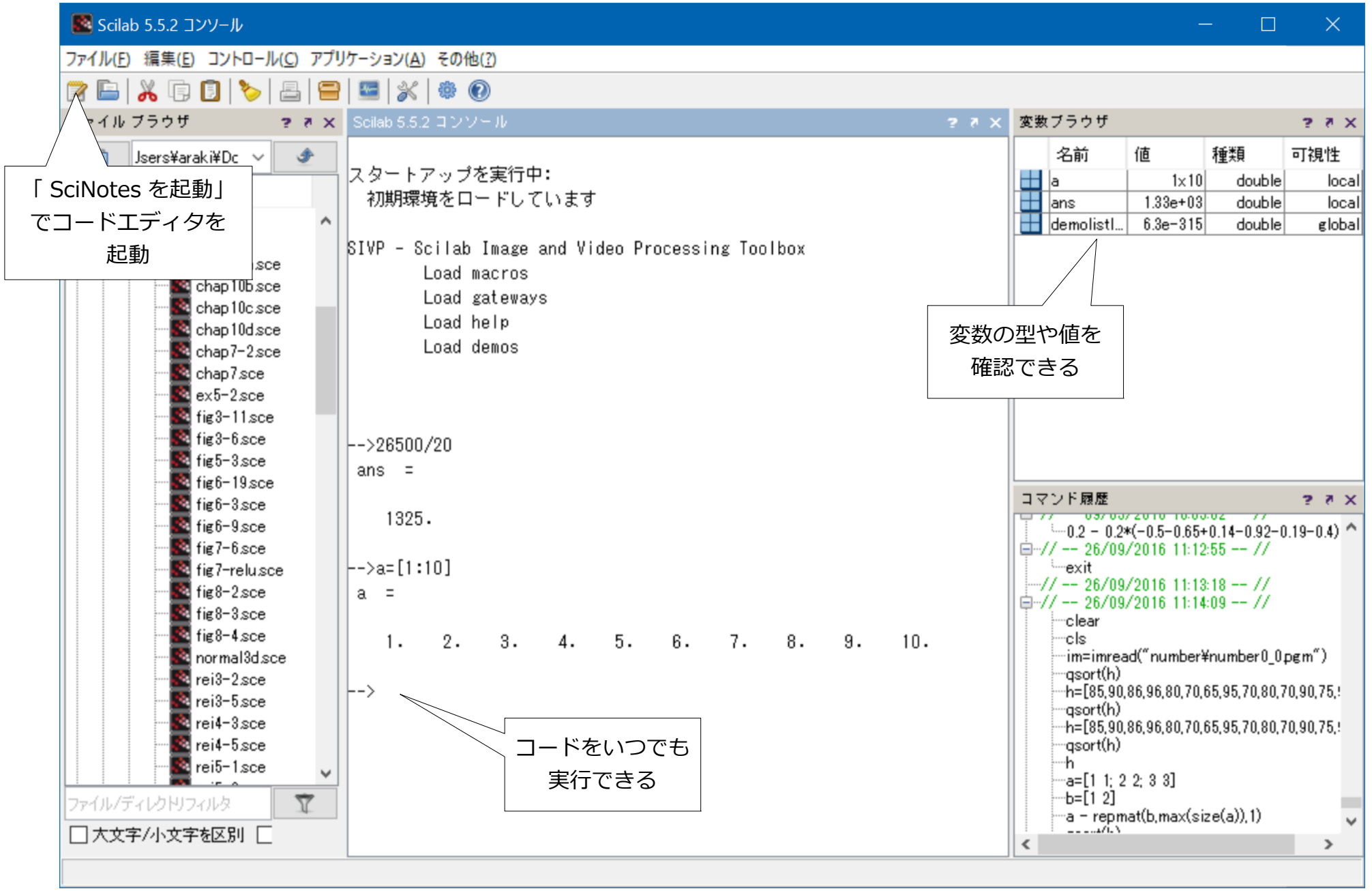
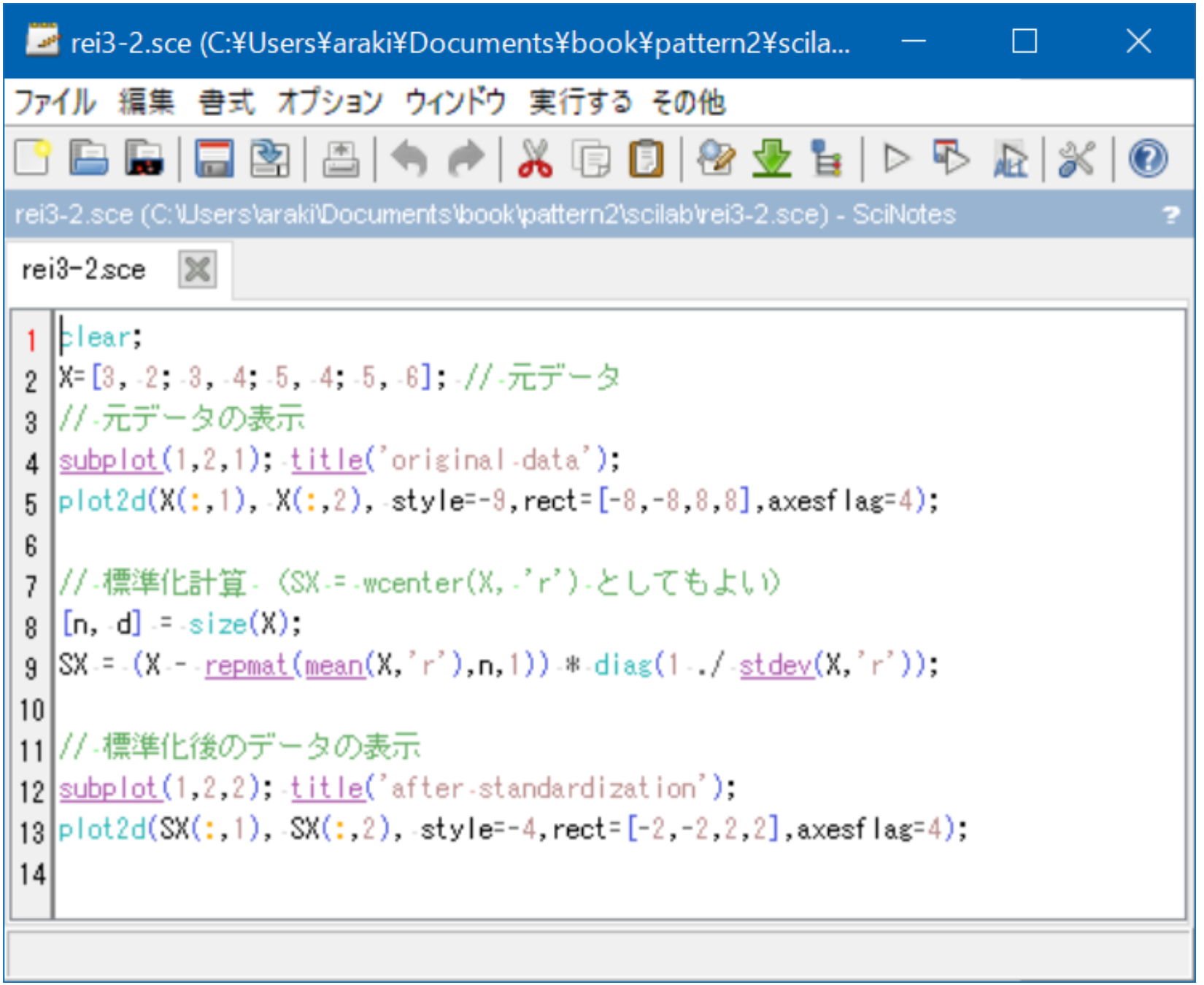
Section 5

- 数値計算用言語Scilab
- 機械学習ツールWeka

Scilab 入門

- Scilab とは
  - 数値計算を伴う問題の解決手順を記述するのに適したプログラミング言語
  - ベクトルや行列を変数の値とすることができ，それらの間の演算が可能
  - データをグラフにして表示する可視化が簡単に行える


