MineSet™ 3.0 Enterprise Edition Tutorial for Windows®

ドキュメント番号 007-4006-001JPN

編集協力者

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

MineSet™ 3.0 Enterprise Edition Tutorial for Windows®
ドキュメント番号 007-4006-001JPN

目次

このマニュアルについて　v
このマニュアルの対象者　v
このマニュアルの前提条件　v
このマニュアルの構成　vi
表記上の決まり　vi

1. **データマイニングの基本　1**
データマイニングとは　1
このマニュアルで使用している専門用語　2
データマイニングの手法　2
データマイニングアルゴリズム　3
　教師付き (Supervised) モデリング　4
　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

2. **データマイニングのプロセス　9**
データの識別　9
データの準備　10
　データの変換　11
モデルの構築　12
モデルの評価　12
モデルの展開　12
特定のデータベースへのプロセスの適用　12

3. **解約** (churn) **データセットを使用するチュートリアル　13**
元データについて　13
MineSet の起動　14
レコードの表示　15

エビデンス・クラシファイア (Evidence Classifier) の構築　19
スプラット・ビジュアライザ (Splat Visualizer) による確率の表示　23
地理的分布の可視化　28
決定木クラシファイア (Decision Tree Classifier) の作成　33

4. MineSet **のその他の機能　37**
データクラスタ　37
モデルの項目と軸の対応付け　40
クラスタ化されたモデルの重要項目を確認　42
スキャタ・ビジュアライザ (Scatter Visualizer) への割当て　43
デシジョン・テーブル (Decision Table)　46
モデルによる顧客の絞り込み　48
訓練事例の標本作成　49
モデルの適用　50
誤ったクラス判別のコストを削減　56
混同マトリックス (Confusion Matrix) の表示　56
損失マトリックス (Loss Matrix) の定義　59
ROI 曲線 (ROI curve) の表示　60
MineSet のその他の機能　62

A. MineSet **ビジュアライザのナビゲーション　65**
ツリー・ビジュアライザ (Tree Visualizer) でのナビゲーション　65
非ツリー・ビジュアライザでのナビゲーション　67

このマニュアルについて

このマニュアル『*MineSet 3.0 Enterprise Edition Tutorial for Windows*』は、Windows 環境で使用する MineSet について説明しています。MineSet は、データマイニングとデータの可視化を行う広範囲なツールから構成された統合ソフトウェア製品で、データマイニングの概念やプロセスをすぐに理解することができます。このマニュアルでは、ユーザが MineSet をすぐに使い始めることができるように、基本的な作業方法を紹介しています。MineSet のインタフェースに慣れたら、『*MineSet 3.0 Enterprise Edition User's Guide for Windows*』を参照して、他の MineSet 機能についても十分に理解してください。ユーザーズガイドは、MineSet 製品の一部としてオンラインでも提供されています。詳細については、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。

このマニュアルの対象者

このマニュアルは、エンドユーザを対象としています。プログラミング経験や統計についての知識があると内容を理解しやすくなりますが、これらの経験や知識がなくても差し支えありません。ただし、Windows の基礎知識は必要です。

このマニュアルの前提条件

このマニュアルに紹介されている操作を試すためには、システムに MineSet がインストールされているか、または MineSet がインストールされたシステムにアクセスできなければなりません。製品に付属しているサンプルデータは、システムによっては使用できない場合があります。MineSet のインストール方法は、『*MineSet 3.0 Enterprise Edition Instllation Instructions*』と MineSet の Web ページ（http://mineset.sgi.com、日本語版 http://mineset.sgi.co.jp）に説明されています。この Web ページでは、MineSet の評価版をダウンロードすることもできます。

このチュートリアルでは、データベースにアクセスする必要はありません。必要なデータは、MineSet 製品に付属しています。

このマニュアルの構成

第 1 章「データマイニングの基本」では、データマイニングの概念を紹介するとともに、データマイニングを使用して問題を解決する方法について説明しています。データマイニングの一般的な操作は、各種の MineSet ツールを使用して行います。各ツールについての詳細は、以降の章で取りあげます。

第 2 章「データマイニングのプロセス」では、データマイニングのプロセスに関する操作について説明しています。この章では、MineSet を使用したデータマイニングの事例も紹介しています。

第 3 章「解約 (churn) データセットを使用するチュートリアル」では、MineSet を使用してデータマイニングを行う方法について詳しく学習します。このチュートリアルでは、MineSet 製品に付属している「解約 (churn)」というデータセットを使用して、最初の画面から順に MineSet のツールを説明しています。

第 4 章「MineSet のその他の機能」では、より複雑なデータマイニング手法を使用して MineSet の詳細を学習します。

付録 A「MineSet ビジュアライザのナビゲーション」では、各ビジュアライザ・ウィンドウでの移動方法と操作方法を多数紹介しています。

表記上の決まり

このマニュアルでは、次の表記法を用いています。

『』	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。

第 1 章

データマイニングの基本

この章では、次の項目に沿って、データマイニングの方法、モデルの作成と評価、MineSet の役割について説明します。

- 「データマイニングとは」（1 ページ）
- 「このマニュアルで使用している専門用語」（2 ページ）
- 「データマイニングの手法」（2 ページ）
- 「データマイニング・アルゴリズム」（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）

データマイニングとは

データマイニングの目的は、データ内に潜むパターンを発見し、その結果を問題解決の手段として使用することです。強力な可視化機能と統合されたデータマイニングは、知識発見のための新しい方法を提供します。データマイニング・システムは、画期的な洞察方法を導く新しいパターンを自動的に発見し、表示するシステムです。たとえば、さまざまな属性の中から相関を発見したり、異なる特性を持つデータサブセットを識別したり、過去のデータから未来を推測することなどが挙げられます。

通常のデータベース問い合わせやオンライン分析処理 (OLAP: Online Analytical Processing) では、データ要素間の関係をユーザが直接指定する必要があります。データマイニングでは、一般にデータ間の未知の関連性や、ユーザが気が付かない関係などの発見ができます。

解析またはマイニングの対象となるデータは、通常、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。しかし、そのデータは、データマイニング以外の方法では解析が不可能なほど大量なデータ量になる場合があります。このような大量のデータは、多くの場合、データ・ウェアハウスに格納されます。詳細については、「データの準備」(10 ページ) を参照してください。

このマニュアルで使用している専門用語

MineSet で使用するデータファイルは、大きなテーブル（表）と考えることができます。1 行は 1 つの「レコード」を意味し、項目は各レコードの「属性」を意味します。クラス判別作業における「ラベル」の値とは、クラス判別のために選択した属性の特定の値を意味します。たとえば、このマニュアルで使用されているサンプルファイル *churn* では、解約した顧客と解約していない顧客とにレコードをクラス判別する作業を行います。この場合、ラベルの属性（項目）として「解約済み (churned)」があり、使用できるラベルの値は「はい (yes)」と「いいえ (no)」です。

離散型ラベルは、性別や、所得範囲（$40,000 未満、$40,000 ～ $80,000、$80,001 以上など）、年齢別（21 歳未満、21 ～ 35 歳、36 ～ 50 歳、51 歳以上など）のように限られた値だけが入る項目です。連続型ラベルには、年間所得、年間売上、リッター当たりの距離など、範囲の大きい任意の値を持つことできます。

データマイニングの手法

データマイニングの手法は、仮説の実証試験とデータ指向の発見を組合せたものです。仮説の実証試験では、ユーザはデータ本体と照らして仮説の妥当性の真偽を検証します。場合によっては、データ自体が発見を導くことがあります。データ指向の発見では、データ自体に結論を提示させることによって、ユーザはデータから結論を引き出します。データマイニングでは、この両方の手法を組合せることによってしばしば問題を解決します。たとえば、データが提示した結論から新しい仮説を引き出し、その仮説の真偽を検証できます。データマイニングは、統計学と機械学習が統合されたものと考えることができます。

MineSet の各種ツールを使用すると、データを分析、探索、グラフィック表示することによって、データの可視化と探索、さらに深い理解へとユーザを導きます。また、さまざまな方法でデータを編成、検証することもできます。マイニングツールは、自動的にパターンを発見し、可視化ツールを使用して表示できます。可視化ツールをデータに直接適用すると、データをさらに深く直感的に理解でき、隠れていたパターンや重要な傾向を発見できます。

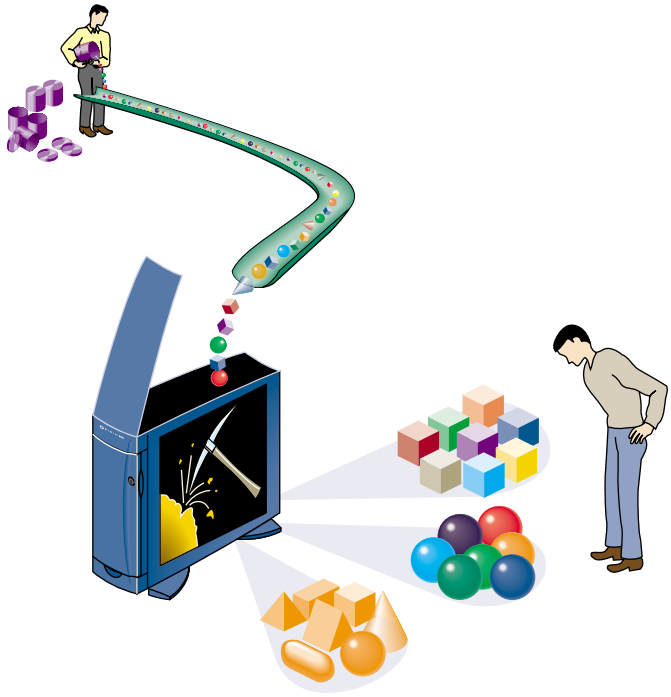


図 1-1 データ内のパターンを発見するデータマイニング

MineSet で一般的なデータマイニングを実行すると、データを記述したモデルが作成され、そのモデルが可視化されます。この可視化では、オブジェクトを対話形式で操作したり、アニメーションを実行できる 3D（3 次元）インタフェースが使用されます。この可視化により、モデルを理解し、複雑なデータパターンを調べることができるため、価値ある洞察を意思決定に反映させることができます。MineSet は、解析アルゴリズムによって可視化が行われる統合システムです。そしてユーザはさらに可視化要素を選択して、データマイニングを進めることができます。

データマイニング・アルゴリズム

データマイニング・アルゴリズムは、データからモデルを自動的に生成します。モデルの生成には、教師付き (*supervised*) アルゴリズムと教師なし (*unsupervised*) アルゴリズムの 2 種類が一般的に使用されます。他の項目の値に基づいて特定の項目の値を予測することを目的とする予測モデルタスクは、「教師付き」タスクと呼ばれます。このタスクは、生徒に質問の答を教える教師の指導に似ています。

記述モデリングの目的は、データからパターンとセグメントを発見することです。このようなタスクは、「教師なし」タスクに当たります。教師なしタスクには正しい答もなく、またあらかじめ結果を予測することもしません。教師なしタスクは、類似したパターンとセグメントを示すことにより、データ全体に対する洞察を手助けします。

教師付き (Supervised) モデリング

教師付きモデリングでは、これから予測しようとする特別な属性、つまり「ラベル」を指定します。このラベルと他の属性との関係をコード化することにより、ラベルのない新しいデータについて予測できます。また、モデル自体を可視化して、ラベルと他の属性との関係を洞察することもできます。たとえば、ある顧客が取引を中止した場合（一般に、顧客離れまたは解約）、今後どのような顧客が解約する可能性があるかを予測し、また、解約に至った理由とパターンを理解できるようなモデルを構築できます。

教師付きモデリングでは、2 つのタスク、クラス判別 (classification) タスクと回帰 (regression) タスクが最もよく使用されます。ラベルが離散型データ（つまり、固定した値のデータセット）の場合、そのタスクはクラス判別タスクです。ラベルが連続型データを持つ（つまり、給与や株価のように連続した範囲の値をとる）場合、そのタスクは回帰タスクです。

クラス判別 (Classification) タスク

クラス判別タスクとは、ラベルされていないレコードに離散型項目の値を割当てるタスクです（用語の意味については 2 ページの「このマニュアルで使用している専門用語」を参照）。この割当てにより、レコードはあらかじめ定義されているグループに分けられます。たとえば、顧客請求レコードを、支払期間が 60 日以内のクラスと 60 日を超えるクラスの 2 つに分けるといったクラス判別が行えます。また、顧客を性別や所得別に分けることもできます。クラシファイアは、ラベルが特定のレコードの特定の値になる確率も予測できます。たとえば、MineSet は、特定の顧客のレコード内の他の属性値がわかると、その顧客が 60 日以内に支払う確率を計算できます。

クラシファイアは、他の属性が与えられているデータセットから 1 つの属性を予測するモデルです。MineSet は、訓練事例（データセットのサブセット）から自動的にクラシファイアを作成できます。MineSet がクラシファイアを作成すると、ユーザがこのモデルの動作を理解し、洞察できるように、このモデルの可視化も行います。クラシファイアが生成されると、ラベルのないレコード（つまり、ラベルという属性が与えられていないレコード）のクラスの確率を判別、予測することができます。この概念の詳細は、第 3 章で説明します。

MineSet には 4 つのクラス判別モデル、すなわち決定木 (Decision Tree)、選択式決定木 (Option Tree)、エビデンス (Evidence、単純ベイズ (Simple Bayes) ともいう)、デシジョン・テーブル ・クラシファイア (Decision Table Classifier) があります。各モデルは、3D ビジュアライザを使用して表示できます。

回帰 (Regression) タスク

回帰 (Regression) タスクは、ラベルが離散型ではない点を除き、クラス判別 (Classification) タスクと同じ教師付きモデリングタスクです。たとえば、給与や株価の予測は回帰タスクですが、給与が一定の範囲内であるか、または株価が上昇するか下降するかの予測などはクラス判別タスクです。

モデルの精度の評価

予測モデルが完全であることはまれなため、データマイニング・プロセスでは予測モデルの精度の評価が重要な位置を占めます。精度の測定に使用するツールは、モデルの種類によって異なります。クラシファイアは通常、それらの誤差率に基づいて評価されます。もっとも一般的なのは、クラス判別の誤差、つまり誤ってクラス判別されたレコードの割合を調べる方法です。モデルの精度を評価する場合は、そのモデルの構築に使用されていないデータで試験する必要があります。MineSet には、誤差を評価する多数の方法が用意されています。詳細については、第 4 章「MineSet のその他の機能」を参照してください。

教師なし (Unsupervised) モデリング

教師なしモデリングの目的は、データ内で同様な特性を持つパターンやセグメント（つまり、クラスタ）を発見することです。教師なしモデリングは記述タスクであり、予測タスクではありません。予測のためにモデルを直接使用することはできず、したがって、クラシファイアの検証に使用するテストセットとしてデータの一部を設定する必要はありません。教師なしモデリングには、相関（Association）タスクとクラスタリング（Clustering）タスクがあります。

相関 (Associations) タスク

相関 (Associations) タスクとは、A は B を内包するといったデータ属性間の含意規則を決定して、相関性を生成するタスクです。相関性はアフィン群の検出に使用して、たとえば、他のアイテムと一緒によく購入されるアイテムを発見します。古典的なアフィン群とは、ある一定のアイテムも一緒に購入される頻度を予測するマーケット・バスケット分析です。たとえば、ベビーフードの購入者は、通常のタバコよりも低タールのタバコを購入する確率が高いことが分かれば、棚の配置変更に役立ちます。

クラスタリング (Clustering) タスク

クラスタリング (Clustering) アルゴリズムは、データを同様な特性を持つレコード群、すなわち、クラスタにセグメント化します。たとえば、生命保険会社では、年齢 20 ～ 45 歳、技術者、子供の数 1 人以下、好きなテレビ番組 SF、可処分年間所得 5,000 ～ 10,000 ドルといった特性によって 1 つのセグメントを抽出することができます。

