Fast Clustering Based Feature Selection Approach for Data

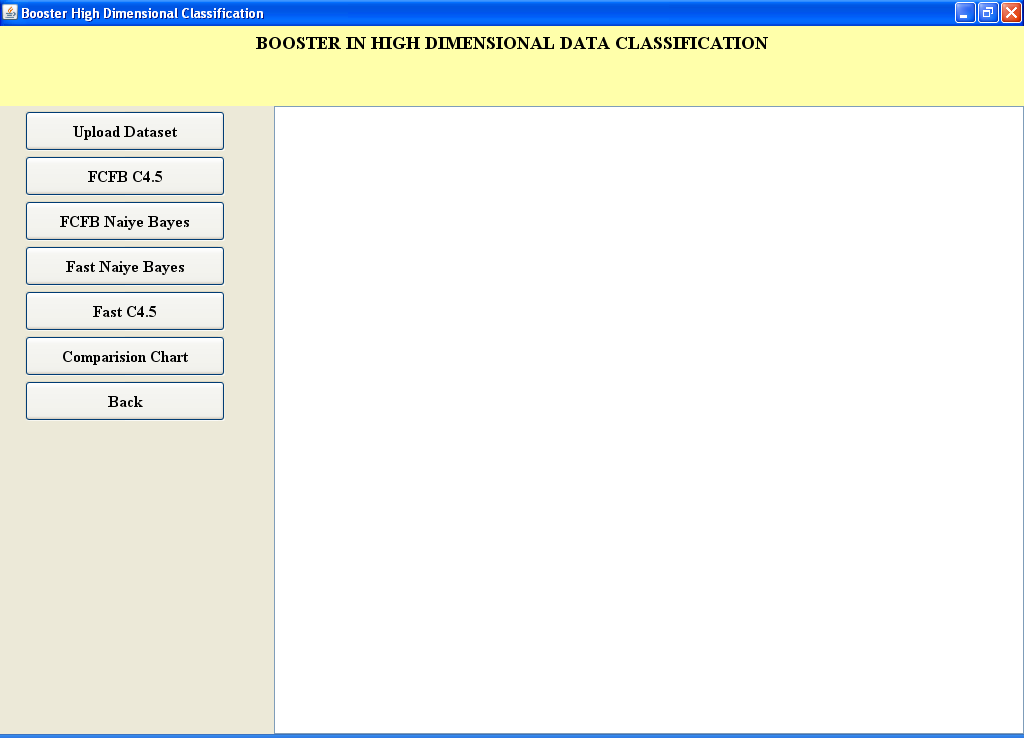
In this paper we are using FAST clustering based feature selection algorithm to select important features from large imbalance dataset and ignore unrelated or unimportant features. Some dataset may contain 1000’s of attributes and for classification if we use all attributes then system execution time will increase and classification accuracy will decrease.

To overcome from above issue attributes feature selection algorithms introduces such as FCFB, Fast etc. FCBF is not efficient so we are using Fast feature selection algorithm which will cluster entire data and choose only those attributes from clusters whose values are more related to dataset class value. Due to this technique only related features will be chosen and classifier accuracy will increase.

