Learning Player-specific Strategies Using Cricket Text Commentary

Thesis submitted in partial fulfilment of the requirements for the award of the degree of

Doctor of Philosophy

in

Computer Science and Engineering

by

Swarup Ranjan Behera

Under the supervision of

Dr. Vijaya Saradhi Vedula

Department of Computer Science and Engineering
Indian Institute of Technology Guwahati
Guwahati - 781039 Assam India

August, 2021

Dedicated to

Dadu, Mama, Bapa, Maa, and Saroj.

It is from your love, care, and support, I derive my strengths.

Acknowledgements

In July 2015, I embarked on an exciting journey: a Ph.D. at IIT Guwahati. I have been fortunate enough to have met several people who have helped me and contributed to my success in various ways during my eventful journey at IIT Guwahati.

First and foremost, I would like to express my sincere gratitude to my supervisor Dr. Vijaya Saradhi Vedula. Thank you for your valuable guidance, inspiration, and encouragement throughout my research work. I imagine I did not always make it easy for you to manage me, but I think you hit the sweet spot between giving me space to work independently and correcting my course whenever I was slacking off with too little progress. I especially appreciate your guidance on producing well-written texts and clear presentations. It has been a memorable experience to work under your supervision.

I want to thank my doctoral committee members, Dr. Amit Awekar, Dr. Rashmi Dutta Baruah, Dr. Purandar Bhaduri, and Dr. Suresh Sundaram, for their insightful comments and critical questions. I also thank my collaborators, Dr. Amit Awekar, Pratik Agrawal, Harshith Reddy Padigela, and Rahul R Huilgol.

I wish to thank the MHRD, Government of India, for supporting my research by providing financial assistantship throughout the Ph.D. program. I also thank the Department of CSE, IIT Guwahati, for providing different facilities to carry out my research work.

I may not have always shown it, but I am grateful to have met excellent teachers like Dr. Vijaya Saradhi Vedula, Dr. Rakesh Mohanty, Mr. Sanjib K. Nayak, Mr. Suresh Srichandan, and Mr. A.K. Nayak. Thank you for all the wisdom you shared and for setting a great example. I owe my gratitude to my alma mater VSSUT Burla and IIT Guwahati, for everything.

I am fortunate to have several wonderful friends who were there to support and celebrate with me during various events of this journey. I particularly thank Shilpa, Hema, Akash, Nayantara, and Saroj for helping me out in my research work on countless occasions. I will surely miss my friends - Chitta, Ranjan, Debashish, Tushar, Biswajit, Dilip, Trushna, Ghalib, and many others. Some friends also deserve a separate shoutout. Saroj, I am convinced that nobody could ask for a better friend than you. I think we hit the sweet spot between messing around and working hard. I especially like our dynamic because we appreciate that there is more to life than research and that sometimes exceptional efforts must be made when confronted with conference deadlines. I sincerely thank you for your support and the fun memories we made over the years.

Of course, my Ph.D. colleagues are not the only people I wish to thank. To my friends Prabhat, Archana, Amit, Abhay, Poonam, Jerry, and Ankur, I am glad that I went through my masters with you and that our friendship has continued beyond it. To my friends Raj, Akash, Shekhar, Ashutosh, and Mukesh, I am glad that I went through my bachelor's with you. To my friends Nayan, Jaga, and Damru, I am thankful to have spent my school days with you. To my uncles Jhas and Yash and aunts Taru, Anu, and Munu, I am thankful to have spent most of my happy days with you.

It goes without saying that this journey would not have been possible without the persistent support, unconditional love, and profound encouragement of my mother and father. I fall short of words to express my gratitude to them.

August 5, 2020 Swarup Ranjan Behera

Declaration

I certify that

- The work contained in this thesis is original and has been done by myself and under the general supervision of my supervisor.
- The work reported herein has not been submitted to any other Institute for any degree or diploma.
- Whenever I have used materials (concepts, ideas, text, expressions, data, graphs, diagrams, theoretical analysis, results, etc.) from other sources, I have given due credit by citing them in the text of the thesis and giving their details in the references. Elaborate sentences used verbatim from published work have been clearly identified and quoted.
- I also affirm that no part of this thesis can be considered plagiarism to the best of my knowledge and understanding and take complete responsibility if any complaint arises.
- I am fully aware that my thesis supervisor is not in a position to check for any possible instance of plagiarism within this submitted work.

August 5, 2020

Swarup Ranjan Behera

Department of Computer Science and Engineering
Indian Institute of Technology Guwahati
Guwahati - 781039 Assam India

Dr. Vijaya Saradhi Vedula
Associate Professor
Email : saradhi@iitg.ac.in
Phone : +91-361-258-2356

Certificate

This is to certify that this thesis entitled "**Learning Player-specific Strategies Using Cricket Text Commentary**" submitted by **Swarup Ranjan Behera**, in partial fulfilment of the requirements for the award of the degree of Doctor of Philosophy, to the Indian Institute of Technology Guwahati, Assam, India, is a record of the bonafide research work carried out by him under my guidance and supervision at the Department of Computer Science and Engineering, Indian Institute of Technology Guwahati, Assam, India. To the best of my knowledge, no part of the work reported in this thesis has been presented for the award of any degree at any other institution.

Date: August 5, 2020
Place: Guwahati

Dr. Vijaya Saradhi Vedula
(Thesis Supervisor)

Abstract

In recent times, as sports get more competitive than ever, players and teams are looking for ways to get an edge over their rivals. The progressive trend of analyzing vast amounts of data has also emerged in cricket, as it brings a significant advantage against other teams. A massive amount of information in the form of scorecards, audio commentary, video broadcasts, and tracking data is generated in every match. Various graphical representations and statistical summaries are obtained from these data to build player-specific strategies. The graphs and statistics summarize player's - batting overview (wagon wheel, ground map, batting average, and strike rate), bowling overview (pitch map, bowling economy, and bowling average), and fielding overview (field position map). While interesting, the focus of these analyses has primarily been at an aggregate level. They capture the game's play at a macroscopic level and do not attend to the minute details. For example, the batting average and the strike rate tell us about the batsman's overall statistics, but not the finer details like how a batsman plays under different conditions. However, the team coach and team management require the knowledge of minute details of the game to build player-specific strategies.

Devising player-specific strategy in cricket needs a microscopic understanding of players' strengths and weaknesses. However, there is no specific computational method for this task. This thesis aims to build computational models that mine rules related to the strengths and weaknesses of a player, thereby helping devise player-specific strategies. The challenge lies in identifying a suitable dataset, defining the strength rule and weakness rule, identifying the learning algorithm, and validating obtained rules. We address these challenges systematically and make the following contributions.

To learn players' strengths and weaknesses, we propose the use of unstructured data, namely cricket text commentary. We have constructed a large-scale dataset consisting of more than one million short text commentaries and over five thousand text commentaries for external factor data spanning over thirteen years (May 2006 to April 2019). We have identified challenges in mining cricket-related text data and addressed those challenges. Finally, we have obtained fine-grained information about each player from the short text commentary data and represented it using domain-specific features, namely batting features and bowling features. Similarly, the external factors affecting the game are represented using external features.

The domain-specific features are represented as a confrontation matrix of batting features and bowling features. In this matrix, rows correspond to batting features, and columns correspond to bowling features. Every element in this matrix records how the batsman confronted the bowlers. For example, how many times the batsman has attacked short-length deliveries. Similarly, the domain-specific features and the external features are represented in a confrontation matrix of batting/bowling features and external features. In this matrix, rows correspond to the batting/bowling features, and columns correspond to the external features. Every element in this matrix records how the batsman or bowler performed on a given external feature. For example, how many times the batsman has faced good length deliveries on day-1 of the Test matches.

In the first approach, we propose to learn strength and weakness rules of cricket players using text commentary data. The presented approach proceeds in three steps for each player. In the first step, we introduce computationally feasible definitions of *strength rule* and *weakness rule*. The

second step employs a dimensionality reduction method on the confrontation matrix of batting features and bowling features to construct semantic relations between the batting and the bowling features. These relations are plotted using biplots in the third step, and human-readable strength and weakness rules are extracted. The proposed approach goes beyond the number of runs and balls and considers all the actions performed by a batsman and a bowler on each delivery. The obtained rules are validated using intrinsic and extrinsic methods. Additionally, baseline comparisons are made using word clouds and association rule mining techniques. Several case studies show that the proposed approach can identify players' strengths and weaknesses in a single match as well as throughout their careers.

We also propose an approach to learn temporal changes in strength and weakness rules. For a given player, we use year-wise confrontation matrices of batting features and bowling features to model the data as a three-dimensional confrontation tensor in which batting features, bowling features, and time (in years) are the dimensions. A dimensionality reduction method is then employed on this tensor to construct semantic relations between the three dimensions mentioned above. These relations are visualized in a line plot to show the year-wise changes in strength and weakness rules. The proposed approach is proven to identify temporal changes in a player's strengths and weaknesses.

External factors, such as the playing condition or the match situation, are outside the scope of the game and yet influence a player's strengths and weaknesses. We propose an approach to learn strength and weakness rules of cricket players in the presence of these external factors. In this approach, we first introduce computationally feasible definitions of *strength rule* and *weakness rule* in the presence of external factors. The second step employs a dimensionality reduction method on confrontation matrices formed by the batting/bowling features and the external features to construct semantic relations between these features. Triplots are used to plot these relations to observe the strength and weakness rules in the presence of external factors. We show that the proposed approach can identify players' strengths and weaknesses in the presence of external factors influencing the game.

Lastly, we propose an approach to find players having similar strength rules or similar weakness rules. For this, the strength vectors and weakness vectors of all the batsmen (or bowlers) are obtained from their strength rules and weakness rules, respectively. The t-SNE algorithm is used to visualize these high-dimensional vectors in a two-dimensional plot in which batsmen (or bowlers) having similar strength rules or similar weakness rules are placed closer. Several case studies show that the proposed approach can identify similar batsmen and similar bowlers based on their strengths and weaknesses.

Contents

List of Figures

List of Algorithms

List of Tables

List of Symbols

Symbols	**Description**
N	Technical Confrontation Matrix
n	Sum of Elements of Technical Confrontation Matrix
r	Row Masses of Technical Confrontation Matrix
D_r	Row Diagonal Matrix of Technical Confrontation Matrix
c	Column Masses of Technical Confrontation Matrix
D_c	Column Diagonal Matrix of Technical Confrontation Matrix
P	Correspondence Matrix
A	Standardized Residuals
F	Principal Components of Rows of Technical Confrontation Matrix
$F^{'}$	First Two Principal Components in F ($F_{m \times 2}$)
G	Principal Components of Columns of Technical Confrontation Matrix
$G^{'}$	First Two Principal Components in G ($G_{n \times 2}$)
α	Amount of Deviation from Independence of Events
$\langle F, G \rangle$	Inner Product of F and G
Δ_{ij}^{2}	Sum of Squared Residuals
$\mathcal{N}$	Technical Confrontation Tensor
$\mathcal{P}$	Tensor of Relative Frequencies
$\mathcal{A}$	Residual Tensor
H	Principal Coordinates of Column-tubes

$\mathcal{X}^2$	Sum of the Squared Distance
X	External Confrontation Matrix
Q	Projection Matrix
A^*	Constrained Correspondence Matrix
TCM_{BAT}	Batsman's Technical Confrontation Matrix
TCM_{BOWL}	Bowler's Technical Confrontation Matrix
$TCM_{BAT}Year_n$	Batsman's Technical Confrontation Matrix on her n^{th} Year of Play
$TCM_{BOWL}Year_n$	Bowler's Technical Confrontation Matrix on her n^{th} Year of Play
TCT_{BAT}	Batsman's Technical Confrontation Tensor
TCT_{BOWL}	Bowler's Technical Confrontation Tensor
ECM_{BAT}	Batsman's External Confrontation Matrix
ECM_{BOWL}	Bowler's External Confrontation Matrix
$ECM_{BAT}Day$	Batsman's External Confrontation Matrix for the External Feature 'Day'
$ECM_{BOWL}Day$	Bowler's External Confrontation Matrix for the External Feature 'Day'
$SV_{BAT}P_n$	Strength Vector of n^{th} Batsman
$WV_{BAT}P_n$	Weakness Vector of n^{th} Batsman
$SV_{BOWL}P_n$	Strength Vector of n^{th} Bowler
$WV_{BOWL}P_n$	Weakness Vector of n^{th} Bowler
$TCM_{BAT}PLAYER_n$	Batting Technical Confrontation Matrix of n^{th} Player
$TCM_{BOWL}PLAYER_n$	Bowling Technical Confrontation Matrix of n^{th} Player

List of Abbreviations

Terms	**Abbreviations**
ODI	One Days International
T20I	Twenty20 International
LBW	Leg Before Wicket
ICC	International Cricket Council
DL	Duckworth and Lewis
HMM	Hidden Markov Model
MFCC	Mel Frequency Cepstral Coefficients
TF-IDF	Term Frequency and Inverse Document Frequency
FD	Feature Definition
EFD	External Feature Definition
TCM	Technical Confrontation Matrix
ECM	External Confrontation Matrix
CA	Correspondence Analysis
SVD	Singular Value Decomposition
CP	Commonality Percentage
ARM	Association Rule Mining
TWCA	Three-Way Correspondence Analysis
TCT	Technical Confrontation Tensor
CCA	Canonical Correspondence Analysis

T-SNE	T Stochastic Neighbour Embedding
SR	Strength Rule
WR	Weakness Rule

Glossary of Terms

Terms	**Description**
Cricket	Cricket is a bat-and-ball game played between two teams of eleven players each on an oval field.
Pitch	The playing area on the field; a rectangular strip.
Batsman	A player who hits the ball to score runs (analogous to a batter in baseball).
Bowler	A player who bowls the ball towards the batsman (analogous to a pitcher in baseball).
Fielder	A player who stops the ball hit by the batsman in the field (analogous to a fielder in baseball).
Test Match	A format of a cricket match, which is played for a maximum of five days and comprises four innings, two innings per team.
Text Commentary	Live text commentaries are written narratives that give a detailed description of a ball-by-ball account of the game while it is unfolding.
Short Texts	Short texts are usually less than a few hundred characters long. Examples of short texts are - microblogs, online chat records, product reviews, and blog comments.
Short Text Commentary	Short text commentary encodes the technical factors corresponding to batting and bowling, which describe the game in greater detail and help in a microscopic understanding of the game.
External Factor Data	A part of the text commentary that describes the external factors like playing conditions and match situations, which are not under the player's control but are crucial to outdoor team sports like cricket.

1

Introduction

In today's world, as sports are getting more competitive than ever, players and teams are looking for ways to get an edge over their rivals [1]. Progressive trend of analyzing vast amounts of data has also emerged in cricket, as it brings a significant advantage against other teams. Cricket is known for recording every detail. A massive amount of data in the form of scorecards, video broadcasts, audio commentary, text commentary, and tracking data are generated in every cricket match. Three particularly prominent types of data in cricket are:

- **Box-score Data:** Box-score data are the discrete data referencing in-game events. These are the summary statistics such as the number of runs scored, the number of balls faced, the fall of wickets, the number of extras. Box-score data are generated by companies such as Opta Sports [2], ESPNcricinfo [3], and CricBuzz [4].

- **Unstructured Data:** Examples of unstructured data in cricket are video broadcast and text commentary. The video data is provided by broadcasting companies such as Sky Sports [5]. Cricket text commentary data is provided by companies such as ESPNcricinfo [3], Opta Sports [2], and CricBuzz [4]. In addition to the above two, audio commentary, social media posts, and news articles add to the multiple sources of unstructured data.

- **Tracking Data:** Recent developments in tracking and sensing technologies make it possible to obtain spatio-temporal information about players and equipment in real-time during play. Tracking data are continuous spatio-temporal motion data generated by multi-camera tracking systems such as Hawk-Eye and SportVU. These data enable coaches and broadcasters to uncover new insights into every shot, run, and wicket. Examples of companies that record tracking data are STATSports [6] and Opta Sports [2].

The cricket data mentioned above can be used for tasks such as:

- **Target Re-setting and Outcome Prediction:** When a match is interrupted due to bad weather, target re-setting plays a major role. Duckworth and Lewis [7] proposed a method

(DL-method) for target re-setting, which is adopted by the International Cricket Council. To account for the recent changes in scoring, the DL-method is updated by Stern [8]. Bailey and Clarke [9] developed a predictive model of the game's outcome by employing the DL-method, and a linear model was used for fitting the resulting target scores. Correlation of winning a game to different batting combinations and run rate was identified by Allsopp and Clarke [10].

- **Player Performance Analysis:** Iyer and Sharda [11] employed neural networks to predict cricket players' performance based on their past performances. Das et al. proposed CricVis [12], a web-based visualization system that utilizes box-score data and tracking data to construct visualizations such as pitch maps and stump maps to analyze the bowling overview and batting overview, respectively. Theodoro et al. [13] proposed a Bayesian hidden Markov model for assessing batting in one-day cricket. Camera motion estimation was carried out by Lazarescu et al. [14] to index cricket videos and to classify shots offered by batsmen based on the estimates. Low-scoring shots are classified accurately compared to high-scoring shots.

- **Team Strength and Tactics:** Davis et al. [15] proposed a match simulator that assesses team strength in Twenty20 cricket. Lemmer [16] and Ahmed et al. [17] proposed search algorithms to select teams. Scarf and Akhtar [18] proposed in Test cricket whether teams should declare at various stages of matches and under various circumstances. Scarf et al. [19] used negative binomial distributions to model the runs scored in inning and partnerships during test matches. Morgan et al. [20] predicted where a specific batsman would hit a specific bowler and bowl type in a specific game scenario.

To build player-specific strategies in cricket, various graphical representations and statistical summaries are obtained. These graphs and statistics summarize players' - batting overview (wagon wheel, ground map, batting average, and strike rate), bowling overview (pitch map, bowling economy, and bowling average), and fielding overview (field position map). While interesting, the focus of these analyses has primarily been at an aggregate level. They capture a game's play at a macroscopic level and do not attend to the minute details. For example, (i) batting average and strike rate looks at the overall statistics of a batsman, but not the finer details like how the batsman plays under different conditions, and (ii) tracking data are limited to specific events of the game and are devoid of the rules and context of the game. Experts, including team coaches and team management, on the other hand, require the knowledge of minute details of the game to build player-specific strategies.

Devising player-specific strategy in cricket needs a microscopic understanding of players' strengths and weaknesses while considering the various factors and conditions in play. These conditions are different for batsmen and bowlers and arise due to the opponent's actions and the external factors such as playing conditions and match situations. From a practical point of view, understanding players' strengths and weaknesses is essential due to the following three reasons:

- **Player Selection:** Identified strengths and weaknesses help the coaches in the selection process and in developing team strategies accordingly. Additionally, the batting order and the bowling order can be determined based on the game's situation using strength and weakness information.

- **Performance Monitoring:** It makes batsmen and bowlers aware of their strengths and weaknesses, which is usually more than just a delivery against which they got out or took a wicket. Based on this, players can improve their performances. The coach can also inspect the players' strengths and weaknesses, give pointers on their weaknesses, and monitor how well they implement the advice.

- **Match Preparation:** Understanding the opponent player's strengths and weaknesses offers tactical advantages. The bowlers of a team will benefit from the knowledge of the strengths and weaknesses of batsmen they are up against. Similarly, the batsmen of a team will benefit from the knowledge of the strengths and weaknesses of the bowlers they are up against.

The strengths and weaknesses of cricket players are understood informally by players themselves, coaches, and team management [21, 22]. However, there is no specific computational method to obtain the strengths and weaknesses of cricket players. This work aims to *build computational models to learn the strengths and weaknesses of cricket players*, thus helping devise *player-specific strategies*. Following are the challenges of this work and the steps taken to address these challenges.

- **Data:** Multiple data sources are available such as video recording, audio commentary, text commentary, social media posts, and news articles. An important question is: **which data is credible, requires less storage, yet is detailed enough for player-specific analysis?**

  We identify that cricket text commentary data has advantages over other data to mine for players' strengths and weaknesses. Cricket text commentary is a rich source of fine-grained details about each delivery of the game. The subjective opinions of the commentators about how players performed on each delivery and external factors such as playing conditions and match situations are recorded in the text commentary. Text commentary data has some unique advantages over the other data sources in cricket. Unlike box-score data, minute details of player actions are present in the text commentary data. Compared to video or audio data, it is publically available and requires less storage and computational resources. Unlike social media posts, it is credible, maintains a consistent style, and maps each delivery to a single text commentary. Unlike news articles, it provides a microscopic view with ball-by-ball details of the match. We have collected a large and first-of-its-kind *cricket text commentary* dataset of over one million deliveries, covering 550 international cricket Test matches in the last thirteen years (May 2006 to April 2019).

- **Text Representation:** For the effective representation of text documents, *stemming* and *stop word removal* are performed as preprocessing steps in the traditional information retrieval context. A differentiating factor specific to cricket text commentary is that most of the technical words used in the cricketing domain are stop words in the conventional text mining literature. In addition, the vocabulary size is large, and one delivery of cricket text commentary has only up to 50 words. This leads to *sparsity*. The key question is: **how to represent cricket text commentary to obtain player's strengths and weaknesses?**

We have identified the limitations in employing well-known text representation methods in the context of cricket text commentary. We have extracted domain-specific features from the text commentary data and proposed a confrontation matrix representation that takes into account both batting features and bowling features present in each text commentary.

- **Machine Learning:** Strengths and weaknesses do not have definite and distinct classes. They can neither be viewed as a text categorization problem nor can they be considered a text clustering problem. Therefore the critical question is: **which learning paradigm fits the problem of effectively identifying player's strengths and weaknesses?**

  We have posed the problem of finding strengths and weaknesses as a dimensionality reduction problem in which the relationships between batting features and bowling features are identified in a two-dimensional space. We have obtained simple rules using a graphical representation, which accounts for players' strengths and weaknesses.

- **Validation:** The obtained rules, though easy to interpret, are difficult to validate. No loss function exists which captures the risk associated with each of the obtained rules. Cricket experts' opinion matters the most in judging the derived rules. An important question is: **how to validate the obtained strength and weakness rules?**

  We have carried out the validation of the obtained strength and weakness rules in two distinct ways: *extrinsic* and *intrinsic*. For extrinsic validation, we verify the identified rules against external sources. The main bottleneck is the absence of trustable gold standard data about player's strengths and weaknesses. However, we could get a couple of such resources available in the public domain where domain experts have directly mentioned some cricket players' strengths and weaknesses. For intrinsic validation, we use the k-fold cross validation method, in which the commentary data is divided into multiple training sets and test sets.

1.1 Contributions of the Thesis

The work reported in this thesis attempts to learn player-specific strategies (strengths and weaknesses of cricket players) using publicly available text commentary data. The major contributions made in this thesis are listed herein.

1.1.1 Contribution 1: Representing Cricket Text Commentary Data

We propose to use unstructured data, namely *cricket text commentary*, for constructing player-specific strategies in cricket. The game of cricket has multiple formats, of which we focus on the *Test cricket* format. We have constructed a large-scale dataset consisting of more than one million text commentaries spanning over thirteen years (May 2006 to April 2019). In addition, over 5000 paragraphs of playing condition descriptions are collected for all the 550 matches between May 2006 and April 2019. ESPNcricinfo [3] is selected as the data source. This data collection is a time-consuming and challenging process that requires considerable amounts of domain expertise and experience. Such a dataset can have an essential role in cricket analytics as minute details of each delivery are captured in these text commentaries.

While the text commentary data is a more practical fit than other data sources to capture player-specific strategies, this does not make the task easy. We have identified challenges inherent to this data. We have obtained fine-grained information about each player and conditions of play from the text commentaries and represented it using domain-specific features such as batting features, bowling features, and external features.

1.1.2 Contribution 2: Mining Strength and Weakness Rules of Cricket Players

Knowledge of players' strengths and weaknesses is the key to team selection and strategy planning in any team sport such as cricket. Computationally, this problem is mostly unexplored. Computational methods to identify player's strengths and weaknesses can have a two-pronged impact. First, such methods can assist domain experts in better team selection and strategy planning. Second, such methods are the foundation of future automated strategy planning in cricket. To this end, the following research question is investigated.

RQ: What are the Strength and Weakness Rules of a cricket player? Can these rules be obtained using cricket text commentaries?

Data Representation: We have extracted domain-specific batting features and bowling features from the text commentary data and proposed a confrontation matrix representation in which rows correspond to batting features and columns correspond to bowling features. Every element in this matrix corresponds to how the batsman confronted the bowlers. For example, how many times a batsman has attacked short-length deliveries.

Learning Model: We employ a dimensionality reduction method specific to the discrete random variable case, namely Correspondence Analysis [23, 24, 25] and construct semantic relations between batting features and bowling features. These relations are plotted using biplot [26]. Human readable strength and weakness rules are extracted from the biplot.

Results: We use the proposed approach to mine strength and weakness rules corresponding to 131 batsmen and 129 bowlers who have batted/bowled more than 2000 deliveries. We highlight some of the obtained strength and weakness rules for batsman Steve Smith and bowler Kagiso Rabada. For batsman Steve Smith, we observe that - (i) Steve Smith attacks slow or short-length or the middle-line deliveries (strength rule). (ii) Steve Smith gets beaten on the deliveries that are either swinging or moving-away or moving-in (weakness rule). For bowler Kagiso Rabada, we observe that - (i) batsmen get beaten on the swing deliveries of Kagiso Rabada (strength rule). (ii) batsmen attack the full-length deliveries of Kagiso Rabada (weakness rule). We have also obtained the strength and weakness rules for batsmen against spin and fast bowling separately. For batsman Steve Smith, we observe that - (i) Against fast bowlers, Steve Smith gets beaten on the moving-in deliveries (weakness rule) and attacks full length deliveries (strength rule). (ii) Against spin bowlers, Steve Smith gets beaten on the moving-away deliveries (weakness rule) and attacks short-length deliveries (strength rule).

1.1.3 Contribution 3: Mining Temporal Changes in Strength and Weakness Rules of Cricket Players

Cricket players' strengths and weaknesses are not constant throughout their careers. Over time players evolve in the sense that they work on their weaknesses and overcome them. So it is of interest to find out the traits (strengths or weaknesses) they lost or acquired over time. In particular, the following research question is investigated.

RQ: How do the strength and weakness rules of a player change over time?

Data Representation: Time granularity in cricket is identified in the increasing order as over, session, day, inning, match, series, season, year, or career. The year-wise analysis is our primary focus. For a given player, the year-wise confrontation matrices of batting and bowling features are constructed and modeled as a three-dimensional confrontation tensor in which batting features, bowling features, and time (in years) are captured.

Learning Model: The obtained three-dimensional confrontation tensor is subject to a dimensionality reduction method, Three-way Correspondence Analysis [23, 27] and semantic relations between batting features, bowling features, and time (years) are obtained. These relations are plotted in a line plot to visualize the year-wise changes in strength and weakness rules.

Results: We use the proposed approach to mine the temporal changes in strength and weakness rules corresponding to 131 batsmen and 129 bowlers who have batted/bowled more than 2000 deliveries. We highlight some of the obtained results for batsman Steve Smith and bowler Kagiso Rabada here. For batsman Steve Smith, we observe that - (i) Steve Smith has shown an increase in the trend of attacking *full-length* deliveries between the years 2013 and 2015 (both inclusive). However, in 2016 he struggled on *full-length* deliveries. He has once again shown an increase in the trend of attacking *full-length* deliveries between the years 2017 and 2018 (both inclusive) (year-wise changes in strength rule). (ii) Steve Smith has never exhibited weaknesses on *full-length* deliveries (year-wise changes in weakness rule). For bowler Kagiso Rabada, we observe that - (i) Kagiso Rabada has shown an increase in weakness on full-length deliveries between the years 2017 and 2019 (both inclusive) (year-wise changes in weakness rule). (ii) Kagiso Rabada has shown a decrease in strength on full-length deliveries from 2015 to 2017 (both inclusive) and once again shown an increase in strength on full-length deliveries from 2017 to 2018 (both inclusive) (year-wise changes in strength rule).

1.1.4 Contribution 4: Mining Strength and Weakness Rules of Cricket Players in the Presence of External Factors

In an outdoor team sport like cricket, external factors like playing conditions and match situations are outside the scope of the game and yet influence the gameplay. Different playing conditions (age-of-ball, pitch condition, and weather condition) or match situations (day, inning, and session) suit different players. For example, a bowler can extract high bounce from a grassy, hard, or intact pitch. Match situations also have indirect and sometimes direct effects. A spinner can be thought to have favorable conditions on the last day of the Test match. These external factors influence the techniques batsmen or bowlers exhibit. To this end, the following research question is investigated.

RQ: What are the strength and weakness rules of a player in the presence of external factors influencing the game?

Data Representation: We have extracted batting features, bowling features, and external features from the cricket text commentary data and proposed confrontation matrix representations of- (i) batting and bowling features and (ii) batting/bowling and external features.

Learning Model: We employ a dimensionality reduction method, Canonical Correspondence Analysis [28], on the confrontation matrices to construct semantic relations between batting features, bowling features, and external features. These relations are plotted in a triplot [29] to visualize the strength and weakness rules in the presence of external factors.

Results: We use the proposed approach to mine strength and weakness rules in the presence of external factors, corresponding to 131 batsmen and 129 bowlers who have batted/bowled more than 2000 deliveries. We highlight some of the obtained results for batsman Steve Smith and bowler Kagiso Rabada here. For batsman Steve Smith, we observe that - (i) On the third day of a Test match, bowlers tend to bowl moving-in deliveries to Steve Smith, and he plays them on the back foot, gets beaten, or loses his wicket in those deliveries. (ii) On the first day and second day of a Test match, bowlers tend to bowl moving-away deliveries on the outside off line to Steve Smith, and he plays them on the front foot, defends, or scores no runs. (iii) Steve Smith gets beaten or loses his wicket on green pitches. For bowler Kagiso Rabada, we observe that - (i) When the ball is more than 30 overs old, Kagiso Rabada tends to bowl moving-away deliveries to batsmen, and the batsmen defend or play them to the thirdman area. (ii) Kagiso Rabada tends to bowl moving-away deliveries to batsmen on the grass pitch, and the batsmen defend them.

1.1.5 Contribution 5: Visualization of Similar Players Based on their Strength and Weakness Rules

Grouping similar players based on their strengths and weaknesses yield in-depth insights. It will help devise strategies for team member selection and ordering of the batsman/bowler during a match (batting lineup and bowling lineup). In particular, the following research question is investigated.

RQ: Which players (batsmen or bowlers) have similar strength rules or similar weakness rules?

Data Representation: The strength vectors and weakness vectors of all the batsmen and bowlers were obtained from their strength and weakness rules.

Learning Model: We employ the t-distributed Stochastic Neighbor Embedding (t-SNE) [30] algorithm to visualize these high-dimensional strength vectors and weakness vectors in a two-dimensional plot in which batsmen (or bowlers) having similar strength rule or similar weakness rule are placed closer.

Results: We use the proposed approach to visualize the similar players in a t-SNE plot corresponding to 131 batsmen and 129 bowlers who have batted/bowled more than 2000 deliveries. We highlight some of the obtained results here. Pairs of similar batsmen based on their strength rule are - (i) Virat Kohli and VVS Laxman, (ii) Paul Collingwood and David Warner, and (iii) Michael Clarke and Brad Haddin. Pairs of similar batsmen based on their weakness rule are - (i) Virender Sehwag and Misbah-ul-Haq, (ii) Jonathan Trott and Kevin Pietersen, and (iii) Salman Butt and Ian Bell. Pairs of similar bowlers based on their strength rule are - (i) Josh Hazlewood and Rahat Ali and (ii) Umesh Yadav and Ben Stokes. Pairs of similar bowlers based on their weakness rule are - (i) Brett Lee and Vernon Philander and (ii) Daren Powell, Dwane Bravo, and Matthew Hoggard.

1.2 Outline of the Thesis

The thesis comprises nine chapters. The chapter-wise organization of the thesis is given below:

Chapter 2 In this chapter, we present the necessary background to understand the subsequent chapters. It first introduces the game of cricket and its data sources. Next, the chapter provides bird's-eye views of the research fields of sports data mining and text mining.

Chapter 3 This chapter proposes to use *cricket text commentary* data for obtaining player's strength and weakness rules. To this end, it first touches on the construction of a large-scale dataset consisting of more than one million cricket text commentaries spanning over thirteen years. Next, the chapter discusses several challenges inherent to the text commentary data that arise when applying machine learning techniques. Next, the chapter discusses cricket text commentary processing and the extraction of domain-specific features from it. Finally, the extracted features are represented using confrontation matrix representations.

Chapter 4 This chapter introduces an approach to identify the strength and weakness rules of individual players using cricket text commentary data. A computationally feasible definition of strength rule and weakness rule is introduced. Next, a visualization method is proposed for interpreting the obtained strength and weakness rules.

Chapter 5 This chapter introduces an approach to learn the temporal changes in the strength and weakness rules of cricket players using the text commentary data. Year-wise changes of strength and weakness rules of batsmen and bowlers are visualized using line plots.

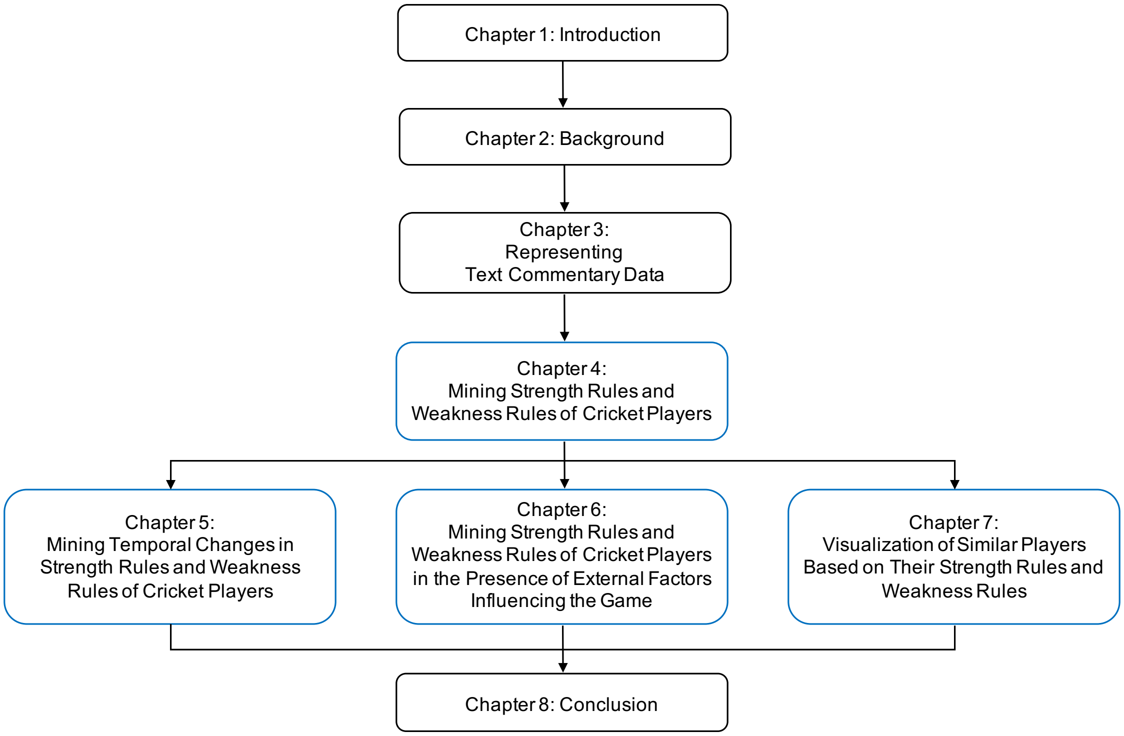

Chapter 1: Introduction
Chapter 2: Background
Chapter 3: Representing Text Commentary Data
Chapter 4: Mining Strength Rules and Weakness Rules of Cricket Players
Chapter 5: Mining Temporal Changes in Strength Rules and Weakness Rules of Cricket Players
Chapter 6: Mining Strength Rules and Weakness Rules of Cricket Players in the Presence of External Factors Influencing the Game
Chapter 7: Visualization of Similar Players Based on Their Strength Rules and Weakness Rules
Chapter 8: Conclusion

Figure 1.1: Flow Diagram of the Thesis Structure.

Chapter 6 This chapter introduces an approach to mine the strength and weakness rules of cricket players in the presence of external factors influencing the game. The influence of external factors such as playing conditions and match situations on player's strength and weakness rules are analyzed.

Chapter 7 This chapter introduces a method to visualize similar players based on their strength and weakness rules. Similar batsmen and bowlers are visualized using the t-SNE plots.

Chapter 8 This final chapter highlights the conclusions and summarizes the contributions made. New avenues for future research are also discussed.

The structure of the thesis is shown in Figure 1.1.

2

Background

In this chapter, we first briefly introduce the game of cricket. Next, we discuss the types of data generated in cricket. Finally, we provide an overview of existing work in the fields of sports data mining and text mining.

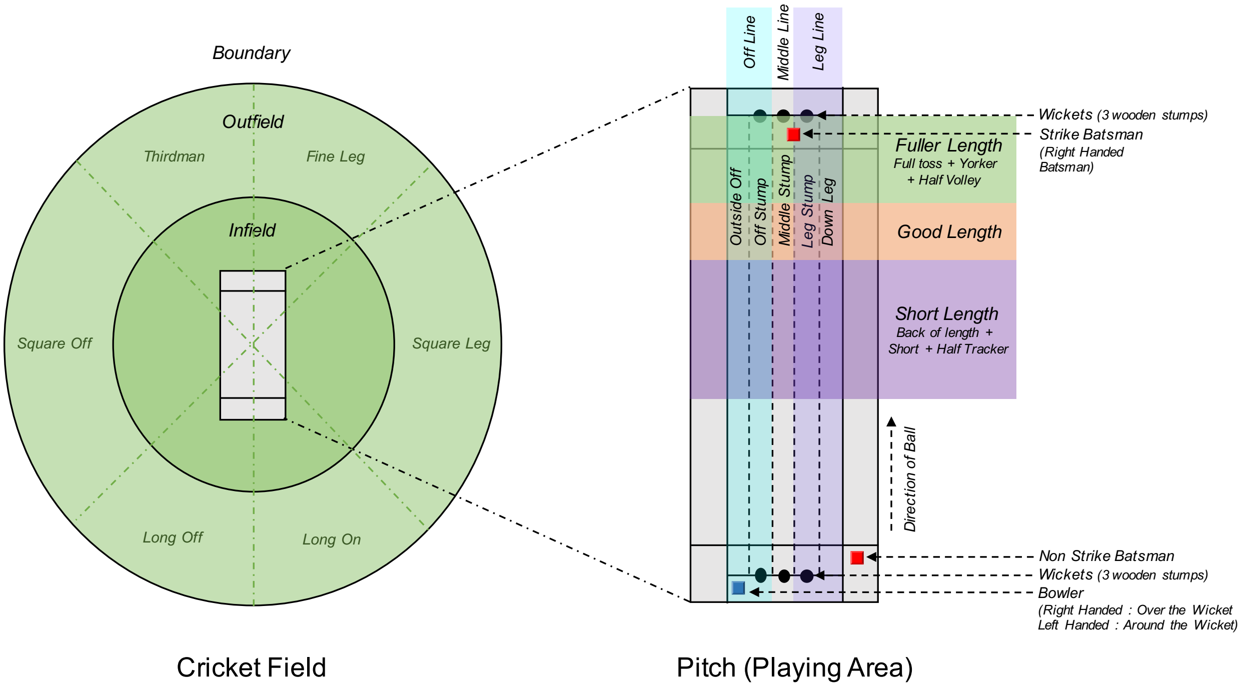
2.1 Cricket

Cricket is a bat-and-ball game played between two teams of eleven players each in a field at the center of which is a rectangular strip called the *pitch*. A standard cricket field with the playing area or pitch is presented in Figure 2.1. A cricket field is divided into - *infield* inside the 30 yard circle and *outfield* from circle to boundary.

In cricket, a player can be a (i) *batsman* who hits the ball to score runs, (ii) *bowler* who bowls the ball towards the batsman, (iii) *fielder* who stops the ball hit by the batsman in the field, and (iv) *wicket-keeper* who stands behind the wicket to collect the ball bowled by the bowler.

Each match is divided into innings. In every innings, one team bats and the opposite team fields (or bowls), which is decided by a coin toss. The bowler bowls on a 22-yard pitch, a hard surface made of clay and has two wickets (3 wooden stumps) on either side. A batsman bats on one side of the wicket, and the bowler bowls from the other side of the wicket. A ball can be delivered onto the batsman in different ways to get the batsman out. Fast bowlers aim to rely on their speed or use the seam of a ball so that it swings or curves in flight. Spinners bowl slowly but with a rapid rotation to change the ball's trajectory on striking the pitch. Each ball also has attributes like length (how far down the pitch the ball is pitched), line (how far to the left or right of the wicket the bowler bowls the ball), type (nature of the delivery), speed (speed of the ball after it is released), and movement (movement of the ball w.r.t. the batsman).

At any moment, two batsmen of a team are present on the pitch. One is called the striker batsman who hits the ball, and another is the non-striker batsman on the opposite side of the striker (or bowler's end). Each batsman continues batting until he is out, which happens when the



Cricket Field

Pitch (Playing Area)

Figure 2.1: A Standard Cricket Field, Showing the Rectangular Cricket Pitch (White) at Center, Infield (Medium Green) inside the 30 Yard Circle, Outfield (Light Green) from Circle to Boundary, and Shot Areas (Thirdman, Square Off, Long Off, Long On, Square Leg, and Fine Leg) for Right Handed Batsman.

batsman hits the ball, but it is caught by a fielder without bouncing (caught) or when the bowler strikes the wickets (bowled) or when the ball would have struck the wicket but was rather blocked by the batsman's body except the hand holding the bat (leg before wicket or LBW) and a few other scenarios. A batsman can score one/two/three/four runs by running between wickets, i.e., both the striker and the non-striker must reach their respective opposite ends a requisite number of times. A batsman may also score four runs (ball hits the ground before hitting/passing the boundary) or six runs (ball passes or hits the boundary without bouncing), without running, by striking the ball to the boundary. Batsmen react to a bowler in a variety of ways. They could defend the ball (block the ball) from hitting the wickets, attack (play aggressive shots) it for a boundary scoring four or six runs, or get beaten (play poor shot) by the bowler. The batsman can play different shots to hit the ball to different regions of the field (shot areas). The cricket field can be divided into six regions such as *thirdman*, *square off*, *long off*, *long on*, *square leg*, and *fine leg*.

A batsman (or a bowler) can be left-handed or right-handed. In Figure 2.1, line of delivery and shot areas are shown for right-handed batsmen. For left-handed batsmen, the shot areas are mirrored. Similarly, these notations are mirrored for left-handed and right-handed bowlers as well.

The completion of an innings depends upon the format of the game. In limited over formats of the game, an inning gets completed when all the overs have been bowled, or 10 out of 11 batsmen of the batting team have been declared out (all-out). The two limited formats of cricket are (i)

Twenty20 International (T20I), which is the shortest format of the game and comprises two innings, one innings per team (each inning is limited to 20 overs, and each over has six deliveries/balls), (ii) One Day International (ODI), which is played for one day and comprises of two innings, one innings per team (each inning is limited to 50 overs). In the first innings, the batting team sets the target for the fielding team, and in the second innings, the fielding team (which is now the batting team) tries to achieve the target. The team which scores the most runs wins the match.

Another format of the game, which is not limited by overs, is Test cricket. It is the longest and purest form of the game because it tests teams' technique and temperament over a more extended time. Test match cricket is played for a maximum of five days (each day has three sessions of two hours each) and comprises four innings, two innings per team. Usually, teams will alternate after each innings. A team's innings ends when (i) team is all-out, (ii) team's captain declares the innings, (iii) team batting fourth scores the required number of runs to win, or (iv) time for the match expires. Let Team-A bat in the first innings and Team-B field. Next, Team-B bat in the second innings. If, after the second innings, Team-A leads by at least 200 runs, the captain of Team-A may order (enforcing the follow-on) Team-B to bat in the next innings. In this case, the usual order of the third and fourth innings is reversed. Now, Team-A will bat in the fourth innings. The team which scores the most runs in its two innings wins the match.

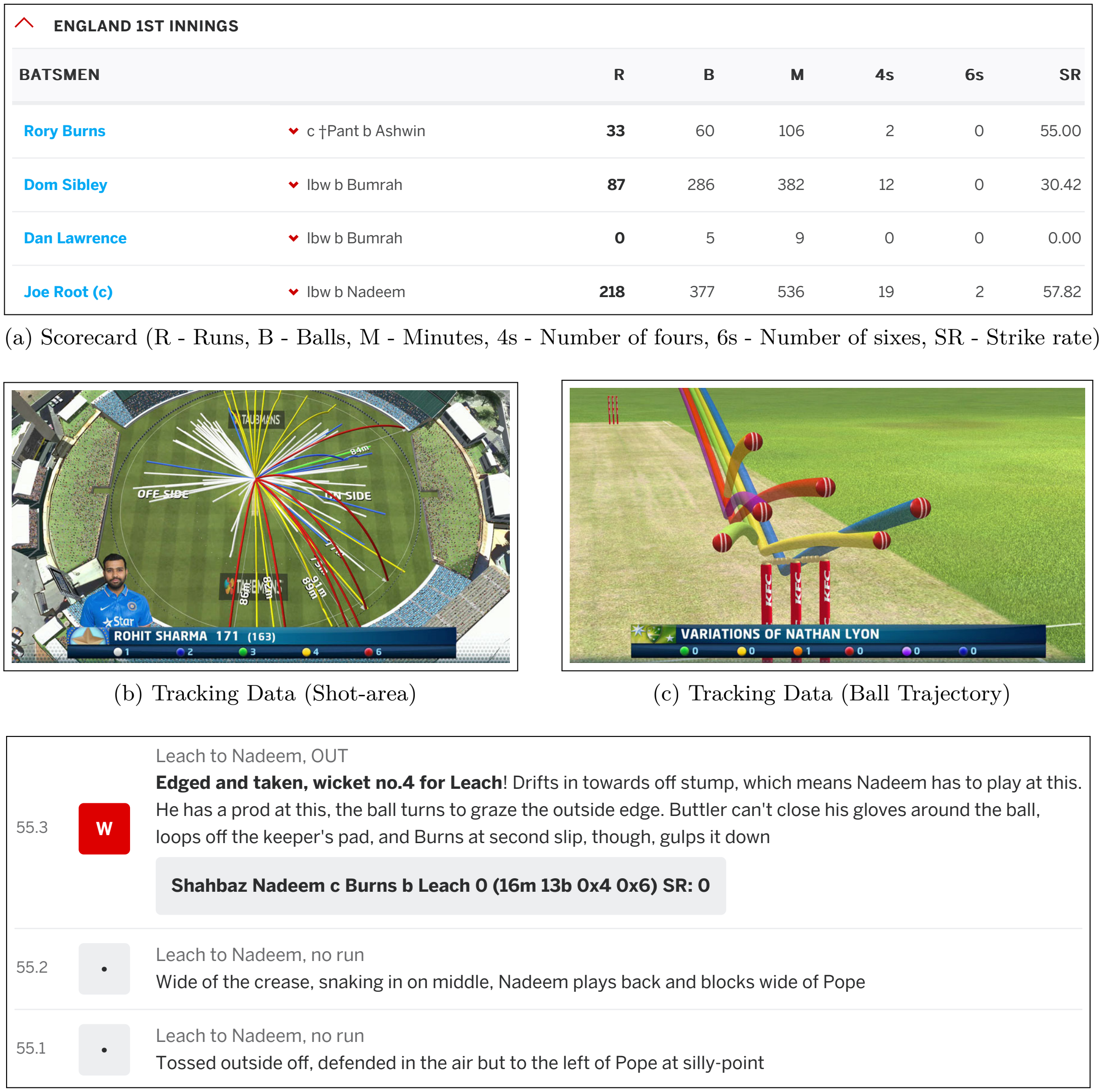
2.2 Cricket Data Sources

Cricket is known for recording every detail. A vast amount of data is generated in every cricket match. Cricket teams are collecting increasing amounts of data during the matches. We give a brief description of the types of data (structured and unstructured) generated in the game of cricket.

2.2.1 Structured Data

Structured data have an inherent structure, which makes the pattern recognition task easier. The following types of structured data are generated in cricket matches:

1. **Box-score Data:** In cricket, box-score data are the first kind of data that is recorded. These are the summary statistics such as the number of runs scored, the number of balls faced, the fall of wickets, the number of extras, etc. The earliest box-score data (scorecards) dates back to the year 1772. Compared to the latest box-score data, they lack a few details: the fall of wickets, the number of balls faced, the number of balls bowled, the number of extras, etc. Even though it is almost 249 years old, the change in the box-score data percolates only for a few places. Accordingly, the change in box-score data analysis made progress with this available additional information. An example of a scorecard is presented in Figure 2.2a. Box-score data are generated by companies such as Opta Sports [2], ESPNcricinfo [3], and CricBuzz [4].



