機械学習を用いた野球の勝敗予測に関する研究

2019SE050 櫻井一成　2019SE059 冨山蓮
指導教員：沢田篤史

1　はじめに

人工知能技術をスポーツに活用する事例が増えている．元来，スポーツ界にはデータを活用する傾向がある．実際に，野球界では打撃力や守備力，走塁力などの要因から総合的にどのくらいチームの勝利に貢献しているのかを数値化した WAR と呼ばれる指標を用いて選手を評価したりしている．

特に，野球界では人工知能を用いて膨大なデータを活用する動きが盛んとなっている．メジャーリーグでは，投球のストライク判定を人工知能に任せている例が実際にある．

一方で，野球というスポーツは人工知能を用いた勝敗予測の場面において，他のスポーツと比べ精度が低い傾向にある [1]．野球はチームスポーツであり各選手の成績がそれぞれ異なる点や様々な球場，天候，戦術など試合の結果に影響を与える要因が多く存在するので，正確な予測が困難である．

本研究の目的は，限られたデータを用いて野球の勝敗予測を高い精度で行う方法を考察することである．現在，高校野球でも勝利に近づくために試合のデータなどを活用しているが，得られるデータは限られている．しかし，野球の勝敗に大きな影響をもたらすデータをチームの勝利に活用することは可能であり，同時に勝敗予測の精度も高めることが可能であると考える．

野球の勝敗予測を行う上での技術課題は，以下の通りである．

1. データを用いた野球の勝敗予測のための一般的な機械学習モジュールの提案
2. 提案したモジュールの妥当性の確認

上記の技術課題に対するアプローチとして勝敗に大きな影響を与えていると予想される投手力と打撃力の各要因を示した特徴量と試合の進捗状況を示した特徴量を選択する．特徴量の時系列を考慮するために LSTM（Long Short-Term Memory）を選択する．

本研究では提案した手法に対して Python を用いて実現する．必要なデータを Web サイト上からスクレイピングを行い収集する．

評価実験において，訓練時には高い精度を得られていたが，テスト時では 50% を下回る結果であった．このような結果となった理由を考察し，よりよい精度を誇る野球の勝敗予測を行うための今後の課題をいくつか挙げた．

2　スポーツの勝敗予測に関する課題

2.1　既存研究

Rahul の研究 [2] はクリケットを対象に 90% を超える予測精度である．Nazim らの研究 [3] ではサッカーを対象に最高で約 80%，平均でも 75% を超える予測精度が得られている．一方で，野球の勝敗予測の精度は 50% から 70% 程度しか得られていない研究がほとんどである [4][5][6]．

横井の研究 [4] では，投手起用を例として勝敗予測を行っている．隠れ層の活性化関数として ReLU 関数，出力層の活性化関数としてソフトマックス関数を用いている．予測精度は 65% である．

Juliette の研究 [5] では，イニング数や対戦相手，試合時間，試合が行われた時点でのチームの順位など、チームスケジュールの各試合に関する統計を引き出し，LSTM に入力した予測方法を取っている．予測精度は 59% である．

Soto の研究 [6] では，予測手法が SVM，ANN，決定木，Lazy-Lerning（KNN）であり，予測精度は 59% である．この研究では，Soto が独自に算出したデータが一部用いられている．

Mei-Ling らの研究 [1] では，様々な打者のデータや投手のデータ計 30 項目のデータを特徴量としている．出力は勝ち，負けの 2 つである．予測手法は 1DCNN，ANN，SVM であり，予測精度は 94% である．高い精度を得られたが，特徴量が多く馴染みのないデータなど入手困難なデータを用いている．

2.2　問題点

野球の勝敗予測の精度が比較的高くないのにはいくつかの問題点があるからである．その 1 つに，各選手の成績や天候や監督の采配，試合スケジュールなど，試合に影響を与える要因が多く存在するからである．

野球はチームスポーツでありながらも，打者と投手の 1 対 1 の場面が存在するなど非常に複雑なルールがあるのも要因の 1 つである．プロ野球では先発ピッチャーにローテーションが組まれていて，基本的に 1 週間に 1 回程度しか登板しない．したがって，日によって先発投手の力量が大きく変わる．つまり，両チームの先発投手の組み合わせが勝敗に大きな影響を与える．さらに，1 対 1 の場面において相性などの要素が大きな影響を与えるので，防御率があまり良くない投手が強力な打線を無失点で抑えてしまうこともよくある．実際，シーズンを通して首位のチームが最下位のチームに直接対決で負け越すこともある．

先行研究では試合前に特徴量を入力して予測をする方法がとられていて，試合間で変わる心境やそれによる試合

の流れが考慮されていない．この問題に対しては，過去のデータを再帰させることで時系列を追うことに特化したLSTMを使用することによって，点差かつイニング数による試合中の点差の変動を情報として取り入れることが可能であり，勝敗予測の精度を高めることができると考える．

