TESTING FOR PRESENCE OF CLUSTERING EFFECT IN MULTILEVEL MODEL WITH HIGH-DIMENSIONAL PREDICTORS

by

Frances Claire S. San Juan

A thesis submitted in partial fulfillment

of the requirements for the degree of

Master of Science (Statistics)

School of Statistics

University of the Philippines

Diliman, Quezon City

ABSTRACT

As big data become more accessible with the boom of data analyzing software, creating value through analytics has grown in demand. Dealing with large data sets in anomaly detection problems, accurate tagging of anomalies is oftentimes lacking and expensive. Unsupervised learning via clustering analysis can be performed to derive labelled data, but used alone, is prone to high false alarm rates. We propose a nonparametric procedure to test presence of clustering effect in a multilevel model with a large set of predictors. Model estimation is done through principal component regression (PCR) and two-way analysis-of-variance (ANOVA), embedded in a backfitting algorithm. Hypothesis test is based on sieve bootstrap. A simulation study showed that the test is effective in detecting high clustering effects, and is optimal when sample size exceeds the number of predictors. The test can be a useful support tool to help address limitations of existing cluster-based methods in anomaly detection.

Keywords : anomaly detection, multilevel model, clustered data, high-dimensional data, principal component regression, nonparametric regression, generalized additive models, backfitting algorithm, bootstrapping