석사학위 논문

인공신경망을 이용한 KBO 프로야구 승부예측 연구

A Study of KBO Professional Baseball Game Prediction using Artificial Neural Networks

2016년 12월

숭실대학교 소프트웨어특성화대학원

소프트웨어전공

노 언 석

석사학위 논문

인공신경망을 이용한 KBO 프로야구 승부예측 연구

A Study of KBO Professional Baseball Game Prediction using Artificial Neural Networks

2016년 12월

숭실대학교 소프트웨어특성화대학원

소프트웨어전공

노 언 석

석사학위 논문

인공신경망을 이용한 KBO 프로야구 승부예측 연구

지도교수 최 재 현

이 논문을 석사학위 논문으로 제출함

2016년 12월

숭실대학교 소프트웨어특성화대학원

소프트웨어전공

노 언 석

노 언 석의 석사학위 논문을 인준함

심 사 위 원 장 박 제 원 인

심 사 위 원 김 종 배 인

심 사 위 원 최 재 현 인

2016년 12월

숭실대학교 소프트웨어특성화대학원

목 차

표 목 차

그 림 목 차

국문초록

인공신경망을 이용한 KBO 프로야구 승부예측 연구

노언석

소프트웨어전공

숭실대학교 소프트웨어특성화대학원

프로야구는 해마다 관중이 증가하고, 또한 2012년 최초로 700만 관중을 돌파하였다. 2013년 9개 구단 체제에서 2015년 KT 위즈가 참여하여 10 구단으로 늘어나는 등 대한민국에서 가장 인기 있는 프로스포츠이다. 대중으로부터 많은 관심을 받는 스포츠이기 때문에 경기에서 생성되는 데이터를 바탕으로 많은 연구가 이뤄지고 있다. 기존 연구는 한국의 프로야구 승률을 예측하는 모델연구는 랜덤포레스트, 로지스틱회귀, 여러 분석기법으로 변인을 팀, 타자, 투수로 나눠 시행하였지만, 예측률이 떨어지고 각 변인의 통계적 유의성을 검증하는 데에 그쳤다. 본 논문에서는 KBO(Korea Baseball Organization) 프로야구경기의 승패 예측을 위해 선수들이 기록한 날짜별 데이터를 기반으로 인공신경망을 이용하여 경기를 예측하는 모델을 제시한다. 야구 경기 9이닝 중 선발 투수가 차지하는 비중이 높음을 고려하여 선발투수의 세부 기록과 나머지 투수들의 기록을 분리하여 적용하여 2014년부터 2016년까지의 데이터를 활용하여 실험하였다.

ABSTRACT

A Study of KBO Professional Baseball Game Prediction Model Artificial Neural Networks

Roh, Eon-seok
Major in Software
Graduate School of Software Soongsil University

Professional baseball is the most popular sport in Korea. Many studies are being made on the basis of data generated by the game. Existing model researches are using analysis techniques to predict baseball games such as random forest, logistic regression and so on.

In this paper, using artificial neural network based on the data recorded for the date players to predict the outcome KBO(Korea Baseball Organization) baseball game presents a model that predicts the match. Considering that the contribution of starting pitchers is higher than other pitchers, applied separated detail records of the starting pitchers and the rest of the pitchers. Data from 2014 to 2016 was used in this study.

제 1 장 서 론

1.1 연구배경

프로야구는 해마다 관중이 증가하고, 2012년 최초로 700만 관중을 돌파하였다. 2013년 9개 구단 체제에서 2015년 KT 위즈가 참여하여 10 구단으로 늘어나는 등 관중의 규모와 리그의 규모도 커지고 있다. 야구는 대한민국에서 가장 인기 있는 프로스포츠이다. 이러한 흥행과 더불어 일반인들은 자신이 좋아하는 팀의 승리예측이나, 승부에 대한 의견을 야구 전문 커뮤니티에서 나누거나 스포츠 토토, 프로토와 같은 투표권을 산다. 이처럼 프로야구의 승부예측을 하는 것은 일반적이다.

야구는 다른 어떠한 스포츠경기보다 많은 데이터를 쏟아내는 스포츠이다. 한 사람의 투구, 타격, 수비, 주루 동작들은 그 동작 그대로가 기록이 된다. 많은 기록을 쏟아내는 스포츠인 만큼 '빅데이터 분석'이 활발히 연구가 진행되고 있는 종목이다. 한 경기에서 쏟아지는 데이터가 많은 만큼 데이터를 기록하는 여러 인터넷 사이트들이 있고, 활발히 이용되고 있다. Sports2i (http://www.sports2i.co.kr)는 KBO 공식 기록 제공업체로 데이터를 제공하고 있고, Statiz (http://www.statiz.co.kr) 또한 한국 프로야구 데이터를 제공한다.

