Hierarchical Clustering for Player Recruitment

Area of Study

I propose to show how Opta's statistics can be transformed to answer recruitment questions in the language that a manager or scout might ask them; moving raw numbers into the background – at least initially – and defining players by describing their peers.

The analysis will answer the question, "can you find me a player like x?" where x is a named player from any league covered by Opta data, or even a hypothetical, invented player. It will allow a manager or coach to instantly see a shortlist of statistically similar alternatives to a player who they need to replace on their own team, or who they would like to sign.

This analysis builds upon and extends the scope of a piece that I published in 2015 under the title "Find me a player like Andrés Iniesta".

I have worked with Opta's XML data before.

Method

Hierarchical clustering is a tree-based data mining algorithm. Applied to Opta data, it ranks the large number variables available to us, in terms of their ability to describe differences between players.

The method first splits players by very obvious factors – an early split might separate goalkeepers from everybody else – and then branches out into a tree, splitting again and again until it has arranged players into small groups, which are statistically similar. E.g. it might use position on the pitch, passing accuracy, chance creation, take-on success and other factors to identify a small group of five or six elite attacking midfielders who all share a similar playing style.

Hierarchical clustering has very useful properties for player recruitment, in that its definitions can be relaxed and filtered, allowing us to create the following workflow:

1. Find me a player like x
2. Show me more options (relax the grouping by moving back up the tree's branches)
3. Show me only players with a pass completion rate above 80%, and aged under 25
4. Show me detailed statistical profiles for each player on this list
5. Create a shortlist for further investigation and scouting

The analysis will also allow for questions that are not based on an existing player. We will be able to take an existing player's profile and adapt it – e.g. "find me a player like x, but with better ball retention", or even ask for a list of current players who are similar to one who has retired.

Alternatively, a scout can define the performance numbers for a completely hypothetical 'ideal'

player who he needs to recruit, and then via lookalike analysis, find statistically similar real-world players.

Hierarchical clustering, together with a workflow that incorporates searches and filtering, extends the type of analysis presented by Will Gurpinar-Morgan in 2015. A key difference to Gurpinar-Morgan's work is the generation of very many more and much smaller clusters of players, grouped using playing position, style *and* success rates, explicitly targeting clubs' recruitment questions. Secondly, hierarchical clustering is better suited than k-means segmentation to the particular questions posed by player recruitment, because it can be so easily adapted by filtering and moving up or down the tree's branches.

As a starting point, the analysis will investigate inclusion of the following variables at a player level.

- Event frequency per 90 minutes (pass, shot, through-ball etc.)
- Success rate or outcome for each event type
- Variance in success rates (is the player consistent?)
- Event location on the pitch (pitch split into zones, optimal number to be determined)
- Proportion of passes played by type

The analysis will focus on outfield players, due to goalkeepers requiring a unique set of metrics and problems with repeatability in goalkeeping statistics.

Output from this analysis will be a presentation covering the methodology, a worked example and discussion of potential further developments. If feasible in the venue, I would also welcome the chance to demonstrate the workflow live using an interactive dashboard, allowing attendees to ask "find me a player like…"

Contact

Neil Charles

Email: neil@hilltop-analytics.com

Phone: XXXXX XXXXXX

www.hilltop-analytics.com