Public Perception of Match Prediction

April 2017:

“Thank the Lord we have the ESPN Win Probability stat to tell us the team that's ahead has a good chance to win.”

-Christian Schneider (@Schneider\_CM) April 17, 2017

“ESPN showing win probability is extremely good. Next up, run expectancy!”

-Neil Weinberg (@NeilWeinberg44) April 4, 2017

“The win probability graphic/discussion on ESPN is literally taking a sword and sticking it through the chest of any fun left in baseball”

-Kenny Ducey (@KennyDucey) April 2,2017

In recent years, win probability has become increasingly prevalent in sports broadcasting. Brian Burke, an NFL commentator, has posted live win-probability graphs of NFL playoff games on his website for the past few years. Earlier this year, ESPN began to post win probabilities atop the score box in televised Major League Baseball games. Despite mixed reactions from fans, as shown above, these developments represent a transition in the modern narrative of sports broadcasting. At it’s simplest, a win probability from any score line communicates how much one team or player is favored to win. Such a metric demonstrates an individual model’s most informed guess at how likely a player is to win the match. Yet, with the recent proliferation of online betting, in-match win probability dictates an entire market of its own. While tennis has yet to broadcast win probabilities, Betfair reportedly traded over $\_\_\_ million in volume over the 2015 Wimbledon Final. Furthermore, around 80% of such betting occurs while matches are in progress. Clearly, in-match win probability is of great concern to all participants in tennis’ betting market. Drawing upon past research and exploring new methods, this paper searches for the most effective approach to in-match forecasting.

However, research papers on in-match prediction remain scattered over the past two decades,

This paper unifies noteworthy prediction models and explores new methods in order to produce the most effective in-match win probabilities.

To comical effect, fans took to Twitter to post wide-ranging reactions such.

To comical effect, fans have taken to Twitter to post a wide range of emotional reactions to televised win-probability forecasts and graphs

-win probability is becoming more prevalent in sports broadcasting today

-the purpose of providing win probability estimates

-data and information regarding tennis betting in high volume across the world

-Brian Burke, nflwin

Introduction

Plenty of research on tennis prediction exists over the past twenty years. In (are points in tennis iid?), Klaassen and Magnus test the assumption that points in a tennis match are independent and identically distributed. Under this assumption, they construct a hierarchical Markov model in conjunction with tennis’ scoring system. From this model, they offer an analytical equation for match-win probability from any score, given each individual player’s probability of winning a point on serve (forecasting). Barnett and Clarke then offer a method of estimating each player’s serve probability from historical data (combining player outcomes). Years later, Bevc proposed updating each player’s serve probability with a beta distribution between each point and computing the corresponding win probability with the above model (predicting the outcome). Recently, Kovalchik assessed performance of 11 different pre-match prediction models on all tour-level ATP matches in 2014.

Over the past several years, Jeff Sackmann (offer footnote to website) has released the largest publicly available tennis dataset via github. This collection contains match summaries of every ATP and WTA match in the Open Era, point-by-point summaries of nearly 100,000 matches—both tour-level and satellite--and a crowd-sourced match-charting project spanning 2800 matches, where volunteers record each shot’s type and direction in every point. While 538 and Kovalchik use Jeff Sackmann’s match data to refine their own elo system for pre-match prediction, none of the aforementioned papers have used his point-by-point dataset.

As these papers have spanned several decades, the datasets referenced among them are not consistent. \_\_\_, \_\_\_\_ both test in-play models on around 500 matches from Wimbledon in 1991-94. Barnett focuses almost exclusively on a marathon match between Andy Roddick and \_\_\_ El Aynoui from the 2002 Australian Open. However, Jeff Sackmann has recently released the largest publicly available tennis dataset via github. This collection contains match summaries of every ATP and WTA match in the Open Era, point-by-point summaries of nearly 100,000 matches—both tour-level and satellite--and a crowd-sourced match-charting project spanning 2800 matches, where volunteers record each shot’s type and direction in every point. While 538 and Kovalchik use Jeff Sackmann’s match data to refine their own elo system for pre-match prediction, no papers on in-match prediction have used his point-by-point dataset. With over 10,000 ATP and WTA tour-level matches from 2010-2017, I attempt to compare existing approaches side-by-side and explore new methods to in-match prediction. Just as Kovalchik does with pre-match prediction models, I attempt to find the most effective in-match prediction models.

This paper combines elo ratings, a wealth of data, and current technology to provide a similar survey of which in-match prediction methods perform the best. I build upon past research by testing variations of previous state-of-the-art methods, and applying several new concepts to these datasets, from successful probability models in football and baseball, to state exploration via hidden Markov Models.

Tennis Scoring

Scoring in tennis consists of three levels: sets, games, and points. At any stage of a match, we can represent the score from player 1’s perspective as (). Consider a match between two entities, p1 and p2. Depending on the match’s format, best-of-three or best-of-five, a player must win two or three sets to win a match.

