Likelihood Ratios for Categorical Evidence

With Applications in Digital Evidence

Rachel Longjohn, Padhraic Smyth, Hal Stern

csafe Center for Statistics and Applications in Forensic Evidence

UCI University of California, Irvine

1. Introduction

- We introduce a *likelihood ratio-based approach* for evaluating the strength of evidence when it is in the form of *categorical count data*.
- In digital evidence, *user-generated event data* can often be expressed as categorical count data.
- Digital evidence research to date, however, has primarily focused on information extraction and reconstruction rather than statistical methods [1,2].
- We demonstrate how our approach can be applied to user-generated event data to *compare user behavior patterns in digital evidence*.

2. Methods: statistical model

Evidence

- Two sets of counts across k categories of forensic interest (e.g., # times each of k apps was opened on the device)

$$\underbrace{r_1 = (r_{11}, r_{12}, ..., r_{1k})}_{\text{known source data}} \quad \underbrace{r_2 = (r_{21}, r_{22}, ..., r_{2k})}_{\text{unknown source data}}$$

Hypotheses

- *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

Model

- Known source counts: $r_1|\theta_1 \sim \text{Multinomial}(N_1, \theta_1)$
- Unknown source counts depend on which hypothesis is true:

$$r_2|H_s, \theta_1 \sim \text{Multinomial}(N_2, \theta_1)$$

$$r_2|H_d, \theta_2 \sim \text{Multinomial}(N_2, \theta_2)$$

- Opt for a Bayesian approach and assign priors to unknown parameters [3]: $\theta_1, \theta_2 \overset{i.i.d.}{\sim} \text{Dirichlet}(\alpha)$

3. Methods: likelihood ratio

- Express the strength of evidence via the likelihood ratio [4]:

$$\underbrace{\frac{Pr(H_s|E)}{Pr(H_d|E)}}_{\text{prior odds}} \cdot \underbrace{\frac{Pr(E|H_s)}{Pr(E|H_d)}}_{\text{likelihood ratio}} = \underbrace{\frac{Pr(H_s|E)}{Pr(H_d|E)}}_{\text{posterior odds}}$$

- Likelihood ratio (LR) can be interpreted as [4]:
  - LR < 1: evidence more probable under H_d
  - LR = 1: evidence does not help distinguish H_d from H_s
  - LR > 1: evidence more probable under H_s
- Multinomial Dirichlet assumptions lead to a closed-form solution for the likelihood ratio in terms of multivariate Beta functions:

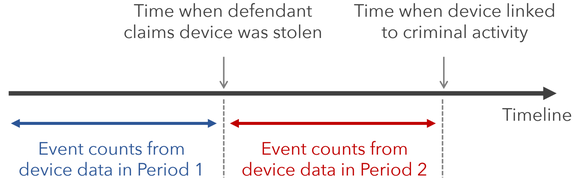
$$LR = \frac{B(\alpha + r_1 + r_2)B(\alpha)}{B(\alpha + r_2)B(\alpha + r_1)}$$

4. Application to digital evidence

User-generated event data

- ID identifies the user, device, or account
- Event is a user-generated action that can be categorized
- Timestamp is the time at which the event occurred

Stolen device scenario

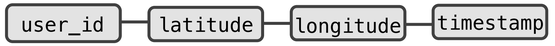


- Our setup is also flexible to other situations in digital forensics [5], e.g., if it's believed a suspect carries two phones, we can compare behavior patterns across the two devices.

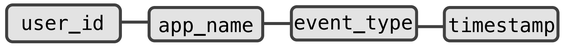
5. Data and experiments

Email communications

Device location pings (geolocated Tweets)

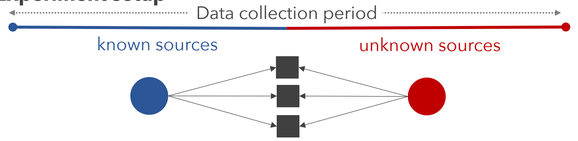


Mobile app records



(Use QR code to learn more about the datasets on the project website.)

Experiment setup



- Evaluate LR using all same-source pairs and a stratified random sample of different-source pairs, where the strata are defined by the amount of data.

6. Results

	TPR@1	FPR@1	AUC
Email, $k = 945$			
Symmetric	0.948	0.090	0.983
Asymmetric	0.947	0.068	0.985
Twitter, $k = 1415$			
Symmetric	0.726	0.067	0.912
Asymmetric	0.724	0.051	0.923
App, $k = 159$			
Symmetric	0.750	0.184	0.823
Asymmetric	0.638	0.098	0.767

TPR@1 and FPR@1 are the true and false positive rates, respectively, using 1 as a threshold for each LR. AUC is the area under the ROC curve. Symmetric vs. asymmetric refers to the choice of Dirichlet prior.

7. Discussion and conclusions

- We observed generally *high TPRs, low FPRs, and high AUC values* across all three datasets.
- Using an *asymmetric Dirichlet prior (prior parameters set via holdout data) resulted in lower FPRs*. This indicates that considering categories' prevalence decreases false positives.
- For the email and Twitter datasets, AUC and TPR were not particularly sensitive to the prior choice.
- Performance on the mobile data was generally worse than the other two datasets. This could indicate that mobile app behavior is not particularly discriminatory. Further investigation is needed.
- Future directions include incorporating the relationships between categories (e.g., spatial or hierarchical) and accounting for the sequence of events.

8. References

[1] SWDGE. 2020a. "Best Practices for Mobile Device Evidence Collection and Preservation, Handling, and Acquisition."
[2] SWGDE. 2020b. "Best Practices for Mobile Device Forensic Analysis."
[3] Puig, Xavier, Martí Font, and Josep Ginebra. "A unified approach to authorship attribution and verification." The American Statistician 70.3 (2016): 232-242.
[4] Aitken, Colin, Franco Taroni, and Silvia Bozza. Statistics and the Evaluation of Evidence for Forensic Scientists. John Wiley & Sons, Inc., 2021.
[5] Casey, Eoghan, et al. 2020. "Structuring the Evaluation of Location-Related Mobile Device Evidence." Forensic Science International: Digital Investigation 32: 300928.

9. Acknowledgements

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.