Smart Media Journal / Vol.7, No.1 / ISSN:2287-1322
http://dx.doi.org/10.30693/SMJ.2018.7.1.16

Deep Neural Network 기반 프로야구 일일 관중 수 예측 : 광주-기아 챔피언스 필드를 중심으로

(Deep Neural Network Based Prediction of Daily Spectators for Korean Baseball League : Focused on Gwangju-KIA Champions Field)

박동주*, 김병우**, 정영선***, 안창욱****

(Dong Ju Park, Byeong Woo Kim, Young-Seon Jeong, Chang Wook Ahn)

요 약

본 연구는 Deep Neural Network(DNN)을 이용하여 광주-기아 챔피언스 필드의 일일 관중 수를 예측함으로써 이를 통해 구단과 관련기업의 마케팅 자료제공 및 구장 내 부대시설의 재고관리에 자료로 쓰임을 목적으로 수행 되었다. 본 연구에서는 Artificial Neural Network(ANN)의 종류인 DNN 모델을 이용하였으며 DNN 모델의 과적합을 막기 위해 Dropout과 Batch normalization 적용한 모델을 바탕으로 총 4종류를 설계하였다. 각각 10개의 DNN을 만들어 예측값의 Root Mean Square Error(RMSE)와 Mean Absolute Percentage Error(MAPE)의 평균값을 낸 모델과 예측값의 평균으로 RMSE와 MAPE를 평가한 Ensemble 모델을 만들었다. 모델의 학습 데이터는 2008년부터 2017년까지의 관중 수 데이터를 수집하여 수집된 데이터의 80%를 무작위로 선정하였으며, 나머지 20%는 테스트 데이터로 사용하였다. 총 100회의 데이터 선정, 모델구성 그리고 학습 및 예측을 한 결과 Ensemble 모델은 DNN 모델의 예측력이 가장 우수하게 나왔으며, 다중선형회귀 모델 대비 RMSE는 15.17%, MAPE는 14.34% 높은 예측력을 보이고 있다.

■ 중심어 : 한국프로야구 ; DNN ; 머신러닝 ; 수요예측

Abstract

In this paper, we used the Deep Neural Network (DNN) to predict the number of daily spectators of Gwangju - KIA Champions Field in order to provide marketing data for the team and related businesses and for managing the inventories of the facilities in the stadium. In this study, the DNN model, which is based on an artificial neural network (ANN), was used, and four kinds of DNN model were designed along with dropout and batch normalization model to prevent overfitting. Each of four models consists of 10 DNNs, and we added extra models with ensemble model. Each model was evaluated by Root Mean Square Error (RMSE) and Mean Absolute Percentage Error (MAPE). The learning data from the model randomly selected 80% of the collected data from 2008 to 2017, and the other 20% were used as test data. With the result of 100 data selection, model configuration, and learning and prediction, we concluded that the predictive power of the DNN model with ensemble model is the best, and RMSE and MAPE are 15.17% and 14.34% higher, correspondingly, than the prediction value of the multiple linear regression model.

■ keywords : Korean Baseball League ; Deep Neural Network ; Machine Learning ; Demand Forecasting

Ⅰ. 서 론

한국프로야구 관중 수는 1982년 1,438,768명으로 시작하여 2017년 8,400,688명으로 성장하면서 36년간 총 146,859,897명을 동원하여 한국프로야구는 국민스포츠로 자리 잡았다.

프로야구는 프로스포츠 시장에 가장 먼저 진입한 장점을 활용하여 프로스포츠 시장 내에서 시장점유율이 가장 큰 종목으로 성장 해왔다[1]. 한국프로야구는 연고를 바탕으로 사회적 역할인 지역통합의 기능도 수행하고 있으며, 지속적인 관람객 증가와 다양한 굿즈 판매 및 프로모션을 통해 수익을 창출하고 있다. 이를 바탕으로 한국프로야구는 리그 수준을 발전시킴과 동시에 국내 대표 프로스포츠시장으로써의 입지를 확장해왔다.

프로스포츠의 가장 큰 자산은 관람객이라 할 수 있다. 관람객이란 관람스포츠 경기장을 직접 찾는 사람들로, 관람객이 많은 구단은 인기구단으로 인정받는다고 할 수 있다. 국내 프로스포츠 또한 연간 관중수가 구단의 가치에 중요한 척도로 인정받고 있다[2]. 프로스포츠에서 관중의 증가는 구단의 존속과 성장을 위해서는 절대적인 과제로서 관중이 외면한 프로스포츠는 존재할 수도 없을 뿐 더러 TV중계, 광고료, 스폰서쉽, 라이센싱, 구

* 준회원, 광주과학기술원 전기전자컴퓨터공학부
** 준회원, 전남대학교 산업공학과
*** 정회원, 전남대학교 산업공학과
**** 종신회원, 광주과학기술원 전기전자컴퓨터공학부
이 성과는 2017년도 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. NRF-2015R1D1A1A02062017).

접수일자 : 2018년 03월 05일
수정일자 : 1차 2018년 03월 21일
게재확정일 : 2018년 03월 27일
교신저자 : 정영선 e-mail : young.jeong@chonnam.ac.kr
공동교신저자 : 안창욱 e-mail : cwan@gist.ac.kr

장 이익금 등을 통한 재정수입 또한 기대하기 어렵다[3]. 이처럼 프로스포츠에서 관중 수는 구단의 직접적인 수익뿐만 아니라 구단의 가치를 평가할 수 있는 지표가 될 수 있으며 결과적으로 구단의 존속 및 발전으로 이어진다.

