QAC 211

Patton Yin

Data that I plan to use:

<https://github.com/statsbomb/StatsBombR>

And this is a part of the instruction:

<https://github.com/statsbomb/open-data/blob/master/doc/Open%20Data%20Events%20v4.0.0.pdf>

My previous plan is to do a detailed analysis on the counter-attack effectiveness, however, the data such as the time that the player had done sprinting, the second player position in the data form are not the data that I can find.

So, I’m thinking about changing my topic.

My plan now is to focus on the clustering techniques:

Basing on different existing concepts classifying different soccer style, I would like to apply multi-dimensional analysis to it.

(However, my concern is that this might not be a standardized research question, I wonder if this is possible as a class project: if in the end I end up giving the visualizations of data by integrating different dimensions which would give us a classification on the style of soccer is acceptable. But anyways, stuff below are my thoughts)

So, ideally, there are in total 4 types of soccer strategies:

1. The team that favors established attacks & the team that executes defense at a high position.
2. The team that favors established attacks & the team that executes defense at a low position.
3. The team that favor counterattacks & the team that executes defense at a high position.
4. The team that favor counterattacks & the team that executes defense at a low position.

And each type will have certain traits that they will show on the pitch, and my goal is to utilize the data that we can get, and try to ultimately classify the teams into 1 of the 4 categories.

Here’re some variables that I am considering:

Possession ratio

The higher the possession ratio one team gets is more likely to be the team that favors established attacks.

Some possible improvements on this variable:

* I can possibly add some other influencing factors such as the difference between the top teams and the bottom teams.
* Also, I can add some other factors as the scores at some instant, because teams of 3-0 and 0-3 will act totally differently.
* Pass accuracy can also be taken into account. If a team gets a higher passing accuracy, this represents a better skill in doing established attacks.

Number of shots – position

The number shots and their associated positions of shooting reveals the difference between strategies of teams.

If a team have most of their shots outside of the box, this might indicate that the team is less capable of doing established attacks

Possible improvements:

* Related events
  + For shots close to the gate:
    - If this shot is caused by a cutback, or a thorough ball just from the outside of the box, this might indicate that the team is better at doing established attacks
    - If caused by a cross, the team might not be better at this.

Number of thorough balls – position

The number of thorough balls one team has made might be an indicator of the team’s ability to do established attacks

Ball recovery - position

The position which a team recovers the ball is a pretty straightforward way of illustrating whether a team prefers to defend at a high position or not.

Possible improvements:

* The talent (ability) difference between two teams.
* The time and score which might influence

And the stuff listed below are some other variables that I’m considering using but not yet have a good idea of how to fit them into my analysis.

Play-pattern – from counter

Player heatmap?

Pressure (counterpress) - position

First 15 minutes of match, most likely indicates the strategy that the coach has assigned.

How to tell what the strategy is?