Heuristic Supervised Approach for Record Linkage

Javier Murillo[1], Daniel Abril[2,3], and Vicenç Torra[3]

[1] CIFASIS-CONICET, Universidad Nacional de Rosario, Argentina
[2] Universitat Autònoma de Barcelona (UAB), Barcelona, Spain
[3] Institut d'Investigació en Intel·ligència Artificial (IIIA), Consejo Superior de Investigaciones Científicas (CSIC), Barcelona, Spain

Abstract. Record linkage is a well known technique used to link records from one database to records from another database which make reference to the same individuals. Although it is usually used in database integration, it is also used in the data privacy field for the disclosure risk evaluation of protected datasets. In this paper we compare two different supervised algorithms which rely on distance-based record linkage techniques, specifically using the Choquet integral's fuzzy integral to compute the distance between records. The first approach uses a linear optimization problem which determines the optimal fuzzy measure for the linkage. While, the second approach is a kind of gradient algorithm with constraints for the fuzzy measures' identification. We show the advantages and drawbacks of both algorithms and also in which situations they will work better.

Keywords: Fuzzy measure, Choquet integral, Record linkage, Heuristic, Optimization.

1 Introduction

Record Linkage is the task of identifying records from different databases (or data sources in general) that refers to the same entity. This technique was firstly used for database integration in [14] and further developed in [24,16], and it is nowadays a popular technique used by statistical agencies, research communities, and corporations. The main applications of record linkage are database and datasets integration [1,10,29,30], data cleaning and quality control [5,31]. An example of the last application is the detection of duplicate records between several datasets [15]. However, more recently, in the context of data privacy [21], record linkage has emerged as an important technique to evaluate the disclosure risk of protected data [26,32]. By identifying the links between the protected dataset and the original one, we can evaluate the re-identification risk of the protected data [12].

Among record linkage approaches we have focused on those based on a distance function between records, that is, it links records by their closeness. There are previous works [28,4,3] that have considered the use of different

V. Torra et al. (Eds.): MDAI 2012, LNAI 7647, pp. 210–221, 2012.

parameterized distances together with a supervised learning approach. This supervised approach relies on an optimization problem which finds the best combination of distance's parameters in order to maximize the number of correct re-identifications. In this paper we compare two different supervised learning approaches relying on distance-based record linkage for data privacy which are based on the Choquet integral [9,27]. Both supervised approaches allow the use of a fuzzy measure to weight the attributes in the datasets. However, one is based on an adaptation of the gradient descent algorithm proposed by Grabisch in [17] and the other is based on a linear optimization problem [2]. That means that the first one will find the parameters of a local minimum in a reasonable time, while the other approach will find the optimal parameters that give the maximum number of re-identifications. The goal of this comparison is to analyse if the Grabisch heuristic method can achieve similar results than the optimization problem. These results are based on the number of correct linkages between the records from two databases, the computational time needed and whether weights are much fitted for the training set, producing overfitting.

The paper is organized as follows. Section 2 introduces the record linkage techniques in the data privacy context. In Section 3 we define both supervised approaches that are compared. Section 4 shows the results of the comparison taking into account all the factors mentioned. Finally, Section 5 concludes the paper and present the future work.

2 Record Linkage in Data Privacy

In data privacy, record linkage can be used to re-identify individuals from a protected dataset. It serves as an evaluation of the protection method used by modeling the possible attack to be performed on the protected dataset.

A dataset X can be viewed as a matrix with N rows (*records*) and V columns (*attributes*), where each row refers to a single individual. The attributes in a dataset can be classified, depending on their capability to identify unique individuals, as follows:

- *Identifiers*: attributes that can be used to identify the individual unambiguously. A typical example of identifier is the passport number.
- *Quasi-identifiers*: attributes that are not able to identify a single individual when they are used alone. However, when combining several quasi-identifier attributes, they can unequivocally identify an individual. Among the quasi-identifier attributes, we distinguish between confidential (X_c) and non-confidential (X_{nc}), depending on the kind of information that they provide. An example of non-confidential quasi-identifier attribute would be the zip code, while a confidential quasi-identifier might be the salary.

