ABSTRACT

Exploring and analyzing large datasets has become an active research area in the field of data mining in the last two decades. There had been several approaches available in the literature to investigate the large datasets that comprise of millions of data. Whereas science and industry have leveled up their data collecting activities, conventional data analysis investigation in statistics and machine learning has been comparatively slow to begin the challenge and much research is still point out on relatively small data sets.

Classification is the essential and critical section that has to be carefully formulated for the process of analyzing the large dataset. Although, some amount of the features can lead to high classification accuracy, the extra features added over there cannot contribute much to the performance but they do not humiliate the general performance. Then, the classifier is expected to classify unlabeled instances into one or more predefined categories based on their content.

An efficient and promising choice to process large data set is to independently learn from a number of moderate-sized subsets and integrate their results through ensemble of classifiers. Ensemble classification is one of the most recent approaches widely used in pattern recognition and machine learning. The main goal of the ensemble is to attain significant classification

accuracy than that offered by its individual classifiers with a lesser complexity.

The present research work aims to investigate the usage of various efficient feature selection and classification approaches for providing significant results. This work describes the application of various feature selection and classification techniques which helps to analyze the large datasets. The present research work also focuses on improving the feature selection and classification accuracy by selecting the best features in the large medical dataset. The work analyzes the medical data records for diagnosing various types of diseases. Hybrid and Ensemble principles have been introduced in feature selection and classification models for handling the high volume, complex medical records.

Hybrid model of KPCA-SVM with enhanced GA has been used as the hybrid feature selection approach. Then, ensemble feature selection model that combine the outputs of multiple feature selector approaches such as Hybrid Genetic Particle Swarm Optimization (HGPSO) and Hybrid Artificial Bee Colony Algorithm (HABCA) to improve the prediction accuracy for the subsequent classifier learning tasks.

The present research work focuses on ensemble classification in large datasets. The work uses both homogeneous and heterogeneous ensemble classification models. k-nearest neighbor algorithm (KNN), Adaptive Neuro-Fuzzy Inference System (ANFIS) Classifier and Fuzzy Rule-Based Classifier (FRB) are the classifiers used in ensemble classification.