NEURAL MODELS AS TOOLS OF SPORT PREDICTION

by

IGOR RYGUŁA[1]

One of most important problems of competitve sports is the proper selection of candidates for training. Its solution is above all related with the determination of such vector of variables (set of athlete personality features), which will bring the greatest quantity of information on the level of sports development of a given athlete. Another aspect of this problem concerns the determination of informative value (diagnostic value and prognostic value) of measuring tools (tests), which may be good indicators of the investigated features. The initial decision on the choice of candidates for sports training should be made on the basis of essential analysis of many features, among them somatic traits, efficiency and laboratory tests.

This work shows that the tool assisting the candidate selection process for sports training may be neural-genetic model and therefore several possible solutions of this subject were proposed. Their construction and evaluation were made on the basis of two-year experimental studies with groups of young girl handball players. The results of our studies confirm earlier findings of Wit and Barton (1996), that neural networks may have special usefulness in theory of sports training. However, as has been shown earlier, the use of neural-genetic models for sport prediction requires relatively great number of observations and small number of variables.

Key words: sport selection, artificial neural networks, genetic algorithms, result prediction.

Introduction

Artificial neural networks (ANN) are mathematical models created in the result of development in biological sciences in scope of composition and functioning of the nervous systems of living organisms (Lula 1999). Neural models may find application in solving many problems: modeling and forecasting, in taxonomy of structures, optimization and polyoptimization.

[1] Assist. Prof., Dep. of Theory of Sports, Acad. of Phys. Educ., Katowice
Correspondence address: 72a Mikołowska Str. 40-065 Katowice, Poland

ANN problems have strong position in world literature since 1943 when the pioneering work of McCulloch and Pitts (1943) has been published, containing mathematical model of the neural cell. Further developments were represented by the works of Hebb (1949), Rosenblatt (1961), or Widrow and Hoff (1960) - inventors of linear network and method of teaching. A feast of enormous influence on further development of ANN was development and generalization of the algorithm of backwards propagation of errors, used for effective teaching of multiplayer networks (Rumelhart, Hinton, Williams 1986).

Artificial neural networks have applications in great number of scientific disciplines, such as engineering (Korbicz 1994, Osowski 1996, Tadeusiewicz 1993), medicine (Lewenstein 1999), economics (Lula 1999), social science (Grabowski 1997) and others. With increased frequency, works are published on biomechanics (Barton, Lees 1993, Svelberg, Herzog 1995) and theory of sport training (Wit, Barton 1996). So far, the applications of neural models in the area of sports sciences are very limited, concerning mainly object recognition (discriminative analysis) and object grouping (taxonomic analysis). These are relatively simple tasks well tackled by neural networks. In sports problems solved so far (Wit, Burton 1996) it sufficed to divide the input set into two groups - teaching set and test set. The application of ANN to sport selections depends not only on the accurate mapping of interdependence between specified groups of input and output variables, but above all - on the ability of network to generalize. It is manifested by facility of determination of right responses on such input data that were not presented during teaching. This means that the network has obtained new skills and is able to increase the base of information on the analyzed phenomenon or to make positive or negative extrapolation of the trend function.

The objective of study

The objective of this study is a search for such set of measurable and heuristic variables that will bring the greatest amount of information on the suitability of girl handball players for sports training, and therefore to obtain answers on the following research questions:

1. Is there a possibility to make operationalization[2] of features leading to maximization of the hoard of information on investigated athletes?
2. Which is the optimum set of variables filling the role of prediction variables in girl handball players?

Material, methods and research tools

In order to obtain answers to the research questions formulated above, the author used the results of empirical studies made in three age groups (athletes born in 1982, 1981, 1980, 1979, 1978) of handball players in competition season 1995/96/97. Three series of measurements were made in age group in half-year intervals (Table 1).

