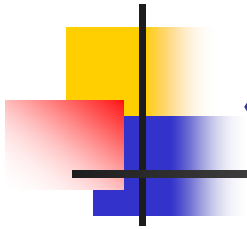
機械学習統合環境 Wekaと ベンチマークデータ

山本章博
情報学研究科 知能情報学専攻
(工学部 情報学科)
2012年12月25日



機械学習とプログラミング

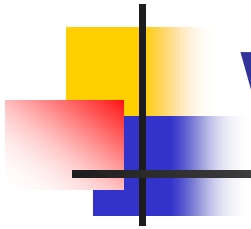
各種の機械学習手法をデータに適用するには

- プログラムを自作する
  - 一般的な言語を用いてすべてを自作
  - ライブラリ群の利用

    dlib C++ library, ELKI,...
- 機械学習向けプログラミング言語
  - R, S, ...
- 機械学習向け統合環境
  - Weka

Wekaとは

- フリーの機械学習統合環境
  - 研究，教育，実務など幅広い利用が可能
  - データの前処理，多様な学習アルゴリズム，評価手法の理解ができる.
  - データの可視化などのGUIを備えている.
- New Zealand の Waikato大学で開発
- 実装はマルチプラットホーム対応のPure Java
  - 配布パッケージは Weindos/Mac OS X/JAR
  - ...なので，速度は期待できない.
- 参考書: I.H. Witten and E. Frank : Data Mining -Third Edition- Morgan-Kaufmann



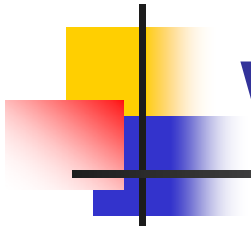
Wekaの配布元

Waikato大学 Weka HP

- http://www.cs.waikato.ac.nz/ml/weka/
  - 配布はsourefoge経由

日本語情報プロジェクト weka-jp

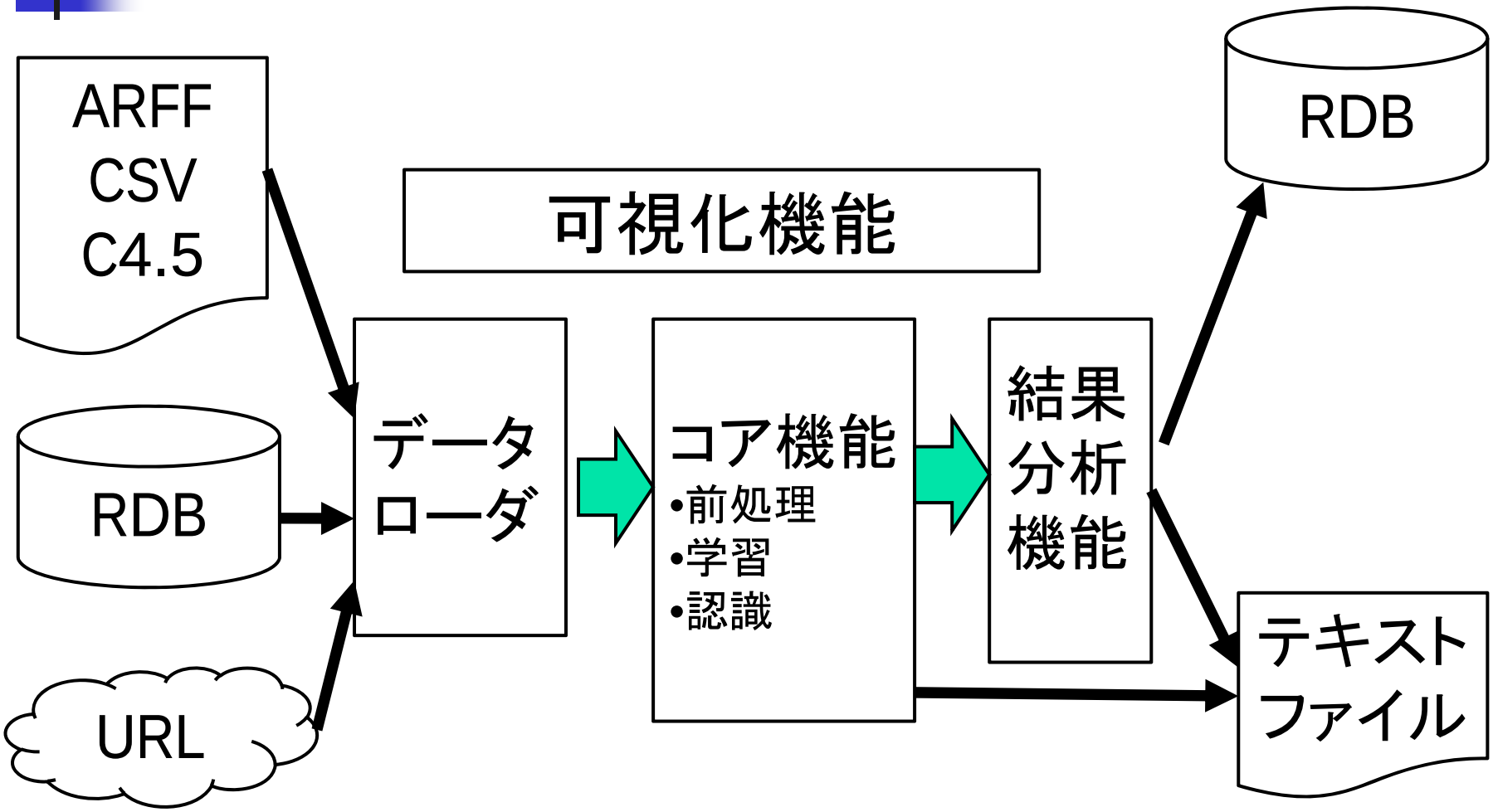
- http://www.weka-jp.info/
  - 文教大学講師 阿部秀尚氏が管理



Wekaのダウンロードとインストール

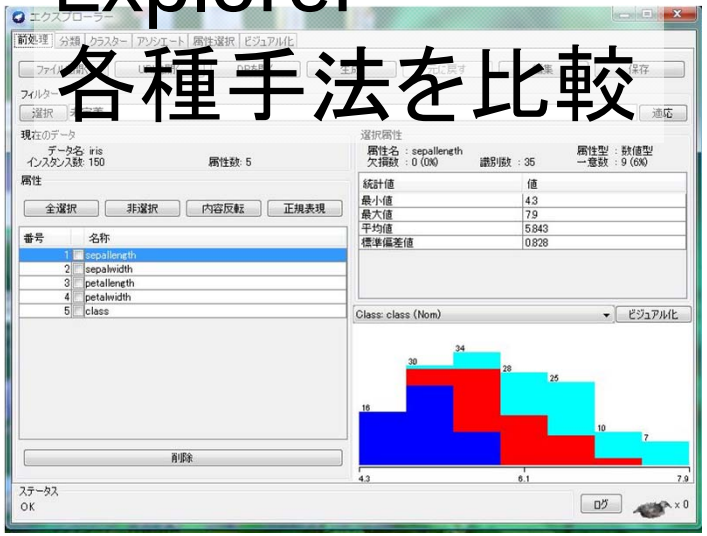
- Weka HPからダウンロード
  - JRE(Sun Java VM 1.5)をPCにインストールしていない場合は，JRE付を選択する
- ウィザードに従ってインストール
- 起動

