阿部秀尚†

キーワード：データマイニングツール，データマイニング，パターン認識，決定木学習

1．まえがき

本稿では，映像メディアデータに対するデータマイニング手法の適用について，データマイニングツールWeka[1)]が活用できるよう，その入手から主にデータの自動処理に向けた利用方法について紹介する．

Wekaでは，データマイニングの適用のため多く用いられる機械学習アルゴリズムをはじめ，従来のパターン認識に用いられるアルゴリズムが実装されている．このため，新たに考案したパターン認識アルゴリズムとの比較にも用いることができるほか，データマイニング適用事例に多く用いられる決定木やif-thenルールのような判別過程が明らかなモデルを生成することも可能である．具体的には，**図1**に示すように「新たに考案した特徴量が有効であるか，有効な適用場面の明示化」「映像中の情報抽出後のデータ処理」などへの適用が可能である．

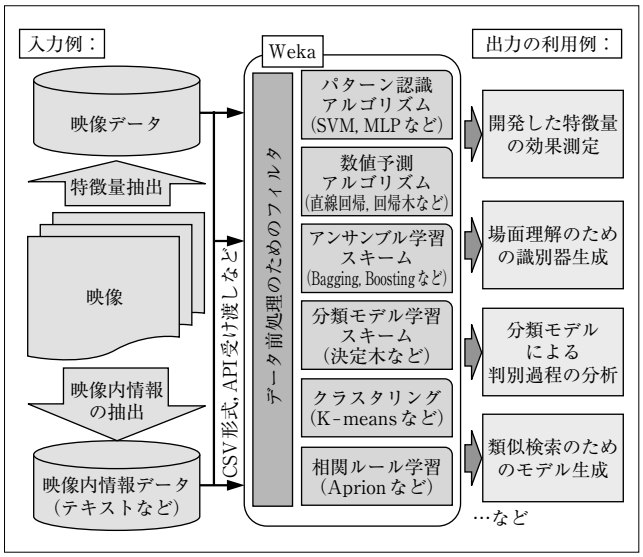

図1　Wekaを利用した映像データ処理の例

2．Wekaの概観

Wekaは，ワイカト大学（ニュージーランド）の機械学習研究グループを中心に開発されている機械学習アルゴリズムを適宜実行，開発するためのソフトウェアである．Wekaは，オープンソース（GPL/LGPL）で開発されたソフトウェアであり，実装言語にはJavaが用いられ，パッケージ以外のライブラリーを要しない．このため，拡張性および他の統合ツールとの親和性が高いという特徴をもっており，KNIME[2)]，RapidMiner[3)]，Pentaho[4)]など，KDNuggets.com[5)]での利用状況調査で利用頻度が上位とされたツールのマイニング処理ライブラリーとしても広く用いられている．

Wekaの機械学習アルゴリズム集には，以下に挙げるように，パターン認識や多変量解析でもよく用いられるものも含め，主要なアルゴリズムが実装されている．

（1）数値予測・分類学習（分類・Classify）
- tree型（J4.8（C4.5），CART，RandomForestなど16種類）
- function型（SVM，ロジスティック回帰，線形回帰，マルチレイヤパーセプトロンなど18種類）
- Bayes型（NaiveBayes，BayesianNetworkなど13種類）
- rule型（PART，OneRなど11種類）
- lazy型（IBk（k-NN）など4種類）
- metaスキーム型（Bagging，Boosting，Stacking，Votingなどのアンサンブル学習スキーム31種類）

（2）クラスタリング（クラスタ・Cluster）
- 混合正規分布（EM），DBScan，k-Means，XMeans，密度に基づく凝集など13種類

（3）アソシエート（Associate）
- Aprioriなど6種類

以上の各アルゴリズムは，単体でコマンドラインから，あるいはAPI呼び出しにて実行可能である．また，Weka自体にも，これらを実行する3種類のGUIが用意されている．

3．入手からインストール

Wekaの開発は，sourceforge.net上で行われ，各環境向けのパッケージについても，同サイトからのダウンロードにより入手できる．まず，WekaのWebページ（http://

