OWA Operators in Data Modeling and Reidentification

Vicenç Torra, *Senior Member, IEEE*

***Abstract*—This paper is devoted to the application of aggregation operators and to the application of ordered weighting averaging (OWA) operators to data mining. In particular, we consider two application of OWA operators in this field: model building and information extraction. The latter application is oriented to the reidentification procedures.**

***Index Terms*—Aggregation operators, learning models, ordered weighting averaging (OWA) operators, privacy preserving data mining, record linkage, reidentification methods, weighted mean.**

I. Introduction

INFORMATION fusion techniques, in general, and aggregation operators, in particular, are currently being used in several scientific fields. In fact, their use is rapidly increasing because, on the one hand, data is gradually obtained in an easier way and, on the other hand, computational power has largely increased so that systems that combine information from several experts or sensors are nowadays feasible. Even more, it is also possible to compute several solutions using different approaches and then combine the alternative solutions. This is the case of bagging, boosting and related approaches used for building data models in machine learning. In these approaches, several models are built using different mechanisms and, then, a decision making module (e.g., voting [2], [16], the weighted mean [17], or the ordered weighting averaging (OWA) operator [42]) is put on top of the models.

In general, in the field of artificial intelligence, data fusion techniques are mainly used for two main purposes: 1) when a system has to make a decision, or 2) when it needs a comprehensive representation of its domain.

In the first case, an alternative has to be selected or built from several ones. The typical case for selection is to consider several criteria for each alternative (this corresponds to a multicriteria decision-making problem) and the best alternative is usually chosen [27] in a two phase process: 1) the aggregation of the degree of satisfaction for all criteria, per decision alternative; and 2) the ranking of the alternatives with respect to the global aggregated degree of satisfaction. Instead, when the alternative has to be built from the existing ones, fusion corresponds to the whole building process and it has to consider the importance and reliability of the alternatives and of the approaches used to build these alternatives. Plan merging can be seen from this point of view.

In the second case, a system builds the representation of its environment from some background knowledge embedded in the system and some knowledge supplied by some information sources (e.g., experts or sensors). Naturally, the knowledge has to be "reliable" and extend on the whole domain of system's actuation. However, the information supplied by a single information source is often not reliable enough and also too narrow in relation to the working domain. In this case, the information provided from several sensors or experts are combined to improve data reliability and accuracy and to include some features that are impossible to be perceived from individual sensors. Note that information fusion for knowledge representation can be either applied at the time of defining the background knowledge (e.g., using several experts' knowledge) or at run-time (either combining different pieces of new information or combining some new information with some knowledge already established in the system).

Nevertheless, other uses of data fusion techniques are conceivable in the artificial intelligence field. In particular, fusion is useful in data mining and knowledge discovery for two types of applications. On the one hand, aggregation operators and, in general, any data fusion model, are suitable for building data models. This capability relies on a result published in [34]. In this paper, it is proved that a model defined in terms of a hierarchy of quasiweighted means (a kind of aggregation operator; see [12] or [39] for details on the operator) is a universal approximator and, as such, it can be used for representing any arbitrary complex function. Therefore, given a data set it is possible to build a model of this data set using aggregation operators. On the other hand, aggregation operators can be used to extract useful information from raw data and, thereafter, other programs can use the structures that otherwise would remain implicit.

In this paper, we consider both approaches: the use of aggregation operators to build data models, and its use to extract implicit information. In the former case, due to the large number of aggregation operators we concentrate in some of them for which learning models are simple (small cost) and, thus, applicable to data mining. In particular, we focus on the OWA operator [47]. In relation to the latter case, we show an application to the reidentification of individuals in data files with noncommon variables. We focus on the use of aggregation operators to extract implicit information from files and their application to reidentification.

The structure of this paper is as follows. In Section II, we review the state of the art of the two approaches. Then, in Section III we review some definitions that are needed later on in

Manuscript received February 12, 2003; revised February 6, 2004. This work was supported in part by the European Community under Contract IST-2000-25069, by the U.S. Census Bureau under Contract OBLIG-2000-029144-0-0, and by MCyt under Contract TIC2001-0633-C03-02.

The author is with the Institut d'Investigació en Intel·ligència Artificial-CSIC, Campus UAB s/n, 08193 Bellaterra, Catalunya, Spain (e-mail: vtorra@iiia.csic.es).

Digital Object Identifier 10.1109/TFUZZ.2004.834814

this work. Section IV is devoted to describe a method to determine models for quasiweighted means. Section V corresponds to the use of aggregation operators to extract information and its application to reidentification. The work finishes in Section VI with some conclusions.

II. State of the Art

In this section, we first review recent work on the use of aggregation operators for building models. Then, we give an overview of reidentification algorithms.

Aggregation operators have been widely studied; see, e.g., [4], [19], and [40] for a detailed state-of-the-art description of the field; see [46] for or a recent overview on aggregation operators.

A. Aggregation Operators for Building Data Models

Theorem 1 in [34] establishes that hierarchies of quasiweighted means are universal approximators. This result implies that quasi weighted means in particular, and aggregation operators in general, are suitable tools for modeling data and, as a consequence, methods can be developed to build models for complex data. When efficient tools are developed, methods can be used for large data bases.

At present, methods to determine the suitable aggregation model have been focused on mechanisms to determine the parameters of the operator once the aggregation operator is selected. Several approaches have been developed in the last years. They can be broadly classified in two groups.

On the one hand, there are some methods based on the assumption that there exists an expert that supplies crucial information that is used later on to extract the parameters of the selected aggregation function. This is the case of Saaty's analytical hierarchy [28] process (used to determine the weighting vector of a weighted mean). A similar approach is the one followed in [22] to determine the weighting vector of the OWA operator. He requires that a user supplies the so-called *orness*, a measure of how much large values influence the outcome of the operator. This method was further developed in [5].

On the other hand, there are some methods that do not require the presence of an expert but the existence of a set of examples. In this case, an example consists on the input values (i.e., values to be aggregated) and the expected result (i.e., the value that the model estimates). From these examples, the parameters of the operators are inferred and the model is fitted on the data. This approach is rooted on estimation theory where two random variables X and Y are considered such that Y is said to depend on X when the distribution $Y|X$ is different from the distribution Y. This is, the distribution *posterior* to observation is different from the *a priori* one. Then, a model is built to estimate the value of Y such that the variance of error is minimized; see [3] and [13] for a more detailed discussion of learning parameters from this perspective.

