2022 IEEE International Conference on Multimedia and Expo Workshops (ICMEW) | 978-1-6654-7218-0/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICMEW56448.2022.9859411

DECIDE THE NEXT PITCH: A PITCH PREDICTION MODEL USING ATTENTION-BASED LSTM

Chih-Chang Yu[*], *Chih-Ching Chang*[†] *and Hsu-Yung Cheng*[‡]
[*]Chung Yuan Christian University and ccyu@cycu.edu.tw; [†]Chung Yuan Christian University and cycu10977011@cycu.org.tw; and [‡]National Central University and chengsy@ncu.edu.tw

ABSTRACT

Information collection and analysis have played a very important role in high-level baseball competitions. Knowing opponent's possible strategies or weakness can help own team plan adequate countermeasures. The purpose of this study is to explore how artificial intelligence technology can be applied to this domain. This study focuses on the pitching events in baseball. The goal is to predict the pitch types that a pitcher may throw in the next pitch according to the situation on the field. To achieve this, we mine discriminative features from baseball statistics and propose a stacked long-term and short-term memory model (LSTM) with attention mechanism. Experimental data come from the pitching data of 201 pitchers in Major League Baseball from 2016 to 2021. By collecting information of pitchers' pitching statistics and on-field situations, results show that the average accuracy rate reaches 76.7%, outperforming conventional machine learning prediction models.

Index Terms— Pitch prediction, LSTM, attention model, sport analysis

1. INTRODUCTION

With the increasing of ICT technology, we have many methods to obtain precise statistics of athletes during their competitions. These gathered data are quite huge so that it is hard to analyzed them manually. Therefore, many machine learning models were applied to analyze these data, wishing to provide useful information to human to make precise decisions. In terms of baseball, there are many ICT technologies involved which help athletes to improve their effectiveness in training. However, to our knowledge, there are few effective methods to provide accountable advises in pitchers pitching strategies. The pitching strategy is very complicated; a good pitcher usually needs to face a lot of batters many times to conclude adequate pitching strategy. However, this requirement is hard to meet for young pitchers. Unexperienced pitchers often rely on coaches' advices. However, a common phenomenon is that coaches' advices varies. Therefore, this study aims to discuss if we can provide trustable pitch prediction from big data with modern artificial intelligence technologies. With a reliable prediction system, unexperienced pitchers can have valuable information to improve their pitching skills.

The information collection and analysis of early baseball has always been a complex and time-consuming work. In tradition, players often try to extract some valuable information from the videos of official games in order to improve their skills. Many studies have provided some prediction models to predict players' traditional baseball statistics such as batting average or earned run average (ERA) [1][2]. In terms of prediction the pitches, Ganeshapillai & Guttag [3] proposed a method in 2012. In their study, they used binary classifiers to predict whether the pitch was a fastball or not, and the accuracy rate was around 70%. In 2015, Hoang [4] proposed a dynamic feature selection method that successfully increases the accuracy rate to 80%. However, in practical applications, it is not very useful to predict a pitch is fastball or not. More importantly, we expect a multi-classification model, that is, to be able to predict multiple types of pitches. Such model can provide more information than predicting a fastball only. In order to classify multiple pitch types, traditional machine learning models such as support vector machine (SVM) and decision tree (DT) are often used. Bock [5] used four binary classifiers to predict four pitch types, the final average accuracy can reach 74%. In the study of Sidle & Tran [6], the authors predict 7 kinds of pitch types by using 7 binary classifiers. However, the accuracy rate is below 60%.

From these studies we can find that the use of multiple binary classifiers will cause users to be easily confused when referring to the model results. If the prediction results of binary classifiers are inconsistent, the voting mechanism is usually used, but this mechanism cannot eliminate the event of a tie for voting. To tackle the above issues, this study proposes a single model solution, which is an attention-based long-short-term memory (LSTM) model to predict multiple pitch types.

The contribution of this paper includes: (i) unlike previous studies which only predicted if the next pitch will be a fastball, the proposed attention-based LSTM model can predict true pitch types in baseball. (ii) we discover valuable features including the pitching status and on-field situation to make promising predictions.

2. PROPOSED METHOD

2.1. Data preprocessing

This study analyzed a total of five years of major league season data from 2015-2018 and 2021. The data is obtained from the Statcast database. In addition, we only consider the data of the regular season, because there may or strategic adjustments in spring training and playoffs, so that the statistics were different from the data of the regular season. As a result, the trained model may be affected.

