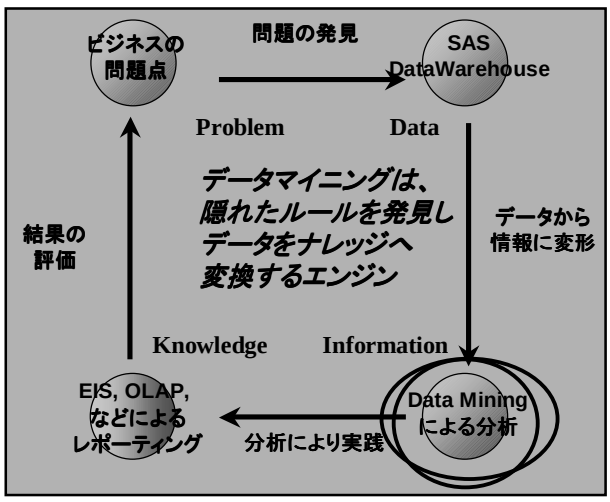
EnterpriseMinerによる データマイニング

SAS Institute Japan Ltd.
CustomerService本部
SolutionPlanningCenter

The Power to Know.

ビジネスにおける情報分析サイクル



2

従来の統計分析との違いは？

○『大規模データ』という言葉を除けば従来の統計分析が目指しているところとかわりはない

○データが大量にあることによって新たな何かをできるようになったことが重要なポイント

○推測せざるを得なかったことが実測データによる検証が可能になってきた

3

なぜマイニングなのか？

- 複雑なデータの場合はデータマイニングが役に立つ
  - 考え方の基本は「クロス分析」を行う
  - しかし、属性が数十個あったらどうなるか
    - 組み合わせの爆発
  - データマイニングの１手法として用いられるディシジョンツリーは、この組み合わせの爆発を避けながら、意味のあるセグメントを抽出する
- OLAP 技術との対比
  - OLAP ではデータを様々な角度から観察することが目的
  - 自由度が高すぎて実際にはデータのジャングルに突入

4

sas e-Intelligence

The Power to Know.

EnterpriseMinerソフトウェア

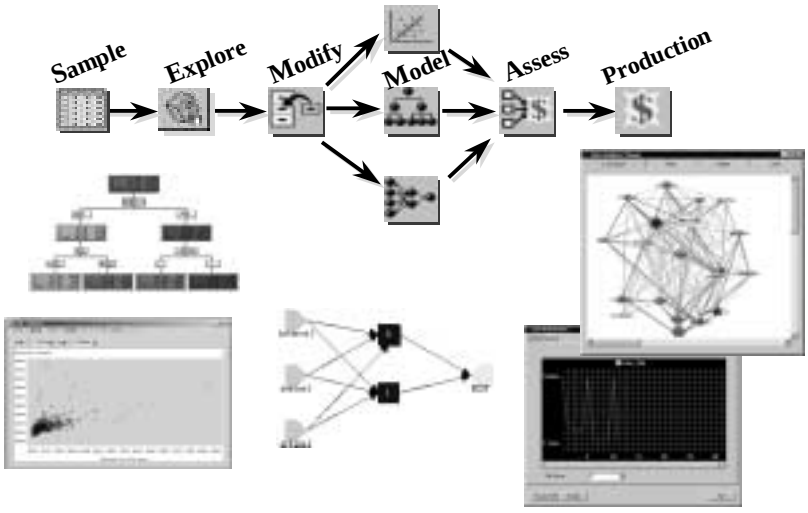
- EnterpriseMinerソフトウェアは、データマイニングの全てのプロセスと機能を統合化したソフトウェアソリューション
- データマイニングのための方法論である、SEMMAモデル(Sample,Explore,Modify,Model,Assess)に基づいてGUIによる操作が可能。
- 理解しやすい操作により分析に集中
- SASデータウェアハウスソリューションとの結合により、スムーズな意思決定を支援
- レポーターノードにより、分析フローをHTML形式などに書き出し、Webによる共有が可能

5

sas e-Intelligence

The Power to Know.

EnterpriseMinerにおける データマイニング方法論 「SEMMAモデル」



6

sas e-Intelligence

The Power to Know.

EnterpriseMinerソフトウェア

- サポート可能な機能・手法
  - 回帰分析(線形回帰分析、ロジスティック回帰分析)
  - ツリー(CHAID、C&RT、C4.5)
  - ニューラルネットワーク
  - クラスタリング(k-means法)
  - 自己組織化マップ(SOM/Kohonen)
  - アソシエーションルール、シーケンシャルルール
  - Bagging、Boostingのリサンプリング手法、Ensembleによる複合モデルの作成
- SASシステムとの統合
  - 様々なRDBMSやファイルの取り込みや、高速なデータ加工
  - クライアント/サーバ環境での動作など







7

sas e-Intelligence

The Power to Know.

