Title: Human Activity Recognition Can Be Accomplished with Surprisingly Few Features

Introduction:

Research focused on human activity recognition using data collected with small wearable devices has a variety of applications such as monitoring for elderly people or others at risk of injury without physical supervision [1]. Studies collecting data with technology available in smartphones have shown that measures from simple devices can be used to distinguish between human activities [1]. This analysis uses quantitative measurements collected with the accelerometer and gyroscope embedded in a Samsung Galaxy S II smartphone to predict which of six activities (walking, walking upstairs, walking downstairs, sitting, standing, and laying down) a subject was performing while wearing the phone on his/her waist [2].

Methods:

Data Collection

The data used in this analysis were provided by Dr. Jeff Leek for his Coursera Data Analysis Course [3]. The original data set included data collected on a group of 30 volunteers ranging in age from 19-48 years old [2]. This analysis focused on a subset of the original data including measurements on 7352 instances of six activities performed by 21 subjects. The data set included measurements for roughly 300-400 instances of activity per subject with the number of activity records per subject ranging from 281 to 409. The data on each activity was given as a 561 feature vector of time and frequency domain variables created through preprocessing of data collected by the accelerometer and gyroscope in a smartphone worn on the subject's waist. Additional details on the data set and the preprocessing of the data are available from the research group that originally collected the data [2].

Unusual Data Features and Transformations

Each of the 561 features in the data set were normalized and transformed to range from 1 to -1 by the data collectors [2] so no further transformations were required for the features. The subjects' activities were recorded in a character variable and this variable was converted to a factor variable or a numeric variable as needed for individual analyses and plots.

Missing Data

There were no missing data in this data set so no special handling was necessary.

Training, Validation, and Test Sets

A potential pitfall with models in general and predictive models in particular is the problem of overfitting [4]. When a model is "overfit" it matches the data too closely to generalize well to new situations. Assessing the fit of a model using the same data used to create it can lead to biased assessments and overfitting [4]. One way to guard against overfitting and to better assess the generalizability of a prediction model is to split the available data into two datasets and use one of these datasets to create the model and the other to assess it [4]. In this

analysis, we split the data into 3 parts – a training set, a validation set, and a test set. Eight subjects were assigned according to requirements specified by Dr. Leek [3] and the other 13 subjects were randomly assigned as follows: (1) All of the measurements for subjects 27, 28, 29, and 30 were reserved for the test set; (2) Subjects 1, 3, 5, 6 were assigned to the test set and then the remaining 13 subjects were randomly assigned to create a training set with all of the measurements for 9 subjects (1, 3, 5, 6, 7, 8, 16, 17, and 23) and a validation set with all of the measurements for the remaining 8 subjects (11, 14, 15, 19, 21, 22, 25, and 26). All of the activity observations for an individual subject were kept together in the same set because we are trying to create a model that will generalize to new people. By keeping all of an individual's activities together we are working with data sets to create and test the model that are much like future data we might encounter.

Exploratory Analysis

Our exploratory analyses focused on the training set to preserve the validation set for error estimates before selecting a final model. Exploration of the training set focused on exploring clusterings of activities and features and selecting variables to distinguish between the six activities for prediction. Plots for each individual colored by activity, k-means clustering, and singular value decomposition [5] were used to explore clusterings of the activities and to identify which of the 561 features to use in a prediction model. A singular value decomposition of the training set showed that the first five right singular vectors accounted for 65% of the variation coded in the 561 features.

Prediction Models

Our analysis focused on tree based classification models because classification trees are useful for sorting through many predictors and can handle interactions and nonlinearities in the data [6]. Two different tree based classification models were fit using the Training Set and the tree() function in R [7]: (1) using all 561 features as predictors; (2) using the top 10 contributing variables for the first five singular vectors from the singular value decomposition of the Training Set done in the exploratory analyses. Each of these candidate trees were pruned based on results from the cv.tree function in R [7] to avoid overfitting to the training data and then tested on the Validation Set to get a sense of the predictive ability of each classification tree [4].

The first tree including all 561 features as predictors used 9 of the features to split the training set into 10 terminal nodes. Tree pruning algorithms in R suggested a tree with six nodes might suffice [7]. Pruning this tree model resulted in a prediction model using only 5 of the feature variables to split the training set into 6 terminal nodes corresponding to the 6 activities of interest. The misclassification rate for the pruned tree applied to the training set was 13% (391 out of 3010). Using this pruned tree for prediction on the validation set resulted in a slightly higher misclassification rate of 17% (478 out of 2857).

The second tree including only the top contributing variables identified via singular value decomposition of the training set as input resulted in a tree with splits using 9 variables to split

the training set into 11 terminal nodes. Pruning this tree resulted in a prediction model using only 5 of the feature variables to split the training set into 7 terminal nodes with two of the terminal nodes being assigned to walking downstairs. The misclassification rate for this pruned tree applied to the training set was 15% (464 out of 3010). Using this pruned tree for prediction on the validation set resulted in a slightly lower misclassification rate of 14% (398 out of 2857).

An additional tree was fit and then pruned using the combined training and validation sets (all 5867 observations on the 17 training and validation subjects) and all 561 features. The resulting tree split using 8 variables into 9 terminal nodes. Pruning this tree resulted in a prediction model using only 5 of the feature variables to split the training set into 6 terminal nodes corresponding to the 6 activities of interest. The misclassification rate for this pruned tree applied to the combined training and validation sets was 11% (636 out of 5867).

