

# WEKA と樹木モデル

## 1. WEKA とは

樹木モデルに関するアルゴリズムは多く提案されているが、Rには CART のファミリーの一族 **tree**、**rpart**、**randomForest** 以外には見当たらない。樹木モデル(決定木)の話題で欠かせないのは C4.5 のアルゴリズムである。そこで、本稿では、C4.5 のアルゴリズムが実装されているデータマイニングのフリーソフト WEKA を紹介する。

WEKA は、ニュージーランドのワイカト大学(University of Waikato) の Ian H. Witten、Eibe Frank を中心とした機械学習の研究者によって開発され続けている、Java 言語によるオープンソースのデータマイニングのフリーソフトである。実は、WEKA はニュージーランドに生息し、飛ぶのが苦手であるが、探究心が非常に強い鳥の名前である。

ソフト WEKA に関する 1 次情報は、次のページから入手できる。

<http://www.cs.waikato.ac.nz/~ml/WEKA/>

上記のサイト、あるいは次のサイトからコンピュータの OS(Windows、Mac、Linux など) にマッチした WEKA を入手することができる。

<http://prdownloads.sourceforge.net/WEKA/WEKA.ppt>

使用しているマシンに Java 言語がインストールされていない場合は、Java が同梱されているバージョンを選んだ方がよい。

WEKA で扱っているデータマイニングのアルゴリズムの基礎に関する本としては、ソフトの開発者の著書(参考文献[1])がある。WEKA は、データの事前処理、分類と予測、クラスタリング、相関ルール、視覚化に関するアルゴリズムの集合体である。

WEKA では、データセットの中の列(変数)を**属性(attribution)**、行(個体)を**インスタンス(instance)**、特定のタスクを実行するアルゴリズムの集まりをスキーム(scheme)、判別・分類を行うスキームを**分類器(classifier)**、樹木モデルを**決定木(decision tree)**と呼ぶ。

本稿で用いた WEKA は Windows 用のバージョン 3.4.3 である。

## 2. WEKA の基本操作

WEKA のダウンロードとインストールの手順は紙面の都合により割愛する。

WEKA を起動すると図 1 のような GUI 画面が開かれる。GUI の鳥が WEKA である。GUI の下部には 4 つのボタンがある。それぞれのボタンを押すとデータを操作するパネルが開かれる。その主な機能を表 1 に示す。

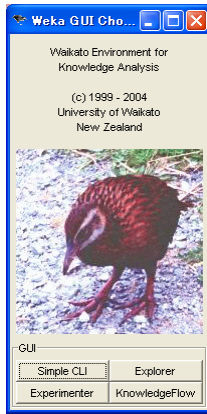


図 1 WEKA の GUI 画面

表 1 GUI のボタンの機能

SimpleCLI	コマンドによる操作環境
Explorer	メニュー選択型の操作環境
Experimenter	学習スキームの間の統計検定などを行う環境
KnowledgeFlow	データ処理・マイニングのプロセスをアイコンで連結してマイニングを行う GUI 環境

WEKA を扱っている書籍 “Data Mining”（参考文献[1]）ではコマンドラインを用いて解説しているが、本稿では GUI の Explorer 環境を用いることにする。WEKA の GUI における [Explorer] ボタンを押すと図 2 のパネルが開かれる。

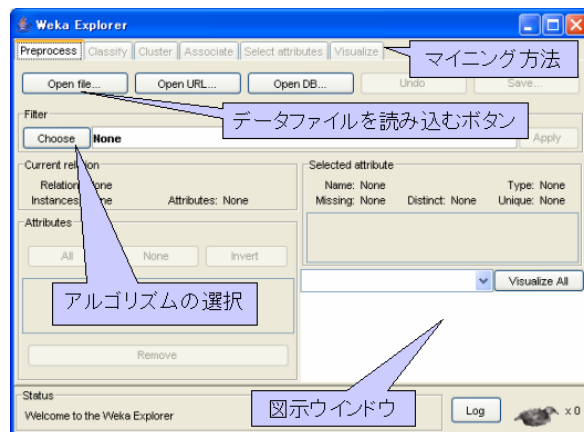


図 2 Explorer 画面

Explorer のパネルの上部には 6 つのタブが設置されている。WEKA のデータ処理・マイニングに関する機能はこの 6 つのタブに分類されている。この 6 つのタブに含まれている機能及び主なスキームを表 2 に示す。

