Pages / Home

Kebing Li's CS251 Final Project

Added by Kebing Li, last edited by Kebing Li on May 11, 2018

Final Project Design:

Data Set: UCI Dataset about DOTA2 Games Results

Data Set Information: Dota 2 is a popular computer game with two teams of 5 players. At the start of the game each player chooses a unique hero with different strengths and weaknesses. The dataset is reasonably sparse as only 10 of 113 possible heroes are chosen in a given game. All games were played in a space of 2 hours on the 13th of August, 2016.

Dataset Attribute Information:

Each row of the dataset is a single game with the following features (in the order in the vector):
1. Team won the game (1 or -1)
2. Cluster ID (related to location)
3. Game mode (eg All Pick)
4. Game type (eg. Ranked)
5 - end: Each element is an indicator for a hero. Value of 1 indicates that a player from team '1' played as that hero and '-1' for the other team. Hero can be selected by only one player each game. This means that each row has five '1' and five '-1' values.

Questions to be answered:

1. What is the hero combination that most likely to win or what heroes are more important than others (PCA)
2. Predict the win rate given a specific combination of heros against a specific combination of heros (Machine Learning)

Final Project:

Abstract: For this final project, I will be using the data described above (the dataset and dataset attribution information section) to basically answer the two questions asked in the design stage (the questions to be answered section).

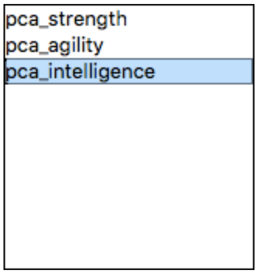
Methods: I will be using PCA Analysis and Classification for my visualizations and analysis.

Actual Implementations:

1) Pre_process data: In order to make my selected dataset fit for the GUI, I have to do some data cleaning before using them. Therefore, I wrote a file called "pre_process", which basically did three things: a) Add headers to the data b) Add types to the data c) Randomly select half of the data points from train data and test data since the original dataset is too large (original data is approximately 22MB)

2) PCA Analysis with respect to hero attributes.

Since Dota2 differentiates all 100+ heroes into three categories: Strength, Agility, and Intelligent, I am trying to use PCA analysis to see which heroes play more important roles in each group. For example, I first did a PCA analysis on all strength heroes, and I clicked "browse the result", which helped me to save the result to a local file called "pca_analysis.csv". From that csv file, we can easily see the eigenvectors that contribute a lot to the total energy. Then, by looking at the presence of each hero in each significant eigenvectors, we can get a sense of which heroes are relatively important within its category group.



a) PCA Analysis on Strength Heroes:

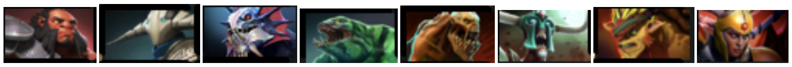
Hero ID: 2, 7, 14, 16, 18, 19, 23, 28, 29, 38, 42, 49, 51, 54, 57, 59, 60, 69, 71, 73, 77, 78, 81, 83, 85, 91, 96, 97, 98, 99, 100, 102, 103, 104, 107, 110

Output File: "pca_analysis0.csv"

Result: The first 24 eigenvectors provide 90% of the energy. Then, I add up all the values for the first 24 eigenvectors for each hero and come up with the result that the following strength heroes are relatively more important in games than other strength heroes:

Hero 2: Axe, Hero 18: Sven, Hero 28: Slardar, Hero 29: Tide Hunter , Hero 54: Life Stealer, Hero 85: Undying, Hero 99: Bristle Back, Hero 104: Legion Commandar (In total 8 heroes)

They are:



b) PCA Analysis on Agility Heroes:

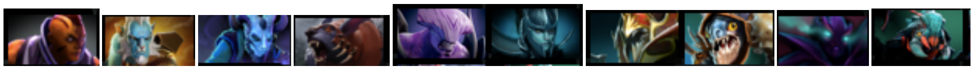
Hero ID:1, 4, 6, 8, 9, 10, 11, 12, 13, 20, 32, 35, 40, 41, 44, 46, 47, 48, 56, 61, 62, 63, 67, 70, 72, 80, 82, 88, 89, 93, 94, 95, 106, 109, 113

Output File: "pca_analysis1.csv"

Result: The first 24 eigenvectors provide 90% of the energy. Then, I add up all the values for the first 24 eigenvectors for each hero and come up with the result that the following agility heroes are relatively more important in games than other agility heroes:

