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

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.

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.

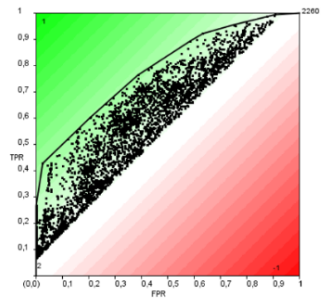

Fig. 1. ROC Curve search depth 2

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:

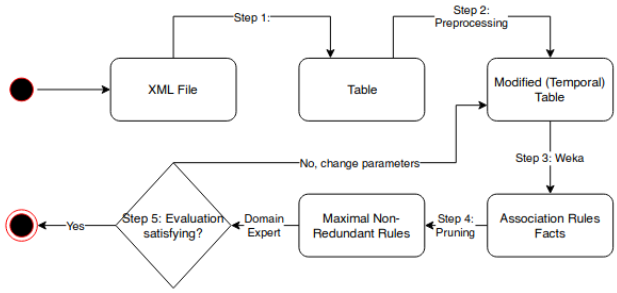

Fig. 2. maximal non-redundant association rules process

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

Using Markov Decision Processes, to predict player head-on match results.

3.2 Clustering for player styles

Using clustering techniques, to mine player style groupings from the ATP data (e.g., offensive, defensive, baseline, volley)

3.3 Matching player rivals

Aggregate and highlight top rivalries in the ATP/WTA tour for the past decade.

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:

- 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:

- https://datahub.io/sports-data/atp-world-tour-tennis-data#resource-player_overviews_unindexed
- https://github.com/JeffSackmann/tennis_atp
- https://github.com/serve-and-volley/atp-world-tour-tennis-data
- https://www.kaggle.com/gmadevs/wta-matches
- https://www.kaggle.com/gmadevs/atp-matches-dataset
- http://www.tennis-data.co.uk/alldata.php
- https://public.opendatasoft.com/explore/dataset/atp/table
- https://github.com/JeffSackmann/tennis_MatchChartingProject

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

- 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

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.