Data-driven set piece analysis in football

Bram van Eerden[1 [s3726991]]

Leiden Institute of Advanced Computer Science, Leiden University, Leiden, The Netherlands

Abstract. This report analyzes set-piece strategies for a Premier League team aiming to improve their free kick effectiveness. Focusing on non-direct free kicks, we utilized event data from the 2017/2018 season. The data was filtered, normalized, and segmented into zones for tactical analysis. Dynamic Time Warping (DTW) measured sequence similarities, and K-Means clustering identified common tactics. Recommendations for the coaching staff are to focus on direct crosses from the left, near-post deliveries from the right, and diagonal crosses or short passes from the center. Implementing these strategies can enhance the team's set-piece performance.

1 Introduction

1.1 Client's demand

A premier league team wants to improve the danger they create with free kicks. They want to take a data-driven approach to find out what kind of tactics work best in the premier league. The head coach wants an analysis on historical data that shows what tactics they should work on in specific situations. The team generates sufficient chances from open play but has identified set pieces as an area with significant room for improvement.

1.2 Approach

To address the client's request, we will do analysis on set-piece situations, specifically focusing on scenarios where the team does not take a direct shot on goal but instead utilizes crosses or a series of passes to create scoring opportunities. We will only look at historical sequences that happen after a free kick, that are successful. In our analysis we will not look at the percentage of a certain tactic that succeeds. We will analyze these successful free kicks, and see if we can find patterns that may help our coach. Our approach will be divided into several key phases:

- Data Collection and Preprocessing:
  We will use the event dataset created by [1]. The data will be cleaned and preprocessed to standardize formats and make sure it fits our analysis techniques.
- Identification of Successful Set Pieces:
  Using dynamic time warping and clustering, we will analyze the data to identify patterns and sequences in set-piece play that lead to scoring opportunities or shots on goal. A set-piece will be deemed 'successful' if it leads to a significant chance creation or a direct shot on target.

– Tactical Analysis:
  For each successful cluster of set-piece patterns, we will analyze whether they can be useful for our coach. We will try to find patterns that are direct, and reproducable for a team.

1.3 Sports Data Science cube

This project touches on the Sports Data Science cube in several ways. The sports discipline in this project is football, sometimes called soccer.
Then, the data type we use is event data. For each event several values are recorded, like the type of event, the time and place. As the place of the event is also recorded, we make use of position data as well.
The application of this project is very much focused on tactics. We analyze the historical data provided, to cluster the free kicks into similar tactics. We will then analyze these tactics found, and give advice to the coach of the team on which tactics are useful.

2 Dataset

2.1 Data structure

The dataset used in this project is developed by Pappalardo et. al. [1]. The authors collected data of several large football leagues in the 2017/2018 season. The dataset consists of several segments:

– Events:
  The events are stored in json files. For each competition, a file is available. This file contains all events of all games in the season. Each event is described by its type, place and time, as well as the player involved in the event.
– Matches:
  The matches files store general data for each match played in that season. It has info on the date, time, teams, formations and the score.
– Player data:
  Info about all players in the dataset. This includes name, birth-data, nationality, position and preferred foot.

2.2 Preprocessing

Before we are able to analyze any sort of tactics used, we must first focus on structuring and preprocessing the data. Quite a few steps are necessary to achieve the desired structure. These steps will be described below.

Filtering
The event-id's in the dataset are not chronologically labeled. We therefore start by finding all the unique match ids in the dataset. Then, for each match, we sort the events by their timestamp. This gives us a chronological set of events in a single match.
The events are all labeled to a certain eventid and subeventid. The sub event id's

31 and 32 indicate 'free kick' and 'free kick cross'. The 'free kick' event indicates an event where the free kick is not crossed, but passed to a teammate.
After finding all free kicks in a match, we take the subsequent events in the next fifteen seconds. If an event in the next 15 seconds leads to a goal or opportunity, we classify the free kick as 'successful'. We then save the sequence of events up until the event that has the tag 'goal' or 'opportunity'.

