

# 潜在能力と暫定成績の乖離から読み取る 成長期待値の高さに着目した推薦システムの提案

神 一輝<sup>†</sup> 張 建偉<sup>††</sup>

<sup>†</sup> 岩手大学大学院総合科学研究科 〒020-8551 岩手県盛岡市上田4丁目3-5

<sup>††</sup> 岩手大学理工学部 〒020-8551 岩手県盛岡市上田4丁目3-5

E-mail: <sup>†</sup>{g0322102,zhang}@iwate-u.ac.jp

**あらまし** 近年スポーツ界隈におけるデータ分析技術は著しく発展し、様々な角度から選手を分析したデータが手に入るようになった。新たなシステムの導入により、選手の成績だけでなく潜在能力も明確に数値化されたことにより、選手の起用や采配に新たな視点が加わるようになったことで、高い能力を持っているにも関わらず成績に伸び悩むような、成長期待値の高い選手もより明白になった。本研究は、数多くの有望選手から忍耐強く育成するべき選手を選定する基準として機能するシステム作りを目的とし、x-means法と階層的クラスタリングを用いて、一流選手と酷似した才能を持つ選手を抽出する手法を提案した。その結果抽出した選手の多くが翌年以降の成績に成長が見られ、結果として前年よりも成長した選手をより多く特定することに成功した。

**キーワード** 推薦システム、スポーツデータ活用、クラスタリング

## 1 はじめに

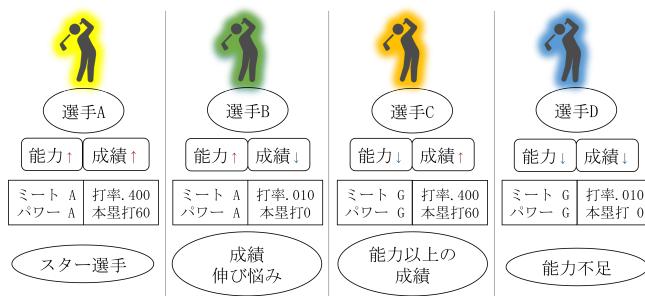


図1 タイプ別選手の定義

近年、インターネットの普及及び検索エンジンの性能向上に伴い、データ分析・解析への参入障壁は著しく下がった[1][2]。大規模データや開発環境・分析手順の解説ページが増加し、更にはスマートフォンの普及からアクセスも容易になったことでインターネットへアクセスする人口が増え[3]、質の高いサイトが生き残るようになっていった[4]。この波はスポーツ界隈にも押し寄せ、豊富なデータセットを便利なアプリケーションを通して分析することにより、より様々な角度からスポーツ技術を向上させる術ができ[6]、得られる情報も有益かつ実用性の高いものとなった[7]。特にアメリカのメジャーリーグでは、専門のデータ解析班を先駆けて取り入れたチームが、チームの劇的な再生に成功し大成功を収めた例[8][9]から、国内外のプロ・アマチームに選手やチームの管理・戦略や評価・概念の考案といったスポーツアナリズムが拡散されていき、今日のスポーツ界隈にはなくてはならないものとなった[10][11]。

本研究ではこの一側面である、選手の評価基準に焦点を置き、

それに特化した研究を行うことで、選手・組織並びに第三者にも納得のいく選手起用の一部形式化を目的とする。それに伴い本研究では選手を、現時点での成績と潜在的な能力の二つの視点に分けて評価する。これまでのスポーツ、引いては様々な事柄での評価は、成績による相対評価が絶対的で、加えて関係者の目分量による能力評価で将来性を計るケースが多く用いられてきた[12][13]。数値化が容易な成績に対し、数値化が難しい能力は、成績への相関関係も不明瞭なものが多い[14]。現にスポーツ界隈でも、成績は多数の指標が存在する一方で能力は、野球を例にとると球速・飛距離・足の速さ以外は目分量で評価されることが多い[15]。能力面の評価が形式化されないからは、様々な視点から納得のいく評価は得られず、私的な温情の入った評価にも繋がってしまう[16]。しかし近年スポーツ界隈では、メジャーリーグで2015年に導入されたstatcast<sup>1</sup>を例に、能力面の数値化が次々に行われ、投手の制球力や野手のミート力など、これまで言語を用いて行われてきた能力的評価も明確に数値化されるようになった[17]。

本研究ではstatcast導入以降両方が定量化されたことから、各選手の能力と成績の乖離に着目し、高い能力に成績が伴っていない選手、逆に高い成績に能力が伴っていない選手を抽出する方法を提案した。選手並びにチームの将来性を考慮した場合、現時点の成績も大切であるが、今後成長する可能性の高い選手を長期的に起用することも大切である[18]。本研究では、前者の高い能力に成績が伴っていない選手を、長期的に起用する対象とするべきとの仮説を立てた。この背景として、アマチュアの数年間は能力と成績の差分が埋まらず高い成績を維持出来るのであるが、プロの数十年間となると、成績が能力に収束していくことが多い[19]。加えてアマチュア内での高い人気も伴っ

1: 試合中の全動作を数値化するシステム

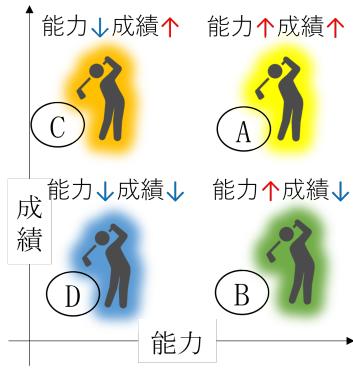


図 2 タイプ別選手の位置関係

てくることで、評価が成績以上に上がってしまい、更にその差が広がってしまう例も多々ある [20]。アマチュアで類を見ない活躍を遂げていた選手が、プロに入った途端に成績に伸び悩み引退を宣告される例が多いのは、そういう背景があるのではないかと考えられる。長期的な活躍・高い成績の維持を求められる世界で、真に問われるのは根本的な能力の部分ではないだろうか。

