Machine Learning Research
2017; 2(3): 86-98
http://www.sciencepublishinggroup.com/j/mlr
doi: 10.11648/j.mlr.20170203.12

Using Soft Computing Techniques for Prediction of Winners in Tennis Matches

Mateus de Araujo Fernandes

Federal Institute of Education, Science and Technology in Sergipe, Aracaju/SE, Brazil

Email address:
mateus.fernandes@ifs.edu.br

To cite this article:
Mateus de Araujo Fernandes. Using Soft Computing Techniques for Prediction of Winners in Tennis Matches. *Machine Learning Research*. Vol. 2, No. 3, 2017, pp. 86-98. doi: 10.11648/j.mlr.20170203.12

Received: February 24, 2017; **Accepted**: March 20, 2017; **Published**: April 10, 2017

Abstract: The forecast of winners in sports brings valuable information for both organizers, media and audience, and this is particularly important in tennis, where the results of a round in a tournament determine which matches will occur in the next round. With that in mind, this work presents a study of the main factors influencing matches predictability and, from this analysis, a new hybrid approach is proposed to calculate the chances of victory of each of the competitors before the start of a match. A Fuzzy Inference System, with its ability to reproduce knowledge of an expert among mixed information, a Neural Network, with the capability of features extraction from examples, and a Strength Equation with optimized weighting factors are the techniques employed. These predictors have as inputs data from previous performances of the players, which in this case try to capture their short, medium and long-term performances, as well as their affinity for the different types of surfaces. Subsequently the results from these predictors are combined by a voting system. The results are encouraging, showing significant gains when comparing to the use of the ATP ranking.

Keywords: Artificial Intelligence, Forecast, Soft Computing

1. Introduction

Tennis is one of the most popular sports in the world, especially when considering the universe of individual sports. With an annual tour consisting of approximately 800 tournaments spread over 70 countries [1-2], where the most important of those attract millions of viewers and distribute millionaire prizes, this sport has a large and loyal legion of fans and its top players are some of the most popular and well-paid [3] sportsmen of the world.

With this popularity Tennis is moving a continuously ascendant sum of money with tickets, advertising contracts, sporting goods and even bets, not to mention the prizes offered by the tournaments and the value of the athletes' images for publicity. Parallel to this increase in commercial interest that permeates not only tennis, but professional sports in general, there is an increasingly strong presence of quantitative scientific methodologies applied to their analysis. These methods have become indispensable for both players and coaches to analyze performance, strategies, weaknesses, strengths [4], and even aspects of physical conditioning and biomechanics [5], as well to organizers, investors and media, so they are provided with important information for business planning and analysis of economic viability.

In this context, the development of predictors for matches outcomes is one of the lines of studies, aiming to generate data that can be of interest not only for informational use or as a source of incomes from betting, but also for planning the tournaments and their coverage. Predicting the most probable matches to occur in the forthcoming rounds and/or their duration times can, for example, assist in the allocation of attractive games in the major courts and at the best times, allow forecasts of public and audience, and even uphold merchandising actions [6-7].

In the literature, several studies deal with the predictability of results employing the most diverse approaches, including point-by-point analyzes during the match and predictions of winners before the start of each match.

The works of Clowes et al. [8] and Klaasen and Magnus [6] are some of those that are based on point-by-point analysis, focusing not only on the forecast before the beginning of the match but also – and especially – during its course, with simulations based on the probability of the player who's serving to win the next point. Knottenbelt et al. [9] also presented a predictor for matches with analysis after every point, however, adding information on the performance of the players involved against a common opponent in the past. This is made in order to eliminate the bias that exists in service statistics: stronger players, because usually they more often advance to the final rounds of tournaments, confront, in an average, stronger opponents.

Clarke and Dyte [10] set a logistic regression model to calculate the probability of winning a set based on the differences in ranking points between players. This model was used to forecast matches outcomes and to simulate tournaments.

These works rely on the hypothesis that the points or sets played are independent and identically distributed (i. i. d.), with this meaning that previous results do not exert influence on the forthcoming results. However, the work of Klaasen e Magnus [11] discuss the validity of this hypothesis, concluding that winning the previous point has a positive influence on winning the current point, and that at pressure points the servers are negatively affected, what seems to be more verisimilar.

In the approach proposed by del Corral and Prieto-Rodríguez [12], consisting of a prediction for winners in Grand Slam matches without sticking to the scoreboard, the analyzed variables of influence on the results of matches were the surface type and physical characteristics of the competitors, in addition to the ranking of both players. The work of McHale and Morton [7] perform predictions using a Bradley-Terry model (based on pairwise data comparisons) adjusted from previous results and on the surface where the matches were played. Meanwhile, Scheibehenne and Bröder [13] show that it is possible to obtain good correct prediction rates only with the recognition of players' names by an audience not necessarily specialized.

In the present work the approach adopted is also intended to give forecasts of matches outcomes before the first ball is thrown up and not taking into consideration any events that may occur during its course. For this purpose, three different predictors are proposed: the first employing a Fuzzy Inference System based on memberships and rules that attempt to mimic the knowledge of an expert, the second using an equation can calculate a "strength" factor for each player at a specific tournament, based on previous performance and optimized weighting factors, and the third using an Artificial Neural Network and exploring its capabilities of learning and feature extraction from training sets composed by a database of matches. To make the best of the power of these techniques, a previous study is done trying to provide an insight on some quantitative performance factors and their correlation with the belief in who will be the winner of a particular match (and with extensions to championships). At the end, the outcomes of these predictors are combined in one by a majority vote system.

The general framework and the dataset acquisition are presented on Section 2.1, followed by the analysis of the influence factor on matches' predictability shown in Section 2.2. Afterwards, the Section 2.3 contains the implementation details of the soft computing techniques employed for prediction. Subsequently, the results and their analysis are presented on Section 3 for the matches' predictors, with comparisons to real results. Finally, Section 4 brings the final discussions and the concluding remarks.

2. Method

2.1. Dataset Acquisition

The development of the quantitative methods proposed relies on a database composed with statistics of 220 active players in the men's professional circuit in the years 2014 and 2015. This data is made available by the Association of Tennis Professionals (ATP) on its official website [1] and consist of:

a) Number of titles accumulated during a player's career;
b) Career Ratio – Fraction of overall matches won throughout a player's career;
c) Grass, Clay and Hard Ratios – Fraction of wins on the different surfaces;
d) Grand Slam Ratio – Fraction of matches won in the four main tournaments, disputed in best of five sets;
e) Last 10 – Fraction of wins in the most recent matches.

These statistics comprise only results from matches played in the main draws of tournaments at ATP and Grand Slam levels, i.e., results in tournaments of lower levels (Challengers, Futures, and Qualifiers) are not considered in order to standardize the difficulty levels of the matches and maintain an equality in the comparisons. The data employed was updated at different moments in order to be consistent with the required forecasts.

With that in hand, the study begins with an evaluation of players' performance data with the purpose of discovering what factors/parameters give the major contribution to more efficient results forecasts. This evaluation will maximize the predictive capabilities of the Soft Computing techniques, while defining the best variables to be used as inputs for the predictors.

An additional database with the results of all matches of these levels played in the most recent seasons was obtained from [14]. This dataset also gives the position and points in the ATP entries ranking for all the players, updated prior to the start of each tournament. During the development of the predictors, with the implementation details discussed in Section 2.3, the dataset relative to the tournaments disputed in 2014 are used for training and adjustments. The dataset relative to the 2015's tournaments is used for testing, being presented only to the final versions of the predictors and allowing a comparative study of their performance on matches outcomes predictions.

