Information Fusion 26 (2015) 144–153

Contents lists available at ScienceDirect

Information Fusion

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

Supervised learning using a symmetric bilinear form for record linkage

Daniel Abril [a,c,*], Vicenç Torra [a,d], Guillermo Navarro-Arribas [b]

[a] IIIA, Institut d'Investigació en Intel·ligència Artificial, CSIC, Consejo Superior de Investigaciones Científicas, Campus UAB s/n, 08193 Bellaterra, Spain
[b] DEIC, Dep. Enginyeria de la Informació i de les Comunicacions, UAB, Universitat Autònoma de Barcelona, Campus UAB s/n, 08193 Bellaterra, Spain
[c] UAB, Universitat Autònoma de Barcelona, Campus UAB s/n, 08193 Bellaterra, Spain
[d] School of Informatics, University of Skövde, 54128 Skövde, Sweden

ARTICLE INFO

Article history:
Received 13 February 2014
Received in revised form 13 November 2014
Accepted 16 November 2014
Available online 26 November 2014

Keywords:
Record linkage
Data privacy
Disclosure risk
Bilinear form
Choquet integral

ABSTRACT

Record Linkage is used to link records of two different files corresponding to the same individuals. These algorithms are used for database integration. In data privacy, these algorithms are used to evaluate the disclosure risk of a protected data set by linking records that belong to the same individual. The degree of success when linking the original (unprotected data) with the protected data gives an estimation of the disclosure risk.

In this paper we propose a new parameterized aggregation operator and a supervised learning method for disclosure risk assessment. The parameterized operator is a symmetric bilinear form and the supervised learning method is formalized as an optimization problem. The target of the optimization problem is to find the values of the aggregation parameters that maximize the number of re-identification (or correct links). We evaluate and compare our proposal with other non-parametrized variations of record linkage, such as those using the Mahalanobis distance and the Euclidean distance (one of the most used approaches for this purpose). Additionally, we also compare it with other previously presented parameterized aggregation operators for record linkage such as the weighted mean and the Choquet integral. From these comparisons we show how the proposed aggregation operator is able to overcome or at least achieve similar results than the other parameterized operators. We also study which are the necessary optimization problem conditions to consider the described aggregation functions as metric functions.

1. Introduction

In this paper we introduce a new variation of the supervised learning approach for record linkage. This consists of a symmetric bilinear form and a supervised learning approach. This aggregation function relies on a symmetric weighting matrix and it can be considered a Mahalanobis-based distance when the weighting matrix is positive semi-definite. We also present a couple of supervised learning approaches adapted to this symmetric bilinear function. Both are different approximations to obtain a semi-definite weighting matrix. Additionally, we study the previously proposed aggregators, the weighted mean and the Choquet integral, and propose different problem formalizations to use them as distances. Finally, we present a comparison in terms of accuracy and time between all disclosure risk approaches. Those are the non-parameterized functions such as the Euclidean distance and the Mahalanobis distance, the literature parameterized aggregators with their corresponding proposed modifications and finally the proposed symmetric bilinear approach.

The outline of this paper is as follows. Section 2 briefly introduces the state-of-the-art and related work. In Section 3, we review some concepts needed in the rest of the paper. Then, in Section 4, we review some standard distances used in record linkage, two parametrized aggregation operators used in previous works and finally the proposed bilinear function. In Section 5, we describe the optimization problem. That is, the supervised learning approach for distance-based record linkage. The evaluation of the method in the context of data privacy is done in Section 6. Finally, Section 7 presents the conclusions of the paper.

2. Related work

Record linkage is the process of finding quickly and accurately two or more records distributed in different databases (or data sources in general) that make reference to the same entity or

* Corresponding author at: IIIA, Institut d'Investigació en Intel·ligència Artificial, CSIC, Consejo Superior de Investigaciones Científicas, Campus UAB s/n, 08193 Bellaterra, Spain.
E-mail addresses: dabril@iiia.csic.es (D. Abril), vtorra@his.se (V. Torra), guillermo.navarro@uab.cat (G. Navarro-Arribas).

http://dx.doi.org/10.1016/j.inffus.2014.11.004
1566-2535/

individual. This term was initially introduced in the public health area in [1], when files of individual patients were brought together using name, date-of-birth, and some other information. In the following years, this idea was deeply developed in [2–4], and nowadays it is a popular technique.

Record linkage is one of the existing preprocessing techniques used for data cleaning [5,6], and for improving data quality [7]. For example, record linkage can be used to scrutinize databases to improve dirty data removing duplicate records [8], correcting data entry mistakes, transcription errors, and solving problems due to lack of standards for recording data fields, etc.

In addition, it is also a popular technique employed to integrate different data sets that provide information regarding the same entities [9,10]. For instance, consider the linkability of a census dataset with health records. A step forward in this direction is the merging of very large databases. A clear example of this database integration is the initiative recently launched by the UK government to make all its data available as RDF (Resource Description Framework) with the purpose of enabling data to be linked together [11]. Similar initiatives also have been applied in the USA [12].

