New Trends and Issues Proceedings on Humanities and Social Sciences

Volume 5, Issue 4 (2018) 52-57

ISSN 2547-8818
www.prosoc.eu

Selected Paper of 7th Cyprus International Conference on Educational Research (CYICER-2018) 07-09 June, 2018, Acapulco Hotel Convention Center in Kyrenia, Cyprus

Estimating the maximal oxygen uptake with new prediction models for college-aged students using feature selection algorithm

Mehmet Fatih Akay*, Department of Computer Engineering, Faculty of Engineering, Cukurova University, 01000 Adana, Turkey.

Mustafa Mikail Ozciloglu, Department of Electrical and Electronics Engineering, Faculty of Engineering, Kilis 7 Aralik University, 79000 Kilis, Turkey.

Ebru Cetin, Department of Physical Education and Sport Teacher, Faculty of Sport Science, Gazi University, 06200 Ankara, Turkey.

Imdat Yarim, Department of Physical Education and Sport Teacher, Faculty of Sport Science, Gazi University, 06200 Ankara, Turkey.

Shahaboddin Daneshvar, Department of Computer Engineering, Faculty of Engineering, Cukurova University, 01000, Adana, Turkey.

Suggested Citation:

Akay, F. M., Ozciloglu, M. M., Cetin, E., Yarim, I. & Daneshvar, S. (2018). Estimating the maximal oxygen uptake with new prediction models for college-aged students using feature selection algorithm. *New Trends and Issues Proceedings on Humanities and Social Sciences.* [Online]. *5*(4), pp 52–57. Available from: www.prosoc.eu

Selection and peer review under responsibility of Prof. Dr. Huseyin Uzunboylu, Near East University, Cyprus

Abstract

Maximum oxygen consumption (VO_2max) is important to observe the endurance of the athletes and evaluate their performance.. Aim is to develop new prediction models for college-aged students using Support Vector Machine (SVM) with Relief-F feature selection algorithm. Ten different models consisting of the predictor variables gender, age, weight, height, maximal heart rate (HRmax), time, speed, Perceived Functional Ability scores (PFA-1 and PFA-2) and Physical Activity Rating score (PA-R) have been created by Relief-F scores for prediction of VO_2max. The prediction models' standard error of estimates (*SEE*'s) and multiple correlation coefficients (*R*'s) have been calculated for evaluating their performances. For comparison purposes, Tree Boost (TB) and Radial Basis Function Network (RBFN) based models have also been developed. The results show that the prediction model including PAR, speed, time, weight, PFA-1, gender and HRmax gives the lowest *SEE* with 6.42 mL.kg^{-1}.min^{-1} and highest *R* with 0.79. Also, this study shows that the predictor variables HRmax and gender play a considerable role in VO_2max prediction.

* ADDRESS FOR CORRESPONDENCE: **Mehmet Fatih Akay**, Department of Computer Engineering, Faculty of Engineering, Cukurova University, 01000 Adana, Turkey.
E-mail address: mfakay@cu.edu.tr / Tel.: +90-322-3387101

Keywords: Maximum oxygen uptake, machine learning, feature selection.

1. Introduction

VO_2max is the maximum rate of oxygen consumption as measured during maximal exercise. VO_2max is very important to observe the endurance of the athletes and evaluate the performance of them in sport science, education and research (Abut, Akay & George, 2016). The most accurate method to assess VO_2max is directly measuring the oxygen uptake during graded, maximal exertion exercise on a treadmill or cycle ergometer in the laboratory (Eler, 2016; Hunn, Lapuma & Holt, 2002;). However, this technique requires expensive laboratory equipment, a great deal of time, continuous medical supervision and highly motivated subjects (Bandyopadhyay, 2013; George et al., 2009).

