NTT DATA
NTT DATA Mathematical Systems Inc.

より手軽に、より正確に、より大規模に

Visual Mining Studio

次期バージョン（V8.0）のご紹介

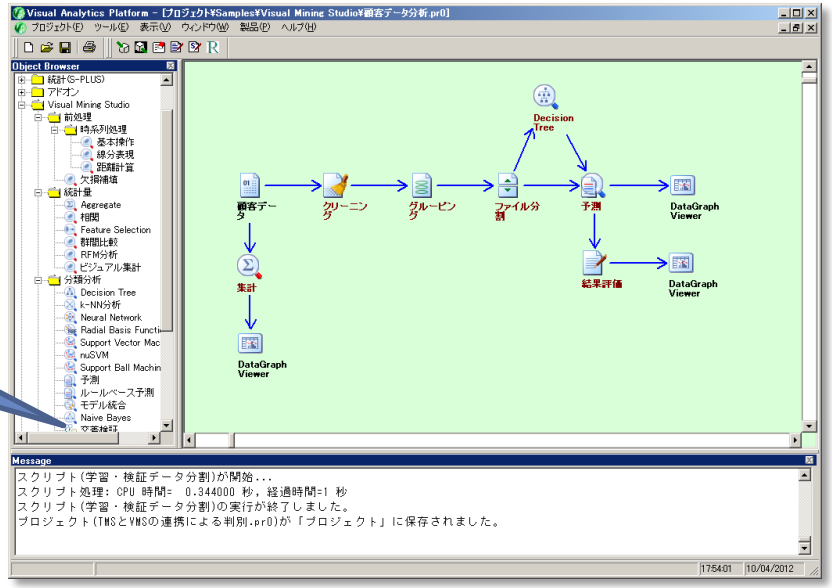
2014年2月リリース

（株）NTTデータ 数理システム
徐 良為

Visual Mining Studioとは

NTT DATA Mathematical Systems Inc.

汎用データマイニングソフト
日本発・完全自社開発



- 特徴
  - 手軽な試行錯誤環境
    - ビジュアルプログラミング
    - データ可視化
  - 豊富なマイニング機能
    - 前処理・予測・データ探索・統計解析
  - 多様な他システム連携機能
    - DBシステム・Excel・R・SPLUS・SAS・Python・MatLab

V8.0 主な新機能(分析、データ処理)

NTT DATA
NTT DATA Mathematical Systems Inc.

- 隠れ(セミ)マルコフモデル:時系列データ分析の定番
  - 学習　モデルパラメータ推定
  - 推論　Smoothing・Filtering・Prediction
  - シミュレーション　　時系列観測データの生成
  - 評価　時系列のモデルへの適合度
- データハンドリング新機能
  - ビッグデータハンドリング
  - データベースハンドリング
  - Hadoop データハンドリング(改善)
- 前処理
  - 軽量化、高速化された並列実行
  - 文字列処理の高速化(高速ハッシュ機能導入)
  - スクリプト上の正規表現
  - マルチコードデータの読込

主な新機能(ユーザビリティ)

NTT DATA
NTT DATA Mathematical Systems Inc.

【スクリプト】 新関数

- optimize　機能改善:目的関数の返り値、初期解のデフォルト値
- **is.substr.reg**　部分文字列の正規表現
- search_reg　正規表現による文字列パターン検索
- replace_reg　正規表現による文字列置換
- read_file(・,encode[・])　言語コードの設定
- **run_script ("R")**　Rの新バージョン対応
- **get_scr_input_table**　**スクリプトアイコン引数の直接引渡**
- chain_table　機能改善、重複列を可能に
- run_sql_on_db　オプション:StopInError追加
- コンスタント値
  - CONST_PI　π 値
  - CONST_INT_MIN　整数最小値
  - CONST_INT_MAX　整数最大値
  - CONST_REAL_MIN　実数最小値
  - CONST_REAL_MAX　実数最大値