이러한 한국 프로야구 데이터들을 바탕으로 많은 연구가 진행되고 있다. 랜덤포레스트, 로지스틱회귀, 의사결정트리, 인공신경망 등의 알고리즘을 사용하여 승패예측모형수립, 세이버메트릭스를 활용하여 한국 프로야구 타자의 경기력 요인 분석한 연구, 마르코프 연쇄를 활용하여 득점예측을 하는 연구 등이 있다. 그동안 프로야구 승패를 예측하는 모델연

구는 많이 진행됐지만 기존의 연구들은 승패를 예측하기 위한 변인들의 유의성을 판단하는 연구와 당일 경기의 승패예측보다는 일정 기간의 승률예측 모델이 더 많았다.

본 논문에서는 실제 정확히 승패를 예측하기 위해 야구데이터 기록사이트인 스탯티즈(http://www.statiz.co.kr)에서 2014년부터 2016년까지의 날짜별 선수별 개인 데이터를 추출한다. 팀의 평균기록, 타자의 기록, 선발 투수의 기록, 선발 투수를 제외한 나머지 투수의 기록을 나누어 각 선수가 기록한 절댓값과 비율을 입력값으로 하고 홈팀과 원정 팀의 승패를 입력하여 인공신경망을 학습시킨다. 최적의 성능을 찾기 위해 입력 뉴런의 개수와 입력 변인을 조정하여 실험을 진행한다. 모델의 성능을 측정하기 위해 실제 경기를 예측하여 본다.

제 2 장 관련연구

2.1 기계학습

기계학습(Machine Learning)은 생물체에 있는 학습 능력과 같은 기능을 컴퓨터로 구현한 것으로 인공지능의 한 분야이다. 데이터를 분석하여 학습가능한 턴을 찾아 내어 새로운 데이터에 패턴을 적용하는 과정을 거친다. 빅데이터 시대의 도래와 컴퓨팅 능력의 향상으로 활용도가 높아지고 있다.[1]

기계학습은 지도학습(Supervised Learnig)과 비지도학습(Unsupervised Learning)으로 나눌 수 있다. 지도학습은 학습 데이터를 이용해 하나의 식을 유추하기 위한 방법이다. 이를 이용한 알고리즘으로 랜덤포레스트, 서포트 벡터 머신, 인공신경망, 회귀 분석등이 있다.

2.1.1 랜덤 포레스트

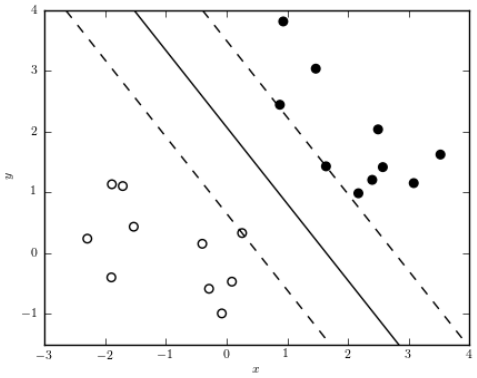
랜덤 포레스트는 분류기법 중 하나로, 의사결정나무를 바탕으로 하나의 나무가 아니라 여러 개의 나무로 확장시킨 여러개의 의사결정나무의 메타학습형태를 갖는 기계학습 기법이다[2]. 랜덤 포레스트를 구성하는 각 의사결정나무는 무작위로 추출된 학습 데이터와 입력 변수들에 의해 만들어진다. 개별의 의사결정나무의 정확도는 떨어질 수 있지만, 나무들을 종합해서 예측을 수행하게 되는 랜덤포레스트의 정확도와 안정성은 높아지게 된다. 랜덤 포레스트의 경우 큰 수의 법칙에 따라 의사결정나무의 개수가 많아질수록 일반화 오류가 특정한 값으로 수렴하게 되어 과적합

(Overfitting)을 방지할 수 있으며, 개별 의사결정나무들을 학습시킬 때 전체 학습 데이터셋에서 무작위로 복원 추출된 데이터를 사용하기 때문에 잡음(Noise)이나 이상치(Outliner)로부터 큰 영향을 받지 않는다. 랜덤 포레스트가 갖는 가장 큰 장점은 모형의 설계자가 입력 변수 선정에서 자유로울 수 있다는 점이다.

2.1.2 SVM

서포트 벡터 머신(Support Vector Machine)은 서로 다른 범주에 속한 데이터 간 간격이 최대가 되는 평면이나 선을 찾아서 이것을 기준으로 데이터를 분류해내는 모델이다.

