Engineering Applications of Artificial Intelligence 72 (2018) 228–240

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai

Using social network analysis and gradient boosting to develop a soccer win–lose prediction model

Yoonjae Cho, Jaewoong Yoon, Sukjun Lee *

Business School, Kwangwoon University, 26 Kwangwoon-gil, Nowon-Gu, Seoul 139-701, South Korea

ARTICLE INFO

Keywords:
Social network analysis
Gradient boosting
Win–lose prediction
Soccer
Machine learning technique

ABSTRACT

We present the conceptual framework of a soccer win–lose prediction system (*SWLPS*) focused on passing distribution data (which is a representative characteristic of soccer) using social network analysis (SNA) and gradient boosting (GB). The general purpose of soccer predictions is to help the field supervisor design a strategy to win subsequent games using the derived information to improve and expand the coaching process. To implement and evaluate the proposed *SWLPS*, actual network indicators and predicted network indicators are generated using passing distribution data and SNA. The win–lose prediction is conducted using the GB machine learning technique. The performance of the *SWLPS* is analyzed through comparison with various machine learning techniques (i.e., support vector machine (SVM), neural network (NN), decision tree (DT), case-based reasoning (CBR), and logistic regression (LR)). The experimental results and analyses demonstrate that the network indicators generated through SNA can represent soccer team performance and that an accurate win–lose prediction system can be developed using GB technique.

1. Introduction

Social network analysis (SNA) has been widely used in various academic fields (e.g., sociology, psychology, anthropology, biology, and medicine). SNA focuses on the structure of ties within a set of social actors, such as persons, groups, organizations, and nations or the products of human activity. Thus, SNA is used to define the structure to be analyzed as a function of the relationship with the participant and to understand the structure of the network based on graph theory, algebra, statistical probability, and simulation (Galaskiewicz and Wasserman, 1993; Wasserman and Faust, 1994).

Nixon (1993) first applied the SNA approach to understand interactions, relationships, and structural changes in sports. This use of SNA was highly acclaimed because of its ability to quantify the attack type and the efficiency of the team. In recent years, many studies have been conducted on the application of SNA for sports, including soccer, basketball, handball, and hockey, in which the individual link structure is a pass. In particular, soccer (one of the most popular sports and the most representative team sport) can be considered a complex and dynamic system that evolves based on the interactions of multiple factors (Gréhaigne et al., 1997). In soccer, the most consequential form of interaction is clearly defined as the pass. In particular, the setting of soccer allows for a direct assessment of the interactions among team members. One of the advantages of investigating soccer is that the boundaries of the teams and the possible interactions of team members are clearly defined (Grund, 2012). In past research on soccer games, it has been considered challenging to predict victory or to quantify the contents of a game since the result of the competition can be changed due to complex factors (e.g., psychological, physiological, and environmental factors) affecting the player (Hughes and Franks, 2004). In addition, since the performance evaluation is determined by the supervisor and the subjective evaluation criteria of the coach, it is necessary to evaluate the nature of the competition through scientific analysis. The general purpose of sports predictions is to help the field supervisor design a strategy to win subsequent games based on the team performance evaluation. In addition, the derived information contributes to improve and expand the coaching process (Gonçalves et al., 2017). Thus, a match result prediction system must be built that considers the characteristics of the sports event and the team's performance.

Here, we propose a conceptual framework for a soccer win–lose prediction system (*SWLPS*) based on SNA and the gradient boosting (GB) machine learning technique; the system is designed for win–lose prediction specifically for Champions League (CL) soccer games. SNA, a method for analyzing organizational performance (Cross et al., 2002), is used to extract network indicators from passing distribution data to

* Corresponding author.
E-mail addresses: yoonjae@kw.ac.kr (Y. Cho), yjw8860@kw.ac.kr (J. Yoon), sjlee@kw.ac.kr (S. Lee).

https://doi.org/10.1016/j.engappai.2018.04.010
Received 13 January 2018; Received in revised form 29 March 2018; Accepted 9 April 2018
Available online 8 May 2018

evaluate how the team's performance affects the likelihood of winning. In the system, network indicators replace data obtained from notational analysis, which has often been used as input in previous research. In fact, passing distribution data can also belong to the category of data obtained via the notational analysis method. However, network indicators are different in that the passing distribution simply means the number of passes, but network indicators measure a team's performance. Thus, network indicators are more suitable than data from notational analysis for prediction of the outcome of a game. Additionally, the system employs GB, which is a family of powerful machine-learning techniques that have had considerable success in a wide range of practical applications (Natekin and Knoll, 2013). In particular, GB performs well in settings in which the number of variables exceeds the number of samples (high-dimensional data) (Lusa, 2017). The number of data points (samples) in this study is insufficient because the game prediction is performed for each season and round. As a result, the system adopting GB yielded excellent performance in game prediction.

We also implement an analysis of variance (ANOVA) analysis to identify the performance of the *SWLPS* through comparison of the performance of various classifiers (i.e., support vector machine (SVM), neural network (NN), decision tree (DT), case-based reasoning (CBR), and logistic regression (LR)) in terms of machine learning techniques for result predictions.

