A Graph-based prediction model with applications

András London[1], Miklós Krész[2], József Németh[3]

Oct 11 2018

[1]University of Szeged (and also Poznan University of Economics and Business)
[2]InnoRenew CoE and University of Primorska (also at University of Szeged)
[3]University of Szeged

Introduction

- Topic: **rating and ranking**
  - entities: nodes of a graph [4]
  - methods: graph algorithms (vs statistical models)
- Main motivation: **making predictions**
  - given an evolving graph (now the nodes are fixed)
  - "forward-looking" type methods
  - graph-based learning
- Possible application: **sport** predictions
  - can be applied to general graph processes

[4] many applications: social networks, recommendation systems, decision making systems, web search, etc.

A general methodology

Notions

- nodes of the graph: $V = (1, \ldots, n)$; rounds/games $r = 1, \ldots R$
- rating after r round: $\phi^r : V \to \mathbb{R}$
- ranking is $\sigma : V \to V$ bijection according to ϕ

Data

- final results, as win–draw–loss
- final results with scores (goals/points)

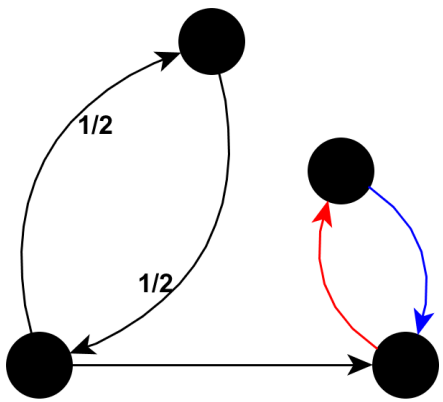
Datasets used

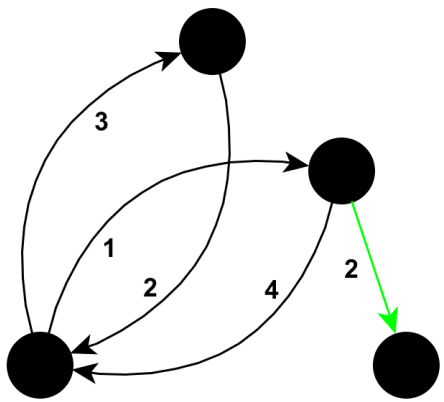
- final score results of several European football championships

Matrix / Graph representation

$$W_{ij} = \#(i \text{ won against } j) + \frac{1}{2}\#(\text{draw between } i \text{ and } j)$$

$$S_{ij} = \sum_{i \text{ and } j \text{ games}} \#(i \text{ scores against } j).$$





Forecasting

Goal: predicting game outcomes $\rightarrow$ "Probabilistic forecasting"

As a graph process it means, that predicting the weight of a novel link

How?

- assigning probabilities to the possible outcomes (weight of the links)
- using only the past results (current state of the graph)
- considering time-dependency, home field advantage, draws (in sport applications)
- without using experts opinions

Baseline 1: Betting odds

Odds are determined by betting agencies (experts) (in case of football they are fixed)

Example: Premier League, 2013, week 10 odds:

Aston Villa – Arsenal 6.00 4.20 1.62

$\frac{1}{6} + \frac{1}{4.2} + \frac{1}{1.62} - 1 = 1.02205 - 1 = 0.02205 \rightarrow$ "overround" (not a fair game)

Assigned probabilities: $\frac{1/\text{odds}(i)}{\sum_{i \in \{w,d,\ell\}} 1/\text{odds}(i)}$

Aston Villa – Arsenal 0.16 0.24 0.6

Baseline 2: The Bradley-Terry model[5]

For every team there is a parameter π_i – intrinsic strength

$$\Pr[i \text{ won against } j] = \frac{\pi_i}{\pi_i + \pi_j}$$

Estimate π_i parameter values $\rightarrow$ maximum likelihood

$$L(\pi_i, \ldots, \pi_n) = \prod_{i<j} \left[\frac{\pi_i}{\pi_i + \pi_j}\right]^{y_{ij}} \left[\frac{\pi_j}{\pi_i + \pi_j}\right]^{n_{ij} - y_{ij}} \rightarrow \max$$

n_{ij} and y_{ij} are the number of games between i and j and the number of games where i won, resp.