Table 1. Mean value and standard deviation of height and body mass and fat tissue content in investigated groups in consecutive stages of pedagogic experiment

		Stage I (A) August 1995			Stage II (B) August 1996			Stage III (C) August 1997		
Birth year [age]	Number [n]	Body height [cm]	Body mass [kg]	Fat tissue content [%]	Body height [cm]	Body mass [kg]	Fat tissue content [%]	Body height [cm]	Body mass [kg]	Fat tissue content [%]
1982 [13]	33	162.21 ±6.3	51.17 ±8.6	20.65 ±3.3	164.61 ±6.1	52.72 ±8.3	19.64 ±2.9	165.39 ±6.0	54.89 ±8.1	20.67 ±1.1
1981 [14]	38	167.58 ±5.5	55.7 ±5.0	20.32 ±1.9	168.66 ±5.5	56.21 ±4.8	19.63 ±1.7	169.24 ±5.3	57.91 ±5.1	19.93 ±1.7
1980 [15]	20	168.6 ±5.3	57.48 ±6.4	22.58 ±3.9	169.5 ±5.0	58.61 ±6.0	55.5 ±3.3	169.95 ±5.1	59.95 ±5.9	22.55 ±3.4
1979 [16]	21	167.62 ±4.3	61.52 ±8.0	21.38 ±4.2	168.05 ±4.3	61.45 ±7.6	20.55 ±3.7	168.81 ±4.1	62.48 ±7.5	21.19 ±3.7
1978 [17]	21	166.53 ±5.7	62.7 ±6.6	21.9 ±2.5	167.1 ±5.8	62.25 ±6.6	21.03 ±2.1	167.6 ±6.3	62.68 ±7.0	21.13 ±2.0

In total, 26 features of 134 athletes in four measuring cycles were statistically analyzed (Table 2).

[2] Under operationalization is understood making algebraic transformations on their values in order to obtain new variables, characterizing other properties of investigated units compared with the case of input features (Grabiński, Malina, Zeliaś 1990).

Table 2. Descriptive parameters of the distribution of investigated features[3] obtained by the athletes in stage I (August 1995) of pedagogic experiment in group A (birth year 1982)

Pos.	VARIABLES	X	S	V	A_s	K_u
1	Body mass [kg]	51.167	8.493	16.6	-1.778	2.768
2	Body height [cm]	162.212	6.163	3.8	-2.789	2.904
3	Fat tissue [%]	20.652	3.211	15.5	-0.015	1.055
4	Maximum power [W/kg]	6.678	0.652	9.8	-2.403	3.601
5	Total energy [J/kg]	83.091	8.490	10.2	-2.398	3.554
6	Power decrease index [%]	2.947	2.316	78.6	0.365	1.973
7	Time to develop maximum power [s]	8.282	1.675	20.2	2.617	2.781
8	Time of keeping maximum power [s]	4.324	1.484	34.3	2.572	2.652
9	Start speed - 5 m [s]	1.206	0.075	6.2	1.412	3.000
10	Maximum speed - 10 m [s]	1.522	0.066	4.3	-1.389	2.962
11	Explosive power of lower limbs [cm]	167.091	14.431	8.6	-0.405	1.202
12	Explosive power of upper limbs [m]	8.558	0.928	10.8	-2.331	3.395
13	Agility [s]	11.902	0.572	4.8	-1.007	2.032
14	Endurance - shuttle run [stages]	7.636	1.356	17.8	-0.207	1.175
15	Suppleness [cm]	24.970	3.512	14.1	-1.620	3.098
16	PWC_{170} [W/kg]	1.879	0.384	20.5	-1.229	2.525
17	Index of basic technique in version I - $W_{TP}1$ [pts]	26.939	4.403	16.3	-0.617	1.507
18	Index of special motor skills $W_{SUM}1$ [s]	67.195	3.751	5.6	0.861	1.747
19	Index of special proficiency $W_{SS}1$ [pts]	0.404	0.081	19.9	-0.583	1.506
20	Index of basic technique in version II - $W_{TP}2$ [pts]	29.152	2.904	10.0	-0.155	1.061
21	Index of special motor skills $W_{SUM}2$ [s]	71.312	5.576	7.8	0.862	1.749
22	Index of special proficiency $W_{SS}2$ [pts]	0.413	0.066	15.9	-0.851	1.950
23	Index of special coordination skills - W_{KZS} [s]	4.117	2.241	54.4	0.857	1.735
24	Index of synthetic special proficiency - W_{SSS} [pts]	0.409	0.070	17.2	-0.708	1.701
25	Intelligence (spatial imagination) [pts]	27.152	8.877	32.7	0.898	1.862
26	Index of play efficiency [ju]	32.858	10.594	32.2	-0.396	1.802
27	Index of competitive development - WRZ [ju]	0.356	0.178	50.0	-0.161	1.599

