Cricket Sabermetrics

A Data Mining Analysis of Cricket

P.A. Gregory, D.H.M.S.N. Herath, D.S.L. Karunasekara, S. Deegalla, A. Bandaranayake
Department of Computer Engineering
Faculty of Engineering, University of Peradeniya
Peradeniya 20400 Sri Lanka

***Abstract*— Where there is sports there is statistics and cricket is no exception to this. The game of cricket has a wide wealth of complex statistical data associated with the game. Analysts and fans throughout history have crunched the numbers, trying to make meaning out of this vast pool of data. The aim of this project is to attempt the same using data mining techniques to hopefully unearth key underlying data patterns that can be used to accurately model in-game performances.**

Keywords—Cricket; Data Mining; Statistics; Indian Premier League;

I. INTRODUCTION

Whether it is the sprinter who finished first or the team that scored more points, it's usually easy to determine who won a sporting event. But finding the statistics that explain why an athlete or a team wins is more difficult and major figures at the intersection of sports and numbers are determined to crack this problem.

One such attempt to make meaning out of statistics was seen in baseball. The aim was to objectively analyze the game of baseball by deviating from traditional performance measures. This approach now commonly known as Sabermetrics, often questioned traditional measures of baseball as they did not provide an accurate representation of in game performance [1]. Instead, Sabermetric researches used statistical analysis to determine performance metrics that actually contribute towards a team win.

Sabermetrics has transformed baseball. The question is, can it do the same for cricket? Cricket is a far more diverse sport spread across three major formats and played across the globe. But like baseball it has a wealth of statistics associated with the game. These traditional statistical measures of cricket are well established performance metrics commonly used to evaluate player performances. But like traditional baseball statistics, we believe they do not provide an accurate representation of in game performance.

This project aims to explore the statistical aspect of cricket in a selected domain, and realize key performance metrics that contribute towards the outcome of a cricket match.

II. LITERATURE REVIEW

Player classification using performance metrics always topped the priority list of researches irrespective of the sport under consideration. However in cricket, there hasn't been extensive research on performance-based player classification apart from the traditional measures such as averages, strike rates and economy rates. These existing player evaluation metrics in cricket are believed to be fundamentally flawed [2]. Alternative performance measure have been proposed, an example being a classification scheme developed for batmen using performance data of One Day International (ODI) matches and Test cricket [3][4]. The proposed method uses a single measure derived from a batsman's average, strike rate and batting consistency to evaluate performances.

The latest addition to cricket is the Twenty-20 format. The fast paced game of T20 has given rise to many lucrative domestic T20 competitions, such as the Indian Premier League (IPL). The IPL has emerged as a focal point for many different disciplines, from Economics and Finance to Statistics and Decision Science. With wide spread growth of T20 cricket, many new attempts to model player performances have been made. Reference [5] discusses the performance of players in the first T20 World Cup. All-rounder performance from the first edition of the IPL was evaluated using the terms batting all-rounder, bowling all-rounder and ideal all-rounder [6]. Even fantasy cricket has had its own share of player performance modelling where existing measures of player performances were used in a binary integer programming model to select players [7]. The idea is to use both static prediction and dynamic prediction to develop a dynamic prediction algorithm based on an aggregate of static predictions.

In 2010 Venky Mysore, was bought on board the IPL team Kolkata Knight Riders to aid the selection process during the player auction. Influenced by MoneyBall: The Art of Winning an Unfair Game, his strategy was to buy wins rather than players [8]. With various predictive analysis capabilities, KKR put together the winning team of IPL 2014 [9].

WASP - "Winning And Score Predictor" is a calculation tool used in cricket to predict scores and possible results of a limited overs match [10]. Cricket Australia's new statistical approach and "Key to Success" statistical analysis method are few examples of them. Since these examples are commercial products, the methods they used in developing these applications are not a matter of public record.

The main challenge lies in identifying performance metrics that actually matter. Past attempts at solving the same, have relied on expert domain knowledge which can introduce a potential bias to the final results. To eliminate this, our

approach uses data mining and machine learning techniques to identify patterns in data that could potentially provide us with an indication on key performance metrics that could be used for player evaluation.

III. METHODOLOGY

A. The Domain

As stated previously, cricket is a diverse sport. It is played globally across three different formats and as a result analyzing the game as a single entity is a difficult if not impossible task to accomplish. Therefore, the domain of choice for our problem is the Indian Premier League (IPL). The IPL being eight editions old, provides us with a decent sample set of data to perform our analysis.

B. Data Set

An exhaustive set of up-to-date statistical data for the IPL domain was obtained. The dataset contained complete statistical details of 501 instances of IPL matches. The data was parsed and stored in a database using an object to relational mapping framework. Classes were created following Object Oriented principles to manipulate the data as needed during feature construction.

C. Methodology Outline

An iterative approach was followed for the analysis. The analysis looped between developing the feature set and improving the data mining model. Two different approaches were used to improve the overall accuracy of the analysis. An outline of the approach is given below.

