arXiv:2010.15377v1 [cs.LG] 29 Oct 2020

Supervised sequential pattern mining of event sequences in sport to identify important patterns of play: an application to rugby union

Rory Bunker[1*], Keisuke Fujii[1,2], Hiroyuki Hanada[1], Ichiro Takeuchi[1,3]

1 RIKEN Center for Advanced Intelligence Project, Tokyo, Japan
2 Graduate School of Informatics, Nagoya University, Nagoya, Aichi, Japan
3 Department of Computer Science, Nagoya Institute of Technology, Nagoya, Aichi, Japan

* Corresponding author
Email: rorybunker@gmail.com

Abstract

Given a set of sequences comprised of time-ordered events, sequential pattern mining is useful to identify frequent sub-sequences from different sequences or within the same sequence. However, in sport, these techniques cannot determine the importance of particular patterns of play to good or bad outcomes, which is often of greater interest to coaches. In this study, we apply a supervised sequential pattern mining algorithm called safe pattern pruning (SPP) to 490 labelled event sequences representing passages of play from one rugby team's matches from the 2018 Japan Top League, and then evaluate the importance of the obtained sub-sequences to points-scoring outcomes. Linebreaks, successful lineouts, regained kicks in play, repeated phase-breakdown play, and failed opposition exit plays were identified as important patterns of play for the team scoring. When sequences were labelled with points scoring outcomes for the opposition teams, opposition team linebreaks, errors made by the team, opposition team lineouts, and repeated phase-breakdown play by the opposition team were identified as important patterns of play for the opposition team scoring. By virtue of its supervised nature and pruning properties, SPP obtained a greater variety of generally more sophisticated patterns than the well-known unsupervised PrefixSpan algorithm.

Introduction

Large amounts of data are now being captured in sport as a result of the increased use of GPS tracking and video analysis systems, as well as enhancements in computing power and storage, and there is great interest in making use of this data for performance analysis purposes. A wide variety of methods have been used in the analysis of sports data, ranging from statistical methods to, more recently, machine learning and data mining techniques.

Among the various analytical frameworks available in sports analytics, we adopt in this paper an approach to extract events from sports matches and analyse sequences of events. The most basic events-based approach is based on the analysis of the *frequencies* of events. These frequencies can be used as performance indicators (PIs) by comparing the frequency of each event in positive outcomes (victory, points scored, etc.) and negative outcomes (loss, goals conceded, etc.) in order to investigate which events are related to positive and negative outcomes. However, frequency-based analyses have drawbacks in that the information contained in the order of events cannot be exploited.

In this study, we consider a sequence of events, and refer to a partial sequence of events a *sequential pattern* or simply a *pattern (of play)*. In sports, the occurrence of certain events in a particular order often has a strong influence on the outcome, so it is useful to use patterns as a basic analytical unit. The computational framework for finding patterns from sequential data that have specific characteristics is known as *sequential mining* in the field of data mining.

The most basic problem setup in sequential mining is to enumerate frequent patterns, which is called *frequent sequential mining*. If a particular series of events occur frequently and in a particular order, the pattern is likely to contain important information for coaches and/or performance analysts. Although the total number of patterns (i.e., the number of ordered sequences of all possible events) is huge, it is possible to efficiently enumerate patterns that appear more than a certain frequency by making effective use of branch-and-bound techniques. Frequent sequential mining is categorised as an *unsupervised learning technique* in the terminology of machine learning.

When applying frequent sequential mining to data from sport, there are several options. The first option is to simply extract the frequent patterns from the entire dataset. The drawback of this approach is that it is not possible to distinguish whether a pattern leads to good or bad results. The second option is to split the dataset into a good result dataset and a bad result dataset, and perform frequent sequential mining on each dataset. The third option is to perform frequent sequential mining on the entire dataset to identify frequent patterns and then create a machine learning model that uses the patterns as features to predict whether the results are good or bad. The disadvantage of the second and third options is that the process of pattern extraction and the process of relating the patterns to the "goodness" of the results are conducted separately.

Unlike unsupervised mining, a mining method that directly extracts patterns that are associated with good or bad outcomes is called *supervised mining*. Roughly speaking, by using supervised mining, we can directly find patterns that have different frequencies depending on the results, and thus we can find more direct effects on the results than by simply combining unsupervised mining as described above.

If an event sequence represents a portion of a match, i.e., *a passage of play*, the sequence label outcome may indicate a success or failure of some sort, e.g., points being scored at the end of the passage of play. Alternatively, one event sequence could represent an entire match, and the sequence could be labelled with a win/loss outcome. We refer to an event sub-sequence that is relevant to either successful or unsuccessful outcomes an *important* sub-sequence. While unsupervised techniques can identify frequent patterns, the identified patterns may not necessarily be important. In order identify important sub-sequences (patterns of play), we take a supervised sequential pattern mining approach where each sequence represents a portion of a match (passage of play) that is labelled with a points scoring outcome.