[3] Description of measuring tools in Table 2 may be found in paper: Ryguła I., Jarząbek R. 2000: Predictive value of chosen body build characteristics, general and specific physical proficiency of girl handball players. Biology of Sport, no. 3.

The basic method used in this study was the application of neural network (Ryguła 2000). The tool used for building neural models was SNN program (Statistica Neural Networks).

Measures of fitting of neural model to real data

The natural measure of the quality of neural model is the sum of squares of deviations between real and theoretical values (SSE):

$$SSE = \sum_{p=1}^{R} \sum_{i=1}^{M} \left(d_{pi} - y_{pi} \right)^2$$

where:
R - number of elements in teaching set,
M - number of neurons in output layer,
d_{pi} - assumed value,
Y_{pi} - value calculated by the network.

The deficiency of the presented measure is the lack of possibility to compare networks describing different phenomena and dependence on the number of observations used for model construction. These deficiencies are partly eliminated in mean square error (MSE), being one of most popular measures describing the quality of a network. Its formula is (Azoff 1994):

$$MSE = \frac{\sum_{p=1}^{R} \sum_{i=1}^{M} \left(d_{pi} - y_{pi} \right)^2}{RM}$$

In this study, the quality of network was evaluated on the basis of two measures: normalized error (NRMSE) and Pearson correlation coefficient (r), calculated between real and computed values of the explained variable based on the model (index of competitive development).

The NRMSE value may be defined as:

$$NRMSE = \frac{RMSE}{s_d}$$

Where RMSE = √MSE

Characteristics of neural regressive model

Selection of the candidates for sport training may be done in many ways. One of them is the application of neural regressive model. The starting point is the development of neural model representing dependence between the characteristics of a candidate for sport training (body build, results of proficiency and laboratory tests) and quantities determining further sport development. The diagram of such model is shown in fig. 1.

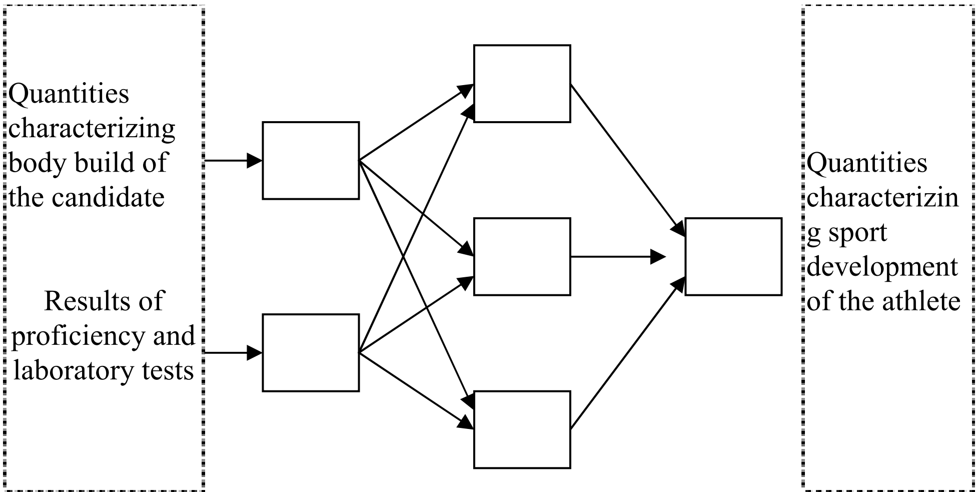

Fig. 1. Diagram of neural regressive model.

