Mini-project:SoccerCPD: Formation and Role Change-Point Detection in Soccer Matches Using Spatiotemporal Tracking Data - MVA 2023/2024

Rania Bennani rania.bennani@ens-paris-saclay.fr
Raphaël Razafindralambo raphael.razafin@gmail.com

January 9, 2024

1 Introduction

In fluid team sports like soccer and basketball, decoding the intricate tactics employed by teams is crucial for gaining insights into their strategies. One fundamental aspect of this analysis revolves around understanding team formations, as they are pivotal in shaping players' movement patterns and interactions on the field. Despite the abundance of sports data, accurately estimating and tracking dynamic changes in team formations and player roles during a match presents a significant challenge. Coaches assign specific roles to players, yet these instructions can undergo alterations throughout the game. Additionally, players may temporarily switch roles with their teammates, and unforeseen events, such as set-pieces, can disrupt the expected team formation. Existing approaches often fall short, assuming a consistent formation throughout the match or overly frequent changes that may not align with tactical intentions. In response to these challenges, this article [KKC+22] introduces SoccerCPD, a change-point detection framework designed to distinguish tactically intended formation and role changes from temporary deviations in soccer matches.

The task at hand involves developing a comprehensive solution to accurately identify and interpret these tactical changes within the dynamic context of a soccer match. Standard approaches either oversimplify the problem by assuming constant formations or lead to excessive changes that do not align with real-world scenarios. SoccerCPD proposes a two-step approach: firstly, a formation change-point detection (FormCPD) method identifies points where the team's formation is assumed to be consistent. This is achieved by representing players' spatial configurations as role-adjacency matrices and employing a nonparametric change-point detection algorithm. Secondly, a role change-point detection (RoleCPD) method is applied within each formation period, capturing instances where players switch roles with one another.

We have equitably worked on this report, both in terms of writing and substance. Each of us delved into the theoretical methods outlined in the article, and each conducted new experiments not covered in the original paper. Due to the complexity of the code and the non-availability of compatible datasets, we used the entire available source code. We modify it to conduct new experiments such as the influence of some hyperparameters in the RoleCPD algorithm. We did not improve the original method as our accuracies tend to be lower than the authors' ones.

2 Method

Given a match time T and a time series $V = (V(t))_{t\in T}$ in $\mathbb{R}^{N\times 2}$ representing the records of the x, y positions of the $N = 10$ outfield players during a game, our goal is to identify times with significant changes. The method is divided into two steps: first, it finds formation change-points, and then it finds role change-points.

2.1 FormCPD

The objective of FormCPD is to find change-points of the distribution of the V_t's. Given a sequence of features $A = \{A(t)\}_{t\in T}$ representing the role topology, we aim to find a partition of **formation periods** $T_1 < \cdots < T_m$ of T such that $t \in T_i \implies A(t) \sim \mathcal{F}_i$ for some distinct formation distributions $\mathcal{F}_1, \ldots, \mathcal{F}_m$.

To build features, the paper computes a role-adjacency matrix $A(t) \in \mathbb{R}^{N\times N}$ at each frame t, from the set $\mathcal{X} = \{X_1, \ldots, X_N\}$ of roles and the Delaunay triangulation. The components are given for all (i, j) in N^2 by $[A(t)]_{i,j} = 1$ if the roles X_i and X_j are adjacent, 0 otherwise. A CPD algorithm is then applied on the series A using g-segmentation of the matrices.

In the paper, a graph-based CPD algorithm named **discrete g-segmentation** is used. This non-parametric method builds a statistic test based on the Manhattan distance. The method aims to find one change point $\tau \in T$ in a sequence of observations. The paper suggests empirical conditions for considering τ as significant.

After generating T_1 and T_2, a recursive framework is constructed to find multiple change-points. The method is applied on the new partitions until no significant change point is detected. This CPD technique is adapted for high-dimensional data with repeated observations/matrices over T. Given the partitions $T_1 < \cdots < T_m$ found by g-segmentation, each formation period T_i is associated with a mean role-adjacency matrix $A(T_i)$ and a mean role location $V(T_i)$. For all i in $\{1, \ldots, m\}$, we have

$$A(T_i) = \frac{1}{|T_i^*|} \sum_{t\in T_i^*} A(t), \quad V(T_i) = \frac{1}{|T_i^*|} \sum_{t\in T_i^*} V(t),$$

and the formation in T_i is the graph $F(T_i) = (V(T_i), A(T_i))$. To assign a label to the m formations, the paper includes them in a data collection of $n = 864$ annoted formation periods to apply a two-steps agglomerative clustering. It gives $K = 7$ formations. The distance function used for clustering is $d((F, A), (F', A')) = d_M(QAQ^T, A')$ where $Q \in \mathbb{R}^{N\times N}$ is a permutation matrix found by the Hungarian algorithm.

