Title: Optimistic prediction rates for human activity is possible using the XYZ coordinates of several Accelerometer signals.

Introduction:

Is it possible to predict human movement? Researchers at the University of California's (UCI) Center for Machine Learning and Intelligent Systems [1] believe so. They have established the UCI Machine Learning Repository, which is a collection of databases [2] such as the Human Activity Recognition (HAR) project. Thirty volunteers between the ages of 19-48 years were recorded as they performed six activities (Walking, Walking-Upstairs, Walking-Downstairs, Sitting, Standing, and Laying) of daily living while carrying a smartphone with embedded sensors [3]. The HAR project leveraged the embedded accelerometer and gyroscope components of the smartphone device to captured sensor signals associated with body motion. Each observation of a volunteer's movement produced 563 unique measurements.

We were interested in exploring how accurate human motion estimates could be using only a few acceleration signals from the HAR database. Using exploratory analysis and statistical modeling we showed that a prediction model based on the Random Forest algorithm [4] could optimistically classify human activity.

Methods:

Data Collection

For our analysis we used a subset of motion observations from 21 of the 30 subjects studied in the Human Activity Recognition project. The data set was downloaded from Amazon EC2/S3 on 25 February 2013 using the R programming language [5].

Exploratory Analysis

Our dataset was comprised of 7352 observations each containing 563 variables. We applied the principals of ETL [6] and Tidy Data [7] to prepare our dataset for analysis. While the dataset meet most of our Tidy Data checks, including complete observations, we did have to perform some data munging tasks.

To preserve variable readability while avoiding name collisions and usability issues within the R programming language, we renamed each of the 563 variables to eliminate embedded parentheses and to append a unique identifier.

We appended a new variable called `activityFactor` to the dataset since the `activity` variable was not of type char-vector and we needed to use this variable in several statistical formulas like `tree()` and `randomForest()` where a factor is required.

Statistical Modeling

Using the Random Forest algorithm, our goal was to train a prediction model that could classify human motion based on several significant variables. A random forest is a highly accurate classification model consisting of a collection of tree structured classifiers {h(x,Θk), k=1, ...} where the {Θk} are independent identically distributed random vectors and each tree casts a unit vote for the most popular class at input x [8].

Reproducibility

This report represents a synthesized version of our research activity. Our analysis was performed using an R markdown file. The artifacts of this analysis have been organized into a reproducible research package, which is available upon request. Alternatively, a complete transcript of our data analysis workflow process is publically available [9].

Results:

Data Selection

There has been significant research done in the area of human body orientation estimations using accelerometers and gyroscopes. Specifically, the Kalman Filter was designed to use the three signals (x, y and z coordinates) derived from an accelerometer for both body and gravity to estimate human movement [10]. We decided to narrow our focus of interest form 563 variables down to six (6) accelerometer measurements. We selected six (6) variables associated with the XYZ Coordinates for mean body and gravity acceleration signals. Specifically, we established a new dataset that included activity outcome observations for `xMeanBodyAcc`, `yMeanBodyAcc`, `zMeanBodyAcc`, `xMeanGravityAcc`, `yMeanGravityAcc`, and z`MeanGravityAcc`. We ran a linear regression model on all 563 variables; these six variables seemed to be insignificant coefficients. Yet when we isolated them in their own linear regression model we observed statistical significance in the confidence of a relationship (`p-value`) for each of the six variables. Our premise was that we could achieve a fair degree of prediction accuracy just with these six variables.

Acceptable Error Rates

One of the issues associated with the training of prediction models is over-fitting. As we add complexity to our model to help it fit our training data, we will pass a threshold whereby the extra complexity causes the model to do a worse job of predicting new data. This is a case of over-fitting the training data.

While we want to avoid over-fitting we also want to strive for consistent accuracy in our prediction model. We desired to find a balance between accuracy and acceptable error rates. As such, we decided to employ an optimism scale for our endeavor.

