Expectation Propagation for Rating Players in Sports Competitions

Adriana Birlutiu [a] Tom Heskes [a]

[a] *Institute for Computing and Information Sciences, Radboud University Nijmegen*
Toernooiveld 1, 6525 ED Nijmegen

Abstract

The full version of this paper appears in the Proceedings of PKDD 2007, Warsaw, Poland.

1 Introduction

Our goal is to develop and evaluate methods for the analysis of paired comparison data. In this paper we illustrate such methods by rating players in sports, in particular in tennis.

We consider the player's strength as a probabilistic variable in a Bayesian framework. Before taking into account the match outcomes, information available about the players can be incorporated in a prior distribution. Using Bayes' rule we compute the posterior distribution over the players' strengths. 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.

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). EP generalizes assumed density filtering (ADF) by iteratively improving the approximations that are made in the filtering step of ADF. Furthermore, we distinguish between two variants of EP: EP-Correlated, which takes into account the correlations between the strengths of the players and EP-Independent, which ignores those correlations.

The question that we want to answer here is: how do different variants of expectation propagation perform for this setting? In particular, does it make sense to perform backward and forward iterations for the approximations and does it help to have a more complicated (full) covariance structure?

2 Experiments

We evaluate the variants of EP on a large tennis dataset. The results are shown in Table 1. We applied a binomial test to check the significance of the difference in performance between the algorithms [4]. We found out that, for this type of dataset, EP does significantly better than ADF (iterative improvement indeed helps) and EP-Correlated does significantly better than EP-Independent (correlations do matter). Further experiments should reveal whether this also applies to other types of data.

Using the posterior probability over the players' strengths we computed the confidence of the predictions. The algorithms perform about the same in estimating the confidence. However, they all tend to be overconfident, as indicated in Figure 1. We can correct this by adding noise to the players' strengths, to account for the fact that a player's strength changes over time.

We also compared the accuracy of the predictions based on the EP ratings with the accuracy of the predictions obtained using the ATP ratings. The ATP rating system gives points to players according to the type of the tournament and how far in the tournament they reached. Averaged over all the years, both EP and ATP ratings, give similar accuracy of predictions for the next, about 62%. In this paper 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.

Table 1: 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	16636 (54.48%)	2395 (7.81%)	17857 (58.46%)	1174 (3.83%)
incorrect	1902 (6.21%)	9620 (31.50%)	945 (3.09%)	10577 (34.62%)

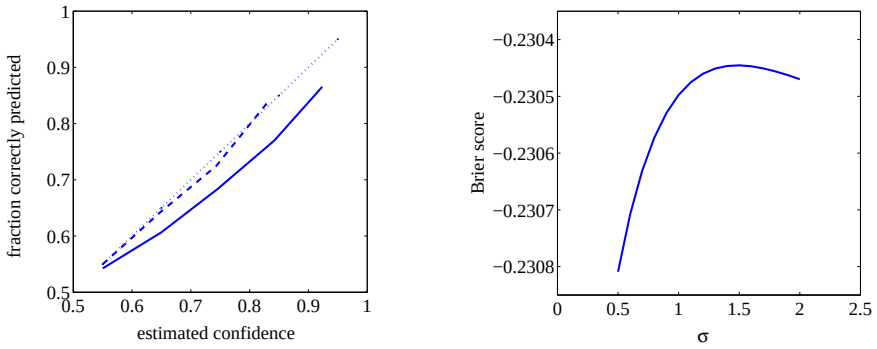

Figure 1: Left: the actual fraction of correctly predicted matches as a function of the predicted confidence; without added noise (solid line) and with noise of standard deviation 1.4 added (dashed line); the dotted line represents the ideal case and is drawn for reference. Right: the Brier score for the confidence of the predictions as a function of the standard deviation of the noise added to each player's strength.

3 Future work

Our results are generalizable to more complex models, e.g. including dynamics over time, which means that a players rating in the present is related to his performance in the past [1]; and team effects: a player's rating is inferred from team performance [2, 3]. Specifically for tennis, the more complex models should also incorporate the effect of surface because the performance of tennis players in a match is influenced by the type of surface they play on (grass, clay, hard court, indoor).

References

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

[2] Ralf Herbrich, Tom Minka, and Thore Graepel. Trueskill: A Bayesian skill rating system. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 569–576. MIT Press, Cambridge, MA, 2007.

[3] Tzu-Kuo Huang, Chih-Jen Lin, and Ruby C. Weng. A generalized Bradley-Terry model: From group competition to individual skill. In Lawrence K. Saul, Yair Weiss, and Léon Bottou, editors, *Advances in Neural Information Processing Systems 17*, pages 601–608. MIT Press, Cambridge, MA, 2005.

[4] Steven L. Salzberg. On comparing classifiers: Pitfalls to avoid and a recommended approach. *Data Mining and Knowledge Discovery*, 1(3):317–328, 1997.