そういう仮説を踏まえ検証を進めるにあたり、本研究で定義する選手を大きく四つに分類する。A は能力・成績共に高い選手、B は能力が高く成績が低い選手、C は能力が低く成績が高い選手、D は能力・成績共に低い選手とする（図 1,2）。先程仮定した最も成長期待値の高い選手を B とし、どのような手順で B を抽出するかに着目し、両手法の解説を行う。最後に本研究の構成を述べる。第 2 章では関連研究で、これまで成長期待値の高い選手を抽出する際に取られていた既存手法について説明する。第 3 章では既存手法を踏まえ、より抽出性能の向上が期待できる提案手法の説明と既存手法との差異、利点について説明する。第 4 章で本研究に用いられるデータセットや特徴量について触れた後、第 5 章では実験結果を述べる。第 6 章ではデータセットを変更して実験した際の結果を述べ、提案手法並びに本研究の汎用性を述べ、第 7 章で全体のまとめと今後の課題を述べる。

## 2 既存研究

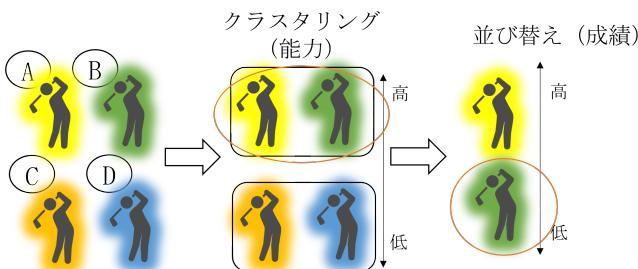


図 3 既存手法の流れ

Ishii 氏 [21] は、k-means 法と階層的クラスタリング [22] を用いて、野球における過小評価選手を特定した（図 3）。ここ

でいう過小評価選手とは、スター選手に酷似した能力を持ち合わせているにも関わらず、未だ成績に結び付いていない選手を指す。Ishii 氏は、データセットを投手に設定し、能力的特徴量（変化量、回転数、リリース位置等）を基準に、投手をクラスタリングした。その際、確立された投手（スター選手）と同じグループに所属する、最も ERA（防御率）が低い投手を、過小評価投手と定義した。抽出した過小評価選手の、翌年以降の成績推移（才能の開花）には、WAR を用いた。

本手法では、データセットに野手を用いるため、Ishii 氏の手法を改変した上で、導出された性能を提案手法と比較する。その改変した内容を以下に示す。まず、選手の能力的特徴量を基準にクラスタリングを行い、能力の傾向が似ている選手同士をグループに分類する。ここまででは Ishii 氏の手法と同様である。次に本研究では、所属する選手の能力値を平均し、各グループの能力平均値を算出する。そしてその能力平均値を基に、グループの並び替えを行い、全体の能力レベルが高いグループは上位、全体の能力レベルが低いグループは下位へと並び替える。

この作業の目的としては、確立された選手の選定基準をより形式化せざることがある。Ishii 氏の研究論文には、確立された投手の選定基準が明記されていない。ここで Ishii 氏と著者の選定基準に差異が生じることで、誤った選手を選ぶ確率を低くするために、スター選手の選定基準を出来るだけ一般化させた。そのため、この並び替えによって上位になった数グループ（閾値）を、次のステップに進む該当グループ（A・B）とした。そして抽出されたグループは、所属する選手の成績合計値を基準に、選手の並び替えを行い、成績合計値が高い選手は上位、個人成績が低い選手は下位へと並び替えた。また同時に、所属する選手の成績合計値を平均し、各グループの成績合計値の平均を算出した。その成績合計値の平均を基準に、下回る選手を各グループから抽出し、それらの選手を成長期待値の高い選手（B）と予測した。この作業の目的としては、まず野手の場合、ERA のような投手の成績を一概に表す指標が存在せず、よりニュアンスを近づける目的で行った。理由として、投手はいろいろなタイプはあるが、基本的に投手の質は防御率に現れる。しかし野手の場合、それに対応する指標は存在しない。足の速い選手は、長打率は低いが打率は高い傾向にあり、一方筋力のある選手は、打率は低いが長打率は高いといった傾向にある。野手を抑えるという一つの目的がある投手に対して、野手は様々な成績の残し方が存在する。少しでも様々なタイプの選手の成績を比較出来ないかと考えた時に、様々なタイプの選手を平等に審査出来るように選別した特徴量の合計値を、成績の基準値として用いた。また、Ishii 氏はグループで最も ERA の低い投手のみを抽出しているのに対し、本研究ではもっと広い意味での過小評価選手を、人数も多く特定する事を目的としているため、成績合計値が閾値以下の選手を抽出するという作業を行った。

まとめると既存手法を改変した形として、能力が高いグループから、それぞれ閾値より成績が低い選手を抽出し、成長期待値の高い選手と仮定した。著者はこの手法を応用させることで、もっと広い意味での過小評価選手を特定出来ると考えた。次項

