Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 122 (2017) 525–532

Procedia Computer Science

www.elsevier.com/locate/procedia

Information Technology and Quantitative Management (ITQM2017)

Cricket Team Prediction with Hadoop: Statistical Modeling Approach

Shubham Agarwal, Lavish Yadav, Shikha Mehta *

Jaypee Institute of Information Technology, Noida, India

Abstract

Cricket is one of the most popular team games in the world. In this paper, we embark on predicting the best suitable Team to be lined for a particular match. We propose statistical modeling approach to predict the perfect players for the match to be played. As cricket is not a very simple sport, there are many factors affecting the line-up and selection of players for a particular match such as Player's Overall Stats, Player Performances with different Teams and the most important Last 5 Performances. All these factors have been considered for selection of players in playing 11 from the Team of 16. This work suggests that 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 player batting and bowling performances, forming the basis of approach used. Career statistics as well as the recent performances of a player have been used to model. Player independent factors have also been considered in order to predict the outcome of a match. Experimental analysis was performed using Hadoop and Hive for Indian players. Results establish that proposed approach is able to obtain up to 91% accuracy as compared to the real results available over WWW.

Peer-review under responsibility of the scientific committee of the 5th International Conference on Information Technology and Quantitative Management, ITQM 2017.

Keyword:Cricket team prediction, Hadoop, Hive, Sports Prediction, Statistical Modelin;

1. Introduction

Cricket is the game that set records galore across broadcast and digital platforms. It is one of the most popular sports in the world, second only to soccer. There are so many questions in the minds of viewers before the beginning of any series such that who all players will be selected for the team of a particular country? Who all players will be part of Team 11? Subsequently during the match although the sequence of players to play the game is fixed to some extent, still depending on the situation sequence is changed. So the question arises who will play next. This presents significant challenges in predicting the accurate results of a game. As a result, very limited efforts have been made by the researchers in this direction in spite of the fact that lot of data about this

* Corresponding author. Tel.: 01202594266
E-mail address: mehtshikha@gmail.com.

1877-0509
Peer-review under responsibility of the scientific committee of the 5th International Conference on Information Technology and Quantitative Management, ITQM 2017.
10.1016/j.procs.2017.11.402

game is publicly available on WWW.

This paper presents an enhanced statistical modeling approach to answer the research question: Who all players will be selected for the team of a particular country? Since cricket is the game that contains many manual decisions, we have tried to consider as many factors possible to predict the perfect selection of players for a match. In reference [1] authors considered player's overall stats and recent performances for predictions. From other references like [2] and [3] it was observed that the main factors affecting the player's performance are batting/bowling average and batting/bowling consistency. Although in previous reference stats were taken to predict the winning probability but after thoroughly analyzing the impacts of the factors affecting the Player selection for a match, algorithm have been improvised by adding many other factors affecting the selection of player in a team and formation of team for a particular match. Section 2 presents proposed methodology. Experiments and Results details are presented in section 3. Conclusion is given in section 4.

2. Proposed Methodology

This section proposes an algorithm used to model the batsmen, bowlers and the all-rounders of a team. As cricket requires many manual and on spot decisions, some assumptions are taken. Firstly, it is considered that each and every player is fit for playing. Secondly, it is independent of player's performance in trials (as there is no information available for the same).

For a player to be selected in a team, all factors about cricket must be clear such as categorizing the players into three different groups, Batsmen, Bowler and All Rounder. There are subcategories of Batsmen i.e. Top-Order Batsmen and Middle--Order Batsmen. Subcategories of Bowlers include Spin, Medium-Pace and Fast bowlers. For the Batting Performance, Batting Scores of a player are calculated. Similarly, for the Bowling Performance Bowling Scores of a player are calculated and for the All Rounder's Performance a threshold unit of Batting and Bowling Score is set. Players having scores more than threshold in both comes under the category of All Rounder.