Several works can be found in the literature about parameter learning from examples. For example, [9] and [10] study the determination of the weighting vector for the OWA operator. In a similar way, [33] studies the learning for both the weighted mean and the OWA operator. [12] and [31] deal with the learning of fuzzy measures for the Choquet integral and [13] compares different approaches for modeling using the Choquet integral.

In the framework of data mining, the second approach is of interest. It allows to build a data model from existing data. However, not all the approaches to build the model are suitable as some methods have a high computational cost. Iterative computation with a cost proportional (linear or even higher) to the number of examples is prohibitive. For example, genetic algorithm-based approaches [20] are, usually, extremely costly as the fitness function needs to compute at each step an aggregated value for all the examples. This is also the case for the method described in [9] and [10] based on gradient descent. However, in this latter case, if it is possible to start with a "quite good solution," then it can be revised as long as time and resources are available. A good alternative for building models in the case of quasiweighted means is to use active set methods. These methods are iterative ones and thus can obtain different solutions according to the "available" time. In this approach, the cost of the initial step is proportional to the number of examples, then, once the iterative process is started, the cost of a single step is proportional to the number of considered variables. Moreover, the number of steps of the iterative process is bounded and thus the best solution can be found in a finite time.

B. Reidentification of Individuals

Reidentification happens when two entities are detected as corresponding to the same object or individual. For example, when some sensitive and confidential data is linked to a particular individual. Record linkage is one of the most general reidentification method. Its goal [21], [26] is to link records in separate files that relate to the same individual or household. These methods were developed to improve the quality of the data and are nowadays used in data cleaning [24] for distributed and nonhomogeneous databases. Such databases typically [41] contain information about the same individuals described using the same variables that, frequently, do not match due to accidental distortion of the data. Record linkage is applied in such cases to find the records that correspond to the same individuals and to make databases consistent. Existing tools for this purpose (e.g., Integrity [14]) use statistical and artificial intelligence techniques to determine matching between records. Multidatabase mining, that intends to extract knowledge from nonhomogeneous databases (see, e.g., [50]), also benefits from these tools.

References [11], [44], and [41] describe the main approaches for record-linkage. The usual case is to consider files that share a set of variables. In this case, the main difficulty is that a matching procedure among pairs of records is usually not enough to link the records. This is so because data files are subject to errors (either due to intentional or to accidental distortion) and thus not only equal values should match but also similar ones. As [45] points out, "the normal situation in record linkage is that identifiers in pairs of records that are truly matches disagree by small or large amount and that different combinations of the nonunique, error-filled identifiers need to be used in correctly matching different pairs of records."

This situation of files sharing a set of variables is usually dealt by probabilistic record linkage [44] or distance-based record linkage [23]. The former is based on estimating (using the EM algorithm; see [6] and [15] for details) conditional probabilities of coincidence of the values of a particular variable when a true match or a true nonmatch is assured. This is, which is the probability that the corresponding values for a particular variable is the same, when two records are known to correspond (or not correspond) to the same individual. Then, given a pair of records the pair is classified as either corresponding to the same individual or not according to an index computed from these probabilities. While first methods assumed conditional independence between variables [15], [21], more recent works avoid such assumption; see [41] for a detailed description of probabilistic record linkage and [45] for a description of current approaches and research topics. The distance-based record linkage consists on linking a record with the more similar one. This approach relies on the existence of a distance function. [41] compares both methods and concludes that the probabilistic one is slightly better (reidentifies more records) for categorical variables and that the distance-based one is more appropriate for numerical variables. Recently, alternative methods based on other assumptions have been introduced in the literature. For example, [1] describes a method based on clustering techniques.

Although most methods follow the ideas explained previously, that files share a set of common variables, other situations are also possible. In particular, it is also of interest the case of files not sharing any variable (or only a few of them). In this case, reidentification is partially possible but being of a different nature because it cannot be based on the comparison of values from records of different files but corresponding to the same variable.

Record linkage for files not-sharing variables is of interest when considering data files with similar information (e.g., economical variables) from consecutive time periods (e.g., two different years) concerning to almost the same individuals (e.g., the companies of a certain region). In this case, although the variables are not the same, "similar" behavior of variables in both files allows for the reidentification of the individuals. Naturally, the more similar the behavior, the better for reidentification. Nevertheless, although this is a subject of increasing interest, no much effort has been devoted to the subject.

In [35], we gave a first approximation to the reidentification for files not sharing variables. A more exhaustive analysis of the approach using real data is described in [8]. In [35], and [8], some basic guidelines are established to allow for reidentification: 1) files share a set of individuals, and 2) some relationships between individuals are kept across files. In [35] and [8], these relationships are established using clustering algorithms. These methods differ from [1], that also uses clustering for reidentification, on the way clusters are defined and how the records are linked once their corresponding clusters are known.

Here, we follow, a different approach suitable for quantitative variables. We show that some aggregation operators (in particular, we focus on the OWA [47] operators) are a suitable way to extract implicit structures from data. This work extends [37]. Here, we give empirical results and we proof that the number of reidentified elements are significant.

III. Preliminaries

In this section, we review some aggregation operators that are used latter on in this work. In particular, definitions for the weighted mean, quasiweighted mean and the OWA operator (the one based on a weighting vector and the one based on nondecreasing fuzzy quantifiers) are given.

Definition 1: A vector $w = (w_1, \ldots, w_n)$ is a *weighting vector* of dimension n if and only if $w_i \in [0,1]$ and $\sum_i w_i = 1$.