The developed model has a prognostic character. On the basis of the quantities characterizing the candidate the model should evaluate the level of his or her future development. The process of network construction should be effected with the use of teaching set containing data for different periods of time (earlier input data and later output quantities). It is assumed that the created model will have wide application capability. Its basic function is forecasting future level of sport development of candidates for training. In such forecasts it should be remembered that model so developed does not takes into account the influence of sport training on the obtained level of development. When evaluating the potential capabilities of the candidate it should be assumed that his or her training will have similar course to the training of athletes considered in teaching set.

The analysis of model sensitivity enables evaluation of the predictive value of the considered characteristics of body traits and results of proficiency tests. The course of investigation is following: after evaluation of the model, the measures of its fitting to available data. Next, the estimated network is launched many times, substituting certain constant value for one variable (e.g. the mean value of this variable, determined on the basis of teaching set). The number of repetitions of the experiment is equal the number of input variables. In each case the measure of model fitting is determined. The comparison of quality measures of the model considering a full set of variables and of the model without one variable enables to evaluate the importance of disregarded input - the bigger worsening of model quality, the more important was disregarded variable.

Modeling the dependence between chosen characteristics of body build and general and specific physical proficiency and the level of sport development

During the computations a set was used containing 133 cases (Table 1), that were divided into randomly three subsets:
teaching subset - 85 cases;
validation subset - 25cases;
test subset – 23 cases.

Construction of the linear model

The aim of further work was building a model describing relation between variables x_1 - x_{26} (shown in Table 2) and the variable representing the level of sport development of girl handball players (Y). The process of construction has begun with an attempt to develop a model using linear neural network (such model is equivalent of the linear regression function). The linear network has relatively small capability, but the construction of it is worthwhile for the following reasons:

a) If the relation between input variables and output variable is linear, the network may describe this relation. In such situation there is no need to construct nonlinear models (the structure and teaching of them may present some problems);

b) If the relation between the investigated variables has nonlinear character, beginning work from the construction of linear model is also founded, for such model will serve as a reference during comparing nonlinear models (we will be able to determine how many times the nonlinear model is better or worse than the linear one).

The linear networks always has two layers. Greater number of layer is unnecessary, as combination of two linear layers will function in the same way as single linear layer.

For all variables minimax conversion has been chosen. This means that before using the data in calculations they are scaled to the range assumed by the user (default range is 0..1).

After defining the network it is subjected to teaching process (that is parameters estimation - weights and threshold levels). Estimation of the parameters of the linear regression function was effected with the use of least square method. Then it has been tested.

Table 3. The values of quality measures of linear model

Network quality measures	Teaching set	Validation set	Test set
NRMSE	0,423722	1,032629	0,5274803
r	0,9115122	0,5209175	0,851949

During the evaluation of regression model two measures described earlier were considered:

- NRMSE - this measure is always nonnegative. The smaller value indicates better quality of the model. For very good model this measure has values in the range 0 .. 0.1. If NRMSE value is greater than 1, using this model is unfounded, for more accurate estimation of variable value is its arithmetic mean determined on the basis of teaching set;
- Correlation factor (r) - this measure has values in the –1,1 range. The higher the value, the better is the model.

Table 3 shows these measures in three columns - first pertains to the teaching set, second - validation set and third to test set.

NRMSE value for teaching set is 0.41. This indicates that linear network is not able to describe correctly the relation between input variables and output variable. The more it cannot generalize the obtained knowledge (1.03 and 0.52 for validation and test set respectively). Small practical value of this model is confirmed by relatively low values of correlation factor. The obtained results indicate that investigated relation in not linear and an attempt should be made to construct a nonlinear model.

Construction of a nonlinear model

Construction of a nonlinear neural model is much more difficult then construction of the linear network. The reasons are:

a) A correct kind of neural network must be chosen for building nonlinear model[4];
b) The structure of the network must be correctly determined;
c) The network of the network must be correctly taught.

