**Dirty jobs: the reality of working with wearable activity tracker data**

Sherman MD.

Department of Computational Medicine and Bioinformatics

University of Michigan

**Abstract**

Obesity is a disease that is a major contributor cardiac disease, type 2 diabetes, and general health and well-being. Therefore, it is important for us to understand how to successfully battle the disease. Using only triaxial armband accelerometer data collected by an ongoing comprehensive multi-phase weight-loss study, this study attempted to build a machine learning classifier that could predict the success of consistent weight-loss. Best practices of cross-validation and data partitioning were implemented to optimize and train a series of machine learning models (Gradient boosted random forest, random forest, support vector machine, multi-layer perceptron, K-nearest neighbor, & logistic regression) on labeled data. The outcome produce highly unstable classifiers that could not reliably predict better than chance (50%). The instability is in part due to lack of a large enough dataset and in part due to the nature of the disease itself. This study shows that consistent weight-loss is contingent on more than just physical activity alone.

**Introduction**

While obesity continues to be an epidemic in the United States, little research has shown exactly the components that play into a subject’s ability or capacity to lose weight. As individuals increase their physical activity; their cardiometabolic risk profiles improve, they become more resistance to non-communicable diseases, they are less likely to succumb to type II diabetes, and generally live longer lives4. To properly test the influence of physical activity regarding weight loss, large-scale studies generally rely on the self-reporting of the participants. Herein lies a bias regarding the disease.

As self-reporting tends to be more reliable when physical activity is more vigorous, light activity is often poorly reported4. However, participants in the study are unlikely to undergo continuous vigorous workout regimens (e.g. High intensity interval training, HIIT) regularly as this leads to early burnout. Successful weight-loss often starts with appropriate scaling of physical activity. This often means starting with light physical activity will both stimulate weight-loss and support broader reaching lifestyle changes.

To objectively and accurately capture data on light activity levels and sedentary behavior, accelerometers are often used. An accelerometer is a small, lightweight device that can be worn on several different positions of the body. Conceptually, all accelerometers use an axial spring mass mechanism5. As acceleration occurs on a given axis, the force of the acceleration is applied to a mass attached to a “spring”. The spring force is then used as a proxy of acceleration. Using a triaxial (vertical, mediolateral, anteroposterior) accelerometer, clinicians can monitor the physical activity level of subject without direct observation, little cost, and little discomfort to the subject. Furthermore, since the device is designed to be wearable over time, accelerometers can capture both the intensity of movement and frequency of activity the subject5.

There are some caveats to using such devices though. The data is inherently noisy due to vibrations of soft-tissue (muscle twitching or leg shaking), external vibration (such as riding in a car), and subject to bias due to device placement. If the device is worn on the waist, it might not capture the intensity of heel strikes or upper body movement. If worn on the ankle, the acceleration of the ankle during a run may disproportionately influence the readings. The same can be said with the wrist and arm. Furthermore, accelerometers cannot discern differences in acceleration due to changes in posture or load (weight) working upon the subject4.

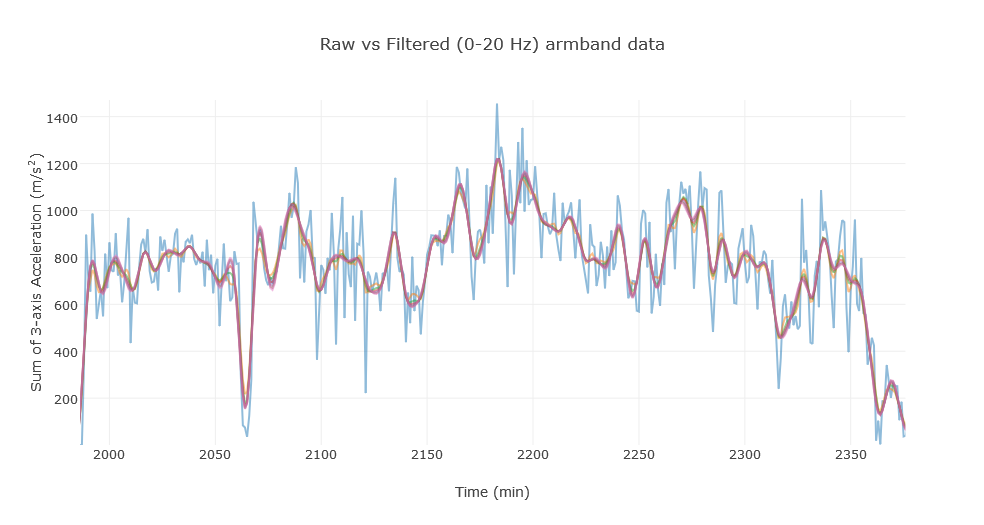
Since the data collected from accelerometers is continuous, time-domain measurements, it stands to reason that contemporary signal processing and machine learning can be applied to discern the differences between successful weight-loss study participants and failures. Furthermore, identified distinctions can then be used to supplement current federal guidelines for physical activity regarding health and fitness and shore the gap of knowledge with respect to light physical activity.

