Risk-utility trade-off for a new method of statistical disclosure limitation based on a mixture model with constraints

Anna Oganian

Georgia Southern University (GA, US)
National Center for Health Statistics (MD, US)

Before releasing databases which contain sensitive information about individuals, data publishers must apply Statistical Disclosure Limitation (SDL) methods to them, in order to avoid disclosure of sensitive information on any identifiable data subject. Transforming unsafe data into safe data can be viewed as an optimization problem whose solution is far from trivial. SDL methods aim to reduce the risk of disclosure of confidential information and at the same time to maximize the utility of the released data. These are conflicting goals.

In this work we describe a new SDL approach, which controls attribute disclosure risk for continuous variables. The proposed method is based on synthetic data generation using mixture model with constraints on parameters of components' spread. The method guarantees that the requirements of an attribute disclosure risk metric called v-dispersion are satisfied.

Experiments with real data show that our method compares very favorably with other methods of disclosure limitation for continuous microdata in terms of utility and risk.

Keywords and phrases: Statistical disclosure limitation (SDL), v-dispersion, mixture models, expectation-maximization (EM) algorithm, constraints.