In literature, only a few studies exist on VO_2max prediction of Turkish athletes. In Kaya, Akay, Cetin & Yarim (2016); SVM, Multilayer Perceptron (MLP) and Single Decision Tree (SDT) were used on a dataset which included the data of 48 students. Age, height, weight, body mass index (BMI), test time (TT) and HRmax were used to predict VO_2max. It was shown that VO_2max of Turkish athletes could be predicted with reasonable error rates by using SVM. Dincer, Akay, Cetin, Yarim & Daneshvar (2016) predicted VO_2max of college-aged students using Multiple Linear Regression (MLR) and hybrid data, which was a combination of exercise data and questionnaire variables. Twenty-six students from the College of Physical Education and Sports at Gazi University participated in the experiments. The dataset included gender, age, height, weight, BMI, HRmax, TT, PFA and PA-R. This study suggested that the prediction equation, $VO_2\text{max} = -(7.42 \times \text{gender}) + (4.26 \times \text{age}) - (1.44 \times \text{BMI}) + (4.31 \times \text{HRmax}) + (3.64 \times \text{TT}) - (0.16 \times \text{PFA-1}) + (0.75 \times \text{PFA-2}) + (0.61 \times \text{PAR}) - 895.26$ yielded the lowest *SEE*. Akay, Cetin, Yarim, Abut & Kaya (2016) established new prediction equations for estimating VO_2max from gender, age, height, weight, BMI, HRmax and TT for college-aged students in Turkey. In more details, 18 students from the College of Physical Education and Sports at Gazi University volunteered for that study. Twelve VO_2max prediction equations had been established by using MLR. The obtained results showed that the regression equation, $VO_2\text{max} = -(12.331 \times \text{gender}) - (0.805 \times \text{age}) + (0.883 \times \text{height}) - (1.167 \times \text{weight}) - (0.052 \times \text{HRmax}) - (0.158 \times \text{TT}) + 6.473$, gave the lowest *SEE* and the highest *R*. Ozciloglu, Akay, Cetin & Yarim (2016) developed submaximal VO_2max prediction models for Turkish college students by using SVM, MLP and MLR. The dataset included data of 65 students from the College of Physical Education and Sport at Gazi University. To predict VO_2max, two categories of prediction models had been formed. In the first category, the common predictor variables in each model were gender, age, height and weight, whereas the models in the second category had common predictor variables gender, age and BMI. Rest of the predictor variables for both categories were time, speed and submaximal heart rate (HRsmax). It was shown that the models consisting of the common predictor variables together with solely time yielded the lowest *SEE*'s for prediction of VO_2max in each category by using SVM. In Akay, Cetin, Yarim & Ozciloglu (2017), MLR-based VO_2max prediction models were developed by using physiological and questionnaire variables. Seven different models including gender, age, weight, height, PFA and PA-R had been used to predict VO_2max. This study suggested that the prediction models including PAR gave significant improvements for VO_2max prediction. In addition, MLR models could be used to predict VO_2max accurately for college-aged sport students in Turkey. Akay, Cetin, Yarim, Bozkurt & Ozciloglu (2017) used SVM, Generalized Regression Neural Network (GRNN), RBFN and Decision Tree Forest (DTF) to predict VO_2max of Turkish athletes. Fifteen different VO_2max prediction models had been created with gender, age, height, weight, HRmax, grade, speed and TT. It was shown that GRNN-based models usually produced much lower *SEE*'s and higher *R*'s than the ones given by SVM-, DTF- and RBFN-based models. On the other hand, the RBFN-based models yielded the worst performance with unacceptable error rates.

The purpose of this study is to develop new prediction models for college-aged students using SVM combined with Relief-F feature selection algorithm. The dataset includes the data of 97 (40 females and 57 males) students, ranging in age from 15 to 33 years, from the College of Physical Education and Sports Science at Gazi University. Ten different models have been created by using Relief-F scores of the predictor variables for prediction of VO_2max. The prediction models' *SEE*'s and *R*'s have been calculated for evaluating their performances. The results show that the prediction model including PAR, speed, time, weight, PFA-1, gender and HRmax gives the lowest *SEE* with 6.42 mL.kg^{-1}.min^{-1} and highest *R* with 0.79. Also, this study shows that the predictor variables HRmax and gender play a considerable role in VO_2max prediction.

The rest of the paper is organised as follows. Section 2 describes dataset generation. Section 3 introduces prediction methods and feature selection algorithm. Section 4 gives results and discussion. Section 5 concludes the paper.

