Expert Systems With Applications 44 (2016) 58–66

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Modeling basketball play-by-play data

Petar Vračar*, Erik Štrumbelj, Igor Kononenko

Faculty of Computer and Information Science, Večna pot 113, Ljubljana 1000 Slovenia

ARTICLE INFO

Keywords:
Forecasting
NBA
Logistic regression
Decision tree
Markov process

ABSTRACT

We present a methodology for generating a plausible simulation of a basketball match between two distinct teams as a sequence of team-level play-by-play in-game events. The methodology facilitates simple inclusion into any expert system and decision-making process that requires the performance evaluation of teams under various scenarios. Simulations are generated using a random walk through a state space whose states represent the in-game events of interest. The main idea of our approach is to extend the state description to capture the current context in the progression of a game. Apart from the in-game event label, the extended state description also includes game time, the points difference, and the opposing teams' characteristics. By doing so, the model's transition probabilities become conditional on a broader game context (and not solely on the current in-game event), which brings several advantages: it provides a means to infer the teams' specific behavior in relation to their characteristics, and to mitigate the intrinsic non-homogeneity of the progression of a basketball game (which is especially evident near the end of the game). To simplify the modeling of the transition distribution, we factorize it into terms that can be estimated with separate models. We applied the presented methodology to three seasons of National Basketball Association (NBA) games. Empirical evaluation shows that the proposed model outperforms the state-of-the-art in terms of forecasting accuracy and in terms of the plausibility of the generated simulations.

1. Introduction

Statistics and mathematical modeling have become an important part of sports and a lot of effort is dedicated to predicting the outcomes of sporting events (Percy, 2015; Stekler, Sendor, & Verlander, 2010). In this paper we focus on a more general task of sports outcome forecasting, where the goal is to predict not only the outcome, but also a more detailed evolution of the sporting event.

One of the most significant changes in the past decade has been the introduction and growing public availability of play-by-play data. In particular, in basketball. Compared to the more traditional box-score summary statistics, play-by-play data offer a richer description of within-match events (see Tables 1 and 2).

Often, such data are used for modeling the outcome of sports matches. Previous experience suggests that bookmaker odds are the best source of probabilistic forecasts for sports matches (Forrest, Goddard, & Simmons, 2005; Song, Boulier, & Stekler, 2007; Spann & Skiera, 2009). It appears that predictions from statistical and other types of prediction models are less accurate than those of fixed-price bookmakers and betting exchanges (which perform better not only because they use all publicly available information but also because of the wisdom-of-the-crowds effect). On the other hand, statistical models can be used to generate credible simulations of the likely progression of a sports match between two specific opponents.

The current state-of-the-art methods produce simulations that reflect the actual teams' win probabilities. However, the simulations are limited to reproducing only one aspect of the game (e.g. scoring statistics) or, in the case of more detailed simulations, exhibit some flaws in their credibility. The main reason for this disadvantage is that the models do not take into account the current context of the game, which affects the dynamics of the game.

We propose a methodology for learning from play-by-play data that deals with the issues outlined above. We assume the Markov property and model state transitions with a Logistic regression model, similar to Štrumbelj and Vračar (2012). We deal with non-homogeneity of the progression of a basketball game by incorporating relevant variables (time, point difference, …) into the state space of the model. The model also includes the transition time, which we model conditional to the predicted transition type. We show that our approach leads to a more credible simulation and more accurate forecasts of game outcomes.

The presence of such a detailed simulator can be beneficial for the further development of expert systems in sports. Having already

* Corresponding author. Tel.: +38631328933.
E-mail addresses: petar.vracar@fri.uni-lj.si (P. Vračar), erik.strumbelj@fri.uni-lj.si (E. Štrumbelj), igor.kononenko@fri.uni-lj.si (I. Kononenko).

http://dx.doi.org/10.1016/j.eswa.2015.09.004
0957-4174/

Table 1
An excerpt from the first quarter of an NBA game between Denver (Away) and Orlando (Home) 2010-03-28.

Time	Team	Player	Action	Score	Quarter
7:21	DEN	N Hilario	2 Point Field Goal	DEN 10–8	1
7:05	ORL	M Barnes	2 Point Miss	DEN 10–8	1
7:04	DEN	N Hilario	Defensive Rebound	DEN 10–8	1
6:54	DEN	J Petro	2 Point Miss	DEN 10–8	1
6:54	DEN	Team	Offensive Rebound	DEN 10–8	1
6:51	DEN	N Hilario	2 Point Miss	DEN 10–8	1
6:48	DEN	A Afflalo	Offensive Rebound	DEN 10–8	1
6:40	ORL	D Howard	Foul	DEN 10–8	1
6:40	DEN	N Hilario	1 Point Free Throw	DEN 11–8	1
6:40	DEN	N Hilario	Missed Free Throw	DEN 11–8	1
6:38	ORL	M Barnes	Defensive Rebound	DEN 11–8	1

Table 2
Box-score data for the NBA game from Table 1. FG - field goals, FGA - field goal attempts, FG% - field goal percentage, 3P - 3-point field goals, 3PA - 3-point field goal attempts, 3P% - 3-point field goal percentage, FT - free throws, FTA - free throw attempts, FT% - free throw percentage, ORB - offensive rebounds, DRB - defensive rebounds, TRB - total rebounds, AST - assists, STL - steals, BLK - blocks, TOV - turnovers, PF - personal fouls, PTS - points.

