산업융합연구 제19권 제4호 pp. 17-21, 2021
ISSN 2635-8875 / e-ISSN 2672-0124
DOI : https://doi.org/10.22678/JIC.2021.19.4.017

데이터 분석을 활용한 신인급 선수 육성 방안 연구

유강수
전주대학교 교양학부 부교수

A Study on How to Nurture New Players using Data Analysis

Kangsoo You
Associate Professor, School of Liberal Arts, Jeonju University

요 약 최근 스포츠 현장에서는 데이터를 활용하여 경기를 진행하고 시즌을 구상하며, 팀을 운영하는 시도가 많아지고 있다. 선수 육성을 하기 위해서 데이터를 활용하고 정확한 분석이 필요하다. 이에 본 연구에서는 데이터를 수집하고 전처리하여 선수들의 데이터 분석과 시각화를 통하여 신인급 선수들의 경기력을 분석하였다. 또한 신인급 선수를 육성하려면 최소한 어느 정도의 기회가 부여되어야 하는지 분석하였다. 아울러 스포츠 현장에서 데이터를 활용하여 선수 육성을 하기 위한 데이터 분석 방안을 제시하였다. 본 연구는 데이터를 활용하여 신인급 선수를 육성하는 데에 이바지할 것으로 기대된다.

키워드 : 데이터 분석, 스포츠 데이터, 한국프로야구, 선수 육성, 장타율

Abstract Recently, in the field of sports, the use of data in conducting games, planning seasons, and operating teams has increased significantly. Also, in order to develop better players, it has become necessary to use data to accurately analyze their performance. Therefore, in this study, various data about rookie players was collected and pre-processed in order to analyze and visualize their performance. Additionally, an analysis was conducted to determine at least how many opportunities should be given to foster rookie players. Then, a data analysis method was presented for nurturing athletes by using data in the field of sports. It is expected that this study will contribute to fostering rookie players by utilizing data.

Key Words : Big Data Analysis, Sports Data, KBO League, Player Fosterage, Slugging Percentage

1. 서론

최근 데이터를 통한 의사결정은 일반적인 일처리 방법이 되었다. 스포츠 분야에서도 데이터를 활용하여 경기를 진행하고 시즌을 구상하며, 팀을 운영하려는 시도가 많아지고 있다. 특히 농구의 APBR메트릭스(APBRmetrics)나 야구의 세이버메트릭스(Sabermetrics)와 같이 기존의 클래식 기록을 넘어서 과학적인 통계방법을 활용해 2차 데이터를 분석하는 것이 대세가 되고 있다. 각종 스포츠 데이터는 스포츠 언론에 게시되고 있으며 일반 팬들도 전문 사이트를 통하여 자주 열람하는 데이터가 되었다. 또한 영상처리 기술의 발달로 인하여 각종 동영상과 정지영상을 분석하여 선수들의 여러 동작과 공의 물리적인 움직임을 분석하기까지 이르렀다. 이에 따라 스포츠 분야에서 수많은 데이터가 다양한 형태로 거대하게 발생하게 되었고 다양한 수요가 창출되고 있다[1-3].

특히 야구는 데이터가 많이 발생하는 종목이다. 경기시간이 3시간 내외로 다른 종목에 비하여 긴 편에 속하고, 프로스포츠 중 경기 수가 가장 많으며, 가장 많은 선수가 동시에 출전한다. 이에 많은 데이터가 파생되고 있으며, 기록을 보는 재미가 쏠쏠하다. 최근에는 선수와

*Corresponding Author : Kangsoo You(gsyou@jj.ac.kr)
Received July 26, 2021
Revised August 18, 2021
Accepted August 20, 2021
Published August 31, 2021

팀의 효율성을 측정하기 위하여 과학적인 통계방법을 활용한 2차 기록인 세이버메트릭스를 활용하여 경기 운영, 시즌 구상, 팀 육성 등을 하기에 이르렀다[4-6].

