인공 신경망 모형을 이용한 한국프로야구 관중 수요 예측

박진욱*, 박상현*†
*연세대학교 컴퓨터과학과
e-mail: parkju536@yonsei.ac.kr

A Prediction of Demand for Korean Baseball League using Artificial Neural Network

Jinuk Park*, Sanghyun Park*†
*Dept. of Computer Science, Yonsei University

요 약

본 연구는 기존의 수요 예측 등의 시계열 분석에서 주로 사용되는 ARIMA 모형의 어려움을 극복하고자 인공신경망(Artificial Neural Network) 모형을 이용하여 한국 프로 야구 관중 수를 예측하였다. 인공신경망의 가장 기본적인 종류인 전방향 신경망(Feedforward Neural Network)의 초모수(Hyperparameter) 선정에 그리드 탐색(Grid Search)을 적용하여 최적의 모형을 찾고자 하였다. 훈련 자료로는 2015년 3월부터 8월까지의 일별 KBO 관중 수 자료를 대상으로 하였고, 예측력 검증을 위해 2015년 9월 관중 수를 예측하여 실제 관측값과 비교하였다. 그 결과, 그리드 탐색법에서 최적 모형이라고 판단한 모형의 예측력은, 평균 절대 백분율 오차(MAPE) 기준으로 평균 27.14% 였다. 또한, 앙상블 기법에서 착안하여 오차율이 낮은 모형 5개의 예측값 평균의 MAPE는 평균 28.58% 였다. 이는 다중회귀와 비교해보았을 때, 평균적으로 각각 14%, 13.6% 높은 예측력을 보이고 있다.

1. 서론

한국프로야구 관중 수는 1982년 140만 명을 시작으로 2015년에는 736만 명으로 성장하였다. 프로 스포츠 발전의 중요한 척도 중 하나는 관중 수이다. 관중 수는 구단의 경영측면과 직결되는 척도로써, 입장료부터 부대시설 이용료 등의 다양한 수입으로 연결되어 구단의 안정적인 재정 운영을 가능케 한다. 또한 수요예측은 기본적인 마케팅과 예산 전략 수립에 활용할 수 있다.

전통적으로 수요예측에 사용되는 방법은 Box Jenkins의 ARIMA 모형이다. 이와 관련된 선행연구들에서는 연간 누적 관중 수를 이용하여 시계열 모형을 식별하고 모수를 추정하였다 [2]. 또한 관중 수 자료 외에 독립변수를 추가적으로 설정하여 다변량 ARIMA 모형을 적용하여 예측값을 비교하였다 [3]. 이러한 선행연구들은 모두 연간 누적 관중 수를 이용하여 시간에 따른 변화를 파악하여 예측값을 도출하는 방법이다. 본 연구에서는 연간 누적 관중 수가 아닌 일별 관중 수를 수집하여 각 홈 경기장별로 일일 관중 수를 예측하는 것을 목적으로 하고 있다. 활용도 측면에서 일일 관중 수 예측은 연간 관중 수 예측보다 더 큰 효용을 가져올 것으로 판단하였다.

하지만 일별 관중 수 자료에 선행 연구들과 같은 전통적인 ARIMA 모형을 적용하는 데에 두 가지 문제점이 있었다. 첫 번째로, 한 홈 구장에서 발생하는 경기의 주기가 일정하지 않는다. 예를 들면, 롯데의 홈 구장인 사직구장의 2015년 4월 경기는 첫 주에는 금, 일요일에 있었다. 하지만 둘째 주에는 금, 토, 일요일이고, 셋째 주에는 화, 수, 목요일에만 존재하였다. 이는 주기적인 추세와 경향을 확인하는 ARIMA 모형에는 적용할 수 없는 문제이다.

