IBM SPSS Modeler 18.1 データベース内マイニング・ガイド

注記

本書および本書で紹介する製品をご使用になる前に、 109 ページの『特記事項』に記載されている情報をお読みください。

本書は、IBM SPSS Modeler バージョン 18 リリース 0 モディフィケーション 0 および新しい版で明記されない限り、以降のすべてのリリースおよびモディフィケーションに適用されます。

お客様の環境によっては、資料中の円記号がバックスラッシュと表示されたり、バックスラッシュが円記号と表示されたりする場合があります。

原典： IBM SPSS Modeler 18.1 In-Database Mining Guide

発行： 日本アイ・ビー・エム株式会社

担当： トランスレーション・サービス・センター

目次

前書き

IBM® SPSS® Modeler は、IBM Corp. が開発した企業強化用のデータ・マイニング・ワークベンチです。SPSS Modeler を使用すると、企業はデータを詳しく調べることにより顧客および一般市民とのリレーションシップを強化することができます。企業は、SPSS Modeler を使用して得られた情報に基づいて利益をもたらす顧客を獲得し、抱き合わせ販売の機会を見つけ、新規顧客を引き付け、不正を発見し、リスクを減少させ、政府機関へのサービスの提供を改善することができます。

SPSS Modeler の視覚的インターフェースを使用すると、特定ビジネスの専門知識を適用し、より強力な予測モデルを実現し、解決までの時間を短縮します。SPSS Modeler では、予測、分類、セグメント化、および関連性検出アルゴリズムなど、さまざまなモデル作成手法を提供しています。モデルを作成した後は、IBM SPSS Modeler Solution Publisher により、企業全体の意思決定者やデータベースにモデルを配布することが可能になります。

IBM Business Analytics について

IBM Business Analytics ソフトウェアは、意思決定者がビジネス・パフォーマンスを向上させるために信頼する完全で、一貫した正確な情報を提供します。ビジネス・インテリジェンス、予測分析、財務実績および戦略管理、分析アプリケーション の包括的なポートフォリオを利用することによって、現在の実績を明確、迅速に理解し、将来の結果を予測することができます。豊富な産業用ソリューション、証明された実践法、それに専門家によるサービスを組み合せることにより、あらゆる規模の会社組織が、最高の生産性を推進し、信頼できる意志決定を自動化し、そして、よりよい結果を実現させることができます。

このポートフォリオの一部として、IBM SPSS Predictive Analytics ソフトウェアを使用する組織は、将来のイベントを予測し、その洞察に基づいて積極的に行動し、より優れた業績を実現することができます。全世界の企業、政府、学術分野のお客様が IBM SPSS の技術を活用し、不正行為を減少させ、リスクを軽減させながら、顧客の獲得、保持、成長において、競争優位を高めることができます。組織は、IBM SPSS ソフトウェアを日常業務に組み込むことにより、予測力を持つ企業になり、意思決定の管理と自動化を可能にすることで、ビジネス目標を達成し、重要な競争上の優位性を実現します。詳細な情報、または営業担当者へのお問い合わせ方法については、http://www.ibm.com/spss を参照してください。

技術サポート

お客様はテクニカル・サポートをご利用いただけます。IBM Corp. 製品の使用方法、または対応するハードウェア環境へのインストールについてサポートが必要な場合は、テクニカル・サポートにご連絡ください。テクニカル・サポートのご利用には、http://www.ibm.com/supportのIBM Corp. Web サイトをご覧ください。支援を要請される場合は、事前にユーザー、会社組織、そして、サポート契約を明確にしておいていただくよう、お願いします。

第 1 章 IBM SPSS Modeler について

IBM SPSS Modeler は、ビジネスの専門知識を活用して予測モデルを迅速に作成したり、また作成したモデルをビジネス・オペレーションに展開して意志決定を改善できるようにする、一連のデータ・マイニング・ツールです。IBM SPSS Modeler は業界標準の CRISP-DM モデルをベースに設計されたものであり、データ・マイニング・プロセス全体をサポートして、データに基づいてより良いビジネスの成果を達成できるようにします。

IBM SPSS Modeler には、マシン学習、人工知能、および統計に基づいたさまざまなモデル作成方法が用意されています。「モデル作成」パレットを利用して、データから新しい情報を引き出したり、予測モデルを作成することができます。各手法によって、利点や適した問題の種類が異なります。

SPSS Modeler は、 スタンドアロン製品として購入または SPSS Modeler Server と組み合わせてクライアントとして使用することができます。後のセクションで説明されているとおり、多くの追加オプションも使用することができます。 詳しくは、https://www.ibm.com/analytics/us/en/technology/spss/を参照してください。

IBM SPSS Modeler 製品

製品と関連するソフトウェアの IBM SPSS Modeler ファミリーの構成は次のとおりです。

- IBM SPSS Modeler
- IBM SPSS Modeler Server
- IBM SPSS Modeler Administration Console (IBM SPSS Deployment Manager に付属)
- IBM SPSS Modeler Batch
- IBM SPSS Modeler Solution Publisher
- IBM SPSS Modeler Server の IBM SPSS Collaboration and Deployment Services

IBM SPSS Modeler

SPSS Modeler はこの製品のすべての機能を搭載したバージョンであり、ユーザーのパーソナル・コンピューターにインストールし、そのコンピューターで実行します。スタンドアロン製品としてローカル・モードで SPSS Modeler を実行するか、大規模なデータ・セットを使用する場合にパフォーマンスを向上させるために IBM SPSS Modeler Server と組み合わせて実行することができます。

SPSS Modeler を使用して、プログラミングの必要なく、正確な予測モデルを迅速かつ直感的に構築することができます。独自のビジュアル・インターフェースを使用すると、データ・マイニング・プロセスを簡単に視覚化することができます。製品に組み込まれている高度な分析の支援を受けて、データ内に隠れたパターンやトレンドを発見することができます。結果をモデル化し、ビジネスチャンスを活用してリスクを軽減できるようになり、それらに影響を与える要因を理解することができます。

SPSS Modeler は SPSS Modeler Professional および SPSS Modeler Premium の 2 つのエディションで使用できます。 詳しくは、トピック 3 ページの『IBM SPSS Modeler のエディション』を参照してください。

IBM SPSS Modeler Server

SPSS Modeler は、クライアント/サーバー・アーキテクチャーを使用して、リソース集中型の操作が必要な要求を、強力なサーバー・ソフトウェアへ分散します。

SPSS Modeler Server は、1 つまたは複数の IBM SPSS Modeler のインストールと組み合わせてサーバー・ホストで分散分析モードで継続的に実行する、別途ライセンスが必要な製品です。このように、SPSS Modeler Server では、メモリー集中型の操作を、クライアント コンピューターにデータをダウンロードせずにサーバー上で実行できるため、大きなデータ・セットで優れたパフォーマンスを発揮します。IBM SPSS Modeler Server は、パフォーマンスと自動化のさらなる利点を提供し、SQLの最適化とデータベース内のモデリング機能をサポートしています。

IBM SPSS Modeler Administration Console

Modeler Administration Console は、SPSS Modeler Server 構成オプションの多くを管理するグラフィカル・ユーザー・インターフェースです。それらの構成オプションは、オプション・ファイルで設定することも可能です。コンソールは、IBM SPSS Deployment Manager に含まれています。コンソールを使用すると、SPSS Modeler Server インストール済み環境をモニターしたり、構成したりできます。SPSS Modeler Server の現在の顧客は、コンソールを無料で利用できます。アプリケーションは Windows コンピューターにのみインストールできますが、サポートされる任意のプラットフォームにインストールされたサーバーを管理できます。

IBM SPSS Modeler Batch

データマイニングは、通常、対話型のプロセスですが、グラフィカル・ユーザー・インターフェースを必要とせずに、コマンドラインから SPSS Modeler を実行することも可能です。例えば、ユーザーの介入なしで実行する長期実行または反復的なタスクがあります。SPSS Modeler Batch は、通常のユーザー・インターフェースにアクセスせずに SPSS Modeler の完全な分析機能のサポートを提供する製品の特別バージョンです。SPSS Modeler Batch を使用するには、SPSS Modeler Server が必要です。

IBM SPSS Modeler Solution Publisher

SPSS Modeler Solution Publisher は、外部ランタイムで実行することができ、外部アプリケーションに埋め込まれる SPSS Modeler ストリームのパッケージ版を作成することができるツールです。このように、SPSS Modeler がインストールされていない環境で使用するための完全な SPSS Modeler ストリームを公開して展開することができます。SPSS Modeler Solution Publisher は、個別のライセンスが必要とされている IBM SPSS Collaboration and Deployment Services - Scoring サービスの一部として配布されています。このライセンスを使用すると、SPSS Modeler Solution Publisher Runtime を受信し、公開されたストリームを実行することができます。

SPSS Modeler Solution Publisher について詳しくは、IBM SPSS Collaboration and Deployment Services の資料を参照してください。IBM SPSS Collaboration and Deployment Services Knowledge Center に『IBM SPSS Modeler Solution Publisher』と『IBM SPSS Analytics Toolkit』というセクションがあります。

IBM SPSS Collaboration and Deployment Services のための IBM SPSS Modeler Server アダプター

さまざまな IBM SPSS Collaboration and Deployment Services 用のアダプターを使用すると、SPSS Modeler および SPSS Modeler Server を IBM SPSS Collaboration and Deployment Services リポジトリーとインタラクティブに機能させることができます。このように、リポジトリーに展開された SPSS

Modeler ストリームは、複数のユーザーで共有したり、シンクライアント アプリケーションである IBM SPSS Modeler Advantage からアクセスしたりできます。リポジトリーをホストするシステムに、アダプターをインストールします。

IBM SPSS Modeler のエディション

SPSS Modeler は次のエディションで使用できます。

SPSS Modeler Professional

SPSS Modeler Professional は、CRM システムで追跡する行動や対話、人口統計データ、購入行動や販売データなど、多くの構造化データを処理するために必要なすべてのツールを提供しています。

SPSS Modeler Premium

SPSS Modeler Premium は、特化したデータ、または構造化されていないテキスト・データを処理するために SPSS Modeler Professional を拡張する、別途ライセンスが必要な製品です。SPSS Modeler Premium には、以下の IBM SPSS Modeler Text Analytics が含まれます。

IBM SPSS Modeler Text Analytics は、高度な言語技術と Natural Language Processing (NLP) を使用して、構造化されていない多様なテキスト・データをすばやく処理し、重要なコンセプトを抽出および組織化し、そしてそのコンセプトをカテゴリー別に分類します。抽出されたコンセプトとカテゴリーを、人口統計のような既存の構造化データと組み合わせ、IBM SPSS Modeler の豊富なデータ・マイニング・ツールを適用する方法で、焦点を絞ったより良い決定を下すことができます。

IBM SPSS Modeler ドキュメント

資料は、SPSS Modeler の「ヘルプ」メニューから参照できます。この「ヘルプ」メニューから SPSS Modeler Knowledge Center を開きます。Knowledge Center は、製品の外部で公に利用できます。

各製品の完全な資料 (インストール手順を含む) は、PDF 形式でも提供されており、製品ダウンロードの一部として、個別の圧縮フォルダーに格納されています。PDF 文書は、Web (http://www.ibm.com/support/docview.wss?uid=swg27046871) からもダウンロードできます。

SPSS Modeler Professional ドキュメント

SPSS Modeler Professional のドキュメント スイート (インストール手順を除く) は次のとおりです。

- **IBM SPSS Modeler** ユーザーズ・ガイド**:** SPSS Modeler の使用への全体的な入門で、データ ストリームの作成方法、欠損値の処理方法、CLEM 式の作成方法、プロジェクトおよびレポートの処理方法と、IBM SPSS Collaboration and Deployment Services または IBM SPSS Modeler Advantage に展開するためのストリームのパッケージ方法が含まれています。
- 「**IBM SPSS Modeler** 入力ノード、プロセス・ノード、出力ノード」。各種形式のデータの読み取り、処理、および出力に使用するすべてのノードの説明です。これは、モデル作成ノード以外のすべてのノードについての説明です。
- 「**IBM SPSS Modeler** モデル作成ノード」。データ・マイニング・モデルの作成に使用するすべてのノードについての説明です。IBM SPSS Modeler には、マシン学習、人工知能、および統計に基づいたさまざまなモデル作成方法が用意されています。

- **IBM SPSS Modeler** アプリケーション・ガイド。このガイドの例では、特定のモデル作成手法および技法について、簡単に対象を絞って紹介します。本ガイドのオンライン バージョンは、「ヘルプ」メニューからも利用できます。 詳しくは、トピック『アプリケーションの例』を参照してください。
- 「**IBM SPSS Modeler Python** スクリプトとオートメーション」。Python スクリプトによるシステムの自動化に関する情報です。ノードおよびストリームの操作に使用できるプロパティーを含めて説明します。
- **IBM SPSS Modeler** 展開ガイド**:** IBM SPSS Deployment Manager のもとで処理されるジョブ内のステップとして IBM SPSS Modeler のストリームを実行することに関する情報。
- **IBM SPSS Modeler CLEF** 開発者ガイド**:** CLEF では、IBM SPSS Modeler のノードとしてデータ処理ルーチンやモデル作成アルゴリズムなどのサード・パーティー製のプログラムを統合できます。
- 「**IBM SPSS Modeler** データベース内 マイニング・ガイド」。サード・パーティー製アルゴリズムを使用してご使用のデータベースの能力を利用してパフォーマンスを向上させ、分析機能の範囲を拡張する方法に関する情報を示します。
- **IBM SPSS Modeler Server** 管理およびパフォーマンス・ガイド**:** IBM SPSS Modeler Server の構成方法と管理方法に関する情報。
- 「**IBM SPSS Deployment Manager** ユーザー・ガイド」。IBM SPSS Modeler Server の監視や構成を行うための Deployment Manager アプリケーションに組み込まれている管理コンソール・ユーザー・インターフェースの使用法に関する情報。
- 「**IBM SPSS Modeler CRISP-DM** ガイド」。SPSS Modeler でのデータ・マイニングに対する CRISP-DM 方法の使用に関するステップバイステップのガイドです。
- 「**IBM SPSS Modeler Batch** ユーザーズ・ガイド」。IBM SPSS Modeler をバッチ・モードで使用するための完全ガイドで、バッチ・モードでの実行およびコマンド・ライン引数の詳細について説明します。このガイドは、PDF 形式のみです。

SPSS Modeler Premium ドキュメント

SPSS Modeler Premium のドキュメント スイート (インストール手順を除く) は次のとおりです。

- 「**SPSS Modeler Text Analytics** ユーザーズ・ガイド」。SPSS Modeler でテキスト分析を使用する場合の情報。テキスト・マイニング・ノード、インタラクティブ・ワークベンチ、テンプレートなどについて説明します。

アプリケーションの例

SPSS Modeler のデータ・マイニング・ツールは、多様なビジネスおよび組織の問題解決を支援しますが、アプリケーションの例では、特定のモデル作成手法および技術に関する簡単で、目的に沿った説明を行います。ここで使用されるデータセットは、データ・マイニング作業者によって管理される巨大なデータ・ストアよりも非常に小さいですが、関係するコンセプトや方法は実際のアプリケーションの規模に応じて拡張できます。

例にアクセスするには、SPSS Modeler の「ヘルプ」メニューで「アプリケーションの例」をクリックします。

データ・ファイルとサンプル・ストリームは、製品のインストール・ディレクトリーの `Demos` フォルダーにインストールされています。詳しくは、 5 ページの『Demos フォルダー』を参照してください。

データベース・モデル作成の例 **:** 例は、『*IBM SPSS Modeler* データベース内マイニング・ガイド』を参照してください。

スクリプトの例 : 例は、『*IBM SPSS Modeler* スクリプトとオートメーション ガイド』を参照してください。

Demos フォルダー

アプリケーションの例で使用されるデータ・ファイルとサンプル・ストリームは、製品のインストール・ディレクトリーの Demos フォルダーにインストールされています (例: C:¥Program Files¥IBM¥SPSS¥Modeler¥<version>¥Demos)。このフォルダーには、Windowsの「スタート」メニューの IBM SPSS Modeler プログラム・グループから、または「ファイル」 > 「ストリームを開く」ダイアログ・ボックスの最近使ったディレクトリーのリストで「Demos」をクリックしてアクセスすることもできます。

ライセンスの追跡

SPSS Modeler を使用すると、ライセンスの使用状況が一定の間隔で追跡され、ログに記録されます。ログに記録されるライセンスメトリックは *AUTHORIZED_USER* と *CONCURRENT_USER* であり、ログに記録されるメトリックのタイプは、SPSS Modeler に使用するライセンスのタイプによって決まります。

作成されたログファイルは IBM License Metric Tool によって処理可能であり、そのファイルからライセンス使用状況レポートを生成できます。

ライセンスログファイルは、SPSS Modeler クライアントログファイルが記録されるディレクトリと同じディレクトリに作成されます (デフォルトでは %ALLUSERSPROFILE%/IBM/SPSS/Modeler/<version>/log)。

第 2 章 データベース内マイニング

データベース・モデル作成の概要

IBM SPSS Modeler Server は、データベース・ベンダーから入手できる、IBM Netezza、Oracle Data Miner、Microsoft Analysis Services などのデータ・マイニングおよびモデル作成のツールとの統合をサポートしています。データベース内でのモデルの作成、スコアリング、および格納は、すべて IBM SPSS Modeler アプリケーション内から実行できます。これにより、ベンダーが提供するデータベース固有のアルゴリズムを活用しながら、IBM SPSS Modeler の分析機能と使いやすさを、データベースのパワーとパフォーマンスに組み合わせることが可能となります。モデルはデータベース内部に構築され、通常の方法で IBM SPSS Modeler インターフェースによって参照、スコアリングすることができます。さらに必要に応じて IBM SPSS Modeler Solution Publisher を使用して展開することもできます。使用可能なアルゴリズムは、IBM SPSS Modeler の「DB モデリング」パレットにあります。

データベース固有のアルゴリズムへのアクセスに IBM SPSS Modeler を使用することには、以下のようにいくつかの利点があります。

- データベース内アルゴリズムはデータベース・サーバーと密接に統合されることが多く、性能が向上することがあります。
- 「データベース内」で作成され格納されるモデルは、展開やデータベースにアクセス可能なアプリケーションとの共有をより簡単に実行できる場合があります。

SQL 生成。 データベース内モデル作成は SQL 生成と異なり、「SQL プッシュバック」として知られています。この機能を使用すると、パフォーマンス向上のためにデータベースに「プッシュバック」できる (データベースで実行できる) ネイティブの IBM SPSS Modeler 操作の SQL ステートメントを生成することができます。例えば、レコード結合、レコード集計、データ選択のノードは、すべて、この方法でデータベースにプッシュバックできる SQL コードを生成します。SQL 生成は、データベース モデル作成と組み合わせて使用すると、データベースの開始から終了までの実行が可能なストリームを生み出す場合があり、同時に、IBM SPSS Modeler で実行するストリームに大幅なパフォーマンスの向上をもたらします。

注 : データベース・モデル作成および SQL 最適化では、IBM SPSS Modeler Server 接続が IBM SPSS Modeler コンピューター上で可能でなければなりません。この設定を有効にすると、データベース・アルゴリズムにアクセスし、IBM SPSS Modeler から SQL を直接プッシュバック、IBM SPSS Modeler Server にアクセスできます。現在のライセンス ステータスを確認するには、IBM SPSS Modeler メニューから次を選択します。

「ヘルプ」 > 「バージョン情報」 > 「その他の詳細」

接続が有効な場合、「ライセンス ステータス」タブにオプション 「サーバーの有効化」 が表示されます。

サポートされているアルゴリズムの詳細は、この後の特定のベンダーの節を参照してください。

必要な準備

DB モデリングを行う前に、以下の設定が必要です。

- 必要な分析コンポーネント (Microsoft Analysis Services または Oracle Data Miner) をインストールした適切なデータベースへの ODBC 接続。
- IBM SPSS Modeler で DB モデリングを「ヘルパー アプリケーション」ダイアログ・ボックスから有効にします (「ツール」>「ヘルパー アプリケーション」)。
- IBM SPSS Modeler Server (使用している場合) と同様に、IBM SPSS Modeler でも「ユーザー オプション」ダイアログ・ボックスで 「**SQL** 生成」 および 「**SQL** 最適化」 の設定を有効にする必要があります。SQL 最適化は、厳密には DB モデリングを使用するためには必要ではありませんが、パフォーマンス上の理由から使用することを強くお勧めします。

注 : データベース・モデル作成および SQL 最適化では、IBM SPSS Modeler Server 接続が IBM SPSS Modeler コンピューター上で可能でなければなりません。この設定を有効にすると、データベース・アルゴリズムにアクセスし、IBM SPSS Modeler から SQL を直接プッシュバック、IBM SPSS Modeler Server にアクセスできます。現在のライセンス ステータスを確認するには、IBM SPSS Modeler メニューから次を選択します。

「ヘルプ」 > 「バージョン情報」 > 「その他の詳細」

接続が有効な場合、「ライセンス ステータス」タブにオプション 「サーバーの有効化」 が表示されます。

詳細は、この後の特定ベンダーの節を参照してください。

モデルの構築

データベースのアルゴリズムを使用した、モデルの構築とスコアリングのプロセスは、IBM SPSS Modeler のほかの種類のデータ マイニングに似ています。ノードとモデル作成「ナゲット」での作業の一般的なプロセスは、IBM SPSS Modeler で作業する場合の他のストリームに似ています。唯一の相違点は、実際の処理とモデルの構築がデータベースにプッシュバックされることです。

データベース・モデリング・ストリームは、概念的には IBM SPSS Modeler 内の他のデータ・ストリームと同じです。ただし、このストリームは、Microsoft Decision Tree ノードを使用したモデルの構築などを含め、データベース内のすべての操作を実行します。このストリームを実行すると、IBM SPSS Modeler は、データベースに構築と、作成されたモデルの保存を指示し、また詳細が IBM SPSS Modeler にダウンロードされます。データベース内実行は、ストリーム内で紫色の影が付いているノードを使用することで示されます。

データの準備

データベース固有のアルゴリズムが使用されているか否かにかかわらず、データの準備をできるだけデータベースにプッシュバックして、パフォーマンスを向上させる必要があります。

- その目的は、元のデータがデータベースに保存されている場合に、必要な上流操作がすべて SQL に変換できることを確認したうえで、データをデータベース内に保持しておくことです。これによって、データの IBM SPSS Modeler へのダウンロードが防止されるため、ゲインを無効にする可能性があるボトルネックが回避され、ストリーム全体がデータベースで実行できるようになります。
- 元のデータが、データベースに保存されていない 場合でも、DB モデリングを使用できます。この場合、データの準備は IBM SPSS Modeler で行われ、準備が完了したデータ・セットは、モデル構築のために自動的にデータベースにアップロードされます。

モデル・スコアリング

データベース内マイニングを使用して IBM SPSS Modeler から生成されるモデルは、通常の IBM SPSS Modeler モデルとは異なります。それらは、生成されたモデル「ナゲット」として、モデル・マネージャーに表示されますが、実際には、リモートのデータ・マイニングやデータベース・サーバーに保持されるリモート・モデルです。IBM SPSS Modeler に表示されるものは、単にこのリモート・モデルへの参照です。つまり、表示される IBM SPSS Modeler モデルは、データベース・サーバー・ホスト名、データベース名およびモデル名などの情報を含む「hollow」モデルです。これは、データベース ネイティブ・アルゴリズムを使用して作成されるモデルを参照したりスコアリングしたりする場合に、理解すべき重要な違いです。

モデルが作成されると、IBM SPSS Modeler で生成される他のモデルのようにスコアリングするために、ストリームに追加できます。たとえ上流の操作でなくても、すべてのスコアリングはデータベース内で行います(パフォーマンスの向上にとって可能ならば上流の操作はデータベースにプッシュバックされますが、これはスコアリングを実行するための要件ではありません)。ほとんどの場合、データベース・ベンダーが提供する標準的なブラウザーを使用して、作成されたモデルをブラウズすることもできます。

参照する場合もスコアリングする場合も、Oracle Data Miner または Microsoft Analysis Services が動作しているサーバーへのライブ接続が必要です。

結果の表示と設定の指定

スコアリングの結果を表示し設定を指定するには、ストリーム領域のモデルをダブルクリックします。代わりに、モデルを右クリックして 「参照」 または 「編集」 を選択することもできます。具体的な設定値は、モデルのタイプに依存します。

データベース・モデルのエクスポートと保存

「ファイル」メニューにあるオプションを使用すると、データベース・モデルおよび要約は、IBM SPSS Modeler で作成された他のモデルと同様の方法でモデル・ブラウザーからエクスポートできます。

1. モデル・ブラウザーの「ファイル」メニューから次のオプションを選択します。

- 「テキストのエクスポート」 を選択してモデル要約をテキスト・ファイルにエクスポート
- 「**HTML** 形式でエクスポート」 を選択してモデル要約を HTML ファイルにエクスポート
- **PMML** をエクスポート (IBM DB2 IM モデルのみをサポート) を選択し、その他の PMML 互換性のあるソフトウェアと使用できる予測モデル・マークアップ言語 (PMML) としてモデルをエクスポート

注: 生成されたモデルは、「ファイル」メニューから 「ノードの保存」 を選択して保存することもできます。

モデルの整合性

生成された各データベース・モデルについて、IBM SPSS Modeler は、そのデータベースに格納されているものと同じ名前のモデルへの参照と共に、モデル構造の説明を保存します。生成されたモデルの「サーバー」タブは、データベース上にある実際のモデルに一致するユニークなキーを表示します。

IBM SPSS Modeler はこのランダムに生成されたキーを、モデルの整合性が維持されているかどうかを検査するために使用します。このキーは、モデルの構築時にモデルの説明に保存されます。展開ストリームを実行する前に、キーが一致することを確認しておくとよいでしょう。

1. データベースに保存されているモデルの説明と IBM SPSS Modeler により保存されているランダム・キーとを比較することで整合性を検査するには、「検査」ボタンをクリックします。そのデータベース・モデルが見つからないか、キーが一致しない場合、エラーが報告されます。

生成された SQL の表示とエクスポート

生成された SQL コードは、実行前にプレビューすることができ、デバッグに役立ちます。

第 3 章 Microsoft Analysis Services によるデータベース・モデリング

IBM SPSS Modeler と Microsoft Analysis Services

IBM SPSS Modeler は、Microsoft SQL Server Analysis Services との統合をサポートします。この機能は IBM SPSS Modeler のモデル作成ノードとして実装され、「DB モデリング」パレットから使用できます。パレットが表示されていない場合は、「ヘルパー アプリケーション」ダイアログ・ボックスから、「Microsoft」タブ上にある「MS Analysis Services の統合」を有効にしてアクティブ化できます。 詳しくは、トピック 13 ページの『Analysis Services との統合を有効にする』を参照してください。

IBM SPSS Modeler は、次の Analysis Services アルゴリズムの統合をサポートします。

- デシジョン ツリー
- クラスタリング
- アソシエーション・ルール
- Naive Bayes
- 線型回帰
- ニューラル・ネットワーク
- ロジスティック回帰
- 時系列
- シーケンス・クラスタリング

次の図は、クライアントから、 IBM SPSS Modeler Server がデータベース内マイニングを管理するサーバーへのデータの流れを説明しています。モデル作成は、Analysis Services を使用して実行されます。作成されたモデルは、Analysis Services により保存されます。モデルへの参照が、IBM SPSS Modeler ストリーム内に維持されます。次に、モデルは Analysis Services から、スコアリングのために Microsoft SQL Server、または、IBM SPSS Modeler のいずれかにダウンロードされます。

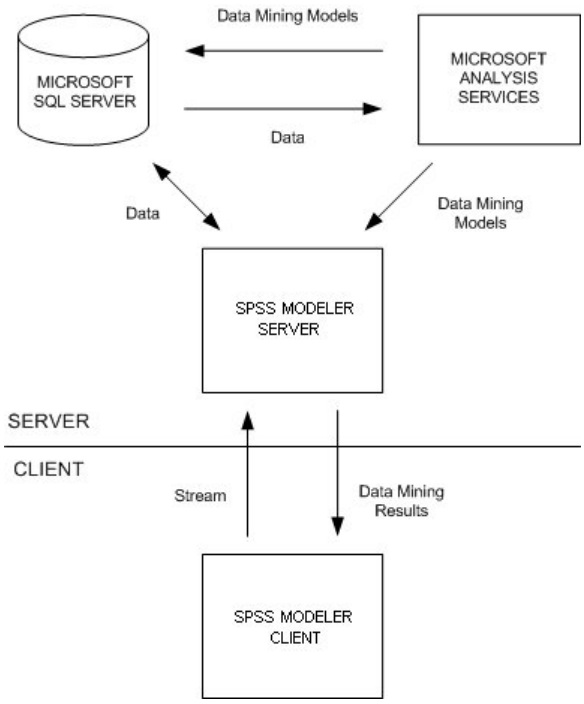

図 *1.* モデル作成中の *IBM SPSS Modeler*、*Microsoft SQL Server*、および *Microsoft Analysis Services* 間のデータ・フロー

注 : IBM SPSS Modeler Server は使用できますが、必須ではありません。IBM SPSS Modeler クライアント自体がデータベース内マイニングの計算を処理することができます。

Microsoft Analysis Services との統合の要件

IBM SPSS Modeler と共に Analysis Services アルゴリズムを使用してデータベース内モデル作成を行なう場合の前提条件を次に示します。場合によっては、これらの条件が満たされているかデータベース管理者に問い合わせて確認してください。

- Windows 上の IBM SPSS Modeler Server インストール済み環境 (分散モード) に対する IBM SPSS Modeler の実行。Analysis Services との統合では、UNIX プラットフォームはサポートされません。

  重要**:** IBM SPSS Modeler ユーザーは、「*IBM SPSS Modeler Server* の追加の要件 」の下に示す Microsoft の URL からダウンロードできる SQL Native Client ドライバーを使用して、ODBC 接続を構成する必要があります。*IBM SPSS Data Access Pack* に付属するドライバ (通常、*IBM SPSS Modeler* で他の用途が推奨されている) は、この目的には推奨できません。 ドライバーは、「統合 **Windows** 認証」が有効になっている SQL Server を使用するように構成する必要があります。IBM SPSS Modeler は、SQL Server 認証をサポートしていないためです。ODBC データ・ソースに対するアクセス権の作成や設定に関する質問がある場合は、データベース管理者に問い合わせてください。

- SQL Server がインストールされている必要があります。これは必ずしも IBM SPSS Modeler と同一のホスト上の必要はありません。IBM SPSS Modeler ユーザーには、データの読み込みおよび書き込み、テーブルとビューのドロップおよび作成に適切なアクセス権が必要です。

  注**:** SQL Server Enterprise Edition が推奨されています。Enterprise Edition は、高度なパラメーターを提供することによって、アルゴリズムの結果を調整することができます。Standard Edition バージョンは、同一パラメーターを提供しますが、高度なパラメーターの中には編集できないものもあります。

- Microsoft SQL Server Analysis Services は、SQL Server と同一ホストにインストールされている必要があります。

IBM SPSS Modeler Server の追加の要件

IBM SPSS Modeler Server と共に Analysis Services のアルゴリズムを使用する場合は、次のコンポーネントが IBM SPSS Modeler Server ホスト・マシンにインストールされている必要があります。

注**:** SQL Server is が IBM SPSS Modeler Server と同一ホストにインストールされている場合、これらのコンポーネントはすでに利用可能になっています。

- Microsoft SQL Server Analysis Services 10.0 OLE DB Provider (使用している OS に適切なバリアントを選択してください)
- Microsoft SQL Server Native Client (使用している OS に適切なバリアントを選択してください)
- Microsoft SQL Server 2008 または 2012 を使用している場合は Microsoft Core XML Services (MSXML) 6.0 も必要となる場合があります。

これらのコンポーネントをダウンロードするには、*www.microsoft.com/downloads* に移動して **.NET Framework** または (他のすべてのコンポーネントの場合は) **SQL Server Feature Pack** を検索し、使用バージョンの SQL Server の最新パックを選択します。

これらは、ほかのパッケージが先にインストールされている必要がある場合があります。そのようなパッケージも、Microsoft Downloads Web サイトからダウンロードして利用できます。

IBM SPSS Modeler の追加の要件

IBM SPSS Modeler と共に Analysis Services アルゴリズムを使用するには、上記と同じコンポーネントの他に、クライアントで以下をインストールする必要があります。

- Microsoft SQL Server Datamining Viewer Controls (使用している OS に適切なバリアントを選択してください) も以下が必要です。
- Microsoft ADOMD.NET

これらのコンポーネントをダウンロードするには、*www.microsoft.com/downloads* に移動して **SQL Server Feature Pack** を検索し、使用バージョンの SQL Server の最新パックを選択します。

注 : データベース・モデル作成および SQL 最適化では、IBM SPSS Modeler Server 接続が IBM SPSS Modeler コンピューター上で可能でなければなりません。この設定を有効にすると、データベース・アルゴリズムにアクセスし、IBM SPSS Modeler から SQL を直接プッシュバック、IBM SPSS Modeler Server にアクセスできます。現在のライセンス ステータスを確認するには、IBM SPSS Modeler メニューから次を選択します。

「ヘルプ」 > 「バージョン情報」 > 「その他の詳細」

接続が有効な場合、「ライセンス ステータス」タブにオプション 「サーバーの有効化」 が表示されます。

Analysis Services との統合を有効にする

Analysis Services との IBM SPSS Modeler の統合を有効にするには、SQL Server および Analysis Services を設定し、ODBC ソースを作成して、IBM SPSS Modeler の「ヘルパー アプリケーション」ダイアログ・ボックスで統合を有効にする必要があります。さらに、SQL の生成および最適化を有効にします。

注: Microsoft SQL Server および Microsoft Analysis Services が利用可能になっている必要があります。詳しくは、トピック 12 ページの『Microsoft Analysis Services との統合の要件』を参照してください。

SQL Server を設定する

SQL Server を設定して、データベース内でスコアリングが実行できるようにします。

1. SQL Server ホスト・コンピューター上で次のレジストリ キーを作成します。

   ```
   HKEY_LOCAL_MACHINE¥SOFTWARE¥Microsoft¥MSSQLServer¥Providers¥MSOLAP
   ```

2. このキーに次の DWORD を追加します。

   ```
   AllowInProcess 1
   ```

3. この変更が完了したら SQL Server を再起動します。

Analysis Services の設定

IBM SPSS Modeler で Analysis Services との通信を開始する前に、まず、「Analysis Server Properties」ダイアログ・ボックスで 2 つの設定を手動で行う必要があります。

1. MS SQL Server Management Studio から Analysis Server へログインします。
2. サーバー名を右クリックし 「プロパティー」 を選択して、「プロパティー」ダイアログ・ボックスにアクセスします。
3. 「詳細プロパティーを **(**すべて**)** 表示する」 チェック・ボックスを選択します。
4. 次のプロパティーを変更します。

- DataMining¥AllowAdHocOpenRowsetQueries の値を True （真）に変更します（デフォルト値は False （偽））。
- DataMining¥AllowProvidersInOpenRowset の値を [all] に変更します（デフォルト値はありません）。

SQL Server の ODBC DSNを作成する

データベースを読み書きするには、ODBC データ・ソースがインストールされていて、該当するデータベースに対して必要に応じて読み取り権限や書き込み権限が設定されている必要があります。Microsoft SQL Native Client ODBC ドライバーが必要で、SQL Server で自動的にインストールされます。*IBM SPSS Data Access Pack* に付属するドライバ *(*通常、*IBM SPSS Modeler* で他の用途が推奨されている*)* は、この目的には推奨できません。 IBM SPSS Modeler と SQL Server が異なるホストにある場合、Microsoft SQL Native Client ODBC ドライバーをダウンロードできます。 詳しくは、トピック 12 ページの『Microsoft Analysis Services との統合の要件』を参照してください。

ODBC データ・ソースに対するアクセス権の作成や設定に関する質問がある場合は、データベース管理者に問い合わせてください。

1. Microsoft SQL Native Client ODBC ドライバーを使用して、データ・マイニング・プロセスで使用される SQL Server データベースを指し示す ODBC DSN を作成します。その他は、デフォルトのドライバ設定を使用します。
2. この DSN を使用するために、「**Integrated Windows Authentication** と共に」 が選択されていることを確かめてください。

- IBM SPSS Modeler と IBM SPSS Modeler Server が、それぞれ異なるコンピューター上で実行されている場合、両方のコンピューターに同じ ODBC DSN を作成します。それぞれのホストに同じ DSN 名が使われていることを確かめます。

IBM SPSS Modeler で Analysis Services Integration を有効にする

IBM SPSS Modeler を有効にして Analysis Services を使用するには、まず「ヘルパー アプリケーション」ダイアログ・ボックスでサーバー明細を準備する必要があります。