3　目的と課題

本研究の目的は，限られたデータを用いて高い精度を誇る勝敗予測を行うことである．[1] の研究のように精度の高い勝敗予測を得られている研究もあるが，用いられているデータの中には Mei-Ling らがデータの取得に用いた Web サイトが独自に算出したデータなどが含まれている．

最近では勝利を得る可能性を高めるためにデータを活用する動きが盛んである．特に，プロ野球チームなどの優れた環境では守備範囲なども定量化されている．しかし，高校野球などではチームによって環境，得られるデータ，データを活用するための人員などに大きな差がある．

そのような環境の差を考慮したうえで勝敗予測の精度を高めるためには，限られたデータの中で勝敗に大きな影響を与える特徴量を選択する必要がある．環境の差に依存することなく，精度の高い勝敗予測を行うことが可能になれば，データの収集力が高くなくとも試合前にどちらのチームが優勢であるのかなどを定量的に判断できるようになる．

目的を達成するために解決すべき技術課題を次のように設定する．

1. データを用いた野球の勝敗予測のための一般的な機械学習モジュールの提案
2. 提案したモジュールの妥当性の確認

1 つ目は，データを扱うことで野球の勝敗予測を行う上での，一般的な機械学習モジュールを提案することである．今までの勝敗予測では試合前に予測を行う機械学習モジュールが多く，選手の調子や試合の流れなどの情報を活用できていない．しかし，各試合のイニング毎の点差の変動のような長期依存するデータに対して，時系列データの予測に特化した LSTM を使用することで勝敗予測の精度を向上させることが可能であると考える．

2 つ目は，提案した機械学習モジュールの妥当性の確認である．1 つ目の技術課題が本研究の目的に沿っているかを確認する必要がある．そのために，本研究で提案する機械学習モジュールにおいて，用いるデータや予測モデルが適切であるかを，実際に野球の勝敗予測システムを設計し，実験を行うことでその妥当性を確認する．

4　課題解決へのアプローチ

4.1　研究方針

本研究では，**3 章**で挙げた課題を解決するために以下のような研究方針で進めていく．

1. 時系列データを用いた野球の勝敗予測モデルの設計
2. 野球の勝敗に大きな影響を与えると考えられる特徴量の設定
3. 提案したモデルの実装
4. 3 で実装したモデルを用いた実験
5. 実験結果から 3 で実装したモデルの評価

4.2　野球勝敗システムの設計

4.2.1　システムの概要

本研究におけるシステムの構成図を**図 1** に示す．図左下

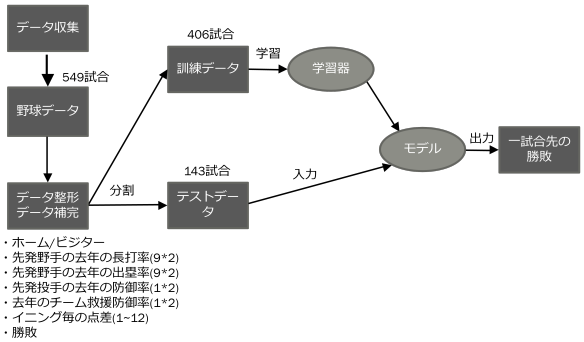

図 1　システムの構成

部のホーム/ビジター，先発投手の去年の出塁率，先発投手の去年の長打率，先発投手の去年の防御率，去年のチーム救援防御率，イニング毎の点差のデータを入力とし，一試合先の勝敗を出力とする．収集した野球データは必要に応じて整形，補完を行う．それらのデータを訓練データとテストデータに分割，訓練データを予測器を用いて学習，学習したデータをモデルとし，分割したテストデータをそのモデルに入力する．

4.2.2　LSTM

本研究では，系列データを扱うのに適したニューラルネットワークを選択する．RNN (Recurrent Neural Network) は，系列データを解析するのを得意としたニューラルネットワークの一種である [7]．RNN はループ構造を持っているので，過去の入力についての情報をネットワーク中に残すことができ，様々な情報を記憶することができる．一方で，RNN には勾配が急速に減少，増加するという勾配消失問題または勾配爆発問題がある．これらの問題があることから，RNN は系列データを扱うのに適しているにも関わらず長期的な記憶を行うのが困難である．

こういった RNN の問題を解決したニューラルネットワークが LSTM である．LSTM と RNN の違いは隠れ層の仕組みにある．LSTM は LSTM セルと呼ばれる複数のユニットから構成されたセルを持ち，それらが多数集まったネットワーク構造である．