두 번째로, ARIMA 모형은 과거 관측값과의 상관관계를 통해 분석을 진행한다. 따라서 시계열 자료에서 자기상관(Autocorrelation)이 존재해야만 한다. 위의 예의 사직구장의 자기상관 분석결과가 <표 1>에 나타나있다. 24시차까지 확인해 보았을 때, 유의 수준 0.05에서 자기상관이 존재하지 않는 다는 귀무가설을 기각하지 못한다. 즉, 과거 관측값이 현재의 관측값에 영향을 미치지 못한다고 판단할 수 있다. 이는 ARIMA 모형의 기반과 반대되는 결과로써, ARIMA 모형을 적용하지 못한다. 따라서 관측값이 각각 독립적인 자료라고 판단하고 독립성 모형인 인공신경망을 적합하는 것을 목표로 하였다. kt의 홈 구장의 경우에는 일부 자기상관이 존재했지만, 전체적인 흐름을 위해 인공신경망을 적용하였다.

※이 논문은 2015년도 정부(미래창조과학부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임.(NRF-2015R1A2A1A05001845)

† 교신 저자: sanghyun@yonsei.ac.kr

<표 1> 사직구장 일별 관중 수의 자기상관 검정

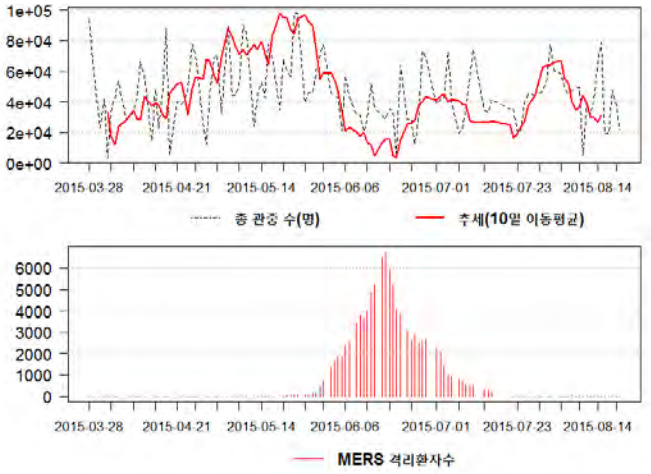
Lag	P-value
6	0.26
12	0.33
18	0.59
24	0.28

2. 분석자료 범위와 자료 수집

<표 2>는 인공신경망의 출력변수와 입력변수로 사용한 변수들을 나타낸다. 본 연구에서 사용한 출력 변수의 범위는 2015년 3월부터 9월까지의 한국 프로 야구 경기별 관중 수로써, 총 702개의 관측값을 가지고 있다. 관중 수 자료는 KBO에서 제공하는 2016년 연감[1]에 수록된 공식 자료를 기준으로 하였다.

인공신경망의 입력 변수 선정은 시간요소, 날씨요소, 지역요소, 팀별 특성 요소, 대중적인(Social) 요소로써 총 5개의 분야를 선택하였다. 시간요소는 날짜와 요일 변수, 그리고 공휴일인 경우와 성수기인 경우를 나타내는 이항변수를 생성하였다. 성수기 변수는 여름 휴가에 해당하는 7월과 8월을 1, 아닌 경우 0으로 표기하였다. 공휴일 변수도 마찬가지로 해당하는 날짜를 1로 생성하였다. 날씨요소는 기상청 데이터베이스의 일별 평균 기온과 평균 습도를 포함한다. 지역요소는 경기가 열리는 구장을 나타낸다. 즉, 팀별 구장으로써, 잠실(두산, LG), 청주(한화), 대전(한화), 광주(KIA), 수원(kt), 마산(NC), 목동(넥센), 문학(SK), 대구(삼성), 포항(삼성), 사직(롯데), 울산(롯데) 총 12개의 범주를 가지고 있다. 팀별 특성 요소에는 각 경기의 홈 팀과 어웨이 팀을 나타내는 변수와 홈/어웨이 팀의 존속 기간을 나타내는 변수를 포함한다. 또한, 같은 구장을 쓰는 LG와 두산의 경우, 라이벌 구단임을 표시하는 이항변수를 생성하였다. 경기 실력에 대한 변수로는 각 관측값별로 해당 일을 기준으로 그 전 날의 순위와 누적 승수, 누적 승률 변수가 존재한다. 관련 자료는 KBO 기록실을 참조하였다. 대중적 요소로는 Naver[4]에서 제공하는 각 경기 별로 진행하는 사전 투표와 응원 댓글 개수를 선정하였다. 또한 홈/어웨이 팀 별 변수와 총 투표, 댓글 수를 변수로 선정하였다.