Before releasing the data, a protection method ρ is applied, leading to a protected dataset X'. This protection method will protect the non-confidential quasi-identifiers $X'_{nc} = \rho(X_nc)$. However, to ensure the privacy, identifiers are either remover or encrypted. The confidential quasi-identifiers are not modified because they are interesting for third parties. Therefore, the protected dataset, $X' = X'_{nc}||X_c$ can be published and made available. This scenario, which was first used in [12] to compare several protection methods, has also been adopted in other works like [26].

In data privacy, record linkage can be used to re-identify individuals between the protected dataset and a part or the whole original dataset as an evaluator of disclosure risk. There are two main approaches of record linkage. The **Probabilistic record linkage (PRL)** [20] and the **Distance-based record linkage (DBRL)** [25], which links each record a to the *closest* record in b, by means of a distance function. Both approaches have been used extensively in the area of data privacy to evaluate the disclosure risk of protected data. Nevertheless, the work in this paper is focused on distance-based record linkage, specifically using the Choquet integral as a distance. This is further described in the next section.

2.1 Record Linkage Based on the Choquet Integral

The main point of distance-based record linkage is the definition of a distance. In this paper we consider the parametrization of distance-based record linkage. This distance parameterization allows us to weight data attributes in order to express the importance of the variables in the linkage process.

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 Euclidean distance used for attribute-standardized data as a weighted mean of the distances for the attributes.

We will use $V_1^X, \ldots, V_n^X$ and $V_1^Y, \ldots, V_n^Y$ to denote the set of variables of file X and Y, respectively. Using this notation, we express the values of each variable of a record a in X as $a = (V_1^X(a), \ldots, V_n^X(a))$ and of a record b in Y as $b = (V_1^Y(b), \ldots, V_n^Y(b))$. $\overline{V_i^X}$ corresponds to the mean of the values of variable V_i^X.

In a formal way, we redefine the Euclidean distance as follows,

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

In addition, we will refer to each squared term of this distance as,

$$d_i^2(a,b) = \left(\frac{V_i^X(a) - \overline{V^X}_i(a)}{\sigma(V_i^X)} - \frac{V_i^Y(b) - \overline{V^Y}_i(b)}{\sigma(V_i^Y)} \right)^2$$

Using these expressions we can define the squared of the Euclidean distance as follows.

Definition 1. *Given two datasets X and Y the square of the Euclidean distance for attribute-standardized data is defined by:*

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

where AM is the arithmetic mean $AM(c_1, \ldots, c_n) = \sum_i c_i/n$.

In general, any aggregation operator $\mathbb{C}$ [27] might be used in the place of arithmetic mean. So, we can consider the following generic distance.

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

From this definition, it is straightforward to consider weighted versions of the Euclidean distance. In this case we have focused on fuzzy measures of the Choquet integral, these permit us to represent, in the computation of the distance, information like redundancy, complementariness, and interactions among the variables, which are not used in other parametrized distances. Therefore, tools that use fuzzy measures to represent background knowledge permit us to consider variables that, for example, are not independent.

Definition 2. *Let μ be an unconstrained fuzzy measure on the set of variables V, i.e. $\mu(\emptyset) = 0$, $\mu(V) = 1$, and $\mu(A) \leq \mu(B)$ when $A \subseteq B$ for $A \subseteq V$, and $B \subseteq V$. Then, the Choquet integral distance is defined as:*

$$d^2CI_\mu(a,b) = CI_\mu(d_1(a,b)^2,\ldots,d_n(a,b)^2) \tag{1}$$

where $CI_\mu(c_1,\ldots,c_n) = \sum_{i=1}^{n}(c_{s(i)} - c_{s(i-1)})\mu(A_{s(i)})$, given that $c_{s(i)}$ indicates a permutation of the indexes so that $0 \leq c_{s(1)} \leq \ldots \leq c_{s(i-1)}$, $c_{s(0)} = 0$, and $A_{s(i)} = \{c_{s(i)},\ldots,c_{s(n)}\}$.

