

# 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. 3.9.2 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

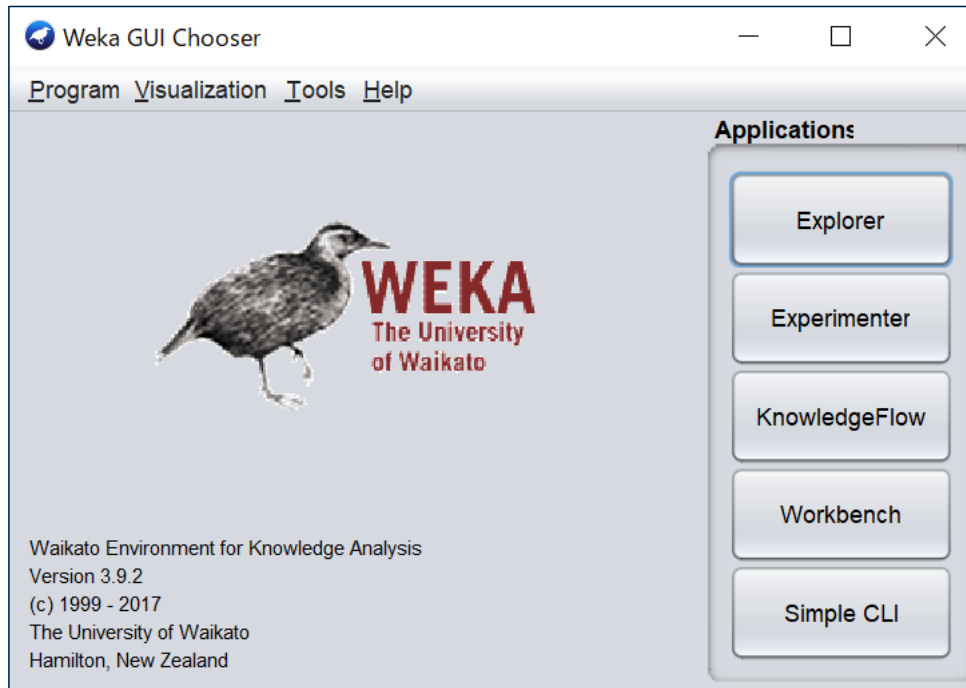
# 勉強のためのデータセット

表 2.2 Weka 付属のデータ (一部)

データ名	内容	特徴	正解情報
breast-cancer	乳癌の再発	カテゴリ	クラス (2 値)
contact-lenses	コンタクトレンズの推薦	カテゴリ	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)
ReutersCorn	記事分類	文字列	クラス (2 値)
supermarket	スーパーの購買記録	カテゴリ	なし
weather.nominal	ゴルフをする条件	カテゴリ	クラス (2 値)
weather.numeric	ゴルフをする条件	混合	クラス (2 値)

# 起動

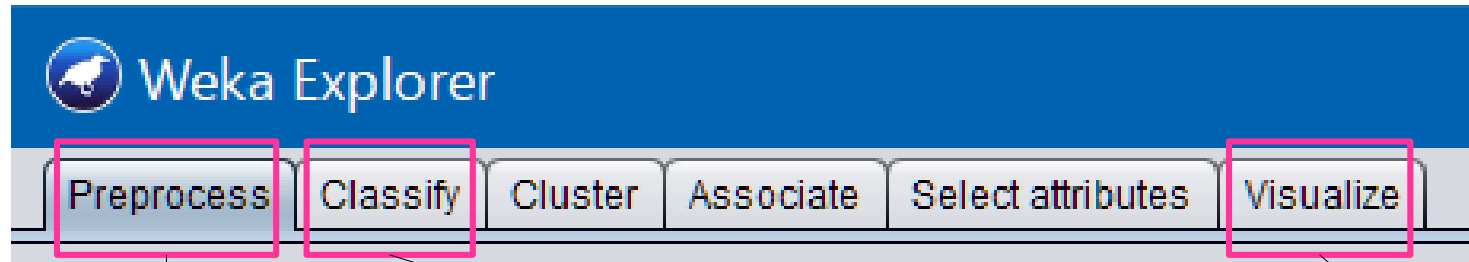
- アプリケーションの選択



• **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 (カスタマイズ可能)
- **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

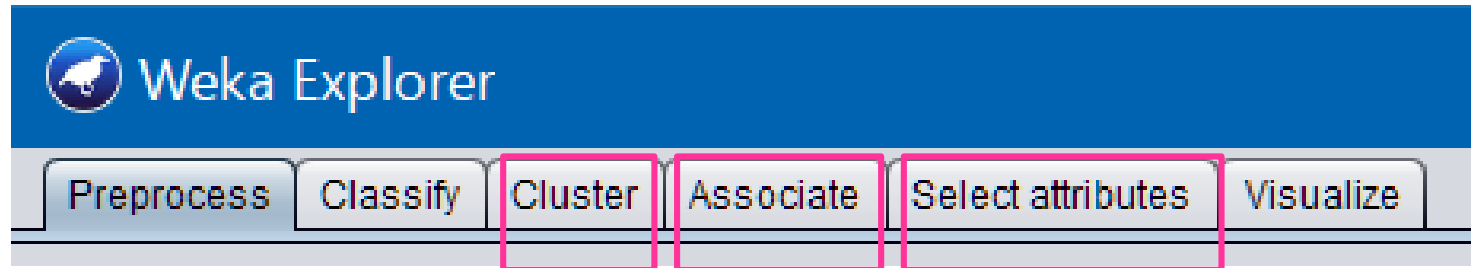
- 識別

- 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)

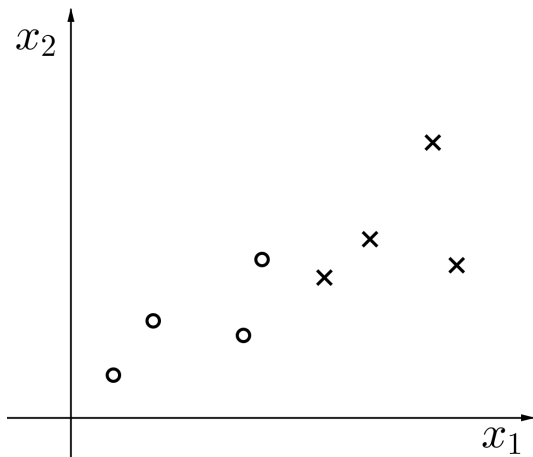
- CSV ファイルの読み込み
  - ファイル形式の前提
    - 1 行目は特徴名である
    - クラスラベルは文字列で表現されている
  - そうでないときは読み込み時に options dialog で指定

	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

# 主成分分析の考え方



共分散行列 $\Sigma$ の計算

$\bar{x}_1, \bar{x}_2$  : 平均値、 $N$  : データ数

対角成分は分散、  
非対角成分は相関を表す

$$\Sigma = \frac{1}{N} \begin{pmatrix} \sum (x_1 - \bar{x}_1)^2 & \sum (x_1 - \bar{x}_1)(x_2 - \bar{x}_2) \\ \sum (x_1 - \bar{x}_1)(x_2 - \bar{x}_2) & \sum (x_2 - \bar{x}_2)^2 \end{pmatrix}$$

$\Sigma$ は

半正定値(→固有値がすべて0以上の実数)

対称行列(→固有ベクトルが実数かつ直交)

であるので、以下のように分解できる

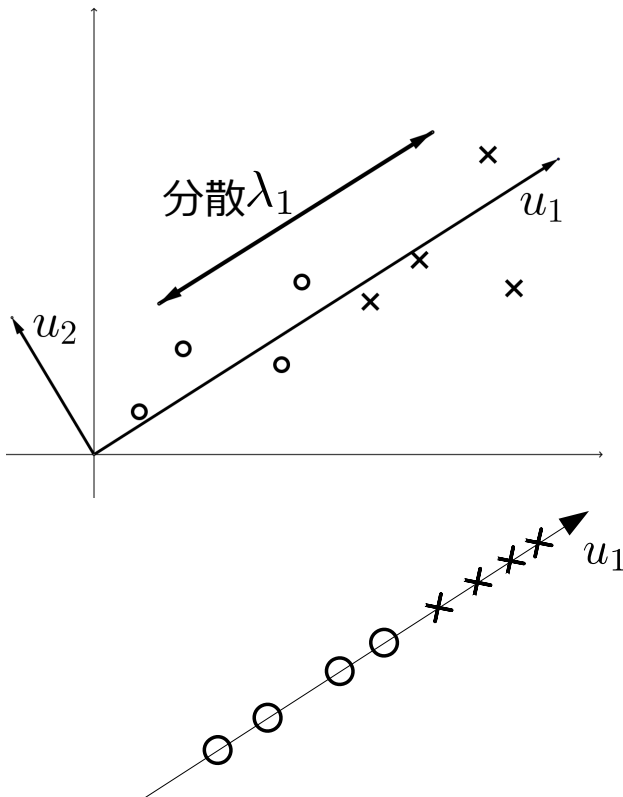
$$\Sigma' = U^T \Sigma U = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}$$

$\lambda$ は固有値の大きい順、 $U$ は対応する  
固有ベクトル $U_1, U_2$ を並べたもの

$\lambda_1$ に対応する固有ベクトル $U_1$ で  
2次元データを1次元に射影

$$u_1 = U_1^T x$$

$$\text{寄与率} = \frac{\lambda_1}{\lambda_1 + \lambda_2}$$



# 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 also highlighted with a pink box and labeled '適用' (Apply). The 'Current relation' section shows 'Relation: ex7-1-weka.filters.unsuper...' and 'Instances: 15'. The 'Attributes' section shows a list of attributes: f1, f2, and class. The 'Selected attribute' section shows statistics for attribute f1: Name: f1, Missing: 0 (0%), Distinct: 11, Type: Numeric, Unique: 7 (47%). A table of statistics 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) and '標準偏差 1' (StdDev 1). Below the statistics, a horizontal bar chart shows the distribution of values for attribute f1, with a range from -1.08 to 1.86. The status bar at the bottom shows 'OK' and a 'Log' button.

