Inferring Human Mobility Patterns from Taxicab Location Trace

Presented By: Vedika Babu
Index

- INTRODUCTION
- EXISTING MODEL
- PROPOSED MODEL
- EVALUATION
- APPLICATIONS
INTRODUCTION

Problem
Why Improve?
Dataset
PROBLEM STATEMENT

To identify trip(s) made by taxicab. Such location traces are a rich source of information and can be used for:

- congestion pricing
- taxicab placement
- Improved city planning

The use of a graph theory concept - *stretch factor* in a novel manner with a Hidden Markov Model based algorithm can identify trips.
DATASET

Two independent taxicab based sensor deployments in two cities:

- Stockholm (Sweden)
- Shanghai (China).

Each sample comprises of $x_i$

- taxi ID
- location (latitude/longitude)
- timestamp (UTC ms).
<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Stockholm</th>
<th>Shanghai</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sampling rate</td>
<td>1/min</td>
<td>2/min</td>
</tr>
<tr>
<td>Number of cabs</td>
<td>~ 2000</td>
<td>~ 10,000</td>
</tr>
<tr>
<td>Privacy</td>
<td>No sampling when taxi hired</td>
<td>None</td>
</tr>
<tr>
<td>Timeline</td>
<td>1 month</td>
<td>1 month</td>
</tr>
<tr>
<td>Total number of trips</td>
<td>570,690</td>
<td>1,335,360</td>
</tr>
</tbody>
</table>
EXISTING METHOD

Speed Based Classification
Clustering Based Classification
SPEED BASED CLASSIFICATION

- Identify end point from continuous location samples.
- Check avg speed in a time window < threshold.
- Normalize the speed of a taxicab by historical speeds of all taxicabs on neighboring road segments.
  1. compensate for different speed limits in different road segments,
  2. remove false positives – misclassifying slow down due to traffic light or stop signs at a junction as a candidate trip end-point
  3. ensure that the normalization is robust against transient congestion in road traffic.
- SVM to classify over a taxi cab’s normalized speed. Classifier labeled each $x_i$ as an endpoint or not.

- Recall and precision is 0.16. Reason for such low precision and recall is that:
  - Average speed often fails to be strong enough differentiator between slow moving traffic and endpoints of a particular trip.
Clustering Based Classification

- Expected number of cabs situated at a trip’s endpoint is likely to be larger than that in the immediate neighborhood.
- The count on the number of taxicab sightings in a spatial bounding box $b$ as $\text{count}(b)$
- $\text{count}(b) \geq \gamma \times \text{count(parent}(b))$, where $\gamma \leq 1$
- Spatial clusters based on the optimal parameter $\gamma^*$; subsequently, a SVM for classifying a cluster into a trip endpoint.
● SVM to classify after clustering is done. Classifier labeled each $x_i$ as an endpoint or not. Recall and precision of 0.25.
● Hidden semi markov model to classify after decision tree based classifier is used. Recall and precision of 0.40-0.60.
(a) Trip $x_a$ to $x_b$ has no other endpoints

(b) Trip $x_a$ to $x_b$ has other endpoints
PROPOSED MODEL

Stretch Factor

Hierarchical Segmentation

Hidden Markov Model (HMM)

- HMM-Naive
- HMM-Stretch Factor

Post Processing the Output
Proposed Model - Flow Chart

- Taxi id, Location, Time
  - Stretch Factor
    - Hierarchical Segmentation
    - HMM-Naive
      - Post Process The Output
    - Hmm-Stretch Factor
STRETCH FACTOR:

**Distance Stretch Factor:**

\[
\text{Actual distance traversed on a given trip} \quad \frac{1}{\text{Shortest distance between the endpoints of that trip}}
\]

**Time Stretch Factor:**

\[
\text{Actual time traversed on a given trip} \quad \frac{1}{\text{Shortest time between the endpoints of that trip}}
\]
Shanghai Dataset: 10,000 trips

Stockholm: 2000 trips

(a) Distance stretch factor  
(b) Time stretch factor
Proposed Model - Flow Chart

\[\text{<Taxi id, Location, Time>} \rightarrow \text{Stretch Factor} \rightarrow \text{Hierarchical Segmentation} \rightarrow \text{HMM-Naive} \rightarrow \text{Post Process The Output}\]

\[\text{HMM-Stretch Factor} \]
HIERARCHICAL SEGMENTATION:

\((lx_1, lx_2), \ldots, (lx_i, lx_{i+1}), \ldots, (lx_{n-1}, lx_n)\) and progressively merges two overlapping shortest path segments.

The stretch factor of the actual path taken by the taxi \((lx_i, lx_{i+1}, \ldots, lx_k)\) is smaller than a threshold \(\beta\) (for some \(\beta \geq 1\)).

Two consecutive trips are combined if the combined trip has a stretch factor that is smaller than a given threshold.
Proposed Model - Flow Chart

< Taxi id, Location, Time > → Stretch Factor

Hierarchical Segmentation

HMM-Naive

HMM-Stretch Factor

Post Process The Output
HIDDEN MARKOV MODEL

A HMM is characterized by the following:

- $N$: the number of hidden states
- $M$: the number of distinct observation symbols per state
- $\alpha_{N \times N}$: state transition probabilities
- $B_{N \times M}$: Observation symbol probability distribution for each state (also, known as emission probability)
- $\pi_{N \times 1}$: initial state distribution.

