Progress Report

Daniel Cloutier Sagar Patel

October 22, 2020

1 TODO

Side note: I have text wrapping on in case the lines go super far off screen, that's why.

References if needed. I assume we'll use APA or something. It's pretty easy to do it in LaTeX. I'll have to look it up though, but I know a few people who've done it before.

2 Problem Statement

In this era, there is a massive amount of data collected in a large variety of different topics, of which lots of information can be extracted. The issue here comes with the fact that we cannot simply extract the information by visual inspection. There is currently too much data.

One such domain that contains a large amount of data is the game of chess. Many players like to study chess games to try and improve their play, but it's very hard to find games that are relevant or interesting to learn from amidst the plethora of games to choose from. Although there are many famous known games such as the Opera game of Paul Morphy vs Duke Karl or game 6 from the 1972 world championship match between Robert Fischer vs Boris Spassky, these are usually kept as beginner studies. Also, the fact that they are already well known doesn't necessarily help us discover anything new.

On top of these numerous known games, surely there are many more games that you could learn from. In fact, the website lichess.org logged 68,027,862 games at all levels of play in September of 2020 alone. Some of these could be interesting to study, irrespective of the players ratings not being at the level of Grandmaster. All of this to say, there must

be support from algorithms or machine learning techniques to be able to group similar games of interest together, or show games that lie far outside the norm, which could be interesting studies.

3 Problem Analysis

In order to tackle this problem, we must try and quantify what makes a game more or less similar to others and perhaps what other parameters might be used to determine whether a game would be worth studying or not. (Information from the paper inserted here with their distance function, or how they did the supervised version.)

4 Literature Review

http://pami.uwaterloo.ca/pub/sunym/overview.pdf

'An overview of associative classifier

$https://www.chessprogramming.org/Point_Value_by_Regression_Analysis$

'Very important for piece value changes.

Unsupervised:

They used only 5000 GM games, this is good but we can do better. Lichess database has 1.5 Billion games available to us. Although they are at faster speeds. Standard FIDE clocks are 90+30, 30 minutes added at move 40. The slowest standard lichess game is 30+20. However, we can definitely filter our games because we have access to these.

The distance function. This is fundamentally good, but evaluating chess games on only one dimension to try and group or find outliers is not that telling. Why? Well:

1. Chess Openings (ECO Codes): Games with similar ECO codes will by default be similar. For example, ECO codes B20-B99 are all different variations of which the first moves are 1. e4 c5. This is relevant because if, for example, a game with opening B20 (Sicilian Defense) somehow has a similar distance than multiple games with code C53 (Giuoco Piano; 1. e4 e5 2. Nf3 Nc6 3. Bc4 Bc5 4. c3), that game could turn out to be interesting.

2. Player rating (ELO): With our data, games will be spread out over multiple different skill levels. Because of this, it is important to sort our games by skill level

3. Game Quality: It is possible to be able to sort our games by the "quality" of the game; that is we would define a function that looks at a move and determines whether it is good or bad. Either we can make our own, or we can use something existing. One thing to note is that we're changing the piece values so it is possible to at least attempt to make our own evaluation function if we want to.

4. Game Clock: Some games also have not only the total time for the game, but also have the amount of time a player took per move (starting April 2017 on Lichess). This information can also be used to further classify games. For example, you may not want to look at games in Bullet format (1 minute total to make all moves) if you're trying to improve at a slower time format. But if you want to improve in that format specifically, this might actually be something of value.

5 Subproblems, Statement and Analysis

Subproblem: Not all games have position evaluations, so we will possibly have to do our own position evaluations if we want to have that as a parameter for the unsupervised algorithm. That or we can scrap that idea as it isn't really necessary considering player ELO ratings will be already a good indicator of the quality of games.

6 Algorithmic Sketch, Illustration of Solution

Not sure what we would do here, but we can probably draw something pretty quick in paint (or figure out how to do graphs in LaTeX) as our sketch. It probably isn't that complex... Probably a famous last words moment though.