Wekaの構成


ARFF
CSV
C4.5
RDB
URL
可視化機能
データ
ローダ
コア機能
•前処理
•学習
•認識
結果
分析
機能
RDB
テキスト
ファイル

Wekaのインタフェース

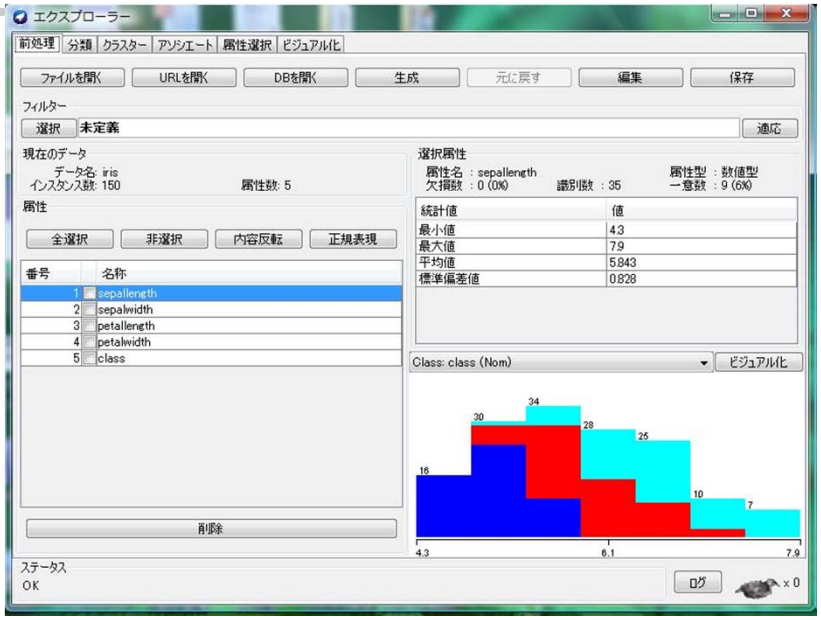
Explorer
各種手法を比較



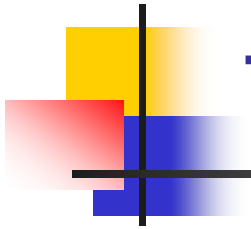
Experimenter:
複数のデータセットに対して複数のアルゴリズムを実行し比較

Knowledge Flow
MLアプリケーションの流れを可視化して実行

WekaにおけるDMのプロセス



- 前処理: データの理解
  - データサイズ，外れ値
    欠損値,…
- フィルター:データの準備
  - 欠損値処理，離散化，サンプリング…
- モデル選択: 学習
  - 分類，クラスタリング，相関規則,…



前処理

- 扱えるデータ
  - 入力形式

    テキストファイル，SQL, URL,…
  - データ形式

    ARFF, CSV, C4.5, …
- 前処理のツールは“filter”とよばれている

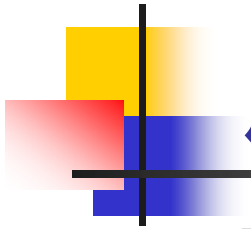
ARFFの概要

```
@RELATION iris
@ATTRIBUTE sepallength   NUMERIC
@ATTRIBUTE sepalwidth    NUMERIC
@ATTRIBUTE petallength   NUMERIC
@ATTRIBUTE petalwidth    NUMERIC
@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
4.7,3.2,1.3,0.2,Iris-setosa
```

データセット名

属性名と
属性の型

CSV形式の
データ
欠損値は?で表す



属性の選択

- 目的変数(クラス属性)に寄与の低い特徴量(属性)を使わない
  - フィルタ法:

    目的変数(クラス属性)と各特徴量(属性)との関係を利用して選択
  - ラッパ法:

    交叉検定法などで求めた誤差を最小にする特徴量(属性)の部分集合を選択

Example data “iris”

- アヤメの品種分類
  - Fisherが判別分析法を紹介するために利用(1936)
- クラス:

  setosa, versicolor, virginica
- 特徴量(属性)

