Hyperparameter tuning and pipeline optimization via grid search method and tree-based AutoML in breast cancer prediction

ABSTRACT

Automated machine learning (AutoML) has been recognized as a powerful tool to build a system that automates the design and optimizes the model selection machine learning (ML) pipelines. In this study, we present a tree-based pipeline optimization tool (TPOT) as a method for determining ML models with significant performance and less complex breast cancer diagnostic pipelines. Some features of pre-processors and ML models are defined as expression trees and optimal gene programming (GP) pipelines, a stochastic search system. Features of radiomics have been presented as a guide for the ML pipeline selection from the breast cancer data set based on TPOT. Breast cancer data were used in a comparative analysis of the TPOT-generated ML pipelines with the selected ML classifiers, optimized by a grid search approach. The principal component analysis (PCA) random forest (RF) classification was proven to be the most reliable pipeline with the lowest complexity. The TPOT model selection technique exceeded the performance of grid search (GS) optimization. The RF classifier showed an outstanding outcome amongst the models in combination with only two pre-processors, with a precision of 0.83. The grid search optimized for support vector machine (SVM) classifiers generated a difference of 12% in comparison, while the other two classifiers, naïve Bayes (NB) and artificial neural network—multilayer perceptron (ANN-MLP), generated a difference of almost 39%. The method's performance was based on sensitivity, specificity, accuracy, precision, and receiver operating curve (ROC) analysis.

Keyword: Machine learning; breast cancer; Genetic programming; Tree-based pipeline optimization tool