WEKA は、カンマ区切りの CSV 形式、C4.5 形式、ARFF 形式などを読み込むことができる。

CSV 形式は表計算ソフト Excel でも簡単に作成できる。ARFF 形式のデータファイル概観を示すため、データセット Iris の ARFF 形式の一部分のコピーを図 3 に例示する。

表 2 WEKA-3-4-4 のタブとスキーム

Preprocess	データの選択と修正などのための前処理に関するフィルタ環境で、44 種類(supervised 7 種類、unsupervised 37 種類)のアルゴリズムがある
Classify	分類と予測に関する環境で、71 種類(Bayes 7 種類、function 12 種類、lazy 5 種類、meta 23 種類、misc 3 種類、trees 11 種類、rules 10 種類)のアルゴリズムがある
Cluster	クラスタに関する環境で 5 種類のアルゴリズムがある
Associate	相関ルールに関する環境で、3 種類のアルゴリズムがある
Select attributes	属性の選択に関する環境で、20 種類(Attribute Evaluator 12 種類、search Method 8 種類)のアルゴリズムがある。
Visualize	データの 2 次元グラフの環境である

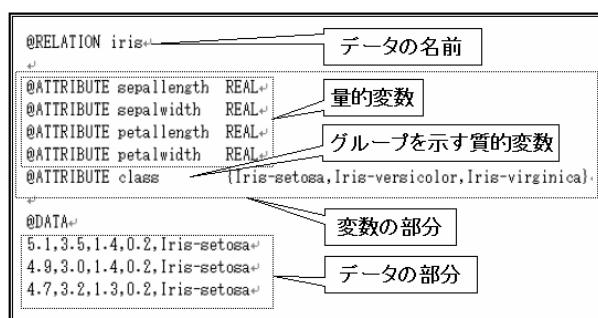


図 3 ARFF ファイルの形式

ARFF ファイルは、データの属性部とデータの部分に分けて記述する。属性の部分は、属性のデータの性質について具体的に記述し、データの部分は属性の順序の順にカンマでデータを区切る。データはローカルディスク、インターネット、データベースから直接読み込むことができる。

### 3. WEKA と決定木

WEKA における決定木は `classify` のタブの中に分類されている。`classify` のタブの中には、分類と回帰に関連するスキームが7つのグループに分けられている。その中の `trees` グループの主な決定木を表 3 に示す。

表 3 WEKA の主な決定木

DecisionStump	決定木の切り株(stump)を生成する
LMT	ロジステックモデルの木を構築する
ID3	ID3 アルゴリズムに基づいた未剪定の木を生成する
J48	Quinlan の C4.5 に基づいた決定木を生成する
NBtree	Naive Bayes 分類器による決定木を生成する
RandomForest	Leo Breiman が 2001 年に提案した「ランダム森」のモデルを構築する
RandomTree	属性をランダムに用いた未選定の決定木を生成する
REPTree	Gini と分散の情報を用いた快速決定・回帰木を生成する

## (1) データを読み込む

Preprocess タブの [Open file] ボタンを押し、データが置かれているフォルダを開き、ファイルを指定して、開かれているパネルの「開く」ボタンを押すとファイルが読み込まれる。WEKA には、幾つかのデータセットがインストールする際に作成された `data` というフォルダに置かれている(例えば、`C:\Program Files\Weka-3-4\data`)。ここではその中の `iris` データセットを用いて説明する。図 4 にフォルダ `data` の中の `iris.arff` を読み込んだパネル画面を示す。