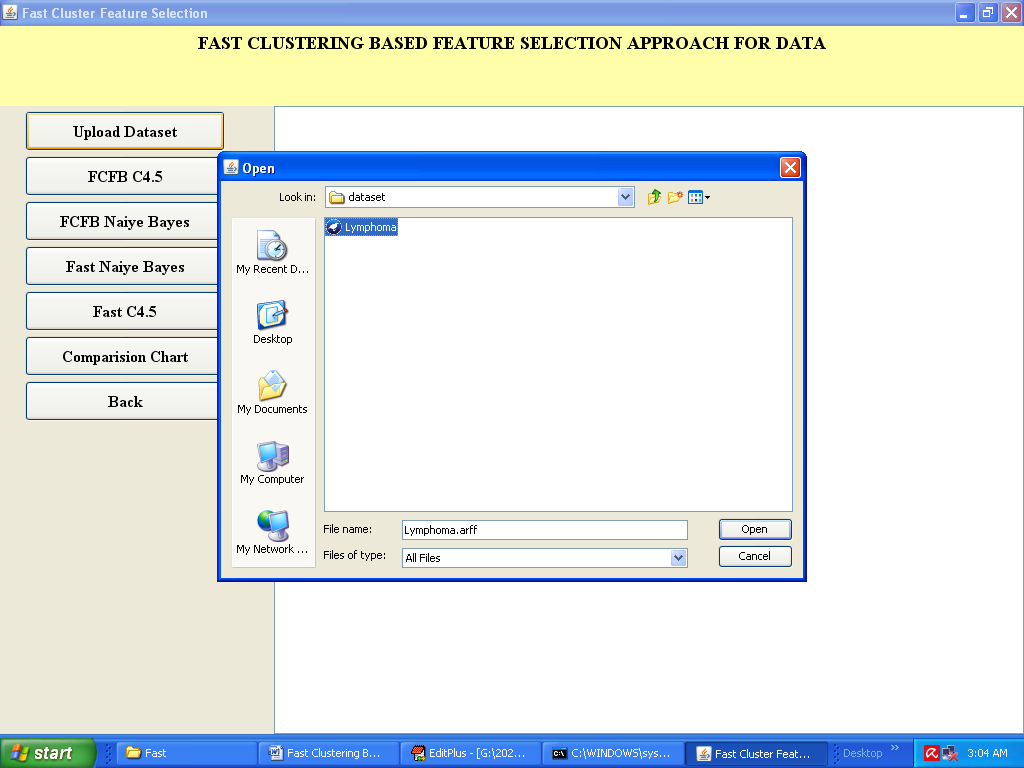
Here we are using Micro Array Genes dataset which contains more than 4000 features but these feature selection algorithms will choose only 100 to 80 important features.

Screen shots

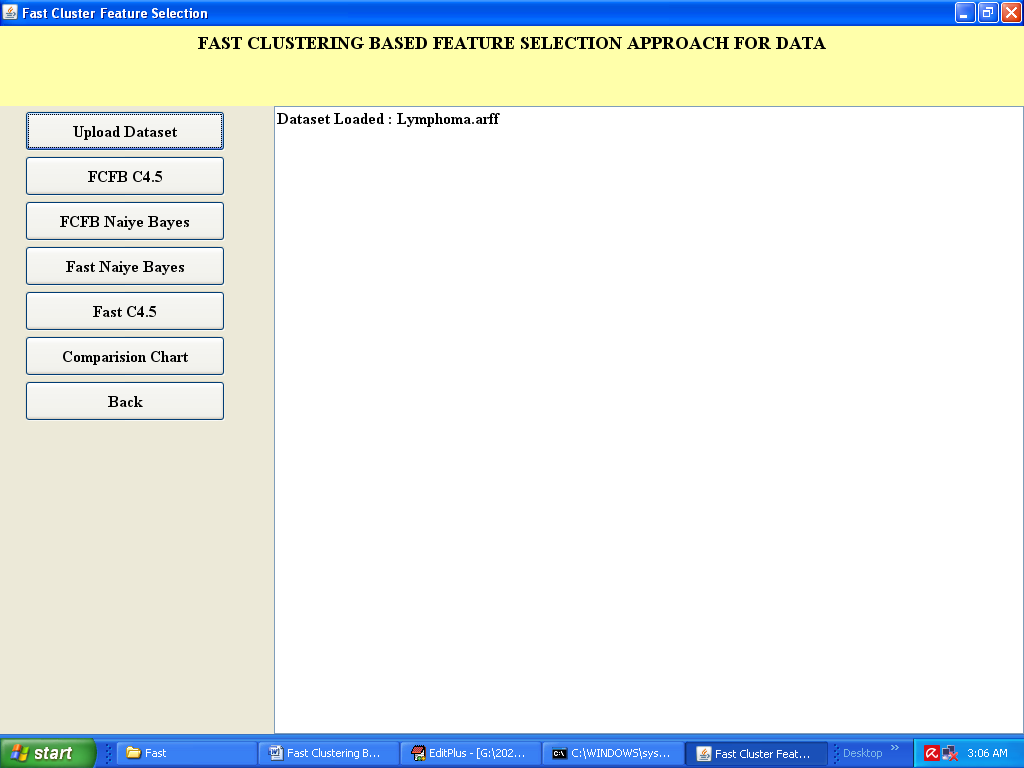
To run project double click on ‘run.bat’ file to get below screen



