Performance Analysis of a Selected Cricket Team Using Data Mining and Machine Learning

Nilucshan Siva, Ashan Perera, Gayan Thejawansha,
Asitha Bandaranayake and Sampath Deegalla

Department of Computer Engineering, Faculty of Engineering, University of Peradeniya, Peradeniya 20400, Sri Lanka

***Abstract*— Cricket is one of the most popular game and contains large amount of statistical data associated with it. National teams and franchise club owners continuously tend to improve their teams' performance due to its high popularity and competitiveness. Statistical models along with advanced data mining techniques are applied based on many aspects in order to improve the performance of a team, predict end results of a match and efficiently handling constraints such as game rules and budget. It is critical to identify the right set of attributes which impose a high impact on performance of a team. This study attempts to analyze performance of a selected cricket team using data mining and machine learning. Several works have been carried out to address this problem. Majority of the works were considering Indian Premier League (IPL) datasets due to its high popularity and financial value. Since IPL consists of twenty overs, teams and players often come up with different strategies from their usual ones. Players of an IPL team varies dramatically in each season and therefore it is critical to identify attributes behind a successful team. This work considers all forms of cricket for the analysis. Sri Lankan cricket team was selected for the performance analysis with match details from 2006 to 2017. Analysis was divided into two categories, team wise performance evaluation and player wise performance evaluation. In team wise evaluation, selected team was analyzed with past data without considering its players. Analysis was performed on ten commonly available attributes in past data. Player wise evaluation was based on international twenty twenty and IPL matches since these datasets are rich in player statistics. Player wise evaluation was further divided into two categories. One part of the player wise evaluation was dealing with defining effective performance metrics to evaluate batting, bowling and fielding abilities of a player and ranking best players. Other part was analyzing past data of players and identifying patterns and trends over the years. Identifying the optimal batting, bowling position was the ultimate goal of the latter. Different machine learning algorithms were applied, and Naïve Bayes and Support Vector Machine (SVM) were giving promising results. In conclusion, this study discuses on team wise evaluation and player wise evaluation and their collaborative impact on a cricket team.**

Keywords– Cricket, Attributes, IPL, Data Mining, Machine Learning, Naïve Bayes, Support Vector Machine

I. INTRODUCTION

Data mining is the process of analyzing large amount of data to extract data patterns and relationships. Bayes' theorem and regression analysis were used in early times to identify patterns in data. Dramatic increase in data collection, storage and manipulation resulted with the invention of computer technologies. Modern data mining consists of advanced methods such as genetic algorithms, decision tress, decision rules, support vector machines, cluster analysis and neural networks. Extracted data patterns are often used to predict future outcomes. Data mining is used in almost all fields. Unlike other fields, sports have been keeping records for centuries to performance evaluation and official requirements. Since large amount of data is available, many researches and experiments have been carried out throughout the past considering many aspects. Cricket is one of the most popular sports in the world. It has a huge fan base being only second to Soccer. Cricket is very popular in England, Australia, New Zealand, West Indies, South Africa, Zimbabwe and Asian countries. Heavy engagement of cricket fans in recent times gave rise to league matches and franchise clubs. Countries and franchise club owners are required to find methods which make them to dominate in this highly competitive sport. Teams come up with many data analysis approaches to increase their winning possibility. However most of these analysis approaches are commercial products and hence not exposed to public knowledge.

A. Background

In the 1980s, work by Bill James on statistical analysis of Baseball named Sabermetrics gained a major popularity. Sabermetrics is study of the game baseball through scientific and objective analysis. The term was derived from the acronym SABR which stands for Society of American Baseball Research. Recognition of

this work led to a revolution in the game of Baseball. Oakland Athletics baseball team went on to set the record for twenty consecutive wins in the 2002 season using Sabermetrics. Many teams and clubs adapted this approach in the following years and got succeeded. Many sports including cricket tend to follow similar approaches in their games after witnessing the transformation of game baseball through sabermetrics and its positive outcomes. Kolkata Knight Riders, a franchise club in Indian Premier League Cricket series was able to win the season in 2012 and 2014 by using predictive analysis and techniques similar to sabermetrics.

