

MineSet™ 3.0 Enterprise Edition User's Guide for Windows®

ドキュメント番号 007-4005-001JPN

編集協力者

執筆：Sandra Motroni、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

© 2000 Silicon Graphics, Inc.— All Rights Reserved

本書の内容の一部あるいは全部について（ソフトウェアを含む）、Silicon Graphics 社から事前に文書による明確な許諾を得ず、いかなる形態においても複写、複製することは禁じられております。

LIMITED AND RESTRICTED RIGHTS LEGEND

Use, duplication, or disclosure by the Government is subject to restrictions as set forth in the Rights in Data clause at FAR 52.227-14 and/or in similar or successor clauses in the FAR, or in the DOD, DOE or NASA FAR Supplements. Unpublished rights reserved under the Copyright Laws of the United States. Contractor/ manufacturer is Silicon Graphics, Inc., 1600 Amphitheatre Pkwy., Mountain View, CA 94043-1351.

Silicon Graphics は Silicon Graphics, Inc. の登録商標であり、SGI、MineSet および SGI のロゴは Silicon Graphics, Inc. の商標です。Oracle は、Oracle Corporation の登録商標です。Excel、Windows および Windows NT は Microsoft Corporation の登録商標です。Microsoft SQL Server は Microsoft Corporation の商標です。MATLAB は The Matchworks, Inc. の商標です。SPSS は SPSS, Inc. の登録商標です。DBMS/COPY は Conceptual Software, Inc. の商標です。

ツリー・ビジュアライザ (Tree Visualizer) は、米国特許番号 5,528,735、5,555,354、5,671,381、5,861,885 で特許を取得しています。スプラット・ビジュアライザ (Splat Visualizer) は、米国特許番号 5,861,891 で特許を取得しています。マップ・ビジュアライザ (Map Visualizer)、スキヤタ・ビジュアライザ (Scatter Visualizer)、およびスプラット・ビジュアライザ (Splat Visualizer) の 2D スライドについては特許出願中です。エビデンス・ビジュアライザ (Evidence Visualizer)、デシジョン・テーブル (Decision Table)、およびスプラット・ビジュアライザ (SplatViz) のアニメーションについては特許出願中です。

MineSet™ 3.0 Enterprise Edition User's Guide for Windows®

ドキュメント番号 007-4005-001JPN

目次

図目次 xi

表目次 xv

このマニュアルについて xvii

対象読者 xvii

MineSet 情報について xvii

このマニュアルの構成 xviii

このマニュアルの図 xix

参考文献 xx

表記上の決まり xx

1. データマイニングと MineSet ツールの概要 1

データマイニングの用語 1

データマイニングとデータマイニング手法について 2

データマイニング 3

教師付き (Supervised) モデリング 3

教師なし (Unsupervised) モデリング 5

データの可視化 6

データマイニングのための MineSet ツール 7

データマイニング・プロセスの概要 8

データの識別 9

データの準備 9

モデルの構築 11

モデルの評価 11

モデルの展開 11

2. MineSet によるデータのアクセス 13

MineSet が格納されているディレクトリ 13

MineSet ソフトウェアの動作 14

MineSet とユーザ・アプリケーション	15
MineSet の起動	15
「 <i>Tool Manager</i> 」ウィンドウの使い方	18
レコードビューワ (Record Viewer) による元データの表示	19
レコードビューワの項目の変更	20
レコードビューワでのデータのフィルタリング	21
レコードビューワでのデータの保存	22
統計量ビジュアライザ (Statistics Visualizer) によるレコード統計の表示	22
ボックスプロットとは	22
ヒストグラムとは	23
統計量ビジュアライザの起動	24
サンプル・データファイルの使用	25
ヒストグラム・ビジュアライザの起動	25
MineSet 3D ビジュアライザでのナビゲーション	27
ツリー・ビジュアライザでのナビゲーション	27
非ツリー・ビジュアライザでのナビゲーション	29
ヘルプ情報	32

3. データの整形 33

データの整形の必要性	33
Tool Manager によるデータの変換	34
項目の削除と追加	35
項目に対する階級の変更または新規作成	37
集計処理による新規項目の作成	41
フィルタによる項目内容の限定	43
項目の型や名前の変更	44
モデルの適用	47
データの標本抽出	47
テーブル履歴による手順のトレース	48
レコードの重み付け	50
重要項目の検索	51

4.	スキャタ・ビジュアライザとスプラット・ビジュアライザによるデータの分析	55
	スキャタ・ビジュアライザとスプラット・ビジュアライザの概要	55
	スキャタ・ビジュアライザの概要	56
	スプラット・ビジュアライザの概要	57
	スキャタ・ビジュアライザやスプラット・ビジュアライザのためのデータ変換	60
	スプラット・ビジュアライザの処理技法	62
	スキャタ・ビジュアライザとスプラット・ビジュアライザの起動	64
	スキャタ・ビジュアライザでの可視化要素のマッピング	64
	スプラット・ビジュアライザでの可視化要素のマッピング	66
	スキャタ・ビジュアライザとスプラット・ビジュアライザでの結果の表示	69
	表示モード	70
	スキャタ・ビジュアライザのスライドの作成	74
	スキャタ・ビジュアライザとスプラット・ビジュアライザでのアニメーションの作成	75
	ビジュアライザ・サマリウィンドウを使ったアニメーション	75
	スキャタ・ビジュアライザでのアニメーション軌跡の表示	76
	スプラット・ビジュアライザでのアニメーション	79
	スプラット・ビジュアライザでのサマリウィンドウの解釈	80
	スキャタ・ビジュアライザとスプラット・ビジュアライザでの結果の操作	82
	表示の変更	82
	スキャタ・ビジュアライザの選択とドリルスルー	84
	スプラット・ビジュアライザの選択とドリルスルー	85
	「形状」メニューによるスキャタ・ビジュアライザ表示の変更	87
	「形状」メニューによるスプラット・ビジュアライザ表示の変更	87
5.	ツリー・ビジュアライザによるデータの表示	89
	ツリー・ビジュアライザ (Tree Visualizer) の概要	89
	ツリー・ビジュアライザの起動	90
	ツリー・ビジュアライザでの可視化要素のマッピング	92

- ツリー・ビジュアライザによる結果の表示 94
 - ツリー・ビジュアライザの例 95
 - ツリー・ビジュアライザでデータの詳細を見るには 98
 - 「概観 (Overview)」による階層全体の表示 101
 - 検索パネルによる特定オブジェクトの検索 102
 - 「マーク」パネルによる重要な場所のフラグ付け 103
 - 「フィルタ」パネルによるデータのフィルタリング 106
 - ツリー・ビジュアライザ表示の微調整 107
- 6. マップ・ビジュアライザによるデータの表示 109**
 - マップ・ビジュアライザの概要 110
 - マップ・ビジュアライザの起動 113
 - データの準備 114
 - データの集計処理 114
 - マップ形状の選択 115
 - マップ・ビジュアライザの可視化要素のリンク 115
 - マップ・ビジュアライザの説明 120
 - 表示モード 121
 - マップ・ビジュアライザでのアニメーションの作成 123
 - マップ・ビジュアライザでの結果の操作 124
 - マップ・ビジュアライザの表示の変更 124
 - 選択とドリルスルー 125
- 7. 予測モデリングの理解 127**
 - 予測モデリングの概要 127
 - モデルの生成 128
 - エビデンス・モデル (Evidence Model) 128
 - 決定木モデル (Decision Tree Model) 130
 - 選択式決定木モデル (Option Tree Model) 131
 - デシジョン・テーブル・モデル (Decision Table Model) 133
 - 回帰ツリーモデル (Regression Tree Model) 136

予測モデルの評価	137
すべてのデータを使用したクラス判別	137
予備法による誤差推定	139
相互検証の誤差推定	140
学習曲線 (Learning Curve) の作成	141
予測モデルの適用	145
モデルの選択	145
モデルの適用	146
この後の学習順序	147
8. 決定木、選択式決定木、回帰ツリーによるモデリングと予測	149
決定木、選択式決定木、回帰ツリーの概要	150
決定木分析 (Decision Tree)	150
選択式決定木 (Option Tree)	151
回帰ツリー分析 (Regression Tree)	154
決定木、選択式決定木、回帰ツリーの起動	155
決定木ビジュアライザによる結果の表示	156
決定木のメインウィンドウを使ったレコードのクラス判別	157
その他の便利なオプション	160
選択式ツリー・ビジュアライザによる結果の表示	160
回帰ツリー・ビジュアライザによる結果の表示	161
回帰ツリーのメインウィンドウを使った値の予測	162
回帰ツリーにおける誤差推定	162
その他の便利なオプション	163
決定木、選択式決定木、回帰ツリーによる予測	163
9. デシジョン・テーブル・クラシファイアとビジュアライザによるモデリングと予測	165
デシジョン・テーブル・クラシファイアの概要	165
デシジョン・テーブルの起動	167
デシジョン・テーブル・ビジュアライザによる結果の表示	169
「デシジョン・テーブル」パネルの説明	170
「ラベル確率」パネル	171
デシジョン・テーブルの例	171
デシジョン・テーブルによる予測	174

10.	エビデンス・クラシファイアとビジュアライザによるモデリングと予測	175
	エビデンス・クラシファイアとビジュアライザの概要	175
	「エビデンス・ビジュアライザ」ウィンドウ	176
	エビデンス・ツールの起動	178
	エビデンス・ビジュアライザによる結果の表示	179
	エビデンス表示	180
	確率表示	183
	バー表示	184
	エビデンス・ビジュアライザ表示の変更	186
	エビデンス・クラシファイアによる予測	187
11.	予測モデリングの改良	189
	モデルの精度の検査	189
	モデルのテスト	189
	モデルへのデータの適合	192
	バックフィッティング (backfitting) による誤差の評価	194
	ブースティング (boosting) による精度の向上	194
	混同マトリックス (Confusion Matrix) と損失マトリックス (Loss Matrix) によるモデルの 微調整	196
	混同マトリックスによる誤差の調査	196
	混同マトリックスの表示	198
	損失マトリックスの定義	201
	改善曲線 (<i>Lift Curve</i>) と ROI 曲線 (ROI Curve) によるモデルの評価	204
	改善曲線による予測の表示	204
	投資利益率曲線の使用によるマーケティング・メリットの発見	206
12.	クラスタリングによるデータのセグメント化	211
	クラスタリングの概要	211
	Tool Manager によるクラスタリングの起動	212
	サンプル・クラスタリング・ファイルの使用	214
	クラスタ・ビジュアライザのメインウィンドウ	215
	クラスタリングを可視化する別の方法	216

13.	相関規則分析によるデータの分析	217
	相関規則分析 (Association Rule Generator) と可視化の概要	217
	相関規則分析	218
	規則の可視化	219
	相関規則分析の起動	221
	相関の設定	221
	レコードの重み付け	223
	相関規則の属性と可視化要素のマッピング	223
	スキャタ・ビジュアライザによる相関規則の説明	224
	ドリルスルー	227
	多重相関規則	227
	MineSet ユーザの用語集	231
	索引	239

図目次

図 1-1	サンプル・データテーブル 3
図 1-2	データマイニング・プロセス 8
図 2-1	Tool Manager のログイン・ウィンドウ 16
図 2-2	「データファイルを開く」ウィンドウ 17
図 2-3	「Tool Manager」ウィンドウ 18
図 2-4	レコードビューワによる表示 19
図 2-5	レコードビューワのフィルタパネル 21
図 2-6	統計量ビジュアライザで表示した数値型項目 23
図 2-7	統計量ビジュアライザで表示した離散型項目 24
図 2-8	統計量ビジュアライザが選択された「データの可視化／マイニング」 パネル 25
図 2-9	ヒストグラム・ビジュアライザが選択された「データの可視化／マイ ニング」パネル 26
図 3-1	「データの変換」パネル 34
図 3-2	「項目の追加」ダイアログ・ボックス 36
図 3-3	「階級の生成」ダイアログ・ボックス 38
図 3-4	階級生成を行う詳細オプションパネル 39
図 3-5	「集計処理」ダイアログ・ボックス 42
図 3-6	「フィルタ (Filter)」ダイアログ・ボックス 44
図 3-7	データ型のポップアップリスト 45
図 3-8	「標本」ダイアログ・ボックス 47
図 3-9	「履歴の表示」ダイアログ・ボックス 49
図 3-10	「重要項目」タブ 51
図 3-11	「重要項目」の「詳細モード」 53
図 4-1	スキャタ・ビジュアライザの例 56
図 4-2	1次元サマリスライダを設定したスプラット・ビジュアライザの 例 58
図 4-3	スキャタ・ビジュアライザが選択された「データの可視化／マイニン

	グ」パネル 65
図 4-4	スプラット・ビジュアライザでの可視化要素のマッピング 67
図 4-5	マッピング用の整数型項目の追加 69
図 4-6	選択された要素情報の表示 71
図 4-7	データ上でのピックドラッグの操作 73
図 4-8	スキャタ・ビジュアライザのチューブ・モーション軌跡の例 78
図 4-9	スプラット・ビジュアライザのサマリウィンドウとスライダ・コントロール 79
図 4-10	スライダの移動による可視化の変化(図 4-2 と比較) 81
図 4-11	スキャタ・ビジュアライザとスプラット・ビジュアライザの表示メニュー 82
図 4-12	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「フィルタ・ダイアログ」 83
図 4-13	スキャタ・ビジュアライザの「選択」メニュー 84
図 4-14	スプラット・ビジュアライザの「選択」メニュー 85
図 5-1	ツリー・ビジュアライザのメインウィンドウでの表示例 90
図 5-2	ツリー・ビジュアライザが選択された Tool Manager の「データの可視化／マイニング」パネル 92
図 5-3	ツリー・ビジュアライザによる店舗データセットの初期画面 96
図 5-4	オブジェクトとその元になっている情報の強調表示 99
図 5-5	選択された(スポットライトが当てられた)オブジェクトの例 101
図 5-6	ツリー・ビジュアライザの「概観」ウィンドウ 102
図 5-7	ツリー・ビジュアライザでの検索結果の例 103
図 5-8	ツリー・ビジュアライザの「マーク」パネル 104
図 5-9	「マークの作成」ダイアログ・ボックス 104
図 5-10	マークを表すフラグが表示されたメインウィンドウ 105
図 6-1	地理形状を表示したマップ・ビジュアライザの例 110
図 6-2	人口の比較をバーグラフで地図上に表示したマップ・ビジュアライザの例 111
図 6-3	米国の2地点間を結ぶ線を表示したマップ・ビジュアライザの例 112
図 6-4	マップ・ビジュアライザによる集計処理 114
図 6-5	マップ・ビジュアライザでの項目と可視化要素のマッピング 116
図 6-6	マップ・ビジュアライザの「オプション」ダイアログ・ボックス 119
図 6-7	Year スライダを 1990 年に移動したときの population.use.mapviz の

	例	121
図 7-1	エビデンスによって生成された churn データセットのエビデンスの可視化	129
図 7-2	決定木によって生成された churn データセットの決定木モデル	130
図 7-3	選択式決定木によって生成された cars データセット	132
図 7-4	デシジョン・テーブルによって生成された mushroom データセットのデシジョン・テーブル・モデル	133
図 7-5	回帰ツリー分析によって生成された adult データセットの回帰ツリー分析モデル	135
図 7-6	アヤメの分類の間違いの例	139
図 7-7	「誤差推定のオプション」の予備比率	140
図 7-8	「誤差推定のオプション」の相互検証オプション	141
図 7-9	ラベルを churned に設定した churn データセットの学習曲線	143
図 7-10	学習曲線のオプション	144
図 7-11	「モデルのテストと適用」ウィンドウ: クラシファイアの選択	146
図 7-12	「モデルの適用」パネル	147
図 8-1	cars データセットの決定木モデルの可視化	151
図 8-2	cars データセットの選択式決定木モデルの可視化	153
図 8-3	adult データセットの回帰ツリーの可視化	154
図 8-4	Tool Manager の「データの可視化／マイニング」パネルと「クラス判別」タブ	156
図 8-5	iris データセットの決定木	159
図 9-1	mushroom データセットのデシジョン・テーブル	166
図 9-2	「クラス判別」タブを表示した Tool Manager の「データの可視化／マイニング」パネル	167
図 9-3	mushroom データセットのデシジョン・テーブルの可視化	169
図 9-4	ドリルダウンした mushroom データセットのデシジョン・テーブル	172
図 9-5	ドリルダウンしたデシジョン・テーブルの拡大	173
図 10-1	iris データセットを使用したエビデンス・ビジュアライザ	176
図 10-2	確率を表示したエビデンス・ビジュアライザ	177
図 10-3	Tool Manager の「データの可視化／マイニング」パネル	179
図 10-4	エビデンス・ビジュアライザによる churn データセットのケーキグラフ	180

図 10-5	エビデンス・ビジュアライザで表示した mushroom データセット	182
図 10-6	エビデンス・ビジュアライザの円グラフ	184
図 10-7	cars データセットでラベルの値「Japan」を選択	185
図 11-1	「モデルのテスト」パネル	191
図 11-2	「モデルへのデータ適合」パネル	193
図 11-3	mushroom データセットの混同マトリックス	197
図 11-4	「混同マトリックスの表示」が選択された「クラシファイアオプション」パネル	199
図 11-5	間違ったクラス判別を示す mushroom データセットの混同マトリックス	200
図 11-6	損失マトリックスの編集パネル	201
図 11-7	mushroom データセットの混同マトリックスと損失マトリックス	203
図 11-8	改善曲線	205
図 11-9	ROI 曲線	208
図 12-1	「クラスタ」タブ	212
図 12-2	反復型 k-means を使ったクラスタリング	213
図 12-3	クラスタ・ビジュアライザのメインウィンドウ	214
図 13-1	相関規則ビジュアライザのメインウィンドウの詳細表示	220
図 13-2	相関規則分析の最初の「Tool Manager」ウィンドウ	222
図 13-3	相関規則のマッピング	223
図 13-4	brand.rules.scatterviz を指定したときの最初の相関規則の表示	225
図 13-5	相関規則を表すバーの上にカーソルがある状態	226
図 13-6	多重相関生成を設定する最初の「Tool Manager」ウィンドウ	228

表目次

表 2-1	レコードビューワの項目に対する操作	20
表 2-2	ツリー・ビジュアライザ (Tree Visualizer) のナビゲーション・ボタン	27
表 2-3	ツリー・ビジュアライザの調整スライダとダイヤル	28
表 2-4	非ツリー・ビジュアライザのナビゲーション・ボタン	29
表 2-5	非ツリー・ビジュアライザの調整スライダとダイヤル	31
表 3-1	Tool Manager の「データの変換」パネルのボタン	35
表 3-2	階級生成を行う詳細オプション	39
表 3-3	項目データ型の意味	46
表 3-4	テーブル履歴の意味	48
表 4-1	スプラット・ビジュアライザの可視化要素にマッピングできる項目型	61
表 4-2	スキャタ・ビジュアライザの可視化要素にマッピングできる項目型	61
表 4-3	スキャタ・ビジュアライザでの可視化要素のマッピング	66
表 4-4	スプラット・ビジュアライザでの可視化要素のマッピング	68
表 4-5	スキャタ・ビジュアライザとスプラット・ビジュアライザの表示メニュー選択肢	82
表 5-1	ツリー・ビジュアライザの可視化要素	93
表 5-2	店舗データのマッピング例	97
表 5-3	ツリー・ビジュアライザの表示設定パラメータ	107
表 6-1	ドリル方法	122
表 6-2	マップ・ビジュアライザの表示メニューの項目	124
表 6-3	マップ・ビジュアライザの選択メニューの項目	125
表 7-1	学習曲線に対するアクション	145
表 10-1	エビデンス・ビジュアライザ表示の変更	186
表 11-1	「モデルのテスト (Test Model)」パネルのオプション	191
表 11-2	「モデルへのデータ適合」オプション	193

表 12-1	クラスタリング方法のオプション	212
表 13-1	相関規則の構成要素	219
表 13-2	相関規則と可視化要素のマッピング	224

このマニュアルについて

『*MineSet 3.0 Enterprise Edition User's Guide*』では、MineSet のマイニングツールや可視化ツールの特徴と機能について説明します。MineSet 製品の最新情報は、MineSet の Web ページ（<http://mineset.sgi.com>、日本語版 <http://mineset.sgi.co.jp>）を参照してください。

対象読者

MineSet 製品およびこのマニュアルは、データマイニングの専門家でなくても使用できます。しかし、データやその意味を理解していれば、結果を解釈する上で助けになります。データマイニング技法にすでに精通している場合でも、MineSet アルゴリズムがどのように働き、その適用や可視化がどのように行われるかをこのマニュアルから知ることができます。

Tool Manager を使ってデータベースから MineSet ツールにデータをインポートする場合は、主に『*MineSet 3.0 Enterprise Edition Interface Guide*』を参照して関連する操作を行ってください。データが可視化ツールにすでにロードされている場合には、データベースやプログラミングの知識は必要ありません。

MineSet 情報について

このマニュアルは MineSet の実行に必要な操作を説明しています。したがって、このマニュアルのほとんどの部分は、さまざまなツールの使い方の説明に当てられています。「このマニュアルの構成」に各章の要約を記載しています。

技術的な詳細や詳しい説明については、『*MineSet 3.0 Enterprise Edition Reference Guide*』を参照してください。特定の MineSet アルゴリズムの特に詳しい情報については、http://www.sgi.com/software/mineset/mineset_data.html を参照してください。

MineSet をほかの場所やアプリケーションにエクスポートしたり、コマンド行を使用する場合のように、システム管理に関連することを行う場合は、『*MineSet 3.0- Enterprise Edition Interface Guide*』を参照してください。

さらに、MineSet では、AcPro などのサードパーティ製品をアプリケーションにプラグインできます。

このマニュアルの構成

最初の 3 つの章はデータマイニングと MineSet 製品の概要です。それ以後の各章では、個々のツールや手順を説明しています。

第 1 章「データマイニングと MineSet ツールの概要」

この章では、データマイニングの原理について簡単に説明し、用語を説明し、さらに、MineSet ツールによる分析や可視化に必要な手順の概要を説明します。

第 2 章「MineSet によるデータのアクセス」

この章では、MineSet の起動方法と、基本的なツールを使ってデータを調べる方法について説明します。

第 3 章「データの整形」

この章では、Tool Manager を使ってデータを変換する必要性とその手順について説明します。さらに、データに潜む特性を探索するための重要項目ツールやクラスタリング・ツールについても説明します。

第 4 章「スキャタ・ビジュアライザとスプラット・ビジュアライザによるデータの分析」

この章では、スキャタ・ビジュアライザとスプラット・ビジュアライザのインタフェースについて説明します。これらのツールは、多次元データを統計的に見たり、可視化するときに役立ちます。

第 5 章「ツリー・ビジュアライザによるデータの表示」

この章では、ツリー・ビジュアライザのインタフェースについて説明します。このツールは、階層データを見るときに役立ちます。

第 6 章「マップ・ビジュアライザによるデータの表示」

この章では、マップ・ビジュアライザのインタフェースについて説明します。このツールは、データを地理的または空間的に見るときに役立ちます。

第 7 章「予測モデリングの理解」

この章では、記述モデリングに対するものとしての予測モデリングと、MineSet のさまざまなクラシファイアについて説明します。

第 8 章「決定木、選択式決定木、回帰ツリーによるモデリングと予測」

この章では、決定木モデル、選択式決定木モデル、回帰ツリーの生成方法と使用方法について説明します。これらのツールは、いくつかの属性を使ってデータをクラス判別し、それらの属性に基づいて一連の決定を下すときに役立ちます。選択式決定木は、一度に複数の属性に及ぼす分割の影響度を示します。回帰ツリーは、実生活で扱うような連続値に基づいて属性を予測するときに有益です。

第 9 章「デシジョン・テーブル・クラシファイアとビジュアライザによるモデリングと予測」

この章では、デシジョン・テーブルのインタフェースについて詳しく説明します。このツールは、データをクラス判別してクラシファイアを作成して下した決定事象を可視化するときに有益です。

第 10 章「エビデンス・クラシファイアとビジュアライザによるモデリングと予測」

この章では、エビデンス・クラシファイアの生成方法と使用方法について説明します。このツールは、ある属性に基づいて、データをクラス判別し、指定した結果が起る確率を調べるときに役立ちます。

第 11 章「予測モデリングの改良」

この章では、混同マトリックス、損失マトリックス、ROI 曲線、改善曲線の目的と、これらのツールを使って予測モデリングと実生活の事象を関連付ける方法について説明します。

第 12 章「クラスタリングによるデータのセグメント化」

この章では、クラスタ・ビジュアライザのインタフェースについて説明します。このツールは、データを調べ、クラスタリング・パターンを発見するときに役立ちます。

第 13 章「相関規則分析によるデータの分析」

この章では、相関規則ビジュアライザについて説明します。このツールは、大きいデータセットからデータの相関を発見するときに役立ちます。

MineSet ユーザの用語集

このマニュアルで頻繁に使用されている技術的な用語の意味を説明します。

このマニュアルの図

このマニュアルの印刷物ではすべてのスクリーン・ショットや図が白黒で示されていますが、オンライン版ではフルカラーで表示されます。特定のグラフィックやスクリーン・ショットが印刷物では分かりにくい場合は、オンライン版を参照してください。

参考文献

データマイニングや関連する手法の一般的な説明については、次の書籍を参照してください。

- 『*Data Mining Techniques for Marketing, Sales, and Customer Support*』 Michael Berry および Gordon Linoff 著、John Wiley & Sons 社
- 『*Data Mining Solutions: Methods and Tools for Solving Real World Problems*』 Christopher Westphal および Teresa Braxton 著、John Wiley & Sons 社

データマイニングの技術的な詳細について興味がある場合は、『*MineSet 3.0 Enterprise Edition Interface Guide*』の付録 A の参考資料を参照してください。

表記上の決まり

このマニュアルでは、次の表記法を用いています。

『 』	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。
イタリック体	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。

ほかのマニュアルへのリンクや、アプリケーションなどの実行可能な語句は赤く表示されます。

本書のほかの章、節、または図などへのリンクは青く表示されます。

データマイニングと MineSet ツールの概要

この章では、データマイニングと MineSet の使い方について概要を説明します。MineSet の知識がすでにある場合は、この章を読まなくてもかまいません。この章の構成は次のとおりです。

- 「データマイニングの用語」(1 ページ)
- 「データマイニングとデータマイニング手法について」(2 ページ)
- 「データマイニング」(3 ページ)
- 「データの可視化」(6 ページ)
- 「データマイニングのための MineSet ツール」(7 ページ)
- 「データマイニング・プロセスの概要」(8 ページ)

MineSet はデータ分析用の統合ツールセットです。MineSet のグラフィカル・ユーザ・インタフェースは、マーケティングや販売の専門家、財務アナリスト、保険業者など、データ分析を必要とする人ならだれでも使用できます。さらに、アプリケーション開発者がビジネスや科学技術用途の意思決定支援アプリケーションを作成したい場合にも、MineSet のインタフェースを使用できます。

データマイニングの用語

基本的なデータマイニング用語の意味については、この節とこのマニュアルの最後の用語集で概略を説明します。データマイニングでは、データはレコードの集まりです。解析（マイニング）の対象となるデータは、顧客への課金データ、医薬品の試験データ、POS システムの取引データといった業務処理や学術的処理から入手します。これらのレコードは、MineSet 用に行と項目からなるテーブル形式に編成されます。テーブルには、特定企業のすべての顧客を表すレコードという意味で、たとえば「顧客」という名前を付けることができます。データマイニングでは、項目を属性と呼んだり、行をインスタンスやレコードと呼ぶことがあります。

MineSet ではアルゴリズムや数式を使用して分析を行います。モデル自体は、データの行や項目を分析するのでクラシファイアと呼ばれます。したがって、行や項目からなるレコード内容を取り出し、いろいろな方法でソートまたはクラス判別ができます。

データマイニングとデータマイニング手法について

データマイニングの目的は、データに潜むパターンを発見し、その知識を利用して問題を解決することです。データマイニングを使って解決する典型的な問題には次のものがあります。

- 不正な検出
- 解約分析
- 通話パターン分析
- ターゲット・マーケティング
- マーケティング・セグメンテーションの決定
- 運用手順の改善
- 医療サービスの改善
- マーケット・バスケット分析

クエリやオンライン分析処理 (OLAP: Online Analytic Processing) を使ってデータベースからデータを取り出す場合には、データ要素間の関係を直接指定する必要があります。たとえば、クエリを使ってすべての売上げを地域別に取り出すとします。この方法では、売上げが地域によって異なることを知っていなければなりません。したがって、その仮説をデータに照らし、妥当性の真偽を検証する必要があります。

OLAP が発見指向のデータマイニングの方向へ向かうのは、ユーザが存在することを知らない場合の関係性を網羅できないという点です。この方法では、データ自身が発見者に結論を示唆することのみ許されます。

データの可視化ではデータが啓発的なビジュアル形式で表現されるため、可視化されたモデルを見るだけで傾向や特徴を知ることができます。複雑なデータパターンを可視化し、検討できることは、マーケティングの意思決定にとって計り知れない利益になることがあります。このような可視化は、測定や定量化された既存のデータを説明するという意味で「説明的」といえます。

データマイニングでは、まず、アルゴリズムを使ってデータからモデルを自動的に抽出します。次に、選択されているアルゴリズム手法によっては、これらのアルゴリズムモデルをさまざまな方法でデータに適用します。その後、データ解析の結果を可視化したり、それをまだ解析していないデータでテストしたり、モデルをまったく新しいデータセットに適合したりします。解析データマイニングでは、このモデルを使って、新しく提示されるデータの特性を予測します。

データマイニングでは、データからモデルを作成します。このときに、「教師付きモデリング」と「教師なしモデリング」という、広い意味で2つの方法が使用されます。

MineSet では、可視化とマイニングモデルを組み合わせて使用の方がそれぞれの技法を個別に使用するよりもより深い分析結果が得られます。

データマイニング

データマイニングでは、データは項目と行からなるテーブル（表）として表します。さらに、データレコードはテーブルの行で、データ属性は項目でそれぞれ表します（[図 1-1](#)を参照）。

記述属性					ラベル
	年齢	未婚 / 既婚 / 離婚	性別	勤務年数	クレジットリスク
レコード 1	35	D	M	3	> 60
レコード 2	45	M	F	19	< 60
レコード 3	29	S	M	45	> 60

図 1-1 サンプル・データテーブル

教師付き (Supervised) モデリング

教師付きモデリングの目的は、レコードのある項目（属性）の内容または値を同じレコードの他の項目（属性）の値に基づいて予測することです。つまり、目的は特別な属性または特性を予測することです。この属性をラベルと呼びます。教師付きモデリング・アルゴリズムを使って作成したモデルでは、ラベルとその他の属性の相関を指定することによって、ラベルのない新しいデータを予測します。

次に、モデルを可視化すれば、ラベルと他の属性の相関を知ることができます。たとえば、顧客が他社に移った（解約した）場合には、モデルを構築して、どの顧客が解約する可能性があるかを予測したり、解約に至った理由やパターンを理解できます。

最もよく使われる教師付きモデリングタスクにはクラス判別 (classification) タスクと回帰 (regression) タスクがあります。ラベルが、yes や no などの固定値からなる離散型の

場合はクラス判別タスクと呼びます。ラベルが、収入や株価など、連続した範囲の値を持つ連続型データの場合は回帰タスクと呼びます。

クラス判別 (Classification) タスク

クラス判別タスクとは、ラベルのないレコードに離散型ラベルの値を割り当てる操作です。このとき、レコードはあらかじめ決められたグループに分割されます。たとえば、簡単なクラス判別では、顧客請求レコードが2つのクラスにグループ化されます。一方は60日以内に支払う顧客、他方は支払いが60日より長くなる顧客です。あるいは、顧客を性別や収入で分割します。クラス判別では、ラベルが特定の値をとる確率も予測できます。たとえば、ある人が60日以内に支払う確率を計算できます。

クラス判別タスクでは、MineSet のアルゴリズムを使ってデータからモデルを生成します。MineSet では訓練事例からクラシファイアを自動的に生成できます。訓練事例は、元のデータからランダムに抽出したサブセットです。たとえば、データの40%だけを使ってモデルを作成するとします。モデルを生成したら、それを使ってラベルの値が分かっていないレコードのクラス確率をクラス判別または予測できます。たとえば、テーブルに新しいレコードを追加するときに、ラベルの値が分かっていないことがあります。その場合には、モデルを使ってラベル項目の値を予測できます。後でラベルの値が予測値と同じであることが分かれば、このモデルは正しいことになります。生成したモデルが正しいかどうかを検証する場合には、元のデータセットのうち訓練事例に使用しなかった部分を使用できます。

MineSet では4つのクラス判別モデルを作成できます。モデルは、ビジュアルライザで表示できます。これらのモデルの違いは、アルゴリズムがデータのソートや分析をどのように行うかにかかっています。たとえば、決定木 (Decision Tree) では、通話レコードを通話が256分未満と256分以上の顧客にグループ化できます。選択式決定木 (Option Tree) では、解約しそうな顧客を3つの選択肢を使って探索することができます。これらの違いについては、「[データマイニングのための MineSet ツール](#)」(7 ページ) で詳しく説明します。例は、この後の各章で示します。

回帰 (Regression) タスク

回帰タスクは、ラベルの値が離散型ではなく連続型であることを除けば、クラス判別タスクに似た教師付きモデリングタスクです。たとえば、給料や株価を予測することは回帰タスクですが、給料がいくつかの範囲（たとえば、\$60K 未満、\$60K から \$120K まで、\$120K 超など）に入っているかどうかや、株価の騰落を予測することはクラス判別タスクです。

MineSet では、訓練事例やサンプルデータから回帰ツリー分析モデルを自動的に生成できます。回帰ツリー分析モデルが生成されたら、それを使って連続属性の値を予測できます。たとえば、給料が \$60,420 と予測するような場合です。MineSet の回帰モデルは、現在、回帰ツリー (Regression Tree) だけです。MineSet で作成する他のモデルと同じように、回帰ツリーも MineSet で可視化できます。

モデル精度の評価

教師付きモデルの精度を評価することは、データマイニング処理の重要な部分です。通常、クラシファイアは、誤差率、つまり誤ってクラス判別されたレコードの割合に基づいて評価されます。回帰ツリー分析モデルは、予測誤差における標準偏差の絶対誤差に基づいて評価されます。モデル精度の評価では、モデルの作成に使用しなかったデータでモデルをテストする必要があります。

教師なし (Unsupervised) モデリング

教師なしモデリングでは、同じような特性をもつデータから規則やセグメントを発見します。正しい答えとか、一般に認められた明確な尺度はありません。教師なしモデリングでは、同じようなパターンやセグメントが示されるため、データ全体を洞察できます。モデルを使って予測を直接行うことはできませんので、クラシファイアを作成するためにデータの一部を使って訓練事例を作成する必要はありません。最もよく使われる教師なしモデリングのアルゴリズムは、相関 (associations) タスクとクラスターリング (clustering) タスクです。

相関 (Associations) タスク

相関の生成では、どのような規則で A が B を内包するのが記述されます。古典的な相関グループ化では、マーケット・バスケット分析を行って、ある品目が同時に購入される頻度を出力します。離乳食を買う顧客はレギュラタバコよりも低タールのタバコを買う確率が高いことが分かれば、それに合わせて棚の編成を変えることができます。

クラスタリング (Clustering) タスク

クラスタリング・アルゴリズムは、データを同じ特性を持つレコードグループ、すなわち、クラスタにセグメント化します。たとえば、あるセグメントが次の特性からなることを保険会社が発見したとします。つまり、このセグメントの特性は、20 歳から 45 歳の技術者、子供は 2 人未満、テレビの SF ファン、可処分所得が年 \$5,000 から \$10,000 であるとしします。保険会社は、SF 物語にテレビ広告を打ち、この人たちに適した生命保険パッケージを用意すれば、このセグメントを効果的に攻めることができます。

データの可視化

データマイニング・アルゴリズムを補足するものとして、データ可視化技術が使用できます。人間の脳には驚くべきパターン認識能力があります。MineSet には、次のビジュアライザがあります。

- マップ・ビジュアライザ (Map Visualizer) — データが地図形式で表示されます。
- スキャタ・ビジュアライザ (Scatter Visualizer) — データポイントが 1、2、または 3 次元で表示されます。さらに、他の属性を色とサイズに対応させることができます。また、他の 2 つの属性をスライダに対応させれば、アニメーションやフライスルーを行うことができます。8 つめの属性を色コードに対応させれば、興味深いと思われるアニメーション変数の値の組み合わせを使ってアニメーションを行うことができます。MineSet の重要項目 (Column Importance) 機能を使うと、可視化においてどの次元に何を対応させるかについて明らかになる場合があります。
- スプラット・ビジュアライザ (Splat Visualizer) — スキャタ・ビジュアライザと似ていますが、スプラット・ビジュアライザでは、データ密度がさまざまな不透明度で表示されます。個々のデータポイントをレンダリングしたような効果が得られます。このビジュアライザは、スキャタ・ビジュアライザで表示するにはデータベースのポイント数が多すぎるときに特に便利です。
- ツリー・ビジュアライザ (Tree Visualizer) — データがノードにマップされるため、データの分類を階層的に見ることができます。デシジョン・テーブル (Decision Table)、選択式決定木 (Option Tree)、回帰ツリー (Regression Tree) では、データをさまざまな分岐ツリーによる可視化で表示できます。

データマイニングのための MineSet ツール

クラス判別 (Classification)、回帰 (Regression)、相関 (Association)、クラスタリング (Clustering) を必要とするデータセットには、次の MineSet ツールが便利です。

- 決定木分析／ビジュアライザ (Decision Tree Inducer and Visualizer) — クラシファイアを作成し、決定木として可視化します。
- 選択式決定木分析／ビジュアライザ (Option Tree Inducer and Visualizer) — 決定木分析／クラシファイアに似たクラシファイアを作成しますが、別の選択枝を作成し、クラス判別中にそれらの選択枝を平均化することによって、通常精度を向上させます。
- エビデンス分析／ビジュアライザ (Evidence Inducer and Visualizer) — 独自のクラシファイアを作成し、与えられたデータを基にエビデンスを可視化します。
- デシジョン・テーブル分析／ビジュアライザ (Decision Table Inducer and Visualizer) — 各階層レベルを対の次元で可視化します。コンテキストを維持しながら、階層をすばやくドリルアップ、ドリルダウンすることができます。
- 回帰ツリー分析／ビジュアライザ (Regression Tree Inducer and Visualizer) — 実際に値の入った属性を予測する回帰ツリー分析モデルを作成します。つまり、あらかじめ決められた特定の限界ではなく、値の段階的変化を予測します。回帰ツリー分析モデルの構造はツリー・ビジュアライザで表示されます。
- 相関規則 (Association Rule) — A が B を内包する場合の相関を支配する規則を生成します。これは一般にマーケット・バスケット分析と呼ばれます。規則は、スキャタ・ビジュアライザ (Scatter Visualizer) を使って、通常、バーやディスクで表示されます。
- クラスタリング・アルゴリズム (Clustering Algorithm) — 特性の相似に基づいてデータをグループ化して、統計量ビジュアライザ (Statistics Visualizer) のような一連のボックスプロットとヒストグラムで表示します。デフォルトでは、クラスタリング・アルゴリズム (Clustering Algorithm) はクラスタ・ビジュアライザ (Cluster Visualizer) を使用して結果を表示しますが、他のビジュアライザも使用することができます。
- 重要項目 (Column Importance) — ラベルの値を判別する場合の特定の項目の重要度を決定します。このツールは、変数を変更した効果を観察する場合や、スキャタ・ビジュアライザ (Scatter Visualizer) とスプラット・ビジュアライザ (Splat Visualizer) の軸に割当てる項目を指定する場合などに使用します。

MineSet にはさらに、知識発見プロセスを支援する次のツールがあります。

- 統計量ビジュアライザ (Statistics Visualizer) — データを項目ごとに 1 つずつボックスプロットや柱状グラフの形で表示します。連続型項目はボックスプロットとして、離散型項目は柱状グラフとして表示します。
- ヒストグラム・ビジュアライザ (Histogram Visualizer) — データを項目データごとに 1 つずつヒストグラムの形で表示します。必要なら、連続型項目の階級を生成します。
- レコードビューワ (Record Viewer) — 元データをスプレッドシートで表示します。

データマイニング・プロセスの概要

ここでは、知識発見プロセスに必要な手順について概要を説明します。このプロセスは [図 1-2](#) のように循環します。通常、新しいパターンを発見し、データの理解が深まったら前の方の段階に戻ります。



図 1-2 データマイニング・プロセス

このプロセスでは、次の手順を行います。

1. データのソースを識別する。「[データの識別](#)」(9 ページ) を参照してください。
2. データを準備する。「[データの準備](#)」(9 ページ) を参照してください。

3. モデルを作成する。「[モデルの構築](#)」(11 ページ) を参照してください。
4. モデルを評価する。「[モデルの評価](#)」(11 ページ) を参照してください。
5. モデルを展開する。「[モデルの展開](#)」(11 ページ) を参照してください。

データの識別

データの識別では、問題を解決するためにどのようなデータが必要かをまず決める必要があります。たとえば、よくあることですが、顧客の行動を予測したいとします。その場合には、問題の観点から解決に必要なデータを識別し、データのソースが他にないか調べる必要があります。

入手しにくい場所にデータがあったり、データ形式がはっきりしないことがあります。あるいは、いくつかの初期データベースが相互の整合性に欠ける場合があります。さらに、データが不十分であったり、不完全な場合は、さらに多くのデータが必要になります。新しいデータをどのようにして収集するかは、既存データの形式によって異なります。MineSet は、商用データベース Oracle、Informix、SQL に対するネイティブ・インタフェースと ODBC インタフェースをサポートするほか、タブで区切られたフラットファイル、MineSet バイナリファイル、Excel、SPSS、MATLAB など、さまざまなファイル形式のデータを読み取ることができます。Conceptual Software, Inc. の DBMS/COPY などのツールを使えば、100 以上の形式のデータを MineSet の形式に変換できます。さらに、Tool Manager の「ファイル」メニューから使用する MineSet のインポートデータ (MineSet Import Data) 機能でも、同じような変換を行うことができます。詳細については、『[MineSet 3.0 Enterprise Edition Reference Guide](#)』と『[MineSet 3.0 Enterprise Edition Interface Guide](#)』を参照してください。

データの準備

データを MineSet にロードする前にデータを変更しなければならない場合があります。この手順は、しばしばクリーニングと呼ばれます。データの変更が必要になるのは、よく見られる次の問題の場合です。

- データの形式が MineSet のデータ形式と互換性がない。たとえば、古いメインフレーム・コンピュータのバイナリデータ、符号化データ、EBCDIC 文字列などです。
- データのスペルが正しくなかったり、データの値が不完全または間違っている。

- フィールドの記述が不明瞭または紛らわしい。あるいは、ソースによって意味が異なる。たとえば、注文日は、送信された日なのか、消印が押された日なのか、受信した日なのか、データを入力した日なのかなどです。
- データが古い。たとえば、顧客が引っ越してしまったり、家族構成が変わったり、支出パターンが変わっているかもしれません。

クリーンデータでさえ、マイニングや可視化に適したものにするには、変換が必要になることがあります。

データの変換

データを変換すると、モデルのパフォーマンスが大幅に向上します。たとえば、電話会社のデータを解析した場合、顧客動向を予測するためには個々の要素よりも、長距離電話をかける割合の方が予測材料として便利であることがわかるかもしれません。データ変換は、信頼できるモデルを構築するための中心的な作業です。また作業を進めながら、前に戻ってデータを別な方法で変換することもできます。データ変換には、次の方法を利用できます。

- 項目を追加する。このためには、通常、既存のデータに数式を入力して新しいフィールドを作成します。
- 不適切な項目、重複した項目、明らかに使用されない予測要素が入った項目などを削除する。
- 項目値のブール表現に基づいてデータをフィルタリングして、モデルや可視化に影響を及ぼす。たとえば、最も強い規則や最も収益性の高い顧客セグメントを見たい場合です。
- データを階級生成する。連続した範囲のデータを、[1 ～ 10]、[11 ～ 20] のような離散型セグメントに分割します。
- データを集計処理する。つまり、レコードをグループ化し、合計値や最大値、最小値、平均値を見つける。
- データの標本を抽出して、割合またはカウントごとにランダムなデータサブセットを生成する。
- 前に作成したクラシファイア、回帰ツリー分析ツール、クラスタリング・モデルを適用して、新しいレコードにクラスのラベルを付けたり、あるラベル値の確率を推定する。

MineSet では、これらの変換のほとんどを Tool Manager の「データ変換」パネルを使って行います。変換機能や特殊なモデルを作成するプラグイン・モジュールを利用できる場合もあります。

モデルの構築

知識発見プロセスの中心はモデルの構築（作成）ですが、これはデータマイニング・アルゴリズムによって自動的に行われます。モデルの構築では、すべてのデータを使用することもできますし、後でモデルの精度をテストするために一部を残しておくこともできます。モデルの構築では、モデルの生成方法に関連していくつかのオプションを使用できます。どのオプションを選択するかによって、可視化がどのように行われるかだけでなく、クラシファイアがアルゴリズムによってどのように作成されるかが決まります。これらのオプションについては、[第 7 章「予測モデリングの理解」](#)でさらに説明します。

モデルの評価

モデルの精度を評価すると、モデルとその有効性についての理解が向上します。決定木 (Decision Tree) モデルや選択式決定木 (Option Tree) モデルなど、モデルの中にはモデルの別々の部分を評価し、それを可視化をとおして直接表示するものがあります。

MineSet のモデル評価方法には、誤差推定、混同マトリックス、改善曲線、ROI (投資利益率) 曲線の 4 つがあります。これらの評価方法については、[第 11 章「予測モデリングの改良」](#)で説明します。

モデルの展開

モデルを展開するには、モデルを新しいデータに適用します。新しいデータを使用すると、新しい疑問が発生し、モデルをさらに改善しなければならない場合があります。

『*MineSet 3.0 Enterprise Edition Tutorial*』の電話会社の例では、解約しそうな顧客（つまり、この電話会社から離れそうな顧客）を判定するモデルが作成されています。このモデルを顧客レコードに適用すれば、解約する確率が最も高い顧客を識別できます。このような顧客には、他の会社に移るのを思いとどまらせるような奨励措置を取ることができます。

churn（解約）データセットに対し知識発見プロセスを適用できます。このデータセットはあらかじめ用意された電話会社の顧客データベースで、MineSet ソフトウェアに添付されています。このマニュアルの例を見る際には、ビジネスや科学研究における業務がどれほど前進するかを考えながら、完全にテストされた分析データマイニング・モデルが得られるまで、繰り返しモデルを改善してください。

第 2 章「[MineSet によるデータのアクセス](#)」では、MineSet とサンプルファイルを実際に使用します。

MineSet によるデータのアクセス

この章では、MineSet の起動方法や、Tool Manager と呼ぶグラフィカル・ユーザ・インタフェースを使ってデータを表示する方法を説明します。この章の構成は次のとおりです。

- 「MineSet が格納されているディレクトリ」(13 ページ)
- 「MineSet ソフトウェアの動作」(14 ページ)
- 「MineSet とユーザ・アプリケーション」(15 ページ)
- 「MineSet の起動」(15 ページ)
- 「「Tool Manager」ウィンドウの使い方」(18 ページ)
- 「レコードビューワ (Record Viewer) による元データの表示」(19 ページ)
- 「統計量ビジュアライザ (Statistics Visualizer) によるレコード統計の表示」(22 ページ)
- 「統計量ビジュアライザの起動」(24 ページ)
- 「ヒストグラム・ビジュアライザの起動」(25 ページ)
- 「MineSet 3D ビジュアライザでのナビゲーション」(27 ページ)
- 「ヘルプ情報」(32 ページ)

さらに、いろいろなツールの練習をするときは、『*MineSet 3.0 Enterprise Edition Tutorial*』を参照してください。独自のデータベースを使用する場合は、『*MineSet 3.0 Enterprise Edition Interface Guide*』の MineSet のセットアップに関する説明を参照してください。

MineSet が格納されているディレクトリ

MineSet は、CD またはインターネットからインストールして使用します。MineSet をインストールしたら、MineSet 製品に付属しているフォーマット済みのデータファイルを使用して MineSet をただちに起動できます。このファイルは、MineSet がインストールされているディレクトリの data ディレクトリにあります。MineSet の操作について

は、次の節で簡単に説明します。すぐに起動したい場合は、「[MineSet の起動](#)」(15 ページ) を参照してください。

MineSet ソフトウェアの動作

MineSet はクライアント / サーバ方式で動作します。サーバは、クライアントと同じシステムにあっても、それとは別のさらに強力なシステムにあってもかまいません。サーバは、クライアントからアクセスできるところにインストールされていなければなりません。サーバでは DataMover が動作します。DataMover は、データベース・ファイルにアクセスしたり、データ変換を行ったり、マイニング操作を行ったり、可視化ファイルを生成したりします。この可視化ファイルはクライアントに転送されます。

グラフィカル・ユーザ・インタフェース (GUI: graphical user interface) を提供するクライアント・プロセスには、Tool Manager を使用します。MineSet との対話型操作のほとんどは、Tool Manager 使って行います。Tool Manager では、データのソースや、データに適用する一連の変換、使用するマイニングツールや可視化ツール、作業結果の保存方法などを指定します。

このような指定が終わったら、MineSet に対し、必要なデータにアクセスし、Tool Manager に指定したアクションを行うように要求します。Tool Manager は、この情報をサーバの DataMover プロセスに送信します。分析するデータがフラットファイルにある場合は、MineSet が設定ファイル (.schema) を読み取ります。この設定ファイルには、テーブルや、実際のデータがある .data ファイルの名前に関するメタ情報が入っています。分析するファイルが Oracle や Sybase、Informix データベースの場合は、DataMover がデータベースに対しクエリーを実行して、必要なデータセットを取得します。

.schema ファイルには、入力データファイルの形式が記述されています。このファイルについては、『*MineSet 3.0 Enterprise Edition Interface Guide*』に詳しい説明があります。

注記： MineSet のツールを使用する前に、MineSet のリリースノートに記載されているインストールとライセンスの情報をお読みください。システム管理者は DataMover の設定ファイルをセットアップする必要があるかもしれません。セットアップについては、『*MineSet 3.0 Enterprise Edition Interface Guide*』に詳しい説明があります。

MineSet とユーザ・アプリケーション

MineSet のアプリケーション・プログラミング・インタフェース (API: Application Programming Interface) を使えば、MineSet のビジュアライザやデータマイニング・エンジンを独自のアプリケーションに組み込むことができます。詳細は、『*MineSet 3.0 Enterprise Edition Interface Guide*』や MineSet の Web サイトを参照してください。

MineSet の起動

MineSet を起動する最も簡単な方法は、*mineset.exe* のショートカットをデスクトップに作成する方法です。それには、MineSet がインストールされているディレクトリの *\bin* ディレクトリを開き、*mineset.exe* ファイルをクリックします。「ファイル」メニューから「ショートカットの作成」を選択します。

あるいは、デスクトップの「スタート」->「プログラム」->「*MineSet 3.0 Enterprise Edition*」メニューでも MineSet にアクセスできます。マウスの左ボタンをクリックして MineSet を起動するか、または右ボタンを押したままデスクトップにドラッグし、ボタンを離します。

1. MineSet のショートカット・アイコンをクリックすると、MineSet の Tool Manager が起動します。
2. デフォルトで「サーバへログイン」ダイアログ・ボックス (図 2-1) が表示されます。表示されない場合は、「ファイル」->「サーバに接続」を選択します。現在のシステムをクライアントとサーバの両方で使用する場合は、「このマシンを現在のユーザとします」をクリックします。サーバとして別のシステムを使用する場合は、サーバ名、ログイン名、パスワードを入力します。パスワードがない場合は省略できます。

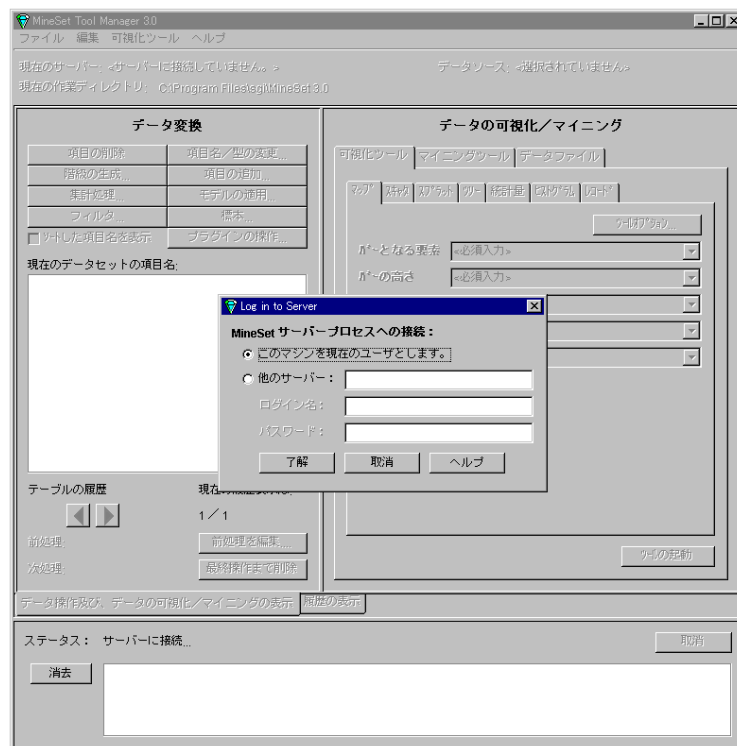


図 2-1 Tool Manager のログイン・ウィンドウ

- 「Tool Manager」ウィンドウで「ファイル」->「新しいデータファイルを開く」を選択します。「データファイルを開く」ウィンドウが表示されますので、使用するデータへのパスを入力します。MineSet に付属しているサンプルファイルを試してみたい場合には、MineSet をインストールした *MineSet 3.0\data* ディレクトリにそれらのファイルがあります。

サンプルファイルを使用するには、「データファイルを開く」ウィンドウで「*churn.schema*」を選択します。このファイルは、MineSet に付属しているデータのサンプルテーブルです。図 2-2 のように、右側の「前の項目」パネルにはデータの項目名が表示されます。「開く」をクリックします。

churn.schema ファイルは、電話会社の顧客データセットにアクセスするときに使用します。MineSet を次に実行するときには、最後に終了したときの状態に自動的に戻り、前に選択したオプションが復元されます。

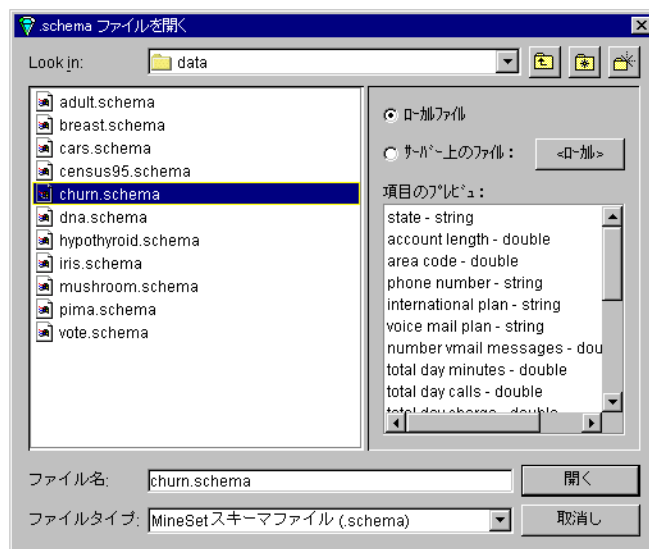


図 2-2 「データファイルを開く」 ウィンドウ

練習のために使用できるその他のサンプルファイルのリストが『*MineSet 3.0 Enterprise Edition Reference Guide*』に載っています。各ファイルの内容については、このマニュアルの「設定ファイルとデータファイルのサンプル」の項を参照してください。

「Tool Manager」ウィンドウの使い方

この Tool Manager からすべての操作を開始します。したがって、この部分を簡単に説明します。

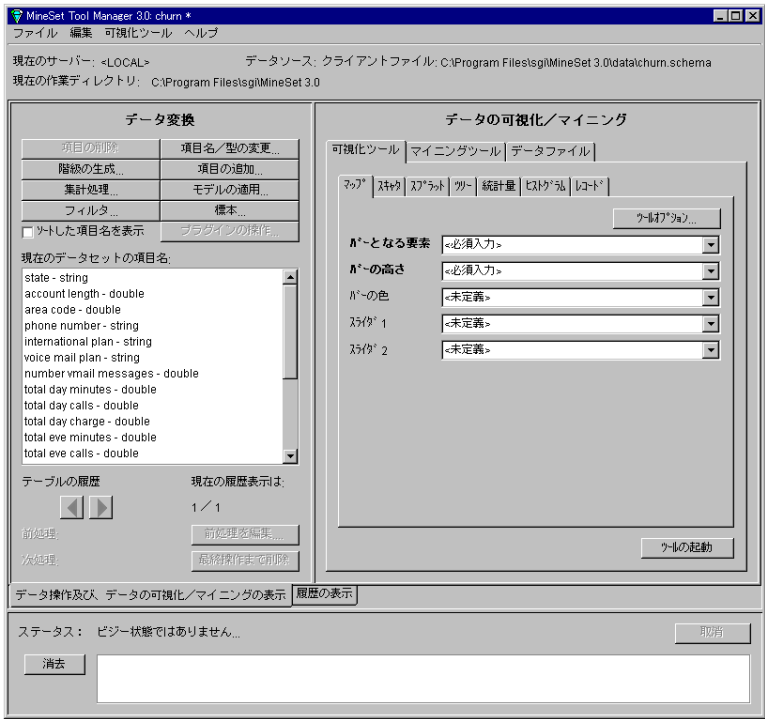


図 2-3 「Tool Manager」ウィンドウ

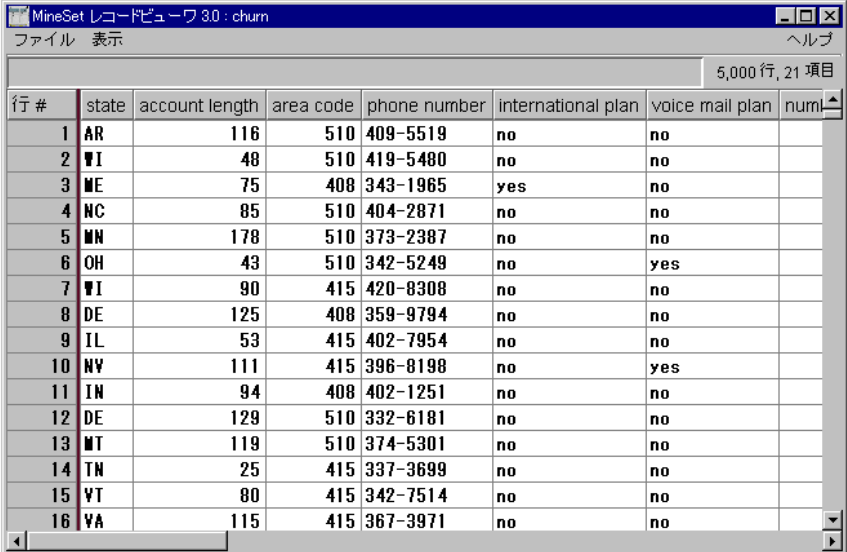
左側の「データ変換」パネルは、すべてのデータの変換を行うところです。上部のボタンを使って行うアクションは、その下の「現在のデータセットの項目名」に適用されます。下部のボタンでは、変換の履歴を前後に移動します。最下部の「ステータス」ウィンドウには、データセットに対して行われているアクションが表示されます。右側の「データの可視化／マイニング」パネルでは、可視化ツールやマイニングツールを選択したり、現在の状態をデータファイルに保存します。これらの機能については、次の各節で説明します。

