Winter 12-5-2020

A Two-Stage Real-time Prediction Method for Multiplayer Shooting E-Sports

Jiaxin Liu
Sichuan University, China, 1007297545@qq.com

Jiaxin Huang
Sichuan University, China, 1435958090@qq.com

Ruyun Chen
Sichuan University, China, 562789829@qq.com

Tianchang Liu
Sichuan University, China, 190758398@qq.com

Liang Zhou
Sichuan University, China, zhouliang_bnu@163.com

Liu, J.X., Huang, J.X., Chen, R.Y., Liu, T.C. & Zhou, L. (2020). A two-stage real-time prediction method for multiplayer shooting e-sports. In *Proceedings of The 20th International Conference on Electronic Business* (pp. 9-18). ICEB'20, Hong Kong SAR, China, December 5-8.

A Two-Stage Real-time Prediction Method for Multiplayer Shooting E-Sports

(Full Paper)

Jiaxin Liu, Sichuan University, China, 1007297545@qq.com
Jiaxin Huang, Sichuan University, China, 1435958090@qq.com
Ruyun Chen, Sichuan University, China, 562789829@qq.com
Tianchang Liu, Sichuan University, China, 190758398@qq.com
Liang Zhou*, Sichuan University, China, zhouliang_bnu@163.com

ABSTRACT

E-sports is an industry with a huge base and the number of people who pay attention to it continues to rise. The research results of E-sports prediction play an important role in many aspects. In the past game prediction algorithms, there are mainly three kinds: neural network algorithm, AdaBoost algorithm based on Naïve Bayesian (NB) classifier and decision tree algorithm. These three algorithms have their own advantages and disadvantages, but they cannot predict the match ranking in real time. Therefore, we propose a real-time prediction algorithm based on random forest model. This method is divided into two stages. In the first stage, the weights are trained to obtain the optimal model for the second stage. In the second stage, each influencing factor in the data set is corresponded to and transformed with the data items in the competition log. The accuracy of the prediction results and its change trend with time are observed. Finally, the model is evaluated. The results show that the accuracy of real-time prediction reaches 92.29%, which makes up for the shortage of real-time in traditional prediction algorithm.

Keywords: real-time prediction; machine learning; two-stage algorithm; E-sports.

*Corresponding author

INTRODUCTION

The prediction of E-sports competition has an important impact on the development of e-sports industry (Keiper *et al.*, 2017) and the optimization of game content (Breiman, 2001). There are some related prediction algorithms about the existing popular games and popular E-sports projects (Bosc *et al.*, 2014). At present, there are three kinds of game prediction algorithms: neural network algorithm, AdaBoost algorithm based on Naive Bayesian (NB) classifier and decision tree algorithm. These three algorithms have their own advantages and disadvantages. The neural network model is suitable for the scene with large amount of data and the intrinsic relationship between parameters. Its advantage is that the parallel distributed processing ability is strong, it can extract data features and approximate complex nonlinear relations. The disadvantage is that the learning time is too long when a large number of parameters are needed. The learning process cannot be observed and the output results are difficult to explain. AdaBoost algorithm can take different classification algorithms as weak classifiers, which makes good use of weak classifiers to cascade. AdaBoost has high accuracy. Compared with bagging algorithm and random forest algorithm, AdaBoost fully considers the weight of each classifier. However, the problem is that the number of weak classifiers is not easy to set, so cross validation is needed to determine. And the data imbalance will lead to the decline of classification accuracy. The advantage of the decision tree algorithm is that it is convenient for us to intuitionistic decision rules, at the same time, it can deal with the nonlinear characteristics and consider the interaction between variables. The disadvantage is that it is easy to over fit and it is very difficult to deal with missing data. However, the rankings predicted by these algorithms are the final rankings at the end of the game, and they do not predict the behavior of the game in real time. The frequency of data monitoring required by real-time prediction is more difficult. Although the prediction of the results is more accurate, the requirements of the algorithm are also higher. The prediction needs to obtain the player's status at each time, including the equipment the player obtains, the player's behavior characteristics and so on. At the same time, there are some unpredictable and nonstandard factors for the results, such as the influence of random refresh mechanism on players' psychology, the influence of players' personal level, etc., which will cause deviation to the results. But the real-time prediction can dynamically display the competition situation. The real-time prediction takes into account the changes of the importance of each index in different stages of the competition with the development of the competition process (Ziegler & König, 2014). However, the current E-sports ranking prediction algorithm has not considered the real-time prediction problem, so this paper focuses on the real-time prediction of E-sports competition.

