머신러닝을 이용한 한국프로야구 관중 수 예측모델

서원빈 · 길이만

성균관대학교 소프트웨어대학

Prediction Model of the Number of Spectators in Korean Baseball League Using Machine Learning

WonBin Seo · RheeMan Kil

College of Software, Sungkyunkwan University

E-mail : devin7@skku.edu, rmkil@skku.edu

요 약

본 연구는 기존 관중 수 예측에 주로 사용되는 ARIMA 모형과 다른 GKFN(Network with Gaussian kernel functions) 모델을 시계열 모델로 제안하고 여러 변수 간의 상관관계를 분석한 MLP(Multilayer Perceptron) 모델을 각각 따로 만들어 두 가지 RMSE값의 가중치를 결합한 새로운 모델을 최종적으로 제안한다. GKFN 모델은 phase space 분석을 위해 smoothness measure를 측정하고 커널 개수를 늘려가며 학습시키는 방법이다. 또한, MLP 모델은 관중 수에 영향을 주는 여러 변수(날짜, 날씨 등 팀과 관련된 특징들)의 상관관계를 correlation coefficient 값을 이용해 분석하고 높은 상관관계를 가지는 변수들을 이용해 MLP 모델을 만들어 학습하는 것이다. 이를 통해 프로야구 팀 기아 타이거즈의 일일 단위 관중 수를 예측하고자 하였다. 관중 수 예측을 통해 구단과 관객 모두 긍정적인 활용이 가능할 것이다. 훈련 자료는 2010년부터 2018년까지 9년 동안 기아 타이거즈의 일별 관중 수를 자료로 하였다.

I. 서 론

한국프로야구는 국내 인기 스포츠로서 2016년 833만 명, 2017년 840만 명, 2018년 807만 명에 이르기까지 최근 3년간 800만 관중을 돌파하는 등 관중 신기록을 경신하며 성장하고 있다. 각 팀마다 지역연고제를 바탕으로 지역통합이란 사회적 역할도 수행하고 있으며 관람수익, 굿즈 판매, 각종 광고 등을 바탕으로 많은 수익을 내고 있다. 또한 각 팀들이 경기장으로 쓰는 홈구장은 각 지역의 랜드마크로 자리 잡아서 주변 상권 활성화, 관광객 유치 등 지역경제에 많은 도움이 되고 있다. 이처럼 한국 프로스포츠에서 관중 수는 매우 중요한 요소이다. 그래서 관중 수를 미리 예측할 수 있다면 구단들에게 큰 도움이 될 것이다. 관중 수가 많을 것으로 예측되는 날에 이벤트를 여는 등 활용 가능할 것이다.

그래서 이번 연구에서 관중 수를 예측하는 모델을 만들어 보고자 한다. 시계열 모형을 이용하여 년 단위 관중 수를 예측하는 모델은 많이 제시되었지만 본 연구에서는 일일 단위 관중 수를 예측하는 모형을 만들어 보려고 한다. 년 단위 관중 수 예측보다 일일 단위 예측이 훨씬 많은 활용가치가 있을 거라고 생각된다. 먼저 시계열 조건을 고려해 예측 모델을 만들 것이다. 기존 시계열 분석 방법과 차별화된 GKFN 모델을 이용해 커널을 이용한 분석 모델이 될 것이다. 이 외에도 관중 수에 영향을 끼치는 여러 변수를 고려한 모델도 있다. 관중 수는 요일별로 차이가 생길 것이고 같은 요일이라도 상대 팀이 누구인지, 그 날 날씨가 어떤지, 분석 당시 팀의 환경에 따라서도 달라질 수 있다. 이렇듯 여러 변수 간의 관계를 이용해 MLP 모델을 만들 것이다. 이를 GKFN 모델과 결합해서 최적의 결과를 갖는 모델을 만드는 것이 최종 목표이다.

II. 관련 연구

2.1 시계열 모델

본 연구와 관련하여 관중 수를 예측하는 연구를 살펴보면 지난 수년간 관중 수 자료를 이용해 ARIMA 모형을 만들어 년 단위 관중 수를 예측하는 것이다. 이 방법은 시계열 자료의 자기상관특성을 이용하는 방식으로 일정한 주기를 가지는 년 단위 모델에 적합하다[1]. 그래서 일일 단위 예측에서 주기를 이용할 수 없기 때문에 ARIMA 모델은 유용하게 쓰일 수가 없었다.

