Predicting an Optimal Sri Lankan Cricket Team for One Day International Matches According to the Nature of the Game

[1]W.A.S.C. Perera, [2]H.A. Caldera

Abstract— *this paper focuses on predicting an optimal Sri Lankan cricket team for One Day International (ODI) matches according to the nature of the game. In general, the team selection process in One Day International is based on performance measures such as batting and bowling averages. These measures have several numbers of limitations. The number of runs scored by batsmen and wickets taken by bowlers serves as a natural way of quantifying the performance of a cricketer. However, the factors such as scoring runs against a strong bowling line-up or delivering a brilliant performance against a team with a strong batting line-up, etc. deserves more credit. In this paper, we present a new method of prediction by scanning the dependencies applied in the game such as the average performances of the players, the ground, the opposition team and the match outcome. Due to the complexity in the data set in size and the dimension, and analysis required, advanced analysis techniques such as Clustering and Association Rule Mining has been used to predict the players. The study concludes by predicting teams (eleven players per each match) for thirty-five matches played in between 2013-2018. The final outcome shows that the Sri Lankan cricket team can win the match with 88% by predicting players using our system.*

Index Terms— *Association Rule Mining, Clustering, Cricket, Game conditions*

I. INTRODUCTION

Cricket can be considered as the most popular game in Sri Lanka. The Sri Lankan national cricket team has gained vital importance and prestigious recognition in the country [1]. It is the responsibility of the Sri Lankan Cricket Selection Committee to rank the players and select the national team as well as the required squads. The general criteria on which the selectors will be considered are, Current form, Past performances (batting average, strike rate, bowling average, and economy rate), Balance of the team, Health/fitness, Contribution to the team environment, and investing in youth development.

This manual process consumes more effort and time. More importantly, selectors just consider few factors only for the selection process. Basically, they consider about the average performances of the players only. The number of runs scored by batsmen and wickets taken by bowlers serves as a natural way of quantifying the performance of a cricketer. It is accepted that these measures have several numbers of limitations in assessing the true performances of players [2]. When selecting a team, consists of eleven players, plenty of information should be considered: the performances, the ground, the opposition team, the match outcome, and etc. This each element can make a big difference to the final outcome.

The paper is focused on predicting an optimal Sri Lankan cricket team for One Day Internationals according to

[1] *Department of Physical Science, Vavuniya campus of the University of Jaffna, Vavuniya, Sri Lanka, anneshehari@gmail.com*
[2] *University of Colombo School of Computing, Colombo, Sri Lanka, hac@ucsc.cmb.ac.lk*

DOI: 10.37200/IJPR/V24I4/PR201034
Received: 22 Oct 2019 | Revised: 13 Sep 2019 | Accepted: 15 Jan 2020

the game's nature.

Choosing a team is more than just picking the best players from a pool based on the batting and bowling averages. The team should be balanced and the balance should reflect closely the tactics having for winning the match. For that, a plenty of factors should be considered.

The ground is directly affected the team selection. A pitch consisted of loose clay or sand (dusty pitch) favors the spinners to get a good amount of spin and bounce from the pitches. Green pitch is a challenge for even the best batsmen as they have to judge the movement of the ball after pitching in a short time. Dead pitches favor batsmen a lot. So, the state of the pitch is one of the primary considerations that should be taken into account.

More importantly, the team has an advantage when the pitch is domestically located. The statistics have shown that a home-field is affected by the home teams to win 57% of all matches.

The team selection should be always depended on the opponent. Different players are familiar with opponent teams in a different manner. They show the performances against different teams in a different manner. The personal experiences of each player are highly affected to the above point.

So, there should be an advanced analysis technique that checks all the dependencies for the selection. The current process, which is based on a few factors, is not being able to produce a standard team since the most important factors are hidden. These concealed factors might be able to change the modern game strategies totally.

The objective of this paper is to develop a new method by scanning the dependencies applied in the game and finally predict the optimal teams according to the situations by expecting that the results will have important implications for the Sri Lankan cricket team.

The study considers only the Sri Lankan cricket players who played more than twenty matches as members of the Sri Lankan national cricket team (One Day International matches) or who play for the first class matches domestic or internationally. Current players, who play for the matches at the end of 2018, are taken into account.

Only One Day International (ODI) cricket matches, played in between the year 2013-2018, are considered for both collecting data and predicting teams.

This study is based on the discipline of Data Mining which extracts or mines knowledge from large amounts of data and data is collected mainly from the espncricinfo website.

II. RELATED WORKS

A comprehensive review of the literature regarding the performance analysis of the players reveals the following findings. Lemmer [3] has shown that, in order to be fair, the calculation formulas using for batting and bowling such as batting averages and bowling averages cannot be used in the case of a small number of matches played. He has developed some other formulas to analyze the performances in those cases. Saikia and Bhattacharjee [4] compared the performance of both Indian and foreign cricketers in the Indian Premier League (IPL). They showed the differences between the player performances when they played the IPL and the national team. They have proposed a model by considering the characteristics such as the number of innings, the strike rate, and the batting average to measure the player performance. Both Staden [5] and Bracewell and Ruggiero [6] used graphical measures to illustrate player performances. These researches are based on some mathematical or graphical models. None of these researches have focused on the data mining technique for assessing the players' performances. However, Iyer and Sharda [7] used a neural network approach to predict each cricketer's future performance based

DOI: 10.37200/IJPR/V24I4/PR201034
Received: 22 Oct 2019 | Revised: 13 Sep 2019 | Accepted: 15 Jan 2020

upon their past performance.