Scilab 5.5.2 コンソール
ファイル(F) 編集(E) コントロール(C) アプリケーション(A) その他(?)
「 SciNotes を起動」でコードエディタを起動
chap10b.sce
chap10c.sce
chap10d.sce
chap7-2.sce
chap7.sce
ex5-2.sce
fig3-11.sce
fig3-6.sce
fig5-3.sce
fig6-19.sce
fig6-3.sce
fig6-9.sce
fig7-6.sce
fig7-relu.sce
fig8-2.sce
fig8-3.sce
fig8-4.sce
normal3d.sce
rei3-2.sce
rei3-5.sce
rei4-3.sce
rei4-5.sce
rei5-1.sce
ファイル/ディレクトリフィルタ
大文字/小文字を区別
Scilab 5.5.2 コンソール
スタートアップを実行中:
初期環境をロードしています
SIVP - Scilab Image and Video Processing Toolbox
Load macros
Load gateways
Load help
Load demos
-->26500/20
ans =
1325.
-->a=[1:10]
a =
1. 2. 3. 4. 5. 6. 7. 8. 9. 10.
-->
コードをいつでも実行できる
変数ブラウザ
名前 値 種類 可視性
a 1x10 double local
ans 1.33e+03 double local
demolistl... 6.3e-315 double global
変数の型や値を確認できる
コマンド履歴
0.2 - 0.2*(-0.5-0.65+0.14-0.92-0.19-0.4)
// -- 26/09/2016 11:12:55 -- //
exit
// -- 26/09/2016 11:13:18 -- //
// -- 26/09/2016 11:14:09 -- //
clear
cls
im=imread("number¥number0_0.pgm")
qsort(h)
h=[85,90,86,96,80,70,65,95,70,80,70,90,75,
qsort(h)
h=[85,90,86,96,80,70,65,95,70,80,70,90,75,
qsort(h)
h
a=[1 1; 2 2; 3 3]
b=[1 2]
a - repmat(b,max(size(a)),1)


rei3-2.sce (C:¥Users¥araki¥Documents¥book¥pattern2¥scila...
ファイル 編集 書式 オプション ウィンドウ 実行する その他
rei3-2.sce (C:\Users\araki\Documents\book\pattern2\scilab\rei3-2.sce) - SciNotes
rei3-2.sce
clear;
X=[3, 2; 3, 4; 5, 4; 5, 6]; //元データ
//元データの表示
subplot(1,2,1); title('original data');
plot2d(X(:,1), X(:,2), style=-9,rect=[-8,-8,8,8],axesflag=4);
//標準化計算 (SX = wcenter(X,'r')としてもよい)
[n, d] = size(X);
SX = (X - repmat(mean(X,'r'),n,1)) * diag(1 ./ stdev(X,'r'));
//標準化後のデータの表示
subplot(1,2,2); title('after standardization');
plot2d(SX(:,1), SX(:,2), style=-4,rect=[-2,-2,2,2],axesflag=4);

Weka 入門

- Weka とは
  - Waikato Environment for Knowledge Analysis
  - 機械学習のアルゴリズムを実装した Java ライブラリ
  - データファイルを直接操作できる GUI を持つ
  - ライセンスは GNU GPL
    - プログラムの実行・改変・再配布が自由
    - ただし二次的著作物に対しても GNU GPL が適用される
  - この解説では開発版である ver. 3.9.1 を使用

Weka に関する資料

- 開発者による機械学習一般の解説書
  - Ian H. Witten et.al.: Data Mining: Practical Machine Learning Tools and Techniques, Fourth Edition, Morgan Kaufmann, 2016.

- web 教材
  - Waikato 大学 Mooc: Data Mining with Weka
    - http://www.cs.waikato.ac.nz/ml/weka/mooc/dataminingwithweka/
    - ビデオやスライドを公開

Weka 付属の学習用データ

表 2.2　Weka 付属のデータ

データ名	内容	特徴	正解情報
breast-canser	乳癌の再発	ラベル	クラス (2 値)
contact-lenses	コンタクトレンズの推薦	ラベル	クラス (3 値)
cpu	CPU の性能評価	数値	数値
credit-g	融資の審査	混合	クラス (2 値)
diabetes	糖尿病の検査	数値	クラス (2 値)
iris	アヤメの分類	数値	クラス (3 値)
Reuters-Corn	記事分類	テキスト	クラス (2 値)
supermarket	スーパーの購買記録	ラベル	なし
weather.nominal	ゴルフをする条件	ラベル	クラス (2 値)
weather.numeric	ゴルフをする条件	混合	クラス (2 値)

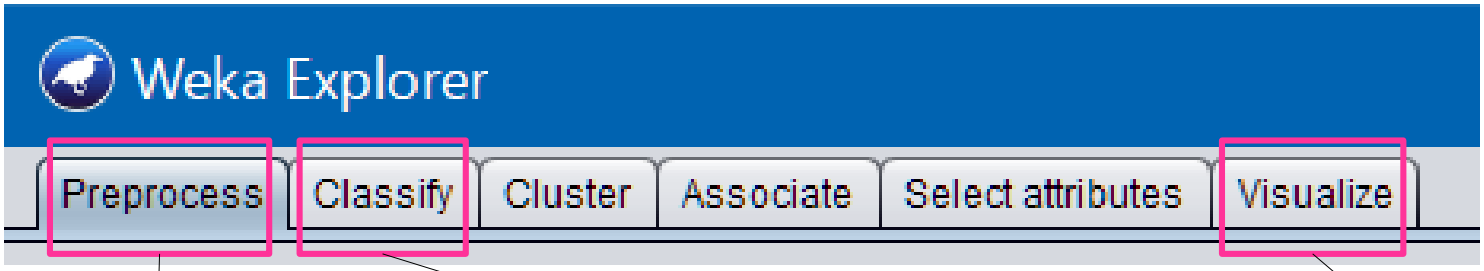
起動

- アプリケーションの選択

- **Explorer** アプリケーション
  データの読み込みから、特徴選択・学習・評価を試行錯誤的に行うのに適した操作を提供

- Experimenter : ハイパーパラメータ等を変えて性能を比較実験
- **KnowledgeFlow** : 実験プロセスを GUI で組み立て
- Workbench : すべてのアプリケーションをまとめた GUI
- SimpleCLI : コマンドラインインタフェース

