A Perspective on Analyzing IPL Match Results using Machine Learning

Gagana S[1] K Paramesha[2]
[1]M.Tech Student [2]Associate Professor
[1,2]Department of Computer Science and Engineering
[1,2]VVCE, Mysuru, India

Abstract— Indian Premier League (IPL) is a T20 league which was started in the year 2008 and it is the most belaud T20 cricket carnival in the world. Since IPL has huge popularity, it is needful to inspect the possible predictors that affect the overall result of the matches. The expected solution depends on time series analysis and Machine learning techniques which minimizes the use of domain knowledge. This paper attempts to predict the runs for every ball by using the runs scored by the batsman previously as the observed data. Data from all past IPL matches is collected for the analysis and the problem is modeled as a classification problem. In this work, we build predictive models for predicting the runs based on the observed data. We have used Recurrent Neural Network (RNN) and Hidden Markov Model (HMM) for generating the models for the problem. Finally, the model created using RNN and HMM provided the highest prediction accuracy.

Keywords: IPL, Machine Learning, HMM, RNN, Runs Prediction

I. Introduction

Sports have gained much importance in both national and international level. Cricket is one such game, which is marked as the prominent sports in the world. T20 is one among the forms of cricket which is recognized by the International Cricket Council (ICC). Because of the short duration of time and the excitement generated, T20 has become a huge success. The T20 format gave a productive platform to the IPL, which is now pointed as the biggest revolution in the field of cricket.

IPL is an annual tournament usually played in the month of April and May. Each team in IPL represents a state or a part of nation in India. IPL has taken the T20 cricket's popularity to sparkling heights [1]. It is the most attended cricket league in the world and in the year 2010, IPL became the first sporting event to be broadcasted live. Till date, IPL has successfully completed 11 seasons from the year of its inauguration [2]. Currently, there are 8 teams that compete with each other, organized in a round robin fashion during the stages of the league. After the completion of league stages, the top 4 teams in the points table are eligible to the playoffs. In playoffs, the winner between 1st and 2nd team qualifies for the final and the loser gets another opportunity to qualify for the finals by playing against the winner between 3rd and 4th team. In the end, the 2 qualified teams play against each other for the IPL title [3]. The significance is that IPL employs television timeouts and therefore there is no time constraint in which teams as to complete the innings.

This game is exceedingly unpredictable because at each phase of the game, the momentum changes to one of the teams between the two. Many times the results will be decided in the last ball of the match where the game gets really closer. Considering all these aspects, there is immense interest among the viewer to make predictions either at the beginning of the match or during the match [11]. IPL games can't be easily predicted only by making use of statistics and teams past match's data. Forecasting from the previous data is highly personalized and requires remarkable expert decisions.

In this paper, we have examined various elements that may affect the outcome of IPL match in determining the runs for each ball by considering the runs scored by the batsman in the previous ball as the labeled data. The suggested prediction model makes use of RNN and HMM to fulfill the objective of the problem stated. Few works have been carried out in this field of predicting the outcomes in IPL. In our survey, we found that the work carried out so far is based on Data Mining for analyzing and predicting the outcomes of the match. Our work novelty is to predict runs for each ball by keeping the runs scored by the batsman in the previous ball as the observed data and to verify whether our prediction fits into the desired model.

II. Literature Survey

Many exploration works have been accomplished to analyze and predict the outcome of the matches using various techniques of Machine Learning and Data Mining. We have examined some of those techniques and algorithms with summarized models accuracy.

Pranavan Somaskandhan et al [1] aim to identify the optimal set of attributes, which impose high impact on the results of a cricket match. The proposed solution relies on statistical analysis and Machine Learning. Different Machine Learning algorithms were employed and Support Vector Machine (SVM) achieved the best accuracy in the evaluation.

Prince Kansal et al [2] as built several prediction models for predicting the selection of a player in IPL based on each players past performance. Various Data Mining algorithms are applied namely Decision Tree, Naïve Bayes and Multilayer Perceptron (MLP) on the dataset to fulfill the objective. MLP gave the best accuracy among all other algorithms.

Rabindra Lamsal et al [3] as proposed a linear regression based solution to calculate the weight age of a team based on the past performance of its players who have appeared most for the team using 2 Machine Learning algorithms: Multiple Regression and Random Forest and the classification results are satisfactory.

A N Wickramsinghe et al [4] created a model to predict cricket match outcome using Machine Learning algorithms such as SVM, Logistic Regression, Naïve Bayes and Random Forest. Final results indicated that twitter based model is better than natural parameter based model.

