


Appariements probabilistes - méthode de Fellegi et Sunter

Lucas Malherbe

18 novembre 2021

Introduction

- Les appariements dits "probabilistes" dérivent tous du cadre décrit par Fellegi et Sunter en 1969¹.
- Ils se caractérisent par le calcul d'une probabilité pour chaque paire considérée et par un processus d'inférence bayésienne.

¹Fellegi, Ivan P and Alan B Sunter. 1969. "A theory for record linkage." Journal of the American Statistical Association 64(328):1183-1210. 

Outline

- 1 Calcul des prédictions
- 2 Les poids
- 3 Règle de décision
- 4 Estimation des paramètres
- 5 Appariement probabiliste et indexation

Outline

- 1 Calcul des prédictions
- 2 Les poids
- 3 Règle de décision
- 4 Estimation des paramètres
- 5 Appariement probabiliste et indexation

Notations

- À chaque paire est associé un couple (γ, M) correspondant à la réalisation de deux variables aléatoires :
 - $\gamma = (\gamma_1, \dots, \gamma_K)$ représente le vecteur de comparaison des paires sur K champs présents dans les deux fichiers à appairer.
 - M représente le vrai statut de la paire (1 pour un *match*, 0 sinon)
- Par exemple, avec des comparaisons exactes,
 $\gamma = (\gamma_{nom}, \gamma_{prenom}, \gamma_{date_naiss}, \gamma_{com_naiss}) = (1, 0, 1, 0)$ signifie que les deux individus ont les mêmes nom et date de naissance, mais des prénom et commune de naissance différents.
- Cette partie est consacrée au calcul de $P(M = 1|\gamma)$.

Notations

Deux probabilités conditionnelles jouent un rôle fondamental dans le modèle :

$$m(\gamma) = P(\gamma|M = 1)$$

$$u(\gamma) = P(\gamma|M = 0)$$

- m mesure la qualité d'un champ identifiant.
 u mesure la probabilité d'observer la même valeur sur un champ identifiant par hasard.
- Par exemple, $m_{prenom}(1) = 0,9$ signifie que des erreurs surviennent 1 fois sur 10 sur le champ *prénom*.
La probabilité u dépend de la cardinalité de chaque champ.
Par exemple, $u_{mois_nais} \approx 1/12$ tandis que u_{nom} sera très faible.

Exemple avec un seul champ identifiant

- En l'absence de champ identifiant :
 - γ est vide et la probabilité devient $P(M = 1)$. On note λ cette quantité.
 - Elle correspond à la proportion de *matches* parmi l'ensemble des paires, de l'ordre de $1/n$. Il s'agit de l'*a priori*.
- Avec un unique champ identifiant :
Par le théorème de Bayes,

$$\begin{aligned}
 P(M = 1 | \gamma_1) &= \frac{P(\gamma_1 | M = 1) \cdot P(M = 1)}{P(\gamma_1 | M = 1) \cdot P(M = 1) + P(\gamma_1 | M = 0) \cdot P(M = 0)} \\
 &= \frac{m_1(\gamma_1) \cdot \lambda}{m_1(\gamma_1) \cdot \lambda + u_1(\gamma_1) \cdot (1 - \lambda)}
 \end{aligned}$$

Hypothèse d'indépendance conditionnelle

Hypothèse

On suppose que les composantes du vecteur de comparaison γ sont mutuellement indépendantes conditionnellement au vrai statut de la paire M .

Cette hypothèse simplifie nettement les calculs et implique notamment :

$$m(\gamma) = m_1(\gamma_1)m_2(\gamma_2) \dots m_K(\gamma_K)$$

$$u(\gamma) = u_1(\gamma_1)u_2(\gamma_2) \dots u_K(\gamma_K)$$

Exemple avec deux champs identifiants

Une première manière d'envisager les choses est de façon séquentielle :

$$P(M = 1 | \gamma_1, \gamma_2) = \frac{\tilde{\lambda} \cdot m_2(\gamma_2)}{\tilde{\lambda} \cdot m_2(\gamma_2) + (1 - \tilde{\lambda}) \cdot u_2(\gamma_2)}$$

$$\text{avec } \tilde{\lambda} = P(M = 1 | \gamma_1) = \frac{\lambda \cdot m_1(\gamma_1)}{\lambda \cdot m_1(\gamma_1) + (1 - \lambda) \cdot u_1(\gamma_1)}$$