レコードビューワ (Record Viewer) による元データの表示

レコードビューワを使えば、選択したデータセットの内容を簡単に見ることができます。データベース・レコードや項目内のデータ値がスプレッドシート形式で表示されます。

「[MineSet の起動](#)」の手順に従って Tool Manager を起動してから、次の手順に従います。

1. 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。
2. その下にある一連のタブから「レコード」を選択して「レコードビューワ」にアクセスします。
3. パネルの右下隅にある「ツールの起動」をクリックすると、レコードビューワが表示されます ([図 2-4](#))。



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

図 2-4 レコードビューワによる表示

表示方法を変えたい場合は、ソート、番号変更、サイズ変更などをレコードビューワに指定できます。

レコードビューワの項目の変更

表 2-1 の方法を使えば、レコードビューワの項目の表示方法を変えることができます。

表 2-1 レコードビューワの項目に対する操作

目的	操作
項目の表示サイズの変更	項目の右枠をクリックし、必要な方向へドラッグします。
項目の再編成	タイトルセルをクリックし、必要な場所へドラッグします。
項目の非表示	「表示」メニューから「項目の表示／非表示」を選択し、非表示にする項目の選択を解除します。
項目の値によるレコードのソート	項目のタイトルセルをクリックします。
複数項目によるレコードのソート	<Ctrl> キーを押したまま、ソートに使用する項目タイトルを、ソートする順に 1 つずつクリックします。
ソートの反転	項目のタイトルセルを再びクリックします。
元の順序を復元	行番号項目のタイトルをクリックします。
行番号の変更	「表示」メニューから「行番号の振り直し」を選択します。この処理を元に戻すことはできません。
数字または値の検索	「表示」プルダウン・メニューから「検索」パネルを選択し、「検索」フィールドに値を入力し、検索に使用する項目を強調表示し、「次を検索」か「前へ検索」をクリックします。複数項目を選択する場合は <Shift>- クリックを、連続していない項目を選択する場合は <Ctrl>- クリックをそれぞれ使用します。

レコードビューワでのデータのフィルタリング

一定範囲の値だけを見るためにデータをフィルタリングすることができます。

1. 「レコードビューワ」ウィンドウで「表示」->「フィルタパネル」を選択します。
2. フィルタ表現を作成するために使用する項目と表現を選択します。たとえば、cars データセットから立方インチ容量 (ci) が 400 ci より大きいレコードだけをフィルタリングする場合は、表現 'cubicinches' > 400 を作成します (図 2-5 を参照)。

表現を作成するには、左側から要素を選択し、右矢印をクリックするか、テキストフィールドに表現を直接入力します。

3. 「適用」をクリックすると、フィルタ処理がただちに始まります。パネルは、閉じるまで表示されています。

フィルタリングの結果を削除するには、表現ウィンドウで表現を消去し、「適用」を選択するか、「レコードビューワ」ウィンドウで「表示」->「フィルタの削除」を選択します。



図 2-5 レコードビューワのフィルタパネル

フィルタパネルはいくつでも開けます。複数のフィルタを同時に適用したい場合は、まず 1 つのフィルタを適用してから行番号を変更し、次のフィルタを適用します。

注記： 変更した行番号を元に戻すことはできません。元データに戻りたい場合は、ファイルを開き直す必要があります。

レコードビューワでのデータの保存

レコードビューワでファイルを保存するには、「ファイル」->「保存」か「ファイル」->「名前を付けて保存」を使用します。

「名前を付けて保存」では、データをバイナリ、ASCII、HTML、またはテキストで保存できます。バイナリか ASCII 形式で保存すると、データファイル（テーブル）とスキーマファイルが両方とも保存されます。スキーマファイルは、Tool Manager や DataMover に対してデータファイルの内容を記述したものです。HTML 形式では、ファイルが HTML テーブルとして保存されます。テキスト形式では、タブで区切られた形式でファイルが保存されます。最初の行が項目タイトルです。

統計量ビジュアライザ (Statistics Visualizer) によるレコード統計の表示

統計量ビジュアライザを使えば、データセットのレコードについてさらに多くのことが分かります。ある種の統計はデータセットのレコード数に基づいて計算されます。項目のデータ型が数値か離散値かによって、統計値はボックスプロットかヒストグラムで表示されます。

ボックスプロットとは

ボックスプロットでは、ある列に属する数値の最小値、最大値、平均値、中央値、2 つの四分位数（25 と 75 の百分位数）が、色の付いた垂直バーに対し横線で表示されます。データセット母数の標準偏差は $\pm$ 値で表されます。四分位数は、個別値の数が 50,000 より少ないときに表示されます（[図 2-6](#) を参照）。項目に属する個別値の数が 50,000 より多いと、統計値はグレイの垂直バーとして表示されます。用語集の「平均値 (Mean)」、「中央値 (Median)」、「標準偏差 (Standard Deviation)」を参照してください。

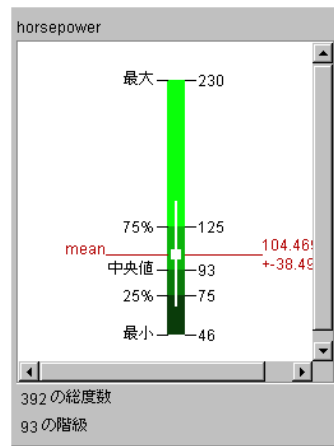


図 2-6 統計量ビジュアライザで表示した数値型項目

ヒストグラムとは

ヒストグラムには、文字列、階級など、数値以外の項目を表示します（[図 2-7](#)を参照）。つまり、この項目には、「yes」、「spore_color」などの文字列や、「10-70」などの階級値が入っています。項目には、このような個別エントリを最高 100 まで持つことができます。文字列値（離散型）属性のデフォルト順序は数値の降順ですが、「表示」プルダウン・メニューを使えば、他のソート方法も選択できます。文字列をソートする方法としては、数値順とアルファベットによる名前順の 2 つがあります。重み付けが選択されている場合は、数値順の代わりに重み順が使用されます。個別カテゴリの数が 100 以下の場合は、項目パネルには、個別値のカウントがさらに含まれます。値の階級が生成されている場合は、ソート方法が何であれ、階級順序が保持されます。

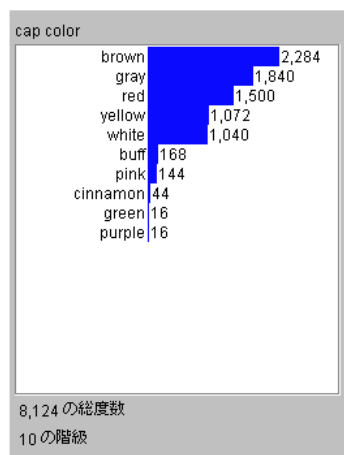


図 2-7 統計量ビジュアライザで表示した離散型項目

データの可視化を行った後でヒストグラムの切捨てられたテキスト情報を見るには、そのフィールドの上にマウスを移動します。

統計量ビジュアライザの起動

統計量ビジュアライザを起動するには、データセットを選択してから次の手順に従うのが最も簡単です。

1. 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。
2. その下にある一連のタブから「統計量」タブを選択して、統計量ビジュアライザにアクセスします（図 2-8 を参照）。
3. パネルの右下にある「ツールの起動」をクリックします。

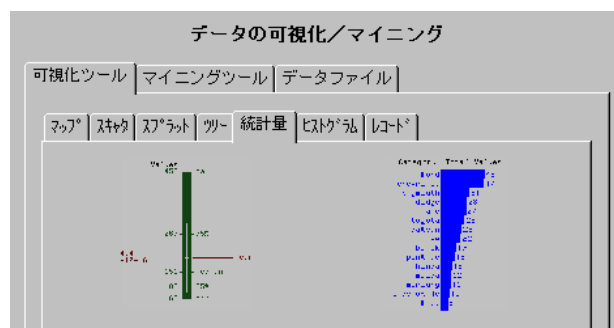


図 2-8 統計量ビジュアライザが選択された「データの可視化／マイニング」パネル

このツールを起動する他の方法については、『*MineSet 3.0 Enterprise Edition Reference Guide*』の統計量ビジュアライザの説明を参照してください。

サンプル・データファイルの使用

統計量ビジュアライザの特徴や機能を知りたい場合は、サンプル・データファイルを使用してください。サンプルファイルは、MineSet を最初にインストールした *MineSet 3.0\examples* ディレクトリにあります。

Tool Manager のメニューバーにある「可視化ツール」メニューから「統計量ビジュアライザ」を選択し、次に「ファイル」->「開く」プルダウン・メニューを使っていずれかの *.statviz* ファイルを開きます。

ヒストグラム・ビジュアライザの起動

ヒストグラム・ビジュアライザは、連続値（つまり、数値）が含まれているすべての項目の階級を生成し、結果を統計量ビジュアライザに送信して表示します。用語集の「階級生成 (Binning)」を参照してください。

ヒストグラム・ビジュアライザを起動するには、次の手順が最も簡単です。

1. 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。
2. ツール選択メニューから「ヒストグラム ビジュアライザ」を選択します（[図 2-9](#)を参照）。
3. 次のオプションを設定します。オプションを設定しない場合は、次の手順へ進みます。
 - 階級数を指定します。あるいは、MineSet に任せます。
 - 端数の切捨を設定します。これは、階級を作る前に極端な値を除外することを意味します。デフォルトは 0.05 です。この場合には、最も極端な値のインスタンスが、下位から 2.5%、上位から 2.5%、合計で 5% だけ除外されます。通常、切捨を使うと、区間の生成に与える外れ値の影響が少なくなります。外れ値とは、データの塊の外にある単独のインスタンスです。

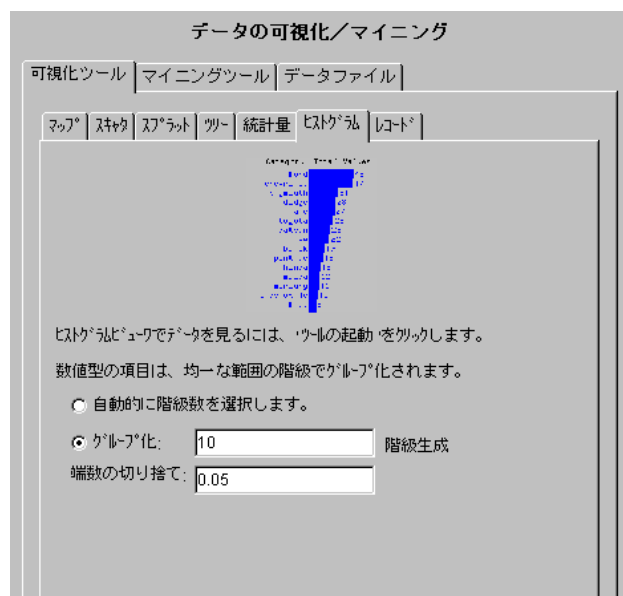


図 2-9 ヒストグラム・ビジュアライザが選択された「データの可視化／マイニング」パネル

4. 右下の「ツールの起動」をクリックします。

「*Tool Manager*」ウィンドウの下部にある「ステータス」ウィンドウには、アルゴリズムがデータをどこまで処理したかを示す進捗状況が表示されます。その後、現在のデータがヒストグラムとして表示され、階級生成されたデータの状態が一目で分かります。「[ヒストグラムとは](#)」(23 ページ) を参照してください。

MineSet 3D ビジュアライザでのナビゲーション

この節では、さまざまな MineSet ビジュアライザのナビゲーション・コントロールについて説明します。ツリー・ビジュアライザと非ツリー・ビジュアライザにはそれぞれ固有のコントロールがあります。次の「[ツリー・ビジュアライザでのナビゲーション](#)」と「[非ツリー・ビジュアライザでのナビゲーション](#)」(29 ページ) でそれぞれのコントロールを説明します。

ツリー・ビジュアライザでのナビゲーション

ツリー・ビジュアライザ (Tree Visualizer) の表示内容は、カメラを通し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）で利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 2-2](#) にナビゲーション・ボタンの一覧を示します。

表 2-2 ツリー・ビジュアライザ (Tree Visualizer) のナビゲーション・ボタン

ボタ ン	操 作
	ホーム表示として指定されたサイズおよび位置にグラフを戻します。デフォルトでは、ビジュアライザを最初に起動したときのサイズと位置にグラフが表示されます。次のボタンを使用し、ホーム位置を変更することができます。
	グラフの新しいホーム表示を設定します。特定の表示または位置を保存する場合も、このボタンを使用します。
	グラフをウィンドウ中央に移動して全体を表示します。
	Web ブラウザの「戻る」ボタンのように、前回の移動操作を取消します。
	Web ブラウザの「次」ボタンのように、取消した移動操作をもう一度実行します。
	ツリーのルートに向かって 1 ノード分だけ移動します。

表 2-2 (続き) ツリー・ビジュアライザ (Tree Visualizer) のナビゲーション・ボタン

ボタ ン	操 作
	1 ノード分または 1 バー分だけ左側に移動します。
	1 ノード分または 1 バー分だけ右側に移動します。
	ツリーの左側のパス上で、1 ノード分だけ下に移動します。
	ツリーの右側のパス上で、1 ノード分だけ下に移動します
	現在のノードから移動できるパスを示すポップアップ・メニューを表示します。

表 2-3 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。これらの操作のほとんどは、ビジュアライザ・ウィンドウのいずれかのコントロールまたはマウスを使用して実行することができます。

表 2-3 ツリー・ビジュアライザの調整スライダとダイヤル

操 作	調整スライダまたはダイヤ ル	マウス操作
シーンの表面上をフライオー バーする	該当なし	マウスの左ボタンと右ボタン（また は中ボタン）を押したままマウスを 動かします。
バーの高低を調整して差異を 強調する	「Height」スライダ（左 上）	該当なし
視点を上下に移動する	「H」ダイヤル	マウスの右ボタンを押したままマウ スを上下に動かします。
視点を左右に移動する	「<-->」ダイヤル	マウスの左ボタンと右ボタン（また は中ボタン）を押したままマウスを 端から端へ動かします。。
視点を前後に移動する	「Dolly」ダイヤル	マウスの左ボタンと右ボタン（また は中ボタン）を押したままマウスを 上下に動かします。
カメラの上下の傾きを変更す る	「Tilt」ダイヤル	該当なし

表 2-3 (続き) ツリー・ビジュアライザの調整スライダとダイヤル

操作	調整スライダまたはダイヤル	マウス操作
ポイント方向前方へ移動する	該当なし	<Alt> キーと、マウスの左ボタンと右ボタン（または中ボタン）を押したままマウスを動かします。前方に移動すると、視点は現在の傾きを基準に下に移動します。同様に後方に移動すると、視点は現在の傾きを基準に上に移動します。
ノードの子ノードを選択する	該当なし	<Ctrl> キーを押したまま親ノードをマウスの右ボタンでクリックし、次にその子ノードをマウスの左ボタンでクリックして移動するか、または分岐ナビゲーション・アイコンを使用します。

非ツリー・ビジュアライザでのナビゲーション

エビデンス・ビジュアライザ (Evidence Visualizer)、デシジョン・テーブル・ビジュアライザ (Decision Table Visualizer)、マップ・ビジュアライザ (Map Visualizer)、スキャタ・ビジュアライザ (Scatter Visualizer)、およびスプラット・ビジュアライザ (Splat Visualizer) 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。表 2-4 にナビゲーション・ボタンの一覧を示します。

表 2-4 非ツリー・ビジュアライザのナビゲーション・ボタン

ボタン	名前	操作
	ピック	プログラムをピックモード（矢印）に切替えます。ピックモードでは、グラフの要素を強調表示（ブラシオーバー）または選択（クリック）することができます。

表 2-4 (続き) 非ツリー・ビジュアライザのナビゲーション・ボタン

ボタン	名前	操作
	ハンド モード	プログラムをハンドモード（手のひら）に切替えます。ハンドモードでは、ウィンドウ内でグラフを移動することができます。 — ウィンドウ内でグラフを移動するには、マウスの右ボタンを押したままマウスを動かします。 — グラフを回転させるには、マウスの左ボタンを押したままマウスを動かします。 — グラフの視点を前後に移動（ドリーイン、ドリーアウト）するには、マウスの左ボタンと右ボタン（または中ボタン）を押したままマウスを動かします。
	ホーム	ホーム表示として指定されたサイズおよび位置にグラフを戻します。デフォルトでは、ビジュアライザを最初に起動したときのサイズと位置にグラフが表示されます。「ホームの設定」アイコンを使用することによってホーム位置を変更することができます。
	ホームの設 定	グラフの新しいホーム表示を設定します。このボタンを使用して特定の表示または位置を保存します。
	すべて表示	グラフをウィンドウの中心に移動して全体を表示します。
	ズーム	選択したポイントをウィンドウの中央に移動してズームします。マウスカーソルが目的の表示になったら詳しく表示したい個所にカーソルを移動し、マウスの左ボタンをクリックします。
	3D	3D 遠近法を切替えます。
	上面表示	グラフ表示を上面表示に切替えます（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	グラフ表示を前面表示に切替えます（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）。
	側面表示	グラフ表示を側面表示に切替えます（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）。

非ツリー・ビジュアライザで利用できる調整スライダとダイヤルの一覧を表 2-5 に示します。

表 2-5 非ツリー・ビジュアライザの調整スライダとダイヤル

操作	スライダまたはダイヤル	マウスまたはキーボードの操作
選択（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、またはナビゲーション・ボタンを使用します。
シーンを移動する	該当なし	マウスの右ボタンでグラフをクリックして押したまま、移動したい方向にカーソルを動かします。
ケーキグラフ、円グラフ、またはバーの高低を調整して差異を強調する	「 <i>Height</i> 」スライダ（左上）	該当なし
X 軸を中心にシーンを回転する	「 <i>Rotx</i> 」ダイヤル	マウスの左ボタンでグラフをクリックして押したまま、回転させたい方向にカーソルを動かします。
Y 軸を中心にシーンを回転する	「 <i>Rotx</i> 」ダイヤル	マウスの左ボタンでグラフをクリックして押したまま、回転させたい方向にカーソルを動かします。
シーンをズームインまたはズームアウトする	「 <i>Dolly</i> 」ダイヤル	マウスの左および右ボタン（または中ボタン）でグラフをクリックして押したまま、マウスを動かします。ズームインするにはマウスを下方向に動かし、ズームアウトするにはマウスを上方向に動かします。

表 2-5 (続き) 非ツリー・ビジュアライザの調整スライダとダイヤル

操作	スライダまたはダイヤル	マウスまたはキーボードの操作
詳細レベルをドリルダウンする (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	特定のグラフ上 (またはすべてのグラフを対象にする場合は背景) にマウスの矢印カーソルを置き、マウスの右ボタンをクリックします。
詳細レベルをドリルアップする (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	特定のグラフ上 (またはすべてのグラフを対象にする場合は背景) にマウスの矢印カーソルを置き、<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 (またはマウスの中ボタンをクリックします) 。

ヘルプ情報

「[MineSet 情報について](#)」 (xvii ページ) に示した資料のほかに、MineSet に関する次の情報が利用できます。

- ツールの使い方の概要については、『*MineSet 3.0 Enterprise Edition Tutorial*』を参照してください。
- 文脈に関連したヘルプが各ツールにあります。ヘルプを表示したい領域の上で F1 を押してください。
- 一部のツールについては、短い時間で概要を知ることができるオンライン・デモムービーが <http://www.sgi.com/software/mineset/demos.html> から入手できます。

データの整形

この章では、MineSet によるデータの整形について説明します。この章の処理は、次の章で説明するクラシファイアの作成に先立って行う必要があります。この章の構成は次のとおりです。

- 「データの整形の必要性」(33 ページ)
- 「Tool Manager によるデータの変換」(34 ページ)
- 「テーブル履歴による手順のトレース」(48 ページ)
- 「レコードの重み付け」(50 ページ)
- 「重要項目の検索」(51 ページ)

データの整形の必要性

なじみのないデータセットに対しては、まずレコードビューワ (Record Viewer) や統計量ビジュアライザ (Statistics Visualizer) を使って内容を調べようとするでしょう (第 2 章「MineSet によるデータのアクセス」を参照)。この章では、その続きとして元データを整形または変換します。たとえば、意味のある結論を引き出すには情報が多すぎるという場合があります。可視化を簡潔にするには、項目の追加や削除、結合を行うのが最も簡単です。このような操作をしても、元のデータは変更されません。作業中のセッション・ファイルが変更されるだけです。ユーザが適用する一連の変更を Tool Manager セッションの履歴といいます。

MineSet 構成要素との対話のほとんどは、Tool Manager のグラフィカル・ユーザ・インタフェース (Graphical User Interface) を通して行います。MineSet の個別のツールを使えば、既存のデータソースを選択したり、データの変換や分析をしたり、結果を可視化したりすることができます。

MineSet ビジュアライザとの対話方法の詳細については、『*MineSet 3.0 Enterprise Edition Reference Guide*』の各ツールの項目を参照してください。

注記： MineSet 以外のもので作成されたデータファイルを使用する場合は、Tool Manager で手作業が必要になることがあります。詳細は、『*MineSet 3.0 Enterprise Edition Interface Guide*』を参照してください。

Tool Manager によるデータの変換

図 3-1 のボタンの意味は表 3-1 で説明します。データを変換するには、「現在のデータセットの項目名」パネルから項目を選択し、必要なボタンをクリックします。

- 1. MineSet を起動します。第 2 章の「MineSet の起動」の項を参照してください。
- 2. Tool Manager のウィンドウで「ファイル」->「新しいデータファイルを開く」を選択し、プルダウン・メニューからデータセットを選択します。「データファイルを開く」ダイアログでファイルをクリックすると、「データ変換」パネルの「現在のデータセットの項目名」ウィンドウにデータセットの項目名が表示されます (図 3-1)。

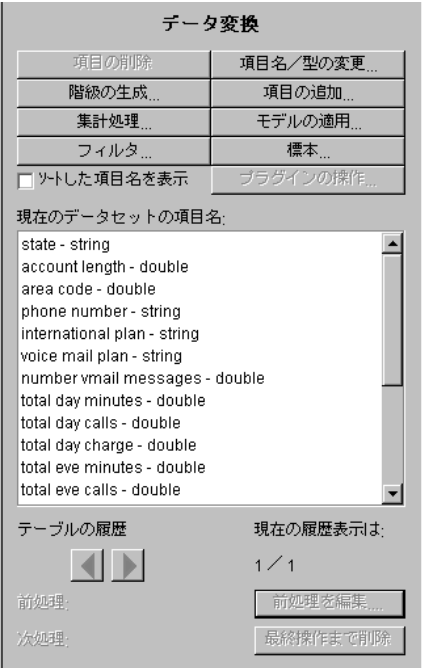


図 3-1 「データ変換」パネル

表 3-1 Tool Manager の「データの変換 (Data Transformations)」パネルのボタン

ボタン名	操作
項目の削除	選択されているすべての項目を削除します。
階級の生成	各レコードをある範囲（階級）に割り当てます。
集計処理	項目の合計、平均、最小、最大、カウントの集計処理を行いながら、他方ではかの項目を使ってデータをグループ化します。集計処理した項目に対し配列索引を使用することもできます。
フィルタ	ブール表現に基づいてデータのサブセットを作成します。表現の評価結果が真になるレコードだけが保持されます。
データ型の変更	項目の名前と型を変更します。たとえば、浮動小数点から整数に変更します。表 3-3 の説明を参照してください。
項目の追加	数式に基づいて新しい項目を追加します。表現には、ほかの項目を含むことができます。
モデルの適用	既存のモデルを使って新しいレコードにラベルを付けるか、データをバックフィットします。
サンプル	大きなデータセットからデータのサブセットをランダムに選択します。
ソートした項目名を表示	データセットの項目をアルファベット順にソートします。

項目の削除と追加

項目の削除や追加を行うには、Tool Manager の「データ変換」パネルで次の手順に従います。

項目の削除 — 可視化やマイニングに必要な項目を選択し、「項目の削除」ボタンをクリックします。不要な項目を削除すれば、計算時間が短くなります。連続していない複数の項目を同時に削除する場合は、<Ctrl> キーを押したまま項目を順に選択していきます。連続した項目の場合は、<Shift> キーを使用します。

項目の追加 — 「項目の追加」ボタンをクリックしてダイアログ・ボックスを表示し、新規項目の名前と数式を指定します（図 3-2 を参照）。たとえば、項目「age」に基づいて「minor_age」という新規項目を追加する場合は、表現に「if age is less than or equal to 18 then minor_age is true; else minor_age is false;」と指定します。この表現は、「if ('age' <= 18) then ('minor_age' = 1) else (minor_age = 0);」と書くこともできます。



図 3-2 「項目の追加」ダイアログ・ボックス

「項目の追加」ボタンをクリックすると、図 3-2 のダイアログ・ボックスが表示されます。

1. ダイアログ・ボックスに新規項目の名前を入力し、その下にポップアップ・メニューを使って項目型（整数、文字列、浮動小数点など）を指定します。項目名にスペースが含まれていてはなりません。
2. 大きいテキスト入力域に表現の定義を入力します。
 - 使用可能な項目や演算子が左側のスクロール・リストに表示されますので、そこから名前や演算子を素早く選択できます。
 - 表現に項目名や演算子を挿入するには、スクロール・リストの項目名や演算子をダブルクリックするか、項目名や演算子を選択してからリスト右の矢印をクリックします。

3. 「表現をチェック」ボタンをクリックして、表現の構文が正しいか検査します。ステータスボックスを消すには、「了解」をクリックします。もう一度「了解」をクリックすると、「項目の追加」ダイアログ・ボックスが消去されます。

「項目の追加」ダイアログ・ボックスでは、型の互換性がさらにチェックされます。数値表現を文字列項目に割り当てたり、その反対を行うと、警告メッセージが表示され、新規項目の型は自動的に正しいものに變更されます。

項目に対する階級の変更または新規作成

モデルや可視化の準備では、レコードを複数の階級にグループ化しなければならない場合がよくあります。とりわけ、アニメーション用のスライダを作成するときにはこれが必要になります。可視化を簡潔にするためには、階級の境界を變更したり、項目の数や範囲を試してみる必要があります。

たとえば、年齢グループとして 0-18、19-30、31-50、51-60、61- を設定するとします。これは、(...18],[18...30],[30...50],[50...60],[60...] と表すことができます。数字の横の丸かっちは、その数字が範囲に含まれないことを示します。角かっちは数字が範囲に含まれることを示します。さらに、用語集の「階級生成 (Binning)」と『*MineSet 3.0 Enterprise Edition Reference Guide*』の階級生成の説明を参照してください。

Tool Manager の「データ変換」パネルで「階級の生成」をクリックしてダイアログ・ボックスを表示し、階級生成オプションを指定します (図 3-3 を参照)。



図 3-3 「階級の生成」ダイアログ・ボックス

1. ダイアログ上部のパネルで、階級を生成する項目をクリックします。デフォルトは「階級生成を行いません。」です。階級の生成方法を選択すると、階級を生成する項目の名前が「新しい項目名」フィールドに表示されます。新しい項目名を入力することもできます。複数の項目を選択すると、「新しい項目名」がアクティブになりません。

元の項目を削除する場合は、「元の項目名を削除」ボックスをクリックしてください。

2. 「階級自動生成」、「区間の指定」、または「均一な区間設定」を選択します。「均一な区間設定」は、階級を等間隔にすることを意味します。
 - ・ 「階級自動生成」を選択すると、MineSet が自動的に階級区間を決めます。
 - ・ 「区間の指定」を選択する場合は、「区間」フィールドに階級区間を指定します。
 - ・ 均一な区間設定により、範囲のはじめと終わり、各階級のサイズを指定できます。

「詳細を表示」矢印をクリックすると、階級生成を行う詳細オプションのパネルが表示されます (図 3-4)。

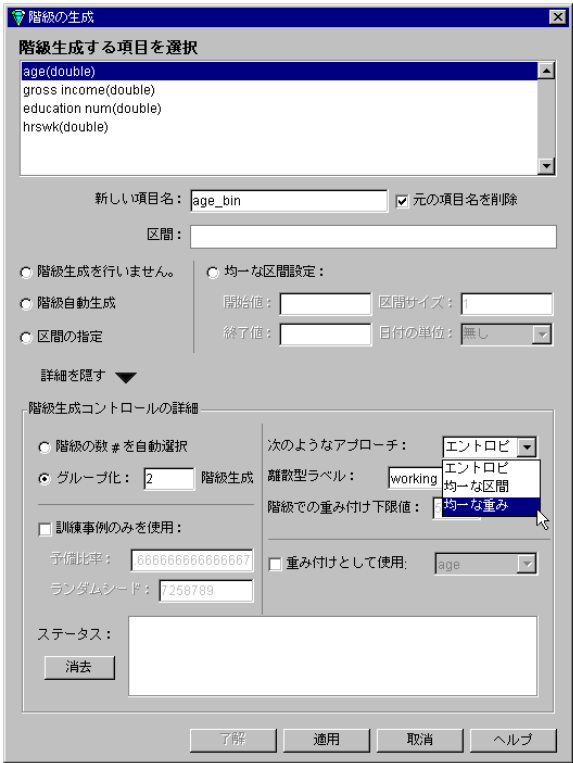


図 3-4 階級生成を行う詳細オプションパネル

3. 階級生成を行う詳細オプションパネルでは次のオプションを選択できます。

表 3-2 階級生成を行う詳細オプション

選択	操作
階級の数 # を自動選択	最も適切な階級数が自動的に決められます。
グループ化: __ 階級	データセットは、ユーザが指定する階級数に分割されます。
次のようなアプローチ	階級作成に対するアプローチをメニューから 1 つ選択します。

表 3-2 (続き) 階級生成を行う詳細オプション

選択	操作
エントロピー	階級内のラベルの値が類似したものになるように階級が生成されます。これにより、エントロピ、または階級を生成するラベルの間の不規則性が最小化されます。
均一な区間	範囲全体に均一に分散された階級に分割されます。
均一な重み付け	階級ごとのデータの重み付けが同じになるように階級が生成されます。つまり、階級ごとのインスタンス数が同じになるような階級に分割されます。
端数の切り捨て	「均一な区間」と「均一な重み付け」のときだけ使用します。階級区間から外れ値を除去できます。外れ値とは、データの塊の外にある単独のインスタンスです。
訓練事例のみを使用	データセットをサブセットに分割し、予備比率だけを使用します。
予備比率	階級を作成するために使用するデータセットの割合を指定します。
ランダムシード	データセットの分割をどこから始めるかを指定します。ランダムシードの数値は時刻によって異なります。次の実行時にまったく同じデータを使用したい場合は、同じランダムシードを指定する必要があります。

「区間の指定」を選択して独自の区間を設定する場合は、さらに次の入力ができます。

• 区間

テキストフィールドに、区間の各範囲をコンマで区切って指定します。数字が範囲に含まれない場合は、数字の横に丸かっこを、範囲に含まれる場合は、角かっこをそれぞれ指定します。たとえば、年齢が 18 より上で 30 以下の場合は、(18...30] のように表します。

「指定の区間を使用」フィールド（入力領域）に日付を入力する場合は、日付を二重引用符で囲む必要があります。引用符で囲まないと、日付は自動的に加算に使用されます。

"1/1/99", "2/1/99", "3/1/99", "4/1/99", "5/1/99", "6/1/99"

"05/22/1999" のような mm/dd/yyyy 形式も使用できます。

- 均一な区間設定

このボタンをクリックすると、その下のテキストフィールドがアクティブになります。階級を生成する範囲全体の始めと終り、階級のサイズをそれぞれ適切なフィールドに入力します。

「均一な区間設定」の場合は、日付を引用符で囲みません。日付形式は、1/1/99 でも 01/01/99 でもかまいません。

詳細は、『*MineSet 3.0 Enterprise Edition Reference Guide*』の階級生成の説明を参照してください。

集計処理による新規項目の作成

集計処理を使って既存項目の内容から新規項目を作成することができます。たとえば、churn データベースから州当たりの平均解約数が入った項目を新たに作成し、データのサブセットを作成できます。その他にも、さまざまな条件を使って項目の分布を作成できます。集計処理ではデータの形状が変わります。この機能を完全に理解するためには、『*MineSet 3.0 Enterprise Edition Reference Guide*』を参照し、その背景知識として配列や分布について理解する必要があります。さらに、用語集の「集計処理 (Aggregate)」を参照してください。

集計処理では元のデータセットの行が別々の集計行にグループ化されるため、全体の行数は減少します。

集計処理を行うには、次の手順に従います。

1. Tool Manager の「データ変換」パネルで「集計処理」をクリックしてダイアログ・ボックスを表示し、簡単な集計処理を行うか、配列を作成するか、項目を分割させるかを指定します (図 3-5 を参照)。最初に現在のデータセットの項目が中央の「項目によるグループ化」テキスト・フィールドに表示されます。

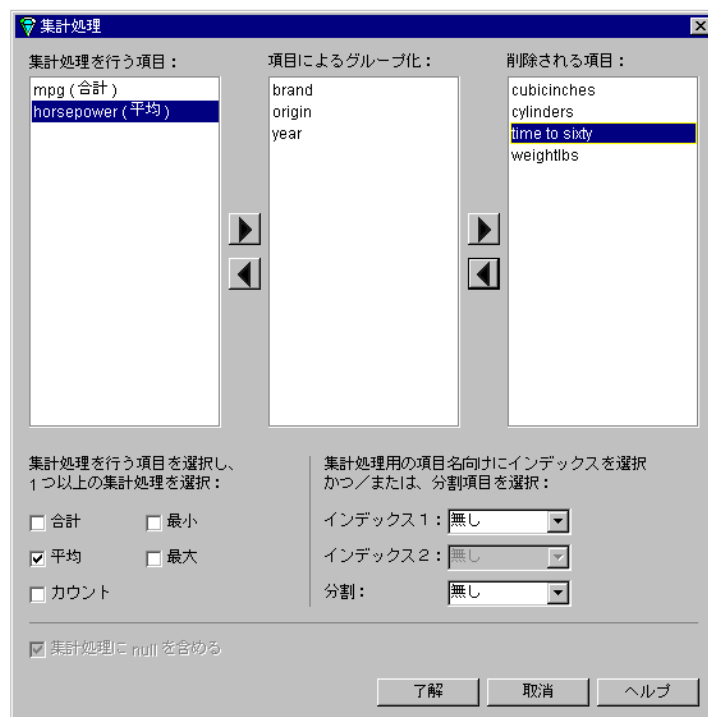


図 3-5 「集計処理」ダイアログ・ボックス

2. 集計処理する項目を選択し、左矢印ボタンをクリックして項目を左へ移動します。「項目によるグループ化」に使用するエントリ（インデックスに使用する項目）は中央のパネルに残し、それ以外を右に移動します。複数の項目を選択するには、<Ctrl> キーを押したまま項目を選択します。

集計処理を行う項目 — 集計処理に使用する項目をここに移動します。その下にあるボックスのどれがチェックされているかにより、集計処理する項目の値に対し合計、平均、最小、最大、カウントの操作が行われます。

項目によるグループ化 — ここに残っている項目は操作の間も変更されません。同じ値の組み合わせをもつレコード群からは、1 レコードだけが結果テーブルに出力されます。

注記：「項目によるグループ化」の項目が多数あったり、数値項目であると、結果テーブルには予想より多くの行が出力されることがあります。一般に「項目によるグループ化」には、値の数が少ない項目をいくつか使用します。

削除される項目 — 集計処理では、削除する項目は使用しません。

3. その下のチェックボックスをクリックして、集計処理で値をどのように結合するかを指定します。合計、平均、最小、最大、カウントの操作が選択できます。最小と最大はまれにしか使用しません。

数値項目の場合はどの組合せでも使用できますが、ほかの型の場合は「カウント」しか使用できません。結果はオプションごとに別の項目に出力されます。たとえば、平均と最大を選択すると、1つの項目に平均値が、別の項目に最大値が出力されます。どの操作が行われたかは、デフォルトの項目名でわかります。たとえば、`count_state`、`avg_weightlbs` などです。

4. 右下のポップアップ・メニューを使って、インデックス（結果が配列の場合）と分割項目（結果を分割する場合）を指定します。

配列はスライダを作成するときに便利です。いくつかのビジュアライザで利用できるアニメーションにはスライダが必要です。配列のインデックスに使用する項目は、階級が前もって生成されていなければなりません。これは、ある特性が別の特性との対比で（別の特性によるインデックスによって）どう違うかを示すためです。配列と分割の詳細は、『*MineSet 3.0 Enterprise Edition Reference Guide*』の集計処理の説明を参照してください。

5. 「了解」をクリックします。
6. 「データ変換」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、上の条件が適用され作成された新しい項目名が表示されているか確認してください。

集計処理の使用例については、『*MineSet 3.0 Enterprise Edition Tutorial*』を参照してください。州別の平均解約が作成されています。

フィルタによる項目内容の限定

項目値に基づく表現を使ってデータのサブセットを選択することができます。たとえば、年齢が 25 歳未満とか、ガロン当たりの燃費が 25 マイル未満のレコードだけを選択できます。フィルタ処理を行うには、Tool Manager を使ってデータセットをロードした後で、次の手順に従います。

1. Tool Manager の「データ変換」パネルで「フィルタ」をクリックしてダイアログ・ボックスを表示し、表現を使ってデータのフィルタリングを行います（図 3-6 を参照）。



図 3-6 「フィルタ (Filter)」 ダイアログ・ボックス

2. 左側から項目名と演算子を選択して、右側に表現を作成します。あるいは、「表現による定義」テキストフィールドに表現を直接入力することもできます。結果テーブルには、表現が真（数値の場合は非ゼロ）になるレコードだけが出力されます。さらに、『*MineSet 3.0 Enterprise Edition Reference Guide*』のフィルタパネルの説明を参照してください。表現定義言語の詳細は、『*MineSet 3.0 Enterprise Edition Interface Guide*』を参照してください。.

項目の型や名前の変更

データベースの中には、数値を文字列として格納するものがあります。たとえば、Oracle データベースは、整数や実数などすべての数値を同じ形式で格納します。Tool Manager は、デフォルトでこのような数値のデータ型を倍精度型とします。これを変更したい場合は、「データ型の変更」ボタンを使用します。項目名を変更する場合にも同じボタンを使用します。

項目のデータ型の変更

1. Tool Manager の「データ変換」パネルで「項目名／型の変更」をクリックします。ダイアログ・ボックスが表示され、そこに項目名とデータ型が表示されます (図 3-7)。
2. 変換する項目を選択し、「新しい型」メニューをクリックします。メニューにデータ型の一覧が表示されます。無効なデータ型はグレイで示されます。表 3-3 に、このメニューの拡張を示します。

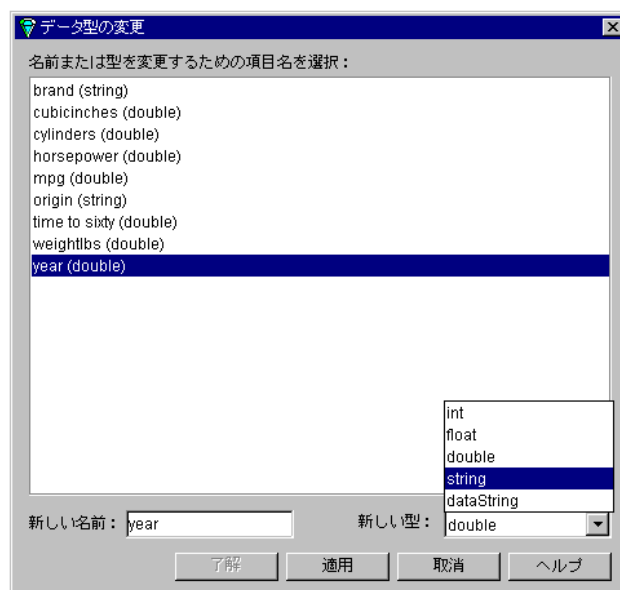


図 3-7 データ型のポップアップリスト

表 3-3 項目データ型の意味

データ型	意味
整数	32 ビット符号付き整数
浮動小数点	単精度浮動小数点数。小数点は任意
倍精度	倍精度浮動小数点数。小数点は任意
dataString	<i>dataString</i> の値は共通の文字列テーブルには保持されません。したがって、同じ値がいくつあっても、そのデータ文字列は個別に格納されます。まれにしか使用しません。
文字列	共通の文字列テーブルに格納される文字列です。メモリーの節約になります。文字列を使用すべきか <i>dataString</i> をすべきかわからない場合は、文字列を使用してください。
日付	そのデータベースの日付型として指定されます。
階級	階級生成操作で作成された項目

- 3. 新しいデータ型を選択したら、「適用」をクリックして変更を有効にします。
- 4. 「閉じる」をクリックしてダイアログ・ボックスを閉じます。

名前など、適切でないフィールドを数値に変換すると、結果の値は全体がゼロになります。

項目名の変更

項目名の変更は、項目のデータ型の変更と同じダイアログ・ボックスで行います。

- 1. 「データ型の変更」ダイアログボックスで、名前を変更する項目を選択し、「新しい名前」に新しい名前を入力し、「適用」をクリックします。
- 2. 「閉じる」をクリックしてダイアログ・ボックスを閉じます。

モデルの適用

モデルがすでに作成されていれば、「モデルの適用」ボタンを使って、現在のテーブルにある新しいレコードにラベルを付けたり、ラベルの値の確率を推定したり、現在のテーブルに対しモデルの性能を評価したり、現在のテーブルを既存のモデルにバックフィットしたりすることができます。第 7 章「予測モデリングの理解」にこれらの例があります。

データの標本抽出

データセットが大きすぎて作業効率が悪い場合は、ランダムに標本抽出したサブセットを使用できます。その場合は、データファイルをロードしてから、Tool Manager の「データ変換」パネルで次の手順を行います。

1. 「標本」ボタンをクリックして「標本抽出」ダイアログ・ボックスを表示します (図 3-8)。
2. 次のどちらかのオプションを選択します。

「パーセンテージ」チェックボックスをクリックし、現在のデータセットから標本抽出したい標本抽出率をテキスト・フィールドに入力します。標本抽出率はおおよそその値です。

「カウント」チェックボックスをクリックし、標本に含めるレコード数を指定します。

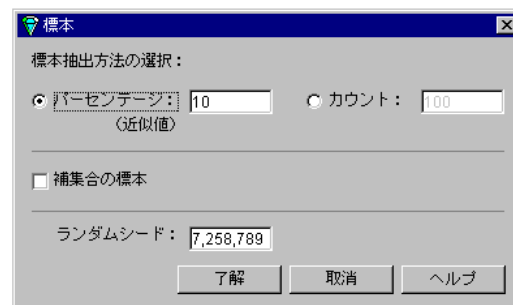


図 3-8 「標本」ダイアログ・ボックス

- 3. 「補集合の標本」チェックボックスをクリックして、ランダム・サンプルに含まれないすべてのレコードを取得します。たとえば、「補集合の標本」ボックスがオフのときに 10% の標本を要求し、その後でこのボックスをオンにすると、残りの 90% のデータが得られます。「了解」をクリックします。
- 4. (オプション) ランダムな標本抽出に使用するシード値を「ランダムシード」フィールドに入力します。シード値を指定しないと、現在の時刻に基づく数値がシード値として使用されます。異なるランダムな標本抽出が必要なときは、異なるランダムシードを指定します。データセットのマイニングのたびに同じシード値を使えば、同じランダム標本を使用できます。発見がデータセット全体に対してどの程度有効であるかを調べたい場合は、シード値を変更します。用語集の「ランダム・シード (Random Seed)」も参照してください。

テーブル履歴による手順のトレース

「データ変換」パネル下部にある 2 つの「テーブル履歴」ボタンを使えば、操作の履歴を表示できます。あるいは、考えが変わったり、誤りを発見した場合には、前に戻ることができます。左矢印ボタンをクリックすると、以前の手順のテーブルが表示されます。右矢印ボタンをクリックすると、テーブルは現在の状態に戻ります。表 3-4 にボタンの意味を示します。

表 3-4 テーブル履歴の意味

フィールド	意味
現在の履歴表示は	行われた変更の回数と、現在の手順がそのどれに当たるのかを示します。
前処理 : および次処理 :	取られたアクションを示し、追跡を容易にします。
前処理を編集	「前処理 :」フィールドに表示されている操作を編集するためのダイアログ・ボックスが表示されます。後続の操作で使用されている操作を削除するときには、注意が必要です。
最終操作まで削除	現在の手順より後のすべての操作を削除します。すでに履歴の最後に達している場合を除き、ボタンをクリックした後で「データの可視化／マイニング」パネルが再びアクティブになります。
履歴の表示タブ	データ変換テーブルの完全な履歴を表示します (図 3-9)。
データ操作及び、データの可視化／マイニングの表示タブ	履歴のある時点に戻ります。Tool Manager はそのときの状態に戻ります。

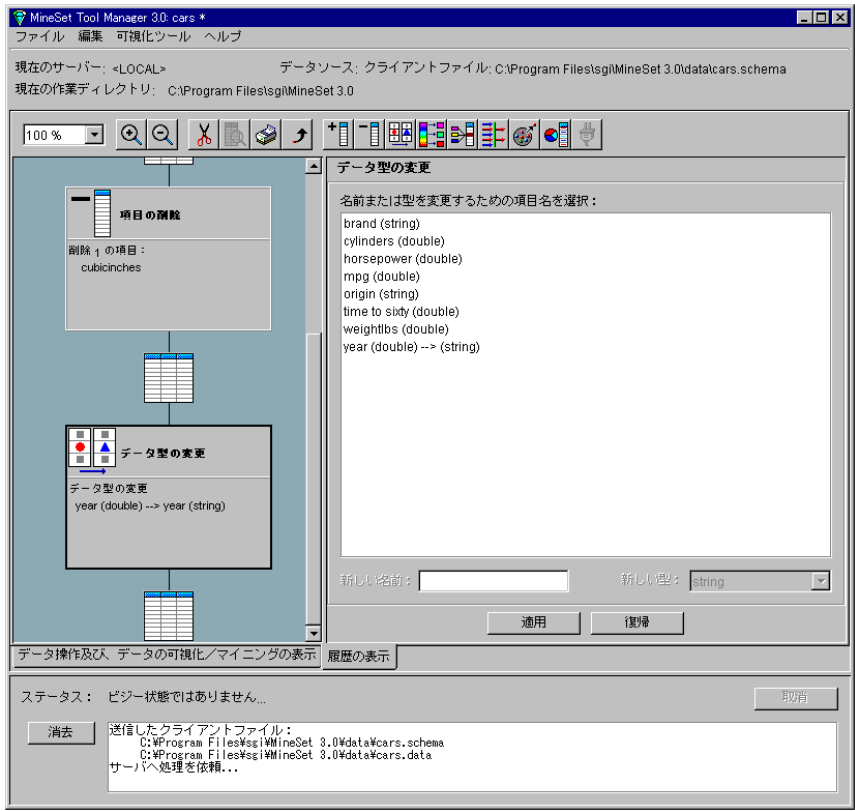


図 3-9 「履歴の表示」ダイアログ・ボックス

「前処理を編集」の場合のように、ある操作を変更すると、通常、履歴の後の操作に影響が及びます。場合によっては、後の操作が無効になることもあります。特定の操作を選択して、編集、追加、表示を行うことができます。変更が履歴に影響を与えるときには、「履歴の表示」に警告が表示されます。さらに、新しい履歴が発生するたびにその履歴がここに表示されます。以前の操作に戻るには、ダイアグラムをクリックします。ある状態と状態の間のテーブルに戻るには、操作と操作の間のテーブルをクリックします。

Tool Manager を以前の状態に戻すには、「データ操作及び、データの可視化／マイニングの表示」タブをクリックします。

レコードの重み付け

データ変換のもう 1 つの方法にレコードの重み付けがあります。重み付けでは、必要に応じてあるレコードを他のレコードより重要としたり、重要でないとしたりすることができます。たとえば、電話会社はすべての不正通話をデータベースに保存しますが、正常な通話はほんの一部しか保存しません。レコードの重み付けを使用すれば、顧客全体における各レコードの真の割合を知ることができます。

データセットによっては、集計処理がすでに行われていて、レコードにそれと関連する「カウント」が自然に備わっていることがあります。たとえば、米国の市に関する統計には、通常、市の人口が含まれています。このカウント属性を重み付けに対応付ければ、レコードをカウントの数だけ複製したのと同じことになります。

レコードを重み付けすると、重み付け 2 のレコードは重み付け 1 のレコード 2 つと同じことになります。浮動小数点の重み付けも使用できます。

レコードの重み付けは、どの分析ツールの「詳細設定」でも行うことができます。分析ツールについては、用語集の「分析 (Inducer)」を参照してください。重み付けを行うには、データセットをロードした後で次の手順に従います。

1. Tool Manager の「データの可視化／マイニング」パネルで「マイニングツール」を選択し、「クラス判別」タブをクリックします。
2. 「モード」ポップアップ・メニューから分析ツールをどれか選択します。
3. 「詳細設定」ボタンをクリックしてダイアログ・ボックスを表示し、「重み付けとして使用」チェックボックスをクリックします。
4. 重み付けに使用する項目を選択します。

- 手順4で使用した属性を通常の属性としても使用する場合は、「属性としても使用」チェックボックスをクリックします。このボックスをクリックしなければ、この重み付け属性は分析処理に影響を与えません。
- 「了解」をクリックしてオプションを使用し、「実行」をクリックして、選択した分析ツールを実行します。

重要項目の検索

重要項目は、選択したラベル項目のさまざまな値を予測する上でどれが最も重要な項目かを発見するときに使います。重要項目とクラスタリングの違いは、重要項目では、項目の重要度を判定するときにどのラベルを使用するかをユーザが決める必要があるのに対し、クラスタリングでは、データ自体が識別要素を明らかにしてくれます。ただし、ラベルは生成されません。クラスタリングについては、[第12章「クラスタリングによるデータのセグメント化」](#)を参照してください。

たとえば、重要項目は、スキャタ・ビジュアライザ (Scatter Visualizer) の軸に割り当てる最も重要な3つの項目を見つけるときに使用できます。この場合、ラベルを選択し、ツールを実行するだけで、最も重要な識別要素として3つの項目がウィンドウに表示されます。これらの項目がさまざまなラベルをどのくらい識別できるかは、0から100までの「寄与率」と呼ぶ尺度で示されます。項目をさらに追加すると、寄与率は上がります。

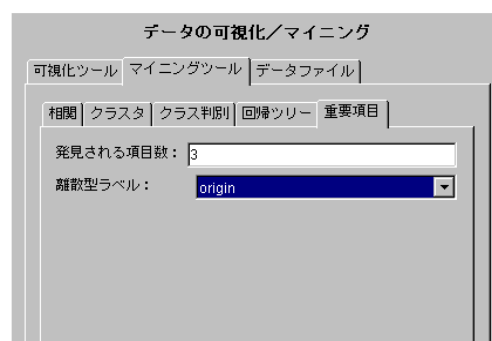


図 3-10 「重要項目」タブ

1. Tool Manager の「データ変換」パネルで「マイニングツール」タブを選択し、「重要項目」タブをクリックします (図 3-10 を参照)。
2. 標準モードで実行する場合は、ポップアップ・メニューから離散型ラベルを選択し、「発見される項目数」フィールドに数値を入力し、「実行」をクリックします。下部のステータス・ウィンドウに結果が表示されます。

項目の選択を制御する場合は、「詳細モード」ボタンをクリックして「重要項目」ダイアログ・ボックスを表示します (図 3-11)。
3. このダイアログ・ボックスでは、「重み付けとして使用」チェックボックスをクリックし、「次の項目に関連する重要項目」ポップアップ・メニューから項目を選択できます。これを行うと、その項目が、重要度を判定するための正規属性となるかどうかを判断する助けとなります。さらに、「[レコードの重み付け](#)」(50 ページ) も参照してください。ツール操作の詳細は、『*MineSet 3.0 Enterprise Edition Reference Guide*』の重要項目の説明を参照してください。
4. ダイアログ・ボックスには 2 つの項目名リストがあります。左のリストには使用可能なすべての属性が、右のリストには、ユーザまたは「重要項目」アルゴリズムが重要と認めた属性がそれぞれ表示されます。項目を選択した後でそれを左から右に移動するには、右矢印をクリックします。

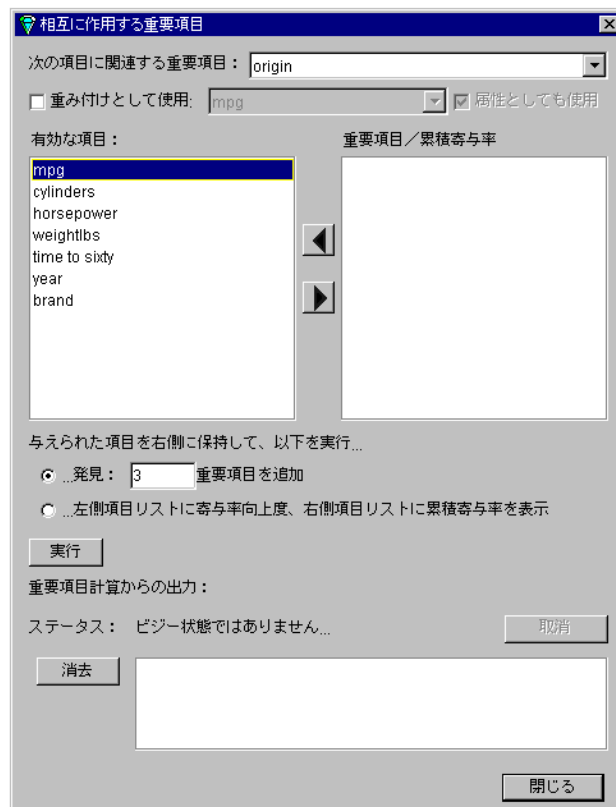


図 3-11 「重要項目」の「詳細モード」

5. ダイアログ・ボックスの中央には、2つの詳細モード選択肢があります。

- いくつかの重要な属性を検索する

「... 発見: [数個] 重要項目を追加」ボタンをクリックします。それ以上の変更をせずに「実行」をクリックすると、結果は標準モードの場合と同じになります。指定した数の重要項目が検索され、右側の項目に移されます。累積寄与率が与えられます。

ただし、右矢印をクリックして項目名を左から右に移動すれば、組み込む項目をさらに指定し、追加することができます。

「実行」をクリックすると、各項目の累積寄与率が、リストにあった前の累積寄与率とともに表示されます。寄与率が 100 というのは、その項目を使用すれば、データセット内の異なるラベルの値を完全に識別できることを意味します。

- 使用可能な属性をランク付けする

項目を追加することで、右側に表示されている寄与率がどのくらい増加するかをテストできます。たとえば、[図 3-11](#) に示す cars データセットを使用する場合、項目 *cylinders* を右側のリストに移動します。このようにして、左側の列に残っている各項目が寄与率の向上に貢献するかどうかを計算できます。すでに重要性が認められている右側の項目に対し累積寄与率が計算され、結果が右側の項目に表示されます。

「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」ボタンをクリックします。このサブモードを使えば、この処理を細かく制御できます。2つの項目のランク付けが非常に接近している場合は、一方を優先的に使用できます。たとえば、安価に収集できるとか、信頼性が高いとか、理解が容易といった場合です。

重要項目は、スキャタ・ビジュアライザ (Scatter Visualizer) やスプラット・ビジュアライザ (Splat Visualizer) の最も重要な 3つの軸を見つけるときに便利です。さらに、重要項目は、ツリー・ビジュアライザ (Tree Visualizer) を使って識別力のある階層（異なるラベルの値を分離する階層）を見つけたり、ラベルをツリー・ビジュアライザのキーとして選択するときにも便利です。

「重要項目」でサンプル・ファイルを使用する方法については、『*MineSet 3.0 Enterprise Edition Reference Guide*』の重要項目の説明を参照してください。

スキャタ・ビジュアライザとスプラット・ビジュアライザによるデータの分析

スキャタ・ビジュアライザ (Scatter Visualizer) とスプラット・ビジュアライザ (Splat Visualizer) の外観と使い方はよく似ています。スキャタ・ビジュアライザでは、データが一連の個別要素として表示されるのに対し、スプラット・ビジュアライザでは、データポイントの集約がさまざまな不透明度と色付きの雲として 3 次元ランドスケープに表示されます。これらのビジュアライザを詳しく知りたい場合は、『*MineSet 3.0 Enterprise Edition Reference Guide*』を参照してください。この章の構成は次のとおりです。

- ・ 「スキャタ・ビジュアライザとスプラット・ビジュアライザの概要」 (55 ページ)
- ・ 「スキャタ・ビジュアライザやスプラット・ビジュアライザのためのデータ変換」 (60 ページ)
- ・ 「スキャタ・ビジュアライザとスプラット・ビジュアライザの起動」 (64 ページ)
- ・ 「スキャタ・ビジュアライザとスプラット・ビジュアライザでの結果の表示」 (69 ページ)
- ・ 「スキャタ・ビジュアライザとスプラット・ビジュアライザでのアニメーションの作成」 (75 ページ)
- ・ 「スキャタ・ビジュアライザとスプラット・ビジュアライザでの結果の操作」 (82 ページ)

スキャタおよびスプラット・ビジュアライザの特徴と機能を示すサンプルの設定ファイルとデータファイルが提供されています。サンプルファイルは、MineSet を最初にインストールした *MineSet 3.0\examples* ディレクトリにあります。

スキャタ・ビジュアライザとスプラット・ビジュアライザの概要

この節では、スキャタおよびスプラット・ビジュアライザの使い方と、その使い分けの理由を説明します。

スキャタ・ビジュアライザの概要

スキャタ・ビジュアライザの個々のデータポイントはデータファイルの行に対応しています。この可視化がうまく機能するためには、データポイントの数が 50,000 未満であるか、何らかの処理によりデータが少数の集約データに集計処理されている必要があります。スキャタ・ビジュアライザのスキャタ・プロットをアニメーション表示すれば、関係をさらに明確に見ることができます (図 4-1)。

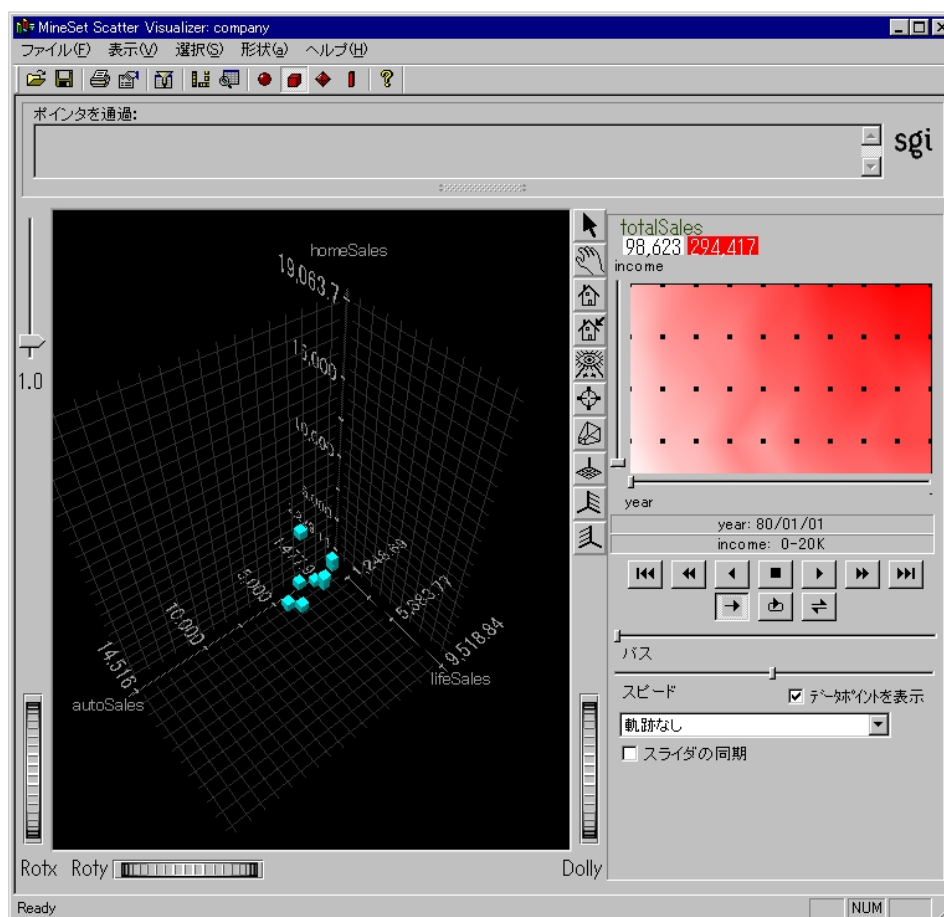


図 4-1 スキャタ・ビジュアライザの例

スキャタ・ビジュアライザでは、各項目のデータが3次元ランドスケープの要素や、軸、サイズ、色などの要素にマッピングされます。1つまたは複数の数値型変数をスライダにマッピングすれば、要素のサイズ、色、位置をアニメーション表示できます。[図 4-1](#)の例では、データはいくつかの会社の売上げを表しています。時間変数をスライダにマッピングし、売上げをサイズにマッピングすると、時間スライダの移動とともに要素が伸びたり縮んだりします。この例は *MineSet 3.0\examples\company.scatterviz* にあります。