1. IBM SPSS Modeler のメニューから次の項目を選択します。

   「ツール」 > 「オプション」 > 「ヘルパー アプリケーション」

2. 「**Microsoft**」タブをクリックします。

- **Microsoft Analysis Services** 統合を有効化**:** IBM SPSS Modeler ウィンドウの下部でデータベース モデリング パレットを有効にし (まだ表示されていない場合)、Analysis Services アルゴリズムのノードを追加します。
- **Analysis** サーバー・ホスト **:** Analysis Services が動作しているマシンの名前を指定します。
- **Analysis** サーバー・データベース**:** 省略符号 (「...」) ボタンをクリックして、使用可能なデータベースを選択できるサブダイアログ ボックスを開き、目的のデータベースを選択します。リストは、指定した Analysis サーバーで使用可能なデータベースに取り込まれます。Microsoft Analysis Services は、データ マイニング モデルを指定のデータベースに保存するため、IBM SPSS Modeler により構築された Microsoft モデルの保存先の適切なデータベースを指定する必要があります。
- **SQL Server** 接続 **:** SQL サーバー・データベースが使用する DSN 情報を指定して、Analysis サーバーに渡されるデータを保存します。Analysis Services データ・マイニング・モデルを構築するためのデータ提供に使用される、ODBC データ・ソースを選択します。フラット・ファイルまたは ODBC データ・ソースにより提供されるデータから Analysis Services モデルを構築する場合、データは、この ODBC データ・ソースが指定している SQL Server データベース内に作成される一時テーブルに自動的にアップロードされます。
- データ・マイニング・モデルの上書き時に警告する **:** データベースに格納されているモデルを、警告なしに IBM SPSS Modeler が上書きしないようにするために選択します。

注 : 「ヘルパー アプリケーション」ダイアログ・ボックスで行った設定は、さまざまな Analysis Services ノードの内部でオーバーライドできます。

SQL の生成と最適化を有効にする

1. IBM SPSS Modeler のメニューから次の項目を選択します。

   「ツール」 > 「ストリームのプロパティー」 > 「オプション」

2. ナビゲーション ペインの 「最適化」 オプションをクリックします。
3. 「**SQL** 生成」 オプションが有効になっていること確認します。この設定は、データベースのモデル作成が機能するために必要です。
4. 「**SQL** 生成の最適化」 と 「その他の実行を最適化」 を選択します (絶対に必要な訳ではありませんが、最適化されたパフォーマンスを得るために、選択することを強くお勧めします)。

Analysis Services でのモデル構築

Analysis Services でモデルを構築するには、学習データ・セットを、SQL Server データベース内のテーブルまたはビューに配置する必要があります。データが、SQL Server にない場合、または、SQL Server では行えないデータ準備作業として IBM SPSS Modeler で処理する必要がある場合、データはモデル構築の前に SQL Server の一時テーブルに自動的にアップロードされます。

Analysis Services モデルの管理

IBM SPSS Modeler によって Analysis Services モデルを構築すると、IBM SPSS Modeler 内にモデルが作成されると同時に、SQL Server データベース内でもモデルの作成や置き換えが行われます。この種類の

IBM SPSS Modeler モデルは、データベース・サーバーに格納されているデータベース・モデルの内容を参照します。IBM SPSS Modeler は、IBM SPSS Modeler モデルと SQL Server モデルの両方に、同一のモデル・キー文字列を生成して格納し、整合性チェックを実行します。

MS ディシジョン・ツリー モデル作成ノードは、カテゴリー属性と連続値属性の両方がある、予測的なモデル作成に使用されます。カテゴリー属性については、ノードで、データセット内の入力列間の関係に基づいた予測が行われます。例えば、どの顧客が自転車を購入するかを予測する計画で、若い顧客は 10 人のうち 9 人までが自転車を購入するのに対し、年長者の顧客は 10 人のうち 2 人しか購入しない場合、ノードにより、年齢は自転車購入の予測に適した予測フィールドであると推論されます。ディシジョン・ツリーでは、このような傾向に基づいて特定の結果に向かう予測が行われます。連続値の属性については、ディシジョン・ツリーを分割する箇所を決めるために線型回帰が使用されます。複数の列が予測値に設定される場合、または入力データが予測値に設定される入れ子のテーブルを含んでいる場合は、ノードにより、それぞれの列予測値に対応する個別のディシジョン・ツリーが作成されます。

MS クラスタリング モデル作成ノードでは、データセット内のケースをクラスター内で似たような特徴を持つグループにまとめる反復的な手法が使用されます。このようなグループ化は、データの検証、データ内の異常性の識別、予測の作成に役立ちます。クラスタリング・モデルで、通常の観察からは論理的に引き出せない可能性がある関係が、データセット内で識別されます。例えば、自転車で通勤する人は、通常は勤務先から遠い場所に住んでいないと論理的に認識できます。ただしアルゴリズムによって、自転車通勤者のそれほど明白とも言えない他の特徴も発見できます。クラスタリング・ノードは、対象フィールドが指定されないという点で、他のデータ・マイニング・ノードとは異なります。クラスタリング・ノードにより、データ内と、ノードに識別されるクラスターからの関係に従って、モデルが正確に調整されます。

MS アソシエーション ルール モデル作成ノードは、購買アドバイス エンジンにとって有用です。購買アドバイス・エンジンで、顧客がすでに購入したり興味を示したりしたアイテムに基づいて、商品が顧客に推奨されます。アソシエーション モデルは、個々のケースとケースに含まれるアイテムの両方の識別子が含まれたデータセットに対して構築されます。ケース内のアイテムのグループは、アイテム・セットと呼ばれます。アソシエーション・モデルは、一連のアイテムセットと、アイテムがケース内でグループ化される方法を記述したルールからなります。アルゴリズムで識別されるルールは、顧客のショッピング カート内にすでに存在するアイテムに基づいて、顧客の将来の購買を予測するのに使用できます。

MS Naive Bayes モデル作成ノードでは、対象フィールドと予測フィールド間の条件付きの確率が計算され、各列が独立していると見なされます。提示されたすべての予測変数は相互に依存関係がないものとして処理されるので、モデルは naïve と命名されます。この手法はほかの Analysis Services のアルゴリズムより計算上の強度が低く、そのため、モデル作成の準備段階で相互関係性を簡単に検出するのに役立ちます。このノードは、データを最初に検索し、その結果を適用して、計算に時間がかかってもより正確な結果が得られる他のノードと組み合わせて追加のモデルを作成するために使用できます。

MS 線型回帰モデル作成ノードは ディシジョン・ツリー・ノードの一種で、`MINIMUM_LEAF_CASES` パラメーターを、ノードがマイニング・モデルを学習するために使用するデータ・セット内の合計件数以上に設定します。パラメーターをこのように設定することによって、ノードは分岐を行うことはなく、線型回帰を実行します。

MS ニューラル・ネットワーク モデル作成ノードは、予測可能な属性の各状態が指定された場合、ニューラル・ネットワーク・ノードが入力属性の可能性がある各状態に対する可能性を計算するという点で、MS ディシジョン・ツリー・ノードと類似しています。後でこれらの可能性を使用し、入力属性に基づいて予測された属性の結果を予測することができます。

MS 線型回帰 モデル作成ノードは、MS ニューラル・ネットワーク・ノードのバリエーションで、HIDDEN_NODE_RATIO パラメーターは 0 に設定されています。この設定により、隠れ層を含まないニューラル・ネットワーク・モデルを作成するため、ロジスティック回帰と同じです。

MS 時系列モデル作成ノードでは、商品の売り上げなど、時間を経過した連続型値の予測に最適化された回帰アルゴリズムを提供します。ディシジョン・ツリーなど、その他の Microsoft アルゴリズムには、新しい情報の追加の列が傾向を予測するための入力として必要ですが、時系列モデルでは必要ありません。時系列モデルは、モデルの作成に使用される元のデータセットにのみ基づいて傾向を予測できます。予測を行う場合、モデルに新しいデータを追加して、自動的に傾向分析の新しいデータを結合します。 詳しくは、トピック 19 ページの『MS タイム シリーズ・ノード』を参照してください。

MS シーケンス・クラスタリング モデル作成ノードは、データ内の順序が指定されたシーケンスを特定し、この分析の結果をクラスタリング手法によって結合し、シーケンスとその他の属性に基づいてクラスターを生成します。 詳しくは、トピック 21 ページの『MS シーケンス・クラスタリング・ノード』を参照してください。

それぞれのノードは、IBM SPSS Modeler ウィンドウの下部にある「DB モデリング」パレットからアクセスできます。

すべてのアルゴリズム・ノードに共通の設定

次の設定は、すべての Analysis Services アルゴリズムに共通です。

サーバー・オプション

「サーバー」タブでは Analysis サーバー・ホストとデータベース、SQL Server データ・ソースを設定します。ここで指定したオプションで、「ヘルパー アプリケーション」ダイアログ ソックスの「Microsoft」タブで指定した設定が上書きされます。 詳しくは、トピック 13 ページの『Analysis Services との統合を有効にする』を参照してください。

注 : このタブの一種も、また、Analysis Services モデルをスコアリングする際に使用することができます。 詳しくは、トピック 22 ページの『Analysis Services モデル・ナゲットの「サーバー」タブ』を参照してください。

モデル オプション

最も基本的なモデルを構築するためには、作業の前に「モデル」タブの設定を指定する必要があります。「エキスパート」タブではスコアリング方法やその他の詳細設定を指定できます。

次の基本的なモデル作成オプションを利用できます。

モデル名**:** ノードの実行時に作成されたモデルに割り当てられる名前を指定します。

- 自動: モデル名は、対象フィールドまたは ID フィールド名、または、クラスタリング・モデル・ノードなどのように対象が指定されていない場合モデル タイプの名前に基づいて自動的に生成されます。
- カスタム: 作成されたモデルのカスタム名を指定できるようになります。

データ区分データを使用。 現在のデータ区分フィールドに基づいて、データを、学習用、テスト用、および検証用の独立したサブセット、つまりサンプルに分割します。1 組のサンプルをモデルの作成に使用し、別の組のサンプルで生成したモデルをテストすることにより、そのモデルが、このデータに似た性質を持つより大きなデータセットにどの程度適用できるかについての目安を得ることができます。データ区分フィールドがストリーム中で指定されていない場合、このオプションは無視されます。

ドリルスルーあり : 表示された場合、このオプションを選択すると、モデルを問い合わせて、モデルに含まれるケースについての詳細を知ることができます。

一意のフィールド : ドロップダウン・リストから、各ケースを一意に識別するフィールドを選択します。通常、これは 「**CustomerID**」 などの ID フィールドです。

MS ディシジョン・ツリーのエキスパート・オプション

「エキスパート」タブで利用可能なオプションは、選択したストリームの構造に応じて変化する可能性があります。各 Analysis Services モデル・ノードで選択したエキスパート・オプションの詳細は、ユーザー・インターフェースのフィールド レベルのヘルプを参照してください。

MS クラスタリングのエキスパート・オプション

「エキスパート」タブで利用可能なオプションは、選択したストリームの構造に応じて変化する可能性があります。各 Analysis Services モデル・ノードで選択したエキスパート・オプションの詳細は、ユーザー・インターフェースのフィールド レベルのヘルプを参照してください。

MS Naive Bayes のエキスパート・オプション

「エキスパート」タブで利用可能なオプションは、選択したストリームの構造に応じて変化する可能性があります。各 Analysis Services モデル・ノードで選択したエキスパート・オプションの詳細は、ユーザー・インターフェースのフィールド レベルのヘルプを参照してください。

MS 線型回帰のエキスパート・オプション

「エキスパート」タブで利用可能なオプションは、選択したストリームの構造に応じて変化する可能性があります。各 Analysis Services モデル・ノードで選択したエキスパート・オプションの詳細は、ユーザー・インターフェースのフィールド レベルのヘルプを参照してください。

MS ニューラル・ネットワークのエキスパート・オプション

「エキスパート」タブで利用可能なオプションは、選択したストリームの構造に応じて変化する可能性があります。各 Analysis Services モデル・ノードで選択したエキスパート・オプションの詳細は、ユーザー・インターフェースのフィールド レベルのヘルプを参照してください。

MS ロジスティック回帰エキスパート・オプション

「エキスパート」タブで利用可能なオプションは、選択したストリームの構造に応じて変化する可能性があります。各 Analysis Services モデル・ノードで選択したエキスパート・オプションの詳細は、ユーザー・インターフェースのフィールド レベルのヘルプを参照してください。

MS アソシエーション・ルール・ノード

MS のアソシエーション・ルール・モデル作成ノードは、購買アドバイス・エンジンにとって有用です。購買アドバイス・エンジンで、顧客がすでに購入したり興味を示したりしたアイテムに基づいて、商品が顧客に推奨されます。アソシエーション モデルは、個々のケースとケースに含まれるアイテムの両方の識別子が含まれたデータセットに対して構築されます。ケース内のアイテムのグループは、アイテムセットといいます。

アソシエーション・モデルは、一連のアイテムセットと、アイテムがケース内でグループ化される方法を記述したルールからなります。アルゴリズムで識別されるルールは、顧客のショッピング カート内にすでに存在するアイテムに基づいて、顧客の将来の購買を予測するのに使用できます。

テーブル形式データの場合、アルゴリズムは生成された各推奨事項 ($M-*field*) の確率 ($MP-*field*) を示すスコアを作成します。トランザクション形式データの場合、アルゴリズムは生成された各推奨事項 ($M-*field*) のサポート ($MS-*field*)、確率 ($MP-*field*) および調整済み確率 ($MAP-*field*) のスコアを作成します。

要件

トランザクション形式アソシエーション・モデルの要件は次のとおりです。

- 一意のフィールド **:** アソシエーション・ルール・モデルには一意にレコードを特定するキーが必要です。
- **ID** フィールド**:** MS アソシエーション・ルール・モデルをトランザクション形式データで作成する場合、各トランザクションを特定する ID フィールドが必要です。 ID フィールドは一意のフィールドと同じように設定できます。
- **1** つ以上の入力フィールド **:** アソシエーション・ルール・アルゴリズムには、少なくとも 1 つの入力フィールドが必要です。
- 対象フィールド**:** MS アソシエーション・モデルをトランザクション形式データで作成する場合、対象フィールドはユーザーが購入した製品など、トランザクション・フィールドでなければなりません。

MS アソシエーション・ルールのエキスパート・オプション

「エキスパート」タブで利用可能なオプションは、選択したストリームの構造に応じて変化する可能性があります。各 Analysis Services モデル・ノードで選択したエキスパート・オプションの詳細は、ユーザー・インターフェースのフィールド レベルのヘルプを参照してください。

MS タイム シリーズ・ノード

MS タイム シリーズ モデル作成ノードは、次の 2 つの予測をサポートしています。

- 将来
- 過去

将来の予測は、過去のデータ以降の指定された期間、対象フィールド値を推定し、これは常に実行されます。過去の予測は、過去のデータに実際の値が指定されている期間、対象フィールド値を推定します。過去の予測とは、過去のデータに実際の値がある指定した数の期間について、推定された対象フィールド値です。過去の予測を使用して、実際の過去の値と予測値を比較することで、モデルの品質を評価できます。予測の始点の値によって、過去の予測が実行されるかどうかが決まります。

IBM SPSS Modeler 時系列ノードとは異なり、MS タイム シリーズ・ノードの前に時間区分ノードは必要ありません。さらに異なるのは、デフォルトで時系列データのすべての過去の行ではなく、予測行にのみスコアが作成されるという点です。

要件

MS タイム シリーズ モデルの要件は次のとおりです。

- 単一のキー時間フィールド：各モデルには、ケース・シリーズとして使用され、モデルが使用するタイム・スライスを定義する数値型フィールドまたは日付フィールドが 1 つ必要です。キー時間フィールドのデータ型は、日付/時刻型または数値型です。ただし、フィールドには連続型の値が含まれている必要があり、各系列の値は一意でなければなりません。
- **1** つの対象フィールド：モデルごとに対象フィールドを 1 つだけ指定できます。対象フィールドのデータ型は連続型の値でなければなりません。例えば、収入、売上、温度などの数値属性が、時間を経てどのように変化するかを予測できます。ただし、購入状況または学歴のレベルなど、カテゴリー型値を含むフィールドを対象フィールドとして使用できません。
- **1** つ以上の入力フィールド：MS タイム シリーズ アルゴリズムには、少なくとも 1 つの入力フィールドが必要です。入力フィールドのデータ型は連続型の値でなければなりません。連続型以外の入力フィールドは、モデルの作成時に無視されます。
- データセットをソートすること：入力データセットは (キー時間フィールドで) ソートする必要があります。ソートしない場合、エラーが発生してモデル作成が中断します。

MS タイム シリーズのモデル・オプション

モデル名**:** ノードの実行時に作成されたモデルに割り当てられる名前を指定します。

- 自動: モデル名は、対象フィールドまたは ID フィールド名、または、クラスタリング・モデル・ノードなどのように対象が指定されていない場合モデル タイプの名前に基づいて自動的に生成されます。
- カスタム: 作成されたモデルのカスタム名を指定できるようになります。

データ区分データを使用。 データ区分フィールドが定義されている場合、このオプションでは学習用データ区分からのデータのみがモデル構築に使用されるようにします。

ドリルスルーあり：表示された場合、このオプションを選択すると、モデルを問い合わせて、モデルに含まれるケースについての詳細を知ることができます。

一意のフィールド：ドロップダウン・リストから、時系列モデルの作成に使用するキー時間フィールドを選択します。

MS タイム シリーズのエキスパート・オプション

「エキスパート」タブで利用可能なオプションは、選択したストリームの構造に応じて変化する可能性があります。各 Analysis Services モデル・ノードで選択したエキスパート・オプションの詳細は、ユーザー・インターフェースのフィールド レベルのヘルプを参照してください。

過去の予測を行っている場合、スコアリング結果に含めることができる過去のステップ数は、(HISTORIC_MODEL_COUNT * HISTORIC_MODEL_GAP) の値によって決まります。デフォルトでは、この制約は 10 で、過去の予測は 10 件だけ行われます。この場合、モデル・ナゲットの「設定」タブの「過去の予測」に -10 より小さい値を入力すると、エラーが発生します (24 ページの『MS タイム シリーズ モデル・ナゲットの「設定」タブ』 を参照)。過去の予測についてもっと知りたい場合、HISTORIC_MODEL_COUNT または HISTORIC_MODEL_GAP の値を大きくすることができますが、モデルの作成時間が長くなります。

MS タイム シリーズの設定オプション

推定の開始：推定を開始する時間を指定します。

- 開始位置 :新しい予測 :将来の予測を開始する時間で、過去のデータの最後の時間からのオフセットとして表されます。例えば、過去のデータが 12/99 に終了し、予測を 01/00 に開始したい場合、1 の値を使用します。ただし、予測を 03/00 に開始したい場合、3 の値を使用します。
- 開始位置 :過去の予測 :過去の予測を開始する時間で、過去のデータの最後の時間からの負のオフセットとして表されます。例えば、過去のデータが 12/99 で終了し、データの終わりの 5 つの期間の過去の予測を行う場合、-5 の値を使用します。

推定の終了 : 推定を終了する時間を指定します。

- 予測の手順を終了 : 予測を終了する時間で、過去のデータの最後の時間からのオフセットとして表されます。例えば、過去のデータが 12/99 で終了し、予測を 6/00 で終了する場合、ここで 6 の値を使用します。将来の予測の場合、値は 「始点」 の値以上でなければなりません。

MS シーケンス・クラスタリング・ノード

MS シーケンス・クラスタリング・ノードでは、シーケンス分析アルゴリズムを使用して、次のパスまたはシーケンス でリンクできるイベントを含むデータを探索します。例えば、ユーザーが Web サイトをナビゲートまたは参照する場合に作成されるクリックパス、または顧客がオンライン ショップのショッピングカートに品物を追加する順序などとなります。アルゴリズムは、同一のシーケンスをグループ化またはクラスタリング して、最も一般的なシーケンスを検出します。

要件

Microsoft シーケンス・クラスタリング・モデルの要件は次のとおりです。

- **ID** フィールド: Microsoft シーケンス・クラスタリング・アルゴリズムには、シーケンス情報をトランザクション形式で保存する必要があります。このため、各トランザクションを特定する ID フィールドが必要になります。
- **1** つ以上の入力フィールド : アルゴリズムには、少なくとも 1 つの入力フィールドが必要です。
- シーケンス・フィールド : アルゴリズムには、シーケンス ID フィールドも必要です。尺度は連続型でなければなりません。例えば、フィールドがシーケンス内のイベントを特定する限り、Web ページの ID、整数、または文字列を使用できます。各シーケンスにはシーケンス ID を 1 つだけ使用でき、また各モデルにシーケンスを 1 種類だけ使用できます。シーケンス・フィールドは、ID フィールドおよび一意フィールドとは別のフィールドでなければなりません。
- 対象フィールド: 対象フィールドは、シーケンス・クラスタリング・モデルを作成する場合に必要です。
- 一意のフィールド : シーケンス・クラスタリング・モデルには一意にレコードを特定するキー・フィールドが必要です。一意のフィールドは ID フィールドと同じように設定できます。

MS シーケンス・クラスタリングのフィールド・オプション

すべてのモデル作成ノードには、「フィールド」タブがあり、そこからモデルの作成に使用するフィールドを指定できます。

シーケンス・クラスタリング・モデルを作成する前に、対象フィールドや入力フィールドを指定する必要があります。MS シーケンス・クラスタリング・ノードの場合、上乳のデータ型ノードのフィールド情報を使用できません。ここでフィールド設定を指定する必要があります。

ID: リストから ID フィールドを選択します。ID フィールドとして使用できるのは、数値またはシンボル値のフィールドです。選択したフィールドでは、一意の値がそれぞれ、ある分析ユニットを示している必要

があります。例えば、マーケット・バスケット分析なら、各 ID が 1 人の顧客を表します。Web ログ分析なら、各 ID が 1 台のコンピューター (IP アドレス) あるいは 1 人のユーザー (ログイン・データ) を表します。

入力: モデルの入力フィールドを選択してください。これらのフィールドには、シーケンス・モデル作成の対象となるイベントが含まれています。

シーケンス : リストからシーケンス ID フィールドとして使用するフィールドを選択します。例えば、フィールドがシーケンス内のイベントを特定する場合、Web ページの ID、整数、または文字列を使用できます。各シーケンスにはシーケンス ID を 1 つだけ使用でき、また各モデルにシーケンスを 1 種類だけ使用できます。シーケンス・フィールドは ID フィールド (このタブで指定) および一意フィールド (「モデル」タブで指定) とは異なるフィールドでなければなりません。

目標: 対象フィールドとして使用するフィールド、値をシーケンス・データに基づいて予測しようとしているフィールドを選択します。

MS シーケンス・クラスタリングのエキスパート・オプション

「エキスパート」タブで利用可能なオプションは、選択したストリームの構造に応じて変化する可能性があります。各 Analysis Services モデル・ノードで選択したエキスパート・オプションの詳細は、ユーザー・インターフェースのフィールド レベルのヘルプを参照してください。

Analysis Services モデルのスコアリング

モデル・スコアリングが SQL Server 内に発生し Analysis Services で実行されます。データが IBM SPSS Modeler 内にあったり、IBM SPSS Modeler 内でデータを準備する必要があったりする場合は、データ・セットを一時テーブルにアップロードすることが必要な場合があります。データベース内マイニングを使用して IBM SPSS Modeler から生成したモデルは、リモートのデータマイニング・サーバーまたはデータベース・サーバー上に保管されています。これは、Microsoft Analysis Services を使用して作成されたモデルを参照およびスコアリングする場合に、理解すべき重要な違いです。

IBM SPSS Modeler では、通常、単一の予測および関連付けられている確率または確信度のみが提供されます。

モデル・スコアリング の例については、 25 ページの『Analysis Services マイニングの例』 を参照してください。

すべての Analysis Services モデルに共通の設定

次の設定は、すべての Analysis Services モデルに共通です。

Analysis Services モデル・ナゲットの「サーバー」タブ

「サーバー」タブで、データベース内マイニングに対する接続を指定します。このタブには、一意のモデル・キーも用意されています。このキーは、IBM SPSS Modeler のモデル内と Analysis Services データベースに格納されているモデル・オブジェクトの説明内との両方で、モデルが作成され、格納されるときに、ランダムに生成されます。

「サーバー」タブでは Analysis サーバー・ホストとデータベース、スコアリング操作に対する SQL Server データ・ソースを設定します。ここで指定した設定は、IBM SPSS Modeler の「ヘルパー アプリ

ケーション」ダイアログ・ボックス または「モデル作成」ダイアログ・ボックスで指定した設定を上書きします。詳しくは、トピック 13 ページの『Analysis Services との統合を有効にする』を参照してください。

モデル **GUID :** モデル・キーがここで示されます。このキーは、IBM SPSS Modeler のモデル内と Analysis Services データベースに格納されているモデル・オブジェクトの説明内との両方で、モデルが作成され、格納されるときに、ランダムに生成されます。

検査 **:** このボタンをクリックすると、Analysis Services データベースに保存されるモデル内のキーを基準にして、モデル・キーが検査されます。これにより、モデルが Analysis サーバー内に存在し、モデルの構造に変化がないことを示すことを検証できます。

注: 「検査」ボタンはスコアリングに備えてストリーム領域に追加されたモデルにのみ使用できます。検査が失敗した場合、モデルが削除されたか、サーバー上の他のモデルにより置き換えられたかを調べます。

表示**:** クリックすると、ディシジョン・ツリー・モデルがグラフィカルに表示されます。Decision Tree Viewer は IBM SPSS Modeler のすべてのディシジョン・ツリー・アルゴリズムが共有し、その機能は同一です。

Analysis Services モデル・ナゲットの「要約」タブ

モデル・ナゲットの「要約」タブで、モデルそのもの (精度分析)、モデルで使用するフィールド (フィールド)、モデルの構築時に使用する設定 (構築の設定)、およびモデルの学習 (学習の要約) についての情報を表示します。

ノードを初めて参照する場合、「要約」タブの結果は閉じられています。目的の結果を表示するには、項目の左側にある展開コントロールを使用して項目を展開するか、または 「すべて展開」 ボタンをクリックしてすべての結果を表示します。見終わった結果を隠すには、展開コントロールを使用して目的の結果を省略するか、または「すべて閉じる」ボタンをクリックしてすべての結果を非表示にします。

精度分析 **:** 特定のモデルについての情報を表示します。このモデル・ナゲットに接続されている精度分析ノードを実行した場合、その精度分析情報もこのセクションに表示されます。

フィールド**:** 対象フィールドおよびモデル構築時の入力として使われるフィールドが表示されます。

構築の設定 **:** モデル構築時に使われる設定情報が表示されます。

学習の要約 **:** モデルの種類、モデルの作成に使われたストリーム、モデルの作成者、モデルの作成日時、およびモデルの構築時間などの情報が表示されます。

MS タイム シリーズ モデル・ナゲット

MS タイム シリーズ モデルは、過去のデータではなく、予測された期間のみのスコアを生成します。

モデルに追加されるフィールドを次の表に示します。

表 *1.* モデルに追加されるフィールド

フィールド名	説明
$M-*field*	フィールドの予測値
$Var-*field*	計算されたフィールドの分散
$Stdev-*field*	フィールドの標準偏差

MS タイム シリーズ モデル・ナゲットの「サーバー」タブ

「サーバー」タブで、データベース内マイニングに対する接続を指定します。このタブには、一意のモデル・キーも用意されています。このキーは、IBM SPSS Modeler のモデル内と Analysis Services データベースに格納されているモデル・オブジェクトの説明内との両方で、モデルが作成され、格納されるときに、ランダムに生成されます。

「サーバー」タブでは Analysis サーバー・ホストとデータベース、スコアリング操作に対する SQL Server データ・ソースを設定します。ここで指定した設定は、IBM SPSS Modeler の「ヘルパー アプリケーション」ダイアログ・ボックス または「モデル作成」ダイアログ・ボックスで指定した設定を上書きします。詳しくは、トピック 13 ページの『Analysis Services との統合を有効にする』を参照してください。

モデル **GUID :** モデル・キーがここで示されます。このキーは、IBM SPSS Modeler のモデル内と Analysis Services データベースに格納されているモデル・オブジェクトの説明内との両方で、モデルが作成され、格納されるときに、ランダムに生成されます。

検査 **:** このボタンをクリックすると、Analysis Services データベースに保存されるモデル内のキーを基準にして、モデル・キーが検査されます。これにより、モデルが Analysis サーバー内に存在し、モデルの構造に変化がないことを示すことを検証できます。

注: 「検査」ボタンはスコアリングに備えてストリーム領域に追加されたモデルにのみ使用できます。検査が失敗した場合、モデルが削除されたか、サーバー上の他のモデルにより置き換えられたかを調べます。

表示**:** クリックすると、時系列モデルがグラフィカルに表示されます。Analysis Services は、完成したモデルをツリーで表示します。対象フィールドの時を経た過去の値と、将来の予測値をともに表示するグラフとを表示することができます。

詳細は、http://msdn.microsoft.com/en-us/library/ms175331.aspx の MSDN ライブラリーのタイム シリーズ・ビューアーの説明を参照してください。

MS タイム シリーズ モデル・ナゲットの「設定」タブ

推定の開始 **:** 推定を開始する時間を指定します。

- 開始位置 **:**新しい予測 **:**将来の予測を開始する時間で、過去のデータの最後の時間からのオフセットとして表されます。例えば、過去のデータが 12/99 に終了し、予測を 01/00 に開始したい場合、1 の値を使用します。ただし、予測を 03/00 に開始したい場合、3 の値を使用します。
- 開始位置 **:**過去の予測 **:**過去の予測を開始する時間で、過去のデータの最後の時間からの負のオフセットとして表されます。例えば、過去のデータが 12/99 で終了し、データの終わりの 5 つの期間の過去の予測を行う場合、-5 の値を使用します。

推定の終了 **:** 推定を終了する時間を指定します。

- 予測の手順を終了 **:** 予測を終了する時間で、過去のデータの最後の時間からのオフセットとして表されます。例えば、過去のデータが 12/99 で終了し、予測を 6/00 で終了する場合、ここで 6 の値を使用します。将来の予測の場合、値は 「始点」 の値以上でなければなりません。

MS シーケンス・クラスタリングのモデル・ナゲット

MS シーケンス・クラスタリング・モデルに追加されるフィールドを次の表に示します (ここで、フィールド とは対象フィールドの名前です)。

表 2. モデルに追加されるフィールド

フィールド名	説明
$MC-*field*	シーケンスが属するクラスターの予測。
$MCP-*field*	このシーケンスが予測クラスターに含まれる確率。
$MS-*field*	フィールドの予測値
$MSP-*field*	$MS-*field* 値が適切である確率。

モデルのエクスポートとノードの生成

モデル集計をエクスポートして、テキストおよび HTML 形式のファイルに構造化できます。適切な場所に、適切な条件抽出ノード、フィルター・ノードを生成できます。

IBM SPSS Modeler のモデル・ナゲットと同様、Microsoft Analysis Services モデル・ナゲットはレコードの直接生成とフィールド操作ノードをサポートします。モデル・ナゲットの「ノードの生成」メニューのオプションを使用して、次のノードを生成できます。

- 条件抽出ノード (項目が「モデル」タブで選択された場合のみ)
- フィルター・ノード

Analysis Services マイニングの例

いくつものサンプル・ストリームがあり、それらが IBM SPSS Modeler を使用して MS Analysis Services データ・マイニング を行う方法を説明します。これらのストリームは、次の IBM SPSS Modeler のインストール・フォルダーにあります。

¥Demos¥Database_Modelling¥Microsoft

注 : Demos フォルダーには、Windows の「スタート」メニューの IBM SPSS Modeler プログラム・グループからアクセスできます。

ストリームの例 :ディシジョン・ツリー

次のストリームは、MS Analysis Services が提供するディシジョン・ツリー・アルゴリズムを使用するデータベース・マイニング・プロセスの例として、順番に使用されます。

表 3. デシジョン・ツリー - ストリーム例

ストリーム	説明
1_upload_data.str	フラット・ファイルのデータを整理して、データベースへアップロードするために使用されます。
2_explore_data.str	IBM SPSS Modeler でのデータ探索の例として使用されます。
3_build_model.str	データベース固有のアルゴリズムを使用したモデルを構築します。
4_evaluate_model.str	IBM SPSS Modeler でのモデル評価の例として使用されます。
5_deploy_model.str	データベース内スコアリングのためにモデルを展開します。

注 : サンプルを実行するには、ストリームを順番に実行する必要があります。さらに、各ストリーム中の入力およびモデル作成ノードは、使用するデータベースの有効なデータ・ソースを参照するように更新する必要があります。

サンプル・ストリームで使用されるデータセットは、クレジット カード申請に関するものであり、カテゴリー型および連続型予測フィールドの混在について、分類上の問題を提示します。このデータ・セットの詳細は、サンプル・ストリームと同じフォルダーにある *crx.names* ファイルを参照してください。

このデータセットは、*ftp://ftp.ics.uci.edu/pub/machine-learning-databases/credit-screening/* にある UCI Machine Learning Repository から入手可能です。

ストリームの例 :データのアップロード

最初のストリーム例、*1_upload_data.str* は、フラット・ファイルのデータを整理して SQL Server へアップロードするのに使用されます。

Analysis Services データ・マイニングは、キー・フィールドを必要とするため、この初期ストリームは、フィールド作成ノードを使用して、一意の値 *1*、*2*、*3* を持つ *KEY* という名前の新しいフィールドをデータ・セットに追加します。これには、IBM SPSS Modeler の `@INDEX` 関数が使用されます。

その次にある置換ノードは、欠損値の処理に使用され、*crx.data* テキスト・ファイルから読み込まれた空のフィールドを*NULL* 値で置き換えます。

ストリームの例 :データの調査

2 番目の例のストリーム、*2_explore_data.str* を使用して、要約統計およびグラフなど、データの概要を取得するデータ検査ノードの使用方法を説明します。

データ検査レポート内のグラフをダブルクリックすると、指定されたフィールドについて、より深く検索した結果を表す詳細なグラフが生成されます。

ストリームの例 :モデルの作成

3 番目のストリーム例、*3_build_model.str* では、IBM SPSS Modeler でのモデル構築を説明します。ストリームにデータベース・モデルを追加したら、追加したモデルをダブルクリックして構築の設定を指定できます。

このダイアログ・ボックスの「モデル」タブでは、次の設定ができます。

1. 「キー」 フィールドを一意の ID フィールドとして指定します。

「エキスパート」タブでは、モデル構築の設定を細かく指定できます。

実行する前に、モデル構築用に正しいデータベースが指定されていることを確認してください。「サーバー」タブを使用すると、設定を変更できます。

ストリームの例 :モデルの評価

第 4 のストリーム例、*4_evaluate_model.str* では、IBM SPSS Modeler を使用したデータベース内モデル作成の利点を説明します。モデルを実行すると、その結果をユーザーのデータストリームに追加したり、IBM SPSS Modeler が提供するいくつかのツールを使用してモデルを評価できます。

モデル作成結果の表示

モデル・ナゲットをダブルクリックすると、結果を確認できます。「要約」タブには、結果を表示するルールツリー・ビューがあります。また、「サーバー」タブにある 「ビュー」 ボタンをクリックして、ディシジョン・ツリー・モデルのグラフ表示を行うことができます。

モデル作成結果の評価

サンプル・ストリームの分析ノードを使用すると、予測フィールドとその対象フィールド間の一致パターンを表す一致行列を作成できます。結果を表示するには分析ノードを実行します。

サンプル・ストリームの評価ノードを使用して、このモデルによりどの程度、精度が改善されたかを示すゲイン・グラフを作成できます。結果を表示するには評価ノードを実行します。

ストリームの例 :モデルの展開

モデルの精度が満足できるものになったら、外部のアプリケーションと共に使用するために展開したり、データベースに保存するために発行できます。最後のストリームの例、*5_deploy_model.str* では、データは表 CREDIT から読み込まれ、スコアリングされてから、データベース・エクスポート・ノードを使用して、表 CREDITSCORES に発行されます。