추가적으로, 2015년도는 질병 MERS가 유행했던 시기이다. 그림 1은 관중수와 MERS 격리환자 수와의 상관관계를 나타낸다. 관중 수가 급 하강하는 5월 26일부터 7월 16일까지의 추세와 격리된 환자수와의 관계가 반비례하는 것을 확인할 수 있다. 또한, 관중수와 환자 수 자료의 상관도를 나타내는 피어슨 상관계수(Correlation coefficient)는 -0.53을 가진다. 이는 뚜렷한 음의 상관관계를 가진다고 판단할 수 있다. 따라서 MERS 첫 사망환자가 발생한 6월 1일부터 마지막 사망 환자가 발생한 7월 10일까지를 MERS 유행기간으로 판단하고, 이를 나타내는 이항변수 MERS (유행=1)를 생성하였다.



(그림 1) 관중 수와 MERS 환자 수와의 상관관계

<표 2> 입력변수와 출력변수의 선택

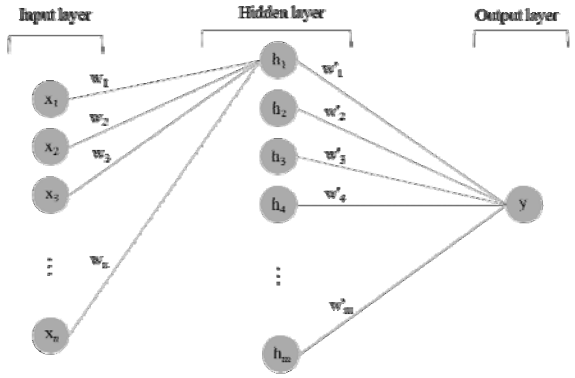
	영향 요소		
입력 변수	시간 요소	날짜	Date
		요일**	Day**
		공휴일*	Holiday*
		성수기*	Peak*
	날씨 요소	평균 온도	Temperature
		평균 습도	Humidity
	지역	경기장**	Stadium**
	팀별 속성 요소	홈 팀**	Home**
		어웨이 팀**	Away**
		홈 팀 존속기간	H.year
		어웨이 팀 존속기간	A.year
		LG-두산*	Rival*
		홈 팀 순위	H.rank
		어웨이 팀 순위	A.rank
		홈 팀 누적승수	H.cumWin
		어웨이 팀 누적승수	A.cumWin
		홈 팀 승률	H.cumWinRate
		어웨이 팀 승률	A.cumWinRate
	대중적 요소	홈 팀 댓글 수	H.comment
		어웨이 팀 댓글 수	A.comment
		총 댓글 수	Comment
		홈 팀 투표 수	H.vote
		어웨이 팀 투표 수	A.vote
		총 투표 수	Vote
		MERS 유행기간*	MERS*
출력 변수	경기별 관중 수		Y

*이항 변수(1, 0) **범주형 변수

3. 인공신경망 모형 설계

3.1 인공신경망 개요

그림 2는 인공신경망의 기본적인 모형인 전방향 신경망(Feedforward Neural Network)의 구조를 보여주고 있다. 입력 변수들을 나타내는 입력층(Input layer), 1개의 출력변수를 나타내는 출력층(Output layer), 은닉 노드들의 집합인 은닉층(Hidden layer)이 존재한다. 출력변수(y)와 입



(그림 2) 인공신경망의 구조

력변수의 경우, 변수들의 범위를 0과 1사이로 바꿔 주었다. 또한 범주형 변수의 경우, 더미 변수를 생성하였다.