Tejinder Singh et al [5] created a model that predicts the score of 1st inning and the outcome of the match in the 2nd inning. Implementation is done using Linear Regression and Naïve Bayes. It was found that the accuracy of Naïve Bayes in predicting the match outcome is more.

Kalpdrum Passi et al [6] attempted to predict the performance of players. They have used Naïve Bayes, Random Forest, Multiclass SVM and Decision Tree classifiers to generate the prediction models for the problem. Random Forest classifier was found to be most accurate.

Shimona S et al [9] aim at analyzing the IPL cricket match results from the dataset collected by applying existing Data Mining algorithm to both balanced and imbalanced dataset. The model was built successfully with accuracy rate of 97% for the balanced dataset and error rate was found to be more in imbalanced dataset when compared to that of the balanced dataset.

GuYu et al [10] presented a generic, scalable and multi layer framework based on HMM for sports game detection.

Mayank Khandelwal et al [11] aimed at choosing the players for the match by calculating Most Valuable Player (MVP) using Decision Tree, Bipartite Cover and Genetic Algorithm. This will give more option for the captain to utilize his batting and bowling strength during the match.

III. Models for Predicting Runs using RNN and HMM

A. Proposed Solution

The main objective of this work is to predict the runs using labeled data. A large amount of data has to be analyzed in order to get a reliable accuracy and hence the first step in achieving the purpose is to gather data of all IPL matches. Data that provides detailed information about each IPL match was collected. Then the mandatory data was extracted and refined. The attributes are the gathering of IPL match details which uses less amount of domain knowledge. We decided to model the problem as a supervised learning problem since labels can be attached to match details. The labels are discrete and hence this work can be modeled as a classification problem.

B. Data Collection and Data Preprocessing

We collected data from dataworld.com which contains details about 577 matches with 21 attributes. The dataset has been divided into training and testing datasets. The logic behind partitioning the dataset is to give a clear outline of each innings. In this analysis, match details with respect to the 1st innings of all the matches is considered as the training dataset and match details regarding 2nd innings of all matches is considered to be the testing dataset.

Match id, inning, batting team, bowling team, over, ball, batsman, bowler, runs etc are the attributes in the dataset. Current Run Rate (CRR), Required Run Rate (RRR), Wickets taken, Batsman Strike Rate, Bowler Average is the additional attributes that are computed in order to achieve the goal. Runs are considered to be the independent feature and the remaining attributes are taken as the dependent features in the dataset. We have taken runs as the labeled data for predicting the outcome of the match.

C. Calculating Attributes

Run Rate: It is the number of runs that a team scores in one over (6 balls). It is obtained by dividing the total number of runs scored at some point of time by the number of over's played [6].

Run Rate = Total number of runs scored/Number of over's bowled

1) Required Run Rate:

It is the number of runs per over the batting side must score in order to win the current match [6].

Required Run Rate = Total runs required to win/Total over's left

2) Batsman Strike Rate:

It is the average number of runs scored per 100 balls faced [6].

Batsman Strike Rate = (Runs scored/Total balls faced)*100

3) Bowler Average:

It is the number of runs conceded by a bowler per wicket taken [6].

Bowler Average = Runs conceded/Wickets taken

Attributes	Description
Match id	Number assigned to each match
Inning	Division of a match
Batting team	The team which is currently batting
Bowling team	The team which is currently bowling
Over	The number of over's bowled at a particular stage of the batting team
Batsman	Person who is batting
Bowler	Person who is bowling
Total runs	Final score of the team
Current Run Rate	Number of runs that a team scores in one over
Required Run Rate	Number of runs per over the batting side must score in order to win the current match
Batsman strike rate	Average number of runs scored per 100 balls faced
Bowler Average	Number of runs conceded by a bowler per wicket taken

Table 1: Description of the attributes

Table 1 gives the description of the attributes present in the IPL dataset. All the attributes are common in both the training and testing dataset. Attribute values may vary according to the match situation.

Subsequently, using an open source python library we applied several Machine Learning algorithms.

D. Models for Prediction

1) Recurrent Neural Network (RNN):

It is a class of Artificial Neural Network (ANN) where connections between nodes form a directed graph along a sequence. It uses internal state or memory to process sequence of input.

Back propagation algorithm is used to train RNN and RNN remembers its input and involves sequential data and they are the only ones with an internal memory. It produces predictive results in sequential data that other algorithms can't. RNN can map 1 input to many outputs, many to many and many to 1. Commonly used type of RNN is Long Short Term Memory (LSTM) which is much better at capturing long term dependencies [7].

