Learning Strength and Weakness Rules of Cricket Players using Association Rule Mining

Swarup Ranjan Behera and V. Vijaya Saradhi

Indian Institute of Technology Guwahati
Guwahati, Assam, India
{b.swarup,saradhi}@iitg.ac.in

Abstract. Association rule mining is an important data mining technique that finds association rules by mining frequent attributes. This work aims to construct association rules that determine cricket players' strengths and weaknesses. We propose an approach to learn the association of strengths or weaknesses exhibited by batters (or bowlers) with the type of delivery they have faced (or bowled). In essence, the bowling (or batting) features that may be associated with the batter's (or bowler's) strengths or weaknesses are investigated. Each delivery is represented as a set of bowling and batting features, similar to the set of items representing a transaction in association rule mining. Apriori algorithm of association rule mining is used to obtain the strength association rules and weakness association rules. Cricket text commentary data are obtained from the EspnCricInfo website and utilized for finding player's strength and weakness rules. Rules for more than 250 players are constructed by analyzing text commentaries over one million deliveries for 13 years (2006-2019). The data, codes, and results are shared at https://bit.ly/3rj6k6c.

Keywords: Sports text mining · Cricket analytics · Association rule mining.

1 Introduction

With the rapid increase in technology consumption, the amount of data generated and published has grown exponentially over the past few years. As a result, newer avenues have opened up where data mining techniques can prove indispensable to make sense of a large amount of data. We identify sports text commentary analysis as a field with enormous potential, leading to a better understanding of the game dynamics and characterization of various elements. We focus on the game of cricket to extract information from online archives of text commentary.

Cricket as a sport is renowned for recordkeeping. Cricket statistics have been widely analyzed for years. However, information carried by statistics is limited. It only captures the gameplay at the macroscopic scale and fails to capture the details. For instance, although statistics provide a perspective on how fluent

batters are (in terms of averages, runs, etc.), they fail to give an insight into how the batters played or performed. However, cricket text commentary is rich in description and contains a lot of information about the minute details of the gameplay. It captures the commentators' opinion about how the batter played, how the bowler bowled, and other auxiliary information.

In this work, we discuss the application of Association Rule Mining (ARM) in constructing rules that account for individual player's strengths and weaknesses from the cricket text commentaries. Specifically, we build rules that explain the strengths of a batter (or bowler) and rules that explain the weaknesses of a batter (or bowler). In addition, we make the following research contributions.

- We have collected a large and first-of-its-kind dataset of over one million deliveries, covering all international cricket Test matches for 13 years.
- We propose several domain-specific features to represent each delivery with fine-grained details.
- We provide the computationally feasible definition of strength and weakness rule and propose to use ARM to obtain these rules.

The paper is organized as follows. In Section 2, we introduce the game of cricket. In Section 3, we present the literature related to ARM. The methodology is briefed in Section 4. In Section 5, we describe the cricket text commentary data and associated challenges. Unigram and bigram modeling is presented in Section 6. Features extraction is discussed in Section 7. Strength and weakness rules construction is presented in Section 8. The work is finally concluded in Section 9.

2 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 Fig. 1. 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) *batter* who hits the ball to score runs, (ii) *bowler* who bowls the ball towards the batter, (iii) *fielder* who stops the ball hit by the batter 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. Batter 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 batter in different ways to get the batter 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