Potential Confounders

The goal of the current analysis is prediction rather than interpretation, so we are not overly concerned with confounding variables. However, if we were to build on this analysis and move toward interpreting our classification trees, we would need to think carefully about potential confounders before making any interpretations.

Results:

Since the two candidate trees developed on the training set yielded similar misclassification errors on the validation set, we opted to select the pruned tree developed on the combined training and validation sets as our final prediction model (see Figure 1). Our prediction model uses only one variable to split the stationary activities (laying, sitting, and standing) from the mobile activities (walking, walking up, and walking down). The actual feature variables used for the splits were:

Frequency Domain 1 = fBodyAccJerk.bandsEnergy...1.16
Time Domain 1 = tGravityAcc.min...X
Angle Mean = angle.Y.gravityMean.
Frequency Domain 2 = fBodyAccMag.energy..
Time Domain 2 = tGravityAcc.arCoeff...Y.2

Roughly speaking, the time domain variables measure how quickly the subject was moving, the frequency domain variables measure how frequent the movements were, and the angle variables measure the direction of movements. Specifics about what these measurements mean are beyond the scope of the current analysis but are available from the original data collectors [1]. After selecting our final model, we used our tree to predict the 1485 activities of the 4 subjects in the test set. The misclassification rate for our final prediction model applied to the test set was 14% (205 out of 1485). Our model perfectly predicted all 293 instances of laying down and almost perfectly predicted walking (226 out of 229, 99%). Predictions for sitting and standing were both over 80% accurate (86% and 81% respectively) and all of the

errors for these two stationary activities were predictions for the other – that is, all of the errors for sitting predicted standing and all of the errors for standing predicted sitting. Not surprisingly, our model was least successful with predicting walking down stairs (79% accurate) and walking up stairs (69% accurate). As expected, all of the errors for walking up and walking down stairs were incorrectly categorized as other walking activities.

Conclusions:

Not surprisingly, the misclassification rate for the test set was higher than the rate when using the tree to predict the data on which the model was developed. However, we were surprised and encouraged to find that a model using only 5 of the 561 reported features could achieve such a respectable classification rate on new data. These results are promising, suggesting that more sophisticated models developed on larger data sets might be very useful in predicting human activity collected with a small wearable device like the Samsung smartphone used to collect these data.

One potential problem with the data used in this analysis is the lack of information about the subjects. For example, it is possible that by chance our sample contained all of the younger subjects and that activity data differ by a subject's age (or other subject characteristics). Another weakness of this analysis is we considered simple prediction models and did not take advantage of more sophisticated methods for model creation and refinement (e.g. random forests [8], support vector machines [9], the caret package in R [10], leave-one-out cross validation [4]). More advanced techniques were not used for the current analysis because we did not have the expertise to employ them thoughtfully. Our future analyses of these data and/or the problem of human activity recognition should focus on ways of combining the various features to improve classifications, more advanced techniques for feature selection and model building, and on finding ways to include and/or control for characteristics of the subjects to further improve predictions.

References

[1] Davide Anguita, Alessandro Ghio, Luca Oneto, Xavier Parra and Jorge L. Reyes-Ortiz. Human Activity Recognition on Smartphones using a Multiclass Hardware-Friendly Support Vector Machine. International Workshop of Ambient Assisted Living (IWAAL 2012). Vitoria-Gasteiz, Spain. Dec 2012

[2] Davide Anguita, Alessandro Ghio, Luca Oneto, Xavier Parra and Jorge L. Reyes-Ortiz. UCI Machine Learning Repository. URL: http://archive.ics.uci.edu/ml/datasets/Human+Activity+Recognition+Using+Smartphones

[3] Samsung Data for Data Analysis Coursera course taught by Dr. Jeff Leek, January-March 2013. URL: https://class.coursera.org/dataanalysis-001/class/index

[4] Andrew Moore. Cross-validation for detecting and preventing overfitting. Downloaded March 5, 2013. URL: http://www.autonlab.org/tutorials/overfit.html

[5] Kirk Baker. Singular Value Decomposition Tutorial, March 29, 2005 (Revised January 14, 2013). downloaded 3/2/2013; URL: http://www.ling.ohio-state.edu/~kbaker/pubs/Singular_Value_Decomposition_Tutorial.pdf

[6] Classification and Regression Trees, 36-350, Data Mining 6, November 2009. Downloaded March 2, 2013. URL: http://www.stat.cmu.edu/%7Ecshalizi/350/lectures/22/lecture-22.pdf.

[7] Package 'tree'. February 15, 2013. Downloaded March 1, 2013. URL: http://cran.r-project.org/web/packages/tree/tree.pdf

[8] Leo Breiman and Adele Cutler, Random Forests. URL: http://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm

[9] Alexandros Karatzoglou, David Meyer, and Kurt Hornik, Support Vector Machines in R. Journal of Statistical Software, April 2006, Volume 15, Issue 9.

[10] Max Kuhn. Building Predictive Models in R Using the caret Package, Journal of Statistical Software, November 2008, Volume 28, Issue 5.