

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

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

```
1 clear;
2 X=[3,-2;-3,-4;-5,-4;-5,-6]; //-元データ
3 //-元データの表示
4 subplot(1,2,1); title('original data');
5 plot2d(X(:,1), X(:,2), -style=-9, rect=[-8,-8,8,8], axesflag=4);
6
7 //-標準化計算 (SX=-wcenter(X, 'r')としてもよい)
8 [n, d] = size(X);
9 SX = (X - repmat(mean(X, 'r'), n, 1)) * diag(1 ./ stdev(X, 'r'));
10
11 //-標準化後のデータの表示
12 subplot(1,2,2); title('after standardization');
13 plot2d(SX(:,1), SX(:,2), -style=-4, rect=[-2,-2,2,2], axesflag=4);
14
```

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-cancer	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

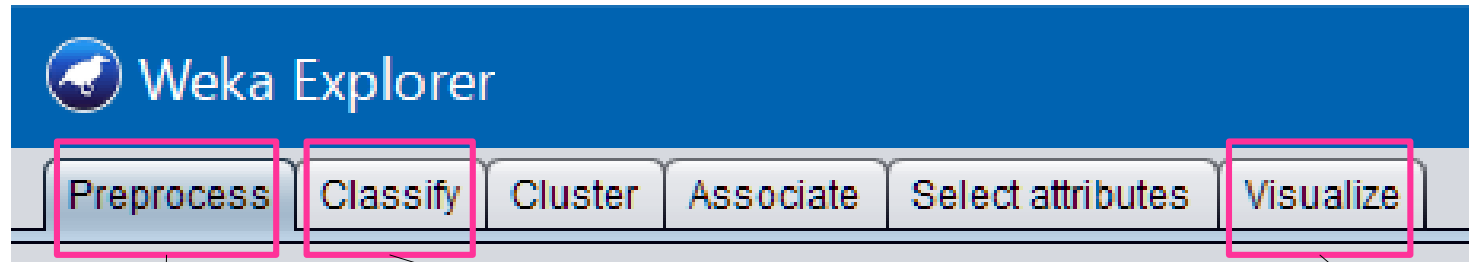
- アプリケーションの選択



- **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

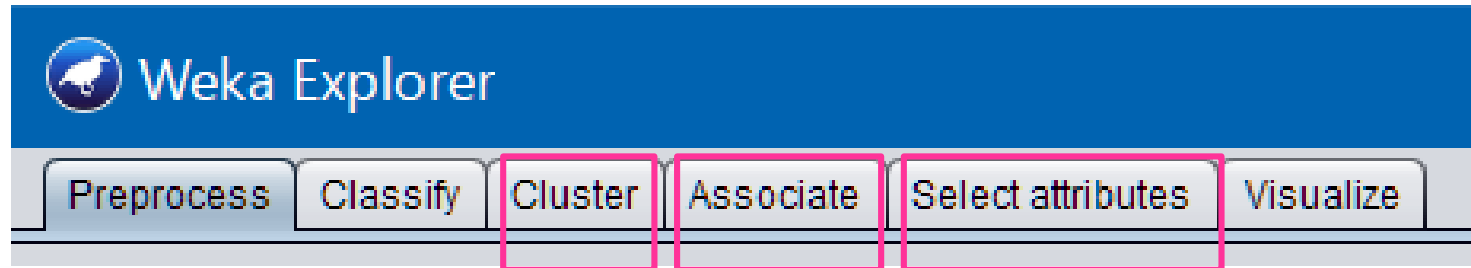
- 識別

- 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

The screenshot shows the Weka Explorer application window. The 'Preprocess' tab is selected. The 'Open file...' button is highlighted with a pink box and labeled '読み込み' (Load). The 'Edit...' button is highlighted with a pink box and labeled 'データの表示' (Data display). The 'Filter' section has a 'Choose' button highlighted with a pink box and labeled 'フィルタ' (Filter). The 'Current relation' section shows 'Relation: ex7-1' and 'Instances: 15'. The 'Attributes' section has a table with columns 'No.' and 'Name'. The 'Selected attribute' section shows 'Name: f1' and 'Type: Numeric'. The 'Class' is set to 'vowel (Nom)'. A horizontal bar chart is displayed at the bottom right, showing the distribution of the 'vowel' class across the range of 'f1' values.