Before the real-time prediction, the real-time prediction algorithm must first get a weight model of the factors that influence the game results, and then put the data into the prediction according to the new method on the basis of this model. Therefore, the research stage is divided into two parts. In the first stage, the problem to be studied is to predict the players' final ranking according to the overall performance of the players in the game, and calculate the influence weight of the main factors in the game for the players to win. The second stage is to predict the players' game rankings in real time, that is, according to the players' performance in a short period of time, track the game rankings in real time and predict their final rankings. In real-time prediction, data acquisition and cleaning are also particularly important. The experimental data should ensure perfect

characteristics, uniform format, and enough data to support the accuracy of the results, excluding the influence of accidental events. Therefore, the difficulty of this study is dealing with the large amount of data in the original data set, and the calculation takes a long time, which has certain requirements for the performance of the algorithm. For the case of large number of features, it is necessary to weight different features according to their importance, or generate additional features through analysis. In the problem of model selection, ensemble learning is needed to improve the accuracy of the algorithm. The research on real-time prediction not only improves the theory of prediction algorithm in machine learning, but also promotes the development of e-sports industry. In recent years, with the rapid development of the electronic game industry (Rahman *et al.*, 2020), the e-sports industry is also developing in the direction of professional sports (Kinkade *et al.*, 2015). Many games have been held online and offline all over the world, and many games have been listed as official events in the Asian Games. Data analysis plays an extremely important role. The official team and professional team of each game need professional personnel to collect, process and analyze the game data, so as to adjust and optimize the game content or help the team win the game (Pluss *et al.*, 2019). Therefore, this paper aims to predict the player ranking in the game from a new perspective, so as to help the relevant personnel in the e-sports industry.

Firstly, this paper summarizes the main methods and existing problems of game prediction algorithm. Then, the design of our research (real-time prediction algorithm based on random forest) is described from four aspects: research ideas, experimental preparation, experimental process and experimental results. Finally, we evaluate our model algorithm from the aspects of accuracy, implementation ranking and prediction fluctuation, discuss the innovation points and shortcomings of this experiment, and put forward the areas that can be improved, and look forward to the future research.

LITERATURE REVIEW

At present, there are three kinds of game prediction algorithms: neural network algorithm, AdaBoost algorithm based on NB classifier and decision tree algorithm.

Neural Network Algorithm

At present, there are many directions of improvement about neural network prediction algorithm. Because sports competition and E-sports have a lot of similarities in competitive rules, this paper also refers to the prediction algorithm of team events in sports competition. Using the deep neural networks (DNNS) and artificial neural networks (ANN), MD. Ashiqur Rahman (Xiong *et al.*, 2014) established an effective framework to predict the results of football matches. The data set is used for ranking, team performance, results of all previous international football matches, etc. ANN and DNN are used to mine and process the motion data, and the predicted values are generated. The data set is divided into training, verification and testing. Using the proposed DNN structure, the corresponding model performs well in predicting the 2018 World Cup. The prediction accuracy of the model is 63.3%. From the perspective of prediction accuracy, it has reached the pass line, but due to the problems of data set selection and team information acquisition channel, the prediction accuracy is limited to a certain extent. This accuracy can be improved with appropriate data sets and more accurate team information.

AdaBoost Based on NB Classifier

Pulung Nurtantio Andono, Nanang Budi Kurniawan, Catur Supriyanto (Pereira *et al.*, 2016) used naive Bayes (NB) as classifier to predict the winning team in data mining algorithm. However, NB has some disadvantages such as data imbalance, so this study proposes some methods to implement AdaBoost based on NB, which are discretization and Gaussian distribution kernel function. Both are used to handle numeric properties. The experimental results show that the prediction accuracy of win with AdaBoost plus Gaussian distribution kernel can reach 80%. However, the training of AdaBoost algorithm is time-consuming, and the best segmentation point of current classifier needs to be re selected each time. It is time-consuming and laborious to train the model for many times to improve the prediction accuracy.

Decision Tree Algorithm

Firstly, the decision tree algorithm processes the data, generates readable rules and decision tree by inductive algorithm, and then analyzes the new data by using decision. As for the PUBG game, Yonghan Qiu uses classical machine learning algorithms such as linear regression and decision tree to analyze how different factors affect the winning rate of the game, and establishes a decision tree prediction model to predict the level of players. First of all, the author carries on the correlation analysis, obtains the data attribute which has the high correlation with the player rank. Then, due to the different types of players' data have some differences, these data reflect the level of players' game. According to the ranking score of each player, players are divided into four categories. In this paper, the decision tree is constructed based on the information gain value. By calculating the information entropy difference of different data sets to select the node to be split, the decision tree prediction model is successfully constructed and predicted. This is consistent with our research topic, but the author predicted the final ranking of players in the game, and did not track and predict players' behavior in real time.

RESEARCH METHOD

The research problem of this paper is the prediction of the electronic sports game - PUBG. The reason why we take PUBG as an example is that as a tactical competitive shooting sandbox game, PUBG has become popular all over the world. There are millions of active players every month, so we have a huge data base to support our experiments. At the same time, the research results are of great significance to many players and E-sports related personnel. The specific prediction direction is to predict the players' ranking in a game under the conditions of given equipment and landing point. We study the problem of regression,

y value is the prediction of the ranking. The first place is 1, the last one is 0, ranking as percentile. Because the actual competition contains too many factors, so our research is divided into two stages. In the first stage, the problem to be studied is to predict the player's final chicken eating ranking according to the player's overall performance in the game. At the same time, the weight of the main factors included in the game, such as the number of kills, the number of treatments, and the moving distance, is calculated. The second stage is the real-time prediction of players' chicken eating ranking. According to the players' performance in a short period of time, we can track the match rankings in real time and predict their final chicken eating ranking.

