Patterns of Play: Predicting tennis match outcomes and player styles

Hunter Hobbs
University of Colorado, Boulder

Nivetha Kesavan
University of Colorado, Boulder

Pratik Revankar
University of Colorado, Boulder

Dmitri Tarasov
University of Colorado, Boulder

ABSTRACT

This paper presents a system of analysis that matches Open Era tennis match statistics for players in the WTA and ATP circuits to predict future match outcomes amongst players and highlight some key rivalries. Players would also be grouped based on playing styles, using clustering techniques, while also modelling an "ideal player" based player data.

ACM Reference Format:
Hunter Hobbs, Nivetha Kesavan, Pratik Revankar, and Dmitri Tarasov. 2020. Patterns of Play: Predicting tennis match outcomes and player styles. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 PROJECT MOTIVATION

The introduction of technology has revolutionized the way we consume and enjoy sports. Instant replays, "Hawk-eye" verdicts, match predictions, forecasts, and player analyses have, in many ways, improved the sport, and the decision making involved. Tennis has recently seen an increase in the use of technology, with major tech players, like IBM and Accenture investing and creating partnerships with ATP, WTA, and all major Grand Slams. With huge investments and use of technology and real-time match statistics, it has become imperative that this data is analysed, understood fully, and made available to the appropriate audiences.

Technology has created a disruption in the sport, and several top ranking players have invested and begun to include analytics to their team, to better understand their game and their opponents. Current world leader, Novak Djokovic is known to leverage match statistics, and his team of physiotherapists, coach and strategy analysts used AI algorithms to improve his game, which played a part in helping him win the 2019 Wimbledon Grand Slam. Judy Murray, coach and mother of Andy Murray, is also a supporter of technology and has heavily relied on a data-driven approach to her coaching.

With an evolving digital audience, data and technology have been leveraged to provide a better experience and give key insights. Data mining and analysis have shown great promise in improving various aspects of the sport, such as decision making by the chair umpire, available data for real-time commentary, pre-match predictions for marketing, creating traction, media hype, and post-match player analysis.

Having the domain knowledge and being huge fans of the sport, we hope to use data mining techniques and the data available to highlight some key insight and showcase the disruptive potential that data and technology has in tennis.

2 LITERATURE REVIEW

2.1 Player rankings for match prediction

Clarke and Dyte [1] used ATP rankings to predict the player's chance of winning in a head to head match. The player rankings are derived from a set of rating points, the rating points of two players prior to the tournament is collected and using the Logistic Regression model they predict the winner of the head to head based on the difference in points rating.

McHale and Morton[6] developed an alternative player ranking based on the past match results, number of points won by each player in a match, how recent the match was played, etc. This ranking was used to predict the winner of a match and it performed better than using ATP rankings. This new ranking focused on ease of win, opponent quality, how recent the match was and the absolute performance of the player whereas ATP ranking gives higher weight-age to how frequently a player participated in different events rather than absolute match performance.

2.2 Points won in a serve

Klassen and Magnus[3] used rankings and points won by players on service to predict the match outcome at the beginning of the match as well as during the match using a graph technique based on the current score and current server.

Liu[5], Newton[7] modelled the probabilities of a player winning a set and match based on the probability of a player winning a point while serving.

2.3 Common Opponent Strategy

Knottenbelt et al. [4] used a hierarchical Markov model to predict the probability of each player winning a match. The model uses the match statistics of an opponent that has been encountered by both players and using the statistics, it computes the probability of each player winning a serve and the match.

2.4 Tactical Analyses in Professional Tennis

Leeuw et. al [2] performed tactical analyses for a specific player using Subgroup Discovery. They used Subset Discovery for the data-set of point,match, and stroke characteristics with 4,13, and 27 features respectively. Using this method, the authors identified the characteristics of a successful serve point of a specific player and they used it to identify the difference between generic successful serve point and serve point of the specific player.

The best subgroup with the highest ROC curve with an area of 0.79.