4.2.3　予測モデル

本研究では各試合のデータを時系列データとして扱うので **4.2.2 節**より，予測器に LSTM を選択する．中間層の数

は汎化性と演算量を考慮した結果 200 がよいと考えた．活性化関数は中間層では ReLU（Rectified Linear Unit）関数を用いる．ReLU は非線形であり，線形変換と比べより複雑な関数を作ることができる．勝敗を 1，0.5，0 と出力するので回帰をする際に適している linear 関数を出力層で用いる．損失関数の値を最小化するために，最適化手法として Adam を選択する．損失関数はは主に回帰問題で使用される MSE（Mean Squared Error）を用いる．

4.3 対象データ

本研究では，野球において勝敗に影響が大きいものを特徴量として選択することで勝敗予測を行う．特徴量とするのは次の 7 項目である．

- 各先発野手の昨年の出塁率
- 各先発野手の昨年の長打率
- 昨年の先発投手の防御率
- 昨年のチーム救援防御率
- イニング毎の点差
- ホーム/ビジター
- 勝敗

これらの特徴量は，チームの打撃力と投手力，試合の進捗状況，球場を考慮するために抽出した．

各先発野手の昨年の出塁率，長打率はチームの打撃力を考慮するうえで必要だと考えた．昨年の先発投手の防御率，昨年のチーム救援防御率はチームの投手力を考慮するうえで必要だと考えた．

試合の状況を考慮する上では，イニング毎の点差を使用する．点差は勝敗予測に大きな影響を与えると考えた．また，イニング毎の点差のデータは，先ほど挙げた特徴量との相互関係があると考えた．

ホーム/ビジターのデータも特徴量とした．このデータは，対象チームがホーム/ビジターそれぞれでの勝ち負けの傾向を図れると考えたので選択した．

過去の勝敗のデータは過去数試合の出場選手の成績やイニング毎の点差の傾向から，各試合における勝敗がどうであったかを学習することで精度の高い勝敗予測を行えると考えた．

上述したそれぞれのデータにおいて，いくつかデータの補完を行う必要がある．出塁率や長打率，防御率のデータは，昨年度のデータを使用するが，そのままデータを使用すると正しい勝敗予測を行えないと考えた．例えば打者データでは，昨年度の規定打席 1/2 以上を達成していない選手は昨年度の所属チームの出塁率，長打率のデータ用いる．新加入外国人打者は長打を期待されているので昨年度のチーム長打率に 3 分加えるといった補完を行う．また，イニング毎の点差のデータも補完を行う．

本研究では，混合行列の 1 つである Accuracy（正解率）を用いて予測システムの評価を行う．この Accuracy の値によって，提案したシステムが有効であるかを判断する．

5 プロトタイプシステムを用いた実験と評価

5.1 開発環境

提案したシステムを実装するための環境として，Google Colaboratory[*1]（以下 Colab と呼ぶ）を使用する．Colab は Google が提供している，ブラウザ上で Python を実行できるサービスである．

使用言語は Python，機械学習ライブラリは Keras を用いる．また，その他ライブラリとして Numpy や Pandas，Scikit-learn を用いる．

5.2 データセットの作成

LSTM を用いて勝敗予測を行うためには，大量のデータが必要となる．そのため，本研究では **4.3 節**で挙げた特徴量のデータを収集し，1 つのデータセットにまとめる．

データは Web スクレイピングを行って収集する．大量のデータを取得する際，手作業ではなくスクレイピングを行うことで素早く取得することが可能である．スクレイピングを行うにあたってプログラムを Python で作成した．

なお，データを取得するにあたって本研究では，阪神タイガースを対象にデータを収集する．1 つのチームに絞ることで各日ごと，各シーズンごとの時系列を考慮できるのではと考えた．

データの取得にはそれぞれ以下の Web サイトを使用する．

- nf3-Baseball Data House[*2]
- プロ野球データ Freak[*3]
- プロ野球 Freak[*4]

5.3 実験と評価

本研究では，**5.2 節**で作成したデータセットを用いて実験する．作成したデータのうち，2019，2020，2021 年シーズンのデータを訓練データ，2022 年シーズンのデータをテストデータとして使用する．過去 6 試合のデータを入力し，一試合先の試合を出力する．

実験結果は**図 2**，**図 3** となった．

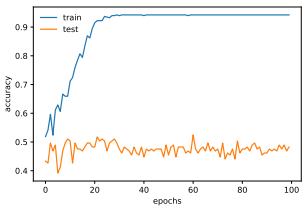

図 2 accuracy

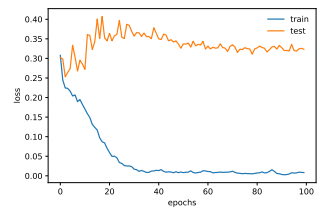

図 3 loss

実験を行った結果，訓練時は 90% を超える精度を得ら

[*1] https://colab.research.google.com/
[*2] https://nf3.sakura.ne.jp/
[*3] https://baseball-data.com/
[*4] https://baseball-freak.com/

