

Variable **Ranking** and **Selection** via Random Forests

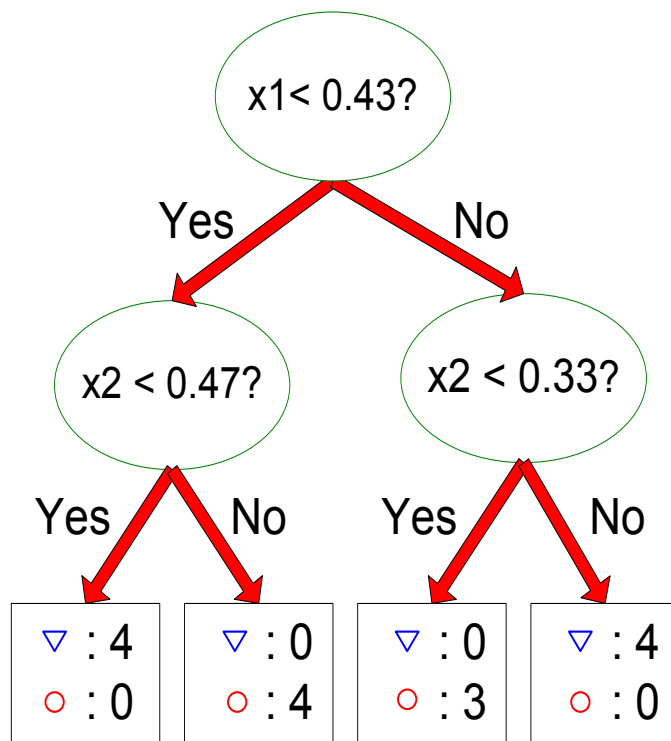
Houtao Deng

Overview

- Objectives
 - Advantages and Disadvantages of Random Forests on variable selection and ranking
 - New improvements
- Agenda
 1. Random Forests
 2. Variable ranking
 3. Variable selection

Random Forests

Decision tree



Splits are often based on information **gain** measured by entropy or Gini index .

Gini index of Y:

$$G(Y) = \sum_i 1 - p_i^2$$

Gini index after observing X:

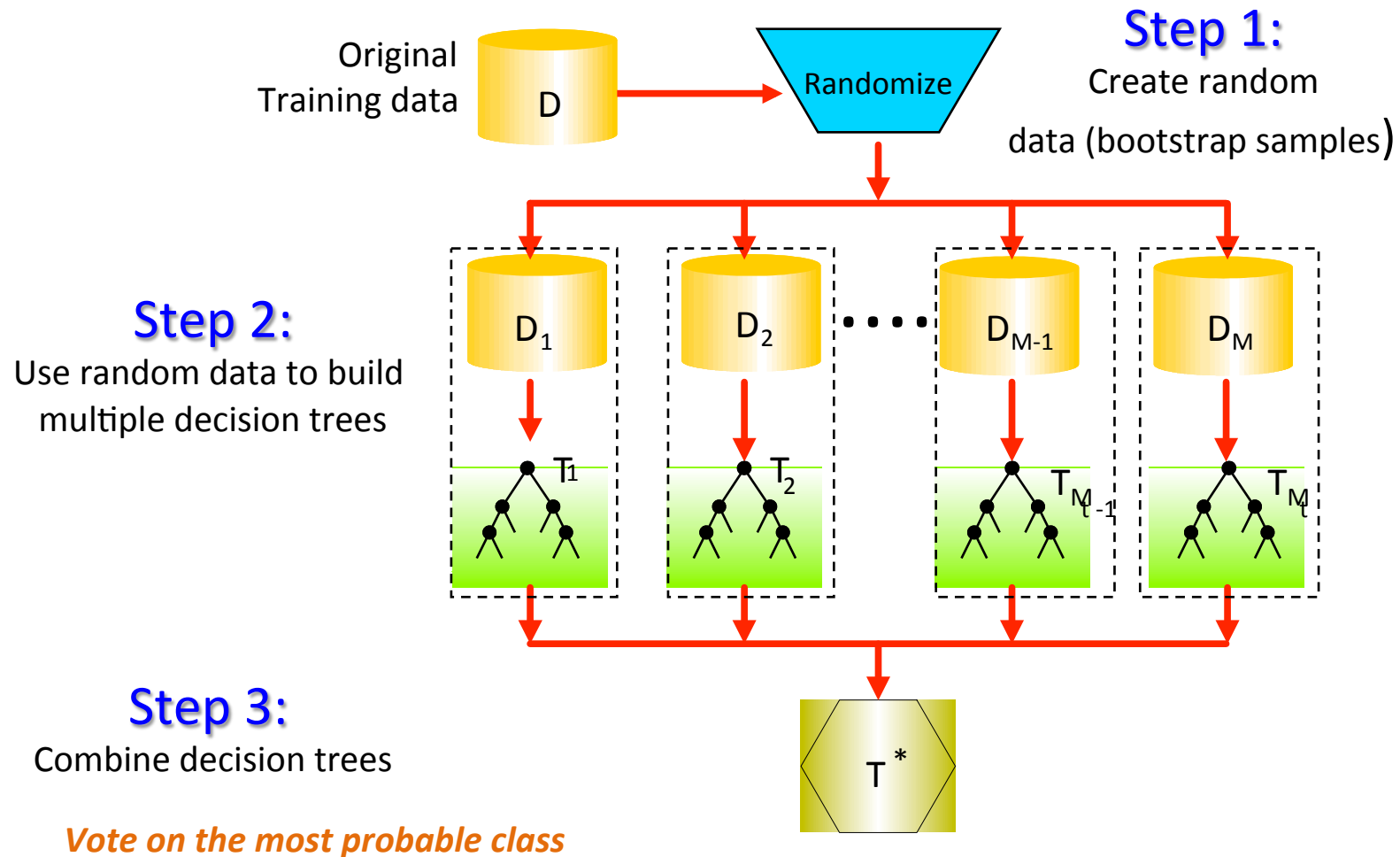
$$G(Y|X) = \sum_i P(X=v_i) G(Y|X=v_i)$$




Gini decrease:














































$$\Delta I = G(Y) - G(Y|X)$$

Random Forests

Random forests



Some characteristics of different learning methods. Key:  = good,  = fair, and  = poor.

Characteristic	Neural Nets	SVM	Trees	Random forest	k-NN, Kernels
Natural handling of data of “mixed” type					
Handling of missing values					
Robustness to outliers in input space					
Insensitive to monotone transformations of inputs					
Computational scalability (large N)					
Ability to deal with irrelevant inputs					
Ability to extract linear combinations of features					
Interpretability					
Predictive power					

Variable Ranking

Overview of variable ranking

- In supervised learning (classification and regression), variable ranking is to rank the predictor variables according to their contributions regarding predicting the response variable.
- The key for variable ranking is to define an importance metric for each variable

Variable Ranking

Importance metrics

- Univariate:
 - Information Gain, Gini Index, Chi-square, correlation coefficient, etc.
 - Disadvantages: consider one variable's contribution without other variables' influences
- Logistic regression
 - $\ln[p(Y=1)/(1-p(Y=1))] = \alpha + \beta X$, where $\beta = (\beta_1, \beta_2, \dots, \beta_K)$ can be used to evaluate the importance of each variable
 - Disadvantages:
 - Assumption of linearity in terms of the logit function versus the predictor variables
 - Different units of variables make it a little bit harder to interpret the coefficients
 - Categorical variables usually need to be transformed to binary variables
- Decision trees like C4.5 and CART
 - Decision trees intrinsically evaluate each variable's importance when selecting the best variable to split.
 - Advantages:
 - Handle nonlinear interactions, different variable units, mixed numerical and categorical variables, etc.
 - Disadvantages:
 - The accuracy of C4.5/CART is limited, and thus the importance score from these classifiers may be not reliable

Variable Ranking

Variable importance score from random forests

- Gini index:

- Gini index Y: $G(Y) = \sum_i 1 - p_i^2$
- $G(Y|X) = \sum_i P(X=v_i) G(Y|X=v_i)$
- Gini decrease due to observing X: $\Delta I = G(Y) - G(Y|X)$