アニメーション・パスを再生すると、さまざまに変わるデータポイントのサイズ、色、モーションから傾向や異常値を見つけることができます。3次元ランドスケープの表示をいろいろ変えると、興味深い瞬間を見つけたり、変数値のスケールを変えて強調表示したりすることができます。シーンを簡素にしたい場合は、表示をフィルタリングして一定の条件に合う要素だけを表示できます。

スプラット・ビジュアライザの概要

スプラット・ビジュアライザでは、いくつかの変数間の関係を視覚的に分析できます。アニメーション機能を使用すると、この関係が明確になる場合があります。スプラット・ビジュアライザでは、スプラットと呼ぶグラフィカル・オブジェクトを使用します。これはデータポイントを集計したものです。スプラットの色や不透明度はアニメーションで変化しますが、位置は変化しません。

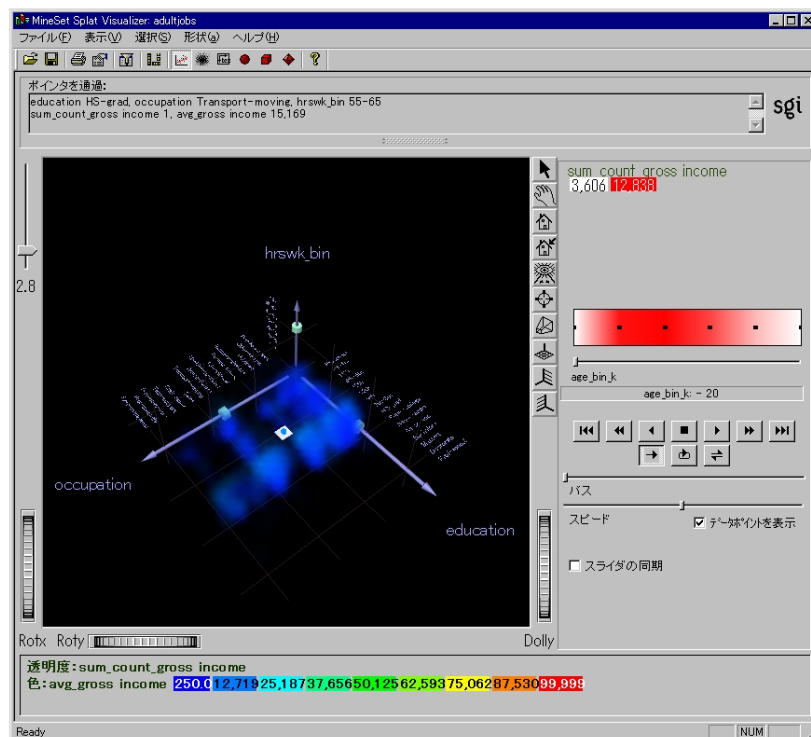


図 4-2 1次元サマリスライダを設定したスプラット・ビジュアライザの例

図 4-2 では、スプラット・ビジュアライザの 3 次元ランドスケープに、サンプル・データセット *adult94* の項目が軸、スライダ、色、不透明度にマッピングされています。これはスキャタ・ビジュアライザに似ていますが、スキャタ・ビジュアライザではデータポイントを個別に描画するのに対し、スプラット・ビジュアライザでは、相互に近い（同じ階級に入る）データポイントが集計処理され、単一のスプラットとして描画されます。結果は、各ポイントのスキャタで描画したら得られるであろうイメージに似ています。結果のイメージは、3D カラー・ヒストグラムとみなすことができます。

スプラット・ビジュアライザの結果から次のことが分かります。

- ・ アニメーション・パネルを使って、データの全体的な変化や傾向を知ることができます。色や不透明度を変えると、実際に動いているような感覚になります。このマニュアルのオンライン版で、色付きの図をご覧ください。

- 3次元ランドスケープの上をフライオーバーすることによって、特定の次元や視点の表示を強調できます。
- メインウィンドウ左上のスケールスライダを使ってスプラットの不透明度を増減できます。データ密度の濃い領域は、多数の値の平均に基づいているため、色の変化はあまりありません。
- 表示をフィルタリングして一定の条件に合うスプラットだけを表示できます。軸、スライダ、重み付け、色に対応する項目をフィルタできます。
- ボリューム内の個々のスプラットのテキスト情報を取り出すことができます。
- ボックスセレクトで選択した領域を定義して、オリジナルデータに至るまでドリルスルーしたり、Tool Manager に送信することができます。

たとえば、[図 4-2](#)の左の軸には、平均収入でソートされた職業が軸に沿って表示されています。軸の終りの方にある executive-managerial（役員 / 管理職）は、平均収入が最も高く、値が自然に伸びています。一方、[図 4-2](#)の右の軸にある education の値の順序は、大体、低から高に渡っていますが、いくつかのケースでは順序に正常でないものが見られます。この予期しない順序は重要な点かもしれません。これは予測に合致しないデータの場所を示しているからです。

スキャタ・ビジュアライザやスプラット・ビジュアライザのためのデータ変換

データ変換の決定には、データの綿密な計画や少なからぬ知識が必要です。データセットを構成する項目を軸や色、スライダにマッピングする場合には、その前に何らかの操作や変換が必要になることがよくあります。さらに、結果を見てから前に戻り、データを別の形に変換することもよくあります。

MineSet では、ほとんどの変換は Tool Manager の「データ変換」パネルで行います。[「Tool Manager によるデータの変換」\(34 ページ\)](#) を参照してください。考慮すべき変換には次のものがあります。

- 項目を追加する。それには、通常、表現を既存データに適用します。
- 適切でない項目や重複している項目、あたりまえの予測や興味のない予測となりそうな項目を削除する。これは、データセットのサイズを削減し、処理の速度を速めるのに役立ちます。
- データをフィルタリングする。たとえば、特定の属性に対し一定範囲の値を持つレコードだけを使用するような場合です。
- 項目の名前や型を変更する。
- データの階級を生成する。つまり、連続したデータを個別のセグメントにグループ化します。
- (スプラット・ビジュアライザのみ) データを集計処理する。Tool Manager を使ってデータを手作業で集計処理しないと、集計処理は Tool Manager によって自動的に行われます。しかし、集計処理はできるだけ Tool Manager で明示的に行ってください。そうすれば、処理がサーバで行われるため、集計処理されたデータセットはクライアント・システムに送信されません。[「スプラット・ビジュアライザの処理技法」\(62 ページ\)](#) を参照してください。
- (スキャタ・ビジュアライザのみ) データを集計処理する。集計処理を行えば、大きなデータセットを少数の集計処理データに削減できます。集計処理後のテーブルの各行がスキャタの要素に対応します。集計処理手順では、アニメーション用の配列項目を作成できます。

スプラット・ビジュアライザにおける可視化の対話性能は、表されるデータポイントの数とは関係ありません。対話性能は、それぞれの軸次元の階級に依存します。データセットが非常に大きい場合は、集計処理を Tool Manager で明示的に行ってください。そうすれば、Tool Manager がストリーミング操作を使ってサーバでこの処理を行いますので、データセット全体がクライアントに送信され、そこで処理されることはありません。「[スプラット・ビジュアライザの処理技法](#)」(62 ページ)を参照してください。

[表 4-1](#) に、スプラット・ビジュアライザの可視化要素にマッピングできる項目型を示します。

表 4-1 スプラット・ビジュアライザの可視化要素にマッピングできる項目型

要素	項目型	要件
軸	数値	手動または自動的に Tool Manager によって階級が生成されます。
	階級	直接マッピングできます。
	文字列	直接マッピングできます。
色	数値	直接マッピングできます。
不透明度	数値	何もマッピングしないと、デフォルトでレコード数が使用されます。
スライド	数値	手動または自動的に Tool Manager によって階級が生成されなければなりません。
サマリ	数値	直接マッピングできます。

[表 4-2](#) に、スキヤタ・ビジュアライザの可視化要素にマッピングできる項目型を示します。

表 4-2 スキヤタ・ビジュアライザの可視化要素にマッピングできる項目型

要素	項目型	要件
軸	配列を含む任意の型	直接マッピングできます。
要素 - サイズ	数値または配列	直接マッピングできます。
要素 - 色	配列を含む任意の型	直接マッピングできます。
要素 - ラベル	配列を含む任意の型	直接マッピングできます。

表 4-2 (続き) スキャタ・ビジュアライザの可視化要素にマッピングできる項目型

要素	項目型	要件
スライダ	階級	集計処理時に暗黙的にマッピングできます。
サマリ	数値	集計処理時に暗黙的にマッピングできます。

文字列項目を軸にマッピングすると、階級がその項目の個別値になります。これらの値は、色にマッピングされている属性と対応するように並べられます。色にマッピングされている属性がない場合は、不透明度と対応します。文字列値の軸に沿って色が変わるのを見れば、その項目と色にマッピングされた項目がどのくらい相関しているかが分かります。

スプラット・ビジュアライザの処理技法

Tool Manager で集計処理を行えば、クライアントでの処理を避けることができます。その場合、ほとんどの作業はサーバで行われます。一般にサーバはクライアントよりも強力なマシンです。サーバではストリーミング操作を行いますので、データセット全体をメモリーにロードする必要はありません。さらに、集計処理を Tool Manager で行うと、大きなデータセットをクライアントに送信する必要がなくなります。集計処理を行うには、次の手順に従います。

- 1. Tool Manager を使って、軸やスライダに使用する数値型項目の階級を生成します。
- 2. 色にマッピングする項目をカウントと平均で集計処理しながら、軸項目とスライダ項目でグループ化します。
- 3. 結果のカウント集計処理を不透明度にマッピングします。
- 4. 結果の平均集計を色にマッピングします。

MineSet に添付されている *adult94* データセットを使用すると、上に示した処理の手順は次のようになります。

- 1. *age* と *hours_per_week* の階級を生成します。前者をスライダに、後者を階級生成された数値の例として軸にそれぞれ使用します。スライダには、階級生成された値が必要です。
- 2. *gross_income* をカウントと平均を使って集計処理します。*education*、*occupation*、*age_bin*、*hours_per_week_bin* を「項目によるグループ化」パネルに保持し、他のすべての項目を削除します。これによって、*gross_income* が他の 3 つの条件に従ってどのように変化するかを可視化できます。

3. `education`、`occupation`、`hours_per_week_bin` をそれぞれの軸にマッピングします。
4. `avg_gross_income` を色に、`count_gross_income` を不透明度に、`age_bin` をスライダにそれぞれマッピングします。これにより、軸にマッピングされている条件に従って、色と不透明度が年齢に応じた収入（平均とカウント）の変化を示します。

ツールを起動するときには、すべての処理がサーバで行われることと、データファイル *adult94.splatviz.data* にはオリジナルデータの行を集計処理したものが行として含まれていることに注意してください。そのため、[図 4-2](#) に示す可視化と同じものが表示されます。

スキャタ・ビジュアライザとスプラット・ビジュアライザの起動

スキャタ・ビジュアライザやスプラット・ビジュアライザを起動する最も簡単な方法は Tool Manager から起動する方法です。

1. Tool Manager を起動し、サーバとデータソースを選択します。これについては、「[MineSet の起動](#)」(15 ページ) を参照してください。この最初の例では、*churn* データセットを使用します。
2. 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」タブを選択し、その下に並んでいるタブからスキャタ・ビジュアライザかスプラット・ビジュアライザを選択します (図 4-3 を参照)。

スキャタ・ビジュアライザでの可視化要素のマッピング

1. Tool Manager の「データ変換」パネルで、「現在のデータセットの項目名」をスクロールし、項目を選択することによって「現在のデータセットの項目名」を「可視化要素」にマッピングします。たとえば、*churn* データベースを使用する場合には、軸 1 に number of vmail message (ボイスメールメッセージの数) を、軸 2 に total day charge (昼間の合計料金) を、軸 3 に total eve charge (夜間の合計料金) をそれぞれ選択します。

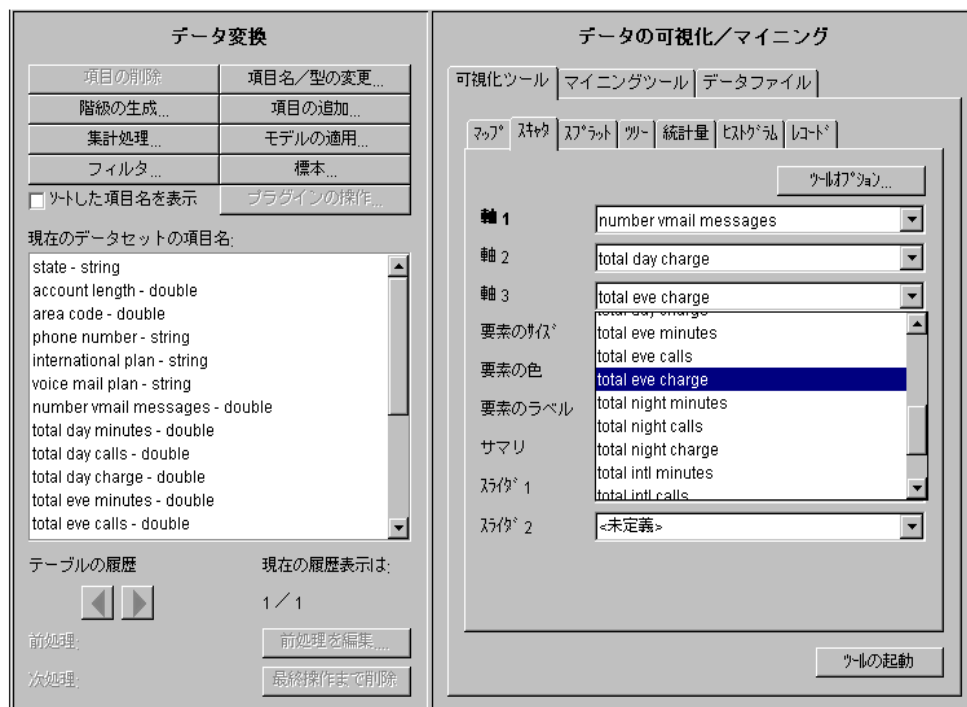


表 4-3 に、スキャタ・ビジュアライザにおける可視化要素のマッピング効果を示します。

表 4-3 スキャタ・ビジュアライザでの可視化要素のマッピング

可視化要素	操作:
軸 1	通常、データを最初の軸に割り当てただけでは役に立ちません。3 次元グラフを生成するには、データをすべての軸に割り当てる必要があります。
軸 2	
軸 3	
要素のサイズ	サイズ、色、ラベルを要素に割り当てます。「スキャタビジュアライザのオプション」ダイアログ・ボックスでこれらのマッピングのオプションを指定してください。このダイアログを表示するには、スキャタ・ビジュアライザの「ツールオプション」ボタンをクリックします。
要素の色	
要素のラベル	
サマリ	サマリスライダの背景色を決めます。
スライダ 1	項目を 1 つまたは 2 つのアニメーション・スライダに直接マッピングします。
スライダ 2	

- 2. マッピングを取消すには、要素を選択し、使用可能な項目のリストから <未定義> を選択します。独自のデータベースを使用する場合は、可能なマッピングを判別するために「重要項目」ツールを実行してみてください。これについては、[「重要項目の検索」](#)を参照してください。
- 3. 「データの可視化／マイニング」パネル右下の「ツールの起動」をクリックします。

スプラット・ビジュアライザでの可視化要素のマッピング

スキャタ・ビジュアライザと同じように、スプラット・ビジュアライザでも起動前に項目を可視化要素にリンクする必要があります。項目によっては、操作を行わないと、可視化要素の条件に合わないものもあります。この例では、項目を操作してマッピング条件を満たすにはどうすればよいかを、付属しているサンプルデータの 1 つ「キノコ (Mashroom)」データセットを使って説明します。まず、[「スキャタ・ビジュアライザとスプラット・ビジュアライザの起動」](#) (64 ページ) のようにして Tool Manager を起動し、「キノコ (Mashroom)」データセットを選択してからスプラット・ビジュアライザを選択してください。

1. Tool Manager の「データ変換」パネルで、「現在のデータセットの項目名をスクロールし、項目を選択することによって「現在のデータセットの項目名」の項目を「可視化要素」の項目にマッピングします。

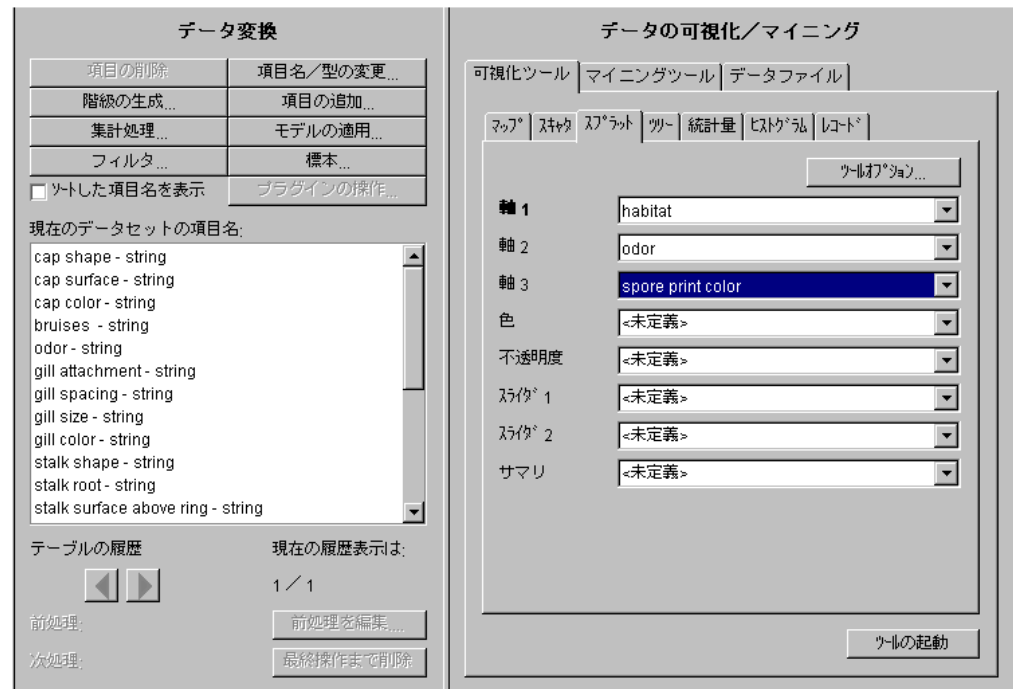


図 4-4 スプラット・ビジュアライザでの可視化要素のマッピング

この例では、軸 1 に habitat（生息環境）を、軸 2 に odor（匂い）を、軸 3 に spore print color（胞子プリント色）をそれぞれ選択します。

表 4-4 に、スプラット・ビジュアライザのマッピング効果を示します。

可視化要素	アクション
軸 1 軸 2 軸 3	通常、データを最初の軸に割り当てただけでは役に立ちません。3 次元グラフを生成するには、データをすべての軸に割り当てる必要があります。
色	スプラットの色を決めるには、数値型項目が必要です。色に何もマッピングしないと、結果はモノクロになります。
不透明度	デフォルトでは、レコードのカウントが使用されます。Tool Manager でカウントの集計処理や重み付けの合計を行う場合や、カウントや重み付けに基づいて作成された項目がすでにデータセットにある場合は、その項目が使用されます。
スライダ 1 スライダ 2	項目を 1 つまたは 2 つのスライダに直接マッピングします。項目の型は、数値か階級でなければなりません。
サマリ	サマリスライダの背景の値を決めます。サマリ項目がマッピングされていない場合は、カウントが使用されます。サマリ項目がマッピングされている場合は、その項目の重み付け平均がサマリに表示されます。

項目のマッピングで数値型項目が必要な場合は、2 つの値をもつ文字列型項目から新しい数値型項目を作成できます。その場合には、たとえば `('stringCol'=="value1")? 1:0` のような表現を使用します。詳細は、第 3 章の「[項目の削除と追加](#)」の項を参照してください。

この例では、食用に適する確率を表すために整数項目を追加し、その特性を色にマッピングします。「キノコ (*Mashroom*)」データセットには文字列型項目しかありません。文字列のような非連続型属性を平均化することはできないため、これを色にマッピングすることはできません。

2.
- Tool Manager の「データ変換」パネルで「項目の追加」ボタンをクリックします。表示されたダイアログ・ボックスで新しい項目名 `p_edible` を型 `int` で作成します。表現には `('edibility' == "edible")?100:0` を使用します。引用符に注意してください。

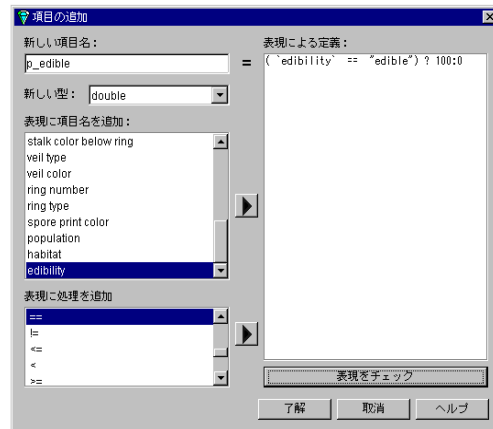


図 4-5 マッピング用の整数型項目の追加

3. p_edible を色にマッピングして、キノコが食べられるかどうかを色で表示します。

マッピングを取消するには、プルダウン・メニューから<未定義>を選択します。

4. 「データの可視化／マイニング」パネル右下の「ツールの起動」をクリックします。可視化により、キノコに関する興味深い事実が分かります。

次の節では、MineSet に付属しているサンプルデータにアクセスするもう 1 つの方法を説明します。

スキャタ・ビジュアライザとスプラット・ビジュアライザでの結果の表示

どのビジュアライザでもそうですが、設定ファイルを指定しないでスキャタ・ビジュアライザやスプラット・ビジュアライザを起動すると、「ファイル」と「ヘルプ」プルダウン・メニューしか使用できません。設定ファイルとは、名前が .scatterviz や .splatviz で終るファイルです。ビジュアライザのメインウィンドウにすべてのメニューとコントロールを表示するには、設定ファイルを開く必要があります。「ファイル」->「開く」を選択して、設定ファイルのリストを表示してください。これらのファイルは、MineSet を最初にインストールした MineSet 3.0\examples ディレクトリにあります。

有効な設定ファイルを開くと、3 次元のランドスケープが表示されます。図 4-1 の *company-total.scatterviz* には、生命保険、自動車保険、家庭保険の売上げが収入階層に対応して時間の経過とともに表示されています。

表示モード

スキャタとスプラット・ビジュアライザの表示モードには、「ハンド」と「選択」があります。このモードを相互に切り替えるには、カーソルをメインウィンドウに移動し、<Esc> キーを押します。あるいは、ビジュアライザ・ウィンドウのメインパネルの枠上の矢印アイコンまたはハンドアイコンをクリックします。

スキャタ・ビジュアライザの選択モード

スキャタ・ビジュアライザの選択モードでは、次の手順を使用します。

- 要素の情報をスキャタ・ビジュアライザ上部のテキスト・フィールドに表示するには、マウスカーソルをオブジェクトの上に移動します。
- 要素を選択するには、オブジェクトの上でマウスの左ボタンをクリックします。複数の要素を選択する場合は、<Ctrl> キーを押したままマウスの左ボタンをクリックします。
- 選択を解除するには、黒い背景をクリックします。ナビゲーションの詳細は、「[MineSet 3D ビジュアライザでのナビゲーション](#)」(27 ページ) を参照してください。

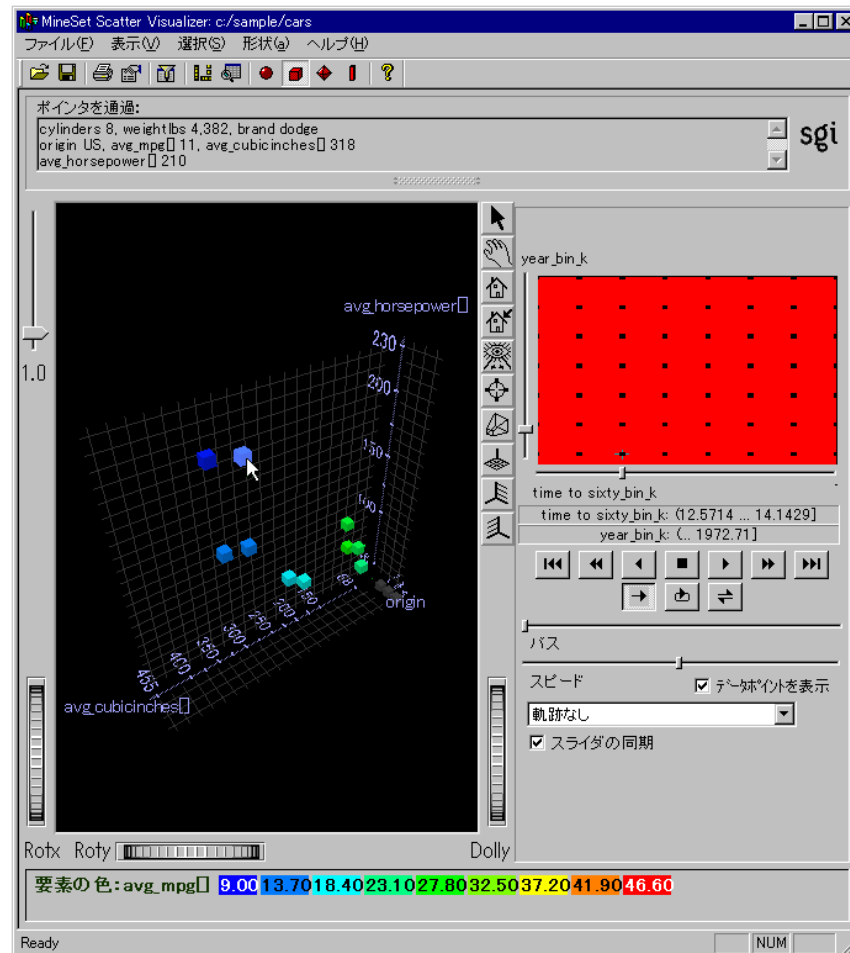


図 4-6 選択された要素情報の表示

コマンド行からスキャタ・ビジュアライザやスプラット・ビジュアライザを指定する方法については、『*MineSet 3.0 Enterprise Edition Reference Guide*』のそれぞれの項を参照してください。

スプラット・ビジュアライザの選択モード

スプラット・ビジュアライザの選択モードでは、濃いポイントの雲の中で 3-D ピックドラッグを移動することで、シーン内のさまざまな領域の情報を表示できます。ピックドラッグは円筒と四角形からなります。

- 円筒の軸と並行に移動するには、円筒をクリックし、マウスカーソルを必要な方向にドラッグします。
- 四角形の平面に限定された移動を行うには、四角形をクリックし、ドラッグします。平面内のどちらかの軸に沿った移動に限定するには、<Shift> キーを使用します。
- 平行モードと限定モードを切り替えるには、カーソルをドラッグ上に置いて <Ctrl> キーを押します。マウスボタンを押す必要はありません。
- あるいは、各軸には、ピックドラッグの位置に合わせた円盤があります。軸上の円盤を動かせばピックドラッグが移動し、ピックドラッグを移動すれば円盤が移動します。

ピックドラッグをデータ上に移動すると、円筒の色がその下のスプラットの色に変わり、その領域の情報が表示域上部に表示されます（表 4-7 を参照）。データがない場合は、円筒の色は淡いグレイのままで、ナビゲーションを助けるために位置情報が表示域上部に表示されます。

ドラッグが終り、マウスボタンを離すと、その位置にあるスプラットのメッセージが一番上の選択ウィンドウに表示されます。この選択情報はアニメーション・スライドの動きに合わせて更新されます。マウスを使えば、この情報をレポートやデータベースなど他のアプリケーションにカット & ペーストできます。

ピックドラッグをシーンから消去するには、「選択」->「ピックドラッグの表示」を選択し、これをオフにします。

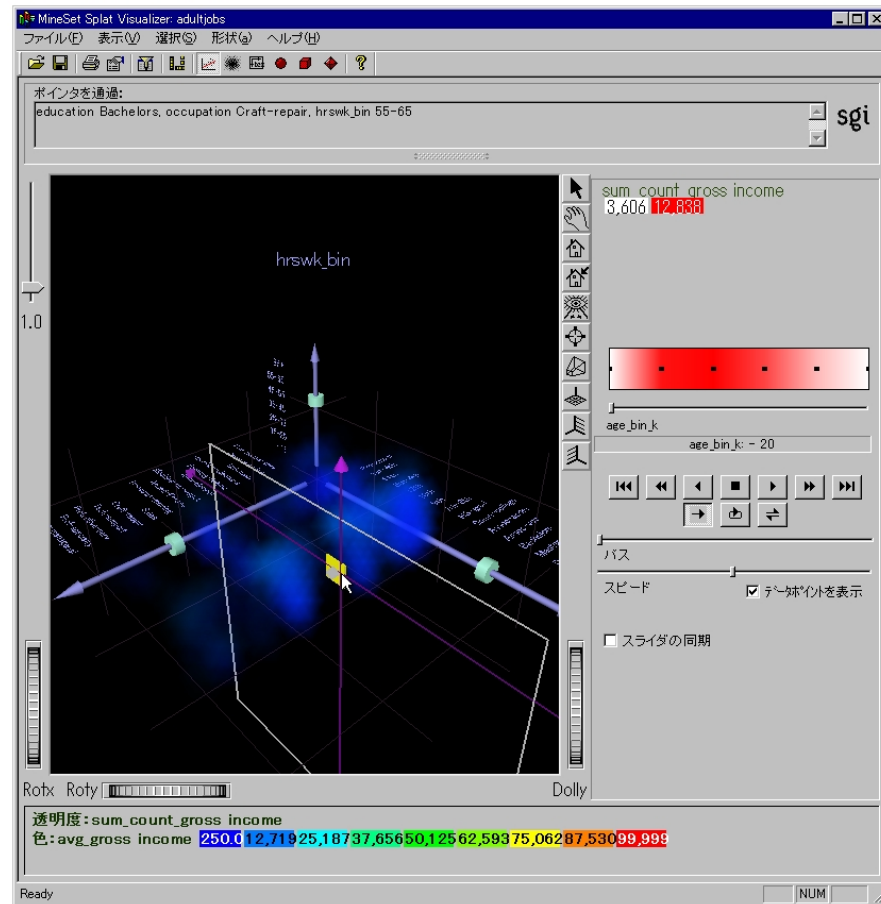


図 4-7 データ上でのピックドラッグの操作

ウィンドウ・アイコンの説明は、「[MineSet 3D ビジュアライザでのナビゲーション](#)」(27 ページ)にあります。ビジュアライザ・オプションの詳しい説明については『[MineSet 3.0 Enterprise Edition Reference Guide](#)』のスキャタ・ビジュアライザとスプラット・ビジュアライザの説明を、色の変化の詳細についてはカラーブラウザの説明をそれぞれ参照してください。

スキャタ・ビジュアライザのスライドの作成

値が別の項目の値に従って変わるように思われる場合は、項目をスライドにマッピングできます。項目をスライドにマッピングするためには、項目の型が数値（整数、浮動小数点、倍精度のどれか）か階級でなければなりません。項目の階級がすでに生成されていれば、名前の後に `_bin` が付いています。「現在のデータセットの項目名」リストには、項目の型が `total day calls - double` のように項目名の後に表示されます。ほとんどの場合は、Tool Manager を使って項目をスライドにマッピングするだけで、スライドが自動的に作成されます。

スライドの自動作成

集計処理手順でスライドが暗黙的に指定されていないと、Tool Manager は、階級自動生成と集計処理を通してスライドを作成します。このような自動操作は「ツールの起動」をクリックした後に行われます。固有要素の数は、削除されていない項目や可視化要素にマッピングされていない項目を使って決められます。つまり、これらの項目は自動集計処理での「項目によるグループ化」項目になります。Tool Manager の「ツールオプション」をクリックすると、現在のスライド・インデックスが「ツールオプション」ダイアログ・ボックスに表示されます。

可視化要素にマッピングされている数値型項目はすべて集計処理されます。集計処理の種類は「ツールオプション」パネルで選択できます。異なる要素を別々の方法で集計処理する場合は、スライドを手動で作成する必要があります。

スライドの手動作成

予期どおりの結果を得るには、Tool Manager を使って階級と集計処理を明示的に行う方が確実です。それには、平均、カウント、合計を使って、可視化要素にマップする項目を集計処理し、他の文字列や項目でグループ化した後、アニメーション・スライドにしたい項目で「配列のインデックス付け」を行います。それによって、集計処理されたすべての項目に対する配列項目が、スライド変数でインデックス付けされます。このようにして指定したスライドをスライド要素に直接マッピングすることはできません。集計処理の詳細は、『*MineSet 3.0 Enterprise Edition Reference Guide*』の集計処理の説明を参照してください。

現在のデータテーブルに配列項目があると、スキャタ・ビジュアライザは、その項目のインデックスに従ってスライドを作成します。インデックスが1つの場合は1次元の X スライドを、2つの場合は X スライドと Y スライドをそれぞれ作成します。現在のテーブルのすべての配列項目は同じインデックスを持っていなければなりません。そうしないと、スライドは作成されません。配列項目の作成については、[第 3 章](#)の

「[集計処理による新規項目の作成](#)」の項と『[MineSet 3.0 Enterprise Edition Reference Guide](#)』を参照してください。

スキャタ・ビジュアライザとスプラット・ビジュアライザでのアニメーションの作成

アニメーションの作成には、メイン・ビジュアライザ・ウィンドウの右側にあるアニメーション・コントロール・パネルを使用します。アニメーション・ウィンドウは、スライダがマッピングされている場合だけ表示されます。たとえば、*company.scatterviz* はそのようなデータセットです。『[MineSet 3.0 Enterprise Edition Reference Guide](#)』のアニメーションの説明を参照してください。値が特定の条件に従って変わっていると思われる場合は、その項目をスライダにマッピングできます。これは、たとえば、人口密度が年とともに変化するような場合です。アニメーション・スライダを使用するスプラット・ビジュアライザの例として *adultJobs.splatviz* が付属しています。

スライダや軸には独立属性と思われる属性を選択し、軸、サイズ、色には依存属性を選択するのが最も有効です。*adultJobs.splatviz* や *adult94.scatterviz* のようにスライダが 1 つしかマッピングされていないと、サマリウィンドウは圧縮されます。スライダの次元はその下のラベルに示されます。*brand.scatterviz* のように項目がスライダにまったくマッピングされていないデータセットの場合は、スライダ・コントロールは表示されません。

ビジュアライザ・サマリウィンドウを使ったアニメーション

アニメーションは、スキャタ・ビジュアライザやスプラット・ビジュアライザのアニメーション・コントロール・パネルのサマリウィンドウ部分に作成します。サマリウィンドウには、メインウィンドウに表示されているデータの集計処理に基づくサマリ値の配列が 1 次元か 2 次元で表示されます。サマリスライダ・ウィンドウの黒いドットはそれぞれがサマリ値に対応しています。大きな値にはそのサマリ色が、小さな値には白が割り当てられます。中間の値には、2 色間の色が補間されます。ある位置のサマリ変数がデータをもっていなかったり、NULL 値だけだったりすると、不明であることを示すグレイで表示されます。サマリ値は、そのスライダ位置の左のプロットに表示されているすべてのデータの指定された集計処理を使って計算されます。黒いドットを消すには、アニメーション・パネルの「データポイントの表示」チェックボックスを使用します。

次の例では、付属の *company.scatterviz* ファイルを使ってアニメーションを作成します。

1. Tool Manager のメニューバーから「可視化ツール」->「3D ビジュアライザ」を選択します。
2. メニューから「ファイル」->「開く」を選択します。
3. `company.scatterviz` ファイルを開きます。このファイルは、MineSet を最初にインストールした `MineSet 3.0\examples` ディレクトリにあります。
4. サマリウィンドウで、アニメーション・パスの始点にする黒いドットを選択します。その黒いドット上でマウスの左ボタンをクリックしたままウィンドウ上でカーソルをドラッグします。マウスの左ボタンを離して終点を定義します。
 - あるいは、マウスの左ボタンをクリックして始点を設定し、独立次元のスライダの 1 つをドラッグしてこの次元に沿って直線を引きます。スライダが 2 つある場合は、さらに 2 つ目のスライダを使って、2 つめの軸に沿って直線を引きます。
 - 別の方法では、ドット上で右ボタンをクリックし、別のドット上で中ボタンをクリックして連続したパスセグメントを作成してパスを定義します。このオプションは、マウスボタンが 3 つあるときだけ使用できます。
5. サマリウィンドウの下にあるビデオ機のようなボタンを使ってアニメーションをコントロールします。上の列の再生ボタンをクリックしてから、その下の列のパスボタンをクリックします。コントロールの詳細は、『*MineSet 3.0Enterprise Edition Reference Guide*』のアニメーションの説明を参照してください。

スキャタ・ビジュアライザでのアニメーション軌跡の表示

スキャタ・ビジュアライザでは、移動の軌跡を表示して要素のアニメーション・パスの変化を見ることができます。アニメーションを作成すると、選択された各要素の後ろにその軌跡が、選択された形式で表示されます。スキャタ・ビジュアライザのアニメーション・コントロール・パネル (図 4-8) 右下にあるモーション・オプション・メニューから次のオプションを選択できます。

- 軌跡なし — デフォルト
- ライン軌跡 — 細い色付きのライン
- フェードアウト軌跡 — ライン軌跡に似た透明な色付きライン。最も新しい位置が最も不透明になります。
- チューブ軌跡 — 3 次元の管状の軌跡。その太さは、要素がアニメーション・パスを移動するときのサイズの変化に従って変わります。管状の軌跡が多すぎると、アニメーションが著しく遅くなることがあります。

すべての軌跡は起点の要素に従って色付けされます。サマリウィンドウの下のスライダが動くに従って、要素の色がたとえば赤から青に変わると、対応する軌跡の色も2地点間で次第に変わります。マッピングされていない属性がある2地点間のパスに沿って一定であると、この2地点間に軌跡が引かれます。たとえば、収入が年齢とともにどのように変化するかを表示し、その後に軌跡を描いてある地点から別の地点までの変化の割合を示すことができます。

モーション軌跡を表示する対象としては、通常、少ない項目数でグループ化した集計データの方が優れています。当初、軌跡は、パスによって影響を受ける、スキヤタのすべてのポイントに対して表示されます。要素をクリックして選択すると、選択されたポイントに対してのみ軌跡が表示されます。このようにすると、表示の混雑が緩和されます。NULL位置をもつ要素は軌跡上では中断として表示されます。

図 4-8 に、スキヤタ・ビジュアルライザのチューブ・モーション軌跡の例を示します。

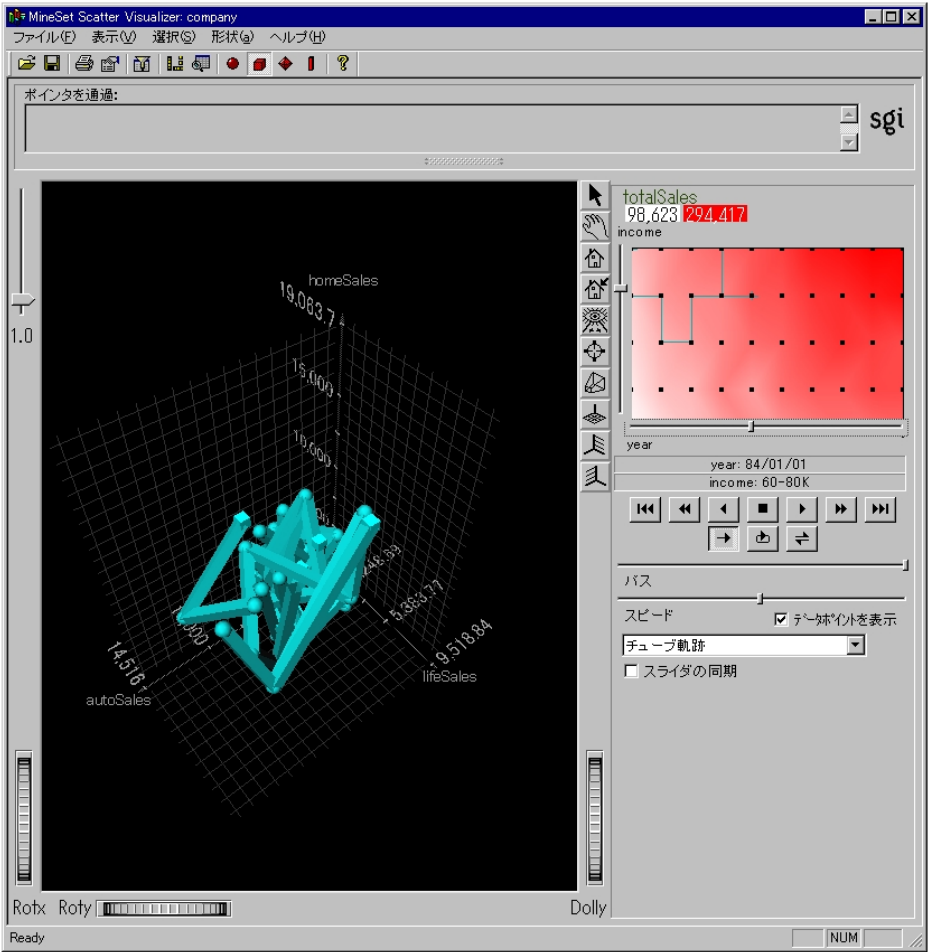


図 4-8 スキャタ・ビジュアライザのチューブ・モーション軌跡の例

スプラット・ビジュアライザでのアニメーション

スプラット・ビジュアライザでのアニメーションは、チューブ軌跡オプションがないことを除けば、本質的にはスキャタ・ビジュアライザと同じです。図 4-9 に *adultJobs.splatviz* の例を示します。

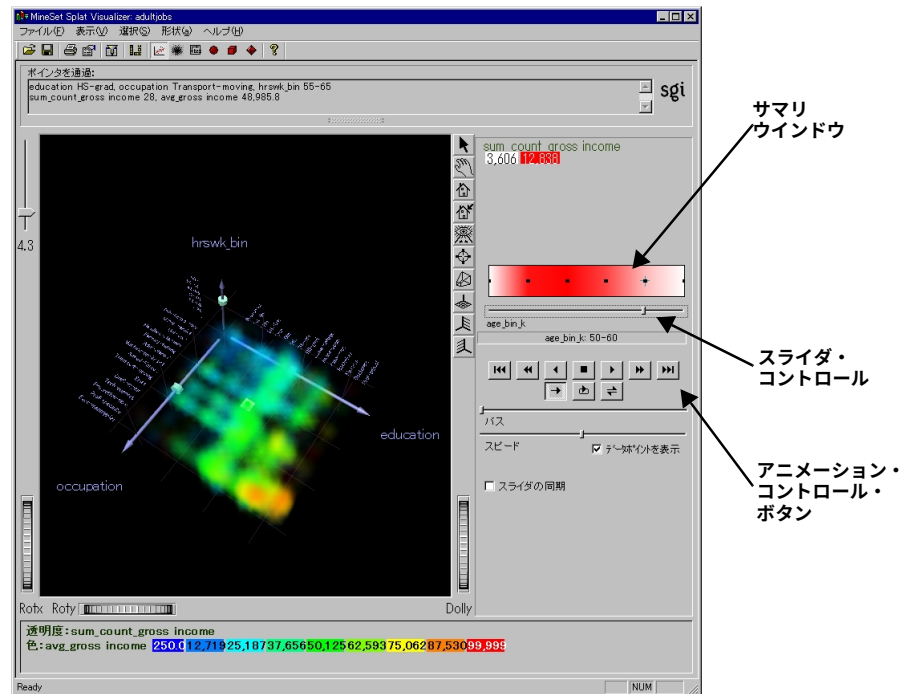


図 4-9 スプラット・ビジュアライザのサマリウィンドウとスライダ・コントロール

スライダ・マッピングが設定ファイルに指定されていないと、スライダ・コントロールは表示されません。

スプラット・ビジュアライザでのサマリウィンドウの解釈

サマリウィンドウ (図 4-9) では、サマリ属性の集計処理がスライダ次元に表示されます。色の密度は、メインウィンドウのスプラットによって表される高いサマリ値を示します。サマリに項目が明示的にマッピングされていない場合は、スライダのどの位置が最も多いデータを表すかがカウントで示されます。

図 4-9 の例では、サマリスライダの純白領域にある黒いドットは、スライダがその位置にあるとき、左のスプラットは 3,606 レコードから構成されていることを示します。スライダの最も赤いところでは、スプラットは 12,838 レコードから構成されています。

等間隔にあいた黒いドットは、離散データポイントの正確な位置を表します。黒いドットを消去するには、アニメーション・パネル下部の「データポイントの表示」ボックスをオフにします。黒いドット間のスライダ位置では、基になっているデータの補間を使ってイメージが生成されます。

サマリウィンドウでの色の密度

adultJobs.splatviz ファイルを使うと、値が色の密度にどのように反映されるかを見ることができます。サマリウィンドウには、白 (左側) から赤 (中)、そして白 (右側) へという色範囲が表示されます。赤はレコード数が多いことを、白はレコード数が少ないことを表します。この例の場合、スライダ中央の最も濃い赤は、20 ~ 50 の年齢層に人が最も集中していることを意味します。

サマリウィンドウでのパスの作成

パスの作成は、スキャタ・ビジュアライザとまったく同じです。コントロールも同じです。ただし、アニメーションが進行するに従い、スプラットのサイズと色がスムーズに変わります。ステータスフィールドには、補間されたデータ値 (interpolated data) が表示されます。スライダの動きを停止しても、スライダ位置は、補間されたデータ値 (interpolated data) が使用されていない最も近くの離散データの位置まで移動します。

サマリスライダにおいては、それぞれの階級生成された値を持つテーブルがあります。元データの集計処理であるこのようなテーブルの各行は、シーン内の 1 つのスプラットを定義しています。隣り合わせの階級に対応するテーブルの行数は、位置によってデータの分布が異なるため、同じでなくてもかまいません。たとえば、図 4-2 (58 ページ) で、スライダを 1 ノッチ右に移動すると (図 4-10 を参照)、可視化表示が 40 ~ 50 の年齢層から 50 ~ 60 の年齢層に変わるとします。この場合、位置によっては、前に表示されていなかったところにスプラットが表示されたり、その反対になったりすることがあります。

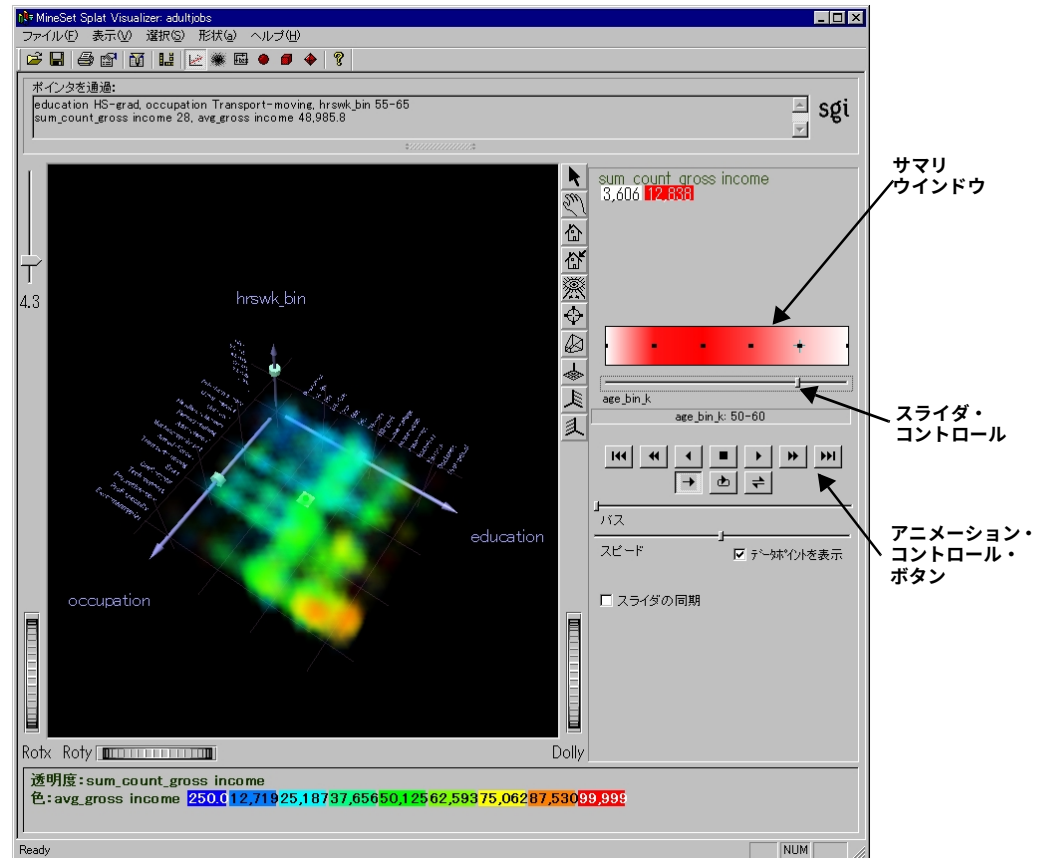


図 4-10 スライダーの移動による可視化の変化 (図 4-2 と比較)

1次元スライダーの補間では、隣接する2つのテーブルがマージされ、この空間的な項目を固有キーとして集計処理されます。重み付けはそのまま補間されます。どちらかのテーブルに特定の行がない場合は、0の重み付けが使用されます。色に使用する平均値も補間されますが、これは重みによって重み付けされます。

補間の詳しい技術的な例については、『*MineSet 3.0 Enterprise Edition Reference Guide*』のスプラット・ビジュアライザの説明を参照してください。

スキャタ・ビジュアライザとスプラット・ビジュアライザでの結果の操作

スキャタ・ビジュアライザやスプラット・ビジュアライザのそのほかの機能にアクセスするプルダウン・メニューが 5 つあります。「ファイル」、「表示」、「選択」、「形状 (Shape)」、「ヘルプ」です。設定ファイルを指定せずにビジュアライザを起動すると、「ファイル」と「ヘルプ」メニューしか使用できません。しかし、ファイルを開くと、他のメニューも使用できるようになります。「ファイル」と「ヘルプ」メニューは標準的なものです。詳細は、『*MineSet 3.0 Enterprise Edition Reference Guide*』を参照してください。

表示の変更

「表示」メニューでは、ビジュアライザ・ウィンドウに表示されている要素のある側面を制御します。[図 4-11](#) にスキャタ・ビジュアライザとスプラット・ビジュアライザの「表示」メニューを示します。

フィルタパネル(F)...	Ctrl+F
背景色の設定(C)...	Ctrl+K
✓ 拡張コントローラーの表示(D)	
✓ アニメーションパネルの表示(A)	
✓ Null 位置の表示(N)	
✓ ツールバー(T)	
✓ ステータスバー(S)	

図 4-11 スキャタ・ビジュアライザとスプラット・ビジュアライザの表示メニュー

次の表で各選択肢を説明します。

表 4-5 スキャタ・ビジュアライザとスプラット・ビジュアライザの表示メニュー選択肢

オプション	説明
フィルタパネル	表示の内容を操作する「フィルタパネル」(図 4-12)を開きます。
背景色の設定	新しい表示背景色を指定する「色の選択」を開きます。
拡張コントローラ	メインウィンドウ周囲の外部コントロールを表示または消去します。

表 4-5 (続き) スキャタ・ビジュアライザとスプラット・ビジュアライザの表示メニュー選択肢

オプション	説明
アニメーション・パネルの表示	アニメーション・コントロール・パネルを表示または消去します。
NULL 位置の表示	軸に沿って NULL 位置や不明な位置をもつ要素を表示または消去します。



図 4-12 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「フィルタダイアログ」

表示されているデータの量を一定条件に従ってフィルタリングする場合は、[図 4-12](#) の「フィルタ ダイアログ」を使用します。それには、「表示」メニューから「フィルタパネル」を選択します。上部のパネルでは、文字列型項目に基づいてデータをフィルタリングします。値を選択するには、それをクリックし、選択解除するには、それを再びクリックします。項目のすべての値を選択する場合は、「すべてに設定」ボタンをクリックします。現在の選択をすべて取消す場合は、「消去」ボタンをクリックします。

下部のパネルでは、文字列型項目と数値型項目両方の値に基づいてデータをフィルタリングします。詳細は、『*MineSet 3.0 Enterprise Edition Reference Guide*』のフィルタパネルの説明を参照してください。

右中央の「フィルタのスケール」チェックボックスでは、メインウィンドウのランドスケープがデータセット全体の領域を表すのか、フィルタリングされたデータの領域だけを表すのかを指定します。

フィルタ処理を始めるには、「適用」ボタンをクリックします。パネルがアクティブのときに <Enter> キーを押すと、フィルタ処理が自動的に始まります。

スキャタ・ビジュアライザの選択とドリルスルー

スキャタ・ビジュアライザの「選択」メニューでは、基になっているデータにドリルスルーできます。

選択ボックスの作成(B)	Ctrl+B
値の表示(V)	
オリジナルデータの表示(O)	Ctrl+D
ToolManagerに送信(T)	Ctrl+T
補集合データのドリルスルー(Q)	
項目選択してドリルスルー(Q)...	