Definition 2: Let w be a weighting vector of dimension n, then a mapping $\mathrm{WM}_w : \mathbb{R}^n \to \mathbb{R}$ is a *weighted mean (WM)* of dimension n if

$$\mathrm{WM}_w(a_1, \ldots, a_n) = \sum_i w_i a_i.$$

Definition 3: Let w be a weighting vector of dimension n, let f be a strictly increasing function, then a mapping $\mathrm{QWM}_w : \mathbb{R}^n \to \mathbb{R}$ is a *quasi-weighted mean (QWM)* of dimension n if

$$\mathrm{WM}_w(a_1, \ldots, a_n) = f^{-1}\left(\sum_i w_i f(a_i)\right).$$

For properly selected functions $f(x)$, QWMs generalize some well-known aggregation operators. See, for example, that with $f(x) = Kx + K'$ we obtain the weighted mean and with $f(x) = Klogx + K'$ we obtain the geometric mean.

Definition 4: [47] Let w be a weighting vector of dimension n, then a mapping $\mathrm{OWA}_w : \mathbb{R}^n \to \mathbb{R}$ is an OWA operator of dimension n if

$$\mathrm{OWA}_w(a_1, \ldots, a_n) = \sum_i w_i a_{\sigma(i)}$$

where $\{\sigma(1), \ldots, \sigma(n)\}$ is a permutation of $\{1, \ldots, n\}$ such that $a_{\sigma(i-1)} \geq a_{\sigma(i)}$ for all $i = 2, \ldots, n$. (i.e., $a_{\sigma(i)}$ is the ith largest element in the collection $a_1, \ldots, a_n$).

This definition of the OWA operator requires a weighting vector of fixed dimension being its dimension the number of elements to aggregate. An alternative definition based on decreasing fuzzy quantifiers exists that can be used for data vectors of arbitrary size. This alternative definition allows the comparison of different size data vectors: comparison with respect to the outcome of the OWA operator. Similarity of two data vectors can then be measured as a function of the differences between the outcomes of the OWA operators applied to the vectors.

Definition 5: [48]. A function $Q : [0,1] \to [0,1]$ is a *nondecreasing fuzzy quantifier* if $Q(0) = 0$, $Q(1) = 1$ and for all x, y in [0,1], $x < y$ implies $Q(x) \leq Q(y)$.

Definition 6: [48], [49]. A mapping $\mathrm{OWA}_Q : \mathbb{R}^n \to \mathbb{R}$ is an OWA operator of dimension n if

$$\mathrm{OWA}_Q(a_1, \ldots, a_n) = \sum_i w_i a_{\sigma(i)}$$

where $w_i = Q(i/n) - Q((i-1)/n)$ and where σ is defined as in Definition 4.

In this way, an OWA operator with a suitable fuzzy quantifier can compute a value with which *all* information sources agree, *at least one* source agree, *about half* agree, etc. Note that the same quantifier allows the computation of the aggregated value

TABLE I
DATA EXAMPLES

a_1^1	a_2^1	...	a_N^1	b^1
a_1^2	a_2^2	...	a_N^2	b^2
$\vdots$	$\vdots$		$\vdots$	$\vdots$
a_1^M	a_2^M	...	a_N^M	b^M

for an arbitrary number of inputs (sources) and that the quantifier only refers to a proportion of the sources but not to an exact number.

Therefore, the OWA operator, specially when defined in terms of a fuzzy quantifier, is a flexible tool that is appropriate when the number of variables is not known or can change in different instantiations of the same problem. This is the case in the reidentification problem described in Section V and similar situations can be envisioned in ensemble methods (new partial models are included in the system without modifying the decision making module). In fact, the flexibility of OWA operators compare positively in relation to other operators as e.g. the weighted mean (where the number of sources has to be fixed beforehand and weights are assigned to sources). Moreover, more complex operators like the Choquet and Sugeno integrals present additional difficulties because the number of parameters tend to be extremely large (2^n where n is the number of sources or variables).

IV. AGGREGATION OPERATORS FOR BUILDING DATA MODELS

This section is focused on mechanisms for learning parameters for aggregation operators from examples. Examples are assumed to be in accordance with Table I. This is, it consists on M different examples, each of them consisting on the values supplied by N information sources and the *correct* outcome that we intend to estimate from these values. Therefore, each example consists on $N+1$ values being $(a_1^i a_2^i \dots a_N^i | b^i)$ the ones for the ith example where a_j^i is the value supplied by the jth information source and being b^i the ideal outcome for the same example.

Given a set of examples and assuming that the function used to aggregate $\mathbb{C}$ is known, the goal is to determine the parameters of $\mathbb{C}$. When $\mathbb{C}$ is the weighted mean, this is to find the weighting vector w so that error is minimized. Error is measured for each example in terms of the difference between the ideal outcome (i.e., b^j) and the real outcome. This is, in the general case

$$\left(\mathbb{C}\left(a_1^j, \dots, a_n^j\right) - b^j\right)^2$$

and when $\mathbb{C}$ is the weighted mean

$$\left(\mathrm{WM}_{\mathbf{w}}\left(a_1^j, \dots, a_n^j\right) - b^j\right)^2.$$

Therefore, following the notation in Table I, the function to minimize is

$$D_{\mathbb{C}}\left(\text{parameters}(\mathbb{C})\right) = \sum_{j=1}^{M}\left(\mathbb{C}\left(a_1^j, \dots, a_n^j\right) - b^j\right)^2. \quad (1)$$

However, this problem is usually a constrained one because there usually exist constraints over the parameters. This is the case of the weighted mean, where w is a weighting vector. In this case, the problem can be formalized in the following way

$$\begin{aligned} &\text{Minimize } D_{\mathrm{WM}}(w) \\ &\text{Subject to} \\ &\qquad \sum_{i=1}^{N} w_i = 1 \\ &\qquad w_i \geq 0 \end{aligned}$$

where $D_{\mathrm{WM}}(w) = \sum_{j=1}^{M}(\mathrm{WM}_w(a_1^j, \dots, a_n^j) - b^j)^2$.

A. Solving the Optimization Problem

The problem formulated previously is a typical optimization problem where a function has to be minimized subject to a set of constraints. There exist several methods to solve these problems according to the function to minimize (quadratic, convex, ...) and the type of restrictions that apply (linear constraints, equality constraints, ...). When the distance to minimize is either $D_{\mathrm{WM}} = \sum_{j=1}^{M}(\mathrm{WM}_{\mathbf{w}}(a_1^j, \dots, a_n^j) - b^j)^2$ or $D_{\mathrm{OWA}} = \sum_{j=1}^{M}(\mathrm{OWA}_{\mathbf{w}}(a_1^j, \dots, a_n^j) - b^j)^2$ the problem to solve is a quadratic one subject to linear constraints. In such a case, several methods apply. For example, [9], [10], and [33] applied two different approaches: [9], [10] used the gradient descent for the OWA operator and [33] used active set methods for the weighted mean and the OWA operator.

