KIPS Tr. Software and Data Eng.
Vol.6, No.12 pp.565~572 pISSN: 2287-5905

https://doi.org/10.3745/KTSDE.2017.6.12.565

A Study on Prediction of Attendance in Korean Baseball League Using Artificial Neural Network

Jinuk Park† · Sanghyun Park††

ABSTRACT

Traditional method for time series analysis, autoregressive integrated moving average (ARIMA) allows to mine significant patterns from the past observations using autocorrelation and to forecast future sequences. However, Korean baseball games do not have regular intervals to analyze relationship among the past attendance observations. To address this issue, we propose artificial neural network (ANN) based attendance prediction model using various measures including performance, team characteristics and social influences. We optimized ANNs using grid search to construct optimal model for regression problem. The evaluation shows that the optimal and ensemble model outperform the baseline model, linear regression model.

Keywords : Artificial Neural Network, Korean Baseball League, Attendance, Hyperparameter, Grid Search, MAPE

인경신경망을 이용한 한국프로야구 관중 수요 예측에 관한 연구

박 진 욱† · 박 상 현††

요 약

본 연구는 기존의 수요 예측 등의 시계열 연구에서 주로 사용되는 ARIMA 모형의 어려움을 극복하고자 인공신경망(Artificial neural network) 모형을 이용하여 한국 프로 야구 관중 수를 예측하였다. 훈련 자료로는 2015년 3월부터 9월까지의 일별 KBO 관중 수 자료를 대상으로 하였다. 전방향 신경망(Feedforward neural network)의 모형 훈련 과정에서, 그리드 탐색(Grid search)을 적용하여 최적의 초모수(Hyperparameter)를 찾고자 하였다. 그 결과, 그리드 탐색법의 최적 모형을 이용한 평균 절대 백분율 오차(MAPE)는 평균 20.9% 였다. 앙상블 기법을 이용한 모형의 MAPE는 평균 20.0%였다. 이는 다중회귀와 비교해보았을 때, 평균적으로 각각 26.3%, 30.3% 높은 예측력을 보인다.

키워드 : 인공신경망, 한국프로야구, 관중 수, 초모수, 그리드 탐색, 평균 절대 백분율 오차

1. 서 론

한국프로야구 관중 수는 1980년대 초반 약 140만 명을 시작으로 2015년에는 약 730만 명으로, 한국 스포츠의 주요 종목으로 성장해왔다. 특히 2016년에는 834만 명으로 역대 최다 관중 수를 갱신하며, 한국프로야구의 입지를 증명하였다. 이처럼 한국 프로 스포츠 발전의 가장 중요한 관심사 중 하나는 관중 수라고 할 수 있다. 관중 수는 구단의 경영 측면과 직결되는 요소로써, 입장료부터 기타 시설의 이용료 등 다양한 수입을 통해 구단의 안정적인 재정 운영을 가능케 한다. 또한, 수요예측은 한국 프로 스포츠의 마케팅과 구단의 예산 전략 수립에 활용될 수 있다[1].

수요예측에 사용되는 전통적으로 모형은 Box Jenkins의 ARIMA 모형[2]이다. ARIMA 모형은 시계열 분석 방법의 대표적인 모형으로써, 시계열 자료의 자기 상관 특성을 이용한 모형이다. 선행 연구들은 이와 같은 시계열 모형을 이용하여 년 단위 누적 관중 수의 시계열 특성을 파악하여 예측값을 도출하였다[3-7]. 본 연구에서는 연간 누적 관중 수가 아닌 관중 수를 일별로 수집하여 각 홈의 경기장 별로 일별 관중 수를 예측하는 모형을 만드는 것을 목표로 하고

※ 본 연구는 2015년도 정부(미래창조과학부)의 재원으로 한국연구재단-차세대정보·컴퓨팅기술개발사업의 지원을 받아 수행된 연구임 (NRF-2015M3C4A7065522).
※ 이 논문은 2017년도 한국정보처리학회 춘계학술발표대회에서 '인공신경망을 이용한 한국프로야구 관중 수요 예측'의 제목으로 발표된 논문을 확장한 것임.
† 준 회 원 : 연세대학교 컴퓨터과학과 석사과정
†† 종신회원 : 연세대학교 컴퓨터과학과 교수
Manuscript Received : July 17, 2017
Accepted : August 17, 2017
* Corresponding Author : Sanghyun Park(sanghyun@yonsei.ac.kr)

있다. 활용도 측면에서 일별 관중 수요 예측은 연간 누적 관중 수 분석보다 더 큰 효용을 가져올 것으로 판단하였다.

하지만 일별 관중 수 분석에는 선행 연구들에서 사용된 전통적인 ARIMA 모형을 구축하는 데에 두 가지 문제점이 있었다. 첫 번째로, 한 홈구장에서 진행되는 경기의 주기가 일정하지 않았다. 예를 들면, 롯데의 홈구장인 사직구장의 2015년 4월 첫 주에는 금, 일요일에 경기가 진행되었다. 하지만 둘째 주에는 금, 토, 일요일, 셋째 주에는 화, 수, 목요일에만 경기가 진행되었다. 따라서 시계열 자료의 주기적인 관측값에 대한 추세와 경향을 바탕으로 모형을 가정하는 ARIMA 분석법을 적용할 수 없다.

두 번째로, ARIMA 모형의 특징인 과거 관측값과의 자기상관관계(Autocorrelation)를 통한 분석을 적용할 수 없었다[2]. 위와 같은 예로, 사직구장의 자기상관 분석결과가 Table 1에 나타나 있다. 24 시차까지의 자기상관을 확인해본 결과, 유의 수준 0.05를 기준으로 '자기상관이 존재하지 않는다.'는 귀무가설을 기각하지 못하였다. 즉, 현재 관측되는 값과 과거의 자료들이 상관관계를 가지지 못한다고 판단할 수 있다. 이는 ARIMA 모형의 가정과 반대되는 결과로, 모형을 이용한 시계열 분석을 적용하지 못한다.