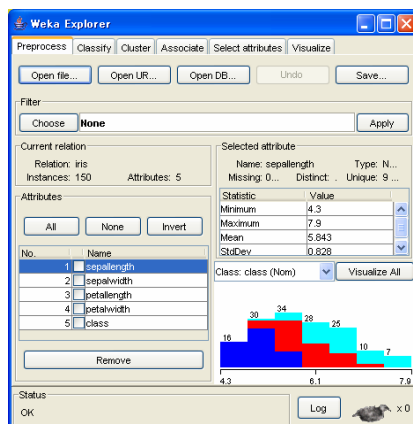


図 4 データセット `iris` を読み込んだ画面

## (2) マイニング方法の選択

目的に応じて、マイニングタブをクリックし、アクティブ化する。決定木の選択は、Classifyのタブをアクティブ化し、[Choose]のボタンを押し、trees フォルダのリストから1つを選ぶ。ここではJ48を選択する。J48が取り入られている画面を図5に示す。[Choose]ボタンの右側にJ48-\*\*\*が表示されたら、正常に読み込まれている。

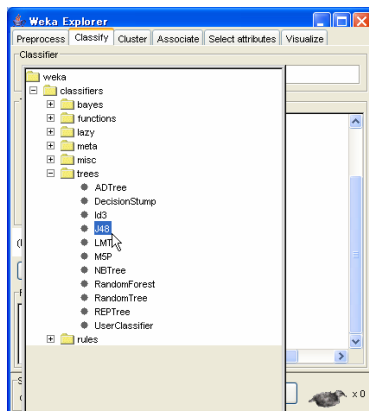


図5 J48が読み込まれている画面

J48はC4.5をWEKAに実装したものである。C4.5では、ゲイン比(gain ratio)を分岐基準としている。ゲイン比は、エントロピーとゲイン値から求められる。

最も簡単なデータの例として、表4のような男女別の意思表示(賛成、反対)に関するデータがあるとする。意思表示の結果を目的・被説明変数、性別を説明変数とする。

表4 データの例

	意思表示		合計
	賛成	反対	
男	4	3	7
女	6	2	8
合計	10	5	15

$$\begin{aligned} \text{info(意思)} &= - \sum_i p_i \log_2(p_i) \\ &= - \frac{10}{15} \log_2\left(\frac{10}{15}\right) - \frac{5}{15} \log_2\left(\frac{5}{15}\right) \\ &\doteq 0.9183 \end{aligned}$$

$$\begin{aligned} \text{info(性別 \& 意思)} &= - \sum_{i,j} \frac{N_{ij}}{N} p_{ij} \log_2(p_{ij}) \\ &= - \frac{7}{15} \left[ \frac{4}{7} \log_2\left(\frac{4}{7}\right) + \frac{3}{7} \log_2\left(\frac{3}{7}\right) \right] \\ &\quad - \frac{8}{15} \left[ \frac{6}{8} \log_2\left(\frac{6}{8}\right) + \frac{2}{8} \log_2\left(\frac{2}{8}\right) \right] \\ &\doteq 0.8925 \end{aligned}$$

$$\begin{aligned} \text{gain(性別)} &= \text{info(意思)} - \text{info(性別 \& 意思)} \\ &= 0.9183 - 0.8925 = 0.0258 \end{aligned}$$

$$\begin{aligned} \text{split info(性別)} &= - \sum_j \frac{N_{ij}}{N} \log\left(\frac{N_{ij}}{N}\right) \\ &= -\frac{7}{15} \log_2\left(\frac{7}{15}\right) - \frac{8}{15} \log_2\left(\frac{8}{15}\right) \\ &\doteq 0.7042 \end{aligned}$$

$$\begin{aligned} \text{gain ratio(性別)} &= \frac{\text{gain(性別)}}{\text{split info(性別)}} \\ &= \frac{0.0258}{0.7042} \doteq 0.0366 \end{aligned}$$

紙面の都合により、ここでは1つの変数(性別)のみのゲイン比を求めた。このように全ての説明変数についてゲイン比を求め、その中で最も大きいゲイン比を持つ「変数」を決定木の分岐に用いる変数の第1候補とする。