Explorer での操作

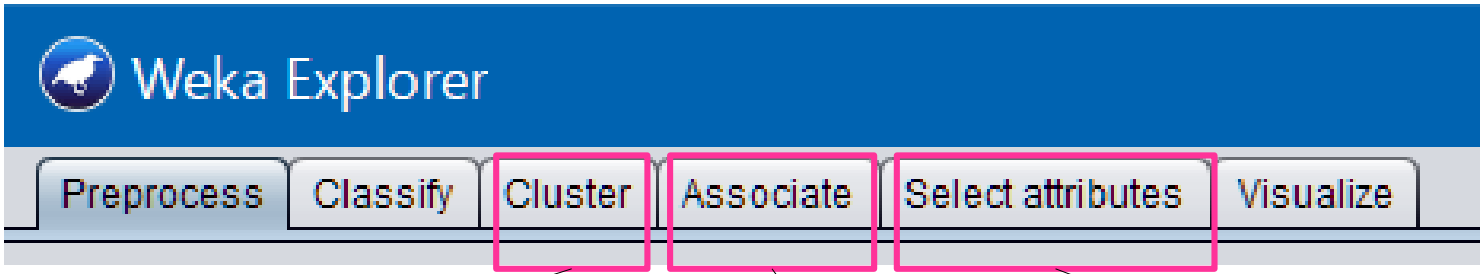


- 前処理
  - データの読み込み
  - 標準化
  - 特徴選択
  - 特徴の分析

- 識別
  - 100 以上の識別アルゴリズムの実装
  - 学習の設定
  - ハイパーパラメータの設定
  - 学習結果の評価

- 可視化
  - データの 2 次元プロット

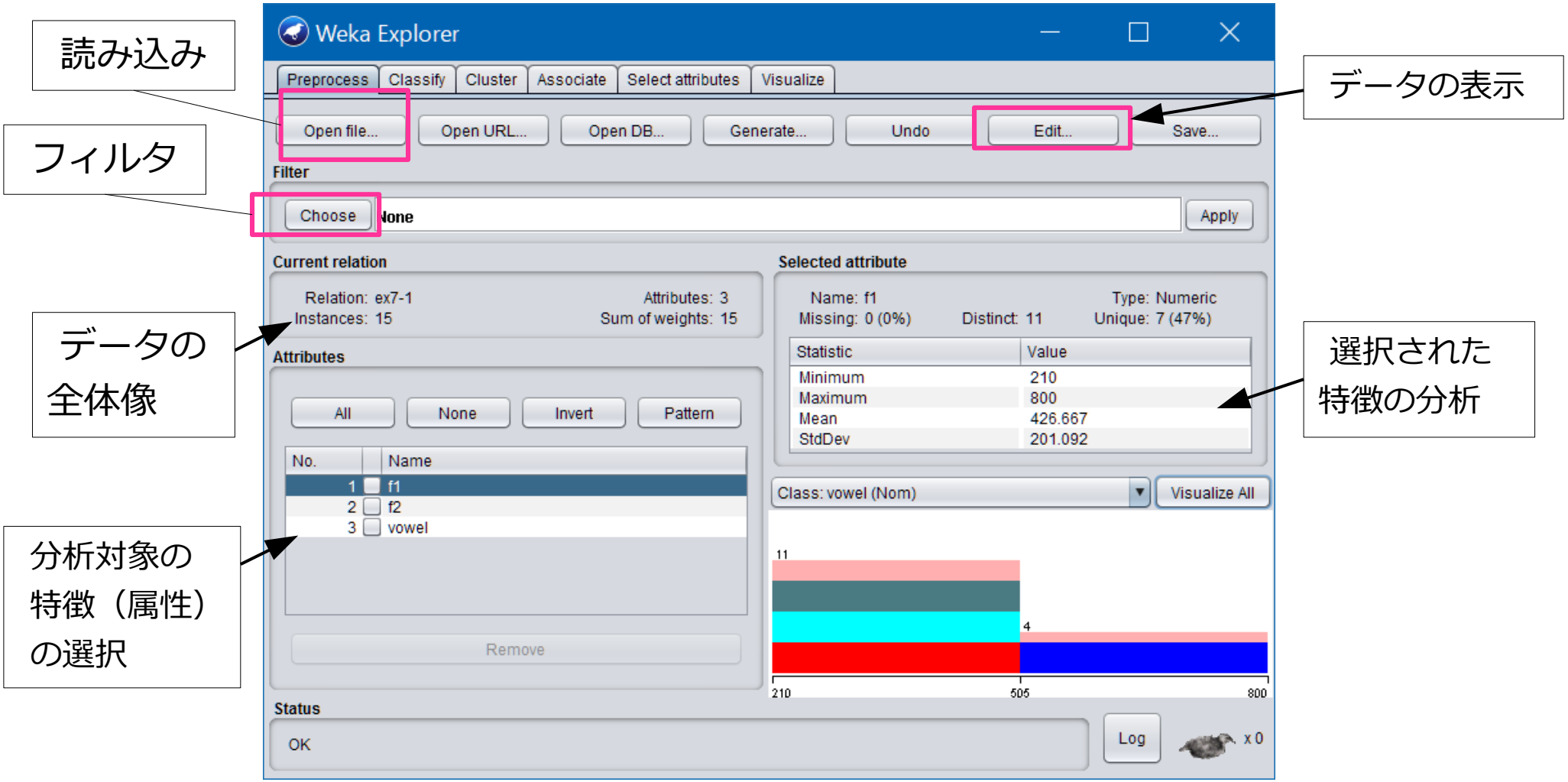
Explorer での操作



- 教師なしクラスタリング
- 規則学習
- 特徴選択

前処理（Preprocess）

- 特徴抽出後のデータを読み込む
  - いくつかの特徴の操作（フィルタの適用）が可能



前処理（Preprocess）

- 読み込み可能なデータ形式
  - ARFF (Attribute Relationship File Format) 形式
  - ヘッダ部とデータ部で構成
    - ヘッダ部
      - @relation : データ集合の名前（ファイル名と同じでよい）
      - @attribute : 特徴の各次元の名前とデータの型を宣言
    - データ部
      - @data 以降に 1 行 1 件のデータを記述
      - 各特徴・クラスラベルはカンマ区切り

前処理（Preprocess）

- ARFF ファイルの例

```
@relation ex7-1

@attribute f1 real
@attribute f2 real
@attribute class {a, i, u, e, o}

@data
700,1100,a
240,1900,i
240,1100,u
440,1700,e
400,750,o
```

連続値データは real

Nominal データは取り得る値のリストを中括弧で囲む

前処理（Preprocess）

- アヤメの分類データ (iris)

```
% 1. Title: Iris Plants Database
@RELATION iris

@ATTRIBUTE sepallength   REAL
@ATTRIBUTE sepalwidth    REAL
@ATTRIBUTE petallength   REAL
@ATTRIBUTE petalwidth    REAL
@ATTRIBUTE class {Iris-setosa,Iris-versicolor,Iris-virginica}

@DATA
5.1, 3.5, 1.4, 0.2, Iris-setosa
4.9, 3.0, 1.4, 0.2, Iris-setosa
…
7.0, 3.2, 4.7, 1.4, Iris-versicolor
6.4, 3.2, 4.5, 1.5, Iris-versicolor
…
6.3, 3.3, 6.0, 2.5, Iris-virginica
5.8, 2.7, 5.1, 1.9, Iris-virginica
…
```

データセット名

特徴名と型

萼・花びらの長さ・幅

アヤメの種類

これ以降、 1 行に 1 事例
(Excel の CSV 形式と同じ)

前処理（Preprocess）

- CSV ファイルの場合
  - １行目は特徴名とする
  - クラスラベルが数字で表現されている場合は Numeric2Nominal フィルタを適用して、 Nominal データに変換

	A	B	C
1	f1	f2	class
2	700	1100	a
3	240	1900	i
4	240	1100	u
5	440	1700	e
6	400	750	o

前処理（Preprocess）

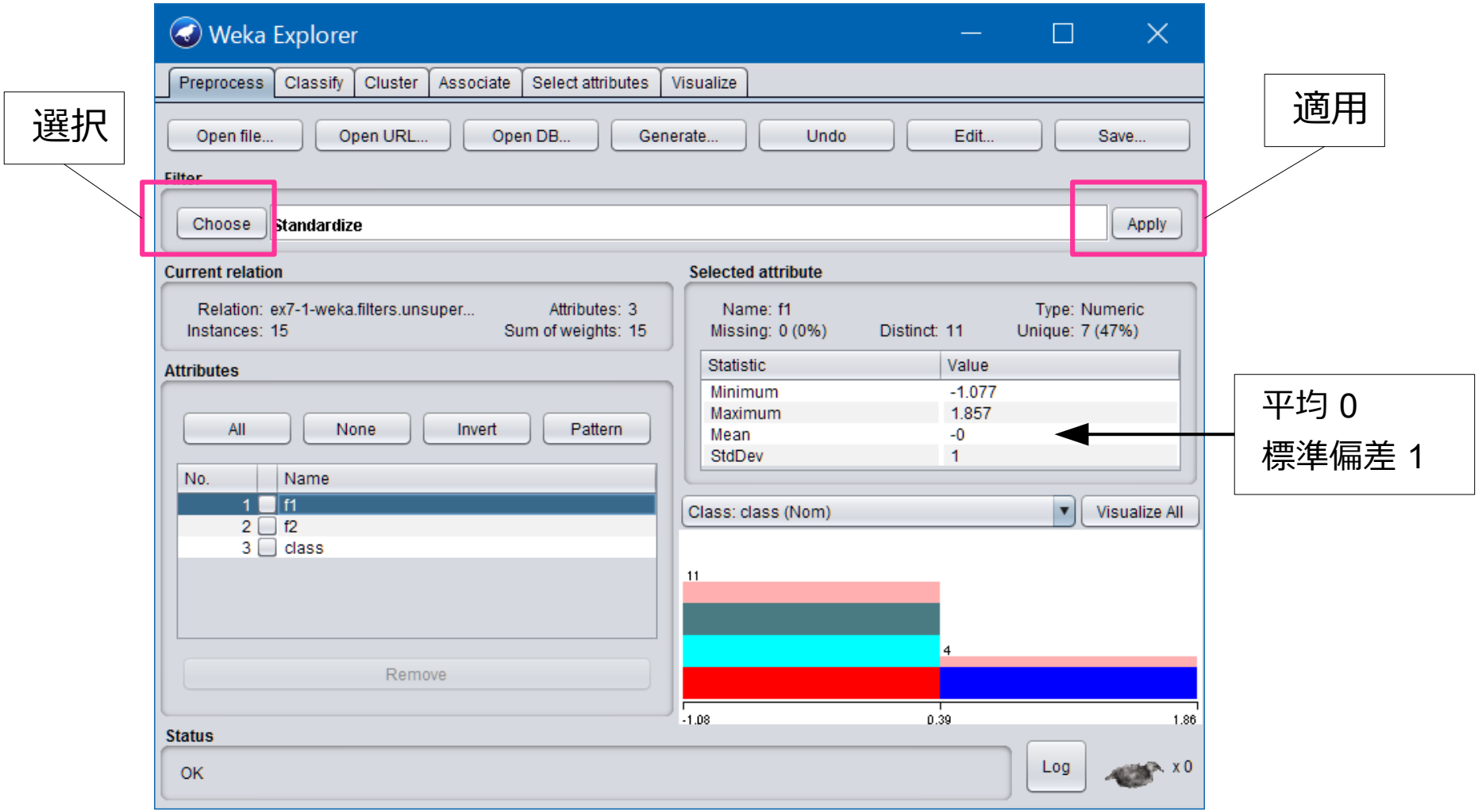
- フィルタの適用
  - 有用なフィルタのほとんどは

    weka → filters → unsupervised → attribute

    の下にある
  - Standardize：標準化（平均 0，分散 1）
    - 各次元に対して平均値を引き、標準偏差で割る
  - Normalize：値を [0,1] に変換
  - PrincipalComponents：主成分分析

