本書では、IBM SPSS Modeler 18.2Jを使用しています。

3 オートメーション（自動数値）

IBM SPSS Modelerでは、モデル作成のための**オートメーション**ノードが含まれています。オートメーションでは、目的や用途に応じた複数のモデリング手法を同時に実行させて、精度の良いものから順番に表示して評価することができます。オートメーションの予測モデルとして、カテゴリ型フィールドを対象とする**自動分類**ノードと、連続型フィールドを対象とする**自動数値**ノードがあります。

Contents

Keyword

オートメーション / 連続型 / 自動数値 / 相関 / 散布図 / ヒストグラム / 予測値 / アンサンブル / C&R Tree / 線型回帰 / 一般化線型 / CHAID /

§3.1.1　自動数値モデルの概要

予測のためのモデリングでは、**カテゴリ型**フィールドを対象とする場合と、**連続型**フィールドを対象とする場合があります。連続型フィールドを対象とするモデリング手法は、個別にそれぞれのノードを用いるか、オートメーションの**自動数値**モデルで実行することができます。自動数値モデルによって指定できるのは、以下の手法です。

	**線型回帰**	線型回帰分析を実行するノードです。対象フィールドは連続型です。
	**1次**	線型モデルを作成するためのノードです。
	**一般化線型**	リンク関数(接続関数)を使用することで、さまざまな分布のデータを扱う一般化線型モデルを作成します。
	**Random Trees**	入力フィールドのランダムサンプリングを行って多数のツリーを構築し、予測精度を向上させます。
	**KNN**	類似性に基づいてレコードを分類して予測を行うNearest Neighbor Modelsを作成します。
	**CHAID**	統計的仮説検定に基づくディシジョンツリーの手法です。多分岐のツリーが構成されます。
	**C&R Tree**	不純度に基づくディシジョンツリーの手法です。常に2分岐のツリーが構成されます。
	**ニューラルネット**	入力層と出力層の間に隠れ層を持つ多層型ニューラルネットワークモデルを構築するノードです。

Table3.1.1　連続型フィールドを対象とする主要なモデリングノード

POINT

自動数値ノードを使用すると、連続型フィールドを対象とするモデリング手法を、さまざまなパラメーター設定で複数実行して、結果の精度の良いモデルを採用することができます。

§3.1.2　ストリームの確認

オートメーションの自動数値によるモデリングの例として、**Chapter3.str**をIBM SPSS Modelerに読み込みます。このストリームには、データソースとしてカンマ区切りの**analysis1.csv**がセットされています。顧客の属性や利用サービスに関するデータが記録されています。フィールド(変数)は16個、8,632行のデータが含まれています。