In the last years, record linkage techniques have also emerged in the data privacy context. Many government agencies and companies are collecting massive amounts of confidential data, such as medical records, income credit rating or even several types of test scores. In order to perform different kind of studies these datasets are analyzed by their owners or more commonly by third parties, creating a conflict between individual's right to privacy and the society's need to know. So, it is fundamental to provide privacy to databases against disclosure of confidential information. Privacy Preserving Data Mining [13] and Statistical Disclosure Control [14] research on methods and tools for ensuring the privacy of these data. The idea behind all of these developed methods is to modify statistical data, also called microdata, so that they can be published without giving away confidential information that can be linked to specific respondents and also achieve it with minimum loss of detail. Record linkage permits the evaluation of the disclosure risk of protected data [15,16]. In this context record linkage could be applied by an attacker, who tries to link his own information (original) with the protected one to obtain some new and unknown information. If the links can be established, the attacker can re-identify individuals from the protected data, and the protection is said to be broken. This is also applicable to model the worst-case scenario, where the attack attempts to link all records from the original data (the most comprehensive information an attacker can use) to the protected data. This gives an estimation of the chances that an attacker will be able to re-identify records in the protected data. The estimation is usually used as a disclosure risk measure of the protection method applied to protect the data. That is, the percentage of correctly linked records between the protected dataset and the original dataset is taken as a measure for the disclosure risk of the data. Thus, the higher the percentage of correctly linked records, the higher the risk of disclosure. This approach to measure the disclosure risk of protected data was initially introduced in [17] and adopted in much of the subsequent literature such as [18,16,15]. Note also that sampling is not taken into account in this approach, which means assuming that the intruder knows the sampled individuals in the data set. This is a common practice in the previously cited works.

In addition, Domingo-Ferrer and Torra [18] defined a general score to qualitatively rank protection methods. This score is the combination of disclosure risk techniques, to evaluate the risk of re-identification, and other techniques, which readily quantifies the information loss of a protected data set using analytical measures (either generic or data-use-specific).

We introduce a new optimization problem for distance-based record linkage and its application to data privacy. The performance of this approach depends critically on a given distance. The choice of a distance over an input space always has been a key issue in many machine learning algorithms. Due to the problems of the commonly used Euclidean distance, which assumes that each feature is equally important and independent from the others, distance metric learning has emerged as a research topic [19]. Although the origins of metric learning can be traced in earlier works, Xing et al. [20] were pioneers within this research area. Similar to our proposal, they parameterize the Euclidean distance using a symmetric positive semi-definite matrix $\Sigma \succeq 0$ to ensure the non-negativity of the metric. Their algorithm maximizes the sum of distances between dissimilar points, while keeping closer the set of distances between similar points. However, despite its simplicity, the method is not scalable, because it has to perform many eigenvalue decompositions. [21] proposed a method for learning distance metrics from relative comparisons such as a is closer to b than a is to c. This relies on a less general Mahalanobis distance learning in which for a given matrix a, only a diagonal matrix W is learnt such that $\Sigma = A'WA$. More recently, [22] proposed a framework for learning the weighted Euclidean subspace based on pairwise constrains and cluster validity, where the best clustering can be achieved. Beliakov et al. [23] considered the problem of metric learning in semi-supervised clustering defining the Choquet integral with respect to a fuzzy measure as an aggregation distance. The authors investigate necessary and sufficient conditions for the discrete Choquet integral to define a metric. Weinberger et al. [24] proposed a new classification algorithm, the Large Margin k-nearest neighbor (LMNN), in which a Mahalanobis distance is learned. This metric is trained with the goal that the k-nearest neighbors always belong to the same class while examples from different classes are separated by a large margin. However, Sun and Chen [25] show that LMNN cannot satisfactorily represent the local metrics which are respectively optimal in different regions of the input space and they propose a local distance metric learning method, a hierarchical distance metric learning for LMNN, which first groups data points in a hierarchical structure and then learns the distance metric for each hierarchy. The authors use a classification algorithm, Large Margin Nearest Neighbor (LMNN) to classify points in the hierarchical structure. The paper concludes that hierarchical distance works well when the number of classes is large but that it does not improve the results of LMNN when the number of classes is small.

One of the most important challenges associated with supervised metric learning approaches, specially in Mahalanobis-based distances is the satisfaction of the positive semi-definiteness. In the literature there are different approximations, from several matrix simplifications to modern *semi-definite programming* methods within the operations research field. Some Σ simplifications force it to be diagonal and so Σ is positive semi-definite if and only if all diagonal entries are non-negative. This simplification reduces the number of parameters drastically and makes the optimization problem a linear program. Higham [26] proposed an algorithm to find the nearest correlation matrix, symmetric positive semi-definite matrix with unit diagonal, to a given symmetric matrix by means of a projection from the symmetric matrices onto the correlation matrices, with respect to a weighted Frobenius form. Semidefinite programming (SDP), is a kind of convex programming which evolved from linear programming. While, a linear programming problem is defined as the problem of maximizing or minimizing a linear function subject to a set of linear constraints, semidefinite programming is defined as the problem of maximizing or minimizing a linear function subject to a set of linear constraints and a "semi-definite" constraint, a special form of