Not All Features Are Created Equal: Player Clustering and Evaluation

Andrew Kang, Scott Powers, Travis Curson

Rice University

Motivation

The player clustering problem is one that has been studied in various publications, most notably Soccerment. However, most clustering methodologies contain a common flaw: they treat all features with equal weight. This flaw is a critical one, since players may be deemed similar under an empirically chosen set of features, while the slightest changes may result in the opposite conclusion. Given an anonymized dataset, we do not have such luxury of validating and fine tuning the generated clusters based on our domain knowledge.

We seek to address this flaw by identifying a hierarchy of features that should not be given equal weight, grouped by relevance. Thus, we propose a two-step recursive K-means algorithm that first clusters player styles by their relative tendencies near the ball (shooting, passing, dribbling, tackling, dueling, etc.) and then creates sub-clusters based on the scout's objective. If the scout's objective is to find a passer like De Bruyne, the scout would run our sub-clustering methodology with passing-related features.

Additionally, this abundance of sub-clusters motivates our new player evaluation method, tailored to understand what each player cluster needs to excel in the game. This method aims not just to identify player abilities, but to contextualize them within the strategic requirements of winning.

Methodology

Objective 1: Construct generalizable, useful framework that needs less validation

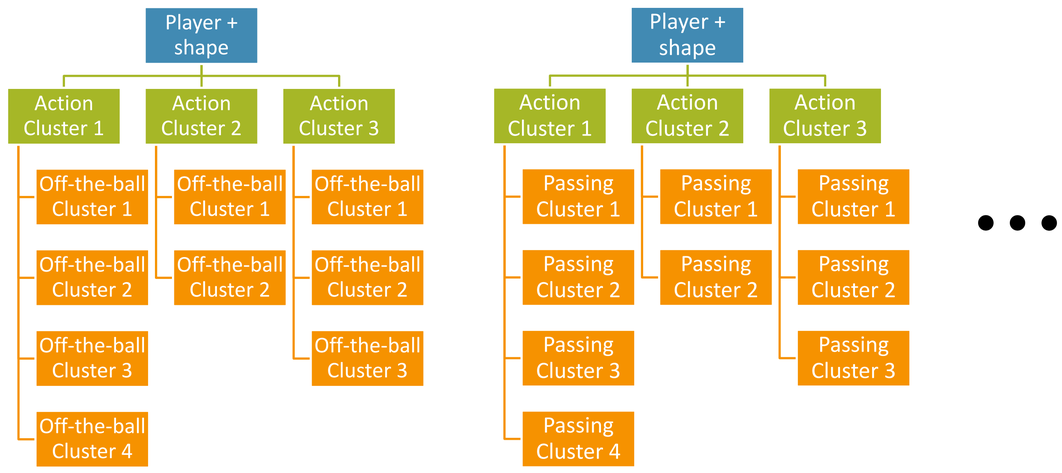
Objective 2: Capture player tendencies independent of skill, leaving room for player development and evaluation stage

Objective 3: Capture player roles under diverse tactical situations, to view each player as a set of several personas using Opta's live team shape data

➔ Hierarchy of Features

➔ Objective-specific Clustering

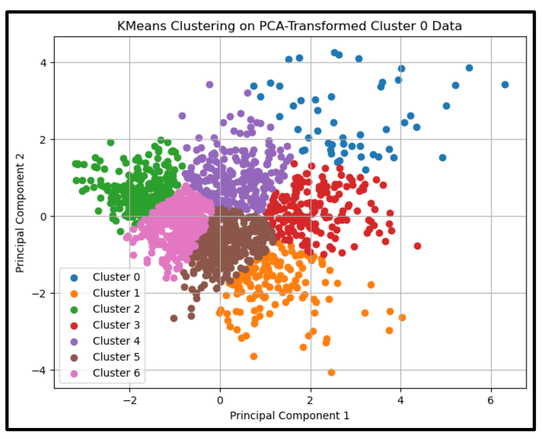
➔ Division of players into personas for each team shape

