PREDICTION OF THE SUCCESSFULNESS OF TENNIS PLAYERS WITH MACHINE LEARNING METHODS

Andrej Panjan[1], Nejc Šarabon[1] and Aleš Filipčič[2]

[1]Laboratory for Advanced Measurement Technologies, Wise Technologies, Ljubljana, Slovenia
[2]Faculty of Sport, University of Ljubljana, Ljubljana, Slovenia

Abstract:

The aim of the study was to examine the predictability of the competitive performance of Slovene tennis players by using the most promising morphological measures and motor tests selected by automatic computer methods and by experienced tennis coaches by means of machine learning methods. The analysis included altogether 1,002 male and female tennis players who had undergone regular testing by the National Tennis Association and were positioned on the ranking list of the Slovene Tennis Association between the years 1993 and 2008. Selections of the most promising variables by means of the two automatic methods yielded similar results, whereas the selection performed on the basis of estimates made by coaches differed considerably. With regard to the analysis by means of classification methods, an accurate predictability of competitive performance for the age category younger than 16 years was observed, while the results of predictions for the age category older than 16 were poor. Among the regression methods, as opposed to the linear regression, which has yielded satisfactory results, regression trees served no useful purpose in practice. Automatic methods for identifying the most promising variables proved to be more successful than those of the coaches, which was most clearly noticeable with regard to the female tennis players and when linear regression was used.

Key words: tennis, identification, selection, predictability, competitive performance, machine learning

Introduction

The development of a tennis player is a long--term and complex process, which takes place in the fields of tactics, playing technique, psychology and physical fitness. In addition to genetic factors, there are numerous external factors which influence the development of a player (the content, quantity and intensity of the workload, training methods, supervision of training and regular progress evaluation, the number and level of tournaments, etc.). The capability of a player can be defined in terms of his/her competitive results and indicated by estimates of future potential (Filipčič, A. & Filipčič, T., 2005a, 2005b). The former is defined by comparing the player's results with other presently active players within his/her age category in the current tournament season (the number of victories, tournaments won, ranking, etc.), whereas the latter is determined by means of measurement procedures in the aforementioned bio-psycho-social parameters of the competitor.

In competitive tennis we give priority to absolute competitive performance. Evaluations and predictions of competitive performance of young tennis players at the senior level can be conducted in several ways. A frequently used method in sports is for a specifically qualified expert to watch some performances of a certain athlete and then, based on his/her own experience and expertise, to give his/her opinion on the athlete's future success. A second method extends the comparison to other athletes within the same age category. The third method is to analyse and predict competitive performance on the basis of the measurements of the potential skills of an athlete.

In studying the factors pertaining to competitive performance, our aim is to avoid making subjective estimates as much as possible, by using artificial intelligence methods. Our assumption is that it is possible to improve the objectivity and reliability of predicting the competitive performance of a tennis player by means of machine learning methods. Machine learning (Hand, Mannila, & Smyth, 2001; Kononenko & Kukar, 2007; Witten & Frank, 2005) is an artificial intelligence field which deals with discovering knowledge in data, data analysis, automatic generation of knowledge databases for expert systems, construction of numeric and quali-

tative models, classification, regression, etc. By increasing the number of data in digital form in a very intensive manner, which we have witnessed in recent years, machine learning is becoming an important tool for transforming these data into useful information, since manual processing of such a vast quantity of data has become impossible. The basic principle of machine learning is the automatic modelling of data. Learned models attempt to interpret the data from which the models were constructed. They can assist in making decisions when it comes to studying the modelled process in the future (predictions, diagnosis, control, verification, simulations, etc.).

Studies conducted so far (Banzer, Thiel, Rosen-hagen, & Vogt, 2008; Filipčič, A. & Filipčič, T., 2005a, 2005b; Girard & Millet, 2009; Sánchez--Muñoz, Sanz, & Zabala, 2007) have pointed to a correlation between competitive and potential performance, and we were therefore interested in finding out how it is possible to elucidate competitive performance of tennis players in all male and female age groups in Slovenia by means of artificial intelligence methods. The primary aim of the study was to compare the predictions of competitive performance on the basis of the results of morphological measures and motor tests, selected by means of automatic machine learning methods on the one hand, and by means of subjective estimates made by tennis coaches on the other hand. The secondary aim was to compare the most promising morphological measures and motor tests, selected by machine learning methods and tennis coaches. Additionally, we were interested in the differences in prediction between age categories and between genders. We hypothesized that the prediction power of the machine learning methods would dominate those of subjective estimations. Furthermore, we anticipated the morphological dimensions would play a decisive role in junior categories, while strength and power would be expected to take over the key role when approaching senior categories.

