趨勢を考慮した野球の勝敗予測に関する研究

2020SE003 後藤聡一朗　2020SE010 今枝拓斗
指導教員：野呂昌満

1　はじめに

近年，あらゆる分野で ICT 技術の導入が進んでおり，スポーツでも多く導入されている [1][2][3][4]．スポーツ分野での ICT 技術の代表例として機械学習による勝敗予測がある．野球においても，機械学習による勝敗予測は多くの試みがあるが，それらの精度は低い [5]．野球には選手成績の他に，球場や天候，日程，試合の趨勢 (流れ) のように試合結果に影響を及ぼす要因が数多くある．本研究では，それらの要因のうち，野球において試合の趨勢が試合結果に大きな影響を与えると仮定した．そして，機械学習による野球の勝敗予測の精度が低い理由として，試合の趨勢の考慮が十分でないことが挙げられると考えた．

本研究の目的は，試合データから趨勢を分析し，勝敗予測に与える影響を考察することである．選手成績のように，具体的な数値として現れない趨勢を分析し，勝敗予測の特徴量の一つとして扱うことで，野球の勝敗予測の精度を高めることができると考える．

本研究の技術課題として，RQ1：特徴量の決定，RQ2：ニューラルネットワークの設計，RQ3：ニューラルネットワークの妥当性検証を挙げる．

2　野球の勝敗予測の課題

2.1　関連研究

Tomislav は，スポーツの結果予測における既存の機械学習アルゴリズムをレビューすることを目的に，100 を超える論文を分析した [5]．図 1 のそれぞれのスポーツにおける予測精度の中央値は，フットボールはおよそ 74%，バスケットボールはおよそ 69%，サッカーはおよそ 64%，クリケットはおよそ 63%，野球はおよそ 56% である．図 1 の全体の予測精度の分布をみても，野球の勝敗予測の精度が他のスポーツと比べて低いことが分かる．

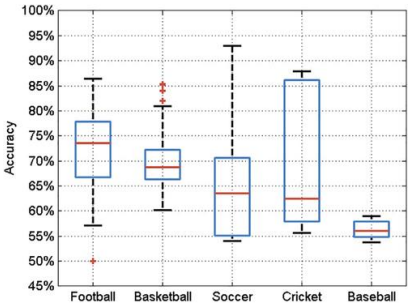

図 1　勝敗予測精度の比較 [5]

Shu-Fen らは MLB(Major League Baseball) における次の試合の結果を予測する研究を行った [6]．予測手法として，1DCNN，ANN，SVM(サポートベクターマシン)，ロジスティック回帰を用いた．MLB の 2015 年から 2019 年までの 30 チームの試合データを収集し，30 のデータセットを作成し，勝敗予測を行った．特徴量は，OBP(出塁率) や OPS(出塁率 + 長打率) などの打者に関するものが 15 個，ERA(防御率) や WHIP(与四死球・被安打/回) などの投手に関するものが 8 個である．結果は，SVM を用いた方法が最も高く，およそ 65% であった．

Andrew は機械学習を用いた MLB の試合結果予測を行った [7]．予測手法として，正則化ロジスティック回帰エラスティックネットを用いた．主な特徴量は，OBP，休養日，ISO(打者の長打力を示す指標)，WHIP，のホームチームとビジターチームのパーセンテージ差である．2001 年から 2015 年のデータを訓練データ，2016 年から 2019 年のデータをテストデータとして勝敗予測を行った．予測精度は 61.77% であった．

Juliette は深層学習を用いた野球の勝敗予測を行った [8]．予測手法として，LSTM を用いた．データセットとして，各年度の要約統計と各チームのスケジュールを作成している．また，イニング数，対戦相手，試合時間，試合が行われた時点でのチームの順位に関する統計も抽出し，試合結果を時系列に，年間の要約統計を行列に再フォーマットしている．最終的な予測精度は 58.6% であった．

2.2　問題点

機械学習を用いた野球の勝敗予測の精度が他のスポーツに比べて低い理由として，試合の趨勢 (流れ) の考慮が十分でないことが挙げられる．得点や投手成績といった試合展開や選手交代，ファインプレーなどの価値あるプレーが趨勢の要因として挙げられる．これらの要因による趨勢は，野球の試合において重要な要素として考えられる．

