刈り係数を増加させて小さいツリーを作成してください。指定できる値の下限は 0 ですが、上限には制限がありません。枝刈り係数を 0（切り詰めなし）に設定すると、単一のノードの誤差率がサブツリーの誤差率と同じかそれ以下である場合に限り、サブツリーが除去されます。

  「生成コストも考慮」オプションでは、ツリーの誤差率（コスト）とツリー内の枝の数（複雑性）のトレードオフ関係を調整することによって、最低コストのツリーよりもサイズの小さいツリーを選択することができます。これによって、許容できる最低コストのツリーよりも標準誤差の数が多くなります。このオプションのパラメータを 0 に設定すると、最低コストのツリーが選択されます。また、このパラメータを 0.5 に設定すると、誤差率が（最低コストのツリーの誤差率 + 標

準誤差 * 0.5）未満の最低サイズのツリーが選択されます。デフォルト値の 0 に設定すると、最低コストの最低サイズのツリーが選択されます。パラメータ値を大きくすると、除去されるサブツリーの数が増えます。データにノイズ（誤差または異常値）が含まれている場合は、パラメータ値を大きくして小さいツリーを作成してください。ツリーが切り詰められて 1 つのノードになってしまう場合は、パラメータ値を小さくして大きいツリーを作成してください（「生成コストの考慮」（65 ページ）を参照）。

「無し」オプションを選択すると、枝刈りが行われません。サブツリーをまったく除去しないと、訓練事例にオーバーフィットする決定木が生成されるためにクラシファイアの誤差率が増加しますが、決定木の完全な構造（制限がない場合の構造）を検証することができます。

- 「二者択一の分割」

  このオプションをオンに設定すると、3 つ以上の値を取り得る離散属性について二者択一の分割を行うことができます。通常、離散属性について分割を行うときは、その属性が取り得るすべての値を対象としてツリーを分割します（「決定木の作成」（78 ページ）を参照）。たとえば、"color" という属性についてツリーを分割するときは、赤、緑、黄、青という値ごとに 1 つの枝を作ります。二者択一の分割では、たとえば、赤および赤以外という 2 つの枝にツリーを分割することができます。すなわち、二者択一の分割を行うと、属性の取り得る複数の値のうち 1 つの値だけが分離されます。データの属性が取り得る複数の値のうち、ラベルの判別に特に適した値が存在する場合は、このような分割方法を採用すると有効です。

- 「ブースト」

  ブースティングはモデルの精度を改善するための手法ですが、非常に長い時間を必要とする処理です。ブースティングとビジュアル処理を同時に実行することはできません。

  ブースティングの実行中は、推定誤差がステータス・ウィンドウに表示されます。ブースティングを実行すると、訓練事例内の個々のレコードに新しい重みが割当てられ、その訓練事例に基づいてクラシファイアを作成する処理が繰返し実行されます。詳細については、「ブースティング」（42 ページ）を参照してください。

検索パネルとフィルタパネル

「表示 」メニュー（IRIX システム上では「表示 (Show)」メニュー）から 「検索パネル 」または 「フィルタパネル」を選択すると、オブジェクトの検索またはフィルタリングに関する条件を指定するためのダイアログ・ボックスが表示されます。決定木分析では検索とフィルタリングで同じオプションが使用されます。これらのオプションを以下で説明します。

検索またはフィルタリングを特定のクラスラベルに限定するには、上部ウィンドウのクラスリスト内の値を選択するか、下部ウィンドウのクラス項目を使用します。こうすると、より強力な比較演算子（Matches など）を使用できるようになります。別の条件を表示するには、下方向にスクロールします。次に、オプションについて説明します。

- 「サブツリー重み」― 指定されたサブツリー重み（重みが設定されていない場合はレコード数）を持つバーまたはベースだけを検索 / フィルタリングの対象とします。検索対象としてバーまたはベースを選択するには、「バー」ラジオボタンまたは「ベース」ラジオボタンを使用します。たとえば、重みが 50 以上のバーだけを検索することができます。
- 「テスト属性 」― 指定されたテスト属性をラベルとするノードだけを検索 / フィルタリングの対象とします。決定木ノードのラベルはテスト属性を表し、リーフノードのラベルは予想されたラベルを表すことに注意してください。たとえば、テスト属性として年齢 (*Age*) を選択した場合は、年齢の値がテストされるノードだけが検索 / フィルタリングの対象となります。
- 「テスト値」― 入ってくるライン上の値が指定の値と一致するノードだけを検索 / フィルタリングの対象とします。
- 「占有率」― 各ノードにおける重み全体のうち指定の割合 (%) を占めるバーだけを検索 / フィルタリングの対象とします。たとえば、特定のクラスが重み全体の 80% より多くを占めるようなノードを見つける場合は、クラスラベルをクリックし、占有率 > 80 と指定します。バーではなくベースを選択した場合、このオプションは意味を持ちません（ベースの重み値は 0 です）。
- 「寄与率」― 指定された寄与率を持つノードだけを検索 / フィルタリングの対象とします。たとえば、大部分のレコードが単一のクラスに属しているようなノードを見つける場合は、寄与率 > 90 と指定します。
- 「テストセットのサブツリー重み」― テストセット (test set) のサブツリー重み（重みが設定されていない場合はレコード数）が指定の範囲に一致するノードだけを検索 / フィルタリングの対象とします。

- 「テストセットの誤差 / 損失」― テストセット (test set) の予想誤差 / 損失が指定の範囲に一致するノードだけを検索 / フィルタリングの対象とします。
- 「平均誤差 / 損失の標準偏差」― テストセット (test set) の予想誤差 / 損失の標準偏差が指定の範囲に一致するノードだけを検索 / フィルタリングの対象とします。
- 「レベル」― 検索 / フィルタリングの対象を特定のレベルまたは一定範囲のレベルに限定します。たとえば、最初の 5 レベルだけを検索することができます。

次のオプションは決定木ではあまり使用されません。

- 「階層」― ルートから伸びるパスの末尾の値が指定の値に一致するノードとラインが検索され、それらのノードの子ノードがマークされます。
- 「Null をゼロとみなす」― 決定木分析では Null 値が生成されないため、このオプションは決定木では使用されません。

検索が完了すると、検索条件に一致するオブジェクトが黄色いスポットライトで強調表示されます。強調表示されたオブジェクトに関する情報を表示するには、マウスのポインタをスポットライト上に移動します。オブジェクトに関する情報は、左上隅の「ポインタを通過」というラベルの下に表示されます。黄色いスポットライトの下のオブジェクトを選択してズームするには、マウスの左ボタンでスポットライトをクリックします。クリックするときに < Shift > キーを同時に押すと、ズームは行われません。

「離散型ラベル (Discrete Label)」メニュー（Discrete Label Menu）

Tool Manager の「クラス判別」タブにある 「離散型ラベル」メニューに離散的な値を取るラベル属性のリストが表示されます。離散的な属性（階級生成値、文字列、数個の整数など）は限られた数の値しか取りません。ラベルとして使用する属性については、できる限り少ない数（理想的には 2 ～ 3 個）の値を取る属性を選択してください。離散的な属性が存在しない場合は、このメニューに「離散型のラベルがありません」と表示され、「実行」ボタンは使用できない状態になります。その場合は、Tool Manager の「データ変換」パネルを使用して新しい項目を追加するか既存の項目を階級生成して、離散的な属性を新たに作成する必要があります。

ドリルスルー

現在選択されているオブジェクトのオリジナルデータ（データソースから取得したデータ）を表示して、操作することが必要になる場合があります。こうしたオリジナルデータの表示や操作を「ドリルスルー」と呼びます。ビジュアライザのウィンドウ内でオブジェクトを選択し、その元データを Tool Manager に送信すると、別のツールを使用してデータの表示や分析を行うことができます。MineSet の全ツールにこのような機能がありますが、このトピックスの最後で説明するように、ツールの種類や操作の履歴によっては一定の制限が課せられることがあります。

「選択」メニューには、ドリルスルーを実行するためのオプションが 2 つ用意されています。

- 「Tool Manager に送信」

  オプションを選択すると、可視化の操作の履歴が Tool Manager に送信されます。この履歴には、ユーザがビジュアライザ内で行った選択操作に使用されたフィルタも追加されます。フィルタは後述の制限に従い、履歴のできるだけ早い時期に挿入されます。

- 「オリジナルデータを表示」

  オプションを選択すると、「Tool Manager に送信 」オプションの場合と同様に、可視化の操作の履歴が Tool Manager に送信され、履歴のできるだけ早い時期にフィルタ操作が挿入されます。ただし、履歴内でフィルタ操作より後にある非フィルタ操作はすべて削除されます。この新しい履歴を使用して、レコードビューワ (Record Viewer) に表示するテーブルが作成されます。フィルタを履歴の先頭に配置できれば、オリジナルのレコードが表示されます。履歴の先頭がフィルタでないときは、データがオリジナルではないことを知らせる警告メッセージが表示されます。Tool Manager の状態は（Tool Manager が現在実行中でない限り）変更されません。

どちらのオプションでも、Tool Manager によって全操作が実行されます。Tool Manager が動作中でない場合は、自動的に起動されます。

ドリルスルーでは、Tool Manager によって作成されたビジュアルデータしか使用できません。ビジュアライザ用の設定 (*.schema*) ファイルには履歴セクションがあり、そのファイルが生成されたプロセスを Tool Manager に知らせるようになっています。学習曲線、混同マトリックス、改善曲線などの特殊な目的の可視化では履歴が作成されないため、ドリルスルーはサポートされません。

ドリルスルーの対象とするオブジェクトを選択すると、可視化されたテーブルに基づくフィルタ・ステートメントが暗黙的に指定されます。可視化の前にテーブルが変換

されている場合は、Tool Manager がフィルタを変更して、履歴の早い段階にフィルタを挿入しなければならないことがあります。たとえば、階級生成された項目に基づくフィルタの場合は、階級生成以前の項目を参照するように、Tool Manager がフィルタを変更します。

Tool Manager が一部の操作に関するフィルタを変更できないために、フィルタを履歴の先頭に置くことができない場合があります。たとえば、「項目の追加」、「集計処理」、または「モデルの適用」によって作成された項目を対象とするフィルタは、項目を作成する操作よりも前には置けません。また、既存の「標本」操作よりも前にフィルタを置くことはできません。「標本」より前にフィルタを置くと、標本抽出の結果が大幅に異なることになります。

「オリジナルデータを表示」オプションでは、できるだけ早い段階のレコードを表示するために履歴が枝刈りされます。ただし、実際に選択された数よりも多くのレコードが表示されないように、既存のフィルタ操作（ユーザが実行したフィルタ、または以前のドリルスルーによって実行されたフィルタ）は削除されません。

ドリルダウンとドリルアップ

ドリルダウンとドリルアップは、オブジェクトをクリックして表示の詳細レベルを増減させる操作です。これらの操作は、マップ・ビジュアライザとデシジョン・テーブル・ビジュアライザだけで実行することができます。

ドリルダウンを行うには、マウスの右ボタンでオブジェクトをクリックします。ドリルアップを行うには、<Ctrl> キーを押しながらマウスの右ボタンでオブジェクトをクリックします（3 ボタンのマウスを使用する場合は、マウスの中ボタンでオブジェクトをクリックします）。グラフの背景をクリックすると、グローバルなドリルダウンまたはドリルアップが行われます。

誤差推定

クラシファイアを構築したときは、そのクラシファイアの誤差率を調べて、その後の予測性能を評価する必要があります。クラシファイアの誤差率に影響を与える要因には次のようなものがあります。

- 訓練事例内に存在するレコードの数

  分析では訓練事例に基づいて学習が行われるため、訓練事例のサイズが大きいほど（レコードの数が多いほど）、クラシファイアの信頼度は高くなります。ただし、訓練事例のサイズが大きくなるほど、クラシファイアの構築にかかる時間が長くなります。訓練事例のサイズが大きくなるに従って、誤差の改善率は低減します（収穫逓減の法則）。

- 属性の数

  属性の数が多くなると、分析によって計算される組合わせの数が多くなるため、問題が複雑になってクラシファイアの作成に長い時間が必要になります。また、数多くの属性間のランダムな相関によって分析が混乱し、精度の低いクラシファイアが作成される可能性があります（このような状態は「オーバーフィッティング」と呼ばれます）。タスクに関連のない属性は、Tool Manager を使用して訓練事例から削除してください。

- 属性に含まれる情報

  ラベルを正確に予測するのに必要な情報が記述属性に含まれていない場合があります（たとえば、人の目の色に基づいてその人の給料を予測する場合など）。そのような場合は、他の記述属性（職業、1 週間の就業時間、年齢など）を追加すると、誤差率が低下することがあります。

- 新しいデータセット内のレコードの分布

  訓練事例内の属性値の分布と新しいデータセット内の（ラベルがない）属性値の分布が異なる場合は、誤差率が高くなる傾向があります。たとえば、自家用車のレコードからなる訓練事例に基づいてクラシファイアを作成した場合、そのクラシファイアはスポーツカーのレコードからなるデータセットのクラス判別には適していません。その理由は、これら 2 つのデータセット内にある属性値の分布が非常に異なるためです。

クラシファイアの誤差率を評価する 2 通りの一般的な方法を下記に説明します。これらの方法は、訓練事例と同じ分布のデータセットから新しいレコードが標本抽出されることを前提としています。

- 予備法 (Holdout)

  全レコードの一定の比率（通常は 2/3）を訓練事例として使用し、残りをテストセット (*test set*) として使用します。分析によって訓練事例が評価され、クラシファイアが作成されます。作成されたクラシファイアを使用してテストセットをクラシファイアとして生成し、そのテストセットに関する誤差率または損失をクラシファイアの推定誤差率または推定損失とみなします。図 1-17 に、予備法による誤差推定を示します。

  バックフィッティングがオンに設定されていて、「訓練事例をディスクとして表示」オプションをオンに設定した場合は、訓練事例の分布がディスクとして表示されます。ディスクの高さはバーの高さに対応します。すなわち、ディスクの高さは「詳細分析オプション」で設定された予備の比率を表します。

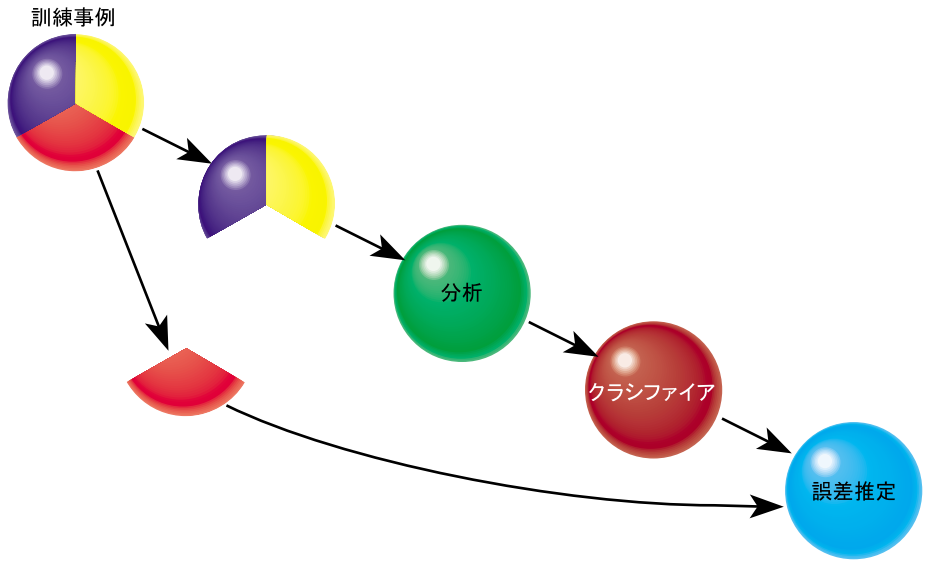

図 1-17 予備法によるクラシファイアの誤差推定

予備法は高速ですが、クラシファイアの構築にデータセットの一定の比率しか使用しないため、学習過程でデータが有効に利用されません。データセットの全体を使用できれば、さらに精度の高いクラシファイアが作成されるはずです。

- 相互検証法

  データセット全体をほぼ同じサイズの *K* 個の互いに独立したサブセットに分割します。1 個のサブセットをテストセット (test set) として使用し、(*k* - 1) 個のサブセットを訓練事例として誤差率の評価を *k* 回繰返し、*K* 個の誤差率の平均値を推定誤差率とみなします。図 1-18 に、*k*=3 のときの相互検証を示します（デフォルト値は *k*=10 であることに注意してください）。

  相互検証は *t* 回繰返すこともできます。*k* 回の相互検証を t 回繰返すと、*k***t* 個のクラシファイアが構築されて評価されます。これは相互検証に必要な時間が *k***t* 倍に増えることを意味します。デフォルトでは、*k*=10、*t*=1 であるため、相互検証を実施するには単一のクラシファイアを構築するときの約 10 倍の時間がかかることになります。

  繰返し回数 (*t*) を増やすと、テストの実施時間が長くなり、誤差率と信頼区間が改善されます。

  k の値は自由に増減することができます。*k* を 3 ～ 5 に減らすと、テストの実施時間が短くなりますが、訓練事例のサイズが小さくなるため、誤差率がかなり高くなる傾向があります。*k* の値を 10 以上に増やすこともできますが、これは非常に小さいデータセットの場合に限ってください

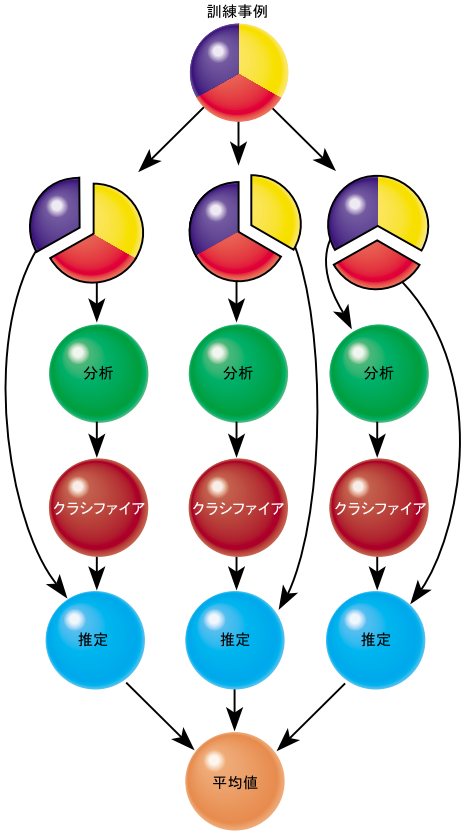

図 1-18 相互検証のクラシファイア (k=3)

一般的に、予備法はクラシファイア構築の準備的な段階で使用するか、レコード数が 5,000 個を越える大きいデータセットに適用してください。相互検証は最終的な段階で使用するか、比較的小さいデータセットに適用してください。

バックフィッティングでは、訓練事例のデータだけを使用して作成されたクラシファイアに対してデータセット全体が適用されます。予備法で誤差を評価するときは、データの一部がテスト用に残されます。クラシファイアの構造をすべてのデータに対してバックフィッティングすると、最終的な誤差を低減することができます。クラシファイアの構造内のレコード数（カウント）、重み（ウェイト）、確率などは、訓練事例だけではなくデータセット全体を反映しています。詳細については、「バックフィッティング」（37 ページ）を参照してください。

エビデンスモデル

エビデンス・クラシファイアは特定の属性に基づいて特定の事象が発生する確率を判別することによって、データセット内の各レコードを特定のクラスに割り当てます。このモデルの構造はエビデンス・ビジュアライザによって表示されます。このビジュアル表示により、クラシファイアにとって重要な属性の検出が容易になり、データをより深く洞察することができます。また、"what if" 型の質問（条件に基づくアクション）への答えが見つけやすくなります。エビデンス・ビジュアライザの詳細については、『*MineSet 3.0 Enterprise Edition User's Guide for Windows*』を参照してください。

エビデンス分析

エビデンス分析によってモデルが構築されるときは、各属性値の確率が特定のクラスに依存しない（各クラスの各属性が互いに独立である）ことが前提となっています。たとえば、*iris* データセットでは、アヤメの 4 つの属性（萼片の縦、萼片の横、花びらの縦、花びらの横 (sepal length, sepal width, petal length, petal width)）がアヤメの各クラス（アイリスセトサ、アイリスバージカラー、アイリスバージニカ (iris-setosa, iris-versicolor, iris-virginica)）に依存しないとみなされます。この単純な前提はあまり現実的ではありませんが、エビデンスはデータの初期分析には適しており、そのクラス判別結果は実用的な業務にも適用することができます。

エビデンス・ビジュアライザでは、個々の属性値（連続的な属性の場合は階級生成された一定範囲の属性値）が正確に 1 つのグラフに対応しており、各グラフによって各クラスラベルの条件付き確率が表されます。特定のレコードをクラス判別するには、各クラスの事前確率に行列内の各行の条件付き確率を掛け合わせて各クラスの相対確率を計算します。この相対確率が最大であるクラスがレコードのラベルの予想値となります。属性に含まれている未知の値（NULL 値）は無視されます。これらの NULL 値は、通常のグラフから少し離れた左側のグラフによって表されます。また、MineSet に付属のサンプルである *iris.schema* ファイル（詳しいパス名は付録 A 「設定ファイルとデータファイルのサンプルファイル」を参照）を開くと、NULL 値は左側に表示されます。NULL 値の詳しい説明については、『*MineSet 3.0 Enterprise Edition User's Guide for Windows*』を参照してください。

iris データセットにエビデンス分析を適用すると、各クラスラベルの事前確率 (*prior propability*) が、画面の右側にあるラベル確率パネル内の円グラフによって表されます。クラスラベルの事前確率とは、属性値に関係なくレコードをランダムに選択した場合に、そのラベルが観察される確率です。数学的に、この値は特定のクラスラベルを持つレコードの数をデータセット内のレコードの総数で除した値になります。メインウィンドウ内のエビデンス（ケーキグラフ）と確率（円グラフ）を切替えるには、右

上にある「エビデンス」ラベルをクリックするか、「表示」メニューの「エビデンスモード」を切替えます。

画面の左側にあるメインウィンドウ内のケーキグラフによって表される条件付き確率は、特定のラベル値が現れるという条件のもとで特定の属性値が現れる相対的な確率を示します。ケーキのスライスのサイズは、特定の属性値が現れるという条件のもとで、クラシファイアが事前確率に追加するエビデンス (evidence) の量を表します。各スライスのサイズが同じである場合は、個々の属性値には関係なく同じ量のエビデンスがすべてのクラスに追加されます。

グラフの各スライスは、クラスラベル L を前提として属性値 A が現れる条件付き確率を表します。すなわち、条件付き確率 P (A|L) は、ラベルが L であるレコードからランダムに選択したレコードが属性値 A を取る確率です。デフォルト設定では、確率はレコードの重みに基づいて計算されます。たとえば、P（0.75 < 花びらの横幅 < 1.65 | アイリスバージカラー (0.75 < petal width < 1.65 | iris-versicolor) ）は、ラベルがアイリスバージカラー (iris-versicolor) である 36 件のレコードのうち花びらの横幅 (petal width) が条件の範囲に合致するレコードが 33 件存在するため、33/36 = 91.6% となります。

ラベルを選択すると、左側のメインウィンドウに棒グラフが表示されます。個々のバーの高さは、選択されたラベルのエビデンスの値を表します。エビデンスの対象は -log[1-P(L)/sum(P(Li))] という表現によって計算され、エビデンスの対象外は -log[P(L)/sum(P(Li))] という表現によって計算されます。「最小因子の排除」オプションのチェックマークを外すと、小さいバーと同様にベースに影響します。

各属性の寄与率は、その属性がラベル値の予測に寄与する度合いを表す尺度です。メインウィンドウを見ると、ラベル値に影響を与える各属性値の寄与率の他に、特定の属性値自身の寄与率も確認することができます。

エビデンス・クラシファイアの作成

エビデンス・クラシファイアを作成するときは、レコード数（重み）に基づいて確率を計算する処理を繰り返します。エビデンス・クラシファイアはデータに基づいて自動的に作成されます。複数のレコードと各レコードに対応するラベルから構成されるデータは、訓練事例と呼ばれます。

確率は次の手順に従って計算されます。

1. 連続的な属性は複数の離散的な範囲に階級生成されます。その際、各範囲でのクラス分布ができるだけ異なるように考慮されます。範囲の数は自動的に決定されます。マルチプロセッサ版の MineSet がインストールされている場合は、並列処理が起動されて複数の属性が並列的に階級生成されます。Tool Manager を使用して手作業で属性を階級生成すれば、自動的な階級生成を変更することができます。
2. 訓練事例内で特定のクラスラベルを持つレコードの数（重み）をレコードの総数（重み合計）で割って、各クラスラベルの事前確率が計算されます。
3. 訓練事例内で特定の属性値を持つレコードの数（重み）を特定のクラスに属するレコードの数（重み）で除して、各属性値の条件付き確率が計算されます（グラフでは各属性値に基づいて正規化された確率値が表示されます）。

1 行に表示されるグラフの数は、分析によって生成される離散範囲の数に一致します。範囲が 1 つしか存在しない場合、その属性はラベルの予測に効果がなかったことを意味します。最初は、ラベルの事前確率が右側のラベル確率パネルに表示されます。

個々の属性値（連続的な属性の場合は階級生成された一定範囲の属性値）は正確に 1 つのグラフに対応しており、各グラフによって各階級におけるラベルの条件付き確率が表されます。

- ラプラス補正

  このオプションを選択すると、個々の確率値が平均値に近づくように補正されるため、極端な値の確率（0 や 1 など）が避けられます。「エビデンス分析詳細オプション」ダイアログ・ボックス（IRIX システム上では「詳細分析オプション (*Further Inducer Options*)」ダイアログ・ボックス）の「ラプラス補正」にチェックマークを付けて、「係数」フィールドに何も入力しないか 0 を入力すると、"1.0/ 訓練事例の重み " を係数として自動ラプラス補正が適用されます。「ラプラス補正」（117 ページ）も参照してください。

- 階級での重み付け下限値

  エビデンス分析では、連続的な属性が離散的な範囲に階級生成されます。このオプションを使用すると、1 つの範囲に入れる重みの下限を設定することができます。「自動」を選択すると、データセットのサイズに基づいて重みの下限が自動的に設定されます。一般的に、データセットが大きくなるほど、1 つの階級の最小レコード数（重み）が増加し、階級の幅が小さくなります。データセットが非常に大きい場合は、必要な数を越える階級が作成される場合があります。階級の数を減らすには、最小重みを増加させてください。

- 項目の自動選択

  このオプションを使用すると、ラベルの予想に役立つ項目（属性）だけが自動的に選択されます。余分な項目を使用するとエビデンス・クラシファイアの予想精度が低下するため、クラシファイアに効果のある項目が自動的に検出されます。ただし、項目の数が多い場合は、実行時間が非常に長くなる可能性があります。このオプションは、予想精度低下の原因となる相関の高い項目を除外するのに効果的です。処理が終わらなくなる可能性があるため、「項目の自動選択」オプションと「ブースト」を同時に使用することはできません。

「項目の自動選択」オプションでは、クラシファイアの誤差を低減させる項目の集合を検出するために、ラッパー法に基づいて様々な属性集合の誤差が推定されます。クラシファイアの誤差は相互検証法によって推定され、誤差推定率に基づいて項目の追加と除外が行われます。デフォルト設定では空集合から項目の検索が始まりますが、「逆方向」オプションを選択すると、最初に大きいモデルが構築されて項目の全集合から検索が始まるため、実行速度がかなり遅くなります。（『*MineSet 3.0 Enterprise Edition Interface Guide*』の「Format of the Evidence Visualizer’s Data File」を参照）

エビデンス・ビジュアライザの起動

エビデンス・ビジュアライザ (Evidence Visualizer) を起動するには次の 4 通りの方法があります。

- Tool Manager の「クラス判別」タブからエビデンス分析を実行します。分析によってクラシファイアが構築されると、エビデンス・ビジュアライザが自動的に起動されます。Tool Manager を通じてエビデンス・ビジュアライザを操作する方法については、下記を参照してください。
- Tool Manager の「可視化ツール」メニューを使用して、エビデンス・ビジュアライザを起動します。
- 使用する設定ファイルが分かっている場合は、その設定ファイルのダイヤモンド形のアイコンをダブルクリックします。こうするとエビデンス・ビジュアライザが起動され、選択した設定ファイルが自動的に読み込まれます。この起動方法を利用できるのは、設定ファイルの拡張子が *.eviviz* である場合に限られます（Tool Manager を使用してエビデンス・ビジュアライザ用の設定ファイルを作成すると、ファイル拡張子が常に *.eviviz* になります）。
- IRIX シェルウィンドウのプロンプトに次のコマンドを入力して、エビデンス・ビジュアライザを起動します。

eviviz [*configFile*]

configFile は任意指定の引数であり、設定ファイルの名前を表します。コマンド行で設定ファイルを指定しなかった場合は、「ファイル」->「開く」オプションを選択してファイル名を指定する必要があります。

IRIX システム上でエビデンス・ビジュアライザを起動するときのオプション

IRIX システム上でビジュアライザの起動時に **-quiet** オプションを指定すると、進行状況を表すダイアログが表示されなくなります。このオプションを常に有効にするには、各ユーザのホーム・ディレクトリ内の *.Xdefaults* ファイルに次の行を追加します。

```
*minesetQuiet:TRUE
```

「警告オプション」（219 ページ）も参照してください。

Windows システム上では、同様のオプションを「3D ビジュアライザ設定」パネルで設定することができます。

エビデンス分析のオプション

「詳細分析オプション」（IRIX システム上では「詳細分析オプション (Further Inducer Options)」）を選択すると、「分析オプション」ダイアログ・ボックスが表示されます。このダイアログ・ボックスには次の 3 つのパネルがあります。

- 一番上のパネルには、Tool Manager の「データの可視化 / マイニング」パネルで設定した項目が表示されます。「誤差推定のオプション」のタイプはモデルに応じて決まります。
- 左下のパネルでは、詳細分析オプション（下記参照）を設定することができます。
- 右下のパネルでは、誤差推定オプションを設定することができます。ただし、「データの可視化 / マイニング」パネルで「クラシファイアのみ」モードを選択した場合、このパネルには何も表示されません。このパネルに表示されるオプションは、選択した誤差推定法に応じて異なります（「誤差推定」（87 ページ）を参照）。

「詳細分析オプション」（Windows システム）（IRIX システムの場合は「詳細分析オプション (Further Inducer Options)」）で損失マトリックスが指定されている場合（デフォルト設定）は、「損失マトリックスを使用」というタイトルのボタンが円グラフの右下に表示されます。損失マトリックスは、円グラフで示される確率を調整するために使用されます。損失マトリックスを使用しない場合の確率を表示するには、「損失マト

リックスを使用」ボタンの選択を解除します。損失マトリックスを編集するときに NULL を予測するための項目を指定した場合は、灰色のスライスが表示されることがあります。灰色のスライスが最大のスライスである場合は、NULL がクラシファイアによる予測結果です。詳細については、「損失マトリックス」（121 ページ）を参照してください。

エビデンス・ビジュアライザのメインウィンドウ内でのオブジェクトの選択

選択モード では、矢印の形をしたカーソルが表示されます。カーソルをオブジェクト（円グラフまたは棒グラフ）の上に移動するとオブジェクトが強調表示され、そのオブジェクトに関する情報がメインウィンドウの上に表示されます。この情報はカーソルがオブジェクトの上にある限り表示されます。

- オブジェクトが円グラフである場合は、次のような形式のメッセージが表示されます。

  < 属性名 >: < 値または 範囲 >
  weight = < 重み >

  < 重み > は、特定範囲の属性値または単一の属性値を取るデータ・ポイントの重みの合計です。グラフの高さは、この重み値に比例します。重みが設定されていない場合、< 重み > はレコード数を表します。

- オブジェクトが棒グラフである場合は、次のような形式のメッセージが表示されます。

  (<属性名> = <値>) ==> Prob(<選択されたラベル>) = x% [low%-high%]
  Evidence=z
  < 選択されたラベル > ==> Prob(< 属性 > = < 値 >) = y% [low%-high%]
  weight = < 重み >

  x% は強調表示された属性値を持つレコードが選択されたラベル値を取る確率です。中カッコで囲まれた範囲 [low%-high%] は、95% の信頼区間を表します。同様に、y% は選択されたラベル値を持つレコードが強調表示された属性値を取る確率です（図 13-14 参照）。バーの高さは確率ではなくエビデンス (Evidence) を表すことに注意してください。すなわち、"Evidence=z" はバーの高さに直接対応しています。エビデンスを合計すると（確率の場合は乗算する必要があります）、予測結果となるクラスを判定することができます。< 重み > は、強調表示された属性値を取るデータ・ポイントの重みです。

「エビデンスの対象」は、次式によって計算されます。

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

「エビデンスの対象外」は、次式によって計算されます。

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

A は属性値、L は選択されたラベル値、N はラベル値の数です。バーの高さを計算するときは、バーが無限に高くなることを避けるために、非常に小さい値が上記のカッコ内部の式に加算されます。メインウィンドウ内の「対象」または「対象外」という単語の周囲にはボックスがあり、各単語をクリックできるようになっています。ボックスをクリックすると、「対象」または「対象外」が切替わります。

バーのベースになっているグレー（灰色）の矩形ベースの高さは、事前確率によって追加されるエビデンスの量を表します。たとえば、ラベルが車のシリンダ数 (cylinders) である場合、3 シリンダ（気筒）の車は非常に少ないため、エビデンスの対象の表示中はベースが低くなり、エビデンスの対象外の表示中はベースが高くなります。ベースの高さに個々のバーの高さを加算することができます。

特定のラベル値の予測に最も効果的な属性値を判断するときは、エビデンスの対象を使用してください。

上記の式から分かる通り、エビデンス (Evidence) の量（バーの高さ）は強調表示された属性値またはラベル値の確率ではなく、他のすべてのラベル値の確率に対する条件付き確率に応じて決まります。

カーソルが属性値の上にあるときにマウスの左ボタンをクリックすると、任意の数の属性値を選択することができます。選択された属性値に応じて右側のラベル確率パネル内の円グラフが変わり、それらの属性値を前提とした場合の事後確率が表示されます。ただし、クラスの表示順序は変わらず、最大のスライスのクラスが右側のリストの一番上に表示されます。エビデンス・ビジュアライザでは新しい事後確率の分布を表示するために、各属性の条件付き確率を乗算した後、その結果に事前確率を乗算し、1.0 を基準値とする正規化を行っています。

このような確率の乗算は、個々の属性が互いに独立であるという仮定に基づいています。この仮定が満たされないときに複数の属性値を選択すると、最終的なクラシファイアが適正であっても、クラスの予想確率が不正な値になる可能性があります。分析の実行時にステータス・ウィンドウに表示される誤差推定値を見れば、この仮定が妥当かどうかを判断することができます。誤差率 / 損失が低い値であれば、仮定は満たされていると推測できます。

ラベル確率パネル内で特定のラベルを選択すると、各属性値のケーキまたは円ではなく、バーがメインウィンドウ に表示されます。バーの上には「エビデンスの対象」というタイトルが表示され、「対象」の回りのボックスがクリック可能な状態になります。

タイトル「エビデンスの対象」の「対象」の部分をクリックすると、「対象」の代わりに「対象外」と表示され、ラベルに対するエビデンスの対象外を表すようにバーの高さが変更されます。

ケーキまたは円を選択する場合と同様に、左側のバーを選択すると、右側のラベル確率パネルの確率分布が変わります。バーの高さは、選択されたラベルに対するエビデンスの対象 / 対象外を表します。エビデンスの計算では対数確率が使用されているため、バーの高さを加算することによってエビデンスが累計されます（確率の場合は乗算する必要があります）。

エビデンス・ビジュアライザのメニュー

エビデンス・ビジュアライザでは、5 種類のプルダウン・メニュー（「ファイル」、「表示」、「選択」、「属性値の順序付け」、「ヘルプ」）を通じて、さまざまな機能を利用することができます。設定ファイルの名前を指定しないでエビデンス・ビジュアライザを起動した場合は、「ファイル 」メニューと「ヘルプ 」メニューだけを使用することができます。

「表示 」メニュー

「表示」メニューでは、エビデンス・ビジュアライザのウィンドウに表示される情報を制御することができます。一般的なメニュー項目の詳細については、「「表示」メニュー」（204 ページ）を参照してください。各プラットフォーム（Windows または IRIX）に固有のオプションもいくつか用意されています。

- 「エビデンス」－エビデンスと確率の表示を切替えます。このオプションにチェックマークを付けると、左側のウィンドウにエビデンスが表示されます。チェックマークを外すと、確率が表示されます。

- 「寄与率でソート」— 選択されたラベルをクラス判別するときの効用（寄与率）に従って属性をソートして表示します。このオプションを選択しない場合は、Tool Manager の「現在のデータセットの項目名 」と同じ順序で各属性が表示されます。
- 「最小因子の排除 」— このオプションが適用されるのは、ラベルが選択されて、バーが表示されている場合に限られます。このオプションをオン（デフォルト）に設定すると、すべてのラベルに共通の（各ラベルのエビデンスのうち最小の）エビデンスが差し引かれます。差し引かれる値は各属性ごとに異なる可能性がありますが、特定の属性値については、ラベル値全体を通じて一定になります。このオプションがオンの場合は、すべてのラベル値の最小公倍数が差し引かれるために、小さい誤差が拡大されることになります。
- 「ラプラス補正の使用」— ラプラス補正のオン / オフを切替えます。特定のラプラス補正値が「詳細分析オプション」ダイアログ・ボックスで指定されている場合は、その値が使用されます。それ以外の場合は、デフォルト値が使用されます。
- 「ランドスケープビューワの使用」— 別の 3D ナビゲーション・モードに切替えます。

「属性値の順序付け 」メニュー

「属性値の順序付け」メニューでは、離散属性の値をどのようにソートするかを制御することができます。メニュー項目の詳細については、「属性値の順序付け」メニュー」（140 ページ）を参照してください。

「選択 」メニュー

「選択」メニューを使用すると、ドリルスルーによって元のデータを細かく分析することができます。ドリルスルーを行うには、最初に特定の属性値またはクラス（あるいはその両方）を選択した後、2 通りのドリルスルー方法のうちの 1 つを選択します。このメニューは複数のツール間で似通っています。メニュー項目の詳細については、「「選択 」メニュー」（167 ページ）を参照してください。

「ファイル」メニュー

ほとんどのビジュアライザの「ファイル」メニューは似通っており、次のようなオプションがあります。

Windows システム

- 「開く」は、設定ファイルを開き、その内容をメインウィンドウに表示します。これまで表示されていたデータは表示されなくなります。新しいデータセットを表示するとき、または既存のデータセットの設定を変更して表示するときは、「開く」を選択してください。
- 「再度開く」は、現在のファイルを再度開きます。設定ファイルまたはデータファイルを更新したときは、このオプションを選択してください。
- 「表示イメージの保存」は、ビジュアライザ・ウィンドウの現在の状態をイメージ・ファイルに保存します。（ボタン類を含む）ウィンドウ全体を保存するか、またはグラフィカル・オブジェクトを含むメインのシーンだけを保存するかを指定することもできます（デフォルトはウィンドウ全体）。
- 「表示イメージの印刷」は、ビジュアライザ・ウィンドウの現在の状態をプリンタに出力します。出力先のプリンタは「表示イメージの印刷」ダイアログ・ボックスで指定することができます（デフォルトはシステムのデフォルトのプリンタ）。「名前を付けて保存」ダイアログと同様に、ウィンドウ全体を印刷するか、またはメインのシーンだけを印刷するかを指定することができます。
- 「印刷プレビュー」は、印刷出力のイメージを表示します。
- 「ページ設定」は、出力先のプリンタ（デフォルトはシステムのデフォルトのプリンタ）と印刷出力のオプションを設定するためのダイアログ・ボックスを開きます。
- 「Web 公開用ファイルの作成」は、現在のファイルを Web 上で公開可能な形式で書出します。
- 「設定」は、「設定」ダイアログ・ボックスを開き、マウスの動作の設定や、コマンド実行時の警告のオン / オフ切替えなどを行えるようにします。ビジュアライザのサウンド効果、UI アニメーション、デフォルトのフォント・サイズなども指定することができます。
- 「Tool Manager の起動」は、Tool Manager を起動します。Tool Manager が既に起動されているときは、ビジュアライザの起動時の初期状態に戻ります。

- 「最近使ったファイル」は、任意のビジュアライザで最近使用されていたファイルを開きます。メニュー リストには、最近使用された 4 個のファイルが表示されます。
- 「終了」はすべてのウィンドウを閉じて、MineSet アプリケーションを終了します。

IRIX システム

- 「開く (*Open*)」は、設定ファイルを開き、その内容をメインウィンドウに表示します。これまで表示されていたデータは表示されなくなります。新しいデータセットを表示するとき、または既存のデータセットの設定を変更して表示するときは、「開く (*Open*)」を選択してください。
- 「再度開く (*Reopen*)」は、現在のファイルを再度開きます。設定ファイルまたはデータファイルを更新したときは、このオプションを選択してください。
- 「名前を付けて保存 (*Save As*)」は、ビジュアライザ・ウィンドウの現在の状態をイメージ・ファイルに保存します。ユーザはファイル名（ツリー・ビジュアライザのデフォルトは *treeviz.rgb*）とフォーマット（デフォルトは *rgb*）を指定できる他、（ボタン類を含む）ウィンドウ全体を保存するか、またはグラフィカル・オブジェクトを含むメインのシーンだけを保存するかを指定することもできます（デフォルトはウィンドウ全体）。
- 「表示イメージの印刷 (*Print Image*)」は、ビジュアライザ・ウィンドウの現在の状態をプリンタに出力します。出力先のプリンタは「印刷 (Print)」ダイアログ・ボックスで指定することができます（デフォルトはシステムのデフォルトのプリンタ）。「名前を付けて保存 (Save As) 」ダイアログと同様に、ウィンドウ全体を印刷するか、またはメインのシーンだけを印刷するかを指定することができます。
- 「Web 公開用ファイルの作成 (*Publish on the Web*)」は、Web 上での公開に適したフォーマットのファイル (*.mtr*) としてビジュアライザ・ウィンドウを保存します。
- 「Tool Manager の起動 (*Start Tool Manager*)」は、Tool Manager を起動します。Tool Manager が既に起動されているときは、ビジュアライザの起動時の初期状態に戻ります。
- 「終了 (Exit) 」は、すべてのウィンドウが閉じて、アプリケーションが終了します。

必要なファイル

MineSet のほとんどのビジュアライザでは、設定ファイルのほかに、少なくとも 2 つのファイル（データファイル (*.data*) とスキーマファイル (*.schema*)）が必要です。Tool Manager を通じてビジュアライザを起動すると、データファイルとスキーマファイルが自動的に作成されます。

データファイル (*.data*) は、タブで区切られた複数のフィールド（項目）を含む行のリストから構成されます。データファイルを作成するには、データソース（Oracle、INFORMIX、Sybase などのデータベース）からデータを抽出し、任意のテキストエディタ（メモ帳、jot、vi、Emacs など）を使用してフォーマットを整える方法もあります。ファイル・フォーマットの詳細については、『*MineSet 3.0 Enterprise Edition Interface Guide*』を参照してください。Tool Manager を通じてビジュアライザを起動したときに作成される各ファイルの拡張子を表 1-10 に示します。

表 1-10 デフォルトのファイル拡張子

ツール	データファイルの拡張子	スキーマファイルの拡張子	設定ファイルの拡張子
相関規則分析	*.rules.data*	*.rules.schema*	*.rules.scatterviz*
クラスタ・ビジュアライザ	なし	なし	*.clusterviz*
デシジョン・テーブル・ビジュアライザ	*.dtableviz.data*	なし	*.dtableviz*
エビデンス分析	なし	なし	*.eviviz*
マップ・ビジュアライザ	*.mapviz.data*	*.mapviz.schema*	*.mapviz*
レコードビューワ	*.data*	*.schema*	なし
スキャタ・ビジュアライザ	*.scatterviz.data*	*.scatterviz.schema*	*.scatterviz*
スプラット・ビジュアライザ	*.splatviz.data*	*.splatviz.schema*	*.splatviz*
統計量ビジュアライザ	なし	なし	*.statviz*
ツリー・ビジュアライザ	*.treeviz.data*	*.treeviz.schema*	*.treeviz*

ビジュアライザを起動するとき、またはファイルを開くときには、データファイルまたはスキーマファイルではなく設定ファイルを指定する必要があります。ただし、レコードビューワは、任意のスキーマファイル (*.schema*) を開くことができます。

データファイルには、ユーザが任意の拡張子を付けることができます（MineSet のサンプルファイルの拡張子は *.data* です）。

「フィルタ」ボタン

Tool Manager の「データ変換」パネル上にある「フィルタ」ボタンを使用すると、数学的な表現に基づいてデータをフィルタリングすることができます。フィルタリングを行うと、表現が真（数値の場合は非ゼロ）となるレコードだけを含むテーブルが生成されます。「フィルタ」ボタンをクリックすると、「フィルタ」ダイアログ・ボックスが表示されます。

このダイアログ・ボックスの左側のパネルで項目名と演算子を選択すると、右側のパネルに式が生成されます。式の記述方法の詳細については、「項目の追加」（1 ページ）を参照してください。

「フィルタ」パネル

「表示」メニューから「フィルタパネル」を選択すると、「フィルタ」パネルが表示されます。ここでは、「フィルタ」パネルのオプションについて説明します。「フィルタ」パネルは、スキャタ・ビジュアライザ、スプラット・ビジュアライザ、マップ・ビジュアライザ、デシジョン・テーブル・ビジュアライザで使用されます。ツリー・ビジュアライザでは類似したフィルタパネルが使用され、エビデンス・ビジュアライザでは独自のフィルタパネルが使用されます。

「フィルタ」パネルでは、特定のフィルタリング条件を指定して、メインの表示領域に表示するデータの数を減らすことができます。このパネルは、表示内容を調整したい場合、特定の情報だけを強調したい場合、または表示する情報量を単に減らしたい場合などに使用することができます。

「フィルタのスケール」オプション（スキャタ・ビジュアライザのみで使用可能）では、グラフィック・オブジェクトの高さをスケーリングするときの基準として、「データセット全体」または「フィルタリングされたデータのみ」を選択することができます。

「フィルタ 」パネルには 2 つのウィンドウがあります。上段のウィンドウでは、文字列型変数に基づいてフィルタリングを実行することができます。変数のすべての値を選択する場合は、「全てに設定 」ボタンをクリックします。選択を取消す場合は、「消去」ボタンをクリックします。値をクリックすると選択され、もう一度クリックすると選択が解除されます。

下段のウィンドウでは、文字列型変数と数値型変数の値に基づいてフィルタリングを実行することができます。フィルタリングに使用できるのは、スライダを動かしても値が変化しない変数に限られます。ウィンドウをスクロールすると、階層関係（含む、等しい、一致、NULL 値）が表示されます。

数値をフィルタリングする場合は、値を入力してから、関係演算子 (=、!=、>、<、>=、<=) を指定します。アルファベットの文字列をフィルタリングする場合は、対象となる文字列を入力します。文字列の検索では、次の 3 種類の条件を指定することができます。

- 「含む」は、指定された文字列を含むことを示します。たとえば、California は文字列 Cal と forn を含んでいます。
- 「等しい」は、文字列が正確に一致することを示します。
- 「一致」では、次のワイルドカードを使用することができます。
  - アスタリスク (*) は、任意の数の文字を表します。
  - 疑問符 (?) は、任意の 1 文字を表します。
  - 角かっこ ([ ]) は、その中に囲まれた文字の 1 つを表します。

たとえば、California は、Cal*、Cal?fornia、Cal[a-z]fornia に一致します。

場合によっては（特に Tool Manager の階級生成機能を使用している場合）、テキスト・フィールドの代わりに、値のオプションメニューが表示されます。このような変数を無視する場合は、オプションメニューで「無視」を選択します。これらのオプション

については、関係演算子（>= など）を使用することができます。こうすれば、指定した値だけでなく、それに後続する値も選択されます。

数値と文字列の比較演算子の他に、`Is Null` という演算子を使用することができます。この演算子は、値が NULL の場合に TRUE となります。

各フィールドの右側には追加のオプション・メニューが用意されており、「論理積 (And)」または「論理和 (Or)」を指定することができます。たとえば、フィルタリング基準として、"sales > 20 And < 40" を指定することができます。特定の変数について任意の数の And と Or を指定できますが、同一変数内で And と Or を混在させることはできません。

「適用」ボタンをクリックすると、フィルタリングが開始されます。フィルタパネルがアクティブなときに *Enter* キーを押すと、フィルタリングが自動的に開始されます。

増加比率

増加比率は相互情報量を分岐のエントロピーで除算した値（ラベル値は無視します）であり、決定木を分岐するときの条件として使用されます。決定木の分岐条件については、「決定木」（77 ページ）を参照してください。

IRIX システムの「ヘルプ (Help)」メニュー（Help （IRIX））

IRIX システムの「ヘルプ (Help)」メニュー（F1 キーのみまたは Shift キーを押しながら F1 キーを押しても表示）には下記の 5 つのオプションがあります。

- 「クリックでヘルプを表示 (*Click for Help*)」を選択すると、カーソルが疑問符 (?) の形に変わります。ビジュアライザのメインウィンドウ内にある任意のオブジェクト上に疑問符カーソルを移動し、マウスのボタンをクリックすると、そのオブジェクトに関するヘルプウィンドウが表示されます。ヘルプウィンドウを閉じると、カーソルは矢印の形に戻り、ヘルプは機能しなくなります。この機能のキーボード・ショートカットは <Shift+F1> です（矢印カーソルをオブジェクト上に置いて F1 キーを押しても同じです）。
- 「概要 (*Overview*)」— ビジュアライザの主な機能や特徴に関する簡単な説明（ファイルを開く方法や表示内容の操作方法など）を表示します。
- 「索引 (*Index*)」— ヘルプの全項目の索引を表示します。このオプションは現在のところ使用できません。

- 「キーとショートカット (*Keys & Shortcuts*)」— ビジュアライザの全機能のうち、アクセラレータ・キーが定義されている機能のキーボード・ショートカットを表示します。
- 「製品情報 (*Product Information*)」— ビジュアライザのバージョンと著作権を表示します。

Windows システムの「ヘルプ」メニュー（Help （Windows））

Windows システムの「ヘルプ」メニュー（(F1 キーを押しても表示）では、上記のオプションのほかに、MineSet 3.0 Enterprise Edition のすべてのマニュアルを参照することができます。また「製品情報 」には、ビジュアライザのバージョンと著作権を示す画面が表示されます。

ヒストグラム・ビジュアライザ

ヒストグラム・ビジュアライザはデータ内の連続型項目を自動的に階級生成して、その階級生成結果を統計量ビジュアライザに送信します（「統計量ビジュアライザ」（184 ページ）を参照)。ヒストグラム・ビジュアライザには次のようなオプションがあります。

- 階級の数は、ユーザが明示的に指定するか、MineSet ソフトウェア内部で自動設定することができます。
- 「端数の切捨て」を指定することができます。このオプションを指定すると、階級が生成される前に、あらかじめ極端な値が排除されます。デフォルトの端数切捨ては 0.05 です。すなわち、すべての値のうち 5% の極端な値（下限の 2.5% と上限の 2.5%）が排除されます。このオプションは、しきい値を決めるときに、外れ値の影響を低減する効果があります。

履歴の表示

Tool Manager の「テーブルの履歴 」ボタンを使用すると、現在のセッションで実行した操作の履歴を表示することができます。詳細については、「「テーブルの履歴」ボタン」（187 ページ）を参照してください。「ファイル」メニューを使用すると、現在のセッションの内容を保存することができます。また、Tool Manager の「データの可視化 / マイニング」パネルの「データファイル」タブをクリックすると、変換作業が完

了した後のデータファイルを保存することができます。作業を再開するときは、該当のセッションまたはファイルを読み込むと、履歴が自動的に表示されます。

予備法

予備法では、全レコードの一定の比率（予備比率）を訓練事例として使用し、残りをテストセットとして使用します。分析によって訓練事例が評価され、クラシファイアが作成されます。作成されたクラシファイアを使用してテストセットをクラス判別し、そのテストセットに関する誤差率をクラシファイアの推定誤差率とみなします。

分析

分析は、ラベルを持つレコードからなる訓練事例に基づいて予測モデルを自動的に構築するアルゴリズムです。学習セットは分析時のデータの一部で分析モデルを構築する方法を“学びます”。分析モデルが作成されると、その構造を可視化したり、それをラベルのないレコードのクラス判別に使用することもできます。分析を適用するときは、CPU と I/O が大量に使用されるため、MineSet の分析は MineSet クライアントではなく MineSet サーバ上で実行されます（図 1-20 を参照）。

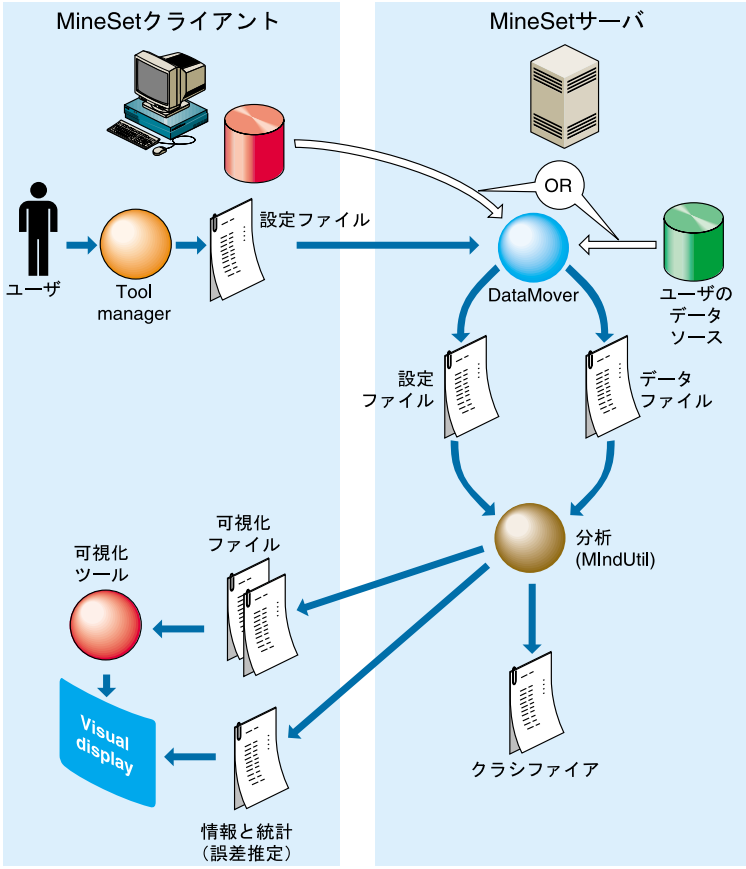

図 1-19

図 1-20　　分析を適用するときのツール実行環境

分析には訓練事例が必要です。訓練事例は複数の属性を含むテーブルであり、そのうち 1 つの属性がクラスラベルとして使用されます。分析モデルが構築されると、その分析モデルに基づいて新しいレコードを個々のクラスにクラス判別することができます。これらの新しいレコードは、分析モデルで使用されるすべての属性（名前と型は訓練事例内の属性と同じ）が収録されたテーブル内に存在しなければなりません。このテーブル内にラベル属性が存在する必要はありません。存在している場合、そのラベル属性は分析モデル処理では無視されます。

Tool Manager で分析を実行するときのモード

Tool Manager を通じて分析を実行するときは、次の 4 種類のモードを選択することができます。

- クラシファイアとエラー
- クラシファイアのみ
- 誤差推定
- 学習曲線

「クラシファイアとエラー」モードでは、予備法を使用してクラシファイアが構築されます。すなわち、全レコードの一定の比率（通常は 2/3）が訓練事例として使用され、残りのレコードがテストセット (test set) として使用されます。予備の比率は、「詳細設定」（Windows システム）（IRIX システムの場合は「詳細分析オプション (*Further Inducer Options*)」）で設定することができます（「誤差推定」（87 ページ）を参照）。このモードはデフォルト・モードであり、初期のテスト段階に適しています。このモードは高速であり、誤差が自動的に評価されます。

「クラシファイアのみ」モードでは、すべてのデータを使用してクラシファイアが構築されます。誤差推定は行われません。データ量が少ない場合、または最終的なクラシファイアを構築する場合に、このモードを選択してください。

「誤差推定」モードでは、すべてのデータを使用して構築されるクラシファイア（「クラシファイアのみ 」モードなど）の誤差が推定されます。このモードでは相互検証が実施されるため、実行に長い時間がかかります（「相互検証」（66 ページ）を参照）。データ量が少ない場合に、このモードを選択してください。帰納されるモデルは、「クラシファイアのみ」モードを選択した場合とまったく同じです。

「学習曲線」モードでは、訓練事例のサイズがクラシファイアの誤差に及ぼす影響が評価されます（「学習曲線」（118 ページ）を参照）。

分析における誤差の取扱い

分析の誤差推定を微調整するために、次のようなオプションが用意されています。利用できる誤差推定オプションは、選択するモードに応じて異なります。

「クラシファイアとエラー」モード（または「回帰ツリーとエラー」モード）および「誤差推定」モードでは、データを訓練事例とテストセット (test set) にどのように分割するかを決めるランダム・シードを設定することができます。ランダム・シードを変更すると、データを訓練事例とテストセットに分割するときの比率が変わります。分割比率に応じて誤差推定値が大きく変化する場合は、分析プロセスが不安定になります。

