

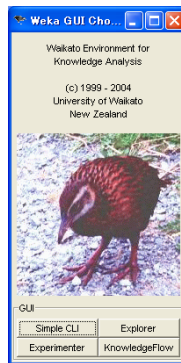
WEKA の Experiment と KnowledgeFlow 環境

1. Experiment の環境

先月号では、複数のデータセットに複数のデータマイニングの方法を用いて比較分析を行った。WEKA の Explorer 環境では、このような作業は、各々のデータに一つひとつの方法を対応させるので効率が悪い。WEKA では、複数のデータと方法をまとめて処理する Experimenter 環境を提供している。その操作の手順を次に説明する。

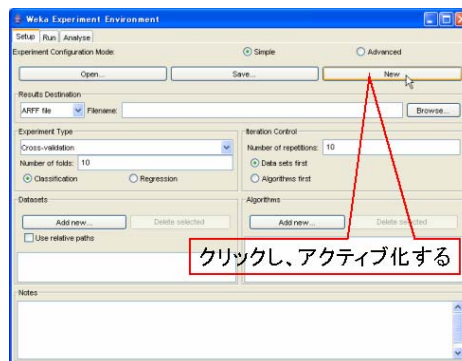
- ① WEKA の GUI の [Experimenter] ボタンを押し、Experiment 環境パネルを開く。

図 1 WEKA の GUI



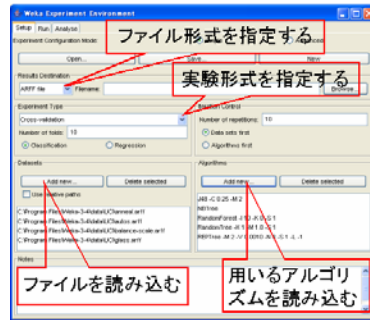
- ② Experiment Environment パネルの [New] ボタンを押す。

図 2 Experiment Environment パネル



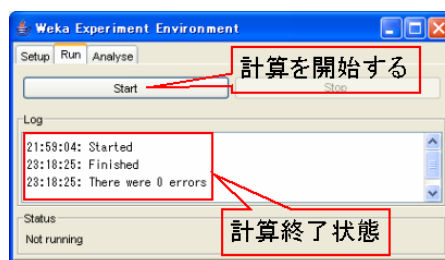
- ③ 条件を設定し、データファイルと分析手法を読み込む。

図 3 複数のデータと方法を読み込んだ画面



- ④ Run タブをアクティブ化し、[Start]ボタンを押し、計算を開始させる。

図 4 計算が終了した画面



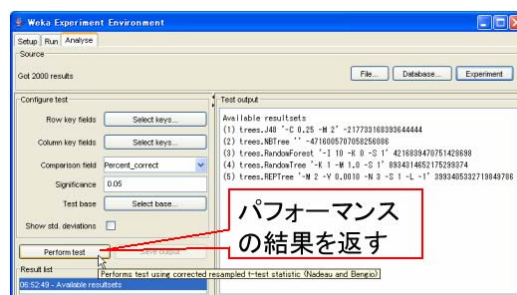
- ⑤ 計算が終了したら、Analyse タブをアクティブ化し、[Experiment]ボタンをクリックする。

図 5 Analyse タブの画面



- ⑥ [Perform test]ボタンを押すと Test output のウィンドウに結果が返される。

図 6 実行結果の画面



さらに、図7のような操作を行うと、図8のような独立したパフォーマンスの結果を返すウィンドウが開かれる。

図7 パフォーマンスの結果画面

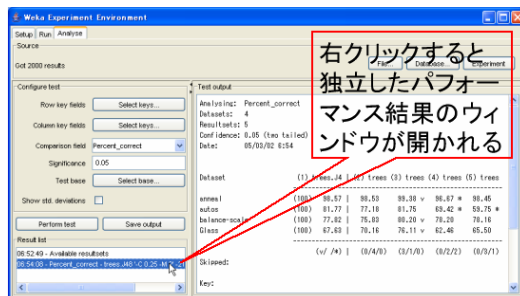
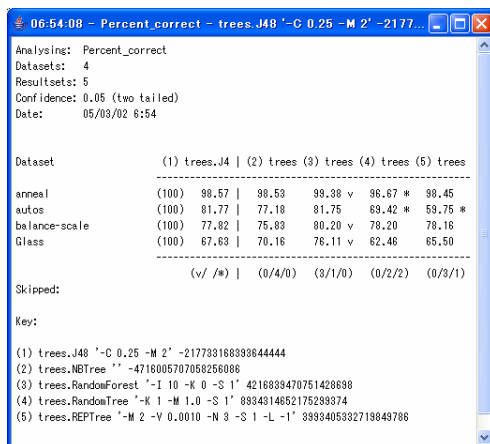


