The Power to Know.

SUGI-J 2001

Enterprise Miner™ SoftwareVersion4.1
ハンズオンワークショップ

2001 年 7 月 26 日～27 日
株式会社 SAS インスティチュート ジャパン
カスタマーサービス本部

The Power to Know.

Ⅰ. Introduce SAS Enterprise Miner™

ⅰ. はじめに

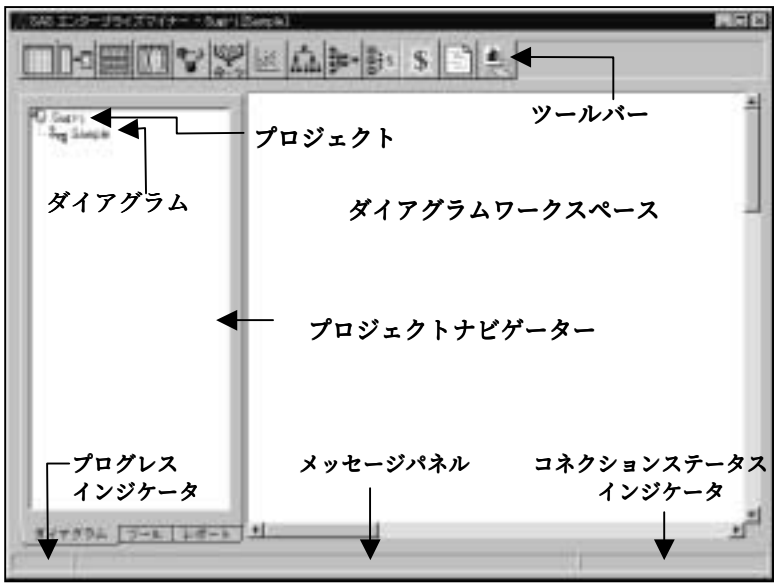
Discover the diamonds in your data with SAS Enterprise Miner™

Enterprise Miner™ソフトウェア はデータマイニングのプロセスを簡便化、合理化させ、様々な手法をサポートしたソリューションを提供します。
データマイニングとは、大規模なデータから有益な情報を導き出す為の一連のプロセスであると言えます。具体的には、分析データの準備、データの分布の調査、変数変換、クラスタリングなどの予備作業の後、回帰分析やニューラルネットワークなどの手法によるモデル化を行い、実行結果の有用性と信頼性を評価するという流れになります。
Enterprise Miner™ソフトウェアでは、この流れを、「サンプリング」・「探索」・「加工」・「モデル化」・「評価」の5つの過程に分類しています。
また SAS Data Warehousing と OLAP テクノロジーを兼ね備え、広範囲な Knowledge Discovery を行います。

ⅱ. Enterprise Miner™ ソフトウェアバージョン 4.1 の拡張点

インターフェース

Enterprise Miner™ソフトウェアバージョン 2.02（以下 EM2 と表記）では、プロジェクト、ノードタイプ、ワークスペースの各ウィンドウが個別に存在していましたが、Enterprise Miner™ソフトウェアバージョン 4.1（以下 EM4 と表記）のインターフェースでは、全てが一つのウィンドウに表示されるようになりました。これによりプロジェクト管理が容易になりました。



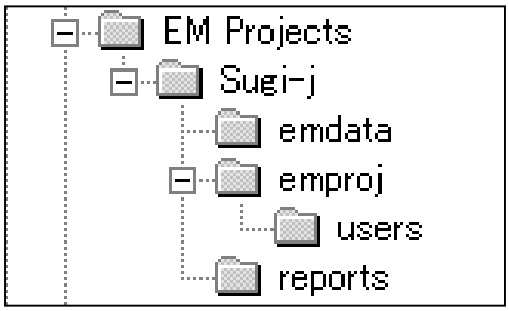
Enterprise Miner™ソフトウェアバージョン 4.1 のインターフェース

プロジェクトナビゲーターはプロジェクトダイアグラムとツールパレット、そしてレポートノードによって作成される HTML ファイル管理をするレポートタブで構成されています。
ツールバーには、プロセスフローを構築する際、頻繁に使われる一部のツールを収容しています。

プロジェクトナビゲーターのツールパレット

プロジェクト

EM2 においては、2 種類の SAS ライブラリを定義する必要がありました。1つはプロジェクト情報を保持しておくプロジェクトライブラリ、もう 1 つは一連のプロジェクトを通して作成される中間データファイルを保持しておくデータライブラリです。そして、これらのライブラリはユーザによりそれぞれにライブラリ参照名を定義する必要がありました。
EM4 では、これらのライブラリを自動的に作成することでユーザ定義設定、並びにプロジェクト管理の簡易化を図りました。これにより、プロジェクト名と分析環境を指定するだけで新しいローカルプロジェクトを作成できるようになりました。また、1プロジェクトに対して 100,000 ダイアグラムまで作成が可能です。



プロジェクトは上図のディレクトリ構造を自動的に作成します。

ディレクトリ「emproj」には、ダイアグラムごと、ノードごとのターゲットプロファイル情報、その他様々な設定情報を保管されます。更に複数のユーザが同一ダイアグラムに同時アクセスすることを避ける為のダイアグラムロックファイル（*.lck）が保管されます。サブディレクトリ「USERS」には現在ダイアグラムを使用しているユーザを明示するファイルが保持されます。

ディレクトリ「emdata」は分析を進める間に作成される、潜在的に大きなサイズのファイルが保管されます。クライアント/サーバ環境で分析を実行した場合はサーバのデータディレクトリを記したファイルが保持されます。

レポーターノードによって作成された HTML レポートはディレクトリ「Reports」に保存されます。

ターゲットプロファイラ

EM4 で追加された機能の1つにターゲットプロファイラがあります。EM4では、ある意思決定を行った時に生じるコストや利益情報をターゲットプロファイルと呼んでいます。これにより、ユーザはコストを最小化もしくは利益を最大化するような意思決定を選択することができ、コスト意識をもったデータマイニングが可能となります。

以下のノードでターゲットプロファイルを定義、または確認することができます。

- 入力データソースノード
- データセット属性ノード
- ニューラルネットワークノード
- ツリーノード
- 回帰分析ノード
- アンサンブルノード
- ユーザ定義モデルノード

予測モデルの精度に応じた影響力を測定するアセスメントノードにおいて、損益関係を対話的に再定義できます。

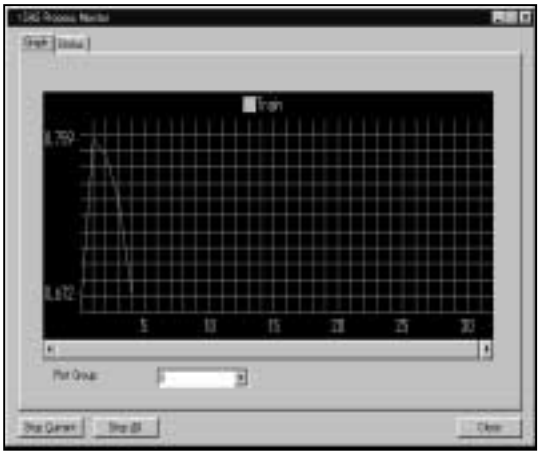
SAS プロセスモニタ

SAS プロセスモニタを使用すると、計算が集中的に行われるプロセスを追跡したり、停止したりすることができます。Enterprise Miner リリース 4.1 では、ニューラルネットワーク、回帰分析、および SOM/Kohonen の各ノードが、SAS プロセスモニタを使用します。SAS プロセスモニタを使用すると、ローカルマシンおよびクライアント／サーバーにおけるプロジェクトの学習をモニタできます。

SAS プロセスモニタを使用するには、Windows の環境変数として「IP_ADDRESS」とその値としてモニタするコンピュータの IP アドレスを設定する必要があります。［エディタ］ウィンドウ、または［プログラムエディタ］ウィンドウにおいて、以下のステートメントを実行してから SAS プロセスモニタを起動します。

options set=IP_ADDRESS "IP-Adress";