For Batting/Bowling Score, weights have been allotted to particular columns as given below:

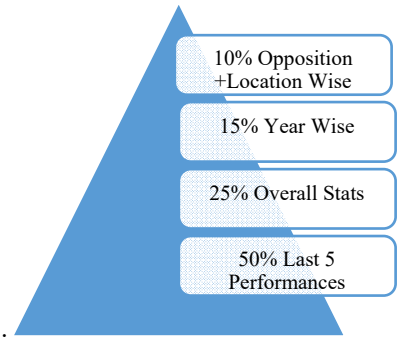

Fig.1. Pyramid of factors with their weights

Batting and Bowling Scores of a Player are calculated by the factors such as Overall stats, Year-wise stats, Opposition-wise stats, Location-wise stats and most importantly the Last 5 performances. All these factors except Last 5 performances are calculated in a similar way i.e. through Batting/Bowling Average, Number of Centuries/Fifties scored/Wickets taken and the number of matches played in that format of cricket. Mean of Last 5 scores is the remaining factor. All the factors are divided by the maximum value of that factor for scaling the value of that factor on a scale of 0-1 like Innings batted(A) is divided by total matches (Table 1), batting average(B) is divided by maximum batting average (Table 1), bowling average(H) is divided by maximum bowling average (Table 1), etc. This scaling is done to select players on their average performance and not only on the number of matches played by them.

Batting Score (Overall Stats) = (40% * Average Number of Centuries/Fifties) + (60% * Average of Batting Average)

Batting Score = (Innings Batted/Total Matches){ (50% * Average of Last 5 Performances) + (25% * Batting Score(Overall Stats)) + (15% * Batting Score(Year-wise)) + (10% * Batting Score(Opposition-wise + Location-wise)) }

Bowling Score (Overall Stats) = (40% * Average Number of Wickets taken) + (60% * Average of Bowling Average)

Bowling Score = (Innings Bowled/Total Matches){ (50% * Average of Last 5 Performances) + (25% * Bowling Score(Overall Stats)) + (15% * Bowling Score(Year-wise)) + (10% * Bowling Score(Opposition-wise + Location-wise)) }

The Batting/Bowling scores for Year-wise, Opposition-wise and Location-wise are calculated similarly as Batting/Bowling Scores of Overall Stats. Table-1 shows the specifics of the algorithm with variables.

Table 1: Calculation of batting and bowling score

Batting Score	Bowling Score
A= Innings_batted/Total_Matches B= Bat_Avg/Max_Avg C=(20 * Num_Centuries + 5*Num_Fities) / Innings_batted D= C/Max(C) E= 0.4*D + 0.6*B E->Batting_Score(Overall_stats) P->Batting_Score(Year_Wise) Q->Batting_Score(Oppostion_Wise) R->Batting_Score(Location_Wise) L= Mean of Last 5/ Max (Mean) **Batting_Score= A*((0.25*E) + (0.15*P) + (0.1* (Q+R))+(0.5*L))**	V=Innings_bowled/Total_Matches F= ((20*Num_5W)+(5*Num_3W))/Innings_bowled G= F/Max(F) H= Bowl_Avg/Max_avg I= 0.4*G + 0.6 *H I ->Bowling_Score(Overall_stats) S->Bowling _Score (Year_Wise) T->Bowling_Score(Oppostion_Wise) U->Bowling_Score(Location_Wise) L1= Mean of Last 5/ Max (Mean) **Bowling_Score=V*((0.25*I) +(0.15*S) + (0.1 * (T+U))+(0.5*L1))**

Predicting Playing 16- For the Team Prediction, this is a general formation of playing 16 (Figure 2). A range has been taken because in Cricket a good amount of decisions made are manual and depend on dynamic requirement. So the players are sorted with respect to their scores and the category they fall in.

