기계학습 기법을 이용한 한국프로야구 승패 예측 모델

A Win/Lose prediction model of Korean professional baseball using machine learning technique

저자 (Authors)
서영진, 문형우, 우용태
Yeong-Jin Seo, Hyung-Woo Moon, Yong-Tae Woo

출처 (Source)
한국컴퓨터정보학회논문지 24(2), 2019.2, 17-24(8 pages)
Journal of the Korea Society of Computer and Information 24(2), 2019.2, 17-24(8 pages)

발행처 (Publisher)
한국컴퓨터정보학회
The Korean Society Of Computer And Information

URL
http://www.dbpia.co.kr/journal/articleDetail?nodeId=NODE07617573

APA Style
서영진, 문형우, 우용태 (2019). 기계학습 기법을 이용한 한국프로야구 승패 예측 모델. 한국컴퓨터정보학회논문지 , 24(2), 17-24

Journal of The Korea Society of Computer and Information
Vol. 24 No. 2, pp. 17-24, February 2019
www.ksci.re.kr
https://doi.org/10.9708/jksci.2019.24.02.017

A Win/Lose prediction model of Korean professional baseball using machine learning technique

Yeong-Jin Seo*, Hyung-Woo Moon**, Yong-Tae Woo***

Abstract

In this paper, we propose a new model for predicting effective Win/Loss in professional baseball game in Korea using machine learning technique. we used basic baseball data and Sabermetrics data, which are highly correlated with score to predict and we used the deep learning technique to learn based on supervised learning. The Drop-Out algorithm and the ReLu activation function In the trained neural network, the expected odds was calculated using the predictions of the team's expected scores and expected loss. The team with the higher expected rate of victory was predicted as the winning team. In order to verify the effectiveness of the proposed model, we compared the actual percentage of win, pythagorean expectation, and win percentage of the proposed model.

▸Keyword: Win/Lose Prediction Model, Baseball Strategy, Machine Learning, Deep Learning, Baseball Data Analysis

I. Introduction

기계학습 기법은 컴퓨터가 수행해야 할 작업을 데이터로부터 스스로 학습하기 위한 기술이다[1]. 이러한 기계학습 기법은 문제의 규모가 크고 복잡하여 사람이 직접 단시간에 해결하기 어려운 문제 해결을 위해 유용하게 이용되고 있다[2]. 기계학습 기법은 인공지능이나 패턴인식 분야뿐만 아니라 스마트폰, 카메라, 소셜 네트워크 분야에서도 응용 분야를 넓혀가고 있다[3]. 최근에 기계학습 기법은 야구와 같은 스포츠 분야에서도 활발하게 이용되고 있다[4-14].

미국 메이저리그 피츠버그 파이어리츠 팀에서는 야구공 트랙킹 장비를 통하여 수집한 야구 빅 데이터를 기계학습 기법을 통해 분석하여 선수 영입에 이용하였다. 이 팀은 선수들의 특징을 기계학습을 통해 분석하여 잠재력이 뛰어난 선수들을 대상으로 장기계약을 체결하였다. 그 결과 20년 연속 승률 5할 미만이었던 피츠버그 파이어리츠 팀은 3년 연속 포스트 시즌에 진출하는 성과를 거두었다. 최근에는 대부분의 메이저리그 팀에서 기계학습을 통해 야구 경기 데이터를 분석하여 선수 영입이나 구종 예측과 같은 경기 전략 수립에 활발하게 이용하고 있다[4].

최근에 국내외에서 기계학습 기법을 이용하여 야구 데이터를 효과적으로 분석하기 위한 연구가 활발하게 진행되고 있다[5-10]. P. Hoang은 투수의 피칭 데이터와 기계학습을 이용하여 다음 피칭 구질을 예측하는 모델을 제안하였다[5]. David M. Hansen는 ANN(Artificial Neural Network) 모델을 이용하여 야구선수가 메이저리그 명예의 전당에 선정될 확률을 예측하였다[6]. Arlo Lyle는 기계학습 기법을 이용하여 득점, 안타, 2루타 등과 같은 야구 결과를 예측하는 모델을 제안하였다[7].

국내에서도 기계학습 기법을 이용하여 야구 경기 결과를 예측하기 위한 연구가 활발하게 진행되고 있다[8-10]. 채진석 등은 야구 데이터에 대한 판별 분석, 로지스틱 회귀분석, 인공신경망을 이용하여 포스트 시즌 진출팀의 승패를 예측하기 위한 모델을 제안하였다[8]. 오윤학 등은 데이터마이닝 기법과 야구

• First Author: Yeong-Jin Seo, Corresponding Author: Yong-Tae Woo
*Yeong-Jin Seo (yjseo@hiball.net), Technical Support Team, Hiball Inc.
**Hyung-Woo Moon (hwmoon@changwon.ac.kr), Institute of Industrial Technology Research Center, Changwon National University
***Yong-Tae Woo (ytwoo@changwon.ac.kr), Dept. of Computer Engineering, Changwon National University
• Received: 2019. 01. 31, Revised: 2019. 02. 25, Accepted: 2019. 02. 25.
• This research is financially supported by Changwon National University in 2017~2018.

