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 Intellència 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

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-parametrized 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 parameterized 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
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 $\Sigma = A^T W A$. 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 $\Sigma$ simplifications force it to be diagonal and so $\Sigma$ 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. Semi-definite 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...
3. Preliminaries

In this section we review some ideas and definitions that are needed to follow the rest of the paper. We explain the notation we use as well as how the record linkage is applied in the data privacy area.

3.1. Record linkage

As stated in the introduction, record linkage is a re-identification method that links records in one file with records in another file that correspond to the same individuals. There are two extensively used approaches of record linkage.

- **Distance based record linkage (DBRL).** This approach [30] links each record a of a file X to the closest record b in a file X′. The closest record is defined in terms of a distance function.

- **Probabilistic record linkage (PRL).** In this case, the matching algorithm uses the linear sum assignment model to choose which pairs of records must be matched. In order to compute this model, the EM (Expectation–Maximization) algorithm [31,32] is normally used. Informally, we consider records a and b of files X and X′, respectively, represented in terms of a set of variables V₁,..., Vₙ. That is, a = (V₁(a),..., Vₙ(a)) and b = (V₁(b),..., Vₙ(b)). Then, we define a coincidence vector γ(a,b) = (γ₁(a,b),..., γₙ(a,b)), where γᵢ(a,b) is defined as 1 if Vᵢ(a) = Vᵢ(b) and as 0 if Vᵢ(a) ≠ Vᵢ(b). According to some criterion defined over these coincidence vectors, pairs are classified as linked pairs (LP) or non-linked pairs (NP). This concrete method was introduced in [33], although probabilistic record linkage was first presented in [4].

To avoid disclosure, when we want to publish a data set X, where X = Xid||Xnc||Xc, a protection method should be applied to X, leading to a protected data set X′ = ρ(X). This protection process is usually as follows; first, a protection method ρ is used to protect the non-confidential quasi-identifiers, i.e., X′nc = ρ(X′nc). Second, to ensure the privacy of the individuals the identifiers are either removed or encrypted and, their confidential quasi-identifiers are not modified because they are the data of interest for third parties. Therefore, the protected data set consists of X′ = Xid||Xc.

In order to evaluate the disclosure risk of releasing ρ(X), we model the behavior of an attacker applying record linkage to the pair (X, ρ(X)). The more records are re-identified, the larger the disclosure risk. This scenario, which was first used in [18] to compare several protection methods, has been adopted in other works like [15].

4. Distance-based record linkage

The main point in distance-based record linkage is the definition of the distance function used to match the records. Different distances can be found in the literature, each obtaining different results. In this section we start reviewing two of the most frequently used distances on record linkage, the Euclidean and the Mahalanobis distances. Then, we introduce the parametrized distances, which we will use together with a supervised learning process to obtain the parameters yielding the highest number of re-identifications. Examples of these parametrized distances are those based on the weighted mean and the Choquet integral. In this vein we introduce a parametrized symmetric bilinear function in which the parameters are represented by a weighting matrix.

We adopt the definition of distance function and metric from [34], where a distance function is defined in a less restrictive way than a metric.

**Definition 1.** Let X be a set. A function d : X × X → R is called a distance (or dissimilarity) on X if, for all a, b, c ∈ X, there holds:

1. d(a,b) ≥ 0 (non-negativity)
2. d(a,a) = 0 (reflexivity)
3. d(a,b) = d(b,a) (symmetry)
4. d(a,c) ≤ d(a,b) + d(b,c) (triangle inequality)
Definition 2. Let $X$ be a set. A function $d : X \times X \rightarrow \mathbb{R}$ is called a metric on $X$ if, for all $a, b, c \in X$, there holds:

1. $d(a, b) \geq 0$ (non-negativity)
2. $d(a, b) = 0$ iff $a = b$ (identity of indiscernibles)
3. $d(a, b) = d(b, a)$ (symmetry)
4. $d(a, b) \leq d(a, c) + d(c, b)$ (triangle inequality)

Note that other works may consider the terms metric and distance function as the same concept (Definition 2). Then, those works are using terms such as pseudo-metric or pre-metric in order to denote Definition 1.

Now that we have reviewed the properties required by a metric and a distance function, we are going to survey some metrics used in record linkage.

We will use $V_1, \ldots, V_n$ and $V'_1, \ldots, V'_n$ 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 $a = (V_1(a), \ldots, V_n(a))$ and of a record $b$ in $Y$ as $b = (V'_1(b), \ldots, V'_n(b))$. $ar{V}_i$ corresponds to the mean of the values of variable $V_i$.

Definition 3. Given two datasets $X$ and $Y$, the square of the Euclidean distance between two records $a \in X$ and $b \in Y$ for variable-standardized data is defined by:

$$d^2ED(a, b) = \sum_{i=1}^{n} \left( \frac{V_i^2(a) - \bar{V}_i}{\sigma(V_i)} - \frac{V'_i(b) - \bar{V}_i}{\sigma(V'_i)} \right)^2$$

where $\sigma(V_i)$ and $\bar{V}_i$ are the standard deviation and the mean of all the values of variable $V_i$ in the dataset $X$, respectively.