  sepallength：がくの長さ

  sepalwidth：がくの幅

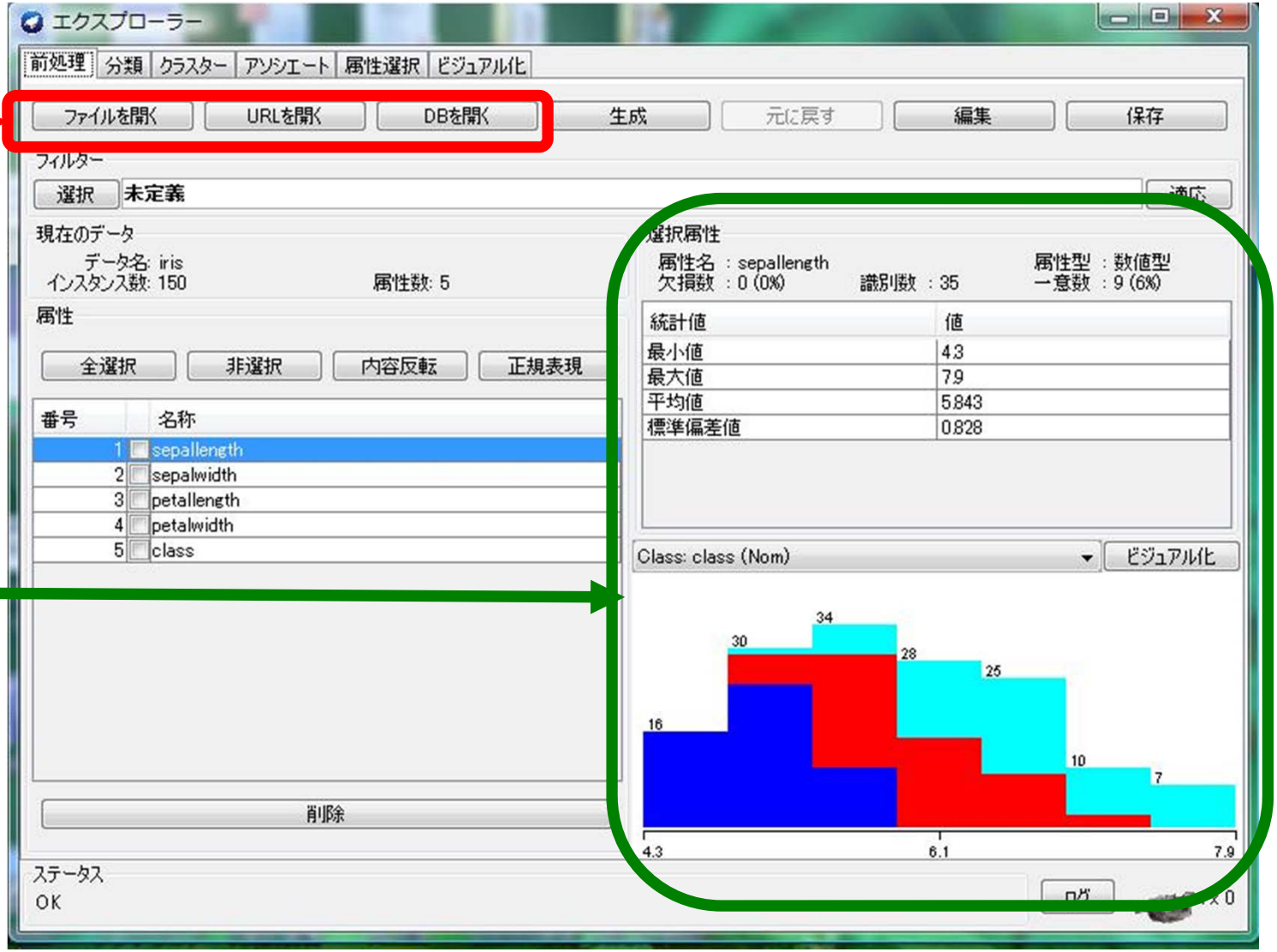
  petallength：花弁の長さ

  petalwidth：花弁の幅

Explorer: 前処理

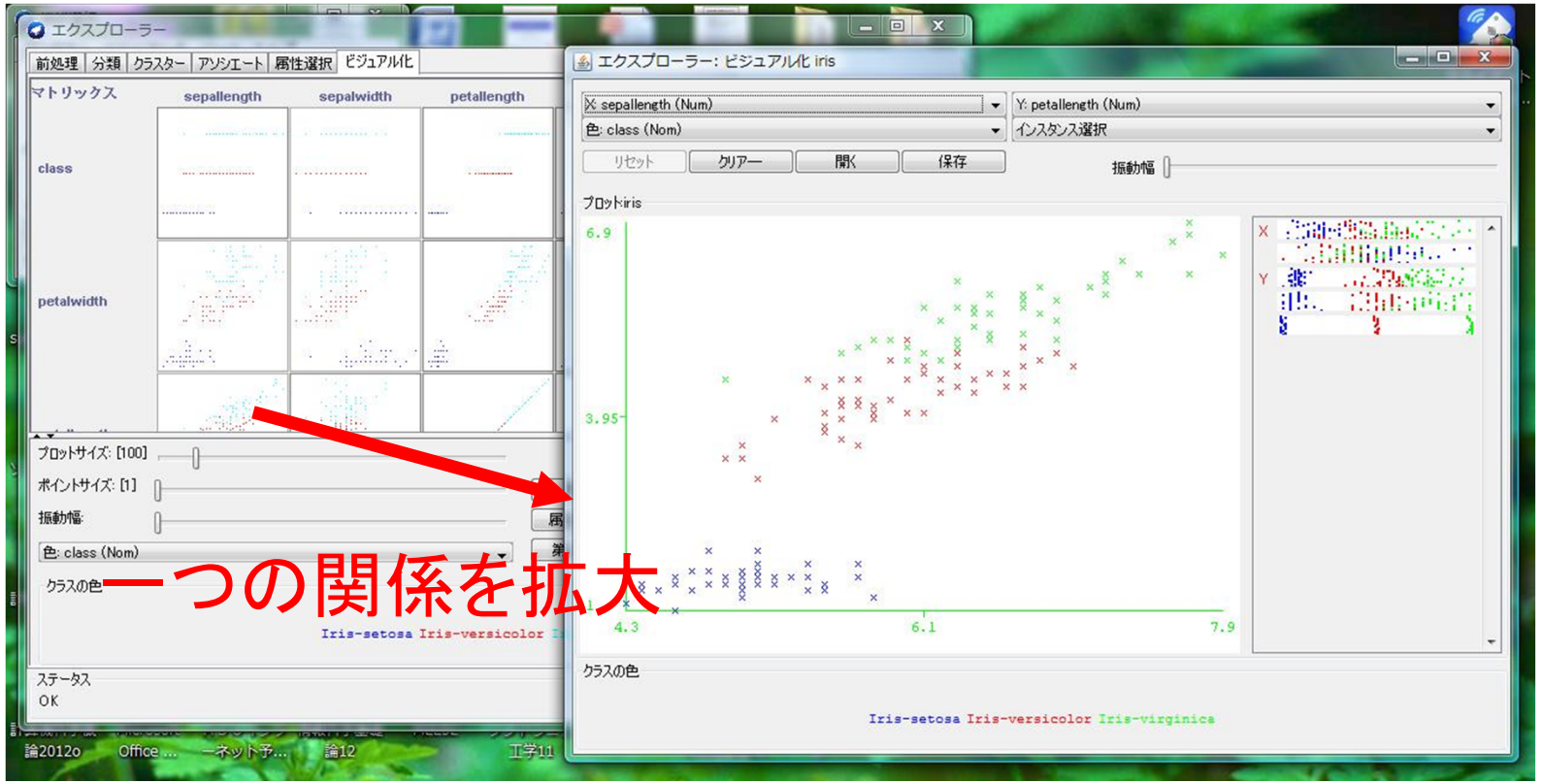
データ
読込

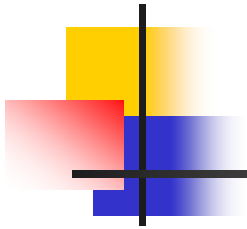
選択した
属性の
統計情報



Explorer: Visualizer

- 特徴量(属性)間の関係を表示





Explorer: Classifiers

- 多様な学習機械が実装

  SVM, ニューラルネット, 決定木,

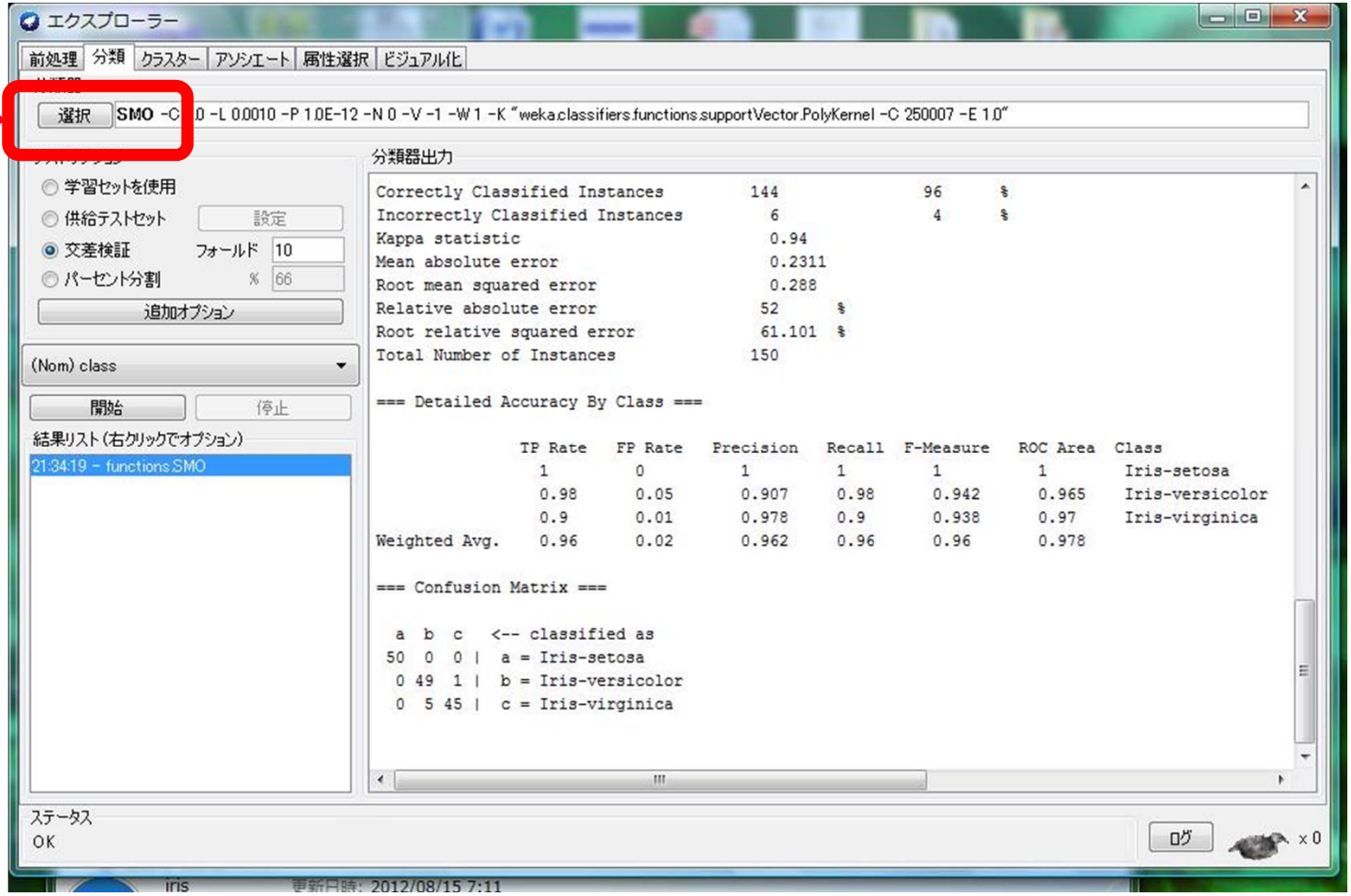
  重回帰分析, ベイジアンネット,...