The interest of this 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, e.g. an attribute where $V_i^X = V_i^Y$. 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 the interaction of coalitions of variables is taken into account by the fuzzy measure.

3 Supervised Learning Approaches for Record Linkage

In this section we describe the two learning processes used on this work. Firstly, we describe the optimization problem approach and then we introduce the heuristic approach, which is based on a gradient descent algorithm. Both approaches take as input a matrix formed by $n+1$ columns (n attributes + target value) and m rows (each row represent one example). The output of both algorithms are the coefficients of the fuzzy measure that maximizes the number of re-identifications.

3.1 Linear Optimization Problem

We start discussing the notation we have used.

Let X represent the original file, and Y the protected file, both with variables $V_1,\ldots,V_n$. Then, $V_k(a_i)$ represents the kth variable of the ith record. Using this notation, for all $a_i \in X$ we have $a_i = (V_1(a_i),\ldots,V_n(a_i))$ and for all $b_i \in Y$ we have $b_i = (V_1(b_i),\ldots,V_n(b_i))$. For the application of the record linkage algorithm we will consider the sets of values $d(V_k(a_i),V_k(b_j))$ for all pairs of records $a_i \in X$ and $b_j \in Y$.

For the sake of simplicity in the formalization of the process, we presume that each record a_i of X is the protected version of b_i of Y. That is, files are aligned.

Then, two records are correctly linked using an aggregation operator $\mathbb{C}$ when the aggregation of the values $d(V_k(a_i), V_k(b_i))$ for all k is smaller than $d(V_k(a_i), V_k(b_j))$ for all $i \neq j$. In optimal conditions this should be true for all records a_i.

We have formalized the learning process into an optimization problem with an objective function and some constraints. As the correct linkage will not always be satisfied because of the errors in the data cause by the protection method a relaxation is needed. The relaxation is based on the concept of blocks. We consider a block as the set of equations concerning record a_i. Therefore, 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. Then, we assign to each block a variable K_i. Therefore, we have as many K_i as the number of rows of our original file. Besides, we need for the formalization a constant C that multiplies K_i to overcome the inconsistencies and satisfies the constraint.

The rationale of this approach is as follows. The variable K_i indicates, for each block, if all the corresponding constraints are accomplished ($K_i = 0$) or not ($K_i = 1$). Then, we want to minimize the number of blocks non compliant with the constraints. 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 these variables K_i and the constant C, we have that all pairs $i \neq j$ should satisfy

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

As K_i is only 0 or 1, we use the constant C as the factor needed to really overcome the constraint. In fact, the constant C expresses the *minimum distance* we require between the correct link and the other incorrect links. The larger it is, the more correct links are distinguished from incorrect links.

Using these constraints and the Choquet integral aggregation operator $d^2CI_\mu(a, b)$, explained in Definition 2, the minimization problem is defined in a generic form as:

$$\begin{aligned}
&Minimize \sum_{i=1}^{N} K_i \\
&Subject\ to: \\
&\qquad \sum_{i=1}^{N}\sum_{j=1}^{N} CI_\mu(d(V_1(a_i), V_1(b_j)), \ldots, d(V_n(a_i), V_n(b_j)))- \\
&\qquad\qquad - CI_\mu(d(V_1(a_i), V_1(b_i)), \ldots, d(V_n(a_i), V_n(b_i))) + CK_i > 0 \\
&\qquad K_i \in \{0, 1\} \\
&\qquad \mu(\emptyset) = 0 \\
&\qquad \mu(V) = 1 \\
&\qquad \mu(A) \leq \mu(B) \text{ when } A \leq B
\end{aligned}$$

where N is the number of records, and n the number of variables. This problem is considered a mixed integer linear problem, because it is dealing with integer and real-valued variables in the objective function and in the constraints, respectively. See more details of the implementation and complexity in [2].

