Control and Cybernetics

vol. **39** (2010) No. 4

Supervised learning for record linkage through weighted means and OWA operators*

by

Vicenç Torra, Guillermo Navarro-Arribas and Daniel Abril

IIIA, Institut d'Investigació en Intel·ligència Artificial
- CSIC, Consejo Superior de Investigaciones Científicas,
Campus UAB s/n, 08193 Bellaterra, Catalonia, Spain

e-mail: {vtorra, guille, dabril}@iiia.csic.es

Abstract: Record linkage is a technique used to link records from one database with records from another database, making reference to the same individuals. Although it is normally used in database integration, it is also frequently applied in the context of data privacy. Distance-based record linkage permits linking records by their closeness. In this paper we propose a supervised approach for linking records with numerical attributes. We provide two different approaches, one based on the weighted mean and another on the OWA operator. The parameterization in both cases is determined as an optimization problem. We evaluate our proposal and compare it with standard distance based record linkage, which does not rely on the parameterization of the distance functions. To that end we test the proposal in the context of data privacy by linking a data file with its corresponding protected version.

Keywords: data privacy, disclosure risk, record linkage, supervised learning, weighted mean, OWA operator.

1. Introduction

Record linkage techniques were developed with the purpose of finding entries from different sources (files, databases, ...), that refer to the same entity. An example is when we consider the join of two datasets that do not have a unique key in common but do refer to the same entities. It is a widely used technique nowadays. For example, consider the linkability of a census dataset with health records. Moreover, business registers are normally constructed from tax and employment databases providing links between names, addresses, and financial information (Colledge, 1995). Recently the UK government launched and initiative to make all government data available as RDF (*Resource Description Framework*) with the purpose of enabling data to be linked together

*Submitted: July 2010; Accepted: Novemver 2010.

(data.gov.uk, 2010), and similar initiatives were previously taken in the USA (data.gov, 2010).

Record linkage was first introduced in Halbert (1964) and further developed in Newcombe et al. (1959), Fellegi and Sunter (1969). It is nowadays a common technique employed by statistical agencies, research communities, and corporations (Batini and Scannapieco, 2006; Winkler, 2003). Record linkage is also implemented in data privacy techniques to determine the risk of a protection method (Torra et al., 2006; Winkler, 2004).

A popular family of record linkage methods, known as distance-based record linkage, attempts to link records by their closeness. A distance function is used to determine how *close* records from different databases are in order to establish their linkage. The selection of a concrete distance function and its parameterization is a key issue in these methods. It is normally difficult and tedious to test and find the most appropriate distance function and moreover, to determine the correct parameters such as weights.

In this paper we introduce a novel approach to parameterize distance based record linkage. We provide a supervised learning approach for OWA operators and weighted mean, which are common aggregators used to determine the distance between records. Ordered weighted averaging (OWA) operator was introduced by R. Yager (1988) and since its introduction, it has been widely used in the computational intelligence field (Yager and Kacprzyk, 1997; Torra, 2004; Bronselaer and De Tre, 2009). We show the suitability of our proposal, testing it in the field of data privacy. To our knowledge there is no similar work to provide supervised learning for parametrized record linkage in the literature.

The paper is organized as follows. Section 2 introduces two distance functions based on the weighted mean and the OWA operator that we will use as a record linkage. In Section 3 we describe our approach to the supervised learning of parameters for the distance function. The experiments and validation of our proposal are described in Section 4. Finally, Section 5 concludes the paper.

2. On record linkage approaches

Given two different data files, record linkage algorithms link each record of one file with another in the other file that is presumed to correspond to the same entity. For example, when record linkage is applied to a data file for customers and a data file for sellers, it is presumed that the algorithm will deliver a list of links establishing the sellers that are also customers.

Record linkage algorithms have been used for a long time for database integration. In addition, these algorithms have also been used in data protection to evaluate the risk of a data protection method.

Different algorithms exist for record linkage. Two main families can be distinguished: probabilistic record linkage and distance based record linkage. We detail these methods below.

In the description we assume that we have two files, A and B, represented as $A = (a_1, ..., a_N)$ and $B = (b_1, ..., b_N)$, respectively. Then, a record linkage algorithm will consider pairs of records (a_i, b_j), each record being described in terms of a set of variables. We will define $V_1^A, \ldots, V_n^A$ and $V_1^B, \ldots, V_n^B$ to denote the set of variables of file A and B, respectively. Also, we will express the values of each variable of a record i as $a_i = (V_1(a_i), \ldots, V_n(a_i))$ and $b_i = (V_1(b_i), \ldots, V_n(b_i))$.