데이터를 이용하여 한국프로야구 경기의 승패 예측 모형을 제안하였다[9]. 김종훈 등은 딥 러닝 기법을 이용하여 팀별 시즌 승률을 예측하는 모델을 제안하였다[10].

하지만 기존의 기계학습 기법을 이용하여 승패를 예측하기 위한 연구는 다음과 같은 문제점이 있다. 첫째, 기존 연구에서는 타율, 삼진 등과 같은 기본적인 야구 데이터를 이용하여 승패나 승률을 예측하였다[8-10]. 하지만 이러한 기본 데이터는 승패에 가장 큰 영향을 미치는 득점이나 실점과의 상관관계가 적은 관계로 정확한 승패 예측이 어렵다. 둘째, 한국 프로야구는 시즌동안 144경기를 진행하는 관계로 선수의 부상이나 부진이 팀 성적에 많은 영향을 미친다. 하지만 기존 연구에서는 시즌 전체를 대상으로 예측하는 관계로 최근 컨디션과 같은 요소들을 적절하게 반영하기 어렵다[8, 9]. 셋째, 야구는 상대에 따라 전력의 차이가 있을 수 있지만 기존 연구에서는 팀별 특성의 차이를 고려하고 있지 않다[9, 10].

본 연구에서는 기계학습 기법을 이용하여 한국 프로야구 경기의 승패를 예측하기 위한 모델을 제안하였다. 제안 모델은 크게 데이터 구성 및 전처리, 데이터 학습 및 예측, 기대 승률 계산 및 승패 예측 단계로 구성된다. 먼저, 데이터 구성 및 전처리 단계에서는 효과적인 승패 예측을 위해 기본적인 야구 데이터와 득점이나 실점과의 상관관계가 높은 세이버메트릭스 지표를 함께 이용하였다. 또한 팀별 특성을 반영하기 위해 다변량 검정 기법을 이용하여 데이터를 분석한 후, 타자와 투수를 팀별 특성에 따라 상세화하였다. 그리고 이동 평균법을 이용하여 팀의 최근 컨디션도 함께 고려하였다. 또한 Z-표준점수를 이용하여 예측 결과에 편향된 영향을 미칠 수 있는 이상치 데이터를 사전에 제거하였다.

학습 및 예측 단계에서는 지도학습 기반의 딥 러닝 기법에 의해 학습하였다. 그리고 학습된 모델 중에서 오류율이 가장 적은 모델을 예측 모델로 선정하여 득점이나 실점을 예측하였다. 마지막으로 기대 승률 계산 및 승패 예측 단계에서는 예상 득점과 예상 실점을 이용하여 기대 승률을 계산하고, 해당 경기의 기대 승률이 높은 쪽을 승리 팀으로 예측하였다.

본 연구에서 제안한 모델의 효율성을 검증하기 위하여 2006년부터 2011년까지 한국 프로야구의 경기 데이터를 이용하여 비교 실험하였다. 그리고 제안 모델의 예측 승률과 피타고리안 승률을 비교하였다[16]. 실험 결과, 제안 모델의 예측 승률이 피타고리안 예측 승률보다 실제 승률을 4.2% 더 정확하게 예측하였다. 따라서 본 연구에서 제안한 모델은 한국프로야구 승패 예측에 효과적인 모델이다.

본 연구에서 제안한 모델은 경기 전에 팀별로 승패 예측 결과를 제시하여 팬들의 흥미를 유발하여 한국프로야구의 저변 확대에 기여할 수 있다고 생각한다. 또한 구단에서는 당일 경기에 대한 승패를 예측하여 경기력 향상을 위한 선수 구성이나 효과적인 경기 전략 수립에도 활용할 수 있다. 앞으로 타자의 타구의 속도나 각도와 같이 승패와 관련된 다양한 데이터를 추가적으로 학습하여 보다 정확하게 승패를 예측할 수 있는 추가적인 연구가 필요하다.

본 논문의 구성은 다음과 같다. 2절에서는 딥 러닝 신경망에 사용된 기술과 세이버메트릭스 데이터에 대해서 소개하였다. 3절에서는 야구 경기 승패 예측 모델에 사용된 데이터 처리 기법과 예측 방법에 대해서 설명하였다. 4절에서는 야구 경기 승패 예측 모델에서 예상 득점과 예상 실점을 예측한 실험 결과를 분석하였다. 마지막으로 5절에서는 결론을 맺는다.

II. Preliminaries

1. Related works

1.1 Definition and use of machine learning

기계학습은 컴퓨터가 수행해야 할 작업을 데이터로부터 스스로 학습하기 위한 기술이다[1]. 특히 이 기법은 규모가 크고 복잡하여 사람이 직접 해결하기 어려운 문제를 해결하는 데 유용하게 이용되고 있다. 최근에 기계학습은 딥 러닝 기술의 개발, 빅데이터의 활성화 그리고 하드웨어의 발전으로 인해 처리 가능한 데이터의 범위가 급속도로 커지면서 다양한 분야에서 응용되고 있다[2].

