야구 기록과 뉴스 데이터를 활용한 타격 예측 모델

A batting average prediction model based on news data analysis and baseball record

저자
(Authors)

상의정, 정창권, 이하정, 한용구, 이영구
Uijeong Sang, Changkwon Jeong, Hajeong Lee, Yongkoo Han, Young-Koo Lee

출처
(Source)

한국정보과학회 학술발표논문집 , 2016.6, 2020-2022(3 pages)

발행처
(Publisher)

한국정보과학회
KOREA INFORMATION SCIENCE SOCIETY

URL

http://www.dbpia.co.kr/journal/articleDetail?nodeId=NODE07018060

APA Style

상의정, 정창권, 이하정, 한용구, 이영구 (2016). 야구 기록과 뉴스 데이터를 활용한 타격 예측 모델. 한국정보과학회 학술발표논문집, 2020-2022

야구 기록과 뉴스 데이터를 활용한 타격 예측 모델

상의정O, 정창권, 이하정, 한용구, 이영구1)
경희대학교 컴퓨터공학과
syj5982@gmail.com, 92fmnsos@gmail.com, jjudy94@hanmail.net, ykhan@khu.ac.kr, yklee@khu.ac.kr

A batting average prediction model based on news data analysis and baseball record

Uijeong SangO, Changkwon Jeong, Hajeong Lee, Yongkoo Han, Young-Koo Lee
Dept of Computer Science and Engineering, Kyung Hee University

요 약

야구는 기록의 스포츠로서 다양한 기록이 만들어지고 쌓인다. 이러한 기록은 선수의 가치를 평가하는 척도 등의 용도로 다양하게 야구계에서 활용되고 있다. 이러한 과거의 기록들은 현재의 상황을 예측하는 척도로서 사용될 수 있다. 하지만 사람의 행동은 객관적인 지표 이외의 감정이나 상황 등에 의해서도 결정될 수 있다. 최근 기계학습 분야에서는 문서의 분류와 감정 분석과 같은 기술들이 개발되고 실제 영화의 평가 분석등의 다양한 분야에서 활용되고 있다. 야구와 같은 스포츠는 타율과 같은 정량적인 지표 뿐 아니라 뉴스와 같은 매체를 통해 확인할 수 있는 정성적인 지표 또한 선수의 상태에 영향을 미칠 수 있다. 본 논문은 기계학습을 통해 분류된 뉴스데이터의 정보를 활용하여 실제로 뉴스 데이터를 통해 나타나는 정성적인 지표를 제안하고 선수의 최근 기록과 함께 예측 모델의 독립변수로 활용하여 예측하고자 하는 타자의 타율의 상승하강 여부를 분류하는 모델을 제안한다. 실험을 통해 기존 연구에 비해 50.4%의 분류 정확도 상승을 보인다.

1. 서 론

야구는 기록의 스포츠로서 수많은 기록이 만들어지고 쌓이고 있다. 그리고 획득한 다양한 기록을 바탕으로 팀과 선수에 대한 분석이 이루어진다. 스포츠 뉴스는 그날 그날의 선수와 팀의 상태와 현황을 알려주는 지표이다. 뉴스를 통해 사람들은 팀과 선수의 흐름을 보는 것이 가능하다. 야구 기록을 활용한 분석방법으로는 세이버메트릭스가 있다. 세이버메트릭스는 기존의 방식인, 기록의 단순 통계로는 선수의 객관적인 능력을 판단하기 어렵다고 보았다. 이를 해결하기 위하여 게임 이론과 통계학적 방법론을 도입한 객관적인 평가 지표를 만들었다. 하지만 게임의 상황에 대한 예측은 불가능하다.

기계학습은 데이터에 대한 일반화를 통해 컴퓨터가 학습할 수 있도록 하는 기술이다. 기계학습의 여러 분야 중 패턴인식은 인간의 추론능력을 모델링하여 외부 대상에 대해 인식하고 패턴을 이해하는 기술을 의미한다. 패턴인식은 자연언어의 구문적 패턴을 분석과 외부 대상을 지각하는 기술 등의 다양한 응용분야를 갖고 있다. SVM[1](support vector machine)은 이러한 패턴인식 학습 모델의 하나로 데이터의 카테고리를 판단하는 분류 모델을 만들며 분류와 회귀분석에 사용된다.