따라서 본 연구에서는 일별 관측값이 각각 독립적인 자료로 판단하고, 독립 모형인 전방향 인공신경망(Feedforward neural network) 모형을 적합하여 미래 관중 수를 예측하는 것을 목표로 한다. 인공신경망의 최적화를 위해 그리드 탐색법을 이용하여 최적의 초모수 집합을 탐색하였으며, 신경망 구현을 위해 H2O 프레임워크를 사용하였다.

Table 1. Autocorrelation Test for Sajik Stadium

Lag	P-value
6	0.26
12	0.33
18	0.59
24	0.28

2. 관련 연구 및 배경 지식

2.1 한국프로야구 관중 수 예측

한국프로야구를 포함한 프로 스포츠 관중과 관련된 연구들은 크게 두 가지 관점으로, 관중 수를 예측하는 관점과 관중 수에 영향을 미치는 요인을 분석하는 관점이 있다. 본 연구는 관중 수를 예측하는 연구로, 선행 연구의 결과를 이용하여 입력 변수들을 선정하였다.

먼저, 관중 수를 예측한 선행 연구들은 대부분 연간 관중 수 자료를 사용하여 ARIMA 모형을 만들고, 연도별 추세를 예측하였다. [3]은 1982년부터 2008년까지의 부산 롯데자이언츠 관중 수를 이용하여 관중예측모형을 만들고, 향후 5년간의 예측값을 추정하였다. [4]는 모든 팀의 합산된 누적 관중 수를 예측하였고, [5]에서는 구단별 모형을 적합하였다.

또한, 추가적인 독립변수를 이용하여 ARIMA 모형을 다변량 모형으로 확장한 연구[5, 6]에서는 단변량 및 다변량 ARIMA와 더불어, 추세모형, 최근린예측모형[6]과 비교하여 최적의 모형을 선택하였다. 1982년 시작한 한국프로야구의 연간 관중 자료의 개수가 제약적이라는 한계를 극복하기 위해, 공변량을 도입한 ARIMAX와 성장곡선(GCX) 모형이 소개되기도 하였다[7]. [8]의 연구에서는 자기회귀모형과 독립변수 외에도 GARCH 모형을 이용하여 조건부 이분산성을 고려하였다. 한편, 관중 수가 아닌 야구 통계량의 장기 추세에 관한 시계열 분석 연구도 제시되었다[9].

관중 수에 영향을 미치는 요인에 대한 연구는 설명력이 높은 회귀모형을 이용한 연구가 많았다. 관중 수를 종속변수로 두고 독립변수들의 회귀 계수를 추정하는 회귀모형을 사용하여 관중 수를 분석하였다[10, 11]. 소득 등과 같은 거시 변수를 사용하여 경기당 평균 관중 수를 분석한 [10]은 입장료, 지역별 차이와 더불어 특정 선수의 효과를 통해 심층 분석을 시도하였다. [11]은 경기별 관중 수를 대상으로, 원정팀과 요일에 따른 중도절단회귀모형을 적합하였다. 이를 통해 공휴일과 원정팀, 요일이 관중 수에 유의미한 영향을 미친다는 것을 밝혔다. 또한, 창단된 신 구단의 여부, 승률 등의 외부적인 요인들이 관중 수에 영향을 미친다는 것이 연구되었다[12-14].

스포츠 경기의 승패를 예측하거나 관련이 있는 요인을 분석한 연구들도 제시되었다[15-17]. 또한, 인공신경망 모형과 로지스틱 회귀모형을 이용하여, 한국프로야구 관전자의 행동을 분석하고 예측한 연구도 진행되었다[18]. 특히 스포츠 경기의 결과를 예측하는 데에 높은 예측력을 나타내는 인공신경망이 제시되고 있다[17]. 이처럼 한국프로야구를 비롯한 스포츠 분야의 연구에서 인공신경망 또는 딥러닝을 적극적으로 활용하려는 움직임이 있다.

2.2 인공신경망 개요

Fig. 1은 인공 신경망의 각 요소들을 나타낸다. 인공 신경망은 입력층(Input layer), 출력층(Output layer), 은닉층(Hidden layer)으로 구성되어 있고, 각 층에는 1개 이상의 노드들이 존재한다.

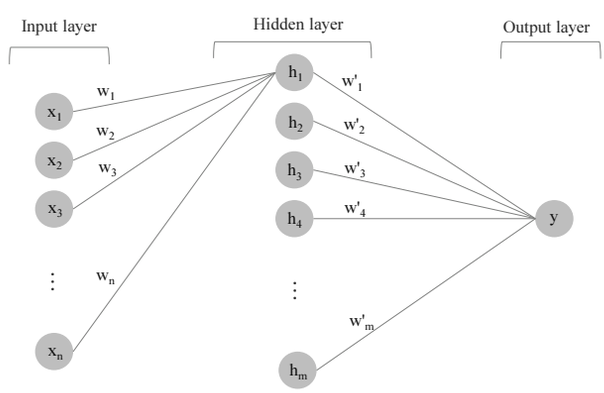

Fig. 1. Structure of Artificial Neural Network

각 노드들의 정보는 각각 가중치를 곱하여 다음 층의 모든 노드로 전달된다. 다음 층의 입력값은 이전의 모든 노드들과 가중치들을 곱한 가중합이 되고, 활성함수(Activation function)를 거쳐 다음 층으로 다시 연결된다. 즉, 입력층을 제외한 각 노드들의 출력값은 Equation (1)과 같이 결정된다.

$$O_j = f(\sum_{i=1}^{n} O_i w_{ij} + b_j) \quad (1)$$

이 때, O_i와 O_j는 각각 입력층의 i번째 노드와 은닉층의 j번째 은닉 노드의 출력값을 나타낸다. w_{ij}는 입력층의 i번째 노드가 은닉층의 j번째 노드로 전달될 때 곱해지는 가중치이다. 또한 b_j는 j 노드가 속한 은닉층의 편향(Bias)이다.

