European Journal of Operational Research 216 (2012) 257–269

Contents lists available at ScienceDirect

European Journal of Operational Research

journal homepage: www.elsevier.com/locate/ejor

Invited Review

Recent advances in optimization techniques for statistical tabular data protection

Jordi Castro *

Department of Statistics and Operations Research, Universitat Politècnica de Catalunya, Jordi Girona 1–3, 08034 Barcelona, Spain

ARTICLE INFO

Article history:
Received 11 June 2010
Accepted 30 March 2011
Available online 12 April 2011

Keywords:
Linear programming
Network flows
Mixed integer linear programming
Statistical disclosure control
Large-scale optimization

ABSTRACT

One of the main services of National Statistical Agencies (NSAs) for the current Information Society is the dissemination of large amounts of tabular data, which is obtained from microdata by crossing one or more categorical variables. NSAs must guarantee that no confidential individual information can be obtained from the released tabular data. Several statistical disclosure control methods are available for this purpose. These methods result in large linear, mixed integer linear, or quadratic mixed integer linear optimization problems. This paper reviews some of the existing approaches, with an emphasis on two of them: cell suppression problem (CSP) and controlled tabular adjustment (CTA). CSP and CTA have concentrated most of the recent research in the tabular data protection field. The particular focus of this work is on methods and results of practical interest for end-users (mostly, NSAs). Therefore, in addition to the resulting optimization models and solution approaches, computational results comparing the main optimization techniques – both optimal and heuristic – using real-world instances are also presented.

© 2011 Elsevier B.V. All rights reserved.

1. Introduction

National Statistical Agencies (NSAs) store information about individuals or *respondents* (persons, companies, etc.) in microdata files. A microdata file V of s individuals and t variables is a $s \times t$ matrix where v_{ij} is the value of variable j for individual i. Formally, it can be defined as a function

$$V : I \rightarrow D(V_1) \times D(V_2) \times \cdots \times D(V_t),$$

that maps individuals of set I to an array of t values for variables $V_1, \ldots, V_t$, $D(\,)$ being the domain of those variables. According to this domain, variables can be classified as numerical (e.g., "age", "net profit") or categorical ("sex","economy sector"). From those microdata files, tabular data is obtained by crossing one or more categorical variables. For instance, assuming a microdata file with information of inhabitants of some region, crossing variables "profession" and "municipality" the two-dimensional *frequency* table of Fig. 1 may be obtained. Instead, the table could provide information about a third variable; these tables are named *magnitude* tables. For instance, the table of Fig. 2 shows the overall salary for each profession and municipality. Formally, a table is a function

$$T : D(V_{i_1}) \times D(V_{i_2}) \times \cdots \times D(V_{i_l}) \rightarrow \mathbb{R} \text{ or } \mathbb{N},$$

l being the number of categorical variables that were crossed. The result of function T (cell values) belongs to $\mathbb{N}$ for a frequency table, and to $\mathbb{R}$ for a magnitude table.

Although tabular data show aggregated information, there is a risk of disclosing individual information. For instance, if the two tables of Figs. 1 and 2 are published, then any attacker knows that the salary of the unique respondent of cell (M_2,P_3) is 22,000€. This is named an *external attacker*. If there were two respondents in that cell, then any of them could deduce the other's salary, becoming an *internal attacker*. Even if there was a larger number of respondents, e.g. 5, if one of them had a salary of, e.g. 18,000€, there would be a disclosure risk, since the contribution of the largest respondent could exceed some predefined percentage of the cell total; this cell would be reported as sensitive by the so-called *dominance rule*. A more dangerous and difficult to protect scenario is named the *singleton problem* or *multi-attacker problem*. This happens, for instance, when two cells with a single respondent (a singleton) appear in the same table relation (e.g., a row or a column of Figs. 1 and 2), such that any of them can deduce the other's contribution. This situation can be generalized to more than two singletons with collusions. The singleton problem has been discussed in Jewett (1993), Robertson (2000), and more recently in Daalmans and de Waal (2010). In all the above situations, prior to publication, NSAs have to apply some tabular data protection method. In short, those methods, whose origins date back to Bacharach (1966), basically either suppress or perturb the table cell values.

A different set of protection techniques apply directly to the original microdata files, instead of the resulting tabular data. These are out of the scope of this work. Some recent improvements on microdata protection methods can be found in Domingo-Ferrer and Mateo-Sanz (2002), Hansen and Mukherjee (2003), Muralidhar and Sarathy (2006), and in the monographs Domingo-Ferrer and Franconi (2006), Domingo-Ferrer and Magkos (2010),

* Tel.: +34 934015854.
E-mail address: jordi.castro@upc.edu
URL: http://www-eio.upc.es/~jcastro

0377-2217/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.ejor.2011.03.050

	P_1	P_2	P_3	P_4	P_5	TOTAL
M_1	20	15	30	20	10	95
M_2	72	20	1	30	10	133
M_3	38	38	15	40	11	142
TOTAL	130	73	46	90	31	370

Fig. 1. Two-dimensional frequency table showing number of persons for each profession and municipality.

	P_1	P_2	P_3	P_4	P_5	TOTAL
M_1	360	450	720	400	360	2290
M_2	1440	540	22	570	320	2892
M_3	722	1178	375	800	363	3438
TOTAL	2522	2168	1117	1770	1043	8620

Fig. 2. Two-dimensional magnitude table showing overall salary (in 1000€) for each profession and municipality.