Data Sources

The data of stage 1 is from the real competition data provided by the open data set of Kaggle. The data set contains more than 65000 anonymous player data of games. The format of the training set data is that each row contains a player's post game statistical data, which is listed as the characteristic value of the data, with 29 data items. The data set includes all types of games: single row, double row, four row, but each game does not necessarily have 100 people, each team has up to 4 members. In the experiment, we will randomly scramble the data set and divide it into training set and test set according to the proportion of 70% and 30%. The method is to use the train in the sklearn Library of Python_ test_ Split() function. The data of stage 2 is from the JSON data of PUBG official API through crawler. First, determine the source platform and date of the game to obtain the match ID within 24 hours, then use the ID to obtain the match data of each game. Finally, the log JSON is obtained from the telemetry URL contained in the match data, which is the data set needed in stage 2. This data set has comprehensive features. There are 44 types of logs in a single game, the highest JSON file is 32MB, and the total number of 561 games is more than 8GB.

Assumptions

The main background and rules of PUBG are: in each competition, 100 players fly and are sent to different places on an island. Players are constantly shrinking and moving inside the "poison circle" while searching for materials, weapons and large escape with other players until the last person or team survives. In this game, PUBG has four maps: green land, desert, rain forest and snow. With different maps, the proficiency level of each team is different, and the ranking may be different. At the same time, there is an airdrop mechanism in the PUBG, and the team that grabs the airdrop has an absolute advantage in terms of materials due to its AWM, Groza and other air drop guns and three-level equipment. Another factor that needs to be considered is that the security zone will be refreshed randomly in the PUBG. The teams in the security zone have a natural advantage and can attack the teams entering the circle by blocking bridges or guarding the circle. Therefore, when forecasting, we should also consider whether we need to consider the factor of security zone refresh. Therefore, according to the above factors, the ranking prediction algorithm used in this paper is based on the following four prerequisites:

1) There are many data features generated in the game, but not all of them are related to the player's ranking. It is necessary to screen out some features that are not related to the prediction results or related but have repetition. After calculating the correlation, those less than 0.09 will be deleted and the features with greater relevance will be retained.
2) Since the data set of Kaggle does not include the data of weather, map type, player coordinates, poison ring coordinates, bombing area coordinates and weapon equipment types, the influence of these factors is not considered temporarily.
3) Without considering the influence of unpredictable factors such as player's personal factors, the result of game prediction is only related to the characteristics we set.
4) Players do uniform movement throughout the course. We don’t consider the starting and ending time of walking, swimming and driving. The distance increases at a constant speed.

Research Design

In the first stage, we set up the random forest model (Carrillo Vera & Aguado Terrón, 2019), Light GBM model and GBR gradient enhanced regression tree model (Hamari & Sjöblom, 2017) to train the weight of various features in the PUBG game. We evaluate the advantages and disadvantages of different models through four evaluation indexes: accuracy, average absolute error, mean square error and goodness of fit, so as to select the most suitable model for our second stage and make the accuracy of the second stage real-time prediction accurate. Finally, we choose the random forest model from the comprehensive perspective of accuracy, mean absolute error, mean square error and goodness of fit.

In the second stage, taking all the members of "chicken eating" team as an example, the ranking prediction model trained by random forest algorithm is used. At the same time, the game log data is obtained from the official PUBG API, and each influencing factor in the data set is mapped and converted with the data items of the game log. Then extract the final player ranking data. Get each player's "winplace" data from match data. Determine the time interval, that is to define the "real-time prediction", put the data into the prediction, and get the ranking of each player. Compare the predicted results of each segment with the real ranking, evaluate and continue to train the model. The accuracy of the prediction results and its change trend with time were observed. Finally, the model was evaluated.

RESEARCH MODEL

Random forest (RF) is an ensemble classifier based on Bagging, which is composed of several fully grown decision trees (Chen *et al*., 2019). In classification prediction, the class label of the sample is determined by the mode of the output class label of these decision trees (Jonasson & Thiborg, 2010). The training set of each decision tree in RF is generated by self-help resampling, that is, n samples are randomly selected from the original training set with N number. Some samples may be

extracted many times under self-service resampling, while others may not be extracted. According to statistics, the training set of each decision tree contains 2/3 samples of the original training set, while the remaining 1/3 samples which are not extracted form out of bag data (OOB data) to calculate the importance of features. When building a decision tree, random forest extracts $\sqrt{d}$ attribute from d attributes randomly. Then, according to Gini gain maximization principle, it selects the attribute with the best classification ability as the splitting attribute, and divides the data of nodes into new sub nodes. Gini value is often used to measure the purity of data D, and its calculation formula is as follows:

$$Gini(D) = 1 - \sum_{k=1}^{|y|} p_k^2 \quad (1)$$

p_k represents the proportion of the kth class label in the data. $|y|$ is the number of category label values. $\mathrm{Gini}(D)$ reflects the probability that two samples are extracted from data set D with different categories. Therefore, the smaller $\mathrm{Gini}(D)$, the higher the purity of dataset D. Gini gain obtained by splitting data set D according to attribute a can be calculated as follows:

$$\Delta I_G(a) = \mathrm{Gini}(D) - \sum_{v=1}^{V} \frac{|D^v|}{|D|} \mathrm{Gini}(D^v) \quad (2)$$