3.2 Gradient Descent Algorithm

Inspired in HLMS (Heuristic Least Mean Squares), a gradient descent algorithm, introduced by Grabisch in [17], we introduce an new record linkage process relying on it. HLMS takes as input a training dataset P like the following:

$$P = \begin{pmatrix} x_1^1 & \dots & x_i^1 & \dots & x_n^1 & T^1 \\ \vdots & & \ddots & & \vdots & \\ x_1^z & \dots & x_i^z & \dots & x_n^z & T^z \\ \vdots & & \ddots & & \vdots & \\ x_1^N & \dots & x_i^N & \dots & x_n^N & T^N \end{pmatrix}$$

where x_i^j is the value of sample j for attribute i, and T^j its target value. The algorithm finds the fuzzy measure μ that minimized the difference $\mathcal{C}_\mu(\{x_1^j, x_2^j, ..., x_n^j\}) - T^j \, \forall j$. The error made in the approximation can be calculated as:

$$E(\mu) = \sum_{j=1}^{N} (\mathcal{C}_\mu(\{x_1^j, x_2^j, ..., x_n^j\}) - T^j)^2$$

The formula represents simply the squared difference between the target T^j and the Choquet integral of sample j using μ, summed over all training examples. The direction of steepest descent along the error surface can be found by computing the derivative of E with respect to each component of the vector μ.

$$\bigtriangledown E(\mu) \equiv [\frac{\delta E}{\delta \mu_{(1)}}, \frac{\delta E}{\delta \mu_{(2)}}, \dots, \frac{\delta E}{\delta \mu_{(n)}}]$$

Since the gradient specifies the direction of steepest increase of E, the training rule for gradient descent is:

$$\mu_{(i)} \leftarrow \mu_{(i)} - \lambda \bigtriangledown E(\mu_{(i)})$$

Here λ is a positive constant called the learning rate, which determines the step size in the gradient descent search. The negative sign is present because we want to move the attributes of the aggregation operator in the direction that decreases E. The record linkage problem cannot be addressed directly with HLMS since the target value is unknown. To simplify notation let $V_k(a_i) = x_k^i$ and $V_k(b_i) = {x'}_k^i$. As in the previous approach we have divided the problem in blocks, so a block D_k is now defined as follows:

$$D_k = \begin{pmatrix} (x_1^k - {x'}_1^1)^2 & \dots & (x_i^k - {x'}_i^1)^2 & \dots & (x_n^k - {x'}_n^1)^2 \\ \vdots & & \ddots & & \vdots \\ (x_1^k - {x'}_1^z)^2 & \dots & (x_i^k - {x'}_i^z)^2 & \dots & (x_n^k - {x'}_n^z)^2 \\ \vdots & & \ddots & & \vdots \\ (x_1^k - {x'}_1^N)^2 & \dots & (x_i^k - {x'}_i^N)^2 & \dots & (x_n^k - {x'}_n^N)^2 \end{pmatrix}$$

The original dataset has N different blocks, one for each row in X. The algorithm must find the fuzzy measure μ that makes for block k that the value of

$$\mathcal{C}_\mu(\{(x_1^k - x'^z_1)^2, \ldots, (x_i^k - x'^z_i)^2, \ldots, (x_n^k - x'^z_n)^2\}) \qquad (2)$$

to be minimum when $k == z$.

The approach used for each block k is the following:
The fuzzy measure is initialized to the equilibrium state ($\mu_i = \frac{|i|}{n}$). The Choquet integral of each row in D_k is calculated. If the minimum of the Choquet integral is for row k, then proceed with the next block. If the minimum of the Choquet integral is not for row k, calculate the gradient direction that makes the value of the Choquet minimum increases and the gradient of the Choquet integral for row k decreases.

The algorithm for this approach is shown in Algorithm (1).