ストリームを実行すると次の SQL が生成されます。

```
DROP TABLE CREDITSCORES

CREATE TABLE CREDITSCORES ( "field1" varchar(1),"field2" varchar(255),"field3" f
loat,"field4" varchar(1),"field5" varchar(2),"field6" varchar(2),"field7" varcha
r(2),"field8" float,"field9" varchar(1),"field10" varchar(1),"field11" int,"fiel
d12" varchar(1),"field13" varchar(1),"field14" int,"field15" int,"field16" varch
ar(1),"KEY" int,"$M-field16" varchar(9),"$MC-field16" float )

INSERT INTO CREDITSCORES ("field1","field2","field3","field4","field5","field6","field7","field8",
"field9","field10","field11","field12","field13","field14","field15","field16",
"KEY","$M-field16","$MC-field16")
SELECT T0.C0 AS C0,T0.C1 AS C1,T0.C2 AS C2,T0.C3 AS C3,T0.C4 AS C4,T0.C5 AS C5,
 T0.C6 AS C6,T0.C7 AS C7,T0.C8 AS C8,T0.C9 AS C9,T0.C10 AS C10,
 T0.C11 AS C11,T0.C12 AS C12,T0.C13 AS C13,T0.C14 AS C14,
 T0.C15 AS C15,T0.C16 AS C16,T0.C17 AS C17,T0.C18 AS C18
FROM (
 SELECT CONVERT(NVARCHAR,[TA].[field1])  AS C0, CONVERT(NVARCHAR,[TA].[field2])  AS C1,
  [TA].[field3] AS C2, CONVERT(NVARCHAR,[TA].[field4])  AS C3,
  CONVERT(NVARCHAR,[TA].[field5])  AS C4, CONVERT(NVARCHAR,[TA].[field6]) AS C5,
  CONVERT(NVARCHAR,[TA].[field7])  AS C6, [TA].[field8] AS C7,
  CONVERT(NVARCHAR,[TA].[field9])  AS C8, CONVERT(NVARCHAR,[TA].[field10])  AS C9,
  [TA].[field11] AS C10, CONVERT(NVARCHAR,[TA].[field12])  AS C11,
  CONVERT(NVARCHAR,[TA].[field13])  AS C12, [TA].[field14] AS C13,
  [TA].[field15] AS C14, CONVERT(NVARCHAR,[TA].[field16])  AS C15,
  [TA].[KEY] AS C16, CONVERT(NVARCHAR,[TA].[$M-field16])  AS C17,
  [TA].[$MC-field16] AS C18
 FROM openrowset('MSOLAP',
  'Datasource=localhost;Initial catalog=FoodMart 2000',
  'SELECT [T].[C0] AS [field1],[T].[C1] AS [field2],[T].[C2] AS [field3],
   [T].[C3] AS [field4],[T].[C4] AS [field5],[T].[C5] AS [field6],
   [T].[C6] AS [field7],[T].[C7] AS [field8],[T].[C8] AS [field9],
   [T].[C9] AS [field10],[T].[C10] AS [field11],[T].[C11] AS [field12],
   [T].[C12] AS [field13],[T].[C13] AS [field14],[T].[C14] AS [field15],
   [T].[C15] AS [field16],[T].[C16] AS [KEY],[CREDIT1].[field16] AS [$M-field16],
   PredictProbability([CREDIT1].[field16]) AS [$MC-field16]
  FROM [CREDIT1] PREDICTION JOIN
   openrowset(''MSDASQL'',
   ''Dsn=LocalServer;Uid=;pwd='','''SELECT T0."field1" AS C0,T0."field2" AS C1,
   T0."field3" AS C2,T0."field4" AS C3,T0."field5" AS C4,T0."field6" AS C5,
   T0."field7" AS C6,T0."field8" AS C7,T0."field9" AS C8,T0."field10" AS C9,
   T0."field11" AS C10,T0."field12" AS C11,T0."field13" AS C12,
   T0."field14" AS C13,T0."field15" AS C14,T0."field16" AS C15,
   T0."KEY" AS C16 FROM "dbo".CREDITDATA T0'') AS [T]
  ON [T].[C2] = [CREDIT1].[field3] and [T].[C7] = [CREDIT1].[field8]
   and [T].[C8] = [CREDIT1].[field9] and [T].[C9] = [CREDIT1].[field10]
   and [T].[C10] = [CREDIT1].[field11] and [T].[C11] = [CREDIT1].[field12]
   and [T].[C14] = [CREDIT1].[field15]') AS [TA]
)  T0
```

第 4 章 Oracle Data Mining によるデータベース・モデリング

Oracle Data Mining について

IBM SPSS Modeler は、Oracle Data Mining (ODM) との統合をサポートします。このリリースでは、Oracle RDBMS 内にデータ・マイニング・アルゴリズム・ファミリーが密接に組み込まれています。これらの機能は、IBM SPSS Modeler のグラフィカル・ユーザー・インターフェースとワークフロー指向の開発環境でアクセスでき、顧客は ODM が提供するデータ・マイニング・アルゴリズムを使用できます。

IBM SPSS Modeler は、Oracle Data Mining の次のアルゴリズムを統合できます。

- Naive Bayes
- Adaptive Bayes
- Support Vector Machine (SVM)
- 一般化線型モデル (GLM)*
- デシジョン ツリー
- O-Cluster
- K-means
- 非負数マトリックス因数分解 (NMF)
- Apriori
- Minimum Descriptor Length (MDL)
- Attribute Importance (AI)

* 11*g* R1 のみ

Oracle との統合の要件

Oracle Data Mining を使用してデータベース内のモデル作成を実行する場合、以下の条件が前提条件となります。場合によっては、これらの条件が満たされているかデータベース管理者に問い合わせて確認してください。

- ローカル・モードで動作している IBM SPSS Modeler、あるいは Windows か UNIX 上の IBM SPSS Modeler Server インストール済み環境で動作している IBM SPSS Modeler Server。
- Oracle データ・マイニング・オプションを搭載した Oracle 10*g*R2 または 11*g*R1 (10.2 データベース以降)。

注 : 10*g*R2 は、一般化線型モデル (11*g*R1 が必要) 以外のデータベース モデル作成アルゴリズムをサポートします。

- 次に説明するように、Oracle への接続に使用する ODBC データ・ソース。

注 : データベース・モデル作成および SQL 最適化では、IBM SPSS Modeler Server 接続が IBM SPSS Modeler コンピューター上で可能でなければなりません。この設定を有効にすると、データベース・アルゴリズムにアクセスし、IBM SPSS Modeler から SQL を直接プッシュバック、IBM SPSS Modeler Server にアクセスできます。現在のライセンス ステータスを確認するには、IBM SPSS Modeler メニューから次を選択します。

「ヘルプ」 > 「バージョン情報」 > 「その他の詳細」

接続が有効な場合、「ライセンス ステータス」タブにオプション 「サーバーの有効化」 が表示されます。

Oracle との統合を有効にする

Oracle Data Mining との IBM SPSS Modeler の統合を有効にするには、Oracle を設定し、ODBC ソースを作成して、IBM SPSS Modeler の「ヘルパー アプリケーション」ダイアログ・ボックスで統合を有効にする必要があります。さらに、SQL の生成と最適化を有効にします。

Oracle の設定

Oracle Data Mining をインストールして構成するには、Oracle の資料 (特に「*Oracle Administrator's Guide*」を参照してください。

Oracle の ODBC ソースの作成

Oracle と IBM SPSS Modeler の接続を有効にするには、ODBC システム・データ・ソース名 (DSN) を作成する必要があります。

DSN を作成する前に、ODBC データ ソースおよび ODBC ドライバーの基礎、さらに IBM SPSS Modeler のデータベース サポートの基礎を理解する必要があります。

IBM SPSS Modeler Server に対して、分散モードで実行している場合、サーバ コンピューターに DSN を作成します。ローカル (クライアント) モードで実行している場合、クライアント・コンピューターに DSN を作成します。

1. ODBC ドライバーをインストールします。これらのドライバーは、このリリースに付属する IBM SPSS Data Access Pack インストール ディスクにあります。*setup.exe* ファイルを実行してインストーラーを起動し、関連するドライバーをすべて選択します。画面上の指示に従って、ドライバーをインストールします。
   a. DSN を作成します。

      注 :メニュー・シーケンスは使用する Windows のバージョンによって異なります。

      - **Windows XP** の場合。「スタート」メニューから、「コントロール パネル」を選択します。「管理ツール」 をダブルクリックし、次に 「データ ソース **(ODBC)**」 をダブルクリックします。
      - **Windows Vista** :「スタート」メニューから、「コントロール パネル」 → 「システム メンテナンス」 を選択します。「管理ツール」 をダブルクリックし、次に 「データ ソース **(ODBC)**」 を選択して 「開く」 をクリックします。
      - **Windows 7 :**「スタート」メニューから、「コントロール パネル」 → 「システムとセキュリティー」 → 「管理ツール」 を選択します。「データ ソース **(ODBC)**」 を選択して 「開く」 をクリックします。

   b. 「システム **DSN**」 タブをクリックしてから、「追加」 をクリックします。
2. **SPSS OEM 6.0 Oracle Wire Protocol** ドライバーを選択します。
3. 「終了」をクリックします。

4. 「ODBC Oracle Wire Protocol Driver セットアップ」画面では、選択したデータ・ソースの名前、Oracle サーバのホスト名、接続用のポート番号、および使用する Oracle インスタンスの SID を入力します。

   tnsnames.ora ファイルで TNS を実装した場合、ホスト名、ポート、および SID は、サーバ マシンにある *tnsnames.ora* ファイルから取得できます。詳細は Oracle 管理者までお問い合わせください。
5. 接続をテストするには、「テスト」ボタンをクリックします。

IBM SPSS Modeler での Oracle Data Mining 統合の有効化

1. IBM SPSS Modeler のメニューから次の項目を選択します。

   「ツール」 > 「オプション」 > 「ヘルパー アプリケーション」
2. 「**Oracle**」タブをクリックします。

Oracle Data Mining 統合を有効化**:** IBM SPSS Modeler ウィンドウの下部でデータベース モデリング パレットを有効にし (まだ表示されていない場合)、Oracle Data Mining アルゴリズムのノードを追加します。

Oracle 接続 **:** 有効なユーザ名とパスワードと共に、モデルの作成と格納に使用するデフォルトの Oracle ODBC データ・ソースを指定します。この設定を個々のモデル作成ノードおよびモデル・ナゲットでオーバーライドすることができます。

注: モデル作成の目的で使用するデータベース接続は、データアクセスに使用する接続と同じであっても、異なっていても問題ありません。例えば、特定の Oracle データベースのデータにアクセスし、クリーニングまたはその他の操作のためにデータを IBM SPSS Modeler にダウンロードし、次にモデル作成を目的としてデータを別の Oracle データベースにアップロードするストリームを利用できます。あるいは、元のデータをフラット・ファイルや他の (Oracle 以外の) ソースに格納できます。この場合、モデルを作成する際に、データを Oracle にアップロードする必要があります。いずれの場合でも、データはモデル作成用のデータベース内で作成された一時テーブルに自動的にアップロードされます。

Oracle データ・マイニング・モデル上書き時に警告する **:**このオプションを選択して、データベースに格納されているモデルが警告なしに IBM SPSS Modeler で上書きされないことを確認します。

Oracle Data Mining モデルを一覧表示する **:** 利用できるデータ・マイニング・モデルを表示します。

Oracle Data Miner の起動の有効化 **:** 有効にすると、IBM SPSS Modeler は Oracle Data Miner アプリケーションを起動できるようになります。詳しくは、 49 ページの『Oracle Data Miner』を参照してください。

Oracle Data Miner 実行可能ファイルのパス**: (**オプション**)** Windows 版 Oracle Data Miner の実行可能ファイルの物理的な位置を指定します (*C:¥odm¥bin¥odminerw.exe* など)。Oracle Data Miner は、IBM SPSS Modeler とともにインストールされません。Oracle の Web サイト (*http://www.oracle.com/technology/products/bi/odm/odminer.html*) から正しいバージョンをダウンロードし、クライアントでインストールする必要があります。

SQL の生成と最適化を有効にする

1. IBM SPSS Modeler のメニューから次の項目を選択します。

   「ツール」 > 「ストリームのプロパティー」 > 「オプション」
2. ナビゲーション ペインの 「最適化」 オプションをクリックします。

3. 「**SQL** 生成」 オプションが有効になっていること確認します。この設定は、データベースのモデル作成が機能するために必要です。
4. 「**SQL** 生成の最適化」 と 「その他の実行を最適化」 を選択します (絶対に必要な訳ではありませんが、最適化されたパフォーマンスを得るために、選択することを強くお勧めします)。

Oracle Data Mining を使用してモデルを構築する

多少の例外はありますが、Oracle モデル作成ノードは、IBM SPSS Modeler の他のモデル作成ノードと同様に動作します。このノードは IBM SPSS Modeler ウィンドウの下部にある「データベース モデル作成」パレットからアクセスできます。

データの考慮事項

Oracle では、カテゴリー・データを文字列形式 (CHAR または VARCHAR2) で格納する必要があります。したがって、IBM SPSS Modeler では、記憶域のフィールドを、ODM モデルの入力値として 「フラグ型」 または 「名義型」 (カテゴリー) が測定の尺度の数値のストレージ・フィールドを指定することは許されません。IBM SPSS Modeler でデータ分類ノードを使用すると、必要に応じて、数値を文字列に変換できます。

対象フィールド**:** ODM 分類モデルでは、出力 (対象) フィールドとして、1 つのフィールドのみを選択できます。

モデル名**:** Oracle 11*g*R1 以降、名前 `unique` はキーワードで、カスタム・モデル名として使用できません。

一意のフィールド **:** 各ケースを一意に識別するためにフィールドを指定します。例えば、これは *CustomerID* などの ID フィールドです。IBM SPSS Modeler には制限があり、このキー・フィールドは数値である必要があります。

注 :このフィールドは、Oracle Adaptive Bayes、Oracle O-Cluster および Oracle Apriori を除くすべての Oracle ノードについてオプションです。

一般的なコメント

- Oracle Data Mining で生成するモデルの場合、IBM SPSS Modeler からは PMML のエクスポート/インポートは提供しません。
- モデル・スコアリングは、常に ODM で実行されます。データが IBM SPSS Modeler 内で発生する (または、データを IBM SPSS Modeler 内で準備する必要がある) 場合は、データ・セットを一時テーブルにアップロードする必要がある場合があります。
- IBM SPSS Modeler では、通常、単一の予測および関連付けられている確率または確信度のみが提供されます。
- IBM SPSS Modeler では、モデル作成とスコアリングに使用できるフィールド数を 1,000 に制限しています。
- IBM SPSS Modeler Solution Publisher を使用して実行用に公開したストリームでは、IBM SPSS Modeler で ODM モデルをスコアリングできます。

Oracle モデルの「サーバー」オプション

モデル作成用データのアップロードに使用する Oracle 接続を指定します。「ヘルパー アプリケーション」ダイアログ・ボックスで指定されたデフォルトの Oracle 接続を上書きするために、必要に応じて、

「サーバー」タブで各モデル作成ノードに対して接続を選択できます。 詳しくは、トピック 30 ページの『Oracle との統合を有効にする』を参照してください。

コメント

- モデル作成に使用する接続は、ストリームの入力ノードで使用する接続と同じであっても別であってもかまいません。例えば、特定の Oracle データベースのデータにアクセスし、クリーニングまたはその他の操作のためにデータを IBM SPSS Modeler にダウンロードし、次にモデル作成を目的としてデータを別の Oracle データベースにアップロードするストリームを利用できます。
- ODBC データ・ソース名は、各 IBM SPSS Modeler ストリームに効果的に埋め込まれます。あるホスト上で作成されたストリームが別のホスト上で実行された場合、データ・ソースの名前はそれぞれのホストで同じである必要があります。また、各入力ノードまたはモデル作成ノードで、「サーバー」タブから異なるデータ・ソースを選択できます。

誤分類コスト

状況によっては、特定の誤りコストが他の誤りコストに比べて高いことがあります。例えば、信用リスクの高い申請者を低リスクに分類した場合 (ある種の誤分類) のコストは、リスクの低い申請者を高リスクに分類した場合 (別種の誤分類) よりも高くなります。誤分類コストでは、さまざまな予測の誤りに対し、相対的な重要度を指定できます。

誤分類コストは、基本的には、特定の結果に対して適用される重みです。これらの重みは、モデルに組み込まれ、(コストの高い誤りを防ぐための手段として) 実際に予測値に影響する場合があります。

C5.0 モデルを例外として、誤分類コストは、モデルのスコアリング時には適用されず、自動分類ノード、評価グラフ、または分析ノードを使用してモデルをランク付けまたは比較する場合には考慮されません。コストを含むモデルは、コストを含まないモデルに比べてエラーが少なく、全体の精度の項目で高くランク付けされません。ただし、コストが少ない エラーにより組み込まれたバイアスがあるため、実際の問題でパフォーマンスが優れる場合があります。

コスト行列には、予測カテゴリーと実際のカテゴリーの可能な組み合わせごとのコストが表示されます。デフォルトでは、すべての誤分類コストが 1.0 に設定されています。コストの値を自分で入力するには、「誤分類コストを使用」 を選択して、コスト行列に独自の値を入力します。

誤分類コストを変更するには、目的の予測値と実際の値の組み合わせに対応するセルを選択して、セルの内容を削除してから、適切なコストを入力してください。コストは自動的に対称的にはなりません。例えば *A* を *B* として誤分類した場合のコストを 2.0 に設定しても、*B* を *A* として誤分類した場合のコストは、変更しない限りデフォルト値 (1.0) のまま変わりません。

注: ディシジョン・ツリー・モデルのみで構築時のコストを指定できます。

Oracle Naive Bayes

Naive Bayes は、分類の問題に対応する有名なアルゴリズムです。提示されたすべての予測変数は相互に依存関係がないものとして処理されるので、モデルは *naïve* と命名されます。Naive Bayes は拡張性のある高速のアルゴリズムであり、複数の属性と対象属性の組み合わせに対して、条件付きの確率を計算します。学習データから、個別の確率が計算されます。各入力変数の各値カテゴリーを計算単位とすると、この確率は各対象クラスの確率を表します。

- モデル作成に使用したデータで、モデルの精度をテストする場合、交差検証が使用されます。この検証方法は、モデル作成に利用できるケース数が少ないときに特に便利です。

- モデル出力は、行列形式で参照できます。行列の数値は、条件に関連する確率であり、予測クラス (列) と予測変数－値の組み合わせ (行) を関連付けています。

Naive Bayes の「モデル」オプション

モデル名**:** 対象または ID フィールド (そのようなフィールドが指定されていない場合はモデル タイプ) に基づいてモデル名を自動的に生成するか、カスタム名を指定することができます。

データ区分データを使用。 データ区分フィールドが定義されている場合、このオプションでは学習用データ区分からのデータのみがモデル構築に使用されるようにします。

一意のフィールド **:** 各ケースを一意に識別するためにフィールドを指定します。例えば、これは *CustomerID* などの ID フィールドです。IBM SPSS Modeler には制限があり、このキー・フィールドは数値である必要があります。

注 :このフィールドは、Oracle Adaptive Bayes、Oracle O-Cluster および Oracle Apriori を除くすべての Oracle ノードについてオプションです。

データの自動準備 **:** (11*g* のみ) Oracle Data Mining の自動データ準備モードを有効化 (デフォルト) または無効化します。このボックスがチェックされている場合、ODM は、アルゴリズムに必要なデータ変換を自動的に実行します。詳細は、『*Oracle Data Mining Concepts* 』を参照してください。

Naive Bayes の「エキスパート」オプション

モデルの作成時に、特定の値または値のペアが学習データ内に十分にない場合、個々の予測属性値または値のペアは無視されます。値を無視する閾値は、学習データ内のレコード数に基づいて分数で指定します。これらの閾値を調整すると、ノイズを減らして、他のデータ・セットに一般化するモデルの能力を改善できます。

- 単一型閾値 **:** 特定の予測属性値に閾値を指定します。特定の値の発生回数が指定した分数以上にならない場合、この値は無視されます。
- ペア単位閾値 **:** 特定の属性と予測値のペアに、閾値を指定します。特定の値ペアの発生回数は、指定した分数以上になる必要があります。下回った場合、ペアは無視されます。

予測確率 **:** モデルに、対象フィールドの出力に対する適切な予測の確率を含めることができます。この機能を有効にするには、「選択」 を選択して 「指定」 ボタンをクリックし、可能性のある出力のいずれかを選択してから、「挿入」 をクリックします。

予測セットを使用 **:** 対象フィールドの可能性のあるすべての出力に対する可能性のあるすべての結果の表を作成します。

Oracle Adaptive Bayes

Adaptive Bayes Network (ABN) は、最小記述長 (MDL) と自動機能選択を使用して、Bayesian Network 分類子を構築します。ABN は、Naive Bayes が適さないような環境でも適切に機能し、また他の多くの環境でも少なくとも同程度の機能を発揮します。ただし、パフォーマンスが低下する可能性はあります。
ABN アルゴリズムでは、簡素化されたデシジョン ツリー (単一機能) モデル、剪定された Naive Bayes モデル、ブーストされた複数機能モデルという、高度な 3 種類の Bayesian ベース モデルを作成できます。

注**:** Oracle Adaptive Bayes アルゴリズムは Oracle 12C で削除されたため、Oracle 12C を使用する場合の IBM SPSS Modeler ではサポートされていません。 http://docs.oracle.com/database/121/DMPRG/release_changes.htm#DMPRG726を参照してください。

生成されたモデル

単一機能の作成モードの場合、ABN は人が判読できるルール・セットに基づいて、簡素化されたデジション・ツリーを作成します。ビジネス・ユーザーまたはアナリストは、このルール・セットを参照して、モデル予測の基本原理を理解できます。また、適切に対応することや、第三者に論理的に説明することができます。この特長は、Naive Bayes や複数機能モデルに対して、大きな利点になっています。IBM SPSS Modeler では、標準的なルール・セットと同様に、これらのルールを参照できます。単純なルール・セットは、次のようになります。

```
IF MARITAL_STATUS = "Married"
AND EDUCATION_NUM = "13-16"
THEN CHURN= "TRUE"
Confidence = .78, Support = 570 cases
```

剪定された Naive Bayes と複数機能モデルは、IBM SPSS Modeler では参照できません。

Adaptive Bayes の「モデル」オプション

モデル名**:** 対象または ID フィールド (そのようなフィールドが指定されていない場合はモデル タイプ) に基づいてモデル名を自動的に生成するか、カスタム名を指定することができます。

データ区分データを使用。 データ区分フィールドが定義されている場合、このオプションでは学習用データ区分からのデータのみがモデル構築に使用されるようにします。

一意のフィールド **:** 各ケースを一意に識別するためにフィールドを指定します。例えば、これは *CustomerID* などの ID フィールドです。IBM SPSS Modeler には制限があり、このキー・フィールドは数値である必要があります。

注 :このフィールドは、Oracle Adaptive Bayes、Oracle O-Cluster および Oracle Apriori を除くすべての Oracle ノードについてオプションです。

モデル タイプ

モデル作成のために、3 種類のモードから適切なモードを選択できます。

- 複数機能 **:** NB モデル、単一機能 / 複数機能の確率モデルを含めて、複数のモデルを作成して、比較してください。単一機能モデルは、最も負担が大きいモードであり、通常、結果を出すまでに最も時間がかかります。ルールが生成されるのは、単一機能モデルが最善であると判断された場合のみです。複数機能または NB モデルが選択されると、ルールは生成されません。
- 単一機能 **:** ルール・セットに基づいて、簡素化されたデジション・ツリーを作成します。各ルールには、条件とそれぞれの結果に関連付けられた確率が指定されています。ルールは相互に排他的であり、人が判読できる形式で提供されます。この特長は、Naive Bayes や複数機能モデルに対して大きな利点になっています。
- **Naive Bayes :** 単一の NB モデルを作成し、以前のグローバル・サンプル (グローバル・サンプル内の対象値の棒グラフ) と比較します。NB モデルが出力として生成されるのは、事前のグローバル・サンプルよりも、NB モデルの方が、対象値の予測に適していると判断された場合に限られます。それ以外の場合は、モデルは出力として生成されません。

Adaptive Bayes の「エキスパート」オプション

実行時間を制限する：最大の作成時間を分で指定するには、このオプションを選択します。このオプションを選択すると、作成したモデルの精度が落ちる可能性がありますが、モデルの作成時間を減らすことができます。アルゴリズムは、モデル作成プロセスの各確認ポイントで、指定された時間内に次の確認ポイントまで完了できるかどうかをチェックし、プロセス続行の可否を判断します。そして、指定された制限を越える場合は、利用できる最適なモデルを返します。

最大予測値：このオプションにより、使用する予測値の数を制限することによって、モデルの複雑化を予防したりパフォーマンスを改善したりできます。予測値は、MDL 法でランク付けされます。この方法は、モデルに含まれている確率を測定する手段として、対象に対する相関関係を示します。

最大 **Naive Bayes**予測値：このオプションで、Naive Bayes モデルで使用する予測値の最大数を指定します。

Oracle Support Vector Machine (SVM)

サポート・ベクター・マシン (SVM) は、マシン学習理論を使用して、データをオーバーフィットすることなしに最高の予測精度を得られるようにする、分類および回帰アルゴリズムです。SVM は、オプションの学習用データの非線形変換を使用し、次にその変換されたデータ中で回帰式を検索して、クラスの分離 (カテゴリー対象の場合)、または対象'l のフィッティング (連続型対象の場合) を行います。Oracle が実装した SVM では、2 つの使用可能なカーネル、線型またはガウスのいずれかを使用して、モデルを作成できます。線型カーネルは、非線型変換をまったく行わないので、生成されたモデルは、本質的に線型回帰モデルです。

詳細は、『*Oracle Data Mining Application Developer's Guide* 』 および 『*Oracle Data Mining Concepts* 』を参照してください。

Oracle SVM の「モデル」オプション

モデル名**:** 対象または ID フィールド (そのようなフィールドが指定されていない場合はモデル タイプ) に基づいてモデル名を自動的に生成するか、カスタム名を指定することができます。

一意のフィールド：各ケースを一意に識別するためにフィールドを指定します。例えば、これは *CustomerID* などの ID フィールドです。IBM SPSS Modeler には制限があり、このキー・フィールドは数値である必要があります。

注 :このフィールドは、Oracle Adaptive Bayes、Oracle O-Cluster および Oracle Apriori を除くすべての Oracle ノードについてオプションです。

データの自動準備：(11*g* のみ) Oracle Data Mining の自動データ準備モードを有効化 (デフォルト) または無効化します。このボックスがチェックされている場合、ODM は、アルゴリズムに必要なデータ変換を自動的に実行します。詳細は、『*Oracle Data Mining Concepts* 』を参照してください。

アクティブな学習：大きな構築セットを処理する方法を示します。アクティブな学習により、アルゴリズムは、完全な学習データセットに適用する前に、 小さなサンプルに基づいて初期モデルを作成し、その結果に基づいてサンプルとモデルを徐々に更新します。モデルが学習データに集束するまで、またはサポート・ベクトルの最大許容数に達するまで、このサイクルは繰り返されます。

カーネル関数**:** 「線型」または「ガウス」を選択、またはデフォルト「決定システム」のままで、システムが最も適切なカーネルを選択できるようにします。ガウス・カーネルは、より複雑な関係を学習できますが、一般に計算する時間も長くなります。まず、線型カーネルを使用して、もし線型カーネルが良い適合度

を発見できなかった場合にのみガウス・カーネルを使用することもできます。これは、もっぱら回帰モデルに当てはまります。回帰モデルではカーネルの選択がより重要になるからです。また、ガウス・モデルを使用して構築された SVM は IBM SPSS Modeler で表示できないことにも注意してください。線型カーネルを使用して構築されたモデルは、標準の回帰モデルと同じ方法で IBM SPSS Modeler でブラウズできます。

正規化**:** 連続型入力フィールドと対象フィールドの正規化方法を指定します。「**Z-**スコア」、「最小**-**最大」、または「なし」を選択できます。「データの自動準備」 チェック・ボックスをオンにすると、正規化が自動的に実行されます。このチェック・ボックスをオフにすると、正規化は手動で行われます。

Oracle SVM の「エキスパート」オプション

カーネル キャッシュ・サイズ **:** 構築処理中に、計算したカーネルを保存するために使用するキャッシュのサイズをバイト単位で指定します。通常、キャッシュ・サイズを増やすと構築速度が速くなります。デフォルト値は 50MB です。

収束の許容範囲 **:** モデルの構築を打ち切るまで許される許容値を指定します。この値は 0～1 の間です。デフォルト値は 0.001 です。より大きな値では、構築速度が速くなり、またモデルの精度が低下する傾向があります。

標準偏差を指定 **:** ガウス・カーネルで使用される標準偏差パラメーターを指定します。このパラメーターは、モデルの複雑さと他のデータ・セットへの一般化 (データのオーバーフィットとアンダーフィット) との間のトレードオフに影響します。標準偏差値が高くなるにつれて、アンダーフィッティングになる傾向があります。デフォルトでは、このパラメーターは、学習用データから見積もられます。

イプシロン **(ε)** の指定 **:** イプシロンは、回帰モデルの場合にのみ、イプシロンに依存しないモデルの構築で許容されるエラーの区間を指定します。つまり、大きなエラー (無視できない) 大きなエラーから、(無視できる) 小さなエラーを区別します。この値は 0～1 の間です。デフォルトでは、学習データから推定されます。

複雑性ファクタの指定 **:** 複雑性ファクタを使用します。複雑性ファクタは、(学習用データについて計測される) モデル エラーと、モデルの複雑さの各値間のトレードオフ関係を制御して、データがオーバーフィッティングまたは、アンダーフィッティングになることを回避します。この値を大きくするとは、エラーに、より大きなペナルティーを課しことになり、データをオーバーフィッティングするリスクが増大します。この値を小さくするとは、エラーに、より小さなペナルティーを課しことになり、アンダーフィッティングになりやすくなります。

外れ値率の指定 **:** 学習データにおける希望の外れ値率を指定します。単一 SVM モデルにのみ有効 **:**「複雑性ファクタの指定」設定と併用できません。

予測確率 **:** モデルに、対象フィールドの出力に対する適切な予測の確率を含めることができます。この機能を有効にするには、「選択」 を選択して 「指定」 ボタンをクリックし、可能性のある出力のいずれかを選択してから、「挿入」 をクリックします。

予測セットを使用 **:** 対象フィールドの可能性のあるすべての出力に対する可能性のあるすべての結果の表を作成します。

Oracle SVM の「モデル」オプション

分類モデルでは、重みを使用して、さまざまな考えられる対象値の関連する重要度を指定することができます。例えば、学習データ内のデータ・ポイントがカテゴリー間で現実的に分布していない場合に役に立ちま

す。重みを使用すると、モデルを偏らせて、データにうまく表れていないカテゴリーの補正を行うことができます。対象値の重みが大きくなると、カテゴリーの適切な予測の割合が大きくなります。

重みを設定する方法は 3 つあります。

- 学習データに基づく**:** これがデフォルトです。重みは、学習データ中のカテゴリーの相対度数に基づいて決定されます。
- すべてのクラスで同じ **:** すべてのカテゴリーの重みが $1/k$ として定義されます。ここで、k は対象のカテゴリーの数です。
- ユーザー設定**:** 独自の重みを指定することができます。重みの開始値が、すべてのクラスで同じに設定されます。各カテゴリーの重みを、ユーザー定義値に調整することができます。特定のカテゴリーの重みを調整するには、そのカテゴリーに対応するテーブル中の重みのセルを選択し、セルの内容を削除してから、適切な値を入力してください。

すべての重みの合計は、1.0 です。合計が 1.0 にならない場合、値を自動的に正規化するオプションと警告が表示されます。この自動調整によって、重みの制約を強制しながら、カテゴリー間の比率が維持されます。この調整は、任意の時点で 「正規化」 ボタンをクリックして行うことができます。すべてのカテゴリーで値を均等化するためテーブルをリセットするには、「均等化」 ボタンをクリックします。

Oracle 一般化線型モデル (GLM)*

一般化線型モデルは、線型モデルによる限定的な仮説を緩和します。例えば、目標変数に正規分布があるという仮説、目標変数に対する予測の効果が本質的に線形となる仮説などです。一般化線型モデルは、対象の分布が、多項分布またはポアソン分布など、非正規分布となる傾向である場合の予測に適切です。同様に、一般化線型モデルは、予測値と目標変数間の関係またはつながりが非線型であると考えられる場合にも役立ちます。

詳細は、『*Oracle Data Mining Application Developer's Guide* 』 および 『*Oracle Data Mining Concepts* 』を参照してください。

Oracle GLM の「モデル」オプション

モデル名**:** 対象または ID フィールド (そのようなフィールドが指定されていない場合はモデル タイプ) に基づいてモデル名を自動的に生成するか、カスタム名を指定することができます。

一意のフィールド **:** 各ケースを一意に識別するためにフィールドを指定します。例えば、これは *CustomerID* などの ID フィールドです。IBM SPSS Modeler には制限があり、このキー・フィールドは数値である必要があります。

注 :このフィールドは、Oracle Adaptive Bayes、Oracle O-Cluster および Oracle Apriori を除くすべての Oracle ノードについてオプションです。

データの自動準備 **:** (11*g* のみ) Oracle Data Mining の自動データ準備モードを有効化 (デフォルト) または無効化します。このボックスがチェックされている場合、ODM は、アルゴリズムに必要なデータ変換を自動的に実行します。詳細は、『*Oracle Data Mining Concepts* 』を参照してください。

正規化**:** 連続型入力フィールドと対象フィールドの正規化方法を指定します。「**Z-**スコア」、「最小**-**最大」、または「なし」を選択できます。「データの自動準備」 チェック・ボックスをオンにすると、正規化が自動的に実行されます。このチェック・ボックスをオフにすると、正規化は手動で行われます。

欠損**'l**の処理 **:** 次のように、入力データの欠損値の処理方法を指定します。

- 「平均値または最頻値との置換」を選択すると、数値型属性の欠損値を平均値と置き換え、カテゴリー型属性の欠損値を最頻値と置き換えます。
- 「完全なレコードのみ使用」を使用すると、欠損値を含むレコードは無視されます。

Oracle GLM の「エキスパート」オプション

行の重みを使用 : 隣接するドロップダウン・リストで、行の重み付け因子を含む列を選択できます。

行の診断をテーブルに保存 : 隣接するテキスト・フィールドで行レベルの診断を含むテーブルの名前を入力できます。

係数の確信度レベル : 対象について予測された値が、モデルによって計算された信頼区間内にあるという 0.0 ～ 1.0 の確信度。確信度の境界は、係数の統計で返されます。

対象の参照カテゴリー: 「カスタム」を選択して参照カテゴリーとして使用する対象フィールドの値を選択するか、デフォルト値「自動」のままにします。

頂上回帰 : 頂上回帰は、変数の相関の程度が高すぎる状況を補正する方法です。「自動」 オプションを選択して、アルゴリズムがこの方法の使用を制御できるようにしたり、「無効」 および 「有効」 オプションを使用して手動で制御できるようにします。手動で頂上回帰を有効にすると、隣接するフィールドに値を入力して、システムのデフォルト値を頂上パラメーターで上書きすることができます。

頂上回帰の **VIF** を作成 : 頂上を線型回帰に使用する場合、分散拡大係数 (VIF) 統計を作成します。

予測確率 : モデルに、対象フィールドの出力に対する適切な予測の確率を含めることができます。この機能を有効にするには、「選択」 を選択して 「指定」 ボタンをクリックし、可能性のある出力のいずれかを選択してから、「挿入」 をクリックします。

予測セットを使用 : 対象フィールドの可能性のあるすべての出力に対する可能性のあるすべての結果の表を作成します。

Oracle GLM の「重み」オプション

分類モデルでは、重みを使用して、さまざまな考えられる対象値の関連する重要度を指定することができます。例えば、学習データ内のデータ・ポイントがカテゴリー間で現実的に分布していない場合に役に立ちます。重みを使用すると、モデルを偏らせて、データにうまく表れていないカテゴリーの補正を行うことができます。対象値の重みが大きくなると、カテゴリーの適切な予測の割合が大きくなります。

重みを設定する方法は 3 つあります。

- 学習データに基づく: これがデフォルトです。重みは、学習データ中のカテゴリーの相対度数に基づいて決定されます。
- すべてのクラスで同じ : すべてのカテゴリーの重みが $1/k$ として定義されます。ここで、k は対象のカテゴリーの数です。
- ユーザー設定: 独自の重みを指定することができます。重みの開始値が、すべてのクラスで同じに設定されます。各カテゴリーの重みを、ユーザー定義値に調整することができます。特定のカテゴリーの重みを調整するには、そのカテゴリーに対応するテーブル中の重みのセルを選択し、セルの内容を削除してから、適切な値を入力してください。

すべての重みの合計は、1.0 です。合計が 1.0 にならない場合、値を自動的に正規化するオプションと警告が表示されます。この自動調整によって、重みの制約を強制しながら、カテゴリー間の比率が維持されま

す。この調整は、任意の時点で 「正規化」 ボタンをクリックして行うことができます。すべてのカテゴリーで値を均等化するためテーブルをリセットするには、「均等化」 ボタンをクリックします。

Oracle デシジョン・ツリー

Oracle Data Mining は、一般的な分類と回帰ツリーのアルゴリズムに基づく古典的なディシジョン・ツリー機能を提供します。ODM ディシジョン・ツリー・モデルは、確信度、サポート、および分割基準を含む、それぞれのノードに関する完全な情報を備えています。それぞれのノードの完全なルールが表示され、さらに、欠損値のあるケースにモデルを適用する場合に代わりに使用すべき各ノードに代理変数属性が適用されます。

