Regression study for thyroid disease prediction Comparison of crossing-over approaches and multivariate analysis

ABSTRACT

Regression analysis is one of the common machine learning method to model the relationship between dependent and independent variables. In this study, we aim to tackle two crucial elements that affect the performance of regression models, which are the type of crossing-over method used for model evaluation and multivariate analysis with the number of predictors. We used the classic thyroid disease dataset from the UCI machine learning repository and compare the crossing-over approaches of k-fold with different folds, bootstrap, Leave One Out Cross-Validation (LOOCV), and repeated k-fold on linear and logistics regression. For multivariate analysis, we compare the performance of the models by using the different combinations of bi-predictors and multi-predictors. Our result shows that models that use kfold cross-validation have greater performance, and a higher number of k does not improve the model performance. For the multivariate analysis, we found that the number of variable is not the key element to determine the performance of a model, rather than a suitable combination of strong predictors. Future studies could explore the effects of cross-validation and multivariate analysis on other machine learning algorithms.