でその詳細について説明する。

### 3 提案手法

#### 3.1 提案手法の概要

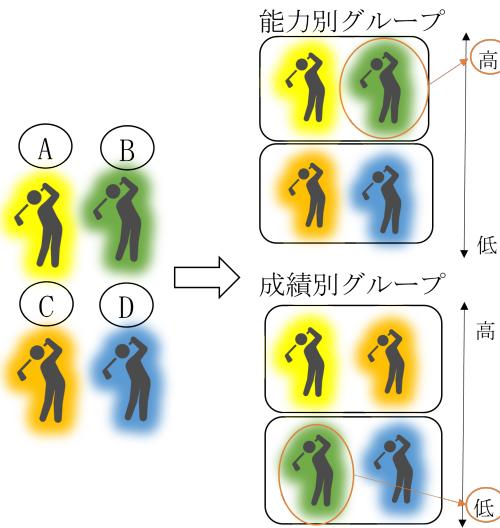


図 4 提案手法の流れ

提案手法は、全選手を選手の能力に応じてクラスタリングをすることは、既存手法と一緒にである。既存手法と異なる点は、第1段階で並行して、全選手を対象に選手の成績に応じたクラスタリングも行う(図4)。つまり提案手法では、第1段階で能力と成績を、各々クラスタリングを行って評価する。クラスタリングの手法としては、x-means法と階層的クラスタリングを用いる。クラスタリング後は同様に、第2段階で各グループに所属する選手の成績の平均値の高い順にグループを並び替える。こうして能力を基準に作られたグループ群と、成績を基準に作られたグループ群が出来上がる。この両グループ群にて、自分が所属する各グループ順位を比較し、成績基準のグループ順位と、能力基準のグループ順位の差分が閾値以上高い選手を抽出し、成長期待値の高い選手とする。既存手法では、能力が高い選手の中で、比較的成績が低い選手を抽出するのに対し、提案手法では、能力より成績が、閾値以上低い選手を抽出する方法を提案した。

#### 3.2 x-means 法

x-means[27]はk-meansを拡張したアルゴリズムで、クラスタ数を自動推定するという特徴がある。以下にアルゴリズムの詳細[23]の一部を引用する。

(STEP0)：分類対象とする  $n$  個の  $p$  次元データを  $x_1, x_2, \dots, x_n$  とする。

(STEP1)：十分小さなクラスタ数の初期値  $k_0$  (特に指定がなければ2) を定める。

(STEP2)： $k = k_0$  として、k-means法を適用する。分割後のクラスタを  $C_1, C_2, \dots, C_{k_0}$  とする。

(STEP3)：この時点での総クラスタ数を  $g$  とする。クラスタ  $C_i$

に対して  $k = 2$  として、k-means法を適用する。分割した後のクラスタを  $C_i^1, C_i^2$  とする。

(STEP4)：分割する前の場合について考える。すべてのデータに対しての  $p$  変量混合正規分布を

$$f(x; \theta) = \sum_{j=1}^g \pi_j \frac{1}{(2\pi)^{p/2} |\Sigma|^{1/2}} \exp \left\{ -\frac{1}{2} (x - u_j) \Sigma^{-1} (x - u_j)^T \right\} \quad (1)$$

と仮定する。ただし、 $\sum_{i=1}^g \pi_i = 1$  とする。分散共分散行列は、 $\Sigma = \text{diag}(\sigma^2)$  と仮定する。分割する前の BIC を以下のように計算する。

$$BIC = -2 \log L(x; \hat{\theta}) + q \log n \quad (2)$$

ここで、 $\hat{\theta} = [\hat{\pi}, \hat{u}, \hat{\Sigma}]$  は、 $p$  変量混合正規分布の推定量であり、 $q$  はパラメータ空間の次元数で、 $q = g - 1 + gp + 1 = g(p + 1)$  である。 $x_i$  はクラスタ  $C_i$  に含まれる  $p$  次元データとする。 $L$  は尤度関数で  $L(\cdot) = \prod_{i=1}^n f(\cdot)$  である。

(STEP5)： $C_{i1}, C_{i2}$  と 2 分割に分けた場合について考える。すべてのデータに対しての  $p$  変量混合正規分布を

$$f(x; \theta) = \sum_{j=1}^{g+1} \pi_j \frac{1}{(2\pi)^{p/2} |\Sigma|^{1/2}} \exp \left\{ -\frac{1}{2} (x - u_j) \Sigma^{-1} (x - u_j)^T \right\} \quad (3)$$

と仮定する。ただし、 $\sum_{i=1}^{g+1} \pi_i = 1$  とする。分散共分散行列は、 $\Sigma = \text{diag}(\sigma^2)$  と仮定する。この 2 分割モデルにおける BIC を以下のように計算する。

$$BIC' = -2 \log L(x; \hat{\theta}) + q' \log n \quad (4)$$

パラメータ数は、 $q' = g + (g + 1)p + 1 = (g + 1)(p + 1)$  となる。 $L$  は尤度関数で  $L(\cdot) = \prod_{i=1}^n f(\cdot)$  である。

(STEP6)： $BIC > BIC'$  ならば、2 分割モデルが適切であると判断して、2 分割を継続すべく  $C_i \leftarrow C_i^1$  とする。 $C_i^2$  については、 $p$  次元データ、クラスタの重心、対数尤度とモデル評価基準を保持して、STEP(3)へ戻る。

(STEP7)： $BIC < BIC'$  ならば、2 分割しないモデルの方がより適切であると判断され、 $C_i$  についての 2 分割を停止する。

(STEP8)： $C_i$  における 2 分割が全て終了。STEP(4)～(7)で作成された 2 分割クラスタが  $C_i$  内で一意になるようにデータに属するクラスタ番号を振り直す。

(STEP9)： $i = 1, 2, \dots, k_0$  に STEP(4)～(8)を繰り返す。

(STEP10)：最初に  $k_0$  個に分割したクラスタ全てにおいて 2 分割が全て終了。全てのデータに対して一意になるようにデータに属するクラスタ番号を振り直す。