読み込み

フィルタ

データの全体像

分析対象の特徴（属性）の選択

データの表示

選択された特徴の分析

No.	Name
1	f1
2	f2
3	vowel

Statistic	Value
Minimum	210
Maximum	800
Mean	426.667
StdDev	201.092

Class: vowel (Nom)

Visualize All

11

4

210 505 800

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

前处理 (Preprocess)

- 標準化

The screenshot shows the Weka Explorer interface with the 'Preprocess' tab selected. The 'Filter' dropdown is set to 'Standardize'. The 'Choose' button is highlighted with a pink box and labeled '選択' (Select). The 'Apply' button is highlighted with a pink box and labeled '適用' (Apply). The 'Current relation' shows 'Relation: ex7-1-weka.filters.unsuper...' and 'Instances: 15'. The 'Attributes' list shows 'f1', 'f2', and 'class'. The 'Selected attribute' section for 'f1' shows 'Type: Numeric' and 'Distinct: 11'. A table of statistics for 'f1' is displayed:

Statistic	Value
Minimum	-1.077
Maximum	1.857
Mean	-0
StdDev	1

An arrow points from the 'Mean' value (-0) to a text box labeled '平均 0' (Mean 0). Another arrow points from the 'StdDev' value (1) to a text box labeled '標準偏差 1' (Standard Deviation 1). A horizontal bar chart at the bottom shows the distribution of 'f1' values, with a peak at 0.39 and a range from -1.08 to 1.86.

前処理 (Preprocess)

- 主成分分析
 - iris データ (4 次元特徴) を 2 次元に

The screenshot shows the Weka Explorer application window. The 'Preprocess' tab is selected. The 'Filter' section shows 'PrincipalComponents -R 0.95 -A 5 -M -1' applied. The 'Current relation' section shows 'Relation: iris_principal components-weka.filters.unsupe...' with 3 attributes and 150 instances. The 'Attributes' section shows a list of attributes, including the principal components and the class attribute. The 'Selected attribute' section shows the selected attribute name and its statistics. The 'Class' dropdown is set to 'class (Nom)'. A histogram is displayed at the bottom right, showing the distribution of the class attribute across the principal components.

Weka Explorer

Preprocess | Classify | Cluster | Associate | Select attributes | Visualize

Open file... | Open URL... | Open DB... | Generate... | Undo | Edit... | Save...

Filter

Choose **PrincipalComponents -R 0.95 -A 5 -M -1** Apply

Current relation

Relation: iris_principal components-weka.filters.unsupe... Attributes: 3
Instances: 150 Sum of weights: 150

Attributes

All | None | Invert | Pattern

No.	Name
1	<input checked="" type="checkbox"/> -0.581petallength-0.566petalwidth-0.522sepalwidth+0.263sepalwidth
2	<input type="checkbox"/> 0.926sepalwidth+0.372sepalwidth+0.065petalwidth+0.021petallength
3	<input type="checkbox"/> class

Remove

Selected attribute

Name: -0.581petallength-0.566petalwidth-0.522sepalwidth+0.263sepalwidth Type: N...
Missing: 0 (0%) Distinct: 147 Unique: 14...

Statistic	Value
Minimum	-3.298
Maximum	2.765
Mean	0
StdDev	1.706

Class: class (Nom) Visualize All

Histogram showing the distribution of the class attribute across the principal components. The x-axis represents the principal component values, and the y-axis represents the frequency of instances. The distribution is skewed to the right, with a peak around -0.27.