Having more data often helps to come up with a good modelling of the game and effectively apply data mining techniques. Sports have been keeping records for many years. This makes sports a wealthy domain in machine learning. Even though number of attributes are recorded it is hard to find relationships among attributes with manual observations. The interesting fact about data mining is that it reveals interesting patterns in data which are not visible otherwise. As a highly competitive sport, application of data mining in cricket is highly inevitable.

Player evaluation and outcome prediction are the two major areas of data mining in sports. Introduction of franchise clubs in cricket resulted in auction based selection of players. Therefore team owners are required to make optimal decisions considering their budget limit. Modern data mining algorithms are used to evaluate batting, bowling and fielding performances much more effectively than traditional measures. Often traditional metrics such as batting average, strike rate and economy rate along with new performance metrics are used to evaluate players. Defining new performance metrics is a crucial task and requires domain knowledge expertise.

B. The problem

If the use of Sabermetrics has transformed a statistically heavy sport like baseball, could it possibly do the same for cricket? Being rich in data make Cricket to be modeled in a variety of ways. Many models have been created considering various aspects such as prediction of the winner, prediction of possible outcomes and evaluation of players. However, it is not always easy to determine which model performs better than others and which evaluation metrics are accurate compared with others. Also it is crucial to validate prediction results in real match. For example, if some model propose 11 players for a team to get its optimal performance, we will not be able to find the truthfulness of this prediction unless that exact 11 players are played in a match.

Over the last few years, there have been many efforts in applying data mining approaches in Cricket. Most of these works heavily depend on their domain knowledge of cricket. Depending on available data, it could be completely random or there could be interesting statistical patterns which could explain the outcomes of a match.

Aim of this work is to analyze data as much as possible and identify patterns and relations, which can be used to efficiently model the game while minimizing domain knowledge.

C. The proposed solution

Performing analysis without considering domain knowledge can lead to two kind of outcomes. One, there could be hidden patterns which could be used to model the game and two being yielding not useful relations between attributes. However, it should be noted that relations which seems to be not useful in the aspect of cricket domain can be useful to researchers, data analysts, business analysts and could also possibly help to create better models of the game in near future. Therefore, this research was carried out in two categories. One being solely based on random analysis while other being based on usage of domain knowledge such as evaluation metrics and intuitive reasoning.

II. Background Study

A. Measuring Player Performance

Craig [1] proposes metrics to evaluate batting, bowling and fielding performance of a player. Proposed metrics overcome drawbacks in traditional metrics present to evaluate general measures such as batting average, strike rate, economy rate and further helps to evaluate fielding performance. It should be noted that there are no standardized methods to evaluate fielding performance of a player. Apart from these general measures, his work introduces weighted summation of attributes to evaluate consistency, form and performance against a particular opposition and venue. Batting performance is measured by defining batting score, home run hitting ability (ability to hit fours and sixes) and milestone reaching ability (ability to score fifty or hundred runs) metrics. Batting performance is also measured by defining similar bowler specific metrics like bowling score, run prevention rate and ability to get four or more wickets. Fielding performance is evaluated considering catches, stumping, run outs and assists in getting a wicket.

V. V. Sankaranarayanan et al. [2] evaluates batsmen using batting average, strike rate, home run hitting ability and milestone reaching ability. By using these four measures, batsmen are divided into five clusters namely opening batsmen, middle order batsmen, all-rounders, wicket keeper and tail-enders. These metrics were used to predict the runs a batsmen will score in an ongoing match. They applied data mining techniques to find the remainder of a game when its instantaneous state is provided. They consider toss outcome, current overs and wickets and game segment (game is divided

into 10 segments) along with batting evaluation to predict the end results. Data analyzed from the past data and instantaneous data was used to model the remaining state of the game. Rigger regression and Attribute bagging algorithms were used on the selected features for the prediction. Their work does not include bowling and fielding performance for the analysis.