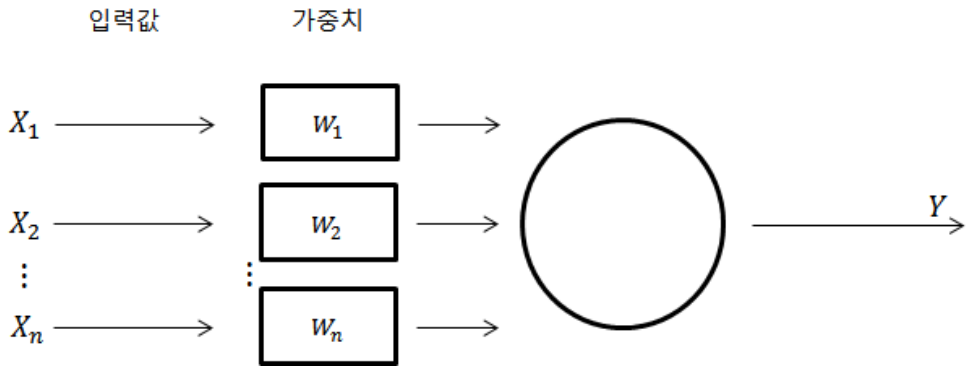
[그림2-1]은 서포트 벡터 머신의 개념으로 검은원과 흰 원은 서로 다른 분류를 뜻하며, 각 범주에 속하는 데이터로부터 같은 간격, 그리고 최대로 멀리 떨어진 평면 또는 선을 찾는다. 이러한 면 또는 선을 Hyperlane이라 한다. 이 Hyperlane으로 분류 할 수 없는 문제는 커널 트릭이라는 방법으로 문제를 해결한다.



[그림 2-1] 서포트 벡터 머신

2.1.2 인공신경망

인공신경망(Artificial Neural Network)은 인간의 두뇌와 신경 시스템을 닮은 정보 처리 소자이다. 인공 신경망 기법은 연결주의(Connectionism) 기법이라고도 하는데, 인간의 두뇌에서 정보처리를 담당하는 세포인 뉴런(Neuron)을 닮은 요소들을 연결하여 문제 해결 모델을 만드는 것이다 [3].



[그림 2-2] 인공뉴런의 구조

하나의 인공뉴런에 여러개의 입력값($X_1 \ldots X_n$)이 주어지면 각각의 가중치($W_1 \ldots W_n$)를 곱한 다음 이들을 모두 더하면 결과값 u가 된다. 이는 식 (1)로 나타난다.

$$u = \sum_{i}^{n} x_i w_i \qquad (3)$$

u값을 최종적으로 활성 함수 시그모이드 함수 f에 적용하여 뉴런의 출력값 y를 얻는다.

$$y = f(u) \qquad (4)$$

시그모이드 함수는 식(5)과 같다.

$$f(u)=\frac{1}{1+e^{-u}} \qquad (5)$$

또 다른 활성 함수인 쌍곡탄젠트 함수는 시그모이드 함수가 0과 1사이의 값을 갖는데 반해 -1과 1사이의 값을 갖는다.

시그모이드 함수 보다 학습 속도가 빠르고 시그모이드 함수는 출력이 1에 가까운 반면에 쌍곡탄젠트 함수는 0에 가깝다. 쌍곡탄젠트 함수의 식은 식(6)과 같다.

$$f(x)=\frac{e^{x}-e^{-x}}{e^{x}+e^{-x}} \qquad (6)$$

Softmax함수는 주로 출력층에 사용되는 활성함수로 인공 신경망의 출력값을 확률 벡터로 얻고자 할 때 사용된다. 최솟값이 0, 최대값이 1이 되도록 해주는 함수이다. 츨력 노드 k의 넷 활성화가 net_k이라고 할 때 식(7)과 같다.

$$\frac{e^{net_k}}{\sum e^{net_i}} \qquad (7)$$

2.1.3 기계학습 예측 연구

기계학습은 데이터를 이용하여 검증과 학습의 과정을 통해 특정 조건에서 예측값을 얻는 과정으로 과거의 추이를 통해 미래의 예측을 한다. 인공신경망, 랜덤포레스트, 로지스틱회귀, 서포트 벡터 머신(SVM)등 많은 알고리즘을 사용하여 예측 연구가 이뤄지고 있다.

김미경등이 시행한 인공 신경망을 이용한 새로운 전력 수요 예측 모델 연구는 인공 신경망 입력 변수로 시간과 날씨요소를 고려하여 변수간의 상관관계를 분석하여 온도와 이슬점등을 중요 요소로 고려하여 높은 적중률을 보였다[4].