최근에 기계학습은 야구, 축구, 농구 등과 같은 스포츠 분야에서도 활발하게 응용되고 있다[11-12]. Andrew D. Blaikie 등은 ANN(Artificial Neural Network) 모델을 이용하여 미식축구의 승패 예측 모델을 제안하였다[11]. Bernard Loeffel Holz 등은 PNN(Probabilistic Neural Network) 모델을 이용하여 미국 프로 농구 경기에서 승패를 예측하는 모델을 제안하였다[12].

1.2 Deep Learning Neural Network

딥 러닝 신경망은 기계학습 분야 중의 하나로 대량의 데이터를 학습하기 위해 다 단계의 신경망을 구성하여 데이터를 학습하는 기술을 의미한다[1]. 이 기법은 기존 신경망에 비해 다 단계의 신경망을 이용하여 학습 능력을 증가시킬 수 있는 장점이 있다[2]. 기존에는 데이터의 양이 소규모인 관계로 딥 러닝 신경망을 구성하는 기술이 활발하게 이용되지 않았다. 최근에는 데이터의 급격한 증가로 인해 고도의 학습 기술이 필요하게 되면서 딥 러닝 기법은 다양한 분야에서 활용되고 있다. 이 기법은 전통적인 인공지능이나 패턴인식 분야뿐만 아니라 스마트폰, 카메라, 소셜 네트워크 등과 같은 분야에서도 활발하게 이용되고 있다[3]. 그림 1은 딥 러닝 신경망의 구성도이다.

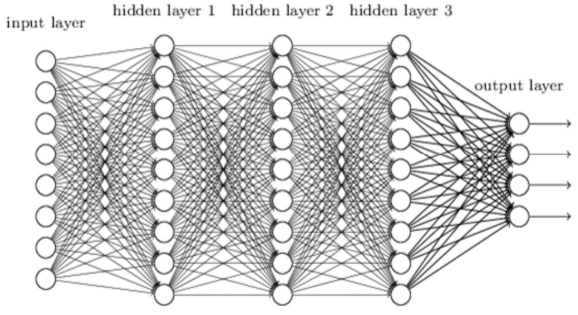
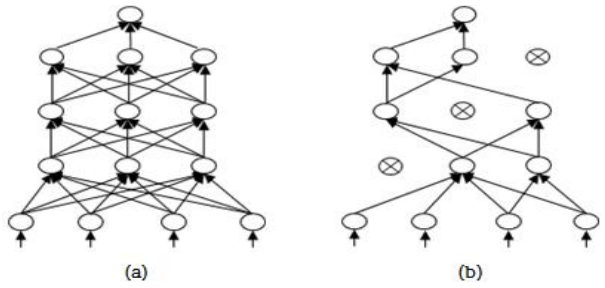


Fig. 1. Deep Learning Neural Network Model

1.3 DROP-OUT

DROP-OUT 알고리즘은 딥 러닝 신경망에서 과잉학습의 문제점을 개선하기 위해 제안된 알고리즘이다[13]. 이 알고리즘에서는 학습 과정에서 모든 노드를 이용하지 않고 일정한 확률로 노드를 제거하면서 학습을 진행하는 방식으로 과잉학습을 방지한다. 이러한 학습 방법에 의해 딥 러닝에서 학습의 안정성과 정확도를 향상시켰다[14]. 그림 2는 학습 과정에서 DROP-OUT 알고리즘을 적용한 경우와 적용하지 않은 경우를 비교한 그림이다[13].



Fig, 2. Compared using drop-out algorithm
(a) Before using drop-out algorithm,
(b) After using drop-out algorithm,

1.4 ReLu

ReLu 활성 함수는 딥 러닝 기법에서 학습량이 많아지는 경우에 발생하는 기울기 사라짐 문제를 개선하기 위하여 제안된 활성 함수이다[15]. 기울기 사라짐 문제는 주로 기존 활성 함수인 비선형 함수에 의해 발생한다. 그림 3은 기존의 sigmoid 활성 함수와 ReLu 활성 함수를 비교한 그림이다.

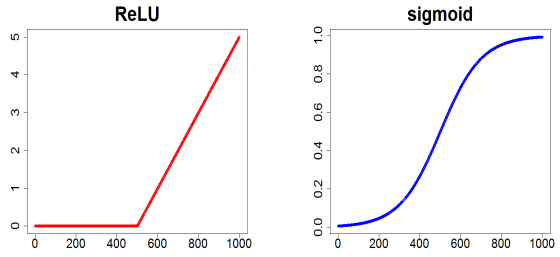


Fig. 3. Compared ReLU activate function and sigmoid activate function

기존의 sigmoid 활성 함수에서는 최적의 답을 찾기 위해 값을 미분하여 다음 과정으로 전달한다. 하지만 전달 과정에서 값이 너무 작아져서 최적의 답을 찾지 못하고 0이나 1로 수렴하게 되는 경우가 발생한다. 이러한 문제를 기울기 사라짐 문제라고 한다. 이러한 기울기 사라짐 문제 때문에 대량의 데이터를 학습하는 딥 러닝 기법의 이용이 제한적이었다. 최근에는 ReLu 활성 함수를 이용하여 기울기 사라짐 문제를 효과적으로 개선하였다.

