より手軽に、より正確に、より大規模に

マイニング自動化への大きな一歩

Visual Mining Studio

次期バージョン（V7.1）のご紹介

2012年1月21日リリース

(株)数理システム
徐 良為

主な新機能

- **モデル最適化：パラメータチューニング、変数選択**
  - 最適なモデル構築に必要なパラメータ調整、及び説明変数の選択を自動的に行える
  - 試行錯誤の負担を大幅に軽減
  - 目的関数をユーザが自由に定義可能
  - メタヒューリステック最適化アルゴリズム採用
- **VMX新機能**
- **ビジュアルプログラミング機能**
  - 行毎実行
  - パラメータ毎実行
  - より手軽に実現
- **自動欠損補填機能（GUIインターフェース）**
  - 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 (ユーザビリティ)

【スクリプト】 optimize関数、最適化関数
【全体】 全処理アイコンの立ち上げ速度改善 (大規模データ)
【行選択】 大規模データでも手軽に処理可能にした
【スクリプト】 下記の関数に、デフォルト値指定可能にした

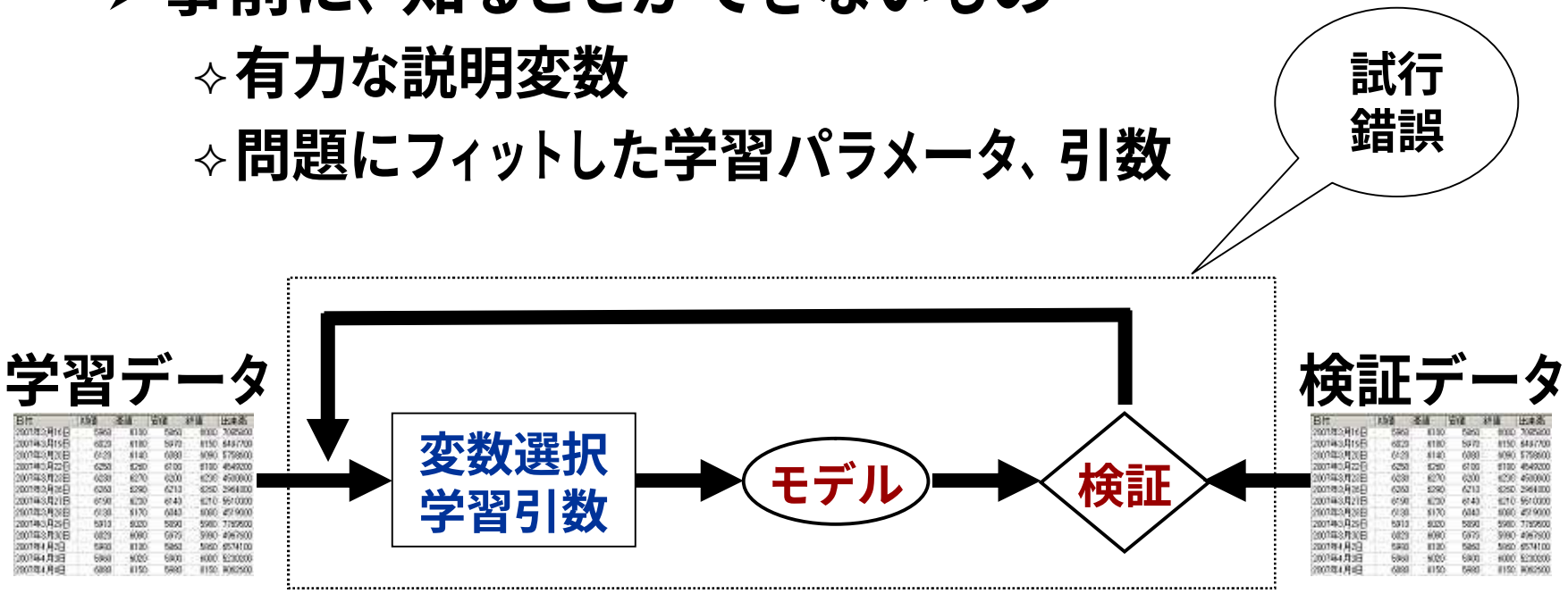
- get_win_env
- get_usr_env
- get_prj_env

【スクリプト】 新規関数 get_scr_env
外部から、sys_proc_opt によって、スクリプトへ値を引き渡す機能を実現するため、スクリプト側に取り入れる機能
【Loop】 処理フロー上の繰り返しループ処理機能
【全体】 処理実行をしながら、他のタスクを干渉しないようにした
【DB連携】 DBに含まれる特殊な区切り文字に対応可能とした
【スクリプト】read_file で固定長のテキストファイル内容を読み込み機能
【スクリプト】長文スクリプト定義 (数万行) の処理速度が遅くなる問題を解消
【その他】 GUI機能改善、不具合フィックスなど多数

