Privacy-Preserving Data Sharing in High Dimensional Regression and Classification Settings

Stephen E. Fienberg, Carnegie Mellon University, Pittsburgh, PA USA
fienberg@stat.cmu.edu
Jiashun Jin, Carnegie Mellon University, Pittsburgh, PA USA
jiashun@stat.cmu.edu

August, 2012
LARC-TR-05-12

LARC Technical Report Series: http://smu.edu.sg/centres/larc/larc-technical-reports-series/

SMU
SINGAPORE MANAGEMENT UNIVERSITY

Carnegie Mellon University

ABSTRACT

We focus on the problem of multi-party data sharing in high dimensional data settings where the number of measured features (or the dimension) p is frequently much larger than the number of subjects (or the sample size) n, the so-called $p >> n$ scenario that has been the focus of much recent statistical research. Here, we consider data sharing for two interconnected problems in high dimensional data analysis, namely the feature selection and classification. We characterize the notions of “cautious", “regular", and “generous" data sharing in terms of their privacy-preserving implications for the parties and their share of data, with focus on the \feature privacy" rather than the \sample privacy," though the violation of the former may lead to the latter. We evaluate the data sharing methods using a *phase diagram* from the statistical literature on multiplicity and Higher Criticism thresholding. In the two-dimensional phase space calibrated by the signal sparsity and signal strength, a phase diagram is a partition of the phase space and contains three distinguished regions, where we have no (feature) privacy violation, relatively rare privacy violations, and an overwhelming amount of privacy violation.