1.5 Sabermetrics

미국 메이저리그 오클랜드 애슬레틱스 팀에서는 세이버메트릭스 기법을 이용하여 선수를 평가하는 머니 볼 이론을 통해 큰 성과를 거두었다[4]. 이 팀은 세이버메트릭스 지표 중의 하나인 OPS가 야구 경기의 승패에 밀접한 관계가 있음을 분석하였다. 이를 이용하여 저평가된 선수들을 저비용으로 영입하여 메이저리그 최초로 20연승이라는 성과를 거두었다. 이러한 세이버메트릭스 기법은 통계적인 기법을 이용하여 야구 데이터를 객관적인 지표로 분석하기 위한 기법이다. 이 기법은 미국 메이저리그 발전을 위해 자발적으로 참여하는 세이버메트리션에 의해 다양한 지표들이 지속적으로 개발되고 있다.

Table 1. Correlation between sabermetrics data and team score

Data	Correlation
AVG	0.822
OBP	0.885
SLG	0.910
IsoP	0.805
OPS	**0.946**
wOBA	**0.942**
XR27	**0.948**

표 1에서처럼 AVG(타율), OBP(출루율), SLG(장타율), IsoP(순수 장타율) 같은 기본적인 야구 데이터보다 세이버메트릭스 지표인 OPS(출루율+장타율), wOBA(득점 기여도) XR27(득점 창출 능력) 데이터가 팀의 득점률과의 상관관계가 더 높았다. 본 연구에서는 기본적인 야구 데이터와 세이버메트릭스 지표를 함께 사용하여 한국 프로야구의 승패를 효과적으로 예측하기 위한 새로운 모델을 제시하였다.

III. Win/Loss Prediction Model

본 연구에서 제안한 승패 예측 모델은 크게 데이터 가공 및 전처리 단계, 학습 및 예측 단계 그리고 기대 승률 계산 및 승패 예측 단계로 구성된다. 먼저, 데이터 가공 및 전처리 단계에서는 다변량 분석을 이용하여 팀별 특성을 분석하여 타자와 투수의 특징에 따라 상세화하였다. 또한 데이터의 변동성을 줄이고 현재 팀의 컨디션을 반영하기 위해 이동 평균법에 의해 경기 데이터를 가공하였다[17]. 그리고 전처리 단계에서는 학습 과정에서 편파성을 일으킬 수 있는 이상치 데이터를 사전에 제거하였다[18]. 그림 4는 본 연구에서 제안한 한국 프로야구의 승패 예측 모델에 대한 개념도이다.

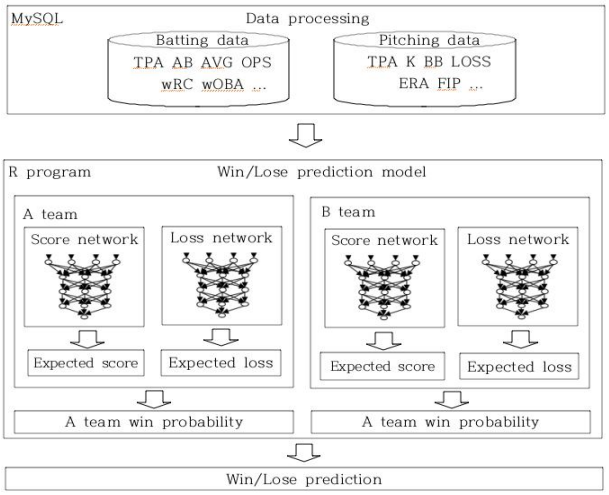


Fig. 4. Win/Loss prediction model

학습 및 예측 단계에서는 전처리 과정에서 구성된 데이터 대상으로 딥 러닝 기법을 이용하여 지도 학습 기반으로 학습하였다. 학습된 모델 중에서 오류율이 가장 적은 모델을 예측 모델로 선정하여 득점과 실점을 예측하였다. 그리고 예측된 득점과 실점을 이용하여 기대 승률을 계산하였다. 마지막으로 기대 승률을 이용하여 두 팀 중에서 기대 승률이 높은 팀을 승리 팀으로 예측하였다.

1. Data Preprocessing Stage

본 연구에서 제안한 모델의 효율성을 검증하기 위하여 iSTAT에서 수집한 한국 프로야구의 2006~2011년도 경기 데이터를 이용하여 실험하였다. 실험에서 사용한 경기 데이터는 타석, 타수, 안타, 홈런과 같이 야구 경기에서 발생하는 데이터이다. 경기 데이터를 학습용 데이터로 구성하기 위하여 MySQL 데이터베이스에 저장하여 이동 평균법으로 가공하였다. 팀별 특성을 고려하기 위하여 다변량 검정을 이용하여 데이터를 상세화하였다. 그리고 Z-표준점수를 이용하여 이상치 데이터를 제거하였다.

