야구 데이터 분석을 통한 데이터 축소 방안 연구

A Study on Reduction of Data Through Baseball Data Analysis

저자 (Authors)
박대서, 김화종
Park Dae Seo, Kim Hwa Jong

출처 (Source)
한국통신학회 학술대회논문집 , 2016.11, 244-245(2 pages)
Proceedings of Symposium of the Korean Institute of communications and Information Sciences , 2016.11, 244-245(2 pages)

발행처 (Publisher)
한국통신학회
Korea Institute Of Communication Sciences

URL
http://www.dbpia.co.kr/journal/articleDetail?nodeId=NODE07082710

APA Style
박대서, 김화종 (2016). 야구 데이터 분석을 통한 데이터 축소 방안 연구. 한국통신학회 학술대회논문집, 244-245

야구 데이터 분석을 통한 데이터 축소 방안 연구

박대서, 김화종*
강원대학교, *강원대학교
gentlevento@naver.com, *hjkim3@gmail.com

A Study on Reduction of Data Through Baseball Data Analysis

Park Dae Seo, Kim Hwa Jong*
Kangwon Univ., *Kangwon Univ.

요 약

최근 빅데이터는 다양한 분야에서 마케팅 전략, 경제 예측 등에 활용되고 있으며 그에 따라 더 많은 데이터를 수집하고 분석하려는 노력이 많이 나타나고 있다. 빅데이터는 측정, 기록 외에는 공유를 통해서 수집하는 경우가 많이 이루어지고 있으나 큰 용량으로 인해 공유에 많은 시간이 소요되며 다양한 정보를 포함하고 있기 때문에 어떤 값들이 포함되어 있는지 파악하기 쉽지 않다. 이러한 문제를 해결하기 위한 방안으로 일부 필드 데이터를 통해 전체 필드 데이터를 예측하고 설명할 수 있다면 일부데이터만을 공유함으로써 빅데이터를 축소하고 보다 빠르게 값들을 파악할 수 있을 것으로 예상된다. 이를 위해 본 논문에서는 국내 프로야구 데이터를 활용할 것이며 데이터 내에서 일부 값들을 선택하고 해당 값들을 통해서 선택되지 않은 값들을 예측할 것이다. 또한, 예측된 값이 실제 값과 어느 정도 일치하는지 계산해보고 성공적으로 예측된 값들은 원시 데이터로부터 제외하는 방법을 통해 빅데이터 축소 방안을 제시한다.

Ⅰ. 서 론

현재 빅데이터는 시간이 지날수록 지속적으로 증가하고 있으며 이것은 기존 속성들에 대해 신규 데이터가 기록되는 것은 물론이고 새로운 속성이 추가됨으로써 하나의 빅데이터가 과거에 비해 아주 큰 크기를 가지게 되었음을 의미한다.[1] 또한, 한눈에 파악하기 어려울 정도의 많은 속성들을 가지게 되었다. 다양한 속성과 데이터는 빅데이터를 전문적으로 분석하는 전문가, 기업 등에서는 유용할 수 있지만 개인이나 숙련도가 낮은 빅데이터 분석가에게는 불필요하거나 데이터를 이해하는데 어려움을 초래할 수 있으며, 빅데이터 공유 활성에 있어서도 장애물이 될 수 있다. 이를 해결하기 위해 빅데이터 내에서 제거할 수 있는 속성들을 판단하여 빅데이터 크기를 줄이는 방법이 필요하며 적절한 기준에 의해 제거할 속성을 판단하기 위해 빅데이터 분석 기법과 분석 도구가 활용되어진다.

빅데이터는 경제, 마케팅, 스포츠 등 많은 분야에 걸쳐 분석에 활용되고 있으며 그 중 스포츠 데이터는 공개·공유하는 사이트가 많아 쉽게 습득하고 이용할 수 있다.[2] 또한, 경기를 분석하는 구단, 기관뿐만 아니라 관심이 있는 개인들도 많이 찾아보고 분석해보기 때문에 활용도가 높다. 그 중 야구는 기록의 스포츠로 불리며 다양한 기록들이 쌓이고 타석, 타수, 안타 등의 기본 기록들로부터 타율, 출루 등 새로운 기록들을 생성하고 있다.

본 논문에서는 국내 프로 야구데이터를 활용하며 선형회귀분석을 통해 독립변수들로 종속변수를 예측하는 모델을 생성하고 모델로부터 생성된 예측값과 실제값간의 차이를 계산한다.[3] 예측값과 실제값간의 차이가 작을 수록 정확하게 예측 할 수 있음을 의미하기 때문에 차이가 작다면 원시 데이터에서 해당 종속변수 데이터를 제거할 수 있다. 이와 같은 방법을 통해서 빅데이터의 크기를 축소하는 방안에 대하여 기술한다.

