Prediction of Korea Baseball Organization Rookie Draft Result through Deep Learning

Wonho Lim
Bugil Academy

Abstract

After COVID-19 struck the world, the fourth industrial revolution, especially Artificial Intelligence, came close to our life rapidly. Artificial Intelligence based on deep learning is used in various aspects of our life: weather forecast, traffic condition prediction, etc. This research approaches sports and conducts an experiment of predicting a KBO rookie draft result using a deep learning model. Because it deals with many complicated data, baseball using data has been in the spotlight in recent years. In this paper, high school baseball records are used to see if deep learning can be used to predict the KBO rookie draft nomination. The records of each Korean high school baseball player over the past five years are used as training data, and the results of the rookie draft are used as answer data. Based on this, the new player's nomination result is predicted. Data from the KBSA(Korea Baseball Softball Association) website is collected by using web crawling, and those data is converted into a training data set. Finally, the MultiLayer Perceptron model is used for prediction and evaluation. In conclusion, predicting a Korean Baseball Organization rookie draft result using deep learning based on high school baseball players' data was possible, with the accuracy of roundabout 75%. Later, other various data like physical information and school information can be collected and analyzed for improvement.

1. Introduction

Baseball using data has been in the spotlight in recent years. Baseball tactics based on analysis of data are much popular than tactics using senses and predictions. For example, these days, the teams' managers consider more detailed data such as the player's relative batting average rather than unconditionally use a right-handed pinch hitter against a left-handed pitcher. Moreover, recently, teams of the big baseball leagues, including MLB, NPB, and KBO, are pursuing efficiency by using data. In Korea, several teams are running data teams separately from performance analysis teams. Generating maximum benefit from efficient data utilization became essential for every team.

This trend is strongly affecting various aspects of Baseball, like match tactics and the rookie draft pick, which has not been changed for several decades. Especially, data analysis has become essential in bringing in suitable players for each club. There are seats for only 100 players in the Korean professional baseball rookie draft, even though thousands of high school students participate. Each club can select only ten players, and they struggle to pick the best ten future players for the club. Because scouts cannot observe thousands of players at once, clubs often rely on records to decide on draft picks. Due to these various circumstances, the results of the rookie draft are not easy to predict. However, it is crucial to analyze the rookie draft as it is a matter of deciding the future of the club and Korean Baseball. Therefore, this research aims to use deep learning to predict which player will be nominated from the KBO rookie draft.Every process, ranging from collecting data, processing input, running model, and gaining results, was designed to be completed in one system. Through this integrative system, dispersed processing time and meaningless work were decreased. Additionally, the range of errors was decreased in several ways.

The system made from this challenging research can conflict with several ethical issues as it predicts players' future. However, this research is worthwhile as it can make Baseball more attractive, cause a change in various aspects of Baseball, and lead the prosperity of the sports.

For accuracy, both players' records and the rookie draft's nomination result had to be collected while the condition was kept. Thus, data from only the recent five years was used for the prediction. MLP(MultiLayer Perceptron) was employed as a deep learning model.

The most important part of deep learning is having enough amount of training data. For this research, all the match records from 2015 to 2020 on the KBSA website (https://www.korea-baseball.com/) are collected. That collected metadata was converted into a training data set for the prediction. Finally, based on the value derived from the deep learning model, the rookie draft nomination probability was deduced.

2. Related Research

Several pieces of research that used machine learning for prediction was investigated before this research was started. They showed that players' record data are useful for making various predictions.

At "의사결정나무 분석 기법을 활용한 프로야구 외국인 투수 재계약 확률 예측(A study on KBO league foreign pitchers' re-sign possibilities using decision tree analysis)," Decision Tree and Random Forest, which are machine learning models, are used to figure out which record

increases the contact extension probability of foreigner pitchers of KBO. The researcher collected various pitcher records to make a model so that the random forest can have many decision trees with a smaller error range. The result shows that ERA is the most significant factor that determines the contract extension.[1]

"선형회귀분석기법을 이용한 고교야구투수의 투구속도 예측(High-School Baseball Pitcher's Pitching Speed Prediction Using Linear Regression Analysis Method)" deals with the linear regression method that can be used to predict the fastball speed of high school baseball player. TensorFlow, a deep learning framework, is used in this research. The researcher generated a linear regression function using training data. The Gradient Descent algorithm is used to improve the accuracy of the model. Finally, the researcher found that the pitcher's fastball speed increases as the stride of the pitcher increases.[2]

"DATA ANALYTICS IN SPORTS: IMPROVING THE ACCURACY OF NFL DRAFT SELECTION USING SUPERVISED LEARNING" uses various Supervised Learning Classification methods and algorithms to make more accurate NFL draft result prediction. The researcher trained Naive Bayes, Multiplayer Perception, Logistic Regression, RBF Network with the draft result and the post year record. Then, the researcher checked the accuracy and predicted the outcome. In conclusion, the researcher concludes that machine learning can be used to predict the probability of the picked player's success in the NFL.[3]

3. Baseball Data Collection and Processing

For this research, data from 2015 to 2020 on the KBSA website was crawled through Python. Crawled data was saved as a CSV file so that it could be used as training data without any additional process. Players' position and the record were considered factors that affected the draft result. Player's position, 20 hitting records, and 14 pitching records (total 35) were collected. The draft result was predicted based on these records. Below is the list of the data.

