Available at www.ComputerScienceWeb.com
POWERED BY SCIENCE DIRECT®

ELSEVIER

Information Sciences 151 (2003) 153–170

INFORMATION SCIENCES
AN INTERNATIONAL JOURNAL

www.elsevier.com/locate/ins

On the connections between statistical disclosure control for microdata and some artificial intelligence tools

Josep Domingo-Ferrer [a], Vicenç Torra [b,*]

[a] *Dept. Enginyeria Informàtica i Matemàtiques (ETSE), Universitat Rovira i Virgili, Av. Països Catalans 26, 43007 Tarragona (Catalonia), Spain*
[b] *Institut d'Investigació en Intel·ligència Artificial—CSIC, Campus UAB s/n, 08193 Bellaterra (Catalonia), Spain*

Received 1 January 2002; received in revised form 8 October 2002; accepted 18 November 2002

Abstract

Statistical disclosure control (SDC) and artificial intelligence (AI) use similar tools for different purposes. This work describes the common elements of both areas to increase their synergy.

SDC is a discipline that seeks to modify statistical data so that they can be published (typically by National Statistical Offices) without giving away the identity of any individual behind the data. When dealing with individual data (microdata in SDC jargon), both SDC procedures and AI knowledge integration procedures use similar principles for different purposes (masking data vs. improving its quality). Similarities can also be found for methods evaluating re-identification risk in SDC and data mining tools for making data consistent.

This paper explores those methodological connections with the aim of stimulating interaction between both fields. In particular, data mining turns out to be a common interest of both fields.

Keywords: Artificial intelligence; Data mining; Re-identification procedures; Statistical disclosure control; Synthesis of information; Official statistics; Data cleaning

* Corresponding author. Tel.: +34-93580-9570; fax: +34-93580-9661.
E-mail addresses: jdomingo@etse.urv.es (J. Domingo-Ferrer), vtorra@iiia.csic.es (V. Torra).
URLs: http://www.etse.urv.es/~jdomingo, http://www.iiia.csic.es/~vtorra.

0020-0255/03/$ - see front matter
doi:10.1016/S0020-0255(03)00064-1

1. Introduction

This work tries to highlight and characterize some relationships between the way statistical disclosure control (SDC) and artificial intelligence (AI) deal with individual data. We show that, although the goals in both areas are completely different, similar techniques are already being applied. The objective of this paper is to stimulate interaction and increase synergy between researchers in both fields.

The production of official statistics by National Statistical Offices (NSOs) can be regarded as a process with three main steps: data collection, data processing and data dissemination. The last step is essential to justify the resources spent and the large amount of information being collected on individuals and organizations. Thus data dissemination should preserve the informational content of collected and processed data as much as possible whilst guaranteeing that particular individuals cannot be re-identified (*disclosure control problem*). If NSOs fail to protect the disseminated data against re-identification, individual respondents will probably complain and/or refuse to collaborate in future data collections. The usual approach to protecting the released data is to distort them in some way before publication. The distortion should be small enough to preserve data utility, but it should be sufficient to prevent confidential information about an individual from being deduced or estimated from the released data. Equivalently, both the information loss and the disclosure risk associated to the released data should be kept small. The methods that attempt to perform such a nontrivial distortion are collectively known as statistical disclosure control methods (or SDC methods for short, [15,58]).

NSOs release two kinds of data through their statistical databases: *tabular data* (tables with cells containing aggregated data) and *microdata sets* (sets of records, each containing information about an individual entity such as a person, household, business, etc.). While there is a long experience in table dissemination, microdata dissemination is a much more recent activity (first attempts in the late 80s). SDC methods for microdata are usually known as *masking methods* and are currently a hot research topic. From what has been said above, masking methods inherit the difficult goal of achieving a balance between information loss and disclosure risk. Up to now, several masking algorithms have been proposed, some of which are analogous to methods used in the AI framework, and more specifically in the areas of machine learning and knowledge acquisition.

In machine learning, models are built from a set of examples. A typical case is to have a set of n examples with $m+1$ attributes (or variables) each. The general approach is to learn the behaviour of the $m+1$th attribute once the values of the other m attributes are already known. A common goal in this setting is to design error-resilient procedures (e.g., see Chapter 5 in [39]) be-

cause only algorithms with such characteristics can be applied to real environments. The interesting point is that examples with a certain number of errors can be viewed as "distorted" examples stemming from a single (nonexistent) correct example. Some methods that are successful in the case of errors in data are the *ensemble of classifiers*. This corresponds to building a model from the combination of several other, say, "partial" models. This approach leads to good results with respect to the trade-off bias/variance (see [32] for a particular method and [29] for a review of the approach). A process similar in nature to the ensemble of classifiers but different as to the procedure can be found in knowledge acquisition.

