Knowledge-Based Systems 36 (2012) 322–339

Contents lists available at SciVerse ScienceDirect

Knowledge-Based Systems

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

pi-football: A Bayesian network model for forecasting Association Football match outcomes

Anthony C. Constantinou *, Norman E. Fenton, Martin Neil

Risk Information Management (RIM) Research Group, Department of Electronic Engineering and Computer Science, Queen Mary University of London, London E1 4NS, UK

ARTICLE INFO

Article history:
Received 2 April 2012
Received in revised form 14 July 2012
Accepted 15 July 2012
Available online 1 August 2012

Keywords:
Bayesian reasoning
Football betting
Football predictions
Soccer predictions
Subjective information

ABSTRACT

A Bayesian network is a graphical probabilistic model that represents the conditional dependencies among uncertain variables, which can be both objective and subjective. We present a Bayesian network model for forecasting Association Football matches in which the subjective variables represent the factors that are important for prediction but which historical data fails to capture. The model (pi-football) was used to generate forecasts about the outcomes of the English Premier League (EPL) matches during season 2010/11 (but is easily extended to any football league). Forecasts were published online prior to the start of each match. We show that:

(a) using an appropriate measure of forecast accuracy, the subjective information improved the model such that posterior forecasts were on par with bookmakers' performance;
(b) using a standard profitability measure with discrepancy levels at $\geqslant 5\%$, the model generates profit under maximum, mean, and common bookmakers' odds, even allowing for the bookmakers' built-in profit margin.

Hence, compared with other published football forecast models, pi-football not only appears to be exceptionally accurate, but it can also be used to 'beat the bookies'.

1. Introduction

Association Football (hereafter referred to simply as 'football') is the world's most popular sport [11,43,12], and constitutes the fastest growing gambling market [7]. As a result, researchers continue to introduce a variety of football models which are formulated by diverse forecast methodologies. While some of these focus on predicting tournament outcomes [36,4,35,26,27] or league positions [34], our interest is in predicting outcomes of individual matches.

A common approach is the Poisson distribution goal-based data analysis whereby match results are generated by the attack and defence parameters of the two competing teams [41,9,38,32]. A similar version is also reported in [10] where the authors demonstrate profitability against the market only at very high levels of discrepancy, but which relies on small quantities of bets against an unspecified bookmaker. A time-varying Poisson distribution version was proposed by [53] in which the authors demonstrate profitability against Intertops (a bookmaker located in Antigua, West Indies), and refinements of this technique were later proposed in [8] which allow for a computationally less demanding model.

In contrast to the Poisson models that predict the number of goals scored and conceded, all other models restrict their predictions to match result, i.e. win, draw, or lose. Typically these are ordered probit regression models that consist of different explanatory variables. For example, [37] considered team performance data as well as published bookmakers' odds, whereas [24,22] considered team quality, recent performance, match significance and geographical distance. Ref. [23] compared goal-driven models with models that only consider match results and concluded that both versions generate similar predictions.

Techniques from the field of machine learning have also been proposed for prediction. In [55], the authors claimed that a genetic programming based technique was superior in predicting football outcomes to other two methods based on fuzzy models and neural networks. More recently, [52] claimed that acceptable match simulation results can be obtained by tuning fuzzy rules using parameters of fuzzy-term membership functions and rule weights by a combination of genetic and neural optimisation techniques.

Models based on team quality ratings have also been considered, but they do not appear to have been extensively evaluated. Knorr-Held [33] used a dynamic cumulative link model to generate ratings for top division football teams in Germany. The ELO rating

* Corresponding author.
E-mail addresses: anthony@constantinou.info (A.C. Constantinou), norman@eecs.qmul.ac.uk (N.E. Fenton), martin.neil@eecs.qmul.ac.uk (M. Neil).

0950-7051/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.knosys.2012.07.008

that was initially developed for assessing the strength of chess players [13] has been adopted to football [3]. In [29], the authors used the ELO rating for match predictions and concluded that the ratings appeared to be useful in encoding the information of past results for measuring the strength of a team, but the forecasts generated were not on par with market odds. Ref. [40] have also assessed an ELO rating based model along with the FIFA/Cocal Cola World rating model and concluded that both were inferior against bookmakers' forecasts for EURO 2008.