2.2 RoleCPD

Initially, each player $p \in P$, where P denotes the set of players in a match time T, is assigned to an initial role $X_p \in \mathcal{X}$ according to a uniform number. The goal is to partition each formation period T_i into several time intervals $T_{i,1} < \cdots < T_{i,n_i}$ named **role periods**.

The **player-to-temporary-role (P-TR) maps** are defined as functions $\beta_t : P \to \mathcal{X}$, where $p \mapsto \pi_t(X_p)$. At time t in T, the permutation π_t defines the temporary roles of the players. If no change is observed at t, π_t is the identity. RoleCPD consists in finding significant change points in the sequence of the role permutations $(\pi_t)_{t\in T}$ using g-segmentation.

The distance used to build a similarity graph is the *Hamming distance* given for all pairs (t, t') by $d_H(\pi_t, \pi_{t'}) = |\{X : \pi_t(X) \neq \pi_{t'}(X), X \in \mathcal{X}\}|$. Only valid permutations with switch rates below 0.7 are considered. The recursive CPD is applied to the sequence from each formation period T_i, leading to $T_{i,1} < \cdots < T_{i,n_i}$.

Given a role period $T_{i,j}$, we set an instructed role for each player. Then, the final step consists in aligning roles and assigning domain-specific position labels (LB,LR,LM, etc) to instructed roles of players. To achieve this, the paper clusters the roles in each formation group using agglomerative clustering and Hungarian algorithm. Domain labels are then set based on domain knowledge.

3 Data

For our experiment, we initially intended to apply the algorithm to a real dataset, such as the positional data of each Manchester City player during a Premier League match. Unfortunately, these data were private, leading us to utilize a sample match dataset provided by the authors, named "17985.ugp."

3.1 "17985.ugp" Dataset

This dataset consists of upgraded GPS (Global Positioning System) data capturing soccer player movements during a South Korean professional soccer league match. It includes various columns like "player id," "session," "gametime," "unixtime," "player period," "duration," "x," "y," and "speed." Each row corresponds to a timestamp, providing detailed insights into player movements.

	player_id	session	gametime	unixtime	player_period	duration	x	y	speed
2020-01-01 19:00:00.100	4181	1	00:00.1	1.577905e+09	1	0.1	4633.0	574.0	4.615196
2020-01-01 19:00:00.200	4181	1	00:00.2	1.577905e+09	1	0.1	4642.0	561.0	4.655973
2020-01-01 19:00:00.300	4181	1	00:00.3	1.577905e+09	1	0.1	4651.0	548.0	4.714858
2020-01-01 19:00:00.400	4181	1	00:00.4	1.577905e+09	1	0.1	4659.0	534.0	4.781228
2020-01-01 19:00:00.500	4181	1	00:00.5	1.577905e+09	1	0.1	4667.0	519.0	4.844462

Figure 1: First rows of the dataset

Specifically, the "gametime" and "unixtime" columns enable insights into player behavior during specific match phases. While spatially, the "x" and "y" coordinates reveal player positioning, aiding in the analysis of formations, strategies, and responses to in-game events.
Notably, this dataset serves as the primary input for the entire code (in the "main" function). Additional datasets provided by the authors, namely "form_periods.pkl," and "role_records.csv," are specific to the formation clustering and role labeling steps. These datasets are detailed above:

3.2 Formation Clustering:

This dataset contains essential information about soccer formations during match periods. Key attributes like "activity_id,","session", "form_period", "start_dt","end_dt", "duration", "coords" and "edge_mat"(edge matix) offer insights into temporal and spatial aspects. (Some rows of the table's data is given in Appendix - Figure 4)

Remark: The authors also provide a CSV file containing the "ground truth" formation and role labels. When analyzing this file, we noticed that the number of formations was limited to only five formations

(instead of the 18 formations recorded). Thus, the provided dataset appears to be poorly chosen. This limitation impacts the performance of the clustering algorithm, and the number of "outliers" (formations not among the five specified ones) is directly neglected. Additionally, during our experiments, we attempted to change the clustering method (K-Means instead of Hierarchical clustering). However, the significant number of outliers prevented us from achieving accurate results.

