International Journal of Medical Informatics 109 (2018) 70–75

Contents lists available at ScienceDirect

International Journal of Medical Informatics

journal homepage: www.elsevier.com/locate/ijmedinf

A new computationally efficient algorithm for record linkage with field dependency and missing data imputation

John Ferguson[a,*], Ailish Hannigan[b], Austin Stack[b,c]

[a] Clinical Research Facility, National University of Ireland, Galway, Ireland
[b] Graduate Entry Medical School, University of Limerick, Ireland
[c] Health Research Institute, University of Limerick, Ireland

ARTICLE INFO

Keywords:
Record linkage
Fellegi/Sunter
Conditional independence
EM-algorithm
Log-linear models

ABSTRACT

Record linkage algorithms aim to identify pairs of records that correspond to the same individual from two or more datasets. In general, fields that are common to both datasets are compared to determine which record-pairs to link. The classic model for probabilistic linkage was proposed by *Fellegi and Sunter* and assumes that individual fields common to both datasets are completely observed, and that the field agreement indicators are conditionally independent within the subsets of record pairs corresponding to the same and differing individuals. Herein, we propose a novel record linkage algorithm that is independent of these two baseline assumptions. We demonstrate improved performance of the algorithm in the presence of missing data and correlation patterns between the agreement indicators. The algorithm is computationally efficient and can be used to link large databases consisting of millions of record pairs. An R-package, corlink, has been developed to implement the new algorithm and can be downloaded from the CRAN repository.

1. Introduction

Record linkage is a term used to describe techniques to link two or more datasets, or sometimes to remove duplicate records from a single database, in situations where there is absence of a single unique identifier [1]. These techniques are of particular interest in health research, where often several related databases need to be linked together to facilitate the conduct of epidemiological studies in population health [2], but are also useful in many other fields including fraud-detection [3], linking various socioeconomic databases collected via census [4] and national security [5]. The applications and utility of record linkage is likely to increase in the future giving the rising interest in exploiting secondary data sources for research, both in health research as well as other fields. For example, a vast volume of data is currently collected through reward cards, smart devices (such as phones, watches and health censors), and website traffic. In general, linking two databases is only possible if they share a number of common fields (quasi-identifiers) that have some discriminative power in identifying unique members of the population in question. In population health research, the fields usually include demographic identifiers such as name, address, age, sex and race.

Record linkage methods can be broadly split into deterministic and probabilistic techniques. Deterministic methods (see [6] for example) determine that two records are a match, when a combination of prespecified fields show a high level of agreement. For instance, a simple deterministic matching strategy might be to decide a pair of records represents a match only if they agree exactly on forename, surname and date of birth. Probabilistic matching is a more sophisticated approach where separate probability models for the agreement patterns of fields for a given record pair are posited, both in the case that the record pair truly constitutes a match, and in the case that the record pair constitutes a mismatch. In general, probabilistic matching gives more reliable and accurate results compared with deterministic matching [7]. Fellegi and Sunter [8] developed the first widely utilized mathematical model for probabilistic record linkage and their theoretical framework and notation has heavily influenced this area of data science (see [9], for example). The probabilistic approach facilitates the creation of a matching score that can be assigned thresholds so that the record pairs having scores above a specific threshold are considered matches. When these procedures are viewed within a Bayesian framework, this score can often be transformed into an estimated posterior probability that the two records match [10].

