Cricket Match Score Prediction using k-Nearest Neighbors Algorithm

T. Suvarna Kumari^1, P.Narsaiah^2

^{1}*Computer Science & Engineering, Chaitanya Bharathi institute of Technology, Hyderabad, Telangana, India*

^{2}*Computer Science & Engineering, Vasavi College of Engineering, Hyderabad, Telangana, India*

Abstract - Cricket is one of the most popular team sports, with billions of fans all over the world. In this paper a model has been proposed to predict the scores of ODI cricket matches using a team-composition based approach before the start of the match. However, a team changes its composition depending on the match conditions, venue, and opponent team, etc. Therefore, in this paper a model has been proposed which takes into account the varying strengths of the individual players and reflects the changes in player combinations overtime. The relative team strength between the competing teams forms a distinctive feature for predicting the winner. Modeling the team strength boils down to modeling individual players batting and bowling performances, forming the basis of the paper. Career statistics as well as the recent performances are used to model a player. Using the relative strength of one team versus the other, along with two player-independent features, namely, the home advantage and the venue average of the match the scores for the first innings and second innings are predicted using the k-Nearest Neighbor Algorithm. The approach is implemented on the past ODI matches from 2000 to 2018 and from the results it has been found that the error rate of first innings in 22% and the error rate of second innings is 18%.

Keywords— Regression; k-Nearest Neighbors; Machine learning; Projected Score.

I. INTRODUCTION

Statistical modeling has been used in sports over decades and has contributed significantly to success on the field. Cricket is one of the most popular team sports in the world, second only to soccer. Various natural factors affecting the game, enormous media coverage, and a huge betting market provide strong incentives to model the game from various perspectives. For instance, Duckworth and Lewis proposed a solution,called D/L method[1], to reset targets in rain interrupted matches which was adopted by the International Cricket Council (ICC) in 1998. However, the complex rules governing the game, the ability of players and their performances on a given day, and various other natural parameters play an integral role in shaping the course of a cricket match. This presents significant challenges in modeling the game The batsman looks for making runs by hitting the ball being bowled to him. The bowler on the other hand tries to get the batsman out. There are certain rules defined to get the batsman out by the bowlers or the fielders. Each batsman keeps on batting until he gets out. So, the innings of the batting team is over when either the 10 batsmen got out or the 50 overs have been bowled by the fielding team; in either of the situation the batting team now gets the chance of bowling and the bowling team gets the chance of batting. The team which scores more runs wins the match. Unlike other sports, cricket stadium's size and shape is not fixed except the dimensions of the pitch and inner circle which are 22 yards and 30 yards respectively. The cricket rules do not mention the size and the shape of the field of the stadium [2]. Pitch and outfield variations can have a substantiate effect on batting and bowling. The bounce, seam movement and spin of the ball depends on the nature of the pitch. The game is also affected by the atmospheric conditions such as altitude and weather. A unique set of playing conditions are created due to these physical differences at each venue. Depending on these set of variations a particular venue may be a batsman friendly or a bowler friendly. Currently, in an ODI match the projected scores can be seen displayed at the score card during the first innings, which is basically the final score of the batting team at the end of that innings if it scores according to the current run rate or a particular rate. Run rate is defined as the amount of runs scored per the number of overs bowled. However, run rate is considered as the only criteria for calculating the final score. But there are other factors too which may affect the final score like number of wickets fallen, the venue and the batting team itself.

In this paper, a method has been proposed in which the final score can be predicted of the first innings and the second innings. K-Nearest Neighbors algorithm is implemented to predict the match scores. Factors like Relative team strength, Home, and Venue average has been considered for the prediction. These past records have been taken from all the non-curtailed ODI matches played from 2000 to 2018. The structure of the paper is as follows. In the following section the related works done in the game of cricket or any other sports have been discussed briefly. In section III, an overview on regression has been given and the algorithms implemented for predicting the final scores have been demonstrated. Section IV focusses on the data collection and preparation while section V discusses about training and testing of data. In section VI, the statistical analysis has been done and it has been found that the error rate of second innings is slightly less than the error rate of first innings. Conclusion and future scope are given in Section VII.

