The Collaborative Center for Statistics in Science C^2S^2

Search for the Smallest Random Forest

Heping Zhang

Yale University

Joint Work with Minghui Wang

Outline

- Background
- Goal
- Key idea
- Method
- Simulation
- Application

Background

Random forests have emerged as one of the most commonly used nonparametric statistical methods in many scientific areas, particularly in analysis of high throughput genomic data.

Background

- A general practice in using random forests is to generate a sufficiently large number of trees, although it is subjective as to how large is sufficient.
- Furthermore, random forests are viewed as a “black-box” because of its sheer size.

Goal

- Explore whether it is possible to find a common ground between a forest and a single tree
  - retain the easy interpretability of the tree-based methods
  - avoid the problems that the tree-based methods suffer from.
- Does a forest have to be large, or how small can a forest be?

Key idea

- Shrink the forest with two objectives
  - maintain a similar (or even better) level of prediction accuracy
  - reduce the number of the trees in the forest to a manageable level

Method

- Three measures are considered to determine the importance of a tree in a forest
  - by prediction
  - by similarity
  - by restricted similarity

Method

- “by prediction” method
  - focuses on the prediction
  - A tree can be removed if its removal from the forest has the minimal impact on the overall prediction accuracy.

Method

- "by prediction" method
  - For tree *T* in forest *F*, calculate the prediction accuracy of forest $F_{(-T)}$ that excludes *T*.
  - $\Delta_{(-T)}$ represents the difference in prediction accuracy between *F* and $F_{(-T)}$.
  - The tree with the smallest $\Delta_{(-T)}$ is the least important one and hence subject to removal.

Method

- “by similarity” method
  - is based on the similarity between two trees.
  - A tree can be removed if it is “similar” to other trees in the forest.

Method

- “by similarity” method
  - The correlation of the predicted outcomes by two trees gives rise to a similarity between the two trees.
  - For tree *T*, the average of its similarities with all trees in $F_{(-T)}$, denoted by ρ_T, reflects the overall similarity between *T* and $F_{(-T)}$.
  - The tree with the highest ρ_T is the most similar to the trees in $F_{(-T)}$ and hence subject to removal.

Method

- "by restricted similarity" method
  - is based on the weighted similarity between two trees.
  - A tree can be removed if it is "similar" to other trees in the forest.

Method

- “by restricted similarity” method
  - Evaluate the pairwise similarity of two trees in forest *F*, according to their predicted outcomes.
  - Select the pair of trees being most similar.
  - Calculate $\wp_T$ for the two trees and the one with higher $\wp_T$ is subject to removal.
  - Distribute the weight of *T* to all other trees in $F_{(-T)}$, proportional to the pairwise similarity in $\wp_T$.

Method

- Select the optimal size sub-forest
  - Let *h(i)*, *i*=1,…N_f-1, denote the performance trajectory of a sub-forest of *i* trees
    - N_f is the size of the original random forest.
  - If we have only one realization of *h(i)*, we select the optimal size sub-forest by maximizing *h(i)* over *i*=1,…N_f-1.
  - If we have multiple realizations of *h(i)*, we select the optimal size sub-forest by using the 1-se rule.
- The size of this smallest sub-forest is called the critical point of the performance trajectory.

Simulation Designs

- Simulation Designs
  - For each data set, we generated 500 observations, each of which has one response variable and 30 predictors from Bernoulli distribution with success probability of 0.5.
  - Chose ν of the 30 variables to determine the response variable.

$$y = \begin{cases} 1, & \text{if } \sum_{i=1}^{v} X_i \;/ v + \sigma > 0.5, \\ 0 & \text{Otherwise.} \end{cases}$$

- Where σ is a random variable following the normal distribution with mean zero and variance .
  - Considered two choices for ν (5 and 10) and two choices of σ (0.1 and 0.3).