ENGLAND 1ST INNINGS

BATSMEN		R	B	M	4s	6s	SR
Rory Burns	c †Pant b Ashwin	**33**	60	106	2	0	55.00
Dom Sibley	lbw b Bumrah	**87**	286	382	12	0	30.42
Dan Lawrence	lbw b Bumrah	**0**	5	9	0	0	0.00
Joe Root (c)	lbw b Nadeem	**218**	377	536	19	2	57.82

(a) Scorecard (R - Runs, B - Balls, M - Minutes, 4s - Number of fours, 6s - Number of sixes, SR - Strike rate)

(b) Tracking Data (Shot-area)

(c) Tracking Data (Ball Trajectory)

55.3 W
Leach to Nadeem, OUT
Edged and taken, wicket no.4 for Leach! Drifts in towards off stump, which means Nadeem has to play at this. He has a prod at this, the ball turns to graze the outside edge. Buttler can't close his gloves around the ball, loops off the keeper's pad, and Burns at second slip, though, gulps it down

Shahbaz Nadeem c Burns b Leach 0 (16m 13b 0x4 0x6) SR: 0

55.2 •
Leach to Nadeem, no run
Wide of the crease, snaking in on middle, Nadeem plays back and blocks wide of Pope

55.1 •
Leach to Nadeem, no run
Tossed outside off, defended in the air but to the left of Pope at silly-point

(d) Text Commentary

Figure 2.2: Cricket Data - (a) Scorecard [3], (b) Tracking Shot-area [31], (c) Tracking Ball Trajectory [31], (d) Text Commentary [3]

2. **Tracking Data:** The use of advanced data capture techniques [1] enable coaches and broadcasters to uncover new insights into every shot, run, and wicket. An example of batsman's shot-area tracking is presented in Figure 2.2b. Hawk-Eye technology (introduced in 2001) is for tracking ball trajectory, speed of the delivery, position of the pitched ball, and bounce of the ball. The visualizations obtained using this data are employed in umpires' decision-making and are used by analysts and experts. Ball trajectory data (Refer to Figure 2.2c) is used to extrapolate the ball trajectory using simulations for taking Leg Before Wicket (LBW)

[1] https://www.ibtimes.co.in/top-10-technological-advancements-that-changed-cricket-forever-624174

related decisions. Similarly, ball pitch position is used to understand the trends in the bowling patterns. Tracking data are generated by companies such as STATSports [6] and Opta Sports. The main bottleneck with the tracking data is that they are not available publicly as they are highly expensive to capture in every match. Besides, tracking data capture only specific game events, not the gameplay in detail.

2.2.2 Unstructured Data

Unstructured data do not have an inherent structure, which makes the pattern recognition task difficult. The following types of unstructured data are generated in cricket:

1. **News Articles:** Sports news articles report the summary of a match after it is over. They provide only a macroscopic view of the game, i.e., include only the game's key events. These are unstructured text data having a length of two to three pages.

2. **Video Broadcast:** Video broadcast is the live video streaming of the match. The video broadcast service for cricket was started in the year 1938. It is usually accompanied by audio commentary by the official cricket commentators, describing the events that are happening during the match. The video data is provided by broadcasting companies such as Sky Sports [5].

3. **Social Media Post:** Social media activity such as Twitter posts during the cricket match can highlight interesting events in the match. These are text data having a few hundred words.

4. **Commentary:** Cricket commentary is a source of rich descriptions of minute details of the game's proceedings. The commentator's opinion about how a bowler bowled and a batsman played on every individual delivery is recorded in the commentary. Commentary is of two types:

   (a) **Audio Commentary:** Live audio commentary associated with the broadcast (television/radio) of the match. Live audio commentary was started in the year 1922.

   (b) **Text Commentary:** Text commentaries are written narratives that provide a detailed description of a ball-by-ball account of the game while it is unfolding. Text commentary was started in 1991, but it has seen an explosion since 2006. Cricket text commentary data is provided by companies such as ESPNcricinfo, Opta Sports, and CricBuzz. Refer to Figure 2.2d for the examples of text commentaries from ESPNcricinfo. The first commentary describes the third delivery in the 56^{th} over of the game in which *Leach* is the bowler and *Nadeem* is the batsman. The outcome of this delivery is OUT, i.e., the batsman is dismissed in this delivery. The rest of the text describes how the ball is delivered and how the batsman played it. For instance, this commentary describes several features of bowling, such as line (*off stump*). Similarly, it describes batting features such as batsman's response (*outside edge*).

Cricket Data	Storage	Computation	Availability	Credibility	Details
Box-score Data	✓	✓	✓	✓	✗
Tracking Data	✗	✗	✗	✓	✓
News Article	✓	✓	✓	✓	✗
Audio Commentary	✗	✗	✗	✓	✓
Video Broadcasts	✗	✗	✗	✓	✓
Social Media Post	✓	✓	✓	✗	✗
Text Commentary	✓	✓	✓	✓	✓

Table 2.1: Comparision of Cricket Data.

2.2.3 Comparison of Cricket Data

The structured data, such as box-score data and tracking data, capture only a limited amount of information specific to the events in matches. On the other hand, unstructured data such as video broadcast and text commentary capture all the information of matches in minute detail. We present a comparison of various cricket data in Table 2.1. The comparison criteria include the efficient storage and computation, availability and credibility of the data, and the richness of information (minute details). From Table 2.1, we can see that text commentary data has some unique advantages over all of them. Compared to audio, video, or tracking data, it is readily available and requires less storage and computational resources. Unlike social media posts, it is credible, maintains a consistent style, and maps each delivery to a single text commentary. Unlike news articles, it provides a microscopic view with ball-by-ball details of the match.

2.3 Related Work

Literature closely related to sports data mining, text representation, short text representation, text visualization, and short text visualization is presented in this section.

2.3.1 Sports Data Mining

Data mining is the process of uncovering hidden trends and patterns from data sources. The data sources can be structured such as databases or unstructured such as videos. Data mining is applied to these structured and unstructured data to learn hidden patterns in different applications domains such as business, medicine, engineering, etc.

A massive amount of data related to player performances, team performances, etc., are collected by organizations and sport-related associations. Various data mining techniques have been applied successfully on the generated sports data in the past few years. Sports data mining refers to the application of data mining techniques on sports data [32]. It is used for player performance analysis, team performance analysis, building game strategies, etc., by teams to have advantage over their opponents. Sports data mining tasks in various sports are presented in Table 2.2. Box-score data, video data, and tracking data are the focus of analysis for sports researchers and analysts [61].

Sports	Data Mining Tasks
American Football	Estimating team strength [33], Evaluating quarterbacks [34], Evaluating place-kickers [35], Forecasting the success [36]
Baseball	Batter's performance [37], Base runner's performance [38], Pitcher's performance [39], Fielder's performance [40]
Basketball	Evaluating player contributions [41, 42], Optimal strategy planning [43], Measuring offensive/defensive player abilities [44, 45]
Cricket	Target resetting [7, 8], Match simulation [9, 10, 46, 47], Evaluating player performance [11, 13, 48, 49], Evaluating team strength [15, 50], Optimal lineups [16, 17, 51], Tactics [18, 19]
Hockey	Assessing player performance [52], Optimal strategy planning [53], Player contribution [54], NHL drafting [55]
Soccer	Model outcomes [56, 57], Team quality [58], Individual player ratings [59], Referee bias [60]

Table 2.2: Data Mining Tasks in Various Sports.

We present below work carried out using each of these data in various sporting domains, including cricket.

Box-score Data Analysis

Box-score data are the discrete data referencing the in-game events. Humans generate most of these events and statistics (e.g., scorecards, points table). Statistics is primarily applied to box-score data to measure player and team performances. In baseball, a massive amount of data and statistics are generated in a traditional tabular presentation of a baseball team's schedule. To make sense of this data requires significant cognitive effort. SportsVis [62] uses the baseline bar display and player map to explore a team's and a player's performance throughout a season, respectively. However, it considers only the aggregate information for any particular game. In basketball, treemap [63] is used to visualize NBA basketball player statistics. In soccer, a gap chart [64] is used to visualize the temporal evolution of ranks and scores of soccer teams participating in a competition. A gap chart is a class of line charts where gaps between teams show the magnitude of their score difference, hence ensures no overlap of tied entries.

In cricket, box-score data is used for target resetting, match simulation, player performance analysis, team strength analysis, optimal lineups prediction, and devising match tactics. We present the literature related to limited over cricket (T20I and ODI) and Test cricket separately.

Limited Over Cricket: When a match is interrupted due to bad weather, target resetting plays a major role. Duckworth and Lewis [7] proposed a method (DL method) for target resetting, which is adopted by the International Cricket Council (ICC). To account for the recent changes in scoring, the DL-method is updated by Stern [8]. Bailey and Clarke [9] developed a predictive model of the game's outcome by employing the DL-method, and a linear model was used for fitting the resulting target scores. Correlation of winning a game to different batting combinations and run rate was identified by Allsopp and Clarke [10]. Traditional statistics such as batting average fails to consider the number of balls faced, and strike rate fails to consider the number of dismissals. Croucher [48] proposed batting index (batting average × strike rate) which takes both into account. Saikia et al. [49] proposed the first quantitative investigation of fielding. To provide a weighted measure of fielding proficiency, they performed subjective assessments of every fielding play, such as catching, ground fielding, and run-outs. Theodoro et al. [13] proposed a Bayesian hidden Markov model for assessing batting in one-day cricket. Iyer and Sharda [11] employed neural networks to predict cricket players' performance based on their past performances. Assessing team strength in cricket is very crucial. Davis et al. [15] proposed a match simulator that assesses team strength in T20I cricket. Jhanwar and Pudi [50] model relative team strength using player's career statistics and recent performances. Using the relative team strength, toss decision, and match venue, they predict the winner of an ODI cricket match. Optimal team selection for a game has a major effect on the game's outcome. Lemmer [16] and Ahmed et al. [17] proposed search algorithms to select teams. The approaches typically permit constraints on team selection, e.g., a fixed number of pure batsmen, all-rounders, and bowlers are imposed when forming a team. Chhabra et al. [51] proposed a team recommendation system in cricket by modelling players into embeddings that represent players' strengths and weaknesses. Their model is based on player's past performances (quantitative factor) and opponent players' strengths and weaknesses (qualitative factor).

Test Cricket: Brooks et al. [46] used an ordered probit model with batting and bowling strengths to predict match outcomes in Test cricket. Scarf and Shi [47] modeled the match outcome probabilities using logistic regression, given the position at the end of the third innings. Scarf and Akhtar [18] extended this to the positions at the end of the first and second innings. Their models have been used to consider the declaration strategy and the follow-on decision in Test cricket. Scarf et al. [19] used negative binomial distributions to model the runs scored in innings and partnerships during test matches.

Video Analysis

Gong et al. [65] proposed a system to parse soccer videos to various play categories. To achieve this, they employed four high-level detectors, namely line mark recognition, motion detection, ball detection, and player's uniform color detection. The plays in the mid-field, penalty zones, and corner areas are identified with high accuracy, whereas for shot-at-goals and corner-kicks, the accuracy is low. Assfalg et al. [66] employed a Hidden Markov Model (HMM) to detect and recognize soccer matches' highlights. Specifically, the authors investigated penalty kicks, free kicks next to the goal post, and corner kicks. These three actions are typical highlights often shown in a soccer game. For the classification task, qualitative features are extracted from the video. The free kicks, penalty kicks, and corner kicks are identified with 80%, 90%, and 100% accuracy, respectively. Highlight recognition is also successfully examined in other sports domains such as tennis [67], basketball [68], and baseball [69]. In cricket, camera motion estimation was carried out by Lazarescu et al. [14] to index cricket videos and to classify shots offered by batsmen based on the estimates. Low-scoring shots are classified accurately compared to high-scoring shots.

Baillie and Jose [71] proposed an audio-based event detection on soccer broadcasts. Mel Frequency Cepstral Coefficients (MFCC) are extracted from the soundtrack of soccer commentary. A high correlation between crowd response to key events was taken as a cue to detect events effectively.

The text data present in the video is utilized in indexing, retrieval, efficient learning, and effective inference. The superimposed text provides vital information about the game's proceedings. Zhang and Chang [72] employed caption-text detection and recognition to identify events in baseball videos. Score, out, and ball-counts related text are detected with high accuracy compared to inning-number.

Xu et al. [73] proposed a combination of video and audio features to detect tennis events. In particular, low-level features in the video, namely motion vector field, texture, and color are extracted. MFCC features and zero-crossing rate are employed to differentiate applause from commentary speech. Audio keywords, which are argued to be mid-level features, are extracted for audio commentary data in various sports. These are, in turn, used to detect semantic events. Nepal et al. [68] used crowd cheer (audio), scorecard (text), and motion detection (video) for event detection in basketball. Sankar et al. [70] used keyword analysis on synchronized text commentary and video to identify an act of the game that a video frame corresponds to (ball being bowled or advertisement). They also looked into interesting commentaries by maintaining a count of interesting words.

Tracking Data Analysis

Recent developments in tracking and sensing technologies make it possible to obtain spatio-temporal information (x and y coordinates at time t) about the players and equipment (e.g., ball, bat, etc.) in real-time during the play. Tracking data are the continuous spatio-temporal motion data generated by multi-camera tracking systems (e.g., Hawk-Eye, SportVU).

Many visualization tools are introduced for finding hidden patterns using spatio-temporal data. In soccer, Wu et al. [74] proposed ForVizor to visualize player formation changes over time and reveal the continuous spatial flows of formations (formation flows) for in-depth analysis. To explain the cause of formation analysis, in addition to formation flows, multiple coordinated components are also designed. ForVizor uses a combination of manual annotation and algorithmic representation. It involves the manual annotation of the entire video by experts, which requires significant effort and may not be scalable. In baseball, Dietrich et al. [75] proposed Baseball4D that uses raw baseball tracking data (player and ball) over time and plots them as events on a dot map to reconstruct the entire game and visually explore each play. It combines time-varying player tracking and ball tracking data streams to generate nontrivial statistics and visualizations. In basketball, Beshai [76] developed Buckets that utilizes basketball shot data (spatial data) to view details about a single-player, compare multiple players, and explore league trends. In cricket, Das et al. [12] proposed CricVis, a web-based visualization system that utilizes box-score data (scorecards) and tracking data (ball tracking) to construct visualizations such as pitch maps and stump maps to analyze the bowling overview and batting overview, respectively. Morgan et al. [20] predicted where a specific batsman would hit a specific bowler and bowl type in a specific game scenario. Spatio-temporal data-based analysis focuses on visualizing low-level information (e.g., player actions). High-level tactical strategies, e.g., team tactics, are hard to infer from this low-level information.

Sports data mining methods mainly focus on box-score data, tracking data, and video data. Box-score data are used for tasks such as target resetting, match simulation, player performance analysis, etc. While interesting, the focus of these methods has primarily been at an aggregate level, i.e., they do not attend to the minute details of the game. Tracking data and video data are successfully applied in many sports-related tasks. However, the main bottleneck with these data is that they are not available publicly as they are highly expensive to capture in every match. It motivates us to propose a model that uses publicly available data, considers the game's minute details, and learns player-specific strategies (strengths and weaknesses). Unstructured data in the form of cricket text commentary is used for this task. In order to effectively use the text commentary data, we look at the literature related to text (and short text) representation and visualization in the following sections.

2.3.2 Text Representation

The text representation literature is grouped into two distinct categories - term frequency-based methods and topic modeling methods.

Term Frequency Methods

Indexing and retrieval are two key tasks for designing information retrieval systems. The vector space model is a popular approach in which each document is represented by a set of indexing terms (Salton and Buckley [77]). Every indexing term is associated with a weight. The frequency of occurrence of a term in a given document is effective for improved recall, well known as term-frequency. For improved precision, the frequency of terms occurring in different documents needs to be accounted for in the index term weighting scheme known as inverse document frequency. The combination of term-frequency and inverse document frequency is attributed towards term discrimination by taking the product of term-frequency and inverse document frequency. An exhaustive study of varying term weighting methods is explored by Salton and Buckley [78] and Leopold and Kindermann [79] for the effectiveness and efficiency of retrieval systems. Wilbur and Kim [80] have shown that by utilizing only inverse document frequency and ignoring term frequency, one obtains classification accuracy similar to that when term frequencies are included in classification systems. The implication of document length on retrieval systems has been identified by Singhal et al. [81].

A pivoted normalization approach was presented for the retrieval system's effectiveness by Singhal et al. [82].

Topic Modeling Methods

Document representation based on term frequency relies on the presence of terms in a given document. Deerwester et al. [83] presented an automatic indexing method that extracts the semantic structure of the term-document association by applying singular value decomposition on the term-document matrix. In particular, a semantic space is constructed in which closely related terms and documents are placed nearer. Hofmann [84] proposed a probabilistic latent semantic indexing known as pLSI based on the likelihood principle. The proposed model is a generative model of the data. This approach aims to maximize the log-likelihood of the joint probability of word and document weighted by term-frequency. Blei et al. [85] identified two shortcomings in the pLSI model, namely number of parameters is directly related to the number of documents, and there is no direct handle in assigning probabilities to documents beforehand. They considered mixture models to account for *exchangeable representation* of words and documents. Laplacian pLSI (Cai et al. [86]) and Locally consistent topic modeling (Cai et al. [87]) considered topic modeling in the intrinsic document manifold by removing the assumption in pLSI and LSI that latent topics are in the Euclidean space. Huh and Fienberg [88] proposed the discriminative topic model that not only pulls neighboring document pairs closer together but also separates non-neighboring document pairs from each other.

2.3.3 Short Text Representation

Experiments on large documents are performed for topic modeling using the above-discussed methods with convincing success. However, effective representation of very short documents (usually less than a few hundred characters long) such as microblogs, news feeds, product reviews web snippets for classification, clustering pose a challenge due to large indexing terms and sparsity of the terms in each document. Short text documents are often noisy and are less topic-centric. Suggested ways to work with short text documents are - (i) expanding the short text using search engines [89, 90, 91] and (ii) using online data repositories such as Wikipedia [92, 93].

Phan et al. [94] employed a framework of collecting large external data for each category termed as *universal dataset*. Classification models are built by utilizing a universal dataset along with the short text document.

Hu et al. [95] explored the impact of social relations on sentiment classification in the online social network. The unigram model is employed for document representation with binary weight for each term representing the term's presence or absence. Stemming and stop word removal operations are not performed.

Ni et al. [96] presented a method to cluster short text snippets. The central idea is to represent each short text snippet as a vertex of a graph. The relationship between two snippets is represented as an edge between them. A weight is associated on the edge, which signifies the similarity of the vertices. A variant of RatioCut called as TermCut is applied on the built graph on the short text snippets. To represent each text snippet, the inverse document frequency of a term is used as the associated term's weight.

Sun [97] performed short text classification by employing very few words. The key idea is to identify representative words which stand for a topic. Representative terms are selected using the term-frequency criteria. Terms having more than a specified threshold are chosen as representative terms. A clarity measure is proposed to understand topically relevant words, which uses Kullback-Leibler divergence distance between a set of documents containing a given word and the entire document collection. The clarity score is multiplied with term-frequency.

Li et al. [98] addressed the limitation of single-topic assumption in the field of short text topic modeling. Previously the assumption, each short text has only one topic was too strong for some datasets. To address this limitation, they proposed a Poisson-based Dirichlet Multinomial Mixture model that allows each short text to be sampled by a limited number of topics (one to three). This model delivered better classification accuracy and topic coherence than the state-of-the-art alternatives in most settings.

Liang et al. [99] proposed a dynamic user clustering topic model to infer user's intention distributions based on the context of user's posts during the current time period and previously inferred distributions. However, they do not consider the inner relationships between words in short texts.

2.3.4 Text Visualization

Text visualization has been extensively studied in the past for many years (Alencar et al. [100]). The visualization methods are categorized based on the input into three categories - single document, document collection, and document collection over time.

Single Document

In the case of visualization methods involving a single document, the frequency of words is considered an essential measure. The idea is to use the bag-of-words representation to visualize frequent words. For instance, Steinbock [101] proposed TagCrowd that creates a tag/word cloud to visualize the words used in a particular text. The font size indicates its frequency in the text, with larger fonts indicating more frequent/important words. Viegas et al. [102] proposed Wordle to enhance the visualization of word clouds by improving the usage of available visual areas.

The word frequency-based methods cannot convey semantic relationships among the words. To overcome this, Wattenberg and Viegas [103] proposed Word Tree, in which a text document is represented as a tree (suffix tree) with nodes representing words and branches linking sequential words. It enables the users to navigate in the text by selecting words and checking all sentences in which they occur. Ham [104] proposed Phrase Nets that creates a graph to visualize the relationship between words where nodes correspond to words and edges correspond to relationships.

Document Collection

To visualize a collection of documents, document-maps and hierarchical strategies are primarily used. Document-maps exploit the spatial relationship among documents and provide a navigation interface to visualize the global document relationships. Skupin [105] proposed Cartographic Maps that generate a visualization that is similar to a geographic map. However, these methods suffer from the overlapping of graphical elements in the navigation interface. On the other hand, hierarchical strategies allow users to view maps at multiple levels of detail, i.e., from large clusters to individual documents. For example, Andrews [106] proposed InfoSky to display the hierarchy in the visual space to allow zoom in or out operations in some regions of the projection.

Document Collection Over Time

To visualize a collection of documents over time, a few methods extend the existing document visualizations to handle time. For example, Lee et al. [107] proposed SparkClouds, which is an extension of tag clouds to incorporate time. In addition to the words, it displays a minimal line chart called sparkline under each word to show its frequency variation over time. Havre et al. [108] proposed ThemeRiver to display the change in the theme of a document collection by highlighting selected topics represented by single words. In this representation, the major topics are represented as colored streams, with flow width indicating the topic's strength at a particular moment.

2.3.5 Short Text Visualization

Short texts are usually less than a few hundred characters long. Unlike conventional texts, short texts are sparse (contains few words), making the job of text visualization difficult. Due to the sparsity, the visualization of a single short text document becomes irrelevant. Thus, the short text visualization methods are categorized as - short text collection and short text collection over time.

Short Text Collection

Opinions (posts and comments) in social forums are generally organized based on their semantic contents (Joty et al. [109]). However, it encourages selective exposure to information and opinion polarization. Gao et al. [110] developed an interface to allow interactive visualization and categorization of Reddit posts about controversial topics that involve people with different attitudes and stances. The aim is to allow users to explore opinions from different stances and ultimately reduce opinion polarisation. Social media text, an unstructured text, is a valuable source of information about the public's interests and opinions. To provide a high-level summary of social media text, word clouds [101] are used. Hu et al. proposed SentenTree [111] to visualize the frequent sequential patterns in social media text to gain a fast understanding of critical concepts and opinions. It uses a node-link diagram where nodes are words, links are the word co-occurrence within the same sentence, and nodes' size indicates their frequency of occurrence.

Short Text Collection Over Time

With the surge in social media, identifying anomalous information spreading patterns is necessary to make data-informed decisions. Ho et al. [112] proposed various methods to measure a user's ability to analyze micro-blog diffusion. Their methods are based on factors such as number of people influenced, propagation speed, and geographic distance. They also provided a visualization system to explore the information propagation via propagation paths and social graphs, influence scores, timelines, etc. However, their methods cannot reveal the diffusion patterns. Zhao [113] designed FluxFlow to detect the diffusion process of rumors on social media. The design presents the temporal diffusion process in a flow to show the overview patterns and individual details. It uses multiple coordinate views to reveal rich contexts such as topics and sentiments. FluxFlow enables a detailed comparison between normal and abnormal diffusion processes.

Cricket text commentary is typically short text with some unique features. None of the existing methods can be employed directly in the context of cricket text commentary as the objective of the present work is distinctly different from the rest of the work in literature. In particular, the relationship between batting and bowling features, which are collectively present in the text document, needs to be captured.

2.4 Summary

This chapter discussed the preliminaries and presented the existing works to introduce the perspective for our work in the context of literature. First, the game of cricket is introduced, and the types of data generated in cricket matches are discussed. A detailed comparison of various cricket data is made, and advantages of text commentary data over others are explained. Next, a bird's-eye view of the existing research work in sports data mining is presented. Several sports data mining methods based on box-score data, tracking data, and video data are presented. Limitations of these data and methods are discussed, and the motivation behind the use of cricket text commentary data is formed. Finally, the literature related to text (and short text) representation and visualization are presented, and their applicability in cricket text commentary data is discussed.

3

Representing Text Commentary Data

In this chapter, we discuss the representation of text commentary data and its challenges. Refer to Figure 3.1 for an overview of this chapter. We first describe the raw text commentary data and its acquisition. We then discuss the engineering challenges and propose steps to process this data. Next, we describe the process of feature extraction from the processed data. Finally, we introduce the confrontation matrix data model.

3.1 Text Commentary Data

Text commentaries are written narratives that give a detailed description of the ball-by-ball account of the game while it is unfolding. It is a source of detailed descriptions of the game's proceedings with minute details. This data is generated live by human commentators who watch live video feeds of cricket games and are maintained by companies such as ESPNcricinfo, Opta Sports, and CricBuzz. The technical details, such as how a bowler bowled a delivery and the way a batsman played the delivery, are recorded in each text commentary. In addition to the focus on players'

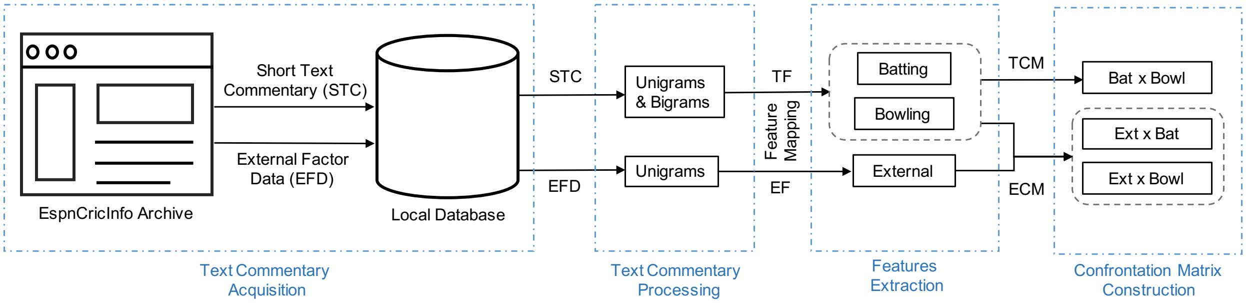


Figure 3.1: Chapter Overview - Representing Text Commentary Data. In this figure, STC (Short Text Commentary), EFD (External Factor Data), TF (Technical Features), EF (External Features), TCM (Technical Confrontation Matrix), ECM (External Confrontation Matrix), Bat (Batting Features), Bowl (Bowling Features), and Ext (External Features).

on-the-ball actions, external factors such as pitch conditions and weather conditions that influence the game are also recorded in the text commentary. Cricket has multiple formats of the game, of which the *Test cricket* format is considered for the present work. Two types of text commentary data are considered for this work: (i) *Short text commentary* corresponding to every delivery of a Test match and (ii) *External factor data* corresponding to every session of a Test match.

3.1.1 Short Text Commentary

Short text commentary encodes the technical factors corresponding to batting and bowling, which describe the game in greater detail and help in a microscopic understanding of the game. A ball can be delivered onto the batsman in different ways to get the batsman out. Fast bowlers aim to rely on their speed or use the seam of a ball so that it swings or curves in flight. Spinners bowl slowly but with a rapid rotation to change the ball's trajectory on striking the pitch. Each ball also has attributes like length, line, speed, and movement. Batsmen react to bowlers in a variety of ways. They could defend the ball from hitting the wickets, attack it for a boundary (scoring four or six runs), play it for a single, or be beaten by the bowler (miss the ball or show imperfection). Batsmen can play different kinds of shots to hit the ball in different regions of the field. These actions and outcomes comprise features that can be used to represent a batsman's or a bowler's action at a microscopic level. Short text commentary for a particular delivery comprises a maximum of fifty words. Consider an example of short text commentary from ESPNcricinfo [3]:

> 106.1, Anderson to Smith, 1 run, 144 kph, England have drawn a false shot from Smith! well done. good length, angling in, straightens away, catches the outside edge but does not carry to Cook at slip.

This commentary describes the first delivery in the 107th over of the game in which *Anderson* is the bowler and *Smith* is the batsman. The outcome of this delivery is one run. The rest of the text describes how the ball is delivered and how the batsman played it. For instance, this commentary describes several features of bowling, such as length (*good length*) and movement (*angling in*). Similarly, it describes batting features such as batsman's response (*outside edge*) pointing to the batsman's imperfection. The word *false shot* emphasizes the imperfection and points to batsman's weakness against a *good length* and *angling in* delivery.

Consider another example of short text commentary given below where the technical word *punch* points to the batsman's perfection or strength against a *short* length delivery.

> 3.2, Finn to Sehwag, FOUR, short of a length, but a little wide, enough for Sehwag to stand tall and punch it with an open face, past Pietersen at point.

3.1.2 External Factor Data

External factors like playing conditions (age-of-ball, pitch condition, and weather condition) and match situations (day of the match, inning of the match, and session of the day) are not under the player's control and are crucial to outdoor team sports like cricket. Pitch condition and weather condition data are available as text commentary at the beginning of each session. The rest of the external factor data are available directly from the website.

Different types of pitch conditions and weather conditions suit different kinds of players. For example, a bowler can extract high bounce from a grassy, hard, or intact pitch. Similarly, a pitch with moisture can help swing the ball in-flight using the ball's heaviness due to moisture. A pitch that is dry and cracked helps a spinner achieve more turn because of the unevenness. Similarly, flat pitches do not assist bowlers and favor batsmen. Here is an excerpt of text commentary from ESPNcricinfo for a game in Perth, Australia describing the characteristics of the grassy pitch.

> The WACA track has loads of grass on it, there should be plenty of bounce and seam movement.

Weather also has its effect on the game. On a hot sunny day, the sun takes away all moisture from the pitch, making it batsman-friendly. On a day with high humidity, it becomes hard for a bowler to grip the ball again, aiding the batsman. Here is another excerpt commentary from ESPNcricinfo on weather condition:

> We don't know as yet whether England have enforced the follow-on or not considering the weather conditions with a chance for showers England might enforce it.

Weather conditions and pitch conditions are not the only external factors affecting the game of cricket. Match situations also have indirect and sometimes direct effects. A pitch once prepared for a match is not repaired after each day. As one gets closer to the end of the game, typically on the last day or two, it becomes dry, and cracks start to develop. A spinner can be thought to have favorable conditions on the last day compared to the first day. A ball is replaced every 90 overs in the game. A new ball offers more pace and movement to fast bowlers.

The above instances of text commentary data show how text commentary capture information about external factors. They also act as evidence of the different external conditions and how they favor batsmen or bowlers.

3.2 Text Commentary Acquisition

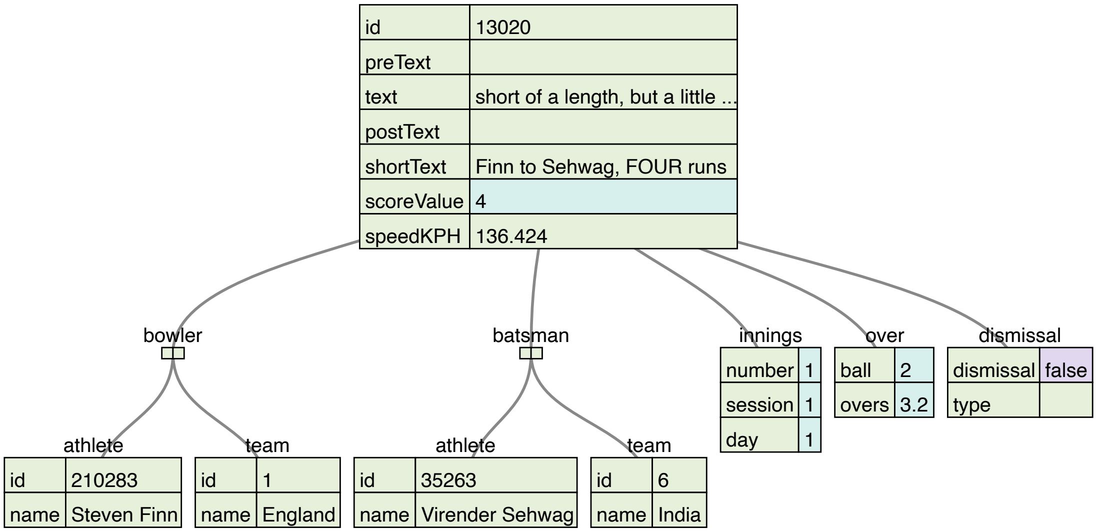
Three prominent vendors of text commentary data are: ESPNcricinfo [3], Opta Sports [2], and CricBuzz [4]. We selected ESPNcricinfo as the data source because (i) it is the pioneer company to record the data, (ii) the data are made publicly available, (iii) it includes additional data of external factors affecting the game, and (iv) the simpler structure of the data. It is a sports news website exclusively dedicated to cricket, and it provides the most comprehensive repository of cricket text commentary. The earliest documented text commentary available for Test matches dates back to 2006. An example of delivery in a cricket match recorded by ESPNcricinfo is presented below:

3.2, Finn to Sehwag, FOUR, short of a length, but a little wide, enough for Sehwag to stand tall and punch it with an open face, past Pietersen at point.

```
{ "id": "13020", "clock": "00: 00", "date": "2012-12-05T09: 00", "playType": { "id": "3", "description": "four"} , "team": { "id": "6", "name": "India", "abbreviation": "INDIA", "displayName": "India"} , "mediaId": 0, "period": 1, "periodText": "1st innings",
"preText": "", "text": "short of a length, but a little wide, enough for Sehwag to stand tall and punch it with an open face, past Pietersen at point", "postText": "", "shortText": "Finn to Sehwag, FOUR runs",
"homeScore": "18/0", "awayScore": "0", "scoreValue": 4, "sequence": 100302,
"athletesInvolved": [{ "id": "210283", "name": "Steven Finn", "shortName": "Finn", "displayName": "Steven Finn"} , { "id": "35263", "name": "Virender Sehwag", "shortName": "Sehwag", "displayName": "Virender Sehwag"} ],
"speedKPH": "136.424", "speedMPH": "84.770",
"bowler": { "athlete": { "id": "210283", "name": "Steven Finn"} , "team": { "id": "1", "name": "England"} , "maidens": 0, "balls": 8, "wickets": 0, "overs": 1.2, "conceded": 13},
"otherBowler": { "athlete": { "id": "8608", "name": "James Anderson"} , "team": { "id": "1", "name": "England"} , "maidens": 0, "balls": 12, "wickets": 0, "overs": 2.0, "conceded": 5},
"batsman": { "athlete": { "id": "35263", "name": "Virender Sehwag"} , "team": { "id": "6", "name": "India"} , "totalRuns": 4, "faced": 7, "fours": 1, "runs": 18, "sixes": 0},
"otherbatsman": { "athlete": { "id": "28763", "name": "Gautam Gambhir"} , "team": { "id": "6", "name": "India"} , "totalRuns": 14, "faced": 13, "fours": 3, "runs": 0, "sixes": 0},
"innings": { "id": "118240", "runRate": 5.4, "remainingBalls": 0, "byes": 0, "number": 1, "balls": 20, "noBalls": 0, "wickets": 0, "legByes": 0, "ballLimit": 0, "target": 0, "session": 1, "day": 1, "fallOfWickets": 0, "remainingOvers": 0.0, "totalRuns": 4, "wides": 0, "runs": 18},
"over": { "ball": 2, "balls": 6, "complete": false, "limit": 0.0, "maiden": 0, "noBall": 0, "wide": 0, "byes": 0, "legByes": 0, "number": 4, "runs": 4, "wickets": 0, "overs": 3.2},
"dismissal": { "dismissal": false, "bowled": false, "type": "", "bowler": { "athlete": { "id": "210283"} }, "batsman": { "athlete": { "id": "35263"} } , "fielder": { "athlete": { } } , "retiredText": ""} }
```

Out of the information provided by ESPNcricinfo, the following attributes are of interest for this thesis (Refer to Figure 3.2):

- **preText and postText:** External factor data such as weather condition and pitch condition
- **text and shortText:** Short text commentary data



id	13020
preText	
text	short of a length, but a little ...
postText	
shortText	Finn to Sehwag, FOUR runs
scoreValue	4
speedKPH	136.424

number	1
session	1
day	1

ball	2
overs	3.2

dismissal	false
type	

id	210283
name	Steven Finn

id	1
name	England

id	35263
name	Virender Sehwag

id	6
name	India

Figure 3.2: Data Collected for a Delivery.

- **scoreValue:** Outcome of the delivery
- **speedKPH:** Speed of the delivery
- **bowler and batsman:** Details of bowler and batsman respectively
- **innings:** Inning number, session number, and day number of the Test match
- **over:** Number of overs and deliveries bowled
- **dismissal:** batsman's dismissal and type of dismissal information

To collect these data for a given Test match, one has to first obtain the season and series of which this particular match is a part. In addition, match IDs, innings IDs, and associated URLs need to be formulated from ESPNcricinfo. These data are used to acquire the text commentaries for a given match. The steps to acquire the short text commentaries and external factor data are presented below:

Step 1. URLs of the Test matches. To begin with, we have made a list of the seasons for which we want the data. In international cricket, the calendar is divided into two seasons - (i) a season that spreads over two calendar years like 2020/21 (starts in October and ends in March), and (ii) a season that is spread over a single calendar year like 2021 (starts in April and ends in September). For each season, the Test matches played in that season are identified. For each Test match, the Test ID and URL of that match are obtained and stored in the *MatchLinks* table of the 'Cricket' database.

Step 2. List of teams and players. A list of teams and a list of players with additional information are prepared from ESPNcricInfo's archive and stored in the *TeamList* table and *PlayerList* table, respectively.

Step 3. Order of batsmen and bowlers. Batting order and bowling order for each match are obtained and stored in the *BattingOrder* table and *BowlingOrder* table, respectively. In addition, the information regarding which team won the toss for each match is also acquired and stored in the *TossWon* table.

Step 4. Number of overs bowled per session. A list of the total number of overs bowled per session for each match is compiled and stored in the *OverList* table.

Step 5. Text commentaries. The short text commentaries and external factor data for each match are obtained and stored in the *Commentary* table. The text commentaries corresponding to the external factor data are stored per session basis in the *EnvironmentalData* table.

This procedure is repeated for 550 international Test matches between May 2006 and April 2019. Total short text commentaries of 1,088,570 deliveries are collected spanning thirteen years. Two types of external factor data are acquired - *match situations* and *playing conditions*. The data related to match situations such as day of the match, innings of the match, session of the day are acquired directly from the respective match web page of the ESPNcricinfo website for all the 1,088,570 deliveries. The data related to playing conditions are the age-of-ball, pitch condition, and weather condition. Age-of-ball data is acquired from the 'match summary' tab of each match's home page. Pitch condition and weather condition data are available at the beginning and end of each session in the text commentary data. The pitch condition and weather condition data present in a session are mapped to every delivery bowled in the same session. Over 5000 paragraphs (consisting of multiple lines with no limit on the number of words) of playing condition descriptions are collected for all the 550 matches between May 2006 and April 2019.

Python programming language is used to build a web crawler that extracts these data from the ESPNcricinfo website. We have made use of the (i) *urllib2* library to obtain the web pages or HTML files of each Test match and (ii) *beautifulsoup* library for pulling data out of the obtained HTML files. A laptop (Macbook Air) with 8 gigabytes of RAM and an Intel i5 CPU was used to crawl the data. The acquired data (MySQL database) and the python code to obtain the data can be accessed at https://www.dropbox.com/sh/4kytwt881kvy3tj/AAAWGsefJ9To_ecrQhLuwgsPa?dl=0. For each of the above steps, we map the python code as (StepNumber)Code.py: (1)MatchLinks.py, (2)PlayerList.py, (3)BatBowlOrder.py, (4)OverList.py, (5)Commentary.py.

3.3 Text Commentary Processing

In this section, we discuss the challenges and proposed steps for processing the short text commentary and external factor data.

3.3.1 Short Text Commentary Processing

Each short text commentary can be divided into two parts: the structured part and the unstructured part. The structured part is located at the beginning of each commentary. It describes the exact over number, delivery number, name of the bowler, name of the batsman, and outcome of the delivery. After this, a text commentary will optionally describe various bowling features such as line, length, and delivery speed. Some text commentaries will also describe batsmen's responses in terms of their footwork and shot selection. In the end, some deliveries have a subjective opinion of the commentator about how the batsman performed. An example of short text commentary structure is presented below.

Structured Text Unstructured Text

39.4, Broad to Paine, OUT, Short ball, Paine pulls - straight to deep square leg! Catching practice for Burns, and Paine is in a world of... well, pain!.

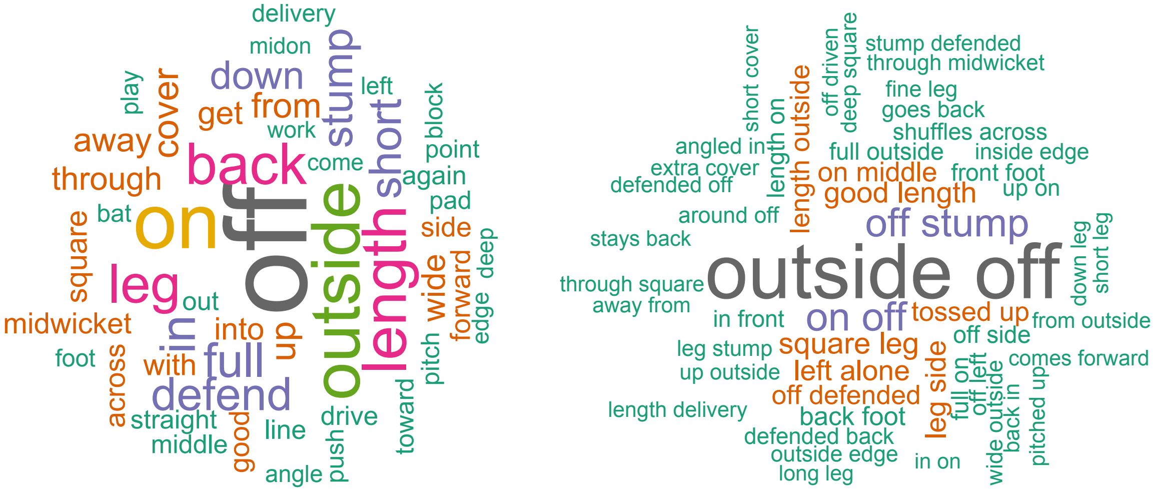
Extraction of information from the structured part is a straightforward task. However, information extraction from the unstructured part requires non-trivial effort. The main challenges are:

- **Stopwords:** For an effective representation of text documents, stopword removal is performed as a preprocessing step in the traditional information retrieval context. A differentiating factor specific to short text commentary is that the majority of the technical words used in the cricketing domain are *stopwords* in the conventional text mining literature. A non-exhaustive list of technical stopwords are: off, on, room, across, behind, back, out, up, down, long, turn, point, under, full, open, good, great, away, etc.

- **Sparsity:** Cricket text commentary has a definite structure in which both bowler's action and batsman's actions are described. Commentators, at times, focus either only on the bowler's action or only on the batsman's action. Moreover, every document (commentary for a particular delivery) comprises a maximum of fifty words. This induces sparsity. Features employed in traditional text mining literature like term frequency and inverse document frequency (TF-IDF) are not suitable due to the sparsity of the data.

Cricket, like any other sport, has a rich jargon. Frequency analysis of the data reveals that commentators frequently use technical words. Figure 3.3 presents the unigram and bigram word clouds [101] of the short text commentaries where Steve Smith is mentioned as a batsman. The domain-specific technical vocabulary is dominantly observed in these word clouds. The high technicality of the vocabulary used, on most occasions, enables us to approach the problem in a more organized manner. Through our analysis, we discover that the domain-specific feature space captures sufficient information from the short sentences and identifies whether the information is for the bowler or the batsman.

To capture the most relevant information, we have used a combination of web resources [114] and the frequency counts of words in the corpus to come up with words (unigrams) most likely to capture the data. However, we found significant fault with many such cases. Consider two examples of short text commentary:



(a) Unigram Word Cloud.

(b) Bigram Word Cloud.

Figure 3.3: Short Text Commentary Word Clouds for Batsman Steve Smith.

Swings in from outside off, well left in the end as it scoots past off stump. (Swing signifies the type of ball)

Short ball over middle stump, Dhoni swings into a pull and takes it down to fine leg. (Swing signifies the way batsman played the ball)

In both of these examples, the word *swing* is used differently; first concerning the bowler and second concerning the batsman. Many such words like *leg*, *short*, etc., are used in both contexts.

There are also instances when a word changes meaning when combined with another word. For example, the word *short* usually refers to a short ball, but when it is used as in *short leg*, *short cover*, etc., it refers to field positions. Consider two examples of short text commentary:

Short on the body, he gets up and nicely plays it to square leg. (Length of delivery)

Full outside off, Dhoni reaches out and pushes it to short cover. (Field position)

Such instances made us look into the possible usage of bigrams along with unigrams. Two words occurring together in a document are called bigrams. They carry more semantic meaning than single words. For example, *swing in*, *swing away*, *swing back*, and *late swing* are all bigrams that specifically address the swinging nature of the ball and removes the ambiguity of association with the batsman. Bigrams are also helpful in the instances when a word changes meaning when combined with another word. For example, word *short* usually refers to a short ball, but when it is used as in *short leg*, *short cover*, *short midwicket*, etc., it refers to the field positions. Thus, bigrams can be used to differentiate between multiple meanings and contexts of a word. However, a significant problem is the identification of all relevant unigrams and bigrams. We build a set of relevant unigrams and bigrams using a combination of *unigram frequency*, *bigram frequency*, *A glossary of cricket terms* [1], and *The Wisden Dictionary of Cricket* [114]. Finally, we represent each short text commentary as a set of unigrams and bigrams.

3.3.2 External Factor Data Processing

External factor data related to playing conditions such as pitch conditions and weather conditions are unstructured text commentaries. These are subjected to *stemming* and *stop word removal* operations. To capture the most relevant information, we use the frequency counts of words in the corpus to come up with words or unigrams most likely to capture the information present. Finally, we represent each text commentary related to external factors as a set of unigrams.

3.4 Feature Extraction

Unigram and bigram representations of text commentary can not be directly used for strength and weakness rule extraction. We have extracted *technical features* from the short text commentary and *external features* from the external factors data.

3.4.1 Technical Features

We have identified a total of 19 batting features (batsman facing the delivery is associated with these features) and 12 bowling features (bowler bowling the delivery is associated with these features) to represent each short text commentary.

Batting Features

The features that characterize batting are- *0 run, 1 run, 2 run, 3 run, 4 run, 5 run, 6 run, out, beaten, defended, attacked, front foot, back foot, third man, square off, long off, long on, square leg,* and *fine leg.* We give a brief description of each of these features with their feature categories below.

1. **Outcome**: Describes the outcome (runs/wicket) of each delivery.
   - Runs: The number of runs scored in a particular delivery. One can score `0, 1, 2, 3, 4, 5,` or `6` runs. Note that each possible run is enumerated for a delivery.
   - Dismissal: Whether the batsman is dismissed or declared `out` in that delivery.
2. **Response**: Describes the response of the batsman on each delivery. The response can be:
   - `Attacked`: These types of shots are executed to score runs, for instance a boundary. Aggressive batting displays strength of batsman (or weakness of bowler) to a particular type of delivery.

[1] https://es.pn/1bAFI9H

- `Defended`: The batsman either plays a blocking shot or leaves the ball. These shots usually do not result in runs and are known as dot balls.
- `Beaten`: Corresponds to a delivery in which a batsman has exhibited weakness (or a bowler has exhibited strength). Situations when a batsman offers a poor shot or a bowler bowls an excellent delivery.

3. **Footwork**: Describes the stance decision a batsman takes when facing a delivery from the opposing bowler. A batsman can play shots off the front foot or back foot. A right-handed batsman's front foot is the left foot, and back foot is the right foot. Similarly, a left-handed batsman's front foot is the right foot, and back foot is the left foot.

    - `Front Foot`: A batsman makes a front foot shot for full length deliveries (ball pitching closer to the batsman).
    - `Back Foot`: A batsman makes a back foot shot for short length deliveries (ball pitching closer to the bowler).

4. **Shot Area**: Describes the region where the shot is played by the batsman. The shot area corresponds to the fielding positions and usually is dependent on the length and line of the delivery bowled. Six shot areas are considered. These are: **`third man, square off, long off, long on, square leg,`** and **`fine leg`** (Refer to Figure 2.1).

Bowling Features

The features that characterize bowling are- *good, short, full, off, leg, middle, spin, swing, fast, slow, move-in,* and *move-out.* We give a brief description of each of these features with their feature categories below.