최근 머니볼이라는 영화를 통하여 일반 팬들도 리빌딩이라는 단어를 많이 접하고 있고, 일부 구단들도 경제적인 논리로 선수들을 가격 대 성능비로 판단하기 시작하였다. 또한 신인급 선수들을 육성하는데 많은 관심을 기울이고 있다. 그러나 대한민국 스포츠는 미국에 비하여 규모가 협소하다. KBO리그는 10개 팀 중에서 5개 팀이 포스트 시즌에 진출하는 구조이다. 따라서 리빌딩이라는 단어로 선수육성에 일정 기간을 투자하기에는 스포츠 구단의 경영진들에게 정치적인 부담이 많이 따르고, 팬들의 성원에 부응할수 있는 방법이 아니게 되었다[1,7-10].

이에 본 연구에서는 신인급 선수를 육성하기 위한 방안으로 데이터 분석을 통하여 어느 정도의 기회를 신인에게 부여하는 것이 효과를 보기 시작하는지 알아보았으며, 이를 통하여 데이터 분석을 통한 신인급 선수 육성 방안을 제안하였다. 특히 본 연구는 장타력을 중요시하는 최근 야구계의 트렌드에 부응하기 위하여 장타력이 있는 신인을 발굴하기 위한 방안으로 초점을 맞췄다.

본 연구는 다음과 같은 구조로 이루어져 있다. 서론에 이어서 2장의 선행 연구에서는 스포츠에서 데이터를 분석하는 방법과 야구의 기록 분석 방법에 대하여 기술하였고, 3장에서는 데이터 셋에 대하여 기술하였다. 4장에서는 연구결과를 통하여 논의를 기술하였고, 5장은 결론 및 제언으로 맺는다.

2. 선행연구

2.1 스포츠에서의 데이터 분석

스포츠에서 데이터를 분석하여 활용하려는 예가 최근에 많이 진행되고 있다. 오영환(2020)의 연구에서는 다중선형회귀 분석기법을 통하여 고등학교 투수의 평균자책점을 예측하려 하였다. 이를 위하여 투수의 이닝당 출루 허용율(WHIP: Walks plus Hits divided by Innings Pitched), 피안타율, 탈삼진율, 평균자책점을 측정하였다. 이를 통하여 선수들의 평가와 신인선수 드래프트에 도움을 줄 수 있도록 하였다[11].

또한 김태훈(2020) 등의 연구에서는 KBO리그의 승패 예측 분석을 하기 위하여 각 이닝별로 가장 정확도가 높은 모델을 통하여 최적 모델을 선정하고 승패 결과를 예측하여 순위표를 작성하게 하였다[1]. 언론에서나 미디어에서도 데이터를 기반한 언급이 많이 되고 있으므로 앞으로도 위에서 제시한 연구 이외에도 더 많은 연구가 앞으로도 진행될 것으로 예상된다.

2.2 관련 야구 기록

본 연구에서 활용하는 기록인 타율은 타자가 타석에 들어서서 안타를 얼마나 생산하는지 측정한 것이다. 여기서 안타는 1루타, 2루타, 3루타, 홈런으로 구성되어 있으며 이들은 각각 득점생산력과 가중치가 다르기 때문에 타율만 보아서는 이 선수가 생산력이 있는 선수인지 알 수 없다. 따라서 최근에는 세이버메트릭스를 통하여 WAR(Wins Above Replacement)이나 wRC+(Weighted Runs Created Plus) 등을 측정하여 팀 승리 기여도나 타격생산력을 제시하고 있다. 그러나 세이버메트릭스에는 복잡한 계산식이 많이 있다. WAR을 예로 들면 파크팩터를 적용하여 wRAA(Weighted Runs Above Average)를 산출하고, 수비 지표(Defensive Runs Saved)를 산출한 다음, 포지션별 수비기여도를 보정하여 최종값을 나타낸다. 그러나 일반인들은 산출되는 과정이 직관적으로 이해하기 쉽지 않다. 따라서 누적 및 평균 기록에 의존하지만 스포츠 언론에서도 많이 활용하는 출루율, 장타율, OPS(On-base Plus Slugging)를 타율과 함께 활용하기도 한다[12-15].

