New Trends and Issues Proceedings on Humanities and Social Sciences

Issue 3 (2017)256-262

ISSN 2421-8030
www.prosoc.eu

Selected paper of 5th Cyprus International Conference On Educational Research (Cyicer-2016) 31 March-02 April 2016, University Of Kyrenia, Kyrenia North Cyprus

On the Application of Artificial Intelligence and Feature Selection in Sports Science Education and Research: A Case Study

Mustafa Mikail Ozcicloglu[a] , Kilis 7 Aralik University, Information Technology Department, Kilis, Turkey
Mehmet Fatih Akay [b] *, Cukurova University, Computer Engineering Department, Adana, Turkey
Dan Heil [c], Montana State University, Human Health and Development Department, Montana

Suggested Citation:

Ozciloglu, M. M., Akay, M. F. & Heil, D. (2017). On the Application of Artificial Intelligence and Feature Selection in Sports Science Education and Research: A Case Study. *New Trends and Issues Proceedings on Humanities and Social Sciences*. [Online]. 03, pp 256-262. Available from: www.prosoc.eu

Selection and peer review under responsibility of Assist. Prof. Dr. Cigdem Hursen, Near East University

Abstract

In sports science education and research, the use of artificial intelligence methods along with feature selection algorithms can be of great help for developing prediction models where experimental studies based on measurements are not feasible. In this paper, we present a case study in regards to how sports science can benefit from the use of artificial intelligence methods combined with a feature selection algorithm. More specifically, the purpose of our study is to develop prediction models for upper body power (UBP), which is one of the most important factors affecting the performance of cross-country skiers during races. The dataset, which includes 75 subjects, was obtained from the College of Education, Health and Development of Montana State University. Multilayer Perceptron (MLP) and Single Decision Tree (SDT) along with the minimum-redundancy maximum-relevance (mRMR) feature selection algorithm were used to produce prediction models for predicting the 10-second UBP (UBP_{10}) and 60-second UBP (UBP_{60}). The predictor variables in the dataset are protocol, gender, age, body mass index (BMI), maximum oxygen uptake (VO_2max), maximum heart rate (HRmax), time and heart rate at lactate threshold (HRLT) whereas UBP_{10} and UBP_{60} are the target variables. Based on the ranking scores of predictor variables assigned by the mRMR, 16 different prediction models have been developed. By using 10-fold cross-validation, the efficiency of the prediction models has been calculated with their multiple correlation coefficients (*R*'s) and standard error of estimates (*SEE*'s). The results show that using less amount of predictor variables than the full set of predictor variables can be useful for prediction of UBP_{10} and UBP_{60} with comparable error rates. The model consisting of the predictor variables gender, BMI, VO_2max, HRLT and time yields the lowest SEE's for prediction of UBP_{10}, while the model including the predictor variables gender, age, BMI and VO_2max gives the lowest SEE's for prediction of UBP_{60}, whichever regression method is used. Using these two models instead of the full set of predictor variables yields up to 4.95% and 6.83% decrement rates in *SEE*'s for MLP and SDT based UBP prediction models, respectively.

Keywords: Multilayer Perceptron; Single Decision Tree; sports science education

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

1. Introduction

Cross-country skiing has increasingly become a popular sport that continues to evolve due to advances in training, equipment and ski techniques. Cross-country skiing is a very exhausting sport because the cross-country skiers intensively use all of upper and lower body musculature. A principle component in this sport is the ability to take advantage of the generating ability of UBP. UBP is defined as the rate at which work can be performed using the arm, shoulder and trunk muscles. UBP is the most fundamental determinant of cross-country ski race performance. (Marsland, 2012; Heil & Camenisch, 2014).

Regardless of the frequency with which research efforts about UBP are presented in literature, the necessary equipment for measuring UBP is expensive and not commonly accessible. Previously conducted experiments regarding UBP measurements have been mainly performed on dedicated ergometers in special research laboratories. Moreover, there is no standardization in measurement

Cross-country skiing has increasingly become a popular sport that continues to evolve due to advances in training, equipment and ski techniques. Cross-country skiing is a very exhausting sport because the cross-country skiers intensively use all of upper and lower body musculature. A principle component in this sport is the ability to take advantage of the generating ability of UBP. UBP is defined as the rate at which work can be performed using the arm, shoulder and trunk muscles. UBP is the most fundamental determinant of cross-country ski race performance. (Marsland , 2012; Heil & Camenisch, 2014).

Regardless of the frequency with which research efforts about UBP are presented in literature, the necessary equipment for measuring UBP is expensive and not commonly accessible. Previously conducted experiments regarding UBP measurements have been mainly performed on dedicated ergometers in special research laboratories. Moreover, there is no standardization in measurement protocol of UBP. Consequently, it may be advantageous to predict rather than measure UBP using actual training data previously collected by tests on ergometers (Alsobrook & Heil, 2009).

