ABSTRACT

Feature selection is a standout amongst the most important techniques in machine learning algorithms. Likewise, feature selection is named attribute selection or variable selection. It is specifically applied to enhance the performance and the prediction accuracy for complex data classification. Feature selection approach selects a subset of the features of a dataset, and it is done in view of a characterized measure to remove the irrelevant and redundant data through data preprocessing method.

After reviewing existing literature, it was found that various studies highlight the several feature selection approaches such as information gain, gain ratio, chi squared test, F-Score, correlation coefficient scores, correlation based feature selection, etc. The challenges posed by these limitations have motivated the author to develop a new methodology to enhance the effectiveness of these learning processes.

The present research work uses multiple datasets with blends of various datasets like medical datasets, forecasting datasets, chemical datasets and
composed numerals ("0"-"9") separated from an accumulation of Dutch utility maps datasets. The dataset comprises of a base numbers to the greatest quantities of attributes and instances. The attribute characteristics are additionally classified into nominal and numeric datasets.

Hybrid feature subset selection for an ensemble method of effective Adaptive Boosting algorithm with decision tree has performed well with higher classification accuracies. These methodologies were compared with several decision trees based classification algorithm, for example C4.5, DecisionStump, Random forest, Naïve Bayes Tree with tenfold cross-validation.

The First phase of the research work incorporates a hybrid feature subset selection for multiple dataset using decision tree based classifiers. The feature selected utilizing Information Gain (IG) is consolidated with the features selected from ReliefF which creates the resultant feature subset. At that point the resultant component subset is thus consolidated with a Correlation-based Feature Selection (CFS) technique to produce the aggregated feature subset. Fusion is completed with intersection and exclusive OR to generate the aggregate feature subset. The fundamental objective of the hybrid feature selection is to enhance the classification accuracy, prediction and to lessen the execution time by utilizing multiple datasets.
In the second phase, a novel ensemble method named eAdaBoost (Effective Adaptive Boosting) which is developed by enhancing the existing AdaBoost algorithm. The eAdaBoost decreases the error rate when contrasted with existing techniques and produces the better precision by reweighting each feature for further process. The exactness of the classifiers and statistical test comparisons are made with different boosting algorithms. The algorithm has been computed with different dataset with various weight thresholds and the execution is investigated. The proposed strategy eAdaBoost gives better classification accuracy, expectation precision, prediction accuracy and the execution time are likewise less when contrasted with different classifiers.

In the third phase, merging of hybrid feature subset selection and ensemble method is framed. This phase works on the intersection and XOR methods with boosting algorithms to find classification precision. The classification accuracy is compared with boosting algorithms to give better precision, prediction accuracy and to lessen the execution time by utilizing the aggregated subset.

The proposed methodology is implemented by the data collected from various Repositories such as I2R Data Mining Department's dataset repository (Kent Ridge Bio-medical Dataset), Bioinformatics Laboratory and UCI machine learning repository of the multiple datasets. Challenges such as
increasing the classification accuracy, reducing the computational time and reducing the high-dimensionality are addressed in this research work. The results of the proposed method and the existing methods are compared through various classifiers.