Optimising Classification of Congestive Heart Failure Using Feature Weight Importance Correlation (FWiC)

Busola Oronti

In this work, a novel method for selecting the optimal set of input features for classifying the presence of congestive heart failure (CHF) using a supervised machine learning approach is presented. Secondary data in the form of long-term (20 hours) electrocardiogram (ECG) Holter monitor data for healthy and CHF patients were obtained from PhysioNet. The datasets were pre-processed and analysed to obtain heart rate variability (HRV) measures in the time, frequency, and non-linear domains. A random forest classifier (RFC) was employed to carry out the binary classification task and two different models were explored. For the first model, the embedded RFC feature importance attribute was employed, and the 10 most important input features selected. The second model adopts a multi-classifier technique which integrates the feature weight importance correlation (FWIC) method. Fourteen commonly used classifiers were aggregated to select the 10 best features, based on the frequency of occurrence of each feature across all classifiers and the associated feature importance score. For each classifier, features were ordered based on rated order of importance and score (1 being the highest and 10 the lowest). Preference was given to features occurring more frequently across all classifiers. Where features report the same frequency across classifiers, a higher value of the mean feature importance score was used to decide their hierarchy in the frequency table. Features occurring more than once were subjected to iterative correlation analysis based on the order of importance of the feature in the frequency table. Correlation values were used to remove features from the dataset once the set correlation iteration threshold (CIT) value (≤0.45 & 0.40 respectively) was exceeded. Results obtained on the test set for Model 1 showed a root mean squared error (RMSE) of 0.25 and a prediction accuracy of 0.94. For Model 2b (where CIT≤0.40), the model attained 100% accuracy. Additionally, stratified shuffle split cross-validation was employed to test the effect of class imbalance and data leaks on model performance. Interestingly, the test accuracy for Model 2b remained at 100% for each stratified fold. Consequently, The FWIC input feature selection method provides a generalised approach to feature engineering for machine learning models irrespective of the algorithm used, thus assuring optimal model performance as well as the relevance of predicted variables to the task at hand.