**Methods**

The data used in this study comes from triaxial armband accelerometer data of 100 samples in an ongoing multi-phase weight-lost study. Each sample has 7000 data points representing one reading per minute over the course of five days. 21 of the subjects were labeled as successes after phase 2 of the study. A success label was derived from both reaching a specific weight-loss percentage threshold in phase 1 as well as reaching a similar threshold in phase 2. For the purposes of this research, we only considered the armband data to determine the dependence of physical activity on weight-loss.

A low-pass third order Butterworth filter was applied to the accelerometer data with a cutoff at 20 Hz (Fig. 1). While most movements are between 0 and 3.5 Hz5, the arm acts as a mass damper to minimize head pitch in the sagittal plane, and therefore counteracts the torque applied to hips by the legs. This means that as leg movement increases, so does arm movement. Therefore, arm movement acts as a proxy measure of whole body activity. To adequately capture the occurrence of HIIT training or running, the upper limit of the filter was increased7.

**Fig. 1**: Example of filtered accelerometer data using a low-pass third-order Butterworth filter

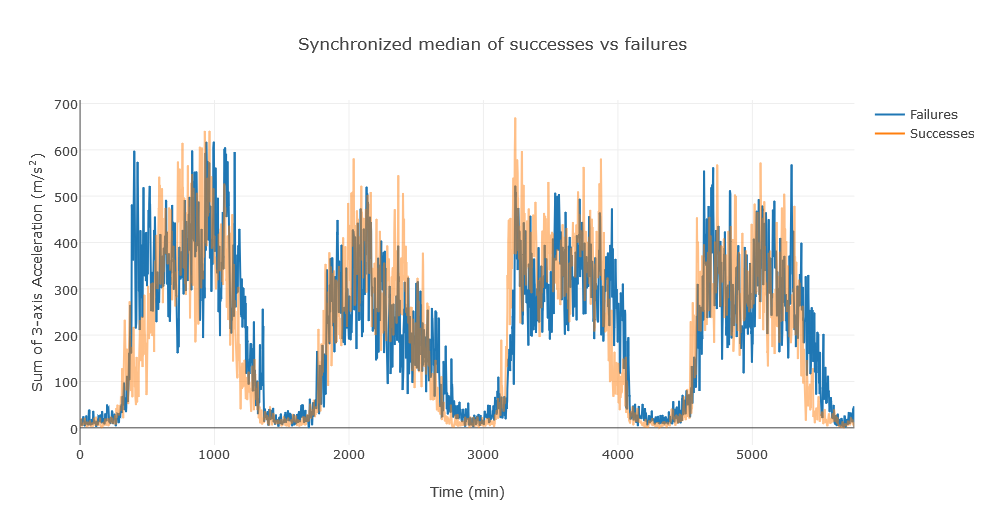


The data encompasses five days of free-living subject activity level, each of the samples were synchronized to a complete 4-day cycle (i.e. coupled sleep-wake rhythm). Since the data was unbalanced between failures and successes, an even data set was produced by taking out all 21 successes and randomly sampling 21 of the failures. The new dataset was 42 samples, half success and half failure. When both group’s median readings were plotted, it appeared as if successes were more active during the day and less active during the night (Fig. 2).

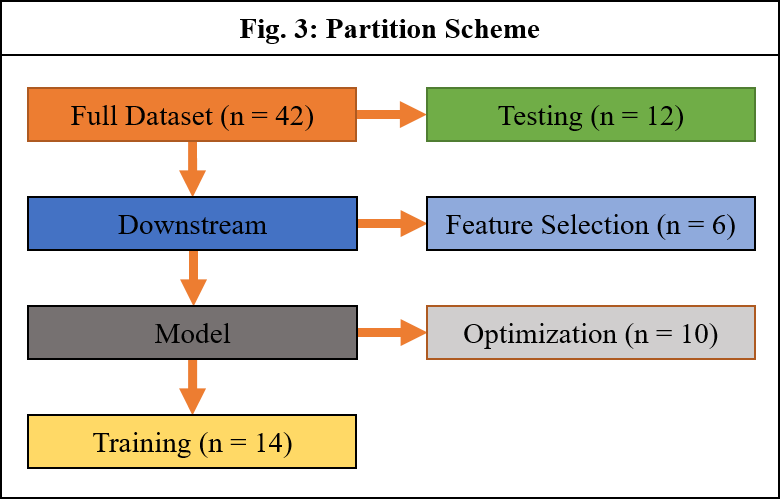
|  |  |
| --- | --- |
| **Table 1** | |
| **Domain** | **Features** |
| Time | Mean, median, variance, skewness, kurtosis |
|  | Max |
|  | Δ max - min |
|  | Autocorrelation |
|  | RMS |
| Frequency | RMS |
|  | Power |
|  | Mean Frequency |
|  | Power Spectral Density |
|  | Spectral Flux |
|  | Spectral Roll-off |
|  | Spectral Centroid |