Status: OK Log x 0

補足 – Select Attributes での主成分分析

The screenshot shows the Weka Explorer interface with the 'Select attributes' tab active. The 'Attribute Evaluator' is set to 'PrincipalComponents -R 0.95 -A 5'. The 'Search Method' is 'Ranker -T -1.7976931348623157E308 -N -1'. The 'Attribute Selection Mode' is 'Use full training set'. The 'Attribute selection output' pane displays the following data:

eigenvalue	proportion	cumulative	
2.91082	0.7277	0.7277	-0.581petallength-0.566petalwidth
0.92122	0.23031	0.95801	0.926sepalwidth+0.372sepalwidth

Below the table, the 'Eigenvectors' section shows the loadings for V1 and V2:

	V1	V2	
	-0.5224	0.3723	sepalwidth
	0.2634	0.9256	sepalwidth
	-0.5813	0.0211	petallength
	-0.5656	0.0654	petalwidth

The 'Ranked attributes' section shows the top two attributes:

0.2723	1	-0.581petallength-0.566petalwidth-0.522sepalwidth+0.263sepalwidth	
0.042	2	0.926sepalwidth+0.372sepalwidth+0.065petalwidth+0.021petallength	

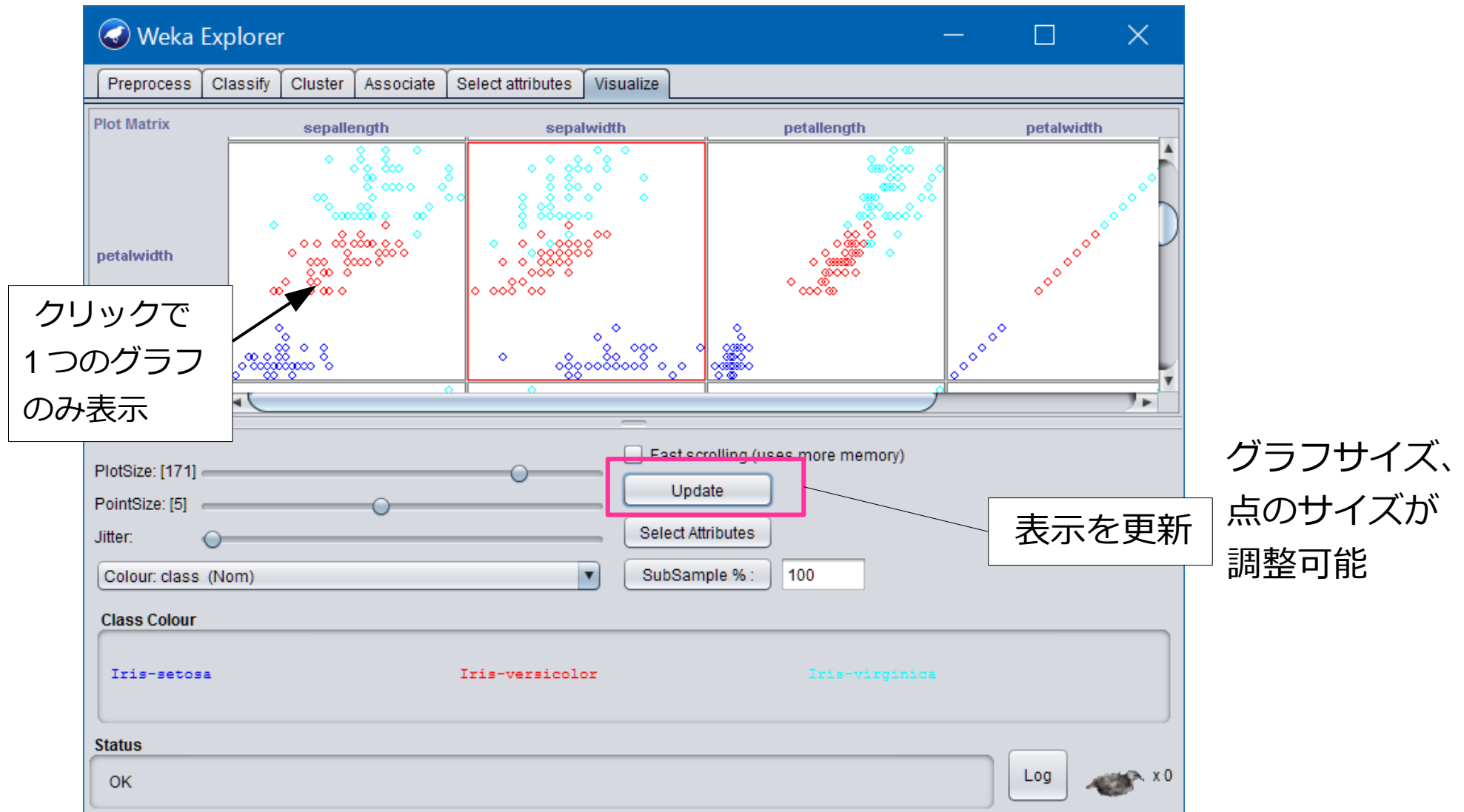
The 'Result list' on the left shows the selected attributes. The 'Status' bar at the bottom indicates 'OK'.

