순환신경망을 활용한 야구승부예측

정경석, 김진학, 한연희[1]
한국기술교육대학교 컴퓨터공학과
e-mail : {jks207, wlrgkr, yhhan}@koreatech.ac.kr

A Prediction of Baseball Game Results Using Recurrent Neural Netowrks

Kyeong-Seok Jeong, Jin-Hak Kim, Youn-Hee Han
*Dept. of Computer Science, Korea-University of Technology and Education.

요 약

최근 딥러닝(Deep-learning) 을 활용한 기상 예측, 심리 예측, 교통상황 예측 등 다양한 분야에 걸쳐 여러 모델의 인공신경망이 활용되고 있다. 본 논문에서는 여러 분야 중 스포츠라는 분야에 접근했으며, 딥러닝 모델을 통해 승부를 예측하는 실험을 진행하였다. 야구의 승부는 선수의 능력치, 기상의 변화, 홈/어웨이 여부, 교체 여부 등 가늠할 수 없이 수많은 데이터들에 의존하고 있다. 그러나 본 논문에서는 이러한 수많은 데이터 중 경기 외적인 데이터를 제외한 데이터를 활용하여 그 다음 경기의 승부를 예측할 수 있을 지를 연구한다. 날짜 별 경기들이 훈련데이터가 되고 목표는 이전 경기들의 영향으로 예측된 다음 경기의 승/패를 예측한다. 즉 순차적인 데이터의 활용에 적합한 모델, Recurrent Neural-Network 을 이용하였다. 이를 위하여 KBreport 에서 데이터를 수집하였고, 수집된 데이터를 훈련 데이터 세트로 만들어 Recurrent Neural Network 를 통해 훈련시켜 다음 경기의 승패를 예측하였다.

1. 서론

최근 딥러닝을 활용한 [1]기상 예측, 교통상황 예측, 물체 인식, 질병 예측 등 미래의 사건(Action) 을 예측하는 연구들이 진행되고 있다. 따라서, 본 논문에서는 여러 분야들 중, 스포츠를 선정, 더 들어가 데이터의 스포츠라고 일컫는 야구를 선정하여 이전 경기의 데이터를 토대로 다음 경기의 승부를 예측하였다.

본 논문에서는 데이터의 수집부터 딥 러닝의 모델에 적합한 Input 값으로 전처리, 딥 러닝 모델의 구동, 결과값 도출에 이르기까지 하나의 시스템에서 진행될 수 있도록 연구를 진행하였다. 또한 이러한 통합시스템에서 수집 - 전처리 - 분석 - 결과 도출 이 될 수 있도록 하여 기존의 분산된 작업 처리, 인력, 시간 등을 최소화하였다. 추가적으로, 주관적인 판단 하에 생길 수 있는 오차의 범위를 줄이려는 노력을 하였다.

이러한 도전적인 연구를 바탕으로 한 시스템은 윤리적인 문제에 부딪칠 수 있지만 야구의 관심을 더욱 이끌 수 있고 선수관리 시스템, 경기 환경조성에 변화를 줄 수 있으며 스포츠의 활발한 번영을 이끌 수 있다.

본 논문에서 가능한 한 수많은 팀들의 그 다음 경기들을 예측하기에는 승/패의 주체가 되는 팀의 경우를 모두 따져 보아야했다. 그러므로 범위를 좁혀 한 화의 팀만을 주체로 승/패를 예측해보는 연구를 하였다. 딥러닝의 모델은 RNN(Recurrent Neural Network) 을 활용하였다. 딥러닝의 모델에는 RNN 의 방식과 비교되는 CNN(Convolution Neural Network) 가 존재한다. 이 두가지 모델에서 RNN 을 선정하게 된 확실한 이유는 야구 승부 예측을 위한 훈련데이터의 형태때문이다. 훈련 데이터는 경기 별로 생성되며 경기는 시계열 데이터들이다. 이전의 경기들의 데이터가 예측해야되는 경기의 결과에 반영이 되어야함으로 적절하게 RNN 을 선정하였다. 어떠한 이미지를 훈련데이터로 주어 픽셀 값의 패턴을 통해 예측하는 CNN 보다 더욱 적절하다고 판단되었다.

딥러닝을 활용하기에 가장 중요한 것은 훈련데이터의 충분한 확보이다. 본 논문에서는 KBreport 사이트 (www.kbreport.com)에서 2012 년도부터 최근까지의 경기기록들을 크롤링하여 수집할 수 있었으며, 이러한 메타데이터들을 훈련데이터로 가공하여 예측을 진행하였으며 RNN 모델에서 도출된 값을 바탕으로 승, 무, 패의 확률들을 도출하였다.