2.2 Analysis of Influence Factors in the Forecasts

2.2.1. Ranking Influence

The ATP entry ranking [1] is responsible for classifying professional tennis players based on their points accumulated in the tournaments played through the last 52 weeks, with the purpose of defining the admissions and the draws for the forthcoming tournaments. In this work, its use is proposed as a medium-term performance measure for the players.

With regard to the ranking information utilized in the predictions, an interesting observation is that, based on statistics of ATP's matches, can be noted a strong tendency that the difficulty encountered by a tennis player to win an opponent seems to increase in steps increasingly wider as the ranking of these opponents approaches the pinnacle. In other words, it is much more common a victory of the 120th classified against the 101th than a win of the 20th against the leader of the ranking, although the difference in positions are the same. Therefore, the strength relationship that seems to exist is not linked only to the ranking position, which induces to think of a model involving a non-linear mathematical relationship.

To better understand this trend, the graph shown in Figure 1 presents curves of the ranking points *versus* the ranking position for five different dates between the years 2012 and 2015 (a period without changes in the criteria for points distribution in tournaments). When these curves are analyzed, the obtained relationship seems to be similar to the aforementioned trend, with differences in points increasingly higher as we approach the top of the ranking. That makes sense in the way the ranking was designed, considering that, as the difficulty of opponents tends to grow rapidly, the ranking points awarded to a player for each advanced round in a tournament grow geometrically [1].

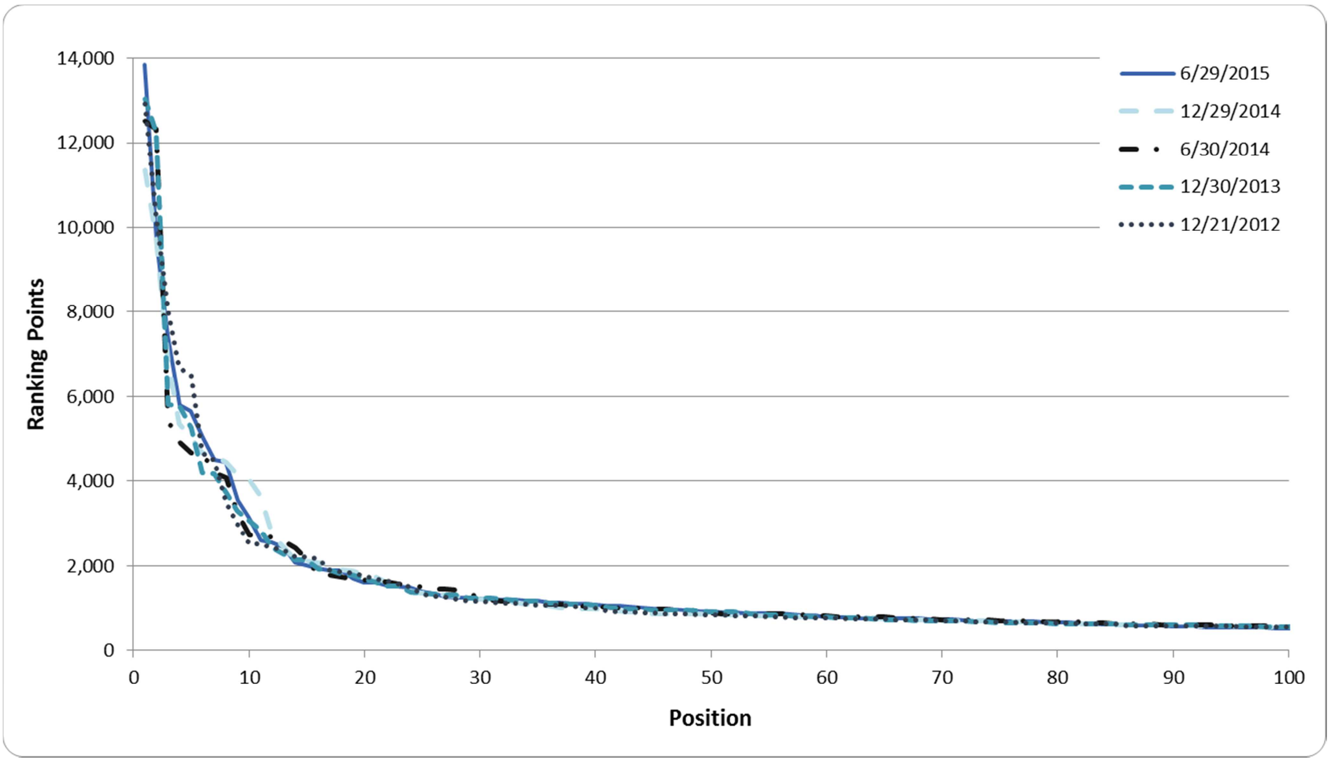

Figure 1. *Ranking points and position relationship for different dates.*

Using curves for a larger number of dates is possible to model this tendency, as made by Clarke and Dyte [10], relating ranking points and position. For the case here studied, the best fit was found employing a power equation where the parameters were adjusted by minimization of squared errors and resulted in (1):

$$\text{Points} = 18157 \cdot \mathit{Position}^{-0.779} \qquad (1)$$

Considering this information, it was proposed for this work, as a way to quantify the ranking dependence in the expected performance of the players, the simple use of their current number of points, normalized relative to the points of the leader.

2.2.2. Long and Short-Term Performance

Based on the obtained dataset, one of the possible ways of quantifying a player's performance through his career is by his victory ratios (in general numbers and on specific surfaces, as made available by ATP); however, these ratios are not always reliable, mainly due to the difference between the numbers of matches played along the career of each athlete. Illustrating with a real case, the young tennis player Jiri Vesely, at the moment of a specific data collection for this study, had played only three matches on the grass in high level tournaments and had won two of them, resulting in a good ratio of 0.667. However, in practical terms, this value should not be more significant than the fraction of 0.656

obtained with 59 victories and 31 defeats by the much more experienced Ivo Karlovic.

For such reasons, forecast models should also consider other factors for a long-term performance measure, and to do this, here are taken into consideration the overall career ratio and the number of titles. The latter appears as a relevant factor to the history of the athlete and, in this work, its application is proposed considering only absolute numbers, with no weighting factors due to their relevance. Quantification to be used as performance factor is made simply by a normalization, having as reference the largest number of titles among the players in activity, in this case, the number of tournaments won by Roger Federer.

The short-term performance here is quantified as the fraction of matches won in the last 10 played immediately prior to the tournament under analysis. This number is based on matches played in the main draws of ATP's and Grand Slam, but without considering weighting factors for victories in different levels of tournaments or against different levels of opponents. These measures, along ranking information, are expected to portray more accurately the career and the current "momentum" of each player.

2.2.3. Surface Influence

Although at the primordium of the sport all tournaments were played on grass courts, tennis now counts on three different floors classes: hard (which encompass a variety of synthetic floors), clay, and the grass itself, currently adopted in a small number of tournaments. Each of these surfaces – considering their influence on game speed, the bounce of the ball and the players' movements on the court – has peculiarities in physical demands, techniques and tactics, requiring great adaptability by the players and often resulting in significant performance differences.

The victory ratio on a specific surface is here considered due to this fact. This is an important factor to aid in the forecasts, because different surfaces require different features from the athletes. For example, on the grass, being that the fastest floor, players who are owners of a good service and greater ability to play aggressively, including net approaches and volleys to shorten the points, usually have in this surface their best performance. In contrast, on clay, the slowest surface, usually the best adapted players are those with good defensive skills and efficient movement in the baseline, what is correlated with performance in longer rallies. These characteristics are evidenced when comparing, for example, styles of play and results on both surfaces of the greatest champions in activity, Roger Federer and Rafael Nadal, being the first owner of a more offensive style and the biggest winner of the professional era on grass courts, while the second, with his efficiency near the baseline, is the greatest champion on clay courts.

