Intransitivities in videogames and matchmaking

CE888 Assignment 2

Raul Axel Suarez Martinez
MSc Student at University of Essex.
Colchester, Essex, UK.
rs16200@essex.ac.uk

***Abstract*— In this paper it is discussed an investigation about intransitivities and how they affect the match ranking. This kind of systems have their base around 30 years ago, based on well stablished statistical models and have been a success in disciplines like chess, but for other types of contests were several factors are taken into account, there could be the possibility of presenting an intransitivity. This project will try to present the scenario where intransitivities are presented, and critical thinking should be applied in order to determine if they are relevant to the context or just are casualties presented due to random matchmaking.**

Keywords—intransitivity; ranking; matchmaking.

I. Introduction

Pairwise comparison or matchmaking (used interchangeably later in the text) is the process of comparing two or more agents in order to select one based on the quantitative value of a feature, or at least sort them relating a ranking. In the area of multiagent systems, it could be used to select the agent with most probability of winning a contest. This is done by modelling the skill to get a value of an agent and compare it with others through a selector. The later example can be easily exemplified through the rankings of two agents compared against each other, where the probability of one agent to outperform another (or the rest of agents in the case of multiple comparison) is determined by a stablished probabilistic distribution guided by a ranking (either probabilistic or normal distributions). But in more complex scenarios this is not the case, due to the difficulty that intransitivities arise in the problem.

An intransitivity is presented in a graph when A → B and B → C but C → A. And a clear example of the later is presented through the well-known game: Rock, Paper, Scissors. In this example, the strongest strategy is impossible to determine due to a paradox in transitivity of power: one strategy always will beat the second, while the second will beat the first; making the optimal policy impossible to deduce by simple inference logic.

Because most of the pairwise comparison methods (between two agents) does not take into account possible intransitivities, they are prone to predict a low probable outcome for some matches that rely on an unconsidered feature given by the lowly ranked agent strengths against a particular most-feasible-to-win opponent. This is a case with regular appearance in the field of multiplayer online videogames, were the game designers strive for general game balance over all the different game classes or gaming policies, in order to overcome the problem of general sense of overwhelming power by the strongest competitors, and resolving on make it possible for the majority of challengers to stand against the meta strategy of the game.

II. Background

The topic of pairwise comparison has been broadly discussed previously, and one of the models that found a direct impact in application and is currently the most used either in its classical form or as the base for more complex approaches is Elo ranking [1]. This system is the base for numerous ranking models and follows the idea that any given player i that show performance $p_i \sim N(p_i;\, s_i, \beta^2)$, and the probability that any given player, lets say player p_1 to uotperform player p_2 is the probability that the performance of $p_1 > p_2$:

$$P(p_1 > p_2 \mid s_1, s_2) = \Phi\left(\frac{s_1 - s_2}{\sqrt{2}\beta}\right)$$

In his work, Elo states that the distribution for the probability for an agent to show its strenght in a match followed a normal distribution, although some other methods use a logistic distribution. Given the positive results of this ranking in chess, the searching for new models which can achieve similar or better performance have been apiering through the years. In [2] they present a bayesian approach to generalize the Elo ranking through factor graphs and using the logistic distribution, achieveing improved performance over the baseline and testing in a competitive videogame environment. This approach is commercially used for the videogame industry.

Another example of rating systems based on the Elo model is the Glicko system [3], proposed by M. Glickman. This variation adds the concept of *reliability* to the model by calculating the rating deviation (RD) of a match to determine how high or low is the quality of the match. The concept behind the later point revolves around considering the number of matches that an agent has taken, relying on the premise that the more games it has, more reliable is going to be its score. The base for RD value of a new player is set on 350, and after each match is necessary to re-calculate it following the next formula:

$$RD_{new} = \min\left(\sqrt{RD_{old}^2 + c^2}, 350\right)$$

The c on this formula is a constant that is responsible for the steepness of the parabolic curve of rating deviation, and is determined depending on the nature of the contest.

These methods however, rely solely on a single scalar to describe the ability of a competitor for a given discipline or game, and there are given circumstances where this does not describe accurately the strengths or weakness that can arise on a particular match.