V is the number of types of value a, and $|D^v|$ is the number of samples corresponding to the value of v. The Gini gain maximization principle is to calculate the Gini gain of all attributes of a node, and select the attribute with the largest Gini gain as the splitting attribute. According to this principle, the split attribute can make the sub node dataset with the highest purity, which shows that the classification performance of this attribute is the best. The better the classification performance, the more important the attribute is in the feature set, so the importance of the feature can be reflected according to the node division of the decision tree. However, due to the double random mechanism of random forest, it is not advisable to use the frequency of attributes in the random forest decision tree to reflect the importance of features. Therefore, in order to reflect the importance of features more accurately, this paper chooses the method based on the classification accuracy of verification set to measure the importance of features. Assuming that there are k decision trees in RF, the importance of feature a can be obtained by the following steps:

1) When $k = 1$, self-help resampling is used to generate training set and out of bag data set, and decision tree T_k is constructed on the training set;

2) Based on T_k, the OOB data are predicted and classified, and the number of samples with correct classification is counted as R_k;

3) A new OOB sample set is obtained by disturbing the value of characteristic a in OOB. Then, TK is used to classify and predict the new OOB sample set, and the number of correctly classified samples is counted and recorded as R'_k;

4) Let k=2, 3, ⋯ , K, Repeat steps 1 to 3;

5) The importance of characteristic a can be calculated as follows:

$$IMP(a) = \frac{1}{K} \sum_{k=1}^{K} (R_k - R'_k) \quad (3)$$

If the value of feature a is disturbed, the classification accuracy does not change much before and after the disturbance, which indicates that feature a plays a small role in classification and the classification performance is low. At this time, the value of $R_k - R'_k$ will be very small, so the higher the imp (a), the better the classification performance of feature a.

DATA ANALYSIS

Data Processing

After obtaining huge data, it is necessary to clean the data, otherwise it will affect the efficiency and results of the experiment. Due to the large amount of data collected, we import the data sets into the database on the cloud server, sort them according to the competition, and export the first 100000 pieces of data as experimental data sets. At the same time, we judge whether there is missing value through is_null() function. If there are less missing values in the experimental data set, the missing data will be deleted and the data will be re-selected if there are more. In the selected experimental data set, the result of missing value check is that there is no missing value. In order to ensure the uniqueness of the collected data, it is necessary to check the duplicate of the data. Because a player can only participate in one team in a game, a new data item can be formed by combining the player ID, team ID and game ID, and the duplicate can be checked through the data item. At the same time, not every game has 100 players. The difference in the number of participants will lead to different meanings of other features. Obviously, it is more difficult to knock down 5 players in a 90 player game than in a 100 player game, which means that the player's strength is stronger. Therefore, to deal with the distribution of the data items affected by the number of participants, the method of reassignment is to subtract the weight of the actual number of participants from 100 people.

The generation of outliers may be due to errors in the collection of data sets, or the use of plug-ins by players during the game. In this study, we do not study plug-ins, so we remove outliers directly from the data set. Because the data value of the data item

is 0, it is not suitable to use the box diagram method to detect the abnormal value. System data items and player's inherent data items should not have abnormal values, and the main object of processing is the data generated by players in the game. In the visualization process, there are obvious outliers in the data: kills is greater than 30, which is the level that ordinary players can't reach; the number of participants is less than 85, such a game has no reference value, so the data of these games are deleted. In addition to the outliers of individual data items, there are also data exceptions when discussing multiple data items. The solutions are as follows: 1) there is no movement in the whole game, but there is kill data. The created data item distance is the sum of ridedistance, swimmdistance and walkdistance, which represents the global moving distance. Find out the data whose kills is greater than 0 and distance is 0 and delete it. 2) The rate of head burst is high and the number of people killed is more. Create data item snapshot_ Rate, which is the quotient of the number of hits and the number of kills. Find out that the kills are greater than 5_ Data with a rate equal to 1 and deleted.

In the experimental data set, there are many feature items, and some of them have little correlation with the research problem. We reduce irrelevant features and redundant features by feature selection. The method used is filter. A filtering method of "correlation statistics" is designed to calculate the features, set a threshold to select, and finally to train the learner. Because our problem is a regression problem, we choose to use the correlation coefficient method for feature selection. The correlation coefficient between each feature and winplaceperc (percentage ranking) was calculated. The threshold value was set as absolute value, and the correlation coefficient between 0-0.09 was no correlation. The final calculation results show that the characteristic items whose absolute value of correlation coefficient is less than 0.09 are killpoints, matchduration, maxplace, numgroups, rank points, road kills, team kills and vehicle destroys Therefore, these characteristics were excluded in the experiment.

