PLAIN News
宇宙航空研究開発機構 宇宙科学研究本部
科学衛星運用・データ利用センター
http://www.isas.jaxa.jp/docs/PLAINnews/
SODA ISAS/JAXA
Center for Science-satellite Operation and Data Archive
2009.11.27 / No. 193

衛星データとデータマイニング
第２回 データマイニングの実際

本田 理恵（高知大学自然科学研究系理学部門 /JAXA 宇宙科学情報解析研究系 客員）

1. はじめに

前回のデータマイニングの概略の説明に引き続き，今回はオープンソースソフトウェアの Weka を用いた学習事例を交えながら代表的な手法について具体的に紹介します。事例としては、はじめての方に衛星データへの活用のイメージをつかんでいただくために、簡潔な例を用いましたが、教科書的記述が多くなってしまったことにはお詫びいたします。また、紙面の制約のため図が小さく見にくくなってしまいましたが，web 版には大きめに掲載されていますのでそちらもご参照ください。

2. 決定木学習

機械学習やデータマイニングでは，“分類”に属する問題が数多く扱われます。“分類“は,複数の属性を持つデータにおいて，ある属性を目的属性，それ以外の属性を従属属性として，従属属性だけから目的属性を決定する問題としてとらえることができます。機械学習の分野では，目的属性をラベルまたはクラス，従属属性をパターンと呼ぶこともあります。

表１に,分類で扱われるデータセットの一例を示します。ここでは，衛星のテレメトリデータを想定して，搭載センサの誤動作の有無，電圧，温度，他のセンサの動作の有無を属性とします（注1）。このようなデータセットから，センサの誤動作がおこる条件を知りたい，という要求があるものとします。この場合,衛星の誤動作の有無を目的属性,それ以外を従属属性とすることになります。

表１　データセットの例-センサ誤動作情報ー

温度	電圧 A	電圧 B	他センサ	誤動作
高	高	高	無	有
高	高	高	有	有
中	高	高	無	無
低	中	高	無	無
低	低	中	無	無
低	低	中	有	有
中	低	中	有	無
高	中	高	無	有
高	低	中	無	無
低	中	中	無	無
高	中	中	有	無
中	中	高	有	無
中	高	中	無	無
低	中	高	有	有

“分類“を扱う手法には,ナイーブベイズ,サポートベクターマシンなど種々のものがありますが，ここでは決定木学習を取り上げます。表１のデータに対して求められた決定木の例を図１に示します。決定木の根または節には従属属性に対する条件式が割り当てられます。一方，葉には目的属性が割り当てられます。よって，目的属性の値が未知で従属属性が既知のデータが与えられた場合，従属属性の値を参照しながら決定木を根から葉へたどっていくことによって，この事例の目的属性値を推定することができます。

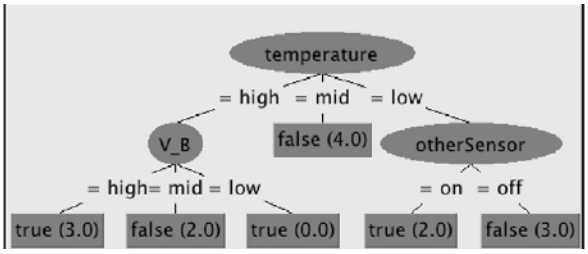

図１．表１のデータに対して得られた得られた決定木の例。
括弧内数字は実際に分類された事例数。

このような木構造を，データからの学習によって形成しようとするのが決定木学習です。決定木学習では，まず全データに対して考えられる全ての分割条件を用いてそれぞれ分割を試行し，目的属性に関しての分割前と分割後のデータセットの多様性指標の変化量を計算します。ここで，多様性指標の減少量がもっとも大きくなる条件を最初の分割条件として採用してデータセットを分割します。この過程を分割後の各データセットが目標属性について均一になるか，それ以上の分割条件がなくなるまで再起的に実施していきます。多様性指標には様々な物が有りますが，代表的なアルゴリズムである Quinlan の C4.5 や C5.0 では情報エントロピーに基づく情報利得比が用いられています。

また，木構造は，if then 形式のルールで表現することもでき，このルール自体が現象に関する知識であって，決定木学習によってこの知識を発掘することができた，と解釈することもできます。図１の決定木では，“if temperature=high and V_B=high then sensor failure=true, if temperature = low and otherSensor=active then failure=true となり，このセンサが不具合を起こすのは，温度が高く B 点の電圧が高いとき，または，温度が低く他のセンサが動作しているとき，ということになります。