（IP-Adress には、モニタするコンピュータの IP アドレスを指定します）



SAS プロセスモニタ

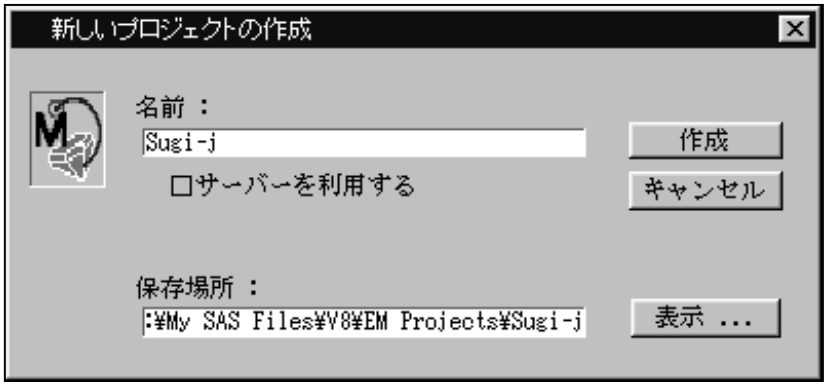
Ⅱ. Set up The Enterprise Miner™ Software

ⅰ. Opening The Enterprise Miner

EM4 を起動するには、**メニュー**の**[ソリューション]→[データ解析]→[エンタープライズマイナー(E)]**を選択します。

ⅱ. Setting Up The Initial Project and Diagram

1. **メニュー**から**[ファイル]→[新規作成]→[プロジェクト]**を選択します。
2. **プロジェクト名**(ここでは**「Sugi-j」**とします)を入力します。



3. **「作成」ボタン**を選択します。これにより**「Sugi-j」**という名称のプロジェクトで**「無題」**のダイアグラムがオープンします。

4. 無題のダイアグラムをクリックし、タイトルを新しく入力(ここでは**「Sample」**と入力)します。

クライアント/サーバ環境で実行する際はチェックボックスを選択します。
またデータライブラリ、プロジェクトライブラリはプロジェクト名を入力すると自動的にディレクトリが作成されますので、任意に作成したい場合は「表示」をクリックし、割り当て作業が必要です。

iii. Data Mining using SEMMA

Enterprise Miner™ソフトウェアが提供するデータマイニングにおける機能は SAS の提唱する SEMMA プロセスに則っています。
SEMMA の意味は以下のとおりです。

Sample（サンプル）- 膨大なデータから傾向を損なわずサンプルを作成し、また学習用、評価用、テスト用のデータに分割します。

Explore（探索）-データの特徴を統計的、視覚的に捉えます。データのグラフ化、記述統計量の計算、変数の選択、アソシエーション分析を可能にします。

Modify（加工）-分析の為のデータを準備します。変数の新たな追加や既存変数の変換、欠損値の補填を行います。さらにクラスター分析、自己組織化マップの作成を可能にします。

Model（モデル）-回帰モデル、決定木モデル、ニューラルネットワークモデル、ユーザ定義モデル等、予測モデルを当てはめます。

Assess（評価）-複数の予測モデルについて反応率、計算された予測反応率、リフトチャート、プロフィットチャートを作成し、予測精度の比較を可能にします。

ユーティリティグループの中にはその他の付加ツールが置かれています。

Overview of the Node

■ サンプルノード

入力データソースノードはデータソースの読み込み、分析プロセスにおける各変数の役割や属性の定義等、様々な役割を担っています。
入力データソースノードはデータセットやデータマートにアクセスすることができます。データマートは SAS Data Warehouse Administrator™ソフトウェアによって定義され、Enterprise Miner Warehouse Add-in によって作成されるものです。入力されたデータの各変数についてのメタデータサンプルを自動的に作成し、各変数の測定水準や分析における役割を自動的に設定します。必要に応じてユーザによる定義設定も可能です。間隔変数と分類変数それぞれの要約統計量を算出します。入力されたデータセットについてのターゲットプロファイル定義が可能です。

サンプリングノードではランダムサンプリング、層別サンプリング、そしてクラスターサンプリングが可能です。サンプリングは特に大きなデータベースを扱う際に、モデルの学習時間を短縮させます。また、サンプルが母集団の傾向を十分に捕捉したものであれば、サンプルにおける分析結果を母集団に当てはめることができます。サンプリングノードはサンプル抽出したレコードをデータセットに書き出し、乱数を発生させる際に使われたシード値を保存しますので、サンプルを再作成する事が可能です。

データ分割ノードは学習用データ、テスト用データ、そして評価用データにデータセットを分割する事します。学習用データは主にモデルの適合に用いられます。評価用データは推定処理におけるモデルのチェックやモデルの重みの微調整、またモデルの精度評価に用いられます。テスト

用データは付加的なモデルの評価に用いることができます。このノードでは単純無作為サンプリング、層別サンプリング、それ以外のユーザ定義の分割方法によりデータを作成することができます。

■ 探索ノード

分布エクスプローラノードは多次元ヒストグラムを描画し、手早く簡易にデータを視覚化するツールです。このノードは 3 変数まで、同時に分布を表示する事ができます。また変数が二値尺度、名義尺度、順序尺度の場合は、チャートから除外したい変数を指定することができます。間隔尺度において異常値(外れ値)を除外することも範囲設定で可能です。またこのノードでは、チャート表示した変数について要約統計量を算出します。

マルチプロットノードは大量のデータの視覚的な探索を可能にするツールです。Insight ノードや分布エクスプローラノードと違い、メニューやウィンドウアイテムを選択することなく自動的に Input 変数や Target 変数の棒グラフや散布図を作成します。このノードではグラフ作成の SAS コードを自動生成しますので、バッチ処理の環境に流用する事が可能です。

INSIGHT ノードは SAS/INSIGHT™ソフトウェアセッションによって対話的にデータの探索や分析を行います。複数のウィンドウにわたってリンクされたグラフや分析結果を用いてデータ探索を行うことができます。一変量および多変量の分析、さらに一般線形モデルの当てはめを行うことができます。

アソシエーションノードはデータ内の相関ルールを見つけ出します。例えば、パンと牛乳の購入関係を見たい場合等です。このノードは時点を表す変数がデータにある場合、逐次発見を行うことも可能です。

変数選択ノードは Target 変数を予測または分類する際、Input 変数の重要性を計算します。重要な変数の選択は、R2 乗値(決定係数)もしくは χ 2 乗値を基準に行われます。Target 変数に関連のない変数は、マイニングのプロセスにおける分析から除外されます。除外された変数もプロセスフローの後続ノードに引き渡されますが、ニューラルネットワークやツリーノードなどのモデル作成ノードでは Input 変数としては使用されません。

■ 加工ノード

データセット属性ノードでは、データセット名、説明、役割のようなデータセット属性を加工することができます。またデータセットに関連のあるメタデータサンプルを加工でき、更にターゲットに対してのターゲットプロファイルが可能です。SAS コードノードで生成されたデータセットのメタデータサンプルを加工したい場合などに用います。

変数変換ノードでは変数の変換行います。平方根、自然対数、Target 変数との相関の最大化、正規性の最大化といった手法で変換を行います。さらにユーザ定義の変換式もサポートし、ビジュア

ルインターフェースによる間隔尺度変数のバケットや分位数毎のグルーピングも簡易に行うことができます。また、このノードでは自動的に決定木のアルゴリズムを使用して間隔尺度変数をビン化します。変数変換は、モデルの最適な当てはめや、モデルの予測精度に大きな影響を与えます。

外れ値のフィルターノードでは外れ値の発見、除外を行います。外れ値はモデルに大きな影響を及ぼすので、事前にチェックすることを推奨します。

データ置き換えノードでは欠損値のあるオブザベーションに代替値を埋め込むことが可能です。間隔尺度変数の置き換えには、平均値、中央値、範囲中央値、中央値-最小値スペース、分布ベースの統計量、Tukey の双 2 次加重、Huber's、Andrew の波形 M 推定値などが用いられます。また置き換える値をツリー補充法により推定することも可能です。分類尺度変数は、最頻値、分布ベース、ツリー補充、または定数で置き換えられます。

クラスタリングノードではデータをセグメント化します。このノードでは K-means 法により類似したオブザベーションが同一クラスターに集約されます。それぞれのオブザベーションに付けられたクラスター番号は、後続のノードで Input、ID、Target などの役割を定義して使用することが可能です。デフォルトではグループ変数として引き渡されます。