EnterpriseMiner4.1の新機能

- 既存ノードへの追加機能
  - 「アセスメント」ノード
    - 予測値プロットを表示
  - 「テーブル属性」ノード
    - 分布の表示が可能
  - 「データ置き換え」ノード
    - 名義変数の置き換え方法の追加
  - 「回帰分析」ノード
    - ターゲット変数の測定水準の追加
- ノード以外の新機能
  - プロセスモニター
    - 回帰分析
    - SOM/Kohonen
    - ニューラルネットワーク
  - 「RESTORE」コマンド
    - 損傷したダイアグラムの修復

- ノードの新規追加
  - 「主成分/DMニューラル」ノード
  - 「2段階モデル」ノード
  - 「時系列」ノード(評価版)
  - 「リンク分析」ノード(評価版)
  - 「メモリー-基本推論」ノード(評価版)- MBR

8

「データ置き換え」ノード

- 名義変数の水準で、トレーニングデータに存在せず、スコアリングデータに存在する場合に、以下の置き換え方法が選択可能
  - 「most frequent value」
  - 「missing value」

トレーニングデータ 変数「地区」
大田区
千代田区
千代田区
新宿区
・

スコアデータ 変数「地区」
中央区
大田区
千代田区
新宿区
中央区
・

左記の場合、トレーニングデータでは「千代田区」が最頻値であるが、スコアデータには「中央区」が存在している。「中央区」はトレーニングデータにとっては、「未知の水準」となるため、スコアリングに反映できない。
この設定により、「未知の水準」に対して、「most Frequent value(最頻値)」や「missing value(欠損値)」などが選択できる。

9

回帰分析ノード

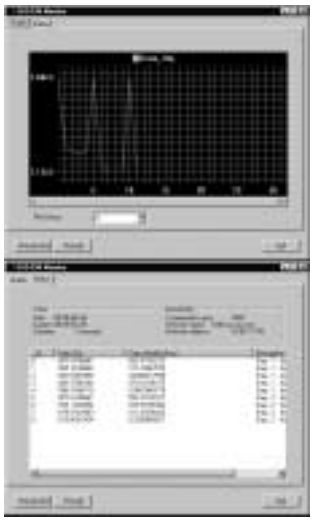
- 回帰分析でもNominal水準をターゲットとする事が可能。
- これで、ニューラル・ツリー回帰分析の全てで、あらゆる水準の変数をターゲットにする事が可能となった。

ターゲット変数:	ニューラル ネットワーク	ツリー	回帰分析
Nominal	Yes	Yes	Yes*
Ordinal	Yes	Yes	Yes
Interval	Yes	Yes	Yes

10

プロセスモニター

- 回帰分析・ニューラルネットワーク・SOM/Kohonenの各ノードで、処理状況を観測し、処理の停止などを指定できる。
- Windowsの環境変数、SASのオプションにIP_ADDRESSを指定する事により、ローカルマシンおよびクライアント／サーバーにおけるプロジェクトの学習をモニタリング可能。



11

「主成分/DMニューラル」ノード

- 変数の数が多く、入力変数間に高い相関があるような場合に有効。
- 主成分分析
  - 主成分分析のみを実行し、結果の主成分スコアを後続のノードに渡す事が可能となっている。
- DMニューラル
  - 主成分分析を行い主成分スコアを算出
  - 算出された主成分スコアをバケット変換し、二値または間隔尺度の変数を予測する非線形モデルを作成。
- 安定したモデル作りを期待できる

12

「2段階モデル」ノード

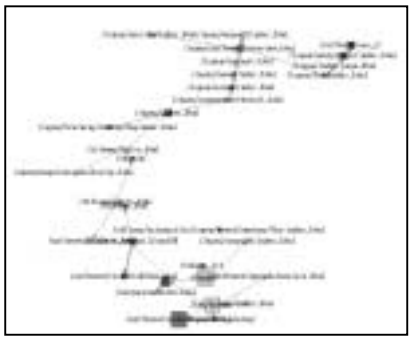
- 予測変数が分類変数と間隔変数の二つのダーゲットを「2段階モデル」で予測する。
- 例えば･･･
  - ターゲット1⇒「レスポンスフラグ」
  - ターゲット2⇒「売上高」
  - 上記二つのターゲットに対して、まず第一段階で「レスポンスがありそうな人間を探し出す」、その後に第二段階で「売上高が高くなりそうな人を探し出す」といった分析が一つのノードで可能。
- 使える手法は？
  - 第一段階
    - Nominal,Ordinal,Binaryのターゲットに対して、
    - ツリー、回帰分析、MLP、RBF、GLIM
  - 第二段階
    - Intervalのターゲットに対して、
    - ツリー、回帰分析、MLP、RBF、GLIM

13

「リンク分析」ノード（評価版）

- 複雑なデータ・システムの中から、有用な結論を導き出す行動パターンを視覚的に発見する事を支援する。
- どのような利用パターンが考えられるか？
  - 不正の発見
  - 犯罪組織におけるネットワーク
  - 通話トラフィックのパターン
  - Webサイトにおける構造と使用法
- アソシエーション分析の補完的な役割としても使える。

Webアクセスログ解析



14

Demonstration!!

sas e-Intelligence

The Power to Know™