Hero 1: Antimage, Hero 12: Phantom Lancer, Hero 32: Riki, Hero 41: Faceless Void, Hero 44: Phantom Assassin, Hero 63: Weaver, Hero 67: Spectre, Hero 70: Ursa, Hero 88: Nyx Assassin, Hero 93: Slark (In total 10 heroes)

They are:



c) PCA Analysis on Intelligence Heroes:

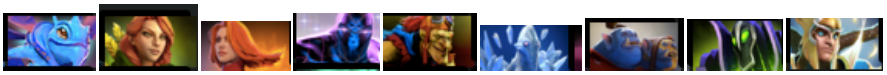
Hero ID: 3, 5, 13, 15, 17, 21, 22, 25, 26, 27, 30, 31, 33, 34, 36, 37, 39, 43, 45, 50, 52, 53, 55, 58, 64, 65, 66, 68, 74, 75, 76, 79, 84, 86, 87, 90, 92, 101, 105, 111, 112

Output File: "pca_analysis2.csv"

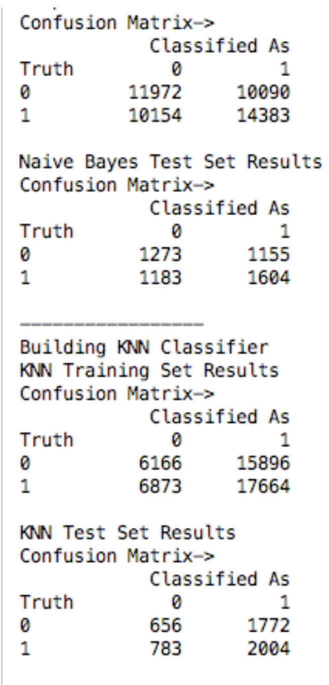
Result: The first 30 eigenvectors provide 90% of the energy. Then, I add up all the values for the first 30 eigenvectors for each hero and come up with the result that the following intelligent heroes are relatively more important in games than other intelligent heroes:

Hero13: Puck, Hero 21: Wind Runner, Hero 25: Lina, Hero 33: Enigma, Hero 65: Bat rider, Hero 68: Ancient Apparition, Hero 84: Orge Magi, Hero 86: Rubick, Hero 101: Skywrath Mage (In total 9 Heroes)

They are:



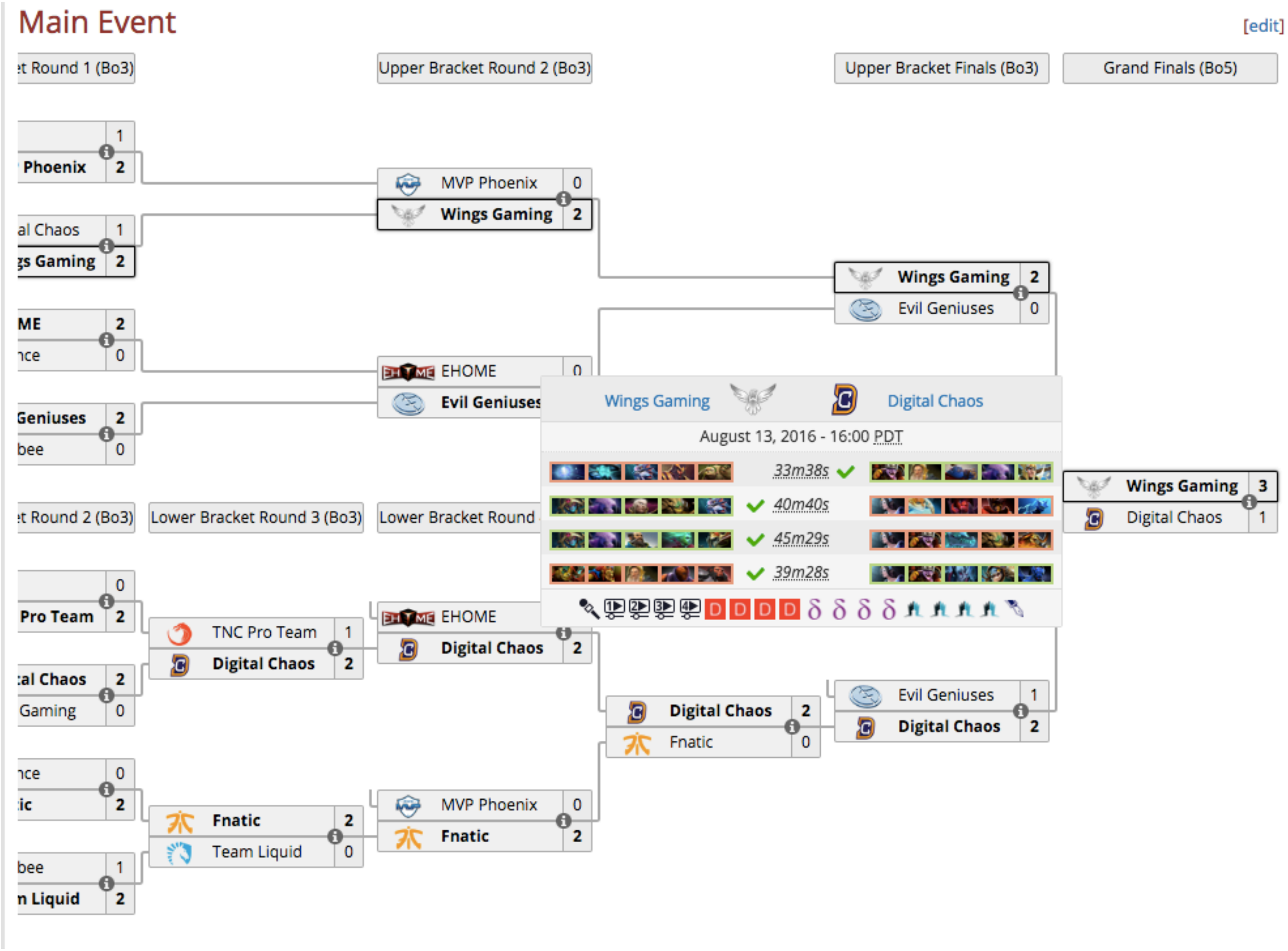
3) Machine Learning Classification: To answer the prediction question, I have to use the machine learning classifiers to do the classification. In this case, I am using the win/lose as two classes and trying to predict the outcome of a match given certain combination of heroes. The following are the confusion matrices:



```
Confusion Matrix->
            Classified As
Truth          0         1
0          11972     10090
1          10154     14383

Naive Bayes Test Set Results
Confusion Matrix->
            Classified As
Truth          0         1
0           1273      1155
1           1183      1604

_________________
Building KNN Classifier
KNN Training Set Results
Confusion Matrix->
            Classified As
Truth          0         1
0           6166     15896
1           6873     17664

KNN Test Set Results
Confusion Matrix->
            Classified As
Truth          0         1
0            656      1772
1            783      2004
```

