

Visual Mining Studio バージョン 8.2 の変更点

2016 年 3 月

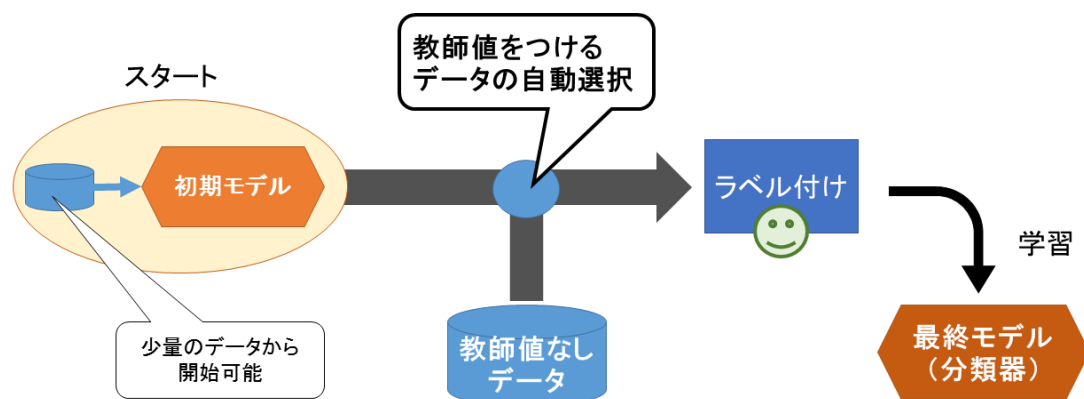
株式会社 NTT データ数理システム

主な変更点

1. 新アイコン「対話型モデル 学習」「対話型モデル 教師値設定」「対話型モデル 予測」の追加

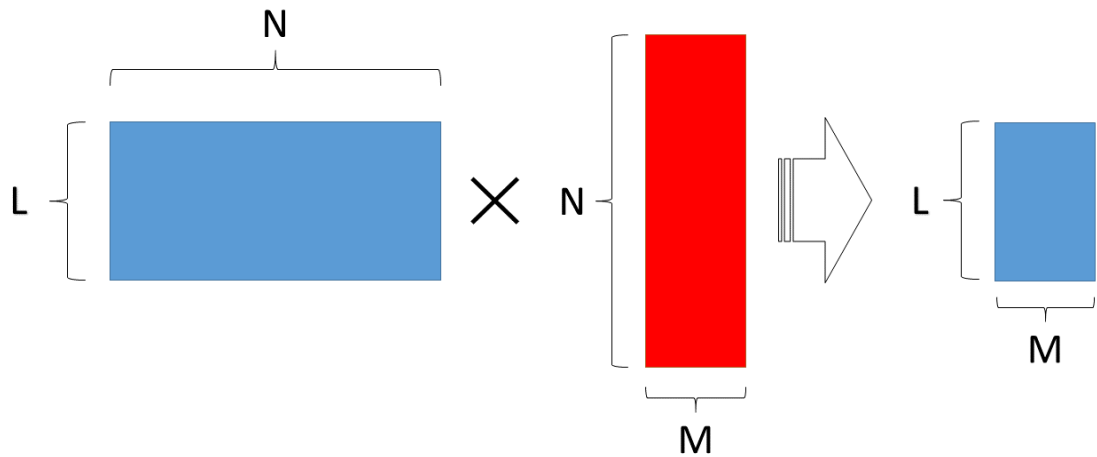
対話型モデル分析機能により、一部のデータにしか教師値(正解値)が付与されていないケースでも分析モデルの構築・効率的な改善および予測が行えるようになります。

通常、大量の学習データに対して教師値を付与する作業は非常に煩雑になり、大きな労力を要します。対話型モデルの分析機能では、少量の教師値データから初期モデルを生成し、予測精度をより大きく向上させるデータを優先的に教師値を付与(ラベル付け)することができるため、同じ労力でより大きな精度向上が見込めます。