- メタ学習機械も提供

  Ada-boosting, stacking,...

SVM in Wekaの設定

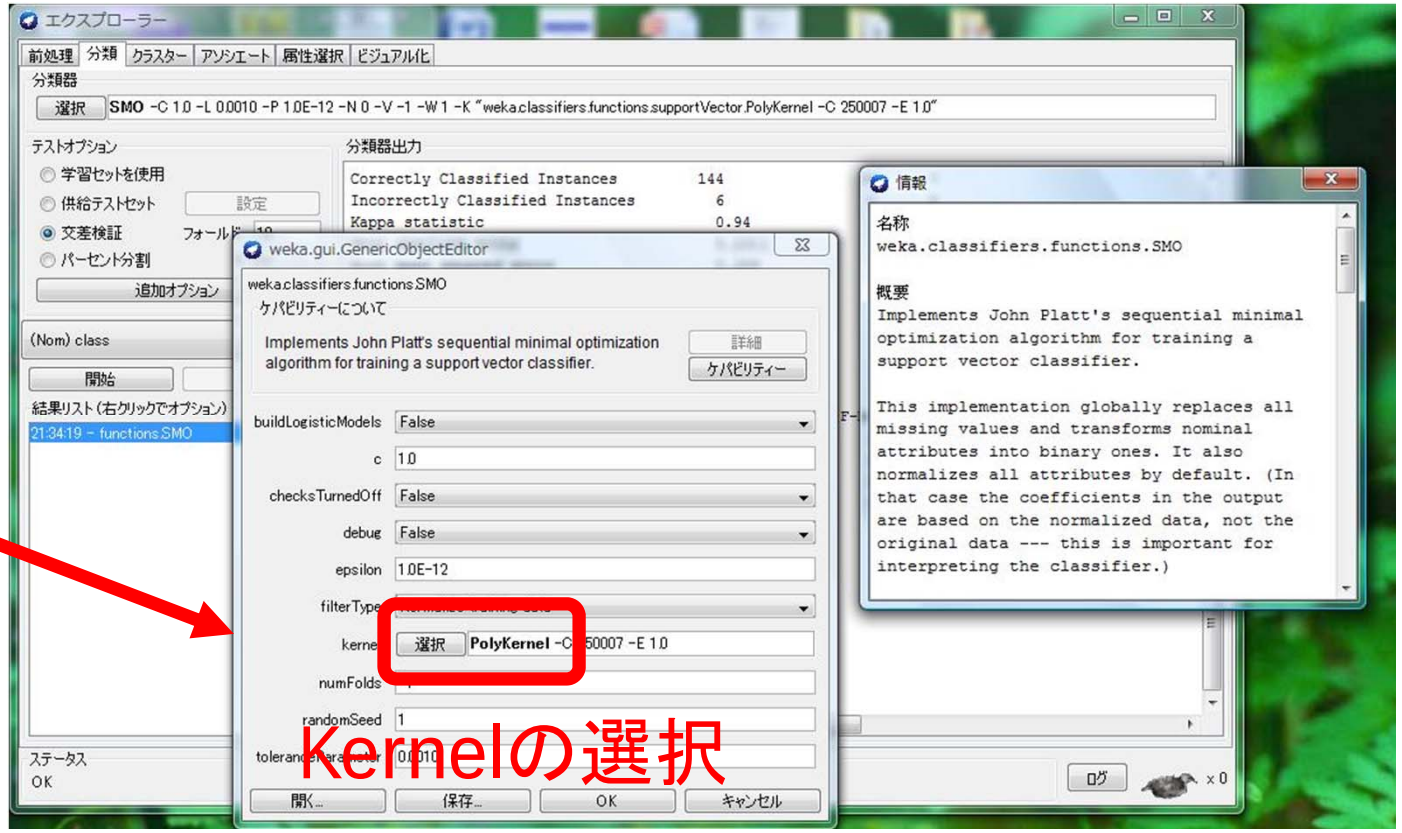
SMOを
選択



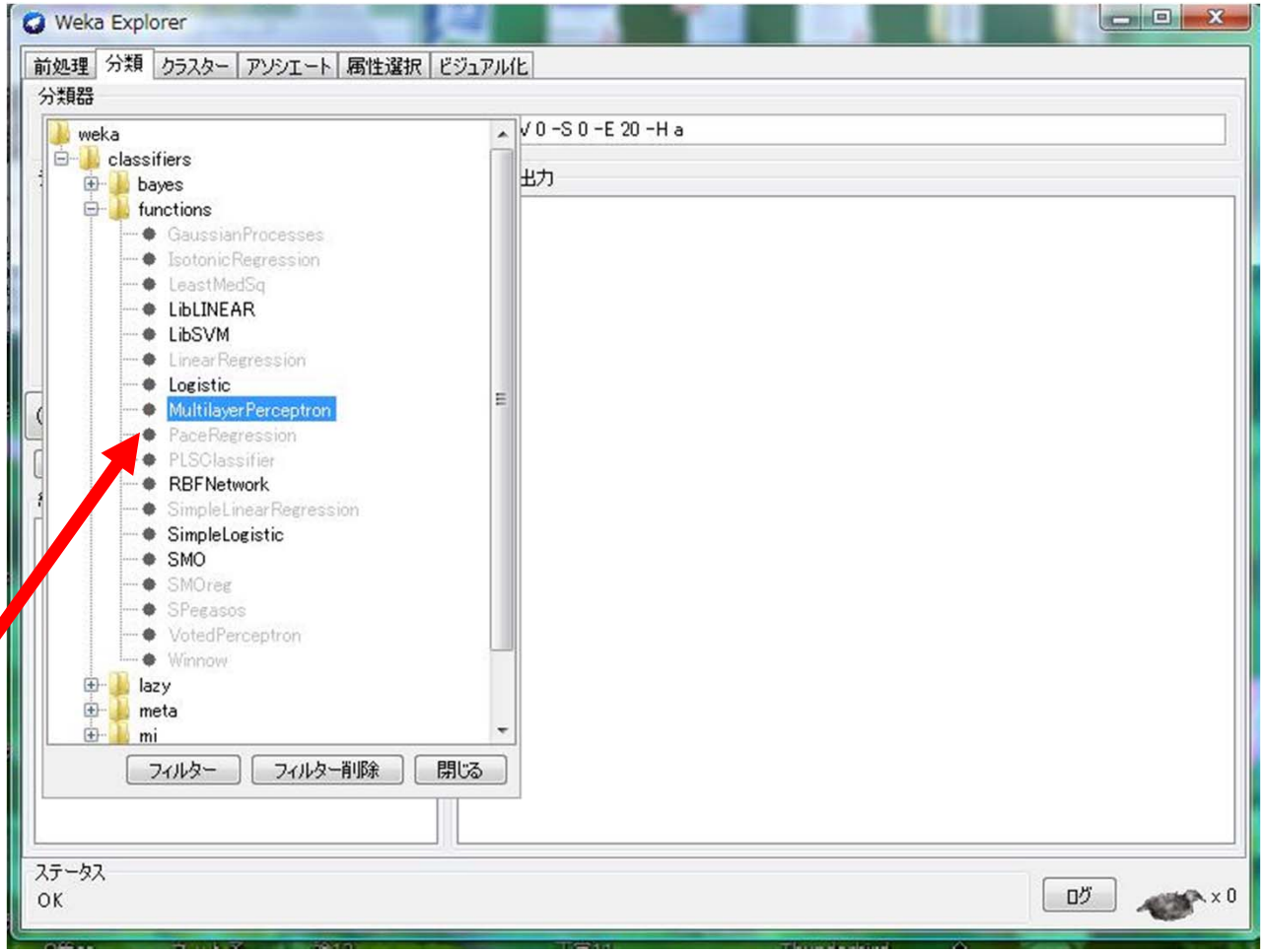
SVM in Wekaの設定

パラメータ設定の説明

SMOをクリックするとパラメータ設定ウインドが現れる



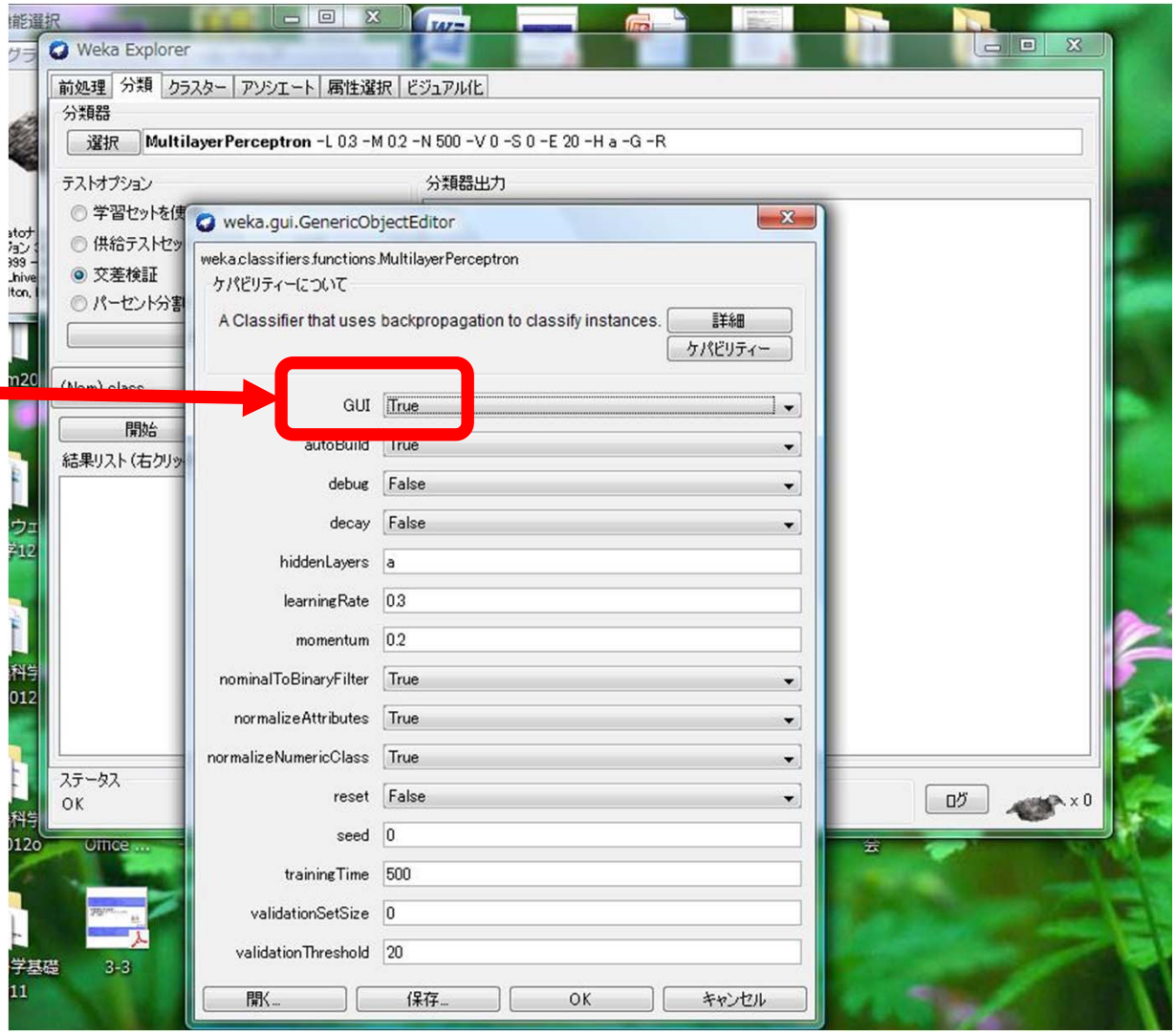
階層型ニューラルネットの設定



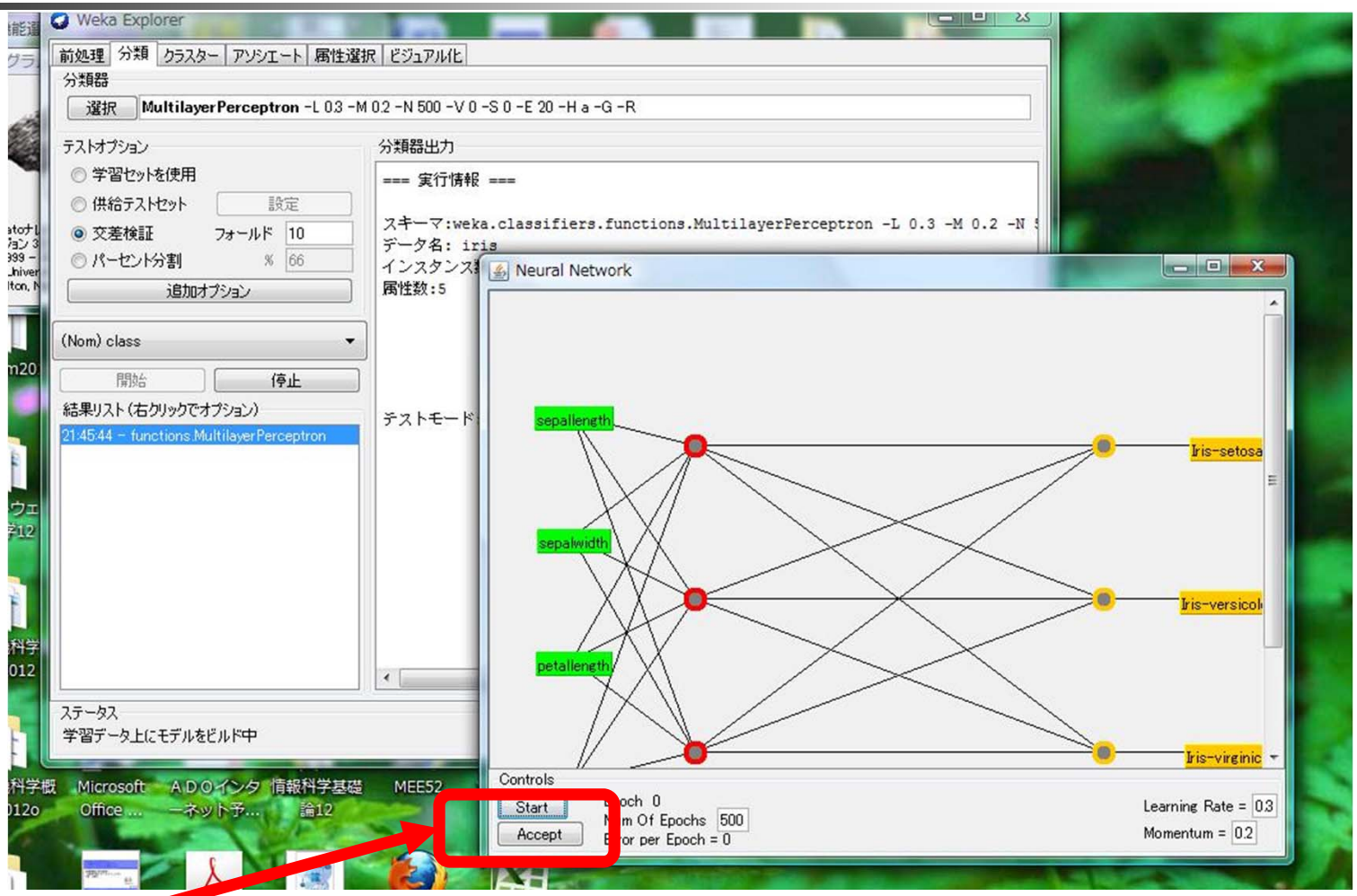
MultilayerPerceptronを選択

階層型ニューラルネットの実行

Multilayer Perceptronをクリックして GUIを選択



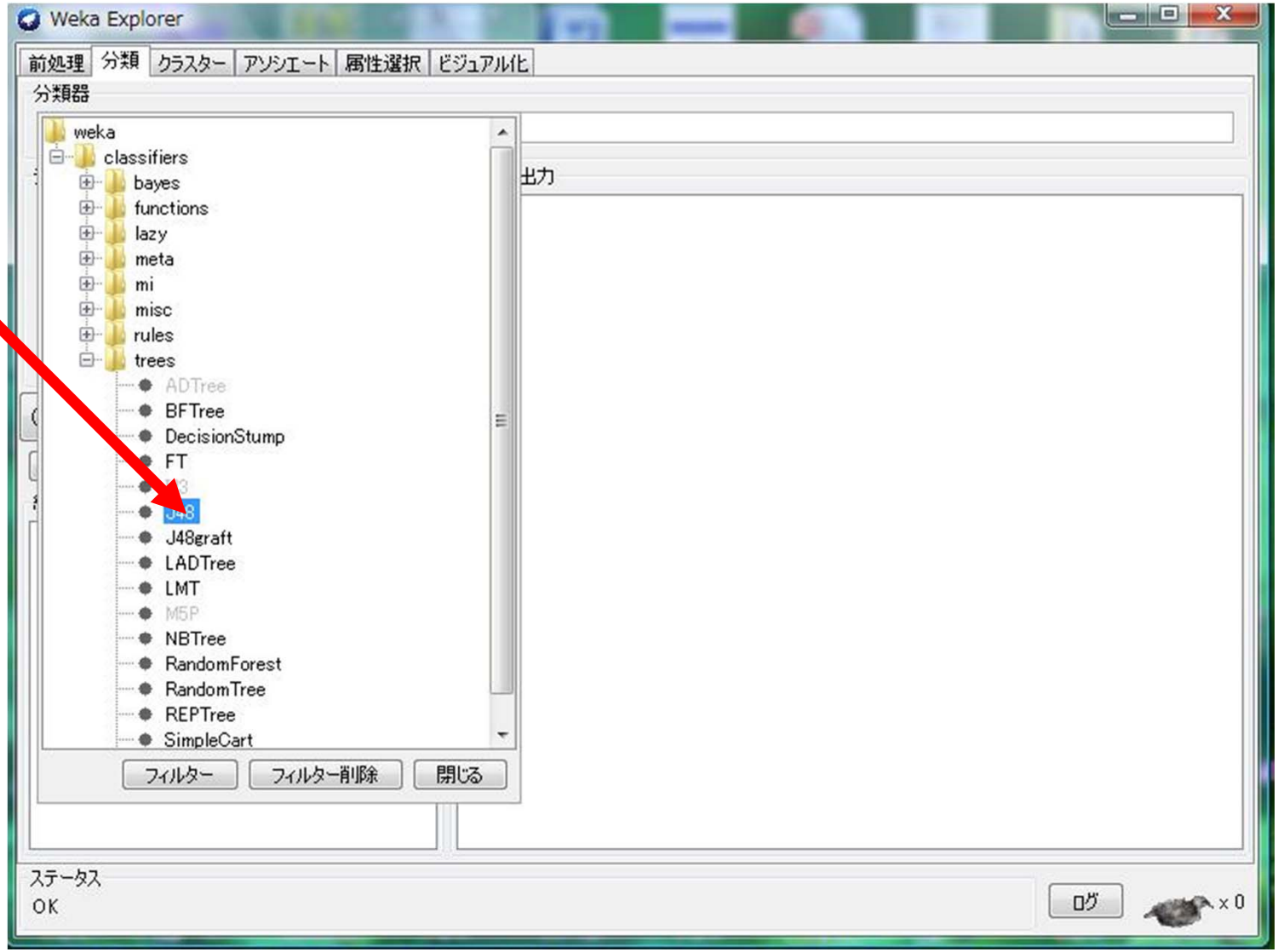
階層型ニューラルネットの実行



Startボタンで開始，Acceptボタンで学習結果の承認

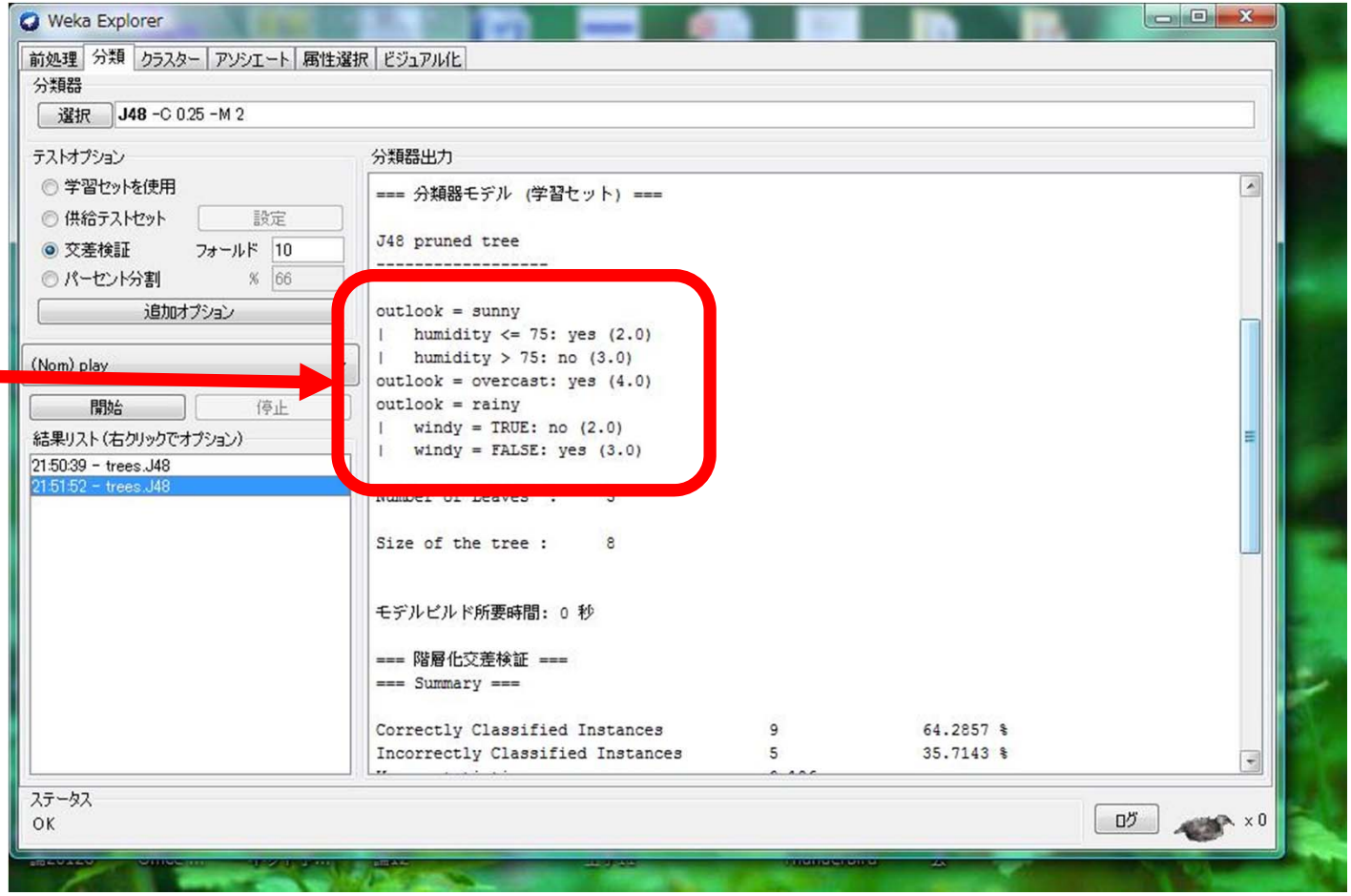
決定木生成の設定

J48を
選択

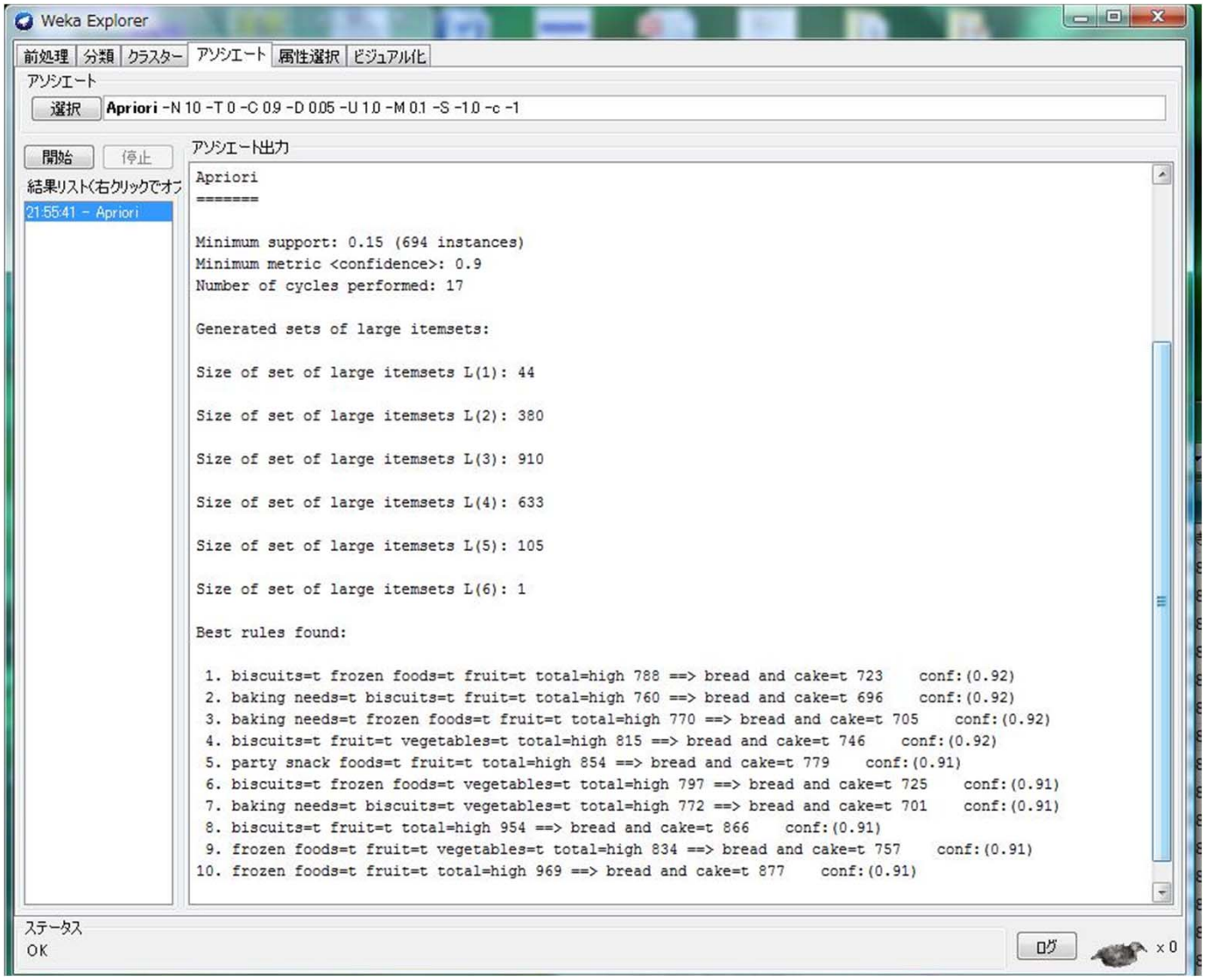


