情報意味論 (5) 回帰, Weka紹介(2)

櫻井彰人
慶應義塾大学理工学部

Weka

- ニュージーランドのワイカト大学が開発 (University of Waikato, New Zealand)
  - Waikato Environment of Knowledge Analysis の略
  - Weka: 探求心旺盛な飛べない鳥
- Java言語で記述
  - 機能追加可能
- フリーソフト
  - 営利目的以外には自由に使用可能。改変可
- 日本語化が比較的容易 (Javaがそうだから)
- 欠点：機能が少ない
  - 特に GUI (graphical user interface) が貧弱
  - 営利目的でない以上、ある程度は我慢すべし
  - 無保証 (これは商用ソフトも似たようなもの)

データの形式

```
@relation 天気とテニス

@attribute 天気予報 {晴, 曇, 雨}
@attribute 気温 real
@attribute 湿度 real
@attribute 風 {強, 弱}
@attribute テニス {行う, 止め}
```

天気とテニス.arff の内容

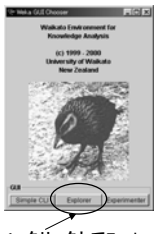
```
@data
晴,29,85,弱,止め
晴,27,90,強,止め
曇,28,86,弱,行う
雨,21,96,弱,行う
雨,20,80,弱,行う
雨,18,70,強,止め
曇,18,65,強,行う
晴,22,95,弱,止め
晴,21,70,弱,行う
雨,24,80,弱,行う
晴,24,70,強,行う
曇,22,90,強,行う
曇,27,75,弱,行う
雨,22,91,強,止め
```

Excel の表形式で書いたもの

天気予報	温度	湿度	風	テニス
晴	29	85	弱	止め
晴	27	90	強	止め
曇	28	86	弱	行う
雨	21	96	弱	行う
雨	20	80	弱	行う
雨	18	70	強	止め
曇	18	65	強	行う
晴	22	95	弱	止め
晴	21	70	弱	行う
雨	24	80	弱	行う
晴	24	70	強	行う
曇	22	90	強	行う
曇	27	75	弱	行う
雨	22	91	強	止め

Wekaの初期画面

- weka.jar または weka.bat



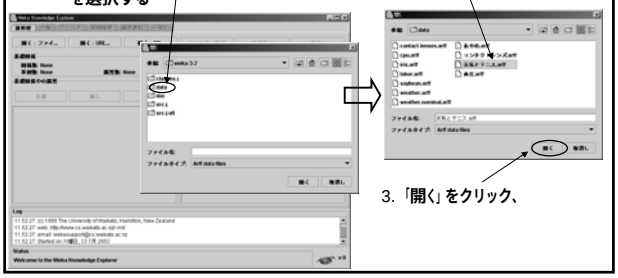
1. クリックしてExplorerを起動

2. クリックしてデータファイルを選択する

対象データファイルの指定

1. クリックしてdataフォルダを選択する

2. クリックして天気とテニス.arffファイル (どこかにある) を選択し、



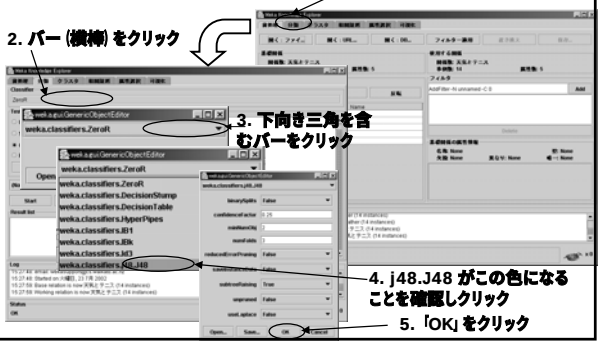
3.「開く」をクリック、

決定木の作成 (計算)