활성함수의 종류로는 선형함수(Linear function), 시그모이드(Sigmoid), tanh, ReLU 함수가 존재한다. Tanh 함수는 시그모이드 함수의 문제점인 훈련이 지연되는 점을 해결할 수 있고, ReLU 함수는 더 빠르게 수렴하고, 그라디언트가 소실되는 것을 막을 수 있다[19]. 본 연구에서는 은닉층의 활성함수로 대중적인 ReLU 함수를 선정하였다.

본 연구는 기존의 선행 연구에서 밝혀진 요인들을 선정하여 일별 관중 수를 예측하는 모형을 만드는 것을 제안한다. 이전 경기의 결과, 승률 등과 같은 내적 요인과 기후, 공휴일 등의 외적 요인을 추가하였다. 또한, 댓글과 투표수와 같은 소셜 요인을 고려하여 모형을 설계하였다. 많은 선행 연구들에서 밝혀낸 홈 팀과 원정팀의 영향을 고려하기 위해, 홈 팀별로 모형을 생성하였다. 일별 관중 수를 예측하는 모형으로 ARIMA 모형을 적용할 수 없다는 한계점을 극복하기 위해 인공신경망을 사용하였다.

3. 분석자료 범위와 자료 수집

본 연구에서 사용한 관중 수의 범위는 2015년 3월부터 9월까지 한국프로야구의 경기별 관중 수로써, 관측값은 총 702개이다. 관중 수와 구단 관련 자료는 KBO에서 제공하는 2016년 연감[20]에 기재된 자료를 기준으로 하였다.

인공신경망의 입력 변수로는 시간 요소, 날씨 요소, 지역 요소, 팀별 특성 요소, 누적 경기 성적, 대중(Social) 요소로써 총 6개의 요소를 선정하였다. 시간 요소는 해당 날짜와 요일, 그리고 공휴일과 성수기를 나타내는 이항변수를 생성하였다. 성수기의 기준은 여름 휴가에 해당하는 7월과 8월에는 1, 그 외 경우 0으로 결정하였다. 마찬가지로, 공휴일 변수는 공휴일에 해당하는 날짜에만 1을 생성하였다.

날씨 요소는 기상청 데이터베이스의 일별 평균 습도과 평균 기온을 포함한다. 지역 요소는 해당 경기가 진행된 구장을 나타낸다. 즉, 구단별 구장으로써, 사직(롯데), 울산(롯데), 잠실(두산, LG), 포항(삼성), 대구(삼성), 청주(한화), 목동(넥센), 대전(한화), 수원(kt), 마산(NC), 광주(KIA), 문학(SK) 총 12개의 범주를 가지고 있다. 팀별 특성 변수에는 각 경기의 홈/원정팀을 나타내는 변수와 홈/원정팀의 존속 기간을 나타내는 변수를 포함한다. Round 변수는 두 팀이 진행하는 경기가 몇 차전 경기인지를 나타낸다. 또한, 같은 구장을 공유하는 두산과 LG의 경우, 라이벌 구단임을 나타내는 이항변수를 생성하였다. 경기 성적에 대한 변수로는 각 경기 일을 기준으로 그 전날까지의 순위와 누적 승수, 누적 승률 변수가 존재한다. 대중 요소로는 Naver[21]에서 경기별로 사전에 진행하는 투표 수 및 응원 댓글 개수를 선정하였다. 이를 참조하여 홈, 원정팀 각각의 투표 수, 응원 댓글 수와 두 팀의 합산 투표, 댓글 수를 변수를 생성하였다.

Table 2는 위에서 언급한 인공신경망의 출력변수와 입력변수로 사용한 요소들을 요약하고 있다. *표시는 이항 변수, **표시는 명목형 변수임을 나타낸다. 명목형 변수는 데이터 분석 시에, One-hot 인코딩 방식을 사용하였다. 즉, 각각의 범주에 해당하는 가변수들과 각 명목형 변수의 결측값을 나타내는 가변수를 생성하여 분석에 활용하였다.

추가적으로, 2015년도는 질병 MERS가 유행했던 시기이다. Fig. 2는 관중 수와 MERS 격리환자 수와의 상관관계를 나타낸다. 관중 수가 급격하게 감소하는 5월 중순부터 7월 중순까지의 추세(이동평균)와 격리된 환자 수와의 관계가

Table 2. Input and Output Variables

	Variables	
Input	Time	Date
		Day**
		Holiday*
		Peak*
	Climate	Average temperature
		Average humidity
	Location	Stadium**
	Characteristic	Home team name**
		Away team name**
		Home team duration year
		Away team duration year
		Rival*
		Rounds
	Performance	Home team ranking
		Away team ranking
		Home team cumulative win
		Away team cumulative win
		Home team win rate
		Away team win rate
	Social Influence	Comments for home team
		Comments for away team
		Total comments
		Votes for home team
		Votes for away team
		Total votes
		MERS*
Output	Attendance per game	

