2015 International Conference on Soft Computing Techniques and Implementations- (ICSCTI)
Department of ECE, FET, MRIU, Faridabad, India, Oct 8-10, 2015

Score and Winning Prediction in Cricket through Data Mining

Tejinder Singh,
Computer Science & Engineering
Thapar University
Patiala, Punjab, India
teji.tsk@gmail.com

Vishal Singla,
Computer Science & Engineering
Thapar University
Patiala, Punjab, India
vsingla160@gmail.com

Parteek Bhatia
Computer Science & Engineering
Thapar University
Patiala, Punjab, India
parteek.bhatia@thapar.edu

Abstract—Currently, in One Day International (ODI) cricket matches first innings score is predicted on the basis of Current Run Rate which can be calculated as the amount of runs scored per the number of overs bowled. It does not include factors like number of wickets fallen and venue of the match. Furthermore, in second innings there is no method to predict the outcome of the match. In this paper a model has been proposed that has two methods, first predicts the score of first innings not only on the basis of current run rate but also considers number of wickets fallen, venue of the match and batting team. The second method predicts the outcome of the match in the second innings considering the same attributes as of the former method along with the target given to the batting team. These two methods have been implemented using Linear Regression Classifier and Naïve Bayes Classifier for first innings and second innings respectively. In both methods, 5 over intervals have been made from 50 overs of the match and at each interval above mentioned attributes have been recorded of all non-curtailed matches played between 2002 and 2014 of every team independently. It has been found in the results that error in Linear Regression classifier is less than Current Run Rate method in estimating the final score and also accuracy of Naïve Bayes in predicting match outcome has been 68% initially from 0-5 overs to 91% till the end of 45th over.

Keywords— Linear Regression; Naïve Bayes; Data Mining; Projected Score; Winning Probability.

I. INTRODUCTION

Cricket is the second most popular sports in the world with billions of fans across India, UK, Pakistan, Africa, Australia, etc. [1]. It 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 Twenty Over International (T20). In ODI each team takes its chance to bat, trying to score as many amount of runs which can be scored in 50 overs while the other team fields for that much amount of overs. Each chance is termed as an innings [2].

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 50 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 stadium's size and shape is 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 [2]. 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 an ODI match 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 amount 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 number of wickets fallen, the venue and the batting team itself.

In this paper, a method has been proposed in which the final score can be predicted of the first innings and the winning probability of the batting team in the second innings can be estimated. In the former case Linear Regression Classifier has been used and in the latter Naïve Bayes Classification has been implemented. Unlike the current procedure for projecting the score, the factors like the venue of the match, the number of wickets fallen and the batting team have been considered in the estimation and in the second innings, the target given to the batting team has been included along with the factors taken in the first innings, for probability estimation. These past records have been taken from all the

non-curtailed ODI matches played among the top eight countries from 2002 to 2014.

The structure of the paper is as follows. In the following section the related works done in the game of cricket or any other sports have been discussed briefly. In section III, an overview on classification has been given and the algorithms implemented for predicting the final score and match's outcome have been demonstrated. Section IV focusses on the data collection and preparation while section V discusses about training and testing of data. In section VI, the statistical analysis has been done and it has been found that the error in Linear Regression classifier is less than that of the existing method of predicting the score and also the accuracy for the Naïve Bayes classifier has been calculated. Conclusion and future scope are given in Section VII.

II. RELATED WORK

Very few have worked in statistically predicting the scores or the outcome of the ODI match. One such work 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 about 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. In the second innings it estimates the winning probability with the given number of balls and wickets remaining, runs scored at the given situation and the target given. The estimates have been made from a dynamic programming [9].

Likewise, Raj and Padma [10] analysed 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 Naïve Bayes classier and association rules and analysed the factors contributing to a win. But they do not estimate the final score of the innings.

While [9] is very much similar to 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 and Naïve Bayes Classifiers have been implemented on the dataset of the ODI matches played from 2002 to 2014 in this method.

A. *Data Mining in Various Sports*

A lot has been analysed 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'. Schultz [6] made a model of selecting the players' combination that are most appropriate for winning the games.

These works have been developed for a particular sport with different algorithms and techniques of data mining.

