A Comparison of the Predictive Performance of Dynamic Updating Methods for Chess Player Ratings

Alec G. Stephenson
CSIRO Mathematics, Informatics and Statistics,
Clayton South, Victoria, Australia.

July 25, 2012

The 11th Australasian Conference on Mathematics and Computers in Sport
17–19 September 2012, Melbourne, Australia

Summary

This article uses a large dataset to analyze and compare the predictive performance of different dynamic updating methods for rating chess players. The analysis shows that the simpler Elo system is outperformed by both the Glicko and Stephenson systems. The analysis also suggests that the K factors used in the current FIDE (World Chess Federation) implementation of the Elo system are smaller than what would be needed for optimum predictive performance.

1 Introduction

Updating systems for rating players (i.e. individuals or teams) in two-player games are fast and surprisingly accurate. The idea is that given games played in time period t, the ratings can be updated using only the information about the status of the system at the end of time period $t-1$, so that all games before t can be ignored. The ratings can then be used to predict the result of games at time $t+1$. Comparing the game predictions with the actual results gives a method of evaluating the accuracy of the ratings as an estimate of a player's true skill. There exists more computationally intensive approaches that use the full gaming history via a time decay weighting function (e.g. Sismanis, 2010). These can be more accurate but will not be considered here.

The result of a game is considered to be a value in the interval $[0, 1]$. For chess data, a value of 1 represents a win for white, a value of 0 represents a win for black, and a value of 0.5 represents a draw. The status of the system is typically a small number of features, such as player ratings, player rating standard deviations, and the number of games played. We focus on comparing variations of three basic systems. In increasing order of mathematical complexity, these systems are: the Elo system (Elo, 1978), the Glicko system (Glickman, 1999), and the Stephenson system (Stephenson & Sonas, 2012), which is currently under consideration by FIDE for implementation as the official system for chess player ratings.

The ratings systems considered here derive from statistical models for paired comparisons (e.g. Bradley & Terry, 1952). Preference of one item over another can be related to player preference in two-player games. Draws can be treated as one win and one loss, with each game given a one-half weighting. It is possible to explicitly account for draws using an additional parameter (Davidson, 1970), but this approach does not work well for the dataset considered here. Similar findings for chess data were reported by Joe (1990).

2 Methods

The dataset we analyze here contains approximately one million games played over the eight year period $1999-2007$ by 41 077 chess players. Each record contains the white and black player identifiers, the game result and the month in which the game was played. The dataset was constructed by Jeff Sonas of Chessmetrics and is used with his kind permission. During this period the reporting of individual game results was not required and therefore it contains only a proportion of all games played by FIDE rated players. We also have a dataset of FIDE ratings for 14 118 chess players active at January 1999 which we use to initialize the ratings systems. It may not be an ideal initialization for all systems, but gives a fair method of comparison between them.

We use games from the period $1999-2005$ as training data, games from 2006 as validation data, and games from 2007 as test data. Parameter estimation is performed by minimizing the binomial deviance criterion on predictions for the 2006 data, and we evaluate the performance of different systems using the same criterion applied to the unseen 2007 data. For a single game, the binomial deviance criterion is defined by

$$-[S\log(P) + (1-S)\log(1-P)] \tag{1}$$

where $S \in \{0, 0.5, 1\}$ is the actual game result and $P \in [0, 1]$ is the predicted score. The minimum value is obtained at $P = S$ but predictions of $P = 0$ or $P = 1$ should only be made in cases of 100% certainty, otherwise an infinite value could be obtained. For drawn games the minimum value occurs at $P = 0.5$ and is therefore equal to $-\log(0.5) \approx 0.69$. For the overall criterion, we present the mean of this value across all predicted games, multiplied by a scaling factor of 100.

The basic form of the Elo system tracks only the rating R for each player at each time period. After each period, the rating of a player is updated using $R := R + K\sum_i(S_i - E_i)$ where the sum is over the games that the player plays within the period, S_i is the actual game result and E_i is the expected game result which is based on the current rating of the player and his or her opponent. The Elo system has one global parameter K which is known as the K factor. The Elo system therefore tracks one system parameter (i.e. the rating) and has one global parameter (i.e. the K factor). In practice the system is often applied by making the K factor dependent on additional information on the player such as the player ratings or number of games played, requiring the use of additional system parameters.

Method	Parameters	Valid (2006)	Test (2007)
Stephenson	$c = h = 9$, $\lambda = 2$	61.46	62.31
Glicko	$c = 15$	61.54	62.40
EloG ($G < 30$)	$K = 32$ or 26	61.64	62.40
EloR ($R < 2300$)	$K = 32$ or 26	61.63	62.41
EloP ($G < 30$, $R^* < 2400$)	$K = 30$ or 20 or 15	61.69	62.42
Elo	$K = 27$	61.71	62.47
EloF ($G < 30$, $R^* < 2400$)	$K = 30$ or 15 or 10	61.96	62.64

Table 1: A comparison of predictive performance of dynamic updating methods for chess player ratings. The Valid and Test columns give the binomial deviance values for predictions on the validation and test data. Details of the different methods are given in the text.