3. 데이터 셋

3.1 연구 대상 및 방법

본 연구를 위해 2019년도의 KBO리그의 타자들과 그 기록을 연구대상으로 하였다. 이를 위하여 KBO리그의 공식 홈페이지에 있는 클래식 기록을 수집하였고, 신인 선수의 데이터를 포지션별로 분류하여 전처리를 실시한 후 구간별로 평균 및 누적값을 분석하고 도표와 Heat Map을 활용하여 시각화를 실시하여 본 연구에서 알고자 하는 바를 제안하였다. 연구 방법은 Fig. 1과 같다.

Fig. 1. Research Process

3.2 데이터 수집

본 연구에서는 외부 환경이 스포츠 기록에 미치는 영향을 수집하기 위하여 양질의 데이터를 수집하고자 하였다. 데이터 셋의 종류는 분석하고자 하는 주제에 따라 달라질 수 있다. KBO리그는 팀당 144경기를 진행한다. 따라서 많은 데이터가 발생하며, 선수마다 데이터 발생 정도의 차이가 있다.

본 연구를 위하여 파이썬에서 활용할 수 있는 웹 테스트 자동화 도구인 셀레니움 라이브러리(Selenium Library)로 데이터를 수집하였다. 수집할 데이터는 선수의 포지션별로 수집하였으며 포수, 내야수, 외야수별로 수집하였다. 또한 선수별 평균 기록을 수집하여 타자는 총 310건의 데이터 셋을 수집하였다. Fig. 2는 데이터를 수집하기 위한 작업 중에서 선수의 평균기록 데이터를 수집하기 위한 의사코드이다.

```
if PLAYERS_POSITION == CATCHER then
        PLAYERS_LIST -> CATCHERS_LIST
elseif PLAYERS_POSITION == INFIELDER then
        PLAYERS_LIST -> INFIELDER_LIST
elseif PLAYERS_POSITION == OUTFIELDER then
        PLAYERS_LIST -> OUTFIELDER_LIST
endif
for PLAYER_NAME in ALL_PLAYER_LIST
        goto PLAYERS_AVG_RECORDS
        if DATA_EXISTS == TRUE then
                RAW_DATA = pandas.DataFrame(RECORD)
                RAW_DATA_CSV("DATANAME.csv")
        endif
endfor
```

Fig. 2. Record Collection Pseudo Code

3.3 데이터 분석

본 연구에서는 타자 육성을 위하여 어느 정도의 출전 기회를 부여해야 하는지 데이터로 분석하려 하였다. 이에 출전한 타석 수를 바탕으로 분류한 후 데이터를 전처리하였다. 따라서 전체 선수를 대상으로 한 데이터와 10타석 이상을 출전한 선수부터 시작하여 경기당 한 타석에 가까운 140타석 이상을 출전한 선수에 이르기까지 타석별로 분류하였다. 전처리한 장면은 Fig. 3과 같다.

	AVG	G	PA	AB	R	H	2B	3B	HR	TB	RBI	SAC	SF
SUM	0.2497	2246	6115	5402	549	1349	228	15	116	1955	651	74	54
PA_all	0.2497	48.8261	132.9348	117.4348	11.9348	29.3261	4.9565	0.3261	2.5217	42.5000	14.1522	1.6087	1.1739
PA_10	0.2504	60.0000	164.5135	145.2973	14.8378	36.3784	6.1622	0.4054	3.1351	52.7568	17.5676	2.0000	1.4595
PA_40	0.2524	72.9310	203.4828	179.3793	18.5517	45.2759	7.6207	0.4828	3.9655	65.7586	21.8966	2.5172	1.8276
PA_70	0.2564	91.7000	273.5000	240.8000	25.4500	61.7500	10.3000	0.6500	5.4500	89.7000	30.0000	3.3500	2.5500
PA_105	0.2600	95.7222	294.6111	259.6111	28.0000	67.5000	11.1111	0.7222	6.0000	98.0556	32.6667	3.3333	2.7222
PA_140	0.2667	111.7692	365.3077	321.9231	35.4615	85.8462	13.6923	1.0000	7.8462	125.0769	41.4615	3.8462	3.5385

Fig. 3. Data Pre-Processing

선수가 출전한 타석에 대비하여 나타날 수 있는 데이터를 분석하기 위하여 필요한 파라미터는 Table 1과 같다. PA(Plate Appearance) 항목에 있는 파라미터는 타석에 따른 분류를 한 것이며, Record Name 항목에 있는 파라미터는 야구에서 타자에게 주어지는 각종 기록의 이름을 나열한 것이다.

