Real Time Tennis Match Prediction Using Machine Learning

Yang "Eddie" Chen, Yubo Tian, Yi Zhong

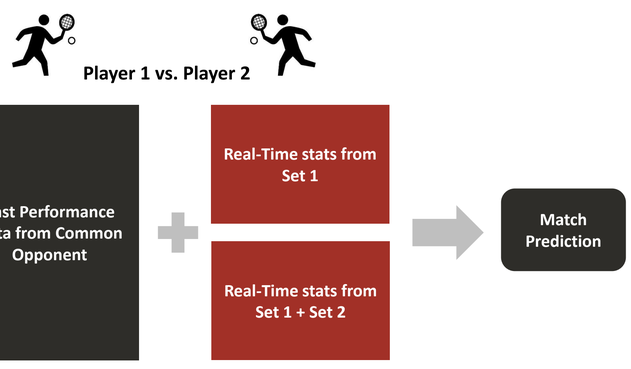
Summary

- Sports bring unpredictability and a lucrative industry trying to predict the unpredictable. Past work on predicting outcome for tennis matches focused on pre-game prediction, we want to apply machine learning to **predict tennis match outcome in-game, after the 1st set and after the 2nd set**
- We have four data models: **historical data only** (pre-game prediction; baseline), **current after set 1 data only** (in-game prediction), **historical + current after set 1 data** (in-game prediction), **historical + current after set 2 data** (in-game prediction).
- **Models explored: logistic regression, support vector classification (SVC) with linear, rbf and poly kernel, neural network, Naïve Bayes, Gauss. Dis. Analysis**
- **Feature selection: recursive feature elimination, principal component analysis**
- **Findings**:
  - **When you are making in-game prediction with 1st set performance data, historical performance data *do not* increase accuracy**
  - **When in-game data are introduced, accuracy and precision both improved**
  - **All models suffer from high bias – feature set does not cover edge cases**
  - ***TTL* (total points won) is too dominant. We need features that cover cases where *TTL* fails to predict**
- **Future work:**
  - **Extract more features on fatigue and age**

Data Source, Cleaning & Transformation

- **Source**: We are using two datasets, "**ATP Tennis Rankings, Results, and Stats**" and "**Tennis Abstract Match Charting Project**". Both datasets are crowdsourced GitHub repository maintained by Jeff Sackmann.
- **Processing**:
  - **Matches from 2000 – 2017 are used as samples (1415 matches)**
  - **No Davis Cup data:** group matches may involve strategic move
  - **Historical performance computed from 14000 matches from 1969 - 2016**
  - **Remove duplicates via match_id**: Since both datasets are manually entered via crowdsourcing, there are some inaccuracy and duplication
  - **Merge Tennis ATP and Match Charting Project**: prediction is made on Match Charting Project, where as historical performance for both players is computed from Tennis ATP. Two datasets are merged on player's first name and last name, with players with the same names removed manually.

Proposed System



Results & Discussion

How well can we predict tennis matches?

Our best model, which uses both historical and in-game performance with SVC predicts with 88% accuracy on dev set when 2 sets have been played

Data Model	Method	Accuracy
Historical Data Only	Logistic Regression	69%
Current, Real-Time Data Only, After Set 1	Logistic Regression	81%
Historical + Current Data (after Set 1)	Linear SVC	81%
Historical + Current Data (after Set 2)	**Linear SVC**	**88%**

Does using current, real-time info improve accuracy?

It does. Moreover, we found that if **we are using current, real-time info, there is really no need for past performance data**. The model that utilizes real-time data after the 1st set is able to have accuracy on par with that plus historical data. This is given the fact the features from current match are fewer than historical data, which have been researched extensively.

Historical Only	TRUE	FALSE
POSITIVE	25.8%	21.4%
NEGATIVE	29.0%	23.7%

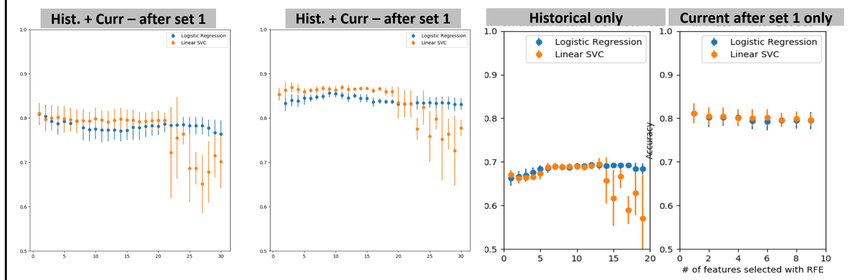
Current Data After Set 1	TRUE	FALSE
POSITIVE	35.0%	7.2%
NEGATIVE	44.1%	13.7%

Historical + After Set 1	TRUE	FALSE
POSITIVE	36.5%	7.0%
NEGATIVE	44.4%	12.2%

Historical + After Set 2	TRUE	FALSE
POSITIVE	40.3%	6.0%
NEGATIVE	42.0%	8.4%

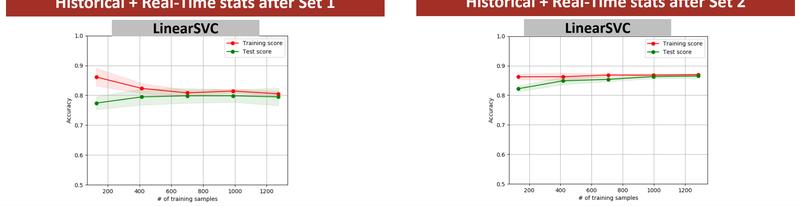
Which features are the most predictive?