The Glicko and Stephenson systems track both the player rating and the player deviation, which is a measure of the accuracy of the player rating as an estimate of true skill. The mathematical details are more complex and are not given here. The Glicko system has a global parameter c which controls the changes in the deviations through time. In the Stephenson system this role is shared by the global parameters c and h. In addition there is a global neighbourhood parameter λ which shrinks the rating of each player to that of his or her opponents, and a global activity parameter b which gives a small per game bonus irrespective of the result. The b parameter improves predictive performance but also creates rating inflation over time. For chess data this is undesirable and so we do not consider it further.

3 Results

The initialization of ratings is an important issue for all systems. It is useful to distinguish between two forms of initialization: the initialization for players who are already known to exist in the player pool before any updates are performed, and the initialization for players who subsequently enter the system during the updates. For the first case, we use FIDE ratings for 14 118 chess players active at January 1999 as our initial ratings. For the second case, we set the rating of any new player to the value 2200. For the Glicko and Stephenson methods, the initial deviation parameters are set to the value 300 for all players. Another issue in chess is that white typically has a small advantage over black, and this can be modelled using a white advantage parameter γ. It is not important to account for this when constructing the player ratings, but it is important to account for it when predicting subsequent games. We use $\gamma = 30$ for this purpose, which seems roughly optimal across all systems.

Table 1 presents the key findings of this article, showing the predictive performance of seven different methods. The seven methods include five different variations of the Elo system, using different methodologies for determining the K Factor. The Elo system is fairly simple, and so several implementations introduce additional complexity by allowing the K factor to depend on additional features. The basic Elo method uses a constant K factor. The methods EloG and EloR use two different K factors. For EloG the K factors are specified according to whether the number of games G played by the player is less than 30, while for EloR they are specified according to whether the player rating R is less than 2300. Lower K factors are typically associated with more experienced or stronger players, so that their ratings have less tendency to change.

The EloF method applies the FIDE implementation of the K factor. This currently specifies $K = 30$ for players with $G < 30$ games, $K = 15$ for players with $G \geq 30$ and whose highest rating ever obtained R^* is less than 2400, and finally $K = 10$ for $G \geq 30$ and $R^* \geq 2400$. Although EloF uses exactly the same K factors as FIDE, it does not implement the initialization system of FIDE, which would require knowledge of the type of tournaments that correspond to the games. Despite this, it can still be used to gain some insight into the FIDE ratings implementation. For all methods other than EloF, the parameters have been chosen to be optimal on validation data predictions (i.e. predictions on games in 2006). The EloP method is the same as the EloF method but where K factor values are optimized on the validation data.

The final column of Table 1 shows the predictive accuracy of each method on the unseen test data (i.e. predictions on games in 2007, using data from the period $1999 - 2006$). We see that Stephenson is best, followed by Glicko, then EloG, EloR, EloP, Elo and EloF. The EloF method has the worst predictive performance. The EloP method outperforms EloF because increasing the K factor by 5 for players who have played 30 or more games gives an increase in predictive accuracy.

The top ten players on 1st January 2007 identified by the Stephenson method are shown in Table 2, selecting from the set of players who have played at least 25 games and have played at least once in 2006. The latter condition removes Garry Kasparov. Figure 1 shows the ratings traced over the period $2001 - 2006$ for these same ten players. Note that all rating systems discussed here are relative rating systems, and therefore the mean of the overall ratings is dependent on the method of initialization used in any particular application. The ranking of both the Glicko (not shown) and Stephenson methods are similar, but for Stephenson the absolute ratings are lower. This is a direct consequence of the neighbourhood parameter λ, which draws player's ratings towards their opponents and therefore prevents spread at both the high and low ends. The histogram of the Stephenson ratings (not shown) is slightly more peaked than for Glicko ratings, and acts more like Elo in the upper tail. When $\lambda = 0$, the overall distributions of Glicko and Stephenson ratings are virtually identical, and therefore λ narrows the spread.

The role of the c parameter in Glicko is to increase the rating deviations over time. In Stephenson this role is shared by c and h, and so c is typically lower in Stephenson than the corresponding parameter in Glicko. This feature appears to make little or no difference to the overall distribution of the ratings, but typically improves predictive performance.

4 Discussion

The Elo system has been in existence for more than 50 years. These results suggest that for chess data, rather than attempting to add complexity to the K factor, a better approach for predictive performance is to use systems such as Glicko or Stephenson, which use a rating deviation value to explicitly model the accuracy of the ratings as an estimate of skill. Under these systems, players who have not played many games may have very high or very low ratings with large rating deviation values. It therefore makes sense to

	Name	Rating	Deviation	Lag
1	Anand, Viswanathan	2759	65	2
2	Kramnik, Vladimir	2757	61	2
3	Topalov, Veselin	2756	59	2
4	Morozevich, Alexander	2755	60	0
5	Ponomariov, Ruslan	2751	60	1
6	Mamedyarov, Shakhriyar	2750	59	1
7	Leko, Peter	2741	61	1
8	Aronian, Levon	2737	60	1
9	Radjabov, Teimour	2731	61	2
10	Polgar, Judit	2728	65	2

Table 2: Stephenson ratings and rating deviations for the top ten chess players, 1st January 2007. The lag value represents the number of months since the player last played a game.

consider a rating official only when the player has played some fixed number of games or when the rating deviation decreases below some fixed threshold.

