Tensorflow를 활용한 야구선수 유형 분석 및 타격 결과 예측에 관한 연구

박채원[1], 박지범[2], 주영준[3], 김현석[4], {이남용, 김영종}*
[1,2,3,4,*]숭실대학교 소프트웨어학부
e-mail: pcwyvonne@gmail.com, mayer0425@naver.com, juyi7282@gmail.com,
sfsfkj@gmail.com, [nylee, youngjong]@ssu.ac.kr*

A Study on Baseball Players' Type Analysis and Prediction of Batting Result by using Tensorflow

Chaewon Park[1], Jibeom Park[2], Yeongjun Joo[3], Hyunseok Kim[4],
{Namyong Lee, Youngjong Kim}*
[1,2,3,4,*]School of Software, Soongsil University

요 약

본 연구는 한국 프로 야구 선수 개인의 수치화된 데이터를 바탕으로 타석의 결과를 예측하고자 하는 데 목적을 두고 있다. 연구의 방법은 2015시즌부터 2018시즌에 활약한 한국 프로 야구 소속의 투수와 타자의 유형을 군집화 하여 지도학습 모델을 만든다. 지도학습 모델과 현재까지 진행된 2019시즌의 결과를 비교·대조한다. 본 연구결과는 한국 프로 야구 10개 구단의 감독의 선수 선발 결정에 기여할 것으로 판단된다.

1. 서론

한국 프로 야구(KBO)가 2019 시즌의 목표 관중을 역대 최대인 878만 명으로 잡은 만큼, 야구는 해마다 큰 인기를 누리고 있다. 야구팬들은 '올해의 우승 구단', '자신이 응원하는 구단의 성적', '어떤 선수가 작년에 비교해 더 나은 성적을 낼지'에 대해 예상하며 자신이 좋아하는 구단 혹은 선수를 응원한다. 본 연구에서는 2015시즌부터 2018시즌의 모든 프로 야구 경기에서 발생한 10개 구단의 결과를 인공지능 접근 방식 중 하나인 Machine Learning 기법을 바탕으로 투수와 타자 간 대결 시 나오는 결과를 예측하고자 한다.

2.본론

2-1. 데이터 수집 및 전처리

한국 프로 야구(KBO) 공식 홈페이지와 Statiz, 레전드닷컴, KBReport와 같은 사설 사이트에서 2015시즌부터 2018시즌까지 총 4시즌에 활약한 선수들의 개별 데이터(타자 당 타율, 출루율, 장타율, OPS, wOBA, wRC+ 등)를 수집한다. 수집한 데이터를 바탕으로 투수의 유형과 타자의 유형을 각 4가지로 나눠 분류한다.

그림 1 투수 유형

그림 1에 나타나 있는 'Power Pitcher' 는 구위를 바탕으로 삼진을 잡는 투수이며, 'Finesse Pitcher' 는 제구력을 바탕으로 범타를 유도하는 투수이다.

또한 'F/B Pitcher' 는 뜬공(Fly ball)을 유도하는 투수이고, 'G/B Pitcher' 는 땅볼(Ground ball)을 유도하는 투수이다.

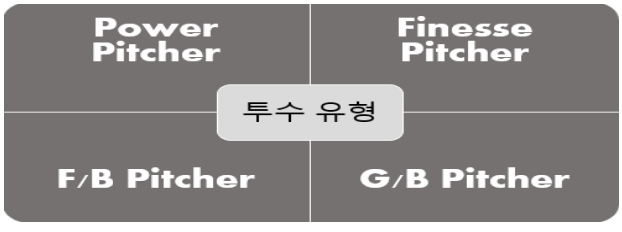
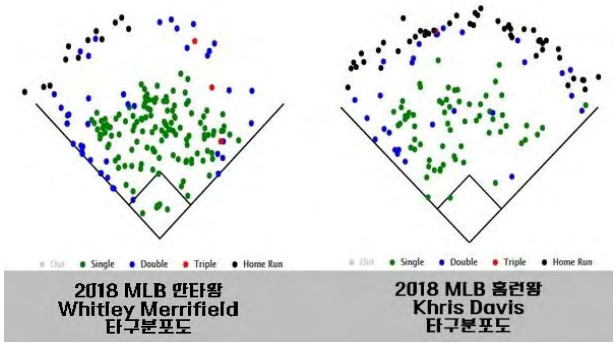