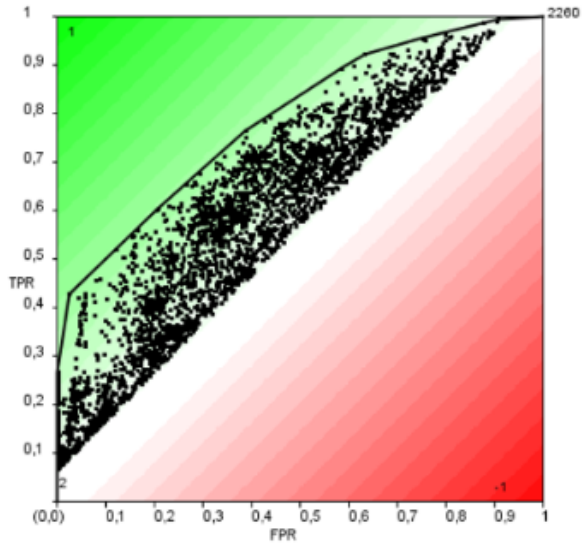

Figure 1: ROC Curve search depth 2

2.5 Modelling ATP Tennis as a Network

Zhang [9] created a network model of ATP matches based on parameters such as win,forehand,backhand, and serve networks. The paper discusses finding common groups with Louvain's Algorithm. Based on the network,the authors discovered the structural roles which revealed the prominent attribute in any individual player's game.

The first experiment was to find the ranking groups based on player ranks which had uninformative results.

2.6 Finding Maximal Non–Redundant Association Rules in Tennis Data

Weidner et.al[8] presents a method of using maximal association rules for tennis data-set. The diagram of the process is described as:

The first two steps is processing the raw data into a time based table of entries. The third step is split into

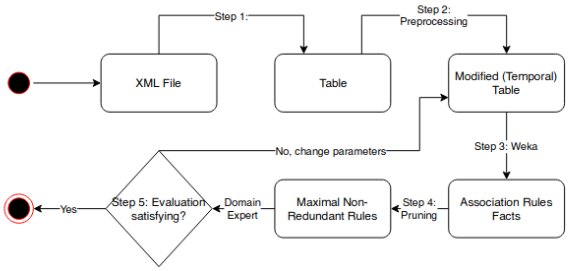

Figure 2: maximal non-redundant association rules process

two parts which are the generation of association rules and the creation of facts. The process of association rules is generated through an algorithm like apriori algorithm. The other part is to use the association rules to establish Prolog facts. The third step is to transform the Prolog facts to maximal non-redundant association rules which might not be present from the apriori algorithm.

3 PROPOSED WORK

As mentioned in Section 1, Tennis data is analyzed by different entities for different use cases. Given the vast amount of Tennis data available over the years, we have identified four different interesting tasks that we hope to explore in this project.

3.1 Match results prediction

Machine learning models are used to predict the match winner. Statistical data for each player, like number of aces, service points, double faults, etc. were used for the prediction. Currently, baselines have been established for models like Support Vector Machines, Decision Trees, K-Nearest Neighbors and AdaBoost. Markov Decision Process will be explored next for the prediction task.

3.2 Clustering for player styles

Used clustering techniques, to mine player style groupings from the ATP data (e.g., offensive, defensive, baseline, volley). The current work that has been completed is a check of the data set which doesn't require cleaning. The current work is aggregating the numeric values for each player. The features would then be clustered to find out if there is any differences between players.

3.3 Matching player rivals

Aggregated player match data to highlight top rivalries in the ATP/WTA tour for the years - 2000 to 2019 (2020 has been excluded since most tournaments after March were suspended due to the coronavirus pandemic). Other metrics for the rival pairs shall be further explored like - number of tournament finals reached, number of titles won, Grand Slam (GS) wins, evolution of player ranking from 2000 to 2019, and court surface performance.

3.4 Player modeling based on playing styles

Model an "ideal player" and a top contender player based on input player stats.