1.1 Moving average

본 연구에서는 데이터의 최근 추세와 데이터의 변동성을 줄이기 위하여 이동 평균법을 이용하여 데이터를 사전에 가공하였다. 이동 평균법은 시계열 데이터의 추세를 분석하기 위해 많이 이용하는 방법으로 해당 데이터 이전의 N 개 데이터의 평균을 이용한다. 이러한 이동 평균법은 데이터의 변동성을 줄이고 데이터를 평활하게 구성할 수 있다[17]. 식 (1)은 이동 평균값을 계산하는 식이다.

$$SMA = \frac{D_{n-1} + D_{n-2} \cdots D_{n-(n-1)}}{D_n} \quad (1)$$

$SMA =$ 이동 평균값
$D_n =$ 표본의 갯수

1.2 Outlier Remove

본 연구에서는 비정상적인 분포를 가지는 이상치 데이터를 제거하였다. 이상치가 포함된 데이터는 분석 결과의 편파성을 일으키는 문제가 발생할 수 있다[18]. 따라서 본 연구에서는 데이터를 학습하기 전에 Z-표준점수를 이용하여 이상치 데이터를 사전에 제거하였다[19]. 데이터를 표준화하여 Z-표준점수가 이상 범주를 벗어나는 데이터를 이상치로 판단하였다. 식 (2)는 이상치를 계산하는 식이다.

$$Z\text{-}score = \frac{D_i - \mu_D}{\sigma_D} \quad (2)$$

$D_i =$ 방어율
$\mu_D = D_i$의 평균값
$\sigma_D = D_i$의 표준편차

일반적으로 통계학에서는 Z-표준점수가 2.5점 이상을 벗어난 데이터를 이상치 데이터로 판단한다[20]. A팀에서는 2006 ~ 2011년까지 517건의 방어율 데이터 중에서 11건이 Z-표준점수 2.5 이상으로 분석되었다. 이에 따라, 11건의 경기 데이터는 이상치 데이터로 판단하여 학습용 데이터에서 제거하였다.

2. A Win/Lose prediction

2.1 Win/Lose prediction

본 연구에서는 승패를 예측하기 위하여 딥 러닝 기법을 이용하여 경기별로 각 팀의 득점과 실점을 예측하였다. 그리고 예상 득점과 예상 실점을 이용하여 기대 승률을 구하였다. 마지막으로 양 팀의 기대 승률을 비교하여 기대 승률이 높은 팀을 승리 팀으로 예측하였다.

2.2 Score prediction

먼저, 각 팀의 득점을 예측하기 위하여 팀별 타격 데이터를 학습하여 각 팀별 예상 득점을 예측하였다. 그림 5는 본 연구에서 예상 득점을 예측하기 위해 구성한 득점 예측 신경망의 개념도이다.

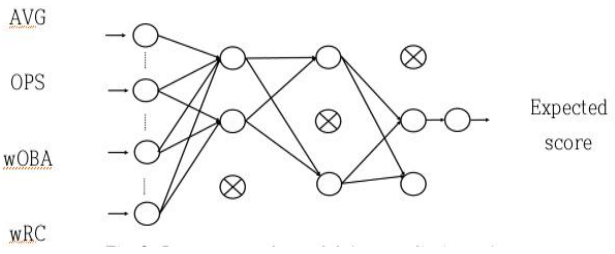


Fig. 5. Deep network model for prediction of score

2.3 Loss prediction

각 팀의 실점을 예측하기 위하여 팀별 수비력 데이터를 학습하여 각 팀별 예상 실점을 예측하였다. 그림 6은 본 연구에서 예상 실점을 예측하기 위해 구성한 실점 예측 신경망의 개념도이다.

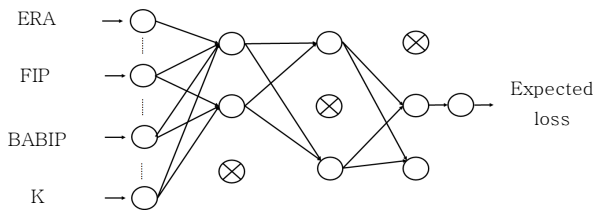


Fig. 6. Deep network model for prediction of loss

2.4 Win probability

본 연구에서는 승패 예측을 위해 딥 러닝 기법을 이용하여 경기별로 각 팀의 득점과 실점을 예측하였다. 그리고 예상 득점과 예상 실점을 이용하여 기대 승률을 구하였다. 마지막으로 양 팀의 기대 승률을 비교하여 기대 승률이 높은 팀을 승리 팀으로 예측하였다.

IV. Experimental Results

1. Experimental environment

본 연구에서 제안한 모델의 효율성을 검증하기 위하여 iSTAT에서 수집한 한국 프로야구의 2006 ~ 2011년도 경기 데이터를 이용하여 실험하였다. 실험에서 사용한 경기 데이터는 타석, 타수, 안타, 홈런과 같이 야구 경기에서 발생하는 데이터이다. 경기 데이터를 학습용 데이터로 구성하기 위하여 이동 평균법으로 가공하였다. 팀별 특성을 고려하기 위하여 다변량 검정을 이용하여 데이터를 상세화하였다. 그리고 Z-표준점수를 이용하여 이상치 데이터를 제거하였다. 학습 및 예측 실험은 R 프로그램의 H2o 패키지를 이용하였다.