図8 独立したパフォーマンスの結果画面



Test output のウィンドウに返された結果は、[Save output]ボタンを用いて、保存することができる。

2. KnowledgeFlow の環境と機能

(1) KnowledgeFlow とは

WEKA の GUI(図1)には、KnowledgeFlow というボタンが設けられている。KnowledgeFlow は、データの処理システムを、コンポーネント(component、構成要素・部品)アイコンを組み合わせる自由な構築するグラフィカル環境である。

KnowledgeFlow は発展途上であり、WEKA の”classifiers”と”filters”のすべての機能が実装されているが、”clustering”の機能の実装は若干遅れている。その一方、Explorer にない機能を持っている。

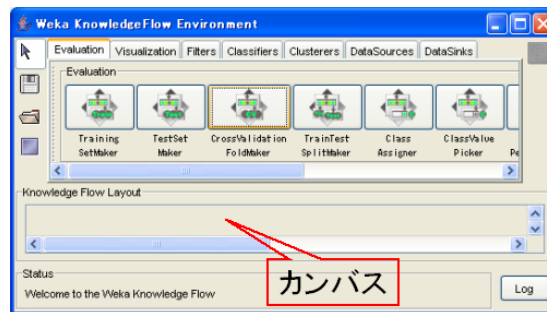
WEKA の Explorer では、データのバッチ処理しかできないが、KnowledgeFlow ではバッチ処理と段階的な処理を行うことができる。KnowledgeFlow は次のような特徴を持っている。

- ◇ データの処理の流れが直感的
- ◇ データの段階的処理とバッチ処理
- ◇ データの平行処理
- ◇ filters の連結
- ◇ 交差確認の結果の視覚化
- ◇ 段階的な処理の視覚化

(2) KnowledgeFlow の起動と基本操作

WEKA の GUI の [KnowledgeFlow] ボタンを押すと図 9 のような KnowledgeFlow 環境のパネルが開かれる。

図 9 KnowledgeFlow 環境パネル画面



KnowledgeFlow 環境パネルには、複数のタブ(Evaluation、Visualization、Filters、Classifiers、Clusterers、DataSources、DataSinks)があり、各タブをクリックするとパネルにコンポーネントアイコンが現れる。その中の一つひとつのアイコンがデータ処理およびマイニングシステムを構築する部品である。

パネルの下部に KnowledgeFlow Layout というキャンバス(canvas)がある。使用者は、コンポーネントアイコンをキャンバス上で連結させ、データ処理とマイニングのシステムを構築する。

データの処理とマイニングを行うためには、まずデータを読み込まなければならない。

データの読み込みは、まず DataSources タブをアクティブ化する。DataSources には、WEKA が扱っているデータファイル形式のコンポーネントがある。コンポーネントのアイコンを左クリックするとコンポーネントが選定される。コンポーネントが選定された状態で、マウスポインタをキャンバスに移し、左クリックするとコンポーネントがキャンバスに取り込まれる(図 10)。取り込まれたアイコンは、マウスの左ボタンで自由に配置位置を換えることができる。アイコンを右クリックするとアイコンを操作するメニューが開かれる(図 11)。操作メニューの項目を表 1 に示す。

図 10 コンポーネントを選定した画面

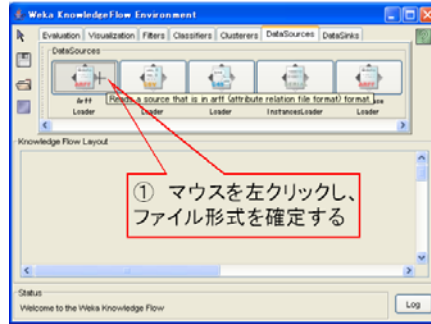


図 11 カンバスにコンポーネントを取り込んだ画面

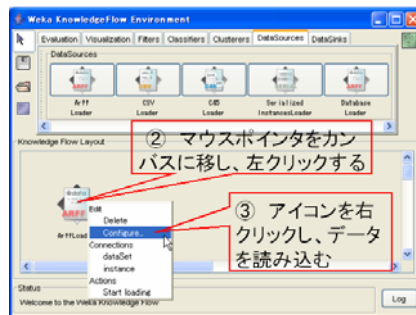


表 1 data アイコンの操作メニュー

Edit	
Delete	アイコンの削除
Configure	データを読み込むなど
Connections	
dataSet	アイコン連結と切断
instance	アイコン連結と切断
Actions	
Start loading	データを流す

データは、data アイコンの操作メニューの **Configure** を左クリックし、データが置かれているフォルダを開きデータファイルをクリックすることで読み込まれる。