変数が連続の量的データの場合は、群間の平方和が最大になる変数を第1候補とする。

### (3) 決定木の生成

Classify パネルの左側の Test options の下には、学習データセット(Use training set)、テストデータセット(Supplied test set)、交差確認法(Cross-validation Folds)、データセットの1部分(Percentage split)がある。決定木を求める前に、この中から1つを選択しなければならない。ここでは、学習データセット(Use training set)を指定する。

出力の結果に関しては[More options...]ボタンを押し、開かれる分類器環境(Classifier evaluation)パネルで自由に指定することができる。ここでは図6のように設定し、[OK]ボタンを押し。

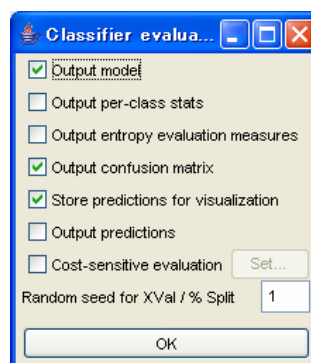


図6 出力に関するオプションパネル

ボタン[More options...]の下部の窓には目的変数を指定する。

上記の設定が終わったら、[Start]ボタンを押すとプログラムが実行され、計算結果が返され

る。その結果の画面を図 7 に示す。

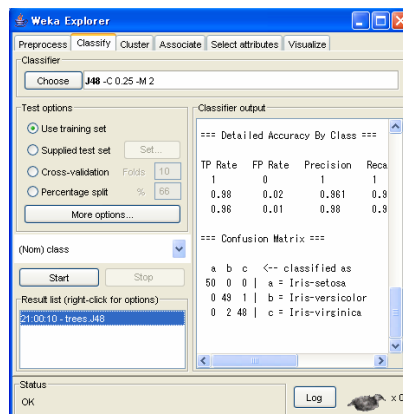


図 7 J48 が実行された画面

パネルの右側の **Classifier output** の下の窓には決定木に関する結果がテキスト形式で出力される。結果は、用いたデータに関する情報、決定木のルール、決定木生成に関する情報の要約、正・誤判別の行列(Confusion Matrix)の順になっている。出力される結果は、分類器環境パネルの設定に依存する。決定木のルール画面コピーを図 8 に示す。

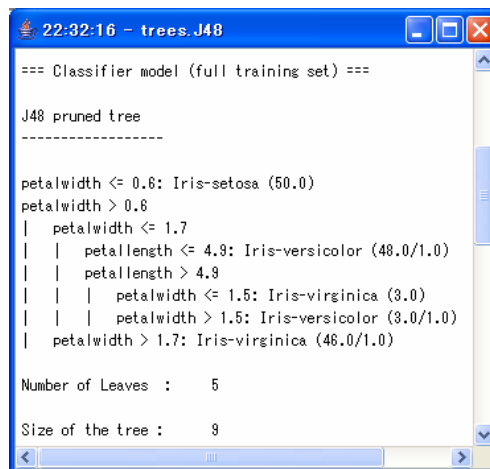


図 8 テキスト結果のウィンドウ

パネルの左下部の **Result list(right-click for options)** の下の窓に行ったマイニングの結果のリスト返される。リストの項目を右クリックすると図 9 のようなメニューが開かれる。

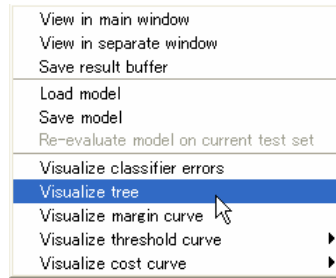


図 9 結果の表示リスト

メニューの中の **View in separate window** を左クリックすると操作パネルと分離されたテキスト結果のウインドウが開かれる。

メニューリストの中の **Visualize tree** を左クリックすると生成した木のグラフが **Tree view** ウインドウで作成される。マウスポインタを木のグラフに合わせ、左ボタンを押したまま移動することで自由にグラフの位置を変えることができる。