[1] 교신저자: 한연희

2. 야구 데이터 수집 및 전처리

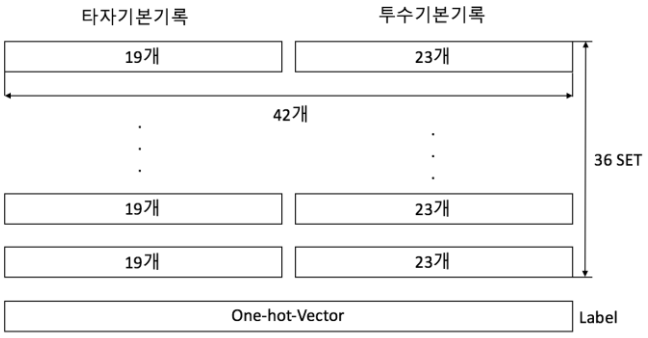
본 연구에서는 KBreport 사이트, 2012 년도부터 경기의 기본데이터들을 python 환경에서 크롤링하여 수집하였다. 크롤링한 데이터들은 직접적으로 데이터베이스에 저장하여 한번 수집된 데이터는 변경되지 않고 훈련데이터로서 활용된다. 야구의 승부에 가장 영향을 주는 요인을 선수(타자, 투수)로 선정하였고 타자의 기본 기록, 투수의 기본 기록을 바탕으로 다음 경기의 승부를 예측해보고자 하였다. 타자기본기록 19 개, 투수기본기록 23 개 (총 42 개) 를 바탕으로 수집을 진행하였다. 데이터는 다음과 같다.

<표 1> 타자 투수 기본기록 리스트

	기본 기록	자료형
타자 기록	TPA(타석)	INT(11)
	AB(타수)	"
	H(안타)	"
	1B(단타)	"
	2B(2 루타)	"
	3B(3 루타)	"
	HR(홈런)	"
	R(득점)	"
	RBI(타점)	"
	BB(볼넷)	"
	IBB(고 4)	"
	HBP(사구)	"
	SO(삼진)	"
	SF(희생플라이)	"
	SH(희생타)	"
	GDP(병살)	"
	SB(도루)	"
	CS(도루실패)	"
	AVG(타율)	DECIMAL(3,8)
투수 기록	W(승리)	INT(11)
	L(패배)	"
	SV(세이브)	"
	HLD(홀드)	"
	BS(블론)	"
	QS(퀄리티스타트)	"
	IP(이닝)	DECIMAL(3,8)
	PA(타자)	INT(11)
	H(안타)	"
	2B(2 루타)	"
	3B(3 루타)	"
	HR(홈런)	"
	R(실점)	"
	ER(자책점)	"
	SO(삼진)	"
	BB(볼넷	"
	IBB(고 4)	"
	HBP(사구)	"
	WP(폭투)	"
	BK(보크)	"
	PK(견제사)	"
	SB(도루)	"
	CS(도루실패)	"

위의 <표 1> 에서의 데이터는 수집한 데이터의 범주 리스트이다. 타자와 투수 기본 기록 내의 동일한 명칭은 주체가 되는 선수 분류에 따라 다른 값을 띄게 되므로 메타데이터들은 모두 독립적인 값이다.

현재까지 수집된 경기 수는 약 7542 개의 경기 수이며 한 경기 당 42 개의 데이터가 나오므로 각각의 메타데이터의 수로 나타내어진다면 총 7542 * 42 개가 된다. 메타데이터의 7542*42 개의 값들은 다음 순서의 딥러닝 모델의 입력 값에 쓰이기 위한 가치 있는 값들이 된다. 이 데이터를 바탕으로 RNN 모델에 사용되기 위한 데이터셋을 가공하였다. 42 개의 타자 투수기본기록들이 1 경기의 데이터이며 36 경기 묶음과 하나의 레이블이 쌍을 이루어 하나의 데이터셋을 생성시켰다. 구조는 다음과 같다.



(그림 1) 데이터셋의 구조

one-hot-vector 의 label 은 다음과 같이 분류된다.