It usually requires sufficient amount of data to train a robust deep neural network. However, a regular starting pitcher generally has only 2000 pitches per season. Because pitching is a sequential process, this study slices the pitching sequence of each plate appearance into subsequences to increase the amount of training samples. More specifically, a pitch sequence with length 3 will be sliced into subsequences {1}, {2}, {3}, {1, 2}, {2, 3} and {1, 2, 3}. By doing so, a pitch sequence which has n pitches will have $n(n+1)/2$ training samples. The maximum length of each training sample is set to 6. We also pad the sequence with 0s for those training samples with length less than 6. There are 13 pitch types used in this study, which are: 4-Seam, 2-seam, Changeup, Curveball, Cutter, Eephus, Fork, Knucklecurve, Knuckle ball, Screwball, Sinker, Slider and Split-Finger. We do not consider the pitchout and intentional walk because these pitch types are few and they usually appeared due to coaching strategies.

2.2. Feature selection

To train the LSTM model, this study adopted 8 features which can be categorized into pitcher-related features and on-court characteristics. Pitcher-related features are the pitch type, pitch location and total pitches. On-court characteristics are the count, number of runners on the base, the score difference, number of innings and current number of outs. Table 1 describes the definitions of features used in this study. Note that we categorized the pitch location into 13 zones, where 9 of them are inside the strike zone.

2.3. LSTM model

The RNN model was proposed to solve the problem of sequential data. However, it was found that the RNN model suffers the vanishing gradient severely. In order to solve this problem, LSTM was proposed to solve such problem [7]. Four new units were designed in the LSTM unit, namely the input gate, forget gate, memory cell, and output gate. With the design of this core mechanism, LSTM can retain the information from the previous sequence. The equations of LSTM are listed as follows:

Table 1. Description of features used in this work.

Feature	explanation	Value type
Previous pitch	Type of the previous pitch. Set to unknown if the predicted pitch is the first pitch	Categorical
Previous location	The location of the previous pitch	Categorical
Pitch count	Accumulated pitches in this plate appearance	Continuous
Inning	current inning	Categorical
Runner on base	Cases of runners on base	Categorical
Score difference	Positive if pitcher's team is leading	Continuous
Ball count	Number of balls and strikes in current plate appearance	Categorical
Outs	Number of outs	Continuous

$$f_t = \sigma_g(W_f x_t + U_f h_{t-1} + b_f) \quad (1)$$

$$i_t = \sigma_g(W_i x_t + U_i h_{t-1} + b_i) \quad (2)$$

$$o_t = \sigma_g(W_o x_t + U_o h_{t-1} + b_o) \quad (3)$$

$$c_t = f_t \circ c_{t-1} + i_t \tanh \sigma_c(W_c x_t + U_c h_{t-1} + b_c) \quad (4)$$

$$h_t = o_t \tanh \sigma_h(c_t) \quad (5)$$

where f, i, o, and c represent forget gate, input gate, output gate, and memory cell, respectively. W, U, and b represent the weight coefficients, and σ is the activation function.

2.3. Attention-based LSTM

Although LSTM significantly improves the performance of RNN-based model, it still has some improvements can be made. For example, involving the attention mechanism. The attention mechanism is often used to solve natural language processing problems. The core idea is that when the human brain observes things, it generally focuses on some important part of them. The more important things, the more attention it should be. The attention mechanism wants the model to simulate human's attention, when the model was learning, it could focus on learning important parts, thereby improving the efficiency of the model. The concept of attention mechanism has been proposed for many years, and there are many kinds of attention mechanisms that can be used so far. This work uses the soft attention mechanism [8]. Fig. 1 illustrates the architecture of the attention-based LSTM proposed in this study. The proposed model has six layers, which are two LSTM layers, one Attention layer, and three fully connected layers.

The pre-processed data was used as feature vectors and fed into the LSTM layer. Then the output processed by the LSTM layer will be combined by the Attention layer to obtain a context vector *C*, followed by two fully connected layers to create the final output. The equations are as follows:

$$\propto_t = softmax(S_i) \quad (6)$$

$$C = \sum_{t=1}^{k} \propto_t o_t \quad (7)$$

where α_t is the attention weights that can be trained from the network, representing the importance of the LSTM output at timestep *t*. *C* represents the output of the Attention layer, which is the weight sum of the LSTM output at timestep *t*.