1.「分類」をクリック

2. バー (横棒) をクリック

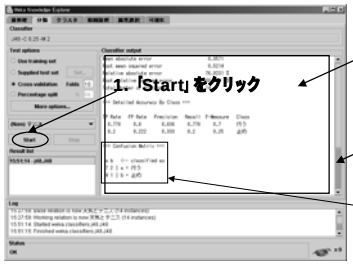
3. 下向き三角を含むバーをクリック



4. j48.J48 がこの色になることを確認しクリック

5.「OK」をクリック

結果の確認



1. 「Start」をクリック

2. 結果はこのウィンドウに表示される

3. このバーを上にドラッグすると、最初の方が見れる

余談：作成した木についてではなく、約半分のサンプルで決定木を作成し、残りのサンプルでテストした結果

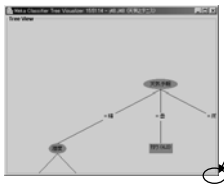
結果の確認と図示

1. 決定木を文字列で表現したもの

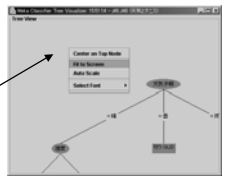
2. この上で「右」クリック

3. 「Visualize tree」の上でクリック

図示された木の変形

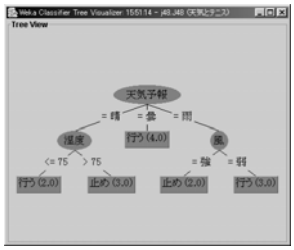


1. マウスカーソルをこの角にもってくると↔に変わる。その状態でドラッグすると、このウィンドウの形・大きさが変更できる



2. このスクリーン上で「右」クリック。Fit to Screen をクリックすると、スクリーンの大きさにあった大きさの木になり、Auto Scale でクリックすると木がコンパクトになる。文字の大きさを変えるには Select Font でクリック
木をドラッグすることもできる

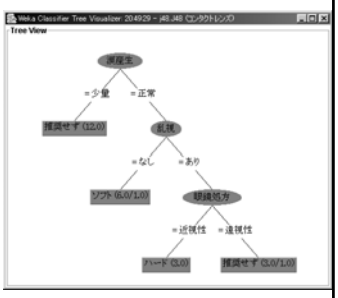
決定木の例



天気予報が雨であれば
　そして風が強ければ、止め
　風が弱ければ、行う
天気予報が曇りであれば、　行う
天気予報が晴れだれば
　そして湿度が75%より高ければ、止め
　湿度が75%以下であれば　行う

コンタクトレンズの例

年齢	眼鏡処方	乱視	涙産生	コンタクトレンズ
若年期	近視性	なし	少量	推奨せず
若年期	近視性	なし	正常	ソフト
若年期	近視性	あり	少量	推奨せず
若年期	近視性	あり	正常	ハード
若年期	遠視性	なし	少量	推奨せず
若年期	遠視性	なし	正常	ソフト
若年期	遠視性	あり	少量	推奨せず
若年期	遠視性	あり	正常	ハード
前老眼期	近視性	なし	少量	推奨せず
前老眼期	近視性	なし	正常	ソフト
前老眼期	近視性	あり	少量	推奨せず
前老眼期	近視性	あり	正常	ハード
前老眼期	遠視性	なし	少量	推奨せず
前老眼期	遠視性	なし	正常	ソフト
前老眼期	遠視性	あり	少量	推奨せず
前老眼期	遠視性	あり	正常	推奨せず
老眼期	近視性	なし	少量	推奨せず
老眼期	近視性	なし	正常	推奨せず
老眼期	近視性	あり	少量	推奨せず
老眼期	近視性	あり	正常	ハード
老眼期	遠視性	なし	少量	推奨せず
老眼期	遠視性	なし	正常	ソフト
老眼期	遠視性	あり	少量	推奨せず
老眼期	遠視性	あり	正常	推奨せず



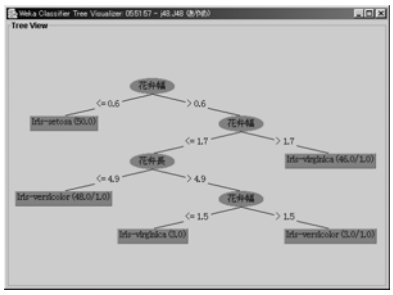
古典的問題：あやめの分類

- 萼片長、萼片幅、花弁長、花弁幅とあやめ（setosa, versicolor, virginica の3種）の値が150組。
- Fisher, R. A. (1936). The Use of Multiple Measurements in Axonomic Problems. Annals of Eugenics 7, 179-188.