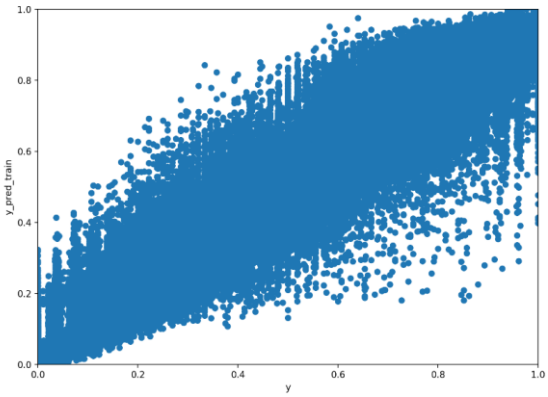
Table 1: Characteristic correlation coefficient table

Characteristic item	correlation coefficient
assists	0.307821213
boosts	0.63847773
damageDealt	0.450121504
DBNOs	0.286132604
headshotKills	0.280731682
heals	0.431013528
killPlace	-0.726740931
killPoints	0.011517781
kills	0.430925331
killStreaks	0.374426971
longestKill	0.410889786
matchDuration	-0.002643288
maxPlace	0.036413025
numGroups	0.037417883
rankPoints	0.015676896
revives	0.243392452
rideDistance	0.354353049
roadKills	0.03527861
swimDistance	0.145687882
teamKills	0.015653821
vehicleDestroys	0.072103196
walkDistance	0.817316682
weaponsAcquired	0.619227354
distance	0.686146091
winPoints	0.005982689

Source: This study.

Experimental Process

In the first stage, the three models are operated in the same way and the results are compared. Firstly, the data set after data cleaning and feature selection are imported. We defined function, the data set is divided into training set and verification set. This experiment set "Val"_ Perc = 0.2 ", that is, the data set is divided into training set and verification set according to the ratio of 8:2. After the data set is divided into training set and verification set, the dimension of the data set is checked. And adjust the model parameters continuously. Adjust the parameters, run the model and get the results. Different features have different importance in the prediction model based on Stochastic Forest algorithm. By calculating the importance index of different features in the current model, the results show that the importance of "killplace", "ridedistance", "boosts" and "weapons acquired" is relatively high for the prediction model. However, the importance of "headshot kills" and "revives" to the prediction model is low. We set the importance index greater than 0.005 as the more important features, and retain these more important features. According to these important eigenvalues, a new data set is generated, and the training set and verification set are redesigned. Then the model algorithm is used to set the parameters, and finally the model training and prediction are carried out again.



Source: This study.

Figure 1: Forecast results

In the second stage, we can only predict one player at a time, so we need to select some players for experiment. In this paper, we select the final chicken eating team (ranking first), predict the real-time ranking of each team member, and evaluate the results. First, the log events are mapped to the dataset to get the following table.

Table 2: Table log events correspond to datasets

Dataset1	Dataset2
assists	LOGPLAYERKILL
boosts	LOGPLAYERKILL
Damage Dealt	LOGPLAYERKILL
DBNOs	LOGPLAYERKILL
Head shot Kills	LOGPLAYERKILL
heals	LOGHEAL
killPlace	LOGPLAYERKILL
kills	LOGPLAYERKILL
killStreaks	LOGPLAYERKILL
LongestKill	LOGPLAYERKILL
revives	LOGPLAYERREVIVE
rideDistance	LOGVEHICLERIDE LOGVEHICLELEAVE
swimDistance	LOGSWIMEND LOGSWIMSTART
walkDistance	LOGPLAYERPOSITIN

	LOGPLAYERPOSITION
Weapons Acquired	LOGITEMPICKUPFROMLOOTBOX LOGITEMPICKUPFROMCAREPACKAGE LOGITEMPICKUP

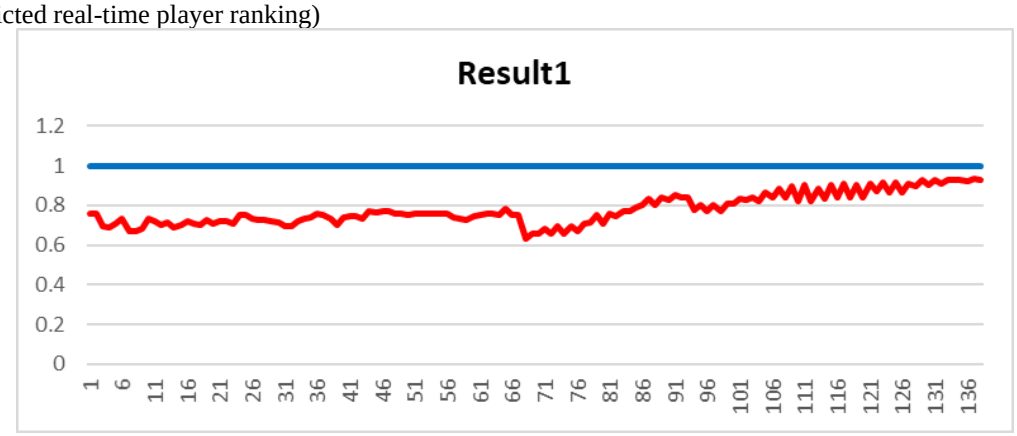
Source: This study.

