Fatih Abut

Development of New Hybrid Models for Prediction of Maximal Oxygen Uptake (VO2max) Using Machine Learning Methods Combined with Feature Selection Algorithms

Doctoral Thesis / Dissertation

GRIN

Bibliographic information published by the German National Library:

The German National Library lists this publication in the National Bibliography; detailed bibliographic data are available on the Internet at http://dnb.dnb.de .

Imprint:

Copyright © 2017 GRIN Verlag
ISBN: 9783346551061

This book at GRIN:

https://www.grin.com/document/1156463

Fatih Abut

Development of New Hybrid Models for Prediction of Maximal Oxygen Uptake (VO2max) Using Machine Learning Methods Combined with Feature Selection Algorithms

GRIN Verlag

ÇUKUROVA UNIVERSITY
INSTITUTE OF NATURAL AND APPLIED SCIENCES

PhD THESIS

Fatih ABUT

DEVELOPMENT OF NEW HYBRID MODELS FOR PREDICTION OF VO_2MAX USING MACHINE LEARNING METHODS COMBINED WITH FEATURE SELECTION ALGORITHMS

DEPARTMENT OF COMPUTER ENGINEERING

ADANA, 2017

ÇUKUROVA UNIVERSITY
INSTITUTE OF NATURAL AND APPLIED SCIENCES

DEVELOPMENT OF NEW HYBRID MODELS FOR PREDICTION OF VO_2MAX USING MACHINE LEARNING METHODS COMBINED WITH FEATURE SELECTION ALGORITHMS

Fatih ABUT

PhD THESIS

DEPARTMENT OF COMPUTER ENGINEERING

We certify that the thesis titled above was reviewed and approved for the award of degree of the Doctor of Philosophy by the board of jury on 24/04/2017.

.....	.....	.....
Assoc. Prof. Dr. M. Fatih AKAY	Assoc. Prof. Dr. Zekeriya TÜFEKÇİ	Assoc. Prof. Dr. Uğur CEM HASAR
SUPERVISOR	MEMBER	MEMBER

.....	.....
Assoc. Prof. Dr. Serdar YILDIRIM	Asst. Prof. Dr. Ali ŞENTÜRK
MEMBER	MEMBER

This PhD Thesis is written at the Department of Institute of Natural And Applied Sciences of Çukurova University.
Registration Number:

Prof. Dr. Mustafa GÖK
Director
Institute of Natural and Applied Sciences

This thesis was supported by the Scientific Research Project Unit of Çukurova University. (Project Number: FDK-2015-5027).

ABSTRACT

PhD THESIS

DEVELOPMENT OF NEW HYBRID MODELS FOR PREDICTION OF VO_2MAX USING MACHINE LEARNING METHODS COMBINED WITH FEATURE SELECTION ALGORITHMS

Fatih ABUT

ÇUKUROVA UNIVERSITY
INSTITUTE OF NATURAL AND APPLIED SCIENCES
DEPARTMENT OF COMPUTER ENGINEERING

Supervisor : Assoc. Prof. Dr. M. Fatih AKAY
Year: 2017, Pages: 117
Jury : Assoc. Prof. Dr. Zekeriya TÜFEKÇİ
: Assoc. Prof. Dr. Uğur CEM HASAR
: Assoc. Prof. Dr. Serdar YILDIRIM
: Asst. Prof. Dr. Ali ŞENTÜRK

The purpose of this thesis is twofold. The first purpose is to develop new hybrid feature selection-based maximal oxygen uptake (VO_2max) prediction models using for the first time the double and triple combinations of maximal, submaximal and questionnaire variables. Several machine learning methods including Support Vector Machine, artificial neural network-based and tree-structured methods combined individually with three feature selectors Relief-F, minimum redundancy maximum relevance (mRMR) and maximum-likelihood feature selector (MLFS) have been applied for model development. The second purpose is to design a new ensemble feature selector, which aggregates the consensus properties of Relief-F, mRMR and MLFS to produce more robust decisions about the set of relevantly identified VO_2max predictors and to create more accurate prediction models. Using 10-fold cross validation on three different datasets, the performance of prediction models has been evaluated by calculating their multiple correlation coefficients (*R*'s) and root mean squared errors (*RMSE*'s). The results show that compared with the results of the other regular feature selection-based models in literature, the reported values of *R* and *RMSE* of the hybrid models in this thesis are considerably more accurate. Furthermore, prediction models based on the proposed ensemble feature selector outperform the models created by individually using the Relief-F, mRMR or MLFS, achieving similar or ideally up to 12.46% lower error rates on the average.