유상록등은 시계열 분석과 인공신경망을 이용하여 해상 교통량 예측을 하였다. 인천항에 출입하는 선박의 톤급별로 최적합 모형과 환율, 인천항 물동량, 국제유가지수 등의 경제지표를 활용하여 해상교통량을 예측하였다[5].

김만식등은 단기 하천수질 예측에 인공신경망을 사용하였다. 상류로부터 유입되는 유입량, 수질인자인 BOD, COD, SS를 입력으로하여 분석하였다[6].

박선등은 인공신경망, 서포트벡터머신을 이용하여 적조 발생 예측을 하였다. 수온, 기온, 강수등 날씨 데이터와 생물밀도 개체수등을 사용하여 인공신경망과 서포트벡터머신 알고리즘을 비교하였다[7]. 정확도를 높이기 위해 다른 타 알고리즘과 비교하고 다른 기계학습 알고리즘과 함께 사용하였다. 이처럼 다양한 분야에서 결과나 미래를 예측하기 위하여 인공신경망을 활용한 연구가 활발히 진행되고 있다.

2.1.4 스포츠 예측 연구

스포츠 또한 승패를 예측하거나, 어떠한 상황이 벌어지는데에 대한 예측이 활발히 연구되고 있다.

김세형은 한국남자프로농구의 승패결정요인을 추정하기위하여 경기기록에서 2점슛, 3점슛등 공격 기록 7가지와 수비리바운드, 가로채기, 블록슛등 수비를 나타내는 기록 7가지를 사용하여 판별분석과 로지스트회귀분석을 통하여 승패 결정 요인을 추정하였다[8].

박해원등은 10가지의 남자 프로농구 기록을 변수로 독립표본 t검정과 로지스틱회귀 분석을 통하여 각 팀이 성공할 수 있는 성공변수를 판별해내었다[9].

김종훈등은 인공신경망을 이용하여 평균자책점, 타율, 피안타수 볼넷, 투구이닝, 삼진, 탈삼진, 승률, 홈어웨이 변인을 과거 데이터와 실제 경기 전 3경기 데이터를 입력값으로 주어 9월 6일부터 30일까지의 각 구단의 승률을 예측하였다. 각 구단의 승률과 예측승률이 3.39%의 차이를 보였다[10].

오윤학등은 랜덤포레스트, 로지스틱회귀, 의사결정트리, 인공신경망알고리즘을 사용하여 프로야구의 승패예측모형을 수립하였다. 변인을 타자, 투수, 팀요인으로 나누어 t검정을 실시하여 변인의 유의성을 판별하여 원시데이터, 나눔데이터, 이분데이터로 나누어 시행하였다. 랜덤포레스트가 이분데이터 일 때 오분류율이 15.86%으로 가장 우수한 알고리즘으로 판별하였다[11].

양도업등은 세이버메트릭스 지표를 활용하여 군집분석과 주성분회귀분석을 통해 한국 프로야구 타자의 경기력 요인 분석하였다[12].

이렇든 승패를 분석하는 데에 있어서 미치는 요인을 분석하고 승패를 예

측하는 연구가 많이 진행되었다[13].

본 논문에서는 야구에서 선발투수가 갖는 비중을 고려하여 선발투수와 나머지 투수들의 변인을 따로 두어 실험을 진행하였다.

2.2 빅데이터와 야구

2.2.1 세이버 메트릭스

야구는 한 경기당 약 1000개의 데이터를 쏟아낼 정도로 많은 데이터를 생성하는 스포츠이다. 하지만 이러한 데이터들은 치고, 달리고, 던지는 기록들이기 때문에 선수들이 경기에 미치는 척도를 제대로 평가하지 못한다. 이에 나온 것이 세이버메트릭스라는 것으로 미국야구연구협회(Society of American Baseball Research)와 메트릭스(metrics)의 합성오로 1980년대 초 빌 제임스(Bill James)에 의해 창안 되었다[14]. 다음은 잘 알려진 세이버메트릭스의 공식이다.

