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:

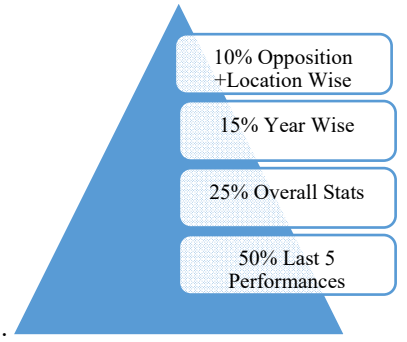

Fig.1. Pyramid of factors with their weights

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.