A Graph-based prediction model with applications

[Extended Abstract]

András London[*]
University of Szeged, Institute of Informatics
Poznan University of Economics, Department of Operations Research

József Németh
University of Szeged, Institute of Informatics

Miklós Krész
InnoRenew CoE
University of Primorska, IAM
University of Szeged, Institute of Applied Sciences

ABSTRACT

We present a new model for probabilistic forecasting using graph-based rating method. We provide a “forward-looking” type graph-based approach and apply it to predict football game outcomes by simply using the historical game results data of the investigated competition. The assumption of our model is that the rating of the teams after a game day correctly reflects the actual relative performance of them. We consider that the smaller the changing of the rating vector – contains the ratings of each team – after a certain outcome in an upcoming single game, the higher the probability of that outcome. Performing experiments on European football championships data, we can observe that the model performs well in general and outperforms some of the advanced versions of the widely-used Bradley-Terry model in many cases in terms of predictive accuracy. Although the application we present here is special, we note that our method can be applied to forecast general graph processes.

Categories and Subject Descriptors

I.6 [**Simulation and Modeling**]: Applications; I.2 [**Artificial Intelligence**]: Learning

1. INTRODUCTION

The problem of assigning scores to a set of individuals based on their pairwise comparisons appears in many areas and activities. For example in sports, players or teams are ranked according to the outcomes of games that they played; the impact of scientific publications can be measured using the relations among their citations. Web search engines rank websites based on their hyperlink structure. The centrality of individuals in social systems can also be evaluated according to their social relations. Ranking of individuals based on the underlying graph that models their bilateral relations has become the central ingredient of Google’s search engine and later it appeared in many areas from social network analysis to optimization in technical networks (e.g. road and electric networks) [16].

Making predictions in general, and especially in sports as well, is a difficult task. The predictions generally appear in the form of betting odds, that, in the case of “fixed odds”, provide a fairly acceptable source of expert’s predictions regarding sport games outcomes [21]. Thanks to the increasing quantity of available data the statistical ranking, rating and prediction methods have become more dominant in sports in the last decade. A key question is that how accurate these evaluations are, more concretely, the outcomes of the upcoming games how accurately can be predicted based on the statistics, ratings and forecasting models in hand.

Statistics-based forecasting models are used to predict the outcome of games based on some relevant information of the competing teams and/or players of the teams. A detailed survey of the scientific literature of rating and forecasting methods in sports is beyond the scope of this paper, we refer only some important and recent results in the topic. For some papers with detailed literature overview and sport applications of the the celebrated *Bradley-Terry model* [3], see e.g. [5, 7, 24]). Other popular approach is the Poisson goal-distribution based analysis. For some references, see for instance [10, 15, 20]. In these models the goals scored by the playing teams follow a Poisson distribution with parameter that is a function of attack and defense “rate” of the respective teams. A large family of prediction models only consider the game results win, loss (and tie) and usually uses some probit regression model, for instance [11] and [13]. More recently, well-known data mining techniques, like artificial neural networks, decision trees and support vector machines have also become very popular; some references - without being exhaustive - see e.g [8, 9, 14, 18].Based on the huge literature it can be concluded that the prediction accuracy strongly depends on the investigated sport and the feature set of the machine learning algorithms used. A notable part of prediction models based on the historical data of game results use the methodology of ranking and rating. Some recent articles in the topic are e.g. [2, 6, 12, 17, 23]. Specifically highlighting [2] the authors analyzed the he predictive power of eight sports ranking methods using only win-loss and score difference data of American major sports. They found that the least squares and random walker meth-

[*]Corresponding author, email: london@inf.u-szeged.hu

ods have significantly better predictive accuracy than other methods. Moreover, utilizing score-differential data are usually more predictive than those using only win-loss data.

In contrast to those techniques that use the actual respective strength of the two competing teams, we provide a graph-based and forward-looking type approach. The assumption of our model is that if a rating of the teams after a game day correctly reflects the actual relative performance, then the smaller the change in that rating after a certain result occurs (in an upcoming single game) the higher the probability of that event occur.

The structure of this paper is follows. After presenting the classical approaches ("Betting Odds" and "The Bradley-Terry Model"), our new model is introduced. Then in Sec. 3 we present our preliminary experimental results, and finally in Sec. 4 we conclude and discuss some possible research directions.