본 논문은 야구 기록과 SVM을 사용한 뉴스 데이터 분석을 통해 타자의 타율 상승하강을 예측할 수 있는 분류 모델을 제안한다. 기존연구에서 제안하는 상대 투수의 정보와 경기장 별 데이터에 비해 타자의 최근 데이터가 타율에 더 큰 영향을 미치는 것을 보인다. 또한 객관적인 선수의 지표인 기록 이외의 뉴스와 같은 정성적인 지표로 타격 분류 정확도를 올린다. 야구 기록들이 타자의 타격확률에 영향을 미치는지에 대해 분석하고 뉴스 데이터를 학습하여 뉴스기사들이 나타내는 팀과 선수의 상태를 타격에 영향을 주는 지표로서 모델에 적용한다. 이를 통해 기존의 통계가 보이지 못했던 당일 타자의 타격 예측결과를 보이고 뉴스에서 나타나는 선수와 팀의 상태와 흐름이 경기에 미치는 영향을 보인다.

2. 관련 연구

기존의 야구 기록을 활용한 예측 모델로는 로지스틱 회귀분석 모델을 통해 가장 안타를 칠 확률이 높은 타자를 예측하는 연구[2]가 있었다. 입력값으로 최근 몇 시합에서의 타율, 최근 몇 시합에서 타자가 공을 배트에 맞춘 평균 횟수, 투수의 최근 피안타율, 타자의 경기장에 따른 타율을 사용하여 로지스틱 회귀분석 모델을 만들어 예측을 수행하였다.

SVM은 분류와 회귀분석에 사용되는 지도 학습 모델이다. 주어진 집합을 바탕으로 두 개의 카테고리를 나누는 선형 분류 모델을 만든다. SVM은 지도학습 모델로 분류 모델을 만들 때의 분류 기준은 사용자가 직접 정한다. 정해진 분류 기준에 따라 각각의 데이터의 특징과 특징 값들을 모델 학습에 사용한다. 새로운 데이터가 들어오면 경계로 표현되는 분류 모델을 기준으로 어떠한 카테고리에 들어가는 데이터인지 판단하게 된다.

TF-IDF[3]는 텍스트 마이닝에서 사용되는 가중치이다. TF는 특정 단어가 문서 내에 어느정도 자주 등장하는지

1) 교신저자

를 나타내는 값으로 값이 높으면 중요도가 높은 단어로 판단할 수 있다. 하지만 문서군 내에서 자주 사용되는 흔한 단어인 경우도 있을 수 있다. 이 가중치는 DF로 이 값의 역수를 IDF라 부른다. TF와 IDF를 곱한 값이 TF-IDF로 이 값이 높은 경우 특정 문서 내에서 단어 빈도가 높고 전체 문서들 중에서는 그 단어를 포함한 문서가 적은 단어로 판단할 수가 있다.

