F-053

機械学習による日本プロ野球の成績予測モデル構築

Model Construction for Predicting Results in Nippon Professional Baseball Organization by Machine Learning

近藤 竜也 † 李 嘉誠 † 能登 正人 †
Tatsuya Kondoh Jiacheng Li Masato Noto

1. はじめに

近年の野球界では，主に MLB (Major League Baseball) における成績予測の研究が進んでいる．MLB と NPB (Nippon Professional Baseball Organization) を比較した観点から見ると，MLB はセイバーメトリクスという野球をデータ分析し，統計学的な根拠に基づいて選手の評価や戦略を考える手法を作成し，Statcast という打球や動作を追跡し解析するシステムが導入された．このように MLB では，チームの勝利に効果的に繋がる采配や育成ができている．一方で NPB では，近年 MLB の研究を取り入れ，野球をデータ分析する客観的な視点からの研究に取り組んでいるが，MLB と比較して成績予測の研究が不十分であり，遅れが見られる．以上の背景から，MLB の先行研究を参考にしつつ，NPB における成績予測の研究を進めることで，より効果的な采配や育成ができ，NPB の更なる競技レベル向上の可能性がある．

本研究では，NPB の競技レベル向上に貢献できる成績予測システム作成を目的とし，個人成績データとチーム成績データから，主成分分析を行い，特徴量の選択，機械学習の予測モデルの選択・調整・最適化などの予測モデリングを繰り返し行うことで最適な予測モデルを構築する．

2. 先行研究

Horvat らは，スポーツにおいて，成績の予測と貴重な情報の抽出が，野球関係者だけでなく，広範な視聴者にとっても，エンターテイメントの面で有益であると考えている [1]．スポーツの成績予測では，機械学習アルゴリズムを使用した研究が多い．そこで，100 を超えるスポーツ結果予測の論文を分析しレビューを行った．

分析と調査の結果，数あるスポーツと比較して，野球の結果予測の精度が低いことが明らかとなった．野球の予測精度が低い原因として，分析した論文の成績予測では，MLB のデータを用いており，MLB は首位のチームでも 6 割程度の勝率が最高で，競争が激しいことがあげられる．また他のスポーツと比較して，勝敗の要因が多いことも考えられる．また分析から，機械学習の中の回帰モデル，分類モデル，ニューラルネットワークモデルが適していることが分かり，今後のスポーツ成績予測は勿論のこと，野球の成績予測における今後の研究課題が示唆された．また Bunker らの研究においても，同様のスポーツ成績予測の研究テーマで主張をしている [2].

3. 提案手法

本研究では，主に個人成績データを使用し，どれだけ個人の成績がチームの勝利に影響しているのかを主成分分析を行い，新年度のチームの勝率を予測する最適な予測アルゴリズムを作成する．最適な予測アルゴリズムを作成するために，まずは既存の機械学習の予測モデルから探索する必要がある．探索方法として，先行研究 [1] と [2] から野球の成績予測に適切なモデルの評価を踏まえつつ，現状で実装が可能であった回帰モデル，ニューラルネットワークモデルを使用し，この中から予測精度が高かったものを比較して明らかにする．

4. 実験

4.1 データセット作成

NPB 公式サイトや NPB の成績をまとめているサイトから，NPB 公式サイトの情報を基準にし，正確なデータを収集する．収集できたデータは，2009 年から 2023 年のチーム成績であり，順位や勝率，野手成績，投手成績を収集した．

収集したデータは CSV ファイルとして格納し，予測アルゴリズムには Python を使用して構築するため，プログラムで処理しやすいように，データ整理や前処理を行う．データセットは，個人成績を野手成績と投手成績に分けてファイルを作成し，チーム成績も別で作成する．

4.2 主成分分析

個人成績の野手成績と投手成績を別々に扱い，野手成績は合計 5068 人分のデータ，投手成績は 4934 人分をデータセットとして用意する．15 年分の各年度と各チームの一軍の公式戦に出場したデータがある．すべての選手の成績を平等に扱ってしまうと，予測に偏りが出てきたしまうため、野手の場合は打席数，投手の場合は投球回の各年度各チームで上位 10 人を抽出する．これによって，チームに最も貢献した 10 人を選定し，より適切な予測が可能となる．次に主成分分析を行い，抽出した選手の成績とチームの勝率との関連性を求め，それぞれの個人成績とチームの勝率との寄与率を求める．こうする