SOM/Kohonen ノードは自己組織化マップ、Kohonen ネットワーク、そしてベクトル量子化ネットワークを生成します。このノードは教師なし学習でデータ構造の学習を行います。このノードの分析結果はクラスターの特徴をマップに描写して表示します。

■ モデル化ノード

回帰分析ノードでは線形回帰モデルとロジスティック回帰モデルを利用できます。Target 変数には間隔尺度、順序尺度、二値尺度の変数を使用でき、Input 変数には間隔尺度と名義尺度の変数を用いる事ができます。また変数の選択方法として Stepwise、Forward、Backward をサポートしています。

ツリーノードは名義尺度、順序尺度、2 値尺度、間隔尺度の Target 変数および Input 変数に対して、カイ2 乗検定、エントロピー減少、Gini 減少の分割方法をサポートしています。ノードの設定により、一般的な分割手法である CHAID、CART、C4.5 アルゴリズムへの近似も可能です。このノードでは自動学習と対話型学習とをサポートしており、自動学習の場合、分割への寄与度により Input 変数を順位付けします。この順位付けは、後続のモデル化ノードにおいて、使用する変数の選択に役立ちます。また、後続のモデルノードで使用できるダミー変数の生成も可能です。対話型学習では、分割ルールの定義機能や不必要な枝の剪定機能を利用でき、インタラクティブな探索や評価が可能です。

ニューラルネットワークノードは多層フィードフォワードニューラルネットワークを構築、学習、評価することができます。デフォルトでは 3 つのニューロンで構成された 1 つの隠れ層を持つ、多層フィードフォワードニューラルネットワークを構築します。ニューラルネットワークノードは多種のネットワーク形態をサポートしています。

ユーザ定義モデルノードは、ユーザが独自に定義したモデル（例えば SAS/STAT™ソフトウェアの LOGISTIC プロシジャを使用したロジスティク回帰モデル）を SAS コードノードで構築し、そのモデルの予測値について評価用の統計量を算出します。この評価用データセットは、入力データソースノードを使用してプロセスフローに加えることができます。

アンサンブルノードは複数のモデルから得られた予測値を平均することにより、新しいモデルを作成します。その後、この新しいモデルを使用して、データがスコアリングされます。アンサンブルモデルは個々のモデルに比べて非常に安定しており、モデルによって結果に差がある場合に効果的です。このノードでは以下のモデル操作をサポートしています。

1. 結合モデル：Target 変数の予測値を平均することにより複数のモデルを結合します。
2. 層別モデル：データの各セグメント別に作成されたモデルのスコアリングコードを結合することにより層別モデルを作成します。層ごとに異なったスコアリングコードを 1 つの DATA ステップにまとめることにより、フローの 1 つのパスで層別変数のすべての水準にスコアを与えることができます。
3. バギングモデル/ブースティングモデル：データからサンプリングを繰り返し行い、サンプルごとにモデルを当てはめます。その後予測値を平均することでアンサンブルモデルを作成します。バギングモデルではオブザーベーションの重複を許した無作為抽出を行うのに対して、ブースティングモデルは適応型再サンプリングを行います。

主成分/DM ニューラルノードはバケット変換した主成分を Input 変数として、二値または間隔尺度の変数を予測する加法的な非線形モデルを作成します。また、主成分分析のみを実行し、結果の主成分得点を後続のノードに渡す事も可能です。このノードのアルゴリズムは、分析に用いる入力変数間に高い相関がある場合に生じる、通常のニューラルネットワークの問題点を克服しています。

2 段階モデルノードは分類変数のターゲット変数と間隔尺度のターゲット変数を予測する 2 段階モデルを計算します。クラスモデルと値モデルは、第 1 段階と第 2 段階で分類 Target 変数と間隔尺度の Target 変数にそれぞれ当てはめられます。このノードのスコアコードは、分類モデルと値モデルを組み合わせたものになります。

■ アセスメントノード

アセスメントノードはモデル化ノード（回帰分析、ツリー、ニューラルネットワーク、ユーザ定義モデル）から算出される予測結果を比較する共通の枠組みを提供します。比較はモデルを適用した場合の予想損益、と実際の損益にもとづいて行います。このノードはモデルの精度を示す為にリフトチャート、期待利益/損失チャート、投資利益チャート、ROC チャート、適合度診断チャート、閾値評価チャートを作成します。

スコアノードは学習されたモデルから、新たな入力データに対する予測値を算出、管理します。EM4 はスコアリングの計算式を EM4 のない環境でも使用できるよう、SAS データステップの形式で生成、管理します。

リポーターノードでは、分析プロセスを、ウェブブラウザで閲覧可能な HTML 形式のリポートにします。作成されるリポートはヘッダー情報、プロセスフローダイアグラムのイメージ、フロー内の各ノード処理結果を含みます。レポートはプロジェクトナビゲータのレポートタブで管理されます。

C*スコアノードは、SAS データステップで書かれたスコアリングコードをプログラミング言語 C によるスコアリング関数に変換します。また、スコアリング関数をテキスト形式で出力しますので C もしくは C++の開発環境で使用することができます。C*スコアノードはコンパイルに使用するためのヘッダファイルを生成します。C*スコアの出力は完全なスコアリングシステムではなく、変換された SAS データステップのスコアリングコードで提供される機能だけでは不十分です。

■ ユーティリティノード

グループ処理ノードは、性別のような分類変数ごとの処理を行いたい場合に有用です。このノードは目的変数の水準数が多様な場合の分析において、同様の設定で複数回の処理を繰り返す場合に用いられます。

データマイニングデータベースノードは、バッチ処理時にデータマイニング用データベース（以下 DMDB）を作成する場合に用います。それ以外でノードに DMDB が必要な場合は、そのノードの実行時に自動的に作成されます。このノードを使用して DMDB を作成すると、その DMDB がプロセスフローダイアグラム上に表示され視覚的に確認することができます。DMDB には、数値変数についての要約統計量、およびカテゴリ変数についての要因水準情報を含むメタデータカタログが格納されています。

SAS コードノードでは、プロセスフローダイアグラムにおいて SAS コードを導入することができます。SAS システムがサポートするプロシジャ等の SAS コードを記述できることで、データマイニングの中に独自の処理を取り入れることができます。また、カスタマイズされたスコアリングコードを SAS データステップで作成でき、条件付データ処理、データの連結などが可能となります。学習用、評価用、テスト用、またはスコアリング用のデータセットや、Input、Target、予測値などの変数を機能的に参照する為のマクロを定義、提供しています。SAS コードノードの実行結果データセットは後続のノードで使用可能です。

コントロールポイントノードは、プロセスフローダイアグラムに制御ポイントを設定してノードの接続数を減らすことができます。例えば、3 つの入力データソースノードが 3 つのモデル化ノードに接続されている場合、入力データソースノードとモデル化ノードの間に 9 つの接続を設定する必要がありますが、このノードを利用することで必要とされる接続数は少なくなります。

サブダイアグラムノードは、プロセスフローダイアグラムの一部分をサブダイアグラムとしてグループ化します。複雑なプロセスフローダイアグラムにおいては、このノードの利用によりデザインの向上や管理の簡便化を図ることができます。

Some General Usage Rules for Node

プロセスフローダイアグラムにおけるノードの設定位置に関して以下の仕様があります。

- 入力データソースノードは如何なるノードの後にも接続させることはできません。
- サンプリングノードはデータセットを出力するノードの後に接続させなくてはなりません。
- アセスメントノードは 1 つ以上のモデル化ノードの後続でなければなりません。
- スコアノードはスコアコードを生成するノードの後続でなければなりません。例えば、モデル化ノードはスコアコードを生成します。
- SAS コードノードはプロセスフローダイアグラムのどこにでも配置できます。SAS コードノードには、入力データソースノードで定義される入力データセットは必要ありません。SAS コードノードを記述して SAS データセットを作成した場合、後続のデータセット属性ノードを使用することで SAS データセットの変数に役割を設定する事ができます。

Enterprise Miner™ソフトウェアバージョン 2.02 プロジェクトの読み込み制限

- **Work ライブラリにプロジェクトを定義し作成された EM2 の DMX ファイルはテンポラリプロジェクトですので読み込めません。**
- **SASUSER もしくは SASHELP をデータライブラリ、プロジェクトライブラリと定義していた EM2 の DMX ファイルは読み込む事ができません。**
- **プロジェクトを読み込む際には EM2 のセッションを必ず閉じておく必要があります。**