Methods

Subjects

The sample of subjects included those Slovene tennis players who were positioned on the ranking list of the Slovene Tennis Association in individual periods and who also had undergone morphological and motor measurements in these individual periods between 1993 and 2008. Measurement data was collected for 593 male tennis players and 409 female tennis players, i.e., 1,002 individual tennis players in total. The data collection procedures met international ethical standards and were consistent with the Helsinki Declaration. The selected subjects were then divided into three age groups. The entire sample of measures and tests was then separated into three age categories for the analysis of the predictability of competitive performance:

- Age category U12/U12: for subjects in the age group under 12 years on the basis of measurements performed in the period under 12 years of age, consisting of 170 male tennis players (age 12.14±1.02 years, body height 153.53±7.95 cm, body weight 42.85±7.58 kg) and 157 female tennis players (11.85±.75 years, 155.74±8.16 cm, 44.02±8.33 kg).
- Age category 12-16/12-16: for subjects in the age group between 12 and 16 years on the basis of measurements performed in the period between 12 and 16 years of age, consisting of 341 male tennis players (14.88±1.20 years, 170.35±10.08 cm, 58.43±11.49 kg) and 215 female tennis players (14.80±1.19 years, 166.65±6.18 cm, 55.93±7.46 kg).
- Age category A16/A16: for subjects in the age group above 16 years on the basis of measurements performed in the period above 16 years of age. This sample consisted of 82 male tennis players (18.87±2.53 years, 182.73±5.72 cm, 73.37±6.69 kg) and 37 female tennis players (18.07±1.78 years, 169.88±6.22 cm, 62.59±7.95 kg).
- Age category 12-16/U12: for subjects between 12 and 16 years, but on the basis of measurements performed in the period under 12 years of age consisting of 89 male tennis players (12.03±1.02 years, 156.34±7.82 cm, 44.82±7.76 kg) and 84 female tennis players (11.96±.71 years, 157.68±7.03 cm, 45.42±8.12 kg).
- Age category A16/U12: for subjects in the age group above 16 years, but on the basis of measurements performed in the period under 12 years of age. This sample consisted of 47 male tennis players (12.09±1.25 years, 157.02±8.26 cm, 45.20±8.51 kg) and 35 female tennis players (11.92±1.18 years, 159.22±6.94 cm, 46.73±7.28 kg).
- Age category A16/12-16 for subjects in the age group above 16 years, but on the basis of measurements performed in the period between 12 and 16 years of age. This sample consisted of 125 male tennis players (14.80±1.27 years, 175.00±9.19 cm, 63.66±11.11 kg) and 79 female tennis players (14.94±1.17 years, 168.29±6.03 cm, 58.88±7.54 kg).

Data collection

Measurements were made on a selection of independent variables, whose usefulness in predicting competitive performance in tennis had already been identified. These measurements test both general and tennis-specific motor abilities of players, as well as morphological variables, and were conducted annually in the laboratories of the Faculty of Sport in Ljubljana between 1993 and

2008. The tests of general and tennis-specific motor abilities examined all key areas of the player's motor abilities and cardiorespiratory functional capacity (strength, speed, agility, flexibility, balance, coordination, endurance). Table 1 presents the composition of this test battery.

The position on the ranking list of the Slovene Tennis Association for an individual year was used as the primary criterion for estimating competitive performance. This ranking list takes into account the results achieved in that competition year. The position on the Tennis Association ranking list is determined on the basis of a coefficient which represents the total number of points won by the individual player divided by the number of tournaments played.

Additionally, a survey was conducted among Slovene professional tennis coaches who were asked to study a previously prepared list of all our morphological measures and motor tests and then to select those variables that in their opinion most influence the competitive performance of individual age categories. Five Slovene coaches aged between 24 and 38 years who had had between two and 18 years of coaching experience took part. They were all familiar with the tennis players whom they evaluated.

Data processing

Prediction of competitive successfulness was estimated with classification and regression methods. Both methods used morphological measures and motor tests as predictor variables and position on the ranking list as a criterion variable (class).