3. 야구 기록과 뉴스 데이터를 활용한 타격 예측 모델

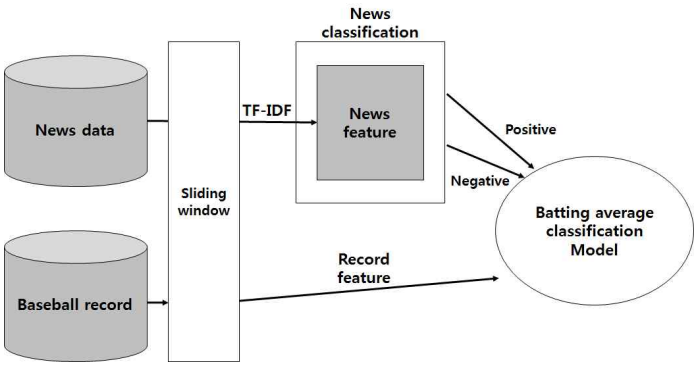

그림 1. 분류를 사용한 타격 예측 모델 구조

그림 1.은 제안하는 타격 예측 모델의 전체적인 구조를 나타낸다. 우선 뉴스 데이터와 야구 기록을 슬라이딩 윈도우의 크기만큼 가져온다. 수집된 뉴스 데이터는 TF-IDF를 사용하여 분류에 사용할 특징들로 변환된다. 변환된 뉴스 특징은 긍정 부정 레이블을 나누어 분류를 시행한다. 분류기를 통해 분류된 뉴스의 긍정과 부정 여부 그리고 야구 기록을 타율 예측 모델에 사용하여 최종적으로 타율 예측을 수행한다.

3.1 TF-IDF와 SVM을 사용한 뉴스 분석

SVM을 사용하여 긍정과 부정의 분류하기 위해서 각각의 분류할 데이터가 갖고 있는 특징들의 값을 정해주어야 한다. 뉴스를 분석하는 경우 뉴스의 특징들은 뉴스에 나오는 단어가 되고 이 특징이 되는 단어들을 값으로 표현해주어야 한다. 이 때 뉴스에 나오는 모든 단어들을 특징으로 사용하게 되면 각각의 뉴스의 고유한 특징을 분류에 사용하는 것이 어려워져 분류의 정확도가 떨어지게 된다.

뉴스를 분류에 사용하기 위하여 각각의 뉴스에 나오는 고유한 특징과 특징의 값을 정해주는 방식으로 TF-IDF를 사용한다. 다수의 야구 뉴스 데이터에서 나타나는 단어들 중 특정 뉴스에서만 나타나는 유의미한 단어들을 파악하여 특징이 되는 단어만을 뉴스의 특징으로 활용한다. 특징으로 사용되는 단어들은 각각 고유한 가중치를 갖고 있음으로 이를 분류에 사용할 고유한 특징의 값으로 활용한다.

그림 2.와 같이 TF-IDF를 사용하여 여러 특징들로 나타내는 것이 가능해진 뉴스 데이터를 SVM을 사용하여 학습한다. 학습을 위한 긍정과 부정의 분류 레이블을 각 뉴스마다 지정해주었다. 학습에 사용될 뉴스 데이터의

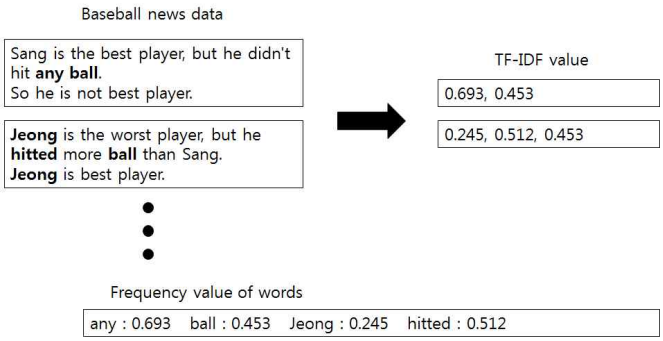

그림 2. TF-IDF를 사용한 뉴스 특징 분석

내용이 긍정적인 경우 +1, 부정적인 경우 -1의 레이블을 부여해준다. 최종적으로 레이블을 지정한 각 뉴스 데이터의 특징을 학습에 사용하게 된다. 최종적으로 생성된 분류 모델을 사용하여 뉴스 데이터를 분류하게 되면 뉴스의 특징들이 갖는 긍정적인 정도와 부정적인 정도 중 어느 부분의 특징을 더 많이 갖고 있는가를 사용하여 뉴스 데이터의 분석 결과가 나오게 된다.

3.2 SVM 분류모델을 활용한 타격 예측 모델

본 논문은 야구 기록과 뉴스 분석 결과를 입력 값으로 사용하여 SVM을 사용한 타율 상승하강 예측 분류모델을 생성한다. 타자의 타격 확률에 가장 큰 영향을 주는 데이터는 타자의 최근 경기 데이터이다. 그렇기 때문에 최근 선수의 경기 상태를 고려할 수 있는 분류 모델을 만든다. 분류 모델에 사용하는 야구기록 특징들로는 최근 경기의 타율 평균과 타석 수, 안타, 홈런 개수, 장타율과 출루율을 사용하였다. 뉴스 데이터 분석을 통해 얻은 긍정적 뉴스와 부정적 뉴스는 뉴스가 나온 날짜의 긍정, 부정 뉴스 개수로 나타내어 분류에 특징으로 사용하였다.

