Roger Federer

DE KANS OM EEN TENNISWEDSTRIJD TE WINNEN

Federer-Nadal in de finale van Wimbledon 2007

De simpelste methode om de winnaar te voospellen, is te kijken naar eerdere prestaties. Dit artikel beschrijft een methode om de winnaar van een tenniswedstrijd te voorspellen, niet alleen bij aanvang van de wedstrijd, maar (juist) ook gedurende de wedstrijd. De kans dat een speler de wedstrijd zal winnen, wordt na elk gespeeld punt bijgesteld en dit leidt tot een kansgrafiek die zich ontrolt tijdens de wedstrijd. De methode is gebaseerd op een snel en flexibel computerprogramma, en op een statistische analyse van een grote dataset van Wimbledon, op wedstrijd- en op puntniveau. We zetten de methode kort uiteen en illustreren hem aan de hand van de Wimbledonfinale tussen Roger Federer en Rafael Nadal in 2007.

Franc J.G.M. Klaassen & Jan R. Magnus

Tijdens een televisie-uitzending van een tenniswedstrijd krijgen de kijkers een aantal statistieken te zien. Uiteraard de stand, maar ook het percentage eerste services die in worden geslagen, het aantal aces, en andere statistieken worden regelmatig getoond op het scherm. De commentatoren bediscussiëren deze statistieken om de kijkers meer inzicht te geven in verschillende aspecten van

de wedstrijd. Een statistiek over het belangrijkste aspect van de wedstrijd, namelijk wie zal winnen, wordt echter niet gegeven. We beschrijven nu een methode om die kans te schatten.

Er bestaat al een aantal methoden om de kans te schatten dat een speler de wedstrijd wint, bij *aanvang* van de wedstrijd. Men kan bijvoorbeeld kijken naar de inleg bij *bookmakers*. Of men kan een statistisch model gebruiken, zoals het model van Clarke en Dyte (2000) dat het aantal punten op de officiële (ATP en WTA) wereldranglijst gebruikt. Stel dat in een wedstrijd tussen speler A en speler B de initiële winkans 70% voor speler A is (en dus 30% voor speler B). Gedurende de wedstrijd komen nieuwe data beschikbaar en die kunnen worden gebruikt om de initiële kans bij te stellen. Als bijvoorbeeld A de eerste set heeft verloren, zal de kans dat A wint afnemen, maar de vraag is met hoeveel. Onze methode (Klaassen en Magnus, 2003) berekent niet alleen de kans dat A wint bij aanvang van de wedstrijd, maar (juist) ook gedurende het verloop van de wedstrijd, bij elk punt. Dit resulteert in een grafiek met elkaar opvolgende kansen dat een speler zal winnen, die langzamerhand zichtbaar worden gedurende de wedstrijd. Als de berekening uitkomt op meer dan 50% kans voor een speler, dan wordt voorspeld dat die speler de wedstrijd zal winnen. Dus de grafiek voorspelt ook de winnaar van de wedstrijd.

De grafiek en de onderliggende kansen zijn informatief voor de televisiekijkers. De score geeft aan wie er momenteel vóór staat in de wedstrijd, maar geeft geen goede indicatie van de mogelijke winnaar van de wedstrijd: een topspeler kan nog steeds de favoriet zijn ook nadat de eerste set is verloren. De score geeft ook maar gedeeltelijke informatie over het verloop van de wedstrijd: een score van 5-5 kan voorafgegaan zijn door 4-4, maar ook door 5-0. De wedstrijd Federer-Nadal die hieronder wordt beschreven is hiervan een goed voorbeeld. Samenvattende statistische gegevens, zoals het percentage eerste services in en het aantal aces dragen hier weinig aan bij. Echter, een schatting van de kans dat A de wedstrijd zal winnen, geeft een directe aanwijzing van de mogelijke winnaar. En de grafiek met hoe groot de kans is dat een speler de wedstrijd wint na elk punt dat gespeeld is, geeft een overzicht van de ontwikkeling van de wedstrijd tot dan toe. Het geeft de informatie in één oogopslag, zodat het nuttig lijkt om de grafiek op televisie te laten zien ter ondersteuning van het commentaar.

Methode

Om de berekening van de kans dat een speler wint, en daarmee de hele grafiek, te bespreken maken we onderscheid tussen de kans vóór de wedstrijd begint (het eerste punt op de grafiek), en de kansen tijdens de wedstrijd (de rest van de grafiek). Om de eerste kans te schatten, gebruiken we een transformatie van de officiële *ranking* (positie op de wereldranglijst) van de spelers. Dit resulteert (bijvoorbeeld) in een kans van 60% dat speler A zal winnen van speler B. Natuurlijk is de *ranking* slechts één van de indicatoren om het onderlinge verschil tussen spelers aan te geven. Als er andere informatie beschikbaar is, zoals het feit dat een speler speciaal goed presteert op gras of het feit dat hij/zij last heeft van een blessure, kunnen aanpassingen plaatsvinden en kan de berekening op basis van de *ranking* worden verfijnd. Uiteindelijk zal er een schatting aan het begin van de wedstrijd zijn van bv. 70%. Klaassen en Magnus (2003) laten zien dat dergelijke aanpassingen uiteraard de grafiek iets doen verschuiven, maar dat het verloop van de grafiek niet veel verandert. Het nut van de grafiek hangt dus niet af van het exacte uitgangspunt.