1. **Length**: Describes how far down the pitch (playing area) the ball is bowled/pitched. The bowler often adopts variations in length to surprise batsmen or make them uncomfortable.

    - `Full Length:` Bowler pitched the ball closer to the batsman.
    - `Short Length:` Bowler pitched the ball closer to himself.
    - `Good Length:` Bowler pitched the ball at an optimal length, in between full and short.

2. **Line**: Describes how far to the left or right of the wicket the ball is bowled by the bowler. Variation in line is used to keep the batsman guessing about the line to avoid pre-meditated shots.

    - `Off Line:` Ball travels on the off-stump line or outside the off-stump line.
    - `Middle Line:` Ball travels on the middle-stump line.
    - `Leg Line:` Ball travels on the leg-stump line or outside the leg-stump line.

Category	Batting Features	Identifying Unigrams and Bigrams
Response	Defense	leave, defend, block, leave alone
	Attack	drive, whip, punch, whack, great timing
	Beaten	miss, struck pad, beat, edge, lbw, poor shot
Footwork	Front	front foot, step out, come down
	Back	back foot, step back, hang back
Shot Area	Third man	third man, late cut, gully, back cut
	Square off	square, cover, point, upper cut, square drive
	Long off	mid off, long off, straight drive, off drive
	Long on	mid on, long on, on drive
	Square leg	short leg, square leg, sweep, hook
	Fine leg	fine leg, long leg, leg glance, paddle sweep

Table 3.1: Examples of Feature Definition (FD) for the Batting Features.

Category	Bowling Features	Identifying Unigrams and Bigrams
Length	Short	short, bouncer, short pitch, back length
	Full	full, overpitch, full toss, toss up, blockhole
	Good	length, good length, length delivery
Line	Off	outside off, pitch off, off stump, from off
	Middle	straight ball, straight line, middle stump
	Leg	down leg, wide leg, outside leg, leg stump
Type	Spin	spin, turn, googly, doosra, legspin, offspin
	Swing	swing in/away, late swing, reverse swing
Speed	Fast	fast, pace, quick, quicker
	Slow	slow, slower
Movement	Move In	move in, swing in, angle in
	Move Away	move away, swing away, angle away

Table 3.2: Examples of Feature Definition (FD) for the Bowling Features.

3. **Type**: Describes the nature of the delivery.

   - `Spin:` Bowler bowl slow deliveries which turn sharply after pitching.
   - `Swing:` Bowler bowl fast deliveries which have movement in the air.

4. **Speed**: Describes the speed of the ball after it is released.

   - `Fast:` Speed of the ball is medium (60 mph to 80 mph) or fast (> 80 mph)
   - `Slow:` Speed of the ball is 40 mph to 60 mph

5. **Movement**: Describes the movement of the ball w.r.t. the batsman.

   - `Move-in:` Ball moves towards the batsman.
   - `Move-away:` Ball moves away from the batsman.

All these features are discrete-valued. To transform each short text commentary to this feature space, we have defined a mapping from unigrams and bigrams to this feature space. Each feature is represented as a *set* of unigrams and bigrams such that the identified set corresponds to the feature in question. For the batting features and bowling features, the corresponding examples of unigrams and bigrams or the *Feature Definition (FD)* are given in Table 3.1 and Table 3.2, respectively. The complete list of FDs can be accessed at `https://www.dropbox.com/s/uqigazg72020rf1/features.txt?dl=0`. This unigram/bigram to feature mapping is obtained by consulting cricket experts. Corresponding to these features, 19 (batting features) and 12 (bowling features) sets of unigrams and bigrams are obtained. *This method of obtaining features has addressed the stop word related problem. The sparsity is addressed by mapping unigram and bigram of the text commentary only to these features.*

Since all the proposed features are discrete-valued, we represent them using one-hot encoding. For example, a bowling feature category *type* has two possible values: *spin* and *swing*. We represent this feature using two bits. For any given delivery, at the most, one of these two bits is set to one. When the value of feature *type* is unknown, both the bits are set to zero. A total of 31 values are possible across all the nine proposed features. Therefore, each delivery is represented using 31 bits.

3.4.2 External Features

We have identified a total of 28 external features. Out of these, 12 features correspond to the match situation, and 16 features correspond to the playing condition.

Match Situation

Twelve features are identified that characterize the match situation. These are: *Day 1, Day 2, Day 3, Day 4, Day 5, Inning 1, Inning 2, Inning 3, Inning 4, Session 1, Session 2,* and *Session 3*. We give a brief description of each of these features with their feature categories below.

- **Day**: Describes the day of a Test match. Each test match lasts a maximum of five days. The day can be `Day 1, Day 2, Day 3, Day 4,` or `Day 5`.
- **Inning**: Describes the inning of a Test match. Each match can have four innings, with each team getting to bat twice. Inning can be `Inning 1, Inning 2, Inning 3,` or `Inning 4`.
- **Session**: Describes the session of the day in a Test match. Each day of a test match has three sessions. So, the session can be `Session 1, Session 2,` or `Session 3`.

Playing Condition

Sixteen features are identified that characterize playing conditions. These are *BallNew, BallMid, BallOld, Green, Grass, Bounce, Pace, Slow, Turning, Swing, Flat, Dry, Rain, Breeze, Moist,* and *Hot*. We give a brief description of each of these features with their feature categories below.

Feature Name	Examples of Identifying Unigrams
rain	rain, overcast, shower, cloudy, gloom, muggy
hot	hot, sunny, bright, sultry, heat
bounce	bouncy, hard, rockey
flat	flat, dead, lifeless

Table 3.3: Examples of External Feature Definition (EFD).

1. **Age-of-ball**: The concept of old and new ball is accounted for in this feature and is measured as the number of overs the ball is used for - less than 10 overs (`BallNew`), from 10 to 30 overs (`BallMid`) and older than 30 overs (`BallOld`).

2. **Pitch condition**: The pitch condition denotes the state of the playing area (pitch). The predominantly found pitch types are **`green, grass, bounce, pace, slow, turning, swing, flat,`** and **`dry`**.

3. **Weather condition**: The weather conditions during the match are described using **`Rain, Breeze, Moist,`** and **`Hot`**.

To transform each external factor text commentary to this feature space, we have defined a mapping from unigrams to the feature space. Pitch condition and weather condition features are represented as a *set* of unigrams such that the identified set corresponds to the feature in question and is represented as *External Feature Definition (EFD)*. For example, the unigrams corresponding to the four external features presented in Table 3.3. The complete list of EFDs can be accessed at https://www.dropbox.com/sh/fwmm5a3x81ypikh/AACtGCD83tWHycSLnBioo1XKa?dl=0.

3.5 Confrontation Matrix Construction

Section 3.2 detailed the process of acquiring the entire text commentary database. To perform the player-specific analysis, one has to obtain a *subset of text commentaries* from this database. Extraction of this subset depends on a *Filter Tuple* ⟨*Player*, *Opponent Player*, *Time*, *Type*⟩ having four elements:

1. **Player:** The player for whom strength and weakness rules are to be extracted.

2. **Opponent Player:** The opponent player (for a batsman, opponents are the bowlers, and vice-versa) can be one player or a set of players.

3. **Time:** Time granularity in cricket is identified in the increasing order as over, session, day, inning, match, series, season, year, or career.

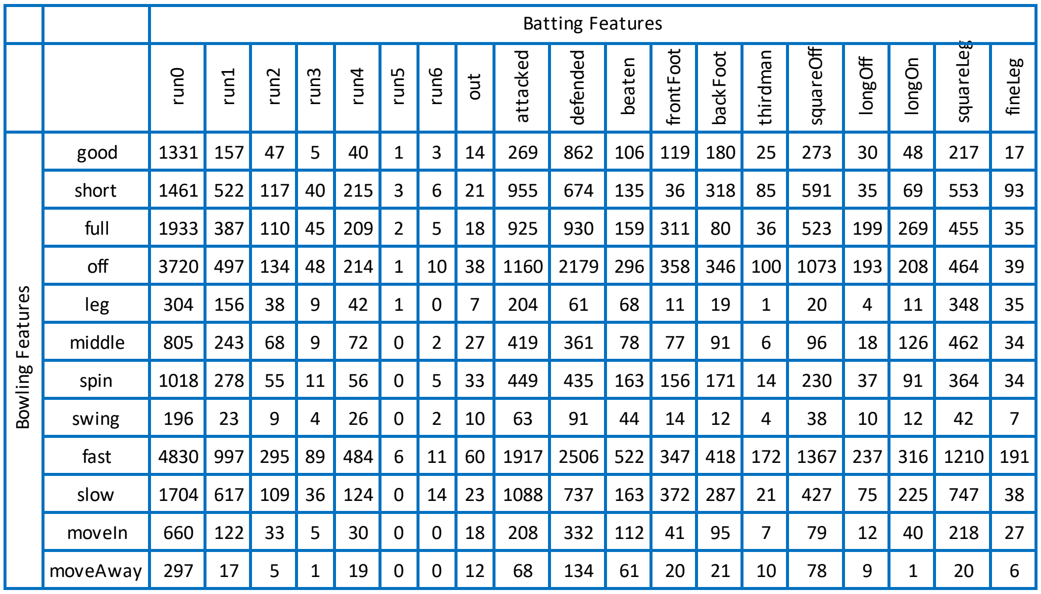
4. **Type:** It can be batting or bowling, depending on whether we want to analyze the batting or bowling of the player in focus. If the type is batting, then all the commentaries where the

Algorithm 1 Construction of Technical Confrontation Matrix (TCM)

```
Require: Filter Tuple ⟨Player, Opponent Player, Time, Type⟩, Feature definitions (FD), BattingOutcome = {0, 1, 2, 3, 4, 5, 6, out}, BattingFeatures = {attacked, defended, beaten, front foot, back foot, thirdman, square off, long off, long on, square leg, fine leg}, BowlingFeatures = {good, short, full, off, leg, middle, spin, swing, fast, slow, move in, move away}, A matrix $TCM_{19\times 12}$ of zeros.
Extract the short text commentaries using *Filter Tuple*
for Every commentary do
    Initialize two sets: $\underline{bat} = \phi$ and $\underline{bowl} = \phi$
    Get the $\underline{\text{outcome}}$ from the structured part of commentary
    for Every i ∈ BattingOutcome do
        if $\underline{\text{outcome}}$ == i then
            $\underline{bat} = \underline{bat} \cup \{i\}$
        end if
    end for
    Get all the unigrams and bigrams from the unstructured part of commentary
    for Every unigram or bigram y do
        for Every i ∈ BattingFeatures do
            if y ∈ $FD_i$ then
                $\underline{bat} = \underline{bat} \cup \{i\}$
            end if
        end for
        for Every j ∈ BowlingFeatures do
            if y ∈ $FD_j$ then
                $\underline{bowl} = \underline{bowl} \cup \{j\}$
            end if
        end for
    end for
    for Every $a \in \underline{bat}$ and $b \in \underline{bowl}$ do
        $TCM[a, b] = TCM[a, b] + 1$
    end for
end for
return Technical Confrontation Matrix (TCM)
```

player is mentioned as a batsman are selected. Similarly, if the type is bowling, then all the commentaries where the player is mentioned as a bowler are selected.

Given the subset of the text commentaries specific to the player, two confrontation matrices are constructed: (i) Technical Confrontation Matrix (TCM) corresponding to the bowling features and batting features and (ii) External Confrontation Matrix (ECM) corresponding to the batting/bowling features and external features.



		Batting Features																		
		run0	run1	run2	run3	run4	run5	run6	out	attacked	defended	beaten	frontFoot	backFoot	thirdman	squareOff	longOff	longOn	squareLeg	fineLeg
Bowling Features	good	1331	157	47	5	40	1	3	14	269	862	106	119	180	25	273	30	48	217	17
	short	1461	522	117	40	215	3	6	21	955	674	135	36	318	85	591	35	69	553	93
	full	1933	387	110	45	209	2	5	18	925	930	159	311	80	36	523	199	269	455	35
	off	3720	497	134	48	214	1	10	38	1160	2179	296	358	346	100	1073	193	208	464	39
	leg	304	156	38	9	42	1	0	7	204	61	68	11	19	1	20	4	11	348	35
	middle	805	243	68	9	72	0	2	27	419	361	78	77	91	6	96	18	126	462	34
	spin	1018	278	55	11	56	0	5	33	449	435	163	156	171	14	230	37	91	364	34
	swing	196	23	9	4	26	0	2	10	63	91	44	14	12	4	38	10	12	42	7
	fast	4830	997	295	89	484	6	11	60	1917	2506	522	347	418	172	1367	237	316	1210	191
	slow	1704	617	109	36	124	0	14	23	1088	737	163	372	287	21	427	75	225	747	38
	moveIn	660	122	33	5	30	0	0	18	208	332	112	41	95	7	79	12	40	218	27
	moveAway	297	17	5	1	19	0	0	12	68	134	61	20	21	10	78	9	1	20	6

Figure 3.4: Technical Confrontation Matrix (TCM_{BAT}) of Batsman Steve Smith (Transposed).

3.5.1 Technical Confrontation Matrix

A TCM of size 19×12 is constructed in which rows correspond to the batting features and columns correspond to the bowling features. For batting analysis, TCM_{BAT} is constructed such that the rows correspond to batting features of the batsman, and the columns correspond to bowling features of opponent bowlers. Similarly, for bowling analysis, TCM_{BOWL} is constructed such that the rows correspond to batting features of the opponent batsmen and the columns correspond to bowling features of the bowler. We present the detailed steps of TCM (TCM_{BAT} when *Type* in the filter tuple is *batting* and TCM_{BOWL} when *Type* in the filter tuple is *bowling*) construction in Algorithm 1.

Every element in the TCM corresponds to how the batsman confronted the bowlers, i.e., count of co-occurrences of the batting features and bowling features. For example, how many times a batsman has attacked the short-length deliveries. An example of TCM_{BAT} (transposed) for batsman Steve Smith is presented in Figure 3.4. This TCM_{BAT} is constructed using the filter tuple ⟨*Steve Smith, All Opponent Players, Career, Batting*⟩. The first entry in this matrix is 1331, which accounts for the number of times Steve Smith scored zero runs in the good length deliveries in his entire career against all opponent bowlers. TCMs for 131 batsmen and 129 bowlers who have batted/bowled more than 2000 deliveries are provided at https://www.dropbox.com/s/ko4xo7g69niqbrg/TCM.zip?dl=0.

3.5.2 External Confrontation Matrix

For batting analysis, ECM_{BAT} is constructed in which rows correspond to bowling features of the opponent bowlers and columns correspond to the external features. For each external feature category such as day, inning, we have constructed separate ECMs such as $ECM_{BAT}Day$,

Algorithm 2 Construction of $ECM_{BAT}Pitch$

```
Require: Filter Tuple ⟨Player, Opponent Player, Time, Type⟩, Feature definitions (FD), External feature definitions (EFD), BowlingFeatures = {good, short, full, off, leg, middle, spin, swing, fast, slow, move in, move away}, $ExternalFeatures_{Pitch}$ = {green, grass, bounce, pace, spin, slow, swing, flat, dry}, A matrix $(ECM_{BAT}Pitch)_{12\times 5}$ of zeros.
Extract the short text commentaries and external factor data using Filter Tuple
for Every delivery do
    Initialize two sets: $ext = \phi$ and $bowl = \phi$
    Get all the unigrams and bigrams from the unstructured part of commentary
    for Every unigram/bigram x do
        for Every $i \in BowlingFeatures$ do
            if x $\in FD_i$ then
                $bowl = bowl \cup \{i\}$
            end if
        end for
    end for
    Get all the unigrams from the external factor data
    for Every unigram y do
        for Every $j \in ExternalFeatures_{Pitch}$ do
            if y $\in EFD_j$ then
                $ext = ext \cup \{j\}$
            end if
        end for
    end for
    for Every $a \in bowl$ and $b \in ext$ do
        $ECM_{BAT}Pitch[a, b] = ECM_{BAT}Pitch[a, b] + 1$
    end for
end for
return External Confrontation Matrix ($ECM_{BAT}Pitch$)
```

$ECM_{BAT}Inning$, etc. We present the steps of $ECM_{BAT}Pitch$ construction in Algorithm 2.

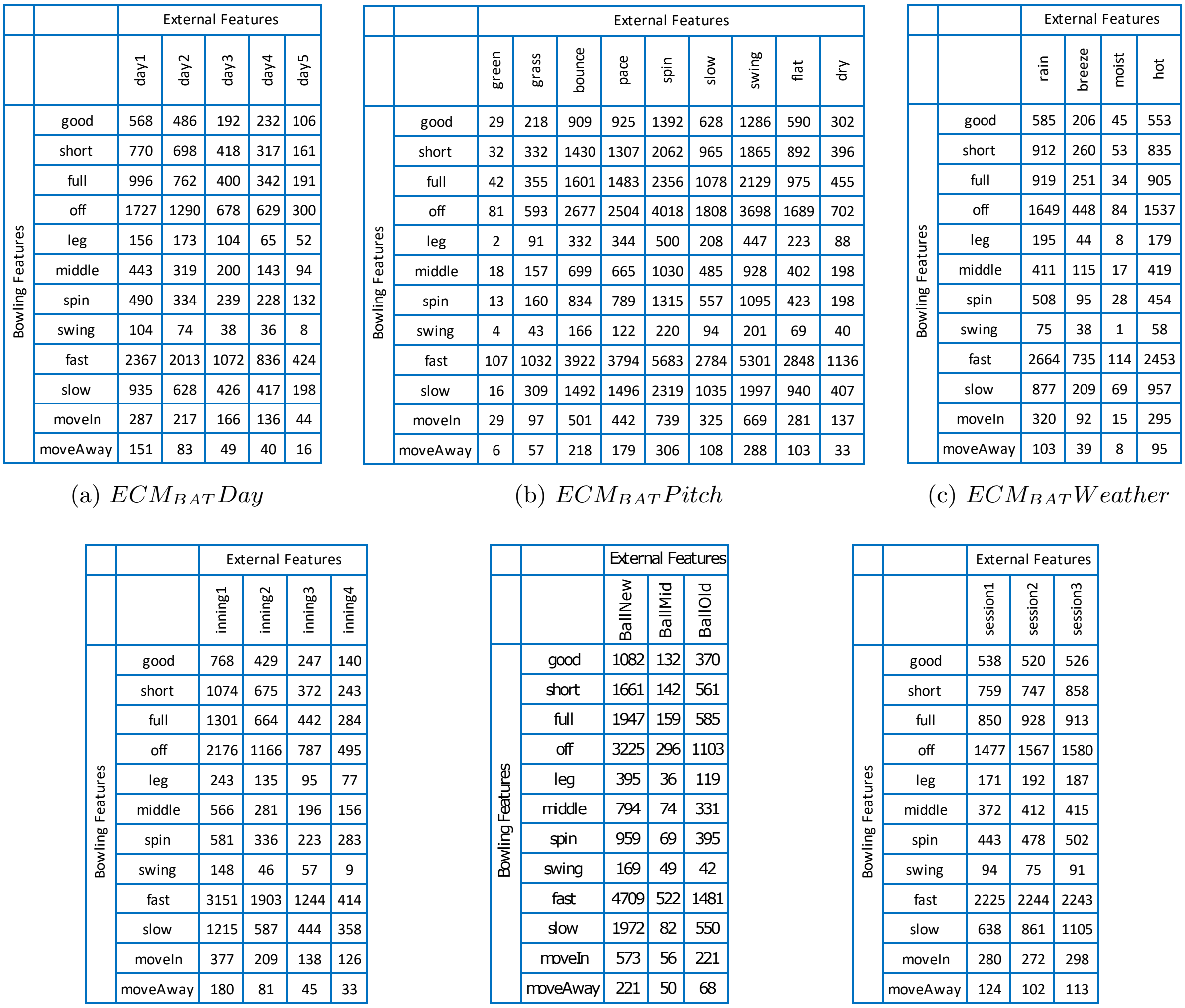
Examples of ECM_{BAT} for batsman Steve Smith are presented in Figure 3.5a - 3.5f. Figure 3.5a is the $ECM_{BAT}Day$ for batsman Steve Smith. The first entry in this matrix is 568, which accounts for the number of *good length* balls faced by Smith on *day-1* of all the Test matches he has played.

For bowling analysis, ECM_{BOWL} is constructed in which rows correspond to batting features

Algorithm 3 Construction of $ECM_{BOWL}Pitch$

```
Require: Filter Tuple ⟨Player, Opponent Player, Time, Type⟩, Feature definitions (FD), External
    feature definitions (EFD), BattingOutcome = {0, 1, 2, 3, 4, 5, 6, out}, BattingFeatures =
    {attacked, defended, beaten, front foot, back foot, thirdman, square off, long off, long on,
    square leg, fine leg}, $ExternalFeatures_{Pitch}$ = {green, grass, bounce, pace, spin, slow, swing,
    flat, dry}, A matrix $(ECM_{BOWL}Pitch)_{19\times5}$ of zeros.
Extract the short text commentaries and external factor data using Filter Tuple
for Every delivery do
    Initialize two sets: $ext = \phi$ and $bat = \phi$
    Get the outcome from the structured part of commentary
    for Every i ∈ BattingOutcome do
        if outcome == i then
            $bat = bat \cup \{i\}$
        end if
    end for
    Get all the unigrams and bigrams from the unstructured part of commentary
    for Every unigram/bigram x do
        for Every $i \in BattingFeatures$ do
            if x ∈ $FD_i$ then
                $bat = bat \cup \{i\}$
            end if
        end for
    end for
    Get all the unigrams from the external factor data
    for Every unigram y do
        for Every $j \in ExternalFeatures_{Pitch}$ do
            if y ∈ $EFD_j$ then
                $ext = ext \cup \{j\}$
            end if
        end for
    end for
    for Every $a \in bat$ and $b \in ext$ do
        $ECM_{BOWL}Pitch[a, b] = ECM_{BOWL}Pitch[a, b] + 1$
    end for
end for
return External Confrontation Matrix ($ECM_{BOWL}Pitch$)
```

of the opponent batsman and columns correspond to the external features. For each external feature category, we have constructed separate ECMs such as $ECM_{BOWL}Day$, $ECM_{BOWL}Inning$,



		External Features				
		day1	day2	day3	day4	day5
Bowling Features	good	568	486	192	232	106
	short	770	698	418	317	161
	full	996	762	400	342	191
	off	1727	1290	678	629	300
	leg	156	173	104	65	52
	middle	443	319	200	143	94
	spin	490	334	239	228	132
	swing	104	74	38	36	8
	fast	2367	2013	1072	836	424
	slow	935	628	426	417	198
	moveIn	287	217	166	136	44
	moveAway	151	83	49	40	16

(a) $ECM_{BAT}Day$

		External Features								
		green	grass	bounce	pace	spin	slow	swing	flat	dry
Bowling Features	good	29	218	909	925	1392	628	1286	590	302
	short	32	332	1430	1307	2062	965	1865	892	396
	full	42	355	1601	1483	2356	1078	2129	975	455
	off	81	593	2677	2504	4018	1808	3698	1689	702
	leg	2	91	332	344	500	208	447	223	88
	middle	18	157	699	665	1030	485	928	402	198
	spin	13	160	834	789	1315	557	1095	423	198
	swing	4	43	166	122	220	94	201	69	40
	fast	107	1032	3922	3794	5683	2784	5301	2848	1136
	slow	16	309	1492	1496	2319	1035	1997	940	407
	moveIn	29	97	501	442	739	325	669	281	137
	moveAway	6	57	218	179	306	108	288	103	33

(b) $ECM_{BAT}Pitch$

		External Features			
		rain	breeze	moist	hot
Bowling Features	good	585	206	45	553
	short	912	260	53	835
	full	919	251	34	905
	off	1649	448	84	1537
	leg	195	44	8	179
	middle	411	115	17	419
	spin	508	95	28	454
	swing	75	38	1	58
	fast	2664	735	114	2453
	slow	877	209	69	957
	moveIn	320	92	15	295
	moveAway	103	39	8	95

(c) $ECM_{BAT}Weather$

		External Features			
		inning1	inning2	inning3	inning4
Bowling Features	good	768	429	247	140
	short	1074	675	372	243
	full	1301	664	442	284
	off	2176	1166	787	495
	leg	243	135	95	77
	middle	566	281	196	156
	spin	581	336	223	283
	swing	148	46	57	9
	fast	3151	1903	1244	414
	slow	1215	587	444	358
	moveIn	377	209	138	126
	moveAway	180	81	45	33

(d) $ECM_{BAT}Inning$

		External Features		
		BallNew	BallMid	BallOld
Bowling Features	good	1082	132	370
	short	1661	142	561
	full	1947	159	585
	off	3225	296	1103
	leg	395	36	119
	middle	794	74	331
	spin	959	69	395
	swing	169	49	42
	fast	4709	522	1481
	slow	1972	82	550
	moveIn	573	56	221
	moveAway	221	50	68

(e) $ECM_{BAT}Ball$

		External Features		
		session1	session2	session3
Bowling Features	good	538	520	526
	short	759	747	858
	full	850	928	913
	off	1477	1567	1580
	leg	171	192	187
	middle	372	412	415
	spin	443	478	502
	swing	94	75	91
	fast	2225	2244	2243
	slow	638	861	1105
	moveIn	280	272	298
	moveAway	124	102	113

(f) $ECM_{BAT}Session$

Figure 3.5: External Confrontation Matrices of Batsman Steve Smith.

etc. We present the detailed steps of $ECM_{BOWL}Pitch$ construction in Algorithm 3. ECM_{BAT} and ECM_{BOWL} of 131 batsmen and 129 bowlers are provided at https://www.dropbox.com/s/btrqii76tntl1ua/Chapter6%20External%20Factor%20Analysis.zip?dl=0.

3.6 Summary

This chapter discussed the challenges and representation of text commentary data. First, the acquisition of raw text commentary data is described. After that, the engineering challenges and proposed steps for text commentary processing are discussed. Next, the feature extraction from the text commentary data is discussed. Finally, the confrontation matrix data model is introduced to represent the extracted features in a matrix form.

4

Mining Strength and Weakness Rules of Cricket Players

Knowledge of player's strengths and weaknesses is the key for team selection and strategy planning in any team sport such as cricket. Computationally, this problem is mostly unexplored. Computational methods to identify player's strengths and weaknesses can have a two-pronged impact. First, such methods can assist domain experts in better team selection and strategy planning. Second, such methods are the foundation of future automated strategy planning in cricket.

This chapter presents an approach to learn the strength and weakness rules of cricket players using short text commentary data. For a given player, the presented approach proceeds in three steps. In the first step, the approach introduces the computationally feasible definitions of strength and weakness rule. The second step employs a dimensionality reduction method specific to the discrete random variable case, namely Correspondence Analysis [23, 24, 25] on player's TCM to construct semantic relations between batting features and bowling features. In the third step, the approach plots these relations using biplot [26] and extracts human readable strength and weakness rules. The proposed approach considers the validation of the extracted rules as an integral part of rule extraction. The obtained rules are validated using intrinsic method and extrinsic method. For intrinsic validation, k-fold cross validation strategy is used, and for extrinsic validation, the identified rules are validated against external sources containing strengths and weaknesses of cricket players. Additionally, baseline comparisons are made using word clouds and association rule mining techniques. The chapter overview is shown in Figure 4.1.

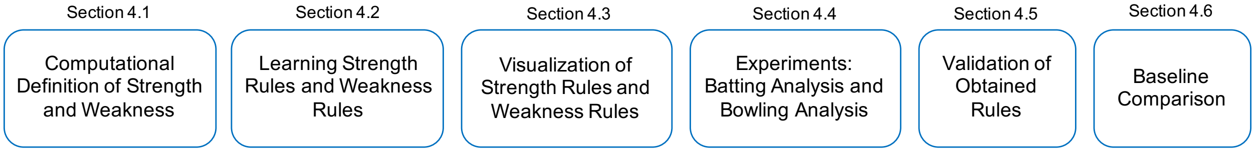

Section 4.1
Computational Definition of Strength and Weakness
Section 4.2
Learning Strength Rules and Weakness Rules
Section 4.3
Visualization of Strength Rules and Weakness Rules
Section 4.4
Experiments: Batting Analysis and Bowling Analysis
Section 4.5
Validation of Obtained Rules
Section 4.6
Baseline Comparison

Figure 4.1: Chapter Overview - Mining Strength and Weakness Rules of Cricket Players.

We use the proposed approach to mine strength and weakness rules corresponding to 131 batsmen and 129 bowlers who have batted/bowled more than 2000 deliveries. The data, code, and result of the experiments can be accessed at https://www.dropbox.com/sh/05zrr3sdz18mep8/AACBn8mjee3t2fyrhCb40wA2a?dl=0. We highlight some of the obtained rules for batsman Steve Smith and bowler Kagiso Rabada. For batsman Steve Smith, we observe that - (i) Smith attacks slow or short-length or middle-line deliveries (strength rule). (ii) Smith gets beaten on the deliveries that are either swinging or moving-away or moving-in (weakness rule). For bowler Kagiso Rabada, we observe that - (i) batsmen get beaten on the swing deliveries of Rabada (strength rule). (ii) batsmen attack the full-length deliveries of Rabada (weakness rule). We have also obtained the rules for batsmen against spin and fast bowling separately. For batsman Steve Smith, we observe that - (i) Against fast bowlers, Smith gets beaten on the moving-in deliveries (weakness rule) and attacks full-length deliveries (strength rule). (ii) Against spin bowlers, Smith gets beaten on the moving-away deliveries (weakness rule) and attacks short-length deliveries (strength rule).

4.1 Computational Definition of Strength and Weakness

There is no universally agreed definition of strength and weakness as different players exhibit strength or weakness at varying instances of deliveries. When players exhibit a particular behavior repeatedly, it amounts to their strength or weakness. For example, batsmen yielding their wicket to short deliveries pitched outside the off-stump consistently or bowlers getting attacked on in-swing deliveries amount to their weakness. Strength is when a batsman or a bowler exhibits perfection on a particular delivery, and *weakness* is when a batsman or a bowler exhibits imperfection on a particular delivery.

To compute the strength or weakness, arriving at a computationally feasible definition of what constitutes a rule is very important. We define the strength rule and weakness rule of a batsman/bowler as follows:

Definition 4.1 *Rule. A rule must comprise of one batting feature and one bowling feature which are dependent on each other.*

Definition 4.2 *Strength Rule of Batsman. In Definition 4.1 when the batting feature corresponds to attacked and involves any of the bowling features.*

Definition 4.3 *Weakness Rule of Batsman. In Definition 4.1 when the batting feature corresponds to beaten and involves any of the bowling features.*

Whenever a batsman exhibits strength on a delivery, it is a weakness for the bowler, and the inverse is also true. Therefore, bowlers' strengths and weaknesses are defined in terms of the batting features of the batsmen they have bowled to. A bowler exhibits strength when the opponent batsman's batting feature is beaten. Similarly, a bowler exhibits weakness when the opponent batsman's batting feature is attacked.

		Batting Feature	**Bowling Feature**
Batting Analysis	**Strength Rule**	Attacked	Any
	Weakness Rule	Beaten	Any
Bowling Analysis	**Strength Rule**	Beaten	Any
	Weakness Rule	Attacked	Any

Table 4.1: Computational Definitions of Strength Rule and Weakness Rule.

Definition 4.4 *Strength Rule of Bowler. In Definition 4.1 when the opponent's batting feature corresponds to beaten and involves any of the bowling features.*

Definition 4.5 *Weakness Rule of Bowler. In Definition 4.1 when the opponent's batting feature corresponds to attacked and involves any of the bowling features.*

Table 6.1 presents a summary of the above definitions. The central idea for rule computation is to subject Definition 4.1 to *independence event* test of probability. Let event e_1 = batsman attacking and e_2 = bowler bowling a good-length delivery. When these two events are *independent*, one event does not influence the probability of the other event. On the other hand, when these two events are dependent, the occurrence of one event influences the probability of the other event. For example, there is a greater chance of a batsman attacking (event e_1) when the bowlers bowl a good-length delivery (event e_2). The deviation from the independence signifies the relationship between batting features and bowling features, which in turn is captured as the strength rule or weakness rule of the player. In other words, every pair of batting and bowling features are subject to the above independence test. The deviation from the independence reveals the relationship between batting features and bowling features, which is expressed as strength rules or weakness rules as given in Definition 4.2 to Definition 4.5. The dependency is captured through the extent of violation (α) of the *independence of events* probability axiom: P(batting feature $\cap$ bowling feature) $= \alpha$ P(batting feature) $\times$ P(bowling feature); where $\alpha = 1$ when the batting feature is independent of the bowling feature. When $\alpha < 1$, then there is a dependency of batting feature on the bowling features. The value of α determines the extent of dependency.

To obtain the deviation from independence or the relationship between batting features and bowling features, the TCM (detailed in Section 3.5.1) is subjected to a dimensionality reduction method. Dimensionality reduction techniques are used to reduce the number of variables under consideration, capture the discriminative variables, and visualize. They represent the data such that transformed data retain original geometric structure and identify associations between rows and columns in the data. They are classified into two broad categories - (i) distance preserving methods: pairwise distances are preserved in the projected space [115, 116] and (ii) topology preserving methods: topology of the original data is preserved while performing projection. Self-organizing maps [117], locally linear embedding [118] and laplacian eigenmaps [119] are well-known methods in this category.

Widely employed dimensionality reduction methods such as laplacian eigenmaps [119] and principal component analysis [120] are not useful in the present context because the key assumption in the above methods is that the variables are continuous random variables. However, in the present context, batting features and bowling features are all discrete random variables. A multivariate statistical technique - Correspondence Analysis (CA) [23, 24] is used for the dimensionality reduction task. The detailed method is described in the following section.

4.2 Learning Strength and Weakness Rules

The input for strength and weakness rules computation is the TCM of the player. Let N be a TCM with I rows (batting features) and J columns (bowling features). An entry in the i^{th} row and j^{th} column, N_{ij}, represents the deliveries that contain both the features (i, j). Let n be the sum of the elements of matrix N. Every element of the matrix N is divided with n to obtain correspondence matrix. An ij^{th} element of the correspondence matrix denotes the joint probability that event i and event j occur simultaneously. Let event e_1 = batsman attacking and e_2 = bowler bowling good-length ball. When these two events are *independent*, the model is written as: $P(e_1 \cap e_2) = P(e_1) \times P(e_2)$. That is $P_{ij} = P_{i.} \times P_{.j}$; where P_{ij} denotes the joint probability of i^{th} row variable event occurring and j^{th} column variable event occurring, $P_{i.}$ denotes the probability of row event i occurring, and $P_{.j}$ denotes the probability of column event j occurring. When the total independence gets deviated, the model is re-written as:

$$P_{ij} = \alpha_{ij} \times P_{i.} \times P_{.j} \tag{4.1}$$

In equation 4.1, α_{ij} denotes the amount of deviation. If $\alpha_{ij} = 1$ then row event i and column event j are independent. When row features (batting features) have a certain relation with respect to column features (bowling features), α_{ij} takes a value less than 1. For every row event and for every column event, ij^{th} entry of the α matrix is given by:

$$\alpha_{ij} = \frac{P_{ij}}{P_{i.} \times P_{.j}} \tag{4.2}$$

Equation 4.2 is well known as Pearson's ratio. Pearson's Chi-square statistic is then expressed as:

$$\mathcal{X}^2 = n \times \sum_i \sum_j P_{i.} \times P_{.j} (\alpha_{ij} - 1)^2 \tag{4.3}$$

Equation 4.3 assumes a smaller value when $\alpha_{ij} \to 1$; which indicates that batting features are independent of bowling features. The higher the value of this quantity, the stronger is the relationship between batting and bowling features. This is the reason why we needed batting feature and bowling feature in Definition 4.1 to Definition 4.5. When $\alpha_{ij} \to 1$, the batting features and bowling features are not dependent on each other, the rule does not hold.

Note, however, that both batting features and bowling features are in high dimensional space.

CA, a multivariate statistical technique, is used to obtain a low dimensional subspace that contains the batting (row) features and bowling (column) features. In order to obtain the low dimensional space, CA minimizes the sum of the squared χ^2-distance (a weighted Euclidean distance given as $\chi^2 = \sum \frac{(observed-expected)^2}{expected}$) from the subspace to each of the row and column profile points. Row profile is a point in #columns (number of columns) dimensional feature space where each element in a row is divided with the corresponding sum of elements. Column profile is a point in #rows (number of rows) dimensional feature space where each element in a column is divided with the corresponding sum of elements. To obtain the solution to this minimization, Singular Value Decomposition (SVD) [142] is applied on the normalized and centered confrontation matrix N to obtain the principal components of row/batting features (F) and principal components of column/bowling features (G).

The steps of CA are presented in Algorithm 4. Let D_r and D_c be diagonal matrices whose diagonal elements are the elements of r and c. Center the correspondence matrix P as $(P - \mathbf{rc}^T)$. From the centered correspondence matrix obtain standardized residual matrix A as given below:

$$A = D_r^{-\frac{1}{2}}(P - \mathbf{rc}^T)D_c^{-\frac{1}{2}} \tag{4.4}$$

The matrix A is subjected to SVD to obtain $U\Sigma V^T$. Where Σ is a diagonal matrix whose elements are referred to as singular values of A. Every row of matrix U is associated with the row categories. Every column of matrix V is associated with the column categories. The principal component of the rows denoted by F is given by:

$$F = D_r^{-\frac{1}{2}}U\Sigma \tag{4.5}$$

Principal components of the columns are denoted by G is given below:

$$G = D_c^{-\frac{1}{2}}V\Sigma \tag{4.6}$$

F (row principal components) retains the batting features and G (column principal components) retains the bowling features as detailed in Algorithm 4. A finer interpretation of the relationship between batting features and bowling features for strength rule and weakness rule construction is described below based on Definition 4.1 to Definition 4.5.

4.2.1 Batting Analysis through CA

Refer to Figure 4.2 for the steps involved in batting analysis through CA. For batting analysis of a player, the TCM_{BAT} (denoted as N) of the player is used to obtain the relationships between batting features and bowling features. CA first obtains the residual matrix A from N. Next, SVD is applied to A to obtain the batting principal components (F) and bowling principal components (G). Then, the first two principal components of F and G (denoted as $F_{m\times 2}\ or\ F^{'}\ and\ G_{n\times 2}\ or\ G^{'}$) are obtained. Finally, the inner product matrix ($\langle F_{m\times 2}, G_{n\times 2}\rangle$) of the first two principal components

Algorithm 4 CA Algorithm

```
Require: A confrontation matrix $N_{I \times J}$
Matrix sum: $n = \sum_{i=1}^{I} \sum_{j=1}^{J} N_{ij}$
Row masses(r): $r_i = \frac{N_{i.}}{n}, i = 1, 2, \cdots, I$
Diagonal matrix: $D_r = diag(r_1, r_2, ..., r_I)$
Column masses(c): $c_j = \frac{N_{.j}}{n}, j = 1, 2, \cdots, J$
Diagonal matrix: $D_c = diag(c_1, c_2, ..., c_J)$
Correspondence matrix: $P = \frac{1}{n}N$
Standardized residuals: $A = D_r^{-\frac{1}{2}}(P - rc^T)D_c^{-\frac{1}{2}}$
Singular value decomposition: $A = U\Sigma V^T$
Principal components of rows: $F = D_r^{-\frac{1}{2}}U\Sigma$
Principal components of columns: $G = D_c^{-\frac{1}{2}}V\Sigma$
return F and G
```

of F and G is obtained.

Strength Rules of a Batsman

To qualify for a strength rule for a batsman, the batting feature as given in Definition 4.2 must contain the *attacked* feature. In order to frame a complete rule as per Definition 4.1, a bowling feature must be identified. The process of obtaining a bowling feature is to take the inner product of $F^{'}_{attacked}$ with every bowling vector of $G^{'}$, i.e., the row vector of *attacked* batting features in the $\langle F_{m\times2}, G_{n\times2}\rangle$ matrix (Refer to Figure 4.2). In other words, compute $\langle F^{'}_{attacked}, G^{'}_{j}\rangle$ for all $j \in \{good, short, full, off, leg, middle, spin, swing, fast, slow, movein, moveaway\}$. Here, $F^{'}_{attacked}$ is the row vector of the *attacked* batting feature in the $\langle F_{m\times2}, G_{n\times2}\rangle$ matrix and $G^{'}_{good}$ is the column vector of the *good-length* bowling feature in the $\langle F_{m\times2}, G_{n\times2}\rangle$ matrix. The bowling vector $G^{'}_{j}$, which yields the highest inner product ($\langle F^{'}_{attacked}, G^{'}_{good}\rangle$, $\langle F^{'}_{attacked}, G^{'}_{short}\rangle$, $\langle F^{'}_{attacked}, G^{'}_{full}\rangle$, $\cdots$, $\langle F^{'}_{attacked}, G^{'}_{moveaway}\rangle$), corresponds to the batsman's first strength rule. Similarly, the bowling vector that yields the second-highest inner product value corresponds to the batsman's second strength rule. The process is continued for all bowling features of the opponent players.

Weakness Rules of a Batsman

To qualify for a weakness rule for a batsman, the batting feature as given in Definition 4.3 must contain the *beaten* feature. In order to frame a complete rule as per Definition 4.1, a bowling feature must be identified. The process of obtaining bowling feature is to take the inner product of $F^{'}_{beaten}$ with every bowling vector of $G^{'}$, i.e., the row vector of the *beaten* batting feature in the $\langle F_{m\times2}, G_{n\times2}\rangle$ matrix (Refer to Figure 4.2). The bowling vector $G^{'}_{j}$, which yields the highest inner product, corresponds to the batsman's first weakness rule. Similarly, the bowling vector that yields the second-highest inner product value corresponds to the batsman's second weakness rule. The process is continued for all bowling features of the opponent players.

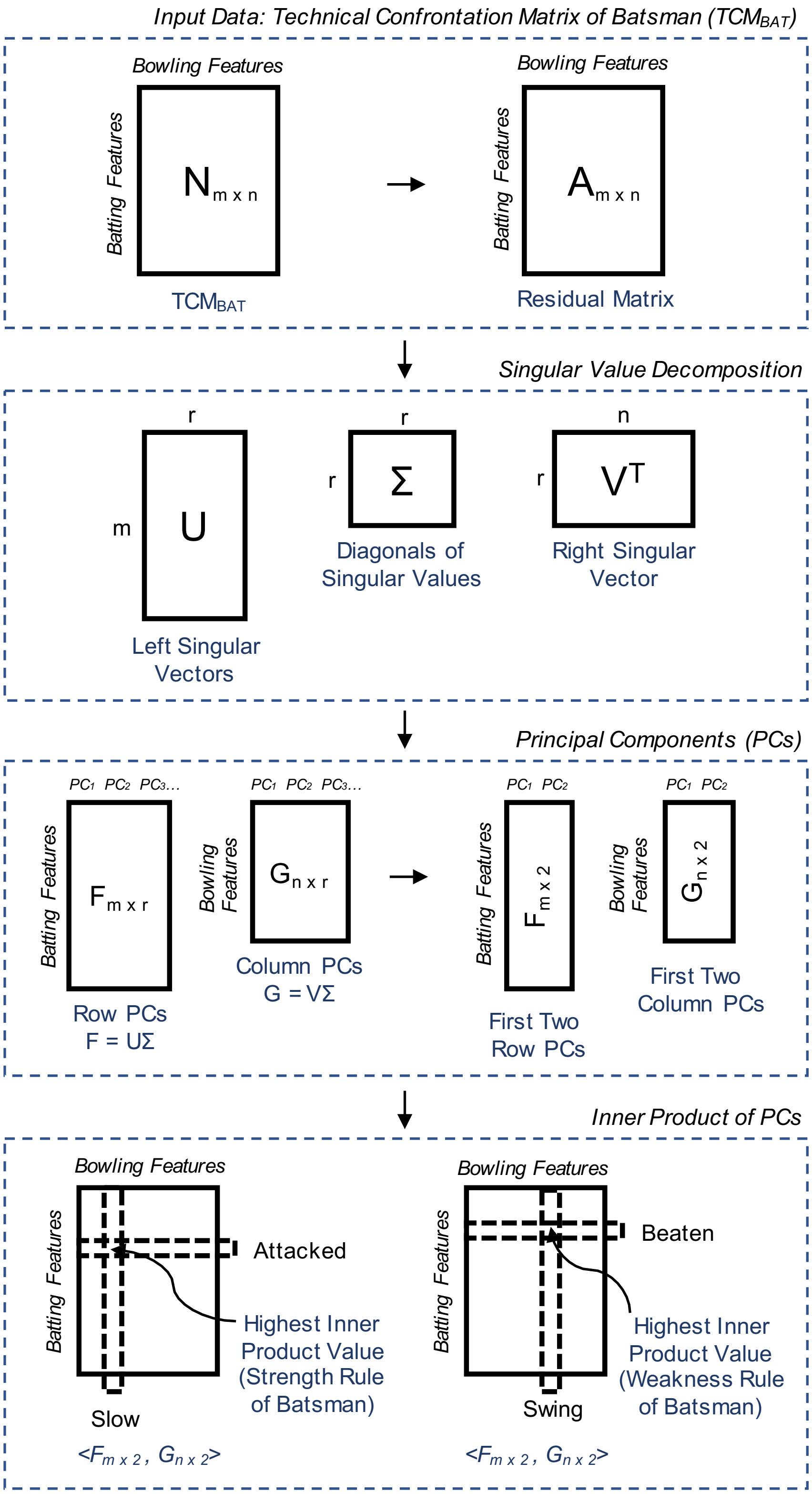


Figure 4.2: Batting Analysis through CA.

Other Rules of a Batsman

In addition to the strength rule and weakness rule, it is important to learn the other rules for batsman corresponding to the *response*, *outcome*, *footwork*, and *shot area* of the batsman. To obtain these rules, except *attacked* and *beaten*, all the other batting features are considered. In order to frame a complete rule for each of these batting features, as per Definition 4.1, a bowling feature must be identified. The process of obtaining a bowling feature is to take the inner product of $F^{'}_{batting\ feature}$ with every bowling vector of $G^{'}$, i.e., the row vector of the selected batting feature in the $\langle F_{m\times 2}, G_{n\times 2}\rangle$ matrix. The rest of the process remains the same as strength rule and weakness rule learning.

4.2.2 Bowling Analysis through CA

Refer to Figure 4.3 for the steps involved in bowling analysis through CA. For bowling analysis of a player, the TCM_{BOWL} (denoted as N) of the player is used to obtain the relationships between batting features and bowling features. CA first obtains the residual matrix A from N. Next, SVD is applied to A to obtain the batting principal components (F) and bowling principal components (G). Then, the first two principal components of F and G (denoted as $F_{m\times 2}\ or\ F^{'}\ and\ G_{n\times 2}\ or\ G^{'}$) are obtained. Finally, the inner product matrix ($\langle F_{m\times 2}, G_{n\times 2}\rangle$) of the first two principal components of F and G is obtained.

Strength Rules of a Bowler

To qualify for a strength rule for a bowler, the batting feature of opponent batsmen as given in Definition 4.4 must have the *beaten* feature. To frame a complete rule as per Definition 4.1, a bowling feature must be identified. The process of obtaining bowling feature is to take the inner product of $F^{'}_{beaten}$ with every bowling vector of $G^{'}$, i.e., the row vector of *beaten* batting feature in the $\langle F_{m\times 2}, G_{n\times 2}\rangle$ matrix (Refer to Figure 4.3). The bowling vector $G^{'}_{j}$, which yields the highest inner product, corresponds to the bowler's first strength rule. Similarly, the bowling vector that yields the second-highest inner product value corresponds to the bowler's second strength rule. The process is continued for all bowling features of the selected bowler.

Weakness Rules of a Bowler

To qualify for a weakness rule for a bowler, the batting feature of opponent batsmen as given in Definition 4.5 must have the *attacked* feature. To frame a complete rule as per Definition 4.1, a bowling feature must be identified. The process of obtaining bowling feature is to take the inner product of $F^{'}_{attacked}$ with every bowling vector of $G^{'}$, i.e., the row vector of *attacked* batting feature in the $\langle F_{m\times 2}, G_{n\times 2}\rangle$ matrix (Refer to Figure 4.3). The bowling vector $G^{'}_{j}$, which yields the highest inner product, corresponds to the bowler's first weakness rule. Similarly, the bowling vector that yields the second-highest inner product value corresponds to the bowler's second weakness rule. The process is continued for all bowling features of the selected bowler.

