Abstract

These days, machine learning algorithms deal with large amount of data. These data contain tens of thousands of instances, each instance being represented by hundreds of features. When dealing with large datasets, it is often the case that the information available is somehow redundant for the purpose of Data Mining applications. The process of identifying and removing the irrelevant features from the original data so that the learning algorithms mainly focus on the relevant data which are useful for analysis and future predictions is called Feature Selection. It can improve the resulting comprehensibility of the resulting classifier significantly. A single feature selection method may be bias. So ensemble of feature selection methods is employed

This is applicable for a wide range of datasets. We are motivated to design and implement an ensemble of feature selection methods. But after selecting a good feature method it is equally important to know about different clusters so that we can accurately and effectively classify data into different data sets or cluster. After that selecting a subset of features we perform classification to find the accuracy.

We also implemented majority voting ensemble feature selection method and validated the output accuracy using the four classifiers Decision tree, Random forest, KNN, SVM.

We have also implemented our proposed algorithm using distributed client server architecture where EFS-MI runs on the client and the client sends the optimal subset of the feature to the server.