그래서 시계열 방법을 적용하기에 커널을 이용한 학습을 이용하기로 한다[2][3]. 먼저 시계열 자료의 경향성을 분석하기 위해 phase space 분석 단계에서

smoothness measure를 구하고 그것을 이용해 어느 위치에 커널을 위치시킬지 결정해 며칠 주기의 자료를 몇 개를 이용하여야 다음 자료를 예측할 수 있는지 알게 되는 것이다. 그리고 전체 자료에서 커널 개수를 조절해가며 커널 함수를 만들어 학습 모델을 완성시킨다.

2.2 변수 모델

여러 관중 수 예측 연구에서는 구단 별로 년 단위 관중 수를 주로 예측하였다. 하지만 일일 단위 관중 수를 예측한 몇몇 연구를 참고하여 예측 모델을 만들어보려고 한다. 일일 단위 예측을 한 연구들은 주로 관중 수에 영향을 미치는 변수들을 기반으로 인공신경망 모형을 주로 만들었다[4]. 입력 변수로는 날짜. 요일, 휴일, 성수기, 온도, 습도, 팀순위, 상대팀순위 등이 있었다[5]. 이 외에도 사회요소로 댓글 수, 투표 수 같은 외부 요인들을 추가한 연구도 있었다[6]. 여러 변수들을 이용해 인공신경망 모형을 만들어서 훈련을 시켰다. 자료를 학습시키기 위한 모델로는 DNN이나 전방향 인공신경망 모델이 있었다. DNN은 입력층과 출력층 사이 여러 개의 은닉층으로 이루어진 인공신경망으로 복잡한 비선형 관계를 모델링하기 좋다.

III. 데이터 수집

1000개 이상의 자료가 요구되어 KBO 홈페이지에서 기아 타이거즈의 2010년~2018년까지의 9년도 경기 관중 수 자료를 수집하였다. 최소 649부터 최대 28500까지의 넓은 범위의 데이터 값을 re-scaling하기 위해 정규화 과정을 거쳤다. 정규화의 목적은 범위가 너무 큰 데이터 값을 이용하면 노이즈가 들어가거나 overfitting될 확률이 높기 때문에 정규화를 통해 범위를 줄여주는 데 목적이 있다. min-max 정규화를 통해 데이터 값은 0부터 1까지의 값으로 정규화 되게 된다.

$$z = \frac{x - \min(x)}{\max(x) - \min(x)}$$

그림 1 min-max 정규화

선행 연구를 참조하여 관중 수에 영향을 미치는 주요 변수들을 선정하여 이들과 관중 수 간의 관계를 수치적으로 보기 위해 상관 계수를 구하는데 이용한다. 각 변수 별로 각각의 날짜에 해당하는 데이터를 기상청과 KBO 홈페이지에서 수집한다.

도표 1 입출력 변수

	변수	
입력	시간	요일, 월, 연도
		공휴일, 성수기
		경기시간
	날씨	온도
		강수량, 상대습도
	기록	팀 순위
		상대팀 순위
출력	관중 수	

IV. 데이터 분석

4.1 Smoothness measure

$$X_{T+P} = f(x(t), x(t-tau), ---, x(t-(E-1)tau))$$

함수 f를 통해 가장 최적의 자료를 예측할 수 있는 E와 tau값을 구하게 된다. 이를 통해 데이터의 반복 정도, 즉 경향성을 파악할 수 있다. 예를 들어 E가 3이고 tau가 4라면 x, x+4, x+8 처럼 8일 동안의 3개의 데이터를 가지고 예측하는 게 가장 근접한 예측값을 가질 수 있다는 뜻이다. smoothness measure를 구하는 프로그램을 통해 관중 수 데이터에서 sm값이 처음으로 양수가 되는 적절한 E=7과 tau=4를 결정한다.

