ISSN: 2508-7894 © 2019 KAIA. http://www.kjai.or.kr
Doi: http://dx.doi.org/10.24225/kjai.2019.7.2.13

A Study on Prediction of Baseball Game Based on Linear Regression

[1]Kwang-Keun LEE, [2]Seung-Ho HWANG
[1, First Author] Professor, Department of Social Welfare, Kyungdong University, Korea
E-mail: kankun@k1.ac.kr
[2 Corresponding Author] Student, Department of Medical IT Marketing, Eulji University, Korea
E-mail: enenahwk100@naver.com

Received: June 24, 2019. Revised: September 03, 2019. Accepted: December 10, 2019.

Abstract

Currently, the sports market continues to grow every year, and among them, professional baseball's entry income is larger than the rest of the professional league. In sports, strategies are used differently in different situations, and the analysis is based on data to decide which direction to implement. There is a part that a person misses in an analysis, and there is a possibility of a false analysis by subjective judgment. So, if this data analysis is done through artificial intelligence, the objective analysis is possible, and the strategy can be more rationalized, which helps to win the game. The most popular baseball to be applied to artificial intelligence to analyze athletes' strengths and weaknesses and then efficiently establish strategies to ease the competition. The data applied to the experiment were provided on the KBO official website, and the algorithms for forecasting applied linear regression. The results showed that the accuracy was 87%, and the standard error was ±5. Although the results of the experiment were not enough data, it would be possible to effectively use baseball strategies and predict the results of the game if the amount of data and regular data can be applied in the future.

Keywords: Machine Learning, Linear Regression, Baseball Game

1. Introduction

In sports, strategies are often used differently in different situations, where the coach or coach will usually decide which direction to set up by analyzing the player based on data. There is a possibility that some people may miss out on their analysis and misinterpret with subjective judgment, so if such data analysis is done through artificial intelligence, it will enable objective analysis, which will help them win the game more rationally (New Kyu-sik, 2014). Among sports, baseball is the most inclined to rely on all-time, or historical data, and among the many types of play in baseball, it randomly selected and tested the type of left-handed pitcher. In the future, different types of athletes will be able to be analyzed in this way. Data has been passed on to players by situation on the KBO homepage, so it is possible to analyze the data accurately since the player you want to analyze is all year after year from his debut (KBO, 2019). The data mainly analyzed the pitcher's hit rate and used Microsoft's Azure Machine Learning Studio (Microsoft, 2019). The predictive model applied Azure Machine Learning Studio's linear regression algorithm (Kang Min-soo, 2018). The reason for this is that the results of the analysis were easy to understand through visualization functions such as statistics and graphs. The results of the analysis were estimated at 30% by 70% in order to avoid overconformity, a phenomenon where learning is so well done in the analysis process, but less accurate in the test data or actual application, the total result accuracy is 87%, and sampling error is ±5%. Regression analysis was performed to predict the outcome of these studies and apply them to tactics to help solve the game.

2. Data Set

2.1. Data Set Description

Sports clubs have a variety of categories to earn income, such as advertising, player trade and attendance fees, but according to a 2015 survey, the team's revenue is 73 billion won, followed by 36 billion won as broadcasting rights (Kim Won, 2017). Taking baseball as an example, the ratings are low because it usually broadcasts only on sports channel cables, but since important games such as the postseason and the Korean Series are broadcast on terrestrial channels, ratings go up overwhelmingly. Therefore, the higher the team's ranking, the higher the value of the team's image, broadcasting fees, and spectators' fees, the higher the income of the team. If the strategy is established through accurate analysis, the team can earn a lot of income by achieving a high winning rate and naturally increasing the team's ranking, which makes it even more important for the team to use rational strategy with solid analysis.

구분	H	2B	3B	HR	BB	HBP	SO	WP	BK	AVG
주자없음	79	12	1	10	20	2	65	0	0	0.253
1루	20	0	0	3	2	0	38	2	0	0.182
1,2루	5	0	0	1	3	0	3	1	0	0.200

Figure 1. Process of data collection

Figure 1 shows four players as Kim Kwang-hyun, Yang Hyun-jong, Jang Won-sam and Jang Won-jun as examples of the data collection process, and the premise is: 1. KBO league players 2. billion-a-year salary 3. starting pitchers 4. For more than 10 years, left-handed pitchers have been set at five conditions The reason why we set the conditions is that we set up veterans to exclude variables outside the game as much as possible. The reason is that rookie players are influenced by environmental or psychological variables, but because veteran players are more experienced and show consistent performance by adjusting the variables to the fullest extent possible. To maintain common ground in the data, we created conditions for starting pitchers and left-handed pitchers, and only high salaries for reliable data. By dataizing their entire lineup in various situations, including type of batter, innings and out count, they can be more useful in establishing strategies such as replacement timing or change of defensive position through analysis (he Jun-hoi, 1998).

2.2. Data Set Item

Table 1 describes the items used in the data.

Table 1. Explanation of attribute

H	2B	3B	HR	BB	HBR	SO	WP	AVG

Table 2 shows the maximum value, maximum value, median value, and average of all four of the statistics, including hits and home runs. The variables were set as innings, runners, free type, and out count. He set the hit rate as the data value because the statistics that can analyze the pitcher's characteristics such as strength and weakness are the percentage of hits, and since the hit rate is a situation where he loses the score, he focused on label.