Table 1 shows the features that were generated for the complete signal, each full day, and each active and inactive period observed in the data. Furthermore, the differences between each of the days (to detect trend of activity) and sleep vs wake period were calculated. In total, 336 features were generated.

As overfitting is always an issue with any machine learning algorithm, the dataset was partitioned for each of the following steps in a stratified fashion: Feature selection, model optimization, model training, and model testing. Fig. 3 shows the partition workflow for the project.

 Feature selection was done by selecting the 12 best features based on mutual information score6 and by recursive feature elimination & K-fold cross validation (RFECV)2. RFECV uses iteratively trains a given model, optimizing the weights a given feature has on the classification based on its receiver operator curve area under the curve score (ROC-AUC), and recursively eliminates the feature with the lowest rank until the smallest number of features exist with the highest accuracy the model can achieve is left. Ultimately, K-best feature selection based on mutual information was used for all other downstream analysis. See discussion for more.

**Fig. 2**: Synchronized median filtered accelerometer values of failures (blue) vs successes (orange)

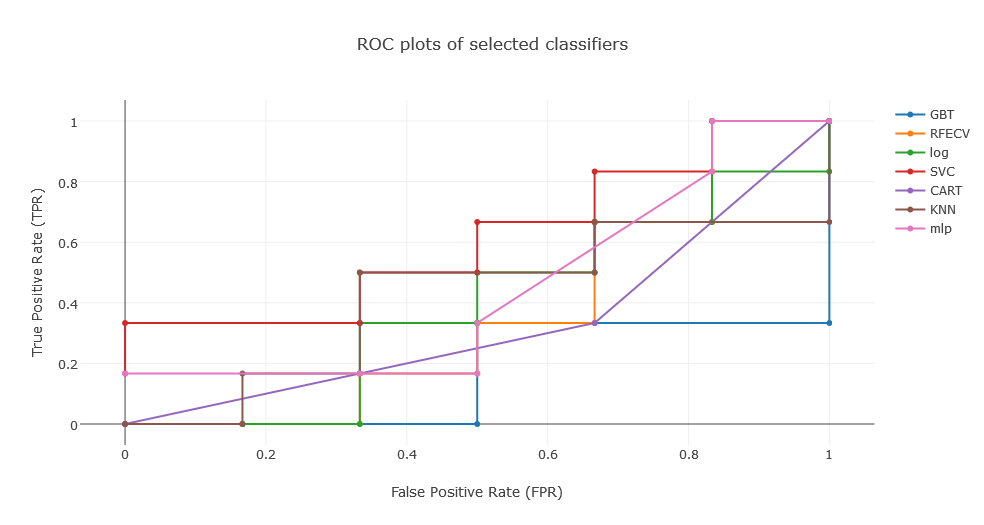


Model optimization was performed by 5-fold cross-validated grid search of optimum hyperparameters (e.g. C value, max number of iterations, number of estimators, number of hidden layers, etc.). The hyperparameters were tuned for each of the following classifiers: Gradient boosted decision tree (GBT), random forest, support vector machine (SVC), logistic regression, K-nearest neighbor, and a multi-layer perceptron (MLP).

Training of the optimized classifiers was then completed on separate data. Finally, test data was applied to each model and their respective ROC was plotted.

**Results**

Using a small, unbalanced, and noisy dataset of triaxial accelerometer data spanning five days, machine learning models were optimized and trained to classify samples as success or failure. After stratified cross-validation over many iterations, all the classifiers used were unable to satisfactorily classify the samples. Despite optimization, all the models were highly unstable. This resulted in classification being at chance or worse. A sample of the ROC plots can be seen in Fig. 4. Ultimately, no single classifier could effectively predict success of weight-loss based on armband data alone.