図 4-13 スキャタ・ビジュアライザの「選択」メニュー

- 「選択ボックスの作成」では、3次元の選択ボックスを作成します。ボリュームの領域を選択するには、このボックスを伸ばしたり、移動したりします。これがアクティブの間は、レコードビューワのテーブルが開かれ、その中の要素によって表される集計データの情報が表示されます。カーソルは選択モードでなければなりません。選択ボックスを移動するには、いずれかの面の上でマウスの左ボタンをクリックし、必要な方向にドラッグします。<Shift> キーを押したままドラッグすると、移動はドラッグ動作が最も近い軸に限定されます。選択ボックスの範囲を変更するには、グレイスケール・タブの1つを必要な方向にドラッグします。ボリュームの境界を越えて移動したり、サイズを変えることはできません。グレイスケール・タブは、一定のスクリーンサイズを維持するために絶えずサイズが変わります。グレイスケール・タブが大きすぎる場合は、ズームインすれば、ボックスと比較したタブの相対的なサイズが小さくなります。
- 「値の表示」では、テーブル内の選択された要素の値をレコードビューワで表示します。

- ・「オリジナルデータの表示」では、選択されているものに対応するレコードを検索し、表示します。レコードは、レコードビューワによってテーブル形式で表示されます。何も選択されていないときは、このメニュー項目はグレイです。
- ・「Tool Manager に送信」では、現在の選択に基づいて、フィルタ操作を Tool Manager 履歴の最初に挿入します。ドリルスルーに使用される実際の表現は、現在の選択によって決まります。何も選択されていないときは、このメニュー項目はグレイです。
- ・「補集合データのドリルスルー」を指定すると、「オリジナルデータの表示」や「Tool Manager に送信」を選択したときに、選択されていないすべてのデータが取り出されます。
- ・「項目選択してドリルスルー」では、ドリルスルーで使用する項目を選択するためのパネルを表示します。他の可視化ツールとは異なり、データには、データのキーとして特に指定された項目はありません。スキャタ・ビジュアライザは、ユーザがドリルスルー表現にどの項目を必要としているかを判断することはできません。たとえば、車種、型式、重量をもつ自動車データがあるとしたします。オリジナルデータまでドリルスルーするときに、車種と型式は対象とするが、重量は対象としないように指定したい場合があります。デフォルトでは、グラフィカル要件にマッピングされたすべての項目がドリルスルーの対象とみなされます。他の項目は対象とされませんが、「設定」ダイアログ・ボックスでそれらの項目を強調表示すれば、対象になります。

詳細は、『*MineSet 3.0 Enterprise Edition Reference Guide*』のドリルスルーの説明を参照してください。

スプラット・ビジュアライザの選択とドリルスルー

スプラット・ビジュアライザの「選択」メニューは、元になっているデータにドリルスルーするオプションの点でスキャタ・ビジュアライザとはわずかに異なります。

選択ボックスの作成(B)	Ctrl+B
オリジナルデータの表示(D)	Ctrl+D
ToolManagerに送信(T)	Ctrl+T
補集合データのドリルスルー(C)	
▼ドリルスルーにスライダを使用(S) ピクドラッグの表示	

図 4-14 スプラット・ビジュアライザの「選択」メニュー

- 「選択ボックスの作成」では、3 次元の選択ボックスを作成します。ボリュームの領域を選択するには、このボックスを伸ばしたり、移動したりします。これがアクティブの間は、レコードビューワのテーブルが開かれ、その中の要素によって表されたすべての集計データの情報が表示されます。このウィンドウを閉じると、現在のすべての選択が消去されます。選択ボックス内の要素や <Shift> キーのクリックで選択された要素は、テーブル・ウィンドウに表示されます。選択ボックスを移動するには、いずれかの面の上でマウスの左ボタンをクリックし、必要な方向にドラッグします。<Shift> キーを押したままドラッグすると、移動はドラッグ動作が最も近い軸に限定されます。選択ボックスの範囲を変更するには、グレイスケール・タブの 1 つを必要な方向にドラッグします。ボリュームの境界を越えて移動したり、サイズを変えることはできません。グレイスケール・タブは、一定のスクリーンサイズを維持するために絶えずサイズが変わります。グレイスケール・タブが大きすぎる場合は、ズームインすれば、ボックスと比較したタブの相対的なサイズが小さくなります。
- 「オリジナルデータの表示」では、選択ボックスで選択されたものに対応するレコードを検索し、表示します。レコードは、レコードビューワによってテーブル形式で表示されます。
- 「Tool Manager に送信」では、現在のボックス選択に基づいて、フィルタ操作を Tool Manager 履歴の最初に挿入します。ドリルスルーに使用される実際の表現は、現在のボックス選択の範囲によって決まります。何も選択されていないと、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」を指定すると、「オリジナルデータの表示」や「Tool Manager に送信」を選択したときに、選択されていないすべてのデータが取り出されます。
- 「ドリルスルーにスライダを使用」では、ドリルスルーを行うときにスライダ位置を使用します。
- 「ピックドラッグの表示」では、ピックドラッグを表示するかどうかを切り替えます。デフォルトはオンです。ボックス選択を起動すると、ピックドラッグは除去されます。ただし、ピックドラッグとボックス選択を同時にアクティブにすることもできます。

詳細は、『*MineSet 3.0 Enterprise Edition Reference Guide*』のドリルスルーの説明を参照してください。

「形状」メニューによるスキャタ・ビジュアライザ表示の変更

スキャタ・ビジュアライザの「形状」メニューでは、スキャタ・ポイントの描画方法を変更します。スキャタ・ポイントは、次の不透明プリミティブで表すことができます。

- 「立方体」では不透明な立方体が表示されます。その大きさは、可視化要素のサイズにマッピングされた属性に比例します。
- 「ダイヤモンド」ではワイヤフレーム三角形が表示されます。方向はその色によって変わります。サイズは、可視化要素のサイズにマッピングされた属性の値に比例します。
- 「球体」では不透明な球体が表示されます。その大きさは、可視化要素のサイズにマッピングされた属性に比例します。
- 「バー」では長細い棒が表示されます。その長さは、可視化要素のサイズにマッピングされた属性の値で決まります。

「形状」メニューによるスプラット・ビジュアライザ表示の変更

スプラット・ビジュアライザの「形状」メニューでは、スプラットの描画方法を変更します。このメニューでは対話性の精度を変更できます。テクスチャスプラットは、どのような方法でも近似化される理想的なガウス密度分布を最も正確に表現します。ほとんどのコンピュータにはテクスチャのハードウェア・アシスト機構がありますので、一般にはテクスチャスプラットが最も優れています。スプラットの描画方法には次の3つがあります。

- 「線形」では、少数の三角形によってガウススプラットの線形近似値が描画されます。
- 「ガウシアン」では、多数の三角形によってガウススプラットの近似値が描画されます。
- 「テクスチャ」では、テクスチャマッピングされた四角形によって最も正確な表示が行われます。テクスチャ・マッピングのハードウェア・アシスト機構がないシステムでは、表示が非常に遅くなることがあります。

あるいは、次の不透明プリミティブを使ってスプラットを表示することもできます。

- 「立方体」では不透明な立方体が表示されます。その大きさは、可視化要素の不透明度（ある場合）またはカウントにマッピングされた属性に比例します。
- 「ダイヤモンド」ではワイヤフレームの三角形が表示されます。方向はその色によって変わります。サイズは、可視化要素の不透明度（ある場合）またはカウントにマッピングされた属性に比例します。
- 「球体」では不透明な球体が表示されます。その大きさは、可視化要素の不透明度（ある場合）またはカウントにマッピングされた属性に比例します。

ツリー・ビジュアライザによるデータの表示

この章は次の節から構成されています。

- 「ツリー・ビジュアライザ (Tree Visualizer) の概要」(89 ページ)
- 「ツリー・ビジュアライザの起動」(90 ページ)
- 「ツリー・ビジュアライザによる結果の表示」(94 ページ)
- 「ツリー・ビジュアライザ表示の微調整」(107 ページ)

ツリー・ビジュアライザの特徴と機能を示すために設定ファイルとデータファイルのサンプルが用意されています。このファイルは、MineSet を最初にインストールした *MineSet 3.0\examples* ディレクトリにあります。

ツリー・ビジュアライザ (Tree Visualizer) の概要

ツリー・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。ツリー・ビジュアライザでは、データ階層がツリーの形で表示されます。ツリーの各レベルは、個別の属性の値に基づいて分岐します。ツリーの各ノードには、その下のサブツリーに属するすべてのデータを含むグラフが表示されます。このグラフはデータ属性に従って高さや色が変わるベースからなり、その上にバーやディスクが表示されます。バーやディスクの数、ラベル、高さ、色も指定されている属性によって変わります。

図 5-1 に示すように、ツリー・ビジュアライザは、階層的につないだノードとしてデータを表示することによって、その定量的な特性や関係を明らかにします。ノードのバーやディスクの高さと色は、データ値の集計処理に対応しています。通常、集計処理には合計や平均、カウントを使用します。ノードとノードをつなぐエッジと呼ぶラインはあるデータとそのサブセットの関係を示しています。

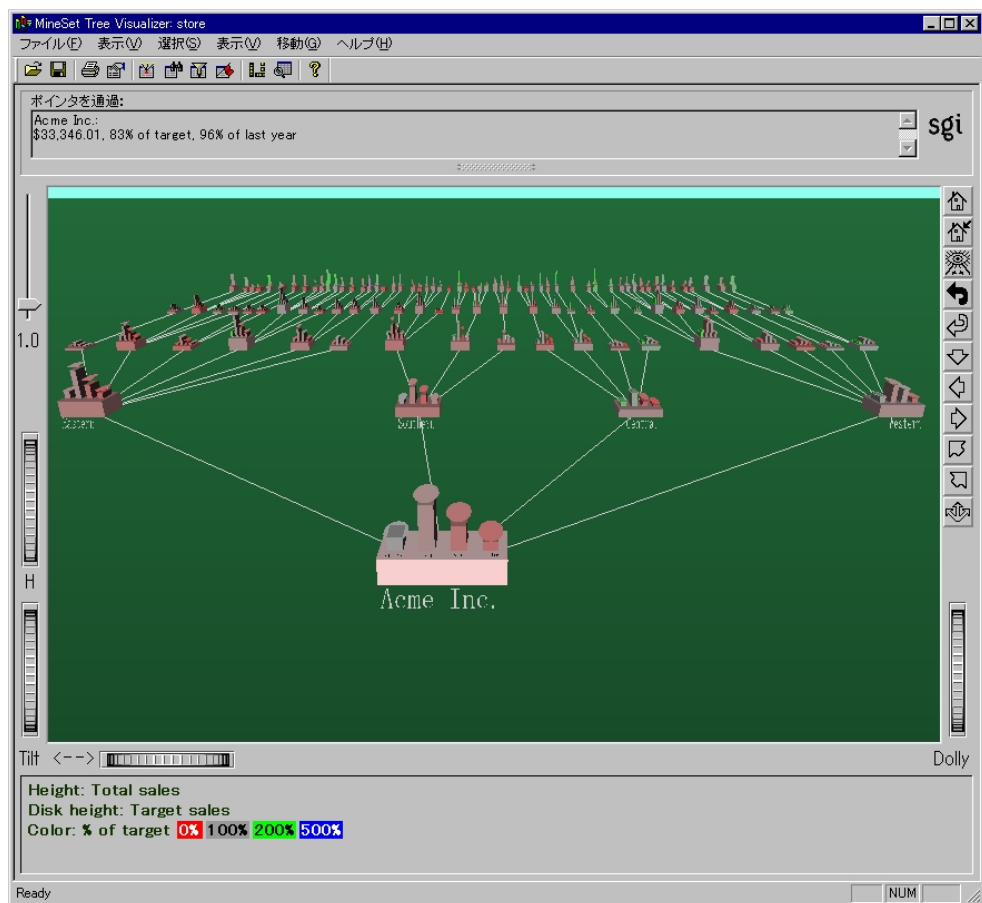


図 5-1 ツリー・ビジュアライザのメインウィンドウでの表示例

ツリー・ビジュアライザの起動

この節では、Tool Manager を使ってツリー・ビジュアライザを設定するにはどうすればよいかを説明します。Tool Manager を使うとツリー・ビジュアライザの設定作業は非常に簡単になりますが、エディタを使って設定ファイルを手作業で作成することもできます。その他にも、ツリー・ビジュアライザを起動する方法はいくつかあります。詳細は、『*MineSet 3.0 Enterprise Edition Reference Guide*』を参照してください。

ツリー・ビジュアライザを Tool Manager から起動するには、次の手順に従います。

1. 「ファイル」->「サーバに接続」を選択し、サーバにログインします。
2. 「ファイル」->「新しいデータファイルを開く」を選択し、必要なファイル名を選択するか入力します。
3. 「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 (図 5-2)。
4. 一連のツールタブから「ツリー・ビジュアライザ」タブを選択します。
5. 可視化要素の横にあるプルダウン・メニューを使って、データセットの属性を「データの可視化／マイニング」パネルの適切な可視化要素にマッピングします。マッピングの詳細は、「ツリー・ビジュアライザでの可視化要素のマッピング」(92 ページ) を参照してください。
6. 「ツールオプション」ボタンをクリックし、ツリー・ビジュアライザの「設定オプション」パネルで「高さの集計処理」と「集計処理の色」を指定します。この手順によって、意味のある可視化かどうかが決まることがあります。
デフォルトでは、親ノードのバーの高さまたは色は、子ノードのすべてのバーの値を合計したものです。データセットによっては、合計よりも平均や最大、最小、カウントの方がよい場合もあります。集計処理は、バー、ベース、ディスクの高さや色に割り当てることができます。集計処理やその他の高度なオプションの詳細は、『*MineSet 3.0 Enterprise Edition Reference Guide*』を参照してください。
7. 「ツールの起動」をクリックしてビジュアライザを起動します。

「Tool Manager」ウィンドウ下部の「ステータス」ボックスに進行状況と結果の統計が表示されます。

ツリー・ビジュアライザでの可視化要素のマッピング

ツリー・ビジュアライザを選択すると「データの可視化／マイニング」パネルに可視化要素が表示されますので、それらの要素にデータの属性をマッピングします (図 5-2)。それには、可視化要素の横にあるプルダウン・メニューから項目を選択します。表 5-1 に、可視化要素とその使い方を示します。ツリー・ビジュアライザを使用する場合には、要素のマッピングをいろいろと試してみて、属性と可視化要素の最善の組み合わせを見つける方がよいことがしばしばあります。

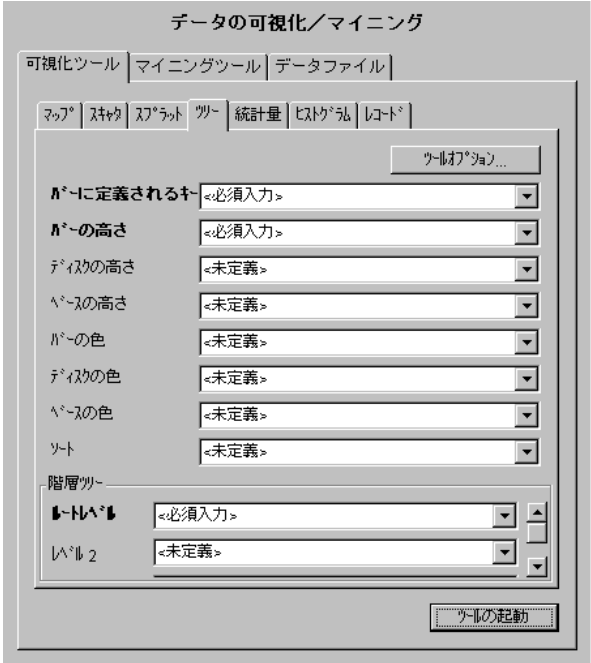


図 5-2 ツリー・ビジュアライザが選択された Tool Manager の「データの可視化／マイニング」パネル

表 5-1 に可視化要素を示します。

表 5-1 ツリー・ビジュアライザの可視化要素

可視化要素	コメント
バーに定義されるキー	バーで何を表すかを指定します。項目が数値の場合は、通常、項目の階級を生成します。階級を生成しないと MineSet が属性の値ごとにバーを作成しますので、可視化が使いにくいものになるおそれがあります。
バーの高さ	バーの高さで何を表すかを指定します。一般には、バーが高ければ高いほど、表されているデータも大きいことを意味します。各レベルで属性を集計処理するオプションについては、「ツールオプション」パネルを参照してください。
ディスクの高さ	ディスクの高さで何を表すかを指定します。省略可能なディスクがバーと同じ場所に表示されます。ディスク属性とバーの高さ属性は測定単位を同じにした方がディスクが意味のあるものになります。たとえば、今年の売上げと去年の売上げのような場合です。このマッピングを指定しないと、ディスクは表示されません。
ベースの高さ	ベースの高さで何を表すかを指定します。このマッピングを指定しないと、バーの高さが使用されます。各レベルで属性を集計処理するオプションについては、「ツールオプション」パネルを参照してください。
バーの色	バーの色で何を表すかを指定します。「ツールオプション」パネルを使って特定の色を割当てすることもできますし、MineSet に色の割当てを任せることもできます。『 <i>MineSet 3.0 Enterprise Edition Reference Guide</i> 』の色の選択に関する説明を参照してください。各レベルで属性を集計処理するオプションについては、「ツールオプション」パネルを参照してください。
ディスクの色	ディスクの色で何を表すかを指定します。「ツールオプション」パネルを使って特定の色を割当てすることもできますし、MineSet に色の割当てを任せることもできます。『 <i>MineSet 3.0 Enterprise Edition Reference Guide</i> 』の色の選択に関する説明を参照してください。このオプションは、ディスクの高さが指定されたときだけ有効です。

表 5-1 (続き) ツリー・ビジュアライザの可視化要素

可視化要素	コメント
ベースの色	ベースの色で何を表すかを指定します。「ツールオプション」パネルを使って特定の色を割当てすることもできますし、MineSet に色の割当てを任せることもできます。『 <i>MineSet 3.0 Enterprise Edition Reference Guide</i> 』の色の選択に関する説明を参照してください。各レベルで属性を集計処理するオプションについては、「ツールオプション」パネルを参照してください。このマッピングを指定しないと、バーの色が使用されます。
ソート	ノードのレイアウトを、選択された属性の値でソートします。デフォルトで、ソート順は左から右への昇順になります。
階層ツリーのルートレベル	データソースからのテーブルを階層にどのように変換するかを指定します。ツリーの第 2 レベルには、選択した属性の値と同じ数のノードが作られます。 属性の値数が多くなるようだと、ツリーが大きくなり、分析が難しくなります。これを避けるには、項目の階級生成が有効な場合があります。
階層ツリーのレベル 2,3,...	データをさらに多くの階層に分割します。デフォルトで、可視化要素のリストは 3 つの階層レベルになります。3 つめの階層レベルを指定すると、ビジュアライザが 4 つめを自動的に追加します。追加のレベルを指定するたびに、ビジュアライザがレベルをもう 1 つ追加します。階層レベルはいくつでも指定できます。 属性の値数が多くなるようだと、ツリーが大きくなり、分析が難しくなります。これを避けるには、項目の階級生成が有効な場合があります。

ツリー・ビジュアライザによる結果の表示

ツリー・ビジュアライザを起動すると、シーンの前方、つまり、メインウィンドウの下の方に階層のルートノードが、ルートノードの後方に下位ノードが表示されます。ノードは、ベースとその上のバーからなります。バーの高さや色で何を表すかは、Tool Manager を使用するか、.treeviz 設定ファイルを変更すれば変更できます。『*MineSet 3.0 Enterprise Edition Interface Guide*』のツリー・ビジュアライザの説明を参照してください。通常、ベースはすべてのバーの集計処理を表します。ベースとベースは、ノードとその下位ノードの接続を表すエッジで接続されます。

サブグループの値は、自動的に合計され、次の上のレベルに表示されます。ベースは、その上にあるすべてのバーの集計処理値の情報を持っています。負の値を表すバーは、ベースの上部から下方向に表示されます。ベースの高さを無効にすると、負の値を表すバーの表示が明瞭になります。「[ツリー・ビジュアライザ表示の微調整](#)」(107 ページ) か、『*MineSet 3.0 Enterprise Edition Reference Guide*』のツリー・ビジュアライザの説明を参照してください。

ツリー・ビジュアライザの例

[図 5-3](#) は、stores.treeviz ファイルを最初に開いたときの画面です。このファイルは、MineSet を最初にインストールした *MineSet 3.0\examples* ディレクトリにあります。このファイルを Tool Manager の「可視化ツール (Visual Tools)」メニューを使って表示するには、この 3D ビジュアライザを開き、ツリー・ビジュアライザの「ファイル」->「開く」を選択します。データファイルのレコードの項目は、地域（東部、南部、中部、西部）、州、市、店舗番号、商品の種類（什器、衣料、電気、家具）、今年の売上げ、去年の売上げ、売上げ目標です。さらに、新しい項目が 2 つ追加されています。売上げ目標に対する今年の売上げの割合と去年の売上げに対する今年の売上げの割合です。[表 5-2](#) に、これらの項目と、ツリー・ビジュアライザの可視化要素とのマッピングを示します。

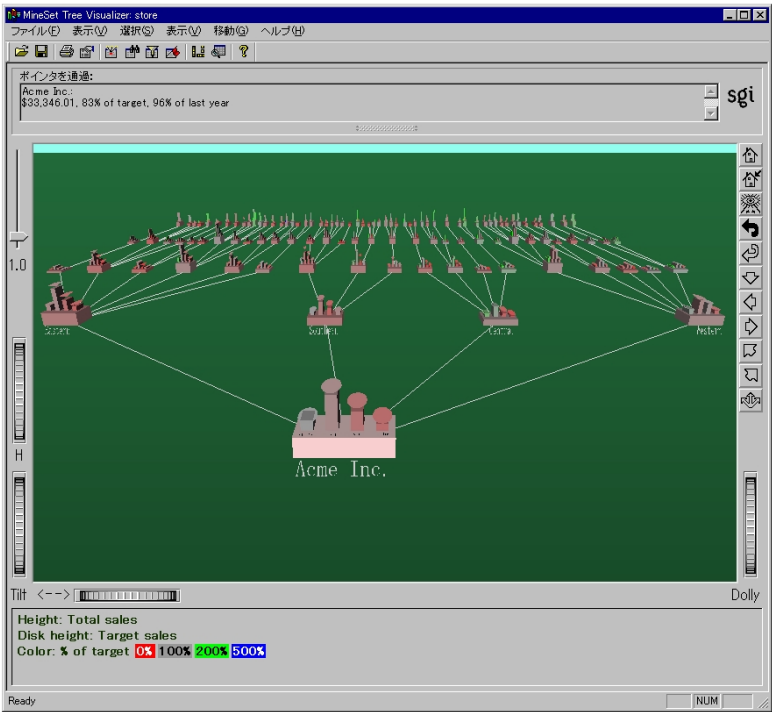


図 5-3 ツリー・ビジュアライザによる店舗データセットの初期画面

このツリーに表示されている最上位ノードのバーは、各製品の売上げ合計を表しています。ディスクは売り上目標げです。次のレベルでは売上げが地域で分割され、その次のレベルでは州で分割されています。以下、個々の店舗に至るまで分割されています。

ツリー・ビジュアライザのシーンを見ると、それぞれの地域、州、市、店舗が売上げ目標を達成しているかどうか商品ごとに一目で分かります。さらに、地域、州、製品の種類などのレベルで売上を比較することもできます。

表 5-2 に、この例における項目と可視化要素のマッピングを示します。.

表 5-2 店舗データのマッピング例

項目	集計処理	マッピング
地域	なし	階層レベル 1
州	なし	階層レベル 2
市	なし	階層レベル 3
店舗番号	なし	階層レベル 4
商品の種類	なし	バーのキー
今年の売上げ	合計	バーの高さ
去年の売上げ	合計	なし
売上げ目標	合計	ディスクの高さ
売上げ目標に対する達成率	なし	バーの色
去年の売上げに対する比率	なし	なし

意味のある可視化となるかどうかはマッピングが左右します。ツールに精通するまでは、さまざまに試行してみてください。

ツリー・ビジュアライザでデータの詳細を見るには

各ノードの詳細を表示するには、次のようにします。

- ノードやオブジェクト（バーやベース）を強調表示して、元になっているデータを表示します。「[オブジェクトやノードの強調表示](#)」を参照してください。
- ノードやオブジェクトを選択して、その場所に視点を合わせズームし、元になっているデータを表示します。「[オブジェクトの選択](#)」を参照してください。同時に複数の場所を選択できます。さらに、ノードやオブジェクトを選択すると、スポットライトが当てられます。スポットライトが当たっていれば、その場所がシーンの遠方に移動しても追跡できます。

オブジェクトやノードの強調表示

オブジェクト（バーやベース）を強調表示するには、オブジェクトの上にマウスを移動します。オブジェクトの情報が表示領域の左上にある「ポインタを通過」に表示されます（[図 5-4](#)）。ノードを強調表示し、ノードの情報を得るには、そのノードに至るラインの上にポインタを置きます。

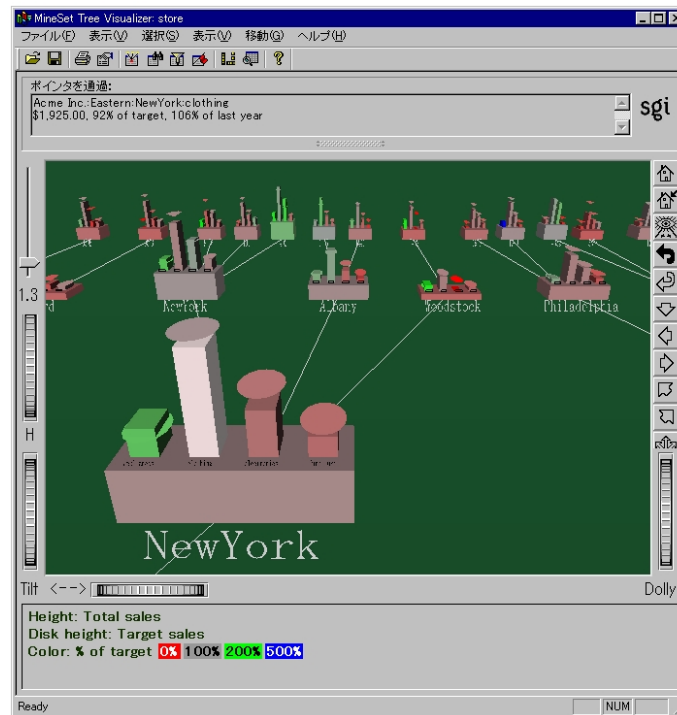


図 5-4 オブジェクトとその元になっている情報の強調表示

オブジェクトの選択

オブジェクトを選択してズームするには、オブジェクトをクリックします。<Shift>キーを押したままオブジェクトをクリックすると、オブジェクトはズームされません。ただし、オブジェクトが選択されていれば、その情報が表示されます。

<Ctrl> キーを押したままオブジェクトをクリックすると、オブジェクトの選択が切り替わります。つまり、オブジェクトが選択されていなければ、クリックによってオブジェクトが選択され、選択されていれば、クリックによって選択が解除されます。この方法を使うと、複数のオブジェクトを同時に選択できます。「ポインタを通過」フィールドの情報は最後に選択されたオブジェクトの情報ですが、すべての選択の値を表示することもできます。それには、それらの選択のオリジナルデータまでドリルスルーするか、「選択」->「値の表示」を選択します。

オブジェクトを選択すると、オブジェクトに白いスポットライトが当てられます (図 5-5)。検索では、スポットライトが黄色になります。「[検索パネルによる特定オブジェクトの検索](#)」(102 ページ) を参照してください。スポットライトは、選択されたオブジェクトが背景の奥にある下位ノードであっても見えます。

スポットライトのエッジはオブジェクトの代わりです。ポインタをスポットライトのエッジの上に移動すると、対応するオブジェクトが強調表示され、オブジェクトの情報が表示域の左上に表示されます。スポットライトのエッジをクリックすると、対応するオブジェクトが選択され、<Shift> キーが押されていない場合はズームされます。スポットライトは、エッジに沿った実線の上だけがアクティブであり、中央の半透明の部分はアクティブではありません。半透明の部分をクリックすると、スポットライトに隠れている部分が選択されます。

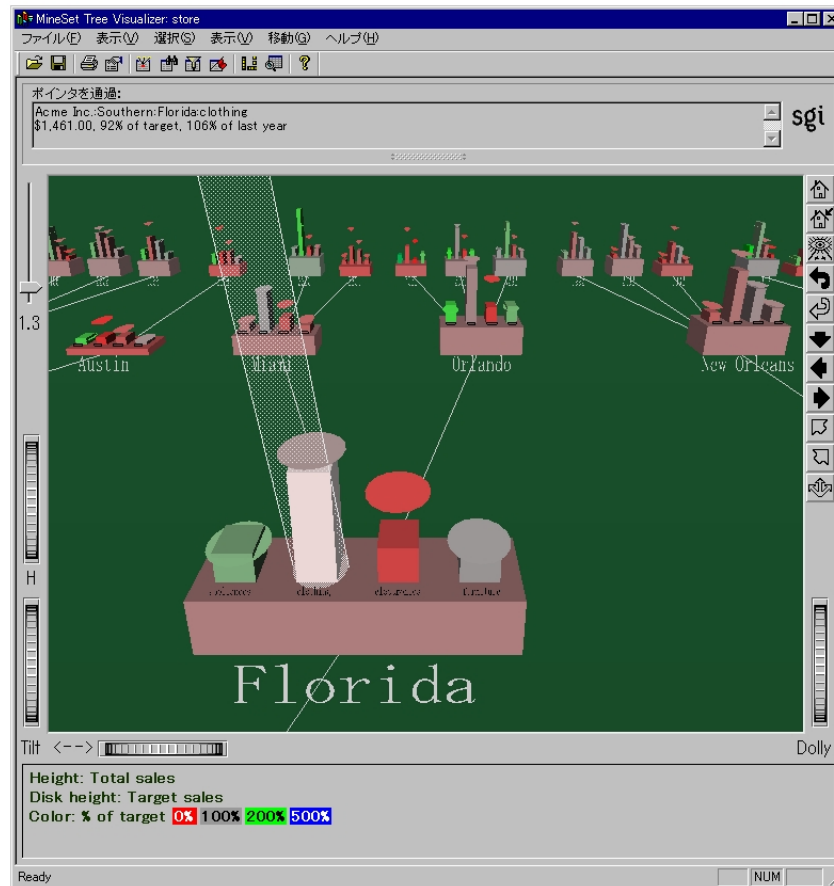


図 5-5 選択された（スポットライトが当てられた）オブジェクトの例

「概観」による階層全体の表示

階層全体を上部から見たウィンドウ（図 5-6）を開くには、「表示」->「概観」を選択します。シーンを表示する度に「概観」を使用する場合は、設定ファイルの「概観」オプションを設定できます。『*MineSet 3.0 Enterprise Edition Interface Guide*』の概観 (Overview) の説明を参照してください。

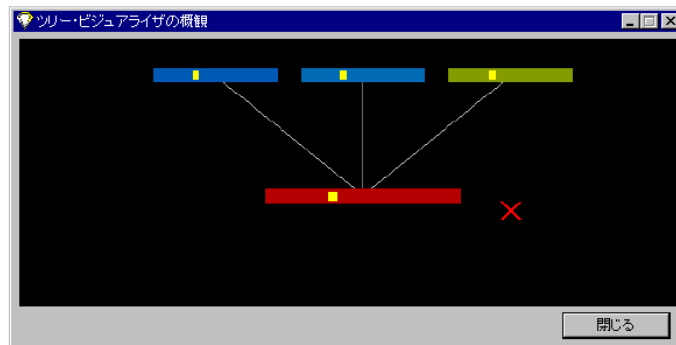


図 5-6 ツリー・ビジュアライザの「概観」ウィンドウ

「概観」ウィンドウの「X」は現在の場所を示しています。「概観」では、場所や視点をシーン全体に渡って追跡したり、特定の場所に素早く移動できます。

- 「概観」でノードを選択し、メイン表示をそこにズームし、スポットライトを当てるには、ノードをクリックします。これはメイン表示でノードをクリックするのと同じです。
- 新しい場所に視点を合わせズームするには、その場所にノードがない場合でも、その場所でマウスの左右ボタンまたは中ボタンをクリックします。

検索パネルによる特定オブジェクトの検索

ツリー・ビジュアライザ・ウィンドウで特定オブジェクトを検索するには、「表示」->「検索パネル」を選択します。次に、検索条件を入力し、「検索」をクリックします。「検索」パネルについては、『*MineSet 3.0 Enterprise Edition Reference Guide*』を参照してください。

検索が終了すると、検索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照らされます(図 5-7)。黄色いスポットライトに照らされたオブジェクトの情報を表示するには、ポインタをスポットライトの上に移動します。情報が左上の「ポインタを通過」に表示されます。黄色いスポットライトに照らされたオブジェクトを選択し、それにズームするには、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>キーを押さなければ、ズームは行われません。

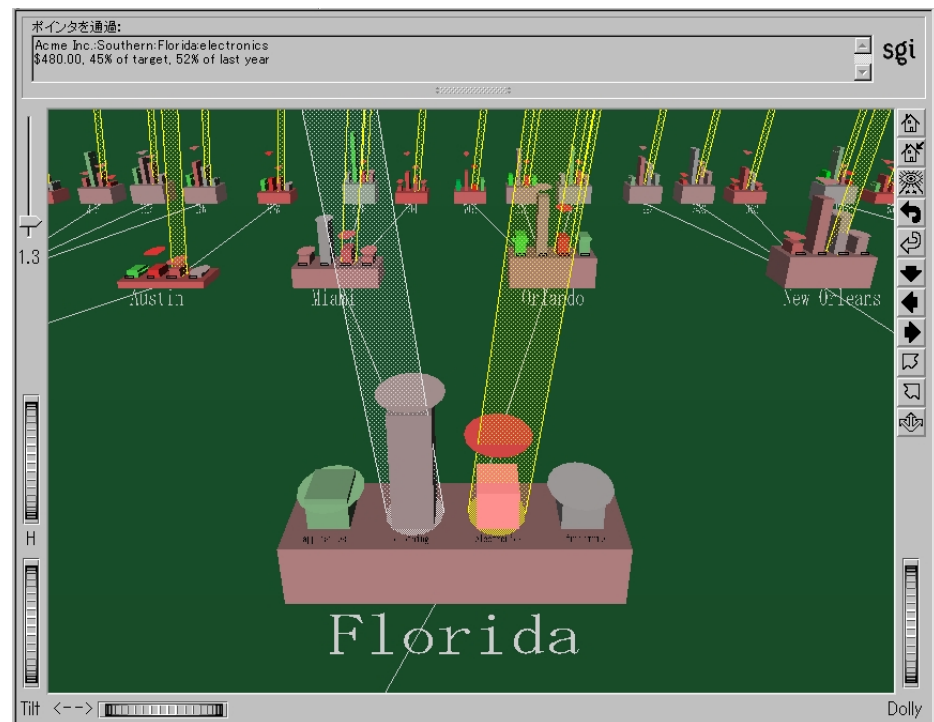


図 5-7 ツリー・ビジュアライザでの検索結果の例

「マーク」パネルによる重要な場所のフラグ付け

「マーク」パネルでは、重要な場所または視点に名前を付け、後で参照するためにそれを格納できます。メイン表示のすべてのマークには、色付きのフラグを付けることができます。マークが、選択されたオブジェクトを表す場合には、フラグはそのオブジェクトに付けられます。マークが視点位置を表す場合には、フラグはその位置に付けられます。マークに移動するには、フラグをクリックします。「表示設定」メニューの「マークフラグ」を選択すれば、すべてのフラグをオンまたはオフにすることができます。「ツリー・ビジュアライザ表示の微調整」(107 ページ) の「マークフラグ」を参照してください。

場所を保存し、それにフラグを付けてマークするには、次の手順に従います。

1. ツリー・ビジュアライザの「表示」メニューから「マーク」パネルを選択します。
図 5-8 のウィンドウが開かれます。マークを初めて設定するのであれば、リストは空です。



図 5-8 ツリー・ビジュアライザの「マーク」パネル

2. 「マーク」ボタンをクリックして、メインウィンドウの現在の場所にマークを設定します。これによって、図 5-9 のダイアログ・ボックスが表示されます。
3. マークに付ける名前を入力し、色を選択します。デフォルト名は、選択されたオブジェクトの名前です。選択した色は、メインウィンドウに表示されるフラグの色になり、マークを表します。



図 5-9 「マークの作成」ダイアログ・ボックス

図 5-10 に、マークを表すフラグが表示されたメインウィンドウの例を示します。

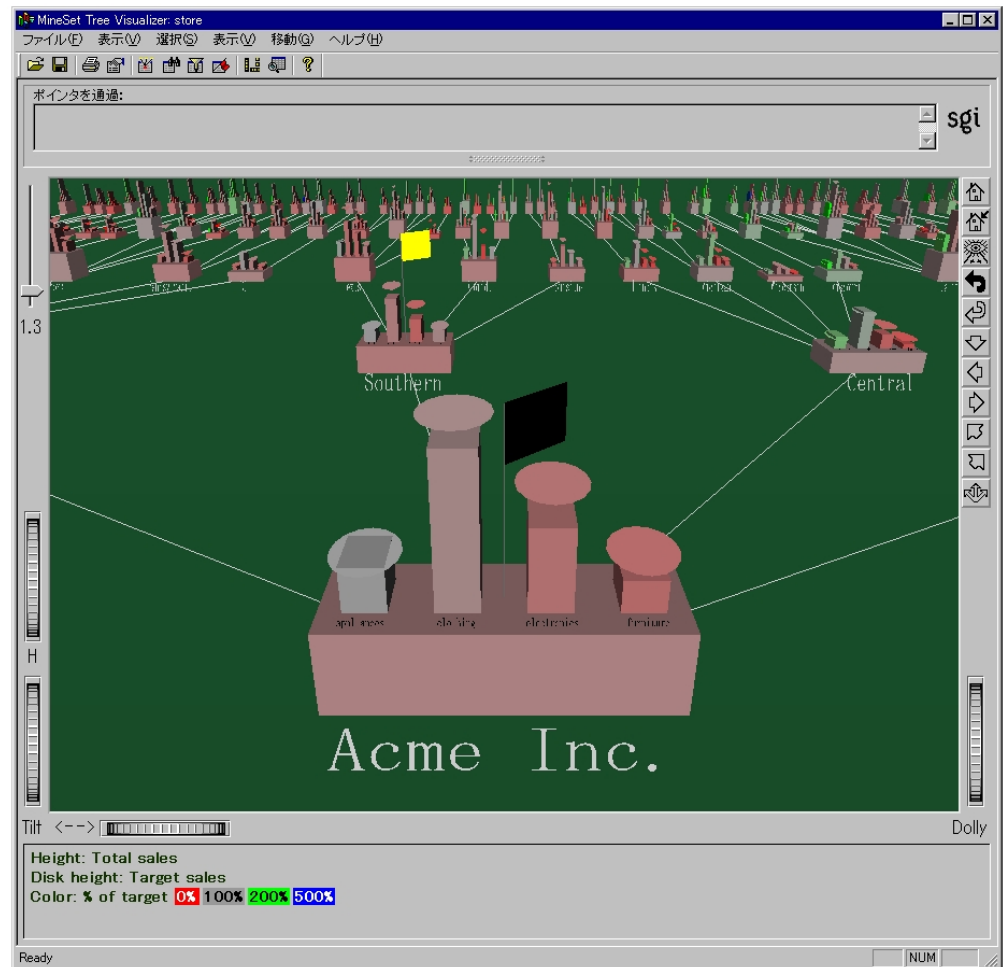


図 5-10 マークを表すフラグが表示されたメインウィンドウ

「マーク」ウィンドウの端に並んでいるボタンでは、次のことができます。

- 選択されたマークの場所に行くには、「マーク」パネルの「移動」ボタンをクリックします。マークをダブルクリックしても同じ結果が得られます。マークによって選択されたオブジェクトがすでに存在していない場合は、オブジェクトがあったところに最も近い場所が表示されます。フィルタによってオブジェクトが除外

されていたり、マークが作成された後でデータが変更されていると、オブジェクトが存在しないことがあります。

- 選択されたマークをパネルから削除するには、「削除」ボタンをクリックします。
- 選択されたマークの名前や色を変更するには、「変更」ボタンをクリックします。
- 選択されたマークのリスト順序を上へ移動するには、「上へ」ボタンをクリックします。
- 選択されたマークのリスト順序を下へ移動するには、「下へ」ボタンをクリックします。
- 「マーク」パネルを終了するには、「閉じる」ボタンをクリックします。

マーク情報を格納するファイルの名前は設定ファイルと同じで、接尾辞が `.marks` になります。マークを変更するたびに、マークはこのファイルに格納されます。すべてのマークを削除すると、`.marks` ファイルは削除されます。パーミッション・エラーなどでマークの変更が格納されない場合は、警告が出されます。この後でマークを変更しても、警告は再び出されません。

「フィルタ」パネルによるデータのフィルタリング

「フィルタ」パネルでは、選択する情報を削除して、表示される階層を微調整します。「フィルタ」パネルを使えば、特定の情報を強調したり、データ量を減らしてパフォーマンスを改善できます。「フィルタ」パネルにアクセスするには、ツリー・ビジュアライザの「表示」メニューを使用します。「フィルタ」パネルの使い方については、『*MineSet 3.0 Enterprise Edition Reference Guide*』のツリー・ビジュアライザの説明を参照してください。

ツリー・ビジュアライザ表示の微調整

表 5-3 に、ツリー・ビジュアライザの表示を微調整するオプションを示します。ツリー・ビジュアライザの「表示」メニューのオプションを使えば、一部の表示パラメータを制御できます。

表 5-3 ツリー・ビジュアライザの表示設定パラメータ

オプション	説明
ベースの高さ (トグル)	ベースの高さをオンまたはオフにします。負の数値を表示したり、バーの高さの比較を容易にするには、このオプションをオフにします。オンにすると、すべてのバーのサマリ情報が表示されます。このトグルの初期値は、設定ファイルの「base height」ステートメントで変更できます。
マークフラグ (トグル)	マークを表すフラグをオンまたはオフにします。 「マーク」パネルによる重要な場所のフラグ付け を参照してください。
ゼロ (サブメニュー)	高さがゼロのオブジェクトをどのように表示するかを制御します。デフォルトでは、このようなオブジェクトは、ほかのオブジェクトと同じように高さゼロの実線立方体（平面）として表示されます。サブメニューでは、このようなオブジェクトを枠で表示したり（空洞の四角形）、完全に隠す（描画しない）ことができます。このトグルの初期値は、設定ファイルの「zero」ステートメントで変更できます。『 <i>MineSet 3.0 Enterprise Edition Reference Guide</i> 』を参照してください。
NULL (サブメニュー)	高さが NULL のオブジェクトをどのように表示するかを制御します。このオプションはゼロメニューのオプションと同じですが、NULL オプションのデフォルトでは、オブジェクトは枠で表示されます。このトグルの初期値は、設定ファイルの「null」ステートメントで変更できます。『 <i>MineSet 3.0 Enterprise Edition Reference Guide</i> 』を参照してください。

マップ・ビジュアライザによるデータの表示

マップ・ビジュアライザ (Map Visualizer) では、データがバーグラフや見やすい地図として 3 次元ランドスケープに表示されます。マップ・ビジュアライザは、データが地理的な意味合いをもっていたり、特定の地理形状をしているときに使用します。この章の構成は次のとおりです。

- 「マップ・ビジュアライザの概要」(110 ページ)
- 「マップ・ビジュアライザの起動」(113 ページ)
- 「データの準備」(114 ページ)
- 「マップ・ビジュアライザの説明」(120 ページ)
- 「マップ・ビジュアライザでのアニメーションの作成」(123 ページ)
- 「マップ・ビジュアライザでの結果の操作」(124 ページ)

マップ・ビジュアライザの特徴や機能を示す設定ファイルとデータファイルのサンプルが用意されています。これらのファイルは、MineSet を最初にインストールした *MineSet 3.0\examples* ディレクトリにあります。

マップ・ビジュアライザの概要

マップ・ビジュアライザは、空間的な関連をもつデータを表示するためのツールです。マップ・ビジュアライザでは、地理形状を示すランドスケープの中を動的に移動するだけでなく、ドリルアップやドリルダウンして全体を表示したり、粒度を増したりすることができます。さらに、アニメーションを使えば、独立した1つまたは2つの次元でデータがどのように変化するかを見ることができます。独立した次元とは、年齢や年度など、別の項目とは無関係に変化する属性です。アニメーション・パネルは、独立次元がスライダにマッピングされているときだけ、メインウィンドウの右側に表示されます(図 6-1 を参照)。

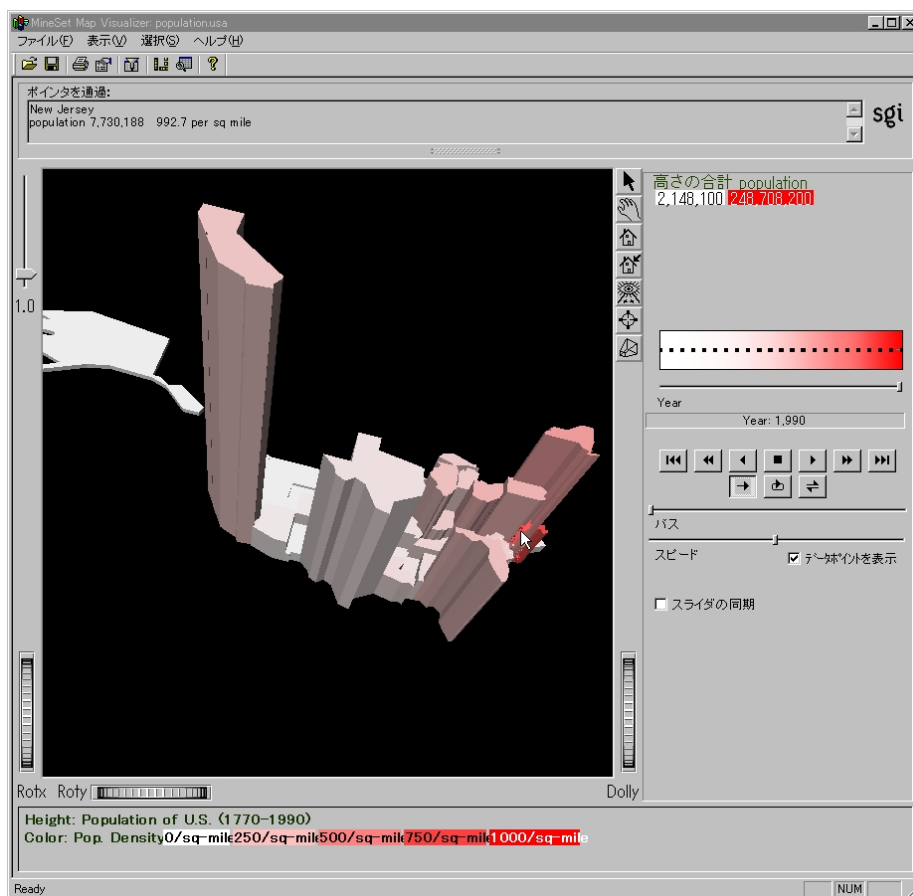


図 6-1 地理形状を表示したマップ・ビジュアライザの例

さらに、別のランドスケープでは、グラフィカル・オブジェクトが単純な輪郭線の平面に描かれ、特定の場所にバーグラフ状の円柱が配置されます（図 6-2 を参照）。

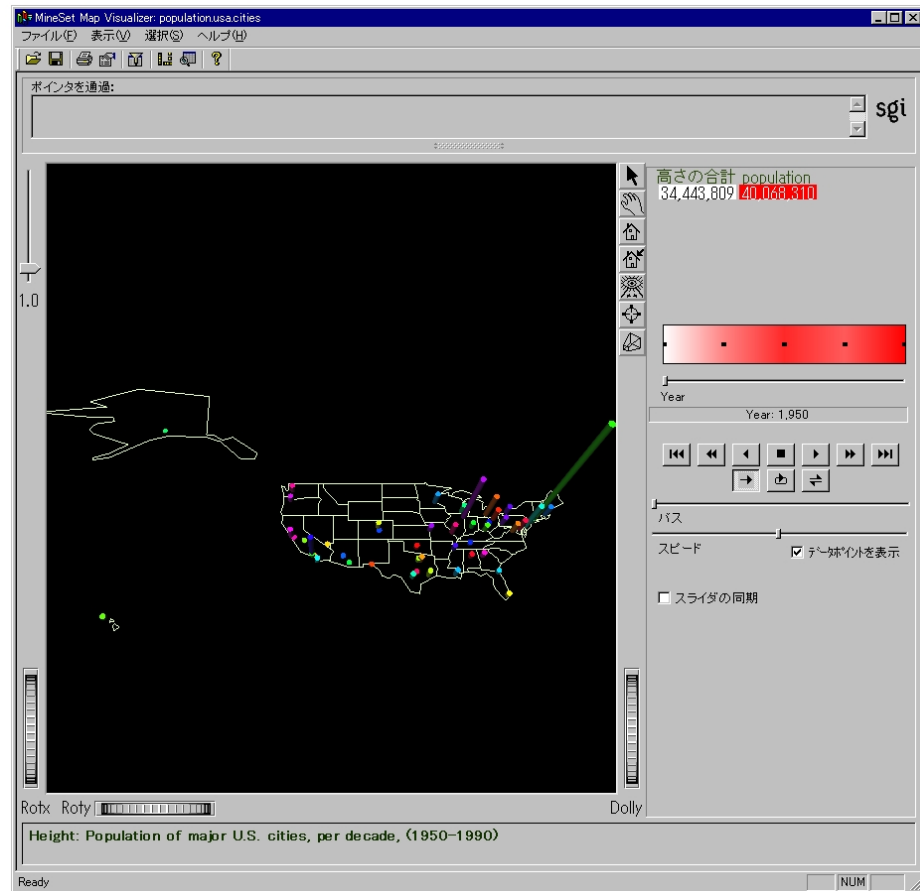


図 6-2 人口の比較をバーグラフで地図上に表示したマップ・ビジュアライザの例

別のランドスケープでは、2 地点が線で結ばれます。線は独自の幅と色で表示されます (図 6-3 を参照)。線の幅や色は、通常のオブジェクトや円柱の高さや色の特性に対応します。

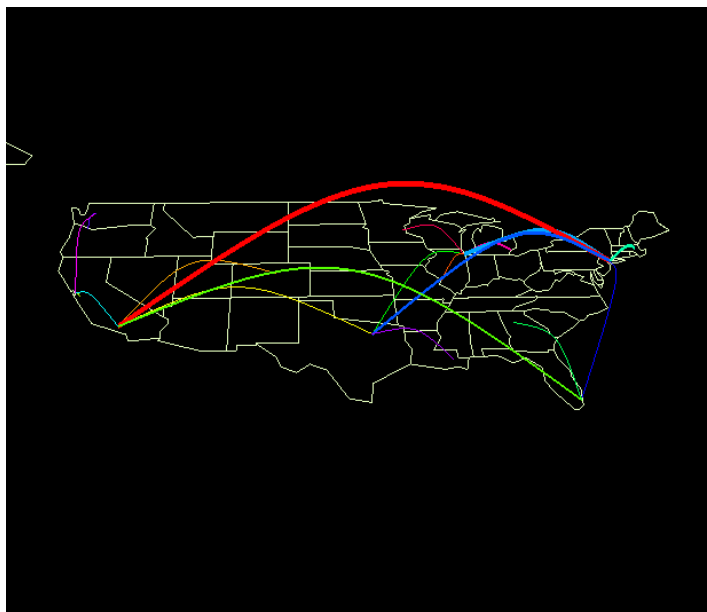


図 6-3 米国の 2 地点間を結ぶ線を表示したマップ・ビジュアライザの例

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、通常、次の手順に従います。

1. Tool Manager を起動し、サーバとデータ・ソースを選択します。必要なら [第 2 章「MineSet によるデータのアクセス」](#) を参照してください。この例では churn データセットを使用します。
2. 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、その下に並んでいるタブから「マップ・ビジュアライザ」を選択します。

データの準備

この節では、マップ・ビジュアライザで表示するために必要な地理データの変換手順について説明します。

データの集計処理

地図上にデータを分布するには、まずデータを小さなデータセットに変換する必要があります。このような変換を「変換」パネルで「集計処理」をクリックします。

1. Tool Manager の「データ変換」パネルで「集計処理」をクリックします。

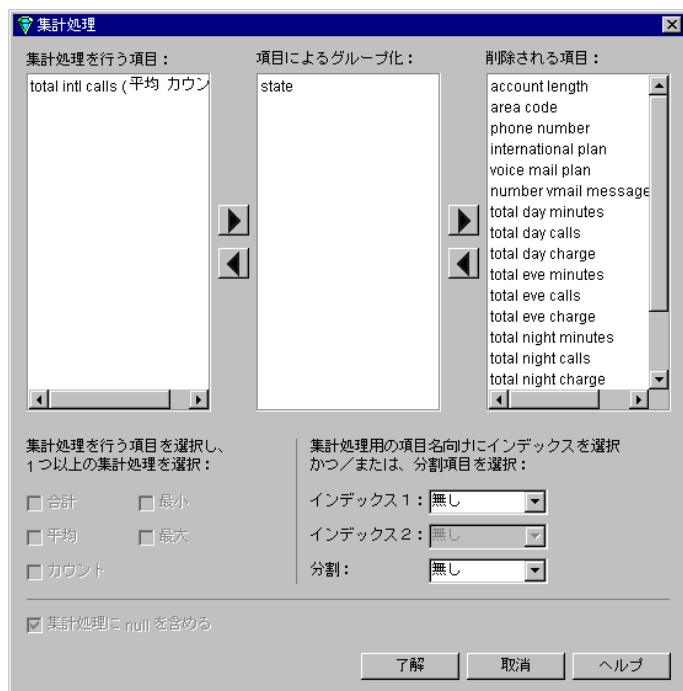


図 6-4 マップ・ビジュアライザによる集計処理

2. 必要な項目を「集計処理を行う項目」リストに移動し、下のチェックボックスを使ってアクションを指定します。地理的なグループ化に使う項目を中央の「項目によるグループ化」フィールドに残し、そのほかのすべての項目を可視化から削除します。アクションの説明は、[図 3-5 \(42 ページ\)](#) を参照してください。

たとえば、*churn* データセットの場合には、「集計処理」ダイアログ・ボックスで *total intl calls* をクリックしてから左矢印をクリックして、この項目を左のリストに移動します。「平均」と「カウント」をオンにし、「合計」をオフにします。*state* を中央のリストに残し、その他のすべての項目を右のリストに移動します。複数の項目を選択する場合は、<Ctrl> キーを押したままクリックします。「了解」をクリックして選択を適用します。

次の手順では、適切な地理形状を検索します。

マップ形状の選択

1. 「データの可視化／マイニング」パネルの「可視化ツール」タブで「ツールオプション」ボタンをクリックして、地理ファイルにアクセスします。マップ・ビジュアライザの「オプション」ダイアログ・ボックスが表示されます。
2. 「要素」テキスト・フィールドの右側にあるボタンをクリックし、いずれかの地理要素ファイルを選択します。これらのファイルは、MineSet を最初にインストールした *MineSet 3.0\config\mapviz\gfx_files* にあります。この例では *usa.state.hierarchy* を使用します。
3. 「了解」をクリックして、選択したファイルを指定し、再び「了解」をクリックしてマップ・ビジュアライザの「オプション」パネルを消します。*gfx_files* の詳細は、『*MineSet 3.0 Enterprise Edition Reference Guide*』のマップ・ビジュアライザの説明を参照してください。

次の手順では、可視化要素と項目をリンクします。

マップ・ビジュアライザの可視化要素のリンク

マップ要素と可視化要素をリンクする際には、項目の操作を行わないと項目が可視化要素の条件に合わないことがあります。

- 項目の離散値の数が多すぎる場合は、[第 3 章の「項目に対する階級の変更または新規作成」](#)の項を参照してください。
- 項目のデータ型が適切でない場合は、[第 3 章の「項目の型や名前の変更」](#)の項を参照してください。

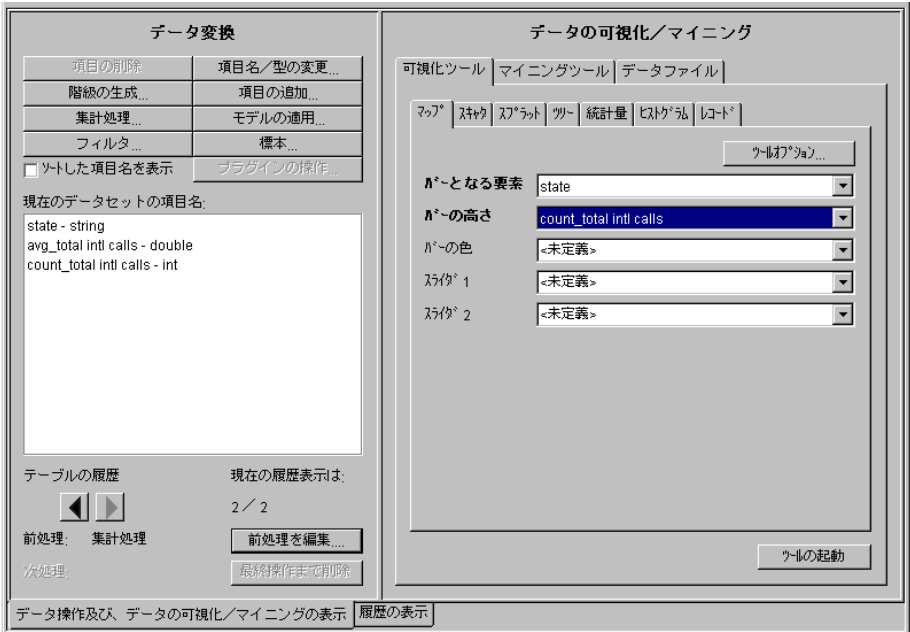


図 6-5 マップ・ビジュアライザでの項目と可視化要素のマッピング

1. Tool Manager の「データ変換」パネルで項目を要素にマッピングします。それには、各要素のポップアップ・メニューから項目を選択します。
この例では、「バーとなる要素」に state を、「バーの高さ」に count_total intl calls をそれぞれ選択します。
その他のマッピングについては、「[スライドへのマッピング](#)」と「[マップへのマッピング](#)」を参照してください。
2. 「ツールの起動」をクリックして、データセットの地理的な分散を表示します。

スライダへのマッピング

ある値が特定の条件に従って変化するように見えることがあります。たとえば、収入が年齢とともに変化するような場合です。その場合には、項目をスライダにマッピングし、他の項目を他の可視化要素にマッピングできます。たとえば、[図 6-1](#) では項目 Year がスライダにマッピングされているため、スライダを移動すれば、データが年ごとにどのように変化するかを見ることができます。

項目をスライダにマッピングするためには、項目の型が数値（整数、浮動小数点、倍精度）か階級でなければなりません。項目の階級がすでに生成されている場合は、接尾辞が `_bin` になっています。「現在のデータセットの項目名」フィールドには、`total day calls - double` のように、項目名の後に項目の型が表示されます。

スライダは、自動的に作成することも、手動で作成することもできます。項目をスライダにマッピングすると、スライダが自動的に生成されます。スライダは、自動的な階級生成と集計処理の組合せによって作成されます。用語集の「階級生成」と「集計処理」を参照してください。これらの自動操作は、「ツールの起動」をクリックした後で行われます。

現在のデータテーブルに配列項目がある場合は、配列項目のインデックスに対応するスライダを作成できます。現在のスライダのインデックスは「ツールオプション」ダイアログ・ボックスに示されます。これらのインデックスを表示するには、Tool Manager の「データの可視化／マイニング」パネルで「可視化ツール」タブをクリックしてから「ツールオプション」ボタンをクリックします。現在のテーブルでは、すべての配列項目のインデックスが同じでなければなりません。そうでないと、スライダは作成されません。

配列項目の詳細は、[第 3 章の「集計処理による新規項目の作成」](#)と『[MineSet 3.0 Enterprise Edition Reference Guide](#)』の集計処理の説明を参照してください。

自動スライダや手動スライダの詳細は、『[MineSet 3.0 Enterprise Edition Reference Guide](#)』の Mapviz、Scatterviz、Splatviz のスライダの作成の説明を参照してください。

マップへのマッピング

MineSet に付属している `.gfx` ファイルや `hierarchy` ファイルには、わかりやすい地理形状やその他の地理的な要素が含まれています。「バーの要素」にグラフィック要素をマッピングしないと、意味のあるマップは作成できません。別の可視化要素や地理マップを表示する場合は、「[マップ形状の選択](#)」（[115 ページ](#)）を次の手順で置き換えてください。

1. Tool Manager の「可視化ツール」タブで「ツールオプション」ボタンをクリックします。
2. ダイアログ・ボックスから必要なオプションを選択します。図 6-6 の「マップビジュアライザのオプション」ダイアログ・ボックスが表示されます。
 - マップ形状の要素の場合は、「要素ファイル」テキスト・フィールドの右側にある「ファイル検索」ボタンをクリックし、マップ・ビジュアライザのメインウィンドウでオブジェクトの表示に使用する *.hierarchy* 地理要素ファイルを選択します。このファイルは */config/mapviz/gfx_files* にあります。
 - 四角形の場合は、「アウトラインファイル」テキスト・フィールドの右側にある「ファイル検索」ボタンをクリックし、描画する輪郭オブジェクトを指定します。このようなオブジェクトは、3D 要素オブジェクトが上に置かれた平面として表示されます。

「要素ファイル」や「アウトラインファイル」フィールドの指定は任意です。「要素ファイル」を指定しないと、要素オブジェクトは、シーンに恣意的に配置された四角形の集まりとして表示されます。

その他のオプションについては、『*MineSet 3.0 Enterprise Edition Reference Guide*』のマップ・ビジュアライザの説明を参照してください。

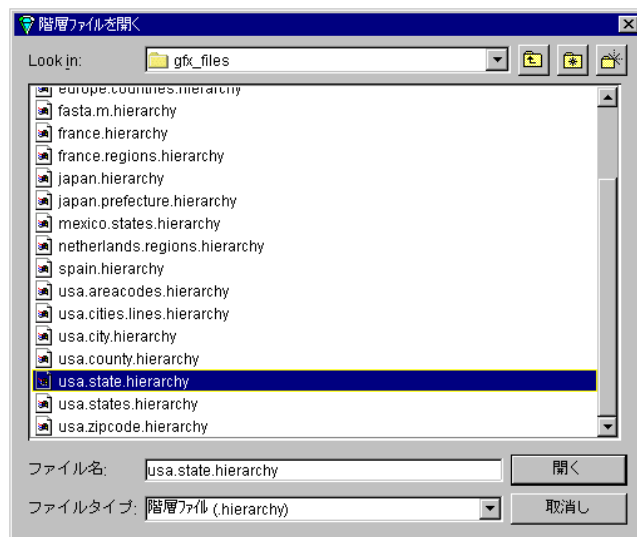


図 6-6 マップ・ビジュアライザの「オプション」ダイアログ・ボックス

3. 「ツールオプション」ダイアログ・ボックスの変更が終わったら「了解」をクリックして Tool Manager のメイン画面に戻ります。ここで「ツールの起動」をクリックすることができます。

マップ・ビジュアライザの説明

他のビジュアライザの場合と同じように、たとえば、Tool Manager の「可視化ツール」プルダウン・メニューを使って、設定ファイルを指定せずにマップ・ビジュアライザを起動すると、メインウィンドウは空で、「ファイル」、「表示」、「ヘルプ」プルダウン・メニューしか使用できません。すべてのメニューとコントロールを表示するには、「ファイル」->「開く」を選択して設定ファイルのリストを表示します。これらのファイルは、MineSet を最初にインストールした *MineSet 3.0\examples* ディレクトリにあります。

有効な設定ファイルが指定されていれば、[図 6-7](#) のような地理形状を表すランドスケープが表示されます。これは、*population.use.mapviz* を指定し、*Year* スライダを右端に移動した例です。