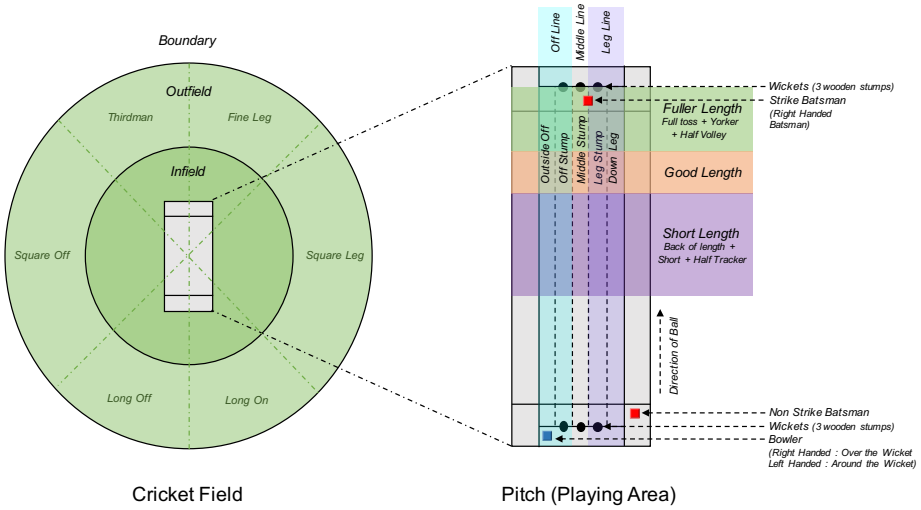


Fig. 1. A standard cricket field, showing the rectangular cricket pitch (white) at center, infield (medium green) inside the 30 yard circle, outfield (light green) from circle to boundary, and shot areas (thirdman, square off, long off, long on, square leg, and fine leg) for right handed batter/batsman.

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 batter). These batting and bowling attributes are listed in Table 1.

At any moment, two batters of a team are present on the pitch. One is called the striker batter who hits the ball, and another is the non-striker batter on the opposite side of the striker (or bowler's end). Each batter continues batting until she is out, which happens when the batter 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 batter's body except the hand holding the bat (leg before wicket or LBW) and a few other scenarios. A batter 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 the requisite number of times. A batter 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. Batters 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 batter 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 batter (or a bowler) can be left-handed or right-handed. In Figure 1, line of delivery and shot areas are shown for right-handed batters. For the left-handed batters, these notations are mirrored. Similarly, the 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 batters 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 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.

3 Literature Review

Association Rule Mining (ARM) is a data mining technique in which the extracted knowledge is in the form of association rules that describe a relationship between different attributes. Agrawal et al. [1] 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 $A \rightarrow B$, where A and B are disjoint sets of items, and B is likely to be purchased whenever A is purchased. ARM is widely used in many domains, such as health care [2], financial transactions [3], and retail [4], etc.

ARM is applied in the sports domain as well [5–7]. In cricket, Raj et al. [8] used ARM to find the association between the factors in cricket matches such as toss outcome and playing conditions with the outcome of the game. UmaMaheswari et al. [9] proposed to model an automated framework to identify find

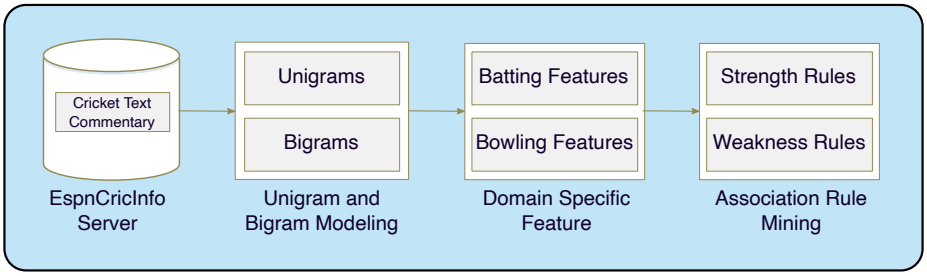


Fig. 2. Overview of our approach.

correlations among play patterns in cricket. ARM has not been applied to the sports text commentary data for detecting player-specific rules in the literature.

In our previous works, we have used the unstructured data, namely cricket text commentary, to visualize (i) cricket player's strengths and weaknesses [10, 12] and (ii) the temporal changes in player's strengths and weaknesses [11, 12].

4 Methodology

An association rule has the form $A \rightarrow B$, where A and B are disjoint sets of items, and the B set is likely to happen whenever the A set occurs. This paper analyzes the association of strength/weakness exhibited by a batter with the type of delivery she has faced. In essence, the paper investigates the bowling features that may be associated with the batter's strength/weakness.

