Title: A study of Human Activity Recognition Using Smartphones Data Set

Introduction:

Smartphones are very popular nowadays. Using their embedded accelerometer and gyroscope, quantitative data of people's body activity can be collected. These data help a lot in human activity recognition.

In this study, body activities are classified into six groups (walking, sitting, standing, laying, walking upstairs, walking downstairs). We try to build a **classification model/function** that predicts what activity a subject is performing based on these quantitative measurements. In our exploratory experiment, 93% of human activity can be correctly recognized.

Methods:

Data Collection

Our dataset came from a Human Activity Recognition project [1]. Find more details about how activities were measured and the dataset were built there.

The dataset has 7352 records. All records are complete, with on missing value. Each record has:
- A 561-feature vector with time and frequency domain variables. They are triaxial acceleration from the accelerometer and the estimated body acceleration, and the triaxial Angular velocity from the gyroscope.
- Its activity label (walking, sitting, standing, laying, walking upstairs, walking downstairs), which tells what the subject was performing.
- An identifier of the subject who carried out the experiment.

For this analysis, our training set includes the data from subjects 1, 3, 5, and 6. Our test set is the data from subjects 27, 28, 29, and 30.

Data Munging

Some data munging procedures are performed 1) variable ***activity*** is coerced to factor; 2) in order to rename duplicate variable names and eliminate special characters in variable names, all 561 numeric variables (except ***subject*** and ***activity***) are renamed to ***X***n, where n is from 1 to 561, indicating the sequence of the variables in each record.

Statistical Modeling

Since this is a typical classification problem [2], our final goal is to find out an effective classifier. Here, we use **Misclassification Error rate** to measure the accuracy of the prediction.

During the study, besides the classifier, we also have interests in following things:
- try different algorithms quickly and pick the one who perform better with this problem.
- find out whether feature selection can help or dimension reduction is necessary in this situation.

Results*:*

Since the dataset is not very big, we can perform different machine learning algorithms and find out which algorithm can perform better very quickly.

Here are summaries of algorithms we tried on the training data, and measured on the testing set.

1) One vs. all classification using Logistic Regression [3]
As the first step, we want to distinguish activity laying from all other 5 activities, so a new logical variable is added, which is true if the activity is laying and false for all others. Then we try the algorithm with 25, 50, 500, and 1000 iterations.

Result: the algorithm can't converge even with 1000 iterations.

So, we think much more training data, say 20000 (approximately 561 * 40), have to be collected for logistic regression to train an acceptable classifier. Another possible way is to reduce the number of variables. We stop working on this method and move to the next one.

2) Bagging of trees [4]
We run the algorithm to build bagging classification trees with 25 bootstrap replications.

Result: Misclassification Error rate is 0.1420875 on testing dataset.

3) SVM [5]
SVM are performed with different gamma parameters. A lower gamma can decrease misclassification error slightly, but too low gamma, say 0.0005, make things worse.

Result: Misclassification Error rate is 1) 0.0989899 with gamma 0.001782531. 2) 0.0962963 if gamma is 0.001.

4) Random Forests [6]
We run random forests algorithm on the training and grow a forest with 500 trees.

Result: Misclassification Error rate is 0.07272727.

This algorithm grows multiple trees and do majority vote for the result. It bootstrap samples for each tree and bootstrap variables at each split. So, it, growing 500 trees, runs much faster than bagging of 25 trees. Error rate decreases dramatically with increasing of number of trees, especially from 1 to 100 trees, and trees after 350 don't do obvious contributions. See figure 1.

