Title: Predicting a subject's activity based on Samsung phone data

Introduction:

When used together in a smartphone, a gyroscope and an accelerometer provide a six-axis interpretation of movement through space. An accelerometer is used to measure sudden increases in speed within a certain range of motion and can sense the orientation of the phone. This technology is often used for application interactions on the smartphone, for example, to switch the screen display from portrait to landscape. A gyroscope is used to interpret the shift in positioning from a set rate of rotation within the X, Y, and Z axis (left/right, up/down, forward/backward). When a smartphone is tilted toward the sky the gyroscope is able to compare this movement to its normal state and calculate the change. When a gyroscope and accelerometer are combined, it is possible to create an accurate measurement of movement and location through space by providing constant, cross-referenced measurements of spatial placement and acceleration.[1] The challenge then becomes, can a model be built using this type of data that accurately predicts which of the 6 activities a subject performs – standing, sitting, laying down, walking, and walking vertically up or down.

Data Used:

The dataset use for this analysis was the Human Activity Recognition Using Smartphones Dataset.[2] The database was built from recording the data produced by the Samsung Galaxy S II smartphone worn on the waist of 30 subjects while they performed normal daily activities, grouped into 6 categories: standing, sitting, laying, walking, and walking vertically up or down (such as a flight of stairs). The data were gathered from the smartphones' embedded accelerometer and gyroscope and then post-processed based on video-records of the subjects, used to label the activities manually.

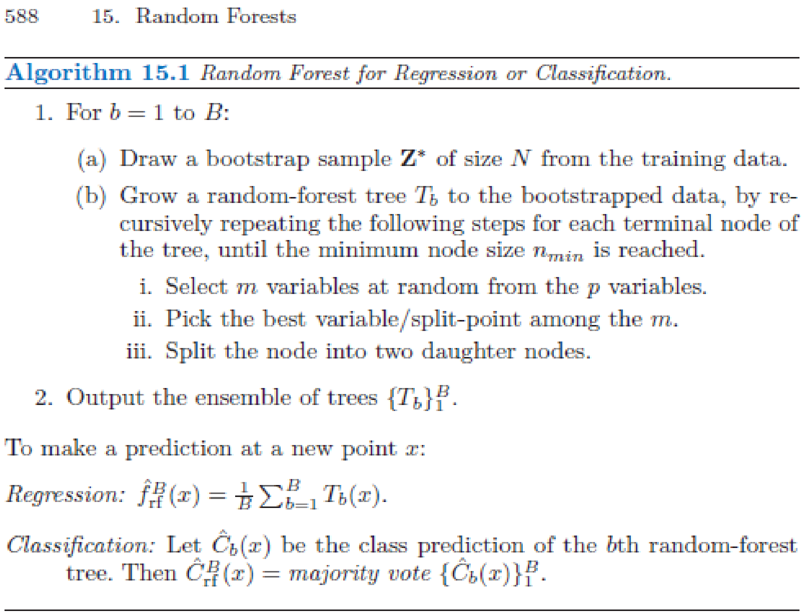
The 3-axial linear acceleration and 3-axial angular velocity signals were captured, and with pre- and post-processing, used to create the quantitative dataset. For each activity recorded for each subject in the experiment, a broad range of data was supplied, including a 561-feature vector with time and frequency domain variables, which was a vector of variables with values ranging from +1 to -1, representing the left/right, up/down, forward/backward positioning. The acceleration signal was separated into body and gravity acceleration signals, then the body linear acceleration and angular velocity were derived in time to obtain jerk signals and the magnitude of these three-dimensional signals, and a set of factors to indicate frequency domain signals were included. There were no missing values in the dataset.[2]

This highly technical and very robust dataset was used to build a model that predicts which of the 6 activities a subject was performing (standing, sitting, laying down, walking, and walking vertically up or down).

