The University of Texas At El Paso

CS 4361 Final Semester Project

Predicting Tennis Matches with the use of Machine Learning

Author:

R Noah Padilla

1. Introduction

Tennis is one of the oldest (originated in the 12th century) and most popular sports that is still being played today on an international level. Like other games systems, *for the most part* the traditional tennis game system consists of sets where each set is made up of 6 games, where each game is made up of 4 points, and the winner is the team that earns the most sets (typically best 2 of 3 sets). Also, the game can be played as a singles (1v1) or a doubles (2v2) match. This short report presents a supervised machine learning approach to predict a singles match winner given previous years statistics and the names of an upcoming match based on points, games and sets to those playing in top tier international matches.

2. Approaching the problem

The goal is to predict the winner of a singles match given the previous year's information and the names in an upcoming match, hence a classification problem at hand. Although there are many ways to approach a classification problem, I am choosing the top 2 algorithms I think are best based on previous experience with classifying the MNIST data set.

- Multilayer Perceptron Classifier (Neural Network)
- Random Forest Classifiers (An ensemble of Decision Trees)

3. Feature Selection

The data collected were ATP singles matches ranging from the years 1968 to 2020 that can be found on GitHub (Sackmann) [1] but we will only be looking at the 2018 and 2019 years to make it a feasible problem. Since we are using algorithms that will be supervised, we will need a set of input features (X) with their respective labelled outputs (y). Since mapping each to match to a one hot representation of the two players (index 0 being the higher ranked player and index 1 being the lower ranked player) seems intuitive. Next, being able to represent the 2 players statistics into 1 representable feature would seem reasonable. Luckily merging these

statistics in a particular way have shown to be sufficient in representing the data [2] This merging process will take the difference in the players predicted ranks [3] and the differences of the average for each feature (winners, aces, double faults, etc.) from the previous year. An example of merging the statistics is shown below-

$$\boldsymbol{Rank = Rank_{player\ 0} - Rank_{player\ 1}}$$

$$\boldsymbol{Feature = Avg\ (Feature_{player\ 0)} - Avg\ (Feature_{player\ 1})}$$

Furthermore, tennis can be played on a variety of surfaces such as grass, hard or clay. Tennis surfaces can heavily influence the performance of the players in a match. For example, Rafael Nadal may be the most dominate player on a clay surface but is not as dominate as Roger Federer on grass. To include surface information in our data set, we follow the same process as the rank and other features.

$$\boldsymbol{Court = Wins\ on\ Predicting\ Court_{player\ 0} - Wins\ on\ Predicting\ Court_{player\ 1}}$$

One thing should be noted – before step 3 I had to discard or ignore all matches that had holes or incomplete statistics. Moving on, a similar summary of features was created to that of Sipko's [2]which is shown below –

Feature	Explanation
Rank	ATP rank
POINTS	ATP points
CT	~Percentage of games won on predicting court
FS	~First serve success percentage
W1SA	~Winning on first serve average
W2SA	~Winning on second serve average
WSP	~ Overall winning on serve percentage

TMW	~Percentage of all matches won
ACES	~Average number of aces per game
DF	~Average number of double faults per game
BP	~Percentage of break points won
PCW	~Predicted court percentage wins

A '~' indicates the use of previous years information otherwise the present or predicting state.

4. Testing the models

Unfortunately, I was unable to complete the data processing step due to too many issues with the data that could not be resolved by the deadline.

5. Conclusion

Although I was unable to finish before the deadline, I was able to tailor a method of sufficiently representing the data from Sipko [2] given statistics from the data sets I was using. What I learned from this class project was that most of the work comes from extracting the key features to use in a desired machine learning model. In class our data was very generous, but predicting real world has shown to be very difficult. Even though I was unable to complete this class project, I will continue working on this project on my spare time and can be found on my GitHub [4].

References

[1] Sackmann, J. (n.d.). JeffSackmann/tennis_atp. Retrieved December 09, 2020, from https://github.com/JeffSackmann/tennis_atp

[2] Sipko, M. (2015). Machine Learning for the Prediction of Professional Tennis Matches (1116040262 843058529 W. Knottenbelt, Ed.). 1-64.

[3] S. R. Clarke and D. Dyte. Using official ratings to simulate major tennis tournaments. International Transactions in Operational Research, 7(6):585–594, 2000.

[4] Padilla, N. (2020). NoahTheGr8 - Overview. Retrieved December 11, 2020, from https://github.com/NoahTheGr8