Exemple avec deux champs identifiants

La seconde manière consiste à calculer directement :

$$\begin{aligned} P(M = 1 | \gamma_1, \gamma_2) &= \frac{P(M = 1) \cdot P(\gamma_1, \gamma_2 | M = 1)}{P(\gamma_1, \gamma_2)} \\ &= \frac{\lambda \cdot m(\gamma_1, \gamma_2)}{\lambda \cdot m(\gamma_1, \gamma_2) + (1 - \lambda) \cdot u(\gamma_1, \gamma_2)} \\ &= \frac{\lambda \cdot m_1(\gamma_1) \cdot m_2(\gamma_2)}{\lambda \cdot m_1(\gamma_1) \cdot m_2(\gamma_2) + (1 - \lambda) \cdot u_1(\gamma_1) \cdot u_2(\gamma_2)} \end{aligned}$$

Cas général

Dans le cas général avec K champs identifiants,

$$P(M = 1|\gamma) = \frac{\lambda m_1(\gamma_1)m_2(\gamma_2)\dots m_K(\gamma_K)}{\lambda m_1(\gamma_1)m_2(\gamma_2)\dots m_K(\gamma_K) + (1 - \lambda) \cdot u_1(\gamma_1)u_2(\gamma_2)\dots u_K(\gamma_K)}$$

Outline

- 1 Calcul des prédictions
- 2 Les poids**
- 3 Règle de décision
- 4 Estimation des paramètres
- 5 Appariement probabiliste et indexation

Transformation de la probabilité en cote

Afin d'interpréter l'impact de chaque variable, il est plus facile de raisonner sur les cotes (*odds* en anglais). Cette transformation de la probabilité fait apparaître des poids, qui représentent l'importance relative des colonnes dans la discrimination des paires.

$$\begin{aligned}\frac{P(M = 1|\gamma)}{1 - P(M = 1|\gamma)} &= \frac{\lambda m_1(\gamma_1)m_2(\gamma_2)\dots m_K(\gamma_K)}{(1 - \lambda)u_1(\gamma_1)u_2(\gamma_2)\dots u_K(\gamma_K)} \\ &= \frac{\lambda}{1 - \lambda} w_1(\gamma_1)w_2(\gamma_2)\dots w_K(\gamma_K)\end{aligned}$$

avec $w_j(\gamma_j) = \frac{m_j(\gamma_j)}{u_j(\gamma_j)}$ le poids associé au champ j .

Exemples de poids

■ Pour le prénom

- Un exemple de valeurs plausibles serait $m_{prenom}(1) = 0,8$ et $u_{prenom}(1) = 0,02$.
- Le poids en cas d'accord sur le prénom s'élève à $w_{prenom}(1) = \frac{0,8}{0,02} = 40$
- En cas de désaccord, il vaut $w_{prenom}(0) = \frac{1-0,8}{1-0,02} \approx 0,20$

■ Pour le genre

- Si le champ est de bonne qualité, on peut imaginer $m_{genre}(1) = 0,99$, et $u_{genre}(1) = 0,5$ sur une population équitablement répartie entre les deux genres.
- Le poids en cas d'accord sur le genre s'élève à $w_{genre}(1) = \frac{0,99}{0,5} = 1,98$
- En cas de désaccord, il vaut $w_{genre}(0) = \frac{1-0,99}{1-0,5} = 0,02$

Interprétation

- Poids en cas d'accord sur le champ

$$w_{prenom}(1) = 40 \gg 1,98 = w_{genre}(1)$$

Un accord sur le prénom est beaucoup plus significatif qu'un accord sur le genre.

- Poids en cas de désaccord sur le champ

$$w_{prenom}(0) = 0,20 \gg 0,02 = w_{genre}(0)$$

Un désaccord sur le genre est beaucoup plus significatif qu'un désaccord sur le prénom car le poids fait largement diminuer la cote, et donc la probabilité.

Outline

- 1 Calcul des prédictions
- 2 Les poids
- 3 Règle de décision**
- 4 Estimation des paramètres
- 5 Appariement probabiliste et indexation

Classement des paires

- En sortie de l'appariement, les paires sont classées dans trois ensembles disjoints : celui des *matches* \mathcal{M} , celui des *non-matches* \mathcal{U} et une zone grise de *matches* possibles \mathcal{P} .
- On matérialise cette règle de décision par une fonction d qui à un vecteur de comparaison γ associe l'une de ces trois catégories.