III. CLASSIFICATION

In machine learning, classification is the technique of recognising to which class a new instance belongs, on the basis of a training set of data containing observations whose class association is known. For example assigning an unknown sample of a flower into the given flower species "Jasmine" or "Rose" based upon its characteristics or features. Classification is a supervised learning in which a training set of correctly identified instances is given. The algorithm that is used for implementation of classification is known as a classifier.

Here,two classifiers have been used, namely Linear Regression (LR) and Naïve Bayes classifier.

A. *Linear regression*

If the data of a class and attributes are numeric then linear regression classifier is used for classification. This is a staple method in statistics. The concept is to get the expression of the class in terms of linear combination of the predetermined weights and attributes which is been given by the equation (i).

$$x = w^0 + w^1.a^1 + w^2.a^2 + \ldots + wv.av \qquad (i)$$

Where x is the class; a_1, a_2, ...a_v are the values of the attributes; and w_0, w_1, ... $w_v$are weights. The weights can be computed from the training data. Here, the notation gets a little heavy, because a way of expressing the attribute values is needed for each training instance [16].

In this paper, the Linear Regression Classifier have been implemented for the first innings datasets where the class attribute 'x' is the 'Score' and the input attribute 'a_i'(i =

1,2,3…) are the current score and wickets fallen at 5-over period starting from 0-5 till 40-45 overs.

B. Naïve Bayes

Naive Bayes classifier is the probabilistic classifier which is based on Bayes' theorem which depends upon the strong (naive) independence assumptions. This classifier assumes that presence (or absence) of a particular feature of a class is independent to the presence (or absence) of any other feature. It considers all of these features to independently contribute to find the probability even if the attributes depend on each other. Depending on the accuracy of the probability, Naive Bayes classifier is used for efficiently training a supervised learning setting [14].It is based on the model of conditional probability in which a given instance has be classified is given by a vector x = (x1,……,xn) which represents some n features (dependent variables) of the n attributes, A1, A2, . . . , An respectively, it assigns the probabilities to each of k possible classes or outcomes [13].

Using the Bayes' theorem, the conditional probability is given

$$P(Ci|X) = \frac{P(X|Ci)\ P(Ci)}{P(X)} \qquad \text{(ii)}$$

by the equation (ii).

Where,P(Ci|X) is the posterior probability of the class, given predictor (attribute),P(Ci) is the prior probability of class, P(X|Ci) is the likelihood which is the probability of predictor given class and P(X) is the prior probability of predictor. The Equation (ii) can also be written of the form as depicted in equation (iii).

$$\text{posterior} = \frac{\text{prior x likelihood}}{\text{evidence}} \qquad \text{(iii)}$$

The evidence, also termed as normalizing constant is equal to the sum of the posteriors. The evidence can be ignored as it is a positive constant. (Normal distributions are always positive)[15]. So, only calculate the numerator of the equation (ii) which is P(X|Ci)P(Ci).

This classifier is mainly applied to find the decision rule. The most common rule is to choose the one which is the most probable and that is known as the Maximum A Posteriori (MAP) decision rule. That is, out of all the probabilities that have been calculated using Naïve Bayes, the one with the maximum probability will be selected for decision making which can be calculated by using equation (iv).

$$posterior(Ck) = \max_{k\in\{1,\dots,k\}} P(Ck) \prod_{i=1}^{n} P(xi|Ck) \qquad \text{(iv)}$$

Now, if numerical variables are there then they are assumed to be normal distributed for numerical variables [17]. Hence, Gaussian Naïve Bayes is used in this case which is similar to Naïve Bayes Classification except it deals with continuous data. The main assumption of the classifier is that continuous values are distributed on the basis of Gaussian distribution which are related to each class. Let x be a continuous attribute that the training data contains. Then firstly, the data is segmented by the class; the mean and the variance of x in each class is then computed. Let μ and σ^2 is the mean and the variance of the values in x related with class c respectively. So, probability distribution of any value of a particular class, P(x=m|c), is computed by inserting m in the equation (v) representing the probability as Normal distribution.

$$P(x = m|c) = \frac{1}{\sqrt{2\pi\sigma^2}} exp - \frac{(m-\mu)^2}{2\sigma^2} \qquad \text{(v)}$$

In this paper, Gaussian Naïve Bayes has been implemented as the data collected for the second innings has all the input attributes as numeric with only the class attributes being nominal. Therefore, equation (v) has been used for posterior calculations and thus finding the winning probability.