#### 3.3 提案手法の優位性

##### 3.3.1 異なるタイプの選手を平等に評価

本研究の優位性を考える上で注目したいのが、タイプの異なる選手への評価である(図5)。スポーツには、ベクトルの異

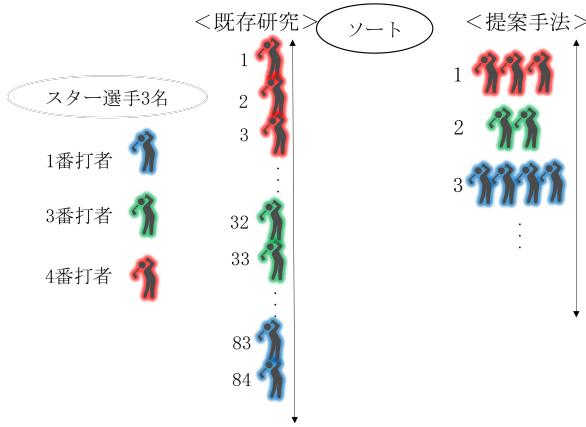


図 5 異なるタイプの選手への平等性の詳細

なる一流選手が多数存在する [24]. 野球を例に取ると、出塁率が高く俊敏性に富んだ 1 番打者タイプの選手もいれば、本塁打が多く腕力に富んだ 4 番打者タイプの選手も存在する. この他にも様々なタイプの一流選手が存在し、これらの選手同士に優劣をつけることは不可能である [25]. しかしそういった選手同士が横一列に並び、能力値によって順位付けをされた場合、能力値に使用されている特徴量の種類や数の偏りによって平等性が失われ、誤った優劣がついてしまう.

一方提案手法では、グループ毎に順位付けをする. 選手 1 人に明確な順位付けをしないため、両選手には、既存手法で発生したような順位の差分はつかず、両者は共に上位グループに所属することができるために、両選手の間に優劣関係が生じることはない. 既存手法のように、能力はグループ、成績は選手 1 人と別の土俵で考えるのではなく、能力・成績共にグループという同じ土俵で比較することで、実験する側が定義する成長期待値の高い選手に、より近い選手を抽出できるという優位性がある.

### 3.3.2 外れ値から受ける影響の軽減

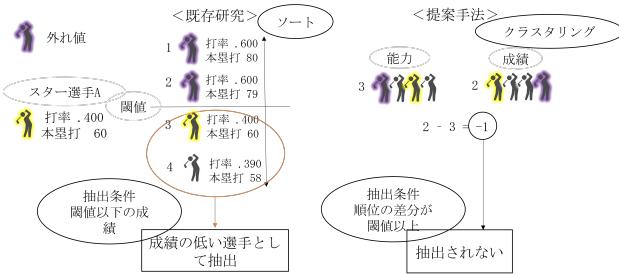


図 6 外れ値から受ける影響の詳細

データセットの全選手の中に、外れ値があった場合の、両手法における影響を比較する（図 6）. 例として、既存手法の第 2 段階までに抽出した上位グループの 1 つに、極めて成績の高

い選手が集中した場合を考える. この場合、その選手達の成績が、グループの平均成績を釣り上げてしまうため、他グループではそれほど成績が低くない選手が、グループの平均成績を上回れずに成長期待値の高い選手として抽出されることがある. 一方提案手法では、選手 1 人 1 人に順位付けをせずに、グループ毎に成績を比較するため、グループ毎の平均値に影響を受けず、全体を見た上で選手の位置付けが確認できる. そのため、パラメータをきつくすればするほど、より能力と成績の差が大きい選手を抽出できる.

## 4 実験

表 1 MLB データセットの詳細

データ	MLB
期間 (年)	2015 - 2022
人数 (人)	5,504
特徴量 (次元)	能力 54 / 成績 95
グループ (個)	能力 20 / 成績 20
評価指標	WAR

能 力		成 績	
Avg Ev	平均打球速度	PA	打席数
BA - xBA	運の良い安打率	H	安打数
SLG - xSLG	運の良い長打率	HR	本塁打数
Z-Swing & Miss %	ストライク球凡退率	BB %	四球率
O-Swing %	ボール球スイング率	AVG	打率
Sweet Spot %	8~32° 打球率	SLG	長打率
Flare/Burner %	60~100km & 0~30°	OPS	出塁率 + 長打率
Barrel %	157km & 26~30°	Off	打撃貢献率
5* %	捕球確率0~25% 捕球率	Def	守備貢献率
Sprint Speed	最大走行速度	UBR	走塁貢献率

図 7 使用する特徴量（一部）

### 4.1 データ

本研究で抽出する方法を提案した対象選手の概要を表 1 に示す. 本研究ではデータや特徴量の種類が豊富な観点から、MLB のデータを使用する. 期間は 2015-2022 年、statcast システム導入以降の計 8 年間、100 以上の規定打席を満たす 5,504 人（同一選手を含む）を対象とする. その際、使用した特徴量の一部を図 7 に示す. それぞれ標準化・正規化を行った能力的特徴量 ( $f_a^1, f_a^2, \dots, f_a^{54}$ )、成績的特徴量 ( $f_s^1, f_s^2, \dots, f_s^{95}$ ) を使用し、クラスタリングを行った結果、両者 20 のグループに自動的に分類された.

尚、本研究では能力と成績の 2 種類に分けて特徴量を収集しているが、両者に依存関係はない. 例えば図 8 にあるように、本研究で用いている特徴量の 1 つに、能力では塁間走のタイム、成績では盗塁数がある. 両者は足の速さという観点の元、採用された特徴量であるが、塁間走のタイムが速さと盗塁数の多さに直接的な繋がりはない. 相関性はあるものの、一方の特徴量を基にもう一方の特徴量の数値が決定されるといった形式にはなっていないため、両者に依存関係はないという前提のもと、特徴量を選別を行った.

