統計的テキスト解析（14）
～ テキストの分類① ～

同志社大学文化情報学部教授
金　明哲（Jin Mingzhe）

■中国生まれ。総合研究大学院大学数物研究科統計科学専攻博士後期課程修了。博士（学術）。1995年札幌学院大学社会情報学部、助教授、教授を経て、2005年４月より現職。E-mail：mjin@mail.doshisha.ac.jp

1. テキストの分類

予めテキストをカテゴリ化したデータを用いてカテゴリを識別するルール（モデル）を作成し、そのルールに基づいてカテゴリが未知であるテキストをいずれかのカテゴリに振り分けることをテキストの分類（text classification）、あるいはテキストのカテゴライゼーション（text categorization）と呼ぶ。予めカテゴリ化したデータを用いてカテゴリを識別するルールを作成することを、学習データに基づいた機械学習と呼び、学習の結果を用いてデータを分類することを予測と呼ぶ。また、分類のアルゴリズム、あるいは分類システムを分類器（classifier）と呼ぶ。

テキストマイニングの目的のひとつは、電子化された大量のテキストを自動的に分類することである。例えば、インターネット上の大量のテキストを何らかの特徴に基づいて分類すること、コールセンターに寄せられた顧客の“声”の内容を要望、批判、質問などに分類することなどが考えられる。

計量文体分析の分野では、著者不明の文章の書き手の判別などに、テキストを著者別に分類する方法が古くから用いられている。例えば、1959年にCox and Brandwoodは、作品における文末の５つの音節に関する32のパターンの数を集計し、プラト（Plato）の作品の判別分析を行った。1963年にはMosteller and Wallaceは、著者の特徴が現れると考えられる単語の使用頻度を用いて、ハミルトン（Alexander Hamilton）の作品とマディソン（James Madison）の作品について著者の判別分析を行った。日本語の作品についての分析では、1965年に菫沢 正が江戸時代に書かれた『由良物語』について、特徴となる語の使用頻度に基づいて著者の判別分析を行った。

今日のテキストマイニングにおけるテキスト分類は、このような計量文体学におけるテキストの判別分析の手法を用いた著者の推定に関する研究のアプローチと基本的には同じである。異なるのはテキストの中から抽出して分類の情報として用いる要素である。

2. 分類の方法

データに基づいて個体を分類する方法は、古典的な統計学では判別分析と呼ぶ。判別分析の中で最も基本的な方法は線形判別分析である。学習データの集合$(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \cdots, (\mathbf{x}_n, y_n)$があるとする。$\mathbf{x}=(x_1, x_2, \cdots, x_p)$は、テキストから集計した$p$個の項目である。このひとつひとつの項目を独立変数と呼ぶ。$\mathbf{y}=(y_1, y_2, \cdots, y_n)$は、それぞれのテキストが属するカテゴリのラベルであり、目的変数と呼ぶ。線形判別分析には、次のような1次関数を用いる。線形判別式の中の変数の重みw_iは、学習データから求める。

$$y = \sum_{i=1}^{p} w_i x_i + w_0$$

線形判別分析では、上記の式の値を用いて判別を行う。例えば、2つのカテゴリA、Bの場合は、上記の式で得られた値がプラスであればカテゴリA、得られた値がマイナスであればカテゴリBに振り分ける。

線形判別分析法は、データによってはパフォーマンスが良くない場合がしばしばある。そこで、距離（あるいは類似度）による判別分析、ベイズ法による判別分析、ルールによる判別分析、k-NN法、サポートベクターマシン、集団学習法による分類法など、多くの分類方法が提案されている。本稿では、決定木と決定木を用いた集団学習法を中心に紹介する。

⑴ 決定木

①決定木とは

決定木は、学習データを用いて変数を分岐させる方法によって分類のルールを構築する。説明の便利のため、図1に示す2つの変数（横軸x_1、縦軸x_2）を用いて個体を3つのカテゴリに分類することを考えよう。1本の直線（線形判別）では3つのカテゴリを正しく分類することは不可能である。しかし、図1のように座標軸と平行する2本の直線を用いると、3つのカテゴリを完全に正しく分類することができる。

図1の2次元平面の区間分割図の構造は、図2のように表現することができる。図2は逆さにした木のような形状をしていることから樹木モデルと呼び、分類の問題では決定木、回帰の問題では回帰木とも呼ぶ。

また、図1と図2は次のようなIF-THENルールで表現することも可能である。

IF $x_1 < a$ THEN C
IF $x_1 \geq a$ and $x_2 \geq b$ THEN A
IF $x_1 \geq a$ and $x_2 < b$ THEN B