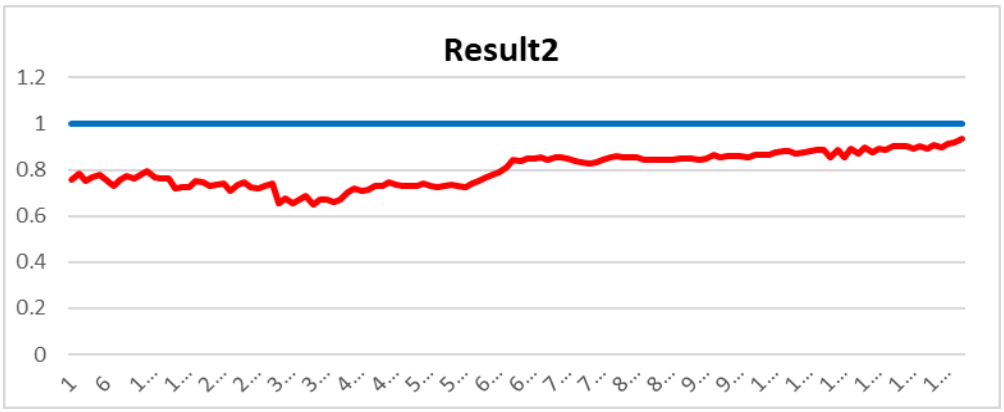
Then the log data of the target player is extracted, and the corresponding data items are obtained or calculated from the log data, and the time items are converted into formats. Finally, they are saved into different data item meta groups. Then get the start and end time and duration of the game. The next step is to encapsulate the function, which is called in real time during prediction. The function is to determine whether there are events in the time period. If so, the corresponding statistical items will be returned to form the data set of the input model. Then determine the time interval for real-time prediction and evaluate the results. Finally, we use Python's chicken diner and code base to visualize the log, generate game playback animation, and intuitively understand the game process.

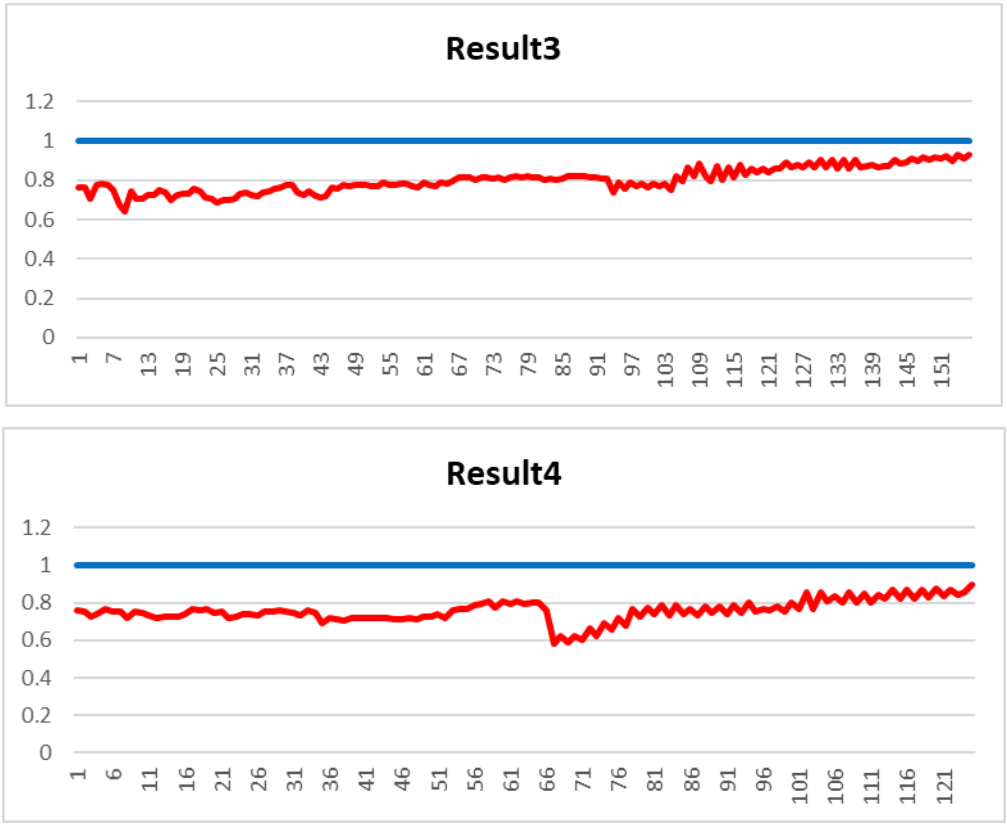
Experimental Results

In the first stage, the prediction accuracy was 0.9068, MAE was 0.06917, MSE was 0.0087 and R^2 was 0.9068. After training, the prediction accuracy was 0.9084, MAE was 0.06873, MSE was 0.0086 and R^2 was 0.9084. To view the importance of different features in the current prediction model based on random forest algorithm, among the 16 features included in the data set, "killplace", "ridedistance", "boost", "weaponsacquired" and other features are more important for the prediction model. However, the importance of "headshot kills" and "revives" to the prediction model is low. The prediction accuracy and Mae error of Light GBM model are 0.8997 and 0.0735 respectively. The final prediction accuracy of GBR gradient enhanced regression tree model was 0.9026, MAE was 0.06977, MSE was 0.00877 and R^2 was 0.9063.