It is well known that in the Euclidean distance all the variables contribute equally to the computation of the distance. Because of that all points with the same Euclidean distance to the origin define a sphere. There are other metrics whose property does not hold. For example, the Mahalanobis distance [35] allows us to calculate distances taking into account a different variable contribution by means of weighting these variables. These weights are obtained from the covariances between data variables. Because of this rescaling, points at the same Mahalanobis distance define an ellipse around the mean of the set of variables.

Definition 4. Given two datasets $X$ and $Y$, the square of the Mahalanobis distance between two records $a \in X$ and $b \in Y$ is defined by:

$$d^2MD(a, b) = (a - b) \Sigma^{-1} (a - b)$$

where $(a - b)$ is the transpose of $(a - b)$ and $\Sigma$ is the covariance matrix, computed by $[\text{Var}(V^2) + \text{Var}(V') - 2 \text{Cov}(V^2, V')]$, where $\text{Var}(V^2)$ is the variance of variables $V^2$, $\text{Var}(V')$ is the variance of variables $V'$ and $\text{Cov}(V^2, V')$ is the covariance between variables $V^2$ and $V'$.

Any covariance matrix is a symmetric positive semi-definite\(^1\) matrix, so $d^2MD$ satisfies the first metric requirement (Definition 2), because all covariance matrices are always positive semi-definite and it is known that the inverse of a positive definite matrix is always positive definite too. That is, for any vector $v$ and a $n \times n$ positive definite matrix $\Sigma$ the following inequality $v' \Sigma^{-1} v > 0$ is always satisfied.

Notice that from this definition it follows that when the covariance matrix is the identity matrix, the Mahalanobis distance is reduced to the Euclidean distance.

Let us now focus on the parametrized distances. We first introduce a generic definition of a distance based on aggregation operators [36] and then consider two particularizations of this generic distance.

The generic distance is based on the fact that the Euclidean distance has the same results when it is multiplied by a constant. Then, we express the Euclidean distance ($d^2ED(a, b)$) given in Definition 3 as a weighted mean of the distances for the variables. For the sake of simplicity we consider the square of the distances although it is clear that it is not a distance itself, because it does not satisfy the triangle inequality.

To make it simple, we first define the difference between two variables from two records taking into account the normalization of the data. That is,

$$\text{diff}_i(a, b) = \frac{V_i^2(a) - \bar{V}_i}{\sigma(V_i)} - \frac{V'_i(b) - \bar{V}_i}{\sigma(V'_i)}$$

In a formal way, we redefine $d^2ED(a, b)$ as follows:

$$d^2(a, b) = \sum_{i=1}^{n} \frac{1}{n} (\text{diff}_i(a, b))^2$$

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

$$d_i^2(a, b) = (\text{diff}_i(a, b))^2$$

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

Definition 5. Given two datasets $X$ and $Y$ the square of the Euclidean distance for variable-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 $AM(c_1, \ldots, c_n) = \sum c_i / n$.

In general, any aggregation operator $C$ [36] might be used in the place of the arithmetic mean. It is important to note that not all aggregation operators will satisfy all the metric/distance properties. However, as we will show, most of the parametrized distances presented in this paper satisfy the distance properties explained in Definition 1, so we call them distances.

We can consider the following generic function.

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

From this definition, it is straightforward to consider weighted versions of the $d^2ED(a, b)$. We briefly revise two of them below.

Definition 6. Let $p = (p_1, \ldots, p_n)$ be a weighting vector (i.e., $p_i \geq 0$ and $\sum p_i = 1$). Then, square of the weighted mean is defined as:

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

where $WM_p = (c_1, \ldots, c_n) = \sum p_i \cdot c_i$.

In the context of supervised learning approaches for disclosure risk evaluation this was first used in [27,28]. The interest of this definition is that it does not assume that all attributes are equally important in the re-identification process, since there is a weight for each attribute expressing its relevance in the re-identification process. However, it is easy to see that when null weights ($p_i = 0$) and the square of the function are considered, the identity of indiscernibles and the triangle inequality (Definition 2) are not satisfied.

Another type of distance is based on the Choquet integral (Definition 7, see below). This was first introduced in the context of data privacy in [29]. From a definitional point of view, its main difference with respect to the weighted distance is the use of fuzzy
measures. Choquet integrals, with fuzzy measures permit us to represent, in the computation of the distance, information like redundancy, complementariness, and interactions among the variables, which are not used in the weighted mean. Therefore, tools that use fuzzy measures to represent background knowledge permit us to consider variables that, for example, are not independent.

**Definition 7.** Let $\mu$ 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 square of the Choquet integral distance is defined as:

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

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

As in **Definition 6**, the Choquet integral based distance cannot be considered a metric because it does not satisfy the triangle inequality and the identity of indiscernibles properties. Nevertheless, it is shown in [37] that the Choquet integral, with respect to a submodular measure, can be used to define a metric. That is, it is possible to use the Choquet integral as a metric just adding the following condition (submodularity) to the fuzzy measure:

$$\mu(A) + \mu(B) \geq \mu(A \cup B) + \mu(A \cap B)$$

for all $A, B \subseteq V$.

In Section 5.2 are presented the necessary problem constraints in order to consider the weighted mean ($d^2 WM$) and the Choquet integral ($d^2 CI$) as two distance functions.

