Pentaho データマイニング

ホワイトペーパー

株式会社 KSK アナリティクス

目次

要約

オペレーションを最適にし、得られる ROI を最大化するための予測分析の事例が多くのビジネスで見られます。データマイニングによって提供される予測分析が、ビジネスのパフォーマンスを向上させることは明確です。しかし、その価値を得るためのプロセスはそれほど明確ではありません。

このホワイトペーパーは、Pentaho BI スイートを使用することで予想モデルの展開のためのフレキシブルなオプションについて概説します。特に、私たちは、急速に変わっていくデータから共通した有益なシナリオを予測したいと考えています。Pentaho の 100%の Java ベースのコンポーネント指向の軽量アーキテクチャーは、カスタマイズ可能なソリューション構築を容易にします。

ユースケース

あなたが例えば、Salesforce.com などの CRM システムでデータを保存しており、予想モデルを使用して販売見込み予測のためのスコアを作成したいというシナリオを考えてみましょう。予想モデルがあると仮定する場合、そのようなシナリオは、顧客と商談のデータを抽出し、モデルによって必要とされた予測フィールドを引き出して、スコアを生成して、次に CRM システムをアップデートするといった流れとなるでしょう。カスタムコードを作成することなく、そのようなプロセスを行うのが理想的です。また、小さいデータから始めてデータ量が増加するのに従って処理能力を拡張するランタイム環境を持つことも必要です。 その上、異なったデータマイニングモデルを「ドロップイン」したり、現行モデルをリフレッシュする機能を持っていることは、ビジネス環境との同期を担保します。

Pentaho のエンタープライズクラスのデータ統合能力と強力なデータマイニングスイートの組み合わせは、そのようなプロセスを簡単にデプロイして、開発し、維持する環境を提供します。 また、Pentaho のコマーシャルオープンソースモデルは ROI（投資利益率）を最大化することができます。

Pentaho データマイニングソリューションの利点

高い費用対効果を達成するために、データマイニングのデプロイは以下を考慮する必要があります。

１．簡単で素早いモデル開発
２．統合を容易にするオープンなアーキテクチャー
３．全体最適なデータマイニングライフサイクルと自動化できるデプロイメントオプション
４．低い総合的な TCO

１．簡単で素早いモデル開発

Weka プロジェクトに基づく Pentaho Data Mining は分析モデルを構成するのに最新の環境を提供します。 それは分類、回帰、クラスタリング、属性選択、およびデータ前処理のための 200 以上のアルゴリズムをデータマイニングツールの包括的なスイートとして提供します。その上、アカデミックとのアクティブなコミュニティと強い結びつきは、最新のツールキットを提供します。 実験的なデータマイニングの全体のプロセスの使い易いグラフィカルインタフェースとサポートサービスは予測モデルの検証を確実にします。
Pentaho Data Mining ツールキットは、一つの Java Jar アーカイブとしてライブラリーとなっており、容易にデプロイすることができます。

２．オープンでスタンダードなアーキテクチャー

Pentaho の主力製品はオープンソースライセンスの下ですべて配布されます。 これが、すべてのソースコードが自由に利用可能であることを意味し、プロプライエタリーなものからのロックインをプロテクトします。 また、Pentaho の製品は、オープンスタンダードをサポートすることを目指しています。例えば、Pentaho DataMining は PMML(Predictive Modeling Markup Language)形式で外部的に作成されたモデルのインポートをサポートします。

３．完全なライフサイクル展開

Pentaho Data Integration(PDI)は予測ソリューションをデプロイする理想的なプラットフォームを提供します。 100%の Java アーキテクチャーは Pentaho Data Mining との統合を容易にします。 PDI は、データ量が増加するのに従ってクラスタリングが可能であり、スコアソリューションを許容する標準機能があります。 PDI の性能は、トレーニングとデータマイニングモデルの完全なプロセス自動化を行います。 これにより、一から手動でモデルを作り直すコストを削減できます。

４．低いトータルコスト

プロプライエタリーなソフトウェアに比べて、大きな初期費用をかけることなく、低額なサポートが得られることは、Pentaho の予測スコアリングのデプロイメントを容易なものにします。 Pentaho 製品は評価を簡単に行うために自由にダウンロードでき、リスクフリーに利用可能です。

Pentaho データ統合とのデータ抽出と準備(PDI)

Pentaho Data Integration のストリーミングアプローチは予測スコアソリューションをデプロイする環境を提供するだけでなく、パワフルな変換・フィルターのオペレーションも提供します。 PDI は、すぐモデル生成を行うため、Weka ネイティブの ARFF 形式で容易にデータセットをエクスポートできます。

.