Normalizing
A nuisance of the way this dataset is setup, is that the perspective of the location data changes based on the team that has the ball. For instance, team A takes the free kick and crosses it towards location (x=80, y=60). The cross is then cleared by a player of team B. Instead of the clearance being at location (x=80, y=60), the position data is inverted towards the offensive perspective of team B: (x=20, y=40). As we only care about the success of a free kick from the perspective of the team that takes the free kick, we normalize these positions, otherwise our analysis in a later stage will be distorted.

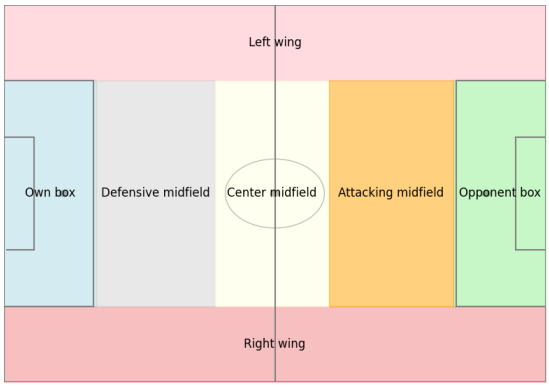

Fig. 1: The 7 zones for our analysis

Zone analysis
To provide the coach with tactics for separate situations, we filter the free kicks into three zones: 'Opponent half left', 'Opponent half center' and 'Opponent half right'. This way we can split the freekicks into similar strarting situations. Them, we can analyze the tactics used for free kicks from different parts of the pitch. We exclude all free kicks that take place on the half of the attacking team, as these do not provide a direct attacking opportunity.
After taking this subset of the free kicks for a certain location, we add more zonal information to the event data. For each event in a free kick sequence, we add a zone, which will allow us to do deeper analysis on the sequence. We split the field into the 7 zones as seen in figure 1.

3 Modelling technique

3.1 Dynamic Time Warping (DTW)

Dynamic Time Warping (DTW) is a technique used to measure similarities between temporal sequences that can vary in speed or timing [2]. In the context of our set-piece analysis, DTW is used in comparing sequences of football events, such as passes, crosses, and movements, which do not necessarily follow a uniform time distribution. This flexibility makes DTW particularly well-suited for analyzing sports data, where the timing of events can vary significantly between different instances of similar play types.
DTW works by finding an optimal alignment between two sequences, which minimizes the distance between them. This alignment is not constrained to be linear, meaning it can map one element of a sequence to multiple elements of another sequence, allowing for differences in timing and speed. The core idea is to "warp" the time dimension of sequences to achieve the best possible match, thereby enabling a more meaningful comparison of sequences that may differ in length or timing.

Feature Extraction
In our analysis, each event in a sequence is converted into a feature vector. Zones on the pitch are mapped to numerical values, transforming categorical spatial data into a numeric format suitable for distance calculations.

DTW Distances
Once the sequences are transformed into feature vectors, DTW is applied to measure the similarity between each pair of sequences. The optimal alignment between sequences minimizes the total distance between the elements.

3.2 Clustering

After computing the DTW distances, we use clustering techniques to group similar sequences together. This combination allows us to identify common tactical patterns and strategies used in successful set-pieces.

Distance Matrix
he DTW distances are converted into a distance matrix, which serves as the input for the clustering algorithm [3]. This matrix represents the pairwise distances between all sequences.

K-Means Clustering
K-Means clustering is applied to the distance matrix to group the sequences into distinct clusters. Each cluster represents a set of sequences that exhibit similar tactical patterns. In our analysis, we chose to create 7 clusters for each segment of starting locations of the free kicks.