生命保険会社は、このセグメント人口に合った生命保険商品のテレビコマーシャルを、たとえば新しく始まる SF 番組に流すことによって、さらに効果を上げることが期待できます。

データの可視化

データマイニング・アルゴリズムは、人間の脳が持つ驚くほどのパターン認識能力をデータ可視化技術によって完全なものとします。MineSet には、次のビジュアライザが用意されています。

- マップ・ビジュアライザ (Map Visualizer) — 一般に地図の形式でデータを表示します。
- スキャタ・ビジュアライザ (Scatter Visualizer) — 1 次元、2 次元、3 次元でデータポイントを表示します。追加の属性をカラー、サイズ、形状に割当てることもできます。最終的には、2 つの追加属性をスライダに割当てて、合計 8 次元のデータに対してアニメーションやフライスルーを行うことができます。また、MineSet の「重要項目」機能を使用すると、特定のタスクに合わせて割当てる重要な次元が識別しやすくなります。
- 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 (Scatter Visualizer) に似ていますが、半透明な雲で表されるカラーの不透明度によってデータ密度が示されます。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。
- ツリー・ビジュアライザ (Tree Visualizer) — データの階層が分かるように、データがノードに割当てられます。

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))、クラスタリング (Clustering) を必要とする問題にデータマイニングを適用する場合は、次の MineSet ツールが便利です。

- 決定木分析／クラシファイア (Decision Tree Inducer and Classifier) — 決定木ビジュアライザで可視化されるクラシファイアを作成します。
- 選択式決定木分析／クラシファイア (Option Tree Inducer and Classifier) — 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、別の選択肢を作成し、クラス判別中にそれらの選択肢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／クラシファイア (Evidence Inducer and Classifier) — 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／クラシファイア (Decision Table Inducer and Classifier) — 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。
- 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。
- 回帰ツリー (Regression Tree) — 実際の値（あらかじめ決められた範囲内の値ではなく段階的変化の値）を予測する回帰モデルを作成します。
- 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主成分を決定します。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、知識発見のプロセスに役立つツールとして、次のものも用意されています。

- 統計量ビジュアライザ (Statistics Visualizer) — ひとつの項目についてボックスプロットとヒストグラムでデータを表示します。連続型項目はボックスプロットとして、離散型項目はヒストグラムとして表示されます。
- ヒストグラム・ビジュアライザ (Histogram Visualizer) — データをヒストグラムの形で表示します。連続型項目は階級生成（複数の範囲に分類）されます。

- レコードビューワ (Record Viewer) — 元データをスプレッドシートとして表示します。

次の 第 2 章では、代表的なデータマイニング・プロセスと、ツールの使用方法について説明します。

第 2 章

データマイニングのプロセス

この章では、知識発見プロセスで行われる個々のタスクについて紹介します。知識発見プロセスは、図 2-1 に示すように、新しいパターンを発見したら元の段階に戻り、データをさらに深く理解していくという作業の繰返しです。

知識発見プロセスは、通常、次の手順で行われます。

1. データソースの識別 — 「データの識別」（9 ページ）を参照してください。
2. データの準備 — 「データの準備」（10 ページ）を参照してください。
3. モデルの構築 — 「モデルの構築」（12 ページ）を参照してください 。
4. モデルの評価 — 「モデルの評価」（12 ページ）を参照してください。
5. モデルの展開 — 「モデルの展開」（12 ページ）を参照してください。

データの識別

データを識別するには、問題を解決するためにどのデータが必要かをまず決定します。たとえば、問題の見地から顧客行動を予測できることが最終目的であることもあります。問題を明確にするには、問題の解決に必要なデータを識別するとともに、他にもデータソースがないかを調べる必要があります。

データが利用しにくい場所に保管されている場合や、古いフォーマットで保存されている場合もあります。また、初期のデータベースには、他と互換性がない場合もあります。データが不十分または不完全な場合には、データをさらに用意する必要があります。新しいデータを収集するためのフォームは、既存のフォームによって決まります。MineSet は、市販のデータベース（Oracle、Informix、SQL）に対するネイティブ・インタフェース、ODBC インタフェース、さまざまなファイル形式（タブで区切ったテキストファイル、MineSet バイナリファイル、Excel、SPSS、Mutable など）からのデータの読込みなどをサポートしています。

図 2-1 データマイニングのプロセス

データの準備

データの中には、MineSet にロードする前にクリーニングと呼ばれる変更が必要なものがあります。データの一般的な問題を次に示します。

- データ形式が、MineSet の形式と互換性がない場合があります。この例として、古いメインフレーム・コンピュータで使用されていたバイナリ、コード化されたデータ、EBCDIC 文字列などがあります。
- データにスペルミスや誤りがある（不完全な値や誤った値が入っている）場合があります。
- フィールドの記述が不明瞭な（紛らわしい）場合や、ソースによって意味が異なる場合などがあります。たとえば、注文日という記述は、品物が送付された日付、消印が押された日付、受領した日付、入力された日付などの可能性があります。
- データが最新でない場合があります。たとえば、顧客情報の場合、引っ越した、家族構成が変わった、消費傾向が変わったなどが考えられます。

クリーニングしたデータであっても、マイニングと可視化が行えるようにあらかじめ変換が必要になる場合があります。

データ変換

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。
- 可視化要素をフィルタリングする。たとえば、最も効力のある規則や収益性が最も高い顧客セグメントだけを表示できます。
- 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。
- データを集計する。レコードをまとめてグループ化し、合計値、最大値、最小値、平均値などを求めます。
- 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。
- すでに作成済みの分析モデルを適用し、新しいレコードにクラスの分類名を付けたり、特定の分類名の値の確率を予測したりする。

MineSet では、これらの変換作業のほとんどは、Tool Manager の「データ変換」ウィンドウを使用して行います。

モデルの構築

知識発見プロセスの中心となるタスクは、モデルの構築です。このタスクは、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。詳細については、第 3 章を参照してください。

モデルの評価

モデルの精度を評価すると、モデルの内容とその実用性についての理解を深め、データのフィルタリング、項目の削除、新しい項目の作成などを通してモデルを向上させることができます。

MineSet には、モデルを評価する方法として、誤差推定 (error estimation)、混同マトリックス (confusion matrix)、改善曲線 (lift curve)、ROI 曲線（return-on-investment curve）の 4 つが実装されています。

モデルの展開

モデルを新しいデータに適用して、さらにモデルを展開させることができます。新しいデータに適用することによって、より高い精度を必要とする疑問が提示されることもあります。

第 3 章の電話会社の例では、電話契約を解除しそうな顧客を調べるモデルを作成します。続いて、そのモデルを使用して顧客レコードを評価し、解約の可能性が最も高い顧客を特定します。これらの顧客には、引き続き契約してもらうための刺激策を施すことができます。

特定のデータベースへのプロセスの適用

第 3 章と第 4 章では、あらかじめ用意された電話顧客情報、churn データセットによる知識発見プロセスを紹介しています。例に沿って作業を進めながら、使用されているプロセスと、図 2-1 に示されている作業の進捗と折返しについて理解してください。

第 3 章

解約 (churn) データセットを使用するチュートリアル

この章では、MineSet に付属する解約 (*churn*) データセットによる知識発見プロセスを紹介します。この章の説明は、ユーザが使用するシステムに、MineSet とすべてのサンプルファイルがインストールされていることを前提にしています。作業ごとに詳しく説明しますが、特に記述がないかぎり次の順に進めてください。

- 「MineSet の起動」（14 ページ）
- 「レコードの表示」（15 ページ）
- 「エビデンス・クラシファイア (Evidence Classifier) の構築」（19 ページ）
- 「スプラット・ビジュアライザ (Splat Visualizer) による確率の表示」（23 ページ）
- 「地理的分布の可視化」（28 ページ）
- 「決定木クラシファイア (Decision Tree Classifier) の作成」（33 ページ）

元データについて

解約 (*churn*) データセットは、電話会社の顧客（電話を定期的に使用している人々）の情報を集めたものです。顧客は、電話サービスを提供する電話会社を選択できます。顧客が電話会社を変更することを「解約 (churn)」と呼び、解約された会社にとっては利益が減少します。電話会社には、一般に、通話情報（発信、着信、日付、通話時間）が入った通話記録データベース、課金データベース、顧客データベース、顧客サービス・データベースなどがあります。顧客に関する情報は、すべてこれらのデータベースに存在します。これらのデータベースを組み合わせると、顧客の特徴が理解できる情報が生成されます。MineSet に用意されている 解約 (churn) データセットは、このようなデータです。データを識別し、顧客の識別情報をレコードとして作成する手順はすでに済んでいます。以降で使用するこのデータセットには、顧客 1 人につき 1 つのレコードが入っています。

MineSet の起動

1. 「開始」->「プログラム」->「*MineSet 3.0 Enterprise Edition*」->「*MineSet*」を選択するか、またはデスクトップ上の「*MineSet*」アイコンをダブルクリックします。
2. 「サーバーへログイン」ダイアログ・ボックス（図 3-1）がデフォルトで表示されない場合、「ファイル」->「サーバーに接続」を選択します。ダイアログ・ボックスが表示されたら、現在のシステムをクライアントとサーバの両方として使用する場合は「このマシンを現在のユーザとします」をクリックしてください。別のシステムをサーバとして使用する場合は、サーバ名、ログイン名、パスワード（設定してある場合）を入力してください。

図 3-1 Tool Manager のログイン・ウィンドウ

3. 「了解」をクリックします。以前に MineSet にログインしている場合は、ここでセッションが復元される場合がありますが、このチュートリアルを進めるためには、次の手順 4 で新しいファイルを開いてください。
4. 「*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。表示されるダイアログ・ボックスにデータ・ディレクトリが表示されない場合、MineSet がインストールされているディレクトリ（デフォルトでは *MineSet 3.0 -> data*）に移動してください。
5. *churn.schema* を選択します。図 3-2 に示すように、右側の「項目のプレビュー」ウィンドウに一連のエントリが表示されます。
6. 「開く」をクリックします。

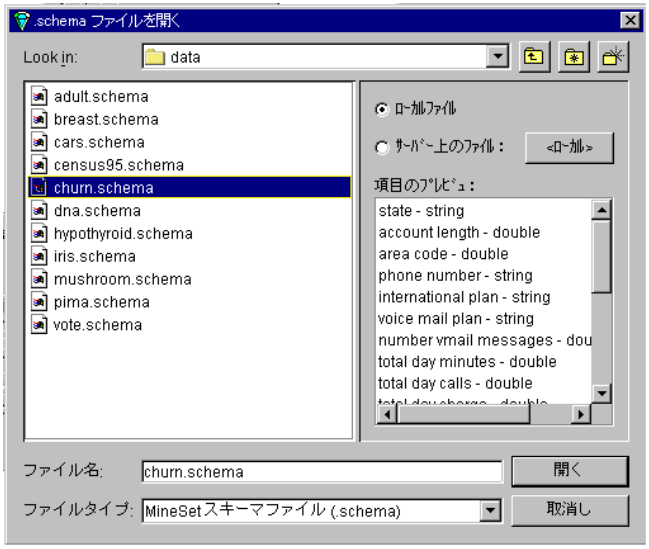


図 3-2　「新しいデータファイルを開く」ウィンドウ

以上で、電話会社の顧客データセットにアクセスできるようになります。次回 MineSet を実行すると、この位置か、MineSet を最後に終了したときの位置に自動的に戻ります。MineSet を実行している間に選択したオプションは保存されます。

レコードの表示

レコードは、次の手順で MineSet の Tool Manager を起動した後、スプレッドシートの形式で表示できます。

1. Tool Manager の「データの可視化／マイニング」ウィンドウの上段タブの中から「可視化ツール」タブをクリックします。続いて、下段の「レコード」をクリックして「レコードビューワ」タブにアクセスします。

   この章で使用している解約 (*churn*) データセットには、顧客ごとに 1 つのレコードが入っています。Tool Manager の左側の「データ変換」ウィンドウには、「州（文字列型）」、「取引期間（倍精度型）」のように、項目名とともにデータ型が表示されます。項目は、数値が入る場合には double（倍精度型）または float（浮動小数点型）、文字が入る場合は string（文字列型）として定義されます。

2. 右隅にある「ツールの起動」をクリックします。

データが、スプレッドシートとして表示されます。次の表 3-1 に、項目とその意味を示します。

表 3-1 MineSet の「レコードビューワ」に表示される解約 (churn) データセットの項目の詳細

項目名	値
「州 (state)」	顧客が居住している米国の州を 2 文字の略語で示したもの
「取引期間 (account length)」	顧客がその電話会社を利用している月数を示した数値
「地域コード (area code)」	電話会社によって指定される 3 桁のコード
「電話番号 (phone number)」	電話会社によって指定される 3 桁 + 4 桁のコード
「国際通話プラン (international plan)」	国際通話の特別価格パッケージ（値「はい (yes)」/「いいえ (no)」で示す）
「ボイスメール プラン (voice mail plan)」	電話会社のボイスメールを利用する顧客向けの特別料金パッケージ（値「はい (yes)」/「いいえ (no)」で示す）
「ボイスメール メッセージの数 (number of voice mail messages)」	1 日あたりのボイスメール・メッセージの平均数
「日中通話の合計時間 (total day minutes)」 「夕方通話の合計時間 (total eve minutes)」 「夜間通話の合計時間 (total night minutes)」 「国際通話の合計時間 (total intl minutes)」	日中、夕方、夜間、国際の各通話に課金された時間の平均（単位 : 分）
「日中通話の合計数 (total day calls)」 「夕方通話の合計数 (total eve calls)」 「夜間通話の合計数 (total night calls)」 「国際通話の合計数 (total intl calls)」	日中、夕方、夜間、国際の各時間帯の通話数の平均

表 3-1 (続き) MineSet の「レコードビューワ」に表示される解約 (churn) データセットの項目の詳細

項目名	値
「日中通話の課金合計 (total day charge)」 「夕方通話の課金合計 (total eve charge)」 「夜間通話の課金合計 (total night charge)」 「国際通話の課金合計 (total intl charge)」	日中、夕方、夜間、国際の各料金の課金平均
「顧客サービス通話の回数 (number customer service calls)」	この半年間でこの顧客が電話会社の顧客サポートにかけた回数
「解約 (churned)」	この半年間でこの顧客が電話会社を変更したかどうか（値「はい (yes)」/「いいえ (no)」で示す）

3. 「レコードビューワ」ウィンドウを閉じます。解約 (*churn*) データソースをまだ使用している「*Tool Manager*」ウィンドウがもう一度表示されます。
4. 「*Tool Manager*」ウィンドウの「データの可視化／マイニング」ウィンドウには、「可視化ツール」タブがまだ表示されています。下段の「統計量」タブをクリックします。
5. 「ツールの起動」をクリックします。

   多数のヒストグラムとボックスプロットが入った「統計量ビジュアライザ」画面が表示されます。ヒストグラムには離散型変数の値の分布が示され、ボックスプロットには連続型変数のサマリ統計量が示されます。

   各ボックスプロット（図 3-3 の右側）には、1 つの項目についての統計量として、最小値、最大値、平均値（赤）、中央値、4 つの四分位数のうちの 2 つ（25 パーセント点と 75 パーセント点）が表示されます。これらの値は線で示され、標準偏差は、+ または - 記号が付いた赤い文字で示されます。

   平均値は、項目内のデータを合計し、それをレコード数で割った値です。中央値は、項目内の値を大きさ順に並べた場合の中央の値です。標準偏差は、項目内のデータの散らばりを表します。