決定木は、枝を増やすことで複雑な分類問

図1　2次元平面の区間分割図

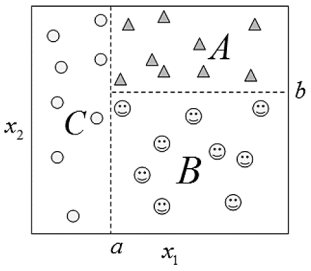
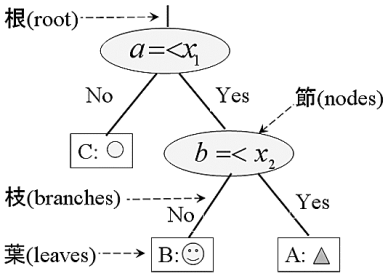


図2　決定木の構造



題や回帰問題に適用することが可能である。決定木の結果は理解しやすいため、データマイニングの方法として広く用いられるようになっている。

決定木に関する研究は1960年代初期までさかのぼるが、今日、広く用いられている決定木はCHAID、C4.5/C5.0/See5、CARTをベースとした3種類のアルゴリズムが主流である。

近年多く用いられているのはC4.5/C5.0/See5とCARTである。C4.5をベースとしたフリーソフトとしてはWekaがあり、Rにはパッケージ Rwekaがある。

CART（Classification And Regression Tree）は2進木を生成する。CARTは集団学習（アンサンブル学習）にも広く用いられているので、本稿ではCARTを用いることにする。

②決定木CARTの基本的仕組み

CARTは木を予め何の制限もせずに成長させ、データと対話しながら木の剪定を行う方法を取っている。

CARTでは分岐する変数を選択する際に不純度（impurity）、または情報量という指標を用いる。不純度は、変数を分岐する前と分岐させた後の誤差の改善の度合を示す指標であり、次の式で定義されている。

$$\Delta GI(t) = P_t GI(t) - P_L GI(t_L) - P_R GI(t_R)$$

式の中の $GI(t)$ は、次に示すノード t におけるGini分散指標（Gini diversity index）である。略してGini（ジニ）係数と呼ぶ。

$$GI(t) = 1 - \sum_k p(k \mid t)^2$$

式の中の $p(k \mid t)$ は、ノード t 内のカテゴリ k が正しく分類されている比率である。

$GI(t_L)$、$GI(t_R)$ は、それぞれノード t の左側と右側の枝のGini係数である。P_t、P_L、P_R は、それぞれ分割する前、分割した後の左側、分割した後の右側の個体の比率である。

不純度を計算する例として、3種類（A：住まい、B：家族、C：友達）の作文33編（それぞれ11編）を用いる。本稿では、作文をテーマごとに分類することを前提とする。したがって、作文の中から名詞を抽出して変数として用いる。合計47個の変数を用いて作成した決定木を図3に示す。また、図3の決定木作成に用いたデータ形式を表1に示す。

この決定木は、47の変数の中から分類に最も役に立つ3つの語（家族、友達、関係）を見つけ出し、分類ルールを構築している。図の中の「家族＞＝1.587」は、語「家族」の使用率が1.587%より高いことを意味する。

図3　3種類の作文の決定木

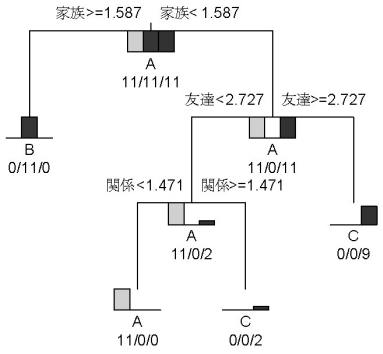


表1　決定木作成に用いたデータ形式

	自分	友達	家族	…	y
akke0	1.176	2.353	0	…	A
ataka0	1.053	0	0	…	A
⋮	⋮	⋮	⋮	⋮	⋮
yuka1	0	0	1.6	…	B
akke2	2.206	16.176	0	…	C
⋮	⋮	⋮	⋮	⋮	⋮
yuka2	0	3.101	0	…	C

図3を用いてGini係数に基づいた不純度の計算過程を説明する。

根の部分の3つのカテゴリの $p(k|t)$ 値は $p(A|t)=1/3$、$p(B|t)=1/3$、$p(C|t)=1/3$なので、Gini係数は

$$GI(t)=1-\{p(A|t)^2+p(B|t)^2+p(C|t)^2\}$$
$$=1-3\frac{1}{9}=0.6667$$

である。ノード1「家族」の左側と右側のGini係数はそれぞれ

$$GI(t_L)=1-\{(\frac{0}{11})^2+(\frac{11}{11})^2+(\frac{0}{11})^2\}=0$$

$$GI(t_R)=1-\{(\frac{11}{22})^2+(\frac{0}{22})^2+(\frac{11}{22})^2\}=0.5$$

であり、分割前の P_t は1、分割後の左側と右側の比率 P_L と P_R はそれぞれ $P_L=11/33=0.3333$、$P_R=22/33=0.6667$である。

