日本オペレーションズ・リサーチ学会
2022年 春季研究発表会

日本プロ野球における混合分布モデルを用いた野手の分類

05001481　順天堂大学　織田大志
01506960　順天堂大学　廣津信義

1. はじめに

野球は、幅広い国々でアマチュアからプロまで存在するメジャーなスポーツである。日本においても、NPB（日本野球機構）傘下の12球団が公式戦を行い、毎年優勝を争っている。

その公式戦で選手が残した記録（長打率、出塁率、OPSなど）をデータとして統計学的見地から分析する手法のことをセイバーメトリクスと呼ぶ。セイバーメトリクスでは、代表的な統計手法の一つ、多変量解析も行われている。

多変量解析において最も多いのが、k-means法によるクラスター分析である。k-means法は、非階層型クラスタリングの一つで、クラスターの平均を用い、対象のデータを主観的に決定したクラスター数k個に分類する。

一方で、クラスター数を客観的に決定するクラスタリング手法も存在する。混合分布モデル（Gaussian Mixture Model）によるクラスター分析では、適切なクラスター数、分散共分散行列の型をBIC値によって決定することができる。

しかしながら、クラスター分析を行う際、分類に使用する変数が複数の場合、一つの問題が生じる。それは、分類結果の解釈が困難になるということである。この問題を解決できるのが主成分分析である。主成分分析とは、相関のある多数の変数から、相関の無い少数で、全体のばらつきを最もよく表す主成分と呼ばれる変数を合成する手法である。元の変数の情報から縮約された主成分の情報を新しく解釈することができる。

本研究では、日本プロ野球選手が公式戦で記録した指標を使って、野球選手をグループ化する一つの枠組みを提示する。ここでいう「枠組み」とは、主成分分析・クラスター分析を単体で実施するのではなく、主成分分析の結果を利用してクラスター分析を行うという「流れ」を指す。

この「流れ」を提示した先行研究には、César Soto-Valero(2017)がある。彼は、FIFAの公式ウェブサイトから入手できる7705人のヨーロッパのサッカー選手の指標に主成分分析を施し、情報を2つの変数に縮小した。次に、その変数を使用して混合分布モデルによるクラスター分析を行い、類似した選手からなる4つのクラスターを生成した。主成分分析とクラスター分析を用いて、サッカー選手をグループ化するための枠組みを示すことが目的であった。本研究はそれを受けて、サッカーではなく日本プロ野球の選手に同様な分析を実施することが目的である。

2. 手法

2.1 主成分分析

相関のある多数の変数から、相関の無い少数で、全体のばらつきを最もよく表す主成分と呼ばれる変数を合成する手法である。今回は主成分分析の結果の解釈を容易にするために、使用する 115 個の打撃指標を株式会社 DELTA が定義した 8 つの項目毎に分析する。

2.2 混合分布モデルによるクラスター分析

任意の混合分布に含まれる$\boldsymbol{G}$個の正規分布の確率密度関数を$\boldsymbol{f_1(x;\theta_1)},\dots,\boldsymbol{f_G(x;\theta_G)}$，これらの混合比を$\boldsymbol{\pi_1},\dots,\boldsymbol{\pi_G}$とする。$\boldsymbol{\theta_g(g=1,\dots G)}$は確率（密度）関数$\boldsymbol{f_g(x;\theta_g)}$に含まれるパラメータからなるベクトルである。また、混合比$\boldsymbol{\pi_1},\dots,\boldsymbol{\pi_G}$については$\boldsymbol{0\le\pi_g\le 1(g=1,\dots G)}, \boldsymbol{\sum_{g=1}^{G}\pi_g=1}$を満たすものとする。このとき、混合分布モデルの確率（密度）関数は次で与えられる。

$$f(x;\theta)=\sum_{g=1}^{G}\pi_g f_g(x;\theta_g)$$

このモデルに含まれるパラメータ$\theta=(\theta_1^T,\dots,\theta_G^T,\pi_1,\dots,\pi_{G-1})^T$を推定するには EM アルゴリズムを用いた。

この EM アルゴリズムの E ステップで用いられる条件付き期待値

$$\gamma_{ig}=E(Z_{ig}|x_i)=Pr(Z_{ig}=1|x_i)$$
$$Pr(Z_{ig}=1|x_i)=\frac{\pi_g f_g(x_i;\theta_g)}{\sum_{h=1}^{G}\pi_h f_h(x_i;\theta_h)}$$

の推定値が最大となる成分へ第i観測値を分類すれば、混合分布を用いてクラスター分析を実施することができる。

2.3 決定木分析

決定木分析とは、母体となる全てのデータを段階的に分割し、決定木と呼ばれるツリー状の分析結果を出力する方法である。クラスター分析の結果を目的変数として、説明変数が分類結果にどの程度影響するかをランダムフォレストにより調べた。