각 층의 노드들은 다음 층의 모든 노드들과 연결되어 있고, 각각 가중치를 가지고 있다. 다음 층의 입력값은 이 가중치들을 이용한 가중합이 되고, 출력값은 활성함수(Activation function)을 거쳐 출력된다. 즉, 입력층을 제외한 각 노드의 출력값은 다음과 같이 결정된다. (수식 1)

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

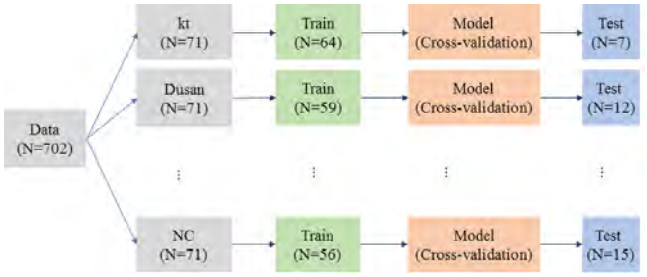
이 때, O_i와 O_j는 각각 이전 층의 i번째 노드와 다음 층의 j번째 노드의 출력값을 나타낸다. w_{ij}는 이전 층의 i번째 노드와 다음 층의 j노드 사이의 가중치이다. 또한 b_j는 j번째 노드가 가지는 편향(Bias)이다.

활성함수로는 선형함수(Linear function), Sigmoid, tanh, ReLU 함수 등이 존재한다. tanh 함수는 Sigmoid 함수의 문제점인 모형 훈련이 지연되는 점을 해결할 수 있고, ReLU는 tanh와 비슷한 성능을 지니면서 더 빠르게 수렴한다 [8]. 따라서 은닉층의 활성함수로 ReLU 함수를 선정하였다. 또한 출력변수의 값이 상수로써 의미를 가지는 회귀 모형을 훈련하는 것이므로, 출력층의 활성함수는 선형함수를 사용하였다.

인공신경망의 훈련은 오류역전파(Backpropagation) 알고리즘을 사용한다. 오류역전파 알고리즘은 목표값과 출력값과의 차이를 이용하여 오차를 줄여나가는 방향으로 가중치들을 갱신하는 방법이다 [5]. 이 때, 초기 가중치는 임의로 결정되므로, 모형 훈련에 영향을 미치게 된다. 따라서 본 연구에서는 이를 최소화하기 위해, 교차검증법(Cross-Validation)을 통해 여러 개의 모형을 만드는 훈련 절차를 적용하였다.

3.2 인공신경망 초모수 탐색 및 훈련

야구 경기장마다 최대 입장 가능한 관중 수가 다르므로, 인공신경망의 성능을 높이기 위해 데이터를 분할하였다. 이 때, 제 2구장에서 열리는 경기는 소수이므로, 경기장 기준이 아닌, 홈 팀 기준으로 데이터를 분할하였다. 분



(그림 3) 데이터 분할과 훈련 절차

할된 데이터에서 3월부터 8월까지는 훈련 데이터에, 9월 데이터는 테스트 데이터로 할당되었다. 모형의 초모수 탐색 및 훈련은 분할된 데이터에서 각각 진행하였다. 그림 3은 전체적인 데이터 분할과 훈련 절차를 나타낸다.

인공신경망은 모형의 성능을 결정하는 은닉층과 뉴런의 개수, 과적합을 방지하는 Regularization, 학습률, 활성함수 등의 초모수를 적절하게 설정해주어야 한다. 하지만 이를 위한 최적화 방법은 정해진 것이 없으며, 실험적으로 오차를 줄여나가는 시행착오법을 사용한다 [6]. 본 연구에서는 가능한 초모수 범위를 선정하여 오차가 최소가 되는 초모수를 탐색하는 그리드 탐색을 적용하였다. 모형 훈련을 위한 그리드 대상과 범위는 <표 3>에 나타나있다.