Key Words: Machine Learning, Feature Selection, Maximal Oxygen Uptake, Prediction

EXTENDED ABSTRACT

PhD THESIS

DEVELOPMENT OF NEW HYBRID MODELS FOR PREDICTION OF VO_2MAX USING MACHINE LEARNING METHODS COMBINED WITH FEATURE SELECTION ALGORITHMS

Fatih ABUT

ÇUKUROVA UNIVERSITY
INSTITUTE OF NATURAL AND APPLIED SCIENCES
DEPARTMENT OF COMPUTER ENGINEERING

Supervisor : Assoc. Prof. Dr. M. Fatih AKAY
Year: 2017, Pages: 117
Jury : Assoc. Prof. Dr. Zekeriya TÜFEKÇİ
: Assoc. Prof. Dr. Uğur CEM HASAR
: Assoc. Prof. Dr. Serdar YILDIRIM
: Asst. Prof. Dr. Ali ŞENTÜRK

Maximal oxygen uptake (VO_2max) is an essential part of health and physical fitness, and is defined as the highest rate of oxygen consumption attainable during maximal or exhaustive exercise. The knowledge of VO_2max is used for many different aims in sport and medical sciences. In sport sciences, it is especially useful as a sign of endurance capacity of an athlete, representing the upper limit for endurance performance. In medical sciences, knowledge of VO_2max is utilized to determine the cardiorespiratory fitness level of an individual, which is highly associated with the risk of heart disease, lung cancer, diabetes and many other diseases.

During the maximal graded exercise test (GXT), measurement of VO_2max is known as the most proper technique for the evaluation of endurance capacity. However, in spite of high level of accuracy, direct measurement of VO_2max is related to a number of practical difficulties and limitations. GXT's require trained staff as well as costly laboratory equipment, and are not convenient for some individuals, as the tests are of strenuous nature which in turn could pose a hazard to older or higher risk individuals. Also, GXT's are time-consuming, and not suitable for measuring VO_2max of large populations outside of the laboratory.

The practical limitations of direct testing have given rise to develop various regression models for predicting VO_2max rather than measuring it. Within the course of the past decade, numerous VO_2max prediction models based on maximal, submaximal or questionnaire variables have been proposed in literature. However, to the best of our knowledge, no study has ever attempted to evaluate the mixture of maximal, submaximal and questionnaire variables along with feature selection

algorithms to reveal the discriminative predictors of VO_2max, with the aim to create more accurate prediction models.

The purpose of this thesis is twofold. The first purpose is to build new hybrid feature selection-based VO_2max prediction models using for the first time the double and triple combinations of maximal, submaximal and questionnaire variables. The rationale is to remove the redundant and irrelevant variables from the sets of maximal, submaximal and questionnaire variables, and combine the most relevant variables from each category in a hybrid prediction model to increase the accuracy over a regular model. Three state-of-the-art feature selectors including Relief-F, minimum redundancy maximum relevance (mRMR) and maximum-likelihood feature selectors (MLFS) have been applied for model creation; whereas seven different machine learning methods including Support Vector Machine (SVM), Multilayer Feed-Forward Artificial Neural Network (MFANN), General Regression Neural Networks (GRNN), Radial Basis Function Neural Network (RBFNN), Tree Boost (TB), Decision Tree Forest (DTF) and Single Decision Tree (SDT) have been used for model development. The second purpose is to design a new ensemble feature selector, named as Majority Voting Feature Selector (MVFS), which aggregates the consensus properties of Relief-F, mRMR and MLFS to produce more robust decisions about the set of relevantly identified VO_2max predictors and to create more accurate prediction models.

