専門家予想の
クラスター分析とロジットモデル

平成25 年6 月6 日

非線形物理学講座　　SP-０９１３５　長洲　雅俊

目 次

1　はじめに

競馬とは1レースごとの１、2、3着を予想し、当たった場合その馬券についている配当がもらえるギャンブルである。配当は（本実験では、一着を当てる単勝馬券の場合を考えている）１レースの単勝馬券合計売り上げ金額の約８０パーセントをそれぞれ馬券の得票率で割ったものがオッズになり、オッズ×購入金額が配当金となる。これはオッズの低いものは馬券購入者集団の得票率が高いと受け取れる。　得票率が高いものは一着になる確率が高い研究結果もでており、競馬は集団知の実績が優れている。本研究の目標はこの優れたオッズよりも精度の良いモデルをつくることである。このモデルをつくるために本研究では日刊競馬新聞の予想者に注目した。

2　解析に用いたデータの詳細

情報として日刊競馬新聞、日刊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（柏木氏、宮崎氏、黒津氏、久保木氏、桧原氏、飯田氏）の予想印が掲載されていて◎、○、▲、∇（新聞では二重三角）、☆、△、無印からなり、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。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、これらの予想の精度、予想者間の相関を調べた。

3　相関係数

相関係数はある二つの対象が似ているか似ていないかの度合いを数値で見ることができる。そして、AR(Accuracy Ratio) は対象情報とレース結果の一着馬だけの相関係数を表している。数値は 0 から１の間で、1 に近づくほど結果と似ており、予想の精度が高いことを表している。オッズの AR が一番高く、次にコンピ指数が高い。競馬予想専門家の中では飯田氏が高く、久

表 1: 各情報の AR

	AR
柏木氏	0.495618
宮崎氏	0.435378
黒津氏	0.504262
久保木氏	0.394078
桧原氏	0.501513
飯田氏	0.564739
odds	0.665113
コンピ指数	0.636621
JRA-VAN	0.477331

保木氏が一番低い。では、競馬専門紙に久保木氏は必要ないのだろうか。ここについては後ほど触れることにする。

4　順位相関係数

次にケンドールの順位相関係数を使って予想者間の相関について調べた。予想がどのくらい似ているかを数値で見ることができる。順位相関係数は－１から１の間で－１は全く似ていない場合にとる値で、１はすべて一致している場合にとる値である。０は完全に独立、二つが全く関係ない場合に現れる。この順位相関係数をそれぞれの予想で表したもの表２である。

