2S-05

深層学習を用いた野球の局面予測

高石 和輝 †　　松澤 智史 ‡　　武田 正之 ‡
東京理科大学 理工学部 情報科学科 † ‡

1. はじめに

スポーツ界ではデータ解析を用いる動きが盛んになっている。しかし、従来のデータ解析は統計手法を用いたものが多く、機械学習が用いられている例は少ない。また、機械学習の中でも任意のデータを学習することで非線形な特徴量を抽出できる深層学習 (ニューラルネットワーク) が近年多くの分野で成果を上げていて、スポーツ分野への応用が期待されている。本研究では、深層学習を用いて人気スポーツである野球の試合の局面予測をするシステムを提案する。研究を進めるにあたり、統計手法を用いた予測よりも高い性能を示すシステムを構築する事を目的とする。

2. 本研究で行う予測

野球の中で特に重要な役割である投手 (ピッチャー)、捕手（キャッチャー)、打者（バッター）の 3 人に着目し、以下の予測を深層学習を用いて行う。

- 投手の投げるボールのコース予測 NN(以後コース NN と呼ぶ)
- 投手の投げるボールの球種予測 NN(以後球種 NN と呼ぶ)
- 打者のスイング予測 NN(以後スイング NN と呼ぶ)

また、本研究では最終的な予測であるスイング予測の性能の評価に重点を置く。

3. データセット

本研究ではデータスタジアム株式会社様から頂いた日本プロ野球の 2 年間 (2016,2017 年) の公式記録を用いる。公式記録の中には以下の情報が含まれている。

- 1 球データ (1 球ごとの投球コース・球種・打球結果・各フラグ (バント・盗塁) 等)
- 1 打席データ (1 打席毎の打球結果・飛球コース・打球捕球者・打点等)
- プロフィールデータ（選手の身体情報・生年月日・所属チーム等）

4. 関連研究とその課題

4.1 統計手法を用いたスイング予測

統計手法を用いた研究ではロジスティック回帰を用いる。ロジスティック回帰モデルの式を式 (1) に示す。

$$\log \frac{\pi_i}{1-\pi_i} = \beta_0 + x_i^T \beta \qquad (1)$$

Pitching and Batting prediction by Neural Network
†Kazuki Takaishi ‡Tomofumi Matsuzawa
‡Masayuki Takeda
† ‡Faculty of Science and Technology Dept of Information Science,Tokyo University of Science

n はデータ数 、$x_i(i = 1; 2; ;; ; n)$ は i 番目の説明変数、$\beta(\beta = (\beta_1, ..., \beta_n)^T)$ はパラメータベクトルを示す。ロジスティック回帰分析で得られる確率 π_i の値が閾値を超えた場合スイングと判定する。
予測に使った説明変数を表 1 に示す。データセットは 2016,2017 年のプロ野球の公式試合のデータを使用する。全てのレコード数 515427 のうち、1 割をテストデータとして使用する。

表 1: 説明変数一覧

イニング	表裏	打者左右打席
打者巡数	打者打席数	ランナー状況
S カウント	B カウント	O カウント
チーム総得点	チーム総失点	投手チーム ID
投手腕	投手投順	当該打者投球数
試合投球数		

4.2 課題

統計手法の課題として、選手視点でスイング有無に重要になる投球コース、球種を入力に使用できない点にある。これは予測には投手が球を投げる前に確定する情報しか使えないためである。そこで深層学習を用いて投球コース、球種に該当する情報を予測し、予測した結果をスイング NN の入力に加えれば精度が向上すると考えた。

5. 提案手法

5.1 概要

本手法ではコース NN、球種 NN、スイング NN の順に行う。提案手法の概略図を図 1 に示す。
それぞれ 100000 を訓練データ、100000 をテストデータとする。

