인공지능 환경에서 이닝별 데이터를 이용한 KBO 승패 예측

김태훈, 임성원, 고진광
순천대학교 컴퓨터공학과
e-mail : taehun@kakao.com, 2tjddnjs2@naver.com, kjg@scnu.ac.kr

KBO Win/Lose Predict Using Innings Data in AI Environments

Tae-Hun Kim, Seong-Won Lim, Jin-Gwang Koh
Dept. of Computer Engineering, Sunchon National University

요 약

과거 몇 년간의 데이터를 기반으로 현재 KBO 승패를 예측하고자 하는 것으로, 경기 초반 페이스가 얼마나 승패에 영향을 미치는지 파악하고자 한다. 경기의 이닝별 데이터로 딥러닝·머신러닝을 이용해 승리 팀을 예측하여 리그 순위를 예측하고, Flask 웹 프레임워크를 통해 입력값을 받아 예측해 주는 웹사이트를 구축하였다.

1. 서론

우리나라에서 가장 인기 있는 스포츠는 야구이다. KBO는 꾸준한 인프라 확장, 다양한 먹거리 문화, 응원 문화, 방송 플랫폼이 발달해 전 국민이 즐길 수 있는 대중적인 스포츠로 자리 잡았다. 야구는 기록에 따라 승패가 갈린다. 2002년 8월 13일부터 9월 4일까지 메이저리그 20연승을 기록한 오클랜드 애슬래틱스(Oakland Athletics)의 빌리 빈 단장의 세이버매트릭스를 시작으로 야구는 구단들 간의 데이터 싸움이 되었다.

본 논문에서는 9이닝까지가 아닌 1·3·5 이닝까지의 결과로 승패를 예측하는 딥러닝·머신러닝 모델을 구축하고, 결과들로 예측 순위를 만들고, 실제 승패와 비교하려 한다. 나아가 Flask 웹 프레임워크를 통해 입력값을 받아 구현한 모델을 거친 결과를 반환하는 웹페이지를 구현하려고 한다.

실험은 2016년 개막전부터 2020년 8월 30일까지의 KIA 타이거즈의 이닝별 데이터를 이용했고, 예측 결과표의 KIA 타이거즈가 아닌 구단끼리의 결과는 실제 결과로 사용하였다.

2. 이닝별 데이터를 이용한 KBO 승패 예측

2.1 데이터 수집 및 전처리

분석에 사용할 데이터는 파이썬을 이용한 셀레니움(Selenium) 프레임워크를 사용했다. 네이버 스포츠와 기아타이거즈 공식 홈페이지에서 2016년부터 2020년까지의 경기마다 이닝별로 데이터 크롤링을 수행했고, 통합하여 하나의 *.csv 파일로 저장했다. 수집한 데이터는 경기 날짜, 상대 구단명, KIA 타이거즈와 상대 구단의 이닝별 홈런/안타/삼진/볼넷/병살타/실책/획득 점수, 홈 원정 여부, 승패 여부다.

KIA 타자 기록

		1	2	3	4	5	6	7	8	9	10	11	12	타수	안타	타점	득점	타율
중	이창진	4구	유병			삼진		좌중2						3	1	0	1	0.351
二	김선빈	유땅		2땅		3땅		삼진						4	0	0	0	0.367
우	터커	중안		우중2		3땅			3땅					4	2	1	0	0.301
지	최형우	삼진		1땅			2땅		삼진					4	0	0	0	0.313
좌	나지완	좌안		3땅			우중홈		삼진					4	2	1	1	0.301
포	김민식		우안		중비		2땅			좌비				4	1	0	1	0.353
一	유민상		중안		중안		삼진			유땅				4	2	0	0	0.310
三	나주환		좌안		우비			중비		4구				3	1	1	0	0.266
유	박찬호		삼진		투땅			삼진						3	0	0	0	0.243
교	오선우									삼진				1	0	0	0	0.263
						합계								34	9	3	3	0.277

(그림 1) 네이버 스포츠 경기 통계

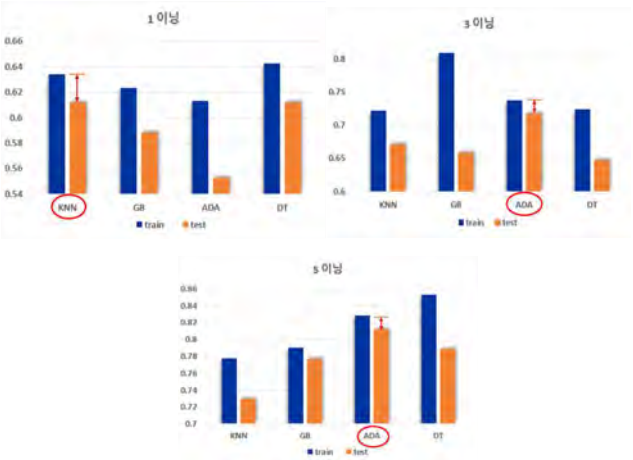
타자와 관련된 변인에서는 장타율이, 투수와 관련된 변인에서는 선발투수의 승률과 볼넷/이닝이 통계적으로 유의하게 나타났다.[1] 따라서 1·3·5이닝별로 SLG(장타율), OBP(출루율), BAT(타율), OPS(장타율+출루율)를 구하여 새로운 컬럼으로 넣고, 같은 의미를 나타내는 단어들로 라벨링하였다.

