ウェブインテリジェンス特論最終レポート課題

ベストナイン賞受賞選手はどの指標が優れているか

2722202018　1 年 12 組 18 番　松山直人　提出日: 2020 年 8 月 6 日

1. 概要・目的

プロ野球では，毎年シーズンで好成績を残した選手をポジションごとに記者投票によって選出される「ベストナイン賞」が存在する．本レポート課題では，プロ野球における様々なデータや指標を用いて，どのような選手がベストナイン賞を受賞しやすいのかについて分析を行った．具体的には逐次特徴選択アルゴリズムとランダムフォレストを用いて，ベストナイン賞を受賞した選手はどの指標が優れているかを予測した．その結果，安打や安打に関連する指標が関係している可能性が示唆された.

2. データ

2.1 データの入手方法

分析に用いたデータはプロ野球 Freak[1]というサイトからスクレイピングによって収集した．こちらのサイトでは 2009 年〜2019 年の選手のプロフィールや個人打撃成績や投手成績等が 12 球団ごとに記録されている．今回は，これらのデータを用いて，その年の選手のプロフィールと個人打撃成績を組み合わせた csv ファイルを作成した[2].

具体的には, まずプロフィールデータにおいて, 年齢や年俸などについている単位を削除し，数値として扱えるようにする．その後，予測に関係ないと思われる列（背番号，生年月日，血液型，出身地）を削除し, csv ファイルとして保存する. 次に, 打撃成績について，こちらはプロ野球 Freak では「基本成績」「総合指標」「その他の指標」の 3 つの表に分かれていたため，これらをスクレイピングしたのち選手名情報を元に統合し，個人打撃成績データを作成した．最後に，プロフィールデータと個人打撃成績データを統合した.

2.2 データの整形

次にデータの整形について説明する．今回収集したデータにおいて，打撃データが記録されていない選手がいたため，その選手の行を削除する処理を行った．また，細かい処理ののち,「守備」や「投打」といった名義特徴量を one-hot エンコーディングによりダミー特徴量を作成した．その後，全チームの打撃成績データを統合した．また，データが Object 型であったため，float 型に直し数値として扱えるようにしている.

ここで，今回の分析ではその年にベストナイン賞を受賞したかどうかという情報も重要になるが, これに関してはデータセットが存在しなかったため，著者が全て手打ちで入力した．なお，ここではベストナイン賞を受賞した選手は 1，しなかった選手は 0 が記入されている．なお，以上の処理によって作成したファイルは 55 個のカラムからなる(表 1).

3. アルゴリズム

本レポートでは逐次特徴選択アルゴリズム，ランダムフォレストの 2 つを用いて分析を行った[3]. そのうち，本レポート課題で用いたアルゴリズムである逐次後退選択（SBS）の手順は以下の通りである[3].

1. アルゴリズムを$k = d$で初期化．（dは全特徴空間x_dの次元数）
2. Jの評価を最大化する特徴量x^-を決定する．（$x \in X_k$）

$$x^- = argmax\, J(X_k - x)$$

3. 特徴量の集合から特徴量x^-を削除

$$X_{k-1} := X_k - x^-; k := k - 1$$

4. kが目的とする特徴量の個数に等しくなれば

表1 分析に用いるデータのカラム名

選手名	best_nine	年齢	年数	身長	体重	年俸(推定)	打率
試合	打席数	打数	得点	安打	二塁打	三塁打	本塁打
塁打	打点	盗塁	盗塁刺	犠打	犠飛	四球	敬遠
死球	三振	併殺打	出塁率	長打率	OPS	NOI	GPA
IsoP	IsoD	RC	RC27	XR	XR37	BABIP	SecA
TA	本塁打率	三振率	四球率	PSN	BB/K	守備_内野手	守備_外野手
守備_投手	守備_捕手	投打_右両	投打_右右	投打_右左	投打_左右	投打_左左	

終了する．そうでなければ，2 に戻る．

4. 環境

本レポートで用いたライブラリは scikit-learn や Matplotlib，NumPy，Pandas，urllib を用いている．また，スクレイピングや機械学習は全て Jupyter notebook 上で行った．

