ESANN 2020 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 2-4 October 2020, i6doc.com publ., ISBN 978-2-87587-074-2.
Available from http://www.i6doc.com/en/.

Mining Temporal Changes in Strengths and Weaknesses of Cricket Players Using Tensor Decomposition

Swarup Ranjan Behera and Vijaya V. Saradhi

Department of Computer Science and Engineering
Indian Institute of Technology Guwahati, India
b.swarup@iitg.ac.in and saradhi@iitg.ac.in

Abstract.
In this work, we present an application of tensor decomposition for discrete random variable tensor. In particular, we construct a tensor using cricket short text commentary data by employing domain-specific features. The aim is to understand the temporal changes in the strength rules and weakness rules of a player. Three-way correspondence analysis (TWCA) is employed to obtain the factors that show dependency between batting features, bowling features, and time respectively. Change in strength rules and weakness rules for Australian batsman Steve Smith (Test Rank #1 ICC player) are presented.

1 Introduction

Cricket is one of the most popular sports in the world with eighteen participating nations. Cricket is known for recording every detail of each match. A considerable amount of data in the form of scorecards (box score data), tracking data, video broadcasts, and coverage articles are generated in every match. Box score data has been used for obtaining a variety of statistical summaries and visualization based on these summaries are used by broadcasters, commentators, match organizers, and audience [1]. However, statistical summary capture only the game's play on a macroscopic level and do not attend to details.

Cricket commentary is a source of rich description about minute details of the game's proceedings. Commentators' opinion about how batsman played on every individual delivery is recorded in the commentary. Commentary is of two types. (i) Audio/Video - the audio/video commentary associated with the broadcast (radio/television). (ii) Text - the text commentary maintained by commercial websites such as EspnCricInfo[1]. The text commentary is introduced in 2006 and is easy to process and analyze compared to audio/video commentary as a reasonable structure is associated with the text commentary.

In the sports domain, box score data and tracking data are widely used for measuring player performance, team performance to convey information to the audience [2, 3]. Use of unstructured text data is limited in the sports domain due to lack of authenticity. In the context of cricket however, cricket text commentary data is an authentic one produced by the commentators who are part of organizing the match.

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

ESANN 2020 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 2-4 October 2020, i6doc.com publ., ISBN 978-2-87587-074-2.
Available from http://www.i6doc.com/en/.

In this work we, for the first time, make use of unstructured cricket text commentary for performing individual player specific analysis. In particular we propose to perform a *distinct* analysis namely *identifying changes in the individual player's strengths and weaknesses over a given time period.* The following are the main contributions of this work: (i) Propose to model the cricket text commentary data as a three-dimensional tensor in which batting features, bowling features, and time are captured. This tensor is termed as confrontation tensor. (ii) Provide a computationally feasible definition for obtaining strength rule and weakness rule. (iii) Use of the Tucker3 decomposition method to factor the confrontation tensor to obtain relationship between batting, bowling features, and time. (iv) Demonstrate the dependency of strength and weakness rules on time.

2 Modeling the Cricket Text Commentary

Every text commentary describes how a bowler bowled and how a batsman responded to the ball. In addition, the outcome of the delivery, and auxiliary information is present in every line of cricket text commentary. This section describes data collection, feature extraction, and building of confrontation tensor for every individual player. The confrontation tensors for every player is made available at `https://bit.ly/2OL0Ujh`.

2.1 Data Acquisition and Description

Short text commentaries related to Test matches (the longest format of the game which lasts up to five days) played between May 2006 to April 2019 are considered. A total of 1,088,570 short text commentaries are collected spanning thirteen years and 550 international Test matches. Consider an example of a short text commentary related to a delivery instance given 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.