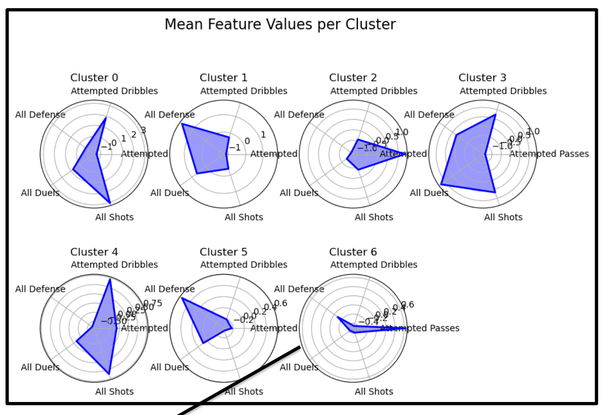


Clustering Results

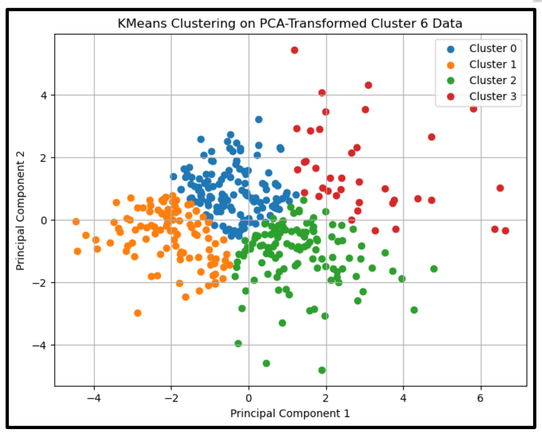
For each step in our recursive clustering, we use PCA on the step's relevant features and then k-means.

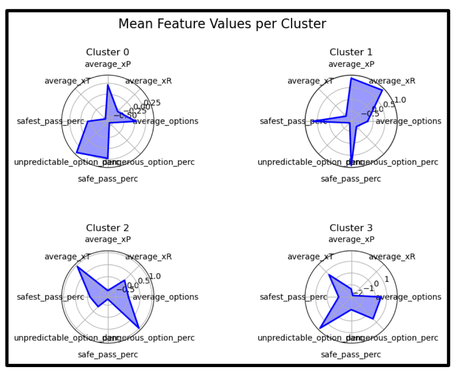
Created Features per Cluster





Step 1: PCA and Initial Clusters





Step 2: PCA and Passers' Passing Sub-clusters

Cluster-specific Player Evaluation

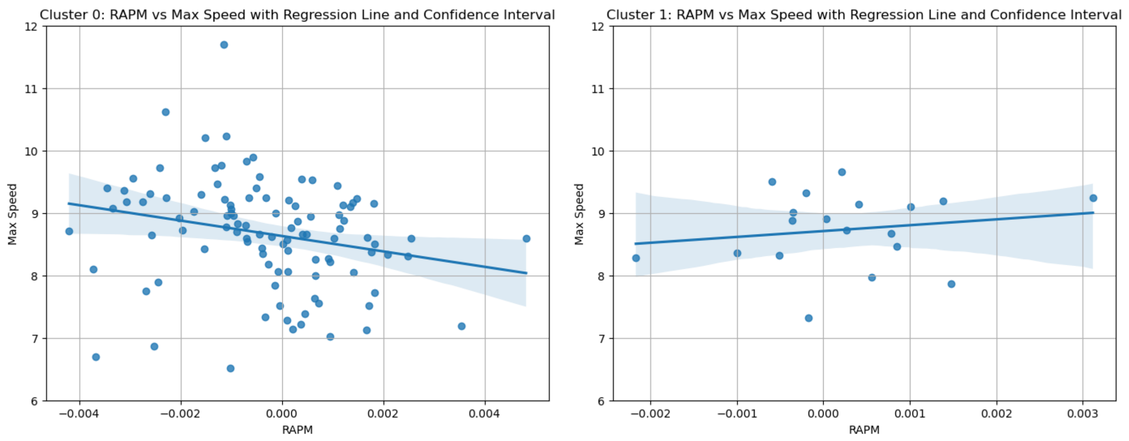
We will adapt Box Plus-Minus (BPM) from basketball analytics: We estimate the contribution of each player to their team's xG differential using Regularized Adjusted Plus-Minus (RAPM).

This methodology works well in basketball, but it generally leads to very noisy results in football because in football there are many fewer substitutions and scoring opportunities than in basketball.

For each player i, this results in a skill rating s_i. Because the estimated player skills from Step 1 are noisy, we perform a linear regression of the RAPM estimates onto box score statistics b_1, ..., b_p (e.g. pass completion percentage in the final third), across all players.

$$\hat{s}_j = \alpha + \beta_1 b_{1j} + \beta_2 b_{2j} + \ldots + \beta_p b_{pj}$$

The scatter plots below (using sprint sub-clusters of forwards) demonstrate an initial step toward building this equation per cluster.



A more refined equation that uses the cluster as an indicator variable is shown below.

$$\text{RAPM} = \beta_0 + \sum_{k=1}^{p} \beta_k b_k + \sum_{i=0}^{n-1} \sum_{k=1}^{p} \gamma_{ik} I(\text{Cluster} = i) b_k$$

Final Scouting Pipeline Example

1. Set Objective: Search for a player who passes like Kevin De Bruyne
2. Which initial cluster is Kevin De Bruyne in?
3. Which passing cluster is Kevin De Bruyne in?
4. Which players in that cluster are the best at what they should be good at?

References

Dan T. Rosenbaum. Measuring How NBA Players Help Their Teams Win, April 2004. http://www.82games.com/comm30.htm.

Joseph Sill. Improved NBA Adjusted +/- Using Regularization and Out-of-Sample Testing, March 2010. MIT Sloan Sports Analytics Conference.