Expectation Propagation for Rating Players in Sports Competitions

Adriana Birlutiu and Tom Heskes

Radboud University Nijmegen, The Netherlands

Abstract

Rating players in sports competitions based on game results is one example of paired comparison data analysis. Since an exact Bayesian treatment is intractable, several techniques for approximate inference have been proposed in the literature. In this paper we compare several variants of expectation propagation (EP). We evaluate the different approaches on a large tennis dataset to find that EP does significantly better than ADF (iterative improvement indeed helps) and EP-Correlated does significantly better than EP-Independent (correlations do matter).

1. Probabilistic Bayesian Framework

We consider the player's strength as a probabilistic variable in a Bayesian framework. Let $\boldsymbol{\theta}$ be an n_{players}-dimensional probabilistic variable whose components represent the players' strengths.

- **Prior**. Information available about the players can be incorporated in a prior.
- **Likelihood**. The Bradley-Terry model [1, 4] expresses the probability of player i winning against player j, as a function of their strengths θ_i and θ_j. We define $r_{ij} = 1$ if player i beats player j, and $r_{ij} = -1$ otherwise;

$$p(r_{ij}|\theta_i, \theta_j) = \frac{1}{1 + \exp[-r_{ij} \cdot (\theta_i - \theta_j)]} \,. \tag{1}$$

- **Posterior**. Using Bayes' rule we compute the posterior distribution over the players' strengths,

$$p(\boldsymbol{\theta}|R) = \frac{1}{d} p(R|\boldsymbol{\theta}) p(\boldsymbol{\theta}) = \frac{1}{d} p(\boldsymbol{\theta}) \prod_{i \neq j} p(r_{ij}|\theta_i, \theta_j) \,. \tag{2}$$

  We take the mean of the posterior distribution as our best estimate of the players' strengths and the covariance matrix as the uncertainty about our estimation.

2. Expectation Propagation

Expectation propagation (EP) [3] is an approximation technique which tunes the parameters of a simpler approximate distribution, to match the exact posterior distribution of the model parameters given the data.

- **Assumed Density Filtering**. In ADF the terms of the posterior distribution are added one at a time, and in each step the result of the inclusion is projected back into the assumed density. As the assumed density we take the Gaussian distribution. This is very similar to the approach taken in [4].
- **Iterative Improvement: EP-Correlated**. After we add a term and project, the Gaussian approximation changes. We call the quotient between the new and old Gaussian approximation a *term approximation*. EP generalizes ADF by performing backward-forward iterations to refine the term approximations until convergence. The final approximation will be independent of the order of incorporating the terms.
- **EP-Independent** The complexity of EP can be reduced if we keep track only of the diagonal elements of the covariance matrix, ignoring the correlations.

3. Experiments

The comparison was made on a large dataset consisting of results of 38538 tennis matches played on ATP events among 1139 players between 1995 and 2006.

Table1: *Comparison between EP-Correlated, ADF and EP-Independent based on the number of matches correctly/incorrectly predicted.*

	ADF		EP-Independent	
	correct	incorrect	correct	incorrect
EP-Correlated				
correct	54.48%	7.81%	58.46%	3.83%
incorrect	6.21%	31.50%	3.09%	34.62%

Figure 1: *A histogram of the players' strengths (means of the posterior distribution) for all years. The bar indicates the average width of the posterior distribution for each of the individual players. The results shown are for EP-Correlated.*

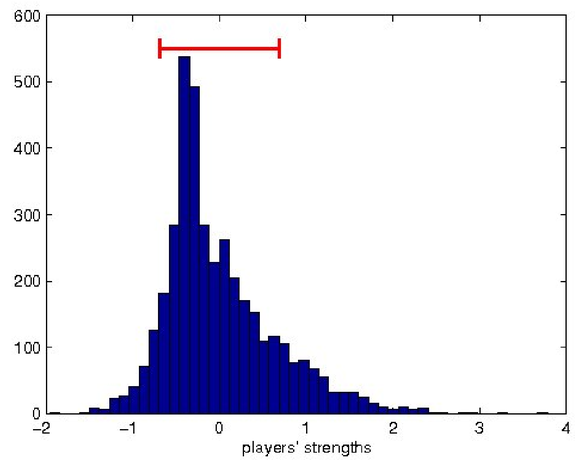

Figure 2: *Comparison between EP-Ranking and ATP for the year 2005.*

	EP (means)	ATP (points)
Federer, Roger	1 (3.78)	1 (6725)
Nadal, Rafael	2 (2.92)	2 (4765)
Hewitt, Lleyton	3 (2.47)	4 (2490)
Roddick, Andy	4 (2.27)	3 (3085)
Agassi, Andre	5 (2.22)	7 (2275)
Gasquet, Richard	6 (1.92)	16 (1506)
Ljubicic, Ivan	7 (1.85)	9 (2180)
Gaudio, Gaston	8 (1.75)	10 (2050)
Gonzalez, Fernando	9 (1.65)	11 (1790)
Nalbandian, David	10 (1.62)	6 (2370)

Acknowledgments

The statistical information contained in the tennis dataset has been provided by and is being reproduced with the permission of ATP Tour, Inc., who is the sole copyright owner of such information.

4. Challenges

Here, we considered the most basic probabilistic rating model; this model performs as good as the ATP ranking system. We would expect that the more complex models could outperform ATP.

- Generalize to more complex models, e.g., including dynamics [4] over time and team effects [2]
- Specifically for tennis, incorporate the effect of the surface the games are played on
- Apply the comparison to other types of data

References

[1] R.A. Bradley and M.E. Terry, Rank Analysis of Incomplete Block Designs: I, the Method of Paired Comparison, Biometrika, 1952

[2] Ralf Herbrich, Tom Minka and Thore Graepel, TrueSkill: A Bayesian Skill Rating System, *NIPS* 2007

[3] Tom Minka, A Family of Algorithms for Approximate Bayesian Inference, PhD thesis M.I.T, 2001

[4] Mark Glickman, Paired Comparison Models with Time Varying Parameters, PhD thesis, Harvard University, 1993