「クラシファイアとエラー」モード（または「回帰ツリーとエラー」モード）では、訓練事例として保持するレコードの比率（予備比率：Holdout Ratio）を指定することができます。デフォルト値は 0.666667 (2/3) です。残りのレコードは誤差を推定するために（テストセットとして）使用されます。

「誤差推定」モードでは、相互検証で使用するサブセットの数と検査の繰返し回数を指定することができます（「相互検証」（66 ページ）を参照）。

分析の詳細オプション

MineSet では、すべての分析について複数の詳細オプションが用意されています。これらの詳細オプションを使用すると、誤差のタイプごとに異なるコスト（損失）を割当てたり、一様でない標本抽出（母集団の特定部分から集中的に標本抽出を行う）を試行したりすることができます。また、計算時間が長い代わりに精度が高い複雑なクラシファイアを構築することもできます。

- バックフィッティング — 「クラシファイアとエラー」モードを使用するときは、すべての分析用の「詳細設定」（Windows システム）（IRIX システムの場合は「詳細分析オプション (*Further Inducer Options*)」）として、「テストセットのバックフィット」オプションを設定することができます。ブースティングが有効になっているときは、このオプションを設定できません。「バックフィッティング」（37 ページ）を参照してください。
- 混同マトリックス— 「クラシファイアとエラー」モードを使用するときは、すべての分析用の「詳細設定」（Windows システム）（IRIX システムの場合は「詳細分析オプション (*Further Inducer Options)*」）として、「混同マトリックスの表示」オプションを設定することができます。「混同マトリックス」（64 ページ）を参照してください。
- 投資利益率曲線 — ROI (Return on Investment) 曲線は改善曲線と似ていますが、損失マトリックスが考慮に入れられ、誤差の観点ではなく利益損失の観点から評価を表示します。すべての分析では「詳細設定」（Windows システム）（IRIX システムの場合は「詳細分析オプション (*Further Inducer Options*)」）として、「ROI 曲線の表示」オプションを設定することができます。ROI 曲線を表示するときは、ラベル

値を選択する必要があります。ラベル値を選択すると、そのラベル値に基づき、ROI 曲線が生成されて表示されます。「投資利益率 (ROI) 曲線」（157 ページ）を参照してください。

- 改善曲線 ― 改善曲線は、特定のラベル値の予測に対する寄与度に関して、ランダムに選択されたレコードと、クラシファイアの予測に基づいてソートされたレコード間の差異を示すグラフです。「改善曲線」（120 ページ）を参照してください。
- 損失マトリックス ― 損失マトリックスを作成すると、様々なタイプの誤りに対して異なるペナルティー（損失額）を割当てることができます。損失マトリックスと混同マトリックスを組み合わせて使用すると、誤りによる損失が最小限に抑えられます。すべての分析では「詳細設定」（Windows システム）（IRIX システムの場合は「詳細分析オプション (*Further Inducer Options*)」）として、「損失マトリックスの使用」オプションを設定することができます。「損失マトリックス」（121 ページ）を参照してください。
- 重みの設定― このオプションでは、各レコードの重みを指定することができます。たとえば、母集団のうち、2 倍の頻度で標本抽出される部分集合に 0.5 という重みを割当て、残りの部分に 1.0 という重みを割当てると、母集団全体が適切に調整されます。すべての分析の「詳細設定」（Windows システム）（IRIX システムの場合は「詳細分析オプション (*Further Inducer Options*)」）の下にあるダイアログ・ボックスの上部近く 「重み付けとして使用」オプションがあります。このオプションをオンに設定した場合は、重みとして使用する項目を選択してください。「属性としても使用」オプションをオンに設定した場合は、重みの項目が通常のデータ属性として分析モデルで使用されます。それ以外の場合は、重みの項目が重み値としてのみ使用されます。重みの項目が現実世界の要素でない場合は、重み値を分析モデルで使用しないでください。
- 学習曲線 ― 学習曲線は、分析によって作成される分析モデルの誤差を、その分析モデルの作成に使用されるレコード数の関数として表したグラフです。通常、分析モデルの作成に使用されるレコード数が多くなるほど、分析モデルの誤差は小さくなります。「学習曲線」（118 ページ）を参照してください。

制限事項

- 配列型の属性は常に無視されます。
- 日付型の属性は文字列とみなされます。離散属性の値の個数に関する制限のために、日付の数が少ない場合を除いて、日付属性は無視されます。したがって、日付を数個のクラスに階級生成してから分析を実行してください。

分析のステータス・ウィンドウ

「データの可視化 / マイニング」パネルの「実行」ボタンをクリックすると、Tool Manager のメインウィンドウの一番下にあるステータス・ウィンドウに、分析の進行状況と分析モデルに関する統計値が表示されます。このステータス・ウィンドウには、生成された分析モデルに固有の情報が表示されます。たとえば、決定木の場合ならば、ノードの数、リーフの数、決定木の深度などが表示されます。これらの情報は、"*.out*" という拡張子を持つセッション・ファイル名でワークステーション上に自動的に保存されます。

- 「 クラシファイアとエラー」モード（または「回帰とエラー」）では、最初の一連の点がファイルの読込みを表し、次に分析モデルの構築過程に関する情報が表示され、最後にテストセットの分析モデルの進行状況が表示されます。
- 「クラシファイアのみ」モードまたは「回帰ツールのみ」）では、テストセットの分析モデルの進行状況が表示されません。
- 「誤差推定」モードでは、相互検証のサブセットの数と繰返し回数に関する情報が表示されます。
- 「学習曲線」モードでは、X 軸上の平均ポイントが 1 行に表示され、その平均ポイント当たりの実行回数が点によって表されます。

「クラシファイアとエラー」モードを選択時のステータス・ウィンドウ

「クラシファイアとエラー」モード（または「回帰ツリーのみ」モード）を選択したときは、ステータス・ウィンドウに次の情報が表示されます。

- データを訓練事例とテストセットに分割するためのランダム・シード
- 分析モデルの作成に使用されるレコード数（訓練事例のレコード数）
- 分析モデルの評価に使用されるレコード数（テストセットのレコード数）
- 正しい予測と誤った予測の数

- 平均で正規化された平均平方誤差（予想確率の精度を表す値）― テストセット内の各レコードの平均平方誤差は、次の式によって計算されます。
  平均平方誤差 =（1 - 正しく予測されたラベル値の予想確率）2 乗 + Σ（(誤って予測されたラベル値の予想確率）2 乗）
  正規化された平均平方誤差は平均平方誤差を 2 で除算した値で、0 〜 1 の範囲を取ります。平均で正規化された平均平方誤差は、正規化された平均平方誤差をテストセット (test set) 内の全レコードによって加重平均した値です。
- クラシファイアでは、クラス判別誤差、すなわち、誤った予測の割合
- 平均平方誤差とクラス判別誤差（平均の標準偏差と平均の信頼区間を表す値）― 信頼区間は、データが同じ分布に従うという前提のもとでクラシファイアから得られる推定値の範囲です。誤差推定については、通常の 2 標準偏差規則よりも正確な公式が使用されます。

「誤差推定」モードを選択時のステータス・ウィンドウ

「誤差推定」モードを選択したときは、ステータス・ウィンドウに次の情報が表示されます。

- 相互検証のサブセットの数と繰返し回数
- ランダムシード
- 推定精度と標準偏差
- 推定精度に関する 95% 信頼区間

回帰分析に関して次の統計量が表示されます。

- 平均二乗誤差と平均絶対誤差の推定値
- 上記の推定値の精度に関する 95% 信頼区間

制限事項

- 配列型の属性は常に無視されます。
- 日付型の属性は文字列とみなされます。離散属性の値の個数に関する制限のために、日付の数が少ない場合を除いて、日付属性は無視されます。したがって、日付を数個のクラスに階級生成してから分析を実行してください。

国際化

Windows システム上では、コントロール・パネルの「地域」アイコンを使用してロケールを設定します。ここでは、IRIX システム上でのロケール設定について説明します。

バージョン 2.6 以降の MineSet では各言語のデータセットがサポートされます。グラフィカル・インタフェースの表記は現在英語になっていますが、2 バイトの項目名やデータ値をそれぞれデータ・エンコーディングで対応した言語で表示することができます。MineSet では、対応する言語がインストールされていれば、日本語、中国語、韓国語の EUC エンコーディングが自動的にサポートされます。これら以外の言語とエンコーディングに関しては、「他の言語とエンコーディングに対する拡張機能（IRIX 専用）」（114 ページ）を参照してください。

IRIX システム上でのロケールの設定

使用する言語のロケールとフォントがクライアント・システムとサーバ・システム、それにリモート表示に使用するシステムにそれぞれ存在していることが必要です。使用するシステムにインストールされているロケールのリストを確認するには、UNIX シェルプロンプトに次のコマンドを入力してください。

```
locale -a
```

ロケールを設定するには、上記コマンドによって生成されたリスト内にある適切なロケールに環境変数 LANG を設定します。たとえば、ロケールを日本語の EUC エンコーディングに設定するには、csh を使用して次のコマンドを入力します。

```
setenv LANG ja_JP.EUC
```

あとは、同じシェルウィンドウで MineSet を起動するだけです。すべてのアプリケーションに対してロケールを永続的に設定する方法については、IRIX のマニュアルを参照してください。

他の言語とエンコーディングに対する拡張機能（IRIX 専用）

インストール・イメージに含まれていないロケールで MineSet を実行する場合は、リソースファイルを適切なディレクトリにコピーして、それらのファイルを変更する必要があります。MineSet の各ビジュアライザでは、2D フォントと 3D フォントの両方で Open Inventor が使用されています。テキストを正しく表示するためには、Type III フォ

ント（CID アウトラインとも呼ばれています）をインストールしておく必要があります。

MineSet のインストール・イメージには、次のロケールのリソースファイルが含まれています。

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

MineSet をロケール *local_name* で実行する手順は次の通りです（インストールされているロケールのリストを表示する方法については、「IRIX システム上でのロケールの設定」（114 ページ）を参照してください）。

1. 通常の手順に従って MineSet をインストールします。
2. root でログインします。
3. 次のリソースファイルを */usr/lib/X11/app-defaults* から */usr/lib/X11/locale_name/app-defaults* にコピーします。
   - *Clusterviz*
   - *Dtableviz*
   - *Eviviz*
   - *Mapviz*
   - *Mineset*
   - *Scatterviz*
   - *Splatviz*
   - *Statviz*
   - *Treeviz*
4. */usr/lib/X11/locale_name/app-defaults* のリソースファイルを編集します。使用したいフォントのリソース名と仕様を把握しておく必要があります (表 1-11 に例を示します)。
5. ロケールを *local_name* に設定して、MineSet を起動します。

例 1-3 韓国語のリソースファイルの編集

韓国語を使用するのに必要な編集例を表 1-11 に示します。この例に示すフォントは次のファイル内のリストから入手したものです。

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

表 1-11 韓国語のリソースファイルの編集例

ファイル	英語のリソース（一部の行は折り返して表示）	韓国語のリソース（一部の行は折り返して表示）
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
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-*:

表 1-11（続き）韓国語のリソースファイルの編集例

ファイル	英語のリソース（一部の行は折り返して表示）	韓国語のリソース（一部の行は折り返して表示）
	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-*:

反復型 k-means（Iterative k-Means）

反復型 k-means は、MineSet のクラスタリング・ツールで使用されるクラスタリング方法です。このクラスタリング方法では、ユーザが指定した範囲（上限と下限）に従ってクラスタの数が自動的に選択されます（クラスタの数を事前に指定する必要はありません）。クラスタリングの複数の候補が生成された後、各クラスタの散らばりとユーザが指定する「ポイントの選択」パラメータに基づいて、最適なクラスタリングが選択されます。

ラプラス補正

エビデンス・ビジュアライザの分析プロセスでは、レコードの数（重み）に基づいて確率が計算されます。「ラプラス補正」オプションを確率の計算に適用すると、極端な値の確率（0 や 1 など）が避けられます。

ラプラス補正を適用するには、Tool Manager の「クラス判別」タブをクリックし、「分析」->「エビデンス」を選択します。「詳細設定 」ボタンをクリックしてダイアログ・ボックスを表示すると、「エビデンスオプション」の下に「ラプラス補正」オプションのチェックマークが表示されます。

ラプラス補正は、「AIDS 検査が陽性である被験者は瀕死の状態である」といった事象に対して、1.0 の確率ではなく 1.0 に非常に近い値を割当てたい場合に使用することができます。1 ではなく、1 に近い確率を割り当てることで、異常な標本値や誤差を正しく評価することができます。「ラプラス補正」オプションを使用すると、エビデンス・ビジュアライザの確率値が平均値に近づくように補正されるため、極端な値の確率（ 0 や 1 など）が避けられます。

したがって、メインウィンドウに表示される各クラスのスライスがゼロになることはありません。階級内のレコード数が少なくなるほど、大幅な補正が行われます。ラプラス補正を指定して（オプションにチェックマークを付けて）、「係数」フィールドに何も入力しないか 0 を入力すると、"1.0/ 訓練事例の重み " を係数として自動ラプラス補正が適用されます。

学習曲線

学習曲線は、分析によって作成される分析モデルの誤差を、その分析モデルの作成に使用されるレコード数の関数として表したグラフです。通常、分析モデルの作成に使用されるレコード数が多くなるほど、分析モデルの誤差は小さくなります。

学習曲線を作成するには、曲線上の各ポイントにおいて特定の数の分析モデルを作成します。個々の分析モデルはランダムに標本抽出されたレコードを使用して生成され、残りのレコード（分析の訓練事例に使用されないテストセット）を使用して分析モデルの誤差が評価されます。

学習曲線の生成には大量の CPU 時間が必要です。訓練事例 i（i は 1 からポイント数までの範囲の整数）に関する分析を学習するのにかかる時間を ti、1 ポイント当たりの実行回数を k とすると、学習曲線の生成にかかる総時間は次の式で表されます。

$$k*\sum_i t_i$$

1 ポイント当たりの実行回数を増やすと実行時間が比例して増加しますが、平均評価は改善されます。実行回数のデフォルト値は 3 です。

スキャタ・ビジュアライザのフィルタパネルを使用すると、表示されるデータ型（平均値、信頼区間、補間値、実験値）の一部をフィルタリングすることができます。たとえば、実験値と信頼区間のデータ・ポイントを非表示にして、平均値と補間値だけを表示することができます。

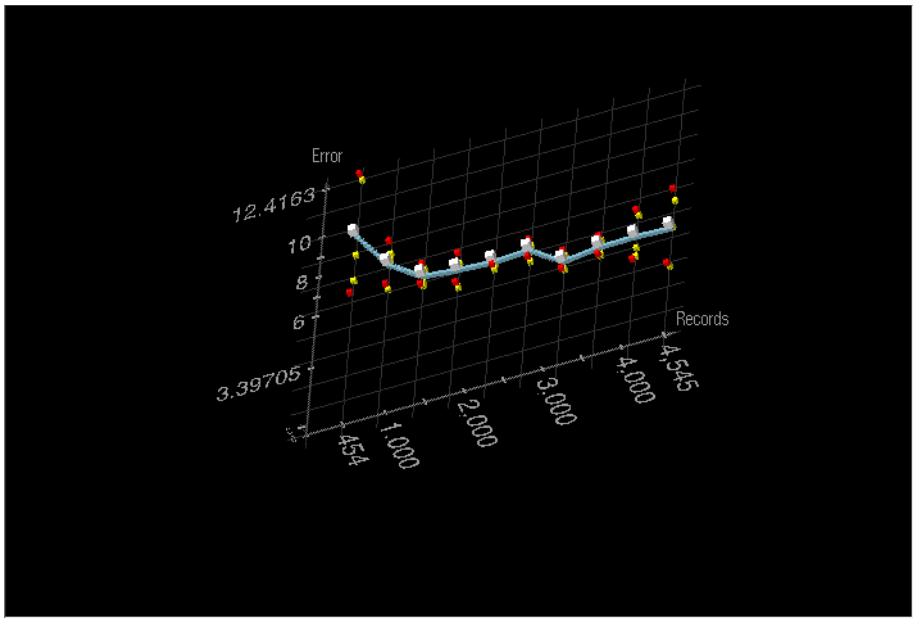

図 1-21 学習曲線

学習曲線は、「マイニングツール」タブの「クラス判別」メニュー（または「回帰」メニュー）で選択できるモードの 1 つであり、任意の分析で使用することができます。「学習曲線」モードを選択したときは、「詳細設定 」（IRIX システムでは「詳細オプション (*Further Options*)」）ダイアログ・ボックス内で次のような学習曲線オプションを指定することができます。

- 学習曲線上のポイントの数
- 1 ポイント当たりの実行回数
- 開始ポイントと終了ポイントで使用するレコード数

個々の中間ポイントで使用されるレコード数は自動的に計算されます。

学習曲線上のポイントの数は、1 以上の値を必ず指定しなければなりません。開始ポイントと終了ポイントのレコード数を指定すると、訓練事例の特定の範囲について学習曲線を生成することができます。これらのオプションを指定しない場合は、学習曲線上のポイント数と訓練事例内の全レコード数に基づいて、開始ポイントと終了ポイントのレコード数が自動的に計算されます。その場合は、訓練事例の全範囲が計算対象となります。たとえば、80,000 件のレコードが収録されたファイルがあるときに、学

習曲線上のポイント数を 3 と指定すると、20,000 番目、40,000 番目、60,000 番目のレコードにポイントが配置されます。通常、これらのオプションは狭い範囲をズームインするために使用します。たとえば、1000 ～ 10,000 番目のレコードだけを対象として学習曲線を生成することができます。

改善曲線

改善曲線は、特定のラベル値の予測に対する寄与度に関して、ランダムに選択されたレコードと、クラシファイアの予測に基づいてソートされたレコード間の差異を示すグラフです。たとえば、契約を解約しそうな顧客を予測し、それらの顧客に何らかの対策を実施する場合などには、改善曲線が役立ちます。

改善曲線の X 軸は標本抽出されたレコードの総数（0 ～ 100 %）を表し、Y 軸は特定のラベル値（この例では `Churn=yes` ）を持つレコードの数を表します。図 1-22 の 2 つの曲線を見てください。

下の曲線（赤）は、レコードをランダムに選択した場合に解約が予想される顧客数を示します。上の曲線（白）は、クラシファイアによる各レコードの予想確率に従ってレコードを選択した場合に解約が予想される顧客数を示します。解約する確率が最も高いと予測される顧客のレコードが最初に表示され、解約する確率が低いと予測される顧客のレコードが最後に表示されます。クラシファイアに基づく判定効果（改善率）は、上側の曲線と下側の曲線の差異として表されます。

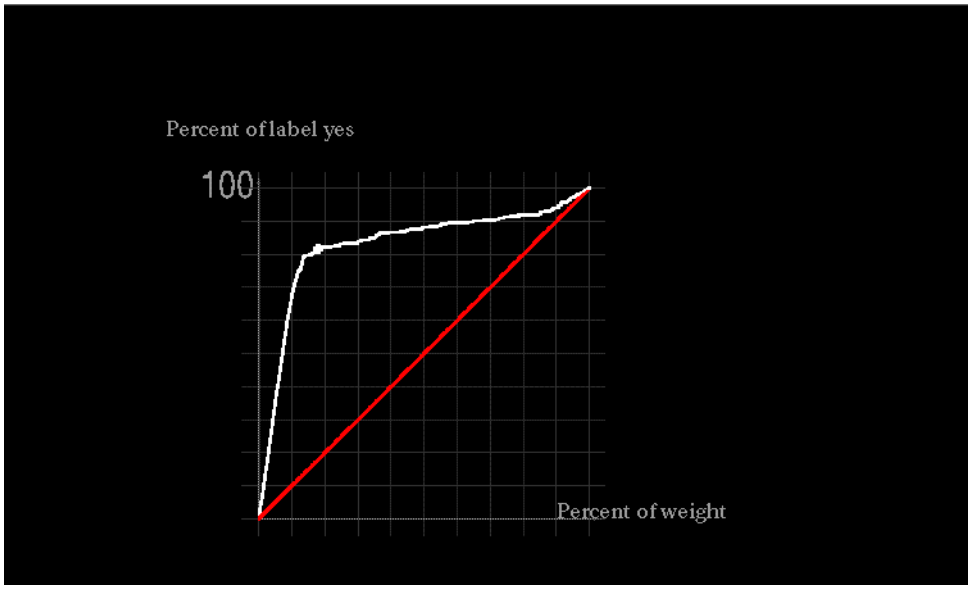

図 1-22 改善曲線

改善曲線を生成するときは、データセットの一定の比率を訓練事例としてクラシファイアを作成した後、データセットの残りの部分（テストセット）に対してクラシファイアを適用します。「クラシファイアとエラー」モードを使用するときは、すべての分析用の「詳細設定」（Windows システム）（IRIX システムの場合は「詳細分析オプション (*Further Inducer Options*)」）として、「改善曲線の表示」オプションを設定することができます。改善曲線を表示するときは、ラベル値を選択する必要があります。ラベル値を選択すると、そのラベル値に基づき、改善曲線が生成されて表示されます。

損失マトリックス

損失マトリックスの目的は、クラシファイアによって生成される誤りの種類を制御することです。通常、誤りの種類に応じて損失の程度が異なります。たとえば、キノコのデータセット (mushroom) を使用して、個々のキノコを毒性のあるものと食用になるものにクラス判別する例を考えます。実際には食用可能なキノコを毒性があるものとクラス判別すると、そのキノコを食べないために 2 ドル（捨てられるキノコの値段）の損失が発生しますが、毒性のあるキノコを食用になると判断して食べると、10,000

ドル（入院費用）の損失が発生します。クラシファイアで損失マトリックスを使用すると、このような高価な損失を避けることができます。

損失マトリックスは、混同マトリックスと組合わせて使用すると極めて効果的です。混同マトリックスでは、クラシファイアによって生成される誤りの種類と程度が詳しく解析されます。数多くの高価な誤りがクラシファイアによって生成されていることが判明した場合は、損失マトリックス内でそれらの誤りの重みを大きく設定すると、クラシファイアの精度が改善されます。損失マトリックスを使用する場合は、「詳細設定」（Windows システム）（IRIX システムの場合は「詳細分析オプション (Further Inducer Options)」）の「損失マトリックスの使用」オプションをオンに設定します。このオプションは、すべての分析で利用することができます。

混同マトリックスによる誤差推定プロセスを下記の例で説明します。図 1-23 に、キノコ (*mushroom*) データセットに関する混同マトリックスを示します。ここでは、データセット全体の 10% を訓練事例として、決定木分析を適用しています。

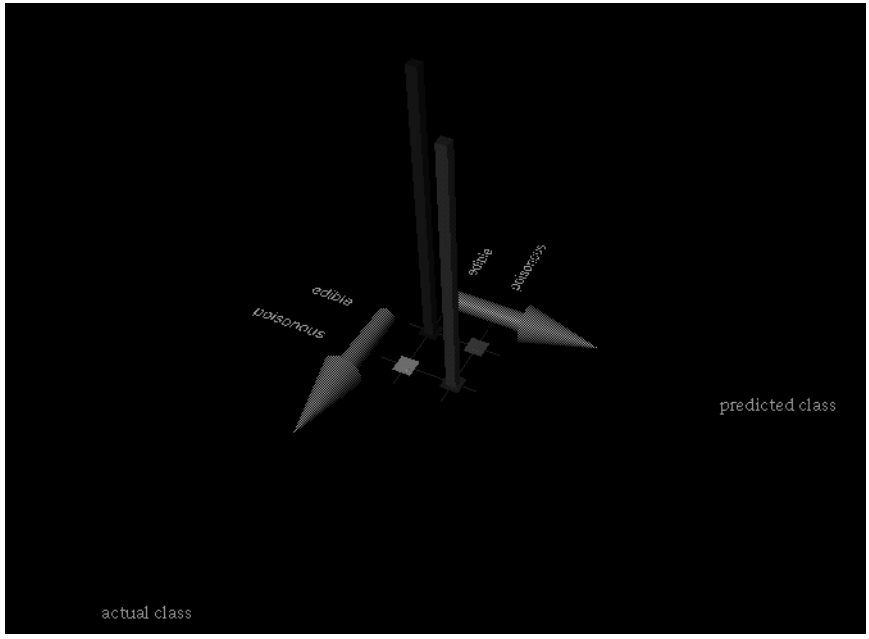

図 1-23 キノコ (mushroom) データセットに関する混同マトリックス（デフォルト設定）

毒性のあるキノコを表す 8 個のレコード (0.1%) が食用可能としてクラス判別され、食用可能なキノコを表す 15 個のレコード (0.2%) が毒性ありとクラス判別されています。

3,793 個の食用可能なキノコと 3,496 個の毒性のあるキノコは正しくクラス判別されています。クラシファイアの誤差率はわずか 0.31%（1.0% 未満）ですが、損失の予想額は、$10,000*8 + $2*15 = $80,030 と多額になります。

図 1-24 に、同じキノコ (mushroom) データセットに関する混同マトリックスを示します。ただし、今回は上記の損失額を反映した損失マトリックス (Loss Matrix) を使用して決定木分析を適用しています。新しいクラシファイアは非常に保守的であり、毒性のあるキノコを食用とクラス判別する誤りは 1 回も発生していませんが、食用可能なキノコを毒性ありとクラス判別する誤りが 1,558 回発生しています。その結果、損失の予想額は $10,000*0 + $2*1,558 = $3,116 となり、損失額の差異を考慮に入れなかった場合の約 3% に抑えられます。

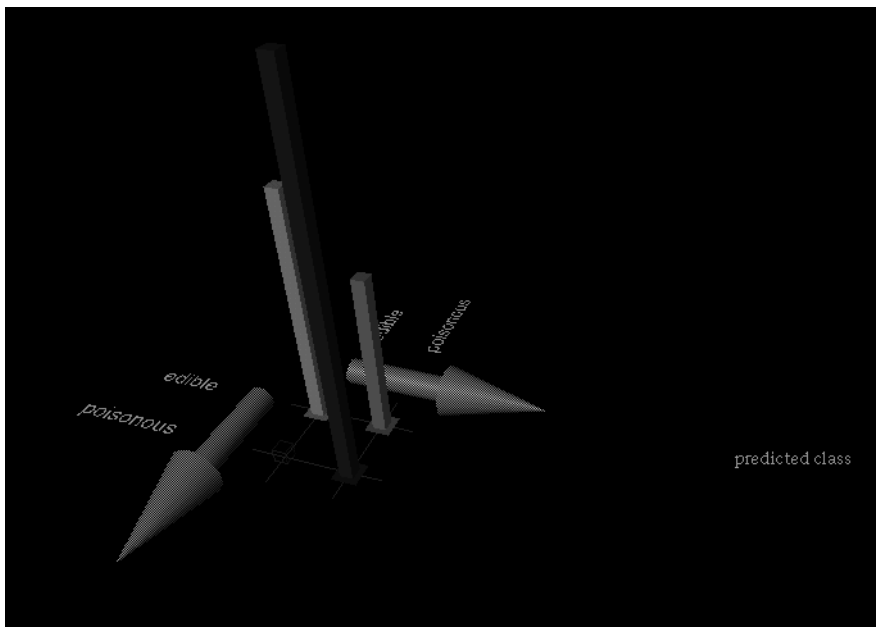

図 1-24 キノコ (Mushroom) データセットに関する混同マトリックス（損失マトリックスを使用）

損失マトリックスでは、疑問符 (?) で示される未知の値（NULL 値）を予測対象とすることもできます。たとえば、キノコに毒性があるかどうかを外部の専門家に判定してもらうのに $1 の費用がかかると想定した場合、一部のキノコは（食用でも毒性でもなく）NULL とクラス判別されます。図 1-25 に、決定木分析によって生成された混同マトリックスを示します。ここでは、1,551 個のキノコが NULL であるとクラス判別さ

れ、食用可能なキノコが毒性ありとクラス判別されたのは僅か 15 件に過ぎません。その結果、損失の予想額は $10,000*0 + $1*1,551 + $2*15 = $1,581 となります。

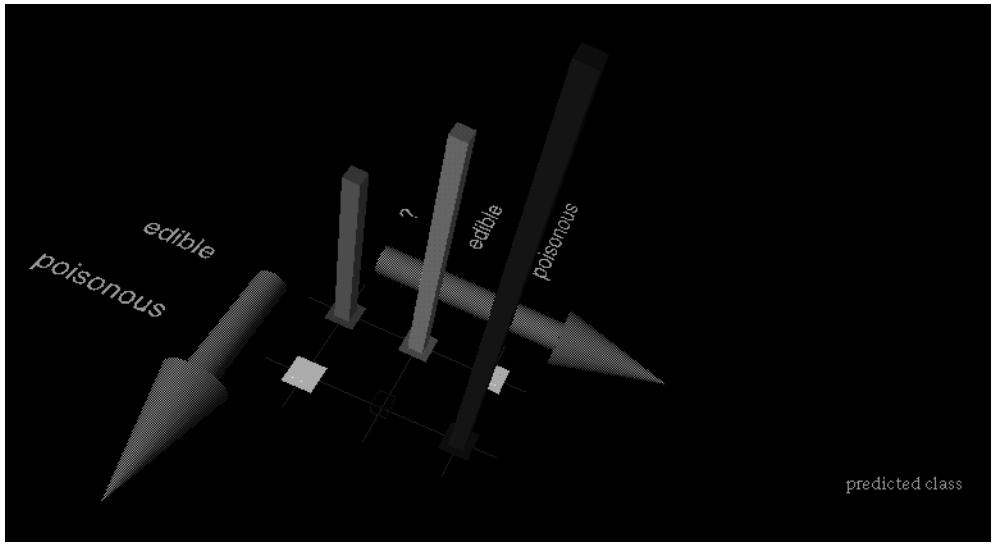

図 1-25 キノコ (Mushroom) データセットに関する混同マトリックス（NULL 値を許可する損失マトリックスを使用）

損失マトリックスはツリーのリーフ部分における予想確率に基づいているため、信頼度の高い予測を行うためには、次のような点に注意する必要があります。

1. 「決定木」および「選択式決定木」の「詳細設定」（Windows システム）（IRIX システムの場合は「詳細分析オプション (Further Inducer Options)」）で、「分割の下限値」オプションの値をデフォルトより大きい値（5 など）に設定してください。一般的に、訓練事例のサイズが大きくなるほど（ノイズが増えるほど）、このオプション値を大きく設定する必要があります。
2. 大きい訓練事例を使用してください。損失額の差異が上記の例ほど極端でない場合は、訓練事例が大きくないと、信頼度の高い予測結果が得られません。

3. 必要に応じて選択式決定木を使用してください。選択式決定木が常に有効とは限りませんが、通常は損失額を抑えるような良好な予測結果が得られます。たとえば、上記の例で $10,000 を $100 に変更し、NULL 値を禁止すると、決定木分析による予想損失額は $1,464 となり、選択式決定木による予想損失額は $662 となります。

すべての分析用の「詳細設定」（Windows システム）（IRIX システムの場合は「詳細分析オプション (Further Inducer Options)」）として、「損失マトリックスの使用 」オプションが用意されています。このオプションをオンに設定すると、「マトリックスの編集」ボタンを使用して損失マトリックスを定義できるようになります。未知の値（NULL 値）を予測対象から外す場合は、損失マトリクス最も高い値で未知の予測項目を埋めます。

エビデンス・ビジュアライザの「詳細設定」（Windows システム）（IRIX システムの場合は「詳細分析オプション (*Further Inducer Options*)」）で損失マトリックスが指定されている場合（デフォルト設定）は、「損失マトリックスの使用」というタイトルのボタンが円グラフの右下に表示されます。損失マトリックスは、円グラフで示される確率を調整するために使用されます。損失マトリックスを使用しない場合の確率を表示するには、「損失マトリックスの使用」ボタンの選択を解除します。損失マトリックスを編集するときに NULL を予測するための項目を指定した場合は、灰色のスライスが表示されることがあります。灰色のスライスが最大のスライスである場合は、NULL がクラシファイアによる予測結果です。

マップ・ビジュアライザ

マップ・ビジュアライザは、データを 3 次元のランドスケープとして表示するグラフィカル・インタフェースです。データは特定の場所に配置されるバーチャート（棒グラフ）の形で表示されます。マップ・ビジュアライザの目的は、地理的データの定量的な特性や関係を明らかにすることですが、空間的に関連性のある任意のデータのグラフ表示に利用することもできます。

個々のデータ項目は、ランドスケープ内のグラフィカルなバーチャート・オブジェクトと関連付けられます。これらのオブジェクトは地理的な形状と位置を表しています。ランドスケープはこれらの地理的なオブジェクトから構成され、各オブジェクトにはそれぞれ高さと色が割当てられます（図 1-26 を参照）。ランドスケープ内部を自由にナビゲートするほかに、ドリルダウン（地理的データを細分化する）またはドリルアップ（地理的データを集計処理する）を行ったり、アニメーションを使用して 1 つの次元または 2 つの次元に沿ってデータがどのように変化するかを確認したりすることもできます。

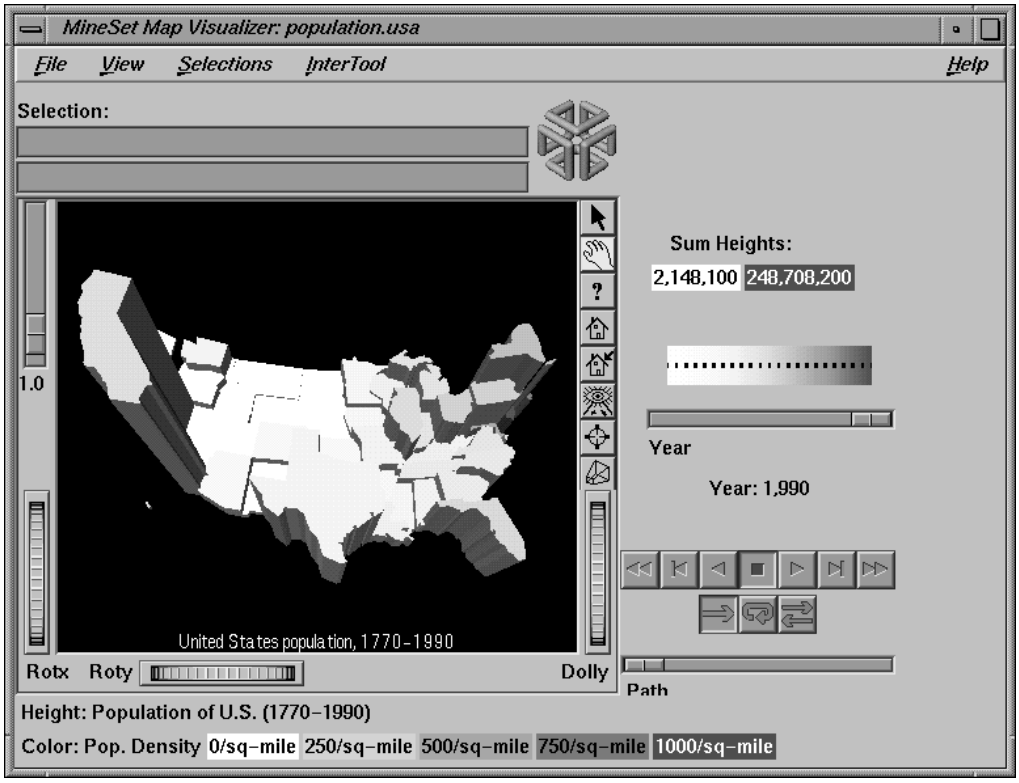

図 1-26 1990 年の米国の人口を表示したマップ・ビジュアライザの例

ランドスケープは単純な輪郭線として描かれる平面状の地理オブジェクトから構成される場合もあります。その場合、バーチャートの円柱は特定の場所に配置されます。

また、終端が特定のポイント位置にある直線から構成されるランドスケープもあります。その場合、個々の直線には独自の幅と色が割当てられます。任意の形状のオブジェクトや円柱は高さと色によって区別されますが、直線は幅と色によって区別されます。

マップ・ビジュアライザに必要なファイル

マップ・ビジュアライザを使用するには、データファイル、gfx ファイル、階層ファイル、設定ファイルが必要です。

- タブで区切られたフィールド行からなるデータファイル
  通常、データファイルは Tool Manager を使用して簡単に作成できますが、Tool Manager を使用しないで作成することもできます（ファイル・フォーマットの定義については、『*MineSet 3.0 Enterprise Edition Interface Guide*』の「Creating Data, Configuration Hierarchy, and GFX Files for the Map Visualizer」を参照してください）。

  データファイルを作成するには、データソース（Oracle、INFORMIX、Sybase などのデータベース）からデータを抽出し、マップ・ビジュアライザで使用できるようにフォーマットを変換します。データファイルの 拡張子はユーザが定義します（マップ・ビジュアライザ用のサンプルファイルの拡張子は *.data* となっています）。

- gfx ファイル
  gfx ファイルには、マップに表示する 1 次元、2 次元、3 次元のオブジェクトの形状と位置を記述します。

  gfx ファイルの拡張子は *.gfx* でなければなりません。MineSet には、郡 (country)、市外局番、郵便番号のレベルまで細分された米国の地図や、州 (province) のレベルまで細分されたカナダの地図など、各種の *.gfx* ファイルが用意されています。*.gfx* ファイルはユーザが手作業で作成することもできます（ファイル・フォーマットの定義については、『*MineSet 3.0 Enterprise Edition Interface Guide*』の「Creating Data, Configuration Hierarchy, and GFX Files for the Map Visualizer」を参照してください）。

- 階層ファイル
  階層ファイルには次の情報を記述します。

  – マップに表示する各種の地理的オブジェクトを表す項目名。

  – 地理的オブジェクトの位置と形状を記述した *.gfx* ファイルの名前。

  – 地理的オブジェクトの階層関係の記述（オプション）。この情報はドリルダウンとドリルアップに使用されます。

ドリルダウンとドリルアップは階層ファイルに基づいて行われます。すなわち、階層ファイルに記述された階層関係に基づき、特定のレベルのオブジェクトに関する情報を集計処理（または逆に情報を細分化）することによって、各オブジェクトを別のレベルで表示できるようにします。たとえば、複数の州からなる地域と各州の関係を示す階層ファイルを使用すると、人口などのデータ値を州レベルと地域レベルの両方で表示することができます。*gfx_files/usa.state.gfx* ファイルには、米国の 50 州の位置と形状が記述されています。階層ファイルである *gfx_files/usa.state.hierarchy* には、個々の州を複数の地域にグループ化し、それらの地域を東部と西部に分割し、東部と西部を米国に集計処理するという階層関係が記述されています。

詳細については、『*MineSet 3.0 Enterprise Edition Interface Guide*』の「Creating Data, Configuration Hierarchy, and GFX Files for the Map Visualizer」を参照してください。

- 入力データのフォーマットとその表示方法を定義した設定ファイル
  通常、設定ファイルは Tool Manager を使用して簡単に作成できますが、Tool Manager の代わりに任意のエディタ（メモ帳、jot、vi、Emacs など）を使用して作成することもできます（ファイル・フォーマットの定義については、『MineSet 3.0 Enterprise Edition Interface Guide』の「Creating Data, Configuration Hierarchy, and GFX Files for the Map Visualizer」を参照してください）。

  設定ファイルの拡張子は *.mapviz* にしなければなりません。これ以外の拡張子にすると、「ファイル」->「開く」オプションを選択したときに、ファイルがリストに表示されません。マップ・ビジュアライザを起動するとき、またはファイルを開くときは、データファイルではなく設定ファイルを指定してください。

マップ・ビジュアライザの起動

マップ・ビジュアライザを起動するには、次の 3 通りの方法があります。

- Tool Manager を使用して、マップ・ビジュアライザを設定して実行します。最初に『*MineSet 3.0 Enterprise Edition User' Guide for Windows*』を参照して、すべての MineSet ツールで共通に使用される Tool Manager の機能を確認してください。
- 使用する設定ファイルが分かっている場合は、その設定ファイルのアイコンをダブルクリックします。こうするとマップ・ビジュアライザが起動され、選択した設定ファイルが自動的に読み込まれます。この起動方法を利用できるのは、設定ファイルの拡張子が *.mapviz* である場合に限られます（Tool Manager を使用してマップ・ビジュアライザ用の設定ファイルを作成すると、ファイル拡張子が常に .mapviz になります）。
- IRIX シェルウィンドウのプロンプトに次のコマンドを入力して、マップ・ビジュアライザを起動します。

mapviz [*configFile*]

configFile は任意指定の引数であり、設定ファイルの名前を表します。コマンド行で設定ファイルを指定しなかった場合は、「ファイル」->「開く」オプションを選択してファイル名を指定する必要があります。

マップ・ビジュアライザ起動するときのオプション

IRIX システム上でビジュアライザの起動時に **-quiet** オプションを指定すると、進行状況を表すダイアログが表示されなくなります。このオプションを常に有効にするには、各ユーザのホーム・ディレクトリ内の *.Xdefaults* ファイルに次の行を追加します。

```
*minesetQuiet:TRUE
```

警告メッセージを表示する warnexecute ステートメントを設定することもできます。詳細については、「警告オプション」（219 ページ）を参照してください。Windows システム上では、「ファイル」メニューから「設定」を選択して、これらのオプションを設定することができます。

Tool Manager によるマップ・ビジュアライザの設定

ここでは、Tool Manager を使用してマップ・ビジュアライザ (Map Visualizer) を設定する方法について説明します。Tool Manager を使用すればマップ・ビジュアライザ用の設定ファイルを簡単に作成できますが、エディタを使用して手作業で設定ファイルを作成することもできます。詳細については、『*MineSet 3.0 Enterprise Edition Interface Guide*』の「Creating Data, Configuration Hierarchy, and GFX Files for the Map Visualizer」を参照してください。

.gfx ファイルと .hierarchy ファイルの作成

マップ・ビジュアライザを使用するには、地理的オブジェクトを定義する次の 2 種類のファイルを作成する必要があります。

- 1 つまたは複数のマップ (*.gfx*) ファイル。マップ・ファイルでは地理的オブジェクトの形状を定義します。
- 1 つの階層 (*.hierarchy*) ファイル。階層ファイルでは複数のマップ (*.gfx*) ファイル間の相互関係を定義します。

これらのファイルは Tool Manager では作成されません。したがって、MineSet パッケージに添付されているサンプルファイルを使用するか、またはユーザが手作業で作成しなければなりません。Windows システム上では、MineSet をインストールしたフォルダの配下の *\mapviz.gfx_files* サブフォルダに、これらのサンプルファイルがあります。IRIX システム上では、*/usr/lib/MineSet/mapviz/gfx_files* ディレクトリに、これらのサンプルファイルがあります。マップ・ファイルと階層ファイルを手作業で作成する場合は、『*MineSet 3.0 Enterprise Edition Interface Guide*』の「Creating Data, Configuration Hierarchy, and GFX Files for the Map Visualizer」を参照してください。

MineSet パッケージには次の *.gfx* ファイルと *.hierarchy* ファイルが添付されています。

- 米国の各州
- 米国の各州によって網羅される地域
- 米国の 5 桁の各郵便番号によって網羅される地域
- 米国の各市外局番によって網羅される地域
- カナダの各州と準州
- メキシコの各州
- オーストラリアの各州と準州
- 西欧と中欧の各国
- フランスとオランダの各地方

マップ・ビジュアライザでは次の項目を持つデータファイルが必要です。

- 地理的オブジェクトを表す 1 つの項目（たとえば、州の項目）。この項目の各行は個々の地理的オブジェクトを表します（州の項目の場合は、各行が 1 つの州を表します）。
- 各地理的バーの高さや色に（算術式を使用して）割当てられた数値を持つ少なくとも 1 つの項目。この項目はスケーラ、1D 配列、または 2D 配列となります。項目が配列の場合は、スライダを使用して特定のデータポイントを選択し、高さと色に割当てる必要があります。

高さと色の両方を 1D 配列または 2D 配列に割当てる場合、配列のインデックスは同じ値でなければなりません（『*MineSet 3.0 Enterprise Edition Interface Guide*』の「Creating Data, Configuration Hierarchy, and GFX Files for the Map Visualizer」を参照）。

スライダとアニメーションの作成

「アニメーション」（10 ページ）を参照してください。

マップ・ビジュアライザのオプション

「ツールオプション」ボタンをクリックすると、新しいダイアログ・ボックスが表示され、マップ・ビジュアライザのオプションのデフォルト値を変更することができます。

ここでは、マップ・ビジュアライザの「ツールオプション」ダイアログ・ボックスにあるボタンとフィールドについて説明します。

「地形」

「地形」セクションの「要素ファイル」フィールドでは、マップ・ビジュアライザのメインウィンドウ内で地理的な要素のオブジェクトを表示するために使用する階層 (*.hierarchy*) ファイルを指定します。

「アウトラインファイル」フィールドでは、平面として描画される輪郭オブジェクトを指定します。3D 要素のオブジェクトはこの平面上に配置されます。

「ファイル検索」ボタンをクリックすると、使用する *.hierarchy* ファイルを検索することができます。

「要素ファイル」と「アウトラインファイル」はオプションのフィールドです。「要素ファイル」が指定されていない場合は、任意のサイズの単純な矩形からなる地理的要素のオブジェクトが生成されて、グラフ内に配置されます。

「高さ」

「高さ」セクションでは、高さの「スケール」の初期値（デフォルト値は 1.0）を設定し、マップ・ビジュアライザのウィンドウの下方に高さの説明を表示するかどうかを指定します。

「色」

「色」オプションを使用するには、事前に「データの可視化 / マイニング」パネルの「* バーの色」に項目を割当てておく必要があります。色の選択と変更の詳細については、「「色の選択」（53 ページ）を参照してください。

「色のリスト」オプションでは、色のリストのラベルの横にあるプラス (+) ボタンを使用して色のリストを指定します。色のリストを指定すると、カラーエディタが起動され、リストに追加する色を指定できるようになります。

「マッピング」オプションでは、表示される色の変化が「連続」的であるか、「離散」的であるかを指定します。「連続」を選択した場合、「色のリスト」フィールドと「マッピング」フィールドで指定した値に従って色が徐々に変化します。

ポップアップ・ボタンの右横にあるフィールドには、色を割当てる特定の値を入力します。このフィールドの値の数は、「色のリスト」フィールドで指定した色の数と同じでなければなりません。

このフィールドには表示に必要な任意の数の色を入力することができます。「* バーの色」に割当てた項目の値の数が、指定した色の数を上回る場合は、マップ・ビジュアライザによって適切な数の色が実行時にランダムに選択されます。色の選択の詳細については、「色の選択」（53 ページ）を参照してください。

「説明オン」オプションでは、色の説明の表示 / 非表示を切替えます。

「正規化オン」オプションでは、色を割当てた項目の最小値と最大値の間で、色を自動的にスケーリングするかどうかを指定します（これを色の正規化と呼びます）。自動的にスケーリングしない場合は、手作業でしきい値を設定する必要があります。「正規化オン」をオンに設定した場合は、しきい値が 0 から 100 の範囲（色を割当てた項目の最小値を最大値で割ったパーセント値）で設定されます。

スライダ

「マップ・ビジュアライザ、スキャタ・ビジュアライザ、スプラット・ビジュアライザのスライダの作成」（168 ページ）を参照してください。

「メッセージ」フィールド

「メッセージ」フィールドには、要素を選択したときに表示されるメッセージを入力します。このフィールドに入力できるメッセージのフォーマットの一覧と説明については、『*MineSet 3.0 Enterprise Edition Interface Guide*』の「Creating Data, Configuration Hierarchy, and GFX Files for the Map Visualizer」の「message ステートメント」を参照してください。

マップ・ビジュアライザ (Map Visualizer) のメインウィンドウの下方に表示されるタイトル文字列を指定します。この文字列は二重引用符で囲まなければなりません。

「実行」フィールド

「実行」フィールドには、要素をダブルクリックしたときに実行されるコマンドを入力します。入力のフォーマットは message ステートメントと似ています。このフィールドにコマンドを入力しなかった場合、要素をダブルクリックしても何も起こりません。

詳細については、『*MineSet 3.0 Enterprise Edition Interface Guide*』の「Creating Data, Configuration Hierarchy, and GFX Files for the Map Visualizer」の「Excute ステートメント」を参照してください。

ツールオプションのリセット

「ツールオプション」ダイアログ・ボックスで設定したオプションをすべてデフォルト値に戻すには、「リセット」ボタンをクリックします。

ツールオプションの保存

「ツールオプション」ダイアログ・ボックスで設定したオプションを確定して保存する場合は、「了解」ボタンをクリックします。「了解」ボタンをクリックすると、Tool Manager のメインウィンドウに戻ります。

マップ・ビジュアライザの設定情報が保存されるファイル

Tool Manager では、マップ・ビジュアライザに関する設定情報が複数のファイルに保存されます。これらのファイルには同じ接頭辞 <prefix> が付きます。

- <*prefix*>.*mapviz.data* にはデータが格納されます。
- <*prefix*>.*mapviz.schema* にはデータファイルが記述されます。
- <*prefix*>.*mapviz* にはマップ・ビジュアライザが必要とする情報が格納されます。
- <*prefix*>.*mineset* には MineSet の他のファイルを作成するのに必要なすべての情報が格納されます。

接頭辞 (prefix) を指定するには、Tool Manager のメインウィンドウで「ファイル」-> 「保存 ...」オプションを選択します。接頭辞を指定しなかった場合は、データソースを表す接頭辞が付きます。

「ツールの起動」ボタンをクリックすると、*.data*、*.schema*、*.mapviz* の各ファイルが必要に応じて更新されます。

「マイニングツール」タブ

Tool Manager の「データの可視化 / マイニング」パネルにある「マイニングツール」タブをクリックすると、次のタブが表示されます。

- 相関 ― 相関規則分析
- クラスター― クラスタリング
- クラス判別― このタブから実行できる分析は、決定木、選択式決定木、エビデンス、デシジョン・テーブル。分析の実行モードとして、クラシファイアとエラー、クラシファイアのみ、誤差推定、学習曲線を選択可能。
- 回帰ツリー ― 回帰ツールの実行モードとして、回帰ツリーと誤差、回帰ツリーのみ、誤差推定、学習曲線を選択可能。
- 重要項目 ― 重要項目ツール
- 任意のプラグイン（ACpro など）

複数のオブジェクトの選択

ほとんどのツールでは、キーボードの <Shift > キーとマウスの左ボタンを使用してオブジェクトを選択します。<Shift> キーを押さずにオブジェクト上でマウスの左ボタンをクリックすると、カーソルの下にあるオブジェクトが選択され、それ以前の選択はすべて解除されます。<Shift> キーを押しながらマウスの左ボタンをクリックすると、それ以前のオブジェクトの選択状態は変わらず、カーソルの下のオブジェクトが新たに選択されます。(「スプラット・ビジュアライザ」で説明するように、スプラット・ビジュアライザでは別のインタフェースが採用されています。)

ビジュアライザ内でオブジェクトを選択すると、そのオブジェクトに関する情報がメインウィンドウに表示されます。ビジュアライザのデフォルト設定では、最後に選択されたオブジェクトに関する情報だけが表示されます。ただし、レコードビューワの独立したウィンドウには、選択されている全オブジェクトに関する情報を示すテーブルが表示されます。ツリービジュアライザとマップ・ビジュアライザでは、「選択」->「値の表示 」メニュー・オプションを使用して、同様のテーブルを表示することができます。

特定のツールについてメッセージが設定されている場合は、そのメッセージもテーブルに表示されます。このテーブルでは、項目間のセパレータをドラッグして項目の幅を調整することができます。また、値をクリックすると、テーブルの一番上にその値の完全なテキストが表示されます。

相互情報量

相互情報量 は子ノードの加重平均純度と親ノード間の純度の変化（エントロピー）であり、決定木を分岐するときの条件として使用されます。加重平均純度は、個々の子ノードに存在するレコードの数に基づいて計算されます。決定木の分岐条件については、「決定木」（77 ページ）を参照してください。

Naive-Bayes

エビデンス分析は、Naive-Bayes または Simple Bayes とも呼ばれます。エビデンス分析によって分析モデルが構築されるときは、各属性値の確率が特定のクラスに依存しない（各クラスの各属性が互いに独立である）ことが前提となっています。たとえば、*iris* データセットでは、アヤメの 4 つの属性（萼片の縦、萼片の横、花びらの縦、花びらの横 (sepal length, sepal width, petal length, petal width)）がアヤメの各クラス（アイリスセトサ、アイリスバージカラー、アイリスバージニカ (iris-setosa,

iris-versicolor,iris-virginica)）に依存しないとみなされます。この単純な前提はあまり現実的ではありませんが、エビデンス・モデルはデータの初期分析には適しており、そのクラス判別結果は実用的な業務にも適用することができます。

ツリー・ビジュアライザ以外のビジュアライザで使用できるナビゲーション・コントロール

ここでは、ツリー・ビジュアライザ以外のビジュアライザ（エビデンス、デシジョン・テーブル、マップ、スキャタ、スプラット）で使用できるナビゲーション・コントロールのクイック・リファレンスとして役立つ 3 つの一覧表を示します。まず、表 1-12 にナビゲーション・ボタンの一覧を示します。

表 1-12 ツリー・ビジュアライザ以外のビジュアライザのナビゲーション・ボタン

ボタン	名前	機能
	ピック	ピック (Pick) モード（矢印カーソル）に切替える。ピックモードでは、グラフの要素を強調表示（ブラシ）または選択（クリック）できる。
	ハンド	ハンドモード（手のひらカーソル）に切替える。ハンドモードでは、ウィンドウ内でグラフを回転・移動できる。 －グラフを回転するには、マウスの左ボタンを押したままマウスを移動する。 －ウィンドウ内でグラフを移動するには、マウスの左ボタンと右ボタン（3 ボタンのマウスを使用できるようにシステムが設定されている場合は、マウスの中ボタン）を押したままマウスを移動する。
	ホーム	ホーム表示として設定されたサイズと位置にグラフを戻す。 デフォルトのホーム表示は、ビジュアライザが最初に起動されたときのグラフのサイズと位置である。ホーム表示を変更するには、「ホームの設定 (Set Home)」ボタンを使用する。
	ホームの設定	グラフのホーム表示（サイズと位置）を変更する。グラフのデフォルトのサイズと位置を変更する場合は、このボタンを使用する。
	全体の概観	グラフを中央に移動し、グラフ全体がウィンドウ内に収まるようにする。

表 1-12（続き）ツリー・ビジュアライザ以外のビジュアライザのナビゲーション・ボタン

ボタン	名前	機能
	ズーム	選択したポイントをウィンドウの中央に移動し、そのポイント付近の領域をズームイン（拡大表示）する。カーソルの形状が照準カーソルに変わったら、ズームインの対象領域にカーソルを移動し、マウスの左ボタンと右ボタン（3 ボタンのマウスを使用できるようにシステムが設定されている場合は、マウスの中ボタン）を押さえる。
	3D	3D 遠近法に切替える。
	上面表示	グラフを上から見下ろす（スキャタ・ビジュアライザとスプラット・ビジュアライザのみ）。
	前面表示	グラフを正面から見る（スキャタ・ビジュアライザとスプラット・ビジュアライザのみ）。
	側面表示	グラフを側面から見る（スキャタ・ビジュアライザとスプラット・ビジュアライザのみ）。

表 1-13 に、ツリー・ビジュアライザ以外のビジュアライザで使用できる調整スライダとダイヤルの一覧を示します。

表 1-13 ツリー・ビジュアライザ以外のビジュアライザの調整スライダとダイヤル

スライダまたはダイヤル	機能
「高さ」スライダ（ウィンドウの左上）	グラフのケーキ、円、バーの高さを調整して差異を強調する。
「詳細」スライダ（エビデンス・ビジュアライザとデシジョン・テーブル・ビジュアライザのみ）	ラベルの分析モデルにあまり効果がない属性を除外する。
「有意水準 % の表示」スライダ（エビデンス・ビジュアライザとデシジョン・テーブル・ビジュアライザのみ）	データセット内にあるレコードの重み合計のうち、指定された割合（最大 2%）を下回る重みの属性値を除外する。
「Rotx」ダイヤル	グラフを X 軸の回りで（上下に）回転する。
「Roty」ダイヤル	グラフを Y 軸の回りで（左右に）回転する。
「Dolly」ダイヤル	グラフのズームインまたはズームアウトを行う。

表 1-14 に、ツリー・ビジュアライザ以外のビジュアライザ内のグラフに対して実行できる操作の一覧を示します。

表 1-14 ツリー・ビジュアライザ以外のビジュアライザ内のグラフに対する操作

操作	スライダまたはダイヤル	マウスまたはキーボード
ピックモードとハンドモードを切替える。	該当なし	< Esc > キーを押すか、ナビゲーション・ボタンをクリックする。
グラフを移動する。	該当なし	ハンドモードで、マウスの右ボタンをクリックして押さえたまま、グラフを移動する方向にマウスをドラッグする。
グラフのケーキ、円、バーの高さを調整して差異を強調する。	「Height」スライダ（ウィンドウの左上）	該当なし
グラフを X 軸の回りで（上下に）回転する。	「Rotx」ダイヤル	ハンドモードで、マウスの左ボタンをクリックして押したまま、グラフを回転する方向にマウスをドラッグする。
グラフを Y 軸の回りで（左右に）回転する。	「Roty」ダイヤル	ハンドモードで、マウスの左ボタンをクリックして押したまま、グラフを回転する方向にマウスをドラッグする。
グラフのズームインまたはズームアウトを行う。	「Dolly」ダイヤル	ハンドモードで、マウスの左ボタンと右ボタン（3 ボタンのマウスを使用できるようにシステムが設定されている場合は、マウスの中ボタン）をクリックして押したまま、マウスを下方向（ズームインの場合）または上方向（ズームアウトの場合）にドラッグする。