Table 1. Parameter for Data Analysis

PA	>= 140 >= 105 >= 70 >= 40 >= 10 >=0
Record Name	AVG, OBP, SLG, OPS PA, AB R, RBI H, 2B, 3B, HR, TB BB, SAC, SF

전처리한 데이터를 바탕으로 Fig. 4와 같이 히트맵을 작성하여 선수들의 데이터에서 출전 빈도에 대한 상관관계를 파악하였다.

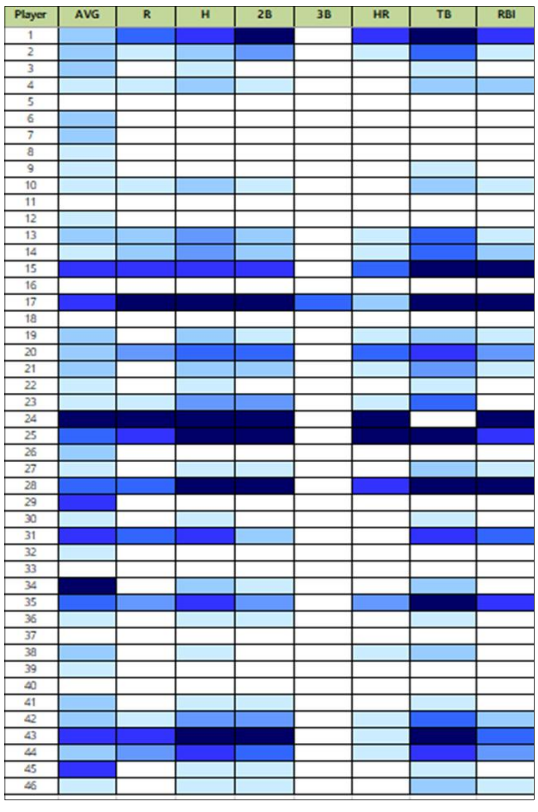

Fig. 4. Heat Map

4. 연구 결과

4.1 타격 성적 변화

본 장에서는 분석한 데이터를 바탕으로 데이터를 시각화하여 논의를 하였다. 예전보다 타율이 경기력을 측정하는데에 있어 중요성이 많이 낮아졌다고 하지만, 선수를 육성하는데 있어서 타격 정확성은 여전히 눈여겨 봐야 할 요소이다.

분석 대상이 되는 시즌의 전체 타자의 평균 타율은 .263이었다. Fig. 5에 따르면 100타석 이하에서는 2할 5푼 대의 성적을 나타내었지만 100타석을 초과하여 출전하였을 때 2할 6푼 대의 성적을 나타내었다. 따라서 100타석 이상의 기회에서 육성의 효과가 나타나기 시작하였음을 알 수 있었다.