ここでは、顧客の**支払金額**を予測するモデル作成の例で、実行手順を確認します。

操作手順

1. **ファイル**メニュー ＞ **ストリームを開く**を選択します。
2. **/ ModelerC / Chpter3.str**を選択します。

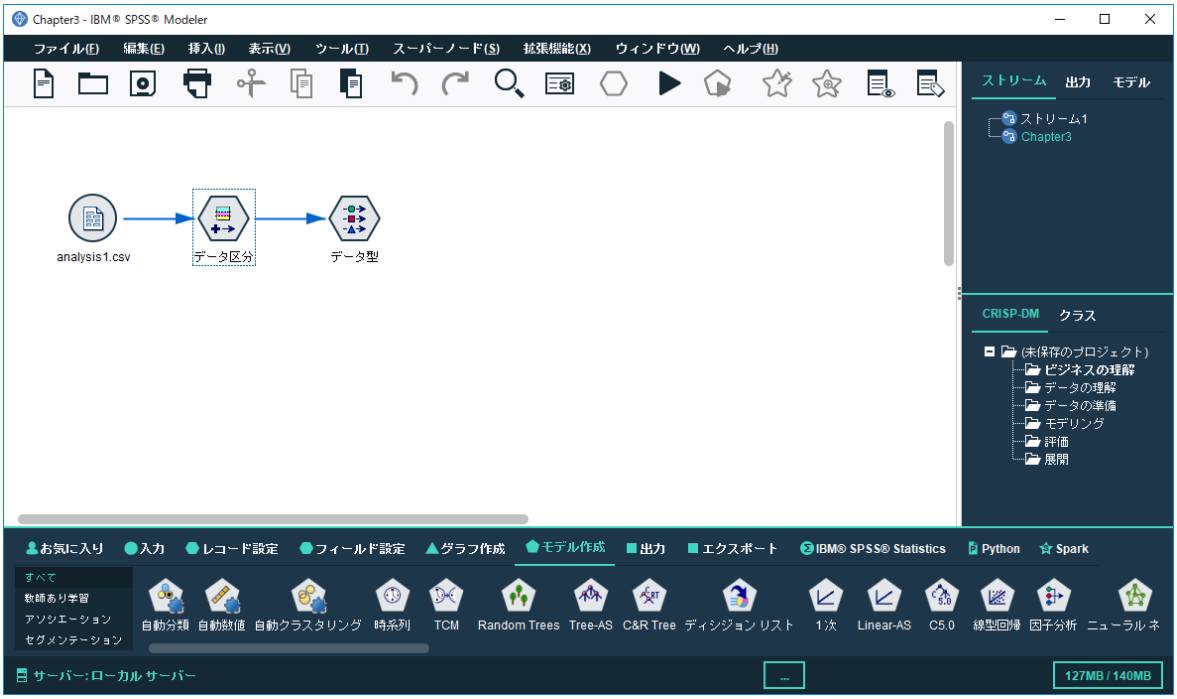

Figure3.1.1　Chapter3.strファイルのストリーム

操作手順

3. ストリームの**データ区分**ノードをダブルクリックして編集画面を表示します。

Figure3.1.2　データ区分ノード

データ区分ノードは、レコードをランダムに**学習**データ区分と**テスト**データ区分に分割する場合に使用します。デフォルトでは、学習データ区分に**50%**、テストデータ区分に**50%**確保されます。また、ランダムシードの設定が有効にすることで、乱数の値が固定されるため、ストリームを実行するごとに異なるレコードが割り当てられることを防ぐことができます。

TIPS

特にレコード数が少ない場合、学習データにはテストデータより多めのサンプルを割り当てます。学習用に70%、テスト用に30%の割合は比較的よく利用されます。

操作手順

4. **OK**ボタンをクリックして、データ区分ノードの編集画面を閉じます。
5. ストリームの**データ型**ノードをダブルクリックして編集画面を表示します。

Figure3.1.3　データ型ノード

データ型ノードで、フィールドの**尺度**と**ロール**を設定します。この例では、**支払金額**を予測の目的フィールドとして**対象**に割り当てています。

顧客番号と**支払方法**、**新規契約**の3つはモデリングには使用しません。

また、**データ区分**のフィールドによって、学習データ区分でモデル作成、テストデータ区分でモデルのテストが自動的に行えるようになっています。

POINT

予測に寄与しないフィールドや、使用することに意味のないフィールドなどはそのロールを**なし**に設定しておきます。

TIPS

入力フィールドや対象フィールドの指定は、各モデリングノードでも設定することが可能ですが、データ型ノードのロールとして設定しておくと、モデリング手法を変える度にフィールドの指定を行う必要がなくなるため効率的です。

操作手順

6. **OK**ボタンをクリックして、**データ型**ノードの編集画面を閉じます。

§3.2.1　オートメーションによる自動数値モデル作成の実行

この例では、顧客の**支払金額**の予測を行う例として、オートメーションの**自動数値**ノードを利用したモデル作成の手順を確認します。

操作手順

1. **モデル作成**パレットの**すべて**サブパレットをクリックします。

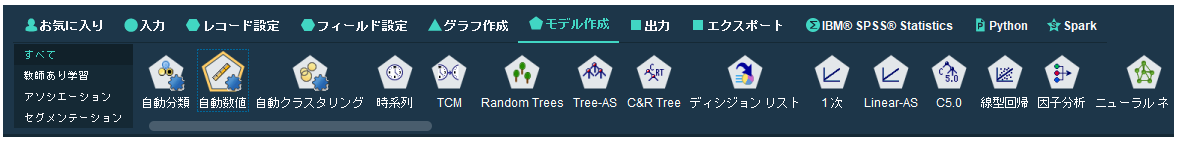

Figure3.2.1　モデル作成パレットのすべてサブパレットの自動数値ノード

操作手順

2. **自動数値**ノードをストリームキャンバスに挿入し、**データ型**ノードからリンクします。

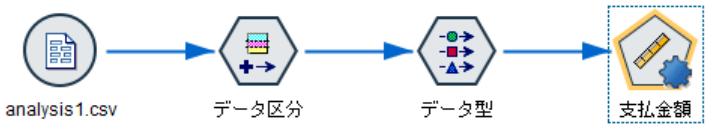

Figure3.2.2　ストリームキャンバスに配置した自動数値ノード

ストリームキャンバスに挿入したノードは、自動的に対象フィールドの名称になります。この例では、**支払金額**になっています。

入力フィールドと対象フィールドの設定は、データ型ノードで完了しているため、このままでもモデリングを実行することができますが、実行前に重要なパラメーターを確認しておきます。