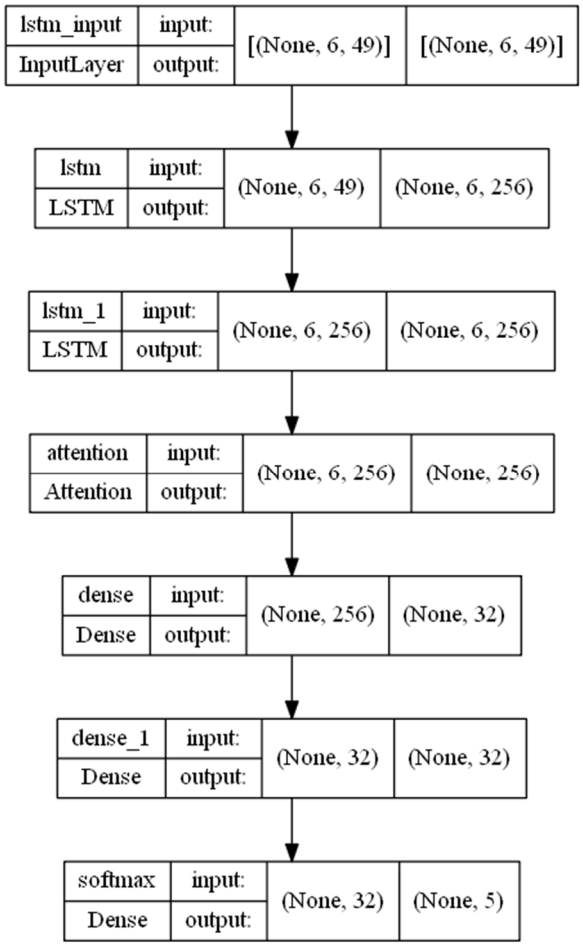

Fig. 1. Proposed Attention-based LSTM structure.

3. EXPERIMENTAL RESULTS

3.1. Data acquisition

This study uses five years of MLB data from 2015-2018 and 2021 for experiments. We exclude pitchers who pitched less than 5,000 in these years. As a result, there are a total of 201 pitchers' data used for evaluating the model. The training set has 3,959,654 samples and size of the test set is 20% of the whole dataset, which is 990,004. As mentioned in section 2.1, there are 13 different pitch types used in this study. Because the available pitch types differ from pitchers, this work trained individual models for each pitcher rather than having a unified model. We believe that this approach is much close to the practical situation. The output categories (i.e. the pitch type) of each model vary from 3 to 9. The average number of pitching types per pitcher is 5.547. All models are implemented and trained using a NVIDIA 2080 Super GPU with Tensorflow 2.8 backbone.

3.2. Parameter optimization

To evaluate the performance of the proposed model, this study designs five models for comparisons. Four of them are LSTM-based models, which are: single-layer LSTM, stacked LSTM, Attention-based LSTM with 1 fully connected (FC) layer and Attention-based LSTM with 3 FC layers. The other one is the popular machine learning model: the XGBoost model [10]. The reason we choose XGBoost in this study is because this model has been proven to have best classification results in many applications.

This study made two experiments. In the first experiment we tuned the LSTM-based model with different number of LSTM units and layers in order to seek for the best accuracy. We selected the pitcher Wade Miley's pitching data for this experiment because his pitching data has the worst prediction accuracy. The number of LSTM units tested for experiments ranges from 128 to 2048. The results are shown in Table 2. It can be seen that the accuracy gradually increases with the increasing of LSTM units. However, the training time of the 2048-unit model is 3.8 times that of the 1024-unit model with only 0.5% improvements. Hence, we think 1024-unit model is a better choice.

Afterwards, we tried different combinations of LSTM units and test the performance of a 2-layed stacked LSTM model. The results are shown in Table 3. From Table 3, it can be seen that the performance of using a combination of (256, 256) is very close to other combinations. In consideration of model complexity, we choose a 2-layer stacked LSTM and set the number of LSTM units to 256 for both layers for the next experiment.

In the second experiment, we use the same parameter configuration obtained from the first experiment, which is 256 LSTM units for both layers, and compare the performances of all models on all pitchers. The results are shown in Table 3. From Table 3 we can see that the LSTM-based model outperforms the XGBoost model. This is not surprising because the XGBoost model only considers features in the previous pitch.