Opting for a Random Forest™ Model:

Expert opinions were reviewed regarding the best type of model to use when there are a large number of variables and the data are complex. It was clear that Random Forest models are quite usefulness in such circumstances. According to Togaware Pty, a Random Forest model is an ensemble of un-pruned decision trees which are often used when there are a very large number of input variables (hundreds or more) – there were 561 in the dataset used. The algorithm is efficient with respect to a large number of variables since it repeatedly subsets the variables available.[3]

Further, *The Elements of Statistical Learning* states: The idea in Random Forests (see Algorithm 15.1 below) is to improve the variance reduction of bagging by reducing the correlation between the trees, without increasing the variance too much. This is achieved in the tree-growing process through random selection of the input variables. Specifically, when growing a tree on a bootstrapped dataset, before each split, randomly select the input variables as candidates for splitting.[4]



588   15. Random Forests

Algorithm 15.1 *Random Forest for Regression or Classification.*

```
1. For $b = 1$ to $B$:
    (a) Draw a bootstrap sample $\mathbf{Z}^*$ of size $N$ from the training data.
    (b) Grow a random-forest tree $T_b$ to the bootstrapped data, by recursively repeating the following steps for each terminal node of the tree, until the minimum node size $n_{min}$ is reached.
        i. Select $m$ variables at random from the $p$ variables.
        ii. Pick the best variable/split-point among the $m$.
        iii. Split the node into two daughter nodes.
2. Output the ensemble of trees $\{T_b\}_1^B$.
```

To make a prediction at a new point x:

Regression: $\hat{f}_{\mathrm{rf}}^B(x) = \frac{1}{B}\sum_{b=1}^B T_b(x)$.

Classification: Let $\hat{C}_b(x)$ be the class prediction of the bth random-forest tree. Then $\hat{C}_{\mathrm{rf}}^B(x) = \textit{majority vote}\ \{\hat{C}_b(x)\}_1^B$.

Since Random Forest[5] models are well suited for dataset like the Human Activity Recognition Using Smartphones Dataset, it was determined that a Random Forest model would be used to build a model that predicts what activity a subject is performing based on the quantitative measurements from the Samsung phone. Since the target/outcome variable (activity) was given and was categorical, a Supervised Classification type of Random Forest model was used. The model training and validation were performed using the R package *randomForest*[6] while utilizing the interactive user interface R package *Rattle*[3].

Preparing the Model:

The Testing dataset for the analysis was defined as the observations associated with the last 4 subjects (subject ids 27 through 30) which contained 1,485 observations that were used to test the model. These records were split from the full dataset and set aside. The remaining data (subject ids 1 through 26) which contained 5,867 observations that were used to build and validate the model were split into three subgroups used for Training (70%), Validation (15%) and the final 15% was used as a "preliminary" Testing dataset. When the model was evaluated, the default evaluation dataset was the Validation data. The same seed (42) was used each time the data was partitioned.

All of the 561 features were explored and measured for correlation, but ultimately all were included in the Random Forest models. Reducing the number of features included in the Random Forest model reduces both the correlation between any two trees in the forest and reduces the strength of each individual tree in the forest. Increasing the correlation between trees increases the forest error rate. Increasing the strength of the individual trees decreases the forest error rate. A tree with a low error rate is a strong classifier. Increasing the number of features included in the model increases both the correlation and the strength[5], which is why all 561 features were used.