Table 2. Data for the left-hander of the year 18

Separation	Minimum	Maximum	Mean	Median
H	5	307	87.65	76
2B	1	56	15.65	14
3B	0	2	0.6	0
HR	0	38	9.6	8
BB	1	79	22.7	17
HBP	0	11	2.9	2
SO	4	198	64.45	54
WP	0	12	2.6	1
AVG	0.178	0.548	0.2974	0.283

3. Experiments and Results

3.1 Preemptive Research and Environment

Using C4.5 Decision Tree Algorithm, a machine learning algorithm, and C4.5 Decision Tree Algorithm to predict the attack line-up, the pitcher's pitching prediction (Jin Seung-woo, 2014) and Deep Learning (Oh Yoon-hak, 2014) are analyzed by analyzing the big data of KBO baseball games to predict victory or defeat. So in this experiment, we used linear regression to learn 70% of each data and forecast 30%, but the amount of data was limited, so we hit the limit similar to the preceding study.

3.2. The result of an experiment

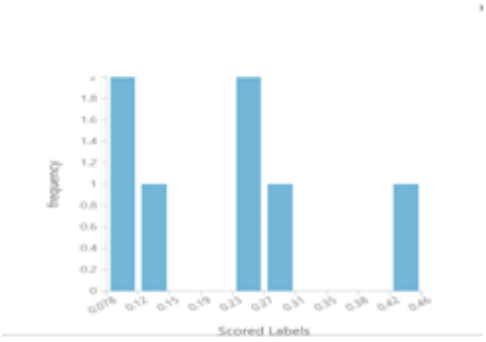

Figure 2. Result of machine learning

Figure 2 shows the seven least accurate result values, which are less accurate because data are created by the limitations of the experiment. The results will also be predictable in the competition situation as long as the accuracy is increased, since it can be compared with other items based on the score values predicted from the result values.

3.3. A review of the results of the experiment

Applying more data on a yearly basis is accurate and provides a full picture of the strengths and weaknesses of the players or types of players, helping to gain a clearer picture of the timing or direction of the strategy and increasing the winning rate by increasing the ability to cope with the crisis of the game. Although data on left-handed pitchers have only been analyzed at this time, this method of applying them to more types will help to use more complex tactics. For example, if the overall defensive position is moved back when there is a high chance of getting a double, it is possible to make it only out or a single, even if the batter is going to go for a double. Then, the long-distance situation will be reduced, and there will be less chance of losing.

4. Conclusion

The above linear regression study allows us to predict the loss of a game. In the analysis of data that we have been doing so far, we can analyze the data more objectively than humans, so we can clearly set the strategic direction to create a high winning rate, but with the most accurate data. As for the characteristics of sports, such as personal situations, internal factors, and environmental factors, it is difficult to analyze and exceptional circumstances can arise, so it is better to use this data analysis for reference, for example, pitchers have different relative histories of hitters, players' psychological states, and environmental factors can be cited as examples of weather or ground conditions. In addition, since the strength and weakness of each individual, such as the old type of left-handed pitcher, are different and have different characteristics, there are concerns that the analysis with the above data may result in premature generalization and decrease accuracy. In addition, the accuracy is low due to overconformity of the data forecast values. So, data variables and environmental variables of the same type can be found in advance research. In the KBO, if the above problems can be solved because there is a huge amount of data like MLB, not only can the problems be solved through additional research, but also the lineup and the winning and losing can be used in various games.

References

Microsoft (2019). Retrieved May 22, 2019, from https://studio.azureml.net

Jin, S.W., Kim, B.C., Um, I.K., & Kim, Y.I. (2014)."Development of Game Prototype for Prediction of Mobile Baseball by Using Data Mining Technique. *Journal of the Korean Association of Information Technology, 12*(2), 135-143.

Huh, J.H., & Jung T.C. (1998). Development of Pitcher Replacement Timing Prediction

Module using Reverse Wave Algorithm in a pro-baseball game predictor Journal of the Korean

Association of Information Science and Technology Announcement, 25(1B), 237-239.

Kang, M.S., Kang, H.J., Yoo, K.B., Ihm, C.H., & Choi., E. S. (2018). *Getting started with Machine Learning*

using Azure Machine Learning studio. seoul, Korea: Hanti media.

KBO (2019). Retrieved May 22, 2019, from https://www.koreabaseball.com/Default.aspx

Kim, J.H., Kim, K.T., & Han, J.K. (2015). Analysis of Big Data in baseball game using Deep Learning based machine learning algorithm. *Journal of the Korea Telecommunications Society Congress*, 22, 262-265

Kim, J.W., Park, S.H., Bang, S.W., Kim, J.K., & Lee, J.H. (2010). Predicting the lineup of baseball games using decision tree. *Journal of AI System Science, 20*(1), 93-95

Kim, W. (2017). Retrieved May 22, 2019, from https://news.joins.com/article/21205179 JoongAng Ilbo

Oh, Y.H., Kim, H.H., Yoon, J.S., & Lee, J.S. (2014). A Study on the Establishment of the Forecasting Model for Korean Professional Baseball by Using Data Manning. *Journal of Industrial Engineering, 40*(1), 8-17.

Park, D.S., & Kim, H.J. (2016). A Study on the Data Reduction Methods through the Analysis of Baseball Data. *Journal of the Korea Telecommunications Society Congress,12,* 244-245

Shin, G.S., & Lee H.C. (2014). The Winning Factors and Pattern Analysis of Korean Professional Baseball by Using the R Program, *Journal of the Korea Industrial Engineering Association Chungye Joint Academic Contest, 11,* 819-824