Multi-stage feature selection in identifying potential biomarkers for cancer classification

Wong, Yit Khee[a]; Chan, Weng Howe[a]; Nies, Hui Wen[a]; Moorthy, Kohbalan A-L[b]
[a] Universiti Teknologi Malaysia, Faculty of Computing, Johor, Malaysia
[b] Universiti Malaysia Pahang, Faculty of Computing, College of Computing and Applied Sciences, Pekan, Malaysia

ABSTRACT

Biomarkers are indicators that show the disease state or its progression of certain health conditions. Identification of biomarkers greatly raises the probability of earlier diagnosis and could be further applied in developing effective treatment for the disease. Besides conducting laboratory analysis, potential biomarkers also can be identified by analysing gene expression data through feature selection and machine learning. Many algorithms have been applied and introduced in this area, yet the challenge of high dimensionality of gene expression data remains and it could lead to the existence of noise that could negatively impact the analysis outcome. Therefore, this study aims to investigate and develop a better feature selection to identify potential biomarkers from gene expression data and construct a deep neural network classification model using these selected features. Thus, a multistage feature selection, namely CIR is proposed, that composed of Chi-square, Information Gain and Recursive Feature Elimination. The dataset used in this study consists of the integration of seven ovarian cancer gene expression datasets from GEO database. Both selected genes and classification model are evaluated through biological context verification and classification performance respectively. The proposed method shows improvements over the existing methods in terms of accuracy (+2.2294%), precision (+8.1415%), recall (+2.2294%), Fl-score (+4.5494%) and AUC scores (+0.2302). The proposed CIR method successfully identified eight genes that could be potential biomarkers for ovarian cancer, including WFDC2,S100A13, PRG4, NRCAM, OGN, B3GALT2, VGLL3, and GATM which are further verified through literature.

KEYWORDS

Bioinformatics; Deep neural network; Feature selection; Gene expression

ACKNOWLEDGMENT

Authors wish to thanks Universiti Teknologi Malaysia for supporting this work. This work was supported by the Ministry of Higher Education under Fundamental Research Grant Scheme (FRGS/1/2019/ICT02/UTM/02/12) and Universiti Malaysia Pahang (www.ump.edu.my) under the grant RDU192218.