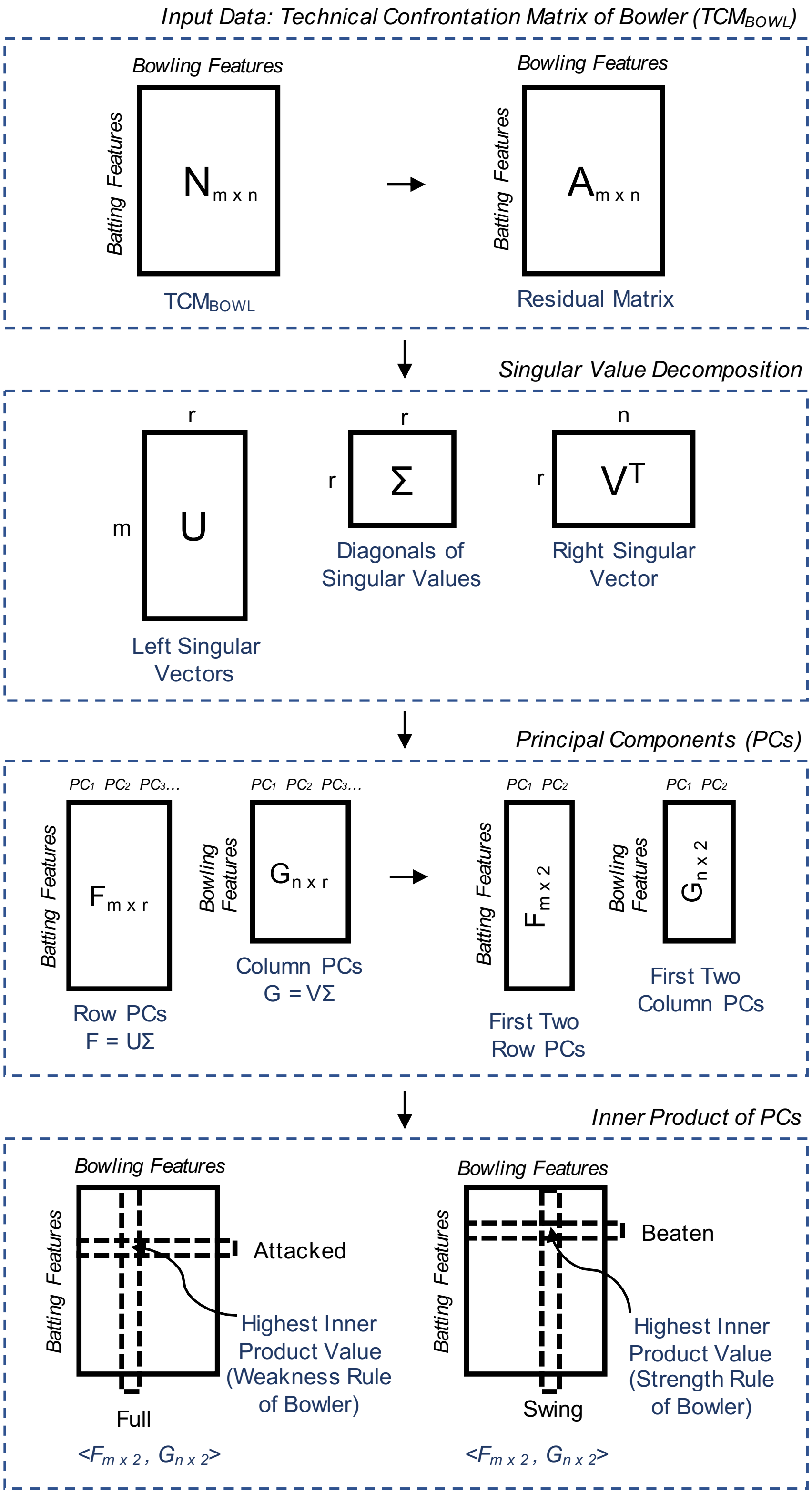


Figure 4.3: Bowling Analysis through CA.

Other Rules of a Bowler

In addition to the strength rule and weakness rule, it is important to learn the other rules for bowler corresponding to the *response*, *outcome*, *footwork*, and *shot area* of the opponent batsman. To obtain these rules, except *attacked* and *beaten*, all the other batting features are considered. In order to frame a complete rule for each of these batting features, as per Definition 4.1, a bowling feature must be identified. The process of obtaining bowling feature is to take the inner product of $F'_{batting\ feature}$ with every bowling vector of G', i.e., the row vector of the selected batting feature in the $\langle F_{m\times 2}, G_{n\times 2}\rangle$ matrix. The rest of the process remains the same as strength rule and weakness rule learning.

4.3 Visualization of Strength and Weakness Rules

In order to visually interpret the relationship between batting features and bowling features, the first two principal directions of F and G ($F'\, and\ G'$) are obtained from Algorithm 4 and plotted on a two-dimensional plot - biplot [26]. Refer to Figure 4.4 for an example biplot. The row (batting) and column (bowling) vectors having the highest inner product values are the closest vectors in the biplot. These two vectors constitute a strength rule or weakness rule. For better visualization, instead of having all the batting features in one plot, only a subset of batting features (each category of batting features such as response, outcome, footwork, and shot area) and all the bowling features are plotted (subset correspondence analysis [25]).

4.4 Experiments

This section presents case studies that illustrate the strength and weakness analysis using the proposed approach. First, we present the batting analysis of batsman Steve Smith. Next, we present the bowling analysis of bowler Kagiso Rabada. Finally, we present the batting analysis of batsman Steve Smith against fast and spin bowlers.

4.4.1 Batting Analysis - Steve Smith

Steve Smith is an Australian international cricketer who is consistently rated as one of the world's top-ranked Test batsmen. In his career (data collected till April 2019), he has played 63 Test matches, batted over 158 days, and faced 11198 deliveries. For Steve Smith, we have 11198 short text commentaries in our data. To perform Steve Smith's batting analysis, we obtain these short text commentaries using the filter tuple ⟨*Steve Smith, All Opponent Players, Career, Batting*⟩. Next, a TCM_{BAT} of size 19×12 is constructed in which rows correspond to Steve Smith's batting features, and the columns correspond to bowling features of opponent bowlers. Employing the proposed approach of batting analysis in Section 4.2.1, a biplot depicting the strengths and weaknesses of batsman Steve Smith is obtained (Refer to Figure 4.4).

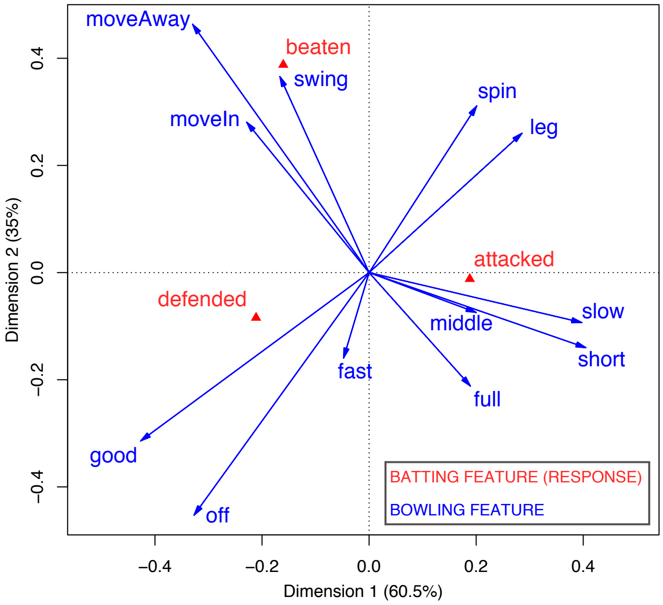
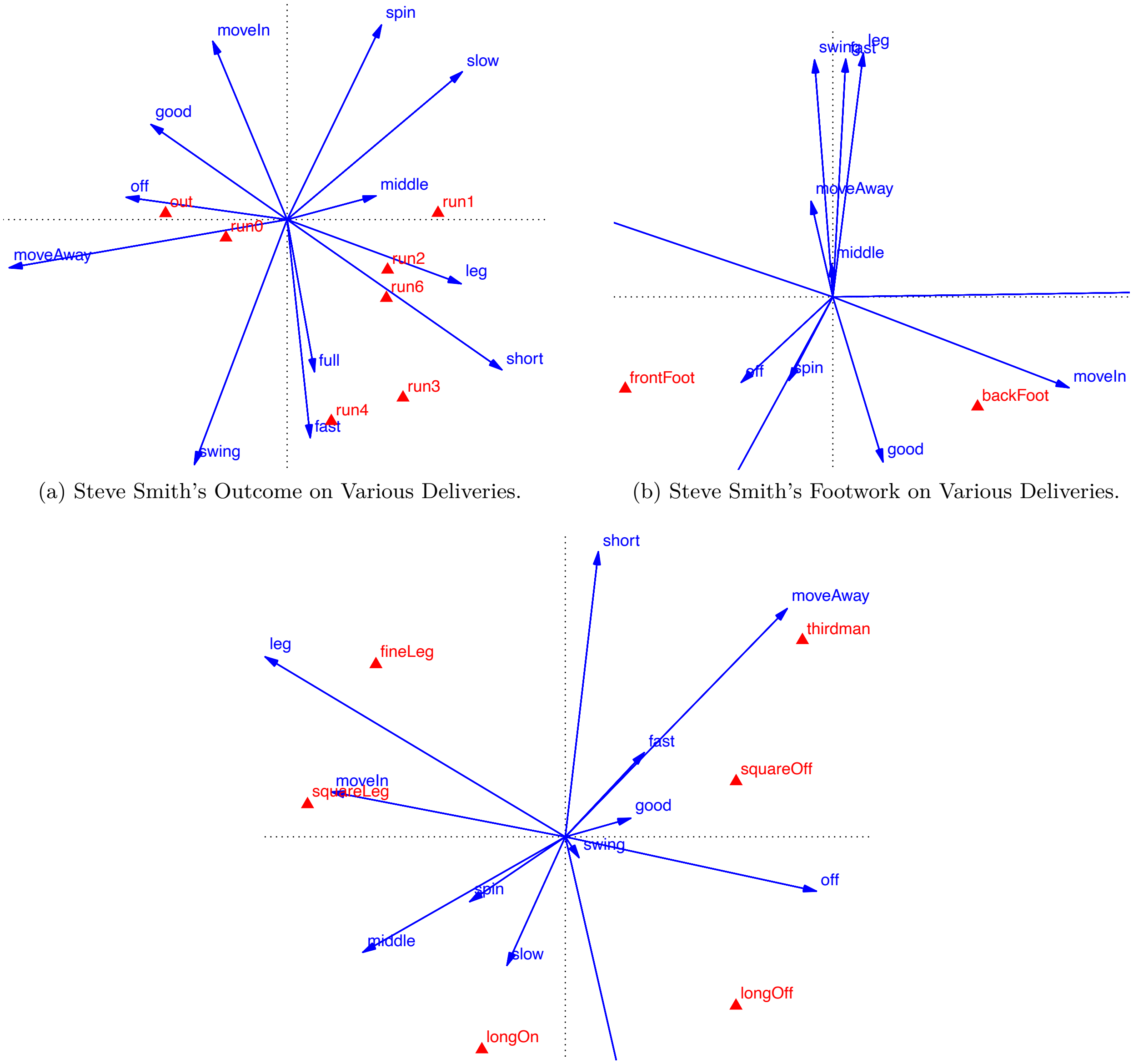


Figure 4.4: Steve Smith's Response on Various Deliveries.

In this biplot, the top three bowling vectors closer to $F_{attacked}$ vector in the decreasing order are G_{slow}, G_{short}, and G_{middle}. Following the Definition 4.2 and Definition 4.1 for strength rule construction, the proposed algorithm obtains the strength rule of Steve Smith as - *Steve Smith attacks slow, short-length, or middle-line deliveries.* The decreasing order of bowling features implies that the probability of Steve Smith attacking slow deliveries is higher than short-length or middle-line deliveries.

The top three bowling vectors closer to F_{beaten} vector in the decreasing order are G_{swing}, $G_{moveAway}$, and G_{moveIn}. Following the Definition 4.3 and Definition 4.1 for weakness rule construction, the proposed algorithm obtains the weakness rule of Steve Smith as - *Steve Smith gets beaten on the deliveries that are swinging, moving-away, or moving-in.* The decreasing order of bowling features implies that the probability of Steve Smith getting beaten on swing deliveries is higher than moving-away or moving-in deliveries.

In addition to the batsman's response, it is essential to note what is the *outcome* of a particular delivery, where the ball is being hit by the batsman (*shot area*), and also their *footwork*. We obtain biplots for Steve Smith's outcome (Refer to Figure 4.5a), footwork (Refer to Figure 4.5b), and shot area (Refer to Figure 4.5c), as well. Some of the rules obtained from these biplots are listed in Table 4.2. Note that Steve Smith is a right-handed batsman, so shot areas are interpreted as presented in Figure 2.1.



(a) Steve Smith's Outcome on Various Deliveries.

(b) Steve Smith's Footwork on Various Deliveries.

(c) Steve Smith's (Right-handed Batsman) Shot Area on Various Deliveries.

Figure 4.5: Steve Smith's Batting Analysis.

4.4.2 Bowling Analysis - Kagiso Rabada

Kagiso Rabada is a South African international cricketer who is consistently rated as one of the world's top-ranked Test bowlers. In his career (data collected till April 2019), he has played 36 Test matches, bowled over 108 days, and delivered 6910 deliveries. For Kagiso Rabada, we have 6910 short text commentaries in our data. To perform Kagiso Rabada's bowling analysis, we obtain these short text commentaries using the filter tuple ⟨*Kagiso Rabada, All Opponent Players, Career, Bowling*⟩. Next, a TCM_{BOWL} of size 19 × 12 is constructed in which rows correspond to batting features of opponent batsmen, and the columns correspond to bowling features of Kagiso

Feature	**Biplot**	**Obtained Rules**
Response	Figure 4.4	Attacks the slow, short-length, or middle-line deliveries Gets beaten on the swinging, moving-away, or moving-in deliveries Defends the good-length deliveries
Outcome	Figure 4.5a	Scores more runs on short-length or full-length deliveries Struggles to score runs and tends to get out on off-stump or moving-away deliveries
Footwork	Figure 4.5b	Plays moving-in deliveries on backfoot Plays off-line deliveries on frontfoot
Shot-area	Figure 4.5c	Plays slow deliveries to long-on area of the field Plays moving-in deliveries to square-leg area of the field

Table 4.2: Rules Obtained from Steve Smith's Batting Analysis.

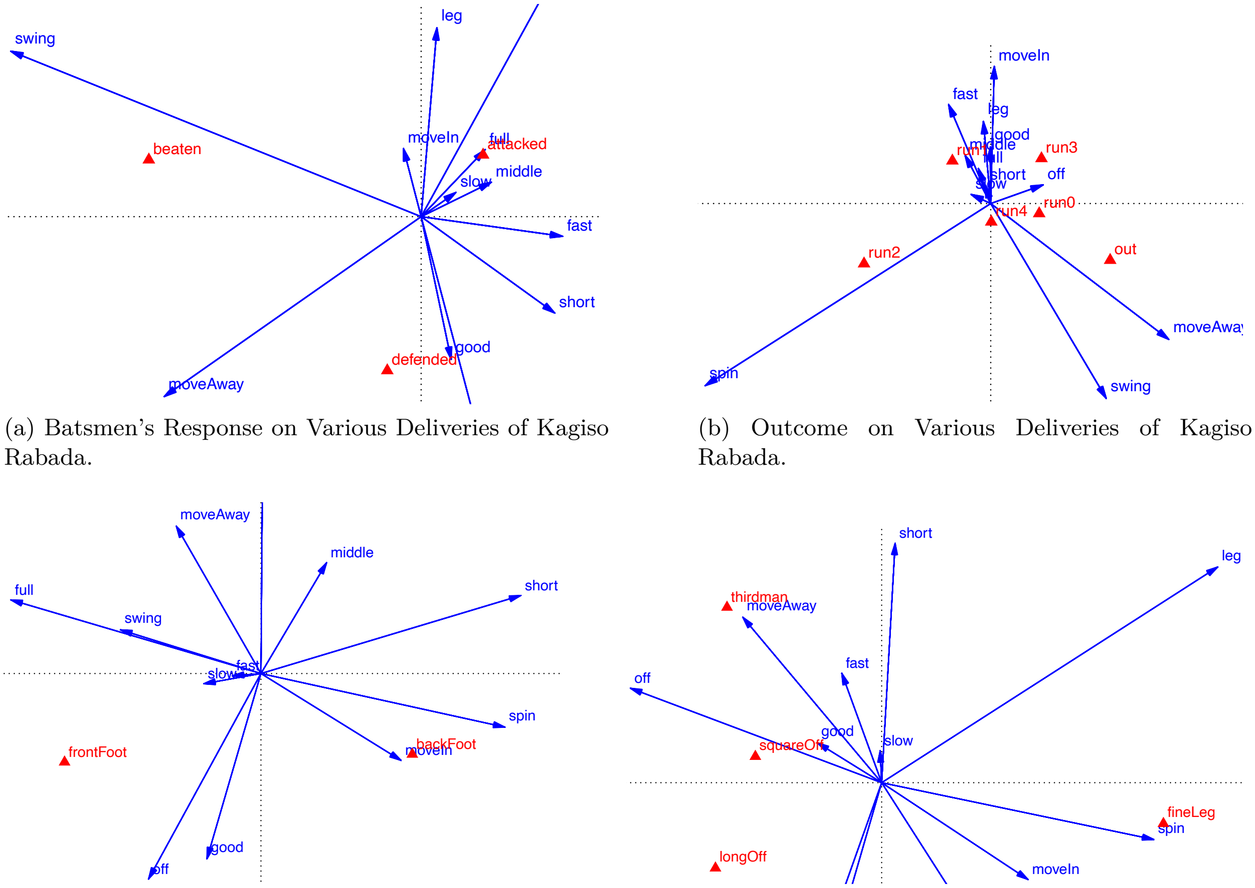
Rabada. Employing the proposed approach of bowling analysis in Section 4.2.2, a biplot depicting the strengths and weaknesses of bowler Kagiso Rabada is obtained (Refer to Figure 4.6a).

In this biplot, the bowling vector closest to F_{beaten} vector is G_{swing}. Following the Definition 4.4 and Definition 4.1 for strength rule construction, the proposed algorithm obtains the strength rule of Kagiso Rabada as - *batsmen get beaten on the swing deliveries of Kagiso Rabada*. The bowling vector closest to $F_{attacked}$ vector is G_{swing}. Following the Definition 4.2 and Definition 4.1 for weakness rule construction, the proposed algorithm obtains the weakness rule of Kagiso Rabada as - *batsmen attack full-length deliveries of Kagiso Rabada*.

In addition to the batsmen's response to Kagiso Rabada's deliveries, it is important to note what is the *outcome* of a particular delivery, where the ball is being hit by the batsman (*shot area*), and also a batsman's *footwork*. We obtain biplots for outcome (Refer to Figure 4.6b), footwork (Refer to Figure 4.6c), and shot area (Refer to Figure 4.6d), as well. Some of the rules obtained from these biplots are listed in Table 4.3.

The maximum number of rules obtained for one player using the proposed method is 228 (19 * 12). Out of these rules, 12 strength rules (inner product of attacked with 12 bowling features) and 12 weakness rules (inner product of beaten with 12 bowling features) are subject to evaluation. Out of these, six (first three strength rules and weakness rules) dominant rules having a high degree of dependency are presented. Strength and weakness rules for more than 250 players can be accessed at https://www.dropbox.com/sh/05zrr3sdz18mep8/AACBn8mjee3t2fyrhCb40wA2a?dl=0.

Effect of Filter Parameter The proposed method of rule mining is not just limited to a single player or match. It can be used to mine strength and weakness rules in various scenarios, such as against a player or a team, against a type of player, for different time frames (session, day, inning, match, season, year, and career), etc. We have to define new criteria in the filter parameters for selecting the set of deliveries to build the TCM. In previous analyses, we have established the



(a) Batsmen's Response on Various Deliveries of Kagiso Rabada.

(b) Outcome on Various Deliveries of Kagiso Rabada.

(c) Batsmen's Footwork on Various Deliveries of Kagiso Rabada.

(d) Batsmen's Shotarea on Various Deliveries of Kagiso Rabada.

Figure 4.6: Kagiso Rabada's Bowling Analysis.

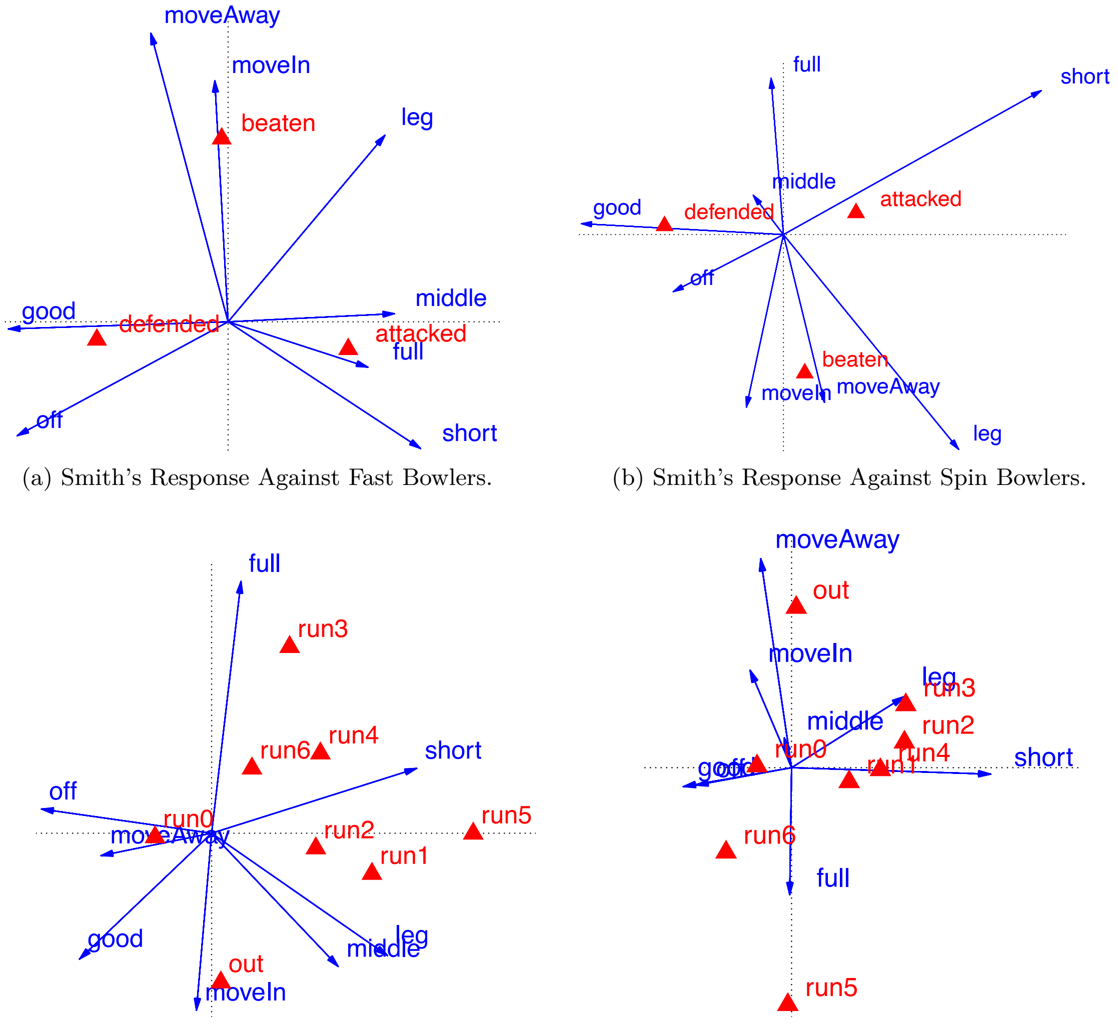
Feature	**Biplot**	**Obtained Rules**
Response	Figure 4.6a	Batsmen attack full-length deliveries of Kagiso Rabada Batsmen get beaten on the swing deliveries of Kagiso Rabada Batsmen defend good-length deliveries of Kagiso Rabada
Outcome	Figure 4.6b	Batsmen tend to get out on Kagiso Rabada's moving-away deliveries
Footwork	Figure 4.6c	Batsmen play Kagiso Rabada's moving-in deliveries on backfoot
Shot-area	Figure 4.6d	Batsmen play Kagiso Rabada's off-line deliveries to square-off area of the field Batsmen play Kagiso Rabada's moving-away deliveries to thirdman area of the field

Table 4.3: Rules Obtained from Kagiso Rabada's Bowling Analysis.

criteria as the set of deliveries played by a particular player. It can be made more specific. For example, to find out the strengths and weaknesses of Steve Smith while facing the bowling by a particular type of bowler (fast bowlers or spin bowlers), the proposed method remains the same. However, now we have to build the TCM using only those deliveries where the batsman is Steve Smith and the bowlers are of the selected type. A player can use such rules to design a very niche strategy while playing against a specific player type. This flexibility makes the proposed approach potentially much more helpful for coaches, players, and administrators in cricket. In the following subsection, we analyze batsman Steve Smith's strength and weakness rules against different bowling types (fast bowling and spin bowling).

4.4.3 Batting Analysis Against Fast Bowlers and Spin Bowlers - Steve Smith

In cricket, bowlers are mainly of two types: (i) *Fast bowlers* who bowl fast deliveries and *Spin bowlers* who bowl deliveries that are slow and turn after pitching. It makes more sense to construct



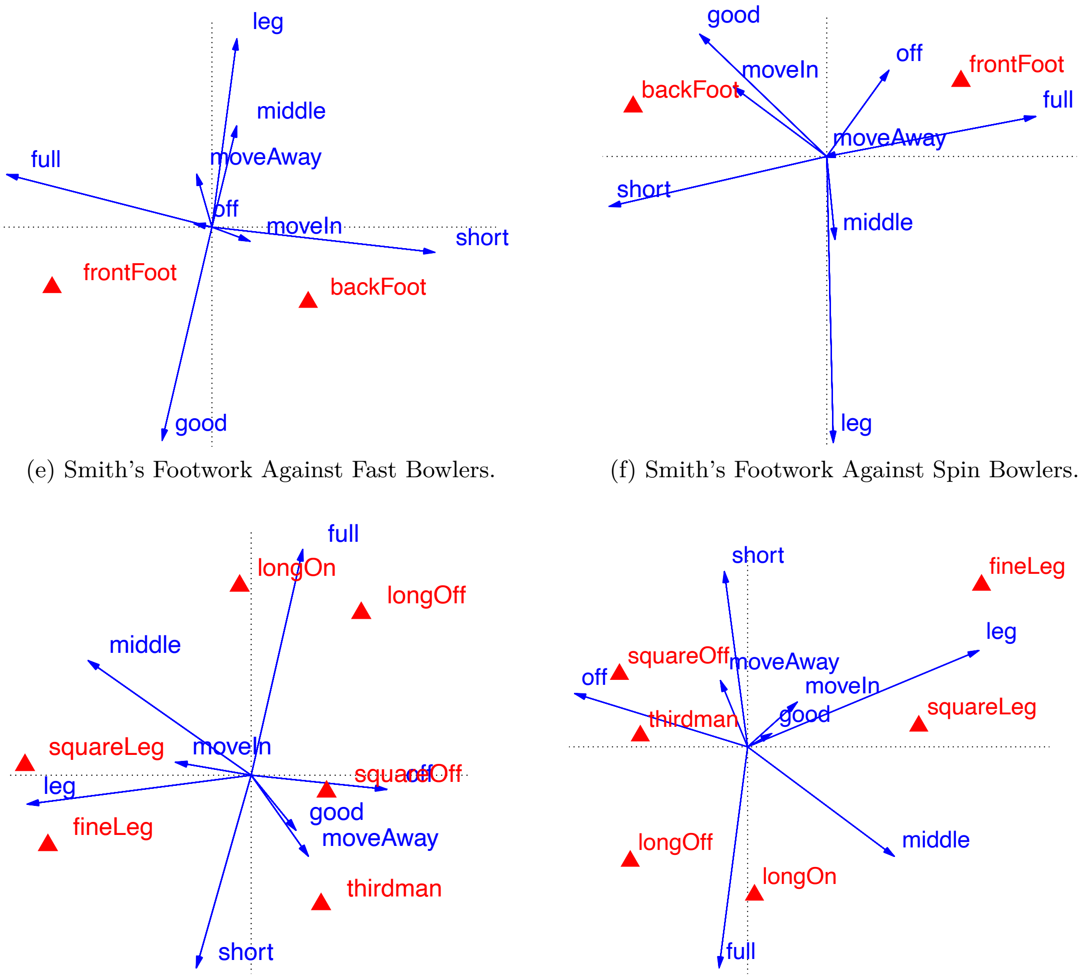
(a) Smith's Response Against Fast Bowlers.

(b) Smith's Response Against Spin Bowlers.

(c) Smith's Outcome Against Fast Bowlers.

(d) Smith's Outcome Against Spin Bowlers.

Figure 4.7: Steve Smith's Batting Analysis Against Fast Bowling and Spin Bowling.



(e) Smith's Footwork Against Fast Bowlers.

(f) Smith's Footwork Against Spin Bowlers.

(g) Smith's Shot-area Against Fast Bowlers.

(h) Smith's Shot-area Against Spin Bowlers.

Figure 4.7: Steve Smith's Batting Analysis Against Fast Bowling and Spin Bowling.

the strength and weakness rules differently against the fast bowlers and the spin bowlers. For this analysis, two confrontation matrices are constructed, one where the deliveries are bowled by fast bowlers (Filter Tuple: ⟨*Steve Smith, Fast Bowlers, Career, Batting*⟩) and another where spin bowlers bowled the deliveries (Filter Tuple: ⟨*Steve Smith, Spin Bowlers, Career, Batting*⟩). The bowling features fast, slow, spin, and swing are omitted from these confrontation matrices as the fast and spin bowling types convey this information. Using the proposed method, the biplots of batsman Steve Smith against fast bowlers and spin bowlers are obtained (Refer to Figure 4.7). Some of the rules obtained from these biplots are listed in Table 4.4. The data and results generated during this analysis (for other batsmen also) can be accessed at `https://www.dropbox.com/s/s2ruurn6doawysw/PACE-VS-SPIN-Analysis.zip?dl=0`.

Feature	**Fast Bowlers**	**Spin Bowlers**
Response	Attacks full-length deliveries Gets beaten on moving-in deliveries	Attacks short-length deliveries Gets beaten on moving-away deliveries
Outcome	Gets out on moving-in deliveries	Gets out on moving-away deliveries
Footwork	Plays moving-in deliveries on the back foot	Plays moving-in deliveries on the back foot
Shot-area	Plays full-length deliveries to the long-on and long-off areas of the field	Plays full-length deliveries to the long-on and long-off areas of the field

Table 4.4: Rules Obtained from Steve Smith's Batting Analysis Against Fast Bowling and Spin Bowling.

4.5 Validation

The obtained strength and weakness rules, though easy to interpret, are difficult to validate. No loss function exists which captures the risk associated with each of the obtained rules. Cricket experts' opinion matters the most in judging the derived rules. We validate the derived rules in two distinct ways: extrinsic and intrinsic.

4.5.1 Extrinsic Validation

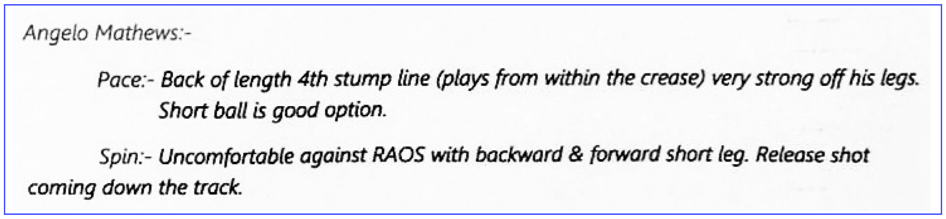
For extrinsic validation, we verify the identified rules against external sources. The main bottleneck is the absence of trustable gold standard data about every cricket player's strengths and weaknesses. However, we could get a couple of such resources where domain experts have directly mentioned strength and weakness rules for some cricket players, which are available in the public domain.

Strategy Sheet

Deccan Chronicle, a well known Indian newspaper, published an article [1] titled - "*Sri Lanka vs India: Unattended document leaks Virat Kohli and Co's Galle Test plans*" on 16th September, 2017. This article contains the Indian cricket team's strategy sheet for the Sri Lankan batsmen during the test matches of the India tour of Sri Lanka, 2017. The strategy sheet lists the strengths and weaknesses of eight Sri Lankan batsmen. It is a rare case where strengths and weaknesses analyzed by an international cricket team are made public. Note that there is no algorithmic method at the disposal of cricket team management to obtain the strategy sheet. These rules are the summary of the experience brought to the table by team coaches and players collectively.

A section of the strategy sheet is shown in Figure 4.8. It contains strategy for batsman *Angelo Mathews* and the strategy is to bowl *back of length* (short-length), *4th stump line* (outside off stump) and RAOS (off-spin) deliveries. Further, he is very strong off his legs (scores a lot of runs).

[1] https://www.deccanchronicle.com/sports/cricket/160917/leaked-heres-the-unattended-document-revealing-virat-kohli-led-team-indias-plans.html



Angelo Mathews:-

Pace:- Back of length 4th stump line (plays from within the crease) very strong off his legs. Short ball is good option.

Spin:- Uncomfortable against RAOS with backward & forward short leg. Release shot coming down the track.

Figure 4.8: A Section of the Strategy Sheet.

Batsman	Commonality Percentage (%)
Angelo Mathews	80.00
Asela Gunaratne	75.00
Dhananjaya De Silva	75.00
Kusal Mendis	75.00
Niroshan Dickwella	66.66
Dilruwan Perera	60.00
Dimuth Karunaratne	50.00
Upul Tharanga	50.00

Table 4.5: Extrinsic Validation Using Strategy Sheet.

We obtain the strength and weakness rules for all the mentioned batsmen using the proposed method. Refer to Table 4.5 which lists the overlap (Commonality Percentage) of the obtained rules with the rules listed in the strategy sheet. We observe that for *Angelo Mathews*, the overlap is 80%. For other batsmen, also we see a high degree of overlap.

Expert Analysis

For expert analysis, the rules identified by using the proposed method are verified against expert sources. One video blog from a domain expert in which the expert shares the strength and weakness rules of Steve Smith is identified and its analysis is presented below. *Sanjay Manjrekar* is a former Indian cricket player with significant domain experience and expertise. He is a regular television commentator for international cricket matches. He has published a video[2] with ESPNCricInfo titled -"*What is Steve Smith's weakness?*" on 1^{st} June, 2017. In this video, the author has provided one strength rule and one weakness rule. The strength and weakness rules are extracted from the video. These rules are then compared with the ones obtained from the proposed method. Following is a transcript of selective parts of the video, along with the rules obtained by the proposed method.

1. Weakness Rule Validation:

   - *Proposed Method:* "Steve Smith attacks deliveries that are bowled on the *middle stump*".
   - *Expert Analysis (Video time* $0.22''$*):* "Bowlers tend to attack him on the stump (*middle stump* line), but then his wonderful angle of the bat carves everything on the leg side.

[2] https://www.espncricinfo.com/video/what-is-steven-smith-s-weakness-1100538

Batsman	$\boldsymbol{\Delta^2_{12_{2013}}}$	$\boldsymbol{\Delta^2_{12_{2014}}}$	$\boldsymbol{\Delta^2_{12_{2015}}}$	$\boldsymbol{\Delta^2_{12_{2016}}}$	$\boldsymbol{\Delta^2_{12_{2017}}}$	$\boldsymbol{\Delta^2_{12_{2018}}}$	$\boldsymbol{\Delta^2_{12_{avg}}}$
Joe Root	0.09	0.18	0.12	0.13	0.17	0.19	**0.15**
Cheteshwar Pujara	0.15	0.14	0.37	0.10	0.16	0.18	0.18
Steve Smith	0.35	0.26	0.25	0.18	0.21	0.39	0.27
Dimuth Karunaratne	0.68	0.28	0.15	0.25	0.20	0.30	0.31
Dean Elgar	0.76	0.58	0.17	0.19	0.14	0.21	0.34
Virat Kohli	0.18	0.71	0.24	0.28	0.41	0.20	0.34
Kane Williamson	0.28	0.16	0.38	0.60	0.42	0.42	0.38
David Warner	0.65	0.47	0.38	0.43	0.19	0.51	0.44

Table 4.6: Intrinsic Validation (K-fold Cross Validation).

Gets lot of runs on the leg side". In other words, Steve Smith scores lot of runs on the leg side for balls bowled on the middle stump.

2. Strength Rule Validation:

- *Proposed Method:* "Steve Smith gets beaten on the move-away and move-in (*seam*) deliveries".
- *Expert Analysis (Video time* $0.49''$*):* "Bowl *seamers* as much as possible".

Both the strength and weakness rules provided by the expert are validated with the proposed method.

4.5.2 Intrinsic Validation

For intrinsic validation, we use the k-fold cross validation strategy, in which text commentary data is divided into k (number of years) subsets, and the holdout method is repeated k times. Each time, one of the k subsets (one year of text commentary data) is used as a test set with the remaining data as a training set. Finally, the average error across all k trials is computed.

The collected text commentaries span over thirteen years, from May 2006 to April 2019. As the data for the year 2019 is incomplete, we have not considered 2019 for the intrinsic validation. The top-ranked batsmen in Test cricket are considered for intrinsic validation. Most of the batsmen in this list started playing Test cricket regularly from the year 2013. Hence, for intrinsic validation, six years (2013 to 2018) are considered. Each time, one year of text commentaries are used as a test set and remaining text commentaries as a training set. For example, if text commentaries for 2013 are selected as a test set, then the remaining text commentaries (from 2014 to 2018) are selected as the training set. This method is repeated for all six years.

The proposed rule mining method is applied to the training data, and a training-biplot is obtained. Similarly, the proposed rule mining method is applied to the test data, and a test-biplot is obtained. Procrustes analysis [122] is used to compare these two biplots. The main idea is to minimize the sum-of-squared differences between the two biplots. This test performs the least square superimposition of one biplot to another reference biplot. The lower the sum of the squared residual, the greater the similarity between the two biplots. The biplots obtained from the training data and test data are compared using the procrustes test. When there is a similarity in the obtained biplots, it is considered that the proposed method is reliable.

Table 4.6 presents the sum of squared residuals for each year ($\Delta^2_{12_{year}}$) and average sum of squared residuals ($\Delta^2_{12_{avg}}$) for eight batsmen. The lowest average sum of the squared residual is 0.15 for batsman Joe Root, implying his training and test data sets have the highest similarity. The highest average sum of the squared residual is 0.44 for batsman David Warner, implying his training and test data sets have the highest dissimilarity.

In both extrinsic and intrinsic validation, we obtain high values in terms of the derived rules suggesting the accuracy of the proposed method in mining individual player's strength and weakness rules. The data and results generated during the validation process can be accessed at https://www.dropbox.com/sh/sa4xmz3pqf6y9np/AABf5TzE1cbU_RgrFj7pxG3pa?dl=0.

4.6 Baseline Comparison

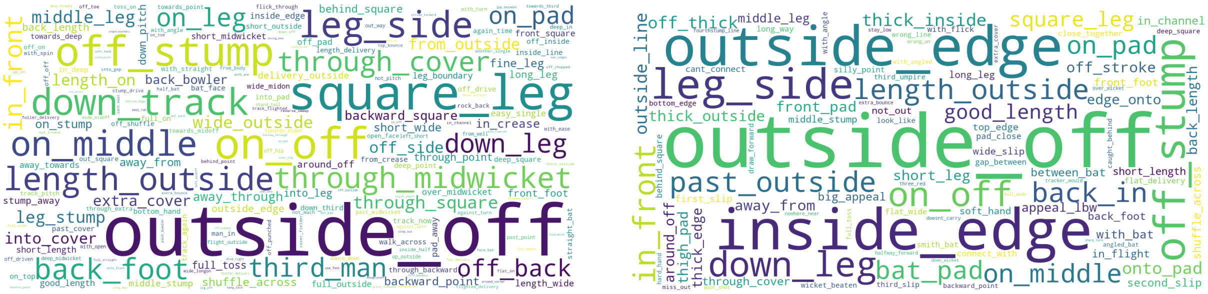
In Section 4.5, the validation of the obtained rules is performed. This section presents baseline comparisons of the strength and weakness rules using the wordclouds and association rule mining.

4.6.1 Strength and Weakness Visualization using Wordclouds

One popular text visualization technique is the word cloud [101]. Word clouds display the relative word frequency or the importance of a word by its font size. It presents a visual summary of document content.

The word cloud for strength is obtained based on the frequency of occurrence of technical features, and strength rules are constructed. The short text commentaries in which the bigrams related to *attacked* feature are present are selected. Using this text data, the strength word cloud is obtained. The strength word cloud for batsman Steve Smith is presented in Figure 4.9a. The top two frequently occurring bigrams in this word cloud are: *outside_off* and *square_leg*. The most frequently occurring bigrams are interpreted as strength rules for Steve Smith. The first strength rule is - *Smith attacks deliveries that are bowled outside off stump*. The second strength rule is - *Smith attacks the balls to the square_leg area.*

Similarly, the weakness word cloud is obtained by considering the short text commentaries in which the bigrams related to *beaten* feature is present. Using this text data, the weakness word cloud is obtained. The weakness word cloud for batsman Steve Smith is presented in Figure 4.9b. The top two frequently occurring bigrams in the word cloud are: *outside_off* and *inside_edge*. The most frequently occurring bigrams are interpreted as weakness rules for Steve Smith. The first weakness rule is - *Smith gets beaten by deliveries that are bowled outside off stump*. The second weakness rule is - *Smith gets an inside_edge frequently.*



(a) Strength Word Cloud.

(b) Weakness Word Cloud.

Figure 4.9: Bigram Word Clouds for Batsman Steve Smith.

Though the obtained rules are interpretable, contradictions are observed in the constructed strength and weakness rules using the word cloud based visualization method. For instance, the first strength rule and first weakness rule for Steve Smith are identical. However, the same rule cannot be interpreted as strength and weakness simultaneously. In cricket, bowlers frequently deliver outside off stump ball, and this is a well-known fact. This makes the constructed rule trivial, and the confidence in such rules will be low in practice. This limits the use of word cloud based visualization methods. Contradictions will not be observed when the proposed method is followed for rule construction and is evident from the strength and weakness rules presented for Steve Smith through biplot. Bigram word clouds for few other players are provided in https://www.dropbox.com/s/srim227j149njjp/Bigram%20Wordclouds.zip?dl=0.

4.6.2 Strength and Weakness Association Rules

We apply the Association Rule Mining (ARM) technique to construct rules that account for a player's strengths and weaknesses using short text commentary data. We analyze the association of strength/weakness exhibited by batsmen with the type of delivery they have faced. In essence, we investigate the bowling features that may be associated with the batsman's batting features.

Agrawal et al. [123] introduced ARM to discover interesting co-occurrence between products in supermarket data (market basket analysis). ARM extracts frequent sets of items that are purchased together and generates association rules of the form $A \implies B$, where A and B are disjoint sets of items, and B is likely to be purchased whenever A is purchased [124, 125, 126]. ARM is widely used in many domains, such as health care [127, 128, 129], financial transactions [130, 131], and retail [132, 133], etc. ARM is applied in the sports domain as well [134, 135, 136]. In cricket, Raj et al. [137] used ARM to find the association between the factors in cricket matches such as toss outcome and playing conditions with the game's outcome. UmaMaheswari et al. [138] modeled an automated framework to identify correlations among play patterns in cricket. Using their model, they have learned association rules for an individual player and generic rules for all players. In the literature, ARM has not been applied on sports text commentary to identify player-specific rules.

For player-specific analysis, we extract a subset of short text commentaries using the filter tuple ⟨Player, Opponent Player, Time, Type⟩. As discussed in Chapter 3, each delivery is represented as a set of extracted bowling and batting features, similar to the set of items representing a transaction in ARM (Example: *fullLength legStump fast attacked*). This is the input for ARM.

We provide computational definitions of the strength and weakness rules and then use ARM to construct player-specific rules given the definitions.

Definition 4.6 *Rule. In the association rule $A \implies B$, when A comprises a set of bowling features and B comprises a batting feature.*

Definition 4.7 *Strength Rule of Batsman. In Definition 4.6, when B or batting feature of the player (batsman) corresponds to attacked.*

Definition 4.8 *Weakness Rule of Batsman. In Definition 4.6, when B or batting feature of the player (batsman) correspond to beaten.*

Definition 4.9 *Strength Rule of Bowler. In Definition 4.6, when B or batting feature of the opponent players (batsmen) corresponds to beaten.*

Definition 4.10 *Weakness Rule of Bowler. In Definition 4.6, when B or batting feature of the opponent player (batsmen) corresponds to attacked.*

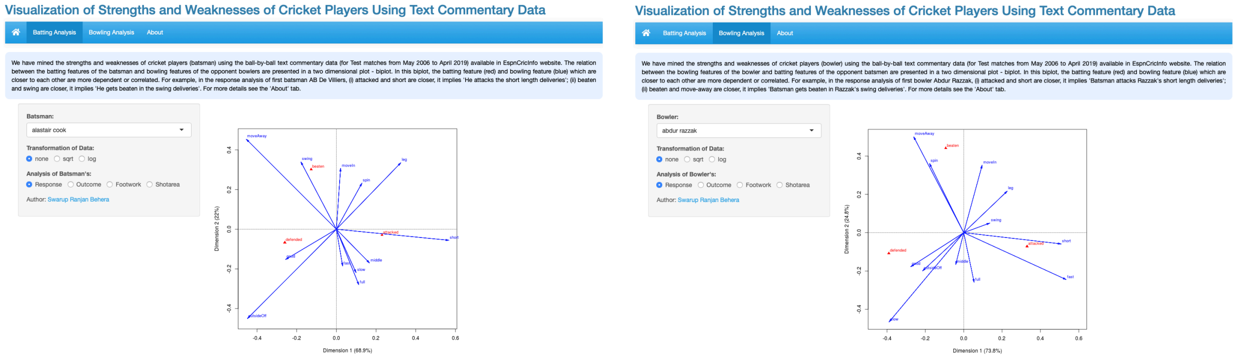
For constructing the strength and weakness association rules, we use the apriori algorithm [123]. The parameters on which the strength of the association of $A \implies B$ is dependent are - (i) *Support* is an indication of how frequently A and B appear in the dataset, (ii) *Confidence* is an indication of how often the rule is true, i.e., the conditional probability of occurrence of B given A, and (iii) *Lift* is the rise in the probability of having B with the knowledge of A being present over the probability of having B without any knowledge about the presence of A. A lift value greater than 1 signifies high association between A and B. In this work, the support for the analysis is varied from 0.001 to 0.1 and the confidence for the analysis is set at 0.5. The analysis has resulted in some interesting results, giving insights into the player's strengths and weaknesses.

The results of the strength and weakness analysis for batsman Steve Smith against all bowlers in Test matches are presented in Table. 4.7. The strength rule of Steve Smith is - *Smith attacks slow and shot-length deliveries.* The weakness rule of Steve Smith is - *Smith gets beaten on the good-length and swinging deliveries.* Both the strength rule and weakness rule are similar to the rules obtained using the proposed method.

Similarly, we can obtain rules other than strengths and weaknesses by choosing the consequent of the association rule as other batting features such as footwork, shot area, and outcome. We present these rules for batsman Steve Smith in Table. 4.7. Similar strength and weakness analyses can be performed for the bowlers as well. The code and result of ARM analysis for more than 250 players are provided in https://www.dropbox.com/sh/dq981ub7gdh3n04/AADIhE6cph8sVxXgL-e6Bt6ba?dl=0.

Association Rule (**A** $\implies$ **B**)	**Support(%)**	**Confidence(%)**	**Lift**
{shortlength, slow} $\implies$ {attacked}	2.6	72.1	1.7
{goodlength, swing} $\implies$ {beaten}	0.1	37.9	3.9
{offstump} $\implies$ {defended}	26.8	59.8	1.3
{fast, shortlength} $\implies$ {backfoot}	9.2	91.4	1.9
{fulllength, offstump} $\implies$ {frontfoot}	8.5	81.9	1.6
{fast, offstump} $\implies$ {0run}	22.8	82.2	1.2
{fast, offstump} $\implies$ {squareoff}	11.2	50.4	1.4
{legstump, slow} $\implies$ {squareleg}	17.8	86.9	2.3
{legstump, movein, spin} $\implies$ {fineleg}	0.01	100	23.7

Table 4.7: Identified Strength and Weakness Association Rules for Batsman Steve Smith.



Visualization of Strengths and Weaknesses of Cricket Players Using Text Commentary Data

We have mined the strengths and weaknesses of cricket players (batsman) using the ball-by-ball text commentary data (for Test matches from May 2006 to April 2019) available in EspnCricInfo website. The relation between the batting features of the batsman and bowling features of the opponent bowlers are presented in a two dimensional plot - biplot. In this biplot, the batting feature (red) and bowling feature (blue) which are closer to each other are more dependent or correlated. For example, in the response analysis of first batsman AB De Villiers, (i) attacked and short are closer, it implies 'He attacks the short length deliveries'; (ii) beaten and swing are closer, it implies 'He gets beaten in the swing deliveries'. For more details see the 'About' tab.