IV. DATA COLLECTION AND PREPARATION

The data has been collected from http://www.espncricinfo. com, where over-by-over data of all the matches are available publicly. The dataset consists of complete matches excluding all the rain-interrupted and rain-abandoned games, played between 2002 and 2014 among the 8 teams namely Australia, India, New Zealand, South Africa, England, Sri Lanka, Pakistan and West Indies. Also it contains the dataset of the matches played on the venues from each country mentioned above.

For each team two separate datasets have been made, one for first innings and the other for the second innings. Similarly for each venue also two datasets are there. Furthermore, each dataset contains the 5 – over period statistics of the matches. Now, for the first innings those complete matches have been taken where the particular team has batted first in the first innings. For example, the first innings dataset of India contains the record of those matches only where India has batted first. From these matches the runs scored, wickets fallen at each 5 over period (like runs scored and wickets fallen at the end of 5th over, 10th over, 15th over and so on till the 50th over or by the fall of tenth wicket) along with the final score at the end of the innings, have been considered.

In the second innings also those matches have been included in which the particular team has batted in the second innings only. For example, the second innings dataset of India contains only those matches in which India has batted in the second innings. In addition to this, from these matches the

runs scored, wickets fallen by the 5 over period (like runs scored and wickets fallen at the end of 5th over, 10th over, 15th over and so on till the 50th over or by the fall of tenth wicket or till the score has been chased) along with the target given to the batting team and the final result in terms of 'Yes' or 'No' depicting win or defeat respectively of that team at the end of the innings, have been considered. Table I shows the description of these attributes that have been taken into consideration in each of the two innings. Except target, all the remaining attributes are common in both the innings, though the values of each will be different according to the innings and situation of the match.

Table 1: Description of the Attributes

ATTRIBUTES	DESCRIPTION
Batting Team	The team which is currently batting.
Current Score	The current score of the match of the batting team after particular overs either in first innings or in second innings.
Overs	The number of overs bowled at a particular stage of the batting team.
Score	The final score of the team at the end of the first innings.
Target	The target given to the team batting in the second innings.
Venue	The stadium where the match is being played.
Wickets Fallen	The number of wickets fallen at a particular situation of the batting team.

V. TRAINING AND TESTING OF DATA

The datasets have been analysed in Weka which contains various machine learning algorithms for implementing the data mining process. Weka includes various tools for data clustering, visualization, classification, pre-processing, regression and association rules [3].

The dataset has been portioned separately into training and testing sets as the ODI matches played till 2013 and the ODI matches played in 2014 respectively. For the first innings of a particular team at a particular venue the Linear Regression Classifier has been implemented on the training dataset, with ten-fold cross validation and the results have been tested with the testing set of matches of that team and venue. Similarly, for the second innings of the same team and venue, Naïve Bayes Classifier has been implemented on the training dataset of the second innings, with ten-fold cross validation and tested the results with the testing set of matches of that team and venue. This way through linear Regression, first innings score of a particular team at the particular venue at different situations of the match can be predicted. On the other hand, Naïve Bayes will estimate the probability of the batting team in the second innings, at a particular venue at different situations of the match.

In 10-fold cross-validation, the given sample is arbitrarily partitioned into 10 equal size subsamples. One subsample is used for testing the model, and the other 9 subsamples are sent for training out of the 10 subsamples. This process is iterated 9 times, with each of 10 subsamples used only once as the testing data. Then, the 10 results from the folds is combined to get a single estimation.

