STAT 542 Final Project Proposal

Wenzhao Xu (Environmental Engineering)
Haoyan Cai (Statistics)

November 18, 2013

1 Introduction

The data are about accelerometer data from mobile devices. The train data contain X,Y,Z axis acceleration values from 387 devices and testing data set consists of 90024 testing sequences. Each test sequence comes with a proposed device Id. The goal is to judge whether the claimed device is the true device that produces the test sequence.

2 Data Set

Package "ff" and SQL database can be used to import data from a large csv file. Unix time is then converted to standard GMT time. 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 of 28475.

In test data, each sequence contains 300 sampling points, roughly lasts for less than 1 minute. We need to identify whether the professed device in the test sequence is the true device. Figure 2 shows two sequence whose professed device are 770. Compared with the train data from device 770 (Fig 1(a)), 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 consistent with training data from device 770.

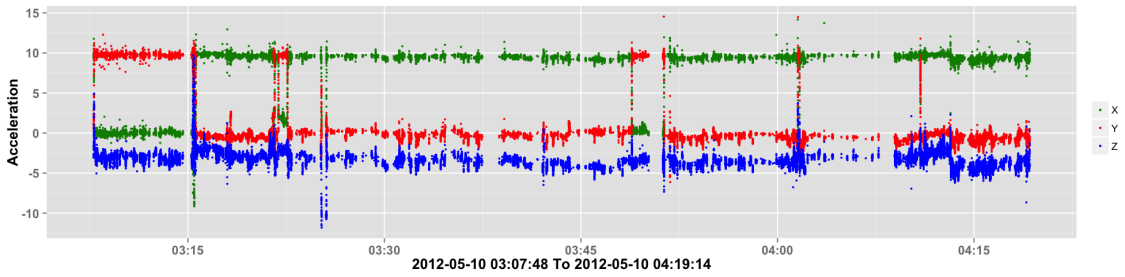
We also came up with Figure 3 and Figure 4 when we conducted exploratory data analysis to have more sense of the data set. In each 2 by 2 plot, the top left graph comes from training data for a given device and the other 3 are sampled from test sequences that are claimed also from the given device. Figire 3 shows another device that its training data and test sequences have gaps in sampling. In fact, such sampling gaps exist common in the whole data set. In Figure 4, one of the test sequences has very large variance (upper left figure) and Z axis acceleration values (blue dots) probably have outliers, which lies in the region occupied by red dots.

3 Potential Approach

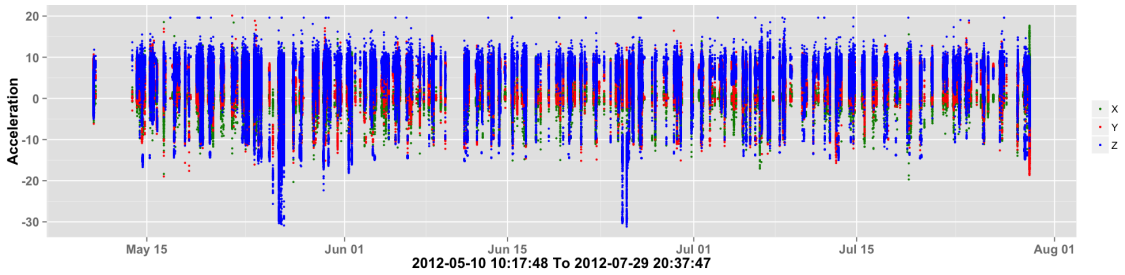
The data involve with time dimension, which is a very important feature. To solve this problem, an assumption should be made that users' activities have time-constraint re-occuring patterns, i.e. users are likely to have similar behavoir during roughly same time of day. For example, users would have similary X,Y and Z time series during today's 4:00 pm and yesterday's 4:00 pm.

