DOTA2 Match Result Prediction Based On Hero Lineups

Team 509
Zhaoyin Zhu[1] and Shuang Zhou[2]

[1]Division of Biostatistics, School of Medicine, New York University
[2]Department of Computer Science, New York University

April 14, 2016

1 Introduction

Dota 2 is a free-to-play multiplayer online battle arena (MOBA) video game developed and published by Valve Corporation. The game was released for Microsoft Windows, OS X, and Linux in July 2013, following a Windows-only public beta testing phase that began in 2011, and is the stand-alone sequel to Defense of the Ancients (DotA), a mod (A mod or modification is the alteration of content from a video game in order to make it operate in a manner different from its original version) for Warcraft III: Reign of Chaos and its expansion pack, The Frozen Throne.[1]

Dota 2 is played in matches between two five-player teams, represented by the name “dire” and “radiant”, each of which occupies a stronghold in a corner of the playing field. A team wins by destroying the other side’s ”Ancient” building, located within the opposing stronghold. Each player controls one of 113 playable ”Hero” characters that feature unique powers and styles of play. During a match, the player collects gold, items, and experience points for their Hero, while combating Heroes of the opposite team.[1]

2 Motivation

ESports are getting growing attention and popularity. An increasing number of tournaments are being held, among which one of the most followed is “The International” (TI). In 2015, the fifth TI tounament (TI5) in Seattle had the largest prize pool in eSports history, with a total of $10.9 million.[2] Therefore, it is very worthwhile to research the strategies and tactics within the game to have a better chance against competitions,yet this is a very new ground awaiting discovery. This is the motivation of our project: we want to propose a programmatic stratedy to pick out a lineup that puts a team in the driver seat before the match even begins.

Each match in Dota starts with the players picking their heroes. This stage of a match is much more important than it seems, and very often a bad lineup will pretty much rule the team out. In this stage, two teams take turns to ban and pick heroes, one at a time, until all 10 players have their corresponding heroes. A lot of strategies can be applied during the banning and picking. To name a few, a team may want to ban the heroes that worked really well for the opposing team before, or they want to counter a specific hero in the opposing lineup by picking one that can contain it with a later pick, or they want to achieve a $1+1 > 2$ effect by designing their lineup in a way that heroes supplement each other. Our project is to materialize those strategies and come up with an ideal lineup.

3 Data

The raw data we have is two database tables with very detailed descriptions of 12000 matches. We will only describe the revelant fields in the scope of this project. Table 1 shows the summary for database table MATCHES, and table 2 shows the summary for database table PLAYERS. The two tables will be joined to form one single table that contains all heroes picked within the matches.

field	type	description
match_id	String	identifier for a match
radiant_win	String	0 if dire win, 1 if radiant win
human_players	Integer	number of human players

Table 1: Summary for table: MATCHES

field	type	description
match_id	String	identifier for a match
hero_id	String	identifier for each hero
win	Integer	0 if the player lose, 1 if win

Table 2: Summary for table: PLAYERS

4 Objective

We have established two capstones for this project. The first capstone and the primary objective is to predict the outcome of a match given only the lineups (We described candidate features in the first paragraph of Section 5). Of course this prediction is made under the assumption that both sides perform around the same level (We are not considering the action of players in this project). After the first capstone, we will try to generate a sequence of picks in response to another sequence of picks to simulate the actual hero selection stage.

4.1 Match Outcome Prediction

The objective here is to make a binary prediction indicating whether "radiant" side can win based on the lineups of two sides.

4.2 Pick Sequence Generation

In the hero selection stage, two sides take turn to pick heroes, one at a time. So the objective here is to recommend picks given two partial, possibly empty, lineups. Each recommendation made will take into account the changes of lineups since last recommendation. We won't use Markov Decision Process in this task. The reason is that we don't want to depend only on the previous action, instead, we want to take into account the whole lineup on both sides. We are planning to use a Sequential Forward Selection algorithm, which is a greedy algorithm on each step: Given the heroes already picked, each newly picked hero will generate a different feature subset, which will evaluate to a different probability of winning. And we will select the hero that gives the highest probability for its side in that step.

5 Methods

In order to achieve desirable prediction accuracy, we will include 3 types of features as potential predictors: single heroes (p_1 = 226), joint combinations of two heroes on the same side (p_2 = 12656), joint combinations of two heroes on the opposite side (p_3 = 12656). All the features are 0/1 variables and with these 3 types of features, we are able to account for both synergistic effect and antagonistic effect.

In our dataset, we only have 12000 matches (n = 12000), and the total number candidate predictors $p = p_1 + p_2 + p_3 = 25538$. In this case, we may encounter some difficulties since dimension p is larger than sample size n. The design matrix X is rectangular, having more columns than rows and the matrix $X^T X$ is huge and singular. The maximum spurious correlation between a covariate and the response can be large because of the dimensionality and the fact that an unimportant predictor can be highly correlated with the response variable due to the presence of important predictors associated with the predictor. To reduce dimensions and remove unnecessary predictors on ultrahigh dimensional data, sure independence screen (SIS) which is based on the correlation between predictors and outcome[3] will be used to reduce p close to n. After that, standard variable selection methods like Lasso[4] and adaptive Lasso[5] will be utilized to fit the final model.

In modeling our data, we will try both nonparametric methods (decision tree, random forest) and parametric methods (SVM, logistic regression). In order to evaluate the performance of different methods, the dataset will be randomly split into training set ($n_{training} =$ 10000) and testing set ($n_{test} = 2000$). Sensitivity and specificity will be reported to assess the accuracy of our model, and running time will be presented as well. The most important package for us is Scikit-Learn, and we plan to rely heavily on it.

References

[1] https://en.wikipedia.org/wiki/Dota_2, Wikipedia, accessed at March 21, 2016.

[2] https://en.wikipedia.org/wiki/The_International_2015, Wikipedia, accessed at March 21, 2016.

[3] Fan, J., & Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(5), 849-911.

[4] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 267-288.

[5] Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476), 1418-1429.