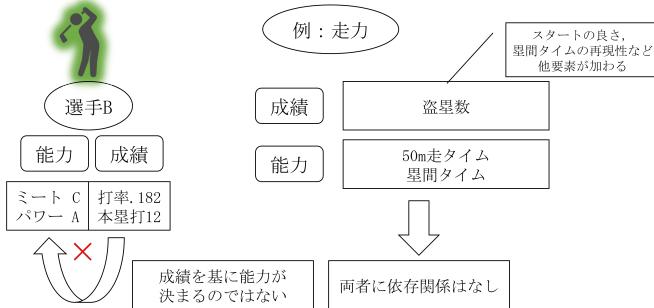


図 8 能力と成績の特徴量の依存関係について

## 4.2 規定打席

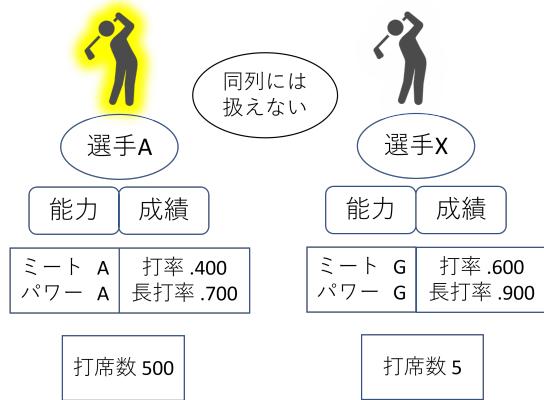


図 9 規定打席について

本実験では対象期間に出場した全選手から、規定打席 100 を設け、その基準をクリアした選手を対象とする。規定打席というのは 1 年を通して、打席に立った回数に制限を設け、達した選手を対象に含める基準のことで、プロリーグでタイトル受賞者（最優秀選手など）を選出する際に用いられている。この規定打席を設ける理由としては、率系指標の平等性確保の意図がある（図 9）。

例として、年間で立った打席の多い選手 A と少ない選手 X を比較する。能力を比較すると選手 A が圧倒的に優れているのが分かるが、成績を比較すると選手 X が圧倒的に優れている結果となっている。この原因は、成績の比較に割合系の指標が用いられていることである。割合系の指標は成功回数を全機会で割ることで求められる指標の為、成功回数が少なくて、そもそも機会も少なければ割合としては高く算出されやすい傾向にある。本塁打数のような回数系の指標であれば規定打席を設けることなく比較することが出来るのであるが、割合系の指標は規定打席を設けなければ、一度の成功や失敗が大きく数値を左右する結果となってしまう。これらの割合系の数値を外して対象選手を増やし、範囲を広げる案も考えたのであるが、現実的にプロ視点での評価には割合系の指標が用いられることが多いため、本実験では規定打席を設けて割合系の指標も用いつつ、選手を評価する方針である。

## 4.3 特徴量

3.3.1 で前述した通り、本実験では扱う特徴量の種類や数の偏りによって、選手のタイプに偏りが出来てしまう恐れがある。本実験ではそれらの偏りを極力減らすことに注力した。

### 4.3.1 能力と成績間のバランス

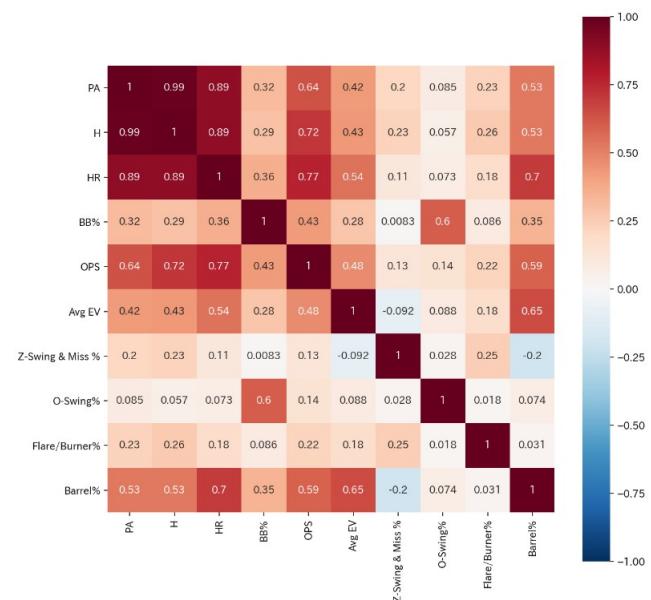


図 10 対生成の様子（一部）

（縦軸、横軸：能力・成績特徴量 5 つずつ）

本実験では特徴量を能力と成績に分類しているが、その両者の特徴量の配分に差があってはならない。例えば能力では俊足選手 A に有利な特徴量が多く、成績では腕力選手 B に有利な特徴量が多いとなった場合、俊足選手 A は能力があるが成績は不足していると判定されることになる。このようなタイプごとの平等性を保つためには、能力と成績に偏りがあってはならない。その問題を解消するために本実験では因子分析を行い、能力の 50 m タイムは成績の盗塁数、能力の平均飛距離は成績の本塁打数といった形で能力と成績の特徴量ごとに對を形成する（図 10）。ただしデータの関係上、現時点では能力的特徴量の数が不足しているため、1 つの能力的特徴量に関連する成績的特徴量を複数組び付けて對を形成する場合もある。對が作れない特徴量に関しては除外することで、両者の偏りを極力なくすることに注力している。

### 4.3.2 特徴量全体のバランス

能力と成績に偏りがなくなったとしても、全体的に偏りが生まれてしまってはいけない。例えば上記で対が 100 個形成されたとすれば、そのうち 80 個が腕力選手 B に有利な対であった場合、やはり結果的に平等な特徴量選択とは言えない。その為に本実験では、対の数も平等になるように調整する。方法としては、本実験の評価指標である WAR を目的変数とした重回帰分析を行う。

