Predicting Smartphone User Activity using Bootstrapped Classification Trees

Introduction

Smartphones are becoming increasingly prevalent [1]. As a result, many people now regularly carry a device which contains multiple sensors, including three-axis accelerometers, ambient light sensors, and Global Positioning Systems (GPS) [2]. Prior to the large scale adoption of smartphones, these kinds of sensors might only have been available as specialized, single purpose devices. The data from these sensors, properly interpreted, have the potential to provide real-time information about the events and objects a smartphone user encounters.

While some applications of this kind of data are focussed on making life more convenient for the individual smartphone user, e.g. by identifying the restaurants closest to them, it also has potentially broader implications. Research that previously required the expensive deployment of specialized measuring technology can now be conducted through people's existing smartphone devices, e.g. using a smartphone app to study traffic patterns rather than having to attach accelerometers to vehicles [3].

However, mapping the data provided by these sensors onto outputs that will be meaningful for the user remains a difficult problem. Activities like walking, sitting and standing might seem vastly different from an intuitive point of view, but prove difficult to separate and classify when looking at the data outputted from an accelerometer. This kind of classification is also potentially limited by the processing power of smartphones, as complex statistical or machine learning approaches to data interpretation may not be feasible in real-time.

In this analysis, we applied classification trees to data from a smartphone accelerometer in order to predict the activity a person is currently undertaking. While an individual classification tree showed decent predictive power, multiple classification trees using different sets of predictors across varying samples were able to achieve a high degree of accuracy.

Methods

Data Collection

The data used for these analyses were a set of accelerometer readings from 21 volunteers performing six different activities: laying, sitting, standing, walking, walking upstairs and walking downstairs. This dataset was downloaded from the Coursera course website [4]. For the purposes of prediction and validation,

the data were divided into a training set containing 11 subjects, a validation set containing 5 subjects, and a test set containing 5 subjects.

Statistical Modelling

Statistical prediction was performed using classification trees [5], as implemented in the `tree` library for R [6]. Trees were applied to bootstrapped samples of the data, and predictions from the individual trees were combined using a simple majority vote criterion, so that the activity label predicted is the one that is predicted in the largest number of the prediction models. Error rates were assessed using misclassification rates, i.e. the proportion of predicted category labels that did not match the actual category label. All statistical analyses and data processing were carried out using the R programming language [7].

Results

The dataset did not contain any missing values. Accelerometer readings had been pre-processed such that they were normalized between -1 and 1, and none of the values for these variables fell outside those ranges. Each subject also had a similar number of data points, with a range from 281 to 409 and a mean of 350.1 across the 21 subjects.

The number of cases for each activity label in the training dataset was also approximately equal, $range = 504 - 655$, $mean = 600.8$. Assuming the training set was representative of the test set, this meant that each activity label should be equally weighted in the prediction model.

An initial classification tree was fitted to the training set using all 561 accelerometer readings as predictors. This model was able to locate predictors which separated many of the activity labels well: for example, the first split was performed on `fBodyAccJerk_std_x`, or the standard deviation of the Jerk signal on the x-axis, passed through a Fast Fourier Transform. This split perfectly partitioned the stationary activities (laying, sitting, standing) from the mobile ones (walking, walking upstairs, walking downstairs) in both the training and validation datasets. However, while this model performed acceptably within the training set, achieving a misclassification rate of 8.54%, its performance out-of-sample was less accurate, with a misclassification rate of 18.88% on the validation data set.

The poor out-of-sample performance of the initial classification tree suggested overfitting, so it was decided to use bootstrap aggregation to minimize this problem. Classification trees were fitted to 20 bootstrapped samples of the training data, again using all 561 accelerometer measures as predictors. The predicted value from these trees was the majority vote over all 20 trees after each data point had been dropped down each of the trees.

The bootstrap aggregated classification trees showed some success in reducing the overfitting, producing an error rate of 17.11% on the validation dataset, even though the training set error was reduced only slightly to 8.16%. However, the still-considerable discrepancy between in-sample and validation error rates suggested that overfitting was still an issue.

To further address overfitting, the existing bootstrapped classification trees were inspected to see which variables were being used to form splits. As shown in Table 1, some variables were heavily relied upon to form the classifiers. In order to create classifiers that relied on a broader range of decision criteria, a new set of bootstrapped trees was created, with the 5 variables that had been used most often in the initial set of trees excluded from the predictors.

Variable	Times Used in Initial Bootstrapped Trees
tGravityAcc_mean_Y	23
tGravityAcc_min_Y	23
tGravityAcc_mean_X	21
tBodyAcc_max_X	18
tGravityAcc_max_Y	12
angle_Y_gravityMean	10
tBodyAccJerk_std_X	9
fBodyAccJerk_std_X	7

Table 1: Variables used most often in the initial bootstrapped trees

The resulting model combined the initial 20 bootstrapped trees with the 20 trees that excluded the 5 most used variables. The values predicted for each data point were again decided by a majority vote across the predictions produced by the 40 trees. This model had an improved in-sample error rate, with a misclassification rate on the training dataset of 6.96%, and a validation sample error rate of 12.14%. Due to this improved out-of-sample error rate, it was decided to use the combined tree model as the final predictive model.

The confusion matrix for the final predictive model applied to the test set are shown in Table 2. The final misclassification error rate achieved was 9.8%. As the confusion matrix shows, misclassification tended to occur between similar categories such as sitting/standing and the different forms of walking.

	laying	sitting	standing	walk	walkdown	walkup
laying	369	0	0	0	0	0
sitting	0	304	56	0	0	0
standing	0	38	301	0	0	7
walk	0	0	0	281	13	11
walkdown	0	0	0	1	232	47
walkup	0	0	0	6	5	206

Table 2: Confusion matrix for the final predictive model applied to the test set

Conclusion

By combining two sets of bootstrapped trees that relied on different predictors, the final predictive model was able to avoid some of the overfitting problems present in the initial bootstrapped tree model. The error rates for the original and final models are compared in Figure 1: it can be seen that the original model had particular trouble with correctly labelling instances of walking in the validation dataset. Features that predicted walking well in the training sample may have been unique to that sample, leading to overfitting that was only overcome when the model was forced to consider other predictors.

References

[1]A. Smith, "46% of American adults are smartphone owners." Pew Internet & American Life Project, 2012.

[2]G.M. Weiss, "Your Smartphone Knows You Better Than You Know Yourself." web, January-2013 [Online]. Available: http://www.insidescience.org/content/your-smartphone-knows-you-better-you-know-yourself/904

[3]R. Bhoraskar, N. Vankadhara, B. Raman, and P. Kulkarni, "Wolverine: Traffic and road condition estimation using smartphone sensors," pp. 1–6, Jan.

[4]J. Leek, "Samsung Accelerometer Data." [Online]. Available: https://spark-public.s3.amazonaws.com/dataanalysis/samsungData.rda%22%3ELink. [Accessed: 27-Feb-2013]

[5]T.J. Hastie, R.J. Tibshirani, and J.J.H. Friedman, *The Elements of Statistical Learning*. Springer-Verlag New York, 2009 [Online]. Available: http://books.google.com.au/books?id=tVIjmNS3Ob8C

[6]B. Ripley, *tree: Classification and regression trees*. 2012 [Online]. Available: http://CRAN.R-project.org/package=tree

[7]R Foundation for Statistical Computing, *R: A Language and Environment for Statistical Computing*. Vienna, Austria: , 2013 [Online]. Available: http://www.R-project.org/