Knowledge acquisition for knowledge-based systems (KBS) is the process of modeling when the model is built by a domain expert [21]. In this process, the knowledge engineer and the expert are involved in the development of a model based on the expertise of the latter. To aid in this task, several algorithms and methodologies have been developed that infer concept descriptions from a given set of training examples. Now, if sets of examples come from several experts rather than from a single one, knowledge integration techniques are required to combine all the information into a single KBS (see [42] for a detailed description of the advantages and disadvantages of considering more than one expert). To do so, the knowledge of all experts is assumed (implicitly or explicitly) to be similar. In this case, differences in descriptions of the same examples can be regarded as distortions caused by different points of view. The different cases a knowledge engineer or an automatic system has to deal with are analyzed in [24].

As shown above, machine learning and knowledge acquisition share with SDC techniques the notion of data distortion. However, while this distortion is intentional in the case of SDC techniques, it is accidental in AI. Thus, while SDC methods are designed to cause distortion, methods in AI try to overcome it.

An additional topic that is present in both areas is re-identification and record linkage (*data cleaning* following data base jargon). Re-identification happens when information from the same individual appearing in different files (for the sake of simplicity we only consider pairs of files) is identified. Record linkage attempts re-identification by linking records corresponding to the same individual.

Re-identification is used by National Statistical Offices to evaluate to what extent a masking method is enough to protect the data. After the masking process, masked data are compared with original data and re-identification procedures are applied to find out whether, given a masked record, the corresponding original one can be identified.

On the other hand, in the fields of data bases and data mining, re-identification is used to make databases consistent. It is a common problem that information is not centralized but distributed among several users. In this case, it

is often the case that the information on a particular individual (supplier, client, ...) is not expressed in a unified way. For example, names of individuals (or companies) or addresses are encoded using different abbreviations or transcribed with some errors. In this case, it is difficult to find all the records that relate to the same individual and, therefore, queries to the system or the knowledge inferred from the database will be incorrect. To relate the records of the files that correspond to the same individuals and to make the files consistent, re-identification techniques are applied.

Therefore, while re-identification is a requirement when building applications on distributed databases, it is a threat for respondent confidentiality in the context of official statistics.

In this paper, we describe some of the methods used in SDC to cause distortion in data and the types of distortion present in databases used for machine learning. We also describe the use of aggregation operators both in SDC and in AI. The interest of this comparison is twofold. On one hand, awareness of methodological similarities strengthens the possibility of synergy between SDC and AI. On the other hand, we show that, in some cases, SDC techniques and AI techniques are inverse.

In Section 2, the use of data distortion both in SDC and AI is discussed and compared. In Section 3, the roles of data aggregation in both disciplines are analyzed. Section 4 is a conclusion.

2. Data distortion

Both in the SDC and the AI contexts, information is expressed using a common structure and differences lie only in the jargon in use. Information is represented as a two-dimensional database where one dimension corresponds to the set of objects (or elements, individuals, persons) and the other is the set of attributes (or variables). The database contains a value for each pair (object, attribute), so that it can be modelled as a function

$$V : \mathbf{O} \rightarrow D(A_1) \times D(A_2) \times \cdots \times D(A_m)$$

where $\mathbf{O}$ denotes the set of objects, $A_1, A_2, \ldots A_m$ are the attributes and $D(A_i)$ denotes the domain of attribute A_i.

In SDC, the goal is to supply the user with a database V' similar to V (i.e., with low information loss) in such a way that

(1) disclosure risk (i.e., risk of re-identification) is low.
(2) The results of user analyses (e.g. regressions, means, etc.) on V' should be "similar" to the results that would be obtained on V.

In machine learning techniques for real applications, only the existence of the database V' is assumed. Equivalently, only a database subject to errors is known.

In knowledge acquisition with a single expert, we have a case analogous to machine learning: a database V' is known which is an approximation of the (unknown) ideal database V. In knowledge acquisition with a set **E** of experts, a set $\mathbf{V}'_{\mathbf{E}} = \{V'_e\}_{e\in\mathbf{E}}$ of databases is available, where V'_e is the database resulting from expert e. Each V'_e is an approximation of the ideal V. Several techniques have been developed to build a single KBS from a set of databases, so that the resulting KBS approximates the one that would be obtained from the ideal database V. For example, the technique [54] builds a single V' from all V'_e coming from $e \in \mathbf{E}$ and this V' approximates V.

Although in machine learning and knowledge acquisition errors are assumed to be accidental, error injection (i.e. data distortion) techniques are also being used for testing purposes:

- Machine learning techniques are often extensively tested under increasing levels of noise to prove their usefulness and determine their robustness for real problems [45]. To do so, several fictitious domains exist that enable a predefined amount of data to be generated for a given percentage of noise: LED [8], Parity [37], SGP/2 [4]. As a result, several databases $V'_1, V'_2, \ldots$ with an increasing percentage of errors are obtained from an initial database V. Data in V'_i are distorted and the distortion (and consequently the loss of information) increases with i.
- Techniques that use information from several experts can be tested for robustness in a similar way. In [56], a methodology for data distortion was introduced which generates several data sets from a single one by applying different distortion methods.

In Subsection 2.1, the main distortion techniques used in SDC are recalled. In Subsection 2.2, distortion techniques that have been used in AI are listed. In Subsection 2.3, the pros and cons of data mining (understood as model building from distorted data) for SDC are explored. Subsection 2.4 highlights some synergies and analogies.