累積寄与率

1次元目

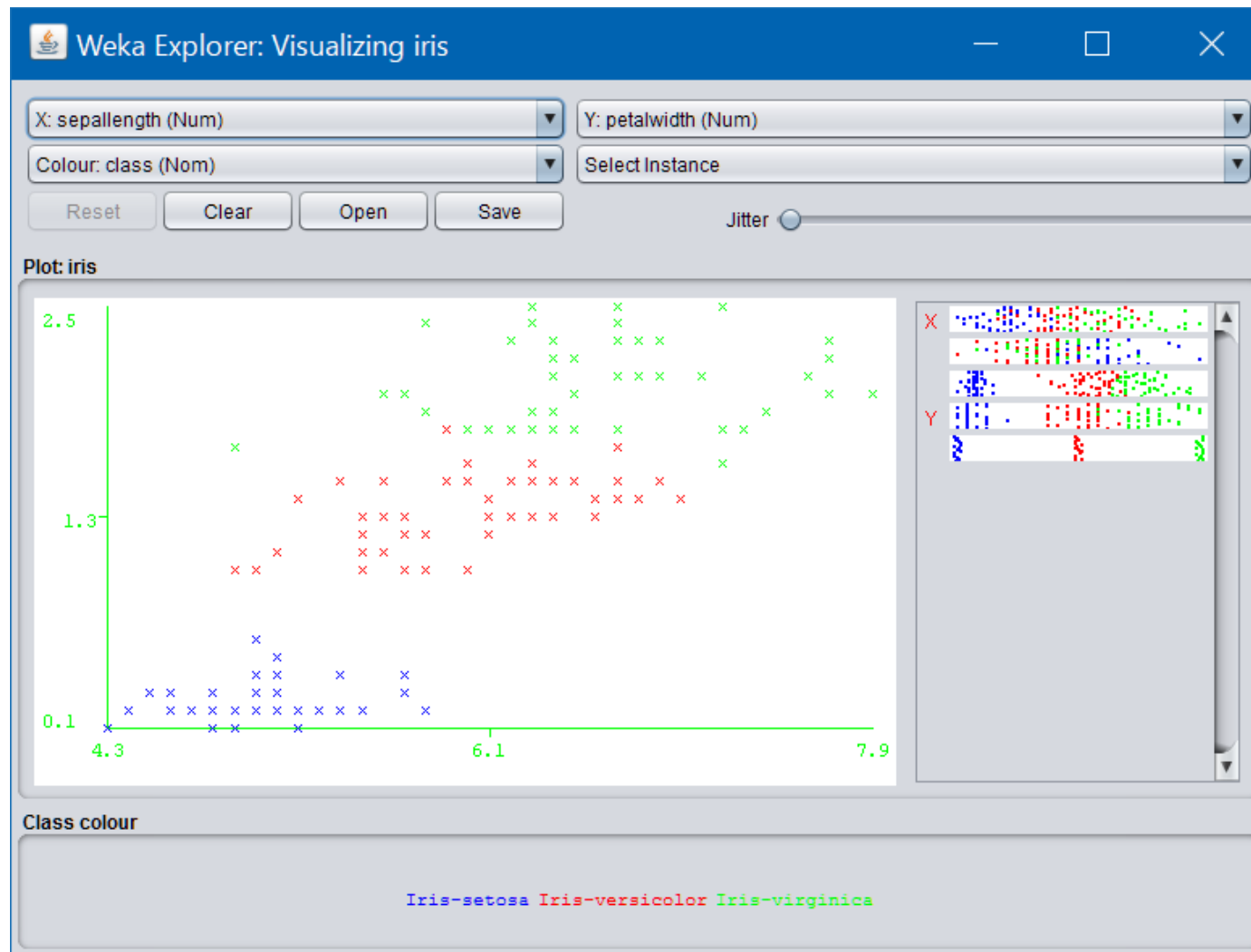
2次元目

データのプロット (Visualize)