Algorithm 1. Description of the heuristic algorithm for record linkage

Let X be the original database and X' the protected one with N samples and n attributes each.

```
———————— Initialization ————————
for $i \in \mathcal{P}(X)$ do
    $\mu_i = \frac{|i|}{|X|}$
end for
——————— For each Block ———————
for $i \in [1..N]$ do
    ———— For each row in $X_i \in X$ ————
    $d_j \leftarrow (X_i - X'_j)^2 \; \forall j \in [1..N]$
    $s = \{j | \mathcal{C}(d_j) \leq \mathcal{C}(d_i) \; \forall \, j \in [1..N]\}$
    ——————— Update step ———————
    for all j $\in$ s do
        Update the fuzzy measure, so that the difference $\mathcal{C}(d_i) - \mathcal{C}(d_j)$ decreases
    end for
    ———— Monotonicity check ————
    Check monotonicity
end for
return $\mu$
```

The algorithm does not guarantee the convergence to a global minimum. Some other minor modifications were done to the algorithm with no significant improvement.

4 Results

In this section we have compared both approaches; the heuristic algorithm for record linkage ($HRLA$) and the Choquet integral optimization algorithm (d^2CI) over different protected files. This comparison is divided in two parts to tackle the optimization problem. In the first part we have focused on the scores' comparison, in terms of the number of correct linkages and also the required times taken from both approaches. In the second part we have focused on the overfitting problem, testing both approaches with a small set for training and a big set for test.

To do our experiments we have applied different protection methods to an amount of records, randomly selected, from the original file. This file was selected from the Census dataset[8] of the European CASC project [7], which contains 1080 records and 13 variables, and has been extensively used in other works, such as [4,13,22].

To solve the Choquet optimization problem , we used the simplex optimizer algorithm from the IBM ILOG CPLEX tool [19], (version 12.1). The problem is first expressed into the MPS (Mathematical Programming System) format by means of the R statistical software[1] , and then, it was processed with the optimization solver. The $HRLA$ was completely programmed in the R statistical software.

4.1 Precision Comparison

The first part of the comparison is made with two different protected files using $Microaggregation$ [11], a well-known microdata protection method, which broadly speaking, provides privacy by means of clustering the data into small clusters of size k, and then replacing the original data by the centroid of their corresponding clusters. The parameter k determines the protection level: the greater the k, the greater the protection and at the same time the greater the information loss.

We have considered two protected files of 400 records, which were protected with two different protection levels.

- $M4-28$: 4 variables, first 2 variables with $k=2$, and last 2 with $k=8$.
- $M5-38$: 5 variables, first 3 variables with $k=3$, and last 2 with $k=8$.

Note that, we have applied two different protection degrees to different attributes of the same file. The values used range from 2 to 8. This is especially interesting when variables have different sensitivity.

Table 1 shows the percentage of re-identifications and the consumed time in the training step of both presented approaches (d^2CI and $HRLA$). It is clear that both supervised approaches have obtained better results than the arithmetic mean (d^2AM). However, if we make a comparison between them, we can see that the $HRLA$ has an error between 2% and 5% respect to the optimum value, achieved by d^2CI. Recall that the $HRLA$ is initialized with an equilibrium fuzzy measure. Therefore, in the first iteration the $HRLA$ is at least as good as the Euclidean distance (d^2AM). It is worth mention that, since $HRLA$ is an algorithm that finds the local minimum of a function, the results shown in that table correspond to the average of ten runs with the same configuration.

We have also compared training computational times of all the approaches. Table 1 shows that in almost all the situations, the time required by the $HRLA$ to achieve similar results than d^2CI is much lower than the optimization algorithm. However, we have to remember that the time's factor of the $HRLA$ approach could be different depending on the learning rate and the number of iterations which are parameters of the algorithm set up in its initialization.

[1] http://www.r-project.org/

Table 1. Percentage of re-identifications and computational time

	Dataset	d^2AM	d^2CI	$HRLA$
% Re-identifications	*M4-28*	68.50	93.75	91.75
	M5-38	39.75	91.25	86.75
Computational Time	*M4-28*	-	30 minutes	20 minutes
	M5-38	-	4 days	20 minutes

4.2 Overfitting

In the last part of this algorithm comparison, we have evaluated the scenario where an attacker would have a small amount of samples of the original database with its correct linkage between those samples and the samples in the public protected database. Therefore, the attacker is able to find the set of weights that achieve more number of linkages between the known samples (training set) and then, with those obtained weights, he/she is able to try the re-identification between the rest of records (test set) of two datasets in order to discover new confidential information.

