Project proposal

Adam Temmel (adte1700)
Fredrik Sellgren (frse1700)
Isak Kjellhard (iskj1700)

2021/02/18

1 Introduction

Dota 2 is a team based strategy game that is infamous for its relative complexity compared to competing titles. In the main gamemode of this game two teams of five players are pitted against each other in a very peculiar arena with several different minor objectives and one major objective; destroying the base of the enemy team. The first team to destroy the enemy base is crowned as the winner, but this requires that the players have a good understanding of the game and even better teamwork. A defining characteristic of the title is the usage of *heroes*. Heroes are classified into three different categories, strength heroes that are durable close range combatants, agility heroes that deal a lot of damage, but are instead not as durable and intelligence heroes that rely on using their spells to impact the game. Each player must choose a hero to play as. This sums up to a total of 10 heroes that end up being played each match. Choosing your hero is a quintessential aspect of playing the game effectively, as certain heroes are generally considered better than others in certain scenarios. Therefore, picking a hero that generally underperforms in most situations is usually a bad choice. Naturally every hero has their place in the game, but some end up being more flexible than others.

2 Problem description

The authors of this project proposal enjoy playing Dota 2, but (most of them) are moderately bad at the game. Therefore, they believe that by mining data from a large dataset of games played, they perhaps might be able to get a better understanding of what heroes to pick. In particular the following topics are things they wish to study:

- Which hero is the *best*, in terms of total winrate? (Games won/Games played)
- Which hero from each attribute is the *best*, in terms of total winrate?
- Create three models which, upon inputing the chosen heroes of two teams, predicts which team is most likely to win, based upon hero choices alone.

3 Preprocessing techniques

The dataset[1] has the following attributes:

- *Cluster ID* The cluster ID of which the game was played on. Different clusters belong to different servers.

- *Game Mode* The mode under which the game was played in, such as all pick.
- *Game Type* The type of game that was played, such as ranked.
- *Heroes* The remaining columns 4 to 116 represent the presence of a hero within the match. If the value is 1, the hero was present on Team 1. If the value is -1, the hero was instead present on Team -1. Otherwise the value is 0, meaning the hero was not present in the match.
- *Sum of attributes* As previously stated, each hero belongs to one of three different attributes. It is therefore of interest to insert this information about the team composition into the dataset, in order to properly measure how different combinations of attributes can impact the game.

The first column of the dataset is reserved for the class, which states which of the two teams that won the match. This column exclusively consists of -1 and 1 values, as only one team can win at a time.

This dataset therefore provides several opportunities for preprocessing, such as:

- Translating from hero ID to hero name. When inserting information about the current game, it is more likely that players will want to input actual hero names instead of hero IDs. It is therefore of interest to provide $ID \rightarrow Name$ and $Name \rightarrow ID$ mappings[2].
- Different game types are played on different levels of engagement. Ranked games are generally taken more seriously than other game types, meaning that data from other game types have a higher risk of including misleading data, as players not playing ranked game types might not care as much about winning.
- Closely connected to the following issue, the difference in game modes may have an impact on how a hero is utilized. Certain game modes do not allow a player to choose their hero, instead giving the player a random hero to play. This may lead to that the players participating in the game might not be as experienced in their choice of hero than if they had picked a hero themselves. The added "randomness" of these game modes can therefore lead to misleading data.

4 Datamining techniques

The three classifier models which will be used in the project are as follows:

- KNN
- Random Forest
- Naive Bayes

5 Evaluation

For all models used, a measurement of accuracy between the predicted classifications of the test data and the actual classifications of the test data will be extracted using a confusion matrix. A ROC curve will also be created for each model in order to properly compare them with each other. As a final evaluation method, cross-validation will also be performed for each model created.

References

[1] *Dataset Description.* URL: `https://archive.ics.uci.edu/ml/datasets/Dota2+Games+Results#` (visited on 02/11/2021).

[2] *Hero name to ID mappings.* URL: `https://github.com/kronusme/dota2-api/blob/master/data/heroes.json` (visited on 02/11/2021).