2018 8th International Conference on Communication Systems and Network Technologies

Predicting Results of Indian Premier League T-20 Matches using Machine Learning

[1] Shilpi Agrawal
Computer Science and Engineering
Anand Engineering College
Agra, India
agarwal.shilpi1@gmail.com

[2] Suraj Pal Singh
Computer Science and Engineering
Eshan College of Engineering
Mathura, India
srjsingh21@gmail.com

[3] Jayash Kumar Sharma
Computer Science and Engineering
Anand Engineering College
Agra, India
jayash.sharma@gmail.com

***Abstract*—Cricket, the most exciting and fascinating game that the people of all age group are very crazy to see and play. It is considered to be the most interesting and uncertain game. For many it becomes a billion dollar market as they speculate financially, hope of being able to earn profit. Every year the gambling market is going to be on hike as there is much great concern about spot fixing. In this paper, we have studied the problem of predicting the uncertainty of who will win the upcoming IPL match based on the individual competency of each player, coordination and team work of whole team evolving and technique followed by each team in each match. In this paper we propose a model using machine learning algorithms that can predict winning team based on past data available. We applied three machine learning algorithms namely Support Vector Machine, CTree and Naïve Bayes and achieved an accuracy of 95.96%, 97.98% and 98.99% respectively.**

***Index Terms*—IPL, Indian Premier League, Cricket, T-20, Support Vector Machine**

I. Introduction

Cricket is most widespread and much-loved game of everyone. It is delighted in by the general population of all age mass as it is exceptionally fascinating and suspicious game. Cricket is also referred as Game of Uncertainty and there is no any precise forecast that a specific team would win in any given conditions. Finally a team wins which multiplies the energy of every team member. There turn into a major jam of cricket darlings in the stadium and TV rooms to see the cricket at whatever point i.e., an international level, national level or any test match. The magnetism of cricket has also included business community that became source of income for them as they gamble over their favorite teams.

Popularity of cricket increased when ICC (International Cricket Council) started concept of fast cricket in the form of twenty-20(T-20) matches. In 2007, first twenty-20 world cup was held in the South Africa that was won by India which increased the popularity of this game in India. BCCI (The Board Of Control For Cricket In India) cashed the opportunity and created a league known as Indian Premier League (IPL) in 2008 and got it approved by ICC. IPL is one of the finest twenty-20 cricket competition in present cricketing world that is based on EPL (English Premier League) football league and NBA (National Basketball Association) Basketball League [5]. During its first edition, IPL gained huge popularity which opened avenues for many stakeholders. In every IPL season there are 8 teams that play with each other in the first stage, after first stage 4 teams go to eliminator round (next stage) and after eliminator round 2 teams go to final match and at last there will be one winner. Each team is owned by a franchise that is owned by group of people. These franchises hire players, evaluate them on the basis of their national, international, T-20 experience and performance and hire them at the time of auctions [1].

Results of every match in the IPL depends on the various conditions like venue, player performance, toss, performance in power play etc. Results of a match can only be predicted to some extent if previous player performance, venue and other match related data is available. In this paper authors predict the result of IPL match using three machine learning algorithms namely Support Vector Machine (SVM), Naïve Bayes and CTree on the basis of previous data available [3] [6].

The rest of the paper is organized as follows: Previous related work related to prediction of matches has been discussed in Section-II. Proposed prediction model is presented in Section-III, Section IV deals with results and subsequent discussions. Conclusion and future work is given in Section V.

II. Related Work

During the past, several researchers have contributed their efforts towards result prediction. Kampakis and Thomas [7] proposed a machine learning model that predicts the English-twenty over county cricket cup. Main motive behind his work was to service the gambling industry by using multi step approach. Authors took the dataset from archive available at cricinfo.com for the season 2009 to 2014 and calculated additional features like strike rate for predicting results. They applied four classifiers namely Naïve Bayes, Logistic Regression, Gradient Boosted Decision Trees and Random Forests and evaluated the benefits of home team and

978-1-5386-5956-4/18/$31.00 ©2018 IEEE
DOI: 10.1109/CSNT.2018.15

away team. Sankaranarayanan et al. [9] defined a prediction model via subsection of match factors with clustering and regression algorithms that evaluates historic Cricket game information and the on the spot conditions of a match to forecast progression of game and the final consequence of One Day International game. To track the progression, Ridge method and arbitrary attribute selection are used to predict the runs scored within the innings and focused on Milestone Reaching Ability (MRA) which is the aggregation of all the qualities of batsmen that consists of opening batsmen, middle-order batsmen, all-rounders, wicket-keeper, and tail-enders.

