Classification Tree Prediction Analysis of Human Activity Using Samsung Data Set

Revised 3/13/2013 19:43:00

Daniel D. Gutierrez[1]
[1]AMULET Analytics a Service Division of AMULET Development Corp.
P.O. Box 241713, Los Angeles, CA, USA 90024
e-mail address: ddgutierrez@alumni.ucla.edu

We use classification tree machine learning techniques for the prediction of a series of six human physical activities (laying, sitting, standing, walking, walking down stairs, and walking up stairs) using the Human Activity Recognition database for the Samsung Galaxy S3 smartphone built from the recordings of 30 subjects performing activities of daily living (ADL) while carrying a waist-mounted smartphone with embedded inertial sensors. The features selected for this database come from the accelerometer and gyroscope 3-axial raw signal's tAcc-XYZ and tGyro-XYZ. A classification tree algorithm is successfully used to predict what activity a subject is performing based on the quantitative measurements from the Samsung phone.

Keywords: Classification trees, machine learning, prediction, smartphone, activity recognition

I. INTRODUCTION

An increasing amount of research is being conducted to monitor human physical activity using smartphones. Modern smartphones are now equipped with inertial sensors such as accelerometers and gyroscopes to collect raw data signals that can be used as data sets for machine learning classifiers to predict a particular human activity. Healthcare research is one area where an understanding of when and where people are active can lead to treatments for obesity and diabetes. Accordingly, there is a push to understand how to increase physical activity and how to promote better habits of activity. Tracking physical activity during the day can be viewed as baseline monitoring, or it could be used to send an encouraging SMS reminder to an overweight patient. Activity monitoring of elderly people for potential use with assisted living technology applications is another area of interest. The key to this healthcare research is the ability to predict human activity from the smartphone raw data.

The goal of this research is to use a supervised Machine Learning (ML) algorithm – classification trees to predict the activity given raw data from using the *Human Activity Recognition* database for the Samsung Galaxy S3 smartphone. Six human activities are considered: laying, sitting, standing, walking, walking down stairs, and walking up stairs. The dataset used for this analysis was built from the recordings of 30 subjects performing activities of daily living (ADL) while carrying a waist-mounted smartphone with embedded inertial sensors. The features selected for this database come from the accelerometer and gyroscope 3-axial raw data signal's tAcc-XYZ and tGyro-XYZ.

A classification tree algorithm is successfully used to predict what activity a subject is performing based on the quantitative measurements from the Samsung phone.

II. METHODS

Data Collection

The experiments were carried out with a group of 30 volunteers within an age range of 19-48 years. Each subject performed six activities – laying, sitting, standing, walking, walking down stairs, walking up stairs – wearing a Samsung Galaxy S3 smartphone on the waist. Using its embedded accelerometer and gyroscope, the experimenters captured 3-axial linear acceleration and 3-axial angular velocity at a constant rate of 50Hz. The experiments were video-recorded to label the data manually.

The sensor signals (accelerometer and gyroscope) were pre-processed by applying noise filters and then sampled in fixed-width sliding windows of 2.56 sec and 50% overlap (128 reading/window). The sensor acceleration signal, which has gravitational and body motion components, was separated using a Butterworth low-pass filter into body acceleration and gravity. The gravitational force is assumed to have only low frequency components; therefore a filter with 0.3 Hz cutoff frequency was used. From each window, a vector of features was obtained by calculating variables from the time and frequency domain.

For reach record in the data set the following features were provided:

- Triaxial acceleration from the accelerometer (total acceleration) and the estimated body acceleration.
- Triaxial angular velocity from the gyroscope.
- A 561-feature vector with time and frequency domain variables.

- Activity label.
- An identifier of the subject who carried out the experiment.

For our analysis we used the data consisting of 7,352 samples, each with 563 features (including the activity label and subject identifier). The Samsung activity data used for this analysis are slightly processed to make them easier to load into R [**2**] and were downloaded from a file sharing website [**2**]. The raw data and a data set description can be obtained from the UCI Machine Learning Repository [**2**].

Exploratory Analysis

Exploratory analysis was performed by examining tables and plots of the observed data. We identified transformations to perform on the raw data on the basis of plots and knowledge of the scale of measured variables. Exploratory analysis was used to:

- Determine whether any missing values were present in the data – none were found.
- Verify the quality of the data.
- Determine correlations between selected variables. We found a number of highly correlated variables such as tGravityAccmaxX and GravityAccmeanX, tBodyAccmaxY and fBodyAccMagstd, etc.
- Convert the character vector for the *activity* variable to a factor class for ease in building statistical models.

Statistical Modeling

We originally used hierarchical clustering to conduct feature engineering in order to determine the most viable sets of variables for the purpose of predicting activity. We used the mean value set for the body acceleration 3-axial signals (tBodyAcc-XYZ) in the X, Y and Z directions, however, the resulting cluster dendrogram did not show conclusive clusters. Maximum acceleration proved to be a better choice for yielding clusters.

We then turned our attention to the use of classification trees to address the non-linearity of the problem domain. We chose to include the entire set of feature variables in fitting the classification tree model, while having *activity* as the outcome variable.

III. RESULTS

In order to train the decision tree model, we included data from subjects 1, 3, 5 and 6 in the training set. For the test set upon which the prediction was made, subjects 27, 28, 29 and 30 were used. The training and test sets did not overlap in subject data. The decision tree was constructed using the 'rpart' [**2**], the recursive partitioning tree package for R. Initially, a tree with 8 leaves (terminal nodes) resulted. We then considered whether the tree could be simplified by pruning. The rpart package uses a *complexity parameter* that controls how much to trim the tree. The plot in Fig. 1 shows an optimal complexity parameter of 0.056 should be used for pruning.