Batsman: alastair cook

Transformation of Data: ● none ○ sqrt ○ log

Analysis of Batsman's: ● Response ○ Outcome ○ Footwork ○ Shotarea

Author: Swarup Ranjan Behera

(a) Batting Analysis.

Visualization of Strengths and Weaknesses of Cricket Players Using Text Commentary Data

We have mined the strengths and weaknesses of cricket players (bowler) using the ball-by-ball text commentary data (for Test matches from May 2006 to April 2019) available in EspnCricInfo website. The relation between the bowling features of the bowler and batting features of the opponent batsmen are presented in a two dimensional plot - biplot. In this biplot, the batting feature (red) and bowling feature (blue) which are closer to each other are more dependent or correlated. For example, in the response analysis of first bowler Abdur Razzak, (i) attacked and short are closer, it implies 'Batsman attacks Razzak's short length deliveries'; (ii) beaten and move-away are closer, it implies 'Batsman gets beaten in Razzak's swing deliveries'. For more details see the 'About' tab.

Bowler: abdur razzak

Transformation of Data: ● none ○ sqrt ○ log

Analysis of Bowler's: ● Response ○ Outcome ○ Footwork ○ Shotarea

Author: Swarup Ranjan Behera

(b) Bowling Analysis.

Figure 4.10: Visualization of Player's Strength and Weakness Rules.

4.7 Web Application

A web-based system (Figure 4.10) is implemented to visualize player's strength and weakness rules. Users can select the batting analysis (Figure 4.10a) or the bowling analysis (Figure 4.10b). The left panel displays a drop-down menu from which the batsman and bowler can be selected, which will load the selected player's TCM from the server's database. The biplots (outcome, response, footwork, and shot-area) of the selected batsman or bowler are displayed in the main panel. The system is live at https://cricketvisualization.shinyapps.io/StrengthWeaknessAnalysis/.

4.8 Discussion

In this chapter, we have learned players' strength and weakness rules. However, the key questions are: (Q1) Which rules are 'fact-like' and which are 'insightful'? (Q2) Are the rules obtained for a particular player similar to the rules obtained for other players? and (Q3) Does the quality of opposition influence the player's strength and weakness rules?

Batsman	Strength (Bowling Features)	Weakness (Bowling Features)
Joe Root	short, full	swing, move-away
Dimuth Karunaratne	slow, short, full	swing
Steve Smith	slow, short, middle	swing, moving-away, moving-in
Cheteshwar Pujara	slow	move-in
Dean Elgar	short, middle	swing, move-away
Virat Kohli	middle, slow	swing
David Warner	short, fast, full	spin
Kane Williamson	short, slow	swing, move-in, move-away

Table 4.8: Strength and Weakness Rules (Maximum Three Bowling Features, in which Batsmen have Shown Strength or Weakness) of Top ICC Ranked Batsmen.

Batsman	Bowling Type	Strength (Bowl-Feat)	Weakness (Bowl-Feat)
Joe Root	Fast	full	leg
	Spin	short	leg
Dimuth Karunaratne	Fast	middle	move-in
	Spin	short	leg
Steve Smith	Fast	full	move-in
	Spin	short	move-away
Cheteshwar Pujara	Fast	middle	move-in
	Spin	short	move-away
Dean Elgar	Fast	full	move-in
	Spin	leg	move-away
Virat Kohli	Fast	full	move-in
	Spin	full	leg
David Warner	Fast	short	move-in
	Spin	short	leg
Kane Williamson	Fast	short	move-in
	Spin	short	leg

Table 4.9: Strength and Weakness Rules (First Bowling Feature, in which the Batsman has Shown Strength or Weakness) of Batsmen against Fast and Spin Bowling.

1. **Which rules are ‘fact-like’ and which are ‘insightful’?** The objective of this work is to construct strength and weakness rules, which are insightful. Proposed methods in this work also help construct rules involving outcome, footwork, and shot area. Of these, rules which include footwork or shot-area might fall in the ‘fact-like’ category. However, this specific distinction has not been carried out in this thesis. We acknowledge the need for identifying ‘fact-like’ rules and the need to differentiate them from ‘insightful’ rules. A detailed investigation of ‘fact-like’ and ‘insightful’ rules is left for future work.

2. **Are the rules obtained for a particular player similar to the rules obtained for other players?** We present strength and weakness rules (maximum three bowling features, in which the batsmen have shown strength or weakness) of top ICC ranked batsmen in Table 4.8. For Steve Smith we observe that he shows strength on slow or short-length or middle-line deliveries. The decreasing order of bowling features implies that the probability of Steve Smith attacking slow deliveries is higher than short-length or middle-line deliveries. For different players, we observe differences in obtained rules. However, some repetitions are present in the type of deliveries, such as slow and swing. To resolve this, we have constructed batsmen's strength and weakness rules separately against fast bowlers and spin bowlers (Section 4.4.3). The focus of the analysis is on the length, line, and movement of the delivery. We present the rules of the selected batsmen against fast bowlers and spin bowlers in Table 4.9. From this table, we can see the difference in strength and weakness rules for different players. From the above discussion, we infer that the proposed strength and weakness rule learner is capable of obtaining distinct rules for each player.

3. **Does the quality of opposition influence the player's strength and weakness rules?** Quality of opponent player may certainly be a factor influencing individual player's strength and weakness rules. In the present work, this factor is not considered for rule construction. We came across specific instances in the literature suggesting the influence of quality of players on team selection and performance of batsmen and bowlers. In particular, the influence of quality of player to model a team recommendation system was proposed by Chhabra et al. [51]. They utilized player's past performance and opponent player's strengths and weaknesses (qualitative factors) to obtain a team recommendation model. Lemmer [139] modeled performance of batsmen and bowlers utilizing strength of the opponent team, along with other factors. A detailed investigation of quality as an additional parameter in strength and weakness rule construction is left for future work.

4.9 Summary

This chapter presented an approach to learn player's strength and weakness rules using short text commentary data. First, the approach introduced the computationally feasible definitions of these rules. Next, it employed CA to construct semantic relations between batting features and bowling features. Finally, the approach plotted these relations using biplots and extracted human-readable strength and weakness rules. Several case studies showed that the proposed approach could identify players' strengths and weaknesses in individual matches and careers. The obtained rules are validated using intrinsic and extrinsic methods. Additionally, baseline comparisons are made using word clouds and association rule mining techniques. The obtained rules will help analysts, coaches, and team management build game strategies. This work has been published in [140].

5

Mining Temporal Changes in Strength and Weakness Rules of Cricket Players

In the last chapter, we presented an approach to learn the strength and weakness rules of cricket players using short text commentary data. It considers players' strengths and weaknesses in their entire careers. It is of interest to understand players' strengths and weaknesses as a function of *time*, primarily because these are not constant throughout their career. Over time players evolve in the sense that they work on their weaknesses and overcome them. So it is of interest to find out the traits they lost or acquired over time.

This chapter presents an approach to learn the temporal changes in cricket players' strength and weakness rules. For a given player, the presented approach proceeds in two steps. The first step involves the construction of a time-dependent confrontation tensor for the temporal analysis. Time granularity in cricket is identified in the increasing order as over, session, day, inning, match, series, season, year, or career. The year-wise analysis is the focus of this chapter. It first constructs the year-wise TCMs and models them as a three-dimensional Technical Confrontation Tensor (TCT) in which batting features, bowling features, and time (in years) are captured. Without this three-dimensional tensor modeling, the multiple interactions among the batting features, bowling features, and time can not be retrieved. In the second step, the obtained three-dimensional TCT is subject to a dimensionality reduction method, three-way correspondence analysis [23, 27] and semantic relations between batting features, bowling features, and time (years) are obtained. These relations are plotted in a line plot to visualize the year-wise changes in strength and weakness rules. Validation of the obtained results is impractical due to the absence of ground truth on year-wise change in player's strengths and weaknesses.

We use the proposed approach to mine temporal changes in strength and weakness rules corresponding to 131 batsmen and 129 bowlers who have batted/bowled more than 2000 deliveries. The data, code, and result of the experiments can be accessed at https://www.dropbox.com/sh/1l3jd7icbx241oj/AADMGaE-tdGNjS6pInnXUDKea?dl=0. We highlight some of the results of year-wise analysis for batsman Steve Smith and bowler Kagiso Rabada on full-length deliveries.

- Batting Analysis of Steve Smith on Full-length Deliveries
  - *Year-wise Changes in Strength Rule:* He has shown an increase in the trend of attacking *full-length* deliveries between the years 2013 and 2015 (both inclusive). However, in 2016 he struggled on *full-length* deliveries. He has once again shown an increase in the trend of attacking *full-length* deliveries between the years 2017 and 2018 (both inclusive).
  - *Year-wise Changes in Weakness Rule:* He has never exhibited weakness on *full-length* deliveries.

  To verify the year-wise results in terms of runs scored by the batsman, a year-wise runs-per-ball metric for full-length deliveries (total runs scored in full-length balls divided by total full-length balls played in a year) is obtained. For Steve Smith, on full-length deliveries, year-wise runs-per-ball values are: 0.43 (2013), 0.63 (2014), 0.77 (2015), 0.66 (2016), 0.49 (2017), 0.45 (2018). This metric has the same pattern (more attack results in more runs) as the year-wise strength rules of Steve Smith between 2013 and 2016 (both inclusive).

- Bowling Analysis of Kagiso Rabada on Full-length Deliveries
  - *Year-wise Changes in Strength Rule:* He has shown a decrease in strength on full-length deliveries from 2015 to 2017 (both inclusive) and once again shown an increase in strength on full-length deliveries from 2017 to 2018 (both inclusive).
  - *Year-wise Changes in Weakness Rule:* He has shown an increase in weakness on full-length deliveries between the years 2017 and 2019 (both inclusive).

  To verify the year-wise results in terms of runs scored by batsmen and wickets taken on Kagiso Rabada's full-length deliveries, year-wise runs-per-ball and wickets-per-ball metrics (total runs given (or wickets taken) in full-length balls divided by total full-length balls played in a year) are obtained. For Kagiso Rabada, on full-length deliveries, year-wise runs-per-ball values are: 0.40 (2015), 0.59 (2016), 0.69 (2017), 0.69 (2018), 0.81 (2019). This metric has the same pattern (more attack results in more runs which implies bowler's decrease in strength) as the year-wise strength rules Kagiso Rabada between 2015 and 2017 (both inclusive). The year-wise wickets-per-ball values are: 0.01 (2015), 0.05 (2016), 0.03 (2017), 0.03 (2018), 0.00 (2019). This metric has the same pattern (fewer wickets implies bowler's weakness) as the year-wise weakness rules Kagiso Rabada between 2017 and 2019 (both inclusive).

5.1 Technical Confrontation Tensor

To measure the changes in player's strengths and weaknesses over the years, we have considered the time granularity as 'year.' To perform the year-wise analysis for a player, we first obtain the subset of short text commentaries for each year that the player has played. This is achieved by taking the

Algorithm 5 Construction of Technical Confrontation Tensor for Batsman (TCT_{BAT})

```
Require: Year-wise TCMs of batsman: $TCM_{BAT}Year_1, TCM_{BAT}Year_2, \cdots, TCM_{BAT}Year_n$. A
    three dimensional tensor of zeros ($TCT_{BAT_{19\times12\times n}}$).
while i ≤ n do
    $TCT_{BAT_{::i}} = TCM_{BAT}Year_i$
    i = i+1
end while
return Technical Confrontation Tensor for batsman ($TCT_{BAT}$)
```

Algorithm 6 Construction of Technical Confrontation Tensor for Bowler (TCT_{BOWL})

```
Require: Year-wise TCMs of Bowler: $TCM_{BOWL}Year_1, TCM_{BOWL}Year_2, \cdots, TCM_{BOWL}Year_n$.
    A three dimensional tensor of zeros ($TCT_{BOWL_{19\times12\times n}}$).
while i ≤ n do
    $TCT_{BOWL_{::i}} = TCM_{BOWL}Year_i$
    i = i+1
end while
return Technical Confrontation Tensor for Bowler ($TCT_{BOWL}$)
```

time parameter of the filter tuple as *year*. Next, year-wise technical confrontation matrices (TCMs) are constructed as described in Chapter 3.5.1. In Chapter 4, we captured the relationship between batting features and bowling features from the TCMs. This chapter aims to capture the relationship between batting features, bowling features, and time (year). To preserve the multiple interactions among these three categories of features, the analysis must be performed without collapsing them into a matrix. 'Time' has to be a dimension by itself. That is why we have modeled these features as a three-dimensional tensor of batting features, bowling features, and time.

For temporal analysis of batting, using the filter tuple ⟨*Player, All Opponent Players, Year, Batting*⟩, we obtain the year-wise TCMs ($TCM_{BAT}Year_1, TCM_{BAT}Year_2, \cdots, TCM_{BAT}Year_n$) for a batsman. Using these year-wise TCM_{BAT}, we construct a Technical Confrontation Tensor (TCT_{BAT}) of size (19 × 12 × number of years) in which rows correspond to batting features of the player, columns correspond to bowling features of the opponent players, and tubes correspond to the time frame (per year) in which the batsman has played. Identical batting features and bowling features are used as given in the previous chapter. Every element in this tensor corresponds to how and when the batsman confronted with the bowlers. For example, how many numbers of times a batsman has attacked short-length deliveries in the year 2019. Similarly, other entries in the TCT_{BAT} represent the count of co-occurrences of batting features and bowling features in a given year. Refer to Algorithm 5 for the steps of TCT_{BAT} construction.

For temporal analysis of bowling, using filter tuple ⟨*Player, All Opponent Players, Year, Bowling*⟩, we obtain the year-wise TCMs ($TCM_{BOWL}Year_1, TCM_{BOWL}Year_2, \cdots, TCM_{BOWL}Year_n$)

for a bowler. Using these year-wise TCM_{BOWL}, we construct a Technical Confrontation Tensor (TCT_{BOWL}) of size (19 × 12 × number of years) in which rows correspond to batting features of the opponent batsmen, columns correspond to bowling features of the bowler, and tubes correspond to the time frame (per year) in which the bowler has played. Identical batting features and bowling features are used as given in the previous chapter. Every element in this tensor corresponds to how and when the bowler confronted with the batsmen. For example, how many numbers of times the opponent batsmen have attacked short-length deliveries of the bowler in the year 2019. Similarly, other entries in the TCT_{BOWL} represent the count of co-occurrences of batting features, bowling features in a given year. Refer to Algorithm 6 for the steps of TCT_{BOWL} construction.

5.2 Learning Temporal Changes in Strength and Weakness Rules

The objective is to obtain relationships between the discrete random variables, namely batting features (row variables), bowling features (column variables), and time (tube variables) present in the TCT. As stated above, to preserve the multiple interactions among these three categories of features, the analysis must be performed without collapsing them into a matrix. Hence, the generalized singular value decomposition [121] is not appropriate in this case. On the other hand, three-way generalizations of singular value decomposition - Tucker3 decomposition model [141] is perfect for this task.

Finally, the tensor factorization is performed using a method known as Three-Way Correspondence Analysis (TWCA) [27, 23], which uses Tucker3 decomposition on the transformed TCT. TWCA tests the independence of events, namely row variables, column variables, and tube variables. If these events are not independent, then equality does not hold; this points to the relationship between the three variables.

Let $\mathcal{N}$ be a three dimensional TCT with I rows (batting features), J columns (bowling features) and K tubes (time in years). An entry in the i^{th} row, j^{th} column and k^{th} tube, N_{ijk}, represents the frequency of those deliveries which contain all three features (i, j, k). Let $n = \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{j=1}^{K} N_{ijk}$ be the sum of the elements of $\mathcal{N}$. Let $\mathcal{P} = \frac{1}{n}\mathcal{N}$ be a tensor of joint relative frequencies with p_{ijk} as its (i, j, k)th element such that $\sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} p_{ijk} = 1$. An element p_{ijk} denotes joint probability that event i, event j, and event k occurring simultaneously. Let the event e_1 be a batsman attacking, e_2 be a bowler bowling off line ball, and e_3 be in the year 2015. When these three events are *independent* then the following equations should hold:

$$P(e_1 \cap e_2 \cap e_3) = P(e_1) \times P(e_2) \times P(e_3) \tag{5.1}$$

$$p_{ijk} = p_{i..} \times p_{.j.} \times p_{..k} \tag{5.2}$$

In equation 5.2, $p_{i..} = \sum_{j=1}^{J} \sum_{k=1}^{K} p_{ijk}$ denotes the probability of row event i occurring. In a similar fashion $p_{.j.} = \sum_{i=1}^{I} \sum_{k=1}^{K} p_{ijk}$ and $p_{..k} = \sum_{i=1}^{I} \sum_{j=1}^{J} p_{ijk}$ are defined. When the total

Algorithm 7 TWCA Algorithm

```
Require: A three dimensional TCT $\mathcal{N}_{I\times J\times K}$
Tensor sum: $n = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{j=1}^{K}\mathcal{N}_{ijk}$
Tensor of relative frequencies: $\mathcal{P} = \frac{1}{n}\mathcal{N}$
Univariate marginal relative frequencies: $p_{i..} = \sum_{j=1}^{J}\sum_{k=1}^{K} p_{ijk}$, $p_{.j.} = \sum_{i=1}^{I}\sum_{k=1}^{K} p_{ijk}$, and $p_{..k} = \sum_{i=1}^{I}\sum_{j=1}^{J} p_{ijk}$
Deviations from the three way independence ($\mathcal{A}$): $\alpha_{ijk} = \frac{p_{ijk}-p_{i..}p_{.j.}p_{..k}}{p_{i..}p_{.j.}p_{..k}}$
Tucker3 decomposition: $Tucker3\left(\mathcal{A}\right) = A_{I\times P}G_{P\times Q\times R}\left(B_{J\times Q}^{T}\otimes C_{K\times R}^{T}\right)$
Principal coordinates of rows: $F = AG_{(P\times QR)}$
Principal coordinates of column-tubes: $H = \left(B\otimes C\right)G_{(QR\times P)}$
return F and H
```

independence gets deviated, the model is re-written as:

$$p_{ijk} = \boldsymbol{\alpha}_{ijk}\times p_{i..}\times p_{.j.}\times p_{..k} \tag{5.3}$$

In equation 5.3, $\boldsymbol{\alpha}_{ijk}$ denotes the amount of deviation. If $\boldsymbol{\alpha}_{ijk} = 1$ then row event i, column event j, and tube event k are independent. When row features (batting attributes) have a certain relation with respect to column features (bowling attributes) and tube features (time) $\boldsymbol{\alpha}_{ijk}$ takes value less than 1. For every row event, for every column event, and for every tube event, ijk^{th} entry of the $\mathcal{A}$ (residual tensor) is given by:

$$\boldsymbol{\alpha}_{ijk} = \frac{p_{ijk}}{p_{i..}\times p_{.j.}\times p_{..k}} \tag{5.4}$$

Equation 5.4 is well known as Pearson's ratio. The three-way association is captured using Pearson's chi-squared statistic for the tensor, which is the deviations from the three-way independence model, i.e.,

$$\boldsymbol{\alpha}_{ijk} = \frac{p_{ijk} - p_{i..}p_{.j.}p_{..k}}{p_{i..}p_{.j.}p_{..k}} \tag{5.5}$$

To obtain a low dimensional subspace that contains the batting features, bowling features, and time, $\mathcal{A}$ is decomposed using Tucker3 [141] decomposition to obtain four factors, namely $\mathcal{G}$ (core tensor), A (retains batting features), B (retains bowling features) and C (retains time feature).

To obtain the association between the batting, bowling, and time features, two variables among the three are coded, i.e., column-tube categories are the coded bowling-time features. It results in the principal components of row/batting features ($F = A\mathcal{G}$) and principal components of column-tube/bowling-time features ($H = (B\otimes C)\mathcal{G}$). The complete algorithm is presented in Algorithm 7. F retains the batting features of the player, and H retains the bowling-time features. $F^{'}$ and $H^{'}$ are the first two principal components of F and H, respectively. The inner product of $F^{'}$ and $H^{'}$ enables us to reconstruct the original three-way TCT and allows for a numerical assessment of the three-way association.

5.2.1 Temporal Analysis of Batting through TWCA

Refer to Figure 5.1 for the steps involved in temporal analysis of batting through TWCA. For batting analysis of a player, the TCT_{BAT} (denoted as $\mathcal{N}$) tensor of the player is constructed using the year-wise TCMs (denoted as $TCM_{BAT}Year_X$). TWCA first obtains the residual tensor $\mathcal{A}$ from $\mathcal{N}$. Next, Tucker3 tensor decomposition is applied to $\mathcal{A}$ to obtain the *batting* principal components (F), *bowling-time* principal components (H). Then, the first two principal components of F and H (denoted as $F^{'}$ *and* $H^{'}$) are obtained. Finally, the inner product matrix ($\langle F^{'}, H^{'}\rangle$) of the first two principal components of F and H is obtained.

Temporal Changes in Strength Rule of a Batsman

To qualify for a strength rule of a batsman, the batting feature as given in Definition 4.2 must contain the *attacked* feature. In order to frame a complete rule as per Definition 4.1, a bowling feature must be identified. The process of obtaining the temporal changes in strength rule of a batsman against a particular bowling feature (let's consider *good length*) is to take the inner product of $F^{'}_{attacked}$ with every *good-year* vector of $H^{'}$; that is, compute $\langle F^{'}_{attacked}, H^{'}_{good2012}\rangle$, $\langle F^{'}_{attacked}, H^{'}_{good2013}\rangle$, $\langle F^{'}_{attacked}, H^{'}_{good2014}\rangle$, $\cdots$, $\langle F^{'}_{attacked}, H^{'}_{good2018}\rangle$. These inner product values can also be obtained from the inner product matrix $\langle F^{'}, H^{'}\rangle$. A higher value of the inner product indicates a high strength of association, while a lower value of the inner product indicates a relatively low strength of association. These inner product values are plotted using line plot to visualize year-wise changes of strength rule against a particular bowling feature. This process is repeated for all the bowling features.

Temporal Changes in Weakness Rule of a Batsman

To qualify for a weakness rule of a batsman, the batting feature as given in Definition 4.3 must contain the *beaten* feature. In order to frame a complete rule as per Definition 4.1, a bowling feature must be identified. The process of obtaining the temporal changes in weakness rule of a batsman against a particular bowling feature (let us consider *good length*) is to take the inner product of $F^{'}_{beaten}$ with every *good-year* vector of $H^{'}$. These inner product values are plotted using line plot to visualize year-wise changes of weakness rule against a particular bowling feature. This process is repeated for all the bowling features.

Temporal Changes in Other Rules for a Batsman

In addition to the temporal changes strength rule and weakness rule, it is important to learn the temporal changes in other rules for batsman corresponding to the *response*, *outcome*, *footwork*, and *shot area* of a particular delivery. To obtain these rules, except *attacked* and *beaten*, all the other batting features are considered. In order to frame a complete rule for each of these batting features, as per Definition 4.1, a bowling feature must be identified. The process of obtaining the temporal changes in other rules of a batsman against a particular bowling feature (let us consider

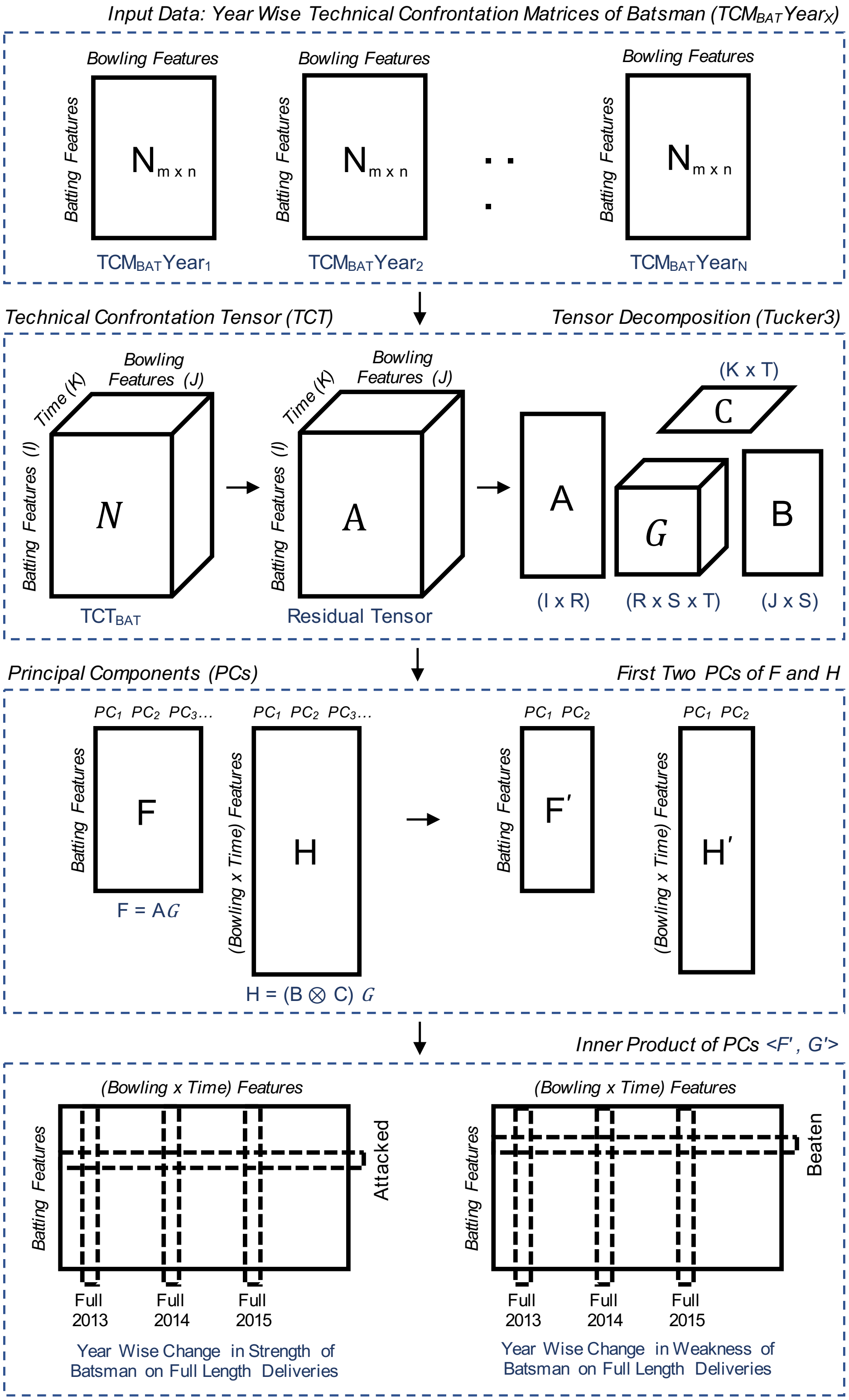

Input Data: Year Wise Technical Confrontation Matrices of Batsman ($TCM_{BAT}Year_X$)
Bowling Features
Batting Features
$N_{m \times n}$
$TCM_{BAT}Year_1$
$TCM_{BAT}Year_2$
$TCM_{BAT}Year_N$
Technical Confrontation Tensor (TCT)
Tensor Decomposition (Tucker3)
Time (K)
Bowling Features (J)
Batting Features (I)
N
TCT_{BAT}
A
Residual Tensor
(K x T)
C
G
B
(I x R)
(R x S x T)
(J x S)
Principal Components (PCs)
First Two PCs of F and H
PC_1 PC_2 PC_3...
F
F = AG
(Bowling x Time) Features
H
H = (B ⊗ C) G
PC_1 PC_2
F′
H′
Inner Product of PCs <F′ , G′>
Attacked
Beaten
Full 2013
Full 2014
Full 2015
Year Wise Change in Strength of Batsman on Full Length Deliveries
Year Wise Change in Weakness of Batsman on Full Length Deliveries

Figure 5.1: Temporal Analysis of Batting through TWCA.

good length) is to take the inner product of $F^{'}_{batting\ feature}$ with every *good-year* vector of $H^{'}$. These inner product values are plotted using line plot to visualize year-wise changes of other rules against a particular bowling feature. This process is repeated for all the bowling features.

5.2.2 Temporal Analysis of Bowling through TWCA

Refer to Figure 5.2 for the steps involved in the temporal analysis of bowling through TWCA. For bowling analysis of a player, the TCT_{BOWL} (denoted as $\mathcal{N}$) tensor of the player is constructed using the year-wise TCMs (denoted as $TCM_{BOWL}Year_X$). TWCA first obtains the residual tensor $\mathcal{A}$ from $\mathcal{N}$. Next, Tucker3 tensor decomposition is applied to $\mathcal{A}$ to obtain the *batting* principal components (F), *bowling-time* principal components (H). Then, the first two principal components of F and H (denoted as $F^{'}$ *and* $H^{'}$) are obtained. Finally, the inner product matrix ($\langle F^{'}, H^{'} \rangle$) of the first two principal components of F and H is obtained.

Temporal Changes in Strength Rule of a Bowler

To qualify for a strength rule of a bowler, the batting feature as given in Definition 4.4 must contain the *beaten* feature. In order to frame a complete rule as per Definition 4.1, a bowling feature must be identified. The process of obtaining the temporal changes in strength rule of a bowler on a particular bowling feature (let us consider *good length*) is to take the inner product of $F^{'}_{beaten}$ with every *good-year* vector of $H^{'}$. These inner product values are plotted using line plot to visualize year-wise changes of strength rule on a particular bowling feature. This process is repeated for all the bowling features.

Temporal Changes in Weakness Rule of a Bowler

To qualify for a weakness rule of a bowler, the batting feature as given in Definition 4.5 must contain the *attacked* feature. In order to frame a complete rule as per Definition 4.1, a bowling feature must be identified. The process of obtaining the temporal changes in weakness rule of a bowler on a particular bowling feature (let us consider *good length*) is to take the inner product of $F^{'}_{attacked}$ with every *good-year* vector of $H^{'}$. These inner product values are plotted using line plot to visualize year-wise changes of weakness rule on a particular bowling feature. This process is repeated for all the bowling features.

Temporal Changes in Other Rules for a Bowler

In addition to the temporal changes strength rule and weakness rule, it is important to learn the temporal changes in other rules for bowlers corresponding to the *response*, *outcome*, *footwork*, and *shot area* of a particular delivery. To obtain these rules, except *attacked* and *beaten*, all the other batting features are considered. In order to frame a complete rule for each of these batting features, as per Definition 4.1, a bowling feature must be identified. The process of obtaining the temporal changes in other rules of a bowler on a particular bowling feature (let us consider *good length*) is

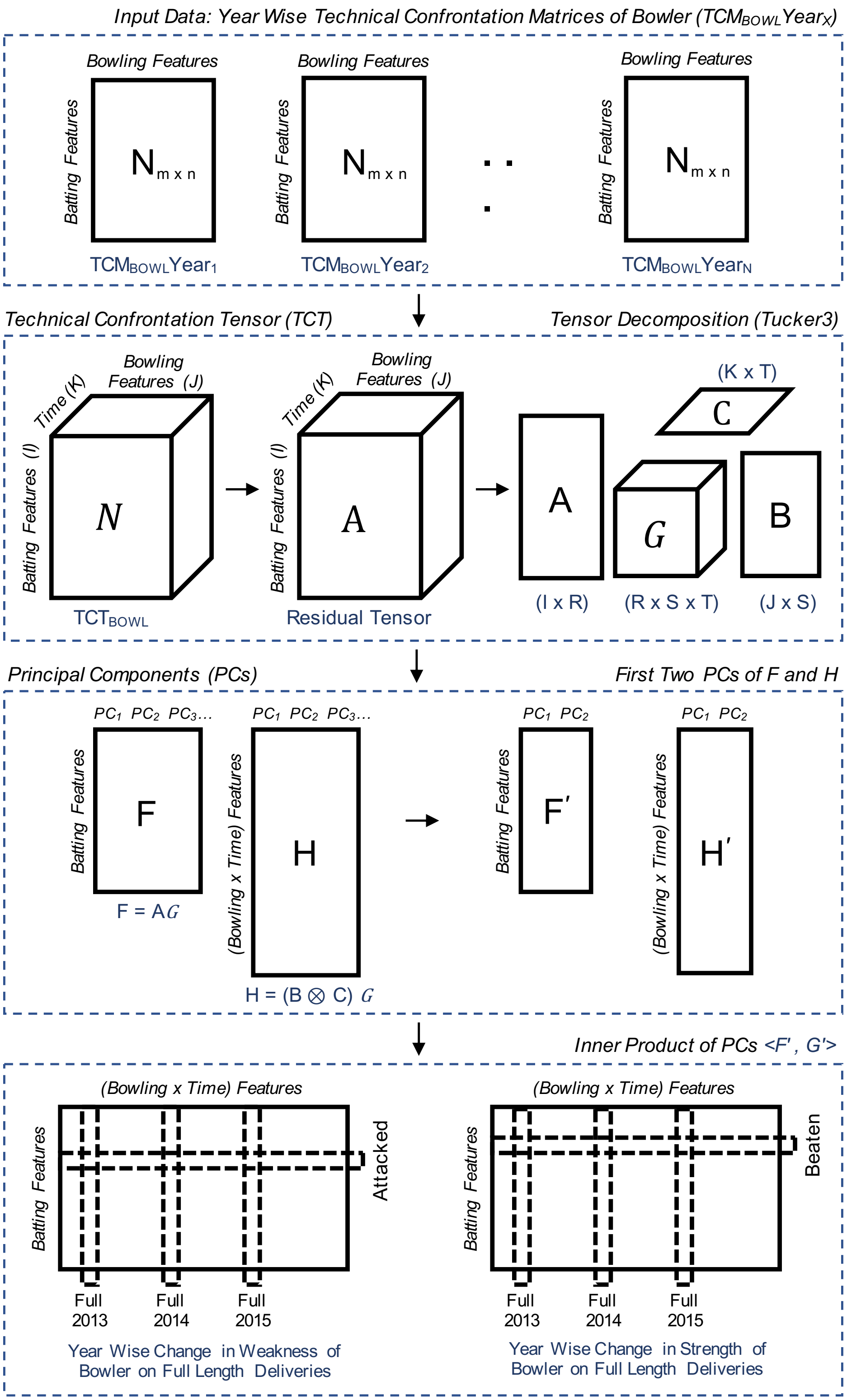

Input Data: Year Wise Technical Confrontation Matrices of Bowler (TCMBOWLYearX)
Bowling Features
Batting Features
Nm x n
TCMBOWLYear1
TCMBOWLYear2
TCMBOWLYearN
Technical Confrontation Tensor (TCT)
Tensor Decomposition (Tucker3)
Time (K)
Bowling Features (J)
Batting Features (I)
N
TCTBOWL
A
Residual Tensor
(K x T)
C
A
G
B
(I x R)
(R x S x T)
(J x S)
Principal Components (PCs)
First Two PCs of F and H
PC1 PC2 PC3…
F
F = AG
(Bowling x Time) Features
H
H = (B ⊗ C) G
PC1 PC2
F′
H′
Inner Product of PCs <F', G'>
(Bowling x Time) Features
Attacked
Beaten
Full 2013
Full 2014
Full 2015
Year Wise Change in Weakness of Bowler on Full Length Deliveries
Year Wise Change in Strength of Bowler on Full Length Deliveries

Figure 5.2: Temporal Analysis of Bowling through TWCA.

to take the inner product of $F'_{batting\,feature}$ with every *good-year* vector of H'. These inner product values are plotted using line plot to visualize year-wise changes of other rules on a particular bowling feature. This process is repeated for all the bowling features.

5.3 Experiments

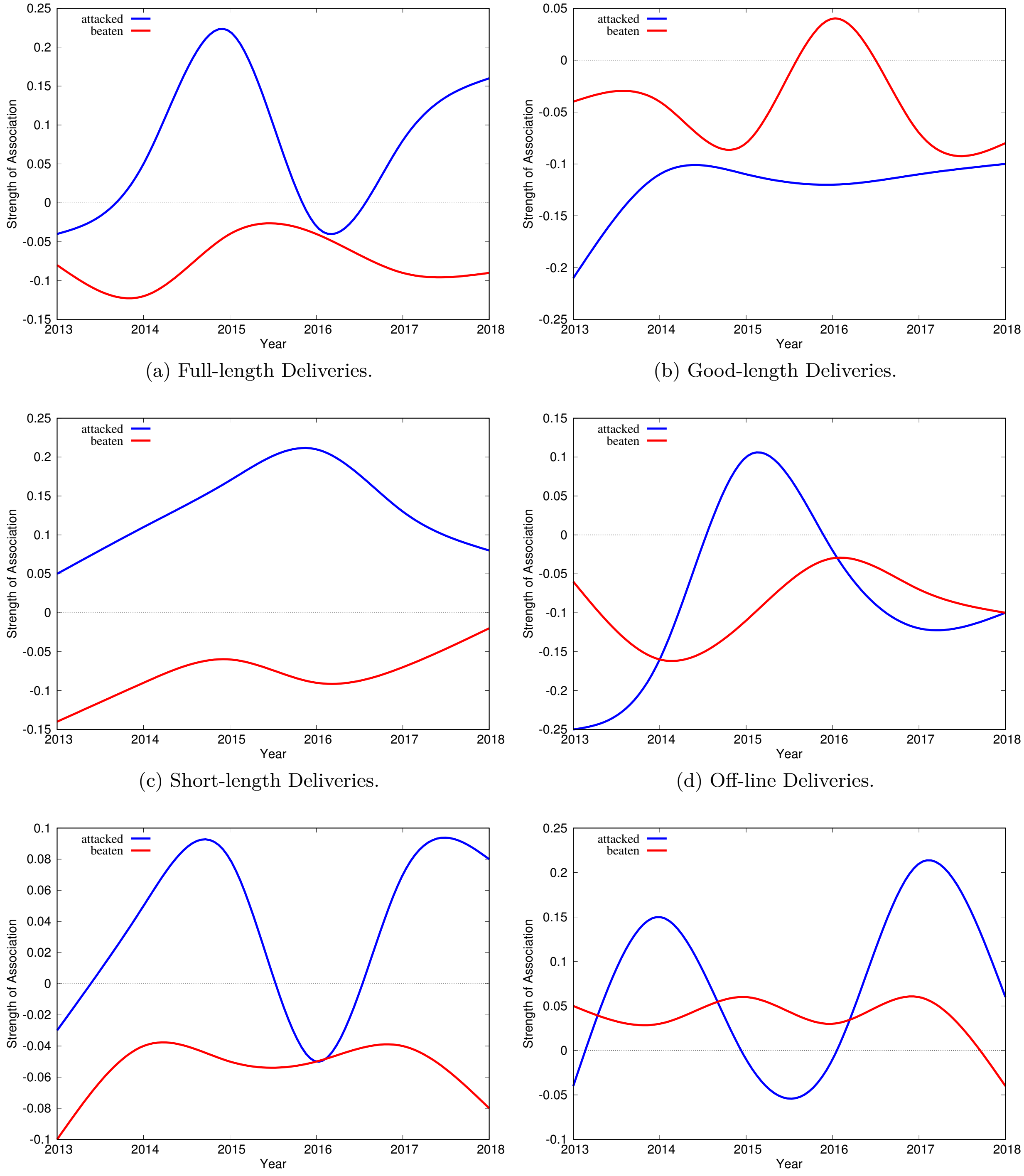
This section presents case studies that illustrate the temporal (year-wise) analysis of player's strengths and weaknesses using the proposed approach. First, we present the temporal analysis of batsman Steve Smith. Next, we present the temporal analysis of bowler Kagiso Rabada. For both batsman and bowler, twelve distinct temporal analyses are presented, namely, *strength and weakness in full-length deliveries over the years*, *strength and weakness in good-length deliveries over the years*, etc. In addition, the year-wise changes in *response*, *outcome*, *shot area*, and *footwork* are also presented for both batsman and bowler.

5.3.1 Year-wise Analysis of Batting - Steve Smith

To perform the year-wise batting analysis of Steve Smith, we first obtain the year-wise TCMs $(TCM_{BAT}Year_{2013}, TCM_{BAT}Year_{2014}, \cdots, TCM_{BAT}Year_{2018})$ for each year of his play (2013 to 2018) using the filter tuple $\langle$*Steve Smith*, *All Opponent Players*, *Year*, *Batting*$\rangle$. Using these year-wise TCM_{BAT}, we construct a Technical Confrontation Tensor (TCT_{BAT}) of size (19 × 12 × 6) in which rows correspond to batting features of Steve Smith, columns correspond to bowling features of the opponent players, and tubes correspond to the time frame (per year) in which Steve Smith has batted. Employing the proposed approach described in Section 5.2.1, the year-wise changes in strength rule and weakness rule of batsman Steve Smith against various deliveries are obtained and presented as line plots (Refer to Figure 5.3).

In these line plots (each plot is for a particular bowling feature), the X-axis represents the years in which Steve Smith has batted, and the Y-axis represents the inner product values or strength of association between the batting features (attacked or beaten) and the bowling-year feature. A higher value of the inner product indicates a high strength of association, while a lower value of the inner product indicates a relatively low strength of association. The blue-colored line represents the change in the strength rule (attacked) over the years, and the red-colored line represents the change in the weakness rule (beaten) over the years.

The year-wise changes in Strength Rule (SR) and Weakness Rule (WR) of batsman Steve Smith on *full length* deliveries are presented in Figure 5.3a. From this figure, we observe that Smith has shown an increase in attacking full-length deliveries between the years 2014 and 2015 (both inclusive). However, in 2016 he struggled on full length deliveries. He has once again shown an increase in the trend of attacking *full-length* deliveries between the years 2017 and 2018 (both inclusive). The red line's negative correlation below the X-axis suggests that Smith has never exhibited weaknesses on *full length* deliveries. These rules also can be written as - *SR (attacked): 2014-2015 (trend - increase), 2017-2018 (trend - increase) and WR (beaten): Never beaten.*



(a) Full-length Deliveries.

(b) Good-length Deliveries.

(c) Short-length Deliveries.

(d) Off-line Deliveries.

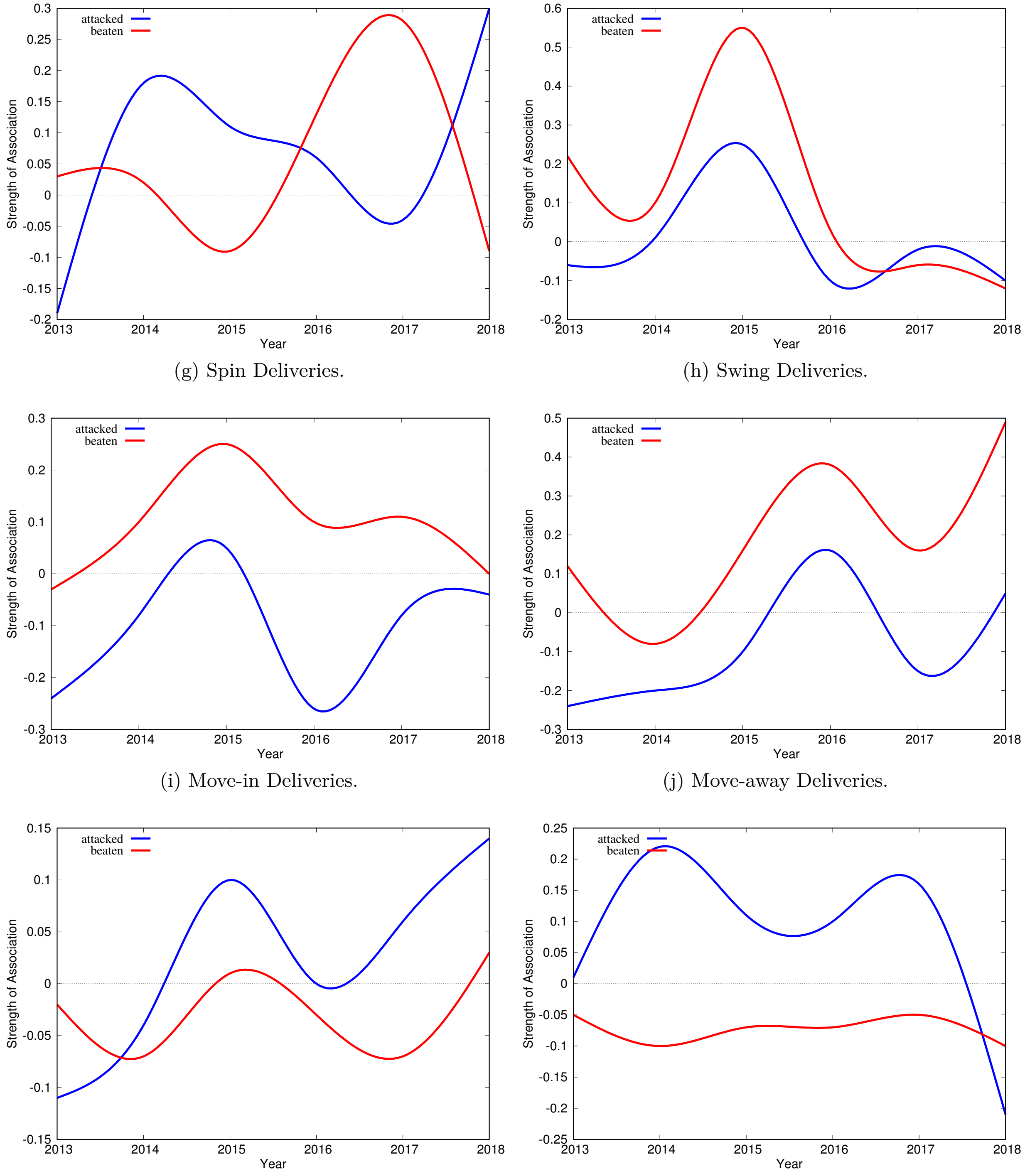
(e) Middle-line Deliveries.

(f) Leg-line Deliveries.

Figure 5.3: Year-wise Change in Strength and Weakness Rules of Steve Smith on Various Deliveries.

The year-wise changes in the SR and WR of batsman Steve Smith on all the bowling features are listed below:

1. *Full Length Deliveries (Figure 5.3a)*



(g) Spin Deliveries.

(h) Swing Deliveries.

(i) Move-in Deliveries.

(j) Move-away Deliveries.

(k) Fast Deliveries.

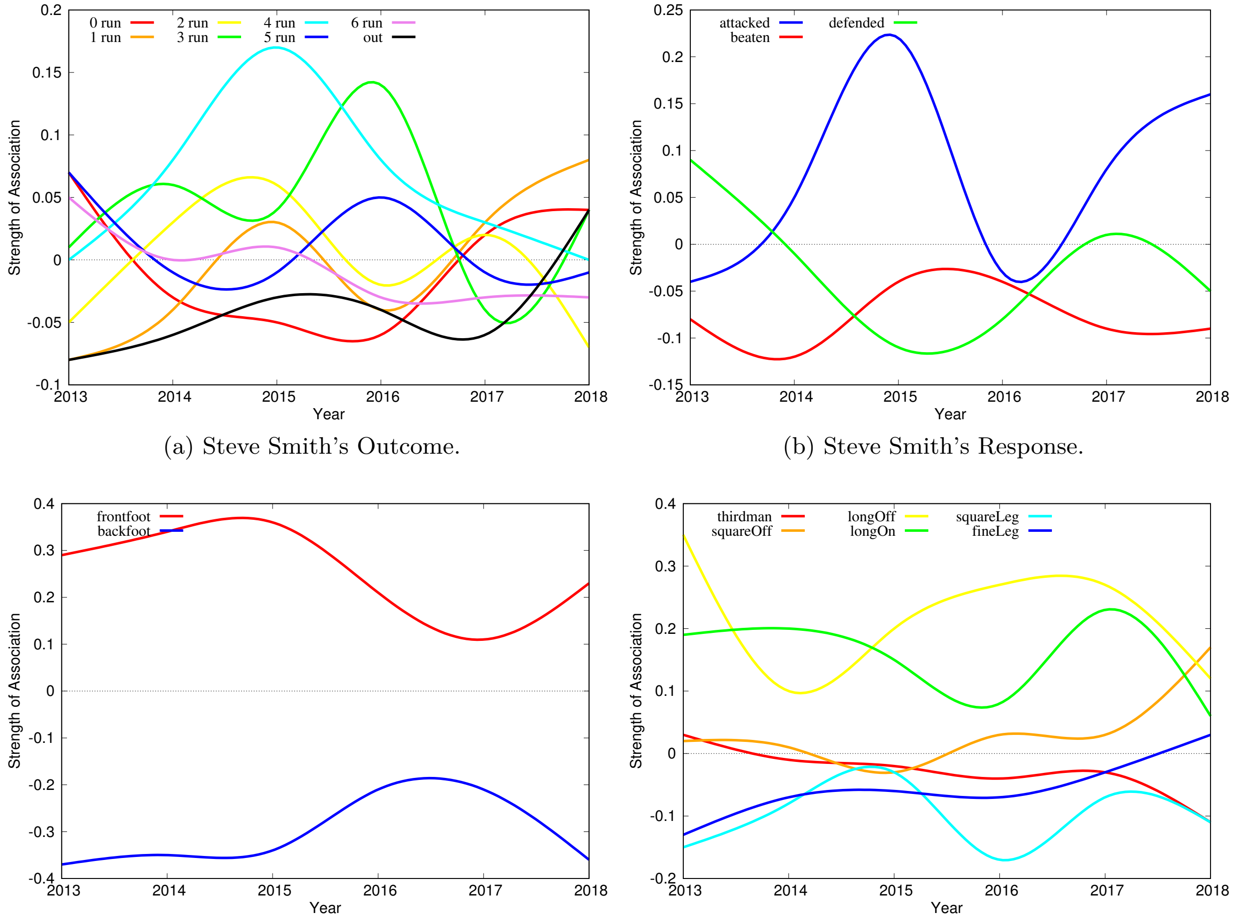
(l) Slow Deliveries.