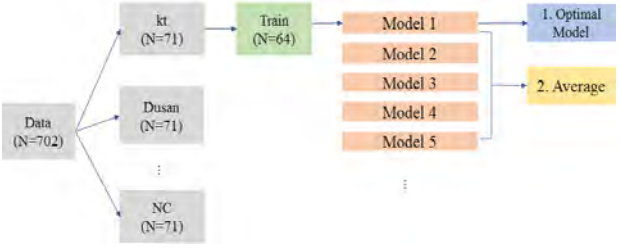
<표 3> 인공신경망의 초모수 탐색 대상과 범위

구분		범위
입력층*		36~40개의 입력 노드
출력층		1개의 출력 노드
은닉 뉴런 개수	은닉층 1개	25, 50, 75, 100
	은닉층 2개	4*4 = 16 (Possible combinations of above neurons)
l1 Regularization		0.0001, 0.005
학습률		0.05
은닉층 활성함수		ReLU
출력층 활성함수		Linear function
평가 기준		RMSE
Cross-Validation		5 Folds Cross-Validation
Epochs		30,000

*범주형 자료는 더미 변수를 생성함, 변화(Variation)가 없는 변수는 훈련 전 제거 (각 분할된 훈련 데이터의 범주에 따라 상이)

4. 결과

그리드 탐색을 수행을 통해 두 가지 예측값을 도출하였다. 첫 번째로, 가능한 모형 중 오차율이 가장 낮은 최적 모형을 선정하였다. 두 번째로는 오차율이 낮은 상위 5개의 모형 예측값의 산술평균을 통해 평균 예측값을 도출하였다. 그림 4는 두 가지 모형 선정 과정을 보여준다.



(그림 4) 최적 모형과 평균 예측값 도출 과정

4.1 최적 모형

<표 5>는 팀 별로 분할된 데이터에서 초모수 탐색을 통해 결정된 모형을 나타낸다. 최적화된 모형의 선택은 교차 검증된 RMSE(Root Mean Square Error)를 기준으로 하였다. 출력변수인 관중 수를 0과 1의 범위로 치환하였기 때문에, RMSE가 1보다 작은 값으로 계산되었다.

평균 절대 백분율 오차(Mean Absolute Percentage Error; MAPE)는 테스트 데이터(9월)를 이용하여 각 팀의 최적 모형을 테스트한 결과이다. MAPE란 예측값과 실제 관측값을 비교하였을 때, 각 관측값에 평균적으로 차이 나는 비율을 나타낸다. 수요 예측의 경우, MAPE를 모형의 예측력을 판단하는 기준으로 하여 직관적인 이해를 돕는다 [7]. MAPE를 계산하는 식은 다음과 같다.(수식 2)

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

4.2 평균 예측값

앙상블 방법은 여러 개의 약한 학습기(Weak learners)를 종합하여 더 좋은 학습기를 얻는 방법이다 [5]. 이를 위해, 오차율이 낮은 상위 5개 모형의 산술 평균 예측값을 계산하였다. <표 6>는 위의 최적 모형, 평균 예측값의 결과와 비교대상인 다중회귀분석의 결과를 나타낸다. 괄호안의 값은 다중회귀의 예측력 대비 향상된 비율이다.

최적 모형으로 선택된 전방향 신경망은 다중회귀 대비 평균적으로 14% 정도 높은 예측력을 보이고 있다. 하지만 롯데와 NC 구장의 경우는 다중회귀분석의 결과보다 안좋은 예측력을 보이고 있다. 또한 SK, KIA의 경우는 다중회귀분석보다는 예측력이 좋으나, 예측력이 높다고 할 수 없다. 이는 추가적인 모형설정이나, 이상치(Outlier)를 처리할 수 있는 방법이 필요하다는 것을 암시한다. 평균 예측값은 두산, 롯데, SK, 한화의 경우는 최적 모형보다 예측력이 낮아지고, 나머지 팀의 경우는 조금 높아졌다.

5. 결론

일별 야구 관중 수 자료는 선행연구와 달리, 주기적이지 않고 자기 상관이 없기 때문에, ARIMA 분석 절차를 적용할 수 없었다. 본 연구는 이러한 한계점을 극복하기 위해 전방향 인공 신경망을 활용하였다.