2.1. Intentional distortion in SDC

There is a wide range of masking methods for microdata. See [1,15,58] for details and a comprehensive survey. We next outline some of them, so that similarities with AI distortion can be appreciated:

Additive noise: If the original data are X, the masked data Y are computed as

$$Y = X + \epsilon$$

where ϵ is independent noise with the same covariance as X (see [6,36,50] for details). With this method, means and covariances can be preserved, but confidentiality may not be satisfactory [23].

Global recoding: Several categories of an attribute are combined to form new (less specific) categories (see e.g. [47,49]).

Local suppression: Certain values of individual attributes are suppressed with the aim of increasing the set of records agreeing on a combination of key values. Ways to combine local suppression and global recording are discussed in [10] and implemented in the μ-Argus SDC package [31].

Microaggregation: Original microdata are grouped into small aggregates or groups. The average over each group is published instead of the original individual values [2,9]. Means are preserved and, if data are sorted using multivariate criteria before forming groups and groups have variable size, the impact on correlations between attributes and the first principal component can be fairly moderate [40]. See [13] for details of microaggregation for quantitative attributes and [14] for a review of results for both quantitative and qualitative attributes.

Resampling: A bootstrap sample $z'_1, \ldots, z'_n$ is obtained by drawing from the original (qualitative) microdata $z_1, \ldots, z_n$, n times and with replacement. It can be shown that the frequencies in the bootstrap sample are expected to be those in the original microdata. See [12] for a survey on resampling methods.

PRAM: The post-randomization method (PRAM) [38] is a probabilistic, perturbative method for disclosure protection of qualitative attributes in microdata files. In the masked file Y, the scores on some qualitative attributes for certain records in the original file X are changed to a different score according to a prescribed probability mechanism, namely a Markov matrix. The Markov approach makes PRAM very general, because it encompasses data perturbation, data suppression and data recoding. PRAM information loss and disclosure risk largely depend on the choice of the Markov matrix and are still (open) research topics [27].

Multiple imputation: This method [46] relies on releasing simulated microdata created by multiple imputation techniques based on the original microdata. A way to perform multiple imputation is on an attribute-by-attribute basis, using a randomized regression (with normal errors) to impute missing values of each quantitative attribute [35].

Camouflage: Vector camouflage [26] is a method for giving unlimited, correct numerical responses to ad hoc queries to a database while not compromising confidential numerical data. No probabilistic assumptions are made and optimization techniques are used to camouflage the sensitive data vector (exact answer) in an infinite set of vectors, thus providing an interval answer. The information loss is the transformation of a point answer into an interval answer.

General methods: The concept of *matrix masking* is a generalization that encompasses a number of methods discussed above (additive noise, microaggregation, PRAM, etc.). If the original data are X, the matrix-masked data Y are computed as

$$Y = AXB + C$$

Table 1
Original records for Example 1

Illness	···	Sex	Marital status	Town	Age
Heart	···	M	Married	Barcelona	33
Pregnancy	···	F	Divorced	Tarragona	40
Pregnancy	···	F	Married	Barcelona	36
Appendicitis	···	M	Single	Barcelona	36
Fracture	···	M	Single	Barcelona	33
Fracture	···	M	Widow	Barcelona	81

Table 2
Masked records for Example 1

Illness	···	Sex	Marital Status	Town	Age
Heart	···	M	Married	Barcelona	33
Pregnancy	···	F	Widow/er-or-divorced	–	40
Pregnancy	···	F	Married	–	33
Appendicitis	···	M	Single	Barcelona	40
Fracture	···	M	Single	Barcelona	36
Fracture	···	M	Widow/er-or-divorced	Barcelona	81

where A is a record-transforming mask (i.e. object-transforming), B is an attribute-transforming mask and C is a displacing mask (noise) [19].

Example 1. Tables 1 and 2 illustrate the application of masking methods. We consider first, in Table 1, a set of original (*unmasked*) records defined over the variables "Illness", "Sex", "Marital Status", "Town" and "Age". Then, in Table 2, the same records are displayed. The following masking methods are considered for building 2: local suppression (for variable "Town"), global recoding (for variable "Marital Status", values "widow/er" and "divorced" are recoded as "widow/er-or-divorced") and rank swapping (for variable "Age").

2.2. Accidental distortion in machine learning and knowledge acquisition

Machine learning techniques usually consider the following types of accidental data distortion:

Intrinsic noise: Data coming from the real world usually convey intrinsic errors. The exact value is usually not known but an approximate value is available instead. Models are built with the available data.

Missing values: Some of the values in the database are missing. Two possibilities are considered: attribute not applicable to the object, or unknown value for a certain example. Different ways for considering missing values are reviewed in [28,43] (for classification using decision trees) and [22] (for EM algorithms).

Low granularity: Values are given in a coarse scale. For example, we could get ordinal temperatures (i.e. values "very cold", "cold", "cool", etc.) instead of numerical temperatures.

Input errors: Filling the database inevitably introduces some errors in the original data (e.g. divergences among *Iris* databases used by several different research groups are reported in [5]). These "typing" errors cannot be distinguished from intrinsic noise.