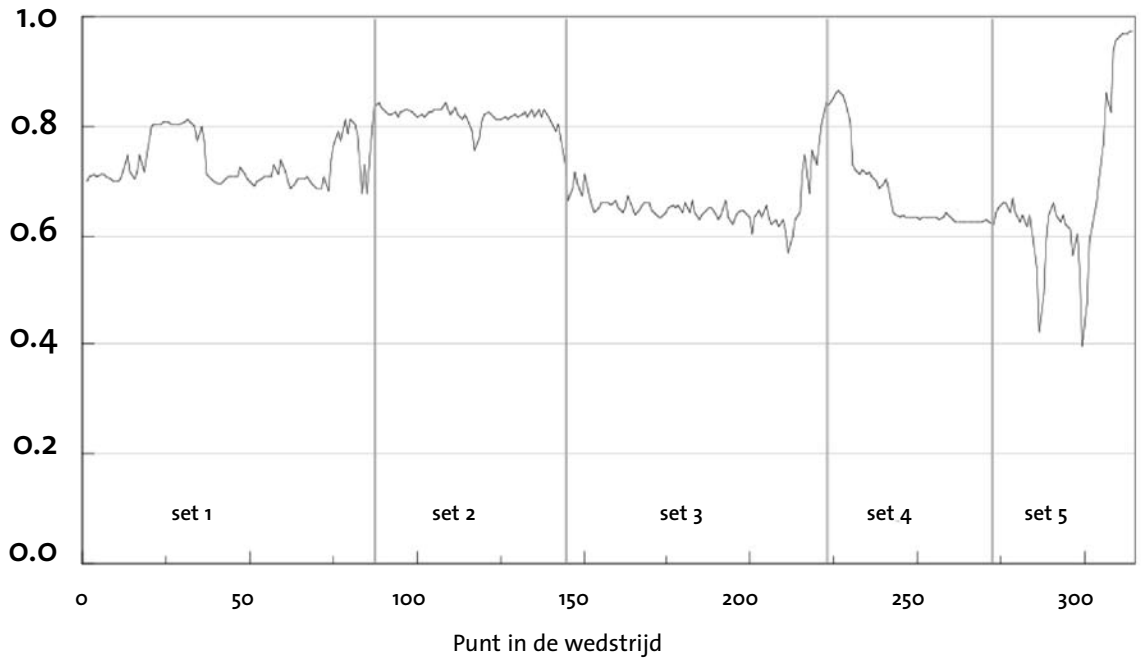
Om de kans te schatten gedurende de wedstrijd, hebben wij een computerprogramma geschreven genaamd Tennisprob. We voeren eerst de specifieke regels van het toernooi in: een wedstrijd om

twee of drie gewonnen sets, wel of geen tiebreak in de laatste set. Gegeven de huidige score, wie er op dat moment serveert, gegeven ook de aanname dat het winnen van een servicepunt een identiek en onafhankelijk verdeeld proces is (zie Klaassen en Magnus, 2001, voor een rechtvaardiging van deze aanname), en gegeven twee in te voeren kansen, berekent Tennisprob de kans dat A wint, en wel op elk gewenst moment in de wedstrijd. Deze kans wordt exact en snel berekend (niet door simulatie, en binnen één seconde). De eerste in te voeren kans is de kans voor de wedstrijd begint, zoals hierboven geschat. De tweede in te voeren kans bestaat uit de som van twee kansen: de kans dat A een punt wint op service en de kans dat B een punt wint op service. Onze methode gebruikt de *ranking* om deze som te schatten. Dit zou bijvoorbeeld 130% kunnen zijn. Deze schatting hoeft niet erg precies te worden geschat omdat de kans die ons interesseert (dat A de wedstrijd wint) hier nauwelijks van afhangt. Deze twee stukjes informatie is alles wat we nodig hebben. In het bijzonder hoeven we niet de onderliggende puntkansen te schatten. Alleen hun som én de initiële kans om de wedstrijd te winnen zijn voldoende. Dit is belangrijk omdat de laatste kansen veel robuuster geschat kunnen worden. Merk ook op dat er geen informatie nodig is over ontwikkelingen in de wedstrijd die na het huidig punt zullen plaatsvinden.