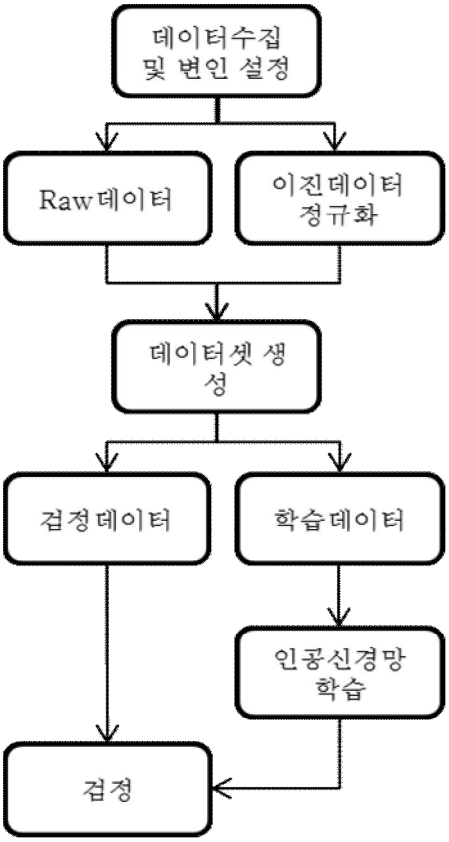
[표 2-1] 주요 세이버 메트릭스 지수

지수	설명
OPS	OPS(On base percentage Plus Slugging percentage) = 출루율 + 장타율 타자의 공격능력을 종합적으로 평가 할 수 있다.
GPA	GPA(Gross Production Average) OPS의 단점을 보완하기 위한 수치로 출루율에 1.8의 가중치를 준다
RC	RC(Run Created) 타자의 출루능력과 주자를 진루시키는 능력을 종합해 팀이 득점을 올리는데 있어서 어느 정도 기여했는지를 알려준다.
XR	XR(Extrapolated Runs) RC와 유사한 개념으로 팀 득점의 공헌도를 나타낸다.
SECA	SECA(Secondary Average) 타율측정에서 제외되는 볼넷과 사구 도루의 가치를 고려해 만든 지수이다.
BABIP	BABIP(Batting Average on Ballls In Play) 타자가 친 공이 페어지역 안에 떨어진 경우만을 나타내는 지수로 타자와 투수에게 모두 적용이 가능하다.
WHIP	WHIP(Walks plus Hits divided by Innings Pitched) 이닝당 안타 + 볼넷 허용률 투수들을 평가하는 기록중 하나

제 3 장 프로야구 승부 예측

3.1 프로야구 승부 예측 학습 과정

인공신경망을 이용한 승부 예측 과정은 [그림3-1]과 같다.
첫 번째 단계에서는 본 연구에서 필요한 데이터를 수집하고 변인을 설정한다. 데이터는 인터넷 웹사이트인 스탯티즈에서 제공한 2014년시즌 개막부터 2016년 시즌 종료까지의 경기일정으로 팀 데이터와 선수별 개인 데이터를 수집하였다. 팀 관련 변인은 승률, 상대 승률을 수집하였고 타자 관련 변인은 야구의 치고 달리는 기본데이터인 타율, 기록들을 타수 대비로 나타내었고 출루율을 사용하였다. 투수 관련 변인은 야구에서 선발투수는 1경기 9.0이닝 중 책임지는 이닝의 비율이 높으므로 투수 변인과 선발투수 변인을 나누어 적용하였다. [표3-1]은 생성한 변인을 나타낸다.
두 번째 단계에서는 인공신경망의 입력값에 적합하게 하기 위해 데이터를 Raw 데이터와 이진 데이터로 정규화한다. 세 번째 단계에서는 인공신경망 학습을 위해 학습데이터 70%와 검정데이터 30%로 나누어 학습데이터를 이용하여 활성화 입력 변인의 개수를 타입을 조절하여 실험을 진행하고 네 번째 단계에서는 학습된 인공신경망 모형의 성능을 검정데이터로 검정한다.



[그림 3-1] 프로야구승부 예측 학습 과정

[표 3-1] 생성한 변인

	변인	변인 생성
팀	승률	승수/무승부를 제외한 총 경기수
	상대 승률	승수/무승부를 제외한 총 경기수
타자	타율	안타/타수
	타수대비홈런	홈런/타수
	타수대비삼진	삼진/타수
	타수대비사구	타수/사구
	타수대비득점	타수/득점
	타수대비타점	타수/타점
	출루율	(사구+안타)/(타수+사구+희타)
	도루성공률	도루성공/도루시도
투수	평균자책점	(자책점*9)/던진이닝
	피안타율	피안타/타수
	피홈런율	피홈런/타수
	이닝당 삼진	삼진/이닝
	이닝당 볼넷	볼넷/이닝
	선발투수평균자책점	(자책점*9)/던진이닝
	선발투수WHIP	(안타+사구)/던진이닝
	선발투수피안타율	피안타/타수

본 논문에서 데이터를 수집하여 생성한 변인 데이터 셋은 2014년 부터 2016년 총 2037개이다. 이 중 70%를 학습에 사용하고 30%를 테스트 데이터로 사용하였다.

3.2 인공신경망 모델 적용

신경망의 활성화 함수 특성상 0과 1사이의 값이 되어야 하기 때문에 투수의 평균자책점과 선발투수의 평균자책점을 정규화 하였다.
투수의 평균자책점은 다음 식(8) 과 같이 0과 1사이의 값이 나오도록 투수의 평균자책점의 최대 값과 최소 값을 구하여 투수 평균자책점에서 최소값을 빼고 최대값과 최소값의 차로 나누어준다.