Works of K. Passi et al. [3] introduces a model which can be used to select players for a team. They defined metrics to evaluate batting and bowling performance of players. An analyzation score was derived applying AHP (Analytic Hierarchy Process) to combination of selected player attributes. Players were categorized into five levels and these levels are used to select players for a team.

Shubham Agarwal et al. [4] propose a model where performance metrics were measured by giving weight considering opposition team, location, year, overall statistics and last five innings performances. A 50% of weight is given for last five innings since it is a good way to measure the form of a player. Players are ranked based on an evaluation score.

An Evaluation based on Data Envelopment Analysis (DEA) was proposed by Gholam R. Amin et al. [5]. For a batsman, highest individual score, average batting performance, strike rate, number of fours and number of sixes were defined as the outputs of DEA. Through DEA it was possible to simultaneously measure the performance of a batman based in these five attributes. 60 batsman from IPL season 4 in 2011 were selected and ranked based on their DEA score. High DEA score depicts high batting performance while low DEA depicts need for improvements. A bowler was evaluated using economy rate, bowling average and bowling strike rate along with number of wickets taken. Inverse values of average, strike rate and economy are selected as the outputs of DEA since having minimum value of these metrics are considered to be good. Bowlers are ranked based on their DEA score.

Faez Ahmed et al. [6] applied multi objective optimization and decision making approaches for selecting a team. They defined objective functions to evaluate players based on batting, bowling and fielding skills. They have chosen batting and bowling performances as the main functional criteria in the initial multi objective optimization study. Subjective criteria have been added during the subsequent decision making phase. NSGA – II (Non-dominated Sorted Genetic Algorithm) is applied for the multi objective optimization process. Their research resulted in a team which was having more performance than the winning champion of IPL season at that time.

B. Identifying Important Attributes

P. A. Gregory et al. [7] analyzed past IPL matches and identified the features which have high impact on end results of a cricket match. They divided the match into three segments namely power play, middle and death overs and analyzed the impact of chosen feature sets. 14 general features were selected which can possibly have impacts on end results. Impact level of chosen features were identified in each segment of the match. Some of the features were impacting all three segments whereas some resulted only in few.

Another research in a similar motive was done by Pranavan Somaskandhan et al [8]. They selected 23 attributes which might affect the end results and found their relative importance. Selected attributes were general features similar to above mentioned work. They concluded that nine attributes were having high relative importance than others toward the end results of a match.

III. METHODOLOGY

Selected cricket team was evaluated under two major categories, team wise evaluation and player wise evaluation. Team wise evaluation considers a team as a single unit regardless of its players. This way of modelling can be effective since each and every player's collaborative effort defines the overall performance of a team rather than few high profile players and it can reveal interesting patterns which a certain team showcase over time. It is obvious that optimal selection of team members results in victory of the team. Therefore it is essential to evaluate players in a team.

A. Team wise evaluation

There were 556 matches played by Sri Lankan Cricket team from 2006 to 2017. Size of the data set was reasonably enough for a decent prediction model. Initially the problem was modeled as a supervised learning problem. 10 attributes were selected from the data set with assumption of these selected attributes will have a significant amount of impact in the performance of the team. Selected attributes are as follows, city, venue, match type, outcomes, number of overs, player of the match, opposition, toss winner, toss decision and winner. All attributes were categorical values. The problem was further modelled as a classification problem since each attribute set can be represented with three discrete labels namely win, lose and draw. 80% of the available data has been used as the training data and 20% of the available data has been used as the testing data for the analysis. Logistic regression, Decision tree, Support Vector Machine (SVM), Naïve Bayes, K-Nearest neighbors and Random forest machine learning algorithms were used in the analysis using R statistical language.

B. Prediction of batting and bowling position

This model is used to predict batting and bowling position for a given player of the opposition team. A team can come up with new decision, strategies based on the results. For example, number of overs a bowler can bowl is constrained in fast phase matches like T20s. In this scenario, high profile bowlers of a team can be reserved for highly skilled batsman of opposition team based on predicted batting position. Also a team can

predict how well a batsman of opposition team will play based on his batting position. This model was built with minimal level of cricket domain knowledge. Data analyzed from 576 international T20 cricket matches and 636 IPL matches. Problem was modelled as supervised learning classification problem.

