Aggregation methods to evaluate multiple protected versions of the same confidential data set

Aïda Valls[1], Vicenç Torra[2], and Josep Domingo-Ferrer[1]

[1] Dept. Comput. Eng. and Maths - ETSE, Universitat Rovira i Virgili
Av Paisos Catalans 26, 43007 Tarragona (Catalonia, Spain)
e-mail: {jdomingo, avalls}@etse.urv.es
[2] Institut d'Investigació en Intel·ligència Artificial - CSIC
Campus UAB s/n, 08193 Bellaterra (Catalonia, Spain)
e-mail: vtorra@iiia.csic.es

Abstract. This work is about disclosure risk for national statistical offices and, more particularly, for the case of releasing multiple protected versions of the same micro-data files. This is, several copies of a single original data file are released to several data users. Each user receives a protected copy, and the masking method for each copy is selected according to the research interests of the user: the selected masking method is such that it minimizes the information loss for his/her particular research.

Nevertheless, multiple releases of the same data increase the disclosure risk. This is so, because coalitions of data users can reconstruct original data and, thus, find the original (non-masked) information. In this work we propose a tool for evaluating this reconstruction.

1 Introduction

Data dissemination is a mandatory requirement for statistical offices: they collect data to be published. However, the release of this data has to be done in such a way that there is no disclosure. In other words, no sensitive data is linked to the original respondent.

For example, the publication of incomes, professions and ZIP codes for the inhabitants of a town should not allow the inference of the exact income of a particular inhabitant. Moreover, publication is forbidden if there is a single person in the town for a given pair (ZIP, profession). This is so because the release of such data implies disclosure (knowing the profession and the ZIP code of a person implies knowing his/her income).

To avoid disclosure, masking methods are applied (see [3], [13] for reviews of masking methods and [4] for a comparative study on masking methods performance). Masking methods introduce distortion to the data prior to its publication so that the information is not disclosured. Distortion should be kept small so that published data is valid for researchers and users (they can infer the same conclusions that would be inferred from the original data)

but on the other hand should be protected *enough* so that disclosure is not possible. Statistical Disclosure Control (SDC) studies methods that attempt to perform such a nontrivial distortion.

1.1 Artificial Intelligence and Soft Computing for Statistical Disclosure Control

The fields of Artificial Intelligence and Soft Computing provide several tools that are useful for Statistical Disclosure Control. These tools can be broadly classified in three categories (a more detailed review is given in [6]):

Methods to overcome distortion: Tools and methods for data mining and machine learning have been developed to be resilient to errors in datafiles (either due to accidental or to intentional distortion). Among other uses, data mining and modeling techniques can be used to correct errors (if data do not follow data models) and to fill missing values. Also, information fusion in general and aggregation operators in particular can be used to increase the accuracy of the data. This is particularly appropriate when there are multiple releases of the same data or there is data from multiple sources.

Methods to evaluate disclosure risk: In general, all methods that can be used to overcome distortion are appropriate to evaluate disclosure risk. The better results of a method to overcome distortion, the worse the protection and the larger the disclosure risk. This is so, because if a method can recontruct the original data it means that data can be disclosured.

Methods for re-identification (see e.g. [17]) also fall in this category. They are used to link records that correspond to the same individual but belong to different files. These techniques are used to compare the masked data with the original one. The more records are re-identified, the worse the protection achieved. Some of the existing re-identification methods are the probabilistic based ones, distance based ones and clustering based ones. Recently, techniques based on soft computing have been used [12].

Methods to cause distortion: Although that most artificial intelligence and soft computing techniques are used in SDC for restoring the original data and establishing the disclosure risk, they can also be used for causing intentional distortion to data. This is the case of using aggregation operators [5] for masking numerical data, and machine learning techniques for increasing the performance of a particular masking method [9].

In this work we focus on the first two categories. In particular, we consider the evaluation of disclosure risk in the case of multiple releases of the same data. This is, several copies of a single original data file are released to several data users. Each user receives a protected copy, and the masking method for each copy is selected according to the research interests of the user: the

selected masking method is such that it minimizes the information loss for his/her particular research.