- ***TTL* (total points won) thus far in the game:** This is computed using recursive feature elimination across our data models and different prediction model. Adding more features only leads to minimal improvement at best. In most case, having ***TTL*** alone almost always yields the best performance. Other top features include: ***RCV*** (% of return points won) from real-time data, ***BPP*** (breaking point saving %) if only historical performance data are used as features



What's going? Diagnostics with learning curves: High Bias

We observed the problem of **high bias** (high training error; Small gap between training and test error) after plotting the learning curves below:



Method

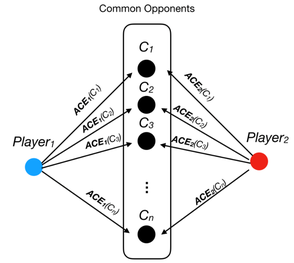
Symmetric Feature Representation

- **Two features for the same metric (*RANK1, RANK2*), or difference between 2 players (*RANKDIFF = RANK1 – RANK2)*?** Based on research [8] [2], we use the difference model for historical performance. For current match performance, we evaluate with both models and found n
used the difference model for all analysis presented unless note

Common Opponent Model for Past Performance

- **Identify a set of common opponents between the two players of a given match**
- **Compute the average performance of both players against the common opponent (*Ci*) for selected features**
- **Take the difference between average performance**

This method is proposed by Knottenbelt and extended by Sipko to construct difference variables capturing player characteristics and past performance, by using their common opponents to achieve a fair comparison.



Edge Cases

- **At least one of the players is new**: remove the samples, which is 84 samples out of 1421 matches
- **When players do not have any common opponent**: we compute the average for each player (regardless of opponent) and take the difference, which is 74 out of 1421 matches.

Feature Lists:

Historical Game Feature	Description
DURATION	Match duration in minutes
SAME_HANDEDNESS	Tracking, on average, the frequency of facing opponents who have the same dominant hand
RANK	ATP rank
RANK_PTS	ATP rank points
HEIGHT	Player height
FSP	First serve success percentage
ACE	Average number of aces per game
DF	Average number of double faults per game
BP_FACED	Average number of break points faced per game
BP_SAVED	Average number of break points saved per game
BPP	Break point saving percentage
SVGM	Average number of serve games
W1SP	Winning on first serve percentage
W2SP	Winning on second serve percentage
SVPT	Average number of serve points per game

Current Game Feature	Description
SAME_HANDEDNESS	1 for same handed, 0 for different handed
FSP	First serve success percentage
ACE	Sum of aces
DF	Sum of double faults
WIN	Number of winners
W1SP	Winning on first serve percentage
W2SP	Winning on second serve percentage
RCV	Percentage of receiving points won
TTL	Total points won
SURFACE	Dummy variable for three types of court
GS	Dummy variable for if tournament is Grand Slam

Future Work

- **To solve for high bias problem, we need more features.** Features to be extracted before the final report include:
  - **Fatigue**: inferred from historical data, through average winning match time and average losing match time, and current match "time" (through number of points played)
  - **Age**: inferred from historical data, through average age of opponents lost to and average age of opponents winning, and the relative age difference at time of match

Reference

[1]Stephen R. Clarke and David Dyte. Using official ratings to simulate major tennis tournaments.International Transactions in Operational Research, 7(6):585 - 594, 2010. [2]O'Malley A. James. Probability formulas and statistical analysis in tennis.Journal of Quantitative Analysis in Sports, 4(2):1-23, April 2008. [3]et al. Jeff Sackmann. The tennis abstract match charting project, 2017. [4]William J. Knottenbelt, Demetris Spanias, and Agnieszka M. Madurska. A common-opponent stochastic model for predicting the outcome of professional tennis matches.Computers and Mathematics with Applications, 64(12):3820 - 3827, 2012. Theory and Practice of Stochastic Modeling. 7 [5]Agnieszka M. Madurska. A set-by-set analysis method for predicting the outcome of professional singles tennis matches. MEng computing- Final year project, Imperial College London, amm208@doc.ic.ac.uk, June 2012. [6]F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python.ournal of Machine Learning Research, 12:2825-2830, 2011. [7]Jeff Sackmann. Atp tennis rankings, results, and stats, 2017. [8]Michal Sipko. Machine learning for the prediction ofprofessional tennis matches. MEng computing - Final year project, Imperial College London, June 2015