申請先学部　基礎工学部
採択番号　　No.3

平成 27 年度学部学生による自主研究奨励事業研究成果報告書

ふりがな 氏　　名	なかむら　しゅんすけ 中村　駿佑	学部 学科	基礎工学部 情報科学科	学年	2　年
ふりがな 共　同 研究者名		学部 学科		学年	年
					年

アドバイザー教員 氏名	伊森　晋平	所属	基礎工学研究科
研究課題名	統計手法及び機械学習を用いた競馬におけるデータの解析		
研究成果の概要	研究目的、研究計画、研究方法、研究経過、研究成果等について記述すること。必要に応じて用紙を追加してもよい。		

1　研究目的

本研究では競馬のレース予想の助けとなる情報を示すことを目的とする．そのために様々な統計的手法や機械学習を用いて競馬に関するデータを解析し，レース結果の予想において重要なものを明らかにする．また，それらを競馬ファンに伝えるためのウェブアプリケーションも作成する．

競馬には関連する様々なデータがあり，例えば競走馬については前走までの成績や血統，レースについては距離，場所，馬場の状態などがある．また，近年では Twitter などの SNS における競馬ファンによる情報や感想，意見もある．これらの多様なデータからレース結果に影響を与える重要な情報を識別し，それを元にレースの正しい結果を予想することは困難である．こういった予想の困難さが新規競馬ファン参入の妨げになっているとも考えられる．したがって，統計的手法や機械学習を用いて競馬に関するデータを解析し，新規競馬ファンによるレース結果の予想を助けるための情報を示すことで，このような困難さを解消することができると考えた．

このような方法でレース結果やオッズを予測することは新しい試みであり，本研究の特色である．本研究の結果がレース結果の予想の助けとなれば，競馬ファンがレースの予想に取り組みやすくなり，その結果として競馬ファンの拡大につながることも期待される．

2　研究計画・方法

本研究ではデータの収集，データの解析・予測モデルの構築，解析結果の可視化の順に研究を進める．

(1) データの収集

データの解析を行うためには，まずデータを収集する必要がある．競馬は日本国内においては中央競馬や地方競馬，また海外でも多くの国で開催されているが，本研究では中央競馬（日本中央競馬会（JRA）が主催する競馬）のレースをデータ解析の対象とする．解析に用いるデータは過去のレースの結果及び競馬に関するツイート（Twitter での投稿）である．過去のレースの結果はインターネット上のウェブサイトから収集する．競馬に関するツイートは Twitter 社が公開している API を用いて，競馬に関連するキーワード（馬名，レース名など）を含むツイートをデータベースに保存するプログラムを作成し，それによってツイートを収集する．

(2) データの解析・予測モデルの構築

上記の方法によって集めたデータに対して，どの変数がレース結果の予想において重要であるかを調べる．データに対して変数選択を行うことにより，前走の成績や血統，レースの距離や場所といった様々な説明変数の中でどれが重要なものであるかを判別する．ツイートに関しては，ツイートから得られる情報（例えば，ツイート本文中に馬名が含まれる回数）を，説明変数として用いる．これらを元にレースの着順及び最終オッズを予測するモデルを作成する．具体的にはロジスティック回帰分析，順序付きロジスティック回帰分析やランダムフォレストなどの統計手法や機械学習の様々な手法を使用する．これらの手法は R のパッケージや Python のライブラリ（たとえば Scikit-learn（[1]）など）として実装されており，本研究ではこれらのソフトウェアも活用して解析を行う．

(3) 解析結果の可視化

上記の方法によって得られた解析結果の理解を容易にするために，レース結果の予測などをグラフなどによって可視化し，それらをウェブブラウザから閲覧可能な簡易的なウェブアプリケーションを作成する．

3　研究経過

研究はおおむね研究計画通りに進めたが，ツイートの解析の方法に関して変更があった．ツイートから読み取れる情報をレース結果の予測の説明変数の一つして扱うことを計画していたが，研究期間中に取得するできた競馬に関するツイート数が想定していたよりも少なかったため，ツイートから得られる説明変数がレース結果の予測に対して重要であるかどうかを判断することは困難であると考えた．したがって，ツイートから得られる説明変数をレース結果の予測のために用いることはせず，その代わりに出走馬名を含むツイート中のポジティブな単語とネガティブな単語の数から，各出走馬に対するイメージを数値化することを試みた．

4　研究成果

(1) レース結果を予測するモデルの構築

(1.1)　予測モデルの構築

まずはレース結果を予測するモデルを作成した．レース結果の中でも，1 着になる馬を予測することは他の順位を予測するよりも重要であると考えられるので，1 着になる出走馬を予測するモデルを作成した．

ここでは L1 正則化付きロジスティック回帰モデル（[2]）を用いたものについて記述する．他の手法も試してみたが，L1 正則化付きロジスティック回帰モデルに比べてあまり良い結果が得られなかった．目的変数は各出走馬が 1 着になるかどうか（1 着を 1，それ以外を 0 の二値）として，説明変数としては各出走馬の騎手名，年齢，性別，調教師名，枠番号，馬番号，斤量，負担重量，前走（直近の出走したレース）の距離，前走との距離差，前走の順位，出走するレースが行われる競馬場の枠番号における平均着順などを用いた．前走のデータを用いるので，前走のデータがない馬を含むレースは推定するデータから除外した．ここで質的変数はダミー変数として扱った．L1 罰則を用いることによって数ある説明変数の中から重要な変数選択をする．

