Indian Institute of Science Education and Research Bhopal

Department of Electrical Engineering and Computer Science

Winner Prediction in a Cricket Match

Author:
Deep Pooja (17074)

Supervisor:
Dr. Kundan Kandhway

Submitted in partial fulfillment of the requirements for the BS degree in Type of Degree of IISER Bhopal

May 15, 2021

Abstract

Cricket is the most popular sport in India and second most popular sport in the world after soccer. Due to its popularity, Cricket is a big money market, there is a great demand for prediction of the match winner, which can be used for betting purposes however predicting the outcome of a cricket match prior to it begins depends on complex rules governing the game, team strength, along with numerous natural parameters such as pitch condition, weather, winning toss and batting first etc. In this project, we predict the outcome of a cricket match (especially ODI matches) prior to it begins based on the player's performance of each team. To calculate the performance of each player, we use the ball by ball data of all matches in which the player has already played. We choose four performance measures i.e two batting performance measures viz. Strike rate and batting average and two bowling performance measure viz. Economy rate and bowling average then we will use this performance matrix as feature set to feed into machine learning models for prediction.

Contents

Chapter 1

Introduction

Cricket is a game played with bat and ball wherein two teams, each of 11 players, compete against each other on a 20 metre pitch by scoring runs and taking wickets. At a given time, the batting team has 2 players on the pitch and the bowling team has 11 players on the field one of which is a bowler. Each player on the batting team scores runs by running between the pitches or hitting boundaries(4 or 6 runs). The team score is the total score of each individual player. A given match is of selected number of overs, that is each team will have a maximum number of balls to play for. The aim of the batting team is to score as many runs as possible. The bowling team aims to restrict the score of the batting team, as well as, looks to take wickets. Wickets are taken by getting the batsman caught, bowled, stumped, leg-before-wicket, or run-out. Each team has 10 wickets, so the aim of the bowling team is to get the 10 wickets as soon as possible. The goal is to outperform the other team by scoring more runs. This Figure 1.1 is a typical snapshot of a cricket match.

After Soccer, Cricket is the second most popular sport in the world. There are around 104 countries playing this sport officially being a part of International Cricket Council. Cricket is a huge business market, especially in India. There are hundreds of statistics to compare and analyze from, each being important in their own

Figure 1.1: A snapshort of a cricket match

way. Cricket being a big money market, there is a great demand for prediction of the match winner, which can be used for betting purposes. Cricket data of various leagues and matches is available in different formats in various sports websites like ESPN, Yahoo Sports etc. The data is also available in a cleaned format on various other platforms like [2]. A successful model would be able to determine the results of the match or performance of individual player with a better accuracy. The objective of this project is to explore the viability of predicting the outcomes of a match for a team, and also explore the various challenges involved in predicting the performance of individual players. We have compared the performances of various classification algorithms like Logistic Regression Classifier, Decision Trees, Random Forests, Support Vector Machines, and Deep Neural Networks. We have used Python 3.6 for our project. We have implemented our Machine Learning algorithms using sklearn package. To implement the Neural Network we have used Keras with Tensor Flow as backend.

Chapter 2

Background

Data-driven approach to predict the winner in a cricket match has already been studied by many people. One of these papers is CricAI [3], which came in 2010, they built a software tool which predict the winner in ODI matches using machine learning techniques specifically bayesian classifiers with given factors as day/night effect, toss won bat-first and home advantages, noticeably they have not incorporated player's performance in their model. Another paper Auto-play [4], which came in 2014, they did ODI cricket match simulation and prediction i.e they feed the model with two instances first being the history of the team's performance (historical features) and instantaneous state of the match (instantaneous features). The historical features they used are: average run scored in an innings, average number of wickets lost in an innings, average run conceded in an innings, frequency of being all-out, frequency of getting opposition all-out, average number of opponent wickets taken in an innings. These features do measure team's strength as a whole but fail to capture individual player capability or performance which is crucial here as players evolve dramatically over matches or get retired over time. The reason we are focusing on player's performance is because we hypothesize that player's skills and team's strategy (viz consequence of player's capability) is the most important factor to win the game.

Lastly the paper team composition based approach [5], their approach lies more or less on what we are trying to do here but with significant differences. They used carrier statistics (data they mined from suitable websites) and recent performance of players. They also divided the team into two groups batsman and bowler and calculated batsman's score and bowler score for each team. Batsman score is the sum of career score and recent score which is nothing but average of runs scored in recent matches and similarly bowler score is career score, they have not considered the recent performance of bowlers due to lack of data. Their features are toss, venue and relative team strength that is they took the ratio of bat_strength to ball_strenght of each team and subtracted the ratios. In their algorithm to compute a player's performance, they have defined their own measures with constant multipliers without giving any proper explanation of why this way, also they do normalize the scores to avoid overfitting of models. Instead we choose to use standard defined measures.