KIA 타이거즈는 1982년 한국프로야구 출범과 함께 광주, 전라도를 연고로 해태 타이거즈로 창단하였으며 2017년 한국시리즈 우승을 차지하여 2017년 현재까지 11회 우승으로 한국프로야구 최다 우승 기록을 보유하고 있다. KIA 타이거즈는 1982년부터 2013년 까지는 광주 무등경기장 야구장을 홈구장으로 사용하였고 2014년부터 현재까지는 광주-기아 챔피언스 필드를 홈구장으로 사용하고 있다.

KIA 타이거즈는 2013년 이후 관중수가 꾸준히 증가하는 추세를 보였으며, 2017년 한국 시리즈를 우승하여 앞으로도 관중수가 증가할 것으로 판단된다. 더불어 2014년 광주-기아 챔피언스 필드로 홈구장을 이전 한 후 전년도에 비해 관중 수가 41% 상승하여 KIA 타이거즈를 제외한 다른 구단들의 평균증가율(0.3%)에 비하여 큰 폭으로 증가함을 보였다. 또한 광주-기아 챔피언스 필드는 구장 안정화 및 관중 유치를 위해 매년 리모델링을 하고 있다. 2016년의 경우 부대시설 설치, 테이블석 확충 및 백스톱석 고급화로 인해 22,244명이었던 관중석을 20,500명으로 줄이기도 하였다. 이처럼 광주-기아 챔피언스 필드 관중 수는 꾸준히 증가하고 있으며 앞으로도 증가할 잠재가능성을 가지고 있다. 광주-기아 챔피언스 필드는 타 구단의 구장에 비하여 사용한 기간이 짧아 관중수 예측이 어려우며, 구장 리모델링에 의해 최대수용인원인 27000명까지 탄력적으로 운영될 수 있기에 광주-기아 챔피언스필드를 일일 관중 수 예측 대상으로 선정하였다.

본 연구에서는 2008년부터 2017년까지 광주 무등경기장 야구장 및 광주-기아 챔피언스 필드의 일별 관중 수 및 다양한 변수들을 바탕으로 Deep Neural Network 기반 광주-기아 챔피언스 필드의 일일 관중 수를 예측하는 모델을 제안하려고 한다. 연간 누적 관중 수가 아닌 일일 관중 수를 예측하는 것은 구장 내 이벤트, 홍보, 팬 프로모션 등 구단 및 관련기업의 마케팅 자료로 이용될 수 있으며 구장 내 부대시설의 재고관리 측면에서 연간 누적 관중 수 예측 보다 큰 효용이 있을 것이라고 사료된다.

Ⅱ. 본 론

1. 이론적 배경

가. Deep Neural Network

보편적으로 딥러닝으로 명칭되는 심층 구조 학습은 2006년 이후 기계학습 연구 분야의 새로운 영역으로 떠올랐다[4]. Deep Neural Network(DNN)는 다양한 정의가 존재하지만 일반적으로는 입력층(Input layer)과 출력층(Output layer) 사이에 많은 수의 은닉층(Hidden layer)들로 이루어진 Artificial Neural Network(ANN)의 일종으로 비선형문제를 해결하는데 사용되는 기계학습 방법이다. ANN의 기본구조는 그림 1과 같다.

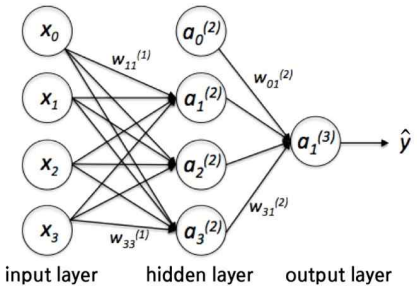

그림 1. ANN의 기본구조

ANN 이란 인간의 신경체계와 유사한 성능과 특성을 갖는 지도 / 비지도 학습을 기반으로 정보를 처리하는 기계학습 방법 중 하나이다. ANN 방식은 인간의 뇌가 대량의 데이터를 효율적, 병렬적으로 처리 및 학습할 수 있다는 사실에 근거하여, 인간의 생물학적 신경세포를 모델링하여 구현하는 방식으로서 신호처리, 제어, 패턴인식, 음성, 문자인식, 예측 등의 분야에서 널리 응용되고 있다[5][6][7].

ANN에서 은닉층을 늘린 DNN은 층마다 다른 층위의 특징을 학습할 수 있으며, 낮은 층위의 단순하고 구체적인 특징에서부터 복잡하고 추상적인 더 높은 층위의 특징을 추출함으로써 데이터의 잠재적인 구조를 파악할 수 있다. 이러한 심층신경망이 최근 급격하게 이슈화 되고 연구될 수 있었던 이유는 이미지 처리 기술능력의 발달, 컴퓨터 하드웨어의 개발에 따른 연산비용 감소, 새로운 기계학습 정보처리 기법이 있다. 심층신경망이 활발하게 연구됨에 따라 복잡하고 비선형함수로 구성된 문제를 효과적으로 해결할 수 있고, 분류되지 않은 자료에 대한 학습성능도 뛰어나기 때문에 인공지능뿐만 아니라 그래픽 모델링, 최적화, 패턴인식, 신호처리 등 다양한 분야에서 활용되고 있다[8][9][10].