In literature, numerous UBP prediction models were developed using different machine learning methods including Support Vector Machines (SVM), MLP, Generalized Regression Neural Network (GRNN), Radial Basis Function Neural Network, Decision Tree Forest, Cascade Correlation Network, Tree Boost, Gene Expression Programming and SDT (Akay, Abut, Daneshvar & Heil, 2015; Akay, Daneshvar, Isoglu & Heil, 2014; Akgol, Akay & Heil, 2015). The results showed that among the machine learning methods, the performance of UBP prediction models based on SVM was superior. Additionally, in (Akay, Akgol, Turhan & Heil, 2014; Akay, Abut, Ozciloglu & Heil, 2015; Ozciloglu, Akay & Heil, 2015; Ozciloglu, Abut, Akay, & Heil, 2015) machine learning methods were combined with feature selection algorithms including Relief-F and mRMR (Peng & Long, 2015) to build UBP prediction models and investigate the effect of predictor variables on UBP prediction. In general, the models developed by utilizing feature selection algorithms showed higher performance than the ones obtained without feature selection algorithms.

The purpose of this paper is to extend the study (Ozciloglu, Akay & Heil, 2015) by developing new UBP prediction models and investigating the effect of predictor variables on 10-second UBP (UBP_{10}) and 60-second UBP (UBP_{60}) prediction using a dataset in which BMI is used instead of height and weight. More specifically, the ranking of the predictor variables has been calculated using the mRMR algorithm and 16 UBP prediction models have been developed utilizing MLP and SDT with respect to the ranking of the predictor variables. Upon comparing the results presented in this paper with the ones given in (Ozciloglu, Akay & Heil, 2015), it can be inferred that UBP prediction models including the predictor variables height and weight perform better than the ones including the predictor variable BMI.

The rest of the paper is organized as follows. Section 2 describes dataset generation. Section 3 introduces MLP and SDT based models. Section 4 gives results and discussion. Section 5 concludes the paper.

2. Dataset Generation

The dataset, which includes 75 subjects, was obtained from the College of Education, Health and Development of Montana State University (MSU). Subjects came three times to the MSU Movement Science/Human Performance Laboratory. In the first coming, researchers recorded subjects' physical information including weight, height and age. After that, subjects performed three trials of a 30 seconds test and then a 60 seconds test on a custom-built double poling ergometer.

Firstly, subjects exercised on the double poling ergometer for 5 minutes. After that, subjects relaxed for 3 minutes before performing three successive trials of a 30 seconds exercise. The skier speeded up power output for the first 20 seconds of the test before double poling at best effort the last 10 seconds. The mean power output which was calculated by the ergometer for the last 10 seconds was named UBP$_{10}$. Before subjects gave 100% effort for the last two trials, they used the first of three trials as a practice, also warm-up, using roughly 80% of maximal effort during the last 10 seconds. Successive UBP$_{10}$ trials were divided to 3 minutes rest periods. Before exercising a single 60 seconds test, subjects relaxed for an additional 5 minutes. After that, subjects performed a single 60 seconds test during which the aim was to achieve the maximum average power output over the entire 60 seconds when starting from the last stop (Alsobrook & Heil, 2009).

Statistical information about the dataset is shown in Table 1.

Table 1. Statistics of variables

Predictor Variable	Minimum	Maximum	Mean	Standard Deviation
Protocol	0	1.00	0.77	0.43
Gender	0	1.00	0.49	0.50
Age	15.00	25.00	18.52	2.29
BMI (kg/m^2)	17.83	27.90	22.08	1.94
VO$_2$max (ml.kg^{-1}.min^{-1})	46.00	79.01	62.33	8.38
HRmax (bpm)	180.00	213.00	196.70	7.34
Time (s)	5.10	13.70	11.55	1.61
HRLT (bpm)	161.00	210.00	181.00	9.92
UBP$_{10}$ (W)	110.00	350.00	225.40	71.08
UBP$_{60}$ (W)	92.00	285.00	172.40	54.03

3. Methodology

Sixteen different models have been developed with the ranking scores of predictor variables calculated by the mRMR feature selection algorithm. Table 2 shows the ranking among the predictor variables for UBP_{10} and UBP_{60} obtained by the mRMR algorithm.