그림 2 타자 유형

'Power Hitter' 는 중장거리 타구 생산에 특화된 타자이며, 'Finesse Hitter' 는 단타와 출루에 집중하는 타자이다. 'F/B Hitter' 는 발사각이 높은 홈런성 타구를 생산하는 타자이며, 'G/B Hitter' 는 발사각

이 낮은 땅볼 타구가 형성되는 타자이다.

그림 3 은 2018시즌 메이저리그(MLB)의 타자인 Single Hitter인 Whitley Merrifield와 Power Hitter 인 Kristopher Davis 의 타구분포도이다. 이 자료를 통해 타자의 유형을 시각적으로 확인할 수 있다.



2-2. 선수 유형 분류 및 데이터 분석 방법

표 1 유형 분석에 사용되는 타자의 지표

구분	지표
타자	IsoP(Isolated Power, 절대장타율)
	BB(볼넷 출루)
	k(삼진)
	FO/GO(땅볼 아웃 대비 뜬공 아웃)
	XH/H(타석 대비 장타 개수)

야구의 특성상 각기 다른 세부 통계도 상관관계가 존재한다. 이를 극복하고 명확한 유형 분류 결과를 얻기 위해 주성분 분석(PCA) 알고리즘을 사용한다. 이를 통해 부적절한 분류 결과를 방지하고, k-평균 알고리즘(K-means algorithm)등의 기법을 이용하여 선수의 유형을 clustering한다.

표 2 유형 분석에 사용되는 투수의 지표

구분	지표
투수	PFR(Power finesse ratio) (삼진+볼넷)/이닝
	FO/GO(땅볼 아웃 대비 뜬공 아웃)
	FB%(Fly Balls, 뜬공 비율)
	k/9(9이닝당 탈삼진수)
	BB/9(9이닝당 볼넷 허용수)

clustering한 결과를 통해 타자-투수간 유형별 상대전적 데이터를 구축할 수 있다. 이에 더하여 각 선수의 상대 전적 데이터, 투수의 구종가치[1]와 구종/위치별 구사율, 타자의 Hot and Cold Zone[2] 등을 이용해 실제 타석의 결과를 예측한다.

Artificial Neural Network(ANN)은 유기물의 학습 메커니즘을 시뮬레이션하는 범용적인 기법이다. ANN은 입력 뉴런에서 출력 뉴런으로 계산된 값을 전파하고 가중치(weight)를 중간 파라미터로 사용하여 입력의 함수를 계산한다. 학습은 뉴런에 연결된 가중치의 값에 변화를 주며 일어난다.

우리는 ANN을 이용하여 타석 결과를 예측하기 위한 신경망 구현 모듈로 Tensorflow를 사용하고자 한다. 그리고 Tensorflow을 통해 지도학습 모델을 학습 시켜 새로운 투수와 타자가 매칭이 될 때의 결과를 예측한다.

2-3.학습 후 평가방법

2019시즌 5월까지 진행된 투수와 타자의 각 유형별 기록을 수집해서 결과를 추출한다. 그 자료를 이전에 지도학습을 통해 생성한 모델을 바탕으로 투수와 타자간 대결 결과(안타율)를 비교·분석한다. 도출해낸 결과를 1회에 그치지 않고, 여러 번에 걸쳐 비교·대조를 함으로써 가장 높은 예측률을 가진 모델을 찾는다.

ACKNOWLEDGMENT

"본 연구는 과학기술정보통신부 및 정보통신기술진흥센터의 SW중심대학사업의 연구결과로 수행되었음 (2018-0-00209-001)"

참고문헌

[1] Neural Networks and Deep Learning, Charu C.Aggarwal, Springer, 2018

1) 구종의 가치. 높을수록 타자가 아웃될 확률이 높다.

2) 스트라이크 존을 9등분 하여, 각 위치의 투구를 타격했을 시의 타율을 나타낸 표.