VI. RESULTS AND DISCUSSIONS

A. *Implementation of Algorithms*

*1) Linear Regression:*The Linear regression classifier has been applied on one of the teams using equation (i) for predicting the first innings score when the team has played just 0-5 overs (out of 50 overs).

$$\text{Score} = 0.6608 * \text{current\_score} - 23.7432 * \text{wickets\_fallen} + 246.434 \quad \text{(vi)}$$

In equation (vi), Score is the projected score by the end of the innings, current_score is the current score and wickets_fallen is the number of wickets fallen of the batting team in the first innings between 0-5 overs and the corresponding constant values are the weights (refer equation (i)). Similarly, eight more such equations have been generated through Linear regression classifier for the remaining 5 over intervals (which are 5-10, 10-15,……..40-45 overs).This way for each team and for each venue nine equations each have been generated from the datasets to predict the first innings score using equation (i).

2) Naïve Bayes: For each five over intervals for a particular team and venue in the second innings, the values for posterior(yes) and posterior(no), with 'yes' representing the matches won and 'no' representing the matches lost have been calculated. Thus, the following calculations are done:

For posterior (yes), p (current score|yes), p(wickets fallen|yes), p(target|yes) has been calculated through equation (v).Then,

$$\text{posterior(yes)} = \frac{\text{p(current score|yes)}.\text{p(wickets fallen|yes)}.\text{p(target|yes)}.\text{p(yes)}}{\text{evidence}} \quad \text{(vii)}$$

by using (iii) and (iv), the equation (vii) has been formed.

Similarly, for posterior (no), p(runs|no), p(wickets|no), p(target|no) have been calculated through equation (v) and then using equation (iii) and (iv), equation (viii) has been formed.

$$\text{posterior(no)} = \frac{\text{p(runs|no)}.\text{p(wickets|no)}.\text{p(target|no)}.\text{p(no)}}{\text{evidence}} \quad \text{(viii)}$$

Where, evidence is given by equation (ix) from equation (iii).

By substituting the values in equation (vii) and (viii), posterior (yes) and posterior (no) respectively are calculated and the one with the maximum value will be the most probable. Since the model is proposed to find the winning percentage or probability. So, posterior (yes) has been considered in the observations.

3) Projected Score Performance: The ODI matches of Australia played in 2014 have been taken for testing and the first innings projected score has been compared which is given by Linear Regression (LR) classifier with the projected score given by the Current Run Rate (CRR), in each 5 over period up to 45 overs (since only 45 overs out of 50 overs have been considered for predicting the final score of the innings). Then the Error by Linear Regression (Error LR) and Error by Current Run Rate (Error CRR) have been calculated which are given by the equations (x) and (xi) respectively. The value of n given as, n=5, 10, 15, ...45 and |a| means absolute value of a.

$$\text{Error LR} = |(\text{Final Score of the match} - \text{Projected Score by LR at the end of nth over})| \quad \text{(x)}$$

$$\text{Error CRR} = |(\text{Final Score of the match} - \text{Projected Score by CRR at the end of nth over})| \quad \text{(xi)}$$

Following Line Charts, refer Fig. 1, have been obtained for various 5 overs period of the first innings with the X- axis representing the number of matches played by Australia in 2014 and Y-axis is the absolute error. Out of the two lines, orange line is representing the Error CRR and the blue line (dotted line) is for Error LR.

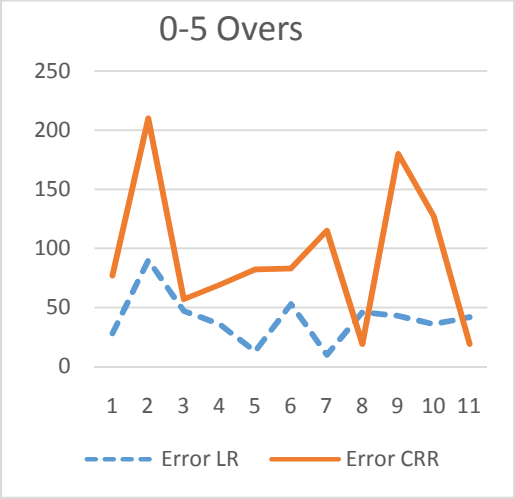

Fig. 1.1. Accuracy of 0-5 overs.

$$\text{evidence} = p(\text{runs}|\text{yes}).p(\text{wickets}|\text{yes}).p(\text{target}|\text{yes}).p(\text{yes}) + p(\text{runs}|\text{no}).p(\text{wickets}|\text{no}).p(\text{target}|\text{no}).p(\text{no}) \quad \text{(ix)}$$