2.2 머신러닝·딥러닝 모델 구축

데이터 수집 및 전처리로 얻은 *.cvs 파일을 바탕으로 머신러닝·딥러닝 모델의 정확도를 비교하기 위해 2016년부터 2019년도까지의 데이터를 Train 데이터로 설정하고, 2020년도의 수집할 수 있는 날짜까지의 데이터를 Test 데이터로 설정했다.

머신러닝 모델로는 Adaboost, Decision Tree, KNN, Random Forest, Gradient Boosting 모델을 이용했고, 딥러닝 모델은 Keras의 Sequential 모델을 선정했다.

머신러닝 모델 구축 결과 1이닝에서는 KNN, 3이닝과 5이닝에서는 AdaBoost가 Train 정확도와 Test 정확도가 가장 적은 오차를 가지면서 높은 정확도를 보였다.



(그림 2) 머신러닝 모델 정확도

```
model=Sequential()
model.add(Dense(400,input_dim=24, activation='relu'))
model.add(Dropout(0.5))
model.add(Dense(200, activation='relu'))
model.add(Dropout(0.5))
model.add(Dense(100, activation='relu'))
model.add(Dropout(0.5))
model.add(Dense(20, activation='relu'))
model.add(Dense(10, activation='softmax'))
model.add(Dropout(0.3))
model.add(Dense(1, activation='sigmoid'))

# 모델 컴파일
model.compile(loss='binary_crossentropy', optimizer='adam', metrics=['accuracy'])

model.fit(X_train,y_train, epochs=100, batch_size=30, verbose=0)

# 트레인 셋에 모델 적용
print("Train Accuracy: %.4f"%(model.evaluate(X_train,y_train)[1]))
# 테스트 셋에 모델 적용
print("Test Accuracy: %.4f"%(model.evaluate(X_test,y_test)[1]))
```

```
21/21 [==============================] - 0s 2ms/step - loss: 0.3547 - accuracy: 0.8948
Train Accuracy: 0.8948
3/3 [==============================] - 0s 2ms/step - loss: 0.3530 - accuracy: 0.8941
Test Accuracy: 0.8941
```

(그림 3) 딥러닝 모델 정확도

신경망은 독립변인의 입력층, 종속변인의 출력층, 그리고 은닉 노드(Hidden Node)들의 은닉층으로 구성된다.[2] 실험의 입력층에서는 Relu 활성함수를 사용하였고 출력층에서는 Sigmoid 함수를 사용하여 정확도가 출력될 수 있게 실험했다.

(그림 3)의 딥러닝 Sequential 모델은 Train(89.48%)과 Test(89.41%)가 모두 높은 정확도를 보였다.

2.3 Flask를 이용한 웹페이지 제작

머신러닝·딥러닝 알고리즘으로 구축한 모델을 바탕으로 입력값을 받고, 결과를 반환하는 웹페이지를 구축했다. 머신러닝·딥러닝은 파이썬으로 구현했기 때문에 웹 프레임워크도 파이썬을 지원하는 Flask로 구현했다.

(그림 4) 웹페이지 값 입력 및 결과

머신러닝과 딥러닝 중 사용할 모델과 이닝을 선택하고, *.csv 파일에서 결괏값을 제외한 모든 값을 사용자에게 입력받아 서버로 전송하게 되면, Flask 파일에서 해당하는 모델에 입력값을 넣어 승·패를 반환한다. 반환된 값은 다시 웹페이지로 넘어와 기아와 상대 구단에 대한 결과를 이미지로 표시해 준다.

실제 데이터에서 SLG, OBP, BAT, OPS를 구하기 위해서는 타수(타자가 타석에 들어서서 타격을 완료

한 횟수)가 필요한데, 입력값을 받는 곳에서는 타수를 일정한 값으로 넣기가 어렵다. 따라서 SLG, OBP, BAT, OPS를 제외한 입력값을 통해 예측을 수행하였다.

3. 결론

본 연구에서는 경기의 이닝별 데이터로 딥러닝·머신러닝을 이용해 승리 팀을 예측하여 리그 순위를 예측하고, Flask 웹 프레임워크를 통해 입력값을 받아 예측해 주는 웹사이트를 구축하였다.