・グラフィカルなドラッグ・アンド・ドロップ開発環境に関するメタデータ駆動のアプローチ
・100 以上のオブジェクトによるリッチなトランスフォーメーションライブラリ
・RDBMS ベースの、または、ファイルベースのリポジトリ
・ほとんどの商用オープンソースデータベース、Excel、CSV、XML、テキスト、ウェブサービスファイルなどを含む広範なサポート
・クラスタリング、ビッグデータ対応する ETL エンジン
・Pentaho プラットフォームとの統合によるスケジュール、セキュリティ、監査、モニタリング

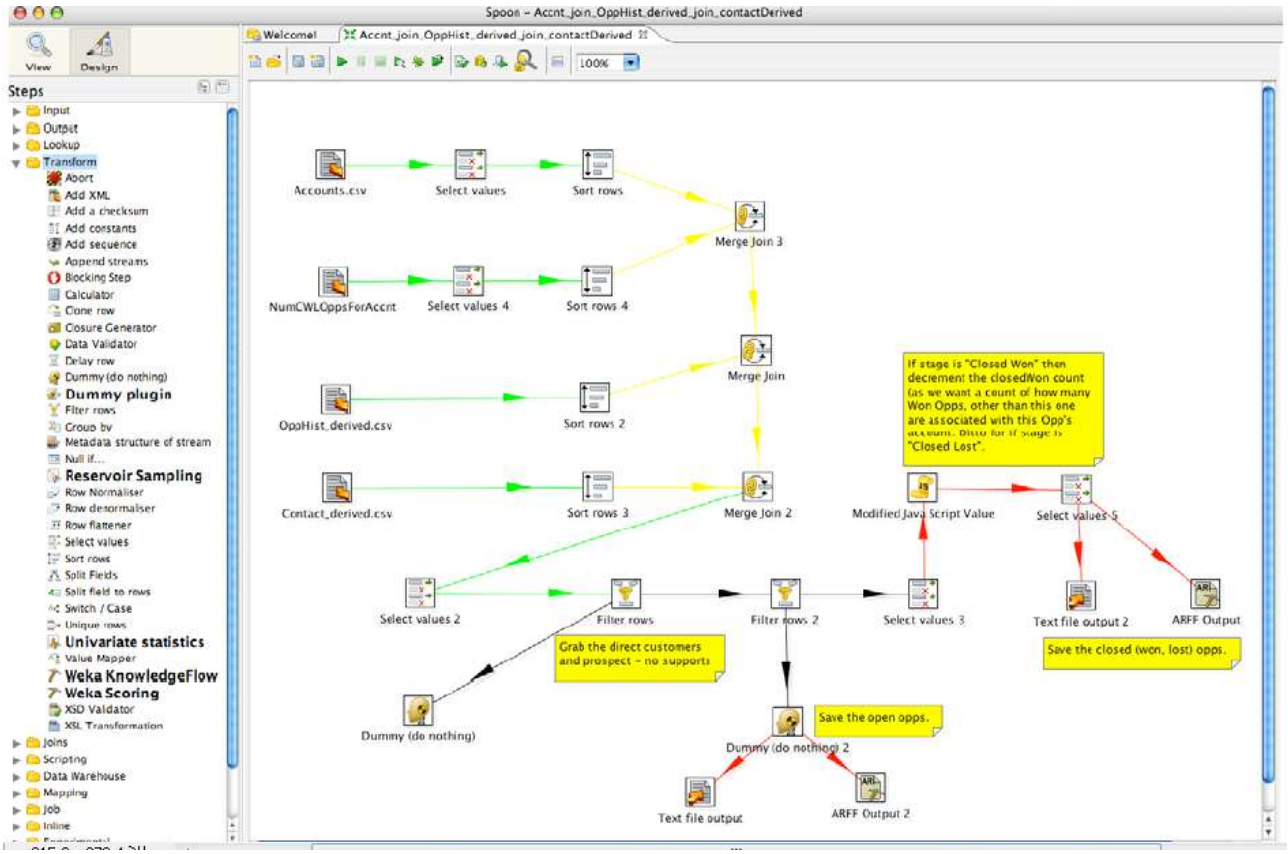

図 1 PDI のグラフィカルな ETL 開発環境

Pentaho データマイニングでのモデル開発

Pentaho Data Mining は、CRISP-DM(Data Mining のための CRoss Industry Standard Process)などの業界基準方法論に応じて、予想モデルを開発するために完全な環境を提供します。 使い易いグラフィカルなフロントエンドによる高範囲のデータマイニングアルゴリズムは、モデルを容易に開発するのを可能にします。 Pentaho Data Mining の機能:

.

・69 個のデータ前処理フィルター
・116 分類/回帰アルゴリズム
・11 クラスタリングアルゴリズム
・18 属性/サブセット検証+12 検索アルゴリズム
・Explorer と Knowledge Flow アプリケーションはそれぞれの開発プロセスを効率化します。
・交叉検証やシグニフィカントテスト等の統計的手法を使用する学習アルゴリズム
・大規模な実験的比較のための Experimenter アプリケーション
・散布図マトリクスや ROC/リフトチャート、ツリー、グラフなどのグラフィカルなビジュアル化
・バイナリーまたは XML 形式のモデルでのデータマイニングプロセスのエクスポート
・PMML 形式における外部的に作成されたモデルのインポート