ツリーノードを使用した C4.5 への近似法

C4.5 に一番近似したツリーノードのオプション設定は次のとおりです。

- **［プログラムエディタ］ウィンドウから次の%LET ステートメントをサブミットします。**

  %let split_1=criterion=ERATIO;

  ツリーノードでは、基準を ERATIO に設定するオプションはありません。
- **［基本設定（Basic）］タブで［ノードからの枝数の最大値（Maximum number of branches from a node）］を入力内の名義尺度の値を最大数に設定します（最大 100）。**
- **［基本設定（Basic）］タブで［各ノードに保存する代理ルール（Surrogate rules saved in each node）］オプションを 0 に設定します。**
- **［基本設定（Basic）］タブで［有意水準（Significance level）］を 0 に設定します。**
- **［詳細設定（Advanced）］タブで［サブツリー（Subtree）］を［最良のアセスメント値（Best assessment value）］に設定します。**
- **ヒューリスティックな検索を強制するために、［詳細設定（Advanced）］タブで［徹底した分岐の検索を行う上限回数（Maximum tries in an exhaustive split search）］を 0 に設定します。**
- **［詳細設定（Advanced）］タブで［分岐の検索に対する最大オブザベーション数（Observations sufficient for split search）］をデータセットのサイズ（最大で 32,000）に設定します。**
- **先行する検証用データセットをツリーノードに読み込みます。検証用データセットは、［Best assessment value］サブツリー法と共に使用します。**

Ⅲ. Start Hands-on!!

ⅰ. Clustering Tools

分析のテーマ

現在のビジネス環境におけるCRMの具体的な手法を考えると、属性データを利用した顧客のセグメンテーションは、現状把握、そしてキャンペーン等のオペレーションに有効な手法です。
今回は、従来の EM2 から使用していたクラスター分析と、更に新しく加わった SOM/Kohonen ノードを使い、EM4 で可能なセグメンテーションの手法を紹介しながら、顧客データのセグメンテーションからセグメント毎の特徴を掴むことを分析のテーマに置きます。
＜SOM＞
SOM（Self-Organizing Maps:自己組織化マップ）は、データの構造を明快に表現する為に有用な手法です。それは、多くの変数で構成される、つまり高次元で表現されるオブザーベーションの特性を低次元（主に平面）で表現し、オブザーベーション間の類似度の把握を容易にする次元縮約手法です。

使用するデータ

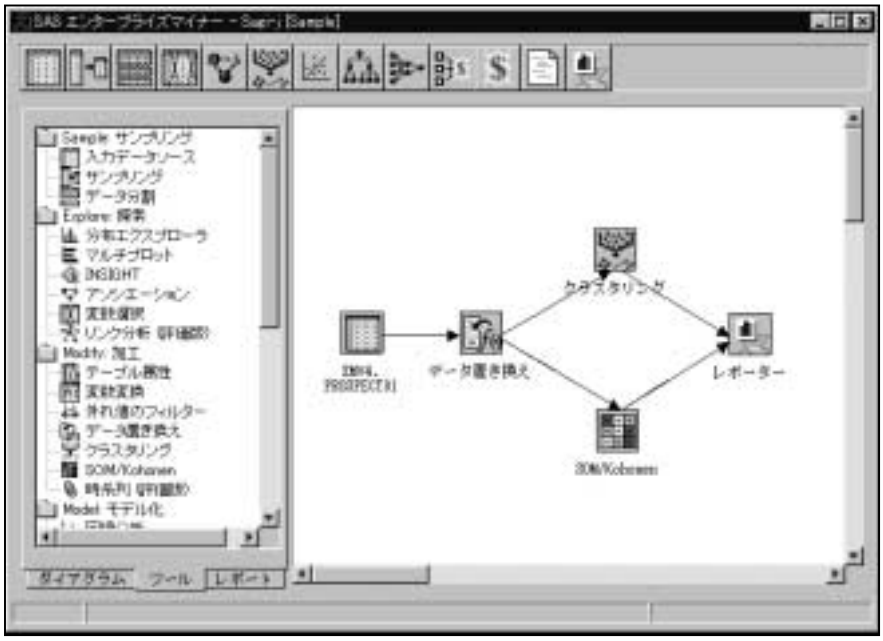
カタログ販売の会社が、ある製品に対して潜在的な購買力を保持する顧客を見つけ出したいと考えています。経験的に、顧客の選択する製品が、地理的条件や地域の人口動勢といった要因に起因するところが大きい事を把握しています。
そこで、見込み顧客を属性データからセグメンテーションすることにします。

変数名	役割	測定水準	ラベル
ID	Id	Nominal	ID 番号
AGE	Input	Interval	年齢
INCOME	Input	Interval	収入
SEXJ	Input	Binary	性別
MARRIED	Input	Binary	未婚
FICO	Input	Interval	クレジットスコア
OWNHOME	Input	Binary	持家の有無
LOC	Rejected	Nominal	居住地域、A-H の値
CLIMATE	Input	Nominal	居住地域の風土

データセット　PROSPECT01　(obs=5055)

分析フローの構築

今回の分析の目的は見込み顧客のセグメンテーションにあります。そこで今回は、クラスタリングノードとSOM/Kohonen ノードを使用してセグメント化のフローを作成し、それぞれのセグメンテーションの特徴を見ていきます。



クラスタリングと SOM/Kohonen を使用した分析フロー

■ 入力データソースノード

最初に**[入力データソース]**ノードでライブラリ参照名**「EMV4」**のデータセット名**「PROSPECT01」**を選択し読込みます。

[変数]タブをクリックし分析で使用する変数の役割を設定します。今回は居住地域を表す分類変数**「LOC」**を**「Rejected」**に設定します。
設定の後、入力データソースノードを保存して閉じます。

■　データ置き換えノード

次に**[データ置き換え]**ノードを入力データソースノードに接続します。
ここでは、欠損値の置き換えを目的としたデータ加工を行います。
オープンして**[標準設定]**タブの**[補充方法]**サブタブをクリックします。
間隔変数における欠損値の補充方法が**「mean」**、分類変数が**「most frequent value（count）」**が選択されていることを確認します。
以上の確認が終了後データ置き換えノードを保存し終了します。

■　クラスタリングノード

データ置き換えノードに**[クラスタリング]**ノードを接続します。
クラスタリングノードをオープンすると**[変数]**タブが表示されます。
変数タブにおいて**[標準化]**の方法で**「標準偏差」**にチェックを入れます。
以上の設定が終了後、クラスタリングノードを保存して閉じます。

■　SOM/Kohonen ノード

次にクラスタリングノードと並列になるようにデータ置き換えノードに**[SOM/Kohonen]**ノードを接続します。
SOM/Kohonen ノードをオープンし以下の設定をします。

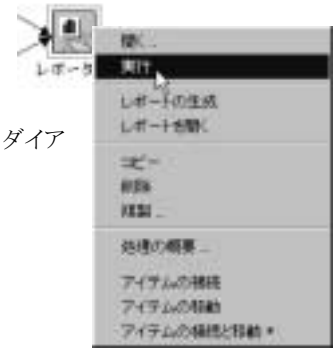
1. **[変数]**タブにおいて標準化の方法で**「標準偏差」**を選択します。

2. **[一般]**タブをクリックし**[マップの行と列]**数をそれぞれ設定します。
   行:2
   列:3
   この設定された行列サイズに応じたトポロジカルマップが作成されます。

3. 最後に**[シード]**タブにおいて**[初期の選択手法]**を選択します。下矢印のコントロールボタンを押しプルダウンメニューから今回は**「MacQueen」**を選択します。

