Module 6: Exact-Mapping for Entity Resolution

Rebecca C. Steorts

joint with Brian Kundinger and Jerry Reiter

Reading

- Binette and Steorts (2020)
- Newcombe et al. (1959)
- Fellegi and Sunter (1969)
- Kundinger, Reiter, and Steorts (2024)

Notation

Assume two databases A and B (or sometimes called X_1 and X_2.)

- The relative size of each database is N_1 and N_2.
- Assume there are duplicates across the databases but not within them. This is called a bipartite record linkage assumption.
- Assume $f = 1, \ldots F$ fields (or attributes).
- Let L_f denote the number of categories for field f.

Motivation

- Record pairs that refer to the same entity should be similar.
- Records pairs that refer to different entities should be dissimilar.

We can compare record pairs using similarity scores (or distance functions).

Examples: Jaccard, Edit, Jaro, Jaro-Winkler.

Comparison Vectors (or Data)

This motivates the comparison vector or comparison data.

Consider

$$\boldsymbol{\gamma}_{ij} = (\gamma_{ij}^1, \ldots, \gamma_{ij}^F),$$

where

γ_{ij}^f compares field f for record $i \in A$ and $j \in B$.

Collect all the comparison vectors as

$$\boldsymbol{\gamma} = \{\boldsymbol{\gamma}_{ij}\}_{i=1,j=1}^{N_1,N_2}$$

Comparison Vectors (or Data)

The above notation is used in the literature as it is compact and short!

What do these vectors look like in practice?

Comparison Vectors

Let

$$i = 1, 2, \ldots, N_1 \times N_2$$

enumerate the set of all record pairs in $A \times B$.

Comparison Vectors

Let

$$i = 1, 2, \ldots, N_1 \times N_2$$

enumerate the set of all record pairs in $A \times B$.

- For the ith pair of records, we compute a corresponding **comparison vector**

$$\gamma_i = (\gamma_i^{(1)}, \gamma_i^{(2)}, \ldots, \gamma_i^{(F)}).$$

Comparison Vectors

Let

$$i = 1, 2, \ldots, N_1 \times N_2$$

enumerate the set of all record pairs in $A \times B$.

- For the ith pair of records, we compute a corresponding **comparison vector**

$$\gamma_i = (\gamma_i^{(1)}, \gamma_i^{(2)}, \ldots, \gamma_i^{(F)}).$$

- Each γ_i^j compares the jth field of the records.

Comparison Vectors

How can we visualize the comparison vectors?

$$\gamma_1 = (\gamma_1^{(1)}, \gamma_1^{(2)}, \ldots, \gamma_1^{(F)}) \tag{1}$$

$$\gamma_2 = (\gamma_2^{(1)}, \gamma_2^{(2)}, \ldots, \gamma_2^{(F)}) \tag{2}$$

$$\vdots \tag{3}$$

$$\gamma_{(N_1 \times N_2)} = (\gamma_{(N_1 \times N_2)}^{(1)}, \gamma_{(N_1 \times N_2)}^{(2)}, \ldots, \gamma_{(N_1 \times N_2)}^{(F)}) \tag{4}$$

Let

$$\gamma = (\gamma_1^{(1)}, \gamma_2^{(2)}, \ldots, \gamma_{(N_1 \times N_2)}^{(F)})$$

Exact Mapping

We will walk through an approach to create a more efficient representation of the comparison vectors (but not do any dimension reduction).

Notation

- Let P be the number of exact agreement patterns in γ, which is bounded above by $\prod_{f=1}^{F}(L_f + 1)$.
- Consider the following function

$$h_f^{(i,j)} = I_{obs}(\gamma_{ij}^f) 2^{\gamma_{ij}^f + I(f>1) \times \sum_{e=1}^{f-1}(L_e - 1))}, \tag{5}$$

which maps a record pair for a field f to a unique integer.

Summing over fields $f = 1, \ldots, F$ for record pair (i, j) results in

$$h^{(i,j)} = \sum_{f=1}^{F} h_f^{(i,j)}.$$

Notation

- Enumerate unique hashed agreement patterns from 1 to P.
- Denote each unique mapped agreement pattern as $h_p = (h_p^1, \ldots, h_p^F) \implies$ that when record pair (i, j) has agreement pattern p, we write $\gamma_{ij} = h_p$.
- Collect all the agreement patterns as $P = \{h_p \mid p \in [P]\}$.

Notation

- When performing computation, it will be useful to represent a more representative version of h_p, known as a one hot encoding.
- $e(h_p)$ denotes the $\sum_{f=1}^{F} L_f$ length vector where the $\ell + \sum_{f=1}^{F} L_f$ component is 1 when $h_p = \ell$ and otherwise is 0.

Example

Consider five fields with binary agreement patterns (and potential missingness). Suppose that records $(5, 7)$ in have agreement pattern $(1, 1, 1, NA, 2)$, which means

- agreement in the first three fields.
- missingness in the fourth field.
- complete disagreement in the fifth field.

Example

1. Find $h(\gamma_{5,7})$
2. Find $e(h(\gamma_{5,7})$.

Example

Recall that

$$h_f^{(i,j)} = I_{obs}(\gamma_{ij}^f)2^{\gamma_{ij}^f + I(f>1)\times\sum_{e=1}^{f-1}(L_e - 1))} \implies$$

$$h_1^{(5,7)} = 2^1 \tag{6}$$

$$h_2^{(5,7)} = 2^3 \tag{7}$$

$$h_3^{(5,7)} = 2^5 \tag{8}$$

$$h_4^{(5,7)} = 0 \tag{9}$$

$$h_4^{(5,7)} = 2^{10} \tag{10}$$

This implies $h^{(5,7)} = 1066$. Assume this maps to unique integer 42.

Example

How do we create the one hot encoding?

For a binary comparison, we have the following options:

1. (1,0): represents complete agreement of the record pairs.
2. (0,1): represents complete disagreement of the record pairs.
3. (0,0): represents missingness in the record pairs.

Given this,

$$e(h^{(5,7)}) = (1, 0, 1, 0, 1, 0, 0, 0, 0, 1).$$