At present, there are a wide range of probabilistic methods for record linkage, and the underlying probabilistic model usually is a simple extension of the Fellegi and Sunter model, [8]. This assumes conditional independence given match status in the agreement patterns of the two

[*] Corresponding author.
E-mail address: john.ferguson@nuigalway.ie (J. Ferguson).

https://doi.org/10.1016/j.ijmedinf.2017.10.021
Received 25 April 2017; Received in revised form 8 September 2017; Accepted 31 October 2017

records across different fields. It has been observed, however, that this assumption is often not true for example in databases containing names and addresses, as strong dependencies have been observed between agreements on fields such as surname, house number, street name, and phone number [16]. When conditional independence is assumed in these situations, the EM algorithm does not yield accurate estimates of the parameters associated with the underlying probability distributions. Moreover, if these parameters are used, then decision rules are sub-optimal and error rates cannot be accurately determined [17]. Although some extensions of the Fellegi and Sunter model have been suggested to allow dependence ([11–14]), software is not widely available, and as a result these methods may be difficult to implement for most users of record linkage. Furthermore, missing data is also a challenge for record linkage with record pairs that have missing data in any record linking field often removed or ignored by commonly used linkage algorithms. This may therefore result in valid matching record pairs being missed [18].

Herein, we propose an innovative computationally efficient record linkage procedure to incorporate field dependency into estimated posterior probabilities based on an EM algorithm. We apply the procedure to real and simulated data and compare this approach to the basic model of [8], both in the presence of absence of conditional independence. Comparisons are made regarding classification performance (using receiving operating characteristic curves) as well as calibration accuracy of estimated posterior probabilities generated from the algorithm. We also propose a flexible data-imputation algorithm as a solution to missing fields in particular records, which is designed to be run as a pre-processing step before the main algorithm. Together, these new developments represent a complete algorithm for record linkage which can be implemented using a new R package corlink.

The rest of this paper proceeds as follows. In Methods and Materials, we introduce notation and describe the models and methods used for our algorithms. In Results, we demonstrate the application of these methods to the linkage of hospital laboratory and mortality data based on a number of noisy demographic identifiers. The potential improvement possible from correctly modelling the correlations between the different identifiers is demonstrated with simulations. In the final section, we summarize the novel aspects of our approach, discuss related approaches and suggest some possible improvements to the algorithm for future investigation.

2. Material and methods

2.1. Notation

For consistency, we use some of the notation from [9]. Solving a matching problem requires us to find the subset of members of some set A that represent the same individual as a member, b, of another set B. The sets A and B usually represent different databases that need to be linked – the members of each set representing records in the database. More formally, one can represent the Cartesian product as a partition $\boldsymbol{A} \times \boldsymbol{B} = \boldsymbol{M} \cup \ \ \boldsymbol{U}$, with $\boldsymbol{M}$ representing the record pairs, one from $\boldsymbol{A}$ and one from $\boldsymbol{B}$, that constitute the same individual and $\boldsymbol{U}$, the record pairs that do not. A number of ‘fields’: 1,…,K, are available to help match records in $\boldsymbol{A}$ and $\boldsymbol{B}$. We assume these fields, which may represent identifiers such as name, address, date of birth, are present in both the databases. For a particular record pair, j, where $j \in \{1, ..., |\boldsymbol{A} \times \boldsymbol{B}|\}$, we write $\gamma_j^k = 1$ if field k, $(i \in \{1,...,K\})$ agrees for the two records constituting pair j, and 0 if the information in the field doesn't agree. For example, if field 1 represents an individual's forename, we would write $\gamma_j^1 = 1$ if the recorded forenames for record pair j are identical.

Algorithm A: Missing Data Algorithm

Due to incomplete data entry or patient non-report, not all identifiers may be available for a particular record. Our first step is a type of data-imputation for record pairs where at least one of the K fields are missing. Consider the vector $(Z_j^1, ..., Z_j^K) \in \{0,1,2\}^K$, with $Z_j^k = 2$ representing missingness of field k ($k \le K$) in at least one of the records from record pair j; $Z_j^k = 1$indicating that field k is observed and in agreement for both records in record pair j; and 0 representing field k being observed in both records in record pair j but having different values. Note that we could extend the definition of Z_j^k by creating values of Z_j^k to represent missing in database A, but observed in database B, or missing in database B but observed in database A, or missing in both databases. In practice however, a reduced state space having 3^K, rather than 5^K, combinations will be more computationally efficient when K is large. In addition, the event that a field is missing on both records is likely to be rare, so potentially not much information is lost. Probabilities of missingness for identifiers used in the linkage of real data are given in the Results Section.