Estimation des erreurs

- Le **taux de faux négatifs** peut être estimé sur l'ensemble des paires par :

$$\begin{aligned} E[\mathbb{1}[d(\gamma) \in \mathcal{U}] | M = 1]) &= \sum_{j=1}^n P(\gamma^j | M = 1) \mathbb{1}[d(\gamma) \in \mathcal{U}] \\ &= \sum_{d(\gamma) \in \mathcal{U}} m(\gamma^j) \end{aligned}$$

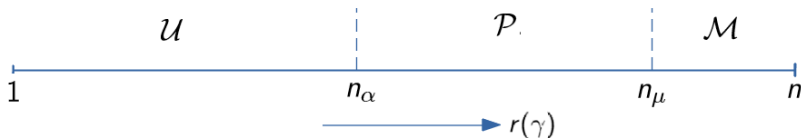
- De même, le **taux de faux positifs** est estimé par :

$$E[\mathbb{1}[d(\gamma) \in \mathcal{M}] | M = 0]) = \sum_{d(\gamma) \in \mathcal{M}} u(\gamma^j)$$

Définition de la règle de décision

- Les paires sont réindexées dans l'ordre décroissant du rapport

$$r(\gamma) = \frac{m(\gamma)}{u(\gamma)}$$



- Étant donné un **niveau toléré de faux négatifs** α , n_α est le plus petit indice vérifiant $\sum_{j > n_\alpha} m(\gamma^j) < \alpha$
- Pour un **niveau toléré de faux positifs** μ , n_μ est le plus grand indice vérifiant $\sum_{j < n_\mu} u(\gamma^j) < \mu$

Optimalité et lien avec la théorie des tests d'hypothèse

- Cette règle de décision est **optimale**, dans le sens où elle minimise la taille de la zone grise à niveaux d'erreur α et μ donnés.
- En considérant le statut d'une paire M comme un paramètre, le classement d'une paire comme un *match* est équivalent à la réalisation d'un **test du rapport de vraisemblance de l'hypothèse $H_0 : M = 1$ contre $H_1 : M = 0$ de niveau α** .
- La statistique de test s'écrit : $\frac{P(\gamma|M=1)}{P(\gamma|M=0)} = \frac{m(\gamma)}{u(\gamma)}$
- D'après le lemme de Neymann-Pearson, **ce test est le plus puissant de niveau α** .
- Le raisonnement est identique pour le **test de l'hypothèse $H_0 : M = 0$ contre $H_1 : M = 1$ de niveau μ** .

Cohérence avec la probabilité estimée

- Le classement des paires s'effectue donc selon l'ordre du rapport $\frac{m(\gamma)}{u(\gamma)} = w_1(\gamma_1)w_2(\gamma_2) \dots w_K(\gamma_K)$
- Il s'agit, à un facteur multiplicatif près, de la côte associée la probabilité $P(M = 1|\gamma)$
- Le passage d'une cote à une probabilité étant une transformation croissante, la règle de décision proposée équivaut à classer les paires selon l'ordre des probabilités.

Outline

- 1 Calcul des prédictions
- 2 Les poids
- 3 Règle de décision
- 4 Estimation des paramètres**
- 5 Appariement probabiliste et indexation

Algorithme EM

- L'algorithme Espérance-Maximisation (EM)² est une adaptation de l'estimation par maximum de vraisemblance en présence de **variables latentes non observables**.
- Dans le cas de l'appariement probabiliste, **les variables observées sont les composantes du vecteur de comparaison γ et la variable latente est le vrai statut de la paire M** .

²Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society: Series B (Methodological), 39(1), 1-22.

Une méthode de référence

- L'estimation des paramètres par l'algorithme EM, proposée par Jaro (1989)³, s'est imposée comme référence dans la littérature des appariements probabilistes.
- Elle s'effectue de façon non supervisée : elle ne nécessite pas d'informations supplémentaires sur les jeux de données.

³Jaro, M. A. (1989). Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida. Journal of the American Statistical Association, 84(406), 414-420.