- existing extensions: draws, home-field, time dependency
- in experiments we used advanced versions and implementations

[5] Bradley & Terry, *Biometrika* (1952); earlier: Zermelo, *Fundamenta Mathematicae* (1929)

General forward-looking type models

- The rating vector after r rounds is $\phi^r(V) = (\phi_1^r, \ldots, \phi_n^r)$
- Consider the game between i and j in the next round

Main idea:

The **relative probability of an o_{ij} game result depends on that how the rating vector changed** by "adding" o_{ij} to the graph (i.e. new links with certain weights)

General model - example

- Use the Euclidean[6] distance between the old and new rating vector after game k (that we are predicting): δ^k
- then for a game between i and j

outcome for i	...	after game k	...
win	...	δ_w^k	...
draw	...	δ_d^k	...
loss	...	δ_ℓ^k	...

- "Training" period: use only results of the previous T game days before the predicted game k [**time window**]

[6] or something else

Model parameters

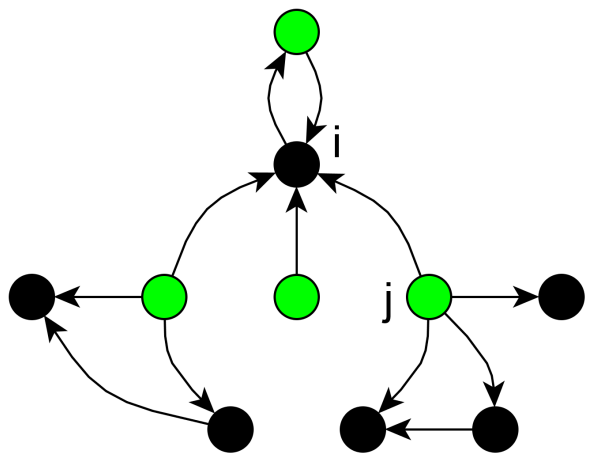
- What is the ϕ^r rating function?
- Which distance function is used?
- How long is the training period (i.e. time window)?
- How to determine the probabilities (based on δ's)?

Ratings and graphs – PageRank

Recursion:

$$PR(i) = \frac{1-\lambda}{n} + \lambda \sum_{j \in N^+(i)} \frac{PR(j)}{d^-(j)}$$

$N^+(i)$ is the out-degree of i, $\lambda \in [0, 1]$ is a parameter[7].



[7] it is not difficult to check that **PR** is the solution of an eigenvalue equation for a special eigenvalue

The proposed method

- Let $\phi^r = \mathbf{PR}^r$ be the PageRank vector [8]
- Euclidean distance between ratings
- Determine the probabilities (home team point of view)
  - based on δ rating vector distances

[8] Precisely a time-dependent variant defined by *London et al.* (2014)

Determine probabilities – technical

What is the distribution of δ_* for the occurred and non-occurred events? $\rightarrow$ learning (function fitting)

1. $\{\delta_1, \ldots, \delta_m\}$ – the distance values obtained by considering different results $\{E_1, \ldots, E_m\}$ of a game between i and j
2. now $E_i \in \{0:0, 1:0, 1:1, \ldots, 5:5\}$
3. f^+, f^- – the probability density function of δ_i (as a random variable) where the event (game result) E_i occurred or not occurred, resp.[9]

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

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

Finally we have

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

[9] we used the gamma function in experiments

Probabilities obtained by the PageRank model

Example for calculated probabilities (from 0 : 0 to 5 : 5) and forecasting error

```
P =

    0.0779    0.0736    0.0635    0.0509    0.0384    0.0275
    0.0591    0.0606    0.0563    0.0482    0.0385    0.0291
    0.0348    0.0384    0.0379    0.0343    0.0287    0.0226
    0.0170    0.0199    0.0207    0.0196    0.0171    0.0139
    0.0071    0.0088    0.0096    0.0094    0.0085    0.0071
    0.0027    0.0035    0.0039    0.0040    0.0037    0.0031

50 Aston Villa -     Arsenal  1-2  6.00 4.20 1.62
Bookmaker probability:  0.16 0.23 0.60  Error:0.24
PageRank probability:   0.24 0.21 0.55  Error:0.30
Expected win if PageRank is right: 0.46 -0.13 -0.11
```

Error= $(\text{real(win)-prob(win)})^2+(\text{real(draw)-prob(draw)})^2+(\text{real(loss)-prob(loss)})^2$

Experiments

Table: Accuracy (= average error) results on football data sets.

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

Conclusion, notes

Short summary:

1. New graph based model for forecasting
2. Applied well to football game predictions
   - without parameter fine-tuning

Future possibilities:

1. Can be useful to compare different rating/ranking methods
   - which reflects better the actual strength of teams (according to our definition)
2. the model is general and can be useful to investigate graph processes
   - new potentials in link prediction

Thank you for your attention!

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

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