# 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 lists three attributes: 1.  $-0.581\text{petal length} - 0.566\text{petal width} - 0.522\text{sepal length} + 0.263\text{sepal width}$ , 2.  $0.926\text{sepal width} + 0.372\text{sepal length} + 0.065\text{petal width} + 0.021\text{petal length}$ , and 3. 'class'. The 'Selected attribute' section shows the selected attribute name and statistics: Minimum: -3.298, Maximum: 2.765, Mean: 0, StdDev: 1.706. The 'Class' is 'class (Nom)'. A histogram at the bottom shows the distribution of the selected attribute across three classes: class 1 (cyan, 10 instances), class 2 (red, 51 instances), and class 3 (blue, 50 instances).

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	$-0.581\text{petal length} - 0.566\text{petal width} - 0.522\text{sepal length} + 0.263\text{sepal width}$
2	$0.926\text{sepal width} + 0.372\text{sepal length} + 0.065\text{petal width} + 0.021\text{petal length}$
3	class

Remove

Selected attribute

Name:  $-0.581\text{petal length} - 0.566\text{petal width} - 0.522\text{sepal length} + 0.263\text{sepal width}$  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

10 51 36 3 50

-3.3 -0.27 2.76

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:

	Rank	Attribute
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'.

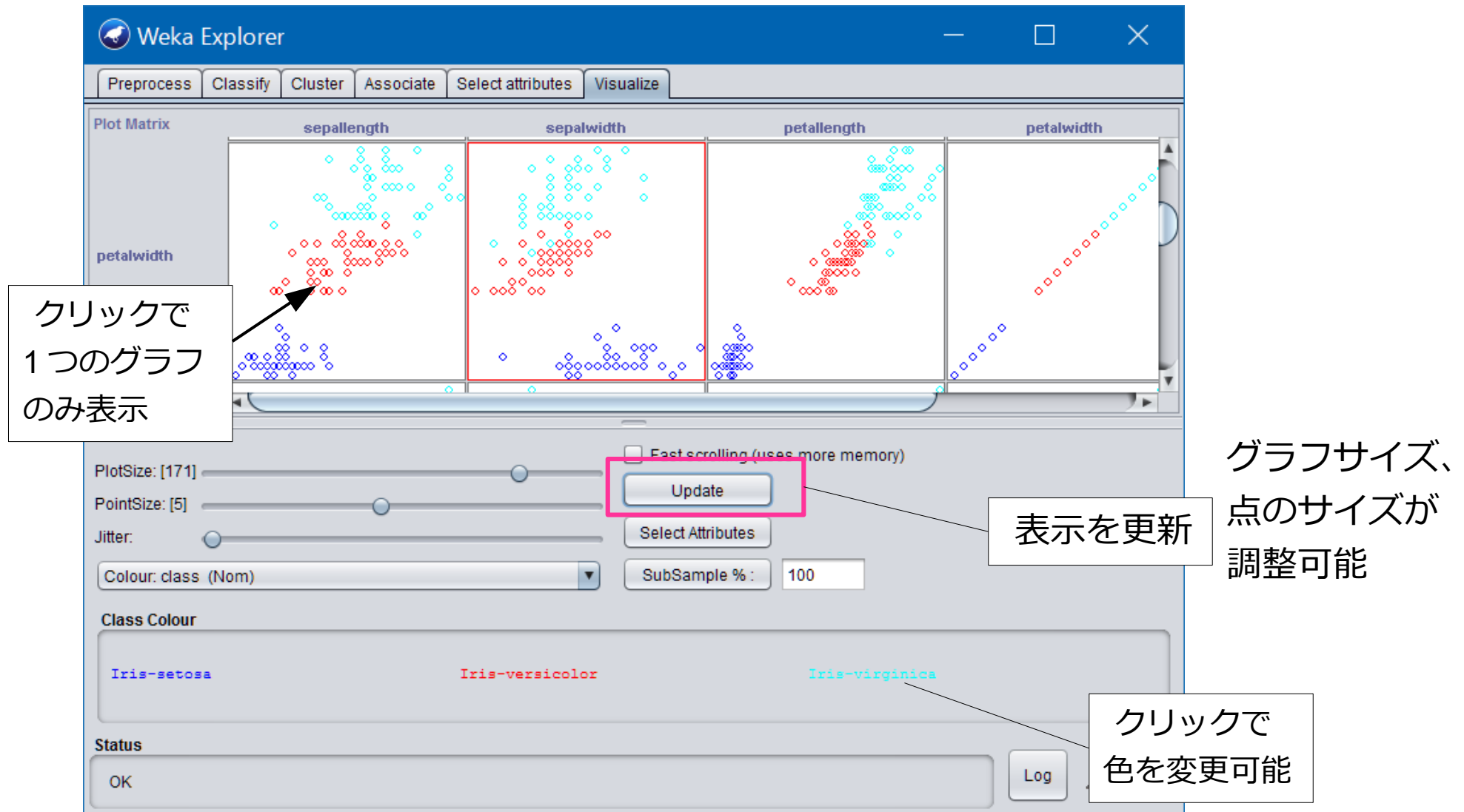
累積寄与率

1次元目

2次元目

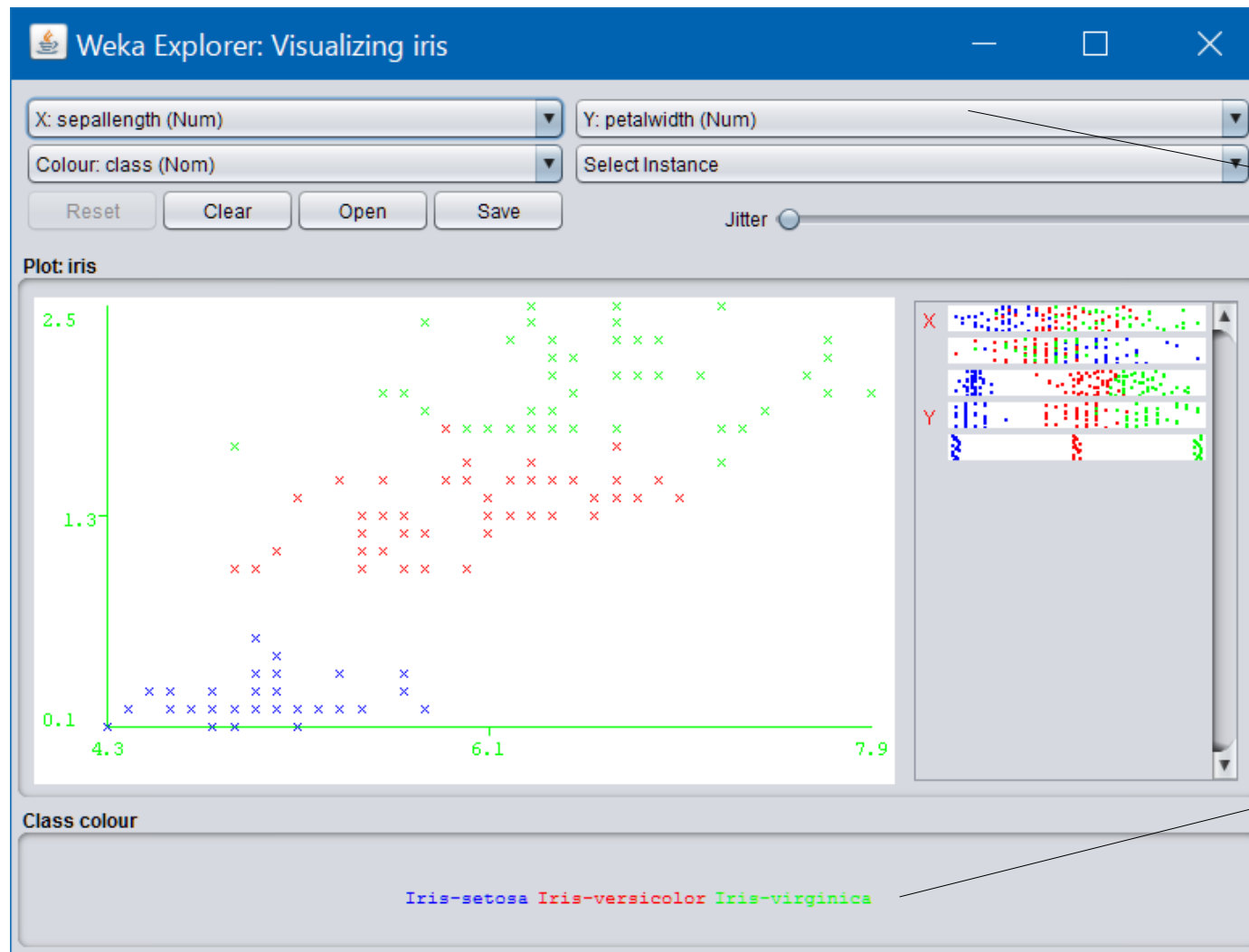


# データのプロット (Visualize)



# データのプロット (Visualize)

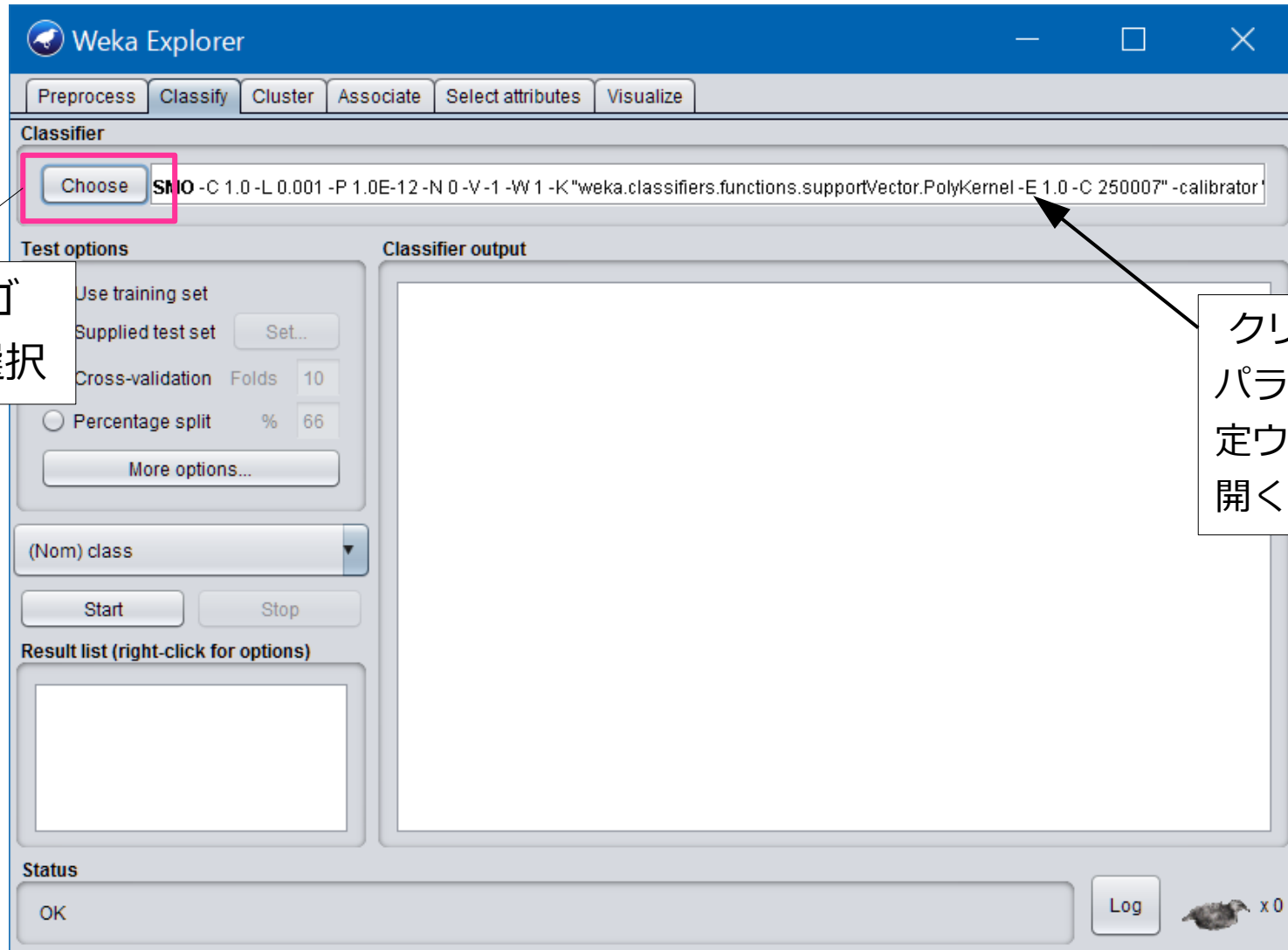
- 1つのグラフのみ表示



x 軸、y 軸、色の  
基準が選べる

クリックで  
色を変更可能

# 識別器の学習 (Classify)

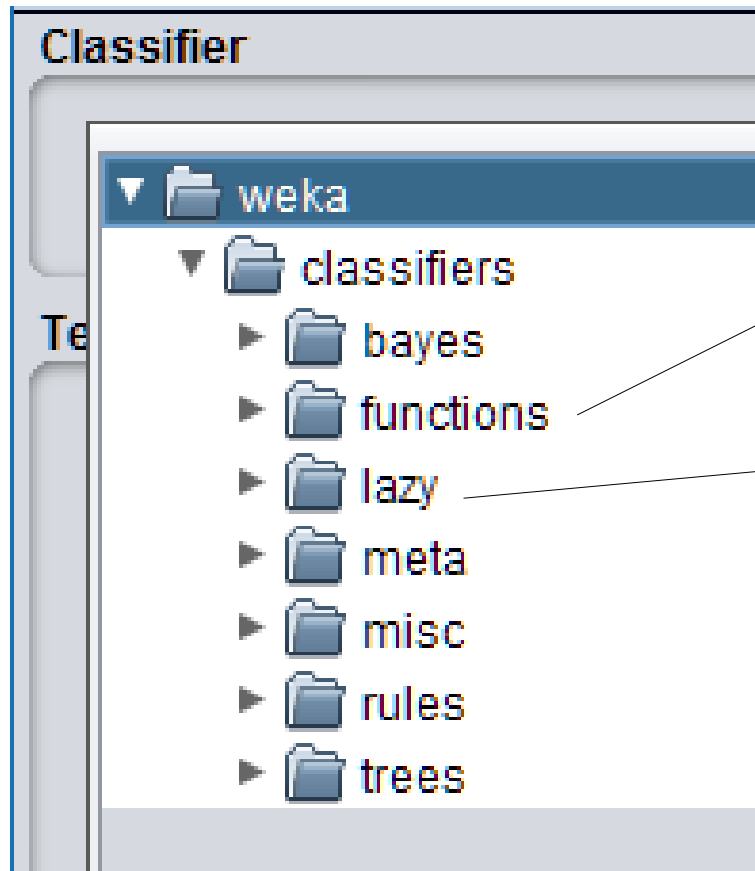


識別アルゴ  
リズムの選択

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

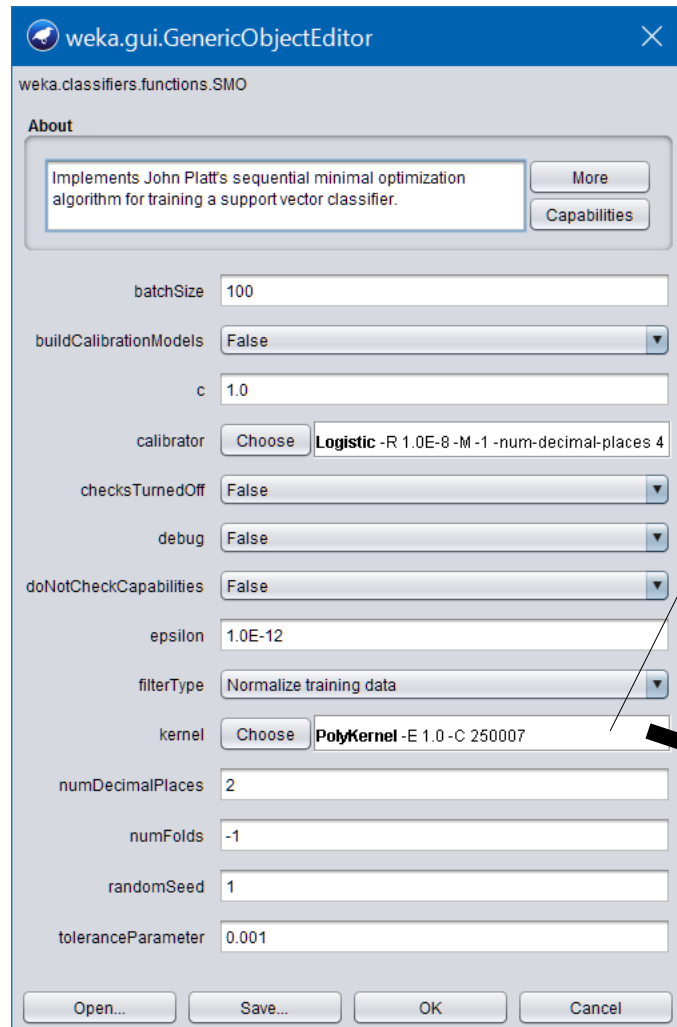
The screenshot shows the 'weka.gui.GenericObjectEditor' window for the 'weka.classifiers.lazy.IBk' classifier. The 'About' section describes it as a 'K-nearest neighbours classifier.' with 'More' and 'Capabilities' buttons. The main configuration area includes the following parameters:

- KNN**: 1 (indicated by a callout 'k')
- batchSize**: 100
- crossValidate**: False
- debug**: False
- distanceWeighting**: No distance weighting (indicated by a callout '距離による重み付けの有無')
- 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'.