In [4] they present a learning model which can potentially learn intransivity preference relations in a multidimentional space for pairwise comparison problems. That is one of the few methods which take into account more than single scalar, and they are base on the Bradley-Terry model. Unfortunately, they do not focus on identifying intransivities in the matches but assumed that this intransitivities exists to validate their research of modeling skill in multiple dimensions.

A work conducted in coevolution algorithms that does measure intransitivities [5] provide the metrics and their implementations to measure intransitivity. In this work, they propose 2 ways of doing the measurement: with a transitivity index and with a KL Divergence. The first one consists in a simple index of the number of 3-dags in a tournament divided by the number of posible 3-dags, where a 3-dag is a transitive graph between 3 agents:

$$\tau: \frac{T_3}{\binom{n}{3}}$$

Althouhgt the simplicity of the method, is straightforward to analyse this measurement: the higher it is to 1, there are less intransitivities. As the authors specified, the problem with this index is that it highly reactive and the minor change in the tournament results could affect significantly the number. This is why the second approach is also used, because an stadistical measure could lead to more stable results.

III. Methodology

The selected programming language to produce the experiment is Python 2, as well the use of data analysis frameworks like Numpy and Pandas were minded. The dataset used for the experiments is the StarCraft Alilugac database [6] updated to the 22 of February 2017, which holds records of 400K games and 200K matches of competitive matchups as well as event and player information. StarCraft is an online real-time strategy game created by Blizzard Entertainment [7] which can be played individually or in multiplayer format.

This database contains the information about the players, scores and rankings. The ranking is represented as an id due to the rankings are not the officials by Blizzard but ones generated by the creators of the database and they change through time so a table of only rankings is generated, the method of such rankings is highly based on the Glickman method [7].

The structure for the experiments is divided in 6 steps described below:

Extraction: The information is taken from the database through a connection object provided by the postgress driver for python.

Pre-processing and Preparation: This step is highly coupled with the later one, because the pre-processing was done directly at the moment of extraction. One possibility was loading all the information in memory through a pandas dataframe, but this would imply a huge use of resources for something that would not be necessary (the extra information will be dropped afterwards) so instead, the matches table was joined with the ranking one, and just the pertinent columns were selected through a query.

Identifying intransitivities: For this step the concepts of graphs were represented with 2 dictionaries. The first one is a representation of the players that a certain player previously defeated, with the id as the key and a set of ids of all the opponents defeated as the value. The second one follows the same idea but this time represents all the players that previously defeated a certain player. This way, to find an intransitivity in each match is only necessary to find the intersection of the players that beat the winner of the match which at the same time are defeated by the loser. Likewise, the transitivities are found given the addition of the intersection of the players defeated by the winner that also defeat the loser, and the intersection of the players that defeat both loser and winner and the one where lose to both winner and looser.

Predicting outcome with Glicko model: As previously mentioned, the selected database came with player rankings corresponding to a specific match, this under the Glicko model so instead of developing an Elo ranking process these were taken into account to predict the outcome of a match. The prediction was made under a cumulative logistic distribution, as mentioned in [3], which take as input the difference of the player's rankings.

Calculating KL-Divergence and Transitivity Index: Because the environment where the data was collected is derived from a real scenario and not a controlled experimental one, the number of matches played for a pair of agents is highly variable (from 5 to 56) and not everyone played against every other player like in a Round-Robin tournament, some adjustments had to be done. For the transitivity index, is important to consider that the intransitivities were taken into account even if repeated (by appearance not by statistics), but the actual number could also be higher due to not every player facing all of the others. Considering this, the number of transitive graphs is also included just to be compared and have a reference point. In the case of the KL-Divergence, because not only an expected probability distribution is necessary but also an observed distribution is used, the data had to be filtrated to those matches in which the players face each other at least 30 times in all the database records. This way the actual observed probability can be significant. This last metric is calculated with the entropy function between two distributions.

Analysis: The results from the previous steps were observed and compared against the initial intuition after getting to know the database. This comparison led to some conclusions and improvements for the overall methodology.