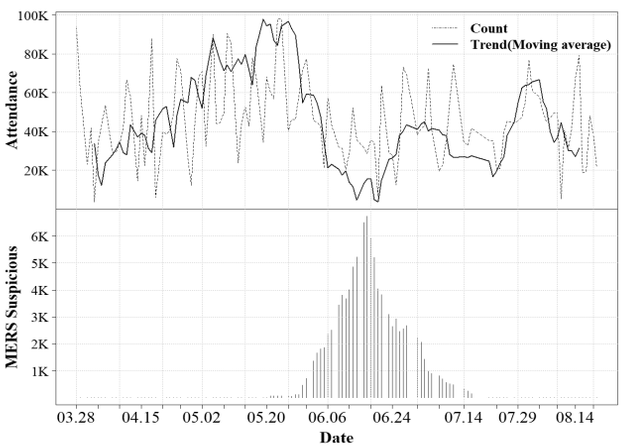

Fig. 2. Relationship between Attendance and MERS Patients

반비례하는 것을 확인하였다. 또한, 관중 수와 환자 수 자료의 상관계수(Correlation coefficient)는 -0.53이고, 이는 일정 수준의 음의 상관관계를 가진다고 판단할 수 있다. 이는 MERS 환자 수가 증가함에 따라 사회적인 공포가 형성되어 관중 수가 급감하였다고 해석할 수 있다. 따라서 MERS로 인한 첫 번째 사망 환자가 발생한 6월 1일부터, 마지막 사망 환자가 발생한 7월 10일까지를 MERS 유행 기간으로 판단하고, 이항변수 MERS (유행=1)를 생성하여 이를 나타내었다.

4. 모형 훈련

4.1 인공신경망 초모수 탐색 및 훈련

인공신경망 훈련은 실제값과 출력값과의 차이를 통한 오류역전파(Backpropagation) 알고리즘을 사용한다. 이는 목표값과의 오차를 줄이는 방향으로 가중치들을 갱신하는 방법이다[22]. 이 때, 선택되는 훈련 자료와 초기에 선정되는 가중치는 모형 훈련에 큰 영향을 끼치게 된다. 따라서 본 연구의 모형 훈련에서는 이를 최소화하면서 초모수 집합을 비교하기 위해, 교차검증법(Cross-validation)을 적용하여 초모수를 탐색하고자 하였다.

또한, 구단별 야구 경기장마다 최대 입장객 수가 다르므로, 모형의 성능을 높이기 위해 데이터를 홈 팀 별로 분할하였다. 이는 제 2구장에서 열리는 경기는 소수이기 때문에, 경기장이 아닌 홈 팀을 기준으로 데이터를 나누었다. 모형의 초모수 탐색과 훈련 과정은 나뉜 데이터에서 각각 진행하였다. Fig. 3은 전반적인 데이터 분할과 훈련 과정을 나타낸다.

인공신경망의 초모수란 모형의 성능을 결정하는 부분으로, 은닉층과 뉴런의 개수, 과적합을 방지하는 정규화 방법, 학습률, 활성함수 등이 존재한다. 하지만 이를 위한 최적화 방법으로는 정해진 것이 없으며, 사용자가 시행착오법을 통해 실험적으로 오차를 줄일 수 있는 초모수를 선택한다[23]. 본 연구에서는 초모수 범위를 선정하여, 최적의 초모수를 찾는 그리드 탐색을 적용하였다[24]. 모형 훈련을 위한 초모수 그리드의 대상과 범위는 Table 3에 나타나있다.

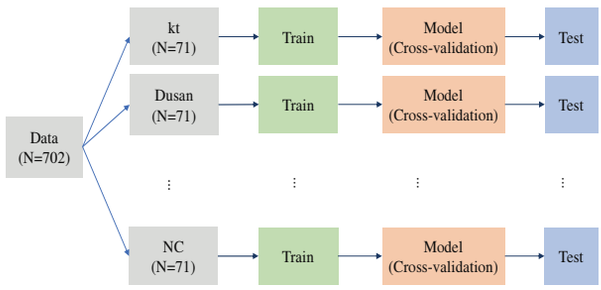

Fig. 3. Segmentation and Training Procedure

Fig. 3에서 나타난 입력 노드의 개수는 홈 팀 기준으로 분할된 서브 데이터 세트에 따라 다르다. 분할된 각 데이터의 입력 변수 중 제 2구장의 유무와 LG-두산을 나타내는 라이벌 구단 변수의 One-hot 인코딩으로 인해 가변수의 개수가 달라진다. 따라서 제 2구장이 없고, LG 또는 두산이 아닌 구단의 입력 노드는 44개이고, 제 2구장이 있거나 LG, 두산의 경우는 47개의 입력 노드를 가진다. 제 1구장, 제 2구장, 그리고 결측값을 나타내는 경우로 각각 One-hot 인코딩하였기 때문에 3개의 노드가 차이 난다.

은닉층과 은닉 노드의 개수는 Table 3과 같이 탐색 범위를 설정해주어 그리드 탐색을 통해 결정하고자 하였다. 은닉층과 은닉 노드의 개수가 많으면 훈련 시간이 오래 걸리고, 과적합으로 인해 모형 성능이 떨어진다는 단점이 있다. 이를 해결하기 위해, L1 정규화를 병행하여 많은 은닉 노드를 설정하여 생기는 과적합을 방지하고자 하였다. 따라서 입력 노드의 절반의 근사치인 20개를 최소 개수로 선정하고, 이를 배수로 늘려 최대 4배인 80개까지를 범위로 설정하였다. 은닉층이 2개인 경우는, 모든 가능한 경우에 대해서 은닉 노드의 개수를 탐색을 시도하였다. L1 정규화 범위는 {0.001, 0.0001}로 설정하여 탐색하였다. 학습률은 일반적으로 사용하는 0.05로 설정하였다.