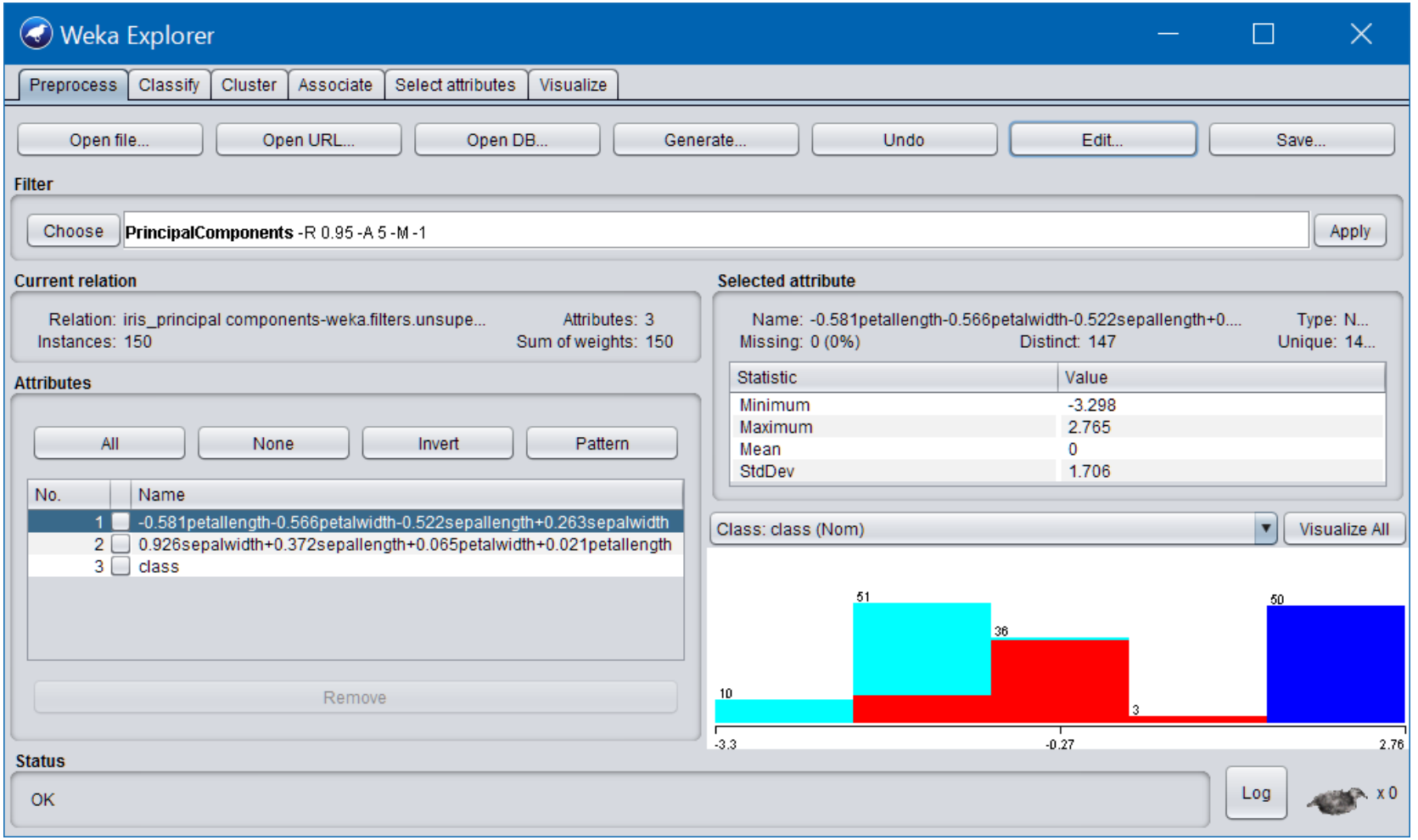
前処理（Preprocess）

- 標準化

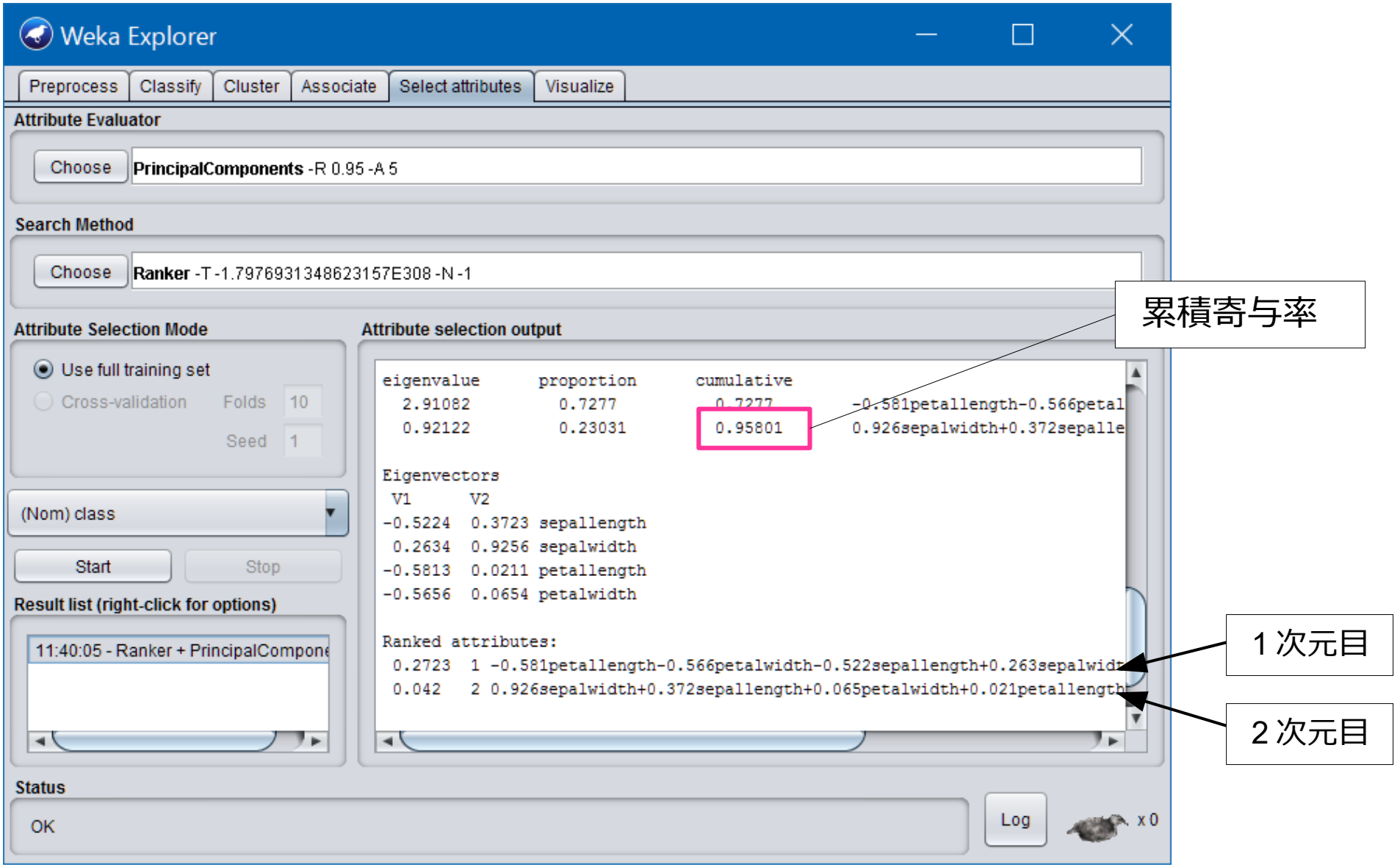


前処理（Preprocess）

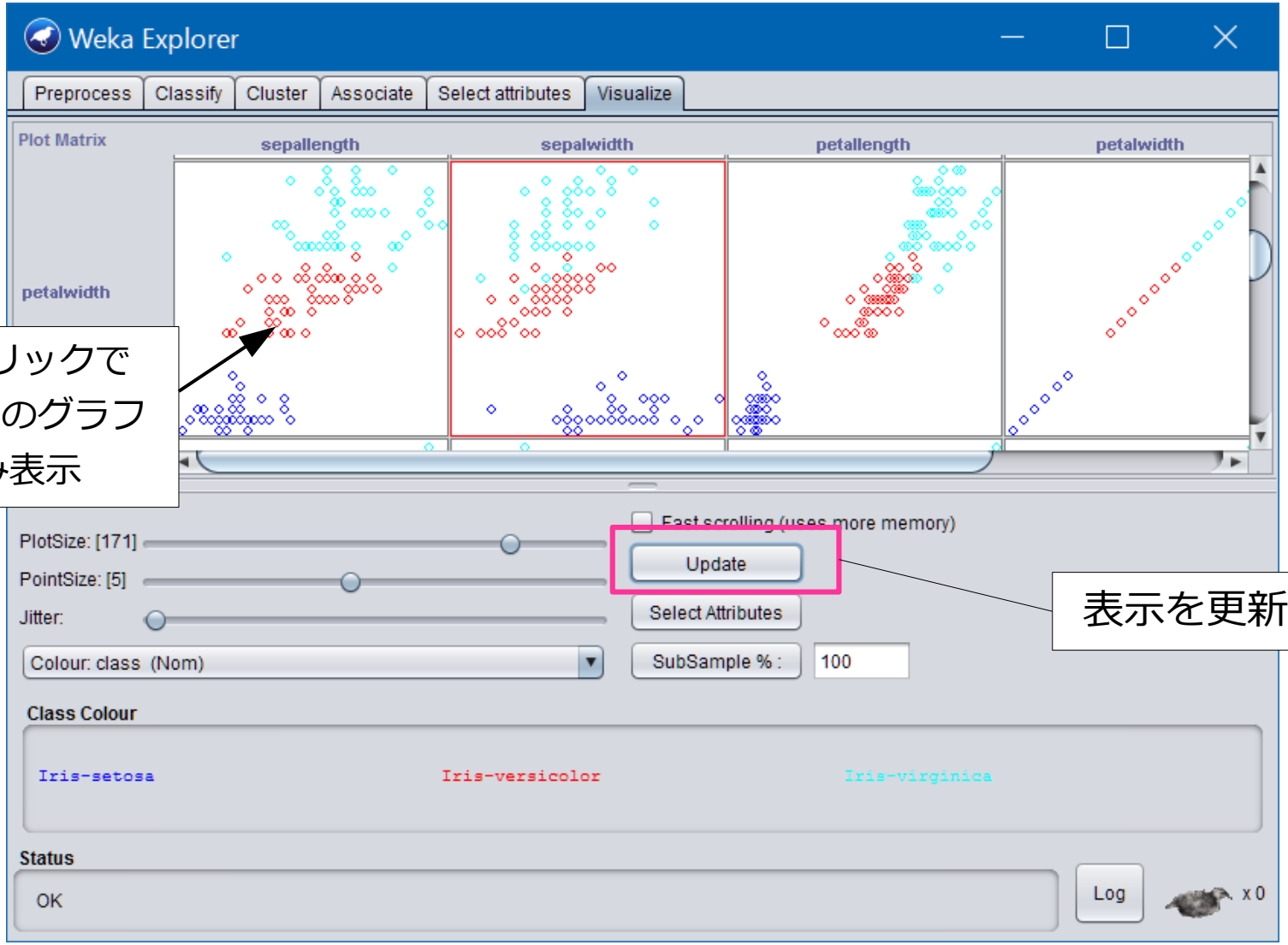
- 主成分分析
  - iris データ (4 次元特徴) を 2 次元に