There is no universally agreed-upon definition of strength and weakness as different players exhibit strength or weakness at varying instances of deliveries. When a batter or bowler exhibits a particular behavior repeatedly, it amounts to her strength or weakness. For example, a batter yielding wicket to deliveries pitched outside the off-stump consistently amounts to her weakness. *Strength* is when a batter or a bowler exhibits perfection for a particular delivery. Similarly, *weakness* is when a batter or a bowler exhibits imperfection for a particular delivery.

Please refer to Fig. 2 for an overview of our approach. We collect text commentary data from the web and perform extensive processing to extract useful information from this text. We propose several domain-specific discrete-valued features to represent each delivery and represent each player's batting (or bowling) features in this feature space. Finally, we identify the relationship between batting and bowling features of each player using ARM and construct the strength and weakness rules.

5 Data and Challenges

This section first describes the data and then highlights several challenges encountered when analyzing this data.

5.1 Data

Cricket has multiple formats of the game, of which we focus on the *Test cricket* format, which is considered as cricket's highest standard. In *Test cricket*, each team bats for two innings for five days. Every Test match generates a large amount of data, namely scorecard, video broadcast, and tracking data. Scorecards provide summary statistics and have been widely used for analyzing players' and teams' performance. However, it does not provide any specifics about the technique player has exhibited during the game's play. Video broadcast and tracking data have a detailed description of the play. However, these are not publicly available and are also expensive to process. In addition to the above forms of data, cricket matches generate text commentary pertaining to every match ball. This data is publicly available and is inexpensive to analyze. This data describes the ball-by-ball proceedings of the game with minute details. *Text commentary* corresponding to every delivery/ball of every match are acquired from the ESPNCricInfo[1] archive.

Consider an example of text commentary:

> 3.2, Finn to Sehwag, Four run, 136 kph, short of a length, but a little wide, enough for Sehwag to stand tall and punch it with open face, past Pietersen at point.

This commentary describes the second delivery in the fourth over of the game. Bowler Finn has bowled this delivery to the batter Sehwag. The outcome of the ball is four runs. The speed of the delivery is 136 kph (kilometer per hour). The rest of the text is unstructured and describes how the ball is delivered and how the batter played it. For instance, this commentary describes several features of bowling, such as length (short of a length) and line (a little wide). Similarly, it describes batting features such as response (punch it) pointing to the batter's strength.

Consider another example of text commentary given below where the technical word *outside edge* points to the batter's imperfection or weakness against *good length* and *angling in* delivery.

> 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.

We can analyze a large number of deliveries played by a batter. If we consistently observe good performance on similar deliveries, we can conclude that playing such deliveries is a batter's strength. Such detailed strength (/weakness) rules are far more expressive and valuable than simple statistics such as batting averages.

[1] `https://www.espncricinfo.com/`

To collect the text commentaries associated with a given Test match, one has to first obtain the season and series in which this particular match is a part. In addition, match and innings IDs and associated URLs need to be formulated from ESPNCricInfo's archive. This information is used to acquire the text commentaries for a given match. This procedure is repeated for all the matches played between May 2006 and April 2019. Total text commentaries of 1,088,570 deliveries are collected spanning thirteen years and stored in a local database. The collected deliveries account for a total of 550 international Test cricket matches. The acquired data are stored in a MySQL database and can be accessed at `https://bit.ly/3tOJHZ3`. The python code to obtain this data can be accessed at `https://bit.ly/3sjVD4W`.

