Analytics: Human Activity Recognition Using Smartphone data

**Purpose**

The exact purpose is to predict the human activity such as standing, sitting, lying, walking, walking downstairs, and walking upstairs using bagged SVM model with feature selection using Random forest and scaled data. The data is initially rescaled, then feature selection is done using Random forest – select features that when removed improves the error rate, and finally the result from bagged SVM using rescaled data and selected features reduces over fitting and improves accuracy.

**State-of-the-art**

Who does this now? How well? Could it be done better? (3-5 pages or more).

I rescaled the features by subject. Everyone walks, stands, sits, lays, walks upstairs and walks downstairs in his own way, so we need to handle this information to be able to generalize the outcome. The easiest way is to rescale features by subject.

Remove outlier subjects with accelerometer measurements that are outliers and those that cannot be generalized in the training data. For this random forest is used to get the accuracy of prediction and based on a predetermined prediction accuracy threshold the subjects are eliminated.

The bagging of SVM gives me a good prediction accuracy.

The model can further be improved by fixing known activity prediction errors.

**Method**

What is your plan to address the problem? What are the hardest parts (bottleneck analysis)? How can you mitigate risk? (e.g., an alternate simpler analytic method, or a narrowing of scope). (3-4 pages or more)

The toughest part is to optimize the parameters such as “mtry” for random forest and setting the threshold for eliminating the subjects based on prediction accuracy.

As the above settings affect the overall accuracy of the model they need to be carefully selected and for that I have selected the best “mtry” parameter for random forest by selecting “mtry” that gives the least error. The threshold for feature selection was done by trying various values of cutoff and selecting the best cutoff value.

Bagging of SVM was done 27 times for prediction.

**Data**

The raw data for this analysis are available from the Machine Learning Repository at the University of California at Irvine http://archive.ics.uci.edu/ml/datasets/Human+Activity+Recognition+Using+Smartphones

There were a number of duplicate column names in the downloaded data set (the ‘bandsEnergy’ variables) – these were renamed in a unique manner. Other variable names included parentheses, dashes or commas, all of which caused problems with certain R packages. These illegal characters were removed or changed to allowed characters.

The data set has been cleaned before publication, since there were no missing values in any channel.

Each subject had observations of each of 6 activity classes: laying, sitting, standing, walk, walk downstairs, and walk upstairs. Each observation consists of 561 accelerometer measurements, including max, min, mean and many others in each of three (XYZ) directions. The training data consists of 562 columns. The 562nd column is the activity class name. The test data set consists of 561 accelerometer measurements. Measurements have been normalized, with all values of all channels lying between -1.0 and 1.0.

This analysis attempted to predict the activity from the accelerometer measurements using bagged SVM.

The data set was divided into training and testing subsets. The training set consists of 6107 observations from subjects 1, 2, 3, 4, 6, 9, 10, 12, 13, 15, 16, 19, 20, 21, 22, 24, 29, and 30 and the test set consists of 4194 observations from subjects 5, 7, 8, 11, 14, 17, 18, 23, 25, 26, 27, and 28.

Where can data be found to perform your analysis? What type of data? Is it consistent with the method? (e.g. If Wipro has a data base of past SW projects, including success rates, delays if any, and differences in SW development methodology, you have an ideal data set to learn correlations and best practices, but if no such data is available, then this project may not be doable. Usually one is somewhere between ideal and no data). (2-3 pages to describe data, provenance, etc.)

**Results**

What happened when you applied your methods to the data? (e.g. were there any correlations, or predictions?) How can you tell the accuracy (e.g. cross-validation methods)? Just present the results including effective visualization techniques, and do significance testing, if possible. (10 pages or more)

The variable ‘tBodyAcc\_max\_X’ (maximum body acceleration in the X direction) can be used to separate the ‘walking’ activities from the stationary activities. Within the stationary activities, ‘laying’ can easily be distinguished using the ‘tGravityAcc\_mean\_X’ variable (the component of gravitational acceleration aligned with the X direction). This makes perfect sense if the X axis is aligned with the direction the subject is facing. The walking activities involve motion (and hence acceleration) in the X direction while the stationary activities do not. Furthermore, if the subject is laying on his/her back, the X axis is now pointing opposite to the acceleration of gravity, so negative values of ‘tGravityAcc\_mean\_X’ would be expected. The ‘tBodyAcc\_max\_X’ and ‘tGravityAcc\_mean\_X’. These two variables neatly distinguish the ‘laying’ category, and separate ‘sitting/standing’ from the walking activities.

Together with rescaled data, feature selection using random forest and bagging of SVM gives me accuracy of 96% on human activity predictions of test data.

**Analysis**

The random forest method was successful in helping me with feature selection.

The bagged SVM model gives me an accuracy of 96%. This accuracy can be further enhanced by fixing known errors on the test data.

What do the results tell you? (e.g. need to collect more data, or conclusive correlations, or promising matches between resumes and job posting, but need better text mining techniques to address equivalent expressions (e.g. “expert in Java” with “10+ years experience in Java”, or whatever conclusions, shortcoming, improvements, and lessons learned). (10 pages or more)

**Appendices**

Please attach complete source code, with proper commenting and indenting.

The source code and primary\_set texts and secondary\_set texts are attached to the e-mail.