ヒストグラムには、個々の離散値（州の名前や「はい (yes)」/「いいえ (no)」で示された値）が表示されます。この画面で下方向へスクロールすると、「解約 (*churned*)」項目のヒストグラムが表示されます（図 3-3 の左側を参照）。この図は、顧客 5,000 人の内 707 人が電話会社との契約を解除したことを示しています。「解約 (*churned*)」項目は、このチュートリアル全体で重要なポイントになります。

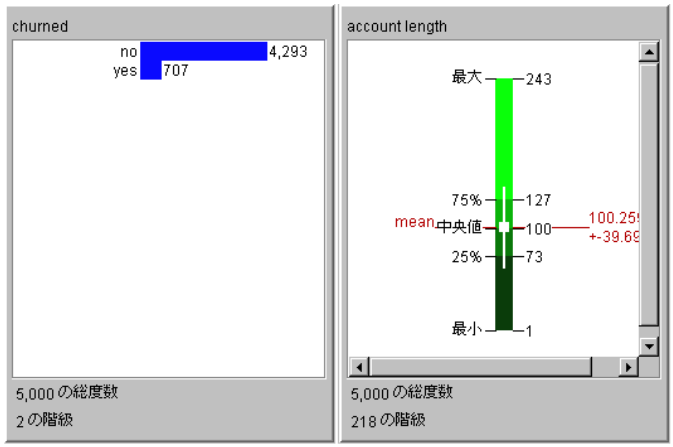


図 3-3 「統計量ビジュアライザ」による代表的なヒストグラムとボックスプロット

6. 「統計量ビジュアライザ」ウィンドウを閉じ、「*Tool Manager*」ウィンドウに戻ります。

エビデンス・クラシファイア (Evidence Classifier) の構築

以上の作業で、データマイニングの準備が整います。MineSet が適切なサーバに接続されており、データソースが *churn.schema* であることを確認してください。セッション間で MineSet を終了した場合は、もう一度起動すれば、履歴ファイルによって自動的に最後に表示されていた画面に戻ります。

1. Tool Manager の「データ変換」ウィンドウで、次の項目を削除します。項目を 1 つクリックして <Ctrl> キーを押せば、他の項目を順にクリックできます。これらの項目が選択されたところで、「項目の削除」をクリックしてください。

   「電話番号 (phone number)」
   「日中通話の合計時間 (total day minutes)」
   「日中通話の合計数 (total day calls)」
   「夕方通話の合計時間 (total eve minutes)」
   「夕方通話の合計数 (total eve calls)」
   「夜間通話の合計時間 (total night minutes)」
   「夜間通話の合計数 (total night calls)」
   「国際通話の合計時間 (total intl minutes)」
   「国際通話の合計数 (total intl calls)」

   「電話番号 (phone number)」項目には、予測値が入っていません。「合計時間 (total minutes)」項目と「通話の合計数 (total calls)」項目は、「通話の課金合計 (total charge)」項目と関連しており、その他の情報も少し入っています。これらの項目を削除すると、モデルを作成する処理時間を短縮し、より簡潔に可視化できます。

2. 「*Tool Manager*」ウィンドウの「データの可視化／マイニング」ウィンドウの上段タブの中から「マイニングツール」タブをクリックします。

3. 下段の「クラス判別」タブをクリックし、プルダウン・メニューから次のように選択します。

   「モード」:「クラシファイアとエラー」

   「分析」:「エビデンス」

   「離散型ラベル」:「解約」

   次は、解約しそうな顧客の特徴を明らかにするために、エビデンス・クラシファイア を作成します。デフォルト・モードの「クラシファイアとエラー」は、データの 3 分の 2 から分析モデルを作成し、残りは誤差率を評価するためのテストセットとして残す予備法を使用します。

4. 「実行」をクリックします。

   「州 (state)」項目の値が多すぎる場合は、分析によるデータ読込みが進行している間、「州 (state)」項目を削除する警告メッセージが表示されます。この場合、Tool Manager の「ファイル」->「設定」メニューを選択し、属性値のデフォルトの最大を 100 に変更してください。

Tool Manager の一番下の「ステータス」ウィンドウに、モデル生成プロセスの進行状況とサマリ情報（推定誤差率 11.40% ± 0.78% など）が表示されます。モデル生成プロセスが終了すると、エビデンス・ビジュアライザ (Evidence Visualizer) が自動的に呼出されて、モデルを表示します（図 3-4）。

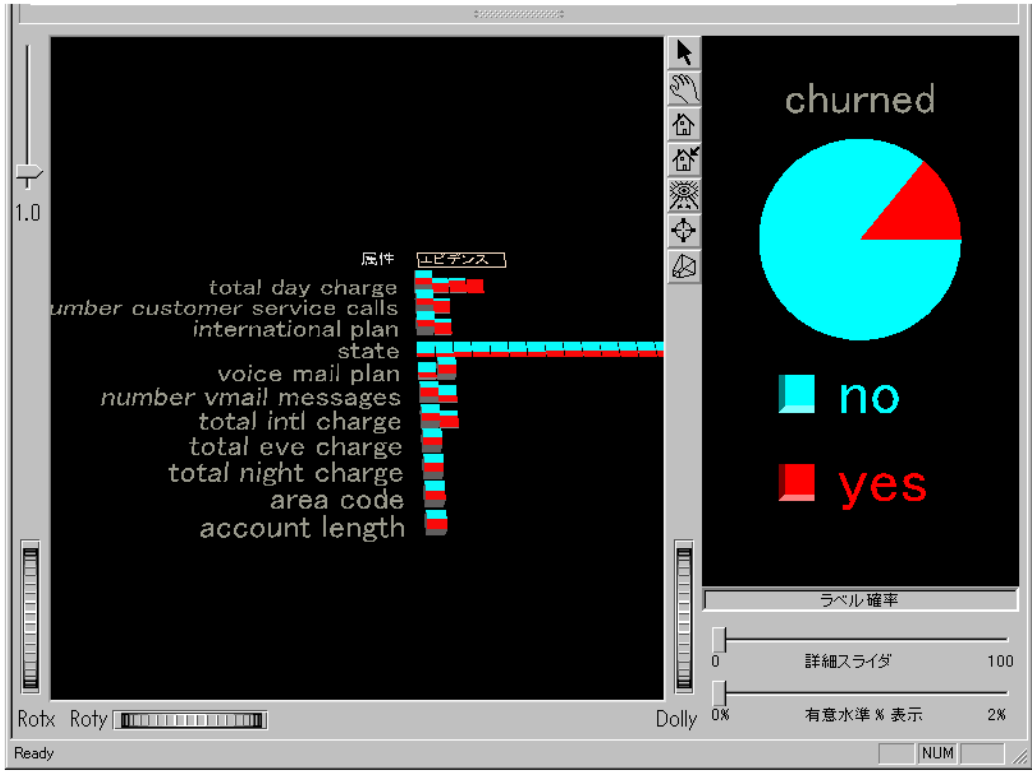


図 3-4 「エビデンス・ビジュアライザ (*Evidence Visualizer*)」ウィンドウ

エビデンス・ビジュアライザ (Evidence Visualizer) は、ラベル「解約 (churned)」に対する識別力順に項目をソートします。表示をすばやく調整するには、「*Dolly*」ダイヤル

を使用してください。このチュートリアルでは、左側のウィンドウのケーキグラフも右側のウィンドウの円グラフも「グラフ」と呼びます。

「ラベル確率」ウィンドウ（図 3-4 の右側）には、事前確率を示す円グラフが表示されます。事前確率とは、属性値を考慮せずに計算された、「はい (yes)」（赤の部分）または「いいえ (no)」（青の部分）の解約値を持つランダムなレコードの確率を意味します。数学的には、このクラスのラベルを持ったレコード数を合計レコード数で割った値です。

「エビデンス」ウィンドウ（図 3-4 の左側）のグラフには、データセット内の各項目が 1 つの値または値の範囲と共に示されます。カーソルモードをハンドモード（手）からピックモード（矢印）に切り替え、「エビデンス」というボックスをクリックしてください。すると、ケーキグラフに条件付き確率が表示されます。これは、円グラフで示されている属性値（「ボイスメール・プラン (voice mail plan)」項目の値「はい (yes)」など）を持つ顧客の解約値が「はい (yes)」である確率です。

モデルに基づいた期待確率を表示するため、左側のグラフをクリックし右側のウィンドウを更新してください。

画面内をナビゲートするには、画面の外枠にあるダイヤルを使用するか、またはマウスボタンと <Ctrl> キーをさまざまな組合せで使用します。ナビゲーション制御の詳細については、付録 A「MineSet ビジュアライザのナビゲーション」を参照してください。

図 3-4では最初と 2 番目の行で解約を示すスライスが左から右へ伸びて解約要因が示されるため、重大な問題は明白に分かります。この電話会社のサービスをよく利用する顧客の解約率も高めです。この会社は、最も大切な顧客を失いつつあります。

クラスの分類名（「解約 (churn)」の値が「はい (yes)」など）を確認するには、右側の「ラベル確率」ウィンドウで値を選択し、「はい」ボタンをクリックします。すると、エビデンス (evidence) がバーとして示されます。これらのバーをポイントすると、推定確率が示されます。

ここに示されている分析属性は、スキャタ・ビジュアライザで軸を選択するためにも使用できます。属性リストで比較的高い値が示されている属性「州 (state)」からは、地理的な関係が推測できます。

エビデンス (Evidence) モデルは、属性を個々に取り上げて表示します。しかし、データセットの多くの属性は単独ではなく、したがって分類名の決定には属性のセット（組合せ）の方が好都合です。ステータス・ウィンドウの誤差推定は、クラシファイア (Classifier) が約 12% の誤差率を予測していることを示しています。後で、誤差がより

少ない決定木 (Decision Tree) を作成します。ここで、エビデンス・ビジュアライザ (Evidence Visualizer) を閉じてください。次は、スプラット・ビジュアライザ (Splat Visualizer) の作業を行います。

スプラット・ビジュアライザ (Splat Visualizer) による確率の表示

スプラット・ビジュアライザ (Splat Visualizer) を使用する場合、色に割当てられた項目は数値でなければなりません。「解約 (*churned*)」項目のデータ型は string（文字列型）であるため、スプラット・ビジュアライザに割当てる前に数値への変換が必要です。このために、「解約の確率 (p_churned)」という新しい項目を作成します。

1. 「データの可視化／マイニング」ウィンドウの上段タブの中から「可視化ツール」タブをクリックします。続いて、下段の「スプラット」タブをクリックします。スプラット・ビジュアライザ (Splat Visualizer) が表示されます。
2. 「データ変換」ウィンドウで、「項目の追加」をクリックします。

図 3-5 新しい項目の追加

3. 「項目の追加」ダイアログ・ボックス（図 3-5）の「新しい項目名」テキスト・フィールドに、新しい項目名として **`p_churned`** と入力します。この項目は、「解約 (churned)」項目を基にした数値型の項目として使用します。

   「表現による定義」テキスト・フィールドで、表現 **`('churned' == "yes") ? 100 : 0`** を作成します。この表現は、左側の 2 つのスクロールリスト「表現に項目名を追加」と「表現に処理を追加」を使用して作成することも、直接入力することもできます。この表現は、「解約 (churned)」項目の値が「はい (yes)」であれば p_churned に値 100 を与え、「はい (yes)」以外であれば値 0 を与えよ」という意味です。この表現は、文字列（「はい (yes)」または「いいえ (no)」を数値に変換するためのものです。「新しい型」テキスト・フィールドが double（倍精度型）に設定されたことを確認してください。

   「表現をチェック」をクリックし、構文エラーがないことを確認してください。「了解」をクリックしてダイアログ・ボックスを閉じ、「了解」をクリックしてこの項目を追加します。

4. 「データ変換」ウィンドウの「スプラット」タブで、各要素の隣りのプルダウン・メニューで項目を選択して、可視化要素に項目を割当てます。このチュートリアルでは、結果が図 3-6 のようになるようにプルダウン・メニューから次のように選択してください。

   「軸 1」:「日中通話の課金合計 (total day charge)」

   「軸 2」:「顧客サービス通話の回数 (number customer service calls)」

   「軸 3」:「国際通話プラン (international plan)」

   「色」:「解約の確率 (p_churned)」

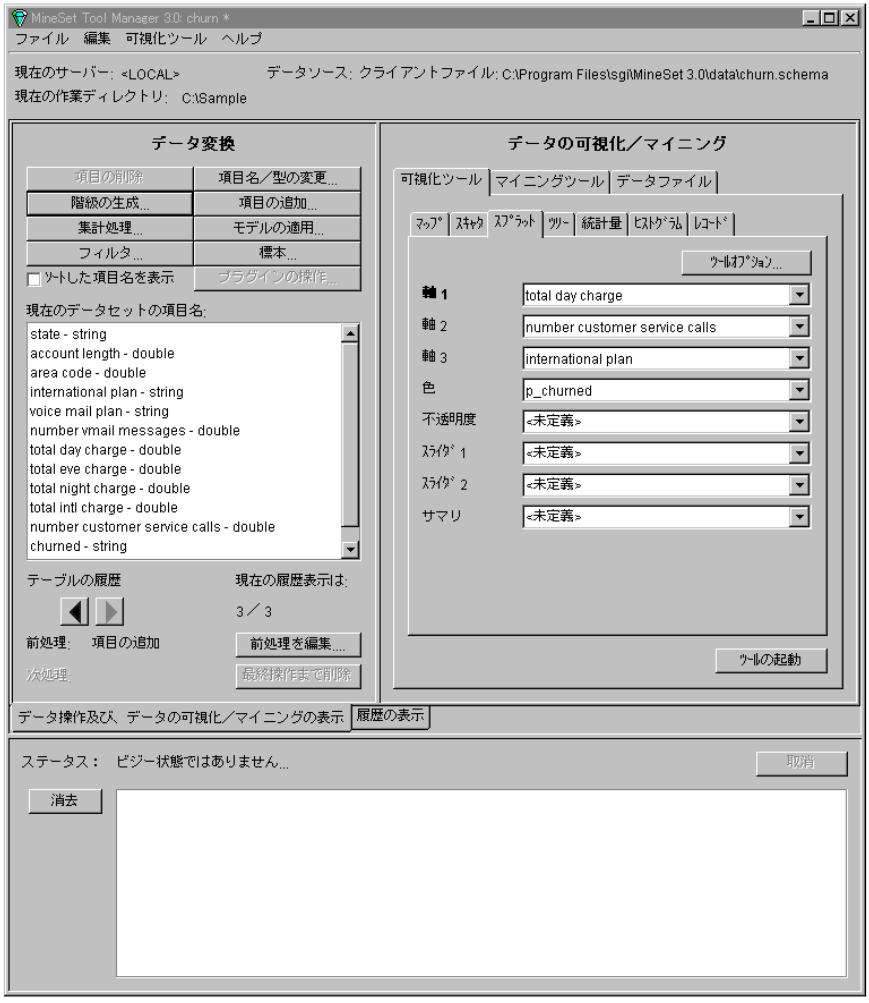


図 3-6 「スプラット ビジュアライザ」の可視化要素に対する項目の割当て

5. 「ツールの起動」をクリックします。

   図 3-7 に示すように、データが「スプラット ビジュアライザ」ウィンドウにプロットされます。左上にあるスライダバーは、色の濃さを変えるために使用します。ウィンドウ操作については、付録 A「MineSet ビジュアライザのナビゲーション」を参照してください。画面をナビゲートし、別の領域を確認するには、左右のマウスボタンをクリックしたままカーソルを画面上で移動させます。スプラット・ビジュアライザ (Splat Visualizer) では、さまざまな動作を複数の次元で表示することにより、複雑なデータを解析できます。