In this paper, we consider that the two data files are described in terms of the same variables and that variables are aligned.

Probabilistic record linkage. This approach for record linkage assigns an index to each pair of records (a_i, b_j) with $a_i \in A$ and $b_j \in B$. Then, using two thresholds, pairs are classified as either a linked-pair, an unlinked pair, or a clerical pair.

The index is computed in terms of probabilities, and the thresholds are computed taking into account the conditional probabilities of false positives and false negatives.

Distance-based record linkage. This approach links each record in A to the *closest* record in B. The *closest* record is defined in terms of a distance.

Both approaches have been tested extensively in the area of data privacy to evaluate the disclosure risk of protected data.

In probabilistic record linkage, the parameters of the method are determined using the expectation-maximization algorithm. Determining the weights in this way has the advantage of only requiring two parameters, corresponding to the probabilities of false positives and false negatives, as mentioned above. All other probabilities and values are obtained automatically from the data and these two probabilities.

In distance-based record linkage, the determination of parameters is not so easy. The main point is the definition of a distance. Nevertheless, different distances can be defined, each yielding different results. Different distances have been considered and tested in the literature. We review them below. To make things easier, we will use the notation a and b when referring to a concrete record a_i or b_i from their respective files, A or B.

Euclidean (DBRL1): The Euclidean distance is used for attribute-standardized data. Accordingly, given the notation above, the distance between two records a and b is defined by:

$$d(a,b)^2 = \sum_{i=1}^{n} \left(\frac{V_i(a) - \overline{V_i^A}}{\sigma(V_i^A)} - \frac{V_i(b) - \overline{V_i^B}}{\sigma(V_i^B)} \right)^2$$

where $\sigma(V_i^A)$ is the standard deviation of V_i^A and $\overline{V_i^A}$ is the average of all the values that the variable V_i^A takes.

Euclidean (DBRL2): This is an alternative definition, also based on the Euclidean distance. In this case, the Euclidean distance is used for attribute-standardized data. Formally, the distance is defined as follows:

$$d(a,b)^2 = \sum_{i=1}^{n} \left(\frac{V_i(a) - V_i(b)}{\sigma(V_i^A - V_i^B)} \right)^2 .$$

Mahalanobis (DBRLM): The Mahalanobis distance is used and applied to the original data with no standardization:

$$d(a,b)^2 = (a-b)'[Var(V^A) + Var(V^B) - 2Cov(V^A, V^B)]^{-1}(a-b)$$

where $Var(V^A)$ is the variance of attributes V^A, $Var(V^B)$ is the variance of attributes V^B and $Cov(V^A, V^B)$ is the covariance between attributes V^A and V^B. In this equation a' corresponds to the transpose of vector a. The computation of $Cov(V^A, V^B)$ poses one difficulty: how records in A are lined up with records in B to compute the covariances. Two approaches have been considered in the literature.

DBRLM-COV In a worst case scenario, it would be possible to know the correct links (a, b). Therefore, the covariance of attributes might be computed with the correct alignment between records.

DBRLM-COV0 It is not possible to know a priori which are the correct matches between pairs of records. Therefore, any pair of records (a, b) are feasible. If any pair of records (a, b) are considered, the covariance is zero.

Kernel (KDBRL): A kernel-distance is considered. That is, instead of computing distances between records (a, b) in the original n dimensional space, records are compared in a higher dimensional space H. Thus, let $\Phi(x)$ be the mapping of x into the higher space. Then, the distance between records a and b in H is defined as follows:

$$\begin{aligned} d(a,b)^2 &= ||\Phi(a) - \Phi(b)||^2 = (\Phi(a) - \Phi(b))^2 = \\ &= \Phi(a)\cdot\Phi(a) - 2\Phi(a)\cdot\Phi(b) + \Phi(b)\cdot\Phi(b) = \\ &= K(a,a) - 2K(a,b) + K(b,b) \end{aligned}$$

where K is a kernel function (*i.e.*, $K(a,b) = \Phi(a)\cdot\Phi(b)$).

Experiments have been carried out for the kernel functions of the form $K(x,y) = (1 + x\cdot y)^d$ for $d > 1$. Note that with $d = 1$, the kernel record-linkage reduces to the distance-based record linkage with the Euclidean distance.

Taking all this into account, the distance between a and b is defined as:

$$d(a,b)^2 = K(a,a) - 2K(a,b) + K(b,b)$$

with a kernel function K.

In this paper we consider a variation of the Euclidean distance. On the one hand we consider the use of the OWA operator to aggregate partial distances, and on the other, we consider a weighted distance.