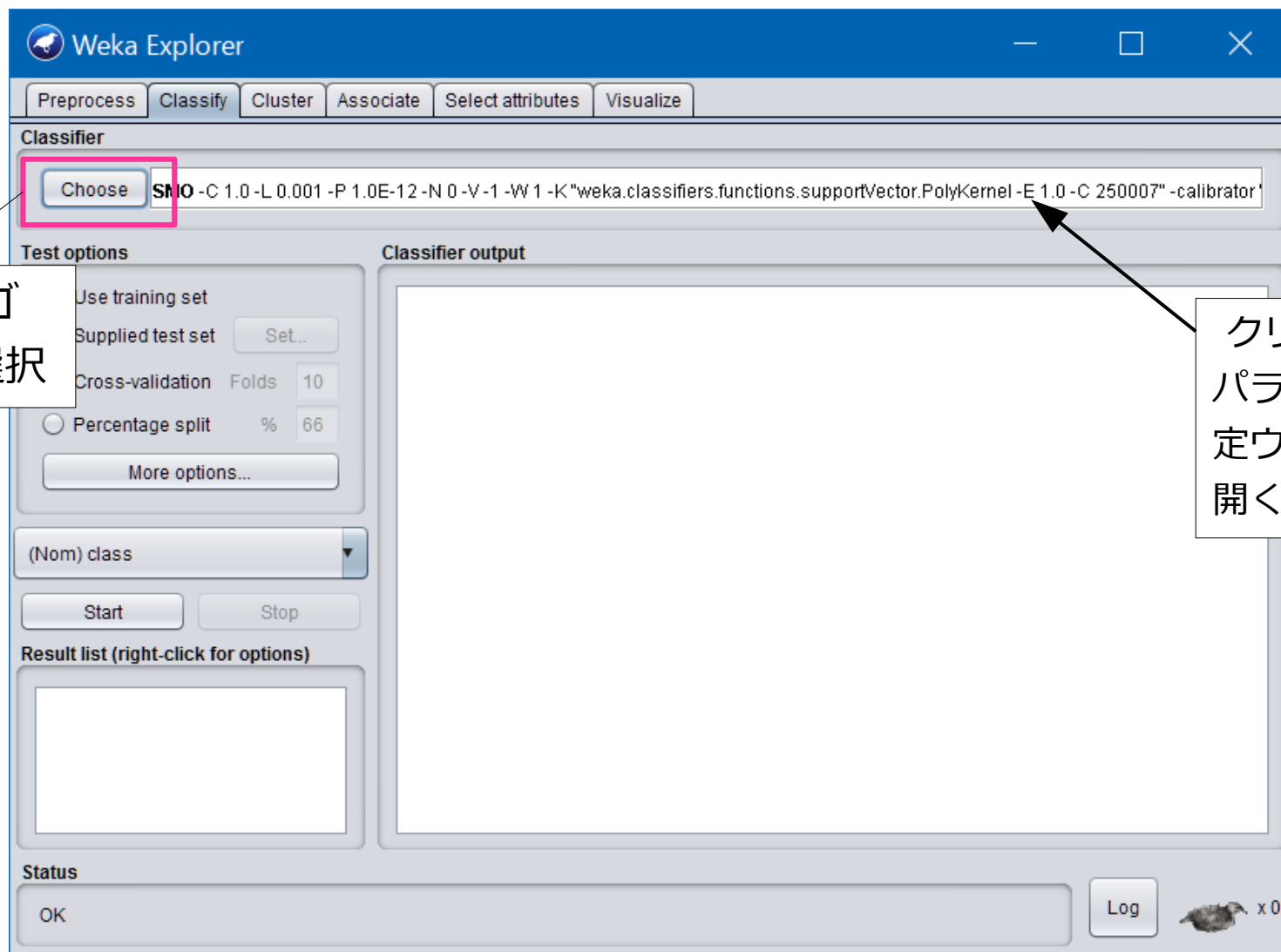
データのプロット (Visualize)