主要新機能のイメージ

モデル最適化支援関数

- モデル構築時の難点
  - 手元にある観測サンプルのみで母集団を推定
  - 通常、サンプルデータは誤差を含み、雑多である
  - 事前に、知ることができないもの
    - 有力な説明変数
    - 問題にフィットした学習パラメータ、引数



モデル最適化支援関数

- モデル構築作業の定式化

$$\max_{v \subseteq V, p \in S} f(v, p, train, test)$$

f：モデル構築、検証、モデルの評価値を算出
V: 説明変数全候補
p: モデル作成時に使用する引数
S: 引数候補集合
train: 学習用データ
test: 検証用データ

モデル最適化支援関数

- VMStudioの新関数

optimize(f, par, table, …)

$$= \max_{opt} f(par, table, ...)$$

f：モデル構築関数
(ユーザ定義可能な任意のスクリプト関数)

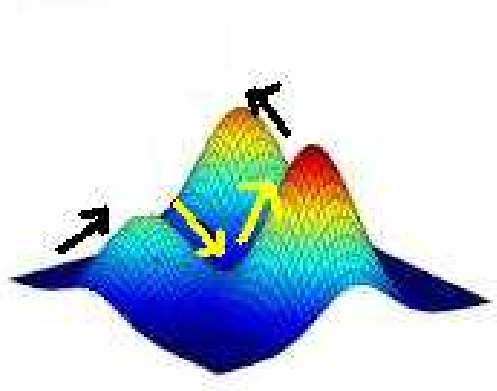
par: 変数候補、モデル作成引数候補など

table: データテーブル(複数可)

optimize: f が最大となるような opt セット (トップN)

モデル最適化支援関数

- optimize の実装について
  - f を単純な数式で表すことができない、微分不可
  - 厳密な最適解を求めるには、NP困難である
    多項式の計算時間内で解くことの出来ない問題
  - 近似解を求めるしかない
  - 一種のメタヒューリステックの山登り手法を採用



「optimize」の適用例

- 最適制御モデル（製造、風力発電、電力消費等）

$$y = f(x, z)$$

y：出力（例：発電効率）

x：制御可能要因

S：xの選択可能な範囲（例：向き、大きさ）

z：制御不可要因（風力、気温）

「optimize」の適用例

- 最適制御

  一般に、f が理論上計算不可能(理想状態除く)
  - データない場合、実験計画法 (Experimental Design)
  - データが集められた場合、マイニングのモデル作成
- 最適化