2.1. A parametric distance for record linkage

It is well known that the multiplication of the Euclidean distance by a constant will not change the results of any record linkage algorithm. Due to this, we can express the distance DBRL1 given above as a weighted mean of the distances for the attributes.

In a formal way, we redefine the DBRL1 as follows:

$$d(a,b)^2 = \sum_{i=1}^{n} \frac{1}{n} \left(\frac{V_i(a) - \overline{V_i^A}}{\sigma(V_i^A)} - \frac{V_i(b) - \overline{V_i^B}}{\sigma(V_i^B)} \right)^2$$

Now, defining

$$d_i(a,b)^2 = \left(\frac{V_i(a) - \overline{V_i^A}}{\sigma(V_i^A)} - \frac{V_i(b) - \overline{V_i^B}}{\sigma(V_i^B)} \right)^2$$

we can rewrite this expression as

$$d(a,b)^2 = AM(d_1(a,b)^2, \ldots, d_n(a,b)^2),$$

where AM is the arithmetic mean $AM(c_1, \ldots, c_n) = \sum_i c_i/n$. See e.g. Miyamoto and Suizu (2003), Chiang and Hao (2003) for details on kernel functions.

In general, any aggregation operator $\mathbb{C}$ might be used:

$$d(a,b)^2 = \mathbb{C}(d_1(a,b)^2, \ldots, d_n(a,b)^2).$$

From this definition, it is straightforward to consider a weighted version of the DBRL, and also a variation of it based on the OWA operators. Their definition is as follows.

DEFINITION 1 *Let $p = (p_1, \ldots, p_n)$ be a weighting vector (i.e., $p_i \geq 0$ and $\sum_i p_i = 1$), and given two records a and b. Then,*

- *the weighted distance is defined as:*

$$d^2WM_p(a,b) = WM(d_1(a,b)^2, \ldots, d_n(a,b)^2),$$

  where $WM = (c_1, \ldots, c_n) = \sum_i p_i \cdot c_i$.
- *the OWA distance is defined as:*

$$d^2OWA_p(a,b) = OWA(d_1(a,b)^2, \ldots, d_n(a,b)^2),$$

  where $OWA = (c_1, \ldots, c_n) = \sum_i p_i \cdot c_{\sigma(i)}$ with σ defining a permutation of $\{1, \ldots, n\}$ such that $c_{\sigma(i)} \geq c_{\sigma(i+1)}$ for all $i > 1$.

The interest of the first variation is that we do not need to assume that all the attributes are equally important in the re-identification. This would be the case if one of the attributes is a key-attribute. In this case, the corresponding weight would be assigned to one and all the others to zero. Such an approach would lead to 100% of re-identifications.

Moreover, as we will see later, this definition permits us to apply a supervised learning approach to determine the parameters of the method. In this way, we can tune the distance to have a better performance.

The interest of the second definition based on the OWA operator is that while the weighted mean permits to assign relevance to attributes, the OWA (because of the ordering σ) permits to assign relevance to either larger or smaller values. In this way, we might give importance to extreme values or central values. Note that extreme values might represent outliers, and in the case of re-identification algorithms, such values might be useful for achieving a better performance.

3. Supervised learning for record linkage

The goal of this paper is to determine the best weights for achieving the best possible performance in record linkage. To do so, we assume that a particular parameterized distance is used and consider the problem of finding the optimal weights for such parameterization.

In this section we describe a supervised learning approach for the determination of such weights. Then, in the next section we will describe some experiments to validate our approach. To make the experiments we consider an application in data privacy. It consists in using a data file, and a protected version of it. Then, the goal of an intruder would be to link own original records (say, A) with the records of the public but protected file (say, B).

In the rest of this section we will use the notation A and B, where A stands for the original file and B for the protected and public file. In the supervised approach we assume that we know the correct links, and this knowledge is used to determine the optimal weights. In the real world, this scenario would occur if an agency wanted to have a maximum bound of an estimation of disclosure risk before releasing a file.

For the sake of simplicity, we presume that each record of A, $A_i = (a_1, \ldots, a_N)$, is the original record of B, $B_i = (b_1, \ldots, b_N)$. That is, files are aligned. Then, if $V_k(a_i)$ represents the value of the kth variable of the ith record, we will consider the sets of values $d(V_k(a_i), V_k(b_j))$ for all pairs of records a_i and b_j.