Chapter 3

Contribution

Our data set consists of ODI matches. It contains ball by ball data of every match, giving information about runs scored, wickets taken, type of wicket taken, bowler and batsman, number of overs for every ball. This data set can be used to analyze and view team information and performance over the years, we can also visualize how a batsman, or a bowler has performed or statistics about the highest run scorer or highest wicket taker. We decide to go one step further with this data, we wanted to leverage the fact that this data set gives us ball to ball information of every match. This was a crucial information, with which multiple predictions and conclusions were possible. As we had data about the team matches played, it was possible to predict the winner based on past matches. Many features given in the data set like venue, bowler to batsman ball by ball performance, can be used to somewhat predict the winner of the match. Thus, we could predict the winner of the next International ODI match based on the right feature selection.

Data Visualization We have also explored the dataset by visualising.
1. Average Runs scored by Indian Cricket team in ODI over the years
We plotted Figure 3.1 the average number of runs scored in a match for each year from 2006 to 2017.

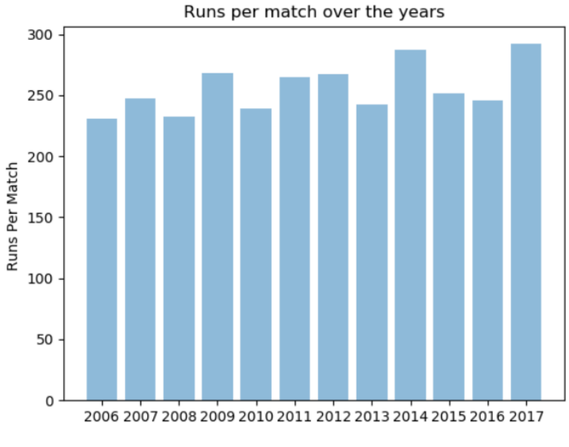

Figure 3.1: Years

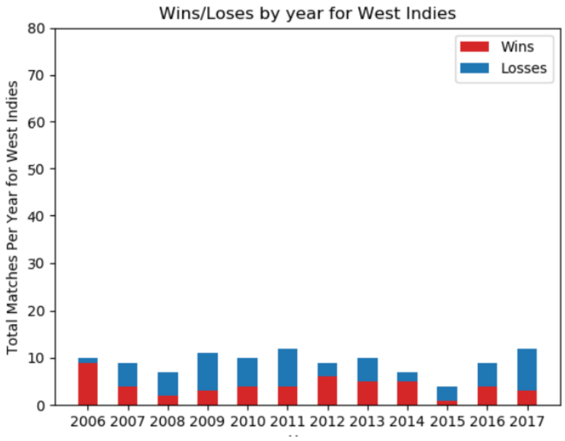

Figure 3.2: Years

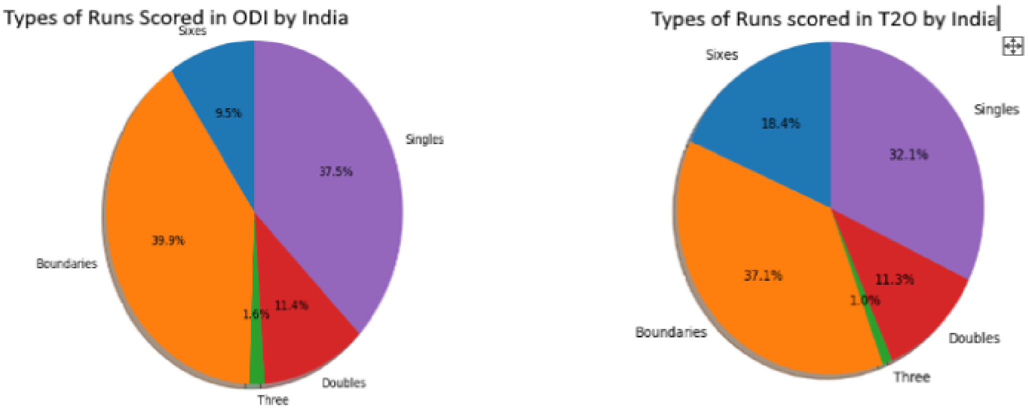

Figure 3.3: ODI and T20 Comparison for India

2. Wins/Losses of West Indies Cricket Team over the years
In the above Figure 3.2 of wins/loses of West Indies over the years, we can see a decrease in the number of wins for the West Indies from the year 2006 to 2017. We can see that West Indies played 11 matches in the year 2006 and won 9 of them, but in the year 2017 it played 12 matches and lost 9 matches.

