STAT 542 Final Project Proposal

Wenzhao Xu, Haoyan Cai

November 17, 2013

1 Introduction

The data are about accelerometer data from mobile device. The train data contain X,Y,Z acceleration data from 387 devices and testing data set have 90024 testing sequences. In addition, for each test sequence, we have information of a possible device. The goal is to judge whether this possible device is the true device that generate the test sequence.

2 Data Set

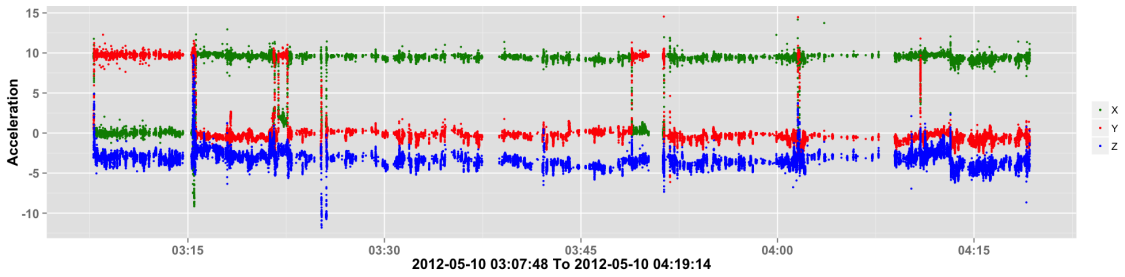
Package "ff" is used to extract data from a large csv file and unix time is converted to GMT time. There are totally 387 training devices, and 90024 testing sequences. A typical visulization of data is shown in Figure 1, in which Device 7 is a device with long sampling time and total 523187 points while Device 770 has limited sampling points as 28475.

In test data, we need to identify the professed device is the true number. Figure 2 shows two sequence whose professed device is 770. Compared with the train data with device 770, it seems sequence 838966 is likely belong to Device 770 while sequence 690194 is not since the range of X,Y,Z in sequence 838966 is in consistnace with training data of device 770.

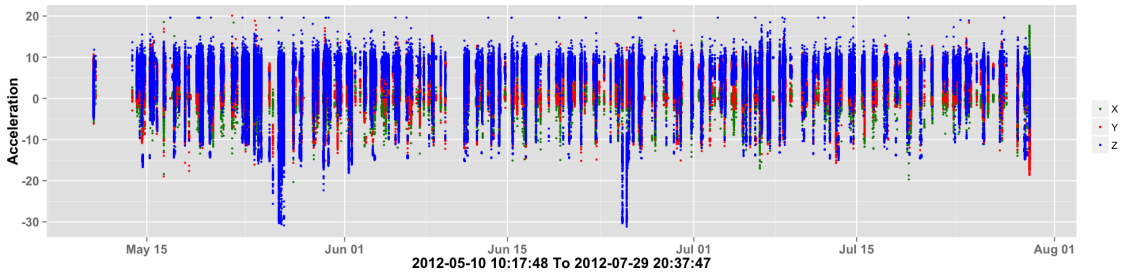
3 Potential Approach

The data involve with time dimension, which is a very important feature. First, we need to define features from the raw data. Some simple but might be useful features include the range of X,Y,Z, the difference between X and Y, between X and Z, and between Y and Z, autocorrelation of X,Y and Z, and correlations between X,Y and Z, etc. We plan to do some literature research about what activities the data might reveal in order to extact features that represent users personal habbit such as how he/she runs. Then,an assumption should be made that users activities have similarity. Similarity means users are likely to have similar behavoir during similar time of day. For example, users would have similary X,Y and Z time series during today's 4:00pm and during yesterday's 4:00pm.

One big problem is that how we define the probability for a test sequence to be rejected. First, since data are generated by different time, training data might have different patterns during different time period. Given a test sequence, denote its day time period as T_{test}. Then search all the training data and extract the data that generated at the similar day time. For a test sequence, a professed device D is known. So we can label the data from device D as 1 and the data from other device as 0. Train the model with these training data and get the classifier. Then fit the test sequence to see wether it is 0 (test sequence is belong to other device) or 1 (test sequence is belong to device D). If possible, identify what activity the test sequence represents such as running or walking or sleeping. If the test sequence is from running,then find the device with similar running patterns.

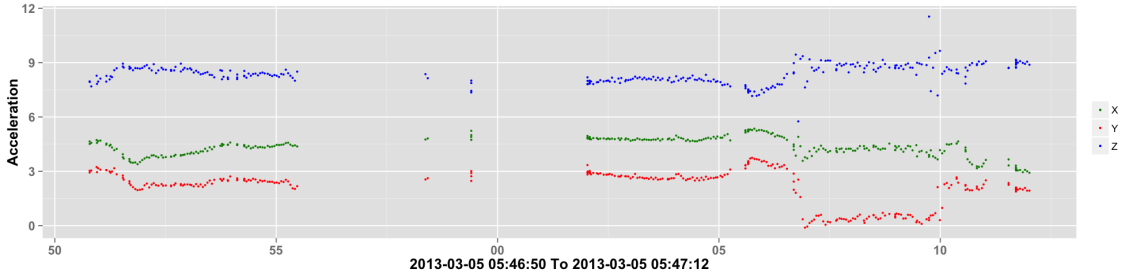


(a) Device 770

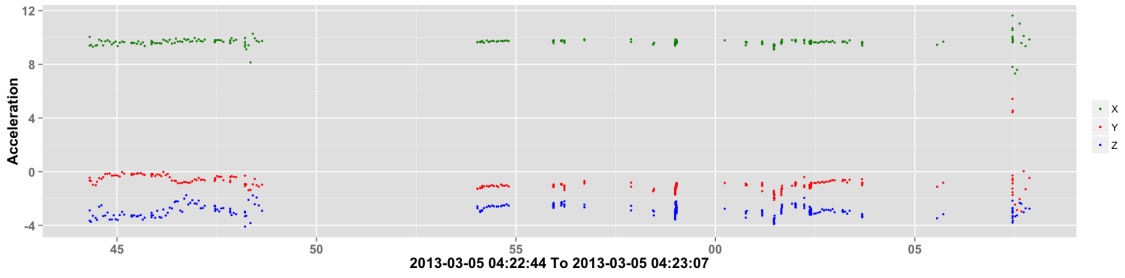


(b) Device 7

Figure 1: Acceleration Data Along Time



(a) Sequence 690194



(b) Sequence 838966

Figure 2: Test Sequence with Professed Device as 770