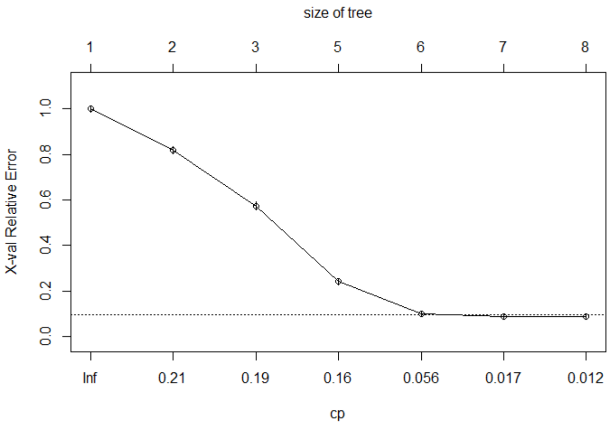

FIG. 1. Determining the optimal complexity parameter value shows 0.056 can be used to prune the tree.

The pruned tree is shown in Fig. 2. Notice that each *activity* is represented by a leaf node. The optimized tree shows that the following feature variables contribute to the predictive classification power of the model:

- tBodyAccmaxX – maximum body acceleration time domain signal in the X direction.
- tGravityAccmeanX – mean gravitational acceleration time domain signal in the X direction.
- tGravityAccminY – minimum gravitational acceleration time domain signal in the Y direction.
- tGravityAccmaxY – maximum gravitational acceleration time domain signal in the Y direction.
- fBodyAccMagstd – standard deviation body acceleration frequency domain signal.

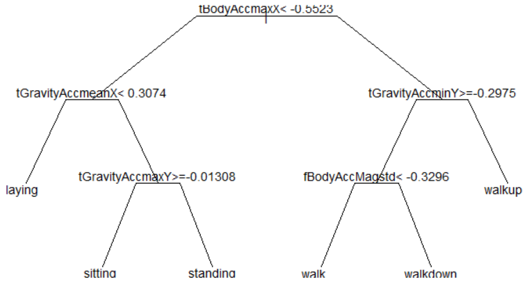

FIG. 2. Pruned classification tree used to predict activity based on inertial body signals of the subject.

The dual plot shown in Fig 3 indicates on the left-hand side that the 5th split offers the most information. The (1 – apparent error) and (1 – relative error) show how much is gained with additional splits. On the right-hand side of interest is the smallest error. As is often true with modeling, simpler is usually better [**2**].

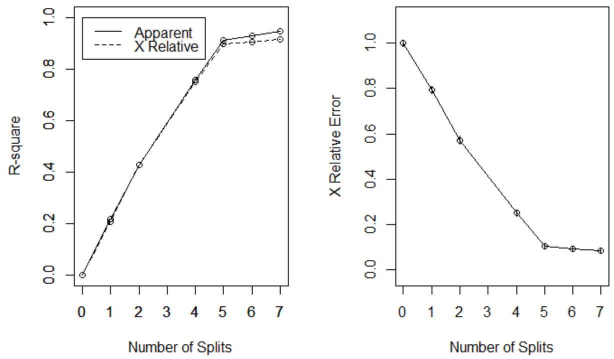

FIG. 3. The right-hand plot is a Relative Error Error(cross-validation) +/- 1-SE from cross validation versus the number of splits while the left-hand plot is the R^2 versus number of splits (apparent and apparent - from cross validation).

The confusion matrix in Table 1 shows the resubstitution errors from the pruned classification tree (fewer leaves), indicating good accuracy for the prediction model although it may make a few more mistakes than the full model with all leaves. So even though the pruned model may not be optimal for the observed data set, the cross validation results suggest that on a new data set the model will perform better that the prediction we would get from the full model. The reason is that the simpler model that is less perfectly tuned (having less overfit) to the training data set will predict better on new data.

```
          laying sitting standing walk walkdown walkup
 laying      221       0        0    0        0      0
 sitting       0     159       39    0        0      0
 standing      0       0      227    0        0      0
 walk          0       0        0  256        7      3
 walkdown      0       0        0    7      183      3
 walkup        0       0        0    8       28    174
```

TABLE 1. Confusion matrix of the classification results for activity on the test data set. Each column represents instances in the predicted class, which each row represents instances in the actual class. The values on the diagonal indicate correctly predicted activities. The values off the diagonal are mistakes in prediction.

We also performed a tree bagging process using the ‘ipred’ [**2**] package for R. The out of bag estimate of misclassification error was a low 2.5%.

IV. CONCLUSIONS

Our analysis suggests the following conclusions:

- The proposed classification tree model was able to predict human activity in the test data set (new data).
- Overfitting was not an issue when moving from the training set to test set.
- The confusion matrix describing the results showed acceptable accuracy.

An extension of these results for predicting human activity would be to utilize an ensemble approach. Ensembles provide the means to combine multiple classifiers by averaging/voting to increase accuracy.

1. [**1**] R Core Team (2012). ”R: A language and environment for statistical computing.” URL: http://www.R-project.org.
2. [**2**] https://spark-public.s3.amazonaws.com/dataanalysis/samsungData.rda%22%3ELink, retrieved 2/25/2013.
3. [**3**] UCI Machine Learning Repository, http://archive.ics.uci.edu/ml/datasets/Human+Activity+Recognition+Using+Smartphones.
4. [**4**] Ripley, B. (2012), Package ‘tree’. Available at CRAN: http://cran.r-project.org/web/packages/tree/index.html.
5. [**5**] Hastie T, Tibshirani R, and Friedman J: The Elements of Statistical Learning, Springer, 2001
6. [**6**] Peters, A., and Hothorn, T. (2012), Package ‘ipred’. Available at CRAN: http://cran.r-project.org/web/packages/ipred/index.html.