Predicting and Analyzing the Performance of the IPL Cricket Using Regression Models

KASUKURTI RAVITEJA, GANESH KUMAR MACHA, Dr. GR ANANTHARAMAN

1 M.Tech Student, Department of CSE, Malla Reddy Institute of Engineering &Technology, Telangana, India.

2 Assistant Professor, Department of CSE, Malla Reddy Institute of Engineering &Technology, Telangana, India.

3 Professor, Department of CSE, Malla Reddy Institute of Engineering &Technology, Telangana, 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. 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. 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. We are using SVM, Random forest, Decision tree classifiers and to generate the prediction models for the problem. The prediction of the matches based on the toss wins and predicting the outcomes who will be winning the matches, addresses the problem of predicting the outcome of an IPL cricket match.
1. INTRODUCTION

Cricket is an extremely famous sport in our country, and it has a huge fan-base, irrespective of the type of format it is played in. As fans, the people make their own predictions while watching a particular match based on the data given they have and then, they make a call on who will win the match. The data includes different statistics and records like number of runs scored by a batsman, the overall strike-rate of a batsman, the number of wickets taken by a bowler, the number of catches taken by a fielder and so on. This research is based on predicting the winner of a cricket match based on the data available from the match. This research aims at predicting the result of an ongoing cricket match in an over-by-over basis based on the information and data that is available from each over. The data that is available in each over includes the number of runs scored in the over, the number of fours and sixes hit in the over, the number of wickets taken in the over, the number of wickets taken so far in the match and the number of extras in the over. We will be performing prediction for all the matches that have taken place in the IPL. This is done by using machine learning algorithms for performing the prediction of the results of the matches.

Data mining in today's world is very important as it helps in extracting important pattern from a given data set. Association, classification, clustering, and outlier deduction are various techniques for effective data mining. Association rule mining is applied to find associations in a volumetric data. Earlier this technique was applied to discover consumer purchasing behaviors. However, today it is used in several other fields such as marketing (Xu, Frankwick, & Ramirez [1]), finance (Olson, & Wu [2]), telecommunication (Le [3]), and management (Choi, Chan, & Yue, 2017). Association rule mining is also used for sports management (Zhao & Chen [4]). Ofoghi, Zeleznikow, MacMahon & Raab [5] developed data mining tools to architecturally link the game and data mining areas. Asanka [6] identified the factors that influence the result of the extra deliveries. Bhattacherjee, Sahoo & Goswami [7] employed association rule mining method in cricket to find the fundamental unidentified relations of the factors impacting the performance of a player. Singh, Singla & Bhatia [8] proposed a model using Linear Regression Classifier and Naïve Bayes Classifier. First innings score was predicted using linear regression classifier on the basis of current run rate, number of wickets fallen and location of the game. The result of the game in the second innings was predicted using naïve bayes classifier.
Verma & Izadi [9] proposed a Cricket Prognostic System, which is an analytical research framework for cricket on the basis of ball-by-ball data and historic data using machine learning and statistical methods. Zhao & Chen [10] developed a sport model by combining multiple small models with different data mining techniques for improving the accuracy of sport model result. Nair [11] developed a model for predicting target score. Satao, Tripathi, Vankar, Vaje, & Varekar [12] developed a prediction system using the K-means clustering algorithm. The system used historical match data, performance of the player and spectator to score prediction and match result. The general framework of Association rule mining assumes that all items have the same significance which is not always the case. Some products may be more profitable, as compared to others. Hence rules regarding them are of greater value. In order to handle such situations weights are assigned to the variable based on their significance. The paper presents an algorithm for weighted association rule mining problem and demonstrates its use in cricket management.

2. RELATED WORK

The research works related to this problem in several sports are explained below. Ahilandeswari et al. [1] explains more about live data extraction from www.twitter.com. They used the concept of Hadoop Map-Reduce to extract the required attributes in an attempt to predict the winner of the cricket match. There search work by Anjali et al. [2] explains the extraction of different types of cricket data and information using the concept of Big Data. The Hadoop Distribution File System (HDFS) is used in this process. The data will be available in the form of Tweets from www.twitter.com, and using HDFS, the required data will be extracted from the tweets. Haghighat [3] described the various prediction techniques including Nave Bayes which can be used to predict the best playing eleven for the team. While the research work was aimed towards basketball, it was not general towards it alone, and can be extended to any sport. This research work used basics of Python and the Scikit-learn API to implement the project. The different types of machine learning algorithms used in this research work included Support Vector Machine and Gaussian Naive Bayes. The research work by Nazim Razali et al. [4] explains predicting the winner of a Football match in the English Premier League, which is a famous football tournament. This implemented by using Gaussian Naive Bayes, a machine learning algorithm. We have used similar to this in predicting the winner of the cricket match. This research work uses different models required to predict the winner of the match at

http://cij.org.in/CurrentissuesDownload.aspx
any phase. The research work by Rory P. Bunker and Fadi Thabtah [6] explains generating models and using machine learning algorithms such as Artificial Neural Networks to determine which team will win the match. This research work was generalized to all sports, and so can be utilized for different sports including cricket. The research work by Xiangyan Zeng [8] described the method to implement feature selection using Recursive Feature Elimination. The reason Recursive Feature Elimination was used is because it selects features by recursively considering smaller and smaller sets of features. The same feature selection algorithm was used to eliminate the features with lesser weight in the project. Since the algorithm works by checking each subset of the given data, RFE is one of the most preferred feature selection algorithms available. Prakash et al. [5] proposed three variations of predictive models using Support Vector Machines to predict the winners of IPL matches. Singh and Kaur [7] used K Nearest Neighbour to predict the winners of IPL matches. However, these approaches have not focused on prediction at any point of time during the match. They predict the winners before the match starts based on the history.