In order to avoid the inconvenience of dealing with the inequality and equality constraints ($w_i \geq 0$ and $\sum_i w_i = 1$) when applying gradient descent to the OWA operator [9], [10] reformulated the problem. Instead of considering the learning of the weighting vector w, they considered the learning of a vector $\mathbf{\Lambda} = (\lambda_1 \dots \lambda_N)$ from which weights were extracted as follows:

$$w = \left(\frac{e^{\lambda_1}}{\sum_{j=1}^{N} e^{\lambda_j}} \dots \frac{e^{\lambda_N}}{\sum_{j=1}^{N} e^{\lambda_j}}\right).$$

In this way, any vector $\mathbf{\Lambda} \in \mathbb{R}^{\mathbf{N}}$ leads to a weighting vector. Therefore, the problem of learning weights for the OWA operator (we assume again that available data follows the description in Table III) is equivalent to the minimization of

$$D_{\mathrm{OWA}}(p) = \sum_{j=1}^{M} \mathrm{OWA}_w\left(a_1^j, \dots, a_n^j\right) - b^j\Big)^2 \quad (2)$$

where $\mathbf{w} = (\mathbf{e}^{\lambda_1}/\sum_{\mathbf{j=1}}^{\mathbf{N}} \mathbf{e}^{\lambda_j} \dots \mathbf{e}^{\lambda_N}/\sum_{\mathbf{j=1}}^{\mathbf{N}} \mathbf{e}^{\lambda_j})$

Reference [33] presents with great detail an alternative approach based on active set methods. Active set methods rely on the simplicity of computing the solution of quadratic problems with linear equality constraints. Based on this, iterative algorithms have been developed in which at each step inequality constraints are partitioned into two groups: Those that are to be treated as active (considered as equality constraints) and inactive (essentially ignored). Once a partition is known, the algorithm proceeds moving on the surface defined by the working set of

constraints (the set of active constraints) to an improved point. In this movement some constraints are added to the working set and some others are removed. This process is repeated until the minimum is reached. When the function to minimize is convex (as they are D_{WM} and D_{OWA}) the method finds the minimum and although the method is iterative the final minimum is not influenced by the initial weighting vector.

Gradient descent requires the computation of the gradient at each step; and the computation of the gradient in successive steps require the evaluation of the examples. As the set of examples is usually very large in data mining domains, it is difficult to apply this approach in this field unless an initially good solution is considered and the iterative process is limited to refine the initial solution (this was the case in [38]). Other disadvantages for gradient descent are its slow convergence and the problem of having more than one $\mathbf{\Lambda}$ vector that correspond to the same weighting vector. This, together with the fact of being an iterative process provokes that the final result depends often on the initial weighting vector.

Active set methods require an initial computation of a square matrix from the original data. Although the dimension of this matrix is the number of variables, its construction cost is on the order of $N*N*M$ (being, as before, N the number of variables and M the number of elements) and thus it is a time-consuming task proportional to the number of elements. However, once this matrix is built, each step of the iterative process has a polynomial cost on N.

Thus, for the weighted mean and the OWA operator is more efficient to use the second approach. However, when the function to minimize is not quadratic (as is the case of the weighted OWA operator [32]), active set methods are complex and difficult to implement because it is not easy to find the minimal solution at each step and, instead, gradient descent can be used. Software packages can be used at this time.

Both learning methods have been applied to determine the weights for the weighted mean and the OWA operator (in practice, the only difference when learning the weights for the OWA operator in relation to the learning for the weighted mean is that each row in the data matrix has to be reordered according to the permutation σ).

The same approach can be used to learn the weights for a selected quasiweighted mean (see [39]). This is, a quasiweighted mean with a known generator function. Let f be the generator of the quasiweighted mean, then the parameters are determined considering in the distance to minimize the following expression:

$$\left(\sum_i w_i f\left(a_i^j\right) - f(b^j)\right)^2$$

instead of the original one

$$\left(f^{-1}\left(\sum_i w_i f\left(a_i^j\right)\right) - b^j\right)^2.$$

This is so because both expressions would lead to a similar distance when weights are learned, and the former is easier to minimize. It allows to compute in an initial step the values $f(a_i^j)$ for all i and j, and then the methods described for the weighted mean can be applied. Considering the other expression, the minimization problem becomes a nonquadratic problem and, thus, more difficult to solve. In a recent work, Beliakov [3] has also considered such an optimization problem.

B. Examples

Example 1: Three examples are considered later. They are taken from the machine learning repository [18]: The *iris* data file (four variables and 150 examples), the *abalone* data file (eight variables and 4177 examples) and the *ionosphere* data file (34 variables and 351 examples—one, that is always zero, is removed). To use these files some preliminary work was required. First, we had to replace all symbolic variables by numerical ones. These were the changes performed: the classes iris-setosa, iris-versicolor, and iris-virginica in the *iris* data file were replaced by numerical values 1.0, 2.0, and 3.0; the three categories M, F, and I (infant) in the variable Sex in the *abalone* data file were also replaced by the numerical values 1.0, 2.0 and 3.0; and the classes "g" and "b" in the *ionosphere* data file were replaced by 1.0 and 0.0. Second, we normalized all the variables in the [0,1] interval.

For each of the resulting files, active set methods were applied and two models were built: one for the weighted mean and the other for the OWA operator. The corresponding distances are given in Table II.

This example shows the suitability of the approach and its cost make it appropriate for large data files. Note that in the ionosphere case, the number of variables is large (34 variables) but the computational cost in the iterative process only depends on the number of variables.

In [33], this approach was compared against [10] for some toy examples described in this latter work and results showed a better performance of our approach (error was reduced from 0.002 156 to 0.001 256).