表 1-14（続き）ツリー・ビジュアライザ以外のビジュアライザ内のグラフに対する操作

操作	スライダまたはダイヤル	マウスまたはキーボード
細かいレベルにドリルダウンする（デシジョン・テーブル・ビジュアライザとマップ・ビジュアライザのみ）。	該当なし	特定のグラフ（または全グラフの背景）上にマウスの矢印カーソルを移動し、マウスの右ボタンをクリックする。
粗いレベルにドリルアップする（デシジョン・テーブル・ビジュアライザとマップ・ビジュアライザのみ）。	該当なし	特定のグラフ（または全グラフの背景）上にマウスの矢印カーソルを移動し、< Ctrl > キーを押しながらマウスの右ボタンをクリックする（3 ボタンのマウスを使用できるようにシステムが設定されている場合は、マウスの中ボタンをクリックする）。

ツリー・ビジュアライザで使用できるナビゲーション・コントロール

ツリー・ビジュアライザの表示内容は、カメラを通じて見るシーンに例えることができます。表示内容を変更するには、カメラ（視点）の位置を変更します。ここでは、ツリー・ビジュアライザ（ツリー、決定木、選択式決定木、回帰ツリー）で使用できるナビゲーション・コントロールのクイック・リファレンスとして役立つ 2 つの一覧表を示します。まず、表 1-15 にナビゲーション・ボタンの一覧を示します。

表 1-15 ツリー・ビジュアライザのナビゲーション・ボタン

ボタン	機能
	ホーム表示として設定されたサイズと位置にグラフを戻す。 デフォルトのホーム表示は、ビジュアライザが最初に起動されたときのグラフのサイズと位置である。ホーム表示を変更するには、次のボタンを使用する。
	グラフのホーム表示（サイズと位置）を変更する。グラフのデフォルトのサイズと位置を変更する場合は、このボタンを使用する。
	グラフを中央に移動し、グラフ全体がウィンドウ内に収まるようにする。
	直前の移動操作を元に戻す（Web ブラウザの「戻る」ボタンと同様）。
	元に戻した移動操作をやり直す（Web ブラウザの「進む」ボタンと同様）。

表 1-15（続き）ツリー・ビジュアライザのナビゲーション・ボタン

ボタン	機能
	ツリーのルートに向かって 1 ノード分だけ進む。
	1 ノード分または 1 バー分だけ左側に移動する。
	1 ノード分または 1 バー分だけ右側に移動する。
	ツリーの左側のパス上で 1 ノード分だけ下に移動する。
	ツリーの右側のパス上で 1 ノード分だけ下に移動する。
	現在のノードから移動できるパスを示すポップアップ・メニューを表示する。

表 1-16 に、ツリー・ビジュアライザで使用できる調整スライダとダイヤルの一覧を示します。

表 1-16 ツリー・ビジュアライザの調整スライダとダイヤル

スライダまたはダイヤル	機能
「Height」スライダ（ウィンドウの左上）	グラフのバーの高さを調整して差異を強調する。
「H」ダイヤル	カメラ（視点）を上下に移動する。
「Tilt」ダイヤル	カメラ（視点）を上下方向に傾ける。
「<-->」ダイヤル	カメラ（視点）を左右に移動する。
「(Dolly」ダイヤル	カメラ（視点）を前後に移動する。

属性値の順序付け」メニュー

エビデンス・ビジュアライザとデシジョン・テーブル・ビジュアライザの「属性値の順序付け」メニューでは、離散属性の値をどのようにソートするかを制御することができます。このメニューには、次のオプションがあります。

- 「アルファベット順」は、属性値を左から右に（または下から上に）アルファベット順でソートします。

- 「重み順」は、レコードの重みが大きい属性が左側に並ぶように、属性値を左から右にソートします。
- 「ラベル確率順」（デフォルト）は、クラスの 1 つを表すスライスのサイズに基づいて、属性値を左から右にソートします。ラベルが階級生成された属性である場合は、最大値の階級のクラスがソート基準となります。ラベルが離散的な属性である場合は、事前確率の円グラフでスライスが最大であるクラスがソート基準となります。特定のクラスを選択して、ラベル確率によるソートを要求すると、選択したクラスがソート基準となります。いずれの場合も、NULL 値は一番左側に最初の値として配置されます。

正規化相互情報量

決定木では、デフォルトの分岐条件として正規化された相互情報量が使用されます。正規化された相互情報量は、相互情報量を子ノードの数の対数（底は 2）で除算した値です。決定木の分岐条件については、「決定木」（77 ページ）を参照してください。

NULL 値（Nulls ）

スプラット・ビジュアライザやマップ・ビジュアライザなどの一部のビジュアライザでは、未知のデータ値（NULL 値）を含むフィールドをビジュアル属性（色など）に割当てたときに、特別な表現形式が使用されます（NULL 値の詳細については、『*MineSet 3.0 Enterprise Edition Interface Guide*』の「Nulls in MineSet 」を参照してください）。スプラットの色に割当てた項目について、階級生成 (bin) 内のすべてのレコードが NULL 値のときは、そのスプラットがグレーで表示されます。色に割当てた項目について、集計内の 1 つまたは複数のレコードが非 NULL 値である場合は、それらのレコードの値が色の算出に使用されます。（通常の）値と NULL 値の合計は NULL 値になりますが、値と NULL 値の平均はその値になります（すなわち、value + Null = Null になり、avg(val, Null) = val になります）。

「ピック」ウィンドウ、「選択」フィールド、「ポインタを通過」エリアなど、ビジュアライザのさまざまな表示ウィンドウ内では、NULL 値が疑問符 (?) で表されます。

軸にマッピングされた数値型項目に NULL 値が含まれている場合は、軸で定義される範囲の下に特別な NULL 位置が表示されます。そのため、NULL 値が他の値と非連続であることがすぐに分かります。数値型項目の軸の NULL 位置は、「表示」メニューの「NULL 位置の表示」オプションを使用して非表示にすることができます。軸にマッピングされた文字列型項目の場合、NULL 値（'?' で表示）は他の通常の値と同様に扱われます。

マップ・ビジュアライザでは、NULL 値が次のような場合に発生します。

- データベースまたはデータファイルに NULL 値が含まれている場合。
- Tool Manager を使用して階級生成 (bin) に基づく配列を作成したときに、特定の階級に当てはまるデータが存在しない場合。たとえば、30-40 歳の人口に該当するデータが存在しない場合、その階級（30-40 歳）は NULL になります。
- Tool Manager を使用して配列を作成するときに、null enum オプションを指定した場合。この場合は、NULL 値の階級をすべて集計処理した配列要素が作成されます。NULL 値の階級には疑問符 (?) が付きます。この階級の値を表示するには、スライダを一番左側に移動します。この NULL 値の階級に当てはまるデータが存在しない場合は、階級自体の値が NULL になり、その階級を表すグラフィック・オブジェクトは「NULL オブジェクト」として表示されます。
- 表現または集計処理に NULL 値が含まれている場合。（「MineSet の NULL 値」を参照）
- マップ・ビジュアライザでは、NULL 値をビジュアル属性に割当てると、特別な表現形式が使用されます。NULL 値を高さに割当てると、高さがゼロで色が濃いグレーのオブジェクトが作成されます。NULL 値を色に割当てると、（高さに割当てられた値に応じた）適切な高さで色が濃いグレーのオブジェクトが作成されます。

•

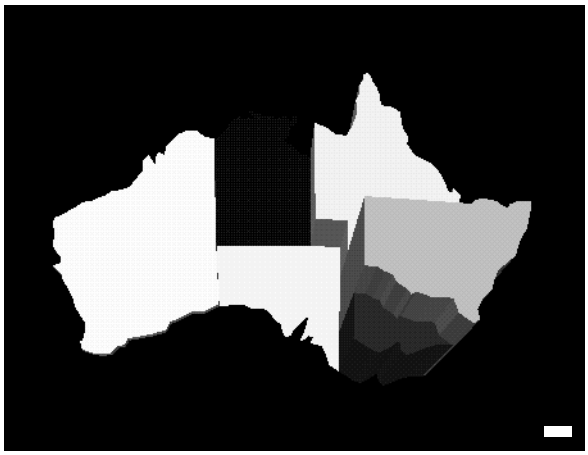

図 1-27 NULL 値を高さに割当てた場合（中央上のオブジェクト）と色に割当てた場合（右下のオブジェクト）

NULL 値を持つオブジェクトを選択した場合、選択フィールドには疑問符 (?) が表示されます。NULL 値を高さに割当てた場合と色に割当てた場合の外観を図 図 1-27 に示します。

選択式決定木

選択式決定木は予測モデルであり、従属属性（値が判明している属性）の値に基づいてラベル（未知の属性）の値が予測されます。選択式決定木によるクラス判別では、最初にデータセット内の個々のレコードが特定のクラスに割当てられます。このクラス判別で使用される基本的な構造は、「決定木」（77 ページ）で説明した決定木です。選択式決定木は、選択肢ノードを使用することによって通常の決定木を拡張したものです。選択肢ノードは、ツリー内の個々の決定木ノードで選択できる複数の選択肢（属性）を表します。選択式決定木は通常の決定木よりもかなり複雑で大きいモデルになる傾向があります。

選択式決定木の作成

選択式決定木モデルはデータに基づいて自動的に作成されます。複数のレコードと各レコードに対応するラベルから構成されるデータは、訓練事例と呼ばれます。

必要なファイル

選択式決定木分析 には訓練事例が必要です。訓練事例用のファイルを作成するには、様々なデータソース（MineSet の ASCII ファイルまたはバイナリファイル、Oracle、INFORMIX、Sybase などのデータベースのテーブルなど）からデータを抽出します。作成されたクラシファイアを適用するには、その分析モデルで使用される属性（ただし、ラベル属性は不要）を持つレコードからなるデータセットが必要です。

選択式決定木クラシファイアの作成

選択式決定木クラシファイアはデータに基づいて自動的に作成されます。最初にサーバにログインし、通常の手順に従ってデータソースを選択します。

Tool Manager の「クラス判別」タブで、「分析」ポップアップ・メニューから「選択式決定木」を選択してください。必要な場合を除き、特別なオプションを設定する必要はありません。単に「実行」ボタンをクリックすると作成されます。選択式決定木では、ツリー・ビジュアライザを使用して分析結果が表示されます。

IRIX システム上での並列化処理

マルチプロセッサ版の MineSet をインストールした場合は、決定木の枝に 1,000 個を越えるレコードが含まれているときに、ツリー・アルゴリズムを並列的に計算することができます（「IRIX システム上での並列化処理」（146 ページ）を参照）。スレッドの数を調整するには、Tool Manager の「設定 」パネルで「並列化処理」モードを変更します（「ファイル」メニュー」を参照）。この並列化オプションは、IRIX システム上だけで使用することができます。

選択式決定木のオプション

「詳細設定」（IRIX システム上では「クラシファイア詳細オプション (Further Classifier Options)」）を選択すると、「クラシファイアオプション」ダイアログ・ボックスが表示されます。このダイアログ・ボックスには次の 4 個のパネルがあります。

- 一番上のパネルには、Tool Manager の「データの可視化 / マイニング」パネルで設定した項目が表示されます。
- 上から 2 番目のパネルでは、損失マトリックスと重み属性を設定することができます。詳細については、「損失マトリックス」（121 ページ）と「レコードの重み付け」（151 ページ）を参照してください。
- 左下のパネルでは、詳細な分析オプションを設定することができます（下記参照）。
- 右下のパネルでは、誤差推定オプションを設定することができます。ただし、「データの可視化 / マイニング」パネルで「クラシファイアのみ」モードを選択した場合、このパネルには何も表示されません。このパネルに表示されるオプションは、選択した誤差推定法に応じて異なります（「誤差推定」（87 ページ）を参照）。

選択式決定木の生成アルゴリズムを微調整するために、決定木 (Decision Tree) の分析オプション（「決定木」（77 ページ）を参照）の他に、次に示す詳細な分析オプションが用意されています。

- 最大 # ルート選択肢数

  ルートノードに作成できる選択肢の数を制限する整数（デフォルト値は 5）です。分析 (Inducer) によって作成される選択肢の数は、各属性の優劣に基づいてこの値以下に制限されます。

- 減少数

  決定木ノードの各レベルで選択肢を減少させる数を定義する整数（デフォルト値は 2）です。たとえば、「最大 # ルート選択肢数」がデフォルト値の 5 である場合は、決定木ノードの 2 番目のレベルには最大で 3 つ (5-2=3)、3 番目のレベルには最大で 1 つ (3-2=1) の選択肢が存在することになります。決定木ノードの 4 番目以降のレベルには 1 つの選択肢だけが存在します。

- 最小の適合比率

  属性を選択肢から除外する条件を決める比率です。各属性の適合性が分析によって判断されるときは、最適な属性の他に選択肢として適した属性が選択されます。この最小の適合比率は、最適な属性を除く他の属性が選択肢となるために必要な適合性を定義します。最小の適合比率が f である場合、選択肢となる属性は少なく

とも (*1*-f)*b にランク付けされなければなりません（*b* は最適な属性の評点です）。最小の適合比率が 1 である場合は、すべての属性が選択肢となります（したがって、分割対象の属性が複数存在する場合は、上記の制限条件が早い時点で満たされます）。最小の適合比率が 0 である場合は、通常の決定木が構築されます（選択肢ノードは作成されません）。デフォルト値は 0.9 です。

選択式決定木の作成にかかる時間は、作成される選択肢ノードの数と密接な関係があります。通常、選択肢ノードはツリーの最上部付近（分かり易さや誤差低減のために最適な位置）に作成されるため、選択式決定木の作成にかかる時間の適正な概算値は、子ノードを持たない選択肢ノードの数に決定木の構築時間を掛け合わせた値になります。デフォルト設定では、ルートノードに作成できる選択肢の最大数が 5 つ、その下の子ノードに作成できる選択肢の最大数が 3 つであるため、選択肢ノードの最大数は 15(5*3) になります。最大 # ルート選択肢数 を 6 に増やすと、作成できる選択肢ノードの総数は 48(6*4*2) になります。同様に、最大 # ルート選択肢数を 7 に増やすと、作成できる選択肢ノードの総数は 105(7*5*3) になります。また、「最大 # ルート選択肢数 」を 5 に維持したまま、減少数 を 1 に変更すると、作成できる選択肢ノードの総数は 120(5*4*3*2) になります。最後の例では、選択式決定木の構築時間は、通常の決定木の構築時間と比較して階乗の次数で増大することになります。通常、「最小の適合比率」を低くすると、これらの制限値を調整する場合よりも、作成される選択肢の数が減少するため、ツリーの構築時間が短縮されます。

IRIX システム上での並列化処理

IRIX システム用のマルチプロセッサ版の MineSet には、64 ビット対応の並列処理に関するオプションが用意されています。この並列化オプションを使用すると、大量の計算処理を伴うタスクを並列的に実行することができます。並列処理を行う場合は、並列サーバをインストールし、Tool Manager の「設定 (Preferences)」メニューから「並列化 (parallel version)」を選択してください。

DataMover はサーバ上で動作するプロセスであり、データベース内やフラットファイル内に保存されたデータを利用できるようにする機能と、データマイニングや可視化に適した形式にデータを変換する機能があります。

IRIX 6.4 以降では、大容量メモリ（64 ビット）がサポートされています。IRIX 6.2 を使用している場合、32 ビットのデータマイニング・ユーティリティは使用できますが、64 ビット対応とスレッド (p-thread) を利用するためには、IRIX 6.5 にアップグレードする必要があります。64 ビット・アドレス指定を完全に利用するには、システム構成に応じて systune リソース・パラメータを変更する必要があります。

systune リソース・パラメータは、使用可能なシステムリソースのデフォルトの制限値を設定するものです。表 1-17 に、SGI 社が推奨する **systune** パラメータ値を示します（詳細については、systune (1M) のマニュアルページを参照してください）。

表 1-17　　systune パラメータ

パラメータ	定　義	推奨値
rlimit_pthread_cur	スレッド数の制限値	1024
rlimit_rss_cur	メモリ使用の制限値	マシンの物理メモリ量
rlimit_vmem_cur	仮想メモリ使用の制限値	マシンの論理スワップスペースまたは物理メモリの約 2 倍
rlimit_nofile_cur	開くファイルの数の制限値	1024 またはスレッド数の制限値

注記： systune パラメータの値を変更した後は、マシンをリブートする必要があります。

マルチプロセッサ版の MineSet をインストールした場合は、枝に 1,000 個を越えるレコードが含まれているときに、ツリー・アルゴリズムを並列的に計算することができます。ツリーの個々のレベルにおける最適な分岐処理がツリー上の各ノードによって分析され、それらの処理が並列的に実行されます。プログラムが生成 (spawn) できるスレッドの最大数は、デフォルト・パラメータによって自動的に決定されます。スレッドの数を調整するには、Tool Manager の「設定 (Preferences)」パネルで「並列化処理 (parallelization)」モードを変更します。並列処理を実行するとメモリのフラグメンテーション（断片化）が発生するため、マルチプロセッサ上で並列的に計算できるデータセットの最大サイズはシングル・プロセッサ上で計算できるデータセットよりも小さくなる可能性があります。

予測値

相関規則を分析するときに、右辺 (RHS) と左辺 (LHS) を 2 つの軸としてグリッドを可視化すると、相関規則の予測値は、LHS の全事象のうち、RHS の対応する事象が発生する割合として表されます。すなわち、予測値は、普及率（Prevalence：RHS の事象と LHS の事象が共に発生する度数）を LHS の事象が発生する度数で割った値です（RHS と LHS の説明については、「相関規則の可視化」（31 ページ）を参照）。たとえば、事象 X と事象 Y の相関規則の予測値が 50% である場合は、事象 X が発生する全レコードの 50% において、事象 Y が同時に発生することが期待されます。したがって、特定のレコードで事象 X が発生することが分かれば、そのレコードで事象 Y が同時に発生する確率は 50% ということになります。

普及率

相関規則を分析するときに、右辺 (RHS) と左辺 (LHS) を 2 つの軸としてグリッドをグラフ化すると、相関規則の普及率 は、LHS の事象と RHS の事象が共に発生する割合として表されます（RHS と LHS の説明については、「相関規則の可視化」（31 ページ）を参照）。すなわち、普及率は事象 X と事象 Y が共に発生する度数をレコードの総数で割った値です。たとえば、普及率が 1% である場合は、レコード総数の 1% において事象 X と事象 Y が同時に発生します。

枝切り

枝切り は、決定木用の分析オプションの 1 つです。枝切り係数の値を調整すると、ツリーの高さと分割の下限値が変更されます。枝切り係数のデフォルト値は 0.7 ですが、この値を大きくすると、より多くのサブツリーが除去されます。詳細については、「決定木」（77 ページ）を参照してください。

ランダムシード

MineSet のいくつかのダイアログ・ボックスには、ランダム・シードを指定するためのオプションがあります。ランダム・シードは乱数生成アルゴリズムの開始点を設定する値であり、生成された乱数はデータセットの標本抽出に使用されます。たとえば、標本抽出を行うたびに同じランダム・シードを使用すると、分析モデルの調整時に常に同じデータ集合を選択することができます。ランダム・シードを変更すると、同じデータセットから別のデータ集合が選択されます。

レコードビューワ

レコードビューワを使用すると、データを直接表示することができます。このツールは、項目と項目内のデータ値を確認する上で便利です。レコードビューワには次のような機能があります。

- 項目の各種操作（項目のサイズ変更、配置変更、表示 / 非表示の切替えなど）
- 特定の項目の値に基づくデータのソートまたはフィルタリング
- 複数の項目の値に基づくデータのソート（Windows システムのみ）
- ソートまたはフィルタリングを実行した後で行番号を変更する機能
- 特定の値の検索
- 操作したファイルをさまざまなフォーマットで保存する機能

レコードビューワの起動

レコードビューワを起動するには複数の方法があります。

- Tool Manager を使用する方法：
  - 「データの可視化 / マイニング」パネルの「可視化ツール」タブをクリックします。
  - Windows システム上では、「レコード」タブをクリックします。IRIX システム上では、「ツール」メニューから「レコードビューワ」を選択します。
- Tool Manager の「可視化ツール」メニューから「レコードビューワ」を選択した後、レコードビューワの「ファイル」メニューを使用してファイルを開きます。
- .schema ファイルのアイコンをダブルクリックします（Windows システムのみ）。

- IRIX システム上では、UNIX シェルのコマンド行プロンプトに対して次のコマンドを入力します。

  recordview [*filename*]

行番号の振り直し

レコードビューワでは、行番号をいつでも変更することができます。ソートまたはフィルタリングの後に行番号を変更した場合、その操作を取消す (Undo) ことはできません。元のデータに戻すには、ファイルを再度開く必要があります。

行番号を変更するには、「表示」メニューから「行番号の振り直し」を選択します。

レコードビューワでの検索

レコードビューワでは、データ内の特定の値を検索することができます。「検索」パネルを開くには、「表示」メニューから「検索パネル」を選択します。特定の値を検索するには、その値を入力し、検索対象の項目を強調表示して、「次を検索」または「前へ検索」をクリックします。

データの保存

レコードビューワでは、データに加えた変更も含めて、すべてのデータをファイルに保存することができます。データを保存するには、「ファイル」メニューから「保存 」または「名前を付けて保存」を選択します。

「保存」を選択すると、データは元のファイル名とフォーマットで保存されます。データを初めてファイルに保存する場合、ファイルは MineSet のバイナリ・フォーマットで保存されます。「名前を付けて保存」を選択すると、「データを保存」画面が表示されます。この画面では、任意のファイル名とフォーマットを指定することができます。

「名前を付けて保存」を使用する場合は、バイナリ、ASCII、HTML、テキストの 4 種類のフォーマットでデータを保存することができます。バイナリまたは ASCII フォーマットを選択すると、データファイルとスキーマファイルの両方が保存されます。HTML フォーマットを選択すると、ファイルが HTML 形式のテーブルとして保存されます。テキスト・フォーマットを選択すると、項目のタイトルが最初の行になり、2 行目以降はタブで区切られた形式でデータが保存されます。

レコードの重み付け

分析モデル作成の試行段階では、母集団の特定部分が他の部分よりも頻繁に標本抽出される場合があります。たとえば、母集団の 1% のレコードを標本抽出するときに、元の母集団の 0.1% しか占めない部分から 1% のレコードを抽出すると、0.1 * 0.01 = 0.001% の標本しか得られないことになります（適正に処理するには標本数が少なすぎます）。レコードの重み付けを行うと、個々のレコードに重みを割当てることができます。たとえば、母集団のうち、2 倍の頻度で標本抽出される部分集合に 0.5 という重みを割当て、残りの部分に 1.0 という重みを割当てると、母集団全体が適切に調整されます。

別の例として、電気通信事業者が不正な通話のすべてをデータベースに記録する一方で、正常な通話は一部分しか記録していない場合は、各レコードの重み付けを行って、母集団における実際の占有率を各レコードに割当てることができます。

また、一部のデータセットが既に集計処理されていたり、レコードに潜在的なカウント属性が含まれている場合があります（たとえば、米国の都市に関する統計値には人口が加味されています）。カウント属性を重みに変換すると、個々のレコードを複製するのと同じ効果が得られます。

レコードの重み付けの概念では、重みが 2 である 1 つのレコードは、重みが 1 である 2 つのレコードと同じです。重みの値として浮動小数点数を指定することもできます。「重み付け」（220 ページ）も参照してください。

「回帰」タブ

「回帰」タブでは、MineSet の回帰ツリーを利用することができます。「回帰」タブを表示するには、Tool Manager の「データの可視化 / マイニング」パネルにある「マイニングツール」タブをクリックします。

回帰ツリー

回帰ツリーは回帰を実行する予測モデルです。回帰とは、1 つまたは複数の説明属性に基づいて連続的なラベル値を予測する処理です。回帰とクラス判別は似ていますが、両者の違いは、クラス判別によって予測されるラベルが少数の離散的な値だけを取るのに対して、回帰モデルによって予測されるラベルは連続的な範囲内で任意の値を取ることです。

MineSet では回帰モデルが構築されると、その回帰モデルの構造を視覚的に分かりやすく表現するグラフが同時に作成されます。このグラフを見ると、回帰モデルの構造や推論の仕組みをよく理解できるとともに、データ自身の特性を細かく解析することもできます。作成された回帰モデルをデータセットに適用すると、ラベルのないレコードのラベル値を予測することができます。

回帰ツリーの作成

MineSet では、訓練事例 に基づいて回帰モデルが自動的に作成されます。「回帰ツリーのみ」以外のモードを選択すると、データが分割されて訓練事例が作成されます。訓練事例はデータセット内のレコードのうち、連続的なラベルの値が判明しているレコードから構成されます。たとえば、個々の説明属性（年齢、最終学歴、業種、週の労働時間など）ごとに 1 つの項目と、ラベル（総収入）を表す 1 つの項目を持つデータベーステーブルを訓練事例として使用することができます。最初にサーバにログインし、通常の手順に従ってデータソースを選択してください。

Windows **システム** :

Tool Manager の「マイニングツール」タブをクリックし、表示される「回帰」タブをクリックします。

IRIX **システム** :

Tool Manager の「回帰」タブをクリックすると、「分析」メニューに「回帰ツリー」オプションが自動的に表示されます。

必要な場合を除き、特別なオプションを設定する必要はありません。各オプションは決定木とまったく同じですが、ラベルは離散値ではなく連続値になります。「ツールの起動」ボタンまたは「実行」ボタンをクリックすると、回帰ツリーが作成されます。回帰ツリーでは、ツリー・ビジュアライザを使用して分析結果が表示されます。

「連続型ラベル」メニュー

「連続型ラベル」メニューには、使用可能な連続型のラベルのリストが表示されます。このリストには、値が数値（連続）であるすべての属性が表示されます。回帰モデルの予測対象とするラベル属性をこのリストから選択してください。たとえば、総収入を予測する回帰モデルを構築する場合は、ラベル属性として "gross income" を選択します。回帰モデルは連続型のラベルについてのみ構築できます。データセット内に連続的な属性（ラベル）が存在しない場合は、このメニューに「連続型のラベルがありません」と表示され、「実行」ボタンは使用できない状態になります。その場合は、Tool Manager の「データ変換」パネルを使用して連続的な属性を追加する必要があります。

回帰ツリーのオプション

「詳細設定」（IRIX システム上では「詳細分析オプション (*Further Classifier Options*)」を選択すると、「回帰ツリーオプション」ダイアログ・ボックスが表示されます。このダイアログ・ボックスには次の 4 個のパネルがあります。

- 一番上のパネルには、Tool Manager の「データの可視化 / マイニング」パネルで設定した項目が表示されます。
- 上から 2 番目のパネルでは、損失マトリックスと重み属性を設定することができます。詳細については、「損失マトリックス」（121 ページ）と「重み付け」（220 ページ）を参照してください。
- 左下のパネルでは、詳細な分析オプションを設定することができます（下記参照）。
- 右下のパネルでは、誤差推定オプションを設定することができます。ただし、「データの可視化 / マイニング」パネルで「クラシファイアのみ」モードを選択した場合、このパネルには何も表示されません。このパネルに表示されるオプションは、選択した誤差推定法に応じて異なります（「誤差推定」（87 ページ）を参照）。

回帰ツリーの分析アルゴリズムを微調整するために、次に示す回帰ツリー分析オプションが用意されています。

- ツリーの高さ制限

  デフォルトでは、回帰ツリーのレベル数（高さ）に制限はありません。レベル数を制限するときは、このチェックボックスをクリックして最大レベル数を入力します。回帰ツリーのレベル数を制限すると分析の実行が高速になり、数多くのノードに気を取られないで回帰ツリーを解析することができます。レベル数を制限すると実行時間は短縮されますが、誤差率が増加する可能性があります。このオプションを設定しても、最大レベル数よりも前のレベルで選択された属性は影響を受けません。

- 分割の基準

  回帰ツリーの構築中に分割対象となる属性が複数存在する場合に、どの属性を選択して分割を行うかを決める基準を指定します。MineSet の回帰ツリーでは、次の 4 種類の分割基準が用意されています。

  – 分散

    「分散」を指定すると、ツリーの各レベルにおけるノード内部の分散が最小になるような分割が行われます。枝を生成するときは、その枝に到達するレコードのラベル値の平均が枝の予測値とされます。分散は各ラベル値と平均値の差の二乗和をラベル値の総数で割った値です。これは統計的に最も頻繁に使用されるモードです。

  – 絶対偏差

    「絶対偏差」を指定すると、ツリーの各レベルにおけるノード内部の絶対偏差が最小になるような分割が行われます。枝を生成するときは、その枝に到達するレコードのラベル値のメジアン（中央値）が枝の予測値とされます。

  – 正規化された分散

    「正規化された分散」を指定すると、分散の代わりに正規化された分散が使用されます（複数方向の分割を防ぐ効果があります）。正規化された分散は、分散を子ノードの数の対数（底は 2）で割った値です。

  – 正規化された絶対偏差

    「正規化された絶対偏差」を指定すると、絶対偏差の代わりに正規化された絶対偏差が使用されます（複数方向の分割を防ぐ効果があります）。正規化された絶対偏差は、絶対偏差を子ノードの数の対数（底は 2 ）で割った値です。

個別の問題に最も適した分割基準を判断するのは困難です。すべての分割基準を試して、誤差推定が最も小さくなる基準を選択するか、最も分かりやすい回帰ツリーが生成される基準を選択してください。

- 分割の下限値

  このオプションの値は、ノードの子ノードのうち、少なくとも 2 つのノードで設定する必要のある重み（重みが設定されていない場合はレコード数）の下限です。このオプションのデフォルト値は 5 です。たとえば、ノードを 3 方向に分割する場合は、3 つのうち少なくとも 2 つの子ノードに 5 以上の重み（重みが設定されていない場合は 5 つ以上のレコード）を割当てる必要があります。これは回帰ツリーのサイズを制限する代替手段になります。

  分割の下限値を増加させると、各枝上のレコード数（重み）が増えるため、予想確率の精度が改善される傾向があります。また、小さいツリーが構築されて、分析の実行時間が短縮されます。データにノイズ（誤差または異常値）が含まれていると思われる場合は、分割の下限値を増加させてください。データセットが非常に小さい（レコード数が 100 未満）場合は、この下限値を減少させてもかまいません。

- 生成コストを考慮した枝刈り

  このオプションでは、ツリーの誤差率（コスト）とツリー内の枝（Leaf）の数（複雑性）のトレードオフ関係を調整することによって、最適なサイズのツリーを生成します。コストと複雑性のバランスを調整するために、訓練事例は学習事例と枝切りセットに分割されます。最初に、学習事例に基づいてツリー全体が構築された後、そのツリーが枝切りされて、それほど複雑でない複数のツリーが作成されます。次に、枝切りセットに基づいて、複数のツリーの中から最もコストの低いツリーが選ばれ、そのツリーのサイズが記録されます。最後に、学習事例と枝切りセットを再結合した訓練事例に基づいてツリーが構築され、そのツリーが最低コストのツリーのサイズに枝切りされます。

  このオプションのパラメータ（枝切り係数）を指定すると、最低コストの（誤差率が最も低い）ツリーよりもサイズの小さいツリーを選択することができます。この枝切り係数は、一定範囲の標準誤差（コスト = 誤差率の増加）を許容して、ツリーの複雑性を抑えるものです。枝切り係数をゼロに設定すると、最低コストのツリーがそのまま選択されます。枝切り係数を 0.5 に設定すると、誤差率が（最低コストのツリーの誤差率 + 標準誤差 * 0.5）未満であるツリーのうち、最もサイズの小さいツリーが選択されます。枝切り係数を 1.0（デフォルト値）に設定すると、誤差率が（最低コストのツリーの誤差率 + 標準誤差 * 1.0）未満であるツリーのうち、最も小さいサイズのツリーが選択されます。枝切り係数を大きくすると、除去されるサブツリーの数が増えツリーのサイズは小さくなります。データにノ

イズ（誤差または異常値）が含まれている場合は、枝切り係数を増加させて小さいツリーを作成してください。ツリーが枝切りされて 1 つのノードになってしまう場合は、枝切り係数を減少させて大きいツリーを作成してください。

枝切りによるツリーの簡素化では、ツリー全体が構築されてからサブツリーが除去されるため、ツリーの高さ制限 または分割の下限値による簡素化よりも実行時間は長くなります。ただし、サブツリーが選択的に除去されるため、より正確な回帰モデルが生成されます。

回帰モデルにおける誤差推定

クラシファイアの一般的な評価基準としては誤差率（誤ったラベルがクラシファイアによって予測された標本の数）が使用されます。損失マトリックスを作成すると、様々なタイプの誤りに対して異なるペナルティー（損失額）を割当てることができます。その場合は、損失額が適切な評価基準となります。

連続的な実数値を予測する回帰モデルについては、普遍的な評価基準は存在しません。よく使用される一般的な評価基準として、平均二乗誤差と平均絶対誤差があります。平均二乗誤差は、予測されたラベル値から実際のラベル値を引いた差の二乗和をラベル値の個数で割った値（誤差の二乗和の平均値）です。平均絶対誤差は、予測されたラベル値から実際のラベル値を引いた差の絶対値の和をラベル値の個数で割った値（誤差の絶対値の平均値）です。

回帰モデル名

生成される回帰モデル名の接頭語は（Tool Manager で指定される）セッション・ファイル名になり、接尾語は *-rt.regress* になります。デフォルトでは、すべての回帰モデルがサーバ上の *file_cache* ディレクトリ（Windows システム上のデフォルトは *MineSet Files*、IRIX システム上のデフォルトは *mineset_files*）に保存されます。Windows システム上では、回帰可視化ファイル（接尾語は *-rt.treeviz*）が現在の作業ディレクトリに保存されます。

回帰モデルを使用すると、ラベルのないデータセットのラベルを予測することができます。保存された回帰モデルを適用するには、その回帰モデルで使用される属性（ただしラベル属性は不要）を持つレコードからなるデータセットを指定する必要があります（「モデルの適用」（17 ページ）と「バックフィッティング」（37 ページ）を参照）。回帰モデルによって、各レコードの新しい連続型ラベル値が予測されます。

「項目の削除」ボタン

モデルまたはグラフを簡素化する必要がある場合は、データセットから項目を削除しないで、計算の処理対象から項目を除外することができます。Tool Manager の「データの変換」パネルを開き、「現在のデータセットの項目名」ボックスから目的の項目を選択して「項目の削除」ボタンをクリックしてください。こうすると、項目が計算の処理対象から除外されますが、データセットからは削除されません。

投資利益率 (ROI) 曲線

投資利益率 (ROI: Return on Investment) 曲線は改善曲線と似ていますが、損失マトリックスが考慮に入れられ、誤差の観点ではなく損失の観点からモデル（アルゴリズム）の精度が評価されます。

Tool Manager で ROI 曲線を設定するときは、分析の「詳細設定」パネルの「誤差推定のオプション」セクションにあるチェックボックスを使用します。回帰ツールでは、ROI 曲線を使用できません。

ROI 曲線上の各ポイントは、各レコードの特定のラベル値に関する予想損失額に基づいて並べられます。また、各ポイントの高さは、そのポイントに至るまでの全レコードの収益（損失の逆）を累積した値を表します。精度（誤差の逆）を累積した値ではありません。

予想損失額を計算するには、損失マトリックスの特定のラベルの下にある金額と、クラシファイアによって該当クラスに割当てられた確率を掛合わせます（「損失マトリックス」（121 ページ）を参照）。したがって、クラシファイア の予測精度が高い場合は、予想損失額が小さくなり、レコードが ROI 曲線の左側付近に表示されます。

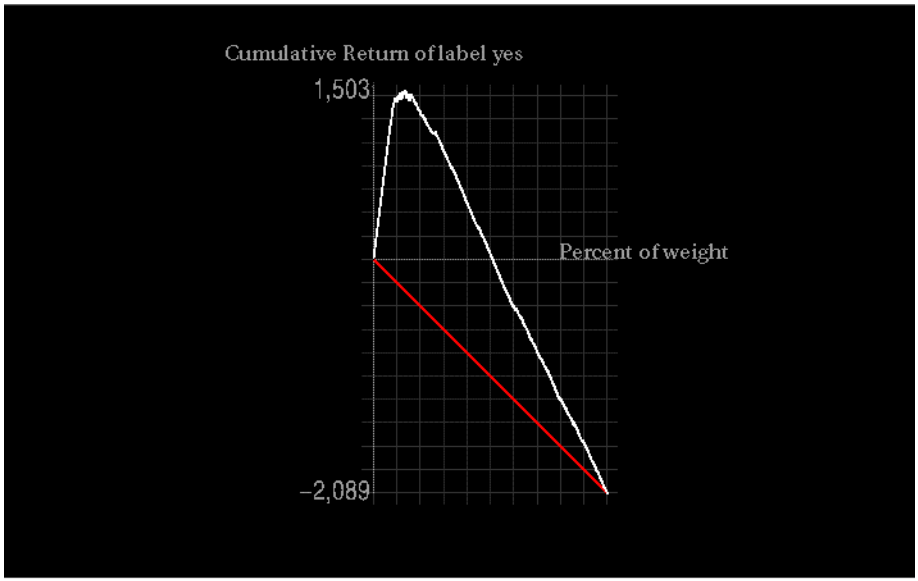

図 1-28 投資利益率 (ROI) 曲線

ROI 曲線では、データセット内の個々のレコードに対して、特定のラベル値に基づくアクションを実行することが前提となっています。たとえば、churn データセットでは、「はい (yes)」というラベル値に基づいて、特定の顧客に販売資料を送付し、その顧客が解約しないようなアクションが実行されます。ただし、アクションを無分別に実行すると、高いコストがかかることになります。ROI 曲線のピーク点は、クラシファイア に従って販売資料の送付先を選別した場合に、どの程度の費用が節約できるかを示しています。

ROI 曲線で使用される損失マトリックスに値を入力するときは、特別な注意を払う必要があります。予測される特定のラベル下の項目によって、そのラベル値に関する ROI 曲線が決定されます。この項目の値は、すべてのクラスについてラベル値に基づくアクションを実行した場合の予想収益または予想損失を表す値でなければなりません。たとえば、項目 "prediction yes"（解約すると予測される）と行 "actual value no"（実際には解約しない）が交差する箇所の値が 2 である場合は、解約しない顧客に対して販売資料を送付するアクション（"yes" に基づくアクション）の費用が 2 ドルであることを示します。.

ファイルの保存

Tool Manager の「ファイル」メニューを使用すると、現在作業中のセッションを保存することができます。また、Tool Manager の「データの可視化 / マイニング」パネルにある「データファイル」タブをクリックすると、特定のデータ変換方法を保存することができます。作業に戻るときは、Tool Manager の「ファイル 」メニューを使用してセッション・ファイルを読込んでください。そうすると、最後に保存した状態からセッションを再開することができます。

サンプルファイルのディレクトリ

MineSet にはさまざまなサンプルファイルが用意されていますが、それらが保存されているディレクトリはプラットフォーム（Windows または IRIX）に応じて異なります。

Windows システム上では、MineSet をインストールしたフォルダの配下の *\examples* サブフォルダにサンプルファイルがあります。

IRIX システム上では、*/usr/lib/MineSet/examples* ディレクトリにサンプルファイルがあります。

スキャタ・ビジュアライザ

スキャタ・ビジュアライザを使用すると、複数の変数間の関係をビジュアルに解析することができます。静的な表示とアニメーションによる表示が行えます。スキャタ・ビジュアライザは、レコード数がそれほど多くないデータセット（50,000 未満）の個々のデータポイントを調べる場合、または数多くのレコードを少数の離散的な階級生成に集計処理した場合に特に便利なツールです。レコード数が非常に多い場合は、スプラット・ビジュアライザを使用してください。スキャタ・ビジュアライザでは次の手法を使用して解析を行います。

- 3 次元 (3D) ランドスケープ
- 2 次元 (2D) スライダを備えたアニメーション・コントロール・パネル
- グラフィカル・オブジェクト（「要素」とも呼ばれます）― 3D ランドスケープ内のアニメーションによる表示を行うときに、各グラフィカル・オブジェクトのサイズ、色、位置を変化させることができます。

スキャタ・ビジュアライザ は、データセット内の各レコードまたは各行を 3D ランドスケープ内の要素（グラフィカル・オブジェクト）にマッピングして、データを可視化します。データ変数は要素のサイズ、色、位置にマッピングすることができます。さらに、1 つまたは 2 つの数値型変数をアニメーション・コントロール・パネルのスライダにマッピングすることもできます。要素のサイズ、色、または位置に割当てた変数がスライダに割当てた変数に依存する場合は、スライダを使用してアニメーションを動かすことができます。たとえば、いくつかの企業の売上高を時系列的に示すデータでは、時間変数をスライダにマッピングし、売上高を要素のサイズにマッピングすれば、時間スライダの動きに従って要素が伸張・収縮します。

スキャタ・ビジュアライザでは、可視化されたデータをさまざまな方法で解析することができます。アニメーション・コントロール・パネルでは、1D または 2D のアニメーション・パスに沿ってアニメーションを表示することができます。データに潜む特定の傾向や偏差を検出するには、アニメーション・パスを再生して、要素のサイズ、色、動きなどを観察します。3D ランドスケープでは、特定の次元や視点を強調表示することができます。また、特定の変数の値をスケーリングすると、その変数の特性が強調されます。フィルタリングを行って、一定の基準に合致する要素だけを表示することもできます。

必要なファイル

スキャタ・ビジュアライザを使用するには、データファイルと設定ファイルが必要です。

- タブで区切られたフィールド行からなるデータファイル
  通常、データファイルは Tool Manager を使用して簡単に作成できます。データファイルを手作業で作成する場合は、『*MineSet 3.0 Enterprise Edition Interface Guide*』の「Creating Data and Configuration Files for the Scatter Visualizer」を参照して、ファイル・フォーマットの定義を確認してください。

  データファイルを作成するには、データソース（Oracle、INFORMIX、Sybase などのデータベース）からデータを抽出し、スキャタ・ビジュアライザで使用できるようにフォーマットを変換します。データファイルの拡張子はユーザが定義します（スキャタ・ビジュアライザ用のサンプルファイルの拡張子は *.data* となっています）。

- 入力データのフォーマットとその表示方法を定義した 設定ファイル
  通常、設定ファイルは Tool Manager を使用して簡単に作成できますが、Tool Manager の代わりに任意のエディタ（メモ帳、jot、vi、Emacs など）を使用して作成することもできます（ファイル・フォーマットの定義については、『*MineSet 3.0*

Enterprise Edition Interface Guide』の「Creating Data and Configuration Files for the Scatter Visualizer」を参照してください)。

設定ファイルの拡張子は *.scatterviz* にしなければなりません。スキャタ・ビジュアライザを起動するとき、または ファイルを開くときは、データファイルではなく設定ファイルを指定してください。

スキャタ・ビジュアライザの起動

スキャタ・ビジュアライザを起動するには、次の 4 通りの方法があります。

- Tool Manager の「可視化ツール」タブを使用して、スキャタ・ビジュアライザを起動します。Tool Manager を通じてスキャタ・ビジュアライザを操作する方法については、「スキャタ・ビジュアライザの設定」(161 ページ)を参照してください。
- Tool Manager の「可視化ツール」メニューを使用して、スキャタ・ビジュアライザを起動します。すべての MineSet ツールで共通に使用される Tool Manager の機能については、本書の「Tool Manager」と『*MineSet 3.0 Enterprise Edition User' Guide for Windows*』を参照してください。
- 使用する 設定ファイルが分かっている場合は、その設定ファイルのアイコンをダブルクリックします。こうするとスキャタ・ビジュアライザが起動され、選択した設定ファイルが自動的に読み込まれます。この起動方法を利用できるのは、設定ファイルの拡張子が *.scatterviz* である場合に限られます(Tool Manager を使用してスキャタ・ビジュアライザ用の設定ファイルを作成すると、ファイル拡張子が常に .scatterviz になります)。
- IRIX シェルウィンドウのプロンプトに次のコマンドを入力して、スキャタ・ビジュアライザを起動します。

  `scatterviz` [*configFile*]

  configFile は、任意指定の引数であり、設定ファイルの名前を表します。コマンド行で設定ファイルを指定しなかった場合は、「ファイル」->「開く」オプションを選択してファイル名を指定する必要があります。

スキャタ・ビジュアライザの設定

- Windows システム上で Tool Manager を使用する場合:

  Tool Manager の「データの可視化 / マイニング」パネルで「可視化ツール」タブをクリックして、「スキャタ 」を選択します。こうすると、項目にマッピングできるさまざまな可視化要素がスキャタ・ビジュアライザのパネルに表示されます。各

テキスト・フィールドの右側にあるポップアップ・メニューから項目を選択してください。各ポップアップ・メニューから選択できるのは、各可視化要素に適したタイプの項目に限られます。

- IRIX システム上で Tool Manager を使用する場合：

  Tool Manager の「データの可視化 / マイニング」パネルで「可視化ツール」タブをクリックします。ポップアップ・メニューから「スキャタ・ビジュアライザ」を選択すると、パネルの右側にスキャタ・ビジュアライザのマッピング要件が表示されます。「可視化要素」のリスト内で先頭にアスタリスク (*) が付いている項目はマッピングが必要です。「現在のデータセットの項目名」リストから項目を選択して、パネルの右側の可視化要素にマッピングしてください。

- テキストエディタを使用して設定ファイルを作成する場合：

  Tool Manager を使用するとスキャタ・ビジュアライザを簡単に設定できますが、任意のテキストエディタを使用してスキャタ・ビジュアライザ用の設定ファイルを手作業で作成することもできます。設定ファイルのフォーマットの定義については、『*MineSet 3.0 Enterprise Edition Interface Guide*』の「Creating Data and Configuration Files for the Scatter Visualizer」を参照してください。

スキャタ・ビジュアライザ用のスライダの作成

時間の経過に応じて変化するデータのように、データセット内の項目のデータ値が独立に変化する場合は、スキャタ・ビジュアライザ用のスライダを作成すると便利です。スライダはアニメーションに必要なオブジェクトです。スライダを作成するには、手作業と自動作業の 2 通りの方法があります。詳細については、『*MineSet 3.0 Enterprise Edition User's Guide for Windows*』のスライダの作成に関する説明を参照してください。

スキャタ・ビジュアライザのオプション

スキャタ・ビジュアライザのさまざまなオプションを設定するには、Tool Manager の「データの可視化 / マイニング」パネルに戻り、「可視化ツール」タブをクリックして「スキャタビジュアライザ」を選択します。次に「ツールオプション 」ボタンをクリックすると、新しいダイアログ・ボックスが表示され、スキャタ・ビジュアライザのオプションのデフォルト値を変更できるようになります。

スキャタ・ビジュアライザの「ツールオプション」ダイアログ・ボックスには、次の 4 個のセクションがあります。

- 要素
- スライダ
- 軸
- その他

「要素」オプション

このセクションでは可視化要素の特性を指定して、スキャタ・ビジュアライザのグラフィック画面の外観を調整します。

- 「要素の説明オン」オプションでは、要素の説明の表示 / 非表示を指定します。
- 「要素のサイズ」オプションでは、要素のサイズの決定方法として、最大サイズ 、スケールサイズ、調整無し のいずれかを選択します。要素の説明の隣りに要素のサイズを表示するかどうかも指定することができます。
- 「要素の色」オプションでは、要素を表示するときの色を指定します。次のような指定方法があります。
  - 使用する色のリストを指定する。
  - マッピングの種類を指定する。
  - 色のリストを値のリストにマッピングする。
  - 色の説明の表示 / 非表示を指定する。
  - 色を要素にマッピングする。
- 「要素の形状」オプションでは、要素の形状（立方体、バー、球体、ひし形）を指定します。
- 「要素のラベルの色 」オプションをクリックすると、「色の選択」ダイアログ・ボックスが表示され、要素のラベルの色を変更することができます。
- 「要素名のサイズ」オプションをクリックすると、要素のラベルのサイズを変更することができます。小さい値を指定すればサイズが小さくなり、大きい値を指定すれば大きくなります。

「色」オプションを使用するには、事前に「データの可視化 / マイニング」パネルで「* 要素の色」に項目を割当てておく必要があります。色の選択と変更の詳細については、「色の選択」（53 ページ）を参照してください。

「色のリスト」オプションでは、色のリストのラベルの横にあるプラス (+) ボタンを使用して色のリストを指定します。色のリストを指定すると、カラーエディタが起動され、リストに追加する色を指定できるようになります。

「色のマッピング」オプションでは、グラフィック画面に表示される色の変化が「連続」的であるか、「離散」的であるかを指定します。「連続」を選択した場合、「色のリスト」フィールドと「色のマッピング」フィールドで指定した値に従って色が徐々に変化します。

ポップアップ・ボタンの右横にあるフィールドには、色を割当てる特定の値を入力します。色に割当てる値を指定しない場合は、色の変数と同じ範囲の値が使用されます。色の選択の詳細については、「色の選択」（53 ページ）を参照してください。

「サマリ」オプション

サマリスライダを使用すると、1 つまたは 2 つの追加の変数に基づくアニメーションを作成することができます。サマリスライダ上の各位置には、サマリ にマッピングされた変数（項目）の集計値を表す色が付けられます。「サマリ」オプションでは、サマリウィンドウに表示される変数の色を指定したり、サマリの説明（値の説明）の表示 / 非表示を切替えたりすることができます。アニメーションの詳細については、「アニメーション」（10 ページ）を参照してください。

2 次元配列型の値を使用する場合は、X スライダと Y スライダを指定することができます。オプションの横のポップアップ・ボタンには使用可能なキーの一覧が表示され、スライダとして使用するキーを選択することができます。

「スライダ」オプション

「スライダ」オプションでは、スライダに値をマッピングする方法を指定します。詳細については、「スキャタ・ビジュアライザ用のスライダの作成」（162 ページ）を参照してください。

「軸 」オプション

「軸」オプションでは、個々の軸について次の事項を指定します。

- 軸のラベル（このフィールドを空白にした場合、項目名が軸のラベルになります。）
- 軸の色
- 軸のサイズの決定方法（「最大サイズ」、「スケールサイズ」、または「調整無し」）
  - 「最大サイズ」を選択すると、データの値とは関係なく、軸が指定のサイズになります。たとえば、1 つの軸の最大サイズを他の軸の 2 倍と指定した場合、その軸のサイズは実際のデータ値とは関係なく、他の軸の 2 倍になります。このオプションは、単位が異なる複数の軸（たとえば、収入の軸と年齢の軸）を比較するときなどに便利です。このオプションは数値型のデータだけに適用されます。
  - 「スケールサイズ 」を選択すると、軸の最大値に基づいてその軸のサイズが決定されます。2 つの軸が同じ「スケールサイズ」であっても、一方の軸の最大値が他方の軸の 2 倍であるならば、前者の軸のサイズは後者の 2 倍になります。このオプションは、単位が同じ複数の軸（たとえば、収入の軸と支出の軸）を比較するときなどに便利です。このオプションは数値型のデータだけに適用されます。
  - 「調整無し」は「スケールサイズ」を 1.0 に設定した場合と同じです。
- 軸のサイズ
- ゼロの値を含むように軸を拡張するかどうかの選択

「他のオプション」

ダイアログ・ボックスの一番下にある「その他のオプション」には次のフィールドがあります。

- 「メッセージ」フィールドには、要素を選択したときに表示されるメッセージを入力します。このフィールドに入力できるメッセージのフォーマットの一覧と説明については、『*MineSet 3.0 Enterprise Edition Interface Guide*』の 「Creating Data and Configuration Files for the Scatter Visualizer」の「Message Statement」を参照してください。
- 「実行」フィールドには、要素をダブルクリックしたときに実行されるコマンドを入力します。入力のフォーマットは message ステートメントと似ています。このフィールドにコマンドを入力しなかった場合、要素をダブルクリックしても何も起こりません。詳細については、『*MineSet 3.0 Enterprise Edition Interface Guide*』の

「Creating Data and Configuration Files for the Scatter Visualizer」の 「Execute Statement」を参照してください。

- 「ラベルを隠す距離」オプションでは、要素のラベルが見えなくなるときの距離を指定します。この値を小さくするとパフォーマンスが向上しますが、ラベルはすぐに見えなくなります。値を大きくすると、ラベルが見えなくなるまで距離が延びます。
- 「軸の名称サイズ」オプションでは、軸のラベルのサイズを指定します。小さい値を指定するとラベルは小さくなり、大きい値を指定するとラベルは大きくなります。
- 「グリッド X サイズ」、「グリッド Y サイズ 」、「グリッド Z サイズ」オプションでは、各軸のグリッド線の間隔を指定します。小さい値を指定すると間隔は小さくなり、大きい値を指定すると間隔は大きくなります。ゼロ (0) を指定すると、グリッド線が表示されません。
- 「グリッドの色」オプションでは、グリッドの色を指定します。このオプションをクリックすると、「色の選択」ダイアログ・ボックスが表示され、グリッドの色を変更することができます。

ツールオプションのリセット

「ツールオプション」ダイアログ・ボックスで設定したオプションをすべてデフォルト値に戻すには、「リセット」ボタンをクリックします。

ツールオプションの保存

「ツールオプション」ダイアログ・ボックスで設定したオプションを確定して保存する場合は、「了解」ボタンをクリックします。「了解」ボタンをクリックすると、Tool Manager のメインウィンドウに戻ります。

アニメーション・コントロール・パネル

メインウィンドウの右側に表示されるアニメーション・コントロール・パネルには、サマリウィンドウ（その隣りに最大 2 つのスライダ）、情報フィールド、アニメーション・ボタン、アニメーション・スライダがあります。詳細については、「アニメーション」（10 ページ）を参照してください。

スキャタ・ビジュアライザにおける NULL 値の取扱い

スキャタ・ビジュアライザ (Scatter Visualizer) では、未知のデータ値（NULL 値）を含むフィールドをビジュアル属性（色など）に割当てたときに、特別な表現形式が使用されます（NULL 値の詳細については、『*MineSet 3.0 Enterprise Edition Interface Guide*』の「Nulls in MineSet」を参照してください)。NULL 値を要素のサイズに割当てると、要素が立方体の輪郭になります。NULL 値を色に割当てると、要素が濃いグレーになります。「選択」フィールドや「ポインタを通過」エリアでは、NULL 値が疑問符 (?) として表示されます。

要素の X 位置、Y 位置、または Z 位置に NULL 値を割当てた場合の表示は、「表示 」メニューの「NULL 位置付きの要素の表示」オプションの設定に応じて異なります (「表示」メニュー」（216 ページ）を参照)。このオプションをオンに設定すると、NULL 位置を持つ要素が該当の軸のすぐ下に表示されます。このオプションがオフである場合、NULL 位置を持つ要素は表示されません。

設定ファイルとデータファイルのサンプルファイル

スキャタ・ビジュアライザの特長や機能を紹介するために、設定ファイルとデータファイルのサンプルファイルが用意されています。これらのファイルの詳しい説明については、付録 A「設定ファイルとデータファイルのサンプルファイル」を参照してください。

「選択 」メニュー

「選択」メニューは、ほとんどのビジュアライザで同じ内容のものが使用されます。特定のビジュアライザで使用できないオプションは、そのビジュアライザのメニューに表示されません。「選択」メニューを使用すると、ドリルスルー (Drill Through) によって元のデータを細かく分析することができます。ドリルスルーを行うには、最初に特定の要素を選択した後、2 通りのドリルスルー方法のうち 1 つを選択します。

- 「選択ボックスの作成」オプションでは、伸張・移動して広い領域を選択できる 3D 選択ボックスが作成されます。このオプションを選択すると、レコードビューワ (Record Viewer) 形式のテーブルが表示され、各要素で表されるすべての集計データに関する情報が表示されます。テーブルを閉じると、3D 選択ボックスは消えますが、要素の選択は解除されません。レコードビューワ形式のテーブルには、3D 選択ボックスの内部にある要素、または <Shift> キーを押しながらクリックした要素が表示されます。選択ボックスを移動するには、ボックスの 1 面をマウスの左ボタンでクリックして、任意の方法にドラッグします。<Shift> キーを押しな

がらドラッグすると、ドラッグする方向に一番近い軸に移動することができます。選択ボックスを拡張・縮小するには、グレースケールのタブの 1 つを任意の方向にドラッグします。選択ボックスをウィンドウの境界の外側に伸張・移動することはできません。グレースケールのタブのサイズは、画面サイズを一定に保つように自動的に調整されます。タブのサイズが大きすぎる場合は、ズームインを行うと、選択ボックスのサイズを基準としてタブが縮小されます。

- 「値の表示」オプションでは、選択したすべての要素の値が表示されます。
- 「オリジナルデータの表示」オプションでは、選択された要素に対応するレコードが取出されて表示されます。レコードはテーブルビューワに表示されます。何も選択されていない場合、このオプションは使用できません（グレー表示されます）。
- 「Tool Manager に送信」オプションを選択すると、Tool Manager による履歴の作成開始時に、現在のボックス選択に基づいてフィルタ操作が挿入されます。ドリルスルーに使用される実際の式は、現在のボックス選択の内容によって決まります。何も選択されていない場合、このオプションは使用できません（グレー表示されます）。
- 「補集合データのドリルスルー」オプションを選択すると、「オリジナルデータの表示 」と「Tool Manager に送信」を使用した際に、選択されていないデータがすべて取出されます。
- 「項目選択してドリルスルー」オプションを選択すると、ドリルスルーで使用する項目を指定するためのウィンドウが表示されます。ビジュアライザではデータ分析にとって重要な項目（属性）が自動的には選別されないため、ドリルスルー式で使用する項目をユーザが明示的に指定する必要があります。たとえば、自動車のデータセットにメーカー、型番、重量の項目がある場合は、メーカーと型番だけを選択してドリルスルーを行うことができます。デフォルトでは、グラフィック要素にマッピングされたすべての項目が重要とみなされてドリルスルーの対象となります。その他の項目はドリルスルーの対象となりませんが、「項目選択してドリルスルー」ダイアログ・ボックスで項目を強調表示すると、その項目をドリルスルーで使用することができます。