In the present study, we apply supervised sequential pattern mining to data consisting of event sequences from rugby. In particular, we make use of safe pattern pruning (SPP) [**?**, 1], which combines a convex optimisation technique called safe screening [3] with sequential pattern mining. Safe screening pre-screens a subset of variables for whether they are required for the optimal model before solving the optimisation problem, which can then be solved more efficiently. This is an important property when the number of potential sub-sequences is large. SPP has been applied to datasets consisting of animal trajectories [**?**]; however, compared with animal trajectories, sports data often contains a greater diversity of events. As a basis for comparison, we also compare the SPP-obtained sub-sequences with those obtained with the well-known unsupervised PrefixSpan algorithm [4] when it is applied to subsets of the original labelled data. To demonstrate its effectiveness, we apply the approach to data consisting of event sequences from one professional rugby union team's matches in the 2018 Japan Top League.

The main contributions of this work are as follows. First, we aim to highlight the potential utility of *supervised sequential pattern mining* as a flexible analytical framework for performance analysis in sport to identify patterns of play that are of importance to particular positive and/or negative outcomes. In addition, we introduce the use of sequential pattern mining as an analytical tool for performance analysis in rugby specifically.

Related Work

We briefly review studies that have applied sequential pattern mining techniques to sport, as well as rugby-related studies that have considered PIs, analysis of sequences or that have applied advanced analytical methods.

Sequential pattern mining [5] has been applied in a number of different application domains [6].

Unsupervised sequential pattern mining techniques have been applied to data from sport, e.g., in judo [7], cycling training [8], and in combination with clustering in football [9]. [10] analysed event sequences from a single match in the 2012 European Championship between Italy and Germany, with sequences defined to begin when a team took possession and to end when the team lost possession. The field was divided into 30 zones, and the passing patterns based on zone-to-zone passing that were effective for creating goal-scoring opportunities for Germany were identified (the definition of effectiveness for goal scoring, e.g., whether this was creating a shot on goal or a shot on target, was not particularly clear from the study). T-patterns [11], which share similarities with sequential pattern mining, have been applied to sports including soccer and basketball [12–16] as well as judo [17].

The importance of frequency-based performance indicators [18] (PIs) has been studied in rugby through the use of statistical techniques such as the Wilcoxon signed rank test and discriminant analysis [19–24], and more recently, with random forests from machine learning [25]. Rugby has also been analysed at the more granular sequence level by analysing the duration of sequences. The duration of the sequences of plays leading to tries at the 1995 Rugby World Cup (RWC) were studied by [26]. In a study of the 2003 RWC, [27] found that teams that were able to create movements that lasted longer than 80 seconds were more successful.

Given that rugby is a complex, dynamic and interactive sport, it has been noted that the analysis of frequencies alone (e.g., through PIs) cannot be expected to solve difficult multivariate problems [28]. The necessity for more advanced methods in rugby union was also highlighted by [24] and [29], who suggested that methods from dynamical systems, machine learning, social network analysis, interpersonal distance and group behaviour, which have been applied in basketball and soccer, remain largely unused in rugby. Machine learning models have been used for the prediction of results in rugby [25, 30–32], while [33] and [34] used Self-Organising Maps [35] to identify important PIs and effective playing styles in New Zealand provincial rugby. Network centrality has been applied to identify tactical and leadership structures and to improve the description of complex passages of play at the 2015 RWC [36], while [37] applied K-modes cluster analysis to identify particular patterns of play that led to tries in the 2018 Super Rugby season. Recently, [38] used convolutional and recurrent neural networks to predict the outcomes (territory gain, retaining possession, scoring a try, and conceding/being awarded a penalty) of sequences of play, based on event order and their on-field locations.

Materials and Methods

Data

XML data, derived from video tagged in SportsCode by the performance analyst of one of the Top League teams (not named for reasons of confidentiality), was obtained and subsequently converted into labelled event sequences. Written consent was obtained to use the data for research purposes. All of the team's matches in the 2018 Japan Top League season were considered, consisting of 490 event sequences in total. As mentioned, we take an event sequence $\boldsymbol{g}_i$ to represent a passage of play, i.e., a portion of a match. In particular, a passage of play was defined to start with either a kick restart, scrum, or lineout (note that it is possible to define a passage of play in other ways, e.g., when possession changes hands. However, the granularity of our data was such that this definition made more sense). Then, when one of these three events occur again (except for a scrum reset where a scrum follows another scrum), this event becomes the first event in a new event sequence; otherwise, if a try is scored or a kick at goal occurs, a new passage of play also begins. In the original dataset, there were 12 unique events for the team and the opposition teams, i.e., $m = 24$ in total. The XML data also contained a more granular level of data than these events represent (i.e., with more detailed events—in other words, a larger number of events); however, in order to reduce computational complexity, the higher level of the data was considered.

Methods

Safe Pattern Pruning

SPP derives its name since by invoking a tree structure that is defined among all possible patterns in a database, the algorithm prunes the tree in such a way that if a node corresponding to a particular pattern is pruned, it is guaranteed that all patterns corresponding to its descendant nodes are not required for the predictive model.