以上の設定終了後、SOM/Kohonen ノードの設定を保存し、閉じます。

■ レポーターノード

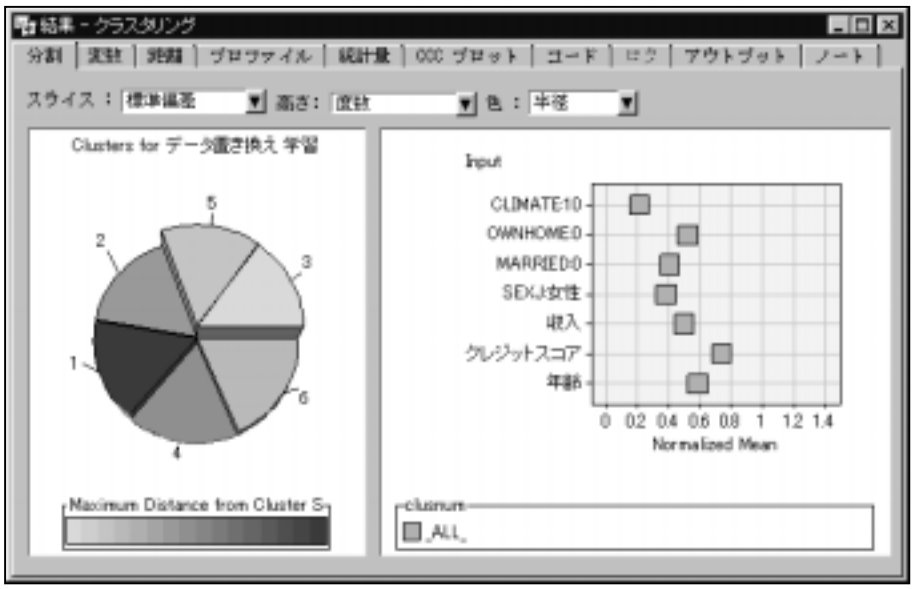


クラスタリングノード、SOM/Kohonen ノードに**[レポーター]**ノードに接続します。その後、ノードを右クリックしてプルダウンメニュー→**「実行」**を選択し、ダイアグラムワークスペースに作成した分析フローを実行します。

作成されたクラスタの特徴を見てみよう。

■ クラスタリングノードの結果

クラスタリングノードを右クリックしてプルダウンメニューから**「結果」**を選択すると、下図**[結果]**ウィンドウの**[分割]**タブが表示されます。



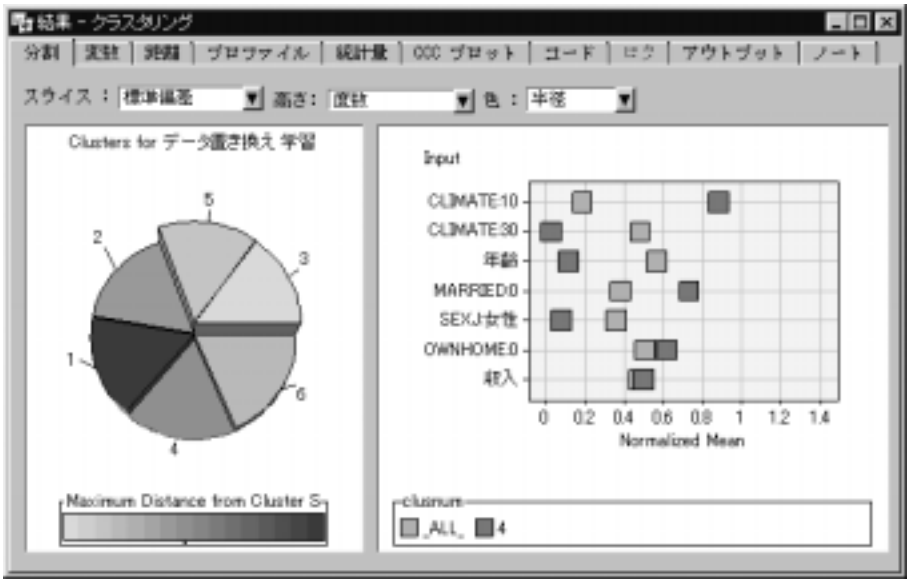
まず左側の円グラフから 6 つのセグメントが作成されたことがわかります。
この 6 つのセグメント毎の特徴を見るために、右図の**「入力平均グラフ」**を参照し、各セグメントの情報と学習用データ全体の平均値を比較します。

1つずつセグメントを選択して、学習用データの平均値と比較します。

ツールバーの  **[ポイントの選択]**ボタンをクリックしてから、円グラフにおける**「特定のセグメントをクリック**（ここでは**セグメント番号 4** のクラスター）**」**して選択します。

次に、ツールバーの  **[入力変数のプロット]**ボタンをクリックします。

自動的に入力平均グラフ上に学習用データの全平均に対するクラスター4 の相対平均がプロットされます。

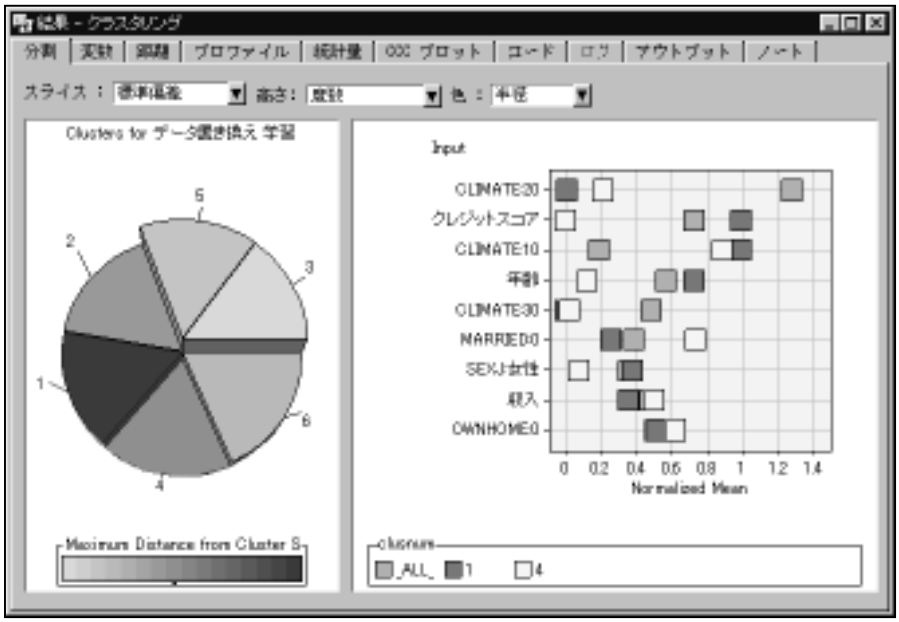


この入力平均グラフからクラスター4 に属する顧客の特徴を全体平均と比較して説明する事ができます。
例えば、年齢に関しては全体の平均より若いですが、収入に関しては全体の平均とほぼ同水準といえます。
クラスター分析においては、分類変数はダミー変数として扱われます。
例えば、居住地域の風土（CLIMATE）について、クラスター4 に属する顧客は居住地域の風土が 10 である割合が高く、居住地域の風土が 30 に属する顧客はほとんど存在しないことがわかります。

同様に他のクラスターと全体の平均を比較してみましょう。

クラスターを複数選択して全体平均と比較したい場合は、ツールバーの  **[ポイント選択]**ボタンをクリックし、**[Ctrl キー]**を押しながら選択したいセグメントをクリックしていきます。その後、ツールバーの  **[入力変数のプロット]**ボタンをクリックします。

それでは、1 と 4 の 2 つのクラスターと全体平均のプロット図を表示させてみましょう。



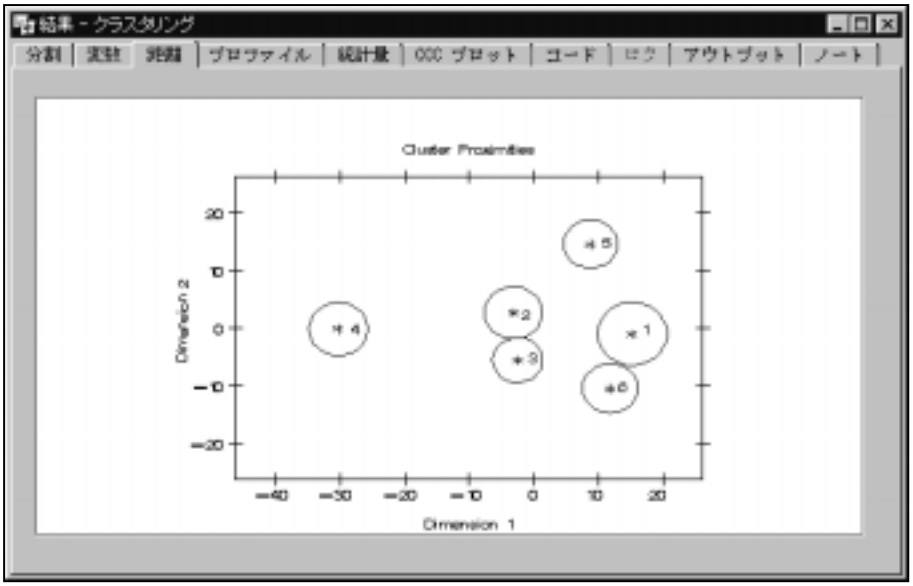
入力平均グラフを見てクラスター1 と 4 の共通点、異なる点から特徴を比較しましょう。