In this experiment we have anonymized the whole original file (Census) by means of four different protection methods with several degrees of protection. The selected protection methods are briefly explained below; $RankSwapping$ [23], where the values of a variable V_i are ranked in ascending order; then each ranked value is swapped with another ranked value randomly chosen within a restricted range. $AdditiveNoise$ [6] which consists of adding Gaussian noise to the original data to get the masked data. If the standard deviation of the original variable is σ, noise is generated using a $N(0, \rho\sigma)$ distribution. Finally, we have also considered the $JPEG$ [18], The idea is to regard a numerical microdata file as an image (with records being rows, variables being columns, and values being pixels) and then use this lossy compression algorithm, and then the compressed image is interpreted as a masked microdata file.

We suppose that the attacker has a prior knowledge, so, a linkage of 200 records between the original and the protected files (labeled training set) could be made. Then, using a supervised approach the set of Choquet integral coefficients are learned to re-identify the rest of records (880 records), i.e., the test set.

Table 2 shows the results of the training and the test steps. Note the lack of training results in the Euclidean distance approach, since it does not require a learning step. Besides, the hyphen indicates that the corresponding computation was not finished, because it needed more than 300 hours. In the training process evaluation we have considered the time need to learn the parameters and the precentage of re-identifications. The minimum consumed times are in bold, most achieved by the $HRLA$, so the optimization problem has needed more than 14 minutes. However, it has achieved the best performance in the training set (9% of improvement at most). With respect to the test step the heuristic algorithm for record linkage has achieved an improment of at most 6% compared with the optimization problem, this is a clear indicator of overfitting. Nevertheless, $HRLA$ has achiveved similar re-identification results than d^2AM. This is due

Table 2. Percentage of re-identifications and time consumed

	d^2AM		d^2CI			$HRLA$		
Dataset	Train	Test	Time	Train	Test	Time	Train	Test
Rankswap-20	14.00	2.61	–	–	–	14*min*	14.50	2.73
Rankswap-15	24.50	9.89	–	–	–	14*min*	26.00	8.98
Rankswap-12	43.50	17.50	–	–	–	14*min*	44.50	17.73
Rankswap-5	94.00	78.86	**4min**	97.5	77.61	14*min*	94.50	79.20
Rankswap-4	95.50	85.23	**9sec**	100.00	80.91	14*min*	97.00	85.11
Mic3-9	83.00	60.23	18*min*	89.50	57.16	**14min**	83.00	60.11
Mic3-5	91.00	77.39	**1.5min**	96.50	74.66	14*min*	93.00	76.93
Mic3-8	82.50	65.00	**5min**	91.00	62.95	14*min*	83.00	65.11
Mic4-4	84.50	61.48	**2min**	88.00	58.52	14*min*	84.50	61.70
Mic4-8	70.00	37.27	**13min**	75.50	35.68	14*min*	70.00	37.16
Mic4-5	80.00	52.50	37*min*	85.00	50.45	**14min**	80.00	52.50
Micz-3	0.00	0.23	**3sec**	0.00	0.11	14*min*	0.00	0.23
MicMull-3	54.50	22.50	2.5*days*	64.50	21.70	**14min**	58.00	23.52
Noise-16	87.00	70.11	1*days*	92.50	67.50	**14min**	87.00	70.11
Noise-12	92.00	86.59	22*min*	97.00	80.57	**14min**	93.00	86.82
Noise-1	100.00	100.00	**4sec**	100.00	99.66	14*min*	100.00	100.00
Jpeg-80	84.50	76.93	2.5*hours*	94.50	73.30	**15min**	85.50	76.48
Jpeg-65	58.50	40.00	15*days*	67.00	36.59	**15min**	58.50	40.00

to the fact that $HRLA$ is initialized with the equilibrated weights and they were slightly changed by this algorithm. Although all the protection processes are different, they mainly rely on the addition of noise to each variable, so a distance function as the Euclidean distance can clearly re-identify some of the records, obviously always depending on the amount of noise added, that is the protection degreed applied for the method.