V. Aggregation Operators to Extract Information From Data

It is a well-known fact that an aggregation operator summarizes the information that the information sources supply. In particular, practical use of these operators is motivated by their capability in reducing the uncertainty associated to the data, compensating redundancy and, in general, to extract relevant information. These properties are the ones that make these operators interesting for extracting information from raw data.

Also, these properties are interesting for reidentifying individuals from data files not sharing variables as, a priori, in such situations the only available information is the one available in the files. In this case, if files contain the "same information" a working hypothesis is that when the aggregation is applied to two records corresponding to the same individual, the relevant information emerges from the raw data. According to this, we assume that for each record a representative that is *somehow* independent of the actual data can be computed. This independence has to hold so that the two representatives (one for each file but from the same individual) are similar. Also, ideally, the

TABLE II
OPTIMAL DISTANCES FOR THE IRIS, ABALONE AND IONOSPHERE DATA FILES USING LEARNED PARAMETERS FOR WEIGHTED MEAN D_{WM} AND OWA OPERATOR D_{OWA}

	iris	abalone	ionosphere
D_{WM}	3.1544	37.5189	55.9145
D_{OWA}	6.3197	43.8082	62.9477

representatives are similar although variables (and the corresponding values) are different. In fact, this hypothesis holds if variables are correlated (or, more precisely, if one set of variables *as a whole* is "correlated" with respect to the other set of variables). This is, for example, the case of "income" and "size of household." Therefore, two files, one containing the information corresponding to "income" and the other with the information corresponding to "size of household" could be used for reidentification of individuals.

However, as the representative value has to be *somehow* independent of the variables, not all aggregation operators can be applied. In fact, it seems that the best ones are those that are commutative (i.e., $\mathbb{C}(a_1,\ldots,a_n) = \mathbb{C}(a_{\sigma(1)},\ldots,_{\sigma(n)})$ for any permutation σ) because particular variables do not have any influence on the output. According to this, the weighted mean and the weighted OWA [32] are not applicable here. In fact, the OWA operator is the only commutative Choquet integral. An additional element to be taken into account is OWA's flexibility with respect to the number of parameters (this was described in Section III). For all this, we have used OWA operators defined in terms of fuzzy quantifiers. It has to be said that Sugeno integral [30] with commutative fuzzy measures could also be used. However, these integrals are usually applied to data belonging to ordinal scales instead of numerical scales and, in this paper, we limit our approach to the case of numerical data.

According to what has been introduced here, the assumptions listed here direct this work. In the following, we assume that we want to link records that belong to two different files.

Hypothesis 1: Both files share a large set of common individuals.

Hypothesis 2: Data in both files contain, implicitly, similar structural information.

Hypothesis 3: Structural information can be expressed by means of numerical representatives for each individual.

Hypothesis 4: Aggregation operators can summarize the information of each individual.

The first hypothesis implies that reidentification is possible, as there are records to be linked because they correspond to the same individual. The second hypothesis is to say that there are similarities between different individuals that are kept *more or less* constant in both files. We call these similarities *structural information*. The third hypothesis and the fourth one are the ones that justify the use of aggregation procedures.

A. Using OWA Operators to Extract Information

To use OWA operators for information summarization, we need the settlement of OWA operators. First of all, one of the two definitions has to be selected. As previously described, due to the fact that the number of variables can be different in both files, it is appropriate to use the quantifier based definition (i.e., Definition 5). This is so because the same quantifier can be used in combination with any input vector of arbitrary dimension. Using the other approach would require the extension of an n-dimensional weighting vector into a m-dimensional one ($n \neq m$). Although methods exist to do this extension (e.g., the construction of the quantifier in [36]) it is simpler to start with the definition based on the quantifier. Moreover, the use of fuzzy quantifiers allows us to consider families of parameterized quantifiers.

TABLE III
SUMMARIZATION STRUCTURE FOR FILE A

A	P_1	P_2		P_t
R_1^A	$c_{1,1}^A$	$c_{2,1}^A$	$\cdots$	$c_{t,1}^A$
$\cdots$	$\cdots$	$\cdots$		$\cdots$
R_n^A	$c_{1,n}^A$	$c_{2,n}^A$	$\cdots$	$c_{t,n}^A$

It is a well-known fact that different quantifiers lead to different results of the OWA operator. In fact, different parameterizations correspond to different representatives of the individual. As *a priori*, it is not known which of the representatives is the best one for reidentification, we have considered a set of them. This is, we apply the OWA operator with several parameterizations (we have considered a family of fuzzy quantifiers) obtaining in this way for each file a two dimensional table that follows the one in Table III. For each individual R_i^A in a file A and for each parameterization P_j, we have the corresponding aggregated value $c_{i,j}^A$. This is, $c_{i,j}^A$ is the result of applying the OWA operator with the jth parameterization to the ith record in file A.

The same process (the same OWA operator with the same quantifiers) is applied to both files. As the number of parameterizations is the same in both files, the same structure is built for both files. Then, usual record linkage techniques can be used to link the new records (now records share the same set of variables). The application of this method and the results obtained are explained in the next section.

B. Examples

Example 2: To analyze the feasibility of our approach we have analyzed three artificial problems. These problems have been generated using the publicly available information from the UCL repository [18] we have already used in Section IV-B. This is the iris, abalone, and ionosphere data files. The selection of this data set for structure determination is based on their use of continuous variables and the fact that being public data files experiments can be reproduced.

To use this data for reidentification, two alternatives were possible: reidentification of the examples and reidentification of the variables. In the first alternative, the original file would be split in such a way that all examples but only half of the variables are present in both files. In the second alternative, the original file would be split so that all variables but only half of the examples are present in both files. We have followed the latter approach because it is not sure that half of the variables have enough information about the examples to allow for reidentification. Instead, two randomly chosen subsets of about half of the initial examples (about 175 examples in the case of ionosphere, 2000

TABLE IV
NUMBER OF REIDENTIFIED VARIABLES (IN THE IONOSPHERE FILE ONE OF THE VARIABLES IS ALWAYS ZERO AND IT WAS NOT CONSIDERED IN THE REIDENTIFICATION PROCESS)

initial file	re-identified variables	number of variables
iris	0	4
abalone	6	8
ionosphere	10	33

examples in the case of abalone, and 75 examples in the case of iris) should give enough information about the structure of the variables. In fact, subsets of these examples are usually used in machine learning [29] because they assume that these subsets have still enough information to model the variable behavior. In other words, the second approach has been used because we assume more redundancy in the examples than in the variables.

