Introduction

This paper compares two predictive models on Samsung Galaxy 3 accelerometer and gyroscope data gathered from 21 participants who each exhaustively performed several activities with the phone on their waist, including: standing, siting, laying down, walking on a flat surface, walking up stairs and walking down stairs. Each of these physical activities should have an identifiable behavioural trace that is recorded by the accelerometer and gyroscope readings. The accelerometer measures movement in three directions of the x, y, and z axes; while the gyroscope provides a means to orient the phone's movement with respect to the gravitational earth-down position. Detailed descriptions of the methodology for producing the summary data and **data transformations, including the 561 position-related variables, are provided elsewhere (see 1)**. The benefits of such a prediction model are evident in keeping an automated log of physical activity for improved fitness and weight loss.

Methods

Twenty-one (N=21) participants wore a Samsung Galaxy 3 smartphone on their waist while exhaustively performing several physical activities. A **total of 5867 observations** of different activities (sessions) were included in the dataset. There were **no missing observations or unusual features** in the data, as the dataset was constructed with extensive pre-processing (1). The purpose of the analyses presented below is to compare two predictive models used to identify those activities a person was performing solely from the accelerometer and gyroscopic data. First, a Tree model in R using the rpart package (2) was used to identify which elements among the 561 positional variables were helpful for predicting participant activity. Second, a random forest model using the randomForest package in R (3) was used to attempt a more accurate prediction. The Tree classification has good parsimony, and provides one tree diagram to illustrate the prediction method. The random forest is less parsimonious and does not provide a single tree-diagram to illustrate the process of prediction. Nevertheless, the random forest has lower model-specification bias as it is not constrained to one model solution, but instead bags many different tree models to generate predictions.

Results

The Tree Model

A tree model was fitted with the rpart package in R using the default settings (2) to predict activity of participants (laying, sitting, standing, walking, walking up and walking down) from the gyroscope and accelerometer measurements (1). The model was constructed from a training set of the first 17 participants, with another set of participants; including **subjects numbered 26, 27, 28 and 29; held out as a test set**. The classification tree fitted is illustrated in Figure 1 below.

Table 1 below shows the classification accuracy of the Tree model for both the training and the test set. Misclassification in the training set was 10.00% of observations that were assigned to the wrong categories by the tree (see PANEL A). The misclassification was larger in the test set at 13.00%, which may represent a small degree of overfit for the model. Only the training set was used to construct

and parameterize the model, and thus the test set is a potentially better measure of prediction accuracy.

Table 1. Confusion Matrix for Tree Model

PANEL A
Training Set (actual= rows, predicted= columns)

	laying	sitting	standing	walk	walkdown	walkup	error
laying	1114	0	0	0	0	0	.0000
sitting	0	947	129	0	0	0	.1199
standing	0	75	962	0	0	0	.0723
walk	0	0	0	871	41	64	.1076
walkdown	0	0	0	19	658	64	.1120
walkup	0	0	0	107	87	729	.2102

Missclassification = 10.00%

PANEL B
Test Set (actual= rows, predicted= columns)

	laying	sitting	standing	walk	walkdown	walkup	error
laying	293	0	0	0	0	0	.0000
sitting	0	226	55	0	0	0	.1957
standing	0	38	228	0	0	0	.1429
walk	0	0	0	220	9	27	.1406
walkdown	0	0	0	0	158	22	.1222
walkup	0	0	0	9	33	167	.2010

Misclassification = 13.00%

Random Forest

To improve prediction accuracy from the Tree model (2), a random forest model (3) was applied using the same training set while the resulting unmodified model was also used to predict the activity values in the test set. Table 2 below shows the classification accuracy for both sets. The training set, as shown in Panel A, had very good accuracy with only 1.64% out of bag error (misclassification). The model showed much lower accuracy (5.05% misclassified) for the test set.

Table 2. Confusion Matrix for the Random Forest Model

PANEL A
Training Set

	laying	sitting	standing	walk	walkdown	walkup	error
laying	1114	0	0	0	0	0	.0000
sitting	0	994	27	0	0	1	.0274
standing	0	38	1053	0	0	0	.0348
walk	0	0	0	985	6	6	.0120
walkdown	0	0	0	4	776	6	.0127
walkup	0	0	0	1	7	849	.0093

Misclassification (out of bag error) = 1.64%

PANEL B
Test Set

	laying	sitting	standing	walk	walkdown	walkup
laying	293	0	0	0	0	0
sitting	0	225	28	0	0	0
standing	0	39	255	0	0	0
walk	0	0	0	228	3	1
walkdown	0	0	0	0	194	0
walkup	0	0	0	1	3	215

Misclassification = 5.05%

Conclusion

The analyses compared the classification accuracy of a Tree model using the rpart package in R (2) and a random forest model using the randomForest package in R (3). The Tree model showed some accuracy in the training set (90% correct) and a potentially small amount of overfit as demonstrated by lower accuracy in the test set (87% correct). The random forest model, while more far more complex, allowed for very good accuracy in the training set (98% correct). Less accuracy was achieved in the test set (95% correct), however, suggesting that the at least some of the good performance is due to overfit.

It is important to recognise that a **potential confounding factor** is a lack of consideration for subjects as an important source of variance. All observations are treated equally in these analyses, but we should expect that within-subject observations should be more similar than between-subjects observations (e.g., people have distinctive styles of walking or sitting).

The relative merits of the Tree and Random Forest models are evident in these analyses. While overfit appeared to be a problem in the random forest model, accuracy was still much improved from the simpler Tree model. The Tree model suffered less from overfit, and had the advantage of clearly identifying the method and key variables by which the prediction is made. As such, the Tree model produced some greater understanding of the essential elements of prediction, whereas the random forest provided the greatest prediction accuracy.

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 see: https://www.dropbox.com/s/rrsm7nv5j7y3rsi/analysis2info.pdf

(2) T.M Therneau and E.J Atkinson. An introduction to recursive partitioning using the rpart routines. Divsion of Biostatistics 61, Mayo Clinic, 1997.

(3) A. Liaw and M. Wiener (2002). Classification and Regression by randomForest. R News 2(3), 18—22.