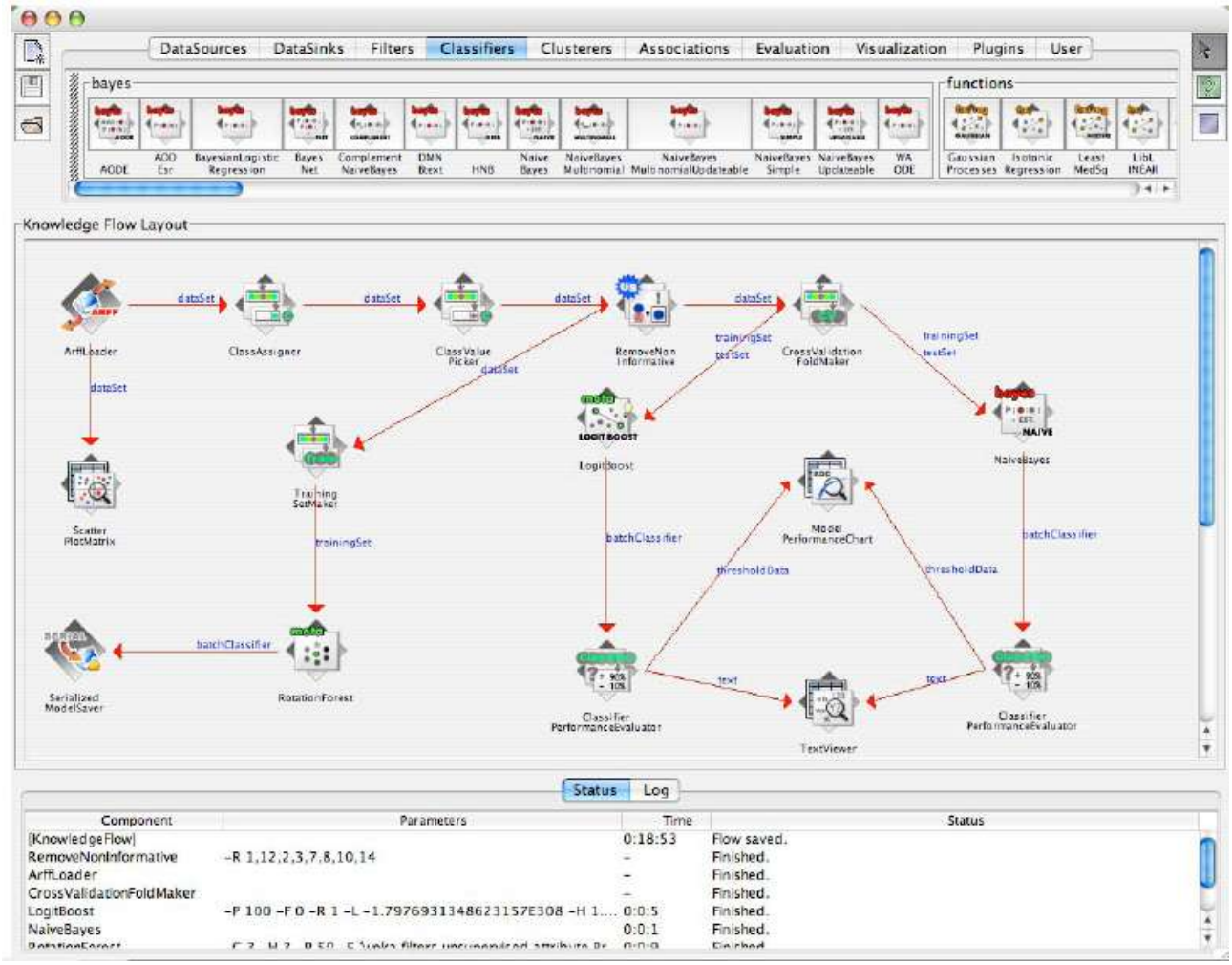

図 2 Pentaho Data Mining の Knowledge Flow でのデータマイニングプロセスの開発

Pentaho データ統合でのモデルデプロイとリフレッシュ…

ETL プロセスの一部として PDI で予想モデル(固有の Weka モデルか PMML に表されたもののどちらか)をデプロイするのは、Weka Scoring トランスフォーメーションノードを使用することで容易に達成されます。 このコンポーネントの機能は以下を提供します。:

・連続した Weka モデルのためにバイナリー、XML か PMML 形式のいずれかでサポート
・モデルはファイルシステムまたリポジトリのメタデータとして PDI データ変換として取り込み
・モデルによって使用されたものに対する入って来るフィールドの自動マッピングとタイプの照合
・分類、回帰、およびクラスタリングモデルのサポート
・ラベル(分類かクラスタリング)としての予測か確率分布のアウトプット