True Prediction Error = Training Error + Training Optimism

Training Optimism is basically a measure of how much worse our model does on new data compared to the training data. The more optimistic we are, the better our training error will be compared to what the true error is and the worse our training error will be as an approximation of the true error [11]. We decided to use an optimism scale whereby we defined an acceptable level of optimism for our prediction model:

Optimism = True Prediction Error - Training Prediction Error

True Prediction Error is the performance measure of our model on the test population and Training Prediction Error pertains to our models performance on the training data. For our project we desired to produce an optimistic prediction model where:

Optimistic: ← Optimism < 1.5
Highly Optimistic ← Optimism >= 1.5

Prediction Model Benchmark

Our probability benchmark for our data was based on an approximation equal to (0.5) ^ size-of-sample-set. As such, we understood that we had 0% probability for a perfect classification! Yet we established parameters for acceptable results.

Environment Preparation

Given a corpus of 21 subjects, we created five (5) groups to be used in the training and testing of our prediction model. Several factors were considered when creating these groups. First, the sponsors of this project pre-selected the populations of the two baseline groups. Second, our exploratory analysis revealed that the distribution of activities across the entire corpus of 21 subjects was unevenly distributed, namely Min(walkdown)=986 and Max(laying)=1407. This was compounded by the fact that there was also an imbalance in the range of activity tests per subject, namely Min(subject#8)=281, Max(subject#25)=409.

Data Set	Subjects	Observations	Purpose
Training Baseline	1, 3, 5, 6	1315	Minimum set of subjects to use for training model.
Training Expanded	1, 3, 5, 6, 14, 15, 16, 17, 19, 22, 23	3753	Broadest set of training subjects possible while avoiding overlap.
Training Test	7, 8, 11, 25	1314	The subjects we used for testing our model during training.
Validation Baseline	27, 28, 29, 30	1485	Minimum set of subjects used to validate accuracy of the model.
Validation Expanded	21, 27, 28, 29, 30	1893	Maximum set of subjects used to validate accuracy of the model.

Table 1: Data Segmentation

Model Training

We ran nine (9) training experiments. For each experiment we tweaked one or more aspects of a random forest prediction model. We captured the results of each experiment, namely error rate and accuracy per activity class in a data frame so that we could plot the aggregate of all experiments upon completion. For eight (8) of the tests, we used the following formula:

activity ~ xMeanBodyAcc + yMeanBodyAcc + zMeanBodyAcc + xMeanGravityAcc + yMeanGravityAcc + zMeanGravityAcc

Experiment	Training Data	Test Data	Tuning Attributes	Results
T1	Training Baseline	Training Test	Trees = 500 Splits = 3	Error Rate: 8.06% Accuracy Range: 15% - 100%
T2	Training Baseline	Training Test	Trees = 1000 Splits = 3	Error Rate: 7.91% Accuracy Range: 15% - 100%
T3	Training Baseline	Training Test	Trees = 500 Splits = 4	Error Rate: 7.83% Accuracy Range: 15% - 100%
T4	Training Baseline	Training Test	Trees = 1000 Splits = 4	Error Rate: 7.68% Accuracy Range: 15% - 100%
T5	Training Expanded	Training Test	Trees = 500 Splits = 4	Error Rate: 9.43% Accuracy Range: 37% - 100%
T6	Training Expanded	Training Test	Trees = 1000 Splits = 4	Error Rate: 9.46% Accuracy Range: 37% - 100%
T7	Training Expanded	Training Test	Trees = 1000 Splits = 2 Used only z coefficients in formula	Error Rate: 49.05% Accuracy Range: 10% - 63%
T8	Training Down Sampled Expanded	Training Test	Trees = 1000 Splits = 4 Resampled Size = 3012	Error Rate: 10.46% Accuracy Range: 37% - 100% but 5 of 6 above 50%
T9	Training Up Sampled Expanded	Training Test	Trees = 1000 Splits = 4 Resampled Size = 4272	Error Rate: 6.67% Accuracy Range: 37% - 100% but 5 of 6 above 50%

Table 2: Training Experiments

As depicted by Table 2, in order to achieve results beyond those observed in experiment T6, we needed to address the imbalance in our expanded training data. We observed a disparity in observations per activity between 712 (laying) and 544 (walkup) with a mean of 625 for all activities. We decided to resample our expanded training group to improve the distribution of

classes across observations. The net benefit here was that up-sampling yielded a greater quantity of balanced training data.

Model Testing

We ran two (2) validation experiments using the model used in training experiment T9. For each experiment we captured the results and appended them to the data frame we used during training.