3. 相関分析

発生した事象の共起関係を調べたい場合には相関分析が有用です。相関分析は，トランザクションに現れるアイテム間の共起関係を分析します。顧客の購買分析とのアナロジーで説明すると,トランザクションは顧客の買い物籠(バスケット)，アイテムはバスケットにはいった商品としてとらえることができます。

このようなデータから相関ルールと呼ばれる共起関係を抽出します。相関ルールは A → B という形で表記され，A という事象が発生すれば高い確率で B という事象も発生することを示します。相関ルールの重要性の指標には支持度(support)，確信度（confidence）があります。Support(A → B) は全トランザクション数に対する A と B が共起するトランザクション数の割合を表し，confidence(A →

［裏へ続く］

B) はAを含むトランザクション数に対するAとBが共起するトランザクション数の割合を示します。よってconfidence(A → B) と Support(A → B) が大きいルールが重要なルールということになります。

ルールの抽出にあたっては，アイテムの数が増加することによって組み合わせ数が膨大になって計算が困難になるという問題がありますが，IBMアルマディン研究所のAgrawalによって開発されたAprioriアルゴリズムによって大規模データから効率よくルールを抽出することが可能になっています。

４．クラスタリング

この他，一般的な用途によく用いられる手法としてクラスタリングがあります。クラスタリングは，数値やカテゴリ値からなる多次元ベクトルを対象とし，類似度に基づいたグループ分けを行う手法です。数値属性のベクトルについては，類似度の代わりに相違度としてベクトル間のユークリッド距離がよく使用されます。クラスタリングは，決定木学習のようにあらかじめ目標属性のようなお手本が与えられないため，教師無し学習の一つです。

具体的な手法には，階層的な手法と非階層的な手法があります。階層併合的クラスタリングは，１データを１クラスタに割り当てた状態を初期状態とし，データを相違度の小さいものから順番に併合して最終的に一つのグループになるまで反復します。併合過程のデータ集合をまとめたデータ集合は樹状図をなし，様々な階層レベルでグループを評価することができます。

一方，非階層的な手法にはK-means法などがあります。K-means法ではあらかじめグループの個数kを定め，K個のクラスタの重心の初期値をランダムに与えます。次に各データについて，全てのクラスタ重心との距離を計算して最も距離のクラスタにデータを割り付けます。全てのデータの割り付けの完了後，各クラスタの属するデータセットからクラスタの重心を決定し直します。この過程を決められた回数，またはデータの重心の変化がなくなるまで反復することによって，尤もらしいグループ分けを得ることができます。

K-means法は非常に明快なアルゴリズムですが，Kをあらかじめ与えなければならない，初期値によって結果が大きく異なる，といった欠点も持ちます。使用時にはこうした特徴に対する注意が必要です。

この他，多変量正規分布の混合モデルでの近似もよく用いられます。学習（パラメータ推定）には EMアルゴリズムが用いられます。

なお，数値属性データではクラスタリングの実施前に，異なる属性値間の寄与を等しくするために各属性の間で正規化を行うことが必要な場合も有ります。

５．Wekaによるデータマイニング実例

それでは前回紹介したニュージーランド大で開発されたデータマイニングや機械学習のための統合的ソフトウェアWekaを使用して，前述の代表的な手法によるデータの学習の実例を紹介します。

インストール　まずは，http://www.cs.waikato.ac.nz/ml/wekaから使用環境に対応したソフトウェアをダウンロードしてインストールします。2009年11月現在，最新のstable バージョンは3.6.1で，Windows，Mac OS，Linuxなどの OS向けの実行形式が準備されています。実行にあたってはJava VM 1.5以降を必要としますので適宜準備をしてください。なお，上記サイトにはチュートリアル，FAQ，サンプルデータ，メーリングリストへの参加法など各種情報が掲載されています。

データの準備　Wekaの入力データの標準的なフォーマットはARFFです。図２に表１のARF形式のファイル”sensor.arff”を示します（ファイルの拡張子にarffをつけます)。ARFFファイルはヘッダ部とデータ部からなり，ヘッダ部は@dataまでです。