WAR はシーズン MVP（最優秀選手）を選出する際も非常に目安となる基準 [25] で、異なるタイプの選手同士を比較できる指標となっている。WAR の計算手順は各サイトによって違うだけでなく、計算に使われている指標も、複合指標同士を組み合わせるなど量が膨大で追いきれない為、重回帰分析により、各特徴量と WAR の関連性を調べ、先程形成された対の合計値が閾値よりも高い対を優先的に採用した。とはいって実際は、打者の WAR 自体が長距離打者に有利なように作られているとの声があったり、守備走塁が得意な選手よりも長距離打者の方がクビになりにくく現状もあったりする [26] ため、一般的な評価としても長距離打者の方が有利になるような現状にあるのは確かである。それを本研究にも反映させ、各タイプの野手に平等になるように特徴量を選別するものの、平等になりすぎないように気を付けながら、世間的な評価とのバランスを考慮して特徴量を選択した。

#### 4.4 評価方法

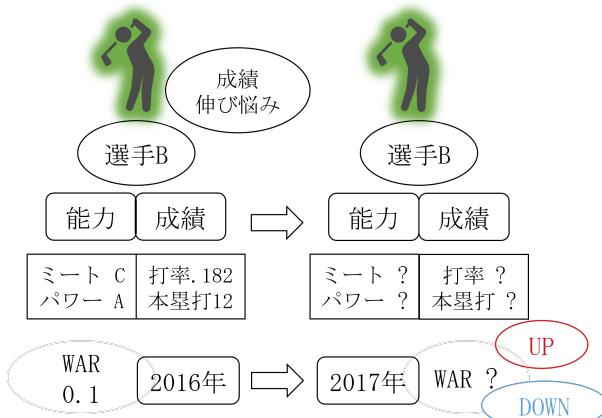


図 11 抽出後の選手の評価方法（1年）

選手成長率を図る目安として、評価には WAR を使用した。本研究は、成績は能力に収束するという仮説の元行っているわけであるが、能力が成績に反映されるまでの期間も重要であろう。本研究は 2015 年から MLB に導入された statcast というトラッキングシステムによって算出された数値を能力的特徴量として扱っているため期間は限られているが、この計 8 年間の中で、最適な反映期間を探したいと考えている。後の実験結果では、1 年から 4 年の 4 つのパターンの結果を示している。

1 年の場合、例えば 2015 年に成長期待値の高い選手として抽出された場合、2016 年の WAR が 2015 年の WAR を閾値以上更新した場合真とし、前年と同値かそれ以下であった場合は偽とする。また 4 年の場合、例えば 2015-2018 年の能力と成績の平均値から成長期待値の高い選手として抽出された場合、2019-2022 年の WAR 平均が 2015-2018 年の WAR 平均を閾値以上更新した場合真とし、前年と同値かそれ以下であった場合は偽とする（図 11）。また冒頭でも触れたが、既存研究との違い（使用するデータセットや特徴量、クラスタリング手法）は、本研究で使用されるものを既存研究にも適応させた上で実験し直し、そこで算出された結果との比較を行った。

## 5 実験結果

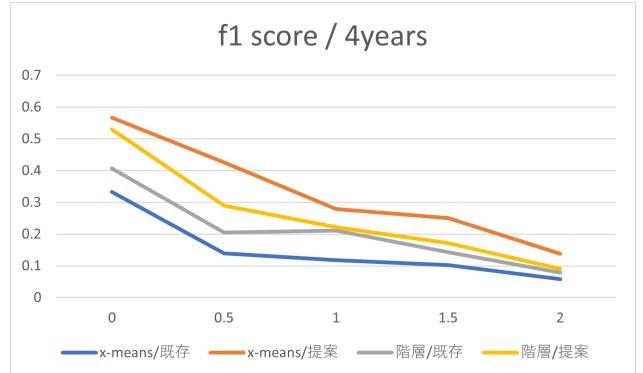


図 12 4 年後に WAR が閾値以上伸びる選手の特定性能  
(縦軸 : f1 値, 横軸 : WAR の差分の閾値)

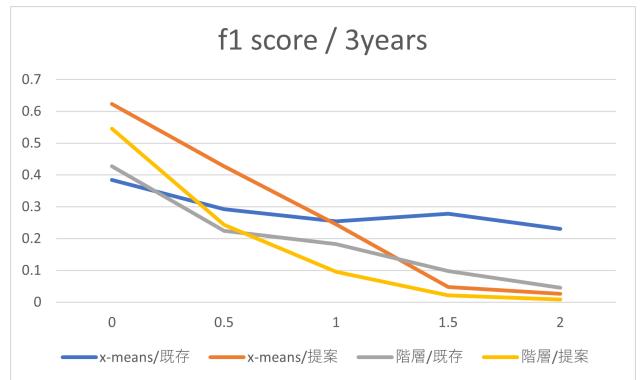


図 13 3 年後に WAR が閾値以上伸びる選手の特定性能  
(縦軸 : f1 値, 横軸 : WAR の差分の閾値)

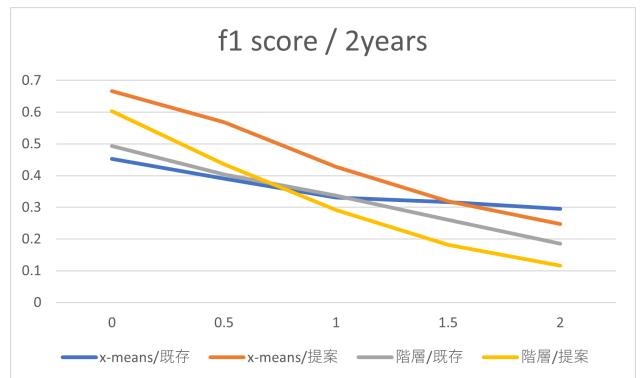


図 14 2 年後に WAR が閾値以上伸びる選手の特定性能  
(縦軸 : f1 値, 横軸 : WAR の差分の閾値)