Experiment procedure included the following phases:

Table 1. Applied morphological measures and motor tests. (For detailed description see (Filipčič, 1996; Filipčič, A. & Filipčič, T., 2005b)

Abbreviation	Measure and Test	Ability/Dimension
M-1	Body height	Morphology
M-2	Body weight	Morphology
M-3	Body mass index	Morphology
M-4	Fat tissue percentage	Morphology
M-5	Muscle tissue percentage	Morphology
M-6	Bone tissue percentage	Morphology
P-1	Sargent test	Explosive power - lower ext.
P-2	Medicine ball throw (2 kg)	Explosive power - upper ext.
P-3	Four-jumps test	Explosive power - lower ext.
P-4	Sit-ups	Muscular endurance - trunk
S-1	20-metre sprint	Sprint acceleration
A-1	9 x 6-metre sprint test	Agility
S-2	Reaction pole	Reaction time
S-3	Foot tapping	Alternative movements' frequency - lower ext.
S-4	Hand tapping	Alternative movements' frequency - upper ext.
F-1	Forward bend	Passive flexibility - lower ext.
F-2	Sprain with a stick	Passive flexibility - upper ext.
F-3	Launge	Active flexibility - lower ext.
A-2	Fan	Agility
A-3	Hexagon test	Agility
C-1	Stamping test	Coordination - lower ext.
C-2	Obstacle course backwards	Coordination - whole body
C-3	Racquet ball handling	Coordination - tennis-specific
A-4	Figure-of-eight sprint with bending	Agility - tennis-specific
B-1	Balance beam turn-arounds	Dynamic balance
B-2	Balance beam walk with racquet ball handling	Dynamic balance
B-3	Side steps on balance beam	Dynamic balance
E-1	2400-metre run	Endurance

Legend: ext. - extremity

- data pre-processing (class discretization – only for classification algorithms),
- selection of the most promising variables (all variables used, ReliefF, wrapper approach – included in the learning phase, variables selected by coaches),
- learning phase, in which the model was generated from a learning set (70% cases from the sample), and
- test phase, in which the model was tested and accuracy was calculated.

An automatic selection of the most promising variables was conducted by means of the ReliefF method and the wrapper approach. Both algorithms select the most promising variables for each sample separately. While ReliefF was set to select the best five variables, the wrapper approach selects any number of variables that give the best score. On the basis of the selections made by the coaches, the five most commonly selected variables (S-1, A-2, M-1, S-4 and S-4) were selected for males and females for further analysis. The analysis of predictions of competitive performance with the selected variables was conducted using both classification and regression methods.

ReliefF (Kononenko & Kukar, 2007; Robnik-Šikonja & Kononenko, 2003) works independently from the learning algorithm and assumes neither *apriori* nor the conditional independence of the variables. Consequently, it works efficiently also when dependent variables are involved. The wrapper approach (Kohavi & John, 1997) conducts a search in the space with one of the search algorithms and adds or removes one or several variables in each iteration. Each iteration also includes a test on selected variables of the learning algorithm and calculations of the learning performance. In this study, the "hill-climbing" algorithm was used as the search algorithm in the wrapper approach while the cross-validation method was used as a measure for evaluating the learning performance.

In analysing the predictions of competitive performance with classification algorithms, it was necessary to discretize the class variable, since classification algorithms do not work with continuous classes (Bratko, 2001; Hand, et al., 2001; Kononenko & Kukar, 2007; Witten & Frank, 2005). The class was determined by the position on the ranking list of the Slovene Tennis Association, and was divided into two parts, i.e., the top ten players and others. The reason for this being our aim was to separate top players from the rest, as only top players can succeed on the international level. Classification was performed by means of several methods: the naive Bayes classification method, decision tree, the C4.5 algorithm (Quinlan, 1993), the *k*-nearest neighbour, support vector machine (SVM), and logistic regression. These methods use considerably different approaches (Kononenko & Kukar, 2007)

to identify a discrete class (with the exception of the decision tree and the C4.5 algorithm, since the C4.5 algorithm is a variation of the decision tree). The simplest method is the *k*-nearest neighbour method, whereas the most complex one is SVM (Witten & Frank, 2005). The evaluation of the performance of classifiers was conducted with classification accuracy (the number of correctly classified cases divided by the number of all cases in the sample), and surface under the ROC (receiver operating characteristic) curve, as well as by means of the *k*-fold cross validation method (Kononenko & Kukar, 2007), *k* equal to 10.