4 DATA SET

The data-sets that are used for the project are the following hyperlinks:

- ATP singles and doubles matches
- Jeff Sackmann data set
- ATP serve and Volley data-set
- WTA matches
- ATP matches data-set
- Tennis Betting Data
- Open Data Soft
- Jeff Sackmann Tennis Match Charting Project

The links to the data are as follows:

- ATP world tour tennis data
- Jeff Sackmann tennis data
- serve and volley
- WTA matches
- ATP matches
- tennis data
- opendatasoft
- Match Charting Project

5 EVALUATION METHODS

The evaluation metrics for each of the four tasks are different which is listed as:

- Predict Match outcomes: have a holdover set for evaluation
- Player Styles: Manual evaluation to see the usefulness of clusters
- Player Rivals: Manual Evaluation
- Player Modeling: Manual Evaluation

6 TOOLS

The tools that would be used for the project is summarized with the list:

- pandas
- sklearn
- tensorflow
- numpy
- matplotlib

7 MILESTONES

7.1 Schedule

- week 1 (18th May - 24th May) - Project Description
- week 2 (25th May - 31st May) - Data Collection and exploration
- week 3 (1st June - 7th June) - Data Exploration and pre-processing
- week 4 (8th June - 14th June) - Data pre-processing and literature review
- week 5 (15th June - 21st June) - Match results prediction and clustering
- week 6 (22nd June - 28th June) - Clustering and matching player rivals
- week 7 (29th June - 5th June) - Player modelling based on player styles
- week 8 (6th July - 12th July) - Player modelling based on player styles
- week 9 (13th July - 19th July) - Project Proposal and evaluation
- week 10 (20th July - 26th July) - Progress Report
- week 11 (27th July - 2nd August) - Refine Paper and work on presentation
- week 12 (3rd August - 9th August) - Project Submission

7.2 Completed

As of 07/24/2020

- Project Description
- Data Collection and exploration
- Data Exploration and pre-processing
- Data pre-processing and literature review
- Project Proposal and evaluation
- Progress Report
- Match results prediction - Baseline
- Partial work on clustering Charting Project data
- Identifying player rival pairs from 2000 to 2019

7.3 TODO

As of 07/24/2020

- Improve on baseline performance for match prediction, Markov-Chain predictions
- Clustering and further metric analysis for rival pairs
- Player modelling based on player styles
- Refine Paper and work on presentation
- Project Submission

8 RESULTS

8.1 Match results prediction

To predict the winner of the ATP matches, Jeff Sackmann's data for the years 2000 to 2020 were used and to predict the winner of WTA matches, Kaggle's WTA data for the years 2000 to 2016 were used. Features like number of aces, service points, double faults, break points, first serves etc.by each player were used for the prediction. both ATP and WTA data have been trained and tested on four different machine learning models - Support Vector Machines, Decision Trees, K Nearest Neighbors and AdaBoost. Grid Search CV was used for Hyperparameter Tuning of the above models.
Performance of the different models for winner prediction on the ATP test data is given below:

Table 1: Model performance on ATP test set

Model	Accuracy	Recall	Precision
Decision Tree	0.783537	0.742337	0.807299
SVM	0.793303	0.803151	0.787237
AdaBoosting	0.785054	0.791282	0.778924
KNN	0.768038	0.809245	0.744365

Performance of the different models for winner prediction on the WTA test data is given below:

Table 2: Model performance on WTA test set

Model	Accuracy	Recall	Precision
Decision Tree	0.658692	0.415689	0.8140531
SVM	0.678462	0.904905	0.6204432
AdaBoosting	0.670779	0.443755	0.8073743
KNN	0. .669386	0.915930	0.611633

As we can see from the tables above, Support Vector Machines has higher accuracy than the other three model.

8.2 Matching player rivals

Player rival pairs from the ATP circuit, for the years, 2000 to 2019 were identified, and tabulated. Data from the *Tennis Betting Data* (http://www.tennis-data.co.uk/alldata.php) was used for this analysis.

Player1	Player2	Players	H2H	Total	Year	Decade
Djokovic N.	Nadal R.	Djokovic-Nadal	28 / 24	52	2019	2010s
Djokovic N.	Federer R.	Djokovic-Federer	26 / 21	47	2019	2010s
Federer R.	Nadal R.	Federer-Nadal	17 / 24	41	2019	2010s
Djokovic N.	Murray A.	Djokovic-Murray	25 / 10	35	2017	2010s
Ferrer D.	Nadal R.	Ferrer-Nadal	6 / 26	32	2019	2010s
Federer R.	Wawrinka S.	Federer-Wawrinka	24 / 3	27	2019	2010s
Berdych T.	Djokovic N.	Berdych-Djokovic	3 / 23	26	2017	2010s
Murray A.	Nadal R.	Murray-Nadal	8 / 17	25	2016	2010s
Federer R.	Murray A.	Federer-Murray	14 / 10	24	2015	2010s
Federer R.	Roddick A.	Federer-Roddick	21 / 3	24	2012	2010s