3. METHODS

The structure of the architecture diagram is shown in Figure 1. It explains the basic data flow of the entire methodology and explains each of the individual components of our approach. The different components are data collection, data extraction, feature selection, model generation, model selection and prediction.

over 500 matches having taken place. The required dataset needed for this work has been collected from the internet web source. From this obtained dataset of the IPL matches, only the attributes that are essential to this work needed to be extracted, and so an algorithm to obtain these features alone has been used. The
data is then extracted using a Java program which converts the ball-by-ball format into an 
over-by-over data format. The data is extracted 
from the website, www.cricsheet.org which 
consists of ball-by-ball information about the 
IPL matches. The attributes include the two 
teams playing the match, the venue of the 
match, which team won the toss, which team is 
batting first, the umpire names, the name of 
the cricket stadium, the number of runs scored 
by the striker and the non-striker, the wickets 
taken in the over and in the match so far, the 
number of extras in an over, the number of 
boundaries hit in the over and the classifier, 
which team won the match. The attributes that 
are essential to the project are divided into two 
categories, the common attributes and the 
Over-by-Over attributes. The Common 
attributes are the ones that are common for 
the entire project and they are the Opposition 
Team, Venue of the Match, Toss and which 
team won the match. The Over-by-over 
attributes are the ones that repeat for all the 
twenty overs of an innings. These include the 
runs scored, the fours and sixes hit in the over, 
the wickets taken in the over, the wickets taken 
so far in the match and the extras in the over. 
To extract only these attributes, Data Filtering 
on the entire dataset was performed, and the 
acquired data is then finally stored in a CSV 
file.

B. Feature Selection Feature Selection 
becomes important when number of features 
are very high. It refers to selecting the most 
important features that results in accurate 
results than the one with all the features. One 
of the important feature selection methods is 
Recursive Feature Elimination (RFE). RFE 
works by recursively building models based on 
the feature subsets and after the model is built, 
it removes the feature with very low priority 
and again build a model using the remaining 
feature subsets. The process is continued till 
all the features are exhausted. First, the entire 
set of features are taken as input. Then the 
number of attributes to be extracted has to be 
specified. The data is then trained using a 
logical estimator, using the LogisticRegression 
inbuilt function. If the total features are 
greater than the given input, then the 
coefficient of all the features given as input is 
calculated. The values are then squared to 
ensure they are all positive values, and then 
they are sorted and stored in a rank list. The 
feature at the zeroth index is removed, and 
then this process is continued until the 
required number of attributes are obtained.

C. Model Generation Based on the data 
obtained, various models required for this 
research work was generated. The models are 
generated based
on different phases in the match. The different phases in the match are 2-Overs, 5-Overs, 8-Overs, 12-Overs, 16-Overs and 20-Overs. These models are used to predict which team will be the winner of the ongoing match. The dataset is loaded first from the corresponding CSV file, and for each phase in the match, a model is created using a different machine learning algorithm, such as Naive Bayes, SVM, KNN and Random Forest. Then the best model is selected using Cross Validation, and this model is then used to predict the winning team. D. Prediction Once the predictive model is created, the test data is given as input to it and the output is predicted. This is then compared with the ground truth defined in the dataset obtained from the website. The dataset is loaded from the corresponding CSV file. A part of the dataset is given as input without feature selection and a part of it is given as input with feature selection to the generated models. Based on this, different models will be generated based on the different phases of the match, which are 2-Overs, 5-Overs, 8-Overs, 12-Overs, 16-Overs and 20-Overs. Based on this, a prediction on which team wins the ongoing match at that particular phase of the match is obtained.

4. RESULTS

The results are obtained for the test set and the accuracy was measured for the particular machine learning model used to predict the winner of the match. Results are predicted based on the different teams of the IPL, and is predicted in an over-by-over analysis so that the winner of the match can be predicted in almost any situation of the match. The accuracy for a selected number of attributes for each team using feature selection was also measured. Figure 2 displays the final results without feature selection. For each generated model along the X-axis and accuracy along the Y-axis, the highest accuracy of each of the teams in the IPL was depicted. Thus, the graph displays which team has the highest accuracy in each generated model.

5. CONCLUSION

Our work will be extended further by increasing our dataset so that it not only includes the games conducted in the IPL, but also matches from other famous Cricketing events such as the Big Bash League as well as matches from International Cricket. The other
feature selection techniques may be implemented in future and explored if the accuracy increases when different algorithms are used. The hyper parameters for the classifiers may be fine tuned in future to get a better performance.

6. REFERENCES


