EasyChair Preprint

№ 14049

Identifying Team Playing Styles Across Phases of Play: a User-Specific Cluster Framework

Samuel Moffatt, Ritu Gupta, Suman Rakshit and Brad Keller

July 20, 2024

Identifying Team Playing Styles Across Phases of Play: A User-Specific Cluster Framework *

Samuel J Moffatt[1][0000−0002−2048−1556], Ritu Gupta[1][0009−0008−6753−7730], Suman Rakshit[1][0000−0003−0052−128X], and Brad S Keller[2][0000−0003−2684−8926]

[1] Curtin University, Perth, Australia
[2] Fremantle Football Club, Perth, Australia

Abstract. Investigating team performance at a granular level within matches allows for the analysis of phase-specific team playing styles. The cluster framework described in this paper assists in identifying phase-specific team playing styles within team invasion sports that are vital for guiding match analysis. This paper develops a novel clustering framework, proposing a composite clustering assessment index for selecting the optimal feature transformation technique, clustering algorithm and number of clusters. The proposed composite index allows the integration of subject matter expert knowledge, ensuring that the resulting clusters are chosen optimally in alignment with the analysis goals. The clustering framework is applied in the context of Australian football to identify an interpretable number of clusters that represent the inherent grouping of team playing styles during match phases.

Keywords: Clustering framework · Playing style · Team sport · Australian football

1 Background Information

Within team invasion sports, teams develop strategies within periods of the match to better manage match complexity [39]. Observing team behaviour is crucial for improving overall performance [14]. Media, coaches, and sports enthusiasts commonly refer to the overriding behaviour as a team's playing style or style of play. The pervasive and colloquial use of the playing style expression has not been matched by scientific research and measurement [14].

Current literature investigates playing styles at a whole match level. Match-level offensive and defensive playing styles have been identified within team sports such as Australian football [14, 25, 26], association football [11, 12, 35, 36], netball [10], and rugby league [40, 41]. To understand the complexity of team behaviour within invasion sports, a team's actions and events can be broken down into smaller periods of play [6, 11, 14, 28]. Identifying team playing styles within smaller periods of play expands current literature and allows for granular performance analysis, improving match analysis compared to whole match analysis.

* Supported by Curtin University and the Fremantle Football Club.

This paper outlines a clustering framework identifying phase-specific playing styles within team invasion sports. The framework is applied to cluster team actions and identify phase-specific team playing styles within the context of Australian football.

2 Cluster Framework

The methodological framework described below objectively identifies team playing styles within Australian football. The aim of the clustering process is important, guiding decisions when implementing the clustering framework [17]. The clustering aim is to identify an interpretable number of clusters that represent the inherent grouping within the Australian football transactional match data. The framework also outlines the feature engineering process and the selection of important performance features. Summarised match possession chains are separated into the four phases of play (offence, defence, transition and out-of-play), where cluster analysis is applied and evaluated using a composite cluster assessment index. The optimal clustering hyper-parameters are identified using the composite cluster assessment index.

2.1 Data Requirements

Transactional match data, commonly available for team invasion sports, is required to identify phase-specific playing styles [7]. The data contains timestamped event information on the type of action performed and the individual performing the action. A sample of the transactional match data collected during an Australian football match is displayed in Table 1.

Table 1. Sample of the transactional match data collected during an Australian football match.

Match	Time	Team	Player	Event
100001	5	T_A	P_{17}	Gather
100001	7	T_A	P_{17}	Short kick
100001	8	T_B	P_{32}	Mark
100001	17	T_B	P_{32}	Long kick
100001	20	T_B	P_{36}	Mark

2.2 Data Preprocessing

Data preprocessing is divided into data cleaning and feature engineering, producing an analysis-ready data set. The raw data is cleaned following best practice data cleaning processes [42]. The important stage of feature engineering, guided by subject matter experts (SMEs), creates new features identifying performance indicators and the state of possession [18, 23]. The data is summarised into possession chains, represented regarding chain duration.

Feature Engineering

Performance Features: Absent/present calculations are performed to engineer performance indicator features (defined in [29]) from the cleaned raw data. When an event that reflects a performance indicator occurs, a one or otherwise, a zero is coded for the respective feature (Table 2).

Table 2. Sample of transactional match data displayed in Table 1 with engineered performance features (bold).

