VO2 Max Predication Model Validation and Potential Improvements Report

-

Quintin Xu

IoT and Embedded Systems Team

Readback Operations

1. Overview

This report aims to briefly introduce the existing VO2 prediction ML models implemented by the previous Data Analytics (DA) team, provide some potential improvements based on the sensors' data that IoT and Embedded Systems Team is going to be acquired in this trimester and propose a plan to validate those existing ML models with the new data.

2. Existing Applied ML Models and Used Datasets

The previous DA team has been done with two different research on the VO2 prediction using different ML models.

2.1 Oxygen Uptake Prediction

The corresponding detailed research doc can be found here.

The dataset utilised by this research comes from a Kaggle data source, which includes time, power, oxygen (ml/min/kg), cadence, HR (Heart Rate – heart beats per min) and RF (Respiration Frequency – number of breaths per minute) data. The data was gathered from 7 recreational cyclists (6 males, 1 female) who were required to complete 3 different intensity level exercises with a bicycle ergometer (more details can be found in this doc).

Detail Compact Column 6 of 6 columns

# time	# Power	# Oxygen	# Cadence	# HR	# RF
1 – 2819	0 – 333 (133.20 - 166.50, Count: 702)	168 – 4.58k	0 – 90	72.4 – 189 (177.34 - 189.00, Count: 506)	11.6 – 61
1	0	318.4	0	75.6	20.1
2	0	356.1666667	0	75.66666667	19.75
3	0	403.2857143	0	75.71428571	19.42857143
4	0	456.25	0	75.75	19.125
5	0	478.9259259	0	75.74074074	18.96296296
6	0	480.4	0	75.7	18.9
7	0	480	0	75.7	18.6
8	0	482.7166667	0	75.69166667	18.375
9	0	488.55	0	75.675	18.225
10	0	497.5	0	75.65	18.15

Figure 1 – Kaggle dataset

Three different ML models have been applied on this dataset, such as Linear Regression (LR), MLPRegressor Neural Network (NN) and Decision Tree (DT).

The target variable is Oxygen, the rest of variables are used as feature variables. The best R2 scores for LR and DT models are **0.97 and 0.99**, respectively, and the best RMSE rate is **163** for MLP NN model.

2.2 Long Short-Term Memory (LSTM) NN Research

The corresponding research doc can be found here.

Except for the Kaggle dataset mentioned in 2.1 above, another 2 datasets from the University of Malaga and the University of Uruguay have also been used for this research. More dataset details can be found here. All of three datasets were acquired from different number of participants and different types of exercises, such as bicycle and running.

	time	Power	Oxygen	Cadence	HR	RF	Participant	Method
0	1.0	0.0	318.400000	0.0	75.600000	20.100000	1.0	0.0
1	2.0	0.0	356.166667	0.0	75.666667	19.750000	1.0	0.0
2	3.0	0.0	403.285714	0.0	75.714286	19.428571	1.0	0.0
3	4.0	0.0	456.250000	0.0	75.750000	19.125000	1.0	0.0
4	5.0	0.0	478.925926	0.0	75.740741	18.962963	1.0	0.0

	Time	RF	HR	Oxygen	Participant	Speed	Exercise
0	0.0	19.74	1.051	875.824024	1	0	0
1	0.06999999999999999	9.97	3.134	927.258849	1	0	0
2	0.1	24.00	0.550	480.327309	1	0	0
3	0.13999999999999999	23.17	0.685	637.904689	1	0	0
4	0.16	20.62	0.904	789.366586	1	0	0

	time	Speed	HR	VO2	VCO2	RR	VE	ID_test	ID
0	0	5.0	63.0	478.0	360.0	27	13.3	2_1	2
1	2	5.0	75.0	401.0	295.0	23	10.3	2_1	2
2	4	5.0	82.0	449.0	319.0	29	12.2	2_1	2
3	7	5.0	87.0	461.0	340.0	28	12.8	2_1	2
4	9	5.0	92.0	574.0	417.0	28	14.6	2_1	2

Figure 2 - Kaggle vs Uruguay vs Malaga datasets

All of three datasets have been merged into one big set for building the Deep Learning (DL) models under this research. Only 3 variables, such as Time, HR and RF, have been applied as feature variables, and the Oxygen Consumption has been applied as the target variable.