Then, the optimal performance of record linkage using an aggregation operator $\mathbb{C}$ is achieved when the aggregation of the values $d(V_k(a_i), V_k(b_i))$ for all k is smaller than the aggregation of the values $d(V_k(a_i), V_k(b_j))$ for all $i \neq j$, i.e.

$$\mathbb{C}(d(V_1(a_i), V_1(b_i)), \ldots, d(V_n(a_i), V_n(b_i))) < \mathbb{C}(d(V_1(a_i), V_1(b_j)), \ldots, d(V_n(a_i), V_n(b_j))) \quad (1)$$

for all $i \neq j$.

Note that the proposed technique uses the same number of variables to link one file to another. This does not imply a constraint on the initial dataset, but on the number of variables one chooses to make the linkage. The fact that the variables are lined up is just an assumption we make to simplify the explanation, but in a very general case one could check all combinations of variables to see which yields a better linkage (with the computational cost that this might imply). It is important to note, however, that most record linkage techniques do rely on the use of the same variables, because they are already known or can be easily guessed. In our case scenario, since we rely on the specific application of record linkage for evaluating the disclosure risk in data privacy, we know the correct alignment of the variables and thus we can avoid the combinatorial problem.

Although we focus our work on numeric data, other types of data (categorical, sequential, ...) could be used as long as we are able to define (and compute) a distance function between the attribute values.

We considered two approaches for learning the weights. We describe them below.

3.1. First approach: minimizing the errors

The first approach consists in transforming Equation (1) into (2) using a new variable $Y_{(i,j)}$ to solve the inconsistencies in the data, and then it is expected that the variable will be as small as possible:

$$\mathbb{C}(d(V_1(a_i), V_1(b_j)), \ldots, d(V_n(a_i), V_n(b_j))) - \mathbb{C}(d(V_1(a_i), V_1(b_i)), \ldots, d(V_n(a_i), V_n(b_i))) + Y_{(i,j)} > 0 \quad (2)$$

for all $i \neq j$.

We formalize the problem as the minimization of the error $Y_{(i,j)}$, taking into account the constraints. However, besides the equation above, we require the weights to be positive and add to one as usual in the weighted mean. Note that these requirements about the weights are also mandatory if the resulting expression has to be a distance (positive and monotonic). In this way we obtain the following optimization problem:

$$Minimize : \sum_{i=1}^{N} \sum_{j=1}^{N} Y_{(i,j)}$$

Subject to :

$$\begin{aligned}
&\sum_{i=1}^{N}\sum_{j=1}^{N} P(d(V_1(a_i),V_1(b_j)),\ldots,d(V_n(a_i),V_n(b_j)))- \\
&\qquad P(d(V_1(a_i),V_1(b_i)),\ldots,d(V_n(a_i),V_n(b_i))) + Y_{(i,j)} > 0 \\
&Y_{(i,j)} > 0 \\
&\sum_{s=1}^{n} p_s = 1 \\
&p_s \geq 0.
\end{aligned} \tag{3}$$

Note that in this optimization problem, we minimize the error in the sense that the error of each constraint, $Y_{(i,j)}$, is made as small as possible.

Although this formalization seems natural and it is expected that the number of violated constraints be small, the approach does not work properly. Note that a single record j violating the constraint for record i, even with a small $Y_{(i,j)}$, implies that record i is incorrectly linked. So, in the case of all records i having just one record j violating a constraint would result in all records being incorrectly linked.

This caused that minimal solutions with respect to small $Y_{(i,j)}$ resulted in a large number of incorrect links. In fact, the number of correct links was less than the number of links when we just used the standard (non-weighted) record linkage with the Euclidean distance.

To solve this problem, we developed a second approach. It is explained in the next section.

3.2. Second approach: minimizing the number of incorrect links

As we have seen above, the first approach does not work as expected. For this reason, we considered another solution. In this one, we consider a variable K for each block. We define a block as the set of all the distances between one record of the original data and all the records of the protected data. Therefore, we have as many K as the number of rows of our original file. Besides, we need a constant C that multiplies K to avoid the inconsistencies and satisfy the constraint.

The rationale of this approach is as follows. The variable K indicates, for each block, if all the corresponding constraints are satisfied ($K = 0$) or not ($K = 1$). Then, we want to minimize the number of blocks non compliant with the constraints. Then, in this way, we can find the best weights that minimize the number of violations, or, in other words, we can find the weights that maximize the number of re-identifications between the original and protected data.

Using this variable, the constraint is defined as follows:

$$\mathbb{C}(d(V_1(a_i), V_1(b_j)), \ldots, d(V_n(a_i), V_n(b_j))) - \mathbb{C}(d(V_1(a_i), V_1(b_i)), \ldots, d(V_n(a_i), V_n(b_i))) + CK_i > 0 \quad (4)$$

for all $i \neq j$.