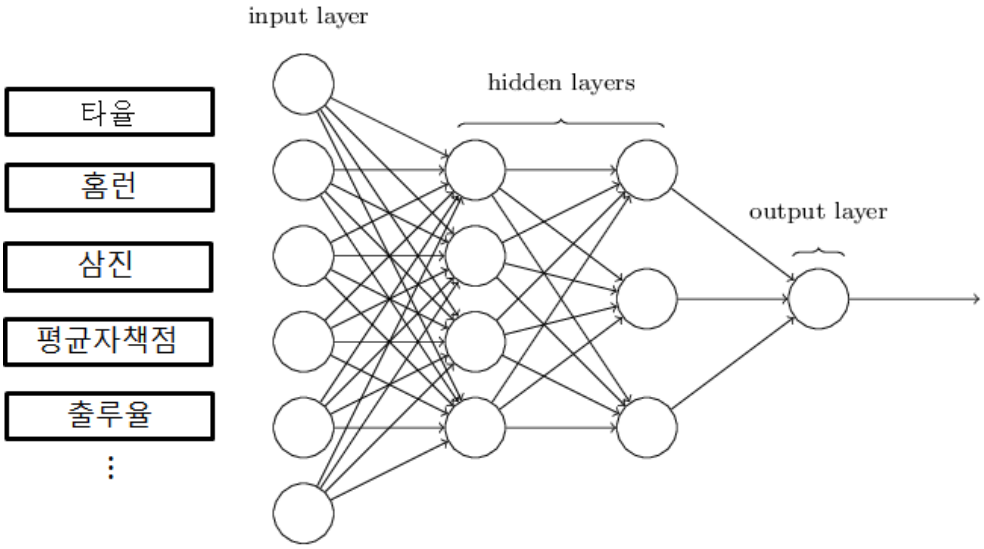
$$\frac{x-\min}{\max-\min} \qquad (8)$$

선발투수의 평균자책점은 2014년 선발투수로 10경기 이상 뛴 선수들의 자책점을 본 결과 3.5이하 이면 리그 최상위 투수, 4.2이하면 팀의 1선발 투수등의 구간이 나누어진다. 이 구간을 바탕으로 6개의 구간으로 나누어서 입력 값으로 정규화 하였다. [표 3-2]와 같이 구간을 나누었다. 또한 선발투수는 경기에서 9이닝중 5~6이닝을 책임지는 투수로 입력값에 가중치를 주었다.

[표 3-2] 선발투수의 평균자책점의 정규화

선발투수의 평균자책점 구간	입력 값
선발투수평균자책점 ≤ 3.5	0.85
선발투수평균자책점 ≤ 4.2	0.8
선발투수평균자책점 ≤ 4.6	0.7
선발투수평균자책점 ≤ 5.2	0.6
선발투수평균자책점 ≤ 6.0	0.5
선발투수평균자책점 ≤ 6.0	0.25

입력 변인을 인공신경망에 학습하여 출력 값이 1이면 홈팀 승리 0이면 어웨이팀 승리로 승패를 예측한다. 본 연구에서는 세 개의 은닉노드로 구성된 한 개의 은닉층으로써 모형을 구성하였다.



[그림 3-2] 인공신경망

제 4 장 실험 및 결과

4.1 실험 설계

본 연구에서는 2014년에서 2016년까지의 일자별 데이터를 바탕으로 다음 경기의 승패를 예측하는 것이기 때문에 스탯티즈에서 수집한 자료를 바탕으로 [표3-1]과 같이 변인을 생성하여, 자료를 생성하였다. 학습 데이터 생성의 예로 2014년 7월 9일 홈팀 SK 데이터는 [표4-1]과 같이 나타난다. 이와 같은 방법으로 2014년부터 2016년까지 총 2037개의 데이터 셋을 수집하였고 선발 투수 자책점, 투수 자책점을 정규화 과정을 통해 정규화하였다.

인공 신경망은 야구경기뿐만 아니라 다른 스포츠 경기의 승패 예측, 승률 예측 등의 목적을 위해 널리 사용되는 기계 학습 알고리즘 중 하나이다. 신경망은 독립변인의 입력층, 종속변인의 출력층, 그리고 은닉 노드(hidden node)들의 은닉층으로 구성된다. 실험의 입력층에서는 쌍 곡 탄젠트 활성함수를 사용하였고 출력층에서는 softmax 함수를 사용하여 정확도가 출력될 수 있게 실험하였다. 신경망의 과적합을 막기 위해 최적화 알고리즘으로 경사 하상식 알고리즘을 사용하였다. 또한, 모델의 정확도 향상을 위해 실험은 변인을 모두 사용하여 분석한 실험과 변인을 줄여 수행한 실험과 변인을 이진 데이터로 나타내어 실험하였다.