・共通点としては、どちらのクラスターも居住地域の風土が 30 に該当する顧客は非常に少なく、収入が全体平均に近いことがわかります。

・異なる点としては、クレジットスコアに関してクラスター4 は非常に低い値であるのに対し、クラスター1 は高い値をとっています。また、年齢についてもクラスター4 では、全体平均やクラスター1 に比べると非常に若いことがわかります。

このように各クラスターには特徴があることから、その位置関係にも特徴があることが考えられます。

[距離]タブをクリックしてみましょう。

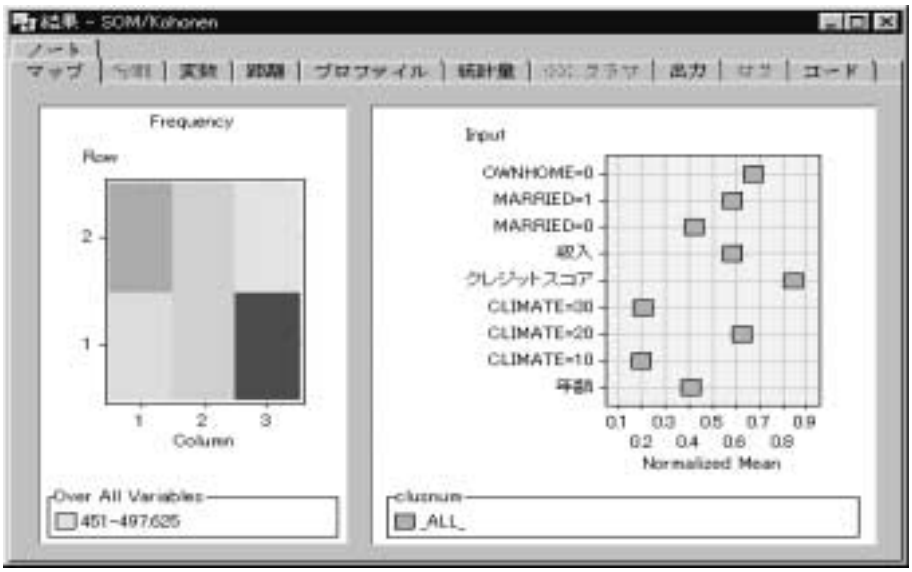


距離プロットで明示されるように各クラスターの中心位置がクラスター毎の特徴（居住地域の風土、クレジットスコア、年収等）となり入力平均グラフ上で表示されます。

このようなクラスター間の位置関係を考慮しながら各セグメントの特徴を分析したい場合に SOM/Kohonen ノードの自己組織化マップによるクラスター説明を使用します。

■ SOM/Kohonen ノードの結果

SOM/Kohonen ノードを右クリックして、プルダウンメニューから**「結果」**を選択すると**[結果]**ウィンドウの**[マップ]**タブが表示されます。



クラスタリングノードの結果によく似た SOM/Kohonen ノードの結果ウィンドウですが、最大の違いは左図です。ここではトポロジカルマップに配置されて各クラスターが表示されます。
今回の分析ではマップを 2 行、3 列に設定しています。

右図は学習用データにおける全データの平均をプロットしたグラフで、クラスタリングノード同様、クラスター毎の正規化平均をプロットすることができます。

またマップは各クラスターに属するレコード数で色分けされており、各クラスターの位置はマップにおける位置関係で把握することができます。対角線にあるクラスター1 とクラスター6 は全く異なる特徴をもったクラスターであり、逆に隣接したクラスターは距離が近く、よく似た特徴をもっています。

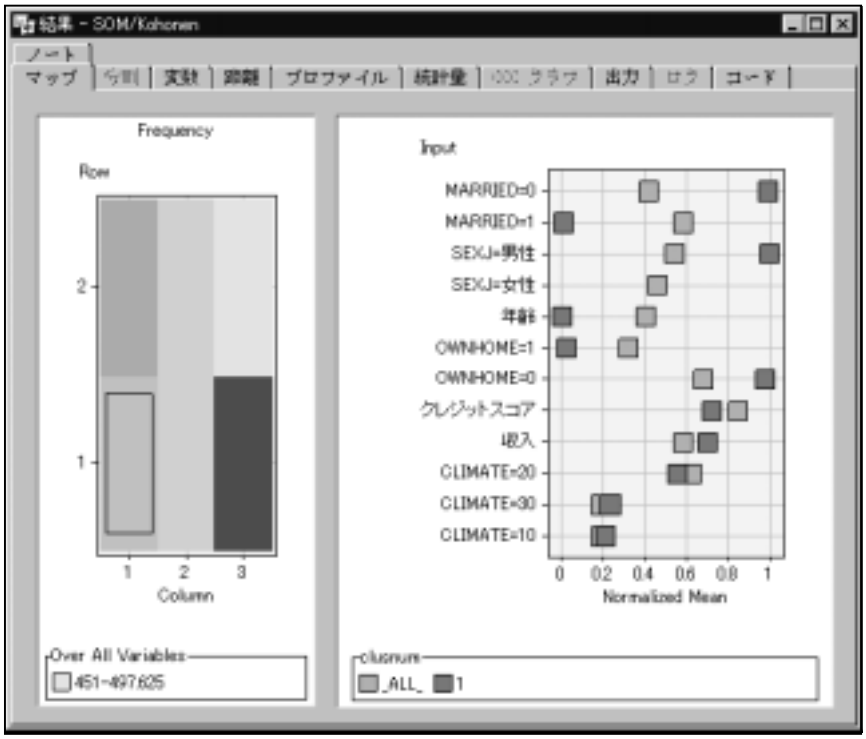
各クラスター平均と全体の平均を比較する場合は、以下の操作により入力平均グラフを用いてプロットします。

ツールバーの  **[ポイントの選択]**ボタンをクリックしてから、マップ上の**「特定のセグメントをクリック」**して選択します。

次にツールバーの  **[入力変数のプロット]**ボタンをクリックします。

自動的に入力平均グラフ上に選択クラスターの正規化平均がプロットされます。

それでは、クラスター1（行：1、列：1）の特徴を全体平均と比較してみましょう。

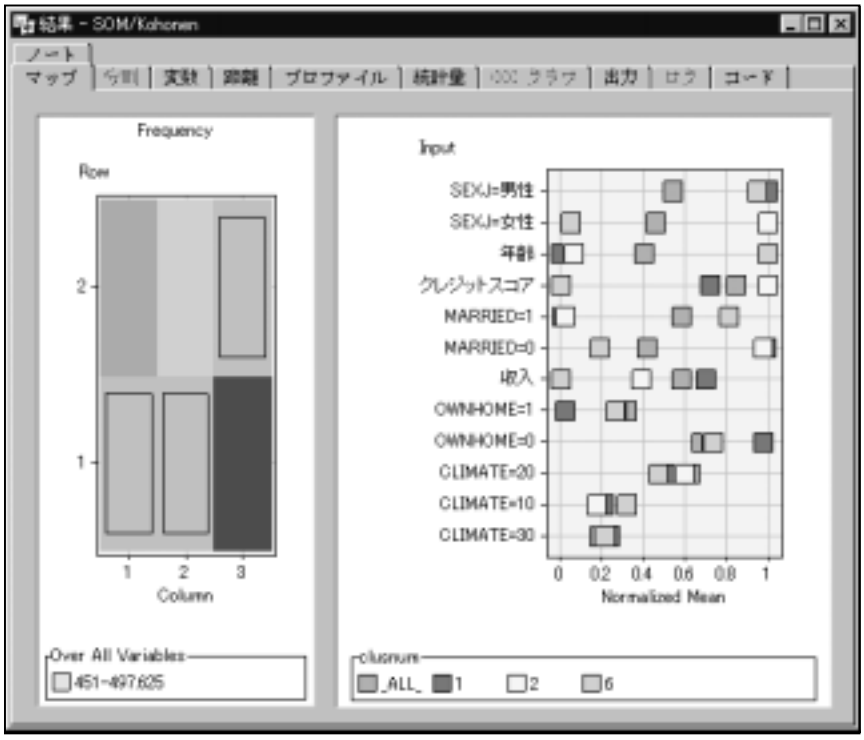


クラスター1 に関して全体平均との比較からわかる特徴は、変数 MARRIED からクラスター1 に属する顧客は未婚者であることがわかり、更に変数 SEXJ から男性であることがわかります。年齢は全体平均に比べ若い集団であり、変数 OWNHOME から持家でないと言えます。クレジットスコアは全体平均に比べればやや低いのですが、収入では全体平均を上回っている顧客のセグメントであると言えます。