ヘッダ部の@relationの後にはテープルの名前を，@attributeの後には属性の名前とタイプを記述します。カテゴリ値の場合は{high, mid, low}のように列挙して示します。数値属性の場合は,”@attribute temperature numeric”のように記述します。なお，数値属性のタイプにはreal, integer, numericを指定することができます。データ部はデータ毎に各行に記述された“,”で区切られた属性値の羅列であり，CSVファイルのフォーマットと同じとなっています（属性値の順番はヘッダに記述された属性の順番に従います)。よって，CSVファイルの先頭にヘッダを不可するだけで，入力ファイルを簡単に作成することが可能です。

```
@relation sensor
@attribute temperature {high, mid, low}
@attribute V_A{high, mid, low}
@attribute V_B{high, mid, low}
@attribute otherSensor {on,  off}
@attribute failure {true, false}
@data
high,high,high,off,true
high,high,high,on,true
mid,high,high,off,false
low,mid,high,off,false
low,low,mid,off,false
low,low,mid,on,true
mid,low,mid,on,false
high,mid,high,off,true
high,low,mid,off,false
low,mid,mid,off,false
high,mid,mid,on,false
mid,mid,high,on,false
mid,high,mid,off,false
low,mid,high,on,true
```

図２　入力用データ sensor.arff の例（ARFF 形式）

起動　Wekaはインストールディレクトリにある鳥のイメージの Weka3-6-1というアイコンをクリックすることで起動できます。起動画面には”Explorer, Experimenter, KnowledgeFlow, Simple-CLI”の４つのボタンが表示されますが，ここではGUIを使用して対話的に処理を行うことができる”Explorer”を選択します。各処理は最上部のボタンで選択できるようになっており，それぞれ，Preprocess(前処理)，Classify（分類，決定木学習を含む)，Cluster（クラスタリング)，

Associate（相関分析），Select attribute（属性選択），visualize（可視化）となっています（図3参照）。なお，起動時にはPreprocessのみ選択可能となっています。

データ読み込み　まずPreprocess（前処理）を選択して，上から２段目左側の“Open file...”ボタンをクリックしてarff（ここでは”sensor.arff”）ファイルを指定して読み込みます。

図3に読み込み直後の画面を示します。この画面ではデータの属性や分布についての特徴が表示されます。右下に表示されたヒストグラム左上のリストで目標属性（クラス）を選択して，右側の”visualize all”ボタンをクリックすると，各属性値のデータに占める目標属性値（クラス）毎のデータ分布を，色付きヒストグラムで一覧することができます。

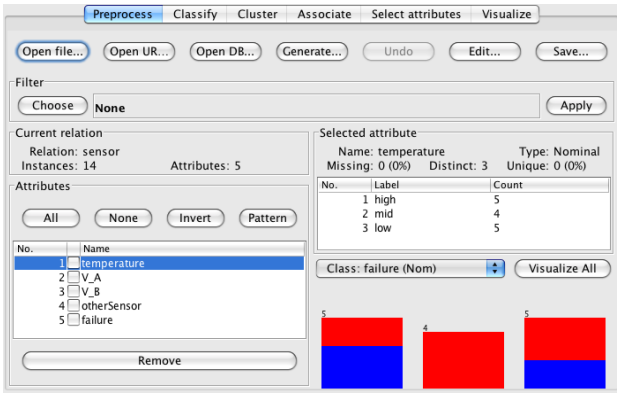

図3　Weka前処理画面例

決定木学習　決定木学習を実施するには，最上段の“Classify”ボタンを選択して図4のような分類画面に切り替えます。次に２段目の”Classifier”欄からtrees，J48（C4.5に相当）を選択します。

左側の”Test option”では学習の方法として交叉検定法などが選択できますが，ここではひとまず全データを使用して学習を行う”Use training set”を選択します。詳細なオプションはその下の”More options”で指定することができます。また，その下のリストで，目的属性を選択します。ここではセンサの誤動作を示す“failure”を選択します。

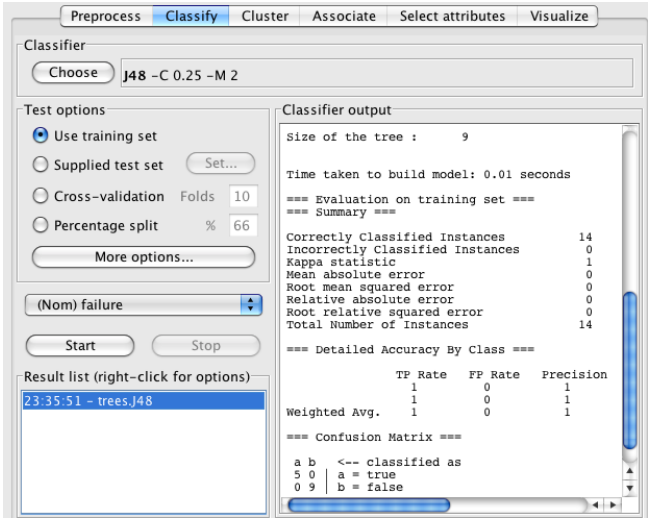

図4　決定木学習の実施画面例

次にその下側にある”Start”ボタンをクリックすると学習が開始されます。学習結果は得られた決定木に対する評価値とともに右側のウィンドウに表示されます。得られた決定木の尤もらしさは，TP（true positive rate），FP rate（false positive rate），precision，recallなどで表示されます。詳細は省略しますが，この場合の正答率は100%で性能のよい決定木が得られたことになります。