4 Results

We will analyze the results of our analysis separately for each starting point of the free kicks. in our analysis we noticed that most clusters show that often successful sequences are interrupted by a clearance or a header by the opponent team, after which the attacking team still finds an opportunity. As this consists more of open play than free kick tactics, we omit these from our analysis. We therefore only discuss the clusters that show a direct tactic from these set pieces.

4.1 Left side of the pitch

The subset of free kicks that start on the left side of the pitch contains 195 sequences. This is the largest subset, the total amount of successful free kicks in our dataset contains 424 sequences.
The largest cluster of 88 sequences shows the free kicks that are directly played into the opponents box. This cluster contains free kicks that are directly dangerous, so at most 2 events are in between the free kick and the shot/opportunity. 97% of these free kicks were crosses directly into the box.
One other cluster in this subset was interesting, as it showed 15 clear sequences that employ an alternative tactic where the ball is played to the other side of the box through a sequence of short passes, which result in a shot on goal. These instances are show in figure 2

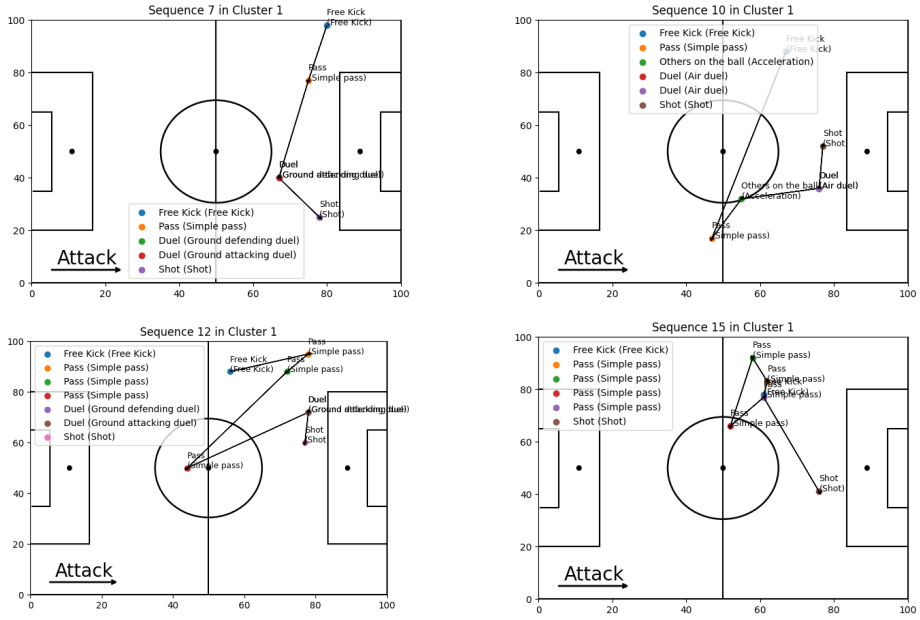

Fig. 2: Four out of fifteen instances where a freekick is passed around to find a shot outside of the box

4.2 Right side of the pitch

The subset of free kicks taken from the right side of the pitch contains 175 free kick sequences. The largest cluster here contains 84 sequences, which consist mainly of free kicks directly into the box. Interestingly, about half of the free kicks here are not direct crosses, but passes to a teammate. What stands out as well, is that over 65% is played towards the near post of the opponents goal (y coordinate of the second event is 40 or lower). Four instances of this tactic are show in figure 3