C. Player performance evaluation and ranking

Batting performance of a player was evaluated based on 11 attributes. Batting average, strike rate, run opportunities created, runs produced, number of fours, number of sixes, no of dismissal, number of balls faced, dots and byes were used for the analysis. Weights for selected attributes were calculated analyzing their importance along past data. Indian Premier League datasets were used to build the model since these datasets are rich in player statistics. IPL data from year 2008 to 2017 were collected. Dataset yielded with 13000 deliveries of 460 players.

IV. Experimental Setup and Implementation

Data was collected from online sites Cricsheet, ESPNcricinfo and online data mining community Kaggle. These sources are official and frequently updated and maintained by cricket councils and statisticians. Ball by ball details of all the matches played by Sri Lankan cricket team during 2006 to 2017 were collected. Currently active thirty players in Sri Lankan cricket team have been selected and their related career details were collected.

A. Team wise evaluation using Machine Learning

Collected data was in YAML format. Team evaluation part of the work has been carried out by using R statistical language. R is a programming language for statistical computing and consisting of numerous libraries and packages for performing machine learning analysis. Since R development platforms do not provide extensive support for YAML formats, a Java program was written to convert data from YAML to CSV format. 10 attributes were chosen and each attribute set was labelled as win, lose or draw. The problem has been modeled as a classification problem.

1) Logistic Regression: Binomial logistic regression and multinomial logistic regression were used. Half of the draw labels were marked as win and other half was marked as lose for binomial logistic regression. ISLR package was used for building the model.

2) Decision tree: Decision trees in most programming languages support only up to 32 categorical values. City and venue attribute in the dataset were having 72 and 99 levels respectively. A new attribute country was introduced considering relevant country of city or venue for the decision tree model. rpart, caret and e1071 packages in R were used for the implementation.

3) Naïve Bayes: Naïve Bayes algorithm is considered to perform very well on larger data sets. It assumes that the presence of one attribute in a class is completely unrelated to the presence of all other attributes. e1071 package was used for the classification.

4) K-Nearest Neighbors: Final value of K used for the model was 9. dplyr, caret and caTools packages were used.

5) Random Forest: A similar issue to decision tree model was present in modelling random forest classifier. Random forest were supporting up to 53 levels only. Country attribute was used for the analysis. randomForest, dplyr packages were used.

6) Support Vector Machine (SVM): Linear, radial, radial basis function (RBF), polynomial and sigmoid kernels were used to build the SVM classifier.

B. Prediction of batting and bowling position

Collected data was in CSV format. Dataset yielded with 5463 player instances in international T20 matches and 6299 player instances in IPL matches. Ball by ball score, batsman name, bowler name and number of overs attributes were extracted from the dataset. Through this model it was possible to predict batting and bowling position of a given player and possible performance in a match. Decision trees, random forest and K-Nearest Neighbors algorithms were used for building the model using Python programming language.

1) Batting Position: Model consisted of three predictor variables and one response (label) variable. Innings of a batsman was divided into three segments power play (first 6 overs), middle overs (next 10 overs) and death overs (last 4 overs) and runs scored in these segments were used as the three predictor variables. Batting position was considered as the response variable.

2) *Bowling Position:* Model consisted of three predictor variable and four response variables. Innings of a bowler was divided into three segments power play, middle overs and death overs and runs given in these segments were used as the three predictor variables. Maximum possible four bowling positions were the response variable.

C. Player performance evaluation and ranking

In this model a player was evaluated with the aid of performance evaluation metrics based on previous research works. Performance of the team comprises of each players individual performance. This phase consists of evaluating only batsman of a team. Batsman evaluation was done based on three key metrics: average, strike rate and evaluation score.

$$Average = \frac{total\ runs\ scored}{total\ no\ of\ innings - (0.5) no\ of\ non\ dismissed\ innings}$$

$$Proposed\ strike\ rate = \frac{player\ score + (0.5) bye\ runs}{total\ no\ of\ balls\ faced}$$

Evaluation score is the probability of a player to achieve the average performance where it's given that the other players have achieved their average performance. Based on an overall score

players were ranked and their collective score was used to give score for the team. Higher score indicates a high performance team.