したがって、ノード1「家族」の不純度は

$$\Delta GI(t)=1\times0.6667-0.3333\times0-0.6667\times0.5$$
$$=0.3334$$

になる。

ノード2「友達」の不純度の計算過程を次に示す。

$$GI(t)=1-\{(\frac{11}{22})^2+(\frac{0}{22})^2+(\frac{11}{22})^2\}=0.5$$

$$GI(t_L)=1-\{(\frac{11}{13})^2+(\frac{0}{13})^2+(\frac{2}{13})^2\}=0.2604$$

$$GI(t_R)=1-\{(\frac{0}{9})^2+(\frac{0}{9})^2+(\frac{9}{9})^2\}=0$$

$P_t=22/33=0.6667$、$P_L=13/22=0.5909$、

$P_R=9/22=0.4091$

$$\Delta GI(t)=0.6667\times0.5-0.5909\times0.2604$$
$$-0.4091\times0=0.1795$$

このようにノード1「家族」の不純度はノード2「友達」の不純度より高い。CARTでは、用いた変数についてこのように不純度を計算し、不純度が最も高い変数を選択して分岐を行い、木を成長させる。

木の枝を増やすことで、分類精度をいくらでも高めることは可能であるが、そのモデルをテストデータに当てはめたときに精度が良くなるとは限らない。モデルの適応性を高めるためには、成長しすぎた枝を適切に切り落とす剪定作業が必要である。剪定方法に関しては誌面の余裕がないので割愛する。

③データを用いた決定木の例

11人が3つのテーマ（住まい、家族、友達）について書いた作文33編の中から名詞を抽出し、相対頻度に置き換えたデータを、読者のために http://mjin.doshisha.ac.jp/data/sb3.csv に掲載している。データ形式は表1のとおりである。表1は分類分析に用いるデータの一般形式である。各行が1つの作文であり、各列が作文で使用された語の相対度数である。行列の左から48番（最後）のy列はテキストがどのカテゴリに属するかを示すラベルである。

RにはCARTに関連するパッケージtree、rpart、mvpart、C4.5に関するパッケージRWekaがある。本稿ではmvpartを用いることにする。図3の決定木を作成するコマンドを次に示す。

```
>install.packages("mvpart");
>library(mvpart)
>sb3<-read.csv("c:/temp/sb3.csv",head=T,
row.names=1)
>(sb3.rp<-mvpart(y~.,sb3,size=4))
```

sb3.rpの中に保存されている結果は分類ルールに関連する情報である。分類と変数との関係を考察するのには決定木を用いるのが便利である。

分類ルールの構築は、学習データを用いて構築したルールにより、カテゴリ所属が不明であるテキストがどのカテゴリに属するかを判別することが目的である。上記の例では、データセットの中のすべてを学習に用いたため、学習結果をテストするデータはない。

④学習と予測

データセットから幾つかをテスト用として取り除き、残りを用いて決定木を作成し、その結果を用いてテストデータについて予測してみる。ここでは6つをテスト用として取り除くことにする。Rでは関数**sample**を用いてサンプリングすることができる。サンプリングしたデータを用いて、学習用のデータとテスト用のデータを次のように作成する。

```
>set.seed(1); (sam<-sample(1:33,6))
[1]  9 12  18 28  6 26
```

サンプルの番号6、9、12、18、26、28をデータセットから抽出する行の番号とすることができる。ここではテーマごとに2つの作文が選ばれているが、これは偶然である。trainを学習用のデータ、testをテスト用のデータとする。

```
>train<-sb3[-sam,]; test<-sb3[sam,]
```

学習用のデータtrainを用いて分類ルールを作成し、そのルールに基づいてテスト用のデータtestについて関数**predict**を用いて予測を行う。結果を見やすくするため、テスト用のデータのカテゴリ属性と予測結果のクロス表（混同行列とも呼ぶ）を関数**table**で作成するコマンドを次に示す。混同行列の対角線の数が正しく判別されている数である。

```
>sb3.rp2<-mvpart(y~.,train)
>sb3.pr<-predict(sb3.rp2,test[,-48],type
="class")
>table(test$y,sb3.pr)
  sb3.pr
    A B C
  A 2 0 0
  B 0 2 0
  C 0 0 2
```