   Tool Manager の現在の状態（特殊なオプションなど）は、「ファイル」->「現在のセッションを別名保存」を選択し、*churn1.mineset* を指定して保存できます。

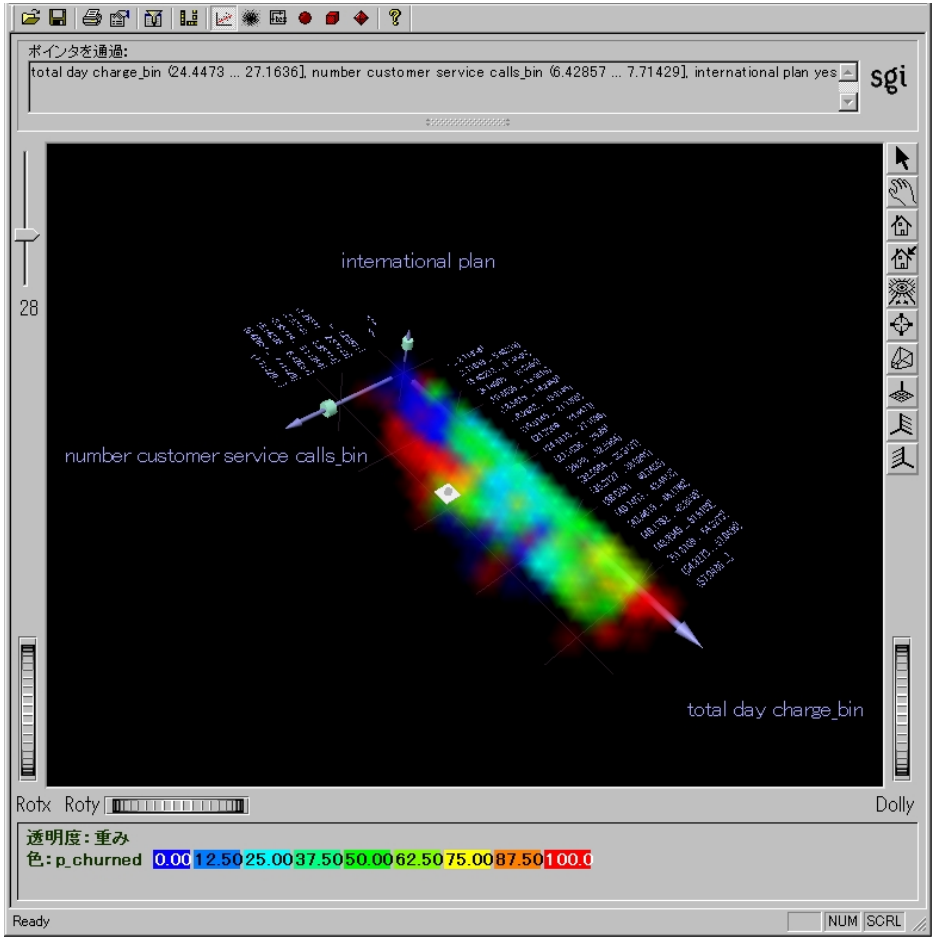


図 3-7 「スプラット ビジュアライザ」ウィンドウ

図 3-7 に示したウィンドウでは、解約が最も起きやすい位置が 2 か所あります。1 つは、「日中通話の課金合計 (total day charge)」が高くなる黄から赤にかけての位置（この図の下の方）です。もう 1 つは、「日中通話の課金合計 (total day charge)」が低く、かつ「顧客サービス通話の回数 (number customer service calls)」が高くなる位置（この図の左上端近く）です。つまり、利用料金は低いが顧客サービス通話の利用が多いという顧客は解約しがちです。このような顧客は、経費がかかり収益がほとんど引き出せないため、引き留める必要はありません。ターゲットにするとよいのは、この図の下の方に示されている利用料金が高い顧客です。

6. スプラット・ビジュアライザ (Splat Visualizer) を閉じ、「*Tool Manager*」ウィンドウに戻ります。

地理的分布の可視化

図 3-4 （20 ページ）に示すように、主成分分析 (Evidence) モデルは「州 (state)」が代表的な属性であることを示しています。使用しているモデルがこのように表示されない場合は、「エビデンス・クラシファイア (Evidence Classifier) の構築」（19 ページ）の手順 4 で説明している最大値の変更がなされていない可能性があります。この節では、解約が州によってどのように異なるかを示すため、前の計算に基づいてデータを地理的に表示します。

すでに、データセットの既存の項目を使用して「解約の確率 (p_churned)」項目の追加は行いました。次は、データを、州ごとの平均解約率が入った小さなデータセットに変換します。このような変換は、集計処理 (Aggregation) と呼ばれます。

1. 「*Tool Manager*」ウィンドウの「データ変換」ウィンドウで、「集計処理」をクリックします。

   「集計処理」ダイアログ・ボックスで、「解約の確率 (p_churned)」を強調表示して左向きの矢印をクリックして、左の項目に移動させてください。新しいウィンドウで p_churned を強調表示し、「平均」と「カウント」がオンにチェックされ、「合計」、「最小」、「最大」がチェックされていないことを確認してください。「州 (state)」を中央の項目に置いたまま、残りをすべて右側の項目に移動させてください（複数の項目を選択するには <Ctrl> キーを押したままクリック）。画面が図 3-8 のようになったことを確認し、「了解」をクリックして選択を確定してください。

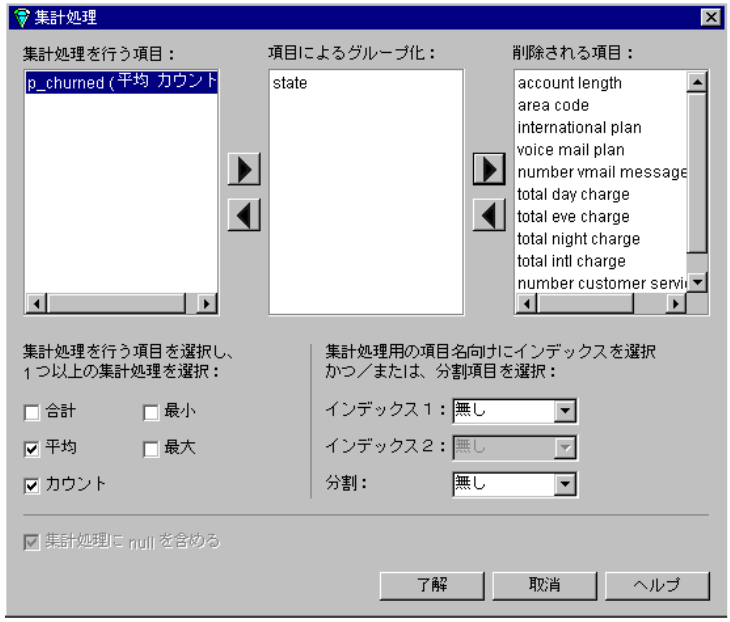


図 3-8 「集計処理」ダイアログ・ボックス

2. 「データの可視化／マイニング」ウィンドウの上段タブの中から「可視化ツール」をクリックし、続いて「レコード」タブをクリックします。このタブで「ツールの起動」をクリックし、平均解約率と解約した顧客の合計数が示された各州のレコードを確認します。
3. 「レコードビューワ」ウィンドウを閉じ「*Tool Manager*」ウィンドウに戻ります。次に、このデータを米国の地図に割当てます。
4. Tool Manager で「データの可視化／マイニング」ウィンドウの下段タブから「マップ ビジュアライザ」タブをクリックします。続いて、「ツールオプション」ボタンをクリックします。「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。(図 3-9)
5. 「要素ファイル」テキスト・フィールドの右側にあるボタンをクリックします。ここから、MineSet がインストールされているディレクトリに移動し、*config\mapviz\gfx_files\usa.state.hierarchy* を選択します。

図 3-9 「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

6. 「開く」をクリックしてこのファイルを開き、「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」パネルを閉じます。

   次は、可視化要素を項目に割り当てます。

7. 可視化要素の隣りのプルダウン・メニューから次のように選択し、「データ変換」ウィンドウの現在の項目を、「データの可視化／マイニング」ウィンドウの要素に割当てます（図 3-10）。

   「バーとなる要素」: 「州 (state)」
   「バーの高さ」: 「解約数 (count_p_churned)」
   「バーの色」: 「平均解約率 (avg_p_churned)」

図 3-10 「マップ ビジュアライザ」の可視化要素に対する項目の割当て

8. 「ツールの起動」をクリックし、解約した顧客を州ごとに示した地図分布を表示します（図 3-11）。

可視化されたこの図から、米国全体の解約した顧客の分布状態が分かります。各州の色は解約の確率を示し、高さはその州で解約した顧客数を示します。たとえば、図 3-11 ではメイン州が選択されており、この州の平均解約率は 18.4466% ですが、この

値は解約数 103 を基にしたものです。つまり、この平均解約率はたった 103 人の解約した顧客から計算されたものにすぎません。ウェストバージニア州は、158 人の顧客を基にした最も高い解約率を示しています。明るい色で示された州（テキサス、モンタナ、ワシントン、カリフォルニア、ニュージャージー）の平均解約率は 21% を超えています。この図から、州によって解約率が異なるものの、解約と地理的要素の間に明白な関係はないことが分かります。

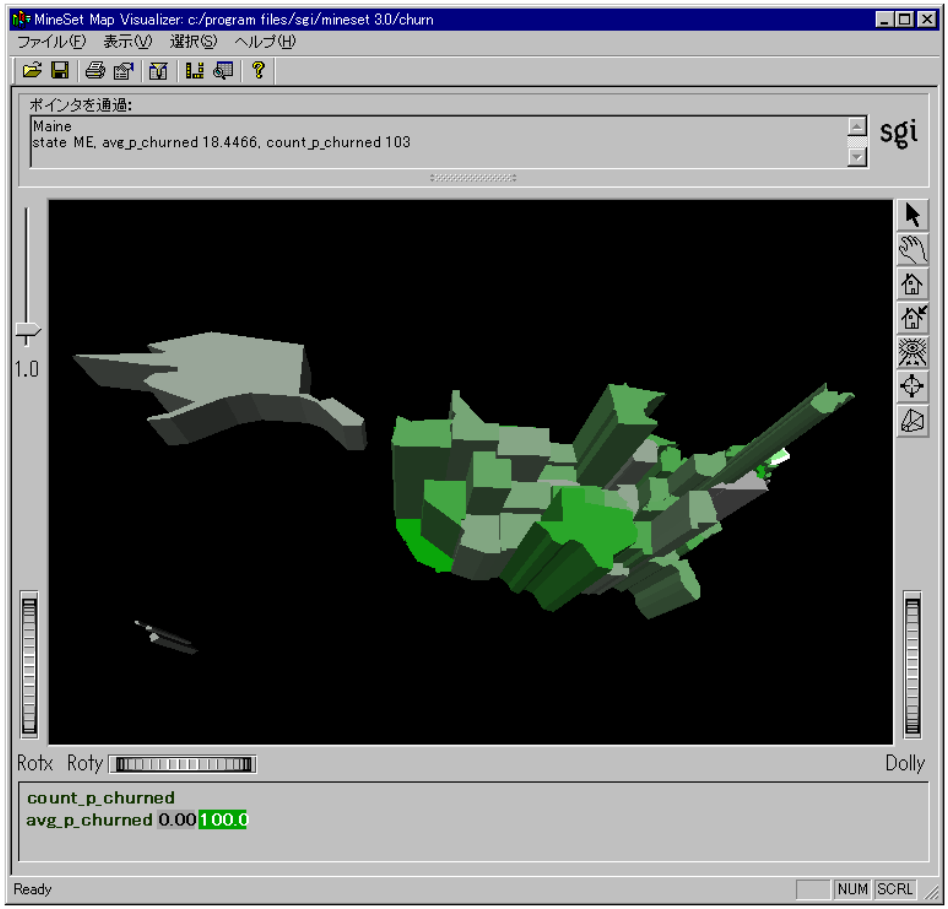


図 3-11 平均解約率を示した「マップ ビジュアライザ」ウィンドウ

「ファイル」->「終了」を選択して、マップ・ビジュアライザ (Map Visualizer) を閉じてください。次の節では、同じデータセットを決定木クラシファイア (Decision Tree Classifier) に適用して、別の可視化を行います。

決定木クラシファイア (Decision Tree Classifier) の作成

エビデンス・クラシファイア (Evidence Classifier) と違い、決定木クラシファイア (Decision Tree Classifier) は、属性間の相関（クラス判別に影響を与える属性値の組合せ）を示すことができます。この節では、まず履歴モードを使用します。次の手順に従って決定木クラシファイアを構築し、可視化してください。

1. Tool Manager の「データの可視化／マイニング」パネルの底部にある「履歴の表示」タブをクリックし、履歴モードに切り替えます。
2. 履歴の一覧で「項目の追加」操作と「集計処理」操作をクリックして「削除」を選択し、これらの操作を削除します。
3. 「データ操作及び、データの可視化／マイニングの表示」に戻します。画面には、「現在の表示は : 2/2」と表示されているはずです。
4. 「データの可視化／マイニング」パネルの上段タブの中から「マイニングツール」タブをクリックします。
5. 下段の「クラス判別」タブをクリックし、プルダウン・メニューから次のように選択します。

   「モード」:「クラシファイアとエラー」

   「分析」:「決定木分析」

   「離散型ラベル」:「解約 (churned)」
6. 「実行」をクリックします。

   図 3-12 のように、決定木 (Decision Tree) モデルが作成されます。図 3-4 のエビデンス・ビジュアライザ (Evidence Visualizer) よりも、推定誤差率が大幅に下がっています (6.36% ± 0.60%)。これは、属性間の相関が重要な意味を持つという当初の仮説が正しいことを裏付けています。図 3-12 では、決定木内の各ノードには、ラベルの値ごとに 2 つのバーがあります。バーをポイントすると、そのラベルの値のレコード数と比率が表示されます。各ノードのベースは、その値に達しているレコードの数と色、およびサブツリーの推定誤差率を示します（この図の底部にある説明を参照）。

   この例では、決定木のルートは「日中通話の課金合計 (total day charge)」です。これは、顧客が日中通話に支払った金額が最も重要な要因であり、分かれ目となる区間が 44.96 であることを示します。