<표 2> label 분류

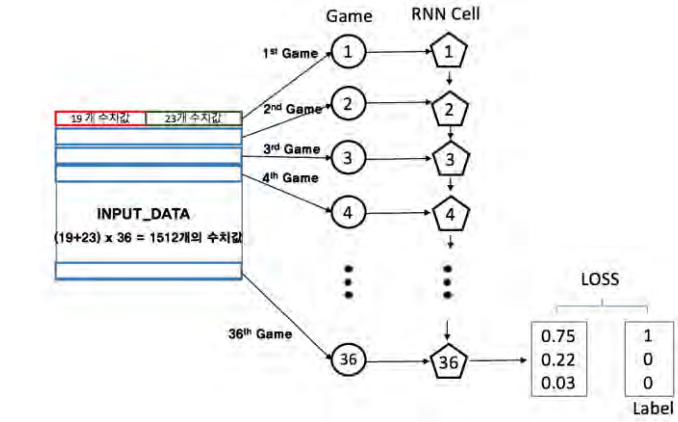
경기 결과	label
승	1 0 0
무	0 1 0
패	0 0 1

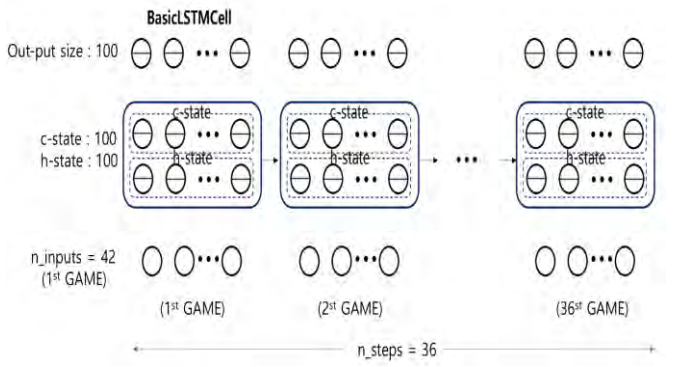
<그림 1>의 label 은 이렇게 3 가지로 정해진다. 본 연구에서는 이전의 36 경기들로 그 다음 경기를 예측할 수 있는지를 실험하기에 위 <그림 1>의 label 은 결국 다음 경기의 승부가 된다.

레이블데이터를 제외한 42 개의 기본기록들에서 영향력 있는 데이터와 영향력이 약한 데이터가 존재할 수 있다. 하지만 본 연구에서는 2012 년부터의 데이터를 사용하더라도 MNIST 의 훈련데이터셋의 개수보다 확실히 적었기때문에 모든 기록들을 훈련데이터셋에 포함시켰다.

3. RNN 모델

본 논문에서는 야구 승부 예측을 위한 훈련모델로 RNN 을 이용하였다. RNN 은 CNN 과 더불어 각광받고 있는 모델이다. [2] RNN 은 히든 노드가 방향을 가진 엣지로 연결되 순환구조를 이룬다. 음성, 문자 등 순차적으로 등장하는 데이터 처리에 적합한 모델이다. 추가적으로 RNN 은 시퀀스 길이에 관계없이 input 과 output 을 받아들일 수 있는 네트워크 구조이기 때문에 다양하고 유연하게 구조를 만들수 있다.





(그림 2) RNN Input-Data feeding

위 (그림 2)에서는 이전의 (그림 1)의 데이터셋을 RNN 모델에 feeding 시켜주는 과정을 보여준다.

기존 신경망 모형은 이전 데이터를 사용하여 예측한 결과값을 현재 예측에 사용하지 않는다. RNN 은 이를 보완한 것으로 예측에 현재 데이터 뿐만 아니라 이전 데이터를 사용한 결과값도 입력된다.

[3] 이론적으로, 기본적인 구조의 RNN 은 장기의존성을 다룰 수 있다. 하지만 입력데이터와 결과데이터의 거리가 멀수록 그런 능력이 떨어진다. 즉 이전의 데이터가 현재에 미치는 영향이 현저히 떨어진다.

이런 문제를 해결하고자 LSTM(Long Short Term Memory Networks)이라는 RNN 모델이 나오게 되었다. 해당 모델은 앞서 말했던 RNN 의 장기 의존성 문제를 해결하기 위한 모델이다. 단순히 tanh 로 셀을 계산하는 RNN 과 달리 LSTM 에는 망각 게이트, 입력 게이트, 출력 게이트가 추가 되어있다. 해당 게이트들로 인해 RNN 은 이전 데이터들을 기억하여 현재에 반영할 수 있게 된다.

본 논문의 승부예측모델은 시계열의 야구경기 데이터를 사용하고 있다. 따라서 해당 모델은 이전의 데이터가 잊혀지지 않고 계속해서 사용될 수 있는 LSTM 모델을 사용하였다.

4. 실험 및 평가

2 장에서 설명한 데이터셋들을 LSTM 에 넣어 훈련시켜보았다. 데이터의 수에 따라 모델의 성능이 변하는지 알아보기 위해 데이터셋의 가로길이(경기 당 데이터 수)와 세로길이(한 데이터셋에 포함된 경기 수)를 변경시키며 결과를 출력해 보았다. 기본 데이터셋 구조인 42*36 에서 변경시킬 것이며 예측하고자 하는 경기는 17 년도 9 월 13 일의 한화 경기이다.