補足 – Select Attributes での主成分分析

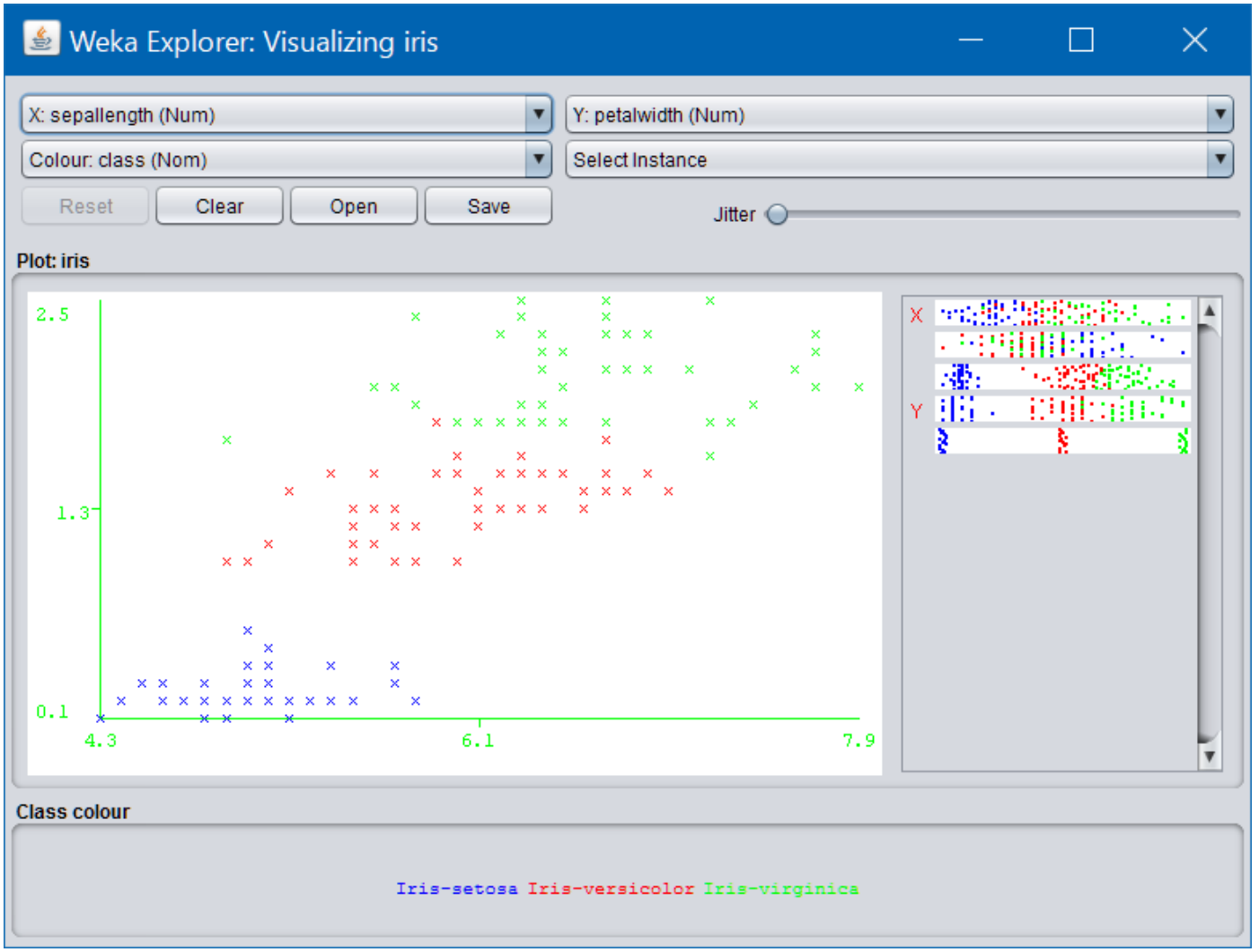


データのプロット（Visualize）



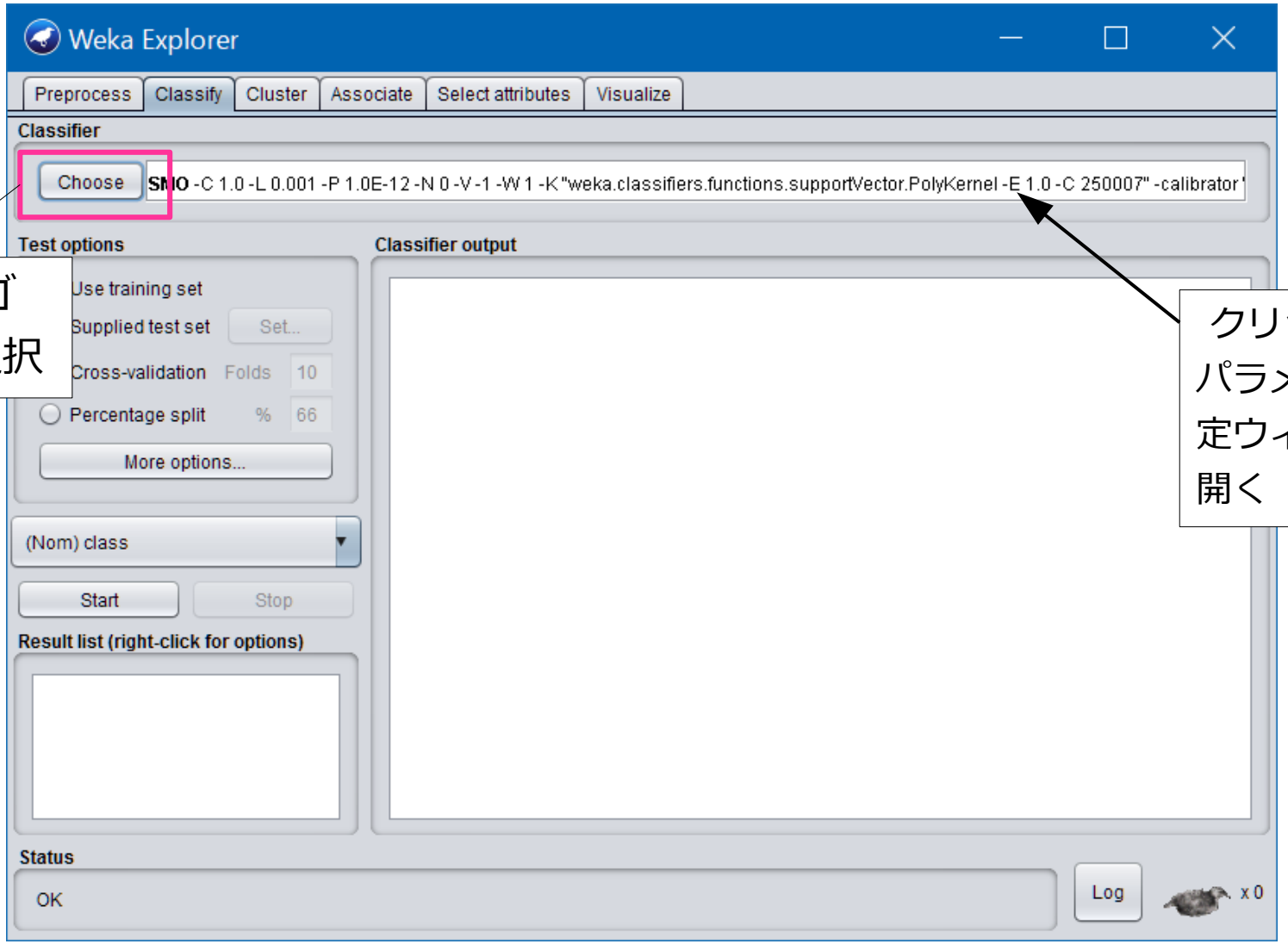
データのプロット（Visualize）

- 1つのグラフのみ表示



x軸、 y軸、色の基準が選べる

識別器の学習（Classify）