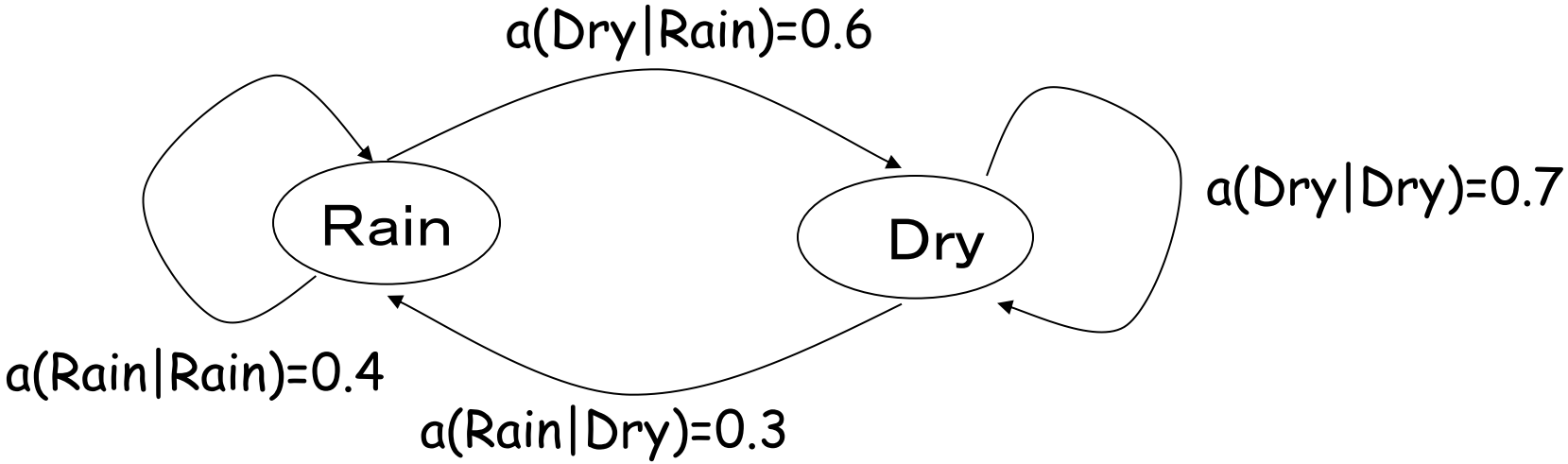
【その他】 VAP実行環境改善(多数)

主要新機能の概要紹介

隠れ(セミ)マルコフモデル(HSMM)機能

NTT DATA
NTT DATA Mathematical Systems Inc.

- HSMM＝Hidden Semi Markov Model
- Markov Chain: 順序付き状態変化(遷移)
- 状態= {Rain, Dry} 、a: 遷移確率、初期状態分布

