Stage 1:

At this stage we tried to do the procedure once from beginning to the end in the simplest way:

A subset of 100 training objects was chosen which contained equal number of four classes (RNA, DNA, DRNA, nonDRNA).

A collection of <#features> features were computed for the training objects by using the websites referred to in the class. These features were selected by Akila according to the criteria he found in the literature.

In Rapidminer, a decision tree model was selected and appropriate blocks for importing data from csv file, imputing missing places, specifying label data, and cross validation was added to the design. with this settings, the results of cross validation was reported as follows:

Stage 2:

We thought that probably a large number of features selected not only had no contribution to the predictive performance, they were even decreasing predictive performance by adding noise. Therefore a good subset of features should have been selected. This could be done by either implementing the methods discussed in the class (filter based approach and wrapper based approach), or automatically by rapidminer. We found out that rapidminer have some modules for this job named “optimization”. It seems that it works as a wrapper-based approach because it trains and tests the model multiple times and each time gauges the performance of the prediction. After doing this, the result was improved considerably.

<Results after feature selection>

Stage 3:

We tried different predictive models to see which of them performs better on the data. We also used voting to combine the results of different classifiers

<Results for different models including voting based ensemble method>

Stage 4:

Since some of the features were continuous which may create problems with some of the methods including decision tree, we decided to discretize the data. This improved the performance further to 79%

<Results after discretization>

Stage 5: stratified sampling for

For unknown reason, when I enabled the stratified sampling for cross validation, the performance, increased considerably.