Now, we present the symmetric bilinear form. Given a vector space $V$ over a field $F$, a bilinear form is a function $B : V \times V \to F$ which satisfies the following axioms for all $w, v, u \in V$:

1. $B(v + u, w) = B(v, w) + B(u, w)$
2. $B(w, v + u) = B(w, v) + B(w, u)$
3. $B(\alpha v, w) = B(v, \alpha w) = \alpha B(v, w)$
4. $B(v, w) = B(w, v)$

Given a square matrix $\Sigma$, we define a bilinear form for all $v, w \in V$ as $B(v, w) = v^T \Sigma w$. This form satisfies the axioms because of the distributive laws and the ability to pull out a scalar in matrix multiplication. Note that the matrix $\Sigma$ of a symmetric bilinear form must be itself symmetric. The symmetric bilinear functions can be considered a generalization of the Mahalanobis distance.

Then, we can use this symmetric bilinear form on the light of previous definitions as:

**Definition 8.** Let $\Sigma$ be a $n \times n$ symmetric weighting matrix. Then, the square of a symmetric bilinear form is defined as:

$$d^2 SB(a, b) = SB_2(diff_1(a, b), \ldots, diff_n(a, b))$$

where $SB_2(c_1, \ldots, c_n) = (c_1, \ldots, c_n)^T \Sigma (c_1, \ldots, c_n)$.

Learning the symmetric weighting matrix $\Sigma$ allows us to find which are the attributes and tuples of attributes that are more relevant in the re-identification process. That is, the diagonal expresses the relevance of each single attribute, while the upper or lower values of the weighting matrix correspond to the weights that evaluate all the interactions between each pair of attributes in the re-identification process.

If the matrix $\Sigma$ satisfies the symmetry and the positive definiteness property all the distance properties of **Definition 1** are satisfied. On the contrary, if this matrix restriction is weaker, the matrix is positive semi-definite, the identity of indiscernibles is not fulfilled. Thus, there will be situations where $d(a, b) = 0$ for all $a \neq b$, and then, the defined operator cannot be considered a distance anymore, it is a pseudo-distance. A clear example is when $\Sigma$ is completely null. Finally, if $\Sigma$ is neither positive definite neither positive semi-definite, i.e. negative definite, only one metric property is satisfied, the symmetry.

As we do not want negative distance values, the only requirement on $\Sigma$ we have considered is that it should be at least a positive semi-definite matrix.

Unlike the standard methods, such us the arithmetic mean (**Definition 5**), the interest of using **Definitions 6–8** is that they give different degrees of importance to variables in the re-identification process. This would be the case if one of the variables is a key-variable, e.g. a variable where $V_i = V^i$. In this case, all the variable weights should be zero except for the key-variable weight which should be assigned to one. Such an approach would lead to 100% of re-identifications. This is a clear example that justifies our decision to consider null weights. Taking into account null weights it is possible to analyze which of the variables are completely useless in the re-identification process. However, this choice forces us to renounce the identity of indiscernibles metric property.

Note that in **Definition 7 and 8** the interactions of different variables are taken into account by means of the fuzzy measure and the matrix $\Sigma$ respectively. Otherwise, in **Definition 6**, the weighting vector can only weight the variables individually.

Fig. 1 illustrates the classification of the different distances that we have explained in this section. As you can see the arithmetic mean is a special case of the weighted mean and at the same time the weighted mean is also a special case of both the Choquet integral and the Mahalanobis distance. Some more details about these relationships can be found in [38].

5. Supervised learning for record linkage

In this section we review the general formalization of the stated optimization problem for record linkage as well as the three described aggregation operators. Section 5.1 reviews the general optimization problem which was first introduced in [28]. Afterwards, in Section 5.2 we present the formalization problems for the weighted mean and the Choquet integral operators, i.e. **Definitions 6 and 7**. We also discuss the necessary problem modifications in order to satisfy all distance properties. Finally, Section 5.3 describes the optimization problem formalization for the proposed symmetric bilinear function.

5.1. General supervised learning approach for record linkage

We describe the formalization of the general supervised metric learning problem for distance-based record linkage. This formalization is presented as a generalization of the record linkage problem independent of any parameterized function. Defining the problem in a general form allows us to create multiple variations
of the problem depending on the parameterized distance function used. This problem variations rely on the specific parameterized function requirements that should be added to the problem as a set of constraints.

The problem is modeled as a Mixed Integer Linear mathematical optimization (MILP). More formally, the stated problem is expressed with a linear objective function and it is subject to a set of linear equalities and inequalities constraints. The difference between MILP and Linear Programming (LP) lies in the type of the variables considered. LP just considers real-valued variables whereas, MILP involves problems in which only some variables are constrained to be integers and the other variables are allowed to be non-integers (real). This fact makes MILPs harder problems. That is, LPs can be solved in polynomial time while, MILPs there are NP-complete problems [39] and therefore, there is no known polynomial-time algorithm.

For the sake of simplicity in the formalization of the process, we assume that each record \( b_i \) of \( Y \) is the protected version of \( a_i \) of \( X \). That is, files are aligned. Then, two records are correctly linked using a parameterized aggregation function, \( C_p \), when the distance between the records \( a_i \) and \( b_i \) is smaller than the distance between the records \( a_i \) and \( b_j \) for all other \( j \) different than \( i \). So, records belonging to the same entity are considered less distant in terms of the aggregation function. Fig. 2 shows an illustration of this scenario. Formally, we have that a record \( a_i \) is correctly matched when the following equation holds for all \( i \neq j \).

\[
C_p(a_i, b_j) < C_p(a_i, b_i)
\]  

(1)

In optimal conditions these inequalities should be true for all records \( a_i \). Nevertheless, we cannot expect this to hold because of the errors in the data caused by the protection method. Then, the learning process is formalized as an optimization problem with an objective function and some constraints.