<Table 1> Player Record Result

	Basic Record	Note
Position	P(0),IF(1), C(2), OF(3)	
Hitting Record	AVG	Real Number
	Games	
	TPA	
	At Bats	
	Run	
	#Hit	
	Doubles	
	Triples	
	Homerun	
	Total Bases	
	RBI	
	Stolen Bases	
	Sacrifice Fly	
	Sacrifice Hit	
	BB	
	Strike Out	
	GIDP	
	SLG	Real Number
	OBP	Real Number
	OPS	Real Number
Pitching Record	ERA	Real Number
	Games	
	Win	
	Lose	
	Innings	Real Number
	#of Hitters	
	Hits	
	Home Runs	
	BB	
	Strike Out	
	Runs	
	Earned Runs	
	Win%	Real Number
	WHIP	Real Number

<Table 1> is a category of data that was collected. As long as the record with the same name will have different values depending on the classification of players, collected metadata are all independent values.

The number of collected player is about 4000, and there are 35 data per player. So, the number of total metadata would be about 4000 * 35. Dataset for a deep learning model was generated based on this data.

Labels of one-hot-vector for the answer is classified as below.

<Table 2> label classification

Prediction Result	label
Nominated	1
Unnominated	0

Answer data was collected by searching for recent five year's rookie draft results. The influence of each 35 data except answer data might vary. However, because there was a smaller number of training datasets than MNIST has, all the

황규인 Kyu-in Hwang (2018). "의사결정나무 분석 기법을 활용한 프로야구 외국인 투수 재계약 확률 예측. A study on KBO league foreign pitchers' re-sign possibilities using decision tree analysis. 국내석사학위논문, 고려사이버대학교 융합정보대학원. In partial Fulfillment of the Requirements for the Master Degree, Korea Cyber University Graduate School of Interdisciplinary Information Studies.

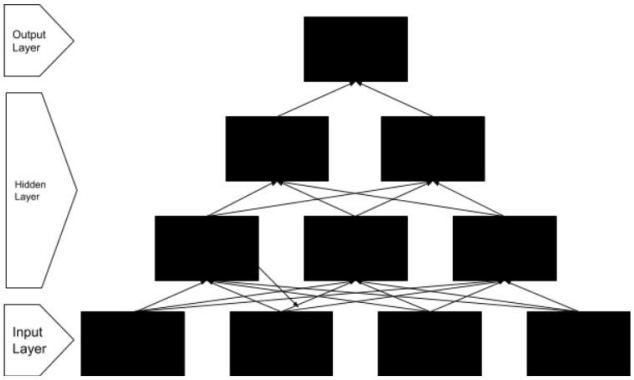
[2] 오영환 Yungwhan Oh (2019). 선형회귀분석 기법을 이용한 고교야구투수의 구속도 예측. High-School Baseball Pitcher's Pitching Speed Prediction Using Linear Regression Analysis Method. 한국지식정보기술학회 논문지, Korea Knowledge Information Technology Society,14(4), 381- 390.

[3] Gary McKenzie (2015). DATA ANALYTICS IN SPORTS: IMPROVING THE ACCURACY OF NFL DRAFT SELECTION USING SUPERVISED LEARNING. In Partial Fulfillment of the Requirements for the Degree Master of Science, University of Missouri-Columbia.

dataset records were included.

4. MLP Model

In this research, MultiLayer Perceptron(MLP) is used as a deep learning model for prediction. MLP is composed of several layers of an artificial neural network. MLP passes the input data to several hidden layers so that it can solve the complex problem. The limitations of single layer perceptron, which has various disadvantages in processing nonlinear data, is overcome by MLP.

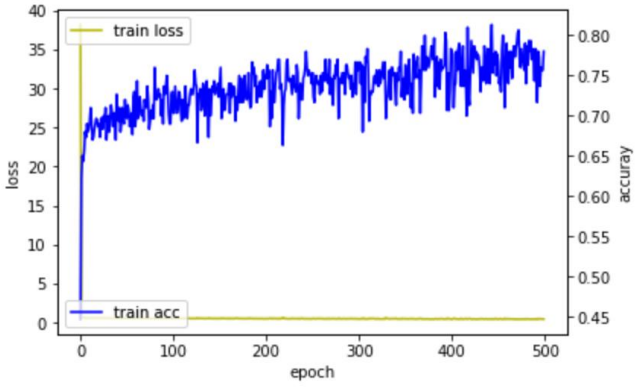


(Figure 1) MultiLayer Perceptron Visualization

5. Experiment and Evaluation

Datasets explained in chapter 3 is used to train the MLP model. 35 records and position is used as input data, and the rookie draft result is used as an answer data.