それでは、クラスタリングと同様に複数のクラスターと全体平均を比較してみましょう。

クラスター1（行:1、列:1）、クラスター2（行:1、列:2）、クラスター6（行:2、列:3）を選択し、クラスター1 とクラスター1 に近い位置にあるクラスター2。クラスター1 と遠い位置にあるクラスター6。そして全体平均を入力平均グラフに表示してみましょう。

クラスターを複数選択してそれぞれと全体平均を比較したい場合は、ツールバーの  **[ポイント選択]**ボタンをクリックし、**[Ctrl キー]**を押しながらマップ上で選択したいセグメントをクリックします。その後、ツールバーの **[入力変数のプロット]**ボタンをクリックします。



プロット表示されたそれぞれのクラスターの特徴をみると、左図のマップにおいて位置的に近いクラスターはよく似た特徴をもっていることがわかります。
次の特徴についてそれぞれのクラスターがどうであるか、確認してみましょう。

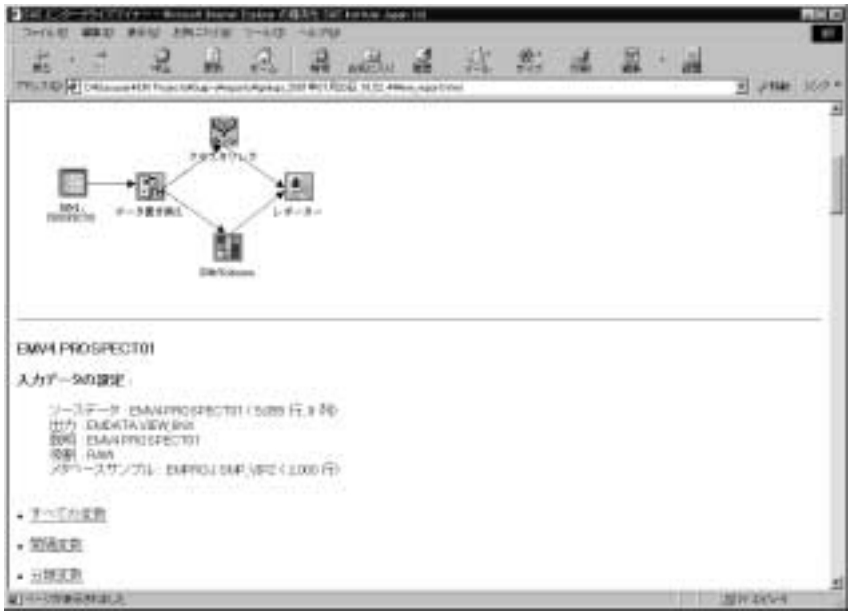
それぞれのクラスターに属する顧客の
性別は？
年齢は？
クレジットスコアは？
収入は？
居住地域の風土（CLIMATE）は？

以上、クラスタリングおよび自己組織化マップを利用したSOM/Kohonenノードにより、このカタログ販売会社の顧客をセグメンテーションすることができました。
経験的に、地理的条件および人口動勢などの要因と潜在的な購買力との関係があると考えていたこの会社において、地理的条件および人口動勢的な観点から顧客をセグメンテーションできたことは、今後様々なオペレーションにおいて対象チャネルの明確化につながります。

■　レポーターノードを利用した結果表示

ここまでの分析結果は、レポーターノードで作成した HTML 形式のレポートでも見ることができます。

[レポーター]ノードを右クリックして、プルダウンメニューから**「レポートを開く」**を選択すると作成された HTML 形式のレポートが Web ブラウザを起動し、表示されます。



レポーターノードにより作成された HTML ファイルレポート

ⅱ. 主成分/DM ニューラル　& アンサンブル

分析のテーマ

現在のマーケティングにおいても、在庫管理においても、将来を予測することが、事業計画に大きな判断材料となっております。
更に、問題の焦点をその予測の精度にまで広げたいと思います。
正確に予測できることで例えば、潜在的購買力のある顧客を見つけだし、より ROI の高い事業を展開する際の意思決定を支援します。
今回は、マーケティングにおける顧客のプロファイリングを行い、DM の発送コストを抑えながらヒット率を上げることを前提にした分析を行います。

使用するデータ

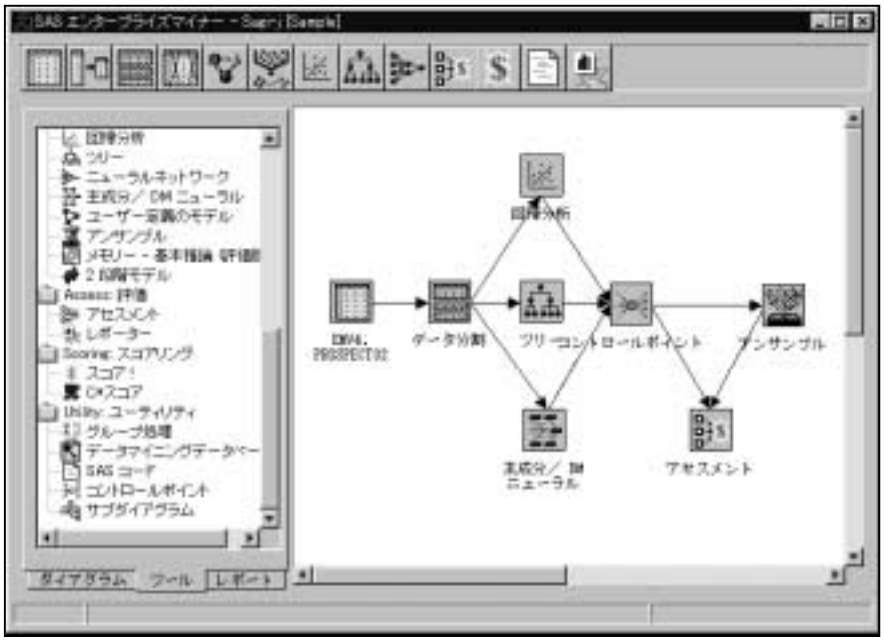
データセットPROSPECT02を使用します。これは、39，779人分の顧客がダイレクトメールに応答したかどうかを調べたターゲット・マーケティングの例です。全顧客のうちの14％から応答がありました。顧客ごとに、応答の有無を表す変数 RESPOND と以下の12個の入力変数があります。

変数名	役割	測定水準	ラベル
RESPOND	**Target**	**Binary**	**反応**
AGE	**Input**	**Interval**	**年齢**
INCOME	**Input**	**Interval**	**収入**
SEX	**Input**	**Binary**	**性別**
MARRIED	**Input**	**Binary**	**未婚**
FICO	**Input**	**Interval**	**クレジットスコア**
OWNHOME	**Input**	**Binary**	**持家の有無**
LOC	**Rejected**	**Nominal**	**居住地域、A-H の値**
BUY6	**Input**	**Interval**	**過去6ヶ月の購入個数**
BUY12	**Input**	**Interval**	**過去12ヶ月の購入個数**
BUY18	**Input**	**Interval**	**過去18ヶ月の購入個数**
VALUE24	**Input**	**Interval**	**過去24ヶ月の総購入個数**
COA6	**Input**	**Binary**	**過去6ヶ月内の転居の有無**