Based on this assumption, we design the following approaches:

- First, appropriate features from the raw data should be defined. Some simple but probably 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,

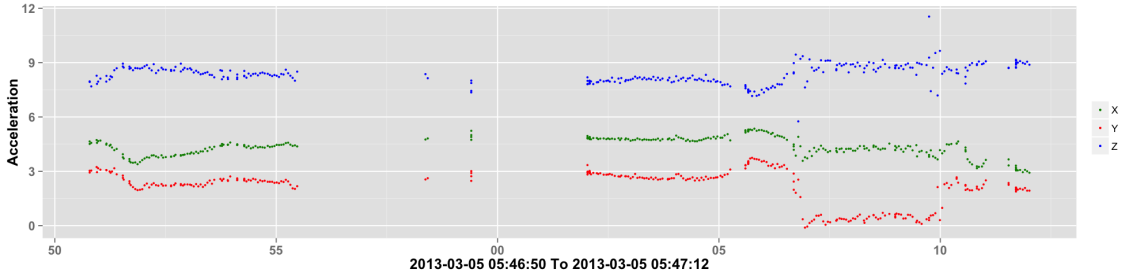


(a) Device 770

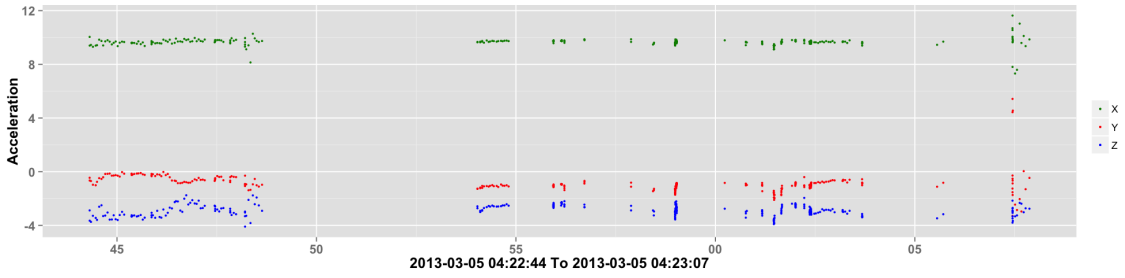


(b) Device 7

Figure 1: Acceleration Data Along Time



(a) Sequence 690194



(b) Sequence 838966

Figure 2: Test Sequence with Professed Device as 770

etc. In addition, We plan to do some literature research about what activities the data might reveal in order to extract features that represent users personal habbit such as how he/she runs.

- Second, Supervised or Unsupervised?
  - Treat it as a time-constrained classification problem, more specifically, a multi-classification problem, i.e. each device is a unique class. It seems that one-all should suffice this task. For a professed device Id, label 1 if true and 0 if its other device's Id. Since data are generated in different time intervals, training data might have different patterns during different time period. So given a test sequence X_{test}, we denote its time of day as T_{test} (e.g. $T_{test} = 19:00$). Then search all the training data and extract available training sequences (denote as X_{train}) that were fuzzily close to T_{test}. For the test sequence X_{test}, a professed device D is known. We can label the training sequences from device D as 1 and the training sequence from other device as 0. Based on the features we extract, train some classifiers with these training examples. Finally, apply the classifier to 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).

  - Learn a standardized ranking based probability metric indicating the similarity of a given test sequence and the associated training device. The heuristics behind this is if a test sequence comes from the claimed device, it should be most similar (defined by some similarity metrics like Euclidean distance, cosine similarity and so on) to this device instead of another. And the extracted features are only used to learn the weights of similarity features, which is different from a classification approach.

Note: identifying what activity the test sequence represents such as running or walking or sleeping is also important. If the test sequence is from running but the running pattern from professed device is totally different with what test sequence reveals, then the professed device is probably a wrong device.