図 12～図 15 からそれぞれ結果を比較すると、全体として提案手法が既存手法を上回る性能を出す事に成功した。特に x-means に関しては、基本的に高い性能を維持出来ている事が分かる。しかし、2 年や 3 年のグラフを見ると、WAR の差分の閾値が上がった時に、既存手法に逆転される事が分かったため、現時点での提案手法は、前年に比べて少しでも成長が見られ

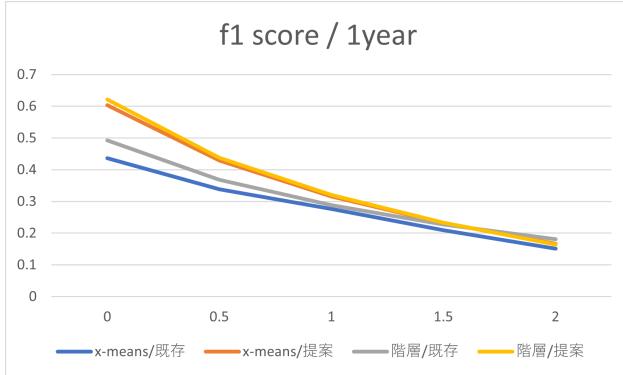


図 15 1 年後に WAR が閾値以上伸びる選手の特定性能  
(縦軸: f1 値, 横軸: WAR の差分の閾値)

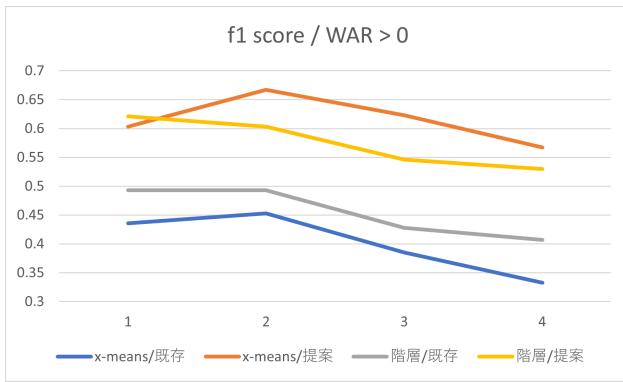


図 16 選手の WAR 向上に必要な年数  
(縦軸: f1 値, 横軸: 必要な年数)

る選手を特定するのに適していると言える。今後は WAR の差分の閾値が上がっても高い性能を維持する事に注力したいと考える。

また図 16 は、能力が成績に反映されるまでの適切な年数を反映したグラフである。例として前年よりも WAR が少しでも向上している場合の各手法の性能を比較すると、どの手法も 1 年か 2 年で能力が成績に反映されている選手が多い事が分かる。ここから分かる事として、能力が成績に反映されるまでの期間は比較的早い傾向にあり、3 年以上成績が出ていない選手は、別の部分に課題がある可能性が高いことが考えられる。具体的には、精神的な部分やコンディションの調整不足等が考えられる。3 年以上成績が低迷している選手は、それらを強化する事が成績低迷の根本的解決であると考えられる。

## 6 汎用性の考察

本研究では MLB の野手に標準を置いた上で実験を行ってきた。この提案手法の汎用性を確認する為に、異なるデータセットに手法を適応し、提案手法の優位性が確認できるかの汎用性を確かめる。用意したデータセットは Netflix (映画) で、スポーツ選手と異なり、同じ選手が能力の成長によって数年後に変化するといったことはないため、図 17 の能力的特徴量を基に、興行収入が閾値を超えるかどうかで各手法実験を行い、その結果を比較する。表 2 に示すデータセットは 4,915 本の特徴

量を能力・成績共に 6 次元ずつ取り入れ、クラスタ数は両者約 20 グループずつに分かれた。

表 2 Netflix データセットの詳細

データ	Netflix
映画 (本)	4,915
特徴量 (次元)	能力 6 / 成績 6
グループ (個)	能力 20 / 成績 19
評価指標	gross (興行収入)

能 力		成 績	
製作費	主演の評価	世間認知度	観客動員数
助演1の評価	助演2の評価	閲覧率	支持率
監督の評価	全キャスト評価	コメント数	総合評価

図 17 Netflix データセットの特徴量

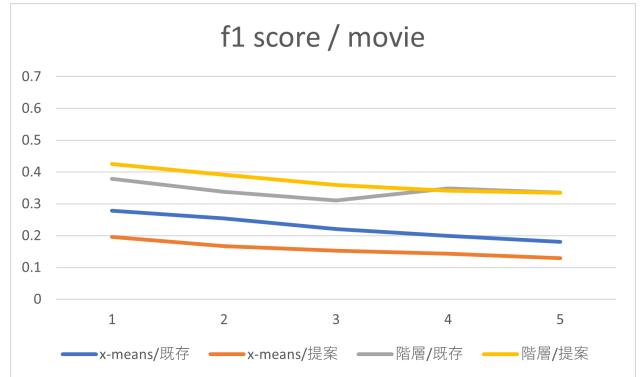


図 18 各手法の興行収入の予測性能  
(縦軸: f1 値, 横軸: 興行収入の閾値)

図 18 より、映画のデータセットに関しても、提案手法の階層的クラスタリングが最も高い性能を出すことに成功した。課題としては、スコアが先程のデータセットと比較して全体的に低い事があげられる。この主な原因としては、能力的特徴量として主に用いている出演陣・製作陣の評価が、映画の面白さや興味深さ・人気等に直結しにくい事が考えられる。昨今は SNS の普及で映画の本質が問われ、低予算で無名キャスト陣が出演する B 級映画でも、興行収入の向上が期待できる世の中になった。出演陣や製作陣がいくら豪華でも、流行しなければ興行収入に繋がらない事が、結果として顕著に表れたのではなかと考える。今後の課題として、映画の場合は数年後に能力が成績に反映されるといった時系列での成長がないため、そういった時系列で能力と成績の関係性を評価できるデータセットへの適応を考えられる。

