랜덤 포레스트 기법을 이용한 한국 프로야구 승부 예측에 관한 연구

이재익, 이종혁, 김응모
성균관대학교 정보통신대학
e-mail: yjipotter@naver.com, jhl0866@skku.edu, ukim@skku.edu

A Study on Result Prediction of Korean professional baseball using Random Forest Method

Jaeik Yi, JongHyeok Lee, Ung-Mo Kim
College of Information & Communication Engineering,
Sungkyunkwan University

요 약

야구는 흔히 기록의 스포츠라는 별명으로 많이 불린다. 그만큼 야구라는 운동이 갖는 기록의 종류는 무척 다양하고 또한 기록의 활용 가능성 역시 무궁무진하다. 이러한 별명에 걸맞게 미국에서는 야구에 대한 다양하고 방대한 정보를 수집하고 활용하고 있다. 그러나 한국 프로야구에 대한 정보의 수집과 활용은 아직까지 크게 부각되지 못하는 것이 현실이다. 랜덤 포레스트 기법을 이용하여 경기의 승부를 예측함으로써 한국 프로야구 데이터의 수집과 활용을 증대 시키는 효과를 기대 해 본다. 본 논문에서는 2014년 한국 프로야구의 승부 예측을 주제로 어떠한 누적 스포츠 데이터집단이 가장 유효한지를 실험 하였다. 승부 예측을 하기위해 사용된 누적 스포츠 데이터는 2014년 선수와 팀 기록, 2013부터 2014년까지의 선수와 팀 기록, 2012년부터 2014년까지의 선수와 팀 기록이다. 이들 세 그룹의 데이터를 이용하여 이분데이터 모형에 랜덤 포레스트 기법을 사용한 승부예측 알고리즘에 적용 시킨 후 어느 그룹의 데이터가 가장 실제 2014 한국 프로야구 경기결과와 맞을 확률이 높음을 구하여 가장 유용한 데이터 그룹이 어떤 그룹인지 연구 하였다.

1. 서론

현대 사회에서 스포츠는 하나의 문화로 자리 잡았다. 직접 운동을 하며 즐기는 것뿐만 아니라 스포츠 경기를 관람하고 자신이 좋아하는 팀을 응원하며 또한 그 팀과 관련되어 있는 다른 활동들을 할 수 있다는 점에서 스포츠는 정신적, 육체적 여가활동의 역할을 동시에 수행 할 수 있는 매력적인 여가활동이다. 이 논문에서는 한국에서 대중적인 스포츠 중 하나인 프로야구에 대해서 다루었다.

변수가 많은 팀 스포츠 경기에서는 명확히 드러나는 기록 외에도 잘 드러나진 않지만 상관관계가 있는 변수들 또한 고려되어야 한다. 이는 어려운 일이지만 야구는 타 스포츠에 비해 데이터 분석과 예측에 강점을 가지고 있다. 공격과 수비가 명확히 구분된다는 점, 투수와 야수, 타자, 주자 등 각 선수의 역할이 부여된다는 점, 시간제한이 없는 스포츠라는 점, 팀마다, 타자마다 동일한 3아웃, 9이닝의 기회를 부여받는다는 점 등이다. 이 때문에 동일한 조건 하에서 변수들을 설정하고 데이터를 분석 할 수 있다. 또한 프로야구 타 프로 스포츠에 비해 경기수가 많다. 한국 프로야구는 2014년 각 팀당 정규시즌 경기수가 128게임, 미국 프로야구는 2014년 각 팀당 정규시즌 경기수가 162경기로, 이는 한국 프로축구의 팀당 38경기, 미국 프로축구의 팀당 36경기와 비교하면 3배가 넘는 경기 횟수이다. 많은 경기를 토대로 축적된 데이터를 통해 더욱 신뢰성 있는 분석 결과를 도출 할 수 있다.