† 島根大学　医学部　医学科　医療情報学講座
"Data Mining Tool: Weka" by Hidenao Abe (Dept. of Medical Informatics, School of Medicine, Shimane University, Shimane)

映像情報メディア学会誌 Vol. 65, No. 10, pp. 1398～1401（2011）


```
weka
+core(データセットなどの基本クラスパッケージ)
+gui(各種GUI)
+filters(データ前処理用クラスのパッケージ)
  +supervised
    +attribute
    +instance
  +unsupervised
    +attribute
    +instance
+classifires(分類・数値予測アルゴリズム群)
  +trees(ツリー型)
    -J48(C4.5決定木学習アルゴリズムの実装)
    -SimpleCart
    …
    +functions(関数型)
    +bayes(ベイズ型)
    +rules(ルール生成型)
    +lazy(事例ベース型)
    +meta(アンサンブル学習スキーム)
+clusterers(クラスタリングのアルゴリズム群)
+associations(相関ルール生成アルゴリズム群)
+AttributeSelection(属性選択法)
```

図2　Wekaのクラスパッケージ階層の概観

www.cs.waikato.ac.nz/ml/weka/)のメニューから"Download"のページを辿り，Snapshot(開発)版の下に，Stable(現行)版のパッケージとしてWindows(32ビット/64ビット)，Mac OS X, ZIPアーカイブファイルへのリンクがある．ここから，各自の環境に合わせてパッケージをダウンロードする．なお，Weka3.6.x以降が要求するJavaの実行環境は，Oracle(Sun)版1.5以上である*1．

WindowsおよびMac OS Xでは，それぞれの環境に合わせたパッケージをダウンロード後，通常のソフトウェアと同様にインストールを行うことで起動が可能となる．また，ZIPアーカイブファイルは，適宜ZIP形式のファイルを展開できるソフトウェアを用いて，ファイルを展開し，4章で示すように，コマンドラインからJVMでweka.jarあるいは各処理のクラスを起動する．

図2にWekaのクラスパッケージ階層を示す．各クラスに関する説明は，インストールフォルダのdoc以下に収録されている．

4．Wekaの起動法

Wekaのパッケージをインストールすると，メニュー等にWekaが登録される．この登録されたメニュー項目からWekaを起動すると，最初に起動されるのが，**図3**に示すGUI Chooser(機能選択)である．ZIPアーカイブ版を展開した場合，あるいは，直接コマンドラインからこのGUI Chooserを起動するには，Wekaのインストールされたフォルダに移動し，以下のようにコマンドを実行する．

*1　Windows向けにはjava実行環境付きのパッケージも用意されている．

図3　Windowsで起動したWeka3.6.5のGUI Chooser

```
$ cd "c:¥Program Files¥weka-3-6"
  (Windowsの場合)
$ java -jar weka.jar
```

GUI Chooserからは，以下の4種類のGUIによる実行環境を起動することができる．

(1)エクスプローラ(Explorer)：対話的に前処理・分類学習・数値予測・クラスタリング・相関ルール・属性選択・データ視覚化の操作を適用するために用いる．Wekaに用意されるすべての処理や視覚化の機能が，GUIを通して実行可能である．実行時に必要なオプションをGUIで設定すると，テキスト形式として表示されるので，コマンド化をする際に参照が可能である．

(2)検証(Experiment)：アルゴリズムあるいはデータセットの違いによる精度の比較を行う．検証用の結果データを読み込んで平均正解率の差を検定し，比較表を各種形式で作成することができる．

(3)ナレッジフロー(Knowledge Flow)：エクスプローラなどで対話的に実行した一連の流れを，各処理に対応したアイコンを配置し，矢印を接続することによって，実行フローを作成できる．作成した一連のフローは，ファイルとして保存することにより，共有が可能である．

(4)コマンドライン(CLI)：コマンドラインでの実行をシミュレートするために用いる．

以上のように，それぞれのGUIは，想定される用途が異なっており，利用者の用途によって使い分けができるように用意されている．以降では，エクスプローラでの操作と，シェルやスクリプトからコマンドを実行する方法について紹介する．

