Discovering Influential Variables and Their Applications: A Review

Shaw-Hwa Lo

Columbia University, New York, U.S.A
slo@stat.columbia.edu

Abstract

A trend in all scientific disciplines, based on technology, is the increasing availability of high dimensional data in which are buried important information. We review a general computer intensive approach, based on an earlier method (Lo and Zheng 02, 04) for detecting which, of many potential explanatory variables, have an influence on a dependent variable Y. This approach is suited to detect influential variables, where causal effects depend on the confluence of values of several variables. It has the advantage of avoiding a difficult direct analysis,involving possibly thousands of variables, by dealing with many randomly selected small subsets from which smaller subsets are selected, guided by a measure of influence I. The main objective is to discover the influential variables, rather than to measure their effects. Once they are detected, the problem of dealing with a much smaller group of influential variables should be vulnerable to appropriate analysis. If time permits, applications and findings on IBD and breast cancer dataset will be presented.