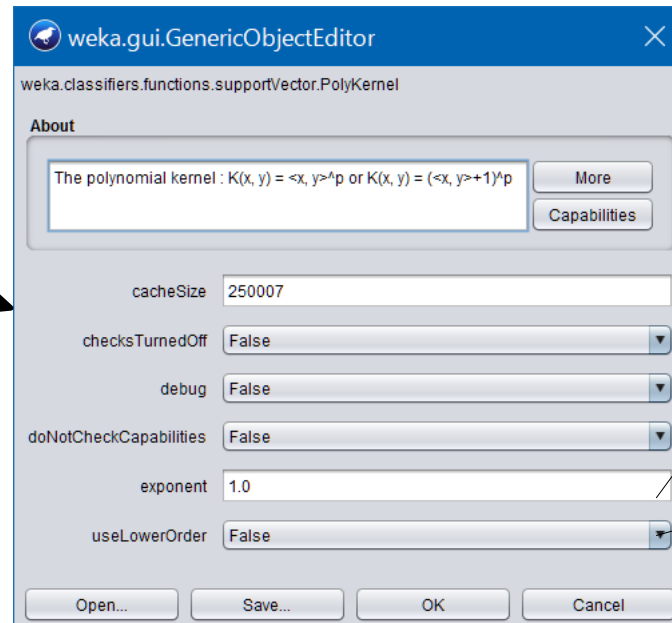
# 識別器の学習 (Classify)

- 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

☒ Cross-validation Folds

☐ Percentage split %

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

正解率

93.3333 %  
6.6667 %

縦方向が正解、横方向が予測  
対角成分が正解数

# 結果の可視化

- 学習したモデル
  - 式、木構造、ネットワークの重み、 etc.
- 性能
  - 正解率、精度、再現率、 F 値
  - グラフ
    - パラメータを変えたときの性能の変化
    - 異なるモデルの性能比較

# 結果の可視化

- 混同行列

	予測+	予測-
正解+	true positive(TP)	false negative(FN)
正解-	falsepositive(FP)	true negative(TN)