In the second stage, the final winning rate of four players is 0.9303726, 0.93701894, 0.92822162 and 0.89602357 respectively. The fluctuation chart of real-time winning rate is as follows: (the blue broken line is the final player ranking; the red broken line is our predicted real-time player ranking)







Source: This study.
Figure 2: The fluctuation chart of real-time winning rate

RESULTS AND DISCUSSION

After comparison, we finally choose the model of random forest. For the random forest, the first criterion for evaluating the model is whether the importance of features is distinguished in the process of training, and whether the accuracy of prediction results is improved. From the results, the accuracy of the first experiment is 0.9068. After improving the feature ratio, the accuracy is 0.9084. Compared with the first test, the accuracy is improved by 0.0016. In other words, through the training of random forest, our prediction accuracy has been improved. At the same time, root mean square error is also an important index in error analysis. MSE was 0.0087 after the first calculation and 0.0086 in the second calculation. In contrast, the error is reduced by 0.0001, so the error has been controlled by random forest training. As for the goodness of fit, the R^2 of the first calculation is 0.9068. Then we sift the features to see how muchcontribution each feature makes to each tree in therandom forest, and then take an average value to compare the contribution between the features.In this process, the features of high shadow loudness are "killplace", "ridedistance", "boosts" and "weeks acquired", etc.; while the features of low influence degree are "headshot kills" and "revives". After removing unnecessary features and retaining the features with high correlation, the second calculation of R^2 is 0.9084, increased by 0.0016.

Table 3: Comparison table of prediction results of three models

prediction algorithm	Accuracy	MAE	MSE	R^2
Random forest	0.9084	0.0687	0.0086	0.9084
Light GBM	0.8997	0.0735		
GBR gradient enhanced regression tree	0.9026	0.0698	0.0088	0.9063

Source: This study.

In the aspect of real-time prediction, 10 seconds is taken as a time point to predict the player's ranking at this time. The average absolute error is obtained by comparing with the final ranking. We have got 4 result sets, and made a ranking fluctuation chart of players to analyze the factors affecting the fluctuation. As can be seen from the ranking fluctuation chart, there is a big gap between the real-time ranking and the final result at the beginning of the game. As time goes on, players' real-time ranking approaches the final ranking. This is because in the beginning of the game, players get less equipment, and the density of players is small in a large range of safety area. The probability of fire fighting between different teams is not as high as that of shrinking safety area in the later stage of the game. Therefore, in the process of ranking prediction, the feature related data acquisition is limited, which leads to a large error between the real-time ranking prediction and the final ranking in the early stage of the game, but very close to the final ranking in the later stage.

CONCLUSIONS AND RECOMMENDATIONS

For different features, we use random forest to screen out important features, and rank the features according to the influence degree to improve the algorithm. The results show that the "position of enemies", "moving distance", "number of props used", "degree of weapon destruction" are the most important features in the prediction ranking; the number of heads burst and the number of rescue members are the characteristics of lower importance in the prediction ranking. According to the experimental results, we get the final ranking of players under the given equipment and landing point, and the accuracy of prediction is about 92.29%.

we hope that the model can be further improved to make it more suitable for the actual situation, and the hypothesis premise in the hypothesis space can be more in line with our thinking mode when playing games.

CONTRIBUTIONS AND IMPLICATIONS

This paper aims to predict the player ranking in the game from a new perspective, so as to help the relevant personnel in the e-sports industry. The main contributions of this study are as follows:

Firstly, this paper proposes a real-time prediction method of E-sports. In the aspect of real-time prediction, we can see that the ranking of prediction is limited by the integrity of features at the beginning of the experiment, and the error is very large. In the process of the experiment, the predicted ranking tends to be stable and close to the actual ranking. We track the players in the game in real time, and analyze the impact of their behavior on the results, so as to help players adjust and improve their ranking in the game purposefully and strategically.

Secondly, this paper establishes a model of feature selection, and obtains the weight and change of the characteristics that affect the final ranking in the competition. From the results we get, the influence of characteristics on winning also changes with the number of features. In a certain range, the more weapons and healing items players pick up, the greater the proportion of winning; however, when the number exceeds this range, the winning proportion will not increase with the increase of the nu mber of these two kinds of equipment. The weight of these characteristics and the situation of weight changing with the competition process are of great significance to guide the strategic selection of the team in the E-Competition (Sánchez-Ruiz & Miranda, 2017).