5.2 Challenges

In order to construct strength and weakness rules from the text commentary data, following are the main challenges:

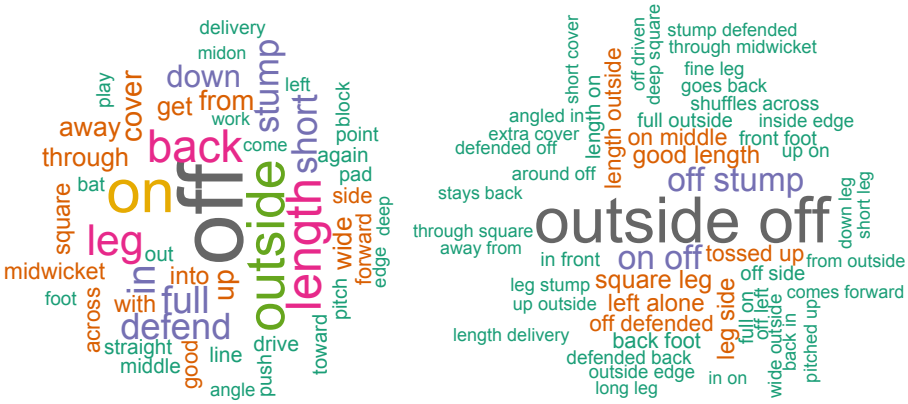
Data representation Every ball of text commentary comprises a maximum of 50 words and contains cricket technical vocabulary. The cricket technical vocabulary majorly overlaps with stop words in the conventional information retrieval application domain. This causes difficulty in adapting the off-shelf text data models in the present work. In addition, text data representation models suffer from sparsity problems (fewer words per document and a large vocabulary).

Rule definition Strength and weakness rules are highly subjective and often debatable. Even experts differ with the individual's opinions about strengths and weaknesses. Given that there is no universal agreement, defining what constitutes a strength rule and a weakness rule itself is challenging. Note that such a definition must be agreeable to every stakeholder.

Computational method Finding a suitable algorithm or computational method that constructs players' strength and weakness rules given the rule definition is another challenge.

6 Unigram and Bigram Modeling

In this section, we discuss the challenges and proposed steps for processing the text commentary data. Each 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 batter, 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 the batter's response in terms of her footwork and shot selection. In the end, some deliveries have a subjective opinion of the commentator about how the batter performed. An example of short text commentary structure is presented below.



(a) Unigram wordcloud. (b) Bigram wordcloud.

Fig. 3. Steve Smith's text commentary analysis.

Structured Text	Unstructured Text
75.3, Woakes to Smith, FOUR runs,	Back of a length, and a scudding pull skims over Root's head at mid-on!

Extraction of information from the structured part is a straightforward task. However, information extraction from unstructured part requires non-trivial efforts. 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 text commentary is that 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 batter's actions are described. Commentators, at times, focus either only on the bowler's action or only on the batter'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 rich jargon. Frequency analysis of the data reveals that commentators frequently use technical words. Figure 3 presents the unigram and bigram (two words occurring together) word clouds [13] of the

text commentaries where Steve Smith is mentioned as a batter. The importance of each unigram or bigram is shown with font size and color. 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. Our analysis reveals that domain-specific feature space captures sufficient information from the short sentences and identifies whether the information is for the bowler or the batter.

To capture the most relevant information, we have used a combination of web resources [14] 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:

> 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 batter played the ball)

In both of these examples, the word *swing* is used differently; first concerning the bowler and second concerning the batter. 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 batter. 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*,

Table 1. Batting and bowling features.

Category	**Feature** (Feature Description)
	Batting Features
Outcome	**0**, **1**, **2**, **3**, **4**, **5**, **6** runs, **out** (Outcome of a delivery - runs or wicket)
Response	**Attacked** (Batter plays aggressive shots or exhibits strength) **Defended** (Batter blocks or leaves the ball) **Beaten** (Batter plays poor shots or exhibits weakness)
Footwork	**Front foot** (Stance decision of a batter for full length deliveries) **Back foot** (Stance decision of a batter for short length deliveries)
Shot area	**Third man**, **Square off**, **Long off**, **Long on**, **Square leg**, and **Fine leg** (Region where the batter plays the shot)
	Bowling Features
Length	**Short** (Bowler pitches the ball closer to himself) **Full** (Bowler pitches the ball closer to the batter) **Good** (Bowler pitches the ball between full and short)
Line	**Off** (Ball travels on the off-stump line or outside the off-stump line) **Middle** (Ball travels on the middle-stump line) **Leg** (Ball travels on the leg-stump line or outside the leg-stump line)
Type	**Spin** (Bowler bowls slow deliveries which turn sharply after pitching) **Swing** (Bowler bowls fast deliveries which have movement in the air)
Speed	**Fast** (Speed of ball upon release: more than 100 kph) **Slow** (Speed of ball upon release: less than 100 kph)
Movement	**Move-in** (Ball moves towards the batter) **Move-away** (Ball moves away from the batter)

