Hierarchical clustering for player recruitment

Neil Charles
neil@hilltop-analytics.com

Why use clustering for player recruitment?

Clustering helps us to ask questions in plain English rather than immediately diving into detailed statistics. It can also take account of far more data than we are able to manually calculate.

Rather than trying to filter Opta Data to find a player who has a final third pass success rate above 80%, good ball retention and shoots more than twice per 90 minutes, we could say “find me a player like Wayne Rooney”.

Clustering will also bring a lot more data to this conversation, closely matching other metrics such as defensive contribution, movement on the pitch and passing accuracy from different locations.

Once we have a shortlist of similar players, it’s much easier to manually impose our own constraints. The question might become “Find me a player like Rooney who is under 25 and playing in England, Spain or Germany.”

Hierarchical clustering is very flexible

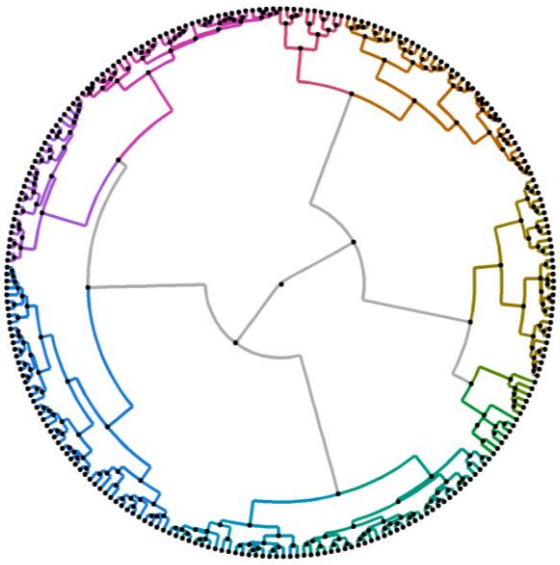
A key benefit of Hierarchical clustering over other techniques is that it is flexible. The algorithm creates a branching tree, with just a few similar players at the end of each branch. If we need more options, we can move back up the tree’s branches to find a longer list of players who are a little less similar.

Many variables could be used to compare similar players, including more advanced metrics such as expected goals or expected assists.

The model that you can see today incorporates…
Touches in 13 pitch areas, pass accuracy in different locations, clearances, crosses and success rates, assists, goal attempts from distance and close, shot on target rates by distance, tackles made and suffered, takeons and success rate, times dribbled past and times offside.

All variables are ‘per 90’ minutes, or per touch to ensure that players are comparable.

A branching tree of 200 players, built using hierarchical clustering



Clustering helps to identify benchmarks

Comparing Opta statistics between players can be difficult. A player’s pass accuracy will depend on the role that he’s asked to fulfil – well beyond simple positional splits like “defender” or “striker”.

Clustering helps with this problem. We benchmark against a set of players who are carrying out similar roles. E.g. rather than benchmarking against all strikers, a player might be compared to others who are asked – like he is – to also drop deeper and defend.

Rooney’s pass success rate in comparison to a set of similar players, identified by clustering