Table 2. Model accuracy of single layer LSTM with different units

Units	Accuracy (%)	Time (sec.)
128	65.1	74.82
256	68.2	120.47
512	68.8	312.67
1024	69.9	1091.5
2048	71.4	4160.0

Table 3. Model accuracy of stacked LSTM model with different combination of LSTM units.

(units in layer1, units in layer2)	(512, 256)	(256, 256)	(256, 512)	(512, 512)
Accuracy (%)	75.9±3.4	76±3.5	76.6±3.3	76.3±3.6

The Attention-based LSTM with 3 FC layers has the best average accuracy of 76.7% among all models. Max accuracy is achieved on pitcher R.A. Dickey, where 84.1% of his pitches is Knuckle ball. The prediction result of XGBoost classifier is very close to this percentage. On the contrary, although the lowest accuracy is achieved on pitcher Wade Miley, who pitched 6 different types in his games. Among all of his pitches, 27.2% are 4-seam fastball, 18.3% are changeup, 9.6% are curveball, 16.9% are cutter, 15.8% are sinker and 12% are slider. Despite the wide divergence in his pitching tendency, the accuracy rate still reaches 69.4%, which outperforms the XGBoost classifier. These results show that Attention-based LSTM not only has higher accuracy than other models, but also has a smaller standard deviation between different pitchers so that we believe that this model can make accurate and stable predictions.

Table 3. Prediction accuracy among different models

Model	Max accuracy (%)	min accuracy (%)	Average accuracy (%)	# of params
XGBoost	85.2	28.4	45.7±9.8	-
LSTM(1024)	92.1	67.0	73.8±4.1	4.4M
Stacked LSTM	91.9	69.1	75.7±3.4	840k
Attention-LSTM	92.3	69.4	76.0±3.5	840k
Attention-LSTM-3FC	92.4	69.4	**76.7±3.4**	850k

4. CONCLUSIONS

Benefiting from the Statcast application which collects various statistics in Major League Baseball games since 2015, we can easily get a lot of pitching data for analysis. In the past, it required experienced coach or scouts to discover the pitching strategies of various pitchers. By leveraging deep learning models, the accuracy of predicting pitch types of pitchers in this study is greatly improved over conventional machine learning models. With the proposed approach, the pitching strategies can be quickly analyzed through the model and the results can be trusted. Coaches and players can refer to the prediction results to improve their pitching strategies during training.

There are still some improvements that can be made. In our opinion, batters' status is also a factor that affects the pitching strategy. In the future, we would like to add more features which reflect the batters' status and test if the model can be further improved. In addition, the popular transformer model which is expected to improve the accuracy can also be considered.

5. ACKNOWLEDGMENT

This work is funded by the Ministry of Science and Technology, Taiwan (R.O.C.) with project no. MOST 110-2627-H-155 -001-

6. REFERENCES

[1] S. R. Bailey, J. Loeppky, and T. B. Swartz, "The prediction of batting averages in major league baseball," Stats, vol. 3, no .2, pp. 84-93, MAR, 2020.

[2] D. Jordan, "Measuring Baseball Defensive Value using Statcast Data," M.S. thesis, Dept. Statistical Science, DU., North Carolina, United States of America, 2017.

[3] G. Ganeshapillai, and J. Guttag, "Predicting the next pitch," presented at SSAC, Boston, MA, 2012.

[4] P. Hoang, "Supervised learning in baseball pitch prediction and Hepatitis C Diagnosis," M.S. thesis, Dept. Applied Mathematics, NC State Univ., North Carolina, United States of America, 2015.

[5] J. R. Bock, "Pitch sequence complexity and long-term pitcher performance," Sports, vol. 3, no. 1, pp. 40-55, MAR, 2015.

[6] G. Sidle, H. T. Tran, "Using multi-class classification methods to predict baseball pitch types," M.S. thesis, Dept. Applied Mathematics, NC State Univ., North Carolina, United States of America, February, 2018.

[7] S. Hochreiter, and J. Schmidhuber, "Long short-term memory," Neural computation, vol. 9, no. 8, pp. 1735-1780, November, 1997.

[8] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," In ICML, 2015, pp. 2048-2057.

[9] T. Chen, and C. Guestrin, "Xgboost: A scalable tree boosting system," In Proceedings of the 22nd ACM SIGKDD, 2016, pp. 785-794