The Player Kernel: Learning Team Strengths Based on Implicit Player Contributions

Lucas Maystre[1], Victor Kristof[1], Antonio J. González Ferrer[1,2], and
Matthias Grossglauser[1]

[1] School of Computer and Communication Sciences, EPFL, Switzerland
[2] Escuela Técnica Superior de Ingenieros Informáticos, UPM, Spain

Abstract. In this work, we draw attention to a connection between
skill-based models of game outcomes and Gaussian process classification
models. The Gaussian process perspective enables *a*) a principled way
of dealing with uncertainty and *b*) rich models, specified through kernel
functions. Using this connection, we tackle the problem of predicting
outcomes of football matches between national teams. We develop a
player kernel that relates any two football matches through the players
lined up on the field. This makes it possible to share knowledge gained
from observing matches between clubs (available in large quantities) and
matches between national teams (available only in limited quantities). We
evaluate our approach on the Euro 2008, 2012 and 2016 final tournaments.

1 Introduction

Statistical models of game outcomes have a rich and diverse history, going back
almost a century: as early as 1928, Zermelo [8] proposed a simple algorithm that
infers the skill of chess players based on observed game outcomes. Zermelo's ideas
have since been rediscovered and refined multiple times, and have been successfully
applied to various sports-related prediction problems and beyond. On the occasion
of the Euro 2016 football tournament, we revisit these ideas and highlight their
connections to modern machine learning techniques. In particular, we show how
Zermelo's model can be cast as a Gaussian process classification model. The
Gaussian process framework provides two key advantages. First, it brings all
the benefits of Bayesian inference. In particular it provides a principled way to
deal with the uncertainty associated to noisy observations and to predictions.
Second, it opens up new modeling perspectives through the specification of kernel
functions.

Equipped with this, we investigate the problem of predicting outcomes of
football matches between national teams. We identify two key challenges, *a*) that
of *data sparsity* (national teams usually play no more than ten matches per year),
and *b*) that of *data staleness* (the team roster is constantly evolving). Taking
inspiration from the observation that national teams' players frequently face
each other in competitions between clubs (see Figure 1), we show that these
two difficulties can be tackled by the introduction of a *player kernel*. This kernel
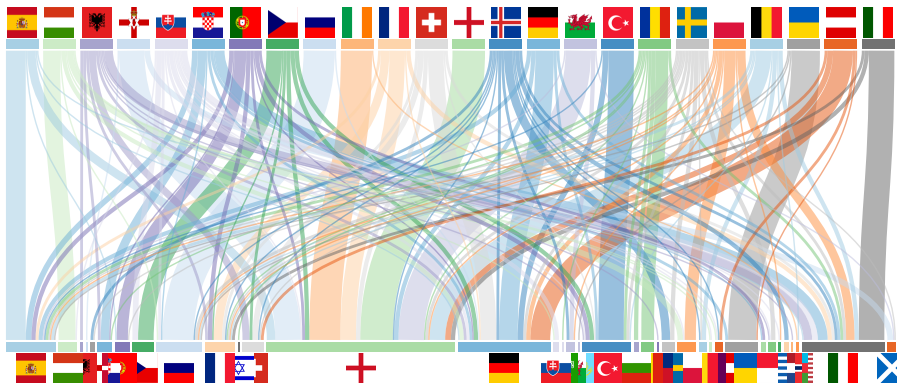
Fig. 1. Players of national teams qualified for the Euro 2016 (top row) are playing in clubs across Europe and beyond (bottom row). The English, German and Italian club championships contain the most selected players.

relates any two matches through the players lined up on the field, and makes it possible to seamlessly use matches between clubs to improve a predictive model ultimately used for matches between national teams. In contrast to national teams, clubs play much more frequently, and more data is available.

The remainder of this short report is organized as follows. We review related work at the end of the present section. In Section 2, we formalize the link between Zermelo's ideas and Gaussian processes, and present our player kernel. Then, in Section 3, we evaluate our predictive model on the Euro 2008, 2012 and 2016 final tournaments.

1.1 Related Work

More than two decades after Zermelo's seminal paper [8], his model for paired comparisons was rediscovered and popularized by Bradley and Terry [1]. Nowadays, the model is usually referred to as the Bradley–Terry model. In the context of skill-based game modeling, the same model (associated to a simple online stochastic gradient update rule) is also known as the Elo rating system [3]. It is used by FIDE to rank chess players[3] and by FIFA to rank women national football teams[4], among others.