이러한 타자의 타격에 영향을 주는 지표들은 시간이 지남에 따라 변화한다. 이러한 시간에 따른 변화를 나타내기 위하여 n일 단위의 슬라이딩 윈도우를 만든다. 종속변수로 매일의 타율, 윈도우 사이즈인 n일 안의 뉴스와 기록을 독립변수로 사용한다. 과거의 데이터를 사용하여 n일 단위의 독립변수들과 타율의 증감을 분류 클래스 레이블로 사용하여 분류 모델을 생성한다.

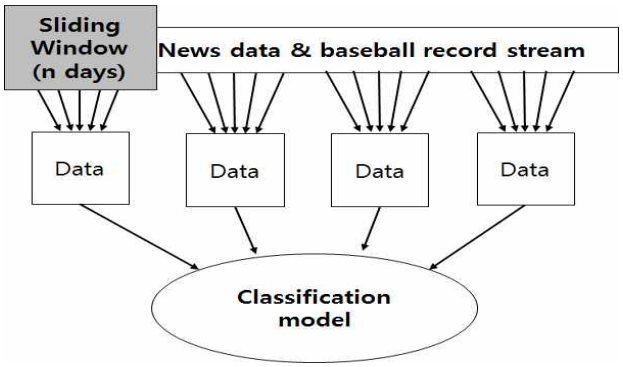

그림 3. 슬라이딩 윈도우를 사용한 분류분석 모델

그림 3.과 같이 슬라이딩 윈도우를 이동시키면서 분류

에 사용되는 데이터들을 모은다. 하나의 슬라이딩 윈도우에는 정해진 시간만큼의 뉴스 데이터 분석결과와 기록이 분류모델의 독립변수로 사용된다. 이를 통해 시간의 변화에 따른 타자의 타율 변화를 고려할 수 있게 된다.

4. 성능 평가

실험을 위해 MLB 뉴스 데이터와 MLB 선수 기록 데이터[4]을 사용하였다. 실험에 사용된 환경은 Intel(R) Core(TM) i5-3570 CPU @ 3.40GHz, RAM 8.00GB, OS Windows7에서 기계학습 라이브러리 WEKA[5]를 이용하였다. 아래 표는 실험 데이터 셋에 대한 통계치를 나타낸다.

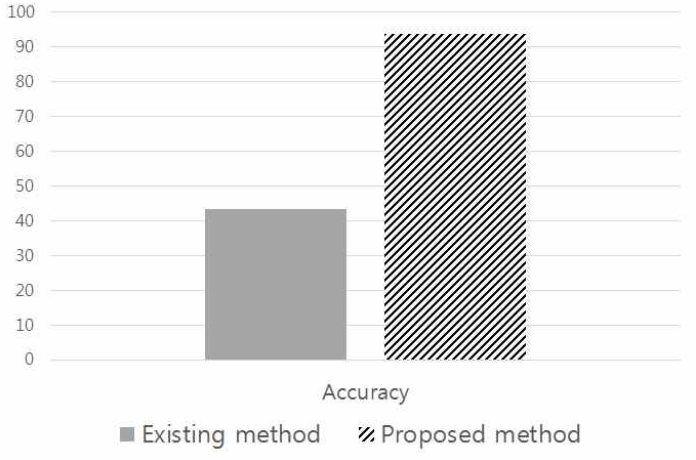

그림 4. 기존 연구 모델과 제안하는 모델의 정확도 비교

그림 4.를 통해 기존 연구에서 사용한 특징만으로 분류한 분류 모델에 비해 뉴스데이터를 추가로 특징으로 사용한 분류 모델이 약 50.4%의 분류 정확도 증가를 보였다. 이를 통해 타자의 최근 경기 결과들과 뉴스 데이터가 구장별 선수의 기록과 투수의 컨디션과 같은 기록에 비해 타율 예측에 큰 영향을 주고 있는 것을 알 수 있다.