- 正解率  $Accuracy = \frac{TP + TN}{TP + FN + FP + TN}$

- 精度  $Precision = \frac{TP}{TP + FP}$

- 再現率  $Recall = \frac{TP}{TP + FN}$

- F 値  $F-measure = 2 \times \frac{Precision \times Recall}{Precision + Recall}$

正解の割合  
クラスの出現率に  
偏りがある場合は不適

正例の判定が  
正しい割合

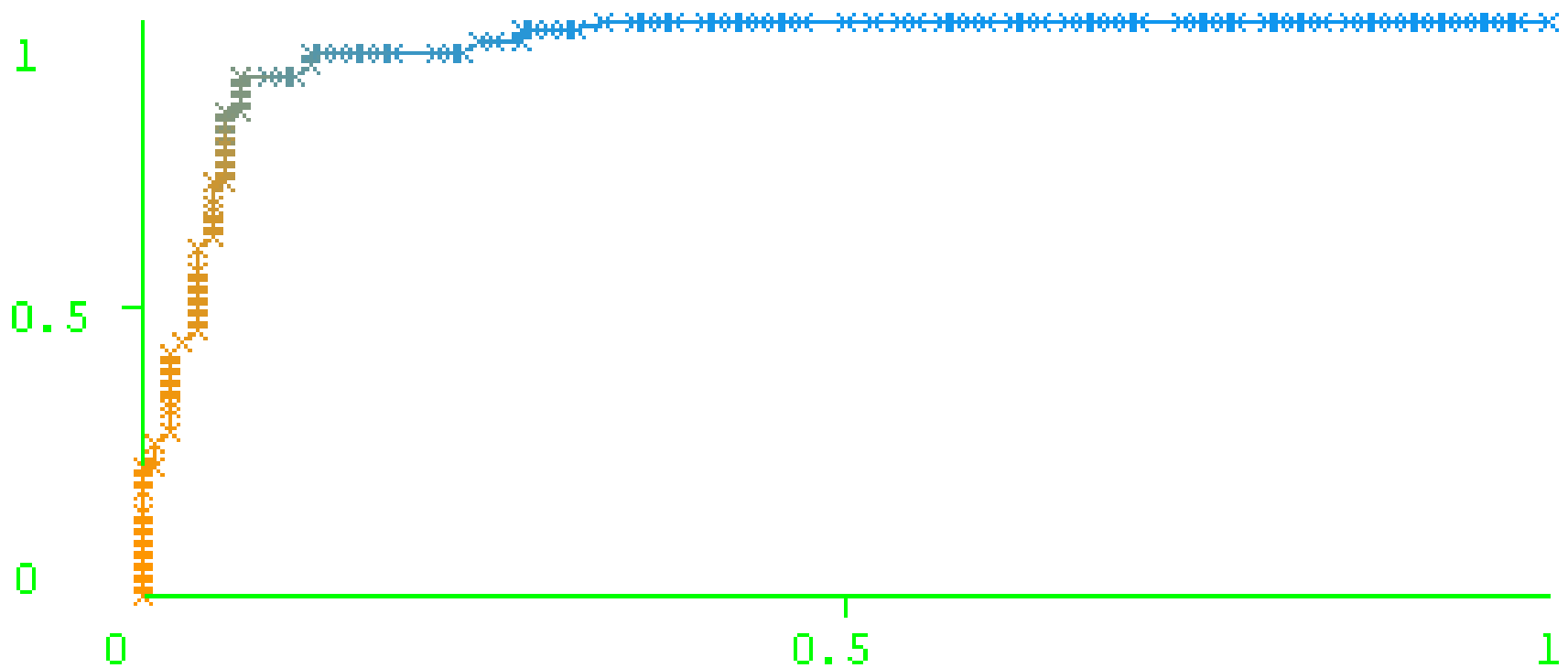
正しく判定された  
正例の割合

トレードオフ

精度と再現率の  
調和平均

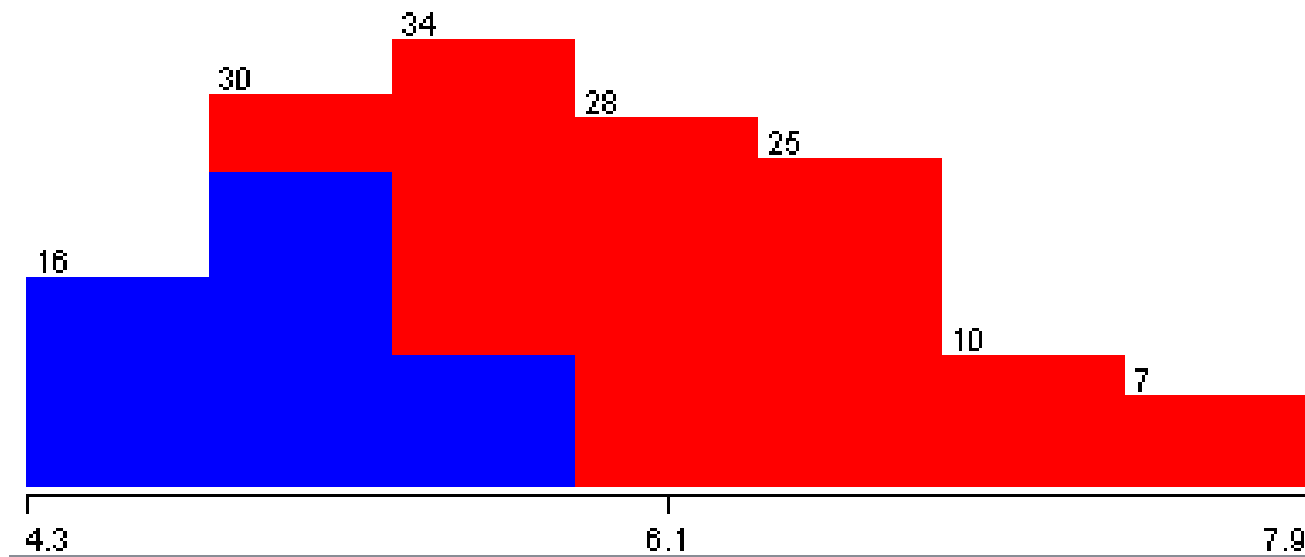
# 結果の可視化

- 精度と再現率のトレードオフ
  - ROC 曲線



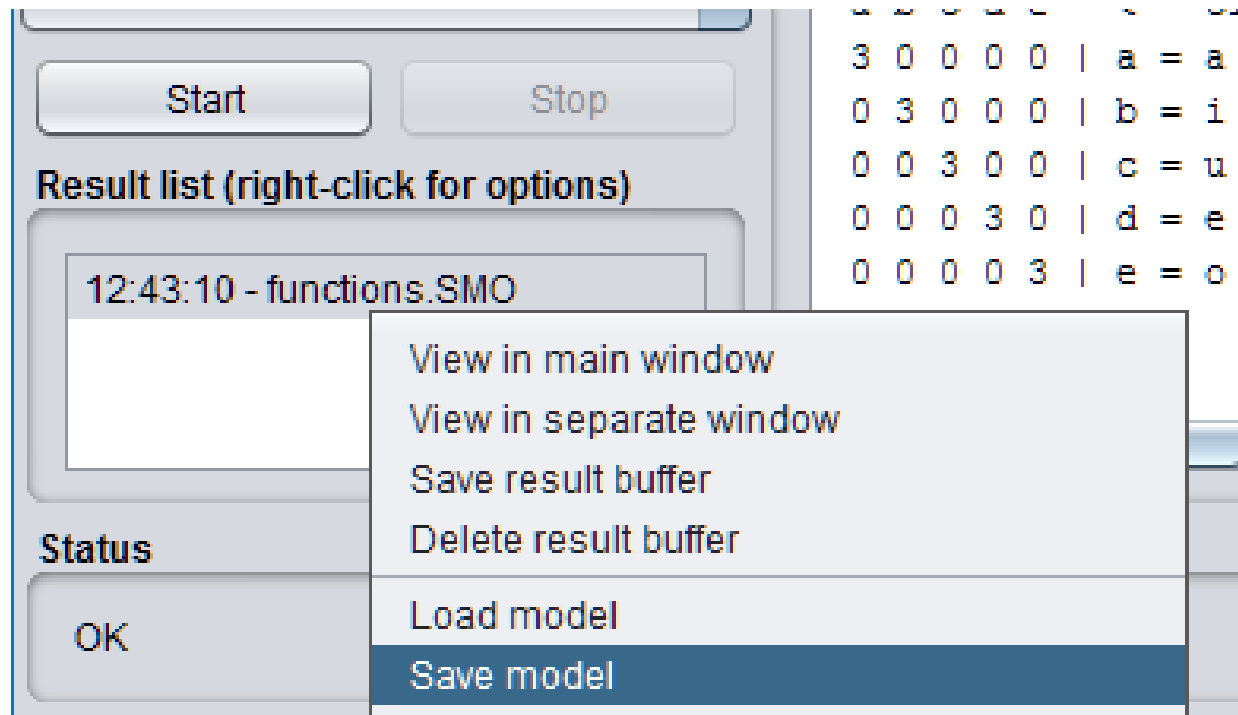
# 結果の可視化

- 識別のための閾値の設定
  - sepallength 特徴による Iris-setosa の識別



# 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



# 識別面の可視化 (Boundary Visualizer)

- GUI Chooser→Visualization から呼び出す

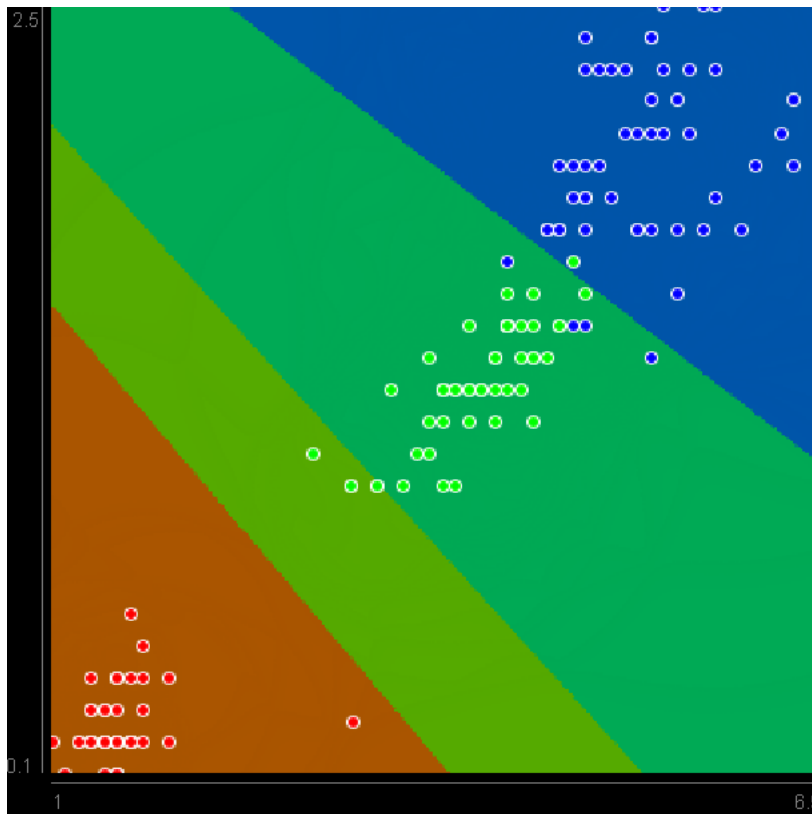
The screenshot shows the BoundaryVisualizer application window. It has several sections: 'About' at the top, 'Dataset' and 'Classifier' sections in the middle, 'Class Attribute' and 'Visualization Attribute' sections below them, 'Class color' section, a large plot area, and control sections on the right. Annotations with Japanese text point to specific parts of the GUI:

- データの指定** (Data specification) points to the 'Open File' button in the Dataset section.
- 識別器の指定** (Classifier specification) points to the 'Choose' button in the Classifier section.
- 表示する特徴の指定** (Specification of features to display) points to the 'X: petallength (Num)' and 'Y: petalwidth (Num)' dropdowns in the Visualization Attribute section.
- 実行** (Execute) points to the 'Start' button in the Plotting section.

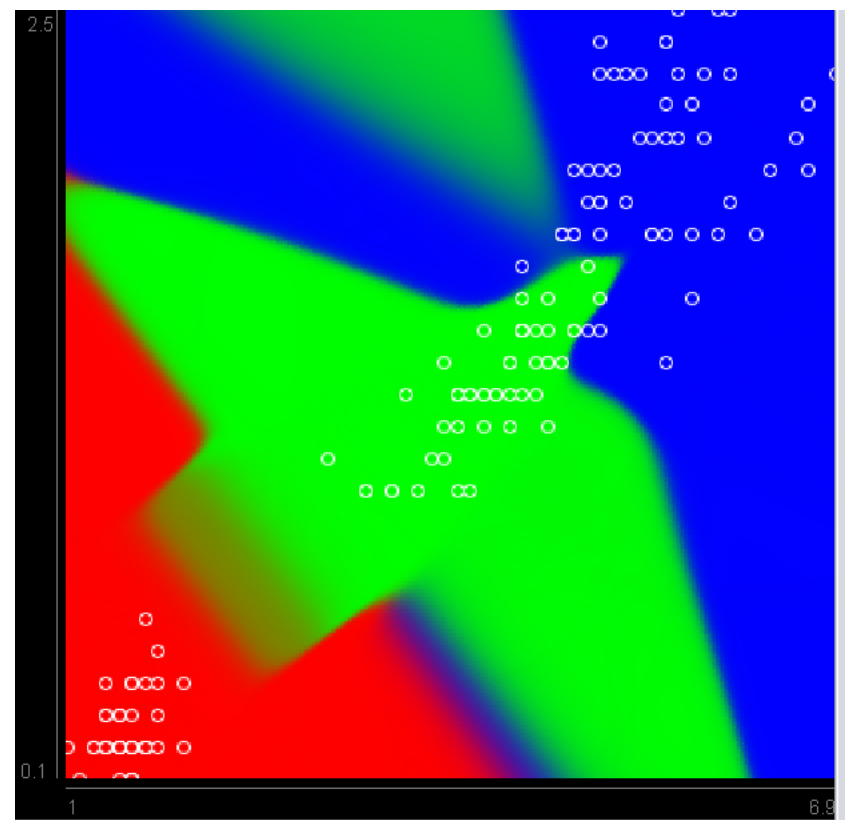
The plot area displays a 2D scatter plot with decision boundaries for three classes: Iris-setosa (red), Iris-versicolor (green), and Iris-virginica (blue). The axes are labeled with values like 0.1, 1, 2.5, and 6.8.

# 識別面の可視化 (Boundary Visualizer)

- iris2D データの識別面



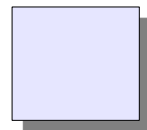
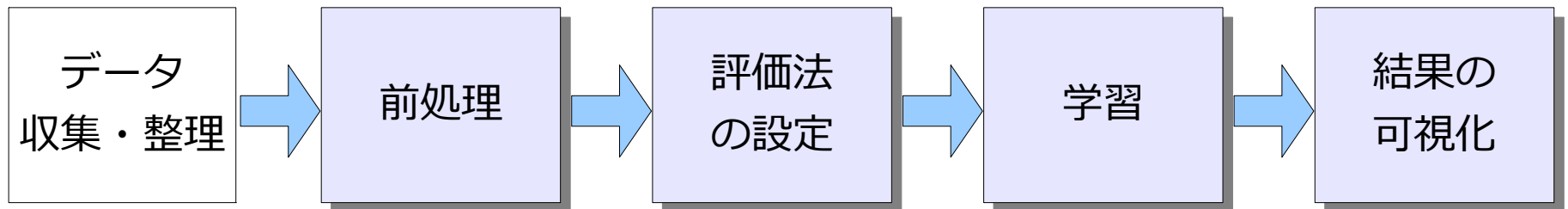
SMO Polykernel 1 次