$$\max_{x \in S} y = f(x, z)$$

微分不可、式表現なし、近似解法

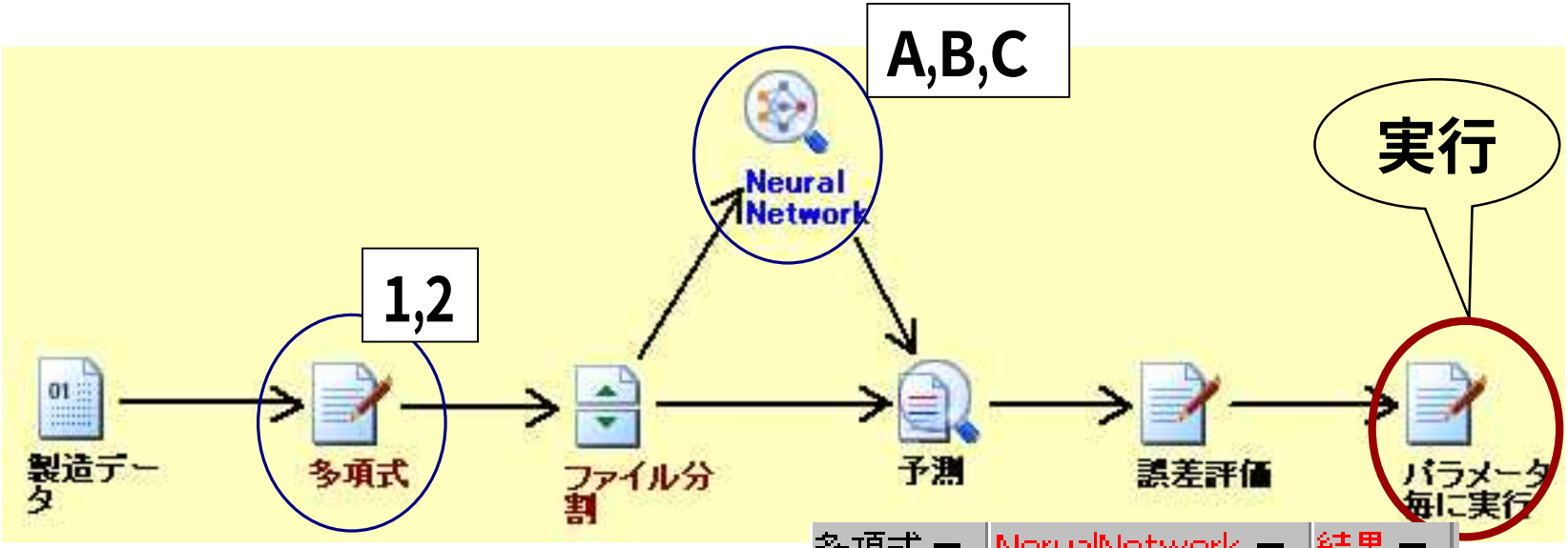
VMX - Visual Mining eXpress新機能

1. モデルの自動チューニング機能を導入！
   (VMStudio の optimize 機能参照)
2. ユーザビリティの大幅向上 (VAP機能)