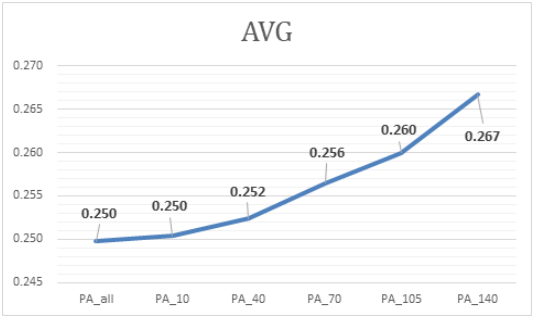

Fig. 5. Visualization for AVG Analysis

4.2 장타율

최근 발사각으로 대표되는 장타의 중요성이 높아짐에 따라 선수 육성에서 장타에 대한 잠재력을 보는 경우가 많아졌다. 이에 본 연구에서는 장타에 대한 데이터로 장타율을 살펴보았다. 분석 대상이 되는 시즌의 전체 타자의 평균 장타율은 .394이었다. Fig. 6에 따르면 타율과 다르게 전체 타자의 장타율을 넘어서지는 못하였지만 140타석 이상의 기회를 부여받았을 때 리그 평균의 장타율에 거의 가까워졌음을 알 수 있었다. 그러나 100타석 근처에서는 성장세가 두드러지지 않았음을 알 수 있다. 이에 장타에 대한 육성에서 정확도에 대한 육성보다 시간이 더 걸렸음을 알 수 있었다.

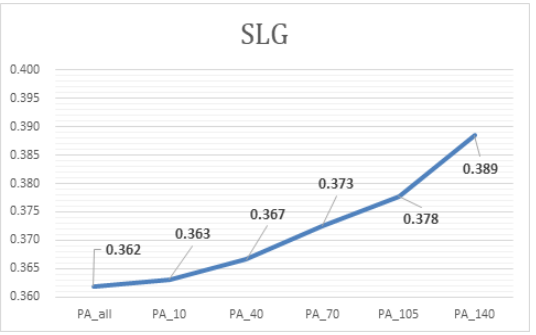

Fig. 6. Visualization for SLG Analysis

4.3 논의

최근 스포츠 구단마다 데이터를 분석하여 선수 육성 및 구단 운영에 활용하고 있는 추세를 보았을 때 본 연구에서와 같이 데이터를 분석하는 것은 선수 육성의 시점을 결정하는 데 도움이 된다고 논의할 수 있다. 분석 대상이 되는 선수들의 데이터 크기에 대하여 반론을 제기할 수 있다. 그러나 대한민국 프로스포츠의 규모를 미루어보았을 때 미국처럼 큰 스포츠 리그와는 다르게 리빌딩이라는 것이 사실상 어렵다. 따라서 성적과 육성을 동시에 진행할 수 밖에 없기 때문에 후보 선수들은 주전 선수들에 비하여 많은 기회를 부여받을 수 없다. 따라서 통상적으로 한 경기당 한 타석 꼴의 기회도 충분히 많은 기회라고 볼 수 있다. 따라서 140타석 정도의 기회를 통하여 선수들은 기회를 살려야 할 것이다.

또한 본 연구에서 진행한 방법을 기반으로 하여 선수들의 경기력을 분석하는 방안을 제시할 수 있었다.

5. 결론 및 제언

본 연구에서는 데이터를 통하여 신인급 선수의 육성 방안과 스포츠 현장에서 활용할 수 있는 분석방법론을 연구하고 제안하였다. 이를 위하여 셀레니움 라이브러리를 활용하여 데이터를 수집하고, 전처리 과정을 통하여 나온 데이터를 기반으로 분류한 후 선수들의 데이터를 분석하였다.

본 연구에서 분석한 데이터를 통하여 신인급 선수를 육성하기 위한 기회는 한 경기당 한 번의 타석에 해당하는 140타석 정도의 기회를 주어 장타에 대한 잠재력까지 검토할 수 있어야 한다. 그렇지 않다면 최소한 100타석 정도의 기회를 주어 최소한의 타격 정확성을 확인할 수 있도록 하여야 한다.

본 연구의 결과를 통하여 스포츠 구단에서 활용 가능한 선수 데이터 분석 방안을 제시할 수 있었다.

향후 연구과제로는 기존의 클래식 데이터인 1차 데이터뿐만 아니라 세이버메트릭스로 대표되는 2차 데이터와 외부 환경을 대입한 3차 데이터까지 대입하여 분석하고, 선수 개인별 육성 방안에 맞는 분석 모델을 제시하는 것이다.

REFERENCES

[1] T. H. Kim, S. W. Lim, J. G. Koh & J. H. Lee. (2020). A study on the win-loss prediction analysis of korean professional baseball by artificial intelligence model. *The Korea Journal of BigData, 5(2),* 77-84. DOI : 10.36498/kbigdt.2020.5.2.77

[2] S. M. Kim. (2020). *The effect of daily average temperature on the batter's performance in baseball game : focused on big data analysis.* Master's Thesis. The Graduate School of Hoseo University, Asan, Chungnam.

