ISSN 1746-7659, England, UK
Journal of Information and Computing Science
Vol. 15, No. 2, 2020, *pp*.113-123

A Pattern Recognition and Performance Index Evaluation Model of Football Team based on Principal Component Analysis and PageRank Algorithm

Linjie Wu[1], Yujie Zheng[2] and Yunfei Fan[2]
[1]Department of Mathematics, Jinan University, Guangzhou, 510632, China
[2]Jinan University–University of Birmingham Joint Institute, Guangzhou, 510632, China
(Received May 10, 2020, accepted June 13, 2020)

Abstract. With the increasing knowledge integration and task complexity, individual ability demands a highly cohesive interdisciplinary team to amplify. To study the elements of successful team cooperation and explore valuable team strategies, this paper present a network pattern recognition model based on PageRank algorithm and principal component analysis method. Further, a team cooperation performance model based on group dynamic theory is built to capture the individual contribution and teamwork characteristic as a supplementary evaluation. By applying the model into football competition, we found our model has 73.68% accuracy, proving its outstanding adaptability. Based on the model, we can get the information of the inner network structure of a team, know the most significant contributors pertinent with team success, and make further justification plans and suggestions to achieve teamwork improvement.

Keywords: teamwork, network, cooperation performance, football

1. Introduction

In this era of more fierce competition and more miscellaneous challenges with high complexity, success is no longer only in favor of all-around individuals, but also more in favor of teams that are experts in division of work and cooperation. A successful team is always good at stimulating members' potentials and balancing their skills to work together as a whole to solve problems that are not attainable to an individual.

In competitive team sports, the significance of team strategies is self-evident. Football, particularly, requires 11 players and 3 substitutes, who are different in physical quality, technical expertise and carry out their own roles in the field to have extremely good cooperation between each other under the restraint of complicated rule and limited time, which is a rather demanding team work.

Due to high complexity of football and the changing situation of the game, finding a personalized but applicable way of football team cooperation is not only attractive but also challenging enough. Therefore, there is a need to build a specific controllable but adaptive model of team cooperation. This includes the analysis of the team's receiving and passing network, the establishment of group performance evaluation index system, we should also explore the value of the model in other team cooperation besides the controllable setting of such competitive team sports as football.

The Numbers Game [1] by Chris Anderson and David Sally studied how to use data analysis to understand the football game, arrange tactics and run club. At the end of the book, the authors predict that geometry "spatial, vector, triangular, and dynamic network" will be the focus of top-notch analysis. However, due to the complexity of football, the network analysis of football still needs to be developed for a long time.

E. Arriaza-Ardiles[2] analyzed the interactions between players in a football team from the point of view of graph theory and complex networks using data of a La Liga team. Clustering coefficient and centrality metrics (closeness and betweenness) are used to characterising the contribution of the players to the team. However, the paper only discusses the problem of passing without considering other contributions and lacks the analysis of game dynamics.

J. M. Buldú, J. Busquets, I. Echegoyen& F. Seirul.lo [3] analyzed the performance of Barcelona in the 2009-2010 season by using data on passing and possession. They focused on the temporal nature of football passing networks and calculated the evolution of all network properties along a match, instead of ignoring their average.

They found that the clustering coefficient, shortest path length, largest eigenvalue of the adjacency matrix, the algebraic connectivity and centrality distribution of Barcelona is better than the other teams. However, the variables selected for analysis in the paper may not be suitable for all teams. Similarly, the paper lacks sufficient research on the performance of teams over time. Further research is needed on which variables are more important for a particular style of team.

2. Models

2.1. Basic Model Description

In order to investigate the information of teamwork, we firstly use the pre-processed data of the a football team Huskies to establish a passing network and **Network Pattern Recognition Model** to learn its network features and then constructed a performance indicator system and Team **Cooperation Performance Model** to figure out key aspects for a team to success. After that, based on the result of models, we give several advice to the coach of the Huskies. Finally, we generalize the model out of the controlled setting of a team spot and into a normal kind of teamwork based on the **Group Dynamic theory**.

2.2. Data Pre-processing

2.2.1. Data Overview

The data is from problem D of The 2020 ICM(The Interdisciplinary Contest in Modeling) and can be downloaded on Internet. The Huskies have provided data with detailing information of last season, including all 38 games that they played against their 19 opponents. Overall, the raw data covers 23,429 passes between 366 players (30 Huskies players, and 336 players from opposing teams) and 59,271 game events (including duel, foul, free kick, goalkeeper leaving line, interruption, offside, others on the ball, pass, save attempt, shot and substitution).

2.2.2. Data Analysis and Pre-processing

Although most of the data is complete and reasonable, incomplete and abnormal data still often exists in a large amount of raw data, which may severely affect the efficiency of modeling and the accuracy of conclusions. It is therefore very important to pre-process the data. We listed the abnormal data and our corresponding pre-process methods and analysis as follows:

- Interference Data

After each shot, the coordinates of the destination of the ball are always (100, 100) or (0, 0). However, the goal situated at the middle near the edge of the field so that any destination of a successful shot (According to final score, it does exist) is impossible to be (100, 100) or (0, 0). Therefore these coordinates can be considered as interference data, which makes it difficult to identify which shot get scores and the deviation of each unsuccessful shot, which is a hinderance of modeling.

- Blank Data:

Blank missing data will affect the analysis process of the model, so it needs to be processed. We observe that the blank values in the data file generally appear in the four columns “EventOrigin_x”, “EventOrigin_y”, “EventDestination_x”, and “EventDestination_y”. Then we classified them into two kinds to tackle：

a) For the blank value due to data missing, we supplement it with the coordinate change before and after it. (Interpolation method)
b) For blank values due to "Substitution" such as column I, J, K, L in row 752 of “fullevents.csv”, , we will not fill it.

- Outliers

If one value in a set of data is more than twice the standard deviation of the average value, we call it the outlier. Statistically, we can use **block diagrams** to identify outliers. For the outliers, we use the average of the two adjacent observations to correct them.

- Wrong classification

Some data are classified wrongly in “fullevents.csv” and “passingevents.csv”, which leads to wrong selection of data if following the original classification standard. For example, in “fullevents.csv”, passes begin with Huskies_M8 were mistakenly classified as “Opponent7” in the column “team ID”（Obviously, it should be “Huskies”） in the 29th game of last season. We corrected these wrong classification.