剪枝算法初步测试实验

Experiments:

**Dataset:** bairong\_prepared.csv, DPFINAL\_prepared.csv

training set: 70%

test set: 30%

**Evaluation:**

SEN: sensitivity, recall

SPE: specificity

ACC: overall accuracy

AUC: area inder curv

Kappa: cohen\_kappa\_score

MCC: matthews\_corrcoef

TIME(s): time used(second)

DEPTH: depth of a tree

LEAVES: leaves of a tree(leaf nodes)

**Schedule:**

1. Experiment of blank control(default parameters):

max\_depth = None

min\_samples\_split = 2

min\_samples\_leaf = 1

2. Experiment of tuning the hyper-parameters of decision tree

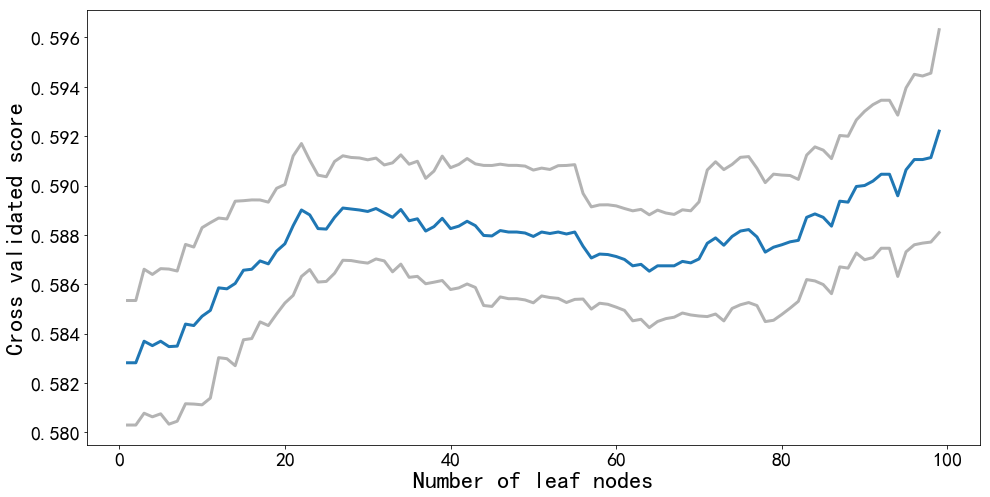
max\_depth = ‘None’ append list(range(10, 40, 2))

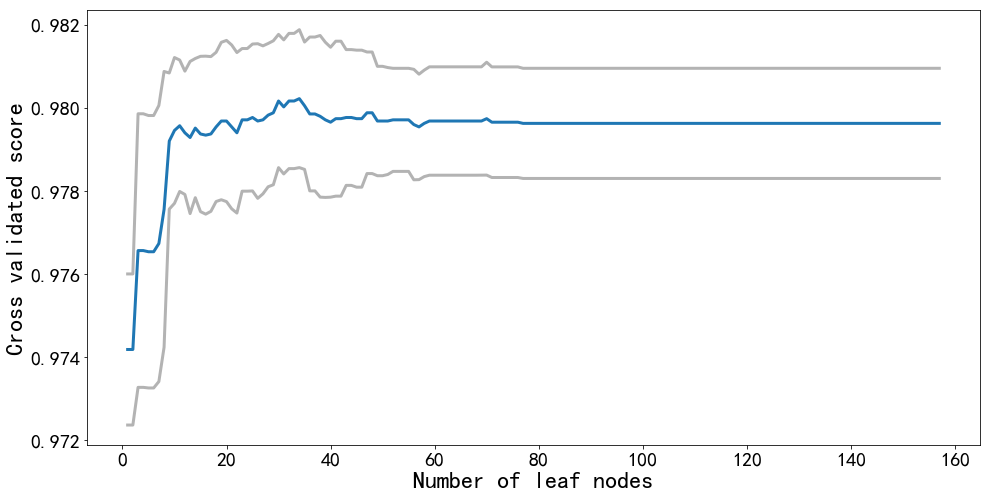
min\_samples\_split: (min(0.00015\*N+50,200)) (fixed as 60)

min\_samples\_leaf: (0.5\*min\_samples\_split) (fixed as 30)

Calculated by the formula (by Pro.Wang)

3. Experiment of post-pruning the decision tree





(A) (B)

fig1. Post-prune Exp. results(A ; bairong B: DPFINAL)

