A Nonparametric Method for Early Detection of Trending Topics

Zhang Zhang

Advisor: Prof. Aravind Srinivasan

Winter Presentation

- Latent Resource Data Model
- Theoretical Guarantee
- MapReduce Method
- Signal Construction Strategy
- Optimal Parameters Strategy
- Project Schedule

Weighted Majority Voting

$$\hat{L}^{(T)}(s,\gamma) = \begin{cases} + & \text{if} \quad \sum_{r\in\mathcal{R}_+} e^{-\gamma d^{(T)}(r,s)} \geq \sum_{r\in\mathcal{R}_-} e^{-\gamma d^{(T)}(r,s)} \\ - & \text{otherwise} \end{cases}$$

$$d^{(T)}(r,s) = \sum_{t=1}^{t} (r(t) - s(t))^2 = \|r - s\|_T^2$$

$$d^{(T)}(r,s) = \min_{\Delta\in\{-\Delta_{max},\ldots,0,\ldots,\Delta_{max}\}} \sum_{t=1}^{T} (r(t+\Delta) - s(t))^2 = \min_{\Delta\in\{-\Delta_{max},\ldots,0,\ldots,\Delta_{max}\}} \|r * \Delta - s\|_T^2$$

Latent Source Model

1. Sample a latent source V from $\mathcal{V}$ uniformly at random. Let $L \in \{+1, -1\}$ be the label of V.
2. Sample integer time shift Δ uniformly from $\{0, 1, ..., \Delta_{max}\}$.
3. Output time series S to be the latent source V advanced by Δ time steps, followed by adding noise signal E, i.e., $S(t) = V(t + \Delta) + E(t)$ for $t \geq 1$.

$$\Lambda^{(T)}(s;\gamma) \triangleq \frac{\sum_{r_+ \in \mathcal{R}_+} exp(-\gamma(\min_{\Delta_+ \in \mathcal{D}} \|r_+ * \Delta_+ - s\|_T^2))}{\sum_{r_- \in \mathcal{R}_-} exp(-\gamma(\min_{\Delta_- \in \mathcal{D}} \|r_- * \Delta_- - s\|_T^2))}$$

Theoretical Guarantee

$$S = V * \Delta' + E'$$

$$R = V * \Delta'' + E''$$

$$\mathbb{P}(\text{missclassification of S using its first T samples})$$
$$= \mathbb{P}(\hat{L}_\theta = -1 | L = +1)\mathbb{P}(L = +1) + \mathbb{P}(\hat{L}_\theta = +1 | L = -1)\mathbb{P}(L = -1)$$

$$\mathbb{P}(\hat{L}_\theta = -1 | L = +1) = \mathbb{P}(\frac{1}{\Lambda^{(T)}} \geq \frac{1}{\theta} | L = +1) \leqslant \theta \mathbb{E}[\frac{1}{\Lambda^{(T)}} | L = +1]$$

$$\mathbb{E}[\frac{1}{\Lambda^{(T)}} | L = +1] \leqslant \max_{r_+ \in \mathcal{R}_+, \Delta_+ \in \mathcal{D}} \mathbb{E}[\frac{1}{\Lambda^{(T)}(r_+ * \Delta_+ + E; \gamma)}]$$

Theoretical Guarantee

$$S = V * \Delta' + E'$$

$$R = V * \Delta'' + E''$$

$$\mathbb{P}(\text{missclassification of S using its first T samples})$$
$$= \mathbb{P}(\hat{L}_\theta = -1 | L = +1)\mathbb{P}(L = +1) + \mathbb{P}(\hat{L}_\theta = +1 | L = -1)\mathbb{P}(L = -1)$$

$$\mathbb{P}(\hat{L}_\theta = -1 | L = +1) = \mathbb{P}(\frac{1}{\Lambda^{(T)}} \geq \frac{1}{\theta} | L = +1) \leqslant \theta \mathbb{E}[\frac{1}{\Lambda^{(T)}} | L = +1]$$

$$\mathbb{E}[\frac{1}{\Lambda^{(T)}} | L = +1] \leqslant \max_{r_+ \in \mathcal{R}_+, \Delta_+ \in \mathcal{D}} \mathbb{E}[\frac{1}{\Lambda^{(T)}(r_+ * \Delta_+ + E; \gamma)}]$$

Theoretical Guarantee

$$\Lambda^{(T)}(s;\gamma) \geq \frac{exp(-\gamma||r_+ * \Delta_+ - s||_T^2)}{\max_{r_- \in \mathcal{R}_-, \Delta_- \in \mathcal{D}} exp(-\gamma||r_- * \Delta_- - s||_T^2}$$

$$\frac{1}{\Lambda^{(T)}(r_+ * \Delta_+ + E;\gamma)}$$

$$\leq \max_{r_- \in \mathcal{R}, \Delta_- \in \mathcal{D}} \{exp(-\gamma||r_+ * \Delta_+ - r_- * \Delta_-||_T^2) exp(-2\gamma\langle r_+ * \Delta_+ - r_- * \Delta_-, E\rangle_T)\}$$

$$\leq \Delta_{max} n_- exp(-(\gamma - 4\sigma^2\gamma^2)G^{(T)})$$

MapReduce Method

The *Map* and *Reduce* functions of *MapReduce* are both defined with respect to data structured in (key, value) pairs. *Map*takes one pair of data with a type in one data domain, and returns a list of pairs in a different domain:

Map(k1,v1) → list(k2,v2)

The *Map* function is applied in parallel to every pair in the input dataset. This produces a list of pairs for each call. After that, the MapReduce framework collects all pairs with the same key from all lists and groups them together, creating one group for each key

The *Reduce* function is then applied in parallel to each group, which in turn produces a collection of values in the same domain

Reduce(k2,list(v2))→list(v3)

Each *Reduce* call typically produces either one value v3 or an empty return, though one call is allowed to return more than one value. The returns of all calls are collected as the desired result list.