그림 2와 같이 DNN에서의 층(Layer)이 깊어짐에 따라 여러 가지 문제점도 발생한다. 첫 번째로는 과적합(Overfitting)문제가 있다. 과적합이란 학습된 모델이 학습에 활용된 데이터에 대해서는 정확하게 예측하지만 학습과정에 사용되지 않은 새로운 데이터에 대해서는 예측정확도가 떨어지는 것을 말한다. 이러한 문제를 극복하기 위해 Dropout이나 Regularization과 같은 기법들을 사용하고 있다. 두 번째는 기울기 소실 문제

Smart Media Journal / Vol.7, No.1 / ISSN:2287-1322
http://dx.doi.org/10.30693/SMJ.2018.7.1.18

(Vanishing gradient problem)이다. DNN은 학습 시 오차역전파 알고리즘(Back propagation algorithm)으로 출력층의 오류를 입력층까지 전달하는데 이때 은닉층의 개수가 늘어날수록 오차가 전달되지 않는 기울기 소실 문제가 나타난다. Nair 등은 [11] 활성함수로 Rectified Linear Unit(ReLU)을 사용하여 학습 시간을 줄이고 기울기 소실 문제를 해결하였다. 이외에도 가중치(Weight)의 초기값에 따라 학습속도나 결과가 달라지는 문제도 존재하지만 Restricted Boltzmann Machine(RBM)이나 Xavier initialization 그리고 He's Initialization과 같은 방법이 등장함에 따라 많은 부분이 극복되었다.

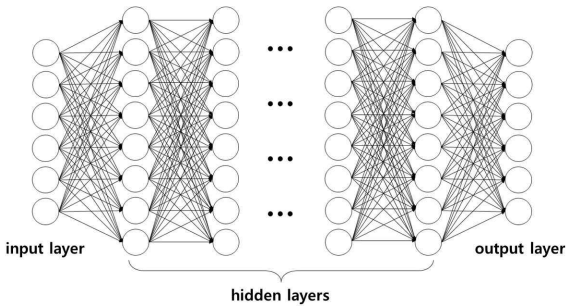

그림 2. DNN의 구조

(1) Dropout

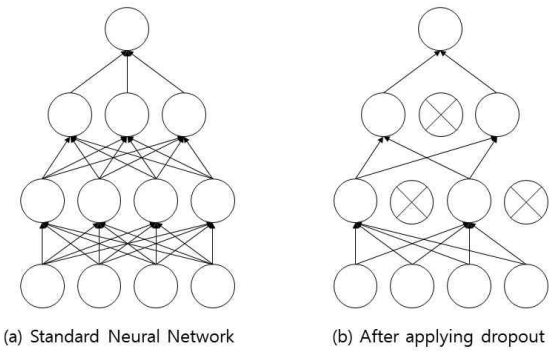

그림 3. (a) Dropout을 적용하지 않은 DNN (b) Dropout을 적용한 DNN

Dropout은 과적합 문제를 해결하기 위한 방법 중 하나이다. 그림 3의 (a)와 같은 DNN을 학습할 때 각 레이어의 모든 노드에 대해 학습하는 것이 아니라 그림(b)와 같이 은닉층의 노드를 무작위로 선택하여 학습을 하는 방법이다. 생략된 노드는 학습에 영향을 끼치지 않고, 일반적으로 50~80%정도의 노드를 사용한다[12]. 한번 학습을 하고나면 다음 학습 시에는 다시 노드를 무작위로 선택하여 학습을 하고 테스트를 할 때는 모든 노드들을 사용하여 테스트를 한다.

(2) Batch Normalization

기계학습에서 널리 쓰이는 최적화 방법은 미니 배치마다 신경망 파라미터를 갱신하는 미니 배치 Stochastic Gradient Descent (SGD)이다. Loffe와 Szegedy[13]는 SGD의 학습과정 중 신경망 내 Activation 값들의 분포가 미니 배치마다 달라 학습의 어려움을 증가시킨다고 제기한다.

이러한 문제를 해소하기 위해 각 층의 입력마다 새로운 층을 추가하고, 수행되는 작업은 식1과 같다.

$$BN(x_i) = x_i^{BN} = \gamma(\frac{x_i - E(X_B)}{\sqrt{Var(X_B)^2 + \epsilon}}) + \beta \quad (1)$$

X_B는 배치정규화 층의 입력들 (배치 단위), x_i는 X_B의 원소이다. 단순 정규화만 적용하게 되면 정확도가 저하되는 문제가 있어, 이를 보정해주는 학습 파라미터 β와 γ가 추가되었다. β와 γ의 크기는 입력과 같은 차원의 벡터이다. 배치정규화 신경망은 학습률을 기존보다 더 크게 설정해도 안정적인 학습이 가능하고, 수렴도 빠르게 이루어진다[14].

(3) Ensemble Method

Ensemble Method는 주어진 훈련데이터를 이용하여 학습모델을 구축할 때 여러 개의 모델을 구축하고 다양한 모델들을 결합하여 하나의 최종 모델을 만드는 방법이다. 다양한 모델의 예측 결과를 결합함으로써 단일의 모델보다 정확도를 향상 시키는 것이 목표이며 과적합 방지에도 도움이 된다. Ensemble Method는 여러 가지 기법이 존재하는데 대표적으로 배깅(Bagging), 범핑(Bumping), 밸런싱(Balancing), 에이다부스트(Adaboost) 등이 있다. 본 연구에서는 각 모델들의 평균으로 결과를 예측하는 배깅 방법을 이용하였다.

나. 수요예측

수요예측이란 모든 계획의 전제가 되는 것으로 불확실한 미래에 발생할 수 있는 다양한 대안 중에서 가장 발생 잠재력이 높은 대안을 예견하기 위한 의사결정 방법이다[15][16].