(1.2)　予測モデルの評価実験

(1.1)で作成したモデルを評価するための実験を行った．2010 年~2013 年のデータを学習用のデータとし，それを用いて L1 正則化付きロジスティック回帰モデルの回帰係数を推定した．正則化パラメータはクロスバリデーションによって決定し，モデルの評価の基準は正判別率を用いた．ここでの正判別率とはレースごとの出走馬のうち最も 1 着になる確率が高い馬が実際に 1 着になった割合である．2014 年のデータをテストデータとし，それに対して推定された回帰係数を用いて 1 着であるかどうかを予測することにより，モデルの評価をする． 2014 年のレースの予測の結果は，全 3026 レースのうち，的中したのは 742 レースで，正判別率は 24.52%であった．

(1.3) モデルの解釈

(1.2)では L1 罰則を用いて推定を行っているので，回帰係数が 0 となるものがあり，それに対応する説明変数は予測に関係していないことになる．つまり回帰係数が 0 でないものに対応する説明変数はそうでないものよりも重要であると考えることができる．(1.2)で決めたパラメータのとき，0 でない回帰係数は 153 個で，0 の回帰係数は 566 個であった．(1.1)で挙げた説明変数のうち，各出走馬の騎手名（64 人分），年齢，性別，調教師名（61 人分），枠番号，馬番号，斤量，負担重量が 0 でない回帰係数に対応する説明変数で，各出走馬の騎手名（258 人分），調教師名（288 人分），前走の順位，出走するレースが行われる競馬場の枠番号における平均着順が 0 の回帰係数に対応する説明変数であった．

(2) Twitter のデータの解析

ツイートの解析に関しては，MeCab というソフトウェアを用いて収集したツイートの本文を形態素解析し，ある出走馬とともにツイート本文中に現れるポジティブな単語とネガティブな単語をカウントすることにより，ツイートからその出走馬のイメージが良いものであるか悪いものであるかを数値で表した．

(3) Web アプリケーションの作成

Web アプリケーションについては，Python のウェブフレームワークである Flask を用いて作成した．(1) によって作成したレース結果を予測するモデルによる各出走馬の 1 着になる確率の予測値をグラフにしたものや，それと単勝のオッズから算出される払戻率の期待値をグラフにしたものなどを表示する簡単なものを作成した．図 1 はこの Web アプリケーションのスク

リーンショットである．

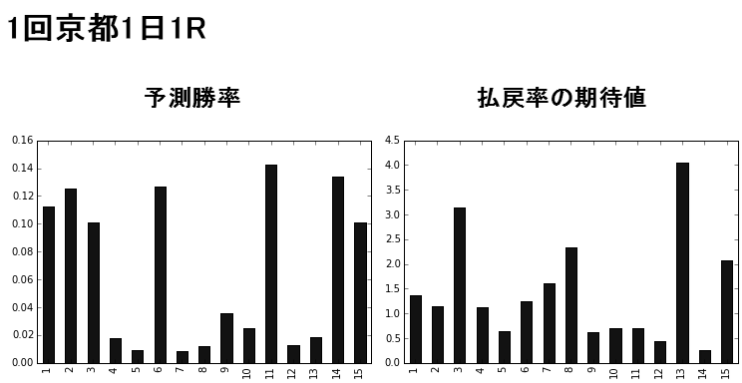

図 1　作成した Web アプリケーションのスクリーンショット．左図の横軸は馬番号，縦軸は予測勝率を表す．右図の横軸は馬番号，縦軸は単勝の払戻率の期待値を表す．「1 回京都 1 日 1R」は第 1 回京都競馬開催の第 1 日・第 1 レースを表す．

4．今後の研究課題

ツイートの解析に関して，ツイートの収集方法の改善が今度の課題である．今回用いた手法では，馬名が略称や愛称などでツイート本文中に現れるものを収集することができないことや，馬名がツイート中には含まれているもののその馬名が一般的に使われる単語と一致する（あるいは単語の部分文字列となる）とき競馬とは全く関係のないツイートも収集してしまうといった問題点がある．こういった問題点を改善することにより十分なデータを集めることができれば，今回の研究では扱うことができなかったツイッターから読み取れる情報を予測モデルに組み込むことにも取り組んでいきたい．

5．参考文献

[1]. Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, Édouard Duchesnay, Scikit-learn: Machine Learning in Python, The Journal of Machine Learning Research, 12, p.2825-2830, 2011.

[2]. Trevor Hastie , Robert Tibshirani , Jerome Friedman 著，杉山 将，井手 剛，神嶌 敏弘，栗田 多喜夫，前田 英作監訳，井尻 善久，井手 剛，岩田 具治，金森 敬文，兼村 厚範，烏山 昌幸，河原 吉伸，木村 昭悟，小西 嘉典，酒井 智弥，鈴木 大慈，竹内 一郎，玉木 徹，出口 大輔，冨岡 亮太，波部 斉，前田 新一，持橋 大地，山田 誠　翻訳，『統計的学習の基礎』，共立出版，2014，p.146-147.

[3].C.M.ビショップ著，栗田 多喜夫，樋口 知之，松本 裕治，村田 昇 監訳，神嶌 敏弘，杉山 将，小野田 崇，池田 和司，鹿島 久嗣，賀沢 秀人，中島 伸一，竹内 純一，持橋 大地，小山 聡，井手 剛，篠田 浩一，山川 宏　翻訳『パターン認識と機械学習』，丸善出版，2012.