2. Dataset generation

All subjects were informed prior to the maximal exercise test and they signed a consent participant form before participating in the tests. During the exercise test that was performed on a treadmill (HP COSMOS, Germany), a subject had been forced until he/she showed maximal performance. In other words, the test continued until the subject was exhausted.

During the maximal test using the maximal stepwise running exercise protocol, each subject's HRmax was measured and registered every 15 seconds. The maximal oxygen consumption capacities of participants were measured with the Cosmed Quark CPET system (Cosmed Quark CPET; Rome, Italy) by breath-by-breath technique. In addition to HRmax, tidal volume, VO_2max and respiratory exchange ratio were also recorded every 15 seconds. VO_2max test protocol started with running at 0° incline and at a speed of 8 km/h for women and at a speed of 10 km/h for men. Speed was incremented by 1 km/h every minute until 15km/h speed level was reached. Upon reaching 15 km/h speed, the incline started to increase by 1.5° each minute and the test continued until the athlete got exhausted. Statistical information about the dataset is shown in Table 1.

Table 1. Statistics of the dataset

Predictor variable	Minimum	Maximum	Mean	Standard deviation
Gender	0	1.00	0.59	0.49
Age (year)	15.00	33.00	20.82	3.18
Weight (kg)	44.00	95.00	65.82	10.96
Height (cm)	153.00	193.00	173.01	7.71
HRmax (bpm)	131.00	144.00	139.41	2.26
Time (s)	2.26	11.35	4.93	1.71
Speed (km)	6.00	14.00	8.37	1.56
PFA-1	2.00	8.00	5.04	1.46
PFA-2	1.00	9.00	4.11	2.05
PAR	1.00	10.00	6.10	2.80
VO_2max (ml.kg-1.min−1)	35.21	87.95	52.67	10.42

3. Methodology

By using the Relief-F feature selection algorithm, ranking of the predictor variables has been calculated. Then, based on these ranking scores, ten different models have been developed by removing the predictor variable with the lowest score at a time. The Relief-F ranking results are shown in Table 2.

Table 2. Relief-F scores of the predictor variables

Variables	Relief-F scores
PAR	0.01680
Speed	0.01615
Time	0.01089
Weight	0.01067
PFA-1	0.00366
Gender	−0.00110
HRmax	−0.00423
Age	−0.00452
PFA-2	−0.00958
Height	−0.01154

The accuracy of an SVM prediction model depends on the value of cost (C), type of the kernel function and parameters of this function (Jaganathan, Rajkumar & Kuppuchamy, 2012). The RBF kernel requiring the optimisation of the parameter gamma (γ) is selected for creating the SVM prediction models. There is no way to determine in advance which C and γ are ideal for a regression problem. Hence, one needs an effective search algorithm to find the best values of these parameters. Grid search has been implemented in order to find the optimal values of C and γ. A cross validation within the grid search is utilised in order to develop the generalisation capability of the SVM prediction models.

The success of RBFN relies on numerous factors (Flyer, Wright & Fornberg, 2014). The neuron number of the hidden layer has to be specified before the parameter selection for the RBF. After the neuron number of the hidden layer is selected, the success of the RBF depends on the maximal number of neurons, radius of the RBF and lambda (Nicolaos & Karayiannis, 2003).

TB is a series of trees and used to combine the subsequent tree with the output of preceding tree in the series. The error obtained in the first tree is minimised, and then is added to the subsequent tree. To increase the accuracy of predictive function, many trees can be added to the series. The maximum numbers of trees used in series, the depth of individual trees and the minimum size node to split affect the performance of TB models (Nassif, Capretz, Ho & Azzeh, 2012).