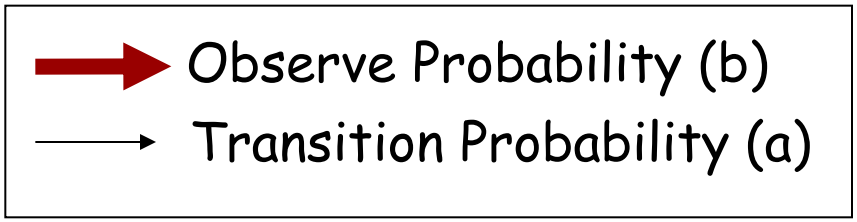


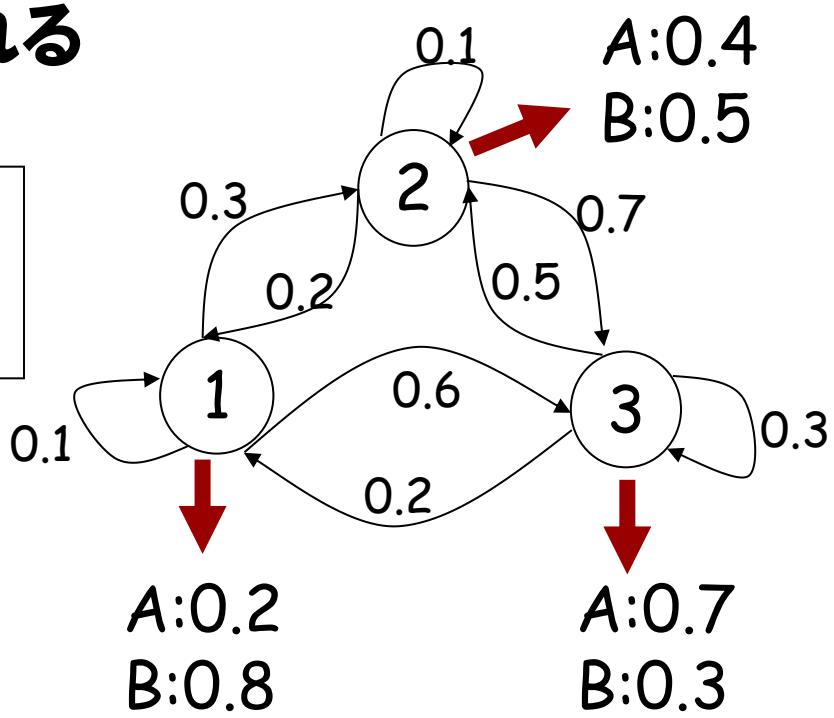
Markov Property(マルコフ性)
現状態は、一つ前の状態にのみ依存

HSMM(Hidden Semi Markov Model)

NTT DATA Mathematical Systems Inc.

1. S= {1,2,3} : Un-Observable States
2. V= {A,B} : Observation Symbols
3. 開始状態、終了状態の確率分布
4. 滞在時間分布を与えられる





HSMMの適用例(１)

NTT DATA
NTT DATA Mathematical Systems Inc.

- 音声認識、音声合成
- テキストマイニング：文章の形態素解析
- 人間行動分析(Human Activity Recognition)
- 手書き文字認識
- ネットワークトラヒック・異常検出
- 医療データ(脳電図、心電図)分析
- DNAに含まれる遺伝情報分析
- 言語認識
- 地上ターゲットトラッキング
- 製造データからの異常検出
- イメージデータ圧縮・分類

HSMMの適用例(２)

NTT DATA
NTT DATA Mathematical Systems Inc.

- 時系列(数値)データの異常検出
- タンパク質の構造予測
- 天気データ分析
- 植物のパターン分岐と開花
- 衛星伝搬チャンネルモデリング
- ビデオ録画からのイベント認識
- モバイルネットワークの移動追跡
- 画像分類
- 金融時系列のモデル
- 音楽分類
- リモートセンサリング
- 空気中の特異物質の検出