Figure 3: Top 10 rival pairs from 2000 to 2019

Figure 3 shows the top 10 rivalries, mostly in the *2010* decade, till the year 2019, which has Djokovic-Nadal as the leading rival pair, having faced each other 52 times. Djokovic is leading 28-24 against Nadal. It's interesting to note that the top 3 rivalries are shared between three players in the top 5 seeds, namely, Novak Djokovic(1), Rafael Nadal(2) and Roger Federer(5). It's also interesting to note, that at least one of these players are in every top 10 pair.

Initial exploratory data analysis (EDA) for the *Tennis Betting Data*, revealed the following observations, with regards to the data fields, as shown in Figure 4 and Figure 5.

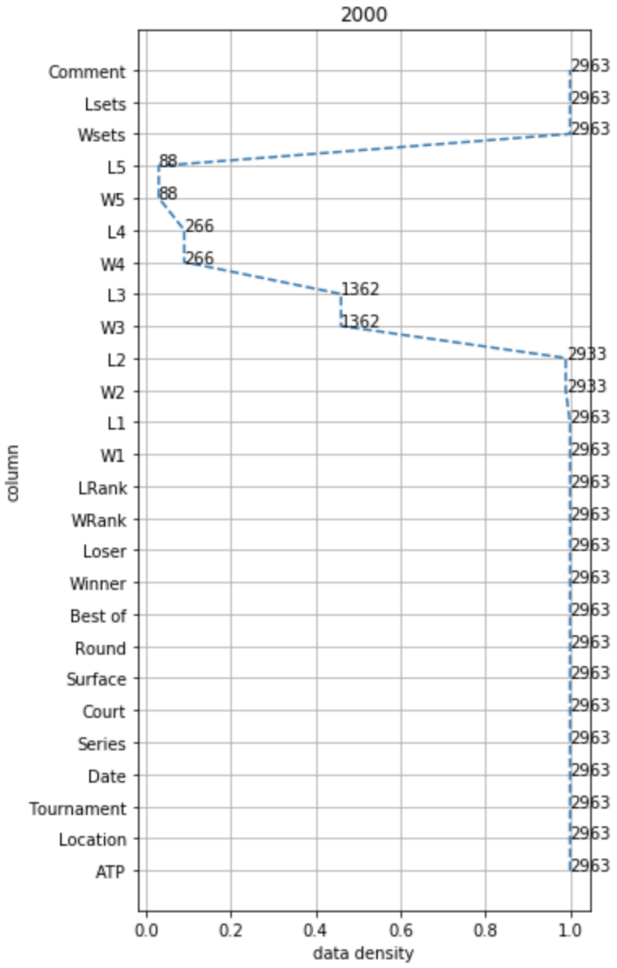

Figure 4: EDA for 2000

Figure 4 shows the distribution of column values for the *Tennis Betting Data*. This shows fairly populated