IV. Experiment Output

The following figures describe the output of the pipeline across different parts of the experiment:

Table 1: representation of the first 10 rows of the matches dataframe (some columns relating rating adjustment were omitted)

pla_id	plb_id	sca	scb	ra	rb
324	438	0	1	0.00573	-0.01224
1100	163	1	3	0.57341	0.356222
5602	317	1	0	0.630058	0.879652
10953	362	2	0	0.307009	0.367283
493	555	1	2	-0.07617	-0.19623
270	416	1	0	0.19239	0.136472
182	1016	2	3	0.219193	0.10327
41	72	2	1	0.724542	0.580553
671	1223	2	0	0.033435	-0.01031
442	151	2	1	-0.05384	0.284978

Table 2: number of graphs detected in both sets

	#Transitive G.	#Intransitive G.
Complete Dataset	9,884,526	2,460,735
Reduced Dataset	6,262	2,081

Table 3: transitivity index for reduced dataset

Transitivity Index	0.1952157598499062

Table 4: description of a dictionary which contains information about the frequencies of different types of matches after reducing the dataset

Different types	49
Mean (# of games)	36.73469
std	7.702206
min	30
25%	31
50%	33
75%	41
max	56

Table 5: KL-Divergence calculated on reduced dataset

KL-Divergence	0.052765963138623917

V. Discussion

From the previous tables is possible to observe that the number of intransitivities although present, are far from representing a majority. Is notable that the transitivity index on the reduced dataset is closer to the relation *intransitive graphs / transitive graphs* on the complete dataset than on the reduced one. This only reflects the suboptimal use of this metric on this specific set of data: because the number of graphs taken into account are counted per game and not calculated by statistical means (not enough number of matches to determine if one is better over the other for all players), the number of 3-dags formed does not represent the true value intended for this index. On the other hand, the relation *intransitive / transitive* previously mentioned does reflect in a way how many times unexpected results take place. This relation is not completely accurate because there is still a random component in the game which can help a sub-skilled player to beat a stronger one, but a game like Starcraft is highly based on ability and strategy other than luck in order to say that this analysis is completely out of place.

The most important result from the experiment is reflected on the KL-Divergence. Considering the numbers from the later index, is logical to expect a divergence of at least 0.1 from the data, but according to Table 5 this number only reach half of this. There are several reasons why this is happening, the most fact-based would be that because those matches have more instances on average than considering the whole dataset, the true nature of the game stands more clearly due to statistical significance. Nonetheless, a second conjecture can be described: because this data is about professional players which are playing from years, those with the most number of matches or with the highest repeated match against another generally are the oldest ones (in the game) so the system clearly can model their skill in a more precise way than the average player.

VI. Conclusion

In conclusion, the setup and methodology followed in the experiment fails to reflect an optimal analysis on the information presented in the dataset. Due to lack of information (round robin format of matches) the metrics decided for this experiment might not be the better fitted. Is worth to mention that the steps from this point can allow to gain more significance through practices like bootstrapping to generate the missing data. But this would come with more statistical validation.

Another open possibility to improve the experiment could be found using different ranking methods, having more than one approach is more reliable in order to draw a conclusion about the impact of intransitivities in general matchmaking. And a natural extension of the later thought would point to trying to analyse different kinds of games. Taking those steps could lead to a deeper analysis of the behaviour of intransitivities in the terms videogames matchmaking.

VII. References

[1] A. Elo, The Rating of Chess Players, Past and Present, Ishi Press, 2008.

[2] R. Herbrich, T. Minka and T. Graepel, "True Skill: A Bayesian Skill Rating System," *Proceedings of the 19th International Conference on Neural Information Processing Systems*, pp. 569-576, 2006.

[3] M. E. Glickman, "Parameter estimation in large dynamic paired comparison experiments," *Applied Stadistics*, pp. 377-394, 1999.

[4] S. Chen and J. Thorsten, "Modeling Intransitivity in Matchup and Comparison Data," *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining* , pp. 227-236, 2016.

[5] S. Samothrakis, S. Lucas, T. P. Runarsson and D. Robles, "Coevolving Game-Playing Agents: Measuring Performance and Intransitivities," *Spyridon Samothraki*, vol. 17, pp. 213-226, 2013.

[6] "Aligulac," [Online]. Available: http://aligulac.com/about/db/. [Accessed 22 February 2017].

[7] "Starcraft 2 Battle . NET," Blizzard Entertainment, [Online]. Available: http://us.battle.net/sc2/en/game/. [Accessed 22 February 2017].