2. MODELS

Let $V = (1, \dots, n)$ be the set of n *teams* (or players) and let R be the number of *game days* in a competition among the teams in V. A *rating* is a function $\phi^r : V \to \mathbb{R}^n$ that assigns a score to each team after each game day r ($r = 1, \dots, R$). This is considered as the quantitative "strength" of the teams. A *ranking* $\sigma^r : V \to V$, after game day r, is an ordering of the teams that is simply obtained by sorting the teams according to the rating ϕ^r. Using the game results data set, one can define a directed multigraph (i.e. a graph where multiple links are allowed), where nodes represent teams, while links between them represent outcomes of games they played. The links are directed and each of them is going from the loser team to the winning team. If ties are also considered they can be represented by two directed links with opposite directions and half weight. An edge weighting can be naturally considered if the final scores of the games are given

2.1 Betting Odds

Bookmakers determine *betting odds* for the games according to their expectations of outcome probabilities. Here we deal with fixed odds, means that they do not vary over time depending on the betting volumes. These "fixed-odds" represent the predictions of bookmakers [21]. The meaning of the betting odds for an upcoming game is the following: Assume that the betting odds between team i and team j are odds(i) and odds(j), respectively. It means that if one bets \$1 to i's win and it comes out, he wins odds(i) dollars, while if j wins, then the bettor loses his \$1. We can calculate the probabilities of the respective events as

$$\Pr(i \text{ beats } j) = \frac{1/\text{odds}(i)}{1/\text{odds}(i) + 1/\text{odds}(j)}$$

and

$$\Pr(j \text{ beats } i) = \frac{1/\text{odds}(j)}{1/\text{odds}(i) + 1/\text{odds}(j)}.$$

We should note here that odds provided by betting agencies do not represent the true chances (as imagined by the bookmaker) that the event will or will not occur, but are the amount that the bookmaker will pay out on a winning bet. The odds include a profit margin meaning that the payout to a successful bettor is less than that represented by the true chance of the event occurring. This means mathematically that $1/\text{odds}(i) + 1/\text{odds}(j)$ is more than one. This profit expected by the agency is known as the "over-round on the book".

2.2 The Bradley-Terry Model

The *Bradley-Terry model* [3] is a widely-used method to assign probabilities to the possible outcomes when a set of n individuals are repeatedly compared with each other in pairs. For two elements i and j, the probability that i beats j defined as

$$\Pr(i \text{ beats } j) = \frac{\pi_i}{\pi_i + \pi_j},$$

where $\pi_i > 0$ is a parameter associated to each individual $i = 1, \dots, n$, representing the overall skill, or "intrinsic strength" of it. Equivalently, π_i/π_j represents the odds in favor i beats j, therefore this is a "proportional-odds model". Suppose that i and j played N_{ij} games against each other with i winning W_{ij} of them, and all games are considered to be independent. The likelihood is given by

$$L(\pi_i, \dots, \pi_n) = \prod_{i<j} \left[\frac{\pi_i}{\pi_i + \pi_j}\right]^{W_{ij}} \left[\frac{\pi_j}{\pi_i + \pi_j}\right]^{N_{ij} - W_{ij}}.$$

Then the log-likelihood is

$$\begin{aligned}\ell(\pi_i, \dots, \pi_n) &= \sum_{1 \le i \ne j \le n} \left[W_{ij} \log \pi_i - W_{ij} \log(\pi_i - \pi_j)\right] \\ &= \sum_{i=1}^{n} W_{ij} \log \pi_i - \sum_{1 \le i < j \le n} N_{ij} \log(\pi_i + \pi_j)\end{aligned}$$

which need to be maximized.

One possible derivation of the model assumes team i produces an unobserved score S_i, no matter which is the opposing team, with the cumulative distribution function

$$S_i \sim F_i(s) = \exp[-e^{-(s - \log \pi_i)}].$$

It follows that distribution of the difference $S_i - S_j$ follows a logistic distribution function

$$S_i - S_j \sim F_{ij}(s) = \frac{1}{1 + e^{-(s - (\log \pi_i - \log \pi_j))}},$$

which implies that

$$\begin{aligned}\Pr(S_i > S_j) = \Pr(S_i - S_j > 0) &= 1 - \frac{1}{1 + e^{\log \pi_i - \log \pi_j}} \\ &= \frac{\pi_i}{\pi_i + \pi_j}.\end{aligned}$$