Eq. (1) should be relaxed so that the solution violates some equations. 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 satisfy 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. Table 1 shows a graphical example of the problem division and the information needed for the learning process, i.e., the labels of the correct links, the ones that correspond to the same individuals.

The rationale of our formalization is that if for a record \( a_i \), Eq. (1) is violated for a certain record \( b_j \), then, it does not matter that other records \( b_h \), where \( h \neq j \neq i \), also violate the same equation for the same record \( a_i \). This is so because record \( a_i \) will not be re-identified.

Using these variables \( K_i \), and the constant \( C \), we have that all pairs \( i \neq j \) should satisfy

\[
C_p(a_i, b_j) - C_p(a_i, b_i) + CK_i > 0
\]

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 we can formalize the optimization problem that finds the set of parameter values defined for a given aggregation operator \( C \) that minimizes the number of incorrect links. That is,

\[
\text{Minimize } \sum_{i=1}^{N} K_i
\]

Subject to :

\[
C_p(a_i, b_j) - C_p(a_i, b_i) + CK_i > 0, \quad \forall i, j = 1, \ldots, N, \quad i \neq j
\]

(3)

\[
K_i \in \{0, 1\}
\]

(4)

This is an optimization problem with a linear objective function and linear constraints (Eqs. (3) and (4)). However, depending on which aggregation operator \( C_p \) we decide to use, we will have to add some additional constraints related to that aggregation operator and its parameters. In addition, we have to pay special attention to which is the polynomial degree of the aggregation operator we want to use and the parameter constraints, because it could lead us to deal with non-linear or non-quadratic programming problem.

If \( N \) is the number of records, and \( n \) the number of variables of the two data sets \( X \) and \( Y \). Then, the objective function, Eq. (2), consists of a summation of \( N \) control variables, one per each defined distances’ block, i.e., \( K_i \) for all \( i = 1, \ldots, N \). With respect to the total number of problem constraints; there are \( N(N - 1) \) constraints concerning to Eq. (3) and \( N \) constraints defining the control variable, Eq. (4). Therefore, there are a total of \( (N(N - 1)) + N \) constraints. Note that depending on the aggregation function \( C_p \) used, there will be more constraints in the problem. We will discuss the number of such constraints in the particular problems below.

**Table 1**

Data to be considered in the learning process.

<table>
<thead>
<tr>
<th>Block</th>
<th>Aggregator function</th>
<th>Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ki</td>
<td>C_p(a_i, b_j)</td>
<td>Must-link</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>K_i</td>
<td>C_p(a_i, b_i)</td>
<td>Cannot-link</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>K_N</td>
<td>C_p(a_i, b_N)</td>
<td>Cannot-link</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

**Fig. 2.** Distances between aligned records should be minimum.
5.2. Learning the optimal weights using $d^2\text{WM}$ on $d^2\text{CI}$

We outline in Table 2 the necessary extra constraints to formalize the general optimization problem (Eqs. (2)–(4)) for the weighted mean and the Choquet integral operators (Definitions 6 and 7 respectively). The table also includes the number of constraints of the optimization problem in each case. More details and deeper explanations can be found in the following works [28,29].

As was mention in Section 4 none of the aggregators showed in Table 2 satisfy completely the distance properties. Therefore, following the instructions given in that section we show in Table 3 which are the set of corresponding changes that have to be applied to each optimization problem. Thus, both the weighted mean and the Choquet integral can be considered distance functions.

Note that in all cases the additional constraints are linear. They are mixed integer linear problems (MILP), because they are dealing with integers and real-valued. Note, that we only have considered aggregation operators with real-valued weights.

5.3. Learning the optimal weights using a symmetric bilinear form

In this section we define the optimization problem and the specific constraints when $\mathbb{C}$ is based on Definition 8 (a symmetric bilinear function). The minimization problem is expressed as:

\[
\text{Minimize } \sum_{i=1}^{N} K_i \quad (5)
\]

Subject to:

\[
d^2\text{SB}(a_i, b_i) - d^2\text{SB}(a_i, b_i) + CK_i > 0, \, \forall i, j = 1, \ldots, N, \, i \neq j \quad (6)
\]

\[
\Sigma \geq 0 \quad (7)
\]

\[
K_i \in \{0, 1\} \quad (8)
\]

where as before, $N$ is the number of records, and $n$ the number of attributes of the input files.

One of the required distance properties for the matrix $\Sigma$ in Definition 8 was its positive semi-definiteness. To ensure this property, we can solve the problem with Semi-definite Programming (SDP) or using other methods that ensure the symmetry of the matrix and also that the matrix has non-negative eigenvalues [40]. Nevertheless, none of these approaches are technically feasible with linear constraints (they can only be formalized with nonlinear constraints). To avoid the non-linear constraints we have considered two approximations.