The number of unique event symbols are denoted as m and the set of those event symbols is denoted as $\mathcal{S} := \{s_1, \dots, s_m\}$. Let n denote the number of sequences in the database. Sequences with the labels 1 and -1 are denoted as $\mathcal{G}_+, \mathcal{G}_- \subseteq [n]$ and are of size $n_+ := |\mathcal{G}_+|, n_- := |\mathcal{G}_-|$, respectively. In this study, sequences are taken to represent passages of play (portions of a match), and we use points scoring outcome (whether a try was scored or kick at goal was attempted, or not) as the sequence labels. The training dataset for learning the SPP classifier is

$$\{(\boldsymbol{g}_i, y_i)\}_{i \in [n]},$$

where $\boldsymbol{g}_i$ represents the i-th sequence/passage of play. Each sequence $\boldsymbol{g}_i$ takes a label from $y_i \in \{\pm 1\}$. and can be written as

$$\boldsymbol{g}_i := \langle g_{i1}, g_{i2}, \dots, g_{iT(i)} \rangle, i \in [n],$$

where g_{it} is the t-th symbol of the i-th sequence, which takes one of the event symbols in $\mathcal{S}$, and $T(i)$ indicates the length of the i-th sequence, i.e., the number of events in this particular sequence. We refer to a contiguous part of a sequence as a *sub-sequence* or *pattern of play* (although SPP can be extended to non-contiguous patterns, the version of the algorithm used in the present study considers contiguous patterns). Patterns of play are denoted as $\boldsymbol{q}_1, \boldsymbol{q}_2, \ldots$, each of which is also a sequence of event symbols:

$$\boldsymbol{q}_j := \langle q_{j1}, q_{j2}, \ldots, q_{jL(j)} \rangle, j = 1, 2, \ldots,$$

where $L(j)$ is the length of pattern $\boldsymbol{q}_j$ for $j = 1, 2, \ldots$. A sequence $\boldsymbol{g}_i$ contains a sub-sequence $\boldsymbol{q}_j$ if

$$\exists \{1 \leq i_1 < \ldots < i_{L(j)} \leq T(i)\} \text{ such that } q_{j1} = g_{i_1}, q_{j2} = g_{i_2}, \ldots, q_{jL(j)} = g_{i_{L(j)}},$$

This relationship is represented as $\boldsymbol{q}_j \sqsubseteq \boldsymbol{g}_i$. The set of all possible patterns contained in any sequence $\{\boldsymbol{g}_i\}_{i \in [n]}$ is denoted as $\mathcal{Q} := \{\boldsymbol{q}_i\}_{i \in [d]}$, where d is the number of possible patterns. Note that, due to the permutations in terms of the number of potential patterns of play, the size of $\mathcal{Q}$ is quite large in general. This is where the aforementioned pruning mechanism comes in useful.

A classifier based on a sparse linear combinations of patterns can be written as

$$f(\boldsymbol{g}_i; \mathcal{Q}) := \sum_{\boldsymbol{q}_j \in \mathcal{Q}} w_j I(\boldsymbol{q}_j \sqsubseteq \boldsymbol{g}_i) + b, \tag{1}$$

where I is an indicator function that takes the value 1 if sequence g_i contains sub-sequence g_i and 0 other otherwise; and $w_j \in \mathbb{R}$ and $b \in \mathbb{R}$ are parameters of the linear model, which are estimated by solving the following minimisation problem

$$\min_{\boldsymbol{w}, b} \sum_{i \in [n]} \ell(y_i, f(\boldsymbol{g}_i; \mathcal{Q})) + \lambda \|\boldsymbol{w}\|_1, \tag{2}$$

where $\boldsymbol{w} := [w_1, \ldots, w_d]^\top$ is a vector of weights, $\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ is a loss function and $\lambda > 0$ is a tuning parameter that can be tuned by cross-validation. The minimisation problem (1) was, in the present study, solved with an L1-regularised L2-Support Vector Machine. The magnitudes of the resulting weights w_j obtained from solving the minimisation problem can be interpreted as the contribution of a particular pattern of play (sub-sequence) to a particular labelled outcome.

Procedures

The original dataset was then divided into two datasets. In the first, which we call the scoring dataset, we consider the case where the sequences are from the team's scoring perspective. In this dataset, the label $y_i = +1$ represents points being scored or attempted. Note that while a try scored was certain in terms of points being scored, the kick at goal represents either a penalty goal, drop goal or conversion being attempted but not necessarily resulting in points. In our data, only the kick at goal being attempted (event id 6) was available—not whether the goal was actually successful or not. However, since it is more important to be able to identify points-scoring opportunities than whether or not the kick was actually successful, we assume 100% of kicks at goal resulted in points. The label $y_i = +1$ represents points being scored or attempted represents a try being scored or a kick at goal occurring by the team at the end of passage i, while $y_i = -1$ represents no points being scored or attempted by the team at the end of passage i. In the second, which we call the conceding dataset, we consider the case where the sequences are from the team's conceding perspective, or equivalently, from the opposition teams' scoring perspective. In this dataset, $y_i = +1$ represents points being scored, i.e., a try being scored or a kick at goal being attempted, by an *opposition* team at the end of passage i, while $y_i = -1$ represents no points being scored or attempted by an opposition team at the end of passage i. The list of events for the three datasets are presented in Table 1. In the scoring dataset, the events that relate to the team scoring—Try scored (event ID = 11) and Kick at goal (event ID = 6)—were removed from the event sequences, since the sequence label identified whether the team scored or not. Similarly, in the conceding dataset, the events that relate to the opposition team scoring—O-Try Scored (event ID = 23) and O-Kick at goal (event ID = 18)—were removed from the event sequences, since the sequence label identified whether the opposition team scored or not.