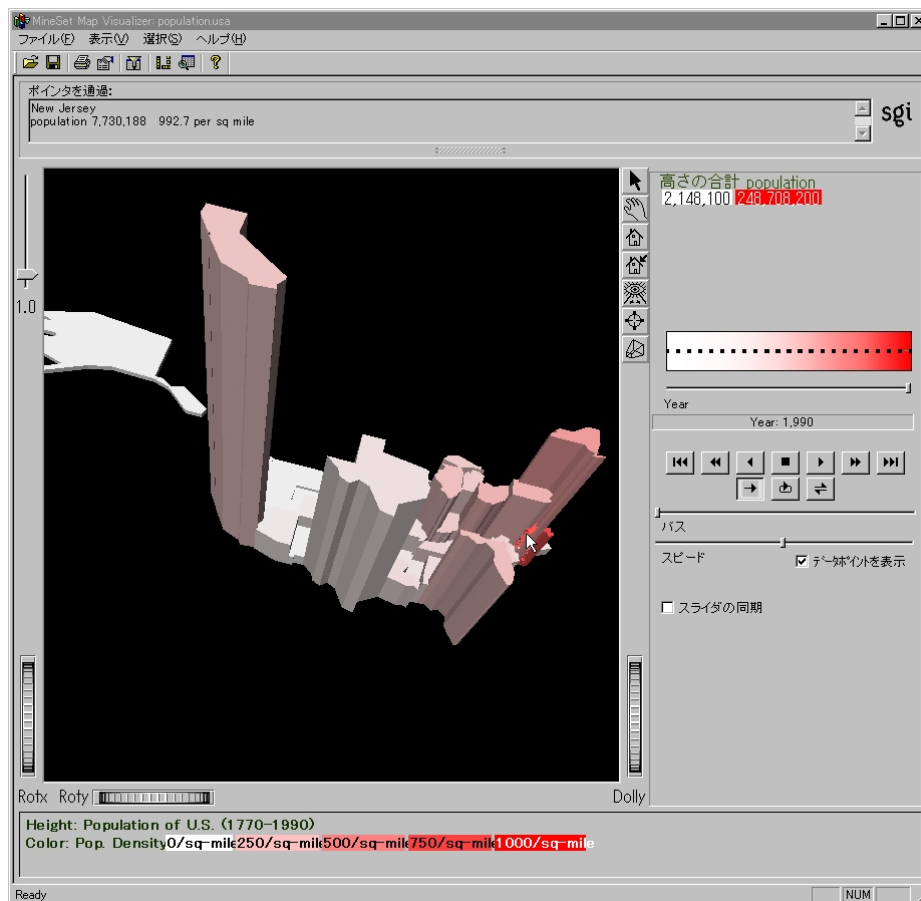


図 6-7 Year スライダを 1990 年に移動したときの population.use.mapviz の例

この例では、米国の州ごとに人口と人口密度が表示されます。州の人口は図形の高さで表されます。高さは、メインウィンドウ右側にあるアニメーション・コントロール域全体における州間の相対的なものです。

表示モード

表示モードには「ハンド」と「選択」があります。このモードを相互に切り替えるには、カーソルをメインウィンドウに移動し、<Esc> キーを押します。あるいは、ウィ

ンドウの枠上で矢印アイコンまたはハンドアイコンをクリックします。詳細は、[70 ページ](#) を参照してください。

ドリルアップとドリルダウン

.data ファイルや .hierarchy ファイルがドリルスダウンやドリルアップをサポートする場合は、選択モードでドリルダウンすると詳細度が増し、ドリルアップすると表示の概略度合いが増します。マップ・ビジュアライザのサンプルファイルでは、ドリルアップとドリルダウンが両方ともサポートされます。

表 6-1 ドリル方法

目的	操作 :
オブジェクトをドリルダウンする	オブジェクトの上でマウスの右ボタンをクリックします。
すべてのオブジェクトをドリルダウンする	背景の上でマウスの右ボタンをクリックします。
オブジェクトをドリルアップする	オブジェクトの上で <Ctrl> キーを押したままマウスの右ボタンまたは中ボタンをクリックします。
すべてのオブジェクトをドリルアップする	背景の上で <Ctrl> キーを押したままマウスの右ボタンまたは中ボタンをクリックします。

ドリルダウンは、データがサポートする最も詳細なレベルまで繰り返し行うことができます。オブジェクトの数が多ければ多いほど、ドリル操作の後でマップ・ビジュアライザがシーンを再構築するための時間が長くなり、アニメーション・コントロールを動かすときの性能が悪くなります。

デフォルトでは、オブジェクトは最初に最も詳細なレベルで表示されます。したがって、最初は、詳細度合いが減少するドリルアップだけがアクティブです。

NULL データの可視化

データが存在しないことを示す NULL があると、その位置に疑問符が表示されます。マップ・ビジュアライザでは、データベースやデータファイルに NULL が含まれている場合やその他の原因で NULL が発生します。『*MineSet 3.0 Enterprise Edition Reference Guide*』の NULL の説明を参照してください。

マップ・ビジュアライザでのアニメーションの作成

アニメーションの作成は、メインウィンドウの右側にあるアニメーション・コントロールを使って行います。アニメーション・ウィンドウは、1つまたは2つのスライダがマッピングされているときだけ表示されます。『*MineSet 3.0 Enterprise Edition Reference Guide*』のアニメーションの説明を参照してください。

マップ・ビジュアライザでの結果の操作

そのほかのマップ・ビジュアライザの機能にアクセスするには、「表示」と「選択」のプルダウン・メニューを使用します。

マップ・ビジュアライザの表示の変更

「表示」メニューには、表 6-2 のオプションがあります。詳細は、『*MineSet 3.0 Enterprise Edition Reference Guide*』の表示メニューの説明を参照してください。

表 6-2 マップ・ビジュアライザの表示メニューの項目

オプション	説明
フィルタパネル	「フィルタパネル」を表示します。このパネルでは、選択した条件に基づいてメインウィンドウに表示する要素の数をフィルタリングします。
拡張コントローラ	メインウィンドウ周囲の外部コントロールを表示または消去します。
アニメーション・パネル	メインウィンドウの右側のアニメーション・コントロール・パネルを表示または消去します。
データポイント	2D サマリウィンドウのグリッドを表示または消去します。このグリッドは黒いドットからなり、データの正確な値を表します。
ランダムカラーの使用	設定ファイルに指定された色のマッピングをアクティブまたは非アクティブにします。
X-Y 座標の表示	.gfx ファイルの開発や改良に使用します。これはデータの分析用ではありません。詳細は、『 <i>MineSet 3.0 Enterprise Edition Reference Guide</i> 』のマップ・ビジュアライザの説明を参照してください。

選択とドリルスルー

「選択」メニューでは、選択したデータに対し表 6-3 のアクションを行うことができます。詳細は、『*MineSet 3.0 Enterprise Edition Reference Guide*』のマップ・ビジュアライザの説明を参照してください。

表 6-3 マップ・ビジュアライザの選択メニューの項目

オプション	説明
値の表示	選択されたオブジェクトの値をレコードビューワのテーブルとして表示します。
オリジナルデータの表示	選択されたレコードを検索し、レコードビューワのテーブルとして表示します。
Tool Manager に送信	現在の選択に基づくフィルタ操作を Tool Manager 履歴の最初に挿入します。
ドリルスルーにスライダを使用	ドリルスルー表現を作成するときにスライダ位置を使用します。デフォルトでは、ドリルスルーはスライダの位置によって定義されるレコードに限定されます。
補集合データのドリルスルー	「オリジナルデータの表示」や「Tool Manager に送信」によって選択されていないすべてのデータを取り出します。

ドリルスルーの詳細は、『*MineSet 3.0 Enterprise Edition Reference Guide*』のドリルスルーの説明を参照してください。

マップ・ビジュアライザ用のその他の標本抽出については、『*MineSet 3.0 Enterprise Edition Reference Guide*』の標本抽出の説明を参照してください。

予測モデリングの理解

この章では、予測モデリングの概要と、MineSet を使ってデータセットから予測モデルを生成する方法を説明します。MineSet が生成する可視化では、モデルがどのように動作するのがわかります。ドリルスルーでは、元になっているデータを調べることができます。この章の構成は次のとおりです。

- 「予測モデリングの概要」(127 ページ)
- 「モデルの生成」(128 ページ)
- 「予測モデルの評価」(137 ページ)
- 「予測モデルの適用」(145 ページ)
- 「この後の学習順序」(147 ページ)

可視化とモデルについては、この後の各章で詳しく説明します。技術的な詳細は、『*MineSet 3.0 Enterprise Edition Reference Guide*』を参照してください。

予測モデリングの概要

MineSet には、予測モデルを生成するデータ・マイニング・アルゴリズムがいくつか含まれています。予測モデルでは、レコード内の他の属性の値に基づいてラベルと呼ぶ属性の値を予測します。他の属性の情報が十分であれば、モデルはラベルを正確に予測できます。予測されたラベルは、レコードでは知られていない未知の特性を示します。たとえば、クレジット履歴データセットで属性「age」、「sex」、「occupation」の値が与えられれば、予測モデルを使って属性「credit_risk」の値を予測できるかもしれません。

ラベルが分からない場合は予測モデリングを使用しますが、ラベルが分かっているデータを可視化する場合は、記述モデリングの方が適しています。MineSet の予測モデリングにはクラス判別と回帰があります。クラス判別では離散値を、回帰では連続する範囲内の値をそれぞれ予測します。これらについては、この後の各章で説明します。この章では、すべての予測モデリングに共通する処理を説明します。

モデルの生成

予測モデルを生成する MineSet のアルゴリズムは、予測する属性のデータ型に依存します。クラス判別では、MineSet のアルゴリズムはエビデンスモデル (Evidence Model)、決定木 (Decision Tree) モデル、選択式決定木 (Option Tree) モデル、デシジョン・テーブル (Decision Table) モデルを生成します。回帰タスクでは、回帰ツリー (Regression Tree) を生成します。

次の 5 つの項では、MineSet がサポートする各種の予測モデルについて、その概要と生成方法を説明します。モデル生成アルゴリズムを調整するには、Tool Manager の「データの可視化／マイニング」パネルにある「詳細設定」ボタンを使用します。これらのオプションがモデルの生成にどのような影響を与えるかについては、『*MineSet 3.0 Enterprise Edition Reference Guide*』の Tool Manager の分析モードに関する説明を参照してください。

個々の分析に関する資料については、『*MineSet 3.0 Enterprise Edition Interface Guide*』の「付録 A」を参照してください。

エビデンス・モデル (Evidence Model)

この章では、エビデンス・クラシファイアと呼ぶ予測モデルの生成方法を説明します。詳細は、第 10 章「エビデンス・クラシファイアとビジュアルライザによるモデリングと予測」を参照してください。

予測モデルを作成するには、まず、Tool Manager を使ってデータセットを選択する必要があります。詳細は、第 2 章の「MineSet の起動」を参照してください。この例では、churn データセットを使って、他の電話会社に移りそうな顧客を予測するモデルを作成します。

1. Tool Manager の「データの可視化／マイニング」パネルで「マイニングツール」タブをクリックします。
2. 「クラス判別」タブをクリックし、ポップアップ・メニューから次のように選択します。

「モード」：「クラシファイアのみ」

「分析」：「エビデンス」

「離散型ラベル」：「churned」（例）

ここでは、解約しそうな顧客の特徴を明らかにするエビデンスモデルを作成します。この例ではすべてのデータを使ってモデルを作成しますので、精度をテストするためにデータを別に保存する必要はありません。

3. 「実行」をクリックします。

Tool Manager 下部の「ステータス」ウィンドウに分析処理の進捗とサマリ情報が表示されます。分析が終了すると、エビデンス・ビジュアライザが自動的に起動され、モデルが可視化されます (図 7-1 を参照)。

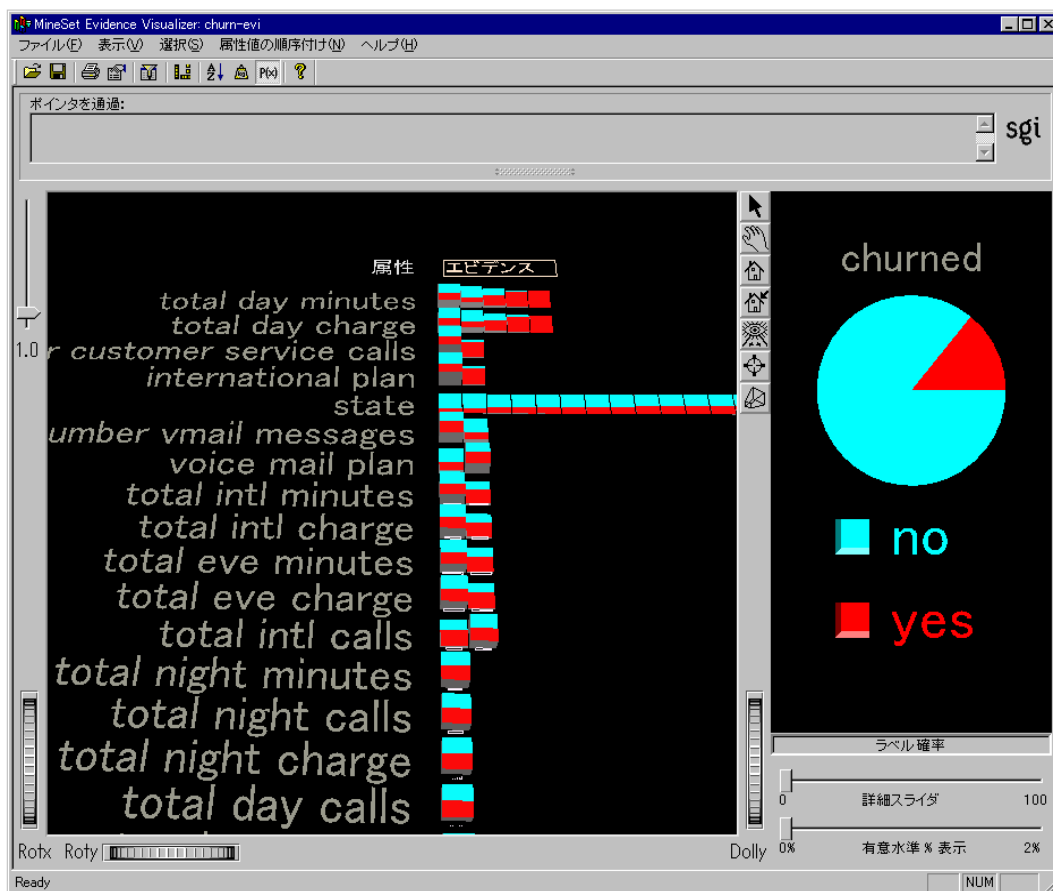


図 7-1 エビデンスによって生成された churn データセットのエビデンスの可視化

このモデルは、各属性が churn の値をどのくらいよく予測するかを表しています。このモデルを使って行動を予測するにはどうすればよいかの説明は、[第 10 章「エビデンス・クラシファイアとビジュアライザによるモデリングと予測」](#)を参照してください。

決定木モデル (Decision Tree Model)

前の例では、エビデンスという方法を使ってクラシファイアの導出と可視化を行いました。決定木では、さまざまな属性の相互作用を表示できます。つまり、属性値のさまざまな組み合わせが、予測されるラベルにどのような影響を与えるかを表示できます。決定木では、データが後続のノード（決定点）でどのように分割されるかは、前のノードで行われた決定に依存します。[図 7-2](#)に示す例では、ルートのノードは、「日中通話の合計時間 (total day minutes)」属性が 264.45 より大きい小さいかによって分岐しています。これは、これ以後の分岐点における値の分割に影響を与えます。この可視化では、左側のバーが明らかに高くなっています。

決定木モデルの生成手順は、「クラス判別」タブで「分析」の代わりに「決定木」を選択するのを除けば、エビデンス・モデルの手順と同じです。この手順については、[第 8 章「決定木、選択式決定木、回帰ツリーによるモデリングと予測」](#)で詳しく説明します。[図 7-2](#)にこの方法の結果を示します。ルートから分岐しているサブツリーでクラスの分割が異なっているのは、属性間の相互作用が大きいことを示しています。

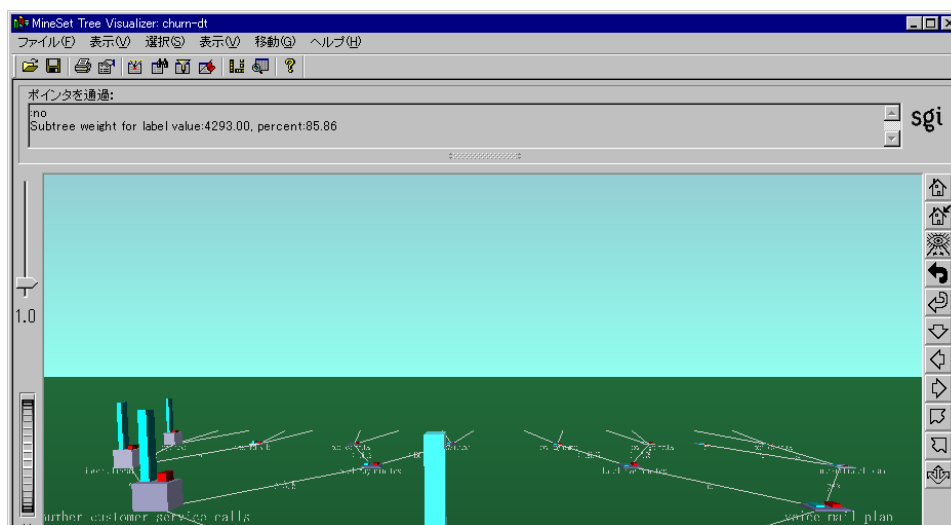


図 7-2 決定木によって生成された churn データセットの決定木モデル

この決定木モデルの各ノードのバーはラベルの値（クラス）の分割を表しています。バーの上にカーソルを移動すると、そのラベルの値に対するレコードの重みと割合が画面上部のステータス域に表示されます。ノードのベースの高さは、レコードの重みを表します。重みを使用されていない場合は、カウント（レコード数）が使用されます。

この例では、決定木のルートになっているブロックに「total day minutes（日中通話の合計時間）」というラベルが付いています。これは解約を予測する最も重要な要素であることを示します。顧客の通話時間を示すこのラベルは、区間 264.45 分で分岐します。決定木分析については、[第 8 章「決定木、選択式決定木、回帰ツリーによるモデリングと予測」](#)で詳しく説明します。

選択式決定木モデル (Option Tree Model)

決定木では、サブツリーごとに「最善の」属性を 1 つ選択します。しかし、よい属性がいくつかあり、それぞれで分岐したい場合があります。そのような場合には、選択式決定木を使って選択ノードを作成します。選択式決定木は決定木に似ていますが、その生成には選択式決定木を使用します。

最上位ノードは「選択枝ノード」です。この選択枝ノードでは、ルートで選択できるいくつかのよい属性が示されています。各選択枝ノードで生成できる選択枝の数は設定できます。それぞれの選択枝はそのポイントにおける最善の属性を表します。それぞれの選択枝で分岐が生成されます。[図 7-3](#) の例では cars データセットを使用しています。選択式決定木については、[第 8 章「決定木、選択式決定木、回帰ツリーによるモデリングと予測」](#)で詳しく説明します。

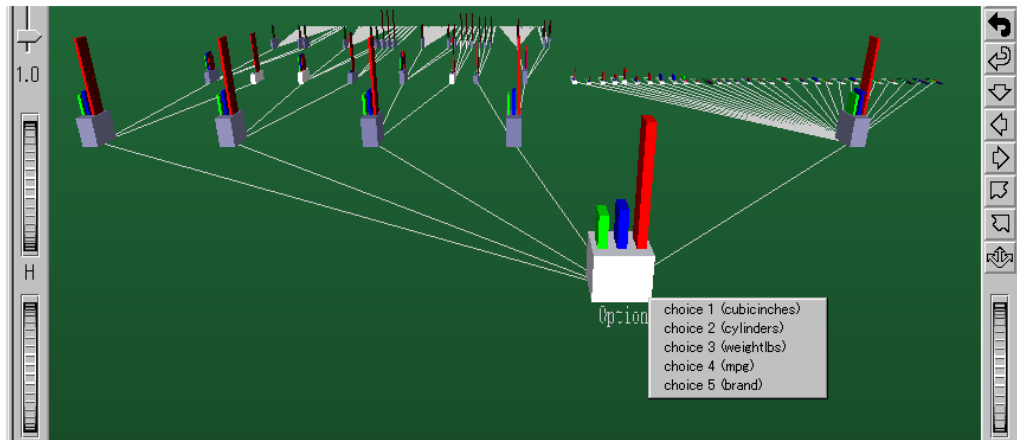


図 7-3 選択式決定木によって生成された cars データセット

このデータセットでは（図 7-3）、自動車がヨーロッパで製造されたのか、日本で製造されたのか、米国で製造されたのかを予測します。決定木ではルートに立方インチを選択しましたが、選択式決定木では、その代わりに立方インチ、シリンダ数、重量、燃費、車種をルートの選択枝として選択します。選択枝を知るには、ルートで <Ctrl> キーを押したままマウスの右ボタンをクリックします。

選択枝ノードはルート以外の場所にも設定できます。デフォルトでは、選択枝ノードはルートかその 1 つ下のレベルだけに設定されます。

選択式決定木は決定木よりも作成に 10 ～ 15 倍の時間がかかり、サイズもずっと大きくなります。しかし、選択式決定木には、次の 2 つの利点があります。

- 柔軟性 — 選択枝ノードでは、いくつかの類似した選択枝を見ることができます。したがって、単一の属性で満足する代わりにいくつかの属性から選択できます。選択式決定木では、理解しやすい選択枝や問題の背景知識に合った選択枝を選択できます。
- 正確性 — 通常、選択式決定木分析の誤差率は決定木分析よりも低くなります。選択式決定木分析では、選択枝ごとにラベルの値の「投票」を行い、投票を平均することによってクラス判別します。これは、それぞれの専門家が独自の主要条件に基づいてラベルを予測するのに似ています。選択枝ノードではすべての専門家の投票を平均化しますので、クラシファイアの安定性が増し、危険度が減少します。

デシジョン・テーブル・モデル (Decision Table Model)

デシジョン・テーブルでは、2つの属性間の相関を、層別化した階層で表すモデルを生成します。決定はデシジョン・テーブルによって決定木と同じようにして行われますが、属性は、特定のサブツリーではなく、ツリーのすべてのレベルに渡って評価されます。結果は、ツリーではなく階層テーブルとして表示されます。いずれかのブロックをクリックすると、ケーキと呼ばれる四角形のブロックがどのような構成になっているのかが分かります（図 7-4 を参照）。

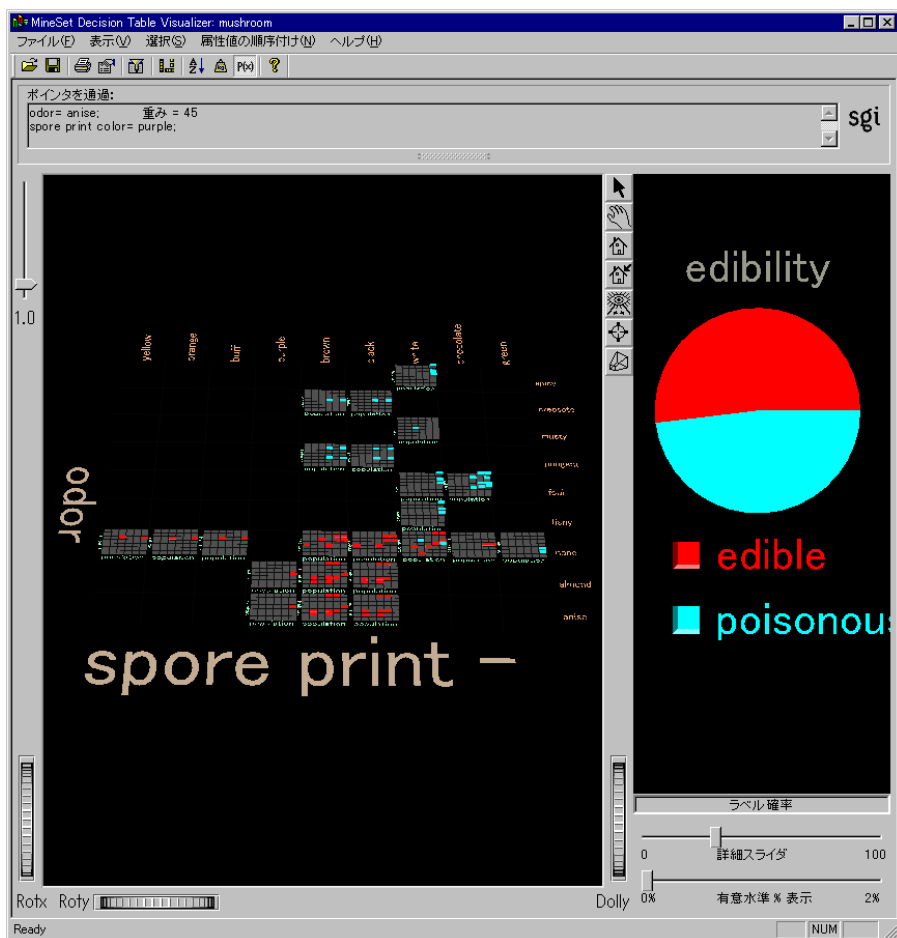


図 7-4 デシジョン・テーブルによって生成された mushroom データセットのデシジョン・テーブル・モデル

このモデルは、キノコが食用に適しているかどうかの予測でのラベル確率を示します。

デシジョン・テーブルの使い方については、[第 9 章「デシジョン・テーブル・クラシファイアとビジュアライザによるモデリングと予測」](#)を参照してください。

回帰ツリーモデル (Regression Tree Model)

回帰ツリー分析 (Regression Tree Inducer) では、回帰ツリー分析モデルと呼ぶ、クラシファイアに似た予測モデルを生成します。回帰ツリー分析モデルでは年棒など連続する範囲のラベルと値を予測するのに対し、クラシファイアでは yes や no とか、20 ～ 100 のような特定の範囲を示す離散値を予測します。

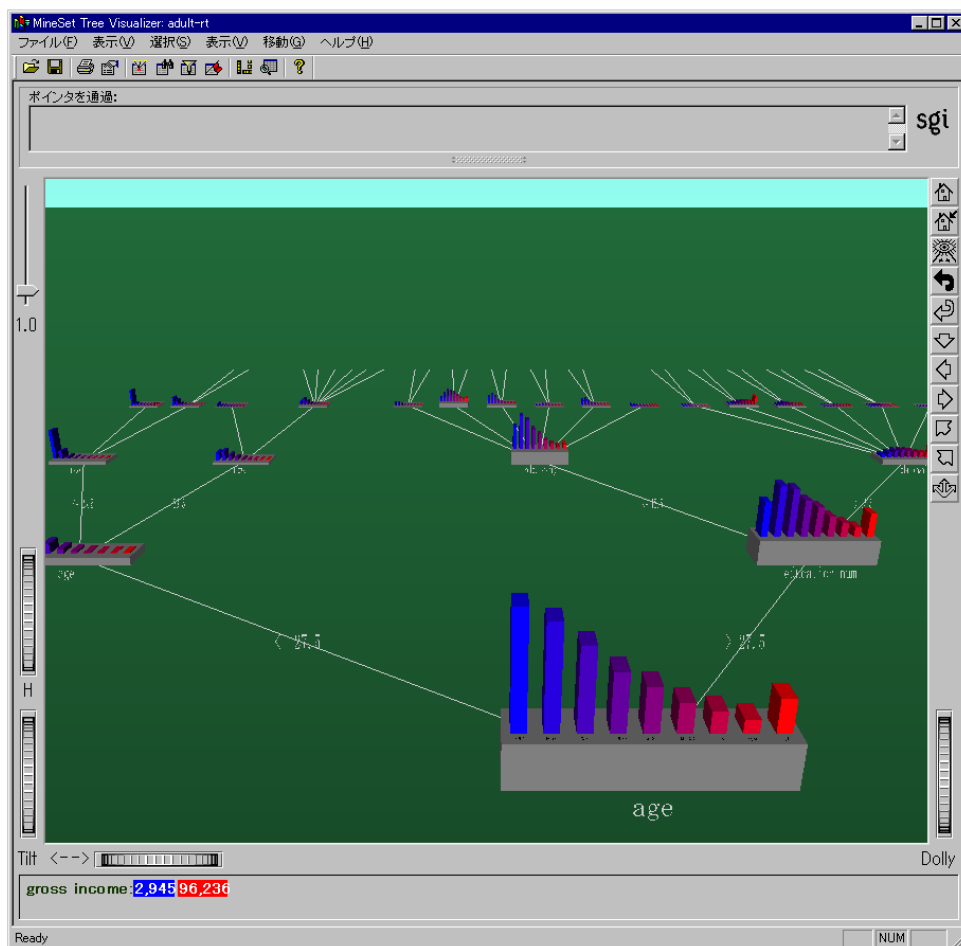


図 7-5 回帰ツリー分析によって生成された adult データセットの回帰ツリー分析モデル

次の例は、回帰ツリー分析モデルの生成方法を示したものです。この手順については、[第 8 章「決定木、選択式決定木、回帰ツリーによるモデリングと予測」](#)で詳しく説明します。

予測回帰ツリー分析モデルを作成するには、まず、Tool Manager を使ってデータセットを選択する必要があります。詳細は、[第 2 章「MineSet によるデータのアクセス」](#)を参照してください。この例では、*adult* データセットを使用します。

1. Tool Manager の「データの可視化／マイニング」パネルで「マイニングツール」タブをクリックします。
2. 「回帰ツリー (Regress)」タブを選択し、ポップアップ・メニューから次のように選択します。

「モード」：「回帰ツリーのみ」

「分析」：「回帰ツリー分析」

「連続型ラベル」：「gross income」（例）

ここでは、データセットのレコードで表される成人の総収入の特徴を明らかにする回帰ツリー分析モデルを生成します。この例ではすべてのデータを使ってモデルを作成します。

3. 「実行」をクリックします。

Tool Manager 下部の「ステータス」ウィンドウに分析処理の進捗とサマリ情報が表示されます。ダイアログ・ボックスにも、進捗を示すインジケータが表示されます。分析手順が終了すると、回帰ツリー・ビジュアライザが自動的に起動され、モデルが可視化されます ([図 7-5](#) を参照)。データセットによっては、他のデータセットに比べ生成時間が長くなる場合があります。

回帰ツリー分析モデルを生成する場合は、データセットに連続属性が含まれていなければなりません。つまり、データセットには、連続する範囲のデータをもつ項目が必要です。このような項目がない場合は、Tool Manager の「データ変換」パネルで連続項目を新たに追加します。[第 3 章の「項目の削除と追加」](#)の項を参照してください。

学習曲線クラシファイア (Learning Curve classifier) モードについては、[「学習曲線 \(Learning Curve\) の作成」 \(141 ページ\)](#) で説明します。

予測モデルの評価

予測モデルの目的は予測を行うことであり、したがって、予測モデルの有用性は予測の精度にはっきり依存します。モデルの精度は誤差率によって評価します。この節では、MineSet のモデリングモード、すなわち、すべてのデータを使用し、誤差推定を行わない「クラシファイアのみ (Classifier Only)」モードと、誤差推定を行う 2 つのモード（予備法と相互検証法）について説明します。

誤差推定を行う予備法では、データセットのある部分（通常は 3 分の 2）をモデルの生成に使用します。分析ではこの訓練事例内のラベルを使ってモデルを生成します。残りのデータは、モデルのテストや誤差率の推定に使用します。

クラシファイアの誤差を評価する相互検証法では、データセットをある数（通常は 10）の層またはサブセットに分割し、同じ数のクラシファイアを作成します。このプロセスを何回か繰り返せば、評価の精度が向上します。

どちらの方法でも、将来のレコードは訓練事例と同じデータ分布から標本抽出できなければなりません。詳細は、『*MineSet 3.0 Enterprise Edition Reference Guide*』の「誤差推定」と「相互検証」の説明を参照してください。.

すべてのデータを使用したクラス判別

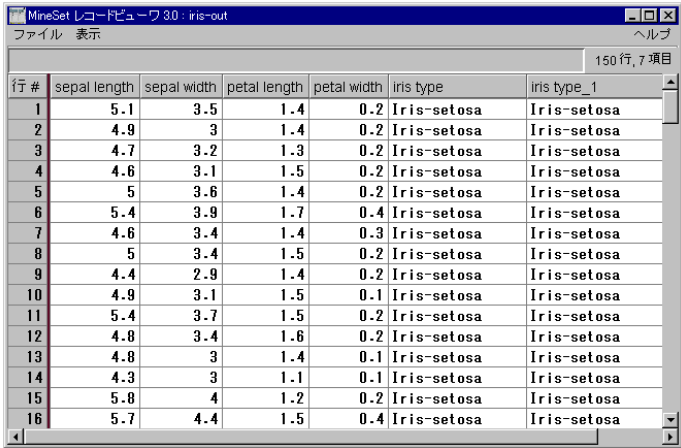
「クラシファイアのみ (Classifier Only)」モードでモデルを作成すると、使用可能なすべてのデータを使ってクラシファイアが生成されます。このモードは、可視化だけを見たい場合など、予測モデルの誤差率を推定する必要がないときに使用します。

たとえば、「クラシファイアのみ (Classifier Only)」モードを使って iris データセットの決定木モデルを作成するとします。次の例では、モデルの作成と適用に必要な一連の手順を行います。次の手順に従えば、モデルのクラス判別の間違いを発見できます。

1. Tool Manager のウィンドウで「ファイル」->「新しいデータファイルを開く」を選択し、プルダウン・メニューからデータセットを選択します。ここでは、*iris.schema* データセットを選択してください。
2. Tool Manager の「データの可視化／マイニング」パネルで「マイニングツール」->「クラス判別」タブを選択し、次のように指定します。
 - 「モード」：「クラシファイアのみ」
 - 「分析」：「決定木分析」
 - 「離散型ラベル」：「iris type」

使用可能なすべてのデータを使って決定木モデルを作成します。このモデルでは、どの属性がアヤメ (iris) の種類を分類するかを判定します。

3. 「実行」をクリックして決定木を実行します。可視化を調べてから消去します。このモデルではすべてのデータを使用しました。次の手順では、このような場合に正しい予測と正しくない予測をどのようにして評価するかを調べます。
4. Tool Manager の「データ変換」パネルで「モデルの適用」をクリックします。アヤメのデシジョン・テーブル・クラシファイアとして「*iris-dt.class*」を選択し、「了解」をクリックして、デフォルトの新規項目名 *iris type_1* を受け入れます。この項目には、予測されたラベルが含まれています。
5. 「項目の追加」ボタンをクリックして、誤差を識別する別の新規項目を作成します。この項目を *iris_fault* とします。これは、表現 '*iris type*' != '*iris type_1*' を持つ、データ型が整数の項目として定義します。
この表現を作成するには、左のリストから表現を選択して右矢印をクリックするか、表現をテキストフィールドに直接入力します。
6. 「表現をチェック」->「了解」をクリックします。新規項目 *iris_fault* が *iris type_1* として追加されます。新しい項目は、クラシファイアがクラス判別を間違えると 1 に、クラス判別が正しいと 0 になります。
7. 結果を検証します。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」をクリックし、「ツール」ポップアップ・メニューから「レコードビューワ」を選択し、「ツールの起動」をクリックします。結果の例を図 7-6 に示します。



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

図 7-6 アヤメの分類の間違いの例

結果をスキャタ・ビジュアライザに適用すれば、新規項目を色にマッピングできます。たとえば、0（誤差なし）を緑に、1（誤差あり）を赤に設定します。スキャタ・ビジュアライザの例については、[第 4 章「スキャタ・ビジュアライザとスプラット・ビジュアライザによるデータの分析」](#)を参照してください。

予備法による誤差推定

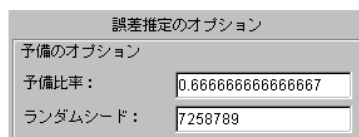
すべてのデータを使ってモデルを作成する代わりに、データセットの一部を訓練事例にしてクラシファイアを生成できます。MineSet のデフォルトモードである「クラシファイアとエラー」モードでは、データセットが独立した訓練サブセットとテストサブセットに自動的に分割されます。訓練事例として使用するデータの比率は、必要に合わせて変更できます。

選択したデータセットに予備法による誤差推定を適用するには、次の手順に従います。

- 1. `iris_schema` データセットを選択します（[137 ページ](#)を参照）。
- 2. Tool Manager の「データの可視化／マイニング」パネルで「マイニングツール」->「クラス判別」タブを選択し、次のように指定します。
 - n 「モード」：「クラシファイアとエラー」
 - n 「分析」：「決定木分析」
 - n 「離散型ラベル」：「iris type」

ここでは、iris データセットを 2 つのデータセットに分割し、訓練事例に 3 分の 2 を割り当てます。これはデフォルトの予備比率です。訓練事例を使って決定木モデルを作成しますが、残りの 3 分の 1 のレコードはモデルの誤差率の推定に使用します。

3. 「詳細設定」ボタンをクリックし、図 7-7 のテキストパネルを使って訓練事例の予備比率を変更します。さらに、訓練事例やテストセットに入れるレコードを変えたい場合は、ランダムシードを変更します。用語集の「ランダムシード (*Random Seed*)」を参照してください。「了解」をクリックしてパネルを消去します。



誤差推定のオプション	
予備のオプション	
予備比率:	0.666666666666667
ランダムシード:	7258789

図 7-7 「誤差推定のオプション」の予備比率

4. 「実行」をクリックして決定木分析を実行します。

残りの 3 分の 1 のデータからなるテストセットを使って推定した分析誤差率がメインウィンドウ下部のステータス領域に表示されます。スクロールすれば、すべての情報を見ることができます。

相互検証の誤差推定

相互検証は、最終クラシファイアを作成するときや、小さなデータセットに対して使用します。この理由はそのプロセスを見れば明白です。相互検証は、より正確な誤差推定を得るための方法です。 n 層の相互検証では、データセットは n 個の独立したサブセットに分割されます。ここで n は必要な任意の数であり、層はデータを分割して作成するいくつかのサブセットです。これらのサブセットは保存され、残りの $n-1$ のサブセットは訓練事例として結合されます。 $n-1$ は、最初に指定した数から 1 を引いたものです。結果のモデルは、保存されたサブセットを使って推定されます。次に、独立して行われた n 個の推定が平均化され、データが結合されて最終モデルが作成されます。 n 層の相互検証は、「クラシファイアとエラー」や「クラシファイアのみ」の方法に比べおよそ $n+1$ 倍の時間がかかります。

選択したデータセットに相互検証を適用し、層の数を設定するには、次の手順に従います。

1. Tool Manager の「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->「クラス判別」をクリックします。
2. 「モード」ポップアップ・メニューから「誤差推定」を選択して誤差を推定します。これは、「クラシファイアのみ」を選択してすべてのデータを使用する場合と同じです。ただし、相互検証を使用するため、実行時間は通常長くなります。データが非常に少ない場合は、「誤差推定」を使用します。生成されるクラシファイアは、「クラシファイアのみ (Classifier Only)」モードで生成したものとまったく同じです。
3. 「詳細設定」をクリックして相互検証の動作方法を変更します。「誤差推定のオプション」(図 7-8 を参照) で「層数」や「回数」を設定すれば、誤差推定のプロセスを繰り返し行うことができます。変更を有効にするには、「了解」をクリックします。

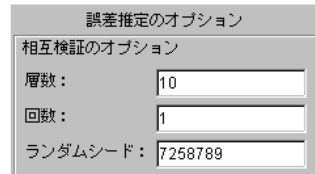


図 7-8 「誤差推定のオプション」の相互検証オプション

4. 「実行」をクリックしてアルゴリズムを実行します。

誤差推定の詳細は、『*MineSet 3.0 Enterprise Edition Reference Guide*』の該当する項を参照してください。

学習曲線 (Learning Curve) の作成

学習曲線では、分析によって生成したクラシファイアの誤差が、クラシファイアの実行に使用したレコードの数との対比で表示されます。一般に、クラシファイアの生成に使用するレコードの数が増えれば、誤差はそれだけ少なくなります。

学習曲線を作成するには、調べる訓練事例のサイズの範囲と、それぞれの訓練事例のサイズでクラシファイアを作成する回数を指定します。訓練事例のサイズを指定するには、レコードの下限值（最小）と上限値（最大）、および調べるポイントの数を入力します。それぞれのクラシファイアがレコードのランダムな標本を使って生成され、その誤差が、訓練事例に使用されなかった残りのデータを使って評価されます。

1. [130 ページ](#)で使用した *churn.schema* データセットを選択します。
2. Tool Manager の「データの可視化／マイニング」パネルで「マイニングツール」->「クラス判別」タブを選択し、次のように指定します。
「モード」：「学習曲線」
「分析」：「決定木分析」
「離散型ラベル」：「churned」（例）
電話番号項目の固有値が多すぎるため、この項目は計算されない旨のメッセージが出力されます。項目を削除するか、「了解」をクリックします。
3. 「実行」をクリックします。
学習曲線のポイントごとにクラシファイアが生成されるため、処理には長い時間がかかります。ステータス・ウィンドウには、パスごとに訓練事例のサイズ、訓練が行われる回数、平均誤差が表示されます。

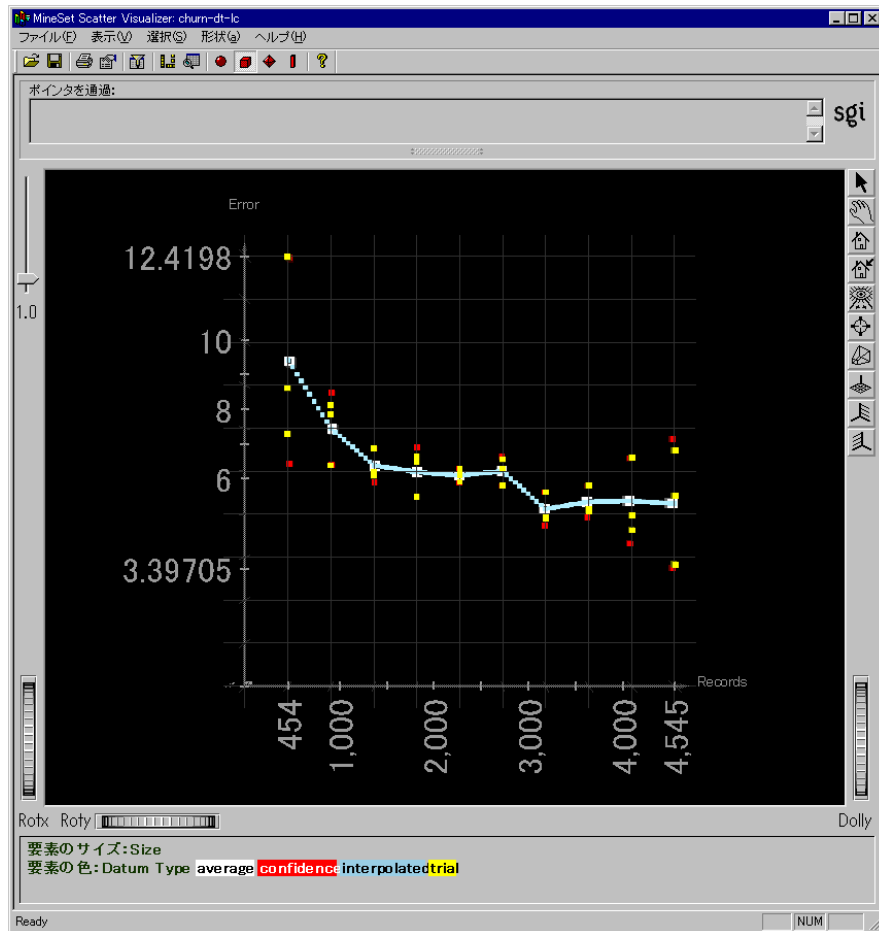


図 7-9 ラベルを churned に設定した churn データセットの学習曲線

図 7-9 は、churn データセットのレコード数に対応する誤差率を示しています。454 レコードを使って学習すると、誤差率は 9.41% です。ほぼ全部のレコードを使って学習すると、誤差率は 5.27% です。誤差率がそれ以上減少しないポイントを判定すれば、特定の誤差率を持つクラシファイアを作成するのに必要な学習レコード数がわかります。この場合には、3,000 レコードの学習で十分ことが分かります。サンプルサイズは、学習曲線のポイントから妥当な精度で選択できます。

グラフには 4 種類のポイントが表示されます。

- 黄色いポイントは、実行回数（アルゴリズムを実行した回数）から得られた実際の誤差推定です。
- 白いポイントは平均です。
- 青いポイントは白いポイント間を補間したものです。このポイントは青い線のように見えることがあります。
- 赤いポイントは、実行ごとの実際の誤差推定に基づいて計算された、平均に関する 95% 信頼区間を示します。

デフォルトではすべてのレコードを使って学習曲線を作成しますが、他のモードを指定することもできます。

1. Tool Manager の「データの可視化／マイニング」パネルで「学習曲線」、「決定木」、「churned」を選択してから、「詳細設定」を選択します。
2. 「詳細設定」ダイアログ・ボックスで次の「学習曲線のオプション」を指定します（[図 7-10](#) を参照）。
 - 学習曲線のポイント数。1 よりも大きくなければなりません。
 - ポイント当たりの実行回数
 - 始点と終点で使用するレコード数

途中のポイントで使用するレコード数は自動的に計算されます。

図 7-10 学習曲線のオプション

範囲かポイント数を省略すると、学習曲線のポイント数と訓練事例の合計レコード数から自動的に計算されます。このデフォルトでは、訓練事例の全範囲が使用されます。たとえば、80,000 レコードのファイルがあるとしします。学習曲線のポイント数に 3 を指定すると、ポイントは 2,000、4,000、6,000 のレコード数で生成されます。小さい範囲に「ズームイン」すると便利な場合があります。たとえば、学習曲線を 1,000 から 10,000 レコードの範囲だけで生成することもできます。

ポイント当たりの実行回数を増やすと、実行時間はそれに比例して増加しますが、誤差の推定は向上します。実行回数のデフォルトは3です。

学習曲線はどの分析に対しても生成できます。実行には、デフォルトを使うこともできますし、曲線のポイント数、実行回数、レコードの範囲を指定することもできます。[表 7-1](#) に学習曲線に対する選択肢の効果を示します。

表 7-1 学習曲線に対するアクション

この結果を得るには	このようにします
信頼区間の短縮（精度の向上）	実行回数を増やす。 テスト・データセットのサイズを増やす。
処理速度の向上	小さい標本を使用する。「データの標本抽出」(47 ページ) を参照

分析操作の詳細は、『*MineSet 3.0 Enterprise Edition Reference Guide*』の分析に関する説明を参照してください。

予測モデルの適用

予測モデルを作成したら、これを他のレコードに適用してそれらのラベルを予測できます。たとえば、アヤメの種類を予測するクラシファイアを作成し、このクラシファイアを記述属性だけをもつレコードに適用すれば、予測したアヤメの種類の新規項目を追加できます。

モデルの選択

「モデルのテストと適用」ダイアログ・ボックス ([図 7-11](#)) の「有効なモデル」パネルからモデルを選択します。右側のパネルに、この分析モデルに必要な項目の名前とデータ型が表示されます。

1. Tool Manager の「データ変換」パネルで「モデルの適用」ボタンをクリックします。
2. 「モデルのテストと適用」ダイアログ・ボックス ([図 7-11](#)) の「有効なモデル」パネルからモデルを選択します。右側のパネルに、このクラシファイアに必要な項目の名前とデータ型が表示されます。

これらの要件が、メモリにロードされている現在のテーブルと合致すれば、各タブの下部にあるボタン（「了解」、「テストの実行」、「データの適合」）がアクティブになります。要件が合致しないと、不足している項目が右側のリストから選択され、下部のボタンが非アクティブにされます。

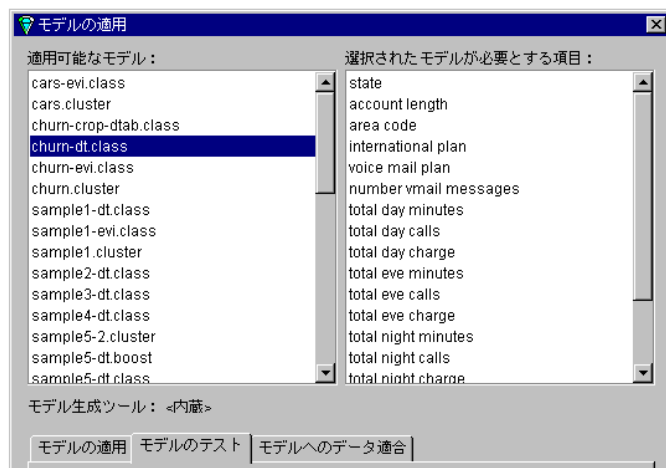


図 7-11 「モデルのテストと適用」ウィンドウ：クラシファイアの選択

モデルの適用

モデルを選択したらモデルを適用します。

1. 「モデルのテストと適用」ダイアログ・ボックスで「モデルの適用」タブ（図 7-12）をクリックし、次のどちらかを選択します。

予測値：たとえば、顧客の解約を予測する項目を追加します。新規項目では、モデルが新しいデータに対してどのくらいよい予測を行うかが分かります。

推定確率値：各レコードがあるラベルの値を持つ確率を推定します。ポップアップ・メニューから有効な値を選択します。正確さの確率を数値として推定する項目が追加されます。

2. テキスト・フィールドに新規項目の名前を入力します（たとえば、p_churned）。これは顧客が解約する確率を表します。
3. 「了解」をクリックします。

- このモデルを項目テーブルとしてグラフィック表示するには、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」をクリックし、「ツール」ポップアップ・メニューから「レコードビューワ」を選択し、「ツールの起動」をクリックします。

ラベルを付けられた項目 `churn` と `p_churned` が表示されますので、予測と結果を比較します。

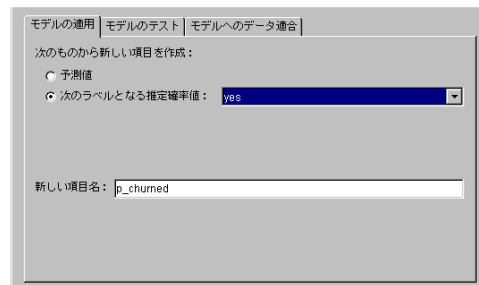


図 7-12 「モデルの適用」パネル

モデルの適用は、たとえば、マーケティング・キャンペーンで使用します。ある市でキャンペーンを行い、そのときの反応に従ってラベルの値を記録する項目を生成し、モデルを作成します。モデルを適用すれば、他の市では、クラシファイアによって反応の可能性が高いとラベルされた人だけにキャンペーン・メールを送信できるため、メール費用を節約できます。

この後の学習順序

次のいくつかの章では、データを調べるためのさまざまな方法を説明し、それぞれのビジュアライザの利点を示します。通常は、変換されていない元データをレコードビューワで見ることから始め、統計量ビジュアライザやヒストグラム・ビジュアライザに進み、次に重要項目 (Column Importance) ツールで重要な項目を検索して、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・ビジュアライザで既知データの有益な可視化を行います。

どこから開始するか特に決まっていない場合は、クラスタリングから始め、この方法に従って知識発見を行ってください。

モデルのテストと適用やモデルの精度の判定を行う準備が整っている場合は、[第 11 章「予測モデリングの改良」](#)を参照すれば、いくつかのアプローチが得られます。

決定木、選択式決定木、回帰ツリーによるモデリングと予測

決定木 (Decision Tree)、選択式決定木 (Option Tree)、回帰ツリー (Regression Tree) のアルゴリズムはかなり異なりますが、それらの外観と感覚は似通っています。この章では、ツリーに基づく 3 つのモデルを使用する上での類似点と相違点を説明します。解析や使用可能なオプションで使用するアルゴリズムの詳細は、『*MineSet 3.0 Enterprise Edition Reference Guide*』を参照してください。

この章の構成は次のとおりです。

- 「決定木、選択式決定木、回帰ツリーの概要」(150 ページ)
- 「決定木、選択式決定木、回帰ツリーの起動」(155 ページ)
- 「決定木ビジュアライザによる結果の表示」(156 ページ)
- 「選択式ツリー・ビジュアライザによる結果の表示」(160 ページ)
- 「回帰ツリー・ビジュアライザによる結果の表示」(161 ページ)
- 「決定木、選択式決定木、回帰ツリーによる予測」(163 ページ)

ビジュアライザの特徴と機能を示すために、設定ファイルとデータファイルのサンプルが付属されています。これらのファイルは、*MineSet* を最初にインストールした *MineSet 3.0\examples* ディレクトリにあります。

決定木、選択式決定木、回帰ツリーの概要

この節では、3 種類のツリーの使い方と、どういう場合にどれを選択するかを説明します。ツリーツールは、モデルを作成する分析と、結果のツリーを表示するビジュアライザからなります。決定木と選択式決定木にはクラシファイアが、回帰ツリーには回帰モデルが使用されます。

決定木 (Decision Tree)

決定木は予測モデルです。決定木では、従属または既知の属性値を使ってラベルまたは未知の属性を判別することで予測を行います。値（通常は「yes」や「no」などの文字列）や、少数の値しかとらない属性を予測することをクラス判別といいます。決定木クラシファイアでは、各レコードのラベルを予測することによってデータをクラシファイアします。分析に使用する基本構造は図 8-1 のような決定木です。決定木でクラシファイアが作成されると、ツリー・ビジュアライザがその構造を表示します。

ツリーの最初の要素は、すべてのデータを表すルートノードです。ツリーはここから 2 つまたはそれ以上の分岐に分割され、それぞれが特定の属性（項目）のデータを表し、異なる値を持ちます。たとえば、図 8-1 は cars データセットの決定木の可視化です。ここでは、レコードが立方インチでクラス判別されています。最初の分岐では、169.5 立方インチ以下が右に、169.5 より上が左に分岐しています。ツリーは、同じ属性を使って複数のノードで分岐できます。さらに、ノードは 3 つ以上に分岐できます。たとえば、mushroom データセットでは、レコードが食用と非食用にクラス判別され、最初のノードが 9 つの匂いに従って 9 つに分岐しています。

オブジェクトは、すべての（またはほとんどすべての）レコードが同じクラス（ラベル）を持つ最終分岐のノード（リーフノード）に達します。図 8-1 では、クラスは、決定木ウィンドウの最深部に並んでいる異なる範囲の立方インチ値です。

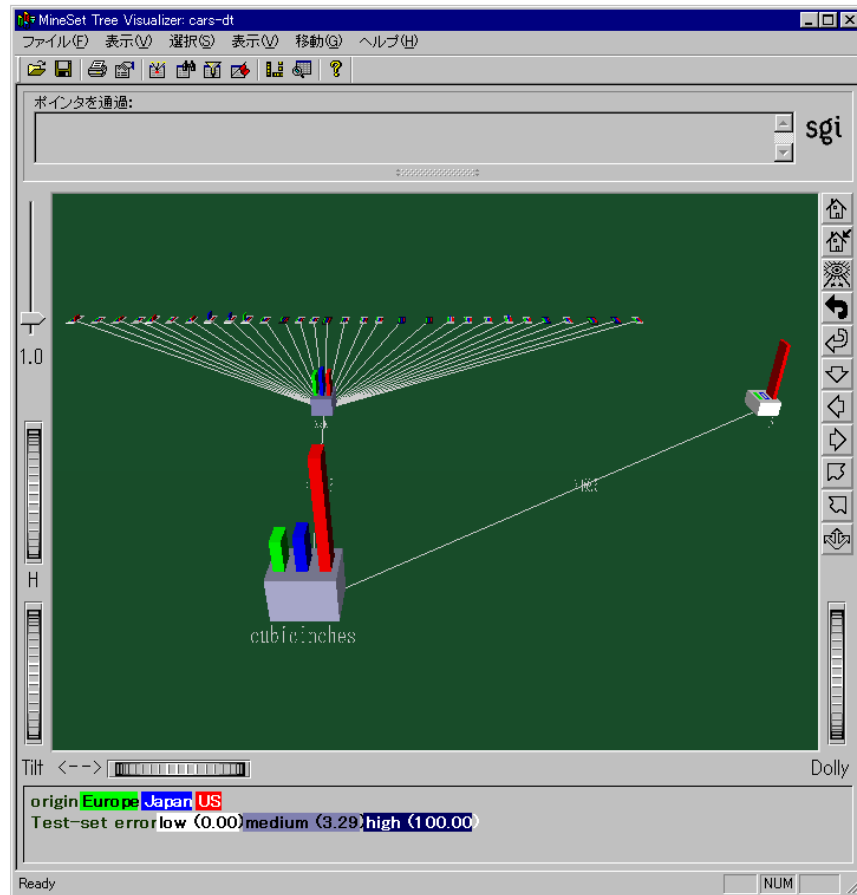


図 8-1 cars データセットの決定木モデルの可視化

選択式決定木 (Option Tree)

決定木モデルと同じように、選択式決定木も各レコードをクラスに割り当てます。前の項で説明したように、クラス判別に使用する基本構造は決定木です。選択式決定木分析でクラシファイアが作成されると、ツリー・ビジュアライザがその構造を表示します。選択式決定木は実際にはいくつかの決定木からなっています。選択式決定木では、ルートノードの分岐に使用する属性を1つだけ選択する代わりにいくつかを選択し、それぞれに対して決定木を作成します。

図 8-2 に cars データセットの選択式決定木を示します。この目的は、自動車の原産国（米国、日本、ヨーロッパ）を予測することです。選択式決定木は、選択枝ノードを追加することによって決定木クラシファイアを拡張したものです。選択枝ノードには、ツリーの決定ノードで選択できるいくつかの選択枝があります。たとえば、図 8-2 では、ルートが選択枝ノードであり、その下位に立方インチ、シリンダ数、重量（ポンド）、燃費、車種の 5 つの選択枝があります。

選択枝ノードには次の目的があります。

- いくつかの選択枝を示すことによって、階級ラベルの判別に影響を与える要素の範囲を広げます。つまり、ノードで 1 つの属性を示す代わりに、選択枝ノードではいくつかの選択枝を示します。ツリーを調べるときには、理解し易いと思われる選択枝や経験から判断して予測に適していると思われる選択枝、推定誤差に基づいて選択した選択枝などを選択できます。

図 8-2 の cars データセットでは、たとえば、*cylinders* や *weightlbs* サブツリーを選択します。前者は値の数が少く、後者は推定誤差が低い (1.53) からです。推定誤差はあくまで推定です。一般に、2 つの選択枝の誤差の差が平均標準偏差の 2 倍未満であれば、統計的に両者の誤差に差異はありません。

- 選択枝のサブツリーの投票を平均化することによって、不正確なクラシファイアを作成する危険を少なくします。各選択枝には、「専門家」ともいえるサブツリーが作成されます。選択枝ノードでは「専門家」の投票が平均化されるため、作成されるクラシファイアの誤差率が改善します。

図 8-2 に示す cars データセットでは、ルートノードの推定誤差率は 0.76% です。これはどの子ノードの誤差率よりも低い値です。ときには、明らかな選択枝が必ずしも最善の選択枝ではないことがあります。たとえば、原産国の予測には車種が当然の選択枝と思われる場合でも、訓練事例にすべての車種が含まれていないことがあります。実際、この例ではすべてが含まれているわけではありません。未知の車種があると、決定木はこれを多数派のクラス（米国）とみなし、2 つの間違いを犯します。しかし、他の選択枝があれば、すべての選択枝が平均化されることで誤差は減少します。