In the literature of team selection, Thakare, Sachin, Suyal, and Pandav [8] followed the association rule mining for enhancing the team selection process by considering the attributes such as age, running capacity, experience, and Achievements. But these researchers have done their studies regarding the Handball. Sharp, Brettenny, Gonsalves, Lourens, and Stretch [9] quantifying cricket player's performance based on his ability to score runs and take wickets. By using these performance measures, they have developed an integer program in order to determine the optimal team. Amin and Sharma ([10]) used data envelopment analysis techniques. The proposed Data Envelopment Analysis (DEA) method can be used to select a national cricket team from club players or top-ranked players. None of these researches considered the game's nature. They have found an optimal team which is common to all the matches irrespective of the game's condition.

III. METHODOLOGY

Data Mining (Knowledge discovery) refers to the nontrivial extraction of implicit, previously unknown and potentially useful patterns, associations or interesting knowledge from data in databases. The dataset for analysis was collected from espncricinfo website. The ODI matches played in between 2013 January to 2018 December were taken into account. 246 players (46 national players and 200 first class players) and 143 matches' data have been considered. Basically, four data sets, average batting performances of players, average bowling performances of players, bating performances of players in each match, and bowling performances of players in each match, have been collected throughout the study. These data sets consist of both nominal and numeric data. Due to the complexity in the data set in size and the dimension, and analysis required, it was decided to apply the Data Mining technique on them.

As shown in Fig. 1, the entire study can be mainly divided into two stages. In the first stage, the data sets are analyzed based only on the batting and bowling averages of each player. As a result, the pool of the most talented batsmen and bowlers are obtained. As the second stage, this pool again has been analyzed based on the nature of games such as the ground and the opposition team. By using the results achieved from the second stage, the optimal teams have been predicted according to the conditions of games.

IV. PERFORMANCE ANALYSIS

The performance analysis is first conduct on two groups of batsmen and bowlers. It was decided to conduct the cluster analysis to identify the pool of potential best batsmen and bowlers by using K-means partitioning clustering algorithm [11].

K-means is an unsupervised learning algorithm. It uses means in grouping a given data set through a certain number of clusters which is represented by the variable K. The algorithm works iteratively to assign each data point to one of K groups based on the features that are provided. Data points are clustered based on the similarity of features [11].

DOI: 10.37200/IJPR/V24I4/PR201034
Received: 22 Oct 2019 | Revised: 13 Sep 2019 | Accepted: 15 Jan 2020

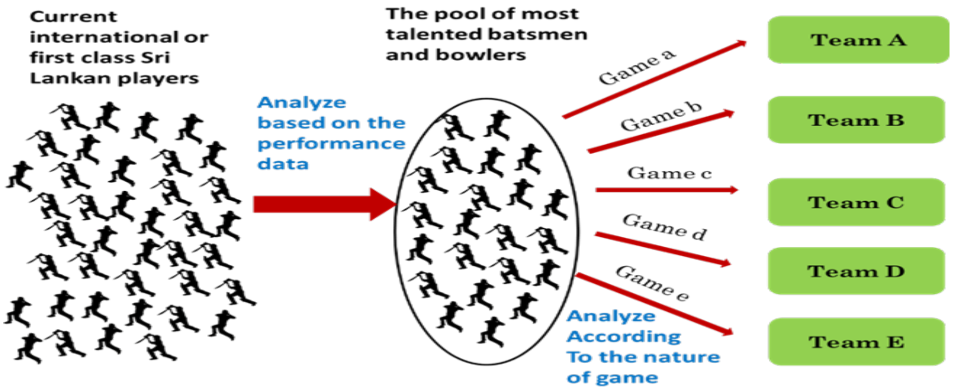

Current international or first class Sri Lankan players
Analyze based on the performance data
The pool of most talented batsmen and bowlers
Game a
Game b
Game c
Game d
Game e
Analyze According To the nature of game
Team A
Team B
Team C
Team D
Team E

Figure 1: An overall image of the study

The biggest challenge with K means clustering is finding the most optimal number of clusters. The Elbow finding technique was decided to use for this purpose [12]. In this technique, the sum of squared error is calculated for the different values of k (number of clusters) and selects the k value at the elbow as the optimum k.

As illustrated above, this stage is about finding the pool of best batsmen and bowlers.

A. Batting Performance Analysis

The sum of squared error calculated for the different values of k are shown in the Fig. 2. As the abrupt change occurs at K=4, it was decided to use four as the optimized number of clusters that should be used to analyze the performances of batsmen.

The two batting statistics often used as a measure of a player's performance; the batting average rate and the batting strike rate, are used for the batting performance analysis through K-means. Since the national players performances and first-class players performances cannot be comparable, analysis was carried out independently for each of those two. Fig. 4 and Fig. 5 show the two-cluster analysis performed against the batsmen's data set by using the Batting Average and the Batting Strike Rate attributes.

Fig. 4 shows the clusters of national batsmen's Batting Strike Rate based on Batting Average. A batsman with high values of batting average and batting strike rate is considered to be a good player [9]. Based on that concept, it was decided to select the batsmen who are within the cluster 1. Other clusters were ignored. So, nine players have been selected for the best players' pool as batsmen.

Fig. 5 shows the clusters of first-class batsmen's Batting Strike Rate based on Batting Average. As mentioned above a batsman with high values of batting average and batting strike rate is considered to be a good player. Based on that, cluster 0 has been selected. So, six players have been selected for the best players' pool as batsmen. Other players were ignored.

Finally, all together 15 players have been selected for the best players pool as batsmen.