Extension with Home advantage and Tie. A natural extension of the Bradley-Terry model with "home-field advantage", according to [1], say, is to calculate the probabilities as

$$\Pr(i \text{ beats } j) = \begin{cases} \frac{\theta \pi_i}{\theta \pi_i + \pi_j}, & \text{if } i \text{ is at home} \\ \frac{\pi_i}{\pi_i + \theta \pi_j}, & \text{if } j \text{ is at home} \end{cases}$$

where $\theta > 0$ measures the strength of the home-field advantage (or disadvantage). Considering also a tie as a possible final result of a game, the following calculations, proposed in [22], can be used :

$$\Pr(i \text{ beats } j) = \frac{\pi_i}{\pi_i + \alpha\pi_j},$$

$$\Pr(i \text{ ties } j) = \frac{(\alpha^2 - 1)\pi_i\pi_j}{(\pi_i + \alpha\pi_j)(\alpha\pi_i + \pi_j)}$$

where $\alpha > 1$. Combining them is straightforward. In our experiments, we used the Matlab implementations found at `http://www.stats.ox.ac.uk/~caron/code/bayesbt/` using the *expectation maximization* algorithm, described in detail in [7].

2.3 Rating-based Model with Learning

Our new model is designed as follows. We will use the term "game day" in each case when at least one match is played on the given day. For any game day in which we make a forecast, we consider the results matrix that contains all the results of the previous $T = 40$ game days. For the 40 game days time window, the entries of the results matrix S are defined as $S_{ij} = \#\{$scores team home-i achieved against team away-j$\}$. To take into account the home-field effect, for each team i we distinguish team home-i and team away-i. Thus, we define a $2n \times 2n$ results matrix, which, in fact, describes a bipartite graph where each team appears both in the home team side and the away team side of the graph. For rating the teams, a time-dependent PageRank method is used. The PageRank scores are calculated according the time-dependent PageRank equation

$$\phi = \mathbf{\Pi} = \frac{\lambda}{N}[I - (1-\lambda)S^t_{mod}(\mathbf{1}\mathbb{1}^t)^{-1}]^{-1}\mathbb{1}, \qquad (1)$$

defined in [19]. The damping factor is $\lambda = 0.1$, while we may multiply each entry of S with the exponential function 0.98^α to consider time-dependency and obtaining S_{mod}, where α denotes the number of game days elapsed since a given result occurred (and stored in S). Note, that a home team and an away team PageRank values are calculated for each team. We would like to establish a connection between team home-i and team away-i using the assumption that home-i is not weaker than away-i. In our implementation we assumed that home-i had a win $2:1$ against away-i to give a positive bias for home-i at the beginning. In our experiments this setup performed well, but it was not optimized precisely.

Using the above-defined results matrix S and the PageRank rating vector ϕ, we assign probabilities to the outcomes $\{$home team win, tie, away team win$\}$ of an upcoming game in game day r between home-i and away-j as follows. Before the game day in which we make the forecast, let the calculated PageRank rating vector be $\phi_{40}^{r-1}(V)$. We use δ^r_{xy} to measure how the rating vector of the teams changes if the result of an upcoming game between teams i and j is $x : y$, where $x, y = 0, 1, \ldots$ are the scores achieved by team i and team j, respectively[1]. We define δ^r_{xy} as the Euclidean distance between $\phi_{40}^{r-1}(V)$ and $\phi_{40}^{r}(V)$ that is the rating vector for the new results matrix obtained by adding x to S_{ij} and y to $S_{n+j,i}$. In the results graph interpretation this simply means that an edge from node away-j to node home-i with weight x and an edge from node home-i to node away-j with weight y are added to the graph, respectively. Our assumption is that if an outcome $x : y$ has a high probability and it occurs, then it causes a small change in the PageRank vector; hence δ_{xy} will be small. To simplify the notations let $\{\delta_1, \ldots, \delta_m\}$ be the distance values obtained by considering different results $\{E_1, \ldots, E_m\}$ of the upcoming game between i and j. The goal now is to calculate the probability that a certain result occurs if $\{\delta_1, \ldots, \delta_m\}$ is given. To do this, we use the following simple statistics-based machine learning method. Let $f^+()$ be the probability density function of δ_i random variable where the event (game result) E_i occurred. In our implementation $E_i \in \{0:0, 1:0, 1:1, \ldots, 5:5\}$, assuming that the probability of other results equals 0. Similarly, let $f^-()$ be the probability density function of δ_i random variable in which case the event (game result) E_i did not occur. To approximate the $f^+()$ and $f^-()$ functions, for each game we use the training data set contains all results and related δ_i $(i = 1, \ldots, m)$ values of the preceding $T = 40$ game days of the considered game. In our experiments, the gamma distribution (and its density function) turned out to be a fairly good approximate for $f^+(\delta)$ and $f^-(\delta)$.