E: 7.000000, tau: 2.000000, sm: -0.025078
E: 7.000000, tau: 3.000000, sm: -0.080925
E: 7.000000, tau: 4.000000, sm: 0.001418

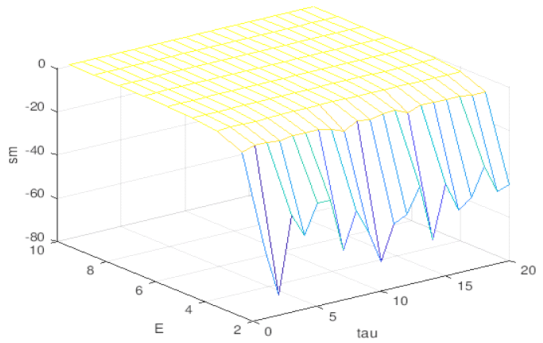

그림 2 smoothness measure 3차원 그래프

4.2 Correlation coefficient

$$\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

그림 3 Correlation coefficient

수집한 변수를 가지고 x를 변수의 값, y를 관중 수로 두고 상관계수 공식을 이용하여 correlation coefficient 값을 구한다. 각각의 변수 별로 구해진 값을 토대로 숫자가 높을수록 상관 관계가 크다는 뜻이므로 큰 순서대로 우선순위를 정해 상위 5개의 변수를 선정하도록 한다. 선정된 변수는 공휴일 여부, 요일, 경기시작시간, 팀순위, 상대습도 5가지이다. 예를 들어 평일과 공휴일은 관중 수에서 엄청난 차이를 보이게 된다. 공휴일 여부를 이항변수로 설정하여 경기가 열리는 날이 평일이라면 0으로 판단하고 공휴일이라면 1로 판단하여 모델링을 하게 된다. 또한, 상대습도 같은 경우 상대습도가 높을수록 관중 수가 줄어드는 추세를 보이게 된다. 이렇게 선정된 변수들을 이용해 MLP model을 생성할 것이다.

V. GKFN 모델

가지고 있는 데이터를 training, test 데이터로 나누어 전체 데이터의 90%를 training 데이터로 프로그램을 학습시킨다. 학습된 프로그램에 나머지 test 데이터를 이용하여 비교한 다음 오차를 따져 RMSE값을 찾을 수 있다. 여기서 사용되는 학습 방법은 커널을 이용한 방법을 따른다. 커널 각각에서 벡터를 만들어 앞에서 도출한 E와 tau값을 이용해 최적의 커널 위치에서 커널을 만들어가며 데이터값을 학습시키는 것이다. 커널 개수를 늘려가며 학습시킨다면 training error는 계속하여 떨어지게 된다. 하지만 test error는 어느 순간 높아지게 되는데 이 때 최적의 test error를 가지는 커널 개수를 토대로 RMSE값을 구할 수 있다. 오차값을 비교하기 위해서는 RMSE(root mean square error)값을 이용한다. 실제로 해보면 커널 개수를 10개부터 늘려가면 training error와 test error가 모두 낮아지지만 어느 순간 커널이 50개를 넘어가는 지점부터 test error는 상승하기 시작한다. 프로그램 결과 커널 50개에서 최적의 RMSE값 0.202421이 나왔다.

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i-\hat{y})^2}$$

그림 4 RMSE

Phase 3 step rmse = 0.211599, rsq = 0.253392
Phase 3 step rmse = 0.211201, rsq = 0.256197
Phase 3 step rmse = 0.211063, rsq = 0.257174
rmse: 0.202421, R2: 0.101872

VI. MLP 모델

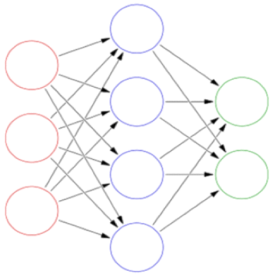

그림 5 MLP 모델

위에서 선정한 변수들로 MLP 모델을 생성한다. 각 층마다 node들로 구성되어 원 모양의 노드가 뉴런을 나타내고 화살표는 다른 뉴런으로의 입력을 의미한다. 첫 번째 층에 입력된 데이터를 다음 층으로 전달하는 과정을 반복하여 마지막 층에서 최종 출력이 나온다. 추가로 각 뉴런에서 정의된 조건에 따라 가중치가 할당되어 가중치가 더해져서 최종 출력이 나온다. 입력 노드는 변수 5개가 되고 은닉층은 256개의 노드를 가진다. 이 은닉층을 지나 최종 출력 1개의 예측값이 나오는 것이다. Python 라이브러리 중 sci-kit learn의 MLPRegressor를 이용해 MLP 모델을 설계한다. 정규화를 한 관중 수 데이터와 선정된 변수별 데이터를 train data와 test data로 나누어 MLP 모델의 노드로 입력하고 출력 노드에서 나오는 예측값과 오차를 비교한다. 이 때 RMSE 값은 0.19가 나온다.

VII. 모델 결합

GKFN, MLP 모델 각각에서 얻어진 RMSE값을 토대로 두 가지 모델을 결합하여 최적의 RMSE값을 갖도록 한다. 각각의 모델에 가중치를 다르게 하여 최적의 값을 갖도록 결합시킨다. 두 가지 모델이 결합한다면 가장 최적의 RMSE값을 갖는 예측 모델이 완성된다.

도표 2 변수별 correlation coefficient 결과

변수	Correlation coefficient	순위	비고
요일	0.346028	2	월(0) -- 일(6)
월	-0.109825	7	
연도	0.112621	6	
공휴일	0.368540	1	0, 1
성수기	-0.079542	10	0, 1(7,8월)
온도	-0.066107	11	
강수량	-0.103444	8	
상대습도	-0.125308	5	
팀순위	-0.144831	4	
상대팀순위	-0.083058	9	
경기시작시간	-0.312914	3	

VIII. 결론

관중 수 예측 모델은 시계열 분석과 변수 분석 모델을 결합시켜 성능을 향상시켰다. 먼저 시계열 분석 모델은 선행연구를 통해 phase space 분석을 위한 smoothness measure를 구했다. 앞의 결과에서 E=7, tau=4를 얻게 된다. 즉, 관중 수 자료를 4일 주기로 24일 씩의 자료를 통해 다음 자료를 예측할 수 있다는 것이다. 이 자료를 토대로 GKFN 프로그램을 이용해 RMSE값을 찾는다. GKFN 프로그램에서 RMSE는 0.202421을 얻었다. 그리고 MLP 모델을 위해 관중 수에 영향을 미치는 변수를 가지고 관련 자료를 수집하였다. 각 변수 별로 변수와 관중 수의 상관관계를 상관계수를 이용해 수치 비교를 한 후 주요 변수들을 선정하여 활용하였다. 주요 변수로는 공휴일여부, 요일, 경기시작시간, 팀순위, 상대습도가 있다. 이 변수들을 가지고 MLP regression 모델을 생성한 후 RMSE는 0.19를 얻었다. 두 가지 모델의 결합을 통하여 최종 예측 모델을 생성하였다. 최종 예측 모델은 두 개 각각의 모델보다 더 좋은 정확성을 가지고 있음을 보였다. 본 연구는 선행 연구에서 부족했던 일일 단위 관중 수 연구를 수행하고 변수 모델을 시계열 모델과 결합하여 최적의 결과를 얻는 데에 의미가 있다. 향후 이것을 활용하여 여러 분야에서 응용이 가능할 것이라고 기대된다.

References

[1] Jinseok Chae, “Prediction Model for Korean Professional Baseball Spectators,” Korean Journal of Sport Science, Vol.23, No.4, pp.892-905, 2012

[2] Dong Kyu Kim and Rhee Man Kil, “Stock Price Prediction Based on a Network with Gaussian Kernel Functions”, College of Information and Communication Engineering

[3] Rhee M. Kil, “Function Approximation Based on a Network with Kernel Functions of Bounds and Locality : an Approach of Non-Parametric Estimation”

[4] Seunghoon Jeong, "Professional Baseball Spectator's Analysis and Prediction by Using Artificial Neural Networks Model and Logistic Regression Model", Korean Journal of Sport Science, Vol.26, No.1, pp.104~121, 2015

[5] Dongju Park, Byeongwoo Kim, Youngseon Jeong, Changwook Ahn, "Neural Network Based Prediction of Daily Spectators for Korean Baseball League : Focused on Gwangju-KIA Champions Field" Smart Media Journal, Vol.7, No.1, ISSN:2287-1322, 2018

[6] Jinuk Park, Sanghyun Park, "A Study on Prediction of Attendance in Korean Baseball League Using Artificial Neural Network", KIPS Tr. Software and Data Eng. Vol.6, No.12, pp.565~572, 2017