Predicting Activity by Phone Accelerometer and Gyroscopic Measurements

Introduction

Every day we engage in a wide variety of activities, from the active to the sedentary. How much do these activities vary in our basic movements? Is it possible to predict what someone is doing based on information about his or her acceleration and position alone? Now that more and more of us have phones in our pocket and now that our phones are increasingly sophisticated, it might be possible to determine what everyone in the world is doing at once using phone accelerometer and gyroscopic data. In this paper, I seek to predict the activity of a subject using nothing more than the accelerometer and gyroscopic information from their phone.

Methods

Data Production

30 volunteers (ages 19-48) volunteered to preform six different activities (walking, walking upstairs, walking downstairs, sitting, standing, and laying) while wearing a Samsung Galaxy S II on their waist. 3-axial linear acceleration and angular velocity as well as gyroscopic information was captured using the built-in accelerometer of the phone. The data was then made publicly available for anyone to use.

Data Collection

The data was accessed on 8 December 2013 from the UCI Machine Learning Repository[1] using the R programming language.

Statistical Modeling

The goal of this project is to predict, as accurately as possible, the type of activity the subject was performing from the given accelerometer data. To do this, a training set was made of 12 subjects and the "trees" method was used on this set to create a predictive formula. This formula was then tested on a validation set of four subjects that was not used in the training set. Tree pruning was used to refine the predictive formula using the validation set in order to protect from overfitting. Lastly, the final pruned tree model was used on the test set in order to gauge how accurate the predictive formula was at estimating activity type based on accelerometer data.

Analysis

Descriptive Analysis

The dataset contained 561 different measures of motion via the accelerometer, as well as one variable that contained the label of the activity being performed and another variable that contained the label of the subject. Each measure of motion was made from combining X factors – a accelerometer/gyroscopic variable (body acceleration, gravity acceleration, body jerk, angular momentum, energy, and entropy), a dimension (X, Y, or Z), and a summary statistic (mean, median, standard deviation, kurtosis, maximum, minimum, and skewness).

Nine subjects were not found in the data set and accordingly their data could not be used for analysis. Furthermore, not every subject had a measurement for every measure of motion. Other than that, there were no missing values in this analysis and data came already pre-cleaned and pre-transformed. The only additional transformation required was removing duplicate variable names in the data set. All variables appear to be within normal ranges.

Initial Trees Analysis Using the Training Set

From the 21 available subjects, a training set was made using a random subset of 12 of those subjects with the constraints provided by the assignment that (1) participants 1, 3, 5, and 6 had to be in the training set and (2) participants 27, 28, 29, 30 could not be used as they must be reserved for the test set.

A tree model was fit to the training set using all possible variables except the actual activity label. The model ended up using 13 variables and resulted in 13 terminal modes, with a residual mean deviance of 0.334 and a misclassification error rate of 5.804% (241 misclassifications out of 4152 total classifications).

Tree Pruning Using the Validation Set

In addition to the training set, a validation set was made using a random subset of 4 additional subjects with the constraints that (1) none of the subjects in the validation set could also be in the training set and (2) participants 27, 28, 29, 30 could not be used as they must be reserved for the test set.

The original tree model from the training set was then applied to the validation set. The model misclassified 16.134% of the activities (227 misclassifications out of 1407 total classifications). The specific predictions for each activity can be seen in Table 1, which shows that our model is having trouble distinguishing sitting from standing and distinguishing walking, walking down, and walking up.

This poor performance indicates that there may be a problem of overfitting, where the tree model fits too tightly to the particular irrelevant features of the training data set rather than the features that generalize across the entire data set. To correct for overfitting, I employ tree pruning, or cutting the tree to use less variables than the original 13.

To find the correct tree pruning, I pruned off one variable at a time until I had cut back from a 13 variable tree model (the original model) to a 2 variable tree model. I then recorded the classification performance of each model, which can be seen in Table 2. After this test, it was found that a variable tree model using between seven to ten of the variables provide the optimal predictive power in the validation data set, only misclassifying 14.357% of the data (202 misclassifications out of the 1407 total classifications). A seven variable tree model was used on the test set because of the desirable simplicity of having less variables in the model.

Table 3 shows the performance of this new seven variable tree model on the validation set. Compared to the original 13 variable model, performance improves, but not by much. The model has not gotten much better at distinguishing sitting from standing and has gotten worse at distinguishing walking from walking down, but has improved rather well at distinguishing walking up from the other variables.

Final Result Using the Test Set

Lastly, a test set was made using the original data using the four remaining participants (27, 28, 29, 30) that were specifically reserved for the test set as required by the assignment. The seven variable tree model was then applied to this test set and was found to misclassify 12.525% of the data (186 misclassifications out of the 1485 total classifications). Individual activity classifications are shown in Table 4 and we can see the same problem with distinguishing sitting from standing and distinguishing walking, walking up, and walking down, though overall performance is good. Figure 1 shows the overall seven variable tree model and the variables it uses.

Conclusions

Overall, a pruned tree model was able to successfully classify all but 12.525% of the activity data in a test set of four participants after being trained on the data of 12 participants and validated on an additional dataset of four participants. This is a pretty good basis for being able to figure out people's activities based on accelerometer data, successfully distinguishing three factors of activity – laying, sitting/standing, and walking/walking down/walking up – from each other with very high accuracy. It's unlikely that any confounding variables will be present, though machine learning will certainly be complicated by the fact that not everyone performs the activity in precisely the same way.