5．入力データの加工

Wekaの各種アルゴリズムを直接起動する場合には，入力ファイルがARFF形式であることが要求される．ARFF

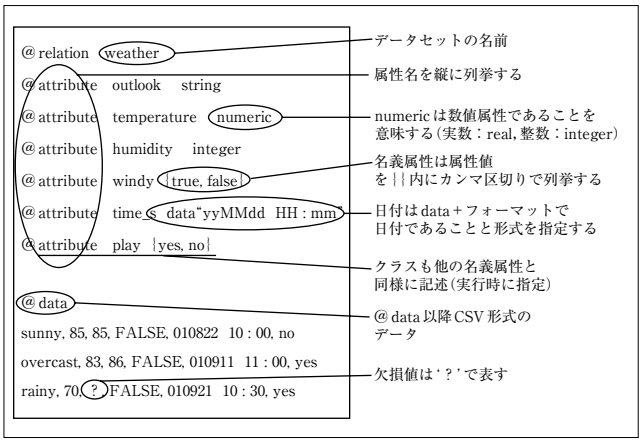

図4　ARFF形式の例と属性上の表記法

図6　エクスプローラでの分類アルゴリズム（決定木学習J4.8）の実行後の画面

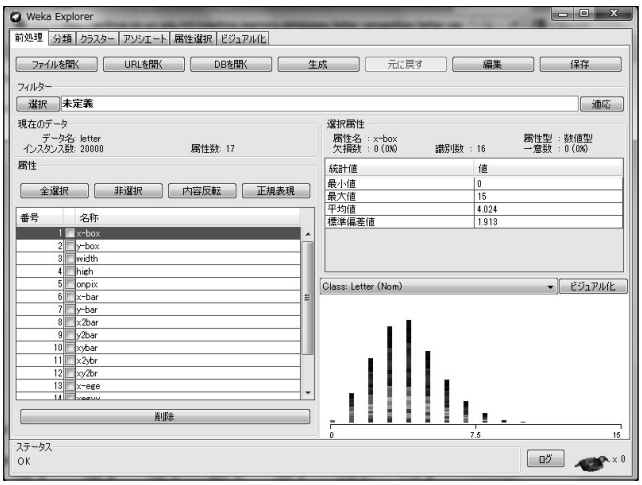

図5　エクスプローラの画面（データ読込後）

形式の概要を**図4**に示す．ARFF形式は，Wekaの直接入力として用いられる形式であり，データを保持するデータベースや他のソフトウェアからのデータは，変換処理を実行する必要がある．

エクスプローラでは，**図5**に示す“前処理（Preprocess）”パネルから，ARFF以外にもCSV形式ファイル，C4.5形式ファイル等を「ファイルを開く（Open File）」ボタンからファイル選択を行って読込む．また，リレーショナルデータベースからの読込みは，「DBを開く（Open DB）」ボタンからSQLクエリによって行える．読込んだデータセットは，直接，あるいはフィルタで各種処理を適用した上，「保存（Save）」ボタンでARFF形式等での保存が可能である．

データの前処理に使われる各種処理は，「フィルタ（Filter）」ボタンで適用するフィルタを選択し，要素の設定で各パラメータの設定後に「適応（Apply）」ボタンで実行する*2．フィルタの種別は，例えば，ヒストグラムの階級設定のように，数値による変数の質的属性への変換（Discretize）は目的変数（クラス）を利用しないため，unsupervisedのグループに属するフィルタを適用する．一方，クラスを利用するデータ加工処理は，supervisedのグループに属する．

以上の変換および処理手順は，Explorerの前処理パネルでの操作を，バッチファイルなどとしてスクリプト化することで，自動化が可能である．

6．各アルゴリズムの実行

Wekaにおいてマイニング処理の実行は，分類（classification），クラスタリング（clusterer），相関（association）の各クラスで実装されている．各区分は，クラス階層，エクスプローラ，ナレッジフローで共通して表現されている．

