6th IEEE Global Conference on Signal and Information Processing

This research is based upon work supported by the National Science Foundation under Grant No. ECCS-1737443

What matters the most? Optimal Quick Classification of Urban Issue Reports by Importance

Yasitha Warahena Liyanage[*], Mengfan Yao[+], Christopher Yong[+], Daphney–Stavroula Zois[*], Charalampos Chelmis[+]

[*]*Electrical and Computer Engineering Department*

[+]*Computer Science Department*

University at Albany, SUNY

29 November 2018

Motivation

- Civic engagement platforms
  - enable citizens to participate in collecting, analyzing and sharing knowledge about their local environments (e.g., measure air quality [Dutta2009])
  - interact with local governments to resolve urban issues, such as potholes and noise complaints (e.g. SeeClickFix [Mergel2012])

SeeClickFix
power to the community

- Reported issues should be **timely processed** and **addressed** to maintain citizens' satisfaction with local governments

FixMyStreet

Related Work

- Prior work
  - **ignores citizens' implicit endorsement** of urban issues that are "important" to them (e.g., [Budde2014])
  - requires **large–scale annotation** to achieve good accuracy (e.g., [Hirokawa2017])
  - relies on **fixed set of features** (e.g., [Budde2014], [Hirokawa2017])
  - **ignores scalability** and **timeliness** (e.g., [Budde2014], [Hirokawa2017])

- Currently, reported issues are acknowledged and assessed by a city official for routing to appropriate agency

  - We propose to **classify importance** of urban issues **as fast as possible without sacrificing accuracy** using optimal subset of features in an **online fashion**

Problem Formulation

- Each urban issue i consists of
  - Title
  - Description
  - Address
  - Timestamp
  - Photo(s)
  - Comment(s)
  - Vote(s)

$\xrightarrow{\text{extract}}$ feature vector $\mathbf{f}_i = [f_1, f_2, \ldots, f_K]^T$

$\xrightarrow[\text{hypotheses}]{\text{two possible}}$ $\begin{cases} H_{\mathcal{H}} : \text{high importance} \\ H_{\mathcal{L}} : \text{low importance} \end{cases}$

- **Urban issue importance**: # of votes and comments received
- Feature cost $c_n > 0, \ n \in \{1, \ldots, K\}$
- Misclassification costs $M_{kj} \geqslant 0, k \in \{\mathcal{H}, \mathcal{L}\}, j \in \{1, \ldots, L\}$ with L decision choices

Optimization Problem

- **Goal:** minimize **number of features used** for inferring **importance** of an issue **without sacrificing accuracy**

the feature f_R that the framework stops at

$$\min_{R, D_R} J(R, D_R)$$

the possibility to select among L decision choices

$$J(R, D_R) = \mathbb{E}\Bigg\{ \underbrace{\sum_{n=1}^{R} c_n}_{\text{Cost of evaluating features}} + \underbrace{\sum_{j=1}^{L} \sum_{k=\mathcal{H},\mathcal{L}} M_{kj} P(D_R = j, H_k)}_{\text{Misclassification cost}} \Bigg\}$$

Optimal Classification Strategy

- Rewrite the objective function using π_n

$$J(R, D_R) = \mathbb{E}\left\{ \sum_{n=1}^{R} c_n + \sum_{j=1}^{L} \big(M_{\mathcal{H}j}\pi_R + M_{\mathcal{L}j}(1-\pi_R)\big)\mathbf{1}_{\{D_R=j\}} \right\}$$

a posteriori probability

$$\pi_n \triangleq P(H_{\mathcal{H}}|f_1, \ldots, f_n)$$

- **Optimal classification strategy**

$$D_R^{optimal} = \arg\min_{1 \leqslant j \leqslant L}\big[M_{\mathcal{H}j}\pi_R + M_{\mathcal{L}j}(1-\pi_R)\big]$$

  - Results to the smallest average cost

$$\widetilde{J}(R) \triangleq J(R, D_R^{optimal}) = \mathbb{E}\left\{ \sum_{n=1}^{R} c_n + g(\pi_R) \right\}$$

where $g(\pi_R) \triangleq \min_{1 \leqslant j \leqslant L}\big[M_{\mathcal{H}j}\pi_R + M_{\mathcal{L}j}(1-\pi_R)\big]$

