**PROBLEM STATEMENT**

Consider a pair of user-generated event series

\[ M = (A, B) = \{ (t_i, m(t_i)) : j = 1, \ldots, n \} \]

where \( t_j \in \mathbb{R}^+ \) is the time and \( m(t_j) \in \{A, B\} \) is the type of the \( j^{th} \) event. We want to quantify the likelihood that the pair was generated by the same source.

**MEASURES OF ASSOCIATION**

**Score Functions using Nearest Neighbors**

- **Coefficient of Segregation** [3]: function of the ratio of the probability that a reference point (i.e., a randomly selected event in \( A, B \)) and its nearest neighbor have different marks to the same probability for independent marks.

\[ S(A, B) = 1 - \frac{PA \cdot PB \cdot \min(1, N_{k}))}{PA \cdot PB \cdot \min(1, N_{k}))} \in [0, 1] \]

- **Minning Index** [4]: mean fraction of points among the \( k \) nearest neighbors whose type is different than that of the reference point

\[ M_k(A, B) = \frac{1}{nk} \sum_{j=1}^{n} \sum_{i=1}^{k} \mathbb{1}(m(t_i) \neq m(t_j)) \in [0, 1] \]

**Score Functions using Inter-Event Times**

Assume that \( n_A < n_B \) and fix series \( B. \) We then measure the time from each event in \( B \) to the closest event in series \( A \) in either direction

\[ T_{BA} = \{ t_{BA,j} : j = 1, \ldots, n_B \} \]

where \( t_{BA,j} = \min_{k \in \{1, \ldots, n_A\}} |t_A - t_{Bk}| \)

- Mean inter-event time from \( B \) to \( A \)

\[ T_{BA} = \frac{1}{n_A} \sum_{j=1}^{n_B} t_{BA,j} \in (0, \infty) \]

- Median inter-event time from \( B \) to \( A \)

\[ \text{med}(T_{BA}) \in (0, \infty) \]

**RESAMPLING APPROACH**

**Given**

- Pair of interest: \( (A^*, B^*) \)

**Score function**: \( \Delta \)

**Method**

- Two competing hypotheses:
  
  \( H_0: (A^*, B^*) \) came from the same source
  
  \( H_1: (A^*, B^*) \) came from different sources

- Use sample \( M_i = (A_i, B_i) \) for \( i = 1, \ldots, N \) to estimate the score-based likelihood ratio

\[ SLR_\Delta = \frac{g(\Delta(A^*, B^*), H_1)}{g(\Delta(A^*, B^*), H_0)} \]

**DIFFERENT INTERPRETATIONS OF THE DENOMINATOR [1]**

**COMPARISON OF APPROACHES**

**Given**

- Pair of interest: \( (A^*, B^*) \)

**Score function**: \( \Delta \)

**Method**

- Focus on the denominator of \( SLR_\Delta \)

**Coincident match probability**: probability that a different-source pair with observed score \( \Delta(A^*, B^*) \) exhibits association by chance

\[ CMP_\Delta = Pr(\Delta(A, B) < \Delta(A^*, B^*), H_0) \]

- Use resampling in time to simulate different-source pairs \( (A^0, B^0) \) and estimate

\[ CMP_\Delta = \frac{1}{n_{sim}} \sum_{i=1}^{n_{sim}} \mathbb{1}(\Delta(A^0, B^0) < \Delta(A^*, B^*)) \]

**REFERENCES**


**QUANTIFYING THE ASSOCIATION BETWEEN DISCRETE EVENT TIME SERIES**

CHRIS GALBRAITH\(^\ddagger\), PADHRAI SMYTH\(^\ddagger\) & HAL S. STERN\(^\ddagger\)

\(^\ddagger\)Department of Statistics, \(^\ddagger\)Department of Computer Science

**CONCLUSIONS**

- Resampling approach shows promise in situations where no reference data is available
- Population-based SLR is preferred, given
  - Better performance for weakly associated pairs
  - Similar performance for strongly associated pairs
- Well-established in forensic investigation

**CASE STUDY**

- Data from a 2013-2014 study at UCI that placed logging software on 124 students’ computers that recorded all browser activity for one week [2]
- Event series created by dichotomizing browsing events to Facebook versus non-Facebook urls
- Only considered 55 students with at least 50 web browsing events of each type

**COMPUTER SERIES**

**FORUM**

**CONTACT INFORMATION**

Web: www.ics.uci.edu/~galbraic
Email: galbraic@uci.edu