Bibliography

Bradley, R. A. and Terry, M. E. (1952) The rank analysis of incomplete block designs: I, The method of paired comparisons. *Biometrika*, **39**, 324–345.

Davidson, R. R. (1970) On extending the Bradley-Terry model to accommodate ties in paired comparison experiments. *J. Am. Statist. Ass.*, **65**, 317–328.

Elo, A. (1978) *The Rating of Chessplayers, Past and Present.* Arco. ISBN 0-668-04721-6

Glickman, M. E. (1999) Parameter estimation in large dynamic paired comparison experiments. *Appl. Statist.*, **48**, 377–394.

Joe, H. (1990) Extended use of paired comparison models, with application to chess rankings. *Appl. Statist.*, **39**, 85–93.

Sismanis, Y. (2010) How I won the “Chess Ratings - Elo vs the Rest of the World” competition. arXiv:1012.4571v1

Stephenson, A. G. and Sonas, J. (2012) PlayerRatings: Dynamic Updating Methods For Player Ratings Estimation. R package version 1.0-0.

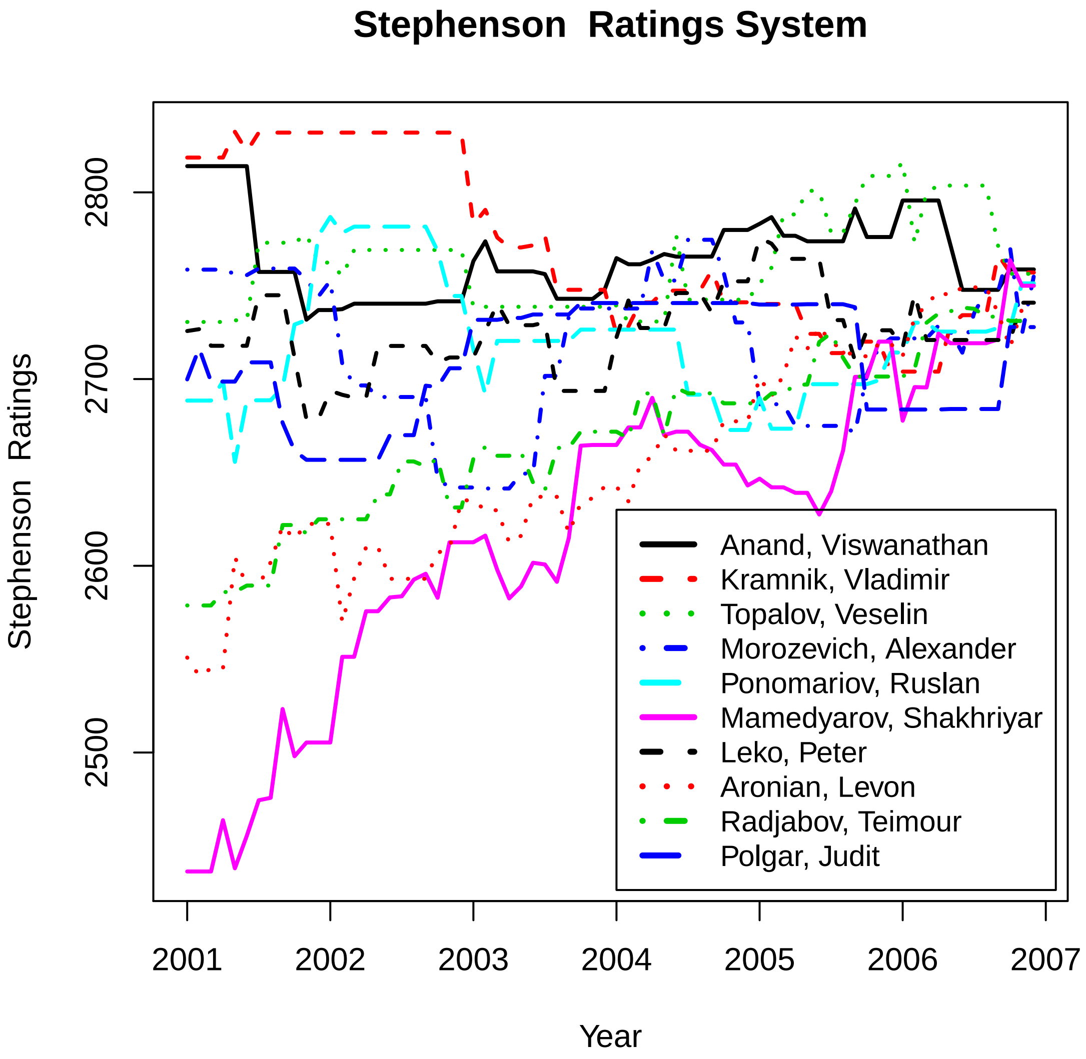

Figure 1: Ratings over time for the ‘current’ (1st Jan 2007) top ten players.