Table 2. Scores calculated by mRMR

mRMR scores for UBP_{10}		mRMR scores for UBP_{60}	
Predictor Variable	Score	Predictor Variable	Score
Gender	0.79	Gender	0.93
BMI	0.44	Age	0.89
VO_2max	0.38	BMI	0.80
HRLT	0.37	VO_2max	0.75
Time	0.26	HRLT	0.75
Age	0.22	HRmax	0.68
HRmax	0.21	Protocol	0.53
Protocol	0.16	Time	0.48

Two different machine learning methods were used to create the models for predicting UBP_{10} and UBP_{60} of cross-country skiers. MLP is a feed-forward artificial neural network model that maps sets of input data onto a set of convenient outputs. Multiple layers of nodes generate MLP in a directed graph, with each layer totally linked to the next one (Delashmit, Walter & Manry, 2005). The activation functions of the hidden layer and output layer as well as the amount of neurons in the hidden layer determine the performance of MLP based models. SDT is most commonly preferred in data mining issues to produce a model for guessing the value of a target attribute based on several input attributes. It follows the top down approach considering all attributes from root to leaf (Hamid, Ivanovich & Hamid, 2014). The parameters affecting the performance of SDT based models are minimum rows in a node, minimum size node to split and maximum tree levels.

Table 3 lists the ranges for values of the utilized parameters for MLP and SDT methods.

MLP and SDT methods.

Table 3. Values of the utilized parameters for MLP and SDT

Methods	Parameters	Value
MLP	Hidden layer neuron selection	2 - 20
	Hidden layer activation function	Logistic - Linear
	Output layer activation function	Logistic - Linear
SDT	Minimum rows in a node	2 - 10
	Minimum size node to split	6 - 10
	Maximum tree levels	7- 10

4. Results & Discussion

Table 4 and Table 5 show the SEE's and R's of MLP and SDT 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 UBP_{10} prediction models for MLP and SDT

Models	Predictor Variables	MLP		SDT	
		SEE	*R*	*SEE*	*R*
Model 4	Gender, BMI, VO_2max, HRLT, Time	33.56	0.88	40.30	0.82
Model 1	Gender, BMI, VO_2max, HRLT, Time, Age, HRmax, Protocol	34.43	0.87	42.76	0.80
Model 2	Gender, BMI, VO_2max, HRLT, Time, Age, HRmax	35.77	0.86	42.89	0.79
Model 3	Gender, BMI, VO_2max, HRLT, Time, Age	37.93	0.84	43.29	0.79
Model 5	Gender, BMI, VO_2max, HRLT	38.14	0.84	43.31	0.79
Model 7	Gender, BMI	38.87	0.84	43.37	0.78
Model 6	Gender, BMI, VO_2max	39.82	0.83	44.73	0.77
Model 8	Gender	40.17	0.82	45.09	0.77

Table 5. SEE and R values of UBP60 prediction models for MLP and SD

Models	Predictor Variables	MLP		SDT	
		SEE	*R*	*SEE*	*R*
Model 14	Gender, Age, BMI	23.44	0.90	26.51	0.87
Model 12	Gender, Age, BMI, VO_2max, HRLT	23.74	0.90	26.65	0.87
Model 13	Gender, Age, BMI, VO_2max	23.95	0.89	26.92	0.86
Model 9	Gender, Age, BMI, VO_2max, HRLT, HRmax, Protocol, Time	24.60	0.89	28.32	0.85
Model 10	Gender, Age, BMI, VO_2max, HRLT, HRmax, Protocol	24.79	0.89	28.94	0.85
Model 11	Gender, Age, BMI, VO_2max, HRLT, HRmax	25.14	0.88	29.68	0.83
Model 15	Gender, Age	25.25	0.88	30.75	0.82
Model 16	Gender	29.26	0.84	31.14	0.81

The following discussions can be made regarding the results obtained:

- In general, MLP based prediction models show better performance than SDT based models.
- The *SEE's* of UBP_{60} prediction models are in average 50.19% lower than the *SEE's* of UBP_{10} prediction models.
- The prediction model comprising of the predictor variables gender, BMI, VO_2max, HRLT and time (Model 4) yields the lowest *SEE*'s and highest *R's* for prediction of UBP_{10}, while the model consisting of the predictor variables gender, age and BMI (Model 14) leads to the lowest *SEE*'s and highest *R's* for prediction of UBP_{60}, regardless of whether MLP or SDT has been used.
- Using Model 4 instead of the full set of predictor variables yields 2.59% and 6.10% decrement rates in *SEE*'s for MLP and SDT. Similarly, when Model 14 is used instead of the full set of predictor variables, the decrement rates in *SEE*'s for MLP and SDT are 4.95%, 6.83%, respectively.
- For predicting UBP_{10} and UBP_{60}, MLP based models give in average 16.00% and 14.49% lower *SEE's* than the ones given by the SDT based models.
- The prediction models consisting of a single variable (i.e. gender) yield the highest *SEE's* and the lowest *R's* for prediction of UBP_{10} and UBP_{60}, regardless of whether MLP or SDT has been used.