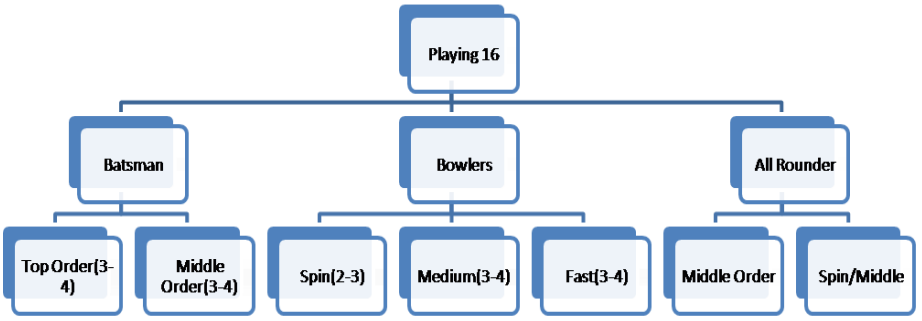

Fig. 2. Categorization of playing 16

All Rounder's are not just the ones who can bowl and bat but the ones who can be accounted for batting as well as bowling when required. For ex. RavindraJadeja is considered as All Rounder because his batting as well as bowling average is above the threshold and this player can be counted for both the roles whenever required to play any of them. All-rounders are selected in maximum numbers as they play two roles and complete the requirement of bowler as well as a batsmen simultaneously.

Generally (in almost every case) all-rounders are Middle-Order Batsmen and Spin/Medium-Pace bowlers. So they are accounted under Middle-Order category as well as under Spin/Medium-Pace category.
In all the remaining categories players are sorted with respect to their scores and the players leading their category are chosen for the playing 16.

Predicting Playing 11- It depends on two factors i.e. opposite team's line-up and type of ground on which the match is to be played. As no dynamic line-up of any team is available, it is predicted on the basis on the type of ground. Type of ground is divided into two Categories i.e. Batting with Pace and Batting with Spin. The optimal batting order is chosen [4].

Formation of Playing 11 (Batting + Pace): In this requirement of spin is covered by all-rounders like Ravichandran Ashwin and RavindraJadeja (Fig. 3).

Fig. 3. Batting + Pace

Formation of Playing 11(Batting + Spin): In this requirement of Pace is covered by all-rounders like Hardik Pandya, etc. (Figure 4).

Fig. 4. Batting + Spin

3. Big Data Frameworks: Hadoop and Hive

To perform the experiments, there were no benchmark datasets available. The datasets were crawled from the WWW. Since the data crawled is big, so to analyze data like this, Hadoop is a good option. Another important reason of choosing Hadoop was of future use as the data crawled is just the data of Indian Players, for further prediction of the teams of all countries, huge data is to be added. Hadoop is an open-source framework to store and process huge volumes of data in a distributed environment across clusters of computers. It is capable of performing complete statistical analysis for huge amounts of data. As our application process is analytical and not transactional, it is a better option. Since stats of each player involving large number of tuples have been collected and the data increases as the count of player increases, when there are many such teams taken in consideration the amount of data is huge. For such kind of big data Hive Framework over Hadoop provides an ease in computing the analysis and calculations by running algorithms in Map-Reduce Jobs which analyse the data in less time by splitting the input data-set into independent chunks which are processed by the map tasks in a completely parallel manner. Rather if these computations would have been carried out on a Normal SQL database, the computations on the data would have taken a lot more time to get the results. Moreover, Map-Reduce process makes the analysis a lot easier and faster to analyse the data, which was one of the reason to choose Hadoop.

The data crawled for the application is stored in a data warehouse as thefeatures of warehouse like Partitioning, etc helps to store the data in a good way and facilitate fast processing. Hive is a good Warehouse as its commands are similar to SQL and operations are easy to execute as well as understand. Hive is a data warehouse infrastructure tool to process structured data in Hadoop.It resides on top of Hadoop to summarize Big Data, and makes querying and analyzing easy [5]. It stores schema in a database and processed data into HDFS. It is designed for OLAP and provides SQL type language for querying called HiveQL or HQL. One of the reason of using Hive is because of the relational data as the stats are dependent on the name of the country, name of the player, etc.