データセット　PROSPECT02　(obs=39779)

分析フローの構築

分析の目的は、入力変数からターゲット（顧客の応答：RESPOND）を予測するモデルを構築することです。
ここでは、回帰分析、決定木、主成分/DM ニューラルを用いた予測モデルと、これら 3 つのモデルを結合したアンサンブルモデルの構築を紹介します。その後、構築したモデルの評価も行います。



主成分/DM ニューラルとアンサンブルを使用した分析フロー

■　入力データソースノード

[入力データソース]ノードで、「EMV4」ライブラリのデータセット「PROSPECT02」を指定して読み込みます。

[変数]タブを選択し、変数「RESPOND」の役割を「Target」に設定します。
また、各 Input 変数の測定水準が正しく設定されているかどうか確認してください。変数「BUY6」、「BUY12」、「BUY18」の測定水準が Ordinal になっている場合は、「Interval」に変更します。

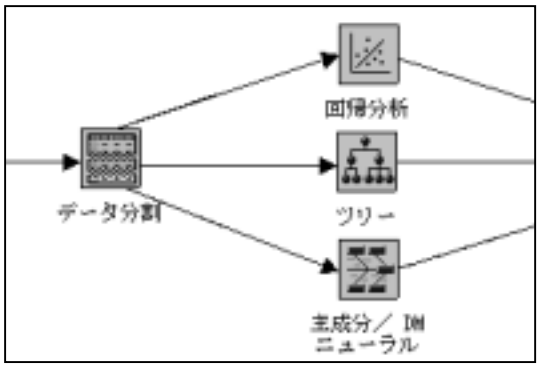
■　データ分割ノード

入力データソースノードに[データ分割]ノードを接続します。

[分割]タブを選択します。
「手法:」ラジオボックスで「単純な無作為抽出法」を選択し、学習データ「70%」、検証データ「30%」の割合になるよう設定します。

■　モデル化ノード

[回帰分析]、**[ツリー]**、**[主成分/DM ニューラル]**ノードをそれぞれデータ分割ノードに接続します。



＜主成分/DM ニューラルノード＞
主成分/DM ニューラルノードは EM4 から追加された新しいノードです。まず主成分分析を実行し、Target 変数と関連の高い主成分スコアを入力変数としたニューラルネットワークを実行します。
このノードの利点は以下のとおりです。

● 分析に用いる入力変数間に高い相関があり、ヘッセ行列が特異もしくは特異の状態に近くなると、目的関数の形状が長く平らでなだらかな谷形の状態になり、非線形最適化の計算における収束が緩やかになります。このような状態では、最適化のための反復計算をいつ終わらせたら適切な解となるのかの判断が難しくなります。このノードでは元のデータから主成分分析を行い、その主成分得点を新たな入力変数として使用することにより、この問題を解決することができます。

● 通常のニューラルネットワークで利用される非線形最適化では、1回の反復においてオブザベーション 1 つ 1 つに対する計算が必要です。そのため、オブザベーション数が多くなると計算量が多くなり計算時間がかかります。このノードでは、主成分得点から少数のグリッドを抽出し、そのグリッドだけをデータとしてニューラルネットワークを実行しますので、計算時間を短縮できます。

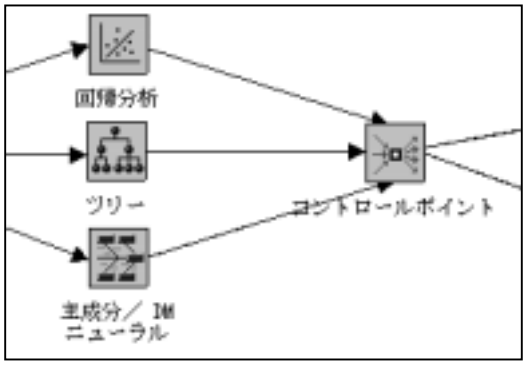
● 従来のニューラルネットワークでは、反復計算の開始値によって収束した後の解が大きく左右されました。このノードで学習される DM ニューラルネットワークは、推定するパラメータ数が少ないので、初期値のグリッド検索を行うことが可能です。

今回は、デフォルトの設定でモデル化ノードを実行します。

■　コントロールポイントノード

このノードはプロセスフローダイアグラムにコントロールポイントを設置することを目的としたノードです。このノードを使用することで、接続の数を減らすことができます。

ダイアグラムワークスペースに**[コントロールポイント]**ノードをドラックアンドドロップで配置し、**[回帰分析]**、**[ツリー]**、**[主成分/DM ニューラル]**ノードからコントロールポイントノードに接続します。



■　アンサンブルノード

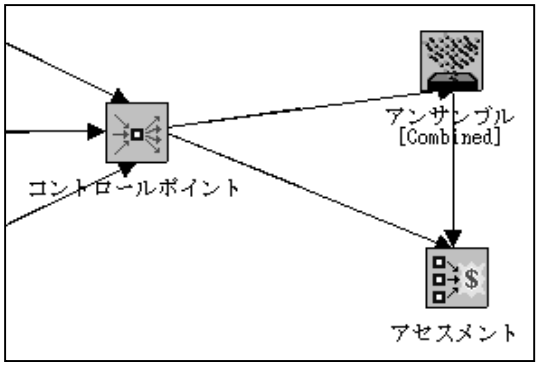
[コントロールポイント]ノードに**[アンサンブル]**ノードを接続してください。このノードでは、3つの予測モデルによるスコアを平均することで、結合モデルによるスコアリングを行います。

ノードを開き、**[設定]**タブを選択します。
[アンサンブルモード]ラジオボタンで**「組み合わせ」**を選択すると結合モデルを作成します。この場合の結合確率関数は**平均**です。

■　アセスメントノード

回帰分析、ツリー、主成分/DM ニューラル、アンサンブルのそれぞれのノードに**[アセスメント]**ノードを接続してください。アセスメントノードでは上記4つのモデルを同時に評価することができます。

フローの構築が完了したら、**[アセスメント]**ノードを右クリックしてプルダウンメニューより**「実行」**を選択します。



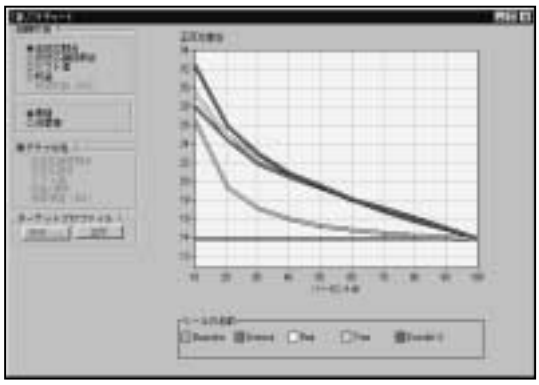
モデルの評価

アセスメントノードの実行後、結果を表示させます。

実行した4つのモデルを同時に選択し、メニューバーより**「ツール」**→**「リフトチャート」**を選択します。

各モデルの正反応割合グラフが表示されます。

データの 10%点を見ると、アンサンブルモデルの反応割合が最も高くなっています。この場合、個々のモデルと比較すると、複数のモデルを結合することでモデルの精度が上がっていることがわかります。



このモデルの結果が示していることは、全顧客の中から単純無作為に10%の顧客を抽出し、DMをうった場合結果は 14%のヒット率しかないでしょう。しかし、モデルを適応させて全顧客から DM に反応してくれそうな顧客上位 10%を抽出した場合、ヒット率は約 30%になるでしょう、ということを示しています。
コストが同じであれば、見込まれる反応が 2 倍あると算出される顧客をターゲットにするほうがベターです。

また、この分析においては単体のモデルによる予測に比べてアンサンブルを使用した結合モデルの精度が一般的に良い結果が出ます。

Enterprise Miner™ SoftwareVersion4.1 ハンズオンワークショップ

株式会社 SAS インスティチュートジャパン
〒104　東京都中央区勝どき 1-13-1 イヌイビルカチドキ 8F
TEL　03(3533)3890
〒530　大阪市北区堂島浜 1-4-16　アクア堂島西館 12F
TEL　06(6345)5700

The Power to Know.