AN EFFICIENT WAY OF USING WRAPPERS IN BIG DATA CLASSIFICATION

M.N.F. Fajila

Department of Mathematical Sciences, South Eastern University of Sri Lanka
fajilanisper@gmail.com

Data is dramatically growing with the growth of time. However, the value of data forces the scientists to find patters to use the high dimensional data efficiently. Dimensionality reduction is an essential technique in data science when handling big data. Although always the techniques are being introduced, applying correct technique at right position still seems to be challenging. One such technique is wrappers for machine learning. Feature selection plays a major role in classification of big data. A feature can be more informative in the presence of another feature. Thus, no feature should be removed without assessing. Wrappers select all the possible combinations of feature subsets, and finally provide the most informative subset which classifies the data with a higher accuracy. But, compared to filters wrappers are much slower and consume a huge amount of time when applied to big data. Therefore, in the proposed approach, wrapper is applied after the application of filter in order to get rid of the computational complexity. This approach uses gain ratio filter followed by classifier subset evaluator, the wrapper for feature sub set selection. The proposed technique is validated and evaluated on two high dimensional microarray data sets namely; lung cancer data set and breast cancer data set. It provided 97.10% accuracy (only with two misclassifications) and 78.78% accuracy for lung cancer and breast cancer data sets respectively. Thus, the results show that the proposed approach is extremely efficient in terms of accuracy and computational time too.

***Keywords**: Big data, Classification, Dimensionality reduction, Microarray, Wrapper.*