Abstract No: 156 ***ICT, Mathematics and Statistics***

A COMPARISON BETWEEN GENETIC ALGORITHM AND LOGISTIC REGRESSION ON VARIABLE SELECTION: A CASE STUDY

A.M.S.M.C.M. Aththanayake[1*], W.B. Daundasekera[1], T.S.G. Peiris[2], F. Noordeen[3] and M.V.M. Divaratna[3]

[1]Department of Mathematics, Faculty of Science, University of Peradeniya, Peradeniya, Sri Lanka
[2]Department of Mathematics, Faculty of Engineering, University of Moratuwa, Katubedda, Sri Lanka
[3]Department of Microbiology, Faculty of Medicine, University of Peradeniya, Peradeniya, Sri Lanka
[*]chathurimalee@gmail.com

Identifying a combination of variables causing infections or infectious diseases is one of the main tasks in clinical models in medicine. Logistic Regression (LR) has been widely used to identify such variable under the assumptions of linearity of independent variables and absence of multi-collinearity. Experimental data which has a large number of variables may not meet these assumptions. Thus, the method of LR may fail to identify variables that cause infections or infectious diseases. Hence, the Genetic Algorithm (GA), which does not depend on pre-defined assumptions, can be applied under such circumstances. By evaluating the prediction rates of LR and GA techniques, this studys' objective was to perform binary LR and GA on a sample of clinical data and compare the goodness of fit statistics to identify the best variable reduction method. Two models were built for 40 independent variables (3 non-categorical and 37 categorical) for a sample of 497 observations collected from a suspected Respiratory Syncytial Virus (RSV) infected children under five years of age, who were hospitalized in the Kegalle Base Hospital from May 2016 to July 2018. The goodness of fit on the two models was compared using statistical methods: 2log-likelihood, Cox & Snell R-square, Nagelkerke R square, correctly classified percentage, specificity, and sensitivity. A total of 162 children were tested RSV positive by an RSV antigen detection method. Except for specificity, GA shows better goodness of fit measurements compared to all other considered statistical methods. However, GA performs better in predictions when sensitivity and specificity were taken together. Moreover, the GA method filtered 17 independent variables to predict RSV infection status, while the LR method filtered 9 independent variables. This case study suggests that GA shows better performance in analysis when the predefined assumptions were not satisfied and solving high dimensional classification problems in a large or complex searching space in the background of the study.

Keywords: Clinical Data, Fitness Function, Genetic Algorithm, Logistic Regression, Sensitivity