2. Experimental results

2.1 Moving average experimental results

본 연구에서는 사용 데이터의 최근 추세와 데이터의 변동성을 줄이기 위하여 이동 평균법을 이용하여 데이터를 사전에 가공하였다. 표 2는 A 팀에 대한 타율 데이터를 이동 평균법으로 가공한 결과이다.

Table 2. A team data processing result using moving average

Game	Before game	Before game AVG	Result
July 7, 2011	July 1, 2011	0.280	**0.274**
	July 2, 2011	0.281	
	July 3, 2011	0.281	
	July 5, 2011	0.281	
	July 6, 2011	0.248	
July 10, 2011	July 5, 2011	0.281	**0.272**
	July 6, 2011	0.248	
	July 7, 2011	0.281	
	July 8, 2011	0.280	
	July 9, 2011	0.271	

2.2 Outlier remove experimental results

본 연구에서는 편향된 학습 결과를 유발할 수 있는 이상치 데이터를 Z-표준점수를 이용하여 사전에 제거하였다. 경기당 방어율 데이터를 Z-표준점수로 산출하여 표준 범위를 벗어나는 데이터를 이상치로 판단하였다. 표 3은 A 팀에 대한 방어율을 표준화하여 이상치를 가지는 방어율을 판단한 결과이다.

Table 3. A team era data convert to Z-score

Game	ERA	Z-score
July 12, 2006	13.5	2.58
August 15, 2006	13.5	2.58
July 29, 2007	**18**	**3.91**
September 13, 2007	14	2.73
April 17, 2009	14	2.73
May 14, 2009	14	2.73

일반적으로 통계학에서는 Z-표준점수가 2.5점 이상을 벗어난 데이터를 이상치 데이터로 판단한다[30]. 표 3에서처럼 2007년 7월 29일 경기에서 A 팀은 가장 높은 18의 방어율을 기록하였고, Z-표준점수도 가장 높은 3.91점을 기록하였다. A 팀에서는 2006 ~ 2011년도까지 517건의 방어율 데이터 중에서 11건이 Z-표준점수 2.5 이상으로 이상치 데이터로 판단되었다. 이에 따라 11건의 경기 데이터는 이상치 데이터로 판단하여 학습용 데이터에서 제거하였다.

2.3 Score/Loss prediction experimental results

표 4는 2011년 8월 2일 경기에서 6팀의 팀별 예상 득점과 실점을 예측한 결과이다. 표 4에서 팀 1의 득점은 해당 경기에서 예상 득점을 의미하고, 팀 2의 실점은 해당 경기에서 예상 실점을 의미한다. 예를 들어, 팀 C2의 예상 득점은 2011년 8월 2일 경기에서 가장 높은 7.47점으로 예측되었다. 그리고 팀 A1의 예상 실점은 2011년 8월 2일 경기에서 가장 낮은 1.57점으로 예측되었다.

Table 4. August 2, 2011 games score/loss prediction by team

Team 1			Team 2		
Team	Score	Loss	Team	Score	Loss
A1	1.63	**1.57**	A2	4.83	2.95
B1	3.60	5.65	B2	5.68	5.58
C1	3.52	7.82	**C2**	**7.47**	7.58

2.4 Win probability experimental results

본 연구에서는 예상 득점과 예상 실점에 의해 팀별로 승패를 예측하기 위하여 기대 승률을 이용하였다. 표 5는 2011년 8월 2일 경기에서 팀별로 기대 승률을 계산한 결과이다.

Table 5. August 2, 2011 games win probability by team

Team 1				Team 2			
Team	Score	Loss	Win probability	Team	Score	Loss	Win probability
A1	1.63	1.57	52%	**A2**	**4.83**	**2.95**	**82%**
B1	3.60	5.65	32%	B2	5.68	5.58	51%
C1	**3.52**	**7.82**	**23%**	C2	7.47	7.58	49%

표 5에서처럼 팀 1 득점은 해당 경기에서 팀별 예상 득점을 의미하고, 팀 2 실점은 해당 경기에서 팀별 예상 실점을 의미한다. 기대 승률은 해당 팀의 예상 득점과 예상 실점에 의해 계산한 승률을 의미한다. 팀 A2는 예상 득점과 예상 실점은 4.83점과 2.95점으로 예측되어 82%의 가장 높은 기대 승률을 기록하였다. 팀 C1의 예상 득점과 예상 실점은 3.52점과 7.82점으로 예측되어 23%의 가장 낮은 기대 승률을 기록하였다. 팀 A1은 가장 낮은 1.57점의 예상 실점으로 예측되었지만 예상 득점 또한 가장 낮은 1.63으로 예측되어 52%의 기대 승률을 기록하였다.