The model and related inference algorithms have been extended in various ways; one direction that is of particular interest is the handling of uncertainty of the estimated skill parameters. Glickman [4] proposes an extension that simultaneously updates ratings and associated uncertainty values after each observation. Herbrich et al. [5] propose TrueSkill, a comprehensive Bayesian

---

[3] See: https://ratings.fide.com/.

[4] See: http://www.fifa.com/fifa-world-ranking/procedure/women.html.

framework for estimating player skill in various types of games. The models and methods described in this paper are fundamentally similar to TrueSkill, as will be discussed in Section 2. Finally, in the context of learning users' preferences from pairwise comparisons, Chu and Ghahramani [2] present a Gaussian process approach that is comparable to our work.

2 Methods

In this section, we first show how the model of pairwise comparisons proposed by Zermelo [8] and popularized by Bradley and Terry [1] and Elo [3] can be expressed in the Gaussian process framework. Second, we present the player kernel, a covariance function that relates matches through lineups.

2.1 Pairwise Comparisons as Gaussian Process Classification

Suppose that we observe outcomes of comparisons between two objects (e.g., two players or two teams) in a universe of objects denoted $1, \ldots, M$. We begin by restricting ourselves to binary outcomes, i.e., we assume that one of the two objects wins. Zermelo [8] postulates that each object u can be represented by a parameter $w_u \in \mathbf{R}_{>0}$, indicative of its relative chances of winning against an opponent. Given these parameters, the probability of observing the outcome "u wins against v" (denoted by $u \succ v$) is given by $w_u/(w_u + w_v)$. Using the reparametrization $w_u = e^{s_u}$, this can be rewritten as

$$P(u \succ v) = \frac{1}{1 + \exp[-(s_u - s_v)]} = \frac{1}{1 + \exp(-\boldsymbol{s}^\top \boldsymbol{x})}, \tag{1}$$

where $\boldsymbol{s} = [s_i]$ and $\boldsymbol{x} \in \mathbf{R}^M$ is such that $x_u = 1$, $x_v = -1$ and $x_i = 0$ for $i \neq u, v$. As such, the pairwise comparison model can be seen as a special case of logistic regression, where the feature vector simply indicates the winning and losing objects. Furthermore, logistic regression is itself a special case of Gaussian process classification [7, Ch. 3]. A Gaussian process $f(\boldsymbol{x}) \sim \mathcal{GP}(m(\boldsymbol{x}), k(\boldsymbol{x}, \boldsymbol{x}'))$ is defined by a mean function $m(\boldsymbol{x})$ and a positive semi-definite covariance (or kernel) function $k(\boldsymbol{x}, \boldsymbol{x}')$. Given any finite collection of points $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$, the Gaussian process sampled at these points has a multivariate Gaussian distribution

$$\begin{bmatrix} f(\boldsymbol{x}_1) \ldots f(\boldsymbol{x}_k) \end{bmatrix} = \mathcal{N}(\boldsymbol{m}, \boldsymbol{K}),$$

where $m_i = m(x_i)$ and $K_{ij} = k(\boldsymbol{x}_i, \boldsymbol{x}_j)$. It is not hard to show that if $\boldsymbol{s} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$, then $f(\boldsymbol{x}) = \boldsymbol{s}^\top \boldsymbol{x}$ is a Gaussian process with $m(\boldsymbol{x}) = 0$ and $k(\boldsymbol{x}, \boldsymbol{x}') = \sigma^2 \boldsymbol{x}^\top \boldsymbol{x}'$. This enables the interpretation of (1) as the likelihood of a Gaussian process classification model with the logit link function.

The Gaussian process viewpoint shifts the focus from the representation of the function $f(\boldsymbol{x})$ (in the case of (1), a linear function) to the correlation between two function evaluations, as defined by the kernel function $k(\boldsymbol{x}, \boldsymbol{x}')$. Intuitively, the model can simply be specified by how similar any two match outcomes

are expected to be. Furthermore, the Gaussian process viewpoint also makes it possible to take advantage of the vast amount of literature and software related to accurate, efficient and scalable inference.

Handling draws. Rao and Kupper [6] propose an extension of the pairwise comparison model for ternary (win, draw, loss) outcomes. In this extension, the two different types of outcomes have probabilities

$$P(u \succ v) = \frac{1}{1 + \exp[f(\boldsymbol{x}) - \alpha]} \text{ and } P(u \equiv v) = (e^{2\alpha} - 1)P(u \succ v)P(v \succ u),$$

where $\alpha > 0$ is an additional hyperparameter controlling the draws. Because a draw can be written as the product of a win and a loss, model inference can still be performed using only a *binary* Gaussian process classification model, with minimal changes needed to the link function.