本研究では，選手成績や Home/Visitor に加え，趨勢を用いて勝敗予測を行うことで，試合の趨勢の考慮が十分でないという問題を解消する．先行研究では，試合前時点での勝敗予測を行っているが，本研究では趨勢を考慮した勝敗予測を行うので，試合進行と同時での勝敗予測を行う．

3　課題解決のアプローチ

本研究では，試合データ (各イニングの得点・点差・打者数) から趨勢を分析し，分析した趨勢と選手成績・Home/Visitor から勝敗予測を行う．本研究の全体図を図 2 に示す．

まず，選手成績のように単純に求めることができない趨勢を試合データから分析する．試合データによる勝敗予測を行い，出力された各イニングの勝利確率から各イニングの趨勢を算出する．次に趨勢を考慮した勝敗予測を行う．

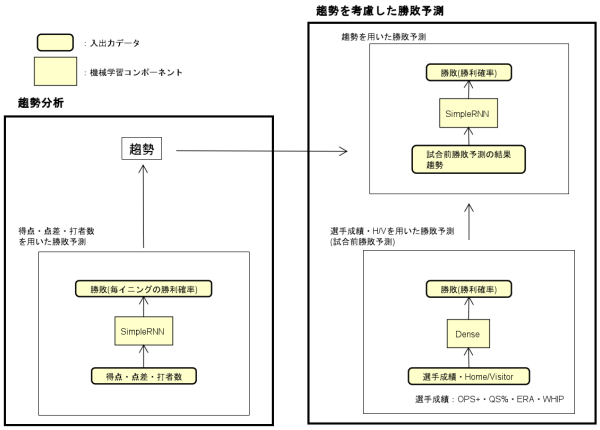

図 2　全体図

選手成績と Home/Visitor による試合前勝敗予測を行った後，試合前勝敗予測の結果と分析した趨勢を用いて勝敗予測を行う．

3.1　趨勢分析

本研究では，WPA(Win Probability Added) という指標を参考に，毎試合の趨勢を分析する．WPA はプレーした状況によって選手をダイナミックに評価する指標であり [9]，評価する対象をイニングに変更することで，各イニングの趨勢を求めることができると考える．試合データ (各イニングの得点・点差・打者数) から毎イニング終了時の勝利確率を求め，求めた勝利確率から趨勢を求める．野球の試合では，趨勢変化の要素は大量に存在する．したがって，すべての要素をもとに趨勢を分析することは不可能に近い．そこで本研究では，趨勢を得点・点差・打者数の 3 要素に着目し，趨勢を分析する．趨勢を求める計算式を以下に示す．

- 1 回の表

$$\text{趨勢} = (1\text{ 回表の勝利確率}) - 0.5$$

- その他のイニング (n 番目のイニング)

$$\begin{aligned}\text{趨勢} = &(\text{n 番目のイニングの勝利確率}) \\ &- (\text{n} - 1\text{ 番目のイニングの勝利確率})\end{aligned}$$

趨勢分析のために行う，得点・点差・打者数による勝敗予測では，時系列データによる勝敗予測を行うので，中間層として SimpleRNN 層を選択した．ユニット数は 32 とした．活性化関数は，中間層では ReLU 関数を，出力層では sigmiod 関数を選択した．最適化手法は Adam を選択した．損失関数は，二値分類問題で多く用いられる二値交差エントロピーを選択した．

3.2　趨勢を考慮した野球の勝敗予測

選手成績と Home/Visitor から試合前勝敗予測を行い，その結果と求めた趨勢を用いて趨勢を考慮した勝敗予測を行う．

試合前勝敗予測では，野球において勝敗の影響が大きいと考えられる選手成績と Home/Visitor を特徴量として選択し，勝敗予測を行う．試合前勝敗予測では，中間層として全結合層 (Dense 層) を選択した．ユニット数は 32 とした．活性化関数として，中間層で ReLU 関数を，出力層では sigmoid 関数を選択した．最適化手法は Adam を選択した．損失関数は二値交差エントロピーを選択した．

