IS624 - Final Project

Aaron Palumbo

Friday, July 17, 2015

Objective

Sensor data is all around us. From the Gartner website: >The Internet of Things (IoT), which excludes PCs, tablets and smartphones, will grow to 26 billion units installed in 2020 representing an almost 30-fold increase from 0.9 billion in 2009, according to Gartner, Inc. Gartner said that IoT product and service suppliers will generate incremental revenue exceeding $300 billion, mostly in services, in 2020. It will result in $1.9 trillion in global economic value-add through sales into diverse end markets. The ability to make sense of the data streaming from all these new devices is a key aspect of the growth in this industry.

The UCI dataset, "Human Activity Recognition Using Smartphones Data Set" is "built from the recordings of 30 subjects performing activities of daily living (ADL) while carrying a wasit-mounted smartphone with embedded inertial sensors." Our goal is to be able to predict user activity based on that sensor data.

From IEEE Volume: 15 Issue: 3, Q3 2013, there is a paper surveying state-of-the-art performance in HAR (Human Activity Recogintion). Table VII performance measurements for several model techniques:

TABLE VII
SUMMARY OF STATE-OF-THE-ART IN ONLINE HUMAN ACTIVITY RECOGNITION SYSTEMS.

Reference	Activities	Sensors	ID	Obtrusive	Experiment	Energy	Flexibility	Processing	Features	Learning	Accuracy
Ermes [44]	AMB (5)	ACC (wrist, ankle, chest)	PDA	High	N/S	High	SPC	High	TD, FD	DT	94%
eWatch [24]	AMB (6)	ACC, ENV (wrist)	Custom	Low	LAB	Low	MNL	Low	TD, FD	C4.5, NB	94%
Tapia [22]	EXR (30)	ACC (5 places), HRM	Laptop	High	LAB	High	Both	High	TD, FD, HB	C4.5, NB	86% (SD), 56% (SI)
Vigilante [46]	AMB (3)	ACC and VS (chest)	Phone	Medium	NAT	Medium	MNL	Low	TD, FD, PR, TF	C4.5	92.6%
Kao [23]	AMB, DA (7)	ACC (wrist)	Custom	Low	N/S	Medium	MNL	Low	TD, LDA	FBF	94.71%
Brezmes [31]	AMB (5)	ACC (phone)	Phone	Low	N/S	Low	SPC	High	TD, FD	KNN	80%
COSAR [30]	AMB, DA (10)	ACC (watch, phone), GPS	Phone	Low	NAT	Medium	MNL	Mediun	TD	COSAR	93%
ActiServ [29], [32]	AMB, PHO (11)	ACC (phone)	Phone	Low	N/S	Low	SPC	High	$\bar{y}, \sigma_y^2$	RFIS	71% - 98%

Our goal will be to test XX models against these data and compare that against the table presented above. We will also discuss the steps and choices involved in cleaning the data, feature selection, model tuning, and final model selection.

Background Information

From the data dictionary:

> 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 readings/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. See 'features_info.txt' for more details.

This means that we are not dealing with time domain vectors. Each observation represents 561 measurements of a 2.56 second window. We will attempt to ascertain what activity the user was engaged in from this 2.56 second window.

Environment / Dependencies

Dependencies

```
library(knitr)
# Suppress messages and warnings in all chuncks
opts_chunk$set(message=FALSE, warning=FALSE)

# Libraries
library(caret)
library(randomForest)

# Processing
library(doParallel)
registerDoParallel(cores = 4)
```

Working Directory / File Paths

```
## Working Directory
proj_dir <- "IS624_Final_Project"
if(basename(getwd()) == proj_dir){
  setwd("./code")
}

if(!(basename(getwd()) == "code")){
  break
}

## File Paths
dataDir <- "../data"
trainDataDir <- file.path(dataDir, "UCI HAR Dataset", "train")
testDataDir <- file.path(dataDir, "UCI HAR Dataset", "test")
```

Custom Functions

Data Partitioning

According to the file README.txt, the data was randomly partitioned into two sets where 70% of the volunteers were selected for generating the training data and 30% the test data. Splitting the data this way, by volunteer, is a good idea in that we want our algorithm to be general enough to work from person to person, not just be good at identifying activities of a particular person. This would offer a good starting point for a device of this type. After it is purchased, it might make sense to tune the parameters to the specific person. We will tune all our models on the training set, and then run a final comparison on the test set.

Load Training Data

```
## Sensor Data
df.sns <- read.csv(file.path(trainDataDir, "X_train.txt"),
                   header=FALSE, sep="")
snsColNums <- 1:ncol(df.sns)

features <- readLines(file.path(dataDir, "UCI HAR Dataset", "features.txt"))

## Subject ID
sub <- as.factor(readLines(file.path(trainDataDir, "subject_train.txt")))

## Activity ID
act <- as.factor(readLines(file.path(trainDataDir, "y_train.txt")))
```

Load Test Data

```
## Sensor Data
df.sns.test <- read.csv(file.path(testDataDir, "X_test.txt"),
                        header=FALSE, sep="")
# snsColNums <- 1:ncol(df.sns)

## Subject ID
sub.test <- as.factor(readLines(file.path(testDataDir, "subject_test.txt")))

## Activity ID
act.test <- as.factor(readLines(file.path(testDataDir, "y_test.txt")))
```