4 Timeline

- Research on literatures with acceleration data
- Data preprocessing to remove potential outliers and extract important features
- Initial result done (before the end of thanksgiving holiday)
- Discussion with instructors, improve results and write up

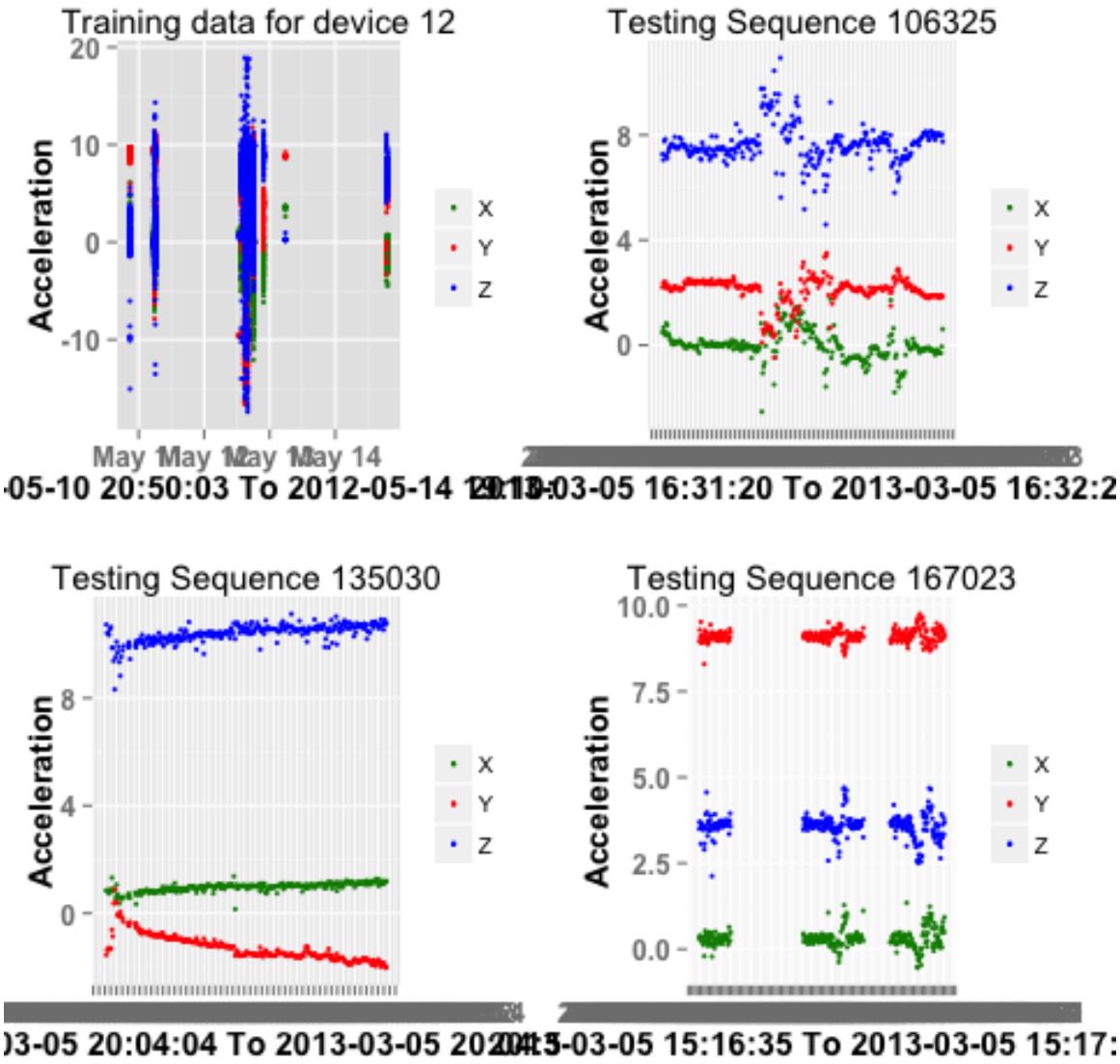


Figure 3: Training data and test sequence labeled as Device 12

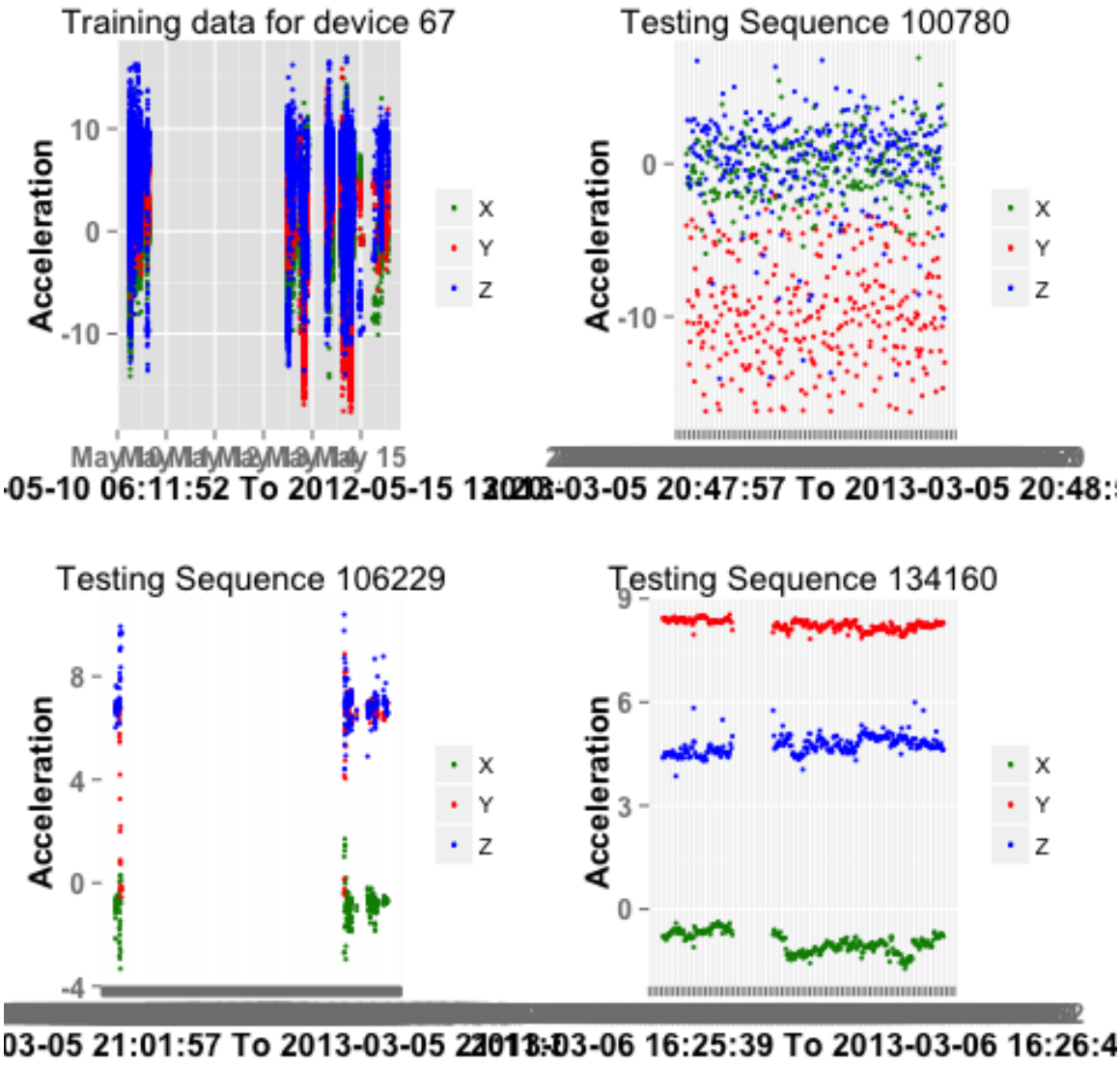


Figure 4: Training data and test sequence labeled as Device 67