B. Bowling Performance Analysis

DOI: 10.37200/IJPR/V24I4/PR201034
Received: 22 Oct 2019 | Revised: 13 Sep 2019 | Accepted: 15 Jan 2020

To measure the bowling performances, the study decided to consider two bowling statistics often used as a measure of a player's performance, bowling average and the bowling economy rate.

According to the elbow finding technique, the abrupt change occurs at three for the bowlers' data set, as shown in the Fig. 3. So, it was decided to use three as the optimized number of clusters that should be used to analyze the performances of bowlers.

As mentioned earlier, the national players' performances and first-class players' performances cannot be comparable. So, analysis was carried out independently for each of those two. Two cluster analysis were performed against the bowlers' data set by using the Bowling Average and the Bowling Economy Rate attributes as shown in Fig. 6 and Fig. 7.

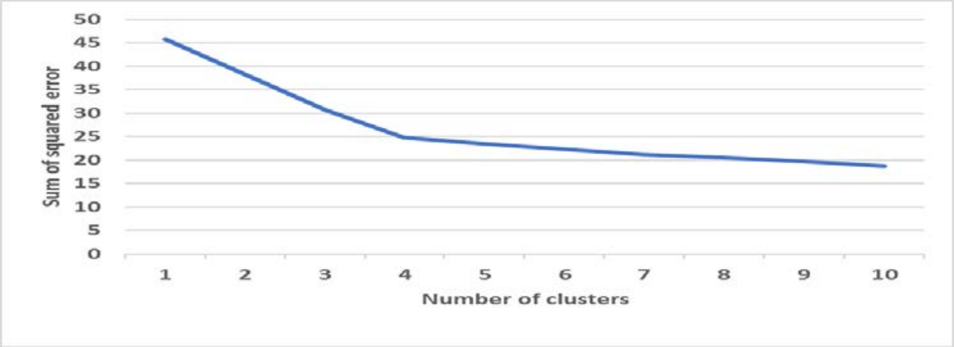


Figure 2: Number of clusters vs. Sum of squared error for the Batsmen's data set

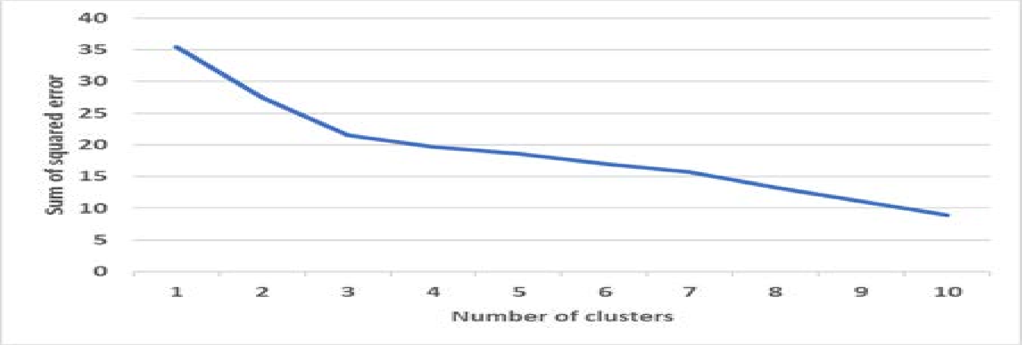


Figure 3: Number of clusters vs. Sum of squared error for the Bowler's data set

Fig. 6 shows the clusters of national bowlers' Bowling Economy Rate based on Bowling Average. A bowler with low values for bowling average and bowling economy rate is considered to be a good player [9]. Based on that concept, it was decided to select the bowlers who are within the cluster 1. Other clusters were ignored. So, five players have been selected for the best players' pool as bowlers.

Fig. 7 shows the clusters of first-class bowlers' Bowling Economy Rate based on Bowling Average. As mentioned above, a bowler with low values of bowling average and bowling economy rate is considered to be a good player. Based on that, cluster 1 has been selected. So, twelve players have been selected for the best players' pool as bowlers. Other players were ignored.

Finally, all together seventeen players have been selected for the best players' pool as bowlers.

The selected players who have been played in between year 2013-2018 are shown in the Table I. Since the

DOI: 10.37200/IJPR/V24I4/PR201034
Received: 22 Oct 2019 | Revised: 13 Sep 2019 | Accepted: 15 Jan 2020

Angelo Mathews is selected as both batsman and a bowler, the final team can be concluded with 31 players.

V. Team prediction overview

The second half of the study considers about the team prediction. The selected pool of players from the previous two steps is analyzed again based on the nature of the game such as, the ground and the opposition team and the most suitable eleven players for the particular game is predicted.