To apply the method previously described, we have considered an initial normalization step following the usual approach, i.e., translation of the initial value x in the $[\min, \max]$ interval into to the value x' in [0,1]

$$x' = \frac{(x - \min)}{(\max - \min)}$$

After normalization, the file was partitioned into two sets of approximately the same number of records (records were selected at random).

Then, for each variable in each file, the OWA operator has been applied using ten different parameterizations. Selected quantifiers are: $Q_i(x) = x^{i/5}$ for $i = 1, \ldots, 10$. In this way, for each of the initial data file we have obtained two files, with the ten representatives for each variable each. At this point, to reidentify the variables, a record linkage algorithm was applied to each pair of files. We have applied the record linkage algorithm developed by W. Winkler [43] at U.S. Census Bureau. The number of variables that were correctly reidentified are given in Table IV. In the case of the iris data file, only one link was suggested (but it was incorrect), the other variables were not considered as related.

1) Evaluation of the Results: Although the results we have obtained so far are not as good as we would like and, in the particular case of the ionosphere data file, the number of corrected links is less than a half of the total number of variables, results are quite better than they seem. To evaluate the results we consider the probability of having more than a certain number of correct links, say k, in a random permutation of n individuals. We later study this probability and we then compute the probabilities for the abalone and ionosphere data files.

Perfect reidentification when two files A and B have the same n individuals correspond to finding a particular permutation π such that, for each record i in A, $\pi(i)$ is assigned to each corresponding record j in B. This is, $\pi(i) = j$ can be understood as the linkage of individual i in file A with the individual j in file B.

Using this notation, we can compute the following.

1) The number of possible reidentifications: $n!$
2) The number of permutations such that there are exactly r elements correctly reidentified: These permutations with exactly r correct links can be built considering the following steps (we later use $k := n - r$). First, we take k elements from the correct permutation π and we permutate this k elements in such a way that there is no one that keeps its original position. Therefore, there are exactly r elements correctly reidentified (the elements not selected). To compute the cardinality of this set of permutations, we need to know that the number of sets that can be taken with k elements is: $n!/(k!(n-k)!)$. Also, that a *permutation without fixed point* (see, e.g., [25]) generated from π' is a permutation π'' in such a way that there is no element that keeps its original position (i.e., $\pi'(i) \neq \pi''$ for all i). The number of permutations of k elements without fixed point is [25]:

$$(\hat{p})(k) = k! \sum_{v=0}^{k} \frac{(-1)^k}{v!}.$$

According to this, the number of permutations such that there are exactly r elements in the correct position is

$$\frac{n! \sum_{v=0}^{k} \frac{(-1)^k}{v!}}{(n-k)!}.$$

3) The probability of finding at random a permutation with exactly r elements in the correct position

$$\frac{\sum_{v=0}^{k} \frac{(-1)^k}{v!}}{(n-k)!}.$$

In Table V, probabilities for the case of having $n = 33$ (the ionosphere case) are given. Note that the probability of obtaining ten or more correct links (the ones obtained in the aforementioned example for the ionosphere data file) is $1.013\,777\,15E-7$. Similarly, Table VI displays the probabilities for the case of $n = 8$. This is the case of the abalone data file. In this case, the probability of having more than 6 correct links (the ones obtained in this example) is $7.192\,460\,5E-4$. Table VII displays the corresponding probabilities for the iris file $(n = 4)$.

An additional aspect to be considered for the evaluation of this result is that standard techniques for record-linkage when files share a set of common variables do not lead to 100% reidentifications. In fact, [41] describes about 300 record-linkage experiments (one third using numerical data and the rest using categorical data) and the corresponding percentage of reidentifications for both probabilistic and distance-based record linkage approaches are listed. The following averages can be computed from the listings: 26.12% (for distance-based record linkage for numerical data), 19.72% (for probabilistic one for numerical data), 59.30% (for distance-based one for categorical data), and 57.93% (for probabilistic one for categorical data). In the approach presented here, we got 75% reidentifications for the abalone file and 30.30% for the ionosphere file.

According to all this, the abalone and ionosphere examples show that, although reidentification for files not sharing variables is far from perfect, the approach proposed here is appropriate because it obtains meaningful and relevant links.

TABLE V
PROBABILITIES OF HAVING r CORRECT LINKS, AND OF HAVING MORE OR EQUAL THAN r LINKS FOR 33 INDIVIDUALS

r	probability $\|links\| = r$	probability $\|links\| \geq r$
0	0.36787942	1.0
1	0.36787942	0.63212055
2	0.18393971	0.26424113
3	0.06131324	0.0803014
4	0.01532831	0.018988157
5	0.003065662	0.0036598467
6	5.109437E-4	5.941848E-4
7	7.299195E-5	8.3241146E-5
8	9.123994E-6	1.0249197E-5
9	1.0137771E-6	1.1252026E-6
10	1.01377715E-7	1.1142548E-7
11	9.216156E-9	1.0047766E-8
12	7.680129E-10	8.316107E-10
13	5.907792E-11	6.359777E-11
14	4.2198515E-12	4.5198524E-12
15	2.8132344E-13	3.0000107E-13
16	1.7582715E-14	1.8677634E-14
17	1.0342773E-15	1.0949201E-15
18	5.745985E-17	6.064281E-17
19	3.0242027E-18	3.1829554E-18
20	1.5121014E-19	1.5875276E-19
21	7.200482E-21	7.5426254E-21
22	3.2729465E-22	3.4214245E-22
23	1.4230203E-23	1.4847793E-23
24	5.929247E-25	6.1758905E-25
25	2.3717165E-26	2.4664351E-26
26	9.121372E-28	9.471847E-28
27	3.3801082E-29	3.5047438E-29
28	1.202626E-30	1.246356E-30
29	4.241236E-32	4.3729944E-32
30	1.2566625E-33	1.317584E-33
31	6.080625E-35	6.092142E-35
32	0	1.1516336E-37
33	1.1516335E-37	1.1516336E-37

VI. CONCLUSION

In this paper, we have considered the use of aggregation operators in data mining. We have considered two different uses. First, we have shown its application in the process of building data models. We have argued that some of the learning techniques are applicable to large databases because the cost is proportional to the number of variables. An example that considers 33 variables has been given. Second, we have shown that aggregation operators are useful for extracting implicit information from raw data and we have applied them to a reidentification problem. We have shown that OWA operators are suitable in this task. We have compared our approach with random selection of linked pairs and with the results usual in reidentification using standard approaches. In both comparisons, our method is well rated and, thus, it shows that the method is suitable for reidentification for files not sharing variables.

