A machine learning-based approach to analyse player performance in T20 Cricket Internationals

Yash Jadwani*, James Denholm-Price**, Gordon Hunter***

School of Computer Science and Mathematics, Kingston University, KT1 2EE, UK
* yash.jadwani1998@gmail.com
** J.Denholm-Price@kingston.ac.uk
*** G.Hunter@kingston.ac.uk

Abstract

Cricket is one of the most-followed sports around the World. T20 is a short version of the game, growing in popularity over the past 20 years due to high profile tournaments such as the Indian Premier League. There is much demand for analysis of player performance, but traditional measures of this – batting and bowling averages, strike and economy rates - have limitations. We created a novel role-based performance metric using machine learning, allowing comparisons between players with similar roles in different teams. Using *ESPNCricinfo* data on T20 international matches, we calculated these new and the traditional performance metrics. Clustering was used to find "natural" classes of player types, a Random Forest classifier employed to identify the features most indicative of each cluster then PCA used to obtain the new performance indicator. Finally, we compared the results of our novel approach to the classical player performance metrics and player classifications provided by experts.

1 Introduction

T20 cricket is a fast-paced and exciting format of the game that demands players to score runs quickly and take wickets within a short span of time. However, the current ranking system for T20 cricket is not a fair and accurate assessment of T20 players. This is because the current system relies on traditional metrics, such as players' career batting and bowling averages, strike rates and economy rates, without considering the distinct roles players undertake within a T20 team. Sports analytics has benefited greatly from the development and acceptance of machine learning methods (Deep et al., 2016). The traditional player performance metrics can help with these comparisons, but it has been well-documented (e.g. Attanayake and Hunter 2015) that each of these has limitations. This paper proposes novel performance metrics and a ranking system that takes into account player roles. The proposed system would assign different weights to different batting and bowling roles, based on the importance of those roles to a T20 team. This would allow for a more accurate assessment of T20 players, as it would take into account the different ways that players can contribute to a team. The proposed ranking system would be a valuable addition to the analytics framework for T20 cricket. It would foster a fairer evaluation process by capturing the diverse ways players can contribute to a team's success. This improvement would be beneficial to all stakeholders, including players, coaches, the media, and sports fans.

2 Related previous work

Traditional methods of comparing cricket performances rely on batting and bowling averages. However, these metrics have limitations as they oversimplify the evaluation process and don't account for the different roles played by players. For example, one of the few players to "average" over 100 in an English First Class season was the Australian bowler W.A. "Bill" Johnston in 1953, who scored 102 runs in 17 innings, but "averaged" 102.00 due to only being out once! In the context of shorter format cricket, where the primary objective for batters is to score runs quickly and for bowlers to concede as few runs as possible per over, traditional metrics such as batting and bowling averages may not provide a complete picture and "strike rates" and "economy rates" are frequently used instead. However, these also have their shortcomings, For example, the Indian fast bowler Zaheer Khan had the apparently excellent batting strike rates of 73.46 and 130.00 in one-day and T20 internationals respectively, but averages only 12.00 in the former and 6.50 in the latter. Although he scored runs quickly, he

didn't score many. It is important to recognize that players have different roles within the team, such as opening batsmen are expected to score more runs than specialist bowlers, who are expected to bowl economically. To assess player performances in shorter format cricket, we should use multiple metrics to consider these various aspects of the game and each player's role in the team.

The cricket commentator and former Australian player Eddie Cowan (2011), following a suggestion by former Australia batting star Mike Hussey, suggested using "magic numbers" that find the sum of a batter's average and strike rate to evaluate their efficiency. Attanayake and Hunter (2015) instead proposed combining runs scored (or conceded), wickets lost (or taken) and balls faced (or bowled) in a multiplicative model, identifying previous metrics as special cases of this. Damodaran (2006) presented a Bayesian technique to deal with not-out scores in cricket as an alternative to batting average. Lemmer (2002) proposed a bowling method called CBR, combining traditional metrics using a formula: CBR = 3[1/Bowling Average + 1/Economy + 1/Strike Rate]. In subsequent papers, Lemmer analyzed bowling and batting performance, including T20 internationals, where new metrics were needed. He developed performance indicators and a formula to evaluate batters, considering their scores when out and not out. Spencer et al. (2016) used k-means clustering and Random Forest to cluster player/team profiles in the Australian Football League.

Deep et al. (2016) created DPI, a machine learning based approach to evaluate T20 players. It uses traditional metrics like economy, strike rate, and average, as well as new indicators like hard hitter, finisher, and running between the wickets (RWB). Deep et al. (2022) proposed DPPI, a more complex approach that considers player's form and role in the team. It uses predefined KPIs and adds new ones like Boundary index and Big Innings index.

