Behavior Modeling Based on Measurements of Subject Movements

Introduction

Many consumer devices such as cell phone and game controllers contain hardware to measure the movements of the user. This data can be used to make playing to game a more enjoyable activity. In the cell phone applications can be developed for a number of purposes such as tracking healthful activities. This analysis uses data from the Samsung phone to identify user activities.

Methods

Data Collection

The data documentation and data used for this analysis is available at http://archive.ics.uci.edu/ml/datasets/Human+Activity+Recognition+Using+Smartphones. The Data was preprocessed by Dr. Leek at the Johns Hopkins Bloomberg School of Public Health for the Coursera course "Data Analysis" offered in early 2013 and was downloaded on February 25, 2013. There are 7,352 observations from 21 subjects in each of six activities for 561 different variables. There was no missing data. However there were several duplicate variable names. These variables however contained differing values for each observation. A review of the variable names leads to credence that they were variables for the x, y, and z dimensions that were not properly labeled in the data set. They were subsequently renamed and retained in the analysis.

Exploratory Analysis

The data was split into three sets with eleven subjects in the training group, four in each of a validation and testing data sets. All of the initial exploratory analysis was done on the training data set. The full training, validation, and testing data sets contained 4373, 1494, and 1485 observations respectively. These data sets contained 13, 4, and 4 subjects respectively. No outlier observations were detected in the analysis.

Exploratory plots showed that the orientation of the X, Y, and Z axis were generally consistent among the subjects. There is some evidence of small rotational differences and scaling between subjects. However we decided there were an insufficient number of subjects to develop transformations that could adjust for this issue. Initial exploratory analysis using the first nine variables in the data set indicated that tree models at the individual subject level performed better than models using the entire training data set. Simple tree models on subjects 1, 3, 4, and 6 produced a misclassification rate of 0.32. Models run for each subject separately produced misclassification rates from 0.12 to 0.19. A natural improvement in the misclassification rate is expected due to the smaller number of observations for the subject level trees, but the improvement seen is larger than what would be expected. These observations indicate that there are potential confounders missing from the data set that could identify

reasons for differences between subjects. Obvious variables that could impact the accelerometer readings are such items as weight, detailed age, and sex of each subject.

Several data transformations were considered. Variable transformations measuring the mean and variability of all of the data elements for a given observation for an individual were computed. These showed some promise but were ultimately rejected as ineffective in the classification process.

Statistical Modeling

Variable identification and preliminary model development were done in parallel. The variable set was divided into 20 non-overlapping sets. An additions five sets were created based on similar classes of variables. For example, one set was all observations in the X-dimension.

The observations were classified into one of three groups with a great deal of accuracy without the need to use any complex modeling. The groups are:

1. Lying
2. Sitting and Standing
3. Walking, Walking Up, and Walking Down

(Note that I have used here the correct form of the work "lie" throughout this report. The original data set uses the incorrect form of the word). Tree models were run on the full set of observations, for each of the variable sets, for each of the three groups and for the three groups combined. The variable measuring the mean gravity in the X-dimension was able to identify correctly all but two of the observation for lying. Five variables were identified that each correctly separated observations into groups 3 and 4. These included four of the accelerometer jerk measurements and one of the body accelerometer measurements. This allowed for the effective identification of the three groups. The model to this stage as well as the variable list is shown in the flow chart provided in Figure 1b. The specific variables are identified in the figure.

Models and variable sets were identified to classify observations within groups 2 and 3. Tree models were run separately for both groups, for each of the 25 sets of variables referred to above. The best performing variables for each of these runs were then combined to test various models. This proved to be effective in identifying variables to correctly classify group 2 observations into the classes sitting and standing. However it did not work to classify between walking, walking up and walking down. A random forest model was better able to classify within group 3 where it gave a 1.8% error rate in the training data set. The full set of variables used in these models is provided in the documentation files.

Various methods and variable choices were examined for splitting the "standing" and "sitting" activities. These included different subsets of the variables with logistic regression, random forest and the tree algorithms. A tree algorithm ultimately worked well. An automated statistics based method worked best to select the variables. The mean and variance was computed for the 561 variables for the "sitting" and "standing" observations. A pseudo t-test statistic was computed on each variable to test the difference in the mean values. These are not a true t-statistics since the variables are constrained to the closed

interval [-1, 1] and the usual normality assumption does not hold true. The 65 variables with the largest pseudo t-statistic were chosen for the final model. A plot of the values of these t-statistics in provided in Figure 1a. This identified a subset of the 51 variables that were “most different” for the classes lying and sitting in the basis of the mean and variance or the observations for each group. The tree algorithm selected 12 of these variables for the classification.

