Classification of fielders in Nippon Professional Baseball using a Gaussian mixture clustering model

申請中 Juntendo University ODA Taishi
01506960 Juntendo University HIROTSU Nobuyoshi

1. Introduction

The purpose of this study is to present a framework for grouping players in Nippon Professional Baseball (NPB). In order to group them, we could apply a method proposed by Soto-Valero (2017). He utilized principal component analysis (PCA) on 7,705 European footballers using data from FIFA's official website, and shrank their information such as Dribbling, Handling and Kicking into two variables. Then, he conducted cluster analysis using a Gaussian mixture model (GMM). After classifying them into four clusters consisting of players who have similar characteristics, excellent players were identified among them.

In our study, we use a similar method. We apply PCA to 129 baseball players and shrink their performances into a small number of variables. Then, we conduct cluster analysis by GMM using 127 indicators such as Slugging percentage, On-base percentage and On-base plus slugging in the 2020 season taken from service "1.02 ESSENCE OF BASEBALL" provided by DELTA Co., Ltd. In each group, we evaluate the players according to their principal component scores corresponding to their characteristics and discuss the usefulness of this method.

2. Principal components analysis

One of the most difficult problems in performing multivariate analysis is reduction and selection of variables used in the analysis. This is a method of synthesizing a small number of uncorrelated variables called principal components that best represent the overall variation from many correlated variables.

Assuming that $X_{n\times p}$ is a dataset consisting of n individuals and p variables, the composite variable is the following linear combination equation that contracts p-dimensional data to a lower k-dimensional (k ≤ p).

$$z_j = a_{1,j}x_1 + a_{2,j}x_2 + a_{3,j}x_3 + \cdots + a_{p,j}x_p,\quad (j = 1,\cdots,k)$$

The coefficient $a_{i,j}(i = 1,\cdots,p)$ at this time is called the main component. In PCA, this principal component is obtained under the constraint of $\sum_{i=1}^{p} a_{i,j} = 1$ so that the variance of z_j is maximized.

3. Gaussian mixture clustering model

Mixture clustering is a method of finding the parameters of the original probability distribution, assuming that the continuous variables at hand are generated from several different probability distributions.

Now, let G probability density functions be $\boldsymbol{f_1(x;\theta_1),\dots,f_G(x;\theta_G)}$, and their mixed ratios be $\boldsymbol{\pi_1,\dots,\pi_G}$. However, $\boldsymbol{\theta_g(g=1,\dots G)}$ is a vector consisting of parameters included in the probability (density) function $\boldsymbol{f_g(x;\theta_g)}$. For the mixing ratios $\boldsymbol{\pi_1,\dots,\pi_G},\ \boldsymbol{0 \le \pi_g \le 1(g=1,\dots,G), \sum_{g=1}^{G}\pi_g = 1}$ shall be satisfied. At this time, the probability (density) function of the mixture distribution model is given as follows.

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

The EM algorithm is used to estimate the parameters $\theta = (\theta_1^T,\dots,\theta_G^T,\pi_1,\dots,\pi_{G-1})^T$ included in this model. Cluster analysis can also be performed using a mixture distribution model. Conditional expectation used in the E step of the EM algorithm (see reference [2] P178,179) for which data each observation belongs to

$$\gamma_{ig} = E(Z_{ig}|x_i) = Pr(Z_{ig} = 1|x_i)$$

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

The i-th observed value is classified into the component that maximizes the estimated value of these formulas.

4. Decision tree

Decision tree analysis is a method of dividing data in stages and outputting tree-like analysis results. It is an algorithm that sets a branch according to the condition, traces from the root, and divides into the one that best meets the condition. When the result of cluster analysis is used as the objective variable, it is possible to know how much the explanatory variables such as slugging percentage and on-base percentage affect the classification result.

5. Data

We used the data of 127 results recorded by fielders who come up to bat 157 or more times among the players who participated in the official games in 2020. The number of players who come up to 157 or more times is a maximum value that does not exceed the number of indicators. When performing PCA, the number of observations (number of players) must be larger than the number of original variables.

As shown in Table1, the results of the decision tree analysis are described with the highest features, along with the mean and standard deviation. For missing values, “0” is substituted for all values this time.