Thirdly, the research results have great practical significance for ordinary users of E-sports games, and can improve their understanding and experience of the game. At the same time, it can also consolidate the mass base of the e-sports industry and promote the development of the whole industry (Ding, 2018). For the e-sports industry, E-sports is an industry with a huge mass base and the number of people who pay attention to it continues to rise (Nascimento Junior *et al.*, 2017). Among various types of games, shooting games represented by "eating chicken" are welcomed by 48.4% of users, second only to MOBA games. We choose the prediction of e-sports, and the results can help a lot of people. For the vast number of game users, by helping users analyze the problems that may be encountered in the game, such as where the first jump is safer, where the materials are the most abundant, and what is the refresh rule of the security zone (Hapfelmeier & Ulm, 2013). The research results help game users improve their game experience. At the same time, it makes the game happy and simple, easier to score. For professional E-sports related personnel, data analysis provides some potential game information for professional players and clubs to help them get higher scores. E-games are becoming more and more popular, so for the spectators watching professional league matches (Jenny *et al.*, 2017), by analyzing the starting position of the team and obtaining material performance, they can more accurately carry out entertainment quiz.

LIMITATIONS AND FUTURE RESEARCH

But our research also has some limitations. Because there are uncontrollable interference factors in the game, such as players' psychological state, players' network fluctuation factors and so on. These data limit the accuracy of the prediction results to a certain extent. We can't get exact control when we control variables, so we don't include uncontrollable factors in this model.

In the future research, the distribution, selection and definition of features in different maps should be updated with the updating of maps. And the source and collection of data set should be updated in time, so that the integrity of prediction results is better, the representativeness of training set is stronger, and the accuracy of test set is higher. To sum up, the research on real-time prediction of player's game ranking makes up for the deficiency of current game prediction algorithm to a certain extent, and provides help for the theoretical supplement of machine learning related algorithm and relevant personnel engaged in e-sports industry.

REFERENCES

[1] Bosc, G., Kaytoue, M., Raïssi, C., Boulicaut, J. F., & Tan, P. (2014, August). Mining balanced sequential patterns in RTS games 1. In *ECAI 2014-21st European Conference on Artificial Intelligence*. https://hal.inria.fr/hal-01100933/

[2] Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5-32.

[3] Carrillo Vera, J. A., & Aguado Terrón, J. M. (2019). The eSports ecosystem: Stakeholders and trends in a new show business. *Catalan Journal of Communication & Cultural Studies*, *11*(1), 3-22.

[4] Chen, P., Niu, A., Jiang, W., & Liu, D. (2019, January). Air pollutant prediction: comparisons between LSTM, Light GBM and Random Forest. In *Geophysical Research Abstracts* (Vol. 21).

[5] Ding, Y. (2018). Research on operational model of PUBG. In *MATEC Web of Conferences* (Vol. 173, p. 03062). EDP

Sciences. https://doi.org/10.1051/matecconf/201817303062
[6] Hamari, J., & Sjöblom, M. (2017). What is eSports and why do people watch it? *Internet Research*, 27(2), 211-232.
[7] Hapfelmeier, A., & Ulm, K. (2013). A new variable selection approach using random forests. *Computational Statistics & Data Analysis*, *60*, 50-69.
[8] Jenny, S. E., Manning, R. D., Keiper, M. C., & Olrich, T. W. (2017). Virtual(ly) athletes: where eSports fit within the definition of "Sport". *Quest*, *69*(1), 1-18.
[9] Jonasson, K., & Thiborg, J. (2010). Electronic sport and its impact on future sport. *Sport in society*, *13*(2), 287-299.
[10] Keiper, M. C., Manning, R. D., Jenny, S., Olrich, T., & Croft, C. (2017). No reason to LoL at LoL: the addition of esports to intercollegiate athletic departments. *Journal for the Study of Sports and Athletes in Education*, *11*(2), 143-160.
[11] Kinkade, N., Jolla, L., & Lim, K. (2015). Dota 2 win prediction. *Univ Calif*, *1*, 1-13. http://jmcauley.ucsd.edu/cse255/projects/fa15/018.pdf
[12] Nascimento Junior, F. F. D., Melo, A. S. D. C., da Costa, I. B., & Marinho, L. B. (2017, October). Profiling successful team behaviors in League of Legends. In *Proceedings of the 23rd Brazillian Symposium on Multimedia and the Web* (pp. 261-268).
[13] Pereira, R., Wilwert, M. L., & Takase, E. (2016). Contributions of sport psychology to the competitive gaming: An experience report with a professional team of league of legends. *International Journal of Applied Psychology*, *6*(2), 27-30.
[14] Pluss, M. A., Bennett, K. J., Novak, A. R., Panchuk, D., Coutts, A. J., & Fransen, J. (2019). Esports: the chess of the 21st century. *Frontiers in psychology*, *10*, 156.
[15] Rahman, M. A. (2020). A deep learning framework for football match prediction. *SN Applied Sciences*, *2*(2), 165.
[16] Sánchez-Ruiz, A. A., & Miranda, M. (2017). A machine learning approach to predict the winner in StarCraft based on influence maps. *Entertainment Computing*, *19*, 29-41.
[17] Xiong, S., Zuo, L., & Iida, H. (2014). Quantifying engagement of electronic sports game. *Advances in Social and Behavioral Sciences*, *5*, 37-42.
[18] Ziegler, A., & König, I. R. (2014). Mining data with random forests: current options for real-world applications. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *4*(1), 55-63.