3. データ

プロ野球に関するデータ分析を扱う株式会社 DELTA が提供するサービス”1.02 Essence of Baseball”から入手できるデータを用いた。具体的には、2020 年の公式戦に出場した野手 327 人の 115 種類の打撃指標を分析した。打撃指標はその意味合いによって詳細な項目 8 つに分かれている。以下にその項目と、各項目の代表的な指標の平均と標準偏差を示す。

表 1 各項目の代表的な指標の平均と標準偏差

項目	指標	平均	標準偏差
Standard	PA	161.199	162.275
	RBI	17.138	22.022
Advanced	K%	23.245	12.304
	OPS	0.593	0.237
Batted Ball	Hard%	30.646	12.315

	IFH%	6.872	11.357
Win Probability	WPA+	2.783	3.262
	WPA-	-2.687	2.639
Pitch Type	SLv	128.669	2.784
	CB%	7.044	3.803
Pitch Value	wCB	0.289	1.388
	wCB/C	0.980	9.162
Plate Discipline	SwStr%	12.285	6.448
	Contact%	74.998	10.766
Value	Replacement	5.050	5.095
	Batting	-0.075	9.094

4. 結果

主成分分析と混合分布モデルによるクラスター分析を用いて、特徴あるクラスターに分割するまでの「枠組み」を構築することができた。

8 つの項目毎に主成分分析を実施し、累積寄与率が約 7 割を目安に主成分を選択すると、表 2 のような主成分個数なった。

表 2 選択した主成分の数

Standard	Advanced	Batted Ball	Win Probability
3	3	4	2
Pitch Type	Pitch Value	Plate Discipline	Value
6	5	2	2

主成分得点を用いて、混合分布モデルによるクラスター分析を実施した。表 3 は、それぞれの項目で最も BIC が高かったクラスター数と分散共分散行列の型である。

表 3 クラスター数と分散共分散行列の型

Standard	Advanced	Batted Ball	Win Probability
3, EII	2, EII	2, VII	2, EII
Pitch Type	Pitch Value	Plate Discipline	Value
2, VII	2, VII	1	2, VII

元の指標の中で、分類結果に大きく影響を及ぼしている変数を調べるために、クラスター分析の結果を目的変数、元の変数を説明変数とし、ランダムフォレストを適用した。その結果が図 1 である。それぞれの項目の中で、重要度が高かった指標上位 2 つを示している。

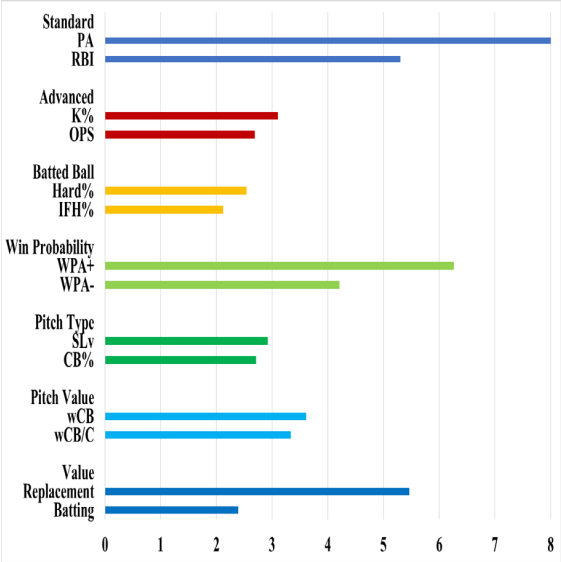

図 1 各指標の重要度（項目別）

所属クラスターがわかるように散布図を描くと、選手の分類が確認できる。図 2 は、Standard で重要度が高かった PA と RBI をそれぞれ x 軸、y 軸として作図したものである。

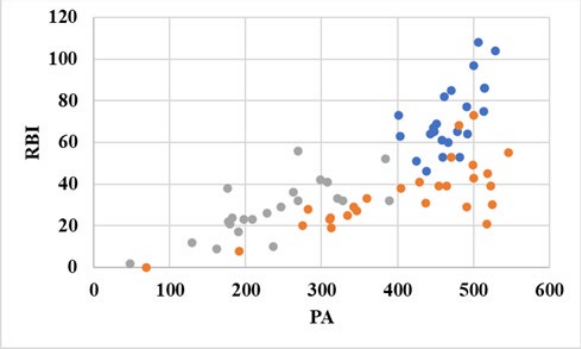

図 2 クラスター別散布図（Standard）

5. おわりに

主成分分析と混合分布モデルによるクラスター分析を使用して、野球選手をグループ化する一つの枠組みを提示することができた。今後は野手だけでなく、投手を分類する枠組みを提示していきたい。また、守備評価に関する指標が不足しているので、それらを加えて分析していきたい。

参考文献

[1] 大野高裕. 多変量解析入門. 同友館, 2000, p. 81-108, p. 214-222.

[2] 金明哲. R によるデータサイエンス. 森北出版, 2014, p. 66-77, p. 121-126, p. 271-275.

[3] César S-V. A Gaussian mixture clustering model for characterizing football players using the EA Sports' FIFA video game system. RICYDE. Revista Internacional de Ciencias del Deporte. 2017, vol. 13, no. 49, p. 244-259.