3.3 Role Labeling:

The "role_records" dataset is fundamental for soccer role labeling, featuring information crucial for assigning domain-specific role labels to players within different formations during various match periods. With 21,128 rows, it includes columns such as "activity_id," "player_period," "form_period," "role_period," "session," "start_dt," "end_dt," "duration," "player_id," "squad_num," "player_name," "base_role," "x," "y," "formation," and "aligned_role."This latter column represents soccer-specific role labels assigned through a role alignment process, aligning players' roles with labels like 'LWB,' 'LCB,' 'CB,' 'RCB,' 'RWB,' 'RCM,' 'LCM,' 'LM,' 'CF,' and 'RM,' based on identified formations. The role alignment relies on a predefined mapping of roles to formations. (Some rows of the table's data is given in Appendix - Figure 5)

4 Results

4.1 Hyperparameter tuning

In this section, we focus on examining the impact of various hyperparameters in the model. Our attention is particularly centered on those from the Change Point Detection part. The three hyperparameters we are going to study are the matrix distance used in FormCPD, the permutation distance used in RoleCPD, and the maximum switch rate. Their evaluation will be based on the result on one sample match, as we mentionned earlier in the Data part. While this approach might lead to limited accuracy in comparing different hyperparameter values, it remains the most feasible method available to us at the moment.

FormCPD matrix distance: The role adjacency matrices are compared by using the *Manhattan distance*. Here, we define distances that are induced by these well-defined norms on the matrix space:

- **Nuclear norm (Ky Fan 'n'-norm)**: $\|A\| = \sum_{i=1}^{N} \sigma_i(A)$ where the singular values of $A \in \mathbb{R}^{N\times N}$ are denoted by $\sigma_1, \ldots, \sigma_N$
- **Frobenius norm**: $\|A\| = \sqrt{\mathrm{tr}(A^T A)} = \sqrt{\sum_{i=1}^{N} \sigma_i^2(A)}$

To make the comparison, we set the maximum switch rate at $r = 0.8$. While the Manhattan distance predicted $m = 4$ formation periods, the other ones predicted two. Along with these newly introduced formation periods, we also have new role periods, the number of which has been reduced by one.

In figures 6 and 7 we measured the prediction accuracies of (1) team formation and (2) player position. Namely, we calculated the ratios of correctly detected one-minute segments to the total number of segments (total minutes played). In order to calculate the accuracy, we had to annotate the groups formed by the clustering step by ourselves, based on our knowledge. Then, we compared to ground-truth labels provided by the author of the paper. We observe that Manhattan

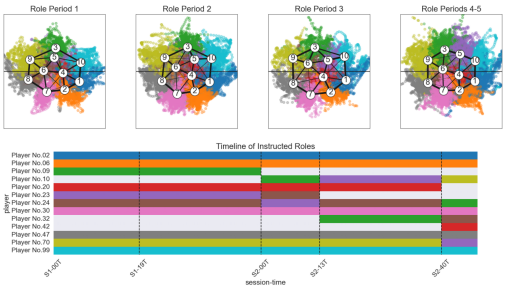

Figure 2: FormCPD & RoleCPD on Manhattan distance

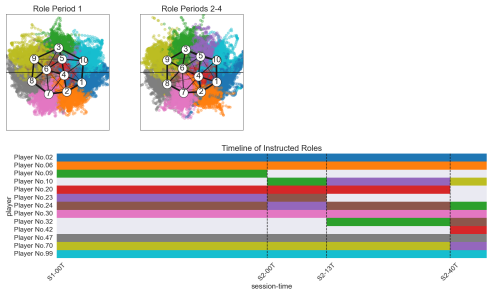

Figure 3: FormCPD & RoleCPD on nuclear and Frobenius distances

Distance achieve a better performance regarding correctness on formations and roles prediction. This shows that the final result can be very sensitive to the choice of the matrix distance. But the reverse is also true, because Frobenius and nuclear distances lead to the same result.