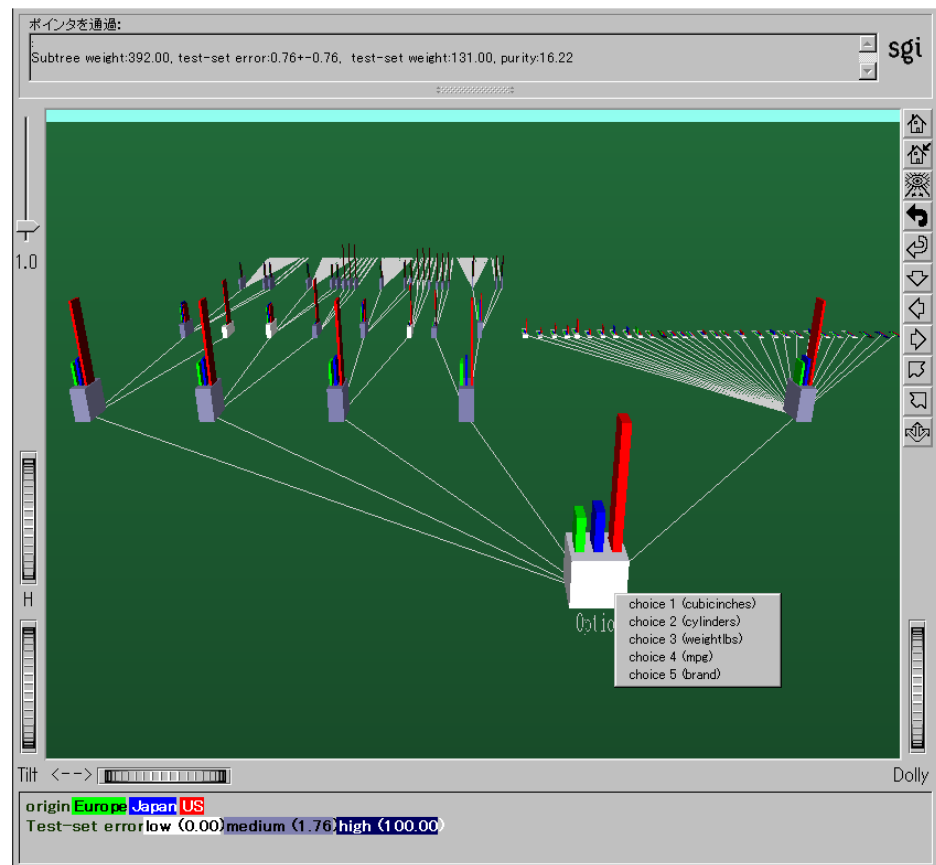


図 8-2 cars データセットの選択式決定木モデルの可視化

しかし、選択式決定木には2つの欠点があります。

- デフォルト設定では、選択式決定木を作成する時間が決定木の作成に必要な時間の約10から15倍になります。
- 作成されるツリー・ビジュアライザ・ファイルが非常に大きなものになります。ノード数が通常の決定木の10から15倍になります。

データセットの選択式決定木分析を行って、包含性や誤差率の利点が分析時間の長さを正当化できるか判断してください。決定木を作成するときどの属性を削除または使用すべきかに関し、新たな情報を得られる場合があります。

回帰ツリー (Regression Tree)

回帰では、一連の記述属性に基づいて連続型ラベルの値を予測します。回帰はクラス判別と似ていますが、クラス判別では、予測されたラベルは少数の離散値をとります。たとえば、iris は、種類によって iris-setosa、iris-versicolor、iris-virginica にクラス判別されます。回帰では、予測されるラベルは連続した範囲内のどの値でもかまいません。たとえば、年収はゼロより大きい任意の値を取ることができます。

回帰モデルが生成されると、選択式決定木ビジュアライザがツリーを表示します。この可視化では、回帰の意味や予測方法が分かるだけでなく、データ自体に対する貴重な洞察が得られます。回帰モデルを作成したら、それを使ってまだラベルになっていないレコードのラベルの値を予測できます。

回帰に使用する基本構造は、[図 8-3](#) のような回帰ツリーです。

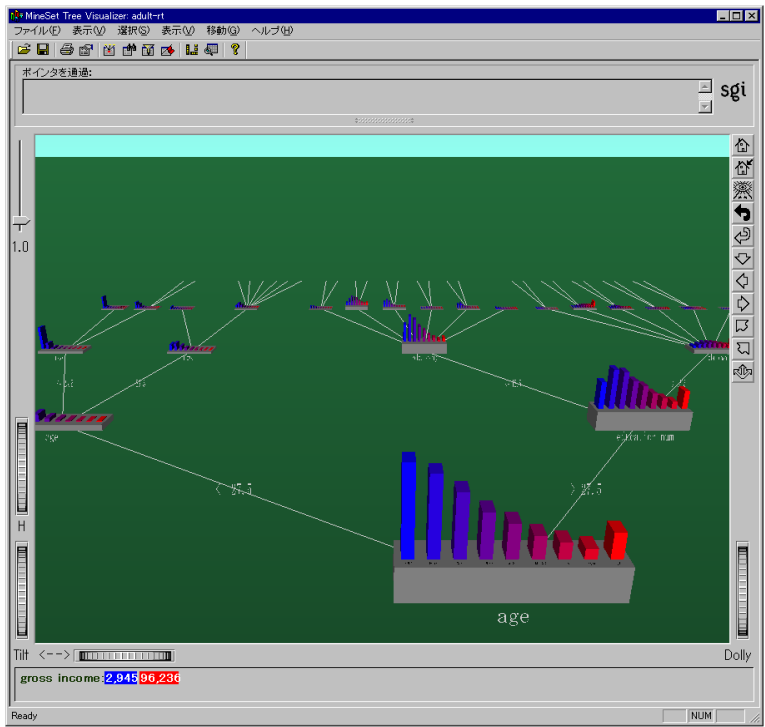


図 8-3 adult データセットの回帰ツリーの可視化

決定木、選択式決定木、回帰ツリーの起動

これらツリーツールの起動は、Tool Manager から行うのが最も簡単です。

1. Tool Manager の「ファイル」メニューからサーバに接続し、ログインします。同じメニューから「新しいデータファイルを開く」を選択し、必要なファイル名を選択または入力します。
2. 「データの可視化／マイニング」パネルで「マイニングツール」([図 8-4](#))をクリックし、その下に並んでいるタブから「クラス判別」か「回帰ツリー」をクリックします。
 - 決定木と選択式決定木の場合は、「クラス判別」タブをクリックします。
 - 回帰ツリーの場合は、「回帰ツリー」タブをクリックします。
3. 「モード」ポップアップ・メニューからモードを選択します。有効な4つのモードの詳細は、『*MineSet 3.0 Enterprise Edition Reference Guide*』の誤差推定の説明を参照してください。
4. 「分析」ポップアップ・メニューから分析ツールを選択します。
5. 「離散型ラベル」ポップアップ・メニューから、ラベルに使用する属性（項目）を選択します。ただし、回帰ツリーの場合は、「連続型ラベル」ポップアップ・メニューから選択します。たとえば、*mushroom* データセットの場合は *edibility* を、*churn* データセットの場合は *churned* を選択します。

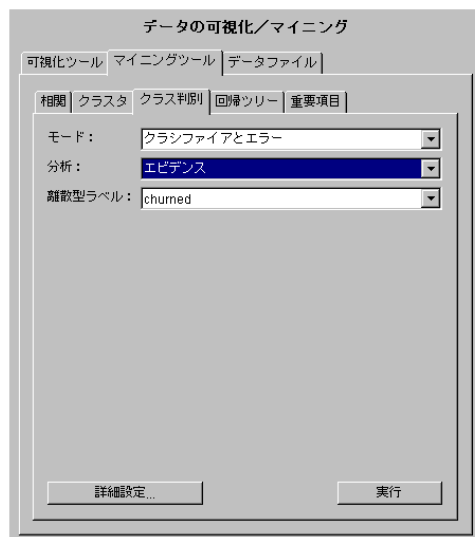


図 8-4 Tool Manager の「データの可視化／マイニング」パネルと「クラス判別」タブ

6. 「実行」をクリックして分析ツールを起動します。「*Tool Manager*」ウィンドウ下部の「ステータス」パネルに進行状況と統計結果が表示されます。計算が終了すると、ビジュアライザが表示されます。

決定木ビジュアライザによる結果の表示

決定木と選択式決定木の可視化は似ています。両方とも、決定ノードとリーフノードがエッジ（ライン）で結ばれます。

決定ノードのラベルは、そのノードでテストされる属性を表しています。属性のテストに使用する値（または値の範囲）はエッジに表示されます。属性の各値は 1 つのエッジに対応します。たとえば、図 8-1 の決定木モデルのルートでは属性 *weightlbs* をテストしています。ルートから出ている 2 つのエッジは、属性の値の範囲（3018 以下と 3018 より大）を示しています。したがって、すべての値は右か左の分岐に入ります。

決定木のリーフノードはクラスを表しています。図 8-1 でルートから右の分岐をたどり (...16.5] というラベルの付いたリーフに行くと、レコードが *weightlbs* > 3018 ポンド、*horsepower* > 141、*cubicinches* > 311.5、*mpg* ≤ 16.5 マイルで分析されていることが分かります。

ベースの色はそのサブツリーの誤差推定を表しています。藍色は大きい誤差を、グレイは中程度の誤差を、白は小さい誤差をそれぞれ表しています。テストセットのレコードがノードに到達しない場合、つまり、誤差推定が存在しない場合、ベースの色は黒です。

ノード上の垂直バーは、そのノードにおけるクラスの分散を表しています。ノードのベースには高さと色があります。高さは、このノードに到達した訓練事例レコードの重みに対応しています。重みが設定されていない場合は、レコード数に対応しています。一般には、この重みが高ければ、ノードにおけるクラス分散の信頼度はそれだけ高くなります。レコードの重み付けについては、[第 3 章の「レコードの重み付け」](#)の項を参照してください。

マウスのポインタをノード上に移動すると、次の情報が表示されます。

- サブツリー重み — ポインタが指しているノードより下のサブツリーに属する訓練事例のレコードの重み。この値はベースの高さにマッピングされます。
- テストセットの誤差 / 損失 — サブツリーの誤差推定。ただし、損失マトリックスの場合は、損失推定です。 $\pm$ の後の数字はその推定値の標準偏差です。標準偏差が高ければ高いほど、誤差推定は不正確です。レコードがほとんどないリーフの場合や、テストセットの誤差が 0% または 100% に近い場合には、誤差推定 / 損失や標準偏差の信頼度は減少します。
- テストセットの重み — このノードに達したテストセットのレコードの重み。ただし、重みが設定されていない場合は、レコード数です。
- ピュリティ (Purity) — そのノードにおけるラベルの値の分布の歪みを表す 0 から 100 までの数字。歪みについては、用語集の「歪み (Skewness)」を参照してください。ノードのレコードがすべて同じクラスに属する場合、ピュリティは 0 です。すべてのラベルの値が同じ重みを持つ場合、ピュリティは 100 です。ピュリティは、バックフィッティングの後で計算されます。バックフィッティングについては、用語集の「バックフィッティング (Backfitting)」を参照してください。

テストセットの誤差 / 損失や重みを生成できるのは「クラス判別」と「エラー」モードだけです。「クラシファイアのテスト」オプションを使えば、既存のクラシファイアとテストセットに基づいて可視化できます。

決定木のメインウィンドウを使ったレコードのクラス判別

ラベルが分からないレコードをクラス判別するには、ツリーのルートから始め、そのレコードの属性値に従って分岐をたどります。レコードの属性値に従って適切なエッジをたどれば、リーフノードに達します。リーフノードに対応付けられたラベルまたはクラスが、予測されたレコードのクラス判別です。

決定までには、すぐに結論が出るため短いパスしか必要ない場合と、長いパスを必要とする場合があります。一般に、リーフは、ルートからそのリーフに至るすべての決定ノードで行われたすべてのテストとエッジ上のすべての値との結合である 1 つの規則に対応しています。

iris データセット用に作成した決定木（[図 8-5](#)）は非常に単純ですが、説明用としては十分です。このファイルを開くには、Tool Manager の「可視化ツール」->「3D ビジュアライザ」->「`examples\iris-dt.treeviz`」を選択します。このツリーのルートでは、誤差率が 6%、標準偏差が 3.39% です。標準偏差が大きいのは、このファイルが小さく、テストセットに 50 レコードしかないためです。プュリティは 0.0 で、分割が一樣であることを示しています。

ルートの左の子ノードは、テストセットの誤差が 0、プュリティが 100 です。これは、`petal_length` が 2.6 インチ以下のすべてのレコードが *iris-sotosa* クラスに属しているためです。したがって、*iris-sotosa* の予測は、おそらく `petal_length` が 2.6 インチ以下のすべてのレコードに対しては非常に正確です。ルートの右の子ノードは誤差推定が 8.57% です。`petal_length` が 2.6 インチより大きいレコードを含むこの子ノードには、*iris-sotosa* クラスに属するレコードはありません。したがって、このクラスはおそらく *iris-versicolor* か *iris-virginica* です。このノードには 2 つの可能性しかないため、プュリティはルートより高くなります (36.91)。

決定木のリーフノードでは、データは同じクラス判別規則（リーフまでのパス）を持つクラスタに分割されます。リーフノードを見れば、同じ特性を持つクラスタを知ることができます。

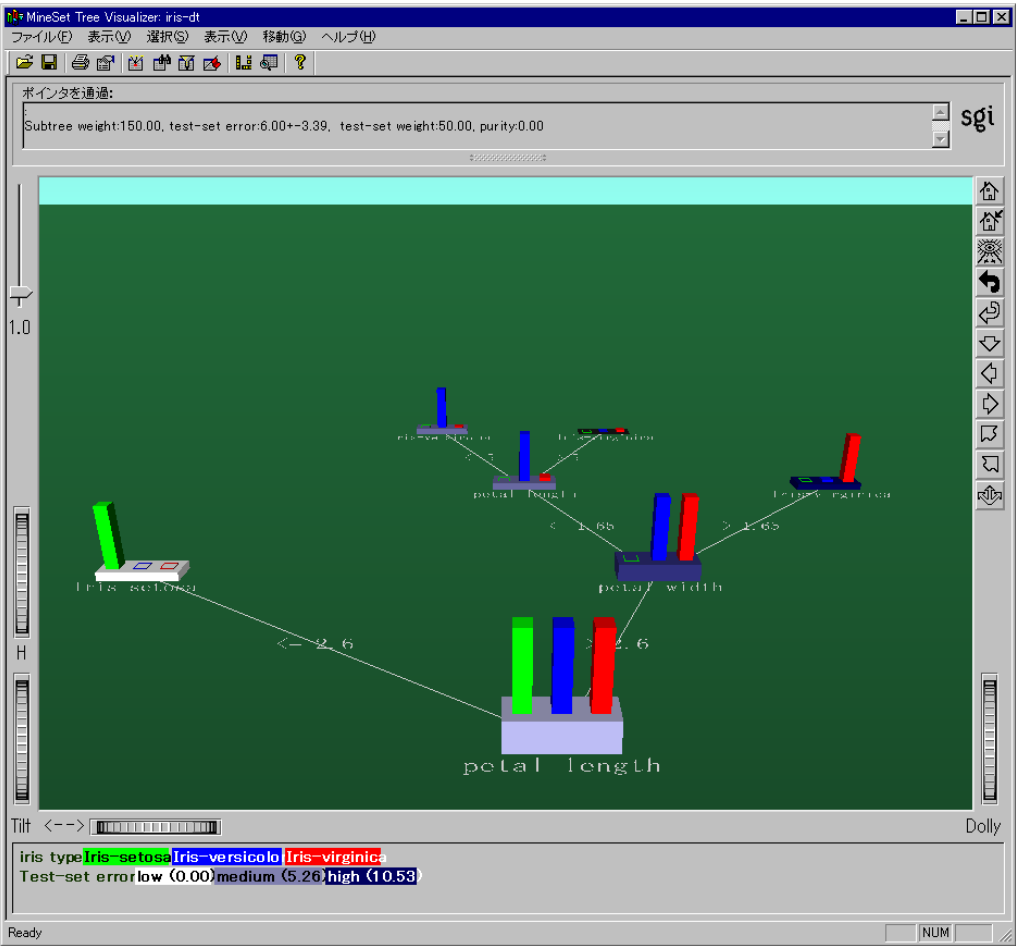


図 8-5 iris データセットの決定木

その他の便利なオプション

決定木の可視化には、ツリー・ビジュアライザを使って表示を行います。ツリー・ビジュアライザには、Tool Manager の「表示」メニューからアクセスする検索パネル、フィルタパネル、概要ウィンドウなど、いくつかの便利な機能があります。これらの機能については、第 5 章「ツリー・ビジュアライザによるデータの表示」や『*MineSet 3.0 Enterprise Edition Reference Guide*』のツリー・ビジュアライザの説明を参照してください。

さらに、決定木には、生成アルゴリズム、枝刈り、評価基準などを調整するためのさまざまなオプションがあります。これについては、『*MineSet 3.0 Enterprise Edition Reference Guide*』の決定木の説明を参照してください。

選択式ツリー・ビジュアライザによる結果の表示

決定木と選択式決定木の可視化は本質的には似ていますが、相違点がいくつかあります。

- 選択式決定木の左端の選択枝は、決定木であれば唯一の選択枝です。右に行くに従い、選択枝は適合性のスコアに従って降順にランクされます。適合性スコアは、表示されているテストセットの誤差とは必ずしも一致しません。これは、分析ツールが使用するスコア関数が完全ではないためです。さらに、当然ながら、テストセットによる推定はさまざまです。テストセットが大きければ大きいほど、推定は正確になります。
- 選択枝ノードの誤差率は、そのいずれの子ノードの誤差率とも異なることがあります。これは、選択枝ノードでは子ノードの予測が平均化されるためです。場合によっては、選択枝ノードの誤差はどの子ノードの誤差よりも小さくなります。これは平均化が役立つことを証明しています。
- バーで示されるインスタンスの分割は、選択枝ノードのどの子ノードでも選択枝ノードのものとまったく同じです。これは、選択枝ノードでは決定が行われなかったためです。それらの選択枝は子ノードとして表示されます。
- 選択式決定木の可視化には、ツリーの中を移動するのに非常に便利な機能があります。選択枝ノードの上でマウスの右ボタンをクリックすると、子ノードのリストが表示されます。これらの子ノードは選択枝を表しています。

単純ですが、見る価値のある選択式決定木が *MineSet 3.0\examples\cars-odt.treeviz* にあります。選択式決定木の可視化には、ツリー・ビジュアライザを使って表示を行います。ツリー・ビジュアライザには、検索パネル、フィルタパネル、概要ウィンドウなど、

いくつかの便利な機能があります。これらの機能については、第 5 章「ツリー・ビジュアライザによるデータの表示」を参照してください。

さらに、選択式決定木には、生成アルゴリズム、枝刈り、評価基準などを調整するためのさまざまなオプションがあります。これについては、『*MineSet 3.0 Enterprise Edition Reference Guide*』の「選択式決定木」についての説明を参照してください。

回帰ツリー・ビジュアライザによる結果の表示

回帰ツリーと決定木の可視化は似ています。両方とも、決定ノードとリーフノードがエッジで結ばれます。

決定ノードには、そのノードでテストする属性が指定されています。属性のテストに使用する値（または値の範囲）はエッジに表示されます。属性の各値は 1 つのエッジに対応します。たとえば、図 8-3 の回帰ツリーのルートでは属性 age をテストしています。ルートから出ている 2 つのエッジは、属性の値の範囲（27.5 より小と 27.5 以上）を示しています。したがって、すべての値は右か左の分岐に入ります。値が未知であり、かつ疑問符のラベルが付いたエッジがない場合は、現在のノードの平均または中央のラベルの値が予測値になります。

回帰ツリーのリーフノードでは値が予測されます。有益な例が `\examples\adult-rt.treeviz` にあります。図 8-3 でルートから左端の分岐をたどり 4002.300 というラベルの付いたリーフに行くと、週に 35.5 時間未満働く 19.5 歳未満の人は平均で総収入が \$4002.30 であると予測されていることが分かります。

ノードの各バーは、連続型ラベルの値の各小範囲に対応します。連続型ラベルの値の範囲はノードによって異なります。ノードのバーは、レコードの重み（数）がこの範囲でどのように分布しているかを示すヒストグラムになっています。バーの数は、ルートにおけるレコードの重みによって決まります。左端のバーは常に最も小さい値に対応します。したがって、サイズと中間点はノードごとに異なります。バーの色は、バーが表す小範囲の中間点を表します。最大範囲は左端の青から右端の赤で表されます。限られた範囲のラベルの値だけを持つレコードのノードでは、ヒストグラムは青から赤までの色を持っていません。

ノードのベースには高さがあります。高さは、このノードに到達した訓練事例のレコードの重みに対応しています。重みが設定されていない場合は、レコード数に対応しています。一般には、この重みが高ければ、ノードにおける分割の信頼度はそれだけ高くなります。

マウスのポインタをノード上に移動すると、次の情報が表示されます。用語の定義については、用語集を参照してください。

- サブツリー重み — ポインタが指しているノードより下のサブツリーに属する訓練事例のレコードの重み。この値はベースの高さにマッピングされます。
- 平均値 — 連続型ラベルの平均値
- 標準偏差 — 連続型ラベルの標準偏差。標準偏差が高ければ、モデルの信頼度はそれだけ低くなります。
- 中央値 — 連続型ラベルの中央値。
- 絶対偏差 — 連続型ラベルの絶対偏差。絶対偏差が高ければ、モデルの信頼度はそれだけ低くなります。

回帰ツリーのメインウィンドウを使った値の予測

レコードの値を予測するには、ツリーのルートから始め、そのレコードの属性値に従って分岐をたどります。レコードの属性値に従って適切なエッジをたどれば、リーフノードに達します。リーフノードに対応付けられた予測値が、予測されたレコードの値です。

回帰ツリーにおける誤差推定

クラシファイアの評価では、評価基準は当然ながら誤差です。誤差とは、クラシファイアが間違ったラベルを予測するサンプルの数です。損失マトリックスが指定されている場合は、不正なクラス判別誤差の種類によって、対応する費用が異なることがあります。このような場合は、損失が当然の評価基準です。詳細は、[第 11 章の「損失マトリックスの定義」](#)の項を参照してください。

実際値を予測する回帰ツリーでは、明確な評価基準はありません。よく使われる基準は標準偏差と絶対偏差です。標準偏差は平均平方誤差の平方根です。絶対偏差は、予測されたラベルの値と実際のラベルの値との差異の絶対値を平均化したものです。

その他の便利なオプション

回帰ツリー・ビジュアライザはツリー・ビジュアライザを使って表示を行います。ツリー・ビジュアライザには、検索パネル、フィルタパネル、概要ウィンドウなど、いくつかの便利な機能があります。これらの機能については、[第 5 章「ツリー・ビジュアライザによるデータの表示」](#)や『*MineSet 3.0 Enterprise Edition Reference Guide*』のツリー・ビジュアライザの説明を参照してください。

さらに、回帰ツリーには、生成アルゴリズム、枝刈り、評価基準などを調整するためのさまざまなオプションがあります。これについては、『*MineSet 3.0 Enterprise Edition Reference Guide*』の「回帰ツリー」の説明を参照してください。

決定木、選択式決定木、回帰ツリーによる予測

この章では、既知の結果を持つデータのクラス判別について説明しました。既知の結果とは、たとえば、解約するか解約しないか、食用か有毒かなどです。しかし、決定木、選択式決定木、回帰ツリーは、未知または将来の結果を予測するときに最も役立ちます。つまり、クラスが分かっているデータに基づいてモデルを作成し、そのモデルを使ってクラスが分かっていない新しいデータをクラス判別できます。モデルの作成と適用については、[第 7 章「予測モデリングの理解」](#)と[第 11 章「予測モデリングの改良」](#)を参照してください。

デシジョン・テーブル・クラシファイアとビジュアライザによるモデリングと予測

この章では、デシジョン・テーブル (Decision Table) の特徴と機能について説明します。この章の構成は次のとおりです。

- 「デシジョン・テーブル・クラシファイアの概要」(165 ページ)
- 「デシジョン・テーブルの起動」(167 ページ)
- 「デシジョン・テーブル・ビジュアライザによる結果の表示」(169 ページ)
- 「デシジョン・テーブルによる予測」(174 ページ)

デシジョン・テーブルの特徴と機能を示すために、設定ファイルとデータファイルのサンプルが用意されています。これらのファイルは、*MineSet* を最初にインストールした *MineSet 3.0\examples* ディレクトリにあります。

デシジョン・テーブル・クラシファイアの概要

デシジョン・テーブルとは、クラス判別を行う予測モデリングツールです。クラシファイアと予測モデリングについては、[第 1 章「データマイニングと MineSet ツールの概要」](#)と[第 7 章「予測モデリングの理解」](#)を参照してください。分析アルゴリズムには、分析ツールとビジュアライザが統合されています。分析ツールはデシジョン・テーブル・モデルを生成するアルゴリズムです。エビデンス・モデルとは異なり、デシジョン・テーブル・モデルでは、属性は独立しているとはみなされません。

デシジョン・テーブルはデータを階層的に分類したもので、階層の各レベルには 2 つの属性があります。デシジョン・テーブル分析がデータのクラス判別に最も重要な属性 (項目) を識別すると、そのビジュアライザが生成されたモデルを一連のケーキグラフとして表示します。可視化される各ケーキは、次の最も重要な属性ペアを表すケーキに小さく分割できます。可視化では、重要性が次第に減っていく属性をいくつかのレベルで表すことができます。[図 9-1](#) に、*mushroom* データセットのデシジョン・テーブルを可視化した最上位レベルを示します。このレベルでは、キノコが食用に適しているかどうかを判断する最も重要な 2 つの属性は匂い (odor) と孢子プリント色 (spore print color) です。

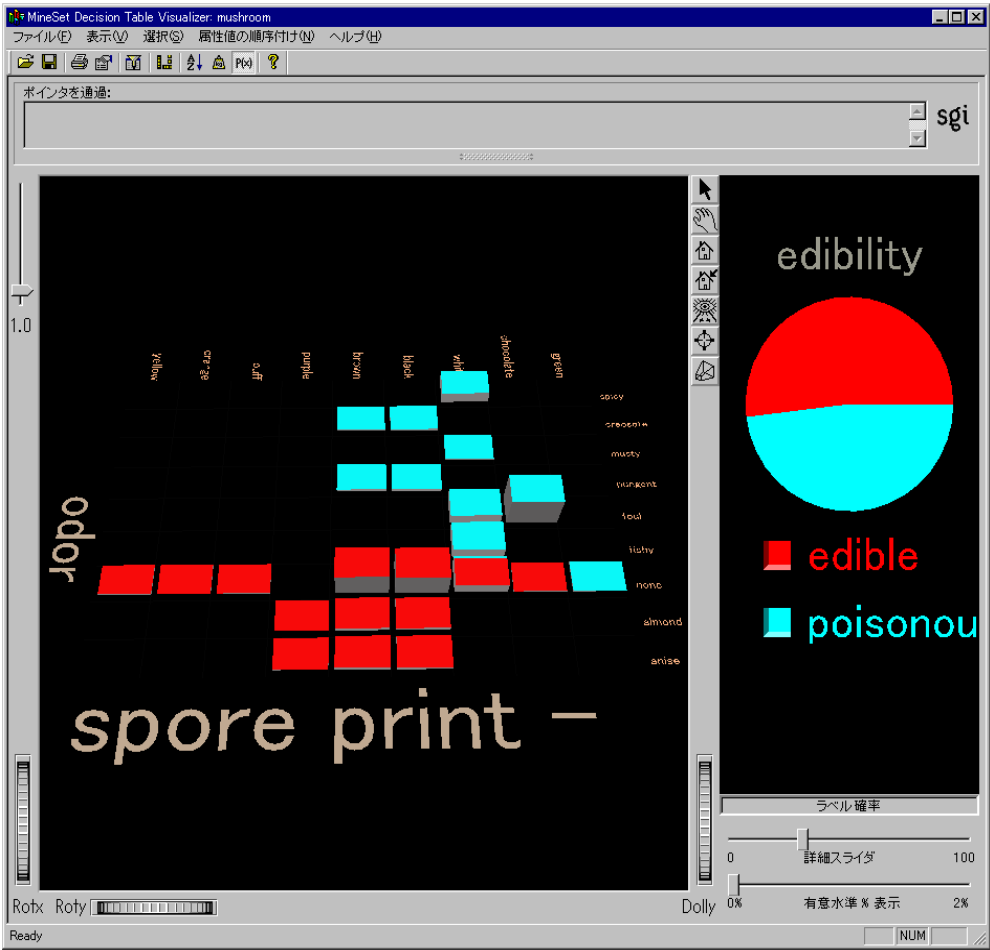


図 9-1 mushroom データセットのデシジョン・テーブル

デシジョン・テーブルの起動

デシジョン・テーブル・クラシファイアの作成は、Tool Manager から行うのが最も簡単です。その他のオプションについては、『*MineSet 3.0 Enterprise Edition Reference Guide*』を参照してください。

1. Tool Manager の「ファイル」メニューからサーバに接続し、ログインします。
2. 同じメニューから「新しいデータファイルを開く」を選択し、必要なファイル名を選択または入力します。
3. 「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（図 9-2）をクリックし、その下に並んでいるタブから「クラス判別」を選択します。

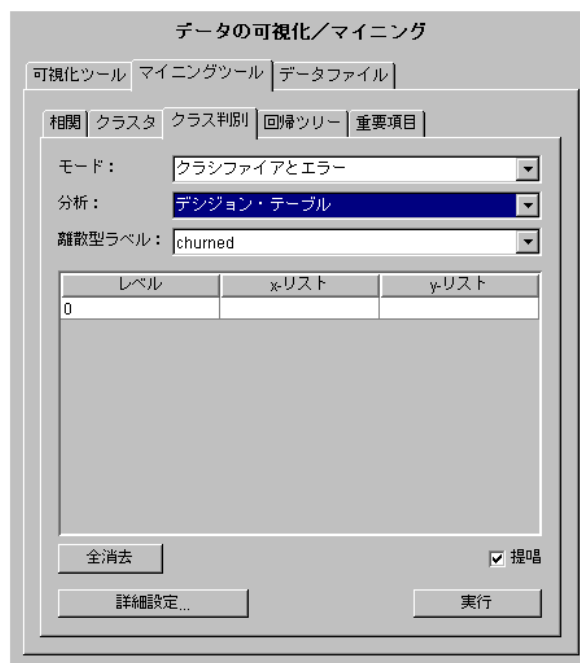


図 9-2 「クラス判別」タブを表示した Tool Manager の「データの可視化／マイニング」パネル

4. 「モード」ポップアップ・メニューからモードを選択します。4 種類のモードについては、[第 7 章「予測モデリングの理解」](#)を参照してください。
5. 「分析」ポップアップ・メニューから「デシジョン・テーブル」を選択します。
6. 「離散型ラベル」ポップアップ・メニューから、離散型ラベルとして使用する属性を選択します。
7. 「x- リスト」と「y- リスト」プルダウン・メニュー ([図 9-2](#)を参照) に、表示する属性を指定します。属性の指定には 3 つの方法があります。
 - 「提唱」ボックスをオンにして、属性の選択を MineSet に任せる。MineSet がクラスの予測に最も有益な属性を識別し、表示します。
 - 「データの可視化／マイニング」パネルのいろいろなプルダウン・メニューから属性を選択する。どこから始めればよいかを知りたい場合は、「重要項目」を使って最も重要な属性を検索してください。[第 3 章の「重要項目の検索」](#)の項を参照してください。
 - 一部の属性を選択し、残りを MineSet に任せる。それには、上のようにして属性を選択してから「提唱」ボックスをオンにします。残りの適切な項目を MineSet が選択します。
8. 「実行」をクリックして分析ツールを起動します。

「Tool Manager」ウィンドウ下部の「ステータス」パネルに進行状況と結果の統計が表示されます。属性自動提唱の処理を中断するには、進捗ダイアログの「取消」か「viz ツールの表示」をクリックします。「可視化ツールの表示」をクリックすると、現在のサーバ処理が停止され、それまでにマッピングされた項目のデシジョン・テーブルが中間結果から作成されます。

注記: デシジョン・テーブルのその他のオプションについては、『*MineSet 3.0 Enterprise Edition Reference Guide*』の「デシジョン・テーブル」の説明を参照してください。

デシジョン・テーブル・ビジュアライザによる結果の表示

デシジョン・テーブル・ビジュアライザでは、「デシジョン・テーブル」パネルと「ラベル確率」パネルが左と右に表示されます。図 9-3 に *mushroom* データセットのデシジョン・テーブルの可視化の例を示します。ここには、最初の 2 つの詳細レベルが表示されています。

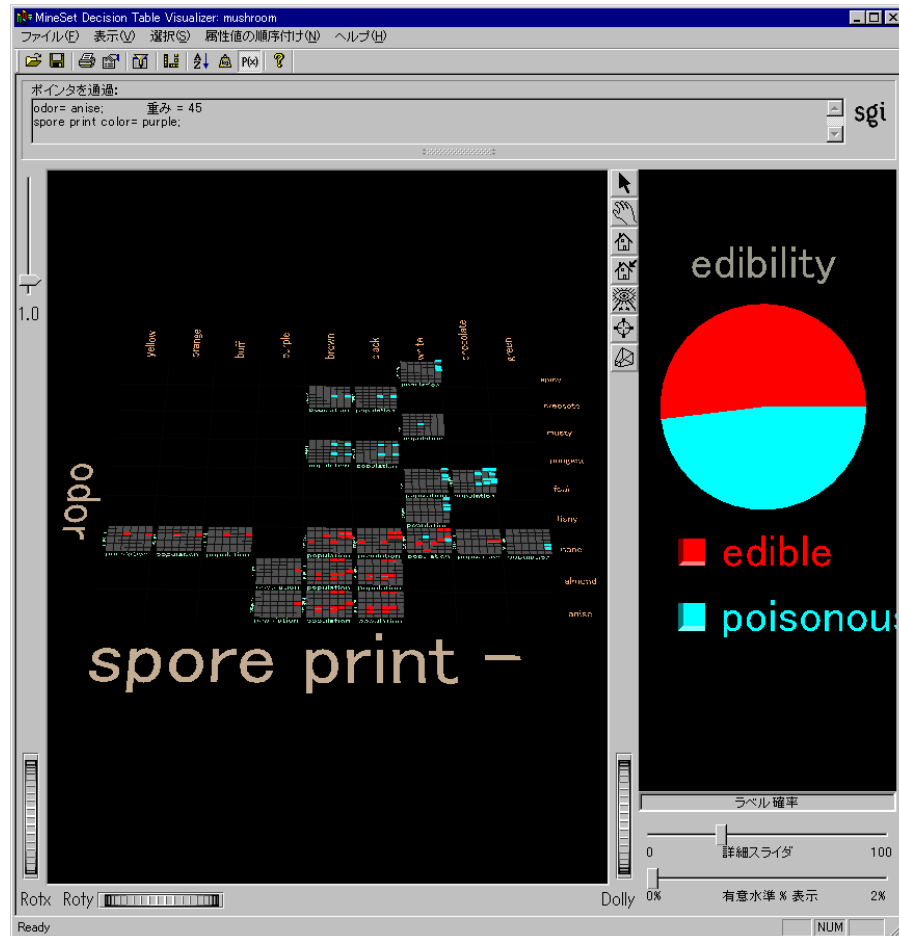


図 9-3 mushroom データセットのデシジョン・テーブルの可視化

「デシジョン・テーブル」パネルの説明

左の「デシジョン・テーブル」パネルには、四角形のケーキグラフが表示されます。色付きの各スライス、一定の属性値を持つレコードのラベル確率を表しています。ラベル確率とは、指定された属性に対し一定の値を持つレコードが一定のクラスに属する可能性を表したものです。たとえば、図 9-3 は、*mushroom* データセットに対しデシジョン・テーブルを実行したものです。このグラフは、胞子プリント色が白で魚の匂いがするキノコが有毒である確率は 100% であることを示しています。しかし、胞子プリント色が白でも匂いがなければ、有毒である確率は 7.69%、食べられる確率は 92.31% です。これらの数字を見るには、左のパネルでケーキを選択し、ラベル（食用または有毒）の横の色付きボックスの上にマウスのポインタを移動します。メニューバーとメインウィンドウの間に百分率が表示されます。

「デシジョン・テーブル (Decision Table)」パネルの要素をマウスの右ボタンでクリックすると、この要素が小さいケーキグラフにさらに分割されます。この処理をドリルダウンといいます。ケーキグラフをさらに細かく調べるには、次の手順に従います。

- 現在の詳細レベルにある 2 つの属性の値を表示するには、必要なケーキグラフの上にマウスポインタを移動します（選択モード）。属性の値とレコードの重みがメニューバーとメインウィンドウの間に表示されます。ケーキグラフの高さは重みに比例します。用語集の「重み (Weight)」を参照してください。
- 次の詳細レベルにドリルダウンするには、必要なケーキグラフの上にマウスポインタを移動して右ボタンをクリックするか、背景をクリックしてすべてのケーキを全体的にドリルダウンします。図 9-5 に、*mushroom* データセットの細分化されたケーキの拡大表示を示します。
- レベルを繰り返しドリルアップするには、<Ctrl> キーを押したままマウスの右ボタンまたは中ボタンをクリックします。一部の領域をドリルアップすることもできますし、背景で <Ctrl> キーを押したままマウスの右ボタンまたは中ボタンをクリックすれば、全体をドリルアップすることもできます。
- 1 つ上の階層レベルにあるケーキの 2 つの属性値とレコードの重みを見るには、マウスポインタをベースの上に移動します。ベースは、ケーキの下にあるグレイのブロックです。
- 特定のケーキグラフを定義する値を見るには、最上位のレベルにあるベースを強調表示してから順に詳細レベルを下げてゆき、目的のケーキグラフの真下にあるベースをポイントします。そのケーキグラフに対応する値のペアが、メニューバーの下選択パネルに表示されます。

それぞれの詳細レベルでは、一方の属性の名前がケーキグラフ配列の左に、その値が右にそれぞれ表示されます。そして、他方の属性の名前が配列の下に、その値が上にそれぞれ表示されます。属性の合計数が奇数の場合は、最下位のレベルに属性が1つだけ表示されます。

「ラベル確率」パネル

「*Tool Manager*」ウィンドウ右側の「ラベル確率」パネルには、データセット全体のラベル確率を示す円グラフが表示されます。すべてのクラスラベルが円グラフの下に表示されます。

ラベル確率をさらに細かく調べるには、次の手順に従います。

- 特定の属性セットのラベル確率を見るには、必要なケーキグラフをクリックします。「ラベル確率」パネルの円グラフに、そのケーキで表される属性セットを持つレコードのラベル確率が表示されます。
- 属性の組合せのラベル確率を見るには、必要なケーキグラフを<Ctrl>キーでクリックします。<Ctrl>キーのクリックでは、異なる詳細レベルから複数のケーキを選択できます。「ラベル確率」パネルに、選択されたレコードセットのラベル確率が表示されます。
- ラベル固有の百分率や信頼度レベルを見るには、必要なラベルの横にある色付きボックスの上にマウスポインタを移動します。それらの数字がメニューバーと「デシジョン・テーブル」パネルの間に表示されます。

デシジョン・テーブルの例

デシジョン・テーブルの使用例として食用キノコと有毒キノコを判別するとします。それには、まず *mushroom* データセットからデシジョン・テーブルを作成します (図 9-1 を参照)。左のパネルには、デシジョン・テーブルの最上位の詳細レベルが表示されます。この詳細レベルには、匂いと胞子プリント色が選択されています。これらは、「重要項目」アルゴリズムによって最も重要な2つの属性と判別されたものです。複数のクラスがあるのは最上位レベルのブロックだけです。この例では、これは「匂いがなく、胞子プリント色が白」のブロックです。グラフの上でマウスの右ボタンをクリックして次のレベルにドリルダウンすると、属性として生息環境 (habitat) と生息数 (population) が表示されます (図 9-4 を参照)。ドリルダウンしたこの領域を拡大したものを図 9-5 に示します。このレベルには混合ケーキはありません。したがって、胞子プリント色が白で匂いのないキノコが他のキノコとともに森に生息しているのを見つけたら、これはかなり高い確率で有毒です。

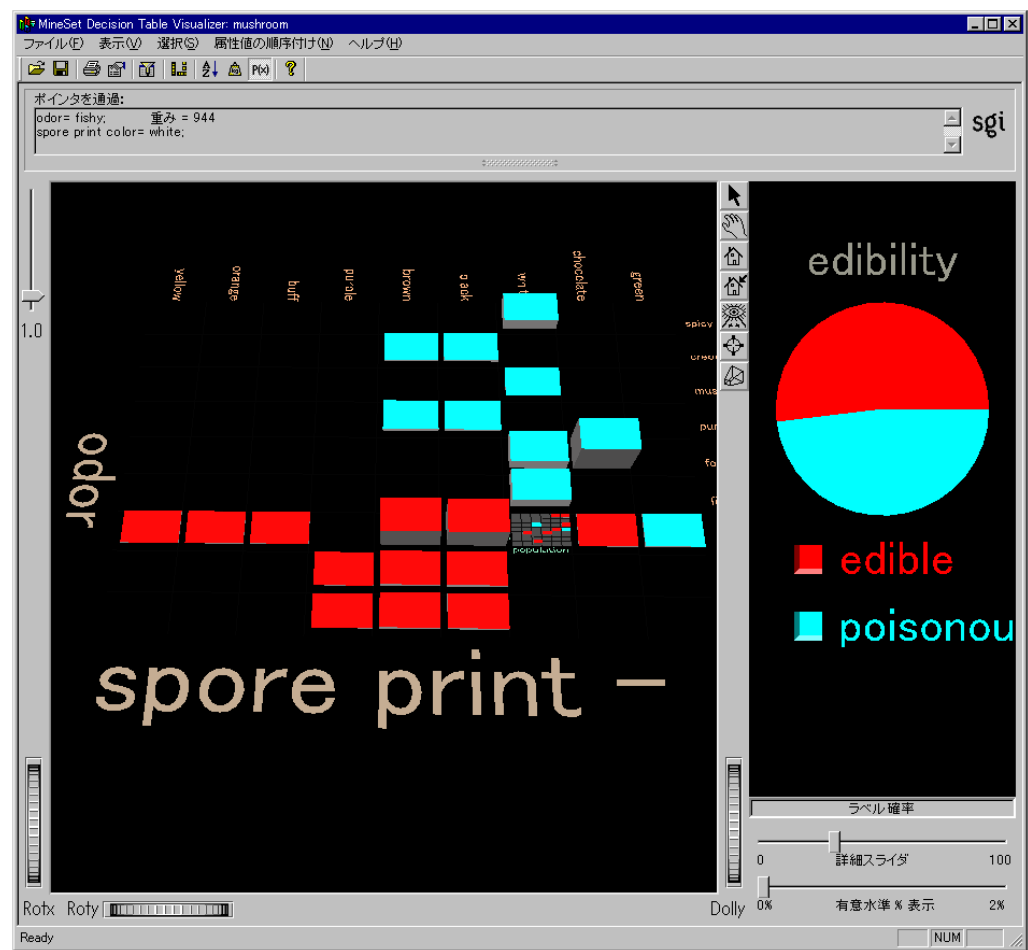


図 9-4 ドリルダウンした mushroom データセットのデシジョン・テーブル

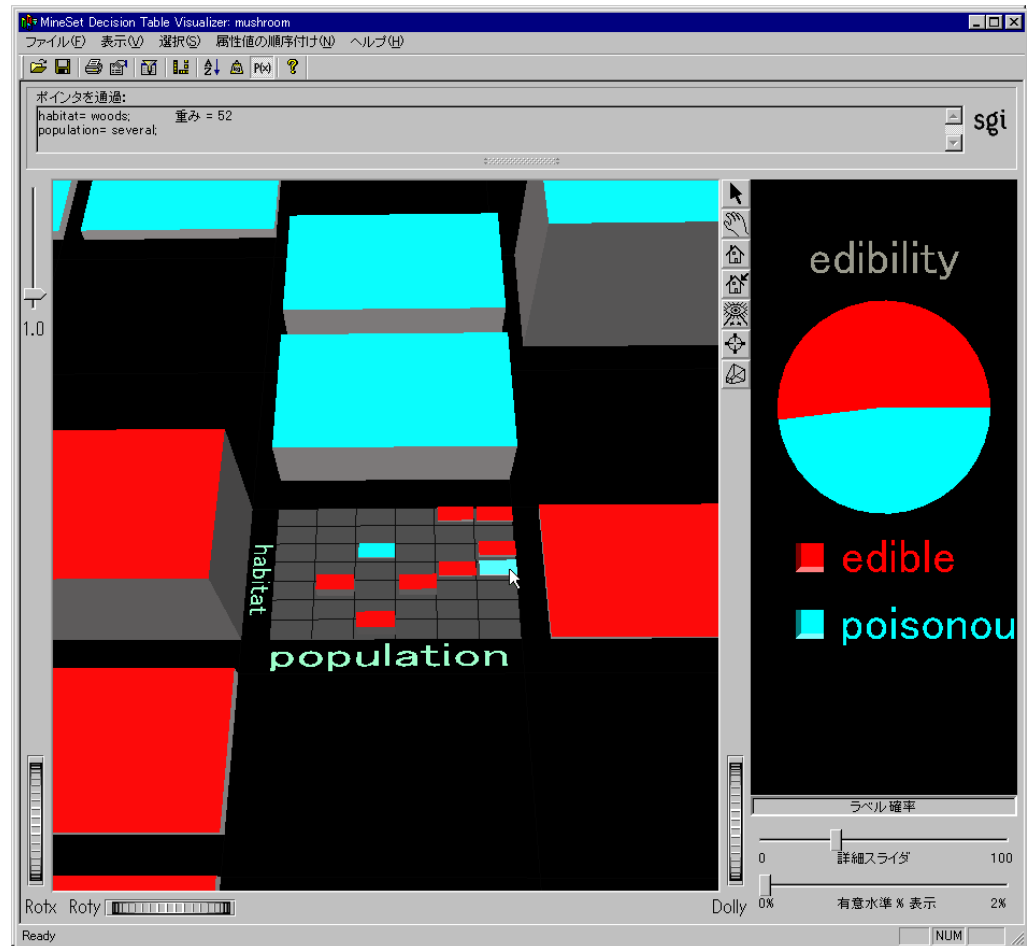


図 9-5 ドリルダウンしたデシジョン・テーブルの拡大

デシジョン・テーブルによる予測

この章では、既知の結果を持つデータのクラス判別について説明してきました。既知の結果とは、たとえば、食用か有毒かなどです。しかし、デシジョン・テーブルは、未知または将来の結果を予測するときに最も役立ちます。つまり、クラスが分かっているデータに基づいてモデルを作成し、そのモデルを使って、クラスが分かっている新しいデータをクラス判別できます。モデルの作成と適用については、[第 7 章「予測モデリングの理解」](#)と[第 11 章「予測モデリングの改良」](#)を参照してください。

メニューの詳細は、『*MineSet 3.0 Enterprise Edition Reference Guide*』のデシジョン・テーブルの説明を参照してください。

エビデンス・クラシファイアとビジュアライザによるモデリングと予測

この章では、エビデンス・クラシファイア (Evidence Classifier) とビジュアライザの特徴と機能について説明します。この章の構成は次のとおりです。

- 「エビデンス・クラシファイアとビジュアライザの概要」(175 ページ)
- 「エビデンス・ツールの起動」(178 ページ)
- 「エビデンス・ビジュアライザによる結果の表示」(179 ページ)
- 「エビデンス・クラシファイアによる予測」(187 ページ)

エビデンス・ビジュアライザの特徴と機能を示すために、設定ファイルとデータファイルのサンプルが用意されています。これらのファイルは、MineSet を最初にインストールした *MineSet 3.1\examples* ディレクトリにあります。

エビデンス・クラシファイアとビジュアライザの概要

エビデンスツールは、クラス判別を行う予測モデリングツールです。クラシファイアと予測モデリングについては、[第 1 章「データマイニングと MineSet ツールの概要」](#)と[第 7 章「予測モデリングの理解」](#)を参照してください。分析アルゴリズムには、分析ツールとビジュアライザが統合されています。分析ツールは Naive Bayes モデルを生成するアルゴリズムです。デシジョン・テーブル・モデルとは異なり、エビデンス・モデルでは、属性は独立しているものとみなされます。しかし、独立していなくても、妥当な結果が生成されます。

エビデンス・ビジュアライザは、クラシファイアからの情報を [図 10-1](#) のように表示します。これにより、クラス判別に対する特定の属性（項目）の重要性を知ることができます。

「エビデンス・ビジュアライザ」 ウィンドウ

クラシファイアが作成されると、結果が「エビデンス・ビジュアライザ (Evidence Visualizer)」 ウィンドウに表示されます。最初は、左のパネルに、クラシファイアが使用した属性のケーキグラフが並んで表示されます (図 10-1)。ケーキグラフは割合を示すという点で円グラフに似ていますが、長方形のスライスからなる四角形をしています。エビデンスを表すケーキグラフ、確率を表す円グラフ、選択されたラベルの情報を表示するバーグラフは、交互に切り替えることができます。詳細は、「エビデンス・ビジュアライザによる結果の表示」 (179 ページ) を参照してください。

図 10-1 にエビデンスの表示を、図 10-2 に確率の表示をそれぞれ示します。

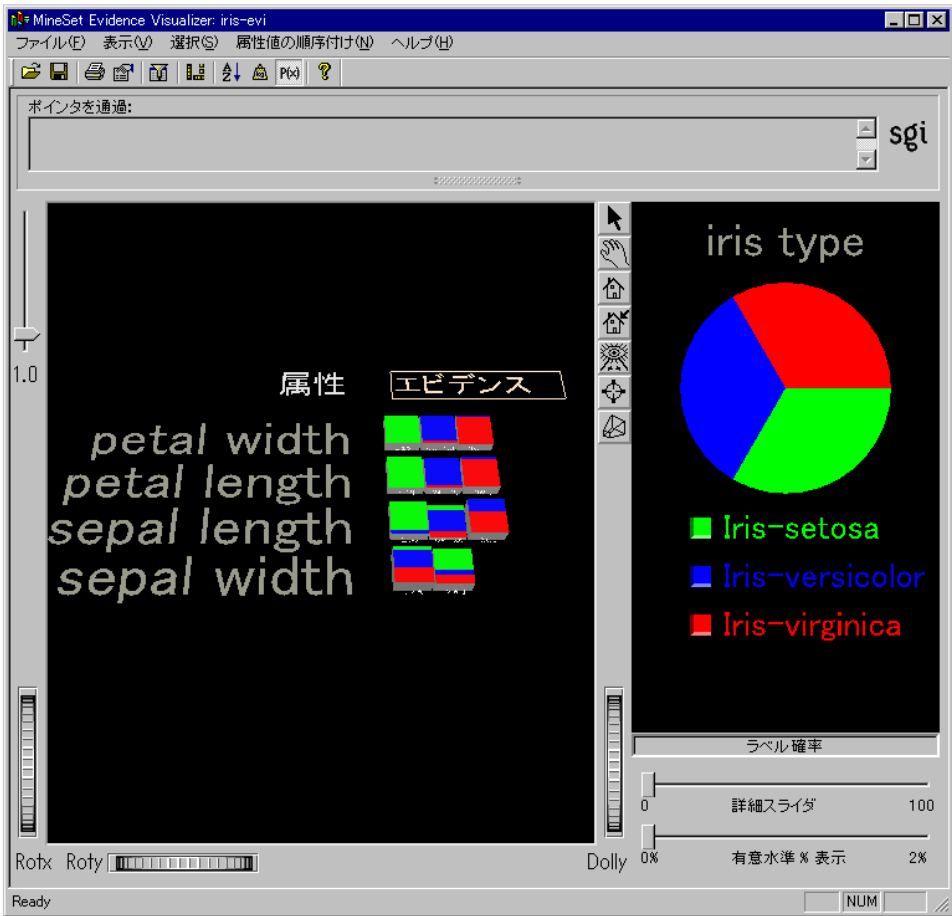


図 10-1 iris データセットを使用したエビデンス・ビジュアライザ

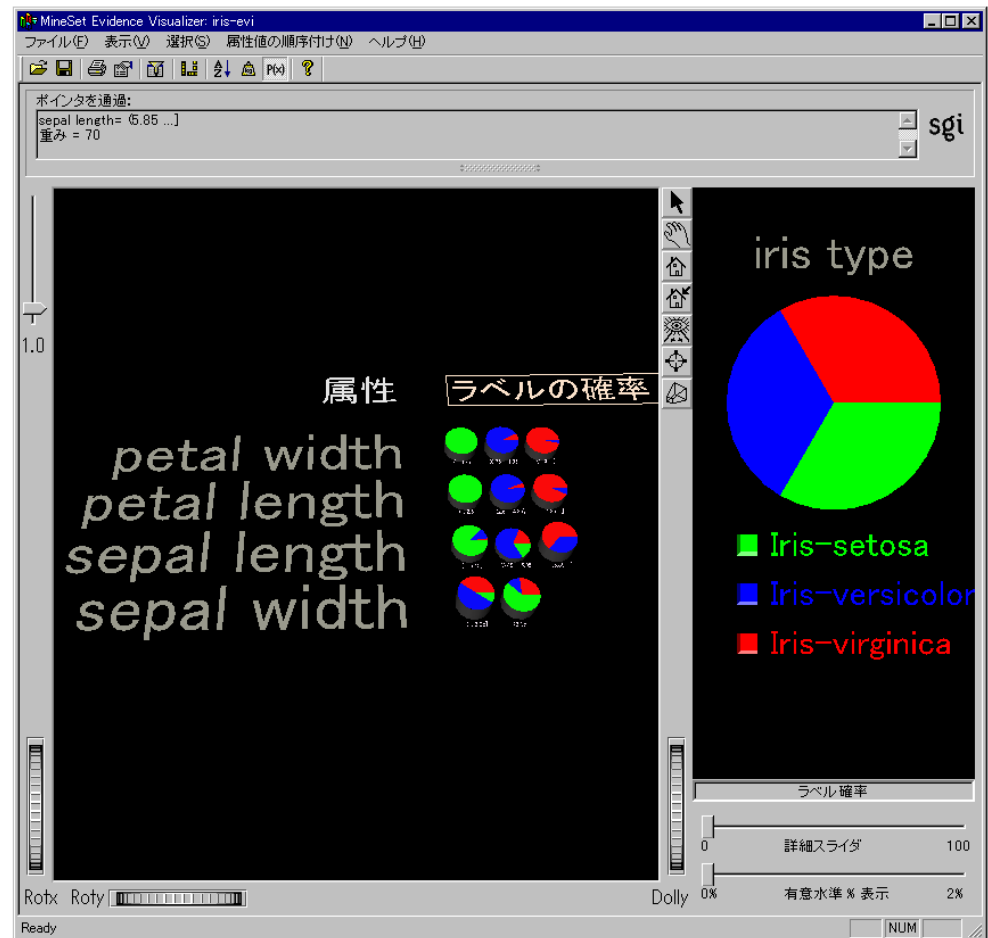


図 10-2 確率を表示したエビデンス・ビジュアライザ

エビデンス・ツールの起動

エビデンス・ツールの起動は、Tool Manager から行うのが最も簡単です。その他のオプションについては、『*MineSet 3.0 Enterprise Edition Reference Guide*』を参照してください。

1. Tool Manager の「ファイル」メニューからサーバに接続し、ログインします。
2. 同じメニューから「新しいデータファイルを開く」を選択し、必要なファイル名を選択または入力します。
3. 「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（図 10-3）をクリックし、その下に並んでいるタブから「クラス判別」を選択します。
4. 「モード」ポップアップ・メニューからモードを選択します。4 種類のモードについては、第 7 章「予測モデリングの理解」を参照してください。
5. 「分析」ポップアップ・メニューから「エビデンス」を選択します。
6. 「離散型ラベル」ポップアップ・メニューから、離散型ラベルとして使用する属性を選択します。たとえば、mushroom データセットの場合は edibility（食用）を選択します。
7. 「実行」をクリックして解析を起動します。

「Tool Manager」ウィンドウ下部の「ステータス」パネルに進行状況と結果の統計が表示されます。

注記：エビデンスのその他のオプションについては、『*MineSet 3.0 Enterprise Edition Reference Guide*』の「エビデンス・オプション」の説明を参照してください。

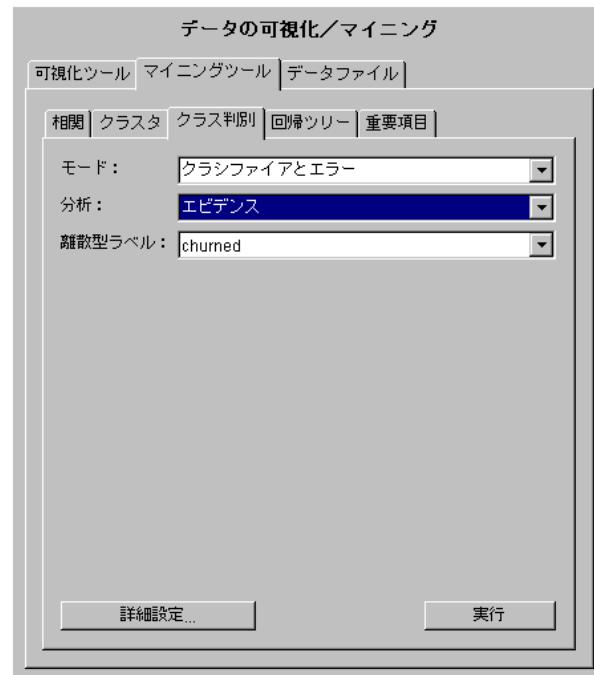


図 10-3 Tool Manager の「データの可視化／マイニング」パネル

エビデンス・ビジュアライザによる結果の表示

エビデンス・ビジュアライザを起動したら、結果をいくつかの方法で表示できます。ビジュアライザ・ウィンドウの左側のパネルでは、エビデンス表示、確率表示、バー表示を交互に切り替えることができます。

- エビデンス表示では、エビデンスを表すケーキグラフが表示されます。「[エビデンス表示](#)」(180 ページ) の項を参照してください。
- 確率表示では、確率を表す円グラフが表示されます。「[確率表示](#)」(183 ページ) の項を参照してください。
- バー表示には、特定のラベルに有利なまたは不利なエビデンスを表すバーグラフが表示されます。「[バー表示](#)」(184 ページ) の項を参照してください。

ビジュアライザ・ウィンドウの右側の「ラベル確率 (*Label Probability*)」パネルには、データベース全体に渡るラベルの値の分布を表す円グラフが表示されます。これにつ

