統計的機械学習による都市インテリジェンス研究

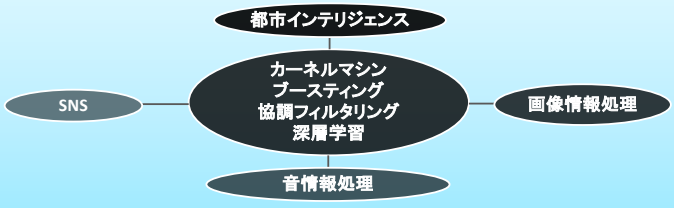
松井　知子　　データ科学研究系　教授

【概要】

本研究室では統計的学習機械を用いて、音声/音楽/画像/SNSなどを処理する方法について研究しています。具体的にはカーネルマシン、ブースティング、協調フィルタリング、深層学習の手法を用いて、

1. 音声・話者認識
2. 音楽情報処理
3. 画像識別
4. SNS解析
5. WEBユーザビリティ評価
6. 都市インテリジェンス　など

の研究課題に取り組んでいます。



本研究室では統計的機械学習とその応用研究に興味のある学生さんを募集しています！

【統計的機械学習】

- 統計科学を用いて、
  - データから、内在する数学的な構造を発見する。
  - その数学的な構造に基づいて、予測や判別などの情報処理を行う。
- 帰納的アプローチ

  v.s.
- 自然科学でよく見られる演繹的アプローチ
  - 仮説をたて、推論し、実験的または理論的に検証する。

- カーネルマシン
  - 自動的な特徴(/モデル)選択機構を含む。
  - 非線形の扱いに優れている。
  - サポートベクターマシン(SVM)、罰金付ロジスティック回帰マシン
- いろいろな確率モデルによる方法
  - 混合ガウス分布モデル
  - 隠れマルコフモデル
- ガウス過程状態空間モデル　など

【WiFiを利用した都市公共交通利用者数の推定】

Current approaches

- Manual counting
  - Expensive to get significant amount of data
  - Cannot track individual passengers
- "Tap on" using t:card – an electronic traveling card
  - Only used when entering bus
- Surveillance camera with face detection
  - Costly to purchase and install

New approach: WiFi detection

- Most passengers have a mobile device with WiFi capabilities turned on
- Even when not in use, a WiFi device will send out *frames* from time to time
- The frames will contain a unique identifier known as a MAC-address
- We want to count passengers based on unique MAC-addresses.

WiFi detection in depth

- Every 802.11 base station sends a beacon frame at regular intervals, typically every 100 milliseconds.
- The beacon frame is 50 bytes long and contains timestamps for synchronization, Service Set Identifier (SSID) etc.
- A probe request is a special frame sent by a client station requesting information from either a specific access point, specified by SSID, or all access points in the area, specified with the broadcast SSID
- Since probe requests are sent even when not connected to a network, these are the frames we will be looking for
- Typically, only one percent of the detected frames are probe requests

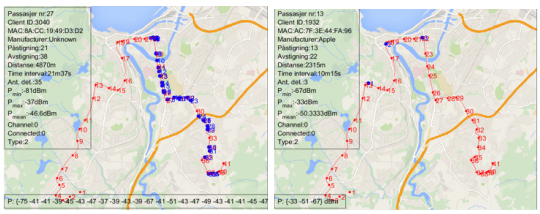
Typical output from wireless sniffer after filtering. Each line contains a timestamp, MAC-address, Received signal strength indication (RSSI), SSID, Anonymous MAC flag, Hidden SSID Flag, Channel number.

```
2016-05-26T21:07:18.145462 [redacted] -88 Nintendo_3DS_continuous_scan_000  True  False 6
2016-05-26T21:07:18.212162 [redacted] -69 <<<HIDDEN>>>  False True  1
2016-05-26T21:07:18.286180 [redacted] -75 <<<HIDDEN>>>  False True  2
2016-05-26T21:07:18.343039 [redacted] -57 <<<HIDDEN>>>  False True  6
2016-05-26T21:07:18.335377 [redacted] -67 <<<HIDDEN>>>  False True  1
2016-05-26T21:07:18.401981 [redacted] -61 <<<HIDDEN>>>  False True  6
2016-05-26T21:07:18.402138 [redacted] -75 ptdepot False False 2
2016-05-26T21:07:18.535995 [redacted] -68 <<<HIDDEN>>>  False True  1
2016-05-26T21:07:18.615717 [redacted] -90 Nintendo_3DS_continuous_scan_000  True  False 6
2016-05-26T21:07:18.659058 [redacted] -65 <<<HIDDEN>>>  False True  1
```