3 Role-based player performance analysis in T20Is

The Player Performance Model developed by McHale et al. (2012) has proven successful in assigning a single score to players, irrespective of their specialty, based on their contributions to winning performances in the premier league and championship, the top two divisions of English football. Our study aimed to achieve the following objectives:

- Build a statistical index devoid of subjective opinions and understandable.
- Compare and evaluate players based on their respective roles within the team.
- Strike a balance between model simplicity and complexity.
- Prioritize runs for batters and wickets for bowlers when creating Key Performance Indicators (KPIs), as these represent the primary objectives for each player.

In our study, we conducted preliminary analysis, exploratory data analysis, and outlier removal to extract Key Performance Indicators (KPIs) for batters and bowlers. The assignment of player roles was based on the derived KPI values, without considering specific roles during the KPI establishment phase.

1	Boundaries Per Ball	Total boundaries (4s + 6s) / Total balls faced
2	Boundary Index	Total boundaries (4s + 6s) / Total innings batted
3	Finishing Index	Number of times batter remained not out / Total innings
4	Runs Without Boundary Index	Number of runs scored without boundaries / Total innings
5	Big Match Index	(2 * Centuries + half centuries) / Total innings batted

Table 1: Extracted KPIs for Batters

1	Balls Bowled per Innings	Number of balls bowled / Total innings bowled in
2	Wickets Index	Number of wickets taken / Total innings bowled in
3	Big Impact Index	Number of times bowler took 3 wickets or more / Total innings
4	Short Impact Index	Wickets taken without Big Impact Innings / Total innings
5	Runs Index	Number of Runs Conceded by bowler / Total innings bowled in

Table 2: Extracted KPIs for Bowlers

We cleaned the data and applied normalization techniques to ensure unbiased clustering. K-means clustering was used to group players with similar roles. This generated a target vector that was used for further steps. To determine the importance of each characteristic in relation to player roles, classification algorithms such as one-vs-all along with Random Forest were used. The feature importance values were multiplied by the normalized KPIs to derive Role-Based scores using Machine Learning (RBML) using a Random Forest classifier.

Additionally, Principal Component Analysis (PCA) was conducted on the normalized data, considering the correlations between the KPIs. The coefficients of the first component were used as feature importance values, which were multiplied by the normalized KPIs for comprehensive performance evaluation.

4 Data used and design of experiments

This study includes both active and retired players. We have used the dataset of 4000 entries from ESPNCricinfo, 2000 each for batters and bowlers. Data was gathered starting with the first T20 International match and continuing through April 17th, 2022. Preliminary metrics (PM) calculated for batters and bowlers using formulas by Basevi et al. (2007). PM for batters was (Average * Strike rate) / 100, and for bowlers, it was (Average * Economy) / 6. Prelim Rankings (PR) were assigned based on criteria such as PM values and runs for batters, and PM values and wickets for bowlers. The rankings included categories like Best, Good, Average, and Poor.

Criteria	Ranking
PM > 30 and runs >= 500	Best
PM > 30 and runs < 500	Good
PM between 20 and 30	Good
PM between 10 and 20	Average
PM < 10	Poor
Runs < 100	Poor

Table 3: Prelim Ranking Criteria for Batters

Criteria	Ranking
PM < 30 and Wickets >= 25	Best
PM < 30 and Wickets < 25	Good
PM between 30 and 40	Good
PM between 40 and 65	Average
PM > 65	Poor
Wickets < 5	Poor

Table 4: Prelim Ranking for Bowlers

4.1 Results and findings

After ranking the players, we extracted key performance indicators (KPIs) and applied feature scaling to mitigate possible machine learning algorithm bias. Subsequently, we utilized K-means clustering to determine player roles, treating these roles as the target vector for classification algorithms. To determine the importance of each KPI for each role, we employed a one-vs-all approach alongside a random forest classifier.

We identified four distinct batter roles: Specialist Batters (SB), Finishers (F), Floaters (FL), and Lower Contribution Batters (LCB). SB players excel in all KPIs except the finishing index, F players demonstrate quick scoring abilities while remaining not out, FL players exhibit high averages, good strike rates, and significant impact in big innings, while LCB players, with limited opportunities, focus on concluding innings strongly.

KPIs (Batters)	Feature importance is determined to identify the key features for each role			
	Specialist Batters (SB)	Finishers (F)	Floaters (FL)	Lower Contribution Batters (LCB)
Average	0.266	0.084	0.171	0.152
Strike Rate	0.065	0.212	0.075	0.383
Boundary Per Ball	0.061	0.166	0.074	0.199
Boundary Index	0.244	0.149	0.276	0.161
Finishing Index	0.027	0.158	0.096	0.027
Runs Without Boundary Index	0.085	0.176	0.186	0.054
Big Match Index	0.251	0.055	0.122	0.025

Table 5: Role-based feature importance for batters

Four distinct roles were identified for bowlers: Specialist Bowlers (SB) excel in all key performance indicators (KPIs), Short Performance Bowlers (SPB) have high averages and strike rates but concede more runs and have a lower wicket index than SB, Impact Bowlers (IB) maintain better economy rates than SPB while effectively restricting runs but bowl fewer overs than SB, and Lower Contribution Bowlers (LCB) bowl fewer overs than expected and take fewer wickets in comparison.