Regression analysis works with a continuous class (Hand, et al., 2001), and therefore no additional data pre-processing was necessary. Here, the class was also represented by position on the ranking list of the Slovene Tennis Association. The sample data was analysed by linear regression and by regression trees. Linear regression is commonly used in practice and is based on modelling the relation between dependent and independent variables, so that a linear model is obtained. In principle, regression trees are the same as decision trees, the difference being that they are able to predict a continuous class (Kononenko & Kukar, 2007). The evaluation of performance was conducted by means of the root mean square error, the mean absolute error and the relative absolute error (Orange, 2010).

Results

Regarding selection by means of the ReliefF method, the variables A-1 and C-3 stood out, whereas the wrapper approach selected C-3 and A-2 most commonly. On the other hand, the coaches selected S-1 and A-2 as the most significant variables, which is evident from Graph 1. Of the morphological variables, according to the selection made by coaches, M-1 proved to be the most promising, whereas the ReliefF method, gave preference to M-4 and M-3.

Graphs 2 and 3 present the results for the naive Bayes method, which on average proved to be the most reliable method for predicting competitive performance. By means of the naive Bayes method, predictions of performance were accurate in U12/U12, and in 12-16/12-16, for male and female tennis players. Predictions of performance in A16/U12, for male tennis players, were also accurate; however, due to the small sample used in the test, this result can hardly be given much weight.