Table 1: Examples of typical indicators

	Mean	SD		Mean	SD
wRC	41.29	23.61	Replacement	9.39	3.46
OPS	0.73	0.11	FAv	144.97	0.85
Batting	4.62	13.66	CB%	7.41	2.39
Offense	9.97	11.11	RngR	3.6	3.58
WPA/LI	1.18	1.26	BB%	9.17	3.38
wRAA	9.73	10.69	CBv	117.92	1.71
SLv	128.9	1.3	Soft%	23.19	5.59
1000	7.4	6.55	O-Swing%	28.53	5.65
1200	8.88	7.86	BABIP	0.3	0.04
Fielding	4.64	4.53	FT%	6.71	1.71
UZR	4.64	4.53	Zone%	44.63	2.97
CTv	137.79	1.17	BB	32.3	19.7
+WPA	6.19	2.9	K%	18.64	4.93
BB/K	0.53	0.29	OBP	0.33	0.05
PA	334.93	119.73	SL%	16.1	3.15
RBI	37.64	22.42	Defense	7.86	5.96

6. Results

Looking at Table 2, the contribution ratio up to the 10th principal component is about 60%. Since the previous study Soto-Valero uses up to the second principal component which has the contribution ratio of about 60%, we use up to the 10th principal component in this study.

Table2: Results of PCA

	Comp1	**Comp2**	**Comp3**	**Comp4**	**Comp5**
Standard deviation	5.198	3.791	2.773	2.428	2.355
Proportion of Variance	0.213	0.113	0.061	0.046	0.044
Cumulative Proportion	0.213	0.326	0.386	0.433	0.477
	Comp6	**Comp7**	**Comp8**	**Comp9**	**Comp10**
Standard deviation	2.125	1.991	1.779	1.736	1.685
Proportion of Variance	0.036	0.031	0.025	0.024	0.022
Cumulative Proportion	0.512	0.543	0.568	0.592	0.614

Table 3: Means of typical indicators in each cluster

Cluster	**Number of players**	**wRC**	**OPS**
1	37	34.8	0.731
2	35	41.3	0.713
3	13	89.9	0.951
4	30	40.5	0.718
5	13	13.7	0.593
6	1	28.7	0.625

The results of applying a mixture model using these principal component scores are as follows. The advantage of using gaussian mixture model is that, unlike other methods, the number of clusters can be determined using BIC. In this study, 6 clusters are selected.

Then, we apply to the classification results as the objective variable and the other variables as the explanatory variables. Figure 1 shows the variable importance. We found wRC and OPS influenced the result of cluster analysis.

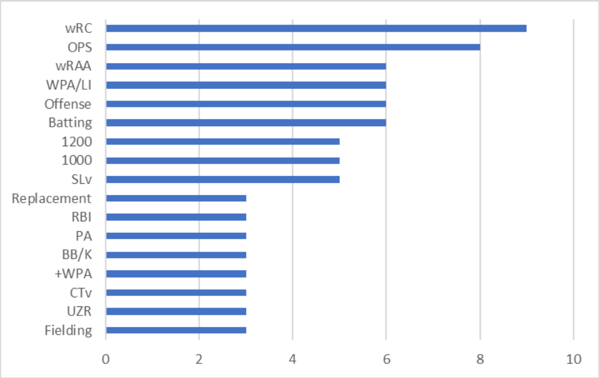

Figure 1: Variable Importance

7. Conclusions and further study

We were able to present a framework for grouping players using PCA, Gaussian mixture model, and decision tree analysis, and to clarify the differences in the types of players in NPB. This time we used the indicators recorded by the fielders, but we would like to analyze the indicators recorded by the pitchers as well. “0” is substituted to make up for the missing values in the dataset this time, but we would like to try other methods such as the average substitution method and the listwise method. In calculating the features, we used decision tree, the simple model, but next time we would like to use the gradient boosting decision tree as Soto-Valero.

References

[1] DELTA Co., Ltd. “Online Baseball Analysis Course Baseball Analytics Now”, https://peatix.com/event/1561186/view (viewed on August 5, 2020).

[2] O Matsui, K Koizumi, “Statistical model and guess”, Koudansha (2020). (in Japanese)

[3] S Miyamoto “Introduction to Cluster Analysis-Theory and Application of Fuzzy Clustering”, Morikita Publishing (2006). (in Japanese)

[4] K Nishiuchi “With Yasuhito Endo, the team will have 117% of points”, SoftBank Shinsho (2012). (in Japanese)

[5] C Soto-Valero “A Gaussian mixture clustering model for characterizing football players using the EA Sports’ FIFA video game system”, RICYDE. Revista International de Ciencias del Deporte (2017).