Comparing Skill Modeling Frameworks in Esports

Alex Bisberg and Kyli McKay-Bishop

School of Computing, University of Utah, Salt Lake City, UT, USA

1 Problem Definition and Motivation

On November 15, 2002 Microsoft introduced Xbox Live, an online gaming arena.[7] Since its release date Xbox Live has only grown in popularity, there are currently 65 million active subscriptions.[9] One of the critical features of Xbox Live is matchmaking; the matching of online players of similar skill. Matching of dissimilar players can result in disengagement or un-enjoyable game play. Successfully matching players requires defining each player's skill.

One of the earliest attempts at skill rating was the Elo model designed by Arpad Elo in 1959. The Elo model was originally developed as a rating system for chess.[2] In the Elo model, player's skills are considered point estimates and the game outcomes are modeled as a probability that one player's game performance exceeds the others. The player's skills are then updated based on the game outcome. A recent variation of the Elo model, that we explore in this work, is SCOPE.[1]

In 2006, a skill model called TrueSkill was introduced by Herbrich et al. In the TrueSkill model the player's skill is modeled as a Gaussian distribution instead of as a point estimate.[4] Updates, to the player's skill are done using a factor graph and are based on the individual player's performance. This model also has the flexibility to model multi-player and team games.

In this work we aim to better understand skill rating systems by comparing SCOPE and TrueSkill. We compare the accuracy, calibration and log loss for each of these models using Call of Duty World League data. In addition, we examine the convergence of these models to a synthetic ground truth data set of a player's skill.

This work is important because the TrueSkill model is more complex than the SCOPE model. TrueSkill is a probabilistic graphical model which adds the ability to quantify skill uncertainty, but makes the final skill inference more mathematically and computationally intensive. By comparing these two models we hope to make a statement on the necessity of such a complex model.

We used probabilistic techniques because a player's skill is inherently uncertain. An example of this uncertainty is an upset, when the player with the lower skill wins. A point estimate for the skill would imply that this player should never win the game. Where as a distribution implies that it is possible for this player to win, all be it, improbable. A distribution reflects reality, where in fact less skilled players do win on occasion.

2 Our Solution

2.1 Data

We compared SCOPE and TrueSkill using two datasets. First we compared the accuracy, calibration, and log loss of each model (defined below) using data from our collaborator Activision, the publishers of the Call of Duty franchise. They provided clean, open source data from the Call of Duty World League, a first person shooter es sport.[10].

Second, we created a synthetic ground truth data set. From the skill value in each rating system a probability of winning can be calculated. We wanted to model a player that had an increased skill over 500 games. To do this we selected a base skill and increased it after the first 200 games and then the next 100 games up to 500. We chose this pattern of games in an attempt to emulate the experiment in Winn's Model Based Machine Learning [5]. For each skill we calculated the probability of winning and used this to sample from a Bernoulli distribution to get the game outcomes the would be the input to our models. The SCOPE and TrueSkill model are on different scales so we selected the skill levels to keep the win probabilities close. The win probabilities across each set of games, separately for SCOPE and TrueSkill, are in Table 1. For SCOPE the skill range is larger and has lower variance.

Table 1: Win probabilities for each model calculated from corresponding skill.

SCOPE		TrueSkill	
Skill	Win %	Skill	Win %
1650	64%	110	78%
1750	85%	120	93%
1850	95%	130	98%
1850	98%	140	99%

2.2 Models

We compared the probabilistic modeling framework of TrueSkill to a somewhat simpler model, SCOPE. Bisberg et al. recently proposed SCOPE and showed comparable accuracy to TrueSkill2 [1]. SCOPE is an acronym that stands for Selective Cross-validation Over Parameters for Elo. The idea behind this model is to systematically tune various parameters of the Elo model to best fit the data using grid search cross-validation. In an attempt to make the model comparison as similar as possible, we used a similar strategy to find the ideal model parameters for TrueSkill [8].

There are multiple implementations of the TrueSkill framework. The original creators from Microsoft support an open source library called Infer.NET that can be used to build and run inference on factor graphs. They provide some examples, but not a full implementation of the TrueSkill algorithm. We chose to work with a Python implementation, found at `https://trueskill.org/` [6].

2.3 Model Assessment Metrics

We used three metrics to measure the performance of these two models on the historical data.