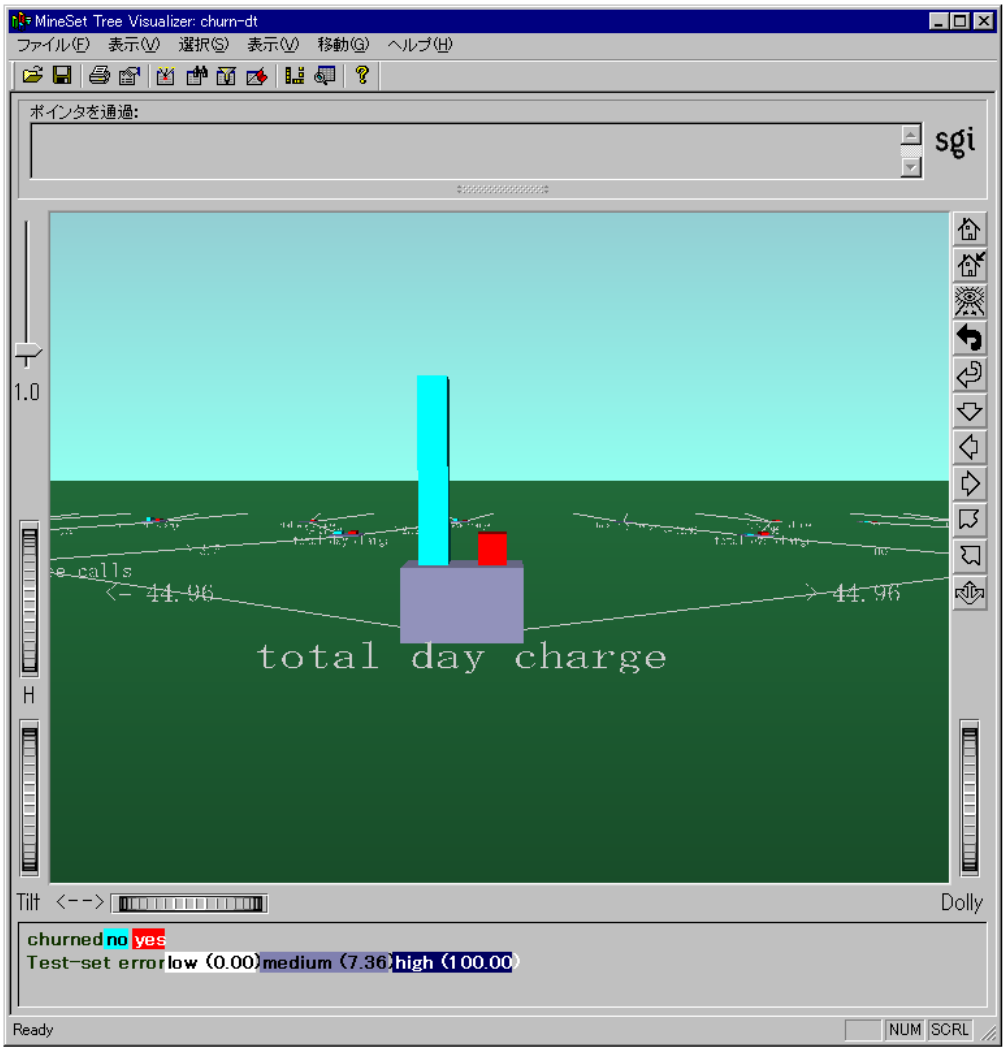


図 3-12 「ツリー ビジュアライザ」ウィンドウ

ツリー・ビジュアライザ (Tree Visualizer) の画面は、画面上の「*Dolly*」ダイヤルかマウスボタンの組合せを使用して移動できます。ナビゲートの詳細については、付録 A「MineSet ビジュアライザのナビゲーション」を参照してください。ルートの右側の赤いバー（「解約 : はい (churned: yes)」）を選択すると、14.14% の顧客が解約することが分かります。赤い線（「日中通話の課金合計 (total day charge) > $44.96」をクリックして、日中通話の利用が一番多かった顧客を含む子ノードに移動してください。これらの顧客の 59.31% は解約することが分かります。左右どちらかの線をもう一度クリックし、日中通話の利用が一番多い顧客のボイスメール・プランの有無を示すノードに移動してください。これらの顧客の内、ボイスメールを利用している顧客は 9.33% という非常に少ない解約率を示します。このことから、ボイスメールの推奨によって解約が減少する可能性がうかがえます。

このツリーはデータから自動的に作成されたことに注意してください。ノードとして選択される属性と区間は、モデル生成プロセスによって自動的に決定されます。

データのドリルスルーと表示を行うには、ノードのベースまたはバーを選択し、「選択」->「オリジナルデータの表示」を選択します。「レコードビューワ 」に、選択されたノードに対応するレコードが表示されます。

次の 第 4 章「MineSet のその他の機能」では、MineSet のその他の機能とクラシファイアの適用方法について学習します。

第 4 章

MineSet のその他の機能

この章では、MineSet のその他のツールを学習します。この章は、第 3 章「解約 (churn) データセットを使用するチュートリアル」のタスクが終了していることを前提としています。この章で学習する MineSet 機能は次のとおりです。

- 「データクラスタ」（37 ページ）
- 「デシジョン・テーブル (Decision Table)」（46 ページ）
- 「モデルによる顧客の絞り込み」（48 ページ）
- 「誤ったクラス判別のコストを削減」（56 ページ）
- 「MineSet のその他の機能」（62 ページ）

データクラスタ

内容がよく把握できていないデータセットであっても、クラスタリング・アルゴリズムを使用して興味深い属性や特性を発見できます。この非予測的なアルゴリズムは、レコードをさまざまな点で似かよったクラスタに分割します。この節では、「*Tool Manager*」ウィンドウに戻り、*churn.schema* ファイルをもう一度開いて新しい履歴から開始します。

1. 「データの可視化／マイニング」ウィンドウの上段タブの中から「マイニングツール」タブをクリックします。
2. 下段の「クラスタ」タブをクリックし、次のように選択します。

   「メソッド」:「単一 k-means」

   「クラスタ数」:「3」

3. Tool Manager の「データ変換」ウィンドウにある「現在のデータセットの項目名」ウィンドウから、次に示す項目を選択して削除します。これらの項目を強調表示し、「項目の削除」をクリックしてください（図 4-1 を参照）。

   「州 (state)」
   「取引期間 (account length)」
   「地域コード (area code)」
   「電話番号 (phone number)」
   「国際通話プラン (international plan)」（「国際通話の合計時間 (total intl minutes)」に関連があるため）
   「ボイスメール・プラン (voice mail plan)」（「ボイスメール メッセージの数 (number of voice mail messages)」に関連があるため）
   <Ctrl> キーを使用すると、複数の項目を選択できます。

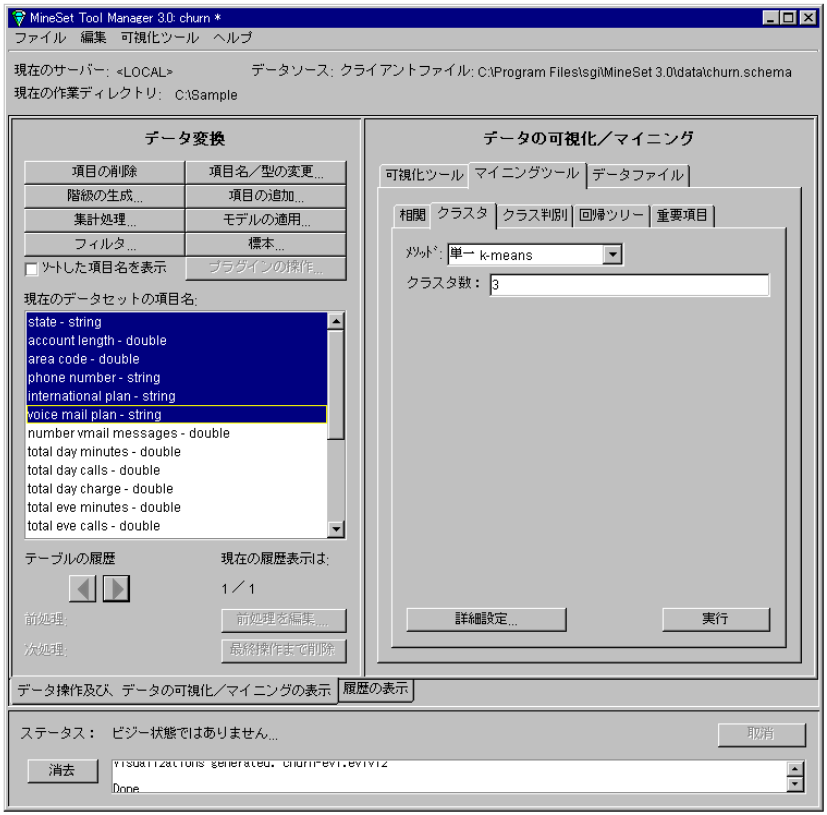


図 4-1 クラスタ化の前に行う項目の削除

削除した項目は、クラスタ化に実質的な影響を与えそうにないものばかりです。「解約 (churned)」項目は、結果の説明に使用できるようにそのまま残します。データセットをさまざまに試してみるうちに、どの項目を削除したらよいかが分かってきます。

4. 「詳細設定」をクリックし、属性の重みを設定します。

   デフォルトでは、各項目の重みは 1 に設定されています。これは、各項目の寄与率が同じであることを意味します。この例では、「解約 (churned)」項目を 0 に設定して、データセットをクラスタ化するとこの属性が自動生成されるかを確認します。「設定」、「了解」の順にクリックしてください。

5. 「*Tool Manager*」ウィンドウの右側の「実行」をクリックします。

   レコードをクラスタ化（グループ化）するための基準となる重要な特性をアルゴリズムが選択している間、Tool Manager の底部にあるステータス・ウィンドウには、クラスタ化の進行状況が表示されます。

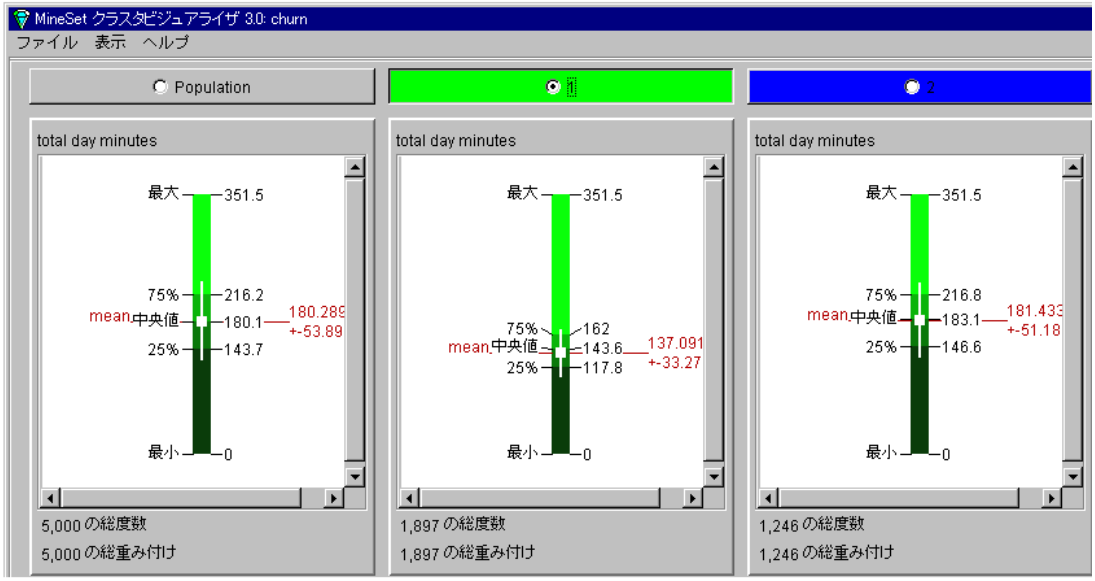


図 4-2 「クラスタ ビジュアライザ」が生成するボックスプロット

図 4-2 は、クラスタ・ビジュアライザ (Cluster Visualizer) による表示結果の一部を示しています。すべてのクラスタの全幅を見るには、ウィンドウを調整してください。項目は、クラスタ間の違いを識別する能力順にソートされます。「ボイスメール メッセージの数 (*number of voice mail messages*)」、「日中通話の合計時間

(*total day minutes*)」、「日中通話の課金合計 (*total day charge*)」は、明らかに最も重要な項目です。上位の項目ではクラスタによって色と平均値が大きく異なりますが、下方向にスクロールするにつれこの違いはわずかになります。

6. 画面上部のクラスタ番号の横にある円をクリックします。すると、そのクラスタを他のクラスタから識別する上で重要な属性の順序が変わります。
7. 「クラスタ ビジュアライザ」ウィンドウで、「ファイル」->「終了」を選択します。このウィンドウが閉じ、「*Tool Manager*」ウィンドウに戻ります。

モデルの項目と軸の対応付け

クラスタ・ビジュアライザ (Cluster Visualizer) を使用すると、データセット内の個々の属性を観察し、重要な属性や属性間の違いなどを調べることができます。しかし、クラスタ間で属性がどのように関連しているかを調べるには、スキャタ・ビジュアライザ (Scatter Visualizer) を使用した方が便利です。クラスタ化されたモデルをスキャタ・ビジュアライザに適用するには、それぞれの軸にどの項目を割当てるべきかを決定する必要があります。

1. Tool Manager の「データ変換 」ウィンドウで、「モデルの適用」をクリックします。続いて、利用できるモデルのリストから *churn.cluster* を選択し、「了解」をクリックします。

   クラスタ・ビジュアライザ (Cluster Visualizer) は 3 つの項目を最重要項目として示しましたが、各クラスタの重要度の順序は他のクラスタには関係がなく、属性間の相関は示しません。この時点では、重要項目ツールが有益です。

2. 「データの可視化／マイニング」ウィンドウの上段タブの中から「データファイル」タブをクリックし、続いて「サーバ」チェックボックスをクリックします。テキスト・フィールドでファイル名 **`churn-crop`** を入力し、「ファイルの作成」をクリックします。これで、この後の作業で使用する解約 (churn) データセットの簡略版が作成されます。

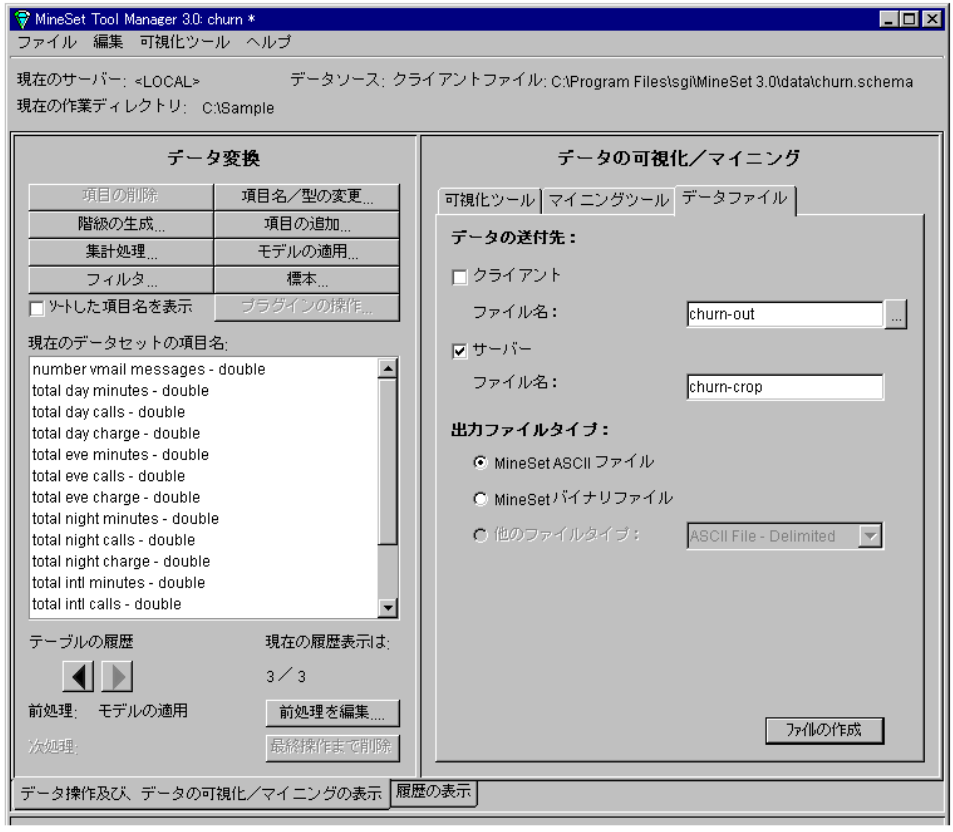


図 4-3 サーバへのデータファイルの保存

クラスタ化されたモデルの重要項目を確認

1. Tool Manager の「データの可視化／マイニング」ウィンドウで、「マイニングツール」タブをクリックし、続いて「重要項目」タブをクリックします。デフォルトでは、このツールは最も重要な項目を 3 つ選択します。「離散型ラベル 」は、「クラスタ (Cluster)」です。