2. 新アイコン「ランダム次元圧縮」の追加

ランダム次元圧縮は、行列積演算により高速に次元数(列数)を減らすことができます。スパースなデータの場合、データの情報量に対して分析処理にかかる時間が長くなってしましますが、次元圧縮によって情報量をできるだけ保ったまま高速に処理を行うことができます。行同士の距離(特にユークリッド距離)の大小関係を高確率で維持したまま次元数を減らすことができるため、k-NN 分析やクラスタ分析の前処理として有効です。



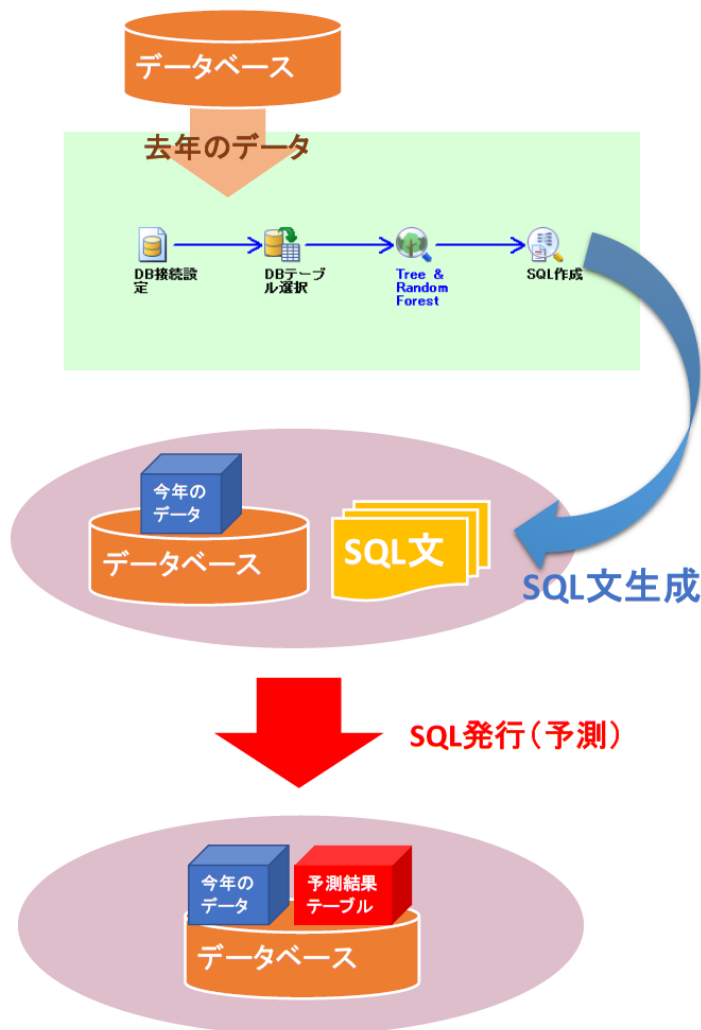
3. 新アイコン「SQL 作成」の追加

SQL 作成アイコンにより、VMS 上のモデリングアイコン¹からデータベース上で直接予測が行える SQL 文を生成することができます。例えば、データベースにあった今年のデータからモデルを作成し、翌年に新しくデータベースに追加されたデータを用いて予測する場合、今までは新しく追加されたデータを VMS / VAP 内に新たにに取り込む必要がありました。SQL 作成アイコンにより作成された SQL 文があれば、データベースに予測用データを準備し、作成された SQL 文を実行するだけで簡単に予測を行うことができます。予測結果はデータベース内のテーブルとして生成されます。

次ページの図は SQL 作成アイコン使用方法の概略図です。

¹ SQL 作成アイコンが対応しているモデリングアイコンは、以下の 4 アイコンです。

- Decision Tree アイコン
- Tree&RandomForest アイコン (ただし、Radom Forest は除く)
- クラスアソシエーション アイコン
- アソシエーション分析 アイコン