Estimation par l'algorithme EM

- En notant $\theta = (m, u, \lambda)$ le vecteur des paramètres à estimer, la log-vraisemblance du modèle s'écrit :

$$\log \mathcal{L}(\theta, M_1, \dots, M_n, \gamma^1, \dots, \gamma^n) = \sum_{j=1}^n M_j \log(\lambda \cdot m(\gamma^j)) + (1 - M_j) \log((1 - \lambda) \cdot u(\gamma^j))$$

- Le processus est **itératif**, alternant les phases de **calcul d'espérance de la vraisemblance** et d'ajustement des paramètres par **maximisation de cette quantité**.
- L'**hypothèse d'indépendance conditionnelle** simplifie les calculs et permet d'obtenir des **formules en forme close**.

Autres méthodes d'estimation

- D'autres méthodes existent, la plupart faisant intervenir des **calculs de fréquence**.
- Cependant **elles nécessitent en général des informations** comme la fréquence des noms / prénoms dans une langue.
- Certaines de ces méthodes permettent de **prendre en compte la fréquence des modalités**, de façon à ce qu'un accord sur un nom rare soit associé à un poids plus important qu'un accord sur un nom très fréquent dans la population.

Outline

- 1 Calcul des prédictions
- 2 Les poids
- 3 Règle de décision
- 4 Estimation des paramètres
- 5 Appariement probabiliste et indexation**

Effet de l'indexation sur les paramètres





- La phase d'indexation réduit le champ des paires étudiées et **modifie la distribution** du vecteur de comparaison γ et du statut des paires M .
- Quelle que soit sa forme (blocage ou filtrage plus complexe), **elle modifie la valeur des paramètres estimés**.
- Sans adaptation de l'algorithme d'estimation des paramètres, **les taux d'erreur ne sont plus fiables**.
- **L'indexation reste néanmoins nécessaire** lorsque les fichiers deviennent grands :
 - d'abord pour des raisons opérationnelles de temps de calcul,
 - mais aussi car lorsque proportion de *matches* dans le produit cartésien devient trop faible, l'estimation des paramètres est biaisée.

Adaptations - Jaro

- L'indexation induit particulièrement des variations sur les valeurs des $u_i(\gamma_i)$ et sur l'*a priori* λ parce qu'elle retire essentiellement des paires dont le vrai statut est négatif.
- En général, si l'indexation retire un nombre significatif de paires, λ augmente fortement. Les $u_i(\gamma_i)$ augmentent légèrement car les paires retenues sont globalement plus similaires qu'une paire aléatoire du produit cartésien.
- Jaro (1989) propose d'estimer les paramètres $u_i(\gamma_i)$ sur le produit cartésien et les paramètres $m_i(\gamma_i)$ sur les données indexées.

Adaptations - Murray

- Murray (2016)⁴ étudie de façon approfondie l'effet de l'indexation sur l'estimation des paramètres.
- Les paramètres estimés sont effectivement différents, en revanche le classement des paires (l'ordre des probabilités) peut être conservé sous certaines conditions. C'est le cas par exemple lorsque l'indexation ne fait intervenir que des éléments du vecteur de comparaison γ .
- L'article propose une adaptation de la phase d'estimation des paramètres, principalement en repérant les modalités du vecteur γ qui ne peuvent pas apparaître dans les paires conservées en raison de choix d'indexation et en considérant ces combinaisons comme des zéros structurels.

⁴Murray, J. S. (2016). Probabilistic record linkage and deduplication after indexing, blocking, and filtering. arXiv preprint arXiv:1603.07816.    

Adaptations - Fortini

- Fortini (2020)⁵ traite des défis liés au volume des fichiers à apparier et propose deux adaptations indépendantes au modèle classique :
- D'abord, lors de la phase de comparaison des paires, il s'agit de conserver toutes celles pour lesquelles au plus un champ identifiant diffère, puis de procéder à une estimation par échantillonnage pour les autres modalités du vecteur γ .
- Ensuite, il introduit un algorithme d'estimation EM robuste, qui reste non-biaisé même lorsque la proportion de matches dans le produit cartésien devient très faible. Il consiste à donner plus d'importance aux paires qui correspondent sur la plupart des champs dans l'étape de maximisation.

⁵Fortini, M. (2020). An Improved Fellegi-Sunter Framework for Probabilistic Record Linkage Between Large Data Sets. *Journal of Official Statistics*, 36(4), 803-825.

Conclusion

- Le modèle de Fellegi et Sunter est **ancien mais pourtant encore très présent**, en particulier en statistique publique.
- **La littérature est active sur le sujet**, notamment pour permettre au modèle de gérer des fichiers de volume important.