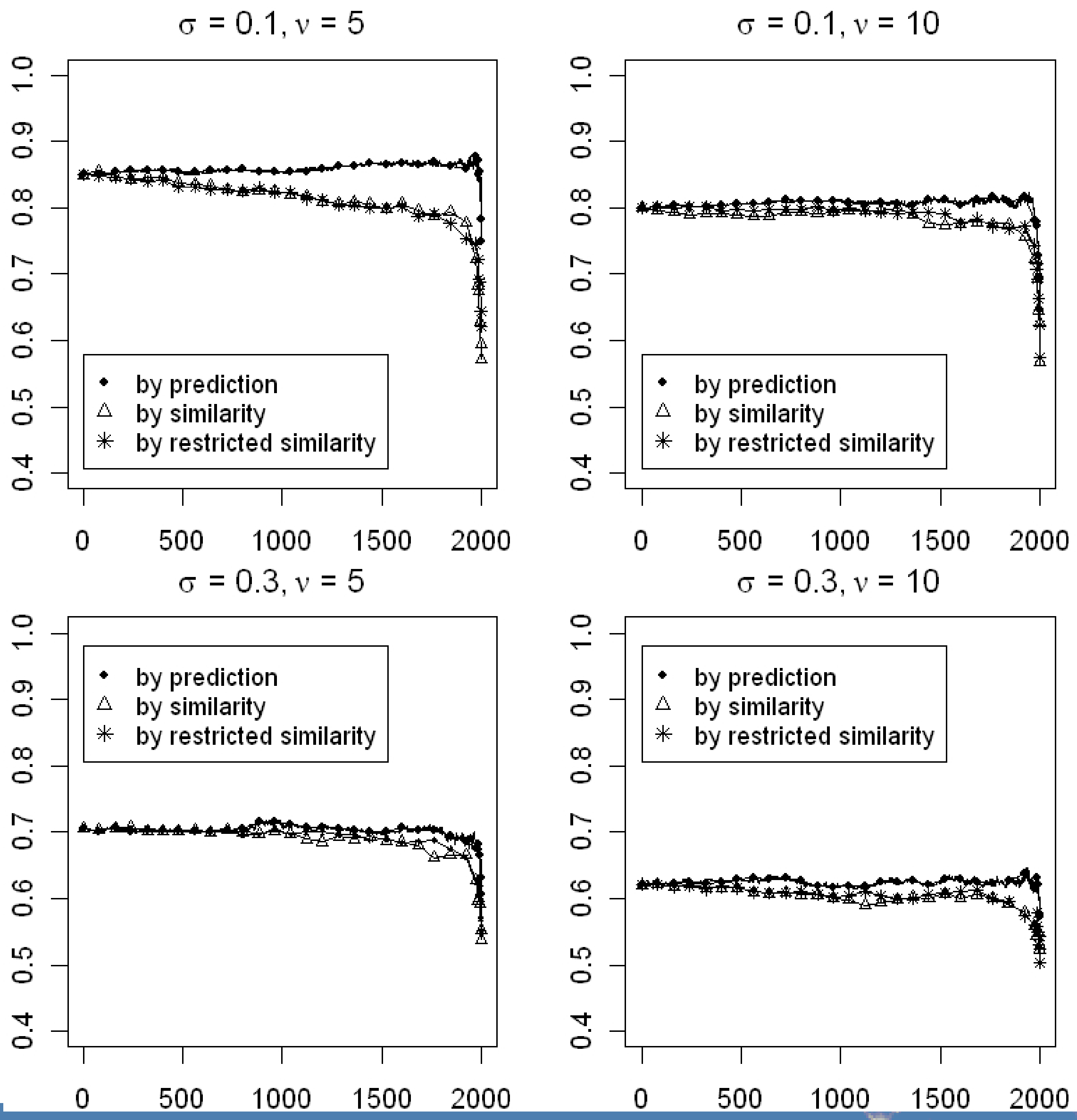
Simulation Designs

- To perform an unbiased comparison of the three tree removal measures, we simulated three independent data sets
  - The training set is used to train the initial random forest
  - The execution set is used to delete trees from the initial forest to produce sub-forests
  - The evaluation set is used to evaluate the prediction performance of the sub-forests
- The generation and use of these three data sets constituted one run of simulation, and we replicated 100 times.