ディシジョン・ツリーはあらゆる場合に適用可能であり、利用しやすくわかりやすいため、一般的に使用されています。デシジョン・ツリーは、可能性のある入力属性を検査して、最適な「スプリッター」、つまり、下流データ・レコードを等質な母集団に分割する属性の分割点 (例: `年齢 > 55`) を検索します。それぞれの分割ディシジョンのあと、ODM は、ツリー全体を成長させ、レコード、項目、または人々の同種の母集団を表すターミナルの「葉」を作成するというプロセスを繰り返します。ルート ツリー・ノード (例: 母集団全体) からみると、デシジョン・ツリーは `IF A, then B` 文について、人が判読できる規則を提供します。このようなディシジョン・ツリーは、それぞれのツリー・ノードに関するサポートと確信度も提供します。

Adaptive Bayes Network はそれぞれの予測を説明するのに便利な簡潔でシンプルなルールも提供するのに対し、ディシジョン・ツリーはディシジョンを分割するための Oracle Data Mining ルールを提供します。ディシジョン・ツリーは、最良の顧客、健康的な患者、不正に関連する因子などの詳細なプロファイルを開発するのにも役立ちます。

ディシジョン・ツリーのモデル・オプション

モデル名**:** 対象または ID フィールド (そのようなフィールドが指定されていない場合はモデル タイプ) に基づいてモデル名を自動的に生成するか、カスタム名を指定することができます。

一意のフィールド **:** 各ケースを一意に識別するためにフィールドを指定します。例えば、これは *CustomerID* などの ID フィールドです。IBM SPSS Modeler には制限があり、このキー・フィールドは数値である必要があります。

注 :このフィールドは、Oracle Adaptive Bayes、Oracle O-Cluster および Oracle Apriori を除くすべての Oracle ノードについてオプションです。

データの自動準備 **:** (11*g* のみ) Oracle Data Mining の自動データ準備モードを有効化 (デフォルト) または無効化します。このボックスがチェックされている場合、ODM は、アルゴリズムに必要なデータ変換を自動的に実行します。詳細は、『*Oracle Data Mining Concepts* 』を参照してください。

不純度メトリック **:** ノードごとにデータを分割するための最良のテスト質問を求めるのに使用するメトリックを指定します。最適なスプリッタと分割値は、ノードのエンティティーの対象値の等質性が最も増大する場合に得られます。等質性はメトリックに従って測定されます。サポートされるメトリックは、**gini** と **entropy** です。

ディシジョン・ツリーのエキスパート・オプション

最大深度 **:** 構築するツリー・モデルの最大深度を設定します。

ノード内のレコードの最小パーセンテージ **:** ノード内の最小レコード数のパーセンテージを設定します。

分割用レコードの最小パーセンテージ：モデルを学習するのに使用するレコードの総数のパーセントとして表される親ノード内の最小レコード数を設定します。レコード数がこのパーセンテージ未満の場合、分割は行われません。

ノード内の最小レコード数：返されるレコード,l最小数を設定します。

分割用最小レコード数：数値として表される親ノード内の最小レコード数を設定します。レコード数がこの値未満の場合、分割は行われません。

ルール識別子：このオプションを選択すると、モデルに含まれる文字列が、特定の分割が行われるツリー内のノードを識別します。

予測確率：モデルに、対象フィールドの出力に対する適切な予測の確率を含めることができます。この機能を有効にするには、「選択」 を選択して 「指定」 ボタンをクリックし、可能性のある出力のいずれかを選択してから、「挿入」 をクリックします。

予測セットを使用：対象フィールドの可能性のあるすべての出力に対する可能性のあるすべての結果の表を作成します。

Oracle O-Cluster

Oracle O-Cluster のアルゴリズムは、データ母集団内の自然発生的なグループ化を識別します。直交データ区分クラスタリング (O-Cluster) は、階層グリッドベースのクラスタリング・モデルを作成する Oracle 独自のクラスタリング・アルゴリズムであり、すなわち、入力属性空間に軸並行 (直交) データ区分を作成します。このアルゴリズムは回帰的に動作します。結果として得られる階層構造は、属性空間がモザイク風のクラスターのように見える不規則なグリッドを表します。

O-Cluster アルゴリズムは数値属性とカテゴリー属性の両方を取り扱い、ODM は最適なクラスター定義を自動的に選択します。ODM はクラスター詳細情報、クラスター・ルール、クラスター重心値を提供し、母集団をスコアリングするのに使用できます。

O-Cluster の「モデル」オプション

モデル名: 対象または ID フィールド (そのようなフィールドが指定されていない場合はモデル タイプ) に基づいてモデル名を自動的に生成するか、カスタム名を指定することができます。

一意のフィールド：各ケースを一意に識別するためにフィールドを指定します。例えば、これは *CustomerID* などの ID フィールドです。IBM SPSS Modeler には制限があり、このキー・フィールドは数値である必要があります。

注 :このフィールドは、Oracle Adaptive Bayes、Oracle O-Cluster および Oracle Apriori を除くすべての Oracle ノードについてオプションです。

データの自動準備：(11*g* のみ) Oracle Data Mining の自動データ準備モードを有効化 (デフォルト) または無効化します。このボックスがチェックされている場合、ODM は、アルゴリズムに必要なデータ変換を自動的に実行します。詳細は、『*Oracle Data Mining Concepts* 』を参照してください。

最大クラスター数: 生成されるクラスターの最大数を設定します。

O-Cluster の「エキスパート」オプション

最大バッファ：最大バッファ サイズを設定します。

重要度 : 新しいクラスターを分離するのに必要なピーク濃度を指定する割合を設定します。この割合はグローバルな均一濃度に関連付けられます。

Oracle K-Means

Oracle K-means のアルゴリズムは、データ母集団内の自然発生的なグループ化を識別します。K-means のアルゴリズムは距離に基づくクラスタリング・アルゴリズムであり、データをあらかじめ決められた数のクラスターに分割します (ただし、明確に区別できるケースが十分にあることを前提とします) 。距離に基づくアルゴリズムは距離メトリック (関数) に依存してデータ・ポイント間の類似性を測定します。データ・ポイントは、使用する距離メトリックに従って最も近いクラスターに割り当てられます。ODM は、K-means の拡張バージョンを提供します。

K-means は階層クラスターをサポートし、数値属性とカテゴリー属性を取り扱い、母集団をユーザー指定のクラスター数に分割します。ODM はクラスター詳細情報、クラスター・ルール、クラスター重心値を提供し、母集団をスコアリングするのに使用できます。

K-Means の「モデル」オプション

モデル名: 対象または ID フィールド (そのようなフィールドが指定されていない場合はモデル タイプ) に基づいてモデル名を自動的に生成するか、カスタム名を指定することができます。

一意のフィールド : 各ケースを一意に識別するためにフィールドを指定します。例えば、これは *CustomerID* などの ID フィールドです。IBM SPSS Modeler には制限があり、このキー・フィールドは数値である必要があります。

注 :このフィールドは、Oracle Adaptive Bayes、Oracle O-Cluster および Oracle Apriori を除くすべての Oracle ノードについてオプションです。

データの自動準備 : (11*g* のみ) Oracle Data Mining の自動データ準備モードを有効化 (デフォルト) または無効化します。このボックスがチェックされている場合、ODM は、アルゴリズムに必要なデータ変換を自動的に実行します。詳細は、『*Oracle Data Mining Concepts* 』を参照してください。

クラスター数: 生成されるクラスターの数を設定します。

距離関数 : K-Means クラスタリングに使用する距離関数を指定します。

分割基準 : K-Means クラスタリングに使用する分割基準を指定します。

正規化: 連続型入力フィールドと対象フィールドの正規化方法を指定します。「**Z-**スコア」、「最小**-**最大」、または「なし」を選択できます。

K-means の「エキスパート」オプション

反復: K-means アルゴリズムの反復数を設定します。

収束の許容範囲 : K-means アルゴリズムの収束許容範囲を設定します。

ビン数: K-means で作成される属性ヒストグラムにおけるビン数を指定します。それぞれの属性のビンの境界は、学習データセット全体において包括的に計算されます。データ分割手段は固定幅です。ビンが一つしかない単一値の属性を除き、属性はすべて同じ数のビンを備えています。

ブロックの成長 : クラスター・データを保持するために割り当てられたメモリの成長因子を設定します。

最小パーセント属性のサポート : クラスターのルール詳細に属性を含めるために、ヌル以外の値にする必要がある属性値の小数部を設定します。欠損値のあるデータに設定したパラメーター値が高すぎると、ルールが非常に短くなったりあるいは空になることさえあります。

Oracle 非負数マトリックス因数分解 (NMF)

非負数マトリックス因数分解 (NMF) は、大きなデータセットを代表的な属性に縮小するのに便利です。NMF は、主成分分析 (PCA) と同様に、相加的モデル表現において多数の属性を処理できますが、多様な事例について使用できる広範囲な強力で最先端技術によるデータ・マイニング・アルゴリズムです。

NMF は、用例テキスト・データなど、多数のデータをより小さく簡潔な表現に縮小するのに使用することが可能であり、データの規模を低減できます (同じ情報量ならば、はるかに少ない変数で保存できます)。NMF モデルの出力は、SVM などの監視学習手法またはクラスタリング技法などの非監視学習手法を用いて分析できます。Oracle Data Mining は、NMF および SVM のアルゴリズムを使用して非構造テキスト・データをマイニングします。

NMF の「モデル」オfvション

モデル名: 対象または ID フィールド (そのようなフィールドが指定されていない場合はモデル タイプ) に基づいてモデル名を自動的に生成するか、カスタム名を指定することができます。

一意のフィールド : 各ケースを一意に識別するためにフィールドを指定します。例えば、これは *CustomerID* などの ID フィールドです。IBM SPSS Modeler には制限があり、このキー・フィールドは数値である必要があります。

注 :このフィールドは、Oracle Adaptive Bayes、Oracle O-Cluster および Oracle Apriori を除くすべての Oracle ノードについてオプションです。

データの自動準備 : (11*g* のみ) Oracle Data Mining の自動データ準備モードを有効化 (デフォルト) または無効化します。このボックスがチェックされている場合、ODM は、アルゴリズムに必要なデータ変換を自動的に実行します。詳細は、『*Oracle Data Mining Concepts* 』を参照してください。

正規化: 連続型入力フィールドと対象フィールドの正規化方法を指定します。「**Z-**スコア」、「最小**-**最大」、または「なし」を選択できます。「データの自動準備」 チェック・ボックスをオンにすると、正規化が自動的に実行されます。このチェック・ボックスをオフにすると、正規化は手動で行われます。

NMF の「エキスパート」オプション

フィールド数の指定 : 抽出するフィールド数を指定します。

ランダム シード。NMF アルゴリズムのランダム シードを設定します。

反復数: これにより、指定された反復数の後にモデル評価の作成を停止できます。NMF アルゴリズムの反復数を設定します。

収束の許容範囲 : NMF アルゴリズムの収束許容範囲を設定します。

すべての機能を表示: 最善の機能のみの値ではなく、すべての機能の機能 ID および確信度を表示します。

Oracle Apriori

Apriori アルゴリズムは、データのアソシエーション・ルールを発見します。例えば、顧客がひげそりとアフター・シェーブ ローションを購入した場合、その顧客は 80 % の確信度でシェービング クリームを購入します。アソシエーション・マイニングの問題は、2種類の下位問題に分解できます。

- 最小範囲よりも大きな範囲を持つ、多頻度アイテムセットと呼ばれる項目のすべての組み合わせを検索します。
- 多頻度アイテムセットを使用して希望のルールを生成します。これは、例えば ABC と BC が多頻度である場合、「A は BC を意味する」というルールは、support(BC) に対する support(ABC) の比率が少なくとも最小確信度と同じである場合に、有効であるという考えに基づきます。ABCD が多頻度であるため、このルールは最小範囲を備えることになります。ODM アソシエーションは、結果として生じる単一のルールのみをサポートします (ABC は D を意味します)。

多頻度アイテムセットの数は、最小範囲パラメーターで管理します。作成されるルールの数は、多頻度アイテムセット数と確信度パラメーターで管理します。確信度パラメーターの設定が高すぎる場合、ルールではなくアソシエーション・モデルに多頻度アイテムセットがあるかもしれません。

ODM は Apriori アルゴリズムの SQL ベースの実装です。候補の作成とサポート カウントのステップは、SQL クエリーを用いて実装されます。特化されたインメモリー・データ構造は使用しません。SQL クエリーは、データベース・サーバーで効率的に動作するようにさまざまなヒントで微調整されます。

Aprioriの「フィールド」オプション

すべてのモデル作成ノードには、「フィールド」タブがあり、そこからモデルの作成に使用するフィールドを指定できます。

Apriori モデルを作成する前に、アソシエーション・モデル作成の対象項目として使用するフィールドを指定する必要があります。

データ型ノードの設定を使用: このオプションを選択すると、上流のデータ型ノードからのフィールド情報がこのノードで使用されます。これがデフォルトです。

ユーザー設定を使用: このオプションを選択すると、上流のデータ型ノードからのフィールド情報ではなく、ここで指定したフィールド情報がこのノードで使用されます。このオプションを選択した後、トランザクション形式を使用しているかどうかによって、ダイアログ内の残りのフィールドを指定します。

トランザクション形式を使用しない場合、次を指定します。

- 入力: 入力フィールドを選択してください。これは、データ型ノードのフィールドの役割を 「入力」に設定するのと似ています。
- データ区分: このフィールドでは、モデル構築の学習、テスト、および検証の各ステージ用に、データを独立したサブセット (サンプル) に分割するフィールドを指定できます。

トランザクション形式を使用する 場合、次を指定します。

トランザクション形式を使用 : アイテムごとの行からケースごとの行にデータを変換する場合に、このオプションを使用します。

このオプションを選択すると、このダイアログ・ボックスの下部にあるフィールド・コントロールを次のように変更します。

トランザクション形式の場合、次を指定します。

- **ID:** リストから ID フィールドを選択します。ID フィールドとして使用できるのは、数値またはシンボル値のフィールドです。選択したフィールドでは、一意の値がそれぞれ、ある分析ユニットを示している必要があります。例えば、マーケット・バスケット分析なら、各 ID が 1 人の顧客を表します。Web ログ分析なら、各 ID が 1 台のコンピューター (IP アドレス) あるいは 1 人のユーザー (ログイン・データ) を表します。
- 「内容」。モデルの内容フィールドを指定します。このフィールドには、アソシエーション・モデル作成で関心の対象となる項目が含まれています。
- データ区分**:** このフィールドでは、モデル構築の学習、テスト、および検証の各ステージ用に、データを独立したサブセット (サンプル) に分割するフィールドを指定できます。1 組のサンプルをモデルの生成に使用し、別のサンプルで生成したモデルをテストすることにより、そのモデルが、このデータに似た性質を持つより大きなデータセットにどの程度適用できるかについての良い目安を得ることができます。データ型ノードまたはデータ区分ノードを使用することで、複数のデータ区分フィールドが定義されている場合、データ区分を使用する各モデル作成ノードの「フィールド」タブで単一のデータ区分フィールドを選択する必要があります。(1 つのデータ区分だけが存在している場合、データ分割を有効にすると、そのデータ区分が必ず自動的に使用されます)。また、選択したデータ区分を分析に適用するには、そのノードの「モデル・オプション」タブでもデータ区分が有効である必要があります。(このオプションの選択を解除すると、フィールドの設定を変更せずにデータ区分を無効にできます)。

Apriori の「モデル」オプション

モデル名**:** 対象または ID フィールド (そのようなフィールドが指定されていない場合はモデル タイプ) に基づいてモデル名を自動的に生成するか、カスタム名を指定することができます。

一意のフィールド **:** 各ケースを一意に識別するためにフィールドを指定します。例えば、これは *CustomerID* などの ID フィールドです。IBM SPSS Modeler には制限があり、このキー・フィールドは数値である必要があります。

注 :このフィールドは、Oracle Adaptive Bayes、Oracle O-Cluster および Oracle Apriori を除くすべての Oracle ノードについてオプションです。

データの自動準備 **:** (11*g* のみ) Oracle Data Mining の自動データ準備モードを有効化 (デフォルト) または無効化します。このボックスがチェックされている場合、ODM は、アルゴリズムに必要なデータ変換を自動的に実行します。詳細は、『*Oracle Data Mining Concepts* 』を参照してください。

ルールの最大長 **:** ルールの前提条件の最大数を 2 ～ 20 の整数で設定します。これにより、ルールの複雑さを制限します。ルールが複雑すぎるまたは特有である場合、またはルール・セットが長すぎて学習できない場合、この設定の数値を減らしてください。

最小確信度 **:** 確信度の最小値を 0 ～ 1 の値で設定します。指定された基準より低い確信度のルールは破棄されます。

最小範囲 **:** 閾値の最小範囲を 0 ～ 1 の値に設定します。Apriori は閾値の最小範囲を超える度数のパターンを検出します。

Oracle 最小記述長 (MDL)

Oracle 最小記述長 (MDL) アルゴリズムは、対象属性に最も大きな影響力を持つ属性を識別するのに役立ちます。多くの場合、最も影響力のある属性を知ることが事業をよく理解して管理するのに役立ち、モデル作成作業の簡素化を促進します。さらに、このような属性はモデルを拡張するために追加するデータのタイプを示します。MDL は、製造部品の品質、顧客離れに関連の因子、または特定の疾患の治療に最も関連していると思われる遺伝子を予測するのに最適なプロセス属性を検索するのに使用できます。

Oracle MDL は、対象を予測する場合に重要でないものと見なす入力フィールドを破棄します。残りのフィールドで、Oracle Data Miner で表示可能な、Oracle モデルに関連する未調整モデル・ナゲットを作成します。Oracle Data Miner でモデルを参照すると、残りの入力フィールドを、対象を予測する際の重要度の順に示すグラフが表示されます。

負の順位は、ノイズを示します。0 以下にランクされている入力フィールドは予測に貢献せず、データから削除する必要があります。

グラフを表示するには

1. 「モデル」パレットにある未調整モデル・ナゲットを右クリックして、 「参照」 を選択します。
2. モデル・ウィンドウで、ボタンをクリックして Oracle Data Miner を起動します。
3. Oracle Data Miner に接続します。 詳しくは、トピック 49 ページの『Oracle Data Miner』を参照してください。
4. Oracle Data Miner のナビゲータ・パネルで 「モデル」、「属性の重要度」 を展開します。
5. 関連する Oracle モデルを選択します (IBM SPSS Modeler で指定した対象フィールドと同じ名前です)。どれが適切なモデルがわからない場合、Attribute Importance フォルダーを選択して、作成日によってモデルを検索します。

MDL のモデル・オプション

モデル名**:** 対象または ID フィールド (そのようなフィールドが指定されていない場合はモデル タイプ) に基づいてモデル名を自動的に生成するか、カスタム名を指定することができます。

一意のフィールド **:** 各ケースを一意に識別するためにフィールドを指定します。例えば、これは *CustomerID* などの ID フィールドです。IBM SPSS Modeler には制限があり、このキー・フィールドは数値である必要があります。

注 :このフィールドは、Oracle Adaptive Bayes、Oracle O-Cluster および Oracle Apriori を除くすべての Oracle ノードについてオプションです。

データの自動準備 **:** (11*g* のみ) Oracle Data Mining の自動データ準備モードを有効化 (デフォルト) または無効化します。このボックスがチェックされている場合、ODM は、アルゴリズムに必要なデータ変換を自動的に実行します。詳細は、『*Oracle Data Mining Concepts* 』を参照してください。

Oracle Attribute Importance (AI)

属性の重要度の目的は、結果に関連するデータセットの属性、および最終的な結果に影響を与える程度を見出すことです。Oracle Attribute Importance モデル作成ノードは、データを分析、パターンを検索、関連レベルの確信度の結果を予測します。

AI のモデル・オプション

モデル名: 対象または ID フィールド (そのようなフィールドが指定されていない場合はモデル タイプ) に基づいてモデル名を自動的に生成するか、カスタム名を指定することができます。

データ区分データを使用。 データ区分フィールドが定義されている場合、このオプションでは学習用データ区分からのデータのみがモデル構築に使用されるようにします。

データの自動準備 : (11*g* のみ) Oracle Data Mining の自動データ準備モードを有効化 (デフォルト) または無効化します。このボックスがチェックされている場合、ODM は、アルゴリズムに必要なデータ変換を自動的に実行します。詳細は、『*Oracle Data Mining Concepts* 』を参照してください。

AI の選択オプション

「オプション」タブで、モデル・ナゲット内の入力フィールドを選択または除外するデフォルトの設定を指定できます。その後、以後のモデル構築作業で使用するフィールドのサブセットを選択するために、ストリームへモデルを追加できます。または、モデルの生成後にモデル・ブラウザー内で追加のフィールドを選択したり選択を解除したりして、このような設定を上書きすることもできます。ただし、デフォルトの設定はそれ以上変更しなくてもモデル・ナゲットに適用できるので、スクリプトを作成する目的に対しては、特に有用です。

使用可能なオプションは次のとおりです。

ランク付けされているすべてのフィールド : 「重要」、「境界」、または 「重要ではない」のランクに基づいてフィールドを選択します。各ランクと、レコードにランクを割り当てるために使用される分割値のラベルは、編集できます。

フィールドの上位数 : 重要度に基づいて上位 *n* 件のフィールドを選択します。

次より大きな重要度 : 指定された値よりも高い重要度のすべてのフィールドを選択します。

対象フィールドは、この選択にかかわらず、常に保存されます。

AI モデル・ナゲットの「モデル」タブ

Oracle AI モデル・ナゲットの「モデル」タブには、すべての入力フィールドのランクと重要度が表示されるので、フィルタリングするフィールドを、左の列のチェック・ボックスを使用して選択できるようになります。ストリームを実行すると、対象の予測とともに、チェック マークが付けられたフィールドだけが保存されます。その他の入力フィールドは、廃棄されます。デフォルトの選択はモデル作成ノード内で指定されたオプションに基づきますが、必要に応じて追加のフィールドを選択したり、選択を解除したりできます。

- ランク、フィールド名、重要度、またはその他の表示された列でリストをソートするには、列見出しをクリックします。または、ツールバーを使用して、「ソート項目」ボタンの隣のリストから該当する項目を選択し、上方向矢印と下方向矢印を使用してソートの方向を変更します。
- ツールバーを使用してすべてのフィールドにチェックを入れたり外したりできます。また、「フィールドのチェック」ダイアログ・ボックスを利用してランクまたは重要度でフィールドを選択できます。Shift キーまたは Ctrl キーを押してフィールドをクリックすると、複数選択することもできます。
- 重要度が高い、境界、重要度が低い、として入力フィールドをランク付けするための閾値は、テーブルの下の凡例に表示されます。これらの値は、モデル作成ノード内で指定されます。

Oracle モデルの管理

Oracle モデルは、他の IBM SPSS Modeler のモデルのように「モデル」パレットに追加され、ほとんど同じように使用できます。ただし、IBM SPSS Modeler 中で生成された各 Oracle モデルは実際にはデータベース・サーバー上にあるモデルへの参照であるなどの、重要な違いがいくつかあります。

モデル・ナゲットの「サーバー」タブ

ODM モデルの構築は、IBM SPSS Modeler を通じて IBM SPSS Modeler 内でモデルを作成し、さらに、Oracle データベース内でモデルを作成または置換します。この種類の IBM SPSS Modeler モデルは、データベース・サーバーに格納されているデータベース・モデルの内容を参照します。IBM SPSS Modeler は、IBM SPSS Modeler モデルと Oracle モデルの両方に、同一のモデル・キー文字列を生成して格納し、整合性チェックを実行します。

それぞれの Oracle モデル用のキー文字列は、「モデルの一覧」ダイアログ・ボックスの「モデル情報 」列に表示されます。IBM SPSS Modeler モデルのキー文字列は、IBM SPSS Modeler モデルの「サーバー」タブの「モデル・キー」として表示されます (ストリームに置かれた場合)。

モデル・ナゲットのダイアログ・ボックスにある「検査」ボタンは、IBM SPSS Modeler モデルと Oracle モデルのキーが一致するかどうかを検査するために使用できます。Oracle に同じ名前のモデルがないかモデル・キーが一致しない場合は、Oracle モデルは削除されているか、または、IBM SPSS Modeler モデルの作成によって再構築されています。

Oracle モデル・ナゲットの「要約」タブ

モデル・ナゲットの「要約」タブで、モデルそのもの (精度分析)、モデルで使用するフィールド (フィールド)、モデルの構築時に使用する設定 (構築の設定)、およびモデルの学習 (学習の要約) についての情報を表示します。

ノードを初めて参照する場合、「要約」タブの結果は閉じられています。目的の結果を表示するには、項目の左側にある展開コントロールを使用して項目を展開するか、または 「すべて展開」 ボタンをクリックしてすべての結果を表示します。見終わった結果を隠すには、展開コントロールを使用して目的の結果を省略するか、または「すべて閉じる」ボタンをクリックしてすべての結果を非表示にします。

精度分析 : 特定のモデルについての情報を表示します。このモデル・ナゲットに接続されている精度分析ノードを実行した場合、その精度分析情報もこのセクションに表示されます。

フィールド: 対象フィールドおよびモデル構築時の入力として使われるフィールドが表示されます。

構築の設定 : モデル構築時に使われる設定情報が表示されます。

学習の要約 : モデルの種類、モデルの作成に使われたストリーム、モデルの作成者、モデルの作成日時、およびモデルの構築時間などの情報が表示されます。

Oracle モデル・ナゲットの「設定」タブ

モデル・ナゲットの「設定」タブで、モデル作成ノードの特定のオプションの設定を、スコアリングの目的で優先させることができます。

Oracle デシジョン・ツリー

誤分類コストを使用 : Oracle Decision Tree モデルで誤分類コストを使用するかどうかを指定します。 詳しくは、トピック 33 ページの『誤分類コスト』を参照してください。

ルール識別子 **:** 選択した場合、ルール識別子の列が Oracle Decision Tree モデルに追加されます。ルール識別子は、特定の分割が作成されるツリーのノードを識別します。

Oracle NMF

すべての機能を表示**:** 最善の機能のみの値ではなく、Oracle NMF モデルの全機能の機能 ID および確信度を表示します。

Oracle モデルのリスト作成

「Oracle Data Mining Model の一覧」ボタンをクリックすると、既存のデータベース・モデルの一覧を表示するダイアログ・ボックスが表示され、モデルを削除できます。このダイアログ・ボックスは、「ヘルパー アプリケーション」ダイアログ・ボックスおよび ODM 関連のノードの構築、ブラウズ、および適用の各ダイアログ・ボックスから起動できます。

各モデルについて、次の情報が表示されます。

- モデル名**:** リストをソートするのに使用されるモデルの名前
- モデル情報**:** 構築日時と対象列から構成されたモデル・キー情報
- モデル・タイプ**:** このモデルの構築に使用されたアルゴリズムの名前

Oracle Data Miner

Oracle Data Miner は Oracle Data Mining (ODM) に対するユーザー・インターフェースであり、以前の IBM SPSS Modeler のユーザー・インターフェースを ODM 用に置き換えます。Oracle Data Miner は、ODM アルゴリズムの適切な活用においてアナリストの成功率を上げるように設計されています。このような目標は、いくつかの方法で対処します。

- ユーザーは、データの準備とアルゴリズムの選択の両方に対処する手法を適用する場合により多くの支援を必要とします。Oracle Data Miner は、データ・マイニング活動を提供して適切な手法をユーザーに示すことにより、この必要性を満たします。
- Oracle Data Miner は、改良および拡張された発見的手法をモデルの構築と変換のウィザードに含めて、モデルと変換の設定を指定する場合におけるエラー発生機会を低減します。

Oracle Data Miner の接続の定義

1. Oracle Data Miner は、「**Oracle Data Miner** の起動」ボタンを介して、Oracleの「構築」、「ノードの適用」、および「出力」のどのダイアログ・ボックスからでも起動できます。

図 2. 「*Oracle Data Miner* の起動」ボタン

2. Oracle Data Miner の「接続の編集」ダイアログ・ボックスは、Oracle Data Miner の外部アプリケーションが起動する前にユーザーに示されます (ただし、「ヘルパー アプリケーション」オプションが適切に定義されていることが前提です)。

   注: このダイアログ・ボックスは、定義された接続名がない場合にのみ表示されます。

   - Data Miner の接続名を指定し、適切な Oracle 10gR1 または 10gR2 のサーバー情報を入力します。Oracle サーバーは IBM SPSS Modeler で指定されているのと同じサーバーにする必要があります。

3. Oracle Data Miner の「接続の選択」ダイアログ・ボックスは、使用する接続名 (上記のステップで定義) を指定するためのオプションを示します。

Oracle Data Miner の要件、インストール、および使用に関する詳細は、Oracle の Web サイトの Oracle Data Miner を参照してください。

データの準備

Oracle Data Mining により提供される Naive Bayes、Adaptive Bayes,、および Support Vector Machine のいずれかを使用してモデルを作成する場合、2 種類のデータを準備すると便利です。

- データ分割とは、連続したデータを使用できないアルゴリズムのために、連続した数値範囲フィールドをカテゴリーへの変換することです。
- 正規化とは、ある数値の範囲が似かよった平均値と標準偏差を持つようにするために、数値範囲に適用される変換です。

データ分割

IBM SPSS Modeler のデータ分割ノードは、データ分割操作を実行するための数多くのテクニックを提供しています。同じデータ分割操作が、単一のフィールドにも複数のフィールドにも適用できるように定義されています。データ・セットに対してデータ分割操作を実行すると、閾値が生成され、IBM SPSS Modeler の フィールド作成ノードが作成できるようになります。フィールド作成操作は、 SQL に変換でき、モデルの構築およびスコアリングの前に適用できます。このアプローチでは、モデルと、データ分割を実行するフィールド作成ノードの間に依存関係が生じますが、データ分割の指定は、複数のモデル作成タスクで再利用できます。

正規化

Support Vector Machine モデルへの入力として使用される連続型 (数値範囲) フィールドは、モデルを構築する前に正規化する必要があります。回帰のモデルの場合、正規化は、モデル出力からスコアを再構築するためにも必要です。SVM モデルの設定では、「**Z-**スコア」、「**Min-Max**」、または「なし」を選択できます。正規化係数は、Oracle によりモデル構築プロセスの 1 ステップとして作成さます。そして、その係数は IBM SPSS Modeler にアップロードされ、そのモデルにも格納されます。適用時に、係数は IBM SPSS Modeler のフィールド作成式に変換され、スコアリング用のデータをモデルに渡す前に、そのデータの準備に使用されます。この場合、正規化は、モデル作成タスクと密接に関係しています。

Oracle データ・マイニングの例

IBM SPSS Modeler と共に ODM を使用する方法について解説する、数多くのサンプル・ストリームが含まれています。これらのストリームは、*¥Demos¥Database_Modelling¥Oracle Data Mining¥* の IBM SPSS Modeler インストール・フォルダーにあります。

注 : Demos フォルダーには、Windows の「スタート」メニューの IBM SPSS Modeler プログラム・グループからアクセスできます。

次の表に示すストリームは、Oracle Data Mining に組み込まれているサポート・ベクター・マシン (SVM) アルゴリズムを使用したデータベース・マイニング処理の例として、まとめて順に使用することができます。

表 4. データベース・マイニング - ストリーム例

ストリーム	説明
1_upload_data.str	フラット・ファイルのデータを整理して、データベースへアップロードするために使用されます。
2_explore_data.str	IBM SPSS Modeler でのデータ探索の例として使用されます。
3_build_model.str	データベース固有のアルゴリズムを使用したモデルを構築します。
4_evaluate_model.str	IBM SPSS Modeler でのモデル評価の例として使用されます。
5_deploy_model.str	データベース内スコアリングのためにモデルを展開します。

注 : サンプルを実行するには、ストリームを順番に実行する必要があります。さらに、各ストリーム中の入力およびモデル作成ノードは、使用するデータベースの有効なデータ・ソースを参照するように更新する必要があります。

サンプル・ストリームで使用されるデータセットは、クレジット カード申請に関するものであり、カテゴリー型および連続型予測フィールドの混在について、分類上の問題を提示します。このデータ・セットの詳細は、サンプル・ストリームと同じフォルダーにある *crx.names* ファイルを参照してください。

このデータセットは、*ftp://ftp.ics.uci.edu/pub/machine-learning-databases/credit-screening/* にある UCI Machine Learning Repository から入手可能です。

ストリームの例 :データのアップロード

最初のストリーム例 *1_upload_data.str* は、フラット・ファイルのデータを整理して Oracle へアップロードするのに使用されます。

Oracle Data Mining には、一意の ID フィールドが必要なので、この初期ストリームは、フィールド作成ノードを使用して、ユニークな値 1、2、3 を持つ *ID* という名前の新しいフィールドをデータセットに追加します。このフィールド作成ノードでは、IBM SPSS Modeler の `@INDEX` が使用されています。

置換ノードは、欠損値の処理に使用され、*crx.data* テキスト・ファイルから読み込まれた空のフィールドを *NULL* 値で置き換えます。

ストリームの例 :データの調査

2 番目の例のストリーム、*2_explore_data.str* を使用して、要約統計およびグラフなど、データの概要を取得するデータ検査ノードの使用方法を説明します。

データ検査レポート内のグラフをダブルクリックすると、指定されたフィールドについて、より深く検索した結果を表す詳細なグラフが生成されます。

ストリームの例 :モデルの作成

3 番目のストリーム例、*3_build_model.str* では、IBM SPSS Modeler でのモデル構築を説明します。データベース入力ノード (CREDIT) をダブルクリックして、データ・ソースを指定します。構築設定を指定するには、構築ノード (最初のラベルは CLASS、データ・ソースを指定すると FIELD16 に変更) をダブルクリックします。

ダイアログ・ボックスの「モデル」タブで、次の作業を行います。

1. 一意のフィールドとして「**ID**」を選択します。
2. カーネル 関数として「線型」を選択し、正規化の方法として「**Z-**スコア」を選択します。

ストリームの例 :モデルの評価

第 4 のストリーム例、*4_evaluate_model.str* では、IBM SPSS Modeler を使用したデータベース内モデル作成の利点を説明します。モデルを実行すると、その結果をユーザーのデータストリームに追加したり、IBM SPSS Modeler が提供するいくつかのツールを使用してモデルを評価したりできます。

モデル作成結果の表示

テーブル・ノードをモデル・ナゲットに適用して、結果を検証します。**$O-field16** フィールドに各ケースの *field16* の予測値が表示され、**$OC-field16** にこの予測の確信度値が表示されます。

モデル作成結果の評価

分析 ノードを使用すると、予測フィールドとその対象フィールド間の一致パターンを表す一致行列を作成できます。結果を表示するには分析ノードを実行します。

評価ノードを使用して、このモデルによりどの程度、精度が改善されたかを示すゲイン・グラフも作成できます。結果を表示するには評価ノードを実行します。

ストリームの例 :モデルの展開

モデルの精度が満足できるものになったら、外部のアプリケーションと共に使用するために展開したり、データベースに保存するために発行できます。最後のストリームの例、*5_deploy_model.str* では、データはテーブル `CREDITDATA` から読み込まれ、スコアリングされてから、*deploy solution* という Publisher ノードを使用して テーブル `CREDITSCORES` に発行されます。

第 5 章 IBM Netezza Analytics によるデータベース・モデリング

IBM SPSS Modeler と IBM Netezza Analytics

IBM SPSS Modeler では、IBM Netezza® Analytics との統合をサポートしており、IBM Netezza サーバーで高度な分析を実行する機能を提供します。これらの機能は、IBM SPSS Modeler のグラフィカル・ユーザー・インターフェースとワークフロー指向の開発環境で使用することができ、IBM Netezza 環境で直接データ・マイニング・アルゴリズムを実行できます。

IBM SPSS Modeler は、IBM Netezza Analytics の次のアルゴリズムの統合をサポートします。

- デシジョン・ツリー
- K-Means
- ベイズ・ネット
- Naive Bayes
- KNN
- 分裂クラスタリング
- PCA
- 回帰ツリー
- 線型回帰
- 時系列
- 一般化線型

アルゴリズムの詳細は、『 *IBM Netezza Analytics* 開発者ガイド 』および『 *IBM Netezza Analytics* リファレンス・ガイド 』を参照してください。

IBM Netezza Analytics との統合の要件

IBM Netezza Analytics を使用してデータベース内のモデル作成を実行する場合、以下の条件が前提条件となります。場合によっては、これらの条件が満たされているかデータベース管理者に問い合わせて確認してください。

- Windows または UNIX (IBM Netezza ODBC ドライバーを使用できない zLinux を除く) 上の IBM SPSS Modeler Server インストール済み環境に対して稼働する IBM SPSS Modeler。
- IBM Netezza Analytics パッケージを実行する IBM Netezza Performance Server。