Accuracy For this metric, we count a *correct* prediction when the team that is expected to win over 50% of the time actually wins. The *accuracy* of the model is the total number of correct predictions divided by the total number of predictions we made over every series.

$$\text{correct} = \begin{cases} 1 & \text{if } \mathrm{E}[S]_x > 0.5 \wedge S_x = 1 \\ 0 & \text{otherwise} \end{cases}$$

$$\text{accuracy} = \frac{\sum^{n} \text{correct}}{n}$$

Calibration This is the sum of the win probabilities for each team predicted to have over a 50% chance of winning divided by the number of times those teams actually won. If the ratio is over one, better teams are predicted to win more often than they actually do, measuring over-prediction.

$$\text{correct} = \begin{cases} 1 & \text{if } \mathrm{E}[S]_a > \mathrm{E}[S]_b \wedge S_a = 1 \\ 0 & \text{otherwise} \end{cases}$$

$$\text{calibration} = \frac{\sum^{n} \mathrm{E}[S]_a}{\sum^{n} \text{correct}}$$

Log Loss This is a ratio of two quantities. The numerator is the sum of (a) the product of when the player actually won times the log of the prediction accuracy and (b) the product of when the player lost times the log of the prediction. A log loss closer to 0 is better.

$$\text{log loss} = \frac{-\sum^{n} S_a \log \mathrm{E}[S]_a + S_b \log \mathrm{E}[S]_b}{n}$$

Convergence We determined convergence by running each model on the synthetic game outcomes described in 2.1 and measuring the relative mean squared error (RSE) to determine the difference between the ground truth skill level and the skill level calculated by each model. The RSE was calculated for each game using the equation below and then averaged.

$$\text{RSE} = \frac{\sum_{j=1}^{n} \left(P_j - T_j\right)^2}{\sum_{j=1}^{n} \left(T_j - \overline{T}_j\right)^2}$$

Where P_{ij} is the predicted value for game i, T_j is the true value for game i, and $\overline{T}_j$ is the average of the true value. [3]

3 Experimental Evaluation

3.1 Accuracy, Calibration, and Log Loss

Below are the results of cross validation over SCOPE and the TrueSkill models. Table 2 shows the results of the SCOPE model and different iterations of the TrueSkill models we used. In our midterm we mentioned that we planned on expanding our TrueSkill model that has a single skill distribution for each (TS_Team) to modeling the skill of each individual on a team. We tried two separate player based models. One where the TrueSkill factor graph sums the skills of all individuals on a team (TS_Player) and another where only the player with the maximum skill on the team is used for outcome prediction (TS_MaxPlayer). Table 3 shows the highest accuracy model parameters for each TrueSkill model. The initial mean and standard deviation of the model are μ and σ. β is the point at which which guarantees about 76% chance of winning. The recommended value is a half of σ. τ is the dynamic factor which restrains a fixation of rating.

In all of the model assessment metrics SCOPE performed equal or better to all of the TrueSkill models. Since the midterm report, we expanded our TrueSkill models to individual players. This improved the model accuracy, increasing by about 2.4%. However, these models had worse calibration and log loss. This could be explained by a larger amount of variance added to the system since we now have four distributions instead of one. The MaxPlayer model has the worst log loss. This could be interpreted as this model tends to make more incorrect predictions. This provides evidence that the the rest of the players could add incremental gains in win probability, but may not be enough to tip the actual prediction. This is evidenced by similar accuracy in the TS_Player and TS_MaxPlayer models.

Table 2: Best metrics across compared models

Metric	Model			
	SCOPE	TS_Team	TS_Player	TS_MaxPlayer
Accuracy	**0.684**	0.646	0.670	0.670
Calibration	**1.01**	**1.01**	0.986	0.900
Log Loss	**0.094**	0.196	0.232	0.356

Table 3: Parameters used in highest accuracy TrueSkill models

Parameter	Model		
	TS_Team	TS_Player	TS_MaxPlayer
μ	1500	1500	1500
σ	100	150	500
β	100	500	100
τ	10	5	2.5

3.2 Convergence

In Figure 2, the results of the convergence experiments are shown. The ground truth skill over time is shown in red, and the skill predictions for each model are shown in blue. In the graph on the left we observe that the TrueSkill model underestimates the ground truth skill. In the graph on the right we observe that the SCOPE model overestimates the ground truth skill. Based on RMSE the SCOPE model estimates the ground truth skill better. The RMSE for SCOPE is 0.52 and the RMSE for TrueSkill is 1.42. We found convergence to be sensitive to the parameters of the model, but the general trend of TrueSkill under-predicting skill and SCOPE over-predicting tended to hold constant over different parameters. Given this observation we didn't find it necessary to do a rigorous search for optimal parameters.