KPIs (Bowlers)	Feature importance is determined to identify the key features for each role			
	Specialist Bowlers (SB)	Short Performance Bowlers (SPB)	Impact Bowlers (IB)	Lower Contribution Bowlers (LCB)
Average	0.092	0.128	0.213	0.072
Strike Rate	0.092	0.06	0.104	0.061
Economy	0.038	0.122	0.082	0.052
Balls Bowled per Innings	0.204	0.233	0.141	0.399
Wicket Index	0.279	0.141	0.104	0.156
Big Impact Index	0.027	0.013	0.005	0.01
Short Impact index	0.199	0.085	0.073	0.145
Runs Index	0.068	0.218	0.278	0.105

Table 6: Role-based feature importance for bowlers

After that, we applied Principal Components Analysis (PCA) to verify the performance of clustering, reducing the data dimensionality, and obtained a performance indicator by multiplying the coefficient values of the first component with the KPIs. Although not role-based, this performance indicator allows us to evaluate the success of our supervised and unsupervised learning experiments.

The scatter plots below show the PCA components and group the points based on their assigned roles from clustering. For batters, higher values on the First component indicate better batters. For bowlers, lower values on the First component indicate better bowlers. Both plots effectively represent all the clusters, demonstrating the success of K-Means clustering in grouping the points.

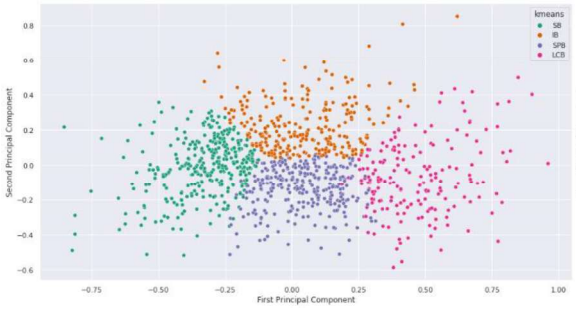

Figure 1: PCA for Batters

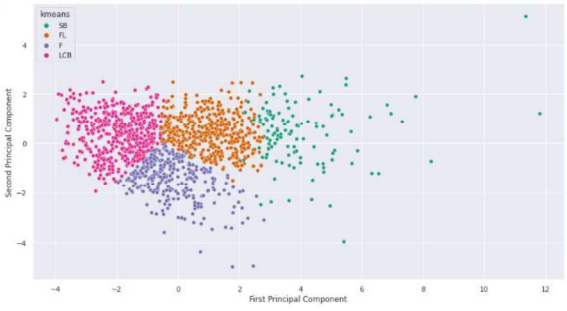

Figure 2: PCA for Bowlers

In summary, the scatter plot provides a visual representation of the role-based performance for both batters and bowlers, showcasing the effectiveness of K-Means clustering in grouping the points and highlighting different aspects of performance in cricket.

Player Name (Country)	Role given by K-means clustering	Preliminary Score	Prelim Rank	RBML Scores
M Hayden (AUS)	Specialist Batter	73.874	Good	2.903
Virat Kohli (IND)	Specialist Batter	70.900	Best	2.494
RA Jadeja (IND)	Finisher	27.036	Good	0.227
MK Pandey (IND)	Floater	55.897	Best	1.015
KC Sangakkara (SL)	Floater	37.539	Best	0.983
A Rashid (IND)	Specialist Bowler	27.479	Best	0.62
T Natarajan (IND)	Specialist Bowler	22.123	Good	0.771
LS Livingstone (ENG)	Short Performance Bowler	23.235	Good	0.621
Yuvraj Singh (IND)	Short Performance Bowler	20.968	Best	0.60
M Theekshana (SL)	Impact Bowler	29.377	Good	0.673
Sohail Tanvir (PAK)	Impact Bowler	32.214	Good	0.653

Table 7: Some players, along with their roles, RBML scores, and preliminary ranks

5 Discussion, conclusions and future work

Before clustering, the dataset was shuffled, and a subset of 98 batters and 82 bowlers was selected. This subset included Prelim Ranks (Best, Good, Average, Poor). Among the 'Poor' batters, 32 were classified as LCB, 12 as Finishers, 5 as Floaters, and 1 as SB. None of the 'Good/Best' batters were LCB, but there were more SB batters in the 'Good' ranks than 'Best'. Notably, ED Silva, initially ranked 'Poor', was classified as SB, suggesting the need for a broader performance evaluation. Among the 'Good' bowlers, 22 were SB, 8 were SPB, 7 were LCB, and 6 were IB. Out of the 'Best' bowlers, 9 were SB, while 1 each were SPB and IB. This highlights the importance of role-based classification over traditional rankings for bowlers, and that traditional metrics such as average, economy, and strike rate should not be the only performance indicators for a player.