수요예측의 중요한 목적은 주어진 상황에서 확률적으로 최상의 수요수준을 추정하는 것이다. 수요예측은 경쟁적, 다변적인 경영환경 속에서 미래를 사전에 예견할 수 있도록 하여 경영의 사결정과 관련된 불확실성과 위험을 줄이고, 경영자에게 합리적인 의사결정을 제공하는 중요한 역할을 하고 있다[17]. 그러나 아무리 정교한 방법으로 모델을 구축하고 수요를 예측했다 하더라도 미래에 대한 불확실성으로 인해 예측치와 실제결과에는

오차가 존재하기 마련이며, 이러한 오차를 최소화 하는 것이 문제의 핵심이 된다[16].

본 연구에서는 DNN모델과 비교를 위해 인과모델 중 하나인 다중선형회귀와 시계열모델인 ARIMA모델을 사용하였다.

(1) 다중회귀분석

회귀분석은 현상과 항목들의 인과 관계에 의해 나타나는 관계를 수학적으로 설명하기 위한 통계적인 분석방법으로 원인의 역할을 하는 설명변수 즉 독립변수와 결과를 나타내는 반응변수 즉 종속변수 간의 함수관계를 규정하여 모델이 도출된다. 이때 종속변수와 하나의 독립변수 사이의 선형모델을 단순선형회귀모델이라 하고, 하나의 종속변수와 두 개 이상의 독립변수들 사이의 선형모델을 다중선형회귀모델이라고 한다. 선형회귀분석 외에 함수형태에 따라 곡선추정 회귀분석, 로지스틱 회귀분석, 자기회귀분석 등이 활용되고 있다[18]. 다중선형회귀모델은 식2와 같다.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_i x_i + \varepsilon_i \qquad (2)$$

y : 종속변수 관찰치
x_i : 독립변수 i번째 관찰치
β_i : 모형의 회귀계수
ε_i : 모형의 i번째 오차항

(2) ARIMA

통합자기회귀이동평균법(Auto-Regressive Integrated Moving Average : ARIMA)을 지칭하는 것으로, 여러 형태의 단변량 시계열 데이터를 확률과정모형 즉 AR, MR, ARMA, ARIMA 및 계절 ARIMA 에 접목시켜 효과적으로 시계열을 분석하는 것을 말한다[19]. ARIMA모델은 다른 시계열 모델에 비하여 비교적 복잡한 산술적, 통계적 과정을 거치게 되고, 때로는 사용자의 주관적 판단이 요구되기도 한다[20].

ARIMA의 일반적인 표현은 단일시계열분석일 경우 ARIMA(p, d, q), 계절변동을 포함할 시에는 ARIMA(p, d, q)(P, D, Q)$_S$ 로 표현된다. 여기에서 (p, d, q)부분은 모형의 비계절 부분을, (P, D, Q)부분은 계절 부분을, S는 계절당 기간수를 나타낸다. 또한 p와 P는 AR모형의 차수이며, q와 Q는 MA모형의 차수 그리고 d와 D는 차분차수(Degree of differencing)를 의미한다. 따라서 ARIMA과정이란 시계열의 패턴을 정확하게 반영해 줄 수 있는 차수 p, d, q 혹은 P, D, Q를 적절하게 찾아내는데 달려 있다고 할 수 있다[21]. 이러한 개념식을 수학적으로 표현하면 식3과 같다.

$$\phi(B)\Phi(B^s)(1-B^S)^D(1-B)^d\hat{y_t} = \theta(B)\Theta(B^S)e_t \qquad (3)$$
$$\hat{y_t} = y_t - \mu$$

t : 시차
y_t : 종속변수 또는 차분변수
μ : 종속변수의 평균
d : 비계절적 차분횟수
D : 계절적 차분횟수
B : 후향연산자 , $BX_t = X_{t-1}$
S : 계절당 기간수
$\phi(B)$: 비계절적 AR모형
$\phi(B) = 1 - \phi_1 B - \ldots - \phi_p B^p$
$\Phi(B)$: 계절적 AR모형
$\Phi(B) = 1 - \Phi_1 B^S - \Phi_2 B^{2S} - \ldots - \Phi_P B^{PS}$
$\theta(B)$: 비계절적 MA모형
$\theta(B) = 1 - \theta_1 B - \ldots - \theta_q B^q$
$\Theta(B)$: 계절적 MA모형
$\Theta(B) = 1 - \Theta_1 B^S - \Theta_2 B^{2S} - \ldots - \Theta_Q B^{QS}$
e_t : 오차항(백색잡음).

다. 선행연구 고찰

스포츠분야에서는 주로 관중의 재 관람 의도나 관람동기, 관람태도 및 구단의 충성도등 경기관람에 관련된 인과관계를 규명하는 연구가 다수이나 소규모로 관중수와 관련된 예측모델로 자귀회귀누적이동평균(ARIMA)모델과 일반화 자기회귀 조건부 이분산모델(GARCH)을 활용하여 프로야구와 축구를 중심으로 시계열데이터 분석 방법을 이용한 관중 수 예측에 관련된 논문이 발표 되고 있다[3][22][23].

관중 수 예측과 관련된 선행연구들은 대부분 연간 누적 관중수를 이용하여 향후 연도에 대한 관중 수를 예측하는 연구들이었다. 본 연구에서는 연간 누적 관중 수가 아닌 일일 관중 수를 수집하여 일일 관중 수를 예측을 목적으로 하고 있다.

하지만 연간 누적 관중 수가 아닌 일일 관중 수에 대해서는 선행연구에서 사용된 ARIMA 모델을 사용하는데 문제가 있다. ARIMA 모델 같은 경우 주기적인 추세와 경향이 존재해야 하는데 프로야구의 경우 모든 연도가 경기 수가 같지 않을 뿐더러 매년 월별 경기 수가 다르고 경기를 하는 요일도 달라 경기의 주기가 일정하지 않음을 알 수 있다[24]. 실제 데이터에서는 일부 자기상관이 존재 하여 ARIMA 모델과 다중선형회귀모델을 비교한 후 더 적합한 모델과 DNN 모델을 비교 하였다.