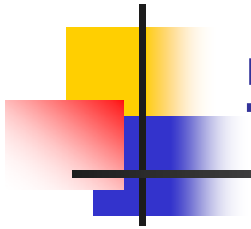
決定木生成の実行

決定木が
表示される



相関規則生成の実行

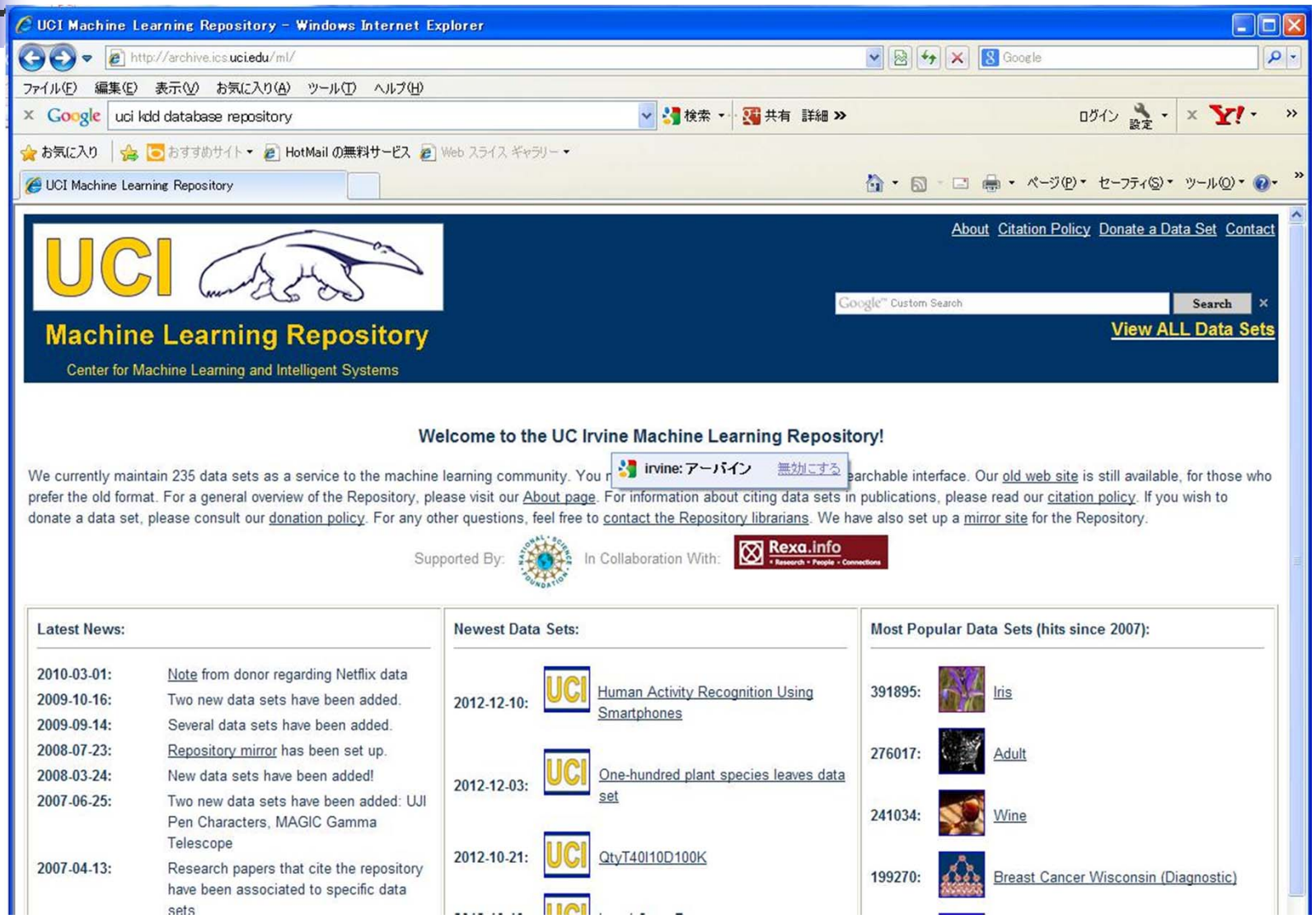

Weka Explorer
前処理
分類
クラスター
アソシエート
属性選択
ビジュアル化
アソシエート
選択
Apriori -N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -c -1
開始
停止
結果リスト(右クリックでオプ
21:55:41 - Apriori
アソシエート出力
Apriori
=======
Minimum support: 0.15 (694 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 17
Generated sets of large itemsets:
Size of set of large itemsets L(1): 44
Size of set of large itemsets L(2): 380
Size of set of large itemsets L(3): 910