모든 스포츠선수들은 우리가 흔히 말하는 기복을 가지고 있다. 선수 생활 중 한 시즌 20홈런을 친 타자나 한 시즌 15승을 거둔 투수는 같은 기록을 세울 잠재력을 가지고 있다. 비록 현재 부진 한 선수라도 중요한 순간에 해결사가 될 수 있는 것이고 만약 과거에 그런 경험이 있다면 그 확률은 더욱 높아진다. 이는 통산 기록이 훌륭한 베테랑 선수가 좋은 대우를 받고 승부처에 투입되는 것과 일맥상통한 일이다.

그렇기 때문에 이번시즌 성적 데이터만을 사용하는 것이 아니라 직전 시즌과 그 전 시즌의 데이터까지 사용하여 승부 예측의 정확도를 측정하는 실험을 통해 축적된 스포츠 데이터의 유효한 기간이 어느 정도까지인지 알아보았다. [1]에서 사용되었던 이분데이터 모형에 랜덤포레스트 기법을 사용한 알고리즘을 이용하여 승부 예측 결과를 도출 해 실제 경기 결과와의 적중률을 분석 했다. 분석에 쓰일 데이터는 3가지 그룹으로 나누어 실험 하였고. 이 세 그룹 데이터의 분석 결과와 실제 경기의 결과와 비교해서 가장 실제 결과에 가까운 그룹을 찾고 유효성을 분석 할 것이다. 이를 통하여 현재부터 어느 정도 이전까지의 스포츠 데이터가 승부에 직접적인 영향을 끼치는 지

분석이 가능할 것이라 예상했다.

2. 제안 기법

본 논문에서는 [2]가 수립했던 랜덤포레스트 기법을 이용한 이분데이터 모형을 사용했다. 이모형에서 사용된 변인은 크게 팀, 타자, 투수로 나뉜다.

변인들을 설정 한 후 2014년 데이터, 2013년과 2014년 데이터, 2012년부터 2014년까지의 데이터 세 그룹으로 나누어 실험을 실시한다. 사용한 데이터는 한국 야구위원회(KBO) 홈페이지 기록실 제공한 선수자료들과 경기일정을 바탕으로 수집하였다. 타율, 출루율, OPS등의 비율 데이터는 비율을 계산하여 사용 했지만 타점, 볼넷, 삼진 등 수치 데이터는 각 경기 시의 2014년 비율에 맞추어 변환했다. 예를 들면 SK 와이번스 팀의 이재원 선수는 2013년 69경기 247타석에 출전하여 57안타, 8홈런, 17볼넷 등을 기록 했다. 이를 2014년 8월 28일 LG 트윈스와 한 경기의 결과 예측을 위해 2013년의 기록을 변환 해 보면, 직전경기까지 이재원 선수가 들어간 타석은 409타석이다. 이를 2013년과 비교 해 본다면 409/247의 비율만큼의 타석을 들어간 것이다. 따라서 409/247의 결과인 1.65587을 2013년의 각 기록에 곱하여 반올림하면 89안타, 12홈런, 27볼넷의 변환 된 수치가 나온다. 이를 이용해 2013년과 2014년 데이터 그룹으로 사용 할 경우 변환된 2013년 수치와 2014년 경기 직전까지의 수치를 평균 내어 반올림 한 후 109안타, 12홈런, 36볼넷의 기록으로 사용 할 수 있다. 다만 한 시즌 출전 경기 수가 팀이 출전 한 전체 경기수의 1/3 이하인 시즌이 있는 선수의 데이터는 사용하지 않고 2014 시즌의 기록만 사용했다. 왜냐하면 한 시즌 성적이 예를 들어 2타석 1타수 1안타 1사구인 선수의 경우 수치 변환 시 타율이 1.00이 되는 사태가 발생 할 수 있고 또한 사용하려는 데이터의 크기가 충분히 크지 않아 신뢰하기 힘들다고 생각했기 때문이다.