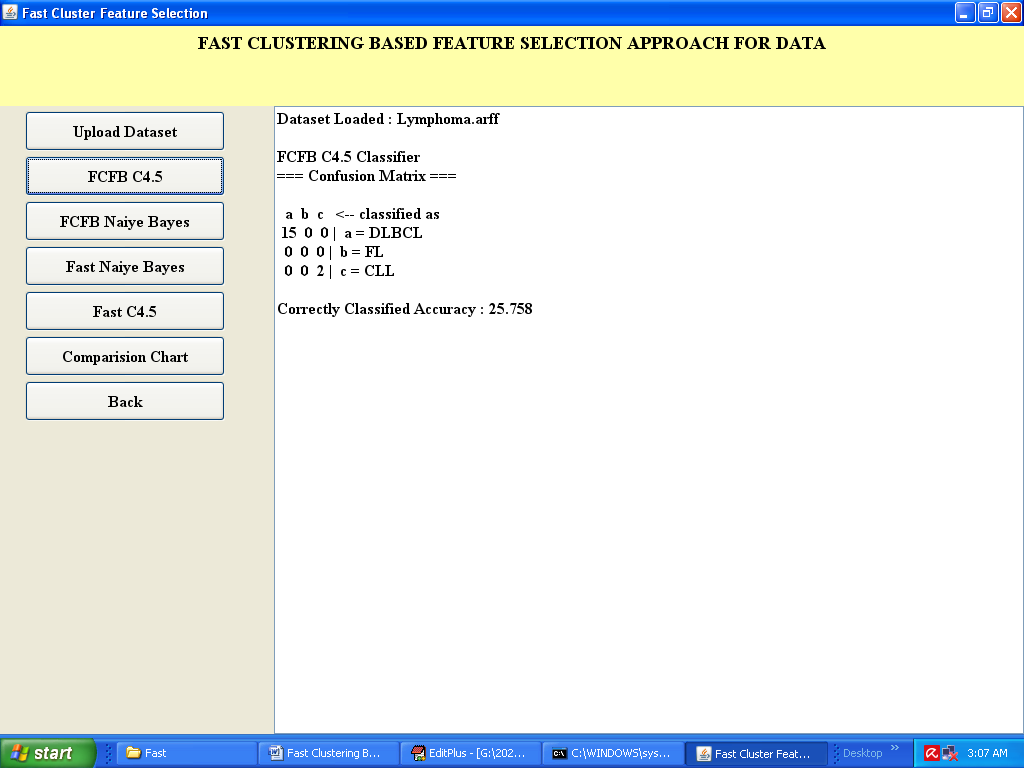
In above screen click on ‘Upload Dataset’ button to upload dataset



In above screen I am uploading ‘Lymphoma’ genes Micro Array dataset which contains more than 4000 attributes. After uploading dataset will get below screen



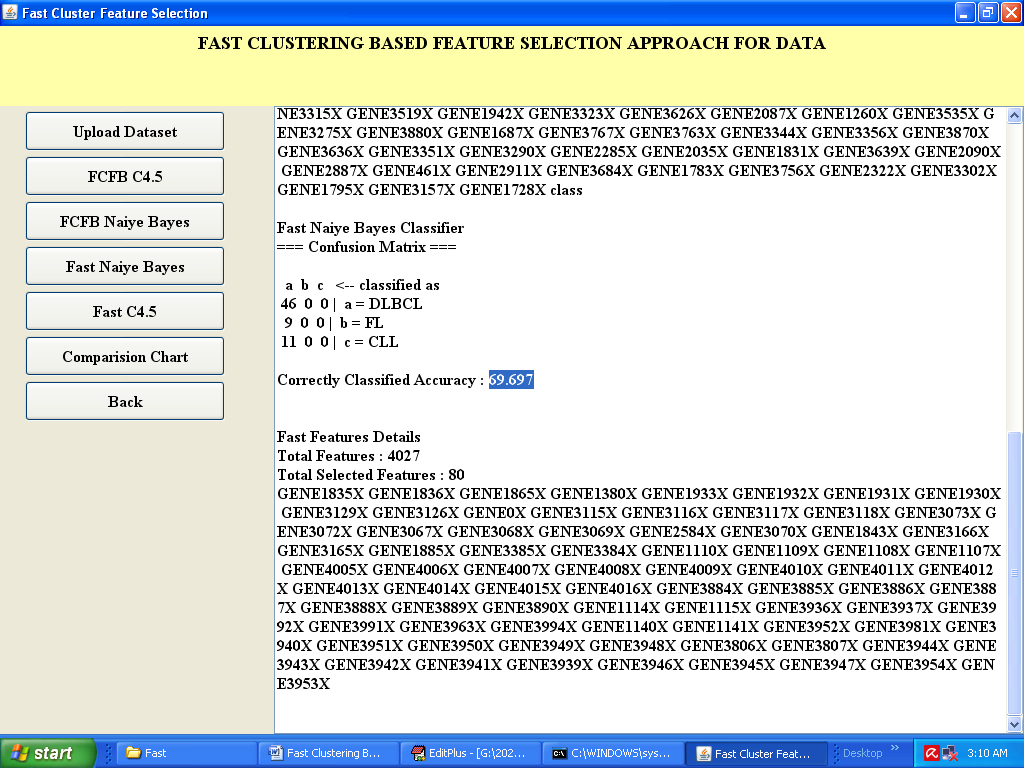
Now click on ‘FCBF C4.5’ button to run c4.5 classifier and apply FCBF Fast Clustering based features