5 Conclusions

In this paper we have introduced an adaptation of the gradient descent algorithm proposed by Grabisch in order to use it as a disclosure risk evaluation in the data privacy context. The use of this heuristic algorithm was motivated on the high computational times required to find the fuzzy measures by another previously presented non-heuristic method which relies on a linear optimization problem. We have evaluated and compared both of them in two different ways.

The first part of the evaluation is focused on a scenario where original and protected files are available, and an evaluation of the protected dataset is performed. This is the worst scenario, where all the information is known, so, a good estimation of the disclosure risk is obtained. This comparison shows that although the linear optimization process (d^2CI) guarantees the convergence to the optimal solution, it requires a lot of time, from seconds to hours or even days depending on the level of protection applied, while the time required by the $HRLA$ remains low and stable. Regarding to the results in this comparison we have achieved an error rate from 2% to 5% higher for $HRLA$.

The second part of this work cope with the overfitting problem. In this scenario the results show that when the training dataset is small, the linear optimization problem get better results for training data than $HRLA$, while for test data the results are worst. This suggest that there is an overfitting of the data.

To sum up, if we have an exhaustive disclosure risk evaluation and we have enough computational resources and time it is recommended to use the optimization approach so we will have the optimal weights to analyse the risk and we can also analyse more efficiently if there is some attribute or a set of them that disclose more information than the others. Otherwise, if the resources needed are not available we can use the heuristic approach, that provide a good approximation to the optimal solution.

In view of the results, some additional tasks remains as future work. Firstly, to program the *HRLA* approach in $C++$ and be able to make a fairer comparison between two compiled approaches. Lastly, to use the fuzzy measures returned by *HRLA* as a first solution of the linear optimization process, to see if the amount of time required to solve the hard datasets reduces.

Acknowledgements. Partial supports by the Spanish MICINN (projects TSI2007-65406-C03-02, ARES-CONSOLIDER INGENIO 2010 CSD2007-00004, TIN2010-45764 and TIN2011-27076-C03-03) and by the EC (FP7/2007-2013) Data without Boundaries (grant agreement number 262608) are acknowledged. Some of the results described in this paper have been obtained using the Centro de Supercomputación de Galicia (CESGA). This partial support is gratefully acknowledged.

References

1. Statistics Canada. Record linkage at statistics canada (2010), `http://www.statcan.gc.ca/record-enregistrement/index-eng.htm`
2. Abril, D., Navarro-Arribas, G., Torra, V.: Choquet integral for record linkage. Annals of Operations Research, 1–14, 10.1007/s10479-011-0989-x
3. Abril, D., Navarro-Arribas, G., Torra, V.: Supervised learning using mahalanobis distance for record linkage. In: Bernard De Baets, R.M., Troiano, L. (eds.) Proc. of 6th International Summer School on Aggregation Operators - AGOP 2011, pp. 223–228 (2011), `Lulu.com`
4. Abril, D., Navarro-Arribas, G., Torra, V.: Improving record linkage with supervised learning for disclosure risk assessment. Information Fusion 13(4), 274–284 (2012)
5. Batini, C., Scannapieco, M.: Data Quality: Concepts, Methodologies and Techniques (Data-Centric Systems and Applications). Springer-Verlag New York, Inc. (2006)
6. Brand, R.: Microdata Protection through Noise Addition. In: Domingo-Ferrer, J. (ed.) Inference Control in Statistical Databases. LNCS, vol. 2316, pp. 97–116. Springer, Heidelberg (2002)
7. Brand, R., Domingo-Ferrer, J., Mateo-Sanz, J.: Reference datasets to test and compare sdc methods for protection of numerical microdata. Technical report, European Project IST-2000-25069 CASC (2002)
8. U.S. Census Bureau. Data extraction system
9. Choquet, G.: Theory of capacities. Annales de l'institut Fourier 5, 131–295 (1953)
10. Colledge, M.: Frames and business registers: An overview. Business Survey Methods. Wiley Series in Probability and Statistics (1995)