試合前勝敗予測を行った後，試合前勝敗予測の結果と趨勢分析にて求めた趨勢を特徴量とし，趨勢を用いた勝敗予測を行う．趨勢を用いた勝敗予測では，時系列データによる勝敗予測を行うため，中間層として SimpleRNN 層を選択した．ユニット数は 32 とした．活性化関数として，中間層で ReLU 関数を，出力層では sigmoid 関数を選択した．最適化手法は Adam を選択した．損失関数は二値交差エントロピーを選択した．

3.3　対象データ

本研究では, OPS+, QS%, WHIP, ERA, Home/Visitor に加え，趨勢を取り扱い，勝敗予測を行う．

打撃力を考慮するために OPS+ を用いる [10]．打者を評価する指標として用いられる OPS をパークファクターにより正規化した OPS+ は，球場や年度の異なる選手を平等に評価できると考えた．投手力を考慮するために QS% と WHIP，ERA を用いる [11][12]．QS% は試合に大きく影響を与える先発投手を評価する成績として用いる．WHIP と ERA は投手の役割に関係なく，全投手を評価する成績として用いる．OPS+ と QS% は，少ない打席数や登板数による成績の上振れを考慮するために，一部成績の補正を行う．特徴量として用いる選手成績のうち，WHIP と ERA は補正を行わない．WHIP と ERA はベンチ入り投手の平均を特徴量として扱う．

趨勢を求めるためのデータとして以下の 3 点を取り扱う．

- 得点

  得点はそのイニングの趨勢を容易に分析できる要素であるといえる．簡単に考えれば，分析するイニングにおいて，攻撃側の趨勢が良ければ (守備側の趨勢が悪ければ)，そのイニングでは得点が入り，反対に，攻撃側の趨勢が悪ければ (守備側の趨勢が良ければ)，そのイニングでは得点が入らないといえる．

- 点差

  野球は 9 イニング (延長・コールドなしの場合) の間に，多くの得点を取ったチームが勝利するスポーツであり，毎イニング終了時の点差は趨勢に大きく関わる．1 イニングで 3 点を取った場合でも，そのイニング終了時にチームが勝っているのか負けているのか，あるいは同点なのかにより，そのイニングの趨勢が異なる．

- 打者数

  打者数により，1 イニングで費やした打者の数や得点

も考慮することで残塁数を把握することができ，攻撃の質を知ることができる．

4　プロトタイプの実験と評価

4.1　開発環境

データセット作成に必要なスクレイピングを行うための環境として，JupyterLab を使用する．使用言語は Python を用いる．データ処理を行うライブラリとして Pandas を用いる．

プロトタイプを実装するための環境として，Google Colaboratory を使用する．Google Colaboratory とは，Google が提供しているウェブブラウザ上で python を実行できるサービスである．機械学習ライブラリとして Keras を用いる．また，その他ライブラリとして Pandas や NumPy，scikit-learn を用いる．

4.2　データセットの作成

集めるデータは，過去 5 年分の中日ドラゴンズの公式戦 621 試合の試合データ (引き分け・延長・コールドを除く) と選手成績である．データ収集には，nf3-Baseball Data House[*1]と NPB.jp 日本野球機構[*2]，プロ野球データ Freak[*3] を使用する．

作成したデータセットは 621 行 ×81 列である．データセットの列の内訳は，試合情報，勝敗，Home/Visitor，OPS+，QS%，WHIP，ERA，1 回表から 9 回裏までの得点・点差・打者数である．

4.3　実験と評価

作成したデータセットを用いて実験を行う．作成したデータセットを 訓練データ : テストデータ = 8 : 2 の比率で分割した．実験結果を表 1 と図 3，図 4，図 5 に示す．

表 1　実験結果

	Accuracy
試合前勝敗予測	0.5680000185966492
趨勢分析のための勝敗予測	0.7995555400848389
趨勢を用いた勝敗予測	0.785263180732727

