Journal of Biomedical Informatics 44 (2011) 648–654

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

Controlling false match rates in record linkage using extreme value theory

Murat Sariyar *, Andreas Borg, Klaus Pommerening

Institute of Medical Biostatistics, Epidemiology and Informatics (IMBEI), University Medical Center of the Johannes Gutenberg University Mainz, Germany

ARTICLE INFO

Article history:
Received 26 August 2010
Available online 23 February 2011

Keywords:
Data cleansing
Fellegi–Sunter model
Generalized Pareto distribution
Statistics of extreme values
Mean excess plot

ABSTRACT

Cleansing data from synonyms and homonyms is a relevant task in fields where high quality of data is crucial, for example in disease registries and medical research networks. Record linkage provides methods for minimizing synonym and homonym errors thereby improving data quality. We focus our attention to the case of homonym errors (in the following denoted as 'false matches'), in which records belonging to different entities are wrongly classified as equal. Synonym errors ('false non-matches') occur when a single entity maps to multiple records in the linkage result. They are not considered in this study because in our application domain they are not as crucial as false matches. False match rates are frequently computed manually through a clerical review, so without modelling the distribution of the false match rates a priori. An exception is the work of Belin and Rubin (1995) [4]. They propose to estimate the false match rate by means of a normal mixture model that needs training data for a calibration process. In this paper we present a new approach for estimating the false match rate within the framework of Fellegi and Sunter by methods of Extreme Value Theory (EVT). This approach needs no training data for determining the threshold for matches and therefore leads to a significant cost-reduction. After giving two different definitions of the false match rate, we present the tools of the EVT used in this paper: the generalized Pareto distribution and the mean excess plot. Our experiments with real data show that the model works well, with only slightly lower accuracy compared to a procedure that has information about the match status and that maximizes the accuracy.

1. Introduction

Record linkage is a mean for cleansing (personal) data and deals either with deduplication of data in one source or with linking data from different sources into one main target file. Record linkage is applied in biomedical databases such as disease registries or Master Patient Indices and in general whenever medical data about one patient needs to be linked with the same patient's data from other sources. Our motivation for research in this area is the use in epidemiological cancer registries where duplicate entries (e.g. due to typing errors) lead to incorrect incidence estimations. As Duvall et al. [1] point out, unlinked patient data can also have a negative impact on medical treatment in a clinical context. Apart from the medical domain, record inkage methods are used in the context of other personal data such as customer databases. For a general introduction to the problem of duplicate detection and a literature survey we refer to [2].

The goal of record linkage is to remove synonym and to prevent homonym errors. Synonym errors are made when two or more records belonging to a single entity are classified as representing different entities. Homonym errors occur when data of two or more different entities are classified as equal. In the following, synonym and homonym errors are denoted by the more technical terms 'false non-match' and 'false match', respectively. Preliminary steps for record linkage are creating record pairs and transforming these record pairs into comparison patterns. After these steps, deduplication of one dataset and linking of two datasets are handled in the same way as there is no reference to the sources from then on.

The problem of record linkage can be handled by means of non-stochastic classification methods or by application of probability theory. The basic model for probabilistic methods in record linkage is outlined by Fellegi and Sunter [3]. They assume the existence of conditional probabilities $P(\gamma|Z=1)$, $P(\gamma|Z=0)$, where $\gamma=(\gamma_1,\ldots,\gamma_n)$ is a comparison pattern with components γ_j, $j=1,\ldots,n$, and Z a binary variable that assumes the values 1 for matches and 0 for non-matches. The usual approach for controlling false match rates in the Fellegi–Sunter model needs a stochastic independency assumption of the attributes γ_j, $j=1,\ldots,n$, and does not work very well in practice (see [4]), therefore a clerical review is often performed for the determination of the false match rates. We can dispense with the independency assumption through the application of an EM-algorithm for a discrete probability distribution for γ (we refer to [5] for further details).

* Corresponding author. Address: Institut für medizinische Biometrie, Epidemiologie und Informatik, Universitätsmedizin der Johannes Gutenberg-Universität Mainz, Gebäude 902, Obere Zahlbacher Straße 69, 55131 Mainz, Germany.
E-mail address: murat.sariyar@unimedizin-mainz.de (M. Sariyar).

1532-0464/$ - see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.jbi.2011.02.008

A logarithm of the likelihood ratio, called global weight, is used to determine the matching status of object pairs:

$$w_\gamma = \log\left(\frac{P(\gamma|Z=1)}{P(\gamma|Z=0)}\right).$$

Concerning this global weight, a threshold has to be set in order to classify the underlying comparison vector. A vector will be classified as match if the corresponding weight is greater than the threshold and as non-match otherwise. A frequently used variant of the outlined approach consists of the definition of two thresholds, which allows a range of non-determination, i.e. a clerical review would be necessary to determine the match status of cases with weights between the lower and the upper threshold.

The results of a record linkage procedure are called 'links' and 'non-links'. Links and non-links derive from classifying weights as representing matches and non-matches, respectively. False matches are links that are non-matches in fact. False non-matches are non-links that are matches in fact.