  注: 必要な Netezza Performance Server (NPS) の最小バージョンは、必要な INZA のバージョンによって異なり、次のようになります。

  – NPS 6.0.0 P8 より大きなバージョンはすべて、2.0 より前の INZA のバージョンをサポートします。
  – INZA 2.0 以上を使用するには、NPS 6.0.5 P5 以上が必要です。

  Netezza 一般化線形および Netezza 時系列が機能するためには、INZA 2.0 以上が必要です。その他すべての Netezza データベース内ノードには、INZA 1.1 以降が必要です。

- IBM Netezza データベースに接続するための ODBC データ ソース。 詳しくは、トピック『IBM Netezza Analytics との統合の有効化』を参照してください。
- IBM SPSS Modeler で有効化された SQL の生成および最適化。 詳しくは、トピック『IBM Netezza Analytics との統合の有効化』を参照してください。

注 : データベース・モデル作成および SQL 最適化では、IBM SPSS Modeler Server 接続が IBM SPSS Modeler コンピューター上で可能でなければなりません。この設定を有効にすると、データベース・アルゴリズムにアクセスし、IBM SPSS Modeler から SQL を直接プッシュバック、IBM SPSS Modeler Server にアクセスできます。現在のライセンス ステータスを確認するには、IBM SPSS Modeler メニューから次を選択します。

「ヘルプ」 > 「バージョン情報」 > 「その他の詳細」

接続が有効な場合、「ライセンス ステータス」タブにオプション 「サーバーの有効化」 が表示されます。

IBM Netezza Analytics との統合の有効化

IBM Netezza Analytics と統合する手順は、次のとおりです。

- IBM Netezza Analytics の構成
- ODBC ソースの作成
- IBM SPSS Modeler で統合を有効にする
- IBM SPSS Modeler での SQL の生成と最適化を有効にする

これらについては、後の項で説明します。

IBM Netezza Analytics の構成

IBM Netezza Analytics のインストールと構成については、IBM Netezza Analytics の資料 (特に「*IBM Netezza Analytics Installation Guide*」) を参照してください。ガイドの「データベース権限の設定」に、IBM SPSS Modeler ストリームをデータベースに書き込むことを許可するために実行する必要のあるスクリプトの詳細について記載されています。

注: 行列の計算に依存するノード (Netezza PCA および Netezza 線型回帰) を使用する場合、`CALL NZM..INITIALIZE();` を実行して Netezza Matrix Engine を初期化する必要があります。実行しないと、ストアード・プロシージャーの実行が失敗します。各データベースごとに一度のセットアップ手順で初期化できます。

IBM Netezza Analytics の ODBC ソースの作成

IBM Netezza データベースと IBM SPSS Modeler の接続を有効にするには、ODBC システム・データ・ソース名 (DSN) を作成する必要があります。

DSN を作成する前に、ODBC データ ソースおよび ODBC ドライバーの基礎、さらに IBM SPSS Modeler のデータベース サポートの基礎を理解する必要があります。

IBM SPSS Modeler Server に対して、分散モードで実行している場合、サーバ コンピューターに DSN を作成します。ローカル (クライアント) モードで実行している場合、クライアント・コンピューターに DSN を作成します。

Windows クライアント

1. *Netezza Client* の CD の *nzodbcsetup.exe* ファイルを実行してインストーラーを起動してください。画面上の指示に従って、ドライバーをインストールします。詳細は、『IBM Netezza ODBC、JDBC、および OLE DB インストールおよび設定ガイド』を参照してください。
   a. DSN を作成します。

      注 :メニュー・シーケンスは使用する Windows のバージョンによって異なります。

      - **Windows XP** の場合。「スタート」メニューから、「コントロール パネル」を選択します。「管理ツール」 をダブルクリックし、次に 「データ ソース **(ODBC)**」 をダブルクリックします。
      - **Windows Vista** :「スタート」メニューから、「コントロール パネル」 → 「システム メンテナンス」 を選択します。「管理ツール」 をダブルクリックし、次に 「データ ソース **(ODBC)**」 を選択して 「開く」 をクリックします。
      - **Windows 7 :**「スタート」メニューから、「コントロール パネル」 → 「システムとセキュリティー」 → 「管理ツール」 を選択します。「データ ソース **(ODBC)**」 を選択して 「開く」 をクリックします。

   b. 「システム **DSN**」 タブをクリックしてから、「追加」 をクリックします。
2. リストから「**NetezzaSQL**」を選択し、「完了」をクリックします。
3. 「Netezza ODBC ドライバ セットアップ」画面の 「**DSN** オプション」 タブで、選択したデータ・ソース名、Netezza サーバーのホスト名または IP アドレス、接続のポート番号、使用している IBM Netezza インスタンスのデータベース、データベース接続のためのユーザー名およびパスワードの詳細を入力します。フィールドについての説明は「ヘルプ」 ボタンをクリックします。
4. 「テスト接続」 ボタンをクリックして、データベースに接続できることを確認します。
5. 正常に接続が行われたら、ODBC Data Source Administrator の画面が終了するまで繰り返し「**OK**」 を押します。

Windows サーバー

Windows サーバーの手順は、Windows XP のクライアントの手順と同じです。

UNIX または Linux サーバー

以下の手順は、UNIX または LINUX サーバー (IBM Netezza ODBC ドライバーを使用できない zLinux を除く) に適用されます。

1. Netezza Client CD/DVD から、関連する `<platform>cli.package.tar.gz` ファイルをサーバー上の一時的な場所にコピーします。
2. **`gunzip`** および **`untar`** コマンドを使用してアーカイブ・コンテンツを抽出します。
3. 実行権限を抽出された *unpack* スクリプトに追加します。
4. スクリプトを実行し、画面上のプロンプトに応答します。
5. `modelersrv.sh` ファイルを編集して以下の行を含めます。

   ```
   . <SDAP Install Path>/odbc.sh
   LD_LIBRARY_PATH_64=$LD_LIBRARY_PATH:/usr/local/nz/lib64; export LD_LIBRARY_PATH_64
   NZ_ODBC_INI_PATH=<SDAP Install Path>; export NZ_ODBC_INI_PATH
   ```

   以下に例を示します。

```
. /usr/IBM/SPSS/SDAP/odbc.sh
LD_LIBRARY_PATH_64=$LD_LIBRARY_PATH:/usr/local/nz/lib64; export LD_LIBRARY_PATH_64
NZ_ODBC_INI_PATH=/usr/IBM/SPSS/SDAP; export NZ_ODBC_INI_PATH
```

6. ファイル /usr/local/nz/lib64/odbc.ini を検索し、その内容を、SDAP とともにインストールされている odbc.ini ファイル ($ODBCINI 環境変数によって定義されたファイル) にコピーします。

   注 : 64 ビット Linux システムの場合、**Driver** パラメーターは誤って 32 ビット・ドライバーを参照します。前の手順で odbc.ini コンテンツをコピーする場合、次の例のようにパラメーター内のパスを編集します。

   ```
   /usr/local/nz/lib64/libnzodbc.so
   ```

7. Netezza DSN 定義のパラメーターを編集して使用されるデータベースを反映させます。
8. IBM SPSS Modeler Server を再起動して、クライアント上の Netezza データベース内マイニング・ノードの使用をテストします。

IBM SPSS Modeler で IBM Netezza Analytics の統合を有効にする

1. IBM SPSS Modeler のメイン・メニューから次の各項目を選択します。

   「ツール」 > 「オプション」 > 「ヘルパー アプリケーション」

2. 「**IBM Netezza**」タブをクリックします。

Netezza Data Mining との統合を有効化**:** IBM SPSS Modeler ウィンドウの下部でデータベース モデリング パレットを有効にし (まだ表示されていない場合)、Netezza Data Mining Integration アルゴリズムのノードを追加します。

Netezza 接続**:** 「編集」ボタンをクリックし、ODBC ソース作成時に以前設定した Netezza 接続文字列を選択します。 詳しくは、トピック 54 ページの『IBM Netezza Analytics の ODBC ソースの作成』を参照してください。

SQL の生成と最適化を有効にする

非常に大きいデータ・セットを扱う確率が高いため、パフォーマンス上の理由により、IBM SPSS Modeler で SQL 生成および最適化を有効にする必要があります。

1. IBM SPSS Modeler のメニューから次の項目を選択します。

   「ツール」 > 「ストリームのプロパティー」 > 「オプション」

2. ナビゲーション ペインの 「最適化」 オプションをクリックします。
3. 「**SQL** 生成」 オプションが有効になっていること確認します。この設定は、データベースのモデル作成が機能するために必要です。
4. 「**SQL** 生成の最適化」 と 「その他の実行を最適化」 を選択します (絶対に必要な訳ではありませんが、最適化されたパフォーマンスを得るために、選択することを強くお勧めします)。

IBM Netezza Analytics によるモデル構築

サポートされているアルゴリズムのそれぞれに、対応するモデリング・ノードがあります。ノード・パレットの「データベース・モデリング」タブから、IBM Netezza モデリング・ノードにアクセスできます。

データの考慮事項

データ・ソース内のフィールドは、モデリング・ノードに応じて、さまざまなデータ型の変数を含めることができます。IBM SPSS Modeler では、データ型は測定の尺度とも呼ばれます。モデリングのノードの「フィールド」タブでは、入力フィールドと対象フィールドに許可されている測定の尺度の種類を示すアイコンを使用しています。

対象フィールド: 対象フィールドは、値を予測しようとしているフィールドです。対象を指定できる場合、ソース・データ・フィールドのうち 1 つだけを対象フィールドとして選択できます。

レコード **ID** フィールド: 各ケースを一意に識別するために使用するフィールドを指定します。例えば、これは *CustomerID* などの ID フィールドです。ソースデータに ID フィールドが含まれていない場合は、次の手順に示すように、フィールド生成ノードを使用してこのフィールドを作成することができます。

1. 入力ノードを選択します。
2. ノード・パレットの「フィールド設定」タブから、フィールド生成ノードをダブルクリックします。
3. 領域内のアイコンをダブルクリックしてフィールド生成ノードを開きます。
4. 「派生フィールド」 フィールドで、例えば `ID` を入力します。
5. 「**CLEM** 式」 フィールドで、`@INDEX` と入力して 「**OK**」 をクリックします。
6. フィールド生成ノードをストリームの残りに接続します。

注**:** `NUMERIC(18,0)` データ型を使用して Netezza データベースから long 数値データを取得する場合、SPSS Modeler はインポート時にデータを切り上げることがあります。この問題を回避するために、データは `BIGINT` または `NUMERIC(36,0)` のいずれかのデータ型を使用して保管します。

注**:** 使用できるフィールド・タイプの制限により、「レコード ID」の役割およびデータ型不明の「測定の尺度」があるフィールドは、Netezza データベース内モデル作成ノード (例えば、K-Means) には表示されません。

ヌル値の処理

入力データに null 値が含まれている場合、いくつかの Netezza ノードを使用することによってエラー・メッセージが表示されたりストリームの実行時間が長くなる可能性があるため、null 値を含むレコードを削除することをお勧めします。以下の方法を使用します。

1. 条件抽出ノードを入力ノードに接続します。
2. 条件抽出ノードの 「モード」 を 「破棄」 に設定します。
3. 「条件」 フィールドに以下を入力します。

   `@NULL(`*`field1`*`) [or @NULL(`*`field2`*`)[... or @NULL(`*`fieldN`*`]])`

   すべての入力フィールドを含むようにします。
4. 条件抽出ノードをストリームの残りに接続します。

モデル出力

Netezza モデル作成ノードを含むストリームは、わずかに異なる結果を実行のたびに生成する可能性があります。それは、データがモデル作成の前に一時テーブルに読み込まれるため、ノードがソース・データを読み取る順序が必ずしも同じではないからです。ただし、この効果によって生まれる相違点は無視できます。

一般的なコメント

- IBM SPSS Collaboration and Deployment Services では、IBM Netezza データベース・モデリング・ノードを含むストリームを使用してスコアリング設定を作成することはできません。
- Netezza ノードで作成されたモデルの場合、PMML エクスポートまたはインポートを行うことはできません。

Netezza モデル - フィールド・オプション

「フィールド」タブで、上流のノードですでに定義されているフィールドの役割設定を使用するか、手動でフィールドの割り当てを行うかを選択します。

事前定義された役割を使用: 上流のデータ型ノード (または上流の入力ノードの「データ型」タブ) の役割設定 (対象、予測など) を使用します。

カスタム・フィールド割り当ての使用: この画面で対象、予測、およびその他の役割を手動で割り当てる場合、このオプションを選択します。

フィールド**:** 矢印ボタンを使用して、このリストの項目を画面右側のさまざまな役割フィールドに手動で割り当てます。アイコンは、各役割フィールドの有効な測定の尺度を示します。

「すべて」 ボタンをクリックしてリスト内のすべてのフィールドを選択するか、各測定の尺度のボタンをクリックして、その測定の尺度のすべてのフィールドを選択します。

目標: 1 つのフィールドを予測の対象として選択します。一般化線形モデルの場合、この画面の 「試行回数」 フィールドも参照してください。

「レコード **ID**」。一意のレコード ID として使用されるフィールド。

予測値 **(**入力**):** 1 つ以上のフィールドを予測の入力として選択します。

Netezza モデル - サーバー・オプション

「サーバー」タブでは、モデルが構築される IBM Netezza データベースを指定します。

Netezza DB サーバーの詳細:ここで、モデルに使用するデータベースの接続の詳細を指定します。

- 上流の接続を使用: (デフォルト) データベース入力ノードなど、上流のノードで指定した接続の詳細を使用します。注 : このオプションは、すべての上流ノードが SQL プッシュバックを使用できる場合にのみ有効です。この場合、SQL が完全にすべての上流ノードを実装するため、データベース外のデータを移動する必要はありません。
- 接続するデータを移動:ここでしてデータベースにデータを移動します。データを移動することにより、データが別の IBM Netezza データベース、他のベンダーのデータベースにある場合、またはデータがフラット・ファイルにある場合でもモデリングを実行できるようにします。また、ノードが SQL プッシュバックを実行していないためデータが抽出されていない場合、データはここで指定されたデータベースに戻されます。接続を参照して選択するには、「編集」ボタンをクリックします。注意 : IBM Netezza Analytics は通常非常に大きいデータ・セットで使用されます。データベース間、またはデータベース内外に多くのデータを移行するのは非常に時間がかかる場合があるため、可能な限り避けることをお勧めします。

注: ODBC データ・ソース名は、各 IBM SPSS Modeler ストリームに効果的に埋め込まれます。あるホスト上で作成されたストリームが別のホスト上で実行された場合、データ・ソースの名前はそれぞれのホストで同じである必要があります。また、各入力ノードまたはモデル作成ノードで、「サーバー」タブから異なるデータ・ソースを選択できます。

Netezza モデル - モデル・オプション

「モデル・オプション」タブで、モデルの名前を指定するか、自動的に名前を生成するかを選択できます。スコアリングのオプションのデフォルト値を設定することもできます。

モデル名: 対象または ID フィールド (そのようなフィールドが指定されていない場合はモデル タイプ) に基づいてモデル名を自動的に生成するか、カスタム名を指定することができます。

名前が使用されている場合は既存モデルを置換。このチェック ボックスを選択する場合、同じ名前の既存モデルは上書きされます。

スコアリングに使用できるようにする。モデル・ナゲットのダイアログに表示されるスコアリング・オプションのデフォルト値を設定できます。オプションの詳細については、その特定のナゲットの「設定」タブのヘルプトピックを参照してください。

Netezza モデルの管理

IBM SPSS Modeler によって IBM Netezza モデルを構築すると、IBM SPSS Modeler 内にモデルが作成されると同時に、Netezza データベース内でもモデルの作成や置き換えが行われます。この種類の IBM SPSS Modeler モデルは、データベース・サーバーに格納されているデータベース・モデルの内容を参照します。IBM SPSS Modeler では、IBM SPSS Modeler モデルと Netezza モデルの両方に、同一のモデル・キー文字列を生成して格納することで、整合性チェックを実行できます。

各 Netezza モデルのモデル名は、「データベース・モデルの一覧表示」ダイアログ・ボックスの「モデル情報」列に表示されます。 IBM SPSS Modeler モデルのモデル名は、IBM SPSS Modeler モデルの「サーバー」タブの「モデル・キー」として表示されます (ストリーム内に配置された場合)。

「検査」ボタンを使用すると、IBM SPSS Modeler モデルと Netezza モデルのモデル・キーが一致するかどうかを検査できます。 Netezza に同じ名前のモデルがないか、モデル・キーが一致しない場合、Netezza モデルは削除されているか、または、IBM SPSS Modeler モデルの作成によって再構築されています。

データベース・モデルの一覧表示

IBM SPSS Modeler には、IBM Netezza に格納されたモデルを一覧表示するダイアログ・ボックスがあり、そこでモデルを削除できます。このダイアログ・ボックスには、IBM「ヘルパー アプリケーション」ダイアログ・ボックスと、IBM Netezza Data Mining 関連ノードの構築、参照、および適用ダイアログ・ボックスからアクセス可能です。各モデルについて、次の情報が表示されます。

- モデル名 (リストをソートするのに使用されるモデルの名前)
- 所有者名。
- モデルで使用されるアルゴリズム。
- モデルの現在の状態 (例えば「完了」)。
- モデルが作成された日付。

Netezza 回帰ツリー

回帰ツリーは、数値型対象フィールドの値に基づいて同じ種類のサブセットを派生させるために、ケースのサンプルを繰り返し分割するツリーベースのアルゴリズムです。ディシジョン・ツリーと同様に、回帰ツリーはツリーの葉が十分に小さいか、均一なサブセットに対応するサブセットにデータを分解します。分割は、葉の平均値で合理的に予測できるよう、対象の属性の値のばらつきを減少させるために選択されます。

Netezza Netezza 回帰ツリーの作成オプション - ツリーの成長

ツリーの成長とツリーの剪定の作成オプションを設定できます。

ツリーの成長には、以下の作成オプションを使用できます。

最大ツリー深さ。 ルート・ノード下でツリーが成長可能な最大レベル数 (サンプルが再帰的に分割される回数)。デフォルトは 62 で、モデリングのための最大ツリー深度です。

注: モデル・ナゲットのビューアーにモデルがテキスト表示される場合、最大 12 のツリー・レベルが表示されます。

分割の基準。これらのオプションは、ツリーの分割をいつ停止するかを制御します。デフォルト値を使用しない場合は、「カスタマイズ」をクリックして値を変更します。

- 評価指標の分割。このクラス評価測定では、ツリーの分割に最適な場所が評価されます。

  注: 現在、使用可能なオプションは「分散」のみです。
- 分割の改善度の最小変化量。新しい分割がツリーで作成される前に不純度を減少する必要のある最小数。ツリー構築の目標は、似かよった出力値を持つサブグループを作成して、それぞれのノード内における不純度を最小にすることです。ブランチが適切に分割されて不純度が分割基準によって指定された値を下回ると、ブランチは分割されません。
- 分割の最小インスタンス数。分割可能な最小レコード数。分割されていないレコードがこの数より少ない場合、これ以上分割は行われません。このフィールドを使用すると、ツリー内に小さいサブグループが作成されないようにすることができます。

統計: このパラメーターは、モデルに含まれる統計の数を定義します。以下のいずれかのオプションを選択します。

- 「すべて」。列に関連したすべての統計と値に関連したすべての統計が含まれます。

  注: このパラメーターを使用すると統計の最大数が含まれるため、システムのパフォーマンスに影響する場合があります。モデルをグラフィカル形式で表示しない場合は、「なし」を指定します。
- 「列」。列に関連した統計が含まれます。
- なし: モデルのスコアリングに必要な統計のみが含まれます。

Netezza 回帰ツリーの作成オプション - ツリーの剪定

剪定オプションを使用して、回帰ツリーの剪定基準を指定できます。剪定の目的は、新しいデータに対して予期した精度が改善されない成長しすぎたサブグループを削除することによって、オーバーフィットのリスクを軽減することにあります。

剪定の測定: 剪定の測定により、ツリーから葉を削除した後、モデルの推定精度を許容限度内に保つことができます。。次のいずれかを選択できます。

- **mse:** 平方平均誤差 - (デフォルト) 近似直線がデータ・ポイントにどれだけ近いかを測定します。

- **r2.** R-squared - 回帰モデルによって説明される従属変数の変動の比率です。
- **Pearson :** Pearson の相関係数 - 正規分布されている線形従属変数間の関係の強さを測定します。
- **Spearman:** Spearman の相関係数 - ピアソンの相関従って弱いと思われるが、実際には強いと考えられる非線形な関係を検出します。

剪定するデータ。新しいデータに対する想定された精度を推定するには、学習データの一部またはすべてを使用できます。また、指定されたテーブルの別の剪定データセットを使用できます。

- すべての学習データを使用。このオプション (デフォルト) は、モデルの精度を推定するためにすべての学習データを使用します。
- 剪定に次のパーセンテージ学習データを使用。 剪定データに指定した割合で、一方は学習用、一方は剪定用です。

  ストリームを実行するごとにデータを同じ方法で区分するようランダム シードを指定する場合、「結果を再現」 を選択します。「剪定に使用するシード」 フィールドで整数を指定するか、または 「生成」 をクリックすると、擬似無作為の整数を作成します。
- 既存のテーブルのデータを使用。モデルの精度を推定するために個別の剪定データセットのテーブル名を指定します。学習データを使用するより信頼性が高いと見なされます。ただし、このオプションにより、学習セットから大きなサブセットのデータが削除され、ディシジョン・ツリーの品質が損なわれます。

Netezza 分裂クラスタリング

分裂クラスタリングは、アルゴリズムが指定された停止ポイントに到達するまで、サブクラスターにクラスターを分割するために繰り返し実行されるクラスター分析の手法です。

クラスター形成は、すべての学習インスタンス（レコード）を含む単一のクラスターから始まります。アルゴリズムの最初の繰り返しでは 2 つのサブクラスターにデータセットを分割し、後続の反復でそれらをさらにサブクラスターに分割します。停止条件は繰り返しの最大数、データセットが分割されているレベルの最大数、およびさらにデータ区分のためのインスタンスの必要最小限数として指定されます。

結果として得られる階層クラスタリング・ツリーを次の例で使用して、ルート・クラスターからそれらを伝播させることによって、インスタンスを分類することができます。

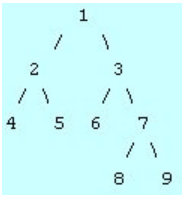

図 3. 分裂クラスタリング・ツリーの例

各レベルで、最良のマッチング・サブクラスターがサブクラスター中心からのインスタンスの距離を基準に選択されます。

葉は負の数にで指定されているため、インスタンスが -1（デフォルト）の適用階層レベルでスコアリングされている場合、スコアリングは葉クラスターのみを返します。例では、これはクラスター 4、5、6、8、または 9 のいずれかになります。ただし、階層レベルが 2 に設定されている場合、例えばスコアリングは、ルート・クラスターより下の 2 番目のレベル、すなわち 4、5、6、または 7 のいずれかのクラスター返します。

Netezza 分裂クラスタリングのフィールド・オプション

「フィールド」タブで、上流のノードですでに定義されているフィールドの役割設定を使用するか、手動でフィールドの割り当てを行うかを選択します。

事前定義された役割を使用: 上流のデータ型ノード (または上流の入力ノードの「データ型」タブ) の役割設定 (対象、予測など) を使用します。

カスタム・フィールド割り当ての使用: この画面で対象、予測、およびその他の役割を手動で割り当てる場合、このオプションを選択します。

フィールド**:** 矢印ボタンを使用して、このリストの項目を画面右側のさまざまな役割フィールドに手動で割り当てます。アイコンは、各役割フィールドの有効な測定の尺度を示します。

「すべて」 ボタンをクリックしてリスト内のすべてのフィールドを選択するか、各測定の尺度のボタンをクリックして、その測定の尺度のすべてのフィールドを選択します。

「レコード **ID**」。一意のレコード ID として使用されるフィールド。

予測値 **(**入力**):** 1 つ以上のフィールドを予測の入力として選択します。

Netezza 分裂クラスタリングの作成オプション

「作成オプション」タブで、モデルを構築するすべてのオプションを設定します。もちろん、「実行」 ボタンをクリックすると、すべてデフォルト・オプションのモデルが構築されますが、通常は、それぞれの目的で作成をカスタマイズする必要があります。

距離測度 :データ・ポイント間の距離の測定に使用する方法です。距離が大きいほど、相違が大きくなります。オプションは次のとおりです。

- ユークリッド:(デフォルト) 2 点間の距離は、それらを直線で結ぶことによって計算されます。
- **Manhattan**:2 点間の距離は、それらの座標間の絶対距離の合計として計算されます。
- **Canberra**:Manhattan の距離と同じですが、原点に近いデータ・ポイントに対してより感度が高くなります。
- 最大値 :2 点間の距離は、座標の次元に沿ったそれらの相違の最大値として計算されます。

最大反復回数:アルゴリズムは、同じプロセスを何度か反復することによって実行します。指定された回数だけ反復した後、モデルの学習を中止します。

クラスター・ツリーの最大深度**:** データセットを分割することができるレベルの最大数。

結果を再現: 分析を複製できるようになります。整数を指定、または 「生成」 をクリックすると、擬似無作為の整数を作成します。

分割の最小インスタンス数: 分割可能な最小レコード数。分割されていないレコードがこの数より少ない場合、これ以上分割は行われません。このフィールドを使用して、クラスター・ツリー内の非常に小さいサブグループが作成されないようにすることができます。

Netezza 一般化線型

線型回帰は、数値型入力フィールドの値に基づいてレコードを分類する長きにわたって確立された統計手法です。線型回帰は、予測された出力値と実際の出力値の違いを最小限にする直線または面に適合します。線形モデルは、学習とモデル・アプリケーションにおいて、その単純さにより、さまざまな実世界の現象をモデル化するのに有用である。しかし、線形モデルは、従属 (ターゲット) 変数の正規分布及び従属変数に対する独立 (予測) 変数の線形衝突を想定します。

線形回帰が役立つ状況は数多くありますが、上記の想定は適用されません。例えば、様々な製品間における消費者の選択をモデリングする場合、従属変数は多項分布である可能性が高くなります。同様に、年齢に対する収入をモデリングすると、通常は年齢が上がるごとに収入は高くなりますが、2 つの間のリンクは直線ほど単純でなくなります。

こうした状況の場合、一般化線形モデルを使用できます。一般化線型モデルは、指定したリンク関数によって従属変数が予測変数に関連するよう、線形回帰モデルを拡張したものです。さらにこのモデルでは、非正規分布にポアソンなどの従属変数を使用できます。

アルゴリズムは、指定された反復数を最大値として、最適なモデルを反復して探します。適合度を計算する場合、従属変数の予測値及び実際の値との間の差の平方和によって誤差が示されます。

Netezza 一般化線形モデルのフィールド オプション

「フィールド」タブで、上流のノードですでに定義されているフィールドの役割設定を使用するか、手動でフィールドの割り当てを行うかを選択します。

事前定義された役割を使用: このオプションでは、上流のデータ型ノードまたは上流の入力ノードの「データ型」タブの役割設定 (対象や予測など) を使用します。

カスタム・フィールド割り当ての使用: この画面で対象、予測、およびその他の役割を手動で割り当てる場合、このオプションを選択します。

フィールド**:** 矢印ボタンを使用して、このリストの項目を画面右側のさまざまな役割フィールドに手動で割り当てます。アイコンは、各役割フィールドの有効な測定の尺度を示します。

「すべて」 ボタンをクリックしてリスト内のすべてのフィールドを選択するか、各測定の尺度のボタンをクリックして、その測定の尺度のすべてのフィールドを選択します。

目標: 1 つのフィールドを予測の対象として選択します。

「レコード **ID**」。一意のレコード ID として使用されるフィールド。このフィールドの値は各レコードで一意である必要があります (例: カスタマー ID 番号)。

インスタンスの重み。インスタンスの重みを使用するフィールドを指定します。インスタンスの重みは、入力データの行あたりの重みです。デフォルトでは、すべての入力レコードの重要度は相対的に等しいと想定されています。個々の重みを入力レコードに割り当てることにより、重要度を変更できます。指定するフィールドには、入力データの各行の数値の重みが含まれている必要があります。

予測値 **(**入力**):** 1 つ以上の入力フィールドを選択します。この操作は、データ型ノードのフィールドの役割を「入力」に設定するのと似ています。

Netezza 一般化線形モデル・オプション - 全般

「モデル・オプション」タブで、モデルの名前を指定するか、自動的に名前を生成するかを選択できます。モデル、リンク関数、入力フィールドの反復 (あれば)に関連する様々な設定を行い、スコアリング・オプションのデフォルト値を設定します。

モデル名**:** 対象または ID フィールド (そのようなフィールドが指定されていない場合はモデル タイプ) に基づいてモデル名を自動的に生成するか、カスタム名を指定することができます。

フィールド・オプション。モデル構築用の入力フィールドの役割を指定できます。

一般的な設定**:** アルゴリズムの停止基準に関連します。

- 最大反復数**:** 最小値は 1、デフォルトは 20 です。
- 最大エラー **(1e):** アルゴリズムが適合度モデルの検索を停止する最大誤差の値 (科学的表記)。最小値は 0、デフォルトは -3、つまり、1E-3 または 0.001 となります。
- 有意でないエラー値の閾値 **(1e):** 誤差が 0 として扱われる値 (科学的表記)。最小値は -1、デフォルトは -7、つまり 1E-7 (または 0.0000001) を下回る誤差の値が有意でないとカウントされます。

分布設定**:** これらの設定は、従属 (ターゲット) 変数の分布に関連します。

- 応答変数の分布**:** 分布のタイプは、ベルヌーイ (デフォルト)、ガウス、ポアソン、二項分布、負の二項分布、ワルド (逆ガウス)、およびガンマのいずれかです。
- パラメーター: (ポアソンおよび二項分布のみ) 「パラメータの指定」フィールドに以下のオプションのいずれかを指定する必要があります。
  - データからパラメーターを自動的に推定させるには、「デフォルト」を選択します。
  - 分布疑似尤度の最適化を許可するには、「準**(Quasi)**」を選択します。
  - パラメーター値を明示的に指定するには、「明示 **(Explicit)**」を指定します。

  (二項分布のみ) 二項分布の要求に応じて試行フィールドとして使用される入力テーブル列を指定する必要があります。この列には、二項分布の繰り返し試行数が入ります。

  (負の二項分布のみ) デフォルトの -1 を使用するか、または別のパラメーター値を指定することができます。

リンク関数設定**:** これらの設定は、従属 (ターゲット) 変数を予測変数に関連させるリンク関数に関連します。

- リンク関数: 使用する関数は、**Identity**、**Inverse**、**Invnegative**、**Invsquare**、**Sqrt**、**Power**、**Oddspower**、**Log**、**Clog**、**Loglog**、**Cloglog**、**Logit** (デフォルト)、**Probit**、**Gaussit**、**Cauchit**、**Canbinom**、**Cangeom**、**Cannegbinom** のいずれかです。
- パラメーター: (Power または Oddspower リンク関数のみ) リンク関数が **Power** または **Oddspower** の場合、パラメーター値を指定できます。値を指定するか、デフォルトの 1 を使用するかを選択します。

Netezza 一般化線形モデル・オプション - 相互作用

相互作用パネルには、相互作用を指定するオプション (入力フィールド間の倍数の影響) が表示されます

列の相互作用 **:**このチェック・ボックスを選択すると、入力フィールド間の相互作用を指定します。相互作用がない場合は、ボックスをオフにします。

ソース・リスト内の1つ以上のフィールドを選択し、相互作用リストにドラッグして、相互作用をモデルに入力します。作成する相互作用の種類は、選択項目をドロップするホットスポットによって異なります。

- 主相互作用**:** ドロップされたフィールドは、交互作用リストの一下部に別の主交互作用として表示されます。
- **2** 次。ドロップされたフィールドのすべての可能なペアが、2 次交互作用として交互作用リストの下部に表示されます。
- **3** 次。ドロップされたフィールドのすべての可能なトリプレットが、3 次交互作用として交互作用リストの下部に表示されます。
- ***.** ドロップされたすべてのフィールドの組み合わせは、相互作用リストの一番下にある単一の相互作用として表示されます。

切片を含める。通常、切片はモデルに含まれます。データが原点を通ると仮定できる場合は、切片を除外できます。

ダイアログ・ボックスのボタン

表示の右にあるボタンを使用すると、モデルで使用する用語に変更を加えることができます。

図 *4.* 「削除」ボタン

削除したい条件を選択し、削除ボタンをクリックして、モデルから用語を削除します。

図 *5.* 「並べ替え」ボタン

順序を変更する条件を選択し、上向きまたは下向きの矢印をクリックして、モデル内の項目を並べ替えます。

図 *6.* 「カスタム交互作用」ボタン

カスタム項目の追加

*n1*x1*x1*x1..* の形式でカスタム項を指定できます。「フィールド」 リストからフィールドを選択し、右方向矢印をクリックしてフィールドを 「カスタム項目」 に追加し、「乗算*」 をクリックし、次のフィールドを繰り返して、右方向矢印をクリックして、を繰り返します。カスタム相互作用を構築したら、「項目を追加」 をクリックして「相互作用」パネルに戻ります。

Netezza 一般化線形モデル・オプション - スコアリング・オプション

スコアリングに使用できるようにする。モデル・ナゲットのダイアログに表示されるスコアリング・オプションのデフォルト値を設定できます。詳しくは、トピック 91 ページの『Netezza 一般化線形モデル・ナゲット -「設定」タブ』を参照してください。

- 入力フィールドを含める: 予測のほかモデル出力に入力フィールドを表示する場合、このボックスをオンにします。

Netezza ディシジョン・ツリー

ディシジョン・ツリーは、分類モデルを示す階層構造です。ディシジョン・ツリー・モデルによって、学習データのセットから将来の観測値を予測または分類する分類システムを開発できます。分類は、ブランチが分類の分割点を示すツリー構造の形式です。分割は、停止点に達するまでデータをサブグループに繰り返し分割します。停止点のツリーノードは、葉と呼ばれます。各葉は、クラス・ラベルというラベルを、サブグループのメンバー、またはクラスに割り当てます。

インスタンスの重みとクラスの重み

デフォルトでは、すべての入力レコードとクラスの重要度は相対的に等しいと想定されています。各重みをこれらの項目のいずれかまたは両方のメンバーに割り当てることによって変更できます。例えば、学習データ内のデータ・ポイントがカテゴリー間で現実的に分布していない場合に役に立ちます。重みを使用すると、モデルを偏らせて、データにうまく表れていないカテゴリーの補正を行うことができます。対象値の重みが大きくなると、カテゴリーの適切な予測の割合が大きくなります。

ディシジョン・ツリー・モデル作成ノードで、2 つの種類の重みを指定できます。インスタンスの重みは、入力データの各行に重みを割り当てます。通常、重みはほとんどの場合に 1.0 に指定され、次の表に示すように、過半数よりも重要度が高いか低いケースにのみ高い値または低い値が割り当てられます。

表 5. インスタンスの重み例

レコード **ID**	対象	インスタンスの重み
1	drugA	1.1
2	drugB	1.0
3	drugA	1.0
4	drugB	0.3

クラスの重みは、次の表に示すように対象フィールドの各カテゴリーに重みを割り当てます。

表 6. クラスの重み例

クラス	クラスの重み
drugA	1.0
drugB	1.5

2 つのタイプの重みを同時に使用できます。その場合、お互いを掛け合わせ、インスタンスの重みとして使用されます。そのため、2 つの前述の例を同時に使用すると、アルゴリズムは次の表に示すようにインスタンスの重みを使用します。

表 7. インスタンスの重みの計算例

レコード **ID**	計算	インスタンスの重み
1	1.1*1.0	1.1
2	1.0*1.5	1.5
3	1.0*1.0	1.0
4	0.3*1.5	0.45

Netezza ディシジョン・ツリーのフィールド・オプション

「フィールド」タブで、上流のノードですでに定義されているフィールドの役割設定を使用するか、手動でフィールドの割り当てを行うかを選択します。

定義済みの役割を使用**:** このオプションを選択すると、上流のデータ型ノード (または上流の入力ノードの「データ型」タブ) の役割設定 (対象、予測など) が使用されます。

カスタム・フィールド割り当ての使用: 対象、予測、およびその他の役割を手動で割り当てるには、このオプションを選択します。

フィールド**:** 矢印ボタンを使用して、このリストの項目を画面右側のさまざまな役割フィールドに手動で割り当てます。アイコンは、各役割フィールドの有効な測定の尺度を示します。

リスト内のすべてのフィールドを選択する場合は「すべて」ボタンをクリックし、特定の尺度のすべてのフィールドを選択する場合は各尺度のボタンをクリックします。

目標: 1 つのフィールドを予測の対象として選択します。

「レコード **ID**」。一意のレコード ID として使用されるフィールド。このフィールドの値は各レコードで一意である必要があります (例: カスタマー ID 番号)。

