An intelligent optimized ReliefF model for autism gene selection from microarray data

G. Anurekha[1,*] and P. Geetha[2]

[1]Department of Computer Science and Engineering, Vel Tech Rangarajan Dr Sagunthala R&D Institute of Science and Technology, Chennai 600 062, India

[2]Department of Information Science and Technology, College of Engineering, Anna University, Guindy, Chennai 600 025, India

Data science is an interdisciplinary domain that has vast applications. One among them is the bioinformatics domain. Bioinformatics is an essential field of science that incorporates computational, statistical and mathematical aspects for various experimental studies. Microarray experiments are vital in identifying the genetic cause of acute diseases. But the massive volume of data generated from microarray experiments makes analysis tedious and time-consuming for researchers. Hence, an efficient method for selecting relevant gene features is essential. This article proposes a new model for autism gene subset selection termed optimized reliefF (ORelifF). The proposed ORelifF employs a new fuzzy multi-verse optimizer (FMVO) to optimize the selection of neighbours. The filter-based feature selection model obtains the gene subset with high classification accuracy and a low error rate. Benchmark datasets of National Center for Biotechnology Information, USA with accession numbers GSE25507, GSE26415 and GSE2704 are used for implementation. The experimental findings show that the proposed autism gene selection model outperformed the state-of-the-art techniques.

Keywords: Autism, dimensionality reduction, feature selection, gene selection, metaheuristic, microarray.

DATA science is a discipline that extracts knowledge and large number of features per sample. Several researchers classify the process of identifying relevant features in high-dimensional data as non-deterministic polynomial-time hard (NP-hard) data[2]. The existing literature shows that filter[3,4], wrapper[5,6] and hybrid methods[7,8] were used to select relevant genes. Contrary to the wrapper and hybrid approaches, the filter-based method does not interact with the classifier. Yet it is the least complex, most effective, and easy to use method[9].

Literature also shows that researchers have preferred relief-based filter approaches[4,10] to select relevant features for several real-time applications in the past. The basic version of the relief algorithm was proposed by Kira and Rendell[11]. It worked well for binary classification problems. Later, Robnik-Šikonja and Kononenko[12] proposed the ReliefF and RReliefF algorithms for noisy data, regression, and multiclass problems. An Iterative Relief algorithm was proposed by Sun[13]. The Iterative Relief incorporates an expectation-maximization algorithm to mitigate the drawback of the existing Relief algorithm. Multi-class margin definition is employed to extend the Iterative Relief. An online learning algorithm is defined to control computational costs. Cai *et al.*[14] used a local hyper-linear approach with a Relief algorithm to identify the informative genes from a high-dimensional dataset. This approach estimates all the features by local approximation with the local