Experiment	Training Data	Test Data	Results
V1	Training Up Sampled Expanded	Validation Baseline	Error Rate: 6.67% Accuracy Range: 40% - 100%
V2	Training Up Sampled Expanded	Validation Expanded	Error Rate: 6.67% Accuracy Range: 40% - 100%

Table 3: Validation Experiments

Observations

Our model demonstrated consistent accuracy (40% or more per activity class) when we tested it on our validation groups. We can infer from this data that body and gravity motion signals from an accelerometer are good coefficients for predicting human movement (see Figure 1 Panel A). Furthermore, our prediction model exceeded expectations with respect to our optimism scale for error rate tolerance. As depicted by Figure 1 Panel B, our error rate results (6.67%) on the validation groups were well within our parameters for acceptable optimism.

Conclusions:

Using exploratory analysis and statistical modeling we developed a prediction model based on the Random Forest classification algorithm and the XYZ coordinate measurements for body and gravity acceleration from an accelerometer. We showed that such a model could be used to classify human activity. We were able to achieve >40% classification accuracy while maintaining acceptable error rates of 6.67%.

While our results were promising we must remember that we did not have access to the entire corpus of 30 subjects. Ideally, data from a larger population would have been better for training our model.

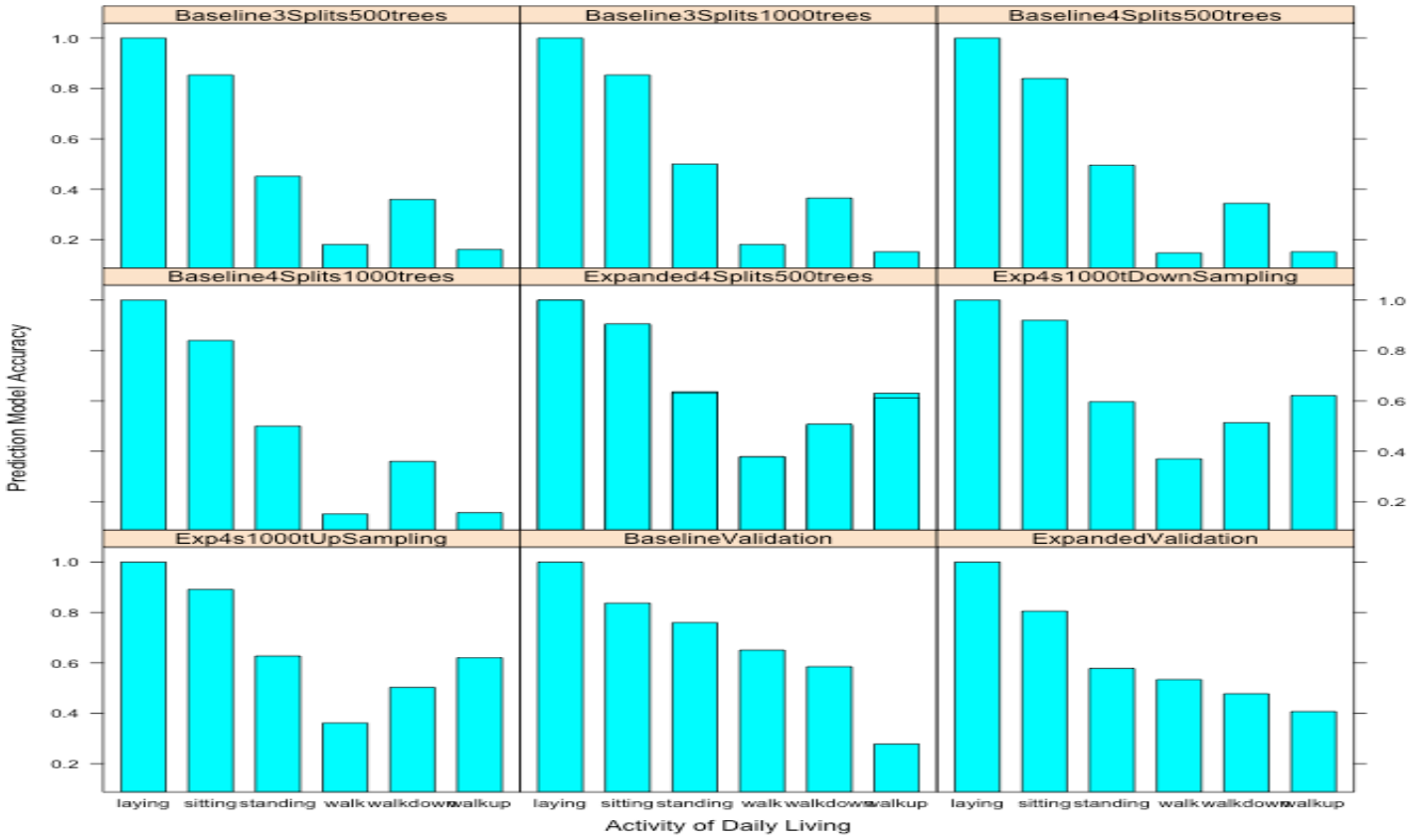
After up-sampling our training data, our prediction model maintained consistent accuracy above 50% for 80% of the classes and it never yielded an accuracy level below 40% for any class. We avoided over-fitting while maintaining error rates that were acceptable based on our optimism scale. By using random forests we able to identify outliers during our training and adjust the model accordingly. Our ability to interpret the results of our model is tied to the complexity of the model. Since we narrowed our variables of interest from 563 down to 6 we allowed for better interpretability.