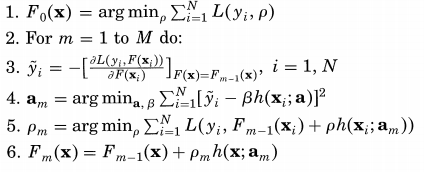


**Fig. 4**: ROC plot for selected classifiers. GBT (blue), RFECV (orange), logistic regression (green), SVC (red), Random Forest (purple), KNN (burgundy), and MPL (pink)

**Discussion**

GBT weights weak learners (random decision trees that have classification accuracy between 0.5 < x < ~0.9) based on gradient descent. The algorithm can be seen in Fig. 51. It was believed that due to the small differences in the data, a series of weak learners would be the best approach to classify successes and failures based on accelerometer data. The scikit-learn implementation of the GBT comes out-of-the-box ready for regularization, subsampling, and loss function selection—all of which protect against overfitting6. However, GBT requires a lot of samples for it to become sufficient in classification. This was something overlooked early on in model selection.

**Fig. 5**: Gradient Boosting algorithm6



RFECV also provided no reprieve despite its promise. The number of features required for RFECV to adequately optimize on greatly outnumbered the number of samples. To protect from overfitting. K-best feature selection was used in its place.

Despite correcting for sample balance and conservative protection against overfitting, no other machine learning model in the scikit-learn arsenal could consistently predict above chance (50%).

**Conclusion**

The classification of success or failure of weight-loss based on triaxial accelerometer data could not be sufficiently done with any of the classifiers in the scikit-learn library. There are many potential causes for this result. The first is the sheer lack of sample data. Ideally, more data is always better. In the case of this study, the number important/informative easily outnumbered the number of samples. By reducing the feature space to protect against overfitting, a comprehensive and informative feature space could not be compiled.

The second issue lies in the features themselves. There are potentially better features that were not considered in this study that may have been more informative to the models.

The third issue is the nature of the data itself. Success at weight-loss is more than just exercise. Genetics, diet, and environment all play a major role in weight-loss. It stands to reason that even if the best features were selected, and hundreds of thousands of samples were collected, one’s ability to lose weight may require more than just physical activity. For us to accurately predict ability to lose weight might require additional multi -*omic* dataset, diet & intake information, heredity & lineage samples, and health & psychological records.

**Future Works**

Additional physiological sensors that measure heart rate, respiration, and muscle contraction may also add another layer of predictive power. Providing privileged information such as psychological records and diet information could help provide a more comprehensive picture of the subject’s progress. Furthermore, a genome-wide association study of the successes may identify some potential biomarkers that could edify the genetic background of what supports weight-loss or genetic changes that occur during productive a weight-loss regimen.

**Supplementary Data**

Both the Jupyter notebook and raw data can be found at this [link](https://github.com/betteridiot/Machine_learning.git) or this address:

<https://github.com/betteridiot/Machine_learning.git>

**Work Cited**

1. Friedman JH. (2001). Greedy Function Approximation: A Gradient Boosting Machine. *Annals of Statistics,* 29(5): 1189-1232.

2. Guyon I, Weston J, Barnhill S, Vapnik V. (2002). Gene selection for cancer classification using support vector machines. *Machine Learning*, 46: 389–422.

3. He K, Zhang X, Ren S, Sun J. (2015) Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. ICCV ‘15:1026-1034. doi:10.1109/ICCV.2015.123

4. Lee I, Shiroma EJ. (2014). Using accelerometers to measure physical activity in large-scale epidemiological studies: issues and challenges. *Br J Sports Med.* 48:197–201. doi:10.1136/bjsports-2013-093154

5. Mathie MJ, Coster ACF, Lovell NH, Celler BG. (2004). Accelerometry: providing an integrated, practical method for long-term, ambulatory monitoring of human movement. *Physiol. Meas*. 25: R1–R20. doi: 10.1088/0967-3334/25/2/R01

6. Pedregosa et al. (2011). Scikit-learn: Machine Learning in Python. *JMLR* 12: 2825-2830.

7. Pontzer H, Holloway III JH, Raichlen DA, and Lieberman DE. (2009). Control and function of arm swing in human walking and running. *Journal of Experimental Biology* 212: 523-534. doi:10.1242/jeb.024927

8. Ross BC. (2014). Mutual information between discrete and continuous data sets. *PLoS ONE* 9(2): e87357. doi:10.1371/journal.pone.0087357

9. Yu HF, Huang FL, Lin CJ. (2011). Dual coordinate descent methods for logistic regression and maximum entropy models. *Mach Learn*. 85: 41. doi:10.1007/s10994-010-5221-8