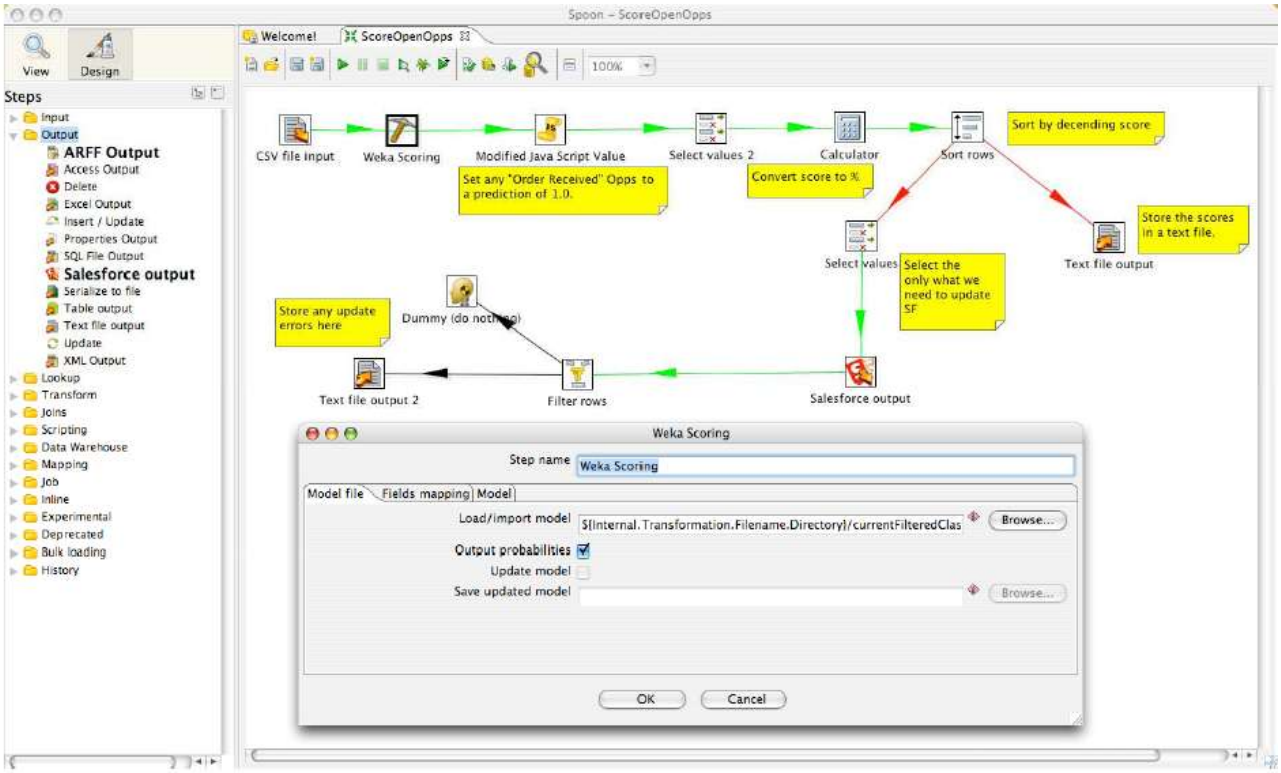

図 3 Pentaho Data Mining 予想モデルを使用することで販売機会をスコアリングする PDI データ変換

時間の経過とともに、新しいデータが集められるのに従って、モデルの予測性能は減退するかもしれません。 これは、データの変化によって引き起こされる場合があり、しばしば「概念ドリフト」と呼ばれます。例えば、私たちのビジネスが発展するのに応じて、売れている顧客のタイプは、時間がたつにつれて、徐々に変化するかもしれません。 これは、最新のマテリアルを使用することで予想モデルの周期的な構築やリフレッシュを必要とします。 予定されている ETL ジョブの一部としてこのプロセスを自動化するのは Weka KnowledgeFlow トランスフォーメーションノードを使用します。

.

・Knowledge Flow プロセスは、ファイルシステムから直接読みこむこともできます。またはリポジトリにおける PDI データ変換としてインポートすることができます。

.・ETL トランスフォーメーションからの受信データを Knowledge Flow プロセスにつなぐことができます。または単に Knowledge Flow プロセスをトリガーことができます。(ソースのデータはネイティブのままで)