Visual Mining Studio 上のHSMMの特徴

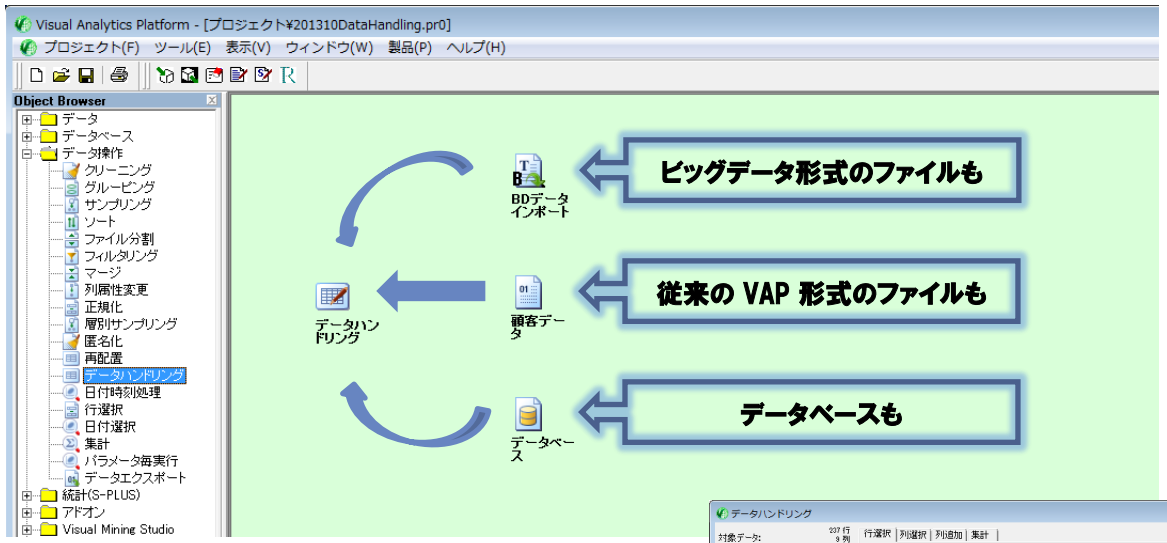
NTT DATA
NTT DATA Mathematical Systems Inc.

- **学習**
  初期モデル＋学習時系列から、パラメータ推定
- **推論＝時系列観測値に対して、**
  Smoothing　過去欠損補填
  Filtering　現在推定
  Prediction　将来予測
- **シミュレーション**
  モデルに沿った観測列の自動生成
- **評価**
  時系列観測値のモデルへの適合度
- **対応観測値**
  - 観測値＝数値（平均回帰）
  - 観測値＝数値（線形回帰）
  - 観測値＝カテゴリ（カテゴリ分布）
- **観測確率分布の外部指定**

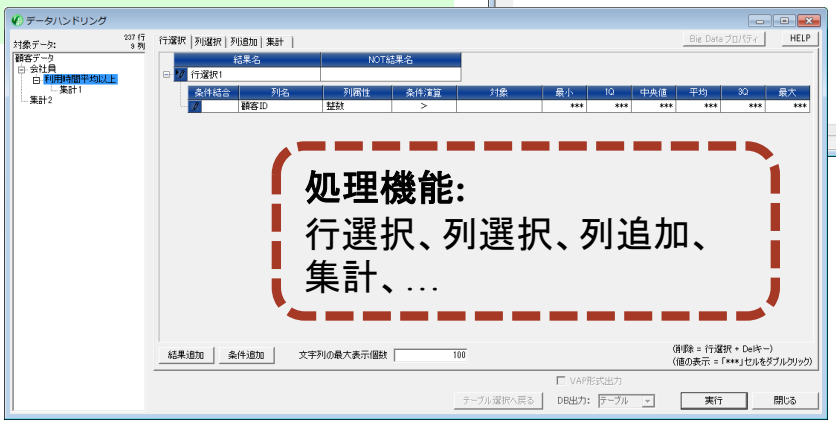
データハンドリング機能拡張

NTT DATA Mathematical Systems Inc.

- ビッグデータからデータベースまで
  同じインターフェースで利用することが可能



同じアイコン、
同じインターフェースで
操作可能！



処理機能:
行選択、列選択、列追加、
集計、…

**あらゆるデータの前処理を
データハンドリングで！**

後継バージョン展望

NTT DATA
NTT DATA Mathematical Systems Inc.

- 開発目標

  より高速、より手軽に、より正確に

- モデル作成の高度な自動化
  - パラメータチューニングの精度向上
  - 高度な並列化
  - 複数マシン間の分散処理
- 新しいデータ分析・機械学習機能の追加
- 決定木のユーザインターフェース改良

引き続き、皆様のデータ分析のお役に立てるよう、
開発者一同頑張っています。
ご期待ください！