Table 1. Unique event lists for the original, scoring and conceding datasets.

event ID	original	scoring	conceding
1	Restart Receptions	Restart Receptions	Restart Receptions
2	Phase	Phase	Phase
3	Breakdown	Breakdown	Breakdown
4	Kick in Play	Kick in Play	Kick in Play
5	Penalty Conceded	Penalty Conceded	Penalty Conceded
6	Kick at Goal		Kick at Goal
7	Quick Tap	Quick Tap	Quick Tap
8	Lineout	Lineout	Lineout
9	Error	Error	Error
10	Scrum	Scrum	Scrum
11	Try Scored		Try Scored
12	Line Breaks	Line Breaks	Line Breaks
13	O-Restart Receptions	O-Restart Receptions	O-Restart Receptions
14	O-Phase	O-Phase	O-Phase
15	O-Breakdown	O-Breakdown	O-Breakdown
16	O-Kick in Play	O-Kick in Play	O-Kick in Play
17	O-Penalty Conceded	O-Penalty Conceded	O-Penalty Conceded
18	O-Kick at Goal	O-Kick at Goal	
19	O-Quick Tap	O-Quick Tap	O-Quick Tap
20	O-Lineout	O-Lineout	O-Lineout
21	O-Error	O-Error	O-Error
22	O-Scrum	O-Scrum	O-Scrum
23	O-Try Scored	O-Try Scored	
24	O-Line Breaks	O-Line Breaks	O-Line Breaks
label	-	Points Scored	O-Points Scored
	n=490	n_+=86, n_-=404	n_+=44, n_-=446

From the team's scoring perspective, the team's points scored (event 6 and 11) determine the class label for the sequences, while opposition points scored (event 18 and 23) are treated as events in the sequences. From the team's conceding perspective, opposition points scored (events 18 and 23) determine the class label for the sequences, while opposition points scored (events 6 and 11) are treated as events in the sequences. Events prefixed by "O-" indicate that the events relate to the opposition team; those that are not are events that relate to the team. Recall that the scoring dataset label Points Scored = 1 if Try Scored and/or Kick at Goal are in the sequence, 0 otherwise. Similarly, the conceding dataset label O-Points Scored = 1 if O-Try Scored and/or O-Kick at Goal are in the sequence, 0 otherwise.

The SPP algorithm (software is available at `https://github.com/takeuchi-lab/SafePatternPruning`) was applied to the scoring and conceding datasets, with 10-times 10-fold cross-validation used to tune the parameter λ. The maximum possible length of a sub-sequence to search for was set to be 20. In pattern mining, one refers to the "support" of a pattern as the frequency with which it occurs. The obtained sub-sequences were then filtered to those that had a minimum support of five, and then the Top five sub-sequences with the largest weight contributions w_j for both the scoring and conceding datasets were recorded. The choices of the largest five weights and the minimum support of five are somewhat arbitrary; the intention was to identify patterns that are not only frequent but also *important*.

As a basis for comparison, we compare the results from SPP with those obtained by PrefixSpan, an unsupervised (i.e., unlabelled) algorithm. The PrefixSpan algorithm begins by exploring the search space of sequential patterns based on a depth-first search, and then, starting from the sequential patterns that contain just a single event, discovers longer patterns by recursively appending events to the existing patterns [4]. The prefixspan package in Python (`https://pypi.org/project/prefixspan/`) was utilised. As well as being a widely used sequential pattern mining algorithm, PrefixSpan was selected as a basis for comparison with SPP because it is itself a component of SPP. In the unsupervised setting, while support values (frequencies) of the patterns of play can be obtained, since the data is unlabelled, we cannot obtain weights for the sub-sequences. However, for a more realistic comparison, we assume prior knowledge of the sequence labels, and thus apply PrefixSpan only to the sequences in the scoring and conceding datasets that have the label y_i=1. PrefixSpan was applied to the dataset, which we call scoring+1, containing the sequences where the team actually scored, and to the conceding+1 dataset, containing the sequences where the team actually conceded points (i.e., the opposition team scored points).

The procedures to create the datasets and labels, and apply and compare the algorithms, are illustrated in Fig 1.