- 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)

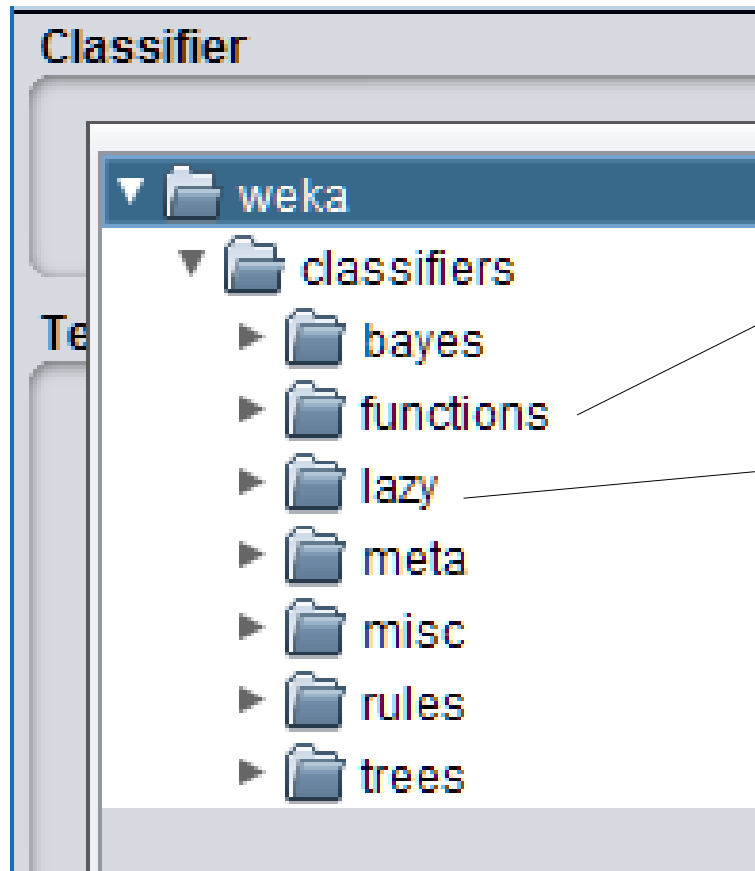


識別アルゴリズムの選択

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

The screenshot shows the 'weka.gui.GenericObjectEditor' window for the 'weka.classifiers.lazy.IBk' classifier. The 'About' section describes it as a 'K-nearest neighbours classifier.' with 'More' and 'Capabilities' buttons. The main configuration area includes the following parameters:

Parameter	Value
KNN	1
batchSize	100
crossValidate	False
debug	False
distanceWeighting	No distance weighting
doNotCheckCapabilities	False
meanSquared	False
nearestNeighbourSearchAlgorithm	Choose LinearNNSearch -A "weka.core.EuclideanDis
numDecimalPlaces	2
windowSize	0