Table 3. Values of the utilised parameters for SVM, RBFN and TB

Method	Parameter	Value
SVM	Cost (C)	[1–100]
	Gamma (γ)	[0.00–50]
	Kernel Function	RBF
RBF	Maximal number of neurons	8
	Radius of the RBF	[0.001–400]
	Lambda (λ)	[10–100]
TB	Maximum number of trees used in series	[200–550]
	Depth of individual trees	[5–7]
	Minimum size node to split	[8–23]

The performance of the prediction models has been evaluated using *SEE* and *R*, the formulas of which are given in Eqs. (1) and (2), respectively. In Eqs. (1) and (2), Y is the measured VO_2max, Y' is the predicted VO_2max, $\overline{Y}$ is the average of the measured values of VO_2max and N is the number of subjects in the dataset.

$$SEE = \sqrt{\frac{\sum (Y - Y')^2}{N}} \quad (1)$$

$$R = \sqrt{1 - \frac{\sum (Y - Y')^2}{\sum (Y - Y')^2}} \tag{2}$$

4. Results and discussion

Table 4 shows the *SEE*'s and *R*'s of SVM-, TB- and RBF-based models along with the predictor variables. The prediction models are sorted by *SEE* values in rising order.

Table 4. *SEE* and *R* values of VO_2max prediction models

Models	Predictor variables	SVM		TB		RBF	
		SEE	*R*	*SEE*	*R*	*SEE*	*R*
Model 4	PAR, Speed, Time, Weight, PFA-1, Gender, HRmax	6.415	0.785	7.771	0.662	7.740	0.661
Model 2	PAR, Speed, Time, Weight, PFA-1, Gender, HRmax, Age, PFA-2	6.632	0.768	7.823	0.656	8.921	0.509
Model 3	PAR, Speed, Time, Weight, PFA-1, Gender, HRmax, Age	6.675	0.765	7.862	0.652	8.975	0.501
Model 1	PAR, Speed, Time, Weight, PFA-1, Gender, HRmax, Age, PFA-2, Height	6.702	0.763	7.881	0.649	9.325	0.437
Model 5	PAR, Speed, Time, Weight, PFA-1, Gender	6.799	0.755	7.895	0.648	9.559	0.387
Model 6	PAR, Speed, Time, Weight, PFA-1	7.840	0.654	9.148	0.471	9.841	0.315
Model 9	PAR, Speed	8.214	0.610	9.266	0.449	10.670	0.305
Model 8	PAR, Speed, Time	8.494	0.574	9.937	0.285	10.881	0.269
Model 7	PAR, Speed, Time, Weight	8.622	0.556	9.988	0.268	11.257	0.245
Model 10	PAR	9.509	0.399	10.013	0.259	11.291	0.232

Regarding the results obtained, the following discussions can be made:

- SVM-based prediction models show better performance than the prediction models based on other machine learning methods. In particular, SVM-based models yield in average 13.34% and 22.91% lower *SEE*'s than the *SEE*'s of TB- and RBFN-based models for VO_2max prediction, respectively.
- The outcomes indicate that Model 4 including PAR, speed, time, weight, PFA-1, gender and HRmax gives the lowest *SEE* with 6.42 $mL.kg^{-1}.min^{-1}$ and highest *R* with 0.79. In contrast, Model 10 including PAR yields the worst performance for all machine learning methods.
- When Model 5 including PAR, speed, time, weight, PFA-1 and gender, and Model 4 including the predictor variables PAR, speed, time, weight, PFA-1, gender and HRmax, are compared, it can be observed that HRmax provides a significant improvement for prediction of VO_2max. In more detail, the inclusion of HRmax in the aforementioned model leads in 5.64%, 1.57% and 19.05% reduction in *SEE* for SVM, TB and RBF, respectively.

5. Conclusion

In this study, Relief-F feature selection algorithm has been evaluated to calculate ranks of all predictor variables. Based on these ranking scores, ten different VO_2max prediction models for Turkish college students have been developed by removing the predictor variable with the lowest score at a time. The results show that the model including PAR, speed, time, weight, PFA-1, gender and HRmax yields the best performance. Also, this study shows that the predictor variables HRmax and gender play a considerable role in VO_2max prediction. Future work can involve using different feature selection algorithms with different machine learning methods to advance the accuracy of VO_2max prediction.

Acknowledgement

The authors would like to thank Cukurova University Scientific Research Projects Center for supporting this work under Grant no. FBA-2016-5537.