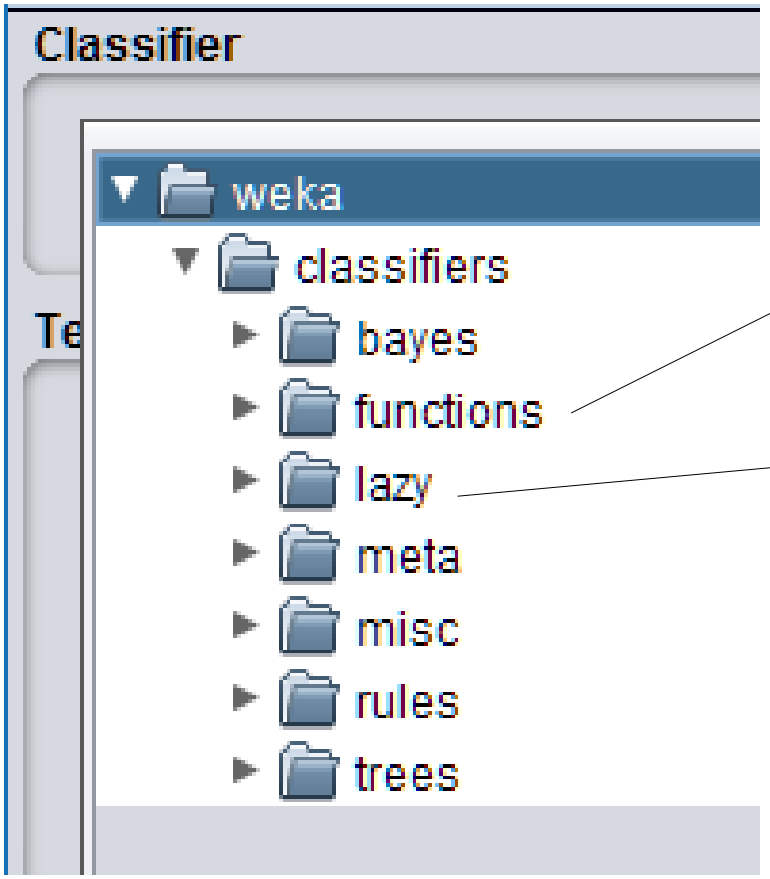


識別アルゴリズムの選択

クリックして、パラメータの設定ウィンドウを開く

識別器の学習（Classify）

- 勉強した識別器

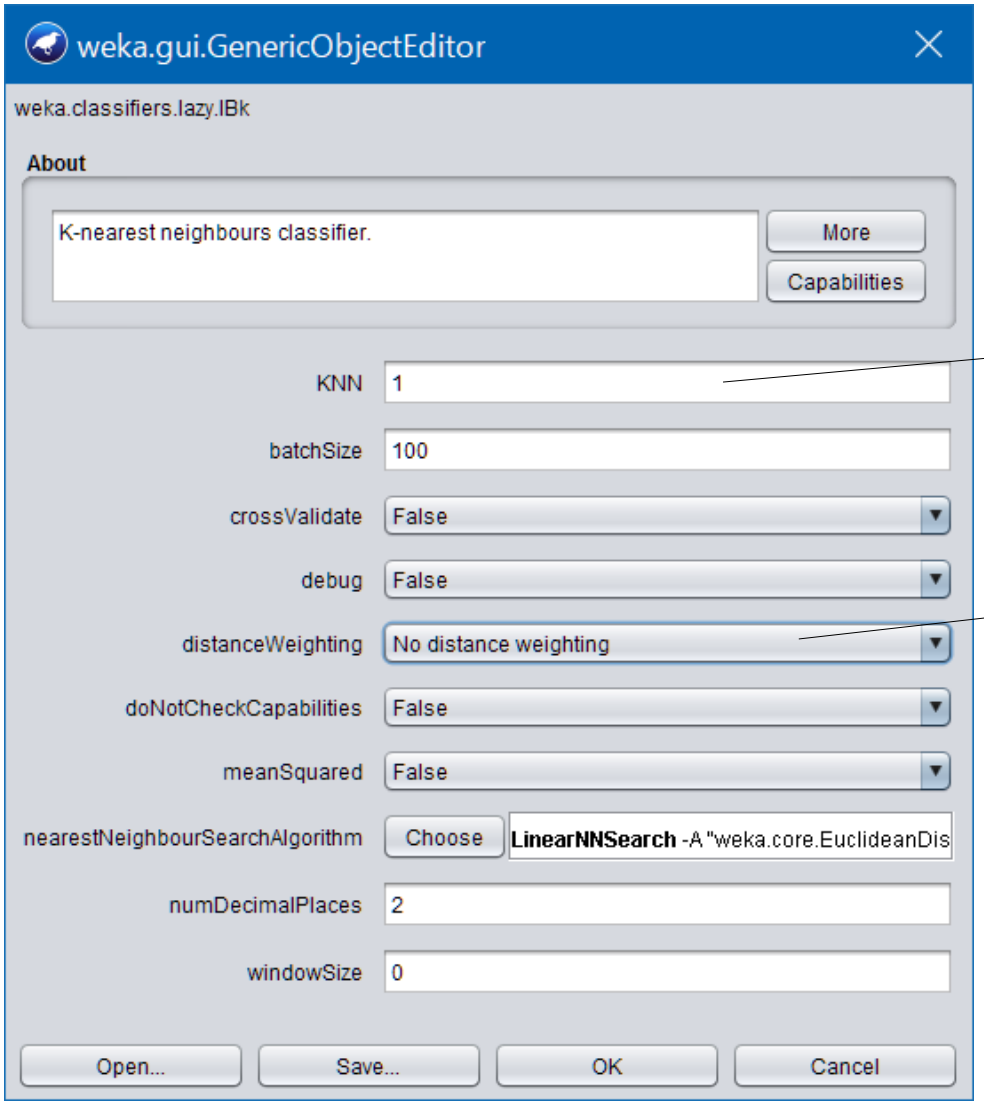


- MultilayerPerceptron (ニューラルネット)
- SMO (SVM)

- IBk (k-NN)