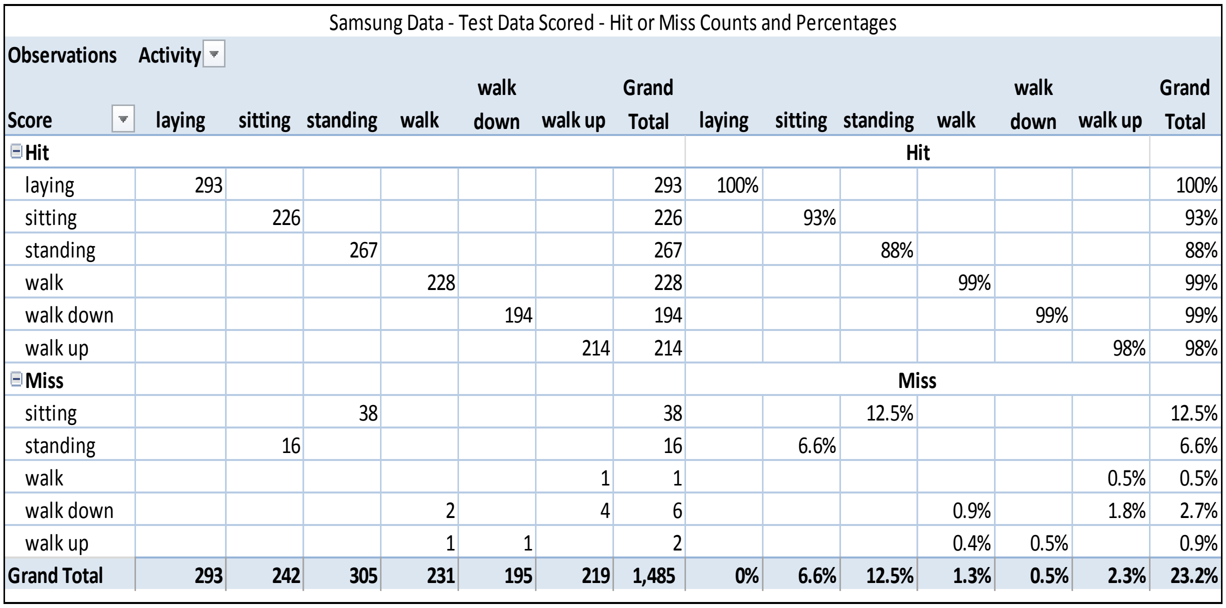
While working with the interactive user interface R package *Rattle*[3], 500 is the recommended number of trees to build. The number of features to choose from at each node is automatically calculated as the square root of the number of all variables available to the model. In the dataset used, the number of features to choose from at each node was set at 23 (√561). However, variations in the number of variables were used just to evaluate additional model performance – the quantities evaluated were 13, 23, 33, 43, and 53 – the default +/- increments of 10 variables. Generally speaking, the resulting models were not very sensitive to the changes in this parameter.

The Random Forest algorithm builds multiple decision trees from different samples of the dataset, and while building each tree, random subsets of the available variables are considered for splitting the data at each node of the tree. A simple majority vote is then used for prediction with a classification model. Random Forests models do not overfit.[5]

An estimate of the error rate was provided as the out-of-bag (OOB) estimate.[7] The OOB process applies each tree to the data that was not used in building the tree to give an estimate of the error rate.[3] Error Rates are measured and reported as each subsequent tree is generated, up to the default quantity of trees. With the model built using the Human Activity Recognition Using Smartphones Dataset, the Error Rates for the OOB and all activities decreased substantially within the first 100 trees built (see Figure 1). The OOB estimate of error rate calculated for this model was 2.34% on the Training data.

Results:

Once the Random Forests model was created, the final step in the modeling process was to Score the Test data using the model. The following Hit or Miss Table quantifies the results:



Samsung Data - Test Data Scored - Hit or Miss Counts and Percentages

Observations	Activity													
Score	laying	sitting	standing	walk	walk down	walk up	Grand Total	laying	sitting	standing	walk	walk down	walk up	Grand Total
⊟Hit								Hit						
laying	293						293	100%						100%
sitting		226					226		93%					93%
standing			267				267			88%				88%
walk				228			228				99%			99%
walk down					194		194					99%		99%
walk up						214	214						98%	98%
⊟Miss								Miss						
sitting			38				38			12.5%				12.5%
standing		16					16		6.6%					6.6%
walk						1	1						0.5%	0.5%
walk down				2		4	6				0.9%		1.8%	2.7%
walk up				1	1		2				0.4%	0.5%		0.9%
Grand Total	**293**	**242**	**305**	**231**	**195**	**219**	**1,485**	**0%**	**6.6%**	**12.5%**	**1.3%**	**0.5%**	**2.3%**	**23.2%**

The Hit and Miss percentages illustrate how laying, walking, and walking vertically up or down, were all classified with a high degree of accuracy, while sitting and standing were more often misclassified. Laying, walking, and walking vertically up or down all achieved an accuracy rating above 97%. The activities that were more challenging for the model to predict (sitting and standing) achieved an accuracy rating of 87.5% and 93.4%, respectively.