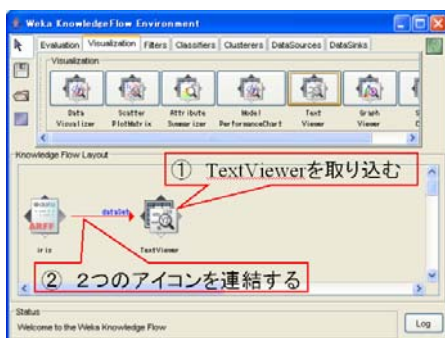
上記の操作で、データが正しく読み込まれているかを確認するため、**Visualization** タブをアクティブ化し、テキストの表示コンポーネント **TextViewer** をカンバスに取り込み、図 12 のようにコンポーネントアイコンを連結する。

アイコンの連結は、基本的にはデータの流れの先後の順に行う。アイコンを右クリックするとアイコンの操作メニューが開かれる。メニューの **Connection** の中の **dataSet** を選択し、マウスのポインタを連結すべきアイコンまで引き、つなぎ合うことでアイコン同士が連結される。連結の切断は、アイコン操作のメニューを開き、連結されている項目を左クリックする。

アイコンの連結操作が終わったら、data アイコンの操作メニューを開き、**Actions** 下の **Start loading** を左クリックするとデータが流される。**TextViewer** アイコンの操作メニューを開き、

Show result を左クリックすると、読み込んだデータを返すテキストウィンドウが開かれる。

図 12 2つのアイコンを連結した画面



読み込んだデータは、図 13 のように Visualization タブの中の DataVisualizer、ScatterPlotMatrix、AttributeSummarizer のコンポーネントを用いて、散布図、対散布図、ヒストグラムを作成することができる。

図 13 作図アイコンを連結した画面

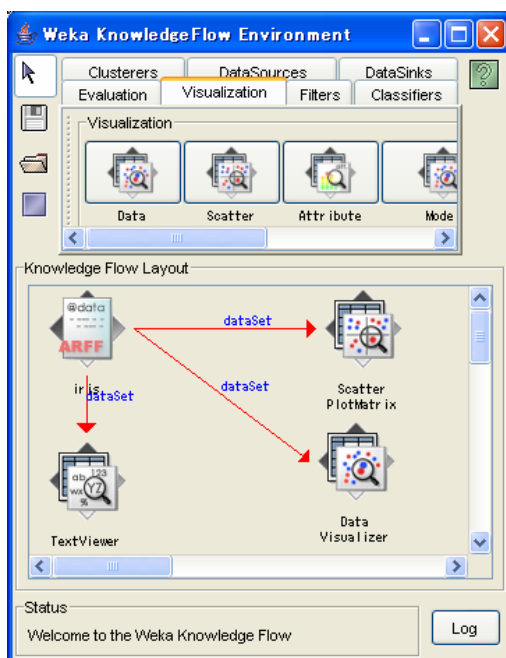
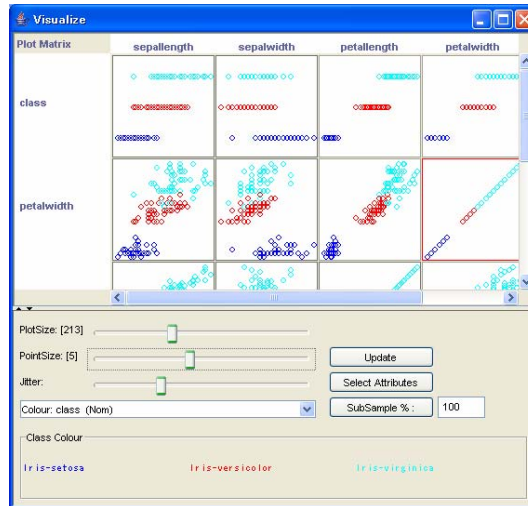


図 13 のようにアイコン同士の連結が終わったら、data アイコンのメニューの Start loading をクリックし、作図アイコンのメニューの Show plot をクリックすると図が返される。例として、ScatterPlotMatrix による対散布図の画面を図 14 に示す。散布図のマークのサイズや色などは開かれている Visualize のパネルで調整することができる。

図 14 対散布図のパネル画面



(2) タブの機能

① Evaluation タブ

Evaluation タブは、データセットの分割と作成 (TrainingSetMaker、TestSetMaker、CrossValidationFoldMaker、TrainTestSplitMaker、ClassValuePicker、ClassValuePiker)、結果の表示など (ClassifierPerformanceEvaluator、IncrementalClassifierEvaluator、PredictionAppender) に関するコンポーネントにより構成されている。

② Visualization タブ

Visualization タブには、データの表示 (TextViewer)、データの図示 (DataVisualizer、ScatterPlotMatrix、AttributeSummatizer)、解析結果の図示 (GraphViwer、StripChart) などのコンポーネントが実装されている。