Simulation Results

- Randomly selected one run of simulation and presented the stepwise change in the prediction performance in Figure 1.
- The “by prediction” method is preferable
  - It can identify a critical point during the tree removal process in which the performance of the sub-forest deteriorates very rapidly.
- The performance of the sub-forests may begin to improve before the critical point.

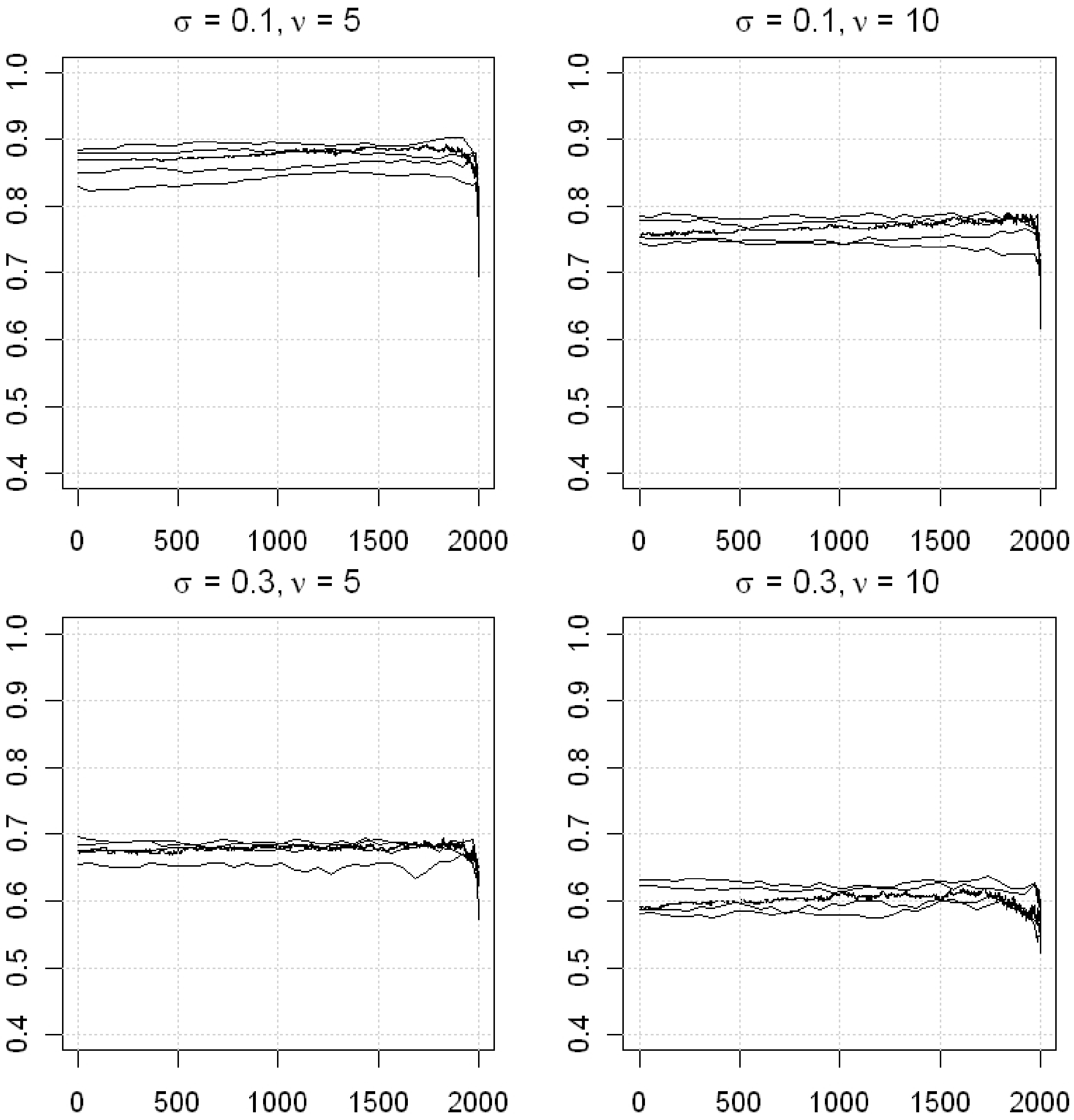
Prediction performance of sub-forests produced from different datasets and methods