II. RELATED WORK

Very few have worked in statistically predicting the scores or the outcome of the ODI match. In literature, Duckworth and Lewis proposed a solution, called D/L method [1], to reset targets in rain interrupted matches. It is designed so that neither team benefits or suffers from the shortening of the game and so is totally fair to both. It is easy to apply, requiring nothing more than a single table of numbers and a pocket calculator, and is capable of dealing with any number of interruptions at any stage of either or both innings. “Winning And Score Predicting (WASP)”, which has been done by Scott Brooker and Seamus Hogan at University of Canterbury as part of the PhD research project [2]. It estimates about how well the average batting team will do against the average bowling team under given conditions and the current state of the game. In the first-innings it estimates the additional runs that can be scored with the given number of balls and wickets remaining. In the second innings it estimates the winning probability with the given number of balls and wickets remaining, runs scored at the given situation and the target given. The estimates have been made from a dynamic programming [2].

A.Kaluarachchi and A.S.Varde[3] studied several interesting factors including home game advantage, day/night effect, winning the toss and batting first. Further, they used artificial intelligence techniques, more specifically Bayesian classifiers in machine learning, to predict how these factors affect the outcome of an ODI cricket match.

U.B.Swartz, P.S.Gill, D.Beaudoin, etal. [4] proposed a method for optimal or nearly optimal batting orders in one day cricket by conducting a search over the space of permutations of batting orders where simulated annealing is used to explore the space. H.H.Lemmer [5] proposed a method to measure strangling, a dramatic form of choking in cricket. In limited overs cricket, the team batting first sets a target for the team batting second. The latter team may win the match, draw the match or lose it by not reaching the target. M.Bailey and S.R.Clarke[6] used past data to create a range of variables that could independently explain statistically significant proportions of variation associated with the predicted run totals and match outcomes.

III. REGRESSION

In machine learning, Regression[7] is a statistical processes for estimating the relationships among variables. It includes many techniques for modeling and analyzing several variables, when the focus is on the relationship between a dependent variable and one or more independent variable. More specifically, regression helps one understand how the typical value of the dependent variable changes when any one of the independent variables is varied, while the other independent variables are held fixed. Regression is a supervised learning in which a training set of correctly identified instances is given.

K-Nearest Neighbor Algorithm

In pattern recognition, the k-nearest Neighbor algorithm (k-NN) is a non-parametric method used for classification and regression.[8] In both cases, the input consists of the k closest training examples in the feature space. In k-NN regression, the output is the property value for the object. This value is the average of the values of its k nearest Neighbor.

In this paper, k-Nearest Neighbor algorithm have been implemented for the first innings and second innings datasets where the class attribute ‘X’ is the ‘Score’ and the input attribute ‘a_i’ (i =1,2,3…) are the relative team strength, home/away and venue average.

IV.DATA COLLECTION AND PREPARATION

The data has been collected from http://stats.espncricinfo.com[9], where over-by-over data of all the matches are available publicly. The dataset consists of complete matches excluding all the rain-interrupted and rain-abandoned games, played between 2000 and 2018 among the ODI playing teams like Australia, India, New Zealand, South Africa, England, Sri Lanka, Pakistan and West Indies etc.. Also it contains the dataset of the matches played on the venues from each country.

For each team two separate datasets have been made, one for first innings and the other for the second innings. Similarly for each venue also two datasets are there one for average first innings scores and second for the average of second innings scores. Now, for the first innings those complete matches have been taken where the particular team has batted first in the first innings. For example, the first innings dataset of India contains the record of those matches only where India has batted first. From these matches the runs scored, opposition, home team, venue, along with the final score at the end of the innings, have been considered.