操作手順

3. **自動数値**ノードをダブルクリックして編集画面を開きます。
4. **モデル**タブを開きます。

Figure3.2.3　自動数値ノードのモデルタブ

モデルタブでは、実行したモデルの選択の基準を設定します。**データ区分データを使用**が有効になっており、上流のデータ区分ノードの結果を受けて、学習用データとテストデータに区分した処理が行われます。

モデルのランク付けは、複数のモデルの精度を評価する基準の選択です。デフォルトでは、**相関**(実測値と予測値の相関)が選択されています。その他には、使用フィールド数、相対誤差が選択可能です。**使用モデル数**は、上位いくつまでのモデルを採用するかの設定であり、デフォルトでは**3**個です。

この例では、デフォルトのまま進めます。

操作手順

5. **エキスパート**タブを開きます。
6. **モデルの選択**ドロップダウンリストから**Modelerで実行（目的を個別に指定）**を選択します。

Figure3.2.4　自動数値ノードのエキスパートタブ

エキスパートタブでは、実行するモデリング手法の選択と各手法のパラメーター設定を行うことができます。デフォルトでは、**線型回帰**、**一般化線型**、**C&R Tree**、**CHAID**、**線型**、**ニューラルネット**が選択されており、それぞれ1つのモデルがデフォルトのパラメーター設定によって作成されます。数パターンのパラメーターを設定することによって、同じモデリング手法でも複数のモデル作成を行えるようになっています。

ここでは、**線型回帰**を例にパラメーター設定の変更手順を確認してみます。

操作手順

7. **線型回帰**の**モデルパラメータ**のセルをクリックします。
8. **指定**を選択します。

Figure3.2.5　線型回帰のモデルパラメータの指定

POINT

モデルタイプは、**モデルパラメータ**を個別に指定することができ、さまざまな設定によるモデル作成を同時に試行することができます。

5 クラスタリング

クラスタリングは、入力フィールドの類似したパターンを持つレコードをクラスター(セグメント)に分割するための分析手法です。入力フィールドのみに関心があるため、クラシフィケーションの手法と異なり対象フィールドの概念を持ちません。クラスタリングによって、例えば、既存の顧客のレコードを顧客セグメントに分類することができます。

Contents

Keyword

クラスタリング / クラスター分析 / K-Means / Two-Step / Kohonen / 異常値検査 /
自動クラスタリング / クラスターのプロフィール /

§5.1.1　クラスタリングの概要

クラスタリングは、入力フィールドの類似したパターンを持つレコードをクラスター(セグメント)に分割するための分析手法です。**入力**フィールドのみに関心があるため、クラシフィケーションの手法と異なり**対象**フィールドの概念を持ちません。クラスタリングによって、既存の顧客のレコードを潜在的な顧客セグメントに分類することができます。

	**自動クラスタリング**	複数のクラスタリングモデリングを同時に試行して、精度の高い結果を採用する場合に使用します。
	**K-Means**	K個のクラスター数を指定することで実行される代表的なクラスタリングのモデリングノードです。
	**Two-Step**	Two-Stepでは、2段階のクラスター化手法が用いられ、最適なクラスター数を自動的に決定されます。
	**Kohonen**	ニューラルネットワークの手法で、ネットワークの学習によってレコードを類似セグメントに分類します。
	**異常値検査**	Two-Stepの方法によりセグメント化を行い、正常なパターンに合致しないレコードを特定します。

Table5.1.1　クラスタリング(セグメンテーション)のノード

IBM SPSS Modelerでは、クラスタリングのためのノードとして、**自動クラスタリング**、**K-Means**、**Two-Step**、**Kohonen**、**異常値検査**の5つがあり、**モデル作成**パレットに含まれています。