ドリルスルーの詳細については、「ドリルスルー」（85 ページ）を参照してください。

マップ・ビジュアライザ、スキャタ・ビジュアライザ、スプラット・ビジュアライザのスライダの作成

項目をスライダに割当てると、特定の条件に応じて値がどのように変化するかを表示することができます。スライダに割当てることができる項目は、数値型（int, float,

double）の値または階級生成された値を持つ項目です。項目が既に階級生成されている場合は、項目名の最後に _bin という接尾語が付きます。項目の型は、「現在のデータセットの項目名」フィールドで項目名の後に表示されます。
例：total day calls - double

マップ・ビジュアライザ、スキャタ・ビジュアライザ、スプラット・ビジュアライザでは、Tool Manager の「データ変換」パネルにある「現在のデータセットの項目名」フィールドから数値型の項目または階級生成された項目を選択すると、スライダが自動的に作成されます。詳細については、「アニメーション用のボタンとスライダ」（14 ページ）を参照してください。

項目名のソート

データセットは変更しないで、項目名をアルファベット順にソートすることができます。こうすると、データが見やすくなります。項目名をソートするには、Tool Manager の「データ変換」パネルで、「項目をソートして表示」チェックボックスをクリックするか（Windows システムの場合）、「項目名をソート (Sort Column Names)」ボタンをクリックします（IRIX システムの場合）。

スプラット・ビジュアライザ

スプラット・ビジュアライザ (Splat) を使用すると、複数の変数（属性）間の関係をビジュアルに解析することができます。静的な表示と アニメーションによる表示が行えます。スプラット・ビジュアライザは、レコード数が非常に多いレコードセットの解析に適しています。レコード数があまり多くないレコードセットの個々のデータポイントを調べたいときは、スキャタ・ビジュアライザ を使用してください。スプラット・ビジュアライザでは、次の手法を使用して解析を行います。

- 3 次元 (3D) ランドスケープ
- 2 次元 (2D) スライダを備えたアニメーション・コントロール・パネル
- スプラット (*Splat*) と呼ばれるグラフィカル・オブジェクト ― 集計処理されたデータポイントを表します。3D ランドスケープ内でアニメーションによる表示を行うときに、スプラットの色と不透明度を変化させることができます（位置とサイズは不可）。

スプラット・ビジュアライザでは、軸、スライダ、色、不透明度に項目（属性）をマッピングすることによって、データがビジュアルに表現されます。その結果、各データポイントが別々にプロットされるスキャタ図に似た 3D ランドスケープが作成されます。ただし、スキャタ図とは異なり、互いに隣接する（同じ階級に属する）データポイントが集計処理されて単一のスプラットとしてプロットされます。

軸またはスライダに割当てられる個々の数値型項目は、事前に階級生成されていなければなりません。事前に階級生成されていない項目に対しては、Tool Manager によって自動的に均一な階級生成が実行されます（「階級生成」（39 ページ）を参照）。文字列型項目は軸に直接マッピングすることができます。数値型項目は色にマッピングすることができます。スプラットの色は、色にマッピングされている項目の値について、同一の階級に属する全データポイントの値を平均して算出されます。スプラットの不透明度は、同一の階級に属するデータポイントの数を重み付けして算出されます。不透明度にマッピングされている項目が存在しない場合は、レコード数に基づいて不透明度が決定されます。生成されるビジュアル図の操作性能は、軸に表示される階級の数のみに影響を受け、データポイントの数は影響を持ちません。

非常に大きいデータセットを扱うときは、Tool Manager を使用して明示的に集計処理を実行してください。そうすれば、サーバ側で集計処理が実行されるため、データセット全体をクライアントに送信して集計処理する必要がなくなります（「集計処理」（4 ページ）を参照）。

アニメーション・コントロール・パネルのスライダには、最大 2 つの数値型項目をマッピングすることができます。アニメーション・コントロール・パネルでスライダをパスに沿ってポイントからポイントへ移動すると、スプラットの色と不透明度が変化します。スキャタ・ビジュアライザとは異なり、スプラットの位置とサイズは変化しません。スプラットの位置とサイズは固定されており、均一に散布されます。変化するのは色と不透明度だけですが、これらが変化するとスプラットが実際に動いているように見えます。

文字列型項目が軸に割当てられている場合、その項目の個々の異なる値が 1 つの階級になります。文字列型の値が軸に並べられる順序は、色に割当てられた項目の平均値に基づいて個々の値をソートすることによって自動的に決定されます。文字列値の軸

に沿った色の変化を観察すれば、軸に割当てられた項目と色に割当てられた項目の相関の度合いを判定することができます。色に割当てられた項目が存在しない場合は、不透明度に基づいて個々の値が軸に並べられます。

スプラット・ビジュアライザの不透明度

不透明度にマッピングする項目は、レコードの重み付けに使用する項目（またはレコード数の項目）でなければなりません。スプラットの不透明度 (α) は、この項目に基づき、次の式に従って計算されます。

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

ここで、weight（重み）は不透明度にマッピングされた項目（不透明度にマッピングされた項目が存在しない場合はレコード数）です。この不透明度関数の形状を見れば分かる通り、weight（重み）の値が大きくなるに従って、不透明度が 1（完全な不透明）に近づきます。変数 u はスケーリング係数です。この係数の値は、メインウィンドウの左側にある不透明度スケールスライダによって調整されます。図 1-29 に、u の値が大きい場合と小さい場合の不透明度関数の形状を示します。図 1-30 には、同じ u の値に対応するグラフを示します。

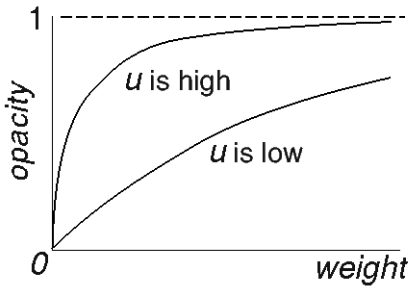

図 1-29 *u* の値が大きい場合と小さい場合の不透明度関数

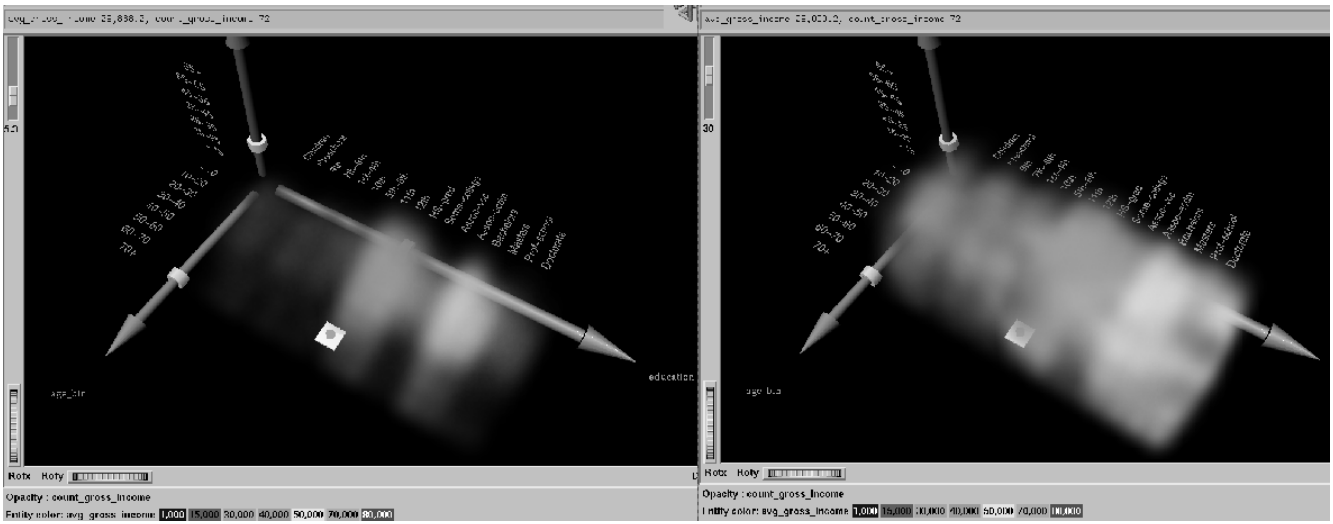

図 1-30 *u* = 5.3 と u = 30 の場合のグラフ

不透明度にマッピングされた項目が存在しない場合は、レコード数を表す項目が 集計処理時に自動的に生成されます。その場合、レコードの重み付けはすべて均一になります。レコード数を表す項目の値は合計されます。色にマッピングされた項目については、軸とスライダの項目ごとにグループ化され、それらの項目の値が平均されます。レコード数と色の項目を除くすべての項目は不必要であるため削除されます。各レコードに 1 以外の重みを割当てる場合を除いて、不透明度に項目をマッピングする必要はありません。

クライアントでなく Tool Manager を使用した集計処理の詳細については、『*MineSet 3.0 Enterprise Edition User' Guide for Windows*』の「スプラット・ビジュアイザの処理方法に関する説明」を参照してください。

ツールを起動すると、すべての処理がサーバ上で行われ、オリジナルデータの集計行が含まれる *adult94.splatviz.data* のデータファイルが処理されます。

特定の項目を使用してレコードの重み付けを行う必要がある場合もあります。下記の例では、人口の項目 (population) と平均給与 (avg_income) の項目を持つデータセット（サンプルファイル：adult94.splatviz.data）を使用しています、最初に、人口と平均給与を掛け合わせて新しい項目 (temp) を作成した後、その新しい項目と人口の項目を軸とスライダの項目ごとにグループ化して合計します。次に、平均給与合計を人口合計で加重した平均給与を表す新しい項目 (avg_salary) を作成します。最後に、人口合計を

不透明度にマッピングし、加重平均給与を色にマッピングします。加重平均給与の項目（avg_salary）は元の項目 (avg_income) と同様に平均給与を表していますが、その値は各行の人口に基づいて重み付けされています。こうして、平均給与の項目は引続き全人口の平均給与を表すことになります。

データセットのサイズが大きいために、クライアント側での集計処理やデータ保存を避けたい場合は、次の手順に従い、Tool Manager を使用して同じ集計処理を実行することができます。

1. temp = population*avg_income という定義式を使用して、新しい項目を作成します。
2. 軸とスライダの項目ごとにグループ化して、population（人口）と temp を合計します。
   この集計処理によって、合計値の項目 sum_population と sum_temp が作成されます。
3. 次の定義式を使用して、新しい項目を作成します。
   avg_salary = sum_temp/sum_population
   この項目は、平均給与合計を人口合計で加重した平均給与を表します。
4. sum_population を不透明度にマッピングし、avg_salary を色にマッピングします。

ユーザが Tool Manager を使用して集計処理を明示的に実行しない場合は、上記と同じ集計処理が（クライアント上の）スプラット・ビジュアライザによって自動的に実行されます。ただし、大量の集計処理は Tool Manager を通じてサーバ上で実行するほうが効率的です。そうすれば、クライアントに送信されるファイルのサイズもかなり小さくなります。

必要なファイル

スプラット・ビジュアライザを使用するには、データファイルと設定ファイルが必要です。

- タブで区切られたフィールド行からなるデータファイル
  通常、データファイルは Tool Manager を使用して簡単に作成できます。データファイルを手作業で作成する場合は、『*MineSet 3.0 Enterprise Edition Interface Guide*』の「Creating Data and Configuration Files for the Splat Visualizer」を参照して、ファイル・フォーマットの定義を確認してください。

データファイルを作成するには、データソース（Oracle、INFORMIX、Sybase などのデータベース）からデータを抽出し、スプラット・ビジュアライザで使用できるようにフォーマットを変換します。データファイルの拡張子はユーザが定義します（スプラット・ビジュアライザ用のサンプルファイルの拡張子は *.data* となっています）。

- 入力データのフォーマットとその表示方法を定義する設定ファイル
  通常、設定ファイルは Tool Manager を使用して簡単に作成できますが、Tool Manager の代わりに任意のエディタ（メモ帳、jot、vi、Emacs など）を使用して作成することもできます（ファイル・フォーマットの定義については、『*MineSet 3.0 Enterprise Edition Interface Guide*』) の「Creating Data and Configuration Files for the Splat Visualizer」を参照してください）。

  設定ファイルの拡張子は *.splatviz* でなければなりません。スプラット・ビジュアライザを起動するとき、またはファイルを開くときは、データファイルではなく設定ファイルを指定してください。

スプラット・ビジュアライザの起動

スプラット・ビジュアライザを起動するには、次の 4 通りの方法があります。

- Tool Manager を使用し、スプラット・ビジュアライザを設定して実行します。詳細については、『*MineSet 3.0 Enterprise Edition User's Guide for Windows*』を参照してください。
- Tool Manager の「可視化ツール」メニューから「スプラット・ビジュアライザ」を選択し、「ファイル」メニューの「開く」オプションを使用して設定ファイルを開きます。
- 使用する設定ファイルが分かっている場合は、その設定ファイルのアイコンをダブルクリックします。こうするとスプラット・ビジュアライザが起動され、選択した設定ファイルが自動的に読み込まれます。
  この起動方法を利用できるのは、設定ファイルの拡張子が *.splatviz* である場合に限られます（Tool Manager を使用してスプラット・ビジュアライザ用の設定ファイルを作成すると、ファイル拡張子が常に .splatviz になります）。
- IRIX シェルウィンドウのプロンプトに次のコマンドを入力して、スプラット・ビジュアライザを起動します。

  splatviz ［*configFile*］

  configFile は任意指定の引数であり、設定ファイルの名前を表します。コマンド行で設定ファイルを指定しなかった場合は、「ファイル」->「開く」オプションを選択してファイル名を指定する必要があります。

スプラット・ビジュアライザを起動するときの IRIX オプション

IRIX システム上でスプラット・ビジュアライザを起動するときに **-quiet** オプションを指定すると、進行状況を表すダイアログが表示されなくなります。このオプションを常に有効にするには、各ユーザのホーム・ディレクトリ内の *.Xdefaults* ファイルに次の行を追加します。

```
*minesetQuiet:TRUE
```

Windows システム上では、「ファイル」メニューから「設定」を選択して、このオプションを設定することができます。

「スプラット」オプション

このオプションでは、スプラット・ビジュアライザによって表示されるスプラットのさまざまな特性を指定します。

- 「スプラットの色」— スプラットに使用する色を指定します。次のような指定方法があります。
  - 使用する色のリストを指定する。
  - マッピングの種類を指定する。
  - 色のリストを値のリストにマッピングする。
- 「スプラットの形状」— スプラットをプロットするときの形状として、「線形 」、「ガウス」、「テクスチャ」、「球体」、「立方体」、「ダイヤモンド」のいずれかの形状を選択します。各形状の詳細については、「「形状」メニュー」（181 ページ）を参照してください。

色に関する下記のオプションを使用する場合は、事前に「データの可視化 / マイニング」パネルの「* 色」に項目を割当てておく必要があります。色のリストに何も入力されていない場合は、デフォルトのカラーマップが使用されます。デフォルトのカラーマップは青（最小値）から赤（最大値）までの連続スペクトルです。色の選択と変更の詳細については、「色の選択」（53 ページ）を参照してください。

「色のリスト」オプションでは、色のリストのラベルの横にあるプラス (+) ボタンを使用して色のリストを指定します。色のリストを指定すると、カラーエディタが起動され、リストに追加する色を指定できるようになります。

「色のマッピング」オプションでは、グラフに表示される色の変化が「連続」的であるか、「離散」的であるかを指定します。「連続」を選択した場合、「色のリスト」フィールドと「色のマッピング」フィールドで指定した値に従って色が徐々に変化します。

ポップアップ・ボタンの右横にあるフィールドには、色を割当てる特定の値を入力します。色に割当てる値を指定しない場合は、色の項目と同じ範囲の値が使用されます。色の選択の詳細については、「色の選択」（53 ページ）を参照してください。

「サマリ」オプション

「サマリ」オプションでは、サマリウィンドウに使用する色を選択します。このオプションを使用できるのは、サマリに項目をマッピングした場合だけです。

「その他のオプション」

ダイアログ・ボックスの一番下にある「他のオプション」には次のフィールドがあります。

- 「ラベルを隠す距離」— 軸のラベル（文字列型の値を持つ軸のラベル）が見えなくなるときの距離を指定します。この値を大きくすると、ラベルが見えなくなるまでの距離が延びます。
- 「軸の名称サイズ」— 軸のラベルのサイズを指定します。小さい値を指定するとラベルは小さくなり、大きい値を指定するとラベルは大きくなります。
- 「グリッドの色」— グリッドの色を指定します。このオプションをクリックすると、「色の選択」ダイアログ・ボックスが表示され、グリッドの色を変更することができます。
- 「グリッド X サイズ」、「グリッド Y サイズ」、「グリッド Z サイズ」— 各軸のグリッド線の間隔を指定します。小さい値を指定すると間隔は小さくなり、大きい値を指定すると間隔は大きくなります。ゼロ (0) を指定すると、グリッド線が表示されません。

ツールオプションのリセット

「ツールオプション」ダイアログ・ボックスで設定したオプションをすべてデフォルト値に戻すには、「リセット」ボタンをクリックします。

設定情報が保存されるファイル

Tool Manager では、スプラット・ビジュアライザに関する設定情報が複数のファイルに保存されます。これらのファイルには同じ接頭辞 <prefix> が付きます。

- *<prefix>.splatviz.data* にはデータが格納されます。
- *<prefix>.splatviz.schema* にはデータファイルが記述されます。
- *<prefix>.splatviz* にはスプラット・ビジュアライザが必要とする情報が格納されます。

ツールオプションの現在の設定情報とともにセッション全体を保存するには、Tool Manager の「ファイル 」メニューから次のいずれかのオプションを選択します。

- 「現在のセッションを保存 ...」オプション ― データソースの名前に基づいて、デフォルトの接頭辞が設定されます。
- 「現在のセッションを別名保存 ...」オプション ― ユーザが独自の接頭辞を指定します。

保存されるファイルの名前は *<prefix>.mineset* となります。このファイルには、MineSet を現在の状態に戻すのに必要な情報がすべて保存されます。

「ツールの起動」ボタンをクリックすると、*.data*、*.schema*、*.splatviz* の各ファイルが必要に応じて更新されます。

スプラット・ビジュアライザにおける NULL 値の取扱い

スプラット・ビジュアライザでは、未知のデータ値（NULL 値）を含むフィールドをビジュアル属性（色など）に割当てたときに、特別な表現形式が使用されます（NULL 値の詳細については、『*MineSet 3.0 Enterprise Edition Interface Guide*』の「Nulls in MineSet」を参照してください）。スプラットの色に割当てた項目について、階級生成 (bin) 内のすべてのレコードが NULL 値のときは、そのスプラットがグレーで表示されます。色に割当てた項目について、集計値の 1 つまたは複数のレコードが非 NULL 値である場合は、それらのレコードの値が色の算出に使用されます。（通常の）値と NULL 値の合計は NULL 値になりますが、値と NULL 値の平均はその値になります（すなわち、value + Null = Null になり、avg(val, Null) = val になります）。

「ピック」ウィンドウ、「選択」フィールド、「ポインタを通過」エリアなど、ビジュアライザのさまざまな表示ウィンドウ内では、NULL 値が疑問符 (?) で表されます。

軸にマッピングされた数値型項目に NULL 値が含まれている場合は、軸で定義される範囲の下に特別な NULL 位置が表示されます。そのため、NULL 値が他の値と非連続であることがすぐに分かります。数値型項目の軸の NULL 位置は、「表示」メニューの「NULL 位置の表示」オプションを使用して非表示にすることができます。軸にマッピ

ングされた文字列型項目の場合、NULL 値（'?' で表示）は他の通常の値と同様に扱われます。

スプラット・ビジュアライザ用のスライダの作成

サマリ ウィンドウの隣りに表示されるスライダの数は、設定ファイルで指定されたスライダのマッピング情報によって決まります。各スライダには、その特性（マッピングされているスライダ次元の数など）を示すラベルが付きます。

「スライダ 1」と「スライダ 2」にマッピングされる項目によって、スライダのインデックスが設定されます。これらの項目は数値型 (int, float, double) でなければなりません。スライダにマッピングされた項目が既に階級生成されている場合、階級自動生成は実行されず、その項目がスライダのインデックスとして使用されます。一方、項目が階級生成されていない場合は、自動的に均一な階級生成が実行されます（「階級生成」（39 ページ）を参照）。階級自動生成に使用された項目は現在のテーブルから削除されます。

アニメーション・コントロール・パネル

メインウィンドウの右側に表示されるアニメーション・コントロール・パネルには、サマリウィンドウ（その隣りに最大 2 つのスライダがある）、情報フィールド、アニメーション・ボタン、アニメーション・スライダがあります。詳細については、「アニメーション」（10 ページ）を参照してください。

スライダ上の個々の階級位置（サマリウィンドウ内では黒い点で示されます）には、メモリ内にある 1 つのデータテーブルが対応しています。1 次元のスライダを補間するときは、2 つの隣接したテーブルが結合された後、空間軸項目を固有キーとしてテーブルのデータが集計処理されます。1 つの階級位置から別の位置にスライダを移動すると、個々のスプラットの重み値（後で不透明度にマッピングされる値）が補間されます（その際、テーブルの特定の行が欠けている場合は、重みが 0 とみなされます）。スプラットの色に使用される平均値も補間されますが、重み値（またはレコード数）に基づく重み付けが適用されます。

例 1-4 補間処理

この例では、補間処理の技術的な側面を詳しく説明します。40-50 歳代のテーブルと 50-60 歳代のテーブルがスライダ上の 2 つの位置に対応しており、それらのテーブルのデータ値を補間する処理を考えます。表 1-18 に 40-50 歳代のテーブルのデータを示し、表 表 1-19 に 50-60 歳代のテーブルのデータを示します。

表 1-18 **40-50** 歳代のテーブル

education	occupation	hours_worked	income	weight
HS-grad	Exec-Man.	15-25	25000	2
HS-grad	Mach-op	15-25	30000	1
Masters	Technician	25-35	35000	3

表 1-19 **50-60** 歳代のテーブル

education	occupation	hours_worked	income	weight
HS-grad	Exec-Man.	15-25	70000	1
Vocational	Mach-op	35-45	40000	2

スプラット・ビジュアライザによる補間処理の仕組みを以下に説明します。

まず、40-50 歳代の表 1-18 には、*(1-t) (weight)* で定義される新しい重み（レコード数）項目と、*(1-t) (weight) (income)* で定義される新しい重み付け値項目が追加されます。50-60 歳代のテーブルには、*(t)(weight)* で定義される新しい重み（レコード数）項目と、*(t) (weight) (income)* で定義される新しい重み付け値項目が追加されます。次に、これら 2 つのテーブルが結合されます。

結合されたテーブルは空間軸項目をキーとして集計処理され、2 つの新しい項目が合計されます。このため、すべての空間軸について、同じ階級生成値を持つ行が重複する可能性がなくなります。最後に、値（この例では weight*income）の合計値を重みの合計値で割って、収入 (income) の補間値を算出します。t=.5 とした場合、表 1-20 に示すテーブルが生成されます。

表 1-20　40-50 歳代のテーブルと 50-60 歳代のテーブルの補間結果

education	occupation	hours_worked	income	weight
HS-grad	Exec-Man.	15-25	40000	1.5
HS-grad	Mach-op	15-25	30000	.5
Masters	Technician	25-35	35000	1.5
Vocational	Mach-op	35-45	40000	1

外部スライダが 2 次元の場合は、双線形補間が実行されます。

この人口調査データセットには約 150,000 行のデータがあります。外部スライダの目的は、データセットを効率的に分析して、データ内の新しい次元の要約情報を表示することです。赤い領域は要約値が大きい範囲を表し、白い領域は要約値が小さい範囲を表します。スライダを黒い点に置くと、補間されていないデータ値が表示されます。スライダの下の VCR コントロール・ボタンを使用すると、スライダ上のパスをたどるアニメーションを表示することができます。

次に、1990 年から 1997 年までの 8 年間のデータ（サマリウィンドウには 8 個のデータポイントが表示されます）を使用して、アニメーションの生成プロセスを説明します。最初に、スライダを別の年に移動して、スプラットがどのように変化するかを観察します。1990 年には、特定の位置にあるスプラットの値が 20 であり（この値は色にマッピングされます）、その重みは 2 である（2 つのレコードが存在する）と仮定します。また、1991 年にはこのスプラットの値が 40 になり、重みは 200 になると仮定します。

1991 年のスプラットは 1990 年よりもずっと不透明になります。これは、1991 年のスプラットが多数のレコード（重み付けが大きいレコード）の集計値を表しているためです。スライダを 1990 年から 1991 年に移動すると、重みは 2 と 200 の間で線形補間されて変化します。

補間値は、レコードの重み値で重み付けされた 2 つの値の平均値となります。たとえば、1990 年と 1991 年の中間点では、重みは (2+200)/2 = 101 になり、値は ((1-.5)*2*20+.5*200*40)/((1-.5)*2+.5*200) = 39.8 になります。1992 年に近づくに従って、値は 40 に近づきます。

アニメーションを離散データポイントの中間で停止することはできません。「パス (Path)」スライダを離散データポイント間の位置にドラッグして停止することもできません。サマリウィンドウ内のデータポイントは、データファイル内にある実際のデータに対応するスライダ位置を示します。たとえば、値 20 と 40 は実際のデータの集計値ですが、39.8 は加重平均値であり実際のデータの集計値ではありません。

スプラット・ビジュアライザのプルダウン・メニュー

スプラット・ビジュアライザには「ファイル」、「表示、「選択」、「形状」、「ヘルプ」の 5 種類のプルダウン・メニューが用意されており、さまざまな機能を利用することができます。「形状」以外のメニューの説明については、「「ファイル」メニュー」（100 ページ）、「表示」メニュー」（216 ページ）、「「選択 」メニュー」（167 ページ）、「IRIX システムの「ヘルプ (Help)」メニュー（Help （IRIX））」（105 ページ）を参照してください。

「形状」メニュー

スプラット・ビジュアライザではスプラット (Splat) を使用して、微小なポイントの群をモデリングします（Lee Westover 著 "Footprint Evaluation for Volume Rendering" in *Proceedings of SIGGRAPH '90*, Vol. 24, No. 4, pages 367-376 を参照）。

「形状」メニューでは、スプラットの描画（プロット）方法を指定します。正確性を増すと操作性能が犠牲になり、操作性能を向上させると正確性が犠牲になります。さまざまな分析で頻繁に使用される理論的なガウス分布（正規分布）を最も正確に表現するのは、テクスチャによるスプラットです。ほとんどのコンピュータではハードウェアによるテクスチャ・マッピングがサポートされるため、一般にはテクスチャによるスプラット（「テクスチャ」）を使用するのが最適な選択肢です。SGI プラットフォームのうち、ソフトウェアによる低速なテキスチャ・マッピングが実行されるのは Indy またはそれ以前のマシンだけです。次の 3 種類のスプラットが用意されています。

- 「線形」— 少数の三角形を描画し、ガウス分布を線形的に近似します。
- 「ガウシアン」— 数多くの三角形を描画し、ガウス分布を近似します。
- 「テクスチャ」— テクスチャ・マッピングされた長方形を使用して、ガウス分布を極めて正確に表現します。ハードウェア・テクスチャ・マッピングをサポートしていないマシンでは処理が非常に遅くなります。

なお、次に示す不透明なオブジェクト（プリミティブ）を選択することもできます。

- 「球体」— 不透明な球体を描画します。球体の半径はレコード数（または重み値）の立方根に応じて変化します。
- 「立方体」— 立方体を描画します。立方体の幅はレコード数（または重み値）の立方根に応じて変化します。

- 「ダイヤモンド」— ワイヤフレーム形の三角形を描画します。三角形のサイズはレコード数（または重み値）の平方根に応じて変化します。

設定ファイルとデータファイルのサンプルファイル

スプラット・ビジュアライザの特長や機能を紹介するために、設定ファイルとデータファイルのサンプルファイルが用意されています。これらのファイルの詳しい説明については、付録 A「設定ファイルとデータファイルのサンプルファイル」を参照してください。

分割の下限値

「分割の下限値」は、決定木分析と回帰ツリー分析 のサイズと精度を調整するためのオプションです。この値を大きくするとツリーのサイズが小さくなりますが、モデルの精度が低下する可能性があります。

「分割の下限値」オプションの値は、ノードの子ノードのうち、少なくとも 2 つのノードで設定する必要のある重み（重みが設定されていない場合はレコード数）の下限です。決定木分析では、このオプションのデフォルト値が 2 です。たとえば、ノードを 3 方向に分割する場合は、3 つのうち少なくとも 2 つの子ノードに 2 以上の重み（重みが設定されていない場合は 2 つ以上のレコード）を割当てる必要があります。これは決定木のサイズを制限する代替手段になります。

分割の下限値を増加させると、各枝上のレコード数（重み）が増えるため、予想確率の精度が改善される傾向があります。また、小さいツリーが構築されて、分析の実行時間が短縮されます。データにノイズ（誤差または異常値）が含まれていると思われる場合、またはツリーを使用して確率を予測する場合は（「モデルの適用」（17 ページ）を参照）、分割の下限値を 5 以上に増加させてください。データセットが非常に小さい（レコード数が 100 未満）場合は、この下限値を 1 に減少させてもかまいません（「決定木」（77 ページ）を参照）。

分割の基準

決定木の「分割の基準」オプションでは、ツリーの分岐に関する 3 種類の分割基準を選択することができます。下記の定義は技術的な説明であり、個別の問題に最も適した分割基準を判断するのは困難です。すべての分割基準を試して、誤差推定が最も小さくなる基準を選択するか、最も分かりやすい決定木が生成される基準を選択してください。

「相互情報量」オプションは、子ノードの加重平均純度と親ノード間の純度の変化（エントロピー）です。加重平均純度は、個々の子ノードに存在するレコードの数に基づいて計算されます。

デフォルトである「正規化された相互情報量」オプションは、「相互情報量」を子ノードの数の対数（底は 2）で除算した値です。

「増加比率 」オプションは、「相互情報量」を分割のエントロピーで除算した値です（ラベル値は無視します）。

「正規化された相互情報量」と「増加比率」は、数個の値しか取らない属性に適しています。

回帰ツリーの「分割の基準」オプションでは、分割対象となる属性（項目）が複数存在する場合に、どの属性を選択して分割を行うかを決める基準を指定します。

「分散」を指定すると、ツリーの各レベルにおけるノード内部の分散が最小になるような分割が行われます。枝（Leave）を生成するときは、その枝に到達するレコードのラベル値の平均が枝の予測値とされます。

「絶対偏差」を指定すると、ツリーの各レベルにおけるノード内部の絶対偏差が最小になるような分割が行われます。枝を生成するときは、その枝に到達するレコードのラベル値のメジアン（中央値）が枝の予測値とされます。

「正規化された分散」(デフォルト) を指定すると、分散の代わりに正規化された分散が使用されます。正規化された分散は、分散を子ノードの数の対数（底は 2）で割った値です。

「正規化された絶対偏差」を指定すると、絶対偏差の代わりに正規化された絶対偏差が使用されます。正規化された絶対偏差は、絶対偏差を子ノードの数の対数（底は 2 ）で割った値です。

統計量ビジュアライザ

統計量ビジュアライザを起動するには、Tool Manager の「データの可視化 / マイニング」パネルにある「可視化ツール」タブを使用します。統計量ビジュアライザでは、Tool Manager の「現在のデータセットの項目名」リスト内にある個々の項目（属性）について、その統計情報がメインウィンドウ内の 1 つの小さいパネルに表示されます（1 つの属性ごとに 1 つのパネルが表示されます）。ただし、メインウィンドウに一度に表示できる項目（属性）の数には制限があるため、側面のスクロールバーを使用して別の項目を表示するか、またはメインウィンドウを横方向または縦方向に拡大して表示可能な項目の数を増やしてください。

統計量ビジュアライザでは、Tool Manager で取り扱うデータセット内のレコード数に応じて特定の統計情報が表示されます。項目のパネルの形式は、項目のデータ型と、その項目に含まれている個別値の数に応じて異なります。一般的に、項目は「数値型」および「離散型」という 2 つのタイプに分けられます。数値型項目はボックスプロットで表示され、離散型項目はヒストグラム（柱状グラフ）で表示されます。

統計量ビジュアライザの使用方法

ボックスプロットは、数値型項目の値に基づいて描画されます。数値型項目は、整数 (integer)、浮動小数点数 (float)、倍精度浮動小数点数 (double)、日付 (date) のいずれかの値を取ります。ボックスプロットの各パネルには、単一の項目のデータに関する統計値が表示されます。統計値を構成するのは、最小値、最大値、平均値、中央値、2 種類の四分位数（25% と 75%）などです。これらの値は垂直バー（緑の目盛りのあるバー）を横断する直線として表示されます。母集団の標準偏差は +/- 付きの値として表示されます。項目の個別値の数が 50,000 個以下の場合には、四分位数が表示されます（図 1-31）。項目に 50,000 個を上回る個別値がある場合、統計値はグレーの垂直バーで表示されます。

平均値は、項目の全レコードの合計値をレコード数で割った値です。中央値（メディアン）は、特定の項目のレコード値を大きさの順に並べたときに中央に位置する値です。標準偏差は、項目の全レコードの散らばり（平均値からの乖離）を表す尺度です。

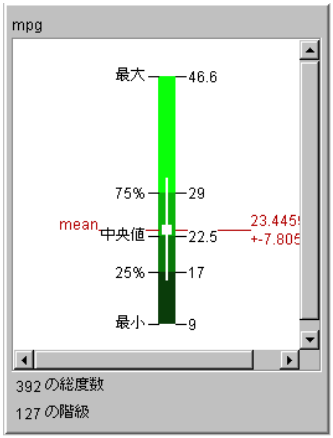

図 1-31 統計量ビジュアライザによる数値型項目の表示

ヒストグラム（柱状グラフ）は、離散型（名目型）項目の値に基づいて描画されます（図 1-32）。離散型項目は、数値以外の型（文字列、階級生成 (bin)、列挙型）の値を取ります。離散型項目のパネルには最大 100 個の個別値が示され、それらの個別値の度数がヒストグラムとして表示されます。デフォルトでは、個別値が度数の多い順に（降順で）表示されますが、「表示」プルダウン・メニューを使用すると昇順でソートすることもできます。個別値が 100 個以下の場合は、個別値のカウントも表示されます。

離散型の値（yes/no の値、州の名前など）を表示するときは、常にヒストグラムが使用されます。ボックスプロットとヒストグラムの各ボックスでは、データセット内の全レコード数が "Total vals" というラベルで示され、各ボックスで表される個別値の数が "Distinct vals" というラベルで示されます。

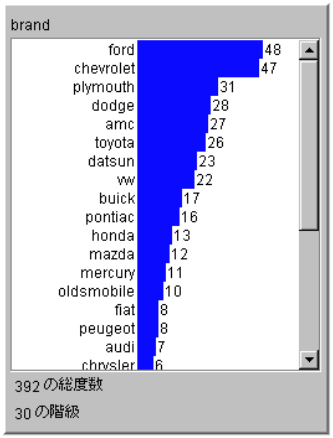

図 1-32 統計量ビジュアライザによる離散型項目の表示

統計量ビジュアライザをアイコンから起動した場合は、メインウィンドウに「ファイル」メニューと「ヘルプ」メニューしか表示されません。すべてのメニューとコントロールをメインウィンドウに表示するためには、設定 (*.statviz*) ファイルを開く必要があります。設定ファイルの一覧を表示するには、「ファイル」->「開く」を選択します。

統計量ビジュアライザのプルダウン・メニュー

統計量ビジュアライザには 3 種類のプルダウン・メニュー（「ファイル」、「表示」、「ヘルプ」）が用意されており、さまざまな機能を利用することができます。統計量ビジュアライザをアイコンから起動した（すなわち、設定ファイルを指定しないで起動した）場合は、「ファイル」メニューと「ヘルプ」メニューしか使用できません。「表示」以外のメニューの説明については、「ファイル」メニューと「ヘルプ」メニューを参照してください。

統計量ビジュアライザの「表示」メニュー

統計量ビジュアライザの「表示」メニューでは、ボックスプロットとヒストグラムで各項目をどのようにソートするかを指定します。

- 「属性値の度数順によるソート」を選択すると、離散型項目の個別値が度数の多い順（降順）にソートされ、ヒストグラムとして表示されます。
- 「属性値の名前順によるソート」を選択すると、離散型の個別値が名前順（アルファベット順）にソートされ、ヒストグラムとして表示されます。

「テーブルの履歴」ボタン

MineSet では一連の操作をデータテーブルに適用してデータテーブルを変換することができます。これらの一連の変換操作は記録されており、Tool Manager の「データ変換」パネルの一番下にある 2 つの 「テーブルの履歴」ボタンを使用して、特定の変換操作を確認・編集することができます。たとえば、いくつか前の段階に戻って、誤った操作を修正することができます。左向き矢印をクリックすると、項目のウィンドウに以前の状態のテーブルが示されます。右向きの矢印をクリックすると、テーブルは最新の状態に近づきます。

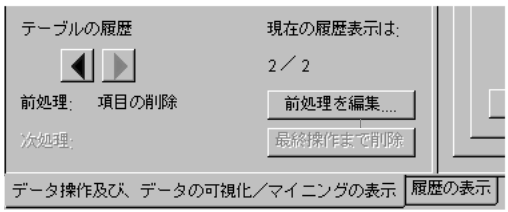

図 1-33 「テーブルの履歴」ボタン

「現在の履歴表示は」フィールド

「テーブルの履歴」ボタンの右側には「現在の履歴表示は」情報フィールドがあり、ユーザがテーブルに対して実行した変換操作の総回数と、現在表示されているテーブルの操作段階が示されます。このフィールドに表示される 2 つの数字を見ると、これまでテーブルに対して実行した一連の変換操作のうち、現在表示されているテーブルがどの段階にあるかを確認することができます。たとえば、テーブルを 2 回変更した場合は、最初のテーブル (1 of 3)、初回の変更後のテーブル (2 of 3)、2 回目の変更後のテーブル (3 of 3) を選択して表示することができます。

「前処理 :」フィールドと「次処理 : 」フィールド

「テーブルの履歴」ボタンを使用してテーブルの以前の状態を表示する場合、矢印ボタンの下にある「前処理 :」フィールドと「次処理 :」フィールドを見ると、現在の状態がテーブルの履歴のどの段階に該当するかを簡単に把握することができます。現在のテーブルを基準として、「前処理 : 」フィールドには直前の変換操作の種類が表示され、「次処理 : 」フィールドには直後の変換操作の種類が表示されます。

「前処理を編集」ボタン

「前処理を編集」ボタンを使用すると、「前処理」フィールドに表示された変換操作を編集することができます（「現在の履歴表示は」フィールドの値が「*1 of* *」であるときは直前の状態のテーブルが存在しないため、このボタンは使用できません）。「前処理を編集」ボタンをクリックすると、直前の操作を示すダイアログ・ボックスが表示され、操作内容を変更できるようになります。たとえば、以前の操作が項目の階級生成 (bin) であった場合、このボタンをクリックすると「項目の階級生成」ダイアログ・ボックスが表示されます。

直前の操作を変更すると、その後で実行した操作に影響が及ぶ場合があります。たとえば、後続の階級生成操作で使用していた項目を削除すると、その階級生成操作は無効になります。「履歴の編集」ボタンを使用すると、このような問題を回避することができます。

「最終操作まで削除」ボタン （Windows システムのみ）

テーブルの履歴をたどって以前の状態のテーブルに戻ると、テーブルが最新の状態（履歴の最後）ではなくなるため、Tool Manager のメインウィンドウの右側にある「データの可視化 / マイニング」パネルがグレー表示（使用不可能）になります。「データの可視化 / マイニング」パネルを使用可能にするためには、テーブルの履歴の最後に移動するか、または「最終操作まで削除」ボタンをクリックして現在のテーブルより後に適用した全操作を削除する必要があります。

「操作の履歴」タブ (Windows)

「操作の履歴」タブ（IRIX システム上では「履歴の表示 (*View History*)」ボタン）をクリックすると、「現在のデータセットの項目名」と「データの可視化 / マイニング」を示すウィンドウに代わって、「データ変換」テーブルの完全な履歴を示すウィンドウが表示されます（図 1-34 を参照）。このウィンドウでは、テーブルの各バージョンが 1 つのボックスとして表示されます。各ボックスの中には項目のリストが表示され、さら

に小さいボックス（テーブルに対して実行された操作を表すボックス）によって次のバージョンにリンクされています。

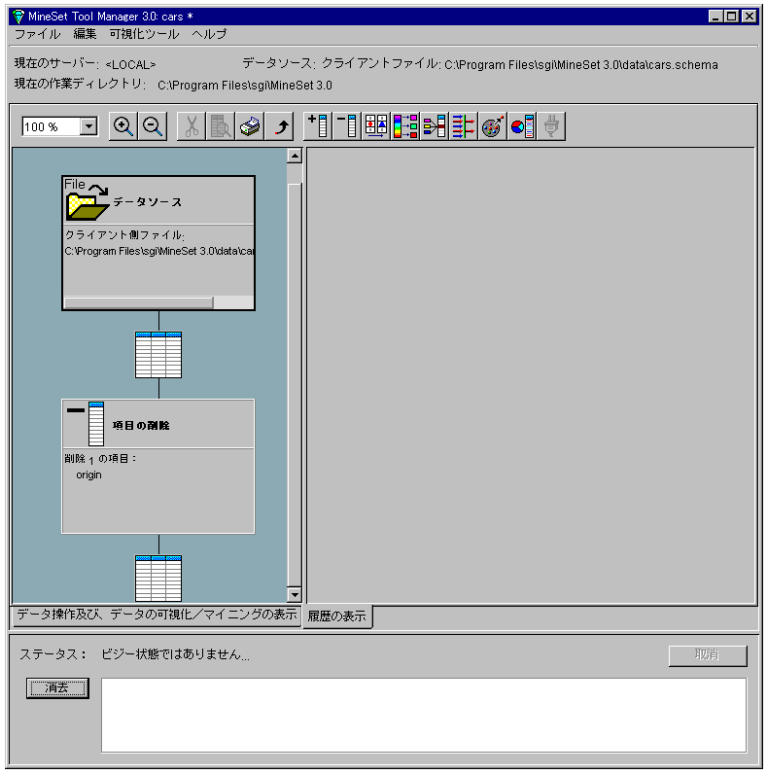

図 1-34 「履歴の表示」ダイアログ・ボックス (Windows)

「操作の履歴」タブのツールバー上にあるアイコンを使用すると、Tool Manager を通じてデータセットに適用したさまざまな操作を確認したり、データセットに対して操作を実行したりすることができます。「データの操作及び、データの可視化 / マイニングの表示」タブ（IRIX システム上では「単一操作 / 目的の表示 (*View Single Ops/Dest.*)」ボタン）をクリックすると、履歴を示すウィンドウが閉じて、現在の項目とデータの可視化 / マイニングを表示するウィンドウに戻ります（図 1-34 と図 1-35 を参照）。「操作の履歴」タブ（IRIX システム上では「履歴の表示」ボタン）をクリックすると、変換操作の履歴を示すウィンドウが再表示されます。

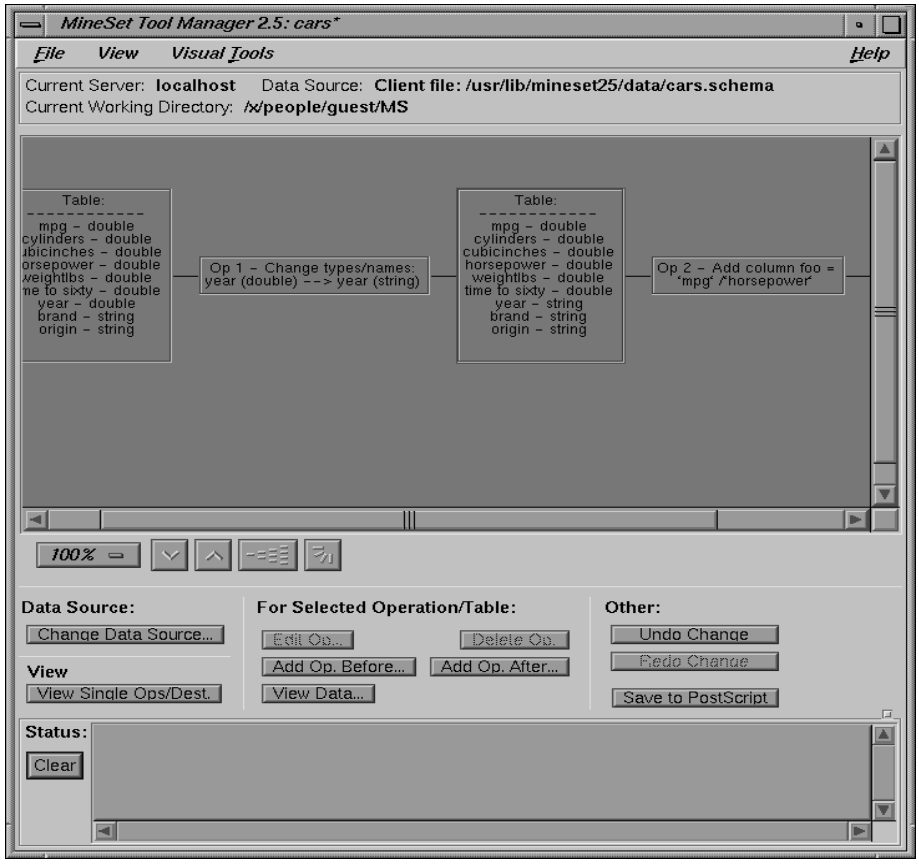

図 1-35 「履歴の表示」ダイアログ・ボックス (IRIX)

「前処理を編集」ボタンの場合と同様に、以前の操作を変更すると、履歴内の後続の操作に影響が及ぶ可能性があります（後続の操作が無効になることもあります）。以前の操作の変更によって後続の操作が影響を受ける場合は、警告メッセージが表示され、新しい履歴が示されます。

「履歴の表示 (View History)」ダイアログ・ボックス（IRIX システム）の下に並んでいるボタンを使用すると、上部のウィンドウに表示されているボックスのサイズと方向を変更することができます。

Tool Manager

Tool Manager は、MineSet ツール用のデータや設定を指定するためのグラフィカル・ユーザ・インタフェース (GUI) です。ここでは、Tool Manager による一般的な処理内容を説明します。詳細については、『*MineSet 3.0 Enterprise Edition User's Guide for Windows*』を参照してください。

Tool Manager は MineSet クライアント上で動作します。Tool Manager を使用するときの一般的な処理の流れは次の通りです。

1. Tool Manager がネットワークを介して DataMover に接続します。DataMover は MineSet サーバ上で動作します。MineSet のサーバとクライアントは、同じワークステーション上または別々のワークステーション上にインストールされています。
2. ユーザが Tool Manager を使用して次の事項を指定します。
   - データが格納されたデータベース、テーブル、バイナリファイル、ASCII（テキスト）ファイルなど（これらはクライアント上またはサーバ上に存在します。）
   - 使用するマイニングツールまたは可視化ツール
   - データを表示する方法（ツールオプション）
   - 作業履歴を保存するためのセッション・ファイル

   上記の操作では、DataMover を通じて取り出された情報が使用されます。Tool Manager は指定された情報に基づき、設定ファイルを作成します。設定ファイルには、下記の手順を実行するときに必要なユーザ定義のパラメータが格納されています。
3. ステップ 2 で作成された設定ファイルのコピーが Tool Manager から DataMover に送信されます。DataMover は設定ファイルの情報に基づいて、次の処理を実行します。
   - データベースまたはフラットファイルにアクセスする。
   - 指定されたデータ変換を実行する。
   - 要求に従ってマイニングツールを実行する。
   - 要求に従って可視化ファイルを生成する。

   可視化ファイルには、MineSet ツールで読取り可能なフォーマットで作成されたユーザデータが入っています。可視化ファイルのコピーは MineSet クライアントに転送されます。
4. Tool Manager によって適切な可視化ツールが起動されます。

5. 可視化ツールが可視化ファイルにアクセスして、データを表示します。
6. ユーザがモデルを作成している場合は、そのモデルを新しいデータに適用することができます。

Tool Manager の設定

Tool Manager の「設定」ダイアログ・ボックスでは、次のオプションを設定することができます。

- 「起動時、自動的にセッションの読込みを実行します」— MineSet にログインしたときに直前のセッションを自動的に復元します。MineSet ではセッション履歴が保存されており、最後にログアウトしたときと同じ状態でファイルが開かれます。
- 「バイナリ データファイルを使用」— MineSet のデータファイルを ASCII 形式ではなくバイナリ形式で保存します。こうすると、一般的に処理時間が短縮されます。
- 「最大の属性値」— データセット内の 1 つの項目（属性）で許容される個別値の最大数を設定します。個別値の数がこの値より多い項目（属性）は、MineSet による計算処理で使用されません。
- 「並列化処理」— IRIX システム上での並列処理に関するオプションを設定します。

訓練事例

訓練事例は複数の属性を含むテーブルであり、そのうち 1 つの属性が階級のラベルとして使用されます。ラベルとは、生成される分析モデルの予測対象となる既存の属性です。

次の例の目的は、アヤメの記述属性（sepal length（萼片の縦）、sepal width（萼片の横）、petal length（花びらの縦）、petal width（花びらの横））に基づいて、アヤメの種（iris-setosa（アイリスセトサ）、iris-versicolor（アイリスバージカラー）、iris-virginica（アイリスバージニカ））を予測することです。図 1-36 に、この例で使用する訓練事例に含まれる複数のサンプルレコードを示します。

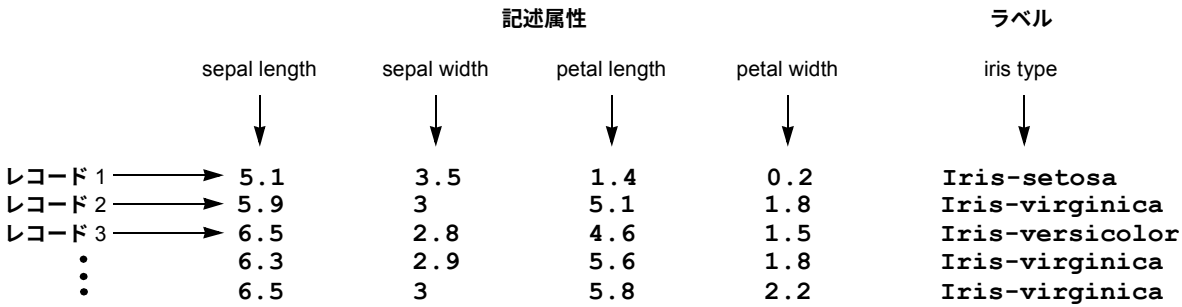

図 1-36 訓練事例内のサンプルレコード

クラシファイアが構築されると、そのクラシファイアに基づいて新しいレコードのラベル値を予測することができます。これらの新しいレコードは、クラシファイアで使用されるすべての属性（名前と型は訓練事例内の属性と同じ）が収録されたテーブル内に存在しなければなりません。このテーブル内にラベル属性が存在する必要はありません。存在している場合、そのラベル属性はクラス判別では無視されます。

ツリー・ビジュアライザ

ツリー・ビジュアライザは、データを 3 次元 (3D) のランドスケープとして表示するグラフィカル・インタフェースです。データは階層ブロック（ノード）とディスク付きのバーで表示されます。ユーザはこのランドスケープ中を自由に探索して、データセットの特定部分や全体を見ることができます。ツリー・ビジュアライザの詳しい使用方法については、『*MineSet 3.0 Enterprise Edition User's Guide for Windows*』を参照してください。

ツリー・ビジュアライザでは、階層的に接続されたノードとしてデータを表示することによって、データの定量的な特性や関係を明らかにします。各ノード上には複数のバーが表示されます。各バーの高さ、色、ディスクはデータの集計値を表しています。ノード間を接続する直線は、データセットとそのサブセット間の関係を示しています。

サブグループの値は、そのすぐ上のレベルで自動的に集計処理されて表示されます。バーの下のベースには、すべてのバーの集計値に関する情報が表示されます。負の値を表すバーはベースの下に表示されます。ベースの高さを無効にすれば、負の値をはっきりと見ることができます（「ツリー・ビジュアライザの表示 メニュー」（211 ページ））、『*MineSet 3.0 Enterprise Edition Interface Guide*』の「Creating Data and Configuration Files for the Tree Visualizer」の「Base Height Statement」を参照）。

必要なファイル

ツリー・ビジュアライザを使用するには、データファイルと設定ファイルが必要です。

- タブで区切られたフィールド行からなるデータファイル
  通常、データファイルは Tool Manager を使用して簡単に作成できます（『*MineSet 3.0 Enterprise Edition User's Guide for Windows*』を参照）。データファイルを手作業で作成する場合は、『*MineSet 3.0 Enterprise Edition Interface Guide*』の「Creating Data and Configuration Files for the Tree Visualizer」を参照して、ファイル・フォーマットの定義を確認してください。

  データファイルの拡張子はユーザが定義します（ツリー・ビジュアライザ用のサンプルファイルの拡張子は *.data* となっています）。

- 入力データのフォーマットと入力データを階層に変換する方法を定義する設定ファイル
  通常、設定ファイルは Tool Manager を使用して簡単に作成できますが（『*MineSet 3.0 Enterprise Edition User's Guide for Windows*』を参照）、Tool Manager の代わりに任意のエディタ（メモ帳、jot、vi、Emacs など）を使用して作成することもできます（ファイル・フォーマットの定義については、『*MineSet 3.0 Enterprise Edition Interface Guide*』の「Creating Data and Configuration Files for the Tree Visualizer」を参照してください）。

  設定ファイルの拡張子は *.treeviz* でなければなりません。ツリー・ビジュアライザを起動するとき、またはファイルを開くときは、データファイルではなく設定ファイルを指定してください。

ツリー・ビジュアライザの起動

ツリー・ビジュアライザを起動するには、次の 4 通りの方法があります。

- Tool Manager を使用し、ツリー・ビジュアライザを設定して実行します。詳細については、『*MineSet 3.0 Enterprise Edition User's Guide for Windows*』を参照してください。
- Tool Manager の「可視化ツール」メニューから「ツリービジュアライザ」を選択し、「ファイル」メニューの「開く」オプションを使用して設定ファイルを開きます。
- 使用する設定ファイルが分かっている場合は、その設定ファイルのアイコンをダブルクリックします。こうするとツリー・ビジュアライザが起動され、選択した設定ファイルが自動的に読み込まれます。この起動方法を利用できるのは、設定ファイルの拡張子が *.treeviz* である場合に限られます（Tool Manager を使用してツ

リー・ビジュアライザ用の設定ファイルを作成すると、ファイル拡張子が常に .splatviz になります)。

- IRIX シェルウィンドウのプロンプトに次のコマンドを入力して、ツリー・ビジュアライザを起動します。

  treeviz [*configFile*]

  configFile は任意指定の引数であり、設定ファイルの名前を表します。コマンド行で設定ファイルを指定しなかった場合は、「ファイル」->「開く」オプションを選択してファイル名を指定する必要があります。

警告システムを有効にして、ビジュアライザの起動時にダイアログ・ボックスが表示されないようにすることもできます。「警告オプション」(219 ページ)を参照してください。

ツリー・ビジュアライザのオプション

「ツールオプション」ボタンをクリックすると、「設定オプション」ダイアログ・ボックスが表示され (図 1-37 と 図 1-38)、ツリー・ビジュアライザのオプションのデフォルト値を変更できるようになります。

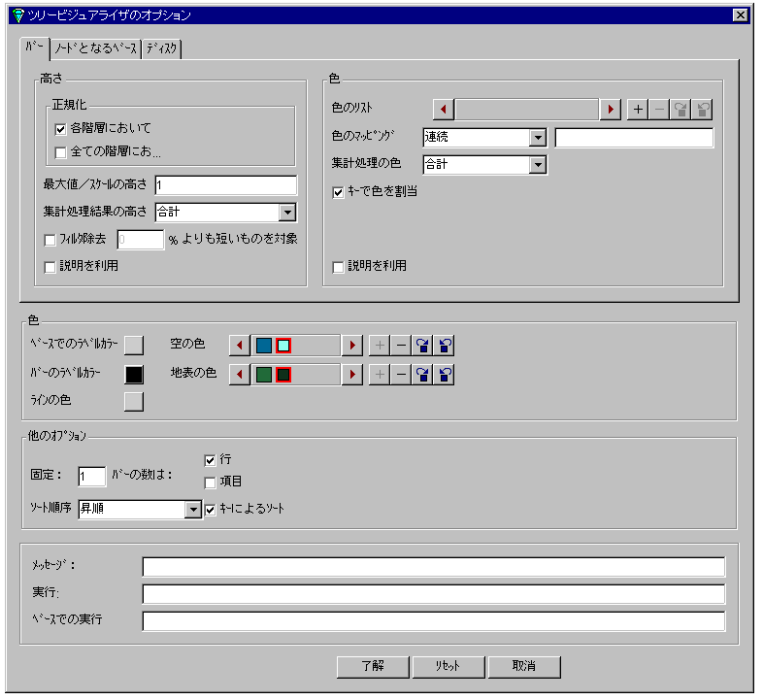

図 1-37 ツリー・ビジュアライザの「設定オプション」ダイアログ・ボックス (Windows システム)

Windows 版のダイアログ・ボックスの一番上には、「バー」、「ノードベース」、「ディスク」という 3 つのタブがあります。これらの各タブをクリックすると、ツリー・ビジュアライザの 3D ランドスケープの外観を細かく設定するための複数のオプションが表示されます。色の選択と変更の詳細については、「色の選択」（53 ページ）を参照してください。グラフィック・オブジェクトの高さの調整方法については、「高さの正規化」（198 ページ）を参照してください。