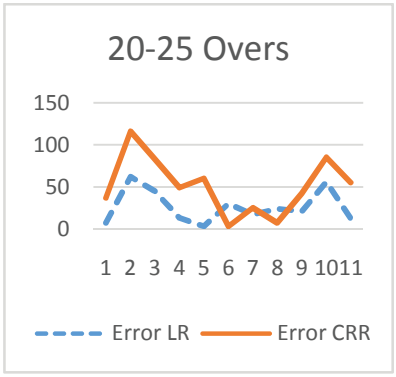

Fig. 1.2. Accuracy of 20-25 overs.

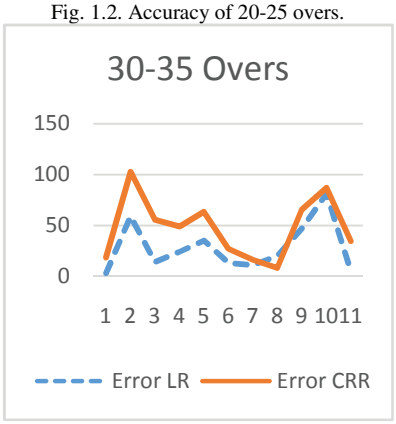

Fig. 1.3. Accuracy of 30-35 overs.

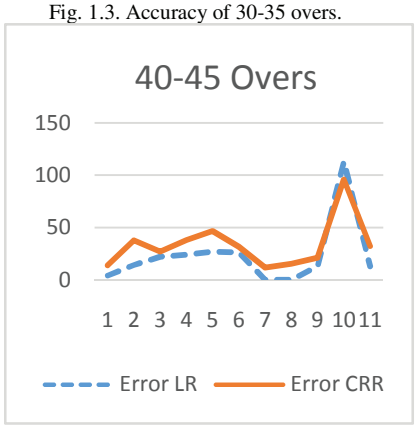

Fig. 1.4. Accuracy of 40-45 overs.

Fig. 1.A comparison between the Error LR and Error CRR of the matches played by Australia in 2014 at various 5 over interval.

It can be observed from Fig. 1 that in almost every match at each chart, Error LR is less than the Error CRR. So, this shows that the projected score calculated by Linear Regression classifier is more accurate than by the Current Run Rate. Thus, LR classifier tends to give better estimation for the final score at any situation of the match.

B. *Winning Percentage Performance*

India's ODI matches in which India have batted in the second innings have been taken. Then the accuracy given by the Naïve Bayes classifier has been calculated which is actually the accuracy in identifying the final outcome of the match at 5 overs interval and Table II shows the accuracy of the Naïve Bayes model corresponding to each 5 over period for the ODI matches played by India.

Table 2: Accuracy of the Naïve Bayes Classifier at different range of overs for the matches played by India.

Overs	Accuracy (%)
0-5	68.6813
5-10	75.8242
10-15	75.8242
15-20	79.6703
20-25	80.2198
25-30	81.3187
30-35	80.2198
35-40	87.3626
40-45	91.2088

Fig. 2 depicts line chart that have been generated by plotting the data of Table II with X-axis taken as overs and Y-axis as the accuracy in terms of percentage.

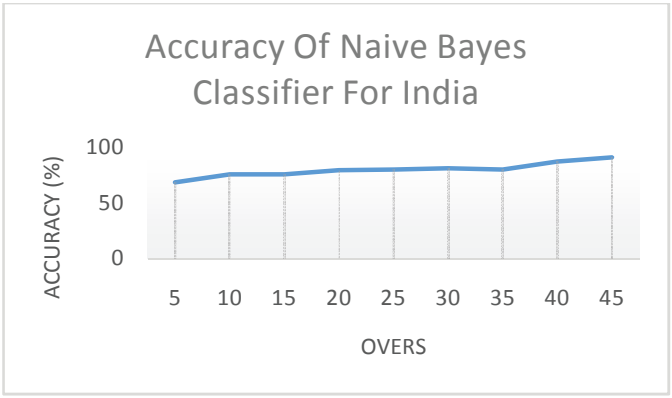

Fig. 2. Accuracy of the Naïve Bayes Classifier at different range of overs for matches played by India.