Multilayerperceptron  
hiddenLayers=20  
trainingTime=50000



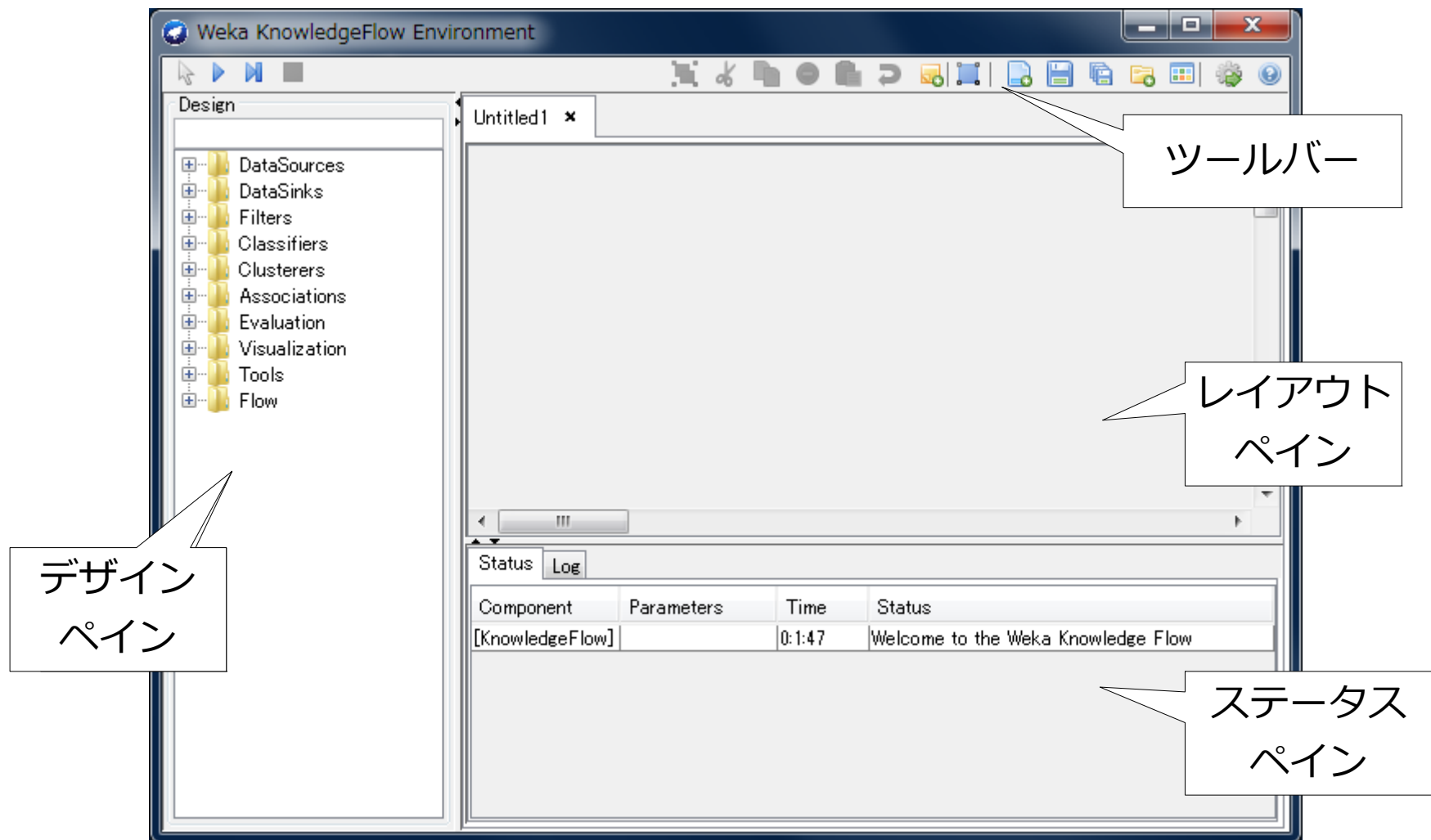
# KnowledgeFlow による自動化



: Weka による支援が可能

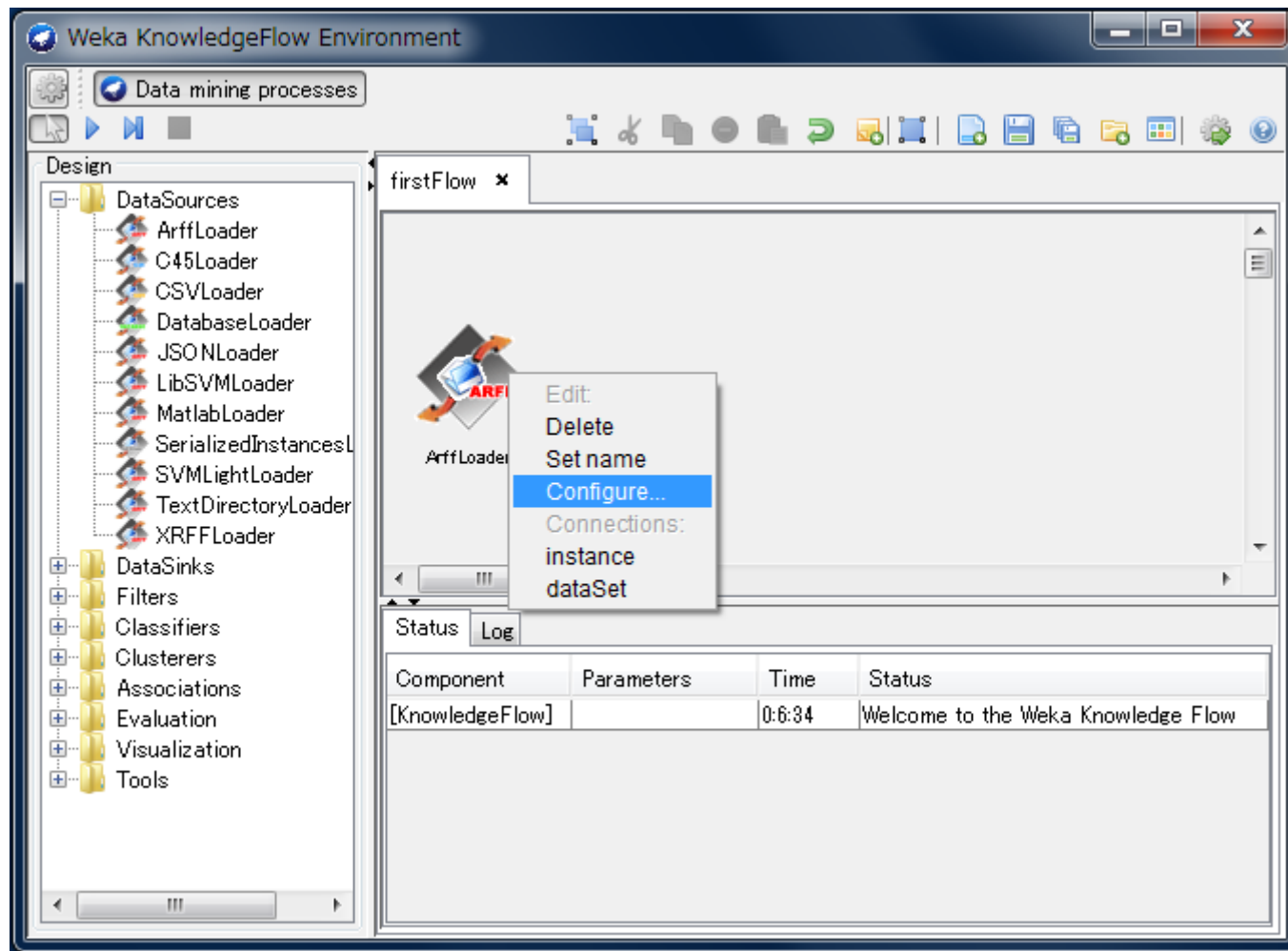
# KnowledgeFlow による自動化

- KnowledgeFlow インタフェースの構成



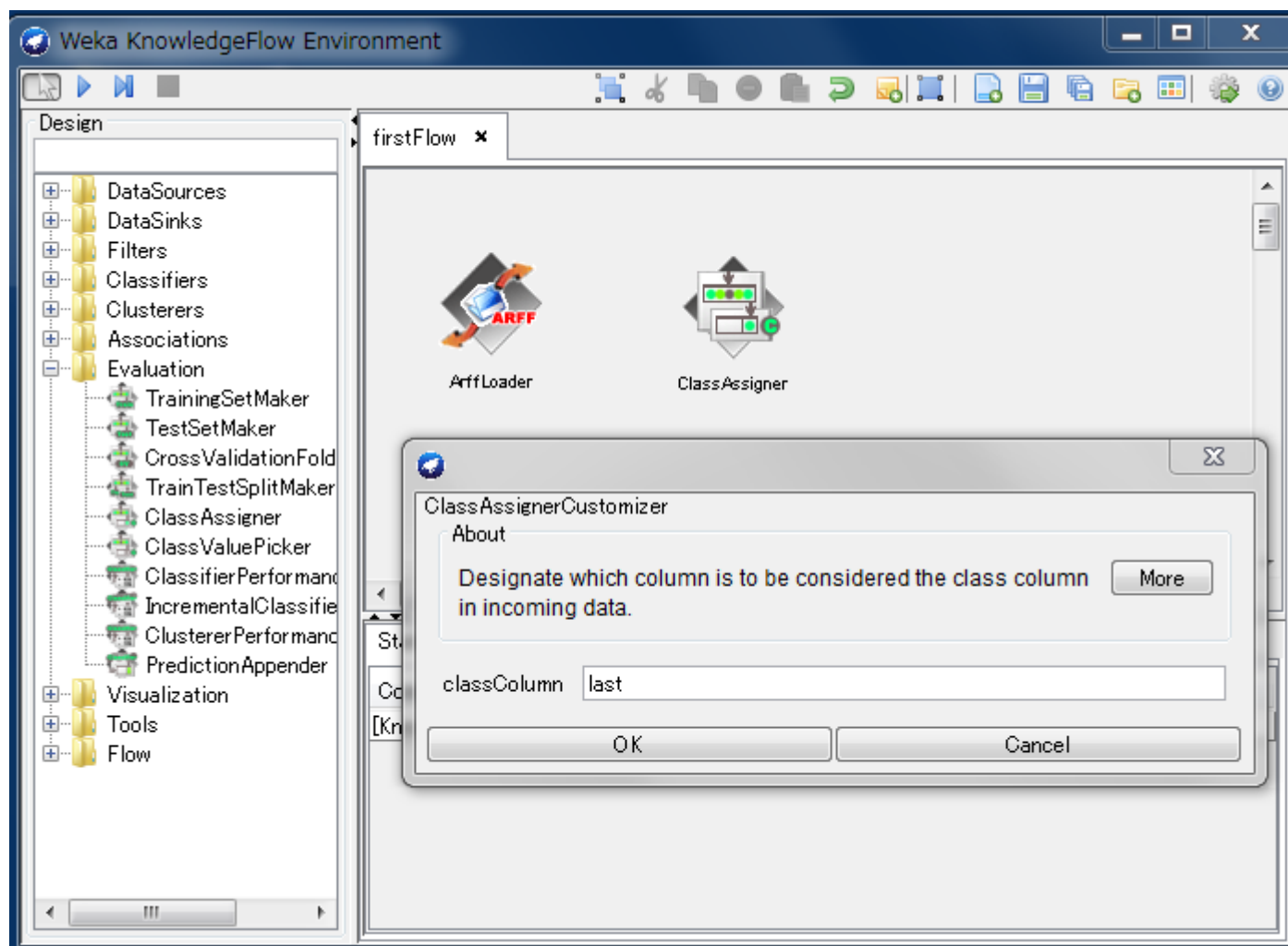
# 前処理

- ArffLoader の配置と設定



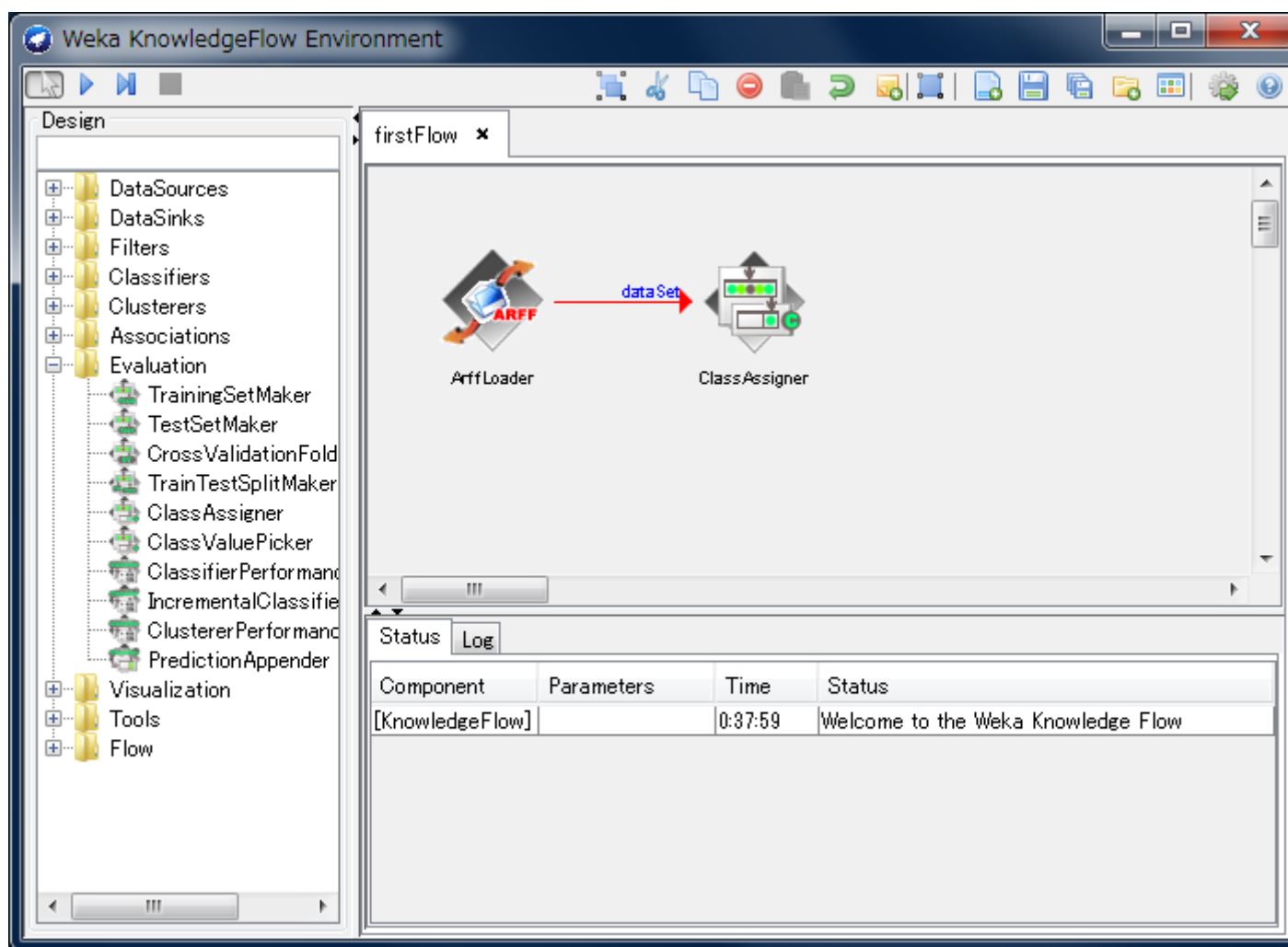
# 前処理

- ClassAssigner の配置と設定



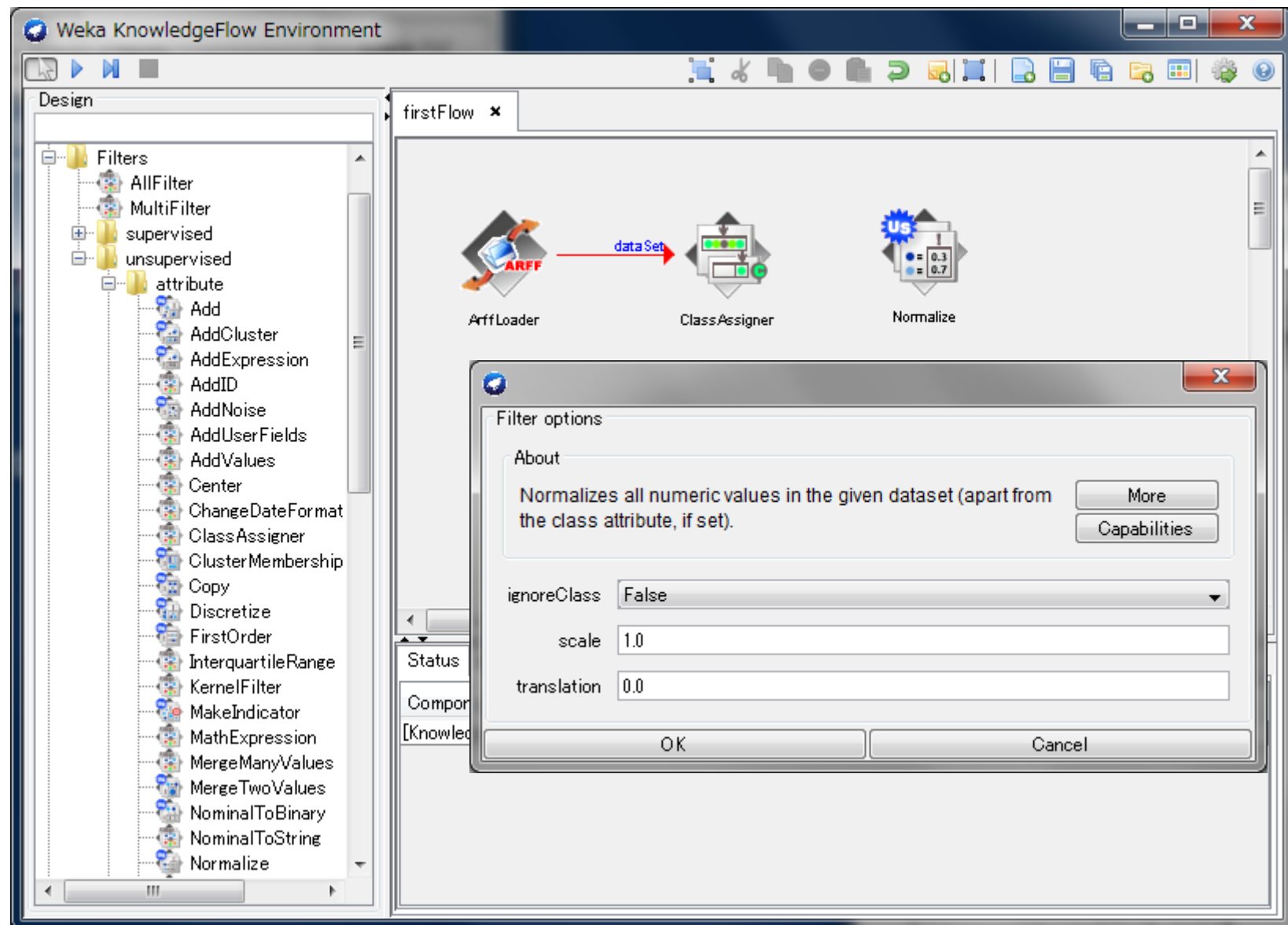
# 前処理

- 部品の結合
  - 受け渡す情報に気をつける

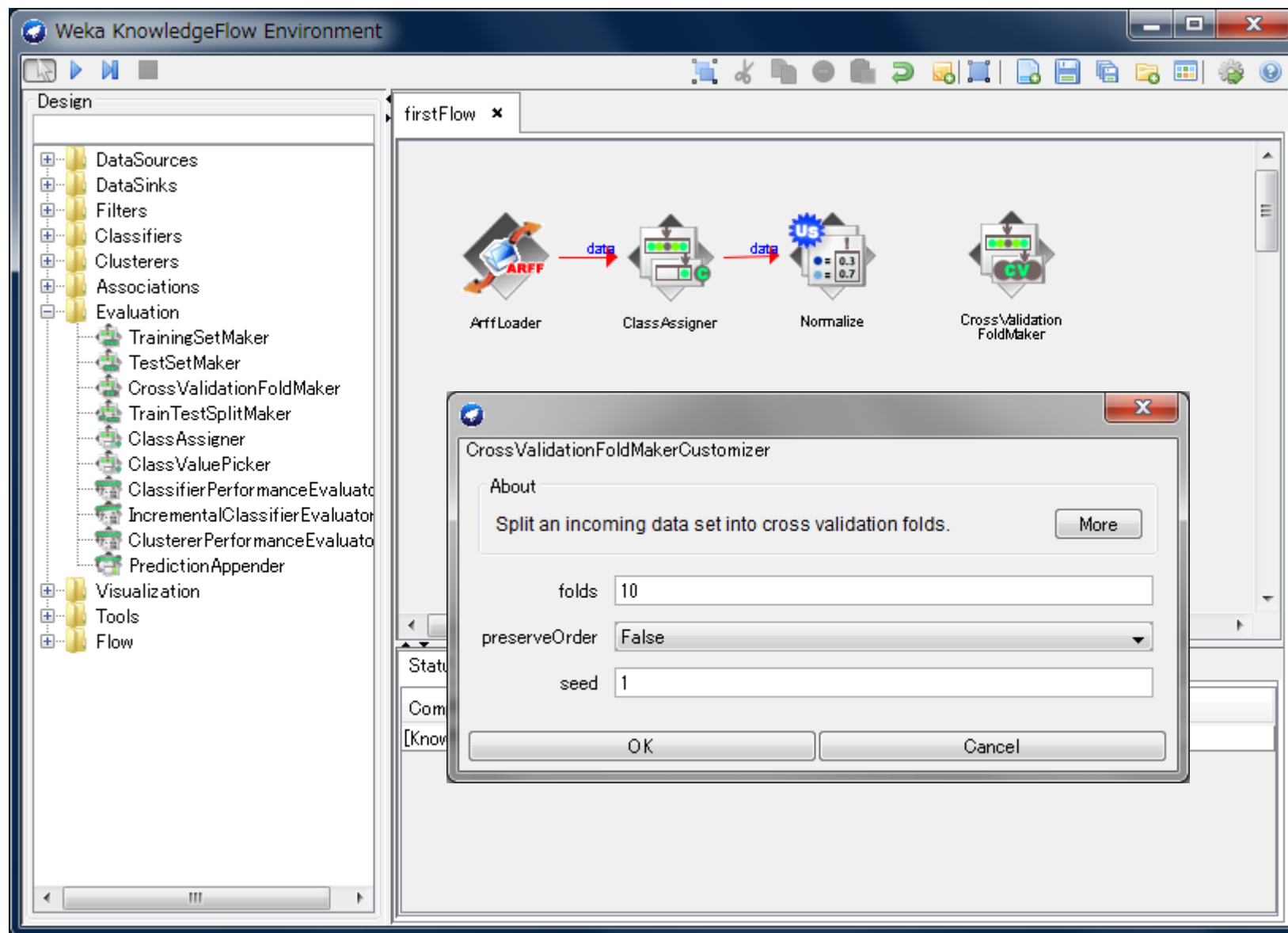


