Chapter 8: Ranking Skill
Chapter Overview

- calculating the skill level from win statistics
- ELO-Ranking
- True Skill
- Team Skill
Models for play level

idea: Skill level can be deduced from past victories and defeats.

model: Every player $i$ has a skill level $s_i$. If $s_i > s_j$ then $s_i$ is very likely to win in a competition.

applications:

• **matchmaking**: choose interesting opponents with comparable skill level

• **ladders/rankings**: creating public rankings as an expression of prestige (compare Tennis, SC2, WOW arena, Halo2, …)

• **organizing tournaments**: assistance for draw, qualification, clearing disputes.
The ELO System

Introduced by Arpad Elo in 1970 and adopted by the World Chess Federation.

**Assumption:** player $i$’s performance $p_i$ is normal distributed around his skill level with variance $\beta^2$.

$\Rightarrow s_i: p_i = N(s_i, \beta^2)$

$\Rightarrow s_i > s_j$ does not necessarily mean $i$ is losing against $j$ 

**but rather** $Pr(i$ wins against $j) > 50\%$

**task:** compute $Pr(p_i > p_j | s_i, s_j)$ (probability of $i$ playing better than $j$)

$\Rightarrow$ Difference of 2 normal distributed variables with the same variance $\beta^2$ is normal distributed with mean $(s_i - s_j)$ and variance $\beta^2$.

Let $\Phi$ be the accumulated density function of a normal distribution with anticipated value of 0 and a variance of 1, then follows:

$$P(p_1 > p_2 | s_1, s_2) = \Phi\left(\frac{s_1 - s_2}{\sqrt{2\beta}}\right)$$
Updating the ELO Ranking

- positions have to be adjusted as soon as new results are available.
- changes follow the zero-sum principle. \( s_1^{\text{new}} + s_2^{\text{new}} = s_1 + s_2 \)
- difference \( \Delta \) is supposed to increase the likelihood of the observation within the model
- match result: \( y \in \{0, -1, 1\} \) (Win:1, Loss:-1, Draw:0)
- updating ELO Scores with the result \( y_i \):
  \[
  \Delta = \alpha \beta \sqrt{\pi} \left( \frac{y_i + 1}{2} \right) - \Phi \left( \frac{s_1 - s_2}{\sqrt{2} \beta} \right)
  \]
  \(\alpha\): weighing factor for a match \(0 < \alpha < 1\) (approx. 0.07 for chess)
- ELO scores need comparatively many matches to stabilize. (ca. 20)
- properties:
  - chronological order of updates is important: good for long intervals between measurements, but bad performance for tournaments, where a players skill presumably stays constant.
  - ELO system does not allow for conclusions about individual performance in team games.
  - restricted representation of results. No differentiated treatment of events with a ranking for result (e.g. racing, ...).
True Skill

factor graphs
bi-partite graph with factor nodes and variable nodes.

- variable nodes: describe distribution functions
- factor nodes: model the interaction of variables
- edges: description of variables interacting for a factor

**example:** Factor Graph for ELO System

- **True Skill:** extension of ELO Systems used for XBOX360 Live (e.g. HALO2 ranking)
- **considers:**
  - skill uncertainty
  - allows conclusions for team-members in team games (additive performance $t_1$)
  - result presentation as order of play results ($t_1 \geq t_2 \geq .. \geq t_m$)
Factor graph for True Skill

Example: 4 Players, 3 Teams: \{ (s_1), (s_2,s_3), (s_4) \}
Result: \( t_1 > \varepsilon + t_2, \ t_1 > \varepsilon + t_3, \ \varepsilon > |t_2 - t_2| \)
Factor Graph use for True Skill

- factor graph represents the distribution for $Pr(s,p,t|r,A)$
  - $r$: ranking result, $A$: team composition
  - $s$: player skill, $p$: player performance, $t$: team rating
- compute the distribution of player skill $s$ conditional to the observations $r$ and $A$:
  $$
  Pr(s \mid r, A) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} Pr(s, p, t \mid r, A) dp \, dt
  $$
  $s_i$ is normal distributed with mean value $\mu_i$ and standard deviation $\sigma_i$
- With the given factor graph and the current values of $\mu$ and $\sigma$ for the participating players $\Pi(d_1 > \varepsilon)$ and $\Pi(\left|d_2\right| \leq \varepsilon)$ can be estimated.
- Comparing the prediction with the actual result, one can propagate the error back to $\mu$ and $\sigma$ and adapt the model accordingly.
- Propagating probabilities and parameter updates on a factor graph are also called message-passing or belief propagation.
Training scheme for True Skill

1. Forward propagation: estimate the results
2. Update of Team-performance: redistribution of results to teams
3. Update of a-posteriori Distributions: propagates update-messages as far as parameters $\mu$ and $\sigma$. 

\[
\begin{align*}
N(\mu_1, \sigma_1^2) & \quad N(\mu_2, \sigma_2^2) \\
N(s_1, \beta^2) & \quad N(s_2, \beta^2) \\
p_1 & \quad p_2 \\
\Pi(t_1 = p_1) & \quad \Pi(t_2 = p_2 + p_3) \\
t_1 & \quad t_2 \\
\Pi(d_1 = t_1 - t_2) & \quad \Pi(d_2 = t_2 - t_3) \\
d_1 & \quad d_1 \\
\Pi(d_1 > \varepsilon) & \quad \Pi(|d_2| \leq \varepsilon) \\
\end{align*}
\] 

A-priori-Distr. \quad Perf. Distr. 