POINT

クラスタリングは、多数の入力ノードに基づいて類似グループ(クラスター/セグメント)を識別することができる分析手法です。

§5.3.2　K-Meansによるクラスタリングの結果の解釈

クラスタリングの実行が完了すると、ストリームにモデルナゲットが作成されます。このモデルに、クラスタリングの結果が含まれています。

操作手順

1. **K-Means**のモデルナゲットをダブルクリックして編集します。

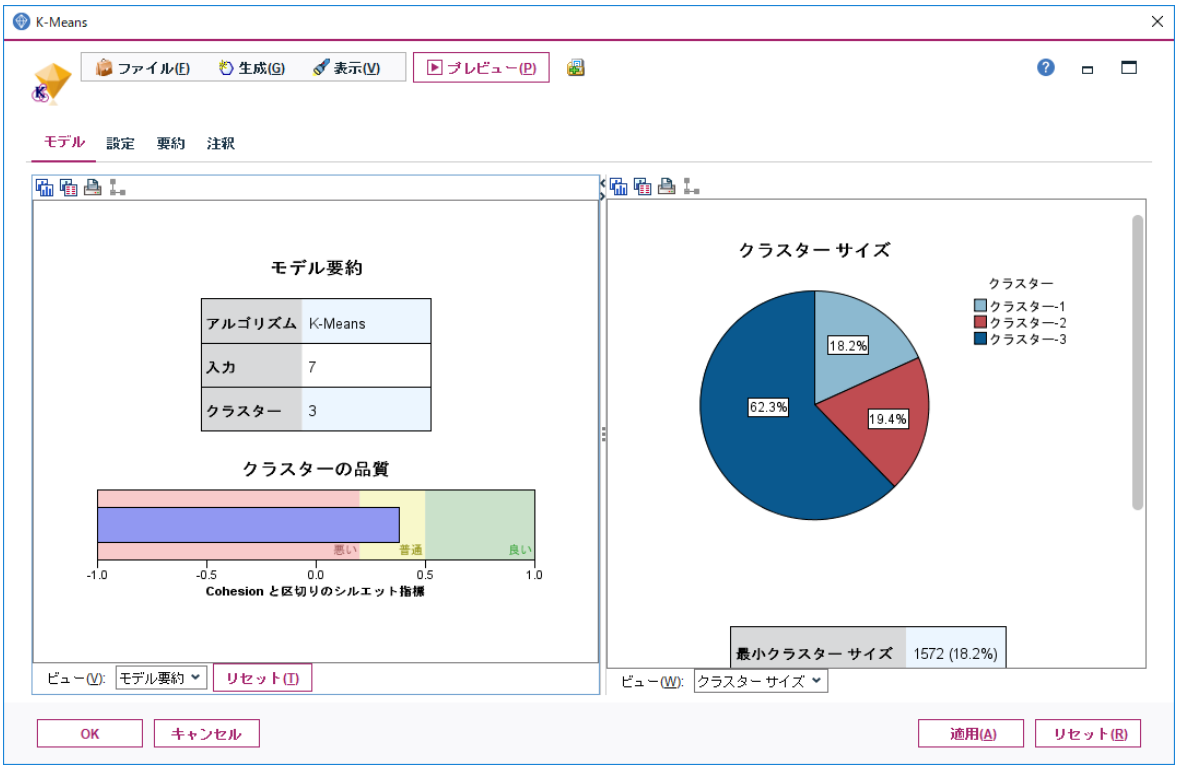

Figure5.3.7　生成されたK-Meansモデル

K-Meansモデルビューアーが表示されます。左側のメインビュー領域にモデル要約が表示されます。この例では、**7**個のフィールドに基づいて**3**個のクラスターに分類されています。また、クラスターの品質はクラスタリングの精度の評価の参考となります。

クラスターの品質は、レコードとクラスター中心の距離を計算しており、値が1に近いほど所属するクラスターと最近隣クラスターとの差があることを意味し、精度が高いと評価します。このスナップショットを使用して、クラスターの精度が悪いかどうかをすばやく確認できます。

TIPS

すべてのレコードに対するシルエット平均は、(B-A) / max(A,B) となります。Aはクラスター中心へのレコードの距離、Bはレコードが属さない最近隣クラスター中心へのレコードの距離です。シルエット係数**1**は、すべてのケースはクラスター中心に直接配置されているということを意味します。値**-1**は、すべてのケースが他のクラスターのクラスター中心にあることを意味します。平均の**0**の値は、ケースが自身のクラスター中心と、その他の最近隣クラスターとの間で等距離にあることを意味します。