Figure 5.3: Year-wise Change in Strength and Weakness Rules of Steve Smith on Various Deliveries.

- SR (attacked): 2014-2015 (trend - increase), 2017-2018 (trend - increase)

- WR (beaten): Never beaten

2. *Good Length Deliveries (Figure 5.3b)*
    - SR (attacked): Never attacked
    - WR (beaten): 2016
3. *Short Length Deliveries (Figure 5.3c)*
    - SR (attacked): 2013-2016 (trend - increase), 2016-2018 (trend - decrease)
    - WR (beaten): Never beaten
4. *Off Line Deliveries (Figure 5.3d)*
    - SR (attacked): 2015
    - WR (beaten): Never beaten
5. *Middle Line Deliveries (Figure 5.3e)*
    - SR (attacked): 2014-2015 (trend - increase), 2017
    - WR (beaten): Never beaten
6. *Leg Line Deliveries (Figure 5.3f)*
    - SR (attacked): 2014, 2017-2018 (trend - decrease)
    - WR (beaten): 2013-2017
7. *Spin Deliveries (Figure 5.3g)*
    - SR (attacked): 2014, 2015-2016 (trend - decrease), 2018
    - WR (beaten): 2013-2014, 2016-2017 (trend - increase)
8. *Swing Deliveries (Figure 5.3h)*
    - SR (attacked): 2015
    - WR (beaten): 2013-2014 (trend - decrease), 2014-2015 (trend - increase), 2015-2016 (trend - decrease)
9. *Move-in Deliveries (Figure 5.3i)*
    - SR (attacked): 2015
    - WR (beaten): 2014-2015 (trend - increase), 2015-2018 (trend - decrease)
10. *Move-away Deliveries (Figure 5.3j)*
    - SR (attacked): 2016, 2018
    - WR (beaten): 2013, 2015-2016 (trend - increase), 2016-2017 (trend - decrease), 2017-2018 (trend - increase)



(a) Steve Smith's Outcome.

(b) Steve Smith's Response.

(c) Steve Smith's Footwork.

(d) Steve Smith's Shotarea.

Figure 5.4: Year-wise Analysis of Steve Smith's Batting on Full Length Deliveries.

11. *Fast Deliveries (Figure 5.3k)*
    - SR (attacked): 2015, 2017-2018 (trend - increase)
    - WR (beaten): 2015, 2018

12. *Slow Deliveries (Figure 5.3l)*
    - SR (attacked): 2013-2014 (trend - increase), 2014-2016 (trend - decrease), 2016-2017 (trend - increase)
    - WR (beaten): Never beaten

In addition to the temporal changes in batsmen's strength and weakness rules (attacked/beaten), it is essential to note the temporal change in the *outcome* of a particular delivery, where the ball is being hit by batsmen (*shot area*), and also their *footwork*. Figure 5.4 presents the line plots for the response, outcome, shot area, and footwork analysis of Steve Smith on *full-length* deliveries over the years. The rules observed from these line plots are presented below:

1. *Outcome (Figure 5.4a)*
    - 0 run: 2013, 2017-2018 (trend - increase)
    - 1 run: 2015, 2017-2018 (trend - increase)
    - 2 run: 2014-2015 (trend - increase), 2017
    - 3 run: 2013-2014 (trend - increase), 2014-2015 (trend - decrease), 2015-2016 (trend - increase), 2017
    - 4 run: 2013-2015 (trend - increase), 2015-2018 (trend - decrease)
    - 5 run: 2013, 2016
    - 6 run: 2013, 2015
    - Out: 2018
2. *Response (Figure 5.4b)*
    - Attacked: 2014-2015 (trend - increase), 2017-2018 (trend - increase)
    - Beaten: Never beaten
    - Defended: 2013 and 2017
3. *Footwork (Figure 5.4c)*
    - Frontfoot: 2013-2015 (trend - increase), 2015-2017 (trend - decrease), 2017-2018 (trend - increase)
    - Backfoot: Never played on backfoot
4. *Shotarea (Figure 5.4d)*
    - Thirdman: 2013
    - Square off: 2013, 2014, 2016-2018 (trend - increase)
    - Long off: 2013-2014 (trend - decrease), 2014-2017 (trend - increase), 2017-2018 (trend - decrease)
    - Long on: 2013, 2014-2016 (trend - decrease), 2016-2017 (trend - increase) , 2017-2018 (trend - decrease)
    - Square leg: Never played
    - Fine leg: 2018

Given the additional batting features that change over time, the proposed temporal analysis is capable of learning evolution of such features. One use case would be to learn the evolution of batsmen's strengths and weaknesses in the presence of external features such as pitch conditions and weather conditions.

5.3.2 Year-wise Analysis of Bowling - Kagiso Rabada

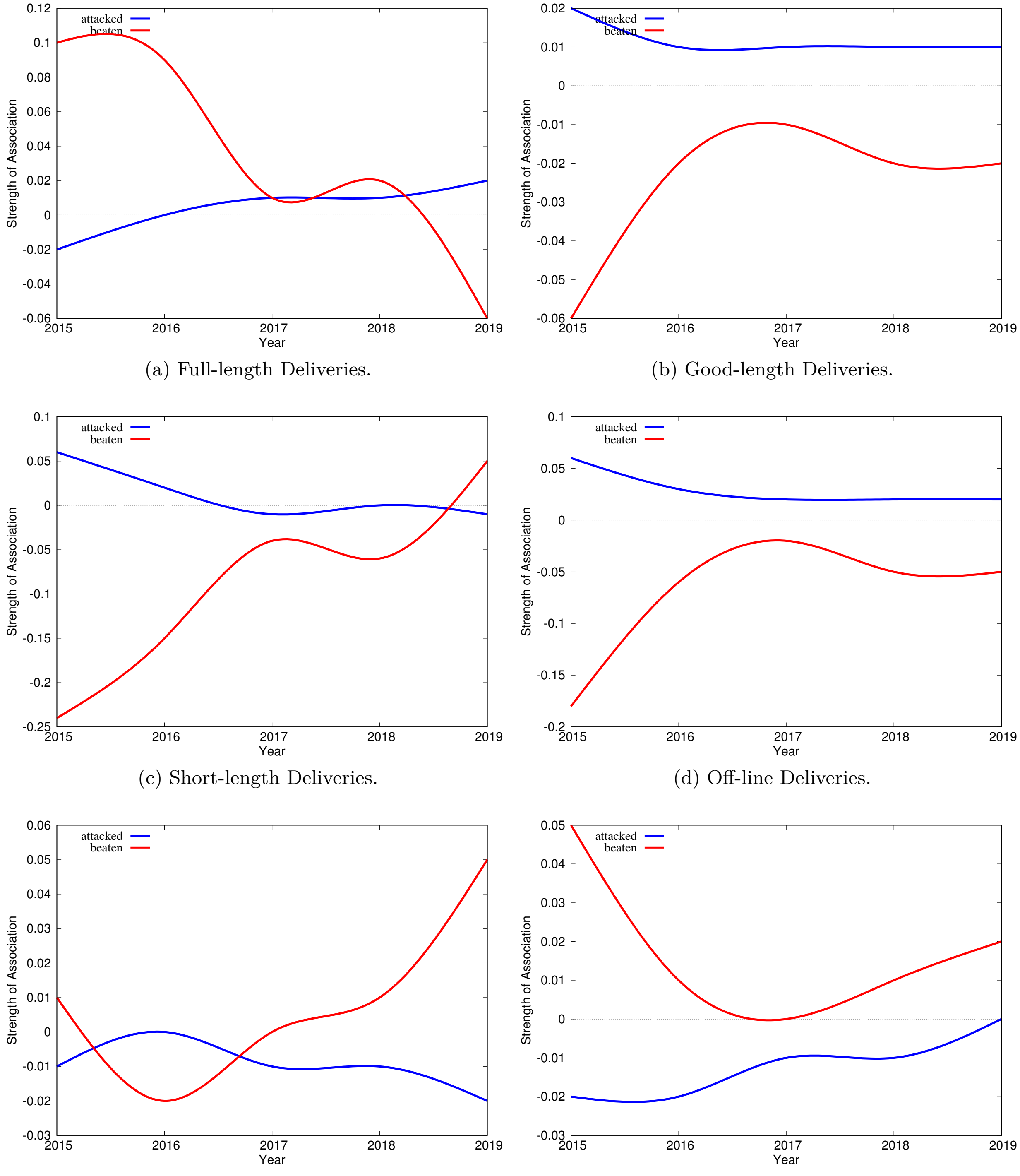
To perform the year-wise bowling analysis of Kagiso Rabada, we first obtain the year-wise TCMs ($TCM_{BOWL}Year_{2015}, TCM_{BOWL}Year_{2016}, \cdots, TCM_{BOWL}Year_{2019}$) for each year of his play (2015 to 2019) using the filter tuple ⟨*Kagiso Rabada*, *All Opponent Players*, *Year*, *Bowling*⟩. Using these year-wise TCM_{BOWL}, we construct a Technical Confrontation Tensor (TCT_{BOWL}) of size ($19 \times 12 \times 5$) in which rows correspond to batting features of opponent batsmen, columns correspond to bowling features of Kagiso Rabada, and tubes correspond to the time frame (per year) in which bowler Kagiso Rabada has bowled. Employing the proposed approach described in Section 5.2.2, the year-wise changes in strength rule and weakness rule of bowler Kagiso Rabada on his various deliveries are obtained and presented as line plots (Refer to Figure 5.5).

In these line plots (each plot is for a particular bowling feature), the X-axis represents the years in which Kagiso Rabada has bowled, and Y-axis represents the inner product values or strength of association between the batting features (attacked or beaten) and the bowling-year feature. A higher value of the inner product indicates a high strength of association, while a lower value of the inner product indicates a relatively low strength of association. The blue-colored line represents the change in the weakness rule (*attacked* by batsmen) over the years, and the red-colored line represents the change in the strength rule (batsmen got *beaten*) over the years.

The year-wise changes in Strength Rule (SR) and Weakness Rule (WR) of bowler Kagiso Rabada on his *full length* deliveries are presented in Figure 5.5a. From this figure, we observe that batsmen attacked his full-length deliveries from 2017 to 2019, i.e., Rabada has shown an increase in weakness on full-length deliveries between the years 2017 and 2019 (both inclusive). Batsman got beaten in his full-length deliveries from 2015 to 2018, i.e., Rabada has shown a decrease in strength on full-length deliveries from 2015 to 2017 (both inclusive) and once again shown an increase in strength on full-length deliveries from 2017 to 2018 (both inclusive). These rules also can be written as - *WR (attacked): 2017-2019 (trend - increase), SR (beaten): 2015-2017 (trend - decrease), 2017-2018 (trend - increase).*

The year-wise changes in the SR and WR of bowler Kagiso Rabada on all the bowling features are listed below:

1. *Full Length Deliveries (Figure 5.5a)*
   - WR (attacked): 2017-2019 (trend - increase)
   - SR (beaten): 2015-2017 (trend - decrease), 2017-2018 (trend - increase)
2. *Good Length Deliveries (Figure 5.5b)*
   - WR (attacked): 2015-2016 (trend - decrease), 2016-2019
   - SR (beaten): Never beaten
3. *Short Length Deliveries (Figure 5.5c)*
   - WR (attacked): 2015-2016 (trend - decrease)



(a) Full-length Deliveries.

(b) Good-length Deliveries.

(c) Short-length Deliveries.

(d) Off-line Deliveries.

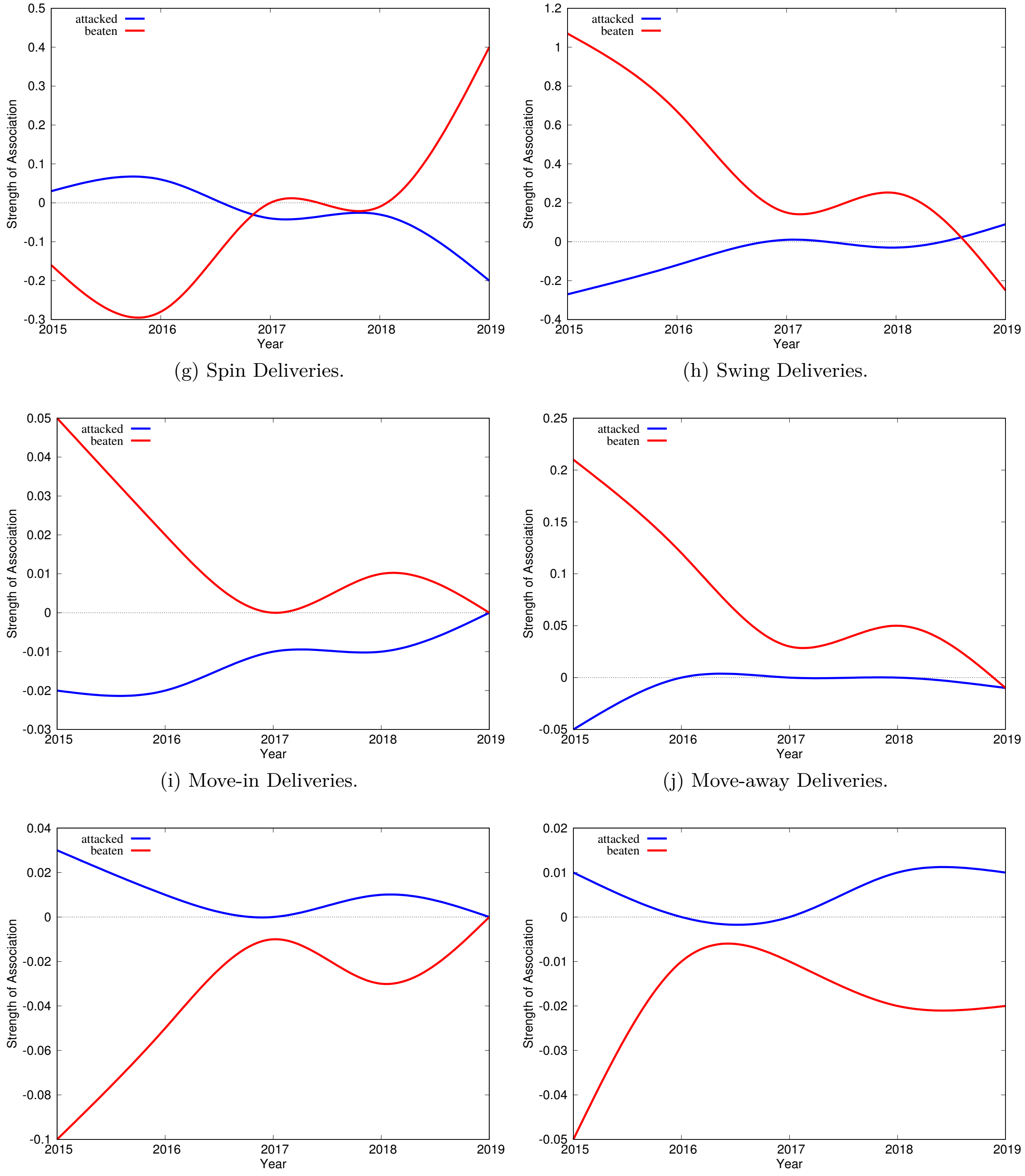
(e) Middle-line Deliveries.

(f) Leg-line Deliveries.

Figure 5.5: Year-wise Change in Strength and Weakness Rules of Kagiso Rabada on Various Deliveries.

- SR (beaten): 2019

4. *Off Line Deliveries (Figure 5.5d)*



(g) Spin Deliveries.

(h) Swing Deliveries.

(i) Move-in Deliveries.

(j) Move-away Deliveries.

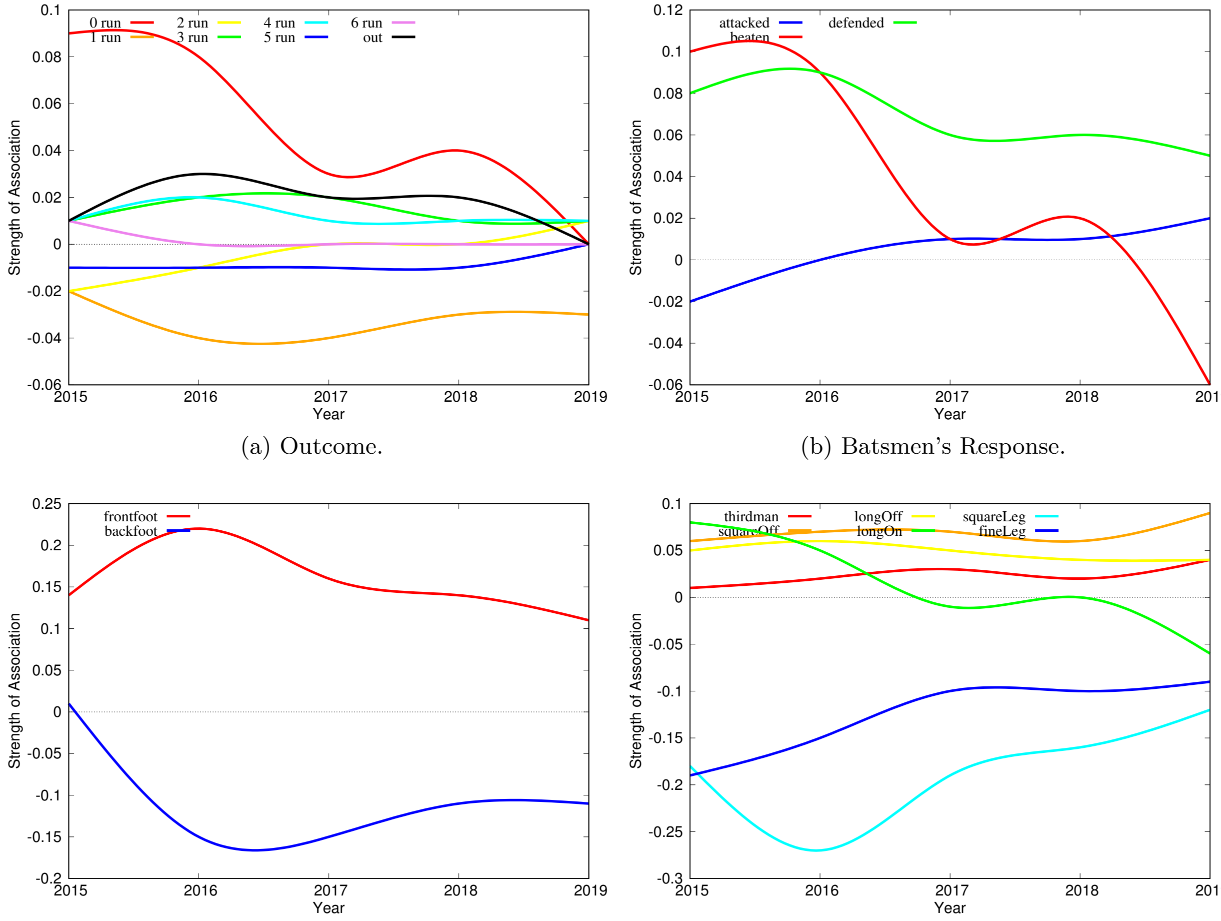
(k) Fast Deliveries.

(l) Slow Deliveries.

Figure 5.5: Year-wise Change in Strength and Weakness Rules of Kagiso Rabada on Various Deliveries.

- WR (attacked): 2015-2017 (trend - decrease), 2017-2019
- SR (beaten): Never beaten

5. *Middle Line Deliveries (Figure 5.5e)*
    - WR (attacked): Never attacked
    - SR (beaten): 2015, 2018-2019 (trend - increase)
6. *Leg Line Deliveries (Figure 5.5f)*
    - WR (attacked): Never attacked;
    - SR (beaten): 2015-2016 (trend - decrease), 2018-2019 (trend - increase)
7. *Spin Deliveries (Figure 5.5g)*
    - WR (attacked): 2015-2016 (trend - increase)
    - SR (beaten): 2019
8. *Swing Deliveries (Figure 5.5h)*
    - WR (attacked): 2019
    - SR (beaten): 2015-2017 (trend - decrease), 2017-2018 (trend - increase)
9. *Move-in Deliveries (Figure 5.5i)*
    - WR (attacked): Never attacked
    - SR (beaten): 2015-2017 (trend - decrease), 2017-2018 (trend - increase), 2018-2019 (trend - decrease)
10. *Move-away Deliveries (Figure 5.5j)*
    - WR (attacked): Never attacked
    - SR (beaten): 2015-2017 (trend - decrease), 2017-2018 (trend - increase), 2018-2019 (trend - decrease)
11. *Fast Deliveries (Figure 5.5k)*
    - WR (attacked): 2015-2017 (trend - decrease), 2017-2018 (trend - increase), 2018-2019 (trend - decrease)
    - SR (beaten): Never beaten
12. *Slow Deliveries (Figure 5.5l)*
    - WR (attacked): 2015-2016 (trend - decrease), 2018-2019 (trend - increase)
    - SR (beaten): Never beaten



(a) Outcome.

(b) Batsmen's Response.

(c) Batsmen's Footwork.

(d) Batsmen's Shotarea.

Figure 5.6: Year-wise Analysis of Kagiso Rabada's Bowling on his Full Length Deliveries.

In addition to the temporal changes in bowlers' strength and weakness rules (beaten/attacked), it is essential to note the temporal change in the *outcome* on a particular delivery, where the ball is being hit by batsmen (*shot area*), and also their *footwork*. Figure 5.6 presents the line plots for the response, outcome, shot area, and footwork analysis on Kagiso Rabada's *full-length* deliveries over the years. The rules observed from these line plots are presented below:

1. *Outcome (Figure 5.6a)*

   - 0 run: 2015-2017 (trend - decrease), 2017-2018 (trend - increase), 2018-2019 (trend - decrease)
   - 1 run: Never scored
   - 2 run: 2019
   - 3 run: 2015-2017 (trend - increase), 2017-2019 (trend - decrease)
   - 4 run: 2015-2016 (trend - increase), 2016-2017 (trend - decrease), 2017-2019 (trend - steady)

- 5 run: Never scored
- 6 run: 2015
- Out: 2015-2016 (trend - increase), 2016-2019 (trend - decrease)

2. *Response (Figure 5.6b)*

   - Attacked: 2017-2019 (trend - increase)
   - Beaten: 2015-2017 (trend - decrease), 2017-2018 (trend - increase)
   - Defended: 2015-2016 (trend - increase), 2016-2019 (trend - decrease)

3. *Footwork (Figure 5.6c)*

   - Frontfoot: 2015-2016 (trend - increase), 2016-2019 (trend - decrease)
   - Backfoot: 2015

4. *Shotarea (Figure 5.6d)*

   - Thirdman: 2015-2017 (trend - increase), 2017-2018 (trend - decrease), 2018-2019 (trend - increase)
   - Square off: 2015-2016 (trend - increase), 2016-2018 (trend - decrease), 2018-2019 (trend - increase)
   - Long off: 2015-2016 (trend - increase), 2016-2019 (trend - decrease)
   - Long on: 2015-2016 (trend - decrease)
   - Square leg and Fine leg: Never played

Given the additional bowling features that change over time, the proposed temporal analysis is capable of learning evolution of such features. One use case would be to learn the evolution of bowlers' strengths and weaknesses in the presence of external features such as pitch conditions and weather conditions.

The data, code, and result of the temporal analysis for more than 250 players can be accessed at https://www.dropbox.com/sh/1l3jd7icbx241oj/AADMGaE-tdGNjS6pInnXUDKea?dl=0.

5.4 Web Application

A web-based system is implemented to visualize the year-wise changes in player's strength and weakness rules. The left panel displays a drop-down menu from which the batsman and bowler can be selected. The line plots presenting the year-wise change in the selected batsman or bowler's strengths and weaknesses are displayed in the main panel. The system is live at https://cricketvisualization.shinyapps.io/StrengthWeaknessTemporal/.

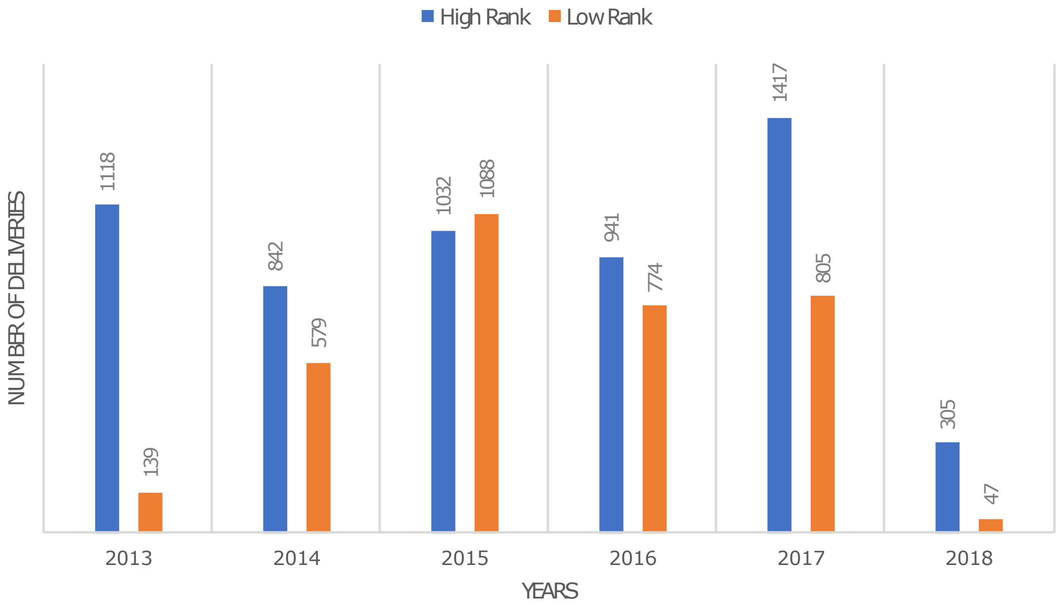


Figure 5.7: Number of Deliveries Faced by Batsman Steve Smith against High Ranked (Rank 1 - Rank 25) and Low Ranked (Rank 26 - Rank 100) Bowlers over the Years.

5.5 Discussion

In this chapter, we learned temporal changes in players' strength and weakness rules. However, the key question is: *Does the quality of opposition influence player's strengths and weaknesses over time?*

In order to understand the effect of quality of a player, we considered Steve Smith and the ICC world rank of bowler in a given year (rank of bowler is considered a proxy for quality of player). A bowler is considered a high ranked bowler when the rank in a particular year is between 1 and 25. Otherwise, the bowler is considered a low ranked bowler. Frequency of balls faced by Steve Smith against high-ranked bowlers and low-ranked bowlers in each year is computed and is presented in Figure 5.7 (complete drill-down analysis on the year of play for batsman Steve Smith can be accessed at https://www.dropbox.com/s/7cltwxgbg94psqe/SmithBowlersRank.pdf?dl=0). From this figure, we note that except in the year 2015, Steve Smith faced high-rank bowlers throughout. Correlating the year 2015 to the temporal strength and weakness rules of Steve Smith as presented in Figure 5.3, we observe that Steve Smith attacked:

- Full-length deliveries the most in the year 2015 (Figure 5.3a)
- Off-line deliveries the most in the year 2015 (Figure 5.3d)
- Middle line deliveries the most in the year 2015 (Figure 5.3e)

We also observe that Steve Smith got beaten the least on good-length deliveries in the year 2015 (Figure 5.3b). This reinforces the belief that inclusion of 'quality of opponent player' in player-specific analysis can provide new insights.

From the above discussion, we infer that the attribute quality has a clear influence on players' strengths and weaknesses over time. A detailed investigation of quality as an additional parameter in temporal analysis of strength and weakness rules is left for future work.

5.6 Summary

This chapter presented an approach to learn the temporal changes in player's strength and weakness rules. First, the approach used the year-wise TCMs to model the data as a three-dimensional TCT in which batting features, bowling features, and time (in years) are captured. Next, it employed the TWCA, which uses tucker3 decomposition on TCT to obtain the relationship between batting features, bowling features, and time. These relations are plotted in a line plot to visualize the year-wise changes in strength and weakness rules. Several case studies showed that the proposed approach could identify the temporal changes in players' strengths and weaknesses. The obtained rules are interpretable and are of value to coaches and team management. This work has been published in [143].

6

Mining Strength and Weakness Rules of Cricket Players in the Presence of External Factors Influencing the game

In Chapter 4, we looked at an approach to learn the strength and weakness rules of cricket players using short text commentary data. In Chapter 5, we obtained the temporal changes in the strength and weakness rules. In an outdoor team sport like cricket, external factors like playing conditions and match situations are outside the scope of the game and yet influence the gameplay. Different playing conditions (age-of-ball, pitch condition, and weather condition) and match situations (day, inning, and session) suit different players. For example, a bowler can extract high bounce from a grassy, hard, or intact pitch. A pitch with moisture can help swing the ball in-flight using the ball's heaviness due to moisture. A pitch that is dry and cracked helps a spinner achieve more turn because of the unevenness. Similarly, flat pitches do not assist bowlers and favor batsmen. Match situations also have indirect and sometimes direct effects. A pitch once prepared for a match is not repaired after each day. As one gets closer to the end of the game, typically on the last day or two, it becomes dry, and cracks start to develop. A spinner can be thought to have favorable conditions on the last day compared to the first day. A ball is replaced every 90 overs in the game. A new ball offers more pace and movement to fast bowlers. These external factors influence the techniques batsmen or bowlers exhibit.

This chapter presents an approach to learn the strength and weakness rules of cricket players in the presence of external factors. These factors are outside the scope of the game and yet influence a player's strengths and weaknesses. For a given player, the presented approach proceeds in two steps. In the first step, the approach introduces the computationally feasible definitions of strength and weakness in the presence of external factors. The second step employs a dimensionality reduction method, Canonical Correspondence Analysis (CCA) [28], on the TCM and ECM to construct semantic relations between batting features, bowling features, and external features. These relations are plotted in a triplot to visualize the strength and weakness rules in the presence

		Batting Feature	**Bowling Feature**	**External Feature**
Batting	**Strength Rule**	Attacked	Any	Any
	Weakness Rule	Beaten	Any	Any
Bowling	**Strength Rule**	Beaten	Any	Any
	Weakness Rule	Attacked	Any	Any

Table 6.1: Computational Definitions of Strength Rule and Weakness Rule in the Presence of External Factors Influencing the Game.

of external factors. Validation of the obtained results is impractical due to the absence of ground truth on player's strengths and weaknesses in the presence of external factors.

We use the proposed approach to mine strength and weakness rules in the presence of external factors, corresponding to 131 batsmen and 129 bowlers who have batted/bowled more than 2000 deliveries. The data, code, and result of the experiments can be accessed at https://www.dropbox.com/s/btrqii76tntl1ua/Chapter6%20External%20Factor%20Analysis.zip?dl=0. We highlight some of the obtained results for batsman Steve Smith and bowler Kagiso Rabada here. For batsman Steve Smith, we observe that - (i) On the third day of a Test match, bowlers tend to bowl moving-in deliveries to Steve Smith, and he plays them on the back foot, gets beaten, or loses his wicket in those deliveries. (ii) On the first day and second day of a Test match, bowlers tend to bowl moving-away deliveries on the outside off line to Steve Smith, and he plays them on the front foot, defends, or scores no runs. (iii) Steve Smith gets beaten or losses his wicket on green pitches. For bowler Kagiso Rabada, we observe that - (i) When the ball is more than 30 overs old, Kagiso Rabada tends to bowl moving-away deliveries to batsmen, and the batsmen defend or play them to the thirdman area. (ii) Kagiso Rabada tends to bowl moving-away deliveries to batsmen on a grassy pitch, and the batsmen defend them.

6.1 Computational Definition of Strength and Weakness in the Presence of External Factors

In this section, we provide a computational definition for the strength rule and weakness rule in the presence of external factors. We define the strength rule and weakness rule of batsman/bowler in the presence of external factors as follows:

Definition 6.1 *Rule. A rule must comprise of one batting feature, one bowling feature, and one external feature which are dependent on each other.*

Definition 6.2 *Strength Rule of Batsman. In Definition 6.1, when the batting feature corresponds to attacked and involves any of the bowling features and any of the external features.*

Definition 6.3 *Weakness Rule of Batsman. In Definition 6.1, when the batting feature corresponds to beaten and involves any of the bowling features and any of the external features.*

Whenever a batsman exhibits strength on a delivery, it is a weakness for the bowler, and the inverse is also true. Therefore, bowlers' strengths and weaknesses are defined in terms of the batting features of the batsmen they have bowled to. A bowler exhibits strength when the opponent batsman's batting feature is beaten. Similarly, a bowler exhibits weakness when the opponent batsman's batting feature is attacked.

Definition 6.4 *Strength Rule of Bowler. In Definition 6.1, when the opponent's batting feature corresponds to beaten and involves any of the bowling features and any of the external features.*

Definition 6.5 *Weakness Rule of Bowler. In Definition 6.1, when the opponent's batting feature corresponds to attacked and involves any of the bowling features and any of the external features.*

6.2 Learning Strength and Weakness Rules in the Presence of External Factors

In this section, we propose a computational method that identifies the strength and weakness rules of the individual player in the presence of external features. In Section 4.2, the rule learner (CA method) finds directions in an unconstrained manner. CA's objective is to minimize the sum of the squared $\mathcal{X}^2$ distance under no constraints. In the presence of external features, the objective is to minimize this $\mathcal{X}^2$ distance along the weighted linear combination of external features. By doing so, it not only achieves the dependency between batting and bowling features as given in Definition 4.1 - Definition 4.5, but also constrains the search space along the weighted linear combination of external features as given in Definition 6.1 - Definition 6.5. Redundancy Analysis (RDA) [144] and Canonical Correspondence Analysis (CCA) [28] methods are used for this task. RDA is an extension of PCA to include external features. Similarly, CCA is the extension of CA include the external features. As we have previously ruled out PCA for our analysis (batting features and bowling features are discrete random variables), we can not use RDA. CCA is employed to incorporate external features to refine rule learning. CCA projects the data onto a subspace defined by the external features and performs CA. The steps of CCA are presented separately for batting analysis and bowling analysis in the following sections.

6.2.1 Batting Analysis through CCA

Refer to Figure 6.1 for the steps involved in batting analysis through CCA. For batting analysis, the relationships between batting features and external features are obtained through the bowling features. To achieve this, TCM of bowling features × batting features (denoted as N) and ECM of bowling features × external features (denoted as X) are constructed. CCA first obtains the residual matrix A from N. The projection matrix Q is obtained from X. Then A is projected onto Q to get the projected matrix A^* $(= QA)$, which is the interpretable part of X by the external features. SVD is applied to A^* to obtain the bowling principal components (F), batting principal components (G)

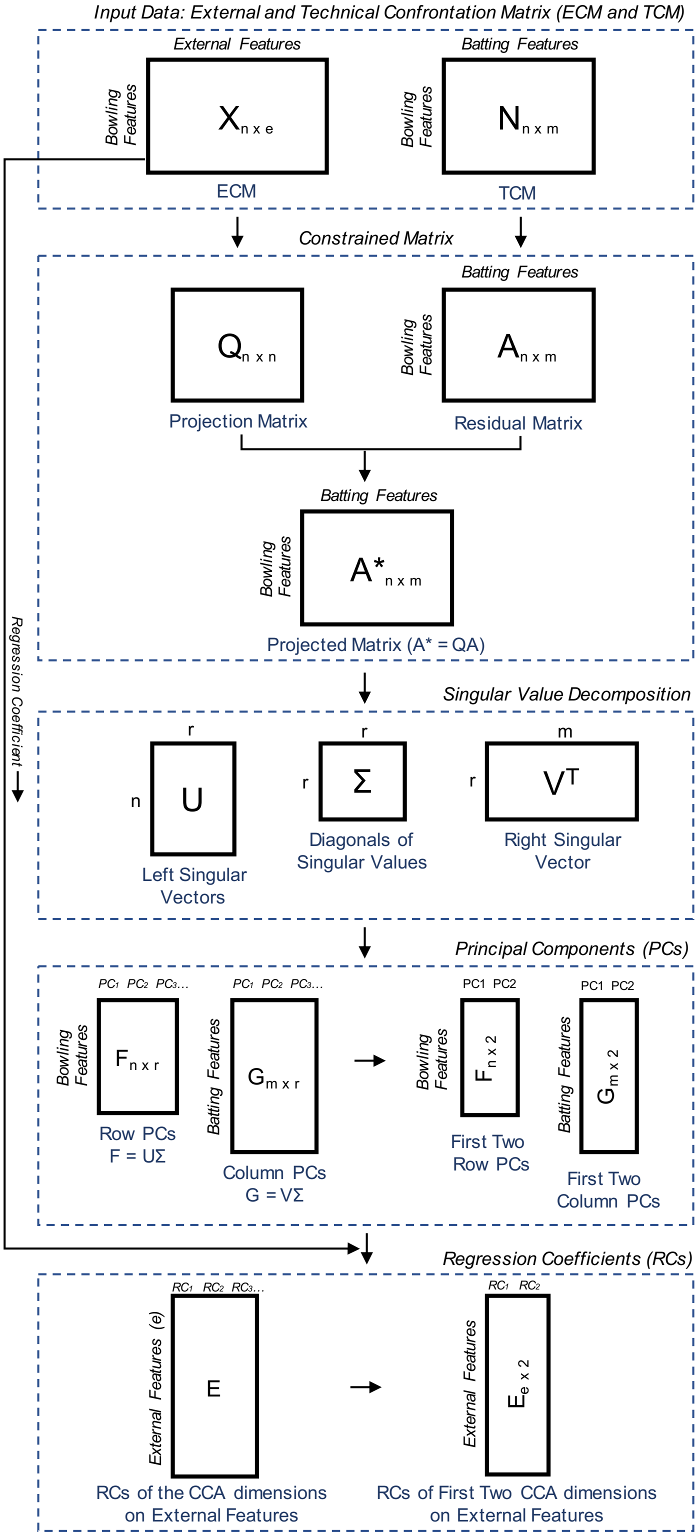


Figure 6.1: Batting Analysis through CCA.

Algorithm 8 CCA Algorithm (Batting Analysis)

```
Require: $TCM_{bowl \times bat}$ $(N_{I \times J})$ and $ECM_{bowl \times ext}$ $(X_{I \times M})$
Matrix sum: $n = \sum_{i=1}^{I} \sum_{j=1}^{J} N_{ij}$
Row masses(r): $r_i = \frac{N_{i.}}{n}, i = 1, 2, \cdots, I$
Diagonal matrix: $D_r = diag(r_1, r_2, ..., r_I)$
Column masses(c): $c_j = \frac{N_{.j}}{n}, j = 1, 2, \cdots, J$
Diagonal matrix: $D_c = diag(c_1, c_2, ..., c_J)$
Correspondence matrix: $P = \frac{1}{n} N$
Standardized residuals: $A = D_r^{-\frac{1}{2}} (P - rc^T) D_c^{-\frac{1}{2}}$
$I \times I$ projection matrix: $Q = D_r^{-\frac{1}{2}} X (X^T D_r X)^{-1} D_r^{-\frac{1}{2}}$
Project A onto Q to obtain constrained correspondence matrix: $A^* = QA$
Singular value decomposition: $A^* = U \Sigma V^T$
Principal components of rows: $G = D_r^{-\frac{1}{2}} U \Sigma$
Principal components of columns: $F = D_c^{-\frac{1}{2}} V \Sigma$
return F and G
```

in the constrained space of external features. Additionally, coordinates for the external features (E) are obtained, which are the standardized regression coefficients obtained by performing the weighted least square regression of the external features on the two principal axes. The CCA algorithm for batting analysis is presented in Algorithm 8. To plot all the three features in a two-dimensional plot, triplot [28] is used, where bowling features and batting features are represented by points and external features are represented by arrows. For the bowling features and batting features, the biplot interpretation holds. That is, the closer they are, the more dependent they are. The relationship between the batting features and external features is through the bowling features they have in common.

6.2.2 Bowling Analysis through CCA

Refer to Figure 6.2 for the steps involved in bowling analysis through CCA. For bowling analysis, the relationships between bowling features and external features are obtained through the batting features. To achieve this, TCM of batting features × bowling features (denoted as N) and ECM of batting features × external features (denoted as X) are constructed. CCA first obtains the residual matrix A from N. The projection matrix Q is obtained from X. Then A is projected onto Q to get the projected matrix A^* ($= QA$), which is the interpretable part of X by the external features. SVD is applied to A^* to obtain the batting principal components (F), bowling principal components (G) in the constrained space of external features. Additionally, coordinates for the external features (E) are obtained. The coordinates for the external features (E) are the standardized regression coefficients obtained by performing weighted least square regression of the external features on the two principal axes. The CCA algorithm for bowling analysis is presented in Algorithm 9. To plot all the three features in a two-dimensional plot, triplot [28] is used, where bowling features and batting features are represented by points and external features are represented by arrows. For the

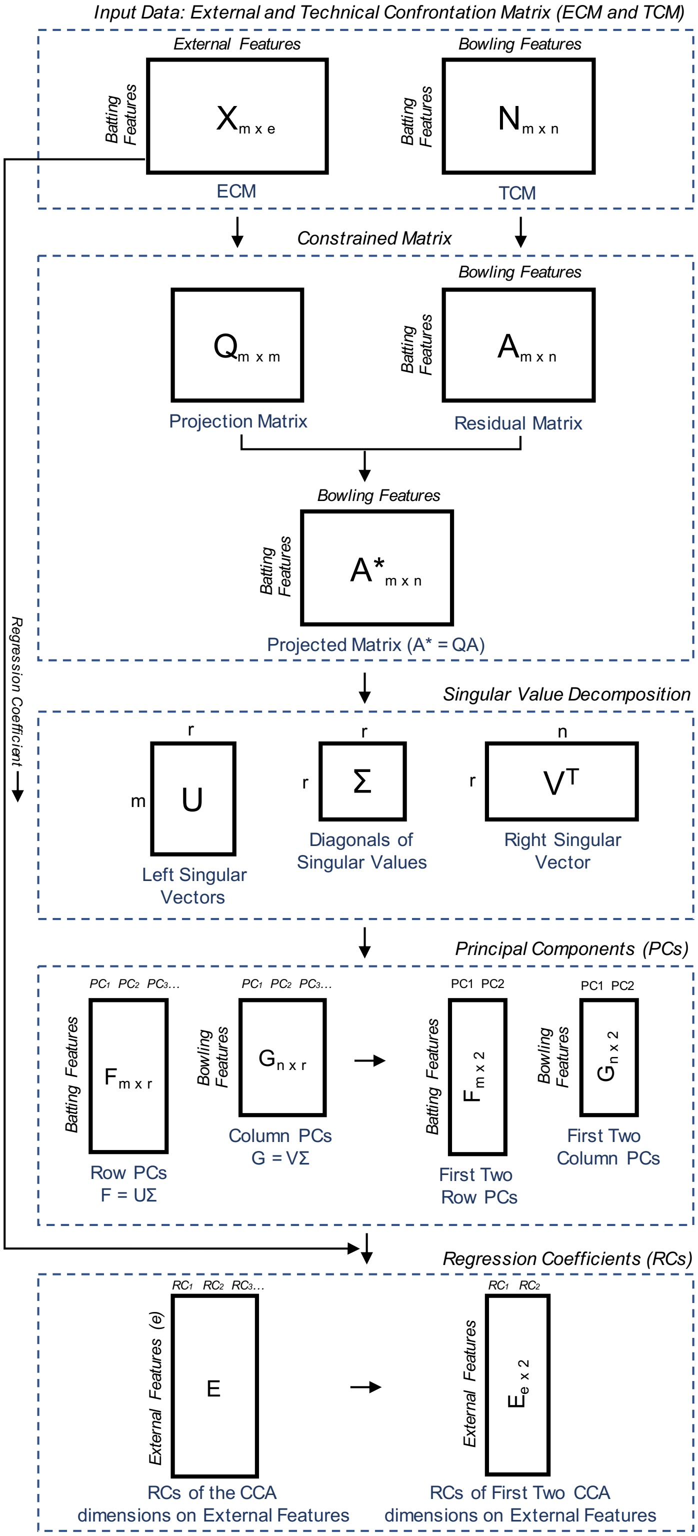


Figure 6.2: Bowling Analysis through CCA.

Algorithm 9 CCA Algorithm (Bowling Analysis)

```
Require: $TCM_{bat \times bowl}$ ($N_{I \times J}$) and $ECM_{bat \times ext}$ ($X_{I \times M}$)
Matrix sum: $n = \sum_{i=1}^{I} \sum_{j=1}^{J} N_{ij}$
Row masses(r): $r_i = \frac{N_{i.}}{n}, i = 1, 2, \cdots, I$
Diagonal matrix: $D_r = diag(r_1, r_2, ..., r_I)$
Column masses(c): $c_j = \frac{N_{.j}}{n}, j = 1, 2, \cdots, J$
Diagonal matrix: $D_c = diag(c_1, c_2, ..., c_J)$
Correspondence matrix: $P = \frac{1}{n}N$
Standardized residuals: $A = D_r^{-\frac{1}{2}}(P - rc^T)D_c^{-\frac{1}{2}}$
$I \times I$ projection matrix: $Q = D_r^{-\frac{1}{2}}X(X^T D_r X)^{-1}D_r^{-\frac{1}{2}}$
Project A onto Q to obtain constrained correspondence matrix: $A^* = QA$
Singular value decomposition: $A^* = U\Sigma V^T$
Principal components of rows: $G = D_r^{-\frac{1}{2}}U\Sigma$
Principal components of columns: $F = D_c^{-\frac{1}{2}}V\Sigma$
return F and G
```

bowling features and batting features, the biplot interpretation holds. That is, the closer they are, the more dependent they are. The relationship between the bowling features and external features is through the batting features they have in common.

6.3 Experiments

This section presents case studies that illustrate the strength and weakness analysis in the presence of external factors. First, we present the batting analysis of batsman Steve Smith. Next, we present the bowling analysis of bowler Kagiso Rabada. The data, code, and result of the experiments for more than 250 players can be accessed at https://www.dropbox.com/s/btrqii76tntl1ua/Chapter6%20External%20Factor%20Analysis.zip?dl=0.

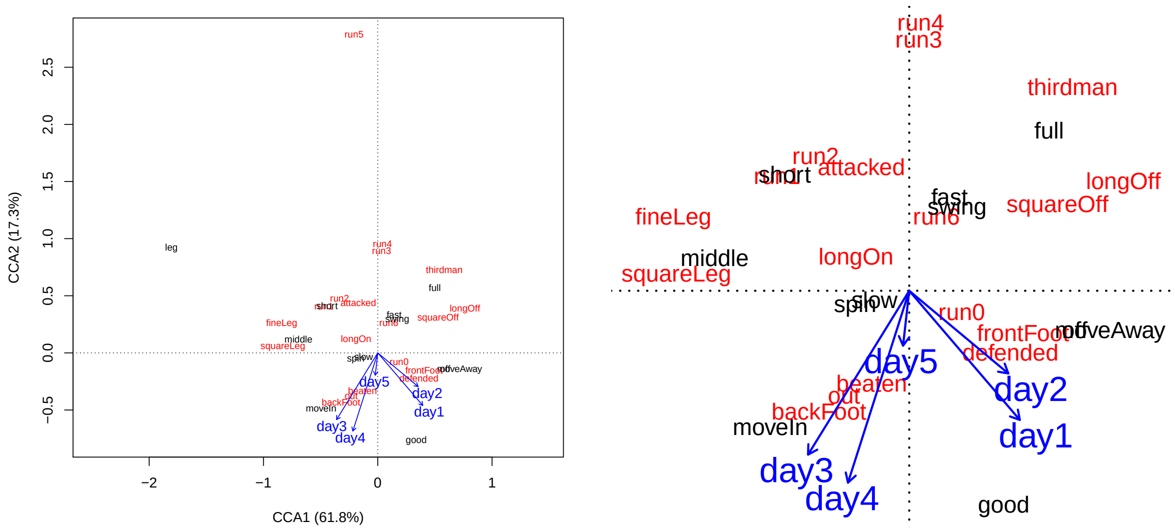
6.3.1 Batting Analysis - Steve Smith

We present the strengths and weaknesses of batsman Steve Smith in the presence of external factors such as day, age-of-ball, inning, session, pitch, and weather.