2.2 The Player Kernel

We now consider an application to football and propose a method to quantify how similar two match outcomes are expected to be. Denote by P the number of distinct players appearing in a dataset of matches. We define a team's *lineup* as the set consisting of the 11 players starting the match. For a given match, let $\mathcal{W}$ and $\mathcal{L}$ be the lineups of the winning and losing teams, respectively. Define $\boldsymbol{z} \in \mathbf{R}^P$ such that $z_p = 1$ if $p \in \mathcal{W}$, $z_p = -1$ if $p \in \mathcal{L}$ and $z_p = 0$ otherwise. We then define the player kernel as

$$k(\boldsymbol{z}, \boldsymbol{z}') = \sigma^2 \boldsymbol{z}^\top \boldsymbol{z}'.$$

Intuitively, the function is positive if the same players are lined up in both matches, and the same players win (respectively lose). The function is negative when players win one match, but lose the other. Finally, the function is zero, e.g., when the lineups are completely disjoint.

This kernel implicitly projects every match into the space of players, and defines a notion of similarity in this space. In the case of national teams qualified to Euro final tournaments, we find that this approach is very useful: a significant part of national teams' players take part in one of the main European leagues and play with or against each other. International club competitions (such as the UEFA Champions League) further contribute to the "connectivity" among players. Figure 2 illustrates the similarity of matches across different competitions in 2011–2012.

It is interesting to note that the player kernel corresponds to a linear model over the players. That is, it is equivalent to assuming that there is one independent skill parameter per player, and that the strength of a team is the sum of its players' skills. Such a model contains a massive number of parameters (possibly much more than the number of observations), and there is little hope to reliably estimate every parameter. In fact, we observe that the model is "weakly" parametric: the number of distinct players usually grows with the number of matches observed. The kernel-based viewpoint that we take emphasizes the fact that estimating these parameters is not necessary.
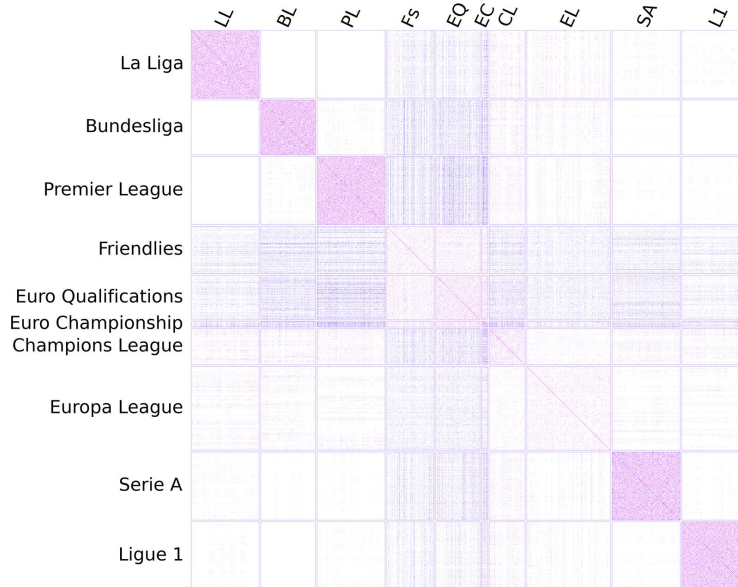
Fig. 2. Heatmap of the magnitude of the kernel matrix for 3184 matches played over the year preceding Euro 2012. White indicates zero correlation, a saturated color indicates more correlation. Matches between national teams exhibit non-zero covariance with matches of all other competitions.

Relation to TrueSkill. Our Gaussian process model coupled to the player kernel is very similar to TrueSkill [5]. The most important difference is that we take advantage of the dual representation and operate in the space of matches instead of the space of players. Beyond the conceptual reasons outlined above, it makes inference significantly less computationally intensive for the datasets that we consider.

3 Experimental Evaluation

In this section, we evaluate our predictive model on the matches of the Euro 2008, 2012 and 2016 final tournaments and compare it to several baselines.