또한, 본 연구의 인공신경망 학습은 통계학에서 주로 사용되는 IBM SPSS 20을 사용하여 실험을 진행하였다.

[표 4-1] 변인에 따른 2014년 7월 9일 SK 생성 데이터

	변인	SK 데이터
팀	승패	1
	승률	0.408
	팀간 승률	0.4
타자	타율	0.285
	타수대비홈런	0.023
	타수대비삼진	0.213
	타수대비사구	0.119
	타수대비득점	0.157
	타수대비타점	0.151
	출루율	0.361
	도루성공률	0.77
투수	평균자책점	0.2236
	피안타율	0.288
	피홈런율	0.034
	이닝당 삼진	0.1036
	이닝당 볼넷	0.581
	선발투수평균자책점	0.25
	선발투수WHIP	0.141
	선발투수피안타율	0.288

4.2 실험 결과

[표 4-2] 실험 1의 데이터

분류				
표본	감시됨	예측		
		0	1	정확도(%)
훈련	0	411	253	61.9%
	1	314	480	60.5%
	전체 퍼센트	49.7%	50.3%	61.1%
검정	0	77	48	61.6%
	1	48	97	66.9%
	전체 퍼센트	46.3%	53.7%	64.4%

[표 4-3] 실험 2의 데이터

분류				
표본	감시됨	예측		
		0	1	정확도(%)
훈련	0	460	168	73.2%
	1	222	549	71.2%
	전체 퍼센트	48.7%	51.3%	72.1%
검정	0	116	42	73.4%
	1	46	118	72.0%
	전체 퍼센트	50.3%	49.7%	72.7%

[표 4-4] 실험 3의 데이터

분류				
표본	감시됨	예측		
		0	1	정확도(%)
훈련	0	501	137	78.5%
	1	150	614	80.4%
	전체 퍼센트	46.4%	53.6%	79.5%
검정	0	224	63	78.0%
	1	71	276	79.5%
	전체 퍼센트	46.5%	53.5%	78.9%

본 논문에서는 학습에 70% 검증에 30%의 데이터를 사용하였으며, 성능을 확인하기 위해 실험 결과를 바탕으로 분류 훈련과 검정 정확도를 확인하였다.

실험 1은 home팀 18개 away팀 18개 총 36개 모든 변인을 사용하여 신경망에 분석한 것으로 정확도 61.1% ~ 64.4%를 기록하였다.

[표 4-5] 실험 1 모형요약

훈련	교차 엔트로피 오차	958.463
	퍼센트 잘못된 예측	38.9%
	사용된 중지 규칙	오차 감소 없이 1 연속 단계
	훈련 시간	0:02:37.15
검정	교차 엔트로피 오차	170.338
	퍼센트 잘못된 예측	35.6%

실험 2는 정확도를 올리기 위해 인공신경망의 계산 복잡도를 낮추고 유사한 변인, 비중이 작은 변인 피안타율과 도루성공률 변인을 제외하고 실험하였다. 정확도가 72.1% ~ 72.3%를 기록하였다. 앞서 모든 변인을 사용한 실험 1보다 정확도가 약 8% 가량 상승하였다.

[표 4-6] 실험 2 모형요약

훈련	교차 엔트로피 오차	775.513
	퍼센트 잘못된 예측	27.9%
	사용된 중지 규칙	오차 감소 없이 1 연속 단계a
	훈련 시간	0:01:57.99
검정	교차 엔트로피 오차	180.980
	퍼센트 잘못된 예측	27.3%

실험 3은 변인을 home팀과 away팀의 변인을 비교하여 높은 값을 가지면 변인에 1, 낮은 값을 가지면 0, 같은 값을 가지면 0.5와 같이 값을 이진화 하여 변인을 가공하여 실험 하였다. 정확도가 78.8% ~ 79.5%로 실험 2보다 약 7%로 상승 하였고 실험 1보다 약 15% 상승하였다. 인공신경망의 계산속도도 다른 실험들보다 빠르고, 정확도도 우수한 것으로 나타났다.

