Advanced Science and Technology Letters
Vol.62 (Sensor 2014), pp.37-40
http://dx.doi.org/10.14257/astl.2014.62.10

Analyzing and Predicting Patterns in Baseball Data using Machine Learning Techniques

Wu-In Jang[1], Aziz Nasridinov[2], Young-Ho Park[1,*]

[1]Department of Multimedia Science, Sookmyung Women's University,
Cheongpa-ro 47 gil 100,Yongsan-gu, Seoul, 140-742, Korea
{leader0710, yhpark}@sm.ac.kr
[2]School of Computer Engineering, Dongguk University at Gyeongju
123 Dongdaero, Gyeongju, Gyeongbuk, 780-714, Korea
aziz@dongguk.ac.kr
*Corresponding Author

Abstract. The popularity of professional baseball has drastically increased mainly due to the success of South Korean national team in international arena. This success happened because of the correct analysis of baseball players when they are selected to the national team. In this paper, we propose to predict whether particular player can join to the national team by analyzing the past data of baseball players. Specifically, we take a baseball dataset of nine baseball players who made a final entry of South Korea national team, and predict if a new candidate can enter to the national team's final using k-nearest neighbor (KNN) algorithm. As far as we know, this is the first attempt to analyze the entry of a baseball player to the final list of national team.

Keywords: baseball data, machine learning, k-nearest neighbor algorithm.

1 Introduction

The professional baseball has significantly developed in South Korea since it was founded in 1982. This trend can be seen from the latest success of South Korean baseball in international level that includes becoming finalist of World Baseball Classic, WBC, in 2006, Olympic gold medal in Beijing in 2008, and Asian Games gold medal in Guangzhou in 2010. Baseball is a game that is based on statistical records. This success happened because of the correct analysis of baseball players when they are selected to the national team. In Major League Baseball, MBL, a method called *sabermetrics*, that is a method to analysis baseball data by using computing power, is widely used in order to perform various statistical and mathematical analyses on data of baseball players [1]. However, this method is only useful to decide the wages or grades of baseball players and predict the output of the baseball game.

There have been a number of researches that focused on the output prediction of a baseball game, lineup prediction of a baseball team, and various relationship analyses among baseball players' attributes and features. In the paper, we propose a method for analyzing and predicting patterns in baseball data using machine learning techniques.

ISSN: 2287-1233 ASTL

Specifically, we propose to predict whether particular player can join to the national team by analyzing the past data of baseball players. We take a baseball dataset of 9 baseball players who made a final of South Korea national team, and predict if a new candidate can enter to the national team's final entry using *k*-*N*earest *N*eighbor, *kNN* algorithm. As far as we know, this is the first attempt to analyze whether the baseball player selects or not to the final entry of national team by using kNN algorithm as one of machine learning algorithms.

The rest of the paper proceeds as follows. Section 2 discusses the related work. Section 3 describes the proposed method. Section 4 highlights conclusions.

2 Related Work

There have been a number of researches that focused on the output prediction of a baseball game, lineup prediction of a baseball team, and various relationship analyses among baseball players' attributes and features. We briefly review each of these approaches in this section.

In [2], the authors used a heuristic model in order to predict the output of baseball game. This method uses the combination of well-known decision tree classification method, ID3 algorithm, statistical analysis, and constructing a heuristic neural network. In the proposed method, the outcome of a baseball game, which is calculated using heuristic model, is the input to the neural network. Similarly, in [3], the authors proposed a hybrid machine learning algorithm in order to predict the output of a baseball game. The heuristic function is used in order to improve the accuracy of prediction, and hybrid model is used in order to reduce the dimension of input data for learning algorithm. The authors also reduced the complexity of a back-propagation algorithm that is used as a learning algorithm.

In [4], the authors proposed to use a decision tree algorithm in order to predict the lineup of a baseball team based on past lineups. The proposed method divides the lineup of a baseball team to four categories according to the roles of the players, and uses a decision tree algorithm to predict who nine starters of the game. In [1], the authors proposed to analyze the relationship between players experience and wages of South Korea Professional Baseball League batters by using the sabermetrics method. The proposed method first collects the batters game and wages information, and then constructs an index for the sabermetrics and calculates relationships by principal component analysis.