萼片長	萼片幅	花弁長	花弁幅	種別
5.1	3.5	1.4	0.2	Iris-setosa
4.9	3	1.4	0.2	Iris-setosa
4.7	3.2	1.3	0.2	Iris-setosa
4.6	3.1	1.5	0.2	Iris-setosa
5	3.6	1.4	0.2	Iris-setosa
5.4	3.9	1.7	0.4	Iris-setosa
4.6	3.4	1.4	0.3	Iris-setosa
5	3.4	1.5	0.2	Iris-setosa
4.4	2.9	1.4	0.2	Iris-setosa

（横軸：萼片長、縦軸：花弁幅）

分類結果



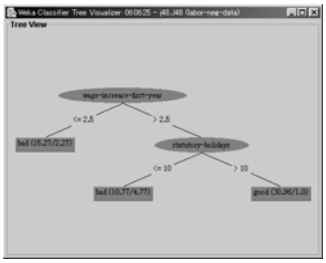
労使間交渉の決着状況

- カナダ労使間交渉の決着状況を、賃金・手当等との組みで表したもの
- 欠損値が多い（ごく普通の状況）：理論的・アルゴリズム的に困難な課題

属性	型	1	2	3	…	40
継続期間	(年数)	1	2	3		2
賃上げ（第1年）	百分率	2	4	4.3		4.5
賃上げ（第2年）	百分率	?	5	4.4		4
賃上げ（第3年）	百分率	?	?	?		?
生活費保証	(none, tcf, tc)	none	tcf	?		none
労働時間/週	時間数	28	35	38		40
年金	(none, ret-allw, empl-cntr)	none	?	?		?
stand-by pay	百分率	?	13	?		?
変則勤務手当て	百分率	?	5	4		4
教育手当て	(あり、なし)	あり	?	?		?
土曜休業	休日数	11	15	12		12
休暇	(平均以下、平均、平均以上)	平均	平均以上	平均以上		平均
長期傷害助成	(あり、なし)	なし	?	?		あり
歯科診療保険助成	(なし、半分、完全)	なし	?	完全		完全
死別助成	(あり、なし)	なし	?	?		ari
健康保険助成	(なし、半分、完全)	なし	?	完全		半分
対応	(良い、悪い)	悪い	良い	良い		良い

（縦横がこれまでと逆なので注意）

労使間交渉データの結果

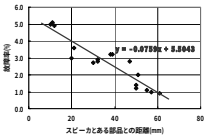


線型回帰

- Excel 付属のツールより多機能
  - 説明変数にカテゴリ変数があること、
  - 一次式（直線）で説明できない場合を扱うことが特徴
  - GUI あり
- Excel の例：あるカラーテレビの故障率

あるカラーテレビの故障率

No	「E」かとある部品との距離(mm)	故障率(%)
1	10	5.0
2	11	5.1
3	55	1.1
4	50	1.2
5	12	4.9
6	20	3.0
7	30	2.7
8	32	2.8
9	61	0.9
10	51	2.0
11	32	2.9
12	50	1.4
13	38	3.2
14	47	2.8
15	57	1.0
16	21	3.6
17	39	3.2



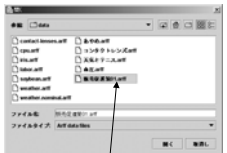
回帰統計	
重相関 R	0.93399
重決定 R2	0.872337
補正 R2	0.863826
標準誤差	0.508842
観測数	17

分散分析表

	自由度	変動	分散	観測された分散比	有意 F
回帰	1	26.53854	26.53854	102.4968	4.25E-08
残差	15	3.88381	0.258921		
合計	16	30.42235			

	係数	標準誤差	t	P-値	下限 95%	上限 95%	下限 95.0%	上限 95.0%
切片	5.504299	0.298473	18.44151	1.02E-11	4.868118	6.14048	4.868118	6.14048
X 値 1	-0.07593	0.0075	-10.1241	4.25E-08	-0.09192	-0.05994	-0.09192	-0.05994