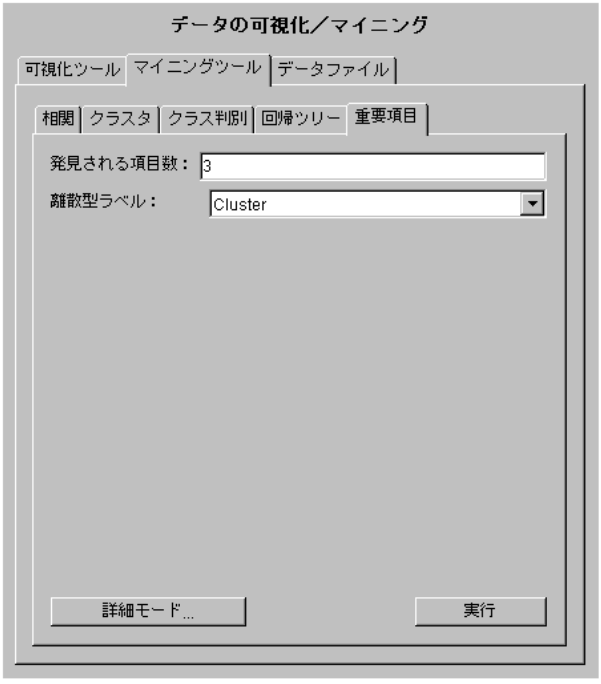


図 4-4 クラスタ化のための重要項目の選択

2. 「実行」をクリックします。

   次の項目を表示するパネルが表示されます。

   1.「ボイスメール メッセージの数 (number of voice mail messages)」
   2.「日中通話の合計時間 (total day minutes)」
   3.「夕方通話の合計時間 (total eve minutes)」

   ステータス・ウィンドウの表示から、日中の通話時間は要因、他の項目は相関を示すことが分かります。次は、これらの項目をスキャタ・ビジュアライザ (Scatter Visualizer) の軸に割当てます。

スキャタ・ビジュアライザ (Scatter Visualizer) への割当て

1. 「データの可視化／マイニング」ウィンドウの上段タブの中から「可視化ツール」タブをクリックします。続いて、下段の「スキャタ」タブをクリックし、スキャタ・ビジュアライザ (Scatter Visualizer) を表示します。

図 4-5 スキャタ ビジュアライザ (*Scatter Visualizer*) の軸に対する項目の割当て

2. 「データの可視化／マイニング」ウィンドウでプルダウン・メニューを使用し、項目に対して可視化要素を次のように割当てます（図 4-5 を参照）。

   「軸 1」:「ボイスメール メッセージの数 (number of voice mail messages)」
   「軸 2」:「日中通話の合計時間 (total day minutes)」
   「軸 3」:「夕方通話の合計時間 (total eve minutes)」
   「要素の色」:「クラスタ (*Cluster*)」（これはモデルの適用時に作成される）

3. 「ツールの起動」をクリックします。

   図 4-6 の「スキャタ ビジュアライザ」ウィンドウでは、色別にクラスタが表示されています。青のスキャタ・キューブはクラスタ 2 を示し、平らなパンケーキの形状は、赤（クラスタ 1）と緑（クラスタ 3）に均等に分割されています。このパンケーキは、「ボイスメール メッセージの数 (number of voice mail messages)」が非常に少ないことを示しています。「日中通話の合計時間 (total day minutes)」と「夕方通話の合計時間 (total eve minutes)」の間には、明らかに相関があることが分かります。興味のあるポイントをクリックすると、元データが表示されます。次の作業に進むため、「スキャタ ビジュアライザ」ウィンドウを閉じ、Tool Manager に戻ってください。

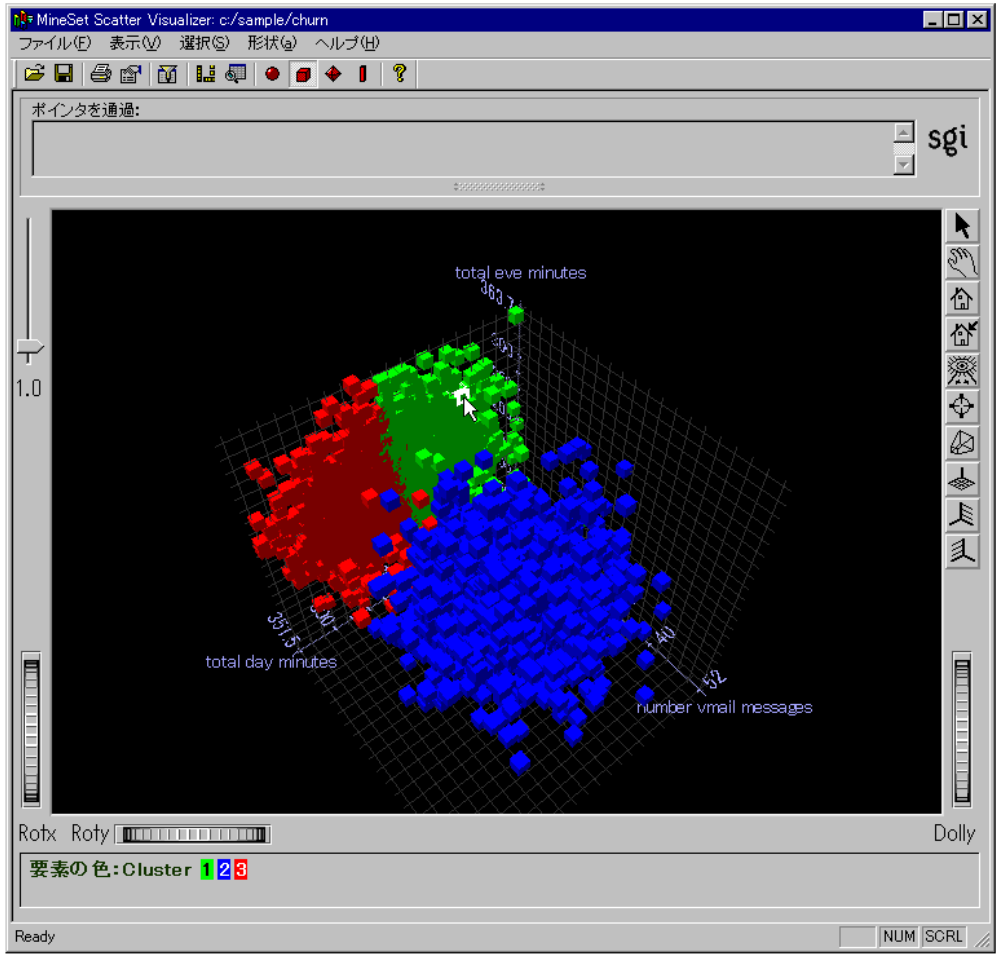


図 4-6 クラスタ化でプロットされたスキャタ ビジュアライザ (*Scatter Visualizer*)

デシジョン・テーブル (Decision Table)

先の例ではクラスタ化されたデータをスキャタ・ビジュアライザ (Scatter Visualizer) で表示しましたが、この例では同じデータをデシジョン・テーブル (Decision Table) として可視化します。

1. 「*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。
2. 「*.schema* ファイルを開く」ウィンドウで、「サーバファイル」ボタンをクリックし、*churn-crop.schema* を選択します。これは、先に保存しておいたファイルです。セッション間で MineSet を終了した場合は、最後に開いていた画面が自動的に表示されます。これは、Tool Manager の「ファイル」->「設定」メニューで設定できます。
3. ファイルをクリックして「開く」をクリックします。
4. 「データの可視化／マイニング」ウィンドウの上段タブの中から「マイニングツール」タブをクリックします。
5. 下段の「クラス判別」タブをクリックし、プルダウン・メニューから次のように選択します。

   「モード」:「クラシファイアとエラー」

   「分析」:「デシジョン・テーブル」

   「離散型ラベル」:「解約 (churned)」

   離散型ラベルが正しく選択されているか確認してください。次に、デシジョン・テーブルを実行し、X 軸と Y 軸に割当てる最重要項目をアルゴリズムから導きます。
6. 「提唱」チェックボックスがチェックされていることを確認し、「実行」をクリックします。

デシジョン・テーブル (Decision Table) ツールが適切な割当てを決定し、各軸に項目が割当てられていく様子が表示されます。Tool Manager の底部にあるステータス・ウィンドウには、進行状態とモデル生成プロセスのサマリ情報（クラス判別の誤差率など）が表示されます。モデル生成プロセスが終了すると、デシジョン・テーブル・ビジュアライザ (Decision Table Visualizer) が自動的に起動され、モデルが可視化されます。「*Dolly*」ダイヤルまたはマウスボタンを使用して画面上をナビゲートすることができます。（付録 A「MineSet ビジュアライザのナビゲーション」を参照）。

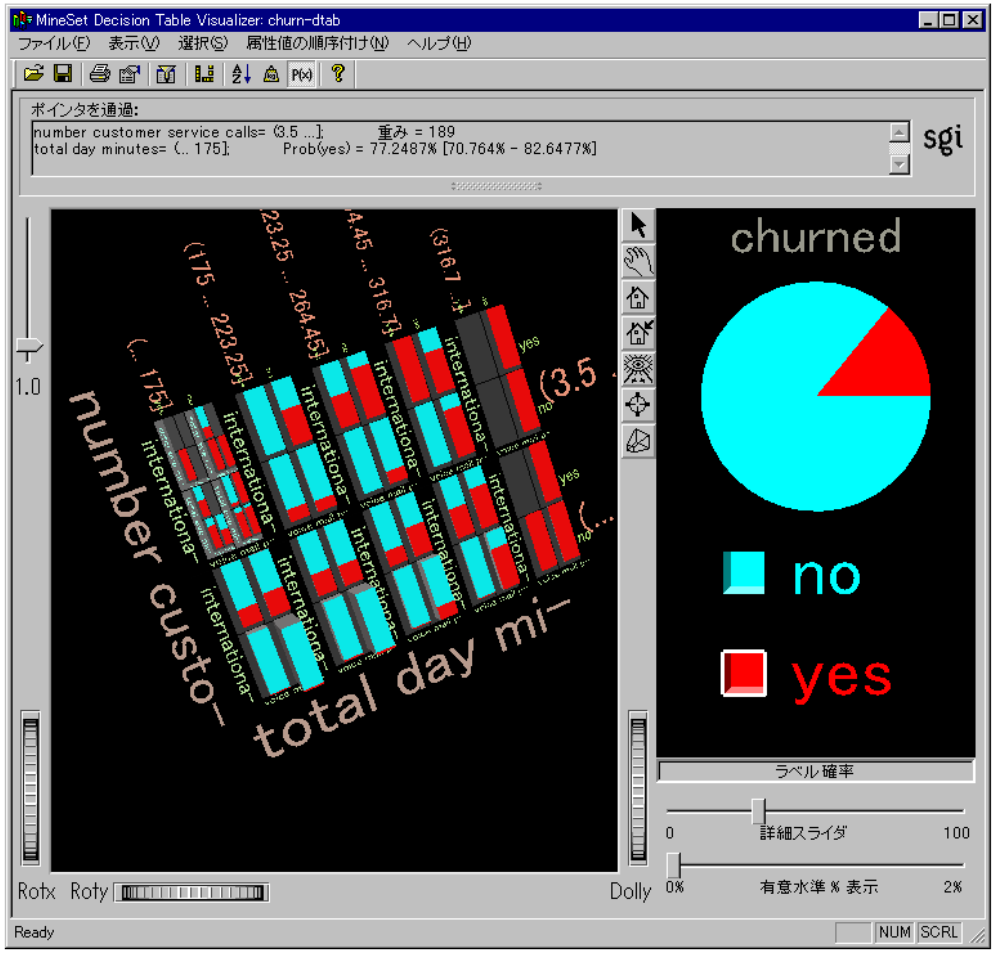


図 4-7 クラスタ化された解約結果を示すデシジョン・テーブル・ビジュアライザ (*Decision Table Visualizer*)

図 4-7 には、エビデンス・ビジュアライザ (Evidence Visualizer) と同様に、「ラベル確率」ウィンドウに全体の解約率を示す円グラフが表示されています。左側のウィンドウには、データのこのサブセット内の解約率を示すケーキグラフが示されています。このグラフから、日中通話の合計時間が高い顧客は解約の傾向が強いことが分かります。

デシジョン・テーブル (Decision Table) では、判定材料として最初はわずかな項目を使用し、徐々に詳細を追加していくというように、さまざまなレベルでデータを表示できます。カーソルモードをハンドモードからピックモードに切替えてグラフ上を移動し、ウィンドウ上部にデータを表示してください。「日中通話の合計時間 (total day minutes)」が 175 未満で、「顧客サービス通話の回数 (number of customer service calls)」が 3.5 を超える期待パターンから外れているバーに注目してください。ドリルアップとドリルダウンを行うには、マウスボタンを使用します（付録 A「MineSet ビジュアライザのナビゲーション」を参照）。デシジョン・テーブルの確認を終えた後このウィンドウを閉じ、Tool Manager に戻ってください。

モデルによる顧客の絞り込み

これまでは、解約しそうな顧客を予測するモデルをいくつか作成してきました。このようなモデルができたところで、次は解約が発生する前に解約しそうな顧客を絞り込めると便利です。このタスクには、改善曲線 (lift curve) が便利です。

改善曲線 (lift curve) は、X 軸が 0 ～ 100% のレコード数を示し、Y 軸が特定の分類名の値（この場合「解約 = はい」(`Churn=yes`)）を持つ顧客のレコード数を示すプロットです。図 4-10 のグラフには、2 つの曲線が表示されています。下の赤い曲線は、レコードをランダムに並べた場合に解約が見込まれる顧客の数を示しています。上の白い曲線は、各レコードに対するクラシファイアのスコア（推定確率）に基づいて並べた場合に解約が見込まれる顧客の割合を示しています。最初に表示されるのは、クラシファイアが最も解約率が高いと認める顧客のレコードです。解約があまり見込まれない顧客のレコードは最後に表示されます。クラシファイアによる順序付けの利点は、このモデルの曲線とランダム曲線間の違いが確認できることです。

改善曲線 (lift curve) を作成するには、選択されたモデルをテストセットに適用します。以下の例では、訓練事例としてデータセットの一部のセグメントを使用します。作成されたモデルは、データセットの残りの部分に適用します。改善曲線はクラシファイアの「詳細設定」から「改善曲線」を選択して簡単に生成できますが、このチュートリアルでは、標本の抽出と、データセットへのモデルの適用によるやや複雑な方法を紹介しています。

訓練事例の標本作成

この例では、Tool Manager の初期ウィンドウに戻り、新しい履歴を開始するため、「ファイル」->「新しいデータファイルを開く」を選択してローカルファイル *churn.schema* に戻ってください。

1. 「データ変換」ウィンドウで、「標本」をクリックします。「標本抽出」ダイアログ・ボックスで、標本抽出の割合として **40** と入力し、「了解」をクリックします。

   この設定によって、データセット全体の中から 40% のデータがランダムに選択されます。クラシファイアは、選択されたこの 40% のデータから生成されることになります。

図 4-8 テストセットの標本の選択

2. 「データの可視化／マイニング」ウィンドウの上段タブの中から、「マイニングツール」タブをクリックします。続いて、下段の「クラス判別」タブをクリックし、プルダウン・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」:「決定木」

   「離散型ラベル」:「解約」

   ランダムな 40% の標本抽出に基づいて決定木クラシファイアを作成し、訓練事例であるため「クラシファイアのみ」を選択します。テストセットは、40% の標本レコードを除いたデータセットの残りの部分に当たります。

3. 「実行」をクリックします。

作成された決定木 (Decision Tree) モデルは、次の段階で必要となるモデルを表します。標本のサイズがデータセット全体のサイズよりも小さいため、ルートの重み付けは実質的には減っています。各ノードのベースに色は表示されていません。これは、誤差推定が利用できないことを示しています。