In this definition we have a constant C. This constant is used to express the *minimum distance* we require between the correct link and the other, incorrect links. It is used in the same way as variable $Y_{(i,j)}$ in Equation (2), but now it can be further parameterized. The bigger it is, the more the correct links are distinguished from the incorrect links.

Using the constraints of the form above, and taking into account what has been explained before, the problem is as follows:

$$\textit{Minimize} : \sum_{i=1}^{N} K_i$$

Subject to:

$$\sum_{i=1}^{N}\sum_{j=1}^{N} P(d(V_1(a_i), V_1(b_j)), \ldots, d(V_n(a_i), V_n(b_j))) - P(d(V_1(a_i), V_1(b_i)), \ldots, d(V_n(a_i), V_n(b_i))) + CK_i > 0$$

$$K_i \in \{0, 1\} \quad (5)$$

$$\sum_{s=1}^{n} p_s = 1$$

$$p_s \geq 0.$$

4. Experiments

For our experiments we have used the "Census" dataset, which contains 1080 records with 13 numerical attributes, and has been extensively used in other works (Domingo-Ferrer et al., 2006; Laszlo and Mukherjee, 2005; Domingo-Ferrer and Torra, 2005; Yancey et al., 2002; Domingo-Ferrer et al., 2001). We have tested the supervised learning approaches with the original dataset and three different protected versions of the same dataset. The protected datasets are generated by microaggregation of the original one (see Section 4.1 for a description of microaggregation). As outlined in Section 3, we consider a possible intruder whose goal is to link the records of the public and protected dataset with own original records.

We have tested both the weighted mean and the OWA operator based distances between records (as defined in Section 2.1). All tests have been performed using R and IBM ILOG CPLEX. R (R, 2010) (version 2.9.2) is a GNU project that provides a language and environment for statistical computing. On the other hand, the IBM ILOG CPLEX (IBM, 2010) is a mathematical

programming optimizer that enables analytical decision support for improving efficiency, reducing costs, and increasing profitability. Specifically we have used the simplex (Dantzig, 1963) optimizer algorithm that ILOG CPLEX version 12.1 provides.

The next section describes the microaggregation technique used to protect the datasets, and then we show the results of our experiments.

4.1. Microaggregation

Microaggregation is a statistical disclosure control technique, which provides privacy by means of clustering the data into small clusters and then replacing the original data by the centroids of the corresponding clusters.

Privacy is achieved because all clusters have at least a predefined number of elements, and therefore there are at least k records with the same value. Note that all the records in the cluster replace a value by the value in the centroid of the cluster. The constant k is a parameter of the method that controls the level of privacy. The larger the k, the more privacy we have in the protected data.

Microaggregation was originally (Defays and Nanopoulos, 1993) defined for numerical attributes, but later extended to other domains, e.g., to categorical data in Torra (2004) (see also Domingo-Ferrer and Torra, 2005), and in constrained domains in Torra (2008).

From the operational point of view, microaggregation is defined in terms of partition and aggregation:

- **Partition.** Records are partitioned into several clusters, each of them consisting of at least k records.
- **Aggregation.** For each of the clusters a representative (the centroid) is computed, and then original records are replaced by the representative of the cluster to which they belong to.

From a formal point of view, microaggregation can be defined as an optimization problem with some constraints. We give a formalization below using u_{ij} to describe the partition of the records in the sensitive data set X. That is, $u_{ij} = 1$ if record j is assigned to the ith cluster. Let v_i be the representative of the ith cluster, then a general formulation of microaggregation with g clusters and a given k is as follows:

$$\begin{aligned} &\textit{Minimize} \quad SSE = \sum_{i=1}^{g}\sum_{j=1}^{n} u_{ij}(d(x_j, v_i))^2 \\ &\textit{Subject to} \quad \sum_{i=1}^{g} u_{ij} = 1 \text{ for all } j = 1, \ldots, n \\ &\qquad\qquad 2k \geq \sum_{j=1}^{n} u_{ij} \geq k \text{ for all } i = 1, \ldots, g \\ &\qquad\qquad u_{ij} \in \{0, 1\}. \end{aligned}$$

For numerical data it is usual to require that $d(x, v)$ be the Euclidean distance. In the general case, when attributes $\mathbf{V} = (V_1, \ldots, V_s)$ are considered, x and v are vectors, and d becomes $d^2(x, v) = \sum_{V_i \in \mathbf{V}} (x_i - v_i)^2$. In addition, it is also common to require for numerical data that v_i be defined as the arithmetic mean of the records in the cluster, i.e., $v_i = \sum_{j=1}^{n} u_{ij} x_i / \sum_{j=1}^{n} u_{ij}$. As the solution of this problem is NP-Hard (Oganian and Domingo-Ferrer, 2000) when we consider more than one variable at a time (multivariate microaggregation), heuristic methods have been developed. A popular heuristic algorithm for multivariate microaggregation is MDAV (Domingo-Ferrer and Mateo-Sanz, 2002) (Maximum Distance to Average Vector). The implementation of MDAV for categorical data is given in Domingo-Ferrer and Torra (2005).