ファイルの選択

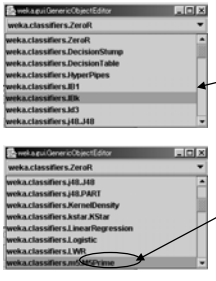


1. 販売促進01.arffファイル（どこにある）クリック、

```
@relation '販売促進策01'
@attribute 月 real
@attribute 日 real
@attribute 曜日 {日, 月, 火, 水, 木, 金, 土}
@attribute 天候 {晴, 雨, 曇り}
@attribute 客数 real
@attribute 備考 {オートコール, 通常}
@data
7, 1, 金, 曇り, 491, 通常
7, 2, 土, 雨, 432, 通常
7, 3, 日, 晴, 514, 通常
7, 4, 月, 晴, 457, 通常
7, 5, 火, 曇り, 451, 通常
7, 6, 水, 雨, 441, 通常
7, 7, 木, 雨, 604, 通常
7, 8, 金, 曇り, 467, 通常
7, 9, 土, 晴, 408, 通常
7, 10, 日, 雨, 457, 通常
7, 11, 月, 雨, 484, 通常
7, 12, 火, 雨, 506, 通常
7, 13, 水, 曇り, 474, 通常
7, 14, 木, 晴, 666, 通常
7, 15, 金, 雨, 479, 通常
7, 16, 土, 曇り, 478, 通常
7, 17, 日, 晴, 640, 通常
7, 18, 月, 晴, 497, 通常
7, 19, 火, 晴, 473, 通常
7, 20, 水, 晴, 468, 通常
7, 21, 木, 晴, 875, オートコール
7, 22, 金, 晴, 829, オートコール
```

月	日	曜日	天候	客数	備考
7	1	金	曇り	491	通常
7	2	土	雨	432	通常
7	3	日	晴	514	通常
7	4	月	晴	457	通常
7	5	火	曇り	451	通常
7	6	水	雨	441	通常
7	7	木	雨	604	通常
7	8	金	曇り	467	通常
7	9	土	晴	408	通常
7	10	日	雨	457	通常
7	11	月	雨	484	通常
7	12	火	雨	506	通常
7	13	水	曇り	474	通常
7	14	木	晴	666	通常
7	15	金	雨	479	通常
7	16	土	曇り	478	通常
7	17	日	晴	640	通常
7	18	月	晴	497	通常
7	19	火	晴	473	通常
7	20	水	晴	468	通常
7	21	木	晴	875	オートコール
7	22	金	晴	829	オートコール
7	23	土	晴	597	通常
7	24	日	雨	633	通常
7	25	月	曇り	476	通常
7	26	火	晴	480	通常
7	27	水	晴	408	通常
7	28	木	晴	544	通常
7	29	金	雨	365	通常
7	30	土	晴	380	通常
7	31	日	晴	448	通常

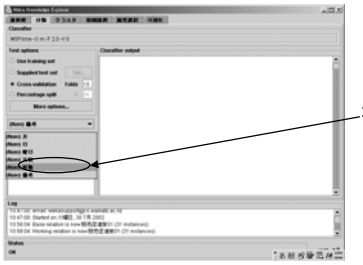
使うアルゴリズムの選択



1. 右にあるこの緑バーをクリックして、メニュー（プルダウンメニュー）の下の方をみる

2. M5Prime というのを選択する

被説明変数の指定



1. 「**客数**」の上でクリック

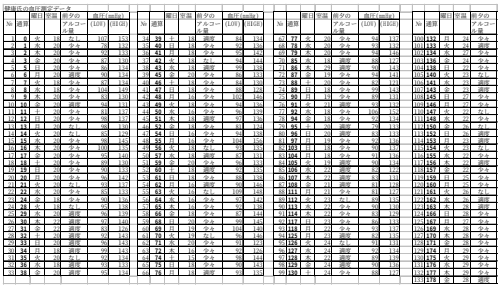
黙っているとデータ(表)のなかの最も右の属性が用いられる。
今回は、「最も右」ではないのでここで指定する