Example 2. A machine learning system applied to the data in Table 2 for building a model for the variable "Illness" will be faced with these types of data distortion: low granularity (due to global recoding in variable "Marital status"), input errors (due to rank swapping in variable "Age") and missing values (due to local suppression in variable "Town").

In addition to the above kinds of accidental data distortion, intentional distortion is sometimes used to reduce the granularity of the information. This is the case of discretization (or quantization) of continuous attributes. The reason for discretizing is that some machine learning algorithms cannot accept continuous attributes, because they have been designed for symbolic domains, where the attributes are defined by means of linguistic attributes. There is a broad literature on discretization methods (some of which are reviewed in [57]).

2.3. Overcoming distortion: re-identification and data mining

Data mining [30] is the search for relationships and global patterns that exist in large databases, but are "hidden" among the vast amounts of data, such as the relationship between patient data and their medical diagnosis. These relationships represent valuable knowledge about the individuals in the database and, if data have been well collected, of the real world recorded by the database. Data mining systems use machine learning techniques to create models that simplify but help to understand a given environment.

Two data mining techniques for overcoming data distortion are:

Re-identification: As pointed out in the introduction, re-identification happens when data on the same individual but from different data files are successfully linked. Record linkage is a re-identification mechanism used to link records that correspond to the same individual. These techniques are currently applied by NSO and by companies. The typical application in both cases is when the files share a set of common variables. In this case, difficulties in the re-identification appear due to the presence of distorted data. See [44] for a review on the problems these methods have to face and [16] for the main approaches considered in the literature. In particular, re-identification procedures based on probability distributions (*probabilistic record linkage*) and based on similarity functions (*distance based record linkage*) have been developed. See [20,59] for a

review and a state of the art (including a formal description) of probabilistic record linkage.

Recent developments in the field of data mining consider re-identification for non-common variables. This is of interest when considering data files sharing a set of individuals that describe similar information (e.g., correlated variables). Re-identification for non-common variables is based on the existence of some relationships between individuals that are kept across files. These relationships imply some underlying structures in both files that (i) can be obtained through the manipulation of the data and (ii) can be, afterwards, related through a re-identification mechanism. For example, in [7,53], the underlying structure considered was a set of partitions of the individuals. These relationships (i.e., partitions) were established using clustering algorithms. After applying the same clustering algorithms to both files, re-identification procedures linked the clusters obtained for one file with the clusters obtained for the other file. A different approach was described in [52] based on building different prototypes for each individual.

Model building: A typical application of data mining techniques is to obtain the model for a given variable in terms of the others using some machine learning techniques.

Example 3 illustrates data mining uses for re-identification.

Example 3. Tables 3 and 4 display records from two different data files that contain information from the same individuals. However, re-identification is not straightforward due to the presence of different variables and the existence of errors in the variables. The availability of models on the variables would help in the re-identification process. For example, it is conceivable to think of a model relating variables "Name" and "Sex", and of a model relating variables "Birthday" and "Age". Such models could either be supplied by the user or extracted from data.

Table 3
Records for Example 3 corresponding to file A

Illness	···	Room number	Sex	Marital status	Town	Age
Heart	···	201	M	Married	Barcelona	32
Pregnancy	···	305	F	Divorce	Tarragona	39
Pregnancy	···	309	F	Married	Barcelona	35
Appendicitis	···	215	M	Single	Barcelona	35
Fracture	···	412	M	Single	Barcelona	32
Fracture	···	412	M	Widow	Barcelona	80
···	···					
···	···					

Table 4
Records for Example 3 corresponding toc file B

Name	Surname	Address	Town	Birthday
David		c/ Antoni Gaudi	Barcelona	5/Aug/1970
Anna		c/ Ausias March	Tarragona	15/Apr/1963
Anna		c/ Pau Casals	Barcelona	9/Oct/1967
Felip		c/ Casanovas	Barcelona	6/Mar/1967
Ricard		c/ Joan Miro	Barcelona	4/Feb/1970
Rafel		c/ Joan Miro	Barcelona	23/Apr/1922
...				
...				

Data mining can be both a threat or a help to statistical confidentiality, depending on how and by whom it is used [11]:

- With statistical databases, it is straightforward to see that data mining techniques threaten statistical confidentiality. This is the case when using the models on the variables in Example 3. Another situation corresponds to the case when a model is discovered that specifies a relationship between a sensitive attribute (e.g. "Illness") and some seemingly innocuous attributes (e.g. "Room Number" in a hospital). Then, even if the sensitive attribute is disclosure-protected (e.g. the variable "Illness" is removed from Table 3), its original values can be inferred using the model constructed by the data mining system. Re-identification techniques that link original data and masked data can also result in undesired disclosure.
- Data mining can also be beneficial if data mining tools are routinely included in SDC software packages. In this way, protection of a data set could be done in two stages:
  (1) The data protector first protects the original data set using the statistical disclosure control techniques offered by the package.
  (2) The data protector uses the data mining tool on both the original and the "protected" data set. The goal is to find models that yield the original values of sensitive attributes in terms of the attributes in the protected data set. If data mining yields good approximations to, say, suppressed cell values, then the data protector should return to stage 1.