Relative absolute errors of the regression trees method were higher than 1.0, which means that using this method for predictions of competitive performance of tennis players in practice serves no practical purpose (Kononenko & Kukar, 2007). Table 2 therefore presents the results for the linear regression method. Results of performance predictions were most accurate when all variables
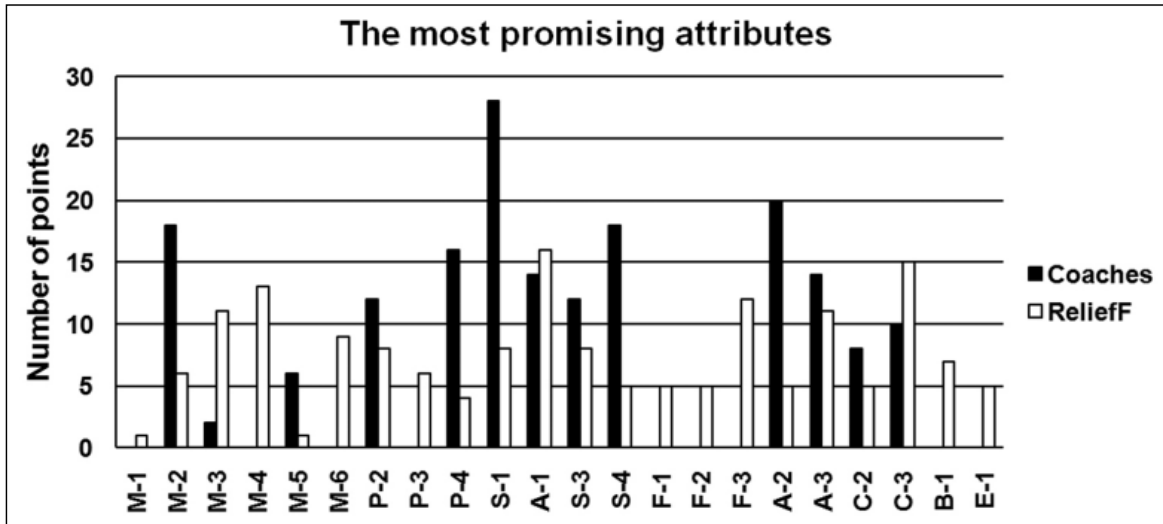
Graph 1. Compares the frequency of selection of the most promising morphological and motor variables for all categories using the ReliefF algorithm with the selection made by coaches.
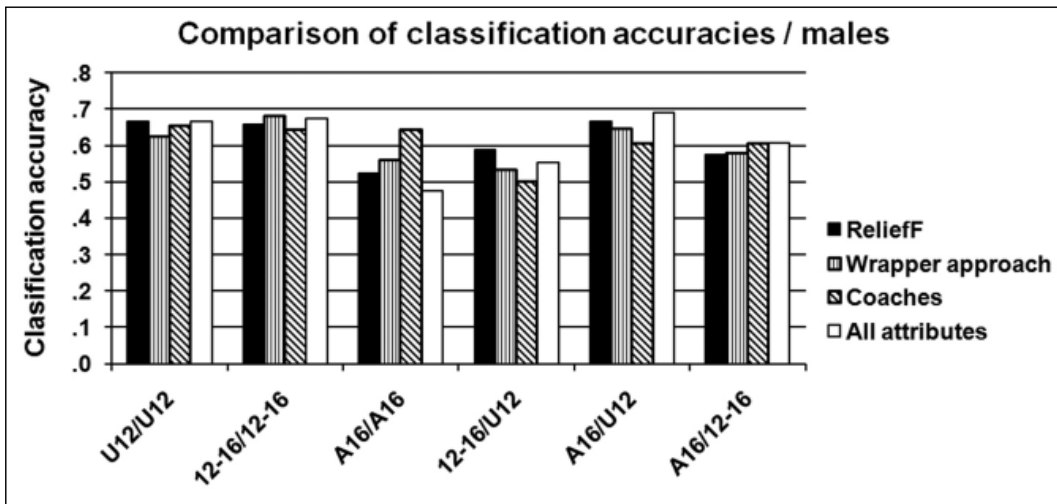
Graph 2. Comparison of classification accuracies of the naive Bayes method on all variables and on variables selected by means of ReliefF, the wrapper approach, and by coaches for male tennis players.
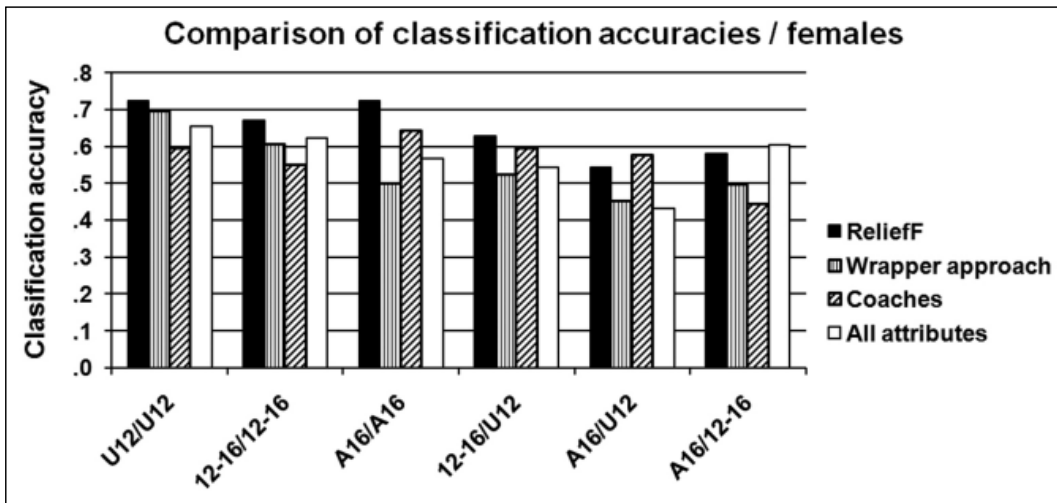
Graph 3. Comparison of classification accuracies of the naive Bayes method on all variables and on variables selected by means of ReliefF, the wrapper approach, and by coaches for female tennis players.

Table 2. Comparison of relative absolute errors of the linear regression method on all attributes and on attributes selected by means of ReliefF, and by coaches for male and female tennis players (there are no results for A16, as the sample was too small).

Method	U12/U12		12-16/12-16	
	Male	Female	Male	Female
ReliefF	.81	.72	.79	.91
Coaches	.81	.89	.97	.99
All attributes	.59	.63	.76	.77

were used, whereas they were least accurate when the linear regression only used variables selected by coaches.

Discussion and conclusions

S-1 and A-2 were selected by coaches as the most significant variables, which is evident from Graph 1. The selection of the S-1 in particular makes sense, as the acceleration significantly describes the criterion variable variance. Such observations have been made by various authors (Bunc, Dlouha, Höhm, & Safarik, 1990; Filipčič, A. & Filipčič, T., 2005a; Müller, 1989) who have recorded the significance of the importance of speed in explaining competitive performance in tennis.