With FCBF C4.5 we got 25% accuracy and the available classes in dataset are DLBCL, FL and CLL. Now run FCBF Naïve Bayes

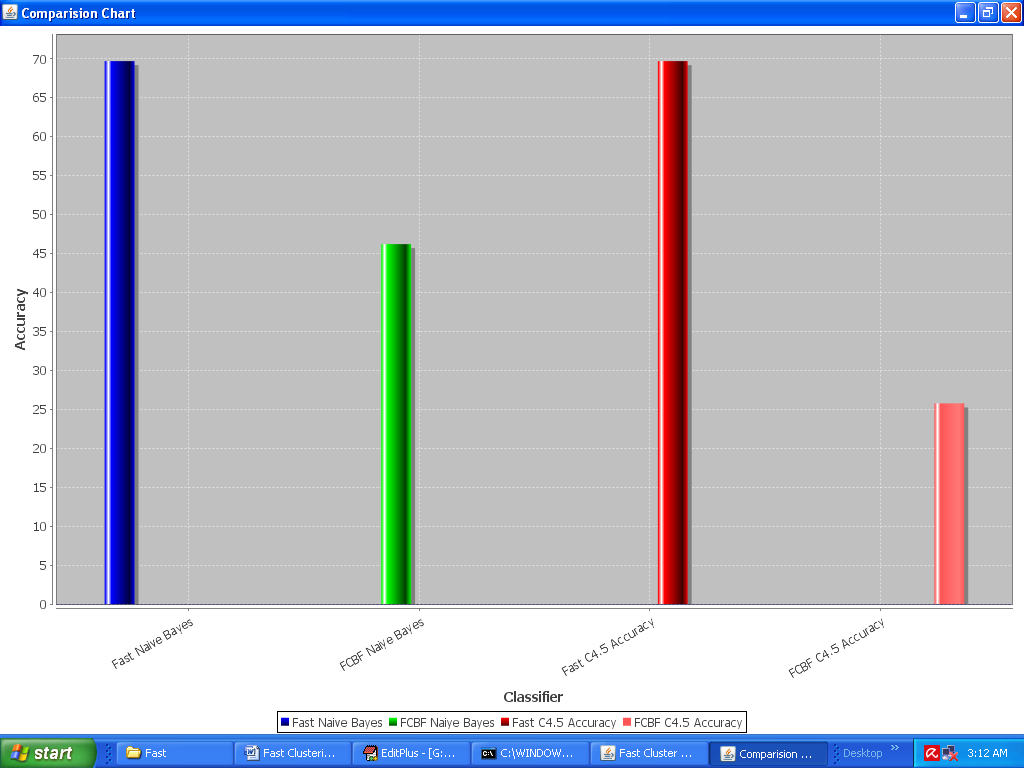


In above screen with FCBF Naïve Bayes we got 46 % accuracy and total features are 4027 and selected features are 111 and below showing names of all selected features. Now run Fast Naïve Bayes



In above screen with Fast Naïve Bayes we got nearly 70 % accuracy and selected attributes are 80. Now run Fast C4.5 algorithm

In above screen with fast C4.5 also got 70% accuracy now click on ‘Comparison Graph’ button to get accuracy graph



In above graph x-axis represents technique name as Fast and FCBF with two classifier Naïve Bayes and C4.5 and y-axis represents accuracy. From above graph we can say Fast is achieving more accuracy compare to FCBF