11. Defays, D., Nanopoulos, P.: Panels of enterprises and confidentiality: The small aggregates method. In: Proc. of the 1992 Symposium on Design and Analysis of Longitudinal Surveys, Statistics, Canada, pp. 195–204 (1993)
12. Domingo-Ferrer, J., Torra, V.: A quantitative comparison of disclosure control methods for microdata. In: Doyle, P., Lane, J., Theeuwes, J., Zayatz, L. (eds.) Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies, pp. 111–133. Elsevier (2001)
13. Domingo-Ferrer, J., Torra, V.: Ordinal, continous and heterogeneous anonymity through microaggregation. Data Mining and Knowledge Discovery 11(2), 195–212 (2005)
14. Dunn, H.: Record linkage. American Journal of Public Health 36(12), 1412–1416 (1946)
15. Elmagarmid, A., Ipeirotis, P.G., Verykios, V.S.: Duplicate record detection: A survey. IEEE Transactions on Knowledge and Data Engineering 19(1), 1–16 (2007)
16. Fellegi, I., Sunter, A.: A theory for record linkage. Journal of the American Statistical Association 64(328), 1183–1210 (1969)
17. Grabisch, M.: A new algorithm for identifying fuzzy measures and its application to pattern recognition. In: Fourth IEEE International Conference on Fuzzy Systems, Yokohama, Japan, pp. 145–150 (1995)
18. J. P. E. Group. Standard IS 10918-1 (ITU-T T.81) (2001), `http://www.jpeg.org`
19. I. IBM ILOG CPLEX. High-performance mathematical programming engine. International Business Machines Corp. (2010)
20. Jaro, M.A.: Advances in record-linkage methodology as applied to matching the 1985 census of tampa, florida. Journal of the American Statistical Association 84(406), 414–420 (1989)
21. Lane, J., Heus, P., Mulcahy, T.: Data access in a cyber world: Making use of cyberinfrastructure. Transactions on Data Privacy 1(1), 2–16 (2008)
22. Laszlo, M., Mukherjee, S.: Minimum spanning tree partitioning algorithm for microaggregation. IEEE Trans. on Knowl. and Data Eng. 17(7), 902–911 (2005)
23. Moore, R.: Controlled data swapping techniques for masking public use microdata sets. U.S. Bureau of the Census (1996) (unpublished manuscript)
24. Newcombe, H.B., Kennedy, J.M., Axford, S.J., James, A.P.: Automatic linkage of vital records. Science 130, 954–959 (1959)
25. Pagliuca, D., Seri, G.: Some results of individual ranking method on the system of enterprise acounts annual survey. Esprit SDC Project, Delivrable MI-3/D2 (1999)
26. Torra, V., Abowd, J.M., Domingo-Ferrer, J.: Using Mahalanobis Distance-Based Record Linkage for Disclosure Risk Assessment. In: Domingo-Ferrer, J., Franconi, L. (eds.) PSD 2006. LNCS, vol. 4302, pp. 233–242. Springer, Heidelberg (2006)
27. Torra, V., Narukawa, Y.: Modeling Decisions: Information Fusion and Aggregation Operators. Springer (2007)
28. Torra, V., Navarro-Arribas, G., Abril, D.: Supervised learning for record linkage through weighted means and owa operators. Control and Cybernetics 39(4), 1011–1026 (2010)
29. USA Government, `http://data.gov` (2010)
30. UK Government, `http://data.gov.uk` (2010)
31. Winkler, W.E.: Data cleaning methods. In: Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2003)
32. Winkler, W.E.: Re-identification Methods for Masked Microdata. In: Domingo-Ferrer, J., Torra, V. (eds.) PSD 2004. LNCS, vol. 3050, pp. 216–230. Springer, Heidelberg (2004)