† 神奈川大学, Kanagawa University

ことで，膨大な個人成績を圧縮して，5 要素にして説明変数を作成する．

4.3 使用する予測モデル

使用する予測モデルは，回帰モデル，ニューラルネットワークモデルである．回帰モデルでは，線形回帰，Ridge 回帰，Lasso 回帰の 3 種類のモデルを使用した．ニューラルネットワークモデルは，多層パーセプトロン，CNN（畳み込みニューラルネットワーク），RNN（再帰型ニューラルネットワーク）を使用した．これらの機械学習モデルを使用した理由としては，多くの特徴量と，判定ではなく数値の詳細な予測を行うため，これらのモデル使用に至った．使用する機械学習モデルを表 1 に示す．

表 1: 使用する機械学習モデル

使用モデル	機械学習の種類
線形回帰	回帰型
Ridge 回帰	回帰型
Lasso 回帰	回帰型
多層パーセプトロン	ニューラルネットワーク型
CNN	ニューラルネットワーク型
RNN	ニューラルネットワーク型

4.4 予測モデルの構築

回帰型とニューラルネットワーク型のモデルの構築方法は，Scikit-learn や TensorFlow，Keras などのライブラリやフレームワークを Python で使用する．使用するデータとしては，2009 年から 2023 年の個人とチーム成績を扱う．その中で，2009 年から 2022 年のデータから主成分分析を行い，求めた寄与率を説明変数として予測モデルに学習させ，2023 年のデータをテストデータとする．学習後，作成したモデルで 2023 年の勝率を予測し，勝率予測に基づいて 2023 年のチーム順位を予測する．学習させる際には，必要に応じて，ハイパーパラメータチューニングなどの最適化を行う．

4.5 評価方法

評価方法として，予測した順位と実際の順位を比較して，順位の的中率を求め，それを予測精度とする．求め方は，相対誤差を用いて，求まったパーセント誤差を最大 100%で表す．予測モデルの性能を評価するために MAE (Mean Absolute Error) を使用する．MAE を使用する理由としては，当初，性能評価候補であった RMSE (Root Mean Square Error) と比べて，外れ値の影響を受けにくいため，より適切な評価が可能になると考えたためである．

5. 結果・考察

6 個の機械学習モデルを用いて，勝率成績予測モデルを構築し，勝率の予測を行った．その結果を表 2 と表 3 に示す．野手成績を使用した場合は，RNN が一番高い精度となった．投手成績を使用した場合は，Lasso 回帰が一番高い精度となった．全体を通して MAE は比較的どれも同じ値となった．結果から，全体的に 90%近く高精度で予測ができており，データセットを変更しても予測精度として大きな違いはなかった．また多層パーセプトロンの精度が比較して低くなったのは，求める勝率と個人成績との関係が線形の関係であり，非線形ではないことが理由と考察する．

表 2: 野手個人成績を使用した予測結果

使用モデル	予測精度 [%]	MAE
線形回帰	89.11	0.004485
Ridge 回帰	89.11	0.004485
Lasso 回帰	89.39	0.004575
多層パーセプトロン	70.01	0.009651
CNN	89.25	0.004533
RNN	89.40	0.004414

表 3: 投手個人成績を使用した予測結果

使用モデル	予測精度 [%]	MAE
線形回帰	89.05	0.003865
Ridge 回帰	89.05	0.003865
Lasso 回帰	89.39	0.004575
多層パーセプトロン	83.73	0.007589
CNN	87.10	0.004588
RNN	88.33	0.004038

6. おわりに

本研究では NPB に貢献できる成績予測システム作成を目的とし提案した．課題として，研究方針の観点から，現在は新年度の特徴量を使用して予測しているため，シーズンが開始する前に，特徴量を作り予測する手法を検討する必要がある．データセットの観点から，今回は野手と投手を別々に扱ったが，同時に使用して予測すれば，さらなる精度向上の可能性がある．アルゴリズムの観点だと，最適化関数やハイパーパラメータチューニングなどモデルの最適化方法を再確認し，適切な手法を追加する必要がある．以上の課題を解決することで，予測精度の改善の余地はある．

参考文献

[1] Horvat, T. and Job, J.: The Use of Machine Learning in Sport Outcome Prediction: A Review, *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, Vol. 10, No. 5, p. e1380 (2020).

[2] Bunker, R. and Susnjak, T.: The Application of Machine Learning Techniques for Predicting Match Results in Team Sport: A Review, *Journal of Artificial Intelligence Research*, Vol. 73, pp. 1285–1322 (2022).