Nevertheless, in this situation the following property has to be taken into account:

Property 1. [6] 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.

This is so, because coalitions of data users can reconstruct original data and, thus, find the original (non-masked) information. Therefore, the following property also holds:

Property 2. [6] 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.

In this work, we describe our information fusion system ClusDM and show its application to reconstruct the original data from multiple releases of the same data. This works extends our previous results presented in [7].

To do so, Section 2 describes our information fusion tool and Section 3 describes the application to a set of multiple protected data. The work finishes in Section 4 with some conclusions.

2 Our information fusion approach

Although our system, ClusDM, has been developed for its application in decision making environments, some of its components can be used for other information fusion applications. In this work we describe its application for fusing multiple releases of the same data. In this section, we give an overview of the general capabilities of the system.

From an abstract point of view, the data fusion component is applied to data matrices V that contain values for each pair (object, attribute). These matrices can be modeled as a function:

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

where $\mathbf{O} = \{O_1, \cdots, O_n\}$ denotes the set of objects, $A_1, A_2, \cdots A_m$ are the attributes and $D(A_i)$ denotes the domain of attribute A_i.

Given a matrix V of this form, the data fusion component builds a new attribute A_C that corresponds to the aggregation of $A_1, \cdots, A_m$. When data is numerical, the attribute A_C is numerical. Instead, if the data is categorical (or some attributes are numerical and other are categorical) the new attribute is categorical.

In the case of numerical information, the system applies the principle of irrelevant alternatives. This is, the aggregated value for each object ($A_C(O_i)$) only depends on the values for that object ($A_1(O_i), \cdots, A_m(O_i)$). This is, there exists a function $\mathbb{C}$ such that:

$$A_C(O_i) = \mathbb{C}(A_1(O_i), \cdots, A_m(O_i))$$

Our system implements several aggregation operators $\mathbb{C}$. Among others, it includes the weighted mean and the OWA operator [14].

In the case of considering categorical and mixed information, the system does not satisfy the condition of irrelevant alternatives. This condition is usually applied (e.g. in [8]) for technical reasons because it simplifies the computations as each object is operated without considering the values of the others. However, while this condition is acceptable for numerical data (in particular, when values correspond to measurements), this is not so true for categorical values. In the categorical case, the values establish equivalences between objects (they are indistinguishable according to a given criteria) and, in the case of ordinal scales, preferences between objects. Therefore it seems natural to keep these relationships in the aggregated attribute.

In order to keep these similarities, the condition of irrelevant alternatives is dropped and we apply clustering for obtaining an aggregated attribute. Once the set of clusters is obtained, the system assigns linguistic labels to each cluster.

According to this, our approach to information fusion considers two steps. We give some additional details of these steps:

Clustering: To obtain the clusters for categorical data, the system assumes an underlying semantics for linguistic labels based on negation functions following [10]. This is, negation functions that are not a one-to-one mapping as in multivalued logic but a one-to-many functions. From the point of view of the user, these negation functions can be interpreted as antonyms following [2].
The clustering of the data is directed by the attributes and their number of categories.

Assignment of linguistic labels: For the assignment, the system is able to select the most appropriate vocabulary considering the ones used in each variable. In this application, we select the vocabulary of the original criterion, because we are trying to see if we can re-identify the original categories. The next step consists of selecting a category to describe each cluster and, if it is needed, to adapt the vocabulary by splitting the category [16]. This assignment process is explained in detail in [15].

3 Experimentation

We have considered data from the *American Housing Survey 1993* [1] and applied our approach to multiple realeases of a single variable (the variable

DEGREE). In the rest of this section we review the application of the approach to a set of 20 records.