- Upon comparing the results presented in this paper with the ones given in (Ozciloglu, Akay & Heil, 2015), it can be inferred that UBP prediction models including the predictor variables height and weight perform better than the ones including the predictor variable BMI. More specifically, GRNN based models including height and weight yield in average 26.64% and 29.26% lower *SEE*'s than MLP based models including BMI for predicting UBP_{10} and UBP_{60}, respectively.

5. Conclusion

This study demonstrates the fact that in sports science education and research, the use of artificial intelligence methods along with feature selection algorithms can be of great help for developing prediction models where experimental studies based on measurements are not feasible. In this study, various feature selection based models have been produced to predict UBP_{10} and UBP_{60} of cross-country skiers by using the MLP and SDT methods. In general, MLP based models give the lowest *SEE*'s and the highest *R's*. Future work can include using different machine learning methods combined with different feature selection algorithms to improve the accuracy of UBP prediction.

Acknowledgment

The authors would like to thank Cukurova University Scientific Research Projects Center for supporting this work under grant no. FDK-2015-5093.

References

Akay, M. F., Abut, F., Daneshvar, S. & Heil, D. (2015). Prediction of upper body power of cross-country skiers using support vector machines. *Arabian Journal for Science and Engineering, 40* (4), 1045–1055.

Akay, M. F., Abut, F., Ozciloglu, M. & Heil, D. (2015). Identifying the discriminative predictors of upper body power of cross-country skiers using support vector machines combined with feature selection. Neural Computing and Applications, doi: 10.1007/s00521-015-1986-9 .

Akay, M. F., Akgol, D., Turhan, I. & Heil, D. (2014). Prediction of upper body power of cross-country skiers using support vector machines combined with feature selection. *In Proceedings of the Second International Symposium on Engineering*, Artificial Intelligence & Applications, North Cyprus, 5-7 Nov 2014 (7-9).

Akay, M.F., Daneshvar, S., Isoglu, O. & Heil, D. (2014). Predicting the upper body of cross-country skiers using support vector machines. *In Proceedings of the 7th Engineering and Technology Symposium*, Ankara, Turkey, 15-16 May 2015 (21-24).

Akgol, D., M. F., Akay, M. F. & Heil, D. (2015). Development of new models for predicting upper body power of cross-country skiers using machine learning methods. *In Proceedings of the 1st International Symposium on Sport Science*, Engineering and Technology, Istanbul, Turkey, 10-13 May 2015 (34-40).

Alsobrook, N. G. & Heil, D. (2009). Upper body power as a determinant of classical cross-country ski performance. *European Journal of Applied Physiology*, *105* (4), 633–641.

Delashmit, W. H. & Manry, M. T. (2005). Recent developments in multilayer perceptron neural networks. In Proceedings of the 7th Annual Memphis Area Engineering and Science Conference, Memphis, Tennessee, 11 May 2005 (1-15).

Hamid, K., Ivanovich, E. V. & Hamid, K. (2014). An alternative approach for assessing sediment impact on aquatic ecosystems using Single Decision Tree (SDT). *Journal of Water Sustainability*, *4* (3), 181-204.

Heil, D. P. & Camenisch, K. (2014). Static flexibility as a correlate of skate roller skiing economy within collegiate cross-country skiers. *International Journal of Sports Science*, *4* (5), 188–197.

Heil, D. & Willis, S. (2012). In Science and Skiing V: Determinants of both classic and skate cross country ski performance in competitive junior and collegiate skiers, Germany: Meyer & Meyer Sport.

Lawson, S. K., Reid, D. C. & Wiley, J. P. (1992). Anterior compartment pressures in cross-country skiers. A comparison of classic and skating skis. *The American Journal of Sports Medicine, 20* (6), 750–753.

Marsland, F., Lyons, K., Anson, J., Waddington, G., Macintosh, C. & Chapman, D. (2012). Identification of cross-country skiing movement patterns using micro-sensors. Sensors (Basel, Switzerland), *12* (4), 5047–5066.

Ozciloglu, M., Akay, M. F. & Heil, D. (2015). New Prediction Models for Upper Body Power of Cross-Country Skiers Using Machine Learning Methods with Minimum-Redundancy Maximum-Relevance Feature Selection Algorithm. In Proceedings of the Third International Symposium on Engineering, Artificial Intelligence & Applications, North Cyprus, 4-6 Nov 2015 (16-17).

Ozciloglu, M., Akay, M. F., Abut, F. & Heil, D. (2015). Feature Selection For Prediction Of Upper Body Power Of Cross-Country Skiers Using Different Machine Learning Methods. In Proceedings of the 1st International Symposium on Sport Science, Engineering and Technology, Istanbul, Turkey, 4-6 Nov 2015 (1-5).

Peng, H. F. & Long, C. D. (2005). Feature selection based on mutual information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy. IEEE Transactions on Pattern Analysis and Machine Intelligence, *27* (8), 1226–1238.