ELO Outside of the Competitive Gaming Realm

Ryan Compton
UC-Santa Cruz
rcompton@ucsc.edu

ABSTRACT

This paper will examine the ELO ranking algorithm outside of a competitive gaming scenario. Two experiments are run on ELO using the YouTube slam records. The first experiment examines what ELO score will be given to videos between a variation of competition inputs. The order of competitions between the videos will be varied and the score between the same video will be compared across the different inputs. The second experiment examines the effect of ELO ranking when the same data is read multiple times. This further increases the number of comparisons made across of videos, thus providing more information for an ELO ranking to be made.

Categories and Subject Descriptors

D.3.3 [**Programming Languages**]

General Terms Ranking, ELO score

Keywords: ELO, Ranking, Prediction

1. INTRODUCTION

ELO is a rating system that is used to calculate the relative skill levels of players. This system was originally designed for two player competitive games, like chess [1]. ELO's creation was intended to improve the system of chess player rankings. Within ELO, a player has a ranking that is intended to represent their ability within the game. When a competition occurs between two players, the difference in their rankings is supposed to serve as a predictor of the outcome of a match. A winner is predicted through who has the higher rank; the higher the ranking of a player, the higher their ability.

Performance isn't a concrete measurement with ELO; it is inferred through the player's number of wins and losses. The player's rating is completely dependent on the rating of those players they won to, as well as who they lost too. Scores are instantiated to players that have no recorded games; they are considered a new player. The average score is given to new players, common values for average scores are 1000-1600, which can be chosen arbitrarily. If a player wins a game, their score will be updated using the following formula:

$$R'_A = R_A + K(S_A - E_A).$$

Players can also decrease in score if they lose. The change is score for a loss is calculated from how much the winner's score increases. Thus a losing player will only decrease the amount the winning player's score increases. This keeps the distribution of all scores centered on the average score that is given to new players.

When using ELO to predict outcomes, the expectation of a game between two players *A, B* with rankings of R_A, R_B are found through the following:

$$E_A = \frac{1}{1 + 10^{(R_B - R_A)/400}}.$$

$$E_B = \frac{1}{1 + 10^{(R_A - R_B)/400}}.$$

These are the expectation values for *A* and *B.* They will give the likelihood that the player will win. Since this is a likelihood calculation, the following property exists:

$$E_A + E_B = 1$$

There does remain some questions when using ELO outside of the realm of competitive games like chess. Can ELO be used as a reliable algorithm in predicting the comparison of two objects? Also are there ways to increase the predictive power of ELO toward who will win within a comparison or competition? Finally, will a varied input of data change the ranking of competitors?

This paper will explore these two questions of ELO by running an experiment on ELO using data from YouTube Slam, a competition to find the best YouTube video by having users compare two videos and vote which one they prefer.

2. METHOD

ELO parameters were set before each experiment. The average score was 1600 and a K value was chosen depending on the rank of the winner:

- Winner below 2100: K = 32
- Winner between 2100 and 2400: K = 24
- Winner above 2400: K = 16

These parameters are the proposed parameters by the United States Chess Federation.

YouTube slam also has more than two possibilities with a competing pair of videos. A choice of neither is allowed, however since this is outside of the traditional ELO algorithm, this case will not affect ELO score calculation in these experiments and will be ignored.

2.1 Experiment One

This experiment was to answer the question, does using a different permutation of the original sequence of comparison change the outcome of the ELO ranking?

To conduct this experiment, a file that contained 657 comparisons from YouTube Slam was used as input. This file was shuffled to 20 different permutations. Permutation distance was calculated between each input sequence through the use of the Sequence Matcher object within the Python library difflib. Sequence Matcher has a function called ratio that will calculate the difference between two sequences. Score differences between the different permutations on the same video were also measured.

2.2 Experiment Two

This experiment was to answer the question, can increasing the number of comparisons increase the predictive power of ELO?

The experiment was conducted through reading in a file of video comparisons multiple times. For each read through a prediction was made on which video would be chosen. For each read through the accuracy was found of the predictions made from ELO.

Furthermore, the distribution of scores was examined to see the differences between reading in a file once and reading a file multiple times.

3. RESULTS

3.1 Experiment 1

The amount of score difference in video distributions was found to not be dependent on the permutation distance between each sequence of comparisons.

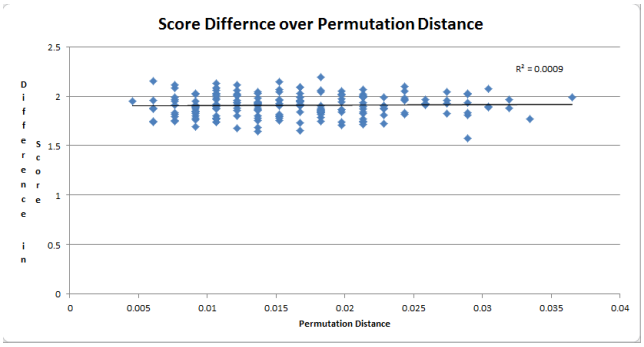

Figure 1. Score Difference over Permutation Distance

Score difference did not change when the comparisons were run in a different order; therefore there exist a distribution of possible scores for a given video that has a very small variation from the video score.

3.2 Experiment 2

The score distribution of reading through the file 20 times had a much higher variance than the distribution of one read through.