The first approximation ($d^2\text{SB}$) consists in changing Eq. (7) of the previous formalization by the following linear constraint:

\[
d^2\text{SB}(a_i, b_i) \geq 0, \, \forall i, j = 1, \ldots, N \quad (9)
\]

Eq. (9) forces the distance to be semi-positive for all pairs of records $(a_i, b_i)$ in the input set. Although, the $\Sigma$ positive semi-definite is not ensured, this approximation ensures that non-negativity will be satisfied for the input dataset.

The second approximation ($d^2\text{SBnc}$) does not consider the matrix restriction in the formalization of the optimization problem. Thus, the problem consists of a linear objective function Eq. (5) and two linear constraints Eqs. (6) and (8). Thus, this approach consists of solving the stated optimization problem and then, do a post-processing of the resulting matrix $\Sigma$. We apply the Higham’s algorithm [26] to the matrix $\Sigma$. This method computes the nearest positive semi-definite matrix from a non-positive definite.

Both proposed approximations have to determine the same number of parameters, $n(n + 1)/2$. They correspond to the diagonal and the upper (or lower) triangle of the matrix $\Sigma$. The first approach consists of a linear objective function plus $N(N-1) + N^2$ constraints. That is, the general plus all constraints related to Eq. (9). While, the second approach considers the same number of constraints as the general optimization problem: $N(N - 1) + N$.  

6. Evaluation

Given an original file and its masked version, we pre-process and build the problem structure by means of a series of $R$ functions, then following this formalized structure the problem is expressed into MPS (Mathematical Programming System) file format. MPS is a file format to represent and store Linear Programming (LP) and Mixed Integer Programming (MIP) problems. Then, each file is processed with an optimization solver. We solve our experiments with one of the most used commercial solvers, the IBM ILOG CPLEX tool [41] (version 12.1). Thus, for each formalized problem this solver finds the corresponding parameter values that maximize the number of correct links between the original and the masked data.

The experiments were performed in the Finis Terrae computer from the supercomputing center of Galicia [42]. Finis Terrae is composed of 142 HP Integrity rx7640 computing nodes with 16 Itanium Montvale cores and 128 GB of memory each, one HP Integrity Superdome node, with 128 Itanium Montvale cores and 1.024GB of memory, and 1 HP Integrity Superdome node, with 128 Itanium 2 cores and 384 GB of memory. From the Finis Terrae computer we used 16 cores and 32 GB of ram memory.

6.1. Test set

A data file was protected by a perturbative approach called microaggregation [9], a well-known microdata protection method, which broadly speaking, provides privacy by means of clustering the data into small clusters of size at least $k$, and then replacing the original data by the centroid of their corresponding clusters. This 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 files with the following protection parameters:

- **M4-33**: 4 variables microaggregated in groups of 2 with $k = 3$.
- **M4-28**: 4 variables, first 2 variables with $k = 2$, and last 2 with $k = 8$.

### Table 2

<table>
<thead>
<tr>
<th>Additional constraints</th>
<th>$d^2\text{WM}$</th>
<th>$d^2\text{CI}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\sum_{i=1}^{N} \mu_i = 1$</td>
<td>$\mu(B) = 0$</td>
<td>$\mu(V) = 1$</td>
</tr>
<tr>
<td>$\mu_i &gt; 0$</td>
<td>$\mu(A) &lt; \mu(B)$ when $A \subseteq B$</td>
<td>$\mu(A) + \mu(B) &gt; \mu(A \cup B) + \mu(A \cap B)$</td>
</tr>
<tr>
<td>Total constr.</td>
<td>$N(N - 1) + N + n$</td>
<td>$N(N - 1) + N + 2 + \sum_{k=2}^{n} \binom{n}{k} + \left( \frac{n}{2} \right)$</td>
</tr>
</tbody>
</table>

### Table 3

<table>
<thead>
<tr>
<th>Additional constraints</th>
<th>$d^2\text{WM}_{\text{M4-33}}$</th>
<th>$d^2\text{CI}_{\text{M4-28}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\sum_{i=1}^{N} \mu_i = 1$</td>
<td>$\mu(B) = 0$</td>
<td>$\mu(V) = 1$</td>
</tr>
<tr>
<td>$\mu_i &gt; 0$</td>
<td>$\mu(A) &lt; \mu(B)$ when $A \subseteq B$</td>
<td>$\mu(A) + \mu(B) &gt; \mu(A \cup B) + \mu(A \cap B)$</td>
</tr>
<tr>
<td>Total constr.</td>
<td>$N(N - 1) + N + n$</td>
<td>$N(N - 1) + N + 2 + \sum_{k=2}^{n} \binom{n}{k} + \left( \frac{n}{2} \right)$</td>
</tr>
</tbody>
</table>
- **M4-82**: 4 variables, first 2 variables with $k = 8$, and last 2 with $k = 2$.
- **M5-38**: 5 variables, first 3 variables with $k = 3$, and last 2 with $k = 8$.
- **M6-385**: 6 variables, first 2 variables with $k = 3$, next 2 variables with $k = 8$, and last 2 with $k = 5$.
- **M6-853**: 6 variables, first 2 variables with $k = 8$, next 2 variables with $k = 5$, and last 2 with $k = 3$.

For each case, we have protected 400 records randomly selected from the Census dataset [43] from the European CASC project [44], which contains 1080 records and 13 variables, and has been extensively used in other works [45–47].

