Detecting a Wide Diversity of Associations in Very Large Data Sets

Christopher Pardy
Prince of Wales Clinical School, UNSW Medicine, Sydney, Australia,
cpardy@unsw.edu.au

Susan R Wilson*
Prince of Wales Clinical School, UNSW Medicine, School of Mathematics &
Statistics, UNSW Science, Sydney, Australia
Mathematical Sciences Institute, Australian National University, Canberra, Australia,
sue.wilson@anu.edu.au

Major challenges arising from today's "data deluge" include how to handle the commonly occurring situation of different types of variables (say, continuous and categorical) being simultaneously measured, as well as how to assess the accompanying flood of questions. Based on information theory, a bias-corrected mutual information (BCMI) measure of association that is valid and estimable between all basic types of variables has been proposed. It has the advantage of being able to identify non-linear as well as linear relationships. Based on the BCMI measure, a novel exploratory approach to finding associations in large data sets has been developed. These associations can be used as a basis for finding clusters and networks, for example, in large data sets in which different types of variables have been collected on each individual (or unit). The application of this exploratory approach is very general. Comparisons will be made with the recently proposed measure, maximal information coefficient (MIC). Illustrative examples include exploring relationships (i) between social, economic, health and political indicators from the World Health Organisation and partners, and (ii) in genomic (say, gene expression and genotypic) and clinical data.

Key Words: Mutual information, exploratory data analysis, high-dimensional data, identifying non-linear relationships