MapReduce Method

Map.py

Reduce.py

One tweet data in JSON

Post_process.py

Get_array_time.py

Let’s run and see the results!!

Signal Construction Strategy

Topics: Hashtag

Trending Topics:

Filter out topics whose rank was never better than 3

Filter out topics did not trend for a long time and topics trend multiple times

Signal Construction Strategy

Trending Topics Reference Signal:

A slice terminates at the first trending onset time;

Non Trending Reference Signal:

Select Slice with random start and end times. Fixed size.

Detection Strategy

Aligned Problem:

$$d(\mathbf{r}, \mathbf{s}) = \min_{k=1,\ldots,N_{ref}-Nobs+1} d(\mathbf{r}_{k:k+N_{obs}-1}, \mathbf{s})$$

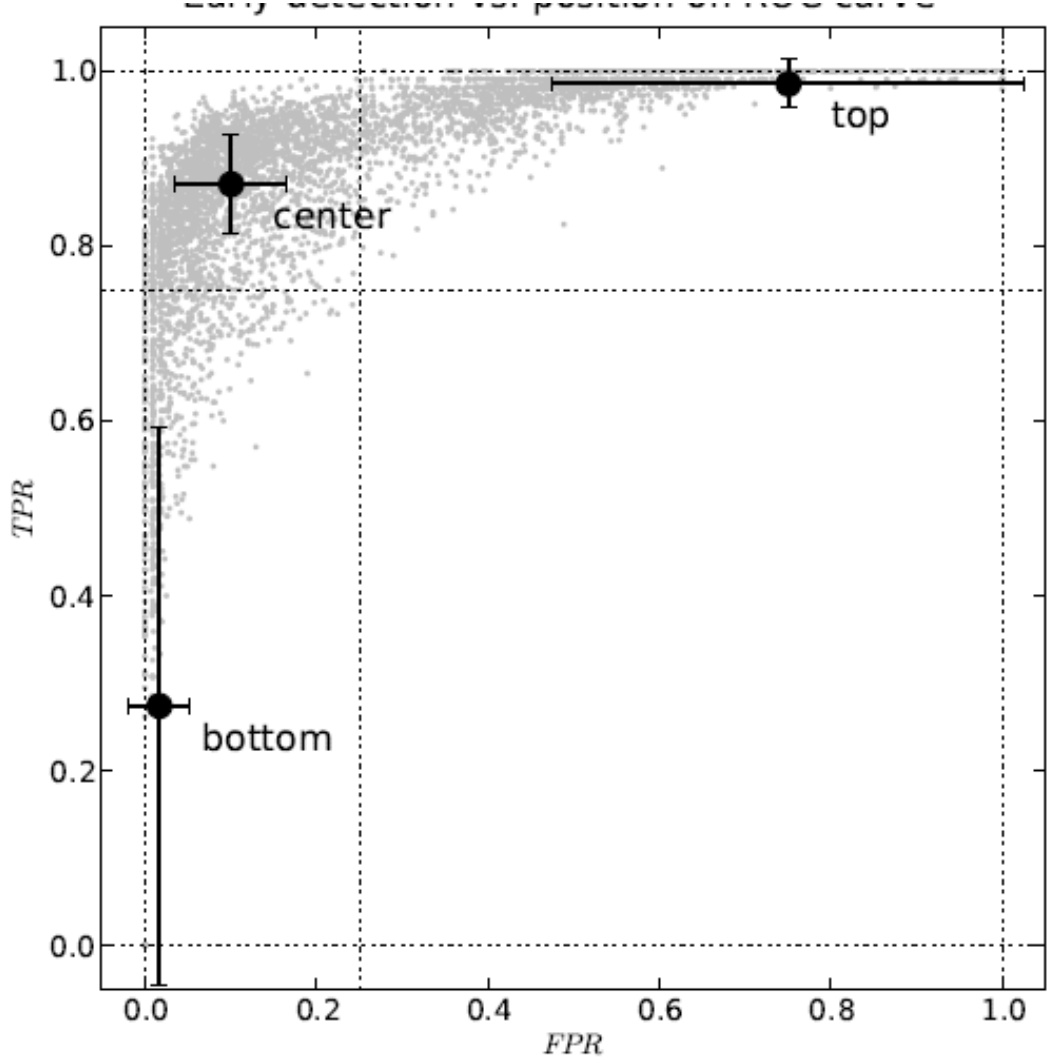
Trended Test Signal:

Detection window is spanned by 2h_ref hours centered at onset

Optimal Parameters

- γ: 0.1, 1, 10
- N_{obs}: 10, 80, 115, 150
- N_{smooth}: 10, 80, 115, 150
- h_{ref}: 3, 5, 7, 9
- D_{ref}: 1, 3, 5
- θ: 0.65, 1, 3

Optimal Parameters



Thank you!

Questions?