X=Clusters	Y=Batting Average

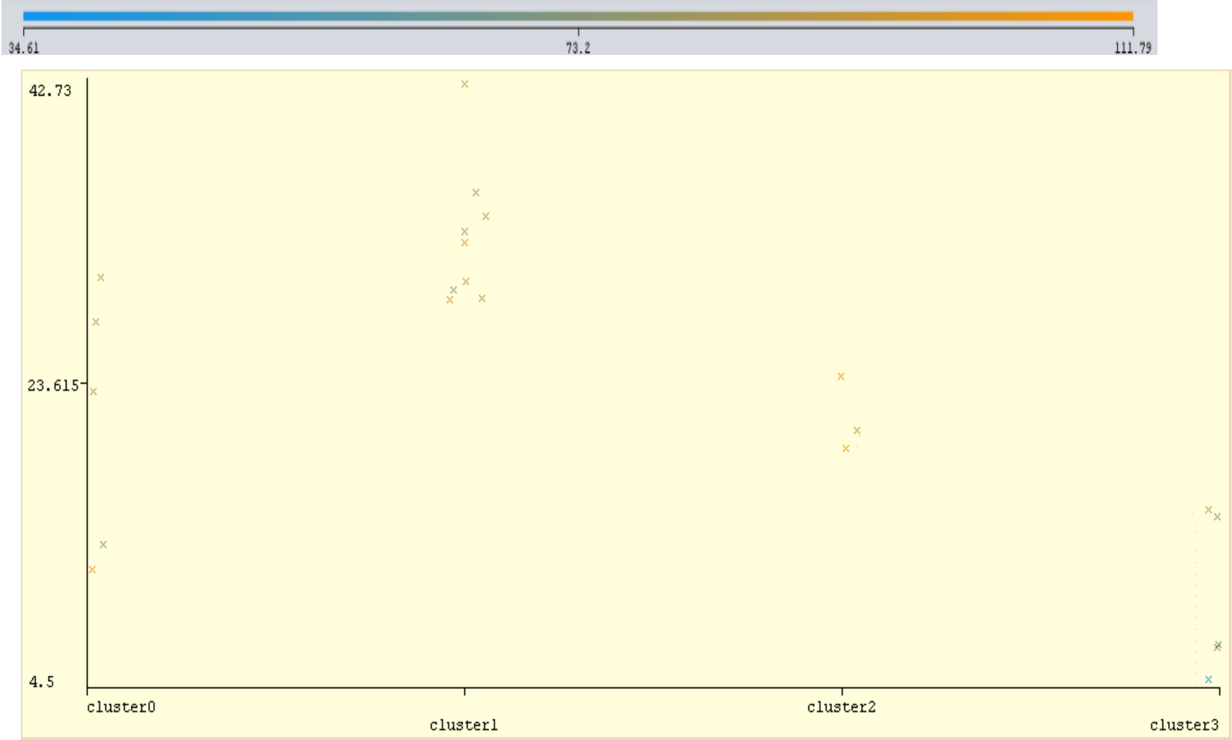
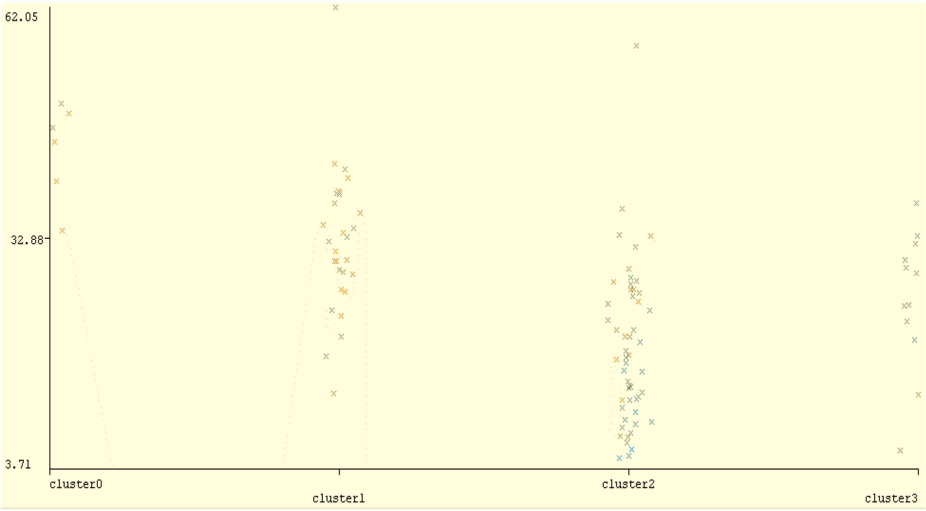


Figure 4: Clusters of national players’ Batting Strike Rate based on Batting Average.



DOI: 10.37200/IJPR/V24I4/PR201034
Received: 22 Oct 2019 | Revised: 13 Sep 2019 | Accepted: 15 Jan 2020

X=Clusters	Y=Batting Average

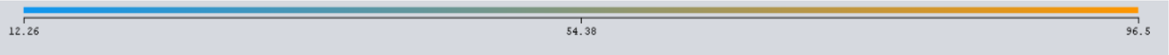


Figure 5: Clusters of first-class players' Batting Strike Rate based on Batting Average.