**Tree view** ウインドウの空白部分を右クリックすると図 10 に示すような、グラフの配置スタイルや文字列のサイズを調整するメニューが開かれる。

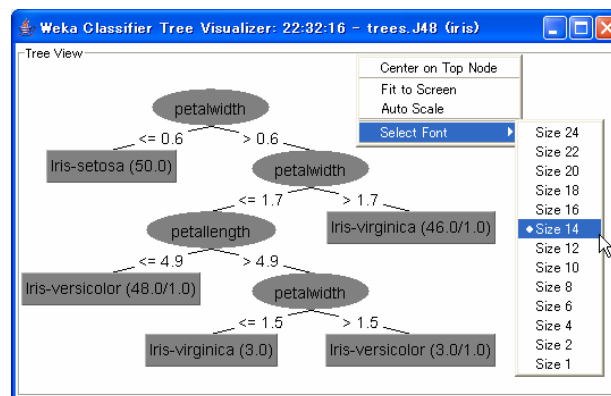


図 10 決定木のグラフウインド

#### (4) パラメータの調整

**Classify** パネルの [**Choose**] ボタンの右にある文字列の窓をクリックするとその方法のオプションパネルが開かれる。図 11 に J48 のオプションパネルを示す。

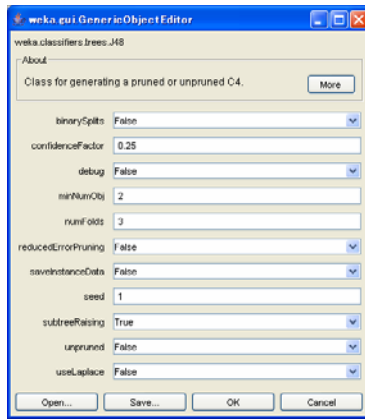


図 11 J48 のオプションパネル

返された J48 の決定木は、開かれているパネルに示されている条件の下で生成されている。表 5 にパネルのパラメータを簡潔に説明する。

表 5 J48 のパラメータ

binarySplits	名義(nominal)尺度の変数を 2 分岐するか(True)しないか(False)の指定
confidenceFactor	剪定のための信頼要因。値が小さいほどより多く剪定される
debug	コンソールに返す情報をコントロールする
minNumObj	葉における最少の個体数
numFolds	データを分割する組数。1 組を REP(reduced-error pruning)という剪定に、その残りを木の生成に用いる
reducedErrorPruning	C4.5 の剪定の代わりに REP を用いるかどうかの指定
saveInstanceData	視覚化のために学習したデータを保存するかどうかの指定
seed	REP を行うときに無作為化に用いる種(seed)
subtreeRaising	剪定を行うとき部分木の扱いの指定
unpruned	剪定を行うかどうかの指定
useLaplace	葉の計算はラプラス(Laplace)における平滑に基づくかどうかの指定

オプションの設定例として、オプションの `minNumObj` を 4 に設定した決定木のグラフを図 12 に示す。図で分かるように葉の中に含まれている個体数が 4 以下の葉を刈り切った木が出力されている。

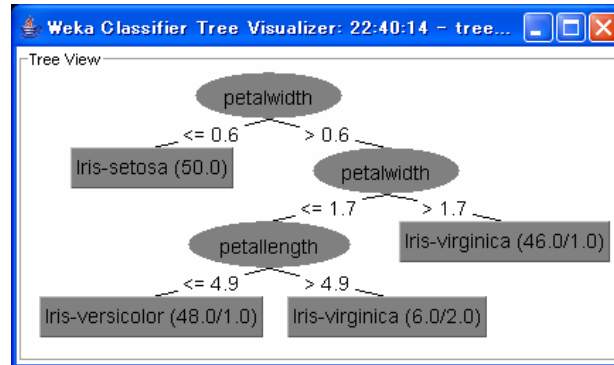


図 12 個体数が 4 以下の葉を切り取った iris の決定木

静岡大学工学研究科阿部秀尚氏のホームページに WEKA のインストールを含む決定木などに関するパワーポイントの資料が公開されている。2005年2月現在の URL を次に示す。

<http://panda.cs.inf.shizuoka.ac.jp/~hidenao/work/weka/>

#### 参考文献

- [1] I.H. Witten, E. Frank: Data Mining: MORGAN KAUFMANN: ISBN 1-55860-552-5