2019 5th International Conference for Convergence in Technology (I2CT)
Pune, India. Mar 29-31, 2019

Data Mining and Machine Learning in Cricket Match Outcome Prediction: Missing Links

Manuka Maduranga Hatharasinghe
Informatics Institute of Technology
No 57, Ramakrishna Road, Colombo 6, Sri Lanka
maduranga.manuka3@gmail.com

Guhanathan Poravi
Informatics Institute of Technology
No 57, Ramakrishna Road, Colombo 6, Sri Lanka
Guhanathan.p@iit.ac.lk

***Abstract* - Using Computer Intelligence to analyze and model the game of Cricket is a promising research area. The increased popularity and financial benefits have made Cricket an interesting sport to be subjected to statistical analysis and machine learning. The dynamic nature of Cricket, complex rules governing Cricket makes the task a challenging one. The various approaches taken and what has been disclosed from available work is neither very clear nor properly documented due to the differences in the approaches. If the good and the drawbacks of the existing work is properly analyzed and documented, it will assist in future researches. This paper presents an analysis of the existing work related to match outcome prediction in the Cricket domain. This paper is a result of an ongoing research, by the end of the research we hope to address the missing links and the drawbacks that will be explored in this paper.**

Keywords – Cricket, Sabermetrics, Data Mining, Statistics, Social Media, Twitter, Machine Learning

I. INTRODUCTION

This section introduces the domain "Cricket match outcome prediction" with background information and why it is such a promising research area.

A. Background

Cricket is one of the most popular sports in the world, second only to football [2]. Cricket is played globally across 106 members of the International Cricket Council (ICC). A Cricket match is played between two countries mainly in three formats, Test Cricket, One Day International (ODI) and Twenty20 (T20). Test Cricket match is played over 5 days with 90 overs per day and One Day Cricket is played with 50 overs per each side while Twenty20 Cricket is played with 20 overs per each side. A 50-Over Cricket World Cup is held once every 4 years and a Twenty20 World Cup is held once every 2 years. In between the World Cups,
teams tour other Cricket playing nations to play Test, ODI, T20 Cricket. Recently, with the introduction of domestic franchise Cricket tournaments like Indian Premiere League (IPL), Caribbean Premiere League (CPL), Pakistan Super League (PSL) etc., Cricket has become a heavy financial sport with billions of dollars involved in player auctions and prize money.

B. Motivation

Cricket can be considered as a very unpredictable sport. The whole outlook of a Cricket match can be changed within a few minutes. Due to the complex nature of the game, decisions on team selection, player performance prediction, match outcome prediction can be tough. As any other sport, every Cricket match leave behind a huge set of data that can be analyzed and modeled to extract data driven insights of the game. These insights can be very helpful to anyone who's involved in any decision-making process related to the game.

In Section 2, we explore the ways in which we gathered data and statistics in order to complete the task. In Section 3, we analyze the Cricket match outcome prediction in depth and the problems involved. In Section 4, we explore the existing solutions proposed under the category of approaches they have taken. In Section 5, we evaluate the existing solutions on a few evaluation metrics. In Section 6, we explore the identified missing links and drawbacks and in the last section we conclude the paper with identifying any possible future work.

II. STUDY SETUP

We started gaining the required domain knowledge with the literature survey and we identified 3 main problem domains that are linked with Cricket match outcome prediction,

1. Cricket Player Performance analysis.
2. Cricket match simulation.
3. Cricket Team Selection.

We explored more than 25 researches done under these domains and we identified 2 main approaches researchers have used in the past, which are discussed in Section 6. Under these approaches we identified 6 main solutions developed by researches, which are also discussed in Section 6.

III. CRICKET MATCH OUTCOME PREDICTION

Cricket is a dynamic game. A team might seem to be way ahead at the half way stage or at any stage of the game but an extraordinary performance from one player on the other team can change the outcome of the match within a few minutes. Also, various factors such as natural elements,

978-1-5386-8075-9/19/$31.00 ©2019 IEEE

complex rules regulating the game and the performance of players on a given day etc. play a pivotal role in the outcome of a match. Given the array of factors affecting the game and also its dynamic nature, predicting the outcome of a Cricket match is a challenging task [3].