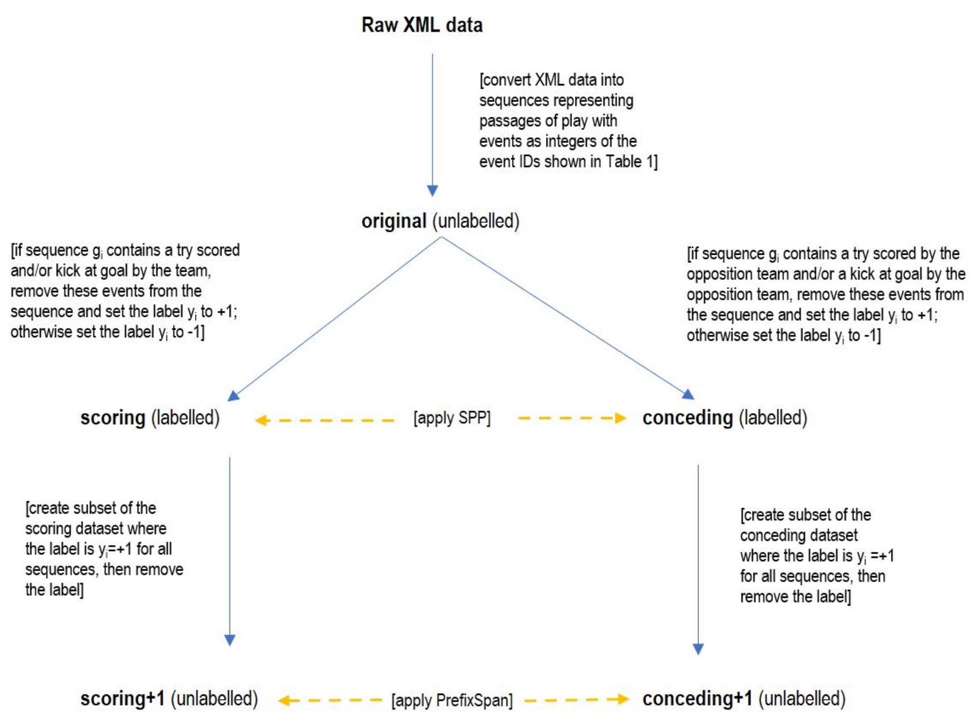

Fig 1. Dataset creation. Illustration of the procedures to create the datasets and labels, and apply and compare the algorithms. Actions are in square brackets and datasets are in bold.

Results

In this section, we present our results obtained from applying SPP to both the scoring and conceding datasets. In addition, as a basis for comparison, we present the results from applying the unsupervised PrefixSpan algorithm [4] to positively-labelled subsets of the scoring and conceding datasets (sub-section).

Analysis of sequence lengths

There were an average of 10.6 events in each sequence in the scoring dataset, and 10.8 events in the conceding dataset. The shortest sequence contained two events, and the longest contained 48 events (Table 2). The slight differences in mean sequence lengths between the scoring and conceding datasets is a result of the removal of the try and kick at goal events from the sequences in order to create the sequence outcome label. The distributions are positively skewed and non-normal (Fig 2), which was confirmed by Shapiro-Wilk tests. By comparing these distributions, we can see that the number of sequences in which points were scored was higher in the scoring dataset than the conceding dataset, which is reflective of the strength of the team in the 2018 season. From the team's scoring perspective, 86 out of the 490 passages of play (18%) resulted in points being scored by the team, while from the team's conceding perspective, 44 out of the 490 passages of play (9%) resulted in points conceded. The sequences in which the team scored points were slightly longer, containing 12.8 events on average compared to those where the team didn't score, which contained 10.2 events, on average. Sequences in which the team conceded points contained 11.2 events on average, while those where the team didn't concede points contained 10.8 events, on average.

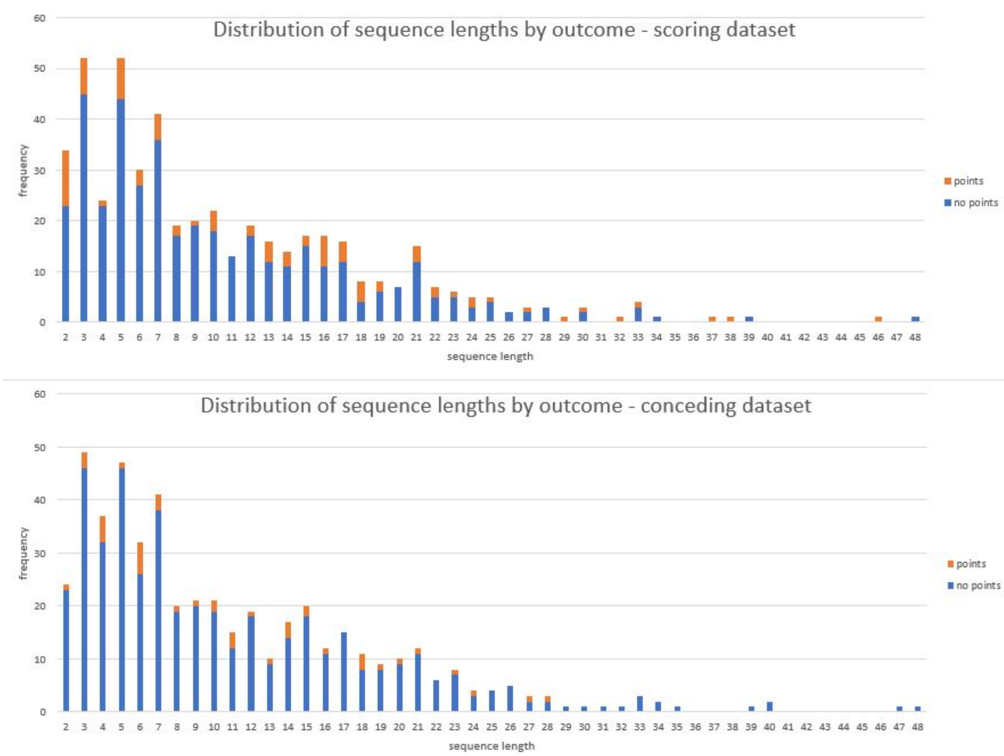

Fig 2. Sequence length distributions. Distribution of sequence lengths by points-scoring outcome for the scoring and conceding datasets. Sequence length is defined as the number of events in each sequence (excluding the outcome label).