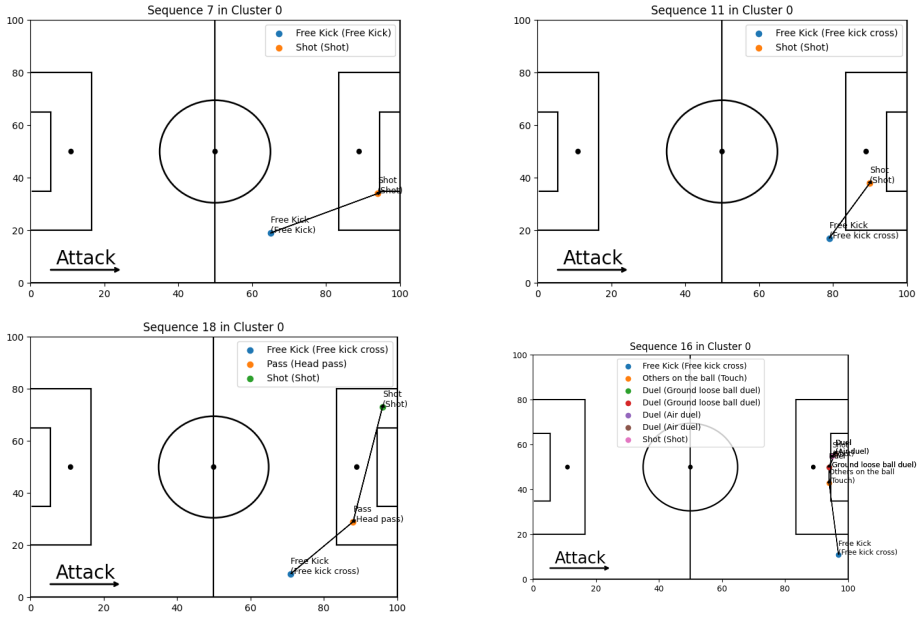

Fig. 3: Four instances where a free kick is played directly towards the near post of the opponents goal

4.3 Center of the pitch

The subset of free kicks taken from the center of the pitch is the smallest, with 54 sequences. This may well be because free kicks in this area are often seen as good chances to shoot on goal directly, a tactic we omit from our analysis. The first interesting cluster in this subset is the largest cluster, where 18 sequences are crossed diagonally into the box. Not all sequences lead to a direct shot, but it is interesting that this tactic is the most effective in this subset, seen in figure 4.

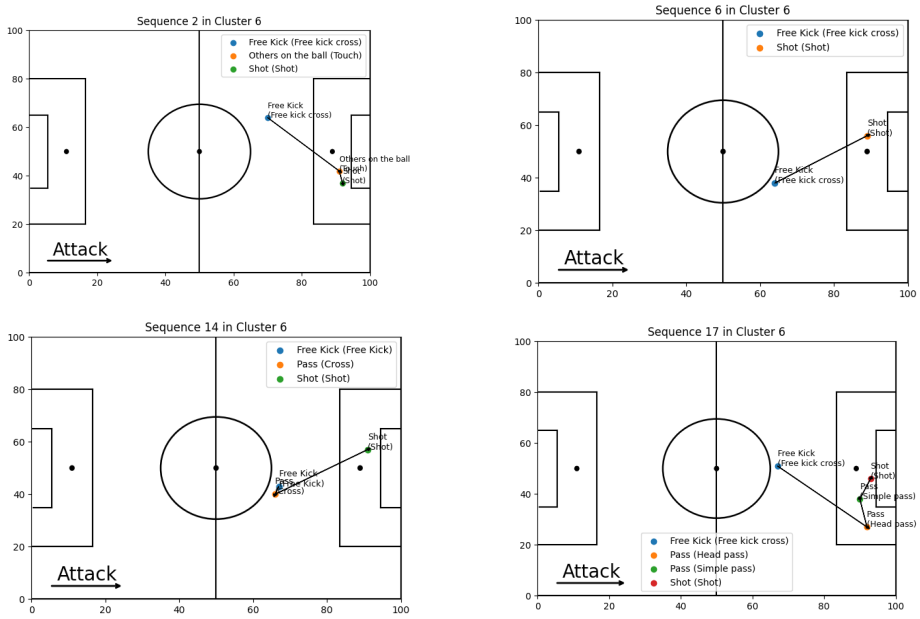

Fig. 4: Four instances where a free kick is crossed from the center of the pitch, diagonally into the opponents box

