Project

Gustav Sternelöv

25 april 2016

Disposition

1. Introduction
   Things like main idea with the project, purpose with the study. What more?
2. Background
   Former studies
   Background info about the data
   Football info, like type of strikers
3. Methods
   Describe the chosen methods and motivate why they have been chosen
4. Results
   Present the results
5. Discussion
   Discuss the results with respect to the aims and problems stated in the introduction(?).

The data - WhoScored and Opta

Defintion of the metrics (shots, key passes etc.)
Time period for the data. 2015/16 season up to round...
Limitations such as the number of games played, only time when used as a striker etc.
How data is collected
All metrics are per 90 minutes. Average contribution/performance per 90 minutes.
https://www.whoscored.com/Glossary
http://optasports.com/news-area/blog-optas-event-definitions.aspx
http://statsbomb.com/2013/08/an-introduction-to-the-per-90-metric/

Transformations of data, does makes sense to scale data? http://rtutorialseries.blogspot.se/2012/03/r-tutorial-series-centering-variables.html
http://stackoverflow.com/questions/20256028/understanding-scale-in-r

Purpose

Why only strikers (narrow down the project), why only top leagues (availability of data)
Investigate if the strikers can be divided into different kinds of strikers. Are there clear differences between strikers and how many different types can be found.
Use for analysis of strikers, in the context of scouting, as well as for player development. "This player is supposed to be a target player but only wins ... aerials per 90 minutes."

Types of strikers

https://chiefsiddharth.com/2014/10/01/top-strikers-in-world-football-an-analysis-of-roles/
Lone striker
Provider
Poacher
All rounder

http://www.guidetofootballmanager.com/tactics/strikers
Target man
Complete forward
Trequartista
False nine
Defensive forward
Deep-lying forward

http://businessdayonline.com/2014/12/five-dynamics-of-modern-soccer-strikers/
Target man
Complete forward
Trequartista
Poacher
Advanced forward

Mine would be *target man*, *complete forward*, *trequartista*, *poacher* and *false nine*. Some more link could be useful.
Would like to say that there are at least three different kinds of forwards and at most six different types. Does then imply that the k-means should be prespecified with k equal to either of 3, 4, 5 or 6.

Characteristics of methods

Actual methods are k-means, hierarchical clustering and DBScan. Is of interest to look more closely at the characteristics of each of these methods. Then, of course, also of importance to motivate the chosen methods with respect to the characteristics of the methods and the dataset.
K-means:
- Works best for compact and well separated clusters
- Finds spherical and convex shaped clusters
- Guarantees to find the local minimum

Although
- Unable to handle noisy data and outliers since all points must belong to a cluster. The outliers may impact the mean values a lot and the clustering will suffer.
- Results depends on the initial centroids
- Sensitive to the choice of K and the initial centroids
- Need to specify number of clusters is a clear disadvantage of the method

Hierarchical clustering:
Distance measures
- Single link. Hierarchical clusters defined by local proximity.
- Complete link. Really useful if true clusters are rather compact and of approximately equal size. Tends to find clusters opting for global closeness.
- Both complete link and single link represents extremes in measuring the distance between clusters. Both tent to be overly sensitive to outliers and noisy data.
- Average link. Can handle both numerical and categorical data.
- Mean and average link are compromises of single and complete link and are not as sensitive to outliers.

Agglomerative(AGNES)

- From scratch all points forms their own cluster
- At every step are the closest clusters merged
- Gives a hierarchy of clusters
- Need to specify the distance between the clusters
- Is called AGNES
- Requires at most n iterations

DBScan:
Can find clusters of arbitrary shapes. All points are necessarily not assigned to a clusters. Is therefore robust to noise and outliers.
- Takes parameters Eps (max radius of neighborhood) and minPts (min points in the specified neighborhood).
- The algorithm is sensitive to the choice of parameter settings

How many methods do I want to use? Is the aim to compare the results given for one algorithm, just changing the settings, or is the aim to use several methods where just one setting is used for every algorithm?

The article about the basketball players seem to show that k-means can be used for splitting players into different group. Could be reasonable to think that there are players that are outliers? If so, another algorithm like DBScan might be interesting to compare with.

Background

Article about NBA players where one of the aims were to investigate how they could be splitted into different groups using game statistics. This was examined by using clustering with the k-means method. The aim with this paper is to investigate if the same thing can be done, but instead of basketball i look att football players and game statistics from football matches.
Think this part, maybe thru subheads, also could include the info about strikers and the dataset.

Introduction

Should include things like *purpose* and main idea with the paper. Why it is interesting etc.

Results

Perform the clusterings with the chosen methods. Present the results with some clever visualisation. Perhaps the heatmaps are a good choice. An alternative could be regular grouped barplots or some multivariate plots (like radars, one per cluster).
Also give som typical examples of the players that belong to the respective cluster to get a feeling for the results in terms of what players that it is. If it is surprising or not. If it is big stars, players from the same league or some other splitting that has been made.
Number of players from the respective leagues in each cluster

Discussion

In order to perfrom this, the problems stated in the background/introdution need to be more clearly specified I think. Otherwise the discussion won't give so much. Need to create something interesting to discuss (search for earlier work about the topic! Could be other sports as well. Use FIFA rankings to find the All-stars and the normal strikers?)