Team	FG	FGA	FG%	3P	3PA	3P%	FT	FTA	FT%
DEN	42	80	.525	5	16	.313	8	11	.727
ORL	39	79	.494	11	29	.379	14	22	.636
Team	ORB	DRB	TRB	AST	STL	BLK	TOV	PF	PTS
DEN	9	30	39	18	3	3	9	21	97
ORL	10	27	37	25	3	4	9	14	103

been used in several decision-making processes (Ballı & Korukoğlu, 2014; Dadelo, Turskis, Zavadskas, & Dadeliene, 2014; Papić, Rogulj, & Pleština, 2009), such expert systems can incorporate the simulator to provide coaches with insight into the performance of their teams during a game. The simulator would act as a sandbox for studying scenarios which may arise in different circumstances on the playing field/court. This functionality will help coaches understand how a team can increase their odds of winning, how individual playing skills affect team performance, and what performance can be expected using different approaches. Through further analysis of the simulation models (e.g. using the method of Štrumbelj & Kononenko (2010)), an expert system can provide its users with an explanation of how different factors affect the possible outcomes of a sequence of events, which could extract new knowledge about basic principles of the sport in question.

The rest of the paper is structured as follows. In the next section we briefly present the related work. In Section 3 we describe our approaches to modeling NBA Play-by-Play data. Section 4 provides the results of the experimental evaluation. The paper concludes with Section 5 where we discuss the results and give some ideas for further work.

2. Related work

Play-by-Play data are now routinely captured for most major team sports and competitions. However, the appropriate tools for modeling and simulating play-by-play data have not been developed. Most related work on modeling the progression of a sports game was done in basketball, which is one of the sports most suitable for such analyses, because of a high frequency of relevant events and better availability of historical statistical data. Stern (1994) used a Brownian motion process to model the evolution of a basketball score between the Home and Away teams, using 1st quarter, half-time, 3rd quarter and final scores. Goldman and Rao (2012) used a similar model to study the effects of 'pressure' on players' performance. They modeled the sensitivity of the win probability to small changes in the score and used it as a proxy for the importance of a particular situation within a game.

The above studies model the progression of a basketball score but not within-game events. Shirley (2007) proposed a more detailed model. He used a homogeneous Markov model with states based on within-game events. In particular, which team has possession (Home or Away), how the possession was obtained, and how many points were scored on the previous possession. Štrumbelj and Vračar (2012) included team-specific variables to Shirley's to account for individual teams strength. The proposed model was good at predicting the final score of a game. However, the game of basketball is neither time-homogeneous nor points-difference homogeneous (see (Štrumbelj & Vračar, 2012) for details). Therefore, a homogeneous model is not realistic.

The duration of transitions between states (that is, of the time that passes between two within-match events) has received little attention. Shirley (2007) and Štrumbelj and Vračar (2012) did not model time explicitly. Instead, they used the average number of state transition to determine the length of a basketball game. Gabel and Redner (2012) built a computational random-walk model to describe several statistical properties of scoring in basketball games. They showed that the distribution of time elapsed between scoring events has an exponential tail. They also argued that the intrinsic strengths of teams play a small role in the random-walk picture of scoring. Merritt and Clauset (2014) modeled the scoring dynamics between featureless teams using two stochastic processes. The first process produces scoring events, while the second process determines which team wins the points. Relying exclusively on the observed patterns in scoring events, they fitted a generative model that accurately reproduces the observed dynamics in lead-sizes over the course of games in several team sports (American football, hockey, and basketball). They also found that the model is able to make highly accurate predictions of game outcomes, after observing only the first few scoring events of the game.

The recent availability of optical player tracking data has opened new prospects in sport modeling. Cervone, D'Amour, Bornn, and Goldsberry (2014) used the player locations on the court to provide a real-time estimation of the expected number of points (EPV) obtained by the end of a possession. They also discussed some potential applications of EPV in revealing novel insights into players' decision-making tendencies. Oh, Keshri, and Iyengar (2015) used diverse data sources (player-tracking data, team lineup data, and play-by-play game log data from the matches played in the 2013–2014 NBA season) to model the progression of a basketball match on individual player level. They treated a game as a random sequence of transitions between discrete states that represent individual players on the court, their actions, and event outcomes. The model simulates the ball movement of every play and subsequent game events based on the player level interaction. However, the simulations are limited to the observed lineups only, since the methodology provides no means of inferring transition probabilities for a hypothetical lineup of players from different teams.

Therefore, while there are similar related works, our contribution is unique in that we facilitate the modeling of in-game events for two distinct teams. As such, a direct comparison is only possible with our previous work (Štrumbelj & Vračar, 2012), which we extend by proposing a model that allows us to incorporate broader in-game context (time, points difference) into the Markov state space, making the assumption of homogeneity more plausible.

3. Modeling NBA basketball data

A basketball game is a realization of a random process that depends on the underlying fundamentals of the sport and the characteristics of the competing teams. We can approximate this process and produce simulations by using a historical set of play-by-play data. We