Match	Time	Team	Player	Event	**Gather**	**ShortKick**	**Mark**	**LongKick**
100001	5	T_A	P_{17}	Gather	**1**	**0**	**0**	**0**
100001	7	T_A	P_{17}	Short kick	**0**	**1**	**0**	**0**
100001	8	T_B	P_{32}	Mark	**0**	**0**	**1**	**0**
100001	17	T_B	P_{32}	Long kick	**0**	**0**	**0**	**1**
100001	20	T_B	P_{36}	Mark	**0**	**0**	**1**	**0**

Possession State Features: The state of possession is identified for each event during a match. A rolling possession status is then calculated, with the rolling possession status not changing until a new event results in a change of possession. A chain of possession is defined as a group of events that occur while one team is in possession, the ball is in dispute or out-of-play. The chain duration is calculated from the time elapsed during the possession chain. The engineered performance features are summated to summarise the actions performed during the possession chain (Table 3).

Table 3. Sample of the transactional match data (Table 1) summarised into possession chains. The performance features (bold) represent the total number of actions performed within the possession chain.

Match	Chain	Duration	Team	Phase	**Gather**	**ShortKick**	**Mark**	**LongKick**
100001	C_2	3	T_A	Offence	**1**	**1**	**0**	**0**
100001	C_3	12	T_B	Offence	**0**	**0**	**2**	**1**

Chain Duration Representation: Representing performance features as total counts of actions causes issues because the features correlate with the duration of the possession chain. For example, a possession chain of 120 seconds typically contains more actions than a possession chain of 10 seconds. The rate of each feature is calculated to overcome the correlation with chain duration by dividing features by the chain duration. The rate at which performance indicators are performed (number of actions per second) will reflect the intent and strategy of a team (Table 4).

Table 4. Sample of the summarised possession chain data (Table 3) with each performance feature transformed concerning the duration of the chain.

Match	Chain	Duration	Team	Phase	**Gather**	**ShortKick**	**Mark**	**LongKick**
100001	C_2	3	T_A	Offence	**0.33**	**0.33**	**0**	**0**
100001	C_3	12	T_B	Offence	**0**	**0**	**0.17**	**0.08**

2.3 Mathematical Notation

The analysis-ready data set is notated as $\mathbf{X} = \{x_{ij} : i = 1, ..., n, j = 1, ..., p\}$. x_{ij} is the observation corresponding to row i and column j. $\mathbf{X}$ has n rows (each row is a chain of possession) and p columns (reflecting engineered performance features). Column j reflecting the engineered performance features is defined as $\mathbb{C}_j = \{x_{ij} : i = 1, ..., n\}, j = 1, ..., p$ where p is the number of engineered features. As the clustering goal is to reflect playing styles within specific phases of play, $\mathbf{X}$ is separated into phase-specific data sets based on the possession state feature. This results in $\mathbf{X}_{\text{off}}$, $\mathbf{X}_{\text{def}}$, $\mathbf{X}_{\text{tran}}$ and $\mathbf{X}_{\text{out}}$.

2.4 Feature Selection

In discussion with SMEs, the process of feature selection is applied to $\mathbf{X}_{\text{off}}$, $\mathbf{X}_{\text{def}}$, $\mathbf{X}_{\text{tran}}$, $\mathbf{X}_{\text{out}}$ to create user-informed data sets with features assisting in the description of team playing styles. Removing irrelevant features will reduce the required data storage and improve computation time without negatively impacting the analysis [5]. The filter feature selection method [5] is applied to a subset of $\mathbb{C}_j$ selected by SME. All features with $\text{var}(\mathbb{C}_j) = \epsilon$ ($\epsilon < \delta$, a threshold close to zero) are removed as these features have constant variance and will not impact the results.

2.5 Clustering Process

No optimal clustering algorithm exists for all possible data sets as different clustering algorithms provide different solutions [38]. To account for the different clustering results, three cluster algorithms, k-means clustering, k-medoids clustering and hierarchical agglomeration clustering using Ward's linkage, are applied to each $\mathbf{X}$. The clustering process, guided by the clustering aim, investigates the discovery of similar possession chains across $\mathbf{X}_{\text{off}}$, $\mathbf{X}_{\text{def}}$, $\mathbf{X}_{\text{tran}}$, $\mathbf{X}_{\text{out}}$. The clustering process involves the application of multiple clustering algorithms ($\mathcal{C}$) across a range of cluster numbers ($\mathcal{K}$) to identify the optimal $\mathcal{C}$ and k for each $\mathbf{X}$.