IV. EXISTING WORK

A. *Computer Intelligence on other sports.*

There has been substantial number of researches in the past about predicting the outcome of sports events, especially in the context of Soccer and Basketball. Bhandari et al. [4] developed a Scout system to find hidden patterns in Basketball games which is now used by the NBA teams. Luckner et al. [5] predicted the results of the FIFA World Cup 2006 using live Prediction Markets. Gartheepan et al. [6] built a data-driven model for Baseball games to help decide when to 'pull a starting pitcher'. Since these researches heavily depend on the sport, they can't be used in the context of Cricket.

B. *Computer Intelligence on Cricket*

The existing work on predicting the outcome of a Cricket match is discussed in this section.

In the recent past there has been substantial interest around the above identified problem. Researchers have taken various approaches to solving the problem, in which we identified 2 main approaches used.

1. Using historical Cricket data.
    a. Classification using team and categorical data.
    b. Match Simulation.
    c. Classification using team composition data.
2. Using Collective knowledge (Social Media).

1) *Using historical Cricket data*

In Cricket Statistics approach researchers use the Cricket data available to develop statistical prediction models to make predictions about up coming matches. This approach is one of the most successful approaches since it takes advantage of the already existing Cricket data to learn and model the game. The disadvantages of this model are that since it is heavily depended on the available cricket data, the models perform poorly when it comes to the International Cricket arena, since most teams don't play each other enough to provide sufficient training data [5]. Also lack of past data about new players can also affect the prediction models.

a.) ***Classification using team and categorical data.***

Kampakis et al. [3] (2018) predicted the outcome of English County Cricket matches using two models. One model using team data and the other model using a combination of team data and player data. They also used a hierarchy of features where level 1 features consisted of raw cricket metrics and level 2 features consisted of combination of two level 1 features and level 3 features consisted of combination of level 2 features.

They experimented with variety of classification algorithms and concluded that Naïve Bayes learner gave them the highest accuracies. With significant data preprocessing, feature selection and complex hierarchical features they were able to correctly predict the outcome of the match in two thirds of instances.

They also compared the accuracy level of their model with the accuracy benchmarks in the betting industry and stated their model was able to consistently beat the accuracy levels in the betting industry.

Singh et al. [6] (2015) developed a model to predict the first innings score of a Cricket match and to predict the final outcome of the match in the Second innings of ODIs. They developed two separate models, one to predict the first innings score and the other to predict the outcome of the match.

They used Linear Regression to predict the first innings score using a 5 over interval approach and achieved much better accuracy levels in predicting the first innings score than the traditional run rate metric.

A Naïve Bayes classifier was used by them to predict the final outcome of the match. They noticed the classifiers accuracy increased as the match progressed, starting at 70% and reaching 91% as the 40th over of the match was reached.

b.) **Simulation based Approach.**

Simulation based approach is similar to classification approach, in that both approaches use past Cricket data to model the game. Simulation approach is different when it comes to the underlying mechanisms of how the data is used. In simulation approach the game progression (no. of runs scored) is predicted continuously where as in classification, the final outcome is given as a 'WIN' or 'LOOSE'.

Sankaranarayanan et al. [7] developed a model to simulate game progression and predict match outcome in ODIs. They developed separate models for home games and away games using historical Cricket data and also instantaneous data extracted from past matches. Their model was able to simulate the rest of the Cricket match given a particular stage of the match, thus predicting the winner of the match. They stated their model was able to achieve highest reported accuracy in outcome prediction in ODI Cricket. They have also shown the model developed by them was able to outperform the model developed by Bailey et al [8].

c.) ***Team Composition approach.***

Predicting the outcome of Cricket matches by calculating relative team strengths has proven to be the most effective and accurate approaches taken by researchers. This approach calculates team strength by modeling the batting potency and bowling potency of individual players in the team. The player combination of both the teams can play a pivotal role in the outcome of a match. However, the player combination of a team can vary often due to a variety of reasons. In this case, predicting match outcomes by historical data of a particular team alone can lead to inaccurate predictions.