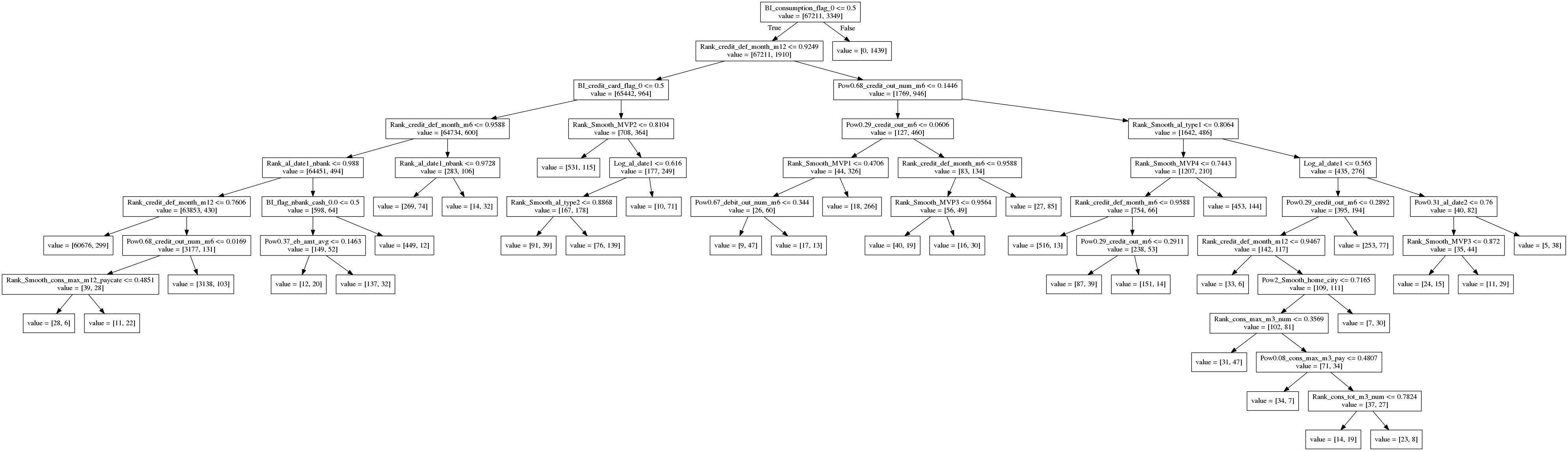


fig2. Final post-pruned tree(bairong)

4. Conclusions（with no VIP features Exp.）

Table1. Comparison of control, tuning and pruning group(bairong)



Table2. Comparison of control, tuning and pruning group (DPFINAL)



1) Pruning is more faster than tuning(67 vs. 632 sceonds)

2) In the dataset of bairong, the performance of test set in pruned model is better than the tuned model, however, the performance of training set is not so good as tuned model(maybe still over-fitting in the tuned model).

In the dataset of DPFINAL, the tuned tree seems to good enough,after post-tune, the performance of the model has deteriorated slightly(also can be seen in fig1 B)

