Title	局面の難しさを考慮したレーティング推定
Author(s)	西川, 就斗
Citation	
Issue Date	2025-03
Type	Thesis or Dissertation
Text version	author
URL	http://hdl.handle.net/10119/19809
Rights	
Description	Supervisor: 池田 心, 先端科学技術研究科, 修士 (情報科学)

Rating Estimation Considering the Difficulty of Board Positions

2310121 Nishikawa Shuto

In recent years, artificial intelligence (AI) technology has been extensively studied across various fields. In the field of games, powerful AI opponents such as dlshogi for shogi and AZdb for chess have achieved strength surpassing that of top human professionals.

Meanwhile, research is progressing not only on creating strong AI opponents but also on AI that entertains humans. For example, some studies focus on AI players that behave like humans to reduce unnatural behavior and create good opponents or cooperative teammates. Other research aims to estimate player skills and adjust enemy strength accordingly or to generate levels with appropriate difficulty. Among these studies on entertaining AI, we focus on rating estimation, a system that assesses player skill.

A rating system is a method for quantifying player skill numerically and is widely used for various purposes, such as player match-making, in-game handicap settings, and analyzing individual player growth. Therefore, it is important to compute player ratings as accurately as possible. Conventional rating systems, such as the Elo rating, update ratings based on the wins or losses of the matches. As a result, determining an appropriate rating often requires dozens or even more matches.

To estimate ratings with fewer matches, several methods have been proposed that utilize *loss*, a metric representing the magnitude of mistakes made in each move. A smaller loss indicates a stronger player, while a larger loss indicates a weaker player. This approach enables the distinction between crushing and close losses, allowing for a more accurate rating estimation that reflects match content even with a small number of games. Baba et al. further improved the accuracy by having players compete against a shogi AI designed to create close game situations. Additionally, they excluded certain positions according to the number of moves and remaining time, reducing the impact of noise in loss values that do not accurately reflect player skill.

In this work, we further focus on the fact that even players of the same skill level may show different loss values depending on the difficulty of the given positions. A player who faced a larger number of difficult positions by chance may show relatively high loss values, potentially leading to an underestimated rating compared to their actual skill level.

Therefore, in this study, we conducted rating estimation considering position difficulty, using Go as the target game. *Difficulty* is often vaguely defined, but in this research, we adopt *expected loss*, which represents how much error a player is expected to make on average in a given position. The expected loss is calculated as the weighted sum of the loss values of possible moves, where the weights are their predicted selection probabilities. We use this value as a measure of position difficulty. A position with an expected loss close to zero indicates an easy position where mistakes are unlikely or not critical. Conversely, a higher expected loss indicates a difficult position where mistakes are more likely and have more severe consequences.

Using this expected loss, we adjust the loss of the actual move based on the position difficulty. If an actual move's loss is moderate in a position with a high expected loss, it is interpreted as avoiding big mistakes in a difficult position, and the loss is adjusted downward. Conversely, if an actual move's loss is moderate in a position with a low expected loss, it is interpreted as making an unnecessary mistake in an easy position, and the loss is adjusted upward.

First, we analyzed the desirable distribution properties of loss values when using loss-based rating estimation. An ideal loss distribution should accurately reflect player skill with minimal variance, regardless of how the game progresses. Additionally, it is preferable that the distributions of different skill levels (i.e., weaker and stronger players) are well-separated. To evaluate whether these conditions are met, we employed three metrics: stochastic dominance, overlap ratio between distributions, and performance in a trinary classification task.

Next, to validate the appropriateness of the evaluation metrics, we reexamined the effectiveness of the method proposed by Kuboki et al., which demonstrated that using the average loss over the first 150 moves, rather than the entire game, led to more reliable estimations. The results confirmed the effectiveness of this method across all three proposed metrics, thereby also justifying the validity of the metrics themselves.

Finally, we conducted an experiment to estimate player ranks (approximately equivalent to ratings) based on ten game records per player, ranging from 10 kyu to 9 dan. The estimation accuracy was evaluated by measuring the deviation from the players' actual ranks. The results showed that using raw loss values resulted in a root mean squared error (RMSE) of 3.975[rank], while applying the method of Kuboki et al. reduced the RMSE to 1.935[rank]. Our proposed method, which incorporates adjustments based on expected loss, further reduced the RMSE to 1.511[rank].

We expect the findings of this work to be applicable not only to rating estimation in competitive games such as Go but also to various other domains. For instance, they may be utilized in AI-driven personalized teaching that fits students' skill levels, as well as in the generation of interesting puzzles.