エクスプローラで分類モデルの生成を行うには，“前処理”パネルでデータを読込み後，“分類”パネルで「分類器」のリストからアルゴリズムを選択して実行する．各アルゴリズムのパラメータは，アルゴリズム名を選択後に横のエリアをクリックして表示されるウィンドウにて行う．「テストオプション」は，実行時の評価データや評価方法を選択するものであり，テストデータを与えるためには学習データと属性情報がすべて同じである必要がある*3．実行は，「開始（Start）」ボタンを押して行い，結果は“分類器出力”のテキストエリアにテキスト形式で出力される．このテキスト出力は，コマンドラインで各マイニングアルゴリズムを実行したときと同一である．**図6**に示すように，結果リストからは，テキスト出力に加え，決定木の木構造表現や分類予測結果の散布図などを得ることができる．

以上の処理は，コマンドラインからも実行可能であり，訓練データ（-t），n回交差検証の回数（-x），テストデータ（-T），クラスとする属性の番号（-c）などの共通オプション

*1　キー属性によるJOIN処理は用意されていないため，事前にデータを一つのファイルとする必要がある．

*3　未知データでは，仮にクラスラベルを与えておく．


```
#学習(訓練)データをARFFに変換
$ java weka.core.converters.CSVLoader Letter.csv > letter.arff

#学習(訓練)データで決定木を評価(-Tでテストデータを指定)
$ java weka.classifiers.trees.J48 ¥
                    -t letter_nom.arff -T letter_nom.arff

#学習(訓練)データの属性値(数値)を名義値として列挙
$ java weka.filters.unsupervised.attribute.NumericToNominal ¥
                    -i letter.arff -o letter_nom.arff

#決定木学習を実行(パラメータを変更せず，10回交差検証を行う)
$ java weka.classifiers.trees.J48 -t letter_nom.arff

#決定木学習の分類予測結果を学習データに付与する
$ java weka.filters.supervised.attribute.AddClassification ¥
        -i letter_nom.arff -o letter_nom2.arff -c last -classificationn ¥
        -W "weka.classifiers.trees.J48"
```

図7　letter.csvに対するJ4.8の実行例

の他に，各アルゴリズムの固有オプションを指定して実行する．

なお，分類予測あるいはクラスタリングの結果を入力ファイルに追加する方法は，フィルタとしてそれぞれAddClassification，AddClusterが用意されている．これらのフィルタでは，入出力ファイルの指定方法は，他のフィルタと同様だが，分類予測やクラスタリングへの引数指定は-Wオプション後に" "で囲って行う．

以上の実行を，UCI機械学習共通データセット[6]のLetter Recognition（16属性，26クラス，2万行）をCSV形式にして実行したとすると，**図7**のようになる．

8. むすび

本稿では，データマイニングツールWekaの概要と，主に対話的なマイニング関連処理の適用方法と，コマンドラインインタフェースを利用した実行方法を解説した．これは，映像データを対象とした研究システムなどで適応的にマイニング処理を組込むことを可能にするものと考えられる．

以上に加え，Wekaのエクスプローラのビジュアル化パネルは，データ分析の初歩的な段階の利用者がデータの概観を把握し，フィルタによる加工処理において，それぞれの処理内容とその結果を理解するために役立つ．また，データ分析の流れを把握している利用者には，商用ツールに見られる実行フローによる実行環境が提供されている．

なお，紙面の都合上，紹介しきれなかった利用法については，筆者が運営するWebページ（http://weka-jp.info/）を参照されたい．

（2011年7月29日受付）

〔文　献〕

1) H.I. Witten, E. Frank: "Data Mining: Practical Machine Learning Tools and Techniques（3rd Edition）", Morgan Kaufmann（2005）
2) KNIME ¦ Konstanz Information Miner, http://www.knime.org/
3) Rapid-i, http://rapid-i.com/content/view/181/196/
4) Pentaho, http://www.pentaho.com/
5) KDNuggets, http://www.kdnuggets.com/
6) University of California, School of Information and Computer Science, UCI Machine Learning Repository, http://archive.ics.uci.edu/ml/

阿部（あべ）　秀尚（ひでなお）　2000年，静岡大学情報学部卒業．2004年，同大学院理工学研究科博士課程修了．同年，島根大学医学部医療情報学講座助手．2007年，島根大学医学部医療情報学講座助教．現在に至る．第26回医療情報学連合大会若手奨励賞受賞．博士（工学）．