Logistic regression model averaging for rare events data

Meng-Fan Huang[1]* (黃孟凡), Chun-Shu Chen[1] and Jin-Hua Chen[2]
[1]Institute of Statistics and Information Science,
National Changhua University of Education
[2]Biostatistics Center and Graduate Institute of Biostatistics, China Medical University

Abstract

In many scientific fields, logistic regression is a popular method for analyzing the binary data accompanied with some covariate variables. When the two classifications are extremely imbalanced, the estimation of model parameters has been shown to be severely biased. Thus, the risk assessment for the rare events would be inaccurate. In this talk, we focus on assessing the risk variations of rare events based on logistic regression models. Instead of selecting a best model based on a particular variable selection criterion, we propose a local model averaging method through a data perturbation technique applied to some information criteria to obtain several risk estimates. Then, an approximately unbiased estimator of Kullback-Leibler loss is proposed to choose among them. The proposed local model averaging method takes the estimation uncertainty and selection uncertainty into account which are generally ignored by usual modeling procedures. Therefore, the proposed method has a superior performance on various situations. We present complete simulations to assess the effectiveness of our proposed method and a real data example about the necrotizing enterocolitis (NEC) is also applied for illustration.

Keywords: estimation uncertainty, imbalanced data, Kullback-Leibler loss, maximum likelihood estimate, risk assessment