HomeActivity: Recognizing home activities using sensor data

Stephen Lee, Patrick Pagus and Dong Chen

December 15, 2015

1   Abstract:

Home activities recognizing allows many smart applications including healthcare and energy efficiency.

2   Introduction

Activity recognition is the problem of determining an individual's or group's activities given some sensor data describing their actions. More formally, activity recognition finds a function that maps a sequence of sensor readings (feature vectors), to a sequence of activity labels that best describes the sensor data. We focus on activity recognition within home based wireless sensor networks. In this case, sensors may include contact, motion, temperature, and humidity sensors located on doors, appliances or monitoring certain rooms. This problem has many applications including healthcare and energy efficiency, where other sensors, such as cameras and wearables, intrude upon privacy or are too expensive or inconvenient. In healthcare, home activity recognition can be used to assess the cognitive and physical capabilities of an elderly person. Home energy efficiency can be improved given a schedule of the occupant's activities by scheduling background loads, such as heating, cooling, and various appliances.

This domain has several challenges. First, there is a dearth of training data because labeling one's activities is a tedious time consuming process, this limits the number of effective techniques. Second, patterns in training data are specific to the home and the individual. Therefore, the performance of a machine learning technique may vary widely over different data sets. Finally, dominant classes may confound machine learning techniques.

To further our knowledge of machine learning, we will conduct a comprehensive empirical analysis of proposed methods in this domain, across a collection of datasets. These methods include Hidden Markov Models, Conditional Random Fields, and Support Vector Machines. We will apply these methods to at least four datasets that vary in duration, home size, and sensor counts, which are not necessarily proportional to home size. By ap-

plying these techniques to real-world datasets and explaining the differences and similarities in their performances, we will gain a deeper understanding of machine learning.

For our group project, we tackle the problem of activity recognition using available sensor information installed in homes. These sensors could be infrared motion detector sensors (to understand when a person entered a room or opened a fridge), to RFID sensors, to microphone based labeling of activities to record the ground truth. These datasets are usually available in the form of timestamp and the corresponding activity or sensor label. So, in this project we would like to predict a label (i.e. the activity) given a sequence of observations.

1. Interesting because it has applications that spans across various domains:

   (a) Healthcare
       - Long term monitoring of activities could provide us interesting insights on degenerative heath
       - Especially useful in monitoring health of elderly people

   (b) Energy savings
       - Understanding the pattern of activities could help us save energy.
       - For example, switching off AC/heater when no one is at home

   (c) From a Security perspective
       - Again, notifications can be sent to alert the home owners if there is an aberrant activity in the door entrance.

   (d) Intelligent Homes
       -

2. Challenging

   (a) Sensor datasets could be noisy
       - the front doors may open and close multiple times; this doesn't may not indicate that the person has left the home.
       - sometimes the sensors itself may have false positives.

   (b) multiple labels may be mapped to a single sensor activity
       - opening of a fridge may mean both; make coffee and make cooking
       - some other information such as how long the person spent in the kitchen may provide some useful insight

(c) Imbalanced class problem
- machine learning algorithms and works best when the number of instances of each classes are roughly equal
- where the total number of a class of data (positive) is far less than the total number of another class of data (negative)
- In this problem, no activity is a perfectly reasonable label. So the prediction model may always output 'no activity' and be accurate 80% of the time.

3   Recognizing Methods

We would like to explore the following techniques:

1. HMM - Hidden Markov Modes (HMM) are a buiqutious tool for modelling time serises data. HMMs can be used as black-box density models on sequences. They have the advantage over Markov models in that they can represent long-range dependencies between observations, mediated via the latent variables.

2. CRF - CRF is a discriminative undirected probabilistic graphical model naturally applicable to sequence labeling tasks. A conditional random field or CRF (Lafferty et al. 2001), sometimes a discriminative random field (Kumar and Hebert 2003), is a version of an MRF where all the clique potentials are conditioned on input features. The advantage of a CRF over an MRF is analogous to the advantage of a discriminative classifier over a generative classifier (see Section 8.6), namely, we dont need to waste resources modeling things that we always observe. Instead we can focus our attention on modeling what we care about, namely the distribution of labels given the data. Another important advantage of CRFs is that we can make the potentials (or factors) of the model be data-dependent. For example, in image processing applications, we may turn off the label smoothing between two neighboring nodes s and t if there is an observed discontinuity in the image intensity between pixels s and t. Similarly, in natural language processing problems, we can make the latent labels depend on global properties of the sentence.

3. SVM - The implementation used in this experiment comes from SVM

4. SSVM -

5. Naive Bayes

The idea of the naive Bayes classifier is to use a generative model of text to estimate
The datasets available to us are:

1. Kasteren dataset - has over a month long sensor information from 3 homes.

2. Tulum dataset - more than six month period from a single home

3

4   Timeline for finishing the project.

- 2.5 weeks for implementing these modeling methods.
- 1 week for analyzing with different dataset; and writing it up. -

5   Experimental Evaluation

5.1   Datasets

5.2   Experimental Setup

5.3   Comparative Analysis Metrics

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2}$$

$$F - Measure = \frac{2 \cdot Precision \cdot Recall}{Precison + Recall} \tag{3}$$

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP) \cdot (TP + FN) \cdot (TN + FP) \cdot (TN + FN)}} \tag{4}$$

5.4   Experiments

Activity recognition is the problem of determining an individual's or group's activities given some sensor data describing their actions. More formally, activity recognition finds a function that maps a sequence of sensor readings (feature vectors), to a sequence of activity labels that best describes the sensor data. We focus on activity recognition within home based wireless sensor networks. In this case, sensors may include contact, motion, temperature, and humidity sensors located on doors, appliances or monitoring certain rooms. This problem has many applications including healthcare and energy efficiency, where other sensors, such as cameras and wearables, intrude upon privacy or are too expensive or inconvenient. In healthcare, home activity recognition can be used to assess the cognitive and physical capabilities of an elderly person. Home energy efficiency can be improved given a schedule of the occupant's activities by scheduling background loads, such as heating, cooling, and various appliances.

This domain has several challenges. First, there is a dearth of training data because labeling one's activities is a tedious time consuming process, this limits the number of effective techniques. Second, patterns in training data are specific to the home and the individual. Therefore, the performance of a machine learning technique may vary widely over different data sets. Finally, dominant classes may confound machine learning techniques.

To further our knowledge of machine learning, we will conduct a comprehensive empirical analysis of proposed methods in this domain, across a collection of datasets. These methods include Hidden Markov Models, Conditional Random Fields, and Support Vector Machines. We will apply these methods to at least four datasets that vary in duration, home size, and sensor counts, which are not necessarily proportional to home size. By applying these techniques to real-world datasets and explaining the differences and similarities in their performances, we will gain a deeper understanding of machine learning.

6   Discussion and Future Work