インスタンスの重み:デフォルトのクラスの重み (対象フィールドのカテゴリーあたりの重み) の代わりに、またはデフォルトに加えてインスタンスの重み (入力データの行あたりの重み) を使用できるようフィールドを指定します。ここで指定するフィールドは、入力データの各行の数値の重みを含むフィールドでなければなりません。 詳しくは、トピック 66 ページの『インスタンスの重みとクラスの重み』を参照してください。

予測値 **(**入力**):** 1 つ以上の入力フィールドを選択します。これは、データ型ノードのフィールドの役割を「入力」 に設定するのと似ています。

Netezza ディシジョン・ツリーの作成オプション

ツリーの成長には、以下の作成オプションを使用できます。

成長の測定。これらのオプションは、ツリーの成長を測定する方法を制御します。

- 不純度の測定**:** この測定では、ツリーを分割するのに最適な場所が評価されます。これは、サブグループまたはデータのセグメントにおける変動性の測定です。不純度の測定値が低い場合は、グループ内のほとんどのメンバーの基準フィールドまたは対象フィールドの値が類似していることを示します。

  サポートされる測定は、「エントロピー」と「**Gini**」です。これらの測定は、ブランチの所属カテゴリーの確率に基づいています。

- 最大ツリー深さ。 ルート・ノード下でツリーが成長可能な最大レベル数 (サンプルが再帰的に分割される回数)。このプロパティーのデフォルト値は 10 であり、このプロパティーに設定できる最大値は 62 です。

  注: モデル・ナゲットのビューアーにモデルがテキスト表示される場合、最大 12 のツリー・レベルが表示されます。

分割の基準。これらのオプションは、ツリーの分割をいつ停止するかを制御します。

- 分割の改善度の最小変化量。新しい分割がツリーで作成される前に不純度を減少する必要のある最小数。ツリー構築の目標は、似かよった出力値を持つサブグループを作成して、それぞれのノード内における不純度を最小にすることです。ブランチが適切に分割されて不純度が分割基準によって指定された値を下回ると、ブランチは分割されません。
- 分割の最小インスタンス数。分割可能な最小レコード数。分割されていないレコードがこの数より少ない場合、これ以上分割は行われません。このフィールドを使用すると、ツリー内に小さいサブグループが作成されないようにすることができます。

統計: このパラメーターは、モデルに含まれる統計の数を定義します。以下のいずれかのオプションを選択します。

- 「すべて」。列に関連したすべての統計と値に関連したすべての統計が含まれます。

  注: このパラメーターを使用すると統計の最大数が含まれるため、システムのパフォーマンスに影響する場合があります。モデルをグラフィカル形式で表示しない場合は、「なし」を指定します。
- 「列」。列に関連した統計が含まれます。
- なし: モデルのスコアリングに必要な統計のみが含まれます。

Netezza ディシジョン・ツリー・ノード - クラスの重み

ここでは、重みを各クラスに割り当てることができます。デフォルトでは、1 の値をすべてのクラスに割り当て、重みを等しくします。異なるクラス・ラベルに異なる数値の重みを指定することによって、アルゴリズムが特定のクラスの学習セットに重みを付けるよう指定します。

重みを変更するには、「重み」 列で変更する重みをダブルクリックし、必要に応じて変更します。

値。対象フィールドの可能な値から派生した、クラス ラベルのセット。

重み。特定のクラスに割り当てられる重み。大きな重みをクラスに割り当てると、モデルは、他のクラスと比べてそのクラスに対して敏感になります。

インスタンスの重みと組み合わせてクラスも重みを使用できます。 詳しくは、トピック 66 ページの『インスタンスの重みとクラスの重み』を参照してください。

Netezza ディシジョン・ツリー・ノード - ツリーの剪定

剪定オプションを使用して、ディシジョン・ツリーの剪定基準を指定できます。剪定の目的は、新しいデータに対して予期した精度が改善されない成長しすぎたサブグループを削除することによって、オーバーフィットのリスクを軽減することにあります。

剪定の測定:デフォルトの剪定の測定、「精度」は、ツリーから葉を削除した後、推定されたモデルの精度が可能な上限内にあるようにします。剪定を適用しながらクラスの重みを考慮に入れる場合、「重みつき精度」 を使用します。

剪定するデータ。新しいデータに対する想定された精度を推定するには、学習データの一部またはすべてを使用できます。また、指定されたテーブルの別の剪定データセットを使用できます。

- すべての学習データを使用。このオプション (デフォルト) は、モデルの精度を推定するためにすべての学習データを使用します。
- 剪定に次のパーセンテージ学習データを使用。 剪定データに指定した割合で、一方は学習用、一方は剪定用です。

  ストリームを実行するごとにデータを同じ方法で区分するようランダム シードを指定する場合、「結果を再現」 を選択します。「剪定に使用するシード」 フィールドで整数を指定するか、または 「生成」 をクリックすると、擬似無作為の整数を作成します。
- 既存のテーブルのデータを使用。モデルの精度を推定するために個別の剪定データセットのテーブル名を指定します。学習データを使用するより信頼性が高いと見なされます。ただし、このオプションにより、学習セットから大きなサブセットのデータが削除され、ディシジョン・ツリーの品質が損なわれます。

Netezza 線型回帰

線型モデルは、対象と 1 つ以上の予測値の間の線型関係に基づいて、連続型対象を予測します。直接の線形関係のモデリングのみに限定しながら、線型回帰モデルは比較的単純であり、簡単に解釈されるスコアリング数式を与えます。その適用はより洗練された回帰アルゴリズムで生成されるものに比べて制限されますが、線形モデルは、高速、高効率で使いやすいです。

Netezza 線型回帰作成オプション

「作成オプション」タブで、モデルを構築するすべてのオプションを設定します。もちろん、「実行」 ボタンをクリックすると、すべてデフォルト・オプションのモデルが構築されますが、通常は、それぞれの目的で作成をカスタマイズする必要があります。

方程式を解くために特異値分解を使用: 元の行列の代わりに特異値分解の行列を使用すると、数値誤差に対してより堅牢であるという利点があり、また計算をスピードアップすることができます。

モデルに切片を含む : 切片を含めると、ソリューションの全体的な精度を向上させます。

モデルの診断を計算する: このオプションを選択すると、様々な診断をモデルで計算します。結果は行列または表に格納されます。 この診断には、R-squared、残差の平方和、分散の推定、標準偏差、*p* 値、および *t* 値が含まれます。

これらの診断は、モデルの有効性と有用性に関連しています。直線性の仮定を満たしていることを保証するために、基礎となるデータに別々の診断を実行する必要があります。

Netezza KNN

最近傍分析は、そのほかのケースに対する類似性に基づいてケースを分類する方法です。マシン学習で、保存されたパターン、またはケースへに完全に一致する必要なくデータのパターンを認識する方法として開発されました。類似したケースはお互いに近く、類似していないケースはお互いに離れています。つまり、2 つのケース間の距離は、それらの非類似度の尺度です。

お互いに近いケースは、「近傍」と呼ばれます。新しいケース (ホールドアウト) が表示されたときに、モデル内の各ケースからの距離が計算されます。最も類似した分類 - 最近傍 - が集計され、新しいケースが、最大数の最近傍を含むカテゴリーに振り分けられます。

検証する最近傍の数を指定できます。この値は *k* となります。図は、新しいケースが 2 つの異なる値の *k* を使用してどのように分類されるかを示します。*k* = 5 の場合、最近傍の大部分はカテゴリー *1* に属するため、新しいケースはカテゴリー *1* にあります。ただし *k* = 9 の場合、最近傍の大部分はカテゴリー *0* に属するため、新しいケースはカテゴリー *0* にあります。

また、最近傍分析を使用して、連続型対象の値を計算することもできます。この場合、最近傍の平均対象値または中央対象値を使用して、新しいケースの予測値が取得されます。

Netezza KNN モデル・オプション - 全般

「モデル・オプション - 全般」タブで、モデルの名前を指定するか、自動的に名前を生成するかを選択できます。また、最近傍の数が計算される方法を制御するオプションを設定し、モデルの強化されたパフォーマンスと精度のオプションを設定できます。

モデル名**:** 対象または ID フィールド (そのようなフィールドが指定されていない場合はモデル タイプ) に基づいてモデル名を自動的に生成するか、カスタム名を指定することができます。

近傍

距離測度 :データ・ポイント間の距離の測定に使用する方法です。距離が大きいほど、相違が大きくなります。オプションは次のとおりです。

- ユークリッド:(デフォルト) 2 点間の距離は、それらを直線で結ぶことによって計算されます。
- **Manhattan**:2 点間の距離は、それらの座標間の絶対距離の合計として計算されます。
- **Canberra**:Manhattan の距離と同じですが、原点に近いデータ・ポイントに対してより感度が高くなります。
- 最大値 :2 点間の距離は、座標の次元に沿ったそれらの相違の最大値として計算されます。

最近傍数 **(k)**: 特定のケースの最近傍の数。より大きな数の近傍を使用すると、必ずしも正確なモデルが作成されるとは限りません。

k を選択することにより、オーバーフィット（特に「ノイズの多い」データの場合は重要な場合があります）の防止と解決（同様のインスタンスに対する異なる予測）のバランスを制御します。各データセットごとに *k* の値 (値は 1〜 数十の範囲) を調整する必要があります

パフォーマンスと精度を向上

距離を計算する前に測定値を標準化**:** 選択した場合、このオプションは、距離の値を計算する前に連続型入力フィールドの測定を標準化します。

コアセットを使用して大規模なデータセットのパフォーマンスを向上させる**:** このオプションは、大規模なデータセットを使用する場合、計算を高速化するためにコアセット・サンプリングを使用します。

Netezza KNN モデル・オプション - スコアリング・オプション

「モデル・オプション - スコアリング・オプション」タブでは、スコアリング・オプションのデフォルト値を設定し、個々のクラスに相対的な重みを割り当てることができます。

スコアリングに使用できるようにする

入力フィールドを含める:入力フィールドがデフォルトでスコアリングに含まれるかどうかを指定します。

クラスの重み

モデルを構築するの個々のクラスの相対的な重要度を変更したい場合は、このオプションを使用します。

注 : このオプションは、分類に KNN を使用している場合にのみ有効です。回帰を実行している場合 (つまり、対象フィールドの型が連続型である場合)、オプションは無効になります。

デフォルトでは、1 の値をすべてのクラスに割り当て、重みを等しくします。異なるクラス・ラベルに異なる数値の重みを指定することによって、アルゴリズムが特定のクラスの学習セットに重みを付けるよう指定します。

重みを変更するには、「重み」 列で変更する重みをダブルクリックし、必要に応じて変更します。

値。対象フィールドの可能な値から派生した、クラス ラベルのセット。

重み。特定のクラスに割り当てられる重み。大きな重みをクラスに割り当てると、モデルは、他のクラスと比べてそのクラスに対して敏感になります。

Netezza K-Means

K-Means ノードは、クラスター分析の方法を提供する *k*-means アルゴリズムを実行します。このノードを使用して、データ・セットをグループにクラスター化できます。

アルゴリズムは距離に基づくアルゴリズムで、距離メトリック (関数) に依存してデータ・ポイント間の類似性を測定します。データ・ポイントは、使用する距離メトリックに従って最も近いクラスターに割り当てられます。

アルゴリズムは、同じ基本プロセスを何度か反復することによって実行します。学習インスタンスは最も近いクラスターに割り当てられます (指定された距離関数に関しては、インスタンスとクラスター中心に適用されます)。すべてのクラスター中心は、特定のクラスターに割り当てられたインスタンスの平均属性値のベクトルとして再計算されます。

Netezza K-means のフィールド・オプション

「フィールド」タブで、上流のノードですでに定義されているフィールドの役割設定を使用するか、手動でフィールドの割り当てを行うかを選択します。

事前定義された役割を使用: 上流のデータ型ノード (または上流の入力ノードの「データ型」タブ) の役割設定 (対象、予測など) を使用します。

カスタム・フィールド割り当ての使用: この画面で対象、予測、およびその他の役割を手動で割り当てる場合、このオプションを選択します。

フィールド**:** 矢印ボタンを使用して、このリストの項目を画面右側のさまざまな役割フィールドに手動で割り当てます。アイコンは、各役割フィールドの有効な測定の尺度を示します。

「すべて」 ボタンをクリックしてリスト内のすべてのフィールドを選択するか、各測定の尺度のボタンをクリックして、その測定の尺度のすべてのフィールドを選択します。

「レコード **ID**」。一意のレコード ID として使用されるフィールド。

予測値 **(**入力**):** 1 つ以上のフィールドを予測の入力として選択します。

Netezza K-Means の作成オプション・タブ

作成オプションを設定することにより、用途に合わせてモデルの作成をカスタマイズできます。

デフォルト・オプションを使用してモデルを作成する場合は、「実行」をクリックします。

距離測度**:** このパラメーターは、データ・ポイント間の距離の測定方法を定義します。距離が大きくなると、非類似度も大きくなります。以下のいずれかのオプションを選択します。

- ユークリッド: ユークリッド測定は、2 つのデータ・ポイント間の直線距離です。
- 正規化ユークリッド **(Normalized Euclidean)**。正規化ユークリッド測定はユークリッド測定に類似していますが、平方標準偏差によって正規化されます。ユークリッド測定とは異なり、正規化ユークリッド測定はスケール不変でもあります。
- **Mahalanobis**。Mahalanobis 測定は、入力データの相関を考慮に入れた一般化ユークリッド測定です。正規化ユークリッド測定と同様に、Mahalanobis 測定はスケール不変です。
- **Manhattan**。Manhattan 測定は 2 つのデータ・ポイント間の距離で、それらの座標間の絶対距離の合計として計算されます。
- **Canberra**。Canberra 測定は Manhattan 測定に似ていますが、原点に近いデータ・ポイントに対してより感度が高くなります。
- 最大値: 最大測定は、2 つのデータ・ポイント間の距離で、座標の次元に沿ったそれらの相違の最大値として計算されます。

クラスター数**:** このパラメーターは、作成するクラスターの数を定義します。

最大反復数**:** アルゴリズムは、同じプロセスを何度か反復します。このパラメーターは、モデルの学習を停止する前の反復数を定義します。

統計: このパラメーターは、モデルに含まれる統計の数を定義します。以下のいずれかのオプションを選択します。

- 「すべて」。列に関連したすべての統計と値に関連したすべての統計が含まれます。

  注**:** このパラメーターを使用すると統計の最大数が含まれるため、システムのパフォーマンスに影響する場合があります。モデルをグラフィカル形式で表示しない場合は、「なし」を指定します。
- 「列」。列に関連した統計が含まれます。
- なし**:** モデルのスコアリングに必要な統計のみが含まれます。

結果の再現: ランダム・シードを設定して分析を複製する場合は、このチェック・ボックスを選択します。整数を指定するか、「生成」をクリックして整数の疑似乱数を作成します。

Netezza Naive Bayes

Naive Bayes は、分類の問題に対応する有名なアルゴリズムです。提示されたすべての予測変数は相互に依存関係がないものとして処理されるので、モデルは *naïve* と命名されます。Naive Bayes は拡張性のある高速のアルゴリズムであり、複数の属性と対象属性の組み合わせに対して、条件付きの確率を計算します。学習データから、個別の確率が計算されます。各入力変数の各値カテゴリーを計算単位とすると、この確率は各対象クラスの確率を表します。

Netezza ベイズ・ネットワーク

Bayesian Network は、モデルで、データセットの変数およびそれらの間の確率的、または条件付き独立性が表示されます。Netezza Bayes Net ノードを使用すると、観測された情報および記録された情報を「常識」という実際の知識を組み合わせることによって確率モデルを作成し、表面的にはリンクしていない属性を使用して発生の尤度を確立できます。

Netezza Bayes ネットワークのフィールド・オプション

「フィールド」タブで、上流のノードですでに定義されているフィールドの役割設定を使用するか、手動でフィールドの割り当てを行うかを選択します。

このノードの場合、対象フィールドはスコアリングのためにのみ必要ですので、このタブには表示されません。このノードの「モデルのオプション」タブで、またはモデルナゲットの「設定」タブで、データ型ノード上で対象を設定または変更することができます。 詳しくは、トピック 85 ページの『Netezza Bayes Net ナゲット -「設定」タブ』を参照してください。

事前定義された役割を使用: 上流のデータ型ノード (または上流の入力ノードの「データ型」タブ) の役割設定 (対象、予測など) を使用します。

カスタム・フィールド割り当ての使用: この画面で対象、予測、およびその他の役割を手動で割り当てる場合、このオプションを選択します。

フィールド: 矢印ボタンを使用して、このリストの項目を画面右側のさまざまな役割フィールドに手動で割り当てます。アイコンは、各役割フィールドの有効な測定の尺度を示します。

「すべて」 ボタンをクリックしてリスト内のすべてのフィールドを選択するか、各測定の尺度のボタンをクリックして、その測定の尺度のすべてのフィールドを選択します。

予測値 (入力): 1 つまたは複数のフィールドを予測の入力として選択します。

Netezza Bayes ネットワークの作成オプション

「作成オプション」タブで、モデルを構築するすべてのオプションを設定します。もちろん、「実行」 ボタンをクリックすると、すべてデフォルト・オプションのモデルが構築されますが、通常は、それぞれの目的で作成をカスタマイズする必要があります。

ベース インデックス: 数値型識別子は、より簡単な内部管理の最初の属性（入力フィールド）に割り当てられる。

サンプル サイズ: 属性の数が多く、非常に長い処理時間を引き起こすことがある場合に採用されるサンプルのサイズ。

実行時に追加の情報を表示: このボックスをオンにすると (デフォルト)、追加情報がメッセージ ダイアログ・ボックスに表示されます。

Netezza 時系列

時系列は、毎日の株価や毎週の営業データなど、連続する (必ずしも正規ではない) 時点で測定した数値データ値のシーケンスです。傾向や季節性（繰り返しパターン）などの行動を強調したり、過去の事象から将来の行動を予測する場合などに、このようなデータの分析が役立ちます。

Netezza 時系列は、次の時系列アルゴリズムをサポートします。

- スペクトル解析
- 指数平滑法
- 自己回帰和分移動平均 (ARIMA)
- 季節的傾向分解

これらのアルゴリズムは、時系列を傾向と季節成分に分類します。予測に使用できるモデルを構築するために、これらのコンポーネントを分析します。

スペクトル解析を使用して時系列の定期的な動作を識別します。複数の基底の周期で構成された時系列の場合、または相当数のランダム ノイズがデータ内に存在する場合、スペクトル解析により、最も明らかに周期的なコンポーネントを特定することができるようになります。この方法により、周波数領域の系列に時間領域からシリーズを変換することによって、定期的な動作の周波数を検出します。

指数平滑化は、前の時系列の観測結果に重み付けされた値を使用して将来の値を予測する方法です。指数平滑化により、観測値の影響は指数関数的な方法で時間の経過に従って短くなります。このメソッドは、新しいデータが入ってくるごとに予測を調整、その追加、傾向、季節性を考慮し、一度に1つのポイントを予想します。

ARIMA モデルは、指数平滑化モデルより洗練された方法で傾向および季節性のコンポーネントをモデリングします。これは、差異の程度と同様に自己回帰および移動平均の順序を明示して指定することと関連します。

注：実際面では、郵送するカタログの数または会社の Web ページのヒット数など予測対象の一連の性質を説明する上で役立つ予測値を含める場合は、ARIMA モデルが最も有用です。指数平滑化モデルは、性質や傾向の理由を理解しようとしないで、時系列の性質や傾向を記述します。

季節的傾向分解では、傾向分析を実行するために、時系列からの定期的な動作を削除し、二次関数などのトレンドの基本的な形状を選択します。これらの基本的な形状は、値が残差の平均二乗誤差（つまり、時系列の近似と実測値の違）を最小化するように決定されるさまざまなパラメーターがあります。

Netezza 時系列の値の補間

補間は、時系列データの欠損値の推定および挿入のプロセスです。

時系列の間隔が定期的で、いくつかの値が存在しない場合、欠損値は線形補間を用いて推定することができます。空港での毎月の乗客数について考えてみましょう。

表 *8.* 空港ターミナルの月間利用者数

月	乗客
3	3,500,000
4	3,900,000
5	-
6	3,400,000

表 8. 空港ターミナルの月間利用者数 (続き)

月	乗客
7	4,500,000
8	3,900,000
9	5,800,000
10	6,000,000

この場合、線形補間は月 5 の欠損値を 3,650,000 (月 4 と月 6 の中間点) と推定します。

不定期な間隔は別の方法で処理されます。次に、温度読み取りについて考えてみましょう。

表 9. 温度読み取り

日付	時間	温度
2011-07-24	7:00	57
2011-07-24	14:00	75
2011-07-24	21:00	72
2011-07-25	7:15	59
2011-07-25	14:00	77
2011-07-25	20:55	74
2011-07-27	7:00	60
2011-07-27	14:00	78
2011-07-27	22:00	74

ここでは、3 日間で 3 つのポイントの読み取り値がありますが、さまざまな場合で、そのうちのいくつかだけがは数日の間で共通です。また、数日のうち 2 日だけが連続します。

この状況は、集計の計算、ステップ・サイズの決定のいずれかの方法で処理できます。

データ集計は、データの意味的情報に基づいた式に従って計算された毎日の集計です。これにより、次のようなデータセットが生成されます。

表 10. 温度読み取り (集計)

日付	時間	温度
2011-07-24	24:00	69
2011-07-25	24:00	71
2011-07-26	24:00	ヌル
2011-07-27	24:00	72

また、このアルゴリズムは、異なる時系列を異なるものとして扱い、適切なステップ・サイズを決定することができます。この場合、アルゴリズムによって決定されたステップ・サイズは 8 時間となり、次のようになります。

表 11. ステップ・サイズを計算した温度の読み取り

日付	時間	温度
2011-07-24	6:00	

表 *11.* ステップ・サイズを計算した温度の読み取り *(*続き*)*

日付	時間	温度
2011-07-24	14:00	75
2011-07-24	22:00	
2011-07-25	6:00	
2011-07-25	14:00	77
2011-07-25	22:00	
2011-07-26	6:00	
2011-07-26	14:00	
2011-07-26	22:00	
2011-07-27	6:00	
2011-07-27	14:00	78
2011-07-27	22:00	74

ここでは、4つの測定値だけが、元の測定に対応していますが、元の時系列の他の値によって、欠損値が再度補間により計算できるようになります。

Netezza 時系列フィールド・オプション

「フィールド」タブで、ソース・データの入力フィールドの役割を指定します。

フィールド**:** 矢印ボタンを使用して、このリストの項目を画面右側のさまざまな役割フィールドに手動で割り当てます。アイコンは、各役割フィールドの有効な測定の尺度を示します。

目標: 1 つのフィールドを予測の対象として選択します。尺度が連続型のフィールドでなければなりません。

(予測値**)** 時点**:** (必須) 時系列の日付または時刻の値を含む入力フィールド。このフィールドには、測定の尺度が連続型またはカテゴリー型のフィールド、および日付、時間、タイム・スタンプ、または数値のデータ・ストレージ・タイプを指定する必要があります。ここで指定したフィールドのデータ・ストレージ・タイプは、このモデリング・ノードの他のタブでいくつかのフィールドの入力タイプも定義します。

(予測値**)** 時系列 **ID:** 時系列 ID を含むフィールド。入力に複数の時系列が含まれる場合に使用します。

Netezza 時系列構築オプション

構築オプションには 2 つのレベルがあります。

- 基本 - アルゴリズムの選択、補間、使用する時間の範囲の設定
- 高度 - 予測の設定

このセクションでは、基本オプションについて説明します。

「作成オプション」タブで、モデルを構築するすべてのオプションを設定します。もちろん、「実行」 ボタンをクリックすると、すべてデフォルト・オプションのモデルが構築されますが、通常は、それぞれの目的で作成をカスタマイズする必要があります。

アルゴリズム

これらは、使用する時系列アルゴリズムに関連する設定です。

アルゴリズム名**:** 使用する時系列アルゴリズムを選択します。利用可能なアルゴリズムは、「スペクトル解析」、「指数平滑化」 (デフォルト)、「**ARIMA**」、または 「季節的傾向分解」 です。 詳しくは、トピック 74 ページの『Netezza 時系列』を参照してください。

傾向**:** (指数平滑化のみ) 時系列の傾向を示す場合は、単純な指数平滑化は適切に実行されません。このフィールドを使用して傾向があれば指定し、アルゴリズムがその傾向を考慮できるようにします。

- 決定システム**:** (デフォルト) システムは、このパラメーターの最適値を見つけようとします。
- なし **(N):** 時系列は傾向を示しません。
- 相加的 **(A):** 時間をかけて着実に増加する傾向。
- 減衰相加的**(DA):** 最終的に消失する相加的傾向。
- 乗法的**(M):** 安定した相加的傾向より通常速く、時間を経て増加する傾向。
- 減衰乗法的**(DM):** 最終的に消失する乗法的傾向。

季節性**:** (指数平滑化のみ) 時系列データの季節的パターンを示すかどうか指定するには、このフィールドを使用します。

- 決定システム**:** (デフォルト) システムは、このパラメーターの最適値を見つけようとします。
- なし **(N):** 時系列は季節的パターンを示しません。
- 相加的 **(A):** 季節変動のパターンは、時間をかけて着実な上昇傾向を示します。
- 乗法的**(M):** 相加的季節性と同様、ただしさらに季節変動の振り幅 (最高点と最低点間の距離) が変動の全体の上方傾向に相対的に増加します。

ARIMA のシステム決定システムを使用**:** (ARIMA のみ) ARIMA アルゴリズムの設定を決定したい場合、このオプションを選択します。

指定 **:** (ARIMA のみ) ARIMA 設定を手動で指定する場合、このオプションを選択してボタンをクリックします。

補間

時系列ソース・データに欠損値がある場合、推定値を挿入する方法を選択して、データの隙間を埋めます。詳しくは、トピック 74 ページの『Netezza 時系列の値の補間』を参照してください。

- 線形**:** 時系列の間隔が規則的であるが、いくつかの値が存在しない場合、この方法を選択します。
- 指数スプライン**:** 滑らかなカーブを既知のデータ・ポイントが高い速度で増加または減少する場所に適合させます。
- 三次スプライン**:** 滑らかなカーブを既知のデータ・ポイントに適合させ、欠損値を推定します。

時間範囲

ここでは、時系列の完全なデータの範囲、またはそのデータの連続したサブセットを使用してモデルを作成するかどうかを選択できます。これらのフィールドの有効な入力は、「フィールド」タブの「時点」で指定されたフィールドのデータ・ストレージ・タイプによって定義されます。 詳しくは、トピック 76 ページの『Netezza 時系列フィールド・オプション』を参照してください。

- データで可能な最も早い時間及び最も遅い時間を使用**:** 時系列データの全体の範囲を使用する場合、このオプションを選択します。
- 時間枠を指定**:** 時系列データの一部だけを使用する場合、このオプションを選択します。「最も早い時間 **(**開始**)**」 と 「最も遅い時間 **(**終了**)**」 フィールドを使用して境界を指定します。

ARIMA 構造

ARIMA モデルの様々な非季節性および季節性要素の値を指定します。それぞれの場合において、演算子を = (等しい) または <= (以下) に設定して、隣接するフィールドに値を指定します。すべての値は、程度を示す負でない整数にする必要があります。

非季節性**:** モデルの様々な非季節性要素の値。

- 自己相関度 **(p)**。 モデル内の自己回帰の次数の数値です。自己回帰の次数は、系列の使用する過去の値を指定し、現在の値を予測します。例えば、自己回帰の次数 2 は、現在の値を予測するために系列の値を過去の 2 期間使用するように指定します。
- 誘導 **(d)**。モデルを推定する前に系列に適用される差分の次数を指定します。トレンドが存在する場合は差分を取る必要があり (トレンドの存在する系列は通常非定常性であり、ARIMA モデルは定常性を前提としている)、その効果を取り除くために行います。差分の次数は、系列のトレンドの次数に対応しています (1 次差分は線型トレンドを表し、2 次差分は 2 次トレンドを表す、など)。
- 移動平均 **(q)**。モデル内の移動平均の次数の数値。移動平均の次数は、過去の値の系列平均の偏差が、現在の値を予測するためにどのように使用されるかを指定します。例えば、移動平均の次数 1 および 2 は、系列の現在の値を予測する際に最近の 2 期間のそれぞれから取得した系列の平均値の偏差を考慮することを指定します。

季節性**:** 季節性自己相関 (SP)、導出 (SD)、移動平均 (SQ) の要素は、非季節性のこれらの要素と同じ役割を果たします。ただし、季節次数については、現在の系列値は 1 つ以上の季節期間で区切られた過去の系列値に影響されます。例えば、毎月のデータ (季節期間 12) については、季節次数 1 は、現在の系列値は現在の期間より 12 期間以前の系列値により影響されることを意味しています。毎月のデータについて、季節次数 1 は、非季節次数 12 を指定するのと同じことになります。

季節性の設定は、季節性がデータで検出された場合、または「詳細」タブの「期間」設定を指定した場合にのみ考慮されます。

Netezza 時系列構築オプション - 詳細設定

詳細設定を使用して、予測のオプションを指定できます。

モデル構築オプションのシステム決定システムを使用**:** 詳細設定を決定する場合、このオプションを選択します。

指定 **:** 詳細オプションを手動で指定する場合にこのオプションを選択します。(アルゴリズムがスペクトル解析の場合、このオプションは無効です)

- 期間/期間の単位**:** 時系列の特徴的な動作が反復する期間。例えば、週間営業成績の時系列について、期間に 1 を指定し、単位に「週」を指定します。「期間」には、負ではない整数を指定する必要があります。「期間の単位」には、「ミリ秒」、「秒」、 「分」、「時間」、「日」、「週」、「四半期」、「年」のいずれかを指定します。「期間」 が設定されている場合、時間のタイプが数値でない場合は「期間の単位」 を指定しないでください。ただし、「期間」 を指定する場合、「期間の単位」 を指定する必要があります。

予測の設定**:** 特定の時点まで、または特定の時点での予測を行うことを選択できます。これらのフィールドの有効な入力は、「フィールド」タブの「時点」で指定されたフィールドのデータ・ストレージ・タイプによって定義されます。 詳しくは、トピック 76 ページの『Netezza 時系列フィールド・オプション』を参照してください。

- 予測値の境界 **:** 予測の終了ポイントの実を指定する場合、このオプションを選択します。予測はこの時点まで行われます。

- 予測時間 : 予測を行う 1 つまたは複数の時点を指定するには、このオプションを選択します。「追加」 をクリックして、時点のテーブルに新しい行を追加します。行を削除するには、行を選択して「削除」 をクリックします。

Netezza 時系列のモデル・オプション

「モデル・オプション」タブで、モデルの名前を指定するか、自動的に名前を生成するかを選択できます。モデル出力のオプションのデフォルト値を設定することもできます。

モデル名: 対象または ID フィールド (そのようなフィールドが指定されていない場合はモデル タイプ) に基づいてモデル名を自動的に生成するか、カスタム名を指定することができます。

スコアリングに使用できるようにする。モデル・ナゲットのダイアログに表示されるスコアリング・オプションのデフォルト値を設定できます。

- 結果に過去の値を含める: デフォルトでは、モデル出力に過去のデータ値 (予測に使用された値) は含まれません。これらの値を含めるにはこのチェックボックスをオンにします。
- 結果に補間値を含める: 出力に過去の値を含めると選択した場合、補間値があれば含めるときはこのボックスをオンにします。補間は過去のデータにのみ機能するため、「結果に過去の値を含める」 が選択されていない場合このボックスは無効です。 詳しくは、トピック 74 ページの『Netezza 時系列の値の補間』を参照してください。

Netezza TwoStep

TwoStep ノードは、大規模データセットにわたってデータをクラスタ化する方法を提供する TwoStep アルゴリズムを実装します。

このノードを使用すると、使用可能なリソース (例えば、メモリーや時間の制約) を考慮しながら、データをクラスタ化できます。

TwoStep アルゴリズムは、次の方法でデータをクラスタ化するデータベース マイニング アルゴリズムです。

1. クラスタ機能 (CF) ツリーが作成されます。このバランスに優れたツリーは、類似の入力レコードが同じツリー ノードの一部となる階層クラスタリングのためのクラスタ機能を格納します。
2. CF ツリーの葉は、最終クラスタリングの結果を生成するためにメモリー内で階層的にクラスタ化されます。クラスタの最適な数は自動的に決定されます。クラスタの最大数を指定する場合、指定された制限内のクラスタの最適な数が決定されます。
3. クラスタリングの結果は、K-Means アルゴリズムに似たアルゴリズムがデータに適用される 2 番目のステップで調整されます。

Netezza TwoStep フィールド・オプション

フィールド オプションを設定することにより、上流ノードで定義されているフィールドの役割の設定を指定できます。フィールド割り当てを手動で行うこともできます。

項目の選択。上流のデータ型ノードまたは上流の入力ノードの「データ型」タブの役割設定を使用するには、このオプションを選択します。役割設定は、例えば対象や予測です。

カスタム・フィールド割り当ての使用: 対象、予測、およびその他の役割を手動で割り当てる場合、このオプションを選択します。

フィールド**:** 矢印を使用して、このリストの項目を右側の役割フィールドに手動で割り当てます。アイコンは、各役割フィールドの有効な測定の尺度を示します。

「レコード **ID**」。一意のレコード ID として使用されるフィールド。

予測値 **(**入力**):** 1 つまたは複数のフィールドを予測の入力として選択します。

Netezza TwoStep 作成オプション

作成オプションを設定することにより、用途に合わせてモデルの作成をカスタマイズできます。

デフォルト・オプションを使用してモデルを作成する場合は、「実行」をクリックします。

距離測度**:** このパラメーターは、データ・ポイント間の距離の測定方法を定義します。距離が大きくなると、非類似度も大きくなります。以下のオプションがあります。

- 対数尤度: この尤度測定により、変数の確率分布を求めます。連続型変数は正規分布しているものと仮定され、カテゴリー変数は多項分布しているものと仮定されます。すべての変数は独立しているものと仮定します。
- ユークリッド: ユークリッド測定は、2 つのデータ・ポイント間の直線距離です。
- 正規化ユークリッド **(Normalized Euclidean)**。正規化ユークリッド測定はユークリッド測定に類似していますが、平方標準偏差によって正規化されます。ユークリッド測定とは異なり、正規化ユークリッド測定はスケール不変でもあります。

クラスタ数。このパラメーターは、作成するクラスターの数を定義します。以下のオプションがあります。

- クラスタ数を自動的に計算。クラスタの数が自動的に計算されます。「最大」フィールドで最大クラスタ数を指定できます。
- クラスタ数を指定。作成するクラスタの数を指定します。

統計: このパラメーターは、モデルに含まれる統計の数を定義します。以下のオプションがあります。

- 「すべて」。列に関連したすべての統計と値に関連したすべての統計が含まれます。

  注**:** このパラメーターを使用すると統計の最大数が含まれるため、システムのパフォーマンスに影響する場合があります。モデルをグラフィカル形式で表示しない場合は、「なし」を指定します。
- 「列」。列に関連した統計が含まれます。
- なし**:** モデルのスコアリングに必要な統計のみが含まれます。

結果の再現: ランダム・シードを設定して分析を複製する場合は、このチェック・ボックスを選択します。整数を指定するか、「生成」をクリックして整数の疑似乱数を作成します。

Netezza PCA

主成分分析（PCA）は、データの複雑さを軽減するために設計された強力なデータ削減技術です。入力フィールドの線型結合が検出されます。成分が互いに直交する (相関しない) 場合に、フィールドのセット全体の分散を把握するのに役立ちます。どちらの手法でも、元の入力フィールド・セットの情報を効果的に要約する少数の派生フィールド (主成分) の検出が目標です。

Netezza PCA フィールド・オプション

「フィールド」タブで、上流のノードですでに定義されているフィールドの役割設定を使用するか、手動でフィールドの割り当てを行うかを選択します。

事前定義された役割を使用: 上流のデータ型ノード (または上流の入力ノードの「データ型」タブ) の役割設定 (対象、予測など) を使用します。

カスタム・フィールド割り当ての使用: この画面で対象、予測、およびその他の役割を手動で割り当てる場合、このオプションを選択します。

フィールド**:** 矢印ボタンを使用して、このリストの項目を画面右側のさまざまな役割フィールドに手動で割り当てます。アイコンは、各役割フィールドの有効な測定の尺度を示します。