Ⅱ. 본론

본 연구의 진행을 위해 2015년 국내 프로 야구 타자 기록 데이터를 데이터 제공사이트로부터 수집하였다. 해당 데이터에서 전처리를 통해 기록이 없는 선수, 기록이 부실한 선수들을 제거하였으며 그림1은 야구 데이터에 포함된 변수를 나타낸다.

```
$ 순   : int  1 2 3 4 5 6 7 8
$ 이름: Factor w/ 274 levels
$ 연도: int  2015 2015 2015
$ 팀   : Factor w/ 18 levels
$ P    : Factor w/ 10 levels
$ 타석: int  595 622 643 566
$ 타수: int  472 528 534 497
$ 득점: int  130 129 126 76
$ 안타: int  180 181 153 138
$ 이타: int  42 35 19 41 42
$ 삼타: int  5 1 1 4 1 0 0 5
$ 홈런: int  47 53 48 11 23
$ 루타: int  373 377 318 220
$ 타점: int  140 146 137 56
$ 도루: int  40 10 22 25 3
$ 도실: int  8 3 6 10 3 0
$ 볼넷: int  103 78 93 59
$ 사구: int  13 12 6 4 2 1
$ 고4 : int  11 6 8 1 2 0
$ 삼진: int  91 161 72 121
$ 병살: int  7 10 10 2 13
$ 희타: int  0 0 1 3 1 1 0
$ 희비: int  7 4 9 3 8 8 9
$ 타율: num  0.381 0.343 0.
$ 출루: num  0.498 0.436 0.
$ 장타: num  0.79 0.714 0.
$ OPS : num  1.288 1.15 0.
$ wOBA: num  0.53 0.481 0.
$ wRC : num  222 182 143 1
$ WAR : num  10.71 7.76 6.
$ WPA : num  7.02 7.82 3.2
```

그림 1. 2015년 국내 프로야구 타자 기록 변수

데이터는 총 31개의 변수로 구성되어 있으며 상위 5개 변수를 제외한 26개의 변수를 이용해 분석을 수행한다. 분석기법을 사용하지 않고 단순하게 26개의 변수 중 일부를 제거 할 수도 있다. 26개의 변수 중 타율, 출루, 장타, OPS, wOBA는 비율값으로 타석 ~ 희비의 변수 중 일부를 사용하여 공식에 대입해 직접 계산할 수 있기 때문에 본 데이터에서 제거해도 무방하다. 이 경우에는 5개의 변수를 제거하여 총 26개 중 21개로 구성된 데이터로 축소 할 수 있다. wRC와 WAR의 경우 공식은 존재하지만 본 데이터의 변수값만으로는 계산할 수 없다. 따라서 이러한 경우 선형회귀분석을 이용해 값을 예측하고 실제값과의 오차에 따라서 종속 변수를 제거할 것인지를 판단한다. 본 연구에서는 선형회귀분석을 통해 wRC, WAR,

WPA 3개의 종속 변수와 타석 ~ 희비까지 18개의 독립변수를 이용해 모델을 생성하고 모델의 결정계수를 이용해 모델을 평가하고 MSE(Mean Square Error)를 통해 예측값의 정확도를 판단한다.[4] 결정계수는 0과 1 사이의 값을 가지며 결정계수가 1인 경우 추정된 회귀선이 변수 사이의 관계를 완전히 설명해 주고 있음을 의미한다.

wRC, WAR, WPA 각각의 값과 독립변수를 통해 생성된 모델의 결정계수와 모델을 통해 예측한 값과 실제값의 차이를 나타내는 MSE의 값은 표 1과 같은 결과를 얻는다.

표 1. 선형회귀분석 결과

종속 변수	모델 (결정계수)	MSE
wRC	0.9928	35.54095
WAR	0.9529	0.1316345
WPA	0.8191	0.3867141

각 종속변수에 대한 세 모델의 결정계수는 wRC, WAR의 경우 아주 높게 나타나고 있으며 모델이 주어진 자료에 적합하다는 것을 알 수 있다. WPA의 경우에는 상대적으로 낮기는 하지만 0.82로 상당히 높은 값이라 볼 수 있다. MSE는 wRC, WAR, WPA의 제거 여부를 판단하는 좋은 값으로 wRC와 WPA는 실제 데이터를 두고 봤을 때 오차가 아주 크지는 않지만 제거를 결정하기에는 어려움이 있다.

반면, WAR의 경우 오차가 작기 때문에 원시 데이터로부터 WAR의 값을 제거해도 독립변수들을 이용해 유사하게 나타낼 수 있을 것으로 기대된다. 표2는 임의 추출한 선수들의 실제 WAR과 예측 WAR을 비교하여 나타내고 있다. 예측-실제은 두값의 차이를 계산한 값으로 순서 72에서 0.45로 가장 크게 나타나고 있으며 대부분 0.1이내에서 차이를 보이고 있다. 이를 통해 MSE의 값이 거짓이 아님을 알 수 있다.