However, within each factor there is less predictive ability and multiple classification errors, though accuracy is still good and gets the right activity in a majority of the cases. It might be possible to improve this prediction with better cross-validation or more powerful statistical techniques, like random forests.

References

1: "Human Activity Recognition Using Smartphones Data Set." UCI Machine Learning Repository. http://archive.ics.uci.edu/ml/datasets/Human+Activity+Recognition+Using+Smartphones. Accessed 8 Dec 2013.

Tables

Table 1: Accuracy of Final Seven Variable Tree Model on the Validate Set, Comparing Actual Activity with Predicted Activity

	Predicted Activity					
Actual Activity	Laying	Sitting	Standing	Walking	Walk Down	Walk Up
Laying	270	0	0	0	1	0
Sitting	0	220	37	0	0	0
Standing	0	82	202	0	0	0
Walking	0	0	0	202	2	15
Walk Down	0	0	0	19	158	4
Walk Up	0	0	0	56	11	128

Table 2: Misclassification Rate of the Tree Model Based on the Number of Best Variables in the Tree

Number of Variables in Tree	**Misclassification Rate**
12	16.1%
11	16.6%
10	14.4%
9	14.4%
8	14.4%
7	14.4%
6	14.7%
5	24.6%
4	35.2%
3	45.1%
2	65.2%

Table 3: Accuracy of Final Seven Variable Tree Model on the Validate Set, Comparing Actual Activity with Predicted Activity

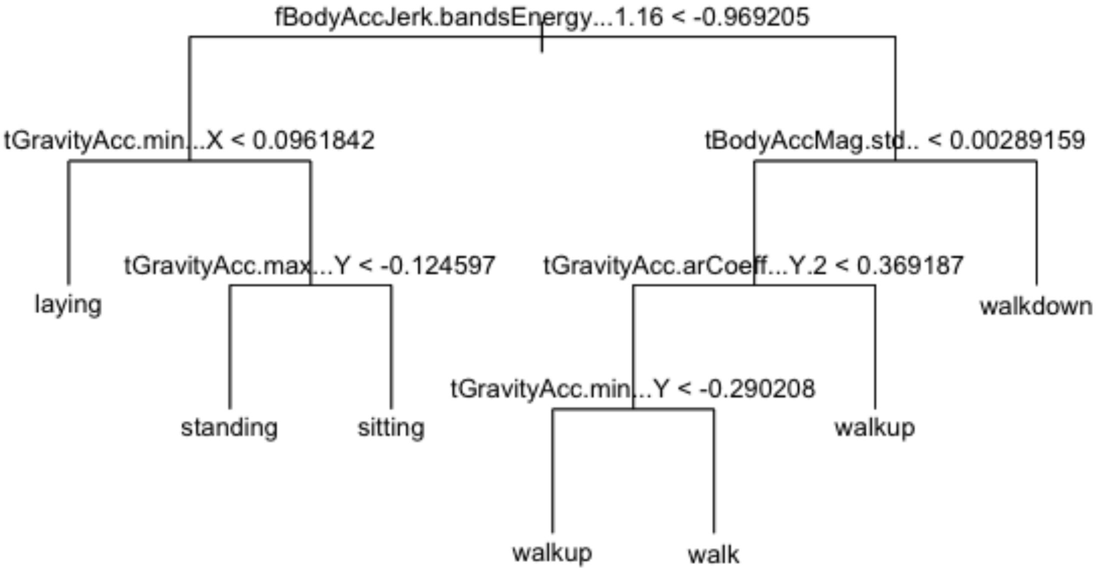
	Predicted Activity					
Actual Activity	Laying	Sitting	Standing	Walking	Walk Down	Walk Up
Laying	270	0	0	0	1	0
Sitting	0	220	37	0	0	0
Standing	0	82	202	0	0	0
Walking	0	0	0	196	1	22
Walk Down	0	0	0	20	151	10
Walk Up	0	0	0	21	8	166

Table 4: Accuracy of Final Seven Variable Tree Model on the Test Set, Comparing Actual Activity with Predicted Activity

	Predicted Activity					
Actual Activity	Laying	Sitting	Standing	Walking	Walk Down	Walk Up
Laying	293	0	0	0	0	0
Sitting	0	227	37	0	0	0
Standing	0	63	220	0	0	0
Walking	0	0	0	226	0	3
Walk Down	0	0	0	33	148	19
Walk Up	0	0	0	14	17	185

Figures

Figure 1: Seven Variable Tree Model for Predicting Activity from Accelerometer Data


fBodyAccJerk.bandsEnergy...1.16 < -0.969205
tGravityAcc.min...X < 0.0961842
tBodyAccMag.std.. < 0.00289159
laying
tGravityAcc.max...Y < -0.124597
tGravityAcc.arCoeff...Y.2 < 0.369187
walkdown
standing
sitting
tGravityAcc.min...Y < -0.290208
walkup
walkup
walk

Variables are categorized according to this decision tree created from a seven variable tree model. If the data satisfies the given condition, it goes down the tree to the left; otherwise, it goes down the tree to the right. (Apologies for the R variable names, but I have no friggin' clue what these variables actually refer to, precisely.)