4. Experimental and Results.

Due efforts were made in order to analyze the performance of proposed approach to predict the team members for a cricket match. The framework we used to make this analysis was Hadoop with Hive Data Warehouse, so to run this framework smoothly, laptops with 8GB RAM were used. Since there were no readily available datasets for presented research, data was crawled from WWW. Respective data crawlers were made for different websites to fetch the data. The data related to the parameters such as overall stats of a player, player performance stats with the opposition, player Location Wise Data and Last 5 performances were crawled from sports.ndtv.com [6]. To predict performance of player at a particular position in the team line-up, data was crawled from thatscricket.com [7]. The statistics of all countries were fetched from espn.crickinfo [8]. Snapshot of the bowling and batting data crawled is shown in Table 2 and Table 3 respectively.

Table 2: Bowling data crawled for player Yuvraj Singh

I	O	M	R	W	Best		3w	5w	Avg	E/R	S/R
Test	35	155.1	14	547	9	2/9 v PAK	0	0	60.77	3.52	103.44
ODI	160	836.2	18	4269	111	5/31 v IRE	6	1	38.45	5.1	45.2
WC	14	92.3	4	462	20	5/31 v IRE	0	1	23.1	4.99	27.75
T20	31	70.4	0	499	28	3/17 v ENG	4	0	17.82	7.06	15.14
IPL	69	140.5	0	1032	35	4/29 v DEL	2	0	29.48	7.32	24.14

Table 3: Batting data crawled for player Yuvraj Singh

M	I	N/O	R	HS		100s	50s	4s	6s	Avg	S/R	Ct
Test	40	62	6	1900	169 v PAK	3	11	260	22	33.92	57.97	31
ODI	296	271	39	8539	150 v ENG	14	51	890	153	36.8	87.64	93
WC	23	21	7	738	113 v WI	1	7	68	13	52.71	90.33	4
T20	58	51	9	1177	77* v AUS	0	8	77	74	28.02	136.38	12
IPL	108	105	13	2335	83 v JAI	0	10	174	133	25.38	130.08	27

The dataset crawled for this analysis is of Indian players. More than 600 files of data similar to the snapshot above were crawled specifying different formats like Overall stats, Last 5 Performances and many more, which converted into 1000's of tuples in multiple tables. The stats are taken till Jan, 2017. So the scores are calculated with respect to given date only. Since the prediction of the team players is dependent on that data, we have tested our predictions on the matches just after Jan, 2017.

Experiments1-This experiment was done to understand the effect of including varied set of parameters for making team predictions. We performed the experiment by making predictions only on the basis of two parameters and five parameters as shown in Figure 5. Although Factors considered by other research papers were not meant to predict team but considering only 2 factors to predict versus five factors in this research paper makes a considerable difference in predictions. It can be observed from the results that using 5 parameters, prediction accuracy improves by around 28% for India vs England and by 30% with respect to India vs Australia series. Therefore various parameters and the weights considered in the proposed algorithm are appropriate for making team predictions. Based on these results next study was performed using 5 parameters.