표 2. 임의 추출한 선수들의 실제 WAR과 예측 WAR

순서	이름	WAR	예측WAR	예측-실제
137	조중근	-0.14	-0.23	0.09
1	테임즈	10.71	10.9	0.19
156	조정원	-0.02	-0.08	0.06
72	이대형	2.18	2.63	0.45
160	박성준	-0.01	-0.05	0.04
214	모상기	-0.08	-0.09	0.01
36	채태인	1.99	1.89	0.1
238	이종환	-0.08	0.05	0.13
229	안정광	-0.31	-0.36	0.05
145	최병연	-0.01	-0.05	0.04
265	손주인	-0.55	-0.55	0
216	김재유	-0.27	-0.08	0.19
174	최경철	-0.23	-0.13	0.1
269	문선재	-0.24	0.11	0.35

선형회귀 분석을 통해서 wRC, WAR, WPA의 값을 예측해 보았으며 wRC와 WPA의 경우 독립변수들로 예측한 결과의 정확도가 다소 떨어지기 때문에 제거하기 여렵다고 결론 내렸으며, WAR은 예측된 데이터와 실제 데이터가 상당히 일치함을 MSE를 통해 예상할 수 있었으며 실제 WAR과 예측 WAR의 데이터를 비교해봄으로써 아주 유사하게 나타나는 것을 확인하였다. 이로써, WAR은 원시 데이터에서 제거 할 수 있으며 추후 필요에 따라 선형회귀분석과 독립변수들을 통해서 WAR을 예측 할 수 있다.

Ⅲ. 결론

본 논문에서는 빅데이터를 축소하는 방안을 제시하기 위하여 국내 프로야구 타자 데이터를 사용하여 회귀분석을 수행하여 독립변수로부터 종속변수를 예측하였다. 예측의 정확도가 높다면 원시 데이터로부터 제거하여도 독립변수를 통해 예측하여 원본에 근접한 데이터를 얻을 수 있다.

이를 통해 원시 데이터에서 예측 가능한 변수를 제거함으로써 원시 데이터의 크기를 축소 할 수 있을 것으로 기대된다. 기존에 많이 소개된 랜덤추출, 계통추출등의 샘플링 기법과는 다르게 예측의 개념을 이용해 새로운 방식의 데이터 축소 방법을 제시하였다.

앞으로의 빅데이터는 더 많은 데이터가 쌓여 크기가 기하급수적으로 증가할 것이며, 방대한 크기로 인해 공유에 많은 시간이 소요되고 트래픽 문제도 심화 될 것으로 예상된다. 본 논문에서 제시한 데이터 축소 방안은 시간, 트래픽, 고용량문제를 해소할 수 있으며 빅데이터를 공유하거나 거래하는 플랫폼에 적용되어 빅데이터 공유・거래 시장에서 기여할 것으로 기대된다.

본 연구에서는 야구 데이터만을 이용해 분석을 수행하였기 때문에 다른 빅데이터에 본 연구의 내용을 적용할 수 있는지 판단하기가 어렵다. 따라서, 추후 연구에서는 연구 데이터를 여러 분야로 확장하여 예측을 통한 데이터 축소 방법을 검증할 필요가 있으며 본 연구에서의 부족한 결과를 보충하기 위해 야구 데이터에서도 변수간의 관계를 더 깊게 파악하여 제거할 수 있는 변수를 찾아야 할 것이다. 또한, 회귀분석뿐만 아니라 다양한 분석기법을 통해서 예측에 국한되지 않고 데이터를 축소할 수 있는 방안을 연구할 것이다.

ACKNOWLEDGMENT

본 연구는 한국정보화진흥원(NIA)의 미래네트워크 선도시험망 (KOREN) 사업 지원과제의 연구결과로 수행되었음 (16-951-00-001).

참 고 문 헌

[1] 이진형, "데이터 빅뱅, 빅 데이터(BIG DATA)의 동향 ", KCA 동향과 전망, 2012, (http://www.kca.kr/open_content).

[2] STATIZ, (http://www.statiz.co.kr).

[3] 심수민, "ICT로 실현되는 야구의 새로운 즐거움" , Issue&Trend, 2013, (http://www.digieco.co.kr/KTFront).

[4] 박선일, 오태호. "상관성과 단순선형회귀분석" 한국임상수의학회지, 제27권 제4호, pp.427-434, Aug 2010.

[5] Mean Squared Error(MSE), (https://en.wikipedia.org/wiki/Mean_squared_error).