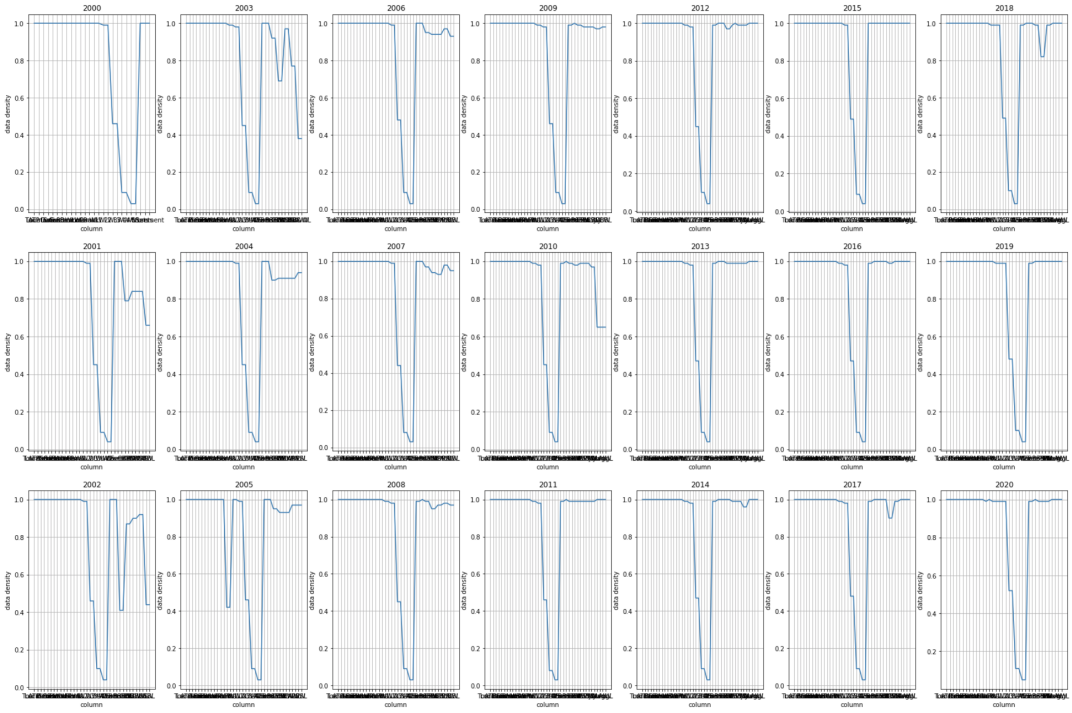

Figure 5: EDA for Tennis Betting Data (2000-2019)

data across all fields, with sparse data for ***W3, L3, W4, L4, W5***, and ***L5***. These columns represent matches that went on to the third, fourth or fifth set and the data suggests that most matches were completed in two sets.

Figure 5 shows similar data distribution across all years from 2000 to 2020.

REFERENCES

[1] S.R. Clarke and D. Dyte. 2000. Using official ratings to simulate major tennis tournaments. *International Transactions in Operational Research* 7, 6 (2000), 585–594. https://doi.org/10.1111/j.1475-3995.2000.tb00218.x arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1475-3995.2000.tb00218.x

[2] Arie-Willem de Leeuw, Aldo Hoekstra, Laurentius Meerhoff, , and Arno Knobbe. 2019. *Tactical Analyses in Professional Tennis.* Stanford University. https://www.researchgate.net/profile/Rens_Meerhoff/publication/335716472_Tactical_Analyses_in_Professional_Tennis/links/5d7790d092851cacdb2e2f10/Tactical-Analyses-in-Professional-Tennis.pdf.

[3] Franc J G M Klaassen and Jan R Magnus. 2001. Are Points in Tennis Independent and Identically Distributed? Evidence From a Dynamic Binary Panel Data Model. *J. Amer. Statist. Assoc.* 96, 454 (2001), 500–509. https://doi.org/10.1198/016214501753168217 arXiv:https://doi.org/10.1198/016214501753168217

[4] William J. Knottenbelt, Demetris Spanias, and Agnieszka M. Madurska. 2012. A common-opponent stochastic model for predicting the outcome of professional tennis matches. *Computers & Mathematics with Applications* 64, 12 (2012), 3820 – 3827. https://doi.org/10.1016/j.camwa.2012.03.005 Theory and Practice of Stochastic Modeling.

[5] Y. Liu. 2001. *Random walks in tennis.* Missouri Journal of Mathematical Sciences, vol.13, no. 3.

[6] Ian McHale and Alex Morton. 2011. A Bradley-Terry type model for forecasting tennis match results. *International Journal of Forecasting* 27, 2 (2011), 619 – 630. https://doi.org/10.1016/j.ijforecast.2010.04.004

[7] Paul Newton and Joseph Keller. 2005. Probability of Winning at Tennis I. Theory and Data. *Studies in Applied Mathematics* 114 (04 2005), 241 – 269. https://doi.org/10.1111/j.0022-2526.2005.01547.x

[8] Daniel Weidner, Martin Atzmueller, and Dietmar Seipel. 2019. *Finding Maximal Non–RedundantAssociation Rules in Tennis Data.* University of Würzburg and Tilburg University. https:

//arxiv.org/pdf/1909.00985.pdf.

[9] Steven Zheng. 2019. *Modelling ATP Tennis as a Network*. Stanford University. http://web.stanford.edu/class/cs224w/project/26380753.pdf.