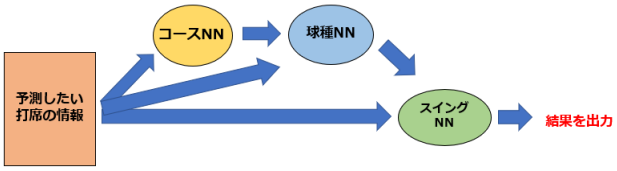

図 1: 提案手法の概略図

5.2 コース NN の学習・予測

入力は表 1 の説明変数と同じものを使用する。出力はストライク、ボール の 2 値とする。与えられたテストデータにコース NN の予測結果を付与する。

5.3 球種 NN の学習・予測

入力は最初のテストデータにコース予測の結果を付与したものである。出力は直球系、曲がる系、落ちる系の 3 値とする。与えられたテストデータに球種 NN の予測結果を付与する。

5.4 スイング NN の学習・予測

入力は最初のテストデータにコース予測と球種予測の結果を付与したものである。出力はスイングする、スイングしないかの 2 値とする。

6. 実験と結果

6.1 評価方法

評価方法については F 値 (F_value) と呼ばれる評価指標を用いた。F 値とは、再現率 (Recall) と適合率 (Precision) と呼ばれる 2 つの評価指標を組み合わせた指標である。予測結果と実際の答えの対応を表 2 にそれぞれの指標の導出式を式 (2)〜(4) に示す。

表 2: 予測結果と実際の答えの対応表

		実際の答え	
		正	負
予測結果	正	TP	FP
	負	FN	TN

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

$$F\_value = \frac{2 \cdot Recall \cdot Precision}{Recall + Precision} \tag{4}$$

6.2 スイング予測 (独立)

6.2.1 概要

コース予測と球種予測が完璧に出来たと仮定した場合、スイング予測がどれくらいの精度で出来るのかを確認するために、球種とコースの正解データを入力として与えたものをコース NN（独立）とし、統計手法と比較した。

6.2.2 結果

深層学習と統計手法の結果の比較を表 3 に示す。

表 3: スイング予測（独立) の精度比較

評価指標	スイング NN	統計手法
再現率	0.81	0.87
適合率	0.85	0.51
F 値	0.83	0.64

6.3 コース予測と球種予測

提案手法に用いるコース予測、球種予測に関しての精度を表 4、5 に示す。

表 4: コース NN（独立) の精度比較

評価指標	ストライク	ボール
再現率	0.46	0.63
適合率	0.47	0.68
F 値	0.47	0.66

表 5: 球種 NN（独立) の精度比較

評価指標	直球系	曲がる系)	落ちる系
再現率	0.90	0.32	0.36
適合率	0.63	0.64	0.62
F 値	0.74	0.42	0.46

6.4 スイング予測（提案手法）

6.4.1 結果

提案手法と統計手法の結果の比較を表 6 に示す。

表 6: スイング予測（提案手法）の精度比較

評価指標	スイング NN	統計手法
再現率	0.81	0.87
適合率	0.53	0.51
F 値	0.64	0.64

7. 考察と今後の展望

7.1 考察

今回の結果ではスイング予測（独立）の結果は統計手法よりも精度が高かったが、スイング予測 (提案手法) と統計手法では大きな差はなかった。原因としては現時点で球種とコース予測の精度が低く、ノイズを持ったデータが多数あったためと考えられる。

7.2 今後の展望

3 つの独立した予測の F 値が 1.0 を大きく下回った理由は以下のデータが不足していたためと考えられる。

- 個人の特徴量を抽出できるパラメータ
- 時系列パラメータ

これらの情報を入力に加えることでさらなる精度の向上が期待できる。

8. まとめ

本研究では、人気スポーツであるプロ野球の投手の投球コース、球種予測と打者のスイング予測を行うシステムを構築した。また、従来使われていた統計手法を用いずに予測性能を高めることができた。その結果、入力のパラメータを深層学習によって生成して学習することによって予測精度が上昇することを示せた。

参考文献

[1] F 値-機械学習の朱鷺の杜 Wiki
http://ibisforest.org/index.php