Simulation Results

- Figure 2 displays a summary plot of prediction performance using the results in five randomly selected runs.
- Although the variation of the trajectories is notable, the sizes of the optimal sub-forests are within a reasonable range (11-36) for the “by prediction” method.

Performance trajectory of the “by prediction” method using the results in five randomly selected runs for four data sets.



The medians of the numbers of trees in the optimal sub-forests in 100 replications.

σ	v	
	5	10
0.1	20(13, 29)	31(20, 40)
0.3	22(15, 32)	18(11, 37)

Simulation Designs

- In practice, we generally have one data set only.
- May not have the execution and evaluation data sets as in previous simulation.
- How do we select the optimal sub-forest with only one data set?

Simulation Designs

- Considered four bootstrap-based approaches and examined them in simulated data sets.
- We have the “golden” standard to be compared with in the simulated data set.

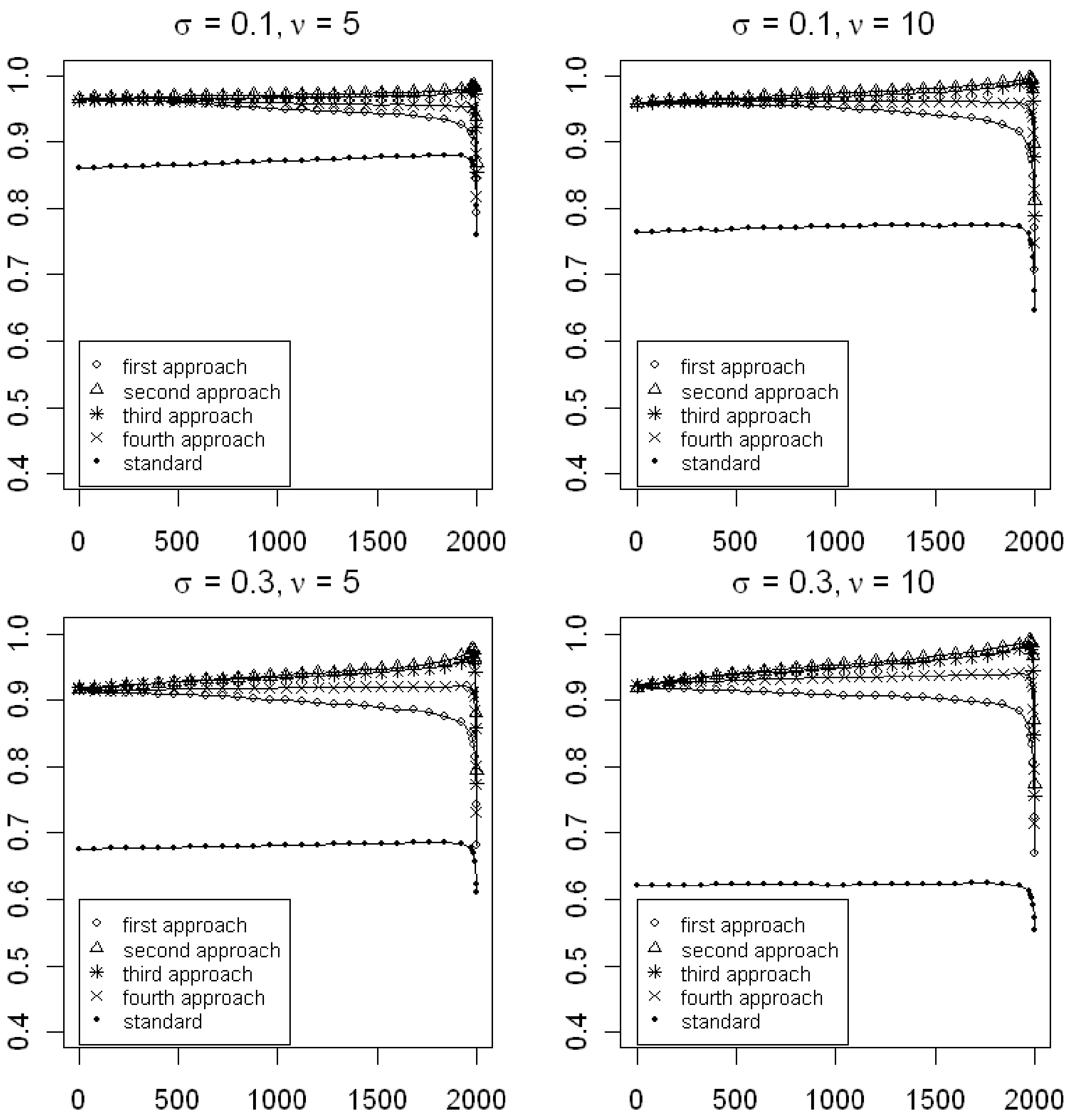
Simulation Designs

- After constructing an initial forest using the whole data set as the training data set
  - use one bootstrap data set for execution and the out-of-bag (oob) samples for evaluation.
  - use the oob samples for both execution and evaluation.
  - use the bootstrap samples for both execution and evaluation.
  - re-draw bootstrap samples for execution and re-draw bootstrap samples for evaluation.

Simulation Results

- Figure 3 compares the performance of the four bootstrap-based approaches in the four simulation data sets.
- The comparison is based on the average performance in 100 runs.