RoleCPD permutation distance: Here we proposed two distances to compare permutations

- L_1-distance defined by $d(\pi^1, \pi^2) = \sum_{i,j=1}^{N} \|\pi_i^2 - \pi_i^1\|$.
- Kendall τ distance [Ken38] that counts the number of pairwise disagreements between two ranking lists. We normalized it to make it have its value in $[0, 1]$ and be equal to 0 if the lists are similar.

To compare them, we fixed the matrix distance as *manhattan* and r as 0.7. However, we got exactly the same results as the hamming case. While the test statistics and the p-values of the tests for g-segmentation had different value between methods (because distance matrices were different), the expected change points τ remained always the same. That's why we observe exactly the same partitions. So we have the same predictions and accuracies.

Maximum switch rate: The concept of the maximum switch rate revolves around the highest acceptable level of tolerance for tactical permutations, such as swapping positions between the attacking midfielder (CAM) and the center forward (CF) in a false nine setup, and so on. A lower switch rate implies a greater tendency to overlook or downplay temporary positional changes, leading to the masking of various game interruptions like injuries, substitutions, and free kicks. When the switch rate is exceptionally low, there's a heightened risk of obscuring real playing time, including individual actions within the game. Thus,a balance had to be found to optimize this rate.
We compared values of 0.7 and 0.8 on the sample match. The number of partitions (formation and role) we get are the same, but the start of the second formation period differs (30 seconds difference). The accuracies are almost identical (1 minute of difference between formation accuracy, and 2*minutes* for role accuracy).
To obtain optimal results, conducting a grid search to identify the best hyperparameters would be a viable approach. However, due to the significant memory/time requirements for each CPD, as well as the need for formation annotation to perform clustering, we were unable to implement this strategy.

5 Appendix:

	activity_id	session	form_period	start_dt	end_dt	duration	coords	edge_mat	cluster	formation
0	12864	1	1	2020-01-01 14:00:00	2020-01-01 14:47:00	2820.0	[[-719.0, 1625.0], [-1342.0, 660.0], [948.0, 1...	[[0.0, 0.977, 0.955, 0.2, 0.023, 0.828, 0.257,...	20	442
1	12868	1	1	2020-01-01 14:00:00	2020-01-01 14:47:00	2820.0	[[-833.0, -1647.0], [14.0, 181.0], [-369.0, -1...	[[0.0, 0.338, 0.684, 0.994, 0.165, 0.042, 0.15...	5	4132
2	12868	2	2	2020-01-01 15:02:00	2020-01-01 15:50:00	2880.0	[[-933.0, -1939.0], [-106.0, 618.0], [-231.0, ...	[[0.0, 0.164, 0.86, 0.993, 0.131, 0.036, 0.237...	20	442
3	12870	1	1	2020-01-01 13:30:00	2020-01-01 14:19:00	2940.0	[[-1408.0, 43.0], [460.0, -1087.0], [-285.0, 5...	[[0.0, 0.22, 0.871, 0.098, 0.367, 0.962, 0.328...	12	352
4	12870	2	2	2020-01-01 14:33:00	2020-01-01 15:23:00	3000.0	[[-1282.0, -147.0], [290.0, -888.0], [-209.0, ...	[[0.0, 0.416, 0.886, 0.054, 0.192, 0.998, 0.27...	-1	others

Figure 4: First rows of the formation clustering dataset

	activity_id	player_period	form_period	role_period	session	start_dt	end_dt	duration	player_id	squad_num	player_name	base_role	x	y	formation	aligned_role
0	1879	1	1	1	1	2020-01-01 16:00:00	2020-01-01 16:35:00	2100	1252	9	P09	1	1696.0	-65.0	433	CF
1	1879	1	1	1	1	2020-01-01 16:00:00	2020-01-01 16:35:00	2100	1759	10	P10	2	136.0	-745.0	433	RCM
2	1879	1	1	1	1	2020-01-01 16:00:00	2020-01-01 16:35:00	2039	1760	14	P14	3	1037.0	1791.0	433	LM
3	1879	1	1	1	1	2020-01-01 16:00:00	2020-01-01 16:35:00	2100	1761	15	P15	4	815.0	-1458.0	433	RM
4	1879	1	1	1	1	2020-01-01 16:00:00	2020-01-01 16:35:00	2100	1237	19	P19	5	-621.0	-94.0	433	CDM