なお，作成された決定木を可視化するには，左下の学習結果のリスト“Result list”に追加された計算結果のリストを右クリックして“Visualize tree”を選択すると，図1のような決定木が別のウィンドウに表示されます。

相関分析　次に，決定木学習と同じデータで相関分析を実施します。図2のデータは本来の相関分析が扱うデータとは少し異なりますが，行をトランザクション，観測される属性値をアイテムと解釈することによって相関分析を行うことが可能です。

最上段の”Associate”をクリックして相関分析用のウィンドウに切り替えます。次の段の”Associator”をクリックして“apriori”を選択します。”apriori……”と表示されたリストの欄をクリックすると，抽出時のsupportの敷居値などを設定することができます（デフォルトでは0.2）。“Start”ボタンを押すと計算が実行され，右ウィンドウに以下のような抽出された相関ルールが表示されます。なお，上記のルールの確信度は1，支持度は14/3で重要性の高いルールといえます。

1. temperature=mid 4 ==> failure=false 4
   conf:(1)（温度が中程度であれば誤動作なし）
2. V_A=low 4 ==> V_B=mid 4　conf:(1)
   （Aの電圧が低いときBの電圧も低い）（以下略）

クラスタリング　最後にクラスタリングを実施します。データにはWekaのインストールディレクトリの下のdataディレクトリに入っている“iris.arff”を利用します。このデータセットには，アヤメの属性（花弁の幅，長さなど4種，数値）がアヤメの種類とともに格納されています。単純なデータではありますが，天文学，惑星科学，地球観測などの分野で一般的である各種スペクトルのグループ化も，この実験から容易にイメージすることが可能と思います。

まず”Preprocess”ボタンを押して，”Open file”でdataの“iris.arff”を指定します。数値属性の場合，必要であれば前処理を実施します。属性ごとの正規化を実施するには，Filter欄の“Choose”ボタンを押して，filters，unsupervised，attributeを順に指定して，”standarize”を選択し，”apply”ボタンを押します。これによって，各属性値は属性毎に0を平均値として標準偏差が1になるように正規化されます。

次に最上部の”Cluster”ボタンをおして，図5のクラスタリング画面に切り替え，“Choose”からアルゴリズムを選択します。ここでは簡単のため”simpleKmeans”を選択します。選択したアルゴリズムが表示されているリスト欄をクリックするとクラスタ数（numClusters，k）などのパラメータを入力するこ

［裏へ続く］

とができます。今回はデータセットの記述からアヤメの種類が 3 個であることが既知なので，numClusters=3 とします。また，学習の際にアヤメの種類をマスクしてクラスタリング結果と比較するために，左側の Cluster mode から“Classes to clusters evaluation”をクリックして，リストで class（アヤメの種類）を指定します。

Start ボタンをクリックすると処理が始まり，右画面にその結果が表示されます。最下部に表示された各データのクラスタ番号とマスクしたアヤメの種類の比較結果から，誤分類は 17/150, 11%で，教師データを与えなくても，まずまずのグループ化ができたたことがわかります。

なお,”Result list” を右クリックして,”visualize cluster assignments”を指定すると，属性毎のクラスタ分布が可視化されます。データ毎のクラスタリング結果を見るには，この画面の save ボタンをクリックして保存用のファイル名を指定してエクスポートします。

図 6 にクラスタリングの結果をクラスタ毎に色を変えたスペクトル形式で表示してみます（エクスポートしたデータからプロット）。この図からもスペクトルの形状に応じた自動的なグループ化がなされていることを見て取ることができます。