By looking at the confusion matrix, we notice that neither classifiers does a great job on predicting the winner. This is probably because the data consists of lots of noises and there're tons of other factors that might affect the result of a game such as the difference of players' abilities, their item selections, the internet conditions and so forth. However, if we calculate the accuracy for both classifiers, we will see that in general, it's harder to predict winning than to predict losing. For Naive Bayes, it has an accuracy of 52.43% to predict winning and an accuracy of 57.55% to predict losing. For KNN, it has an accuracy of 27.02% to predict winning and a 71.91% to predict losing. Also, we notice that KNN is really bad at predicting winning, so if we want a more stable prediction, we will choose to use the Naive Bayes Classifier.

4) Prediction of future outcomes. (Based on The International 2016 data)



Main Event [edit]

The most exciting Dota2 event in 2016, which is the year the data was collected, would be the TI2016, a Dota2 world tournament held in Seattle. The above is a screenshot of its format. I will be using the classifier built from the previous tests to predict the final outcomes (the four matches played by Wings Gaming vs. Digital Chaos). I collected the data from: http://liquipedia.net/dota2/The_International/2016#Results, and created a test_ti_data.csv and a test_ti_data_cat.csv file for testing. The training datasets are the same as before. The result confusion matrix is shown below:

```
Confusion Matrix->
               Classified As
Truth            0          1
0                0          1
1                1          2
```

We could see that this result is pretty much lining up with what we would expect. The accuracy of predicting losing (class1) is higher than the accuracy of predicting winning (class0), and the overall accuracy is 50%, which is a little bit lower due to the small sample size. If we have a larger sample size, the accuracy would be higher than 50% even if not so much higher.

Conclusion:

From the PCA analysis, we can realize that the strength heroes are designed to be more balanced up (least more significant heroes than two other categories) while the agility heroes are less balanced up. From the classification analysis, the classifier would predict with an accuracy slightly higher than 50% for a large sample size. If we want to increase the accuracy, we should either include more variables or improve our classifiers to a more complex one.

Thanks To:

I did this project on my own.

The data is from: https://archive.ics.uci.edu/ml/datasets/Dota2+Games+Results

May.11.2018

Like Be the first to like this

Like Be the first to like this

cs251s18project9