TIPS

悪い結果、普通の結果、よい結果は、クラスター構造の解釈に関するKaufmanとRousseeuw（1990）の研究に基づきます。

操作手順

2. 左側のメインビュー領域の下部にある**ビュー**を**クラスター**に変更します。

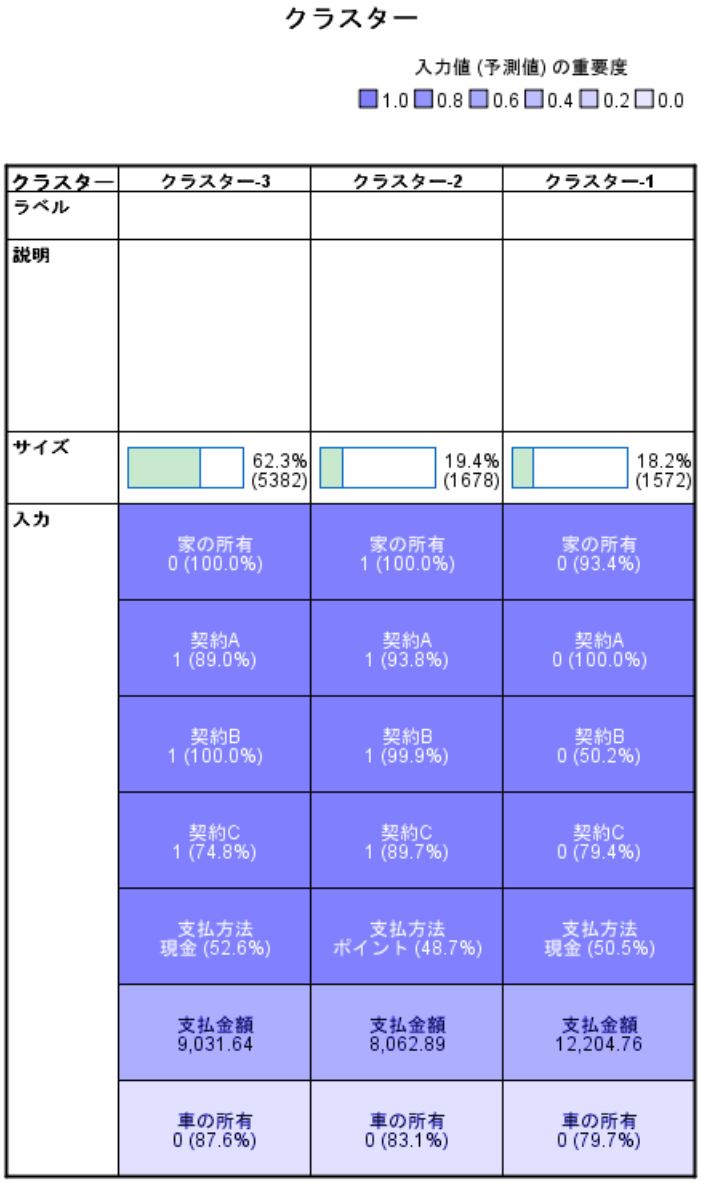

Figure5.3.8　クラスタービューのクラスター中心による表示

クラスタービューでは、各クラスターサイズとして**n**と**%**、各入力フィールドの要約情報を確認することができます。第1クラスターは**1,572**レコードで全体の**18.2%**、第2クラスターは**1,678**レコードで全体の**19.4%**、第3クラスターは**5,382**レコードで全体の**62.3%**です。

TIPS

クラスターの表示順番の設定は、下部のツールバーに含まれる**クラスターをサイズでソート**、**クラスター名でソート**、**クラスターをラベルでソート**の3種類選択することができます。デフォルトではサイズでソートされています。

入力フィールドとして**家の所有**の特徴に注目すると、第3クラスターでは**家の所有=0**が**100%**であり、このクラスターで家の所有がない特徴を持ちます。一方、第2クラスターでは**家の所有=1**が**100%**であり、このクラスターは家の所有がある特徴を持ちます。家の所有は、クラスターの構成に寄与しているようです。

また、**車の所有**の特徴に注目すると、3つのクラスターのいずれも**車の所有=0**が**80%**ほどであり、車の所有によるクラスターの違いはあまりなさそうです。このようなフィールドは、モデリングから除外しても良いかもしれません。

