

Wekaの基本 & Tips集

阿部 秀尚
静岡大学大学院理工学研究科
http://panda.cs.inf.shizuoka.ac.jp/~hidenao/work/weka/
hidenao@ks.cs.inf.shizuoka.ac.jp

目次

- Wekaとは？
- Wekaを使う前に
 - ARFFの書き方
 - ARFF以外を使う
- Wekaを使う
 - Explorerを使う
 - Knowledge Flowを使う

Wekaとは？

- オープンソースのデータマイニングツール^(*)
- 実装はマルチプラットフォーム対応のPure Java
 - 配布パッケージはWindows/MacOS X/JAR (ZIP)
- Waikato大学(New Zealand)が中心になって開発
 - <http://www.cs.waikato.ac.nz/ml/weka/>
 - Weka: Waikato Environment for Knowledge Analysis
- 対話的にDMを実行できるGUIやDMの流れを表現できるGUIを装備！
 - JavaAPIやコマンドラインからも利用可能

(*)データマイニングアルゴリズムを集め、GUI・APIなどから統合的に利用できるソフトウェア

ARFFの書き方

@relation weather	weather	データセットの名前
@attribute outlook {sunny,overcast,rainy}	outlook {sunny,overcast,rainy}	属性名を縦に列挙する
@attribute temperature {numeric}	temperature {numeric}	数値属性を示す (実数, 整数)
@attribute humidity integer	humidity integer	数値属性を示す (実数, 整数)
@attribute windy {true,false}	windy {true,false}	名義属性は属性値 をカンマ区切りで示す
@attribute time_sdate "yyMMdd HH:mm"	time_sdate "yyMMdd HH:mm"	日付はdate+フォーマットで 定義する
@attribute play {yes,no}	play {yes,no}	クラスも他の名義属性と 同様に記述 (実行時に指定)
@data	@data	@data以降CSV形式の データ
sunny,85,85,FALSE,010822 10:00,no	sunny,85,85,FALSE,010822 10:00,no	不明値は'？'
overcast,83,86,FALSE,010911 11:00,yes	overcast,83,86,FALSE,010911 11:00,yes	
rainy,70,?,FALSE,010921 10:30,yes	rainy,70,?,FALSE,010921 10:30,yes	

ARFF以外をWekaに入力

CSVファイルをWekaに入力するにはどうしたらいいの？

Answer:
Explorerの"Open file..."でファイルを開こうとするとLoaderを指定するよう促されますので、CSVLoaderを指定してください。
また、コマンドラインでの変換もできます。
\$ java -cp weka.jar%
weka.core.converters.CSVLoader%
-i foo.csv -o bar.arff

C4.5の形式を入力できますか？

Answer:
Explorerの"Open file..."で開くとLoaderを指定するよう促されますので、C45Loaderを指定してください。
また、CSV同様にコマンドラインでの変換もできます。

データベースに蓄えたデータをWekaに入力したいんだけど...

Answer:
Explorerの"Open DB..."からDBにアクセスします。
JDBCドライバの指定やURL・ユーザ名・パスワードは実行ディレクトリのDatabaseUtils.prop内に記述します。

Explorerを使う (データの読み込み)

1: "Open file..." "Open URL..."でファイルから、"Open DB..."でDBからデータを読み込む

2: "Attributes"で選択した属性の名前、種類、欠損値率、統計値とヒストグラムが表示される

Explorerを使う (データの内容を可視化する)

"Visualize"のタブを選択

領域、点の大きさ、ジッターを調整し、"Update"で反映させる

"Plot Matrix"の1つをダブルクリックして、指定した2属性間の関係を拡大して表示する

Explorerを使う (決定木学習の実行を設定する)

1: "Choose"ボタンを押して出てくるリストから実行するアルゴリズムを選択する

2: "Choose"の横のテキストエリアをクリックし、選択したアルゴリズムのパラメータを設定する

3: "More"ボタンを押すと、各アルゴリズムとパラメータの簡単な解説が表示される

Explorerを使う (決定木学習の実行と結果の表示)

1: 正解率の評価方法とクラス(目的変数)を設定する。テストデータセットを使う場合はここを選択する

2: "Start"によって実行開始

結果を見たい実行を選び、右クリックから"Visualize Classifier Errors"を選択する。(□が誤分類)

結果を見たい実行を選び、右クリックから"Visualize Tree"を選択する

訓練データセットでの決定木(テキスト)と"Test Options"で指定した評価を行った結果が表示される

Knowledge Flowを使う (DM実行手順の設定)

実行フローを保存し、ロードすることも可能

必要なデータマイニング手法のアイコンを選択

データマイニング手法の実行手順を"dataSet"や"trainingSet","Graph"など、データの流れを示す→によって連結して作成する。(→を引こうとすると、連結可能な手法にマーカーが出現する)

Knowledge Flowを使う (可視化したDMプロセスの実行)

1: "Data Source"に読み込むデータセットを設定し、"Start Loading"で実行を開始する

2: "Visualize"から選んだアイコンに結果が流れ込むので、"Show Result"を選択し、結果を可視化する

Wekaに関するFAQ

- Wekaを使うと"Out of Memory Exception"で終了してしまうのですが...
 - javaのオプションとして"-Xmx1024m"のように最大メモリを増やしてください。
 - 数万インスタンス程度のデータセットでも実行可能です(55属性×13,000インスタンスでJ4.8とAprioriの実行を確認)。
- Wekaの実行速度はどれくらいですか?
 - Java Virtual Machineの種類や設定、コンパイル方法に左右されますが、経験的にはC言語での実装の数倍程度という感じです。
- Wekaをインストールせずに使いたいのですが?
 - Automatic Knowledge Minor (<http://www.auknomic.com/>)ではWebからWekaを利用できます。
- WekaはPMMLに対応していますか?
 - Bristol大学(UK)の学生Jiwen Li氏がWekaのPMML拡張を公開しています。 <http://www.cs.bris.ac.uk/home/jl2092/>