2) Hidden Markov Model (HMM):

It is a statistical model for modeling time series data [8]. They are used to model temporal inference, which was built in topological terms. It is also used in recognizing remote events [10].

In the end, the best set of features is identified using feature selection which adds a remarkable impact on the end results.

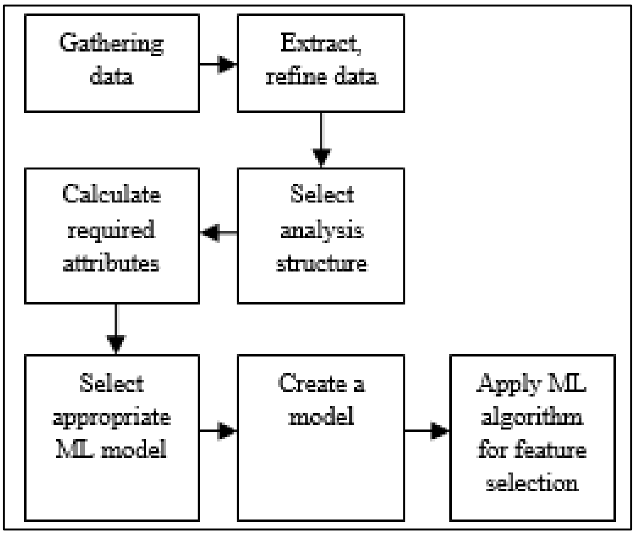
Gathering data
Extract, refine data
Select analysis structure
Calculate required attributes
Select appropriate ML model
Create a model
Apply ML algorithm for feature selection

Fig. 1: Workflow Diagram

Figure 1 represents the flow of the study. As shown in the above diagram, the work started with gathering data, refining the data collected. Then the problem is analyzed and proper Machine Learning structure is created. Necessary attributes like CRR, RRR, Batsman Strike Rate, Bowler Average is calculated and it is treated as a classification problem and it falls under time series analysis since the momentum of the games changes frequently. A model is created with runs as label and all attributes as features. Finally, Machine Learning is applied along with feature selection technique which significantly contributes to end results.

IV. Comparative Analysis

Several works has been carried out to build prediction models using Data Mining tools and algorithms such as Naïve Bayes Classifier, Decision Tree and Random Forest etc [9].

A. Naïve Bayes Classifier

Table 2 shows the prediction accuracy for predicting runs with different sizes of training and testing datasets using Naïve Bayes Classifier.

Dataset	60% train - 40% test	70% train – 30% test	80% train – 20% test	90% train – 10% test
Accuracy (%)	43.08	42.95	42.47	42.50

Table 2: Prediction accuracy of Naïve Bayes for runs prediction

For prediction of runs, Naïve Bayes exhibited highest prediction accuracy of 43.08 with 60% training set and 40% testing set also least prediction accuracy of 42.47 with 80% training set and 20% testing set. From the above table we can conclude that the accuracy of the prediction using Naïve Bayes Classifier increases as we decrease the size of the training set and increase the size of testing set.

Table 3 shows various performance measures of Naïve Bayes Classifier with 60% training and 40% testing data.

Accuracy (%)	Precision	Recall	F1 Score
43.08	0.424	0431	0.418

Table 3: Performance of Naïve Bayes with 60% training and 40% testing data

B. Decision Tree

Table 4 shows the prediction accuracy for predicting runs with different sizes of training and testing datasets using Decision Tree.

Dataset	60% train – 40% test	70% train – 30% test	80% train – 20% test	90% train – 10% test
Accuracy (%)	77.93	79.02	79.38	82.52

Table 4: Prediction accuracy of Decision Tree for runs prediction

For prediction of runs, Decision Tree exhibited highest prediction accuracy of 82.52 with 90% training set and 10% testing set also least prediction accuracy of 77.93 with 60% training set and 40% testing set. From the above table we can conclude that the accuracy of the prediction using Decision Tree increases as we increase the size of training set and decrease the size of testing set.

Table 5 shows various performance measures of Decision Tree with 90% training and 10% testing data.

Accuracy (%)	Precision	Recall	F1 Score
82.52	0.824	0.825	0.824

Table 5: Performance of Decision Tree with 90% training and 10% testing data

C. Random Forest

Table 6 shows the prediction accuracy for predicting runs with different sizes of training and testing datasets using Random Forest.

