An extension of Fellegi-Sunter record linkage model for mixed-type data with application to SNDS

Thanh Huan Vo, Guillaume Chauvet, André Happe, Emmanuel Oger, Stéphane Paquelet, Valérie Garès

An extension of Fellegi-Sunter record linkage model for mixed-type data with application to SNDS

Thanh Huan Vo [1], Guillaume Chauvet [2], André Happe [3], Emmanuel Oger [4], Stéphane Paquelet [5] & Valérie Garès [6]

[1] *INSA (IRMAR) and IRT b-com, Rennes, France, E-mail: than-huan.vo@insa-rennes.fr*
[2] *ENSAI (IRMAR), Campus de Ker Lann, Bruz, France, E-mail: guillaume.chauvet@ensai.fr*
[3] *EA 7449 REPERES, France, E-mail: andre.happe@chu-brest.fr*
[4] *EA 7449 REPERES, France, E-mail: emmanuel.oger@univ-rennes1.fr*
[5] *IRT b-com - Institut de Recherche Technologique b-com, France, E-mail: stephane.paquelet@b-com.com*
[6] *Univ Rennes, INSA Rennes, CNRS, IRMAR - UMR 6625, F-35000 Rennes, France, E-mail: valerie.gares@insa-rennes.fr*

Résumé. Le couplage probabiliste d'enregistrements est un processus de combinaison de données provenant de différentes sources, lorsque ces données se réfèrent à des entités communes et que les informations d'identification ne sont pas disponibles. Fellegi et Sunter ont proposé un cadre de couplage probabiliste d'enregistrements quand les informations d'identification ne sont pas disponibles. Cependant, cette méthode n'utilise qu'une comparaison binaire entre les variables d'appariement. Dans nos travaux, nous proposons une extension de ce modèle notamment lorsque les données d'appariement contiennent différents types de variables (binaires, catégorielles et continues). Nous proposons un modèle de mélange de distributions discrètes pour gérer les variables d'appariement catégorielles avec une faible prévalence, et un modèle de mélange de distributions gamma gonflées en zéro pour gérer les variables d'appariement continues. Les estimations du maximum de vraisemblance pour les paramètres du modèle sont obtenues au moyen de l'algorithme "Expectation Conditional Maximization" (ECM). Grâce à une étude de simulation de Monte Carlo, nous évaluons à la fois l'estimation de la probabilité postérieure pour qu'une paire d'enregistrements soit une correspondance, et la qualité de prédiction des paires d'enregistrements appariés. Les premiers résultats de la simulation indiquent que les méthodes proposées donnent de bons résultats par rapport aux méthodes existantes. La prochaine étape consistera à appliquer la méthode proposée à un jeu de données réel, afin de trouver les patients correspondants dans les données des registres SNDS (Système National des Données de Santé) et GETBO (Groupe d'étude de la Thrombose de Bretagne Occidentale).

Mots-clés. Couplage d'enregistrements probabilistes, algorithme ECM, modèle de mélange, loi gamma gonflée en zéro

Abstract. Probabilistic record linkage is a process of combining data from different sources, when such data refer to common entities and that identifying information is not available. Fellegi and Sunter proposed a probabilistic record linkage framework that takes into account multiple non-identifying information but is limited to simple binary comparison between matching variables. In our work, we propose an extension of this model especially when matching data contains different types of variables (binary, categorical and continuous). We develop a model of mixture of discrete distribution for handling comparison values of low prevalence categorical matching variables, and a mixture of hurdle gamma distribution for handling comparison values of continuous matching variables. The maximum likelihood estimates for model parameters are obtained by means of the Expectation Conditional Maximization (ECM) algorithm. Through a Monte Carlo simulation study, we evaluate both the posterior probability estimation for a record pair to be a match, and the prediction of matched record pairs. The first simulation results indicate that the proposed methods perform well as compared to existing methods. The next step will be to apply the proposed method to real datasets, which aim to find corresponding patients in SNDS (Système National des Données de Santé) and GETBO (Groupe d'étude de la Thrombose de Bretagne Occidentale) register data.