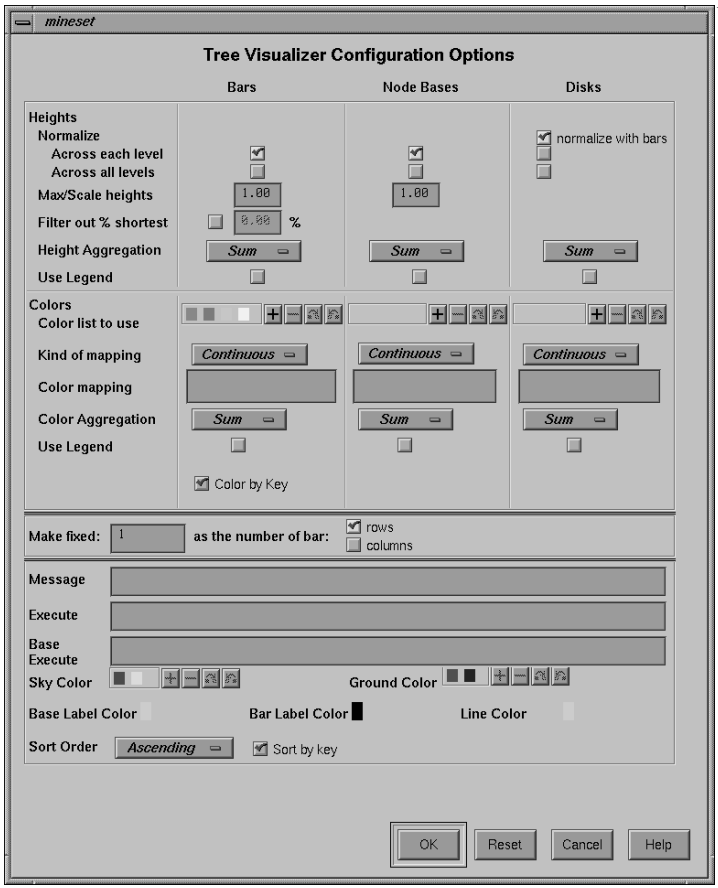

図 1-38 ツリー・ビジュアライザの「設定オプション (Configuration Options)」ダイアログ・ボックス （IRIX システム）

IRIX 版のダイアログ・ボックスの一番上には、「バー (*Bars*)」、「ノードベース (*Node Bases*)」、「ディスク (*Disks*)」という 3 つのタブがあります。ツリー・ビジュアライザに関する全オプションは、このダイアログ・ボックスを使用して設定することができます。

高さの正規化

階層の各レベルで（またはすべてのレベルで）バー、ノードベース、ディスクの高さを正規化するには、「高さ」セクションの「正規化」オプションを使用します。高さを正規化すると、高さの最大値が決定されます（すべての値が高さの最大値を基準として正規化されます）。たとえば、最大値が 30.0 で、バーの高さの最大値が 1.0（単位は任意）である場合、値 15.0 は高さ 0.5 に割当てられます。

「各階層において」を選択すると、階層の各レベルが別々に正規化されます。この設定が最も効果的なのは、階層の各レベルの値が別々に集計処理されている場合です。階層の各レベルを別々に正規化すると、階層の最上位レベルの項目によって最下位レベルの項目が隠されるといった不具合がなくなります。「全ての階層において」を選択すると、階層のレベルに関係なく、すべての項目が一緒に正規化されます。「バー 」タブで「正規化」オプションの 2 つのチェックボックスを両方ともオフにすると、正規化は行われません。

ノードベースはバーとは別に正規化されます。「ノードベース」タブで「正規化」オプションの 2 つのチェックボックスを両方ともオフにすると、バーに適用されるのと同じ方法でノードベースの正規化が行われます（ただし、バーとノードベースの値は別々に正規化されます）。

ディスクが表示されるときに、「バーの設定で正規化」チェックボックスをオンに設定すると、ディスクはバーの設定で正規化されます。同じ値を表すディスクとバーは同じ高さになります。「ディスク」タブで、「各階層において」または「全ての階層において」のいずれかのチェックボックスをオンに設定すると、バーとは別にディスクが正規化されます。すなわち、実際の値とは関係なく、一番高いディスクと一番高いバーの高さが同じになります。

最大値 / スケールの高さ

「最大値 / スケールの高さ」オプションを使用すると、一番高いバーとノードベースの高さを指定することができます。デフォルトは 1.0 です（単位は任意）。表示されるバーまたはノードベースが高すぎる場合（または低すぎる場合）は、このフィールドを使用して高さを調整することができます。たとえば、このフィールドに "2" を入力するとバーの高さがすべて 2 倍になり、"0.5" を入力するとバーの高さがすべて 2 分の 1 になります。

高さの正規化が指定されている場合、このオプションの値は一番高いバーまたはノードベースの高さを表します。正規化が指定されていない場合は、すべての高さがこのオプションの値を倍率として拡大・縮小されます（2 つの異なるデータセットを比較するときに便利な機能です）。

% より短いものをフィルタにより除去

「% より短いものをフィルタにより除去」オプションでは、フィルタリング条件を指定して、短いバーだけを含むノードを除去します。最初に、一番高いバー（高さがレベルごとに正規化されている場合は、各レベルの一番高いバー）が計算されます。続いて、高さが（一番高いバーに対して）一定のパーセントに達するバーを少なくとも 1 つ含むノードだけが表示されます。たとえば、このフィールドに "5%" を入力すると、一番高いバーの 5% 以上に相当する高さのバーを少なくとも 1 つ含むノードだけが表示されます（該当のノードの上位ノードも表示されます）。このオプションは重要性の低い小さいノードを除去する大まかな方法であり、特定の値を持つ特定のノードを正確に識別するメカニズムではありません。このオプションを使用すると、複雑なデータのレンダリングが高速になります。また、ゼロ近くに数多くのバーが集中しなくなるため、画面がすっきりします。

小さいノードは表示画面から除去されますが、階層の計算には組込まれます。

集計処理結果の高さ

デフォルトでは、親ノードのバーの高さは、子ノードの全バーの高さを合計した値になります。ただし、「集計処理結果の高さ」オプションを使用すると、子ノードの全バーの高さの平均値、最大値、最小値、カウント、または他の任意の値を、親ノードのバーの高さに割当てることができます。このオプションは、バーの高さ、ノードベースの高さ、ディスクの高さについて別々に指定することができます。

色

「色」オプションでは、次の事項を指定します。

- 使用する色のリストを指定する。
- マッピングの種類を指定する。
- バー、ノードベース、ディスクに色を割当てる。

「色」オプションを使用するには、「データの可視化 / マイニング」パネルで「バーの色」、「ディスクの色」、「ベースの色」に項目を割当てておく必要があります。色の選択と変更の詳細については、「色の選択」（53 ページ）を参照してください。

「使用する色のリスト」オプションでは、色のリストのラベルの横にあるプラス (+) ボタンを使用して色のリストを指定します。色のリストを指定すると、カラーエディタが起動され、リストに追加する色を指定できるようになります。

「マッピングの種類」オプションでは、グラフィック画面に表示される色の変化が「連続」的であるか、「離散」的であるかを指定します。「連続」を選択した場合、「使用する色のリスト」フィールドと「色のマッピング」フィールドで指定した値に従って、バー、ノードベース、またはディスクの色の値が徐々に変化します。「離散」を選択すると、指定された境界で色が段階的に変化します。

「色のマッピング」オプションでは、色が割当てられる値を指定します。

集計処理の色

デフォルトでは、親ノードのバーの色は、子ノードの全バーの色を合計した値になります。ただし、「集計処理の色」オプションを使用すると、子ノードの全バーの色の平均値、最大値、最小値、カウント、または他の任意の値を、親ノードのバーの色に割当てることができます。このオプションは、バーの色、ノードベースの色、ディスクの色について別々に指定することができます。

キーで色を割当

「キーで色を割当」オプションをオンに設定すると、キーの値に基づいてバーの色が自動的に割当てられます。色の割当てについて別の方法が指定されている場合、このオプションは無視されます。色のリストが指定されていない場合、または指定した色の数が不足している場合は、色がランダムに選択されて追加されます。余分な色を指定した場合、それらの色は無視されます。

固定

デフォルトでは、すべてのバーが 1 行に配置されます。「固定」オプションを使用すると、行数または項目数を変更することができます。行数も項目数も指定されていないか、または 0 が指定されている場合は、行数も項目数も固定されず、平方値に一番近い値が使用されます。

メッセージ

「メッセージ」フィールドには、任意のメッセージを入力することができます。マウスのポインタをオブジェクト上に移動したとき、またはオブジェクトを選択したとき、ここに入力したメッセージが表示されます。デフォルトでは、バーとノードベースで同じメッセージが表示されます。メッセージが入力されていない場合は、全項目の名前と値を示すデフォルトのメッセージが表示されます。

メッセージの書式は、使用されているデータの型に一致していなければなりません。

- 文字列 (string) には、%s を使用します。
- 整数 (int) には、整数の形式（例：%d）を使用します。
- 浮動小数点 (float) と倍精度浮動小数点 (double) には、浮動小数点の形式（例：%f）を使用します。

「メッセージ」フィールドの詳細については、『*MineSet 3.0 Enterprise Edition Interface Guide*』の「Creating Data and Configuration Files for the Tree Visualizer」の「Message Statement」を参照してください。

「実行」と「ベースでの実行」

「実行」フィールドと「ベースでの実行」フィールドには、バーまたはノードベースをダブルクリックしたときに実行されるコマンドを入力します。「実行」フィールドだけにコマンドを入力した場合、そのコマンドはバーとベースの両方に適用されます。両方のフィールドにコマンドを入力した場合、「実行」フィールドのコマンドはバーに適用され、「ベースでの実行」フィールドのコマンドはノードベースに適用されます。入力形式はメッセージのステートメントと似ています。このフィールドにコマンドを入力しなかった場合は、バーやベースをダブルクリックしても何も実行されません。

「実行」フィールドの詳細については、『*MineSet 3.0 Enterprise Edition Interface Guide*』の「Creating Data and Configuration Files for the Tree Visualizer」の「Execute Statement」を参照してください。

空の色

「空の色」オプションでは、グラフの背景（空）の色として 1 つまたは 2 つの色を指定します。色を 1 つだけ指定した場合、空は単色（単一の色調）になります。2 つの色を指定した場合は、最初の色が空の上部の色になり、2 番目の色が空の下部の色になって、それら 2 つの色の間で色調が徐々に変化します。

地表の色

「地表の色」オプションでは、グラフの地表の色として 1 つまたは 2 つの色を指定します。色を 1 つだけ指定した場合、地表は単色（単一の色調）になります。2 つの色を指定した場合は、最初の色が遠方の地平線の色になり、2 番目の色が近くの地面の色になって、それら 2 つの色の間で色調が徐々に変化します。

ベースでのラベルカラー

「ベースでのラベルカラー」オプションでは、ノードベースのラベルの色を指定します。

バーのラベルカラー

「バーのラベルカラー」オプションでは、バーのラベルの色を指定します。

ラインの色

「ラインの色」オプションでは、ノードベース間を接続するライン（直線）の色を指定します。

ソート順序

「キーによるソート」チェックボックスをオンに設定すると、各ノードがキーによってソートされた順序で表示されます。昇順ソートまたは降順ソートの区別は、チェックボックスの左側の「ソート順序」オプションで選択することができます。

ツールオプションのリセット

「ツールオプション」ダイアログ・ボックスで設定したオプションをすべてデフォルト値に戻すには、「リセット」ボタンをクリックします。

ツールオプションの保存

「ツールオプション」ダイアログ・ボックスで設定したオプションを確定して保存する場合は、「了解」ボタンをクリックします。「了解」ボタンをクリックすると、Tool Manager のメインウィンドウに戻ります。

設定情報が保存されるファイル

Tool Manager では、ツリー・ビジュアライザに関する設定情報が複数のファイルに保存されます。これらのファイルには同じ接頭辞 <prefix> が付きます。

- *<prefix>.treeviz.data* にはデータが格納されます。
- *<prefix>.treeviz.schema* にはデータファイルが記述されます。
- *<prefix>.treeviz* にはツリー・ビジュアライザが必要とする情報が格納されます。
- *<prefix>.mineset* には MineSet の他のファイルを作成するのに必要なすべての情報が格納されます。

接頭辞 (prefix) を指定するには、Tool Manager のメインウィンドウで「ファイル (File)」メニューの「現在のセッションを別名保存 ... 」オプションを選択します。接頭辞を指定しなかった場合は、データソースを表す接頭辞が付きます。

「ツールの起動」ボタンをクリックすると、*.data*、*.schema*、*.treeviz* の各ファイルが必要に応じて更新されます。

ツリー・ビジュアライザのプルダウン・メニュー

ツリー・ビジュアライザには 5 種類のプルダウン・メニュー（「ファイル」、「表示」（IRIX システム上では「表示 (Show)」）、「表示 (Display)」、「移動 (Go)」、「ヘルプ (Help)」）が用意されており、さまざまな機能を利用することができます。「ファイル」メニューは他の MineSet ツールと同じであるため、「「ファイル」メニュー」（100 ページ）を参照してください。

「表示」メニュー

「表示」メニュー（IRIX システム上では「表示」メニュー）には、「概要」、「検索パネル」、「フィルタパネル」、「マークパネル」という 4 つのオプションがあります。これらのオプションを選択すると、データを操作するための別のダイアログ・ボックスが表示されます。

オブジェクトの検索

「表示」メニュー（IRIX システム上では「表示 (Show)」メニュー）から「検索」を選択すると、オブジェクトを検索するための条件を指定するダイアログ・ボックスが表示されます（図 1-39 と 図 1-40）。

図 1-39 ツリー・ビジュアライザの「検索」ダイアログ・ボックス (Windows)

階層の一部分（特定のレベル）だけを指定して検索を行うことができます。デフォルトでは、階層全体が検索の対象となります。特定のレベルだけを検索する場合は、オプションメニューから関係演算子（<= など）を選択した後、「レベル」スライダを使用して検索対象のレベルを選択します。レベル 0 は階層のルートレベルであり、レベ

ル 1 はルートのすぐ下のレベルです（以下同様）。たとえば、ルートレベルとその下にある 2 つのレベルを検索する場合は、<= 2 と指定します。

検索対象をバーまたはベースに絞るチェックボックスも用意されています。

「階層」フィールドでは、検索するノードを指定します。「階層」フィールドの下には、個々の項目の検索条件を指定するためのフィールドがあります（検索条件で使用する項目は、Tool Manager の「テーブル処理」タブにある「現在のデータセットの項目名」ウィンドウで選択します）。

検索時に大文字と小文字の区別を無視する場合は、「大文字 / 小文字を区別せずに検索」チェックボックスにチェックマークを付けます。こうすると、たとえば、文字列 "hello" と "HellO" は同じと判断されます。

「NULL をゼロとみなす」チェックボックスのデフォルト設定はオフですが、その場合、検索時に NULL 値を含む値を比較すると、その結果は常に FALSE（偽）になります。このチェックボックスにチェックマークを付けると、NULL 値がゼロとみなされて値の比較が行われます。

バーを検索するときのデフォルト設定では、すべてのバーが検索されます。特定のバーを選択して検索する場合は、検索対象のバーを選択する必要があります。「全てに設定」ボタンをクリックすると、すべてのバーが選択されます。少数のバーだけを除いて残りのすべてのバーを検索するときは、このボタンを使用すると便利です。「消去」ボタンをクリックすると、すべてのバーが選択解除されます。バーがまったく選択されていない場合は、すべてのバーが検索されます。

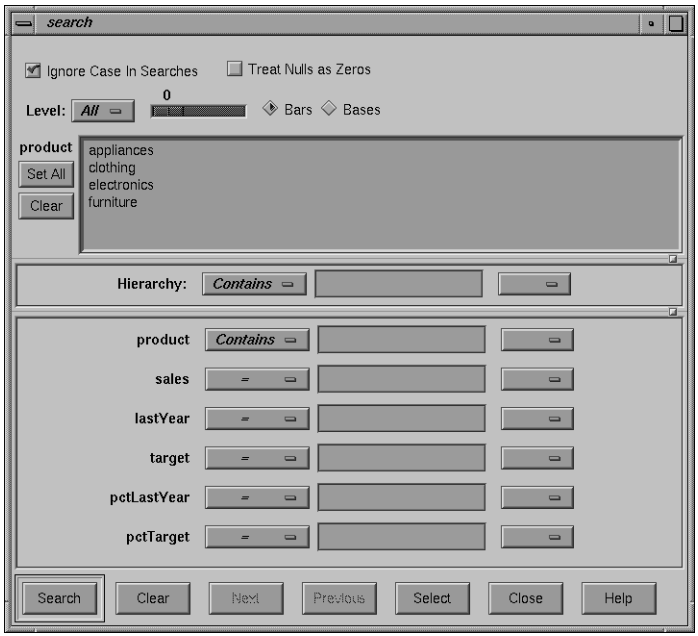

図 1-40 ツリー・ビジュアライザの「検索 (Search)」ダイアログ・ボックス (IRIX)

数値を検索する場合は、数値を入力してから、関係演算子 (=、!=、>、<、>=、<=) を選択します。アルファベットの文字列を検索する場合は、検索する文字列を入力します。文字列の検索では、次の 3 種類の比較条件を指定することができます。

- 「含む」は、指定された文字列を含むことを示します。たとえば、California は文字列 Cal と forn を含んでいます。
- 「等しい」は、文字列が正確に一致することを示します。
- 「一致」では、次のワイルドカードを使用することができます。
  - アスタリスク (*) は、任意の数の文字を表します。
  - 疑問符 (?) は、任意の 1 文字を表します。
  - 角かっこ ([ ]) は、その中に囲まれた文字の 1 つを表します。

  たとえば、California は、Cal*、Cal?fornia、Cal[a-z]fornia に一致します。

場合によっては（特に Tool Manager の階級生成 (binning) 機能を使用している場合)、テキスト・フィールドの代わりに、値のオプションメニューが表示されます。このよう

な変数を無視する場合は、オプションメニューで「無視」を選択します。これらのオプションについては、関係演算子（>= など）を使用することができます。こうすれば、指定した値だけでなく、それに後続する値も選択されます。

数値と文字列の比較演算子の他に、`Is Null` という演算子を使用することができます。この演算子は、値が NULL の場合に TRUE となります。

各フィールドの右側には追加のオプション・メニューが用意されており、「論理積」または「論理和」を指定することができます。たとえば、検索条件として、"sales > 20 And < 40" を指定することができます。特定の項目について任意の数の And と Or を指定できますが、同一項目内で And と Or を混在させることはできません。

階層のレベルによってデータ型が異なる場合（たとえば最上位レベルが文字列型であり、2 番目のレベルが整数型である場合）、「階層」検索フィールドは文字列として扱われ、数値演算ではなく文字列演算が適用されます。

「大文字 / 小文字を区別せずに検索」チェックボックスにチェックマークを付けると、文字列の比較時に大文字と小文字の区別が無視されます。

「検索」ダイアログ・ボックスの一番下には、次のボタンがあります。

- 「検索」ボタンをクリックすると、検索処理が開始されます。パネルがアクティブであるときに Enter キーを押すと、このボタンをクリックしたことになります（デフォルトのボタン）。検索が完了すると、条件に合致したオブジェクトが黄色いスポットライトで強調表示されます。
- 「消去」ボタンをクリックすると、検索結果を示す黄色いスポットライトがすべて解除され、検索フィールドに入力した値が消されます。
- 「次」ボタンをクリックすると、検索基準に一致する次のオブジェクトが（左から右の順に）選択されてズームされます。条件に一致する最後のオブジェクトが選択された後でこのボタンをクリックすると、ホーム画面に戻ります。このボタンが使用可能になるのは、検索条件に一致するオブジェクトが検出された場合だけです。
- 「前」ボタンをクリックすると、「次」ボタンとは逆（右から左）の順序でオブジェクトが選択されてズームされます。
- 「選択」ボタンをクリックすると、検索基準に一致するすべてのオプションが選択されます。選択されたオブジェクトは「選択」メニューで操作することができます。

- 「閉じる」ボタンをクリックすると、「検索」ダイアログ・ボックスが閉じて、検索のスポットライトがオフになります。「検索」ダイアログ・ボックスを次に開いたときは、最後に「閉じる」ボタンをクリックしたときの状態が復元されます。たとえば、「検索」ボタンをクリックすると、前回の検索が再実行されます。

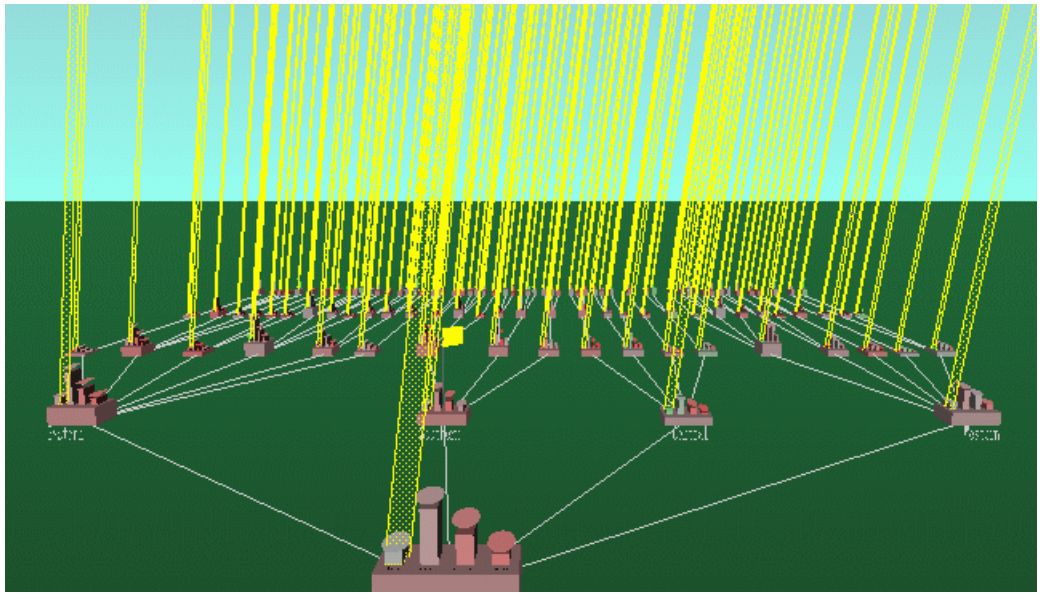

図 1-41 ツリー・ビジュアライザでの検索結果の例

検索が完了すると、検索条件に一致するオブジェクトが黄色いスポットライトで強調表示されます（図 1-41）。強調表示されたオブジェクトに関する情報を表示するには、マウスのポインタをスポットライト上に移動します。オブジェクトに関する情報は、左上隅の「ポインタが次を超えました」というラベルの下に表示されます。黄色いスポットライトの下のオブジェクトを選択してズームするには、マウスの左ボタンでスポットライトをクリックします。< Ctrl > キーを押しながらクリックする s と、ズームは行われません。

フィルタパネル

ツリー・ビジュアライザの「フィルタ」パネルでは、特定のフィルタリング条件を指定して、ツリー階層の表示を細かく調整することができます。このパネルは、特定の情報を強調したい場合や、表示する情報量を減らして操作性能を改善したい場合などに使用することができます。ほとんどの MineSet ツールで類似したフィルタパネルが

使用されています。一般的な説明は、「「フィルタ」パネル」（103 ページ）を参照してください。

ツリー・ビジュアライザの「フィルタ」ダイアログ・ボックスの各フィールドでは、「検索」ダイアログ・ボックスの場合と同様の方法でフィルタリング条件を指定します。たとえば、「大文字 / 小文字を区別せずに検索」チェックボックスにチェックマークを付けると、文字列の比較時に大文字と小文字の区別が無視されます。

ノード、バー、子ノードのいずれもフィルタリング条件に一致しない場合、ノードは表示されません。ただし、フィルタリング条件に一致するオブジェクトは存在しても、階層の上位のオブジェクトが条件に一致しないケースも考えられます。あるいは、同一ノードの他のバーがフィルタリング条件に一致しないケースもあります。階層ツリーでは位置が重要であるため、こうしたケースでフィルタリング条件に一致しないバーを消去するのは必ずしも適切でありません。このような状況に対処するため、オブジェクトの描画方法を指定する 3 つのラジオボタン（「ソリッド」、「アウトライン」、「非表示」）が用意されています。ただし、「表示」メニューの「ゼロ」オプションや「Null」オプションでオブジェクトを非実線で描画することが指定されている場合は、その指定のほうが有効になります。たとえば、NULL 値を非表示にするように指定されている場合は、フィルタリング条件に関係なく、NULL 値が常に非表示になります。

例外として、特定のバーだけを抽出するフィルタリングの場合が挙げられます。この場合は、ラジオボタンの設定に関係なく、条件に合致しないバーは非表示になります。

「高さ」フィルタスライダを使用すると、一定の高さに達しないバーだけを含むノードを除去することができます。高さのフィルタリング条件は、最大の高さに対するパーセントで指定します。最初に、一番高いバー（高さがレベルごとに正規化されている場合は、各レベルの一番高いバー）が計算されます。続いて、高さが（一番高いバーに対して）一定のパーセントに達するバーを少なくとも 1 つ含むノードだけが表示されます。

たとえば、このフィールドに "5%" を入力すると、一番高いバーの 5% 以上に相当する高さのバーを少なくとも 1 つ含むノードだけが表示されます（該当のノードの上位ノードも表示されます）。このオプションは重要性の低い小さいノードを除去する大まかな方法であり、特定の値を持つ特定のノードを正確に識別するメカニズムではありません。特定のノードを正確に検出したい場合は、「検索」パネルを使用してください。「高さ」フィルタを使用すると、複雑なデータのレンダリングが高速になります。また、ゼロ近くに数多くのバーが集中しなくなるため、画面がすっきりします。設定ファイルで「高さのフィルタ」コマンドを指定しても、このフィルタと同じ効果が得られます。

小さいノードは表示画面から除去されますが、階層の計算には組込まれます。

「高さ」スライダの下にある「深さ」スライダを使用すると、同時に表示される階層レベルの数を制限することができます。最上位レベルの階層では、このスライダで指定した数の階層レベルだけが表示されます。横方向の列に並んだノードは最も見やすいように整列されます。階層の下位レベルにあるノードに移動すると、横方向の列が自動的に追加表示され、その上位レベルにあるノードの位置は新たに追加表示されたノードが見えるように自動的に調整されます。このため、一部のノードは移動したように見えます。ただし、全体表示画面では、最上位ノードだけでなく、階層の全ノードが表示されることに注意してください。したがって、全体表示画面のレイアウトとメイン表示画面のレイアウトは一致しない可能性があります。全体表示画面内の "X" 記号は、メイン表示画面内の対応する位置を大まかに示しています（これら 2 つのレイアウトは完全には一致しません）。

フィルタリングを開始するには、「フィルタ」ボタンをクリックします。「フィルタ」パネルがアクティブであるときに *Enter* キーを押すと、このボタンをクリックしたことになります（デフォルトのボタン）。「フィルタ」パネルを閉じるには、「閉じる (*Close*)」ボタンをクリックします。

「選択」メニュー

ツリー・ビジュアライザの「選択」メニューを使用すると、ドリルスルーによって元のデータを細かく分析することができます。このメニューには、次の 5 個のオプションがあります。

- 「値の表示」— 選択したすべてのオブジェクトの値が表形式（レコードビューワ）で表示されます。
- 「オリジナルデータの表示」— 選択したデータに対応するレコードが抽出されて表示されます。レコードはテーブルビューワに表示されます。
- 「Tool Manager に送信」— Tool Manager による履歴の作成開始時に、現在のボックス選択に基づいてフィルタ操作が挿入されます。ドリルスルーに使用される実際の式は、現在のボックス選択の内容によって決まります。何も選択されていない場合は、警告メッセージが表示されます。
- 「補集合データのドリルスルー」— 「オリジナルデータの表示」と「Tool Manager に送信」を使用した際に、選択されていないデータがすべて取出されます。

- 「サブツリーの正規化」— サブツリー内にある要素の高さの最大値を調べ、その最大値を基準としてすべての値を正規化します。

ドリルスルーの詳細については、「ドリルスルー」（85 ページ）を参照してください。

ツリー・ビジュアライザの表示 メニュー

ツリー・ビジュアライザの「表示」メニューでは、複数の表示パラメータを制御することができます。

- 「ベースの高さ」チェックボックスは、ベースの高さのオンまたはオフを切替えます。負の値を表示したい場合や、バーの高さを比較しやすくする場合は、このオプションをオフにします。ベースの高さをオンにすると、すべてのバーの要約情報が表示されます。このオプションの初期値は、設定ファイルの "base height" ステートメントで設定することができます。
- 「マークフラグ」チェックボックスは、マークを表すフラグをオンまたはオフに切替えます（『*MineSet 3.0 Enterprise Edition User's Guide for Windows*』のマークパネル (Marks Panel) に関する説明を参照）。
- 「ゼロ」サブメニューでは、高さがゼロのオブジェクトの表示方法を指定します。デフォルトでは、高さがゼロのオブジェクトは他のオブジェクトと同じように（高さがゼロの立方体（平面）として）表示されます。このサブメニューには、高さがゼロのオブジェクトをアウトラインとして表示する（空白の正方形として表示する）オプションと、完全に隠す（描画しない）オプションがあります。「ゼロ」の初期値は、設定ファイルの zero オプションで設定することができます。
- 「Null」サブメニューでは、高さが NULL のオブジェクトの表示方法を指定します。このサブメニューのオプションは「ゼロ」サブメニューの場合と同じです。ただし、デフォルトでは、高さが NULL のオブジェクトが輪郭として表示されます。「Null」の初期値は、設定ファイルの null オプションで設定することができます。

ツリー・ビジュアライザの「移動」メニュー

「移動」メニューには、メインウィンドウの右上にあるボタンと同じ機能（オプション）があります。一部のオプションについてはキーボード・ショートカットが用意されています。

- 「ホーム」オプションを選択すると、指定したホーム位置（デフォルトでは最初の視点）に現在の視点が移動します。最初の視点（デフォルト）とは、ツリー・ビジュアライザを起動して設定ファイルを指定したときに表示される視点です。た

だし、Tool Manager のセッション中に「ホームの設定」を選択した後で「ホーム」を選択すると、最後に「ホームの設定」を選択したときの視点に戻ります。「ホーム」オプションのキーボード・ショートカットは < Ctrl+H > です。

- 「ホームの設定」オプションを選択すると、現在の視点がホーム位置として設定されます。その後で「ホーム」を選択すると、最後に「ホームの設定」を選択したときの視点に戻ります。
- 「全体の概観」オプションを選択すると、カメラの傾きを維持したまま、階層全体が表示されます。シーン全体を真上から表示するには、カメラを真下に向けてから、「全体の概観」を選択します。
- 「戻る」オプションを選択すると、直前の視点に戻ります。ツリー・ビジュアライザを起動した直後で、ホーム位置から移動していない場合、このボタンは使用できません。このオプションのキーボード・ショートカットは < Ctrl+B > です。
- 「進む」オプションを選択すると、「戻る」を選択した時点の視点に戻ります。「戻る」を選択していない場合、このオプションは選択できません。このオプションのキーボード・ショートカットは < Ctrl+R > です。
- 「親ノード」オプションは、オブジェクトが選択されている場合だけ使用することができます。バーが選択されているときにこのオプションを選択すると、そのバーが属しているベース（バーの親ノード）が選択されます。ベースを選択しているときにこのオプションを選択すると、階層の上位にある親ノードに移動します。ルートノード（階層の最上位レベル）まで到達すると、このオプションはグレー表示されて選択不可能になります。このオプションのキーボード・ショートカットは < Ctrl+U > です。
- 「左に移動」オプションを選択すると、左横にある兄弟オブジェクトが選択されます。バーを選択している場合は、左横のバーが選択されます。ベースを選択している場合は、同じ親を持つ左横のベースが選択されます。オブジェクトがまったく選択されていない場合、または左横に兄弟オブジェクトが存在しない場合、このオプションはグレー表示されて選択できません。
- 「右に移動」オプションを選択すると、右横にある兄弟オブジェクトが選択されます。バーを選択している場合は、右横のバーが選択されます。ベースを選択している場合は、同じ親を持つ右横のベースが選択されます。オブジェクトがまったく選択されていない場合、または左横に兄弟オブジェクトが存在しない場合、このオプションはグレー表示されて選択できません。

- 「最初の子ノード」オプションを選択すると、現在のノードの最初の子ノードが選択されます。オブジェクトがまったく選択されていない場合、バーが選択されている場合、または現在のノードに子ノードが存在しない場合、このオプションはグレー表示されて選択できません。
- 「最後の子ノード」オプションを選択すると、現在のノードの最後の子ノードが選択されます。オブジェクトがまったく選択されていない場合、バーが選択されている場合、または現在のノードに子ノードが存在しない場合、このオプションはグレー表示されて選択できません。

「ヘルプ」メニュー

「ヘルプ」メニューは、すべての MineSet ツールで同じものが使用されます。「IRIX システムの「ヘルプ (Help)」メニュー（Help （IRIX））」（105 ページ）を参照してください。

ツリー・ビジュアライザにおける NULL 値の取扱い

NULL 値は未知のデータ値を表します（NULL 値の詳細については、『*MineSet 3.0 Enterprise Edition Interface Guide*』の「Nulls in MineSet」を参照してください）。

ツリー・ビジュアライザでは、NULL 値が次のような場合に発生します。

- データベースまたはデータファイルに NULL 値が含まれている場合。
- 設定ファイル内で skipMissing オプションが定義されていないときに（『*MineSet 3.0 Enterprise Edition Interface Guide*』の「Creating Data and Cinfiguration Files for the Tree Visualizer」にある「skipMissing」を参照）、キー値のデータが階層内の 1 つのノードに存在するが他のノードには存在しない場合。たとえば、複数の州政府の予算を表示するときに、（他の州のレコードは存在するが）テキサス州の所得税のレコードが存在しない場合、テキサス州の所得税は NULL になります。これは値が 0 のレコードが存在する場合とは異なります。値が 0 のレコードが存在する場合、テキサス州の所得税は 0 になります。
- Tool Manager を使用して階級生成で配列を作成したときに、特定の階級に当てはまるデータが存在しない場合。たとえば、30-40 歳の人口に該当するデータが存在しない場合、その階級（30-40 歳）は NULL になります。
- Tool Manager を使用して配列を作成するときに、null enum オプションを指定した場合。この場合は、各バーチャートの最初のバーごとに、NULL 値を含む階級の値をすべて集計処理した配列が作成されます。NULL 値を含む階級（バー）には疑問

符 (?) が付きます。NULL の階級に属するデータが存在しない場合は、階級自体の値が NULL になります。

注記： NULL の階級に属するデータの値がすべて NULL の場合、ツリー・ビジュアライザではその NULL の階級が無視されて表示されません。

- 表現または集計値に NULL 値が含まれている場合。

ツリー・ビジュアライザでは、NULL 値をビジュアル属性（高さなど）に割当てたときに、特別な表現形式が使用されます。デフォルト設定では、オブジェクトの高さに NULL 値を割当てると、オブジェクトは輪郭で描画されます。ただし、「表示」メニュー（「ツリー・ビジュアライザの表示 メニュー」（211 ページ）を参照）または設定ファイル（『*MineSet 3.0 Enterprise Edition Interface Guide*』の「Creating Data and Cinfiguration Files for the Tree Visualizer」の「Null」を参照）を使用すると、デフォルトの表示形式を変更することができます。デフォルトの表示形式では、バーまたはベースは空白の正方形のように見えます（高さがゼロであるため、立方体には見えません）。ディスクは円のように見えます。オブジェクトの色に NULL 値を割当てると、そのオブジェクトは濃いグレーで表示されます（図 1-42 を参照）。

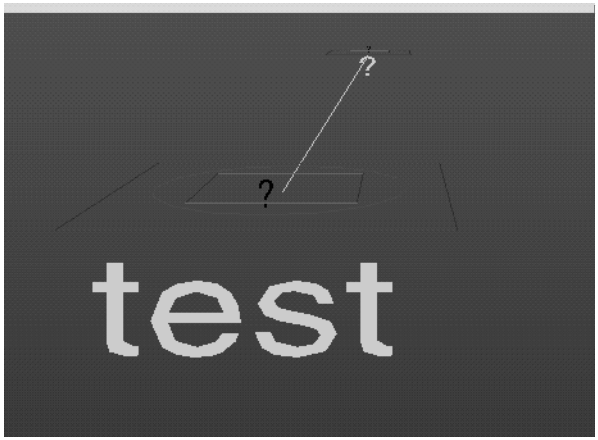

図 1-42 高さ、色、ディスク、ラベルに NULL 値を割当てた場合の表示例

NULL 値を持つオブジェクトを選択すると、選択フィールドに疑問符 (?) が表示されます。

ツリー・ビジュアライザに関する制限事項

ツリー・ビジュアライザの全機能のうち、Tool Manager では次の機能がサポートされません。

- 集計処理されていない階層（データが集計処理されないで直接表示される階層）
- リアルタイムのモニタリング
- 稀にしか使用されないオプション (skipMissing, overview, shrinkage, root label, speed, climb speed, leaf margin, root leaf margin, leaf edge margin, initial position, initial angle, bar label size, base label size, lod)
- 可変長の配列
- 階層を作成した後で計算される式。たとえば、パーセントを集計しても意味が無いため、パーセントは階層を集計処理した後で計算する必要があります。

設定ファイルとデータファイルのサンプルファイル

ツリー・ビジュアライザの特長や機能を紹介するために、設定ファイルとデータファイルのサンプルファイルが用意されています。ツリー・ビジュアライザを使用して決定木分析、選択式決定木分析、回帰ツリー分析をビジュアル表示する例については、付録 A「設定ファイルとデータファイルのサンプルファイル」を参照してください。

端数の切捨て

「端数の切捨て」オプションを指定すると、階級生成処理が実行される前に、あらかじめ極端な値（外れ値）がデータセットから排除されます。デフォルトの端数切捨ては 0.05 です。すなわち、すべての値のうち 5% の極端な値（下限の 2.5% と上限の 2.5%）がデータセットから排除されます。このオプションは、しきい値を決めるときに、外れ値の影響を低減する効果があります。

均一な範囲

「均一な範囲」はデータを自動的に階級生成するときに、値の範囲を均一なサイズの副範囲に分割するオプションです。Tool Manager の「項目の階級生成」ボタンを使用して階級自動生成を実行する方法については、「階級生成」（39 ページ）を参照してください。

均一な重み付け

「均一な重み付け」はデータを自動的に階級生成するときに、値の範囲を均一な重みを持つ副範囲（副範囲の数はユーザが指定した数になります）に分割するオプションです。各レコードの重みが 1 である場合は、個々の階級に含まれるレコード数が同じ（したがって、重みの合計値も同じ）になります。Tool Manager の「項目の階級生成」ボタンを使用して階級自動生成を実行する方法については、「階級生成」（39 ページ）を参照してください。

表示」メニュー

「表示」メニューでは、MineSet ビジュアライザのさまざまな表示オプションを制御することができます。これらの表示オプションは、ほとんどのビジュアライザで同じものが使用されます。プラットフォーム（Windows または IRIX）に応じて、使用できるオプションは若干異なります。

- 「フィルタ」パネルでは、特定のフィルタリング条件を指定して、メインの表示領域に表示するデータの数を減らすことができます。このパネルは、表示内容を調整したい場合、特定の情報だけを強調したい場合、または表示する情報量を単に減らしたい場合などに使用することができます。「フィルタ」パネルの右下にある「ランドスケープをフィルタに設定」チェックボックスでは、メインウィンドウ内のランドスケープにデータセット全体を表示するか、またはフィルタリングされたデータのみを表示するかを指定することができます。フィルタリング条件の指定方法の詳細については、「「フィルタ」パネル」（103 ページ）を参照してください。また、各フィールドにマウスポインタを移動して F1 キーを押すと、そのフィールドに関する簡単な説明が表示されます。
- 「背景色の設定」オプションを選択すると、カラーブラウザが表示され、新しい背景色を指定することができます。

- 「拡張コントローラーの表示」オプションでは、メインウィンドウの周囲に外部コントロールを表示するかどうかを指定します。
- 「NULL 位置の表示」オプションでは、軸の下の専用領域に NULL 値を表示するかどうかを指定します。
- 「アニメーション パネルの表示」オプションでは、アニメーション・コントロール・パネルを表示するかどうかを指定します。独立次元を持たないデータセットの場合、このオプションは使用できません。

可視化ツール

ここでは、MineSet の Tool Manager を通じて利用できる可視化ツールの概要を説明します。可視化ツールを使用すると、さまざまなビジュアル表現を通じてデータを表示・解析することができます。

- クラスタ・ビジュアライザは、クラスタリングによって生成されたクラスタ（グループ）に関する統計値を表示します。これらの統計値はデータセット全体の統計値と並べて表示されるため、各クラスタに固有の特徴が浮き彫りになります。
- デシジョン・テーブル・ビジュアライザは、複数の階層レベルにおける離散項目のデータ分布を表示します。たとえば、ビジネスの収益性を製品の種類、地理的条件、販促活動、営業員に対する報酬制度などの面から分析することができます。階層の各レベルにおいて、データが 2 つの属性に基づいて分割されてゆくため、2 つ一組となった属性を階層の下位レベルに向かってドリルダウンすることができます。
- エビデンス・ビジュアライザは、エビデンス・モデル (Evidence Inducer) によって生成された予測モデルをビジュアルに表示します。エビデンス・ビジュアライザのケーキグラフを見ると、特定のラベル値の出現確率を予測するときに特定の属性値が貢献する度合いを確認することができます（what-if 解析）。たとえば、アヤメ (iris) のデータセットでは、萼片の縦 (sepal length) が 5.45...5.85 である場合、ラベル値（アヤメの種類）がアイリスセトサ (iris-setosa) である確率は 86.54% であることが分かります。
- ヒストグラム・ビジュアライザはデータ内の連続型項目を自動的に階級生成して、その階級生成結果を統計量ビジュアライザに送信します。データはヒストグラムとして表示されます。
- マップ・ビジュアライザは、地理的に有意な区域間に存在するデータの関係をビジュアルに表示します。たとえば、複数の地域を表示して、販促キャンペーンがそれぞれの地域に与える影響を示すことができます。マップ・ビジュアライザの

ドリルダウン機能を使用すると、特定の地域に焦点を合わせ、より小さな地理的要素の中で詳細な解析を行うことができます。

- レコードビューワを使用すると、スプレッドシートのような行列形式でデータを表示することができます。
- スキャタ・ビジュアライザでは、1 次元空間、2 次元空間、または 3 次元空間でデータポイント（属性）が表示されます。また、別の 3 つの属性をグラフィカル・オブジェクト（要素）のサイズ、色、位置にマッピングすることができます。さらに、2 つの属性をスライダにマッピングしてアニメーション表示できるため（スライダを動かすと、個々の属性値が変化してアニメーションが動きます。）、データの動作を最大 8 次元で分析することができます。スキャタ・ビジュアライザは、相関規則の表示にも使用されます。
- スプラット・ビジュアライザはスキャタ・ビジュアライザと同じ機能を数多く備えていますが、データの密度が不透明度（霧）によって表されます。大量のデータを解析する必要がある場合、個々のデータポイントをプロットするのは非効率的であるため、スプラット・ビジュアライザを使用するほうが適切です。
- 統計量ビジュアライザは、現行データセットの要約情報（最大値、最小値、中央値、標準偏差、個別値、四分位数）を計算して表示します。
- ツリー・ビジュアライザは、階層関係にあるデータの解析に最適なツールです。ツリー・ビジュアライザには対話的な「フライスルー」（ナビゲーション）機能があり、階層の各レベルにあるデータ間の関係を効率的に分析することができます。たとえば、製品の製造ラインを分析して、売上総額に占める各製品の割合をグラフィカルに表示することができます。表示される情報は階層のレベルが深くなるにつれて詳細になり、最終的には個々の製品に関する情報が表示されます。

  ツリー・ビジュアライザは、決定木クラシファイア、選択式決定木クラシファイア、回帰ツリー分析モデルによって作成されたモデルを表示するときにも使用されます。分析結果はツリー内の独立したノードとして示されます。

警告オプション

コマンドラインからビジュアライザを起動するときは、次の 2 つのオプションを指定することができます。

- **`-warnexecute`** オプションを指定すると、Excute（実行）ステートメントで指定したコマンドを実行しようとしたときに警告メッセージが表示され、コマンドを実行するかどうかを選択することができます。これは、セキュリティに問題があるケース（Web から入手したインターネット・ファイルを実行するときなど）に対処するためのオプションです。mtr ファイルを使用してコマンドを実行すると、このオプションが自動的に適用されます。（mtr ファイルの詳細については、『*MineSet 3.0 Enterprise Edition Interface Guide*』を参照してください。）

  このオプションを常に有効にする手順は次の通りです。

  – IRIX システムの場合：

    各ユーザのホーム・ディレクトリ内の *.Xdefaults* ファイルに次の行を追加します。

    `*minesetWarnExecute:TRUE`

  – Windows システムの場合：

    Windows レジストリ内にある次のキーの値を変更します。

    `MINESET_WARN_EXECUTE`

    「ファイル」メニューの「設定」ダイアログを使用して、同等の設定を行うことができます。

- **`-quiet`** オプションを指定すると、進行状況を表すダイアログが表示されなくなります。

  このオプションを常に有効にする手順は次の通りです。

  各ユーザのホーム・ディレクトリ内の *.Xdefaults* ファイルに次の行を追加します。

  `*minesetQuiet:TRUE`

  to your *.Xdefaults* file.

Windows システムの場合、「ファイル」メニューの「設定」ダイアログにあるオプションを使用します。

Web 公開

MineSet の Web 拡張機能を使用すると、MineSet ソフトウェアで作成した可視化ファイルやデータを Web 上で表示（公開）することができます。MineSet の *mtr* 拡張機能では、MineSet の設定ファイル、.schema ファイル、.data ファイルがアーカイブ・ファイルに保存されます。このアーカイブ・ファイルは html タグとして Web ページに埋込むことができます。Internet Explorer などの Web ブラウザを使用してその Web ページを読込むと、ブラウザのウィンドウ内部で MineSet 可視化ツールが自動的に起動されます。Web ブラウザを実行するマシンには、MineSet クライアント・ソフトウェアをインストールしておく必要があります。

Web 拡張機能のインストール手順や使用例については、『*MineSet 3.0 Enterprise Edition Interface Guide*』を参照してください。

重み付け

通常、データセットの各レコードの重みは 1 であるため、レコードの総数と重みの合計値は一致します。ただし、データセットの各レコードが均等に標本抽出されていない場合は、レコードに任意の重みを割当てて、各レコードの寄与率を調整することができます。レコードの重み付けの概念では、重みが 2 である 1 つのレコードは、重みが 1 である 2 つのレコードと同じです。重みの値として浮動小数点数を指定することもできます。レコードの重みを指定するには、Tool Manager の「データの可視化 / マイニング」パネルから「マイニングツール」を選択します。分析を使用している場合は、「詳細オプション」（または「詳細モード」）ボタンをクリックし、表示されるダイアログ・ボックス内の「重み付けとして使用」チェックボックスを選択します。「レコードの重み付け」（151 ページ）も参照してください。

西暦 2000 年問題への対応

MineSet では、西暦 2000 年問題に対応した（Y2K 準拠の）日付形式がサポートされています。たとえば、U.S.（米国）ロケールでは、MM/DD/YY または MM/DD/YYYY の形式で日付を入力することができます。MineSet では 2 桁の年表記に関して、X/Open 標準に準拠しています。すなわち、69 以上の 2 桁数値は 1969 〜 1999 年とみなされ、68 以下の 2 桁数値は 2000 〜 2068 年とみなされます。

European（ヨーロッパ）ロケールでは、DD/MM/YY または DD/MM/YYYY の形式で日付を入力することができます。2 桁の年表記の取扱いは上記と同じです。

どちらのロケールでも、2 桁表記の年を入力すると、その値が自動的に 4 桁表記の年に拡張されて画面に表示されます。

設定ファイルとデータファイルのサンプルファイル

MineSet では、さまざまなツールの機能や特長を紹介するために、設定ファイルとデータファイルのサンプルファイルが用意されています。この付録では、各ツール用のサンプルファイルの内容と使用方法を詳しく説明します。以下の各項目はツール名のアルファベット順に並んでいます。

- Windows システム上では、MineSet をインストールした *\examples* 配下のディレクトリにサンプルファイルがあります。
- IRIX システム上では、*/usr/lib/MineSet/examples* ディレクトリにサンプルファイルがあります。

サンプルファイルの説明：

- 「相関規則ビジュアライザ用のサンプルファイル」（225 ページ）
- 「クラスタリング用のサンプルファイル」（226 ページ）
- 「重要項目用のサンプルファイル」（226 ページ）
- 「決定木用のサンプルファイル」（228 ページ）
- 「デシジョン・テーブル用のサンプルファイル」（238 ページ）
- 「エビデンス・ビジュアライザ用のサンプルファイル」（255 ページ）
- 「マップ・ビジュアライザ用のサンプルファイル」（266 ページ）
- 「選択式決定木用のサンプルファイル」（269 ページ）
- 「回帰ツリー用のサンプルファイル」（272 ページ）
- 「スキャタ・ビジュアライザ用のサンプルファイル」（277 ページ）
- 「スプラット・ビジュアライザ用のサンプルファイル」（280 ページ）
- 「ツリー・ビジュアライザ用のサンプルファイル」（282 ページ）

相関規則ビジュアライザ用のサンプルファイル

既製のデータセットを使用して相関規則を可視化できるように、設定ファイルとデータファイル（規則ファイル）のサンプルファイルが用意されています。これらのサンプルファイルの中には、階層型データセットを表すものもあります。規則ファイルのサンプルには、相関規則分析によって導出された規則が格納されています。一般的に、規則を格納したファイルには *.rules.data* という拡張子を付けます。個々の設定ファイルでは、対応する規則ファイルの表示形式が指定されています。設定ファイルの拡張子は、*.scatterviz* でなければなりません。ここで説明するサンプルファイルは、MineSet をインストールした *\examples* 配下のディレクトリ（Windows システム）または */usr/lib/MineSet/scatterviz/examples* 配下のディレクトリ（IRIX システム）にあります。

- *group.rules.data* と *group.rules.scatterviz*

  これらのファイルは、各種の製品グループ（パンとベーカリー、乳製品、炭酸飲料など）について生成された規則ファイルと設定ファイルです。

- *category.rules.data* と *category.rules.scatterviz*

  これらのファイルは、製品グループの中の製品カテゴリ（冷蔵ミルクと非冷蔵ミルクなど）について生成された規則ファイルと設定ファイルです。

- *adult94.rules.data* と *adult94.rules.scatterviz*

  これらのファイルは、国勢調査データベースについて生成された規則ファイルと設定ファイルであり、婚姻状態、教育レベル、年齢、収入などの各種属性間の相関が示されています。

- *germanCredit.rules.data* と *germanCredit.rules.scatterviz*

  これらのファイルは、ドイツの信用データベースについて生成された規則ファイルと設定ファイルであり、信用履歴、雇用、貯蓄などの各種属性間の相関が示されています。

- *cars.rules.data* と *cars.rules.scatterviz*

  これらのファイルは、自家用車のデータセットについて生成された規則ファイルと設定ファイルであり、さまざまな属性間の相関が示されています。

クラスタリング用のサンプルファイル

ここでは、クラスタリングの適用に有効と思われる事例を紹介します。MineSet では、この事例で使用するサンプルファイル（cars データセット）が用意されています。この事例はクラスタリング・ツールの基本的な操作方法やオプションを紹介するものです。

「車 (cars)」データセットは比較的単純な構造であり、重さ、馬力、速度、mpg （ガロン当たりの走行マイル数）などの基本的な属性が格納されています。

クラスタ・ビジュアライザを最初に起動したときは、各クラスタの区別化に対する貢献度に応じて、属性が上から下に順番に並べられます。

クラスタ・ビジュアライザのメイン・ウィンドウ内で特定のクラスタの名前をクリック（選択）すると、選択されたクラスタを基準として他のクラスタ内の属性（棒グラフとヒストグラムで表される属性）が並べ替えられます。たとえば、クラスタ 1 をクリックすると、シリンダー (cylinder)、重さ (weight)、mpg の順に属性が並べ替えられます。この並べ替えは表示上の意味しか持ちません（元のデータセットは変更されません）。他のクラスタの同じ行を比較すると、特定の属性の寄与率がクラスタ間でどのように違うかが分かります。クラスタ 2 をクリックすると、より低いレベルにおける各属性の寄与率の順序が分かります。この場合は、製造元 (origin) が最も重要視され、次にシリンダー (cylinder)、馬力 (horsepower)、mpg の順になっています。

重要項目用のサンプルファイル

ここでは、重要項目の適用に有効と思われる事例を紹介します。MineSet では、この事例で使用するサンプルファイルが用意されています。この事例は、重要項目の使用方法やさまざまなオプションを紹介するものです。

顧客が今まで使用していた電話会社から別の会社に切換えることを「解約 (Churn)」と呼びます。これは電話会社でよくある問題です。この事例では、*churn.schema* ファイルと *churn.data* ファイルが使用されています。これらのファイルは、MineSet をインストールした *\data* 配下のディレクトリ（Windows システム）または */usr/lib/MineSet/data* 配下のディレクトリ（IRIX システム）にあります。

重要項目を「標準 (Simple)」モードで実行すると、次の 3 つの属性が検出されます。

- 昼間通話時間 (Total Day Minutes)
- 顧客サービスコール数 (Number of Customer Service Calls)
- 州 (State)

詳細 モードで「... 左側項目リストに寄与率向上度、右側項目リストに累積寄与率を表示」オプションを実行すると、「昼間料金 (Total Day Charge)」と「昼間通話時間 (Total Day Minutes)」の寄与率ランキングが同じ (48.67) であることが判明します。これら 2 つの項目のいずれか（たとえば「昼間通話時間 (Total Day Minutes)」）を右側に移動し、「... 左側項目リストに寄与率向上度、右側項目リストに累積寄与率を表示」オプションをもう一度実行すると、他方の項目（「昼間料金 (Total Day Charge)」）の値が消失します。すなわち、これら 2 つの項目の間には非常に強い相関関係があります。

右側に移された「昼間通話時間 (Total Day Minutes)」を見れば、次の項目が重要であることが分かります。

- 国際化プラン (International plan) (4.1)
- 顧客サービスコール数 (Number of Customer Service Calls) (8.1)
- 州 (State) (4.7)

ここでは、入手が簡単で比較しやすい「国際化プラン (International plan)」を選択して右側に移動します。

残りの 2 つの属性（「顧客サービスコール数 (Number of Customer Service Calls)」と「州 (State)」）も依然としてかなり重要であるため（実際、これらの属性の寄与率は増加します）、「国際化プラン (International plan)」との相関性はあまり強くないと判断されます。

属性の寄与率をこのような方法で調べると、他の属性（ほとんど同じ寄与率を持つが、より分かりやすく比較しやすい属性）で置換できる属性を見つけることができます。寄与率を調べると、属性の追加による効果を分析することができます。たとえば、上記の例の場合、「州 (State)」を追加すると寄与率がかなり増加します。アヤメ (*Iris*) のデータセットでは、3 番目の属性（「萼片の縦 (*sepal length*)」）を追加しても、寄与率はほんの少ししか増加しません。場合によっては、シンプルな 2 次元のスキャタ図の方が 3 次元のスキャタ図より分かりやすいことがあります。

決定木用のサンプルファイル

ここでは、決定木分析の適用に有効と思われる事例を紹介します。MineSet では、これらの事例で使用するサンプルファイルが用意されています。分析を実行すると、下記で説明する *-dt.treeviz* ファイルが生成されます。

data ディレクトリから該当するスキーマファイル (*.schema*)（*churn.schema* など）を開くと、データファイルが MineSet に読込まれます。分析モデルの可視化ファイル（拡張子 *-dt.treeviz*）は *examples* ディレクトリから開くことができます。

- Windows システムの場合、MineSet をインストールした *data* ディレクトリと *examples* ディレクトリにあります。
- IRIX システムの場合、*/usr/lib/MineSet/treeviz/examples* の *data* ディレクトリと *examples* ディレクトリにあります。

解約 (Churn)

顧客がある電話会社から別の電話会社に移ってしまうことを解約または顧客離れと言います。解約はどの電話会社にとっても共通の重要な問題です。*examples* ディレクトリ内のファイル (*churn-dt.treeviz*）には、この問題について帰納された決定木クラシファイアが収録されています。この可視化ファイルは、ラベルを解約（はい、いいえ）(churn (yes, no)) に設定し、churn.schema ファイルに対して分析を適用して作成されたものです。この例で使用されているデータセットは架空のデータですが、実際の電話会社のデータで見られるパターンを反映しています。

この決定木では、昼間通話時間（total_day_minutes）に基づいてルートが分割されていることに注意してください。1 日の通話時間が 264 分以上の顧客の解約率 (60%) は、それ以外（通話時間が 264 分未満）の顧客の解約率 (11%) よりかなり高くなっていることが分かります。1 日の通話時間が 264 分以上の顧客は、会社に最も利益をもたらす顧客層と考えられます。

ルートの左側のサブツリーは、1 日の通話時間が 264 分未満の顧客を表します。これらの顧客の平均解約率は 11% ですが、顧客サービス通話を 3 回以上利用した顧客の解約率は 49% に跳ね上がっています。