In this example, *3* denotes the number of overs completed and *.2* represents the 2nd ball of 4th over is in progress, *Finn to Sehwag* is the identification of bowler and batsman, i.e., Finn is bowling to Sehwag, *FOUR* is the outcome of the ball, i.e., Sehwag scored 4 runs on that delivery, and rest of the text describes the ball and the way batsman played. When a batsman plays a delivery with confidence that is no technical flaw exhibited while playing the ball commentators explicitly include relevant information in the text commentary. For example, the technical word *punch* points to the batsman's perfection or strength against *short* and *wide* delivery. These details are specific only to the text commentary data and analysis of this data for a given player to mine such strength rules and weakness rules is of value.

For identifying the strengths and weaknesses, relevant features need to be extracted from the short text commentary. Features employed in traditional text mining literature like term frequency and inverse document frequency (TF-IDF) are not suitable due to the following reasons: (i) Each document's length (commentary for a particular delivery) is limited by fifty words, (ii) Technical

ESANN 2020 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 2-4 October 2020, i6doc.com publ., ISBN 978-2-87587-074-2. Available from http://www.i6doc.com/en/.

Table 1: Definition of the Batting and Bowling Features

Features	Description
Batting features	
0, 1, 2, 3, 4, 5, 6(also 6+) runs and **out**	Outcome of a particular delivery
Beaten (exhibits imperfection/weakness), **Defended** (blocks the ball), **Attacked** (plays aggressive shots/shows strength)	Response of the batsman on each delivery
Front foot (ball is played in front of the batsman), **Back foot** (played behind the batsman's wicket)	Footwork of a batsman when facing a delivery
Third man, Square off, Long off, Long on, Square leg, Fine leg	Region where shot is played by batsman (Shot area)
Bowling features	
Short (closer to the bowler), **Good** (optimal length, in between short and full), **Full** (nearer the batsman)	How far down the pitch the ball bounces (Length)
Off (on or outside off-stump), **Middle** (on middle-stump), **Leg** (on or outside leg-stump)	How far to the left or right of the wicket ball is travelling (Line)
Spin (slow deliveries which turn sharply after pitching), **Swing** (fast deliveries with movement in the air)	Nature of the delivery
Fast (medium: 60-80 mph, fast: 80+ mph), **Slow** (40-60 mph)	Speed of the ball after it is released
Move-in (towards batsman), **Move-away** (away from batsman)	Movement of the ball

words of the cricket game fall in the category of stop words in conventional information retrieval literature. Hence there is a need to identify domain-specific features related to batsman and bowler.

2.2 Feature Vectors

Three high-level features are identified which potentially describes a given text commentary related to a delivery. These are: batting features, bowling features, and time. Nineteen features are identified that characterize batting. Twelve features are identified that characterize bowling. Table 1 provides the definitions of batting and bowling features. Each of these feature is defined in terms of a set of unigram and bigram words present in the commentary data. The beaten feature consists of a set of unigram and bigram words all of which refer that batsman got beaten on a specified delivery. A non-exhaustive set for beaten feature is: {miss, beat, edge, confuse, deceive, poor shot, doesnt time, cant connect, knock down, bottom edge, lucky, misjudge}.

Time is another feature that has a significant impact on the way a batsman plays or bowler delivers a ball. Time granularity is identified in the increasing order as per session, per day, per innings, per match, per series, per year, or an entire career. To measure the changes in strength and weakness over the years, we have considered the time granularity as *per year*.

2.3 Confrontation Tensor

To perform the temporal analysis for a player against one or a set of players through a given time frame, one has to obtain a *subset of text commentary* from the local database. We considered a subset of the text commentary in which batsman Steve Smith has played against all the opponent players (bowlers) between the years 2013 and 2018, both inclusive. The considered time period is

ESANN 2020 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 2-4 October 2020, i6doc.com publ., ISBN 978-2-87587-074-2. Available from http://www.i6doc.com/en/.

six years. The confrontation tensor is of size ($19 \times 12 \times 6$) in which rows correspond to batting features of the player, columns correspond to bowling features of opponent players, and tubes correspond to the years in which the player has played. Every element in this tensor corresponds to how and when the batsman confronted with the bowlers. For example, how many numbers of times batsman has attacked short length deliveries in a given year? Similarly, other entries in the confrontation tensor represent the count of co-occurrences of batting features, bowling features in a given year. Note that the confrontation tensor is formed using 31 ($19 + 12$) distinct discrete random variables.

