Chapter 27
Methodology for Record Linkage:
A Medical Domain Case Study

Maria Vargas-Vera
Universidad Adolfo Ibanez, Chile

ABSTRACT

This paper presents a methodology for linking records from several sources each source might contain, missing information. This assumption of missing values has been made, without loss of generality, as the authors has observed that missing information is part of the nature of data in the health domain and also in other domains such as social sciences. The author's methodology is an attempt to deal with the linkage of records of the same patient in several databases. The first phase in her methodology is called homogenization. The homogenization of the databases/datasets is performed by applying a method which fills-in the missing values with the predicted values. The second phase of her methodology is called linking of records. It assesses the similarity between records and implements the linkage of the pairs of records with high level of similarity. Finally, the author presents an evaluation of our methodology. The evaluation of the homogenization phase was carried out using multinomial regression while, the evaluation of the aggregated similarities were performed using Jaccard, Jaro-Winkler and Monge-Elkan similarity metrics.

INTRODUCTION

The problem of matching entity names has been explored in different research communities such as Statistics, Databases and Artificial Intelligence. Each community has proposed a different solution to the problem. For example, the research in the statistics community has been concentrated on record linkage based in the seminal paper of Fellegi & Sunter (Fellegi & Sunter, 1969). Fellegi & Sunter proposed a solution to the entity matching problem as a classification problem, where the goal is to classify entity pairs as two classes matching or non-matching. Their pioneer work was the development of a mathematical model to provide a theoretical framework for a computer-oriented solution to the problem of

DOI: 10.4018/978-1-5225-1759-7.ch027

recognizing records in two files which represent identical persons, objects or events. Fellegi and Sunter proposals have been adopted by other researchers, although often with enhancements of the underlying statistical model (Jaro 1989; 1995; Winkler 1999; Larsen 1999; Belin & Rubin 1997).The Artificial Intelligence community has focussed their efforts in supervised learning, which has been used for learning the parameters of string-edit distance metrics (Ristad & Yianilos 1998; Bilenko & Mooney 2002). Also, the combination of the results of different distance functions has been explored by a good number of researchers. (Tejada et. al, 2001; Cohen & Richman 2002; Bilenko & Mooney 2002).

Some work of the database community on record matching has been based on knowledge approaches (Hernandez & Stolfo 1995; Galhardas et al. 2000; Raman & Hellerstein 2001). However, the use of string-edit distances as a general-purpose record matching scheme was proposed by Monge and Elkan (Monge & Elkan 1997; 1996),

The work presented in this paper is focussed in record linkage from databases. Record linkage represents pairs of entities not by pairs of strings, instead by pairs of vectors of "match features" such as names and categories for variables in survey databases. In our case of study, we work with medical databases as we want to detect records of the same patient from different practices in England and Scotland. Patient identifiers are generated differently in the practices of England and Scotland. Therefore, if the patient has moved between England and Scotland, the same patient could be registered with a different patient identifier in different databases. Another problem encountered was that patients and immunisations datasets have a large percentage of missing values. Then; our main contribution is to propose a methodology for linking records across several databases/datasets. Our methodology, firstly, it takes care of the missing values and secondly, it performs record linkage. The methodology is generic and it could be used in several domains like health or social sciences.

This paper is organized as follows: firstly, it presents a comprehensive state of the art in two streams missing values and matching solutions. Secondly, it describes our linking records methodology by presenting two algorithms. Thirdly, it gives a preliminary evaluation using as case of study an anonymised dataset from the GPRD database (i.e. the immunisation and patient datasets). Finally, it presents conclusions and further work.

RELATED WORK

This section shows the state of the art from two perspectives namely missing values and linking records. Firstly, we describe the state of the art to the problem of missing values in databases and a secondly, the state of the art in linking records.

State of Art in Missing Values

The solutions to the missing values ranges from solutions used already in statistics packages to Artificial Intelligence methods. Although, the problem of missing values is not new one as it has been around in the Database community for several decades. In our view, the proposed solutions appear to be somehow to be ad hoc to the type of data and these solutions could be handled and perhaps solutions could be improved by using other techniques from other fields like, for instance, Artificial Intelligence. Let us start our discussion by describing why the problem of missing data occurs and what we mean by imputation methods.