Team Distr. 

Distribution of score-differences
Discussion True Skill

• Improves the ELO system by:
  • expansion of result representation
  • converges faster using a priori distributions for particular players
  • team assessment

• Disadvantages of True Skill:
  • chronological order is important, even though one can assume that skill does not change between two matches. (Expansion: True Skill Through Time 2008)
  • team skill is considered as the sum of player skills

**But:** In reality player synergy is much more complicated: “Having 5 carries in a Moba does not work!!”
Team Skill

**idea**: Considering not only individual play level, but also team chemistry.

=> Viewing a player’s joint performance compared to his single performance.

=> Some player’s performance increases when combined with specific players.

**given**: A Team $T=\{p_1,...,p_K\}$ with $K$ players. Let $t_k$ be a sub-team of $T$ with $k$-elements. ($t_k \subseteq T \land |t_k|=k$). $\text{Skill}(t_k)$ constitutes sub-team $t_k$‘s skill level. (For example, calculated with ELO or True-Skill.)

**task**: Skill level of team $T$ considering team chemistry?

**approach**: calculating average over computed sub-team ranking
Team Skill-\(k\)

- average play level of a sub team of size \(k\) scaled to team size \(K\).

\[
TS_k(T) = K \cdot \frac{1}{k} \cdot \binom{K}{k} \cdot \sum_{i=1}^{\binom{K}{k}} \text{Skill}(s_{ki}) = \frac{(k-1)!}{(K-1)!} \cdot \sum_{i=1}^{\binom{K}{k}} \text{Skill}(s_{ki})
\]

example:

\(k=1\) and \(K=5\)

\[
TS_k(T) = 5 \cdot \frac{1}{1} \cdot \binom{5}{1} \cdot \sum_{i=1}^{\binom{5}{1}} \text{Skill}(s_{1i}) = \sum_{i=1}^{\binom{5}{1}} \text{Skill}(s_{1i})
\]

\(k=2\) and \(K=5\)

\[
TS_k(T) = 5 \cdot \frac{1}{2} \cdot \binom{5}{2} \cdot \sum_{i=1}^{\binom{5}{2}} \text{Skill}(s_{2i}) = \frac{1}{4} \sum_{i=1}^{\binom{5}{2}} \text{Skill}(s_{2i})
\]
Team Skill-ALLK-LS

Challenges for improving Team Skill-\( k \):

- determining \( k \) is hard \( \Rightarrow \) take all possible sub-teams.
- we don’t have sample observations for all existing sub-teams \( \Rightarrow \) only consider sub-teams with a reliable skill estimation.

Idea: Consider all sub-teams with a reliable estimate and which are not a sub-team of a larger reliably estimated sub-team.

Approach: Consider all sub-teams \( t_{k,i}^* \) for which \( \text{Skill}(t_{k,i}) \) can be reliably computed and for which no sub-team \( t_{k+l,j} \supset t_{k,i} \) exists.

Calculate team performance as \( k \) times the mean of the single player performance.

\[
TS_{ALL-LS}(T) = \frac{K}{\sum_{m \in \{m \mid \exists t_{m}^* \neq \emptyset \}} |m|} \left( \sum_{m \in \{m \mid \exists t_{m}^* \neq \emptyset \}} E(t_{m}^*) \right) = \frac{K}{\sum_{m \in \{m \mid \exists t_{m}^* \neq \emptyset \}} |m|} \left( \sum_{m \in \{m \mid \exists t_{m}^* \neq \emptyset \}} \left( \frac{1}{l} \cdot \sum_{i=1}^{l} \text{Skill}(t_{m,i}^*) \right) \right)
\]
Example: Team Skill ALL-LS

red: pruned area, blue: used sub-teams, grey: pruned sub-teams.

\[ TS_{\text{ALL-LS}}(T) = \frac{4}{3+2} \left( \text{Skill}(t_{BCD}) + \frac{1}{2} \left( \text{Skill}(t_{AC}) + \text{Skill}(t_{AD}) \right) \right) \]
Conclusion

• method for capturing increased success of teams with good chemistry.

• team skill depends on data of as many different team compositions as possible

• approaches for improvement:
  • roles within the team are not required explicitly
  • confidence of the underlying skill estimation is not treated
  • correlation between team skill and player skill is assumed to be uniform

• Team Skill, True Skill and ELO symmetrically value wins and losses.
  => in many casual games a win awards more to the player skill than a loss would reduce it to keep players motivated to play.
Alternative Approach

• rating players not by success, but by skillful behavior:
  1. collect and describe spatial-temporal behavior over the full spectrum of skill.
  2. learn a regression model.
  3. rate player, while playing, for his $k$ last actions.

• this approach is used for dynamic play level adjustment in PVE.

• very suitable if it is known what constitutes successful behavior in the game. (e.g. accuracy in FPS Games, DPS/HPS Numbers in MMORPGS)
Learning goals

• Scope of application for player ranking and matchmaking
• ELO
• True Skill
• Team Skill
Literature