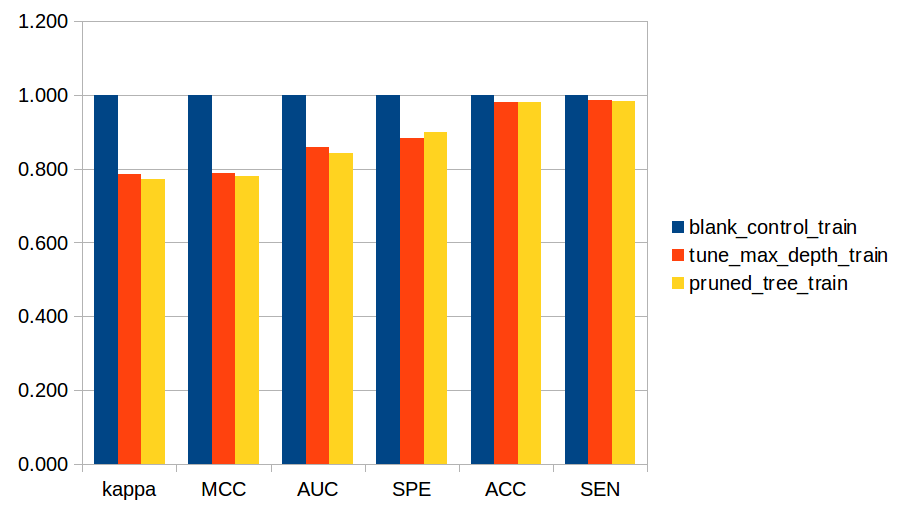
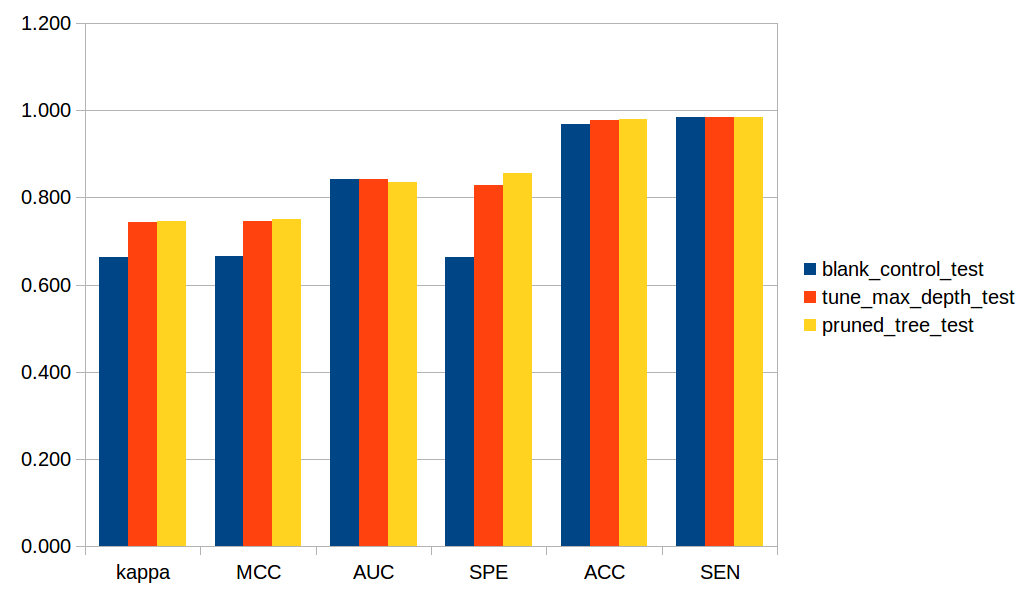


Fig 3. Performance in test and training set(bairong)

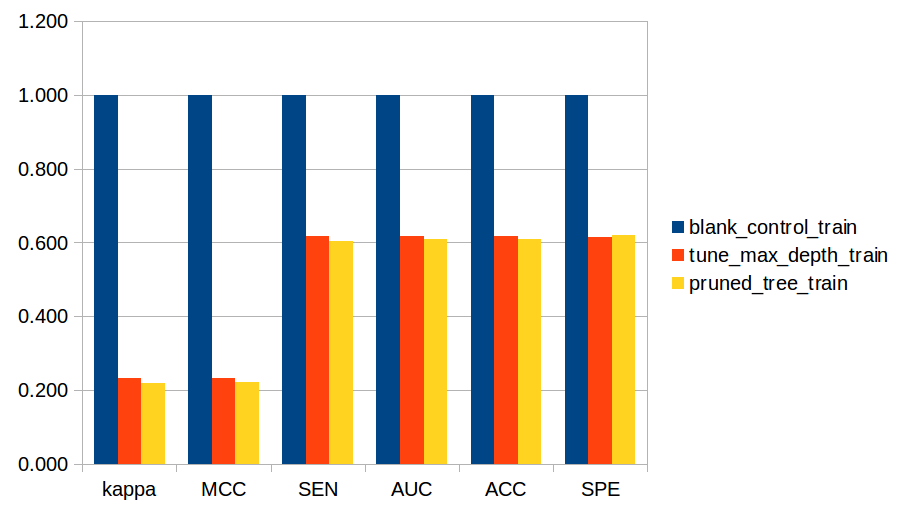
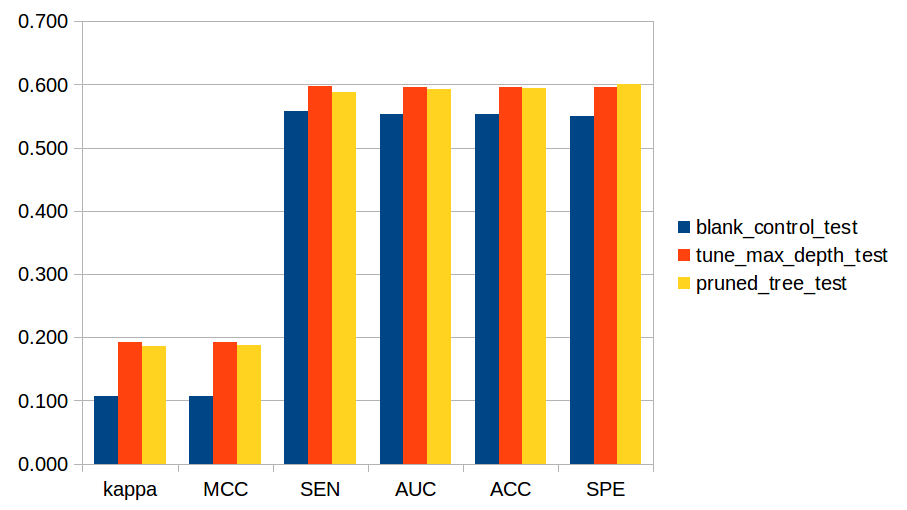