5. 実験

実験は，逐次特徴選択アルゴリズムにおける近傍点数の増減や，ランダムフォレストにおける決定木の個数の増減などを行い，どの特徴量が重要かを分析する．具体的には，逐次特徴選択アルゴリズムでは近傍点数を 5，10 にした時の最小限の特徴部分集合を，ランダムフォレストでは決定木が 500，2000 にした時の特徴量の重要度を算出する．

6. 結果

6.1 逐次特徴選択アルゴリズム

近傍点数が 5, 10 の時の縦軸がスコア，横軸が特徴量の個数で示したグラフが図 1，図 2 である．ここでは, 近傍点数が 5 個の時は「三塁打」「敬遠」「XR」，10 個の時は「盗塁刺」「四球」「敬遠」が最小限の特徴部分集合であった．また，それぞれの正解率を示したものが表 2 である．これらより，近傍点数=5 の時の方が正解率が高く，「三塁打」「敬遠」「XR」が最小限の特徴部分集合であることがわかった．なお，XR は「得点を生み出す能力を

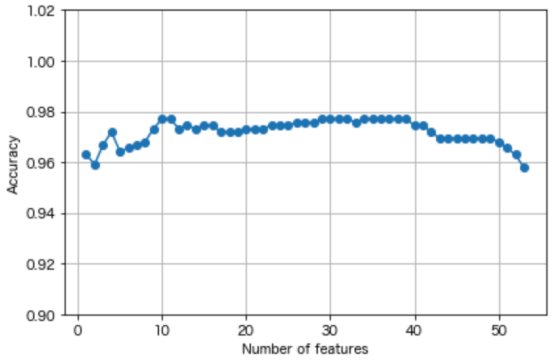

図1 近傍点数=5
（縦軸：スコア，横軸：特徴量の個数）

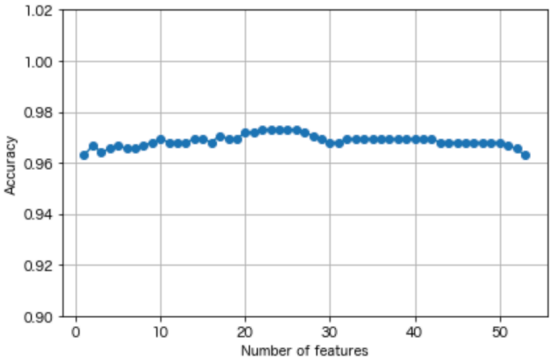

図2 近傍点数=10
（縦軸：スコア，横軸：特徴量の個数）

表 2 最小限の特徴量を用いた時の正解率

	近傍点が 5	近傍点が 10
Training Accuracy	0.973	0.963
Test Accuracy	0.969	0.964

評価する指標」で，「0.50×単打＋0.72×二塁打＋1.04×三塁打＋1.44×本塁打＋0.34×（四球＋死球－敬遠四球）＋0.25×敬遠四球＋0.18×盗塁－0.32×盗塁死－0.090×（打数－安打－三振）－0.098×三振－0.37×併殺打＋0.37×犠飛＋0.04×犠打」で表される（安打の価値が高い）．

6.2 ランダムフォレスト

決定木が 500，2000 の時のそれぞれの特徴量の重要度を算出したグラフが図 3，図 4 である．これらより，多少の差はあるが「XR」「RC」「安打」が重要な特徴量であることがわかった．なお，RC は XR と同じく「得点を生み出す能力を評価する指標」で，「RC＝（2.4×C＋A）×（3×C＋B）÷9×C－0.9×C（A=安打+四球+死球-盗塁死-併殺打，B=塁打+0.26×（四球+死球）+0.53×（犠飛+犠打）+0.64×盗塁-0.03×三振，C=打数+四球+死球+犠飛+犠打）」で表される．一方で，逐次特徴選択アルゴリズムでは高かった「三塁打」は，ランダムフォレストでは 53 個中 44 番目であり，特徴量として重要ではなかった．