The study of Clarke and Dyte [10] compares the preference of players for a certain surface to the home advantage observed in team sports such as football or basketball, given that, for tennis, disputing a tournament in a player's home country usually does not bring a significant advantage for his performance, as pointed out by Holder and Nevill [15].

Moreover, the work of Barnett and Pollard [16] analyzed the performance of players on different surfaces, showing that those with better performance on grass courts hardly have the clay as they second best surface (and vice versa). The hard courts, as the DecoTurf used in the US Open and the Plexicushion used in the Australian Open, are a "halfway" between them.

An analysis in the database used in this work leads to a similar conclusion when quantified the correlations between performances on different surfaces with the use of Pearson's coefficient, also referred to as product-moment correlation coefficient. This measure represents the strength of a linear relationship between paired data, and it is calculated by the Equation (2):

$$r = \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sqrt{\left[\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2\right]\left[\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2\right]}} \tag{2}$$

where x and y are the data vectors, containing n values each. Values of r approaching 1 indicate strong linear relationships, while null values show lack of linear relationship between the vectors [17].

For the studied group, the correlation between the vector composed by the fractions of matches won on clay by the 220 players of the dataset and the analogue vector for hard courts was calculated as 0.688, and for the grass-hard pair, 0.719. These values clearly indicate a stronger correlation than that obtained for the pair grass-clay, calculated as 0.528.

2.2.4. Grand Slam Matches

Another variable of interest is the fraction of matches won in Grand Slam tournaments, class composed by the most traditional and prestigious tournaments in the circuit: Australian Open, Roland Garros, Wimbledon and US Open. These tournaments are the only ones with the main draws composed by 128 players and, for the men, to have their matches played in best of 5 sets. Therefore, money prizes and ranking points awarded to winners are also more generous. It is observed in this case a different behavior in data, which can be related to mental and physical components, as the matches are longer and being part of the biggest events, which draw more attention of public and media.

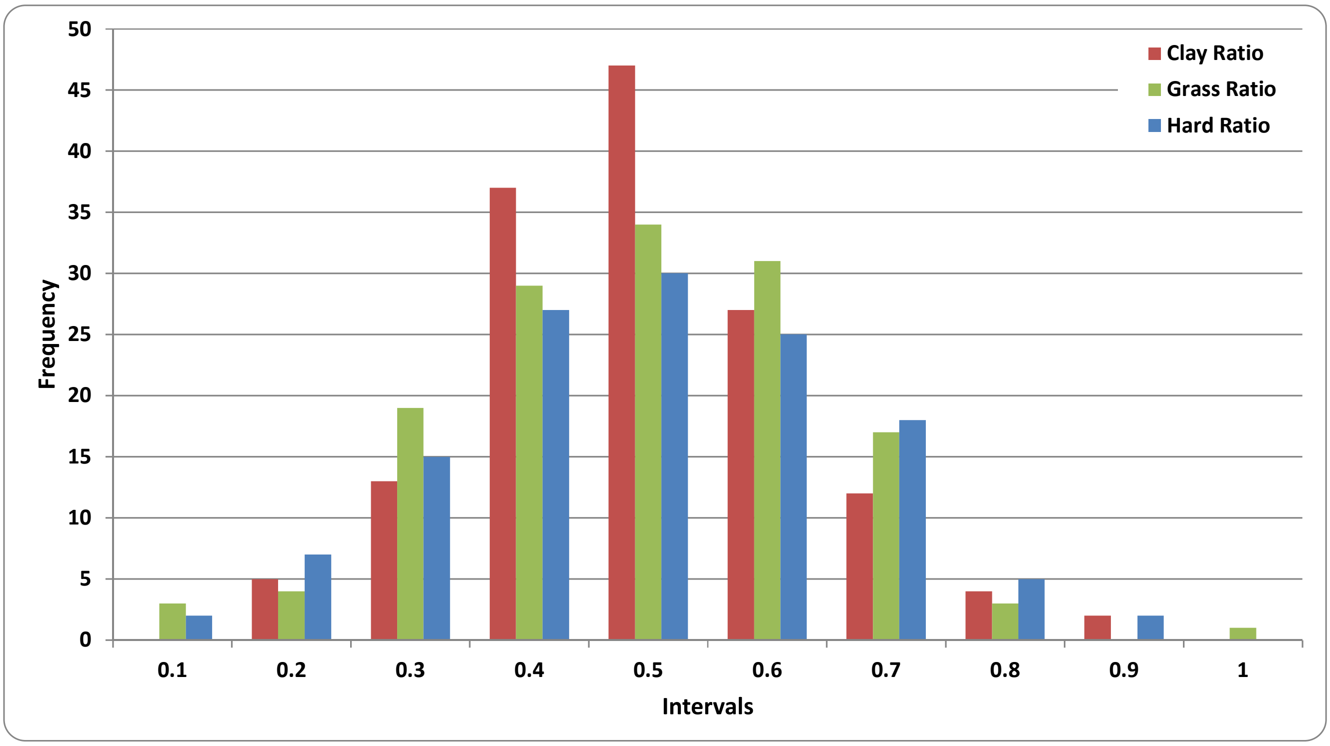

Figure 2. *Frequency distribution of victory ratios on the different surfaces for ATP tournaments.*