--show graphic with sets and arrows

To win a set, a player must win six or more games by a margin greater than or equal to two, with a special “tiebreaker” game played at six games all.

-explain the rules, scoring system

-explain the hierarchical Markov model

-allows us to express the importance of specific points and games

-runs on an assumption that points are independent, which according to Klaassen and Magnus (2001), is a fair enough assumption to make

Pre-match prediction:

Since pre-match forecasts form a starting point for in-match win probability models, we first determine what our best pre-match forecast is for a match between two players in our dataset. this paper is intended to carry over Kovalchik’s analysis to in-match prediction.

Elo Ratings:

-describe the equation for elo, explorations with surface elo, weighting elo 10% (or another amount ???) for grand slam matches

-explain why we omit atp rankings, despite their use in past papers (Klaassen Magnus, forecasting). In general, elo is superior to atp rank

ML approaches:

-feature engineering, architecture is really important

-explain results of logistic regression, NN models

-explain random forest approach (Nettleton and Lock), evolved from Burke’s method

-random forest and KNN models might have a much better shot if you can construct a hierarchy in score feature importance (sets >> games >> points), or something to distance them from each other

-then, you could weave in elo ratings to the multi-dimensional feature space of scores

Probability Models:

-naïve pbp model (Jeff Sackmann, Brian Burke also uses 50-50 initial model)

-Klaassen-Magnus model, with 12-month stats

-could also make season-specific 12-month stats \*\*\*

-beta experiments to incorporate in-match serving performance

-explain the James-Stein estimator and shrinkage of 12-month stats

-compare the pre-match win probabilities of those based on 12-month stats with that of elo; it should be much better with just elo/s\_elo (eg Ferrer vs Djokovic at 2011 French Open example)

-given this advantage of using elo in pre-match forecast, generate approximate s\_p and s\_q for each player and evaluate performance of the hierarchical Markov Model estimator with these serving stats

-according to Klaassen and Magnus (2001), we can assume the s\_p+s\_q ~ 1.28 and then generate serving probabilities for a give pi\_a

Unsupervised Learning:

-Can we gain anything by taking back the assumption that points are independent and order doesn’t matter?

-is there any potential for uncovering state transitions in players over the course of matches using the Viterbi (forward-backward) algorithm?

This is research that has yet to be done…

-can an unsupervised deep learning approach work, given an architecture that encapsulate order of points won?

( Klaassen and Magnus were the first to use a hierarchical markov model to represent the sequence of a tennis match. )

Consider a sports match between two entities, x\_1 and x\_2. Before the match, one can use all relevant information about the two entities to generate P(x\_1), a prior win-probability for x\_1. \_\_\_\_ cite how this has been carried out in other sports, prediction mechanisms, betting odds, etc. \_\_\_cite how this has been done in tennis, referencing papers you have read, eg the neural net paper from UK with 4% returns. Also reference the in-point prediction model with probability of points won on serve, updated throughout the match (how can we approach this from a Bayesian perspective?)

Now consider a time series g\_1,g\_2,…g\_n where g\_k contains all information about the match through time k. …formally explain the sub-modular property to my prediction model that must hold in all new methods I develop…

Throughout, I will employ random forests and neural nets, as well as other common tools from the sci-kit learn arsenal. While these models do not adhere to the above property, random forests make predictions based on randomly generated partitions of training data and neural nets adeptly express non-linear relationships between input variables and output, all models I explore in any depth will maintain this property. While it is the case that losing a point, game, or set can be beneficial to a player in some situations (see “tanking”, add a footnote), these situations are rare enough and against the spirit of tennis and the ATP’s rules of conduct. I think any model that tried to capture such instances would likely be overfitting, with few abstract characteristics to latch onto, especially given the sparseness of our datasets here.

Yet, the majority of match prediction papers in tennis deal with pre-match forecasts. Building upon current models, I attempt to build a model that fits posterior in-match probabilities of the form P(x\_1|g\_k), where g\_k is the sequence of match events through time k.

Big thanks to the statistics department for letting me write a thesis on something I find truly interesting and have spent far too much time watching over the past eight years.

This research has mainly been fueled by personal interest and a desire to provide in-match win probabilities as an avid tennis viewer. In reality, an in-match forecasting system that outperforms the betting market does not even promise anyone a get-rich-quick scheme, as commissions and fees on sports betting websites are substantial. Still, I do hope that my work on this project can contribute to the body of knowledge surrounding tennis match predictions and the betting market. At the very least, I hope knowledge gleaned from this work can ultimately make markets a little more informed efficient, an endeavor I picked up from the fine men, young and old, of Susquehanna International Group LLP.

While (most) papers only concern pre-match forecasts, about 80% of tennis betting occurs during matches (reference from ML paper). With billions of dollars traded over exchanges such as Betfair, it is clear that a survey of established methods should hold interest to the betting public.