Among the morphological characteristics, the coaches selected M-1 as the most promising one. Their selection is a good one, since body height is an advantage in most motor tasks, including tennis. Taller players can reach higher contact points (the point where the racquet hits the ball) in service, forehand, backhand, volley, and smash. Their extra reach enables them to hit distant balls more easily. As far as competitive performance is concerned, body height represents one of the most important morphological characteristics. Body height is also highly correlated with the length of individual body parts. A recent trend in increasing body height has developed and now male players at the top of the world tennis ranking list are 185 cm and more in height. Additionally, a greater body height might indicate accelerated physical and biological development in individual athletes. Two athletes of the same age may have completely different morphological (body) characteristics. Therefore, where the variable M-1 is concerned, the possibility of a faster or slower growth of a tennis player during puberty has to be taken into account (Filipčič, 1996).

Among the morphological characteristics based on the selection by means of the ReliefF method, M-4 and M-3 have proved to be the most promising. As far as somatotype is concerned, tennis is one of the sports which allow more diversity than some other sports (e.g., gymnastics, certain athletics

disciplines, etc.), which is the case especially as regards younger age groups. Čanaki, Sporiš and Leko (2006) analysed the body type of young tennis players aged between 16 and 18 on a sample of 42 subjects. They established that most players belonged to the mesomorph body type, whereas Gualdi-Russo, Brasili-Gualandi and Belcastro (1997) established that tennis players had larger amounts of subcutaneous fatty tissue than other athletes. Sánchez-Muñoz et al. (2007) observed differences in the somatotype between female tennis players who were ranked higher on the ranking list and those who were ranked lower. Female tennis players who were ranked higher on the ranking list were taller and had larger joint diameters. In comparing the quality groups of boys no differences were observed. Morphological characteristics undoubtedly exert an important influence on competitive performance indirectly, and to a large extent define the style of playing, the utilization of biomechanical principles, technical competency, and movement efficiency. Some morphological characteristics (body weight, percentage of body fat and lean body mass) to a large extent also indicate the fitness level of a tennis player, which is among the most significant factors of a competitor's performance.

Results of predicting competitive performance by means of classification methods for individual categories were expected, since the competitive performance in U16 is to a considerable extent influenced by the morphological and motor factors, whereas in A16, tactical and technical competencies, playing and competitive experience, as well as mental abilities play an increasingly more important role. The most promising variables, selected by the automatic machine learning methods, in U16 were C-3, A-1 and A-2, which is not consistent with the first part of our second hypothesis. Owing to bad prediction results in A16, we can neither confirm nor reject the second part of our second hypothesis. Because the ReliefF method and the wrapper approach have both selected C-3 frequently, it can be concluded that the results obtained by each of these two methods are comparable.

With regard to the C-3 test, it has been observed that the test is very specific and based on a mechanism which is not present in all other motor tests. The C-3 test specifically targets hand--to-eye coordination and is the only one that significantly describes a criterion variable variance. It tests the player's precision, suitable perception and evaluation of the ball flight, an ability that is highly important in the tennis game. The level of ability of eye-hand coordination is expressed in the optimal hitting point for each shot in the game. The precision of hitting the ball is expressed in the precision of hitting various parts of the court. A highly developed timing ability allows the player to execute movements or parts of a movement at an

exact moment, which is optimal from the aspect of the playing situation of the player in that instant. In the test used and in a real-life tennis game, the flight path of the ball presents the player with a narrow time frame in which to adjust his/her shot. This time frame during which the ball should be hit is precisely delimited and any deviation from this usually results in an error or the loss of a point (Filipčič, A. & Završki, 2002). To be successful at tennis, players need to develop their coordination, and the ability to execute complex motor tasks with a racquet and a ball precisely and fast.

A high, but not statistically significant value of correlation between the test C-3 and the criterion variable has also been found in a study by A. Filipčič (1996), who has also emphasized the importance of coordination in the development of young tennis players.

A-2 and A-1 are based on a very similar functional mechanism and are also highly correlated. Both tests involve agility, which is another key characteristic of the tennis game. Agility can be defined as the motor ability to carry out acceleration/deceleration types of locomotor movements effectively, including changes in direction. All of these are founded on neuromuscular power, quickness and feet coordination. Studies (Filipčič, 1996; Filipčič, A. & Filipčič, T., 2005a; Šerjak, 2000; Unierzyski, 1994) established that agility tests elucidate competitive performance at a statistically significant level.