結果の解析

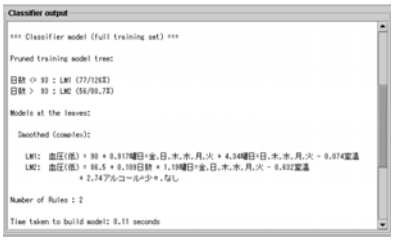
客数 = 490
+ 362**備考**=**オートコール**

オートコールを行った方が客数が増加することがわかる

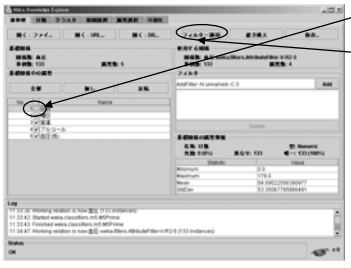
血圧の測定データ



Weka による分析結果



日数をはずす

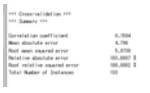


1. **日数のチェックボックスのチェックをはずす**

2. **フィルター適用をクリックする**

3. **「分類」で M5Prime を Start**

日数をはずした場合の結果



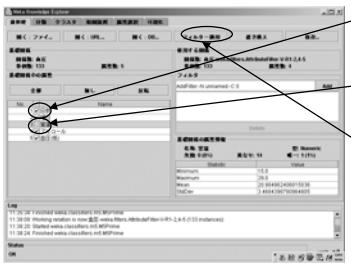
室温 <= 19.5 : LM1 (46/122%)
室温 > 19.5 : LM2 (87/114%)

LM1: **血圧(低)** = 91.9
+ 1.38**曜日**=**金**,**日**,**木**,**水**,**月**,**火**
+ 4.02**曜日**=**水**,**月**,**火**
- 0.112**室温**

LM2: **血圧(低)** = 90.8
+ 0.827**曜日**=**金**,**日**,**木**,**水**,**月**,**火**
- 0.0668**室温**

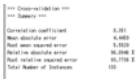
Correlation coefficient 0.1594

室温をはずす



1. **日数のチェックボックスのチェックをつける (元に戻したことになる)**
2. **室温のチェックボックスのチェックをはずす**
3. **フィルター適用をクリックする**

室温をはずした場合の結果



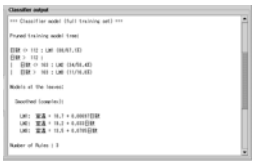
日数 <= 93 : LM1 (77/126%)
日数 > 93 : LM2 (56/84.3%)

LM1: **血圧(低)** = 88.6
+ 0.782**曜日**=**金**,**日**,**木**,**水**,**月**,**火**
+ 4.34**曜日**=**日**,**木**,**水**,**月**,**火**

LM2: **血圧(低)** = 79.4
+ 0.0544**日数**
+ 1.01**曜日**=**金**,**日**,**木**,**水**,**月**,**火**
+ 2.72**アルコール**=**少々**,**なし**

Correlation coefficient 0.351

日数と室温との関係



日数 <= 112 : LM1 (88/67.1%)
日数 > 112 :
| **日数** <= 163 : LM2 (34/58.4%)
| **日数** > 163 : LM3 (11/16.8%)

Models at the leaves:
Smoothed (complex):

LM1: **室温** = 18.7 + 0.00697**日数**
LM2: **室温** = 19.2 + 0.033**日数**
LM3: **室温** = 13.5 + 0.0785**日数**

Correlation coefficient 0.8517

日数と室温をはずすと

Smoothed (complex):
LM1: 血圧(低) = 91.3
Number of Rules : 1
Time taken to build model: 0.03 seconds
=== Cross-validation ===
=== Summary ===
Correlation coefficient -0.1528
Mean absolute error 4.8475
Root mean squared error 6.2984
Relative absolute error 104.7744 %

残りの属性 (曜日と前日のアルコール摂取量) ではうまく説明できないことがわかる

「血圧」の総合的な結論

- **日数がたつにつれ、血圧が上昇している**
- **しかし、それは日数がたったからか、気温が上昇したからかはわからない**
- **土曜日に低い傾向はあるが、確信できず**
- **前日のアルコール摂取量で低い傾向はあるが、確信度はもっと低い**