以上の結果より，安打に関係する指標がベストナイン賞を受賞した選手に関係していることがわかった．

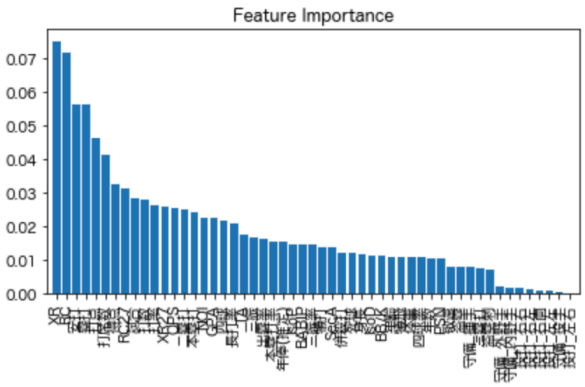

図 3 決定木 500 時の特徴量重要度

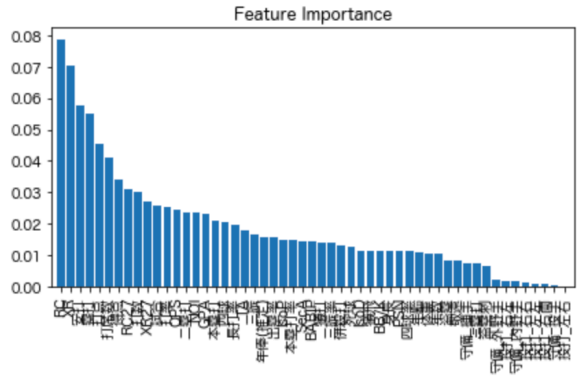

図 4 決定木 500 時の特徴量重要度

7. パーセプトロンのトレーニング

6 章で得られた特徴量を組み合わせて，どの組み合わせがうまく分類できるかを分析する．今回は，「敬遠」「安打」「RC」「XR」から 2 種類ずつ組み合わせ，どの組み合わせが正解率が高くなるかを調査した．その結果のうち，最も正解率の高かった「安打」と「XR」が図 5 である．これより，「安打」と「XR」の組み合わせでは正解率が 0.97 で，分類ミスは 40 個と最も少なかった．以上より，ベストナイン賞受賞選手は「安打」と「XR」の指標が高く，これらの指標が高いとベストナイン賞を受賞する可能性が高いことが示唆された．

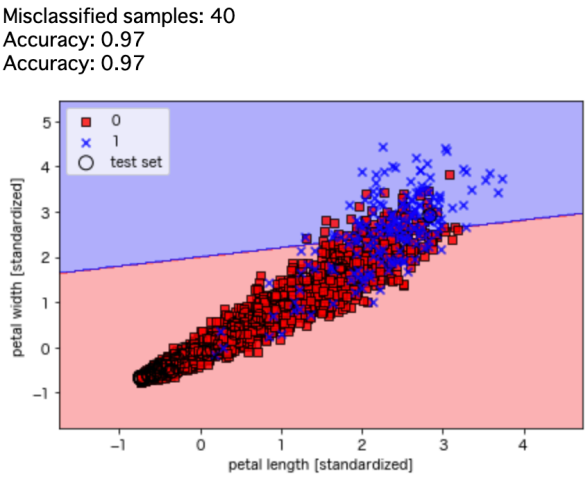

図 5 特徴量を「安打」と「XR」にした時のプロット（0=ベストナイン賞受賞なし，1=ベストナイン賞受賞あり）

8. まとめ

本レポートでは，プロ野球における「ベストナイン賞」に着目し，ベストナイン賞を受賞した選手はどの指標が優れているかについて調査した．逐次特徴選択アルゴリズムとランダムフォレストを用いて分析した結果,「安打」と「XR」が高い選手が受賞していることがわかった．しかし，ベストナイン賞には守備といった要素も関連してくると考えられ，これを含めることで正解率がさらに高まる可能性がある．そのため，今後は守備率等を含めた分析を行う必要があると考えられる．

参考文献

[1]プロ野球 Freak（https://baseball-freak.com/）.
[2] https://note.com/data_science/n/n02f50eae644b
[3]Sebastian Raschka, Vahid Mirjalili. [第 2 版]Python 機械学習プログラミング 達人データサイエンティストによる理論と実践. p.52-58, p.129-135.