2.4. Interim discussion: distortion synergy

It is not difficult to see that there are strong analogies between the distortion algorithms used in SDC and AI. A few connections are next listed:

- The intrinsic noise and the input error considered in machine learning can be actually modelled as additive noise as described in Section 2.1.
- The phenomenon of missing values in machine learning is the accidental version of local suppression in SDC.

- Low granularity and discretization error inherent to the use of some machine learning algorithms are in fact analogous to the global recoding principle used for SDC.

The above methodological connections show that SDC can take quite a few lessons from advances in AI, and conversely.

In SDC, it is important to know that AI methods exist which try to overcome data distortion. Also, new developments on re-identification techniques cause new threats to SDC. In particular, Section 2.3 justifies why data mining methods should be considered by SDC developers.

As mentioned in Section 1, knowledge integration techniques combine (distorted) descriptions from several experts to obtain a single (ideally undistorted) description. Thus, we can state the following properties:

Property 1. *If there is a knowledge integration technique that can reconstruct an original data set out of n different distorted versions of the data set, then statistical confidentiality is compromised if more than n different SDC-protected versions of the same confidential data set are released.*

Property 2. *Information loss in SDC is inversely proportional to the reconstruction capabilities of knowledge integration and re-identification techniques. Disclosure risk is proportional to these reconstruction capabilities.*

An additional use of AI for SDC is related to discretization methods. These methods map continuous attributes onto a set of labels. In fact, labelling methods exist which give semantics to labels by generating intervals or fuzzy sets for each label value. The same idea could be applied when using global recoding for SDC; the new (less granular) values could be associated a semantics in the form of fuzzy intervals defined on the old (more granular) scale.

AI can in turn benefit from SDC by being aware that a very rich array of distortion techniques are available from SDC. As a consequence, new possibilities exist for developing model building techniques in the presence of these new distortions. These techniques have to take into account how distortion is produced by Statistical Offices (i.e., distortion introduced to files is not random) in order to build a model from distorted data which is as similar as possible to the one that would be derived from the original data.

3. Aggregation

Together with distortion, aggregation is another principle that is present both in SDC and AI. Just like for distortion, the role of aggregation is different in each of both fields.

Section 3.1 deals with the use of aggregation to hide information in SDC. Section 3.2 explains how aggregation can be a tool for gaining information in the AI context. In Section 3.3, some synergies are identified.

3.1. SDC: aggregation to hide information

In SDC, aggregation is actually used as a distortion technique. This is the case of microaggregation for quantitative attributes described in Section 2.1. Original microdata are grouped into small aggregates or groups. The average over each group is published instead of the original individual values. The result is that the published microdata set is a distorted version of the original one.

The same aggregation principle is the basis of the SDC methods for protecting tabular data, which are not dealt with in this paper. The idea is that each published cell in a statistical table should correspond to an aggregate of individuals (objects); there should be at least k individuals contributing to the cell value, and no individual should dominate (e.g. contribute too much to) the cell value. In this way, publication of tables of aggregate values does not lead to straightforward disclosure of individual information.

3.2. AI: aggregation to gain information

The use of aggregation techniques in AI is radically different. Their purpose is to *increase* the knowledge of a system, and they are mainly used in two situations: decision making and domain representation.

Decision making: When a decision is to be made by the system, it may happen that there are several decision criteria rather than a single one. This situation, known as the multicriteria decision making problem, is usually solved in two stages:

1. *Aggregation*. For each alternative, the degrees of satisfaction of the various criteria are aggregated.
2. *Ranking*. The alternatives are ranked with respect to the aggregated degree of satisfaction.

Domain representation: To obtain a good representation of an environment, a knowledge acquisition system needs to gain knowledge which is both reliable and comprehensive (encompassing the whole domain or environment). If a single information source (expert, sensor, etc.) is used, reliability may be poor or the view of the domain may be too narrow. A better strategy is to combine the information provided by several sources; doing so leads to higher reliability and accuracy and may provide the system with perceptions which would not be obtainable from a single sensor or expert.

Combination functions used in AI aggregation are defined according to the objects to be synthesized. Consequently, they depend on the formalism of the

knowledge representation. In fact, as pointed out by [25], this dependence causes the selection of a combination function to face the problem of choosing an adequate formalism in addition to the intrinsic problem of deciding which are the properties that should be satisfied by the function. Thus, there exist functions for several formalisms, e.g. for combining probability distributions [25], mass functions, fuzzy sets [17], data matrices [3], rules, preference relations (both quantitative [33] and qualitative [34]) or classifications (in several flavors, such as dendrograms [55] or partitions [41]).

For each formalism, several functions have been defined, each one having proper properties for certain applications.

For example, when the objects to be synthesized are numerical values (e.g. numbers in the [0,1] interval), two classical aggregation functions are considered: the arithmetic mean and the weighted mean. A few years ago, Yager [60] defined the OWA operator, which is an alternative combination function for synthesizing numerical values.