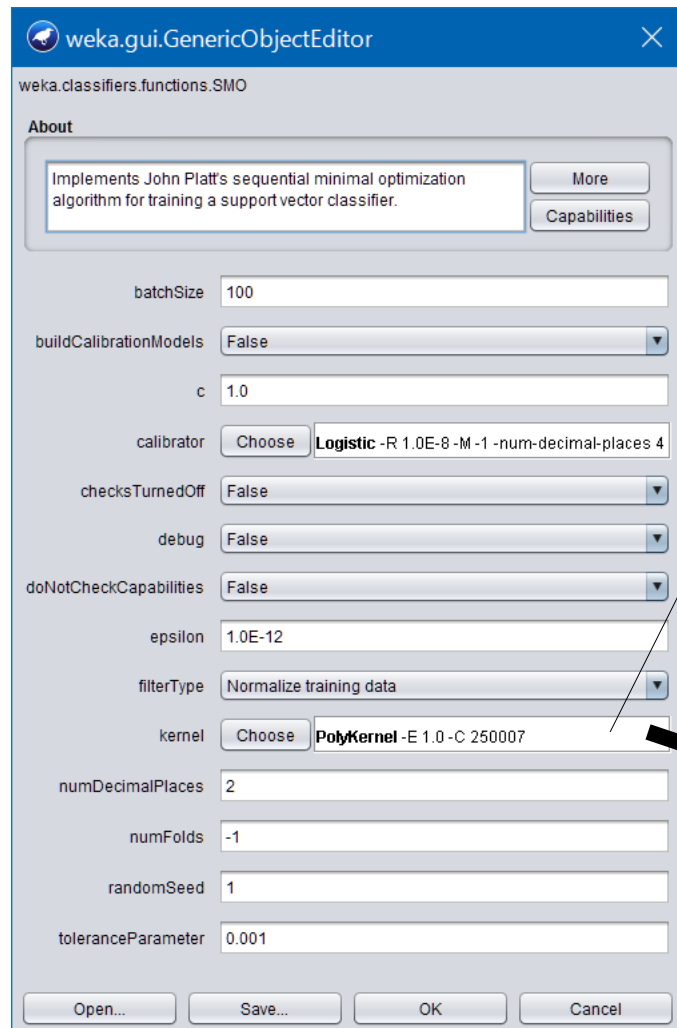
At the bottom are buttons for 'Open...', 'Save...', 'OK', and 'Cancel'.

k

距離による重み付けの有無

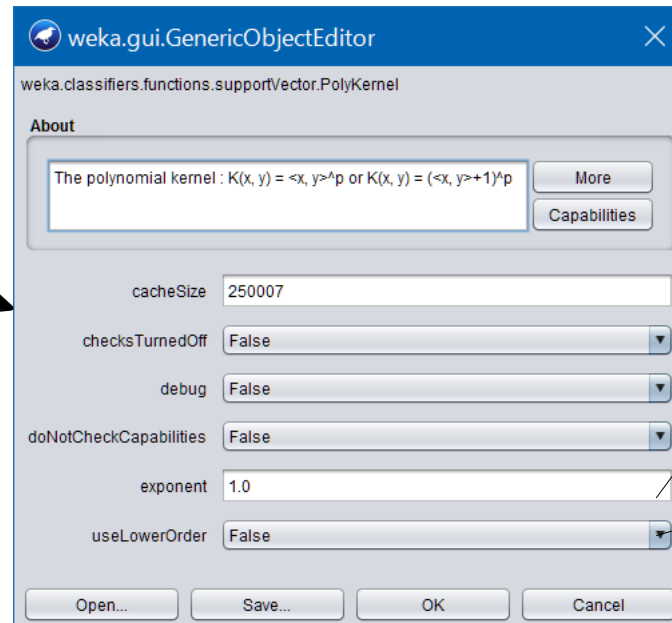
識別器の学習 (Classify)

- SMO のパラメータ



カーネルの設定

多項式カーネルのパラメータ



次数

定数項
の有無

識別器の学習 (Classify)

- MultilayerPerceptron のパラメータ

The screenshot shows the 'weka.gui.GenericObjectEditor' window for the 'weka.classifiers.functions.MultilayerPerceptron' classifier. The window has a title bar with a close button. Below the title bar is a tab labeled 'About' with a text area containing 'A Classifier that uses backpropagation to classify instances.' and buttons for 'More' and 'Capabilities'. The main area contains various parameters for the classifier, each with a label and a value field or dropdown menu. The parameters are: 'GUI' (False), 'autoBuild' (True), 'batchSize' (100), 'debug' (False), 'decay' (False), 'doNotCheckCapabilities' (False), 'hiddenLayers' (a), 'learningRate' (0.3), 'momentum' (0.2), 'nominalToBinaryFilter' (True), 'normalizeAttributes' (True), 'normalizeNumericClass' (True), 'numDecimalPlaces' (2), 'reset' (True), 'seed' (0), 'trainingTime' (500), 'validationSetSize' (0), and 'validationThreshold' (20). At the bottom are buttons for 'Open...', 'Save...', 'OK', and 'Cancel'.