V. Results and Analysis

A. Outcome prediction model

Accuracy on prediction result of K-Nearest Neighbor algorithms is shown in the following table. Best results were gained for the K value of 9.

Table I: K-NN model evaluation

Model feature	Prediction class: draw		Prediction class: lose	Prediction class: win
Sensitivity	0.9010		0.5441	0.6000
Specificity	1.0000	0.5915	0.5443	
Positive prediction	1.0000	0.5606	0.5000	
Negative prediction	0.9275	0.5753	0.6418	
Balanced accuracy	0.5454	0.5678	0.5722	

Variance of accuracy with the mtry parameter (number of variables available for splitting at each tree node) in SVM model is shown in Figure 1.

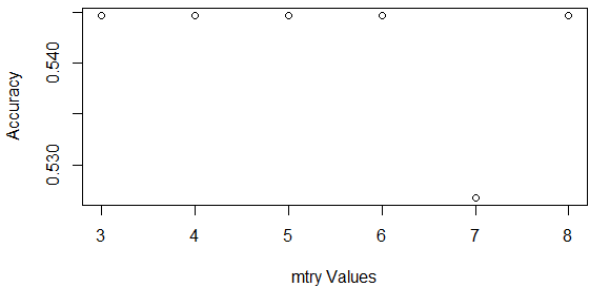

Figure 1: Accuracy with related to mtry values in SVM

Following table shows the comparison of accuracy in different models.

Table II: Accuracy comparison of algorithms

Algorithm	Accuracy
Logistic Regression (Binomial)	0.4536
Logistic Regression (Multinomial)	0.4218
Decision Tree	0.4602
Naïve Bayes	0.6209
K-Nearest Neighbors	0.5324
Random Forest	0.6103
SVM (Radial Kernel)	0.7021

Accuracy was calculated by applying the model in both testing and training data set. Naïve Bayes, Random Forest and SVM algorithms were able to produce better accurate models for both cases.

B. Batting and bowling prediction model

Following tables shows the comparison of accuracy in different models and their underlying algorithms. Second column considers models which treats predicted value equals plus or minus the actual value as correct result.

Table III: Batting prediction model (IPL match data)

Algorithm	Correct predictions	Correct prediction ± 1
Decision tree	0.32	0.78
Random forest	0.31	0.74
K-Nearest Neighbors	0.28	0.69
SVM	0.27	0.74
Gaussian Naïve Bayes	0.19	0.48
Multinomial Naïve Bayes	0.29	0.64
Bernoulli Naïve Bayes	0.28	0.72

Table IV: Batting prediction model (Int. T20 match data)

Algorithm	Correct predictions	Correct prediction ± 1
Decision tree	0.37	0.85
Random forest	0.37	0.81
K-Nearest Neighbors	0.35	0.82
SVM	0.25	0.74
Gaussian Naïve Bayes	0.30	0.77
Multinomial Naïve Bayes	0.29	0.79
Bernoulli Naïve Bayes	0.21	0.70

Table V: Bowling prediction model (IPL match data)

Algorithm	Correct predictions	Correct prediction ± 1
Decision tree	0.29	0.54
Random forest	0.28	0.53
K-Nearest Neighbor	0.27	0.49

Table VI: Bowling prediction model (Int. T20 match data)

Algorithm	Correct predictions	Correct prediction ± 1
Decision tree	0.29	0.54
Random forest	0.28	0.53
K-Nearest Neighbor	0.27	0.49

C. Player performance evaluation model

Weights related with chosen attributes for evaluating a player and their standard error rate is shown in the following table.

