Score-based soccer match outcome modeling – an experimental review

Ondřej Hubáček Gustav Šourek Filip Železný

Czech Technical University in Prague

The Motivation

- 2017 Soccer Prediction Challenge
  - in conjuction with MLJ's SI (Machine Learning for Soccer)
- **What is the best (score-based) model for soccer?**
- lack of accepted dataset led to incomparable results
- large dataset published recently
- match results widely available

Dataset & Models

Table: Sample of the dataset.

League	Season	Date	Home	Away	Score
ENG1	2003	10/6/2004	Arsenal	Chelsea	3 - 1
...	...	...	...	...	...
ITA2	2016	18/5/2016	Ascoli	Ternana	1 - 2

- Type of models
  - statistical models
  - rating systems
  - graph-based models

Double Poisson Model[6]

- The probability of match outcome is given by

$$P(G_H = x, G_A = y|\lambda_H, \lambda_A) = \frac{\lambda_H^x e^{-\lambda_H}}{x!} \cdot \frac{\lambda_A^y e^{-\lambda_A}}{y!}$$

- The scoring rates λ are given by

$$log(\lambda_H) = Str_H - Str_A + H$$
$$log(\lambda_A) = Str_A - Str_H$$

- The parameters Str, H are fitted optimizing log-likelihood

$$L = \prod (P(G_i^H = x, G_i^A = y|\lambda_i^H, \lambda_i^A) \cdot w_i)$$

Double Poisson Model

- \+ gives probability distribution over possible scores
- \+ only one metaparameter
- \+ can be used for other low-scoring sports/games
- assumes independence between score and conceded goals
- – Poisson dist. does not handle over/under dispersed data
- – needs to be reffited after each league round

PageRank[3]

- the leagues can be represented as graphs
  - nodes $\sim$ teams
  - edge $\sim$ matches
- Page P is linked from important pages $\implies$ the page P is important.
- Team T defeated strong teams $\implies$ the team T is strong.
- adjacency matrix given by:

$$M_{ij} = \frac{\sum_m PTS_j(m) \cdot w_m}{\sum_m w_m}$$

- refitted after each round

Elo Rating [5]

- originally developed for rating chess players
- models expected match outcome based on ratings discrepancy

$$E = \frac{1}{1 + c^{(R^A - R^H)/d}}$$

- updates based on actual outcome S and goal difference δ

$$R^H_{t+1} = R^H_t + k(1+\delta)^\gamma \cdot (S - E)$$
$$R^A_{t+1} = R^A_t - k(1+\delta)^\gamma \cdot (S - E)$$

$$S = \begin{cases} 1 & \text{if the home team won} \\ 0.5 & \text{if the match was drawn} \\ 0 & \text{if the home team lost} \end{cases}$$

Steph Ratings[7]

- winning solution from chess ratings competition @Kaggle
- extends another popular rating system (glicko)
- each player has a rating and its variance
- computation of expected outcome similar to Elo
- k factor depends on rating variance

Berrar Ratings[1]

- models goals scored instead of match outcome
- each team has *att* and *def* ratings

$$\hat{g}_h(att_H, def_A) = \frac{\alpha_h}{1 + \exp(-\beta_h(att_H + def_A) - \gamma_h)}$$
$$\hat{g}_a(att_A, def_H) = \frac{\alpha_a}{1 + \exp(-\beta_a(att_A + def_H) - \gamma_a)}$$

- updates ratings according to discrepancy from observed goals

$$att_H^{t+1} = att_H^t + \omega_{att}(g_h - \hat{g}_h)$$
$$def_H^{t+1} = def_H^t + \omega_{def}(g_a - \hat{g}_a)$$

pi-ratings[2]

- models expected goal difference $\hat{gd}$ instead of goals scored
- each team has *home* and *away* ratings
- simplified calculations:

$$\hat{gd}_H = 10^{R_H^{home}/C} - 1$$

$$\hat{gd}_A = 10^{R_A^{away}/C} - 1$$

$$\hat{gd} = \hat{gd}_H - \hat{gd}_A$$

$$\psi = log_{10}(1 + |gd - \hat{gd}|) \cdot C$$

$$R_H^{home} += \lambda \cdot \psi$$

$$R_H^{away} += \gamma \cdot \lambda \cdot \psi$$

Score-based TrueSkill™[4]