A performance summary plot of the “by prediction” method



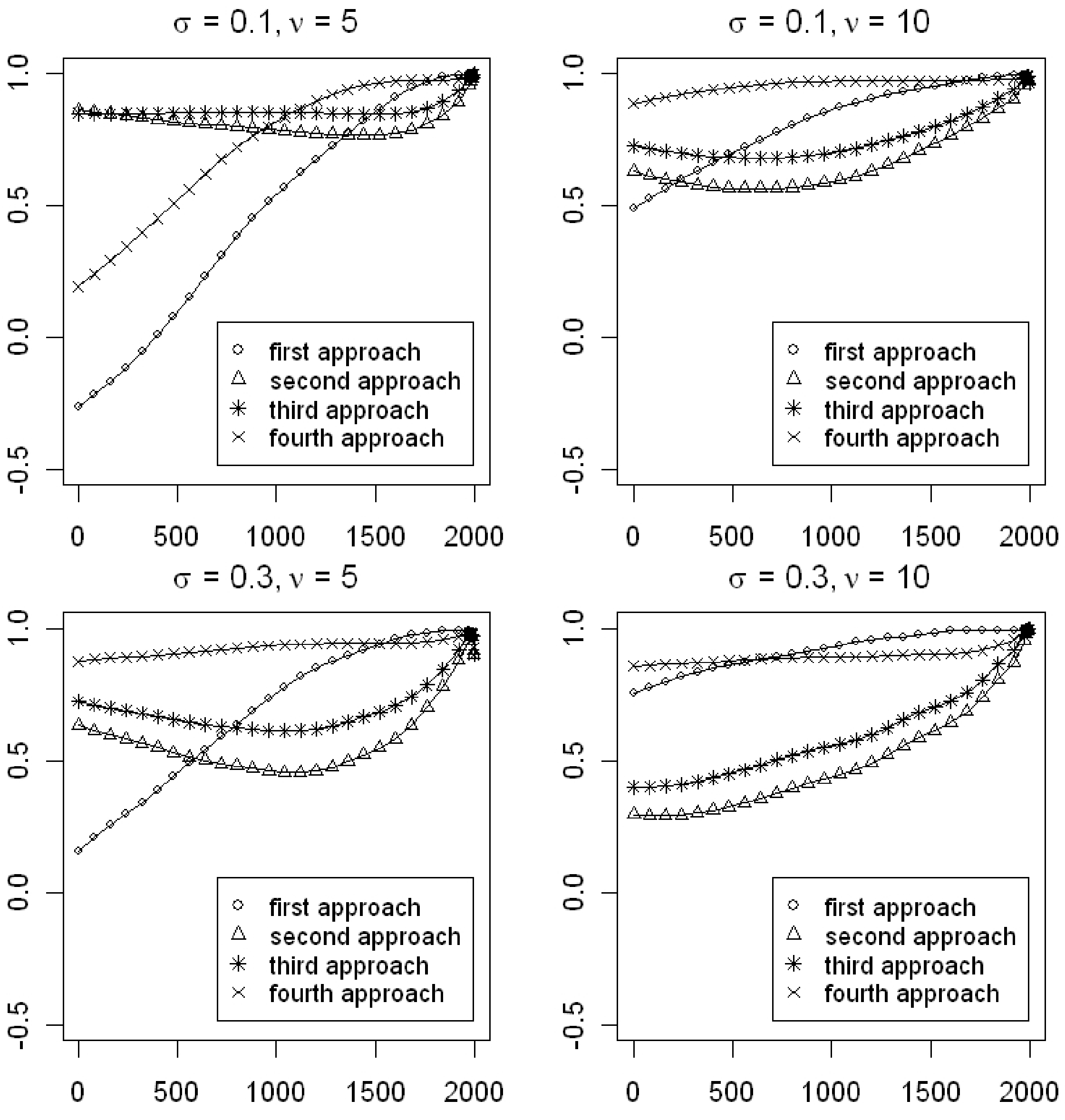
Simulation Results

- The performance trajectories of the four bootstrap-based approaches may not overlap with the “golden” standard.
- For the selection of the optimal sub-forest, the similarity among the trajectories is most relevant, because it could lead to the same or similar sub-forest.

Simulation Results

- In Figure 4, we examined the correlation between the original (the “golden” standard) trajectory and each of the four bootstrap approaches.

The correlation between the performance trend by each of the four bootstrap strategies and the "standard" curve



Simulation Results

- Using the bootstrap samples for execution and the oob samples for evaluation is an effective sample-reuse approach to selecting the optimal sub-forest.

Application

- Dataset
  - the microarray data set of a cohort of 295 young patients with breast cancer, containing expression profiles from 70 previously selected genes.
  - previously studied by van de Vijver *et al*.
- The responses of all patients are defined by whether the patients remained disease-free five years after their initial diagnoses or not.

Application

- Method used
  - The “by prediction” measure
  - The original data set to construct an initial forest
  - A bootstrap data set for execution
  - The oob samples for evaluation.
- The procedure is replicated for a total of 100 times.
  - The oob error rate is used to compare the performance of the initial random forest and the optimal sub-forest.
  - The sizes of the optimal sub-forests fall in a relatively narrow range, of which the 1st quartile, the median, and the 3rd quartile are 13, 26 and 61, respectively.

Application

- The smallest optimal sub-forest in the 100 repetitions with the size of 7 is selected.
- As a benchmark, we used the 70-gene classifier proposed by Vijver, *et al*.