れ，損失も過学習を起こすことなく数値が減っていったのに対し，テスト時は訓練時と同じ振る舞いをすることはなく正しく予測を行えていない．

6　考察

5.3 節のような実験結果になった理由として，データ数の不足がある．テスト時は損失の値が増加や減少を起こしたりしているので，新しいデータに対して十分な予測が行えていなかった可能性がある．予測に用いるデータの複雑さも精度が高まらなかった要因の可能性がある．設定した特徴量，すなわち説明変数に多重共線性が起こった可能性もある．この多重共線性も精度が高まらなかった要因の一つだと考えている．また，本研究の補完方法が適切であるかも見直す必要がある．一方で，実際に予測システムを実装し，予測結果を導出できたことから，データを用いた勝敗予測のための一般的な構造の設計は行えたと評価している．

今後の課題として以下のような改善点が挙げられる．

1. 特徴量の追加
2. 学習データの増加
3. 中間層の数の変更
4. 出力の変更
5. 最適化手法や活性化関数，損失関数の変更

7　おわりに

人工知能技術をスポーツに活用する事例が増えている．元来，スポーツ界はデータを活用する傾向がある．実際に，野球界では打撃力や守備力，走塁力などの要因から総合的にどのくらいチームの勝利に貢献しているのかを数値化した WAR と呼ばれる指標を用いて選手を評価したりしている．

特に，野球界では人工知能を用いて膨大なデータを活用する動きが盛んとなっている．メジャーリーグでは，投球のストライク判定を人工知能に任せている例が実際にある．

一方で，野球の勝敗予測の研究において精度が 50% から 70% 程度のものが多く，精度は比較的低い傾向にある．Mei-Ling らの研究 [1] のように 90% を超える高い精度を得ている研究もあるが，そういった研究は馴染みのないデータや独自に算出したデータを用いての予測を行っている．勝敗予測を研究する人によっては得ることが困難なデータも存在し，それらを考慮しなければならない．

実際に，高校野球でも勝利に近づくために試合のデータなどを活用しているが，チームによって環境，得られるデータ，データを活用するための人員などに大きな差がある．しかし，野球の勝敗に大きな影響をもたらすデータを抽出することでチームの勝利に活用することは可能であり，同時に勝敗予測の精度も高めることが可能であると考える．そのため，限られたデータの中で高い精度を誇る勝敗予測を行うことを本研究の目的とした．

野球の勝敗予測を行う上での技術課題は以下の通りであった．

1. データを用いた野球の勝敗予測のための一般的な機械学習モジュールの提案
2. 提案したモジュールの妥当性の確認

上記の技術課題に対するアプローチとして勝敗に大きな影響を与えていると予想される投手力と打撃力の各要因を示した特徴量と試合の進捗状況を示した特徴量を選択した．特徴量の時系列を考慮するために学習器として LSTM を選択した．

本研究では提案した手法に対して Python を用いて実現した．必要なデータを Web サイト上からスクレイピングを行い収集した．

実験結果として訓練時は高い精度を誇っていたが，テスト時は 50% を下回った．また損失の値もテスト時では増加や減少を繰り返すだけで訓練時のような減少はしない結果となった．

上記のような実験結果になった原因としてデータ数が不足しているのではないかと考えられ，予測モデルの改良が必要である．

今後の課題として特徴量の追加，学習データの増加，中間層の数の変更，出力の変更，最適化手法や活性化関数，損失関数の変更などが挙げられる．

参考文献

[1] Mei-Ling Huang, Yun-Zhi Li, “Use of Machine Learning and Deep Learning to Predict the Outcomes of Major League Baseball Matches”, MDPI, applied sciences, 2021.

[2] Rahul Chakwate, Madhan R A, “Analysing Long Short Term Memory Models for Cricket Match Outcome Prediction”, Cornell University, 2020.

[3] Nazim Razali, Aida Mustapha, Faiz Ahmad Yatim, Ruhaya Ab Aziz, “Predicting Football Matches Results using Bayesian Networks for English Premier League (EPL)”, International Research and Innovation Summit (IRIS2017), 2017.

[4] 横井秀哉，“野球の勝敗予測システムの設計—投手起用を例として—”，南山大学理工学部卒業研究要旨集，2021.

[5] Juliette Love, “Baseball Win Predictions”, Stanford University, 2018.

[6] Soto-Valero, C, “Predicting win-loss outcomes in MLB regular season games-a comparative study using data mining methods”, International Jounal of Computer Science in Sport, 2016.

[7] 手塚 太郎, “しくみがわかる深層学習”, 朝倉書店, 2018.