Scalable Bayesian Record Linkage

Brendan S. McVeigh
Jared S. Murray

Introduction

Probabilistic record linkage (PRL) is the task of merging two or more databases that have entities in common but no unique identifier. Matching must be done based on incomplete information, since features for records may be incorrectly or inconsistently recorded.

Sets of records, A and B with n_A and n_B entries. We use a matrix C to represent which records are linked (correspond to the same entity).

\[ C_{ab} = \begin{cases} 1 & a \sim b \\ 0 & a \not\sim b \end{cases} \]

Comparisons between record pairs used to estimate links

\[ \gamma_{ab} = (\gamma_{a1b}, \gamma_{a2b}, \ldots, \gamma_{anb}) \]

Mixture model with two components, corresponding to comparisons between linked records and non-linked records [1]

\[ m(g) = \Pr(\gamma_{ab} = g | C_{ab} = 1); \quad u(g) = \Pr(\gamma_{ab} = g | C_{ab} = 0) \]

Component membership determines likelihood of comparison vectors

\[ m_{jk} = \Pr(\gamma_{ab} = h | C_{ab} = 1); \quad u_{jk} = \Pr(\gamma_{ab} = h | C_{ab} = 0), \]

for \( 1 \leq j \leq d \) and \( 1 \leq h \leq k_j \), where comparison \( j \) has \( k_j \) possible levels.

Conditional independence assumed between comparisons

\[ m(\gamma_{ab}) = \prod_{j=1}^{d} \prod_{h=1}^{k_j} m_{jk}(\gamma_{ab} = h); \quad u(\gamma_{ab}) = \prod_{j=1}^{d} \prod_{h=1}^{k_j} u_{jk}(\gamma_{ab} = h) \]

MAP Parameter Estimation

Relative likelihood of component membership used to weight links

\[ w_{ab} = \log(m(\gamma_{ab})/u(\gamma_{ab})) \]

Penalized log likelihood discourages over-linking

\[ \ell(C, m, u; \Gamma) = \sum_{a,b \in A \times B} [C_{ab} \log(m(\gamma_{ab})) + (1 - C_{ab}) \log(u(\gamma_{ab}))] - \theta \sum_{ab} C_{ab} \]

\[ = \sum_{a,b \in A \times B} \log(u(\gamma_{ab})) + C_{ab}(w_{ab} - \theta) \]

Equivalent prior over C is \( p(C) \propto \exp(-\theta \ell) \) where \( \theta = \sum_a \sum_b C_{ab} \) [2].

Compute a local mode by alternating maximization of parameters

(1) Maximize \( C \) holding \( m, u \) constant

(2) Maximize \( m, u \) holding \( C \) constant

Maximize C using linear assignment problem

\[ \max_C \sum_{a,b \in A \times B} C_{ab}(w_{ab} - \theta) \]

subject to \( C_{ab} \in \{0, 1\} \); \( \sum_{b \in B} C_{ab} \leq 1 \forall a \in A \); \( \sum_{a \in A} C_{ab} \leq 1 \forall b \in B \)

Maximization of \( m \) and \( u \) can be done in closed form

\[ m_{jk} = \frac{m_{jk} + 1 - C_{ab}}{\sum_{j} \sum_{k} C_{jk} \gamma_{jk}(\gamma_{jk} = h)} \]

\[ u_{jk} = \frac{u_{jk} + 1 - C_{ab}}{\sum_{j} \sum_{k} C_{jk} \gamma_{jk}(\gamma_{jk} = h)} \]

Post-hoc Blocking

Blocking techniques, restrict the set of record pairs considered (e.g. only considering record pairs from the same geographic area).

Post-hoc blocking: New data-driven approach to unsupervised blocking

(1) Generate \( m, u \) and \( C \) via MAP parameter estimation

(2) Compute \( G \), the adjacency matrix with edges \( G_{ab} = 1 \) if \( \gamma_{ab} > \theta \)

(3) Use the connected components of \( G \) as post-hoc blocks

Restricted MCMC

Approximate posterior distribution estimated using a MCMC algorithm

- Fix all record pairs outside of post-hoc blocks to non-link
- Add/drop/swap link updates performed on each post-hoc block
- \( m \) and \( u \) parameters updated via a Gibbs step

Full Bayesian model need not match model for MAP estimates

Simulation Study

Synthetic dataset with four linking fields with two induced errors [3]

- Construct two files with 50,000 records each
- Initially 2.5 billion comparisons with 25,000 true matches
- Initial blocking on postal code (no induced errors)

Comparison vectors utilize string similarity and binary comparisons

- First Name (string) Compared with 4 levels of string similarity
- Last Name (string) Compared with 4 levels of string similarity
- Age (categorical) Binary comparison
- Occupation (categorical) Binary comparison

Selection of \( \theta \) can be based on computational budget

Post-hoc blocks contained almost all linked record pairs

Record Pairs Blocks True Matches Pairs completeness Reduction ratio

<table>
<thead>
<tr>
<th>Post-code blocking</th>
<th>100, 200, 743</th>
<th>29</th>
<th>25,900</th>
<th>99.5%</th>
<th>99.98%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Post-block (( \theta = 4.9 ))</td>
<td>36, 743, 23, 903</td>
<td>24, 577</td>
<td>95.5%</td>
<td>94.6%</td>
<td></td>
</tr>
</tbody>
</table>

Feasible to estimate an approximate posterior with post-hoc blocking

Convergence not achieved using only postal code blocking

Conclusions

- Threshold \( \theta \) allows trade-off between bias introduced into the estimate and computational constraints.
- Post-hoc blocking allows Bayesian inference to be to be performed on larger problems than was previously possible.

Future Work

- Apply to a broader range of PRL models.
- Expand approach to de-duplication problems.

References


Corresponding Author: Brendan McVeigh, bmcveigh@stat.cmu.edu