Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 122 (2017) 910–915

Procedia Computer Science

www.elsevier.com/locate/procedia

Information Technology and Quantitative Management (ITQM2017)

IPL Visualization and Prediction Using HBase

Shubhra Singh*, Parmeet Kaur

Department of Computer Science & IT Jaypee Institute of Information Technology *Noida, INDIA*

Abstract

With the increasing number of matches day by day, it has become difficult to manage or extract useful information from the available data of all the matches. The paper presents a data visualization and prediction tool in which an open-source, distributed, and non-relational database, HBase is utilized to keep the data related to IPL (Indian Premier League) cricket matches and players. This data is then used for visualizing the past performance of players' performance. Additionally, the data is used to predict the outcome of a match through various machine learning approaches. The proposed tool can prove beneficial for the team managements in the player auctions for selecting the right team.

Keywords: match prediction; HBase; machine learning; K-Nearest Neighbors

1. Introduction

The game of cricket is played in various formats, i.e., One Day International, T20 and Test Matches. The Indian Premier League (IPL) is a Twenty-20 cricket tournament league established with the objective of promoting cricket in India and thereby nurturing young and talented players. The league is an annual event where teams representing different Indian cities compete against each other. It was started by the Board of Control for Cricket in India (BCCI) and has now become a giant, remunerative cricket venture. The teams for IPL are selected by means of an auction. Players' auctions are not a new phenomenon in the sports world. However, in India, selection of a team from a pool of available players by means of auctioning of players was done in Indian Premier League (IPL) for the first time.

Due to the involvement of money, team spirit, city loyalty and a massive fan following, the outcome of matches is very important for all stake holders. This, in turn, is dependent on the complex rules governing the game, luck of the team (Toss),the ability of players and their performances on a given day. Various other natural parameters, such as the historical data related to players, play an integral role in predicting the outcome of a cricket match. A way of predicting the outcome of matches between various teams can aid in the team

* Corresponding author.
E-mail address: shubhrasingh1@outlook.com.

10.1016/j.procs.2017.11.454

selection process. However, the varied parameters involved present significant challenges in predicting accurate results of a game. Moreover; the accuracy of a prediction depends on the size of data used for the same.

The tool presented in this paper can be used to evaluate the performance of players. This tool provides a visualisation of players' performances. Using IPL T-20 variables related to statistics of batsmen and bowlers, a number of apt variables have been identified that have elucidative power over auction values. Further, several predictive models are also built for predicting the result of a match, based on each player's past performance as well as some match related data.The developed models can help decision makers during the IPL matches to evaluate the strength of a team against another.

The tool employs HBase, a distributed, open source and non-relational database for storing the data. HBase is increasingly being used for hosting of tables with billions of rows and millions of columns. It allows automatic and configurable sharding of tables for scalability of applications,

The contributions of the presented work are as follows:

- To provide the statistical analysis of players based on different characteristics
- To predict the performance of a team depending on individual player statistics
- To successfully predict the outcome of IPL matches

The paper is structured as follows: Section 2 presents the related work in this field. Section 3 presents the proposed methodology. The obtained results are discussed in Section 4. We conclude this presentation in Section 5.

2. Related Work

Sports analytics has been successfully applied in sports. In soccer, it is common to rely on ratings by experts to assess a player’s performance. However, the experts do not unravel the criteria they use for their rating. The work in [2] attempts to identify the most important attributes of a player’s performance that determine expert ratings. In this work, a series of classifications with three different pruning strategies and an array of machine learning based algorithms are executed. Then, a list of most important performance metrics for each of the four playing positions which approximates the attributes considered by the experts are obtained while assigning ratings.

The authors in [4] analyzed many data mining techniques and the obtained prediction results are compared to arrive at a suitable model for prediction of results for matches played by the Dutch soccer team. These results are built using Naïve Bayes model, a random tree model and a k-nearest neighbor model. Out of these, the most suitable model is selected and its results are analyzed further. In this manner, the variables with the least and most predictive power can be identified from the best prediction model.