Figure 5: First rows of the role labeling dataset

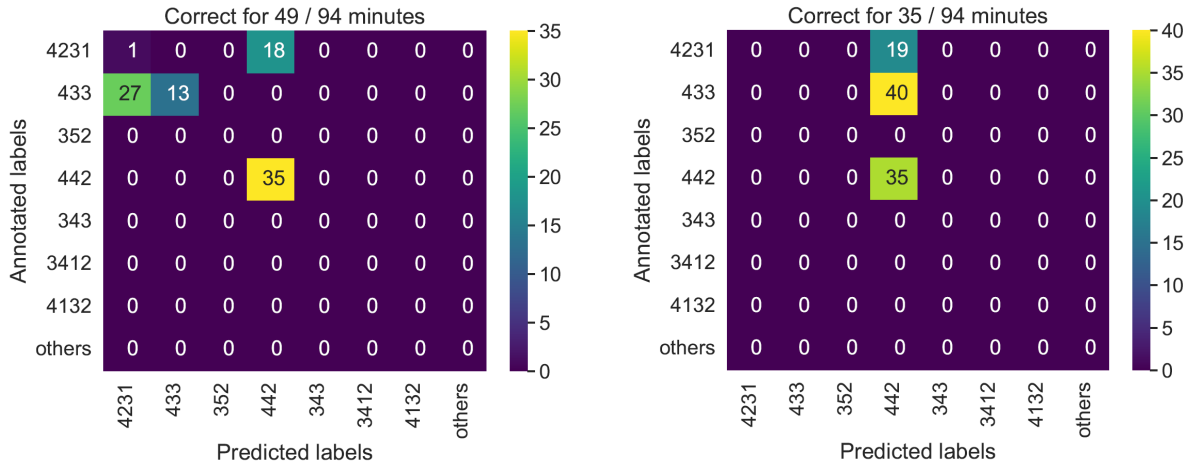

Figure 6: Confusion matrix for formation accuracy (left: manhattan, right: nuclear and frobenius)

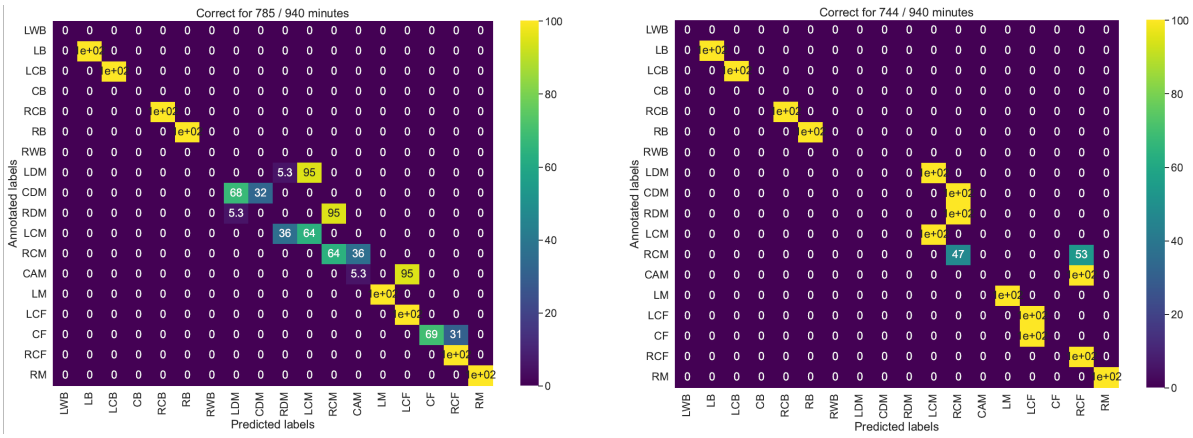

Figure 7: Confusion matrix for role accuracy (left: manhattan, right: nuclear and frobenius)

References

[Ken38] M. G. Kendall. A new measure of rank correlation. *Biometrika*, 30(1-2):81–93, 1938.

[KKC+22] Hyunsung Kim, Bit Kim, Dongwook Chung, Jinsung Yoon, and Sang-Ki Ko. Soccer-cpd: Formation and role change-point detection in soccer matches using spatiotemporal tracking data. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '22. ACM, August 2022.