Numerous studies have considered the impact of specific factors on match outcome. These factors include: home advantage [28], ball possession [28], and red cards [51,56].[1]

Recently researchers have considered Bayesian networks and subjective information for football match predictions. In particular, [31] demonstrated the importance of supplementing data with expert judgement by showing that an expert constructed Bayesian network model was more accurate in generating football match forecasts for matches involving Tottenham Hotspurt than machine learners of MC4, naive Bayes, Bayesian learning and K-nearest neighbour. A model that combined a Bayesian network along with a rule-based reasoner appeared to provide reasonable World Cup forecasts in [42] through simulating various predifined strategies along with subjective information, whereas in [2] a hierarchical Bayesian network model that did not incorporate subjective judgments appeared to be inferior in predicting football results when compared to standard Poisson distribution models.

In this paper we present a new Bayesian network model for forecasting the outcomes of football matches in the distribution form of {p(H), p(D), p(A)}; corresponding to home win, draw and away win. We believe this study is important for the following reasons:

(a) the model is profitable under maximum, mean and common bookmakers' odds, even by allowing for the bookmakers' introduced profit margin;
(b) the model priors are dependent on statistics derived from predetermined scales of team-strength, rather than statistics derived from a particular team (hence enabling us to maximise historical data);
(c) the model enables us to revise forecasts from objective data, by incorporating subjective information for important factors that are not captured in the historical data;
(d) the significance of recent information (objective or subjective) is weighted using degrees of uncertainty resulting in a non-symmetric Bayesian parameter learning procedure;
(e) forecasts were published online before the start of each match [49];
(f) although the model has so far been applied for one league (the English Premier League) it is easily applicable to any other football league.

The paper is organised as follows: section 2 describes the historical data and method used to inform the model priors, section 3 describes the Bayesian network model, section 4 describes the assessment methods and section 5 provides our concluding remarks and future work.

2. Data

The basic data used to inform the priors for the model were the results (home, draw or away) of all English Premier League (EPL) matches from season 1993/94 to 2009/10 inclusive (a total of 6244 occurrences). This information is available online at [17]. The forecasts generated by the model were for season 2010/11, a total of 380 EPL matches.

Table 1
Predetermined levels of team strength.

Total points	>84	80–84	75–79	70–74	65–69	...(intervals of 5 points)	30–34	25–29	<25
Strength	1	2	3	4	5	...	12	13	14

In contrast to previous approaches we use the historical data to generate prior forecasts that are 'anonymous' by using predetermined levels of team-strength, rather than distinct team-names. We achieve this by replacing each team-name in each match in the database with a ranked number that represents the strength of that particular team for a particular season. The team-strength number is derived from the total number of points[2] that the particular team achieved during that particular season as shown in Table 1.

This implies that the same team may receive different ranks for different seasons and that different teams may receive identical ranks within the same season.

For example, the Manchester City at home to Aston Villa match in season 2006–2007 is classified as a ranked 10 versus a ranked 8 team (because in that season Manchester City totalled 42 points and Aston Villa 50 points), whereas in season 2009–2010 the Manchester City at home to Aston Villa match is classified as a ranked 5 versus a ranked 6 team (because in that season Manchester City totalled 67 points and Aston Villa 64 points).

The granularity (of 14 levels of team strength) has been chosen to ensure that for any match combination (i.e. a team of strength x at home to a team of strength y) there are sufficient data points for a reasonably well informed prior for {p(H), p(D), p(A)}. This approach has a number of important advantages:

(a) it enables us to make maximum use of limited data and be able to deal with the fact that every season the set of 20 teams changes (three are relegated and three new teams are promoted). For example, forecasts for teams for which there is little or no historical data (such as those recently promoted) are based on data for different teams but of similar strength;
(b) historical observations do not have to be ignored or weighted since the challenge here is to estimate a team's current strength and *learn* how such a team performed in the past given the specified ground (home/away) and opponent's strength. For example, consider the prior for the Manchester City at home to Aston Villa match in season 2010–2011. Because the historical performances of Manchester City and Aston Villa prior to season 2010–2011 were in no way representative of their strength in season 2010–2011, what matters is not the results of previous matches between Manchester City and Aston Villa (which would be sparse as well as irrelevant), but the results of all previous matches where a rank 4 team played at home to a rank 9 team.
(c) historical observations do not necessarily require weekly updating. The database already consists of thousands of historical match observations, and adding a few more matches every week will not make a major difference (this can be done once a year).
(d) historical observations from one league can be used to pre-

[1] While this work falls within the scope of our interest, other empirical forecasting studies such as attendance demand [46–48,15,20], and the effectiveness of football tipsters [18] do not.

[2] In EPL a total of 20 football teams participate and thus, a team can accumulate a minimum of 0 and a maximum of 114 points.