・リザーバーサンプリングのブルドインサポートは、ETL トランスフォーメーションから、様々なバッチ学習アルゴリズムまで使用することができます。

・Weka フィルターで作り出されたモデルからのアウトプットは川下の PDI トランスフォーメーションステップに向けることもできます。

・埋め込まれた Knowledge Flow エディタは、全体のデータマイニングプロセスが PDI のグラフィカルなデザイン環境の内部であることを許容します。

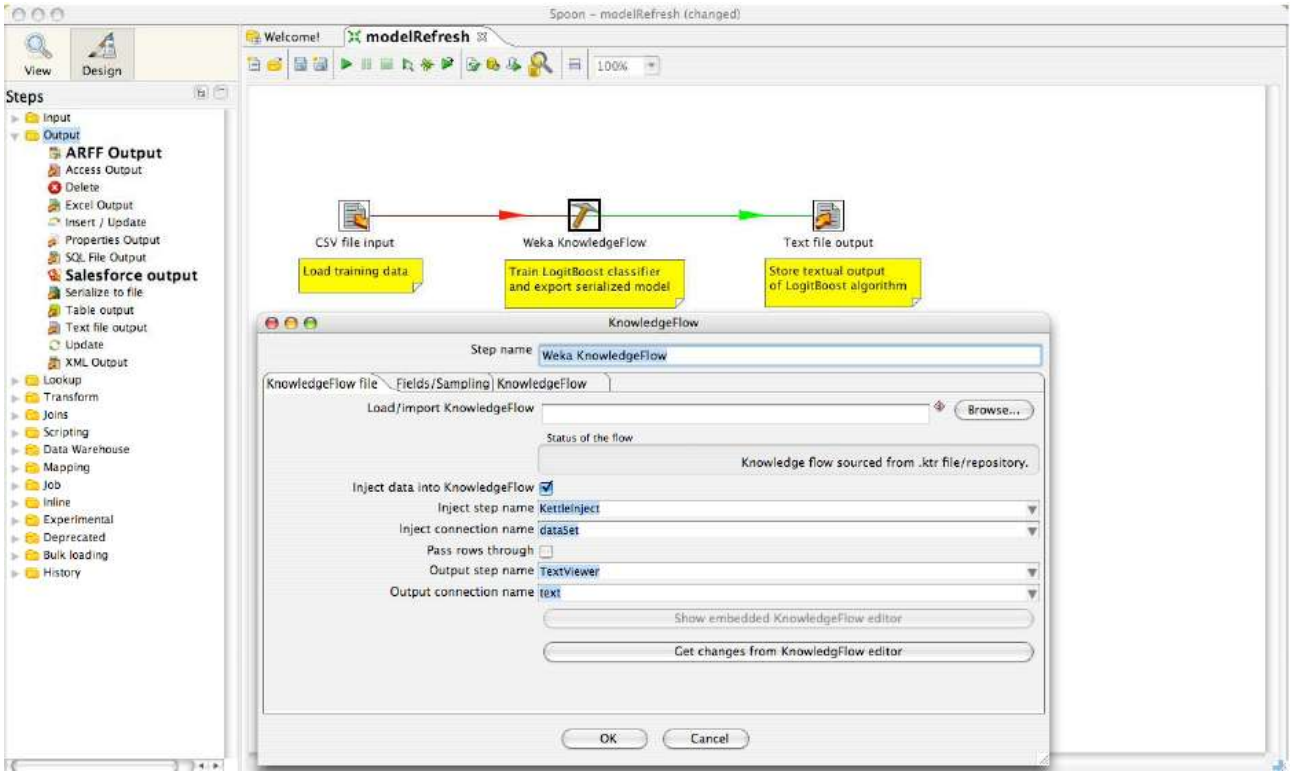

図 4 Knowledge Flow PDI コンポーネントを使用することで予想モデルをリフレッシュ

Pentaho データマイニング顧客事例…

国民健康保険(NHS)イズリントンは、イズリントンのロンドンバラで健康と公共医療を改良するのに責任をもつイギリスでの 150 のローカル NHS 組織の 1 つです。再発防止のためのタイムリーな救命と運用経費を下げるため、入院を必要する高いリスクの患者を特定する必要がありました。

Pentaho BI デプロイの前、イズリントン PCT とドクターは、あるアプリケーションを使用していましたが、ハイリスク患者を特定して予防薬投与で彼らを扱うには限界がありました。 そのソリューションは、高価で、手間がかかり、それがサポートすることができたデータソースのボリュームは制限がありました。

