An analysis of skill tiers in Dota 2

Oisin De Conduin, D12127143

October 16, 2013

1 Introduction

The aims of this project are to analyse the differences between skill tiers in the game Dota 2. This should provide insight into the main differences between the skill tiers, and what the better players are doing differently.

2 Background

Dota 2 is a MOBA type computer game, developed by Valve. It has over 6 million active players playing thousands of games a day. It is a 5 vs 5 team based game, with additional AI controlled units. Each game is separated into 3 skill tiers; Normal, High, and Very High, depending on the Matchmaking Rating (MMR) of the players in the game.

It has also has a large competitive scene, with a recent tournament having a prize pool of over $2.8 million USD.

There are several people who have done analysis on various aspects of Dota 2, and some of their work can be found below:

```
    http://dotabuff.com
http://www.datdota.com
http://dotametrics.wordpress.com/tableofcontents
http://metagamejournal.com/feb-2013-draft-trends
```

3 Aims and Objectives

This project aims to do two things;

Initially, a simple analysis of the main differences between the skill tiers will be done. This will analyse basic statistics such as average Gold Per Minute (GPM), Experience Per Minute (XPM), as well as Kills, Deaths, and Assists per game. Also Hero pick percentages and win percentages will be analysed.

Then a more in depth analysis of the games will be done to see if there is a way of identifying the skill tier of any individual game using the statistics mentioned above.

This will involve creating various classification models and seeing which gives the best prediction. A comparison of the models will also be done.

4   Methodology

An initial dataset of games will need to be created. Valve provides a Web API to access their database of every game ever played. A Python script will be written to query the API and download the relevant data to a database. Relevant information will include GPM, XPM, game length, kills, assists, skill tier and other information. The total number of games for the dataset should be at least 1000, but more could be used.

Once the database has been created, the data must be cleaned. Any game under 20 minutes would be a result of highly unusual circumstances, and should be discarded. This should eliminate games that were not played to completion, or where players left early. Additionally, any game longer than 60 minutes should be checked to see why the game lasted so long.

After the database has been created, and the data cleaned, analysis will be done based on the different skill tiers. The data can be visualised using Python, R, and/or other graphing software. A simple comparison of the main differences between the tiers should suffice. Ideally there should be some identifiable trends that differentiate between the tiers.

Then, an attempt will be made to create a model to classify each skill level using the data. Various methods of classification will be used, including Decision Trees, Neural Networks, and Bayesian Classifications. A comparison of each method will be done to see which is the most effective and why. SAS and R will be used to create these.

5   Deliverables

- A python script used to create the dataset.

- A cleaned and ready to use data set of 10,000 or more games.

- A set visual graphs demonstrating the main differences between skill tiers, and the code used to create them.

- At least three classification models for the data, along with a comparison of them.