Table 3. Grid Search for Hyperparameters

Hyperparameters		Range
Input nodes		44+
Output node		1
The number of nodes for hidden layers	1 layer	4 cases {20, 40, 60, 80}
	2 layer	16 cases (Cartesian product of above nodes)
l1 Regularization		{0.0001, 0.001}
Learning rate		0.05
Activation function for hidden layer		ReLU
Activation function for output layer		Linear function
Stopping metric		RMSE
Cross-validation		5 Folds cross-validation
Epochs		30,000

초모수 탐색 후, 훈련과 평가 데이터를 나누는 과정은 두 가지 방법으로 진행하였다. 경기 날짜와 경기 순번이 입력 변수에 포함되므로, 날짜와 무관하게 랜덤하게 평가하는 방법과 날짜순으로 나눠 평가하는 방법을 사용하였다.

4.2 모형 앙상블

그리드 탐색을 수행하면, 4.1절의 그리드 범위에 해당하는 여러 개의 모형을 생성하게 된다. 그중 교차검증 에러가 최소가 되는 최고의 모형을 포함하여 상위 5개 모형을 사용하여 앙상블을 구축하고자 하였다. 이 방법은 최적 모형보다 성능이 떨어지는 모형을 사용함으로 인해 최종 예측력을 낮출 위험이 있다. 하지만 앙상블 내의 다양성을 증대시키고, 과적합을 방지할 수 있다는 장점을 기대하였다. Fig. 4는 최적 모형을 선정하는 것과 앙상블 모형을 구축하는 모습을 나타낸다.

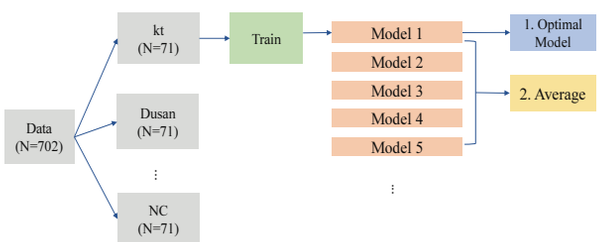

Fig. 4. Optimal and Ensemble Models

5. 결과 및 평가

5.1 초모수 탐색 결과

Table 4는 홈 팀 별로 분할한 데이터에서 그리드 탐색을 통해 결정된 초모수 집합을 나타낸다. 교차 검증된 RMSE (Root mean square error)를 기준으로, 가장 낮은 에러를 나타내는 초모수를 최적 모형으로 선택하였다. 출력변수인 관중 수를 각 경기장의 최대 수용인원으로 나누어 0과 1사이의 비율로 치환하였기 때문에, RMSE가 1보다 작은 값으로 계산되었다.

Table 4. Optimal Hyperparameters for Each Team

Home	Neurons	l1	RMSE(C-V)
KT	(40, 80)	0.0001	0.1102
Dusan	(20, 80)	0.001	0.1271
LG	(60, 40)	0.0001	0.1257
KIA	(80, 60)	0.0001	0.1329
Lotte	(80, 40)	0.0001	0.2001
Samsung	(60, 80)	0.0001	0.1587
SK	(60)	0.0001	0.1603
Hanhwa	(60, 40)	0.0001	0.2226
Nexen	(60, 60)	0.001	0.1606
NC	(40, 40)	0.001	0.1919

모든 최적 모형에서, 입력 노드의 개수에 비해 많은 수의 은닉 노드를 가질 때 가장 낮은 에러를 나타내었다. 이는 L1 정규화를 통해 과적합을 방지한 효과로 생각할 수 있다. L1은 모형에 불필요한 변수들의 가중치를 0으로 만들어, 데이터의 많은 One-hot 인코딩을 처리할 수 있다는 장점이 있다[25].

5.2 교차검증 분석 결과

경기 날짜와 경기 순번이 변수로 입력되므로, 날짜와 순번과 무관하게 랜덤하게 나누어 모형을 평가하였다. 랜덤하게 나누는 과정에서 편향이 발생할 수 있기 때문에, 5개의 독립적인 훈련/평가 데이터로 나눠 교차검증을 시행하였다.

또한 평균 절대 백분율 오차(Mean absolute percentage error; MAPE)를 기준으로 사용하여 최적 모형들을 검증하였다. MAPE란 실제 관측값과 예측된 값을 비교하였을 때, 평균적으로 각 관측값과 차이 나는 에러의 비율을 나타낸다. 특히 수요 예측의 경우, MAPE를 예측력을 판단하는 지표로 사용하여 직관적인 이해를 돕는다[26]. MAPE는 Equation (2)와 같이 계산된다.

$$MAPE = \sum \frac{|y_t - \hat{y}_t| / y_t}{n} \times 100 \quad (2)$$

Table 5는 5.1절에서 선정한 최적 모형, 상위 5개 예측값의 평균인 앙상블 모형을 다중회귀와 비교한 결과를 나타낸다. 각 모형의 에러는 교차검증된 평균 MAPE로 계산되었다. 괄호 안의 값은 다중회귀의 예측력 대비 향상된 비율이다. 다중회귀모형은 데이터의 독립성을 가정하는 대표적인 모형으로서, 비교 대상 모형으로 선정하였다.

최적 모형으로 선택한 전방향 신경망은 다중회귀 대비 평균적으로 26.3% 높은 예측력을 보이고 있다. 앙상블 모형의 경우, 다중회귀 대비 30.3% 높은 예측력을 보이고 있다. 이는 앙상블 모형이 일반화된 예측값을 통해 더 높은 예측력을 얻었다고 판단할 수 있다.