Three different datasets, referred to as VO_2max-set-1, VO_2max-set-2 and VO_2max-set-3, have been used to create the prediction models. The common predictor variables appearing in each dataset are the physiological variables gender, age, height and weight. In addition to these variables, VO_2max-set-1 includes information related to 100 healthy subjects, who were selected from students at the Brigham Young University and workers from the LDS Hospital in Salt Lake City, Utah, and contains the maximal variables maximal heart rate (MX-HR), rate of perceived exertion (MX-RPE), treadmill grade (MX-Grade) and respiratory exchange ratio (MX-RER); the submaximal variables heart rate (SM-HR), ending speed (SM-ES) and stage (SM-Stage) of the treadmill; and finally the questionnaire variables perceived functional ability (Q-PFA) and physical activity rating (Q-PAR). VO_2max-set-2 incorporates data belonging to 440 healthy volunteers, who were selected from the Amherst and Worcester, Massachusetts, communities, and includes the maximal variables MX-HR, MX-RPE, MX-RER, exercise time (MX-Time) and the questionnaire variable active code (Q-AC). Finally, VO_2max-set-3 contains data related to 185 college students from Brigham Young University, and includes the maximal variables MX-HR and MX-RER, and the submaximal exercise times (SM-MIN1, SM-MIN2 and SM-MIN3) and heart rates (SM-HR1, SM-HR2 and SM-HR3) at the 0.5-mile mark, 1-mile mark and 1.5-mile mark, respectively.

By executing the Relief-F, mRMR, MLFS and MVFS algorithms on the three datasets, the relevance rank of every predictor variable has been computed separately and the variables have been sorted in ascending order in compliance with their ranks. The prediction models have been created by iteratively removing

the predictor variable with the lowest rank from the full set of predictor variables until a single variable with the highest rank forms the last model. In this way, several new hybrid VO_2max models for each of the four feature selectors have been created that have never been evaluated before in literature. The multiple correlation coefficient (*R*) and the root mean square error (*RMSE*) have been utilized to assess the performance of the models, whilst the evaluation of their generalization errors has been conducted using 10-fold cross-validation.

The results reveal that compared with the results of the other regular feature selection-based models in literature, that were created by using maximal, submaximal or questionnaire variables only, the reported values of *R* and *RMSE* of hybrid models in this thesis are considerably more accurate. More specifically, it is seen that prediction models giving the highest *R*'s and the lowest *RMSE*'s for the three datasets include at least one predictor variable from each utilized category of maximal, submaximal and/or questionnaire variables. This, in turn, confirms the effectiveness of combining the relevant predictors of VO_2max from different categories in a hybrid model to improve the accuracy over regular models. Particularly, for VO_2max-set-1, the model formed by combining the physiological variables gender, age, height and weight; the maximal variables MX-HR and MX-Grade; the submaximal variables SM-HR, SM-ES and SM-Stage; and finally the questionnaire variables Q-PFA and Q-PAR yields the highest *R*'s and the lowest *RMSE*'s for all utilized machine learning methods. For VO_2max-set-2, the model including all predictor variables produces the highest *R*'s and lowest *RMSE*'s, whereas for VO_2max-set-3, the model containing the physiological variables gender, age, height and weight; the maximal variable MX-HR; and the submaximal variables SM-MIN1, SM-MIN2, SM-MIN3 and SM-HR3 gives the highest *R*'s and the lowest *RMSE*'s, irrespective of the utilized machine learning methods.

Among the set of the three individual feature selectors including Relief-F, mRMR and MLFS, there is no generalizable superiority of one feature selector over the others that consistently performs well over all utilized datasets. However, it is seen that prediction models based on the proposed MVFS lead, on the average, to statistically significant higher *R*'s and lower *RMSE*'s than the models created by individually using the Relief-F, mRMR or MLFS algorithms, irrespective of the utilized datasets or machine learning methods. Particularly, compared with the performance of the three individual feature selectors on prediction of VO_2max, applying the ensemble feature selector on the three datasets always achieves similar or ideally up to 12.46% lower error rates on the average.