2.5 Win percentage prediction experimental results

표 6에서처럼 피타고리안 승률과 실제 승률의 차이는 평균 12.5%를 기록하였다. 그리고 제안 모델에서 예측한 승률과 실제 승률의 차이는 평균 8.3%를 기록하였다. 따라서 제안 모델이 피타고리안 승률보다 약 4.2% 더 정확하게 예측하였음을 알 수 있다.

Table 6. Compared the actual percentage of win, pythagorean expectation and proposed model

Team	Actual percentage	Pythagorean	Proposed model
HAN	30.0%	33.1%	20.0%
SAM	50.0%	55.2%	62.5%
LOT	71.4%	45.7%	71.4%
DOO	41.7%	43.6%	50.0%
NEX	60.0%	37.2%	50.0%
SK	54.5%	55.8%	45.5%
LG	62.5%	46.1%	62.5%
KIA	33.3%	57.0%	50.0%
Average difference percentage		**12.5%**	**8.3%**

그림 7은 실제 승률, 피타고리안 승률 그리고 제안 모델간의 예측 승률의 차이를 비교한 그림이다.

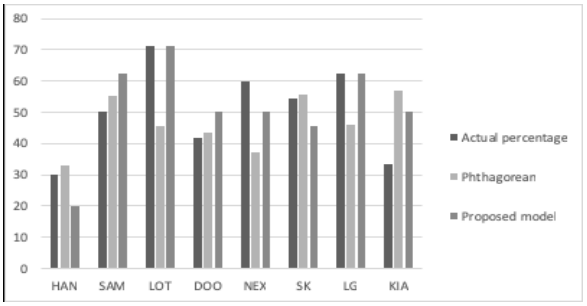


Fig. 7. Compared the actual percentage of win, pythagorean expectation and proposed model

그림 7에서처럼 제안 모델에서 롯데와 LG 팀의 승률을 가장 잘 예측하였다. 그리고 피타고리안 승률은 두산과 SK 팀을 가장 잘 예측하였다. 롯데, 넥센, KIA 팀의 피타고리안 승률과 실제 승률은 평균 24%의 차이를 보여 편차가 크게 나타났다. 하지만 제안 모델의 예측 승률과 실제 승률은 평균 8.3% 정도의 차이를 보여 피타고리안 승률보다 편차가 작았다. 따라서 본 연구에서 제안한 모델은 피타고리안 승률보다 보다 정확하게 승률을 예측함을 알 수 있다.

V. Conclusions

본 연구에서는 기계학습 기법을 이용하여 한국프로야구 경기에서 효과적인 승패 예측을 위한 새로운 모델을 제안하였다. 제안 모델은 기계학습에서 딥 러닝 기법을 이용한 새로운 형태의 승패 예측 모델이다. 제안 모델은 크게 데이터 구성 및 전처리, 데이터 학습 및 예측, 기대 승률 계산 및 승패 예측 단계로 구성된다. 먼저, 데이터 구성 및 전처리 단계에서는 효과적인 승패 예측을 위해 기본적인 야구 데이터와 득점이나 실점과의 상관관계가 높은 세이버메트릭스 지표를 함께 이용하였다. 또한 팀별 특성을 반영하기 위해 다변량 검정 기법을 이용하여 데이터를 분석한 후, 타자와 투수를 팀별 특성에 따라 상세화하였다. 그리고 이동 평균법을 이용하여 팀의 최근 컨디션도 함께 고려하였다. 또한 Z-표준점수를 이용하여 예측 결과에 편향된 영향을 미칠 수 있는 이상치 데이터를 사전에 제거하였다.

학습 및 예측 단계에서는 전처리 과정에서 구성된 데이터 대상으로 딥 러닝 기법을 이용하여 지도 학습 기반으로 학습하였다. 딥 러닝 기법의 문제점인 과잉 학습과 기울기 사라짐 문제를 보완하기 위하여 DROP-OUT 알고리즘과 ReLu 활성 함수를 이용하였다. 학습된 신경망에서 경기별로 팀의 예상 득점과 예상 실점을 예측한 결과를 이용하여 기대 승률을 계산하였다. 계산된 기대 승률을 이용하여 기대 승률이 더 높은 팀을 승리 팀으로 예측하였다.

제안 모델의 효율성을 검증하기 위하여 실제 경기 승률, 피타고리안 승률 그리고 제안 모델의 예측 승률을 비교 실험하였다. 실험 결과, 본 연구에서 제안한 모델이 피타고리안 승률보다 보다 정확하게 승률을 예측함을 알 수 있었다. 본 연구에서 제안한 승패 예측 모델은 경기 전에 팀별로 승패 예측 결과를 제시하여 팬들의 흥미를 이끌어 낼 수 있다. 또한 구단에서는 당일 경기에 대한 경기력을 예측하여 경기 전략 수립에 활용할 수 있다. 앞으로 투수의 투구 동작, 타자의 타구의 속도 등과 같은 야구 승패와 관련된 다양한 데이터를 수집하여 보다 정확하게 승패를 예측할 수 있는 추가적인 연구가 필요하다.

