A Theory for Record Linkage

Ivan P. Fellegi and Alan B. Sunter (1969)

Sep 06, 2016

Model assumptions

Goal: Given two files L_A and L_B, we want to compare the records from these two files and recognize which pairs relate to the same population unit.

Suppose there are two population A and B whose elements are denoted by a and b respectively.

Assume the records in L_A and L_B are generated from these two population with some errors and incompleteness.

Model assumptions

Define two disjoint sets

$$M = \{(a,b)|a = b, a \in A, b \in B\}$$

and

$$U = \{(a,b)|a \neq b, a \in A, b \in B\}$$

We need to make a decision on whether a pair of records belong to M or U for each comparison.

Model assumptions

Decisions are made based on comparisons

$$\gamma[a,b] = [\gamma^1(a,b), \ldots, \gamma^p(a,b)]$$

A decision rule (linkage rule) L maps a comparison onto a set of probabilities:

$$L(\gamma) = \{P(A_1|\gamma), P(A_2|\gamma), P(A_3|\gamma)\}$$

where A_1, A_2, A_3 are the decisions ”matched”, ”possibly matched”, and ”unmatched” respectively.

How to evaluate rules

We assume that (a,b) are randomly selected from population $A \times B$, therefore it is a random variable, and so is the comparison vector $\gamma[a,b]$ Define two conditional probabilities of γ as

$$m(\gamma) = P(\gamma[a,b] | (a,b) \in M)$$

and

$$u(\gamma) = P(\gamma[a,b] | (a,b) \in U)$$

Then we have two types of error associated with a linkage rule

$$P(A_1|U) = \sum_{\gamma} u(\gamma) P(A_1|\gamma)$$

and

$$P(A_3|M) = \sum_{\gamma} m(\gamma) P(A_3|\gamma)$$

Optimal linkage rule

A linkage rule is denoted by $L(\mu, \lambda)$ if

$$P(A_1|U) = \mu$$

and

$$P(A_3|M) = \lambda$$

Among the class of rules at the same level, we say a linkage rule $L(\mu, \lambda)$ is the **optimal rule** , if the relation

$$P(A_2|L) \leq P(A_2|L')$$

holds for every $L'(\mu, \lambda)$

Find the optimal linkage rule at level (μ, λ)

We first order the comparison vectors γ in such a way that the corresponding sequence of

$$m(\gamma)/u(\gamma)$$

is monotone decreasing. When the value is the same, we order them arbitrarily.
We index these vectors as γ_i, $i = 1, 2, \ldots, N_\Gamma$, and also $m_i = m(\gamma_i)$, $u_i = u(\gamma_i)$. Then we choose n and n' such that

$$\sum_{i=1}^{n-1} u_i < \mu \leq \sum_{i=1}^{n} u_i$$

$$\sum_{i=n'}^{N_\Gamma} m_i < \lambda \leq \sum_{i=n'+1}^{N_\Gamma} m_i$$

Find the optimal linkage rule at level (μ, λ)

For γ_i, we assign probabilities to $(P(A_1|\gamma_i), P(A_2|\gamma_i), P(A_3|\gamma_i))$ as

- $(1, 0, 0)$ if $i \leq n - 1$
- $(P_\mu, 1 - P_\mu, 0)$ if $i = n$
- $(0, 1, 0)$ if $n < i \leq n' - 1$
- $(0, 1 - P_\lambda, P_\lambda)$ if $i = n'$
- $(0, 0, 1)$ if $i \geq n' + 1$

where P_μ and P_λ satisfies

$$u_n P_\mu = \mu - \sum_{i=1}^{n} u_i \quad m_{n'} P_\lambda = \lambda - \sum_{i=n'+1}^{N_\Gamma} m_i$$

Theorem

Let $L_0(\mu, \lambda)$ be the linkage rule defined above. Then L_0 is the optimal linkage rule at the levels (μ, λ),

Hypothesis testing view

At the levels (μ, λ), if

$$\mu = \sum_{i=1}^{n} u_i \quad \lambda = \sum_{i=n'}^{N_\Gamma} m_i \quad \text{for some } n < n'$$

the optimal linkage rule becomes

- $(1,0,0)$ if $1 \leq i \leq n$
- $(0,1,0)$ if $n < i < n'$
- $(0,0,1)$ if $n' \leq i \leq N_\Gamma$

And if we define $T_\mu = \dfrac{m(\gamma_n)}{u(\gamma_n)}$ and $T_\lambda = \dfrac{m(\gamma_{n'})}{u(\gamma_{n'})}$, the rule becomes

- $(1,0,0)$ if $T_\mu \leq m(\gamma)/u(\gamma)$
- $(0,1,0)$ if $T_\lambda < m(\gamma)/u(\gamma) < T_\mu$
- $(0,0,1)$ if $m(\gamma)/u(\gamma) \leq T_\lambda$

Application

The large number of possible values of $m(\gamma)$ and $u(\gamma)$ make the computation and storage impractical.
To simplify the model, we assume that all the components of a comparison vector are independent, which gives

$$m(\gamma) = m_1(\gamma^1)m_2(\gamma^2)\dots m_p(\gamma^p)$$

$$u(\gamma) = u_1(\gamma^1)u_2(\gamma^2)\dots u_p(\gamma^p)$$

Suppose the k^{th} component takes n_k different values, then the total number of values of γ is obviously $n_1 n_2 \dots n_p$. With this assumption, we only need to determine $n_1 + n_2 + \cdots + n_p$ values.