Table 2. Unique event lists for the original, scoring and conceding datasets..

	scoring	conceding
Mean	10.6	10.8
Standard deviation	7.8	7.9
Minimum	2	2
25th percentile	5	5
Median	8	8
75th percentile	15	15
Maximum	48	48
Skewness	1.3	1.4

Identification of important patterns of play using SPP

SPP obtained 93 sub-sequences when applied to the scoring dataset, of which 75 had support of 5 or higher. Out of these 75 patterns of play (for which the weight values lie between -1 and 1), 38 had a positive weight (for which the weight values were greater than 0 but less than or equal to 1). In this study, we restricted our analysis to patterns of play that had the highest positive weights. For the scoring dataset, this means the patterns that had a positive contribution to the team scoring. For the conceding dataset, this means the patterns that had a positive contribution to opposition teams scoring. In other words, for the sake of brevity, we did not consider the patterns that had the highest contribution to "not scoring" and "not condeding". These 75 sub-sequences contained an average of 4.5 events, and an average of 5.4 events when considering only those with a positive weight (38 of the 75 patterns). The longest obtained sub-sequence contained 16 events.

Applying SPP to the conceding dataset resulted in a total of 135 sub-sequences, of which 51 had support of 5 or higher. These 51 sub-sequences contained an average of 3.8 events, and an average of 4.4 events when considering only those with a positive weight (31 of the 51 patterns). The longest obtained sub-sequence contained 15 events.

The five patterns with the highest weight contributions from the scoring dataset are listed in Table 3. Line breaks (event id 12) were found to be most associated with the team scoring, with a weight contribution of w=0.919. Line breaks, which involve breaking through an opposition line's defence, advance the attacking team forward and would thus be expected to create scoring opportunities. A lineout followed by phase play (8,2) was the second pattern of play most associated with the team scoring (w=0.808). The pattern of play (2,3,4,2,3) can be interpreted as a kick in play being made by the team and being re-gathered by the team (w=0.796), thus resulting in retained possession. The fourth sub-sequence represents four repeated phase-breakdown plays by the team, followed by them making a kick in play (w=0.732). The fifth pattern of play can be interpreted as the opposition team receiving a kick restart made by the team, attempting to exit their own territory via a kick but not finding touch, thus giving the ball back to the team (w=0.710).

The five patterns with the highest weight contributions from the conceding dataset are listed in Table 4. A linebreak (event ID 24) being made by the opposition team was most associated with them scoring against the team (w=0.613), although the weight magnitude was not as large as for the team scoring from a linebreak against the opposition team (w=0.919). The second sub-sequence (14,9,15) most associated with the team conceding points can be interpreted as the opposition team being in possession of the ball, the team making some form of error, and the opposition team regaining possession (w=0.392). The third pattern of play (20) most associated with the team

Table 3. Top five SPP-obtained patterns of play in terms of weight contribution to the team scoring points: scoring dataset.

event id pattern (q_j)	pattern description	support	λ	weight
12	linebreak	77	1	0.919
8 2	lineout, phase	71	22	0.808
2 3 4 2 3	phase, breakdown, kick in play, phase, breakdown	9	56	0.796
2 3 2 3 2 3 2 3 4	[phase, breakdown]x4, kick in play	9	53	0.732
13 14 15 14 15 16 14 2 3	O-restart received, [O-phase, O-breakdown]x2, O-kick in play, phase, breakdown	6	71	0.710

conceding points was an opposition team lineout (w=0.357). The fourth and fifth sub-sequences most associated with the opposition team scoring both represent repeated phase and breakdown play, with the fifth sub-sequence, for example, representing the opposition team making over six repeated phase/breakdown plays.

Table 4. Top five SPP-obtained patterns of play in terms of weight contribution to the team conceding points: conceding dataset.

event id pattern (q_j)	pattern description	support	λ	weight
24	O-Linebreak	32	15	0.613
14 9 15	O-phase, error, O-breakdown	10	49	0.392
20	O-lineout	86	18	0.357
15 15 14 15	O-breakdown, O-breakdown, O-phase, O-breakdown	5	43	0.339
15 14 15 14 15 14 15 14 15 14 15 14 15	[O-breakdown, O-phase]x6+	16	47	0.261

Comparison of SPP- to PrefixSpan-obtained patterns of play

Tables 5 and 6 show the top five sub-sequences in terms of their support from the scoring+1 and conceding+1 datasets.

Table 5. Top five PrefixSpan-obtained patterns of play in terms of support: scoring+1 dataset.

event id pattern (q_j)	pattern description	support
2	Phase	84
2 3	Phase, Breakdown	60
3	Breakdown	60
2 2	[Phase]x2	59
2 3 2	Phase, Breakdown, Phase	59

The obtained results showed that common events and sub-sequences were detected with PrefixSpan, i.e., breakdowns and phases. Repeated breakdown and phase play is a means retaining possession of the ball and building pressure. Longer repeated breakdown and phases plays were also identified by SPP. However, in the case of the PrefixSpan-obtained results, these sub-sequences are not particularly useful for coaches or analysts since they merely reflect common events rather than *important* events, i.e., those that are important for creation of points scoring opportunities. The supervised approach with SPP, by using sequences representing passages of play labelled with

Table 6. Top five PrefixSpan-obtained patterns of play in terms of support: conceding+1 dataset.

event id pattern (q_j)	pattern description	support
14	O-phase	39
14, 15	O-phase, O-breakdown	33
15	O-breakdown	33
14 14	[O-phase]x2	29
14 15 14	O-phase, O-breakdown, O-phase	29

points scoring outcomes, by virtue of the computed weights, is able to provide a measure of the importance of patterns of plays to these outcomes. In addition, compared to PrefixSpan, the supervised SPP algorithm obtained a greater variety of patterns of play, i.e., not only those containing breakdowns and or phases, and also discovered more sophisticated patterns of play.