「すべて」 ボタンをクリックしてリスト内のすべてのフィールドを選択するか、各測定の尺度のボタンをクリックして、その測定の尺度のすべてのフィールドを選択します。

「レコード **ID**」。一意のレコード ID として使用されるフィールド。

予測値 **(**入力**):** 1 つ以上のフィールドを予測の入力として選択します。

Netezza PCA 作成オプション

「作成オプション」タブで、モデルを構築するすべてのオプションを設定します。もちろん、「実行」 ボタンをクリックすると、すべてデフォルト・オプションのモデルが構築されますが、通常は、それぞれの目的で作成をカスタマイズする必要があります。

PCA を計算する前にデータを集約する**:** （デフォルト）このオプションをチェックした場合、分析前にデータのセンタリングを（または「平均値減算」）を実行します。データ・センタリングは、第 1 主成分が最大分散の方向を記述することを保証するために必要です。そうでない場合、成分がデータの平均値により密接に対応している場合があります。データがすでにこの方法で調製されている場合、通常パフォーマンスの向上のためだけにこのオプションを無効にします。

PCA を計算する前にデータのスケーリングを実行する**:** このオプションは、分析前にデータのスケーリングを行います。そうすることで、別の変数が異なる単位で測定されるとき、分析が恣意的でないようにします。最も単純な形式のデータのスケーリングは、その標準偏差で各変数を分割することによって実現できます。

PCA を計算する精度が低くなってもより高速な方法を使用**:** このオプションは、アルゴリズムが主成分を見つける精度は低くなりますが、より高速な方法（forceEigensolve）を使用することになります。

IBM Netezza Analytics モデルの管理

IBM Netezza Analytics モデルは、他の IBM SPSS Modeler のモデルのようにキャンバスおよび「モデル」パレットに追加され、ほとんど同じように使用できます。ただし、IBM SPSS Modeler 中で生成された各 IBM Netezza Analytics モデルは実際にはデータベース・サーバー上にあるモデルへの参照であるなどの、重要な違いがいくつかあります。ストリームを正しく機能させるためには、モデルが作成されたデータベースに接続する必要があります。モデル テーブルを外部プロセスにより変更してはいけません。

スコアリング IBM Netezza Analytics モデル

モデルは、キャンバス上で金色のモデル・ナゲット・アイコンで示されます。ナゲットの主な目的は、データをスコアリングし、予測を生成、またはモデルのプロパティーの詳細な分析を可能にすることです。スコアは、このセクションで後述するように、ナゲットにテーブル・ノードを接続し、ストリームのそのブランチを実行することによって見えるようにすることができる、1 つ以上の追加データ・フィールドの形式で追加されます。デシジョンツリーや回帰ツリーのダイアログ・ボックスなど、いくつかのナゲットのダイアログボックスには、さらにモデルの視覚的表現を提供する「モデル」タブがあります。

追加フィールドは、対象フィールドの名前に追加された接頭辞 $<id>- によって区別されます。 <id> はモデルによって異なり、追加される情報の種類を識別します。さまざまな識別子について、それぞれのモデル・ナゲットのトピックで説明されています。

スコアを表示するには、次の手順を実行します。

1. テーブルノードをモデル・ナゲットに接続します。
2. テーブル・ノードを開きます。
3. 「実行」をクリックします。
4. テーブル出力ウィンドウの右側にスクロールし、追加フィールドとすれらのスコアを表示します。

Netezza モデル・ナゲットの「サーバー」タブ

「サーバー」タブでは、モデルのスコアリングサーバのオプションを設定することができます。上流に指定されたサーバの接続を使い続けるか、ここで指定した別のデータベースにデータを移動することができます。

Netezza DB サーバーの詳細:ここで、モデルに使用するデータベースの接続の詳細を指定します。

- 上流の接続を使用: (デフォルト) データベース入力ノードなど、上流のノードで指定した接続の詳細を使用します。注 : このオプションは、すべての上流ノードが SQL プッシュバックを使用できる場合にのみ有効です。この場合、SQL が完全にすべての上流ノードを実装するため、データベース外のデータを移動する必要はありません。
- 接続するデータを移動:ここでしてデータベースにデータを移動します。データを移動することにより、データが別の IBM Netezza データベース、他のベンダーのデータベースにある場合、またはデータがフラット・ファイルにある場合でもモデリングを実行できるようにします。また、ノードが SQL プッシュバックを実行していないためデータが抽出されていない場合、データはここで指定されたデータベースに戻されます。接続を参照して選択するには、「編集」ボタンをクリックします。注意 : IBM Netezza Analytics は通常非常に大きいデータ・セットで使用されます。データベース間、またはデータベース内外に多くのデータを移行するのは非常に時間がかかる場合があるため、可能な限り避けることをお勧めします。

モデル名**:** モデルの名前。この名前は通知専用に表示されます。ここで名前を変更することはできません。

Netezza ディシジョン・ツリー・モデル・ナゲット

ディシジョン・ツリー・モデル・ナゲットでは、モデリング操作の出力を表示し、モデルをスコアリングするためのオプションを設定することもできます。

デシジョン・ツリー・モデル・ナゲットを含むストリームを実行すると、デフォルトでは、ノードは 1 つの新しいフィールドを追加します。その名前は、対象名に由来します。

表 *12.* ディシジョン・ツリーのモデルスコアリング・フィールド :

追加フィールドの名前	意味
$I-*target_name*	現在のレコードの予測値。

モデリング・ノードまたはモデル・ナゲットで 「レコードのスコアリングに割り当てられたクラスの確率を計算する」 を選択した場合、さらにフィールドを追加します。

表 13. ディシジョン・ツリーのモデルスコアリング・フィールド - 追加：

追加フィールドの名前	意味
$IP-*target_name*	予測の確信度 (0.0 ～ 1.0)。

Netezza ディシジョン・ツリー -「モデル」タブ

「モデル」タブには、ディシジョン・ツリー・モデルの「予測値の重要度」がグラフィカル形式で表示されます。棒の長さは、予測値の重要度を表しています。

注: IBM Netezza Analytics バージョン 2.x 以前で作業している場合、ディシジョン・ツリー・モデルの内容はテキスト形式でのみ表示されます。

これらのバージョンの場合、以下の情報が表示されます。

- テキストの各行は、ノードまたは葉に対応しています。
- インデントはツリー・レベルを反映します。
- ノードの場合、分割の条件が表示されます。
- 葉の場合、割り当てられたクラスのラベルが表示されます。

Netezza ディシジョン・ツリー -「設定」タブ

「設定」タブで、モデルをスコアリングするためのオプションを設定できます。

入力フィールドを含む:選択すると、元の入力フィールドをすべて下流に渡し、追加のモデリング・フィールドをデータの各行に添付します。このボックスをオフにすると、レコード ID フィールドと追加のモデリング・フィールドのみが渡されるため、ストリームがより早く実行されます。

レコードのスコアリングに割り当てられたクラスの確率を計算する:(ディシジョン・ツリーおよび Naive Bayes のみ) 選択した場合、このオプションは、追加のモデリング・フィールドが確信（確率）フィールドのほか、予測フィールドを含めることを意味します。このチェックボックスをオフにすると、予測のみのフィールドが生成されます。

決定的入力データの使用。このオプションが選択された場合、同じビューの複数のパスを実行するすべての Netezza アルゴリズムが各パスに確実に同じデータ セットを使用するようになります。非決定的データが使用されていることを示すためにこのチェック ボックスをクリアすると、データ区分ノードによって作成されるものなど、処理するデータ出力を保持する一時テーブルが作成されます。このテーブルは、モデルの作成後に削除されます。

Netezza デシジョン ツリー ナゲット -「ビューア」タブ

「ビューア」タブには、SPSS Modeler でデシジョン ツリー モデルが表示されるのと同じ方法で、ツリー モデルのツリー プレゼンテーションが表示されます。

注: モデルが IBM Netezza Analytics バージョン 2.x 以前で構築されている場合、「ビューア」タブは空です。

Netezza K-Means モデル・ナゲット

K-Means モデル・ナゲットには、クラスター化モデルが取得したすべての情報と、学習データと推定プロセスに関する情報が含まれます。

K-Means モデル・ナゲットを含むストリームを実行すると、そのレコードの所属クラスターと割り当てられたクラスターの中心からの距離を含む 2 つの新規フィールドが追加されます。$KM-K-Means という名前の新規フィールドは所属クラスタ用で、$KMD-K-Means という名前の新規フィールドはクラスタの中心からの距離用です。

Netezza K-Means ナゲット -「モデル」タブ

「モデル」タブには、クラスターのフィールドの要約統計量および分布を表示する各種のグラフィック表示があります。モデルからデータをエクスポートしたり、ビューをグラフィックとしてエクスポートしたりすることができます。

IBM Netezza Analytics バージョン 2.x 以前で作業している場合、あるいは距離測度として Mahalanobis を使用してモデルを構築する場合は、K-Means モデルの内容はテキスト形式でのみ表示されます。

これらのバージョンの場合、以下の情報が表示されます。

- 「統計の要約」。最小クラスターおよび最大クラスターの両方について、統計の要約にレコード数が表示されます。「統計の要約」には、これらのクラスターによって占有されるデータ・セットの割合も表示されます。またリストには、最大クラスターから最小クラスターに対するサイズ比が表示されます。
- クラスタリング要約。クラスタリング要約には、アルゴリズムで作成されたクラスターがリストされます。各クラスターについて、テーブルにはクラスターのレコード数、これらのレコードのクラスター中心からの平均距離が表示されます。

Netezza K-Means モデル・ナゲット -「設定」タブ

「設定」タブで、モデルをスコアリングするためのオプションを設定できます。

入力フィールドを含める:選択すると、元の入力フィールドをすべて下流に渡し、追加のモデリング・フィールドをデータの各行に添付します。このボックスをオフにすると、レコード ID フィールドと追加のモデリング・フィールドのみが渡されるため、ストリームがより早く実行されます。

距離測度 :データ・ポイント間の距離の測定に使用する方法です。距離が大きいほど、相違が大きくなります。オプションは次のとおりです。

- ユークリッド:(デフォルト) 2 点間の距離は、それらを直線で結ぶことによって計算されます。
- **Manhattan**:2 点間の距離は、それらの座標間の絶対距離の合計として計算されます。
- **Canberra**:Manhattan の距離と同じですが、原点に近いデータ・ポイントに対してより感度が高くなります。
- 最大値 :2 点間の距離は、座標の次元に沿ったそれらの相違の最大値として計算されます。

Netezza Bayes ネットワークのモデル・ナゲット

Bayes Net のモデルナゲットは、モデルをスコアリングするためのオプションを設定する手段を提供します。

Bayes Net モデル・ナゲットを含むストリームを実行すると、ノードは 1 つの新しいフィールドを追加します。その名前は、対象名に由来します。

表 *14. Bayes Net* のモデルスコアリング・フィールド :

追加フィールドの名前	意味
$BN-*target_name*	現在のレコードの予測値。

モデルナゲットにテーブル・ノードを接続して表のノードを実行することにより、追加フィールドを表示することができます。

Netezza Bayes Net ナゲット -「設定」タブ

「設定」タブでは、モデルのスコアリングのオプションを設定することができます。

目標: 現在の対象とは異なる対象フィールドをスコアリングしたい場合は、ここで新しい対象を選択します。

「レコード **ID**」。「レコード ID」フィールドが指定されていない場合は、ここで使用するフィールドを選択してください。

予測の種類**:** 使用したい予測アルゴリズムのバリエーション。

- 最適 **(**最も相関する近傍**):** (デフォルト) 相関度の高い近傍ノードを使用します。
- 近傍 **(**近傍の重み付き予測**):** すべての近傍ノードの重み付き予測を使用しています。
- **NN** 近傍 **(null** 以外の近傍**):** NULL値を持つノードを無視するという点を除いては、前のオプションと同じです (予測が計算されるインスタンスの欠損値がある属性に対応するノード)。

入力フィールドを含める:選択すると、元の入力フィールドをすべて下流に渡し、追加のモデリング・フィールドをデータの各行に添付します。このボックスをオフにすると、レコード ID フィールドと追加のモデリング・フィールドのみが渡されるため、ストリームがより早く実行されます。

Netezza Naive Bayes のモデル・ナゲット

Naive Bayes のモデルナゲットは、モデルをスコアリングするためのオプションを設定する手段を提供します。

Naive Bayes のモデル・ナゲットを含むストリームを実行すると、デフォルトでは、ノードは 1 つの新しいフィールドを追加します。その名前は、対象名に由来します。

表 *15. Naive Bayes* のモデルスコアリング・フィールド - デフォルト：

追加フィールドの名前	意味
$I-*target_name*	現在のレコードの予測値。

モデリング・ノードまたはモデル・ナゲットで 「レコードのスコアリングに割り当てられたクラスの確率を計算する」 を選択した場合、さらにフィールドを 2 つ追加します。

表 *16. Naive Bayes* のモデルスコアリング・フィールド - 追加：

追加フィールドの名前	意味
$IP-*target_name*	インスタンスのクラスの Bayesian 分子 (前のクラスの確率と条件付きのインスタンスの属性値の確率の積)。
$ILP-*target_name*	後者の自然対数。

モデルナゲットにテーブル・ノードを接続して表のノードを実行することにより、追加フィールドを表示することができます。

Netezza Naive Bayes ナゲット -「設定」タブ

「設定」タブでは、モデルのスコアリングのオプションを設定することができます。

入力フィールドを含む:選択すると、元の入力フィールドをすべて下流に渡し、追加のモデリング・フィールドをデータの各行に添付します。このボックスをオフにすると、レコード ID フィールドと追加のモデリング・フィールドのみが渡されるため、ストリームがより早く実行されます。

レコードのスコアリングに割り当てられたクラスの確率を計算する:(ディシジョン・ツリーおよび Naive Bayes のみ) 選択した場合、このオプションは、追加のモデリング・フィールドが確信（確率）フィールドのほか、予測フィールドを含めることを意味します。このチェックボックスをオフにすると、予測のみのフィールドが生成されます。

小さいまたは大きく偏りのあるデータセットの確率の精度を向上させる**:** 確率の計算時、このオプションでは、推定時に 0 の確立を回避する *m* 推定手法が使用されます。この種類の確率の推定は、速度が遅くなることがありますが、小さいまたは非常に偏りのあるデータセットによりよい結果を与えることができます。

Netezza KNN モデル・ナゲット

KNN のモデルナゲットは、モデルをスコアリングするためのオプションを設定する手段を提供します。

KNN モデル・ナゲットを含むストリームを実行すると、ノードは 1 つの新しいフィールドを追加します。その名前は、対象名に由来します。

表 *17. KNN* のモデル・スコアリング・フィールド：

追加フィールドの名前	意味
$KNN-*target_name*	現在のレコードの予測値。

モデルナゲットにテーブル・ノードを接続して表のノードを実行することにより、追加フィールドを表示することができます。

Netezza KNN ナゲット -「設定」タブ

「設定」タブでは、モデルのスコアリングのオプションを設定することができます。

距離測度 :データ・ポイント間の距離の測定に使用する方法です。距離が大きいほど、相違が大きくなります。オプションは次のとおりです。

- ユークリッド:(デフォルト) 2 点間の距離は、それらを直線で結ぶことによって計算されます。
- **Manhattan**:2 点間の距離は、それらの座標間の絶対距離の合計として計算されます。
- **Canberra**:Manhattan の距離と同じですが、原点に近いデータ・ポイントに対してより感度が高くなります。
- 最大値 :2 点間の距離は、座標の次元に沿ったそれらの相違の最大値として計算されます。

最近傍数 **(k)**: 特定のケースの最近傍の数。より大きな数の近傍を使用すると、必ずしも正確なモデルが作成されるとは限りません。

k を選択することにより、オーバーフィット（特に「ノイズの多い」データの場合は重要な場合があります）の防止と解決（同様のインスタンスに対する異なる予測）のバランスを制御します。各データセットごとに *k* の値 (値は 1〜 数十の範囲) を調整する必要があります

入力フィールドを含める:選択すると、元の入力フィールドをすべて下流に渡し、追加のモデリング・フィールドをデータの各行に添付します。このボックスをオフにすると、レコード ID フィールドと追加のモデリング・フィールドのみが渡されるため、ストリームがより早く実行されます。

距離を計算する前に測定値を標準化: 選択した場合、このオプションは、距離の値を計算する前に連続型入力フィールドの測定を標準化します。

コアセットを使用して大規模なデータセットのパフォーマンスを向上させる: このオプションは、大規模なデータセットを使用する場合、計算を高速化するためにコアセット・サンプリングを使用します。

Netezza 分裂クラスタリングのモデル・ナゲット

分裂クラスタリングは、モデルをスコアリングするためのオプションを設定する手段を提供します。

分裂クラスタリング・モデル・ナゲットを含むストリームを実行すると、ノードは 2 つの新しいフィールドを追加します。その名前は、対象名に由来します。

表 *18.* 分裂クラスタリングのモデル・スコアリング・フィールド：

追加フィールドの名前	意味
$DC-*target_name*	現在のレコードが割り当てられているサブクラスターの識別子。
$DCD-*target_name*	現在のレコードのサブクラスター中心からの距離。

モデルナゲットにテーブル・ノードを接続して表のノードを実行することにより、追加フィールドを表示することができます。

Netezza 分裂クラスタリング ナゲット -「設定」タブ

「設定」タブでは、モデルのスコアリングのオプションを設定することができます。

入力フィールドを含める:選択すると、元の入力フィールドをすべて下流に渡し、追加のモデリング・フィールドをデータの各行に添付します。このボックスをオフにすると、レコード ID フィールドと追加のモデリング・フィールドのみが渡されるため、ストリームがより早く実行されます。

距離測度 :データ・ポイント間の距離の測定に使用する方法です。距離が大きいほど、相違が大きくなります。オプションは次のとおりです。

- ユークリッド:(デフォルト) 2 点間の距離は、それらを直線で結ぶことによって計算されます。
- **Manhattan**:2 点間の距離は、それらの座標間の絶対距離の合計として計算されます。
- **Canberra**:Manhattan の距離と同じですが、原点に近いデータ・ポイントに対してより感度が高くなります。
- 最大値 :2 点間の距離は、座標の次元に沿ったそれらの相違の最大値として計算されます。

適用階層レベル: データに適用する必要がある階層のレベル。

Netezza PCA モデル・ナゲット

PCA のモデルナゲットは、モデルをスコアリングするためのオプションを設定する手段を提供します。

PCA モデル・ナゲットを含むストリームを実行すると、デフォルトでは、ノードは 1 つの新しいフィールドを追加します。その名前は、対象名に由来します。

表 *19. PCA* のモデル・スコアリング・フィールド：

追加フィールドの名前	意味
$F-*target_name*	現在のレコードの予測値。

モデリング・ノードまたはモデル・ナゲットの「主要成分の数...」で 1 より大きい値を選択した場合、ノードは各成分に新しいフィールドを追加します。この場合、フィールド名には -n の接尾辞が付きます。この場合、フィールド名には *-n* という接尾辞が付加されます。*n* は成分の数です。例えば、*pca* という名前で 3 つの成分を含むモデルの場合、新規フィールド名は *$F-pca-1*、*$F-pca-2*、および *$F-pca-3* になります。

モデルナゲットにテーブル・ノードを接続して表のノードを実行することにより、追加フィールドを表示することができます。

Netezza PCA ナゲット -「設定」タブ

「設定」タブでは、モデルのスコアリングのオプションを設定することができます。

投影に使用する主成分の数**:** データセットを削減したい主要成分の数。この値は、属性（入力フィールド）の数を超えてはなりません。

入力フィールドを含める:選択すると、元の入力フィールドをすべて下流に渡し、追加のモデリング・フィールドをデータの各行に添付します。このボックスをオフにすると、レコード ID フィールドと追加のモデリング・フィールドのみが渡されるため、ストリームがより早く実行されます。

Netezza 回帰ツリー・モデル・ナゲット

回帰ツリーのモデルナゲットは、モデルをスコアリングするためのオプションを設定する手段を提供します。

回帰ツリー・モデル・ナゲットを含むストリームを実行すると、デフォルトでは、ノードは 1 つの新しいフィールドを追加します。その名前は、対象名に由来します。

表 *20.* 回帰ツリーのモデルスコアリング・フィールド：

追加フィールドの名前	意味
$I-*target_name*	現在のレコードの予測値。

モデリング・ノードまたはモデル・ナゲットで 「推定分散を計算」 を選択してストリームを実行した場合、さらにフィールドを追加します。

表 *21.* 回帰ツリーのモデルスコアリング・フィールド - 追加：

追加フィールドの名前	意味
$IV-*target_name*	予測値の推定分散。

モデルナゲットにテーブル・ノードを接続して表のノードを実行することにより、追加フィールドを表示することができます。

Netezza 回帰ツリー・ナゲット -「モデル」タブ

「モデル」タブには、回帰ツリー・モデルの「予測値の重要度」がグラフィカル形式で表示されます。棒の長さは、予測値の重要度を表しています。

注**:** IBM Netezza Analytics バージョン 2.x 以前で作業している場合、回帰ツリー・モデルの内容はテキスト形式でのみ表示されます。

これらのバージョンの場合、以下の情報が表示されます。

- テキストの各行は、ノードまたは葉に対応しています。

- インデントはツリー・レベルを反映します。
- ノードの場合、分割の条件が表示されます。
- 葉の場合、割り当てられたクラスのラベルが表示されます。

Netezza 回帰ツリー -「設定」タブ

「設定」タブでは、モデルのスコアリングのオプションを設定することができます。

入力フィールドを含める:選択すると、元の入力フィールドをすべて下流に渡し、追加のモデリング・フィールドをデータの各行に添付します。このボックスをオフにすると、レコード ID フィールドと追加のモデリング・フィールドのみが渡されるため、ストリームがより早く実行されます。

推定分散を計算: 割り当てられたクラスの分散が出力に含まれるべきかどうかを示します。

Netezza 回帰ツリー ナゲット -「ビューア」タブ

「ビューア」タブには、SPSS Modeler で回帰ツリー モデルが表示されるのと同じ方法で、ツリー モデルのツリー プレゼンテーションが表示されます。

注: モデルが IBM Netezza Analytics バージョン 2.x 以前で構築されている場合、「ビューア」タブは空です。

Netezza 線型回帰ツリー・モデル・ナゲット

線型回帰のモデルナゲットは、モデルをスコアリングするためのオプションを設定する手段を提供します。

線型回帰モデル・ナゲットを含むストリームを実行すると、ノードは 1 つの新しいフィールドを追加します。その名前は、対象名に由来します。

表 22. 線型回帰のモデルスコアリング・フィールド：

追加フィールドの名前	意味
$LR-*target_name*	現在のレコードの予測値。

Netezza 線型回帰ナゲット -「設定」タブ

「設定」タブでは、モデルのスコアリングのオプションを設定することができます。

入力フィールドを含める:選択すると、元の入力フィールドをすべて下流に渡し、追加のモデリング・フィールドをデータの各行に添付します。このボックスをオフにすると、レコード ID フィールドと追加のモデリング・フィールドのみが渡されるため、ストリームがより早く実行されます。

Netezza 時系列モデル・ナゲット

モデル・ナゲットは、時系列モデリング操作の出力へのアクセスを提供します。出力は以下のフィールドから構成されています。

表 23. 時系列のモデル出力フィールド

フィールド	説明
TSID	時系列の識別子。モデリング・ノードの「フィールド」タブの「時系列 ID」で指定したフィールドの内容。 詳しくは、トピック 76 ページの『Netezza 時系列フィールド・オプション』を参照してください。
TIME	現在の時系列の時間。

表 23. 時系列のモデル出力フィールド (続き)

フィールド	説明
HISTORY	過去のデータ値 (予測に使用された値)。モデル・ナゲットの「設定」タブでオプション 「結果の過去の値を含める」 を選択した場合にのみ、このフィールドが含まれます。
$TS-INTERPOLATED	使用される補間の値。モデル・ナゲットの「設定」タブでオプション 「結果に補間値を含める」 を選択した場合にのみ、このフィールドが含まれます。補間は、モデリング・ノードの「作成オプション」タブのオプションです。
$TS-FORECAST	時系列の予測値。

モデル出力を表示するには、テーブル・ノード (ノード・パレットの「出力」タブ) をモデル・ナゲットに接続してテーブル・ノードを実行します。

Netezza 時系列ナゲット -「設定」タブ

「設定」タブで、モデル出力をカスタマイズするオプションを指定できます。

モデル名**:** モデリング・ノードの「モデル・オプション」タブで指定した、モデルの名前。

その他のオプションは、モデリング・ノードの「モデル・オプション」タブと同じです。

Netezza 一般化線形モデル・ナゲット

モデル・ナゲットは、モデリング操作の出力へのアクセスを提供します。

一般化線型モデル・ナゲットを含むストリームを実行すると、ノードは 1 つの新しいフィールドを追加します。その名前は、対象名に由来します。

表 24. 一般化線形のモデルスコアリング・フィールド：

追加フィールドの名前	意味
$GLM-*target_name*	現在のレコードの予測値。

「モデル」タブには、モデルに関連するさまざまな統計が表示されます。

出力は以下のフィールドから構成されています。

表 25. 一般化線形の出力フィールド：

出力フィールド	説明
パラメーター	モデルで使用されるパラメーター (予測変数)。切片 (回帰モデルの定数項) のほか、数値型および名義型の列です。
ベータ	相関係数 (モデルの線型要素)。
標準誤差	データの標準誤差。
テスト	パラメーターの妥当性の評価に使用される検定統計。
p-値	パラメーターが有意であると想定する場合のエラーの確率。
残差要約	
残差のタイプ	要約値が表示される予測の残差の種類。
RSS	残差の値。
自由度	残差の自由度。

表 25. 一般化線形の出力フィールド (続き) :

出力フィールド	説明
p-値	エラーの確率。値が大きいほど適合度の低いモデルを示します。値が小さいほど適合度が高くなります。

Netezza 一般化線形モデル・ナゲット -「設定」タブ

「設定」タブで、モデル出力をカスタマイズできます。

オプションは、モデリング・ノードの「スコアリング・オプション」に表示されているものと同じです。詳しくは、トピック 66 ページの『Netezza 一般化線形モデル・オプション - スコアリング・オプション』を参照してください。

Netezza TwoStep モデル・ナゲット

TwoStep モデル・ナゲットを含むストリームを実行すると、そのレコードの所属クラスターと割り当てられたクラスターの中心からの距離を含む 2 つの新規フィールドが追加されます。$TS-Twostep という名前の新規フィールドは所属クラスタ用で、$TSP-Twostep という名前の新規フィールドはクラスタの中心からの距離用です。

Netezza TwoStep ナゲット -「モデル」タブ

「モデル」タブには、クラスターのフィールドの要約統計量および分布を表示する各種のグラフィック表示があります。モデルからデータをエクスポートしたり、ビューをグラフィックとしてエクスポートしたりすることができます。

第 6 章 IBM DB2 for z/OS によるデータベース モデリング

IBM SPSS Modeler および IBM DB2 for z/OS

SPSS Modeler では、DB2 for z/OS との統合をサポートしており、DB2 for z/OS サーバーで高度な分析を実行する機能を提供します。これらの機能には、SPSS Modeler グラフィカル ユーザー インターフェースおよびワークフロー指向の開発環境でアクセスできます。この方法により、DB2 for z/OS 環境で直接、IBM DB2 Analytics Accelerator を活用してデータ マイニング アルゴリズムを実行できます。

SPSS Modeler は、DB2 for z/OS の次のアルゴリズムの統合をサポートします。

- デシジョン・ツリー
- K-Means
- Naive Bayes
- 回帰ツリー
- TwoStep

IBM DB2 for z/OS との統合の要件

DB2® for z/OS® および IBM DB2 Analytics Accelerator for z/OS を使用してデータベース内モデル作成を実行する場合、以下の条件が前提条件となります。場合によっては、これらの条件が満たされているかデータベース管理者に問い合わせて確認してください。

- ローカル モードで動作している IBM SPSS Modeler、あるいは Windows または UNIX 上の SPSS Modeler Server インストール済み環境に対して動作している SPSS Modeler
- DB2 Analytics Accelerator for z/OS バージョン 5 と併用する DB2 for z/OS バージョン 10 以降
- IBM SPSS Data Access Pack V7.1
- SPSS Modeler Server を実行するサーバーでは、以下のいずれかのシステム:
  - IBM DB2 Data Server Driver for ODBC and CLI
  - DB2 for z/OS 用に構成された ODBC データ ソースを使用する任意のバージョンの DB2 for Linux、UNIX、および Windows
- DB2 Connect™ for System z® のライセンス
- SPSS Modeler で有効化された SQL の生成および最適化
- IBM SPSS Modeler Scoring Adapter for zEnterprise® V17.0
- DB2 z/OS データベース内マイニングには、アクセラレーター専用テーブル (AOT) またはアクセラレーテッド・テーブル、および INZA のサポートが必要です。IDAA INZA は IDAA 5.1 で導入されました。つまり、DB2 z/OS データベース内マイニング・ノードは、前のバージョンの IDAA では動作しません。

  IDAA 対応の DSN を Modeler で使用する場合、その DSN 使用してデータベース入力ノードに返されるテーブルのリストに表示される専用テーブルは、AOT またはアクセラレーテッド・テーブルです。

IBM DB2 Analytics Accelerator for z/OS との統合の有効化

DB2 Analytics Accelerator for z/OS と統合する手順は、次のとおりです。

- DB2 for z/OS および DB2 Analytics Accelerator for z/OS の構成
- ODBC ソースの作成
- IBM SPSS Modeler での IBM DB2 for z/OS の統合の有効化
- SPSS Modeler での SQL の生成と最適化を有効にする
- IBM SPSS Modeler Server Scoring Adapter for DB2 for z/OS の有効化
- IBM SPSS Modeler での IBM DB2 クライアントを使用した DSN の構成

IBM DB2 for z/OS および IBM Analytics Accelerator for z/OS の構成

DB2 for z/OS および Analytics Accelerator for z/OS の構成方法は、次の Web サイトで説明されています。

DB2 Analytics Accelerator for z/OS

IBM DB2 for z/OS および IBM DB2 Analytics Accelerator の ODBC ソースの作成

DB2 for z/OS と IBM DB2 Analytics Accelerator の間の接続を有効にする方法について詳しくは、次の Web サイトを参照してください。

- バージョン 4 の場合: DB2 Analytics Accelerator for z/OS 4.1.0
- バージョン 3 の場合: DB2 Analytics Accelerator for z/OS 3.1.0
- Enabling query acceleration with IBM DB2 Analytics Accelerator for ODBC and JDBC applications without modifying the applications
- SQL error from ODBC driver when running a query in DB2 Analytics Accelerator for z/OS

IBM SPSS Modeler での IBM DB2 for z/OS の統合の有効化

SPSS Modeler で DB2 for z/OS の統合を有効にするには、以下のステップを実行します。

1. SPSS Modeler `config` ディレクトリーから `odbc-db2-accelerator-names.cfg` ファイルを開きます。

   このファイルが存在しない場合は、作成する必要があります。
2. すべてのデータ ソースの名前とすべてのアクセラレータの名前を追加します。以下に例を示します。

   ```
   dsn1, acceleratorname1
   dsn2, acceleratorname2
   ```
3. アクセラレータ専用テーブル (AOT) のデフォルト CCSID は Unicode です。これを上書きするには、エンコード文字列をアクセラレータ名に追加して、エントリーを変更します。以下に例を示します。

   ```
   dsn1, acceleratorname1, EBCDIC
   dsn2, acceleratorname2, UNICODE
   ```
4. `odbc-db2-accelerator-names.cfg` ファイルを保存して閉じた後、同じディレクトリーから `odbc-db2-custom-properties.cfg` ファイルを開きます。
5. SPSS Modeler は、SQL を使用して IDAA レジスターを設定します。必要な場合は、SQL を目的の値に変更することにより、これらのエントリーを上書きできます。以下に例を示します。

   ```
   current_query_sql_acc, "SET CURRENT QUERY ACCELERATION = ELIGIBLE"
   current_get_archive_acc, "SET CURRENT GET_ACCEL_ARCHIVE = NO"
   ```

6. デフォルトでは、SPSS Modeler は、SQL を使用して、データベース・キャッシュ用の一時テーブルを作成します。必要な場合は、所定のデータベース名を指定して、これを上書きできます。以下に例を示します。

   ```
   [OSZ]
   table_create_temp_sql_acc, 'CREATE TABLE <table-name> <(table-columns)> IN DATABASE NAME_OF_DATABASE_FOR_AOT'
   ```

7. デフォルトでは、SPSS Modeler は、ODBC 入力ノードで作成された SQL クエリーは再生不可であると判断します。つまり、クエリーを複数回実行すると、異なる結果が返されると判断します。ただし、シナリオによっては、これが原因で Modeler が下流ノードの SQL を生成しない場合があります。この動作を上書きするには、該当する値を `Y` に変更します。次に例を示します。

   ```
   assume_custom_sql_replayable, Y
   ```

8. SPSS Modeler のメインメニューから、「ツール」 > 「オプション」 > 「ヘルパー アプリケーション」をクリックします。
9. 「**IBM DB2 for z/OS**」タブをクリックします。
10. 「**IBM DB2 for z/OS Data Mining** 統合を有効化」を選択して、「**OK**」をクリックします。

注**:** IDAA テーブルと非 IDAA テーブルを Modeler 内に同時に表示することはできません。

SQL の生成と最適化を有効にする

非常に大きいデータ・セットを扱う確率が高いため、パフォーマンス上の理由により、IBM SPSS Modeler で SQL 生成および最適化を有効にする必要があります。

SPSS Modeler を構成するには、以下のステップを実行します。

1. IBM SPSS Modeler メニューから「ツール」 > 「ストリームのプロパティー」 > 「オプション」を選択します。
2. ナビゲーション ペインの 「最適化」 オプションをクリックします。
3. 「**SQL** 生成」 オプションが有効になっていること確認します。この設定は、データベースのモデル作成が機能するために必要です。
4. 「**SQL** 生成の最適化」と「その他の実行を最適化」を選択します (絶対に必要な訳ではありませんが、最適化されたパフォーマンスを得るために、選択することを強くお勧めします)。

IBM SPSS Modeler での IBM DB2 クライアントを使用した DSN の構成

SPSS Modeler で、DB2 用の DB2 クライアントを使用してデータ・ソース名 (DSN) を構成する必要がある場合は、次の手順を実行します。

1. DB2 クライアントを Modeler Server がインストールされているオペレーティング・システムにインストールします (まだインストールしていない場合)。
2. **`db2 catalog`** コマンドを使用してデータベースをカタログした後、DB2 クライアントの `db2cli.ini` ファイルに新しいデータ・ソースを追加します。定義済みのデータベースのエイリアスを指すようにしてください。
3. データ・アクセスを構成します。詳細な手順については、Modeler の資料を参照してください。

   詳しくは、「*Modeler Server* 管理およびパフォーマンス・ガイド」 (ModelerServerAdminPerformance.pdf) のトピック『サーバーアーキテクチャーとハードウェアに関する推奨事項』 > 『データへのアクセス』を参照してください。
4. 手順 2 で定義したデータベースのエイリアスを参照して、`odbc.ini` に新しい ODBC データ・ソースを作成します。

5. Linux ユーザーまたは UNIX ユーザーの場合:
   a. ドライバー・ライブラリーとして、(libdb2.so ではなく) libdb2o.so が使用されていることを確認します。また、新しいデータ・ソースに対して 'DriverUnicodeType=1' が定義されていることを確認します。
   b. IBM SPSS Data Access Pack のインストール済み環境で、DB2 クライアントのライブラリー・パスが odbc.sh に追加されていることを確認します。
   c. Modeler Server で、エンコード方式が UTF-16 である ODBC ドライバー・ラッパー・ライブラリー (これは 'libspssodbc_datadirect_utf16.so' と呼ばれます) が使用されていることを確認します。
6. DB2 に接続するユーザーが、次のクエリーの実行に必要な権限を備えていることを確認します。

```
SELECT ACCELERATORNAME FROM SYSACCEL.SYSACCELERATORS
```

IBM DB2 for z/OS でのモデル構築

サポートされているアルゴリズムのそれぞれに、対応するモデリング・ノードがあります。ノード・パレットの「データベース・モデリング」タブから、DB2 for z/OS モデリング・ノードにアクセスできます。

データの考慮事項

データ・ソース内のフィールドは、モデリング・ノードに応じて、さまざまなデータ型の変数を含めることができます。SPSS Modeler では、データ型は測定の尺度とも呼ばれます。モデリングのノードの「フィールド」タブでは、入力フィールドと対象フィールドに許可されている測定の尺度の種類を示すアイコンを使用しています。

対象フィールド:対象フィールドは、値を予測しようとしているフィールドです。対象を指定できる場合、ソース・データ・フィールドのうち 1 つだけを対象フィールドとして選択できます。

レコード **ID** フィールド:各ケースを一意に識別するためにフィールドを指定します。例えば、これは *CustomerID* などの ID フィールドです。ソースデータに ID フィールドが含まれていない場合は、次の手順に示すように、フィールド生成ノードを使用してこのフィールドを作成することができます。