To predict the outcome of ODI cricket matches, an approach where estimation of the batting and bowling potentials of the 22 players playing the match using their career statistics and active participation in recent games is done in [3]. The player and team strengths are used to render the relative dominance of one team over the other. Taking two other base features into account, along with the relative team strength, supervised learning algorithms are adopted to predict the winner of the match.

The machine learning based approach used in [5] is reached at by an in-depth analysis of T20 cricket features. In order to indicate the players’ performance, a novel index, namely Deep Performance Index (DPI) is derived using the characteristics specific to T20 cricket. The authors extract relevant features using the machine learning algorithm of Recursive Feature elimination for designing the DPI. It is demonstrated that DPI achieves better results in analysis of performance related data for batsmen as well as bowlers in comparison to some other ranking methods for T20 cricket. There exist some other approaches [6,7] which have specifically worked upon IPL data.

The tool presented in this paper utilizes machine learning algorithms for analysis of players' performance in IPL as well as to predict the outcome of matches. It utilizes HBase for storing the required data in order to allow un-structured or semi-structured data as well as features for scalability.

3. Proposed Tool

Around 644 matches have taken place from 2008 to 2017 in IPL, making it one of the biggest leagues in history of T20 cricket with viewership of millions of people. To store this large amount of data which may include ball by ball details, a database for big data, i.e., HBase is used in this work, due to its features as follows:

- Scalability in both linear and modular form
- Automatic as well as configurable sharding of tables
- Provision for distributed storage in the form of HDFS (Hadoop Distributed File System)
- Consistent read and write operations
- Automatic failover support
- Support for Java APIs so that clients can access it easily
- Support for MapReduce for parallel processing of large volume of data
- Back up of Hadoop MapReduce jobs in HBase tables
- Block cache and Bloom filters for easy real time query processing

The methodology followed by the presented tool is described next:

3.1. Data Collection

Data related to players who were part of IPL 2017 was taken for analysis. Collection of data was performed using the *BeautifulSoup*[8]library of the Python programming language from the website www.cricbuzz.com.We used Pandas, a software library to transform the data in the required form. Pandas [9] is an open-source library for manipulation and analysis of data for the programming language, Python. Out of approximately 40 columns scraped from www.cricbuzz.com, we required only 10 (IPL statistics of players only). To obtain these columns, Pandas library was utilized to perform operations on scraped CSV file and to make required CSV file.

3.2. Parsing

The next step is parsing the relevant data from the web pages. This has been done using the *Beautiful Soup[8]*package which creates a parse tree for a page and this parse tree can be used to extract data from the HTML page.

3.3. Data Visualization

The collected data is visualized to get a better understanding of the information regarding the best players for the next auction. The top 10 batsmen, bowlers and all-rounders are plotted against the number of matches played to get the names of players who played best in all the IPLs till date.

The Graphical and Statistical analysis is done with the help of *Tableau* software [10]. This software is used for business intelligence applications as well as for statistical analysis. It has the provision to connect to any database and also for uploading files for performing statistical analysis on them.