A glossary of cricket terms [2], and *The Wisden Dictionary of Cricket* [14]. The glossary and dictionary are the points of reference for cricket-specific unigrams and bigrams. Finally, we represent each text commentary as a set of unigrams and bigrams.

7 Feature Extraction

Unigram and bigram representations of text commentary cannot be directly used for strength and weakness rule extraction. We have identified a total of 19 batting features (batter facing the delivery is associated with these features) and 12 bowling features (bowler bowling the delivery is associated with these features) to represent each text commentary. Nineteen features are identified that characterize batting. These 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*

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

Table 2. Examples of identifying unigrams and bigrams for batting features.

Category	Batting Features	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. Examples of identifying unigrams and bigrams for bowling features.

Category	Bowling Features	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
Movement	Move In	move in, swing in, angle in
	Move Away	move away, swing away, angle away

off, long off, long on, square leg, and *fine leg*. We give a brief description of each of these features with their feature categories in Table 1. Twelve features are identified that characterize bowling. These 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 in Table 1.

All these features are discrete-valued. To transform each 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 are given in Table 2 and Table 3, respectively. The complete list can be accessed at `https://bit.ly/3sjVD4W`. 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.*

Finally, 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.

8 Mining Strength and Weakness Association Rules

In this section, we provide a computational definition of the strength/weakness rule and use ARM to construct strength/weakness rule of individual player given the definition.

Definition 1. *Rule. In the association rule $A \implies B$, when A comprises a set of bowling features and B comprises a batting feature.*

Definition 2. *Strength Rule of Batter. In Definition 1, when B or batting feature of the player (batter) corresponds to attacked.*

Definition 3. *Weakness Rule of Batter. In Definition 1, when B or batting feature of the player (batter) correspond to beaten.*

Whenever a batter exhibits strength on a delivery, it is a weakness for the bowler, and the inverse is also true. Therefore, the bowler's strength and weakness are defined in terms of the batters' batting features she is bowling. A bowler exhibits strength (or weakness) when the opponent batter's batting feature is beaten (or attacked).

Definition 4. *Strength Rule of Bowler. In Definition 1, when B or batting feature of the opponent players (batters) corresponds to beaten.*

Definition 5. *Weakness Rule of Bowler. In Definition 1, when B or batting feature of the opponent player (batters) corresponds to attacked.*

For constructing the strength and weakness association rules, we use the apriori algorithm [1]. 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. 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 player's strengths and weaknesses.

The results of the strength and weakness analysis for batter Steve Smith against all bowlers in Test matches are presented in Table. 4. The first strength

Table 4. Identified strength and weakness association rules for batter Steve Smith.

Association Rule (A $\Longrightarrow$ B)	**Support(%)**	**Confidence(%)**	**Lift**
Strength Rules			
{shortlength, slow} $\Longrightarrow$ {attacked}	2.6	72.1	1.6
{legstump} $\Longrightarrow$ {attacked}	2.5	60.1	1.3
{fast, middlestump} $\Longrightarrow$ {attacked}	2.7	53.9	1.2
{fulllength, middlestump} $\Longrightarrow$ {attacked}	2.0	51.4	1.1
Other Rules			
{goodlength} $\Longrightarrow$ {defended}	10.6	68.7	1.4
{fast, offstump} $\Longrightarrow$ {defended}	19.0	64.4	1.3
{fast, shortlength} $\Longrightarrow$ {backfoot}	9.2	91.4	1.9
{fulllength, offstump} $\Longrightarrow$ {frontfoot}	8.5	81.9	1.6
{fast, offstump} $\Longrightarrow$ {0run}	22.8	82.2	1.2
{fast, offstump} $\Longrightarrow$ {squareoff}	11.2	50.4	1.4
{legstump, slow} $\Longrightarrow$ {squareleg}	17.8	86.9	2.3
{legstump, movein, spin} $\Longrightarrow$ {fineleg}	0.01	100	23.7