# 前処理

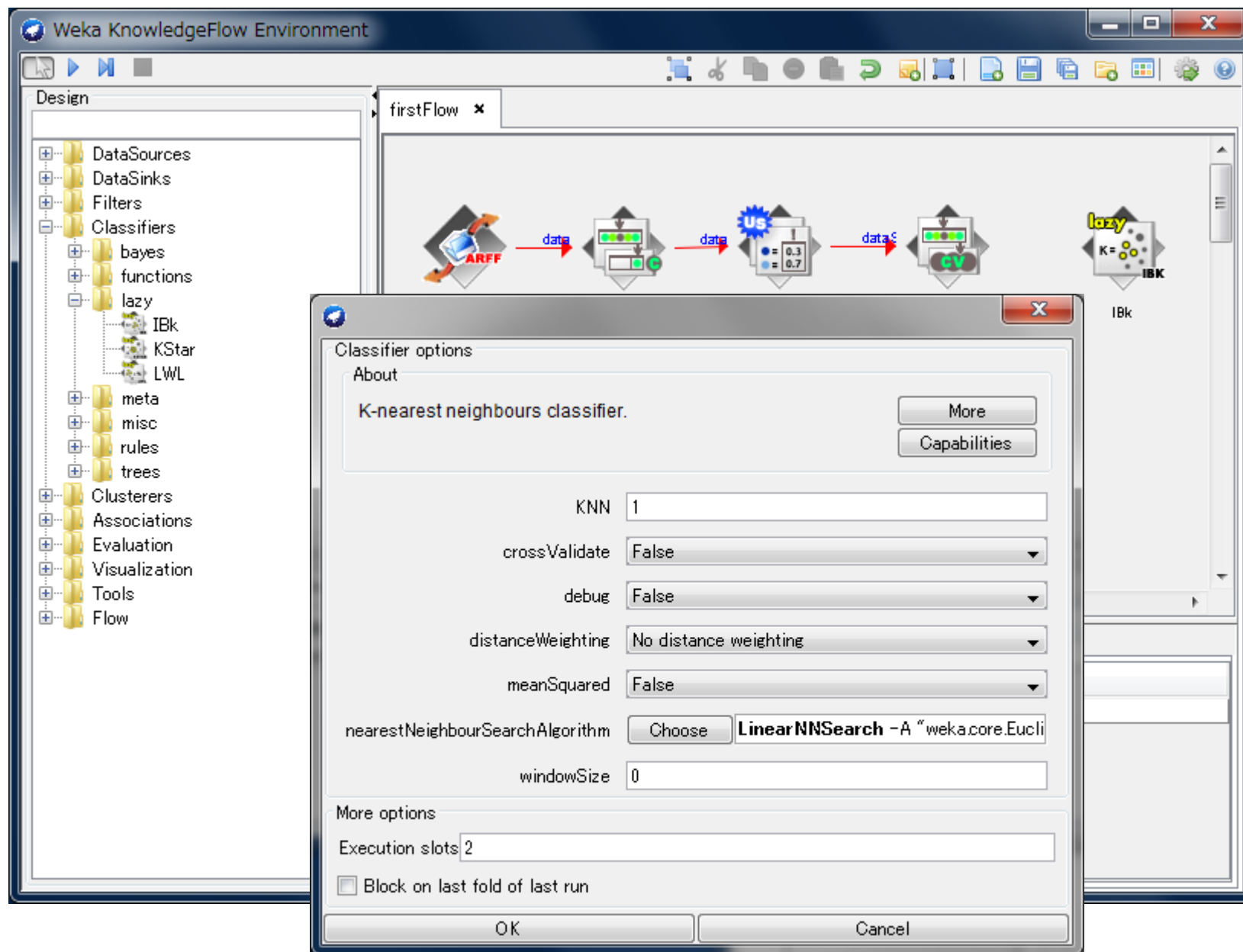
- Normalize の配置と設定



# 評価基準の設定



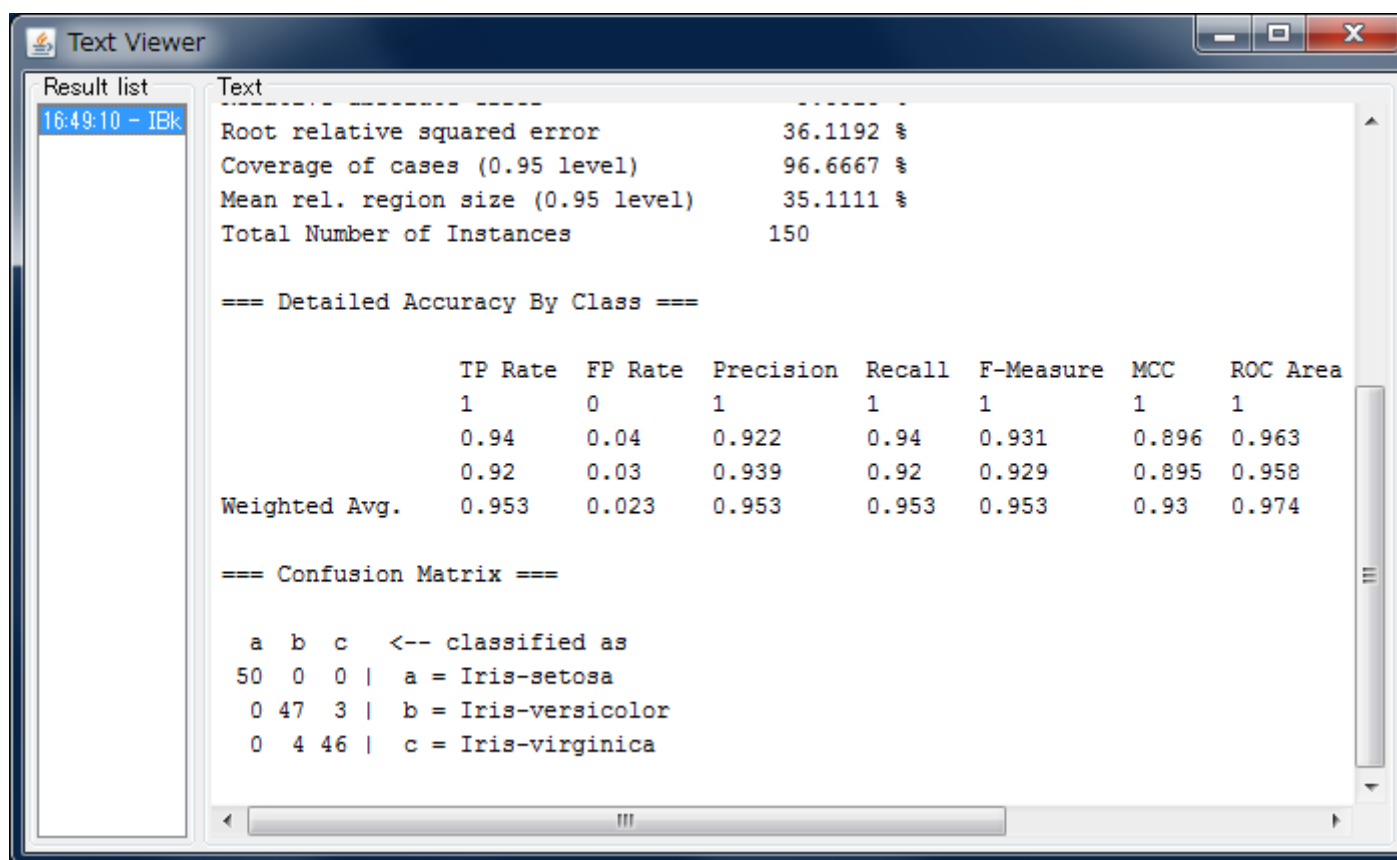
# 學習





# 結果の可視化

- 結果の表示



The screenshot shows a 'Text Viewer' window with a 'Result list' on the left and a 'Text' area on the right. The 'Result list' contains a single entry: '16:49:10 - IBk'. The 'Text' area displays the following output:

```
-----
Root relative squared error      36.1192 %
Coverage of cases (0.95 level)   96.6667 %
Mean rel. region size (0.95 level) 35.1111 %
Total Number of Instances        150

=== Detailed Accuracy By Class ===
```

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area
1	1	0	1	1	1	1	1
	0.94	0.04	0.922	0.94	0.931	0.896	0.963
	0.92	0.03	0.939	0.92	0.929	0.895	0.958
Weighted Avg.	0.953	0.023	0.953	0.953	0.953	0.93	0.974

```
=== Confusion Matrix ===

 a  b  c  <-- classified as
50  0  0 | a = Iris-setosa
 0 47  3 | b = Iris-versicolor
 0  4 46 | c = Iris-virginica
```

# 結果の可視化

- 作成したプロセス