Table 5에서 최적 모형의 MAPE는 팀별로 평균 20.86으로, 최소값인 삼성팀의 13.05부터 최대값인 롯데팀의 36.61까

Table 5. Evaluation Results of Random Analysis

Home	Optimal	Ensemble	Regression
KT	22.93 (+23.7%)	20.30 (+32.4%)	30.04
Dusan	13.90 (+51.0%)	12.29 (+56.7%)	28.39
LG	13.38 (+15.7%)	11.62 (+26.8%)	15.88
KIA	21.85 (+12.1%)	21.68 (+12.8%)	24.87
Lotte	36.61 (+3.4%)	35.36 (+6.7%)	37.90
Samsung	13.05 (+70.8%)	13.29 (+70.2%)	44.62
SK	31.24 (+18.1%)	35.05 (+8.1%)	38.16
Hanhwa	19.56 (+46.1%)	16.6 (+54.2%)	36.27
Nexen	21.68 (+17.7%)	20.56 (+22.0%)	26.35
NC	14.41 (+3.9%)	13.03 (+13.1%)	14.99
Average	+26.3%	+30.3%	

Table 6. Top 5 Residuals of Lotte Predictions

Date	Away	Stadium	Actual	Prediction	|residual|
2015-05-31	Hanhwa	Ulsan	1.00	0.40	0.60
2015-06-28	Nexen	Sajik	0.69	0.10	0.59
2015-09-12	Hanhwa	Sajik	0.38	0.92	0.54
2015-08-29	NC	Sajik	0.94	0.43	0.51
2015-09-20	Samsung	Sajik	0.37	0.85	0.48

지 분포하고 있다. 평균과 비교하여 보았을 때, 롯데와 삼성의 에러가 특별하게 높은 것을 확인할 수 있다. 원인 분석을 위해 두 가지 방법을 병행하였다. 첫 번째로, 이상치 유무를 파악하기 위해 예측값들의 에러를 분석하였다. 또한, 이와 병행하여 인공신경망 모형의 변수 중요도를 산출하여 비교 분석해보았다. 변수 중요도로는 입력 노드와 은닉층의 연결 강도를 바탕으로 계산하는 [27]의 방법이 선택되었다.

Table 6은 가장 MAPE 오차가 큰 롯데의 이상치를 탐색 결과를 나타낸다. 롯데의 최적 모형을 이용하여 예측값의 잔차를 구하고, 그 차이가 큰 상위 5개의 값을 나타내었다. 잔차는 실제값과 예측값의 절대적인 차이를 기준으로 계산하였다. 잔차가 가장 큰 5월 31일의 경기의 경우, 실제로는 울산 경기장이 가득 찰 정도로 관중이 많았지만 모형은 40%의 관중을 예측하였다. 가능한 해석은 학습 자료 부족이다. 실제로 울산 경기장은 롯데 팀의 제 2구장으로서, 경기 수는 9월 말까지 총 10회로 매우 적다. 또한, 울산 경기장의 관중 수는 대부분 0.4를 넘지 못하며, 이것이 크게 반영되었음을 변수 중요도를 통해 확인할 수 있었다. Fig. 5는 롯데 모형의 상위 10개의 변수 중요도를 나타낸다. 6월 28일 넥센과의 경기도 비슷한 원인으로 해석할 수 있다. 롯데가 홈으로 넥센과 붙은 경기의 관중 수들을 살펴보면 0.24로서, 상당히 낮은 관중 기록을 가지고 있다. 울산 경기장과 마찬가지로 넥센도 높은 변수 중요도를 가지고 있으므로, 6월 28일의 0.69을 평균 대비 이상치로 생각할 수 있다.

또한, 롯데는 6월 이후로 경기 성적이 저조하여 꾸준히 8위를 기록하였다. 저조한 경기 성적만큼 관중 수도 적었기 때문에, 실제값도 평균 0.28로 낮다. 따라서 Table 6의 5월, 6월, 8월의 큰 관중 수요는 특이값으로 간주할 수 있다. 반면 9월 12일 토요일, 9월 20일 일요일은 각각 0.38, 0.37로 상대적으로 낮은 관측값을 가진다. 롯데팀의 평균적인 토요일과 일요일의 관중 수는 각각 0.65, 0.51로서, 위의 9월 12일, 20일 관측값과 크게 차이남을 확인할 수 있다. 이처럼 큰 MAPE 값은 다른 데이터와 크게 다른 패턴을 가지는 이상치로부터 발생된 것을 확인하였다.

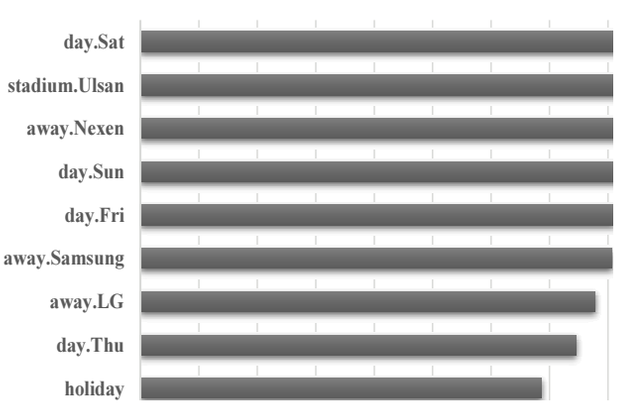

Fig. 5. Variable Importance of Lotte Model

5.3 순차적 분석 결과

시계열 데이터를 평가하기 위해, 훈련 데이터와 평가 데이터를 시간순으로 분할하였다. 이를 위해 3월부터 8월까지는 훈련 데이터에, 9월 데이터는 평가 데이터로 할당하였다. 5.1절에서 찾은 최적화된 초모수를 사용하여 최적 모형을 훈련하고, 5.2절과 동일한 방법으로 앙상블 모형을 구축하여 비교하였다.