ステータス・フィールドの表示から、クラシファイアが *churn-dt.class* という名前で自動的に保存されることが分かります。次は、解約 (churn) データセットの残りの部分にこのクラシファイアを適用します。

モデルの適用

「決定木」ウィンドウを閉じ、「*Tool Manager*」ウィンドウに戻ります。モデルの作成にはデータセットの 40% を使用したため、残りの 60% をテストセットとして使用します。

1. 「データ変換」ウィンドウで、「前処理を編集」をクリックします。「標本抽出」ダイアログ・ボックスが再び表示されます。
2. 「標本抽出」ダイアログ・ボックスで「パーセンテージ」テキスト・フィールドにもう一度 40 と入力し、今回は「補集合の標本」ボックスをクリックします。これは、標本の残りの部分を使用することを意味します。
3. 「了解」をクリックします。
4. 「データ変換」ウィンドウで、「モデルの適用」ボタンをクリックします。
5. 使用できるモデルのリストから、*churn-dt.class* を選択します。これは、解約 (churn) データセットに対して構築した決定木 (Decision Tree) モデルです。
6. ウィンドウの下側にある「モデルのテスト」タブをクリックします。「改善曲線の表示」をオンにし、「ROI/ 改善曲線のラベル」プルダウン・メニューを「はい」に設定します。

以上で、ランダムな標本に基づくクラシファイアの作成が終了しました。次は、これを 解約 (churn) データセットの残りの部分に適用します。

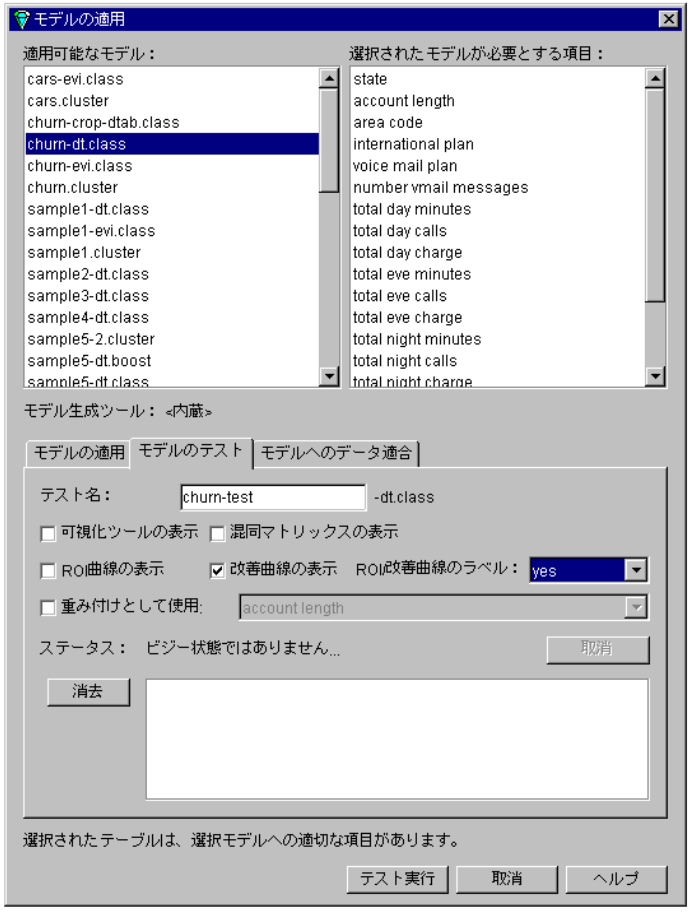


図 4-9 データセット全体を使用してクラシファイアをテストするための準備

7. 「テスト実行」をクリックします。処理に多少時間がかかります。生成された改善曲線 (lift curve) を図 4-10 に示します。選択されたポイントの詳細は、上部のバナーに表示されます。

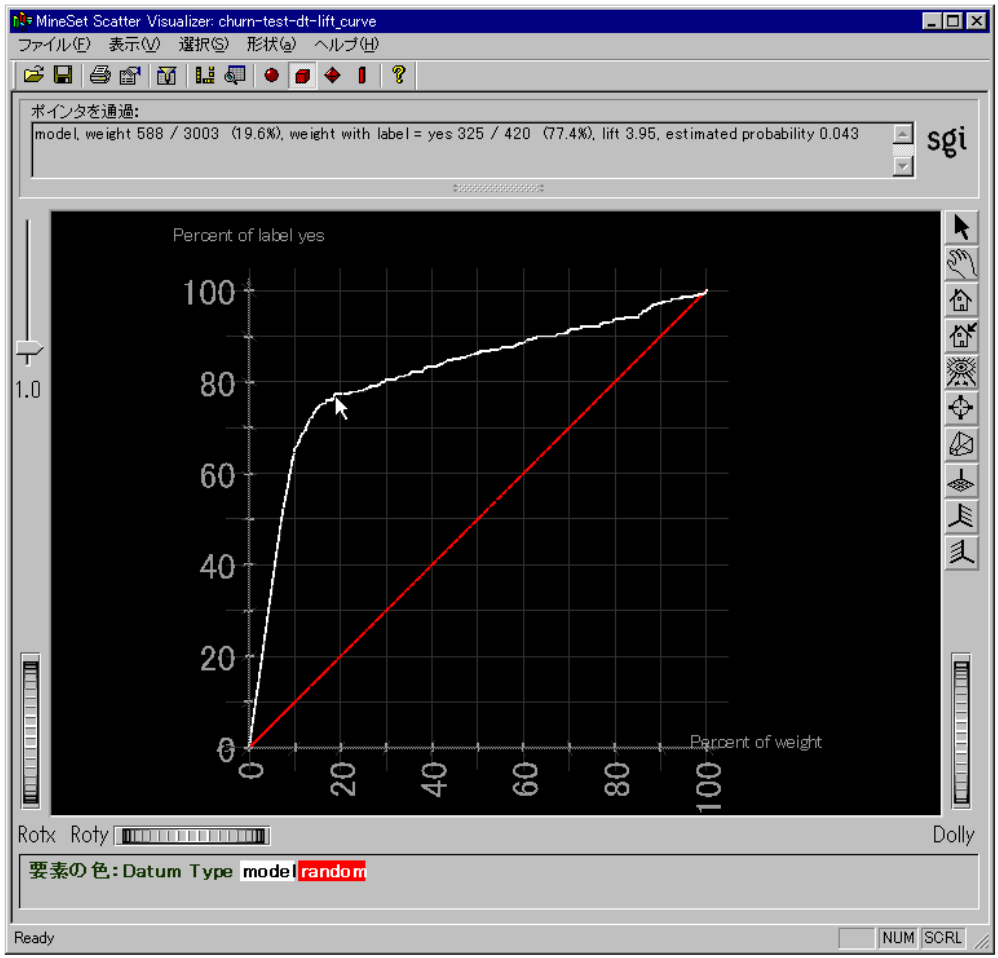


図 4-10 改善曲線 (lift curve)

ポインタを白い（モデル）ラインに沿って動かし、さまざまな位置をクリックすると、「解約 = はい (`churn=yes`)」の顧客の改善状態と割合が表示されます。曲線が大きなカーブを描いている箇所を見つけてください（この例では分析モデルの推定確率が 0.043 の部分）。

これは、解約しそうな顧客に対する刺激策の投資利益率が急激に落ちるポイントです。次は、分析モデルをデータセット全体に適用します。

8. 「モデルの適用」ダイアログ・ボックスに戻ります。「モデルの適用」タブをクリックして *churn-dt.class* を選択し、次のように選択してください。

   「次の分類名となる推定確率値」:「はい (yes)」

   「新しい項目名」: **`p_churned`**（このように入力する）

   「次の分類名となる推定確率値」をクリックする際、「モデルのテスト」タブ内の対応する選択に一致するように「はい」が選択されています。このプロセスで、特定の人が解約する確率 (p_churned) を表す新しい項目を追加します。「了解」をクリックしてください。

9. Tool Manager の「データ変換」ウィンドウで、「フィルタ」をクリックします。「表現による定義」テキスト・フィールドで、表現 **`p_churned > 0.043`** を作成します。表現を確認して「了解」をクリックします。

これは、図 4-10 に示された手順 8 から取得された推定確率値です。この目的は、解約の確率が最も高い顧客だけを選択することです。実用においては、既存の顧客の中から解約しそうな顧客を予測するために、ラベルを付けられていないデータに対してこの操作が行われます。

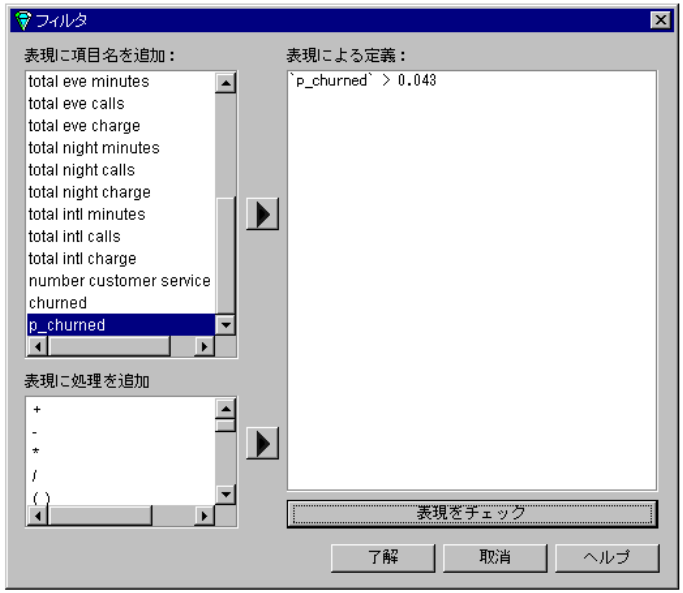


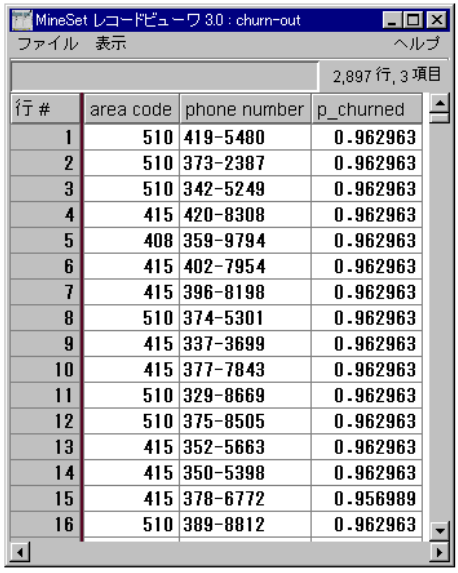
図 4-11 解約の確率を調べるためのフィルタリング

最後に、参照しやすいように不要な項目を削除し、レコードビューワ (Record Viewer) に結果を表示します。

10. Tool Manager の「データの可視化／マイニング」ウィンドウで、「可視化ツール」タブ、「レコード」タブの順にクリックします。「データ変換」ウィンドウで、「地域コード (area code)」、「電話番号 (phone number)」、「解約の確率 (p_churned)」を除くすべての項目を選択し、「項目の削除」ボタンをクリックします。<Shift> キーを使用すると、連続した複数の項目を選択できます。<Ctrl> キーを使用すると、任意の複数の項目を選択できます。

11. 「ツールの起動」をクリックします。

    図 4-12 に示すように、モデルに基づいて解約確率が最も高い顧客の電話リストが生成されます。



行 #	area code	phone number	p_churned
1	510	419-5480	0.962963
2	510	373-2387	0.962963
3	510	342-5249	0.962963
4	415	420-8308	0.962963
5	408	359-9794	0.962963
6	415	402-7954	0.962963
7	415	396-8198	0.962963
8	510	374-5301	0.962963
9	415	337-3699	0.962963
10	415	377-7843	0.962963
11	510	329-8669	0.962963
12	510	375-8505	0.962963
13	415	352-5663	0.962963
14	415	350-5398	0.962963
15	415	378-6772	0.956989
16	510	389-8812	0.962963

図 4-12 レコードビューワ (Record Viewer) の生成結果

レコードビューワ (Record Viewer) には、レコードごとにその顧客の解約確率を予測した値が示されます。フィルタリングによって、最も解約確率の高い顧客だけが表示されています。これにより、刺激策（電話による懇請、メールの送付など）を採るべき、解約の可能性が高い顧客だけが示されます。レコードビューワ (Record Viewer) を閉じて、解約 (churn) データセットによる作業を継続します。

誤ったクラス判別のコストを削減

MineSet では、3 つの重要なツールを使用して、モデル作成における誤ったクラス判別のコスト（損失）を削減することができます。その 1 つ、混同マトリックス (Confusion Matrix) は、誤差や不正な予測を詳細に提示します。損失マトリックス (Loss Matrix) は、間違いの程度差を考慮します。3 つ目の投資利益率曲線 (ROI Curve) は、時間や資金を投資しても無駄な場合を示します。

混同マトリックス (Confusion Matrix) の表示

「*Tool Manager*」ウィンドウに戻り、*churn.schema* をもう一度開きます。

1. 「データの可視化／マイニング」ウィンドウで、「マイニングツール」タブをクリックします。続いて、「クラス判別」タブをクリックし、プルダウン・メニューから次のように選択します。

   「モード」:「クラシファイアとエラー」

   「分析」:「決定木」

   「離散型ラベル」:「解約」

2. 「詳細設定」をクリックします。図 4-13 に示した「クラシファイア詳細オプション」ウィンドウが表示されます。

   処理中に、100 個を超える値が存在するために「電話番号 (phone_number)」が削除されることを示す警告メッセージが表示される場合があります。この場合は、「了解」をクリックして継続してください。

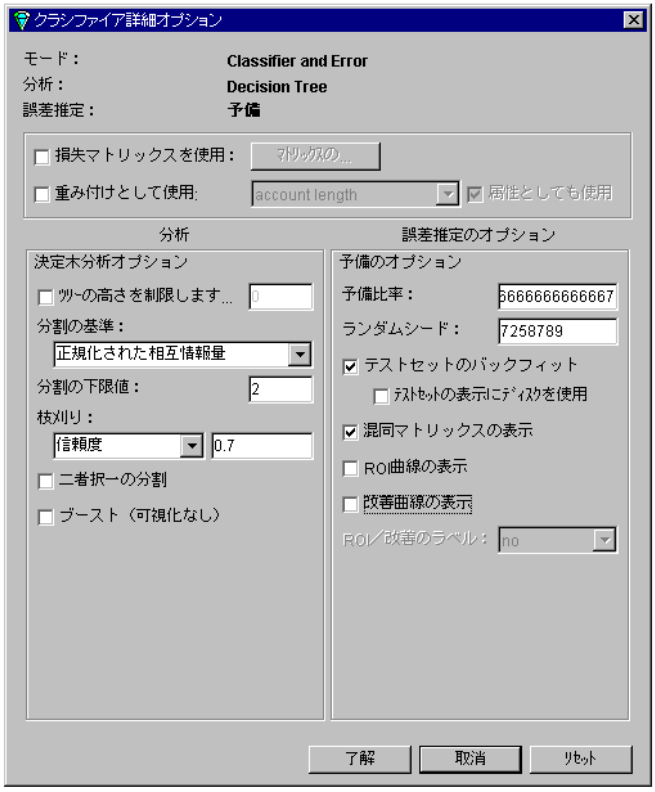


図 4-13　「クラシファイア詳細オプション」パネル

3. 「混同マトリックスの表示」と「テストセットのバックフィット」をオンに設定します。「改善曲線の表示」と「ROI 曲線の表示」がオフに設定されていることを確認し、「了解」をクリックします。
4. Tool Manager の「データの可視化／マイニング」ウィンドウで、「クラス判別」ウィンドウから「実行」をクリックします。