This process resulted in the first classification model. It used one variable to separate group 1 from groups 2 and 3. Five variables were used to split groups two and three. A tree model separated lying from standing in group 2. Finally a random forest model separated walking, walking up, and walking down in group 3. The combined error rate using this model on the full training data set was 1.9%.

A second random forest model was created using the 96 variables identified complex preliminary model. This model performed somewhat better with an error rate of 1.4%. It was able to correctly identify classify the observations into groups 1, 2, and 3 and also provided better at classifications within groups 2 and 3.

Reproducibility

The exploratory work and statistical methods used the R programming language. All of the work was saved as documentation in R scripts. They include outputs from the analysis as well as descriptions of the reasons for the decisions made in the analysis process and the code that supported those decisions. A complete list of the variables identified for each model and for final model is given in the documentation and is too lengthy to include here. The work can be reproduced using these files.

Results

The two models described above were run against the validation data set. The error rate for the mixed tree, random forest model was 12.7%. For the full random forest model the error rate was 11.7%. Analysis of the validation data set led to the deletion of three variables that negatively impacted the random forest model. This reduced the error rate on the validation data set to 11.4%. Deleting variables that the random forest model identified as not significantly important by the model did not improve the model. The original model had 93 variables. Separate models were run retaining only the 60 and retaining only the 15 most important as determined by the random forest algorithm. Both resulted in worst performance on the training data set. The top two performing variables reported by the random forest algorithm were minimum gravity acceleration in the x-dimension and mean gravity acceleration in the y-dimension. The full list of variables is provided in the documentation files.

Additional random forest models were run based on bootstrapping techniques using random subset of subjects and random subsets of the variables. These models had the same error structure as did the original two models. Figure 1c shows a key problem of over fitting revealed by these runs. It provides a histogram of the error rates for 301 random forest models where random subsets of 15 of the 561 variables were selected. In close to 40% of these models the error rate was less than three percent. Thus there are many small subsets of the 561 variables that are good predictors. Selecting the right variables without over fitting can be difficult in this situation.

The original random forest model was selected as the final predictor as it had the lower error rate on the validation data set.

Conclusions

Application of the final random forest model with the 93 variables to the test data set gave error rate of 7.7%. This should be viewed as an approximate error as the rate was 2.6% for one subject and 9.9%, 10.5%, and 10.2% for the other three subjects. Under the assumption that the test cases are a sample from the general population a bootstrap estimate of the standard error on the 7.7% is 0.7%. This is likely an understatement of the true variability as the bootstrap method does not adequately account for the between subject variability. The wide disparity in the error rates between subjects is troubling. It is a further indications of the need for additional data that would facilitate modeling the between subject variability. We believe a better variance estimate that takes into account the between subject variance is the simple variance amount the four subjects in the test data set. This gives a standard error estimate of 2.2% on the estimated 7.7% but is based on only four observations.

The model was able to correctly separate observations into groups 1, 2 and 3. Identification of activities with groups 2 an3 is more problematic. See Table 1 for a summary of classification errors for the test data set. Our experience with the bootstrap models has led us to the conclusion that further improvement in the modeling accuracy will be difficult. A significant issue remaining is the differences between subjects in how they used the device combined with the inability to model these differences due to the size of the subject universe and likely unidentified confounding variables such as sex, age, and weight of the subjects.

Table 1

Cross tabulation of predicted by true classification in the test data set for the selected model

Test Data Set Classifications

Model Prediction	Lying	Sitting	Standing	Walk	Walk Down	Walk Up
Lying	293	0	0	0	0	0
Sitting	0	226	50	0	0	0
Standing	0	38	233	0	0	0
Walk	0	0	0	229	6	14
Walk Down	0	0	0	0	192	5
Walk Up	0	0	0	0	2	197

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] Data analysis Using Regression and Multilevel/Hierarchical Models, Andrew Gelmen, Jennifer Hill, Cambridge University Press 2007.

[3] Pattern Recognition and Machine Learning, Christopher M. Bishop, Springer 2009.

[4] Shih, Stephanie, Random Forests for Classification Trees and Categorical Dependent Variables: an informal Quick Start R Guide. Stanford University | University of California, Berkeley, February 2, 2011. Source:http://www.stanford.edu/~stephsus/R-randomforest-guide.pdf

[5] Livingston, Frederick, Implemettaion of Breiman's Random forest Machine Learning Algorithm, ECE591Q Machine Learning Journal Paper, Fall 2005, URL: http://gogoshen.org/ml2005/Journal%20Paper/JournalPaper_Livingston.pdf

[6]Breiman, Leo(2001). "Random Forests". Machine Learning 45(1):5-32. URL: http://link.springer.com/article/10.1023%2FA%3A1010933404324