Keywords. Probabilistic record linkage, ECM algorithm, mixture model, hurdle gamma distribution

1 Introduction

Electronic health records have become more and more popular in medical fields, and the ability to exchange this information can help in providing better care for patients as well as richer sources for researchers. Record linkage is a process of combining data from different sources that refer to the same entity. The process is straightforward if each record contains a unique identifier such as Social Security Number. However, some large health databases may not contain such identifying information. Therefore, Fellegi and Sunter (1969) proposed a probabilistic record linkage framework that takes into account multiple non-identifying information such as names, and postal code.

Although this model is widely performed in many applications, when unique identifiers are unavailable or when data contain errors, its simple binary comparison has a limitation when some matching variables are binary and with a low prevalence (e.g. medical diagnoses, see Hejblum et al., 2019). Another limitation is that most probabilistic record linkage models only make use of simple binary or categorical comparison values even if the matching variables are continuous.

In this article, we propose a linkage model adapted from Fellegi and Sunter framework and which handle such cases. We aim at better taking into account the nature of matching

variables (e.g., low-prevalence binary, or continuous), so as to improve the performances of record linkage.

2 Probabilistic record linkage model

Consider two databases A and B containing n_A and n_B records respectively, and with elements in common. Following the terminology in Fellegi and Sunter (1969), each possible record pair $(X_{A,i}, X_{B,j})$ with

$$X_{A,i} = (X^1_{A,i}, \ldots, X^K_{A,i}) \in A, i = 1, \ldots, n_A,$$
$$X_{B,j} = (X^1_{B,j}, \ldots, X^K_{B,j}) \in B, j = 1, \ldots, n_B$$

either belongs to the set of true matched pairs noted by M, or to the set of true unmatched pairs noted by U.

The strategy begins by comparing K matching variables of all records $X_{A,i}$, with all records $X_{B,j}$ leading to $n_A \times n_B$ comparison vectors $\gamma_{ij} = \left\{\gamma^1_{ij}, \ldots, \gamma^k_{ij}, \ldots, \gamma^K_{ij}\right\}$, where $\gamma^k_{ij} = h^k(X^k_{A,i}, X^k_{B,j})$ and h^k is a comparison function for the $k-$th matching variable which can be defined in different ways depending on the type of matching variables (see Christen, 2012). The most common way consists in a binary comparison, i.e.

$$\gamma^k_{ij} = h^k(X^k_{A,i}, X^k_{B,j}) = \begin{cases} 1 & \text{if} \quad X^k_{A,i} = X^k_{B,j}, \\ 0 & \text{if} \quad X^k_{A,i} \neq X^k_{B,j}. \end{cases} \tag{1}$$

Because we assumed that each record pair belongs to one of two latent classes (the matched pairs M or the unmatched pairs U), the distribution of comparison vectors γ for each pair is assumed to follow a mixture model

$$\mathbb{P}(\gamma) = \mathbb{P}(\gamma|M)\mathbb{P}(\gamma \in M) + \mathbb{P}(\gamma|U)\left[1 - \mathbb{P}(\gamma \in M)\right]. \tag{2}$$

Once all the parameters of the model are estimated, the record pairs may be ordered and classified into matches, non-matches or possible matches based on either matching weights $\frac{\mathbb{P}(\gamma_{ij}|M)}{\mathbb{P}(\gamma_{ij}|U)}$ or posterior probabilities of matching $q_{ij} \equiv \mathbb{P}(M|\gamma_{ij}) = \frac{\mathbb{P}(\gamma_{ij}|M)p}{\mathbb{P}(\gamma_{ij}|M)p + \mathbb{P}(\gamma_{ij}|U)(1-p)}$. Although the matching scores and the posterior probabilities produce the same ordering for record pairs (Larsen and Rubin, 2001), the posterior probabilities are preferable in our application because they may be useful for further analyses (Lahiri and Larsen, 2005).

