University of
St Andrews

CS5199 - Individual Masters Project

Description, Objectives, Ethics, Resources

September 28, 2018

Supervisor:
John Thomson

Submitted By:
Sizhe Yuen

1 Description

When playing online video games, the identity of the player behind the keyboard is often unknown, even if their account name in game remains the same. This can lead to issues such as account selling and MMR (Matchmaking rating) boosting, where an account's MMR is raised for a price or the account is sold to other players. Furthermore, players who were previously banned for antisocial behaviour can easily make new accounts which takes a long time to detect and ban again. This project looks specifically at the problem of player prediction in the game Dota 2, a popular, free multiplayer online video game by Valve Corporation. Replays of all Dota 2 games are available publicly through Valve's Steam API, which allows full replay `.dem` files to be downloaded. Community tools have also been built around these replays to allow for data gathering.

The goal of this project is to use machine learning models to predict and identify the player behind the keyboard, based on their in-game actions such as mouse movement, item and hero selection. Replays of the same player will be parsed in order to extract data for the models and the different features used for player prediction. The features will be evaluated both independently and in combination to determine which feature or combination of features are most effective in predicting the player in Dota 2.

To extend on the project, once the player is known, it would be interesting to be able to identify their strategy in the game and be able to predict their next actions. Such predications can be an important stepping stone for building sophisticated game AI in online games such as Dota 2 or other strategy games, where there is much hidden information.

2 Objectives

2.1 Primary objectives

2.1.1 Parse Dota 2 replays to generate relevant data

Replay files come in the form of binary `.dem` files, which are meant to be replayed directly using the game. Community tools have been developed to parse replays, but they only act as a libraries for developers to access all the statistics and events that happened in a game. Further processing must be done with these tools to extract the exact data such as mouse movements, gameplay statistics and strategies to be used for the machine learning models.

2.1.2 Binary classification on a single player and hero

Using data from a single player on the same hero, the first step is to run the data through a simple machine learning model such as a Naive Bayes classifier to see how accurately it can predict if it is that player playing on that hero. This simplifies the complexity of the feature space by focusing on the same player and hero to give a strong indication as to which replay features make better predictors.

2.1.3 Multiclass classification of players

The next step after binary classification is to run multiclass classification with a number of players on any hero. Using a Naive Bayes classifier, this will give the probability that the player being analysed is one of the list of players the model was trained on. This will show how the features from binary classification translate to the multiclass scenario.

2.1.4 Evaluation of classification features

Having run classification models with different features, the effectiveness of each feature has to be evaluated both independently and in combination with other features. Looking at features independently is important to determine the difference between game-agnostic features such as player mouse movement and game-specific features such as item or skill selection. This will show which features are more useful in player prediction.

2.2 Secondary objectives

2.2.1 Identification of patterns/strategies from a player's replays

Every player in Dota 2 has different playstyles and strategies specific to them and the state of their current game. For example, buying greedy items at the start of the game or building to counter enemy heroes. Further, there could be unconscious actions such as certain movements or objectives taken by the player. By identifying these additional attributes that are player specific rather than general to all players, the prediction model can be improved and further analysis such as the objective in 2.2.2 can be done.

2.2.2 Prediction of player actions

If the player is accurately identified, it would be interesting to see if their next actions in game can also be predicted based on what is known about the player and the context of the current game. For example from processing the first ten minutes of a game and successfully identifying the player, can the player's build strategy or next objective be predicted.

3 Ethics

This project uses publicly available replays of Dota 2 games to generate data. The data from the replays contain identifiable account names on an opt-in basis. Players who do not consent to allowing external third parties to access their match history appear as anonymous in the data. Furthermore, there is no link between account names in the game and real world ids. The purpose of using this data is to detect people using accounts illegitimately by detecting the change in their in-game behaviour.

4 Resources

This project can be completed using equipment already provided by the school and therefore does not require any special resources.