CRICKET SABREMETRICS: A DATA MINING ANALYSIS OF CRICKET

P.A. Gregory, D.H.M.S.N. Herath, D.S.L. Karunasekera, D.S. Deegalla* and A.U. Bandaranayake

Department of Computer Engineering, Faculty of Engineering, University of Peradeniya

**Corresponding author (email: dsdeegalla@pdn.ac.lk)*

Introduction

Win or loss of a sporting event can be determined by the athlete who finished first or the team who scored the most points. However, finding the statistics that explain why the athlete or the team wins is more difficult. One such attempt to make meaning out of statistics was seen in baseball. 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 used statistical analysis to determine performance metrics that contributed towards a team win. Sabermetrics has transformed baseball; can it do the same for cricket? Cricket, unlike baseball, is a far more diverse sport spread across three major formats and played across the globe. However, 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.

Player classification using performance metrics always topped the priority list of researchers 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 measures have been proposed; an example a classification scheme developed for batsmen using performance data of One Day International (ODI) matches and Test cricket [3]. 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 widespread growth of T20 cricket, many new attempts to model player performances have been made [4]. The main challenge lies in identifying performance metrics that actually matter. Past attempts at solving this, 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. This work 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.

Materials and Methods

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.

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.

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

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. The final set of attributes used in our analysis is given below. Each attribute was built using various combinations of the basic units of information for an innings, namely, runs, balls and wickets.

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

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.

Feature Selection and Modelling Analysis

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

Feature selection was tested using Filter methods or Wrapper methods [5]. The former lacks 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 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 was 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.

Innings Segmentation

While treating a single innings as one complete segment 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. Thus, we divided 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. Initially, the individual predictive accuracy of each feature was recorded for each segment. The feature selection process 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) was analyzed using Wrapper methods.

Results and Discussion

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 for the complete innings are shown in Table 1 and Table 2.

Table 1. First Innings Feature Selection

Algorithm	Optimum Feature Subset	Accuracy (%)
CfsSubsetEval	8,12,5,6,2	70.45
InfoGainAttributeEval	6,7	70.25
ReliefFAttributeEval	5,6,1	70.45

Table 2. Second Innings Feature Selection

Algorithm	Optimum Feature Subset	Accuracy (%)
CfsSubsetEval	11, 8, 1	88.42
InfoGainAttributeEval	13	88.82
ReliefFAttributeEval	8,10,9,1,13	88.02

The accuracy improvements are shown in Table 3 and Table 4.

Table 3. First Innings predictive accuracy

Algorithm	Accuracy (max)
Without feature selection	67.66
With feature selection	70.46

Table 4. Second Innings predictive accuracy

Algorithm	Accuracy (max)
Without feature selection	86.03
With feature selection	88.82

Optimum subset of attributes

The optimum subset of attributes for each inning when the whole innings is considered was identified as follows:

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

Second Innings: Pressure of Losing Wickets

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 5 shows the highest individual predictive accuracy of each attribute across all segments.

Table 5. Maximum prediction accuracy of individual features

Attribute	Segment with Highest Accuracy	Accuracy (%)
Number of Wickets Lost	Powerplay	60.47
Four Hitting Frequency	Complete	60.87
Six Hitting Frequency	Complete	65.26
Boundary Run Percentage	Complete	53.69
Dot Ball Percentage	Complete	61.27
Dot Ball to Runs Ratio	Complete	69.06
Run Rate	Complete	71.05
Average Partnership Score	Complete	64.07
Number of Batting Segments	Complete	53.69
Bat Segment to Wicket Ratio	Powerplay	60.07
Avg Runs in Bat Segment	Complete	66.66
Avg Pressure Factor	Complete	67.66
Pressure of Wickets	Powerplay	58.68
Final Score	Complete	71.45

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 6.

Table 6. Segment feature selection

Segment	Optimum Feature Subset	Accuracy (%)
Powerplay	1	60.47
Middle	5,7	64.07
Death	1,4,5,6,12	65.46

The combined file containing all features (segment + complete) was analyzed using *WrapperSubsetEval*, a selection algorithm that generates random subsets and tests the accuracy of a classification algorithm returning the subset that gave the maximum accuracy. The results are shown in Table 7.

Table 7. 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 71.65 using the J48 decision tree algorithm.

An 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)

Conclusions and Recommendations

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 in a chase. This is obvious when the resultant subset of features is 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 seems to decide the outcome of a match. This behavior 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 for developing the feature set and improving the model.

References

[1] J. Albert. “Sabermetrics: The past, the present, and the future.” Mathematics and sports, 2010, Feb. 12 (43), 15.

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

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

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

[5] I. Guyon, A. Elisseeff. An introduction to variable and feature selection. Journal of machine learning research, 2003, 3(Mar), 1157-82.