③ Filters タブと Classifiers タブ

WEKA の Explorer の “filters” と “classifiers” のすべて機能のコンポーネントが実装されている。

④ Clusterers タブ

Explorer の “Clusterers” の一部のコンポーネントが実装されている。

⑤ DataSources タブ

Explorer で扱えるすべてのデータ形式のコンポーネントが実装されている。

3. KnowledgeFlow の構築の例

通常 of データ処理及びマイニングを行う際のデータの流は、データの読み込み

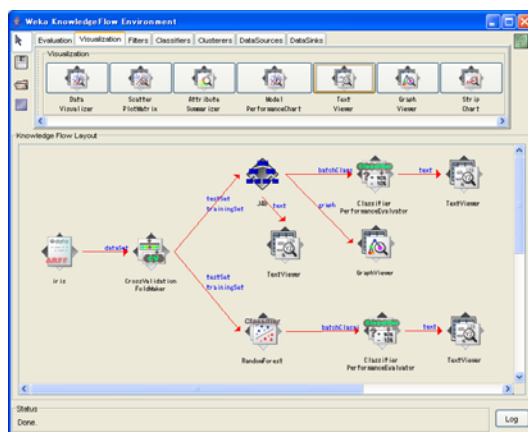
(DataSources)→データの処理(Filters)→データの処理(Classifiers)→結果の表示と図示(Evaluation)の順になる。次に決定木分析を例とした KnowledgeFlow の構築の例を示す。

例：iris データを読み込み、交差確認法による決定木(J48)、RandomForest の分析とその結果を図示するシステムを構築する。

手順：

- ① DataSources タブをアクティブ化し、ArffLoader コンポーネントをキャンバスに取り込む。
- ② 取り込んだ ArffLoader アイコンの操作メニューの Configure を左クリックし、データが置かれているフォルダを開き、iris.arff を読み込む。
- ③ Evaluation タブをクリックし、交差確認法のコンポーネント CrossValidation FoldMaker と、分類の精度に関するパフォーマンス環境のコンポーネント Classifier PerformanceEvaluation をキャンバスに取り込む。キャンバス上の Cross ValidationFoldMaker アイコンの操作メニューの Configure をクリックし、n 重交差確認の n を指定することができる。デフォルトでは n=10 になっている。
- ④ Classifiers タブをアクティブ化し、J48、RandomForest をキャンバスに取り込む。
- ⑤ 結果の観測のため、Visualization のタブをアクティブ化し、TextViewer と GraphViewer をキャンバスに取り組み、図 15 のようにアイコンを連結する。

図 15 アイコンを連結した画面



上記の作業が終わったら、ArffLoader の操作メニューの Start Loading を左クリックすると解析が始まり、計算が終了次第、その結果が最終端末のコンポーネントに記録される。

キャンバス上の J48 と連結されている Text Viewer の操作メニューの Show results をクリックすると J48 の結果のテキストパネルが開かれる。ここでは n 重交差確認(n=10)を行ったので、決定木が 10 個作成され、パネルの左の Result list にリストアップされている。その中の任意の 1 行をクリックするとそれに対応する結果が右の Text ウィンドウに表示される(図 16)。

GraphViewer のメニューの Show results をクリックすると Graph list のパネルが開かれる。さらに Graph list 中の項目をクリックすると、図 18 のような Tree View ウィンドウに決定木のグラフが返される。

アイコン ClassifierPerformanceEvaluation と連結されている TextViewer アイコンの操作メニューの Show results をクリックすると、パフォーマンスの結果のウィンドウが開かれる(図 18)。

図 17 決定木のテキスト結果画面

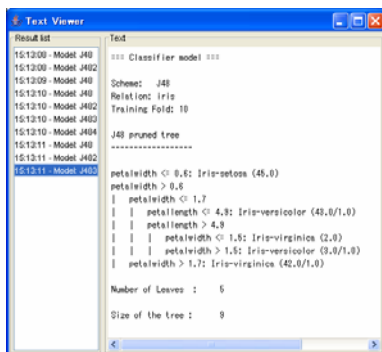


図 18 決定木の画面

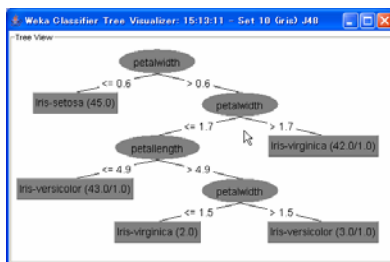


図 19 パフォーマンスのテキスト画面

Correctly Classified Instances	148	95.3333 %
Incorrectly Classified Instances	7	4.6667 %
Kappa statistic	0.93	
Mean absolute error	0.0395	
Root mean squared error	0.1716	
Relative absolute error	0.4592 %	
Root relative squared error	36.2119 %	
Total Number of Instances	150	

構築したデータマイニングの KnowledgeFlow は、KnowledgeFlow Environment パネルの左上のフロッピーアイコンを用いて保存することができる。