Padma et al. [4] studied data of One Day International Cricket matches of the Indian Cricket team and mined different association rules by using market basket tools with attributes. To apply rules they focussed on various things like toss result, toss winner, decision to bat first or not, which two teams are going to play, is this a home team or away team and the result of game. They analyzed matches of 10 years on an average by calculating support and confidence and predicted the unfavorable cases due to which a team loses game. Swartz et al. [8] developed a simulator that simulates each ball in real time. Simulator considers the probabilities of basic features like batting, bowling, over, inning in current scenario. They further used Monte Carlo method with Bayesian Latent Variable Model to predict the result of next ball to be thrown. They predicted that toss winner attribute has no reasonable advantage while playing. Result was calculating by considering both win game and lose game perspective by taking two samples using Binomial test and Bayesian model with different variations. Further logistic regression was used to show that playing on home ground adds some advantage due to familiar with local insights.

III. Proposed Model

This section deals with the general architecture of proposed model. Block diagram of the model has been shown in Figure-1. Proposed methodology includes Dataset Collection, Pre-processing of collected dataset, Feature extraction from raw data, conversion of categorical data into numerical data, partitioning of samples into training and test samples, training and classification. Details of each step is given in the following sections:

A. Collect Dataset

In order to predict IPL T-20 match result, dataset was collected from techgig.com The dataset comprises details of past 500 IPL matches. Each row in the dataset represent ball by ball details played in both the innings with 21 attributes (referred as Deliveries Data) as given in Table-I. Dataset also provides results of each IPL match with additional 14 attributes (referred as Match Results) as shown in table-II. **Deliveries Data**-dataset contains ball by ball record of past 500 IPL matches and also includes the data dictionary associated with it. **Matches Results**- dataset contains extra meta-data of each

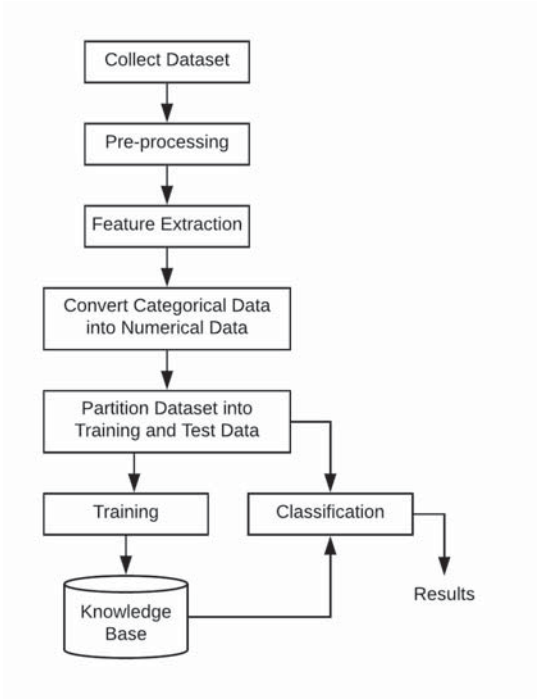

Fig. 1. Block Diagram for Proposed Model

match played. This data file contains additional data for all the 500 IPL matches that are in the Deliveries data such as where it has been played, who won the match, by what margin.