[3] Y. H. Oh, H. Kim, J. S. Yun & J. S. Lee. (2014). Using data mining techniques to predict win-loss in korean professional baseball games, *Journal of the Korean Institute of Industrial Engineers. 40(1),* 8-17. DOI : 10.7232/JKIIE.2014.40.1.008

[4] S. H. Lee & H. J. Choi. (2019). The analysis of pitching result according to the velocity and pitch of pitcher in that case of full-counting on Major League Baseball(MLB). *The Korea Journal of Sports Science, 28(3),* 973-981. DOI : 10.35159/kjss.2019.06.28.3.973

[5] Y. N. Jeon. (2010). *A study of the athletic satisfaction and stress factors of high school baseball players in the case of different positions.* Master's Thesis, The Graduate School of Kyungwon University, Seongnam, Gyeonggi.

[6] S. M. Kim & K. S. You. (2020). The effect of daily average humidity on pitcher's stats of a strike-out : focused on high rankers of winning, hold and save. *Journal of Industrial Convergence, 18(1),* 65-71. DOI : 10.22678/JIC.2020.18.1.065

[7] Rein, R. & Memmert, D. (2016). Big data and tactical analysis in elite soccer: future challenges and opportunities for sports science. *SpringerPlus, 5(1),* 1-13. DOI : 10.1186/s40064-016-3108-2

[8] K. W. Kim. (2020). *A study on professional baseball spectators decision making process using model of goal-directed behavior and leisure constraint negotiation model.* Ph. D. Dissertation, The Graduate School of Kyunghee University, Seoul.

[9] J. W. Lee & C. H. Lee. (2019). A study on the analysis of news data for the improvement of local flower festival. *Journal of Industrial Convergence, 17(4),* 33-38. DOI : 10.22678/JIC.2019.17.4.033

[10] Y. H. Oh. (2019). High-school baseball pitcher's pitching speed prediction using linear regression analysis method. *Journal of Knowledge Information Technology and Systems, 14(4),* 381-390. DOI : 10.34163/jkits.2019.14.4.007

[11] Y. H. Oh. (2020). High-school baseball pitcher's ERA(Earned Run Average) prediction using multi-variable linear regression analysis method. *Journal of Knowledge Information Technology and Systems, 15(4),* 497-506. DOI : 10.34163/jkits.2020.15.4.006

[12] J. Y. Lee & H. G. Kim. (2016). Suggestion of batter ability index in Korea baseball - focusing on the sabermetrics statistics WAR. *The Korean Journal of Applied Statistics, 29(7),* 1271-1281. DOI : 10.5351/KJAS.2016.29.7.1271

[13] J. T. Lee. (2017). Suggestion of batter ability index in Korea baseball - focusing on the sabermetrics statistics WAR. *Journal of the Korean Data Information Science Society, 28(2),* 317-326. DOI : 10.7465/jkdi.2017.28.2.317

[14] T. S. Choi & J. H. Yim. (2016). A study on the excellent operation of the "Korea Baseball Hall of Fame" based on baseball records. *Journal of Korean Society of Archives and Records Management, 16(3),* 157-177. DOI : 10.14404/JKSARM.2016.16.3.157

[15] J. T. Lee. (2015). Long term trends in the Korean professional baseball. *Journal of the Korean Data & Information Science Society, 26(1),* 1-10. DOI : 10.7465/jkdi.2015.26.1.1

유 강 수(Kangsoo You) [종신회원]

· 2005년 8월 : 전북대학교 영상공학과(공학박사)
· 1996년 3월~2006년 8월 : 전주대학교 교양학부 객원교수
· 2006년 9월~현재 : 전주대학교 수퍼스타칼리지 교양학부 교수
· 관심분야 : 스포츠데이터과학, 영상처리, 컴퓨터비전, 소프트웨어교육
· E-Mail : gsyou@jj.ac.kr