Individual assessment indexes (I_j) characterise the separate characteristics of cluster results and are used to assess the results of $\mathcal{C}$ across $\mathcal{K}$. There are a number of I_j within literature, specifically within-cluster sum of squares (I_{wcss}) [20], separation index (I_{sep}) [3], distance to the cluster centroid (I_{distcc}) [1], and density index (I_{dens}) [20] which are used within the clustering framework. To develop a composite cluster assessment index ($\mathcal{A}(\mathcal{C})$), indexes are normalised

(I_j^*) because different I_j have different optimal directions and value ranges to point in the same direction across a range of [0,1], where larger index values are considered better. Equation 1 is applied to I_j where larger values are better, and Equation 2 is applied to I_j where smaller values are better.

$$I_j^* = \frac{I_j - \min I_j}{\max I_j - \min I_j} \tag{1}$$

$$I_j^* = 1 - \frac{I_j - \min I_j}{\max I_j - \min I_j} \tag{2}$$

Using the four assessment indexes highlighted above, a composite cluster assessment index with equal I_j^* weights, $\mathcal{A}(\mathcal{C})_1$ is defined. The maximum $\mathcal{A}(\mathcal{C})_1$ for each $\mathbf{X}$ is used to identify the optimal clustering method $\mathcal{C}$ and k number of clusters.

$$\mathcal{A}(\mathcal{C})_1 = \frac{I^*_{\text{wcss}} + I^*_{\text{sep}} + I^*_{\text{distcc}} + I^*_{\text{dens}}}{4} \tag{3}$$

2.6 Clustering Process Extensions

Feature standardisation Feature standardisation techniques such as z-score and variance-to-range standardisation can be applied before clustering [17, 37]. In addition to $\mathbf{X}$, $\mathcal{C}$ can be applied across $\mathcal{K}$ for the transformed versions of $\mathbf{X}$. $\mathcal{A}(\mathcal{C})$ can be used to assess the effect of the feature transformation on the data and used to select the optimal feature transformation technique.

Index Normalisation A more robust index normalisation method outlined in [4, 19] can be implemented when creating I_j^*. Normalisation involves applying randomised clustering techniques to generate multiple clusterings on the same data. I_j^* for all clusterings are then calibrated using z-score transformation to their expected variation over clusterings of the same data, where a higher number is optimal.

Weighted $\mathcal{A}(\mathcal{C})$ When creating $\mathcal{A}(\mathcal{C})$, [2] recommends the consideration of the dynamics of I_j^* in relation to the clustering goal. I_j^* are weighted (w_j) in discussion with SMEs to bias individual indexes, with a higher weight reflecting a greater importance. For instance, I^*_{distcc} assists in explaining information reduction, reflecting how well-represented the cluster is by the cluster centroids. SMEs determined the information reduction to be of full importance, with $w_{\text{wcss}} = 1$. $\mathbf{X}$ is required to be clustered into a small number of groups of similar profiles, prioritising within-cluster homogeneity over cluster separation. SMEs determined $w_{\text{wcss}} = 1$ and $w_{\text{sep}} = 0.5$. The characteristic that the density of the cluster reduces from the cluster centroid to the outer regions is required to augment $\mathcal{A}(\mathcal{C})$, however, with less impact than the other indexes, defined as $w_{\text{dens}} = 0.25$.

This results in $\mathcal{A}(\mathcal{C})_2$, specific to the goal of clustering playing styles of Australian football. The heuristic weighting of indexes can be further extended by formalising w_j through methods such as the Analytical Hierarchical Process [30].

$$\mathcal{A}(\mathcal{C})_2 = \frac{1 \cdot I^*_{\text{wcss}} + 0.5 \cdot I^*_{\text{sep}} + 1 \cdot I^*_{\text{distcc}} + 0.25 \cdot I^*_{\text{dens}}}{1 + 0.5 + 1 + 0.25} \tag{4}$$

3 Deployment

Transactional match data was provided by Champion Data©, Melbourne, Australia, the official statistical provider of the AFL. Following Section 2.2, chain summary data was created to summarise team performance within specific phases of play. For the offensive phase of play, in discussion with SMEs informed by the clustering aim, an interpretable range of possible clusters is determined, $\mathcal{K} = \{2, ..., 10\}$. The R package `fpc` [21] was used to implement the cluster algorithms and assessment indexes to create $\mathcal{A}(\mathcal{C})$. Selecting the maximum $\mathcal{A}(\mathcal{C})_2$ across $\mathbf{X}_{\text{off}}$ identified five offensive playing styles (Fig. 1).