Jhanwar et al. [9] developed a model to predict outcome of Cricket matches using a team composition-based approach. They stated due to the changing combinations of teams, predicting the winning team based only on historical team data would lead to inaccurate predictions. To overcome this bottleneck, they treated participating players in both the teams as the key feature. They developed novel methods to model batsmen, bowlers and teams, using various career statistics and recent performances.

For batsmen, they calculated values such as 'Career Score' and 'Recent Score' from the batsmen's career statistics. A final value, 'Batsmen Score' was calculated by combining 'Career Score' and 'Recent Score'. For bowlers, 'Bowler score' was calculated using the same approach as for batsmen. A final value for the team was calculated from the combination of 'Bowler Score' and 'Batsmen Score'. Finally, they calculate the relative strengths of the competing teams.

Their model was able to achieve high accuracy levels, with the average accuracy reaching 71%. It is also worthy noting, their model achieved even higher accuracy levels on some teams.

2) Collective Knowledge Approach

Collective knowledge approach is a completely different approach from what we have discussed so far. It is also a very promising one. This approach follows the 'Wisdom of Crowds' concept, which harnesses cumulative opinions of a diverse and a large group of individuals to make predictions about real-world events around the globe. This concept has been used in many researchers in many domains like, elections [11], spread of infectious diseases [10], stock market predictions [12] etc. A number of researches has also used this concept in the context of sports to extract meaningful information to predict outcomes.

Mustafa et al. [13] assessed the effectiveness of harnessing collective knowledge from social media to predict the outcome of Cricket matches. They considered three aspects of Twitter data,

- Twitter Volume (TV) – Number of Tweets for a given team.
- Aggregated fans' sentiments (FS) – Polarity of the sentiment for a given team.
- Score prediction. (SP) – Scores predicted by the fans.

To evaluate their hypothesis they used three classifiers, SVM, NB and LR, SVM was concluded as the best performing out of the three. They further verified their methodology on the games played at the Cricket World Cup 2015 (CWC15) and Indian Premiere League 2014 (IPL14). For CWC15 their methodology was able to beat the odds of the betting industry, and theoretically gain a 67% profit.

They also noticed low accuracy levels for some teams at the CWC15 and argued the lack of tweets for those teams resulted in the lower accuracies and stated that classifier accuracy increases with a larger dataset.

V. ANALYSIS OF EXISTING WORK.

In this section we analyze the existing solutions explored above on 3 main categories,

1. Type of data used.
2. Type of features used.

1) Type of data used.

The following tables shows the comparison of the types of data used in the above explored solutions.

The (1) notation refers to the used data while (0) refers to not used.

	Cricket Statistics approach				Collective knowledge approach
	[3]	[6]	[7]	[9]	[13] *
T20I	0	0	0	0	0
ODI	0	1	1	1	1
Test	0	0	0	0	0
IPL	0	0	0	0	1
County T20	1	0	0	0	0

* Mustafa et al. [13] developed two separate models, one with ODI data and another with IPL data

2) Types of features used and accuracy.

The following table shows the comparison of types of features inputted into the algorithms and the accuracy of the solutions.

	Cricket Statistics approach					Collective knowledge approach
	$[3]_1$	$[3]_2$	[6]	[7]	[9]	[13] *
Raw features	1	1	1	1	1	1
Engineered features	1	1	1	1	1	1
Categorical Features	0	0	1	1	1	0
Accuracy	62.1 %	64.5 %	68.6 %	70 %	71 %	87%

The (1) notation refers to the used data while (0) refers to not used.

Raw features – Traditional Cricket metrics used in Cricket as of today. (Batting average, Win/Loss ratio, Strike rate of batsmen etc.)

Engineered features – Features developed with the combination of raw features. (Batting index – Batting average + Strike rate etc.)

Categorical features – Features that take a defined value. (Toss, Venue etc.)

$[3]_1$ – Model developed with team data only.
$[3]_2$ – Model developed with team data and player data.
* - Accuracy of the model developed with CWC15 data.

VI. MISSING LINKS

In this section we will explore the findings of the study focusing on what has been missing and common shortcomings in the solutions developed.