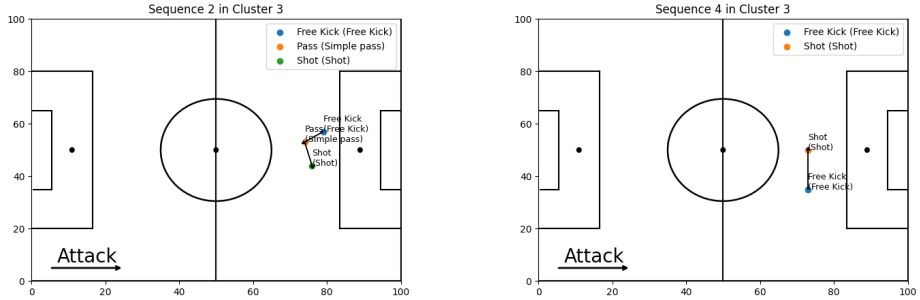

Fig. 5: two instances where a short passing sequence happens before a shot

A second, small cluster of 5 sequences show that some free kicks are successful when the ball is passed short before taking a shot from outside the box. These instances are shown in figure 5.

4.4 Concrete advice

Based on the analysis of of the successful free-kick sequences from different starting points on the pitch, we provide the following concrete advice for the coaching staff to enhance their set-piece strategies:

- **Left side of the pitch**
  The team can focus on two types of approaches: historically, a direct cross into the box is the most effective, and should probably be the main focus for these situations. An alternative tactic could be to switch play to the other side of the pitch by a short passing sequence. This alternative may well be effective because a cross is most often used, and an opponent may not expect it.
- **Right side of the pitch**
  Contrary to free kicks from the left, a direct cross into the box from the right may not be the most successful tactic. A significant portion of successful free kicks from this side is played towards the near post, through a cross or pass over the ground. The team can focus on either option that fits their type of players best.
- **Center of the pitch**
  Not many free kicks are played into the box before a shot is created, compared to free kicks from the wings of the pitch. If the team does opt to play the ball into the box from a central position, the most successful tactic is to cross it into the box diagonally. An alternative can be a short passing sequence, which was shown to be effective in a few cases. This can be used to bypass a wall of players in the box, for instance.

5 Discussion

This analysis shows some concrete and interesting tactics for free kicks in the premier league. The analysis could however be improved in several ways. First of all, we only focus on successful set pieces. We classify successful as a goal or opportunity, which makes the analysis more inclusive as it not only set pieces that led to goals. We don't know however, how often a certain tactic is successful out of the amount of times it has been used. Analyzing how often a tactic is useful out of the times it has been used can help discover underutilized tactics that may be very effective in practice.
Furthermore, this analysis can be done on a deeper level by separating the data into more sub types. For instance, only focusing on free kicks taken short from a smaller part of the pitch may provide even more focused tactics.
Finally, more features and info can be used to enrich this analysis. It would be interesting to see if the preferred foot of the free kick taker has an effect on the tactic, whether the median height of the players influence the type of successful set pieces.
In short, this analysis has led to some interesting insights, and we think many more insights can be obtained in future research.

References

1. Pappalardo, Luca Cintia, Paolo Rossi, Alessio Massucco, Emanuele Ferragina, Paolo Pedreschi, Dino Giannotti, Fosca. (2019). A public data set of spatio-temporal match events in soccer competitions. Scientific Data. 6. 10.1038/s41597-019-0247-7.
2. Giorgino, T.: Computing and Visualizing Dynamic Time Warping Alignments in R: The dtw Package. Journal of Statistical Software, 31(7), 1–24 (2009). URL: https://www.jstatsoft.org/index.php/jss/article/view/v031i07. DOI: 10.18637/jss.v031.i07
3. Tavenard, R.: An introduction to Dynamic Time Warping (2021). URL: https://rtavenar.github.io/blog/dtw.html