X=Clusters	Y=Bowling Average

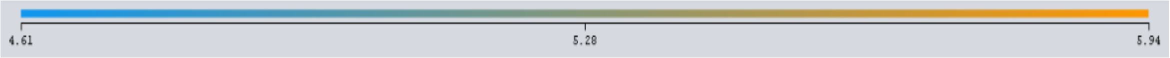
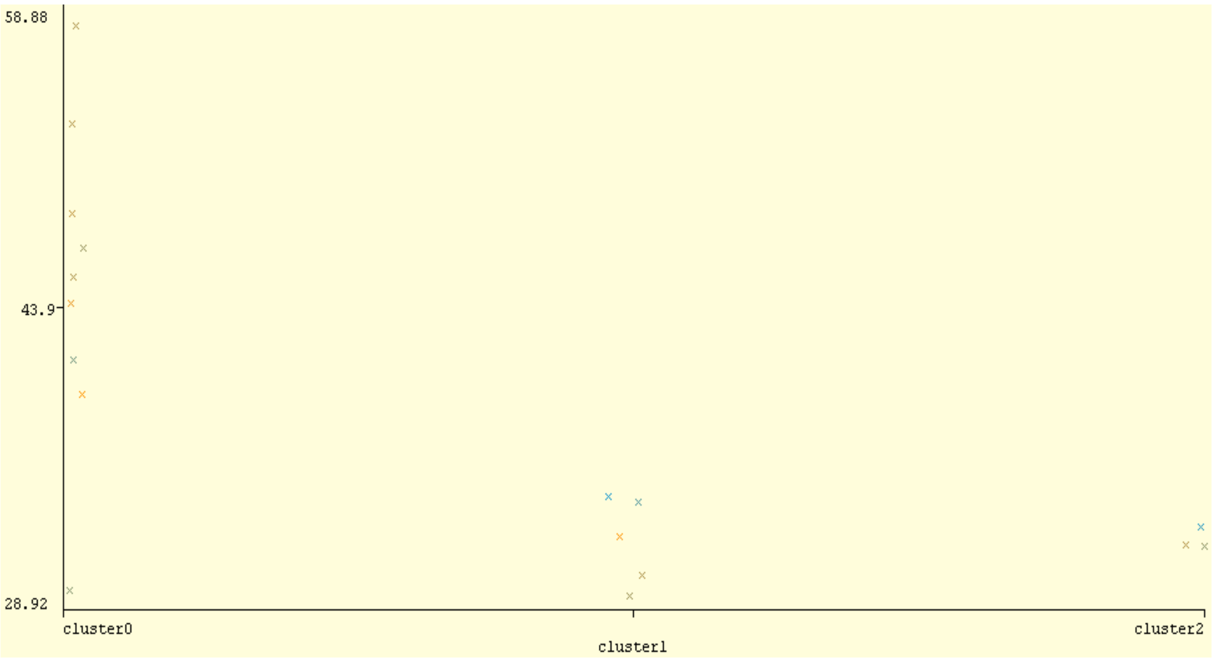


Figure 6: Clusters of national players' Bowling Economy Rate based on Bowling Average



DOI: 10.37200/IJPR/V24I4/PR201034
Received: 22 Oct 2019 | Revised: 13 Sep 2019 | Accepted: 15 Jan 2020

X=Clusters	Y=Bowling Average

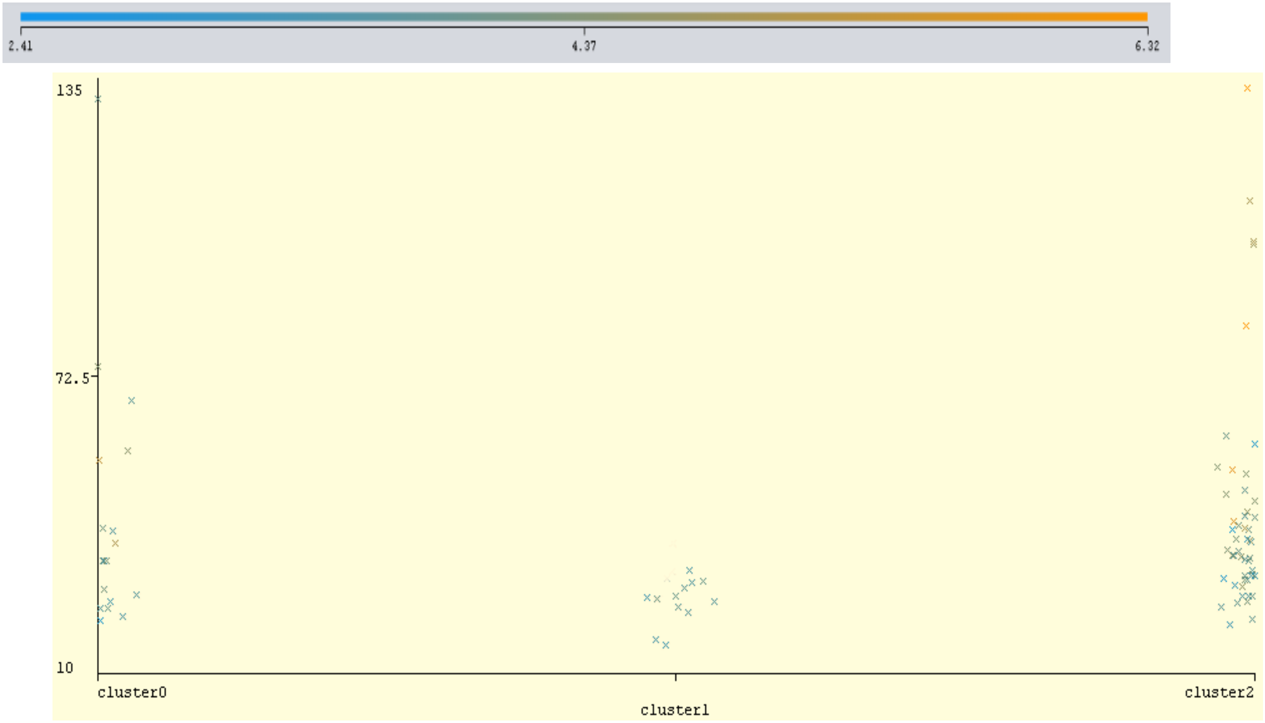
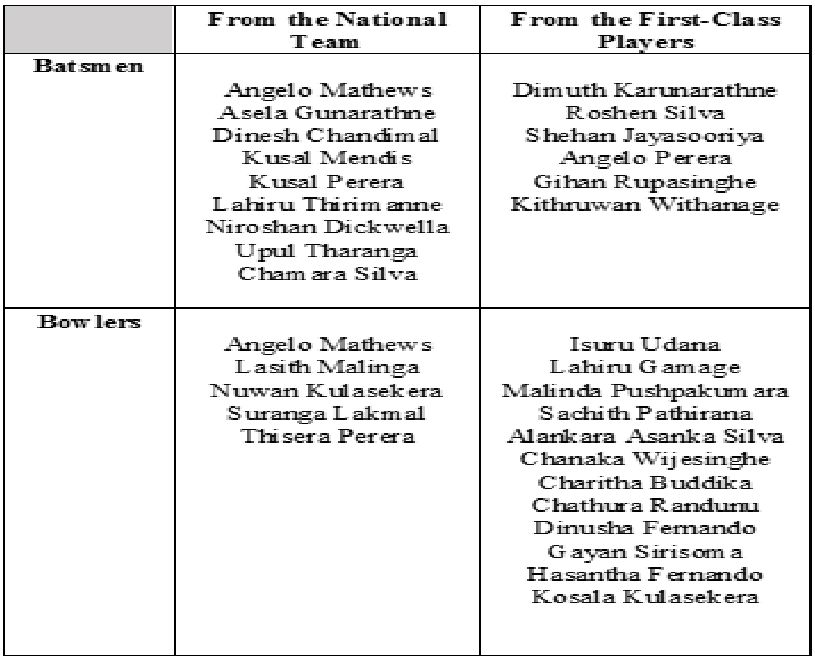