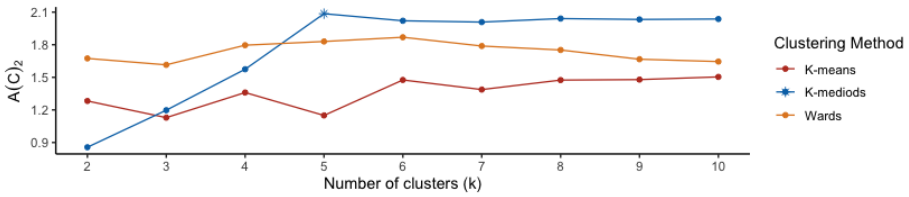

Fig. 1. $\mathcal{A}(\mathcal{C})_2$ assesses the results of three clustering methods across $\mathcal{K}$ for $\mathbf{X}_{\text{off}}$. The maximum $\mathcal{A}(\mathcal{C})_2$ identifies the k-mediod clustering algorithm with five clusters to produce the best clustering result.

4 Conclusion

The clustering framework can be implemented within any team invasion sport to analyse performance. Implementing the framework within Australian football results in the identification of team playing styles within specific periods of a match. The optimal hyper-parameters identified by the maximum $\mathcal{A}(\mathcal{C})$ results in the identification of k playing styles for each specific phase of play within Australian football. The cluster centroids of the optimal clusters can then be used to characterise data points within each cluster, defining the team playing styles within the phase of play. Coaches and analysts can use the identified playing styles to assist with team and opposition analysis, identifying the specific styles of play that teams implement at specific periods of games. This can lead to the adaption of team tactics and strategies within specific periods of a match to maximise the likelihood of success.

References

1. S. Agarwal. Data mining: Data mining concepts and techniques. IEEE. 2013)
2. S. Akhanli. Distance construction and clustering of football player performance data. UCL (University College London). (2019)
3. S. Akhanli and C. Hennig. Comparing clusterings and numbers of clusters by aggregation of calibrated clustering validity indexes. Statistics and Computing. (2020)
4. S. Akhanli and C. Hennig. Clustering of football players based on performance data and aggregated clustering validity indexes. (2022)
5. S. Alelyani, J. Tang and H. Liu. Feature selection for clustering: A review. Data Clustering. (2018)
6. J. P. Alexander, B. Spencer, J. K. Mara and S. Robertson. Collective team behaviour of Australian rules football during phases of match play. J Sports Sci. (2019)
7. Z. Born. Tactical performance insights for Australian rules football using deep learning. The University of Western Australia. (2022)
8. J. Castellano, D. Casamichana and C. Lago. The use of match statistics that discriminate between successful and unsuccessful soccer teams. Journal of human kinetics. (2012)
9. C. M. E. Colomer, D. B. Pyne, M. Mooney, A. Mckune and B. G. Serpell. A qualitative study exploring tactical performance determinants from the perspective of three Rugby World Cup coaches. International Journal of Sports Science and Coaching. (2022)
10. H. Croft, B. Willcox and P. Lamb. Using performance data to identify styles of play in netball: an alternative to performance indicators. International Journal of Performance Analysis in Sport. (2017)
11. J. Diquigiovanni and B. Scarpa. Analysis of association football playing styles: An innovative method to cluster networks. Statistical Modelling. (2019)
12. J. Fernandez-Navarro, L. Fradua, A. Zubillaga, P. R. Ford and A. P. Mcrobert. Attacking and defensive styles of play in soccer: analysis of Spanish and English elite teams. Journal of Sports Sciences. (2016)
13. M. A. Gómez, R. Pollard and J.-C. Luis-Pascual. Comparison of the home advantage in nine different professional team sports in Spain. Perceptual and motor skills. (2011)
14. G. Greenham, A. Hewitt and K. Norton. A pilot study to measure game style within Australian football. International Journal of Performance Analysis in Sport. (2017)
15. C. Hennig and B. Hausdorf. Design of dissimilarity measures: A new dissimilarity between species distribution areas. In: Data science and classification. (2006)
16. C. Hennig and T. F. Liao. How to find an appropriate clustering for mixed-type variables with application to socio-economic stratification. Journal of the Royal Statistical Society: Series C (Applied Statistics). (2013)
17. C. Hennig. Clustering strategy and method selection. Handbook of Cluster Analysis. (2015)
18. C. Hennig, M. Meila, F. Murtagh and R. Rocci. Handbook of cluster analysis. CRC press. 2015)
19. C. Hennig. Cluster validation by measurement of clustering characteristics relevant to the user. Data analysis and applications 1: Clustering and regression, modeling-estimating, forecasting and data mining. (2019)
20. C. Hennig. An empirical comparison and characterisation of nine popular clustering methods. Advances in Data Analysis and Classification. (2022)

