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-