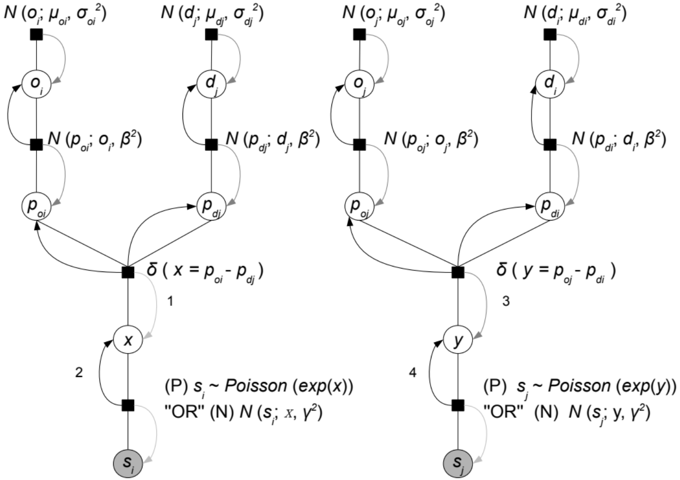

Figure: TrueSkill™model schema taken from the original paper.

From Ratings to Predictions

```python
def optimize_rating(data, res, params)
    ratings = compute_ratings(data, res, params)
    olr = OrderedLogisticRegression()
    olr = olr.fit(ratings, res)
    predictions = olr.predict_proba(ratings)
    loss = RPS(predictions, res)
    return loss.mean()
```

Experimental Setup

- seasons 2000/01-2008/09, 52 leagues
- we omit first season from each league and first 5 rounds of each season
- over 84 000 matches after filtering
- all models re-implemented and check against reference

Results

Table: Experimental results.

	RPS	Ent	Acc
Berrar	0.2088	1.0221	49.03
Elo	**0.2087**	**1.0216**	**49.10**
PageRank	0.2134	1.0349	47.88
pi-ratings	0.2091	1.0236	49.01
Poisson	*0.2088*	*1.0219*	*48.94*
Steph	0.2099	1.0254	48.94
TrueSkill	0.2104	1.0267	48.73

Conclusion

- slightly modified 40 years old Elo model performed best
    - closely followed by 36 years old Poisson model
- performance gap between domain specific and general ratings
- endless options for tuning
- further analysis of the results TBD
- call for contributions

Conclusion

- after nearly 40 years, the Poisson model is still competitive
- performance gap between domain specific and general ratings
- endless options for tuning
- further analysis of the results TBD
- call for contributions

Thank you for your attention.

References I

Daniel Berrar, Philippe Lopes, and Werner Dubitzky.
Incorporating domain knowledge in machine learning for soccer outcome prediction.
Machine Learning, 108(1):97–126, 2019.

Anthony Costa Constantinou and Norman Elliott Fenton.
Determining the level of ability of football teams by dynamic ratings based on the relative discrepancies in scores between adversaries.
Journal of Quantitative Analysis in Sports, 2013.

Anjela Y Govan, Carl D Meyer, and Russell Albright.
Generalizing google's pagerank to rank national football league teams.
2008.

References II

Shengbo Guo, Scott Sanner, Thore Graepel, and Wray Buntine.
Score-based bayesian skill learning.
In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 106–121. Springer, 2012.

Lars Magnus Hvattum and Halvard Arntzen.
Using ELO ratings for match result prediction in association football.
International Journal of Forecasting, 26(3), 2010.

Michael J Maher.
Modelling association football scores.
Statistica Neerlandica, 36(3):109–118, 1982.

References III

Alec Stephenson and Jeff Sonas.
PlayerRatings: Dynamic Updating Methods for Player Ratings Estimation, 2016.
R package version 1.0-1.