TIPS

入力フィールドの表示順番の設定は、下部のツールバーに含まれる**入力値をクラスター内の重要度でソート**、**入力値を名前でソート**、**入力値をデータでソート**の3種類選択することができます。デフォルトではクラスター内の重要度でソートされています。

操作手順

3. 左側の領域の下部にあるツールバーから**セルには絶対分布を表示**をクリックします。

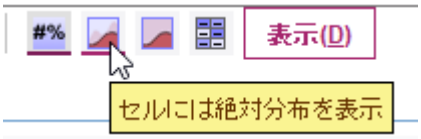

Figure5.3.9　クラスタービューのツールバー

	セルには クラスター中心を表示	連続型フィールドでは平均値、カテゴリ型フィールドの最頻カテゴリの%が表示されます。
	セルには 絶対分布を表示	フィールドの絶対分布を表示します。濃い赤はクラスター分布、淡い赤は全体のデータを表示します。
	セルには 相対分布を表示	フィールドの相対分布を表示します。濃い赤はクラスター分布、淡い赤は全体のデータを表示します。
	セルには 基本情報を表示	フィールド名のみを表示します。よりコンパクトな表示です。

Table5.3.1　セルの内容ツールバーのボタン

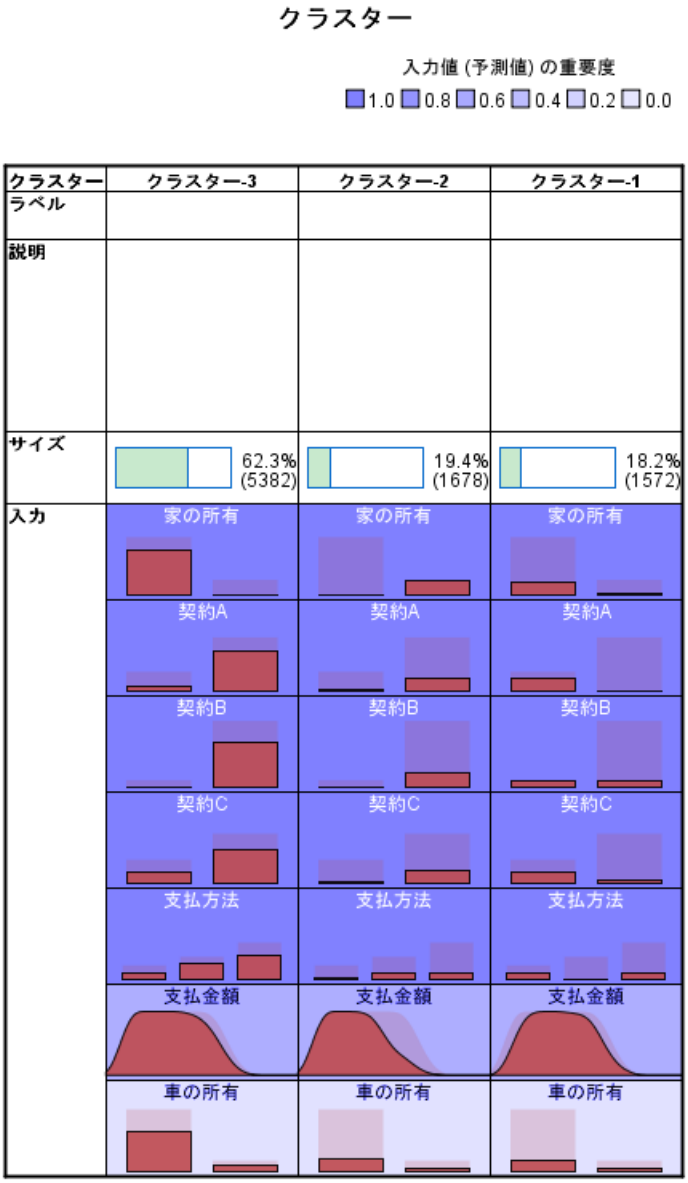

Figure5.3.10　クラスタービューの絶対分布による表示

絶対分布では、全体の分布とクラスターの分布を度数に基づいて表示します。視覚的に確認することができるため、平均値やパーセンテージの表示とあわせて各クラスターの特徴を評価します。