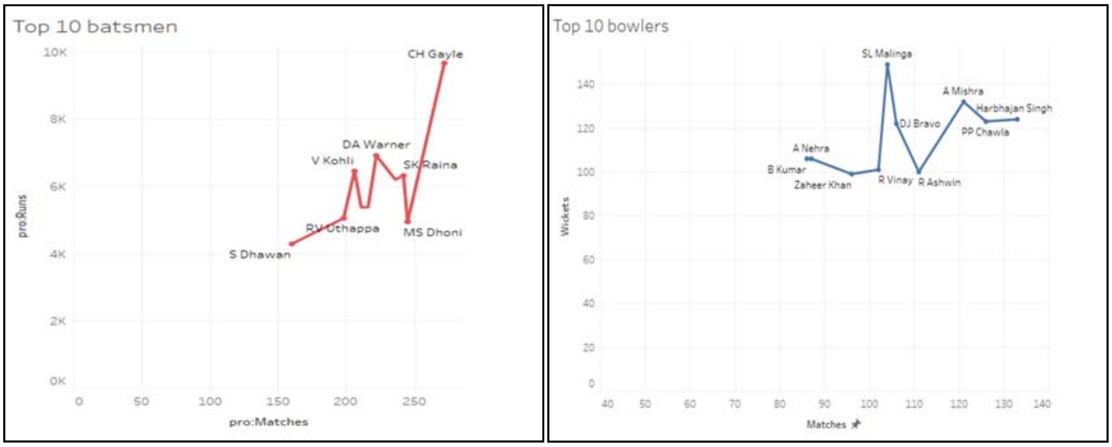

Fig. 1. (a) Top Batsmen (b) Top Bowlers

In Fig1a, the top 10 batsman of all the IPL seasons are given and their runs are compared with the number of matches they have played. Fig 1b depicts the top 10 bowlers of all the IPL seasons and their wickets are compared with the number of matches played.

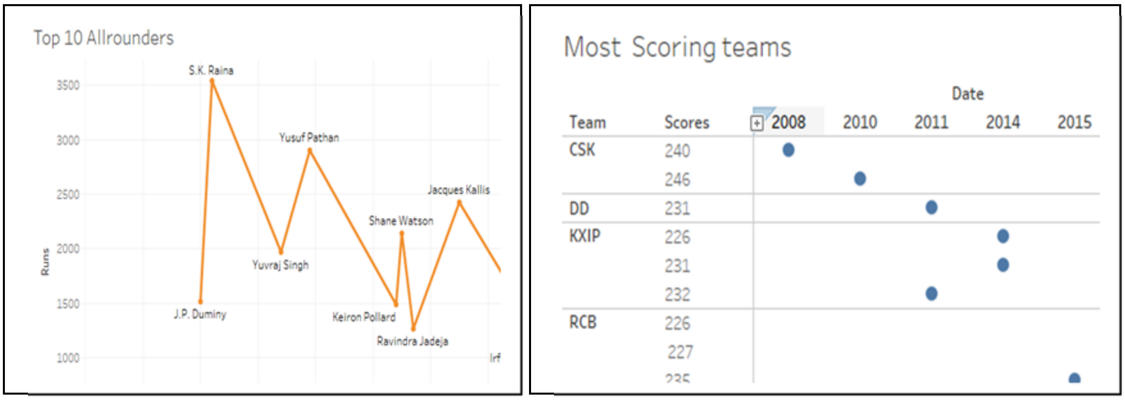

Fig. 2 (a) Top All-rounders (b) Most Scoring Teams

In Fig 2a, the top 10 all-rounders of all the IPL seasons are shown along with their performance in the form of runs scored and number of wickets taken in number of matches they have played. Fig 2b shows the top 10 teams which scored the highest scores in all the seasons.

3.4. Prediction

KNN(k-Nearest Neighbors) algorithm with k=4 has been used for predicting the winner of a match between 2 IPL teams. Further, KNN has been compared with other machine learning algorithms; namely, decision tree, logistic regression, random forest and Support Vector Machine [11] for accuracy. The training examples for KNN are vectors in a multidimensional feature space where there exists a class label for every vector. In the training phase of the algorithm, we store the feature vectors and class labels of the training samples. In the classification phase, k has been taken as 4, and subsequently an unlabeled vector or query point, corresponding to the game whose result has to be predicted, is classified by assigning the most frequent label from the k training samples nearest to that query point. The distance metric used, in general, for continuous variables is the Euclidean distance. Other metrics, such as the overlap metric or Hamming distance can be used for discrete variables, such as text classification. In this work, KNN is used with k=4 i.e. (clusters forming according to 4 of their neighbors). Around 10 fields of IPL statistics of different players are used and prediction of winner of the match is done with the help of 2 other types of criteria i.e. toss and venue of the match.