Finally, the computational speed of our prediction model was not a concern since the Random Forest algorithm is highly optimized and we narrowed our variables of interest. We were able to compute all experiments on a single PC within minutes.

References

1. “Center for Machine Learning and Intelligent Systems”. URL: http://cml.ics.uci.edu. Accessed 2/25/2013.
2. “UCI Machine Learning Repository”. URL: http://archive.ics.uci.edu/ml/about.html. Accessed 2/25/2013.
3. “The Human Activity Recognition Database”. URL: http://http://archive.ics.uci.edu/ml/datasets/Human+Activity+Recognition+Using+Smartphones. Accessed 2/25/2013.
4. Wikipedia “Random Forest” Page. URL: http://en.wikipedia.org/wiki/Random_forest. Accessed 3/3/2013.
5. R Core Team (2012). ”R: A language and environment for statistical computing.” URL: http://www.r-project.org. Accessed 3/9/2013.
6. Wikipedia “ETL” Page. URL: http://en.wikipedia.org/wiki/Extract,_transform,_load. Accessed 2/13/2013.
7. “Tidy Data and Tidy Tools”. URL: http://vita.had.co.nz/papers/tidy-data-pres.pdf. Accessed 3/9/2013.
8. “Random Forests”, Leo Breiman, January 2001. URL: http://oz.berkeley.edu/users/breiman/randomforest2001.pdf. Accessed 3/9/2013.
9. “Data Analysis Workflow Navigation: DAP2 Markdown Results”. URL: https://github.com/vinomaster/dawn/blob/master/guides/rant/sample/dap2/dap2-guide.pdf. Published 3/9/2013.
10. “Estimating Human Movement Using a Three Axis Accelerometer”. URL: http://cope-et-al.com/wp-content/uploads/2009/03/ericcopeqe2007.pdf. Accessed 3/5/2013.
11. “Accurately Measuring Model prediction Error”. URL: http://scott.fortmann-roe.com/docs/MeasuringError.html. Accessed: 3/4/2013.