- 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！

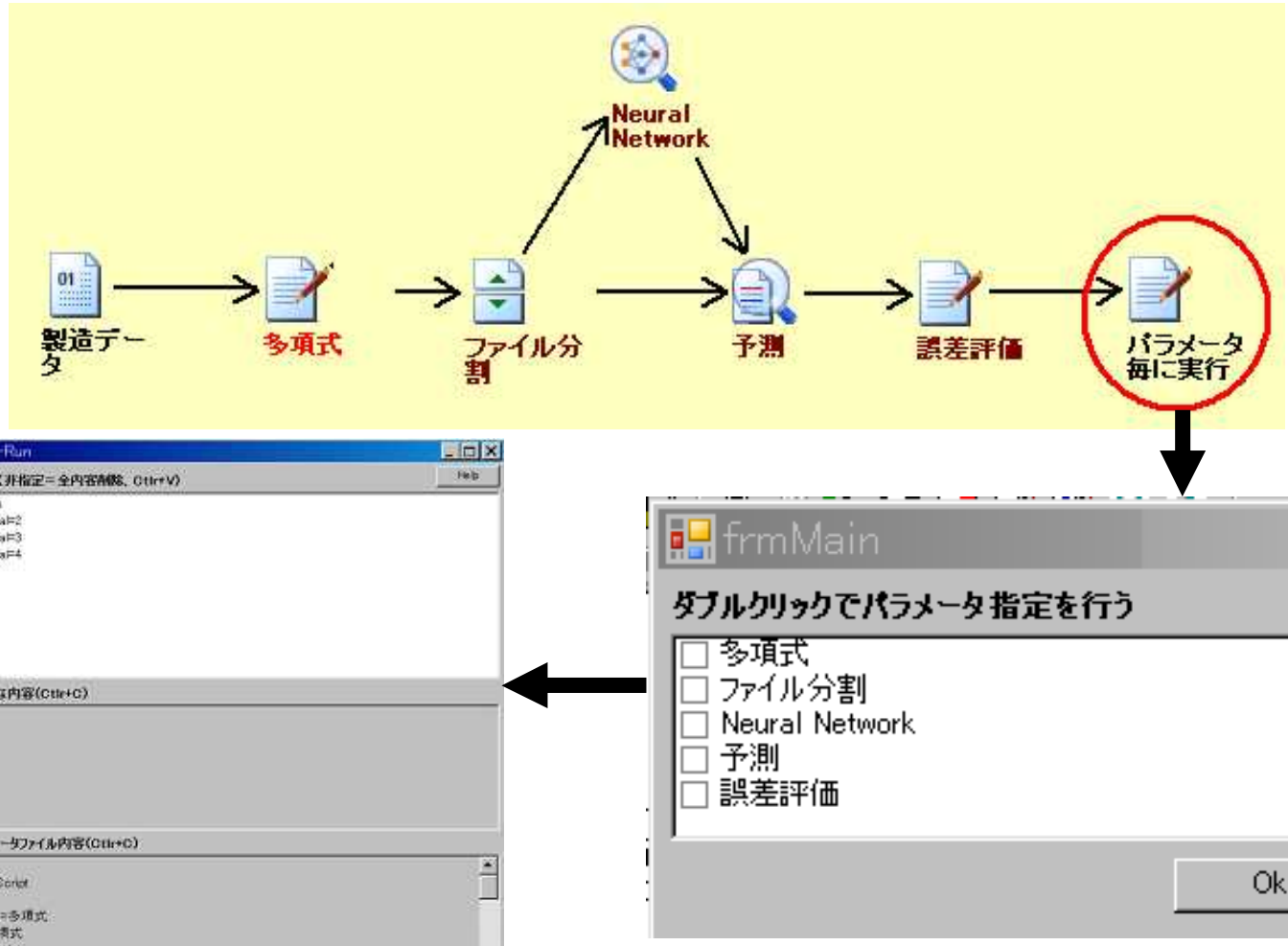
ビジュアルプログラミング

- 処理アイコンに複数選択肢を手軽に試す

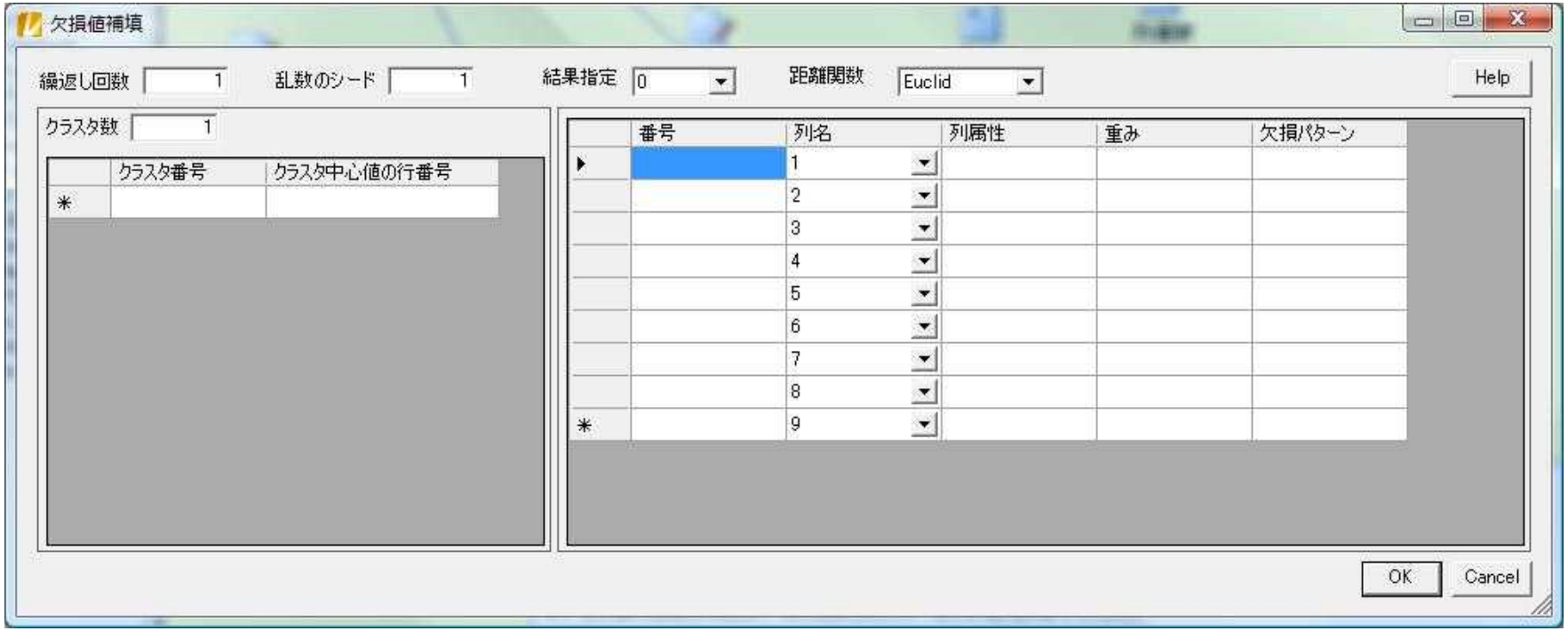


多項式	NerualNetwork	結果
1	A	…
1	B	…
1	C	…
2	A	…
2	B	…
2	C	…

ビジュアルプログラミング



自動欠損補填GUIツール



- 正常データの分布に従い、欠損部分の自動補填
- 欠損処理、データフュージョンに利用
- EMアルゴリズムを採用

後継バージョン

- **より大規模に**
  - GB：ギガバイト
    手軽なマイニング処理
  - TB(PB):テラバイト（ペタバイト）
    Hadoop（MapReduce）との連携機能
  - 無制限
    データストリームマイニング
- **より手軽に**

  VAP機能のGUI機能増強
- **より正確に**

  新しいマイニング機能導入