References

Abut F., Akay M. F. & George J. (2016) Developing new VO(2)max prediction models from maximal, submaximal and questionnaire variables using support vector machines combined with feature selection. *Journal of Computers in Biology and Medicine, 79*, 182–192.

Akay M. F., Cetin E., Yarim, I., Abut, F. & Kaya K. (2016, March 31–April 2). *New regression equations for estimating the maximal oxygen uptake of college of physical education and sports students in Turkey.* In Proceedings of 5th Cyprus International Conference on Educational Research. (pp. 4–7), Kyrenia, North Cyprus.

Akay, M. F., Cetin, E., Yarim I. & Ozciloglu, M. M. (2017, May 3–5). *New prediction models for the maximal oxygen uptake of college-aged students using non-exercise data*. In Proceedings of the 6th Cyprus International Conference on Educational Research, (pp. 1–1), Kyrenia, North Cyprus.

Akay, M. F., Cetin, E., Yarim, I., Bozkurt, O. & Ozciloglu, M. M. (2017, September 16–17). *development of novel maximal oxygen uptake prediction models for Turkish college students using machine learning and exercise data.* In Proceedings of the 9th International Conference on Computational Intelligence and Communication Networks, (pp. 1–1), Kyrenia, North Cyprus.

Bandyopadhyay, A. (2013). Validity of 20 meter multi-stage shuttle run test for estimation of maximum oxygen uptake in male university students. *Indian Journal of Physiology Pharmacol, 57*(1), 77–83.

Dincer, O. F., Akay, M. F., Cetin, E., Yarım I. & Daneshvar S. (2016, October 26–28). *New prediction equations for estimating the maximal oxygen consumption of college-aged students using hybrid data.* In Proceedings of the 1st International Mediterranean Science and Engineering Congress, (pp. 1–1), Adana, Turkey.

Eler, S. (2016). Effects of short term camp periods on aerobic and anaerobic performance parameters in ice hockey national team athletes. *International Journal of Environmental and Science Education*, *11*(5), 973–977.

Flyer, N., Wright, G. B. & Fornberg, B. (2014). Radial basis function-generated finite differences: a mesh-free method for computational geosciences. *Handbook of Geomathematics*. Berlin, Germany: Springer.

George, J. D., Paul, S. L., Hyde, A., Bradshaw, D. I., Vehrs, P. R., Hager, R. L. & Yanowitz, F. G. (2009). Prediction of maximum oxygen uptake using both exercise and non-exercise data. *Journal of Measurement in Physical Education and Exercise Science, 13*(1), 1–12.

Hunn, H. M., Lapuma, P. T. & Holt, D. T. (2002). The influence of pre-test anxiety, personality and exercise on VO_2max estimation. *Journal of Exercise Physiology, 5*(1), 5–14.

Jaganathan, P., Rajkumar, N. & Kuppuchamy, R. (2012). A comparative study of improved F-score with support vector machine and RBF network for breast cancer classification. *International Journal of Machine Learning and Computing, 2*(6), 741–745.

Kaya, K., Akay, M. F., Cetin, E. & Yarım, I. (2016, March 19–20). *Development of new prediction models for maximal oxygen uptake using artificial intelligence methods*. In Proceedings of the International Conference on Natural Science and Engineering, (pp. 986–988), Kilis, Turkey.

Nassif, A. B., Capretz, L. F., Ho, D. & Azzeh, M. A. (2012, December 12–15). *Treeboost model for software effort estimation based on use case points.* In Proceedings of 11th International Conference on Machine Learning and Applications, (pp. 314–319), Boca Raton, FL.

Nicolaos, B. & Karayiannis, M. M. (2003). The construction and training of reformulated radial basis function neural networks. *Journal of IEEE Transactions on Neural Networks, 4*, 835–844.

Ozciloglu, M. M., Akay, M. F., Cetin, E. & Yarim, I. (2016, November 2–4). *Development of new maximum oxygen uptake prediction models for Turkish college students using support vector machines and submaximal data*. In Proceedings of the 4th International Symposium on Engineering, Artificial Intelligence and Applications, (pp. 19–20), Kyrenia, North Cyprus.