4. Results

Table 1. Prediction of Winner

S.No.	Team1	Team2	Venue	Toss	Strength	Winner
1.	RCB	MI	1	1	1.10	MI
2.	RCB	MI	0	1	1.10	RCB
3.	RCB	MI	1	0	1.10	RCB
4.	RCB	MI	0	0	1.10	MI
5.	KKR	SRH	0	0	1.11	SRH
6.	KKR	SRH	1	1	1.11	KKR
7.	KKR	SRH	1	0	1.11	KKR
8.	KKR	SRH	0	1	1.11	KKR
9.	RCB	SRH	0	1	1.89	RCB
10.	RCB	SRH	1	1	1.89	SRH

Table 1 demonstrates the results of the different inputs given to the KNN algorithm. For each case involving two teams and their compositions, we have varied the values for venue (home ground) and toss. A value of 1 denotes the venue or toss is in favour of Team 1 while a 0 denotes it is in Team 2's favour. The columns used for calculation of strength were IPL statistics of matches played till date, such as number of matches played, batting and bowling average, runs scored, wickets taken, bowling economy etc. The data was taken for years 2008-2016. The result is calculated by using 50% data as training set and other 50% as testing set to get a handle on variance as variance is higher on smaller number of test samples.

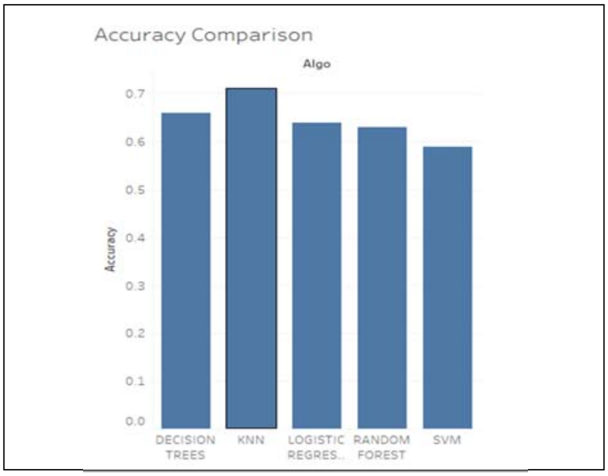

Fig 3 Comparison of Accuracy of Different Algorithms

Fig. 3 shows the accuracy graph of different algorithms used for predicting the winners of matches. It can be observed that KNN has the highest accuracy among all the algorithms used. KNN with k=4 has an accuracy of about 0.71 i.e. 71%. Supervised KNN with k=4 is used which gives accuracy up to 71% as it checks labels with the 4 nearest neighbors' classes of itself which enables it to fit the data most accurately and prevents both over-fitting and under-fitting.

5. Conclusion

The paper addresses the problem of predicting the outcome of an IPL cricket match and also the player profiling system which can be a great help for the team leaders on the auction day. The statistics of 644 matches have been used in the experiments. Factors such as luck and player strength have been used as key features in predicting the winner of a match. The novelty of the proposed approach lies in addressing the problem as a dynamic one and using a suitable non-relational database, HBase for scalability of application. Out of all the machine learning algorithms used, KNN has been observed to be the most accurate.

References

[1] Parker, D., Burns, P., & Natarajan, H. Player valuations in the Indian Premier League, Frontier Economics, 2008,1-17

[2] Gunjan Kumar. Machine Learning for Soccer Analytics. 2013

[3] Madan Gopal Jhawar & Vikram Pudi. Predicting the Outcome of ODI Cricket Matches: A Team Composition Based Approach, European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in

[4] Databases, 2016

[5] Abel Hijmans. Dutch football prediction using machine learning classifiers

[6] C. Deep Prakash, C. Patvardhan & Sushobhit Singh. A new Machine Learning based Deep Performance Index for Ranking IPL T20 Cricketers, International Journal of Computer Applications 137(10):42-49, March 2016

[7] P. Kalgotra, R.Sharda, G.Chakraborty, Predictive Modelling in Sports Leagues:An application in Indian Premier League, SAS Global Forum, 2013

[8] P.K. Dey, D. N. Ghosh, A.C. Mondal. Multi-Criteria Decision Tree Approach to Classify All-Rounder in Indian Premier League, Journal of Emerging Trends in Computing and Information Sciences, 2011. 2 (11), 563-73

[9] https://www.crummy.com/software/BeautifulSoup/#Download

[10] http://pandas.pydata.org/

[11] https://www.tableau.com/

[12] Christopher Bishop. Pattern Recognition and Machine Learning. 2e.