3 An extension of the Fellegi-Sunter model

In this article, we aim at developing the Fellegi-Sunter model by making better use of low prevalence categorical matching variables and of continuous variables.

3.1 Comparison approaches

Let X^k be a categorical matching variable taking L different values, which means that the comparison function for this variable may take L^2 values. For example, a comparison of a binary matching variable may lead to four possible realizations and a comparison function can be defined as follows

$$h^k(0,0) = 0,\ h^k(0,1) = 1,\ h^k(1,0) = 2, \quad \text{and} \quad h^k(1,1) = 3. \tag{3}$$

Because the agreement on the low prevalence value is much more informative than the agreement on the others, our comparison approach aims at using this information while the simple binary comparison (1) method does not distinguish them. Hejblum et al. (2018) propose a Bayesian record linkage framework making use of a similar idea, and which is efficient in case of a large number of low-prevalence binary matching variables. However, their model is designed for binary variables only, while our comparison approach can be combined with other types of matching variables (e.g., continuous).

If the number of matching variables and/or the number of categories is large, the number of parameters to be estimated is L^2-1, which may be too large in practice. This number may be reduced by assigning a same comparison value for the agreement/disagreement of categories which have roughly a same proportion. For instance, we may reduce the comparison values given in (3) as

$$h^k(0,0) = 0,\ h^k(0,1) = h^k(1,0) = 1, \quad \text{and} \quad h^k(1,1) = 2. \tag{4}$$

Now, let us consider the case of a continuous variable X^k. For example date variables (e.g., admission to the hospital, or medical act) are common in medical datasets. They may be seen as continuous counting variables, by converting each date into a duration from a specified origin. Even if an individual is present in both datasets, a lag between dates is likely to appear. The simple binary comparison is therefore not appropriate. In this article, we propose to define $\gamma_{ij}^k = d(X_{A,i}^k, X_{B,j}^k)$, where d is a predefined distance, which can naturally take into account the time lags.

In summary, the comparison vectors in our model can include both categorical and continuous comparison values.

3.2 Estimation of parameters

Let

$$\gamma_{ij} \;=\; \left(\gamma_{ij}^1, \ldots, \gamma_{ij}^{K_1}, \gamma_{ij}^{K_1+1}, \ldots, \gamma_{ij}^{K_1+K_2}\right) \tag{5}$$

be a mixed type comparison vector which includes K_1 categorical comparison values $\gamma_{ij}^1, \ldots, \gamma_{ij}^{K_1}$ and K_2 continuous distances $\gamma_{ij}^{K_1+1}, \ldots, \gamma_{ij}^{K_1+K_2}$. Following the Fellegi-Sunter framework, these comparison vectors are assumed to follow the mixture model (2).

Under the conditional independence assumption between fields of the comparison vector (Winkler, 2000), we have

$$\mathbb{P}(\gamma_{ij}|M) = \prod_{k=1}^{K_1} \mathbb{P}(\gamma_{ij}^k|M) \prod_{k=K_1+1}^{K_1+K_2} \mathbb{P}(\gamma_{ij}^k|M) \tag{6}$$

$$\mathbb{P}(\gamma_{ij}|U) = \prod_{k=1}^{K_1} \mathbb{P}(\gamma_{ij}^k|U) \prod_{k=K_1+1}^{K_1+K_2} \mathbb{P}(\gamma_{ij}^k|U) \tag{7}$$