우선 데이터셋의 가로길이를 10 배수로 변경해가며 모델의 정확도를 측정해 보았다. 가로에 포함되는 데이터 종류는 임의로 설정하여 데이터셋을 만들었다. 또한 마지막 가로길이 40 의 경우 현재 LSTM 에 넣을 수 있는 데이터셋의 최대 가로길이 42 와 큰 차이가 나지 않아 42 로 진행하였다.

<표 3> 데이터셋의 가로길이에 따른 정확도

가로길이	모델 정확도(%)
10	56.03(±2.0)
20	59.01(±2.0)
30	56.93(±2.0)
40(42)	57.71(±2.0)

대체적으로 데이터셋 안의 경기 당 데이터 수가 늘어날수록 모델의 정확도도 점차 오르는 것처럼 보인다. 유의해야할 점은 같은 구조의 데이터셋으로 모델을 훈련시켜도 5%의 정확도 차이가 나는 경우도 있기 때문에 데이터 개수가 모델의 정확도를 증가시킨다고 확언할 수 없다.

다음은 데이터셋의 세로길이를 18 배수로 변경해가며 모델의 정확도를 측정해보았다. 한 팀이 다른 팀과 2 번을 경기한 후 다음 팀과 붙기 때문에 9 배수가 아닌 18 배수로 변경하였다. 또한 위의 가로길이 실험이 유의미한지 알아보기 위해 가로길이 10, 40 에 대해 세로길이를 변경해가며 비교해보았다.

<표 4> 데이터셋의 세로길이에 대한 모델정확도

가로길이	세로길이	모델정확도
10	18	55.64(±2.0)
	36	56.97(±2.0)
	54	58.83(±2.0)
42	18	54.24(±2.0)
	36	57.84(±2.0)
	54	62.01(±2.0)

각 가로길이에 대해 세로길이 변경이 훨씬 증가추세가 눈에 잘 띄었고, 10 보다 42 일 경우가 더 많이

증가하는 것을 볼 수 있다. 즉, 모델의 정확도 증가는 세로길이가 많은 영향을 주지만 가로길이 또한 유의미한 영향을 준다고 볼 수 있다.

데이터셋의 기본크기가 42*36 일 때 가로세로길이가 늘어나면 가로길이는 10n*36 만큼 변하지만 세로길이는 18n*42 씩 늘어나 데이터셋의 크기가 더욱 커져 더 많은 데이터를 처리하기 때문일 수도 있지만, 데이터셋의 경기 수가 늘어 기존보다 더 과거의 데이터를 연관지을 수 있기 때문에 모델의 정확도가 점차 증가하는 것으로도 볼 수 있다.

5. 결과

훈련데이터를 통해 완성된 모델은 미래의 승부를 예측하였고 승부 예측은 17 년 8 월 1 일부터 10 월 3 일까지 진행하여 보았으며 결과는 다음과 같다.

<표 5> 모델 정확도에 따른 승부예측 일치율

정확도	56.2 (±2.0)	66.3(±2.0)
예측성공률 (일치횟수/총 경기)	27 / 48	32 /48

위와 같이 한화의 야구 승부를 예측하는 LSTM 모델의 정확도를 높이는데 세로 길이가 큰 영향을 끼침을 볼 수 있었다. 다만 현재로는 2012 년 이전의 경기 데이터를 구하지 못했고 지금 있는 제한된 데이터로 데이터셋의 세로 길이를 늘리면 총 데이터셋의 수가 줄어들어 모델의 정확도에 영향을 끼칠 수 있기에 무작정 데이터셋 안의 세로 길이를 늘릴 수는 없다. 따라서 앞으로 LSTM 모델의 정확도를 높이기 위해, 2012 년 이전의 경기데이터를 구하기 전까지는 데이터셋에 들어가는 최적의 세로 길이를 알아보고 데이터셋의 확장은 가로길이를 늘리는 것으로 진행할 것이다. 현재 가로 데이터 42 개만큼의 다른 데이터를 수집 중이고 그 외에도 아직 더 남아있는 야구 stat 뿐만 아니라 해당 경기의 날씨 정보나 구장 정보 등 다양한 데이터를 추가 수집 가능하며, 앞서 4 장에서 언급하였듯이 가로데이터수의 부족으로 인해 가로 길이를 늘리는 것의 결과가 좋지 않게 나온 것일 수도 있기 때문이다.

참고논문

[1] 김희언, 배태석, 신지민 “기상인자와 RNN 을 이용한 딥러닝 기반의 강수예측,” 한국측량학회 정기학술발표회 114-115, 2017.

[2] 이은주 “CNN 과 RNN 의 기초 및 응용연구” 방송과 미디어, 제 22 권, 제 1 호, 2017.1.

[3] 김양훈, 황용근, 강태관, 정교민 “LSTM 언어모델 기반 한국어 문장 생성” 한국통신학회논문지, 제 41 권, 제 5 호, 592-601, 2016.5.