Optimal Stopping Strategy

- **Optimal stopping strategy** via dynamic programming
  - Last stage

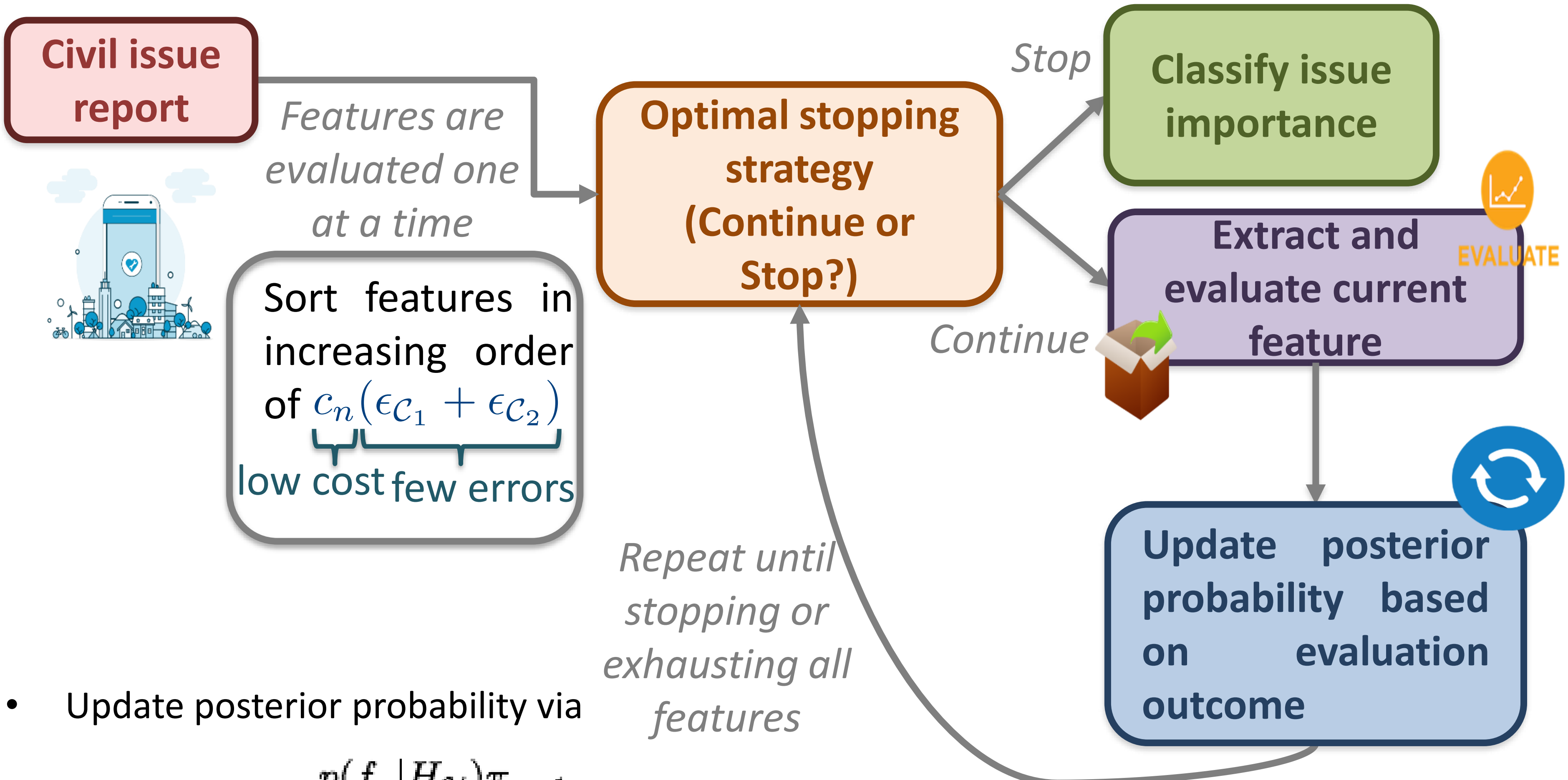
$$\bar{J}_K(\pi_K) = g(\pi_K)$$

  - Any intermediate stage

$$\underbrace{\bar{J}_n(\pi_n)}_{\text{Optimal cost–to–go}} = \min \left[\overbrace{g(\pi_n)}^{\text{Cost of stopping}}, \underbrace{c_{n+1} + \sum_{f_{n+1}} A_n(f_{n+1}) \bar{J}_{n+1}\left(\frac{p(f_{n+1}|H_{\mathcal{H}})\pi_n}{A_n(f_{n+1})} \right)}_{\text{Cost of continuing}} \right]$$

where $A_n(f_{n+1}) \triangleq \pi_n p(f_{n+1}|H_{\mathcal{H}}) + (1-\pi_n) p(f_{n+1}|H_{\mathcal{L}})$