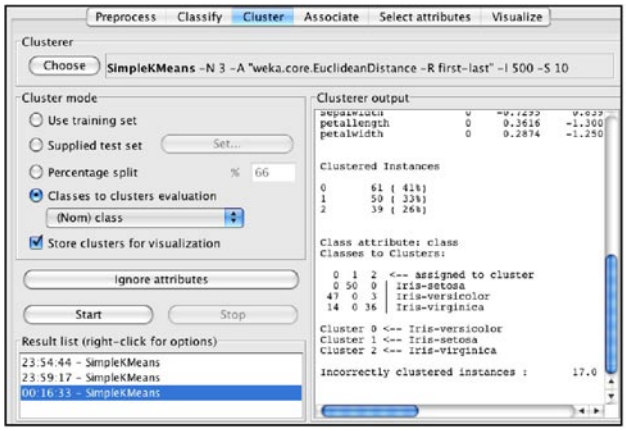

図 5　クラスタリングの実行画面例

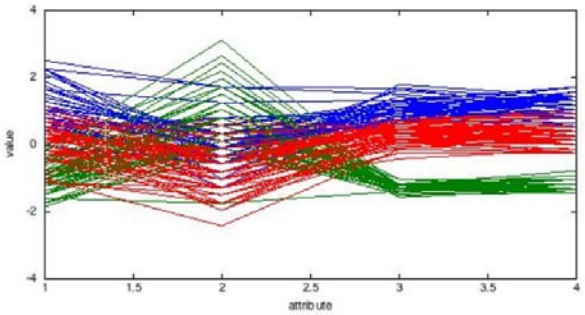

図 6　クラスタリング結果のスペクトル形式での表示（クラスタごとに色づけ）

６．まとめ

今回は代表的な３つの手法，決定木学習，相関分析，クラスタリング（k-means 法）について，toy data に対して Weka を使用した分析例を含めながら説明しました。使用したデータはまさに”toy-おもちゃ”レベルですが，ここで取り上げたような分析のニーズは，様々なシチュエーションで頻出するものであると考えられます。データマイニングは知識の発見という大きな目的を目指すものではありますが，分析指針が定まらず眠っているデータや，煩雑な手作業の自動化などにも活用可能ですので，お試しいただけると幸いです。

次回は，少し進んだモデルや，実際の活用例などについて紹介する予定です。

参考文献等 (1) 元田浩ほか, データマイニングの基礎, オーム社，2006

注1　文献 (1) の掲載例 (ゴルフプレイデータ) を修正して利用。

2009 札幌 ADASS 報告

山内 千里（JAXA 宇宙科学情報解析研究系 /C-SODA）

ADASS とは，Astronomical Data Analysis Software & Systems の略称で，その名前が示すとおり，天文データを扱うためのソフトウェアに関する国際カンファレンスです．毎年開催されており，本年度は札幌で行なわれました．日本での開催は ADASS 史上初であり，多くの日本人開発者・研究者の参加がみられました．

ADASS は，自然科学系の研究会等と同様の口頭発表・ポスター発表のほか，フォーカスデモやフロアデモといった，ソフトウェアを実際に動かしながら紹介する発表もあり，その有用性をアピールする絶好の機会となっています．このほか，BoF といって，あるテーマに関心が高い人達が集まる小ミーティングもあり，毎年，FITS に関するテーマなどの会合があります．我々，C-SODA グループからは，宇宙科学データ公開サービス“DARTS”，次世代 FITS I/O ライブラリ“SFITSIO”をポスターで，天体ナビゲーションツールである“JUDO”,「すざく」XIS 用 QL ツール“UDON”をフォーカスデモで発表しました．“JUDO”と“UDON”のフォーカスデモは，昨年度も行なったのですが，それでも 30 人程度の方々に動作している様子を見てもらう事ができました．“JUDO”は現状ではナビゲーションツールとしてしか使われていませんが，今年度の開発によってより汎用的な QL 画像切り出しシステムに発展していく予定です．ご期待ください．

FITS I/O ライブラリ“SFITSIO”については，いずれ PLAIN ニュースでももっと誌面を使って紹介する予定ですが，CFITSIO よりも圧倒的に簡単に FITS ファイルが扱える C 言語プログラマ向きのライブラリです．今回の ADASS ではポスターの他，FITS の BoF でも座長のご好意で 5 分間の発表をさせてもらう事ができました．チラシとマニュアルのコピーも在庫ゼロとなり，予想以上の関心の高さに驚きました．

来年の ADASS はアメリカのボストンです．自分のソフトウェアをアピールするのにも，いろいろなソフトウェアに関する情報を集めるのにも，たいへん有意義なカンファレンスです．データ処理に深く関わっているそこのあなたも来年は参加してみませんか？

編集発行：宇宙航空研究開発機構 宇宙科学研究本部 科学衛星運用・データ利用センター
〒 229-8510 相模原市由野台 3-1-1　Tel. 042-759-8767　　住所変更等 e-mail：news@plain.isas.jaxa.jp
本ニュースはインターネットでもご覧になれます．http://www.isas.jaxa.jp/docs/PLAINnews

●編集後記:いつのまにか、相模原キャンパスは落ち葉で埋め尽くされています。ふと気づいたら、もう師走が目の前です！(K.E.)