操作手順

4. 左側の領域で、クラスターの名前をCTRLキーを使用してすべて選択します。

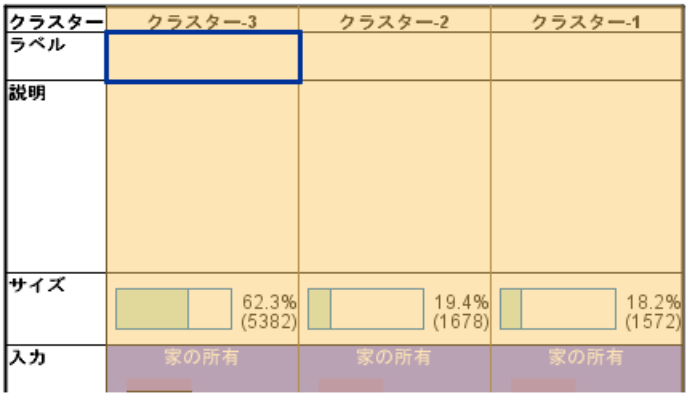

Figure5.3.11　クラスターを選択

クラスターを選択すると、右側の領域にクラスターの特徴をあらわすグラフが表示されます。**カテゴリ型**フィールドでは、最頻カテゴリが円で表示され、そのパーセントが高いものは大きい円で表示されます。この例では、クラスター**3**とクラスター**1**では**家の所有=0**が多く、クラスター**2**では**家の所有=1**が最も多いことが分かります。

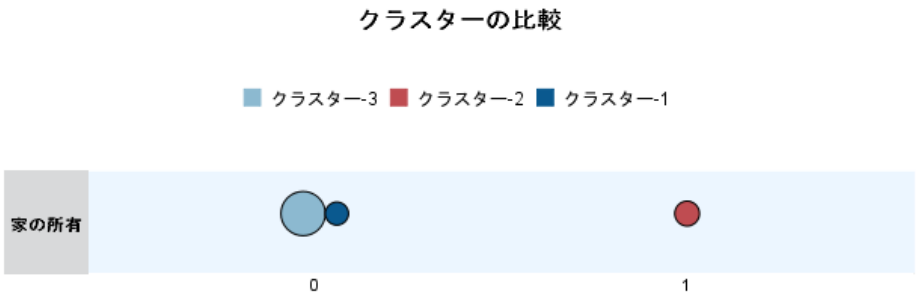

Figure5.3.12　カテゴリ型フィールドの比較の例

POINT

クラスターの比較では、カテゴリ型フィールドは最頻カテゴリが表示されます。

連続型フィールドの場合は箱ひげ図で表示され、全体の中央値と**4分位範囲**を示します。四角形のポイントマーカーと水平線は、それぞれ各クラスターの中央値と4分位範囲を示します。

支払金額フィールドに注目すると、クラスター**1**は全体の中央値より大きく、クラスター**2**とクラスター**3**では全体の中央値より低いことが分かります。

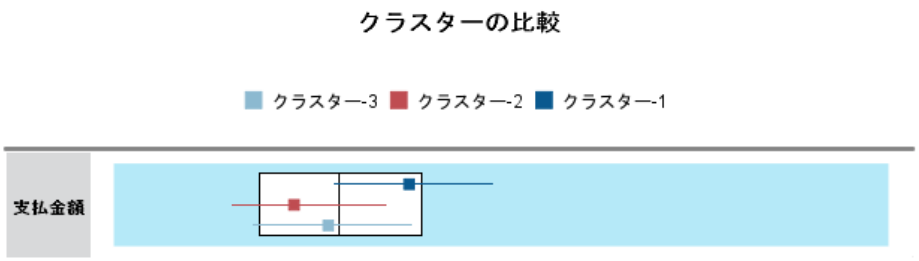

Figure5.3.13　連続型フィールドの比較の例

POINT

クラスターの比較では、**連続型**フィールドは箱ひげ図で表示され、全体の**中央値**と**4分位範囲**を示します。四角形のポイントマーカーと水平線は、それぞれ各クラスターの中央値と4分位範囲を示します。

POINT

4分位範囲は、データをソートして25%、50%、75%で分割した場合の75パーセンタイル値と25パーセンタイルの差を表します。

これらの出力を参考にしながら、クラスターの解釈とプロフィール作成を行います。

操作手順

5. **OK**ボタンをクリックして、モデルビューアーを閉じます。

POINT

クラスタリングは、多数の入力フィールドに基づいて、レコードを類似するグループに分類するための手法であり、分類されたクラスターが説明のつく結果になっているかどうかを評価するのは分析者です。この作業はプロフィール作成と呼ばれます。