Both the weighted mean and the OWA operator combine values according to a set of weights. The main difference is the meaning that the set of weights have in each function. On one hand, the weighted mean allows an aggregated value to be computed by the system from values corresponding to several sources while taking into account the reliability of each information source (assuming that the weight attached to each source measures its reliability). Alternatively, the OWA operator allows weighting the values according to their ordering. Thus, a system using OWA can give more importance to a subset of the input values than to another subset. For instance, the influence of extreme values on the result can be diminished, thus increasing the influence of central values. Furthermore the OWA operator is commutative, and yields the same result on any permutation of the arguments.

We can rephrase the above paragraph in a slightly different manner so that a generalization of OWA will naturally follow. On one hand, weights in the weighted mean measure the importance of an information source regardless of the value that the source has captured. On the other hand, weights in the OWA measure the intrinsic importance of the value (with respect to the other ones) regardless of the information source that captured it. A third operator called WOWA [51] has been defined to combine the advantages of the weighted mean and the OWA operator: the idea is to allow the user to weight the reliability of the information source, as the weighted mean does, and the values with respect to their relative position, as OWA does.

3.3. Interim discussion: aggregation synergies

In Section 2, it was argued that distortion techniques were more developed in SDC than in AI. It is the opposite for aggregation. Techniques like OWA and WOWA could be successfully applied to SDC, and more generally to

official statistics. In the latter disciplines, extreme values are normally a problem, because they either lead to disclosure or at least have an undesirable impact on averages. Therefore, it would make sense to use OWA or WOWA in such a way that the weight of extreme values is very low.

Another strong point of AI aggregation is that it can deal with qualitative data. This is, for example, the case of the Sugeno integral [48] where only ordinal scales are required. Thus, one could envision extending microaggregation to protect qualitative attributes against disclosure.

On the other hand, the application of microaggregation techniques in AI is also possible. In fact, some data mining methods for large databases use techniques analogous to microaggregation (and also other SDC-related methods such as resampling) for scaling down the data set. This is the case of the work by DuMouchel et al. [18] that, roughly speaking, corresponds to the application of microaggregation to reduce the number of records (prior the application of data mining algorithms).

4. Conclusions and directions for future work

Distortion and aggregation have been shown to be two basic principles underlying both SDC and AI techniques such as machine learning and knowledge acquisition.

While SDC introduces intentional distortion (error injection) to protect microdata sets against disclosure, AI techniques try to overcome accidental errors by combining data coming from several sources (machine learning patterns coming from different sensors, descriptions given by several experts, etc.) to reach a common data set. If we disregard the nature of distortion, the goals of SDC and AI turn out to be inverse. The power of AI techniques to reconstruct an original data set from several distorted versions of it implies an upper bound on the number of disclosure-protected versions of the same data set that can be released. Open research topics related to distortion include the following:

- AI knowledge integration techniques would benefit from considering the wide range of distortion techniques that have been designed for SDC of microdata.
- Machine learning techniques (especially those applied to data released by NSOs) should consider SDC intentional distortion in the databases used for learning.
- Data mining is both a threat and a stress test for SDC protection methods. SDC techniques have to be aware of data mining methods to avoid disclosure.

As outlined above, AI tries to neutralize accidental errors by combining data from several sources. Aggregation is the basic tool used in this context. A

number of aggregation functions have been developed in the field of AI; they allow aggregation of qualitative data and also weighted aggregation where the weight of a value does not only depend on the source it comes from, but also on the value itself. In contrast, the use of aggregation in SDC is far less developed, being restricted to releasing averages of quantitative attributes computed over small aggregates rather than the actual attribute values (microaggregation). Open research topics related to aggregation include the following:

- Extend microaggregation for qualitative attributes.
- Apply OWA and WOWA to compute the released averages in a way that the influence of extreme values is reduced (extreme values often turn out to be confidential in official statistics).
- Based on this new field of application for OWA and WOWA, extract new requirements that can lead to new aggregation operators.

From what has been said above, it follows that joint research between the SDC and the AI communities should be encouraged. Knowledge and techniques that may seem classical to one community can result in substantial improvements for the other community's task.

Acknowledgements

This work is partially supported by the European Commission through project "CASC" (IST-2000-25069) and by the Spanish MCyT and the FEDER fund through project "STREAMOBILE" (TIC2001-0633-C03-01/02). Comments by the reviewers are gratefully acknowledged.

References

[1] N.A. Adam, J.C. Wortmann, Security-control methods for statistical databases: a comparative study, ACM Computing Surveys 21 (1989) 515–556.
[2] N. Anwar, Micro-aggregation—The Small Aggregates Method, internal report, Eurostat, 1993.
[3] J.P. Barthélemy, B. Leclerc, B. Monjardet, On the use of comparison and consensus of classifications, Journal of Classification 3 (1986) 187–224.
[4] P.A. Benedict, The use of synthetic data in dynamic bias selection, in: Proceedings of the Sixth Aerospace Applications of Artificial Intelligence Conference, 1990.
[5] J.C. Bezdek, J.M. Keller, R. Krishnapuram, L.I. Kuncheva, N.R. Pal, Will the real Iris data please stand up? IEEE Transactions on Fuzzy Systems 7 (3) (1999) 368–369.
[6] R. Brand, Microdata protection through noise, in: J. Domingo-Ferrer (Ed.), Lecture Notes in Computer Science, Inference Control in Statistical Databases, vol. 2316, Springer, Berlin, 2002, pp. 97–116.
[7] J. Bacher, R. Brand, S. Bender, Re-identifying register data by survey data using cluster analysis: an empirical study, International Journal of Uncertainty Fuzziness and Knowledge Based Systems 10 (5) 589–607.