2. 연구방법 및 실증적 분석

가. 데이터수집 및 전처리

본 연구에서 사용되어지는 관중 수 및 야구경기 관련 데이터들은 KBO에서 제공하는 2009년부터 2017년 연감에 수록된 공

Smart Media Journal / Vol.7, No.1 / ISSN:2287-1322
http://dx.doi.org/10.30693/SMJ.2018.7.1.20

식 데이터를 사용하였으며 부족한 데이터는 기아타이거즈 홈페이지 및 KBO 홈페이지를 참고하여 총 2008년부터 2017년까지 10년간의 광주-기아 챔피언스 필드 및 무등경기장 야구장에서 열린 638개의 경기 데이터를 수집하였다.

표 2. 입력변수와 출력변수

변수	요소	
입력변수	시간 요소	연도**
		월**
		요일**
		공휴일*
		성수기*
		경기시간*
	기상 요소	평균기온
		일 강수량
		평균 상대 습도
	팀 요소	어웨이 팀**
		홈 팀 연승 수
		홈 팀 순위
		어웨이 팀 연승 수
		어웨이 팀 순위
	기타 요소	개막전*
		구장*
출력변수	관중 수	

*이항변수(1, 0), **범주형변수

모델에 사용될 입력변수들은 일반적으로 관중 수에 영향을 미칠 수 있다고 판단되는 요소와 연구자 본인의 주관적 판단에 의한 요소들로 시간요소, 기상요소, 팀 요소, 기타 요소로 선정하였다. 시간 요소에서 연도는 2008년부터 2017년까지, 월은 3월부터 10월까지 그리고 요일은 월요일부터 일요일로 각각 범주형 변수로 선정하였다. 공휴일변수의 경우 공휴일일 경우 1, 아닐 경우 0으로 표기하였으며, 성수기는 여름휴가와 학생들의 방학이 있는 7월과 8월을 구분하기 위해 이항변수로 생성하였고, 경기 시간의 경우 경기가 오후(1시에서 4시 사이)에 시작하는 경우 1, 아닌 경우는 0으로 표기하였다. 기상 요소의 경우 기상청의 기상자료개방포털에서 수집한 평균기온과 일 강수량, 평균 상대습도를 사용하였다. 팀 요소의 어웨이 팀은 홈 팀인 기아 타이거즈를 제외한 팀들은 범주형 변수로 선정하였고, 홈 팀 연승 수와 홈 팀 순위에는 해당일의 기준으로 전날의 기아 타이거즈의 연승 수와 순위를 입력하였으며, 어웨이 팀 연승 수와 어웨이 팀 순위도 마찬가지로 전날의 어웨이 팀의 연승 수와 순위를 입력하였다. 기타 요소로는 광주-기아 챔피언스 필드 및 무등경기장 야구장에서의 개막전을 구분하기 위해 이항변수를 추가하였고, 구장에는 광주-기아 챔피언스 필드 일 경우 1, 무등경기장 일 경우 0으로 표기하였다. 이러한 입력변수와 출력변수들은 표 2와 같다.

무등경기장 야구장과 광주-기아 챔피언스 필드는 시설 리모델링 및 증축에 따른 최대 관중수용인원이 조금씩 변동이 있었는데 일관성을 유지하기 위해 무등경기장 야구장에서 경기한 2008년부터 2013년의 최대 관중수용인원을 12500명으로 하여 초과되는 데이터는 12500명으로 변경시켰으며, 2013년 이후 데이터는 광주-기아 챔피언스 필드의 현재 최대 관중수용인원인 20500명으로 변경시켰다.

이러한 입력변수들과 출력변수는 계산의 편이를 위해 0에서 1사이로 정규화를 하기 위하여 각각의 변수에서 최대 값과 최소 값을 구하여 변수에서 최소 값을 빼고 최대 값과 최소 값의 차로 나누어 주었으며 평가기준에서 Mean absolute percentage error(MAPE)를 계산할 때는 0으로 나눠지는 것을 방지하기 위해 출력변수를 다시 역정규화 시켜서 MAPE를 계산하였다. 또한 범주형 변수의 경우는 더미변수를 생성하였다. 정규화 하는 과정은 아래 식4와 같다.

$$x_n = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \quad (4)$$

x_n : 데이터 값
$x_{\min}$: x 데이터중 최소값
$x_{\max}$: x 데이터중 최대값

나. 모델설계

본 연구를 위해 설계된 DNN 모델은 Google에서 제작한 딥러닝 라이브러리인 TensorFlow를 이용하여 설계하였으며 다중선형회귀의 경우는 Scikit learn을 이용하여 설계하였다.

DNN 모델들은 Dropout과 Batch normalization을 적용하지 않은 모델과, Dropout만 적용한 모델, Batch normalization을 적용한 모델 그리고 Dropout과 Batch normalization을 모두 적용한 모델로 총 4종류이다. Dropout 이 적용된 모델에는 Dropout 비율을 0.5로 하였다.

각 DNN은 입력층과 4개의 은닉층 그리고 출력층으로 총 6개의 층으로 구성되고 입력층은 더미변수를 포함한 입력변수들의 개수인 총 46개의 노드를 가지고 있으며 첫 번째 은닉층은 200개의 노드를, 두 번째는 150개, 세 번째와 네 번째는 각각 100개와 50개의 노드를 가지고 있으며 출력층은 관중 수 예측값인 1개의 노드를 가지고 있다. 은닉층의 노드들의 개수는 시행착오를 통해 가장 적합하다고 생각되는 모델을 선택하였다.