Table 7은 순차적으로 데이터를 분할하여 모형을 평가한 결과이다. 최적 초모수를 사용한 모형과 에러가 낮은 상위 5개의 앙상블 모형은 각각 다중회귀 대비 평균적으로 14.2%, 9.6% 높은 예측력을 나타내고 있다. 5.2절의 교차검증을 통한 훈련 결과와는 다르게, 앙상블 모형의 예측력이 최적 모형보다 낮은 것을 확인할 수 있다. 이는 교차검증을 통하지 않고 편향적인 훈련 데이터를 통해 훈련을 시행했을 경우, 앙상블 모형이 더 낮은 성능을 보임을 시사한다. 즉, 편향된 데이터를 통해 훈련할 경우, 최적 모형보다 성능이 낮은 모형을 예측에 사용하면 낮은 성능이 미치는 영향이 앙상블을 통해 얻는 다양성보다 더 높은 것을 알 수 있다.

<표 7>에서 롯데의 경우는 최적 인공신경망 모형이 다중회귀 모형보다 낮은 예측력을 21.6% 낮은 예측력을 나타낸다. 이에 대한 원인 분석은 5.2절에서 특이값을 통해 밝혔다. 하지만 KIA와 SK의 경우, MAPE가 각각 44.79, 39.02로 상당히 높은 에러를 보인다. 하지만 두 팀 모두 다중회귀에서도 각각 52.26, 49.98로 더 높은 에러를 나타낸다. 즉, 모형의 문제가 아닌, 실제 관중 수의 변동이 커진 것이 원인임을 확인할 수 있다.

Table 7. Evaluation Results of Sequential Analysis

Home	Optimal	Ensemble	Regression
KT	21.72 (+12.4%)	18.87 (+23.9%)	24.79
Dusan	24.01 (+18.0%)	22.82 (+22.0%)	29.27
LG	11.62 (+56.2%)	10.41 (+60.7%)	26.51
KIA	44.79 (+14.3%)	35.32 (+32.4%)	52.26
Lotte	37.37 (-21.6%)	45.85 (-49.2%)	30.73
Samsung	24.2 (+25.4%)	23.17 (+28.6%)	32.44
SK	39.02 (+21.9%)	52.95 (-5.9%)	49.98
Hanhwa	29.33 (+2.0%)	29.33 (+2.0%)	29.94
Nexen	31.17 (+0.4%)	35.65 (-13.9%)	31.31
NC	18.76 (+13.9%)	22.87 (-5.0%)	21.79
Average	+14.2%	+9.6%	

6. 결론 및 향후 연구

일별 야구 관중 수 자료는 선행연구와 달리, 주기성과 자기 상관성이 존재하지 않기 때문에, ARIMA 분석 절차를 적용할 수 없었다. 본 연구에서는 이와 같은 한계점을 극복하기 위해 인공 신경망 모형을 활용하였다.

선행 연구들에서 밝혀진 요소들을 포함시키기 위해 경기 내적 요인과 외적 요인에 해당하는 자료를 수집하였다. 내적 요인으로는 팀별 특성 요소와 경기 성적, 승률 등을 포함하고 있고, 외적인 요인으로는 요일 정보, 소셜 요소 등을 포함한다. 이처럼 다양한 범위의 입력변수를 가지는 인공신경망을 설계하고, 그리드 범위를 설정 및 탐색하여 최적 모형을 찾고자 하였다. 또한 오차율이 낮은 상위 5개를 사용하여 앙상블 모형을 도출하였다.

날짜와 상관없이 랜덤하게 교차검증을 통해 훈련한 최적모형은 다중 회귀의 예측력과 비교하였을 때, MAPE를 기준으로 평균적으로 26.3% 뛰어난 예측력을 보였다. 특히, 삼성의 경우는 성능 향상이 70% 이상으로, 다중 회귀보다 월등히 뛰어난 예측력을 보인다. 시간에 따라 순차적으로 데이터를 나누는 방법으로서, 9월 데이터를 테스트 데이터로 사용한 최적모형의 경우는 평균 14.2%로 높은 예측값을 얻을 수 있었다.

본 연구 결과는 선행 연구에서 사용하지 않은 인공신경망 모형을 이용하여 일별 관중 수를 예측한다는 것에 의미가 있다. 인공신경망을 이용하여 자기 상관이 존재하지 않는 시계열 자료의 예측을 가능케 하였다. 또한 인공신경망의 초모수를 탐색하는 그리드 탐색의 초석을 다지고, 탐색의 활용성을 넓혔다.

본 연구에서는 2015년도 관중 수에 영향을 미치는 변수들을 설정하여 인공신경망을 훈련하였다. 인공신경망 또는 은닉층이 더 많이 존재하는 딥러닝 모형은 많은 데이터가 훈련될수록 더 높은 예측력을 나타낸다. 향후 연구에서는 본 연구에서 사용한 단일 년도 자료인 2015년 관중 수와 더불어 MLB, NPB와 같이 다양한 리그의 관측 값을 사용하여 훈련한다면, 완성도 있는 모형을 구축할 수 있을 것이라 사료된다. 또한 모든 가능한 조합을 고려하는 단순 그리드 탐색에서 시작하였지만, 이를 활용한다면 오차율을 낮추는 방향으로 진화하며 탐색하는 유전(Genetic) 탐색이 가능할 것이라 판단된다.

References

[1] Young-Hoon Lee, “Time-Series Analysis on Attendance in the Korean Professional Baseball League,” *Korean Journal of Sport Science*, Vol.13, No.4, pp.85-102, 2002.

[2] Box, Jenkins, and Gwilym M. Jenkins. “Reinsel. time series analysis, forecasting and control,” Tercera. NJ: Prentice Hall, Englewood Cliffs, NJ, USA, 1994.

[3] Min Cheol Kim, “Model study to predict the number of pro-baseball spectator by time series analysis: About Busan Lotte Giants’ spectator,” *Korean Journal of Sport Management*, Vol.14, No.1, pp.17-25, 2009.