Computation of $m_k(\gamma^k)$ and $u_k(\gamma^k)$

Suppose we know the distribution of the population A and B, as well as the probabilities of different types of error introduced into the files, we can calculate $m(\gamma)$ and $u(\gamma)$ directly.
Recall that

$$m_k(\gamma^k) = P(\gamma^k[a,b]|(a,b) \in M)$$

$$u_k(\gamma) = P(\gamma^k[a,b]|(a,b) \in U)$$

We have

$$\begin{aligned} & m(\text{name agrees and is the j}^{th}\text{ listed value}) \\ =& P(\text{name agrees and is the j}^{th}\text{ listed value}|\text{the two records match}) \\ =& P(\text{pick the j}^{th}\text{ listed value from } A \cap B)(1 - P(\text{error})) \end{aligned} \tag{1}$$

And similar for the other quantities of interest.

Computation of $m_k(\gamma^k)$ and $u_k(\gamma^k)$

If we have access to the files, we can compute the following quantities by simply counting:

- M_h: the proportion of ”agreement” in all components except the h^{th}
- U_h: the proportion of ”agreement” in the h^{th} components
- M: the proportion of ”agreement” in all components

Then we have the following equations:

$$N_A N_B E(M_h) = E(N) \prod_{j \neq h} m_j + [N_A N_B - E(N)] \prod_{j \neq h} u_j$$

$$N_A N_B E(U_h) = E(N) m_h + [N_A N_B - E(N)] u_h$$

$$N_A N_B E(M) = E(N) \prod_{j} m_j + [N_A N_B - E(N)] \prod_{j} u_j$$

Computation of $m_k(\gamma^k)$ and $u_k(\gamma^k)$

We have the following equations:

$$N_A N_B E(M_h) = E(N) \prod_{j \neq h} m_j + [N_A N_B - E(N)] \prod_{j \neq h} u_j$$

$$N_A N_B E(U_h) = E(N) m_h + [N_A N_B - E(N)] u_h$$

$$N_A N_B E(M) = E(N) \prod_{j} m_j + [N_A N_B - E(N)] \prod_{j} u_j$$

where

$$m_h = \sum_{\gamma \in S_h} m(\gamma) \qquad u_h = \sum_{\gamma \in S_h} u(\gamma)$$

and S_h is the set of comparison vectors whose h^{th} component is "agreement".

N_A and N_B are the known number of records in files L_A and L_B and N is the unknown number of matched records.

Blocking

When the comparison space Γ is too large, we need to restrict the comparisons to a subspace Γ^*. This can be achieved by partitioning or ”blocking” and making explicit comparisons only between records in corresponding blocks.
All other γ are considered as ”unmatched”, which gives a different levels of error:

$$\mu^* = \mu - \sum_{\Gamma_\mu \cap \bar{\Gamma}^*} u(\gamma) \quad \lambda^* = \lambda + \sum_{\Gamma_\lambda \cap \bar{\Gamma}^*} m(\gamma)$$

where

$$\Gamma_\mu = \{\gamma : T_\mu \leq m(\gamma)/u(\gamma)\} \quad \Gamma_\lambda = \{\gamma : m(\gamma)/u(\gamma) \leq T_\lambda\}$$

Be careful when you make decisions or construct optimal rules.

Choice of comparison space

In practice, we could have many different comparison spaces. The difference will often be the number of configurations of component vectors in addition to simple ”agreement”—”disagreement” configurations (e.g. ”agreement on name John”).

The choice often depends on the specific problem, and we can evaluate the choices by looking at

$$P(A_2|L) = P(T_\lambda < m(\gamma)/u(\gamma) < T_\mu)$$

where $T_\lambda, T_\mu, m(\gamma), u(\gamma)$ are all functions of the comparison space.

Calculation of threshold values

Recall at the levels (μ, λ), if

$$\mu = \sum_{i=1}^{n} u_i \quad \lambda = \sum_{i=n'}^{N_\Gamma} m_i \quad \text{for some } n < n'$$

and we define $T_\mu = \dfrac{m(\gamma_n)}{u(\gamma_n)}$ and $T_\lambda = \dfrac{m(\gamma_{n'})}{u(\gamma_{n'})}$, the rule becomes

- $(1, 0, 0)$ if $T_\mu \leq m(\gamma)/u(\gamma)$
- $(0, 1, 0)$ if $T_\lambda < m(\gamma)/u(\gamma) < T_\mu$
- $(0, 0, 1)$ if $m(\gamma)/u(\gamma) \leq T_\lambda$

We want to determine (estimate) T_μ and T_λ.

Calculation of threshold values

- We sample γ by sampling the configurations for each component independently. The total size of sample is S.
- For the k^{th} component, we sample with probability $z_1^k, z_2^k, \ldots, z_{n_k}^k$ that are roughly proportional to $\frac{m_k(\gamma_{(1)}^k)}{u_k(\gamma_{(1)}^k)}, \frac{m_k(\gamma_{(2)}^k)}{u_k(\gamma_{(2)}^k)}, \ldots, \frac{m_k(\gamma_{(n_k)}^k)}{u_k(\gamma_{(n_k)}^k)}$
- Then we order these samples by decreasing values of $m(\gamma)/u(\gamma)$, and we index the h^{th} vector as γ_h.
- Then $P(\frac{m(\gamma)}{u(\gamma)} < \frac{m(\gamma_h)}{u(\gamma_h)}|\gamma \in M)$ is estimated by

$$\lambda_h = \sum_{i=h}^{S} m(\gamma_i)/\pi(\gamma_i) \quad \text{where} \quad \pi(\gamma_i) = S \prod_{i=1}^{K} z_{h_i}^i$$

- Let $\lambda = \lambda_h$ to solve h, and $\frac{m(\gamma_h)}{u(\gamma_h)}$ is an estimation for T_λ. Similar for T_μ.