Dataset	60% train – 40% test	70% train – 30% test	80% train – 20% test	90% train – 10% test
Accuracy (%)	89.92	90.27	90.67	90.88

Table 6: Prediction accuracy of Random Forest for runs prediction

For prediction of runs, Random Forest exhibited highest prediction accuracy of 90.88 with 90% training set and 10% testing set also least prediction accuracy of 89.22 with 60% training set and 40% testing set. From the above table we can conclude that the accuracy of the prediction using Random Forest increases as we increase the size of training set and decrease the size of testing set.

Table 7 shows various performance measure of Random Forest with 90% training and 10% testing data.

Accuracy (%)	Precision	Recall	F1 Score
90.88	0.908	0.908	0.908

Table 7: Performance of Random Forest with 90% training and 10% testing data

Among the three discussed Data Mining Algorithms: Naïve Bayes, Decision Tree and Random Forest it is found that the prediction accuracy of Random Forest is more. Decision Tree provides moderate prediction accuracy and Naïve Bayes gives least prediction accuracy in predicting the match outcomes in IPL. The prediction accuracy can be further improved by applying Machine Learning Techniques such as RNN and HMM to the dataset using Rapid Miner which is an integrated environment that supports all steps of Machine Learning process such as data preparation, data processing, model validation etc.

Classifier	Accuracy (%)			
	60% train – 40% test	70% train – 30% test	80% train – 20% test	90% train – 10% test
Naïve Bayes Classifier	43.08	42.95	42.47	42.50
Decision Tree	77.93	79.02	79.38	80.46
Random Forest	89.92	90.27	90.67	90.88

Table 8: Accuracies of algorithms for predicting runs

Table 8 shows the prediction accuracies of various Data Mining Algorithms for predicting runs.

V. Conclusion

In this work, a new model is developed for predicting runs by considering the previously scored runs by the batsman as the observed data. In order to achieve the task, large size of dataset having details about 577 IPL matches was taken into consideration with independent and dependent variables. Research in this paper concludes that RNN and HMM gives the best prediction accuracy for predicting runs in IPL. This model is distinct in its own ways, as Machine Learning techniques was not much used for prediction of IPL match results. The developed model helps in analyzing and predicting IPL match results.

Similar work can be negotiated for other formats of the game such as test cricket, ODI matches and T20 matches. The model can be further refined to reflect necessary characteristics of various other aspects such as weather conditions, injured players etc. that contribute to the end result of the match.

References

[1] Pranavan Somaskandhan, Gihan Wijesinghe, Leshan Bashitha Wijegunawardana, Asitha Bandaranayake and Sampath Deegalla, “Identifying the Optimal Set of Attributes that Imposes High Impact on the End Results of a Cricket Match using Machine Learning”, IEEE, 2017.

[2] Prince Kansal, Pankaj Kumar, Himanshu Arya and Aditya Methaila, “Player Valuation in Indian Premier League Auction using Data Mining Technique”, IEEE, 2014.

[3] Rabindra Lamsal and Ayesha Choudhary, “Predicting Outcome of Indian Premier League (IPL) Matches using Machine Learning”.

[4] A N Wickramasinghe and Roshan D Yapa, “Cricket Match Outcome Prediction using Tweets and Prediction of Man of the Match using Social Networks Analysis: Case Study using IPL Data”, IEEE, 2018.

[5] Tejinder Singh, Vishal Singha and Parteek Bhatia, “Score and Winning Prediction in Cricket through Data Mining”, IEEE, 2015.

[6] Kalpdrum Passi and Niravkumar Pandey, “Increased Prediction Accuracy in the Game of Cricket using Machine Learning”, International Journal of Data Mining and Knowledge Management Process (IJDKP), Vol.8, No.2, March – 2018.

[7] https://towardsdatascience.com/recurrent-neural-networks-and-lstm-4b601dd822a5.

[8] https://en.wikipedia.org/wiki/Hidden_Markov_model.

[9] Shimona S, Nivetha S and Yuvarani P, “Analzing IPL Match Results using Data Mining Algorithms”, International Journal of Scientific and Engineering Research, Volume 9, Issue 3, March – 2018.

[10] GuYu, Yu Fei Ma, Hong Jiang Zhaing and Shiqiang Yang, “A HMM based Semantic Analysis Framework for Sports Game Event Detection”, IEEE, 2013.

[11] Mayank Khandelwal, Jayanth Prakash and Tribikram Pradhan, “An Analysis of Best Player Selection Key Performance Indicator: The Case of Indian Premier League (IPL)”, Springer, 2016.