Om te laten zien hoe het voorspellen in de praktijk werkt, analyseren we de Wimbledon 2007 mannenfinale tussen Roger Federer en Rafael Nadal. Voordat de wedstrijd begint, moeten wij twee kansen invoeren zoals hierboven beschreven. De eerste is de kans dat Federer (speler A) de wedstrijd wint voordat de wedstrijd begint. Omdat hij nummer 1 is op de relevante ATP-wereldranglijst en Nadal nummer 2, krijgen we een kans van 60%. We weten echter dat Federer het toernooi de laatste vier achtereenvolgende jaren heeft gewonnen, en dit suggereert dat de schatting op basis van de *ranking* alleen te laag is. Aan de andere kant, in de onderlinge ontmoetingen met Nadal was op dat moment de stand 4-9 (voornamelijk door Nadals overwinningen op gravel, maar Nadals optredens op snelle ondergronden was ook sterk verbeterd). Met deze informatie in het achterhoofd vonden wij een initiële kans van 70% een redelijk uitgangspunt. De tweede ingevoerde kans in 'Tennisprob' is de som van de kansen van beide spelers om een punt te winnen op service. Onze schatting op basis van *ranking* is 136%, daarmee aangevend dat er gemiddeld 68% kans is om een punt te winnen op de service. Dit lijkt redelijk.

Met deze ingevoerde kansen kunnen we de kans berekenen dat Federer de wedstrijd zal winnen, op elk punt in de wedstrijd. Sterker nog, als het punt is gespeeld en de nieuwe stand bekend is, wordt de nieuwe kans binnen één seconde weergegeven. Zo ontstaat een overzicht van de opeenvolgende kansen tijdens de wedstrijd. Omdat de wedstrijd die we als voorbeeld gebruiken al is gespeeld, kunnen we het complete overzicht weergeven, zie Grafiek 1.

Deze grafiek geeft een duidelijk inzicht in de wedstrijd. De eerste set werd door Federer gewonnen, maar de tweede set door Nadal (die Federer brak bij een stand van 4-5). De derde set was weer voor Federer, de vierde set voor Nadal na twee vroege *breaks*. De grafiek laat zien dat aan het begin van de vijfde set (er zijn 271 punten gespeeld) de kans dat Federer wint is gedaald van 70% naar 60%. De stand van de laatste set, 6-2, suggereert dat het een gemakkelijke overwinning was voor Federer. Maar uit de grafiek blijkt dat dit zeker niet het geval was. Sterker nog, op twee momenten was Federer in grote moeilijkheden. In beide gevallen stond hij achter met twee *breakpoints* (15-40) en werd verwacht dat hij de wedstrijd zou gaan verliezen. Maar, hij werkte de *breakpoints* weg en versloeg Nadal uiteindelijk met 7-6 / 4-6 / 7-6 / 2-6 / 6-2.



Figuur 1: Kans dat Federer de wedstrijd wint

Conclusie

De wedstrijd tussen Federer en Nadal is slechts een voorbeeld. Een grafiek kan worden gemaakt voor elke wedstrijd waar we twee variabelen invoeren voor aanvang van de wedstrijd en informatie over elk punt tijdens de wedstrijd. Daarom is de methode zoals hierboven beschreven een algemeen toepasbare voorspellingsmethode. Door informatie te geven over de mogelijke winnaar en het verloop van de wedstrijd geeft de grafiek (Figuur 1) extra informatie, naast de score en de samenvattende statistieken die men normaliter laat zien op de televisie. De informatie wordt ook in één oogopslag zichtbaar en kan direct worden opgeroepen. Daarom zou het interessant zijn om dit op televisie te laten zien, bijvoorbeeld om de twee games als de spelers van kant wisselen. Commentatoren kunnen de grafiek gebruiken om de wedstrijd te becommentariëren, maar ook om deze na afloop te evalueren.

Literatuur

Clarke S.R. en D. Dyte (2000). Using official rating to simulate major tennis tournaments. *International Transactions in Operational Research*; 7: 585-594.

Klaassen F.J.G.M. en J.R. Magnus (2001). Are points in tennis independent and identically distributed? Evidence from a dynamic binary panel data model. *Journal of the American Statistical Association*; 96: 500-509.

Klaassen F.J.G.M. en J.R. Magnus (2003). Forecasting the winner of a tennis match. *European Journal of Operational Research*; 148: 257-267.

Franc J.G.M. Klaassen *is als universitair hoofddocent verbonden aan de Faculteit Economische Wetenschappen en Econometrie van de Universiteit van Amsterdam en het Tinbergen Instituut. E-mail: f.klaassen@uva.nl*

Jan R. Magnus *is als hoogleraar verbonden aan het Departement Econometrie & OR en het CentER van de Universiteit van Tilburg. E-mail: magnus@uvt.nl*