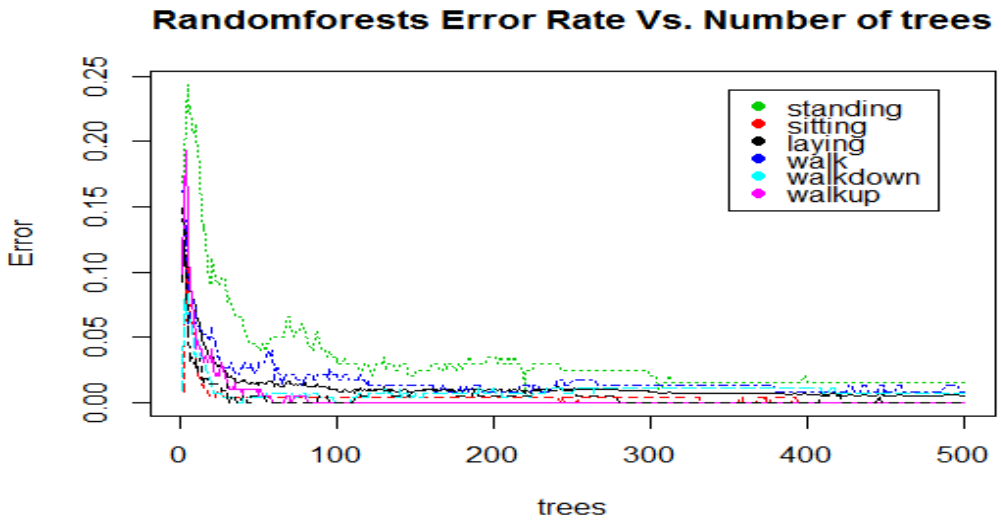

Figure 1 - Error Rate of Randomforests

So far, Randomforests performs better than other algorithms we tried. We would like to dig deeper with Randomforests.

5) Random Forests after removing several outliers
Here we want to know whether outliers have obvious impact on the classifier. We remove the top 4 outliers (934,935,648, and 45) from the training dataset and run the algorithm again. Figure 2 shows random forests' outlier measures.

Result: Misclassification Error rate is 0.07272727.
The result is identical to that in the previous experiment. So, in this classification case, we think Randomforests can handle outliers properly.

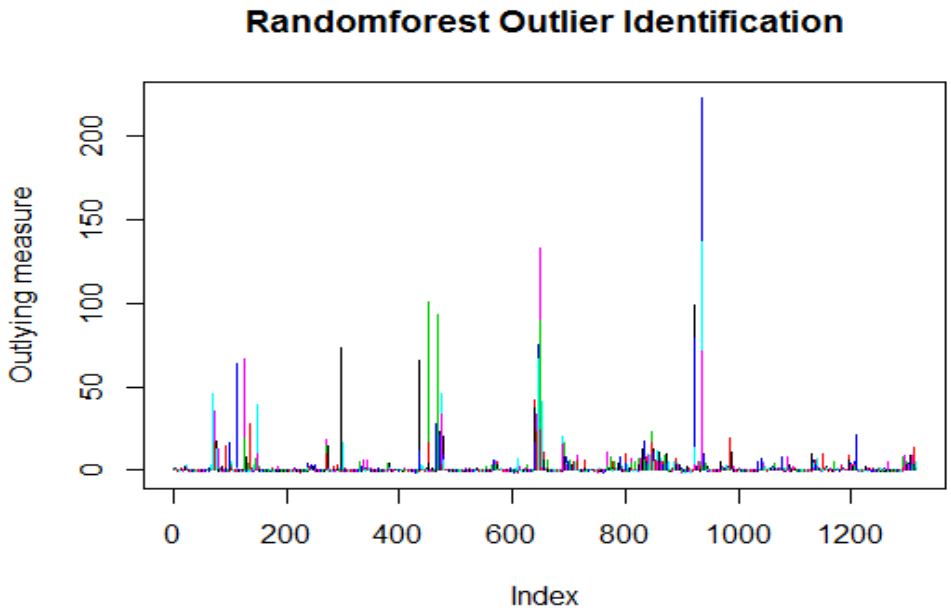

Figure 2 – Randomforest outlier measures

6) Random Forests with top 200 features
Intuitively, all measured variables contain more or less information. If we have enough samples and acceptable running performance, dimension reduction is not quite necessary. But, we want to check whether random forest can perform better with less features.

We first apply SVD[7] to the original data set, and choose 250 features which can keep 99.75% of variance. We run random forests again on the new dataset.

Result: Misclassification Error rate is 0.07205387.
The misclassification error rate is only slightly less than that of experiment 4. So, it's hard to say that dimension reduction can make randomforests better.

Conclusions*:*

In our experiment, randomforests algorithm outperforms others, and 93% of human activities are correctly recognized. Our experiment also shows that it's robust to outliers, and feature picking (or

dimension reduction) is not necessary when running performance is not a problem. Further study can continue to tune the algorithm, like tuning the number of variables selected at each node.

References

1. Human Activity Recognition Using Smartphones Data Set
http://archive.ics.uci.edu/ml/datasets/Human+Activity+Recognition+Using+Smartphones
2. Classification, WikiPedia, http://en.wikipedia.org/wiki/Classification_(machine_learning)
3. Logistic Regression, Wikipedia, http://en.wikipedia.org/wiki/Logistic_regression
4. Bagging, Wikipedia, http://en.wikipedia.org/wiki/Bootstrap_aggregating
5. SVM, Wikipedia, http://en.wikipedia.org/wiki/Support_vector_machine
6. Random Forests, Wikipedia, http://en.wikipedia.org/wiki/Random_forest
7. SVD, Wikipedia, http://en.wikipedia.org/wiki/Singular_value_decomposition