We collect a dataset of matches from *a*) official and friendly competitions involving national teams, and *b*) the most prestigious European club competitions, starting from July 1st, 2006. There are approximately $15\times$ more matches between clubs than there are matches between national teams in our dataset. With respect to the model outlined in Section 2, our final predictive model processes one additional feature that encodes which team played at home (this feature is null for matches played on neutral ground). We train the model using all N matches that were played prior to the start of the competition on which we test. When computing the kernel matrix (whether on training or on test data) we use the starting lineups, usually announced shortly before the start of the game. It is

Table 1. Average logarithmic loss of our predictive model (PlayerKern), a model based on national team ratings (Elo), betting odds (Odds) and a random baseline (Random) on the final tournaments of three European championships. N is the number of training instances, P the number of distinct players and T the number of test instances.

Competition	N	P	T	PlayerKern	Elo	Odds	Random
Euro 2008	4390	7875	31	0.969	**0.910**	0.979	1.099
Euro 2012	15 594	21 735	31	**0.939**	1.003	0.953	1.099
Euro 2016	24 887	33 157	51	1.067	1.102	**1.020**	1.099

interesting to note that the number of distinct players P appearing in the dataset exceeds the number of training instances in each case (the values of N and P are shown in Table 1). We use the GPy Python library[5] to fit the model; inference takes a minute for the 2008 test set (17 minutes for 2016). The predictions come in the form of probability distributions $[p^{\mathrm{W}}, p^{\mathrm{D}}, p^{\mathrm{L}}]$ over the three outcomes (win, draw, loss).

We compare our predictive distributions against three baselines. First, we consider a simple Rao-Kupper model based on national team ratings obtained from a popular Web site[6]. This model is similar to ours, but a) it does not relate matches through player, and thus does not consider club outcomes, and b) as ratings are fixed values, it does not consider uncertainty in the ratings. Second, we consider average probabilities derived from the odds given by three large betting companies. Third, we consider a random baseline which always outputs $[1/3, 1/3, 1/3]$. The predictive distributions are evaluated using the average logarithmic loss over T test instances

$$-\frac{1}{T} \sum_{i=1}^{T} \left[\mathbf{1}_{\{y_i = \mathrm{W}\}} \log p_i^{\mathrm{W}} + \mathbf{1}_{\{y_i = \mathrm{D}\}} \log p_i^{\mathrm{D}} + \mathbf{1}_{\{y_i = \mathrm{L}\}} \log p_i^{\mathrm{L}} \right].$$

The logarithmic loss penalizes more strongly predictions that are both confident and incorrect. Table 1 summarizes the results.

Our predictive model performs well in 2008 and 2012, but slightly less so in 2016. It is noteworthy that the 2016 final tournament has been generally less predictable than earlier editions. The case of the Elo baseline is interesting, as its accuracy varies wildly. Reasons for this might include the noise due to the online gradient updates, and the lack of proper uncertainty quantification in the ratings. Our method, in contrast, seems to produce more conservative predictions, but manages to achieve a more consistent performance

3.1 Conclusion

In this short report, we exposed a connection between a well-known pairwise comparison model and Gaussian process classification, and proposed a kernel that

---

[5] See: https://sheffieldml.github.io/GPy/.
[6] See: http://www.eloratings.net/.

is able to transfer knowledge across different types of football matches—those between clubs and those between national teams. We showed that a predictive model building on these ideas achieves a logarithmic loss that is competitive with betting odds. In future work, we would like to investigate how to incorporate aging into the model, i.e., how to progressively downweight older data.

References

1. Bradley, R.A., Terry, M.E.: Rank Analysis of Incomplete Block Designs: I. The Method of Paired Comparisons. Biometrika 39(3/4), 324–345 (1952)
2. Chu, W., Ghahramani, Z.: Preference Learning with Gaussian Processes. In: ICML 2005. pp. 137–144 (2005)
3. Elo, A.: The Rating Of Chess Players, Past & Present. Arco (1978)
4. Glickman, M.E.: Parameter estimation in large dynamic paired comparison experiments. Applied Statistics 48, 377–394 (1999)
5. Herbrich, R., Minka, T., Graepel, T.: TrueSkill$^{\mathrm{TM}}$: A Bayesian Skill Rating System. In: Advances in Neural Information Processing Systems. pp. 569–576 (2006)
6. Rao, P.V., Kupper, L.L.: Ties in Paired-Comparison Experiments: A Generalization of the Bradley-Terry Model. Journal of the American Statistical Association 62(317), 194–204 (1967)
7. Rasmussen, C.E., Williams, C.K.I.: Gaussian Processes for Machine Learning. MIT Press (2006)
8. Zermelo, E.: Die Berechnung der Turnier-Ergebnisse als ein Maximumproblem der Wahrscheinlichkeitsrechnung. Mathematische Zeitschrift 29(1), 436–460 (1928)