Dealing with the lack of data about certain players: When taking a team composition approach, where the relative strength of a team is calculated from the statistics of the players in the team, a new player having only played few games or a playing making a debut can lead to inaccurate results, this issue was not addressed in any of the studies we explored.

Not giving enough consideration to the conditions the game is played in: As the game of Cricket has showed us time and time again, conditions in which the match is played play a pivotal role in the outcome, where the home team has an advantage over the visiting team. Therefore, it's vital that this is taken into consideration in the feature selection and engineering phase.

Taking an ensemble approach: We identified 2 main approaches taken by researchers so far, using historical cricket data and using collective knowledge from social media. No attempts have been made so far to use these two approaches in tandem to improve the accuracy of the predictions made.

VII. CONCLUSION AND FUTURE WORK

Analyzing the results from the existing work, we derived the following conclusions.

Cricket match outcome prediction remains a new and a promising research area. Due to the complex and dynamic nature of the game, achieving high accuracy scores remains a challenging task, especially in the T20 format. The motivation to model the game of Cricket is shared among many researchers due to the deep and meaningful insights and financial benefits it can entail.

As seen through out this paper, a bigger data set can improve the accuracies of the prediction. This must also be the reason why almost no attempts have been made to use Test Cricket data to make predictions, as very few number of Test matches are played in a year. The best format to use to model the game is T20, due to the vast number of matches played in a year. Also, T20 remains the hardest format to model due to its highly unpredictable nature compared to ODIs or Test matches.

With this knowledge of the existing work and identified missing links, in the near future we hope to develop a model that can accurately predict outcome of Cricket matches.

REFERENCES

[1] J. Surowiecki and P. Silverman, "The wisdom of crowds", American Journal of Physics, Vol. 75, No. 2, pp. 190-192, 2005.

[2] B. E. Sawe, "The Most Popular Sports in the World," *World Atlas*, 16-Sep-2016. [Online]. Available: https://www.worldatlas.com/articles/what-are-the-most-popular-sports-in-the-world.html.

[3] Kampakis, S. and Thomas, W. (2018). Using Machine Learning to Predict the Outcome of English County twenty over Cricket Matches. [online] Arxiv.org. Available at: https://arxiv.org/abs/1511.05837 [Accessed 1 Sep. 2018].

[4] 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.

[5] Brooks, R. D., Faff, R. W., & Sokulsky, D. (2002). An ordered response model of test cricket performance. Applied Economics, 34 (18), 2353-2365.

[6] Singh, T., Singla, V., Bhatia, P.: Score and winning prediction in cricket through data mining. In: International Conference on Soft Computing Techniques and Implementations (ICSCTI), pp. 60–66. IEEE (2015)

[7] Sankaranarayanan, Vignesh Veppur, Junaed Sattar, Laks VS Lakshmanan.: Auto-play: A Data Mining Approach to ODI Cricket Simulation and Prediction. SDM.2014

[8] 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):480–487, 2006.

[9] Jhawar, Madan & Pudi, Vikram. (2016). Predicting the Outcome of ODI Cricket Matches: A Team Composition Based Approach.

[10] M. I. Lali, R. U. Mustafa, K. Saleem, M. S. Nawaz, T. Zia and B. Shahzad, "Finding healthcare issues with search engine queries and social network data", International Journal on Semantic Web and Information Systems, Vol. 13, No. 1, pp.48-62, 2017.

[11] A. Tumasjan, T. O. Sprenger, P.G. Sandner and I. M. Welpe, "Predicting elections with Twitter: What 140 characters reveal about political sentiment", in Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media, pp. 178-185, 2010.

[12] J. Bollen, H. Mao and X. J. Zeng, "Twitter mood predicts the stock market", Journal of Computational Science, Vol. 2, No. 1, pp. 1-8, 2011.

[13] Mustafa, Raza & Nawaz, M. Saqib & Lali, Muhammad Ikram & Zia, Tehseen. (2017). Predicting the Cricket Match Outcome Using Crowd Opinions on Social Networks: A Comparative Study Of Machine Learning Methods. Malaysian Journal of Computer Science. 30. 10.22452/mjcs.vol30no1.5.