表 2: 予想者間の順位相関係数

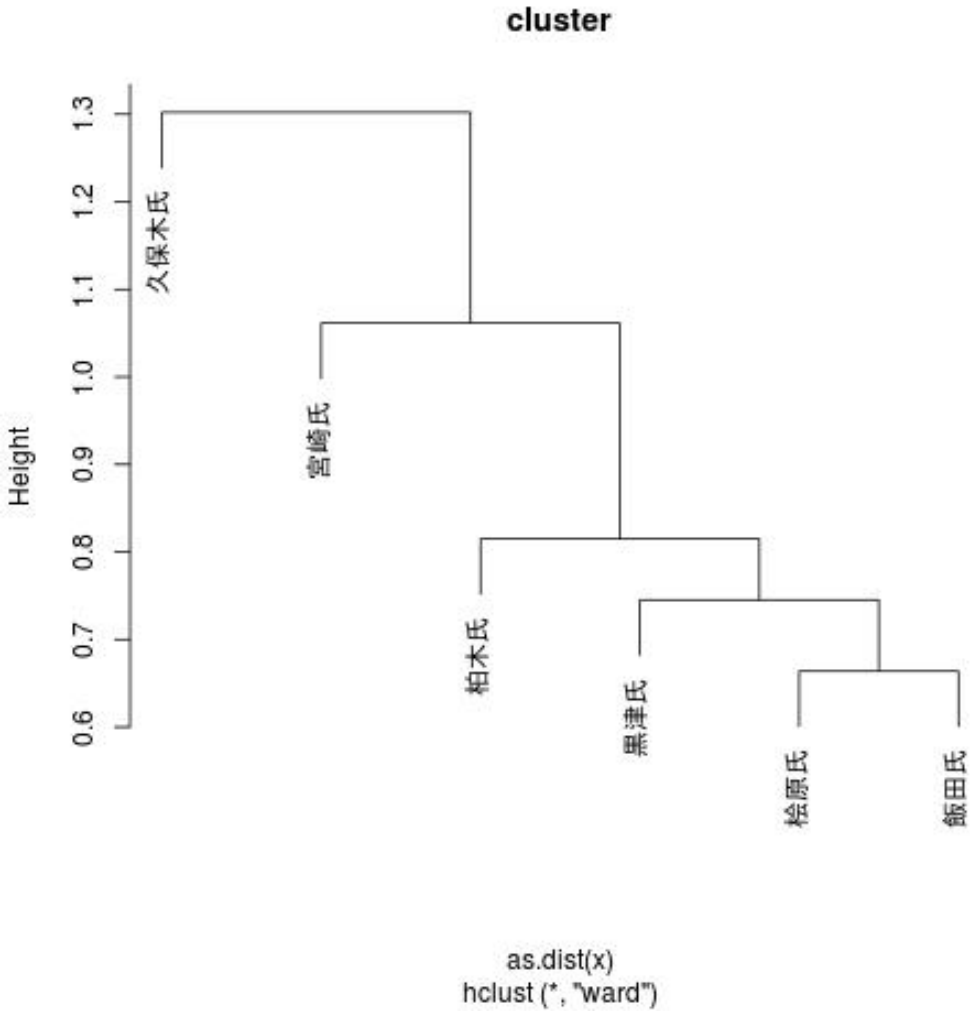
	柏木氏	宮崎氏	黒津氏	久保木氏	桧原氏	飯田氏
柏木氏	1					
宮崎氏	0.528451	1				
黒津氏	0.591621	0.512926	1			
久保木氏	0.444981	0.403474	0.457054	1		
桧原氏	0.598460	0.550397	0.647886	0.448014	1	
飯田氏	0.645960	0.549017	0.627644	0.488374	0.669828	1

5　クラスター分析とロジットモデル

数値では宮崎氏と久保木氏である。局所的に見るのではなく、全体的に見たいと思いクラスター分析を行った。クラスター分析の結果は図１である。今回、クラスター分析をウォード法で解析した。最も予想が似ているのは桧原氏と飯田氏である。その逆で全く似ていないのが飯田氏と久保木氏である。　ここで、久保木氏の必要性について書こうと思う。集団知について話すと必要性がわかる。集団知とは人が集まることで得られる知識である。この集団は考え方の違う人で構成されていなければならない。考え方が違う、または視点の違う人のそれぞれの答えの平均が精確な答えであること知られてい. る。この観点から似ていない久保木氏の予想は一着の的中率が低くても必要なのである。また、集団知から、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。

$$S(h,r) = \sum_{i=1}^{6} w(i)X(h,r,i) \tag{5.1}$$

$$P(h,r) = \frac{exp(S(r,h))}{\sum_{\mathrm{k}=1}^{H} exp(S(r,k))} \tag{5.2}$$



6　まとめ

本研究で作成したロジットモデルはオッズとは少し違う視点でオッズに違う精度まで近づくことはできた。しかし、目標であったオッズより精度のよいモデルをつくることはできなかった。競馬は情報量の多い賭け事である。これが集団知として素晴らしい実績をもたらしているのだと思う。だからと言ってすべての情報をロジットモデルにしたところでオッズに似たモデルができるだけである。いかに数多くある情報から選びだせるかである。今考えているものは馬の脚色とレース会場の適性、レース会場と枠順の有利不利、調教タイムそして買い方である。結果として得られたものが倍率の低いものを二着固定で一着と二着を当てる馬券買ったほうが期待値が高くなるのではないか、など考えれば考えるほど楽しい。競馬についてこれからも研究していこうと思う。