Note that when all variables are considered at once, microaggregation is a way to implement k-anonymity (Samarati, 2001; Sweeney, 2002).

4.2. Results

The results are obtained by linking the original file A with a protected file B_j. Each file contains 13 variables, $V_1(a_i), \ldots, V_{13}(a_i)$ for all records a_i in A, and $V_1(b_i), \ldots, V_{13}(b_i)$ for all records b_i in B_j. And the variables are aligned between A and B.

We consider three different protected files:

- B_1: microaggregation of A taking the variables in groups of three, obtaining four groups of three variables and another with just the last variable, and with $k = 4$.
- B_2: microaggregation of A in two groups of variables (one with 5 and another with 8), and $k = 4$.
- B_3: microaggregation of A in two groups of variables (one with 6 and another with 7), and $k = 20$.

We have tested our learning approach with three different training sets composed of 100, 200, and 300 records for each pair of files (A, B_j). The training sets are denoted T_{100}, T_{200}, and T_{300}, respectively.

As we have outlined before, our first approach, based on minimization of the errors (as described in Section 3.1), produces bad results. Just as an example consider the results from Table 1. There we can see the proportion of correctly linked records (1 means 100% of records correctly linked) with our approach using the weighted mean (d^2WM) as compared to the record linkage using the normalized arithmetic mean DBRL1.

As shown, the results are worse than using a standard record linkage with the distance DBRL1.

However, if we use our second approach, based on minimization of the number of incorrect links from Section 3.2, we can achieve better results. Table 2 compares our approach for the weighted mean d^2WM based record linkage with the DBRL1, and Table 3 compares the OWA operator based distance linkage d^2OWA with the DBRL1.

Table 1. Approach 1 with the weighted mean

	Training set	d^2WM	DBRL1
B_1	T_{100}	0.55	1
	T_{200}	0.56	1
	T_{300}	0.5766667	0.9933

Table 2. Approach 2 with the weighted mean

	Training set	d^2WM	DBRL1
B_1	T_{100}	1	1
	T_{200}	1	1
	T_{300}	1	0.9933
B_2	T_{100}	0.96	0.94
	T_{200}	0.965	0.95
	T_{300}	0.9467	0.93
B_3	T_{100}	0.76	0.73

Table 3. Approach 2 with the OWA operator

	Training set	d^2OWA	DBRL1
B_1	T_{100}	1	1
	T_{200}	1	1
	T_{300}	1	0.996
B_2	T_{100}	0.95	0.94
	T_{200}	0.96	0.95
B_3	T_{100}	0.75	0.73
	T_{200}	0.64	0.595

In both cases our approach using either the d^2WM with d^2OWA performs better than the standard record linkage using the Euclidean distance (DBRL1). Although the difference is not very big, it is clear that we always achieve a higher percentage of correct links. If we compare d^2WM and d^2OWA, we can see that d^2WM performs better than d^2OWA, with a very small difference.

The weights obtained by our methods can be a useful tool for identifying the relevant (or irrelevant) variables for linking records. Moreover, in the field of data privacy, they can constitute a very important indicator of how each variable is protected. Variables with a high weight are more sensitive, since they can be used for the re-identification of records. Ideally, a good protection

method should end up with a uniform distribution of the weights, otherwise if the weights are concentrated in few variables, record linkage can be performed from only these variables without losing much accuracy. It also means that the intruder only needs information about such variables for attacking efficiently the database.

Table 4 shows the weights for the case of the weighted mean. As an example consider the case of B_1, where some variables can be completely ruled out in the record linkage. These variables with weights close to 0 provide very low (or none) use for the linkage. Moreover, we can also see that B_1 has been microaggregated in groups of three variables. For each group there is always one variable with a larger weight, and the general weight of the record is distributed over each group. The weights for the OWA operator based record linkage are shown in Table 5, where we can see that the distribution of weights is more concentrated on specific positions. In general, it shows that the variables with lower values are the ones with the largest importance.

The general interpretation of the weights is, however, highly dependent on the scenario and the data itself (what are we linking).