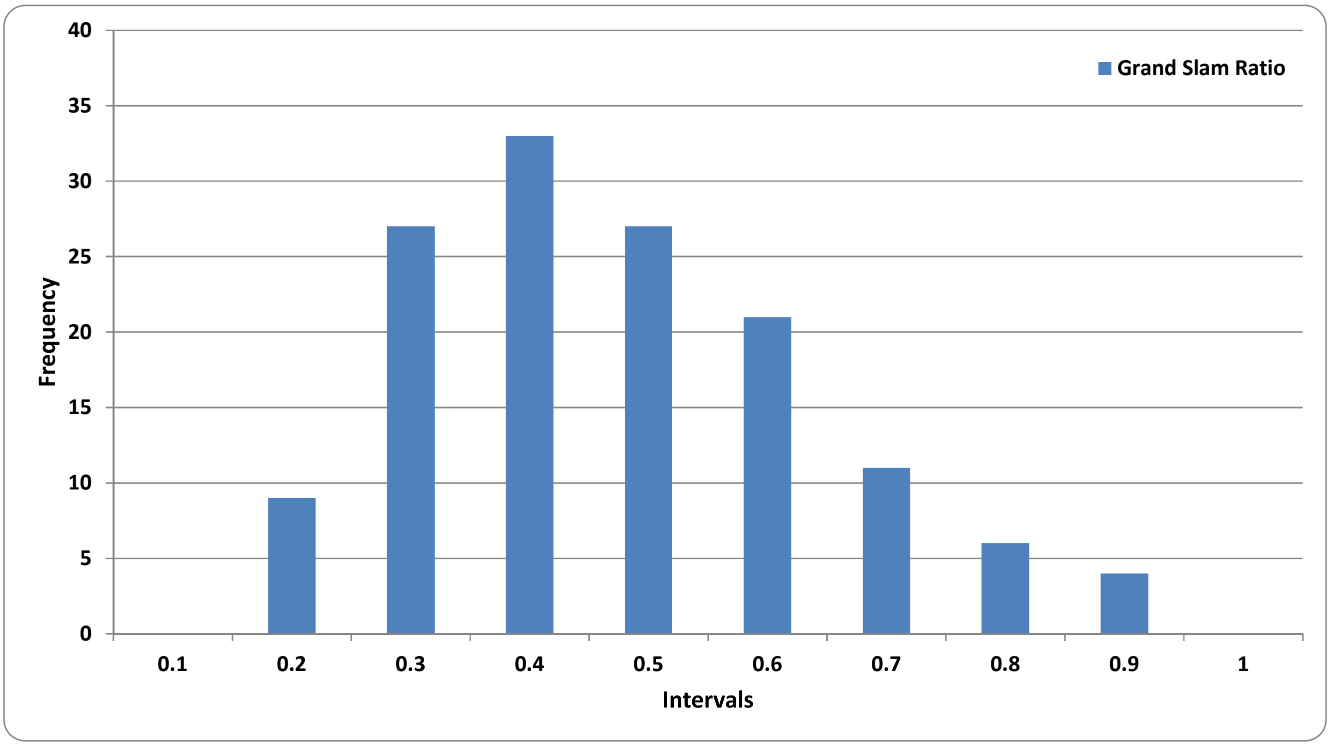

Figure 3. *Frequency distribution of victory ratios for Grand Slam tournaments.*

Evidence of this difference can be seen when comparing the graphs of Figures 2 and 3. The first shows the frequency distributions for the performances (win ratios) of the players in the group examined on the three different classes of surfaces for ATP tournaments, where can be noticed that they approach symmetric Gaussian distributions with mean 0.5. However, the second plot, with performances in Grand Slam matches, has a slightly different look, with its peak shifted toward lower values.

This shift of the peak can be explained by the tendency of the victories of players with lower rankings become scarcer in these tournaments, in other words, a smaller number of players tends to concentrate the success. This greater favoritism confirmation ratio in Grand Slam tournaments is also observed in the model proposed by Clarke and Dyte [10], that highlights that the favorite, being more likely to win sets, will be harder to be beaten in 5 sets than in 3 sets matches. This observation is consistent with the facts, given that in recent years most of the Grand Slam titles (42 in the 52 tournaments disputed between 2004 and 2016) were won by just three players: the Swiss Roger Federer, the Spaniard Rafael Nadal and the Serbian Novak Djokovic. By this feat, these players are already recognized as some of the biggest champions in the history of this sport.

2.3. Development of the Predictors

This section presents the theoretical aspects that underlie the proposed predictors, as well as the details of their design and implementation.

2.3.1. Fuzzy Predictor

Since the seminal work on this subject – the article by Zadeh [18] – fuzzy logic is being employed in a large variety of problems, being the Inference Systems some of its more prominent applications. Introduced by Mamdani, those systems are ruled by the approximated reasoning known as Generalized *Modus Ponens*, based in linguistic variables and IF-THEN implication rules to generate the typical reasoning of the fuzzy systems, by using human experience to develop intelligent algorithms capable of dealing with heterogeneous/imprecise data in a variety of applications [19-20]. The solution adopted in this work is based on a zero-order Sugeno inference system [21], where the consequent of the implication rules is a constant.

The fuzzy predictor here developed utilizes as inputs three variables, each of them being introduced in the form of a difference between values for the respective players involved in a specific match. The first is the difference between the current values in the ranking points, normalized relative to the score of the leader of the ATP entries ranking at that very moment, as cited in Section 2.2.1. The second variable is the difference between the history of the players, with their values quantified by an arithmetic mean of the wins ratio (matches at ATP and Grand Slam levels) accrued throughout their career and the coefficient of titles, calculated as described in Section 2.2.2. The third variable is the difference between the win ratios throughout the career of the players computed only in the same surface of the tournament under consideration. In this predictor, the performance in the Grand Slam tournaments and in the last 10 matches were chosen not to be included in the model, simplifying its design while making its rules more intuitive.

For each one of the matches to be predicted, these three inputs are calculated and subsequently fuzzified, being divided in four categories of values – high negative, low negative, low positive, and high positive – by using triangular membership functions defined in a generic way by (3):

$$trimf(x;a,b,c)=\begin{cases} 0, & x\le a \\ \dfrac{x-a}{b-a}, & a\le x\le b \\ \dfrac{c-x}{c-b}, & b\le x\le c \\ 0, & c\le x \end{cases} \tag{3}$$

and with characteristics shown in Figure 4, for calculating the degrees of compatibility that provide a belief in the antecedents of each rule. Triangular membership functions are chosen because of their mathematical simplicity and efficiency, resulting in a reduced computational cost. There were no improvements observed in the predictions by changing the triangular membership functions for others (as Gaussian), neither by optimizing the number of classes or their parameters for a fine tune.

To generate the set of rules that define the Inference System, the human experience is the primary source of information, and the Fuzzy logic here shows its primary purpose, allowing to express mathematically knowledge that commonly is dealt with in a linguistic form. Thus, setting up the two possible outcomes – victory of Player 1 or victory of Player 2 – a rule base is built to analyze the variables. A sample of some of these rules is shown in Table 1. The AND operators are implemented with the *minimum* function according to (4):

$$\mu_C(x)=\min\left(\mu_A(x),\mu_B(x)\right) \tag{4}$$

where μ_A and μ_B are the chosen membership functions.

Table 1. *Excerpt from the rule base of the proposed inference system.*

	ΔRanking		ΔHistory		ΔSurface		Result
IF	High Positive	AND	High Positive		-	THEN	P1 Wins
IF	Low Positive	AND	Low Negative	AND	Low Positive	THEN	P1 Wins
IF	High Negative	AND	Low Positive	AND	High Positive	THEN	P1 Wins
IF	High Positive	AND	Low Negative	AND	High Negative	THEN	P2 Wins
IF	Low Negative	AND	Low Negative	AND	Low Positive	THEN	P2 Wins
IF	High Negative	AND	High Negative		-	THEN	P2 Wins

