Volume 1, No. 1, Issue 3

Score Prediction in Indian Premier League through Data Mining

Shashwat Kumar Sinha

Department of Computer Application, Babu Banarasi Das University, Lucknow, India
Email: shashwat.sinha00@gmail.com

Abstract

The IPL score is currently projected on the basis of run-rate and there is a lack of intuition on the basis of historical features as T20 cricket is considered unpredictable but through the advancement of data mining algorithms and as the Indian Premier League which has crossed the fifteen years landmark the ball-by-ball data is accessible and can be the subject of predictive analysis. The work we propose is based on historical features which affect the course of prediction quite drastically as the setting of the Indian Premier League has been in India for most of the seasons. A Linear Regression method has been proposed to predict the score of a IPL team on the basis of the current score given by the team and the wickets fallen after a certain interval ranging which impact a T20 match a lot and after the two strategic time-outs. The role of impact player is not considered since the rule has been implemented only for two years. A study has been proposed that how much these breaks affect the outcome of an IPL match and results in the prediction of an IPL match. The prediction accuracy has been increasing remarkably when the intervals of the games have been increased. The attributes considered affect the outcome at a much greater scale as compared to the other attribute and thus reduces the chances of over-fitting of the algorithm.

***Keywords**– Linear Regression; Historical Data of the Indian Premier League; Score Prediction; Final Score*

INTRODUCTION

Indian Premier League is a professional Twenty20 cricket league in India contested during April and May of every year by teams representing Indian cities. The IPL is the most attended cricket league in the world and ranks sixth among all sports leagues. Cricket is an outdoor game played on a cricket field at 22-yard rectangular long pitch between two teams consisting each of 11 players. It is played in three formats namely Test, One Day International (ODI) and T20.In IPL cricket. The batsman looks for making runs by hitting the ball being bowled to him. The bowler on the other hand tries to get the batsman out. There are certain rules defined to get the batsman out by the bowlers or the fielders. Each batsman keeps on batting until he gets out. So, the innings of the batting team is over when either the 10 batsmen got out or the 20 overs have been bowled by the fielding team; in either of the situation the batting team now gets the chance of bowling, and the bowling team gets the chance of batting. The team which scores more runs wins the match. Unlike other sports, cricket stadiums size and shape are not fixed except the dimensions of the pitch and inner circle which are 22 yards and 30 yards respectively. The cricket rules do not mention the size and the shape of the field of the stadium. Pitch and outfield variations can have a substantiate effect on batting and bowling. The bounce, seam movement and spin of the ball depends on the nature of the pitch. The game is also affected by the atmospheric conditions such as altitude and weather. A unique set of playing conditions are created due to these physical differences at each venue. Depending on these set of variations a particular venue may be a batsman friendly or a bowler friendly. Currently, in IPL cricket the projected scores can be seen displayed at the score card during the first innings, which is basically the final score of the batting team at the end of that innings if it scores according to the current run rate or a particular rate. Run rate is defined as the number of runs scored per the number of overs bowled. However, run rate is considered as the only criteria for calculating the final score. But there are other factors too which may affect the final score like the number of wickets fallen and the batting team itself. In this paper, a method has been proposed in which the final score can be predicted based on input given on wickets fallen and the runs scored by the batsman by the applying linear regression on the past data in the IPL. The past records have been taken from the historical data of all nine seasons of the IPL and the algorithm has been applied to that to predict the projected score of the IPL matches. The structure of the paper is a follows: the following section the related works done in the game of cricket has been discussed briefly. In section III, a regression algorithm has been stated for the overview and the implementation has been demonstrated. Section IV focuses on the data collection and preparation while section V discusses about the training and testing of the data. In section VI, the statistical analysis has been done and the error in Linear Regression has less than the existing method of predicting scores in IPL matches. The error rate has also been compared with the present method for score prediction and the results are positive.