ルートの右側のサブツリーは、1 日の通話時間が 264 分以上の顧客を表します。これらの顧客の平均解約率は 59% ですが、ボイスメールを利用している顧客の解約率は 9.3% と低く、ボイスメールを利用していない顧客の解約率は約 75% と高くなっています。

車の原産国

cars データセットには、1970 年代から 1980 年代初頭にかけての様々な車種に関する情報が格納されています。これらの情報には、車の重量、加速度、mpg（ガロン当たりの走行マイル数）などがあります。*examples* ディレクトリ内のファイル (*cars-dt.treeviz*) には、車の原産国（ヨーロッパ、日本、またはアメリカ）を予測するために帰納された決定木クラシファイアが収録されています。この可視化ファイルは、ラベルを origin (Japan, U.S., Europe) に設定し、*data* ディレクトリ内のファイル (*cars.schema*) を分析して作成されたものです。この分析モデルの目的は、車の原産国を特徴付ける属性を判断することです。

- Windows システムの場合、これらのファイルは MineSet がインストールされているディレクトリ *\examples\cars-dt.treeviz* および *\data\cars.schema* にあります。
- IRIX システムの場合、これらのファイルは */usr/lib/MineSet/treeviz/examples/cars-dt.treeviz* および */usr/lib/MineSet/data/cars.schema* にあります。

この決定木の左側ではブランドに基づく分割が行われていますが、ルートでは分割基準としてブランドが採用されていません。その理由は、決定木分析では多方向の分割が不利であり、1 立方インチ (cubic_inches) 属性に基づく分割の方が適切な分析モデルに適していると判断されたためです。Tool Manager の「項目の削除 」オプションを使用してブランドを非表示にすると、さらに興味深い分析を行うことができます。

この決定木クラシファイアを見ると、米国製の車を判別するのに 1 立方インチ (cubic_inches) が最適な属性であることが分かります。大きいエンジン（1 立方インチ (cubic_inches) > 169.5）の車はほとんど米国製ですが、小さいエンジンの車は他国でも製造されています。「選択」メニューから「オリジナルデータの表示」を選択すると、大きいエンジンの車で米国製でないのはメルセデス (Mercedes) だけであることが分かります。決定木のルートノード（すなわち、訓練事例全体）では米国製の車が多くの割合 (62.50%) を占めていますが、1 立方インチ (cubic_inches) 属性に基づいて 1 回の分割を行うと、小さいエンジンの車については製造場所を予測するのが困難になることに注意してください。ルートノードの寄与率は 16.2 であり、1 つの支配的なクラス（この場合は米国）が存在することが分かります。右側のノード 1 立方インチ (cubic_inches) > 169.5 の寄与率は 96.81 であり、エンジンの大きい車はほとんど米国製であること（サブツリーの寄与率が非常に高いこと）が分かります。実際、右側のサブツリーの推定誤差率は 0%（緑色のベース）ですが、左側のサブツリーの寄与率は 0.23 であり、推定誤差率は 31.25%（オレンジのベース）とかなり高くなっています。左側のサブツリーは各クラスのレコード数がほぼ同じであるため、データセット全体よりもラベルの予測が困難になっています。

性別の判断

adult データセットには成人の就労者に関する情報が格納されています。このデータセットは米国人口統計局 (U.S. Census Bureau) のデータベースから抽出したものであり、年間の総収入が $100 以上、1 週間の就業時間が 1 時間以上、年齢が 16 歳以上という条件を満たす就業者に関するデータが収録されています。*adult-sex-dt.treeviz* ファイルには、これらの就労者の性別を判断するために帰納された決定木クラシファイアが収録されています。この可視化ファイルは、ラベルを「性別 (*sex*)」に設定し、*adult.schema* ファイルを分析して生成されたものです。adult データセットには約 50,000 件のレコードが含まれているため、ワークステーション上で決定木分析を実行するのに数分の時間がかかる場合があります。

- Windows システムの場合、これらのファイルは MineSet がインストールされているディレクトリ *\examples\adult-sex-dt.treeviz* および *\data\adult.schema* にあります。
- IRIX システムの場合、これらのファイルは */usr/lib/MineSet/treeviz/examples/adult-sex-dt.treeviz* および */usr/lib/MineSet/data/adult.schema* にあります。

可視化の結果を見ると、次のようなことが分かります。

- 一部のレコードについては婚姻関係の属性から性別を容易に判断することができます。たとえば、通常、夫は男性です（興味深いことに、女性である夫のレコードが 1 件存在します。これは人口統計局のデータ品質の問題であり、同性同士の結婚を意味するものではありません）。同様に、通常、妻は女性ですが、妻を男性と誤って記録したレコードが 3 件存在します。

より興味深い分析を行うために、婚姻関係を表す属性を除去して新しい決定木を作成すると、次のようなことが分かります。

- 最も重要な属性は未婚 / 既婚の区別です。
- ベースの高さから判断して、ほとんどの人が（軍関係ではなく）民間人と結婚しているか、一度も結婚したことがないか、離婚したかのいずれかの状態です。軍関係の配偶者と結婚している人、別居している人、配偶者と死別して再婚していない人なども少数存在します。
- このデータセットではルートにおける男性の割合が高くなっています（このデータセットは成人の就労者に関する情報を表すものであり、人口全体の傾向を反映するものではありません）。

- 一番左側のノードには、離婚した成人就労者のレコードが含まれています。このノードにおける男女の割合（男性 60%、女性 40%）は、ルートにおける男女の割合よりも均等です。2 番目のノードには、結婚している成人就労者のレコードが含まれています。このノードでは 89% が男性であることが分かります。3 番目のノードには、一度も結婚したことがない成人就労者のレコードが含まれています。このノードにおける男女の割合は離婚した成人就労者のグループとほぼ同じですが、男性の割合が多少高くなっています。一番右側のノードには、配偶者と死別して再婚していない成人就労者のレコードが含まれています。このノードでは女性の割合が 81% と高くなっています（おそらく、女性の平均寿命が男性より長いことが原因です）。"widowed"（やもめ）という用語は、配偶者を失った人を意味します。

たとえば、女性就業者を対象として新製品を売込む場合は、検索パネルを使用して、女性の割合が高いセグメントを見つけることができます。その場合の検索条件は次のようになります。

- 性別が女性である。（ウィンドウの最上部の「女性 (female)」をクリックします。）
- サブツリーの重みが 1000 より大きい (subtree weight > 1000)。
- 占有率が 80% より大きい (percent > 80)。

検索条件に一致したノードは 3 つの黄色いスポットライトで強調表示されます。1 つのパス上に 2 つのノードが存在するため、ルートに最も近い（右側の）ノードを見てください。このパスは次の規則を表しています。

```
marital status = Widowed implies that 81.23% are female
```
（配偶者と死別して再婚していない人の 81.23% は女性である。）

```
marital status = Divorced and occupation =
     administrative clerical implies that 87.67% are female
```
（離婚した管理職の 87.67% は女性である。）

この訓練事例では、ルートの 16,192 人のうち、配偶者と死別した 1,233 人と離婚した管理職である 1,045 人の女性が上記の規則に該当します。このセグメントは女性就労者の 14% 以上を占めています。

給料を決める要因

次に、成人就労者のデータセットを分析して給料水準を決める要因を確定し、年間収入が $50,000 以上のクラスと $50,000 未満のクラスにデータセットを分割します。その結果に基づき、"50,000+" または "-50,000" という 2 つの値を取る属性を個々のレコードに追加します。MineSet のクラシファイアを実行すると、給料水準に影響を与える要因を簡単に見つけることができます。*adult-salary-dt.treeviz* ファイルには、この問題について作成された決定木クラシファイアが収録されています。この可視化ファイルは、ラベルを *gross_income_bin*（区間を 50000 と指定して *gross_income* を階級生成した属性）に設定し、*adult.schema* ファイルを分析して作成されたものです。

- Windows システムの場合、これらのファイルは MineSet がインストールされているディレクトリ *\examples\adult-salary-dt.treeviz* および *\data\adult.schema* にあります。
- IRIX システムの場合、これらのファイルは */usr/lib/MineSet/treeviz/examples/adult-salary-dt.treeviz* および */usr/lib/MineSet/data/adult.schema* にあります。

可視化の結果を見ると、次のようなことが分かります。

- 訓練事例全体を表すルートを見ると、成人就業者の 76.07% が $50,000 以上の給料を稼いでいることが分かります。
- 年齢が最も重要な要因です。27 歳未満の就業者のうち、$50,000 以上の給料を稼いでいるのはわずか 3.07% にしかすぎません。ベースの色は非常に正確な規則（誤差率は約 3%）を示す緑色になっていることに注意してください。
- 27 歳以上の就業者の給料を予測するときは、教育レベルが重要な要因になります。人口統計局では個々の就業者に特定の教育レベルを割当てています。決定木クラシファイアは教育レベルが 12.5 の箇所で分割されています。レベル 13 は学士に相当します。教育レベルが学士以上である就労者の約 55% は、$50,000 以上の給料を稼いでいます。
- 年齢が 27 歳以上で教育レベルが高いセグメントについては、給料を予測するにあたって、婚姻関係が重要な要因になります。このセグメントに属する既婚の就労者が $50,000 以上の給料を稼ぐ確率は、男性（夫）の場合は 73%、女性（妻）の場合は 75% に増加します（ただし、既婚の女性を表すノードのベースが小さく、この規則に合致する女性の数は少ないことが分かります）。一方、同じセグメントに属する未婚の就労者が $50,000 以上の給料を稼ぐ確率は、男性（夫）の場合は 27%、女性（妻）の場合は 25% に減少します。

アヤメ (iris) のクラス判別

iris データセットの各レコードには、アヤメの特性を表す 4 つの属性（萼片の縦 (sepal_length)、萼片の横 (sepal_width)、花びらの縦 (petal_length)、花びらの横 (petal_width)）があります。これらの属性に基づいて、アヤメは 3 つのクラス（アイリスセトサ (iris-setosa)、アイリスバージカラー (iris-versicolor)、アイリスバージニカ (iris-virginica)）に分類されます。この例の目的は、アヤメの種類を特徴付ける属性を確定することです。

クラシファイアを作成する前に、Tool Manager の「クラス判別」タブにある「寄与率」タブをクリックした後、「実行」ボタンをクリックしてください。各属性 (petal_width、petal_length、sepal_width、sepal_length) の寄与率がランク付けされたリストが表示されます。これらの属性をスキャタ・ビジュアライザの軸にマッピングし、iris_type を色にマッピングすると、同じ特性を持つクラスタを確認することができます。

iris-dt.treeviz ファイルには、この問題について帰納された決定木クラシファイアが収録されています。この可視化ファイルは、*iris.schema* ファイルを分析して作成されたものです。

- Windows システムの場合、これらのファイルは MineSet がインストールされているディレクトリ *\examples\iris-dt.treeviz* および *\data\iris.schema* にあります。
- IRIX システムの場合、これらのファイルは */usr/lib/MineSet/treeviz/examples/iris-dt.treeviz* および */usr/lib/MineSet/data/iris.schema* にあります。

ツリー・ビジュアライザを実行すると、ルートの寄与率が非常に低い (0) にも関わらず、誤差率が 6% もあることが分かります。寄与率はラベル値の分布の歪みを表す値ですが、ルートにおけるラベル値の分布は完全に一様です（個々のラベル値について 50 個のレコードが存在します）。左側の枝 (petal-length <= 2.6 inches) は、iris-setosas だけを含む緑色のノード（誤差率がゼロ）に属します。その他の枝でも、petal_width に関するテストに基づいて各レコードを特定のクラスにすばやく分類することができます。"petal-length > 2.6 and petal-width 1.65 and petal-length > 5" という条件のパスは、4 つのレコード（3 個の iris-virginica と 1 個の iris-versicolor）を含む低寄与率のリーフノードに到達します。このリーフノードが分割されなかった理由は、それ以上の分割は無意味と判断されたためです（デフォルトでは、個々の分割において、重みが 2 以上である 2 つの子ノードが生成されなければなりません）。このリーフノードの色は黒であるため（ノードに到達したテストセット (test set)・レコードが存在しなかったため）、推定誤差率は計算されません。

まとめ："petal_length <= 2.6 inches" であるアヤメは iris-setosa に分類され、"petal_length > 2.6 inches and <=5 inches and petal_width 1.65 inches" であるアヤメは iris-versicolor に分

類され、"petal_length >2.6 inches and petal_width > 1.65 or petal_length > 5 inches and petal_width <= 1.65" であるアヤメは iris-virginica に分類されます。

決定木では連続的な属性が 2 分割されますが、「重要項目」機能では連続的な属性が離散的な範囲に分割されるため、決定木のルートで分割される属性は、重要項目リストの最初の属性とは異なります（詳細については「重要項目」（58 ページ）を参照）。

キノコ (Mushroom) の分析モデル

mushroom-dt.treeviz ファイルには、キノコを分類するために生成された決定木クラシファイアが収録されています。この可視化ファイルは、*mushroom.schema* ファイルを分析して作成されたものです。

- Windows システムの場合、これらのファイルは MineSet がインストールされているディレクトリ *\examples\mushroom-dt.treeviz* および *\data\mushroom.schema* にあります。
- IRIX システムの場合、これらのファイルは */usr/lib/MineSet/treeviz/examples/mushroom-dt.treeviz* および */usr/lib/MineSet/data/mushroom.schema* にあります。

この例の目的は、食用になるキノコと毒性のあるキノコを判別することです。このデータセットには 8,000 件以上のレコードが含まれているため、分析を実行するのに数秒の時間がかかる場合があります。

個々のキノコには、カサの色、傷、臭いなどの様々な属性があります。決定木クラシファイアを構築した場合は、臭いの属性だけを使用することによって、食用になるキノコと毒性のあるキノコを 50% の確率で判別できることが分かります。臭いのないキノコに毒性がある確率は 3.4% です。次に注目する属性は茎の形状です。茎が先細になっているキノコは食用になりますが、茎が先太になっているキノコは 11.6% の確率で毒性があります（1,032 個のキノコがこのノードに到達します）。ツリー内のノードを奥の方にたどっていくと、判断基準になる他の属性を見つけることができます。

政党への帰属

このデータセット (vote) は選挙のレコードから成ります。この例の目的は、主要な選挙に関するデータに基づいて下院議員が属する政党を判断することです。vote データセットには、議会季刊誌 (*CQA*: *Congressional Quarterly Almanac*) に記載された 16 回の主要な選挙において米国下院の各議員が投票した内容が収録されています。CQA では投票の内容として、賛成 3 種類、反対 3 種類、棄権 3 種類の計 9 種類を規定しています。

決定木クラシファイアを作成する前に、16 回の選挙内容を分析して、重要と思われる特性を考えてみてください。

vote-dt.treeviz ファイルには、この問題について作成された決定木クラシファイアが収録されています。この可視化ファイルは、*vote.schema* ファイルを分析して作成されたものです。

- Windows システムの場合、これらのファイルは MineSet がインストールされているディレクトリ *\examples\vote-dt.treeviz* および *\data\vote.schema* にあります。
- IRIX システムの場合、これらのファイルは */usr/lib/MineSet/treeviz/examples/vote-dt.treeviz* および */usr/lib/MineSet/data/vote.schema* にあります。

乳癌の診断

breast データセットには、乳癌の診断を受けた女性被験者に関する情報が収録されています。このデータセットの各レコードは、細胞の大きさ、細菌塊の厚さ、周辺吸着などの属性を持つ被験者を表します。この例の目的は、被験者の腫瘍が悪性であるか良性であるかを判断することです。*breast-dt.treeviz* ファイルには、この問題について作成された決定木クラシファイアが収録されています。この可視化ファイルは、*breast.schema* ファイルを分析して作成されたものです。

- Windows システムの場合、これらのファイルは MineSet がインストールされているディレクトリ *\examples\breast-dt.treeviz* および *\data\breast.schema* にあります。
- IRIX システムの場合、これらのファイルは */usr/lib/MineSet/treeviz/examples/breast-dt.treeviz* および */usr/lib/MineSet/data/breast.schema* にあります。

決定木クラシファイア を見ると、uniformity_of_cell_size が悪性 / 良性の判定にとって極めて重要な属性であることが分かります。ルートにおけるラベルの分布は約 65% と 35% （寄与率は 7.07）ですが、ルートの 2 つの子ノードは歪みが大きく、左側のノードの誤差率は僅か 1.29% です。ルート単独で十分に正確な判断ができます。ツリーの高さを単一レベルに制限すると、誤差率は 7.3% になります。

甲状腺機能低下症 (Hypothyroid) の診断

hypothyroid（甲状腺機能低下症）データセットの構造は、上記の乳癌データセットの構造と似ています。*hypothyroid-dt.treeviz* ファイルには、甲状腺機能低下症を診断するために作成された決定木クラシファイアが収録されています。この可視化ファイルは、*hypothyroid.schema* ファイルを分析して作成されたものです。

- Windows システムの場合、これらのファイルは MineSet がインストールされているディレクトリ *\examples\hypothyroid-dt.treeviz* および *\data\hypothyroid.schema* にあります。
- IRIX システムの場合、これらのファイルは */usr/lib/MineSet/treeviz/examples/hypothyroid-dt.treeviz* および */usr/lib/MineSet/data/hypothyroid.schema* にあります。

hypothyroid データセットには 3,163 件のレコードが収録されていますが、ほとんど (95.23%) の被験者は陰性です（甲状腺機能低下症にかかっていません）。したがって、特定の被験者を陰性と診断することは高い確率で正解になりますが、最も重要なのは実際に甲状腺機能低下症に罹っている（陽性の）被験者です。陽性の被験者を陰性と診断するのは重大な問題です。「詳細設定」オプションを使用して混同マトリックスを表示すると、陽性の被験者が誤って陰性と診断されたケースが 5 件あることが分かります。

決定木クラシファイアを見ると、ルートノードが緑色（非常に低い誤差率）であることが分かります。ルートにある 1 つの属性 "fti" によって、陰性の被験者を比較的簡単に見分けることができます。すなわち、fti 値が高い被験者の 99.7% は陰性であり、fti 値が不明である被験者（他の測定値から判断して明白であるなどの理由から、医師が fti 値を測定しなかったと考えられます）も陰性です。ただし、残りの 218 人の被験者については細かい分析が必要です（ノードのベースがオレンジ色になっています）。データセット全体には 3,163 件のレコードが含まれていますが、ほとんどのレコードは容易に分類できるため、本当にデータマイニングの必要性があるのはわずか 218 件にすぎません。218 人の被験者のうち、約 66% が陽性、残り 34% が陰性であることが分かります。

決定木を下方にたどりながら、ビジュアライザの左上のスライダをドラッグして高さスケールを調整し、ツリーの様々なレベルを確認してください。陽性の被験者を数多く含むノードは、"fti <= 64.5 and tsh > 5.95" という条件に合致しています。このノードには、151 人の陽性被験者のうち 140 人が含まれています。

ピマ族における糖尿病の診断

pima データセットには、アリゾナ州フェニックスのアメリカ・インディアン（ピマ族）から収集した糖尿病診断に関する統計データが収録されています。この例の目的は、年齢、体重、血圧、血糖値などの医学データに基づいて、被験者が糖尿病かどうかを診断することです。

pima-dt.treeviz ファイルには、この問題について作成された決定木クラシファイアが収録されています。この可視化ファイルは、*pima.schema* ファイルを分析して作成されたものです。

- Windows システムの場合、これらのファイルは MineSet がインストールされているディレクトリ *\examples\pima-dt.treeviz* および *\data\pima.schema* にあります。
- IRIX システムの場合、これらのファイルは */usr/lib/MineSet/treeviz/examples/pima-dt.treeviz* および */usr/lib/MineSet/data/pima.schema* にあります。

DNA 境界

DNA データセットには、分子生物学分野の GenBank 64.1 (*genbank.bio.net*) から抽出された 3,186 件の DNA レコードが収録されています。DNA 連鎖上の点のうち、蛋白質の生成中に余分な DNA が除去される箇所は組継ぎ接合部 (Splice junction) と呼ばれます。この例の目的は、DNA のエクソン / イントロン (exon/intron) 境界（EI サイト）、イントロン / エクソン (intron/exon) 境界（IE サイト）、無境界 (none) を識別することです。IE 境界はアクセプタと呼ばれ、EI 境界はドナーと呼ばれることがあります。DNA データセットの各属性によって、60 種類のヌクレオチドが表されます。これらの属性が 2 分割されるため、接合部の両側にそれぞれ 30 種類のヌクレオチドが存在することになります。

この例では、決定木のルートに 3 つのクラスがあります。各クラスのバーをポイントすると、exon/intron (EI) が約 24%、intron/exon (IE) が 24%、none が 52% というクラス構成になっていることが分かります。ルートノードの前にある "left_01" は、最初に分析する必要のある重要な属性を示しています。"left_01" という表記は、組み継ぎ接合点の左側で検出された最初のヌクレオチドを意味します。この最初のヌクレオチドで選

択できる属性値（通常はその他のヌクレオチドについても同じ）は、"A"、"G"、"T"、"C" の 4 種類のヌクレオチドです。たとえば、"left_01" のヌクレオチドが "G" である場合は、"G" の枝に沿って次のノードに進みます。そのノードのクラス構成は exon/intron (EI) が 34%、intron/exon (IE) が 42%、none が 24% となっており、ルートノードよりも "exon/intron" または "intron/exon" の確率が高いことが分かります。"left_01" のヌクレオチドが "A"、"T"、または "C" である場合は、"A"、"T"、または "C" の枝に沿って次のノードに進むます。これら 3 個のノードでは、"none" の確率がそれぞれ 87%、87%、95% と飛躍的に増加します。このようなテストと分岐を繰返して最後のノードに到達すると、最終的な予想値（"exon/intron"、"intron/exon"、または "none"）が得られます。

このデータセットについては、分子生物学の確率的性質から決定木よりもエビデンス・クラシファイアの方が適しています。このことは推定誤差率を比較すると確認できます。

デシジョン・テーブル用のサンプルファイル

ここでは、デシジョン・テーブルの適用が有効と思われる事例を紹介します。MineSet には、これらの事例で使用するサンプルファイルが用意されています。「詳細設定」の「提唱」をオンに設定してデシジョン・テーブル分析を実行すると、次に説明する *-dtab.dtableviz* ファイルが生成されます。

注記 : データファイル (.data) とそれに対応するスキーマファイル (.schema) はクライアント・ワークステーションの *data* ディレクトリにあります。クラシファイアの可視化ファイル（拡張子 *-dtab.dtableviz*）はクライアント・ワークステーションの *examples* ディレクトリにあります。スキーマファイル (*.schema*) を開くと、それに対応するデータファイル (.data) が自動的に読込まれます。

- Windows システムの場合、MineSet をインストールした *\examples* および *\data* 配下のディレクトリにあります。
- IRIX システムの場合、*/usr/lib/MineSet/treeviz/examples* と */usr/lib/MineSet/data* にあります。

解約 (Churn)

顧客が別の会社に移ってしまうことを「解約」（顧客離れ）と言います。解約は電話会社にとって共通の重要な問題です。この事例では、顧客が解約する原因を解析しています。この事例で使用する元データは、*churn.schema* ファイルに格納されています。Windows システムの場合、これらのファイルは MineSet をインストールした *\Program Files\SGI\MineSet 3.0\data* にあります。IRIX システムの場合、*/usr/lib/MineSet/data* にあります。

churn-dtab.dtableviz ファイルには、「解約 (churned)」属性をラベルとして生成されたクラシファイアの構造が収録されています。このクラシファイアの誤差率は 5.5% です。解約した顧客は全レコードの 14.3% を占めています。最上位レベルとして選択された 2 つの属性は、顧客サービスコール数 (number_of_customer_service_calls) と昼間料金 (total_day_charge) です。これら 2 つの属性の分布を見ると、「昼間料金 (total_day_charge)」が増加するに従って解約率が上昇すること（ただし、total_day_charge < 29.75 の場合は、顧客サービスコール数が 4 回以上であるときに解約率が高いこと）が分かります。昼間料金が 38 未満、かつ顧客サービスコール数が 3 回以下のレコードは、全レコードの約 3/4 を占めています。

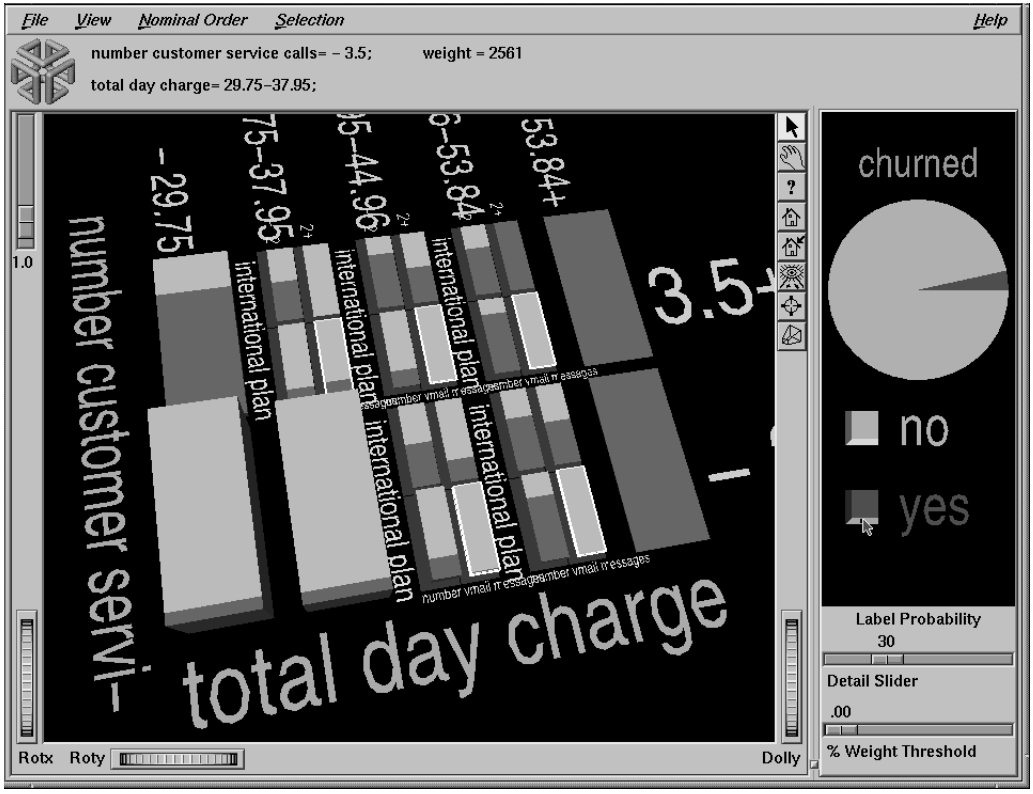

図 A-1 「解約 (Churn)」データセットに対するドリルダウン（IRXI システム）

ここで顧客が解約するときの条件が明確でない領域をドリルダウンしてみます。図 A-1 に、優勢なクラスが存在しない領域の全ケーキをドリルダウンした状態を示します。この図を見ると、次に検討する必要のある重要な属性は「国際化プラン (international_plan)」と「ボイスメール メッセージ数 (number_of_vmail_messages)」であることが分かります。昼間の料金が非常に多い顧客については、「国際化プラン (international_plan)」を契約し、「ボイスメールのメッセージ数 (number_of_vmail_messages)」の少ないことが解約率と高い相関のあることが明白です。ドリルダウンの対象領域の右下にあるケーキを選択した後、右側のラベル確率ウィンドウ内で「解約 = はい (churned=yes)」の左側にあるボックスをマウスでクリックすると、選択した領域内の顧客の解約率はわずか 3.4% であることが分かります。

次に、上記でドリルダウンした各領域の左上にあるケーキをさらにドリルダウンします。そうすると、各領域での属性の分布が非常に似ていることが分かります。すなわち、一貫したパターンとして、「夜間料金 (total_evening_charge)」および「国際通話料金 (total_international_charge)」の多いことが解約率と高い相関があります。別のレベルをドリルダウンして「国際通話数 (total_international_calls)」の影響を調べることもできますが、そうすると推測の基準とするレコード数が少なくなるため、属性の分布について確信的な予測が行えません。「国際通話数 (total_international_calls)」が解約率に与える影響および国際通話数と他の属性との相関を確認する場合は、Tool Manager に戻り、「国際通話数 total_international_calls」を階層の上位レベルに明示的にマッピングし、デシジョン・テーブル分析を再実行してください。

「州 (state)」属性は解約率と高い相関がありますが、値の数が少ない属性を優先するロジックが帰納アルゴリズムに組み込まれているため、「州 (state)」属性は選択されていません。そのため、各レコードを一意に識別する社会保険番号のような属性が選択される状況が避けられます（このような属性を選択すると、訓練事例 (tranining set) の精度は高くなりますが、ラベルの付いていないデータをクラス判別するのに有効なデータは作成されません）。

車の原産国

「車 (*cars*)」データセットには、1970 年代から 1980 年代初頭にかけての様々な車種に関する情報が格納されています。これらの情報には、車の重量、速度、mpg（ガロン当たりの走行マイル数）などがあります。*cars-dtab.dtableviz* ファイルには、車の原産国（ヨーロッパ、日本、またはアメリカ）を予測するために帰納されたデシジョン・テーブル・クラシファイアの構造図が収録されています。この可視化ファイルは、ラベルを origin (Japan, U.S., Europe) に設定し、*cars.schema* ファイルを分析して作成されたものです。

- Windows システムの場合、これらのファイルは MineSet がインストールされているディレクトリ *\examples\cars-dtab.dtableviz* および *\data\cars.schema* にあります。
- IRIX システムの場合、これらのファイルは */usr/lib/MineSet/examples/dtableviz/cars-dtab.dtableviz* および */usr/lib/MineSet/data/cars.schema* にあります。

車の原産国は「ブランド (brand)」属性によって明確に表されるため、分析モデルの構造は非常に単純になります。表示されている属性は、「ブランド (brand)」と「シリンダー (cylinders)」の 2 つだけです。「ブランド (brand)」と「シリンダー (cylinders)」には興味深い関係があります。たとえば、「マツダ (Mazda)」には 21 種類の車種、「ホンダ (Honda)」には 18 種類の車種がありますが、すべての車種が 5 気筒以下です。それに対して、「キャデラック (Cadillac)」の車種はすべて 6 気筒以上です。

最初に「ブランド (brand)」属性を除外すると、この事例はさらに面白くなります。また、「シリンダー (cylinders)」属性を文字列に変換し、階級生成ではなくすべての属性値が使用されるようにしたり、「ブランド (brand)」属性と「 シリンダー (cylinders)」属性を明示的にマッピングして別の属性レベルを作成したりすると、興味深い解析を行うことができます。

性別の判断

「成人 (*adult*)」データセットには成人の就業者に関する情報が格納されています。このデータセットは米国人口統計局 (U.S. Census Bureau) のデータベースから抽出したものであり、年間の総収入が $100 以上、1 週間の就業時間が 1 時間以上、年齢が 16 歳以上という条件を満たす就業者に関するデータが収録されています。

adult-sex-dtab.dtableviz ファイルには、これらの就業者の性別を判断するために作成されたデシジョン・テーブル・クラシファイアの構造図が収録されています。この可視化ファイルは、ラベルを「性別 (sex)」に設定し、（自明なクラシファイアが作成されないように）婚姻関係を表す属性を除去し、*adult.schema* ファイルを分析して作成されたものです。各属性ペアのレコードの分布を確認しやすいように、左側の「スケール」スライダを使用してケーキの高さを調整することができます。

- Windows システムの場合、これらのファイルは MineSet がインストールされているディレクトリ *\examples\adult-sex-dtab.dtableviz* および *\data\adult-sex.schema* にあります。
- IRIX システムの場合、これらのファイルは */usr/lib/MineSet/examples/dtableviz/adult-sex-dtab.dtableviz* および */usr/lib/MineSet/data/adult-sex.schema* にあります。

デシジョン・テーブル・ビジュアライザの「ラベル確率」ウィンドウを見ると、男性の事前確率が女性の事前確率よりも高いことが分かります。エビデンス・ビジュアライザ (Evidence Visualizer) による解析では、「婚姻歴 (marital_status)」と「業種 (occupation)」が性別の判断にあたって非常に重要な属性であることが分かりましたが、これら 2 つの属性間の相関関係は示されませんでした。デシジョン・テーブル・ビジュアライザのグラフの最上位レベルを見ると、複数の相関関係が示されています。

たとえば、職業が「修理業 (craft repair)」であるほとんどの人は既婚の軍人以外との結婚（より正確には 98.6% が既婚男性）ですが、職業が「他のサービス業 (Other-service)」であるほとんどの人は結婚の経験がなく ("Never-married")、そのうち 48% は男性です。

最初は、" 婚姻歴 = 軍人以外との結婚 " (marital_status = Married-civilian-spouse) であるほとんどの人が男性であるという事実が奇妙に思われるかもしれません。ただし、このデータセットが税金統計に基づいていることを考慮すると、それらの男性の働いていない妻のデータが夫のデータとともに収集されたと推測するのが合理的です。

" 業種 = 秘書 " (occupation = Admin-clerical) である人は離婚率が非常に高くなっています。また、" 業種 = その他のサービス業 " (occupation = Other-service) である人も離婚率が高いですが、別居している人の数が "Admin-clerical" よりも多いため、別居を好む傾向の強いことが分かります。

職業が秘書であり ("occupation = Adm-clerical")、かつ未亡人 ("widowed") を前提条件として、その就業者が女性である確率を予想してみます。エビデンス・ビジュアライザ (Evidence Visualizer) では、これら 2 つの属性値をクリックすることによって、女性である確率は約 94.7% という答えが得られました。デシジョン・テーブル・ビジュアライザでは、2 つの属性値の交差位置にあるケーキをマウスの左ボタンでクリックすることによって、95.2% という正確な答えが得られます。

最後に、" 軍人以外との結婚 " かつ " 業種 = その他 " (Occupation = Unknown) という条件のケーキを最下位レベルまでドリルダウンしてみます。この条件のケーキについては、既婚 / 未婚と職業の他の組合わせよりも、若い人が女性であり、年配の人が男性であるという傾向が強く観察されます。

給料を決める要因

次に、成人就業者のデータセットを解析して給料水準を決める要因を見つけます。最初に、連続値の「総収入 (*gross_income*)」属性を 2 つの階級（収入金額が $50,000 未満であるクラスと $50,000 以上であるクラス）に分割し、新しい離散型の総収入属性 (gross_income_bin) を作成します。各レコードにはこの離散型属性のいずれかの値がマッピングされます。MineSet のクラシファイアを実行すると、給料水準に影響を与える要因を簡単に見つけることができます。*adult-salary-dtab.dtableviz* ファイルには、この問題について作成されたデシジョン・テーブル・クラシファイアの構造図が収録されています。この可視化ファイルは、区間を指定して離散型総収入 (gross_income) を 5 つの階級に分割し、*adult.schema* ファイルに対して分析して作成されたものです。

- Windows システムの場合、これらのファイルは MineSet がインストールされているディレクトリ *\examples\adult-salary-dtab.dtableviz* および *\data\adult-salary.schema* にあります。
- IRIX システムの場合、これらのファイルは */usr/lib/MineSet/examples/dtableviz/adult-salary-dtab.dtableviz* および */usr/lib/MineSet/data/adult-salary.schema* にあります。

ラベル (gross_income) が数値型であるため、連続的なスペクトルに基づく色が各階級にマッピングされます（最大値 (50,000+) の階級は赤になります）。各階級のラベルは、確率値の大きい順にラベル確率ウィンドウ内に並べられます。

階層の最上位レベルとして、「家族構成 (relationship)」と「教育番号 (education_num)」という 2 つの属性が選択されます。ただし、「教育番号 (education_num)」は教育年数ではなく様々な最終学歴を表す列挙番号であるため、それほど重要な属性ではありません。ただし、教育水準は給料水準と何らかの相関があります。実際の文字列値を見るには、「教育番号 (education_num)」属性を「最終学歴 (education)」属性で置き換えます。「教育番号 (education_num)」属性を単に除外し、「提唱」を使用して分析を再実行しても、「最終学歴 (education)」属性は階層の最上位レベルとして選択されない可能性があります（属性値の数が非常に多いためです）。

この分析モデルの属性の順序は、精度を上げるように自動的に選択されています。通常、各分野の専門知識に基づいて属性のマッピングを行うと、より効果的なビジュアル・モデルが作成されます。そのような方法で作成されたモデルが *adult-salary3-dtab.dtableviz* ファイルに収録されています（図 A-2 参照）。このモデルでは、最初に、区間を 20,000 および 60,000 と指定して、連続値の「総収入 (gross_income)」属性が 3 つの階級に分割されています。最上位レベルに割り当てられた属性は「家族構成 (relationship)」と「性別 (sex)」)、2 番目のレベルに割り当てられた属性は「最終学歴 (education)」と「業種 (occupation)」、3 番目のレベルに割り当てられた属性は「週の労働時間 (hrswk)」と「年齢 (age)」です。

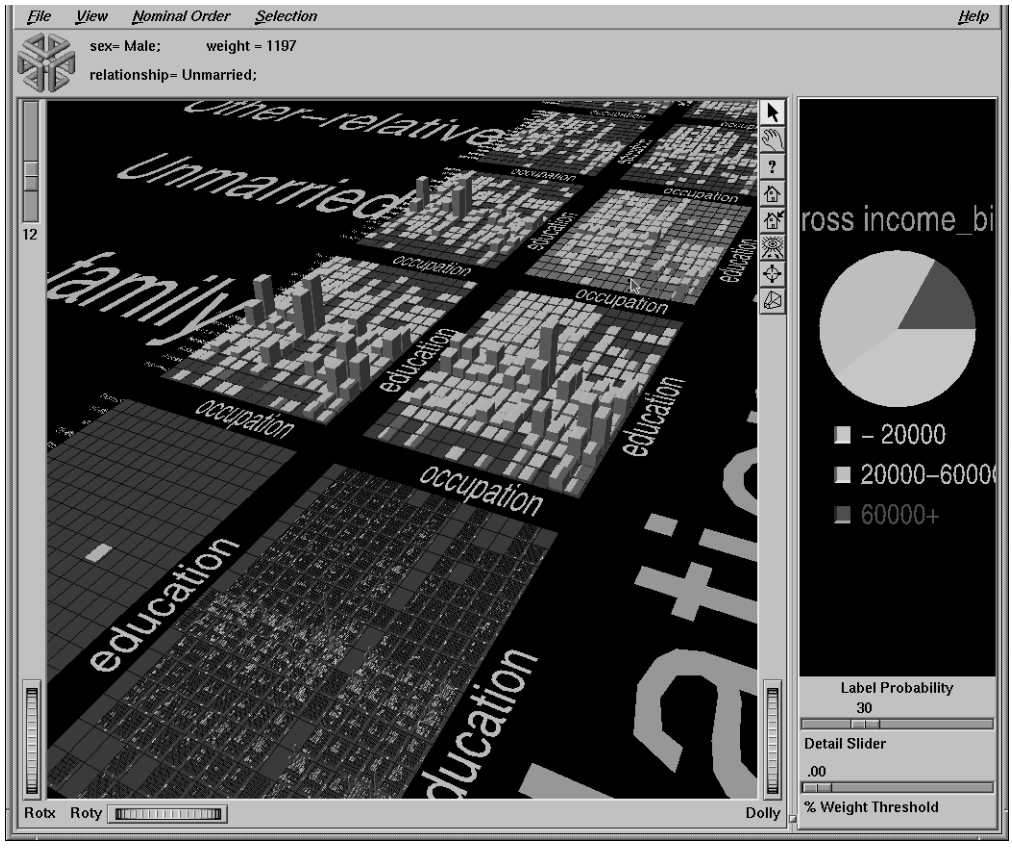

図 A-2 デシジョン・テーブル・ビジュアライザによる「成人 (adult)」データセットの解析（IRXI システム）

最上位レベルの属性である「家族構成 (relationship)」と「性別 (sex)」) には強い相関関係があることが分かっています。すべての「夫 (husbands)」は男性であり、すべての「妻 (wives)」は女性であると考えるのが当然ですが、実際にはそうでないことが即座に確認できます。「男性の妻 (male wives)」を表すケーキにカーソルを移動すると、該当するレコードが 3 件あり、それらのレコードの給料は 20,000 ～ 60,000 の階級に属することが分かります。これらのケーキをドリルダウンすると、異常値のレコードに関する詳細な情報を確認することができます。マウスの左ボタンをクリックしてこれらのケーキを選択し、ドリルダウンを行って元データを詳しく調べてください。

ウィンドウの背景上でマウスの右ボタンをクリックすると、次のレベルにあるすべてのケーキに対してグローバルなドリルダウンが行われます。すなわち、個々のケーキの代わりに、「最終学歴 (education)」属性と「業種 (occupation)」属性のすべての値の組合わせを示すマトリックスが表示されます。表示順序はすべてのマトリックスで同じであり、給与水準との相関の高い順に属性値のペアが並べられます。「属性値の順序付け」メニューから「重み順 」を選択すると、表示順序がレコードの重み順に変わります。最も一般的な教育水準と職業が各マトリックスの左下隅に表示されます。最も一般的な教育水準は「高校卒業 (High school grads)」であり、最も多い職業は「専門家 (Professional-specialty)」ですが、職業が専門家である高校卒業の就業者数はそれほど多くありません。

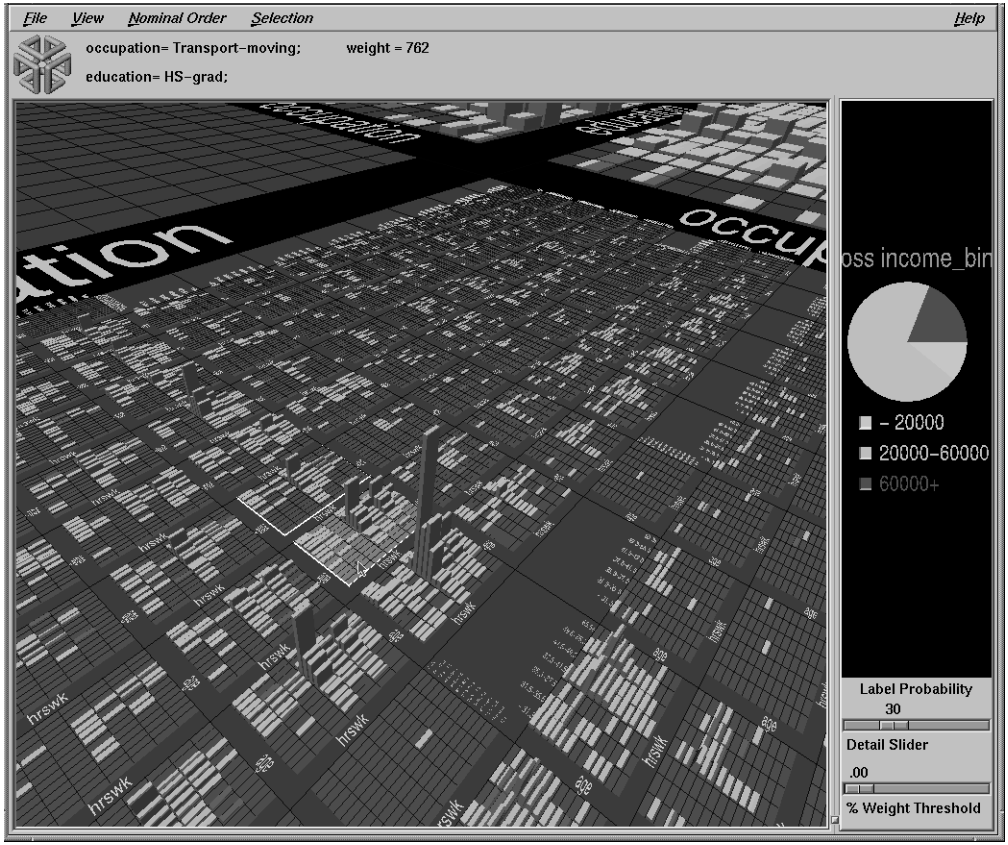

図 A-3 「成人 (adult)」データセットの詳しい解析

給与水準との相関の高い順に属性値を並べ替えると、「家族構成 (relationship)」と「性別 (sex)」の属性値の組合わせごとに分布が異なることが分かります。家族と同居していない (not-in-family) 男女間ではそれほど大きな違いは見られませんが、未婚の男女間では顕著な違いが見られます（図 A-3 参照）。すなわち、男性のマトリックスの左下には、女性のマトリックスには存在しない赤いケーキの塊がはっきりと表示されています。また、ケーキの高さを拡大すると、女性のマトリックスでは " 業種 = 秘書 " "(occupation = admin clerical)" および " 業種 = その他のサービス "(occupation = other service)" の箇所に明白なスパイク（隆起）が見られますが、男性のマトリックスにはそのようなスパイクが見られません。

最もレコード数の多いケーキ（male husbands: 男性の夫）をマウスの右ボタンでクリックしてください。ケーキの次のレベルの全ジオメトリがビジュアライザによって構築されるため、この操作には数分の時間がかかる場合があります。ジオメトリを最初から構築するには非常に長い時間がかかること、最下位に近い詳細レベルをマイニングする必要性はほとんど生じないことなどから、ジオメトリは自動的ではなく要求に応じて構築されるようになっています。ウィンドウの背景を誤ってクリックすると、（この事例のように）次のレベルにあるジオメトリが多い場合は、非常に長い時間にわたって待たされる可能性があります。ドリルダウンに長い時間がかかる場合は、「取消 (Cancel)」ボタンの付いた進行状況バーが表示されます。

「最終学歴 (education)」と「業種 (occupation)」の組合わせごとに、「週の労働時間 (hrswk)」と「年齢 (age)」の組合わせが数多く表示されています。ここで驚くべき事実は、何百ものケーキが存在するにも関わらず、すべての既婚男性の 2.5% に相当する箇所に 1 つのスパイクが存在することです（図 A-3 参照）。典型的な既婚男性の特徴を列挙するとすれば、最終学歴は高卒、業種は修理業、年齢は 41 〜 59 歳、1 週間の労働時間は 38 〜 41 時間と言うことができます。

「高校卒業 (HS-grads)」の既婚男性セールスマンと高卒の既婚男性役員の給料を比較すると、年齢と就業時間の分布がほとんど同じ場合でも、給料が $60,000 を越える確率は、役員が 34% であるのに対して、セールスマンは 27% となっています。これらの確率値をウィンドウの一番上に表示するには、最初に右側のラベル確率のウィンドウ内でクラスラベル "60000+" の左側にあるボタンをクリックした後、左側のケーキの上にカーソルを移動します。

アヤメのクラス判別

「アヤメ (Iris)」データセットの各レコードには、アヤメの特性を表す 4 つの属性（萼片の縦、萼片の横、花びらの縦、花びらの横 (sepal_length, sepal_width, petal_length, petal_width)）があります。これらの属性に基づいて、アヤメは 3 つのクラス（アイリスセトサ、アイリスバージカラー、アイリスバージニカ）(iris-setosa, iris-versicolor, iris-virginica) に分類されます。この事例の目的は、アヤメの各クラスを特徴付ける属性を確定することです。

分析を実行する前に、Tool Manager の「クラス判別」タブにある「重要項目」タブをクリックした後、「示唆」チェックボックスをオンにして「実行」ボタンをクリックしてください。各属性 (petal_width、petal_length、sepal_width、sepal_length) の寄与率がランク付けされたリストが表示されます。これらの属性をスキャタ・ビジュアライザの軸に割り当て、アイリスの種類 (iris_type) を色にマッピングすると、同じ特性を持つクラスタを確認することができます。

iris-dtab.dtablerviz ファイルには、この問題について作成されたデシジョン・テーブル・クラシファイアの構造図が収録されています。この可視化ファイルは、*iris.schema* ファイルを分析して作成されたものです。

- Windows システムの場合、これらのファイルは MineSet がインストールされているディレクトリ *\examples\iris-dtab.dtableviz* および *\data\iris.schema* にあります。
- IRIX システムの場合、これらのファイルは */usr/lib/MineSet/examples/dtableviz/iris-dtab.dtableviz* および */usr/lib/MineSet/data/iris.schema* にあります。

デシジョン・テーブル・ビジュアライザを見ると、分析モデルに最も効果的な属性は「花びらの横 (petal_width)」であることが分かります。「萼片の横 (sepal_width)」を追加すると、"petal width= 0.75-1.65" であるレコードについては、"iris_versicolor" のすべての事象が "sepal width<3.05" の階級に表示されます。

100% 寄与率ではない（全レコードが単一のクラスに属していない）3 つのケーキをドリルダウンしてみます。一番上の 2 つのケーキには iris-versicolor の 1 つの事象が含まれているため、これらのケーキは純粋ではありません。"sepal_width<3.05" のケーキについては、異常値である "iris-versicolor" を分離するのは非常に困難です。ただし、"sepal_width<3.05" のケーキについては、petal_length を使用して iris_versicolor が分離されています。

キノコの分析モデル

mushroom-dtab.dtableviz ファイルには、キノコの食用性を判定するために作成されたデシジョン・テーブル・クラシファイアの構造図が収録されています。この可視化ファイルは、*mushroom.schema* ファイルを分析して作成されたものです。

- Windows システムの場合、これらのファイルは MineSet がインストールされているディレクトリ *\examples\mushroom-dtab.dtableviz* および *\data\mushroom.schema* にあります。
- IRIX システムの場合、これらのファイルは */usr/lib/MineSet/examples/dtableviz/mushroom-dtab.dtableviz* および */usr/lib/MineSet/data/mushroom.schema* にあります。

この事例の目的は、食用になるキノコと毒性のあるキノコを判別することです。このデータセットには 8,000 件以上のレコードが含まれているため、分析をするのに数分の時間がかかる場合があります。デフォルト・モードの予備法でモデルの精度を評価するときは、全レコードの 1/3 がテストセット (test set) として保持されることに注意してください。

個々のキノコには、傘の色、傷、香りなどの様々な属性があります。デシジョン・テーブル・ビジュアライザでは、「香り (odor)」と「柄の形 (stalk-shape)」が最上位レベルの属性として表示されます。「香り (odor)」単独でも食用性を判定する効果的な属性になることに注意してください。食用性の判定が困難になるのは、キノコに香りがなく (odor=none)、柄の形が先太になっている (stalk-shape=enlarging) 場合に限られます。そこで、これら 2 つの属性値を持つケーキをドリルダウンすると、「傷 (bruises)」と「ひだのサイズ (gill-size)」の属性値によって該当のレコードが分解されることが分かります。「ひだのサイズ (gill-size)」と「傷 (bruises)」には相関がありますが、この相関関係は他のクラシファイアではほとんど識別できません。

このデータセットの全属性は離散的であるため、すべての属性値がラベル確率（食用性の判定に対する寄与率）に従って左から右にソートされます。属性値をアルファベット順または重み順にソートすることもできます。その場合は、「属性値の順序付け」メニューから「アルファベット順」または「重み順」を選択してください。「ひだのサイズ (gill-size)」または「傷 (bruises)」のどちらか一方の属性だけでは、食用になるキノコと毒性のあるキノコを完全に判別することはできません。ただし、これら 2 つの属性を組合わせた場合、傷がなくて傘の幅が広いタイプのうち 814 個のキノコは食用可能であり、傷があって傘の幅が狭いタイプのうち 11 個のキノコは毒性のあることが分かります。他の 2 つのケース（傷がなくてひだのサイズが小さい、傷があってひだのサイズが大きい）を明らかにするには、さらにドリルダウンを続ける必要があります。

デシジョン・テーブル・ビジュアライザの「% の有意水準」スライダを右側に移動すると、" 香り = カビ臭い " (odor=musty) という属性値がグラフから除外されます。これは odor=musty に該当するレコードの重み値が 1% 未満であるためです。

政党への帰属

この事例の目的は、主要な票決に関するデータに基づいて特定の下院議員が帰属する政党を判断することです。「政党 (vote)」データセットには、議会季刊誌 (*CQA*:*Congressional Quarterly Almanac*) に記載された 16 回の主要な票決において米国下院の各議員が投票した内容が収録されています。CQA では投票の内容として、賛成 (yes) 3 種類、反対 (no) 3 種類、棄権 (undecided) 3 種類の計 9 種類を規定しています。

クラシファイアを実行する前に、16 回の票決内容を解析して、重要と思われる特性を考えてみてください。続いて、デシジョン・テーブル・ビジュアライザを実行します。このデータセットについては属性値をアルファベット順にソートし、賛成票 (yes) が右上、反対票 (no) が左下に表示されるようにします。

デシジョン・テーブル・ビジュアライザでは、「合成燃料会社の削減 (synfuels_corporation_cutback)」と「医療費の据置き (physician_fee_freeze)」が最上位レベルの属性として表示されます。これら 2 つの属性間には、他の分析モデルではほとんど識別できない微妙な相関関係があります。医療費の据置きに反対した議員は、その全員が民主党員でした。彼らのほとんど全員が合成燃料会社の削減にも反対しました（このパターンに該当しなかったのは、206 人のうち 3 人だけでした）。いっぽう驚くべきことに、賛成した共和党員のほとんど全員（5 人を除く全員）が合成燃料会社の削減に反対しました。このような奇妙な関係は、これらの非常に異なる議題の間に詳しい解析を要する相関が存在することを示唆しています。

最上位レベルの大部分のケーキは、ほとんど純粋です（ほとんどのレコードが単一のクラスに属しています）。ただし、6 件のレコードしか含まない中央のケーキと、議員が両方の議題に賛成したことを示すケーキは純粋ではありません。次のレベルをドリルダウンすると、このグループ内の議員が帰属する政党を判断することができます。そのためには、「対衛星試験禁止令 (anti-satellite test ban)」と「予算決議案の採択 (adoption of the budget resolution)」に基づいて、55 人の議員の特性を細かく解析します。

vote-dtab.dtableviz ファイルには、この問題について作成されたデシジョン・テーブル・クラシファイアの構造図が収録されています。この可視化ファイルは、vote.schema ファイルを分析して作成されたものです。

- Windows システムの場合、これらのファイルは MineSet がインストールされているディレクトリ *\examples\vote-dtab.dtableviz* および *\data\vote.schema* にあります。

- IRIX システムの場合、これらのファイルは */usr/lib/MineSet/examples/dtableviz/vote-dtab.dtableviz* および */usr/lib/MineSet/data/vote.schema* にあります。

乳癌の診断

「乳癌 (Breast)」データセットには、乳癌の診断を受けた女性被験者に関する情報が収録されています。このデータセットの各レコードは、細胞の大きさ、凝集塊厚、周辺癒着などの属性を持つ被験者を表します。この事例の目的は、被験者の腫瘍が悪性であるか良性であるかを判断することです。*breast-dtab.dtableviz* ファイルには、この問題について作成されたデシジョン・テーブル・クラシファイアの構造図が収録されています。この可視化ファイルは、*breast.schema* ファイルを分析して作成されたものです。

- Windows システムの場合、これらのファイルは MineSet がインストールされているディレクトリ *\examples\breast-dtab.dtableviz* および *\data\breast.schema* にあります。
- IRIX システムの場合、これらのファイルは */usr/lib/MineSet/examples/dtableviz/breast-dtab.dtableviz* および */usr/lib/MineSet/data/breast.schema* にあります。

デシジョン・テーブル・ビジュアライザでは、「細胞分裂 (mitosis)」と「細胞形状の不均一 (uniformity of cell shape)」が最上位レベルの属性として表示されます。これら 2 つの属性値が低い場合、その腫瘍は 99.2% の確率で良性です。一方、訓練事例のレコードのうち、これら 2 つの属性値が高い腫瘍はすべて (100%) 悪性でした。

寄与率がそれほど高くない 4 つのケーキをドリルダウンし、「周辺癒着 (marginal adhesion)」と「分離核 (bare-nuclei)」に基づいて良性 / 悪性の判定を行います。このレベルでは各ケーキのレコード数が非常に少なくノイズが多いため、トレンドの検出がかなり困難になります。「周辺癒着 (marginal adhesion)」と「分離核 (bare-nuclei)」の値が高い場合、腫瘍が悪性である傾向が強くなりますが、絶対確実な判定基準ではありません。「分離核 (bare-nuclei)」の最初の値が NULL であることに注意してください。これらの NULL ケーキの分布は信頼性が低いため、「表示」メニューの「NULL 位置の表示」オプションのチェックマークを外して、NULL ケーキを非表示にしてください。

2 つ下のレベルまでグローバルなドリルダウンを行うと、いくつかの興味深い特性が見つかります。ケーキが非常に小さくなり、何もない多次元空間が増えてきます。また、数多くのレコードが集中した小さい領域がいくつか存在します。すべての属性値が減少する 1 つの大きいスパイク（100% 良性）が観察されます。このスパイクだけでデータ全体の約 20% を占めています。

甲状腺機能低下症 (Hypothyroid) の診断

「甲状腺機能低下症 (hypothyroid)」データセットの構造は、上記の「乳癌 (breast)」データセットの構造と似ています。*hypothyroid-dtab.dtableviz* ファイルには、甲状腺機能低下症を診断するために作成されたデシジョン・テーブル・クラシファイアの構造図が収録されています。この可視化ファイルは、*hypothyroid.schema* ファイルを分析して作成されたものです。

- Windows システムの場合、これらのファイルは MineSet がインストールされているディレクトリ *\examples\hypothyroid-dtab.dtableviz* および *\data\hypothyroid.schema* にあります。
- IRIX システムの場合、これらのファイルは */usr/lib/MineSet/examples/dtableviz/hypothyroid-dtab.dtableviz* および */usr/lib/MineSet/data/hypothyroid.schema* にあります。