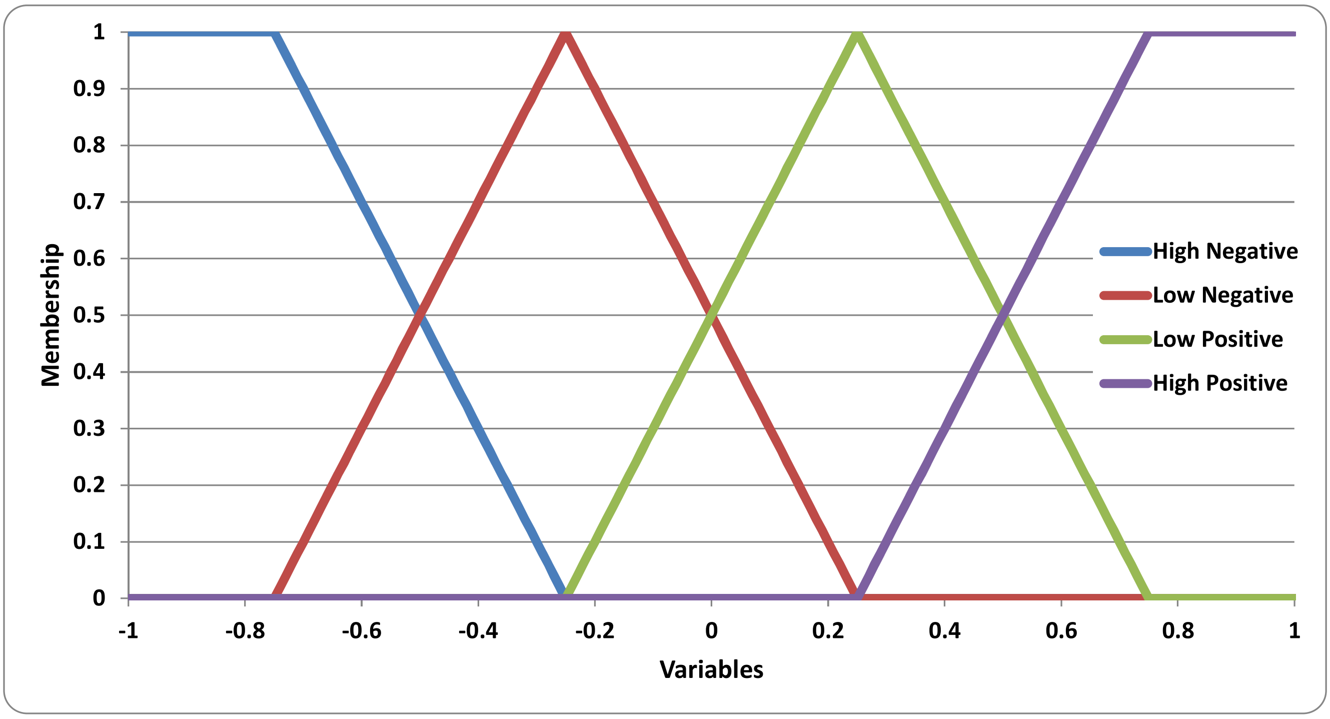

Figure 4. *Membership functions employed by the Fuzzy predictor.*

The method described so far gives the weighting factors for each one of the results. However, the desired outputs here are the beliefs in the membership of the input dataset to each one of the possible outcomes for a match, and is not composed by a single value, as usual in fuzzy inference systems. This leads to the adoption of the weighting factors themselves as the desired beliefs, after suitable normalization to obtain a percentage for each player. The victory is credited to the player with the higher value. The inference system for such task, as described, is illustrated in Figure 5.

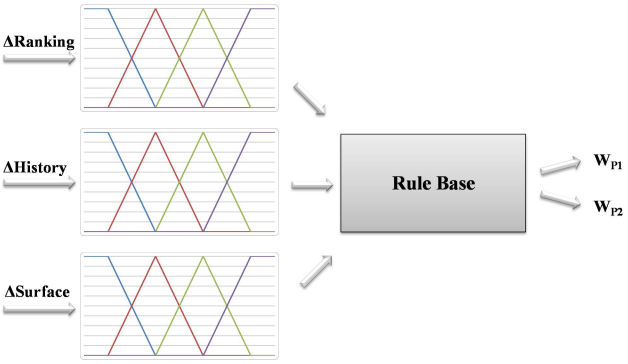

Figure 5. *Fuzzy inference system employed as predictor.*

2.3.2. Neural Network Predictor

An Artificial Neural Network, in the original paradigm inspired by biological neural networks, consists of a set of processing units (also called neurons or nodes) destined to provide an output value within a certain range, based on the weighted sum of its inputs and subsequent application of an "activation function". There are several possible arrangements for the connections between those units to form networks [22], and the most widespread is the Multi Layer Perceptron (MLP), a network with direct signal propagation where the neurons are arranged in sequential layers and the outputs of every neuron in each layer are connected to inputs of the following layer. The "knowledge" of the network is stored in the weights associated with each one of these connections, with their "learning" being made by iterative algorithms that adjust these weights based on examples (pair of inputs-outputs known a priori).

A MLP network with one intermediate layer and one output layer is able to solve some nonlinear problems and to approximate continuous functions, while the addition of one more intermediate layer enables it to implement any function, linearly separable or not, as demonstrated by Cybenko [23]. The number of nodes in each layer is the main responsible for the convergence in the training phase and for the precision of results [22].

Each one of the ANN's nodes contains an activation function, responsible for calculating the node's output from the weighted sum of its inputs. The most usual activation functions are those based on sigmoid functions (*s* shaped), due to their balance between linear and non-linear behavior, and also because they are continuous-valued monotonically increasing functions, differentiable at all points [24]. The sigmoid functions chosen for this case are the one known as logistic, given by (5):

$$f(x) = \frac{1}{1+e^{-x}} \tag{5}$$

and the hyperbolic tangent, given by (6):

$$f(x) = \frac{e^{x}-e^{-x}}{e^{x}+e^{-x}} \tag{6}$$

These functions present the property of compressing the input, with the large positive values asymptotically approaching one and large negative values being squashed to zero. Other examples of activation functions can be found in [24].

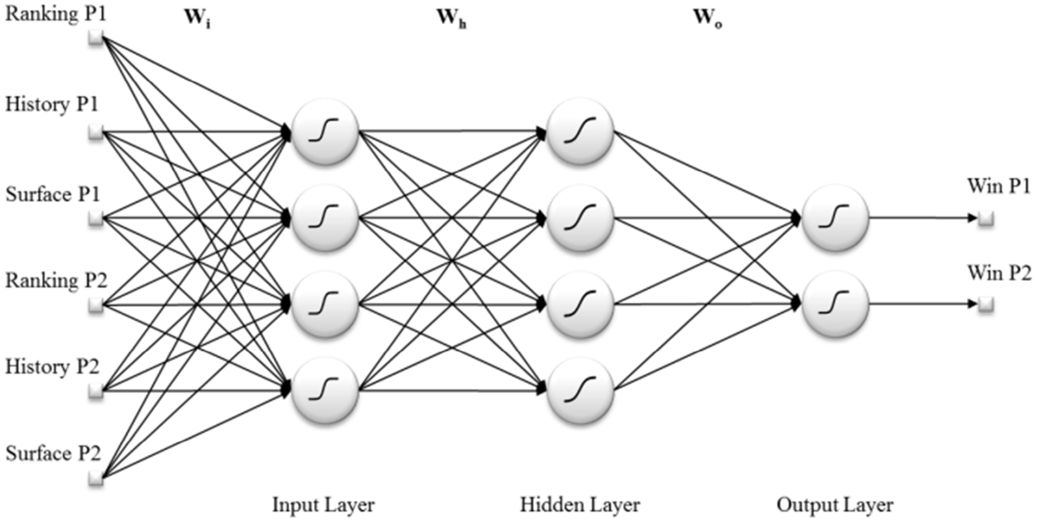

Figure 6. *Architecture of the Neural Network proposed for prediction.*

In this study, such network is then trained for, based on a dataset of matches where the winner is already known as well as data from the history and previous performances of the players, use its the ability of generalization to predict the winner in new matches when given new inputs. The architecture that resulted more appropriate to handle this problem, determined empirically, is the one depicted in Figure 6 and was implemented in *MatLab*.

This ANN is composed of an input layer with four neurons and hyperbolic tangent activation functions, a hidden layer with four neurons and logistic activation functions, and an output layer with two neurons and hyperbolic tangent activation functions. All links between nodes are weighted by a specific weight: W_i vector for the input layer, W_h for the intermediate (hidden) layer, and W_o for the output.

As input variables for the neural network, in the final proposed model, are employed the same three performance measures on which the Fuzzy predictor is based: coefficients for the current score in the ranking, for the player's history (composed by number of titles and career win ratio) and ratio of victories on the same surface of the tournament under analysis. However, in this model are employed individual values for the two players competing against each other in a specific match, resulting in a total of six inputs. The output variables are two, each representing the victory of one of the players in binary values. So, for a victory of the first player is expected that its corresponding output will have unit value and the other output a null value, for example. An excerpt from the training dataset is illustrated in Table 2. Neural models considering also the use of Grand Slam performance data (totalizing eight inputs) were also evaluated. All the variables are normalized for the interval [0, 1].

Table 2. *Excerpt from the training dataset for the Neural Network.*

Tournament	Match	Score	Rk. 1	Hist. 1	Surf. 1	Rk. 2	Hist. 2	Surf. 2	P1	P2
Wimbledon'14	Murray-Dimitrov	1-6 6-7 2-6	0.374	0.757	0.830	0.208	0.470	0.622	0	1
US Open'14	Cilic-Federer	6-3 6-4 6-4	0.144	0.11	0.670	0.587	0.887	0.829	1	0
Paris'14	Djokovic-Raonic	6-2 6-3	0.817	0.826	0.828	0.279	0.546	0.710	1	0
Xangai'14	Nadal-Lopez	3-6 6-7	0.600	0.870	0.776	0.124	0.435	0.509	0	1
Toronto'14	Ferrer-Dodig	1-6 6-3 6-3	0.290	0.682	0.637	0.057	0.312	0.483	1	0

The learning of the ANN in this study was based on the most popular of the training algorithms: the backpropagation. In this method, the weights of the connections between the network's nodes are initialized with random values. After that, sets of input values that result in an output already known are presented to the network (in random order). For each one of these sets, the network output with the current weights is computed, in the so-called "forward phase" of training. The output obtained is then compared to the correct output pattern to allow the calculation of the error between both. This error is propagated through the network in a reverse path ("backward phase", justifying the name of the algorithm). The product of the error of each output by a constant "learning rate" is subtracted from the connections' weights of the respective node in the last layer. The error of each node of the previous layers is calculated using the errors of the nodes from the following layer connected to it, weighted by the weights of the connections between them

[22]. The procedure is repeated, with new pairs of input/output vectors being presented to the network until a stopping criterion is reached: the mean square error becomes smaller than a predetermined limit, a maximum number of iterations is reached, or the error becomes stagnated between iterations. A success in the training phase will result in a network ready for the forecasts.

2.3.3. Strength Equation

Based on the previously presented analysis of the factors that influence the matches' outcomes, an intuitive way of measure them comparatively is by an equation where for each of the studied attributes will be assigned a weighting factor. This equation, here denominated "Strength Equation" also has the objective to quantify the strength of each player for a specific tournament, that is, his ability to succeed based on his current form, his history and his performance on that specific surface. Therefore, the same equation may be used for predictions of matches' outcomes from a belief calculation based on the comparison between the strengths of any two players.

The proposed equation has the following form (7):

$$S_n = w_1 \cdot titles + w_2 \cdot \text{ranking} + w_3 \cdot \text{last 10} + w_4 \cdot career + w_5 \cdot \text{grand slam} + w_6 \cdot \text{surface} \tag{7}$$

where each of the attributes is obtained as described in Section 2.2 and the vector w_i is composed by their respective weighting factors. Figure 7 depicts schematically the process of forecasting the outcome of a match using this equation.

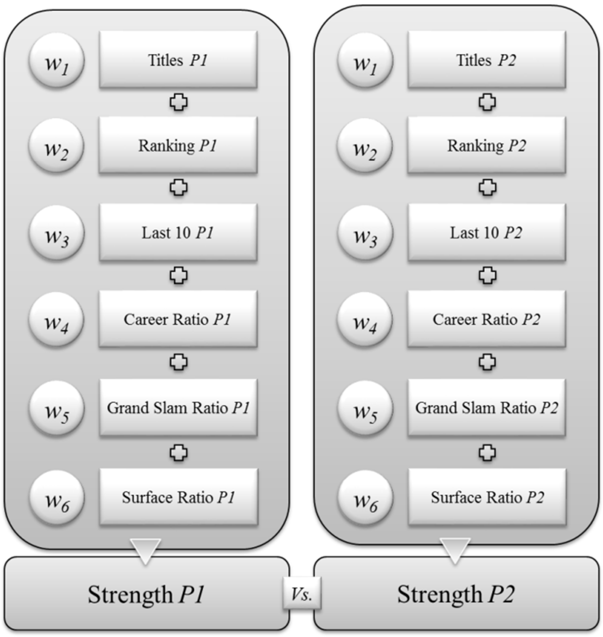

Figure 7. *Representation of the forecast procedure using the Strength Equation.*

The need for adjustment of the weights makes this model dependent of a dataset with matches' results and player's information for "training", as well as in the neural model. With this set, the adjustment can be performed by a combinatorial optimization process, which here is done by means of the evolutionary algorithm available in the *Microsoft Excel's Solver*. This tool makes available a Genetic Algorithm where the user defines the inputs variables, the output, restrictions and the control parameters. The objective is to maximize the number of correct predictions of winners in that set of matches, in other words, to generate a combination of weights such that the winners' strength is greater than the losers' strength in as many cases as possible. For simplicity, the values of the weights were restricted to natural numbers in the [0, 5] interval, with no noticeable performance loss. The control parameters were set as following: Convergence = 0.0001, Mutation Rate = 0.05, Population Size = 200, and Random Seed = 0. Here, two different equations were optimized with data from tournaments played in 2014: one for matches in best of three sets (ATP level) and other for matches in best of five sets (Grand Slam level), with the results shown in Table 3.

From the abovementioned values, the greatest influence of the surface and ranking for the matches of three sets can be seen, where the Grand Slam ratio was restricted to zero. Curiously, optimization led the Last 10 also to zero, even being one of the factors of greater weight in matches of five sets, along with the Ranking and Grand Slam ratio. In the latter case, the influences of career ratio and surface ratio were devalued.

Table 3. *Strength Equation's weights adjustment.*

Attribute	w_i (ATP's)	w_i (Grand Slam)
Titles	1	2
Ranking	4	4
Last 10	0	4
Career Ratio	4	0
Grand Slam Ratio	0	5
Surface Ratio	5	1

2.3.4. Voting System

Having been developed the three predictors based on soft computing, the following step was to develop a system that encompasses their strengths, combining the three outcomes in only one. A simple yet efficient way of achieving that is by a voting system. Having an odd number of predictors, it was chosen for aggregation the simple Majority Vote, what means that the predicted winner of a match will be the one considered the favorite by at least two out of the three independent classifiers. Figure 8 depicts this process.

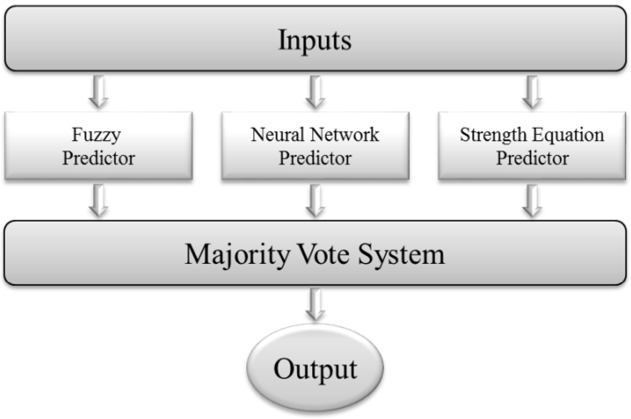

Figure 8. *Representation of the Voting System's framework.*

3. Results

The analysis of the proposed predictors' performance was based on a database containing all matches' results from the last few seasons, available at [14]. That database, with the information of 1744 matches played in tournaments of categories ATP 250, ATP 500, Masters 1000 and ATP Finals, added to 508 Grand Slam matches (all of them played in 2014 and involving over 200 different players), constitutes the training dataset for the Neural Network, and is employed to fine-tune the Fuzzy predictor and to optimize the weights in the Strength Equation.

Validation of the predictors and the statistics of their responses were based on a set with data from 1109 matches played in ATP tournaments added to the 381 matches played in the tournaments Australian Open, Roland Garros and Wimbledon during 2015 season. Again, this group of matches includes more than 200 different players. These are the same datasets employed by the preliminary work presented in [25].