Table 1 includes the original variable DEGREE (in column *o.v.*). The set of linguistic terms used by this variable is $L = \{coldest, cold, cool, mild, mixed, hot\}$. In Table 1, the original values are replaced by the position of the category in the set L. Thus, value 1 stands for coldest, 2 for *cold*, 3 for *cool* and so on. This variable has been masked using 4 different masking methods. In particular, we have applied Top and Bottom coding, Global recoding and Post-Randomization Method. Several parameterizations where considered. Additionally, the original values of record f have been updated so the value for column $P4$ is now equal to 3. Table 1 includes the masked variables for 7 different pairs (techniques, parameterizations). The interested reader is referred to [7] for details.

Now, we have applied the ClusDM method to obtain an aggregated value for each record, in order to check if we can reconstruct the original values from these 7 different released variables.

name	o.v.	B4	T4	G4	R10	P8	P9	P4	a.v.		name	o.v.	B4	T4	G4	R10	P8	P9	P4	a.v.
a	3	&	&	3	3	3	3	3	3		k	3	&	&	3	4	3	3	3	3
b	3	&	&	3	2	3	3	3	3		l	3	&	&	3	2	3	3	3	3
c	3	&	&	3	3	3	3	3	3		m	3	&	&	3	3	3	3	3	3
d	3	&	&	3	3	3	3	3	3		n	2	&	2	2	2	2	2	2	2
e	4	4	&	4	4	4	4	4	4		o	3	&	&	3	3	3	3	3	3
f	4	4	&	4	4	4	4	3	4		p	2	&	2	2	2	2	2	2	2
g	4	4	&	4	3	4	4	4	4		q	3	&	&	3	3	3	3	3	3
h	4	4	&	4	4	4	4	4	4		r	5	&	&	n	5	3	3	4	3
i	4	4	&	4	4	4	4	4	4		s	2	&	2	2	2	2	2	2	2
j	1	&	1	n	1	1	1	1	1		t	2	&	2	2	2	3	4	2	2

Table 1. Records used: First column corresponds to a name for the record; second column is the original value (o.v.); columns 3-9 are masked variables; column 10 is the aggregated value (a.v.)

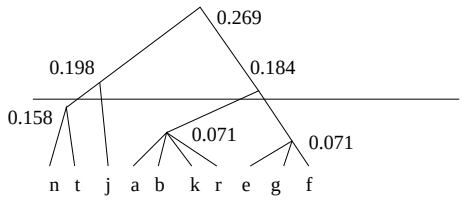

Fig. 1. Dendrogram for the clustering of the records in Table 1

Using the Taxonomic distance and the centroid method, we obtain the dendrogram in Figure 1. Then, an α-cut has to be selected in the tree to obtain a partition of the elements. The α-cut is selected so that the number of clusters is equal to 4 because this is the average number of linguistic labels used in columns *B4*-*P4*. The number of categories used in each column is displayed in Table 2. The selected α-cut is also displayed in Figure 1. The obtained partition is defined by four sets (named A, B, C and D) as follows: $A = \{n, t\}$, $B = \{a, b, k, r\}$, $C = \{j\}$, $D = \{e, f, g\}$. This partition satisfies the conditions required in [7] for a correct partition selection: (i) records with all the variables with the same value should correspond to different clusters (e.g. record a and e) and (ii) clusters should be defined according to the dendrogram.

Note that for the sake of simplicity, we only include in the dendrogram and in the partition one of those elements that are indistinguishable (i.e., it appears the element a but does not appear c because it has the same values for all columns).

Column	B4	T4	G4	R10	P8	P9	P4
Number of used labels	2	3	3	5	4	4	4

Table 2. Number of categories used in each columns

Once the clusters have been obtained, a category has to be assigned to each one. This is done considering the distance between each cluster and the *ideal* element (the one that has larger value for all categories). Then, taking into account this distance and the location of the clusters in relation to the semantics of linguistic labels (see [15]) the following assignment is given: Class C is assigned to category 1, Class A is assigned to 2, Class B is assigned to 3 and Class D is assigned to 4. These assignments are shown in Table 1.