RELATED WORK

Very few have worked in statistically predicting the scores or the outcome of the ODI match. One such work is called Winning and Score Predicting (WASP), which has been done by

Scott Brooker and Seamus Hogan at University of Canterbury as part of the PhD research project [9]. It estimates how well the average batting team will do against the average bowling team under given conditions and the current state of the game. In the first innings it estimates the additional runs that can be scored with the given number of balls and wickets remaining. The estimates have been made from a dynamic programming [9]. Likewise, Raj and Padma [10] analyzed the Indian cricket team's ODI matches data and apply association rules on the attributes namely home or away game, toss, batting first or second and the match result. Swartz et al. [11] use Markov Chain Monte Carlo methods to simulate ball by ball outcome of a match using a Bayesian Latent variable model. Depending on the ability of current batsman, bowler and game situation like number of balls delivered and number of wickets fallen, the outcome of the next ball had been predicted. But the model suffers from severe problems as noted by the authors themselves: the likelihood of a given batsman having previously faced a given bowler in previous games in the dataset is low. Kaluarachchi and Varde [12] implemented both Nave Bayes classier and association rules and analyzed the factors contributing to a win. But they do not estimate the final score of the innings. While [9] is very much like the model that we are making, in terms of the output generated, that is, predicting the final score of the first and winning probability in the second innings. However, they have implemented dynamic programming over the dataset of the matches since 2006.In contrast, data mining concepts like Linear Regression is used for prediction of IPL matches based on the data available and attributes considered by the algorithm. A. Data Mining in Various Sports A lot has been analyzed about the prediction of match results in football, baseball, basketball etc. For example, Bhandari et al. [5] created the Advanced Scout system for identifying various trends from basketball matches. In football, Luckner et al. [7] estimated the outcome of 2006 World Cup FIFA matches using live Prediction Markets. In baseball, Gartheepan et al. [8] made a data driven model that helps when to 'pull a starting pitcher'. Lutz [6] made a model of selecting the players combination that is most appropriate for winning the games. These works have been developed for a particular sport with different algorithms and techniques of data mining.

REGRESSION LINEAR REGRESSION

In machine learning, Linear Regression is a statistical process for estimating the relationships among variables. It includes many techniques for modeling and analyzing several variables when the focus is on one or more independent variables (Predictors). This predictive analysis is done on IPL data set where the independent variables considered are the runs scored by the batsman in intervals ranging from 6 overs,10 overs and 15 overs. The linear regression as a predictive model is represented as a hypothesis function depending on these independent variables. h(theta)= theta0 * x0 +theta1 * x1 + theta2 * x2 where theta is vector of values theta0, theta1 and theta2 and x0,x1,x2 are values of vector X. x0 is the intercept value treated as 1 x1,x2 are values of the runs scored by the batsman and the wickets fallen at that particular interval and on the basis of that the final score has been projected. The model has been represented in a vectorized manner and thus operations are done on the basis of linear algebra. For minimizing the cost function gradient descent approach is implemented for getting the value of theta and thus finding the projected score. Finding theta through normal equation has also been used for cross validating the result and checking the value given by the gradient descent. The input given is the resulting score after a particular interval of various ranges and the wickets fallen after the interval. The ranges can be from 6 to 15 overs.

DATA COLLECTION AND PREPARATION

The data has been collected from http://www.kaggle.com, where ball-by-ball data of all matches are available publicly. The data set consist of all matches including rain interrupted and rain-abandoned games played from the first season of IPL to the 9th season of the IPL. It consist of 13 teams namely Kolkata Knight Riders, Royal Challengers Bangalore , Chennai Super Kings, Kings XI Punjab, Rajasthan Royals, Delhi Daredevils, Mumbai Indians, Deccan Chargers, Kochi Tuskers Kerala, Pune Warriors, Sunrisers Hyderabad, Rising Pune Supergiants and Gujarat Lions. Also it contains the dataset of the each and every ball with player id's and dismissal ids and requires many manipulation to provide the input which has to be processed by the algorithm. For each team dataset are extracted though queries and required data is extracted. Furthermore each dataset can be divided into match periods on which the predictive analysis has to be done (like the runs scored and the wickets fallen at the end of power-play overs (0-6 overs), 0-12 overs after the strategic time-out which is a part of IPL and so on). After extraction of the dataset in a structured way the number of matches played are 576 and the innings will be twice of that will be used in training of the machine learning algorithm. The set of attributes considered are the current score and the wickets taken in the given interval and thus resulting the projected score from the input supplied.

