7Q-01

機械学習を用いた日本プロ野球における順位成績予測

近藤 竜也　　李 嘉誠　　能登 正人
神奈川大学大学院工学研究科工学専攻電気電子情報工学領域

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は進んでおり，より正確で客観的な評価ができ，チームの勝利に繋がる采配や育成ができている．一方で NPB (Nippon Professional Baseball Organization) では，MLB のデータ分析を取り入れ，競技レベル向上に取り組んでいるが，まだ 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．そのため 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，更なる競技レベル向上の可能性がある．

本研究では，NPB の競技レベル向上への貢献を目指し，成績予測システム作成を目的とする．大規模のデータセットを用いたシステムを作成する前に，比較的に少量のデータを用いて，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アルゴリズムを構築する．

2　先行研究

Horvat らは，スポーツにおける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，同様の研究テーマで主張をしている [2].

Predicting Results in Nippon Professional Baseball Organization by Machine Learning
Tatsuya Kondoh, Jiacheng Li and Masato Noto
Field of Electrical, Electronics and Information Engineering, Course of Engineering, Graduate School of Engineering, Kanagawa University

3　提案手法

本研究では，大規模のデータセットを使用して成績予測を行う前段階として，できるだけ少ないデータから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，打撃成績，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．CSV ファイルは 1 チームずつ分け，NPB は 12 球団があるため，合計 12 個のファイルを作成する．

4.2　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，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，使用するデータセットが時系列データであるため，比較的に時系列データからの予測に適しているモデルを

考えることで，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
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

4.3 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．各モデルの基本的な構築する流れとして，最初に NPB のデータセットを作成する．使用するデータとしては，2009 年から 2023 年の年間チーム成績を扱う．その中で，2009 年から 2022 年のデータを説明変数として，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．

ここで，チーム順位は，勝率に基づいて決定しているため，使用する説明変数の特徴量は，勝率を扱い，時系列の傾向を 14 年間の各チームのデータから学習する．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．

4.4 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，式 (2) を用いて，順位が的中したチーム数と全チーム数の商を求め，100 をかけることで，パーセントで的中率を表す．順位付けは，NPB の場合，各リーグで付けられるため，セリーグとパリーグに分けてチームの順位を予測する．加えて，各チームの順位予測に使用した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．

$$\text{予測精度（的中率）} = \frac{\text{的中したチーム数}}{\text{全チーム数}} 100 \qquad (1)$$

5 結果・考察

5 個の機械学習モデルを用いて，順位成績予測モデルを構築し，順位を予測を行った．その結果を表 (2) に示す．予測精度は，Lasso 回帰と CNN が一番高い精度となり，MAE は RNN が一番小さくなった．結果から，性能評価の場合，回帰型は機械学習の中でも同じ型の分類となることもあり，どれも近い値となり，ニューラルネットワーク型は種類によって異なることが分かった．全体的には，精度として 41.7 %が一番高い結果となった．このように少ないデータ数と特徴量から，時系列の傾向のみを学習させ，予測モデルを構築することで，半分近くのチームの順位予測が可能であることが分かった．

表 2: 予測結果

使用モデル	予測精度 [%]	MAE
線形回帰	33.3	0.0428
Ridge 回帰	33.3	0.0434
Lasso 回帰	41.7	0.0485
CNN	41.7	0.110
RNN	33.3	0.0102

6 おわりに

本研究では，比較的に少量のデータを用いて，予測アルゴリズムを構築することを提案した．課題として，特徴抽出方法の点であると，時系列傾向以外で，他にシーズン開始前に，シーズン終了の成績に影響する特徴量を探索し，より予測に効果的な特徴量を見つけ出す必要がある．アルゴリズムの点であると，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多くあると考えられる．

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).