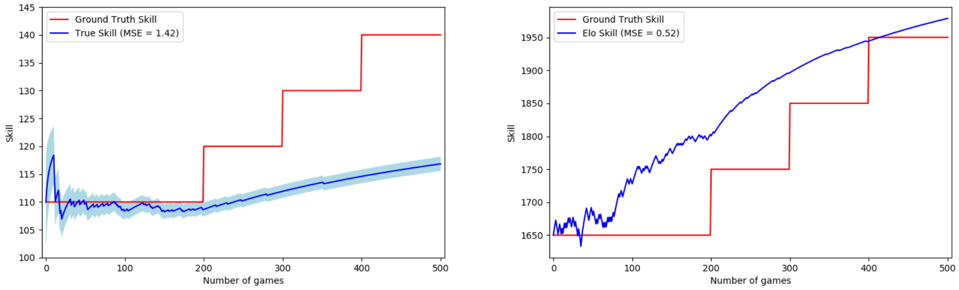

Fig. 2: Convergence results; based on the RMSE the SCOPE model outperforms the TrueSkill model. Notice, that the SCOPE model also appears to give an overestimate where as the TrueSkill model underestimates the ground truth.

4 Future Work

There is an additional TrueSkill model called TrueSkill Over Time that better models how a player's skill changes over time. A good direction for future work would be to compare TrueSkill Over Time to the models explored in this work. The idea behind this model is to add an extra node to the factor graph that accounts for change in skill over time. We hypothesize that this model would have an even lower MSE given the results in Model Based Machine Learning [5].

GitHub Repository: `https://github.com/ajbisberg/trueskill`

References

1. Alexander J. Bisberg, R.E.C.R.: Scope: Selective cross-validation over parameters for elo. In: Proceedings of the Fifteenth Conference on Artificial Intelligence in Interactive Digital Entertainment (2019)
2. Elo, A.E.: The Rating of Chess players Past & Present. Ishi Press (1978)
3. GeneXproTools: https://www.gepsoft.com/gxpt4kb/Chapter10/Section1/SS06.html. Last checked: 2019-12-13
4. Herbrich, R., Minka, T., Graepel, T.: Trueskill^{TM}: A bayesian skill rating system. In: Schölkopf, B., Platt, J.C., Hoffman, T. (eds.) Advances in Neural Information Processing Systems 19, pp. 569–576. MIT Press (2007), http://papers.nips.cc/paper/3079-trueskilltm-a-bayesian-skill-rating-system.pdf
5. John Winn, Christopher M. Bishop, T.D.J.G.Y.Z.: Model Based Machine Learning. Micrsoft (2018), http://mbmlbook.com/index.html
6. Lee, H.: TrueSkill: the video game rating system (2016), https://trueskill.org/, https://trueskill.org/. Last checked: 2019-10-14
7. Microsoft: Xbox live arrives in stores, sparking the next revolution in video games. Microsoft Website (2002), https://news.microsoft.com/2002/11/15/xbox-live-arrives-in-stores-sparking-the-next-revolution-in-video-games/
8. Minka, T., Cleven, R., Zaykov, Y.: Trueskill 2: An improved bayesian skill rating system. Tech. Rep. MSR-TR-2018-8, Microsoft (March 2018), https://www.microsoft.com/en-us/research/publication/trueskill-2-improved-bayesian-skill-rating-system/
9. Satya Nadella, Amy Hood, M.S.: Fourth quareter fiscal year 2019 results, https://view.officeapps.live.com/op/view.aspx?src=https://c.s-microsoft.com/en-us/CMSFiles/SlidesFY19Q4.pptx?version=57904466-7e87-bcd6-1a59-9e4c5e26a761
10. Shacklette, J.: Call of Duty World League Data (2018), https://github.com/Activision/cwl-data, https://github.com/Activision/cwl-data. Last checked: 2019-10-14