International Journal of Applied Engineering Research ISSN 0973-4562 Volume 13, Number 5 (2018) pp. 2814-2816

Predictive Analysis of Sports Data using Google Prediction API

Ujwal UJ[1], Dr Antony PJ[2] and Sachin DN[3]

[1]KVG College of Engineering, Sullia, India.
[2]AJ Institute of Engineering and Technology, Mangalore, India.
[3]Vidyavardhaka College of Engineering, Mysuru, India.

Abstract

Prediction system is very essential to curb the curiosity of anything. Many sports prediction systems are in great demand and data analysis plays a great role in prediction. Previous efforts in sports data analysis have resulted in prediction of sports such as football, prediction of next shot location in tennis, performance of athletes in Olympics, slam dunk shots frequency in Basket Ball and many more. Cricket prediction is comparatively difficult as there are many factors that can influence the result or outcome of the cricket match. Earlier basic prediction systems for cricket match consider only the venue and disregard the factors like weather, stadium size, captaincy etc. The factors like venue of the match, pitch, weather conditions first batting or fielding all play a vital role in predicting the winner of the match. Suitable models are necessary to predict and data mining makes it possible to extract required information from the data files. This paper presents the usage of Google Prediction API to analyse the data of previous cricket matches and predict outcome of a given cricket match.

Keywords: Google Prediction API, Linear Regression, Naive Bayes

INTRODUCTION

The most popular sports in the world after Football is Cricket. Cricket is a very popular game in India, Pakistan, Australia, England, South Africa and many more countries and it has billions of fans across the globe. Especially in India, many fans say "Cricket is my religion". The game is played between two teams, each consisting of 11 players (15 when extra players are included) .The main part of the game takes place in the pitch of 22 yards length. It is played in 3 different formats, viz. One Day International (ODI), Twenty Overs International (T20) and Test. A bowler can bowl 6 successive legal deliveries and such 6 deliveries are called as an "over". Initially depending on the toss one team shall bat first and the other team shall bowl. That will be the first innings. In the second innings, the team which batted first shall bowl and the other team will bat. One Day is played for 50 overs, each innings. Test is played for 5 days and for the first four days a minimum of 90 overs is to be played and on the last day a minimum of 75 overs [1] is to be played. T20 is called limited overs match as only 20 overs per innings. The pitch will be of 22 yards and the inner circle will be of 30 yards.

The batsmen have to score runs by hitting the ball bowled at him by the bowlers. The batsman gets out on certain conditions based on specific rules set. Until the batsmen gets out, he can keep batting till the end of innings. The bowling team gets a wicket when each batsmen is out. An innings gets over when either 10 wickets are down or when the specified number of overs is reached. The runs scored by the first team has to be chased by the second team in order to win. If the team which bats first defends their total, they win.

Prediction system works on the principles of machine learning. There are two types of machine learning namely supervised machine learning and unsupervised machine learning. In supervised machine learning we must train the machine by providing huge data sets and the outcomes. In this paper one such prediction methods is introduced which is used to make predictions of the outcome of a cricket match using Google Prediction API.

Google API is a black box prediction technique [2]. It is a form of supervised learning and hence it is required to provide huge data and train the models. Google Prediction APIs make use of Regression Algorithms when numerical predictions have to be made. And Classifiers when the target output can assume only a limited set of values, either numbers or strings, based on the application content. This API can only account for relatable data. If the attributes are not related to one other, then a correct probability curve will not be drawn. By providing a CSV file of previous cricket matches and using appropriate queries to extract the required data and train the model, predictions can be made.

RELATED WORK

"Winning and Score Predicting (WASP)" [3] is a work done by Scott Broker and Seamus Hogan at University of Canterbury as a part of their PhD research project. It predicts the additional runs that can be scored with the remaining wickets and balls in the first innings. In the second innings it estimates the winning probability.

Tejinder Singh., proposed a model that has two methods to predict. First method predicts the score of first innings using Linear Regression Classifier and the second method uses Naive Bayes Classifier to predict the outcome of the match in second innings. [4]

Swartz et al. made use of Bayesian Latent variable model and Markov Chain Monte Carlo [5] method to simulate ball by ball outcome. Kaluarachchi and Varde implemented association rules and Naive Bayes classifier to analyze the factors that contribute to a team's win. [6]

METHODOLOGY

A. Data Collection

The data collection is the initial step. Data sets of previous cricket matches are available in many websites and we collected the data from www.cricksheet.org and www.espncricinfo.com. The files are in CSV format and this is the required format for Google API. The data set will have all the information of a cricket match such as team names, gender, season, date, venue, city, toss winner, toss decision, player of the match, umpires, TV umpire, winner, winner runs, total score and ball to ball information .

B. Classification

Classification technique is the one that does most of the job by identifying the type of the class an instance belongs to in machine learning. As Google Prediction API is a supervised learning, classifiers are used by it to train data models by giving the data of already correctly identified instances. By analyzing these training data set, the class type of newer instances is predicted.

1) Regression

Regression is used whenever the target output corresponds to numerical values. Linear Regression algorithm works in a ways as to get the expression of the class using attributes and weights. Regression models are used when the output returns a numeric value as the prediction result. Regression model estimates a numeric answer for a question given the closeness of the submitted query to the existing examples.

2) Categorization

Google Prediction API does not just works on numeric data. It also works on string data values. To have such a categorical prediction there are many classifiers and one such classifier is Naive Bayes. Since Google Prediction API is a black block, the exact implementation of the classifiers is not known. The type of model where the output returns a string value is called Categorical Model. Categorical models determines the closest categorical fit for a submitted query among all the example training data sets provided.

