Classification

----Siyu Zhou

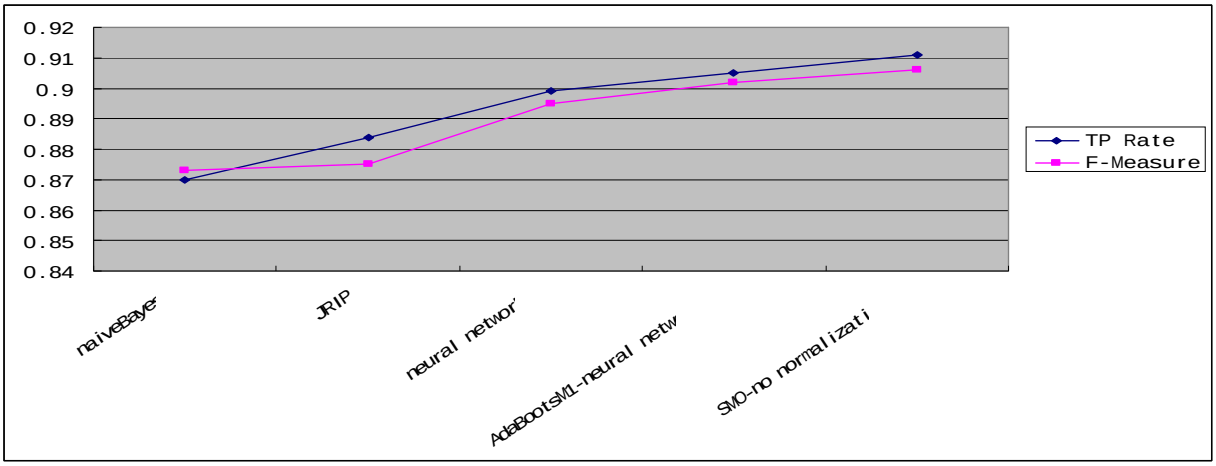
When I contrast these five approaches, I find that Bayesian is the most fast one as 0.02 second. But the prerequisite of this approach is independence between attributes while the dataset in this assignment cannot guarantee it. Since Bayesian is based on probability, it is more suitable to large data. After preprocessing, I just have 671 pieces of data which cannot provide accurate probability to Bayesian.

The running time of JRIP approach is three times over Bayesian while they have closely F-measure values (Graph-1). JRIP uses repeated incremental pruning to produce error reduction while there are seven rules in this dataset. I am afraid this approach will cause over-fitting.

The running time of neural network is 0.88 second which is significantly increased compared to Bayesian. At the same time, TP-rate and F-Measure is increased 0.03 and 0.02. I think the cost of time is valued in this case, but if in a large dataset this method will slow. When I see this model, it didn't show how a conclusion reached. It just tells me there are ten sigmoid nodes, threshold and then the summary of this approach.

AdaboostM1 plus neural network as classifier has the highest TP Rate and F-Measure when compared with AdoboostM1 plus other classifiers. However, the running time is too long as 8.57 second which is tenfold than only neural network classifier. Because this approach pays more attention to the misclassified tuples, I think this might also cause over-fitting.

SMO have the highest TP rate and F-Measure in these five approaches while the running time is just 0.07 second which is a worthy time cost. As SMO has strong mathematical foundation, I think it is precisely. Using kernel, this approach can transform input data into high dimensions and process easily than the original data. I think this is the best one in these five approaches.



Graph-1