For the first term in the right-hand side involving K_1 categorical comparison values of the comparison vector γ_{ij}, we define

$$m_s^k = \mathbb{P}(\gamma_{ij}^k = s|M), \sum_{s\in S^k} m_s^k = 1, \text{ and } u_s^k = \mathbb{P}(\gamma_{ij}^k = s|U), \sum_{s\in S^k} u_s^k = 1$$

for S^k is the set of all possible categorical comparison values for the k^{th} variable. Then we have

$$\mathbb{P}(\gamma_{ij}^k|M) = \prod_{s\in S^k} (m_s^k)^{\mathbb{1}_{\gamma_{ij}^k=s}}, \quad \text{and} \quad \mathbb{P}(\gamma_{ij}^k|U) = \prod_{s\in S^k} (U_s^k)^{\mathbb{1}_{\gamma_{ij}^k=s}} \quad \text{for} \quad k = 1, \ldots, K_1.$$

For the second part in the right-hand side of equations (6) and (7) which involves K_2 continuous values of the comparison vector γ, we define

$$\mathbb{P}(\gamma_{ij}^k|M) \sim f^k(\phi_M^k),$$
$$\text{and} \quad \mathbb{P}(\gamma_{ij}^k|U) \sim f^k(\phi_U^k)$$

for $k = K_1 + 1, \ldots, K_2$. The distribution f^k needs to be postulated, depending on the characteristics of the continuous matching variables and the chosen distance. In our simulation studies, we model f^k by means of a hurdle Gamma distribution. To find the maximum likelihood estimates for parameters, we apply the the Expectation Conditional Maximization (ECM) algorithm (Meng and Rubin, 1993).

4 Simulation studies

For ease of interpretation, our proposed approaches are evaluated and compared to existing approaches over scenarios for binary and continuous variables separately. We will consider two databases A and B containing n_A and n_B individuals and K matching variables. We assume that there is no duplicate in both databases and that all individuals in B have corresponding individuals in A. To be realistic, we also introduced errors for data in B and the distribution of error depends on type of each matching variable.

When there are only binary matching variables, we compare the method proposed by Hejblum et al. (2019) to the Fellegi-Sunter model with different comparison methods (1), (4) and (3). When there are only continuous matching variables, we compare the Fellegi-Sunter model using hurdle gamma distribution for continuous comparison values to this model using discrete distribution for categorical comparison values.

The record linkage procedures are evaluated by means of two common criteria for an imbalance classification problem which are True positive rate (Sensitivity or Recall) and Positive predictive value (Precision). From the record linkage results of different considered cases, there is a significant improvement of the Fellegi-Sunter model with our proposed comparison approaches compared to the model with simple binary comparison.

5 Application

The SNDS database is the French national health database that includes all health insurance and hospital data. It is therefore of major interest for research. The GETBO is a registry database that collects information of venous thromboembolism cases in Brest, France. These databases have common information on demographic data such as month and year of birth, gender and some medical acts. The objective is to link the registry data to SNDS at the patient level, when no common individual identifier is available. Our proposed approach will be performed on these databases.

Bibliography

Christen, P. (2012). Data matching: concepts and techniques for record linkage, entity resolution, and duplicate detection, *Springer Publishing Company, Incorporated*.
Fellegi, I. and Sunter, A. (1969). A theory for record linkage, *Journal of the American Statistical Association*, 64, pp. 1183-1210.
Hejblum, B., Weber, G., Liao, K., Palmer, N., Churchill, S., Shadick, N., Szolovits, P., Murphy, S., Kohane, I., and Cai,T. (2019). Probabilistic record linkage of de-identified research datasets with discrepancies using diagnosis codes, *Scientific Data*.
Lahiri, P. and Larsen, M.B. (2005). Regression analysis with linked data, *Journal of the American Statistical Association*, 100 (469), pp. 222-230.
Larsen, M.B. and Rubin, D.B. (2001). Iterative automated record linkage using mixture models, *Journal of the American Statistical Association*, 96 (453), pp. 32-41.
Meng, X. and Rubin, D. (1993). Maximum Likelihood Estimation via the ECM Algorithm: A General Framework. *Biometrika*, 80(2), pp. 267-278.
Winkler, W.E. (2000). Using the EM algorithm for weight computation in the Fellegi-Sunter model of record linkage. *U.S. Bureau of the Census*.