Effective algorithms for computing covariance and correlation based on privacy-protected data

Authors:

Day, Joshua,
The University of Texas at El Paso, Department of Computer Science

Jalal-Kamali, Ali
The University of Texas at El Paso, Department of Computer Science

Kreinovich, Vladik
The University of Texas at El Paso, Department of Computer Science

Submitted: 2013-07-15 22:52:47

Keywords
privacy protection, estimating correlation, statistical database, interval uncertainty

Discipline Area
Computer Science

Abstract

In medicine (and in many other situations), it is necessary to process data while preserving the patients' confidentiality. One of the most efficient methods of preserving privacy is to replace the exact values with intervals that contain these values. For example, instead of an exact age, a database only contains the information that the age is, e.g., between 10 and 20, or between 20 and 30, etc.

We are interested in computing correlations (and other statistical characteristics) based on this data. For privacy-protected data, different values from the intervals lead, in general, to different estimates for the desired statistical characteristic. Our objective is then to compute the range of possible values of these estimates.

Algorithms for effectively computing such ranges have been developed for situations when intervals come from the original surveys, e.g., when a person fills in whether his or her age is between 10 or 20, between 20 and 30, etc. These intervals, however, do not always lead to an optimal privacy protection; it turns out that more complex, computer-generated "intervalization" can lead to better privacy under the same accuracy - or, alternatively, to more accurate estimates of statistical characteristics under the same privacy constraints. It is therefore necessary to extend the existing efficient algorithms for computing covariance and correlation based on privacy-protected data to this more general case of interval data.

In this work, we describe generalized algorithms, and we provide estimates for their computation time which show that these algorithms are indeed computationally feasible and efficient.

Funding Support
DHS, NCBSI Center

Research Program
NCBSI REU Program

Recommended Citation

Day, Joshua; Jalal-Kamali, Ali; Kreinovich, Vladik. "Effective algorithms for computing covariance and correlation based on privacy-protected data". Abstraacts of the College Office of Undergraduate Research Initiatives (COURI) Symposium, El Paso, Texas, August 3, 2013, Abstract 231