[4] Min Sin Sul, Doo Yong Park and Mi Jeng Lee, “Forecast Study of Korea Pro Baseball Spectators by Using Time Series Analysis (2011-2015),” *Journal of Sport and Leisure Studies*, Vol.45, No.1, pp.375-387, 2011.

[5] Hyung-Don Kim and Jin-Seok Chae, “Prediction of the Number of Spectators for the Pro-baseball Club Using a Time Series Model,” *The Korean Journal of Measurement and Evaluation in Physical Education and Sport Science*, Vol.14, No.3, pp.57-68, 2012.

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

[7] Ga-Hee Han, Jigyu Chung, and Jae Keun Yoo, “A study on prediction for attendances of Korean probaseball games using covariates,” *Journal of the Korean Data & Information Science Society*, Vol.25, No.6, pp.1481-1489, 2014.

[8] Jangtaek Lee and Soyoung Bang, “Forcasting attendance in the Korean professional baseball league using GARCH models,” *Journal of the Korean Data & Information Science Society*, Vol.21, No.6, pp.1041-1049, 2010.

[9] Jangtaek Lee, “Long term trends in the Korean professional baseball,” *Journal of the Korean Data & Information Science Society*, Vol.26, No.1, pp.1-10, 2015.

[10] Young Hoon Lee, “The Decline of Attendance in the Korean Professional Baseball League,” *Journal of Sports Economics*, Vol.7, No.2, pp.187-200, 2006.

[11] Hyeuk Kim, “Prediction of the number of attendances in the home team according to the visiting team and the day in Korean Baseball League,” *Korean Journal of Sport Management*, Vol.21, No.6, pp.85-96, 2016.

[12] J. A. Winfree, J. J. McCluskey, R. C. Mittelhammer, and R. Fort, “Location and attendance in major league baseball,” *Applied Economics*, Vol.36, No.19, pp.2117-2124, 2004.

[13] Juho Lee, Keunhyuk Song, Hongjun Park, and Joonkeun Yum, “A Study on Determinants in Korean Pro-Baseball Spectators,” *Journal of the Korean Data Analysis*, Vol.12, No.6(B), pp.3507-3517, 2010.

[14] J. W. Meehan Jr., R. A. Nelson, and T. V. Richardson, “Competitive balance and game attendance in major league baseball,” *Journal of Sports Economics*, Vol.8, No.6, pp.563-580, 2007.

[15] Jang-Taek Lee and Hyun-Sik Cho, “An Analysis on the Home-Field Advantage in Korean Pro-Baseball with Logistic Regression Model,” *Journal of the Korean Data Analysis,* Vol.11, No.1(B), pp.533-543, 2009.

[16] William A. Young, William S. Holland, and Gary R. Weckman, “Determining hall of fame status for major league baseball using an artificial neural network,” *Journal of Quantitative Analysis in Sports*, Vol.4, Issue 4, Article 4, 2008.

[17] A. McCabe, and J. Trevathan, "Artificial intelligence in sports prediction," in *International Conference on Information Technology: New Generations*, pp.1194-1197, 2008.
[18] Seung-hoon 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.
[19] B. Karlik and A. V. Olgac, "Performance analysis of various activation functions in generalized MLP architectures of neural networks," *International Journal of Artificial Intelligence and Expert Systems*, Vol.1 No.4, pp.111-122, 2011.
[20] 2016 KBO Year Book [Internet], http://www.koreabaseball.com.
[21] Naver Sports [Internet], http://sports.news.naver.com. 2011.
[22] J. Han, M. Kamber, and J. Pei, "Data Mining: Concepts and Techniques," MA: Morgan Kaufmann, 2012.
[23] T. Hegazy, O. Moselhi, and P. Fazio, "Developing practical neural network applications using back-propagation," *Journal of Microcomputers in Civil Engineering*, Vol.9, No. 2, pp.145-159, 1994.
[24] H. Larochelle, D. Erhan, A. Courville, J. Bergstra, and Y. Bengio, "An empirical evaluation of deep architectures on problems with many factors of variation," in *Proceedings of the 24th International Conference on Machine learning*, ACM, 2007.
[25] A. Y. Ng., "Feature selection, L1 vs. L2 regularization, and rotational invariance," in *Proceedings of the 21st International Conference on Machine learning*, 2004.
[26] R. J. Hyndman, and B. K. Anne, "Another look at measures of forecast accuracy," *International Journal of Forecasting*, Vol.22, No.4, pp.679-688, 2006.
[27] Tamás D. Gedeon, "Data mining of inputs: analysing magnitude and functional measures," *International Journal of Neural Systems*, Vol.8, No.2, pp.209-218, 1997.

박 진 욱

http://orcid.org/0000-0003-0424-8225
e-mail : parkju536@yonsei.ac.kr
2016년 서울시립대학교 통계학과(학사)
2017년~현 재 연세대학교 컴퓨터과학과 석사과정
관심분야 : Big Data Mining & Machine Learning

박 상 현

http://orcid.org/0000-0002-5196-6193
e-mail : sanghyun@yonsei.ac.kr
1989년 서울대학교 컴퓨터공학과(학사)
1991년 서울대학교 컴퓨터공학과 (공학석사)
2001년 UCLA 컴퓨터과학과(공학박사)
2001년~2002년 IBM T. J. Watson Research Center Post-Doctoral Fellow
2002년~2003년 포항공과대학교 컴퓨터공학과 조교수
2003년~2006년 연세대학교 컴퓨터과학과 조교수
2006년~2011년 연세대학교 컴퓨터과학과 부교수
2011년~현 재 연세대학교 컴퓨터과학과 교수
관심분야 : 데이터베이스, 데이터마이닝, 바이오인포매틱스, Big Data Mining & Machine Learning