ABSTRACT

Feature selection is critical in machine learning, particularly with high-dimensional datasets and decision-making while handling big data which presents significant challenges. This paper introduces an innovative approach using the NSGA-II (Non-dominated Sorting Genetic Algorithm II) for feature selection, aiming to minimize the number of features and classification error simultaneously. The methodology involved optimizing feature selection on training sets and evaluating on validation sets, identifying Pareto front solutions, and testing on the test set for generalization capability. Results showed effective feature reduction while maintaining or improving classification accuracy. Key metrics, such as average training error, minimum validation error, and test error were tracked to highlight consistency. Non-dominated binary vectors of optimal feature subsets were analyzed for further feature selection via voting. Frequent features are selected after voting with mutual correlation taken into account to ensure that the feature set is most optimum. This framework aims to narrow the disparity between the classification error observed on the training and test datasets, thereby enhancing model generalization and robustness of feature subsets for better decision-making. This study provides a robust framework for feature selection for a set of features having high dimensions, leveraging evolutionary algorithms and cross-validation techniques to enhance machine learning model performance and reduce the error difference between train and test sets while trying to solve the problem of data driven evolutionary computation.