「甲状腺機能低下症 (hypothyroid)」データセットには 3,163 件のレコードが収録されていますが、ほとんど (95.45%) の被験者は陰性です（甲状腺機能低下症に罹っていません）。したがって、特定の被験者を陰性と診断することは高い確率で正解になりますが、最も重要なのは実際に甲状腺機能低下症に罹っている（陽性の）被験者です。陽性の被験者を陰性と診断するのは重大な問題です。

陽性の被験者が陰性と診断されるのを避けるために、陽性と判定される事後確率が高くなるように損失マトリックスを調整する必要があります。実際に病気の人を健康であると診断すると大きな損失が発生しますが、実際に健康な人を病気であると診断しても、その人が精密検査や不要な治療を受けるだけで済みます。

このデータセットにデシジョン・テーブル・ビジュアライザを適用すると、次のようなことが分かります。

- fti が NULL である場合、tbg は NULL ではなく、t3 は高い確率で NULL である（1 つ下のレベルにドリルダウンする）。
- 2 つの事象を除いて、t4u が NULL になるのは、fti が NULL の場合に限られる。
- fti の値が低い場合は、甲状腺機能低下症である確率が高い。

ピマ族における糖尿病の診断

「糖尿病 (pima)」データセットには、アリゾナ州フェニックスのアメリカ・インディアン（ピマ族）から収集した糖尿病診断に関する統計データが収録されています。この事例の目的は、血圧、体重、血糖値、年齢などの医学データに基づいて、被験者が糖尿病かどうかを診断することです。

pima-dtab.dtableviz ファイルには、この問題について作成されたデシジョン・テーブル・クラシファイアの構造図が収録されています。この可視化ファイルは、*pima.schema* ファイル を分析して作成されたものです。

- Windows システムの場合、これらのファイルは MineSet がインストールされているディレクトリ *\examples\pima-dtab.dtableviz* および *\data\pima.schema* にあります。
- IRIX システムの場合、これらのファイルは */usr/lib/MineSet/examples/dtableviz/pima-dtab.dtableviz* および */usr/lib/MineSet/data/pima.schema* にあります。

このデータセットにデシジョン・テーブル・ビジュアライザを適用すると、次のようなことが分かります。

- 2 番目のレベルには、年齢が 28 歳以下で妊娠 6 カ月以上の女性が 4 人表示されている。これらのレコードをドリルスルーすると、これら 4 人の被験者に関する詳しい情報が分かる。
- 「血糖値 (plasma_glucose)」と「肥満指数 (body mass)」が高い場合は、糖尿病である確率が高い。

DNA 境界

dna-dtab.dtableviz ファイルには、この問題について作成されたデシジョン・テーブル・クラシファイアの構造図が収録されています。この可視化ファイルは、*dna.schem*a ファイルを分析して作成されたものです。

- Windows システムの場合、これらのファイルは MineSet がインストールされているディレクトリ *\examples\dna-dtab.dtableviz* および *\data\dna.schema* にあります。
- IRIX システムの場合、これらのファイルは */usr/lib/MineSet/examples/dtableviz/dna-dtab.dtableviz* および */usr/lib/MineSet/data/dna.schema* にあります。

DNA データセットには、3,186 件の DNA レコードが収録されています。DNA 連鎖上の点のうち、蛋白質の生成中に余分な DNA が除去される箇所は組継ぎ接合部 (Splice junction) と呼ばれます。この事例の目的は、DNA の exon/intron 境界（EI サイト）、intron/exon 境界 (IE サイト)、無境界 (none) を識別することです。IE 境界は " アクセプタ " と呼ばれ、EI 境界は " ドナー " と呼ばれることがあります。これらのレコードは元々 GenBank 64.1 (*genbank.bio.net*) から抽出したものです。DNA データセットの各属性によって、60 種類のヌクレオチドが表されます。これらの属性が 2 分割されるため、接合点の両側にそれぞれ 30 種類のヌクレオチドが存在することになります。

デシジョン・テーブル・ビジュアライザを見ると、他の分析モデルではほとんど識別できないような驚くべきパターンを検出することができます。最上位レベルでは、左 _01 (left_01) と 右 _02 (right_02) 間の相関が明示されています。Exon/intron が存在するのは "right_02 = T" の場合に限られ、Intron/extron が存在するのは "left_01=G" の場合に限られています。left_01 と right_01 が他の値である場合、接合部はほとんど存在しません。

次のレベル (left_02 と right_01) に対してグローバルなドリルダウンを行うと、right_02 = T かつ left_01 = G のレコードについては、各エッジに沿って同様のパターンが見られることが分かります。

エビデンス・ビジュアライザ用のサンプルファイル

ここでは、エビデンス分析の適用が有効と思われる事例を紹介します。MineSet では、これらの事例で使用するサンプルファイルが用意されています。エビデンス分析を実行すると、下記で説明する *.eviviz* ファイルが生成されます。

data ディレクトリから対応するスキーマ (*.schema*) ファイル（*churn.schema* など）を開くと、それに対応するデータファイル (*.data*) が自動的に読込まれます。クラシファイアの可視化ファイル（拡張子 *.eviviz*）は、examples ディレクトリから開くことができます。

解約 (Churn)

顧客が別の競合会社に移ってしまうことを「解約」（顧客離れ）と言います。解約はどの電話会社にとっても共通の重要な問題です。この事例では、顧客が解約する原因を分析しています。

この事例で使用する元データは、*churn.schema* と *churn.data* ファイルに格納されています。スキーマファイル (*.schema*) を開くと、それに対応するデータファイル (.data) が自動的に読込まれます。

- Windows システムの場合、これらのファイルは MineSet がインストールされているディレクトリ *\examples\churn.data* および *\data\churn.schema* にあります。
- IRIX システムの場合、これらのファイルは */usr/lib/MineSet/examples/eviviz/churn.data* および */usr/lib/MineSet/data/churn.schema* にあります。

可視化ファイル *churn.eviviz* には、処理 (churned) という属性をラベルとして生成されたクラシファイアの構造図が収録されています。このクラシファイアの誤差率は 12% です。解約した顧客は全レコードの 14.1% を占めています。最も重要な 2 つの属性である昼間通話時間と昼間料金 （total_day_minutes と total_day_charge）には明白な相関関係があります。「詳細設定」から「項目の自動選択」を選択して分析を実行すると、4 つの属性（昼間料金、顧客サービス・コール数、ボイスメール・プラン、ボイスメールのメッセージ数）（total_day_charge、number customer service calls、voice_mail_plan、number vmail messages）だけが使用されて誤差率は 10.5 % に低下します。total_day_charge が 53.78 以上である 29 人の顧客はすべて解約していることが分かります。

顧客サービスコール回数 (number customer service calls) が多い顧客は解約する傾向が強いと言えます。そのような顧客は複雑な機器を使用したり、信頼できないサービスを受けたりすることでフラストレーションが溜まっていると予想されます。国際化プラン (international plan) を利用する顧客も解約する傾向が強いと言えます。一部の州の顧客は他の州の顧客よりも解約率が高くなっています。たとえば、カリフォルニア (California) とニュージャージー (New Jersey) の解約率が最も高く、バージニア (Virginia) の解約率は最も低くなっています。レコード総数の 2% 以上を占める州を確認するために、「有意水準 % 表示」スライダを一番右側に移動してください。こうすると、ほとんどの州の値が画面から消えます。また、「属性値の順序付け」メニューから「重み順」を選択すると、レコード数が最も多いウエストバージニア州 (West Virginia (WV)) が一番左側に表示されます。リストの一番下にあるほとんどの属性（night_charge など）は、解約の原因を分析するのに効果がありません。分析に最も役立つ属性は total_day_charge です。

車の原産国

車 (*cars*) データセットには、1970 年代から 1980 年代初頭にかけての様々な車種に関する情報が格納されています。これらの情報には、車の重量、加速度、*mpg*（1 ガロン当たりの走行マイル数）などがあります。*cars.eviviz* ファイルには、車の原産国（ヨーロッパ、日本、または米国）を予測するために生成されたエビデンス・クラシファイアの構造図が収録されています。この可視化ファイルは、ラベルを origin (Japan, U.S., Europe) に設定し、*cars.schema* ファイルに対して分析 (Inducer) を適用して作成されたものです。なお、階級自動生成を実行しないで元の値をすべて保持するために、シリンダー (cylinders) 項目は文字列 (string) 型に変換されています。

- Windows システムの場合、これらのファイルは MineSet がインストールされているディレクトリ *\examples\cars.eviviz* および *\data\cars.schema* にあります。
- IRIX システムの場合、これらのファイルは */usr/lib/MineSet/examples/eviviz/cars.eviviz* および */usr/lib/MineSet/data/cars.schema* にあります。

このクラシファイアの目的は、車の原産国を特徴付ける属性をデータセットから見つけ出すことです。右側の円グラフで示されるラベル値の分布を見ると、cars データセット内のほとんど車が米国 (62.5%) で製造され、残りの部分が日本 (20.2%) とヨーロッパ (17.3%) で製造されたことが分かります。各ブランドは 1 つの製造国に対応しているため、製造場所を予想するのにブランドが最適な属性であることは明らかです。したがって、ブランドは最も重要な属性としてリストの先頭に表示されています。スライスの高さを見ると、数多くの車が 4 気筒であること、重量は 3000 lbs 未満であること、13 〜 20 秒間で時速 60 マイルに加速できることなどが分かります。

次に、個々の属性値を表すスライスの分布を調べてみます。エンジンサイズが 169 立方インチを越える車は米国製である確率が非常に高い（日本製ではない）と言えます。また、米国製の車の一般的な特性として、6 気筒または 8 気筒である、mpg が低い、馬力が高い（134 以上）、重量が大きい（2981 lbs 以上）、加速度が高いなどが挙げられます。一方、日本製の車の一般的な特性として、3 気筒または 4 気筒（一部は 6 気筒）である、mpg が高い、エンジンが小さいなどが挙げられます。ラベル確率パネル内で "Europe" をクリックすると、ヨーロッパ製の車のエビデンスを表すバーが表示されます。たとえば、5 気筒の車はヨーロッパ製である確率が高いことが分かります。ただし、スライスの高さを確認すると、5 気筒の車はデータセット内に 3 台しか存在しません。mpg が高いことは、ヨーロッパ製の車を示す的確なエビデンスです。MPG が 41 を越える車は、83% の確率でヨーロッパ製です。ヨーロッパ製の車の MPG が 41 を越える確率は 10.4% ですが、同確率の値が日本製の場合はわずか 2%、米国製の場合は 0% になります。

車の mpg が 40、重量が 3000 lbs という事実だけから、その車の原産国を予想する場合を考えてみます。たとえば、mpg=30.95-41.15 および weightlbs=2981.5+ というスライス（またはバー）を選択してください。右側の確率分布は、米国製が 84%、ヨーロッパ製が 16% と示されます。weightlbs > 2981.5 である日本車は訓練事例内に存在しないため、日本製の確率は 0 になります。ただし、「ラプラス補正の使用」オプションをオン（補正値は 0.5）に設定して分析を再実行すると、米国製が 82%、ヨーロッパ製が 16%、日本製が 2% という異なる結果が得られます。これはラプラス補正によって、円グラフのスライスが完全に 0 になる状況が避けられるためです。重量が 2981 lbs を越える車が日本で製造されないという明白な理由はないため、確率を掛け合わせるときに、日本車を予想する確率が除去されません。

性別の判断

adult データセットには成人の就業者に関する情報が格納されています。このデータセットは米国人口統計局 (U.S. Census Bureau) のデータベースから抽出したものであり、年間の総収入が $100 以上、1 週間の就業時間が 1 時間以上、年齢が 16 歳以上という条件を満たす就業者に関するデータが収録されています。*adult-sex.eviviz* ファイルには、これらの就業者の性別を判断するために作成されたエビデンス・クラシファイアの構造図が収録されています。この可視化ファイルは、ラベルを sex に設定し、（自明なクラシファイアが作成されないように）婚姻関係を表す属性を除去し、*adult.schema* ファイルを分析して生成されたものです。

- Windows システムの場合、これらのファイルは MineSet がインストールされているディレクトリ *\examples\adult-sex.eviviz* および *\data\adult-sex.schema* にあります。

- IRIX システムの場合、これらのファイルは */usr/lib/MineSet/examples/eviviz/adult-sex.eviviz* および */usr/lib/MineSet/data/adult.schema* にあります。

エビデンス・ビジュアライザのラベル確率のパネルを見ると、男性の事前確率が女性の事前確率よりも高いことが分かります。

- 婚姻状態 (marital_status) が性別を判断するのに最も重要な属性です。就業者のうち、既婚の民間人は高い確率で男性です。ただし、配偶者と死別した就業者が女性である確率がそれより高い値になっています。
- 2 番目に重要な属性は業種 (occupation) です。性別との相関が高い職業を見つけてください。男性の占める割合が高い順番で職業を左から右に並べると、軍隊 (Armed forces) (100%)、修理業 (Craft-repair) (95%)、運送業 (Transport-moving) (95%)、農・漁業 (Farming-fishing) (94%) などが男性の比率の非常に高い職業であることが分かります。一方、女性の比率が高い職業は、家政婦 (Private-house-service) (94%) や秘書 (Adm-clerical) (67%) などです。ラベル確率パネル内で「女性 (Female)」の左側のボタンをクリックし、"occupation = Adm-clerical"（業種 = 秘書）という属性値にマウスを移動すると、女性の 23% が秘書として働いていることが分かります。また、就業者の職業が秘書であることを前提条件とすると、その就業者が女性である確率は 67% です。

  職業が秘書 ("Adm-clerical") であり、かつ未亡人 ("widowed") を前提条件として、その就業者が女性である確率を予想してみます。そのためには、選択モードに切替え、該当の属性値（職業が秘書、未亡人）を選択した後、ラベル確率パネル内で「女性 (Female)」の左側のボタンをクリックします。こうすると、ウィンドウの一番上のテキスト行に 95% という確率が表示されます。
- 法人組織の自営業者 (self-employed-incorporated) または個人営業の自営業者 (self-employed-not-incorporated) が男性である確率は高くなっています。一方、州公務員 (state-gov) の就業者が女性である条件付き確率は高いですが、あらかじめ事前確率を考慮に入れれば事後確率はそれほど高いとはいえません。属性値をクリックして、右側の事後確率を確認してみてください。州公務員 (state-gov) を選択すると女性のスライスのサイズが大きくなりますが、州公務員の就業者であるという前提条件だけで、その就業者が女性であるという推論には至りません。

  シーンを回転してグラフの高さを観察すると、ほとんどの就業者が民間企業で働いていることが分かります。
- 総収入 (gross-income) の属性を見ると、就業者の収入が高くなるほど、その就業者が男性である確率が高くなることが分かります。
- 一般的に、教育水準は性別を判断する適切な属性ではありません。ただし、博士号の学位を持っている就業者については、男性である確率が高くなっています。

- 職業が異なると、男性と女性の割合も異なります。
- 人種 (race) の属性の条件付き確率を見ると、黒人の女性就業者の割合が他の人種と比較して高くなっています。ただし、属性値をクリックして事後確率を表示すると、男性と女性の割合がほぼ同じであることが分かります。
- このデータセットでは、男性の週の労働時間 (hours per week) が女性の労働時間よりも長くなっています。

給料を決める要因

次に、成人就業者のデータセットを分析して給料水準を決める要因を確定します。最初に、区間を 10,000、20,000、30,000、60,000 と指定して、連続値の属性である総収入 (gross_income) を 5 つの階級に分割し、新しい離散属性 gross_income_bin を作成します。各レコードにはこの離散属性のいずれかの値が割当てられます。MineSet のクラシファイアを実行すると、給料水準に影響を与える要因を簡単に見つけることができます。*adult-salary.eviviz* ファイルには、この問題について作成されたエビデンス・クラシファイアの構造図が収録されています。この可視化ファイルは、ラベルを gross_income_bin（上記の区間を指定して gross_income を 5 つの階級に分割した属性）に設定し、*adult.schema* ファイルを分析して生成されたものです。

- Windows システムの場合、これらのファイルは MineSet がインストールされているディレクトリ *\examples\adult-salary.eviviz* および *\data\adult.schema* にあります。
- IRIX システムの場合、これらのファイルは */usr/lib/MineSet/examples/eviviz/adult-salary.eviviz* および */usr/lib/MineSet/data/adult.schema* にあります。

エビデンス・ビジュアライザでは、各属性が寄与率の順にランク付けされて表示されます。これを見ると、家族構成 (relationship)、婚姻歴 (arital status)、年齢 (age)、業種 (occupation)、最終学歴 (education)、週の労働時間 (hours per week)、性別 (sex) などが最も重要な属性であることが分かります。ラベル (gross_income_bin) が数値型であるため、連続的なスペクトルに基づく色が各階級に割当てられます（最大値 (60,000+) の階級は赤色になります）。各階級でのラベルは、確率値の大きい順にラベル確率パネル内に並べられます。メイン・ウィンドウ内で属性値をクリックすると、確率が最大であるクラスラベルが一番上に表示されるように、各階級でのラベルの表示順序が調整されます。

- 家族構成 (*relationship*) を調べると、未婚の就業者または家族と同居していない就業者よりも既婚の就業者（夫と妻）の方が多くの給料を稼ぐ傾向のあることが分かります。妻の収入の方が夫の収入より多少多くなっています。

- 婚姻歴 (*arital status*) を調べると、多くの割合の就業者が結婚していることが分かります（左から 2 番目のグラフが高くなっています）。既婚の就業者は未婚の就業者よりも多くの給料を稼いでいます。
- 年齢 (*age*) は非常に重要な要因です。61 歳（ほとんどの人が引退する年齢）までは、就業者の年齢が高くなるに従って、$50,000 以上の給料を稼ぐ確率が増加します。
- 業種 (occupation) が異なると、確率分布も異なります。たとえば、重役や専門技術者は年間収入が $60,000 を越える確率が高くなっています。
- 最終学歴 (*education*) も重要な要因です。教育水準だけを考慮すると、$60,000 以上の給料を稼ぐ確率が最も高いのは、学士以上の学歴を持つ就業者または専門学校を卒業した就業者です。
- 週の労働時間 (hours per week) を調べると、就労時間が長くなるほど、多くの給料を稼ぐ確率が高くなります。
- 性別 (sex) を調べると、女性は年間収入が $60,000 未満になる傾向の強いことが分かります。
- 「有意水準 % 表示」スライダを調節して国籍 (native_country) を取り除くと、重みの低い最終学歴 (education) と業種 (occupation) も同時に取り除かれます。

アヤメのクラス判別

アヤメ (iris) データセットの各レコードには、アヤメの特性を表す 4 つの属性（萼片の縦、萼片の横、花びらの縦、花びらの横 (sepal_length, sepal_width, petal_length, petal_width)）があります。これらの属性に基づいて、アヤメは 3 つのクラス（アイリスセトサ、アイリスバージカラー、アイリスバージニカ (iris-setosa, iris-versicolor, iris-virginica)）に分類されます。この事例の目的は、アヤメの各クラスを特徴付ける属性を確定することです。

クラシファイアを実行する前に、Tool Manager の「クラス判別」タブにある「重要項目」タブをクリックした後、「実行」ボタンをクリックしてください。各属性（petal_width、petal_length、sepal_width、sepal_length）の寄与率がランク付けされたリストが表示されます。これらの属性をスキャタ・ビジュアライザ (Scatter Visualizer) の軸に割当て、アイリスの種類 (iris_type) を色に割当てると、同じ特性を持つクラスタを確認することができます。

iris.eviviz ファイルには、この問題について生成されたエビデンス・クラシファイアの構造図が収録されています。この可視化ファイルは、*iris.schema* ファイルを分析して作成されたものです。

- Windows システムの場合、これらのファイルは MineSet がインストールされているディレクトリ *\examples\iris.eviviz* および *\data\iris.schema* にあります。
- IRIX システムの場合、これらのファイルは */usr/lib/MineSet/examples/eviviz/iris.eviviz* および */usr/lib/MineSet/data/iris.schema* にあります。

エビデンス・ビジュアライザを見ると、アヤメのクラス判別に最適な属性は花びらの縦 (petal_length) と花びらの横 (petal_width) であり、萼片の縦 (sepal_length) と萼片の横 (sepal_width) はあまり効果がないことが分かります。「有意水準 % 表示」スライダを右側にドラッグして、sepal_length と sepal_width が最初に消えることを確認してください。

キノコの分析モデル

mushroom.eviviz ファイルには、キノコの食用性を判定するために生成されたエビデンス・クラシファイアの構造図が収録されています。この可視化ファイルは、*mushroom.schema* ファイルを分析して作成されたものです。

- Windows システムの場合、これらのファイルは MineSet がインストールされているディレクトリ *\examples\muschroom.eviviz* および *\data\mushroom.schema* にあります。

- IRIX システムの場合、これらのファイルは */usr/lib/MineSet/examples/eviviz/mushroom.eviviz* および */usr/lib/MineSet/data/mushroom.schema* にあります。

この事例の目的は、食用になるキノコと毒性のあるキノコを判別することです。このデータセットには 8,000 以上のレコードが含まれているため、分析を実行するのに数秒の時間がかかる場合があります。デフォルト・モードの予備法 (Holdout) でモデルの精度を評価するときは、全レコードの 1/3 が検定用として保持されることに注意してください。

個々のキノコには、傘の色 (cap color)、傷 (bruises)、香り (odor) などの様々な属性があります。エビデンス・ビジュアライザでは各属性が寄与率（分類値の予測に寄与する度合い）の順に並べられます。香り (Odor) と胞子の色 (spore print color) がリストの先頭に表示されています。これらの属性の個々の値は各スライスにランダムに分布しているため、キノコの判別にとって重要な属性とみなされます。このデータセットの全属性は離散的であるため、すべての属性値がラベル確率に従って左から右にソートされます。属性値をアルファベット順または重み順にソートすることもできます。その場合は、「属性値の順序付け」メニューから「アルファベット順」または「重み順」を選択してください。毒性のあるキノコの特徴を確認するには、（メイン画面の右上にある矢印アイコンをクリックするか、Esc キーを押して）ポインタを矢印に変更した後、右側のパネル内で該当のクラスラベルのボタンをクリックします。毒性のキノコを表す属性値は、高いバーによって示されます。

エビデンス・ビジュアライザの「詳細」スライダを右側にドラッグすると、寄与率の低い属性がシーンから削除されます。最も重要な属性は香り (odor) であり、その寄与率は 92 です。その他の属性の寄与率は 48 未満です。ほとんどの属性値がキノコの食用性の判定に適していますが、香りがないキノコ (odor=none) の場合は、食用性の判定が困難になります。エビデンス・ビジュアライザを使用すると、属性自身が重要でない場合でも、特定の属性値の重要性を分析することができます。たとえば、ほとんどの場合、つば下部の柄の色 (stalk_color_below_ring) 属性の値は白であるため、この属性は食用性の判定には適していません（この属性値が白であるキノコのうち、食用可能なキノコと毒性のあるキノコはほとんど同じ数であるため、この属性が白であるという事象に基づいて食用性を判定することはできません）。ただし、stalk_color_below_ring 属性の値が灰色または淡黄色である場合は、食用性の重要な判定基準になります。ただし、これらの属性値を持つキノコはほんの少数です。

政党への帰属

この事例の目的は、主要な票決に関するデータに基づいて特定の下院議員が帰属する政党を判断することです。vote データセットには、議会季刊誌 (*CQA*:*Congressional*

Quarterly Almanac) に記載された 16 回の主要な票決において米国下院の各議員が投票した内容が収録されています。CQA では投票の内容として、賛成 (yes) 3 種類、反対 (no) 3 種類、棄権 (undecided) 3 種類の計 9 種類を規定しています。

クラシファイアを実行する前に、16 回の票決内容を分析して、重要と思われる特性を考えてみてください。続いて、エビデンス・ビジュアライザを実行します。このデータセットについては属性値をアルファベット順にソートし、反対票 (no) が左側、棄権票 (undecided) が中央、賛成票 (yes) が右側に表示されるようにします。

特定の問題に関しては、政党間で見解の違いが如実に表れています。たとえば、民主党員は医療費の凍結とエルサルバドルの支援に賛成する傾向があるのに対して、共和党員は予算決議案の採択とニカラグアの支援に賛成する傾向があります。

移民問題は政党間で見解の違いがあまり見られません。移民問題に関しては 235 票のうち棄権は僅か 7 票であったことから、各議員の関心の深さが伺えます。

vote.eviviz ファイルには、この問題について作成されたエビデンス・クラシファイアの構造図が収録されています。この可視化ファイルは、*vote.schema* ファイルを分析して生成されたものです。

- Windows システムの場合、これらのファイルは MineSet がインストールされているディレクトリ *\examples\vote.eviviz* および *\data\vote.schema* にあります。
- IRIX システムの場合、これらのファイルは */usr/lib/MineSet/examples/eviviz/vote.eviviz* および */usr/lib/MineSet/data/vote.schema* にあります。

乳癌の診断

breast データセットには、乳癌の診断を受けた女性被験者に関する情報が収録されています。このデータセットの各レコードは、細胞の大きさ、細菌塊の厚さ、周辺吸着などの属性を持つ被験者を表します。この事例の目的は、被験者の腫瘍が悪性であるか良性であるかを判断することです。*breast.eviviz* ファイルには、この問題について作成されたエビデンス・クラシファイアの構造図が収録されています。この可視化ファイルは、*breast.schema* ファイルを分析して生成されたものです。

- Windows システムの場合、これらのファイルは MineSet がインストールされているディレクトリ *\examples\breast.eviviz* および *\data\breast.schema* にあります。
- IRIX システムの場合、これらのファイルは */usr/lib/MineSet/examples/eviviz/breast.eviviz* および */usr/lib/MineSet/data/breast.schema* にあります。

エビデンス・ビジュアライザを見ると、サンプル・コード番号 (sample_code_number) 属性が等分割された単一の階級になっていることが分かります。これは腫瘍の悪性 / 良性を判断するのに sample_code_number 属性が役立たないことを意味します。

甲状腺機能低下症 (Hypothyroid) の診断

甲状腺機能低下症 (hypothyroid) データセットの構造は、上記の breast（乳癌）データセットの構造と似ています。*hypothyroid.eviviz* ファイルには、甲状腺機能低下症を診断するために作成されたエビデンス・クラシファイアの構造図が収録されています。この可視化ファイルは、*hypothyroid.schema* ファイルを分析して作成されたものです。

- Windows システムの場合、これらのファイルは MineSet がインストールされているディレクトリ *\examples\hypothyroid.eviviz* および *\data\hypothyroid.schema* にあります。
- IRIX システムの場合、これらのファイルは */usr/lib/MineSet/examples/eviviz/hypothyroid.eviviz* および */usr/lib/MineSet/data/hypothyroid.schema* にあります。

Thypothyroid データセットには 3163 件のレコードが収録されていますが、ほとんど (95.45%) の被験者は陰性です。つまり甲状腺機能低下症にかかっていません。したがって、特定の被験者を陰性と診断することは高い確率で正解になりますが、最も重要なのは実際に甲状腺機能低下症にかかっている陽性の被験者です。陽性の被験者を陰性と診断するのは重大な問題です。

6.35 〜 27.5 の "tsh" を表すスライスを見ると、この属性が甲状腺機能低下症の診断にとって重要なエビデンスであることが分かります。ただし、陰性 (negative) の事前確率が非常に高いため、この属性値をクリックしても、事後確率のスライスは陰性 (negative) と表示されます。

このような場合は、陽性の被験者が陰性と診断されるのを避けるために、陽性と判定される事後確率が高くなるように損失マトリックスを調整する必要があります。実際に病気の人を健康であると診断すると大きな損失が発生しますが、実際に健康な人を病気であると診断しても、その人が精密検査や不要な治療を受けるだけで済みます。

エビデンス・ビジュアライザを見ると、"fti" が非常に重要な属性であることが分かります。NULL 値の右側にある最初の 2 つの範囲は、甲状腺機能低下症の診断にとって重要なエビデンスになります。

ピマ族における糖尿病の診断

pima データセットには、アリゾナ州フェニックスのアメリカ・インディアン（ピマ族）から収集した糖尿病診断に関する統計データが収録されています。この事例の目的は、年齢、体重、血圧、血糖値などの医学データに基づいて、被験者が糖尿病かどうかを診断することです。

pima.eviviz ファイルには、この問題について作成されたエビデンス・クラシファイアの構造図が収録されています。この可視化ファイルは、*pima.schema* ファイルを分析して作成されたものです。

- Windows システムの場合、これらのファイルは MineSet がインストールされているディレクトリ *\examples\pima.eviviz* および *\data\pima.schema* にあります。
- IRIX システムの場合、これらのファイルは */usr/lib/MineSet/examples/eviviz/pima.eviviz* および */usr/lib/MineSet/data/pima.schema* にあります。

エビデンス・ビジュアライザを見ると、ほとんどの属性が単独では糖尿病の診断に有効でないことが分かります。血糖値 (plasma_glucose) が高い場合は、糖尿病である確率が高くなります。妊娠期間 (pregnancy) が 6 カ月以上である場合、または年齢 (age) が 27 歳以上である場合も、糖尿病である確率が高くなっています。

DNA 境界

dna.eviviz ファイルには、この問題について作成されたエビデンス・クラシファイアの構造図が収録されています。この可視化ファイルは、*dna.schema* ファイルを分析して作成されたものです。

- Windows システムの場合、これらのファイルは MineSet がインストールされているディレクトリ *\examples\dna.eviviz* および *\data\dna.schema* にあります。

- IRIX システムの場合、これらのファイルは */usr/lib/MineSet/examples/eviviz/dna.eviviz* および */usr/lib/MineSet/data/dna.schema* にあります。

DNA データセットには、分子生物学分野の GenBank 64.1 (*genbank.bio.net*) から抽出された 3,186 件の DNA レコードが収録されています。DNA 連鎖上の点のうち、蛋白質の生成中に余分な DNA が除去される箇所は組継ぎ接合点 (Splice junction) と呼ばれます。この事例の目的は、DNA の exon/intron 境界（EI サイト）、intron/exon 境界（IE サイト）、無境界 (none) を識別することです。IE 境界は " アクセプタ " と呼ばれ、EI 境界は " ドナー " と呼ばれることがあります。DNA データセットの各属性によって、60 種類のヌクレオチドが表されます。これらの属性が 2 分割されるため、接合点の両側にそれぞれ 30 種類のヌクレオチドが存在することになります。

エビデンス・ビジュアライザを見ると、接合点の中心付近の属性が寄与率の高い属性として選択されていることが分かります。接合点から離れた属性は寄与率が低くなっています。

左側のウィンドウ内で "left_01: G" および "left_02: A" を表すグラフをクリックして選択すると、右側のラベル確率パネル内の円グラフが更新され、エビデンス・クラシファイアによって予想された各階級の確率分布が表示されます。この円グラフを見ると、エビデンス・クラシファイアによって予想された確率は、高い値から低い値の順に、"intron/exon"、"exon/intron"、"none" となっていることが分かります。

「項目の自動選択」オプションを使用するとモデルの精度が多少改善されますが、実行時間が非常に長く , 時には数時間に及ぶ場合があります。そのような場合は、「項目の自動選択」機能を一度実行した後、選択された項目（属性）だけを使用してデータマイニングを続けてください。

マップ・ビジュアライザ用のサンプルファイル

マップ・ビジュアライザの機能や特長を紹介するために、設定ファイル (.mapviz) とデータファイル (.data) のサンプルファイルが用意されています。

- Windows システムの場合、これらのファイルは MineSet がインストールされているディレクトリ *\examples\mapviz* にあります。また、*.gfx* ファイルと *.hierarchy* ファイルのサンプルファイルは *\config\mapviz* にあります。
- IRIX システム上では、これらのファイルは */usr/lib/MineSet/examples/mapviz* ディレクトリにあります。また、*.gfx* ファイルと *.hierarchy* ファイルのサンプルファイルは */usr/lib/MineSet/mapviz/gfx_files* ディレクトリにあります。

- *blocks.mapviz*、*blocks.data*、*blocks.gfx*、*blocks.hierarchy.*
  これらの単純なサンプルは 4 つの隣接ブロックを表しています。各ブロックの高さと色は、*blocks.data* ファイル内の元データに応じて変わります。マウスの中ボタンを使用してドリルアップを行うと、上位の 2 つのブロックと下位の 2 つのブロックを集計した値が表示されます。もう一度ドリルアップを行うと、これらの上位ブロックと下位ブロックが単一のブロックに集計処理されます。マウスの右ボタンをクリックしてドリルダウンを行うと、詳細レベルのブロックが再表示されます。

- *population.australia.mapviz*、*population.australia.data*、*australia.states.gfx*、*australia.states.hierarchy*
  データファイル (population.australia.data) には、オーストラリアの 1 つの州と準州ごとに 1 行のデータが入っています。ファイル各行は、タブで区切られた 3 つの項目（州または準州のキーワード名、人口、面積）からなります。

  このサンプルでは、オーストラリアの州と準州における 1991 年の人口と人口密度がグラフィカルに表示されます。オブジェクトの高さは人口を表し、色は人口密度を表します。画面の下方にある説明には、色の範囲と各範囲の値が示されます。

- *population.canada.mapviz*、*population.canada.data*、*canada.provinces.gfx*、*and canada.provinces.hierarchy*
  データファイル (population.canada.data) には、カナダの 1 つの州と準州ごとに 1 行のデータが入っています。ファイルの各行は、ブランクで区切られた 13 個の値（1871 ～ 1991 年の 10 年ごとに 1 個の値）からなります。

  このサンプルでは、カナダの州と準州における 1871 ～ 1991 年の人口と人口密度が 10 年単位でグラフィカルに表示されます。アニメーション・コントロール・パネルを使用すると、時間の経過に従ってデータセットを動的に表示することができます。アニメーションの操作については、「アニメーション・コントロール・パネル」（11 ページ）を参照してください。

- *population.europe.mapviz*、*population.europe.data*、*europe.countries.hierarchy*、*europe.countries.gfx*
  このサンプルでは、西欧各国と中欧各国の 1992 年の人口と人口密度がグラフィカルに表示されます。

- *population.usa.mapviz*、*population.usa.data*、*usa.state.gfx*、*usa.state.hierarchy*
  このサンプルでは、米国の 1770 ～ 1990 年の人口と人口密度がグラフィカルに表示されます。アニメーション・コントロール・パネルを使用すると、人口と人口密度の時系列変化を動的に表示することができます。

- *population.usa.city.mapviz*、*population.usa.city.data*、*usa.state.gfx*、*usa.state.hierarchy*、*usa.city.gfx*、*usa.city.hierarchy*

usa.state.gfx は米国の各州の形状と位置を示すファイルであり、シーンの背景として使用されます。*usa.city.gfx* は、その背景上に表示される各都市の位置を示すファイルです。*.data* ファイルには各都市の人口データが収録されています。

このサンプルでは、米国の主要 48 都市における 1950 ～ 1990 年の人口がグラフィカルに表示されます。色にはデータが割当てられていません。アニメーション・コントロール・パネルを使用すると、人口の時系列変化を動的に表示することができます。

- *perhouse.perage.mapviz*、*perhouse.perage.data*、*usa.state.gfx*、*usa.state.hierarchy*
  このサンプルでは、1988 年 7-8 月から 1991 年 5-6 月までの家計支出データがグラフィカルに表示されます。このデータには、時間および年齢という 2 つの独立次元があります。男女別の家計構成員に別々の色が割当てられています。オブジェクトの高さは、特定の期間と年齢グループについて 1 家計当たりの平均支出額（ドル単位）を表します。サマリ・ウィンドウ内では、最高の支出額が一番濃い色の領域で示されます。最高の支出額は、「1989 年 5-6 月（年齢：30-39）」と「1990 年 5-6 月（年齢：30-39）」です。

- *telecom.mapviz*、*telecom.data*、*usa.city.lines.gfx*、*usa.city.lines.hierarchy*、*usa.state.gfx*、*usa.state.hierarchy*
  このサンプルでは、米国の各都市における電話会社ごとの通話データがアーチ付きの平面地図としてグラフィカルに表示されます。個々のアーチは、電話の発信点と受信点を結んでいます。各アーチの幅と色は変更することができます。このサンプルではアーチの幅と色がランダムになっていますが、発信点と受信点間の回線の料金と通話時間を幅または色にマッピングすることができます。

- *fasta.m.data*、*fasta.m.mapviz*、*fasta.m.gfx*、*fasta.m.hierarchy*

  データファイル (fasta.m.data) には、2 つの完全なゲノムの発生過程を生物学的に比較した解析結果の一部が入っています（このデータは European Bioinformatics Institute の Dr. Tom Flores から提供されたものです）。このデータをグラフィカルに表示すると、2 つのゲノムの類似領域を簡単に識別することができます。このように大量の情報をビジュアルに表示する手法を応用すれば、個々のゲノムに関するより多くのデータを表示することができます。その結果、データの探索が容易になり、生物学的発生の仕組みを効率的に解析できるようになります。

  このサンプルで "map" と表示されるのは、*Mycoplasma genitalium* (MG) と呼ばれる生物学的有機組織を持つ円形のゲノムです。MG ゲノムは 500 個の同じサイズのセグメントに分割されます。各セグメントは、ゲノム内部の 1000 個のヌクレオチド列を表します。スライダをドラッグすると、*Haemophilus influenzae* (HI) と呼ばれる 2 番目のゲノムのセグメントが選択され、2 つのゲノ

ム（MG と HI）を相互比較することができます。アニメーション・コントロール・パネルのサマリ・ウィンドウには最大の類似性を示すセグメントが表示され、スライダを使用して、そのセグメントを詳しく調べることができます。"map" 上のバーの高さと色は、各 MG セグメントと各 HI セグメントの相対的な類似性を示しています（バーが高ければ、類似性が大きいことになります）。類似性は 0.0 ～ 1.0 の値を取る "Reciprocal Evalues"（相互 E 値）によって測定されます。

選択式決定木用のサンプルファイル

ここでは、選択式決定木の適用が有効と思われる事例を紹介します。MineSet では、これらの事例で使用するサンプルファイルが用意されています。分析すると、下記で説明する *-odt.treeviz* ファイルが生成されます。各事例の内容と目的については、「ツリー・ビジュアライザ用のサンプルファイル」を参照してください。ここでは選択式決定木を適用するときの利点と欠点について説明します。

注記：データ・ファイル（拡張子 *.schema*）はクライアント・ワークステーションの *data* ディレクトリにあります。クラシファイアの可視化ファイルは（拡張子 *-odt.treeviz*）はクライアント・ワークステーションの *examples* ディレクトリにあります。スキーマファイル（*.schema*）を開くと、それに対応するデータファイル（.data）が自動的に読込まれます。

- Windows システムの場合、これらのファイルは MineSet がインストールされているディレクトリ *\examples* および *\data* にあります。
- IRIX システムの場合、これらのファイルは */usr/lib/MineSet/examples/treeviz* および */usr/lib/MineSet/data* にあります。

解約 (Churn)

解約 (churn) データセットの選択式決定木を見ると、昼間の合計課金額 (total_day_charge)、昼間の合計利用分数 (total_day_minutes)、顧客サービス通話回数 (customer_service_calls) の各属性がルートノードの選択肢として適していることが分かります。これらの属性の推定誤差率はほとんど同じであるため、各ユーザ独自の判断に基づいて特定のサブツリーをたどることができます。右側のサブツリーは顧客サービス通話回数 (customer_service_calls) から始まっていますが、2 番目のテストは（ルートの左側の選択肢である）昼間の合計課金額 (total_day_charge) または昼間の合計利用分数 (total_day_minutes) に基づいて行われていることに注意してください。ただし、1 つの属性に基づいて既に分割が発生しているため、区間は異なっています。

車の原産国

車 (cars) データセットの選択式決定木を見ると、立方インチ (cubic inches)、シリンダ数 (cylinders)、重量 (weight lbs)、mpg、ブランド (brand) などの複数の属性がルートノードの選択肢として適していることが分かります。ルートの推定誤差率は子ノードの推定誤差率よりも低くなっていることに注意してください。

アヤメのクラス判別

この事例では、決定木のルートの誤差率が 6%、選択式決定木のルートの誤差率が 8% であるため、選択式決定木の予測精度が決定木より劣っているように思われます。ただし、誤差率を推定するときは次のような点に注意してください。

- 推定誤差率の標準偏差がそれぞれ 3.88% および 3.39% とかなり高くなっています。統計学の基本的な規則として、差異が標準偏差の 2 倍未満であれば、その差異は統計的に 95% の信頼水準で有意ではないと判断されます。2% という差異は 1 つの標準偏差にも達していないため、2 つのクラシファイアの誤差率は統計的に 95% の信頼水準で同じであると判断されます。
- 小さいファイル（iris データセットのレコード数は 150）の場合は、ランダム・シードが異なると予測結果も異なります。たとえば、ランダム・シードを 3 に変更すると、選択式決定木クラシファイアの誤差率は 8% から 4% に改善されますが、決定木クラシファイアの誤差率は変わりません（後で忘れずにランダム・シードをリセットしてください）。これは正確なクラシファイアが生成されたのではなく、単に推定誤差率が不安定であることを意味します。検定で使用されるレコードは 50 個だけであるため、レコード 1 つ分の誤差は 2% になります。4% と 8% の差異 (4%) は、レコード 2 つ分の誤差に相当します。

- 小さいファイル（iris データセットのレコード数は 150）については、MineSet の「誤差推定」オプションを使用してください。このオプションを使用すると、信頼区間が狭い良好な推定値が得られます。ステータス・ウィンドウに表示される推定誤差率は、決定木クラシファイアの場合は 4.67% +/- 1.73%、選択式決定木クラシファイアの場合は 4.00% +/- 1.61% です。この差異は有意ではありませんが、選択式決定木の方が多少優れています。
- 選択式決定木の誤差率が高い場合でも、通常、確率を予想するには選択式決定木の方が適しています。iris データセットについては、平均二乗誤差の推定値が決定木の場合は 3.94 であるのに対して、選択式決定木の場合は 3.67 です（ただし、この差異は 95% の信頼水準で有意ではありません）。

キノコの分析モデル

キノコ (mushroom) データセットの選択式決定木を見ると、ルートノードで選択された 5 つの選択肢の推定誤差率がいずれもゼロであることが分かります。予測結果を分析すると、左側の選択肢（傷 (bruises) の属性）が香り（odor）の属性（作成された決定木のルートの属性）よりも正確で分かり易いように思われます。香り（odor）の属性を除外して通常の決定木を構築すると、同じくらい正確な（推定誤差率がゼロの）クラシファイアが得られます。

ただし、ルートの選択肢を除外して決定木を構築したときに、選択式決定木に表示されるクラシファイアと同じ精度のモデルが生成されるとは限らないことに注意してください。除外された属性がツリーの下位レベルで使用されていた可能性があります。たとえば、cars データセットからブランド (brand) 属性を除外すると、ルートの 5 つの選択肢のうち 4 つの選択肢でその属性が使用されていない場合でも、誤差率が大幅に増加します。

政党への帰属

政党 (vote) データセットの特性はアヤメ (iris) のデータセットとよく似ています。選択式決定木分析の誤差率は決定木の誤差率と同じです。「誤差推定」オプションの相互検証 による評価では、誤差率と平均二乗誤差の観点から選択式決定木の方が決定木より多少正確であると判断されます（ただし、95% の信頼水準で有意ではありません）。

乳癌の診断

「クラシファイアとエラー」モードと「誤差推定」モードでは選択式決定木の誤差率が決定木 (*Estimate Error*) の誤差率より多少低くなりますが、その差異は 95% の信頼水準で有意ではありません。

甲状腺機能低下症 (Hypothyroid) の診断

甲状腺機能低下症 (hypothyroid) データセットの誤差率は非常に低くなっていますが（1% 未満）、これは被験者の大部分 (95%) が陰性である（甲状腺機能低下症に罹っていない）ためです。実際には陽性の被験者を陰性と診断する誤りに 100 というペナルティを課す損失マトリックスを作成すると、選択式決定木の損失額（合計 182; 1 レコードあたり 0.17）は決定木の損失額（合計 523; 1 レコードあたり 0.5）よりかなり低いことが分かります。この差異は 95% の信頼水準で有意です。

DNA 境界

DNA データセットについては、選択式決定木の方が決定木よりも多少正確ですが、ルートの選択肢を見ると、左 (left) 1,2 および右 (right) 1,2,5 が選択されていることが分かります。境界に近い属性の方が重要であるという予備知識から判断すると、右 (*right*) 5 での分割を避けた方が適切です。最大 # ルート選択肢数をデフォルト値の 5 から 4 に変更すると、誤差率が 5.65% から 6.59% に増加します。right 5 がルートの選択肢として使用されていないことを考えると、これは意外に思われるかも知れません。選択肢の最大数を 4 に変更するもう 1 つの効果として、減少数 (Decrease) が 2 であるために、下位のツリーに表示される選択肢の数が減少します。このため、残り 4 つのサブツリーの誤差率が増加します。ただし、選択式決定木の誤差率の方が決定木 (Decision Tree) の誤差率（7.06 ± 0.79%）よりもかなり低くなります（その差異は 95% 信頼区間で有意です）。

回帰ツリー用のサンプルファイル

ここでは、回帰ツリー分析の特徴や機能に焦点を当てながら、回帰ツリーの適用が有効と思われる事例を紹介します。MineSet では、これらの事例で使用するサンプルファイルが用意されています。回帰ツリー分析を行うと、次に説明する *-rt.regress* ファイルが生成されます。

注記：データファイル（拡張子 *.schema* ）はクライアント・ワークステーションの *data* ディレクトリにあります。回帰モデルの可視化ファイルは（拡張子 *-rt.treeviz*）はクライアント・ワークステーションの *examples* ディレクトリにあります。スキーマファイル（*.schema*）を開くと、それに対応するデータファイル（*.data*）が自動的に読込まれます。

- Windows システムの場合、これらのファイルは MineSet がインストールされているディレクトリ *\examples* および *\data* にあります。
- IRIX システムの場合、これらのファイルは */usr/lib/MineSet/examples/treeviz* および */usr/lib/MineSet/data* にあります。

解約（Churn）

「解約 (*churn*)」データセットには、電話会社の顧客の通話記録から収集された情報が格納されています。分析モデルの例では、このデータセットを使用して、顧客が他の競合会社に乗り換える原因を解析しました。この回帰モデルの例では、各顧客の 1 日あたりの合計課金額に影響を与える要因を調べます。

可視化ファイル *churn-rt.treeviz* には、昼間料金（total_day_charge）を予測するために「解約 (churn)」データセットから生成された回帰ツリー (Regression Tree) の構造図が収録されています。興味深いことに、回帰ツリーは単一の属性「昼間通話時間 (total_day_minutes)」に基づいて継続的に分割され、この属性が非常に小さい範囲に細分化されています。これは「昼間通話時間 (total_day_minutes)」と「昼間の料金 (total_day_charge)」の間に極めて密接な関係があるためです。すなわち、各顧客の課金額は電話を利用した分数のみに基づいて決定されます。回帰ツリーでは、このような関係を即座に検出して予測に利用することができます。

車の燃費効率

「車 (*cars*)」データセットには、1970 年代から 1980 年代初頭にかけての様々な車種に関する情報が格納されています。これらの情報には、車の重量、速度、mpg（1 ガロン当たりの走行距離 = 燃費効率）などがあります。可視化ファイル *cars-rt.treeviz* には、mpg (miles per gallon) を連続型のラベルとして指定し、「車 (cars)」データセットから作成された回帰ツリーの構造図が収録されています。

最上位ノードをクリックすると、このデータセット内の車の平均 mpg は約 23.5 であることが示されます。回帰ツリーの最初の分割を見ると、mpg に影響を与える最も重要な要因は車の重量であることが分かります。すなわち、重い車ほど mpg が低いという事実が回帰ツリーによって検出されています。この事実は観察者にとっては自明です

が、通常の自動推論モデルでは検出することが困難です。ベースノードの 2 つの子ノードを見ると、右側の子ノードの方が左側の子ノードよりも青いこと（mpg が低いこと）が分かります。ノードを強調表示すると、重量が 3018 lbs 未満の車の mpg は約 28.3 であり、重量が 3018 lbs 以上の車の mpg は約 16.6 であることが示されます。

重い車を調べてみると、2 番目のレベルの分割は車の馬力 (horsepower) に基づいており、馬力の大きい車ほど mpg が低くなる傾向が見られます。3 番目のレベルの分割は車の製造年に基づいており、新しい車ほど mpg が高くなる傾向が見られます。ここで、異常値を示す例外的な車を探してみましょう。「フィルタ (Filter)」パネルを使用して、"mean mpg < 24" かつ "maximum > 30" という条件に合致するノードを検索します。このフィルタを実行すると、ツリーが単一のノード（重量が 3018 lbs 未満、馬力が 77 以上、製造年が 1980 年以前という全条件を満たす車）に縮小されます。このノードに属する車は例外的です。このノードの一番右側のバーをクリックし、「選択」メニューから「オリジナルデータの表示」を選択してノード情報を Tool Manager に送信すると、その車は 1978 年製の Dodge（重量は約 2000 lbs、83 馬力であるが mpg は 33.5 とかなり高い値）であることが分かります。

給料を決める要因

「成人 (*adult*)」データセットには成人の就業者に関する情報が格納されています。このデータセットは米国人口統計局 (U.S. Census Bureau) のデータベースから抽出したものであり、年間の総収入が $100 以上、週の労働時間が 1 時間以上、年齢が 16 歳以上という条件を満たす就業者に関するデータが収録されています。回帰ツリー分析を使用すると、就業者の給料に影響を与える要因を確定できるほかに、他の属性情報に基づいて就業者の給料水準を大まかに予測することができます。

可視化ファイル *adult-rt.treeviz* には、「総収入 (gross_income)」を連続型のラベルとして指定し、「成人 (adult)」データセットから作成された回帰ツリー (Regression Tree) の構造図が収録されています。このデータセットは大きいため（レコード数が約 50,000 件）、ワークステーション上で回帰モデルを生成するのに数分の時間がかかる場合があることに注意してください。

最上位ノードにある縦バーは、「成人 (adult)」データセット内の給料のヒストグラム（度数分布）を表します。利用可能なデータ量は給料水準が高くなるに従って減少することに注意してください。たとえば、年収が $3,000 付近の就業者に関するデータは大量に存在しますが、給料水準が高くなるに従ってデータ量（就業者数）は減少します。ただし、この傾向はヒストグラムの最後の部分で逆転しており、年収が $100,000 付近の就業者に関するデータ量がかなり多くなっています。この矛盾がデータの実際のトレンドに基づくものであるのか、偏ったサンプリングの結果であるのかは不明です。

回帰ツリーは最初に「年齢 (age)」属性に基づいて分割されています。予想通り、若年層の給料は熟年層の給料よりも低くなっています。最上位ノードとその 2 つの子ノードにマウスのカーソルを移動すると、それら 3 つの就業者グループに関する統計情報が表示されます。これらのグループ全体の平均給料は約 $33,500 ですが、27 歳未満の就業者の平均給料は約 $14,300、27 歳以上の就業者の平均給料は約 $40,000 であることに注意してください。

27 歳未満の就業者に対する次の 2 回の分割を見ると、それらの就業者が 23 歳以下と 24 歳以上の 2 つのグループにさらに分割されていることが分かります。興味深いことに、これら 2 回の分割以降は、「週の労働時間 (hours_per_week)」属性に基づいて同じ分割が行われており、両方の年齢グループとも、就業時間が長くなるに従って総収入金額が増えています。

次に、27 歳以上の就業者に焦点を当てると、教育水準に基づいてツリーが即座に分割されていることが分かります。「教育番号 (education_num)」が 13（学士に相当する番号）以上である就業者は給料が高い傾向があります。「教育番号 (education_num)」で分割された 2 つの子ノードを見ると、$90,000 以上の給料を稼ぐほとんどの就業者は高等教育を受けていることが分かります。

「フィルタ」パネルを使用すると、平均年収が $50,000 以上であるカテゴリを簡単に検索することができます。「フィルタ (Filter)」パネルで "mean > 50000" を選択すると、$50,000 以上を稼ぐ就業者は極めて少ないため、最上位レベルのノードが「フィルタ」パネルから消えます。学士の学歴を持つ 27 歳以上の就業者が "mean > 50000" のカテゴリに属します。最初の分割の左側にある枝を最後までたどると、このカテゴリ内に別のグループ（1 週間の就業時間が 35 時間以上で 10 年以上の教育を受けた 36 歳以上の既婚男性）が存在することが分かります。

「フィルタ」パネルをもう一度使用し、絶対偏差が $25,000 以上であるノードを検索すると、各就業者の年収の格差が最も大きいカテゴリが検出されます。この「フィルタ」パネルに残る最初のノードは、学士の学歴を持つ 27 歳以上の就業者です。このノードのヒストグラムを見ると、平均値付近に分布が集中していますが、かなりの人数の就業者が年収 $100,000 程度の給料を稼いでいます。

アヤメ (Iris) の属性値

「アヤメ (iris)」データセットの各レコードには、アヤメの特性を表す 5 つの属性（「花びらの横 (petal_width)」、「花びらの縦 (petal_length)」、「蔕片の横 (sepal_width)」、「蔕片の縦 (sepal_length)」、「アヤメの種類 (iris type)」）が格納されています。この回帰ツリーの目的は、他の属性に基づいて「花びらの横 (petal_width)」を予測することです。可視化ファイル *iris-rt.treeviz* には、花びらの横を予測するために「アヤメ (iris)」データセットから作成された回帰ツリーの構造図が収録されています。

最上位ノードを見ると、「花びらの横 (petal_width)」の値にギャップ（標本が存在しない箇所）のあることが分かります。「アヤメ (iris)」データセットの回帰ツリーでは、最初に「花びらの縦 (petal_length)」に基づいて分割が行われています。「花びらの縦 (petal_length)」が 2.6 未満である場合は、「花びらの横 (petal_width)」の取り得る値が非常に少なくなります。一方、「花びらの縦 (petal_length)」が 2.6 以上である場合は、「花びらの横 (petal_width)」の値が大きくなり、一様な分布が観察されます。「花びらの縦 (petal_length)」が 2.6 未満であるアヤメの「花びらの横 (petal_width)」の平均値は 0.24 ですが、「花びらの縦 (petal_length)」が 2.6 以上であるアヤメの「花びらの横 (petal_width)」の平均値は 1.68 です。「花びらの縦 (petal_length)」が 2.6 以上であるアヤメの枝をたどっていくと、"petal_length = 4.85" という条件でツリーがさらに分割されていることが分かります。同じ属性に基づく 2 回の連続した分割は、「花びらの横 (petal_width)」と「花びらの縦 (petal_length)」の間に限定的な関数関係があることを示唆しています。

「花びらの縦 (petal_length)」が 2.6 未満であるアヤメの枝を調べると、「蔕片の横 (sepal_width)」に基づいて 2 回目の分割が行われていることが分かります。興味深いことに、ツリーのこの部分では「花びらの横 (petal_width)」の分布が散らばっており、「蔕片の横 (sepal_width)」が 3.25 未満であるアヤメの「花びらの横 (petal_width)」の値は 3 つの分離した狭い範囲に分布しています。

ピマ族における糖尿病の診断

「糖尿病 (pima)」データセットには、アリゾナ州フェニックスのアメリカ・インディアン（ピマ族）から収集した糖尿病診断に関する統計データが収録されています。可視化ファイル *pima-rt.treeviz* には、「血糖値水準 (plasma_glucose_level)」を連続型のラベルとして指定し、「糖尿病 (pima)」データセットから作成された回帰ツリー (Regression Tree) の構造図が収録されています。

この回帰ツリーは最初に「糖尿病症状 (diabetes indicator)」に基づいて分割されており、糖尿病症状が陽性である被験者は陰性の被験者よりも血糖値水準が高い傾向（141 対 110）が見られます。2 回目の分割は「2 時間血清インスリン (2-hour serum insulin)」に基づいており、この属性値が 125 を越える被験者は血糖値水準が高くなっています。

回帰ツリーを使用して予測を行うときは、新しいレコードを検査しながらツリー内の決定木ノードを上から下にたどります。たとえば、「2 時間血清インスリン (2-hour serum insulin)」が 110 である被験者は、血糖値水準の予測値が 105 になります。

スキャタ・ビジュアライザ用のサンプルファイル

スキャタ・ビジュアライザの機能や特長を紹介するために、下記の説明の通り、スキャタ・ビジュアライザ用のサンプルファイルが用意されています。*.data* と *.scatterviz* の各サンプルファイルは *examples* ディレクトリ内にあります。スキーマファイル (*.schema*) を開くと、それに対応するデータファイル (*.data*) が自動的に読込まれます。

- Windows システムの場合、これらのファイルは MineSet がインストールされているディレクトリ *\examples* にあります。
- IRIX システムの場合、これらのファイルは */usr/lib/MineSet/examples/scatterviz* にあります。

スキャタ・ビジュアライザのサンプルファイルには、次のものがあります。

- *company.data*
  このファイルには、複数の保険会社の架空の契約高データが入っています。これらの保険会社が扱っている製品は、生命保険、自動車保険、住宅保険の 3 種類です。データは 1 年単位で 10 年分あり、保険被契約者は 5 つの所得層（年収）に分類されています。
- *company.scatterviz*
  このファイル内の指定によって、年度が 1 つのスライダ割当てられ、所得層がもう 1 つのスライダに割当てられています。スキャタ・ビジュアライザでは、生命保険、自動車保険、住宅保険の契約高が 3D ランドスケープで表示されます。サマリ・ウィンドウ内の色の濃度は、全保険会社の全種類の保険の総契約高を表しています。
- *company-total.scatterviz*
  このファイルは *company.scatterviz* と同じですが、各保険会社を表すオブジェクトのサイズがその会社の全種類の保険の総契約高によって決まる点が異なります。