[8] L. Breiman, J.H. Friedman, R.A. Olshen, C.J. Stone, Classification and Regression Trees, Wadsworth Int. Group, Belmont, CA, 1984.
[9] D. Defays, P. Nanopoulos, Panels of enterprises and confidentiality: the small aggregates method, in: Proceedings of the 92 Symposium on Design and Analysis of Longitudinal Surveys 1993, pp. 195–204.
[10] A.G. DeWaal, L.C.R.J. Willenborg, Global recodings and local suppressions in microdata sets, in: Proceedings of the Statistics Canada Symposium 95, 1995, pp. 121–132.
[11] J. Domingo-Ferrer, Pros and cons of new information technologies for statistical data protection, in: Proceedings of the New Technologies and Techniques for Statistics'98, Sorrento, Italy, 1998, pp. 233–240.
[12] J. Domingo-Ferrer, J.M. Mateo-Sanz, On resampling for statistical confidentiality in contingency tables, Computers & Mathematics with Applications (38) (1999) 13–32.
[13] J. Domingo-Ferrer, J.M. Mateo-Sanz, Practical data-oriented microaggregation for statistical disclosure control, IEEE Transactions on Knowledge and Data Engineering 14 (2002) 189–201.
[14] J. Domingo, V. Torra, Aggregation techniques for statistical confidentiality, in: R. Mesiar, T. Calvo, G. Mayor (Eds.), Aggregation Operators: New Trends and Applications, Physica-Verlag, Heidelberg, 2002, pp. 261–271.
[15] J. Domingo-Ferrer, V. Torra, Disclosure control methods and information loss for microdata, Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies, North-Holland, Amsterdam, 2002, pp. 93–112.
[16] J. Domingo-Ferrer, V. Torra, A quantitative comparison of disclosure control methods for microdata, confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies, North-Holland, Amsterdam, 2002, pp. 113–134.
[17] D. Dubois, H. Prade, Combination of information in the framework of possibility theory, in: M. Al Abidi, R.C. Gonzalez (Eds.), Data Fusion in Robotics and Machine Intelligence, Academic Press, New York, 1992, pp. 481–505.
[18] W. DuMocuchel, C. Volinsky, T. Johnson, C. Cortes, D. Pregibon, Squashing flat files flatter, in: Proceedings of the Knowledge Discovery in Data, Association of Computing Machinery (ACM), San Diego, CA, 1999, pp. 6–15.
[19] G.T. Duncan, R.W. Pearson, Enhancing access to microdata while protecting confidentiality: prospects for the future, Statistical Science 6 (1991) 219–239.
[20] I.P. Fellegi, A.B. Sunter, A theory for record linkage, Journal of the American Statistical Association 64 (328) (1969) 1183–1210.
[21] K.M. Ford, J.M. Bradshaw, Knowledge Acquisition as Modeling, Parts I and II, International Journal of Intelligent Systems 8 (1–2) (1993) (special issue).
[22] N. Friedman, Learning belief networks in the presence of missing values and hidden variables, in: Proceedings of the 14th International Conference on Machine Learning, Morgan Kaufmann, San Mateo, CA, 1997, pp. 125–133.
[23] W.A. Fuller, Masking procedures for microdata disclosure limitation, Journal of Official Statistics 9 (1993) 383–406.
[24] B.R. Gaines, M.L.G. Shaw, Knowledge acquisition tools based on personal construct psychology, The Knowledge Engineering Review 8 (1) (1993) 49–85.
[25] C. Genest, J.V. Zidek, Combining probability distributions: a critique and an annotated bibliography, Statistical Science 1 (1986) 114–148.
[26] R. Gopal, P. Goes, R. Garfinkel, Confidentiality via camouflage: the CVC approach to database query, in: J. Domingo-Ferrer (Ed.), Proceedings of the Statistical Data Protection'98, Office des Publications Officielles des Communautés Européennes, 1999.
[27] J.M. Gouweleeuw, P. Kooiman, L. Willenborg, P.-P. De Wolf, Applying PRAM: an account of first experiences, in: J. Domingo-Ferrer (Ed.), Proceedings of the Statistical Data Protection'98, Office des Publications Officielles des Communautés Européennes, 1999.