Domingo-Ferrer and Saigin (2008), Domingo-Ferrer and Torra (2004) and Willenborg and de Waal (2000). Although the number of records in a microdata file r is in general much larger than the number of cells n in a table ($r \gg n \gg 0$), tabular data involve a number of linear constraints m, and in some real-world instances $m \gg 0$. These linear constraints model the relations between inner and total cells, the most usual relation being that the sum of some inner cells is equal to some marginal cell. Microdata protection in general involves few (if not zero) linear constrains. Therefore, tabular data protection methods rely on linear programming (LP), mixed integer linear programming (MILP), and even mixed integer quadratic programming (MIQP) technology, making the protection of complex and large tables a difficult problem. Some huge instances (which result in MILP problems of order of millions of variables and constraints) can be found in http://www-eio.upc.es/jcastro/data.html.

The detection of the sensitive cells to be protected is made by applying some sensitivity rules. Although it is an important step of the data protection process, it is not covered here. Indeed, currently, those rules do not rely on optimization or operations research methodology. Practical details about these rules can be found in Hundepool et al. (2010). Additional information can be found in Domingo-Ferrer and Torra (2002) and Robertson and Ethier (2002).

Although it contains references to recent literature, this paper is not meant to be a comprehensive survey on statistical disclosure control of tabular data. The interested reader is referred, for instance, to the research monographs Domingo-Ferrer and Franconi (2006), Domingo-Ferrer and Magkos (2010), Domingo-Ferrer and Saigin (2008), Domingo-Ferrer and Torra (2004), Willenborg and de Waal (2000), and the recent survey Salazar-González (2008). Guidelines to end-users of tabular data protection methods can be found in the handbook Hundepool et al. (2010). Compared to those previous works, the main contributions of this paper are: (i) it focuses on two of today more relevant techniques for NSAs (namely, cell suppression problem, and controlled tabular adjustment); (ii) not only the resulting optimization models are presented, but also the main solution techniques are sketched; (iii) it reports computational results comparing the available techniques using state-of-the-art software for tabular data protection, showing the lacks and benefits of the different models and solution approaches.

The structure of the paper is as follows. Section 2 shows the different types of tables that can be obtained, and how they are modeled; this background is needed for the subsequent sections. Section 3 introduces tabular data protection methods. Sections 4 and 5 focus on two of the most widely used protection techniques, the *cell suppression* and the *controlled tabular adjustment*, both describing the optimization models and outlining the main solution approaches.

2. Tabular data: types and models

Some protection methods of Section 3 are either only valid or may be specialized (i.e., made more efficient) for some particular type of tabular data. It is thus instrumental to know in advance the type of table to be protected and how to model it.

2.1. Classification of tables

Tables can be classified according to different criteria. Two of the simplest criteria for classification are "cell values" and "sign of cell values". According to the cell values, the two classes of tables were already introduced in Section 1: *frequency tables* – also named contingency tables – and *magnitude tables*. According to the sign of cell values, tables are classified as either *positive* or *general* tables. Cell values of positive tables are non-negative, which is the most usual situation. For instance, all frequency tables and most magnitude tables, like "salary" for "profession" × "municipality", are positive tables. Cell values of general tables can be positive or negative. An example of a general table would be "variation of gross domestic product" for "year" × "state". Assuming a table is general instead of positive can be instrumental in some protection methods. Indeed, those methods usually involve the solution of difficult LP or MILP problems; the lower bounds of some variables are $-\infty$ for general tables (0 for positive ones). This property has been exploited in some efficient heuristics for general tables (Carvalho et al., 1994).

For a modelling point of view, and to exploit the type of table in the resulting LP or MILP from the data protection method, the most important classification criteria is "table structure". Indeed, some protection methods can only be applied to particular table structures. Moreover, the different models in Section 2.2 are tailored for some table structures. According to their structure, tables may be classified as *single k-dimensional*, *hierarchical*, or *linked* tables. A single k-dimensional table is obtained by crossing k categorical variables. For instance, tables of Figs. 1 and 2 are two-dimensional. A hierarchical table is a set of tables obtained by crossing some variables, and a number of these variables have a hierarchical relation. For instance, consider the three tables of Fig. 3. The left subtable shows number of respondents for "region" × "profession"; the middle subtable, a "zoom in" of region R_2, provides the number of respondents for "municipality"(of region R_2) × "profession"; finally the right subtable, "zip code" × "profession", details municipality R_{21}. This table belongs to a particular class named 1H2D, two-dimensional tables with one hierarchical variable. Finally, linked tables are the most general situation. A linked table is a set of tables obtained from the same microdata file. In theory, the set of all tables obtained from a microdata file should be considered together as a (likely huge) linked table. Hierarchical and k-dimensional tables are particular cases of linked tables. Note that, in theory, the only safe way for protecting all the tables from a microfile, is to jointly protect them as a single linked table. Unfortunately, in many cases the size of the

	C_1	C_2	C_3
R_1	5	6	11
R_2	10	15	25
R_3	15	21	36

T_1

	C_1	C_2	C_3
R_{21}	8	10	18
R_{22}	2	5	7
R_2	10	15	25

T_2

	C_1	C_2	C_3
R_{211}	6	6	12
R_{212}	2	4	6
R_{21}	8	10	18

T_3

Fig. 3. Hierarchical table made of three subtables: "region" × "profession", "municipality" × "profession" and "zip code" × "profession".

Download English Version:

https://daneshyari.com/en/article/480164

Download Persian Version:

https://daneshyari.com/article/480164

Daneshyari.com