Parameter	Value
GUI	False
autoBuild	True
batchSize	100
debug	False
decay	False
doNotCheckCapabilities	False
hiddenLayers	a
learningRate	0.3
momentum	0.2
nominalToBinaryFilter	True
normalizeAttributes	True
normalizeNumericClass	True
numDecimalPlaces	2
reset	True
seed	0
trainingTime	500
validationSetSize	0
validationThreshold	20

GUI on/off

隠れ層のユニット数

学習係数

学習回数

識別器の学習 (Classify)

- 評価法の設定

学習データを使
って評価

分割学習法

交差確認法

Test options

☐ Use training set

☐ Supplied test set Set...

☒ Cross-validation Folds 10

☐ Percentage split % 66

More options...

データ分割数

識別器の学習 (Classify)

- 学習結果の見方

```
=== Summary ===
```

Correctly Classified Instances	14
Incorrectly Classified Instances	1
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

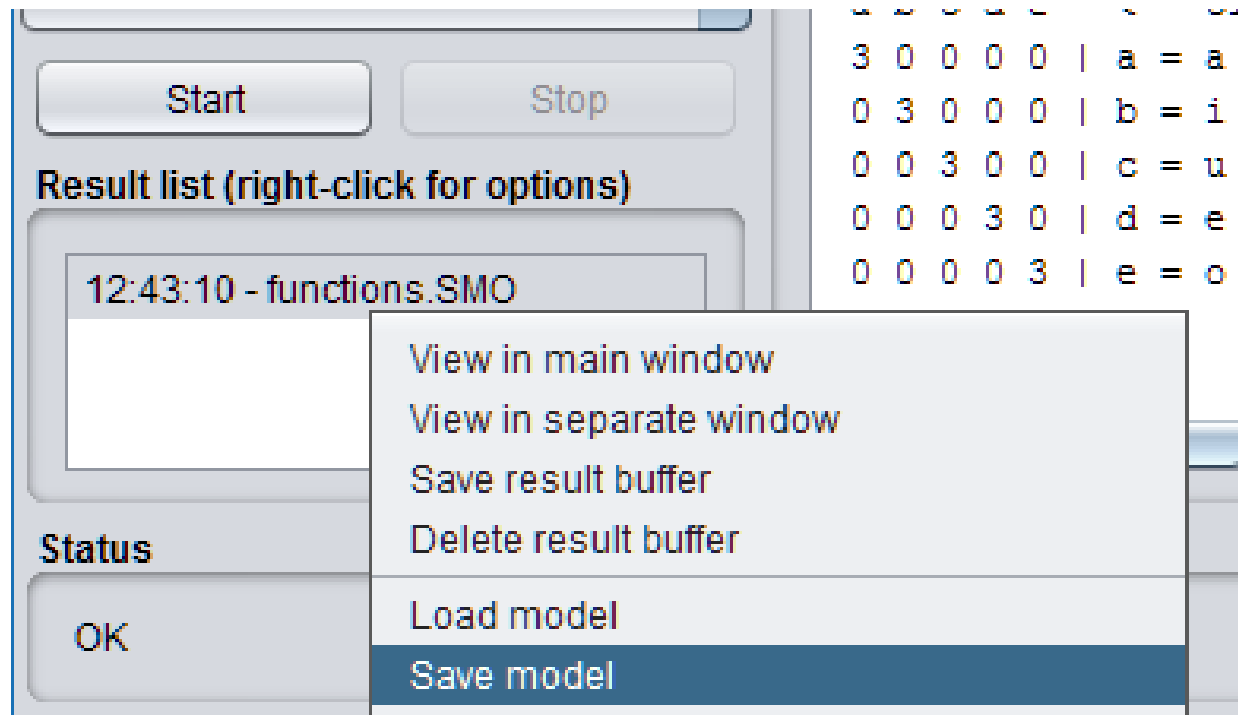
識別率

93.3333 %
6.6667 %

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。