- Feature Set Development
- Feature Selection
- Modelling and Analysis

D. Feature Set

A main component of the analysis is the feature set. The analysis is only as good as the feature set that's fed into the model. Potentially any attribute that can be associated with a team innings can be used in the feature set. Initially a basic set of attributes was used, and as the analysis progressed a more complex set of features were developed to improve the accuracy of the model. Each attribute was built using various combinations of the basic units of information for an innings, namely runs, balls and wickets.

- Number of Wickets Lost (1)
- Four Hitting Frequency (2)
- Six Hitting Frequency (3)
- Boundary Run Percentage (4)
- Dot Ball Percentage (5)
- Dot Ball to Runs Ratio (6)
- Run Rate (7)
- Average Partnership Score (8)
- Number of Batting Segments (9)
- Batting Segment to Wicket Ratio (10)
- Average Runs in a Batting Segment (11)
- Average Pressure Factor (12)
- Pressure of Wickets (13)
- Final Score (14)

The attribute values were calculated for each match across the entire IPL domain. The resulting set of data contained 501 instances of the above attribute set (i.e.: for every match played in the IPL).

E. Feature Selection and Modelling Analysis

The idea behind the analysis was to predict the outcome of a match based on the feature set. The reasoning behind this was, if the prediction accuracy is high, the input feature set can be recognized as an accurate representation of in-game performance metrics.

Feature selection can be carried out using either Filter methods or Wrapper methods. The former is independent of a classifier and ranks the features according to a specific mathematical model. The latter on the other hand, generates subsets of the feature set and calculates the accuracies of each subset against a classification algorithm. Due to computational costs and almost similar results from both approaches, our analysis was carried out using Filter methods.

Three different attribute selection algorithms were tested on the feature set. The resulting subset of features were then fed into a classification model running the J48 decision tree algorithm using ten-fold cross validation. To improve the accuracy various attribute combinations of a given subset were also tested. The subset that provided the highest accuracy was selected as the ideal subset of attributes.

F. Innings Segmentation

Up until now, a single innings was treated as one complete segment. While this allowed us to identify feature impact throughout an innings, it did not provide us with information on the impact at different stages of an innings.

To address this, we proceeded to segment the innings into three main segments:

- Powerplay (1-6 overs)
- Middle (7-15 overs)
- Death (16- 20 overs)

The complete feature set was then calculated for each segment separately. That is, for each of the 501 instances in the dataset, the feature set was evaluated for the three segments. Initially, the individual predictive accuracy of each feature was recorded for each segment. (i.e.: Using a single attribute at a time on the classification model). The feature selection process described under subsection E was then

carried out to identify the optimal subset of features for each segment.

Finally, a combination of all the features (segments and complete innings) were analyzed using Wrapper methods.

IV. RESULTS

A. Feature Selection for Complete Innings

The classification model for the analysis used the J48 decision tree classifier. The feature selection process was carried out using three attribute selection algorithms.

- *CfsSubsetEval* – Selects attributes with high correlation with the class and low inter-correlation
- *InfoGainAttributeEval* – Selects attributes ranked according to information gain
- *ReliefFAttributeEval* – Selects attributes by repeated sampling

The results of feature selection carried out for the complete innings is shown in Table I and Table II. The accuracy improvements are shown in Table III and Table IV.

TABLE I. FIRST INNINGS FEATURE SELECTION

Algorithm	Optimum Feature Subset	Accuracy (%)
CfsSubsetEval	8,12,5,6,2	70.4591
InfoGainAttributeEval	6,7	70.2595
ReliefFAttributeEval	5,6,1	70.4591

TABLE II. SECOND INNINGS FEATURE SELECTION

Algorithm	Optimum Feature Subset	Accuracy (%)
CfsSubsetEval	11, 8, 1	88.4232
InfoGainAttributeEval	13	88.8224
ReliefFAttributeEval	8,10,9,1,13	88.0240

TABLE III. FIRST INNINGS PREDICTIVE ACCURACY

Attribute Selection	Accuracy(max)
Without feature selection	67.6647
With feature selection	70.4591

TABLE IV. SECOND INNINGS PREDICTIVE ACCURACY

Attribute Selection	Accuracy(max)
Without feature selection	86.0279
With feature selection	88.8224

Optimum subset of attributes:

First Innings: Dot Ball to Runs Ratio, Dot Ball Percentage, Number of Wickets Lost
Second Innings: Pressure of Losing Wickets

B. Innings Segmentation

Innings segmentation was done to analyze the feature impact at different stages of the match. This was carried out for the first innings. The innings was divided into three segments, namely Powerplay, Middle and Death. Table V shows the highest individual predictive accuracy of each attribute across all segments.