3 Proposed Method

The objective of the present work 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 confrontation tensor. The tensor factorization is performed using a method known as Three-Way Correspondence Analysis (TWCA) [4, 5], which uses Tucker3 [6] decomposition on the transformed confrontation tensor. TWCA tests the independence of events, namely row variables, column variables, and tube variables. If these events are not independent, then the equality does not hold; this points to the relationship between the three variables.

Let $\mathcal{N}$ be a three dimensional confrontation tensor 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 batsman attacking, e_2 be bowler bowling off line ball, and e_3 be in the year 2015. When these three events are *independent* then the following equation should hold: $P(e_1 \cap e_2 \cap e_3) = P(e_1) \times P(e_2) \times P(e_3)$. That is $p_{ijk} = p_{i..} \times p_{.j.} \times p_{..k}$; where $p_{i..} = \sum_{j=1}^{J} \sum_{k=1}^{K} p_{ijk}$ denote 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 independence gets deviated, the model is re-written as: $p_{ijk} = \boldsymbol{a}_{ijk} \times p_{i..} \times p_{.j.} \times p_{..k}$; where $\boldsymbol{a}_{ijk}$ denotes the amount of deviation. If $\boldsymbol{a}_{ijk} = 1$ then row event i, column event j, and tube event k are independent. When row features have certain relation with respect to column features and tube features, $\boldsymbol{a}_{ijk}$ takes value less than 1. For every row event, for every column event, and for every tube event, ijk^{th} entry of the $\mathcal{A}$ tensor is given by:

$$\boldsymbol{a}_{ijk} = \frac{p_{ijk}}{p_{i..} \times p_{.j.} \times p_{..k}} \tag{1}$$

Equation 1 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

ESANN 2020 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 2-4 October 2020, i6doc.com publ., ISBN 978-2-87587-074-2.
Available from http://www.i6doc.com/en/.

from the three way independence model, i.e.,

$$\boldsymbol{a}_{ijk} = \frac{p_{ijk} - p_{i..}p_{.j.}p_{..k}}{p_{i..}p_{.j.}p_{..k}}. \tag{2}$$

Equation 2 provide the computational definition for the strength rule or weakness rule of individual player. This rule must contain one batting feature, one bowling feature, and one time feature.

To obtain a low dimensional subspace that contains the batting features, bowling features, and time, $\mathcal{A}$ is decomposed using Tucker3 [6] decomposition method to obtain four factors namely $\mathcal{G}$ (the core tensor), A (retains batting features), B (retains bowling features), and C (retains time feature).

To obtain the association between the variables (batting, bowling, and time), two variables among the three are coded, i.e. column-tube categories are the coded bowling and time features. It resulted 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 1. F retains the batting features and H retains the bowling-time features.

Algorithm 1 The Proposed Method

Require: A three dimensional confrontation tensor $\mathcal{N}_{I\times J\times K}$

1: Tensor sum: $n = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{j=1}^{K}\mathcal{N}_{ijk}$
2: Tensor of relative frequencies: $\mathcal{P} = \frac{1}{n}\mathcal{N}$
3: 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 $p_{..k} = \sum_{i=1}^{I}\sum_{j=1}^{J} p_{ijk}$
4: Deviations from the three way independence: $a_{ijk} = \frac{p_{ijk}-p_{i..}p_{.j.}p_{..k}}{p_{i..}p_{.j.}p_{..k}}$
5: Tucker3 decomposition: $Tucker3\left(\mathcal{A}\right) = A_{I\times P}G_{P\times Q\times R}\left(B^{T}_{J\times Q}\otimes C^{T}_{K\times R}\right)$
6: Principal coordinates of rows: $F = AG_{(P\times QR)}$
7: Principal coordinates of column-tubes: $H = \left(B\otimes C\right)G_{(QR\times P)}$
8: **return** F and H