Effectiveness of machine learning methods can be increased by eliminating those variables that adversely affect the successfulness of the learned model (Kohavi & John, 1997; Kononenko & Kukar, 2007). This way we eliminate those variables which do not have an impact on the class and therefore mainly negatively affect successfulness of the learned model. As a rule in our case, nothing is gained by selecting only the most promising variables in classification by means of the ReliefF method and the wrapper approach, in comparison with predicting competitive performance using all the morphological and motor variables. This means either that the ReliefF method and the wrapper approach are not suitable for use regarding the issues in question, or that, in view of a high interdependence among the variables, we are unable to be more successful in predicting competitive performance on the basis of morphological and motor factors.

A comparison of methods for selecting variables – made by using automatic methods and the selection of variables made by coaches – can be made only for U16, for both male and female tennis players, since performance predictions for the other age periods have proved to be inaccurate, and attempts to make such a comparison do not yield useful results. Automatic methods proved to be more reliable than coaches, which was most clearly noticeable with regard to female tennis players. This is also consistent with our first hypothesis. It needs, however, be pointed out that estimates made by coaches are based on a very holistic view on the performance of tennis players, and include an evaluation of players in all key evaluation areas (playing technique, psychology, fitness level).

With regard to linear regression, by selecting only the most promising variables, the relative absolute error increases and predictions of competitive performance are less accurate. This means that the selected variables are either not the ones which contribute most to the player's performance, or that the performance is also influenced by variables which were not selected in the first place. Here it has also been observed that the variables selected by the automatic method were better predictors than the variables selected by coaches, which is consistent with our first hypothesis.

Similar findings to these were recorded by A. Filipčič (1996), who compared the uniformity of estimates made by means of regression analyses in the fields of motor, morphological, and cardio-respiratory functional capacity of tennis players aged between 12 and 14, and estimates of potential performance made using expert modelling. The correlation coefficient between both estimates was .72. According to the author, the rather low correlation between estimates made by means of both procedures can be attributed to the fact that the estimate of the expert system does not reflect current relations between criterion and prediction variables, but is aiming to predict relations that will arise in future.

In future, it would make sense to carry out measurements and data collection in a more organized manner, as it has in certain cases been observed that a large number of values were missing. The reason for the missing results in the higher age categories is attributed to altered measurement procedures, since more sophisticated and accurate measurement procedures are used for A16 tennis players.

Regarding predictions of competitive performance, a future study including a larger number of factors which influence the competitive performance of tennis players would most likely produce even better results. It would also be most interesting to observe improvements with regard to the predictions of the players' competitive performance for several years in advance.

In this study, predictions of competitive performance of tennis players have turned out to be a highly complex issue, as the accuracy of predicted models, based on morphological and motor factors, was relatively poor. However, naive Bayes with all variables, ReliefF or the wrapper approach for U16 tennis players would be the best choice.

References

Banzer, W., Thiel, C., Rosenhagen, A., & Vogt, L. (2008). Tennis ranking related to exercise capacity. *British Journal of Sports Medicine, 42*(2), 152-154.

Bratko, I. (2001). *Prolog programming for artificial intelligence*, 3rd ed. Harlow: Addison-Wesley Longman Publishing.

Bunc, A., Dlouha, O., Höhm, J., & Safarik, J. (1990). Testova baterie pro hodnoceni urivne telesne pripravenosti mladyh tenistu. *Teorie a Praxe Telesne Vychovy, 38*(4), 194 - 203.

Čanaki, M., Sporiš, G., & Leko, G. (2006). Morphological advantages and disadvantages in Croatian u-16 and u-18 tennis players. *Croatian Sports Medicine Journal, 21*(2), 97-101.

Filipčič, A. (1996). *Evalvacija tekmovalne in potencialne uspešnosti mladih teniških igralcev.* [Evaluation of competitive and potential performance of young tennis players. In Slovene.] (Unpublished doctoral dissertation, University of Ljubljana). Ljubljana: Fakulteta za šport.

Filipčič, A., & Filipčič, T. (2005a). The influence of tennis motor abilities and anthropometric measures on the competition successfulness of 11 and 12 year-old female tennis players. *Acta Universitatis Palackianae Olomucensis. Gymnica, 35*(2), 35-41.

Filipčič, A., & Filipčič, T. (2005b). The relationship of tennis-specific motor abilities and the competition efficiency of young female tennis players. *Kinesiology, 37*(2), 164-172.

Filipčič, A., & Završki, S. (2002). Povezava med dvema testoma aerobnih kapacitet in tekmovalno uspešnostjo mladih teniških igralcev. *Kinesiologica Slovenica, 8*(1), 5-9.