Size of set of large itemsets L(4): 633
Size of set of large itemsets L(5): 105
Size of set of large itemsets L(6): 1
Best rules found:
1. biscuits=t frozen foods=t fruit=t total=high 788 ==> bread and cake=t 723 conf:(0.92)
2. baking needs=t biscuits=t fruit=t total=high 760 ==> bread and cake=t 696 conf:(0.92)
3. baking needs=t frozen foods=t fruit=t total=high 770 ==> bread and cake=t 705 conf:(0.92)
4. biscuits=t fruit=t vegetables=t total=high 815 ==> bread and cake=t 746 conf:(0.92)
5. party snack foods=t fruit=t total=high 854 ==> bread and cake=t 779 conf:(0.91)
6. biscuits=t frozen foods=t vegetables=t total=high 797 ==> bread and cake=t 725 conf:(0.91)
7. baking needs=t biscuits=t vegetables=t total=high 772 ==> bread and cake=t 701 conf:(0.91)
8. biscuits=t fruit=t total=high 954 ==> bread and cake=t 866 conf:(0.91)
9. frozen foods=t fruit=t vegetables=t total=high 834 ==> bread and cake=t 757 conf:(0.91)
10. frozen foods=t fruit=t total=high 969 ==> bread and cake=t 877 conf:(0.91)
ステータス
OK
ログ
x 0



学習機械性能評価用

- UCI Machine Learning Repository
  - University California Irvineが提供しているデータアーカイブ
  - http://archive.ics.uci.edu/ml/
- ACM KDDCup
  - ACM SIGKDDが毎年開催するコンテスト用データ
  - http://www.sigkdd.org/kddcup/

UCI ML Repository



Example data “Mushroom”

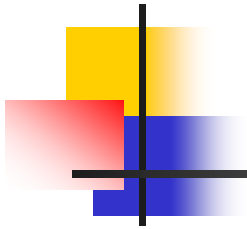
- 北米のキノコの分類
  - 初期の機械学習の開発でベンチマークとして利用
  - 8124タプルは，現在ではサイズが小さい
- クラス:edible, poisonous
- 特徴量(属性)

  cap-shape ：かさの形

  cap-surface ：かさの表面

  など22種

Example data “KDDcup 1999”



- The Fifth International Conference on Knowledge Discovery and Data Mining(KDD 99)のコンテスト用データ
  - データマイニングに関するトップカンファレンスのひとつ
- MIT Lincoln 研究所のネットワークで観測されたTCPアクセスの記録
- クラス: goo, bad
- 743MB
- 特徴量(属性) 31
  - 数値属性と離散属性が混在