Panel A



Panel B

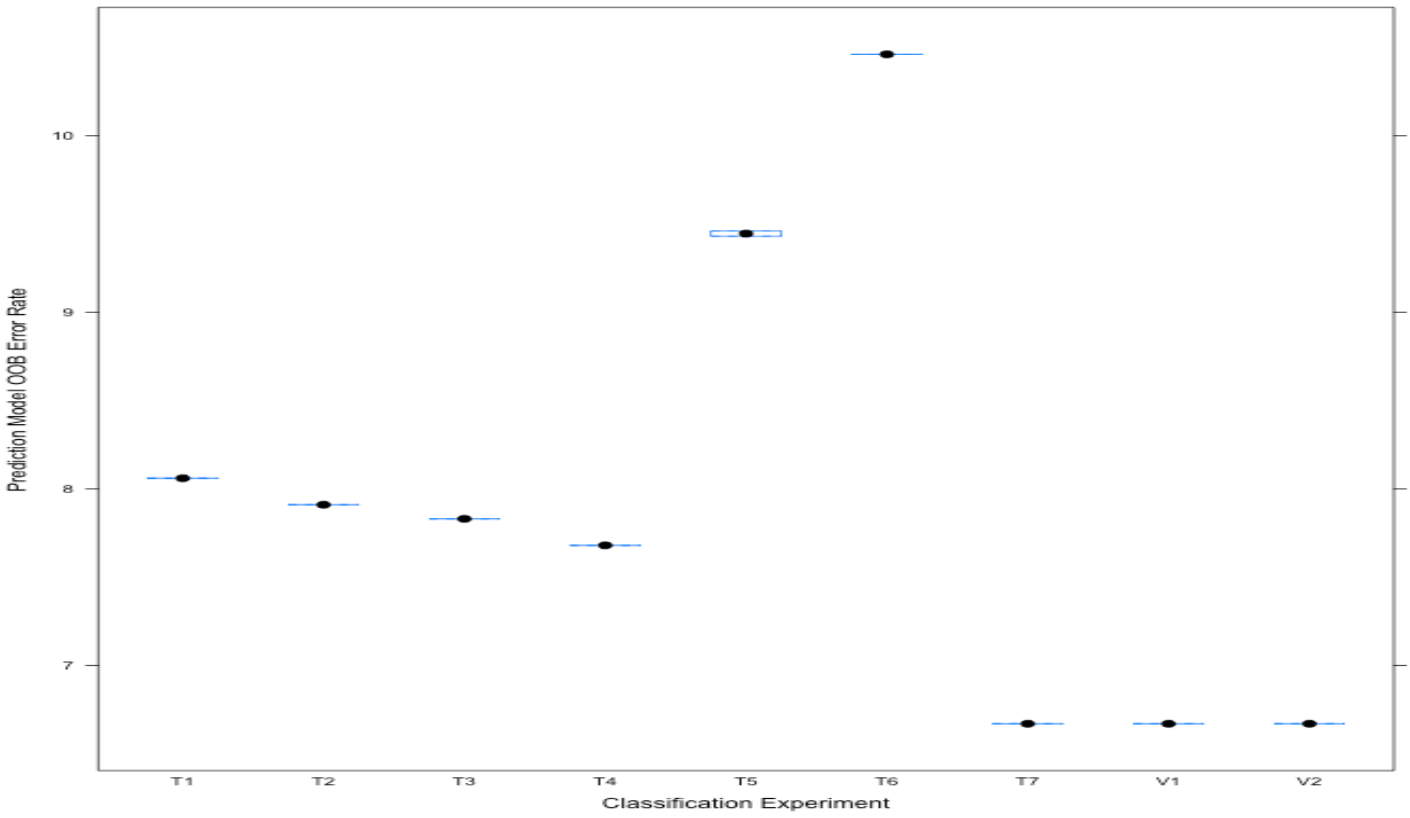


Figure 1: (Panel A): A barplot comparing the results collected from nine experiments. The first seven experiments were training exercises to tune the prediction model. These experiments explored the impact variables such as the number of trees and splits per tree have on the Random Forest classification algorithm. We also compared what impact up and down sampling would have on our model. Our training included testing a baseline set of observations from a preselected group of 4 subjects as well as an expanded set of 11 subjects. The last two experiments pertain to our validation of the prediction model on our test populations. Our validation populations included a predefined baseline that had 4 subjects and a expanded group of 5 subjects. We maintained uniqueness across all training and validation groups. Our model demonstrated consistent accuracy (above 40% per activity class) when we tested it on our validation groups. We can infer from this data that Body and Gravity motion signals from an Accelerometer are good coefficients for predicting human movement. **(Panel B):** We collected OOB Error Rates for all our experiments. Our training experiments (T1-T7), yielded varying error rates as we tuned the prediction model. We experienced an increase in our error rates as we began to use our expanded training group and subsequently up-sampled the data to address an imbalance in the distribution of observations across activity classes. When we applied down-sampling to the expanded training set, we observed a huge improvement in our error rate (+10% down to 6.67%). When using our prediction model with the validation experiments (V1-V2), we observed no change in the error rate (6.67%) This was a very positive outcome as we anticipated an optimism factor of at least 1.5% for our model.