21. C. Hennig. fpc: Flexible procedures for clustering. R. https://CRAN.R-project.org/package=fpc.(2023)
22. L. Hubert and J. Schultz. Quadratic assignment as a general data analysis strategy. British journal of mathematical and statistical psychology. (1976)
23. M. Kuhn and K. Johnson. Applied predictive modeling. Springer. 2013)
24. C. Lago-Peñas, M. Gómez-Ruano and G. Yang. Styles of play in professional soccer: an approach of the Chinese Soccer Super League. International Journal of Performance Analysis in Sport. (2018)
25. J. C. Lane, G. Van Der Ploeg, G. Greenham and K. Norton. Characterisation of offensive and defensive game play trends in the Australian Football League (1999–2019). International Journal of Performance Analysis in Sport. (2020)
26. S. J. Moffatt, R. Gupta, N. French Collier and B. S. Keller. Classifying and quantifying team playing styles in the Australian Football League. International Journal of Performance Analysis in Sport. (2024)
27. I. B. Mohamad and D. Usman. Standardisation and its effects on K-means clustering algorithm. Research Journal of Applied Sciences, Engineering and Technology. (2013)
28. M. J. Rennie, S. J. Kelly, S. Bush, R. W. Spurrs, D. J. Austin and M. L. Watsford. Phases of match-play in professional Australian football: Distribution of physical and technical performance. J Sports Sci. (2020)
29. S. Robertson, R. Gupta and S. Mcintosh. A method to assess the influence of individual player performance distribution on match outcome in team sports. J Sports Sci. (2016)
30. R. W. Saaty. The analytic hierarchy process—what it is and how it is used. Mathematical modelling. (1987)
31. C. E. Shannon. A mathematical theory of communication. The Bell system technical journal. (1948)
32. W. Sheehan, R. Tribolet, A. R. Novak, J. Fransen and M. L. Watsford. An assessment of physical and spatiotemporal behaviour during different phases of match play in professional Australian football. J Sports Sci. (2021)
33. B. Spencer, S. Morgan, J. Zeleznikow and S. Robertson. Clustering team profiles in the Australian Football League using performance indicators. In: Proceedings of the 13th Australasian conference on mathematics and computers in sport, Melbourne. (2016)
34. D. Steinley and M. J. Brusco. A new variable weighting and selection procedure for K-means cluster analysis. Multivariate Behavioral Research. (2008)
35. A. Tenga and E. Sigmundstad. Characteristics of goal-scoring possessions in open play: Comparing the top, in-between and bottom teams from professional soccer league. International Journal of Performance Analysis in Sport. (2011)
36. B. Travassos, K. Davids, D. Araújo and T. P. Esteves. Performance analysis in team sports: Advances from an Ecological Dynamics approach. International Journal of Performance Analysis in Sport. (2013)
37. N. K. Visalakshi and K. Thangavel. Impact of normalisation in distributed k-means clustering. International Journal of Soft computing. (2009)
38. K. Wang, B. Wang and L. Peng. CVAP: validation for cluster analyses. Data Science Journal. (2009)
39. C. Wedding, C. Woods, W. Sinclair, M. Gomez and A. Leicht. Exploring the effect of various match factors on team playing styles in the National Rugby League. International Journal of Sports Science and Coaching. (2021)

40. C. Wedding, C. Woods, W. H. Sinclair, M. A. Gomez and A. S. Leicht. Analysis of styles of play according to season and end of season rank in the National Rugby League. Journal of Science and Medicine in Sport. (2021)
41. C. Wedding, C. T. Woods, W. H. Sinclair, M. A. Gomez and A. S. Leicht. Exploring the effect of various match factors on team playing styles in the National Rugby League. International Journal of Sports Science and Coaching. (2021)
42. H. Wickham. Tidy Data. Journal of Statistical Software. (2014)