REFERENCES

[1] Injung kim “Deep Learning: New trend of machine learning”, *Journal of korea institute of communication sciences*, Vol. 31 No. 11, pp. 52-57, 2014.

[2] Sung Eun Moon, Soo Beom Jang, Jung Huk Lee, Jong Seok Lee, “Machine Learning and Deep Learning Technology Trends”, *Journal of korea institute of communication sciences*, Vol. 33, No. 10, pp. 49-56, 2016.

[3] Y. LeCun, Y. Bengio and G. Hin ton, “Deep learning”, Nature, Vol. 521, No. 7553, pp. 436-444, 2015.

[4] B.Bauer, A. J. Balis, “Sabermetrics Revolution”, Hanbit Biz , pp. 99-156, 2015.

[5] P. Hoang, “Supervised Learning in Baseball Pitch Prediction and Hepatitis C diagnosis”, Doctoral dissertation, North Carolina State University, 2015.

[6] David M. Hansen, “Introducing machine learning via baseball's hall of fame”, Journal of Computing Sciences in Colleges, Vol. 30, No. 4, pp. 7-15, 2015.

[7] Arlo Lyle, “Baseball Prediction Using Ensemble Learning”, Master's thesis, University of Tulsa, 2007.

[8] Jin Seok Chea, Kook Song Jong, “Comparisons of the Outcomes of Statistical Models Applied to the Prediction of Post-season Entry in Korean Professional Baseball”, Korean journal of sport science, Vol. 25, No.1, pp. 92-107, 2014.

[9] Oh Yun Hak, Han Kim, Yoon Jae Seop , Jong Seok Lee, “Using Data Mining Techniques to Predict Win-Loss in Korean Professional Baseball Games”, Journal of Korean Insitite of Industrial Engineers, Vol. 40, No.1, pp. 8-17, 2014.

[10] Jong Hoon Kim, Kyung Tae Kim, Jong Ki Han, “Big Data Analysis based on Deep Learning for Baseball Game Data”, Journal of korea institute of communication sciences, Vol. 2015, No.11, pp. 262-265, 2015.

[11] Andrew D. Blaikie, Gabriel J. Abud, John A. David, and R. Drew Pasteur, “NFL & NCAA Football Prediction using Artificial Neural Networks”, 2011 Midstates Conference for Undergraduate Research in Computer Science and Mathematics, 2011.

[12] Bernard Loeffel holz, Earl Bednar and Kenneth W Bauer, “Predicting NBA Games Using Neural Networks”, Journal of Quantitative Analysis in Sports, Vol. 5, No. 1, 2009.

[13] N Srivastava, G Hinton and A Krizhevsky, “Dropout : A Simple Way to Prevent Neural Networks from Overfitting”, JOURNAL OF MACHINE LEARNING RESEARCH, Vol. 15, No. 2, pp. 1929-1958, 2014.

[14] Hee Yul Choi, Yun Hong Min, “Understanding Dropout Algorithms”, Journal of Korean Institute of Information Scientists and Engineers, Vol. 33, No. 8, pp. 32-38, 2015.

[15] Vinod Nair, G Hinton, “Rectified Linear Units Improve Restricted Boltzmann Machines”, Proceedings of the 27th international, pp.807-814, 2010.

[16] Tom Tango, Mitchel Lichtman and Andrew Dolphin, “The Book: Playing the Percentages in Baseball”, CreateSpace Independent Publishing Platform, 2014.

[17] “MovingAverageMethod,” http://terms.naver.com/entry.nn?docId= 120434&cid=50304&categoryId=50304 (accessed June 1, 2017)

[18] “Outlier,” http://terms.naver.com/entry.nhn?docId=1924352&cid=42125&categoryId=42125 (accessed June 1, 2017)

[19] “Z-Score,” http://terms.naver.com/entry.nhn?docId=512343&cid=42126&categoryId=42126 (accessed June 1, 2017)

[20] Joseph F. Hair Jr, Willam C. Black, Barry J. Babin,Rolph E. Anderson, “Multivariate Data Analysis, 7th Edition,” Pearson, 2010.

Authors

Yeong-Jin Seo received the B.S. and M.S. degrees in Computer Science and Engineering from Changwon National University, Korea, in 2015 and 2017 respectively. Mr. Seo is a Developer in the Technical support team in Hiball Inc. since 2017. He is interested in Data Modeling and Baseball Data Analysis

Hyung-Woo Moon received the B.S. in Computer Engineering from Kosin University, Korea in 2007. He received the M.S. and Ph.D. degrees in Computer Engineering from Changwon National University, Korea, in 2009, 2014, respectively. Dr. Moon is a Researcher in the Institute of Industrial Technology Research Center, Changwon National University since 2014. He is interested in sports data analytics and sports big data architecture.

Yong-Tae Woo received the B.S., M.S. and Ph.D. degrees in Computer Science and Engineering from Kyungpook National University, Korea, in 1982, 1984 and 1995, respectively. Dr. Woo is a Professor in the Department of Computer Engineering, Changwon National University since 1987. He is also CEO of Hibrain.net Co. He is interested in Data Modeling, Internet Business, and Big Data Analysis areas.