CIvIC: Classify urban Issues into Importance Categories



- Update posterior probability via

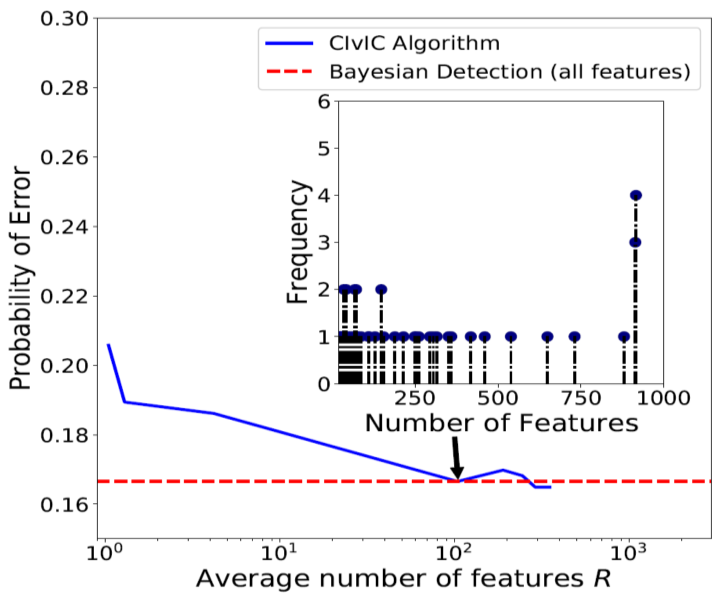
$$\pi_n = \frac{p(f_n|H_{\mathcal{H}})\pi_{n-1}}{\pi_{n-1}p(f_n|H_{\mathcal{H}}) + (1-\pi_{n-1})p(f_n|H_{\mathcal{L}})}, \quad \pi_0 = p(H_{\mathcal{H}})$$

Case Study: The SeeClickFix Platform

- Dataset
  - 2, 195 SeeClickFix issues
  - Metropolitan area surrounding Albany, NY
  - Jan 5, 2010 and Feb 10, 2018
- Features extracted from issues' title, description, address, and reported time
  - E.g., tokenized unigrams, logarithm of the number of words +1, exclamation marks +1, uppercase letters +1
- **Discretized importance** based on predefined thresholds
  - $H_{\mathcal{H}}$ if number of votes $V > \bar{V}$ and number of comments $C > \bar{C}$
  - Otherwise it belongs to $H_{\mathcal{L}}$
- To verify **robustness**, we considered 4 scenarios of varying thresholds $\bar{V}$ and $\bar{C}$

Results

- Baselines
  - Bayesian detection method that uses all features
  - Feature selection method: SVM–FS [Hirokawa2017]
  - Dimensionality Reduction method: SVM–PCA
  - Kernel based method: SVM classifier
  - Tree based classifiers: Random forest and XG-boosting



- ClvIC achieves same error probability as Bayesian detection with all features using only **104 out of 2594 features** on average
- On average **96% reduction** in the number of **features** used

Results

Method	Accuracy	Precision	Recall	Avg. # feat.
CIvIC ($c = 0.25$)	0.794	0.785	0.818	1.05
CIvIC ($c = 10^{-1}$)	0.811	0.789	0.854	1.29
CIvIC ($c = 10^{-2}$)	0.814	0.783	0.873	4.19
CIvIC ($c = 10^{-3}$)	0.833	0.801	0.889	104.10
CIvIC ($c = 10^{-4}$)	0.830	0.807	0.870	189.78
CIvIC ($c = 10^{-5}$)	0.832	0.811	0.867	244.99
CIvIC ($c = 10^{-6}$)	**0.835**	0.819	0.864	289.59
CIvIC ($c = 0$)	0.835	0.819	0.864	350.34
Bayesian Detection	0.833	0.819	0.860	2,594
SVM-FS	0.746	0.701	0.810	20
SVM-linear	0.806	0.801	0.815	2,594
SVM-Gaussian	0.796	0.739	0.916	2,594
SVM-PCA	0.825	0.791	0.886	208
RF (depth=5)	0.815	0.779	0.883	2,594
RF (depth=10)	0.820	0.784	0.886	2,594
XG Boosting	0.827	0.801	0.873	2,594

- CIvIC uses on average 104 and 289 features and achieves **same highest accuracy** (83.3%) and **precision** (81.9%) as Bayesian detection with all features (i.e., **96%** and **88.8% reduction**)

- SVM–Gauss achieves highest recall (91.6%), but **25 times as many features** for a mere 3% improvement compared to CIvIC

Contributions & Future Directions

- Contributions
  - **Optimal stopping theory framework** to **dynamically** infer importance of incoming urban requests
  - **Near–real–time algorithm** that implements optimal solution
- Future directions
  - Extend framework to enable multi–valued importance recognition
  - Devise appropriate learning–to–rank approaches to dynamically order incoming urban issues requests
- Questions?

email: yliyanage@albany.edu

References

[Dutta2009] P. Dutta, P. M. Aoki, N. Kumar, A. Mainwaring, C. Myers, W. Willett, and A. Woodruff, "*Common sense: participatory urban sensing using a network of handheld air quality monitors*," in 7th ACM conference on embedded networked sensor systems. ACM, 2009, pp. 349–350.

[Mergel2012] I. Mergel, "*Distributed democracy: Seeclickfix.com for crowdsourced issue reporting*," 2012.

[Budde2014] M. Budde, J. D. M. Borges, S. Tomov, T. Riedel, and M. Beigl, "*Improving Participatory Urban Infrastructure Monitoring through Spatio-Temporal Analytics*," in 3rd ACM SIGKDD International Workshop on Urban Computing. ACM, 2014.

[Hirokawa2017] S. Hirokawa, T. Suzuki, and T. Mine, "*Machine Learning is Better Than Human to Satisfy Decision by Majority*," in International Conference on Web Intelligence. 2017, pp. 694–701, ACM.