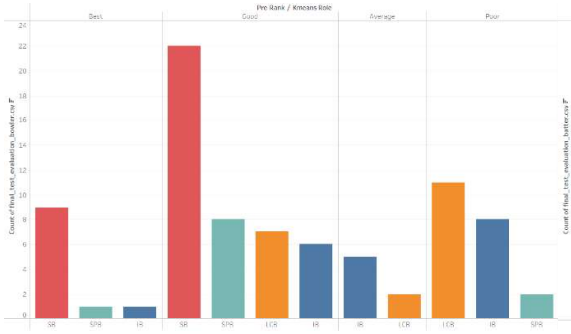

Figure 3: Results for Batters by category

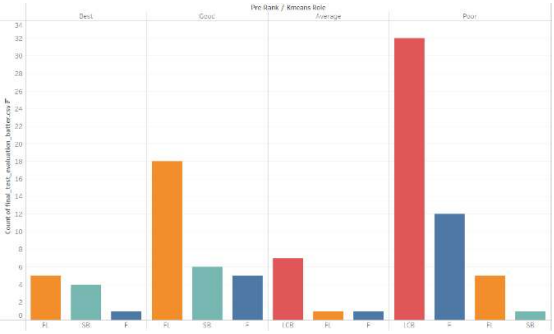

Figure 4: Results for Bowlers by category

In conclusion, the integration of k-Means clustering, classification algorithms, PCA, and role-based ranking provides valuable insights into player performance in T20 cricket. Furthermore, by comparing RBML/PCA ranks with roles assigned through clustering using PR rank and scores, we found that a player's performance cannot be solely determined by their average, strike rate, runs scored, or wickets taken. This approach offers a more comprehensive understanding of player capabilities and performance in the dynamic and fast-paced nature of T20 cricket.

Player evaluation in sports analytics is a growing field. This study has considered role-based clustering and predicted the feature importance of KPIs. However, recent form and pitch conditions can also play a role in player performance. In the future, the model developed by this study will remain valid, and one should be able to improve its performance by obtaining more data.

References

[1] Attanayake, D. and Hunter, G. , 2015. *Probabilistic Modelling of Twenty-Twenty (T20) Cricket : An Investigation into various Metrics of Player Performance and their Effects on the Resulting Match and Player Scores.* In Proceedings of 5th International Conference on Mathematics in Sport (Kay, Owen, Halkon and King., Eds.), Loughborough, U.K., June 2015.

[2] Cowan, E., 2011. *What's so great about a batting average* ? (On-line) https://www.espncricinfo.com/story/ed-cowan-on-more-meaningful-statistics-in-cricket-543061

[3] Spencer, B., Morgan, S., Zeleznikow, J. and Robertson, S. 2016. *Clustering team profiles in the Australian football league using performance indicators*, in: The 13th Australasian Conference on Mathematics and Computers in Sport, Melbourne, Australia, July 2016

[3] Basevi, T. and Binoy, G., 2007. *The world's best Twenty20 players*. [online] ESPNCricinfo. Available at: <https://www.ESPNCricinfo.com/story/the-world-s-best-twenty20-players-311962>

[4] Damodaran, U., 2006. *Stochastic dominance and analysis of ODI batting performance : The Indian cricket team 1989-2005*. Journal of Sports Science & Medicine, Vol. 5, pp. 503-508.

[5] Deep Prakash, C. and Verma, S., 2022. *A new in-form and role-based Deep Player Performance Index for player evaluation in T20 Cricket*. Decision Analytics Journal, 2, p.10002

[6] Deep Prakash, C., Patvardhan, C. and Singh, S., 2016. *A new Machine Learning based Deep Performance Index for Ranking IPL T20 Cricketers.* International Journal of Computer Applications, 137 (10) 5(2), pp. 42-49. Doi : 10.5120/ijca2016908903

[7] Lemmer, H., 2002. *The combined bowling rate as a measure of bowling performance in cricket*. South African Journal for Research in Sport, Physical Education and Recreation, 24(2).
[8] McHale, I., Scarf, P. and Folker, D., 2012. *On the Development of a Soccer Player Performance Rating System for the English Premier League*. Interfaces, 42 (4), pp. 339-351.