In the second innings also those matches have been included in which the particular team has batted in the second innings only. For example, the second innings dataset of India contains only those matches in which India has batted in the second innings. In addition to this, from these matches the runs scored, opposition, home team, venue, first innings score along with the final score at the end of the innings, have been considered.

MySQL databases are used to construct the three important features namely relative team strength, home/away and venue averages. Relative team strength is constructed by summation of the batting strength of all the players in the batting side divided by the summation of the bowling strength of all the bowlers in the bowling side.

Algorithm for modeling Batsmen :

```
1. u=Square root (Innings Played/Matches Played)
2. v=x1*Centuries+x2*Half Centuries
3. w=x3*Batting Average+x4*v
4. Career Score=u*w
5. Recent score=Avg(last5scores)
6. Total Rating=x5*Recent Score+(1-x5)*Career Score
```

Algorithm for modeling Bowlers:

```
1. u=Square root (Innings Played/Matches Played)
2. v=x1*Five Wickets+x2*Wickets
3. w=x3*Bowling Average*Economy
4. Career Score=u*v/w
5. Recent score=Avg(last5scores)
6. Total Rating=x4*Recent Score+(1-x4)*Career Score
```

These algorithms are used to model the batting and bowling strengths of each teams and subsequently the total team strength.

V.TRAINING AND TESTING OF DATA

The dataset has been portioned separately into training and testing sets in Rstudio based on random selection.Initially all the values are normalized to contain the same range 0 to 1. Then the sampling of training and testing takes place. 90% of the dataset is used for training and 10% is used for testing. The training and testing samples have been made both for first innings and second innings prediction. Home/away factor is binary with either 0 or 1 value whereas the other two factors are numeric

VI. RESULTS AND DISCUSSIONS

A. Implementation of Algorithms

1) *Batsmen Model* : Batsmen modeling algorithm has been applied on the batsmen dataset to get the total batting strength feature of each batsmen before the start of the match.

Batsmen_Score = 0.6*Recent_Score +0.4*Career_Score
Recent_Score = Average of last 5 scores/Max(Recent_Score)
Career_Score = 0.7*Batting_Average/Max(Batting_Average) + 0.3*(10*Centuries + 3*Fifties)/Max(10*Centuries + 3*Fifties)

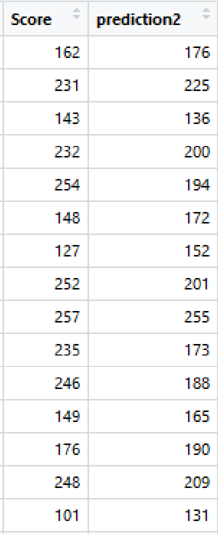
2) *Bowlers Model* : Bowlers modeling algorithm has been applied on the bowlers dataset to get the total bowling strength feature of each bowler before the start of the match.

Bowler_Score = 0.6*Recent_Score + 0.4*Career_Score
Recent_Score = Average of last 5 Wickets/Average*Economy/Max(Recent_Score)
Career_Score = (10*Fifer + Total_Wickets)/(Average*Economy)/Max(Career_Score)

3) *Team Model* : Team modeling algorithm has been applied on the batsmen model and bowlers model to model the total team strength of each team before the start of the match.

Team_Strength = Sum of Batsmen_Score of all 11 players/ Sum of Bowler_Score of all 11 players of opposition team

4) *kNN Regression* : kNN regression algorithm is applied on the final synthetic dataset with the three features. kNN regression is applied with k = 5. The algorithm is implemented successfully and predicted scores of each innings are resulted and these values are rounded to the nearest numeric number.



Score	prediction2
162	176
231	225
143	136
232	200
254	194
148	172
127	152
252	201
257	255
235	173
246	188
149	165
176	190
248	209
101	131

Fig. 1. Predicted scores compared with original scores

B. Prediction Performance

Error Rate Analysis : Error rate of each prediction is calculated as the modulus of difference of original score and predicted score by the original score.