レガシーシステムを Pentaho BI スイートエンタープライズ版の ETL とデータマイニングに取り替えて以来、NHS イズリントンは、患者や非常時の追加の情報のデータを集めて、現在、スムーズに分析できるようになりました。 Pentaho ソリューションは、複数の国・地方のソースからデータを抜粋して、集結されたマイクロソフト SQL サーバーデータベースと統合されます。 Pentaho Data Mining と Pentaho Data Integration と共にデプロイされたロジスティック回帰機能は、どの患者がより高いリスクがあるかを特定するために、現在の投薬や前の入院履歴などの患者属性を広範囲に分析します。 この情報によって、NHS 看護師と町医者は今後の入院を避けるのを助ける適切な再発防止を取ることができました。またそれらのヘルスケアコストを大幅に削減することができました。

Pentaho 社に関して

Pentaho 社はビジネスインテリジェンスの商用オープンソースを提供する企業です。 Pentaho BI スイートエンタープライズ版は包括的なレポーティング,OLAP 分析, ダッシュボード、データ統合データマイニングそして BI プラットフォームを提供します。現在世界の主で広く最もデプロイしているオープンソース BI スイートです。 年間予約申し込みでサポート、サービス、および新機能を提供します。 Pentaho の商用オープンソースビジネスモデルはソフトウェアライセンス料を排除します。 商用オープンソース BI におけるパイオニアとして、Pentaho は小さい組織から Global2000 企業で使用されており、数年間で 300 万回以上ダウンロードされています。