TABLE V. MAXIMUM PREDICTION ACCURACY OF INDIVIDUAL FEATURES

Attribute	Segment with Highest Accuracy	Accuracy (%)
Number of Wickets Lost	Powerplay	60.4790
Four Hitting Frequency	Complete	60.8782
Six Hitting Frequency	Complete	65.2695
Boundary Run Percentage	Complete	53.6926
Dot Ball Percentage	Complete	61.2774
Dot Ball to Runs Ratio	Complete	69.0619
Run Rate	Complete	71.0579
Average Partnership Score	Complete	64.0719
Number of Batting Segments	Complete	53.6926
Bat Segment to Wicket Ratio	Powerplay	60.0798
Avg Runs in Bat Segment	Complete	66.6667
Avg Pressure Factor	Complete	67.6647
Pressure of Wickets	Powerplay	58.6826
Final Score	Complete	71.4571

The optimum subset of features for each segment was then evaluated. This was done using both Filter methods and Wrapper methods with the J48 classification algorithm. The results are shown in Table VI.

TABLE VI. SEGMENT FEATURE SELECTION

Segment	Optimum Feature Subset	Accuracy (%)
Powerplay	1	60.4790
Middle	5,7	64.0719
Death	1,4,5,6,12	65.4691

The combined arff file containing all features (segment + complete) was analyzed using *WrapperSubsetEval*, a selection algorithm that generates random subsets and tests the accuracy against a classification algorithm returning the subset that gave the maximum accuracy. The results are shown in Table VI.

TABLE VII. COMBINED FEATURE SELECTION

Segment	Optimum Feature Subset
Complete	2,9,14
Powerplay	10,3
Middle	4,11,12,5
Death	-

The above subset of features showed a prediction accuracy of 71.6567 using the J48 decision tree algorithm.

Optimum subset of attributes:

First Innings: Four Frequency, Number of Batting Segments, Final Score, Batting Segment to Wicket Ratio (PP), Six Hitting Frequency (PP), Boundary Run Percentage (Middle), Average Runs in Batting Segment (Middle), Average Pressure Factor (Middle), Dot Ball Percentage (Middle), Run Rate (Middle)

I. CONCLUSION AND FUTURE WORK

The initial analysis provided some important distinctions between the first and second innings of a match. The feature selection process identified a different subset of attributes for the two innings. It can be concluded that there are different dynamics to a chase. This is obvious when the resultant subset

of features are analyzed. While wicket loss plays a major part in the result of a chase, it is not as important while setting targets. Similarly, dot ball percentages seem to play a bigger role during the first phase of the match when compared to the second.

The segmentation process resulted in a different subset of features for each segment. For example, according to the above analysis, the most important feature during the powerplays is wickets lost. For the middle overs, dot balls and runs scored has a greater impact. During the death overs though, there is no clear feature that stands out. Rather a combination of different features seem to decide the outcome of a match. This proves that the importance of an attribute varies throughout an innings.

Finally, the combination of features from all segments provided us with the highest predictive accuracy. The combined subset of features contains attributes from the complete analysis as well as the segmentation analysis.

This analysis is by no means complete at this point. Future work would involve exploring new options to improve the model accuracy. One such option is using ensembling, a process that attempts to increase the predictive accuracy of a model by combining different models together. Attention should also be paid to the feature set, since the model is only as good as the feature set. Therefore, future development work will follow an iterative approach between developing the feature set and improving the model.

REFERENCES

[1] Wikipedia, Sabermetrics (2016, August) [Online]. Available: https://en.wikipedia.org/wiki/Sabermetrics

[2] Lewis, A J (2005). "Towards Fairer Measures of Player Performance in One-Day Cricket," The Journal of the Operational Research Society, 56(7), 804- 815.

[3] Lemmer, H (2004). "A Measure for the Batting Performance of Cricket Players," South African Journal for Research in Sport, Physical Education and Recreation, 26(1), 55-64.

[4] [4] Lemmer, H (2006). "A Measure of the Current Bowling Performance in Cricket," South African Journal for Research in Sport, Physical Education and Recreation, 28(2), 91-103.

[5] Lemmer, H (2008). "An Analysis of Players' Performances in the First Cricket Twenty20 World Cup Series," South African Journal for Research in Sport, Physical Education and Recreation, 30(2), 71-77.

[6] Van Staden, P J (2008). "Comparison of Bowlers, Batsmen and All-rounders in Cricket Using Graphical Display," Technical Report 08/01, Department of Statistics, University of Pretoria, South Africa.

[7] Brettenny, W (2010). Integer Optimization for the Selection of a Fantasy League Cricket Team, Unpublished M.Sc Dissertation, Faculty of Science, Nelson Mandela Metropolitan University, South Africa.

[8] How KKR got it right – starsports.com. (2016, August). [Online]. Available: http:/www.starsports.com/cricket/articles/article=how-kkr-got-right/index.html

[9] How SAP Helped KKR Win Pepsi IPL 2014 | cioandleader.com. (2016). [online] Cioandleader.com. Available at: http://www.cioandleader.com/articles/40155/how-saphelped-kkr-win-pepsi-ipl-2014 [Accessed 1 Apr. 2016].

[10] WASP (cricket calculation tool). (2016). [online] Wikipedia. Available at: https://en.wikipedia.org/wiki/WASP_(cricket_calculation_tool) [Accessed 1 Apr. 2016].