Figure 7: Clusters of first-class players' Bowling Economy Rate based on Bowling Average.

Table 1: The selected players who has been played in between year 2013-2018



	From the National Team	From the First-Class Players
Batsmen	Angelo Mathews Asela Gunarathne Dinesh Chandimal Kusal Mendis Kusal Perera Lahiru Thirimanne Niroshan Dickwella Upul Tharanga Chamara Silva	Dimuth Karunarathne Roshen Silva Shehan Jayasooriya Angelo Perera Gihan Rupasinghe Kithruwan Withanage
Bowlers	Angelo Mathews Lasith Malinga Nuwan Kulasekera Suranga Lakmal Thisera Perera	Isuru Udana Lahiru Gamage Malinda Pushpakumara Sachith Pathirana Alankara Asanka Silva Chanaka Wijesinghe Charitha Buddika Chathura Randunu Dinusha Fernando Gayan Sirisoma Hasantha Fernando Kosala Kulasekera

DOI: 10.37200/IJPR/V24I4/PR201034
Received: 22 Oct 2019 | Revised: 13 Sep 2019 | Accepted: 15 Jan 2020

For this purpose, each players' individual performances, opposition team, ground, and the match outcome have been collected separately from the 143 One Day International matches selected.

For the evaluation purpose the data set has been divided into four parts and the most recent ¼ of matches data (35 matches) have been separated out and 108 matches data have used for the analysis.

As mentioned early, first class cricket is played in between inter zonal teams, basically known as clubs, at the national level. Because of that, there is no way that this study can collect the data regarding the international matches (against international teams in international grounds) played by the first-class players. As a result, again the analysis has to be done separately for the national players and the first-class players.

VI. National Player Analysis According to the Nature of the Game

The runs scored in a particular match are the most crucial factor for a batsman in One Day International matches. For the bowlers, the most important factor is the economy rate. By considering the above two facts, it is decided to select the runs and the economy rate as the attributes to measure the performances of batsmen and bowlers in each individual match.

This study has considered the following formula.

$$\bar{x} = \frac{\Sigma x}{N} \qquad (1)$$

where, in the case of batsmen x represents runs and in the case of bowlers x represents economy rate. N represents the number of matches played by each individual player. By applying this formula against each individual player's performances, it is possible to find out the mean value $\bar{x}$ of performances for each player. This $\bar{x}$ can be used to categorize the players' individual performances as high and low as shown in below.

For the batsmen,

```
If   Runs > = $\bar{x}$   then,
Performance=high
Else
Performance=low
```

For the bowlers,

```
If   Economy Rate < = $\bar{x}$   then,
Performance=high
Else
Performance=low
```

This has made another attribute to be added to the data sets as high and low.

DOI: 10.37200/IJPR/V24I4/PR201034
Received: 22 Oct 2019 | Revised: 13 Sep 2019 | Accepted: 15 Jan 2020

These edited data sets are used to find the association of performance with opponent team and the ground by using the Association Rule Mining algorithm [13].

Association rule mining algorithm is a data mining technique which is used to extract associations among a large set of items. Association Rule Mining based on apriori algorithm in WEKA is used to discover the associations.

The minimum support for the analysis is set to 0.01 and the minimum confidence for the analysis is set to 0.5.

The rules obtained from this stage can be used to predict the national players for a particular game according to the opposition team and the ground of the match.

VII. First-Class Player Analysis

Since the data is not available for the first-class players regarding the international matches, first-class player analysis has to be done separately by using some other attributes. The attribute "Role" simply characterizes the players as batsmen, spinner or fast bowler, based on their major talents.

A data set which consists of different grounds in the world along with the preferences of the role of the player is used for the ground attribute.

It was decided to conduct the cluster analysis to group the players according to the role. Clustering based on the Hierarchical clustering algorithm in WEKA is used as the number of clusters is unambiguous. And also, the hierarchical clustering algorithm is both more flexible and has fewer hidden assumptions.

Fig. 8 shows the clusters of first-class players' roles. Three clusters were created since three roles as batsmen, spinner and fast bowler, are identified in the data set. These clusters can be used to predict the first-class players to the international matches according to the venue of the match.