1. 入力ノードを選択します。
2. ノード・パレットの「フィールド設定」タブから、フィールド生成ノードをダブルクリックします。
3. 領域内のアイコンをダブルクリックしてフィールド生成ノードを開きます。
4. 「派生フィールド」 フィールドで、例えば ID を入力します。
5. 「**CLEM** 式」 フィールドで、@INDEX と入力して 「**OK**」 をクリックします。
6. フィールド生成ノードをストリームの残りに接続します。

ヌル値の処理

入力データに null 値が含まれている場合、いくつかの DB2 for z/OS ノードを使用することによってエラー・メッセージが表示されたりストリームの実行時間が長くなる可能性があるため、null 値を含むレコードを削除することをお勧めします。以下の方法を使用します。

1. 条件抽出ノードを入力ノードに接続します。
2. 条件抽出ノードの 「モード」 を 「破棄」 に設定します。
3. 「条件」 フィールドに以下を入力します。

```
@NULL(field1) [or @NULL(field2)[... or @NULL(fieldN]])
```

すべての入力フィールドを含むようにします。

4. 条件抽出ノードをストリームの残りに接続します。

モデル出力

DB2 for z/OS モデル作成ノードを含むストリームは、わずかに異なる結果を実行のたびに生成する可能性があります。それは、データがモデル作成の前に一時テーブルに読み込まれるため、ノードがソース・データを読み取る順序が必ずしも同じではないからです。ただし、この効果によって生まれる相違点は無視できます。

一般的なコメント

- SPSS Collaboration and Deployment Services では、DB2 for z/OS モデリング・ノードを含むストリームを使用してスコアリング設定を作成することはできません。
- DB2 for z/OS ノードで作成されたモデルの場合、PMML エクスポートまたはインポートを行うことはできません。

IBM DB2 for z/OS モデル - フィールド オプション

「フィールド」タブで、上流のノードですでに定義されているフィールドの役割設定を使用するか、手動でフィールドの割り当てを行うかを選択します。

事前定義された役割を使用: 上流のデータ型ノード (または上流の入力ノードの「データ型」タブ) の役割設定 (対象、予測など) を使用します。

カスタム・フィールド割り当ての使用: この画面で対象、予測、およびその他の役割を手動で割り当てる場合、このオプションを選択します。

フィールド**:** 矢印ボタンを使用して、このリストの項目を画面右側のさまざまな役割フィールドに手動で割り当てます。アイコンは、各役割フィールドの有効な測定の尺度を示します。

「すべて」 ボタンをクリックしてリスト内のすべてのフィールドを選択するか、各測定の尺度のボタンをクリックして、その測定の尺度のすべてのフィールドを選択します。

目標: 1 つのフィールドを予測の対象として選択します。一般化線形モデルの場合、この画面の 「試行回数」 フィールドも参照してください。

「レコード **ID**」。一意のレコード ID として使用されるフィールド。

予測値 **(**入力**):** 1 つまたは複数のフィールドを予測の入力として選択します。

IBM DB2 for z/OS モデル - サーバー オプション

「サーバー」タブでは、モデルが構築される DB2 for z/OS システムを指定します。

- 上流の接続を使用**:** (デフォルト) データベース入力ノードなど、上流のノードで指定した接続の詳細を使用します。注: このオプションは、すべての上流ノードが SQL プッシュバックを使用できる場合にのみ有効です。この場合、SQL が完全にすべての上流ノードを実装するため、データベース外のデータを移動する必要はありません。
- 接続するデータを移動:ここでしてデータベースにデータを移動します。データを移動することにより、データが別の IBM データベース、他のベンダーのデータベースにある場合、またはデータがフラット・ファイルにある場合でもモデリングを実行できるようにします。また、ノードが SQL プッシュバ

ックを実行していないためデータが抽出されていない場合、データはここで指定されたデータベースに戻されます。接続を参照して選択するには、「編集」ボタンをクリックします。

コメント:

ODBC データ・ソース名は、各 SPSS Modeler ストリームに効果的に埋め込まれます。あるホスト上で作成されたストリームが別のホスト上で実行された場合、データ・ソースの名前はそれぞれのホストで同じである必要があります。また、各入力ノードまたはモデル作成ノードで、「サーバー」タブから異なるデータ・ソースを選択できます。

IBM DB2 for z/OS モデル - モデル オプション

「モデル・オプション」タブで、モデルの名前を指定するか、自動的に名前を生成するかを選択できます。

モデル名**:** ターゲットまたは ID フィールド (その指定がない場合はモデル タイプ) に基づいてモデル名を生成、またはカスタム名を指定することができます。

名前が使用されている場合は既存モデルを置換。このチェック ボックスを選択する場合、同じ名前の既存モデルは上書きされます。

IBM DB2 for z/OS モデル - K-Means

K-Means ノードは、クラスター分析の方法を提供する *k*-means アルゴリズムを実行します。このノードを使用して、データ・セットをグループにクラスター化できます。

アルゴリズムは距離に基づくアルゴリズムで、距離メトリック (関数) に依存してデータ・ポイント間の類似性を測定します。データ・ポイントは、使用する距離メトリックに従って最も近いクラスターに割り当てられます。

アルゴリズムは、同じ基本プロセスを何度か反復することによって実行します。学習インスタンスは最も近いクラスターに割り当てられます (指定された距離関数に関しては、インスタンスとクラスター中心に適用されます)。すべてのクラスター中心は、特定のクラスターに割り当てられたインスタンスの平均属性値のベクトルとして再計算されます。

IBM DB2 for z/OS モデル - K-Means のフィールド オプション

「フィールド」タブで、上流のノードですでに定義されているフィールドの役割設定を使用するか、手動でフィールドの割り当てを行うかを選択します。

事前定義された役割を使用: 上流のデータ型ノード (または上流の入力ノードの「データ型」タブ) の役割設定 (対象、予測など) を使用します。

カスタム・フィールド割り当ての使用: この画面で対象、予測、およびその他の役割を手動で割り当てる場合、このオプションを選択します。

フィールド**:** 矢印ボタンを使用して、このリストの項目を画面右側のさまざまな役割フィールドに手動で割り当てます。アイコンは、各役割フィールドの有効な測定の尺度を示します。

「すべて」 ボタンをクリックしてリスト内のすべてのフィールドを選択するか、各測定の尺度のボタンをクリックして、その測定の尺度のすべてのフィールドを選択します。

「レコード **ID**」。一意のレコード ID として使用されるフィールド。

予測値 **(**入力**):** 1 つまたは複数のフィールドを予測の入力として選択します。

IBM DB2 for z/OS モデル - K-Means の作成オプション

作成オプションを設定することにより、用途に合わせてモデルの作成をカスタマイズできます。

デフォルト・オプションを使用してモデルを作成する場合は、「実行」をクリックします。

距離測度**:** このパラメーターは、データ・ポイント間の距離の測定方法を定義します。距離が大きくなると、非類似度も大きくなります。以下のいずれかのオプションを選択します。

- ユークリッド: ユークリッド測定は、2 つのデータ・ポイント間の直線距離です。
- 正規化ユークリッド **(Normalized Euclidean)**。正規化ユークリッド測定はユークリッド測定に類似していますが、平方標準偏差によって正規化されます。ユークリッド測定とは異なり、正規化ユークリッド測定はスケール不変でもあります。

クラスター数**:** このパラメーターは、作成するクラスターの数を定義します。

最大反復数**:** アルゴリズムは、同じプロセスを何度か反復します。このパラメーターは、モデルの学習を停止する前の反復数を定義します。

統計: このパラメーターは、モデルに含まれる統計の数を定義します。以下のいずれかのオプションを選択します。

- 「すべて」。列に関連したすべての統計と値に関連したすべての統計が含まれます。

  注**:** このパラメーターを使用すると統計の最大数が含まれるため、システムのパフォーマンスに影響する場合があります。モデルをグラフィカル形式で表示しない場合は、「なし」を指定します。
- 「列」。列に関連した統計が含まれます。
- なし**:** モデルのスコアリングに必要な統計のみが含まれます。

結果の再現: ランダム・シードを設定して分析を複製する場合は、このチェック・ボックスを選択します。整数を指定するか、「生成」をクリックして整数の疑似乱数を作成します。

IBM DB2 for z/OS モデル - Naive Bayes

Naive Bayes は、分類の問題に対応する有名なアルゴリズムです。提示されたすべての予測変数は相互に依存関係がないものとして処理されるので、モデルは naïve と命名されます。Naive Bayes は拡張性のある高速のアルゴリズムであり、複数の属性と対象属性の組み合わせに対して、条件に関連する確率を計算します。学習データから、個別の確率が計算されます。各入力変数の各値カテゴリーを計算単位とすると、この確率は各対象クラスの確率を表します。

IBM DB2 for z/OS モデル - ディシジョン ツリー

ディシジョン・ツリーは、分類モデルを示す階層構造です。ディシジョン・ツリー・モデルによって、学習データのセットから将来の観測値を予測または分類する分類システムを開発できます。分類は、ブランチが分類の分割点を示すツリー構造の形式です。分割は、停止点に達するまでデータをサブグループに繰り返し分割します。停止点のツリーノードは、葉と呼ばれます。各葉は、クラス・ラベルというラベルを、サブグループのメンバー、またはクラスに割り当てます。

IBM DB2 for z/OS モデル - ディシジョン ツリーのフィールド オプション

「フィールド」タブで、上流のノードですでに定義されているフィールドの役割設定を使用するか、手動でフィールドの割り当てを行うかを選択します。

事前定義された役割を使用: 上流のデータ型ノード (または上流の入力ノードの「データ型」タブ) の役割設定 (対象、予測など) を使用します。

カスタム・フィールド割り当ての使用: この画面で対象、予測、およびその他の役割を手動で割り当てる場合、このオプションを選択します。

フィールド: 矢印ボタンを使用して、このリストの項目を画面右側のさまざまな役割フィールドに手動で割り当てます。アイコンは、各役割フィールドの有効な測定の尺度を示します。

「すべて」 ボタンをクリックしてリスト内のすべてのフィールドを選択するか、各測定の尺度のボタンをクリックして、その測定の尺度のすべてのフィールドを選択します。

目標: 1 つのフィールドを予測の対象として選択します。

「レコード **ID**」。一意のレコード ID として使用されるフィールド。このフィールドの値は各レコードで一意である必要があります (例: カスタマー ID 番号)。

インスタンスの重み:デフォルトのクラスの重み (対象フィールドのカテゴリーあたりの重み) の代わりに、またはデフォルトに加えてインスタンスの重み (入力データの行あたりの重み) を使用できるようフィールドを指定します。ここで指定するフィールドは、入力データの各行の数値の重みを含むフィールドでなければなりません。

予測値 (入力): 1 つ以上の入力フィールドを選択します。これは、データ型ノードのフィールドの役割を「入力」 に設定するのと似ています。

IBM DB2 for z/OS モデル - ディシジョン ツリーの作成オプション

ツリーの成長には、以下の作成オプションを使用できます。

成長の測定。これらのオプションは、ツリーの成長を測定する方法を制御します。

- 不純度の測定: この測定では、ツリーを分割するのに最適な場所が評価されます。これは、サブグループまたはデータのセグメントにおける変動性の測定です。不純度の測定値が低い場合は、グループ内のほとんどのメンバーの基準フィールドまたは対象フィールドの値が類似していることを示します。

  サポートされる測定は、「エントロピー」と「**Gini**」です。これらの測定は、ブランチの所属カテゴリーの確率に基づいています。
- 最大ツリー深さ。 ルート・ノード下でツリーが成長可能な最大レベル数 (サンプルが再帰的に分割される回数)。このプロパティーのデフォルト値は 10 であり、このプロパティーに設定できる最大値は 62 です。

  注: モデル・ナゲットのビューアーにモデルがテキスト表示される場合、最大 12 のツリー・レベルが表示されます。

分割の基準。これらのオプションは、ツリーの分割をいつ停止するかを制御します。

- 分割の改善度の最小変化量。新しい分割がツリーで作成される前に不純度を減少する必要のある最小数。ツリー構築の目標は、似かよった出力値を持つサブグループを作成して、それぞれのノード内における不純度を最小にすることです。ブランチが適切に分割されて不純度が分割基準によって指定された値を下回ると、ブランチは分割されません。
- 分割の最小インスタンス数。分割可能な最小レコード数。分割されていないレコードがこの数より少ない場合、これ以上分割は行われません。このフィールドを使用すると、ツリー内に小さいサブグループが作成されないようにすることができます。

統計: このパラメーターは、モデルに含まれる統計の数を定義します。以下のいずれかのオプションを選択します。

- 「すべて」。列に関連したすべての統計と値に関連したすべての統計が含まれます。

  注: このパラメーターを使用すると統計の最大数が含まれるため、システムのパフォーマンスに影響する場合があります。モデルをグラフィカル形式で表示しない場合は、「なし」を指定します。
- 「列」。列に関連した統計が含まれます。
- なし: モデルのスコアリングに必要な統計のみが含まれます。

IBM DB2 for z/OS モデル - ディシジョン ツリー ノード - クラスの重み

ここでは、重みを各クラスに割り当てることができます。デフォルトでは、1 の値をすべてのクラスに割り当て、重みを等しくします。異なるクラス・ラベルに異なる数値の重みを指定することによって、アルゴリズムが特定のクラスの学習セットに重みを付けるよう指定します。

重みを変更するには、「重み」 列で変更する重みをダブルクリックし、必要に応じて変更します。

値。 対象フィールドの可能な値から派生した、クラス ラベルのセット。

重み: 特定のクラスに割り当てられる重み。大きな重みをクラスに割り当てると、モデルは、他のクラスと比べてそのクラスに対して敏感になります。

インスタンスの重みと組み合わせてクラスも重みを使用できます。

IBM DB2 for z/OS モデル - ディシジョン ツリー ノード - ツリーの剪定

剪定オプションを使用して、ディシジョン・ツリーの剪定基準を指定できます。剪定の目的は、新しいデータに対して予期した精度が改善されない成長しすぎたサブグループを削除することによって、オーバーフィットのリスクを軽減することにあります。

剪定の測定:デフォルトの剪定の測定、「精度」は、ツリーから葉を削除した後、推定されたモデルの精度が可能な上限内にあるようにします。剪定を適用しながらクラスの重みを考慮に入れる場合、「重みつき精度」 を使用します。

剪定するデータ: 新しいデータに対する想定された精度を推定するには、学習データの一部またはすべてを使用できます。また、指定されたテーブルの別の剪定データセットを使用できます。

- すべての学習データを使用: このオプション (デフォルト) は、モデルの精度を推定するためにすべての学習データを使用します。
- 剪定に次のパーセンテージ学習データを使用: 剪定データに指定した割合で、一方は学習用、一方は剪定用です。

- ストリームを実行するごとにデータを同じ方法で区分するようランダム シードを指定する場合、「結果を再現」 を選択します。「剪定に使用するシード」 フィールドで整数を指定するか、または 「生成」 をクリックすると、擬似無作為の整数を作成します。
- 既存のテーブルのデータを使用: モデルの精度を推定するために個別の剪定データセットのテーブル名を指定します。学習データを使用するより信頼性が高いと見なされます。

IBM DB2 for z/OS モデル - 回帰ツリー

回帰ツリーは、数値型対象フィールドの値に基づいて同じ種類のサブセットを派生させるために、ケースのサンプルを繰り返し分割するツリーベースのアルゴリズムです。ディシジョン・ツリーと同様に、回帰ツリーはツリーの葉が十分に小さいか、均一なサブセットに対応するサブセットにデータを分解します。分割は、葉の平均値で合理的に予測できるよう、対象の属性の値のばらつきを減少させるために選択されます。

IBM DB2 for z/OS モデル - 回帰ツリーの作成オプション - ツリーの成長

ツリーの成長とツリーの剪定の作成オプションを設定できます。

ツリーの成長には、以下の作成オプションを使用できます。

最大ツリー深さ。 ルート・ノード下でツリーが成長可能な最大レベル数 (サンプルが再帰的に分割される回数)。デフォルトは 62 で、モデリングのための最大ツリー深度です。

注: モデル・ナゲットのビューアーにモデルがテキスト表示される場合、最大 12 のツリー・レベルが表示されます。

分割の基準。これらのオプションは、ツリーの分割をいつ停止するかを制御します。

- 評価指標の分割。このクラス評価測定では、ツリーの分割に最適な場所が評価されます。

  注: 現在、使用可能なオプションは「分散」のみです。
- 分割の改善度の最小変化量。新しい分割がツリーで作成される前に不純度を減少する必要のある最小数。ツリー構築の目標は、似かよった出力値を持つサブグループを作成して、それぞれのノード内における不純度を最小にすることです。ブランチが適切に分割されて不純度が分割基準によって指定された値を下回ると、ブランチは分割されません。
- 分割の最小インスタンス数。分割可能な最小レコード数。分割されていないレコードがこの数より少ない場合、これ以上分割は行われません。このフィールドを使用すると、ツリー内に小さいサブグループが作成されないようにすることができます。

統計: このパラメーターは、モデルに含まれる統計の数を定義します。以下のいずれかのオプションを選択します。

- 「すべて」。列に関連したすべての統計と値に関連したすべての統計が含まれます。

  注: このパラメーターを使用すると統計の最大数が含まれるため、システムのパフォーマンスに影響する場合があります。モデルをグラフィカル形式で表示しない場合は、「なし」を指定します。
- 「列」。列に関連した統計が含まれます。
- なし: モデルのスコアリングに必要な統計のみが含まれます。

IBM DB2 for z/OS モデル - 回帰ツリーの作成オプション - ツリーの剪定

剪定オプションを使用して、回帰ツリーの剪定基準を指定できます。剪定の目的は、新しいデータに対して予期した精度が改善されない成長しすぎたサブグループを削除することによって、オーバーフィットのリスクを軽減することにあります。

剪定の測定**:** 剪定の測定により、ツリーから葉を削除した後、モデルの推定精度を許容限度内に保つことができます。。次のいずれかを選択できます。

- **mse:** 平方平均誤差 - (デフォルト) 近似直線がデータ・ポイントにどれだけ近いかを測定します。
- **r2.** R-squared - 回帰モデルによって説明される従属変数の変動の比率です。
- **Pearson :** Pearson の相関係数 - 正規分布されている線形従属変数間の関係の強さを測定します。
- **Spearman:** Spearman の相関係数 - ピアソンの相関従って弱いと思われるが、実際には強いと考えられる非線形な関係を検出します。

剪定するデータ**:** 新しいデータに対する想定された精度を推定するには、学習データの一部またはすべてを使用できます。また、指定されたテーブルの別の剪定データセットを使用できます。

- すべての学習データを使用**:** このオプション (デフォルト) は、モデルの精度を推定するためにすべての学習データを使用します。
- 剪定に次のパーセンテージ学習データを使用**:** 剪定データに指定した割合で、一方は学習用、一方は剪定用です。

  ストリームを実行するごとにデータを同じ方法で区分するようランダム シードを指定する場合、「結果を再現」 を選択します。「剪定に使用するシード」 フィールドで整数を指定するか、または 「生成」 をクリックすると、擬似無作為の整数を作成します。
- 既存のテーブルのデータを使用**:** モデルの精度を推定するために個別の剪定データセットのテーブル名を指定します。学習データを使用するより信頼性が高いと見なされます。

IBM DB2 for z/OS モデル - TwoStep

TwoStep ノードは、大規模データセットにわたってデータをクラスタ化する方法を提供する TwoStep アルゴリズムを実装します。

このノードを使用すると、使用可能なリソース (例えば、メモリーや時間の制約) を考慮しながら、データをクラスタ化できます。

TwoStep アルゴリズムは、次の方法でデータをクラスタ化するデータベース マイニング アルゴリズムです。

1. クラスタ機能 (CF) ツリーが作成されます。このバランスに優れたツリーは、類似の入力レコードが同じツリー ノードの一部となる階層クラスタリングのためのクラスタ機能を格納します。
2. CF ツリーの葉は、最終クラスタリングの結果を生成するためにメモリー内で階層的にクラスタ化されます。クラスタの最適な数は自動的に決定されます。クラスタの最大数を指定する場合、指定された制限内のクラスタの最適な数が決定されます。
3. クラスタリングの結果は、K-Means アルゴリズムに似たアルゴリズムがデータに適用される 2 番目のステップで調整されます。

IBM DB2 for z/OS モデル - TwoStep フィールド オプション

フィールド オプションを設定することにより、上流ノードで定義されているフィールドの役割の設定を指定できます。フィールド割り当てを手動で行うこともできます。

項目の選択。上流のデータ型ノードまたは上流の入力ノードの「データ型」タブの役割設定を使用するには、このオプションを選択します。役割設定は、例えば対象や予測です。

カスタム・フィールド割り当ての使用: 対象、予測、およびその他の役割を手動で割り当てる場合、このオプションを選択します。

フィールド**:** 矢印を使用して、このリストの項目を右側の役割フィールドに手動で割り当てます。アイコンは、各役割フィールドの有効な測定の尺度を示します。

「レコード **ID**」。一意のレコード ID として使用されるフィールド。

予測値 **(**入力**):** 1 つまたは複数のフィールドを予測の入力として選択します。

IBM DB2 for z/OS モデル - TwoStep 作成オプション

作成オプションを設定することにより、用途に合わせてモデルの作成をカスタマイズできます。

デフォルト・オプションを使用してモデルを作成する場合は、「実行」をクリックします。

距離測度**:** このパラメーターは、データ・ポイント間の距離の測定方法を定義します。距離が大きくなると、非類似度も大きくなります。オプションは以下のとおりです。

- 対数尤度: この尤度測定により、変数の確率分布を求めます。連続変数は正規分布しているものと仮定し、カテゴリー変数は多項分布しているものと仮定します。すべての変数は独立しているものと仮定します。

クラスタ数。このパラメーターは、作成するクラスターの数を定義します。以下のオプションがあります。

- クラスタ数を自動的に計算。クラスタの数が自動的に計算されます。「最大」フィールドで最大クラスタ数を指定できます。
- クラスタ数を指定。作成するクラスタの数を指定します。

統計: このパラメーターは、モデルに含まれる統計の数を定義します。以下のオプションがあります。

- 「すべて」。列に関連したすべての統計と値に関連したすべての統計が含まれます。

  注**:** このパラメーターを使用すると統計の最大数が含まれるため、システムのパフォーマンスに影響する場合があります。モデルをグラフィカル形式で表示しない場合は、「なし」を指定します。
- 「列」。列に関連した統計が含まれます。
- なし**:** モデルのスコアリングに必要な統計のみが含まれます。

結果の再現: ランダム・シードを設定して分析を複製する場合は、このチェック・ボックスを選択します。整数を指定するか、「生成」をクリックして整数の疑似乱数を作成します。

IBM DB2 for z/OS モデル - TwoStep ナゲット - 「モデル」タブ

「モデル」タブには、クラスターのフィールドの要約統計量および分布を表示する各種のグラフィック表示があります。モデルからデータをエクスポートしたり、ビューをグラフィックとしてエクスポートしたりすることができます。

IBM DB2 for z/OS モデルの管理

DB2 for z/OS モデルは、他の IBM SPSS Modeler のモデルのようにキャンバスおよび「モデル」パレットに追加され、ほとんど同じように使用できます。

DB2 for z/OS で直接的にデータをスコアリングするには、以下のステップを実行します。

1. データが配置されている DB2 for z/OS データベースに SPSS Scoring Adapter をインストールします。
2. ストリームが、データが配置されている DB2 for z/OS データベースに接続することを確認します。

IBM DB2 for z/OS モデルのスコアリング

モデルは、キャンバス上で金色のモデル・ナゲット・アイコンで示されます。ナゲットの主な目的は、データをスコアリングし、予測を生成、またはモデルのプロパティーの詳細な分析を可能にすることです。スコアは、このセクションで後述するように、ナゲットにテーブル・ノードを接続し、ストリームのそのブランチを実行することによって見えるようにすることができる、1 つ以上の追加データ・フィールドの形式で追加されます。デシジョンツリーや回帰ツリーのダイアログ・ボックスなど、いくつかのナゲットのダイアログボックスには、さらにモデルの視覚的表現を提供する「モデル」タブがあります。

追加フィールドは、対象フィールドの名前に追加された接頭辞 `$<id>-` によって区別されます。 `<id>` はモデルによって異なり、追加される情報の種類を識別します。さまざまな識別子について、それぞれのモデル・ナゲットのトピックで説明されています。

スコアを表示するには、次の手順を実行します。

1. テーブルノードをモデル・ナゲットに接続します。
2. テーブル・ノードを開きます。
3. 「実行」をクリックします。
4. テーブル出力ウィンドウの右側にスクロールし、追加フィールドとすれらのスコアを表示します。

注: スコアリング・プロセスは、アクセラレーターでは実行されず、DB2 で実行されます。そのため、スコアリングの入力テーブルは物理的に DB2 に配置する必要があります。したがって、スコアリング入力としては、DB2 ベースのテーブルまたはアクセラレーテッド・テーブルのみを使用できます。ストリームがアクセラレーター専用テーブルを使用する場合、`「THE STATEMENT CANNOT BE EXECUTED BY DB2 OR IN THE ACCELERATOR」` というエラーが発生します。

IBM DB2 for z/OS ディシジョン ツリー モデル ナゲット

ディシジョン・ツリー・モデル・ナゲットでは、モデリング操作の出力を表示し、モデルをスコアリングするためのオプションを設定することもできます。

ディシジョン ツリー モデル ナゲットを含むストリームを実行すると、ノードは 2 つの新しいフィールドを追加します。その名前は、対象に由来します。

表 26. ディシジョン・ツリーのモデルスコアリング・フィールド：

追加フィールドの名前	意味
$I-*target_name*	現在のレコードの予測値。
$IP-*target_name*	予測の確信度 (0.0 ～ 1.0)。

注: DB2 for z/OS の制約により、列名は切り捨てられる場合があります。

IBM DB2 for z/OS ディシジョン ツリー ナゲット -「モデル」タブ

「モデル」タブには、ディシジョン・ツリー・モデルの「予測値の重要度」がグラフィカル形式で表示されます。棒の長さは、予測値の重要度を表しています。

IBM DB2 for z/OS ディシジョン ツリー ナゲット - 「ビューア」タブ

「ビューア」タブには、SPSS Modeler でデシジョン ツリー モデルが表示されるのと同じ方法で、ツリー モデルのツリー プレゼンテーションが表示されます。

IBM DB2 for z/OS K-Means モデル ナゲット

K-Means モデル・ナゲットには、クラスター化モデルが取得したすべての情報と、学習データと推定プロセスに関する情報が含まれます。

K-Means モデル・ナゲットを含むストリームを実行すると、そのレコードの所属クラスターと割り当てられたクラスターの中心からの距離を含む 2 つの新規フィールドが追加されます。新規フィールド名はモデル名から派生し、所属クラスターのフィールドには接頭辞の `$KM-`、クラスターの中心からの距離のフィールドには接頭辞の `$KMD-` が付けられます。例えば、モデルの名前が `Kmeans` の場合、新規フィールド名は `$KM-Kmeans` と `$KMD-Kmeans` になります。

注**:** DB2 for z/OS の制約により、列名は切り捨てられる場合があります。

IBM DB2 for z/OS K-Means ナゲット -「モデル」タブ

「モデル」タブには、クラスターのフィールドの要約統計量および分布を表示する各種のグラフィック表示があります。モデルからデータをエクスポートしたり、ビューをグラフィックとしてエクスポートしたりすることができます。

IBM DB2 for z/OS Naive Bayes モデル ナゲット

Naive Bayes モデル ナゲットを含むストリームを実行すると、ノードは 2 つの新しいフィールドを追加します。その名前は、対象名に由来します。

表 *27. Naive Bayes* のモデルスコアリング・フィールド：

追加フィールドの名前	意味
$I-*target_name*	現在のレコードの予測値。
$IP-*target_name*	予測の確信度 (0.0 ～ 1.0)。

注**:** DB2 for z/OS の制約により、列名は切り捨てられる場合があります。

モデルナゲットにテーブル・ノードを接続して表のノードを実行することにより、追加フィールドを表示することができます。

IBM DB2 for z/OS 回帰ツリー モデル ナゲット

回帰ツリー モデル ナゲットを含むストリームを実行すると、ノードは 2 つの新しいフィールドを追加します。その名前は、対象名に由来します。

表 *28.* 回帰ツリーのモデルスコアリング・フィールド：

追加フィールドの名前	意味
$I-*target_name*	現在のレコードの予測値。
$IS-*target_name*	予測値の推定標準偏差。

注: DB2 for z/OS の制約により、列名は切り捨てられる場合があります。

モデルナゲットにテーブル・ノードを接続して表のノードを実行することにより、追加フィールドを表示することができます。

IBM DB2 for z/OS 回帰ツリー ナゲット - 「モデル」タブ

「モデル」タブには、回帰ツリー・モデルの「予測値の重要度」がグラフィカル形式で表示されます。棒の長さは、予測値の重要度を表しています。

IBM DB2 for z/OS 回帰ツリー ナゲット - 「ビューア」タブ

「ビューア」タブには、SPSS Modeler で回帰ツリー モデルが表示されるのと同じ方法で、ツリー モデルのツリー プレゼンテーションが表示されます。

IBM DB2 for z/OS TwoStep モデル ナゲット

TwoStep モデル・ナゲットを含むストリームを実行すると、そのレコードの所属クラスターと割り当てられたクラスターの中心からの距離を含む 2 つの新規フィールドが追加されます。新規フィールド名はモデル名から派生し、所属クラスターのフィールドには接頭辞の $TS-、クラスターの中心からの距離のフィールドには接頭辞の $TSD- が付けられます。例えば、モデルの名前が MDL の場合、新規フィールド名は $TS-MDL と $TSD-MDL になります。

特記事項

本書は米国 IBM が提供する製品およびサービスについて作成したものです。この資料は、IBM から他の言語でも提供されている可能性があります。ただし、これを入手するには、本製品または当該言語版製品を所有している必要がある場合があります。

本書に記載の製品、サービス、または機能が日本においては提供されていない場合があります。日本で利用可能な製品、サービス、および機能については、日本 IBM の営業担当員にお尋ねください。本書で IBM 製品、プログラム、またはサービスに言及していても、その IBM 製品、プログラム、またはサービスのみが使用可能であることを意味するものではありません。これらに代えて、IBM の知的所有権を侵害することのない、機能的に同等の製品、プログラム、またはサービスを使用することができます。ただし、IBM 以外の製品とプログラムの操作またはサービスの評価および検証は、お客様の責任で行っていただきます。

IBM は、本書に記載されている内容に関して特許権 (特許出願中のものを含む) を保有している場合があります。本書の提供は、お客様にこれらの特許権について実施権を許諾することを意味するものではありません。実施権についてのお問い合わせは、書面にて下記宛先にお送りください。

〒*103-8510*
東京都中央区日本橋箱崎町*19*番*21*号
日本アイ・ビー・エム株式会社
法務・知的財産
知的財産権ライセンス渉外

IBM およびその直接または間接の子会社は、本書を特定物として現存するままの状態で提供し、商品性の保証、特定目的適合性の保証および法律上の瑕疵担保責任を含むすべての明示もしくは黙示の保証責任を負わないものとします。国または地域によっては、法律の強行規定により、保証責任の制限が禁じられる場合、強行規定の制限を受けるものとします。

この情報には、技術的に不適切な記述や誤植を含む場合があります。本書は定期的に見直され、必要な変更は本書の次版に組み込まれます。IBM は予告なしに、随時、この文書に記載されている製品またはプログラムに対して、改良または変更を行うことがあります。

本書において IBM 以外の Web サイトに言及している場合がありますが、便宜のため記載しただけであり、決してそれらの Web サイトを推奨するものではありません。それらの Web サイトにある資料は、この IBM 製品の資料の一部ではありません。それらの Web サイトは、お客様の責任でご使用ください。

IBM は、お客様が提供するいかなる情報も、お客様に対してなんら義務も負うことのない、自ら適切と信ずる方法で、使用もしくは配布することができるものとします。

本プログラムのライセンス保持者で、(i) 独自に作成したプログラムとその他のプログラム (本プログラムを含む) との間での情報交換、および (ii) 交換された情報の相互利用を可能にすることを目的として、本プログラムに関する情報を必要とする方は、下記に連絡してください。

IBM Director of Licensing
IBM Corporation
North Castle Drive, MD-NC119
Armonk, NY 10504-1785
US

本プログラムに関する上記の情報は、適切な使用条件の下で使用することができますが、有償の場合もあります。

本書で説明されているライセンス・プログラムまたはその他のライセンス資料は、IBM 所定のプログラム契約の契約条項、IBM プログラムのご使用条件、またはそれと同等の条項に基づいて、IBM より提供されます。

記載されている性能データとお客様事例は、例として示す目的でのみ提供されています。実際の結果は特定の構成や稼働条件によって異なります。

IBM 以外の製品に関する情報は、その製品の供給者、出版物、もしくはその他の公に利用可能なソースから入手したものです。IBM は、それらの製品のテストは行っておりません。したがって、他社製品に関する実行性、互換性、またはその他の要求については確証できません。 IBM 以外の製品の性能に関する質問は、それらの製品の供給者にお願いします。

IBM の将来の方向性および指針に関する記述は、予告なく変更または撤回される場合があります。これらは目標および目的を提示するものにすぎません。

本書には、日常の業務処理で用いられるデータや報告書の例が含まれています。より具体性を与えるために、それらの例には、個人、企業、ブランド、あるいは製品などの名前が含まれている場合があります。これらの名称はすべて架空のものであり、類似する個人や企業が実在しているとしても、それは偶然にすぎません。

商標

IBM、IBM ロゴおよび ibm.com は、世界の多くの国で登録された International Business Machines Corporation の商標です。他の製品名およびサービス名等は、それぞれ IBM または各社の商標である場合があります。現時点での IBM の商標リストについては、http://www.ibm.com/legal/copytrade.shtml をご覧ください。

Adobe、Adobe ロゴ、PostScript、PostScript ロゴは、Adobe Systems Incorporated の米国およびその他の国における登録商標または商標です。

インテル、Intel、Intel ロゴ、Intel Inside、Intel Inside ロゴ、Centrino、Intel Centrino ロゴ、Celeron、Xeon、Intel SpeedStep、Itanium、および Pentium は、Intel Corporation または子会社の米国およびその他の国における商標または登録商標です。

Linux は、Linus Torvalds の米国およびその他の国における登録商標です。

Microsoft、Windows、Windows NT および Windows ロゴは、Microsoft Corporation の米国およびその他の国における商標です。

UNIX は The Open Group の米国およびその他の国における登録商標です。

Java およびすべての Java 関連の商標およびロゴは Oracle やその関連会社の米国およびその他の国における商標または登録商標です。

製品資料に関するご使用条件

これらの資料は、以下のご使用条件に同意していただける場合に限りご使用いただけます。

適用条件

IBM Web サイトの「ご利用条件」に加えて、以下のご使用条件が適用されます。

個人的使用

これらの資料は、すべての著作権表示その他の所有権表示をしていただくことを条件に、非商業的な個人による使用目的に限り複製することができます。ただし、IBM の明示的な承諾をえずに、これらの資料またはその一部について、二次的著作物を作成したり、配布 (頒布、送信を含む) または表示 (上映を含む) することはできません。

商業的使用

これらの資料は、すべての著作権表示その他の所有権表示をしていただくことを条件に、お客様の企業内に限り、複製、配布、および表示することができます。ただし、IBM の明示的な承諾をえずにこれらの資料の二次的著作物を作成したり、お客様の企業外で資料またはその一部を複製、配布、または表示することはできません。

権利

ここで明示的に許可されているもの以外に、資料や資料内に含まれる情報、データ、ソフトウェア、またはその他の知的所有権に対するいかなる許可、ライセンス、または権利を明示的にも黙示的にも付与するものではありません。

資料の使用が IBM の利益を損なうと判断された場合や、上記の条件が適切に守られていないと判断された場合、IBM はいつでも自らの判断により、ここで与えた許可を撤回できるものとさせていただきます。

お客様がこの情報をダウンロード、輸出、または再輸出する際には、米国のすべての輸出入 関連法規を含む、すべての関連法規を遵守するものとします。

IBM は、これらの資料の内容についていかなる保証もしません。これらの資料は、特定物として現存するままの状態で提供され、商品性の保証、特定目的適合性の保証および法律上の瑕疵担保責任を含むすべての明示もしくは黙示の保証責任なしで提供されます。

索引

日本語, 数字, 英字, 特殊文字の順に配列されています。なお, 濁音と半濁音は清音と同等に扱われています。

［ア行］

［カ行］

［サ行］

［タ行］

［ナ行］

［ハ行］

［マ行］

［ヤ行］

［ラ行］

A

B

D

E

G

I

K

M

N

O

P

S

T

Z

Printed in Japan

日本アイ・ビー・エム株式会社
〒103-8510 東京都中央区日本橋箱崎町19-21