TABLE VI
PROBABILITIES OF HAVING r CORRECT LINKS, AND OF HAVING MORE OR EQUAL THAN r LINKS FOR 10 INDIVIDUALS

r	probability $\|links\| = r$	probability $\|links\| \geq r$
0	0.36788195	1.0
1	0.36785713	0.63211805
2	0.18402778	0.26426092
3	0.06111111	0.080233134
4	0.015625	0.019122023
5	0.0027777778	0.0034970238
6	6.9444446E-4	7.1924605E-4
7	0	2.4801588E-5
8	2.4801588E-5	2.4801588E-5

TABLE VII
PROBABILITIES OF HAVING r CORRECT LINKS, AND OF HAVING MORE OR EQUAL THAN r LINKS FOR 4 INDIVIDUALS

r	probability $\|links\| = r$	probability $\|links\| \geq r$
0	0.375	1.0
1	0.333333	0.625
2	0.25	0.291666
3	0.0	0.041666
4	0.041666	0.041666

REFERENCES

[1] J. Bacher, R. Brand, and S. Bender, "Re-identifying register data by survey data using cluster analysis: an empirical study," *Int. J. Uncertainty, Fuzziness Knowledge-Based Syst.*, vol. 10, no. 5, pp. 589–608, 2002.

[2] E. Bauer and R. Kohavi, "An empiricial comparison of voting classification algorithms: bagging, boosting and variants," *Mach. Learn.*, vol. 36, pp. 105–139, 1999.

[3] G. Beliakov, "How to build aggregation operators from data," *Int. J. Intell. Syst.*, vol. 18, pp. 903–923, 2003.

[4] T. Calvo, G. Mayor, and R. Mesiar, Eds., *Aggregation Operators: New Trends and Applications*. Heidelberg, Germany: Physica-Verlag, 2002.

[5] M. Carbonell, M. Mas, and G. Mayor, "On a class of monotonic extended OWA operators," in *Proc. 6th IEEE Int. Conf. Fuzzy Systems (IEEE-FUZZ'97)*, Barcelona, Spain, 1997, pp. 1695–1699.

[6] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Royal Statist. Soc.*, vol. 39, pp. 1–38, 1977.

[7] J. Domingo-Ferrer and V. Torra, "On the connections between statistical disclosure control for microdata and some artificial intelligence tools," *Inform. Sci.*, vol. 153, pp. 153–170, 2003.

[8] ——, "Disclosure risk assessment in statistical microdata protection via advanced record linkage," *Statist. Comput.*, vol. 13, pp. 343–354, 2003.

[9] D. P. Filev and R. R. Yager, "Learning OWA operator weights from data," in *Proc. 3rd IEEE Conf. Fuzzy Syst.*, Orlando, FL, 1994, pp. 468–473.

[10] ——, "On the issue of obtaining OWA operator weights," *Fuzzy Sets Syst.*, vol. 94, pp. 157–169, 1998.

[11] L. Gill, "Methods for automatic record matching and linking and their use in national statistics," Office for National Statistics, London, U.K., 2001.