Fig 4. Performance in test and training set(DPFINAL)

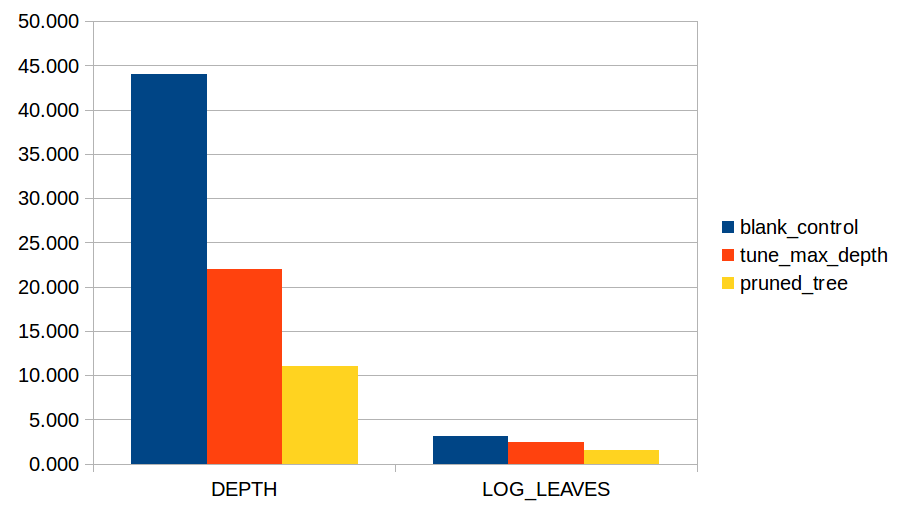
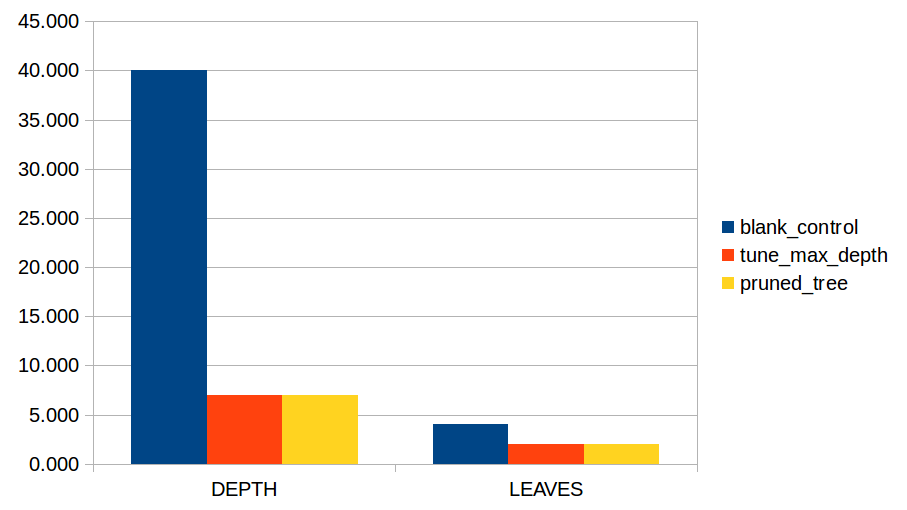


Fig 5 depth and leaves of the trees (bairong and DPFINAL)

3) Both tuned and pruned model are better than blank control