識別器の学習（Classify）

- IBk (k-NN 法) のパラメータ

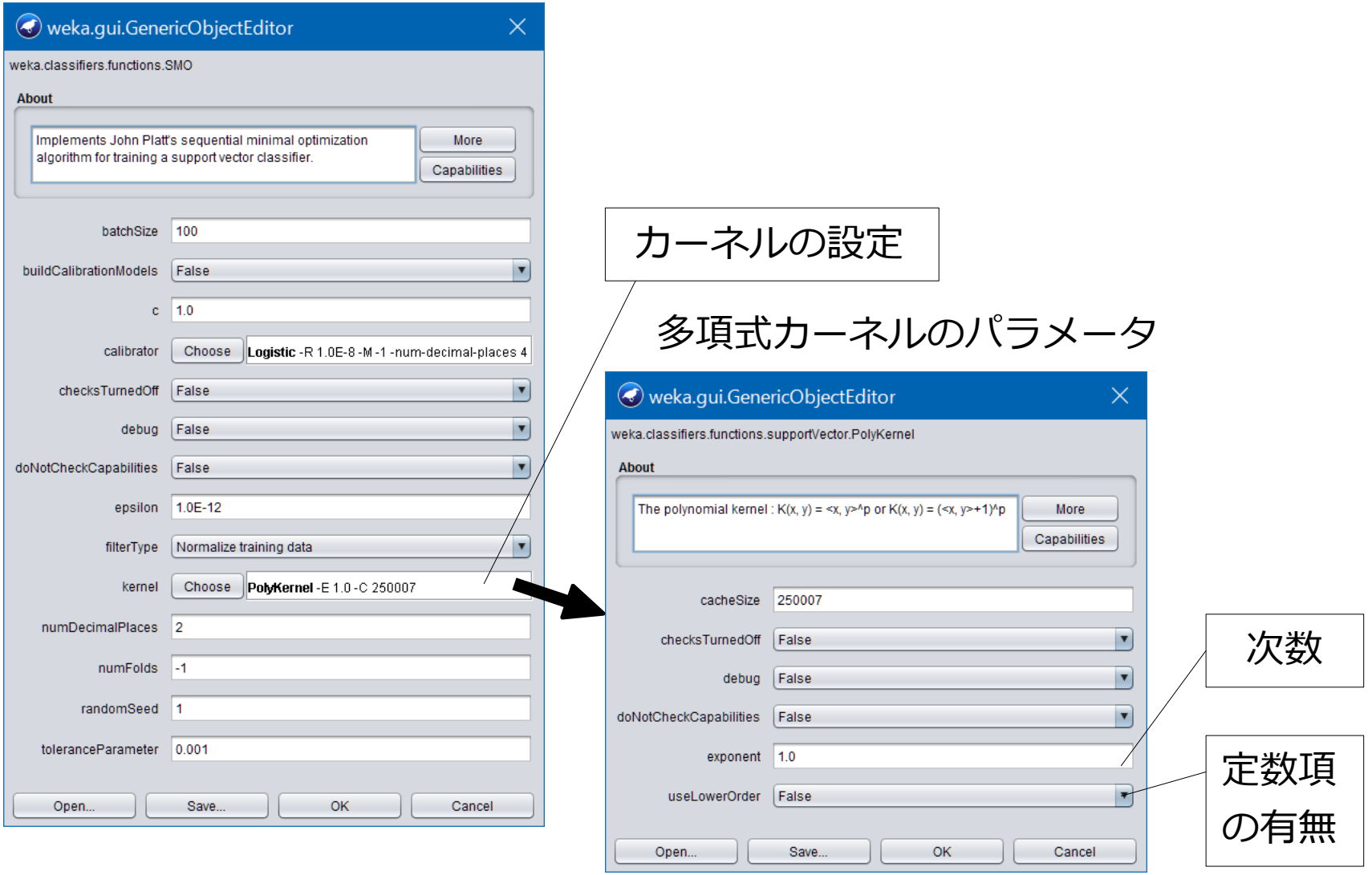


k

距離による重み付けの有無

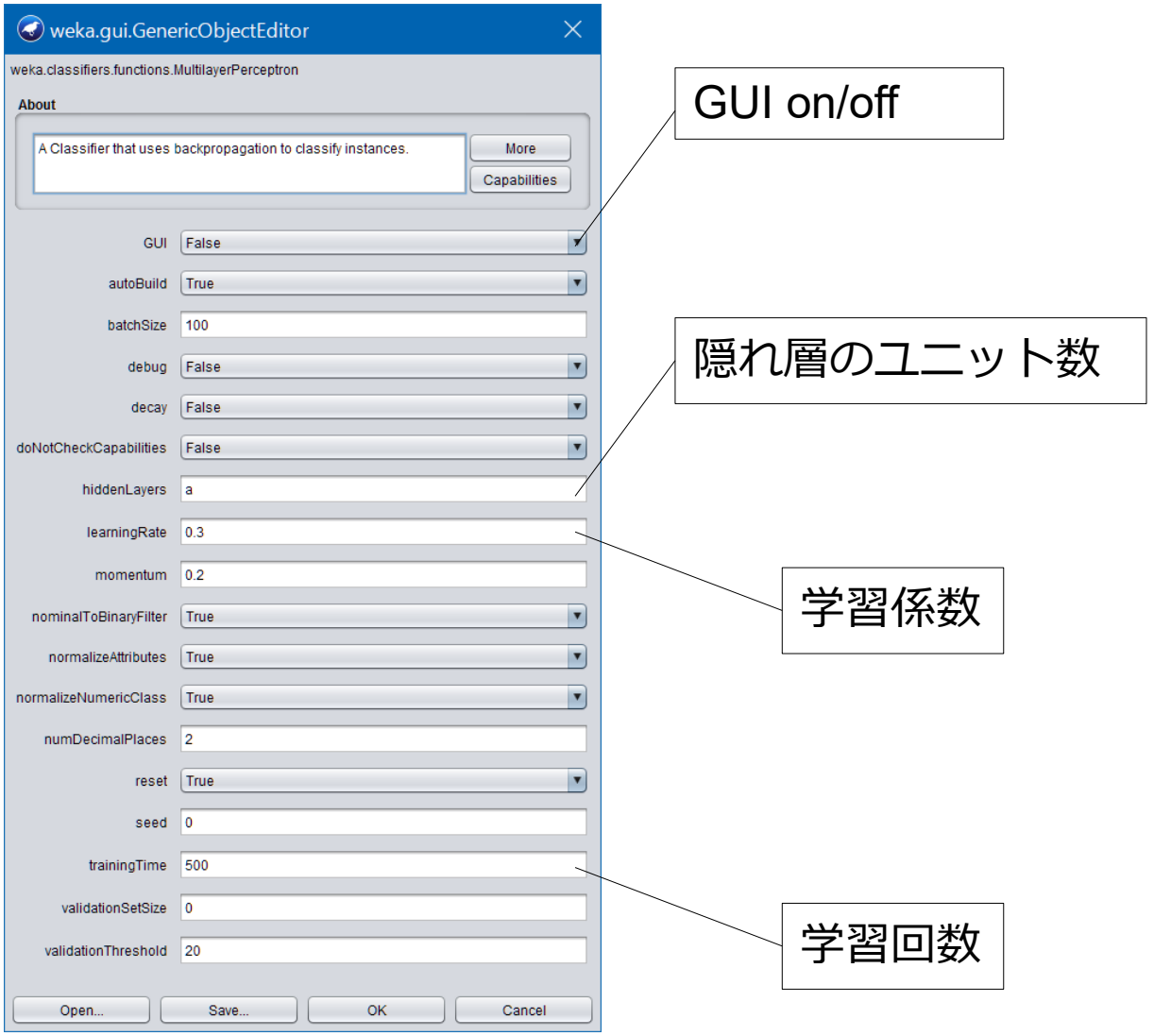
識別器の学習（Classify）

- SMO のパラメータ



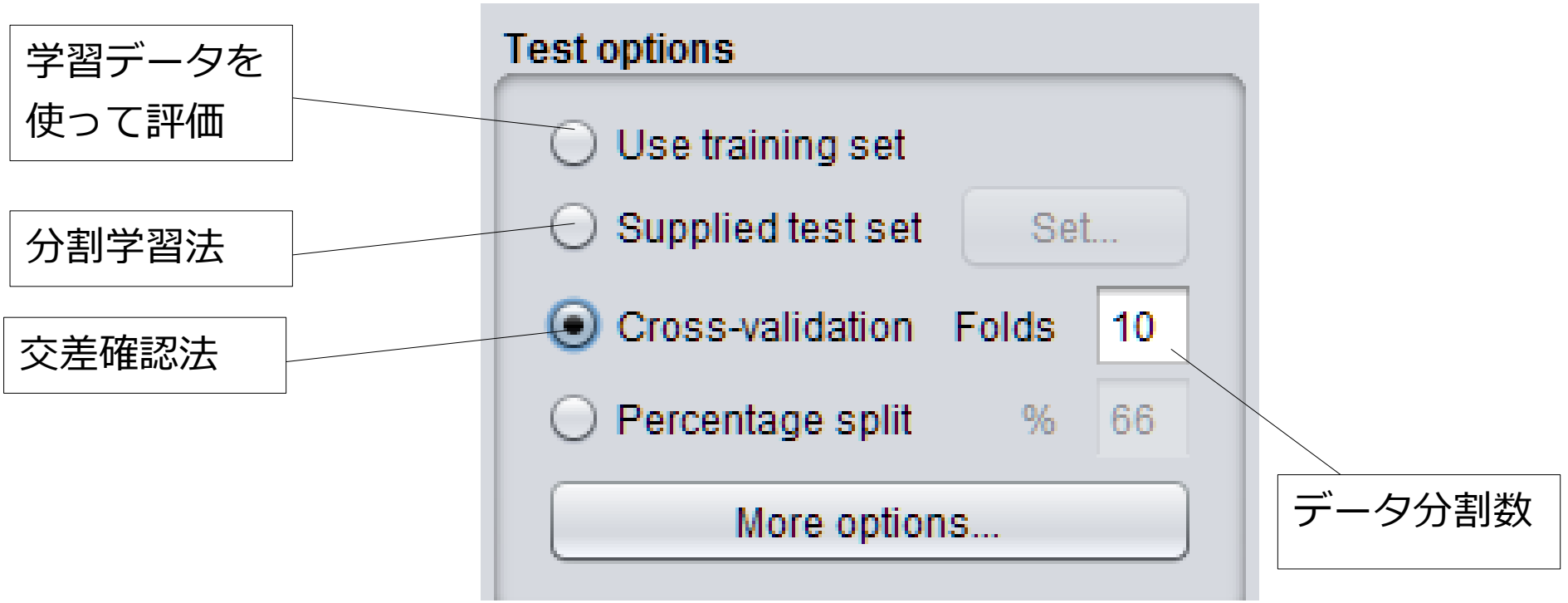
識別器の学習（Classify）

- MultilayerPerceptron のパラメータ



識別器の学習（Classify）

- 評価法の設定



識別器の学習（Classify）

- 学習結果の見方

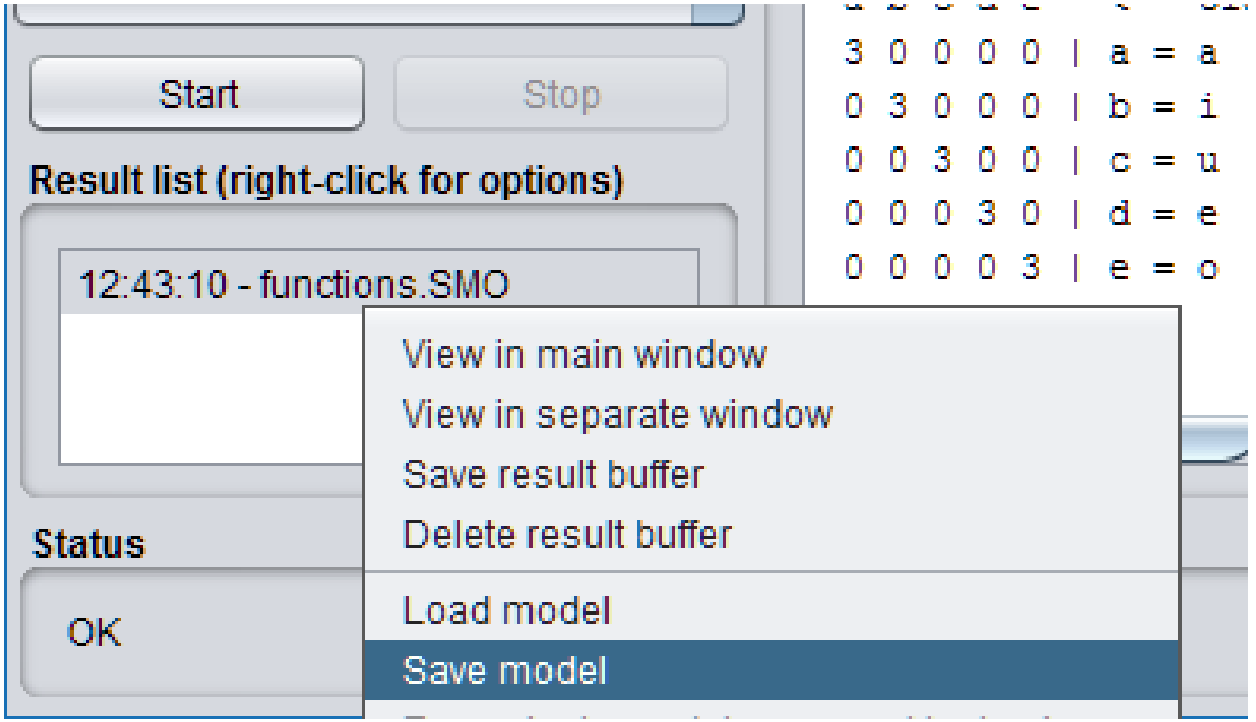
識別率

```
=== Summary ===

Correctly Classified Instances          14               93.3333 %
Incorrectly Classified Instances         1                6.6667 %
Kappa statistic                          0.9167
Mean absolute error                      0.1051
Root mean squared error                  0.1645
Relative absolute error                 31.4161 %
Root relative squared error             39.3051 %
Total Number of Instances               15

...

=== Confusion Matrix ===

 a b c d e   <-- classified as
 3 0 0 0 0 | a = a
 0 3 0 0 0 | b = i
 0 0 3 0 0 | c = u
 0 0 0 3 0 | d = e
 1 0 0 0 2 | e = o
```

縦軸が正解、横軸が出力
対角成分が正解数

識別器の学習（Classify）

- 学習結果の保存
  - Result list の該当行を右クリック → Save model
  - Weka を使う Java プログラムでロード可能



Section5 のまとめ

- Scilab
  - ベクトルや行列を変数の値とすることができる
  - 数式をそのままコードにすることができる
- Weka
  - 勉強に有用なサンプルデータが付属する
  - さまざまな機械学習手法が実装されている
  - 簡単にパラメータを変化させて影響を見ることができる。