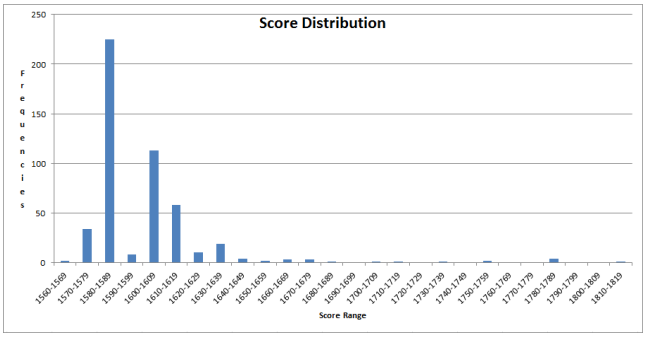

Figure 2. Score Distribution of One Read

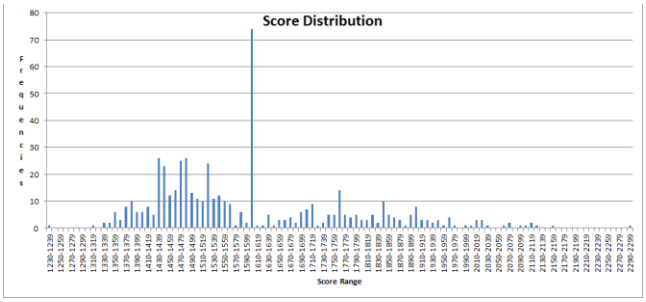

Figure 3. Score Distribution of Multiple Reads

The standard deviation of the single read was around 28.8, while the standard deviation of the 20 reads was around 178.3. Multiple reads is producing a larger variation of scores and therefore allowing more distinguishing ability between a given set of videos.

Accuracy of the predictive model within ELO was able to increase through each read through of the data set. In the first read through, the accuracy was around 51% using the expectation formulas, but this was without any information of the competitors, ELO is intended to be used when information is present, thus the multiple reads of the input. At the 20th read through, the accuracy of the expectation formulas reached around 72%. A large increase in accuracy from the first read through.

Examining the top 12 ranks between the one read through and the multiple read through condition show that different winners were chosen between the two different conditions.

Multiple Reads		One Read, average of score permutations	
Rank	Video	Rank	Video
12th	T99atceVQA4	12th	zgUaX9XMiCw
11th	dMNvQsqZhZE	11th	ZCOPyHVSx9s
10th	zgUaX9XMiCw	10th	KGoS-S-fyF8
9th	yWkPM6_0CTI	9th	ysoD8KG0y90
8th	JxV6JVcEVMc	8th	MF_i2xSc8CM
7th	3bINiS7H3r0	7th	VVgFIlKxZEM
6th	0q4IzXzpseQ	6th	yXlsDjC-hMY
5th	Sp9sgYXDuG8	5th	R-fSDvENqiQ
4th	cNoCf3Lm66U	4th	z7ItSe4Q-gw
3rd	olTBMQq_JM8	3rd	cNoCf3Lm66U
2nd	R-fSDvENqiQ	2nd	maPWrPzo19U
1st	z7ItSe4Q-gw	1st	oeikk3sXlvo

Figure 4. Rankings of one read and multiple read conditions

These two top twelve rankings had a sequence similarity of 0.2695.

Figure 5 shows the results of an additional experiment that was run after finding the previous results of the accuracy and ranking variation between the single read and multiple read conditions. This experiment ran the multiple read conditions until the similarity of one ranking matched 100% of the previous ranking. This experiment found that the ELO ranking will converge on a set of rankings after around 1100 reads of the data set. Predictive power was also measured through each read through and it was found to converge around 78% accuracy after around 900 reads and never changed after this read.

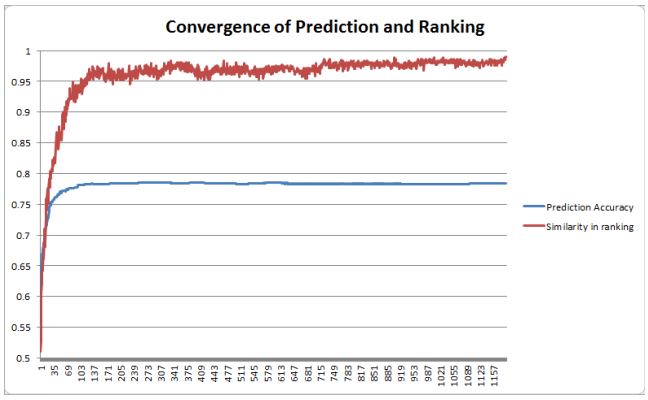

Figure 5. Convergence of Prediction and Ranking

4. DISCUSSION

The results of the two experiments bring an interesting insight into how ELO show be approached if desired to be used as a ranking system outside of the realm of competitive games. The first experiment provides support that ELO is not dependent on the order of the comparison that the outcome of the ranking score will be within a small variation of the average score across multiple possibilities. This experiment also gives an insight into a more effective algorithm to compute scores. Data sets can be repeatedly run on ELO and compute a distribution of a score for a competitor and give insight into the competition.

Despite the first experiments support for ELO being used outside of competitive games, the second experiment provided evidence against using ELO for such a ranking. The second experiment showed that ELO is rather weak given there isn't much data to give a sufficient rank between all competitors. While reading in the data multiple times can produce a more confident outcome, it varies from what the raw data is saying alone and provides a need for further research. While the accuracy increase from multiple reads is interesting, in the manner that this increase was conducted will contain a bias toward what the training data has.

The final experiment does however show a possible solution to this issue. This experiment shows that the rankings and accuracy will converge on a sequence of rankings. This can serve as another alternation to the ELO algorithm to be used on a competition. Finally, the convergence of prediction around 78% does not support using ELO as a predictive algorithm.

ELO may be a good descriptive algorithm as to a set of competitions between players, but as a predictive algorithm, there is not much support. Some future work on ELO can be to modify the algorithm to incorporate a combination of the multiple read condition as well as incorporating a different permutation of the data set for that multiple read. This may create a good variation to increase the convergence of the ELO ranking.

5. REFERENCES

[1] ELO Rating system at Wikipedia.