Assuming that $\delta_1, \ldots, \delta_m$ are independent, using the Bayes theorem and the law of total probability, we can calculate that

$$\Pr(E_i|\{\delta_1, \ldots, \delta_m\}) = \frac{f^+(\delta_i)\prod_{k \neq i} f^-(\delta_k)}{\sum_\ell f^+(\delta_\ell)\prod_{k \neq l} f^-(\delta_\ell)}.$$

We should note here that in this way we assign probabilities to concrete game final results, which is another novelty of our model. Then, for the upcoming game between i and j, the outcome probability of the event "i beats j" is calculated as

$$\Pr(i \text{ beats } j) = \sum_{\substack{k:\ E_k \text{ encodes a result} \\ \text{of team-i win}}} \Pr(E_k|\{\delta_1, \ldots, \delta_m\}),$$

where we sum over those E_k results for which i beats j (i.e. 1:0, 2:0, 2:1, 3:0, 3:1, etc.). The probabilities $\Pr(i$ ties $j)$ and $\Pr(j$ beats $i)$ can be calculated in a similar way.

3. EXPERIMENTAL RESULTS

To measure the accuracy of the forecasting we calculate the mean squared error, which is often called *Brier scoring rule* in the forecasting literature [4]. The Brier score measures the mean squared difference between the predicted probability assigned to the possible outcomes for event E and the actual outcome o_E. Suppose that for a single game g, between i and j, the forecast is $\mathbf{p}^g = (p^g_w, p^g_t, p^g_l)$ contains the probabilities of i wins, the game is a tie and i loses, respectively. Let the actual outcome of the game be $\mathbf{o}^g = (o^g_w, o^g_t, o^g_l)$, where exactly one element is 1, the other two are 0. Noting that the number of games played (and predicted) is N, BS is defined as

$$BS = \frac{1}{N}\sum_{g=1}^{N} ||\mathbf{p}^g - \mathbf{o}^g||_2^2$$

$$= \frac{1}{N}\sum_{g=1}^{N}[(p^g_w - o^g_w)^2 + (p^g_t - o^g_t)^2 + (p^g_l - o^g_l)^2].$$

[1] We should note here that if the result is $0:0$, then $x = y = 1/2$ is used.

The best score achievable is 0. In the case of three possible outcomes (win, lost, tie) we can easily see that the forecast $\mathbf{p}^g = (1/3, 1/3, 1/3)$ (for each game g and any N) gives accuracy $BS = 2/3 = 0.666$. We consider this value as a worst-case benchmark. One question of our investigation is that how better BS values can be achieved using our method, and how close we can get to the betting agencies' fairly good predictions.

The data set we used contained all final results of given seasons of some football leagues, listed in the first two column of Table 1. We tested our method as it was described in Sec. 2.3. We start predicting games starting from the 41th game day; for each single game predictions are made using the results of the previous 40 game day before that game. The Brier scores were calculated using all predictions we made. Our initial results are summarized in Table 1. To calculate the betting odds probabilities we used the betting odds provided by bet365 bookmaker available at `http://www.football-data.co.uk/`. We could see that these predictions gave the best accuracy score (BS) in each case. We highlighted the values where the difference between the Bradley-Terry method and the PageRank method was higher than 0.01. Although we can see that slightly more than half of the cases the Bradley-Terry model gives a better accuracy, the results are still promising considering the fact that the parameters of our method and the implementation are far from being optimized.

4. CONCLUSIONS

We presented a new model for probabilistic forecasting in sports, based on rating methods, that simply use the historical game results data of the given sport competition. We provided a forward-looking type graph based approach. The assumption of our model is that the rating of the teams after a game day is correctly reflects their current relative performance. We consider that the smaller the changing in the rating vector after a certain result occurs in an upcoming single game, the higher the probability that this event will occur. Performing experiments on results data sets of European football championships, we observed that this model performed well in general in terms of predictive accuracy. However, we should note here, that parameter fine tuning and optimizing certain parts of our implementation are tasks of future work.

We emphasize, that our methodology can be also useful to compare different rating methods by measuring that which one reflects better the actual strength (rating) of the teams according to our interpretation. Finally we should add that the model is general and may be used to investigate such graph processes where the number of nodes is fixed and edges are changing over time; moreover it also has a potential to link prediction.

5. ACKNOWLEDGMENTS

This work was partially supported by the National Research, Development and Innovation Office - NKFIH, SNN117879.

Miklós Krész acknowledges the European Commission for funding the InnoRenew CoE project (Grant Agreement #739574) under the Horizon2020 Widespread-Teaming program.

6. REFERENCES

[1] A. Agresti. *Categorical data analysis*. John Wiley & Sons, New York, 1996.