TABLE I
Deliveries data

Match Feature		
Match ID	Inning	Batting Team
Bowling Team	Over	Ball
Batsman	Non Striker	Bowler
Is Super Over	Wide Runs	Bye Runs
Legbye Runs	Noball Runs	Penalty Runs
Batsman Runs	Extra Runs	Total Runs
Player Dismissed	Dismissal Kind	Fielder

TABLE II
Match Results

Match Feature		
Match ID	Season	City
Team1	Team2	Toss Winner
Toss Decision	Result	Winner
Win By Runs	Win By Wickets	Player of the Match
Venue		

B. Pre-Processing and Feature Extraction

Collected dataset has raw data table at its initial stage which needs to be pre-processed for removing irrelevant details. Pre-processing stage cleans the dataset by removing those data that are not useful to get results. Data where results have not been declared or marked are removed during

TABLE III
ADDITIONAL FEATURES

Features Name	Description
Inning Batting Average Strike Rate	Total runs scored/number of balls faced
Inning Bowling Average Run Rate	Total runs given/number of balls thrown
Inning Power play Run Rate	Total runs given in first six overs/number of balls thrown in first six overs
Inning Power play Strike Rate	Total runs scored in first six overs /number of balls faced in first six overs
Winner	Winner of the match(team1, team2,tie)
Team1	Batting team of the match in first inning
Team2	Bowling team of the match in first inning

pre-processing stage. As given features in one dataset are not sufficient to predict the model, moreover, combination of all features given in dataset are not enough to predict results so it is necessary to identify the best features and add on extra features that can play major role in predicting the model. Following key features for both teams (batting and bowling) and for both innings have been identified which can help in result prediction:

- Average Run Rate
- Average Strike Rate
- Power Play Strike Rate

Attribute toss-winner decides who play first inning. For both the innings specific feature vectors shown in Table-III are used. Target attribute value ranges enormously due to participation of many teams in IPL. Every time the result can be in favor of different team. To classify these range of values, we calculated the winner attribute as ***Team-1*** if Team-1 wins, ***Team-2*** if Team-2 wins and ***Draw*** in case of tie.

C. Conversion of data format

Data provided in dataset is categorical in nature due to which classification is quiet complex. It may also affect classification process resulting in wrong prediction. In this step all categorical data in dataset except the target attribute (Winner) has been converted into numeric format and normalized on scale basis.

D. Training and Classification

For training and classification purpose, three predictive modeling classifiers Support Vector Machine (SVM), Naïve Bayes and CTree have been used with proposed model.

1) Support Vector Machine: SVM [2] was first used for text categorization problem that is based on the principle of Risk minimization emphasizing on ascertaining a hypothesis to guarantee minimum error. As it was not known at the initial stage, it was very difficult to calculate correct error. Using error that is calculated at the time of training and intricacy, the actual error can be confined. The foremost emphasis of Support Vector Machine is to reduce true error. Geometrically, it classifies data points into two classes using hyperplane i.e: +1 and -1. +1 represents Normal data and -1 represents the suspicious data. The hyperplane can be expressed as:

$$(W * X) + b = 0 \tag{1}$$

where W = w1,w2, ... ,wn represents the weight vector, X = x1, x2, ... ,xn is the attribute values and b is a scalar quantity. To find a linear optimal hyperplane for maximizing the margin of separation between the two classes is the main aim of Support Vector Machine. SVM trains the system by using the portion of the data.

2) Naïve Bayes classifier: This classifier works on the concept of probability Naïve bayes classifier assumes the class according to the membership probability. Dependent and independent variables relation is observed which derives the conditional probability. Working of Naïve Bayes Classifier can be shown as:

$$P(H|X) = (P(X|H).P(X))/P(H) \tag{2}$$

Here, X represents the data record, H is the hypothesis, P(H) is the prior probability, P(H/X) is the posterior probability and P(X/H) is the posterior probability. Naïve Bayes classifier can be constructed without any complex iterative parameter.

3) CTree Classifier: CTree classifier [9] is based on permutation which takes distribution description of the measures. The conditional distribution of statistics measuring the relationship between responses and covariates perform an exhaustive search over all possible splits maximizing an information measure. In this, every tree is created by selecting different data set randomly. It can tackle the high dimensional data easily.