Table 4. Weights obtained for d^2WM

B_1	T_{100}	0.12	0.00	0.09	0.00	0.00	0.62	0.05	0.00	0.00	0.00	0.12	0.00	0.00
	T_{200}	0.13	0.00	0.10	0.00	0.00	0.06	0.17	0.45	0.00	0.00	0.11	0.00	0.00
	T_{300}	0.00	0.00	0.01	0.00	0.00	0.00	0.31	0.05	0.02	0.00	0.23	0.00	0.37
B_2	T_{100}	0.20	0.08	0.14	0.00	0.26	0.00	0.09	0.04	0.00	0.12	0.00	0.00	0.06
	T_{200}	0.03	0.09	0.07	0.04	0.00	0.06	0.02	0.05	0.13	0.00	0.11	0.00	0.39
	T_{300}	0.02	0.08	0.05	0.04	0.00	0.05	0.05	0.20	0.10	0.00	0.05	0.04	0.32
B_3	T_{100}	0.06	0.03	0.04	0.11	0.03	0.02	0.02	0.02	0.45	0.00	0.08	0.00	0.13

Table 5. Weights obtained for d^2OWA

B_1	T_{100}	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.99	0.00
	T_{200}	0.00	0.00	0.00	0.062	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.93
B_2	T_{100}	0.01	0.12	0.00	0.00	0.00	0.00	0.00	0.31	0.00	0.00	0.00	0.56	0.00
	T_{200}	0.07	0.00	0.015	0.07	0.19	0.00	0.00	0.00	0.66	0.00	0.00	0.00	0.00
B_3	T_{100}	0.09	0.08	0.027	0.14	0.00	0.00	0.26	0.41	0.00	0.00	0.00	0.00	0.00
	T_{200}	0.03	0.04	0.088	0.06	0.01	0.024	0.00	0.06	0.00	0.00	0.21	0.32	0.15

5. Conclusions

In this paper we have introduced a parameterization of distance-based record linkage by means of extending the Euclidean distance, used in standard record linkage, with the weighted mean and the OWA operators, which allows for the parameterization of the distances. Moreover, we have presented a supervised

learning approach to determine the optimum weights for such proposed distances, in order to achieve a better performance in the linkage process between two datasets.

Our experiments, in the field of data privacy, show that the linkage is better when compared with the standard record linkage based on the Euclidean distance. Although the improvement is not very big, we think that in some particular circumstances this difference can be more important. For example, in cases where there is one variable that clearly provides more information for the linkage than the rest. Thus, e.g., in the context of data privacy, it is possible to use different protection degree for each variable. Variables, which are less protected, are more likely to provide more information, so weighing them appropriately can lead to better linkages.

As a result of our approach we also obtain the weight associated to each variable, which is an indicator of the importance (or lack of it) of the variable for the record linkage. This last feature has also important consequences in data privacy. We can identify the variables that provide more information for the linkages, in other words, the variables that have a greater disclosure risk. This fact can help, for example, statistical agencies in evaluating the protection level to be applied to each variable.

As a future work we will extend our proposal to record linkage of other type of data. To that end, we will investigate the use of different distances such as the Jaro-Winkler distance for categorical data or distances on time series for sequential data.

Acknowledgments

The comments from Jordi Castro and the help of Meritxell Vinyals are acknowledged.

Partial support by the Spanish MICINN (projects eAEGIS TSI2007-65406-C03-02, ARES - CONSOLIDER INGENIO 2010 CSD2007-00004, and N-KHRONOUS TIN2010-15764) is also acknowledged.

References

Batini, C. and Scannapieco, M. (2006) *Data Quality - Concepts, Methodologies and Techniques Series: Data-Centric Systems and Applications.* Springer, Secaucus, NJ.

Bronselaer, A. and De Tre, G. (2009) A Possibilistic Approach to String Comparison. *IEEE Transactions on Fuzzy Systems* **17** (1), 208-223.

Chiang, J.H. and Hao, P.Y. (2003) A new kernel-based fuzzy clustering approach: support vector clustering with cell growing. *IEEE Trans. on Fuzzy Systems* **11** (4), 518- 527.

Colledge, M. (1995) *Frames and Business Registers: An Overview. Business Survey Methods.* Wiley Series in Probability and Statistics.

DATA.GOV.UK (2010) UK Government.

DATA.GOV (2010) US Government.

DANTZIG, G.B. (1963) *Linear Programming and Extensions*. Princeton University Press and the RAND Corporation.