X=Clusters	Y=Player name
Spinner	🟦
Batsmen	🟥
Fast bowler	🟩

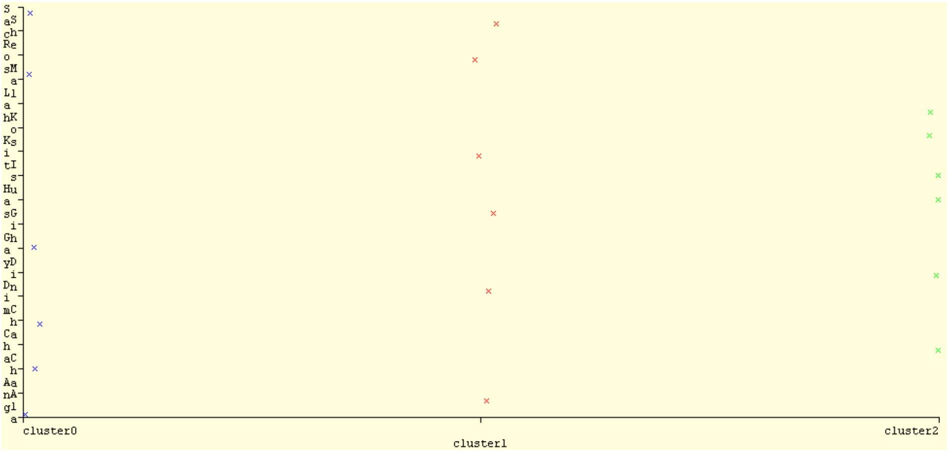


Figure 8: Clusters of first-class players’ role

DOI: 10.37200/IJPR/V24I4/PR201034
Received: 22 Oct 2019 | Revised: 13 Sep 2019 | Accepted: 15 Jan 2020

VIII. EVALUATION

For the evaluation purpose, the data collected regarding the international One Day International matches were divided into four parts and the most recent 1/4 of data (thirty-five matches) were kept without used for the analysis. By using the proposed method, the team members were predicted.

The players which have obtained by predicting for the thirty-five international One Day International matches were evaluated against the real match outcomes. According to the collected data, it was able to calculate the average runs that each batsman can score according to the game's nature. This means, against the particular opposition team in a particular ground, the average score that a particular batsman can score was calculated. The predicted teams consisted of five bowlers and each gets ten overs to bowl. By using these average scores, the total marks that every predicted team could score were calculated as well. These total scores were compared with the real scores of the Sri Lankan team in each match. Fig. 9 shows the comparison and it clearly shows that 88% predicted teams' scores are higher than the real scores.

The same procedure has followed for the bowlers as well. The average economy rate (the average number of runs conceded per over) that each bowler can be obtained according to the game's nature was calculated. This means, against the particular opposition team in a particular ground, the average economy rate that a particular bowler can obtain was calculated.

The predicted teams consisted of five bowlers and each gets ten overs to bowl. By using this average economy rate, the total marks that every predicted team's bowlers concede to the particular opposition teams to score were calculated as well. These total scores were compared with the real scores of the opposition teams for each match. Fig. 10 shows the comparison and it shows that 71% predicted scores are higher than the real scores for the opposition teams.

Finally, both predicted scores (predicted score of the Sri Lankan team that can be obtained and the predicted score of the opposition team which conceded by the Sri Lankan bowlers) for each match were compared with each other. Fig. 11 shows the comparison and the results are quite interesting. It reveals that 88% of the matches can be won by the Sri Lankan team with the predicted players.

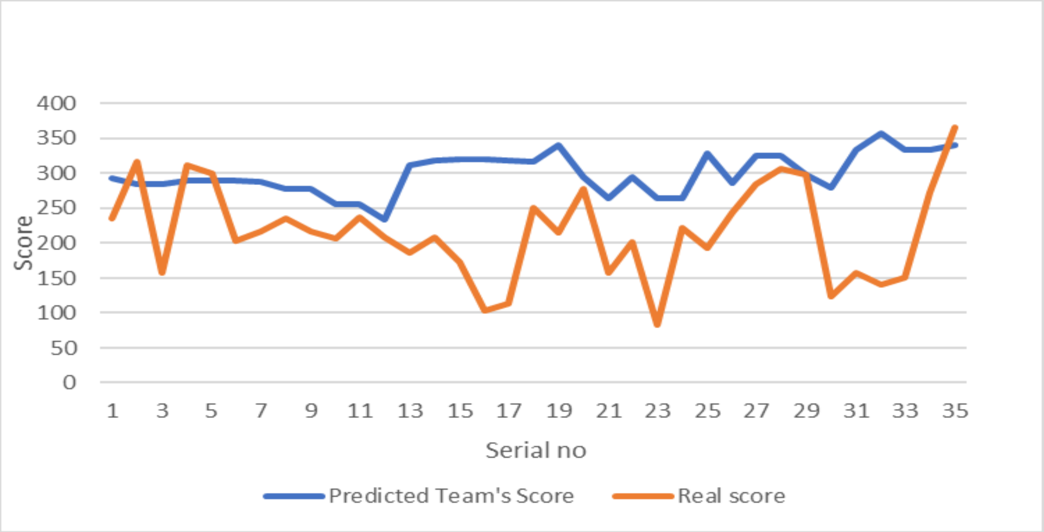


Figure 9: The comparison between the predicted Sri Lankan teams' scores and the real scores

DOI: 10.37200/IJPR/V24I4/PR201034
Received: 22 Oct 2019 | Revised: 13 Sep 2019 | Accepted: 15 Jan 2020