Python programming language is used as a learning environment, and Keras from Tensorflow version 2.2, a deep learning framework, is employed. The learning is repeated 500 times, and because the size of the data was relatively small, the test data set was not divided during the experiment. Following is the result of the experiment.



(Figure 2) Visual graph of the result of the learning

As the picture shows, the learning accuracy was about 75~80%. The experiment was conducted additionally with varying number of MLP Layer (3 to 5~6) and varying optimizer and activation functions like adam, sgd, relu, and sigmoid, but the learning result was similar.

6. Conclusion

In conclusion, it was possible to predict Korean Baseball Organization rookie draft result using deep learning based on high school baseball players' records and data. Prediction result accuracy was about 75%. The numerical value of the prediction result seems reasonably accurate. However, the number of nominated players is much smaller than the number of players that are not nominated. In other words, the number of data that have an answer of "1" is much smaller than the number of data that have an answer of 0. Considering that fact, the reliability of the prediction result is relatively low.

Only position and record data are used in this research. But later, other various data like physical information and school information can be collected and analyzed. By that process, the result would become much meaningful and insightful. Moreover, using different deep learning models other than MLP for learning can also improve the research result later.

References

[1] 황규인 Kyu-in Hwang (2018). "의사결정나무 분석 기법을 활용한 프로야구 외국인 투수 재계약 확률 예측. A study on KBO league foreign pitchers' re-sign possibilities using decision tree analysis. 국내석사학위논문, 고려사이버대학교 융합정보대학원. In partial Fulfillment of the Requirements for the Master Degree, Korea Cyber University Graduate School of Interdisciplinary Information Studies.

[2] 오영환 Yungwhan Oh (2019). 선형회귀분석 기법을 이용한 고교야구투수의 구속도 예측. High-School Baseball Pitcher's Pitching Speed Prediction Using Linear Regression Analysis Method.한국지식정보기술학회 논문지, Korea Knowledge Information Technology Society,14(4), 381- 390.

[3] Gary McKenzie (2015). DATA ANALYTICS IN SPORTS: IMPROVING THE ACCURACY OF NFL DRAFT SELECTION USING SUPERVISED LEARNING. In Partial Fulfillment of the Requirements for the Degree Master of Science, University of Missouri-Columbia.

[4] 홍석미, 정경숙, 정태충. Sukmi Hong, Kyungsuk Jeong, Taechung Jeong (2003). 혼합형 기계 학습 모델을 이용한 프로야구 승패 예측시스템. Win/Lose Prediction System : Predicting Baseball Game Results using a Hybrid Machine Learning Model. 정보과학회논문지:컴퓨팅의 실제 및 레터, Journal of KISS : Computing practices, 9(6), 693-698.

[5] 김송, 이승환, 홍성호, 김재현, 도진우, 윤승식, 강주영 Song Kim, Seunghwan Lee, Sungho Hong, Jaehyun Kim, Jinwoo Do, Seungsik Yun, Jooyung Kang (2018). 텐서플로 기반 야구 점수 예측 모델 설계. Design of Baseball Score Prediction Model using TensorFlow. 한국통신학회 학술대회논문집, The Journal of Korean Institute of Communications and Information Sciences,389-390.

[6] 박상현, 박진욱. Sanghyun Park, Jinwook Park (2017). 인경신경망을 이용한 한국프로야구 관중 수요 예측에 관한 연구. A Study on Prediction of Attendance in Korean Baseball League Using Artificial Neural Network 정보처리학회논문지. 소프트웨어 및 데이터 공학, KIPS transactions on software and data engineering, 6(12), 565-572.

[7] Woodham, Michael & Hawkins, Jason & Singh, Ankita & Chakraborty, Shayok. (2019). When to Pull Starting Pitchers in Major League Baseball? A Data Mining Approach. 426-431.

[8] Bierig, Brian & Hollenbeck, Jonathan & Stroud, Alexander. (2017). Understanding Career Progression in Baseball Through Machine Learning. Class Project for CS229 Fall 2017 at Stanford.

[9] 김혁. Hyuk Kim (2019). 기계학습 방법에 의한 프로스포츠에서의 관중 수 예측과 그 요인들 연구. Study on the Prediction of the Number of Spectators and It's Factors in Pro Sports by Machine Learning Method. Journal of The Korean Data Analysis Society, 21(4), 1867-1880.

[10] Koseler, K., & Stephan, M. (2017). Machine LearningApplications in Baseball: A Systematic Literature Review. *Applied Artificial Intelligence, 31*(9-10), 745-763.