In Table 4 we provide some basic statistical information from the Census dataset, such as the mean and the standard deviation for the first six columns. From it we can see how different are the data attributes in terms of their means, and also how spread out are the data points over a large range of values. In addition, in Fig. 3 is shown a graphical representation of the Pearson correlation coefficient, which indicates a degree of linear relationship between all pairs of attributes.

Note that in our experiments we apply different protection degrees to different variables of the same file. The values used range between 2 and 8, i.e., values between the lowest protection value and a good protection degree in accordance to [18]. This is especially interesting when variables have different sensitivity. We have used the web application [48], which is based on [18], to compute standard scores to evaluate all the protected datasets. These scores are computed by means of a combination of information loss and disclosure risk values, so the best protection method is the one that optimizes the trade-off between the information loss and the disclosure risk. Table 5 shows the average record linkage, the probabilistic information loss and the overall score for all the protected files. The best score is achieved by the M5-38 file, though the other files have a very similar score.

### 6.2. Results

Table 6 shows the linkage percentage using different approaches for record linkage. These percentages determine the maximum number of correctly identified records from the total, so a value of 100 means that all records from the original and the masked data were correctly linked (re-identified). The maximum number of correctly linked records are determined by the CPLEX solver for each generated MILP problem.

The approaches considered are the following ones: the standard record linkage method ($d^AM$); the Mahalanobis distance ($d^MD$); two supervised learning approaches: the weighted mean ($d^WM$) and the Choquet integral ($d^CI$) and their corresponding proposed versions satisfying the distance properties ($d^WMm$ and $d^CIm$), which were described in Section 4; and finally, the new supervised learning approaches, which are based on a symmetric bilinear form ($d^SB$ and $d^SBc$). Recall that whereas $d^SBc$ is the approach formed by Eqs. (5), (6) an (8), $d^SB$ has an extra constraint, Eq. (9).

Recall that due to the lack of constraints in the $d^SBc$ problem formalization, it is possible the solver finds a matrix that does not satisfy the positive semi-definiteness property. In these particular problems, we have applied the Higham’s algorithm [26] to the resulting matrix $\Sigma$. Thus, we are able to compute its nearest positive semi-definite matrix. Then, we check manually the number of correctly linked records with the symmetric bilinear function (Definition 8) and the matrix computed by the Higham’s algorithm. These cases are named $d^SBcH$.

Before tackling the results obtained by the presented supervised approaches we focus on the non-supervised approaches. The most noticeable fact between the standard distance-based record linkage ($d^AM$) and Mahalanobis distance ($d^MD$) is the improvement achieved by the latter method, which in average achieves about 22.6% more correct re-identifications and for the protected file M5-38 achieves a maximum improvement of 48.5%. This improvement and ease computation of Mahalanobis distance makes that $d^MD$ should be strongly considered for the disclosure

### Table 4
Mean and standard deviation ($\sigma$) for each column attribute.

<table>
<thead>
<tr>
<th>Attr.</th>
<th>Mean</th>
<th>Std. dev. ($\sigma$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$V_1$</td>
<td>196039.8</td>
<td>101251.47</td>
</tr>
<tr>
<td>$V_2$</td>
<td>56222.76</td>
<td>24674.843</td>
</tr>
<tr>
<td>$V_3$</td>
<td>3173.135</td>
<td>1401.832</td>
</tr>
<tr>
<td>$V_4$</td>
<td>7544.656</td>
<td>4905.200</td>
</tr>
<tr>
<td>$V_5$</td>
<td>45230.84</td>
<td>21323.470</td>
</tr>
<tr>
<td>$V_6$</td>
<td>2597.184</td>
<td>1826.436</td>
</tr>
</tbody>
</table>

### Table 5
Evaluation of the protected datasets. The best file score is highlighted in bold, while the best average record linkage and probabilistic information loss are highlighted in italics.

<table>
<thead>
<tr>
<th></th>
<th>AvRL(%)</th>
<th>PIL(%)</th>
<th>Score(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>M4-33</td>
<td>42.127</td>
<td>23.85</td>
<td>32.99</td>
</tr>
<tr>
<td>M4-28</td>
<td>33.47</td>
<td>28.40</td>
<td>30.94</td>
</tr>
<tr>
<td>M4-82</td>
<td>32.37</td>
<td>31.80</td>
<td>32.09</td>
</tr>
<tr>
<td>M5-38</td>
<td>26.01</td>
<td>31.92</td>
<td>28.96</td>
</tr>
<tr>
<td>M6-385</td>
<td>35.42</td>
<td>36.91</td>
<td>36.16</td>
</tr>
<tr>
<td>M6-853</td>
<td>30.65</td>
<td>37.76</td>
<td>34.21</td>
</tr>
</tbody>
</table>

### Table 6
Percentage of the number of correct re-identifications. The best scores for each column are highlighted in bold.

