A Data Driven Approach to Characterize Soccer Playing Styles

Author: Mohammad Mustafa Arif | Supervisor: Professor Timothy Chan

Introduction

Soccer is the most popular sport in the world. Today, more than 240 million people play soccer regularly, and a combined 3.5 billion viewers tuned in to the 2018 FIFA World Cup broadcast. Despite its rich history, the first record of analytics in soccer is from 1996, when Opta began recording match statistics for the English Premier League. Since then, the field has rapidly evolved and teams have begun using data-driven approaches for tactical decision making. Recently Liverpool FC, an elite English soccer club, announced the development of "pitch control", a probabilistic model used to quantify space creation during matches in real time.

Background

Traditional methods in soccer analytics involve utilizing high level match statistics such as possession percentages, numbers of shots, passes, crosses, fouls, and offsides. However, these methods are prone to oversimplifying 90+ minute interactions between two teams to a set of global statistics. An increasingly popular approach to add nuance to these methods is the use of expected goals (xG), a metric to measure the quality of shots taken by their probability of resulting in a goal. In addition, modern approaches in literature have attempted to study a squad's passing profile through the distribution of passes amongst the players. The gap, however, exists as the subtleties amongst different kinds of passes are not accounted for. Furthermore, little exploration of team defensive activity has been conducted.

Objectives & Methodology

The first objective of this project is to expand on existing literature characterizing team passing distributions and fill in the gaps by:

1) Augmenting the passing distribution with metrics measuring pass difficulty, such as pass distance, proximity to an opponent's goal, and the lateral and vertical pass components
2) Measuring game buildup speed through the average time between passes
3) Incorporating defensive metrics: the distribution of tackles, interceptions, and fouls

The appropriate metrics will be derived from the Wyscout spatio-temporal soccer event logs[1], a public dataset covering Europe's top 5 domestic leagues for the 2017/18 season. Some metrics will be computed directly from the logs, while others will require the construction of player passing network graphs. The next milestone is to use a multi-dimensional representation of a team derived from these approaches to cluster teams into different playing styles using unsupervised clustering methods such as mixture models, k-means, hierarchical clustering, and DBSCAN. Finally, the project aims to develop a predictive model by leveraging these representations and cluster assignments to predict future performance from past information. For predictive modelling, tree algorithms, linear regressions, and deep neural nets will be tested.

Desired Outcomes

The desired outcome of the project is to arrive at a team classification into different playing styles and analyze the effectiveness of certain playing styles over others. The project aims to leverage this information through a model capable of simulating league rankings.

---

[1] Pappalardo, L., Cintia, P., Rossi, A., Massucco, E., Ferragina, P., Pedreschi, D. and Giannotti, F., 2019. A public data set of spatio-temporal match events in soccer competitions. Scientific Data, 6(1).