Among the set of regression methods, SVM unexceptionally exhibits the best performance by yielding the highest *R*'s and lowest *RMSE*'s; whereas MFANN-based models deliver the second highest *R*'s and lowest *RMSE*'s for prediction of VO_2max. There is no strict order between TB-based and GRNN-based prediction models, but the *RMSE*'s related to TB-based and GRNN-based prediction models are always higher than those of SVM-based and MFANN-based prediction models, and lower than DTF-based prediction models. DTF-based prediction models perform better than RBFNN-based prediction models, and

RBFNN-based prediction models, in turn, outperform SDT-based models which relatively show the worst performance for prediction of VO_2max.

As for the importance of the variables for VO_2max prediction; it turns out that the predictor variables gender, age, MX-HR, MX-Time, SM-ES, SM-HR3, SM-MIN3, Q-PFA and Q-AC have been found to have an improving effect for all utilized datasets and machine learning methods. The rest of predictor variables, on the other hand, can have an improving, negligible or deteriorating effect on prediction of VO_2max, depending on with which other variables they are combined to form the prediction model.

Key Words: Machine Learning, Feature Selection, Maximal Oxygen Uptake, Prediction

DEDICATION

... to my mom and dad

LIST OF PUBLICATIONS

This thesis is based mainly on the following publications of the author:

Journal Publications:

1) **F. Abut**, M.F. Akay and J. George, “Robust ensemble feature selection using rank aggregation for developing new accurate SVM-based VO_2max prediction models”. Turkish Journal of Electrical Engineering & Computer Sciences, vol. 27, no. 5, pp. 3648-3664, 2019.

2) **F. Abut**, M.F. Akay and J. George, “Developing new VO_2max prediction models from maximal, submaximal and questionnaire variables using support vector machines combined with feature selection”. Computers in Biology and Medicine, vol. 79, pp. 182-192, 2016.

3) **F. Abut** and M. F. Akay, "Machine Learning and Statistical Methods for the Prediction of Maximal Oxygen Uptake: Recent Advances". Medical Devices: Evidence and Research, vol. 8, pp. 369-379, 2015.

Conference Publications:

A. Full Research Papers

1) **F. Abut**, M. F. Akay and J. D. George, "Support Vector Machines with Individual Combinations of Relief-F and mRMR for Predicting VO_2max from Maximal and Questionnaire Data". In Proc. of Intl. Conference on Artificial Intelligence and Soft Computing (ICAISC-2016), Barcelona, Spain, 18-19 Aug. 2016, pp. 9-13.

2) **F. Abut**, M.F. Akay and J.D. George, “Development of New Hybrid Models for Prediction of Maximum Oxygen Uptake Using Machine Learning Methods Combined with Feature Selection”. In Proc. of Intl. Symposium on Sport Science, Engineering and Technology (ISSSET-2015), Istanbul, 10-13 May 2015, pp. 163-170.

B. Abstracts / Extended Abstracts

1) **F. Abut**, M. F. Akay and J. D. George, Maximum-Likelihood Feature Selection for prediction of Maximal Oxygen Uptake using Support Vector Machines. In Proc. of Intl. Conference on Educational Research (CYICER-2016), North Cyprus, 4-7 May 2017, p. 48.

2) **F. Abut**, M. F. Akay and J. George, "Minimum-Redundancy Maximum-Relevance Feature Selection for Creating New Hybrid Models for Predicting Maximal Oxygen Uptake Using Support Vector Machines". In Proc. of Intl. Symposium on Engineering, Artificial Intelligence & Applications (ISEAIA-2016), North Cyprus, 2-4 Nov. 2016, pp. 21-22.

3) **F. Abut**, M. F. Akay and J. George, "Data-driven Prediction of VO_2max based on Maximal and Submaximal Variables Using Support Vector Machines Combined with Feature Selection". In Proc. of Intl. Symposium on Engineering, Artificial Intelligence & Applications (ISEAIA-2015), North Cyprus, 4-6 Nov. 2015, pp. 10-11.