DEFAYS, D. and NANOPOULOS, P. (1993) Panels of enterprises and confidentiality: The small aggregates method. *Proc. of the 1992 Symposium on Design and Analysis of Longitudinal Surveys*, Statistics Canada, 195-204.

DOMINGO-FERRER, J., MATEO-SANZ, J.M. and TORRA, V. (2001) Comparing sdc methods for microdata on the basis of information loss and disclosure risk. In: *Preproceedings of ETK-NTTS* 2001 (vol. 2). Eurostat, 807-826.

DOMINGO-FERRER, J. and MATEO-SANZ, J.M. (2002) Practical data-oriented microaggregation for statistical disclosure control. *IEEE Trans. on Knowledge and Data Engineering* **14** (1), 189-201.

DOMINGO-FERRER, J. and TORRA, V. (2005) Ordinal, Continuous and Heterogeneous k-Anonymity Through Microaggregation. *Data Mining and Knowledge Discovery* **11** (2), 195-212.

DOMINGO-FERRER, J., TORRA, V., MATEO-SANZ, J.M. and SEBE, F. (2006) Empirical disclosure risk assessment of the ipso synthetic data generators. In: *Monographs in Official Statistics-Work Session On Statistical Data Confidentiality*. Eurostat, 227-238.

FELLEGI, I. and SUNTER, A. (1969) Fellegi, I., Sunter, A. (1969) A Theory for Record Linkage. *Journal of the American Statistical Association* **64** (328), 1183-1210.

HALBERT, D. (1946) Record Linkage. *American Journal of Public Health* **36** (12), 1412-1416.

IBM (2010) IBM ILOG CPLEX, High-performance mathematical programming engine. International Business Machines Corp. `http://www-01.ibm.com/software/integration/optimization/cplex/`

LASZLO, M. and MUKHERJEE, S. (2005) Minimum spanning tree partitioning algorithm for microaggregation. *IEEE Transactions on Knowledge and Data Engineering* **17** (7), 902-911.

MIYAMOTO, S. and SUIZU, D. (2003) Miyamoto, S., Suizu, D. (2003) Fuzzy c-means clustering using kernel functions in support vector machines. *Journal of Advanced Computational Intelligence and Intelligent Informatics* **7** (1), 25-30.

NEWCOMBE, H.B., KENNEDY, J.M., AXFORD, S.J. and JAMES, A.P. (1959) Automatic Linkage of Vital Records. *Science*, **130**, 954-959.

OGANIAN, A. and DOMINGO-FERRER, J. (2000) On the Complexity of Optimal Microaggregation for Statistical Disclosure Control. *Statistical J. United Nations Economic Commission for Europe* **18** (4), 345-354.

R (2010) R project, software environment for statistical computing and graphics. GNU project. `http://www.r-project.org/`

SAMARATI, P. (2001) Protecting Respondents' Identities in Microdata Release. *IEEE Transactions on Knowledge and Data Engineering* **13** (6), 1010-1027.

SWEENEY, L. (2002) k-Anonymity: a Model for Protecting Privacy. *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems* **10** (5), 557-570.

TORRA, V. (2004) Microaggregation for categorical variables: a median based approach. *Proc. Privacy in Statistical Databases (PSD 2004).* **LNCS 3050**. Springer, 162-174.

TORRA, V. (2004) OWA operators in data modeling and re-identification. *IEEE Trans. on Fuzzy Systems* **12** (5), 652-660.

TORRA, V. (2008) Constrained microaggregation: Adding constraints for data editing. *Transactions on Data Privacy* **1** (2), 86-104.

TORRA, V., ABOWD, J., and DOMINGO-FERRER, J. (2006) Using Mahalanobis distance-based record linkage for disclosure risk assessment. *Privacy in Statistical Databases (PSD 2006).* **LNCS 4302**. Springer, 233-242.

WINKLER, W.E. (2003) Data cleaning methods. *Proc. SIGKDD 2003.* ACM.

WINKLER, W.E. (2004) Re-identification methods for masked microdata. *Privacy in Statistical Databases (PSD 2004)*, **LNCS 3050**. Springer, 216-230.

YAGER, R. (1988) On ordered weighted averaging aggregation operators in multicriteria decision making. *IEEE Trans. Syst. Man Cybern.* **18** (1), 183-190.

YAGER, R. and KACPRZYK, J. (1997) *The Ordered Weighted Averaging Operators: Theory and Applications.* Springer.

YANCEY, W., WINKLER, W. and CREECY, R. (2002) Disclosure risk assessment in perturbative microdata protection. In: *Inference Control in Statistical Databases.* **LNCS 2316**. Springer, 135-152.