C. Steps Involved

Step 1: Build Training Model

The data to must be entered into the Google Cloud Storage. A separate 'Bucket' is created in the Cloud SDK and all the data files are uploaded to that Bucket. Since there are multiple data files, python script is used to extract necessary data from all the files and combine those data into one or fewer CSV files. This helps to have a better accuracy as the comparison becomes easy. Table 1. Shows the attributes that are taken into consideration for prediction. The data entries represent the cricket matches played by one team against rest of all the teams. So all the attributes will be with respect to that fixed team. The first column in the data model will be the result of the match. It will be a string "Win" or "Lose" which says whether that fixed team (Team 1) has won the match or lost it. The next attribute is the name of the team (Team 2) against which the fixed team is going to play, this is again a string value. The next attribute is the decision of the toss. Whether the fixed team goes to bat first or bat second. Next is “Team Strength". Normally we only look in to the current team strengths. But when we provide older data, the team which is strong presently might not have been strong at that point of time. Say 2017 ODI rankings have South Africa at the top one position, but this is not the case in 2013 when South Africa was in the 4th position. The rankings are relative, so even if a particular country does not change its players, the changes made by other countries can make alterations in the rankings table because the newly formed team might perform better and score more points in the rankings table. Though international cricket matches were played as early as 1930's, too much old data is not useful. The players who constituted the team are no longer in the team. So we collected the details of the last 10 years (from 2006 onwards). We took the rankings of the teams year wise [7] and while entering the data, we considered the date of the match. The team strength was then assigned with respect to the year on which the match was played. This assures that the varying team strength is also taken into account. The last attribute is the venue where the match takes place.

Table 1: Attributes in the data set

Sl.No	Attribute Names	Description
1.	Outcome	Whether a team has won the match or lost it
2.	Team 2	The team against which a fixed team is playing
3.	First or Second Batting	Whether the fixed team will bat first or second
4.	Team Strength	Strength of the team against which fixed team is going to play
5.	Venue	Place where the match takes place

Step 2: Calling Prediction API

Once the data is ready in the Cloud it is time to call the Prediction API. Queries are sent to the API to relative to the project created and after authentication is done the data models have to be trained by calling prediction. Trained models insert method to the API. After the training is complete, query is sent to the API to get the desired prediction.

RESULT ANALYSIS

The proposed model will give results based on the data provided previously. The better the data model is trained, the better the results will be. The outcome will be of the fixed team. So when the result is "Win", it means the fixed team wins the match, if the result is "Lose" , then the fixed team will lose the match.

A. Comparison of results.

The training data model is separated from the testing model. After the training is done, the testing data is used to check the accuracy of the prediction system. Table 2. Shows the result of the cricket matches between India and other teams and the prediction of our prediction system.

Table 2: Actual and predicted outcomes

Sl no	Team 1	Team 2	Actual result	Predicted result
1.	India	Afghanistan	Win	Win
2.	India	Bangladesh	Win	Win
3.	India	Australia	Lose	Lose
4.	India	SouthAfrica	Lose	Lose
5.	India	Newzealand	Win	Win
6.	India	Australia	Win	Win
7.	India	Srilanka	Win	Lose
8.	India	Ierland	Win	Win
9.	India	England	Lose	Lose
10.	India	WestIndies	Win	Win

We observe that out of 10 matches, 9 have been predicted correctly. It is also noteworthy that if the data is less, the accuracy is also less. Say in a particular venue very little matches have been played, then it is difficult to predict.

Also the results of the prediction can be stored and used as a reference when the actual outcome of the match is known and then the correctness percentage of the prediction can be calculated.

CONCLUSION

Many works are being done in the field of prediction of sports matches. Analysis of sports data and foretelling the future is a hectic task. Data mining technique is very essential here. Google Prediction API chooses the best classifiers for predicting the values. This paper can help to understand the capabilities of machine learning and data mining and also help in player analysis by the sports committee. Knowing how the match turns out will be a boon for the captains of the team to have a quick change in strategy and make the game even more competent to play and interesting for the audience to watch. This will also help to have a better chance of winning fantasy cricket leagues. The future scope of this paper will be to consider sentimental analysis to understand the mood of the players and combine sentimental analysis and statistical data to provide an even better prediction system.

REFERENCES

[1] The Laws of Cricket https://www.lords.org/mcc /laws-of-cricket/ Accessed 2017

[2] Google Prediction API provides a RESTful interface to build Machine Learning models**.** http://cloudacademy.com/blog/google-prediction-api

[3] Seamus Hogan (2012) Cricket and the Wasp: Shameless self promotion (Wonkish).http://offsetting behaviour.blogspot.co.n z/2012/11/cricket-andwasp-shameless-self.html

[4] Tejinder Singh, Vishal Singla, Prateek Bhatia, Thapar University Punjab "Score and Winning Prediction in Cricket through Data Mining" 2015 International Conference on Soft Computing Techniques and Implementations (ICSCTI) Department of ECE, FET, MRIU, Faridabad, India, Oct 8-10, 2015

[5] T. B. Swartz, P. S. Gill, and S. Muthukumarana. Modelling and simulation for one-day cricket. Canadian Journal of Statistics, 37(2):143{160, 2009}

[6] A. Kaluarachchi and A. Varde. CricAI: A classification based tool to predict the outcome in ODI cricket.In 5th International Conference on Information and Automation for Sustainability, pages 250{255, 2010}

[7] https://web.archive.org/web/20130120040151/ http://www.icccricket.com/match_zone/historical_ranking.php