データマイニングはここまで進化した！
Visual Mining Studio 次期バージョンの
ご紹介

(株)数理システム
徐 良為

主な新機能

- VMX (Visual Mining eXpress)
  - 最新マイニング技術をシナリオ毎に、パッケージ化
  - ブラウザ上で、メニューをマウスクリックするだけで利用可能
  - 初心者にとって、難しい分析アルゴリズム、分析フローを気にすることなく、優れたモデル、分析結果が得られる
- 二項ソフトクラスタリング
  - 「需要」、「供給」間のマッチングの最新手法
  - マッチング： 顧客と商品、Webページと視聴者、人材と仕事、、、
  - 大変有効であることを実データを用いて実証済み
  - レコメンデーションシステムなどに
- 自動欠損補填機能 (スクリプト関数)
  - 正常データの分布に従い、データの欠損部分を自動的に補填する
  - データの欠損処理、データフュージョンに利用される

主な新機能

- 決定木モデルからの説明変数重要度
  - 目的変数に対する説明変数の寄与度
  - 変数選択に使われる
  - 決定木のBoostingやBaggingで説明変数重要度も出力
- R連携
  - 汎用統計ソフトR スクリプトアイコン作成
  - Rスクリプトアイコンの並列実行
  - 外部Rスクリプトの呼び出し
- GUI関連
  - 処理フローのループ化機能：テーブル行毎にフローを実行
  - データインポート機能の大幅増強
  - 処理フロー定義のエクスポート：スクリプトからの呼び出し

主な新機能 (ユーザビリティ)

【スクリプト】 dyadic_soft_clustering 二項ソフトクラスタリング関連
【スクリプト】 cluster_dyadic_data 二項ソフトクラスタリング関連
【スクリプト】 predict_dyadic_data 二項ソフトクラスタリング関連
【スクリプト】 read_file 関連オプション
- unfixed_rec 行に含まれる項目数が異なるデータへの対応
- missing_file_path 欠損箇所指示データの保存場所
- read_title_info 元データの第1行目の内容 (列名) 及び、列属性 (試した結果)) を2列のテーブルとして返す
- force_reading どんなデータでも読めるように「強制読み込み」。読むときに発生したエラーを外部ワーニングファイルに出力
- header_and_detail_row 列名が複数行に渡って存在するケースへの対応

【スクリプト】 run_script のRスクリプトの呼出し
【スクリプト】 ranking キー毎のランク付け
【スクリプト】 filling_missing_data 欠損値自動補填
【スクリプト】 return プロシージャーの外側 (トップレベル) でのリターン

主な新機能 (ユーザビリティ)

【スクリプト】 cells 複数セル内容を取得
【スクリプト】 aggregate 集計時、サマリに対してのマトリックス出力形式
【GUI】 集計アイコン、サマリに対してのマトリックス出力形式
【GUI】 行毎実行機能
【GUI】 Excelの xlsx への対応機能
【GUI】 VAP連携機能
【GUI】 大規模対応データ表示、グラフ描画機能
【GUI】 R-スクリプト定義アイコン

主要新機能のイメージ

VMX - Visual Mining eXpress

- 長年蓄積されたマイニングの最新技術をWebブラウザから利用可能
- 煩雑なパラメータ設定、理解しにくいマイニング概念を一切省いて、初心者にも、データマイニング可能！

VMX - Visual Mining eXpress

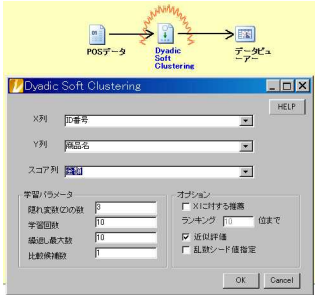
- 弊社が用意したマイニングのシナリオをVMS上で表現
- 弊社開発のVAP (Visual Analysis Platform) 技術を用いて、Web公開
  - 欠損データの自動補填
  - モデル作成時の自動パラメータ選択
  - 最適なモデル作成戦略
  - ユニークなクラスタリング分析
  - 実戦的なアソシエーション分析
  - など

ブラウザ上で、煩雑なパラメータ設定なしで、クリックのみで実現可能

二項ソフトクラスタリング

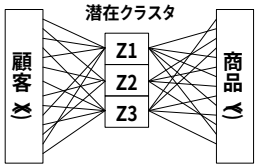
二項目間の自動マッチング
- 顧客 vs 商品
- 視聴者 vs Webページ
- 人材 vs 仕事
  などなど

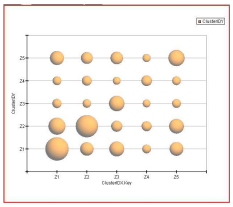
レコメンデーションシステム



顧客	商品名	数量	金額
1	乳製品	1	129
1	パン	1	118
2	調理品	1	171
3	納豆	1	110
3	菓子	1	128
3	菓子	1	128
3	菓子	1	128
3	パン	3	118
4	キャチン	1	189

二項ソフトクラスタリング





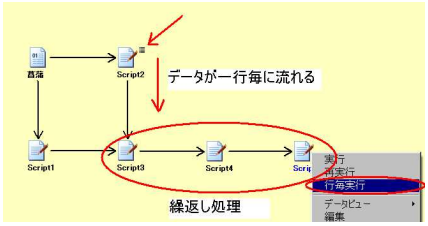
結果を表すテーブル

1. 顧客 (X) とクラスタ (Z) の関連性を表す確率テーブル：$P(X|Z)$、$P(Z|X)$
   → 顧客 (X)がクラスタ Z に属す確率
2. 商品 (Y) とクラスタ (Z) の関連性を表す確率テーブル：$P(Y|Z)$、$P(Z|Y)$
   → 商品 (Y)がクラスタ Z に属す確率
3. 顧客 (X) にお勧め商品 (Y) のトップN

R連携

- RスクリプトをVMSの処理アイコンとして利用
- 並列実行実現
  sys_proc_parallel
- 外部Rソースの呼出し
  run_script("c:/foo.R","R",table,?reuslt);
- Rのグラフ表示

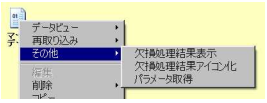
処理フローのループ化機能



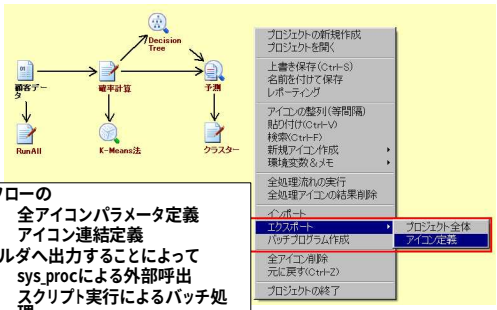
- 「ループ処理」より単純化
- 行毎実行の終端アイコンは任意
- 全ての結果は終端アイコンに収集
- 大量定型処理に便利 (行毎にはファイルパスを含むなど)

新データインポート機能

- 様々なテキスト形式ファイルを柔軟に対応可能
- 欠損（異常）値の自動補填
- データサンプリング
- 各種エンコード形式対応
- 形式エラーになるセルを通知
- 再取り込み、欠損指示ファイル
- パラメータ取得によるスクリプト化可能



処理フロー定義のエクスポート



処理フローの

1. 全アイコンパラメータ定義
2. アイコン連結定義

をフォルダへ出力することによって

A) sys_procによる外部呼出
B) スクリプト実行によるバッチ処理

が可能

来年の後継バージョン

- よりユーザが使い易いもの
- より実戦的なもの
- データストリームマイニング機能

を目指す