[12] M. Grabisch, H. T. Nguyen, and E. A. Walker, *Fundamentals of Uncertainty Calculi With Applications to Fuzzy Inference*. Dordrecht, The Netherlands: Kluwer, 1995.
[13] M. Grabisch, "Modeling data by the Choquet integral," in *Information Fusion in Data Mining*, V. Torra, Ed. New York: Springer-Verlag, 2003, pp. 135–148.
[14] Integrity Online., Jackson, MS. [Online]http://www.integrity.com
[15] M. A. Jaro, "Advances in record-linkage methodology as applied to matching the 1985 Census of Tampa, Florida," *J. Amer. Statist. Assoc.*, vol. 84, no. 406, pp. 414–420, 1989.
[16] C. J. Merz, "Using correspondence analysis to combine classifiers," *Mach. Learn.*, vol. 36, pp. 33–58, 1999.
[17] C. J. Merz and M. J. Pazzani, "Combining regression estimates," *Mach. Learn.*, vol. 36, pp. 9–32, 1999.
[18] P. M. Murphy and D. W. Aha. (1994) UCI repository machine learning databases. Dept. Inform. Comput. Sci., Univ. California, Irvine, CA. [Online]http://www.ics.uci.edu/mlearn/MLRepository.html
[19] M. Sugeno and T. Murofushi, *Fuzzy Measures*. Tokyo, Japan: Nikkan Kogyo Shinbusha, 1993.
[20] D. Nettleton and J. Muniz, "Processing and representation of meta-data for sleep apnea diagnosis with an artificial intelligence approach," *Int. J. Med. Inform.*, vol. 63, pp. 77–89, 2001.
[21] H. B. Newcombe, J. M. Kennedy, S. J. Axford, and A. P. James, "Automatic linkage of vital records," *Sci.*, vol. 130, pp. 954–959, 1959.
[22] M. O'Hagan, "Aggregating template or rule antecedents in real-time expert systems with fuzzy set logic," in *Proc. 22nd Annu. IEEE Asilomar Conf. Signals, Systems, Computers*, Pacific Grove, CA, 1988, pp. 681–689.
[23] D. Pagliuca and G. Seri, "Some Results of individual ranking method on the system of enterprise accounts annual survey," Enterprise, Esprit SDC Project, Deliverable MI-3/D2, 1999.
[24] E. Rahm and H. H. Do, "Data cleaning: problems and current approaches," *Bull. IEEE Comput. Soc. Tech. Comm. Data Eng.*, vol. 23, no. 4, pp. 3–13, 2000.
[25] F. Reinhard and H. Soeder, *Atlas des mathématiques*. Paris, France: Librairie Générale Française, 1997.
[26] J. F. Robinson-Cox, "A record-linkage approach to imputation of missing data: analyzing tag retention in a tag-recapture experiment," *J. Agricultural, Biol., Environ. Statist.*, vol. 3, no. 1, pp. 48–61, 1998.
[27] R. A. Ribeiro, "Fuzzy multiple attribute decision making: a review and new preference elicitation techniques," *Fuzzy Sets Syst.*, vol. 78, no. 2, pp. 155–181, 1996.
[28] T. L. Saaty, *The Analytic Hierarchy Process*. New York: McGraw-Hill, 1980.
[29] V. G. Sigillito, S. P. Wing, L. V. Hutton, and K. B. Baker, "Classification of radar returns from the ionosphere using neural networks," in *John Hopkins APL Tech. Dig.*, vol. 10, 1989, pp. 262–266.
[30] M. Sugeno, "Theory of fuzzy integrals and its applications," Ph.D. dissertation, Tokyo Inst. Technol., Tokyo, Japan, 1974.
[31] A. Tanaka and T. Murofushi, "A learning model using fuzzy measures and the Choquet integral," in *Proc. 5th Fuzzy System Symp.*, Kobe, Japan, 1989, pp. 213–218.
[32] V. Torra, "The weighted OWA operator," *Int. J. Intell. Syst.*, vol. 12, no. 2, pp. 153–166, 1997.
[33] ——, "On the learning of weights in some aggregation operators: the weighted mean and OWA operators," *Mathware Soft Comput.*, vol. 6, pp. 249–265, 1999.
[34] ——, "On some relationships between hierarchies of quasiarithmetic means and neural networks," *Int. J. Intell. Syst.*, vol. 14, no. 11, pp. 1089–1098, 1999.
[35] ——, "Toward the reidentification of individuals in data files with non-common variables," in *Proc. Eur. Conf. Artificial Intelligence (ECAI 2000)*, Berlin, Germany, 2000, pp. 326–330.
[36] ——, "The WOWA operator and the interpolation function W^*: Chen and Otto's interpolation method revisited," *Fuzzy Sets Syst.*, vol. 113, no. 3, pp. 389–396, 2000.
[37] ——, "Re-identifying individuals using OWA operators," presented at the 6th Int. Conf. Soft Computing, Fukuoka, Japan, 2000.
[38] ——, "Learning weights for weighted OWA operators," presented at the IEEE Int. Conf. Industrial Electronics, Control and Instrumentation (IECON 2000), Nagoya, Japan, 2000.
[39] ——, "Learning weights for the quasiweighted mean," *IEEE Trans. Fuzzy Syst.*, vol. 10, pp. 653–666, Oct. 2002.
[40] ——, *Information Fusion in Data Mining*. New York: Springer-Verlag, 2003.
[41] V. Torra and J. Domingo-Ferrer, "Record linkage methods for multidatabase data mining," in *Information Fusion in Data Mining*, V. Torra, Ed. New York: Springer-Verlag, 2003, pp. 101–132.
[42] K. Tumer and J. Ghosh, "Classifier combining: analytical results and implications," presented at the Working Notes From the Workshop Integrating Multiple Learned Models, National Conf. Artificial Intelligence, Portland, Oregon, Aug. 1996.
[43] U. S. Bureau of the Census, "Record linkage software: User documentation,", 2000.
[44] W. E. Winkler, "Advanced methods for record linkage," in *Proc. Section on Survey Research Methods*, 1995, pp. 467–472.
[45] ——, "Matching and record linkage," in *Business Survey Methods*, B. G. Cox, Ed. New York: Wiley, 1995, pp. 355–384.
[46] Z. S. Xu and Q. L. Da, "An overview of operators for aggregating information," *Int. J. Intell. Syst.*, vol. 18, pp. 953–969, 2003.
[47] R. R. Yager, "On ordered weighted averaging aggregation operators in multi-criteria decision making," *IEEE Trans. Syst. Man Cybern.*, vol. 18, pp. 183–190, Jan. 1988.
[48] ——, "Families of OWA operators," *Fuzzy Sets Syst.*, vol. 59, pp. 125–148, 1993.
[49] ——, "Quantifier guided aggregation using OWA operators," *Int. J. Intell. Syst.*, vol. 11, pp. 49–73, 1996.
[50] N. Zhong, Y. Yao, and S. Ohsuga, "Peculiarity oriented multidatabase mining," in *Principles of Data Mining and Knowledge Discovery*, J. Zytkow and J. Rauch, Eds. New York: Springer-Verlag, 1999, vol. 1704, Lecture Notes in Artificial Intelligence, pp. 136–146.

Vicenç Torra (M'96–SM'03) received the B.Sc., M.Sc., and Ph.D. degrees in computer science from the Universitat Politècnica de Catalunya, Catalonia, Spain, 1991, 1992, and 1994, respectively.

He has been an Associate Professor at the Universitat Rovira i Virgili and is currently an Associate Professor (research track) at the Institut d'Investigació en Intel·ligència Artificial (IIIA-CSIC), Bellaterra, Catalonia, Spain. His research interests include information fusion, privacy preserving data mining, and information retrieval. He has coauthored over 80 technical publications. He has been a Guest Editor of the *International Journal of Intelligent Systems* and the *International Journal of Intelligent, Fuzzy, and Knowledge-Based Systems*, and has edited a book on information fusion in data mining. He has written an undergraduate course book on artificial intelligence.

Dr. Torra was a Founding Member of the IEEE Spanish Chapter on Information Theory and of the Catalan Association for Artificial Intelligence (Member of the Board, 1996–2000). Since September 2001, he has been a Member of the Board of the European Society for Fuzzy Logic and Technology. He organized the 1st Catalan Conference on Artificial Intelligence (1998), and is currently co-organizing the Privacy in Statistical Databases (PSD'2004) and Modeling Decisions for Artificial Intelligence (MDAI'2004) conferences.