IV. RESULTS AND DISCUSSIONS

This section emphasizes on the achieved results and subsequently discusses the findings of proposed model. The model dataset originally has 500 records of IPL T-20 matches which has been divided into two sets. First set is referred as Training Dataset (comprising of 400 records) while second set is referred as Test Dataset (comprising of 99 record). One record was ignored during pre-processing due to non-availability of results. Training dataset is trained with SVM, Naïve Bayes and CTree and obtained knowledge is used to predict results of Test dataset. Further achieved results are evaluated on the basis of accuracy, precision, recall and F-measure.

Result is calculated in terms of accuracy that is the percentage of team wins correctly classified by classifiers verses Total number of responses. Table-IV shows accuracy percentage obtained through different classifiers. A comparative analysis of accuracy is depicted in Figure-2.

TABLE IV
ACCURACY COMPARISON

Classifiers	Accuracy
SVM	95.96%
Naive Bayes	98.98%
CTree	97.97%

TABLE V
ACCURACY CALCULATION

Actual Result	Predicted Result	Remarks
Team 1 wins 44 matches	Team 1 wins 46 matches	44 corrected predicted + 2 false prediction
Team 2 wins 54 matches	Team 2 wins 53 matches	53 corrected predicted + 0 false prediction
One Draw Match	0	1 false prediction

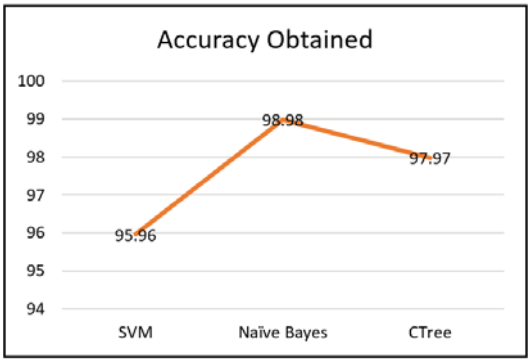

Fig. 2. Accuracy Comparison Graph

As an example let's discuss CTree which is trained using 400 matches data and results are predicted for remaining 99 matches data. Accuracy calculation on CTree is shown in Table-V. Accuracy is calculated as:

Accuracy= (97/99)*100 = 97.9797... =97.98%
Table-VI shows achieved results and evaluation of result on accuracy, precision, recall and F-measure. Results clearly shows that Naïve Bayes outperforms SVM and CTree as it is independent of partially taken attributes and gives importance to each attribute equally.

TABLE VI
COMPARATIVE PERFORMANCE OF CLASSIFIERS

Model Classification (Drawn Matches, Second Team Wins and First Team Wins)				
Classifiers	**Case**	**Precision**	**Recall**	**F-Measure**
SVM	Drawn Match	0.00	0.00	0.00
	Team-2 Wins	96.22	98.07	97.14
	Team-1 Wins	97.77	93.61	95.65
Naive Bayes	Drawn Match	100.00	66.66	80.00
	Team-2 Wins	100.00	100.00	100.00
	Team-1 Wins	97.95	100.00	98.96
CTree	Match Drawn	0.00	0.00	0.00
	Team-2 Wins	100.00	100.00	100.00
	Team-1 Wins	100.00	100.00	98.00

Performance evaluation on three parameters precision, recall and F-measure have been shown in Figure-3, 4, 5 for drawn matches, First batting team and second batting team subsequently. Here, we denote Team-1 as the first batting team and Team-2 as the second batting team.