시간요소, 날씨요소, 지역요소, 팀별 특성 요소, 대중적인 요소의 입력변수를 가지는 인공신경망을 설계하고, 그리드 탐색법으로 최적의 모형을 선정하였다. 또한 오차율이 낮은 상위 5개 모형의 평균 예측값을 도출하였다.

<표 5> 팀 별 최적 모형 선택 결과

팀 구분	최적 뉴런	l1	RMSE(C-V)	MAPE(%)
kt	(25, 25)	0.005	0.1137	18.73
두산	(50, 25)	0.0001	0.1238	20.17
LG	(50, 50)	0.0001	0.1383	15.44
KIA	(25, 25)	0.005	0.1452	35.71
롯데	(100, 75)	0.005	0.2051	39.76
삼성	(50, 50)	0.0001	0.1876	20.10
SK	(50)	0.005	0.1626	42.44
한화	(50)	0.005	0.1925	26.45
넥센	(50, 50)	0.005	0.1920	29.95
NC	(50)	0.0001	0.2207	24.69

<표 6> 각 모형 예측값의 MAPE(%) 비교

팀 구분	최적 모형	5개 평균	다중 회귀
kt	18.73 (+24.4%)	16.80 (+32.2%)	24.79
두산	20.17 (+31.1%)	23.08 (+21.1%)	29.27
LG	15.44 (+41.8%)	15.20 (+42.7%)	26.51
KIA	35.71 (+31.7%)	34.89 (+33.2%)	52.26
롯데	39.76 (-29.4%)	40.26 (-31.0%)	30.73
삼성	20.10 (+38.0%)	16.24 (+49.9%)	32.44
SK	42.44 (+15.1%)	58.21 (-16.5%)	49.98
한화	26.45 (+11.7%)	28.38 (+5.2%)	29.94
넥센	29.95 (+4.3%)	29.20 (+6.7%)	31.31
NC	27.93 (-28.2%)	23.54 (-8.0%)	21.79
평균	+14.1%	+13.6%	

독립성을 가정하는 모형인 다중 회귀의 예측력을 비교하였을 때, 최적 모형은 평균 14% 뛰어난 예측력을 보였다. 상위 5개 평균 모형의 효과는 기대보다 미미했다.

본 연구결과는 선행 연구와 달리 인공신경망을 이용하여 경기장 별로 일일 관중 수를 예측했다는 것에 의미가 있다. 일일 관중 수 예측을 통한 마케팅 등의 활용도가 높을 수 있다. 아울러 영향을 주는 요인을 더 탐색하면 예측력을 높일 수 있을 것이라 사료된다. 또한 인공신경망의 초모수를 탐색하는 그리드 탐색의 초석을 다지고, 그 활용성을 넓혀야 할 것으로 판단된다.

참고문헌

[1] 2016 KBO 연감, http://www.koreabaseball.com, 2017.

[2] 김민철, "시계열분석을 통한 프로야구 관중현황 예측모형연구," *한국스포츠산업경영학회지*, 제14권 제1호, pp.17-25., 2009.

[3] 김형돈, "시계열모형을 이용한 프로야구 구단별 관중수 예측," *한국체육측정평가학회지,* 제14권 제3호: pp.57-68, 2012.

[4] 네이버 스포츠, http://sports.news.naver.com/kbaseball, 2017.

[5] Han, J,. Kamber, M., and Pei, J., *Data Mining: Concepts and Techniques,* MA: Morgan Kaufmann, 2012.

[6] Hegazy, T., Moselhi, O., and Fazio, P., "Developing practical neural network applications using back-propagation," *Journal of Microcomputers in Civil Engineering*, Vol.9, No.2, pp.145-159, 1994.

[7] Hyndman, R. J., and Anne B. K., "Another look at measures of forecast accuracy." *International journal of forecasting.* Vol.22 No.4, pp.679-688, 2006.

[8] Karlik, B., and Olgac, A. V., "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.