각 이닝별로 가장 오차가 작으면서도 정확도가 높은 모델을 이용해 승·패의 수를 구했고, 예측 순위표를 제작했다. 예측 결과표는 2020년 개막전(5월 5일)부터 8월 30일까지의 결과를 바탕으로 제작했다. KIA 타이거즈가 아닌 구단끼리의 결과는 실제 결과로 사용했다.

머신러닝 모델 중에서는 KNN과 AdaBoost가 가장 높은 정확도를 보였다. 실제 순위인 7위와 비교해본 결과 경기를 진행할수록 예측 결과 순위 오차가 작아지는 것을 확인하였다.

[1회 모델]

	예측순위	승	패	승률
1	키움	57	36	0.6129
2	NC	50	34	0.59524
3	KIA	47	38	0.55294
4	롯데	46	38	0.54762
5	두산	48	40	0.54545
6	KT	44	42	0.51163
7	LG	46	44	0.51111
8	삼성	44	45	0.49438
9	SK	29	60	0.32584
10	한화	27	61	0.30682

[3회 모델]

	예측순위	승	패	승률
1	NC	51	33	0.60714
2	키움	55	38	0.5914
3	롯데	45	39	0.53571
4	KIA	50	45	0.52632
5	LG	47	43	0.52222
6	KT	44	42	0.51163
7	두산	45	43	0.51136
8	삼성	42	47	0.47191
9	SK	32	57	0.35955
10	한화	27	61	0.30682

[5회 모델]

	예측순위	승	패	승률
1	NC	52	32	0.61905
2	키움	55	38	0.5914
3	두산	49	39	0.55682
4	KT	47	39	0.54651
5	LG	49	41	0.54444
6	KIA	44	41	0.51765
7	롯데	43	41	0.5119
8	삼성	42	47	0.47191
9	SK	31	58	0.34831
10	한화	26	62	0.29545

(그림 5) 머신러닝 모델 예측 순위표

(그림 6)의 딥러닝 Sequential 모델은 정확도를 89%까지 올렸고, 머신러닝 모델과 마찬가지로 경기를 진행할수록 예측 결과 순위 오차가 작아지는 것을 확인할 수 있다.

[1회 모델]

	예측순위	승	패	승률
1	키움	57	36	0.6129
2	NC	51	33	0.60714
3	두산	48	40	0.54545
4	롯데	45	39	0.53571
5	LG	48	42	0.53333
6	KT	44	42	0.51163
7	삼성	44	45	0.49438
8	KIA	42	43	0.49412
9	SK	31	58	0.34831
10	한화	28	60	0.31818

[3회 모델]

	예측순위	승	패	승률
1	키움	55	33	0.625
2	NC	52	32	0.61905
3	LG	51	39	0.56667
4	두산	48	40	0.54545
5	KT	46	40	0.53488
6	KIA	45	40	0.52941
7	롯데	44	40	0.52381
8	삼성	41	48	0.46067
9	SK	33	56	0.37079
10	한화	27	61	0.30682

[5회 모델]

	예측순위	승	패	승률
1	NC	52	32	0.61905
2	키움	55	38	0.5914
3	KT	49	37	0.56977
4	LG	51	39	0.56667
5	두산	48	40	0.54545
6	롯데	45	39	0.53571
7	삼성	47	42	0.52809
8	KIA	41	38	0.51899
9	SK	33	56	0.37079
10	한화	27	61	0.30682

(그림 6) 딥러닝 모델 예측 순위표

본 논문의 실험 결과는 KBO 승패 예측 뿐만 아니라 여러 분야에서 사용할 수 있을 것으로 사료된다. 야구를 중계하는 방송국에서는 경기 진행 중 이닝별로 승패 확률을 중계 화면에 보여 줌으로써 시청자들로 하여금 흥미를 일으킬 수 있을 것이다. 또한 추후 구단들이 각 구단별로 데이터를 분석하여 이닝별로 승리하기 위한 전략을 세울 수 있도록 활용할 수 있을 것이다.

"본 연구는 과학기술정보통신부 및 정보통신기획평가원의 Grand ICT 연구센터지원사업의 연구결과로 수행되었음" (IITP-2020-0-01489)

참고문헌

[1] Younhak Oh, Han Kim, Jaesub Yun and Jong-Seok Lee, "Using Data Mining Techniques to Predict Win-Loss in Korean Professional Baseball Games" Korean Institute of Industrial Engineersm", Vol. 40, No. 1, pp. 8-17, 2014.

[2] Eonseok No, "A Study of KBO Professional Baseball Game Prediction using Artificial Neural Networks", Thesis, p. 5, 2017.