Making Better Sense of the Demographic Data Value in the Data Mining Procedure

Katherine M. Shelfer, Xiaohua Hu

College of Information Science and Technology, Drexel University Philadelphia, PA 19104, USA

ABSTRACT

Data mining of personal demographic data is being used as a weapon in the War on Terrorism, but we are forced to acknowledge that it is a weapon loaded with interpretations derived from the use of dirty data in inherently biased systems that mechanize and de-humanize individuals. While the unit of measure is the individual in a local context, the global decision context requires that we understand geolocal reflexive *communal* selves who have psychological and social/societal relationship patterns that can differ markedly and change over time and in response to pivotal events. Local demographic data collection processes fail to take these realities into account at the data collection *design* stage. As a result, existing data values rarely represent an individual's multi-dimensional existence in a form that can be mined. An abductive approach to data mining can be used to improve the data *inputs.* Working from the "decision-in," we can identify and address challenges associated with demographic data collection and suggest ways to improve the quality of the data available for the data mining procedure. It is important to note that exchanging old values for new values is rarely a 1:1 substitution where qualitative data is involved. Different constituent user populations may require different levels of data complexity and they will need to improve their understanding of the data values reported at the local level if they are to effectively relate various local demographic databases in new and different global contexts.

1. INTRODUCTION AND OVERVIEW

In 1948, the UN General Assembly declared personal privacy and the associated freedoms of association, belief, inquiry and movement to be universal (UN 1948). They are guaranteed by the U.S. Constitution. According to Andrew Ford of Usenet, *"Without either the first or second amendment, we would have no liberty."* At the same time, data mining of personal demographic data is increasingly used as a weapon to deny opportunity to individuals, even though it is a weapon loaded with interpretations that are derived from the use of dirty demographic data containered in inherently biased systems that mechanize and de-humanize individuals.

When demographic data on individuals is mined, individuals are judged on the basis of data associations that form certain links and patterns. This means that the use of data mining in law enforcement triggers a response based on an individual's presumptive *guilt* even though a nation may espouse the *de jure* presumption of *innocence*. Since presumption of guilt can victimize the innocent even in the most traditional of legal proceedings (Yant 1991), there was a public outcry over the Total Information Assurance initiative that would have authorized secret law enforcement data mining "fishing expeditions" using combined heterogeneous demographic databases (e.g., credit card companies, medical insurers, and motor vehicle databases). DARPA has asserted its intention to protect constitutional guarantees in relation to data mining activities (DARPA 2002). However, given the role that data mining already plays in risk-based assessments of all types (PCI 2004), it is imperative to provide the most accurate interpretation of demographic data values possible.

Regardless of how we approach demographic data mining, we should remain aware that a number of specific federal and state laws and regulations address the control and/or use of personal data. There are reasons for formalizing such protections that take precedence over technological and resource constraints focused on economies of scope and scale. For example, researchers and applied workers cannot seem to agree on the relative merits of the various statistical methodologies that support important data mining applications (e.g., credit scoring) in use at the present time, yet these systems are sometimes used as the sole decision criteria to adjudicate "guilt" and deny opportunity to individuals and classes of individuals in a range of otherwise unrelated risk-based contexts.