<table>
<thead>
<tr>
<th></th>
<th>M4-33</th>
<th>M4-28</th>
<th>M4-82</th>
<th>M5-38</th>
<th>M6-385</th>
</tr>
</thead>
<tbody>
<tr>
<td>$d^AM$</td>
<td>84.00</td>
<td>68.50</td>
<td>71.00</td>
<td>39.75</td>
<td>78.00</td>
</tr>
<tr>
<td>$d^MD$</td>
<td>94.00</td>
<td>90.00</td>
<td>92.75</td>
<td>88.25</td>
<td>98.50</td>
</tr>
<tr>
<td>$d^WM$</td>
<td>95.50</td>
<td>93.00</td>
<td>94.25</td>
<td>90.50</td>
<td>99.25</td>
</tr>
<tr>
<td>$d^{WM}m$</td>
<td>95.50</td>
<td>93.00</td>
<td>94.25</td>
<td>90.50</td>
<td>99.25</td>
</tr>
<tr>
<td>$d^CI$</td>
<td>95.75</td>
<td>93.75</td>
<td>94.25</td>
<td>91.25</td>
<td>99.75</td>
</tr>
<tr>
<td>$d^{CIm}$</td>
<td>95.75</td>
<td>93.75</td>
<td>94.25</td>
<td>90.50</td>
<td>99.50</td>
</tr>
<tr>
<td>$d^SB$</td>
<td>96.75</td>
<td>94.5</td>
<td>95.25</td>
<td>92.25</td>
<td>99.75</td>
</tr>
<tr>
<td>$d^SBc$</td>
<td>96.75</td>
<td>94.5</td>
<td>95.25</td>
<td>92.25</td>
<td>99.75</td>
</tr>
<tr>
<td>$d^SBcH$</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
risk assessment of protected datasets. However, as it is also shown in Table 6, these results can still be overcome by the presented optimization approaches.

We first compare the presented symmetric bilinear function approaches \((d^2SB\) and \(d^2SB_{NC}\)) with the weighted mean \((d^2WM)\) and the Choquet integral \((d^2CI)\) approaches. The results obtained by these supervised approaches show that almost for all the protected files the optimization problem with respect to the symmetric bilinear function \((d^2SB\) and \(d^2SB_{NC}\)) achieves the larger number of correct matches. Their results are slightly followed by the Choquet integral \((d^2CI)\) by a maximum difference of exactly 1% (4 correct matches less) for Mic-4-44 and Mic-5-38 protections. Improvements obtained by the Choquet integral are also slightly followed by the ones obtained by the weighted mean approach \((d^2WM)\), which has a maximum difference of 0.75% (3 correct matches less) for Mic-4-28 protection. In terms of accuracy (number of records correctly re-identified) we can conclude that from the supervised learning approaches the symmetric bilinear, the Choquet integral and the weighted mean are the best methods. Then, we compare their accuracies in those protected files where the standard record linkage approach \((d^2AM)\) achieve the maximum (M-6-853) and the minimum (Mic-5-38) number of re-identifications. We obtained an improvement of 14.76% (by \(d^2SB\)), 14.51% (by \(d^2BM\)) and 14.01% (by \(d^2WM\)) for the M-6-853 file and improvement of 52.5% (by \(d^2SB\)), 51.5% (by \(d^2CI\)) and 50.75% (by \(d^2BM\)) for the M-5-38 file. However, to evaluate all approaches it is also important to bear in mind the problem complexity and its computing time, factor that we analyze below, in Table 7.

We now focus on the results obtained by the weighted mean \((d^2WM)\), the Choquet integral \((d^2CI)\) and their respectively modified versions which satisfy or almost all the metric properties \((d^2WM_a\) and \(d^2Cl_a\)). Table 6 shows how similar they are. Comparing the \(d^2WM\) and \(d^2WM_a\) re-identification percentages we appreciate that both approaches obtain exactly the same values. With respect to the Choquet integral approaches we see the not modified version \((d^2CI)\) achieves slightly better results in some of the datasets tested than \(d^2Cl_a\). Therefore, despite of adding new constraints to the problem there is a slight decrease (or none decrease) in the number of re-identifications.

Finally, we focus on both symmetric bilinear approximations. Let us underline that all matrices by \(d^2SB\) and \(d^2SB_{NC}\) satisfy the positive definiteness property, except for the last dataset (M-6-853), which in either of the two approaches this property was not satisfied. The Higham’s algorithm was applied to the matrix obtained by the solver for the \(d^2SB_{NC}\) approach achieving a new positive definite matrix. \(d^2SB_{PM}\) in Table 6 shows the percentage results for this test case. We note that the percentage of correct re-identifications is slightly lower than for \(d^2SB_{NC}\) but is still higher than the rest of the analyzed methods. Recall that when the obtained matrix is positive definite all distance properties are satisfied as well as the identity of indiscernibles. Using the symmetric bilinear approach with a positive semi-definite matrix achieves better results that the Mahalanobis distance using the covariance matrix compute from the data.