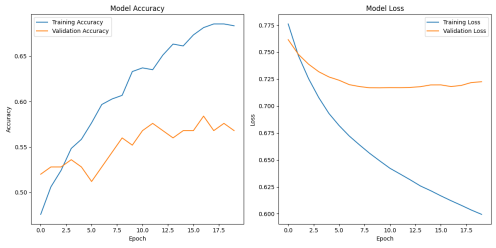

図 3　試合前勝敗予測の結果

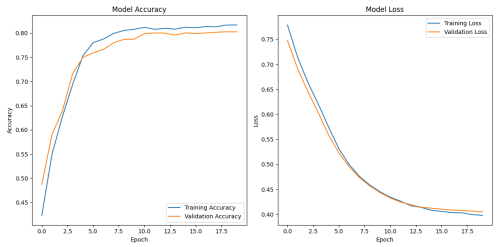

図 4　得点・点差・打者数による勝敗予測 (趨勢分析のための勝敗予測) の結果

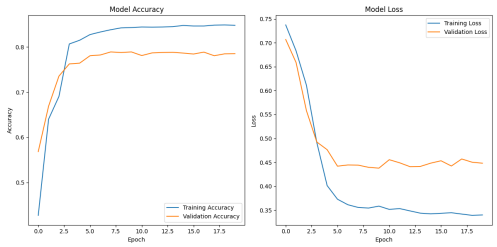

図 5　趨勢を用いた勝敗予測の結果

5　考察

5.1　野球における趨勢の影響

実験の結果，試合前勝敗予測の予測精度は 56.8%，趨勢を用いた勝敗予測の予測精度は 78.5% であった．趨勢を用いた勝敗予測の予測精度は，試合前勝敗予測と比べ精度が高い．また，図 1 の野球の勝敗予測の 56%(中央値) という予測精度と比べても，趨勢を用いた勝敗予測の予測精度は高い．この結果から，趨勢は野球の試合結果に大きく影響するといえる．

本研究では，イニング単位の趨勢に着目し，趨勢を得点・点差・打者数の 3 要素から分析した．しかし，趨勢は打者単位でも考えられ，四死球や先頭打者の結果，エラーなどが要因として挙げられる．本研究で扱った，趨勢分析のための 3 要素は，これらの打者単位での趨勢の要因を含意していると考えており，打者単位の趨勢を用いた勝敗予測を行っても，同様の結果が得られるのではないかと考える．

以上のことから，野球において趨勢が試合結果に大きく影響するのではないかと考えられる．また，趨勢を勝敗予測の特徴量の一つとして扱ったことで，野球の勝敗予測の精度を高めることができたといえる．本研究で行った趨勢を考慮した勝敗予測は，試合中のコーチングへの応用が期待される．

5.2　今後の課題

今後の課題として，特徴量の追加，学習データの増加，アーキテクチャの改善が挙げられる．

5.2.1　特徴量の追加

本研究では，試合前勝敗予測の特徴量として OPS+・QS%・WHIP・ERA・Home/Visitor を，趨勢を用いた勝

[*1] https://nf3.sakura.ne.jp/
[*2] https://npb.jp/
[*3] https://baseball-data.com/

敗予測の特徴量として試合前勝敗予測の結果・趨勢を用いた．試合前勝敗予測の特徴量は，打者に関するものが 1 つ，投手に関するものが 3 つ，その他が 1 つである．これは，先行研究と比べ，扱う特徴量の数が少ないといえる．

打者に関する OBP(出塁率) や IsoP(打者の純粋な長打力を表す指標) などや投手に関する K/9%(9 イニングあたりの奪三振数) や BB/9%(9 イニングあたりの与四死球数)，K/BB%(1 死四球与える間に奪える三振数) といった指標を特徴量として追加することで，選手をより詳細に評価することができると考える．また，Home/Visitor に加え，パークファクターを追加することで，球場ごとの有利不利を判断できると考える．

5.2.2 学習データの増加

本研究では，2019 年から 2023 年の中日ドラゴンズ公式戦 (引き分け・延長・コールドを除く)621 試合を収集しデータセットを作成し，実験を行った．その結果，試合前勝敗予測と趨勢を用いた勝敗予測では，過学習を起こした．過学習を起こす原因の一つとして学習データの不足が挙げられる．**2.1** の先行研究と比べても，データ数が不足しているといえる．したがって，学習データを増やすことで，過学習の防止や予測精度の向上が考えられる．