Table 4. *Correct Predictions by Ranking and Bookmaker's Odds – Previous Years.*

Year	ATP's		Grand Slam	
	% Correct Ranking	% Correct Bets	% Correct Ranking	% Correct Bets
2010	64.72%	67.12%	74.95%	78.59%
2011	66.14%	69.40%	75.00%	78.22%
2012	66.07%	68.92%	74.85%	77.86%
2013	64.00%	66.98%	75.31%	77.97%
2014	66.85%	67.80%	74.07%	75.98%

From the group of matches played in 2014, the survey shows that the percentage of matches won by the best ranked player was close to 67% in ATP tournaments and 74% in the Grand Slam, what is quite consistent with the average observed over previous years, as shown in Table 4. This comparison is important to check that the analyses are not based on atypical events data. The same table also shows, for comparative purposes, the percentage of correct predictions based on bookmakers' odds took from five of the major websites for bets in sports, according to numbers also compiled in [14]. These values represent the fraction of matches that were won by players who were considered the most quoted, what is a good benchmark for a predictor, given that these numbers depict the confidence of bookmakers, who are expected to have some knowledge about players' (past performance, current form and eventual particularities) and tournaments' characteristics.

Unpredictable results, that could be considered outliers, were not removed from the training dataset. This decision is intended to let the predictors try to draw a pattern for these results that would normally be considered as unforeseeable.

As previously mentioned, this work focuses on predictions of winners in tennis matches without considering events during their course or even their final score. Thus, for the three proposed predictors the information required for analysis is a set of inputs for each match, as detailed in Section 2.3, and the winner of this match to compare with the predictors' outputs.

The results of the predictors for the matches in ATP level used as the validation dataset is shown in Table 5, segmented by surfaces. It is noteworthy that, although the results with this segmentation are presented, there were no specific models per surface; the Fuzzy Inference System is the same used in all predictions, the Neural Network is trained with all the matches of ATP tournaments without distinction, and the Strength Equation model has also its coefficients optimized for that class of tournaments. Hit rates are compared to those obtained from mere comparison of rankings at the time of the tournament and also from the bookmakers' odds, as aforesaid.

Table 5. *Performance of Matches' Prediction – ATP's.*

Surface	Number of Matches	% Correct Ranking	% Correct Bets	% Correct Fuzzy	% Correct Neural Net.	% Correct Equation	% Correct Voting Sys.
Hard	474	62.87%	69.62%	65.82%	71.66%	66.88%	67.30%
Clay	483	68.74%	70.73%	70.60%	77.02%	72.26%	72.46%
Grass	152	55.92%	66.77%	62.50%	72.15%	69.08%	69.08%
All	1109	64.47%	69.75%	67.45%	74.06%	69.52%	69.79%

Table 6. *Performance of Matches' Prediction – Grand Slam.*

Grand Slam	Number of Matches	% Correct Ranking	% Correct Bets	% Correct Fuzzy	% Correct Neural Net.	% Correct Equation	% Correct Voting Sys.
Australian Open	127	74.80%	78.74%	77.95%	80.71%	85.83%	81.10%
Roland Garros	127	71.65%	78.15%	76.38%	77.56%	81.89%	78.74%
Wimbledon	127	75.59%	75.59%	74.02%	71.65%	72.44%	74.80%
All	381	74.02%	77.49%	76.12%	76.64%	80.05%	78.22%

From these results, it can be seen that the percentage of correct answers obtained by the Fuzzy predictor represents a gain over the prediction by ranking comparison, but still has a performance inferior to that of the bookmakers, that was nearly equaled by the model using the Strength Equation. On the other side, the Neural Network achieved the greatest accuracy with a very significant margin, which means that it was able to extract relevant features from the training dataset and to quantify them efficiently in the model. The voting system based on the outcomes of the other predictors presented good results, but was inferior to the Neural Network.

The results from the forecasts made for Grand Slam matches are shown in Table 6, presenting the same comparisons. In this case, the Fuzzy predictor had the same modeling and the same rule base previously used for the ATP tournaments, while the Neural Network, although having identical configuration, was trained only having as reference the set consisting of the 508 Grand Slam matches played in the four tournaments disputed in 2014. The same applies to the Strength Equation, which had its coefficients optimized having as reference this same dataset. For the neural model, tests were also conducted including a new input: the Grand Slam victory ratio. However, the addition of this variable to the model did not result in improvements in the quality of the predictions, and because of this the results presented are from a network with the original configuration, previously presented in Figure 6.

For this class of tournaments, as the percentage of matches where the best-ranked won is significantly higher, it is expected the margins of improvement with the use of the predictors to be smaller, and observing the numbers in Table 6 it is what can be noted for most cases. Here, the first two proposed predictors improved the figures obtained from the forecast by ranking, but both were slightly below the performance of bookmakers. The Strength Equation, in turn, was able to obtain correct predictions percentage notably above the others, especially for the first two Grand Slam analyzed, showing that the weights' adjustment was efficient enough to result in a good model for this problem, after adding input variables representing the previous performance in Slams and in the last 10 matches. Once again, the voting system performed better than two of the predictors, but couldn't beat the best.

The difficulty in improving the rates obtained by bettors make clear the limitations of automatic forecasts, since these can never cover all the quantitative and qualitative aspects that a human predictor (as the ones who bet) could take into account. Some examples of these aspects are influence of the crowd, fatigue generated in previous rounds, momentary changes in physical and emotional conditions, extra motivations or pressure etc.

Another metrics for the quality of predictors is the DeFinetti Measure, capable of quantifying the accuracy of predictions when confronted with the results that actually occurred [26]. The importance of this quantification lies on the fact that often the number of correct or incorrect outcomes from a predictor can misled its quality evaluation, while not worrying about the previously estimated error margins. An example of this problem is illustrated by observing the different forecasts for the match between Rafael Nadal and Dustin Brown in Wimbledon 2015, surprisingly won by the German, who at that moment occupied the modest 102nd position in the ATP's entries ranking. The Fuzzy predictor calculated his chances of victory with a probability inferior to 0.001, while the Strength Equation indicated 0.186. While both have missed the winner (in this case even bookmakers gave Brown a low credibility of 0.141), it is clear that the error of the Fuzzy predictor was more serious.

To obtain this measurement for a series of predictions, the DeFinetti distance must be initially calculated for every match by the equation (8):

$$DF = \begin{cases} (p_{w1}-1)^2 + (p_{w2}-0)^2 & \text{if player 1 wins the match} \\ (p_{w1}-0)^2 + (p_{w2}-1)^2 & \text{if player 2 wins the match} \end{cases} \quad (8)$$

where p_{w1} and p_{w2} are the probabilities of victory previously assigned to the players. That distance corresponds geometrically to the quadratic Euclidean distance between the predicted values and the ones that really occurred, when win and loss probabilities are considered elements in a vector. The DeFinetti Measure of the series of predictions can then be obtained by calculating the arithmetic mean of the DeFinetti distances calculated for each match, where a predictor is as best as lower is this average [26].

The values obtained by this means for each of the proposed predictors are shown in Table 7, distinguished by the classes of tournaments – ATP's with matches in best of three sets and Grand Slam with matches in best of five sets. By way of comparison, the table also contains values calculated for a simple predictor by ranking where the probabilities of winning of each player were obtained by weighting their ranking points at that moment. The measure for the Voting System was not computed, due to the absence of a numerical outcome.

Table 7. *DeFinetti's Measure for the Proposed Predictors.*

Level	Number of Matches	Ranking Predictor	Fuzzy Predictor	Neural Network Predictor	Equation Predictor
ATP	1109	0.428	0.410	0.369	0.424
Grand Slam	381	0.351	0.337	0.342	0.335

From the aforementioned results, the first observation to be made is that in all cases the measures are inferior to 0.50, which means that all predictors have performance superior to a "predictor" that assigns 50% chance of winning for each of the tennis players in every match. Moreover, it can be perceived that the three methods showed better results than the inference by ranking, with a positive highlight for the value obtained for the Neural Network with the ATP's and the best performance of the Strength Equation with Grand Slams, which implies in consistency with the results of the percentages of correct outcomes.