いては、「[エビデンス表示による確率の検討](#)」を参照してください。このパネルは、左側パネルの 3 つの表示に対し常に表示されます。

エビデンス表示

エビデンス・ビジュアライザを最初に起動すると、メインパネルにエビデンスが表示されます (図 10-4)。

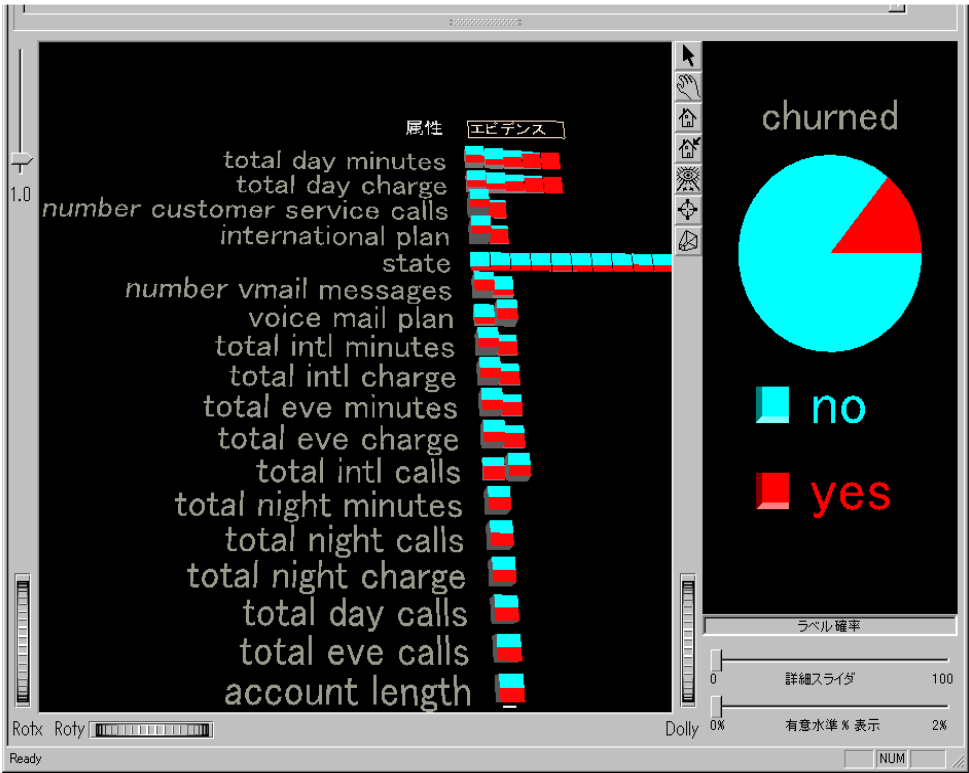


図 10-4 エビデンス・ビジュアライザによる churn データセットのケーキグラフ

エビデンス表示の説明

エビデンス表示では、データセットの属性または項目ごとに1行のケーキグラフが表示されます。各行では、その属性の離散値ごとに1つのケーキが表示されます。属性の値が離散値ではなく連続値の場合は、値が複数の範囲または階級に自動的に分割されます（たとえば、0から5、5から10など）。階級は、隣接するケーキ間の相違が最大になるように行われます。ケーキの高さは、その属性値を持つレコードの重みに比例しています。用語集の「重み (*Weight*)」を参照してください。重み属性が設定されていない場合は、レコード数が使用されます。

ケーキグラフのスライスのサイズは、絶対的な確率ではなく、条件付き確率（エビデンス）を示しています。たとえば、*churn* データセットでケーキが半分「yes」で半分「no」なら、ケーキで表された属性の特定の値を持つ顧客の解約確率は、母集団全体の確率と同じです。つまり、その属性の値には予測価値がまったくないことになります。それに比べケーキが90%「yes」で10%「no」なら、属性の特定の値をもつ顧客の解約確率は母集団全体よりもかなり高くなります。ただし、これは顧客が90%の確率で解約することを意味するわけではありません。全体的な確率は、この属性以外にも多くの変数に依存しているからです。詳細は、『*MineSet 3.0 Enterprise Edition Reference Guide*』の「エビデンス」の説明を参照してください。

行の順序は、ラベルを予測する上での属性の相対的な寄与率を示しています。各属性には、1から100の寄与率が割り当てられます。シーンの属性名の上にマウスポインタを移動すれば、この値を見ることができます。情報は、メニューバーとビジュアルライザパネルの間の領域に表示されます。たとえば、[図 10-5](#) に示す *mushroom* データセットの場合、キノコの匂いは、食用に適しているかどうかを極めてよく表していますので、匂い (*odor*) 属性が寄与率 90.69 でリストの最上位にランクされています。それに比べ茎の形状 (*stalk shape*) は食用に適しているかどうかをまったく表していませんので、この属性は寄与率 .75 で最下位にランクされています。

さらに、カサの色 (*veil-color*) 属性は興味深い結果をもたらします。ほとんどのキノコのカサは白いため、この属性は寄与率の点で最下位近くにランクされています。カサが白いキノコは、食用と有毒にほとんど均等に分割されています。これは、データセットのどのキノコも全体的な確率は有毒であるという前述の確率に非常に近いものです。したがって、統計的にはこの属性がいくつかのクラスに多くの根拠を与えることがないため、この属性は寄与率の点で最下位近くにランクされています。しかし、カサの色の列で1、2、4番めのケーキである茶色、オレンジ、黄色の場合は、事情がかなり異なります。カサが茶色やオレンジのキノコは100%食べられますが、カサが黄色のキノコは100%有毒です。これは説得力がありそうに見えますが、カサが白以外のキノコは数が少ないため、カサの色は有力な判断特性とはなりません。

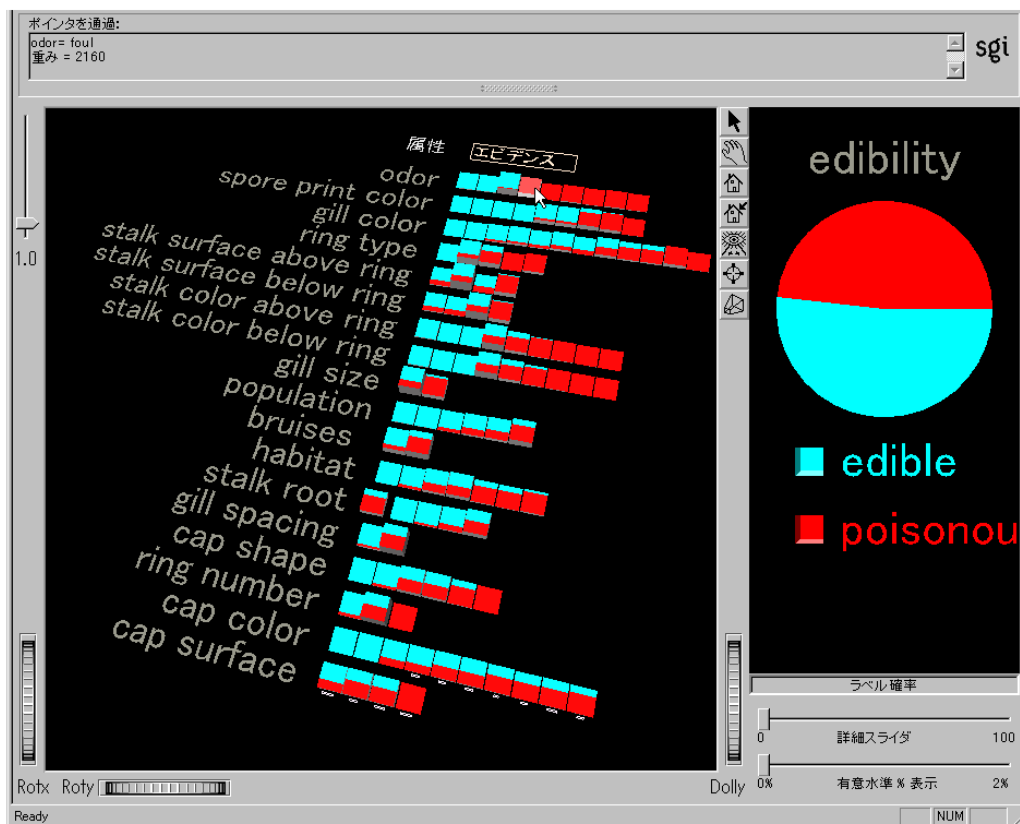


図 10-5 エビデンス・ビジュアライザで表示した mushroom データセット

エビデンス表示による確率の検討

エビデンス・ビジュアライザを最初に起動すると、「ラベル確率 (Label Probability)」パネルにデータセット全体のラベル確率が表示されます。たとえば、*churn* データセットを使った図 10-4 の例では、*churned* ラベルが「yes」のランダムレコードの確率は 14.14% です。いずれかのラベルの横の色付きボックスの上にマウスポインタを移動すると、ウィンドウの左上にそのラベルの確率と重みが表示されます。

属性に対し特定の値を持つレコードの確率を示す円グラフを見たい場合は、その値のケーキをクリックします。属性を選択するには、マウスが選択モードになっていなければなりません。選択した値を持つレコードの確率を示す円グラフが「確率」パネル

に表示されます。グラフの選択を解除するには、メインパネルの背景をどこかクリックします。

ある属性の組合せを持つレコードの確率を見るには、<Ctrl> キーをクリックしていくつかのケーキを選択します。「確率」パネルには、その属性の組合せを持つレコードの確率の円グラフが表示されます。属性の確率はその属性が他の属性から独立しているものとして計算されるため、結合された確率が完全に正確であるとは限りません。このことを示す警告が表示されます。この確率は推定として使用してください。この確率が真に正確であるためには、それらの属性が完全に独立していなければなりません。しかし、そのようなことはまれです。

属性の組合せを選択しているときに、両立しない属性を選択してしまうことがあります。このような場合には、「ラベル確率」パネルにグレイの円グラフが表示されます。これは、その値の組合せをもつレコードがまったくないか、ほとんどないことを示します。

「ラベル確率」パネルをリセットして前の確率を表示するには、メインパネルの背景をどこかクリックします。

確率表示

確率表示 (図 10-6) では、ケーキグラフの代わりに、確率を表す円グラフが表示されます。エビデンス表示から確率表示に移るには、グラフの上部にある白枠ボックス内の「エビデンス」をクリックするか、ビジュアライザの「表示」メニューで「エビデンスとして表示」チェックボックスをクリックします。エビデンス表示に戻るには、「確率」をクリックするか、同じチェックボックスを再びクリックします。

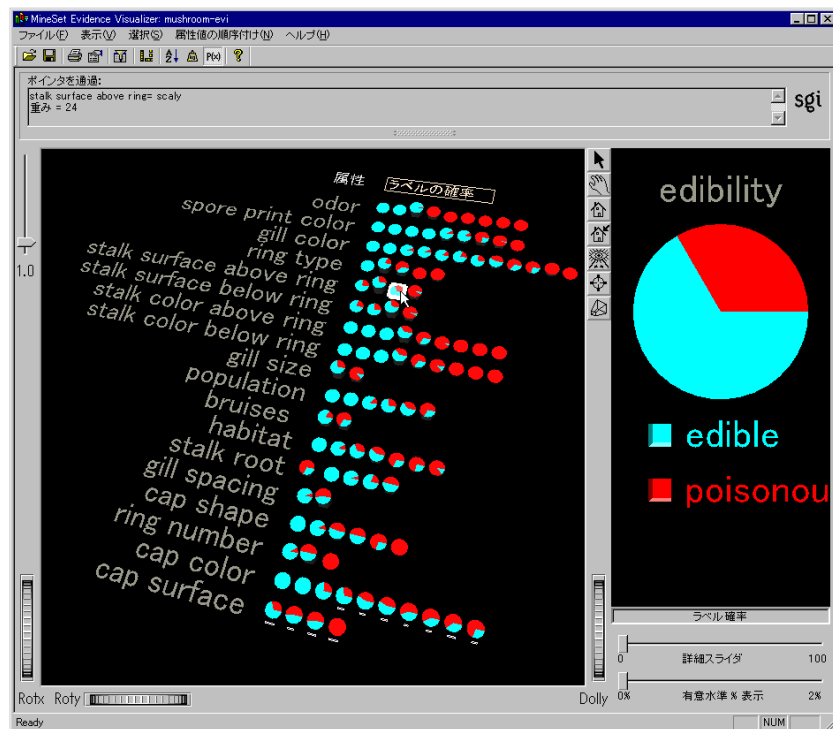


図 10-6 エビデンス・ビジュアライザの円グラフ

バー表示

バー表示には、選択されたラベルの値に有利なまたは不利なエビデンスが表示されます。バー表示に移るには、「ラベル確率」パネルで、必要なラベルの横にある色付きボックスをクリックします。メインウィンドウにバークラフが表示されます(図 10-7)。バーの高さは、選択されたラベルの値に有利なエビデンスを表します。表示を終了するには、色付きボックスを再びクリックします。

バー表示の上部にある白枠ボックス内の「対象」をクリックすると、モードが「対象外」に切り替わります。「対象」に戻るには、「対象外」をクリックします。

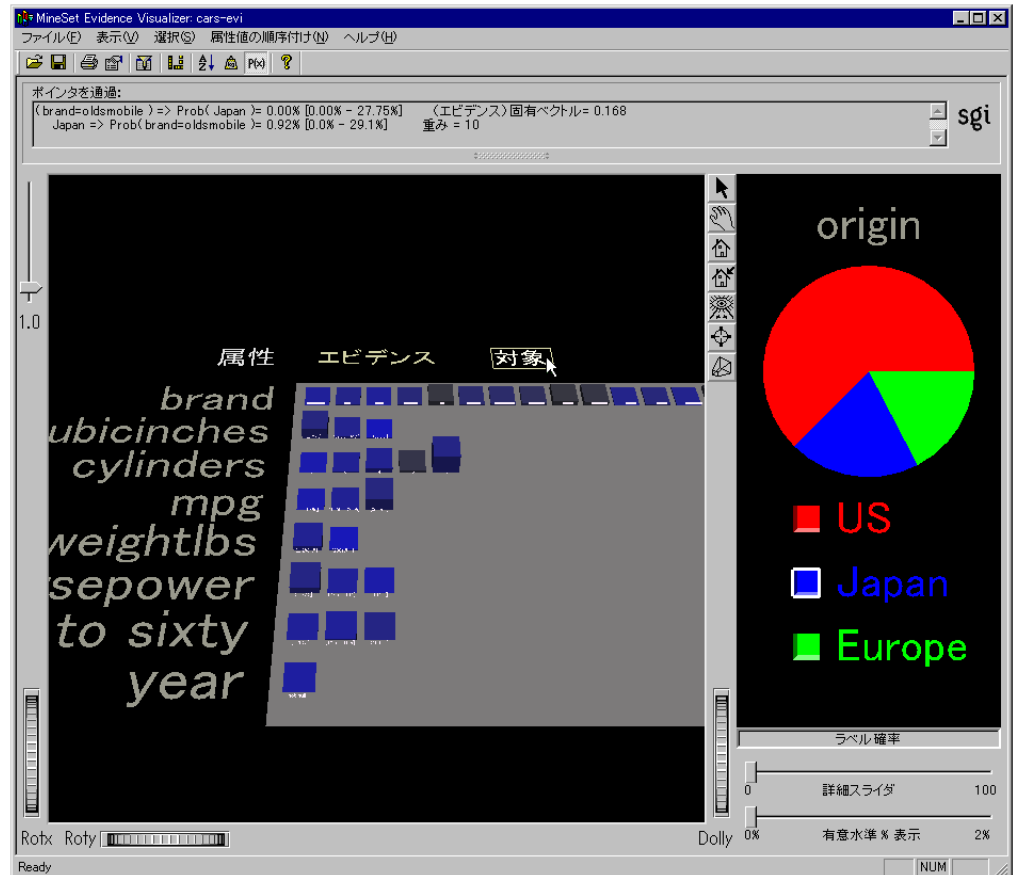


図 10-7 cars データセットでラベルの値「Japan」を選択

「対象」は、特定のラベルの値の予測でどの値が最も役立つかを判定するときに便利です。「対象外」は、その反対を判定するときに便利です。つまり、ラベルがどのようなときに発生しないかを予測する際にどの値が最も役立つかを判定するときに便利です。これらの値がどのように計算されるかについての技術的な説明は、『*MineSet 3.0 Enterprise Edition Reference Guide*』の「エビデンス」の説明を参照してください。

バーの色は、完全な純色から完全なグレイまでいろいろあります。バーのグレイは 95% 信頼区間に基づいています。他方、これはその値の重みに依存しています。したがって、グレイに近いバーは、低い重みと大きい信頼区間を持っています。これらのバーの高さはおそらく正確ではありません。これとは反対に、完全な純色のバーの高さや対応するエビデンス値は信頼できます。この値は、大きな重みに基づいており、多数のレコードを表しているからです。正確なレコード数（重み）は、バーを強調表示すると、テキスト出力行に出力されます。

エビデンス・ビジュアライザ表示の変更

表 10-1 に、結果の表示方法を変更するさまざまなオプションを示します。

表 10-1 エビデンス・ビジュアライザ表示の変更

目的	アクション
ケーキグラフ（エビデンス）と円グラフ（確率）を切り替える	白枠内の「エビデンス」または「確率」をクリックするか、「表示」->「エビデンスとして表示」を選択して切り替えます。
ケーキグラフまたは円グラフとバークラフを切り替える	「ラベル確率」パネルで、必要なラベルの横にある色付きボックスをクリックします。
バークラフとケーキグラフまたは円グラフを切り替える	「ラベル確率」パネルで、必要なラベルの横にある色付きボタンをクリックします。
バー表示の「対象」と「対象外」を切り替える	「対象」または「対象外」を囲む白枠ボックスをクリックします。
特定のケーキまたは円の属性値とレコード重みを表示する	ケーキまたは円の上にマウスポインタを移動します。
ラベル確率を表示する	「ラベル確率」パネルで、そのラベルの横にあるボックスの上にマウスポインタを移動します。
一定の属性値を持つ新しいレコードの予測分割を右のパネルに表示する	その値のケーキをクリックします。
属性の重要度と合計重みを表示する	左の属性名の上にマウスポインタを移動します。
属性値の一定の組合せを持つレコードの推定ラベル確率分布を取得する	<Ctrl> キーをクリックして複数のケーキを選択し、「ラベル確率」パネルの円グラフを調べます。

表 10-1 (続き)エビデンス・ビジュアライザ表示の変更

目的	アクション
選択を解除する	左のパネルの背景をどこかクリックします。
選択した属性値を持つすべてのレコードを表示する	属性値の必要な組み合わせを表すケーキをクリックし、「選択」メニューから「オリジナルデータの表示」を選択します。
選択した属性値を持つレコードを除く、すべてのレコードを表示する	「選択」メニューの「補集合データのドリルスルー」をオンにし、同じメニューから「オリジナルデータの表示」を選択します。
グラフ間の高さの違いを拡大または縮小表示する	メインウィンドウ左側の高さスライダを使用します。
表示する属性行の数を減少または増加する。その際、予測に最も役立たないまたは役立つ属性から減少または増加する。	詳細スライダを使用します。重要度がスライダ値より小さい属性が削除されます。
カウントや重みが小さい属性値をフィルタで除去する	重みスライダを使用します。スライダで示される百分率より重みが小さい値が削除されます。
一部の属性値を知っている新しいレコードのクラスラベルを予測する	これらの属性値に対応するグラフをクリックします。右側の円グラフの最も大きなスライスが予測クラスに対応します。

エビデンス・クラシファイアによる予測

この章では、エビデンス・クラシファイアとビジュアライザを使って、ラベルがすでに分かっているデータを調べることを中心に説明しました。実際には、このツールは、ラベルが分かっていないデータのラベルを予測するときに最も役立ちます。この場合には、既知データを使ってモデルを作成、保存し、ラベルになっていないデータにそのモデルを適用します。モデルの作成と適用については、『*MineSet 3.0 Enterprise Edition Reference Guide*』の訓練事例とモデルの適用に関する説明をそれぞれ参照してください。

ビジュアライザのメニューの詳細は、『*MineSet 3.0 Enterprise Edition Reference Guide*』の
エビデンス・ビジュアライザのメニューの説明を参照してください。

予測モデリングの改良

この章では、前の各章で説明した概念を基に、Tool Manager を使って分析における誤差の費用を調べ、データマイニング操作を実社会のニーズに適用させます。MineSet の高度な解析オプションを使えば、精度の向上を図り、誤りの費用を埋め合わせ、プロジェクト投資の継続が無意味となる段階を判断できます。この章の構成は次のとおりです。

- 「モデルの精度の検査」(189 ページ)
- 「混同マトリックス (Confusion Matrix) と損失マトリックス (Loss Matrix) によるモデルの微調整」(196 ページ)
- 「改善曲線 (Lift Curve) と ROI 曲線 (ROI Curve) によるモデルの評価」(204 ページ)

このマニュアルのオンライン版では、画面がカラーで表示されます。オンライン版は「ヘルプ」メニューから使用します。

モデルの精度の検査

これまでに作成したモデルの精度はいろいろな方法で検査できます。「[予測モデルの評価](#)」(137 ページ) ですでにいくつかの方法を説明しましたが、ここではさらに、異なるソースを使ってモデルを検査する方法を説明します。モデルをまだ作成していない場合は、[第 7 章「予測モデリングの理解」](#)を参照してモデルを作成してください。

モデルのテスト

以前に作成したモデルがすでにある場合は、MineSet の「モデルのテスト」機能を使って、対象となるデータに対しこのモデルをテストできます。データセットには、モデルの作成時に使用したのと同じ名前とデータ型を持つ項目が含まれていなければなりません。

第 7 章の「[予測モデルの適用](#)」で説明する「モデルの適用」の場合とは異なり、「モデルのテスト」では、このテーブルにさらにモデルの作成時に使用したのと同じ名前とデータ型を持つラベルの項目が必要です。これは、正しい答えが分かっているデータを使ってこのテストを行う必要があるためです。テストファイルには、*iris-test* など、接尾辞 *-test* を持つデフォルト名が割り当てられます。

次の例では、*iris* データセットからテストファイルを作成します。サーバにログインしたら、データセットを選択してください。サンプルファイルは、MineSet を最初にインストールした *MineSet 3.0\data\iris.schema* です。

現在のデータセットを使ってモデルをテストするには、次の手順に従います。

1. Tool Manager の「データ変換」パネルで「モデルの適用」ボタンをクリックします。
2. 「モデルのテストと適用」ダイアログ・ボックスで「モデルのテスト」タブをクリックします。
3. ダイアログ・ボックス上部に表示される使用可能なモデルのリストからモデル名を選択すると（たとえば、*iris-dt.class*）、デフォルトのテスト・ファイル名が表示されます（[図 11-1](#)）。

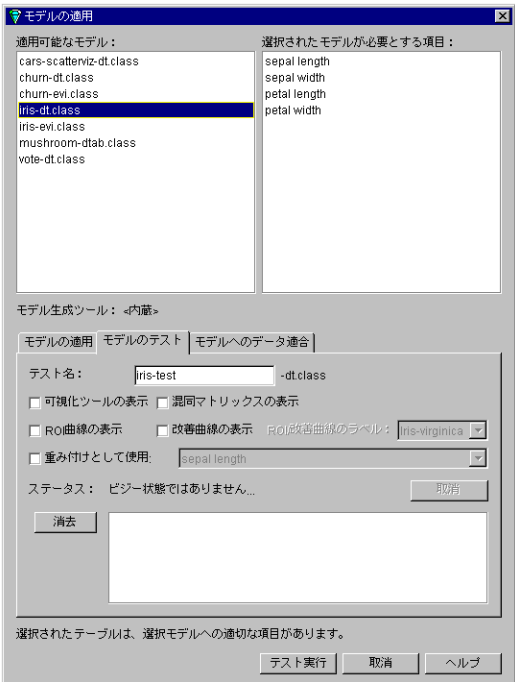


図 11-1 「モデルのテスト」 パネル

現在の表に不足している項目があると、右側の「選択されたモデルが必要とする項目」リストにそれらの項目名が強調表示されます。別のモデルを選択するか、不足している項目を復元してください。「モデルのテスト」パネルには、次のオプションがあります。

表 11-1 「モデルのテスト」 パネルのオプション

選択	操作
テスト名	テストファイルの名前を表示します。この名前は変更できます。
可視化ツールの表示	テーブルをテストセットとして使用して、クラシファイアを可視化します。
混同マトリックスの表示	テーブルレコードに基づいてモデルの混同マトリックスを表示します。 「混同マトリックスの表示」(198 ページ) を参照してください。

表 11-1 (続き) 「モデルのテスト」 パネルのオプション

選択	操作
ROI 曲線の表示	指定されたテーブルレコードに基づいてモデルの ROI 曲線を表示します。「投資利益率曲線の使用によるマーケティング・メリットの発見」(206 ページ) を参照してください。
改善曲線の表示	指定されたラベルの値を使ってモデルの改善曲線を表示します。「改善曲線 (Lift Curve) と ROI 曲線 (ROI Curve) によるモデルの評価」(204 ページ) を参照してください。
ROI/ 改善ラベル	ポップアップ・メニューに使用可能なラベルの値を表示します。
重み付けとして使用	レコードの重み付けに使用する属性をポップアップ・メニューから選択します。さらに、第 3 章の「レコードの重み付け」の項を参照してください。

4. 「テストの実行」をクリックします。結果が「モデルのテスト」パネル下部のテキスト・フィールドに表示されます。さらに、選択したビジュアライザ・オプションが表示されます。

この場合、以前に作成したモデル *iris-dt.class* に対し *iris.schema* のデータセット全体が使用され、150 のテスト・インスタンスに対し 4 つのエラーが検出されました。

モデルへのデータの適合

新しいデータがある場合は、「モデルへのデータ適合」パネル (図 11-2) を使って、現在のテーブル内のデータを以前に作成したモデルに適合します。この結果、元のモデルと同じ構造を持つ新しいモデルが生成されます。ただし、新しいモデルでは、このテーブルのデータに基づいて確率推定が更新されます。これについては、この後の「バックフィッティング (backfitting) による誤差の評価」を参照してください。テーブル内のすべてのデータがモデルの構造に適合されるため、誤差の評価は行われません。「モデルへのデータ適合」は、ブースティングを使って作成したモデルには使用できません。これについては、この後の「ブースティング (boosting) による精度の向上」を参照してください。新しいモデルの性能を別のテストセット (適合データから分離したもの) を使って評価するには、「モデルのテスト」機能を使用します。

既存のモデルにデータを適合するには、次の手順に従います。

1. Tool Manager の「データ変換」パネルで「モデルの適用」ボタンをクリックします。
2. 「モデルのテストと適用」ダイアログ・ボックスで「モデルへのデータ適合」タブをクリックします (図 11-2)。表 11-2 に「モデルへのデータ適合」オプションを示します。

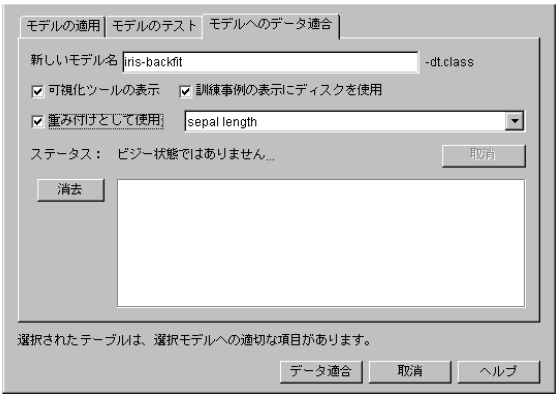


図 11-2 「モデルへのデータ適合」パネル

「モデルへのデータ適合」パネル (図 11-2) では次のオプションを指定します。

表 11-2 「モデルへのデータ適合」オプション

選択	アクション
可視化ツールの表示	新しいモデルを可視化します (ツリーベースのモデルのみ)。
訓練事例の表示にディスクを使用	可視化における訓練事例のデータの割合をディスクを使って表示します。
重み付けとして使用	レコード重みとして使用する属性をポップアップ・メニューから選択します。

バックフィッティング (backfitting) による誤差の評価

モデルをあるレコードセットにバックフィッティングを行うと、確率推定が更新されるだけで、モデルの構造は変更されません。バックフィッティングは、モデルへのデータの適合（デフォルトでバックフィッティングが使用される）と似ています。バックフィッティングは、次の場合に便利です。

- 小さい訓練事例に基づいてモデルの構造を作成し、その構造を大きいデータセットにバックフィッティングする場合。この方が大きいモデル構造を作成するより速いことがしばしばあります。
- 精度を高くしたい場合。モデルの構造内のカウント、重み、確率は、訓練事例だけでなく、すべてのデータを反映しています。

バックフィッティングを有効にするには、次の手順に従います。

1. Tool Manager の「データの可視化／マイニング」パネルで「マイニングツール」タブをクリックします。
2. 「クラシファイア」タブをクリックします。
3. 「クラシファイアとエラー」モードを選択し、「詳細設定」ボタンをクリックして、「詳細設定」ダイアログを表示します。「誤差推定オプション」の中の「テストセットのバックフィット」チェックボックスをクリックします。「ブースティング」が有効になっていると、このチェックマークは無効になります。

ブースティング (boosting) による精度の向上

モデルの選択において誤差率が最も重要な基準になる場合があります。そのような場合には、ブースティングを有効にします。ブースティングとは、いくつかの異なるモデルを作成し、それらの予測を、重み付けした投票方式を使って結合するアルゴリズムです。ブースティングでは、他よりモデル化が難しいデータ例が集中的に生成されるため、モデル精度が向上します。ただし、ブースティングされたモデルを可視化することはできません。

選択した任意のモデルに対しブースティングを有効にするには、次の手順に従います。

1. Tool Manager の「データの可視化／マイニング」パネルで「マイニングツール」タブをクリックします。
2. 「クラス判別」タブをクリックします。

3. 任意のモードから「詳細設定」ボタンをクリックして、「詳細設定」ダイアログを表示します。「分析ツールオプション」セクションで「ブースト（可視化なし）」チェックボックスをクリックします。

ブースティングによって精度が必ず向上するとは限りませんが、向上することもしばしばあります。ブースティングを適用したモデルは可視化できませんが、そのようなモデルの混同マトリックス (Confusion Matrix)、改善曲線 (Lift Curve)、学習曲線 (Learning Curve)、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とまうため、ブースティングを使用しない場合に比べて、分析ツールの処理に通常の 25 倍の時間がかかることもあります。ブースティングを適用したモデルではバックフィッティングは使用できません。

注記： ブースティングの基本理論は 3 つ以上のクラスを持つ問題では一般法則化されていませんが、MineSet では任意の数の値を持つラベルを使ってブースティングを適用できます。このような場合には、ブースティングによって誤差率が改善する可能性は少なくなります。

ブースティングでは、訓練事例に新しい重み分布を割り当て、その重み付けされたデータセットからモデルを作成するという処理が繰り返されます。ブースティングの詳細については、『*MineSet 3.0 Enterprise Edition Reference Guide*』を参照してください。

混同マトリックス (Confusion Matrix) と損失マトリックス (Loss Matrix) によるモデルの微調整

クラス判別の誤差に気付いた場合には、クラシファイアがどのように混乱しているのかや、それに伴う損失がどの程度になるのかが分かれば役立ちます。MineSet にはこの疑問に答えるマトリックスが 2 つあります。相互に補完的な混同マトリックスと損失マトリックスです。

混同マトリックスによる誤差の調査

キノコを食用と有毒にクラス判別するとします。キノコを有毒とクラス判別すると、\$2 の損失になります。これはキノコの費用です。しかし、有毒キノコを食用とクラス判別すると、病院に収容され、\$10,000 ドルを支払うはめになるかもしれません。MineSet の混同マトリックスと損失マトリックスはこのような状況に対処するためのものであり、誤差や不正確な予測の詳しい全体像を明らかにしてくれます。

図 11-3 に、mushroom データセットの混同マトリックスを示します。これはデータセットの 10% を使って学習した決定木に基づいています。ただし、10% という割合は、推奨する訓練事例よりもかなり少ない量です。詳細は、[第 7 章「予測モデリングの理解」](#)を参照してください。

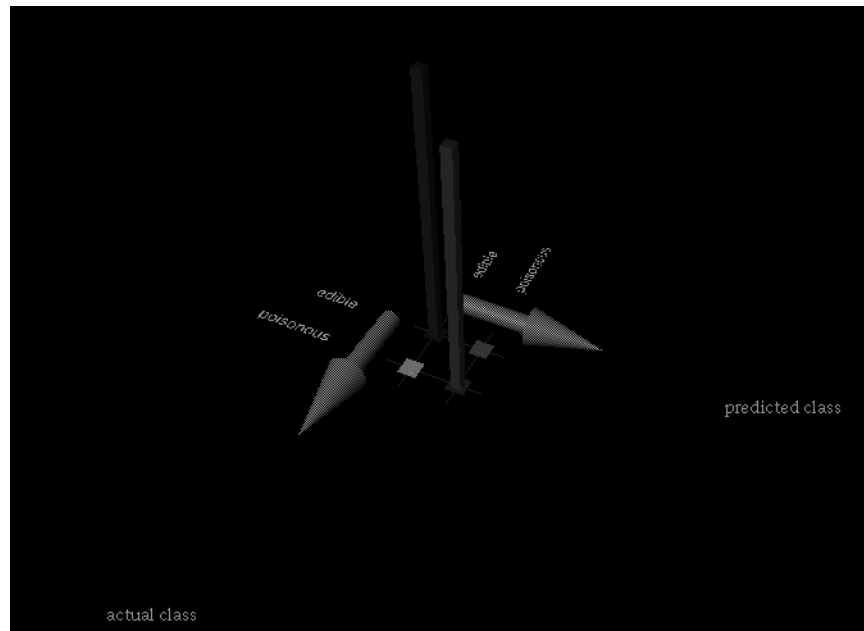


図 11-3 mushroom データセットの混同マトリックス

有毒キノコを表す 8 つのレコードが食用としてクラス判別され (0.1%)、食用キノコを表す 15 のレコードが有毒とクラス判別されています (0.2%)。残りの 3793 種類の食用キノコと 3496 種類の有毒キノコは正しくクラス判別されています。モデルの誤差率は 1% にも満たない 0.31% ですが、上述の仮定に基づいて計算すると推定損失は $\$10000 \times 8 + \$2 \times 15 = \$80,030$ になります。

混同マトリックスでは、誤差の種類を可視化することで、モデルによる誤差の詳細な全体像が得られます。2 つの軸の一方は予測されたクラス値を、他方はテストセット (予備セット) から得られた現在のクラス値を表します。混同マトリックスはバックフィッティングの前に計算されます。次の項では、上のシナリオに基づく混同マトリックスをどのように表示するかを説明します。

混同マトリックスの表示

この例では、「Tool Manager」ウィンドウでデータファイル *mushroom.schema* を開き、キノコを食用と有毒にクラス判別するときの混同を調べます。このファイルは、MineSet を最初にインストールした *MineSet 3.0\data* ディレクトリにあります。

1. Tool Manager の「データの可視化／マイニング」パネルで「マイニングツール」->「分析モデル」タブを選択し、ポップアップ・メニューから次の選択を行います。
「モード」：「クラシファイアとエラー」
「分析」：「決定木分析」
「離散型ラベル」：「edibility」
2. 「詳細設定」をクリックすると、[図 11-4](#) の「クラシファイアオプション」パネルが表示されます。

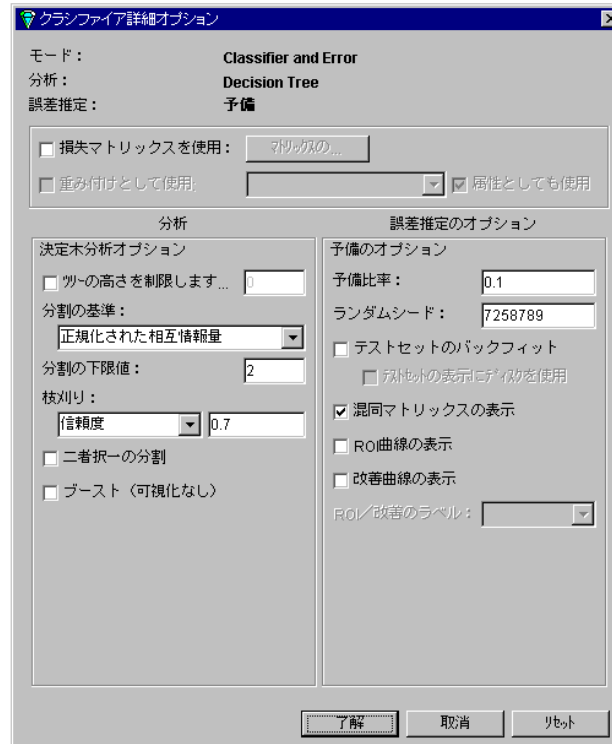


図 11-4 「混同マトリックスの表示」が選択された「クラシファイアオプション」パネル

- 「誤差推定オプション」パネルで、「予備比率」を 0.1 に変更し、「バックフィットテストセット」をオフにし、「混同マトリックスの表示」をオンにし、「了解」をクリックします。

訓練事例のサイズを小さくしましたが、これはクラシファイアに故意に誤りを起こさせるためです。一般には、これよりずっと大きい訓練事例を使用します。

- Tool Manager の「データの可視化／マイニング」パネルで「実行」をクリックします。

決定木が表示されたらこれを終了し、図 11-5 に示す混同マトリックスを調べます。混同マトリックスには、クラシファイアがどこでクラス判別を間違うかが表示されます。そのデータに関して明らかになった情報から損失マトリックスを作成し、より許容できる間違いを選択します。

図 11-5 の混同マトリックスは、この小さい訓練事例では 8 種類のキノコが誤って食用にクラス判別されていることを示しています。これは全体の 0.1% です。「選択リスト」テーブルを表示するには、可視化のバーをクリックし、Tool Manager の「選択」->「値の表示」を選択します。

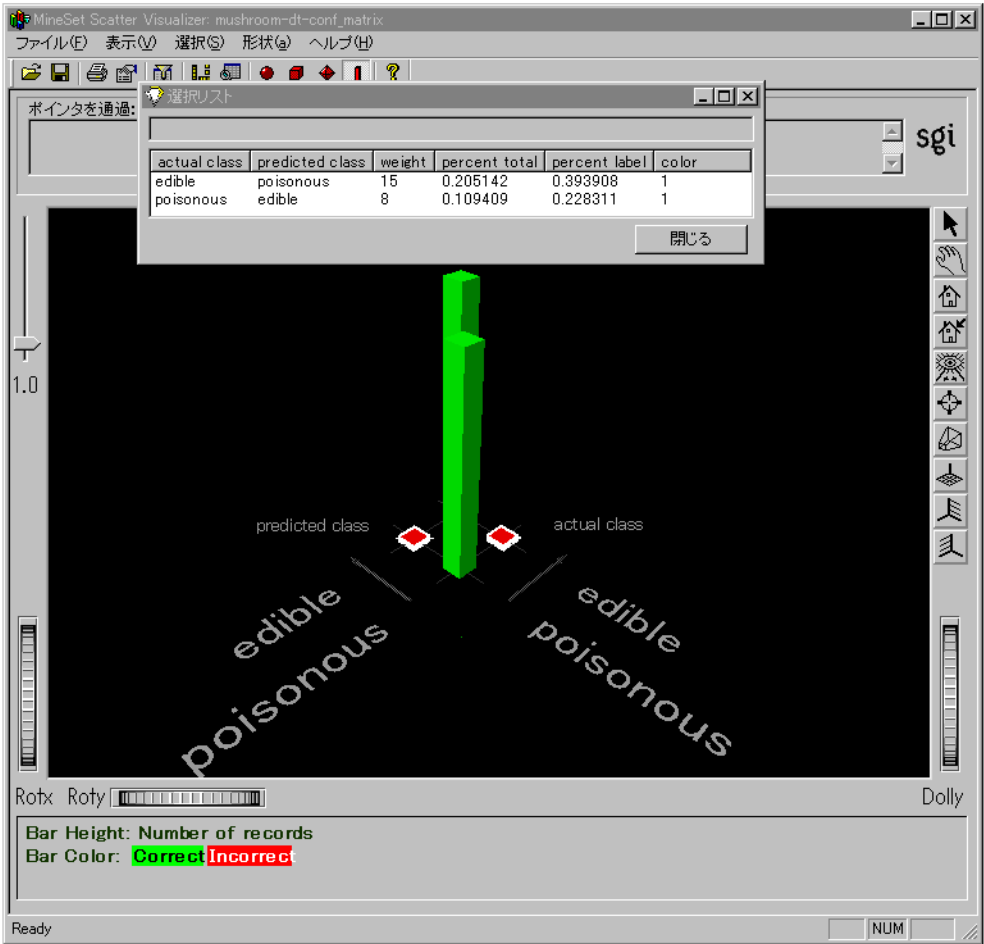


図 11-5 間違ったクラス判別を示す mushroom データセットの混同マトリックス

損失マトリックスの定義

この項では、損失マトリックスを使って相対的に高価なクラス判別誤差を減らすにはどうするかを説明します。この例でも、前の項の条件を引き続き使用します。つまり、mushroom データセット、決定木分析、ラベル edibility のクラス判別を使って損失マトリックスを生成します。

1. 「ファイル」->「終了」を選択して混同マトリックス画面を終了し、「Tool Manager」ウィンドウに戻ります。
2. 「データの可視化／マイニング」パネルの「マイニングツール」タブで「詳細設定」をクリックして「クラシファイアオプション (Classifier Options)」を開きます。
3. 同じパネルの別の場所で「損失マトリックスの使用」をクリックし、しばらく待ちます。
4. 「マトリックスの編集」をクリックし、誤差の費用を重み付けします。図 11-6 のような損失マトリックスが表示されます。



図 11-6 損失マトリックスの編集パネル

5. 「現在の値」の各行に、[図 11-6](#) のように次の値を左から右の順に設定します。

edible: 1—0—1

poisonous: 1—50—0

これにより、有毒キノコを食用と誤ってクラス判別する費用をその反対の場合に比べ 50 倍に重み付けします。

6. 「了解」をクリックして損失マトリックスの値を設定し、「実行」をクリックしてクラシファイアを実行します。

新しいクラシファイアは非常に慎重であるため、有毒キノコを食用にクラス判別する誤りはゼロになりました。それに比べ、食用キノコを誤って有毒とクラス判別した件数は 1558 となりました。したがって、新しい推定損失は、キノコの費用（1 個 \$2）と病院の費用（\$10,000）から $\$10,000 \times 0 + \$2 \times 1,558 = \$3,116$ となります。これは、損失を考慮に入れなかった場合のクラシファイアの費用のわずか 3% です。[図 11-7](#) に、損失を考慮した混同マトリックスを示します。

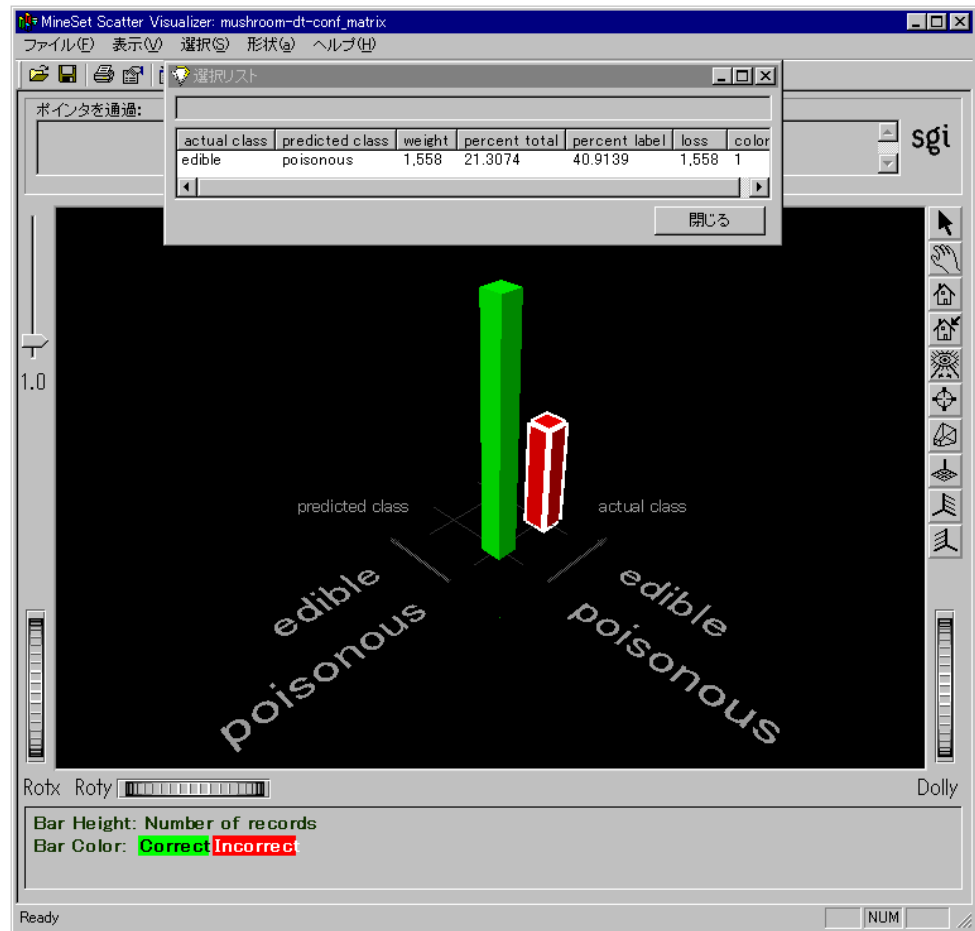


図 11-7 mushroom データセットの混同マトリックスと損失マトリックス

誤差の推定については、『*MineSet 3.0 Enterprise Edition Reference Guide*』を参照してください。

改善曲線 (*Lift Curve*) と ROI 曲線 (ROI Curve) によるモデルの評価

企業が何らかの戦略を打ち出す場合には、その利益の予測が、方針を変えるか、プロジェクトを放棄するか判断に役立つことがしばしばあります。MineSet を使えば、可能性の高い予測を検討し、重大な結果となる前に変更を行うことができます。それには、改善曲線と ROI 曲線を使用します。

改善曲線による予測の表示

改善曲線とは、レコードのランダムな順序と、特定のラベルの値の予測でモデルが生成した順序との差異を示すグラフです。たとえば、モデルを作成してどの顧客が解約しそうかを予測する場合には、そのような顧客が実際に解約する前にこれらの顧客に対しアクションが必要です。改善曲線はそのようなときに使用します。この例では churn データセットを使用します。

改善曲線では、レコード数を 0 から 100% で X 軸にプロットし、一定のラベルの値（この場合は Churn=yes）を持つ顧客のレコード数を Y 軸にプロットします。

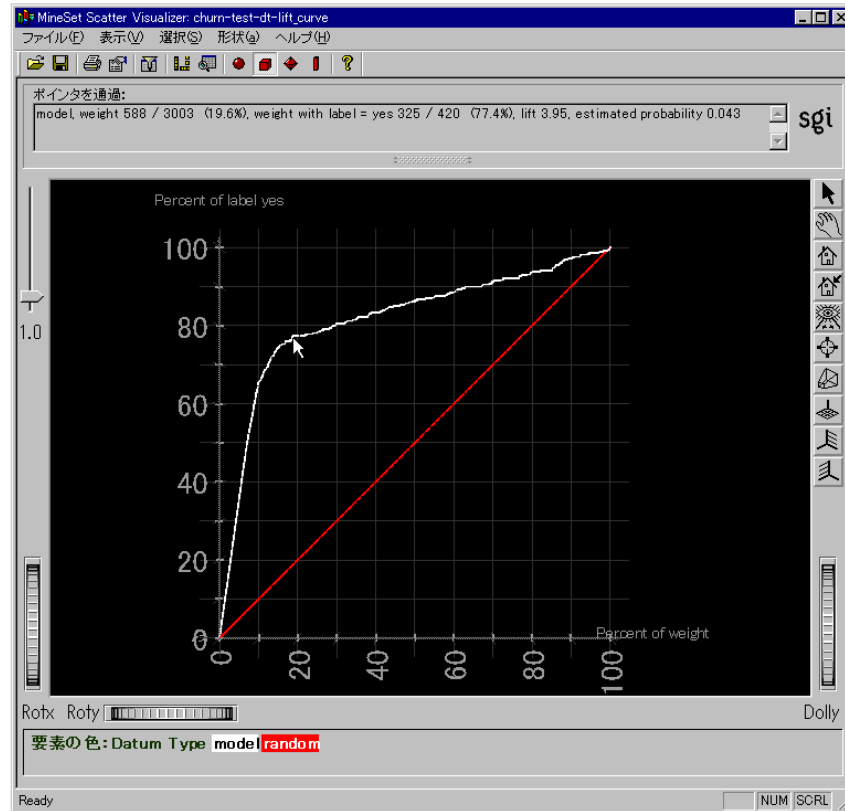


図 11-8 改善曲線

図 11-8 には、2つの曲線が描かれています。下の曲線（赤）は、レコードをランダムに並べたときの、解約が予想される顧客の数を示しています。上の曲線（白）は、レコードごとにモデルが計算したスコア（確率推定）に従ってレコードを並べたときの解約顧客の百分率を示しています。解約の確率が最も高いと判断された顧客のレコードが最初に置かれ、最も低いと判断された顧客が最後に置かれます。モデルによる順序付けの利点は、このモデル曲線とランダム曲線の違いを見ればよく分かります。

改善曲線を作成するには、選択したモデルをテストセットに適用します。この例では、データセットのうち指定するセグメントを訓練事例に使い、残りを、生成されたモデルの実行に使います。

選択したモデルの改善曲線を生成するには、次の手順に従います。

1. Tool Manager の「データの可視化／マイニング」パネルで「マイニングツール」タブをクリックします。
2. 「クラス判別」タブをクリックします。
3. 「クラシファイアとエラー」モードと「決定木」を選択します。
4. 「詳細設定」をクリックし、「クラシファイアオプション」ポップアップ・メニューから、「誤差推定のオプション」の下で「改善曲線の表示」チェックボックスをクリックします。「ROI/ 曲線のラベル」ポップアップ・メニューからラベルの値 (yes) を選択します。
5. 「了解」をクリックして分析オプションを有効にし、ダイアログ・ボックスを終了します。
6. 「実行」をクリックします。

この処理にはある程度の時間がかかるかもしれません。決定木を消してください。改善曲線のポイントを選択すると、その詳細がタスクバーの下で選択ウィンドウに表示されます。ポイントを白（モデル）の曲線に沿って移動し、いろいろな場所でクリックして `churn=yes` の顧客数と百分率を表示し、クラシファイアの推定確率が 0.056 になる曲線の「膝」部分を探してください。解約の可能性のある顧客に販促資料を郵送する場合の投資利益率が急激に減少するポイントはここです。次の手順では、クラシファイアを全データセットに適用します。[第 7 章の「すべてのデータを使用したクラス判別」](#)の項を参照してください。

投資利益率曲線の使用によるマーケティング・メリットの発見

投資利益率 (ROI) 曲線は改善曲線と似ていますが、ROI 曲線では、損失マトリックスの重み付け方法に従って、精度が誤差の代わりに損失で表示されます。損失マトリックスの重み付けについては、[「損失マトリックスの定義」\(201 ページ\)](#)の項を参照してください。モデルが予測を信頼できると判断すると、予測される損失は少なく、レコードは ROI 曲線の左端近くに表示されます。

ROI 曲線が分かったら、データセットの各レコードの結果に対しアクションをとる必要があります。このアクションは、選択したラベルの値に関連するアクションです。たとえば、*churn* データセットの場合、ラベル Yes に関連するアクションは、何らかの販促資料を郵送することかもしれません。これによって解約は少なくなるかもしれませんが、無差別に郵送したのでは費用がかかります。ROI 曲線のピークは、モデルを使って適切な郵送先を予測すれば、テストセットの場合どのくらいの費用が節約できるかを示しています。

ROI 曲線では、一定の誤りをした場合の費用や、アクションをそれ以上続けても意味がないポイントが分かります。「クラシファイアとエラー」を使って *churn* データセットの ROI 曲線を生成するには、次の手順に従います。

1. Tool Manager の「データの可視化／マイニング」パネルで「マイニングツール」タブをクリックします。
2. 「クラス判別」タブ -> 「詳細設定」ボタンを選択します。
3. 「クラシファイアオプション」パネルで、「テストセットのバックフィット」、「混同マトリックスの表示」、「損失マトリックスの使用」がオンになっているか確認します。
4. 上部パネルの「マトリックスの編集」をクリックし、「[損失マトリックスの定義 \(201 ページ\)](#)」のようにして、誤りを犯すことによる費用を重み付けします。ここでは、各行に次のように入力します。

現在の値 : no: 1 0 2

現在の値 : yes: 1 0 -10

「了解」をクリックして損失マトリックスを設定します。

5. 「ROI 曲線の表示」をオンにします。
6. 「ROI/ 曲線のラベル」プルダウン・メニューから yes を選択します (*churn* 例の場合)。
7. 「了解」をクリックします。
8. 「Tool Manager」ウィンドウで「実行」をクリックします。

「決定木」、「混同マトリックス」、「ROI 曲線」という 3 つのウィンドウが表示されます。混同マトリックスでは、クラシファイアが *churn=no* の予測を行うことに慎重であることが分かります。そのため、誤って no にされるケースは少なくなります。さらに、混同マトリックスには、損失を考慮に入れた 4% という重み付けが表示されています。一方の誤差は増加していますが、他方の誤差は減少しています。「ファイル」->「終了」を選択して、「混同マトリックス」と「決定木」ウィンドウを終了し、[図 11-9](#) の「ROI 曲線」ウィンドウを調べます。

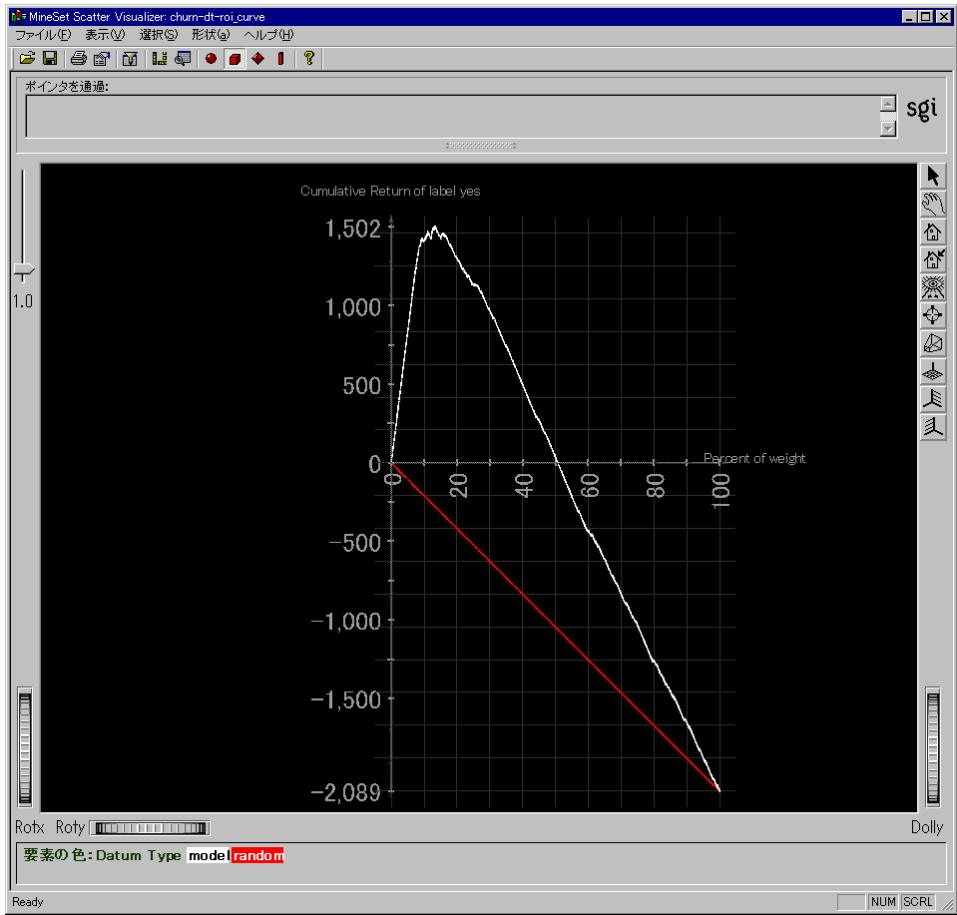


図 11-9 ROI 曲線

ROI 曲線は改善曲線と非常によく似ています。中間の水平線は、損益がゼロになるところを示しています。赤い線は、母数の標本をランダムにとり、メールを郵送する場合の予想成績を示しています。メールには費用がかかるため、すべての人にメールを郵送したのでは損失が出ます。曲線の頂点は投資利益が最大になるポイントを示します。この例では、このポイントは母数の 15.2%、すなわち 1448 人です。

ROI 曲線に使用する損失マトリックスの記入は慎重に行う必要があります。予測するラベルの下の項目がそのラベルの値の ROI 曲線を決定付けます。churn データセットの損失マトリックスにラベル churn を使用する場合には、この項目のエントリは、あり得るすべてのクラスにおいて、そのラベルの値に関連するアクションをとるときに予想される損益を表していなければなりません。たとえば、churn データセットの行「actual value no」の項目「prediction yes」の下のエントリには、値 2 が含まれています。これは、解約するつもりのない人に冊子を郵送する（「yes」に関連するアクション）費用が 2 ドルであることを示します。他方、項目 yes、列 yes の下のエントリには値 -10 が含まれています。これは、顧客が他社に移るのを防止し、メール費用がかかっても会社に 10 ドルの節約をもたらすことを示します。このような費用を推定するのは困難な場合があるため、わずかな差異によって ROI 曲線の頂点の位置が著しく影響されることがあります。

詳細は、『*MineSet 3.0 Enterprise Edition Reference Guide*』を参照してください。最新情報については、MineSet Web サイト (<http://mineset.sgi.com>, 日本語版 <http://minest.sgi.co.jp>) を参照してください。

クラスタリングによるデータのセグメント化

クラスタリングの目的は、データセットの類似している要素や特性を判定することです。クラスタリングは新しいデータセットの探索に便利です。たとえば、`iris` データセットには、アヤメの種類に関し *iris-setosa*、*iris-versicolor*、*iris-virginica* という明確に区別されたクラスタがあります。この章の構成は次のとおりです。

- 「クラスタリングの概要」
- 「Tool Manager によるクラスタリングの起動」
- 「サンプル・クラスタリング・ファイルの使用」
- 「クラスタ・ビジュアライザのメインウィンドウ」
- 「クラスタリングを可視化する別の方法」

クラスタリングの概要

クラスタリングは記述的なデータマイニング操作なので、項目をラベルとして指定しません。そのため、クラスタリングは教師なし学習アルゴリズムに分類されます。クラスタリング・モデルの作成や評価には必ず全データセットを使用します。

MineSet は、類似レコードをグループ化してクラスタを形成し、グループ内の全体的な類似性を最大化しようとします。クラスタリングには、単一 *k*-means と反復型 *k*-means という明確に区別された 2 つのモードがあります。これらのモードについては、『*MineSet 3.0 Enterprise Edition Reference Guide*』のクラスタリングの説明を参照してください。

クラスタリングの実行が終わったら、クラスタ・ビジュアライザを使ってクラスタセンターを直接表示できます。これによってデータセット内の独立属性が表示されますので、最も顕著なものがどれで、それらがどのように異なるかを知ることができます。ただし、クラスタ間で属性が相互にどのように関連するかを見る場合は、スキャタ・ビジュアライザやデシジョン・テーブルの方が明確な結果が得られます。

クラスタ化したモデルをスキャタ・ビジュアライザを使って表示するには、重要項目 (Column Importance) ツールを使ってどの項目をどの軸に割り当てるかを決める必要があります。まず「モデルの適用」機能を使い、使用可能なモデルの中から `<datasetname>.cluster` を選択します。これについては、「[予測モデルの適用](#)」(145 ペー

ジ) の項を参照してください。これによって、新しい「クラスタ」項目が得られます。この項目をスキャタ・ビジュアライザの色に割り当てます。

Tool Manager によるクラスタリングの起動

クラスタリングには前提条件や必須オプションはありません。デフォルトで MineSet は、単一 k-means クラスタリング方法を使ってデータから 3 つのクラスタを見つけます。クラスタリングが終了すると、クラスタリングの評価が表示され、その後クラスタ・ビジュアライザが起動されます。

- 1. Tool Manager の「データの可視化／マイニング」パネルで、[図 12-1](#) のように「マイニングツール」タブ -> 「クラスタ」タブをクリックします。

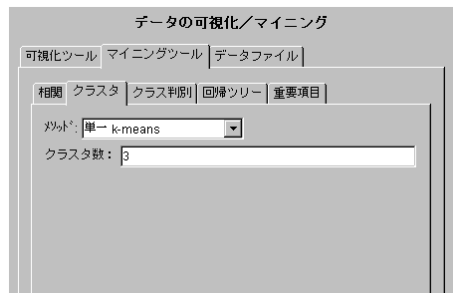


図 12-1 「クラスタ」タブ

- 2. クラスタリング方法を選択するには、「メソッド」ポップアップ・メニューから「単一 k-means」か「反復型 k-means」を選択します。(表 12-1)