Main problems

Data was collected for several bus routes. When logging the WiFi frames, several persons counted the number of passengers entering and leaving the bus. Also, some of the data sets also contain position data (GPS).
Given the collected data, we have the following problems to be solved:

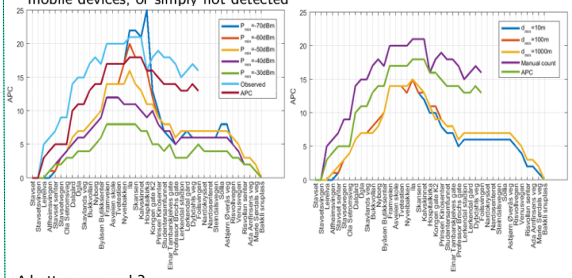
- Detect mobile devices
- Determine which mobile devices are on the bus
- Estimate the number of travelers and their route
- Estimate passenger flow through network



共同研究者:
Tor Andre Myrvoll (NTNU)
Francois Septier (Lille U.)

Basic approach: Hard decision using thresholds

- The mobile devices are classified as ON or OFF the bus according to a set of thresholds on the distance travelled and the RSSI
- In addition a scaling factor is estimated to compensate for people not carrying mobile devices, or simply not detected



A better approach?

- Hard decisions using thresholds is not optimal, especially when several variables are in use
- No model. Variable interaction hard to utilize.
- A probabilistic approach should be attempted
- Initially – focus on the mobile device ON/OFF problem

Model

We formulate an estimator of $o(n)$ as follows,

$$\hat{o}(n) = \sum_{k=0}^{K-1} I(d_k, m_k) \sum_{l=0}^{n} \left(\delta(a_k - l) - \delta(b_k - l)\right).$$

We choose the minimum square error as our metric, resulting in the following loss function

$$L = \sum_{n=1}^{N} (o(n) - \hat{o}(n))^2$$

We want to base $I(d_k, m_k)$ on the posterior probability $P(i_k|d_k, m_k)$, that is, $I(d_k, m_k) = 1$ if $P(i_k|d_k, m_k) > T$ and zero otherwise.

Obtain the model parameters in a maximum likelihood sense

$$\hat{\Lambda} = \arg\max_{\Lambda} P(O, M, D; \Lambda)$$
$$= \arg\max_{\Lambda} \sum_{I} P(I, O, M, D; \Lambda)$$

Since this is a missing data problem we use the EM-algorithm,

$$\Lambda^{(k+1)} = \arg\max_{\Lambda} Q(\Lambda, \Lambda^{(k)})$$

where

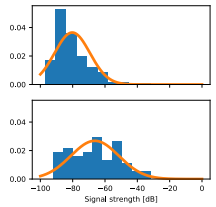
$$Q(\Lambda, \Lambda') = \mathbb{E}_{I|O,D,M;\Lambda'}\{\log P(O, I, D, M; \Lambda)\}$$
$$= \sum_{I} P(I|O, D, M; \Lambda') \log P(C, I, D, M; \Lambda)$$

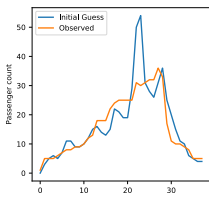
Conclusions

- The methods seems to work well, meaning that incomplete observations is all that is needed to create a working model
- More work is needed on some modeling assumptions
  - Sensitivity with respect to e.g. bus sizes (RSSI variations)
  - Does the model change significantly with time of day and/or seasons?
  - Finer granularity – Use GPS and/or time data instead of bus stops

Experimental results

- We testet our algorithm in the following scenario where we used a complete bus trip to estimate model parameters, and used this to predict the number of passengers per stop on an unrelated bus trip
- The model was initialized by splitting the observed devices into two sets – those seen for more than two stops and those which were not





- We did ten iterations of the EM algorithm
- For every iteration of the EM algorithm, we produced 3000 Gibbs samples, where the first 1000 were considered "burn in" and rejected.
- Using the final model we computed $\hat{o}(n)$ using

$$I(d_k, m_k) = \begin{cases} 1 & P(i_k|d_k, m_k) > T \\ 0 & \text{otherwise} \end{cases}$$

- The prediction is seen to the right for $T \in \{0.7, 0.8, 0.9\}$