Discussion

In this study, a supervised sequential pattern mining algorithm called safe pattern pruning (SPP) was applied to data from professional rugby union in Japan, consisting of sequences in the form of passages of play that are labelled with points scoring outcomes. The obtained results suggest that the SPP algorithm was useful when applied to sequences that represent passages of play (portions of a match) that are labelled with outcomes in terms of whether points were scored or not, in order to detect important sub-sequences (patterns of play) to scoring outcomes. SPP was able to identify relatively complex patterns of play from the original event sequences, the interpretation of which is potentially useful for coaches and performance analysts for own- and opposition-team analysis in order to identify vulnerabilities and tactical opportunities.

By considering both the scoring and conceding perspectives of the team, insight was able to be obtained that would be useful to both the team, as well as opposition teams that are due to play the team. For both the team and their opposition teams, linebreaks were found to be most associated with scoring. For both the team and their opposition teams, lineouts were found to be more beneficial to generate scoring opportunities than scrums. These results are consistent with [37], who found that lineouts followed by a driving maul are common approaches to scoring tries (albeit in a different competition, Super Rugby), and with [39], who found that around one-third of tries came from lineouts in the Japan Top League in 2003 to 2005—the highest of any try source. As well as creating lineouts or perhaps prioritising them over scrums, for opposition teams playing the team, effective strategies may include maintaining possession with repeated phase-breakdown play (over six repetitions), shutting down the team's ability to regain kicks, and making sure to find touch on exit plays from kick restarts made by the team.

As mentioned, compared to the unsupervised PrefixSpan algorithm, the supervised SPP algorithm obtained a greater variety as well as more sophisticated patterns of play. This is likely due to the supervised (i.e., labelled) nature of SPP as opposed to the unsupervised (unlabelled) nature of PrefixSpan, as well as the safe screening mechanism of SPP, which prunes out irrelevant sequential patterns in advance.

The approach highlighted the potential utility of supervised sequential pattern mining as an analytical framework for performance analysis in sport, and more specifically, the applicability of sequential pattern mining in rugby. Although the results obtained are encouraging, a limited amount of data from one sport was used. Also, spatial information such as field position was not available in the data, which may have improved the analysis. Furthermore, although SPP was useful for the specific dataset in this study, its usefulness is likely to be dependent on the structure of the input data and the specific definition of the sequences and labels. For instance, applying the approach to a dataset that consists of entire matches as sequences and win/loss outcomes as the labels is unlikely to produce interesting results since it is self-evident that sequences that contain more scoring events will be more associated with wins. As future work, it would be interesting to apply the approach to a larger amount of data from rugby, as well as to similarly structured datasets in other sports in order to confirm its efficacy.

References

1. Nakagawa K, Suzumura S, Karasuyama M, Tsuda K, Takeuchi I. Safe pattern pruning: An efficient approach for predictive pattern mining. In: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining; 2016. p. 1785–1794.

2. Sakuma T, Nishi K, Kishimoto K, Nakagawa K, Karasuyama M, Umezu Y, et al. Efficient learning algorithm for sparse subsequence pattern-based classification and applications to comparative animal trajectory data analysis. Advanced Robotics. 2019;33(3-4):134–152.

3. Ghaoui LE, Viallon V, Rabbani T. Safe feature elimination for the lasso and sparse supervised learning problems. arXiv preprint arXiv:10094219. 2010;.

4. Pei J, Han J, Mortazavi-Asl B, Wang J, Pinto H, Chen Q, et al. Mining sequential patterns by pattern-growth: The prefixspan approach. IEEE Transactions on knowledge and data engineering. 2004;16(11):1424–1440.

5. Agrawal R, Srikant R. Mining sequential patterns. In: Proceedings of the eleventh international conference on data engineering. IEEE; 1995. p. 3–14.

6. Fournier-Viger P, Lin JCW, Kiran RU, Koh YS, Thomas R. A survey of sequential pattern mining. Data Science and Pattern Recognition. 2017;1(1):54–77.

7. La Puma I, de Castro Giorno FA. Ontology-Based Data Mining Approach for Judo Technical Tactical Analysis. In: The Third International Conference on Computing Technology and Information Management (ICCTIM2017); 2017. p. 90.

8. Hrovat G, Fister Jr I, Yermak K, Stiglic G, Fister I. Interestingness measure for mining sequential patterns in sports. Journal of Intelligent & Fuzzy Systems. 2015;29(5):1981–1994.

9. Decroos T, Van Haaren J, Davis J. Automatic discovery of tactics in spatio-temporal soccer match data. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining; 2018. p. 223–232.

10. Liu T, Fournier-Viger P, Hohmann A. Using diagnostic analysis to discover offensive patterns in a football game. In: Recent Developments in Data Science and Business Analytics. Springer; 2018. p. 381–386.

11. Magnusson MS. Discovering hidden time patterns in behavior: T-patterns and their detection. Behavior research methods, instruments, & computers. 2000;32(1):93–110.

12. Borrie A, Jonsson GK, Magnusson MS. Temporal pattern analysis and its applicability in sport: an explanation and exemplar data. Journal of sports sciences. 2002;20(10):845–852.

13. Camerino OF, Chaverri J, Anguera MT, Jonsson GK. Dynamics of the game in soccer: Detection of T-patterns. European Journal of Sport Science. 2012;12(3):216–224.