- *company-life.scatterviz*
  このファイルは *company.scatterviz* と同じですが、各保険会社を表すオブジェクトの色が総契約高中に占める生命保険の割合を示している点が異なります。

- *store-type.data* と *store-type.scatterviz*
  これらのファイルは、店舗の種類ごとに、各種製品の 3 年間にわたる売上高を示しています。スライダにマッピングされる独立変数は時間だけです。各要素は店舗の種類（食品店、ドラッグストア、ガソリンスタンドなど）を示しています。データファイル (store-type.data) には、各種類の店舗について複数の種類の製品群（アルコール飲料、タバコなど）の総売上高が入っています。データは 1 カ月単位で 36 カ月分あります。

  設定ファイル (store-type.scatterviz) では、時間（月）が単一スライダとして使用されています。軸の 1 つはアルコール飲料の売上高を表し、もう 1 つの軸はタバコの売上高を表しています。3 番目の軸は使用されていません。

- *brand.data* と *brand.scatterviz*
  これらのファイルは、各種店舗における各種ブランドのソフトドリンクの売上高を示しています。このデータセットでは、製品ブランドが要素となり、店舗の種類が軸に割当てられています。総売上高は各製品ブランドのサイズに割当てられています。色のマッピングはランダムです。独立変数は存在しないため、スライダも存在しません。

- *cars.data* と *cars.scatterviz*
  これらのファイルは、複数の型の自動車について、その重量 (weight)、馬力 (horsepower)、年型 (year)、加速 (acceleration) を示しています。軸に割当てられた属性は、立方インチ (cubic_inches)、mpg、加速度（時速 60 マイルになるまでの秒数）です。重要はサイズに割当てられています。

- *people.data* と *people.scatterviz*
  これらのファイルは、特定の人口標本の身長 (height)、体重 (weight)、コレステロール (cholesterol) 値を示しています。

- *nl.births.data* と *nl.births.scatterviz*
  これらのファイルは、オランダの出生パターンを示しています。各地域について、人口密度、出生率、人口が示されています。アニメーションのスライダは、母親の年齢と年度にマッピングされています。

- *adult94.data* と *adult94.scatterviz*
  これらのファイルは、スキャタ・ビジュアライザを adult.data ファイルに適用した場合の複雑な例を示しています。この例では、「週の平均労働時間 (avg_hrswk)」、「平均総収入 (avg_gross_income)」、「平均教育番号 (avg_education_num)」が 3 つの軸にマッピングされています。ただし、「教育番号 (education_num)」は正確な教育年数ではなく、近似値です。右側のスライダは各年齢層を示しています。職業、人

種、性別をグループ化することによって、集計処理値が作成されています。すなわち、3 つの属性値のあらゆる組合わせに対して要素が存在することになります。説明に示されているように色は各種の職業を表し、各要素のサイズはレコードの数を表します。サマリスライダにもデータの密度に応じた色が表示されます。このビジュアル図の作成過程を知りたい場合は、「ファイル」メニューの「Tool Manager の起動」オプションを選択してください。Tool Manager が起動し、このサンプルを作成したセッションが示されます。

シーンの初期画面では、20 歳未満の人の情報が表示されます。週の平均労働時間（約 14 時間）と平均収入（約 4,000 ドル）はかなり低い値になっています。スライダを使用して各年齢層に移動し、（メイン・ウィンドウの右にあるボタンのうち、下の 3 つのボタンを使用して）3 つの正射投影ビューにシーンを表示すれば、さまざまな傾向を観察することができます。たとえば、週の平均労働時間だけを表示するようにシーンを設定すれば、およそ 25 歳までは年齢が高くなるとともに労働時間が増加し、その後は（定年になるまで）週に 49 時間以下程度で安定することが分かります。これに対して、所得は 50 歳になるまで増加し、その後は頭打ちになり、やがて低下していきます。実際の数字は職業やその他の要因によって異なります。

たとえば、craft-repair（修理業）と prof-specialty（専門家）を比較するときは、「表示 (View)」->「フィルタパネルの表示 (Show Filter Panel)」を選択して「フィルタ (Filter)」パネルを開き、職業リストから 「修理業 (craft-repair)」と「専門家 (prof-specialty)」を選択します。これらの職業をアニメーションで表示すれば分かるように、「専門家 (prof-specialty)」の所得は当初低いものの、年齢が高くなると即座に「修理業 (craft-repair)」の所得を上回ります。「専門家 (prof-specialty)」の教育年数は「修理業 (craft-repair)」よりかなり高くなっています。女性だけまたは特定の人種だけというように、比較の対象をさらに絞ることもできます。アニメーションによる表示を行うときに、各グラフィカル・オブジェクトのサイズ、カラー、位置を変化させることもできます。

- *census.data* と *census.scatterviz*

  これらのファイルは、国勢調査データの集計値の散布図を示しています。元のデータセットには約 150,000 件のデータが存在しますが、散布図では性別、教育水準、業種、職業（グループ化に使用された属性）のすべての組合わせについて 1 つの立方体（集計値）が表示されます。

スプラット・ビジュアライザ用のサンプルファイル

スプラット・ビジュアライザの機能や特長を紹介するために、下記の説明の通り、スプラット・ビジュアライザ用のサンプルファイルが用意されています。これらのファイルは *examples* ディレクトリ内にあります。

Windows システムの場合、これらのファイルは MineSet がインストールされているディレクトリ *\examples* にあります。

IRIX システムの場合、これらのファイルは */usr/lib/MineSet/examples/splatviz* にあります。

- *mushroom*
  mushroom.data ファイルには、5,000 種類以上のキノコに関する集計済みデータが入っています。元のデータのグループ化に使用された項目（属性）は、「香り (odor)」、「ひだの色 (gill_color)」、「傘の色 (cap_color)」です。これら 3 つの項目のあらゆる組合わせについて、サンプル数と平均食用性（0 は食用キノコ、1 は有毒キノコ）が示されています。食用性の値が 0 から 1 の間にある集計階級（スプラット）には、1 つ以上の有毒キノコが含まれています（特定のキノコは食用または有毒のどちらか一方に分類され、" 部分的に有毒 " という階級はありません）。

  このデータファイルのビジュアル図では、3 つの項目の個別値が平均食用性に従い、軸に沿ってソートされています。食用かどうかを判断するのに最も有効な項目は明らかに「香り (odor)」です。また、ほとんどのスプラットはすべて 0 であるか、すべて 1 になっています。すなわち、これら 3 つの項目は、キノコを食用と有毒に二分割するのに有効です。実際、軸にマッピングする項目を選択するときに、主成分抽出ツールが使用されています。サンプル数が最も多いスプラットを見つけるには、不透明度スライダを低い値に移動します。不透明度が最も高いスプラットは、「香り (odor)」、「ひだの色 (gill_color)」、「傘の色 (cap_color)」の値が同じである 288 種類のキノコを表しています。

- *adultJobs*
  adultJobs.data ファイルは MineSet パッケージに付属の *adult94* データセットに基づき、「最終学歴 (education)」、「職業 (occupation)」、「週の労働時間 (hrs_worked_per_week_bin)」（階級生成済み）、「年齢 (age_bin)」（階級生成済み）をグルーピング (group-by) 項目として、属性の個別値を集計処理して作成されたデータセットです。スプラット・ビジュアライザの表示では、「年齢 (age_bin)」がスライダにマッピングされ、他の「グルーピング (group-by)」項目は軸にマッピングされています。「総収入 (gross_income)」項目は、カウントと平均によって集計処理されています。「総収入カウント (count_gross_income)」項目は不透明度にマッピングされ、「平均総収入 (avg_gross_income)」項目は色にマッピングされています。

スライダが左端の位置にあるとき、スプラットはほとんど青で埋めつくされます。すなわち、職業、最終学歴、週の労働時間にかかわらず、20 歳未満の人の収入が低いことが分かります。スライダを右方向に動かすと、最終学歴が高くなり職業が軸の右端に近づくに従って、収入の増加度合いが速くなることが分かります。不透明度の分布から明らかなように、最終学歴で一番多いのは「高校卒業 (HS-grad)」、「中退 (some-college)」、「学士号 (Bachelor)」です。

サマリスライダを動かすと、年齢の変化とともに軸上の項目の値が変化し、それに伴って所得の分布も変化することが分かります。

- *adultJobs2*
  adultJobs2 ファイルも *adult94* データセットに基づいて作成されています。このファイルでは、「職種 (working_class)」、「最終学歴 (education)」、「職業 (occupation)」の各項目が軸にマッピングされています。スライダには、「年齢 (age_bin)」（階級生成済み）と「週の労働時間 (hours_worked_per_week_bin)（階級生成済み）がマッピングされています。ここでも「総収入 (gross income)」はカウントと平均によって集計処理され、それぞれ不透明度と色にマッピングされています。2D スライダ上には数多くのポイント（位置）があるため、各ポイントに対応するレコード数は相対的に少なくなります。そのため、色と不透明度の変化度合いが大きくなります。サマリスライダ上で「週の労働時間 (hours_worked_per_week_bin)」次元の中央の領域が赤いことから、大半の人の週当たり労働時間は 35 ～ 45 時間であることが分かります。特定の職業は特定の業種と相関があります。たとえば、職業が「軍隊 (Armed-forces)」である人はすべて「国家公務員 (Federal-gov)」です。

- *censusIncome*
  このサンプルは *adult94* に似たデータセットをベースにしていますが、サイズが大きいため（censusIncome のレコード数は 150,000 件）、MineSet パッケージには収録されていません。総収入と合計収入の差異を調べるために、「総収入 (gross_income)」、「合計収入 (total_income)」、「週の労働時間 (hrs_per_week)」の各項目が軸にマッピングされています。色は年齢を表しています。このデータを可視化すると、「合計収入 (total_income)」=「総収入 (gross_income)」となっているレコードが多いことが分かります。ただし、「合計収入 (total_income)」の値は大きいにもかかわらず、「総収入 (gross_income)」が 0 であるレコードも少なくありません。驚くべきことに、「総収入 (gross_income)」が「合計収入 (total_income)」を上回るケースも数多く観察されます。

  年齢による分布の差異も注目に価します。高齢者（黄色）の多くは「週の労働時間 (hrs_per_week)」=0 となっていますが、これはおそらく退職のためです。青少年（青）の多くは「総収入 (gross_income)」=「合計収入 (total_income)」=0 となっています。両端の領域には不透明度の高いスプラットが見られます。この領域には、軸の最大階級 (maximum bin) に属するポイントがすべて含まれます。たとえば、「合計収入 (total_income)」の最高階級は 70,300 以上であり、値が 703,00 以上のポイントはすべてこの階級に入ります。

密度の変化を確認するために、不透明度スライダを動かしてみます。不透明度のスケールを低い値に移動すると、「総収入 (gross_income)」=「合計収入 (total_income)」であるか、または「合計収入 (total_income)」だけで「総収入 (gross_income)」は 0 であるケースが大半であることが分かります。不透明度のスケールを高い値に移動すると、データがほとんど全域に分布していることが分かります。このデータセットには、150,000 件のレコードがあります。

- *churn*
  顧客が今まで使用していた電話会社から別の会社に切換えることを「解約 (Churn)」と呼びます。churn データセットは、電話会社の「解約 (churn)」データを示しています。この例で使用するデータの様式は、*churn.schema* ファイルで定義されています。

  主成分抽出機能を使用した結果、「昼間料金 (total_day_charge)」、「顧客サービスコール数 (number_customer_service_calls)」、「国際化プラン (international_plan)」が重要な項目であることが分かっています。そのため、これらの項目を軸にマッピングしています。さらに、この例では、解約の場合は Yes (churned==Yes) という値を取る新しい数値型項目 churn を定義して、色にマッピングしています。

  このデータを可視化すると、赤い領域は解約率が高いことが分かります。顧客サービスコール数 (number_customer_service_calls) が 3 回以上で昼間料金 (total_day charge) が低い領域で、解約率が高くなっています。他の顧客よりも多くの料金を払っている顧客には、大きい重みを割当てる必要があります。このような重み付けを行うには、「合計料金 (total_charge)」という新しい項目を作成します。この項目の値は、次の式による値、またはその値の累乗値とします。

  ```
  `total_day_charge`+`total_eve_charge`+`total_night_charge`
  ```

  次に、「合計料金 (total_charge)」項目を不透明度にマッピングします。これによって、すべてのレコードが「合計料金 (total_charge)」によって重み付けされます。「昼間料金 (total_day_charge)」軸の高い値付近には、精査を必要とする新しい領域が表示されます。

ツリー・ビジュアライザ用のサンプルファイル

ツリー・ビジュアライザの機能や特長を紹介するために、下記の説明の通り、ツリー・ビジュアライザ用のサンプルファイルが用意されています。
これらのファイルは examples ディレクトリ内にあります。

- Windows システムの場合、これらのファイルは MineSet がインストールされているディレクトリ *\examples* にあります。

- IRIX システムの場合、これらのファイルは */usr/lib/MineSet/examples/treeviz* および */usr/lib/MineSet/data* にあります。

スプラット・ビジュアライザ用のサンプルファイルには、次のものがあります。

- *store.data* と *store.treeviz*
  これらのファイルを使用すると、架空のチェーンストアの売上データがビジュアルに表示されます。階層はチェーン全体、地域、州、都市、個々の店から構成されています。階層の各レベルには 4 種類の製品が表示されます。高さは売上（ドル単位）を表し、色は目標売上高に対する達成率（% 単位）を表しています。

- *stateRevenue.data* と *stateRevenue.treeviz*
  これらのファイルを使用すると、各州の 1992 年度政府予算の歳入内訳がビジュアルに表示されます。データは米国統計局 (http://www.census.gov/govs/state/stfin92.dat) から入手したものです。高さは税収総額（ドル単位）を表しています。背景の下位ノードは、ルートノードに示された歳入総額に対する各種税収の割合を表しています。

- *beer.data* と *beer2.data*、*beer.treeviz* と *beer2.treeviz*
  これらのファイルを使用すると、マーケット・リサーチに基づく架空のビール消費データがビジュアルに表示されます。階層は次の 3 つのレベルから構成されています。

  1. 最初のレベルは商品の種類です（たとえばビールやエール）。
  2. 2 番目のレベルはブランドのコードです（コードはランダムに割当てられています）。
  3. 3 番目のレベルは個々の商品のコードです。たとえば、1 ダース入りと半ダース入りなどであり、ランダムに割当てられています。

  各チャートには 7 つのバーがあります。各バーはそれぞれ特定の年齢グループを表します。バーの高さは該当の年齢グループがビールを消費した総金額（ドル単位）を表し、色は消費総額における男性と女性の割合（% 単位）を表します。ここで使用されているブランド、商品、データはすべて架空のものです。

  beer.treeviz と *beer2.treeviz* は可視化すれば同じ結果になりますが、その構成過程が異なります。beer.treeviz では、各種類のビールが女性と男性の消費額を示す単一のレコードで表されています。これらの消費額は列挙型の配列に保存されています。

  beer2.treeviz では、各種類のビールについて 7 つのレコードがあり、各レコードが特定の年齢グループを表しています。*beer* ファイルでは年齢グループが設定ファイル (.treeviz) 内で定義されているのに対し、*beer2* ファイルでは年齢グループがデータファイル (.data) に入っています。

beer ファイルは *beer2* ファイルほど大量のディスク領域を必要としません。その代わり、設定ファイルが多少複雑になります。場合によっては、beer2 ファイルの形式でデータを生成した方が作業負荷は小さくなります。

索引

記号

「% より短いものを」オプション , 199
「% より短いものをフィルタにより除去 (Filter out % shortest)」オプション , 199
* ワイルドカード , 104, 206
? ワイルドカード , 206
[] ワイルドカード , 104, 206

数字

「1 ステップ飛越」ボタン , 14
2D 集計 , 13
2 次元集計 , 13
3D ランドスケープ , 170, 193
3 次元 (3D) ランドスケープ , 170, 193
64 ビット対応 , 147
  systune パラメータ , 147

A

adultJobs.data, 280
adult-salary.dtableviz, 243
adult-salary.eviviz, 259, 261, 263, 264, 265
adult-salary.schema, 259
adult.schema, 230, 242, 257
adult-sex.dtableviz, 242
adult-sex-dt.treeviz, 230
adult-sex.eviviz, 257
any キーワード
  色 , 200
australia.states.gfx, 267
australia.states.hierarchy, 267
avg キーワード
  色 , 200

B

beer2.data, 283
beer2.treeviz, 283
beer.data, 283
beer2.treeviz, 283
beer.treeviz, 283
beer2.treeviz, 283
blocks.data, 267
blocks.hierarchy, 267
blocks.gfx, 267
blocks.hierarchy, 267
blocks.mapviz, 267
brand.data, 278
brand.scatterviz, 278
breast.dtableviz, 251
breast-dt.treeviz, 235
breast.eviviz, 263
breast.schema, 251, 263

C

canada.provinces.gfx, 267
canada.provinces.hierarchy, 267
cars.data, 278
cars.scatterviz, 278
cars-dt.treeviz, 229
cars.eviviz, 256
cars.scatterviz, 278
cars.schema, 229, 256
censusIncome データファイル , 281
churn データセット , 239
churn-dt.treeviz, 228
churn.schema, 282
churn データセット , 226, 228, 255, 270, 282
.clusterviz.data ファイル , 52
company.data, 277
company.scatterviz, 277
company-total.scatterviz, 277

D

.data 拡張子 , 103, 127, 160, 174, 194
dna.dtableviz, 254
dna.eviviz, 265
DNA 境界データセット , 272
DNA データセット , 254, 266
DNA データセット , 237

E

europe.countries.gfx, 267
europe.countries.hierarchy, 267
Excute（実行）ステートメント
  スキャタ・ビジュアライザ
    警告コマンドの有効化 , 219
    実行 , 219
Excute（実行）ステートメントの実行
  スキャタ・ビジュアライザ , 219

F

fasta.m.data, 268
fasta.m.hierarchy, 268
fasta.m.gfx, 268
fasta.m.hierarchy, 268
fasta.m.mapviz, 268

G

gfx ファイル , 127
  サンプル , 267
  生成 , 129-131

H

hierarchy ファイル
  サンプル , 267
hypothyroid, 252
hypothyroid.dtableviz, 252, 253, 254, 255
hypothyroid.eviviz, 264
hypothyroid.schema, 236, 252, 264

I

iris.dtableviz, 248, 249, 250
iris-dt.treeviz, 233
iris.eviviz, 261
iris.schema, 233, 261
Is Null 演算子 , 105, 207

L

LANG, 国際化 , 114

M

.mapviz 拡張子 , 128
max キーワード
  色 , 200
MineSet, 32
  ツール
    「概要 (Overview)」, 192
MINESET_WARN_EXECUTE 変数 , 219
MineSet の mtr 拡張機能 , 220
min キーワード
  色 , 200
mtr ファイル , 219
mushroom.data, 280
mushroom.dtableviz, 249
mushroom-dt.treeviz, 234
mushroom.eviviz, 261, 262, 263, 264, 265, 266
mushroom.schema, 234, 249, 261

N

Naive-Bayes, 135
nl.births.data, 278
nl.births.scatterviz, 278
「Null (Nulls)」コマンド , 211
Null 値 , 167, 213
  「スプラット」, 177
  マッピング , 167, 214
  オブジェクト , 211
  「決定木」, 84
  予測 , 123
「Null を 0 とみなす」オプション , 84

P

people.data, 278
people.scatterviz, 278
perhouse.perage.data, 268
perhouse.perage.mapviz, 268
pima.dtableviz, 253
pima-dt.treeviz, 237
pima.schema, 237, 253, 265
population.australia.data, 267
population.australia.mapviz, 267
population.canada.data, 267
population.canada.mapviz, 267
population.europe.data, 267
population.europe.mapviz, 267
population.usa.cities.data, 268
population.usa.cities.mapviz, 268
population.usa.data, 267
population.usa.mapviz, 267

Q

-quiet オプション , 95, 129, 219

R

ROI 曲線 (ROI Curve), 110, 157
「ROI 曲線」オプション , 111
.ruleviz から .scatterviz への変換 , 33

S

.scatterviz 拡張子 , 161
Simple Bayes, 135
.splatviz.data 拡張子 , 177

.splatviz.schema 拡張子 , 177
stateRevenue.data, 283
stateRevenue.treeviz, 283
store.data, 283
store.treeviz, 283
store-type.data, 278
store-type.scatterviz, 278
systune パラメータ
  64 ビット対応 , 147
  rlimit__nofile_cur, 147
  rlimit__rss_cur, 147
  rlimit__vmem_cur, 147
  rlimit_pthread_cur, 147

T

telecom.data, 268
telecom.mapviz, 268
Tool Manager
  オプションの設定
    決定木分析 , 82
    エビデンス・ビジュアライザ , 94
    スプラット・ビジュアライザ , 176
    相関規則 , 27
    ツリー・ビジュアライザ , 203
    マップ・ビジュアライザ , 129, 134
  多重規則
    図 , 29
  ツールオプション
    スキャタ・ビジュアライザ , 162, 166
    スプラット・ビジュアライザ , 176
    ツリー・ビジュアライザ , 195-203
    マップ・ビジュアライザ , 131-134
Tool Manager に送信」コマンド , 253
「Tool Manager に送信」コマンド , 168, 210
  前述の使用方法 , 85
「Tool Manager の起動」コマンド
  ツリー・ビジュアライザ (Tree Visualizer), 100, 101
.treeviz 拡張子 , 194

U

UNIX 起動コマンド
  相関規則 , 27
UNIX コマンド , 133, 165, 201
UNIX コマンドの実行 , 133, 165, 201
usa.cities.gfx, 268
usa.cities.hierarchy, 268
usa.cities.lines.gfx, 268
usa.states.hierarchy, 268
usa.cities.lines.hierarchy, 268
usa.states.hierarchy, 268
usa.states.gfx, 267, 268
usa.states.hierarchy, 267, 268, 267, 268

V

vote.dtableviz, 250
vote-dt.treeviz, 235
vote.eviviz, 263
vote.schema, 263

W

-warnexecute オプション , 219
「Web 公開用ファイルの作成」コマンド , 101
Web ファイル , 219
"what if" 型の質問 , 91

X

X スライダ , 164

Y

Y スライダ , 164

あ

アウトライン , 211
「アウトライン」オプション , 209
「アウトラインファイル」フィールド , 131
値 , 複数選択 , 135
「値の表示」コマンド , 210
「新しい項目名」テキスト・フィールド , 19
新しいデータセット内のレコード , 87
アニメーション , 125, 169
アニメーション・コントロール・パネル , 11, 16
  アニメーションの開始 , 14
  サマリウィンドウ , 13
  ボタン , 14
アニメーション・コントロール・パネル
  サマリウィンドウ , 164
  表示 , 217
アニメーション・コントロール・パネル , 178
  サマリウィンドウ , 176
「アニメーション パネルの表示」コマンド , 217
「アニメーション フロー」ボタン , 14
アヤメの分析モデルデータセット , 233, 248, 261, 270
アルゴリズム
  サイズの変更 , 80, 145
「アルファベット順」コマンド , 77, 141
アルファベットの文字列
  検索 , 206
  フィルタリング , 104

い

「一回だけアニメーションを実行」ボタン , 15
「一致」検索オプション , 104, 206
色 , 53, 199
  キーによる入力 , 200
  グリッド , 166, 176
  「スプラット (Splats)」, 175
  空 , 202
  地表 , 202
  ディスク , 200
  バー , 200
    キーに基づく , 200
    ラベル , 202
  ラベル
    バー , 202
    ベース , 202
  変更 , 56
  ベース , 200
    ラベル , 202
  要素 , 163
  ライン , 202
「色」オプション , 199
色の選択 , 56, 57
「色の選択」ダイアログ・ボックス , 166
色の変更 , 56
「色のマッピング」オプション , 164, 176, 200
「色のリスト」, 132, 164, 175, 200
色の割当て

スキャタ・ビジュアライザ , 164
スプラット・ビジュアライザ , 176
ツリー・ビジュアライザ , 200-202
NULL 値 , 214
マップ・ビジュアライザ , 132
インターネット・ファイルのダウンロード , 219

う

「上に移動」コマンド , 212
売上のサンプルファイル , 278, 283

え

「枝刈り係数」オプション , 81
枝刈り法
決定木分析
「信頼度」, 81
生成コストも考慮 , 81
エビデンス・ビジュアライザ
オブジェクトの選択 , 96
確率 , 91, 92
補正 , 93
「概要」, 91
起動 , 94-95, 129
起動オプション , 95
サンプルファイル , 255
設定 , 94
メインウィンドウ , 98
メニュー , 98-99, 143, 158
予測 , 92
エビデンス・クラシファイア , 91
生成 , 92
エビデンス
重要項目 機能 , 61
「エビデンス」パネル
オブジェクトの選択 , 96
「エラー」オプション , 109-111
演算子
関係 , 206
データのフィルタリング , 104
遠方の地平線 , 202

お

オーストラリアの地図 , 130, 267
大文字 / 小文字を区別せずに検索 , 207
「大文字 / 小文字を区別せずに検索」オプション , 205, 207
「大文字 / 小文字を区別せずにフィルタリング 」オプション , 209
大文字と小文字の区別を無視するフィルタ , 209
オブジェクト
検索 , 83-84, 204-208
選択
NULL 値 , 143
ツリー・ビジュアライザ , 208
選択されているオブジェクトの表示 , 135
高さが NULL の , 211
高さがゼロの , 211
地理的 , 127, 131
メッセージの表示
マップ・ビジュアライザ , 133
スキャタ・ビジュアライザ , 165
ツリー・ビジュアライザ , 201
オブジェクトの強調表示
スキャタ・ビジュアライザ , 167
オブジェクトの選択
NULL 値 , 143, 214
スキャタ・ビジュアライザ , 167
ツリー・ビジュアライザ , 208
「重み順」コマンド , 77, 141
「重み付けとして使用」オプション , 111
「重み付けとして使用」メニュー , 41
「親ノード」ボタン , 212
「オリジナルデータの表示」コマンド , 254

「オリジナルデータの表示」コマンド , 253
「オリジナルデータを表示」コマンド , 168, 210
  前述の使用方法 , 85

か

回帰ツリー分析
  オプション , 153
    コスト / 複雑性を調整する枝刈り , 155
    ツリーの高さ制限 , 154
    分割の下限値 , 155
    「分割の基準」, 154
  「概要 」, 152
  誤差の推定 , 156
回帰分析 , 152
階級生成 , 39
「階級での重み付け下限値」オプション , 93
解析
  関係 , 159, 169
解析
  関係 , 193
改善曲線 , 20, 121
「改善曲線の表示」オプション , 121
改善率 (Lift), 25
階層 , 193
  移動 , 212
「階層」オプション , 84
階層型データ
  サンプルファイル , 225
階層ファイル , 127
  生成 , 129-131
  定義 , 131
「階層」フィールド , 205
「カイ二乗」, 80
「拡張コントローラーの表示」コマンド , 217
拡張子
  スキャタ・ビジュアライザ , 161
確率 , 91
  生成 , 92
  補正 , 93
確率評価 , 38, 44
カナダの地図 , 130, 267
カラーエディタ , 132, 200
カラーブラウザ , 57
  開く , 56
カラーボックス , 53, 56
関係、解析 , 159, 169, 193
関係演算子 , 206
  データのフィルタリング , 104
「ガウシアン」コマンド , 181
学習曲線 (Learning Curve), 111, 118-120
  オプション , 119-120
「学習曲線」モード , 119
訓練事例 , 22, 108

き

キー
  バーへの色の割当て , 200
「キーで色を割当」オプション , 200
「キーによるソート」オプション , 202
規則ファイル , 26
期待される信頼度 , 25
起動 , 70
  エビデンス・ビジュアライザ , 94-95, 129
  スキャタ・ビジュアライザ , 129, 161, 174
  スプラット・ビジュアライザ , 174-175
    デフォルト値のリセット , 177
  相関規則 , 27
  デシジョン・テーブル分析 , 70
  ツリー・ビジュアライザ , 194
  デフォルト値のリセット , 177
  マップ・ビジュアライザ , 128
キノコの分析モデルデータセット , 234, 249, 261,

271, 280
キノコ分析データセット
混同マトリックス , 123, 124
「球体」コマンド , 182
給料を決める要因データセット , 232, 243, 259
寄与率 , 58, 78, 83
テスト , 60
「寄与率」オプション , 83
寄与率尺度 , 61
寄与率（定義済み）, 92
「寄与率でソート」コマンド , 99
寄与率の測定 , 78, 83
切捨て係数 , 106
「逆再生」ボタン , 14
行番号の振り直し
レコードビューワ , 150

く

「区間自動生成」
「均一な重み」, 41
「均一な範囲」, 41
「クラシファイアオプション」ダイアログ・ボックス , 79
「クラシファイアとエラー」モード , 38, 109, 110, 121
出力の表示 , 112
「クラシファイアとエラー」モード
出力の表示 , 112
クラシファイアの精度のテスト , 109
「クラシファイアの適用」オプション , 69
「クラシファイア」のバックフィッティング , 38, 110
「クラシファイアのみ」モード , 109, 112
クラス
レコードの割当て , 77, 91, 143
クラスタリング・アルゴリズム
単一 k-means, 46
反復型 k-means, 48
クラスタリングにおける「重み付けとして使用」, 51
クラスタリングにおける距離の測定法 , 51
クラスタリングにおける「属性の重み」, 51
クラスタリングの実行 , 48
「クラス判別」, 43-44
改善曲線 , 121
学習曲線 , 111, 118
オプション , 119-120
出力の表示 , 112
混同マトリックス , 64-65, 110
主成分抽出 , 61
出力の表示 , 112
生成 , 43, 78, 92, 144
精度 , 87
テスト , 109
損失マトリックス , 111, 125
定義済み , 43
投資利益率 (ROI Return on Investment) 曲線 , 111
バックフィッティング , 38, 110
未知の値の予測 , 123
レコードの重み付け , 111, 151
レコードへの適用 , 38
クラス判別のタイプ , 88, 89
「クラス判別詳細オプション」コマンド , 95, 145, 79
クラスラベル , 108
検索 , 83
車の原産国データセット , 229, 241, 256, 270
グリッド
「色」オプション , 166, 176
線の間隔 , 166, 176
「グリッド X、Y、Z サイズ」オプション , 166, 176
「グリッドの色」オプション , 166, 176

け

警告コマンド , 219
計算済みの項目 , 1
「決定木分析」
  NULL 値 , 84
  オプションの設定 , 79-82, 145-146
  寄与率の測定 , 78, 83
  枝刈り , 81
  検索 , 83-84
  誤差 / 損失評価 , 78, 84
  ノード
    情報の表示 , 78
  表示 , 112
  フィルタリング , 83
  分割 , 80, 183
決定木分析 , 82, 238
  アルゴリズムの調整 , 80
  「カイ二乗」, 80
  「概要」, 77
  枝刈り法
    「信頼度」, 81
    生成コストも考慮 , 81
  サンプルファイル , 228-238
  重要項目 (Column Importance) 機能 , 61
  ジニ集中係数 , 80
  設定 , 82
  ノード情報の表示 , 78
決定木クラシファイア
  オブジェクトの検索 , 83-84
  「概要」, 77
  生成 , 78
検索
  レコードビューワ , 150
「検索」コマンド , 204
検索する , 83-84, 204-208
  検索条件の定義 , 206
  ワイルドカード , 104, 206
「検索」ダイアログ・ボックス , 208
検索のスポットライト , 208
  オフにする , 208
「検索パネル」コマンド , 83
「検索」ボタン , 207
「現在のセッションを別名保存」コマンド , 203
「現在のデータセットの項目名」ウィンドウ , 68
「現在のデータセットの項目名」テキストボックス , 6
現在の履歴表示 , 187
「減少数」オプション , 145

こ

降順ソート , 202
甲状腺機能低下症 (hypothyroid) データセット , 236, 252, 264, 272
項目
  計算済み , 1
  集計処理オプション , 5, 6
  選択 , 58, 60
  表示 , 68
  命名 , 2
「項目によるグループ化」オプション , 6
「項目の階級生成」オプション , 68
「項目の階級生成 」ボタン , 39
「項目の削除」オプション , 68
「項目の自動選択」オプション , 94
「項目の追加」オプション , 69
項目の追加ダイアログ・ボックス , 4
「項目の追加」ボタン , 1
国際化
  リソースファイル , 115
国際化 , 114-117
  LANG, 114
  他の言語とエンコーディングに対する拡張機能 , 114
  リソースファイル

例 , 116
ロケール , 114, 115
リソースファイル , 115
リソースファイルの例 , 116
ロケールの設定 , 114, 115
国勢調査データベースのサンプルファイル , 205, 225
「固定」オプション , 201
子ノード「カイ二乗」
選択 , 212
コマンド行オプション , 27
混同マトリックス , 20, 64-65, 110
「混同マトリックスの表示」オプション , 110
誤差 / 損失評価 , 78
誤差推定
回帰ツリー分析 , 156
「誤差推定」オプション
平均絶対誤差 , 156
平均二乗誤差 , 156
誤差率
精度 , 42
誤差 / 損失評価 , 84

さ

サーバに接続 *, 191
「最後の子ノード」コマンド , 213
「最小因子の排除」コマンド , 99
最小の適合比率 , 145
「最初の子ノード」コマンド , 213
「再生」ボタン , 14
「最速の先送り」ボタン , 14
「最速の巻戻し」ボタン , 14
「最大 # ルート選択肢数」オプション , 145
「最大値 / スケールの高さ」オプション , 198
「再度開く」コマンド , 100, 101
歳入のサンプルファイル , 283
「削除される項目」オプション , 6
「サブツリー重み」オプション , 83
「サブツリーの正規化」コマンド , 211
サマリウィンドウ , 13, 164
サマリウィンドウ , 176
「サマリ」オプション , 164, 176
サマリの説明（値の説明）, 164
サンプルファイル
エビデンス・ビジュアライザ , 255
回帰ツリー分析 , 272
クラスタ・ビジュアライザ , 226
決定木分析 , 228-238
国勢調査データベース , 225
重要項目 , 226
信用データベース , 225
スキャタ・ビジュアライザ , 277
スプラット・ビジュアライザ , 182, 280-282
製品カテゴリ , 225
製品グループ , 225
選択式決定木分析 , 269
相関規則分析 , 225
デシジョン・テーブル・ビジュアライザ , 238-254
ツリー・ビジュアライザ , 215
マップ・ビジュアライザ , 269

し

支持度 , 25
最小区間 , 25
支持度の最小区間 , 25
集計処理 , 4-6, 193
2 次元 , 13
色 , 200
オプション , 5, 6
データポイント , 172, 173
バーの高さ , 199
「集計処理」オプション , 68
「集計処理結果の高さ」オプション , 199

集計処理ダイアログ・ボックス , 4, 5
「集計処理の色」オプション , 200
「集計処理 」ボタン , 4
集計処理ボタン , 4
「集計処理を行う項目」オプション , 6
終端 , 126
終了
  ツリー・ビジュアライザ , 101
「終了」コマンド , 101
終了
  ツリー・ビジュアライザ , 101
重要項目機能
  重要度ランキング , 61
重要項目アルゴリズム , 61
重要項目機能
  依存関係 , 61
  寄与率尺度 , 61
  サンプルファイル , 226
  モード , 59-60
  離散的な属性 , 61
「消去」ボタン
  「検索」ダイアログ , 205, 207
詳細モード , 59-60
「詳細モード」ボタン , 59
昇順ソート , 202
消費に関するリサーチのサンプルファイル , 283
消費のサンプルファイル , 268
進行状況ダイアログ , 無効化 , 95, 129, 219
進行状況ダイアログの非表示 , 219
信用データベースのサンプルファイル , 225
「信頼度」, 24, 25
  期待 , 25
軸
  非表示のラベル , 176
  表示オプション , 165
  ラベルの命名 , 166, 176
「軸」オプション , 165
  最大サイズ , 165
  「スケールサイズ」, 165
  「調整無し」, 165
「軸のラベルサイズ」オプション , 166, 176
事前確率 , 91, 93
「実行」オプション , 133, 165, 201
自動離散化アルゴリズム , 61
ジニ集中係数 , 80
重要度ランキング , 61
条件付き確率 , 92, 93
人口のサンプルファイル , 267

す

推定確率値モード , 18
「スイング」ボタン , 15
数式 , 1
数値
  検索 , 206
  フィルタリング , 104
スキャタ・ビジュアライザ
  NULL 値 , 167
  アニメーション・コントロール・パネル
    サマリウィンドウ , 164
    表示 , 217
  色の割当て , 164
  オブジェクトの選択 , 167
  オプション , 162-166
    保存 , 166
  起動 , 129, 161, 174
  項目の選択 , 60
  サンプルファイル , 277
  相関規則 , 27
  データファイル , 160
  デフォルト値のリセット , 162
  必要なファイル , 160
  ファイルの読み込み , 161
スケーリング

地域 , 132
バー , 198
ベース , 198
要素 , 163
「進む」コマンド , 212
スキャタ・ビジュアライザの「ツールオプション」ダイアログ・ボックス , 163-166
「スピード」スライダ , 15
「スプラット」
「色」オプション , 175
定義済み , 170
表示 , 175
描画オプション , 175, 181
ラベルの命名 , 176
「スプラット」オプション , 175
「スプラットの色」オプション , 175
「スプラットの形状」オプション , 175
「スプラットのタイプ (Splat Type)」メニュー , 181
スプラット・ビジュアライザ (Splat Visualizer), 177, 282
NULL 値 *, 177
アニメーション・コントロール・パネル , 178
サマリウィンドウ , 176
色の割当て , 176
オプション , 176
リセット , 176
起動 , 174, 175
起動オプション , 175
サンプルファイル , 182, 280-282
設定
Tool Manager, 176
データの表示 , 170, 181
データファイル , 173
データポイントの集計 , 172, 173
デフォルト値のリセット , 177
必要なファイル , 173
メニュー , 181
スプラット・ビジュアライザの「ツールオプション」ダイアログ・ボックス , 176
「全てに設定」ボタン
「検索」ダイアログ , 205
スポットライト , 208
オフにする , 208
スライダ
作成 , 162, 169
スプラット・ビジュアライザ用に作成 , 178
「マッピング」オプション , 164
「スライダ」オプション , 164
スライダ制御 , 11-15

せ

「正規化」オプション , 198
「正規化オン」オプション , 132
「正規化された相互情報量」オプション , 80, 183
政党への帰属データセット , 234, 250, 271, 263
精度
ブースティング , 42
精度 (クラシファイア), 87
テスト , 109
製品カテゴリのサンプルファイル , 225
製品グループのサンプルファイル , 225
性別の属性データセット , 230, 242, 257
西暦 2000 年問題に対応した（Y2K 準拠の）, 220
西暦 2000 年問題への対応 , 220
設定
決定木分析 , 82
エビデンス・ビジュアライザ , 94
スプラット・ビジュアライザ
Tool Manager, 176
ツリー・ビジュアライザ
Tool Manager, 203
マップ・ビジュアライザ
Tool Manager, 129-134
設定ファイル , 191
.ruleviz から .scatterviz への変換 , 33

エビデンス・ビジュアライザ
読み込み , 94
スキャタ・ビジュアライザ , 160
サンプル , 277, 278
読み込み , 161
スプラット・ビジュアライザ , 174
読み込み , 174
相関規則のサンプル , 225
ツリー・ビジュアライザ , 194
サンプル , 283
読み込み , 100, 101, 194
マップ・ビジュアライザ , 128
サンプル , 267
読み込み , 128
接続 , 191
「説明オン」オプション , 132
説明（凡例）
サマリ , 164
相関規則 , 31
地域 , 132
要素 , 163
選挙のレコード例 , 234, 250, 263, 271
「線状」コマンド , 181
「選択された項目の区間」, 42
選択式決定木分析 , 143
生成 , 144
選択式決定木分析
アルゴリズムの調整 , 145
「概要」, 143
サンプルファイル , 269
必要なファイル , 144
選択肢ノード
定義済み , 143, 146
「選択」ボタン , 207
「選択」メニュー
エビデンス・ビジュアライザ , 99
ツリー・ビジュアライザ , 210
線の間隔（グリッド）, 166
線の間隔（グリッド）*, 176
「占有率」オプション , 83
「ゼロ」コマンド , 211
「ゼロ」の値
オブジェクト , 211
「全体の概観」コマンド , 212

そ

相関規則 , 23, 23-33
改善率 , 25
期待される信頼度 , 25
起動 , 27
支持度 , 25
最小区間 , 25
「信頼度」, 24, 25
期待 , 25
スキャタ・ビジュアライザ , 27
生成 , 24
設定
Tool Manager, 27-31
多重 , 29
表示 , 29
データのマッピング , 30-31
ドリルスルー , 32
表示 , 31
マーケット・バスケット , 24
レコードの重み付け , 28
相関規則
サンプルファイル , 225
相関規則の起動 , 27
相関規則の生成 , 24
相関規則の設定
Tool Manager, 27
「相関規則の対応付け」パネル , 30
相関規則の表示 , 31
相関規則分析 , 23, 23-27
「概要」, 23
構成要素 , 24, 26

出力 , 24
説明（凡例）の表示 , 31
必要なファイル , 26
相関ファイル
サンプル , 225
相互検証法 , 66, 89, 110
「相互情報量」オプション , 183, 80
ソート , 39, 202
ソート順序
定義 , 202
空の色 , 202
「ソリッド」オプション , 209
損失マトリックス , 111, 125
「損失マトリックスを使用」オプション , 111, 125
「増加比率」オプション , 80, 183
属性 , 43, 84, 108
属性の数 , 87
テスト , 83
離散化アルゴリズム , 61
「属性値の順序付け」メニュー , 140, 77, 99
「属性としても使用」オプション , 111

た

大量メモリー対応 , 147
systune パラメータ , 147
高さの正規化 , 198
「高さ」フィルタスライダ , 209
多次元データポイント , 12
多重規則 , 29
Tool Manager, 29
表示 , 29
「他のオプション」オプション , 165, 176
デシジョン・テーブル分析 , 70
デシジョン・テーブル・ビジュアライザ , 254
サンプルファイル , 238, 254
メニュー , 75
単一 k-means, 46

ち

地域 , 130
スケーリング , 132
説明（凡例）, 132
メッセージ , 133
小さい値 , フィルタ除去 , 209
「地形ファイル」オプション , 131
地表の色 , 202
地理的オブジェクト , 127, 131

つ

ツール
「概要」, 192
複数選択 , 135
ツールオプションの保存
スキャタ・ビジュアライザ , 166
ツリー・ビジュアライザ , 203
マップ・ビジュアライザ , 133
ツールオプションのリセット
スプラット・ビジュアライザ , 176
ツリー・ビジュアライザ , 203
ツリー・ビジュアライザ
データのフィルタリング , 208
次のようなアプローチ
「自動」, 37
「次のようなアプローチ」メニュー
「階級での重み付けの下限値」, 40
「次」フィールド , 188
「次」ボタン , 207
「ツリーの高さ制限」オプション , 80
ツリー・ビジュアライザ
NULL 値 , 213
移動 , 211

色の割当て , 200
印刷 , 100, 101
オブジェクトの検索 , 204-208
オブジェクトの選択 , 208
  NULL 値 , 143, 214
オプション , 195-203
  保存 , 203
  リセット , 203
起動 , 194
強調表示されたオブジェクトに関する情報 , 208
「クラス判別」, 112
項目の選択 , 60
子ノードの選択 , 212
サンプルファイル , 215
終了 , 101
情報の取得 , 208
設定
  Tool Manager, 203
データのフィルタリング , 199, 210
データファイル , 194
デフォルト値の保存 , 203
デフォルト値のリセット , 195
必要なファイル , 194
ファイルの読み込み , 100, 101, 194
メニュー , 203
ツリー・ビジュアライザでの検索結果の例 , 208
ツリー・ビジュアライザの「移動」メニュー , 211
ツリー・ビジュアライザ (Tree Visualizer) のオプション , 195
ツリー・ビジュアライザの「検索」ダイアログ・ボックス (IRIX), 206
ツリー・ビジュアライザの「設定オプション」ダイアログ・ボックス , 196
ツリー・ビジュアライザの「選択」メニュー , 210
ツリー・ビジュアライザの表示メニュー , 211

て

「停止」ボタン , 14
定数コマンド , 181
程度 , 193
テーブル
  「クラス判別」, 108
  処理オプション , 68
  データのセーブ , 67
「テーブル処理」ウィンドウ , 6
「テーブルの履歴」ボタン , 187
テキストフィールド , 42
「適用」ボタン , 105
「テクスチャ」コマンド , 181, 182
「テストセットの誤差 / 損失」オプション , 84
「テストセットのバックフィット」オプション , 110, 38
「テスト属性」オプション , 83
「テスト値」オプション , 83
ディスク
  「色」オプション , 200
    対応付ける , 200
  高さ , 198
「データ型の変更」オプション , 68
データセット , 階層型
  サンプルファイル , 225
データのセーブ , 67
  レコードビューワ , 150
データの表示 , 13
「データ変換」パネル , 68
データファイル
  ツリー・ビジュアライザ
    サンプル , 283
  スキャタ・ビジュアライザ , 160
    サンプル , 277, 278
  スプラット・ビジュアライザ , 173
    サンプル , 280
  選択式決定木分析 , 144
  ツリー・ビジュアライザ , 194
    サンプル , 283
  マップ・ビジュアライザ , 127, 130

サンプル , 267
「データファイル」タブ , 67
「データファイル」パネル , 67
データベース
クラス判別 , 77, 91, 143
データの表示 , 126
3D ランドスケープ , 170, 193
アニメーション・コントロール・パネル , 11
特定の値の検索 , 83-84, 204-208
比較 , 198
標本抽出 , 109
フィルタリング , 199, 208-210
複数値の設定 , 135
保存 , 67
予測 , 43
混同マトリックス (Confusion Matrix), 64
データベース・サーバ
接続 , 191
データポイント , 16
集計処理 , 172, 173
多次元 , 12
データムーバ
接続 , 191
デフォルト , リセット
スキャタ・ビジュアライザ , 162
スプラット・ビジュアライザ , 177
ツリー・ビジュアライザ , 195
マップ・ビジュアライザ , 131
デフォルト値のリセット
スキャタ・ビジュアライザ , 162
スプラット・ビジュアライザ , 177
ツリー・ビジュアライザ , 195
マップ・ビジュアライザ , 131

と

投資利益率 (ROI) 曲線 , 110, 157
「糖尿病」診断データセット , 237, 253 ,265
特定の値の検索 , 204-208
「決定木」, 83-84
「閉じる」コマンド , 101
「閉じる 」ボタン
「検索」ダイアログ , 208
ドリルスルー
相関規則 , 32

な

「名前を付けて保存 」
レコードビューワ , 150
「名前を付けて保存」コマンド , 101

に

「乳癌 (Breast)」データセット , 235, 251, 263, 272

ね

ネットワーク接続 , 191

の

ノード , 193
selecting child, 212
「決定木」
情報の表示 , 78
ディスクの高さ , 198
特定のオブジェクトの検索 , 205
フィルタリング , 209
ベースの高さ , 198

は

配列
  スライダ , 164
  地理的位置 , 131
  分析 , 112, 113
発信点と受信点
  サンプルファイル , 268
反復型 k-means クラスタリング , 47
バー , 193
  「色」オプション , 200
    キーに基づく , 200
    対応付ける , 200
    ラベル , 202
  検索 , 205
  固定する , 201
  スケーリング , 198
  高さ , 198
    集計処理 , 199
  負の値 , 193
  ラベルの命名
    色 , 202
「バーのラベルカラー」オプション , 202
バックフィッティング・モデル , 38
「パス」スライダ , 15
パターンとトレンドの分析 , 23
パラメータ
  表示オプション , 211
パラメータの表示 , 211

ひ

比較
  データベース , 198
  文字列
    フィルタリング・タイプ , 104
ヒストグラム・ビジュアライザ , 106
  枝刈り係数 , 106
「左に移動」コマンド , 212
必要なファイル
  スキャタ・ビジュアライザ , 160
  スプラット・ビジュアライザ , 173
  選択式決定木分析 , 144
  ツリー・ビジュアライザ , 194
  マップ・ビジュアライザ , 127
  スキャタ・ビジュアライザ , 160
日付
  分析 , 112, 113
日付 , 西暦 2000 年問題に対応（Y2K 準拠）, 220
「等しい」検索オプション , 104, 206
非表示
  ラベル , 166, 176
「非表示」オプション , 209
非表示のラベル , 166, 176
「評価エラー」モード , 66, 109, 110
  出力の表示 , 112, 113
表現 (式), 2
「表現をチェック」ボタン , 4
「表示イメージの印刷 」コマンド , 101
表示
  アニメーション・コントロール・パネル , 217
  移動 , 188
  「決定木」, 112
  決定木ノード , 78
  現在 , 187
  スキャタ・ビジュアライザ
    表示オプション , 162, 166
  「スプラット」, 175
  選択されているオブジェクトの表示 , 135
  ツリー・ビジュアライザ
    移動 , 211
    強調表示されたオブジェクトに関する情報 , 208
    表示オプション , 195-203
  データ , 126, 170, 193
    アニメーション・コントロール・パネル , 11-16
  クラシファイアの出力 , 112
  ラベル , 166, 176

マップ・ビジュアライザ , 125, 126
表示オプション , 131-133
メッセージ , 135
スキャタ・ビジュアライザ , 165
ツリー・ビジュアライザ , 201
マップ・ビジュアライザ , 133
要素 , 163
「表示イメージの印刷」コマンド , 100
表示オプション
スキャタ・ビジュアライザ , 162-166
スプラット・ビジュアライザ , 176
ツリー・ビジュアライザ , 195-203
マップ・ビジュアライザ , 131-133
「表示 」メニュー , 204
「表示 」メニュー , 204
エビデンス・ビジュアライザ , 98
スキャタ・ビジュアライザ , 216
デシジョン・テーブル・ビジュアライザ , 76
標準偏差 , 84
標準モード , 59
「標本 」オプション , 69
「開く」コマンド , 100, 101
ビューの移動 , 188

ふ

「ファイル検索」ボタン , 131
フィルタリング
マップ , 104
ファイルの拡張子
ツリー・ビジュアライザ , 103
スプラット・ビジュアライザ , 174
ツリー・ビジュアライザ , 127, 194
ファイルの読み込み
クラスタ・ビジュアライザ , 52
スキャタ・ビジュアライザ , 161
ツリー・ビジュアライザ , 100, 101, 194
「ファイル」メニュー
「Web 公開用ファイルの作成」, 101
ファイルを開く
スキャタ・ビジュアライザ , 161
ツリー・ビジュアライザ , 100, 101, 194
フィールド名 , 2
「フィルタ」オプション , 68
「フィルタのスケール」コマンド , 104
「フィルタ」パネル
ツリー・ビジュアライザ , 208-210
マップ・ビジュアライザ , 104-105
「フィルタパネル」コマンド , 83, 104, 208
「フィルタ」ボタン , 103
フィルタリング
「決定木」, 83
データ , 199, 208-210
「深さ」スライダ , 210
複数値 , 135
複数値の設定 , 135
「含む」検索オプション , 206, 104
浮動小数点数 , 60
負の値 , 193
ブースティング , 42
「分割の下限値 」オプション , 81
「分割の基準 」オプション , 80, 183
分析 , 107-108
「エラー」オプション , 109-111
オプションの設定 , 110
クラスラベル , 108
実行 , 109
実行モード , 109
トラッキングの進行状況 , 112
「分析オプション」ダイアログ・ボックス , 95

へ

「平均誤差 / 損失の標準偏差 」オプション , 84

平均二乗誤差 , 156
平面 , 126
平面地図 , 268
並列化処理 , 146
並列処理 , 79, 144, 147
「ヘルプ」メニュー
  ツリー・ビジュアライザ , 213
編集 , 188
  色 , 56
変数
  フィルタとして , 104
米国の地図 , 130, 267
ベース , 193
  「色 」オプション , 200
    対応付ける , 200
    ライン , 202
    ラベル , 202
  スケーリング , 198
  選択 , 212
  ラベルの命名 , 202
「ベースでの実行」オプション , 201
「ベースでのラベルカラー」オプション , 202
ベースの選択 , 212
「ベースの高さ」コマンド , 211

ほ

ホーム位置 , 212
  設定 , 212
「ホーム」コマンド , 212
「ホームの設定」コマンド , 212
保険のサンプルファイル , 277
「補集合データのドリルスルー」コマンド , 168, 210
母集団の標本抽出 , 151
ボタン
  ツリー・ビジュアライザ
    「検索」ダイアログ・ボックス , 207

ま

「マークフラグ」コマンド , 211
マーケット・バスケット分析 , 24
「前処理を編集」ボタン , 188
「前」フィールド , 188
「前」ボタン , 207
マッピング
  NULL 値 , 167, 214
  相関規則 , 30
  地理的位置 , 130
  文字列 , 170
「マッピング」オプション , 132
「マッピングの種類」オプション , 200
マップ・ビジュアライザ , 269
  NULL 値 , 142
  色の割当て , 132
  オプション , 131-134
    起動 , 129
    保存 , 133
    リセット , 133
  起動 , 128
  起動オプション , 129
  サンプルファイル , 269
  設定
    Tool Manager, 129-134
  地理的位置 , 130
  データの表示 , 126
  データファイル , 127, 130
  デフォルト値の保存 , 134
  デフォルト値のリセット , 131
  必要なファイル , 127
  「フィルタパネル」, 104-105
  メインウィンドウ
    ラベルの命名 , 133
マップ・ビジュアライザの「ツールオプション」ダイアログ・ボックス , 131-133
「マトリックスの編集」ボタン , 125
マルチプロセッサ版 , 79, 144, 146, 147

み

「右に移動」コマンド, 212
未知の値の予測, 123

め

名称サイズの変更, 163
命名
  項目, 2
メイン, 178
メインウィンドウ
  エビデンス・ビジュアライザ, 98
  デシジョン・テーブル, 75
  統計量ビジュアライザ, 186
  マップ・ビジュアライザ
    ラベルの命名, 133
メッセージ, 135
  スキャタ・ビジュアライザ, 165
  ツリー・ビジュアライザ, 201
  マップ・ビジュアライザ, 133
「メッセージ」オプション, 133, 201
メニュー
  エビデンス・ビジュアライザ, 98-99, 143, 158
  スプラット・ビジュアライザ, 181
  ツリー・ビジュアライザ, 203
  デシジョン・テーブル・ビジュアライザ, 75
  統計量ビジュアライザ, 186

も

モードの選択
  エビデンス・ビジュアライザ, 96
文字列
  検索, 206
  比較, 104
  フィルタリング, 104
  マッピング, 170
モデル
  既存ファイルの読込み, 17
  選択, 17
  バックフィッティング, 38
  レコードの重み付け, 38
  レコードへの適用, 21-23
モデルの選択, 17
「モデルの適用」
  推定確率値モード, 18
  離散型予想値モード, 18
「モデルの適用」パネル, 18
「モデルの適用」ボタン, 17
「モデルのテスト」
  改善曲線, 20
  混同マトリックス, 20
「モデルのテスト」パネル, 19
「モデルへのデータ適合」, 20
「モデルへのデータ適合」モード, 20
「戻る」コマンド, 212

よ

要素
  NULL 値, 167
  「色」オプション, 163
  サイズ, 163
  選択, 165
  表示, 163
  ラベルの命名, 163
「要素」オプション, 163
「要素の色」オプション, 163
「要素の形状」オプション, 163
「要素のサイズ」オプション, 163
「要素の説明オン」オプション, 163
要素の選択, 165
「要素のラベルカラー」オプション, 163
「要素ファイル」フィールド, 131

「要素名のサイズ」オプション , 163
要約値 , 193
ヨーロッパの地図 , 130, 267
予算 , 283
予測 , 43, 92
  混同マトリックス , 64
予備比率 , 110
予備法による分析 , 88, 110

ら

ラインの色 , 202
「ラインの色」オプション , 202
「ラプラス補正」オプション , 93
「ラベル確率順 」コマンド , 77, 141
ラベル
  「色」オプション
    バー , 202
    ベース , 202
  距離 , 166, 176
  サイズの変更 , 163
  軸 , 166, 176
  「スプラット」, 176
  バー
    色 , 202
  分析 , 84, 108
  ベース , 202
  メインウィンドウ , 133
  要素 , 163
「ラベルを隠す距離」オプション , 166, 176
ランダムシード , 110
ランダムに選択したサンプル , 109
ランドスケープ , 170
ランドスケープ *, 193

り

離散化アルゴリズム , 61
離散型ラベル , 84
「離散型ラベル」メニュー , 84
離散型予想値モード , 18
「離散」的な色の設定 , 164, 176, 200
離散的な属性 , 61, 84
「離散」的な色の設定 , 132
「リセット」ボタン , 133, 166, 176, 203
履歴ウィンドウ , 188
「履歴の表示」ボタン , 188

る

「ループ」ボタン , 15

れ

レコード
  「クラス判別」, 38
  クラスの割当て , 77, 91, 143
  属性の数 , 87
  モデル , 21-23
  ラベルがない , 87
レコードの重み付け , 38, 111, 151
  相関規則 , 28
レコードビューワ , 149
  「概要 (Overview)」, 149
  起動 , 149
  行番号の振り直し , 150
  検索 , 150
  データのセーブ , 150
  「名前を付けて保存」, 150
レコードビューワの起動 , 149
「レベル」オプション , 84
連続型のカラーオプション設定 , 132, 164, 200, 176

ろ

ロケール
  リソースファイル , 115
    例 , 116
ロケール , 国際化 , 114, 115
ロケールの設定 , 114, 115
「論理積」演算 , 105, 207
「論理和」演算 , 105, 207

わ

? ワイルドカード , 104
ワイルドカード
  ツリー・ビジュアライザ , 206
  マップ・ビジュアライザ , 104