混同マトリックス (Confusion Matrix) に、クラシファイアがクラス判別を間違えた位置が示されます。ツリー・ビジュアライザ (Tree Visualizer) を閉じて、混同マトリックスを確認してください。ここから、データについて得た情報を基に損失マトリックス (Loss Matrix) を作成し、他の誤差と比較して一部の誤差が許容されにくくなるように設定できます。

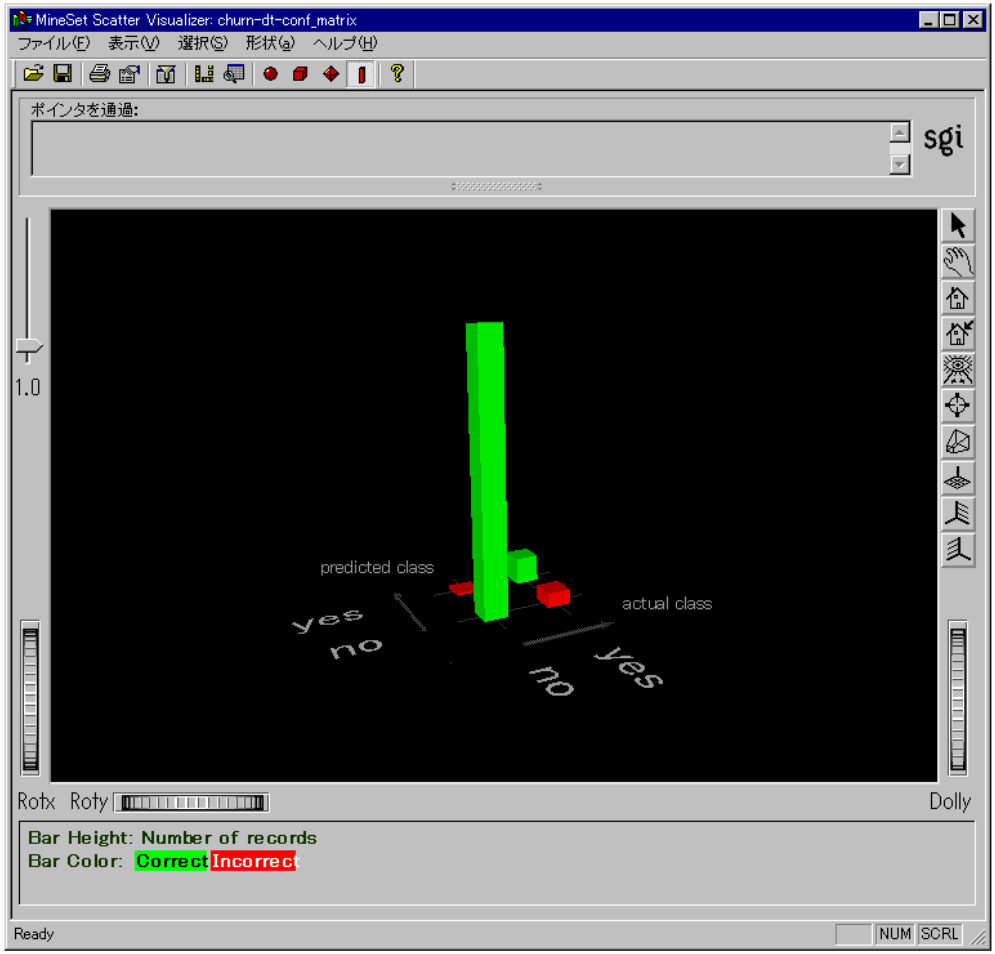


図 4-14 正しいクラス判別と不正なクラス判別を示した混同マトリックス (Confusion Matrix)

図 4-14 に示したウィンドウでは、2 つの緑のバー（高いバーと低いバー）は正しいクラス判別を示します。2 つの赤のバーは不正なクラス判別を示します。赤いバーの大きい方が示すカテゴリ（「予測されたクラス : no」、「現在のクラス : yes」）では、かなりの分析の誤りが発生しています。これらの顧客は解約しないものと予測されましたが実際には解約しており、4.7% という手痛い誤差が生じています（割合を見るには「選択」->「値の表示」を選択）。データから得た情報に基づいて損失マトリックス (Loss

Matrix) を作成して誤差を減らし、この赤いバーが示す誤差にさらに大きな重みを付けることができます。

損失マトリックス (Loss Matrix) の定義

損失マトリックス (Loss Matrix) を作成する目的は、クラシファイアにとって好ましいものとそうでないものの誤差を制御することです。

1. 「ファイル」->「終了」を選択して混同マトリックス (Confusion Matrix) を閉じ、「*Tool Manager*」ウィンドウに戻ります。
2. 「詳細設定」をクリックして、「クラシファイア」オプション・ウィンドウに戻ります。
3. 「クラシファイア詳細オプション」ダイアログ・ボックスで、「損失マトリックスの使用」をオンに設定します。
4. 誤差という損失に重みを付けるため、「マトリックスの編集」をクリックします。図 4-15 のような「損失マトリックス」ウィンドウが表示されます。

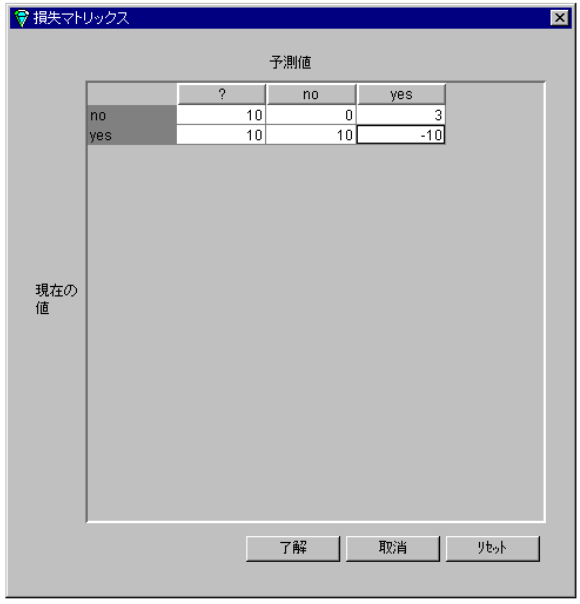


図 4-15 重み付けを示す損失マトリックス (Loss Matrix)

5. 損失マトリックス (Loss Matrix) の行に、左側から順に次のように値を設定します。

   「現在の値」:no: 10--0--3

   「現在の値」:yes: 10--10--(-10)

   疑問符 (?) の下の項目内の値は、分析モデルが「未知」を予測しないようにいくぶん高めに設定する必要があります。

これらの値を使用してある顧客が解約しないと予測し、その予測が正しかった場合、損失も利益もありません（0 が表示される）。ある顧客が解約しないと予測し、刺激策を何も施さず、実際には解約が発生した場合は、損失 10 となります（この値は損失であるため、正の数字として示される）。ある顧客が解約すると予測し、実際は解約しなかった場合、不要なメールを送ったため損失 3 となります。メール対策が利いて解約しそうだった顧客を思いとどまらせた場合、10 を獲得します（負の数字として示される）。次は、投資利益率について調べます。

ROI 曲線 (ROI curve) の表示

ROI 曲線 (ROI curve) は、特定の誤差が生む損失を表示するとともに、顧客への刺激策が効果を出さなくなるポイントを示します。

1. 「クラシファイア詳細オプション」ウィンドウで、「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」がオンに設定されていることを確認します。
2. 「ROI 曲線の表示」がオンに設定されていることを確認します。
3. 「ROI/ 改善曲線のラベル」が「はい」に設定されていることを確認し、「了解」をクリックします。
4. 「*Tool Manager*」ウィンドウの「クラス判別」ウィンドウで「実行」をクリックします。

3 つのウィンドウ、「決定木 」、「混同マトリックス」、「ROI 曲線」が表示されます。「混同マトリックス」は、クラシファイアが「解約 = いいえ (churn=no)」予測に慎重であることを示します。このため、不正な負の値が減っています。片方の誤差は増えていますが、もう一方の誤差は減っています。「混同マトリックス」ウィンドウ、「決定木」ウィンドウとも閉じて、次は「ROI 曲線」ウィンドウを調べます。

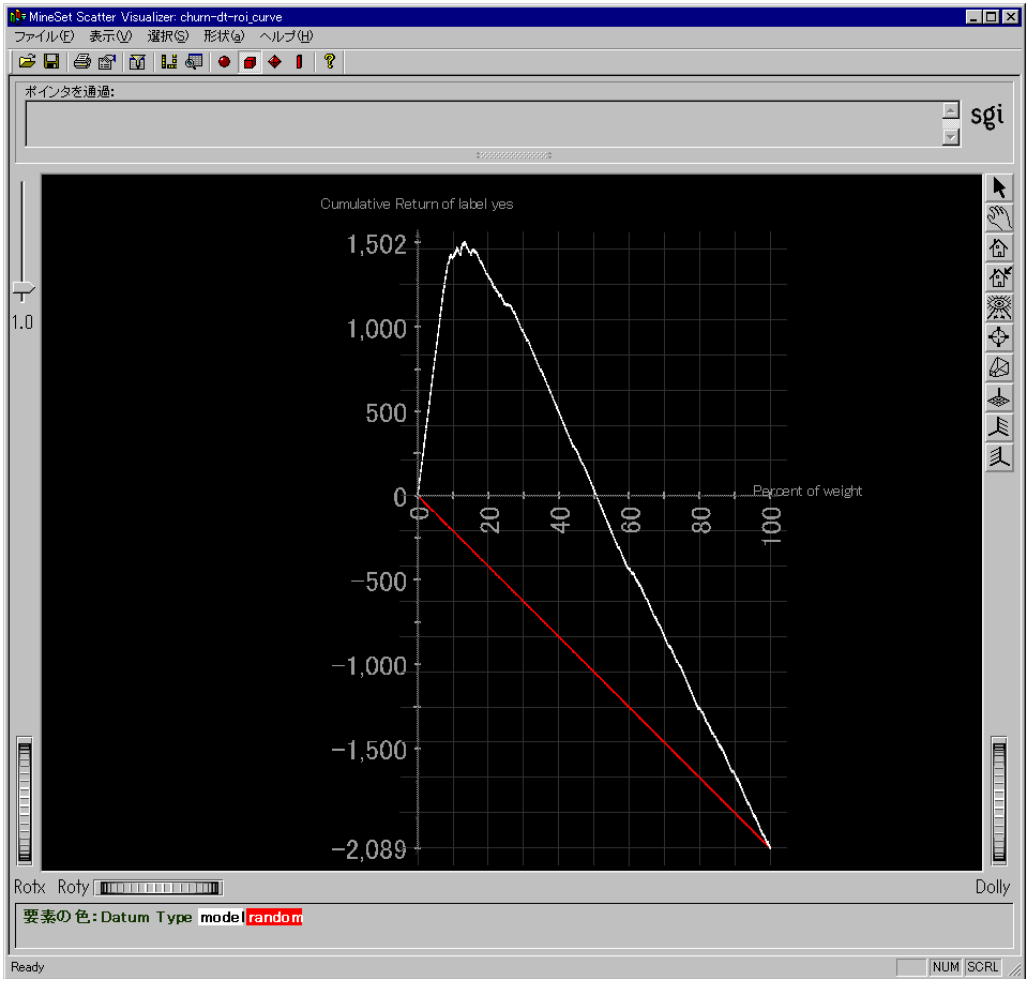


図 4-16 ROI 曲線 (ROI curve)

図 4-16 の ROI 曲線 (ROI curve) は、改善曲線 (Lift curve) によく似ています。中央の水平ラインは、ゼロの損益を示します。赤いラインは、母集団からランダムに標本抽出してメールを送った場合の予測パフォーマンスを示します。顧客全員にメールを送ると、郵便代金のために損失が予想されます。しかし、曲線が大きなカーブを描いている 1,488 人（母集団の 12.6%) を指す位置では、最も高い投資利益率が得られます。

MineSet のその他の機能

MineSet の各種ツールとデータマイニング・アルゴリズムの詳細については、『*MineSet 3.0 Enterprise Edition User's Guide*』、『*MineSet 3.0 Enterprise Edition Reference Guide*』、『*MineSet 3.0 Enterprise Edition Interface Guide*』を参照してください。『*MineSet 3.0 Enterprise Edition User's Guide*』は、「ヘルプ」->「*MineSet User's Guide*」を選択してオンラインで見ることができます。

このチュートリアルでは、MineSet の一連のツールについて簡単に紹介しました。『*MineSet 3.0 Enterprise Edition User's Guide*』では、次の機能について説明しています。

- スキャタ・ビジュアライザ (Scatter Visualizer)
- 階層を可視化するツリー・ビジュアライザ (Tree Visualizer)
- 選択式決定木分析 (Option Tree Inducer)/ クラシファイア (Classifier)
- 相関規則分析 (Association Rules Generator) と相関規則ビジュアライザ (Association Rules Visualizer)
- 離散値ではなく連続値を予測する回帰分析 (Regression)
- 階級生成、分割、配列のインデックスなどのデータ変換
- レコードの重み付け。レコードごとに寄与率が異なるため（たとえば、収益性の高い顧客など）、各レコードに異なる重みを割当てることができます。
- 学習曲線 (Learning Curve)。作成したクラシファイアの精度をあまり低下させずに、知識発見プロセスの速度を上げるために、データセットに対し標本抽出が可能かどうかを決定できます。
- カラー操作、メッセージ・ボックスなどの多様なツールオプション
- 可視化ツール用のアニメーション・スライダ
- バッチ処理。mineset_batch プログラムを使用して、非対話的に演算を実行できます。これは、ジョブを定期的に実行する必要がある場合に便利です（毎晩 1 回実行するなど）。
- 相互検証などの高度な手法による誤差推定

『*MineSet 3.0 Enterprise Edition User's Guide*』には、ファイルとデータ操作についての技術的な詳細も説明されています。

注記：データマイニング・アルゴリズムは、因果関係がないような相関を見つけます。たとえば、靴のサイズと読解力の間は強い関係があるという有名な説があります。これは、靴のサイズが大きくなるにつれ、読解力が向上するというものです。この相関は事実ですが、因果関係はありません。靴のサイズも読解力も年齢とともに向上します。つまり、子供が成長するにつれ、靴のサイズ、読解力とも増えます。したがって、発見された相関に因果関係を結び付けないように注意する必要があります。大きな靴を履いても、読解力がよくなることはないからです。

MineSet ビジュアライザのナビゲーション

ツリー・ビジュアライザ (Tree Visualizer) 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（決定木、選択式決定木、回帰ツリー）で使用できるナビゲーション・コントロールのクイック・リファレンスとして役立つ 2 つの一覧表を示します。まず、表 A-1 にナビゲーション・ボタンの一覧を示します。

表 A-1 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。
	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 A-2 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 A-2 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
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 A-3 にナビゲーション・ボタンの一覧を示します。

表 A-3 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。
	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）。

表 A-3 (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
	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・ビジュアライザで使用できる調整スライダとダイヤルの一覧を表 A-4 に示します。

表 A-4 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
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	「Height」スライダ（左上）	該当なし
X 軸を中心にシーンを回転する	「Rotx」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	「Rotx」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	「Dolly」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 A-4 (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
あまり重要ではない属性をフィルタで除外する	「詳細スライダ」（エビデンス・ビジュアライザとデシジョン・テーブル・ビジュアライザのみ）	該当なし
レコードの重みが、データセット内のレコードの合計の重みに対して指定した割合よりも小さい属性値（最大で 2%）をフィルタで除外する	「有意水準 % 表示」スライダ（エビデンス・ビジュアライザとデシジョン・テーブル・ビジュアライザのみ）	該当なし
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）。