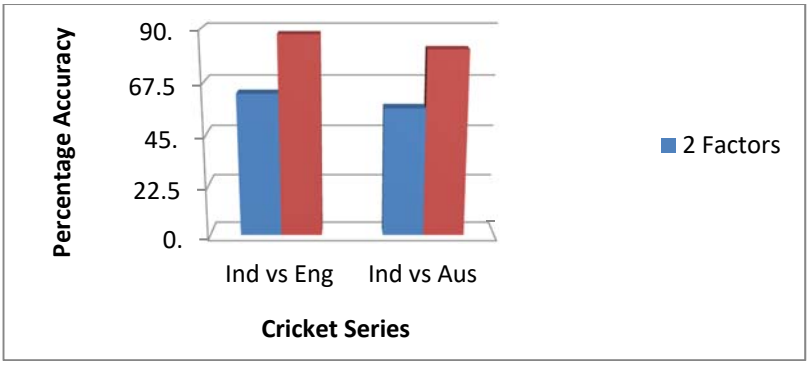

Fig. 5. Tuning of parameters for team Prediction

Experiments2 – This study was performed to evaluate the prediction accuracy of proposed approach firstly for playing 16 and subsequently for playing 11. To perform the experiments, using the datasets mentioned above, Batting/Bowling Scores were calculated through the proposed algorithm. A list of players sorted with respect to their scores and according to their category is made. Then with the toppers of the list players are fitted in proposed team formations of Playing 16 as well as Playing 11. For testing, matches played just after Jan, 2017 are taken in consideration. So England tour of India (ODI series) and Australia tour of India (Test series) were taken as test sets as actual teams' information was available. It was observed that using proposed algorithm, 14 out of 16 team members were predicted correctly(i.e 88% accuracy) for India vs England and 13 out of 16 team members were predicted correctly(i.e 81% accuracy) for Australia vs India. These results show that prediction accuracy of our approach is considerable for prediction team of 16 players. Wrong prediction of the remaining 2-3 players are because of the assumptions that were made earlier i.e. all players are fit for playing as well as no information about their performance in trials. Next study was done to assess match wise performance of our approach.

Table 4: Prediction Accuracy of Indian Players in India vs England Series

India vs England(Odi Match)	Prediction accuracy(%) of Team Players
Match-1	91
Match-2	91
Match-3	82
Match-4	73
Match-5	82

Table 5: Prediction Accuracy of Indian Players in India vs Australia Series

India vs Australia(Test Match)	Prediction accuracy(%) of Team Players
Match-1	91
Match-2	82
Match-3	73
Match-4	82

Table 4 depicts the prediction accuracy of Indian players in India-England ODI for each match and Table 5 depicts the prediction performance of our algorithm for India -Australia for all matches played in the test match. It can be observed from the results that minimum accuracy of the proposed algorithm is above 70% and it reaches up to 91% that means 10 players out of 11 are predicted correctly using our algorithm. The reason for remaining 1-2 players wrongly predicted is because of the situation based decisions made by the team**.** These results establish that proposed algorithm has good potential for making team predictions and may be considered by the team selection committees to automate the initial process of selection and apply expert opinion thereafter. In the end, this analysis is concluded with the theory that the proposed algorithm is a good measure to compare players to choose between them who will prove better for the team. Players leading the chart should be chosen for the formation of the team. And also it is a reliable measure to form a team for an upcoming series or tournaments.

5. Conclusion and Future Work.

This paper presented statistical modeling approach for predicting team members to participate in a particular match/series. Analysis of results establish that factors considered for team prediction are worthy as they provide up to 91% accuracy. In future, this work can be modified on the assumptions made in the starting that every player is fit and their training data is not available, so using training data will increase the accuracy of the prediction. More work can be done in the area to analyze the concept of using substitutes in the middle or last matches of a series. In advance to this real time predictions on any situations can be made with the help of these calculated scores. Making strategies of order of the batting innings or the bowling order can be sorted with these scores.

References

[1] M.G.Jhawar, VikramP, Predicting the Outcome of ODI Cricket Matches: A Team Composition Based Approach, (ECML-PKDD 2016), pp. 1-10.
[2] Barr, G.D.I., Kantor, B.S. A criterion for comparing and selecting batsmen in limited overs cricket, Journal of the Operational Research Society, 2004, 55: 1266-1274.
[3] Beaudoin, David, Tim B. Swartz. The best batsmen and bowlers in one-day cricket, South African Statistical Journal 37.2 (2003): 203.
[4] Norman, John M., and Stephen R. Clarke, Optimal batting orders in cricket, Journal of the Operational Research Society 61.6 (2010): 980-986.
[5] https://www.dezyre.com/hadoop-tutorial/hive-tutorial
[6] https://sports.ndtv.com/cricket/players/6-india
[7] http://www.thatscricket.com/india/players/
[8] http://www.espncricinfo.com/india/content/player/country.html?country=6