TABLE I

DESCRIPTION OF THE ATTRIBUTES

Attributes	Description
Batsman Scored	The runs scored by the batsman
Wickets Fallen	The number of wickets fallen after the intervals

TRAINING AND TESTING DATA

The dataset has been analyzed in Octave through implementing functions of Normal Equations, Cost function, minimization of cost function through gradient descent and has been visualized using Weka. The dataset has been portioned separately into 3 components namely training set, test set, and cross validation set for the IPL matches played throughout the course of 9 seasons played in the IPL. Linear Regression has been implemented on all 3 datasets to calculate the error rate and help in the diagnosis of the error percentage to go down to the least by plotting the learning curves and adding in case of over-fitting of the algorithm. In 10-fold cross validation the given sample is arbitrarily partitioned into 10 equal size subsamples. The result from the fold is combined to get a single estimation.

RESULTS AND DISCUSSIONS

Implementation of Algorithms

A. Linear Regression

The Linear regression has been applied on one of the thirteen teams and a score is calculated from the equation: Score= 0.66*current score + 23.7*wicket+fallen + 246.2 The equation for all thirteen teams can be calculated and the result can vary for different teams based on the historical features present for the particular team. The intervals also can be varied 0-6,0-10,0-12,2-10 overs and so on)

B. Projected Score Performance

The linear regression error rate is compared both for the training dataset, test set and cross validation set as well as the current scheme of run rate is considered and compared with the proposed algorithm.

TABLE II

ACCURACY FOR DIFFERENT RANGE OF OVERS FOR A PARTICULAR TEAM

Interval	Accuracy
0-6	33-47%
0-10	53-65%
0-15	66-81%

CONCLUSION AND FUTURE WORK

The main purpose of the paper is to make a model for predicting the score of an IPL team based on historical features which resulted in efficient learning of the linear regression algorithm reducing the linear regression error by a significant amount as compared to the run-rate method for projection which is currently prevalent. The model being based on the historical features and the fact that most of the IPL matches are played in India predicts the score for teams accurately. The algorithm can be more generalized as it is prone to under-fitting and the fact that IPL data is restricted to a particular country. The attributes can be increased but the author of the future work must keep in mind the problem of over-fitting through adding features which might not affect the game drastically and act as false positives. Thus, keeping these factors in mind, the future model may provide more accurate results. References are important to the reader; therefore, each citation must be complete and correct. If possible, references should be commonly available publications.

REFERENCES

1. F. C. Duckworth and A. J. Lewis. A fair method for resetting the target in interrupted one-day cricket matches. The Journal of the Operational Research Society, 49(3) pp. 220227, 1998.
2. M. Bailey and S. R. Clarke. Predicting the match outcome in one-day international cricket matches, while the game is in progress. Journal of sports Science and Medicine, 5(4):480487, 2006.
3. P. E. Allsopp and S. R. Clarke. Rating teams and analysing outcomes in one-day and test cricket. Journal of the Royal Statistical Society. Series A (Statistics in Society), 167(4) pp. 657667, 2004.
4. Machine Learning Group at the University of Waikato. Weka 3: Data Mining Software in Java.
    a. http://www.cs.waikato.ac.nz/ml/wekaAccessed 12 February 2015

5. I. Bhandari, E. Colet, and J. Parker. Advanced Scout: Data mining and knowledge discovery in NBA data. Data Mining and Knowledge Discovery, 1(1):121125,1997.
6. Schultz- D. Lutz. A cluster analysis of NBA players. In MITSloan Sports Analytics Conference, 2012
7. Luckner, Kratzer STOCCER - a forecasting market for the FIFA World Cup 2006
8. Gartheeban Ganeshapillai, John Guttag. A Data-driven Method for In-game Decision Making in MLB March 1. 2014Hynes Convention Center
9. Hogan, Seamus (2012-11-22). "Offsetting Behaviour: Cricket and the Wasp: Shameless self promotion (Wonkish)". Offsettingbehaviour.blogspot.in. Retrieved 2014-02-03.
10. Antony Arokia Durai Raj K, Padma Panchapakesan. Application of Association Rule Mining: A case study on team India Computer Communication and Informatics (ICCCI), 2013 International Conference
11. Tim B. SWARTZ1*, Paramjit S. GILL2 and Saman MUTHUKUMARANA Use Modelling and simulation for one-day cricket
12. Kaluarachchi and Varde CricAI: A classification based tool to predict the outcome in ODI cricket December 2010