Random Forest models do not have the same concerns for potential confounders as other types of models. However, reducing the number of features used to build the model nodes reduces both the correlation and the strength of the model, while increasing the number increases both the correlation and the strength. Further testing should be conducted to evaluate if the model could be improved by changes in this parameter.

Conclusions:

A Random Forest™ model, which is an ensemble of decision trees without pruning, was used to predict which of the 6 activities a subject in the Human Activity Recognition Using Smartphones Dataset was performing (standing, sitting, laying down, walking, and walking vertically up or down). The Random Forest model was an excellent choice because the algorithm benefits from a large number of variables and yet the model was very fast, returning results in just over 2 minutes. The R package *Rattle*[3], which leverages the R package *randomForest*[6], is highly recommended as a user interface for managing Random Forest modeling since it greatly automates the process while providing a robust toolset for gathering model documentation and results. It integrates well with RStudio for setting more intricate parameters as well as integrating interactive data exploratory R package tools such as *GGobi* and *Latticist*.[8,9]

Further opportunities for improvement may exist since not all activities in the Human Activity Recognition Using Smartphones Dataset were predicted with equal accuracy. Adjustments to the model should be explored. A useful option may be to rerun the model using only those variables that were most important in the original run, or perhaps the error could be balanced by setting different weights for the inputs. Finally, additional data processing could be utilized to transform the outcome variable from categorical to 6 binary variables which could be modeled and evaluated for improvements in prediction performance.

References:

1. How Are a Gyroscope and Accelerometer Used Together? Wise Geek, Clear answers for common questions URL: http://www.wisegeek.com/how-are-a-gyroscope-and-accelerometer-used-together.htm
2. Human Activity Recognition Using Smartphones Data Set, Version 1.0, Jorge L. Reyes-Ortiz, Davide Anguita, Alessandro Ghio, Luca Oneto, Smartlab - Non Linear Complex Systems Laboratory, DITEN - Università degli Studi di Genova, Via Opera Pia 11A, I-16145, Genoa, Italy, www.smartlab.ws URL: http://archive.ics.uci.edu/ml/datasets/Human+Activity+Recognition+Using+Smartphones
3. Rattle 2.6.25 r42 - The R Analytical Tool To Learn Easily (A GNOME Data Miner Built on R) 2013-01-22 Copyright (C) 2006-2013 Togaware Pty Ltd. Users guide and help topics. http://rattle.togaware.com/ and http://cran.r-project.org/web/packages/rattle/rattle.pdf
4. The Elements of Statistical Learning (2nd ed.), Hastie, Tibshirani and Friedman (2008). Springer-Verlag. 763 pages, URL: http://www-stat.stanford.edu/~tibs/ElemStatLearn/ Chapter 15 Random Forests
5. Random Forests, Fortran original by Leo Breiman and Adele Cutler, URL: http://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm
   Random Forests™ is a trademark of Leo Breiman and Adele Cutler and is licensed exclusively to Salford Systems for the commercial release of the software. The trademarks also include RF™, RandomForests™, RandomForest™ and Random Forest™.
6. R package randomForest, R port by Andy Liaw and Matthew Wiener, URL: http://cran.r-project.org/web/packages/randomForest/randomForest.pdf
7. OOB error estimate Random Forests, Leo Breiman and Adele Cutler, URL: http://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm#ooberr
8. R package GGobi, Duncan Temple Lang, Debby Swayne, Hadley Wickham, Michael Lawrence, URL: http://cran.r-project.org/web/packages/rggobi/index.html
9. R package Latticist, author Felix Andrews URL: http://cran.r-project.org/web/packages/latticist/index.html