각 은닉층의 활성화 함수로는 Rectified Linear Unit(ReLU)의 변형형태인 Leaky ReLU를 사용하였고 Leaky ReLU의 음

수 부분의 기울기인 알파값은 0.2로 정하였다. 출력층의 출력변수는 관중 수를 예측하므로 활성화 함수로는 회귀모델을 나타낼 수 있는 선형함수를 사용하였다.

그 외 과적합을 방지하는 L2 Regularization의 Regularization의 강도를 의미하는 lambda 값은 시행착오를 통해 0.001로 하였고 마찬가지로 학습률도 0.001로 정하였다. 모델구성을 정리하면 표 3과 같다.

표 3. DNN 모델의 구성 요소 및 하이퍼 파라미터

구분	설정
입력층 노드 수	46
은닉층 갯수	4
은닉층 노드 수	200, 150, 100, 50
출력층 노드 수	1
학습률	0.001
학습 알고리즘	Adam optimization
가중치 초기화 알고리즘	He's Initialization
L2 Regularization	0.001
미니배치 수	32
은닉층 활성화 함수	Leaky ReLU
출력층 활성화 함수	Linear function
Dropout 비율	0.5
Epoch	10000

다. 연구결과 및 논의

DNN 모델을 평가하기 위해서 먼저 ARIMA 모델과 다중선형회귀모델을 서로 비교하였다. 두 모델의 모두 2008년부터 2017년까지 총 638개의 일별 관중 수 데이터 중 2008년부터 80%인 510개의 데이터를 바탕으로 모델을 만든 후 나머지 20%인 128개의 데이터로 테스트를 하였다. ARIMA 모델의 경우 R프로그램의 forecast 패키지의 auto.arima를 이용하여 가장 적합한 모델인 ARIMA(1,1,2) 모델을 찾았으며 다중선형회귀 모델의 경우 독립변수를 DNN 모델의 입력변수들로 사용하고 종속변수를 일일 관중 수로 모델을 만들었다. ARIMA 모델의 경우 Root Mean Square Error(RMSE)가 0.24833, 다중선형회귀 모델의 경우는 0.16732로 측정되었으며 그 결과 다중선형회귀 모델이 ARIMA 모델보다 더 정확한 예측력을 가진다고 판단되어 DNN 모델을 다중선형회귀 모델과 비교하였다.

DNN 모델과 다중선형회귀 모델을 비교는 ARIMA 모델과 다중선형회귀 모델의 비교와는 달리 연속적인 시계열 데이터가 아니어도 된다고 판단되어 총 638개의 데이터 세트에서 무작위로 80%인 510개의 데이터를 선정하여 학습시킨 후 나머지 20% 데이터인 128개로 테스트를 진행하였다.

데이터 세트 선정과 모델 구성, 학습 그리고 테스트에 대해서는 총 100회 실행하였고 평균을 도출하였다. DNN 모델의 경우 가중치의 초기값에 따라 모델의 성능이 달라질 수 있으므로 성능을 향상시키기 위해 한 데이터 세트에 대해 서로 다른 10개의 모델을 만들어 학습시킨 후 테스트를 하였다. 테스트 결과에서는 그림 4와 같이 10개 모델의 RMSE와 MAPE의 평균과 그림 5와 같이 10개 모델을 배깅 알고리즘을 사용한 Ensemble 모델로 구성하여 각 모델의 출력값들의 평균으로 RMSE와 MAPE를 구하였고 다중선형회귀 모델의 RMSE, MAPE와 비교하였다.

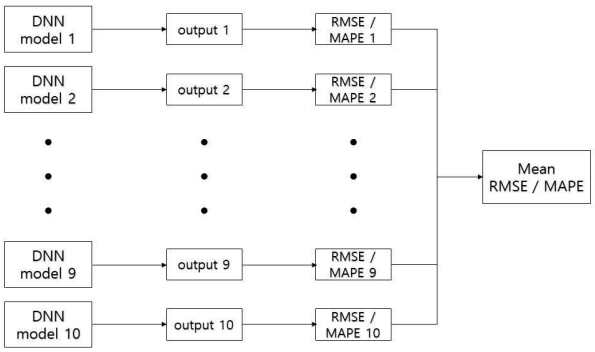

그림 4. Ensemble Method를 사용하지 않은 모델

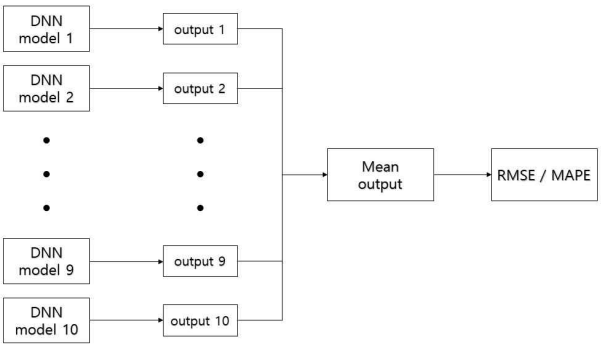

그림 5. Ensemble Method를 사용한 모델

표 4. 각 모델의 RMSE와 MAPE 비교

모델	RMSE	MAPE(%)
다중선형회귀	0.14657	33.2168
Basic_DNN	0.14201 (+3.12%)	31.0624 (+6.49%)
D_DNN	0.14193 (+3.17%)	31.5446 (+5.03%)
B_DNN	0.17545 (-19.70%)	44.8718 (-35.09%)
DB_DNN	0.13571 (+7.41%)	29.5661 (+10.99%)
E_Basic_DNN	0.13188 (+10.03%)	29.4211 (+11.43%)
E_D_DNN	0.13099 (+10.63%)	28.5348 (+14.10%)
E_B_DNN	0.15904 (-8.50%)	42.2374 (-27.16%)
E_DB_DNN	**0.12433 (+15.17%)**	**27.4946 (+14.34%)**