Application

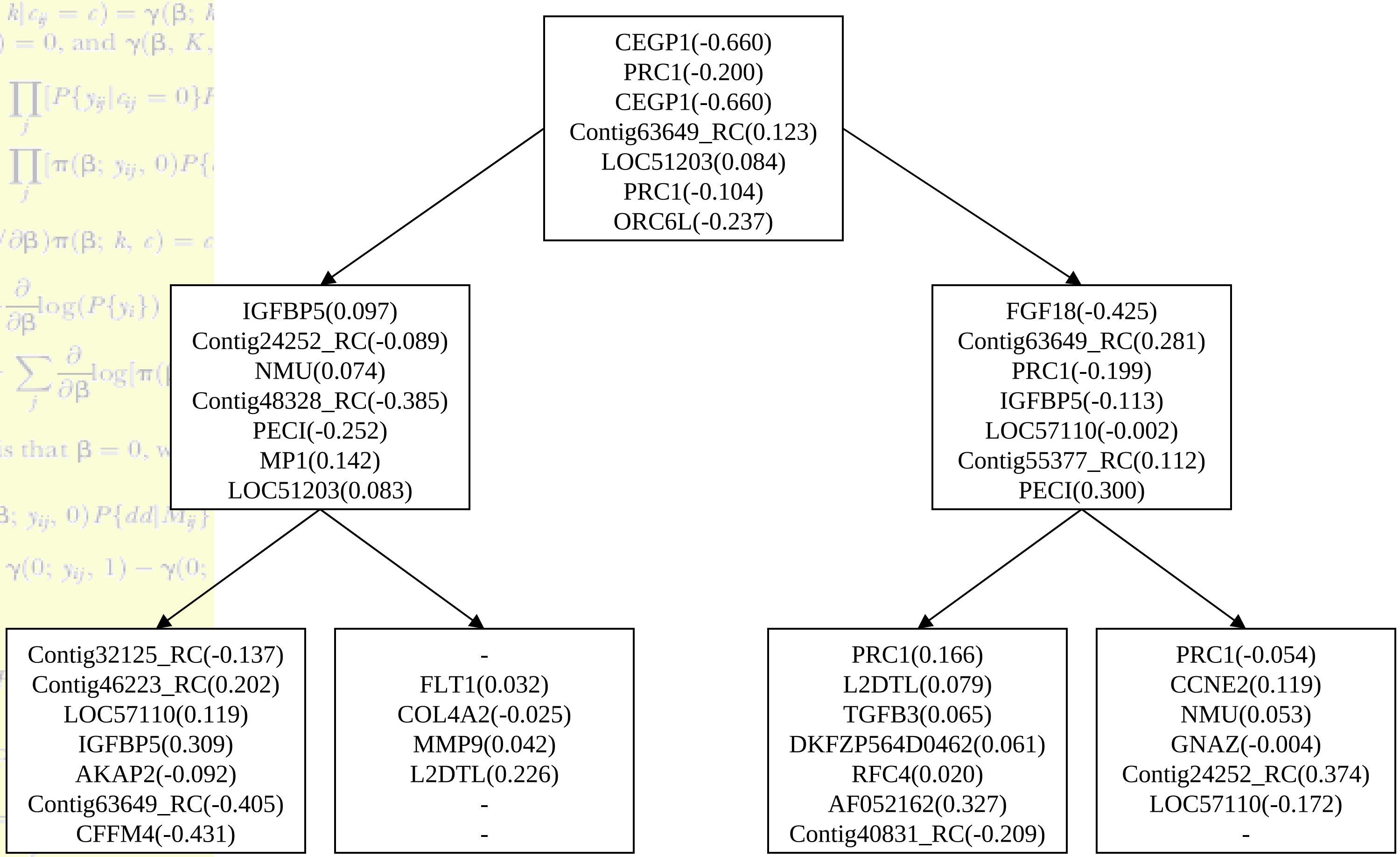
- Table 2 presents the misclassification rates based on the oob samples.
  - The initial forest and the optimal sub-forest achieve almost the same level of performance accuracy.
  - The 70-gene classifier has an out-of-bag error rate which is much higher than those of the forests.