There are no studies related to game result prediction that combine SNA and machine learning. Previous studies in the sports field have usually predicted games using notational analysis data. Among the components of our system that predict the results of future games, there are two important elements that determine the performance of the system. The first element is the generation of appropriate input variables that are able to represent the performance of each team. The second is the compatibility of the classifier that learns the input and predicts the result.

The remainder of this paper is organized as follows. Section 2 briefly introduces the analysis method of sports performance, and Section 3 presents the construction procedure for the *SWLPS*. Section 4 presents the results of an empirical study performed to verify the performance of the *SWLPS*. Finally, conclusions are presented in Section 5.

2. Related work and limitations

Methods for sports performance analysis can be divided into notational analysis, SNA, and result prediction. The details are as follows.

2.1. Notational analysis

Sports performance analysis has been conducted primarily to analyze performance indicators generated through notational analysis, which was first proposed by Charles Reep in 1950 (Pollard, 2002) for football. Notational analysis is a technique for producing a permanent record of the events pertaining to a sporting event and is widely used by sports teams and individuals of various standards (James, 2006). One of the earliest empirical notational analyses on sports was conducted at Reilly and Thomas (1976). They analyzed the number of shots relative to goals scored and considered every move during a soccer game, including the intensity and duration of actions. Thus, they analyzed the computerized notation in terms of evolution. Hughes and Franks (2005) examined the length of passing sequences, the number of passes and team performance using notational analysis, and they found that longer passing sequences produced more goals per possession than shorter passing sequences for successful teams. Lago and Martín (2007) investigated ball possession in soccer and found that determinants affecting possession included the match status (e.g., winning, losing, or drawing), team status (e.g., home team or away team) and style of play. Lago-Ballesteros and Lago-Peñas (2010) analyzed the performance of soccer teams and found specific performance indicators that could be used to discriminate the top teams from the others. They also presented parameters to be used as normative data to collectively design and evaluate practices and competitions to establish peak-performance soccer teams.

2.2. SNA

SNA, which is not a formal theory in sociology but rather a strategy for investigating social structures (Otte and Rousseau, 2002), is a potentially useful method for sports performance analysis. Recently, SNA has received more attention than traditional notational analysis in this area. One of the earliest studies that applied SNA to team sports was conducted by Nixon (1993), who concluded that SNA can provide important insights into the leadership structure of sports teams. However, the few studies that have used SNA in sports settings have focused only on the cognitive or actual interaction between the players during the game (Bourbousson et al., 2010; Cotta et al., 2013; Passos et al., 2011). Grund (2012) studied the issue of within-team network structures and the performance of teams through an analysis of panel data. They found that networks characterized by high intensity and low centralization are indeed associated with better team performance. Pena and Touchette (2012) analyzed the strategy of soccer teams using network theory. They proposed an analysis method to discover the play patterns of each team, including hot spots, potential weaknesses, and the relative importance of each player in games. Cotta et al. (2013) investigated the use of simple graphs and network metrics to analyze the performance and play styles of the Spanish national football team in the World Cup 2010 and explained the results obtained at the complex network level. Fransen et al. (2015) used SNA to provide insight into the leadership structures within sports teams and found that SNA is a valuable tool in this regard. Clemente et al. (2015a) proposed a set of network methods to measure the specific properties of a team and found that network metrics can be a powerful tool to help coaches understand a team's specific properties and support their decision-making to improve the sports training process. Clemente et al. (2015b) analyzed team members' cooperation in basketball using centrality metrics of networks. They found that the specific point guard position is the most prominent position and that SNA is a useful approach to identify the patterns of interactions in basketball.

2.3. Result prediction

In the field of sports, there have been many studies on predicting the outcome of the game and analyzing performance. Koning (2000) and Koning et al. (2003) developed a model that used little prior knowledge and information and was heavily based on pure statistical models, such as ordered probit and Poisson models. Based on the models, they calculated the probability of winning for each team and predicted the most likely winner of a tournament. Rotshtein et al. (2005) developed a model for predicting the result of a football match. In that study, they analyzed the previous results of both teams and tuned fuzzy rules using genetic and neural optimization techniques. Huang and Chang (2010) developed a soccer prediction model using the multilayer perceptron, the backpropagation learning rule and the relative ratio values transformed from game records to be used as input data. The accuracy of the developed model was 76.9%, excluding draws. Halicioğlu (2011) developed a ranking system using the seasonal coefficients of variations of the end of season points and predicted the winner of the league. Snyder (2013) developed a soccer prediction model aimed at betting strategies. They used statistical methods (e.g., the Poisson distribution and multinomial regression), game records (e.g., frequency counts of events and statistical game records), and betting odds offered by various bookmakers and presented the approximate optimal betting strategy for use in simultaneous betting on multiple games with mutually exclusive outcomes.

2.4. Limitations of previous research

These previous studies have suffered from two limitations. First, the studies that have used SNA have analyzed predominantly team performances, strategies, or key players. In addition, they have identified