모델들의 예측 결과는 표 4와 같다. 표 4에서 Basic은 Dropout과 Batch normalization을 모두 사용하지 않은 것을 표기하였으며 D는 Dropout을, B는 Batch normalization을 그

Smart Media Journal / Vol.7, No.1 / ISSN:2287-1322
http://dx.doi.org/10.30693/SMJ.2018.7.1.22

리고 DB는 Dropout과 Batch normalization 모두 사용한 것을 표기하였다. 또한 E는 Ensemble Method를 사용한 모델을 나타낸다. 괄호안의 수치는 다중선형회귀 모델의 예측력 대비 각 모델들의 예측력의 증가 및 하락률을 나타낸다. 예측 결과 Batch normalization 만을 사용한 모델은 다중선형회귀 모델보다 나쁜 예측력을 보였고 그 외 모델들은 더 좋은 예측력을 보였으며 그 중 Dropout과 Batch normalization을 사용하여 Ensemble Method를 사용한 모델이 가장 좋은 예측력을 나타냈다. 이 모델은 다중선형회귀 모델 대비 RMSE는 15.17%, MAPE는 14.34% 증가함을 보였다.

III. 결 론

가. 연구결과 요약

본 연구에서는 DNN 모델을 이용하여 광주-기아 챔피언스필드의 일일 관중 수 예측을 위한 새로운 모델을 제안하였다. 제안 모델은 크게 데이터 수집 및 전처리, 모델설계 그리고 데이터 학습 및 예측으로 구성된다.

먼저, 데이터 수집 및 전처리 과정에서는 2008년부터 2017년까지 10년간의 일별 관중 수와 일일 관중 수에 영향을 미칠 수 있다고 판단되는 요소들을 시간요소, 기상요소, 팀 요소, 기타요소로 나누어 수집하고 0에서 1사이의 값으로 정규화 하였다.

모델설계 과정에서는 DNN 모델의 각 층들의 노드 수와 은닉층의 개수 그리고 학습률과 L2 Regularization의 lambda값과 같은 하이퍼 파라미터와 활성화함수 등을 정하였다.

데이터 학습 및 예측 과정에서는 데이터 수집 및 전처리 과정에서 구성된 데이터세트의 80%를 바탕으로 Adam optimization을 이용하여 DNN 모델을 학습하였다. DNN 모델의 문제점인 과적합을 방지하기 위해 모델에 따라 Dropout과 Batch normalization을 사용하였다. 학습된 DNN 모델에서 학습에 사용되지 않은 나머지 20% 데이터 세트를 이용하여 일일 관중 수를 예측하였다.

DNN 모델의 예측력을 평가하기 위하여 같은 훈련 데이터 세트로 다중선형회귀 모델을 만들어 RMSE와 MAPE를 비교하였다. 그 결과 Dropout과 Batch normalization을 사용하여 Ensemble Method를 사용한 DNN 모델의 예측력이 가장 우수하게 나왔으며, 다중선형회귀 모델 대비 RMSE는 15.17%, MAPE는 14.34% 증가함을 보였다.

나. 연구한계 및 제언

본 연구는 두 가지 한계를 가지고 있다. 향후 보다 발전된 연구를 위하여 다음과 같은 한계를 보완하여야한다. 그 내용을 제언을 함께 요약해보면 다음과 같다.

첫째, 본 연구에서 연구데이터로 선정된 프로야구 관중 수 데이터는 총 10년의 데이터이지만 현재 사용하고 있는 광주-기아 챔피언스필드의 데이터의 양은 4년의 데이터라는 점이다. 무등경기장에서 광주-기아 챔피언스필드로 구장의 변화에 따라 평균 일일 관중 수도 변화하였지만 광주-기아 챔피언스필드의 일일 관중수를 예측하기에 4년간의 데이터가 적다고 판단되어 6년간의 무등경기장의 데이터를 함께 사용하였다. 그에 따라 일일 관중 수 예측에 있어 부정적인 영향을 주었을 것이라고 판단된다. 앞으로 시간이 지날수록 관중 수 데이터들은 쌓이게 될 것이며, 이를 바탕으로 더 정확한 모델 구성 및 예측을 할 수 있을 것이다.

둘째, 모델 설계 및 하이퍼 파라미터의 문제이다. 본 연구에서는 모델 설계 및 하이퍼 파라미터를 일정 범위 내에서 시행착오를 통해 최적이라고 판단되는 것을 선택하였다. 향후 연구에서는 시계열 데이터를 처리할 수 있는 ANN인 Long Short-Term Memory(LSTM)나 Gated Recurrent Unit(GRU)과 같은 모델들을 바탕으로 하이퍼 파라미터의 조정을 통해 더 나은 성능을 기대할 수 있을 것이다.