4 Conclusions

In this work we have reviewed the applicability of artificial intelligence and soft computing techniques for statistical disclosure control. We have shown that information fusion techniques can be used by coallitions of users to overcome distortion in multiple releases of the same data. This is, information fusion techniques can be used for reconstructing an original data set out of n different distorted versions of the same data set.

We have described an example consisting of 20 records with 7 different releases of the same variable. We have applied to this records the system ClusDM.

The results obtained support the Property 2 stated in the introduction: disclosure risk is proportional to the reconstruction capabilities of informa-

tion fusion systems. In this case, we have shown that the original data was reconstructed except for record r that is assigned to Category 3 instead of 5.

Acknowledgements

The authors are partially supported by the EU project CASC: Contract: IST-2000-25069 and CICYT project STREAMOBILE (TIC2001-0633-C03-01/02)

References

1. Census Bureau, (1993), American Housing Survey 1993, Data publicly available from the U. S. Bureau of the Census through the Data Extraction System, `http://www.census.gov/DES/www/welcome.html`
2. de Soto, A.R., Trillas, E., (1999), On antonym and negate in fuzzy logic, Int. J. of Int. Systems, 14:3, 295-303
3. Domingo-Ferrer, J., Torra, V., (2001), A Quantitative Comparison of Disclosure Control Methods for Microdata, 111-133, in Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies, P. Doyle, J. I. Lane, J. J. M. Theeuwes, L. M. Zayatz (Eds.), Elsevier.
4. Domingo-Ferrer, J., Torra, V., (2001), Disclosure Control Methods and Information Loss for Microdata, 91-110, in Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies, P. Doyle, J. I. Lane, J. J. M. Theeuwes, L. M. Zayatz (Eds.), Elsevier.
5. Domingo-Ferrer, J., Torra, V., (2002), Aggregation techniques for statistical confidentiality, in "Aggregation operators: New trends and applications", (Ed.), R. Mesiar, T. Calvo, G. Mayor, Physica-Verlag, Springer.
6. Domingo-Ferrer, J., Torra, V., (2002), On the Connections between Statistical Disclosure Control for Microdata and Some Artificial Intelligence Tools, submitted.
7. Domingo-Ferrer, J., Torra, V., Valls, A., (2002), Semantic based aggregation for statistical disclosure control, submitted.
8. Dubois, D., Koning, J-L., (1991), Social choice axioms for fuzzy set aggregation, Fuzzy Sets and Systems, vol.43, pp.257-274.
9. F. Sebe, J. Domingo-Ferrer, J. M. Mateo-Sanz, V. Torra, Post-Masking optimization of the tradeoff between information loss and disclosure risk in masked microdata sets, Lecture Notes in Computer Science 2316, 163-171.
10. Torra, V., (1996), Negation functions based semantics for ordered linguistic labels, Int. J. of Intelligent Systems, 11 975-988.
11. Torra, Towards the re-identification of individuals in data files with common variables, Proc. of the 14th European Conference on Artificial Intelligence (ECAI2000), Berlin, Germany, 2000.
12. Torra, V., (2000), Re-identifying Individuals using OWA Operators, Proc. of the 6th Int. Conference on Soft Computing, Iizuka, Fukuoka, Japan, 2000.
13. Willenborg, L., De Waal, T., (1996), Statistical Disclosure Control in Practice, Springer LNS 111.

14. Yager, R. R., (1988), On ordered weighted averaging aggregation operators in multi-criteria decision making, IEEE Trans. on SMC, 18 183-190.
15. Valls, A., Moreno, A., Sanchez, D., A multi-criteria decision aid agent applied to the selection of the best receiver in a transplant, Proc. of the 4th Int. Conference on Enterprise Information Systems, ICEIS, 431-438, Ciudad Real, Spain, 2002.
16. Valls, A., Torra, V., (2000), Explaining the consensus of opinions with the vocabulary of the experts, Proc. IPMU 2000, Madrid, Spain, 2000.
17. Winkler, W. E., (1995), Advanced methods for record linkage, American Statistical Association, Proceedings of the Section on Survey Research Methods, pp. 467-472.