Likelihood Ratios for Categorical Evidence

With Applications in Digital Forensics

Rachel Longjohn (rlongjoh@uci.edu) and Padhraic Smyth (smyth@ics.uci.edu)

University of California, Irvine

Background

Digital forensics process

Extract event data as two categorical counts

Source or identity-focused questions

Same vs. different-source hypothesis framework

Same-source hypothesis H_s
Unknown source data **generated** by the known source

Different-source hypothesis H_d
Unknown source data **not generated** by the known source

Strength of evidence via likelihood ratio

$$\frac{Pr(H_s)}{Pr(H_d)} \times \frac{Pr(E|H_s)}{Pr(E|H_d)} = \frac{Pr(H_s|E)}{Pr(H_d|E)}$$

Methods

Evidence

Vector of known source counts

$$r_1 = (r_{11}, r_{12}, ..., r_{1K})$$

Vector of unknown source counts

$$r_2 = (r_{21}, r_{22}, ..., r_{2K})$$

Requires specifying K event types of interest

Model

Model for known source data

$$r_1|\theta_1 \sim \text{Multinom}(\theta_1)$$

Model for unknown source data

$$r_2|H_s, \theta_1 \sim \text{Multinom}(\theta_1)$$

$$r_2|H_d, \theta_2 \sim \text{Multinom}(\theta_2)$$

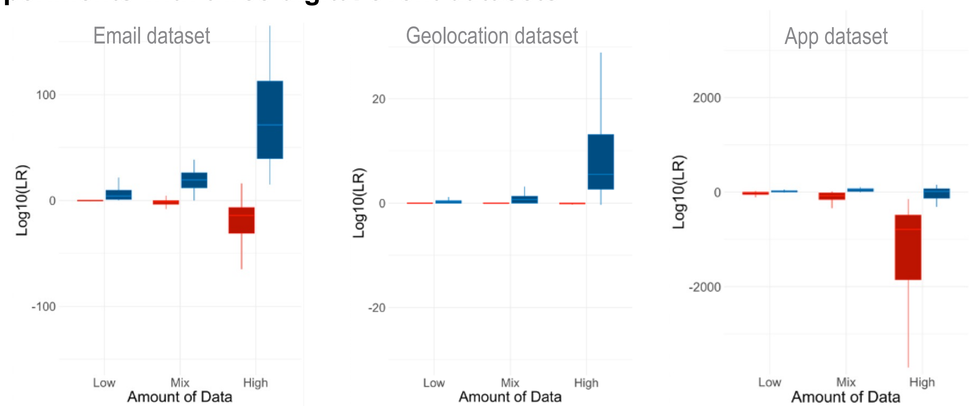
Prior distribution

$$\theta_1, \theta_2 \overset{iid}{\sim} \text{Dir}(\alpha)$$

Closed-form likelihood ratio

$$LR = \frac{B(\alpha + r_1 + r_2)B(\alpha)}{B(\alpha + r_2)B(\alpha + r_1)} = \left(\prod_{k=1, r_{2k} \geq 1}^{K} \prod_{s=0}^{r_{2k}-1} \left(1 + \frac{r_{1k}}{\alpha_k + s}\right)\right)\left(\prod_{s=0}^{N_2-1}\left(1 - \frac{N_1}{c + N_1 + s}\right)\right)$$

Experiments with three digital event datasets



Results

	TPR@1	FPR@1	AUC
Email, $K = 945$			
Noninformative	94.8%	10.1%	98.1%
Informative	94.7%	8.1%	98.5%
Geolocation, $K = 1412$			
Noninformative	72.8%	6.4%	91.5%
Informative	72.7%	5.1%	92.4%
App, $K = 159$			
Noninformative	75.0%	11.5%	86.3%
Informative	63.8%	6.2%	82.5%

TPR@1 and FPR@1 use 1 as a threshold for each LR. Noninformative vs. informative refers to the Dirichlet prior (symmetric vs. asymmetric).

Discussion

- Model demonstrates good discriminative ability (high AUCs, high TPRs, low FPRs)
- Informative priors incorporate the commonality of certain event types (dampen impact of common events on the LR)
- However, it is sensitive to imbalances between the strength of the prior distribution and the number of observed events (geolocation + app boxplots)
- Future work could improve this model by incorporating sequence information, inter-event times, time-varying event patterns, and potential relationships between event types

This work was funded by the Center for Statistics and Applications in Forensic Evidence (CSAFE) through Cooperative Agreements 70NANB15H176 and 70NANB20H019 between NIST and Iowa State University, which includes activities carried out at Carnegie Mellon University, Duke University, University of California Irvine, University of Virginia, West Virginia University, University of Pennsylvania, Swarthmore College, and University of Nebraska, Lincoln.