Interim Analysis Report
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 Current Work

1.1 Feature Generation

We have implemented the feature generation program that will take a raw input file and produce the following two output:

1. The matrix representing the instances from input, X. X is a $n \times d$ matrix where n is the number of matches and d is the total number of features appeared at least once in the input. Each cell has value 1 or 0, indicating whether a feature is present in this instance.
2. The results of matches, y. y is a vector of length n indicating which side has won the match.

1.2 Classifier Trials

After experimenting multiple classification algorithms, SVM, Random Forest, Decision Trees, Boosting, etc. We have landed on two algorithms for now. We choose Logistic Regression as the parametric classifier, and Random Forest as the non-parametric classifier. Both models we choose achieve a similar accuracy around 60 percent without much parameter tuning.

2 Future Work

2.1 On Feature Selection

After feature generation, we have 11970 features right now and some features might negatively affect our model. To improve the accuracy of our methods, we will conduct feature selection and reduce the dimensions of our data. Several feature selection methods will be considered and the performance of each potential method will be assessed on the testing dataset.

1. Since all the features are binary variables, we might remove the features with low appearance frequencies. The number of features that appear less or equal than 3 times is 757, less or equal than 5 times is 902 and less or equal than 10 times is 1348.
2. We will calculate the correlation coefficients between outcome and each feature, and might remove the features with low correlation (< 0.1, < 0.2, $< 0, 3$).
3. For nonparametric models (e.g. random forest), we might use some importance indices like entropy to select variables and then fit the model.
4. For parametric models (e.g. SVM, logistic), we might use Lasso or adaptive Lasso to select variables and estimate parameters simultaneously.

2.2 On Classification Algorithms

One way to improve the performance of classifiers is to tune the hyperparameters of the models. For example, we can adjust number of trees, depth or number of features to consider in each split in random forest model. Or we can adjust number of rounds for boosting algorithms.