### Table 7

<table>
<thead>
<tr>
<th></th>
<th>M4-33</th>
<th>M4-28</th>
<th>M4-82</th>
<th>M5-38</th>
<th>M6-385</th>
<th>M6-853</th>
</tr>
</thead>
<tbody>
<tr>
<td>(d^2WM)</td>
<td>29.83</td>
<td>41.37</td>
<td>24.33</td>
<td>718.43</td>
<td>11.81</td>
<td>17.77</td>
</tr>
<tr>
<td>(d^2WM_a)</td>
<td>3.43</td>
<td>6.26</td>
<td>2.26</td>
<td>190.75</td>
<td>4.34</td>
<td>6.72</td>
</tr>
<tr>
<td>(d^2Cl)</td>
<td>280.24</td>
<td>427.75</td>
<td>242.86</td>
<td>42731.22</td>
<td>24.17</td>
<td>87.43</td>
</tr>
<tr>
<td>(d^2Cl_a)</td>
<td>155.07</td>
<td>441.99</td>
<td>294.98</td>
<td>4017.16</td>
<td>74.93</td>
<td>829.81</td>
</tr>
<tr>
<td>(d^2SB_{NC})</td>
<td>32.04</td>
<td>2793.81</td>
<td>150.66</td>
<td>10592.99</td>
<td>13.65</td>
<td>14.11</td>
</tr>
<tr>
<td>(d^2SB)</td>
<td>13.67</td>
<td>3479.06</td>
<td>139.59</td>
<td>169049.55</td>
<td>13.93</td>
<td>13.70</td>
</tr>
</tbody>
</table>

The computation time taken to learn the optimal weights for each dataset and learning approach can be seen below, in Table 7.

Moreover, we have compared the covariance matrices used in \(d^2MD\) and the inverses of the weighting matrices obtained by the supervised approach using the symmetric bilinear function \(d^2SB_{NC}\) for the first five datasets and the matrix obtained by \(d^2SB_{PM}\) for the last case, because of the positive semi-definiteness. These are supposed to be similar than the covariance matrices or a scaled variation of those. However, when we compare both matrices by means of the mean square error (Eq. (10)), the results show that both matrices are different. See Table 8.

\[
\text{MSE}(V, V') = \frac{\sum_{j=1}^{n} \sum_{i=1}^{n} (V_{ij} - V'_{ij})^2}{n(n-1)}
\]

### Table 8

<table>
<thead>
<tr>
<th></th>
<th>Mean square error</th>
</tr>
</thead>
<tbody>
<tr>
<td>M4-33</td>
<td>18.49</td>
</tr>
<tr>
<td>M4-28</td>
<td>48.75</td>
</tr>
<tr>
<td>M4-82</td>
<td>2784.81</td>
</tr>
<tr>
<td>M5-38</td>
<td>7.26</td>
</tr>
<tr>
<td>M6-385</td>
<td>15.91 \times 10^6</td>
</tr>
<tr>
<td>M6-853</td>
<td>12.77 \times 10^6</td>
</tr>
</tbody>
</table>

7. Conclusions

In this paper we introduced a new supervised learning approach and a parameterized aggregator, a symmetric bilinear function, to solve record linkage problems. This approach is formalized as an optimization problem defined by a set of cannot-link and must-link constraints. Thus, the problem is solved by finding the parameter values of the symmetric bilinear function that maximizes the number of correct links between two datasets. We have compared this supervised learning method with other supervised and non-supervised ones.

Our experiments have been done in the area of data privacy. In this area, record linkage is used to evaluate disclosure risk. It is used to link records of the original and the protected file, modeling the attack of an intruder that wants to disclose information from the protected file. We have focused on the worst case. This is the case in which the person who wants to do the re-identification has the entire original database. Note that as the original data is confidential this scenario is only applicable by the data owner to evaluate the risk of the protected file. The parameterized distance based record linkage is a very useful tool for the data owner, not only to evaluate the disclosure risk of the protected database before its release, but also to know which are the variables or sets of variables that maximize the number of re-identifications and make weaker the protected data. This estimation is based on the number of correct links between the original and the protected data.

The experiments show that the proposed approach is the one that achieves the best results. Although, the improvement is not very high, especially when we compare it with the other parameterized variations, it is relevant for the evaluation of risk of a protected dataset. Moreover, by means of analyzing the weights obtained it is possible to identify the variables and sets of variables that clearly provide more information for an attacker of the database. This is useful in the data protection process. For instance, when a variable provides more information than the others, we would apply a higher degree of protection or even another protection method to make it more secure.
In this paper we also introduced two variations for the weighted mean and Choquet problems in order to be considered metrics \( (d^W_{WM}, d^W_{CI}) \). We can conclude that for our problem, the record linkage for disclosure risk evaluation, they are not promising, so the number of re-identifications slightly decrease when compared with their original approaches \( (d^W_{WM}, d^W_{CI}) \). Besides, in the case of \( d^W_{WM} \) null weights are not considered and so there is a lack of information for those variable which are not relevant in the re-identification process.

As future work we consider developing non-linear program-

ming optimization problems. For example, to consider the case in which the weighting matrix satisfies the positive semi-definiteness property of the covariance matrix. We will compare the results and computational time of such approach with the other presented methods. Furthermore, it would be interesting a comparison between the described supervised learning approach reyeling on the Choquet integral and a similar semi-supervised metric learning research approach proposed by Beliakov et al. in [23].

Acknowledgments

Partial supports by the Spanish MICINN (projects ARES-CO-

SOLIDER INGENIO 2010 CSD2007-00004, TIN2010-15764 and TIN2011-27076-C03-03) and by the European Commission (EC (FP7/2007–2013) Data without Boundaries (Grant agreement number 262608) are acknowledged. Some results described in this paper have been obtained using the Centro de Supercomputación de Galicia (CESCA). This partial support is gratefully acknowledged. The work contributed by the first author was carried out as part of the Computer Science Ph.D. program of the Universitat Autònoma de Barcelona (UAB).

References

[40] CESCA, Centro de supercomputación de galicia <http://www.cesga.es>.