Girard, O., & Millet, G.P. (2009). Physical determinants of tennis performance in competitive teenage players. *The Journal of Strength & Conditioning Research, 23*(6), 1867-1872.

Gualdi-Russo, E., Brasili-Gualandi, P., & Belcastro, M. (1997). Body composition assessment of young tennis players by multi-frequency impedance measurements. *International Journal of Anthropology, 12*(2), 11-20.

Hand, D.J., Mannila, H., & Smyth, P. (2001). *Principles of Data Mining.* Cambridge: The MIT Press.

Kohavi, R., & John, G.H. (1997). Wrappers for feature subset selection. *Artificial Intelligence, 97*(1-2), 273-324.

Kononenko, I., & Kukar, M. (2007). *Machine Learning and Data Mining.* Chichester: Horwood Publishing Ltd.

Müller, E. (1989). Sportmotorische Testverfaren zur Talentauswahl im Tennis. *Leistungssport, 19*(2), 5-9.

Orange. (2010). orngStat: *Orange Statistics for Predictors.* Retrieved January 16, 2010, from http://www.ailab.si/orange/doc/modules/orngStat.htm

Quinlan, J.R. (1993). *C4.5: programs for machine learning.* San Francisco: Morgan Kaufmann Publishers Inc.

Robnik-Šikonja, M., & Kononenko, I. (2003). Theoretical and empirical analysis of ReliefF and RReliefF. *Machine Learning, 53*(1), 23-69.

Sánchez-Muñoz, C., Sanz, D., & Zabala, M. (2007). Anthropometric characteristics, body composition and somatotype of elite junior tennis players. *British Journal of Sports Medicine,* 41(11), 793-799.

Šerjak, M. (2000). *Povezanost izbranih motoričnih sposobnosti in tekmovalne uspešnosti mladih teniških igralk.* [Connection of chosen motor variables with competitive successfulness of young female tennis players. In Slovene.] (Unpublished bachelor's thesis, University of Ljubljana). Ljubljana: Fakulteta za šport.

Unierzyski, P. (1994). Motor abilities and performance level among young tennis players. In: W. Osiński & W. Starosta (Eds.), *Proceedings of the International Conference „Sport Kinetics '93"* (pp. 309-313). Poznan: Institute of Sport, Warsaw.

Witten, I. H., & Frank, E. (2005). *Data mining.* San Francisco: Morgan Kaufmann.

Correspondence to:
Prof. Nejc Šarabon, PhD
Laboratory for Advanced Measurement Technologies
Wise Technologies, d.o.o.
Jarška cesta 10a, SI-1000 Ljubljana, Slovenia
Phone: +386 40 429 505
E-mail: nejc.sarabon@zrs.upr.si

PREDVIĐANJE NATJECATELJSKE USPJEŠNOSTI TENISAČA KORIŠTENJEM METODA STROJNOG UČENJA

Cilj istraživanja bio je testiranje mogućnosti predikcije natjecateljske uspješnosti slovenskih tenisača temeljem najvažnijih morfoloških mjera i testova motoričkih sposobnosti koji su bili odabrani na osnovu automatskih kompjutorskih metoda i od strane iskusnih teniskih trenera, uz upotrebu metoda strojnog učenja. U analizu je bilo uključeno 1002 muškaraca i žena koji su prošli redovita testiranja slovenskog nacionalnog teniskog saveza te su bili pozicionirani na njihovoj rang listi u sezonama od 1993 do 2008. Odabir najvažnijih varijablia uz pomoć dvije automatske metode dao je slične rezultate, dok su se varijable odabrane od strane trenera bitno razlikovale. Dokazano je kako je uz pomoć klasifikacijskih metoda moguća dobra predikcija natjecateljske uspješnosti u uzrastima ispod 16 godina, dok je kod uzrasta starijih od 16 godina ta predikcija loša. Među regresijskim metodama, za razliku od linearne regresije koja je dala zadovoljavajuće rezultate, regresijsko drvo pokazalo se praktično neupotrebljivim. Automatske metode odabira najvažnijih varijabla morfoloških karakteristika i motoričkih sposobnosti pokazale su se superiornima u odnosu na metode odabira koje koriste teniski treneri. Ta je činjenica bila najviše istaknuta kod ženskih tenisačica i kod upotrebe metode linearne regresije.

Ključne riječi: *tenis, identifikacija, selekcija, predviđanje, natjecateljska izvedba, strojno učenje*