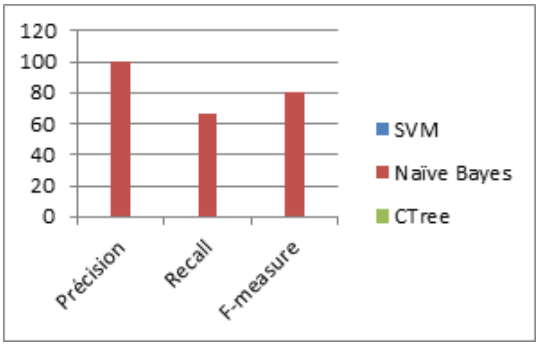

Fig. 3. Comparison of Parameters for Drawn Matches

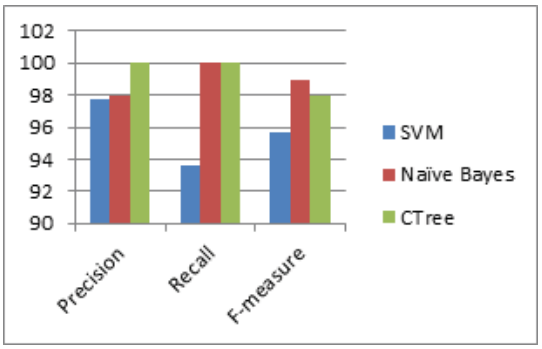

Fig. 4. Comparison of Parameters for Winning Team-First Batting Team

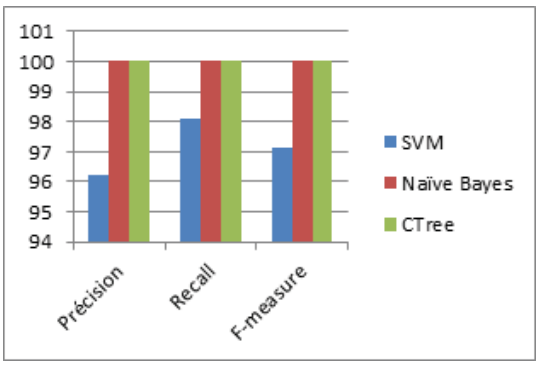

Fig. 5. Comparison of Parameters for Winning Team-Second Batting Team

V. CONCLUSION AND FUTURE WORK

Our main focus is not to support the gambling industry that reveals the uncertainty of the game. Rather we are interested in revealing some useful information through our research.

In this work, historical data has been collected from real IPL cricket matches and useful features have been extracted after pre-processing of data. Further, suitable data is converted to a numeric form and scale it on three parameters win, loss, and tie. This data is trained and classified with three classifier SVM, CTree and Naïve Bayes using R Tool. Outstanding results have been achieved using Naïve Bayes classifier with an overall accuracy of 98.98%.

As our approach well predicts the IPL in current scenario that is based on the history records, it can be further extended in changing environment when many new talents join the team, their history records are made available. Further it can be tested by analyzing IPL 2019 match result. Accordingly, new features vectors can be identified and prediction can be made more accurate.

REFERENCES

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

[2] Joachims T, Nedellec C, and Rouveirol C.(Eds.),"Text categorization with Support Vector Machines: Learning with many relevant features Machine Learning", ECML-98, Springer Berlin Heidelberg, 1998, 137-142.

[3] Kansal P, Kumar P, Arya H and Methaila A, "Player valuation in Indian premier league auction using data mining technique", International Conference on Contemporary Computing and Informatics (IC3I), 2014, 197-203.

[4] K. A. A. D. Raj and P. Padma, "Application of Association Rule Mining: A case study on team India", 2013 International Conference on Computer Communication and Informatics, 2013

[5] Saikia, Hemanta and Bhattacharjee, Dibyojyoti, "On classification of all-rounders of the indian premier league (IPL): A Bayesian approach", Vikalpa, 2011, 36, 51-66

[6] S. Singh, "Measuring the Performance of Teams in the Indian Premier Leaguee",American Journal of Operations Research, 2011, vol 1, No 3, pp 180-184

[7] Stylianos Kampakis, William Thomas,"Using Machine Learning to Predict the Outcome of English County twenty over Cricket Matches", Cornell University, 2015

[8] Tim B. SWARTZ, Paramjit S Gill and S. Muthukumarana,"Modelling and simulation for one-day cricket", Canadian Journal of Statistics, 2009, Vol 37, No 2, pp-143-160

[9] Veppur Sankaranarayanan, Vignesh and Sattar, Junaed and Lakshmanan,"Auto-play: A Data Mining Approach to ODI Cricket Simulation and Prediction",SIAM Conference on Data Mining, 2014