Day Analysis

This external feature describes the day (*day1*, *day2*, *day3*, *day4*, or *day5*) of the Test match on which batsman Steve Smith has batted. Through CCA analysis, we are interested in understanding if there is any influence of the day of play for Steve Smith to exhibit strength rule or weakness rule. Figure 6.3 presents the triplot of day-wise analysis for Steve Smith. The following rules are obtained from this triplot.

- *Strength Rule:* In this triplot, external feature E_{day3}, bowling feature $F_{moving-in}$, and batting



(a) Day Analysis Triplot.

(b) Day Analysis Triplot (Zoomed).

Figure 6.3: Day Analysis Triplot of Batsman Steve Smith.

features $G_{backfoot}$, G_{beaten}, G_{out} are closer to each other. Following the Definition 6.3 and Definition 6.1 for weakness rule construction, the proposed algorithm obtains the weakness rule of Steve Smith as - *On the third day of the Test match, bowlers tend to bowl moving-in deliveries to Steve Smith, and he plays them on the backfoot, gets beaten, or loses his wicket in those deliveries.*

- *Other Rule:* In this triplot, external feature E_{day1}, E_{day2}, bowling feature $F_{moving-away}$, F_{off}, and batting features G_{defend}, G_{run0}, $G_{frontfoot}$ are closer to each other. Thus, *On the first day and second day of the Test match, bowlers tend to bowl moving-away deliveries on the outside off line to Steve Smith, and he plays them on the frontfoot, defends, or scores no runs.*

The number of rules obtained using the CCA method is less than that of the CA method. This is because CCA projects the data onto a constrained subspace defined by the external features (in this case, days of a Test match) and performs CA.

Age-of-ball Analysis

This external feature describes the age of the ball being used in the Test match on which batsman Steve Smith has batted. The considered external features are *BallNew* (less than 10 overs old), *BallMid* (between 10 to 30 overs old), and *BallOld* (more than 30 overs old). Through CCA analysis, we are interested in understanding if there is any influence of the age-of-ball for Steve Smith to exhibit strength rule or weakness rule. Figure 6.4 presents the triplot of the age-of-ball analysis for Steve Smith. The following rules are obtained from this triplot.

- *Other Rule:* When the ball is less than ten overs old (BallNew) or between 10 to 30 overs old

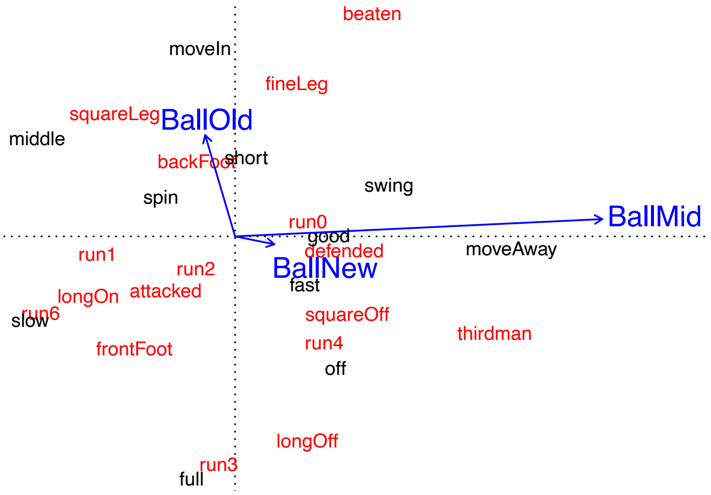


Figure 6.4: Age-of-ball Analysis Triplot of Batsman Steve Smith.

(BallMid), the bowlers tend to bowl good length and moving-away deliveries to Steve Smith, and he defends them or scores no runs.

- *Other Rule:* When the ball is more than 30 overs old (BallOld), bowlers tend to bowl moving-in or short length deliveries to Steve Smith, and he plays them on the back foot.

Inning Analysis

This external feature describes the inning on which batsman Steve Smith has batted in the matches. Every Test match has four innings, of which Steve Smith can play any two innings. The considered external features are *inning1*, *inning2*, *inning3*, and *inning4*. Through CCA analysis, we are

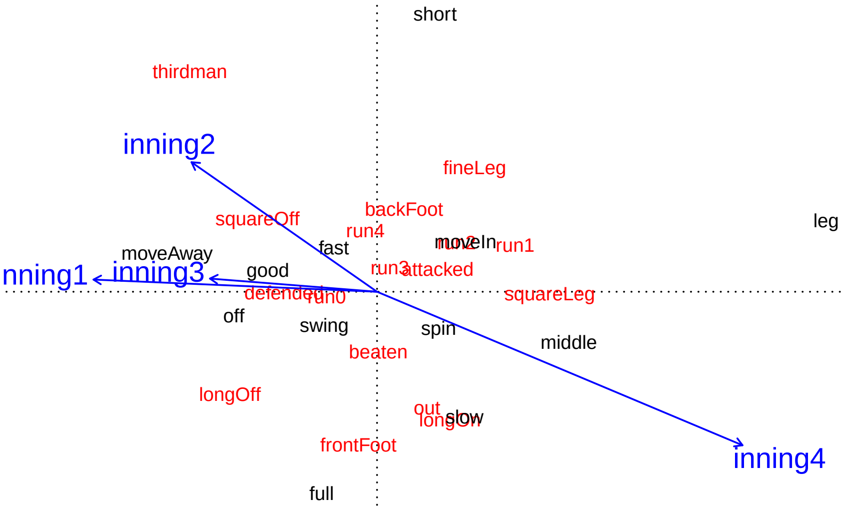


Figure 6.5: Inning Analysis Triplot of Batsman Steve Smith.

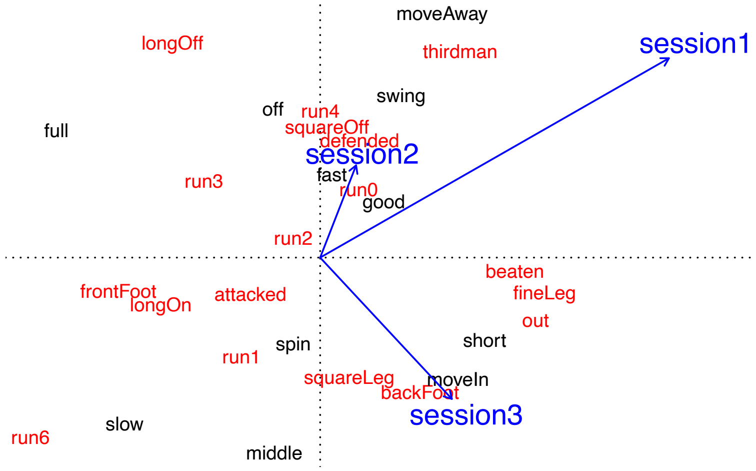


Figure 6.6: Session Analysis Triplot of Batsman Steve Smith.

interested in understanding if there is any influence of the inning of play for Steve Smith to exhibit strength rule or weakness rule. Figure 6.5 presents the triplot of inning wise analysis for Steve Smith. The following rules are obtained from this triplot.

- *Other Rule:* Steve Smith defends or scores no runs on a good length or moving-away deliveries bowled in either first innings or third innings.

Session Analysis

This external feature describes the session on which batsman Steve Smith has batted in the matches. Each day of the Test match has three sessions. The considered external features are *session1*, *session2*, and *session3*. Through CCA analysis, we are interested in understanding if there is any influence of the session of play for Steve Smith to exhibit strength rule or weakness rule. Figure 6.6 presents the triplot of session-wise analysis for Steve Smith. The following rules are obtained from this triplot.

- *Other Rule:* On session 2, bowlers tend to bowl moving-away, swing, or good length deliveries to Steve Smith, and he plays them to the thirdman area.

- *Other Rule:* On session 3, bowlers tend to bowl moving-in deliveries to Steve Smith, and he plays them on the backfoot.

Pitch Analysis

This external feature describes the pitch conditions on which batsman Steve Smith has batted in the matches. The considered external features are *green, grass, bounce, pace, slow, turning, swing, flat, and dry* (Refer to Section 3.4.2). Through CCA analysis, we are interested in understanding if there is any influence of the pitch conditions for Steve Smith to exhibit strength rule or weakness

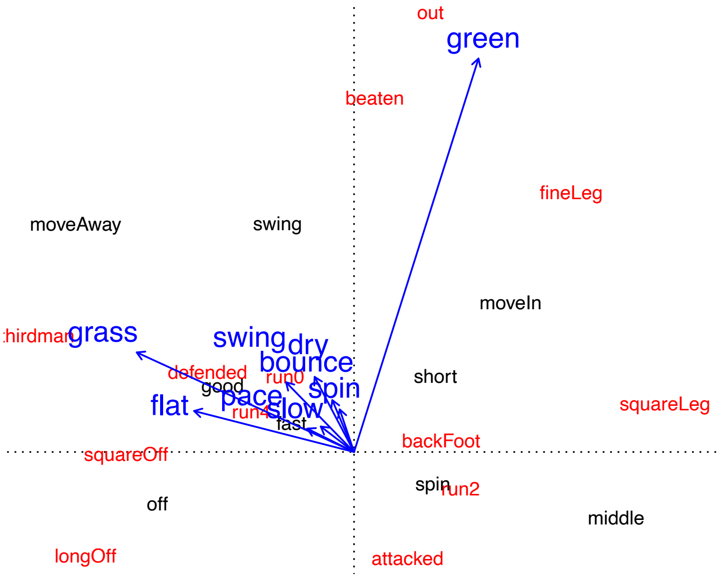


Figure 6.7: Pitch Analysis Triplot of Batsman Steve Smith.

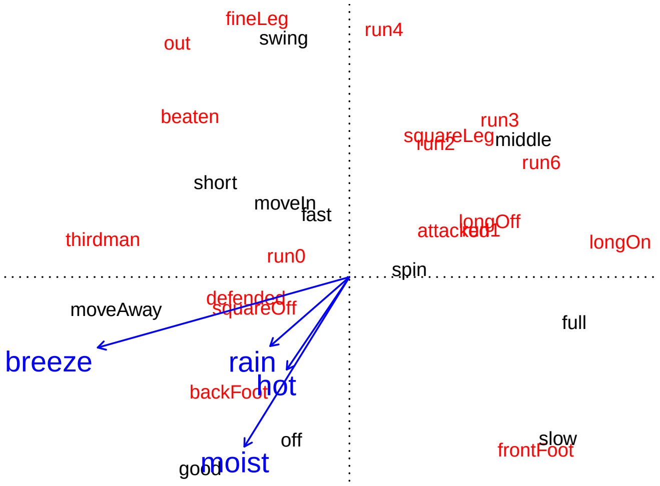


Figure 6.8: Weather Analysis Triplot of Batsman Steve Smith.

rule. Figure 6.7 presents the triplot of pitch-wise analysis for Steve Smith. The following rules are obtained from this triplot.

- *Weakness Rule:* Steve Smith gets beaten or losses his wicket on green pitches.
- *Other Rule:* Steve Smith defends the fast and good length deliveries on the grassy pitch.

Weather Analysis

This external feature describes the weather conditions on which batsman Steve Smith has batted in the matches. The considered external features are *rain*, *breeze*, *hot*, and *moist*. Through CCA analysis, we are interested in understanding if there is any influence of the weather condition for Steve Smith to exhibit strength rule or weakness rule. Figure 6.8 presents the triplot of weather-wise analysis for Steve Smith. The following rules are obtained from this triplot.

- *Other Rule:* In breezy weather, bowlers tend to bowl moving-away deliveries to Steve Smith, and he defends them or plays them to square off position.

6.3.2 Bowling Analysis - Kagiso Rabada

We present the strengths and weaknesses of bowler Kagiso Rabada in the presence of external factors such as day, age-of-ball, inning, session, pitch, and weather. Figure 6.9 presents the triplots depicting the strengths and weaknesses of bowler Kagiso Rabada in the presence of external factors. Following are a few of the rules obtained from these triplots.

1. Age-of-ball Analysis (Figure 6.9a)

    - *Other Rule:* When the ball is more than 30 overs old (BallOld), Rabada tends to bowl moving-away deliveries to batsmen, and the batsmen defend or play them to the thirdman area.

2. Day Analysis (Figure 6.9b)

    - *Other Rule:* On day-1 of a Test match, Rabada tends to bowl fast deliveries to batsmen, and batsmen play them to the thirdman area.

3. Inning Analysis (Figure 6.9c)

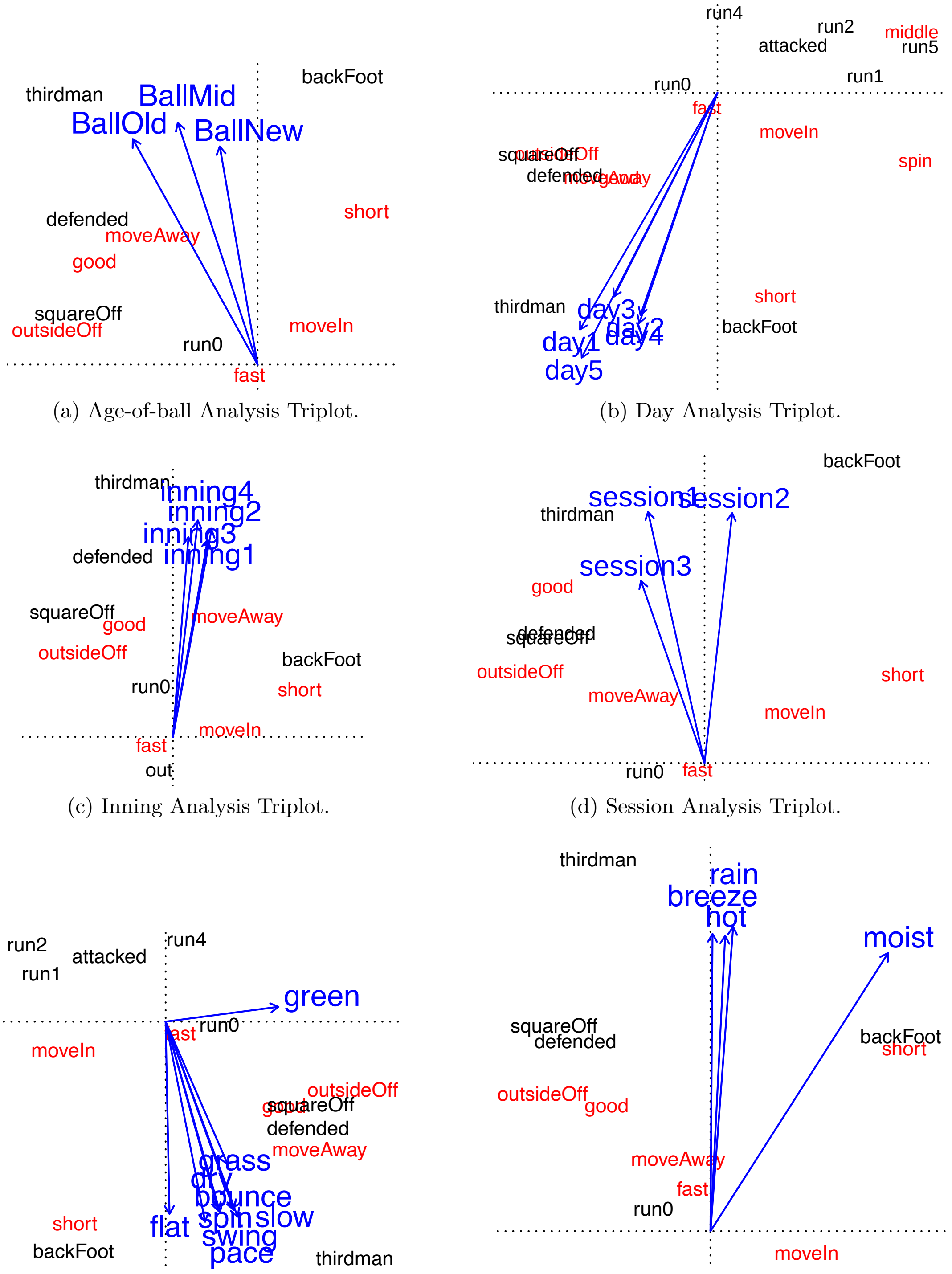
    - *Other Rule:* On inning 1 and inning 2, Rabada tends to bowl moving away deliveries.

4. Session Analysis (Figure 6.9d)

    - *Other Rule:* On session 2, Rabada tends to bowl moving-in deliveries to batsmen, and batsmen play them on backfoot.
    - *Other Rule:* On session 3, Rabada tends to bowl good length deliveries to batsmen, and batsmen play them to the thirdman area.

5. Pitch Analysis (Figure 6.9e)

    - *Other Rule:* On the flat pitch, Rabada tends to bowl fast and short length deliveries to batsmen, and batsmen play them on backfoot.
    - *Other Rule:* On the grass pitch, Rabada tends to bowl moving-away deliveries to batsmen, and batsmen defend them.
    - *Other Rule:* On the green pitch, batsmen mostly score no runs on Rabada's deliveries.



(a) Age-of-ball Analysis Triplot.

(b) Day Analysis Triplot.

(c) Inning Analysis Triplot.

(d) Session Analysis Triplot.

(e) Pitch Analysis Triplot.

(f) Weather Analysis Triplot.

Figure 6.9: Bowling Analysis Triplots of Bowler Kagiso Rabada.

6. Weather Analysis (Figure 6.9f)

- *Other Rule:* On moist weather, Rabada tend to bowl short length deliveries, and batsmen play them on the backfoot.

- *Other Rule:* On breezy weather, Rabada tend to bowl fast and moving away deliveries, and batsmen play them to the thirdman area.

6.4 Web Application

A web-based system is implemented to visualize player's strength and weakness rules in the presence of external factors. Users can select the batting analysis or the bowling analysis. The left panel displays a drop-down menu from which the batsman and bowler can be selected, which will load the selected player's TCM and ECM from the server's database. The triplots (days of a match, inning of a match, session of a match, age of the ball, weather condition, and pitch condition) of the selected batsman or bowler are displayed in the main panel. The system is live at https://cricketvisualization.shinyapps.io/StrengthWeaknessExternal/.

6.5 Summary

This chapter presented an approach to learn the strength and weakness rules of cricket players in the presence of external factors that are outside the game's scope and yet influence the gameplay. First, the approach introduced the computationally feasible definitions of strength and weakness rules in the presence of external factors. Next, it employed Canonical Correspondence Analysis (CCA) to construct semantic relations between batting features, bowling features, and external features. These relations are plotted in a triplot to visualize the strength and weakness rules in the presence of external factors. Several case studies showed that the proposed approach could identify cricket players' strengths and weaknesses in the presence of external factors.

In the proposed method, quality of opposition can be incorporated as a filter parameter (in filter tuple, opposition players can be high ranked or low ranked) or an external feature (player-specific analysis can be constrained on quality of opposition). A detailed investigation of quality as an external feature in strength and weakness rule construction is left for future work.

7

Visualization of Similar Players Based on their Strength and Weakness Rules

In Chapter 4, we looked at an approach to learn cricket player's strength and weakness rules. Grouping similar players based on their strengths and weaknesses yield in-depth insights. It will help devise strategies for team member selection and ordering of the batsman and bowler during a match (batting lineup and bowling lineup). In this chapter, we present an approach to visualize the similar players (batsmen or bowlers) based on their strength and weakness rules. The presented approach proceeds in two steps. In the first step, it obtains the strength vectors and weakness vectors of all the batsmen (or bowlers) from their strength and weakness rules. In the second step, it employs the t-distributed Stochastic Neighbor Embedding (t-SNE) [30] algorithm to visualize these high-dimensional vectors in a two-dimensional plot in which batsmen (or bowlers) having similar strength rule or similar weakness rule are placed closer. Validation of the obtained results is impractical due to the absence of ground truth on similar batsmen and similar bowlers.

We use the proposed approach to visualize the similar players in a t-SNE plot corresponding to 131 batsmen and 129 bowlers who have batted/bowled more than 2000 deliveries. The data, code, and result of the experiments can be accessed at `https://www.dropbox.com/s/d2nyibp5qggp13w/Chapter7%20Similar%20Players.zip?dl=0`. We highlight some of the obtained results below:

- *Pairs of similar batsmen based on their strength rule:* (i) Virat Kohli and VVS Laxman, (ii) Paul Collingwood and David Warner, and (iii) Michael Clarke and Brad Haddin.
- *Pairs of similar batsmen based on their weakness rule:* (i) Virender Sehwag and Misbah-ul-Haq, (ii) Jonathan Trott and Kevin Pietersen, and (iii) Salman Butt and Ian Bell.
- *Pairs of similar bowlers based on their strength rule:* (i) Josh Hazlewood and Rahat Ali and (ii) Umesh Yadav and Ben Stokes.
- *Pairs of similar bowlers based on their weakness rule:* (i) Brett Lee and Vernon Philander and (ii) Daren Powell, Dwane Bravo, and Matthew Hoggard.

7.1 Visualization of Similar Players

This section proposes a computational method that identifies similar batsmen and similar bowlers based on their strength and weakness rules. The proposed approach uses rule learner (CA method) presented in Section 4.2 to learn the rules of batsmen and bowlers. This section first presents an approach to visualize similar batsmen based on their strengths and weaknesses. Next, it presents a similar approach to visualize similar bowlers.

7.1.1 Visualization of Similar Batsmen

In order to visualize similar batsmen, TCM_{BAT} of all batsmen are obtained using a fixed filter tuple ⟨Player, All Opponent Bowlers, Career, Batting⟩. Refer to Figure 7.1 for an overview of our approach. Each of the obtained TCM_{BAT} ($TCM_{BAT}PLAYER_1$, $TCM_{BAT}PLAYER_2$, $\cdots$, $TCM_{BAT}PLAYER_N$) is subject to CA method proposed in Section. 4.2.1. CA first obtains the residual matrix A from $TCM_{BAT}PLAYER_x$. Next, SVD is applied to A to obtain the batting principal components (F), bowling principal components (G). Then, the first two principal components of F and G (denoted as $F_{m\times 2}$ *and* $G_{n\times 2}$) are obtained. Finally, the inner product matrix ($\langle F_{m\times 2}, G_{n\times 2}\rangle$) of the first two principal components of F and G is computed. From this inner product matrix, the first strength rule ($\langle F_{attacked}, G_j\rangle$, where j is the bowling vector which yields the highest inner product value with attacked batting vector) and first weakness rule ($\langle F_{beaten}, G_j\rangle$, where j is the bowling vector which yields the highest inner product value with beaten batting vector) of the batsman are obtained.

For each batsman, the row/batting vector ($F_{attacked}$) and the column/bowling vector G_j corresponding to the first strength rule are obtained. Similarly, for each batsman, the row/batting vector (F_{beaten}) and the column/bowling vector G_j corresponding to the first weakness rule are obtained. Next, the batting vector and bowling vector of strength rule are concatenated as - Strength Vector (SV_{BAT}). Similarly, the batting vector and bowling vector of weakness rule are concatenated as - Weakness Vector (WV_{BAT}). The SV_{BAT} and WV_{BAT} lie in 31 (19 + 12) dimension space representing each batsman's strength and weakness, respectively.

Visualization of Similar Batsmen Based on their Strength Rule

To visualize batsmen who have a similar strength rule, a non-linear dimensionality reduction technique, t-SNE [30], is employed on the high dimensional strength vectors of all batsmen ($SV_{BAT}P_1$, $SV_{BAT}P_2$, $\cdots$, $SV_{BAT}P_N$). The objective of t-SNE is to take a set of high-dimensional data points and obtain a lower-dimensional (typically two-dimension) representation of these points in such a way that similar points are modeled by nearby points, and dissimilar points are modeled by distant points with high probability. For this analysis, the value of perplexity (hyperparameter in t-SNE) is considered as 5. Applying t-SNE on the strength vectors, a two-dimensional plot is obtained in which batsmen having similar strength rules are placed closer.

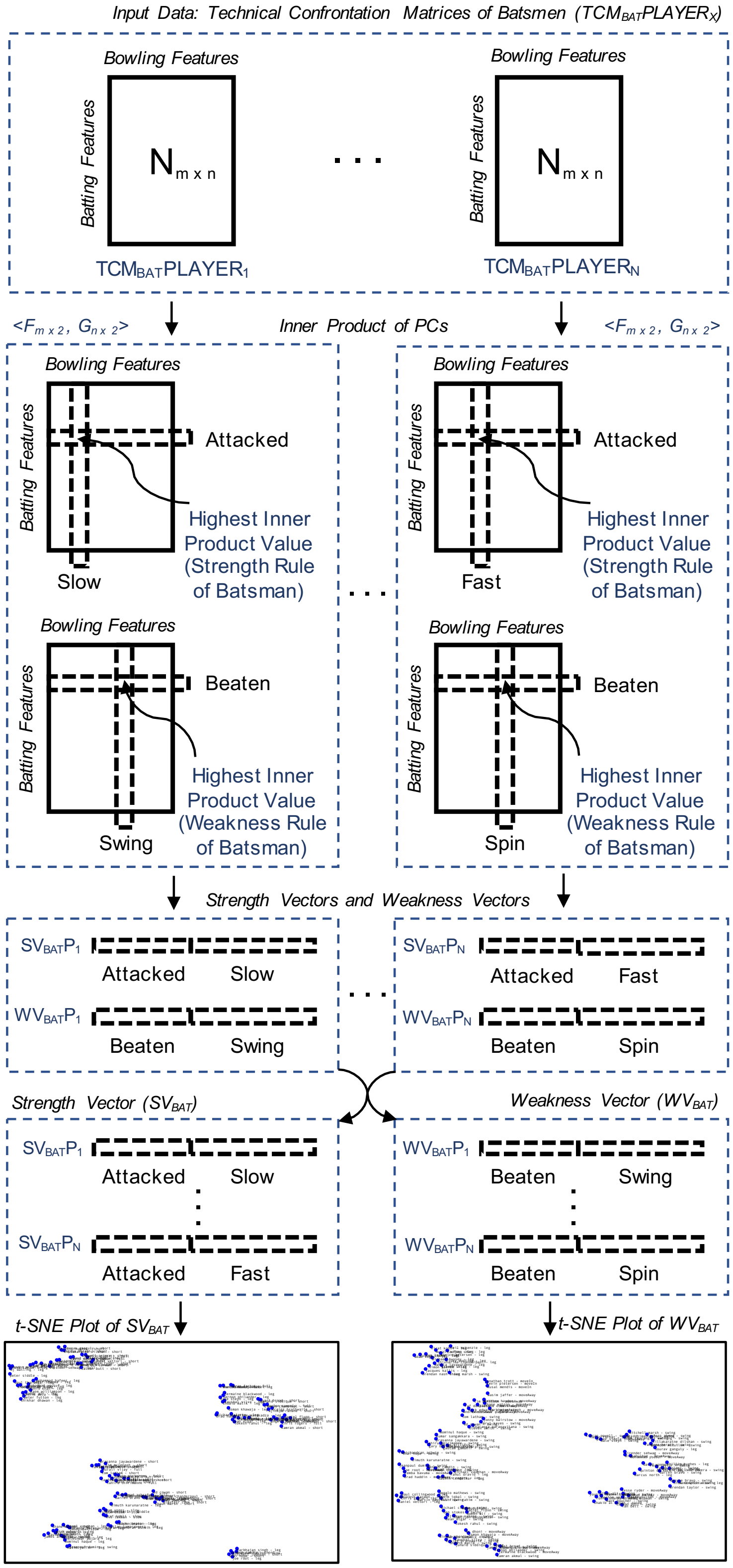

Input Data: Technical Confrontation Matrices of Batsmen ($TCM_{BAT}PLAYER_X$)
Bowling Features
Batting Features
$N_{m \times n}$
$TCM_{BAT}PLAYER_1$
$TCM_{BAT}PLAYER_N$
$<F_{m \times 2}, G_{n \times 2}>$
Inner Product of PCs
Attacked
Highest Inner Product Value (Strength Rule of Batsman)
Slow
Fast
Beaten
Highest Inner Product Value (Weakness Rule of Batsman)
Swing
Spin
Strength Vectors and Weakness Vectors
$SV_{BAT}P_1$
$WV_{BAT}P_1$
$SV_{BAT}P_N$
$WV_{BAT}P_N$
Strength Vector (SV_{BAT})
Weakness Vector (WV_{BAT})
t-SNE Plot of SV_{BAT}
t-SNE Plot of WV_{BAT}

Figure 7.1: Visualization of Similar Batsmen.

Visualization of Similar Batsmen Based on their Weakness Rule

To visualize batsmen who have a similar weakness rule, a non-linear dimensionality reduction technique, t-SNE [30], is employed on the high dimensional weakness vectors of all batsmen ($WV_{BAT}P_1$, $WV_{BAT}P_2, \cdots, WV_{BAT}P_N$). The objective of t-SNE is to take a set of high-dimensional data points and obtain a lower-dimensional (typically two-dimension) representation of these points in such a way that similar points are modeled by nearby points, and dissimilar points are modeled by distant points with high probability. For this analysis, the value of perplexity (hyperparameter in t-SNE) is considered as 5. Applying t-SNE on the weakness vectors, a two-dimensional plot is obtained in which batsmen having similar weakness rules are placed closer.

7.1.2 Visualization of Similar Bowlers

In order to visualize similar bowlers, TCM_{BOWL} of all bowlers are obtained using a fixed filter tuple ⟨Player, All Opponent batsmen, Career, Bowling⟩. Refer to Figure 7.2 for an overview of our approach. Each of the obtained TCM_{BOWL} ($TCM_{BOWL}PLAYER_1$, $TCM_{BOWL}PLAYER_2$, $\cdots$, $TCM_{BOWL}PLAYER_N$) is subject to CA method proposed in Section. 4.2.2. CA first obtains the residual matrix A from $TCM_{BOWL}PLAYER_x$. Next, SVD is applied to A to obtain the batting principal components (F), bowling principal components (G). Then, the first two principal components of F and G (denoted as $F_{m\times 2}$ *and* $G_{n\times 2}$) are obtained. Finally, the inner product matrix ($\langle F_{m\times 2}, G_{n\times 2}\rangle$) of the first two principal components of F and G is computed. From this inner product matrix, the first strength rule ($\langle F_{beaten}, G_j\rangle$, where j is the bowling vector which yields the highest inner product value with beaten batting vector) and first weakness rule ($\langle F_{attacked}, G_j\rangle$, where j is the bowling vector which yields the highest inner product value with attacked batting vector) of the bowlers are obtained.

For each bowler, the row/batting vector (F_{beaten}) and the column/bowling vector G_j corresponding to the first strength rule are obtained. Similarly, for each batsman, the row/batting vector ($F_{attacked}$) and the column/bowling vector G_j corresponding to the first weakness rule are obtained. Next, the batting vector and bowling vector of strength rule are concatenated as - Strength Vector (SV_{BOWL}). Similarly, the batting vector and bowling vector of weakness rule are concatenated as - Weakness Vector (WV_{BOWL}). The SV_{BOWL} and WV_{BOWL} lie in 31 (19 + 12) dimension space representing each batsman's strength and weakness, respectively.

Visualization of Similar Bowlers Based on their Strength Rule

To visualize bowlers who have a similar strength rule, a non-linear dimensionality reduction technique, t-SNE [30], is employed on the high dimensional strength vectors of all bowlers ($SV_{BOWL}P_1$, $SV_{BOWL}P_2$, $\cdots$, $SV_{BOWL}P_N$). The objective of t-SNE is to take a set of high-dimensional data points and obtain a lower-dimensional (typically two-dimension) representation of these points in such a way that similar points are modeled by nearby points, and dissimilar points are modeled by distant points with high probability. For this analysis, the value of perplexity (hyperparameter

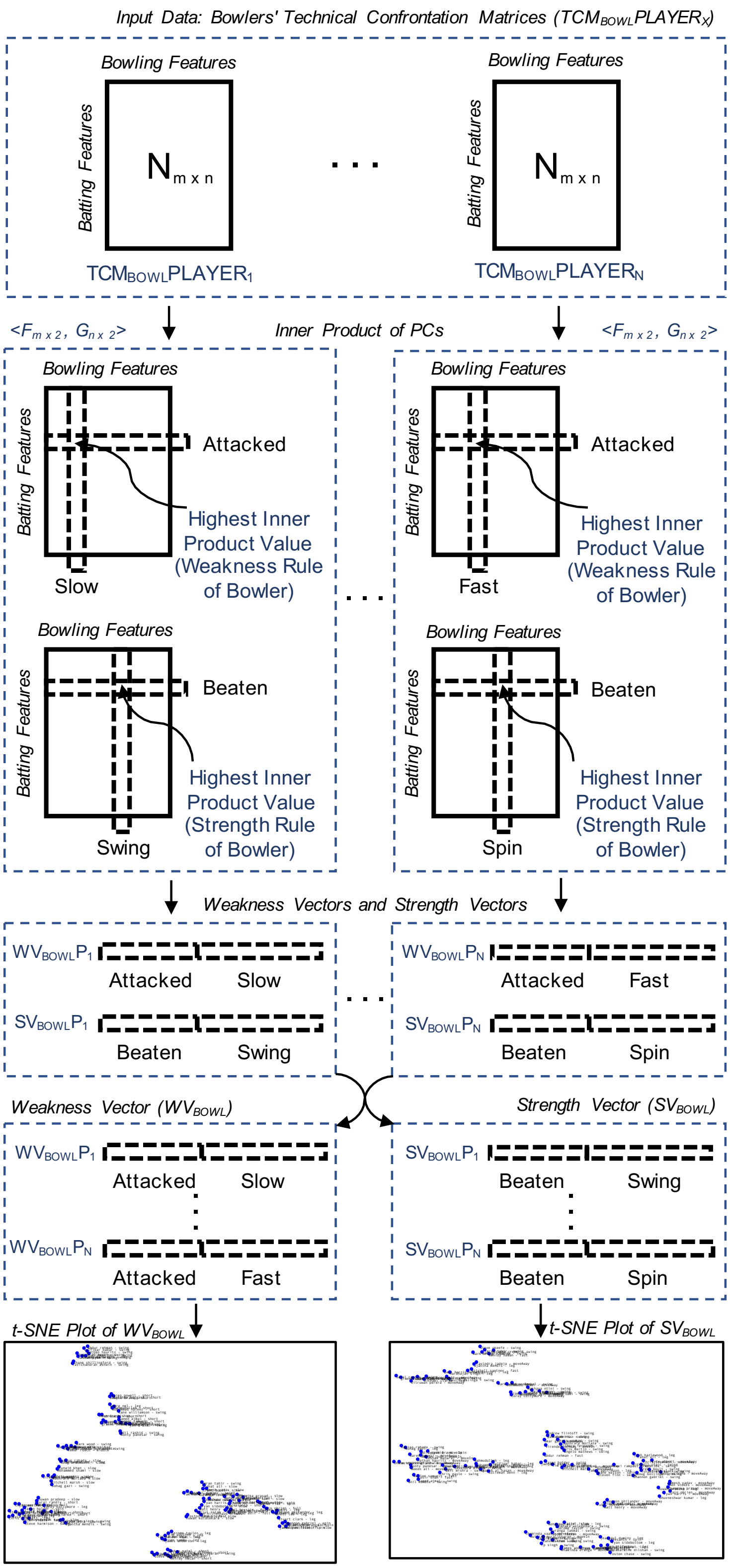


Figure 7.2: Visualization of Similar Bowlers.

in t-SNE) is considered as 5. Applying t-SNE on the strength vectors, a two-dimensional plot is obtained in which bowlers having similar strength rules are placed closer.

Visualization of Similar Bowlers Based on their Weakness Rule

To visualize bowlers who have a similar weakness rule, a non-linear dimensionality reduction technique, t-SNE [30], is employed on the high dimensional weakness vectors of all bowlers ($WV_{BOWL}P_1$, $WV_{BOWL}P_2$, $\cdots$, $WV_{BOWL}P_N$). The objective of t-SNE is to take a set of high-dimensional data points and obtain a lower-dimensional (typically two-dimension) representation of these points in such a way that similar points are modeled by nearby points, and dissimilar points are modeled by distant points with high probability. For this analysis, the value of perplexity (hyperparameter in t-SNE) is considered as 5. Applying t-SNE on the weakness vectors, a two-dimensional plot is obtained in which bowlers having similar weakness rules are placed closer.

7.2 Experiments

This section presents case studies that illustrate similar batsmen and similar bowlers bases on their strength and weakness rules using the proposed approach. First, we present the visualization of similar batsmen based on their strengths and weaknesses. We then present the visualization of similar bowlers based on their strengths and weaknesses. The data, code, and result of the experiments can be accessed at https://www.dropbox.com/s/d2nyibp5qggp13w/Chapter7%20Similar%20Players.zip?dl=0.

7.2.1 Visualization of Similar Batsmen

In this section, we first present the visualization of similar batsmen based on their strength rule. We then present the visualization of similar batsmen based on their weakness rules.

Visualization of Similar Batsmen Based on their Strength Rule

The t-SNE plot of 131 batsmen from 12 countries who have played more than 2000 deliveries is presented in Figure 7.3. Each point in this plot is the strength vector of a batsman. Each of these points is presented in the form of ⟨*batsman name - bowling feature*⟩, i.e., the batsman has exhibited strength on the deliveries which have the mentioned bowling feature. Intriguing batsman clusters are readily apparent, with a group of batsmen exhibiting strength on short-length deliveries and another group on leg-line deliveries in the upper left corner. Similar batsman clusters can be found in the upper right corner also. Identifying the close pairs of batsmen with similar strength rules is of interest. We present a few of the obtained close pair (similar) batsmen in Table 7.1. This table presents the pair of batsmen, the delivery (bowling feature) on which they have exhibited strength, their team, and the number of delivery played in their career.

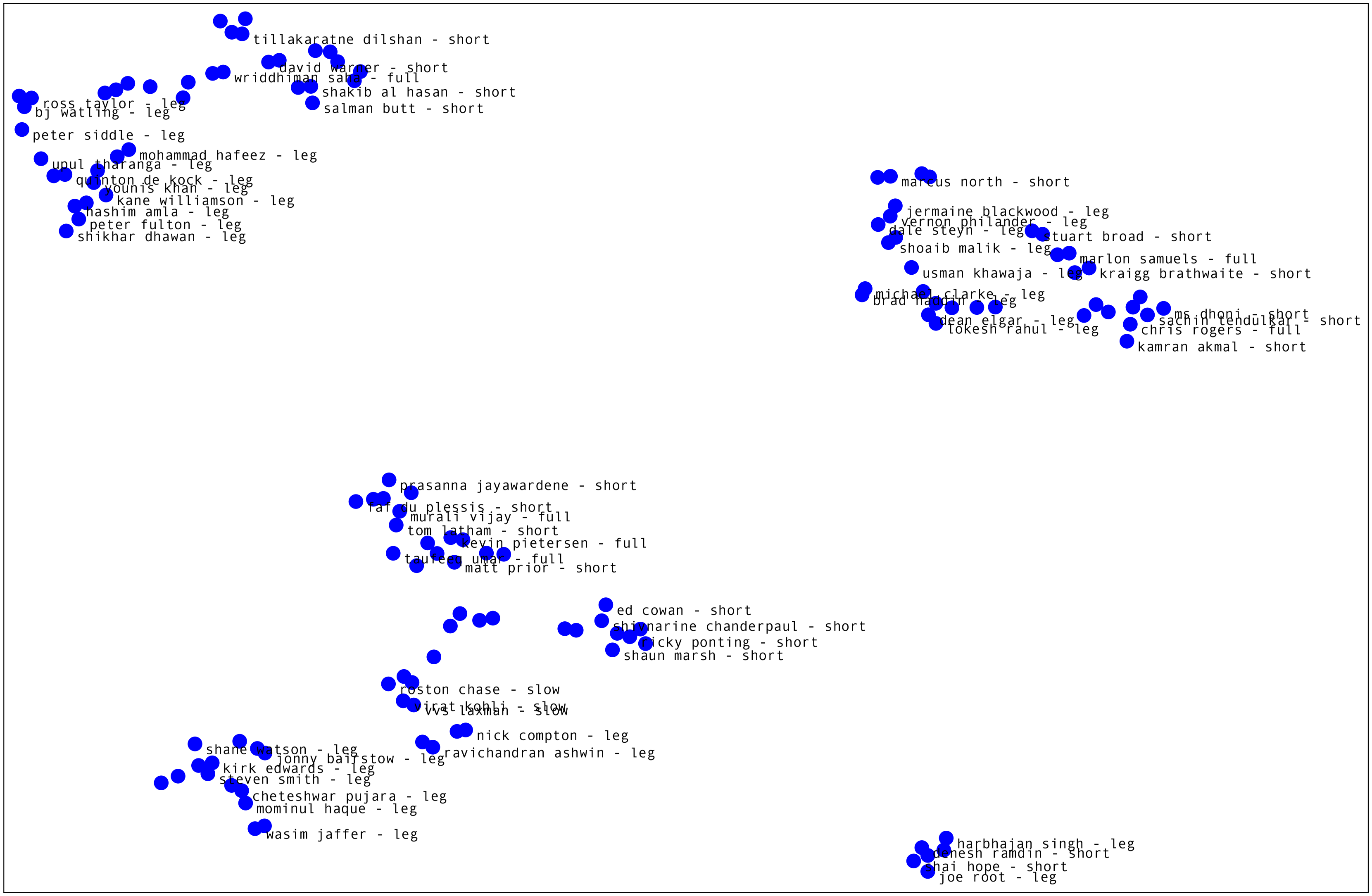


Figure 7.3: Visualization of Similar Batsmen Based on their Strength Rule.

Batsman Pair	**Exhibited Strength**	**Team**	**#balls$_{\textbf{batted}}$**
Virat Kohli	Slow deliveries	India	11568
VVS Laxman	Slow deliveries	India	8882
Paul Collingwood	Short length deliveries	England	7785
David Warner	Short length deliveries	Australia	8556
Michael Clarke	Leg line deliveries	Australia	13434
Brad Haddin	Leg line deliveries	Australia	5602

Table 7.1: Similar Batsman Pairs Based on their Strength Rule.

Batsman Pair	**Exhibited Weakness**	**Team**	**#balls$_{\textbf{batted}}$**
Virender Sehwag	Move-away deliveries	India	5525
Misbah-ul-Haq	Move-away deliveries	Pakistan	11375
Jonathan Trott	Move-in deliveries	England	8160
Kevin Pietersen	Move-in deliveries	England	12012
Salman Butt	Swing deliveries	Pakistan	2602
Ian Bell	Swing deliveries	England	13884

Table 7.2: Similar Batsman Pairs Based on their Weakness Rule.

Visualization of Similar Batsmen Based on their Weakness Rule

The t-SNE plot of 131 batsmen from 12 countries who have played more than 2000 deliveries is presented in Figure 7.4. Each point in this plot is a weakness vector of a batsman. Each of these points is presented in the form of ⟨*batsman name - bowling feature*⟩, i.e., the batsman has exhibited weakness on the deliveries which have the mentioned bowling feature. Intriguing batsman clusters are readily apparent, with a group of batsmen exhibiting weakness on leg-line deliveries and another group on moving-in deliveries in the upper left corner. Identifying the close pairs of batsmen with similar weakness rules is of interest. We present a few of the obtained close pair (similar) batsmen in Table 7.2. This table presents the pair of batsmen, the delivery (bowling feature) on which they have exhibited weakness, their team, and the number of delivery played in their career.

7.2.2 Visualization of Similar Bowlers

In this section, we first present the visualization of similar bowlers based on their strength rule. We then present the visualization of similar bowlers based on their weakness rules.

Visualization of Similar Bowlers Based on their Strength Rule

The t-SNE plot of 129 bowlers from 12 countries who have bowled more than 2000 deliveries is presented in Figure 7.5. Each point in the plot is the strength vector of a bowler. Each of these

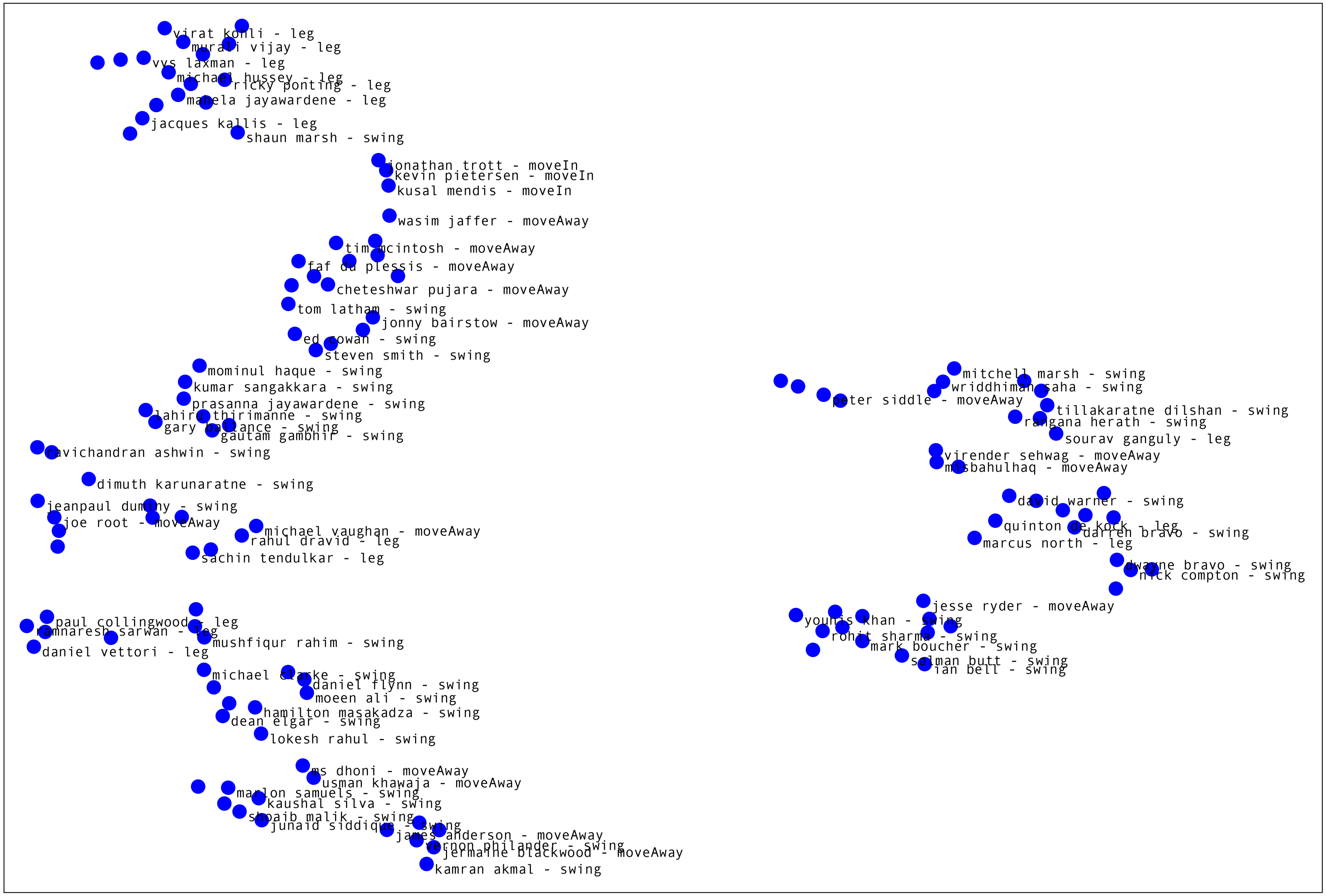


Figure 7.4: Visualization of Similar Batsmen Based on their Weakness Rule.