Table 1: Attribute weights of player evaluation model

Attribute	Weight	Standard Error
Batting Average	0.01442286	0.01394049
Strike Rate	1.23438038	0.08286316
ROC	1.78483827	0.56422746
Run Produced	1.78771888	0.56430196

fours	0.59532027	0.42977277
sixes	2.217309715	0.94357401
innings	1.342090863	1.20595439
No, Dismissal	-1.165790524	1.45965066
balls Faced	-0.06505113	0.06016719
dots	-0.099502378	0.18600119
byes	0.987682686	0.17422221

Following table shows the score given by the model for teams played in IPL 2017 season based on its players and their actual final rankings.

Table 2: Performance scores of IPL teams

Team	Team score	Actual rank
Mumbai Indians	36.3050	1
Rising Pune Supergiant	18.9457	2
Sunrisers Hyderbad	18.9409	3
Kolkata Knight Riders	16.2032	4
Royal Challengers Bangalore	15.2791	8
Delhi Daredevils	14.4269	6
Kings 11 Punjab	9.6930	5
Gujarat Lions	6.4708	7

VI. Conclusions and Future Works

An effective performance evaluation model of a team will significantly contribute to predict outcomes of a match. It will help teams to identify better talents and areas of improvements. This work performs team wise evaluation and player wise evaluation. 10 commonly available attributes were chosen from past data and used to predict win, lose and draw outcomes of a match. Player wise evaluation was done in two approaches. First approach was analyzing past data and predicting batting and bowling position of a given player in a match and their possible performance. Second approach was using effective performance evaluation metrics from previous research works and ranking players of a team. Measuring collective performance of a team through individual player performance was the ultimate goal of the second approach. This approach considered only batting performance up to this level. Different machine learning algorithms have been applied and tested. In general, random forest and SVM algorithms perform better in all models.

In conclusion, given a team its performance can be evaluated based on these three categories. As future works, we aim to carry out the followings:

- Finding optimal batting and bowling position for a given player to give better performance
- Including bowling and fielding performance evaluation for the second approach in player wise evaluation
- Combining these three models and creating a user friendly application (calling R, Python scripts as service from a Microsoft SQL Server)

References

[1] "Cricket Sabermetrics." https://biomechanics101.wordpress.com/2016/09/22/cricket-sabermetrics/. (Accessed on 25/05/2018).

[2] V. V. Sankaranarayanan, J. Sattar, and L. V. S. Lakshmanan, "Auto-play: A data mining approach to ODI cricket simulation and prediction," *Proc. 2014 SIAM Int. Conf. Data Min.*, pp. 1064–1072, 2014.

[3] K. Passi and N. Pandey, "Increased Prediction Accuracy in the Game of Cricketusing Machine Learning," *Int. J. Data Min. Knowl. Manag. Process*, vol. 8, no. 2, pp. 19–36, 2018.

[4] S. Agarwal, L. Yadav, and S. Mehta, "ScienceDirect Cricket Team Prediction with Hadoop : Statistical Modeling Approach," *Procedia Comput. Sci.*, vol. 122, pp. 525–532, 2017.

[5] G. R. Amin and S. K. Sharma, "Cricket team selection using data envelopment analysis," *Eur. J. Sport Sci.*, vol. 14, no. SUPPL.1, pp. 369–376, 2014.

[6] F. Ahmed, K. Deb, and A. Jindal, "Multi-objective optimization and decision making approaches to cricket team selection," *Appl. Soft Comput. J.*, vol. 13, no. 1, pp. 402–414, 2013.

[7] P. A. Gregory, D. H. M. S. N. Herath, D. S. L. Karunasekera, D. S. Deegalla, and A. U. Bandaranayake, "Cricket sabermetrics: A data mining analysis of cricket," *in Proceedings of the 6th YSF Symposium*, pp. 33–38, 2017.

[8] P. Somaskandhan, G. Wijesinghe, L. B. Wijegunawardana, A. Bandaranayake, and S. Deegalla, "Identifying the optimal set of attributes that impose high impact on the end results of a cricket match using machine learning," *2017 IEEE Int. Conf. Ind. Inf. Syst.*, pp. 1–6, 2017.