[28] H.A. Guvenir, I. Sirin, Classification by feature partition, Machine Learning Journal 23 (1996) 47–67.
[29] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning, Springer, Berlin, 2001.
[30] M. Holsheimer, A.P.J.M. Siebes, Data Mining: The Search for Knowledge in Databases, Amsterdam, CWI Report CS-R9406, 1994.
[31] A. Hundepool, The CASC project, in: J. Domingo-Ferrer (Ed.), Lecture Notes in Computer Science, Inference Control in Statistical Databases, vol. 2316, Springer, Berlin, 2002, pp. 172–180.
[32] M.I. Jordan, R.A. Jacobs, Hierarchical mixtures of experts and the EM algorithm, Neural Computation 6 (1994) 181–214.
[33] J. Kacprzyk, M. Fedrizzi (Eds.), Multiperson Decision Making Models Using Fuzzy Sets and Possibility Theory, Kluwer Academic Publishers, Norwell, MA, 1990.
[34] J. Kacprzyk, H. Nurmi, M. Fedrizzi (Eds.), Consensus Under Fuzziness, Kluwer Academic Publishers, Norwell, MA, 1997.
[35] A. Kennickell, Multiple imputation and disclosure protection: the case of the 1995 survey of consumer finances, in: J. Domingo-Ferrer (Ed.), Proceedings of the Statistical Data Protection'98, Office des Publications Officielles des Communautés Européennes, 1999.
[36] J.J. Kim, A method for limiting disclosure in microdata based on random noise and transformation, in: Proceedings of the ASA Sect. on Survey Res. Meth, 1986, 303–308.
[37] K. Kira, L.A. Rendell, A practical approach to feature selection, in: Proceedings of the Ninth International Workshop on Machine Learning, Morgan-Kaufmann Publishers, San Mateo, CA, 1992, pp. 249–256.
[38] P. Kooiman, L. Willenborg, J.M. Gouweleeuw, PRAM: A method for disclosure limitation of microdata, CBS research paper 9705, 1997. Available from http://www.cbs.nl/research.
[39] P.D. Laird, Learning from Good and Bad Data, Kluwer Academic Publishers, Norwell, MA, 1988.
[40] J.M. Mateo-Sanz, J. Domingo-Ferrer, A method for data-oriented multivariate microaggregation, in: J. Domingo-Ferrer (Ed.), Proceedings of the Statistical Data Protection'98, Office des Publications Officielles des Communautés Européennes, 1999.
[41] H. Messatfa, An algorithm to maximize the agreement between partitions, Journal of Classification 9 (1992) 5–15.
[42] C.J. Moore, J.C. Miles, Knowledge elicitation using more than one expert to cover the same domain, Artificial Intelligence Review 5 (1991) 255–271.
[43] J.R. Quinlan, C4.5: Programs for Machine Learning, Morgan Kaufmann, San Mateo, CA, 1993.
[44] E. Rahm, H. Hai Do, Data cleaning: problems and current approaches, Bulletin of the Technical Committee on Data Engineering 23 (4) (2001) 3–13.
[45] D. Riaño, Automatic Construction of Descriptive Rules, PhD Dissertation, Universitat Politècnica de Catalunya, Barcelona, 1997.
[46] D.B. Rubin, Satisfying confidentiality constraints through the use of synthetic multiply-imputed microdata, Journal of Official Statistics 9 (1993) 461–468.
[47] P. Samarati, Protecting respondents' identities in microdata release, IEEE Transactions on Knowledge and Data Engineering 13 (2001) 1010–1027.
[48] M. Sugeno, Theory of Fuzzy Integrals and its Applications, PhD Dissertation, Tokyo Institute of Technology, Tokyo, Japan, 1974.
[49] L. Sweeney, Achieving *k*-anonymity privacy protection using generalization and suppression, Interantional Journal of Uncertainty Fuzziness and Knowledge Based Systems 10 (5) 571–588.
[50] P. Tendik, N. Matloff, A modified random perturbation method for database security, ACM Transactions on Database Systems 19 (1) (1994) 47–63.

[51] V. Torra, The weighted OWA operator, International Journal of Intelligent Systems 12 (1997) 153–166.
[52] V. Torra, Re-identifying Individuals using OWA Operators, in: Proceedings of the International Conference on Soft Computing, Iizuka, Japan, 2000.
[53] V. Torra, Towards the re-identification of individuals in data files with non-common variables, in: Proceedings of the European Conference on Artificial Intelligence (ECAI 2000), 2000.
[54] V. Torra, U. Cortés, Towards an automatic consensus generator tool: EGAC, IEEE Transactions on Systems, Man and Cybernetics 25 (5) (1995) 888–894.
[55] W. Vach, Preserving consensus hierarchies, Journal of Classification 11 (1994) 59–77.
[56] A. Valls, V. Torra, Knowledge acquisition from multiple experts, in: Proceedings of the European Summer School Logic, Language and Information, ESSLLI'96, 1996.
[57] A. Valls, V. Torra, On the semantics of qualitative attributes in knowledge elicitation, International Journal of Intelligent Systems 14 (2) (1999) 195–209.
[58] L. Willenborg, T. De Waal, Elements of Statistical Disclosure Control, LNS 155, Springer, New York, 2001.
[59] W.E. Winkler, The State of Record Linkage and Current Research Problems, U.S. Census Bureau, Research Report 99/04, 1999. Available from http://www.census.gov/srd/www/byname.html.
[60] R.R. Yager, On ordered weighted averaging aggregation operators in multi-criteria decision making, IEEE Transactions on Systems, Man and Cybernetics 18 (1988) 183–190.