4. 新しいアイコン「DB 接続設定」「DB テーブル選択」「DBSQL」の追加

- モダンなユーザーインターフェイスに刷新しました
- ODBC の接続情報が暗号化されて保持されるようになったため、プロジェクトを開き直した際に再指定をする必要がなくなりました
- データベースの文字コードが指定できるようになり、Shift-JIS 以外に UTF-8 のデータベースにも対応しました
- これまでは 64bitOS の場合 64bit の ODBC ドライバと 32bit の ODBC ドライバを設定していただく必要がありましたが、64bit の ODBC ドライバのみを設定すればよいようになりました

新機能一覧

追加されたアイコン

対話型モデル 学習

少量の正解データから初期学習を行います

対話型モデル 教師値設定

「対話型モデル 学習」アイコンから接続して利用し、教師値が未だ振られていない学習データに対してユーザが教師値を指定することで精度を向上させます。この際に提示されるサンプルは精度向上に効果的なサンプルが選択されます。

対話型モデル 予測

「対話型モデル 学習」又は「対話型モデル 教師値設定」アイコンから接続し、教師値が設定されていない未知のデータに対して予測を行います。

SQL 作成

Tree & Random Forest(ただし、Random Forest を除く)、Decision Tree、クラスアソシエーション、アソシエーション分析の結果から SQL 文を作成してファイルに出力します。作成した SQL をデータベース上にある予測対象のテーブルに適用することで、データベース上で直接予測を行うことができます。

ランダム次元圧縮

対象データのサンプル間の距離(ユークリッド距離)を維持したまま、列数を削減します。距離に基づいて行う分析(クラスタ分析など、ユークリッド距離を指定した場合)の前に列数を削減しておくこと処理の高速化が見込めます。

DB 接続設定

DB テーブル選択

DBSQL

ユーザーインターフェイスを刷新しました。

また、64bit OS 環境で ODBC 接続を行う場合、64bit 用 ODBC ドライバと 32bit 用 ODBC ドライバの両方に対して DSN を設定する必要がありました。が、64bit 用 ODBC ドライバに対してのみ DSN 設定を行えばよいようになりました。

新しいスクリプト関数

make_rp_matrix

ランダム次元圧縮の計算に用いる乱数テーブルを生成します。

対象テーブルの列数と同じ行数を持ち、指定した次元数の列数を持つ乱数テーブルを生成します。

sparse_matrix_product

リスト形式テーブルの行列積を計算します。

通常の行列演算(%%)では疎な行列を表現するのに適したリスト形式テーブルのままでは計算できませんでしたが、この関数を使うことでリスト形式テーブルのまま行列積を計算することができます。

その他

Windows10 への対応

右クリックでのアイコン連結

ノート PC などでホイールボタンがなく、左ボタンと右ボタンを同時にクリックできない環境などでも右クリックをドラッグすることでアイコン連結を行うことができるようになりました。

追加機能

改善された機能

クリーニング

- 欠損基準値として、非数・無限を選択できるようになりました。
- 欠損補填値として、欠損値(-INF)を指定できるようになりました。

Tree & RandomForest

- 節点最小データ数、節点の不純度、重み、分位値、シード値の項目を、有効桁数下 16 桁まで入力できるようになりました。
- 最大分岐数の指定が Shift キーや Ctrl キーを押すことで一括指定できるようになりました。
- ランダムフォレスト時の分岐変数候補数の選択肢に $\sqrt{N_{exp}}$ (説明変数選択個数の平方根)が追加されました。

データハンドリング

- 既存の結果名で実行しようとした際に結果を上書きできるようになりました。
- データハンドリング時のデータビューアの表示桁数が VAP のシステムパラメータ設定の値を利用できるようになりました。
- 重み付き集計において、レコード値を変換と統計的重み付けの二種類の計算方法を選べるようになりました。

Aggregate

重み付き集計において、レコード値を変換と統計的重み付けの二種類の計算方法を選べるようになりました。

その他の改善された機能

バッチ実行

VAPConsoleBatch.exe でバッチ実行をする際、正常終了時は 0 を、異常終了時は 0 以外の終了ステータスコードを返すようになりました。