$Y = \{o: \text{occupied}, u: \text{unoccupied}\}$. 
HMM Naive

The transition probabilities $\alpha_{u,u}$, $\alpha_{u,o}$, $\alpha_{o,o}$ and $\alpha_{o,u}$ is computed.

By the likelihood of transitioning between the occupied and unoccupied states based on whether the next point is a candidate endpoint or not.

The HMM-naive approach and use it as a baseline to measure the performance of our algorithm.
HMM Stretch Factor

Compute the likelihood of transitioning between the occupied and unoccupied states based on the stretch factor.

$$\Pr(\{x_{i-\Delta}, \cdots, x_i, \cdots, x_{i+\Delta}\}|y_i = o) \propto \frac{1}{sf(l_{x_{i-\Delta}}, \cdots, l_{x_{i+\Delta}})}$$

where

- for some tunable parameter $\Delta \geq 1$
- $sf(\cdot)$ denotes the stretch factor of a location trace
HMM Stretch Factor-Normalize

Normalize the returned value as:

\[
sf(\{x_{i-\Delta}, \cdots, x_i, \cdots, x_{i+\Delta}\}) / SF
\]

where

\[
SF = \sum_{\{x_{i-\Delta}, \cdots, x_i, \cdots, x_{i+\Delta}\}} sf(\{x_{i-\Delta}, \cdots, x_i, \cdots, x_{i+\Delta}\}) .
\]

Both the HMM algorithms can be extended by post-processing the output (endpoints) of the HMM using the clustering based classifier.
Proposed Model - Flow Chart

<Taxi id, Location, Time> → Stretch Factor → Hierarchical Segmentation → HMM-Naive → Post Process The Output

HMM-Stretch Factor
POST PROCESS THE OUTPUT

If the output of the hierarchical segmentation algorithm is $x_i$ (for an endpoint of the trip).

Clustering algorithm as to which of the three points $(x_{i-1}, x_i, x_{i+1})$ are the most likely endpoints.

Reduces off-by-one errors and improves the accuracy of our algorithm.
EVALUATION

Hierarchical Segmentation

- Beta

Hidden Markov Model (HMM)

- Delta
- Sampling Duration

Distance vs Time stretch factor based model

Time Computation
EVALUATION

Evaluating the section into two subsections,

- evaluate the performance of our algorithms based on stretch factor
- provide an end-to-end evaluation of the mobility pattern detection algorithms

Training: two weekdays and one weekend trace.

We build separate models for weekdays and weekends to capture inherent differences in human mobility patterns.
Evaluation

- performance of our algorithms based on stretch factor
- end-to-end evaluation of the mobility pattern detection algorithms
HIERARCHICAL SEGMENTATION

Stretch Factor of actual path taken by the taxi \((|x_i|, |x_{i+1}|, \ldots, |x_k|) < \beta\) (threshold)
Stretch Factor

(a) Shanghai dataset

(b) Stockholm dataset
HIDDEN MARKOV MODEL

Optimal Delta
Varied Sampling Duration

(a) Shanghai dataset
(b) Stockholm dataset
Varied Distance Stretch Factor
Varied Time Stretch Factor

(a) Shanghai dataset

(b) Stockholm dataset
Time Computation

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Hierarchical Segmentation</th>
<th>HMM naive</th>
<th>HSMM [24]</th>
<th>HMM Stretch Factor</th>
<th>HMM Clustering</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stockholm</td>
<td>2.1</td>
<td>23.4</td>
<td>35.2</td>
<td>41.1</td>
<td>41.3</td>
</tr>
<tr>
<td>Shanghai</td>
<td>3.2</td>
<td>29.4</td>
<td>51.4</td>
<td>63.4</td>
<td>63.7</td>
</tr>
</tbody>
</table>

This approach incurs about 20% additional computation time; but results in 30-40% improvement in accuracy.
Evaluation

- performance of our algorithms based on stretch factor
- end-to-end evaluation of the mobility pattern detection algorithms
Top 10 routes taken in Shanghai

2 main origins: Airports
Endpoints: Downtown
Top 10 routes taken in Stockholm

Main origins: Central Station
Endpoints: Downtown
Percentage of the top 10 trips

The graph shows the percentage of trips for the top 10 trips, with two cities compared: Stockholm and Shanghai. The percentage decreases significantly as the trip id increases.
DISADVANTAGES OF STRETCH FACTOR

When the passengers are picked consecutively the multiple trips may be identified as a single trip.

Certain cab drivers may always choose to take the shortest path irrespective of whether they are occupied or not.
APPLICATION

Applications

Future Work
Applications and Future Work

Mobility traces from about 10,000 private cars in Italy were analyzed to obtain mobility patterns of people.

Human mobility patterns collected from taxicabs were applied to detect flaws in urban planning (in Beijing).

Other types of vehicle data such as delivery trucks (e.g., FedEx, UPS) to identify offloading and onloading of goods and ambulance movement to derive correlation between patient locations and hospitals/clinics.
Q & A
Thank You