5.2.3 アーキテクチャの改善

本研究では，複数のニューラルネットワークを用いることで趨勢を考慮した勝敗予測を実現している．単一のニューラルネットワークによる実現やそれぞれのニューラルネットワークの同士の関係を改良することで，予測精度の向上が見込まれる．

本研究では，勝敗予測を Dense 層と SimpleRNN 層のみで行ったが，LSTM などの異なる機械学習アルゴリズム用いた勝敗予測を行い，比較することで，最適な機械学習アルゴリズムの選択や構造の改善点などが発見でき，予測精度の向上につながると考える．

6 おわりに

機械学習を用いた野球の勝敗予測の試みがあるが，それらの精度は低い．本研究の目的は，試合データから趨勢を分析し，趨勢が試合結果に与える影響を考察することである．勝敗予測の特徴量の一つとして扱うことで，野球の勝敗予測の精度向上を図る．趨勢を考慮した野球の勝敗予測の技術課題は，RQ1：特徴量の決定，RQ2：ニューラルネットワークの検証，RQ3：ニューラルネットワークの妥当性検証である．

本研究では，WPA という指標を参考にイニングを評価することで，趨勢を求め，趨勢を考慮した勝敗予測を行った．実験結果は，試合前勝敗予測は 56% 前後，趨勢を用いた勝敗予測は 78% 前後であった．

図 1 の野球の勝敗予測の 56%(中央値) や試合前勝敗予測と比べ，趨勢を用いた勝敗予測の方が予測精度が高いことから，野球の試合において，趨勢が勝敗に影響を及ぼすといえる．今後の課題として，特徴量の追加，学習データの増加，アーキテクチャの改善が挙げられる．

参考文献

[1] Baseball Geeks, “「トラックマン」とは？最先端の計測機器で取れるデータを紹介！！,” https://www.baseballgeeks.jp/measurement/トラックマンとは？/, (Accessed 2024.1.3)

[2] Rapsodo Japan, “Rapsodo Japan(ラプソード) - 野球, ” https://rapsodo.co.jp/pages/baseball , (Accessed 2024.1.3)

[3] 東京ヤクルトスワローズ, “ホークアイ｜東京ヤクルトスワローズ, ” https://www.yakult-swallows.co.jp/pages/info/players/hawkeye , (Accessed 2024.1.3)

[4] スポーツ報知, “「いい球を打たれたのは何で？」ＭＬＢやヤクルトなどで導入「ホークアイ」でわかること…スポーツテック野球編, ” https://hochi.news/articles/20221228-OHT1T51222.html , (Accessed 2024.1.3)

[5] Tomislav Horvat, “The use of machine learning in sport outcome prediction: A review, ” *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2020

[6] Shu-Fen Li, Mei-Ling Huang, Yun-Zhi Li, “Exploring and Selecting Features to Predict the Next Outcomes of MLB Games, ”, *Entropy*, MDPI, 2022

[7] Andrew Y. Cui, “Forecasting Outcomes of Major League Baseball Games Using Machine Learning, ” *Archived Projects 2020*, University of Pennsylvania: Philadelphia, 2020

[8] Juliette Love, “Baseball Win Predictions, ” *Spring 2020 Submissions*, Stanford University, 2018

[9] 1.02 Essence of Baseball, “WPA(Win Probability Added) ｜ Glossary ｜ 1.02, ” https://1point02.jp/op/gnav/glossary/gls_explanation.aspx?eid=20064, (Accessed 2024.1.8)

[10] すさたまくんと B リーグのブログ, “セイバーメトリクスの OPS ＋について, ” https://susatama.com/2018/02/06/2017ops/, (Accessed 2023.9.21)

[11] プロ野球データパーク, “セイバーメトリクス 指標一覧｜プロ野球データパーク, ” https://baseball-datapark.skr.jp/sabermetrics/glossary/ , (Accessed 2023.9.21)

[12] 日本野球機構, “記録の計算方法｜野球の記録について｜ NPB.jp 日本野球機構, ” https://npb.jp/scoring/calculation.html , (Accessed 2023.9.21)