Creation of single-directional multiplayer network (MLP)

In this work the most popular kind of network was utilized - the one used for modeling relations, namely perceptron network (MLP - multiplayer perceptron). Our network had the following structure: 25-13-1, that is 25 neurons in input layer, 13 neurons in hidden layer and one neuron in output layer, making 1*(25x13+13+13x1+1)=352 connections.

Quasi-Newton algorithm was used for teaching the network.

The evaluation of the network is presented in Table 4.

Table 4.The values of quality measures of the nonlinear model

Quality measures of the network	Teaching set	Validation network	Test network
NRMSE	3.236e–06	1.441	0.698
r	1	0.454	0.766

4 Their characteristics are presented in paper: Ryguła I.2000. Tools of multidimensional analyses of sport training. AWF Katowice.

NRMSE value for the teaching set is very small and indicates that our network has accurately adapted to teaching data. This is confirmed by the value of correlation factor of 1, indicating total correlation between theoretical and real values. NRMSE for validation set is 1.441 which indicates that the built network does not operate correctly for data in validation set. The confirmation of incorrect function of the network is low value of correlation factor (r = 0.45). This gives the basis for opinion that model of this structure will not be able to generalize the knowledge won. Too close fitting of the network to teaching data, accompanied by the lack of ability to generalize is called overteaching. To reduce it, the size of teaching set must be increased (which is impossible in our case), to use the network with greatly reduced number of connections[5] and reduce number of input variables. The aim of our further actions was the reduction of the number of input variables, therefore reduction of the structure of the network built. The problem interesting us is building such structure of the network that maximizes the value of output neuron, independent of the number of inputs (in our case of investigated variables).

This problem has been solved with the use of the genetic algorithm[6]. To choose the optimum set of input variables with SNN, the basic genetic operators were used and following values of procedure parameters were assumed:

- Unit penalty - 0.001
- Population - 1000
- Generations - 750
- Smoothing - 0.31
- Crossover rate - 0.3
- Mutation rate - 1

The best result was obtained for the network with seven inputs (variable no. 4, 9, 13, 15, 21, 23, 25), two neurons in the hidden layer and one neuron in output layer, making 7x2+2+2x1+1=19 connections.

The estimation of the value of this network is shown in Table 5.

---

[5] Some papers suggest that for estimation of *n* parameter network (weights and thresholds), a teaching set should be used with 10 x *n* cases at least,

[6] Its function is among others described in paper of Goldberg D.: Genetic algorithms and their application. WNT, Warszawa 1995.

Table 5. The values of quality measures of the neuro-genetic model

Quality measures of the network	Training set	Verification set	Test set
NRMSE	0.101	0.097	0.089
r	0.967	0.982	0.989

The optimum combination of neuro-genetic model consists of the following variables[7], which in our study serve as predictive variables:

- X_4 - maximum power
- X_9 - start speed
- X_{13} - agility
- X_{15} - suppleness
- X_{21} - index of specific motor skills
- X_{23} - index of specific coordinative skills
- X_{25} - intelligence (spatial imagination)

Variables have chosen by genetic algorithm may be used to building of classical regression model.

The structural form of the model that may serve purposes of sport prediction is as follows:

$$Y = 81{,}277+1{,}767x_4-12{,}031x_9 -3{,}286x_{13}+0{,}344x_{15}-0{,}180x_{21}-1{,}889x_{23}+0{,}141x_{25}$$

where:

Y is an index of competitive development

Variance of random factor	= 60.74
Mean estimation error	= 7.793
Coefficient of random variability	= 20.206
Determination coefficient	= 0.866
Coefficient of multiple correlation	= 0.930

[7] Names and order as in Table 2.

Summary

The basic problem of this work was determination of predictive[8] value of motion tests and tests of the investigated athletes - girl handball players. The meaning of this action is creation of criteria of forecasting success of these athletes in sport training. Another very important element of this research was verification of tools that may be used for diagnosing and control of the effectiveness of training process. On the basis of our present and earlier work (Ryguła, Jarząbek 2000), it may be said that independently of the theory of work link of effective action (diagnose, forecast, planning, realization, control), very important role is played by the mathematical model of these phenomena. The application of it enables to determine the predictability of the separate interpretative variables (structural parameters of regression model), their information hoard on the level of sport development of girl handball players and it may be a means of communication between the separate links (layers) of multiplayer structure of sport training.