It can be seen from Fig. 2 that the accuracy tends to increase as the number of overs increases. For 0-5 overs accuracy is 68%, which is the lowest of all the intervals. This is because at the start of the second innings it is difficult to assess the match situation. Especially for the game like cricket which is highly unpredictable, if accuracy close to 70% is there at the starting and goes on to 91% at the 45 over mark then one can rely upon this method.

VII. CONCLUSION AND FUTURE SCOPE

The main purpose of this paper is to make a model for predicting the final score of the first innings and estimating the outcome of the match in the second innings for the ODIs. Two separate models, one for the first innings and other for the second innings using the Linear Regression classifier and Naïve Bayes classifier respectively on the past ODI matches have been proposed. The error of Linear Regression classifier predictions with the current method for projecting the score by analysing the real ODI cricket data were compared. It was observed that the error in Linear Regression Classifier is less than the Current Run Rate method in predicting the final score at any situation of the match. Also the accuracy of the Naïve Bayes for predicting the match outcome, goes from 70% (initially) to 91% as the match progresses.In future, the focus will be to improve the accuracy for both the models. Furthermore, the other factors like the toss, the ODI ranking of the teams and the home team advantage will be considered in the predictions.

REFERENCES

[1] Khabir Uddin Mughal. Top 10 Most Popular Sports In The World. http://sporteology.com/top-10-popular-sports-world/Accessed 2 February 2015.

[2] Laws of cricket. http://www.lords.org/mcc/laws-of-cricket/Accessed 2 January 2015.

[3] Machine Learning Group at the University of Waikato. Weka 3: Data Mining Software in Java. http://www.cs.waikato.ac.nz/ml/wekaAccessed 12 February 2015.

[4] NarasimhaMurty, M.; Susheela Devi, V. (2011). Pattern Recognition: An Algorithmic Approach.

[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):121{125,1997.

[6] D. Lutz. A cluster analysis of NBA players. In MITSloan Sports Analytics Conference, 2012.

[7] S. Luckner, J. Schroder, and C. Slamka. On the forecast accuracy of sports prediction markets. In Negotiation, Auctions, and Market Engineering, International Seminar, Dagstuhl Castle, volume 2, pages 227{234,2008.

[8] G. Gartheeban and J. Guttag. A data-driven method for in-game decision making in mlb: when to pull a starting pitcher. In Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining, KDD '13, pages 973{979, New York, NY, USA, 2013. ACM.

[9] Seamus Hogan (2012) Cricket and the Wasp: Shameless self promotion (Wonkish). http://offsettingbehaviour.blogspot.co.nz/2012/11/cricket-and-wasp-shameless-self.htmlAccessed 2 March 2015.

[10] K. Raj and P. Padma. Application of association rule mining: A case study on team India. In International Conference on Computer Communication and Informatics (ICCCI), pages 1{6, 2013.

[11] T. B. Swartz, P. S. Gill, and S. Muthukumarana. Modelling and simulation for one-day cricket. Canadian Journal of Statistics, 37(2):143{160, 2009.

[12] A. Kaluarachchi and A. Varde. CricAI: A classification based tool to predict the outcome in ODI cricket.In 5th International Conference on Information and Automation for Sustainability, pages 250{255, 2010.

[13] NarasimhaMurty, M.; Susheela Devi, V. (2011). Pattern Recognition: An Algorithmic Approach.

[14] Naive Bayes Classifier. http://www.ic.unicamp.br/~rocha/teaching/2011s2/mc906/aulas/naive-bayes-classifier.pdfAccessed 12 March 2015.

[15] Milansolanki (2012) Naive Bayes Classifier http://www.codeproject.com/Articles/318126/Naive-Bayes-ClassifierAccessed 20 March 2015.
[16] Ian H. Witten, Eibe Frank, Elsevier, Data Mining Practical Machine Learning Tools (2005).
[17] Dr.SaedSayad (2011) Classification. In: Real Time Data Mining. http://www.saedsayad.com/naive_bayesian.htmAccessed 10 March 2015.