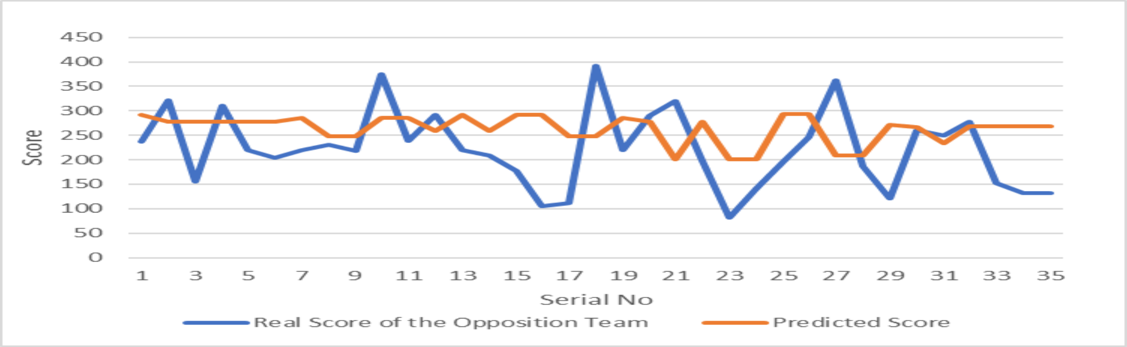


Figure 10: The comparison between the opposition teams' predicted scores and the real scores

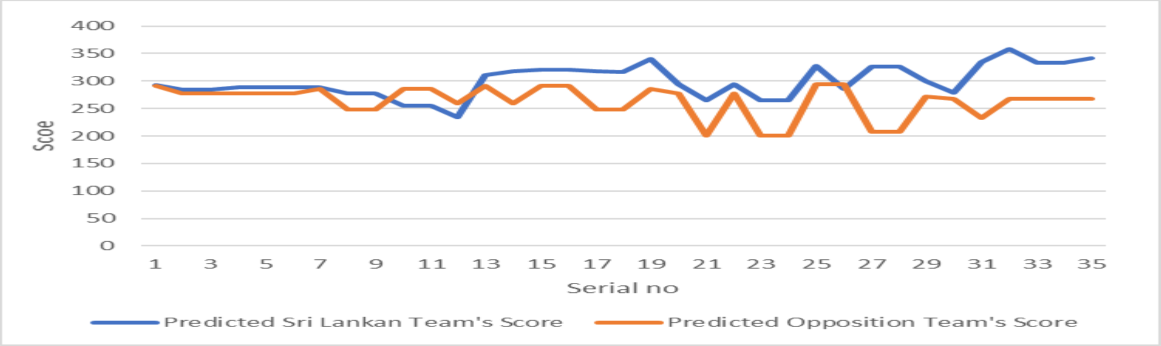


Figure 11: The comparison between the predicted Sri Lankan teams' scores and opposition teams' predicted scores

REFERENCES

1. H. Perera, "Cricket Analytics," Ph.D. thesis, Simon Fraser University, 2015.
2. S. Mukherjee, "Quantifying individual performance in Cricket - A network analysis of Batsmen and Bowlers," Physica A: Statistical Mechanics and its Applications, 393, 2012.
3. H. H. Lemmer, "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, 2008.
4. D. Bhattacharjee and H. Saikia, "On Performance Measurement of Cricketers and Selecting an Optimum Balanced Team," International Journal of Performance Analysis in Sport, 14(1), 2014.
5. P.J. Van Staden, "Comparison of cricketers' bowling and batting performances using graphical displays," Current Science, 96:764–766, 2009.
6. P.J. Bracewell and K. Ruggiero, "A parametric control chart for monitoring individual batting performances in cricket," J Quant Anal Sports, 5(3): 1–19, 2009.
7. S. Iyer and R. Sharda, "Prediction of athletes performance using neural networks: An application in cricket team selection," Journal of Expert Systems with Applications, 36(3),4161-5752 ,2009.

DOI: 10.37200/IJPR/V24I4/PR201034
Received: 22 Oct 2019 | Revised: 13 Sep 2019 | Accepted: 15 Jan 2020

8. Thakare, I. Sachin, S.R. Suyal and K.Y. Pandav, “Performance Evaluation for Sports Team Selection Using Data Mining Techniques,” AADYA-Journal of Management and Technology (JMT) 5,102-108, 2015.
9. G. Sharp, W. Brettenny, J. Gonsalves, M. Lourens and R.A. Stretch, “Integer optimisation for the selection of a Twenty20 cricket team,” Journal of the Operational Research Society, 62(9), 2011.
10. G.R. Amin & S.K. Sharma, “Cricket team selection using data envelopment analysis,” European Journal of Sport Science, 14:suppl 1, S369-S376, 2014.
11. P. K. Singh and M. Ahmad, “Performance Prediction of Players in Sports League Matches,” International Journal of Science and Research (IJSR), 4(1), 2015.
12. R. Tibshirani, G. Walther and T. Hastie, “Estimating the Number of Clusters in a Data Set via the Gap Statistic”, Royal Statistical Society, 411-423, 2001.
13. K. Raj and P. Padma, “Application of association rule mining: A case study on team India,” International Conference on Computer Communication and Informatics (ICCCI), 1-6, 2013.
14. Quang Vinh Tran, Phuong Hong Le, Trung Quang Vo. "Quality Assessment in Systematic Reviews: A Literature Review of Health Economic Evaluation of Hepatitis Studies." Systematic Reviews in Pharmacy 8.1 (2017), 52-61. Print. doi:10.5530/srp.2017.1.10
15. Kak, S.From the no-signaling theorem to veiled non-locality(2014) NeuroQuantology, 12 (1), pp. 12-20.
16. Boyer, R.W.Unless we are robots, classical and quantum theories are fundamentally inadequate(2014) NeuroQuantology, 12 (1), pp. 102-125.

DOI: 10.37200/IJPR/V24I4/PR201034
Received: 22 Oct 2019 | Revised: 13 Sep 2019 | Accepted: 15 Jan 2020