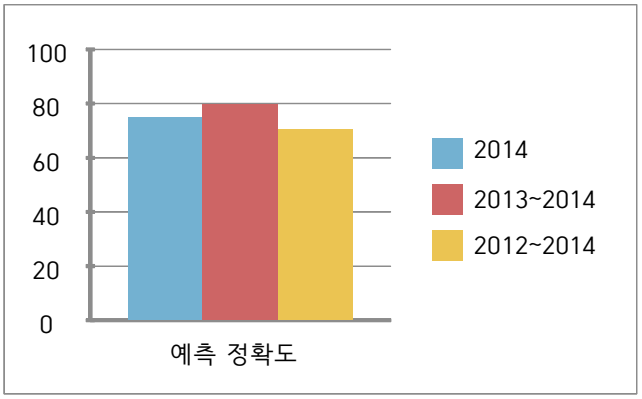
3. 실험

2014 한국 프로야구 2014년 3월 8일 ~ 2014년 8월 28일 까지 경기의 승부를 랜덤포레스트 기법을 이용한 이분데이터 모형을 사용해 예측해 보았다. 이 중 각 팀별 첫 20경기는 유효한 데이터의 축적을 위해 예측을 하지 않았고 각 팀마다 2014시즌 21번째 경기부터 승부를 예측해 보았다. 사용한 랜덤포레스트 기법을 이용한 이분데이터 모형은 승리와 패배만 예측 가능함으로 실제 경기 결과가 무승부가 된 경기는 결과에서 제외했다. 2014년 8월 28일까지의 초반 20경기를 제외 한 각 팀별 경기 수는 아래 <표 1>와 같다. 경기는 총 378경기이다. 이 378경기에 각각의 선수와 팀의 2014년 기록, 2013년과 2014년 기록, 2012년부터 2014년까지의 기록을 변환하고 랜덤포레스트 기법을 이용한 이분데이터 모형에 대입하여 예측 결과를 산출 하였다. 산출된 예측 결과를 각 그룹별로 실제의 경기 결과와 비교해 보았다. 실제 경기 결과와 예측 결과의 데이터 그룹 별 일치 확률은 아래 <표 2>에 나타내었다.

<표 1> 프로야구 각 팀별 경기 수

팀명	삼성	넥센	NC	LG	두산	롯데	SK	KIA	한화	총합
총경기	84	89	88	89	85	87	87	87	84	780
무승부	2	3	3	3	3	3	3	3	1	24
무승부 제외한 경기수	82	86	85	86	82	84	84	84	83	756

<표 2> 데이터 그룹별 예측결과 일치 확률



4. 결론

2014년의 기록만 사용하여 예측한 결과는 총 378경기 중 283경기의 결과와 일치 해 74.87%의 정확도를 보였고 2013년과 2014년의 기록을 사용하여 예측 한 결과는 301경기와 일치해 79.62%의 정확도를 보였다. 마지막으로 2012년부터 2014년까지의 기록을 사용하여 예측 한 결과는 총 378경기 중 266경기의 실제 결과와 일치해 70.37%의 정확도를 보였다. 이를 바탕으로 한국 프로야구의 기록은 이번 시즌과 직전 시즌의 기록의 합이 가장 신뢰성 있고 유용한 데이터 집단이라는 것을 알 수 있었다.

사사표기

본 연구는 미래부가 지원한 2013년 정보통신·방송(ICT) 연구개발사업의 연구결과로 수행되었음. 이 논문은 2013년도 정부(교육부)의 재원으로 한국연구재단의 기초연구사업 지원을 받아 수행된 것임(NRF-2013R1A1A2008578)

참고문헌

[1] 오윤학, 김한, 윤재섭, 이종석, “데이터마이닝을 활용한 한국프로야구 승패예측모형 수립에 관한 연구”, 대한산업공학회지(Journal of the Korean Institute of Industrial Engineers) Vol.40 No.1, 2014

[2] 오윤학 ,김한 ,윤재섭 ,김시명, “데이터마이닝을 활용한 한국프로야구 승패예측 모형 수립에 관한 연구”, 대한산업공학회 추계학술대회논문집 Vol.2013 No.11 ,2013