In a situation where primarily false matches have to be avoided, the right tail of the weights distribution is of concern. By calibrating the threshold we try to attain an error rate below a predefined error level just the same as in common statistical tests. To this end we derive estimates that bound the false match rate. False non-matches are not considered because in our applications they are not as relevant as false matches and because they are a byproduct of calibrating the threshold.

The usage of EVT for the problem of controlling the false match rate seems to be a promising alternative to usual statistic procedures and models. In contrast to 'traditional' statistics, EVT does not deal with the central or most probable areas of a distribution, but with extreme values or with tails. Thus, not the whole empirical distribution, but only its tails, have to be modelled by a theoretical probability distribution.

This paper is inspired by the work of Belin and Rubin [4] and can be viewed as an attempt to overcome some difficulties when using a mixed normal model that needs training data for estimating the false match rate. When applying EVT we do not need training data for the determination of the threshold. In the subsequent sections, we will first of all deal with the definition of the false match rate and the general estimation problem in this context. Then the theoretical outline for estimating the false match rate via concepts and tools of EVT will be given. The description and results of an empirical evaluation with data stemming from a German cancer registry are followed by comments and conclusions of the results. Finally, we shall give a summary of the main results, their implications, and an outlook for further research topics in connection with controlling of false match rates in record linkage.

2. Materials and methods

2.1. False match rates in record linkage

In [4] the false match rate is defined as the proportion of false matches to the whole number of links (i.e. data pairs classified as matches): $\breve{\epsilon}$. At least the same emphasis should be given to the definition of the false match rate as the proportion of false matches to the number of non-matches: ϵ as in [3] or [6]. In the following, we consider both definitions. ϵ seems more appropriate to evaluate methods because $\breve{\epsilon}$ does not give any hints how good a method has separated matches and non-matches. In the 2×2-contingency Table 1 listed below, where the columns indicate the number of entities with the real matching status and the rows the number of cases with assigned results, $\breve{\epsilon}$ is represented by $\frac{b}{a+b}$ and ϵ by $\frac{b}{b+d}$.

When the components γ_j, $j = 1,\ldots,n$, of the comparison pattern assume only values 1 and 0, then γ can assume only finitely many values and therefore the distribution of the weights is discrete (nevertheless, the tail of the discrete distribution can be approximated by a continuous probability distribution). The probability that the random variable W, representing weights, assumes a value W_i, that is associated with the matching status $Z = 0$ for non-matches and $Z = 1$ for matches is

$$P(W = W_i) = P(W = W_i|Z = 1)P(Z = 1) + P(W = W_i|Z = 0)P(Z = 0).$$

Table 1
Theoretical contingency table where the columns indicate the true matching status and the rows the classification results obtained.

Assigned	True status	
	Match	Non-match
Links	a	b
Non-links	c	d

If we set a unique threshold C and decide:

if $W \geqslant C$, then we have a link,
if $W < C$, then we have a non-link,

then the false match rate is given by

$$\breve{\epsilon} = P(Z = 0|W \geqslant C),$$
$$\epsilon = P(W \geqslant C|Z = 0).$$

In common applications where the weights of matches and non-matches overlap, it is necessary for the unsupervised procedure we use to find C_R with

$$C_R = \inf\left\{\widetilde{W}|Z = 1 \Leftarrow (\forall i : W_i \geqslant \widetilde{W})\right\},$$

or C_L with

$$C_L = \sup\left\{\widetilde{W}|Z = 1 \Rightarrow (\forall i : W_i \geqslant \widetilde{W})\right\}.$$

additionally to the classification threshold C in order to bound $P(Z = 1)$ and respectively $P(Z = 0)$. The relations between C_L, C_R and C are exemplified in Fig. 1, where the left bell curve represents the weight distribution of fictive non-matches and the right bell curve stands for the weight distribution of fictive matches.

Estimation of C_R is crucial because our procedure needs to know the area where almost certainly all weights correspond to matches. Under loose conditions, a *conservative* value $\breve{\epsilon}(C_R, C)$ for $\breve{\epsilon}$ is then given by

$$P(W < C_R|W \geqslant C) = \frac{P(C \leqslant W < C_R)}{P(W \geqslant C)}.$$

These preliminaries lead to the following:

Theorem 2.1. *For a C with $C < C_R$ and C_R defined as above it holds: when (i) the number of non-matches to the whole number of cases in the interval $[C, C_R]$ is lower than the number of cases in the interval $[C, C_R]$ to the number of all cases $<C_R$ and (ii) the probability of cases in the interval $[C, C_R]$ is $\leqslant$ the probability of overall non-matches, then a conservative value $\epsilon(C_R, C)$ for the false match ϵ rate is given by*

$$\frac{P(C \leqslant W < C_R)}{P(W < C_R)} \geqslant P(W \geqslant C|Z = 0).$$

Proof. Condition (i) amounts to ($A := \{W \geqslant C \wedge W < C_R\}$ and $B := \{W < C_R\}$)

$$\frac{P(A \wedge Z = 0)}{P(A)} \leqslant \frac{P(A)}{P(B)}.$$