REFERENCES

[1] 정병기, “프로야구 관람결정요인이 관람만족도 및 재구매행동에 미치는 영향”, *한국사회체육학회지*, 제29권, 209-220쪽, 2007.
[2] 송한성, “한국 프로야구단 연고지별 관중 수요예측 연구”, *한양대학교 대학원 석사학위논문*, 2013.
[3] 채한승, 이종호, “프로 스포츠팬 성향 및 경기관람 결정요인에 관한 조사 연구”, *한국스포츠산업 경영학회지*, 제5권, 2호, 137-154쪽, 2000.
[4] Hinton, G., Osindero, S. and Teh, Y., “A Fast Learning Algorithm for Deep Belief Nets”, *Neural Computation*, vol. 18, no. 7, pp. 1527-1554, 2006.
[5] Fausett, L. V., *Fundamental of neural networks: architectures, algorithms, & applications*, NJ: Prentice-Hall, 1994.
[6] Haykin, S. C., *Neural networks: A comprehensive foundation*, NJ: Prentice Hall PTR Upper Saddle River, 1994.
[7] 이유라, 김수형, 김영철, 나인섭, "심층 학습 모델을 이용한 EPS 동작 신호의 인식", *스마트미디어저널*, 제5권, 제3호, 35-41쪽, 2016년 9월.
[8] Li, D. and Yu, D., *Deep Learning: Methods and Applications*, Foundations and Trends® in Signal

Processing, pp. 197-387, 2014.
[9] 문대선, 나인호, 김성호, "풍력발전 고장검출 시스템을 위한 인공 신경망 기반의 모델링 기법 개발", *스마트미디어저널*, 제1권, 제2호, 47-53쪽, 2012년 3월
[10] 문해민, 박진원, 반성범, "역전파가 제거된 CNN과 LDA를 이용한 얼굴 영상 해상도별 얼굴 인식률 분석", *스마트미디어저널*, 제5권, 제1호, 24-29쪽, 2016년 3월.
[11] Nair, V., Hinton, G., "Rectified Linear Units Improve Restricted Boltzmann", *International Conference on Machine Learning*, pp. 807-814, 2010.
[12] Srivastava N., Hinton, G. E. Krizhevsky, A. Sutskever, I. Salakhutdinov, R., "Dropout: A simple way to prevent neural networks from overfitting", *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929-1958, 2014.
[13] Loffe, Sergey, and Christian Szegedy, "Batchnormalization: Accelerating deep network training by reducing internal covariate shift", *International Conference on Machine Learning*, 2015.
[14] 나병국, 윤성로, "GRU 기반 순환 신경망에서의 배치정규화 효과 연구", *한국정보과학회 학술발표논문집*, 663-665쪽, 2016.
[15] 변재진, "관광산업 수요예측 모형에 대한 연구", *대전전문대학 논문집*, 제20권, 103-156쪽, 1994.
[16] 이종원, *경제예측론*, 서울 : 도서출판 해남, 2006.
[17] 정의선, 유정정, 조승현, 중국 인바운드 관광수입의 수요예측 -ARIMA모형에 의한 시계열분석을 중심으로-, *호텔리조트연구*, 제12권, 제1호, 135-157쪽, 2013.
[18] 오승은, "다중회귀분석을 이용한 남한강보 지점에서의 Chlorophyll-a 농도예측에 대한 연구", *을지대학교 대학원 석사학위논문*, 2016.
[19] 오광우, 이우리, *시계열예측 방법과 응용*, 자유아카데미, 1995.
[20] 조광익, *관광수요 예측 및 경제 파급효과 분석 : 강원 역사문화촌을 중심으로*, 한국관광연구원, 1999.
[21] 최영문, 김사헌, "단변량 시계열 관광수요 예측모형의 적정성 비교평가: 내국인 해외관광객수 실측치와 예측치의 비교", 관광학연구, 제21권, 제2호, 111-128쪽, 1998.
[22] 김형돈, 채진석, "시계열모형을 이용한 프로야구 구단별 관중 수 예측", *한국체육측정평가학회지*, 제14권, 제3호, 57-68쪽, 2012.
[23] 김민철, "시계열분석을 통한 프로야구 관중현황 예측모델연구", *한국스포츠산업·경영학회지*, 제14권, 제1호, 17-25쪽, 2009.
[24] 박진욱, 박상현, "인공신경망을 이용한 한국프로야구 관중 수요 예측에 관한 연구", *정보처리학회논문지*, 소프트웨어 및 데이터 공학, 제6권, 제12호, 565-572쪽, 2017.

저 자 소 개

박동주(준회원)

2018년 전남대학교 산업공학과 학사 졸업.
2018년 ~ 현재 광주과학기술원 전기전자컴퓨터공학부 석사 과정.

<주관심분야 : 자연어처리, 생성모델, 기계학습>

김병우(준회원)

2018년 ~ 현재 전남대학교 산업공학과 학사 과정.

<주관심분야 : 자율주행 자동차, 스마트카, IoT>

정영선(정회원)

1997년 전남대학교 산업공학과 학사 졸업.
2001년 고려대학교 산업공학과 석사 졸업.
2011년 뉴저지주립대학교 산업공학과 박사 졸업.

2014년 ~ 전남대학교 산업공학과 조교수.
2018년 ~ 현재 전남대학교 산업공학과 부교수.

<주관심분야 : 기계학습, 빅데이터 분석, 품질경영, 지능형교통시스템>

안창욱(종신회원)

2000년 고려대학교 전파공학과 석사 졸업.
2005년 광주과학기술원 정보통신공학과 박사 졸업.
2005년 ~ 2007년 삼성종합기술원전문연구원.
2008년 ~ 2017년 성균관대학교 컴퓨터공학과 부교수.
2017년 ~ 현재 광주과학기술원 전기전자컴퓨터공학부 교수.

<주관심분야 : 진화알고리즘, 지능형네트워크, 자기-재조립 군집 로봇>