[표 4-7] 실험 3 모형요약

훈련	교차 엔트로피 오차	718.402
	퍼센트 잘못된 예측	20.5%
	사용된 중지 규칙	오차 감소 없이 1 연속 단계a
	훈련 시간	0:00:01.47
검정	교차 엔트로피 오차	327.113
	퍼센트 잘못된 예측	21.1%

제 5 장 결론 및 향후 연구

야구는 대한민국에서 가장 인기 있는 프로스포츠이다. 해마다 관중의 규모와 리그의 규모도 커지고 있다. 이러한 흥행과 더불어 일반인들은 자신이 좋아하는 팀의 승리예측이나, 승부에 대한 의견을 야구전문 커뮤니티에서 나누거나 스포츠 토토, 프로토와 같은 투표권을 사는 것처럼 야구는 많은 사람의 흥미를 끌고 있다. 또한, 야구는 많은 데이터를 쏟아내는 스포츠인 만큼 '빅데이터 분석'이 활발히 연구가 진행되고 있는 종목이다. 본 논문에서는 KBO(Korea Baseball Organization) 프로야구경기의 승패 예측을 위해 선수들이 기록한 날짜별 데이터를 기반으로 인공신경망을 이용하여 경기를 예측하는 모델을 제안하였다. 개인이 날짜별로 기록한 기록을 바탕으로 팀 변인으로는 팀 승률, 팀 간 승률, 타자 변인으로는 타율, 홈런, 타점, 득점, 도루 등 공격 변인을 사용하였고, 투수 변인으로는 일반 구원투수와 선발투수가 갖는 비중이 달라 따로 변인을 두어 인공신경망 학습시켰다. 실험 결과 36개의 모든 변인을 사용한 실험 1의 정확도는 약 61%의 낮은 확률을 기록하였고, 유사한 변인, 중요도가 낮은 변인을 제외하여 실험한 실험 2는 약 72% 정도의 정확도를 기록하였다. 실험 3은 상대 팀과 변인을 비교하여 0과 1로 이진화 하여 변인을 가공하여 실험하여 정확도가 78.8% ~ 79.5%로 인공신경망의 계산속도도 다른 실험들보다 빠르고, 정확도도 우수한 것으로 나타났다.

본 논문에서는 승패 예측을 인공신경망만 만을 이용하여 분석하였지만 향후에는 변인의 조합에 따라 예측의 정확성을 올릴 수 있도록 유전 알고리즘과 같은 탐색알고리즘을 결합하여 사용하는 알고리즘을 같이 사용하면 더욱 우수한 예측 결과 연구가 가능할 것이다.

참고문헌

[1] 임수종, 민옥기. 빅데이터 활용을 위한 기계학습 기술동향,Electronics and Telecommunications Trends, 2012.

[2] Leo Brieman. “Random Forests”. Machine Learning, Vol. 45, issue 1, pp.5-32, Oct. 2006.

[3] 양기철. 인공지능 이론 및 실제. 홍릉과학출판사, 2014.

[4] 김미경, 홍철의(2016). 계절 및 날씨 정보를 이용한 인공신경망 기반 전력수요 예측 알고리즘 개발. 전자공학회논문지, 53(1), 71-78.

[5] 유상록, 김종수, 정중식, 정재용 (2014). 인공신경망과 시계열 분석을 이용한 해상교통량 예측. 해양환경안전학회지, 20(1), 33-41.

[6] 김만식, 한재석 (2002). 단기 하천수질 예측을 위한 신경망모형. 한국지반환경공학회논문집, 3(4), 11-17.

[7] 박선, 김경준, 이진석, 이성로 (2011). 신경망과 SVM을 이용한 적조 발생 예측. 전자공학회논문지-SP, 48(5), 39-45.

[8] 김세형 (2012). 한국남자프로농구 경기기록 분석을 통한 승패결정요인 추정방법 비교. 한국체육과학회지, 21(3), 1347-1360.

[9] 박해원, 김세형 (2016). 허용변수를 고려한 남자프로농구 승패결정요인 분석. 한국체육과학회지, 25(1), 1555-1565.

[10] 김종훈, 김경태, 한종기 (2015). Deep Learning 기반 기계학습 알고리즘을 이용한 야구 경기 Big Data 분석. 한국통신학회 학술대회논문집, 262-265.

[11] 오윤학, 김한, 윤재섭, 이종석. “데이터마이닝을 활용한 한국프로야구 승패예측모형수립에 관한 연구”. 대한산업공학회지 40(1), 2014.2,8-17

[12] 양도업, 조은형, 배상우, 정상원. “한국 프로야구 타자의 경기력요인 분석”. 한국사회체육학회지 60, 2015.5, 305-313.
[13] 채진석, 조은형, 엄한주. “프로야구 포스트시즌 진출 예측을 위한 통계적 모형 비교”. 한국체육측정평가학회지. 제12권 1호, 2010, 33-48.
[14] 양도업, 조은형, 배상우, 정상원. 한국 프로야구 타자의 경기력요인 분석. 한국사회체육학회지, 60, 305-313.