3. Type of Runs scored by India in T20 and ODI
Comparing Figure 3.3 the type of runs scored in ODI and T20 by the Indian Cricket team we can see that, as expected the percentage of sixes scored in T20 is higher than that of ODIS. We can infer that, T20 follows more offensive play than the ODIs

To start with, we choose to work with ODI match format but our approach could easily be extendable to any cricket match format (test matches, T20s IPL ect). As already mentioned, we decided to compute player's performance using standard measures viz bowling average, batting average, economy rate and strike rate which defined as :
Bowling average: the number of runs they have conceded per wicket taken.[lower the the value better the bowler is]

Batting average: the ratio of a player's number of runs to the number of times they have been out.[higher the value better the batsman is]

Strike rate: It is calculated by the runs scored by batsman divided by the number of balls he faced multiplied by 100.[higher the value better the batsman is]

Economy rate: the average number of runs conceded for each over bowled.[lower the value better the bowler is]

To compute the feature matrix of 22x4, that is four performance measures and 22 players, 11 players from each team, we need to scan through each ball by ball match data and compute run scored by each batsman in the match, run conceded by each bowler in the match, wickets taken and number of time a batsman has been out apparently (0 or 1) in a match and stored these values in a dictionary with name as key and list of values as value. After scanning through relevant matches we calculate our features as per the definition and feed these features to machine learning models.
For example, the ODI matches between West Indies and India has only 53 matches that is we have only 53 data points and each data point is a 22x4 matrix, clearly the data is sparse and this is the case for almost any two teams we pick and this might lead to overfitting so to avoid this we have done dimensionality reduction of data by PCA and also only incorporate top 5 batsman and bowlers each team by ranking them subjectively.
Another problem we faced which is the bowler might not have batted so it's batting average and strike rate could be zero however a batsman who has not bowled we can't just put zero in economy rate and bowling average because that implies they overwhelmingly good bowlers but it's the opposite so we put NAN values.
The number of matches between two teams (say India and west Indies) was our training data. We split this into two sets training data and test data. We found the class labels for all our training files, this class label was 1 if the first team won, or 0 if the first team lost. Team1 or Team2 is taken according to user input. Similarly, we found the feature vector for our test data. We validated multiple models on our data, like SVM, Decision Trees and Neural Network, and finally compared the results of these algorithms

We also liked to make Prediction if a given team could win while chasing. In this we needed to keep track of how each team has performed while batting first against our target team. We maintain a similar feature vector as our first case. In our prediction stage, we do not check the 2nd innings score of our team, and on the basis of team batting first's runs scored we make the prediction.

Chapter 4

Experimental Results

Dataset: We got the data of ball by ball matches of all ODI's matches played till date from a publicly available source cricsheet [2].

Result Table

S.No.	Classifier	Acc Team Win(%)	Acc Team Win while chasing(%)
1.	Logistic Regression	55.867	52.13
2.	Decision Trees	53.23	51.81
3.	Support Vector Machines	73.776	61.63
4.	3 hidden layer neural network	67.12	59.22

We have compared the performances of Logistic Regression Classifier, Decision Trees, SVM, and Neural Networks. We can see that SVM gave us the best results, even better than neural network, the reason behind this is that, our training data is very less, and for neural networks to work best we need large training dataset. Thus, it is reasonable to see than SVM has outperformed neural networks in our case. Our best accuracy is around 73% which is very good for a game like cricket where a lot of uncertainties are involved and the chance accuracy is of 50%. Thus models of data were prepared and analyzed, completing the third and fourth step of Modelling and Evaluation

Chapter 5

Conclusion

Generally, cricket matches are very hard to predict, as right from the toss and the weather condition, a lot of uncertainties are involved. One can easily predict the winner of the match at the beginning with a 50% accuracy as, either of the team has the equal chances of winning. Also, the circumstances of the match like the weather, pitch conditions change over the course of the match which are difficult to account for. We attempted to observe the performance of machine learning models on predicting match winners, and we have achieved a score of around 73% which is way better than chance accuracy of 50%. Creation of better datasets which takes into account various details like length, speed, swing of each delivery will enable player performance prediction for a given delivery.

One could also think of making prediction for a batsman vs bowler, that is for a given batsman and a given bowler, predict what would happen on the next ball, would it be four, six or a wicket? But as we got into the data exploration phase, we realized that our data was unsuited for this task. To make such kinds of predictions we need quantitative information of the swing, spin of the ball, the length of the delivery, the pitch condition. We have stadium information, but even that information is insufficient as, the pitch and outfield quality depends upon the current weather, which is missing in our data. Sites like ESPNcricinfo [1] maintain such data, so, parsing data for each match could give us this dataset.

Bibliography

[1] ESPN. ESPNCricinfo http://www.espncricinfo.com/.

[2] S. Rushe. cricksheet. https://cricsheet.org/downloads/#experimental.

[3] https://ieeexplore.ieee.org/abstract/document/5715668/.

[4] https://epubs.siam.org/doi/pdf/10.1137/1.9781611973440.

[5] https://dtai.cs.kuleuven.be/events/MLSA16/slides/06_Madan_Gopal.pdf.