3.1 Strength and Weakness Interpretation

The inner product of the principal components F and H enables us to reconstruct the original three-way confrontation tensor and allows for a numerical assessment of the three-way association. A higher value of the inner product between a batting feature and a coded bowling-time feature indicates a high strength of association, while a lower value indicates a relatively low strength of association. A batsman exhibits strength when he attacks a delivery. In the strength rule for a batsman, *attacked* batting feature must be present. The other two features from bowling and time are identified using the inner product between $F_{attacked}$ and H matrices obtained above. We state the strength rule as (*attacked*, i) which yield maximum inner product value $\langle F_{attacked}, H_i\rangle$. Note that i contains the coded bowling and time features. Similarly, a batsman exhibits weakness when

ESANN 2020 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 2-4 October 2020, i6doc.com publ., ISBN 978-2-87587-074-2. Available from http://www.i6doc.com/en/.

he gets beaten, i.e., the highest value of $\langle F_{beaten}, H_i \rangle$ is considered as batsman's weakness. Whenever a batsman exhibits strength on a delivery, it is a weakness for the bowler, and the inverse is also true. Thus, the strength/weakness of a bowler can be defined in terms of the batting features of batsman he is bowling.

To visualize the year wise changes, we plot the inner product values in a line plot. Fig. 1 show's batsman Steve Smith's change in the strength rule (blue colored line) namely *attack* strategy on *full length* deliveries. He has shown increase in trend of attacking full length deliveries between the years 2013 and 2015 (both inclusive). However, in the 2016 year he struggled on full length deliveries. He once again shown strength on the full length deliveries in 2017 and subsequent year. A similar analysis can be performed on the weakness rule (red colored line) of *beaten* on *full length* deliveries for this player. We have shared the data, code, and results for 264 batsmen and 264 bowlers at `https://bit.ly/2OL0Ujh`.

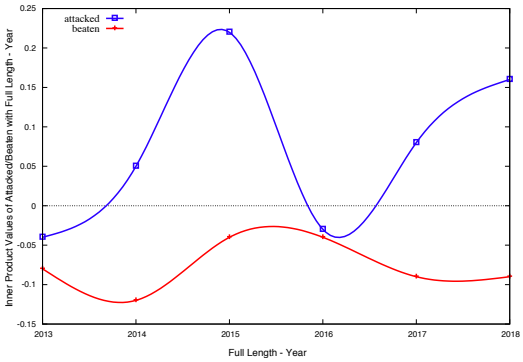

Fig. 1: Smith's strength and weakness on full-length deliveries over the years.

4 Summary

In this work, we have shown the usefulness of discrete random variable tensor decomposition for the temporal analysis of strengths and weakness of cricket players. In particular, we employ three way correspondence analysis (TWCA) which in turn uses tucker3 decomposition on a transformed tensor to obtain the relationship between batting, bowling, and time features. The extracted rules are interpretable and are of value to coaches and team management.

References

[1] J. Albert, M. E. Glickman, T. B. Swartz, and R. H. Koning, "Handbook of Statistical Methods and Analyses in Sports", *Chapman & Hall / CRC Handbooks of Modern Statistical Methods*, CRC Press, Taylor & Francis, 2016.

[2] C. Perin, R. Vuillemot, C. Stolper, J. Stasko, J. Wood, "State of the Art of Sports Data Visualization" *Computer Graphics Forum, Wiley*, vol. 37, no. 3, pages 1-24, 2018.

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

[4] A. Carlier and P.M. Kroonenberg, "Decompositions and biplots in three-way correspondence analysis", *Psychometrika*, vol. 66, pages 355-373, 1996.

[5] E. J. Beh and R. Lombardo, "Correspondence Analysis: Theory, Practice and New Strategies", *Wiley Series in Probability and Statistics*, Wiley, 2014.

[6] L. R. Tucker, "Some Mathematical Notes on Three-Mode Factor Analysis", *Psychometrika*, vol. 30, pages 279-311, 1966.