Denote $M_{z_1,...z_K} = \sum_{j=1}^{|\boldsymbol{M} \cup \boldsymbol{U}|} I\left\{(Z_j^1, ..., Z_j^K) = (z_1, ...z_K)\right\}$ for each possible pattern $(z_1,...,z_K) \in \{0,1,2\}^K$, $I(A)$ being the indicator function of the event A. The vector $\{M_{z_1,...z_K}\} for (z_1,...,z_K) \in \{0,1,2\}^K$represents how many times each possible $(z_1,...,z_K)$ pattern was observed in the data, and can be considered the raw data for our linkage algorithm. In the presence of missing data, the agreement pattern $(i_1,...,i_K) \in \{0,1\}^K$ that would have been observed if there was no missing data for the record can be considered a latent variable. Given this, denote $N_{i_1,...,i_K} = \sum_{j=1}^{|\boldsymbol{M} \cup \boldsymbol{U}|} I\left\{(\gamma_j^1, ..., \gamma_j^K) = (i_1, ...i_K)\right\}$ for each possible pattern $(i_1,...,i_K) \in \{0,1\}^K$. Note that the vector of counts $\{N_{i_1,...,i_K}\}$ represents the frequencies of each agreement pattern that would be observed if no records had missing identifiers. Note that$M_{z_1,...z_K}$ is observed, whereas $N_{i_1,...,i_K}$ is, in general, latent or unobserved (unless of course no records have missing data). Let $p_{i_1...i_K}$ denote the probability of the pattern i_1 … i_K under the assumption of no missing data. The first objective is to predict $N_{i_1,...,i_K}$ or equivalently to estimate $p_{i_1...i_K}$, based on the observed$M_{z_1,...z_K}$. One can estimate $p_{i_1...i_K}$ using the EM algorithm [19], at each step iterating between calculating the expected value, $E(N_{i_1,...,i_K} | p_{i_1...i_K}^{old})$, of $N_{i_1,...,i_K}$ given the current estimates of $p_{i_1...i_K}^{old}$ (that is the E-step) and choosing $p_{i_1...i_K}$ to maximize the expectation of the complete loglikelihood for a saturated log-linear model (that is the M-step):

$$l(\boldsymbol{p}) = \sum_{\{(i_1...i_K) \in \{0,1\}^K\}} E(N_{i_1,...,i_K} \left| p_{i_1...i_K}^{old}) \log(p_{i_1...i_K}) \right.$$

To facilitate the flow of material, we omit the details of the EM algorithm here, and refer the reader to the Supplementary Material for a detailed description of the E-step and M step.

The estimated proportions $\hat{p}_{i_1...i_K}$ and associated predicted counts $\hat{N}_{i_1...i_K}$, from the final step of the EM algorithm are the input for the main linkage algorithm to be described next. For simplicity, we remove the hat symbols when referring to $\hat{N}_{i_1...i_K}$ and $\hat{p}_{i_1...i_K}$ in the following, even though they may be estimated through this initial EM algorithm.

Algorithm B Linkage Algorithm

The goal of probabilistic linkage is to generate posterior probabilities that a record pair constitutes a true match given a particular agreement pattern: $(i_1,...,i_K) \in \{0,1\}^K$. Different probabilistic linkage algorithms vary regarding how these models are specified. With this in mind, we denote:

$\pi_M = |\boldsymbol{M}|/|\boldsymbol{M} \cup \ \ \boldsymbol{U}|$ as the proportion of record pairs that refer to the same individual (1)

$\pi_U = 1 - \pi_M$ as the proportion of record pairs that refer to different individuals.

Suppose, we posit separate probability models for $\left(\gamma_j^1, ..., \gamma_j^K\right)$ when $j \in \boldsymbol{M}$ and when $j \in \boldsymbol{U}$: