Summary of "TrueSkill: A Bayesian Skill Rating System"

Matt Dickenson `mcd31`
STA571/CS590.01 Due: 7 April, 2014

Herbrich et al. present a Bayesian skill rating system that generalizes Elo ratings (used in chess) to multi-player and team-based competitions. Their system, known as TrueSkill, also measures uncertainty about player skills and models draws. Model inference is performed using a factor graph with message passing.

In the Elo model, the outcome of a match is modeled as a function of players' skill levels s_i. A player i's performance in a match is normally distributed around her skill level with variable β^2: $p_i \sim N(s_i, \beta^2)$. Without loss of generality, the probability that player 1 wins can be computed using the standard normal Gaussian cumulative density function Φ:

$$P(p_1 > p_2 | s_1, s_2) \quad = \quad \Phi(\frac{s_1 - s_2}{\sqrt{2}\beta}).$$

After the match s_i is unchanged if there is a draw, and incremented by $y\Delta$ ($y = 1$ if the player won and $y = -1$ if the player lost) where

$$\Delta \quad = \quad \alpha\beta\sqrt{\pi}(\frac{y+1}{2} - \Phi(\frac{s_1 - s_2}{\sqrt{2}\beta})).$$

α serves as a weighting for this game relative to the original estimate, $0 < \alpha < 1$. This model has been very useful for ranking chess players, but faces two major challenges with regard to multi-player online games. The outcomes of such games are often team-based, but the match-making requires estimates of individual skill levels (because team membership may not be constant over time). Outcomes of these games are also more generally, with a full ranking of players/teams rather than just winners and losers.

To address these issues, Herbrich et al. introduce a model for skill ranking. For a match with n players and k teams, the outcome $r = (r_1, \ldots, r_k) \in \{1, \ldots, k\}$ specifies a rank r_j for each team t_j. The posterior probability of an outcome given players' skills and their team assignments $A = \{A_1, \ldots, A_k\}$ can be computed by:

$$P(s | r, A) \quad = \quad \frac{P(r | s, A) P(s)}{P(r | A)}.$$

The prior on skill levels is Gaussian, and each player's performance is assumed to be centered around their skill level with variance β^2 as above. Inference is conducted using expectation propagation on a factor graph of the model.

There are several areas for further development of this model, as well as difficulties interpreting its predictive performance. Like the Elo model, TrueSkill assumes that skill is a uni-dimensional latent trait, but in multi-player games several skill sets (such as game skills and communication skills) could be at play simultaneously. A player's first few games may also have an inordinate effect on their skill level, especially if they lose those games (and are thus matched with lower-ranked players, making it difficult to recover). A key complication in assessing its performance is that rankings are based on match outcomes, but assignment of players and teams to matches is also based on rankings (close matches are desired). Furthermore, human players may seek to game the system by seeking matches to improve their ranking regardless of their actual skill level. Despite these issues, the TrueSkill model is an interesting step forward.