The experiments conducted by us on neuro-genetic models have shown that in case of the construction of neural models of phenomena concerning sport prediction, the networks with relatively small number of parameters should be preferred - number too big causes diminishing ability to generalize. It is also indicated to adapt the suitable algorithm serving to determine the suitable structure of the network. The results of our investigation indicate that specially useful method of choice of the structure are genetic algorithms, because of the global character of searching of the space of structures.

The results of our investigation confirm earlier findings of Wit and Barton (1996) that artificial neural networks may be specially suitable for theory of sport training. However, as has been shown before, the application of neuro-genetic models for sport prediction requires relatively great number of observations and suitable number of parameters (including investigated characteristics).

[8] Prediction is a process of reasoning into the future on the basis of a mathematical model (Ryguła 1976, Hertz J. 1993)

Conclusion

In sport prediction, the greatest suitability show neuro-genetic models with relatively small numbers of parameters, which have great ability to generalize.

In the group of investigated handball players the greatest predictive value has been manifested in such variables as: start speed (x_9), agility (x_{13}) and index of special coordinating abilities (x_{23}).

REFERENCES

Azoff E. M. 1994. Neural Network Time Series Forecasting of Financial Markets. John Wiley & Sons.

Barton G., Lees A. 1993. Development of a connectionist expert system to identify foot problems based on under foot pressure patterns. Clinical Biomechanics, 10.

Grabiński T., Malina A., Zeliaś A. 1990 Methods of analysis of empirical data on the basis of time-section series. AE Kraków (In Polish).

Grabowski M. 1997. The application of self-organizing maps of Kohonen characteristics in data analysis. Section of data classification and analysis of PTS, Z.4 (In Polish).

Hebb D.O. 1949. The organization of Behaviour, a Neuropsychological Theory. Wiley, New York.

Hertz J., Krogh A., Palmer R.G. 1993. Introduction to Theory of Neural Computation. WNT Warszawa (In Polish).

Korbicz J., Obuchowicz A., Uciński D. 1994. Artificial Neural Networks. Principles and applications. Akademicka Oficyna Wydawnicza PLJ, Wrszawa (In Polish).

Lewenstein K., Urbaniak K. 1999. Fahlman's neural network based system for brest cancer diagnosing. Medical & Biological Engineering & Computing, vol. 37.

Lula P. 1999. Feed-forward Neural Networks in the Modeling of Economic Phenomena. A.E. Kraków (In Polish).

Mc Culloch W.S., Pitts W. 1943. Alogical Calculus of Ideas Immanent in Nervous Activity. Bulletin of Mathematical Biophysics, vol. 5.

Osowski S. 1996. Neural Networks in Algorithmic Approach. WNT Warszawa (In Polish).
Rasenblatt F. 1961. Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms. Sportan Books, Washington D.C.
Rumelhart D.E., Minton G.e., Williams R.J. 1986. Learning Representations by Back-Propagating Errors. Nature, vol. 323.
Ryguła I. 1976. Method of initial choice of boys for competition training in high jump. Sport Wyczynowy nr 7 (In Polish).
Ryguła I., Jarząbek R. 2000. Predictive value of chosen characteristics of body build, general and special physical proficiency of girl handball players. Biology of Sport no. 3.
Ryguła I. 2000. Tools of multidimensional analysis of sport training. AWF Katowice.
Svelberg H., Herzog W. 1995. Artificial neural networks used for prediction of muscle force from EMG – patterns. Book of Abstracts. XV-th Congress of the International Society of Biomechanics. University of Jyväskylä.
Tadeusiewicz T. 1993 Neural networks. Akademicka Oficyna Wydawnicza RM. Warszawa (In Polish).
Widrow B., Hoff M.E. 1960. Adaptive Switching Circuits. Wescon Electric Show and Convention Record, New York.
Wit A., Barton G. 1996. Artificial neural networks. Sport Wyczynowy nr 9.