14. Jonsson GK, Anguera MT, Blanco-Villaseñor Á, Losada JL, Hernández-Mendo A, Ardá T, et al. Hidden patterns of play interaction in soccer using SOF-CODER. Behavior Research Methods. 2006;38(3):372–381.

15. Jonsson GK, Anguera MT, Sánchez-Algarra P, Olivera C, Campanico J, Castañer M, et al. Application of T-pattern detection and analysis in sports research. The Open Sports Sciences Journal. 2010;3(1).

16. Lapresa D, Anguera MT, Alsasua R, Arana J, Garzón B. Comparative analysis of T-patterns using real time data and simulated data by assignment of conventional durations: the construction of efficacy in children's basketball. International Journal of Performance Analysis in Sport. 2013;13(2):321–339.

17. Gutiérrez A, Prieto I, Cancela JM. Most frequent errors in judo Uki Goshi technique and the existing relations among them analysed through T-patterns. Journal of sports science & medicine. 2009;8(CSSI3):36.

18. Hughes MD, Bartlett RM. The use of performance indicators in performance analysis. Journal of sports sciences. 2002;20(10):739–754.

19. Ortega E, Villarejo D, Palao JM. Differences in game statistics between winning and losing rugby teams in the Six Nations Tournament. Journal of sports science & medicine. 2009;8(4):523.

20. Bishop L, Barnes A. Performance indicators that discriminate winning and losing in the knockout stages of the 2011 Rugby World Cup. International Journal of Performance Analysis in Sport. 2013;13(1):149–159.

21. Bremner S, Robinson G, Williams MD. A retrospective evaluation of team performance indicators in rugby union. International Journal of Performance Analysis in Sport. 2013;13(2):461–473.

22. Vaz L, Vasilica I, Kraak W, Arrones SL. Comparison of scoring profile and game related statistics of the two finalist during the different stages of the 2011 rugby world cup. International Journal of Performance Analysis in Sport. 2015;15(3):967–982.

23. Hughes A, Barnes A, Churchill SM, Stone JA. Performance indicators that discriminate winning and losing in elite men's and women's Rugby Union. International Journal of Performance Analysis in Sport. 2017;17(4):534–544.

24. Watson N, Durbach I, Hendricks S, Stewart T. On the validity of team performance indicators in rugby union. International Journal of Performance Analysis in Sport. 2017;17(4):609–621.

25. Mosey TJ, Mitchell LJ. Key performance indicators in Australian sub-elite rugby union. Journal of science and medicine in sport. 2019;.

26. Carter A, Potter G. The 1995 rugby world cup finals. 187 tries. 2001) Notational Analysis of Sport III University of Wales Institute Cardiff Cardiff, Wales UWIC. 2001; p. 224–229.

27. van Rooyen KM, Noakes DT. Movement time as a predictor of success in the 2003 Rugby World Cup Tournament. International journal of Performance analysis in Sport. 2006;6(1):30–39.

28. Hughes MT, Hughes MD, Williams J, James N, Vučković G, Locke D. Performance indicators in rugby union. 2012;.

29. Colomer CM, Pyne DB, Mooney M, McKune A, Serpell BG. Performance Analysis in Rugby Union: a Critical Systematic Review. Sports Medicine-Open. 2020;6(1):4.

30. O'Donoghue P, Williams J. An evaluation of human and computer-based predictions of the 2003 Rugby Union World Cup. 2004;.

31. Reed D, O'Donoghue P. Development and application of computer-based prediction methods. International Journal of Performance Analysis in Sport. 2005;5(3):12–28.

32. O'Donoghue P, Ball D, Eustace J, McFarlan B, Nisotaki M. Predictive models of the 2015 Rugby World Cup: accuracy and application. International Journal of Computer Science in Sport. 2016;15(1):37–58.

33. Croft H, Lamb P, Middlemas S. The application of self-organising maps to performance analysis data in rugby union. International Journal of Performance Analysis in Sport. 2015;15(3):1037–1046.

34. Lamb P, Croft H. Visualizing rugby game styles using self-organizing maps. IEEE computer graphics and applications. 2016;36(6):11–15.

35. Kohonen T. Exploration of very large databases by self-organizing maps. In: Proceedings of international conference on neural networks (icnn'97). vol. 1. IEEE; 1997. p. PL1–PL6.

36. Sasaki K, Yamamoto T, Miyao M, Katsuta T, Kono I. Network centrality analysis to determine the tactical leader of a sports team. International Journal of Performance Analysis in Sport. 2017;17(6):822–831.

37. Coughlan M, Mountifield C, Sharpe S, Mara JK. How they scored the tries: applying cluster analysis to identify playing patterns that lead to tries in super rugby. International Journal of Performance Analysis in Sport. 2019;19(3):435–451.

38. Watson N, Hendricks S, Stewart T, Durbach I. Integrating machine learning and decision support in tactical decision-making in rugby union. Journal of the Operational Research Society. 2020; p. 1–12.

39. Sasaki K, Furukawa T, Murakami J, Shimozono H, Nagamatsu M, Miyao M, Yamamoto T, Watanabe I, Yasugahira H, Saito T, Ueno Y. Scoring profiles and defense performance analysis in Rugby Union. International Journal of Performance Analysis in Sport. 2007 Oct 1;7(3):46-53.