There are two different Deep Learning (DL) models Involved: LSTM NN and the linear feedforward NN. The best R2 score and RMSE were reported with **0.73** and **0.11**, respectively.

3. The Fresh Sensor Data We Will Acquire

In this trimester, the IoT and Embedded Systems team is working on two major projects:

smart bike and VO2 max measurement. The following projects-related sensors are about to be delivered to the team soon: Wahoo Speed and Cadence Sensor, Wahoo Heart Rate Sensor, IR11EM CO2 sensor, SGX-4OX Oxygen sensor and SEN0360 Air flow sensor.

These sensors will hopefully yield the sensor data of the followings: speed, cadence, power, HR, RF, Oxygen concentration, CO2 concentration. With these sensors data on hands, since they have covered all essential feature variables used in the previous trimester mentioned above, the team should be able to validate the existing ML models.

4. Potential Improvements for VO2 Predication

For the previous research works relating to VO2 prediction, the following areas can be potentially improved during the team's projects commencing this trimester:

- In the "Oxygen Uptake Prediction" research:
    - The used dataset size was too small: only **2819** samples were used for the work, and the other three sub datasets (ie. sbj_1_II, sbj_1_Wingate and sbj_1_incremental) with around **6000** more samples were not applied for building the ML models at all.
    - The low correlation variables like time, with only 0.1 – 0.3 correlation rate to Oxygen, should not be used for predicting VO2.
    - The RMSE value for the MLPRegressor model seems to be incorrect: a good RMSE metric value should be between **0.1 to 0.5**, the calculated best RMSE were **163** and **203,** which does not make sense on a way larger scale.
- In the "LSTM" research work:
    - It should not build a ML model upon 3 different datasets, even if they have some common feature variables.
        - The 3 different datasets have different scales, with **57983, 6239 and 575087** samples, respectively.
        - These datasets were measured from totally different exercises or protocols, the generated datasets could be vitally different.
    - The research work is also missing many metric checks, such as RMSE, MAE, MSE, and so on.
- Since the VO2 predication is about a regression problem, other popular ML models could also be considered to increase the odds of getting the best predication model, including Logistic Regression, Ridge Regression, Lasso Regression, Random Forest, KNN and Support Vector Machines (SVM).
- Providing that the team is going to gather sensors data this trimester, all of the following data variables should be collected and used for a VO2 ML predication model: speed, cadence, power, HR, RF, Oxygen concentration, CO2 concentration. A features importance/correlation analysis and a further feature selection process might be necessary.
- There are many other factors could be affecting a person's VO2 max, including age, gender, genetics/physiology, altitude, body type/body composition, training status and exercise type (https://www.intelligent-triathlon-training.com/articles/vo2max). To get a more accurate model, the data measurements around these factors might be necessary.
- The final goal is to get accurate **VO2 max** values. After the VO2 data is collected from the "VO2 Max measurement" project and the VO2 predication model is built, it might need some further calculations to reveal the **VO2 max** values for participants

instead of merely O2 concentration values under a certain scenario.

5. Validate the Existing Models and Come Up with the Best Performed Model

To validate the existing ML models built from the previous term and come up with the best performed model, the following steps can be taken:

1) Collect sufficient sensors data from different volunteer cyclists. At minimum, Time, Power, Oxygen, Cadence, HR and RF need to be collected for the existing models' validation.
2) Execute the exact ML models' notebooks with the exact applied datasets to build for the existing ML models from the previous team.
3) Verify the newly produced metrics data matches the metrics presented before in the research docs.
4) Import the new sensors' datasets and predict VO2 using the existing built ML models.
5) Use the same metrics check codes to verify the existing ML models' performance.
6) Re-build the existing ML models and predict by the new sensors' datasets only, record the metrics and compare them with the previous performance data.
7) Re-build, predict and performance check the existing ML models with the extra sensor data variables, except for Time, Power, Oxygen, Cadence, HR and RF.
8) Build, predict and performance check some new ML models with the newly collected sensors data.
9) Compare metrics between different approaches, and conclude the best performed ML model.