rule of Steve Smith is - *Smith attacks slow and shot-length deliveries*. With our confidence threshold, no weakness rule is obtained for Steve Smith.

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 batter Steve Smith in Table. 4. 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://bit.ly/3rj6k6c`.

9 Conclusion

We presented an application of association rule mining for learning the strength and weakness rules of cricket players from the text commentary data. We provided computational definitions for capturing the strength and weakness rules. We fully utilized the ball-by-ball description of the game's proceedings during the Test matches. We established that association rule mining is a suitable method for the computation of strength and weakness rules. The constructed rules will be helpful for analysts, coaches, and team management in building game strategies.

The possible direction for future research is to include the external factors like playing conditions (age-of-ball, pitch condition, weather condition) and match situations (day of the match, inning of the match, session of the day) in the proposed method.

References

1. Agrawal, R., Imielinski, T., & Swami, A. (1993). Mining association rules between sets of items in large databases. SIGMOD '93.
2. Satou, K., Shibayama, G., Ono, T., Yamamura, Y., Furuichi, E., Kuhara, S., & Takagi, T. (1997). Finding association rules on heterogeneous genome data. Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing, 397-408.
3. Hsieh, N. (2004). An integrated data mining and behavioral scoring model for analyzing bank customers. Expert Syst. Appl., 27, 623-633.
4. Brijs, T., Goethals, B., Swinnen, G., Vanhoof, K., & Wets, G. (2000). A data mining framework for optimal product selection in retail supermarket data: the generalized PROFSET model. KDD '00.
5. Puchun, W. (2016). The Application of Data Mining Algorithm Based on Association Rules in the Analysis of Football Tactics. 2016 International Conference on Robots & Intelligent System (ICRIS), 418-421.
6. Liao, S., Chen, J., & Hsu, T. (2009). Ontology-based data mining approach implemented for sport marketing. Expert Syst. Appl., 36, 11045-11056.
7. Sun, J., Yu, W., & Zhao, H. (2010). Study of Association Rule Mining on Technical Action of Ball Games. 2010 International Conference on Measuring Technology and Mechatronics Automation, 3, 539-542.
8. Raj, K.A., & Padma, P. (2013). Application of Association Rule Mining: A case study on team India. 2013 International Conference on Computer Communication and Informatics, 1-6.
9. Umamaheswari, P., & RajaRam, M. (2009). A Novel Approach for Mining Association Rules on Sports Data using Principal Component Analysis: For Cricket match perspective. 2009 IEEE International Advance Computing Conference, 1074-1080.
10. Behera, S.R., Agrawal, P., Awekar, A., & Vedula, V.S. (2019). Mining Strengths and Weaknesses of Cricket Players Using Short Text Commentary. 2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA), 673-679.
11. Behera, S.R., & Vedula, V.S. (2020). Mining Temporal Changes in Strengths and Weaknesses of Cricket Players Using Tensor Decomposition. ESANN.
12. Behera, S.R., & Saradhi, V. (2020). Stats Arent Everything: Learning Strengths and Weaknesses of Cricket Players. Machine Learning and Data Mining for Sports Analytics. MLSA.
13. Steinbock, D. TagCrowd. http://www.tagcrowd.com/blog/about/ Accessed: 2020-11-19
14. Rundell, M. (2009) The Wisden Dictionary of Cricket (3rd ed.), A.& C. Black, 67.