ここでは6つがすべて正しく判別されている（正解率が100%）。しかし、この結果は偶然である可能性がある。より信憑性が高い評価方法について次の節で紹介する。

3. 結果の評価

⑴ 交差確認法

データセットの中から、学習用のデータとテスト用のデータを分けて検証を繰り返す方法として、交差確認法がある。データ標本をn等分し、その中の1つをテスト用、残りのn－1個を学習用とし、すべてについて1回のテストの機会が与えられるようにn回の学習とテストを行う方法をn重交差確認（n-fold cross validation）法と呼ぶ。

⑵ OOB確認法

データセットの中からランダムに一部をテスト用として取り出し、その残りを学習用とする方法もある。取り出したデータをOOB（out-of-bag）データと呼ぶ。学習とテストを繰り返す回数を多くすることで、信憑性が高い結果を得ることが可能である。

⑶ 結果の評価指標

カテゴリC_iにおける分類の結果は、表2の

表2　カテゴリC_iの分類結果

カテゴリ C_i		分類器の結果	
		Yes	No
データ	Yes	a_i	c_i
	No	b_i	d_i

ような形式で示すことができる。

分類器の精度は、正解率$(a_i+d_i)/(a_i+b_i+c_i+d_i)$と誤分類率(1－正解率)で評価することも可能であるが、テキストマイニングの分野では、再現率（R：recall）と適合率（P：precision）を用いる場合もある。

再現率Rは分類器がどれぐらい「漏れ」なく正しく判別しているかに関する度合であり、適合率Pは分類器の分類結果に混入した「ゴミ」がどれだけ少ないかを表す。カテゴリC_iにおける再現率R_iと適合率P_iの定義を次に示す。

$$R_i=\frac{a_i}{a_i+c_i}、\quad P_i=\frac{a_i}{a_i+b_i}$$

カテゴリの総数がmである分類問題では、評価指標として再現率（R）、適合率（P）のマクロ平均（macro-average）

$$\bar{R}_{ma}=\frac{1}{m}\sum_{i=1}^{m}\frac{a_i}{a_i+c_i}、\quad \bar{P}_{ma}=\frac{1}{m}\sum_{i=1}^{m}\frac{a_i}{a_i+b_i}$$

あるいはマイクロ平均（micro-average）

$$\bar{R}_{mi}=\frac{\sum_{i=1}^{m}a_i}{\sum_{i=1}^{m}(a_i+c_i)}、\quad \bar{P}_{mi}=\frac{\sum_{i=1}^{m}a_i}{\sum_{i=1}^{m}(a_i+b_i)}$$

が用いられている。

再現率と適合率を折中した評価指標として、F_β値（F_β-measure）がある。F_β値は次のように定義されているが、通常$\beta=1$であるF_1が多用されている。

$$F_\beta=\frac{(\beta^2+1)\times P\times R}{\beta^2\times P+R}、\quad F_1=\frac{2\times P\times R}{P+R}$$

例えば、分類器による分類結果を集計した混同行列が表3のとおりであるとする。

表3のカテゴリAのみを集計すると、表4のとおりになる。

表3　分類結果の混同行列の例

		分類の結果		
		A	B	C
データ	A	9	1	1
	B	0	10	1
	C	1	0	10

表4　カテゴリAの混同行列

カテゴリ A		分類器の結果	
		Yes	No
データ	Yes	9	2
	No	1	21

カテゴリAの再現率と適合率はそれぞれ次の値になる。

$$R_A=\frac{a_A}{a_A+c_A}=\frac{9}{9+2}=\frac{9}{11}=0.8182$$

$$P_A=\frac{a_A}{a_A+b_A}=\frac{9}{9+1}=\frac{9}{10}=0.9$$

カテゴリBでは$a_B=10$、$c_B=1$、$b_B=1$であるので再現率と適合率はそれぞれ$R_B=10/11=0.9091$、$P_B=10/11=0.9091$であり、カテゴリCでは$a_c=10$、$c_c=1$、$b_c=2$であるので、再現率と適合率はそれぞれ$R_C=10/11=0.9091$、$P_C=10/12=0.8333$である。

これらの再現率のマクロ平均は$\bar{R}_{ma}=(0.8182+0.9091+0.9091)/3=0.8788$、適合率のマクロ平均は$\bar{P}_{ma}=(0.9+0.9091+0.8333)/3=0.8808$である。したがって、表3の混同行列の$F_1$値は$\frac{2\times 0.8788\times 0.8808}{0.8788+0.8808}=0.8798$である。

分類問題においては、再現率のマイクロ平均、適合率のマイクロ平均、マイクロ平均のF_1値、正解率の四者は等しい。