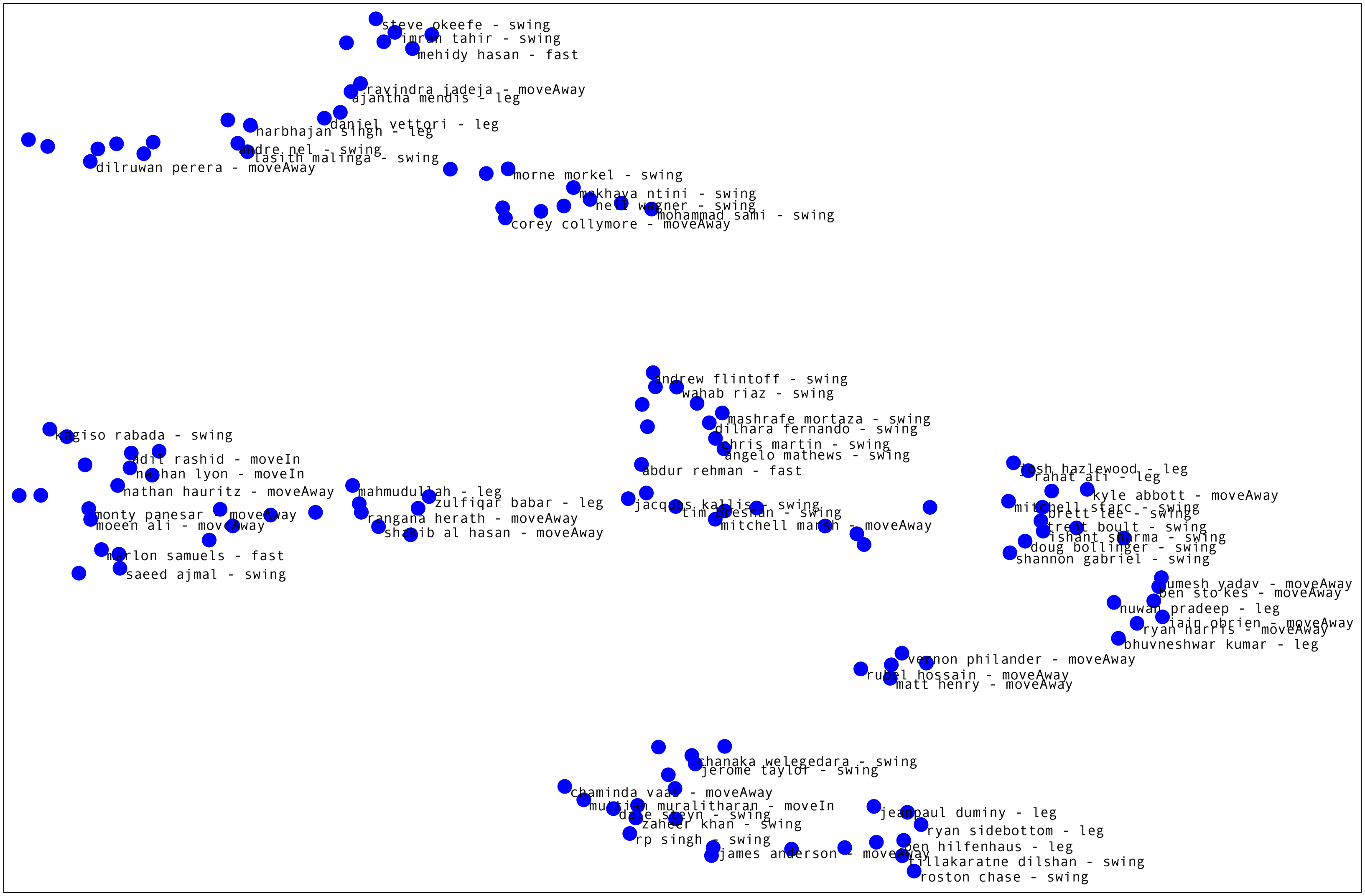


Figure 7.5: Visualization of Similar Bowlers Based on their Strength Rule.

Bowler Pair	**Exhibited Strength**	**Team**	**#balls$_{\textbf{bowled}}$**
Josh Hazlewood	Leg line deliveries	Australia	9585
Rahat Ali	Leg line deliveries	Pakistan	4260
Umesh Yadav	Move-away deliveries	India	6682
Ben Stokes	Move-away deliveries	England	7383

Table 7.3: Similar Bowler Pairs Based on their Strength Rule.

Bowler Pair	**Exhibited Weakness**	**Team**	**#balls$_{\textbf{bowled}}$**
Brett Lee	Leg line deliveries	Australia	5652
Vernon Philander	Leg line deliveries	South Africa	10521
Daren Powell	Short length deliveries	West Indies	3910
Dwane Bravo	Short length deliveries	West Indies	4784
Matthew Hoggard	Short length deliveries	England	3449

Table 7.4: Similar Bowler Pairs Based on their Weakness Rule.

points is presented in the form of ⟨*bowler name - bowling feature*⟩, i.e., the bowler has exhibited strength on the deliveries which have the mentioned bowling feature. Intriguing bowler clusters are readily apparent, with a group of bowlers exhibiting strength on moving-away deliveries in the right and another group on swing deliveries in the middle. Identifying the close pairs of bowlers with similar strength rules is of interest. We present a few of the obtained close pair (similar) bowlers in Table 7.3. This table presents the pair of bowlers, the delivery (bowling feature) on which the bowlers have exhibited strength, their team, and the number of delivery bowled in their career.

Visualization of Similar Bowlers Based on their Weakness Rule

The t-SNE plot of 129 bowlers from 12 countries who have bowled more than 2000 deliveries is presented in Figure 7.6. Each point in the plot is a weakness vector of a bowler. Each of these points is presented in the form of ⟨*bowler name - bowling feature*⟩, i.e., the bowler has exhibited weakness on the deliveries which have the mentioned bowling feature. Intriguing bowler clusters are readily apparent, with a group of bowlers exhibiting weakness on swing deliveries in the upper left corner and another group on leg-line deliveries in the lower middle. Identifying the close pairs of bowlers with similar weakness rules is of interest. We present a few of the obtained close pair (similar) bowlers in Table 7.4. This table presents the pair of bowlers, the delivery (bowling feature) on which the bowlers have exhibited weakness, their team, and the number of delivery bowled in their career.

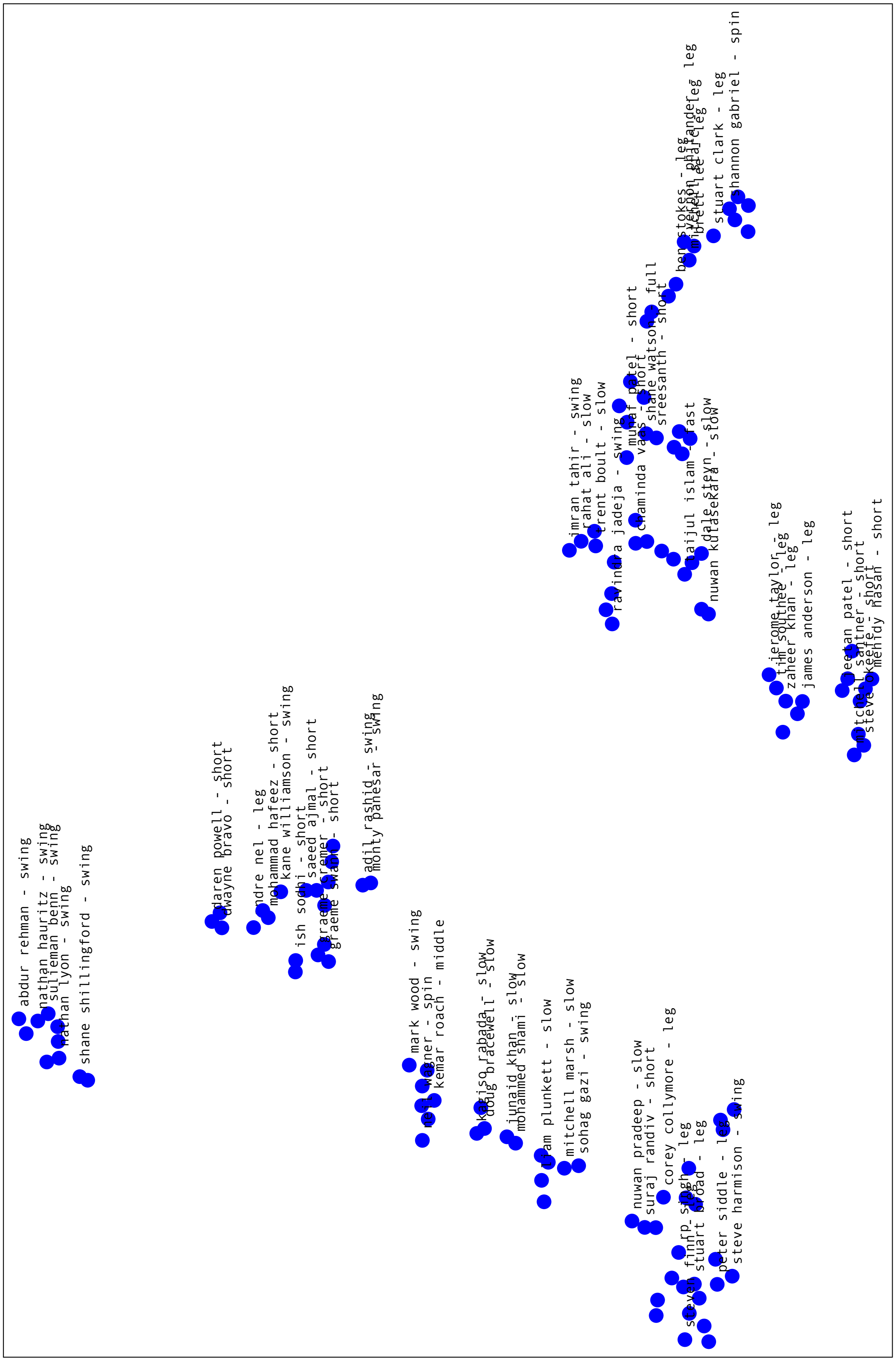


Figure 7.6: Visualization of Similar Bowlers Based on their Weakness Rule.

7.3 Summary

This chapter presented an approach to visualize similar batsmen or similar bowlers based on their strength and weakness rules. First, the approach obtained the strength vectors and weakness vectors of all the batsmen and bowlers from their strength and weakness rules. Second, it employed the t-SNE algorithm to visualize these high-dimensional vectors in a two-dimensional plot in which batsmen (or bowlers) having similar strength rules or similar weakness rules are placed closer. Several case studies showed that the proposed approach could identify similar batsmen and similar bowlers based on their strengths and weaknesses.

In the proposed method, t-SNE is used to visualize the high-dimensional strength and weakness vectors in a two-dimensional plot in which batsmen (or bowlers) having similar strength rules or similar weakness rules are placed closer. The use of different clustering methods on the output of t-SNE remains as future work. Clustering after t-SNE is challenging because t-SNE does not preserve distances nor density. It only preserves nearest-neighbors.

8

Conclusions and Future Directions

In this chapter, we present the conclusions of the work carried out within the scope of this thesis. Experiments aimed at building computational models for mining strength and weakness rules of cricket players were performed. Further, the influence of time and external factors on the obtained strengths and weaknesses were investigated. Using the obtained strength and weakness rules, similar batsmen and similar bowlers were also identified. The next sections describe these experiments briefly and then list some possible future directions of research.

8.1 Representing Cricket Text Commentary Data

In Chapter 3, we proposed the use of unstructured data, namely cricket text commentary, for learning player-specific strategies. We described the types of text commentary data used for this work: (i) *short text commentary* corresponding to every delivery or ball of every match and (ii) *external factor data* available for every session of every match.

We have constructed a large-scale dataset (from the ESPNcricinfo website) consisting of more than one million short text commentaries and over five thousand text commentaries for external factor data spanning over thirteen years (May 2006 and April 2019). We have identified the challenges inherent in processing text commentary data such as stopwords and sparsity and addressed those challenges. We have obtained fine-grained information about each player from the text commentary data and represented it using domain-specific features such as batting features, bowling features, and external features. To perform the player-specific analysis, one has to obtain a *subset of text commentaries* for the player. For this we proposed to use a *filter tuple* having four parameters ⟨*Player*, *Opponent Player*, *Time*, *Type*⟩.

For batsman analysis, the extracted batting features and bowling features are represented as a technical confrontation matrix in which rows correspond to the batsman's batting features and columns correspond to bowling features of opponent bowlers. The filter tuple used here is ⟨*Player*, *All Opponent Players*, *Career*, *Batting*⟩. Every element in this matrix corresponds to how the batsman confronted with the bowlers. For example, how many times the batsman has attacked the short-length deliveries bowled by all opponent bowlers in his entire career.

For bowler analysis, the extracted batting features and bowling features are represented as a technical confrontation matrix in which rows correspond to the opponent batsmen's batting features and columns correspond to the bowler's bowling features. The filter tuple used here is $\langle$*Player*, *All Opponent Players*, *Career*, *Bowling*$\rangle$. Every element in this matrix corresponds to how the bowler confronted with the batsmen. For example, how many times (in his entire career) all opponent batsmen attacked the bowler on his short-length deliveries.

For the analysis of external factors' influence on batting, in addition to a technical confrontation matrix, an external confrontation matrix is constructed in which rows correspond to bowling features of opponent bowlers and columns correspond to the external features.

For the analysis of external factors' influence on bowling, in addition to a technical confrontation matrix, an external confrontation matrix is constructed in which rows correspond to batting features of opponent batsmen and columns correspond to the external features.

8.2 Mining Strength and Weakness Rules of Cricket Players

In Chapter 4, we proposed an approach to learn the strength and weakness rules of cricket players using text commentary data. The proposed approach goes beyond runs and balls. It considers all the actions performed by the batsman and bowler on each delivery.

To compute the strength and weakness rules, arriving at a computationally feasible definition of what constitutes a rule is very important. We have provided the computationally feasible definitions of strength rule and weakness rule. We have employed a dimensionality reduction method, namely Correspondence Analysis, on the technical confrontation matrix to construct semantic relations between batting features and bowling features. We plotted these relations using biplots and extracts human readable strength and weakness rules.

We have applied the proposed approach to mine strength and weakness rules corresponding to 131 batsmen and 129 bowlers who have batted/bowled more than 2000 deliveries. In the experiments, we presented the batting analysis of batsman Steve Smith, the bowling analysis of bowler Kagiso Rabada, and Steve Smith's batting analysis against fast bowlers and spin bowlers.

The obtained rules, though easy to interpret, are difficult to validate. The constructed strength and weakness rules are validated using intrinsic and extrinsic methods. For extrinsic validation, the identified rules are verified against external sources such as a strategy-sheet and an expert. For intrinsic validation, the k-fold cross validation strategy is used, where the training and test data are compared using biplot comparison and rule comparison methods. Additionally, baseline comparisons are made using wordclouds and association rule mining. The obtained strengths and weaknesses will help analysts, coaches, and team management to build game strategies.

8.3 Mining Temporal Changes in Strength and Weakness Rules of Cricket Players

In Chapter 5, we proposed an approach to learn the temporal changes in the strength and weakness rules of cricket players using the text commentary data. Time granularity in cricket is identified in the increasing order as over, session, day, inning, match, series, season, year, or career. To measure the changes in strength and weakness rules over the years, the time granularity is considered as *year*.

To perform year-wise analysis, a time-dependent technical confrontation tensor is constructed from the year-wise technical confrontation matrices. Then the technical confrontation tensor is subject to a dimensionality reduction method, three-way correspondence analysis, and semantic relations between batting features, bowling features, and time (years) are obtained. These relations are plotted in a line plot to visualize the year-wise changes in strength and weakness rules.

We have applied the proposed approach to mine year-wise changes in strength and weakness rules corresponding to 131 batsmen and 129 bowlers who have batted/bowled more than 2000 deliveries. In the experiments, we presented the temporal analysis of batsman Steve Smith and bowler Kagiso Rabada. For both batsman and bowler, twelve distinct temporal analyses are presented, namely, *strength and weakness in full-length deliveries over the years*, *strength and weakness in good-length deliveries over the years*, etc.

8.4 Mining Strength and Weakness Rules of Cricket Players in the Presence of External Factors

In Chapter 6, we proposed an approach to learn the influence of external factors on the strength and weakness rules of cricket players. These external factors such as playing conditions (age-of-ball, pitch condition, and weather condition) and match situations (day, inning, and session) are outside the scope of the game and yet influence the player's strengths and weaknesses.

We have provided the computationally feasible definitions of strength rule and weakness rule in the presence of external factors. We have employed a dimensionality reduction method, Canonical Correspondence Analysis, on the technical confrontation matrix and external confrontation matrices to construct semantic relations between batting features, bowling features, and external features. These relations are plotted in a triplot to visualize the strength and weakness rules in the presence of external factors.

We have applied the proposed approach to mine strength and weakness rules in the presence of external factors corresponding to 131 batsmen and 129 bowlers who have batted/bowled more than 2000 deliveries. In the experiments, we presented the external factor analysis of batsman Steve Smith and bowler Kagiso Rabada.

8.5 Visualization of Similar Players Based on their Strength and Weakness Rules

In Chapter 7, we proposed an approach to visualize similar players (batsmen or bowlers) based on their strength and weakness rules. Grouping similar players based on their strengths and weaknesses yield in-depth insights on team member selection, batting lineup, and bowling lineup.

We have obtained the strength vectors and weakness vectors of all the batsmen (or bowlers) from their strength and weakness rules. We have employed t-SNE to visualize these high-dimensional vectors in a two-dimensional plot in which batsmen (or bowlers) having similar strength rule or similar weakness rule are placed closer.

We use the proposed approach to visualize similar players corresponding to 131 batsmen and 129 bowlers who have batted/bowled more than 2000 deliveries.

8.6 Future Research Directions

The following avenues could be explored as part of the future work of the thesis:

- *'Fact-like' vs. 'Insightful' Rule:* The objective of this work is to construct players' strength and weakness rules. The proposed methods also help construct rules involving outcome, footwork, and shot area. Of these, rules which include footwork or shot area might fall in the 'fact-like' category. However, this specific distinction has not been carried out in this thesis. We acknowledge the need for identifying 'fact-like' rules and the need to differentiate them from 'insightful' rules.

- *Quality of Opposition:* Quality of opponent player may certainly be a factor influencing individual player's strength and weakness rules. In this thesis, this factor is not considered for rule construction. We came across specific instances in the literature suggesting the influence of quality of players on team selection and performance of batsmen and bowlers. A detailed investigation of quality as an additional parameter in strength and weakness rule construction is left for future work.

- *Clustering on the Output of t-SNE:* In this thesis, t-SNE is used to visualize the high-dimensional strength and weakness vectors in a two-dimensional plot in which batsmen (or bowlers) having similar strength rules or similar weakness rules are placed closer. The use of different clustering methods on the output of t-SNE can provide new insights on player similarity. Clustering after t-SNE is challenging because t-SNE does not preserve distances nor density. It only preserves nearest-neighbors.

- *Introduce Multiview Learning into a Reinforcement Learning Framework for Modeling Short Text Commentary:* Reinforcement learning is a machine learning paradigm where the machines learn to complete specific tasks by taking a sequence of actions that leads them through the intermediate states to a goal state. The process of finding this sequence is guided by reinforcement, which is in the form of a reward received at the completion of a task. Q-learning is one of the algorithms that uses reinforcement learning to train the agent. It does not require any knowledge of the environment for the training. Hence, we think it might be suitable in the context of Cricket. As every delivery has two perspectives or views (batsman's view and bowler's view), we would like to introduce multiview learning to Q-learning.

- *Exploit the Dual Nature of Short Text Commentary for Dual Topic Modeling*: The Latent Dirichlet Allocation (LDA) topic model assumes every document is a mixture of topics and every topic is a Dirichlet distribution over words in the vocabulary. It has been applied to analyze documents having only one type of information, like words. LDA-dual extends LDA to be applied on documents containing two types of information (words and author names). To achieve this, it adds the third assumption - every topic is a Dirichlet distribution over all author names. Given the dual nature of information (batting and bowling) in each short text commentary, we think it might be suitable for dual topic modeling.
- *Temporal LDA-dual:* An extension of the LDA-dual model to capture the time-varying trends in topics for each player. The temporal LDA-dual model will capture the low dimensional structure of the data and the change in structure over time.

Bibliography

[1] Lewis, M.: Moneyball: The Art of Winning an Unfair Game. W. W. Norton & Company. (2004)

[2] Opta Sports. http://www.optasports.com. Accessed: 2021-09-14

[3] ESPNcricinfo. http://www.espncricinfo.com. Accessed: 2021-09-14

[4] CricBuzz. http://www.cricbuzz.com. Accessed: 2021-09-14

[5] Sky Sports. https://www.skysports.com/cricket. Accessed: 2021-09-14

[6] STATSports. https://statsports.com/cricket/. Accessed: 2021-09-14

[7] Duckworth, F.C., Lewis, A.J.: A fair method for resetting the target in interrupted one-day cricket matches. The Journal of the Operational Research Society. 49(3), 220-227 (1998)

[8] Stern, S.E.: The Duckworth-Lewis-Stern method: extending the Duckworth-Lewis methodology to deal with modern scoring rates. Journal of the Operational Research Society. 67(12), 1469-1480 (2016)

[9] Bailey, M.J., Clarke, S.R.: Predicting the Match Outcome in One Day International Cricket Matches, while the Game is in Progress. Journal of sports science & medicine. 5(4), 480-487 (2006)

[10] Allsopp, P., Clarke, S.: Rating teams and analysing outcomes in one-day and test cricket. Journal of the Royal Statistical Society Series A. 167(4), 657-667 (2004)

[11] Iyer, S., Sharda, R.: Prediction of Athletes Performance Using Neural Networks: An Application in Cricket Team Selection. Expert Syst. Appl. 36(3), 5510-5522 (2009)

[12] Das, A., Srinivasan, A., Stasko, J.: CricVis: Interactive Visual Exploration and Analysis of Cricket Matches. IEEE VIS. (2017)

[13] Theodoro, K., Saman, M., Dyson, B.C.: A Bayesian stochastic model for batting performance evaluation in one-day cricket. Journal of Quantitative Analysis in Sports. 10(1), 1-13 (2014)

[14] Lazarescu, M., Venkatesh, S., West, G.: On the automatic indexing of cricket using camera motion parameters. In Proceedings. IEEE International Conference on Multimedia and Expo. 1, 809-812 (2002)

[15] Davis, J., Perera, H., Swartz, T.: A Simulator for Twenty20 Cricket. Australian & New Zealand Journal of Statistics. 57(1), 55-71 (2009)

[16] Lemmer, H.H.: Team selection after a short cricket series. European Journal of Sport Science. 13(2), 200-206 (2013)

[17] Ahmed, F., Deb, K., Jindal, A.: Multi-objective optimization and decision making approaches to cricket team selection. Applied Soft Computing. 13(1), 402-414 (2013)

[18] Scarf, P., Akhtar, S.: An analysis of strategy in the first three innings in test cricket: declaration and the follow-on. The Journal of the Operational Research Society. 62(11), 1931-1940 (2011)

[19] Scarf, P., Shi, X., Akhtar, S.: On the distribution of runs scored and batting strategy in test cricket. Journal of the Royal Statistical Society: Series A (Statistics in Society). 174(2), 471-497 (2010)

[20] Morgan, W.G., Dinsdale, D., Gallagher, J., Cherukumudi, A., Lucey, P.: You Cannot Do That Ben Stokes: Dynamically Predicting Shot Type in Cricket Using a Personalized Deep Neural Network. MIT Sloan Sports Analytics Conference. (2020)

[21] ESPNcricinfo. What is Steve Smith's weakness? https://www.espncricinfo.com/video/what-is-steven-smith-s-weakness-1100538. Accessed: 2021-09-14

[22] Deccan Chronicle. Sri Lanka vs India: Unattended document leaks Virat Kohli and Co's Galle Test plans. https://www.deccanchronicle.com/sports/cricket/160917/leaked-heres-the-unattended-document-revealing-virat-kohli-led-team-indias-plans.html. Accessed: 2021-09-14

[23] Beh, E.J., Lombardo, R.: Correspondence Analysis: Theory, Practice and New Strategies, Wiley Series in Probability and Statistics, Wiley (2014)

[24] Greenacre, M.: Correspondence analysis in medical research. Statistical Methods in Medical Research. 1(1), 97-117 (1992)

[25] Greenacre, M.: Correspondence analysis in Practice. Chapman & Hall / CRC Interdisciplinary Statistics Series. CRC Press, Taylor & Francis. (2017)

[26] Gabriel, K.R.: The biplot graphic display of matrices with application to principal omponent analysis. Biometrika. 58(3), 453-467 (1971)

[27] Carlier, A., Kroonenberg, P.M.: Decompositions and biplots in three-way correspondence analysis. Psychometrika. 66, 355-373 (1996)

[28] Braak, C.J.F.T.: Canonical correspondence analysis: A new eigenvector technique for multivariate direct gradient analysis. Ecology. 67(5), 1167-1179 (1986)

[29] Oksanen, J.A.I.: Design decisions and implementation details in vegan. (2016)

[30] Van-der-Maaten, L.J.P., Hinton, G.E.: Visualizing High-Dimensional Data Using t-SNE. Journal of Machine Learning Research. 9, 2579-2605 (2008)

[31] Virtuale Eye. https://virtualeye.tv. Accessed: 2021-09-14

[32] Schumaker, R., Solieman, O., Chen, H.: Sports Data Mining, Springer US. (2010)

[33] David, J.A., Pasteur, R.D., Ahmad, M.S., Janning, M.C.: NFL prediction using committees of artificial neural networks, Journal of Quantitative Analysis in Sports. 7(2) (2011)

[34] von Dohlen, P.: Tweaking the NFL's quarterback passer rating for better results. Journal of Quantitative Analysis in Sports. 7(3), 1-14, (2011)

[35] Pasteur, R.D., Cunningham-Rhoads, K.: An expectation-based metric for NFL field goal kickers. Journal of Quantitative Analysis in Sports. 10(1), 49-66 (2014)

[36] Lock, D., Nettleton, D.: Using random forests to estimate win probability before each play of an NFL game. Journal of Quantitative Analysis in Sports. 10(2), 197-205 (2014)

[37] Albert, J.: Improved component predictions of batting and pitching measures. Journal of Quantitative Analysis in Sports. 12(2), 73-85 (2016)

[38] Baumer, B., Terlecky, P.: Improved estimates for the impact of baserunning in baseball. In JSM Proceedings, Statistics in Sports. ASA. (2010)

[39] Albert, J.: Pitching statistics, talent and luck, and the best strikeout seasons of all-time. Journal of Quantitative Analysis in Sports. 2(1), 1-30 (2006)

[40] Basco, D., Zimmerman, J.: Measuring defense: Entering the zones of fielding statistics. Baseball Research Journal. 39(1) (2010)

[41] Rosenbaum, D.: Measuring how NBA players help their teams win. http://www.82games.com/comm30.htm. Accessed: 2021-09-14

[42] Lock, D., Nettleton, D.: Using random forests to estimate win probability before each play of an NFL game. Journal of Quantitative Analysis in Sports. 10(2), 197-205 (2014)

[43] Annis, D.H.: Optimal end-game strategy in basketball. Journal of Quantitative Analysis in Sports. 2(2) (2006)

[44] Weil, S.: The importance of being open: What optical tracking data can say about nba field goal shooting. MIT Sloan Sports Analytics Conference. (2011)

[45] Franks, A., Miller, A., Bornn, L., Goldsberry, K.: Characterizing the spatial structure of defensive skill in professional basketball. The Annals of Applied Statistics. 9(1), 94-121 (2015)

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

[47] Scarf, P. A., Shi, X.: Modelling match outcomes and decision support for setting a final innings target in test cricket. IMA Journal of Management Mathematics, 16, 161-178 (2005)

[48] Croucher, J.: Player ratings in one-day cricket. In Proceedings of the Fifth Australian Conference on Mathematics and Computers in Sport. 95-106 (2000)

[49] Saikia, H., Bhattacharjee, D., Lemmer, H.H.: A Double Weighted Tool to Measure the Fielding Performance in Cricket. International Journal of Sports Science & Coaching. 7(4), 699-713 (2012)

[50] Jhanwar, M.G., Pudi, V.: Predicting the Outcome of ODI Cricket Matches: A Team Composition Based Approach. MLSA@PKDD/ECML. (2016)

[51] Chhabra, P., Ali, R., Pudi, V.: CRICTRS: Embeddings based Statistical and Semi Supervised Cricket Team Recommendation System. ArXiv, abs/2010.15607. (2020)

[52] Gramacy, R. B., Taddy, M., Tian, S.: Hockey player performance via regularized logistic regression. http://arxiv.org/pdf/1510.02172v2.pdf. Accessed: September 15, (2016)

[53] Morrison, D.G.: On the optimal time to pull the Goalie: A Poisson model applied to a common strategy used in ice hockey. TIMS Studies in Management Sciences. 4, 137-144 (1976)

[54] Gramacy, R.B., Jensen, S., Taddy, M.: Estimating player contribution in hockey with regularized logistic regression. Journal of Quantitative Analysis in Sports. 9, 97-111 (2013)

[55] Tingling, P.M., Masri, K., Martell, M.: Does decision order matter? An empirical analysis of the NHL draft. Sport, Business & Management: An International Journal. 1(2), 155-171 (2011)

[56] McHale, I., Scarf, P.: Modelling soccer matches using bivariate discrete distributions with general dependence structure. Statistica Neerlandica 61. 432-445 (2007)

[57] Titman, A., Costain, D., Ridall, P., Gregory, K.: Joint modelling of goals and bookings in association football. Journal of the Royal Statistical Society: Series A (Statistics in Society). 178, 659-683 (2015)

[58] Lasek, J., Szlavek, Z., Bhulai, S.: The predictive power of ranking systems in association football. International Journal of Pattern Recognition. 1(1), 27-46 (2013)

[59] McHale, I.G., Scarf, P.A., Folker, D.E.: On the development of a soccer player performance rating system for the English Premier League. Interfaces. 42, 339-351 (2012)

[60] Boyko, R., Boyko, A., Boyko, M.: Referee bias contributes to home advantage in English Premiership Football. Journal of Sports Sciences. 25, 1184-1194 (2007)

[61] Perin, C., Vuillemot, R., Stolper, C., Stasko, J., Wood, J.: State of the Art of Sports Data Visualization. Computer Graphics Forum, Wiley. 37(3), 1-24 (2018)

[62] Cox, A., Stasko, J.: Sportsvis: Discovering meaning in sports statistics through information visualization. In Compendium of Symposium on Information Visualization. 114-115 (2006)

[63] Turo, D.: Hierarchical Visualization with Treemaps: Making Sense of Pro Basketball Data. In Conference Companion on Human Factors in Computing Systems. 441-442 (1994)

[64] Perin, C., Boy, J., Vernier, F.: Using Gap Charts to Visualize the Temporal Evolution of Ranks and Scores. IEEE Computer Graphics and Applications. 36(5), 38-49 (2016)

[65] Gong, Y., Sin, L.T., Chuan, C.H., Zhang, H., Sakauchi, M.: Automatic parsing of TV soccer programs. In Proceedings of the International Conference on Multimedia Computing and Systems. 167-174 (1995)

[66] Assfalg, J., Bertini, M., Bimbo, A.D., Nunziati, W., Pala, P.: Soccer highlights detection and recognition using HMMs. In Proceedings. IEEE International Conference on Multimedia and Expo. 1, 825-828 (2002)

[67] Sudhir, G., Lee, J., Jain, A.: Automatic Classification of Tennis Video for High-level Content-based Retrieval. In Proceedings of the International Workshop on Content-Based Access of Image and Video Databases. (1998)

[68] Nepal, S., Srinivasan, U., Reynolds, G.: Automatic Detection of 'Goal' Segments in Basketball Videos. In Proceedings of the Ninth ACM International Conference on Multimedia. 261-269 (2001)

[69] Rui, Y., Gupta, A., Acero, A.: Automatically Extracting Highlights for TV Baseball Programs. In Proceedings of the Eighth ACM International Conference on Multimedia. 105-115 (2000)

[70] Sankar, K.P., Pandey, S., Jawahar, C.: Text Driven Temporal Segmentation of Cricket Videos. In Proceedings of the 5th Indian Conference on Computer Vision, Graphics and Image Processing. 433-444 (2006)

[71] Baillie, M., Jose, J.M.: An Audio-Based Sports Video Segmentation and Event Detection Algorithm. In Conference on Computer Vision and Pattern Recognition Workshop. 110-110 (2004)

[72] Zhang, D., Chang, S. F.: Event Detection in Baseball Video Using Superimposed Caption Recognition. In Proceedings of the Tenth ACM International Conference on Multimedia. 315-318 (2002)

[73] Xu, M., Duan, L., Xu, C., Tian, Q.: A fusion scheme of visual and auditory modalities for event detection in sports video. International Conference on Multimedia and Expo. 1, 1-6 (2003)

[74] Wu, Y., Xie, X., Wang, J., Deng, D., Liang, H., Zhang, H., Cheng, S., Chen, W.: Forvizor: Visualizing spatio-temporal team formations in soccer. IEEE Transactions on Visualization and Computer Graphics. (2018)

[75] Dietrich, C., Koop, D., Vo, H.T., Silva, C.T.: Baseball4d: A tool for baseball game reconstruction visualization. In IEEE Conference on Visual Analytics Science and Technology, 23-32 (2014)

[76] Beshai, P.: Buckets: Basketball Shot Visualization. (2020)

[77] Salton, G., Wong, A., Yang, C.S:. A Vector Space Model for Automatic Indexing. Commun. ACM 18(11), 613-620 (1975)

[78] Salton, G., Buckley. C.: Term-weighting Approaches in Automatic Text Retrieval. Inf. Process. Manage. 24(5), 513-523 (1988)

[79] Leopold, E., Kindermann, J.: Text Categorization with Support Vector Machines. How to Represent Texts in Input Space? Mach. Learn. 46, 423-444 (2002)

[80] Wilbur, W.J., Kim, W.: The ineffectiveness of within-document term frequency in text classification. Information Retrieval. 12(5), 509-525 (2009)

[81] Singhal, A., Salton, G., Mitra, M., Buckley, C.: Document length normalization. Information Processing & Management. 32(5), 619-633, (1996)

[82] Singhal, A., Buckley, C., Mitra, M.: Pivoted Document Length Normalization. In Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 21-29 (1996)

[83] Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., Harshman, R.: Indexing by latent semantic analysis. Journal of the American Society for Information Science. 41(6), 391-407 (1990)

[84] Hofmann, T.: Probabilistic Latent Semantic Indexing. SIGIR Forum. 51(2), 211-218 (2017)

[85] Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet Allocation. J. Mach. Learn. 3, 993-1022 (2003)

[86] Cai, D., Mei, Q., Han, J., Zhai, C.: Modeling Hidden Topics on Document Manifold. In Proceedings of the 17th ACM Conference on Information and Knowledge Management. 911-920, (2008)

[87] Cai, D., Wang, X., He, X.: Probabilistic Dyadic Data Analysis with Local and Global Consistency. In Proceedings of the 26th Annual International Conference on Machine Learning. 105-112 (2009)

[88] Huh, S., Fienberg, S.E.: Discriminative Topic Modeling Based on Manifold Learning. ACM Trans. Knowl. Discov. Data. 5(4), (2012)

[89] Matsuo, Y., Ishizuka, M., Bollegala, D.: Measuring Semantic Similarity Between Words Using Web Search Engines. In Proceedings of the 16th International Conference on World Wide Web, 757-766 (2007)

[90] Sahami, M., Heilman, T.D.: A Web-based Kernel Function for Measuring the Similarity of Short Text Snippets. In Proceedings of the 15th International Conference on World Wide Web, 377-386 (2006)

[91] Yih, W.T., Meek, C.: Improving Similarity Measures for Short Segments of Text. In Proceedings of the 22Nd National Conference on Artificial Intelligence. 2, 1489-1494 (2007)

[92] Banerjee, S., Ramanathan, K., Gupta, A.: Clustering Short Texts Using Wikipedia. In Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 787-788 (2007)

[93] Schonhofen, P.: Identifying Document Topics Using the Wikipedia Category Network. In Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence. IEEE Computer Society, 456-462 (2006)

[94] Phan, X.H., Nguyen, Le.M., Horiguchi, S.: Learning to Classify Short and Sparse Text & Web with Hidden Topics from Large-scale Data Collections. In Proceedings of the 17th International Conference on World Wide Web, 91-100 (2008)

[95] Hu, X., Tang, L., Tang, J., Liu, H.: Exploiting Social Relations for Sentiment Analysis in Microblogging. In Proceedings of the Sixth ACM International Conference on Web Search and Data Mining, 537âĂŞ546 (2013)

[96] Ni, X., Quan, X., Lu, Z., Wenyin, L., Hua, B.: Short Text Clustering by Finding Core Terms. Knowl. Inf. Syst. 27(3), 345-365 (2011)

[97] Sun, A.: Short Text Classification Using Very Few Words. In Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval, 1145-1146, (2012)

[98] Li, C., Duan, Yu., Wang, H., Zhang, Z., Sun, A., Ma., Z.: Enhancing Topic Modeling for Short Texts with Auxiliary Word Embeddings. ACM Trans. Inf. Syst. 36(2), (2017)

[99] Liang, S., Ren, Z., Zhao, Y., Ma, J., Yilmaz, E., Rijke, M.: Inferring Dynamic User Interests in Streams of Short Texts for User Clustering. ACM Trans. Inf. Syst. 36(1), (2017)

[100] Alencar, A.B., Oliveira, M.C.F., Paulovich, F.V.: Seeing beyond reading: A survey on visual text analytics, Wiley Interdiscip. Rev. Data Min. Knowl. Discov., 2, 476-492 (2012)

[101] Steinbock, D.: TagCrowd. http://www.tagcrowd.com/blog/about/ Accessed: 2021-09-14

[102] Viegas, F.B., Wattenberg, M., Feinberg, J.: Participatory visualization with wordle, IEEE Transactions on Visualization and Computer Graphics, 15, 1137-1144 (2009)

[103] Wattenberg, M., Viegas, F.B.: The Word Tree, an interactive visual concordance, IEEE Transactions on Visualization and Computer Graphics, 14, 1221-1228 (2008)

[104] Ham, F.V., Wattenberg, M., Viegas,F.B.: Mapping text with Phrase Nets, IEEE Transactions on Visualization and Computer Graphics, 15, 1169-1176 (2009)

[105] Skupin, A.: A cartographic approach to visualizing conference abstracts, IEEE Computer Graphics Application, 22, 50-58 (2002)

[106] Andrews, K., Kienreich, W., Sabol, V., Becker, J., Droschl, G., Kappe, F., Granitzer, M., Auer, P., Tochtermann, K., The InfoSky visual explorer: exploiting hierarchical structure and document similarities, Information Visualization, 1, 166-181 (2002)

[107] Lee, B., Riche, N.H., Karlson, A.K., Carpendale, S.: SparkClouds: Visualizing trends in tag clouds, IEEE Transactions on Visualization and Computer Graphics, 16(6), 1182-1189 (2010)

[108] Havre, S., Hetzler, E., Whitney, P., Nowell, L.: ThemeRiver: Visualizing thematic changes in large document collections, IEEE Transactions on Visualization and Computer Graphics, 8, 9-20 (2002)

[109] Joty, S., Carenini, G., Ng, R.T.: Topic segmentation and labeling in asynchronous conversations, Journal of Artificial Intelligence Research, 47, 521-573 (2013)

[110] Gao, M., Do, H., Fu, W.: An Intelligent Interface for Organizing Online Opinions on Controversial Topics, Proceedings of the 22Nd International Conference on Intelligent User Interfaces, 5, 119-123 (2017)

[111] Hu, M., Wongsuphasawat, K., Stasko, J.: Visualizing Social Media Content with SentenTree, IEEE Transactions on Visualization and Computer Graphics, 23(1), 621-630 (2017)

[112] Ho, C., Li, C., Lin, S.: Modeling and Visualizing Information Propagation in a Micro-blogging Platform. International Conference on Advances in Social Networks Analysis and Mining, 328-335 (2011)

[113] Zhao, J., Cao, N., Wen, Z., Song, Y., Lin, Y., Collins, C.: FluxFlow: Visual Analysis of Anomalous Information Spreading on Social Media, IEEE Transactions on Visualization and Computer Graphics, 20(12), 1773-1782 (2014)

[114] Rundell, M.: The Wisden Dictionary of Cricket (3rd ed.), A.& C. Black, 67 (2009)

[115] Cox, T., Cox, M.: Multidimensional Scaling, Second Edition. (2000)

[116] SchÃűlkopf, B., Smola, A., MÃijller, K.: Nonlinear Component Analysis as a Kernel Eigenvalue Problem. Neural Computation, 10, 1299-1319 (1998)

[117] Kohonen, T.: Self-Organizing Maps. Springer Series in Information Sciences. (1995)

[118] Saul, L., Roweis, S.: An Introduction to Locally Linear Embedding. (2001)

[119] Belkin, M., Niyogi, P.: Laplacian Eigenmaps for Dimensionality Reduction and Data Representation. Neural Computation, 15, 1373-1396 (2003)

[120] Jolliffe, I.T.: Principal Component Analysis. International Encyclopedia of Statistical Science. (2011)

[121] Abdi, H.: Singular Value Decomposition (SVD) and Generalized Singular Value Decomposition (GSVD). (2006)

[122] Gower, J.C., Dijksterhuis, G.B.: Procrustes problems. ser. Oxford Statistical Science Series. Oxford, UK: Oxford University Press, 30 (2004)

[123] Agrawal, R., Imieliński, T., Swami, A.: Mining Association Rules between Sets of Items in Large Databases. In Proc. of the ACM SIGMOD Conference on Management of Data. 22(2), 207-216 (1993)

[124] Agrawal, R., Srikant, R.: Fast algorithms for mining association rules. In Proc. Int'l Conf. Very Large Data Bases. 487-499 (1994)

[125] Hidber, C.: Online association rule mining. In Proc. Of the ACM SIGMOD International Conference on Management of Data. 28(2), 145-156 (1999)

[126] Pei, J., Han, J., Lu, H., Nishio, S., Tang, S. Yang, D.: H-Mine: Fast And Space-preserving Frequent Pattern Mining In Large Databases. IIE transactions. 39(6), 593-605 (2007)

[127] Satou, K., Shibayama, G., Ono, T., Yamamura, Y., Furuichi, E., Kuhara, S., Takagi, T.: Finding association rules on heterogeneous genome data. Pacific Symposium on Biocomputing. 397-408 (1997)

[128] Gupta, N., Mangal, N., Tiwari, K., Mitra, P.: Mining Quantitative Association Rules in Protein Sequences. In Proceedings of Australasian Conference on Knowledge Discovery and Data Mining-AUSDM. 273-281 (2006)

[129] Nahar, J., Tickle, K. S., Ali, S., Chen, Y. P.: Diagnosis Heart Disease using an Association Rule Discovery Approach. In Proceedings of the IASTED International Conference Computational Intelligence. (2009)

[130] Hsieh, N. C.: An integrated data mining and behavioral scoring model for analyzing bank customers. Expert Systems with Applications. 27(4), 623 - 633 (2004)

[131] Chen, R. S., Wu, R. C., Chen, J. Y.: Data Mining Application in Customer Relationship Management Of Credit Card Business. In Proceedings of 29th Annual International Computer Software and Applications Conference (COMPSAC'05). 2, 39-40 (2005)

[132] Brijs, T., Goethals, B., Swinnen, G., Vanhoof, K., Wets, G.: A Data Mining Framework for Optimal Product Selection in Retail Supermarket Data: The Generalized PROFSET Model. In Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 300-304 (2000)

[133] Brin, S., Motwani, R., Ullman, J. D., Tsur, S.: Dynamic Itemset Counting And Implication Rules for Market Basket Data. In ACM SIGMOD Record. 26(2), 255-264 (1997)

[134] Puchun, W.: The Application of Data Mining Algorithm Based on Association Rules. Analysis of Football Tactics. 418-421 (2016)

[135] Liao, S. H., Chen, J. L., Hsu, T. Y.: Ontology-based data mining approach implemented for sport marketing. Expert Systems with Applications. 36(8), 11045-11056 (2009)

[136] Sun, J., Yu, W., Zhao, H.: Study of Association Rule Mining on Technical Action of Ball Games. Int. Conference on Measuring Technology and Mechatronics Automation. 539-542 (2010)

[137] Raj, K. A. A. D., Padma, P.: Application of Association Rule Mining: A case study on team India. In International Conference on Computer Communication and Informatics. 1-6 (2013)

[138] UmaMaheswari, P., Rajaram, M.: A Novel Approach for Mining Association Rules on Sports Data using Principal Component Analysis: For Cricket match perspective. In 2009 IEEE International Advance Computing Conference. 1074-1080 (2009)

[139] Lemmer, H.: The allocation of weights in the calculation of batting and bowling performance measures. South African Journal for Research in Sport Physical Education and Recreation, 29, 75-86 (2007)

[140] Behera, S.R., Agrawal, P., Awekar, A., Vedula, V.S.: Mining Strengths and Weaknesses of Cricket Players Using Short Text Commentary, 18th IEEE International Conference On Machine Learning And Applications (ICMLA), Boca Raton, FL, USA. 673-679 (2019).

[141] Tucker, L.R.: Some Mathematical Notes on Three-Mode Factor Analysis. Psychometrika, 30, 279-311 (1996)

[142] Meyer, C.D.: Matrix Analysis and Applied Linear Algebra. Other Titles in Applied Mathematics. Society for Industrial and Applied Mathematics (2000)

[143] Behera, S.R., Vedula, V.S.: Mining temporal changes in strengths and weaknesses of cricket players using tensor decomposition. In 28th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN), Bruges, Belgium (2020)

[144] Legendre, P., Anderson, M.J.: Distance-based Redundancy Analysis: Testing Multispecies Responses in Multifactorial Ecological Experiments. Ecological Monographs, 69, 1-24 (1999)

Publications

Conference Papers

- **S. R. Behera**, P. Agrawal, A. Awekar and V. S. Vedula, *Mining Strengths and Weaknesses of Cricket Players Using Short Text Commentary*, 18th IEEE International Conference On Machine Learning And Applications (ICMLA'19), Boca Raton, FL, USA, 2019.

- **S. R. Behera** and V. S. Vedula, *Mining Temporal Changes in Strengths and Weaknesses of Cricket Players Using Tensor Decomposition*, 28th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN'20), Bruges, Belgium, 2020.

- **S. R. Behera** and V. S. Vedula, Stats Aren't Everything; Learning Strengths and Weaknesses of Cricket Players, In: Brefeld U., Davis J., Van Haaren J., Zimmermann A. (eds) Machine Learning and Data Mining for Sports Analytics (MLSA'20), ECML/PKDD'20, Communications in Computer and Information Science, vol 1324, Springer, 2020.

- **S. R. Behera** and V. S. Vedula, Learning Strength and Weakness Rules of Cricket Players using Association Rule Mining, Machine Learning and Data Mining for Sports Analytics (MLSA'21), ECML/PKDD'21.

Conference Posters

- **S. R. Behera** and V. S. Vedula, Video Data Do More, Tracking Data Do Much, Text Commentary Data Do Much More, Carnegie Mellon Sports Analytics Conference (CMSAC'20), Pittsburgh, United States, 2020.

- **S. R. Behera** and V. S. Vedula, Batsman's Kryptonite: Learning Weakness and Strength Rules of Cricket Players using Association Rule Mining, Carnegie Mellon Sports Analytics Conference (CMSAC'20), Pittsburgh, United States, 2020.

- **S. R. Behera**, Performance Analysis of Batsman against Spin Bowling and Fast Bowling in Cricket, Ohio State Sports Analytics Association Conference (OSUSAAC'20), Columbus, United States, 2020. ***(Best Research Award)***

Journals

- **S. R. Behera** and V. S. Vedula, *Mining Unstructured Sports Data - An Example of Cricket Text Commentary.* (In Preparation)
- **S. R. Behera**, Harshith Reddy Padigela, Rahul R Huilgol and V. S. Vedula, *Learning Strengths and Weaknesses of Cricket Players in the Presence of External Factors Influencing the Game.* (In Preparation)
- **S. R. Behera** and V. S. Vedula, *Mining Temporal Changes in Strengths and Weaknesses of Cricket Players.* (In Preparation)

Web Applications

- **S. R. Behera**, *Visualization of Strengths and Weaknesses of Cricket Players Using Text Commentary Data* (Version 1.0) [Web application], 2020. Retrieved from https://cricketvisualization.shinyapps.io/StrengthWeaknessAnalysis/
- **S. R. Behera**, *Visualization of Temporal Changes in Strengths and Weaknesses of Cricket Players* (Version 1.0) [Web application], 2020. Retrieved from https://cricketvisualization.shinyapps.io/StrengthWeaknessTemporal/
- **S. R. Behera**, *Visualization of Strengths and Weaknesses of Cricket Players in the Presence of External Factors* (Version 1.0) [Web application], 2020. Retrieved from https://cricketvisualization.shinyapps.io/StrengthWeaknessExternal/

Vitae

Swarup Ranjan Behera was born on 1st March 1989. He obtained his B.Tech degree in Computer Science and Engineering at Veer Surendra Sai University of Technology, Burla, India, in 2012 and his M.Tech degree in Computer Science and Engineering at Indian Institute of Technology Guwahati, India, in July 2015. Swarup started his doctoral studies at the Department of Computer Science and Engineering of Indian Institute of Technology Guwahati, India, in July 2015 under the supervision of Dr. Vijaya Saradhi Vedula. He received a personal 5-year Ph.D. fellowship from the MHRD, Government of India. Swarup has published/presented papers at machine learning conferences such as ICMLA, ESANN, ECML/PKDD, and sports analytics conferences such as CMSAC and OSUSAAC. He has received the best research award at the 2020 Ohio State Sports Analytics Association Conference. He has a keen interest in pursuing machine learning, text mining, and sports data mining. He enjoys reading poetry, listening to music, watching movies, and playing cricket.

Contact Information

Email	:	b.swarup@iitg.ac.in swarup221b@gmail.com swaruprj.vssut@gmail.com
Web	:	https://swarup-rj.github.io/
Address	:	H.No. 23, At/Po - Tarkera, Via - Birmitrapur, Dist.- Sundargarh, Odisha, India - 770033