Error rate = | (Original score - Predicted score) / Original score|

Average error rate of both innings are calculated and the error rate of 1st innings is 22% and that of 2nd innings is 18%.

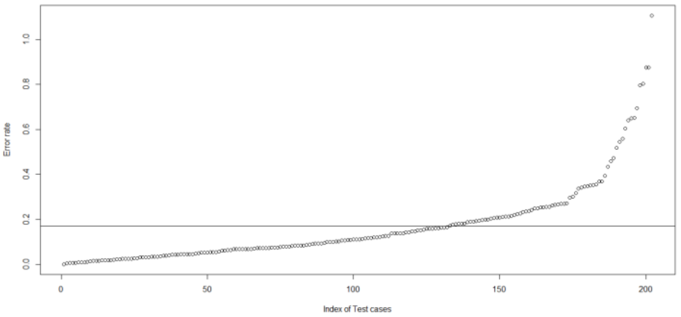


Fig. 2. Error rate plot of 1st innings

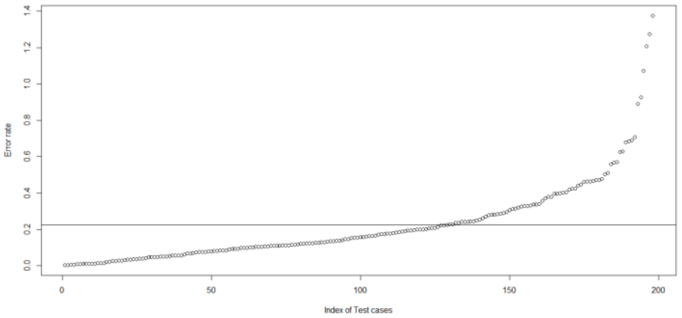


Fig. 3. Error rate plot of 2nd innings

From fig 2 and fig 3 it can be observed that the error rate of 66% of the predictions is less than the average error rate that is 22%and 64% of all predictions in 2nd innings are less than the error rate of 18%.

VII. CONCLUSION AND FUTURE SCOPE

The main purpose of this paper is to make a model for predicting the final score of the first innings and second innings for the ODIs. kNN regression algorithm is implemented to predict the scores of the match based on the factors of team strength of one team over the other, home advantage and venue average scores. The algorithm works pretty good in spite the fact that cricket is rather unpredictable. Furthermore, more features like batsmen partnerships and pitch conditions can also be used to improve the accuracy of the model further.

REFERENCES

[1] F. C. Duckworth and A. J. Lewis, “A fair method for resetting the target in interrupted one-day cricket matches,” Journal of the Operational Research Society, vol.49,no.3,pp.220–227,1998

[2] Seamus Hogan (2012) Cricket and the Wasp: Shameless self promotion (Wonkish).

[3] A.Kaluarachchi and A. S. Varde, “Cricai: A classification based tool to predict the outcome in odi cricket,” in Information and Automation for Sustainability (ICIAFs), 2010 5th International Conference on, pp.250–255, IEEE, 2010

[4] T.B.Swartz, P. S. Gill, D. Beaudoin, et al., “Optimal batting orders in one-day cricket,” Computers & operations research, vol.33, no.7, pp.1939–1950, 2006.

[5] H.H.Lemmer, “Team selection after a short cricket series,” European Journal of Sport Science, vol.13, no.2, pp.200–206, 2013.

[6] M.Bailey and S. R. Clarke, “Predicting the match outcome in one day international cricket matches, while the game is in progress,” Journal of sports science & medicine, vol.5,no.4,480,2006.

[7] https://en.wikipedia.org/wiki/Regression_analysis

[8] https://www.rdocumentation.org/packages/FNN/versions/1.1/topics/knn.reg

[9] stats.espncricinfo.com/ci/engine/records/index.html