■■■専門家予想のクラスター分析とロジットモデル

非線形物理学講座　　SP-０９１３５　長洲　雅俊

競馬新聞には専門家の予想が印という情報で掲載されている。では、その印にはどのような情報がこめられているのだろうか。
本研究では、予想印の精度、専門家間の予想印の相関関係をクラスター分析し、複数の情報を混ぜることにより精度の良い競馬予想の作成を試みた。

1　解析に用いたデータの詳細

情報としては日刊競馬新聞、ニッカンスポーツのコンピ指数、JRA-VAN 提供の予想走破タイム、最終オッズとレース結果を使用している。日刊競馬新聞は 2010 年 3 月 6 日から 2010 年 5 月 30 日までの 12 週 24 日分のデータを 1 部 200 円で購入し、手入力した。他のデータについてはメディアから入手した。日刊競馬新聞のデータは全部で 321 レース 4721 頭の 6 名分の予想印が掲載されていて◎、○、▲、▽（新聞では二重三角）、☆、△、無印からなる。データ入力には◎を 1、○を 2、▲を 3、▽を 4、☆と△を 5、無印を 6 で入力し、予想者間の相関を調べた。予想者の柏木氏を A、宮崎氏を B、黒津氏を C、久保木氏を D、桧原氏を E、飯田氏を F としている。

2　相関係数とクラスター分析

専門家は 1 レースごとに◎＞○＞▲＞▽＞△＞無印と馬をそれぞれ順位づけている。AR(Accuracy Ratio) は予想印の順位と結果との順位相関係数である。0 から 1 の数値をとり、1 に近づくほど信頼度の高い予想であることを意味している。この予想を Kendall の順位相関係数で表した。相関係数は-1 から 1 の数値をとり、0 に近づくほど弱い相関をもち、-1 と 1 に近づくほど強い相関を持つ関係性になっている。

表 2.1: 予想者の順位相関係数

	AR	A	B	C	D	E
A	0.503	1				
B	0.463	0.528	1			
C	0.542	0.591	0.513	1		
D	0.395	0.444	0.403	0.457	1	
E	0.524	0.598	0.550	0.647	0.448	1
F	0.585	0.645	0.549	0.627	0.488	0.669

では、表 1.1 を使ってどのファクターを重視した予想しているか、似た予想者ごとに分類してみる。ward 法でクラスター分析した階層クラスターが図 2.1 である。図 2.1 の縦軸は距離である。この距離が予想の似ている度合いになり、F さんと D さんが最もはなれているので違うファクターを重視して競馬予想していることがわかる。集団のひとりひとりがいろいろな視点から物事を見ているのならば、その平均をとることでその個人よりも良い結果が得られることは集団知として知られている。よって、これらを組み合わせることにより、精度の高いモデルをつくりだすことが可能である。
そして、組み合わせるために使用したモデルがロジットモデルである。あるレース r（１～R）の h（１～H(r)）番目の馬の情報を X（h, r ,i）と　$Y \in (0,1)$　がある。i（1～6）は専門家予想の数。スコア S（h, r）は、ロジットモデルではまず、(１) 式である馬のスコアを計算する。X は予想者の情報であり、w(i) は予想者の重みである。

$$S(h,r) = \sum_{i=1}^{6} w(i)X(h,r,i) \tag{1}$$

$$P(h,r) = \frac{exp(S(r,h))}{\sum_{k=1}^{H} exp(S(r,k))} \tag{2}$$

P は馬の勝つ勝率である。この勝率と結果 $Y \in (0,1)$ との二乗誤差 (error) が 0 に近づくように勾配法をつかって w(i) 最適化していく。

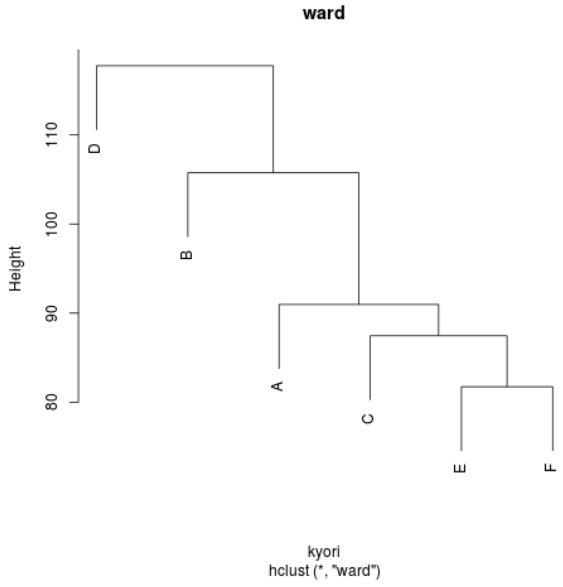

図 2.1: クラスター分析

3　まとめ

今回、学習データ 190 レースをロジットモデルで予想者の重み係数 w(i) を最適化し、確認データで精度を調べた。確認データ 131 レースの結果では　予想者全員の AR に勝つが日刊新聞が独自に算出したコンピ指数、オッズの AR には負けていることがわかった。ロジットモデルにコンピ指数、オッズを追加することでコンピ指数に勝てるかもしれないがオッズには勝てないだろう。これからもいろいろな方向から AR を上げていくこと目標に研究を進めていく。