Comparison of prediction performance of the initial random forest, the optimal sub-forest, and a previously established 70-gene classifier

Method	Error rate	True Predicted	Good	Poor
Random Forest	26.0%	Good	141	17
		Poor	53	58
Sub-forest	26.0%	Good	146	22
		Poor	48	53
70-gene Classifier	35.3%	Good	103	4
		Poor	91	71

Application

- Main motivation
  - seek the smallest possible forest to enable us to examine the forest.
- Figure 5 displays the most critical part (the top three layers) of the optimal sub-forest consisting of the seven trees.
- The selected genes are quite diverse and unique.



CEGP1(-0.660)
PRC1(-0.200)
CEGP1(-0.660)
Contig63649_RC(0.123)
LOC51203(0.084)
PRC1(-0.104)
ORC6L(-0.237)

IGFBP5(0.097)
Contig24252_RC(-0.089)
NMU(0.074)
Contig48328_RC(-0.385)
PECI(-0.252)
MP1(0.142)
LOC51203(0.083)

FGF18(-0.425)
Contig63649_RC(0.281)
PRC1(-0.199)
IGFBP5(-0.113)
LOC57110(-0.002)
Contig55377_RC(0.112)
PECI(0.300)

Contig32125_RC(-0.137)
Contig46223_RC(0.202)
LOC57110(0.119)
IGFBP5(0.309)
AKAP2(-0.092)
Contig63649_RC(-0.405)
CFFM4(-0.431)

-
FLT1(0.032)
COL4A2(-0.025)
MMP9(0.042)
L2DTL(0.226)
-
-

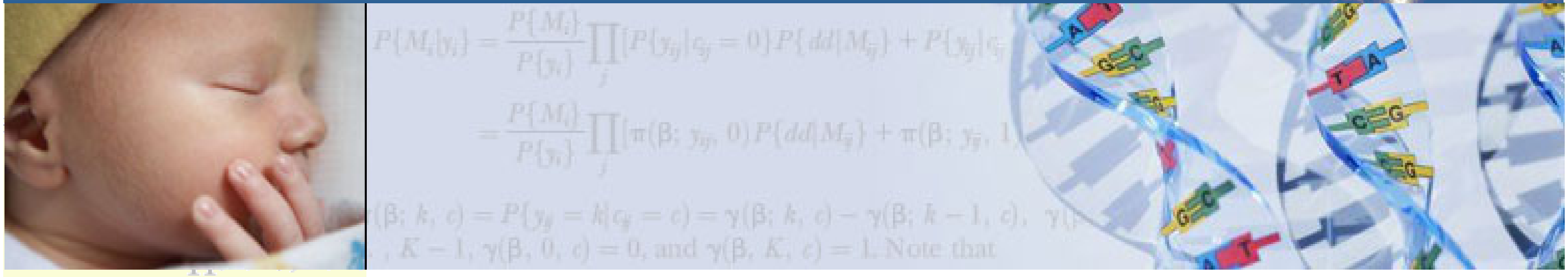
PRC1(0.166)
L2DTL(0.079)
TGFB3(0.065)
DKFZP564D0462(0.061)
RFC4(0.020)
AF052162(0.327)
Contig40831_RC(-0.209)

PRC1(-0.054)
CCNE2(0.119)
NMU(0.053)
GNAZ(-0.004)
Contig24252_RC(0.374)
LOC57110(-0.172)
-

The top three layers of the optimal sub-forest consisting of seven trees

Conclusion

- It is possible to construct a highly accurate random forest consisting of a manageable number of trees.
  - the size of the optimal sub-forest is in the range of tens
  - some sub-forests can even over-perform the original forest in terms of prediction accuracy
- The key advantage
  - the ability to examine and present the forests.
- The limitation
  - future samples and studies are needed to evaluate the performance of the forest-based classifiers.



Thank You!