4. Conclusion

This paper presented a study on the predictability of winners in tennis matches, starting from analysis of players' performance, taking into account their career, their current momentum, and their aptitude on different surfaces. The problem of predicting the matches' outcomes was approached by three different methods: the first an Artificial Neural Network, the second a Fuzzy Inference System and the third a Strength Equation with weighting factors adjusted by optimization. They all rely on classical techniques of Soft Computing, considered relevant for their efficiency and versatility in applications, and the obtained performances (both individually and combined by a Voting System) ratify that. The predictors presented good results, always surpassing the hits rates obtained by simply comparing players' rankings and in some cases even outperforming the – in most cases experts – bookmakers. These predictors can also be used to obtain beliefs in what players will have more chance to succeed prior a given tournament, helping coaches to select teams for competitions like the Davis Cup or the Olympics and even helping the own players to compose their calendar with the tournaments where they could perform better.

The study exposed here, however, is part of a model subject to many imperfections, since it is impossible to quantify dozens of factors that can influence the outcome of matches, as the momentary emotional state, injuries, support from fans, fitness, possible lack of tempo or shape after an absence from the circuit, adaptations to changes in equipment etc. However, it can be noted that there are margins for improvement in predictions, especially by looking at the Neural Network's results for the matches played in three sets, or the Strength Equation's results for the Grand Slam matches, situations where large gains were achieved.

Future work will focus on improvement by using information from new variables, such as head-to-head numbers and the prize money obtained within some specified period of time preceding the tournament under analysis. That is a way of giving more value to the most important victories, as the major tournaments offer more generous prizes and higher monetary values are awarded on victories in the later stages of tournaments. That information, though more difficult to obtain, may allow a better performance in the predictions.

References

[1] ATP. *Official site of men's professional tennis*. 2015. Available online at: <http://www.atpworldtour.com>. Last accessed: November 1, 2015.

[2] ITF. *International tennis federation*. 2015. Available online at: <http://www.itftennis.com>. Last accessed: November 1, 2015.

[3] FORBES. *The world's highest-paid athletes*. 2015. Available online at: <http://www.forbes.com/athletes/list/>. Last accessed: November 1, 2015.

[4] GONZÁLEZ-DÍAZ, J.; GOSSNERB, O.; ROGERS, B. W. Performing best when it matters most: Evidence from professional tennis. *Journal of Economic Behavior & Organization*, n. 84, p. 767– 781, 2012. ISSN 0167-2681.

[5] FERRAUTI, A. *et al*. Diagnostic of footwork characteristics and running speed demands in tennis on different ground surfaces. *Sport Orthopädie Traumatologie*, n. 29, p. 172–179, 2013. Available online at: <http://dx.doi.org/10.1016/j.orthtr.2013.07.017>. Last accessed: November 1, 2015.

[6] KLAASEN, F.; MAGNUS, J. R. Forecasting the winner of a tennis match. *European Journal of Operational Research*, n. 148, p. 257–267, 2003. ISSN 0377-2217.

[7] MCHALE, I.; MORTON, A. A Bradley-Terry type model for forecasting tennis match results. *International Journal of Forecasting*, n. 27, p. 619–630, 2011. ISSN 0169-2070.

[8] CLOWES, S.; COHEN, G.; TOMLJANOVIC, L. Dynamic evaluation of conditional probabilities of winning a tennis match. In: AUSTRALIAN CONFERENCE ON MATHEMATICS AND COMPUTERS IN SPORT, 6. *Proceedings*… Gold Coast, Australia: 6M&CS, 2002. Available online at: <http://hdl.handle.net/10453/6673>. Last accessed: November 1, 2015.

[9] KNOTTENBELT, W. J.; SPANIAS, D.; MADURSKA, A. M. A common-opponent stochastic model for predicting the outcome of professional tennis matches. *Computers and Mathematics with Applications*, n. 64, p. 3820–3827, 2012. ISSN 0898-1221. Available online at: <http://dx.doi.org/10.1016/j.camwa.2012.03.005>. Last accessed: November 1, 2015.

[10] CLARKE, S. R.; DYTE, D. Using official ratings to simulate major tennis tournaments. *International Transactions in Operational Research*, n. 7, p. 585–594, 2000. ISSN 1475-3995.

[11] KLAASSEN, F.; MAGNUS, J. Are points in tennis independent and identically distributed? Evidence from a dynamic binary panel data model. *Journal of the American Statistical Association*, n. 96, p. 500–509, 2001.

[12] DEL CORRAL, J.; PRIETO-RODRIGUEZ, J. Are differences in ranks good predictors for Grand Slam tennis matches? *International Journal of Forecasting*, n. 26, p. 551–563, 2010. ISSN 0169-2070.

[13] SCHEIBEHENNE, B.; BRODER, A. Predicting Wimbledon 2005 tennis results by mere player name recognition. *International Journal of Forecasting*, n. 23, p. 415–426, 2007. ISSN 0169-2070.

[14] TENNIS DATA. *Tennis results and betting odds data*. 2015. Available online at: <http://www.tennis-data.co.uk/alldata.php>. Last accessed: November 1, 2015.

[15] HOLDER, R. L.; NEVILL, A. M. Modelling performance at international tennis and golf tournaments: is there a home advantage? *The Statistician*, n. 46, p. 551–559, 1997.

[16] BARNETT, T.; POLLARD, G. How the tennis court surface affects player performance and injuries. *Medicine and Science in Tennis*, n. 12, v. 1, p. 34-37, 2007. ISSN 1567-2352.

[17] WEISSTEIN, E. W. *Correlation Coefficient*. 2015. Available online at: <http://mathworld.wolfram.com/CorrelationCoefficient.html>. Last accessed: November 1, 2015.

[18] ZADEH, L. Fuzzy Sets. *Information and Control*, n. 8: p. 338-353, 1965. Available online at: <http://dx.doi.org/10.1016/S0019-9958(65)90241-X>. Last accessed: November 1, 2015.

[19] JANG, J.-S.; SUN, C.-T.; MIZUTANI, E. *Neuro-Fuzzy and Soft Computing: A Computational Approach to Learning and Machine Intelligence*. Upper Saddle River, NJ, USA: Prentice-Hall, 1997.

[20] FERNANDES, M. A. *Classificação de alvos utilizando atributos cinemáticos*. Master's Degree Dissertation, ITA, São José dos Campos, Brazil, 2009.

[21] SUGENO, M. *et al*. (Ed.). *Industrial Applications of Fuzzy Control*. New York, NY, USA: Elsevier Science Pub. Co., 1985.

[22] BRAGA, A. P.; CARVALHO, A.; LUDERMIR, T. *Redes Neurais Artificiais – Teoria e Aplicações*. Rio de Janeiro, RJ, Brazil: LTC, 2000.

[23] CYBENKO, G. Approximation by superpositions of a sigmoidal function. *Mathematics of Controls, Signals, and Systems*, Springer Verlag, n. 2, p. 303-314, 1989.

[24] HAYKIN, S. *Neural Networks – A Comprehensive Foundation*. Upper Saddle River, NJ, USA: Prentice Hall, 1998.

[25] FERNANDES, M. A. Inteligência computacional aplicada à previsão de vencedores em partidas de tênis. *Revista Brasileira de Computação Aplicada*, v. 8, n. 2, p. 82–98, 2016. ISSN 2176-6649.

[26] ARRUDA, M. L. *Poisson, Bayes, Futebol e DeFinetti*. Master's Degree Dissertation, USP, São Paulo, Brazil, 2000.

[27] LIMA, B. N. B. *et al*. Probabilidades no esporte. *TRIM: revista de investigación multidisciplinar*, Universidad de Valladolid, n. 5, p. 39-53, 2012. Available online at: <http://uvadoc.uva.es/handle/10324/11665>. Last accessed: November 1, 2015.