3 Proposed Method

In this section, we describe the proposed method. We first describe the baseball data that is used in this paper, and then describe how to apply *kNN* algorithm to this baseball dataset.

We collected a baseball dataset that is publicly available at Korea Baseball Organization, KBO, and Samsung Lions websites. We select 5 baseball players as representative few players with different positions to simply show our algorithm, who have the highest possibilities of entering to the national team, and collect their past data, including Batting average, Home runs, Run Batted In (RBI), Walks, Bases, Mistakes, Most hits, Scores and Bats, beginning from 2010 to present month of 2014. Similarly, we define 9 baseball players that made the finals of Asian Games in 2010, and collect their past data beginning from 2006 to 2010. The main purpose of this paper is predict whether current candidate 5 players can enter to the final of the national team compared to the 9 players that made national team finals of Asian Games in 2010.

In order to predict this phenomenon, we use *kNN* algorithm. *kNN* algorithm is a machine learning technique that classifies the data with precise accuracy. Using *kNN* algorithm we can predict whether current candidate 5 players can enter to the final of the national team compared to the 9 players that made national team finals. It has two parts. The first part defines the nearest neighbors and the second part calculates the distance between those neighbors with other points. Let us assume that we have a training dataset *D* made up x_i training samples. The examples are described by a set of features *F* and any numeric features have been normalized to the range [0, 1]. Each training example is labelled with a class label $y_j \in Y$. Our objective is to classify an unknown example *q*. For each $x_i \in D$ we can calculate the distance between *q* and x_i as follows [5]:

$$d(\mathrm{q}, x_i) = \sum_{f \in F} \omega_f \, \delta(\mathrm{q}_f, x_{i\ f}) \qquad (1)$$

The *kNN* is defined by distance formula calculated as Formula 1. In order to determine the class of *q*, there are several ways. The most straightforward approach is to assign the majority class among the nearest neighbors to the query [5].

First, we implement a multi-attribute *kNN* algorithm to handle various attributes related to baseball wages. Second, through the *kNN* algorithm, we can classify the groups of players with similar wages of players in Korea Baseball League, KBL, in this year. Finally, we determine the new candidate who can enter to the national team's final according to the *kNN* groups. In future work, we will describe the detail algorithm and experimental results by using wages of real players in KBL.

4 Conclusion

In the paper, we have proposed a method to predict whether particular player can join to the national team by analyzing the past data of baseball players. We believe that the proposed method has a positive impact to the development of a pro-baseball in South Korea as it improves the accuracy of selecting the right player to the Korea national team. In the future, we plan to perform a performance evaluation of the proposed method. Particularly, we plan to measure the accuracy and overall processing time of the proposed method compared to the well-known machine learning classification algorithms, including support vector machine (SVM), decision tree, and so on.

Acknowledgement. This work was supported by the IT R&D program of MKE/KEIT. [10041854, Development of a smart home service platform with real-time danger prediction and prevention for safety residential environments].

References

1. Seung, H., Kang, K,-H.: A study on relationship between the performance of professional baseball player and annual salary. J. The Korea Data Information Science Society, 285--298 (2012).
2. Kim, D,-S., Hong, S,-M., Jung, T,-C.: Prediction of win/lose of Professional baseball using Heuristic model. In: Korea Information Processing Society 2000, pp.325—328 (2000).
3. Hong, S., Jung, K., Chung, C.: Win/lose prediction system: Predicting baseball game result using heuristic model. J. KIISE: Computing Practices, 693--698 (2003).
4. Kim, J., Park, S,-H., Bang, S,-W., Kim, J., Lee, J,-H.: Predicting line-up of baseball game using decision tree. In: KIIS spring conference 2010, pp. 93--95 (2010).
5. Lee, J., Jung, M., Lee, S.: KNN/PFCM hybrid algorithm for indoor location determination in WLAN. J. The Institute of Electronics Engineers of Korea- Signal Processing, 146--153 (2010).