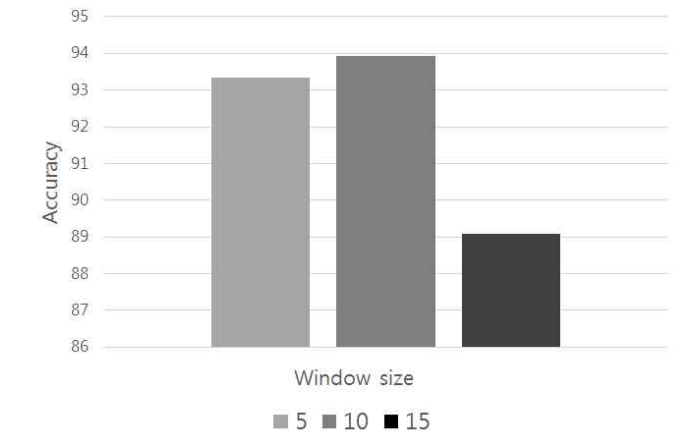

그림 5. 윈도우 사이즈 변화에 따른 정확도 비교

제안하는 방식에서 사용되는 특징은 모두 n의 크기를 갖는 슬라이딩 윈도우만큼의 날짜 간 기록평균을 사용하고 있다. 이 슬라이딩 윈도우의 크기 변화에 따른 정확도 평가 결과 10의 크기를 갖는 슬라이딩 윈도우의 정확도가 가장 높았다. 이는 10의 윈도우 사이즈가 야구 선수의 약 2주가량 경기를 바탕으로 분석하고 최근 2주의 경기가 현재 선수의 기량을 가장 잘 나타내는 지표이기 때문이다.

5. 결론 및 향후 연구

야구에서 기록이라는 정량적인 데이터를 활용한 선수에 대한 분석은 지속적으로 이루어져 왔다. 기록을 활용한 기존연구는 타자의 타율을 예측할 때 투수의 피안타율, 경기장 별 타자의 타율과 같은 기록을 사용했다. 하지만 선수의 타율은 과거의 기록보다는 최근 몇 경기 동안의 타자의 기록과 팀과 선수의 환경과 상태 변화 등 다양한 요인의 영향을 받게 된다. 본 연구는 선수의 타율에 영향을 주는 정성적인 특징들을 야구 뉴스 데이터를 활용하여 분석하고 타율에 큰 영향을 주는 최근 경기의 타자 기록을 활용하여 분류 모델을 만들었다. 분류 모델의 정확도 측정을 통해 정량적 데이터와 정성적 데이터를 모두 활용한 타율 예측이 기존에 비해 높은 정확도를 보였다. 추후 연구를 통해 본 논문의 뉴스 분류 기준인 긍정, 부정 여부 뿐 아니라 야구에 더 특화된 분류 기준을 만들어 뉴스를 분류하고 뉴스가 타자의 타율에 미치는 영향을 더 면밀히 살펴볼 것이다.

※ “본 연구는 미래창조과학부 및 정보통신기술진흥센터(IITP)에서 지원하는 서울어코드활성화지원사업의 연구결과로 수행되었음” (R0613-16-1203)

참 고 문 헌

[1] MANEVITZ, Larry M.; YOUSEF, Malik. “One-class SVMs for document classification.”, the Journal of machine Learning research, 2002, 2: 139-154.

[2] Clavelli, Jason, and Joel Gottsegen. “Maximizing Precision of Hit Predictions in Baseball.”, 2013

[3] PAIK, Jiaul H. “A novel TF-IDF weighting scheme for effective ranking.”, Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval. ACM, 2013. p. 343-352.

[4] MLB.com ¦ The Official site of Major League Baseball, http://mlb.mlb.com/

[5] HALL, Mark, et al. “The WEKA data mining software: an update.”, ACM SIGKDD explorations newsletter, 2009, 11.1: 10-18.