[2] D. Barrow, I. Drayer, P. Elliott, G. Gaut, and B. Osting. Ranking rankings: an empirical comparison of the predictive power of sports ranking methods. *Journal of Quantitative Analysis in Sports*, 9(2):187–202, 2013.

[3] R. A. Bradley and M. E. Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3-4):324–345, 1952.

[4] G. W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3, 1950.

[5] K. Butler and J. T. Whelan. The existence of maximum likelihood estimates in the Bradley-Terry model and its extensions. *arXiv preprint math/0412232*, 2004.

[6] T. Callaghan, P. J. Mucha, and M. A. Porter. Random walker ranking for NCAA division IA football. *American Mathematical Monthly*, 114(9):761–777, 2007.

[7] F. Caron and A. Doucet. Efficient bayesian inference for generalized Bradley–Terry models. *Journal of Computational and Graphical Statistics*, 21(1):174–196, 2012.

[8] A. C. Constantinou, N. E. Fenton, and M. Neil. Pi-football: A bayesian network model for forecasting association football match outcomes. *Knowledge-Based Systems*, 36:322–339, 2012.

[9] D. Delen, D. Cogdell, and N. Kasap. A comparative analysis of data mining methods in predicting NCAA bowl outcomes. *International Journal of Forecasting*, 28(2):543–552, 2012.

[10] M. J. Dixon and P. F. Pope. The value of statistical forecasts in the UK association football betting market. *International Journal of Forecasting*, 20(4):697–711, 2004.

[11] D. Forrest, J. Goddard, and R. Simmons. Odds-setters as forecasters: The case of English football. *International Journal of Forecasting*, 21(3):551–564, 2005.

[12] R. Gill and J. Keating. Assessing methods for college football rankings. *Journal of Quantitative Analysis in Sports*, 5(2), 2009.

[13] J. Goddard and I. Asimakopoulos. Forecasting football results and the efficiency of fixed-odds betting. *Journal of Forecasting*, 23(1):51–66, 2004.

[14] A. Joseph, N. E. Fenton, and M. Neil. Predicting football results using bayesian nets and other machine learning techniques. *Knowledge-Based Systems*, 19(7):544–553, 2006.

[15] D. Karlis and I. Ntzoufras. Analysis of sports data by using bivariate Poisson models. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 52(3):381–393, 2003.

[16] A. N. Langville and C. D. Meyer. *Google's PageRank and beyond: The science of search engine rankings*. Princeton University Press, 2011.

[17] J. Lasek, Z. Szlávik, and S. Bhulai. The predictive power of ranking systems in association football. *International Journal of Applied Pattern Recognition*, 1(1):27–46, 2013.

[18] C. K. Leung and K. W. Joseph. Sports data mining: Predicting results for the college football games. *Procedia Computer Science*, 35:710–719, 2014.

[19] A. London, J. Németh, and T. Németh. Time-dependent network algorithm for ranking in sports. *Acta Cybernetica*, 21(3):495–506, 2014.

[20] M. J. Maher. Modelling association football scores. *Statistica Neerlandica*, 36(3):109–118, 1982.

Table 1: Accuracy results on football data sets. The values where the difference between the Bradley-Terry method and the PageRank method was higher than 0.01 are shown in bold.

League	Season	Betting odds error	Bradley-Terry error	PageRank method error
Premier League	2011/12	0.58934	0.60864	**0.59653**
	2012/13	0.56461	0.59744	**0.58166**
	2013/14	0.54191	**0.55572**	0.59406
	2014/15	0.55740	0.60126	0.60966
Bundesliga	2011/12	0.58945	0.59994	**0.59097**
	2012/13	0.57448	0.59794	**0.58622**
	2013/14	0.55724	**0.57803**	0.60125
	2014/15	0.57268	0.60349	0.60604
La Liga	2011/12	0.54598	**0.57837**	0.58736
	2012/13	0.56417	**0.58916**	0.60205
	2013/14	0.57908	**0.58016**	0.60473
	2014/15	0.52317	0.55888	0.56172

[21] P. F. Pope and D. A. Peel. Information, prices and efficiency in a fixed-odds betting market. *Economica*, pages 323–341, 1989.

[22] P. Rao and L. L. Kupper. Ties in paired-comparison experiments: A generalization of the Bradley-Terry model. *Journal of the American Statistical Association*, 62(317):194–204, 1967.

[23] J. A. Trono. Rating/ranking systems, post-season bowl games, and 'the spread'. *Journal of Quantitative Analysis in Sports*, 6(3), 2010.

[24] C. Wang and M. L. Vandebroek. A model based ranking system for soccer teams. *Research report, available at SSRN 2273471*, 2013.