## 7 ま と め

本研究では、高い能力を持っているにも関わらず、成績に反映されていない選手を、成長期待値の高い選手として定義し、特定と評価を行った。能力が成績に反映されるまでの期間も考慮するため、複数のパターンで実験を行い、全体として提案手法が既存手法の性能を上回ることに成功した。また汎用性の確認として、異なるトピックのデータセットにも各手法を適用させ、こちらも提案手法が既存手法を上回ることに成功した。今後の課題としてはデータセットの拡大や、更に異なるデータセットへの適応、異なるスコアでの手法の評価など、更に予測性能の向上と汎用性の拡大に注力し、実用性の高い知見を得られるように努めたい。

## 文 献

- [1] Kudyba S P, Hoptroff R G. "bigger dilemmas : A critical review." *Journal of Business Research*, 67(9), 2149-2157. 2014.
- [2] Mohammad Saeid Mahdavinejad, Mohammadreza Rezvan, ... , Amit P Sheth. "Machine learning for internet of things data analysis : a survey." *Digital Communications and Network* 2017, 4(3), 161-175. 2018.
- [3] Choi S, Lee C, Park J. "Mobile Internet Access and Use Behavior: Comparing Mobile-only Users with PC Users in Korea." *Telematics and Informatics*, 33(1), 298-308. 2016.
- [4] Wu S, Chen Y. "How smartphone users' motivation and impulsivity influence mobile shopping website usability." *Journal of Retailing and Consumer Services*, 34, 262-270. 2017.
- [5] J C Bradbury. "Valuing Batting and Fielding : How Much Do Hitters and Pitchers Contribute to Team Wins?" *Journal of Sports Economics*, 7(2), 1-22. 2002.
- [6] Talavera J, Martin-Rodriguez J. "Big data applications in sport : A review." *International Journal of Performance Analysis in Sport*, 19(6), 1065-1080. 2019.
- [7] Sam Robertson, Chris Joyce. "Data Science in Sport: A Review." *International Journal of Performance Analysis in Sport*, 18(6), 867-888. 2018.
- [8] Adler P B, Kwon S. "Sign-stealing, whistleblower retaliation, and organizational misconduct: A case study of the Houston Astros." *Journal of Business Ethics*, 169(1), 1-16. 2019.
- [9] L Frisbee, B Gettys, A Mishra, and S, Kunapuli. "The Houston Astros: A Big Data Revolution in Sports," *Big Data*, 6(4), 332-339. 2018.
- [10] David P Hedlund, Kostya P Shy Manovich. "Sports Analytics : A New Data Frontier." *Journal of Big Data*, 6(1), 1-24. 2019.
- [11] Baumer B, Zimbalist A. "The Sabermetric Revolution: Assessing the Growth of Analytics in Baseball." *Journal of Sports Economics*, 18(7), 682-698. 2017.
- [12] Landy F J. "Alternative models of performance appraisal." *Academy of Management Review*, 7(2), 219-231. 1982.
- [13] Dohmen T J, Sauermann J. "Referee bias in professional soccer : Evidence from Colombia." *Journal of Sports Economics*, 17(7), 699-717. 2016.
- [14] Sackett P R, Lievens F. "Personnel selection." *Annual Review of Psychology*, 69, 1-26. 2018.
- [15] Miklasz B. "Eyes of the beholder : Baseball's subjective statistics." *Baseball Research Journal*, 40(1), 90-97. 2011.
- [16] McGarry A, Franks I M. "The effect of expertise on the quality of sport performance assessment." *Journal of sports sciences*, 12(2), 177-186. 1994.
- [17] Rousseau J, Sack D, Welch E. "Using Statcast to analyze player talent evaluation and development in Major League Baseball." *Journal of Sport Management*, 31(4), 384-397. 2017.
- [18] Baim D, Redner S. "Hitting the right spots: Baseball, beer, and the Boston Red Sox." *Journal of Statistical Physics*, 120(5-6), 1197-1208. 2005.
- [19] Burke K L, Joyce C R, Drummond J J. "The impact of aging and experience on the performance of Australian rules footballers." *Journal of aging and physical activity*, 6(4), 358-375. 1998.
- [20] Barker J B, Mellalieu S D. "Behavioural biases of amateur sporting talent selectors: An exploration." *Journal of sports sciences*, 30(14), 1433-1441. 2012.
- [21] Tatuya Ishii. "Using Machine Learning Algorithms To Identify Undervalued Baseball Players." CS229. 2016.
- [22] Jain A K, Murty M N, Flynn P J. "Data clustering: a review." *ACM Computing Surveys*, 31(3), 264-323. 1999.
- [23] Ishioka T. "Extension of the k-means algorithm for automatically determining the number of clusters." *Applied Statistics*, 29, 141-149. 2000.
- [24] Bartlett R M, O'Connor D. "Limits to the validity of the science of performance analysis." *International Journal of Performance Analysis in Sport*, 10(1), 1-7. 2010.
- [25] Scheerder J, Vos S, ... ,D'Hondt E. "Recognizing the complexity and diversity of sports abilities: The need for a holistic and multi-disciplinary approach in talent identification and development in sport." *Journal of Functional Morphology and Kinesiology*, 3(4), 55. 2018.
- [26] Cain J, DeFusco R A. "Scouting and the birth of stars: Evidence from the baseball draft." *Journal of sports economics*, 9(6), 592-605. 2008.
- [27] Pelleg D, Moore A. "X-means: Extending K-means with efficient estimation of the number of clusters." *Journal of Machine Learning Research*, 1, 1-15. 2000.