表 12-1 クラスタリング方法のオプション

方法	説明
単一 k-means	デフォルトの方式。見つけるクラスタ数を指定します（デフォルトは 3）。
反復型 k-means	下限と上限、および「ポイントの選択」を指定する必要があります。このポイントは 0 から 1 の数字で、デフォルトは 0.5 です。

用語集の「ポイントの選択 (Choice Point)」を参照してください。

3. 「実行」をクリックします。

注記：クラスタリングには大量の計算が必要なため、特に反復型 k-means を使用する場合、大きなデータセットでは時間がかかりかかることがあります。データセットのレコード数が 10,000 を超える場合は、データの標本を使ってクラスタリングを行ってみてください。

クラスタ名には、1 から始まる数値が順に割り当てられます。クラスタ名は数値ですが、クラスタの順番はありません。

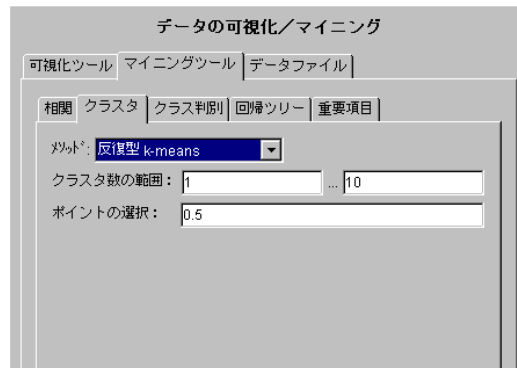


図 12-2 反復型 k-means を使ったクラスタリング

反復型 k-means では、クラスタ名は分割プロセスでの派生に基づいて付けられます。最小限のクラスタ数に基づく最初のクラスタリングでは、単一 k-means と同じように、名前は連続番号になります。クラスタが分割されると、新しい 2 つのクラスタの名前は、分割前のクラスタ名に「A」か「B」を付加したものになります。たとえば、クラスタ名「2-B-A」は最初のクラスタリングでクラスタ 2 から派生したもので、分割が 2 回行われています。

クラスタリング・アルゴリズムの詳細は、『*MineSet 3.0 Enterprise Edition Reference Guide*』のクラスタリングの説明を参照してください。

サンプル・クラスタリング・ファイルの使用

MineSet に付属している cars サンプル・データセットは比較的単純です。おなじみの馬力、車重量、60 mph に達するまでの時間などを扱っています。

Tool Manager の「ファイル」->「新しいデータファイルを開く」プルダウン・メニューから cars.schema ファイルを選択します。このファイルは、MineSet をインストールしたディレクトリの\data にあります。

- 1. Tool Manager の「データの可視化／マイニング」パネルで「マイニングツール」タブをクリックします。
- 2. 「クラスタ」タブをクリックし、「メソッド」を「単一 k-means」に、クラスタ数を 3 にしたまま、右下の「実行」をクリックします。

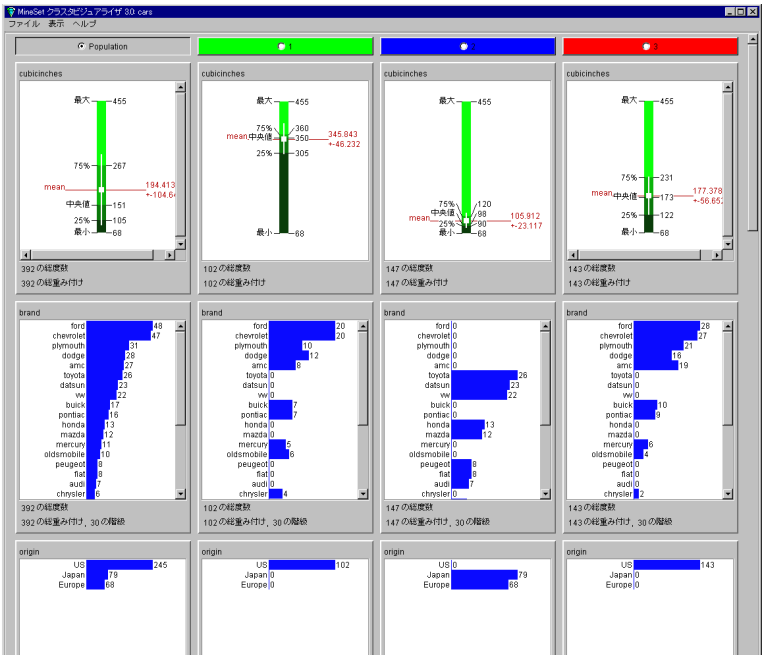


図 12-3 クラスタ・ビジュアライザのメインウィンドウ

- 3. 表示されるクラスタ・ビジュアライザのウィンドウ（図 12-3）で、クラスタ 1 を表す緑のバーをクリックします。このバーは、バーグラフやヒストグラムで表される属性の優先順序を制御します。クラスタ 1 をクリックした場合、属性の重要

度、つまり、ヒストグラムが表示される順序は、上から下へシリンダ数 (cylinders)、立方インチ (cubic inches)、馬力 (horsepower)、重量 (weight lbs)、燃費 (miles per gallon) です。属性の重要度は、この可視化の間だけこのクラスタが制御します。

- 異なるクラスタ間で同じ行を比較し、その属性がクラスタ間でどのように異なるかを調べます。
- クラスタ 2 を選択すると、属性の順序が前と違うことに気付きます。この場合は、原産国 (origin) が最も重要で、その後に車種 (brand)、立方インチ (cubic inches)、シリンダ数 (cylinders)、重量 (weight)、馬力 (horsepower)、燃費 (miles per gallon) が続きます。

クラスタ・ビジュアライザのメインウィンドウ

クラスタ・ビジュアライザには、ボックスプロットが縦横に並んで表示されます。行はデータの属性を、列はクラスタをそれぞれ表します。クラスタの名前は一番上のパネルにあります。「*Population* (個体群の形態)」の列はデータセット全体の統計値です。

「[Tool Manager によるクラスタリングの起動](#)」(212 ページ) の手順を実行すると、[図 12-3](#) のような画面が表示されます。クラスタには番号が順番に付けられています。アルゴリズムではクラスタが何を表すかは分からないため、ユーザの解釈が必要になります。クラスタ・ビジュアライザのウィンドウに属性が表示される順序には意味があります。デフォルトでは、属性はクラスタ間を区別する重要度の順に表示されます。

列の順序を変更するには、次の手順に従います。

- クラスタの名前が表示されているパネルをクリックして、そのクラスタを他のクラスタから区別する重要度の順に属性を再表示します。
- 「*Population* (個体群の形態)」のパネルをクリックして、デフォルトの順序を再表示します。

クラスタリングを可視化する別の方法

クラスタ・ビジュアライザは、クラスタリングを可視化する使い易い基本的なツールです。しかし、これが唯一の方法ではありません。クラスタリングの結果を変換してスキャタ・ビジュアライザで表示する方法が『*MineSet 3.0 Enterprise Edition Tutorial*』に示されています。簡単にいえば、この方法では、新しく作成されたクラスタリング・モデルを適用して新規項目を作成してから、任意のビジュアライザを使用します。クラスタリング結果の表示には、スキャタ、エビデンス、デシジョン・テーブルのビジュアライザが特に適しています。

相関規則分析によるデータの分析

相関規則分析 (Association Rule) ではデータの相関を分析します。つまり、どのデータ・インスタンスが他の特定のインスタンスといっしょに発生する可能性があるかを分析します。これは、同じ買い物かごにいっしょに入れられる商品は何かという意味でマーケット・バスケット分析と呼ばれます。この章の構成は次のとおりです。

- 「相関規則分析 (Association Rule Generator) と可視化の概要」(217 ページ)
- 「相関規則分析の起動」(221 ページ)
- 「スキャタ・ビジュアライザによる相関規則の説明」(224 ページ)
- 「ドリルスルー」(227 ページ)
- 「多重相関規則」(227 ページ)

相関規則分析の特徴と機能を示すために、設定ファイルとデータファイルのサンプルが用意されています。これらのファイルは、MineSet を最初にインストールした *MineSet 3.0\examples* ディレクトリにあります。

相関規則分析 (Association Rule Generator) と可視化の概要

スーパーマーケットの冷凍食品ケースに卵の代用品とワッフルがなぜ並んで置いてあるのか疑問を持ったことがあれば、おそらく相関規則が実際に使われている例を見たことになります。相関規則では、データに潜む関係やパターンを発見します。これは、すぐには識別できない相関を明確にし、可視化して表示する古典的なデータ主導型知識発見プロセスです。相関規則は、大きなデータセットからパターンを発見するときに特に役立ちます。このときには、データセット内に同時に存在する項目の頻度を示す相関規則を使います。相関規則は特定の 1 つの属性に依存するわけではないため、教師なし学習アルゴリズムです。相関規則では、すべての属性を同等に扱い、役立つ規則を全体から探索します。

相関規則の作成には 2 つの段階があります。まず、相関規則分析でデータファイル进行处理し、ビジュアライザで利用できるファイルを作成します。次に、このファイルをビジュアライザで表示します。この例では、スキャタ・ビジュアライザを使用します。

相関規則分析

相関規則分析では、1 対 1 の単純な相関規則と多重相関規則を生成できます。ここでは、単純相関規則について説明します。多重相関規則については、「[多重相関規則](#)」(227 ページ) を参照してください。

単純相関規則では、 X が真なら Y も真である一定の確率が存在すると仮定します。MineSet では、 X を規則の左辺 (LHS)、 Y を規則の右辺 (RHS) と呼びます。

たとえば、店に買い物にきた顧客が一度にいくつかの商品を購入するとします。この場合の規則は、「オムツを購入する客の 80% はベビーパウダーを購入する」であるかもしれません。この割合を規則の信頼度といいます。

この例では、オムツが左辺 (LHS) の商品、ベビーパウダーが右辺 (RHS) の商品です。この場合、

- ベビーパウダーが RHS に現れるなら、LHS はこの商品の売上げを伸ばすにはどうすればよいかの判断に役立ちます。
- オムツが LHS に現れるなら、RHS はオムツの販売を止めたらどの商品に影響があるかの判断に役立ちます。

相関規則分析では、入力ファイルを処理し、規則が入った出力ファイルを生成します。 X と Y がレコードの項目である場合、次の規則

$X \Rightarrow Y$

は、レコードに X があれば、ある頻度で Y が存在することを意味します。

相関の強さは、表 13-1 に示す 4 つの数字で定量化します。

表 13-1 相関規則の構成要素

測度	説明	統計的記述方法
支持度	LHS と RHS が同時に出現する頻度	$P(LHS \cap RHS)$
信頼度	LHS のすべての事象のうち RHS が同時に出現する割合、または指示度を LHS 項目の出現頻度で割ったもの	$P(RHS LHS)$
期待信頼度	RHS 項目の出現頻度	$P(RHS)$
改善	期待信頼に対する信頼度	$P(RHS LHS) : P(RHS)$

用語集の「支持度 (*Support*)」、「信頼度 (*Confidence*)」、「期待信頼度 (*Expected Confidence*)」、「改善 (*Lift*)」の各項を参照してください。技術的な詳細は、『*MineSet 3.0 Enterprise Edition Reference Guide*』の相関規則の説明を参照してください。

注記： Y と $X \Rightarrow Y$ 形式の規則だけが与えられても、X については何も分かりません。規則ではあくまでも LHS から RHS への意味合いを指定します。

規則の可視化

相関規則は、グリッド・ランドスケープにグラフィック形式で表示できます。左辺 (LHS) の項目が一方の軸に、右辺 (RHS) の項目が他方の軸にそれぞれ割り当てられます。図 13-1 に示すように、規則の属性が LHS と RHS 項目の交点に表示されます。表示には、バー、ディスク、ラベルが使用されます。

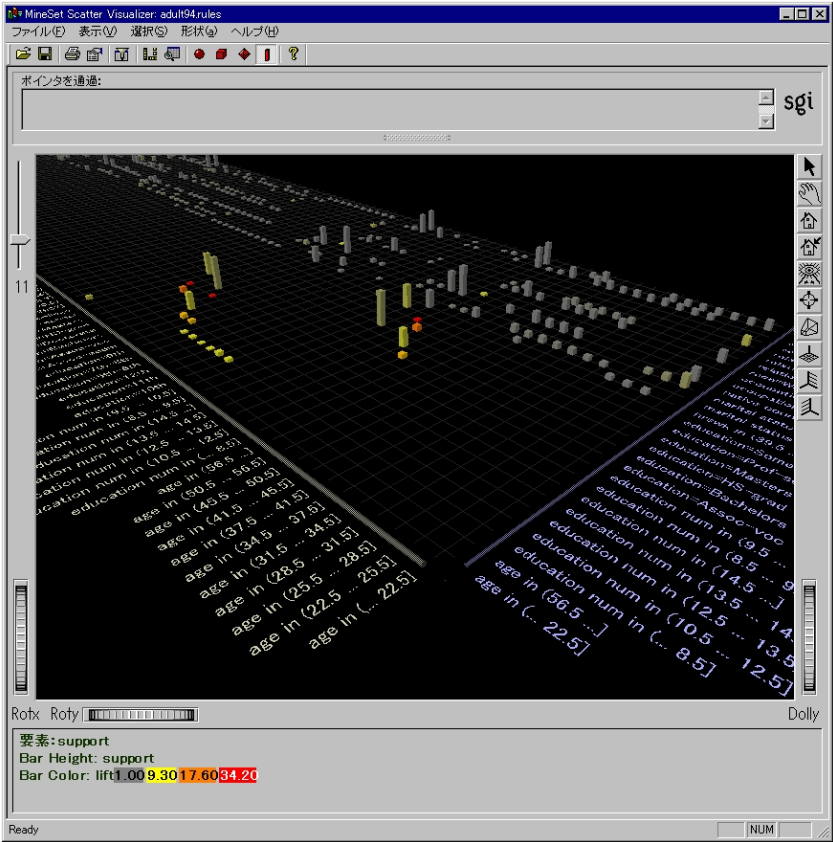


図 13-1 相関規則ビジュアライザのメインウィンドウの詳細表示

バーの高さや色などの表示されている属性と、信頼度や支持度などの元の規則に関連付けられている値とのマッピングを示す説明（凡例）がメインウィンドウ下部に表示されます。

相関規則分析の起動

この節では、Tool Manager を使って相関規則ツールを簡単に設定する方法を説明します。

相関の設定

この例では、cars データセットを使って簡単な相関を設定します。燃費 (miles per gallon)、馬力 (horsepower)、加速 (acceleration)、重量 (weight)、エンジンの大きさ (engin size)、原産国 (origin)、車種 (brand)、製造年 (the year car was built) などの間に相関を発見できるかもしれません。たとえば、燃費は改善したかとか、エンジンの馬力は減ったかなどです。

相関は、次の手順に従えば最も簡単に設定できます。

1. Tool Manager を起動し、サーバーを選択し、データソースとして cars.schema を選択します。「[MineSet の起動](#)」(15 ページ) を参照してください。
2. Tool Manager の「データの可視化／マイニング」パネルで「マイニングツール」タブをクリックします。並んでいるタブの列から「相関」(相関規則) を選択します。[図 13-2](#) が表示されます。

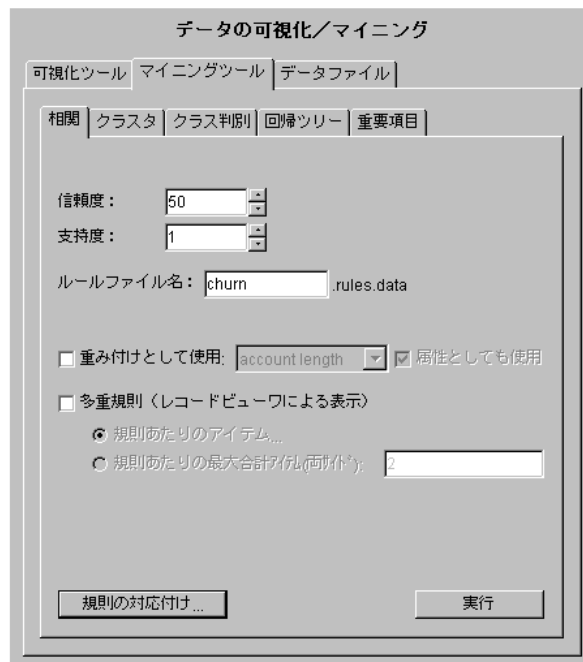


図 13-2 相関規則分析の最初の「Tool Manager」ウィンドウ

3. (オプション) 相関規則のアルゴリズムでは離散値を使用します。したがって、連続型の項目は階級が自動的に生成されます。これとは異なる階級を必要とする場合は、「データの可視化／マイニング」パネルの「階級の生成」ボタンを使用します。詳細は、「[項目に対する階級の変更または新規作成](#)」(37 ページ) の項を参照してください。
4. 「実行」をクリックして相関規則分析を実行するか、次の設定を行います。

信頼度 - 規則の最小信頼区間を指定します。信頼度がこの値より小さい規則は生成されません。デフォルトは 50%、有効な値は 1 から 100 です。

支持度 - 最小支持区間を、レコード総数に対する割合として指定します。支持度がこの値より小さい規則は生成されません。デフォルトは 1%、有効な値は 1 から 100 です。
5. (オプション) 相関規則オプションを選択したら、「規則の対応付け」ボタンをクリックして、支持度、信頼度、期待信頼度、改善などの規則属性を可視化要素にマッピングします。この後の「[相関規則の属性と可視化要素のマッピング](#)」の項を参照してください。

レコードの重み付け

あるレコードの重要度を他のレコードより高くしたり、標本抽出の不均衡を補正したりする場合には、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」ボックスをオンにしないと、すべてのレコードが重み1になります。このボックスは、重み付け可能な項目がデータセットにあるときだけアクティブになります。

レコードの重み付けを有効にするには、「重み付けとして使用」チェックボックスをクリックします。ポップアップ・メニューが表示されますので、レコードの重み付けに使用する項目を選択します。「属性としても使用」ボックスがオンになっていれば、相関規則分析が検出した規則の重み付け項目がこのボックスに表示されます。ボックスがオンになっていなければ、規則の判別においてすべてのレコードに同じ重み付けが与えられます。レコードの重み付けの詳しい説明は、第 3 章の「レコードの重み付け」の項を参照してください。

相関規則の属性と可視化要素のマッピング

相関規則ツールでは、規則の属性を画面の可視化要素にマッピングできます。「規則マッピング」ボタンをクリックすると、図 13-3 の「ルールビジュアライザの要素に規則の対応付け」パネルが表示されます。

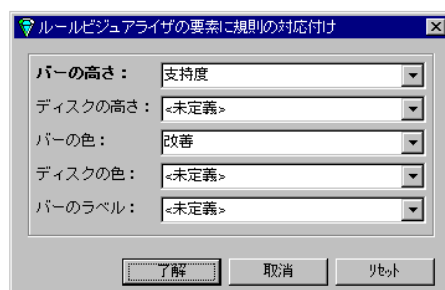


図 13-3 相関規則のマッピング

マッピングに使用できる可視化要素は次のとおりです。前にアスタリスクが付いている要素は省略可能です。

表 13-2 相関規則と可視化要素のマッピング

可視化要素	マッピング
バーの高さ	バーの高さで表すものを指定します。デフォルトは支持度 (Support) です。
* ディスクの高さ	ディスクの高さで表すものを指定します。
* バーの色	バーの色で表すものを指定します。デフォルトは改善 (Lift) です。
* ディスクの色	ディスクの色で表すものを指定します。
* バーのラベル	バーのラベルで表すものを指定します。

スキャタ・ビジュアライザによる相関規則の説明

相関規則ツールは、有効な設定ファイルの指定に従って、規則ファイルのデータをスキャタ・ビジュアライザに表示します。どのようなサンプルファイルがあるかを見るには、MineSet 3D ビジュアライザの「ファイル」->「開く」を選択して設定ファイルのリストを表示します。これらのファイルは、MineSet を最初にインストールした *MineSet 3.0\examples* ディレクトリにあります。たとえば、*brand.rules.scatterviz* を指定すると、[図 13-4](#) に示すイメージが表示されます。相関規則のファイルは、名前の後ろに *.rules.scatterviz* が付きます。

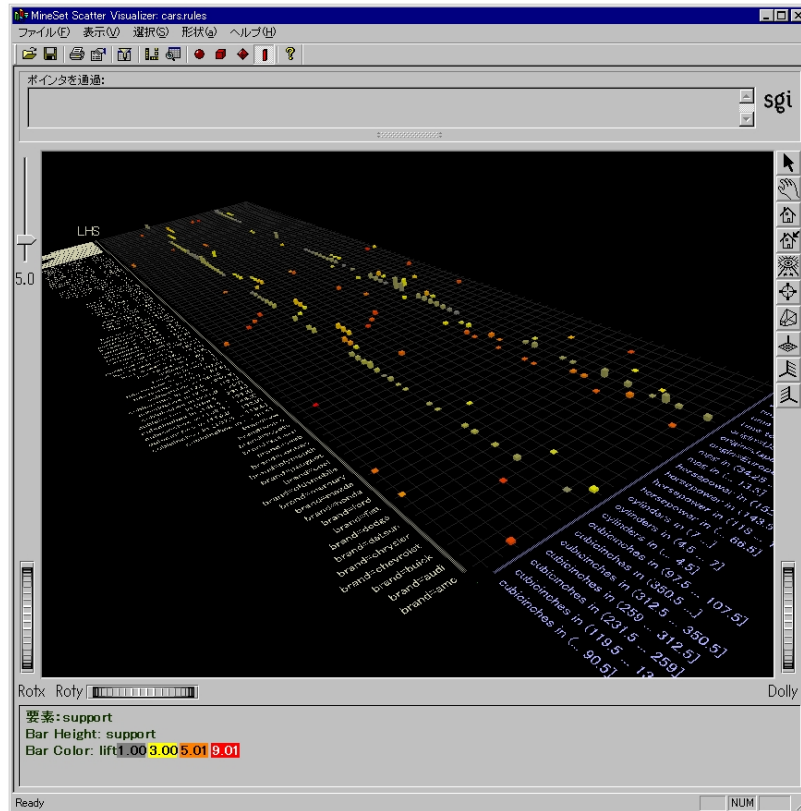


図 13-4 brand.rules.scatterviz を指定したときの最初の相関規則の表示

相関規則は、スキャタ・ビジュアライザのグリッド・ランドスケープに表示されます。左辺 (LHS) の項目が一方の軸に、右辺 (RHS) の項目が他方の軸に割り当てられます。図 13-4 のように、規則の属性が LHS と RHS 項目の交点に表示されます。たとえば、バーの高さが支持度に、バーの色が改善にそれぞれ対応します。表示が小さすぎると、各軸の横の項目ラベルは表示されません。ウィンドウ右下にある「Dolly」ダイヤルを使えば、項目ラベルが表示されるまで表示を拡大できます。さらに、マウスが選択モードのときにマウスポインタを個々のバーの上に移動すれば、特定の規則のラベルを表示できます（図 13-5 を参照）。その規則の詳しい情報が表示領域の左上隅に表示されます。

たとえば、図 13-4 では、バーの高さが支持度に、バーの色が改善にそれぞれ対応しています。

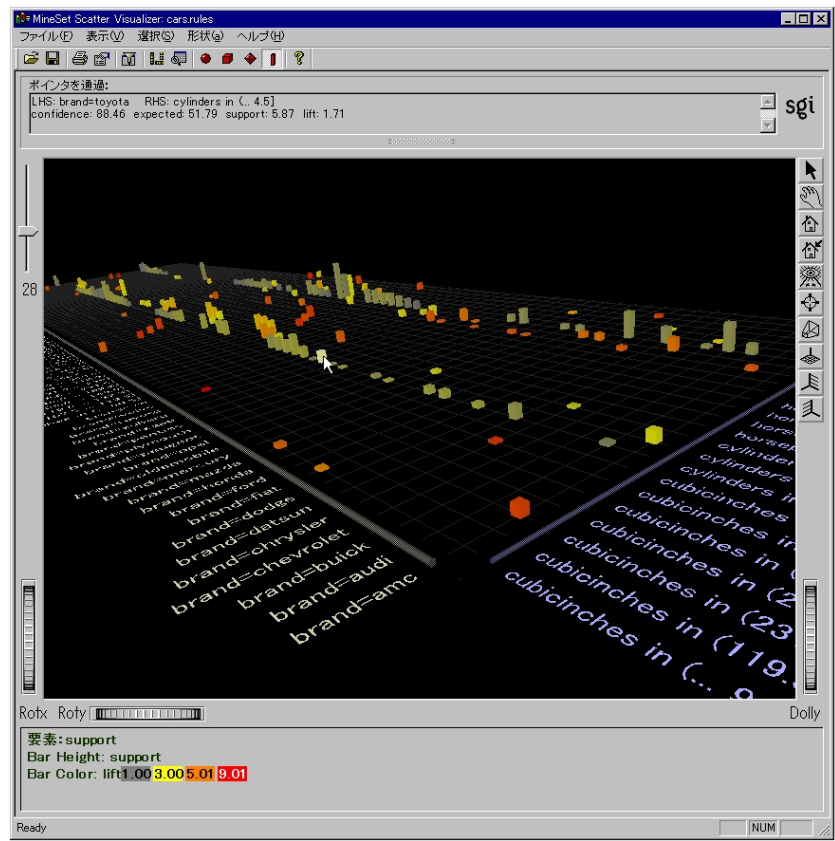


図 13-5 相関規則を表すバーの上にカーソルがある状態

図 13-5 のように、相関規則オブジェクトの上にマウスポインタを移動すると、オブジェクトの情報が表示されます。情報は、ポインタがオブジェクト上にある間だけ表示されます。ポインタがオブジェクト上にあるときに左ボタンをクリックすると、同じ情報がメインウィンドウの上の選択領域に表示されます。さらに、そのバーが選択され、選択されたすべての規則が別のウィンドウに表示されます。複数の規則を選択するには、<Shift> キーを押したままクリックを続けます。

この情報は、他のオブジェクトを選択するか、黒い背景をクリックしてオブジェクトを選択しない限り表示されます。選択ウィンドウのテキスト情報は、カット&ペーストを使って、レポートやデータベースなど他のアプリケーションに取りこむことができます。

ドリルスルー

ドリルスルー表現は、選択された規則の論理積によって決まります。オリジナル・テーブルの項目と *.rules.data* ファイルの項目は一致しないため、相関規則分析では、ドリルスルーが行われたときにフィルタ表現を作成するための特別な項目が生成されます。つまり、*.rules.scatterviz* ファイルでは特別な文字列値項目がすでにドリルスルーにマッピングされているため、ドリルスルー設定パネルを変更しても効果はありません。

1 つまたは複数の規則に対しドリルスルーを行うと、その規則を満足するすべてのレコードが表示されます。

多重相関規則

場合によっては、LHS や RHS に複数の項目を使用する複雑な規則の方が役立つことがあります。このような規則を多重相関規則といいます。[図 13-6](#) に、多重相関規則の生成用に設定した Tool Manager の「相関」パネルを示します。

「多重規則」チェックボックスをオンにすると、LHS や RHS に複数の項目を持つ規則など、支持度と信頼度の最小区間を満たすすべての規則が生成されます。このような規則は、たとえば、「beer and linguini implies potato chips and salsa and wine（ビールとリングィーニ（パスタの一種）はポテトチップとサルサとワインを暗示する）」のような場合です。

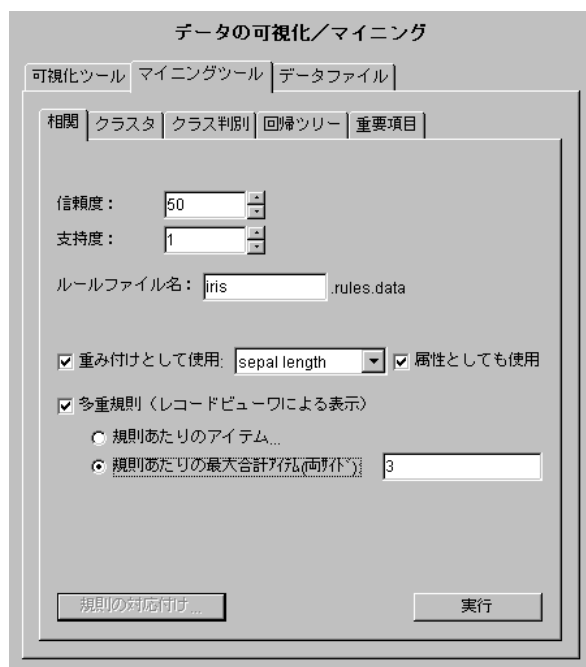


図 13-6 多重相関生成を設定する最初の「Tool Manager」ウィンドウ

多重規則はレコードビューワを使って表示します。スキャタ・ビジュアライザには、このような複雑な規則を表示する方法はありません。多重規則は、1 行に 1 規則ずつ表示されます。テーブルの最初の 2 つの項目には LHS と RHS の項目数が表示されます。次の 4 つの項目には、支持度、信頼度、期待信頼度、改善の各値が表示されます。最後の 2 つの項目には LHS と RHS の項目が表示されます。LHS と RHS 項目では、項目が「and」で区切られます。先ほどの規則の例では、2 つの項目がある LHS は「beer and linguini」で表されます。3 つの項目がある RHS は「potato chips and salsa and wine」で表されます。

生成される規則のサイズを制限するには、「規則あたりの最大合計アイテム (*Max total items per rule*)」フィールドに数字を入力します。この数字は、規則に組み込む項目の最大数です。規則の項目数は LHS と RHS の項目数を合計したものです。先ほどの規則には 3 つの項目があります。単純な規則では項目数は 2 です。

注記：多重規則の生成には長い時間がかかることがあります。反復のたびに生成される規則の数が Tool Manager のステータス・ウィンドウに表示されますので、それに注意してください。生成される規則数が多すぎるようなら、操作をキャンセルしてから支持度や信頼度の最小区間を増やすか、規則当たりの最大項目数を減らします。

アルゴリズム (Algorithm)

何らかのタスクを行うための正式な手順。たとえば、分析アルゴリズムは一連のマシン命令からなり、クラシファイアや回帰ツリー分析モデルの作成方法をコンピュータに指示する。

インスタンス (Instance)

レコードまたは行のこと。

エビデンス (Evidence)

確率推定が最終的な予測に与える影響の程度。エビデンス・クラシファイアは、各属性のエビデンスを合計してラベルの値を予測する。正式には、エビデンスとは正規化された条件付確率の負の対数である。

エントロピ (Entropy)

データの不規則性の程度を示す情報理論の測度。データセットのエントロピが高ければ、その値の多様性や混合度はそれだけ高い。MineSet のマイニング・アルゴリズムの多くは、エントロピを最小化するためにデータを分割する。

重み (Weight)

MineSet では、2 種類の重み付けがサポートされる。

レコード重みは、データの各行に適用される数値である。マイニング・ツールは、重み k の行をその行が k 行あるかのように扱う。非整数値はサポートされない。重みの値は、ユーザが選択するデータ項目で指定する。

属性の重みは、ユーザが指定する項目の重要度を示す測度である。属性の重みは、クラスタリング・アルゴリズムが距離を決めるときに使用する。

回帰モデル (Regressor)

ラベルが連続値をとる予測モデル。クラシファイアとよく似ている。

階級自動生成 (Automatic Binning)

それぞれの階級内のラベルの分割ができるだけ異なるように階級区間を選択するアルゴリズム。正式にいうと、このアルゴリズムは各階級内のエントロピを最小化する。「階級生成 (Binning)」を参照。

階級生成 (Binning)

連続データを離散グループに分割し、実データをカテゴリデータに変換すること。たとえば、ある範囲の連続する年齢を 0-18、19-25、26-35 などの階級に分割すること。グループまたは階級は、重なり合わない一連の連続体領域として定義される。

改善 (Lift)

信頼度と、相関規則分析によって生成された規則の期待信頼度との割合。通常、この数値が高ければ、規則はそれだけ興味深いものである。

改善曲線 (Lift Curve)

モデルの予測精度を評価する 1 つの方法。改善曲線には、クラシファイアを使用した場合の利点が、ラベルをランダムに判定した場合との比較で可視化される。

仮説検定 (Hypothesis Testing)

仮説とは提案された解釈である。ただし、解釈の有効性は検定できなければならない。仮説検定は、事前の考えを検証または反証するトップダウン・アプローチである。仮説の作成では、人は観察された事象のありそうな解釈を考え出す。この拡張として、仮説に基づき、分析するデータが選択される。

規則 (Rules)

相関規則やクラス判別規則など、データの処理に適用するために前もって決められた動作パターン。これは、レコードをクラス判別するための、ルートからリーフに至る決定木の固有パスでもある。規則ベースのシステムは、与えられた手順に従って一連の「if-then」規則を一連の事実に適用することによって推論を行う。

期待信頼度 (Expected Confidence)

RHS 項目がデータセットに出現する頻度。したがって、期待信頼度と信頼度の違いは、LHS 項目が存在することによる予測力の変化を示す測度にある。期待信頼度は、項目の間に関係がなかったら信頼度がどうなるかを示す。

行 (Row)

リレーショナル・テーブルのレコード

均一な重み付け (Uniform Weight)

データの階級自動生成における選択枝の 1 つ。値の範囲は指定された数の階級に分割され、各階級には同じ数のレコードが割り当てられる。レコードの重み付けが有効になっていると、この範囲は、各階級の合計の重み付けが等しくなるように分割される。

均一な範囲 (Uniform Range)

データの階級自動生成における選択肢の 1 つ。値の範囲は同じ大きさの間隔に分割される。

クラシファイア (Classifier)

他の項目（属性）との関連で、ある項目（ラベル）を説明しようとする予測モデル。クラシファイアはラベルがすでに分かっているデータから作成される。その後でこのモデルを使用すれば、ラベルがわかっている新規データのラベルの値を予測できる。内部的に言えば、クラシファイアとは、入力行ごとに 1 つの離散値を予測するアルゴリズムまたは数式である。たとえば、アヤメのデータセットから作成したクラシファイアは、花びらと雄しべの長さや幅を与えられれば、そのアヤメの種類を予測できる。さらに、クラシファイアは、ラベルの各値の確率評価を生成できる。たとえば、自動車のデータセットから作成されたクラシファイアは、特定の自動車が米国で製造された確率を予測できる。このクラシファイアでは離散型ラベルしか使用できない。連続型ラベルを扱うアナログ版のクラシファイアとして回帰ツリー (Regressor) がある。

訓練事例 (Training Set)

MineSet では、本格的なデータ操作を行う前に分離された、クラシファイアやモデルを作成するためのデータのサブセット。記述属性に基づいてラベルがすでに判別されているデータベース・レコードからなり、クラシファイアの作成方法を学習するために使用される。「テストセット (Test Set)」も参照。

ケーキグラフ (Cake Charts)

可視化され、セグメント化されたブロックで、データの分布を表す。円グラフに似ているが、ケーキグラフは四角形をしている。

決定木 (Decision Tree)

分岐した一連のテストからなる予測モデル。各テストではデータの 1 つの項目の値を調べ、それを使って、次に行うテストを決める。すべてのテストの結果によって、予測するラベルが決まる。

誤差率 (Error Rate)

データセットに対してモデルが行う正確または不正確な予測率。通常、誤差率は、学習プロセスではいっさい使用されなかった独立のテストセットを使って決められる。特に、少数のインスタンスからなるデータセットの場合には、相互検証など、より複雑な精度評価技法が使われる。

支持度 (Support)

関連規則 $X \rightarrow Y$ (X が Y を内包する) が与えられた場合に X と Y がファイルに同時に出現する割合。

現する頻度をレコード総数に対する割合として定量化したもの。たとえば、支持度が 1% なら、レコード総数の 1% で X と Y が同時に出現する。支持度は「普及率 (Prevalence)」と同義である。

事前確率 (Prior Probability)

クラスラベルの事前確率は、すべての属性値を無視し、ランダムに選択したレコードのデータにこのラベルが存在する確率である。数学的には、これは、そのクラスラベルを持つレコード数を合計のレコード数で割ったものである。「条件付き確率 (Conditional Probability)」も参照。普及率 (Prevalence) は、MineSet の前のバージョンでサポートと呼ばれていたものである。

集計処理 (Aggregation)

指定されたいくつかの項目に対し 1 つまたは複数の操作を行うことによって複数行のデータを集計する処理。集計処理操作では、サマリ統計の作成、テーブルの置換え、さらには元データに基づく配列の作成を行うことができる。

集計処理をする (Aggregate)

データをいろいろな方法で結合する集計処理を行うこと。上記の「集計処理 (Aggregation)」を参照。

寄与率 (Purity)

主成分抽出 (Column Importance) との関連でいう寄与率とは、ラベル値分布の正確性を示す測度である。累積寄与率とは、それらのラベルの値に関してデータがどのくらい適切に分割されているかを示す測度である。データは、決定木における分割と同じように、重要項目を使って分割される。パーティション内の各セットには独自の寄与率測度があり、パーティション内の寄与率測度はこれらの個々の寄与率測度を組み合わせたものである。パーティション内の 1 つのセットで、レコード全体に各クラスが同等に現れるなら寄与率 0、すべてのレコードが同じクラスなら寄与率 100 である。同じように、パーティション内の各セットで、クラスが同等に現れるなら累積寄与率 0、すべてのレコードが同じクラスなら累積寄与率 100 である。MineSet では、寄与率はエントロピーに基づいている。

条件付き確率 (Conditional Probability)

事象 (B) が起きたときの事象 (A) の確率。 $P(A|B)$ と書き、「 B が与えられたときの A の確率」と読む。条件付確率は、回帰ビジュアライザの左側ウィンドウに四角形グラフで表示される。このグラフには、各ラベルの値が与えられたとき (条件) の各属性値の相対的な確率が表示される。条件付き確率は、与えられたラベルの値のエビデンス (evidence) であるといえる。

信頼度 (Confidence)

相関規則 ($X \rightarrow Y$) の信頼度。 X と Y が同時に出現する頻度を、 X が出現するレコード数に対する割合として定量化する。たとえば、信頼度が 50% なら、 X が出現するレコードの 50% で X と Y が同時に出現する。したがって、あるレコードに X が存在するなら、そのレコードには 50% の確率で Y が同時に存在する。

スキーマ (Schema)

項目の名前やデータ型の情報など、データセットのすべての項目を記述したもの。スキーマは `.schema` 接尾辞のファイルに格納される。

設定ファイル (Configuration File)

`.data` ファイルのデータをビジュアライザ要件にどのようにマッピングするかを指定するファイル。MineSet 設定ファイルの接尾辞には、`.treeviz`、`.scatterviz`、`.splatviz`、`.mapviz`、`.eviviz`、`.dtableviz`、`.statviz`、`.clusterviz` がある。

選択ポイント (Choice Point)

ゼロと、たとえば、クラスタ数の選択を助ける値との間にあるポイント。選択ポイントが高ければクラスタ数が多くなり、低ければ小さくなることを示す。選択ポイント 1.0 では常に上限が選択される。クラスタリングの場合、下限と上限が 1 クラスタと 5 クラスタなら、選択ポイント 0.4 では 2 つのクラスタが、0.8 では 4 つのクラスタがそれぞれ選択される。選択ポイント 1.0 では、常に 5 つのクラスタが選択される。

相互検証 (Cross-Validation)

予測誤差を推定する方法。データセットを、層と呼ぶ k 個（通常は 10）の同じサイズの部分に分割する。 k 個の予測モデルを作成し、各モデルを個別の層でテストする。ただし、このテストの前に、残りの層を使ってモデルに学習させる。この処理を繰り返せば、推定の信頼性が向上する。

属性 (Attribute)

予測モデルの入力として使用される項目。予測モデリングでは、データセットの項目は、入力となる属性と、モデルの出力を表すラベルに分けられる。

知識発見 (Knowledge Discovery)

データに潜む新しく、興味深い、有益なパターンを発見する処理。データマイニングは知識発見のサブセットであり、データから新しい仮説を引き出し、検定する。

中央値 (Median)

数値を大きさの順に並べたときの中央の値。

データセット (Dataset)

行と項目からなるデータのテーブル。行をレコードやインスタンスと呼ぶこともある。MineSet CD には、いくつかのサンプルデータセットが含まれている。たとえば、電話会社に興味がある人向けの *churn*、非常に興味深い *cars*、米国人口統計局からの抜粋情報に基づく *adult* などがある。データセットは、あるスキーマに合致するインスタンスからなるスキーマおよびセットともいえる。通常、インスタンスの順序は考慮されない。

データのクリーニング (Data Cleaning)

たとえば、間違っているデータ値の削除や変更など、データの形式や内容を変更してデータの質をよくなる処理。

データファイル (Data File)

MineSet のデータファイルは元データがタブで区切られたフラットファイルで、モデルや可視化はこのファイルを使って行う。MineSet では、データファイルの接尾辞に *.data* を使用する。データファイルは、小さくて処理が速いバイナリ形式でも、人間が判読可能な ASCII 形式でもよい。

テーブル (Table)

正規化されたリレーショナル・データセット。

テストセット (Test Set)

クラシファイアを作成した後にその誤差率をテストするためにデータセットから別に保存しておくレコード・セット。「訓練事例 (Training Set)」も参照。

投資利益率 (ROI)(Return on Investment)

投資から得られる利益によってプロジェクトの価値を測定する金融用語。ROI ともいう。

ドリルスルー (Drill-through)

可視化に使用したオリジナルデータの一部を取り出す行為。

配列 (Array)

複数の項目を格納できるデータ型。配列型は集計処理によって作成され、ビジュアライザによって使用されるもので、アニメーションのデータを格納する。MineSet では、ベクターと呼ぶ 1 次元配列とマトリックスと呼ぶ 2 次元配列がサポートされる。

端数の切捨 (Trimming Fraction)

階級生成の前に値の範囲から除外する極端な値の割合。デフォルトの端数切捨割合は

0.05 である。この場合、極端な値を持つインスタンスが、最も小さいものから順に 2.5%、最も大きいものから順に 2.5%、合計 5% 除外される。端数の切捨を行うと、通常、区間の生成に対する外れ値の影響が減る。

バックフィッティング (Backfitting)

小さい標本から生成したモデルにデータセット全体を適用すること。適用後のモデルは、元の構造を保持する。バックフィッティングの目的は、モデルの誤差推定を維持しながら、モデルの内部分割がデータセット全体を正確に反映することにある。

範囲 (Range)

項目がとり得る最大値と最小値の差。

標準偏差 (Standard Deviation)

データの分散を示す測度。分散の平方根として定義される。

フィルタ (Filter)

MineSet では、通常、項目の内容の選択や操作を行うためのブール表現。たとえば、「`m.p.g < 25`」では、ガロン当たりの燃費が 25 マイル未満のエントリだけが選択される。つまり、この表現が「真」に評価されるレコードだけが選択される。フィルタとは、データセットレコードのサブセットを指定する一連の選択された条件を保存したものである。

普及率 (Prevalence)

支持度 (Support) を表す別の用語。これは、MineSet の前のバージョンでサポートと呼ばれていたものである。

分析 (Inducer)

訓練事例からモデルを作成するアルゴリズム。MineSet では、4 つの分析（決定木分析、選択式決定木分析、回帰、デシジョン・テーブル）と 1 つの回帰ツリー分析（回帰ツリー）がサポートされる。

平均値 (Mean)

平均値とは、サンプルデータの値を加算し、それをデータの個数で除算したもの。平均は通常 $\bar{x}$ - バーで表し、一般に「平均値 (average)」と呼ばれる。たとえば、 $(1+2+3)/3$ の平均は 2 である。

モデル (Model)

現象を観察し、それを記述したもので、ある種の詳細は省略されることがある。モデルには予測が含まれていることがある。たとえば、ダイレクトメールのモデルが作成された場合、このモデルは「ダイレクトメールの顧客はこのような人たちであると考え

られる」ことを暗示する。ユーザは、値を変更すれば、システム操作に対する変更の影響を観察できる。「クラシファイア (Classifier)」も参照。

歪み (Skewness)

度数分布における調和の欠如。

予測可能性 (Predictability)

信頼度 (Confidence) を表す別の用語。

予備法 (Holdout)

モデルの学習に使用するデータセット部分（通常は3分の2）。残りはテストのために保存される。予備法は、MineSet がサポートする最も簡単な誤差推定法である。

ラベル (Label)

予測モデルが予測しようとする単一の項目。ラベルは、前もってユーザが選択しておく必要がある。たとえば、iris データセットでは、「アヤメの種類 (iris type)」はラベルのよい候補となる。

ランダムシード (Random Seed)

ランダムな標本データの選択を開始する特定のポイントを選択する方法。異なるランダムな標本が必要なら異なるランダムシードを指定する。同じデータセットの別々のマイニング探索で同じシードを使用すると、つねに同じランダムな標本が使用される。探索で発見された特定のモードの安定性をテストしたい場合は、シードを変更する。

離散型属性 (Discrete Attribute)

有限数の個別値を持つ項目。MineSet では、文字列、階級値、日付が離散値とみなされる。クラシファイアのラベルとして使用する場合は、整数値も離散値とみなされる。

リレーショナル・データベース (Relational Database)

データウェアハウスの中心にあるもの。データと、データ間の関係はテーブルに編成される。テーブルは、各エントリが同じフィールドをもつレコードの集合である。一部のフィールドがキーに指定されるため、キーフィールドの値をインデックスとしてデータを高速に検索できる。異なるテーブルの一定のフィールドが同じ値を持っていれば、それらのテーブルをリンクできる。この例としては Sybase、Informix、OLEDB、SQL Server、Oracle などがある。

連続型属性 (Continuous-valued Attribute)

ある範囲全体に渡る連続値をとることができる属性。MineSet では、データ型が整数、倍精度、浮動小数点の項目が連続値とみなされる。

数字

2D の集計処理, 75, 80
2 次元の集計処理, 75, 80
3D グラフ, 66, 68
3D ランドスケープ, 89
3 次元グラフ, 66, 68
3 次元ランドスケープ, 89

A

API, 15

F

F1 ヘルプ, 32

G

「Gaussian」コマンド, 87

K

K-means
 単一, 212
 反復型, 212

M

.marks ファイル名接尾辞, 106
MineSet
 起動, 15
MineSet API, 15
MineSet ツール, 7-8

O

OLAP とデータマイニング, 2

R

ROI 曲線, 206

T

Tool Manager, 18
 項目の階級生成, 37-41
 項目の型の変更, 44-46
 項目の削除, 35
 項目の追加, 36-37
 項目名の変更, 44-46
 設定オプション
 相関規則, 221, 224
 相関規則分析ウィンドウ, 222
 多重規則
 図, 228

ツールオプション

 スキヤッタ・ビジュアライザ, 73

データの標本抽出, 47-48

データのフィルタ, 43-44

データ変換, 34-48

データ変換パネル, 34

ボタン, 34

モデルの適用, 47

「Tool Manager に送信」 コマンド, 85, 86

あ

「値の表示」 コマンド, 86, 84

アニメーション, 110

アニメーション・コントロール・パネル (スキヤッタ・ビジュアライザ)

 サマリウィンドウ, 75

アニメーション・コントロール・パネル (スプラット・ビジュアライザ)

 サマリウィンドウ, 80

アニメーション・コントロール・パネル (マップ・ビジュアライザ), 123

アプリケーション・プログラミング・インタフェース, 15

い

移動の軌跡

 スキヤッタ・ビジュアライザと, 76

 図示, 78

移動ボタン, 105

色

 決定木, 157

印刷マニュアル

 表記上の決まり, xx

う

「上へ」 ボタン, 106

え

エビデンス

 実行, 155, 167

エンドポイント, 112

お

オブジェクト

 選択

 スプラット・ビジュアライザ, 72

 ツリー・ビジュアライザ, 98, 99, 102

 マップ・ビジュアライザ, 125

 ツリー・ビジュアライザ, 102

重み付け、レコードの, 50-51

「オリジナルデータの表示」 コマンド, 85, 86

オンラインヘルプ, 32

か

回帰, 5

 概要, 128-130

回帰ツリー

 概要, 135-136

回帰ツリー分析

 概要, 154

 誤差評価, 162

階級生成

 均一な区間設定, 41

 指定, 40

階級生成、項目の, 37

解析データマイニング, 2-6

改善, 219
改善曲線, 204-206
階層, 89
格納
 データの場所, 103
下降ノード, 94
可視化、データマイニング, 2, 6
関係の分析, 56, 89
概要ウィンドウ (ツリー・ビジュアライザ), 101
学習曲線
 使用, 141-145
「学習曲線」モード, 141-145

き

期待信頼度, 219
起動
 MineSet, 15
 クラスタリング, 212
 統計量ビジュアライザ, 24
 ヒストグラム・ビジュアライザ, 25
 レコードビューワ, 19
キノコ分析データセット
 混同マトリックス, 197, 203
教師付きモデリング, 3-5
 精度の評価, 5
教師なしモデリング, 5-6
強調表示、オブジェクトの
 スキヤタ・ビジュアライザ, 84
 スプラット・ビジュアライザ, 72
 ツリー・ビジュアライザ, 98
 マップ・ビジュアライザ, 125
 スキヤタ・ビジュアライザ, 85
寄与率, 157
寄与率を測る, 157
切り取り、選択情報の, 72

く

区間
 階級指定, 40
 階級生成、均一な区間設定, 41
クラシファイア
 混同マトリックス, 196-200
 損失マトリックス, 201-203
 レコードに適用, 194
「クラシファイアとエラー」モード, 139-140
「クラシファイアのみ」モード, 137-139
クラシファイアのモード, 137
 学習曲線, 141-145
 誤差評価, 140-141
 クラシファイアとエラー, 139-140
 クラシファイアのみ, 137-139
クラス
 指定, 156
クラスタ・ビジュアライザ, 211-215
クラスタリング, 6, 211-215
 起動, 212
 ビジュアライザ, 215
 方法, 212
クラス判別規則, 158

け

傾向, 58
決定木
 概要, 130-131
 寄与率を測る, 157
 誤差 / 損失評価, 157
 ノード
 情報の表示, 157
 ノード情報の表示, 157
 表示, 160
 分析規則, 158
 レコードの分析, 157-158

検索スポットライト, 102

こ

項目

階級生成, 37

削除, 35

重要, 51-54

タイプの定義, 46

タイプの変更, 44

追加, 36

名前の変更, 46

「項目のドリルスルー」コマンド, 85

子ノード

選択式決定木と, 160

混同マトリックス, 196-200

誤差 / 損失評価, 157

誤差評価, 137-141

回帰ツリー, 162

誤差推定オプション

平均絶対誤差, 162

平均平方誤差, 162

「誤差推定」モード, 140-141

誤差率 (選択式決定木分析), 160

さ

最小支持区間, 222

最小信頼区間, 222

削除、項目の, 35

削除ボタン, 106

削除、マークの, 106

作成、モデルの, 11

「サブツリー重み」オプション, 162

サマリウィンドウ (スキャタ・ビジュアルライザ), 75

サマリウィンドウ (スプラット・ビジュアルライザ), 80

サマリ値, 95

し

識別、データの, 9

支持度, 219, 222

最小区間, 222

「下へ」ボタン, 106

集計処理, 41-43, 95

2次元, 75, 80

データポイント, 57

修正、マークの, 106

使用、学習曲線の, 141-145

使用、モデルの, 11

信頼度, 218, 219, 222

最小区間, 222

準備、データの, 9

す

推定、誤差の, 137-141

推定、モデルの, 5, 11, 137

スキャタ・ビジュアルライザ, 55-88

アニメーション・コントロール・パネル

サマリウィンドウ, 75

移動の軌跡, 76

図示, 78

オブジェクトの選択, 84, 85

オプション, 73

概要, 55

データの表示, 75

表示モード, 70

メインウィンドウ, 69-71

メニュー, 82-87

スキャタ・プロット, 58

スプラット
描画オプション, 87
定義, 57
スプラットタイプ・メニュー, 87
スプラット・ビジュアライザ, 55-88
情報の取得, 72
アニメーション・コントロール・パネル
サマリウィンドウ, 80
オブジェクトの選択, 72
概要, 57
データの表示, 80, 87
データの分析, 58
データポイントの集計処理, 57
メインウィンドウ, 69-73
メニュー, 82-88
スポットライト, 100, 102
スライダ
作成, 74

せ

整形、データの
Tool Manager, 33-48
データの整形の必要性, 33
生成、相関規則の, 218
生成、モデルの, 128
精度
評価, 5, 11
設定、相関規則の, 221, 222
Tool Manager と, 221-224
「線形」コマンド, 87
選択、オブジェクトの
スキッタ・ビジュアライザ, 84, 85
スプラット・ビジュアライザ, 72
ツリー・ビジュアライザ, 98, 99
概要ウィンドウ, 102
マップ・ビジュアライザ, 125
ツリー・ビジュアライザ, 102
選択式決定木

概要, 131-132, 150-153
決定木との比較, 153
誤差率, 160
選択式決定木の表示, 160
選択ノード, 152
定義, 152
ランク, 160
選択メニュー
スキッタ・ビジュアライザ, 84
マップ・ビジュアライザ, 125
選択モード
スプラット・ビジュアライザ, 72
選択、モデルの, 145

そ

相関, 5
相関規則, 217-229
改善, 219
期待信頼度, 219
支持度, 219, 222
最小区間, 222
信頼度, 218, 219, 222
期待, 219
最小区間, 222
生成, 218
オプションの設定, 221-222
設定
Tool Manager と, 221, 224
相関の設定, 221-222
多重, 227
多重規則
表示, 228
データ変換, 221-222
ドリルスルー, 227
にデータをマッピング, 223
表示, 219, 224-226
図, 223, 224
マーケット・バスケット, 218
レコードの重み付け, 223

相関規則ビジュアライザ

図, 220

メインウィンドウ, 220

相関規則分析, 217-229

Tool Manager と, 222

オプションの設定, 221-222

出力, 218

表示の説明, 220

相関規則マッピングパネル, 223

損失マトリックス, 201-203

た

多重規則, 227

Tool Manager と, 228

表示, 226

単一 K-means, 212

単純な輪郭線, 111

ち

注釈、データポイントの, 103

つ

追加、項目の, 36

ツリー・ビジュアライザ

オブジェクトの選択, 98, 99, 102

概要, 89

決定木と, 160

視点のマーク, 103, 106

情報の取得, 98

情報にスポットライトを当てる, 100, 102

情報の取得, 100, 102

メインウィンドウ, 90

て

「定数」コマンド, 87

テーブルの履歴, 48-50

適用、モデルの, 11, 47, 145-147

「テキストチャ」コマンド, 87

データ

識別, 9

準備, 9

データ型、項目の, 46

データセット

データの表示, 111, 112

3D ランドスケープ, 89

アニメーション・コントロール・パネル, 75-81

特定値の検索, 102

「データの可視化／マイニングパネル」

クラスタ, 213

データ変換, 34-48

データ変換パネル, 34

データポイント

集計処理, 57

注釈, 103

データマイニング, 2

および OLAP, 2

解析, 2-6

可視化, 2, 6

調査, 8-12

方法, 2

デシジョン・テーブル

概要, 133-134

と

統計量ビジュアライザ, 22-25

起動, 24

ヒストグラムとは, 23

ボックスプロットとは, 22

特定値の検索, 102

閉じるボタン

フィルタ・ダイアログ, 106

ドリルスルー

相関規則と, 227

な

ナビゲーション、ビジュアライザでの, 27-32

名前

視点, 103

の

ノード, 94

決定木

情報の表示, 157

選択式決定木, 160

は

発見、重要項目の, 51-54

貼り付け、選択情報の, 72

反復型 K-means, 212, 213

凡例

相関規則, 220

バー

決定木と, 157

負の値と, 95

場所

マーク, 103-106

バックフィッティング, 194

パターン, 58

パラメータ

表示オプション, 107

ひ

ヒストグラムとは, 23

ヒストグラム・ビジュアライザ, 25-27

起動, 25

表記上の決まり, xx

表示

決定木, 160

決定木ノード, 157

選択式決定木, 160

ツリー・ビジュアイザ

情報にスポットライトを当てる, 100

ツリー・ビジュアライザ

情報にスポットライトを当てる, 102

データ, 89, 111, 112

アニメーション・コントロール・パネル, 75-81

マップ・ビジュアライザ, 110, 111, 112

表示オプション, 119

表示オプション

マップ・ビジュアライザ, 119

表示、相関規則の, 219, 224, 226

表示パラメータ, 107

表示、ヘルプの, 32

表示メニュー

マップ・ビジュアライザ, 124

表示モード

スキヤタ・ビジュアライザ, 70

マップ・ビジュアライザ, 121

標本抽出, 47-48

ビジュアライザ

でのナビゲーション, 27-32

ピックドラッグ, 86

「ピックドラッグの表示」コマンド, 86

ピックドラッグ, 72

ふ

フィルタ, 43-44

- フィルタ、データの
 - レコードビューワ, 21
- フィルタパネル
 - スキャタ・ビジュアライザ, 83
 - レコードビューワ, 21
- フィルタボタン, 84
- 複数オブジェクトの選択, 99
- 負の値, 95
- ブースティング, 194
- 分析, 4
 - 関係, 56, 89
 - パターンと傾向, 58

へ

- 平均平方誤差, 162
- ヘルプ, 32
- 変換、データの
 - Tool Manager, 34-48
 - データの整形の必要性, 33
- 変更、項目の型の, 44-46
- 変更ボタン, 106
- 変更、マークの, 106
- ベース, 94, 95
 - 決定木と, 157

ほ

- 「補集合ドリルスルー」コマンド, 85, 86
- 保存、ツールオプションの
 - マップ・ビジュアライザ, 119
- 保存、データの
 - レコードビューワ, 22
- ボタン、Tool Manager, 34
- ボックスプロットとは, 22

ま

- マークコマンド, 103, 112
- マークパネル, 103-106, 112-??
 - 現在の場所を取得, 105
- マークボタン, 104
- マーケット・バスケット分析, 218
- マッピング
 - 相関規則と, 223
 - 文字列, 62
- マップ・ビジュアライザ
 - NULL 値と, 122
 - アニメーション・コントロール・パネル, 123
 - オブジェクトの選択, 125
 - オプション
 - 保存, 119
 - データの表示, 111, 112
 - 表示モード, 121
 - メインウィンドウ, 120-122
- マップ・ビジュアライザのオプション・ダイアログ・ボックス, 119
- マニュアル
 - 表記上の決まり, xx

め

- メインウィンドウ
 - 回帰ツリー, 161-1620
 - スキャタ・ビジュアライザ, 69-71
 - スプラット・ビジュアライザ, 69-73
 - ツリー・ビジュアライザ, 90
 - マップ・ビジュアライザ, 120, 122
- メニュー
 - スキャタ・ビジュアライザ, 82-87
 - スプラット・ビジュアライザ, 82-88

も

文字列
マッピング, 62
モデリング
概要, 127
教師付き, 3-5
教師なし, 5-6
モデル
作成, 11
使用, 11
生成, 128
精度の評価, 5, 11
選択, 145
適用, 11, 145-147
評価, 5, 11, 137

よ

予測モデリング
概要, 127
予備比率による誤差評価, 139

ら

ランドスケープ, 89

り

リーフノード, 156
量, 89
履歴の表示, 48-50
輪郭, 107
輪郭ファイルのフィールド, 118

る

ルートノード, 94
ルール・ビジュアルライザ
規則の表示, 224, 226

れ

レコード
クラシファイアと, 194
分析, 157-158
レコードの重み付け, 50-51
関連規則と, 223
レコードビューワ, 19-22
起動, 19
項目の操作, 20
データの保存, 22
フィルタパネル, 21