Soccer Analytics: how Data Science is changing the “Beautiful Game”

Luca Pappalardo
Institute of Information Science and Technologies (ISTI)
National Research Council of Italy (CNR)
www.lucapappalardo.com

Soccer statistics are attracting wide interest from a long time ago. Already in the early 1950s Charles Reep collected soccer statistics *by hand* to suggest that “the key to scoring goals and winning games was to transfer the ball as quickly as possible from back to front”, thereby indirectly starting the long-ball movement in English football [1]. Apart from Reep's experience and other sporadic attempts, it is only in the recent years that soccer statistics have developed in an amazing way, thanks to automated sensing technologies that provide high-fidelity data streams extracted from every match, based on video recordings by different cameras or observations by various kinds of fixed and mobile sensors [2]. There are now professional statistical analysis firms, like Wyscout and Opta, which provide soccer data to clubs, coaches and managers, who are interested in such services to ensure they can remain in control of their performances and results as much as possible, by monitoring their players and opponents. The need of unravelling the complexity of the “beautiful game” has fostered the emergence of *Soccer Analytics*, a new discipline that is attracting wide interest from both the academia and the industry.

In the first part of the talk we run through the history of soccer analytics, from Charles Reep to the Big Data era, revising the sensing technologies, the big data available and the open challenges that are interesting to computer scientists. We start by describing the available technology for sensing the so-called *soccer-logs*, the most common data format which describe in great detail all the ball touches that occur during a match [2]. We then describe the most common representation of a soccer match - the passing network [3] - discussing its potential, its limits, and the performance measures that can be extracted from it [4, 5].

In the second part of the talk we focus on the most challenging open problems in soccer analytics: the ranking of soccer players and the data-driven evaluation of their performance. Ranking players means defining a relation of order between them with respect to some measure of their performance over a sequence of matches. In turn, measuring performance means computing a data-driven performance rating which quantifies the quality of a player's performance in a specific match and then aggregate them over the sequence of input matches. This is a complex task since there is no objective and shared definition of performance quality, which is an inherently multidimensional concept. We present a solution to this problem called *PlayeRank* [6], a new-generation data-driven framework for the performance evaluation and the ranking of players in soccer. PlayeRank exploits soccer-logs and machine learning to infer the importance of each ball touch made by a player to win a match, hence describing the quality of the performance of a player as a scalar product between the vector importances and a multidimensional vector describing the performance of a player. An extensive experimentation of PlayeRank over a massive dataset of soccer-logs provided by Wyscout - including 31 millions of ball touches, around 20K matches and 21K players in the last four seasons of 18 soccer competitions - produces a world ranking of soccer players, giving the unprecedented opportunity to investigate the patterns of soccer performance, characterize “outliers” and predict the evolution of a player’s career.

[1] C. Reep, B. Benjamin, "Skill and chance in association football", Journal of the Royal Statistical Society, vol. 131, pp. 581-585, 1968.

[2] Gudmundsson, J., Butte, A. J., Horton, M. Spatio-Temporal Analysis of Team Sports. ACM Computing Surveys 50:2, (2017).

[3] Lopez Pena J. and Touchette H., A network theory analysis of football strategies, arXiv1206.6904L, 2012.

[4] Duch J., Waitzman J. S., Amaral L. A. N., Quantifying the Performance of Individual Players in a Team Activity. PLOS ONE 5:6 (2010).

[5] Cintia, P., Pappalardo, L., Pedreschi, D., Giannotti, F., Malvaldi, M., The harsh rule of the goals: Data-driven performance indicators for football teams. In Proceedings of the 2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA), 1-10 (2015).

[6] Pappalardo L., Cintia P., Ferragina P., Massucco E., Pedreschi D., Giannotti F., PlayeRank: data-driven performance evaluation and player ranking in soccer via a machine learning approach, arXiv:1802.04987, 2018.