- For a single tree:

$$VI(X_i, T) = \sum_{t \in T} \Delta I(X_i, t)$$

Gini decrease due to X_i
at node t of tree T

- For random forest

$$E(X_i) = \frac{1}{M} \sum_{m=1}^M VI(X_i, T_m)$$

Average over all trees

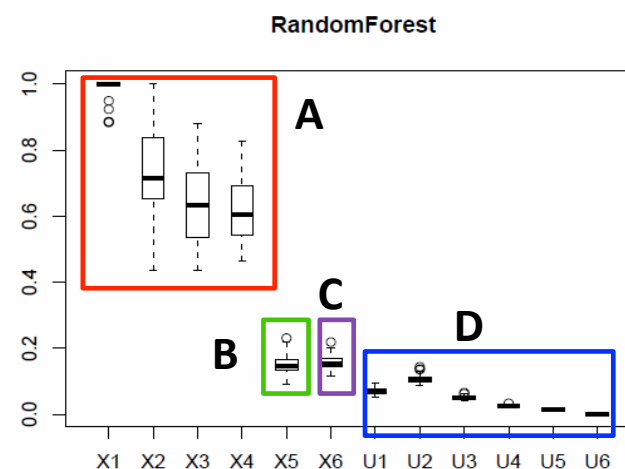
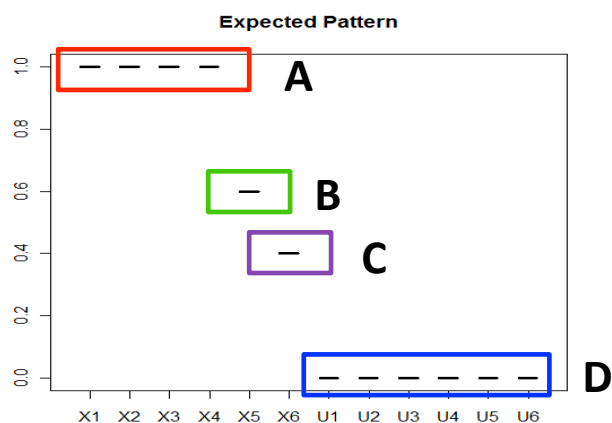
- Advantages

- Accurately capture the relationship between predictor variables and the response variable
- Scale independent
- Naturally handle categorical and numerical variables

Variable Ranking

Bias of Random Forest for multi-valued categorical variables

- Random Forest importance score is biased toward the variables having more values [H. Deng et al. 2011]



Importance:

- $I(X1)=I(X2)=I(X3)=I(X4) > I(X5)>I(X6)>I(Ui)$

Variable types:

- X1 and U1: continuous; Others are categorical

Number of levels:

- $L(X2)>L(X3)>L(X4);$
- $L(X6)>L(X5);$
- $L(U2)>L(U3)>...>L(U6)$

Variable Ranking

Solution 1: OOB Forest [H. Deng, G. Runger 2011]

■ Notation:

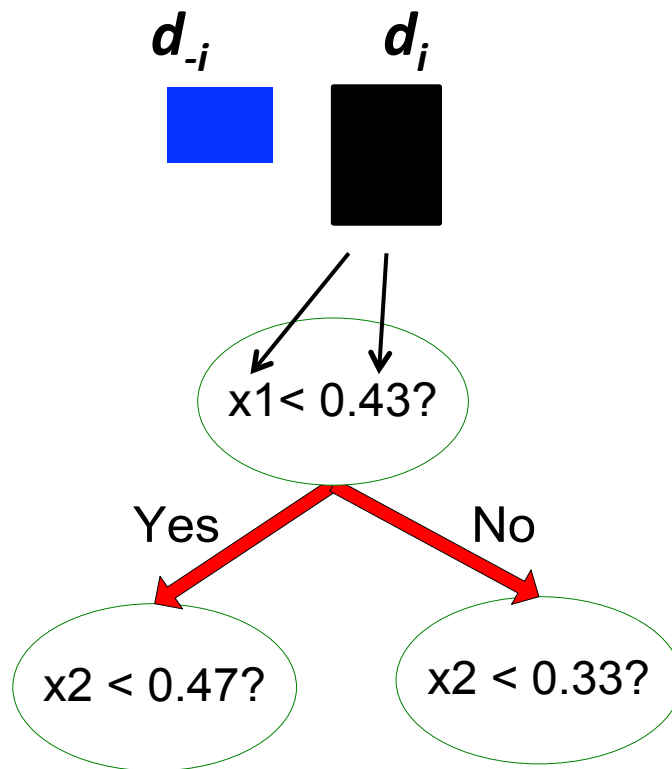
- A random sample \mathbf{d}_i from the training data \mathbf{D} is used growing a random tree i .
- $\mathbf{d}_{-i} = \mathbf{D} - \mathbf{d}_i$ (out-of-bag)

■ Intuition:

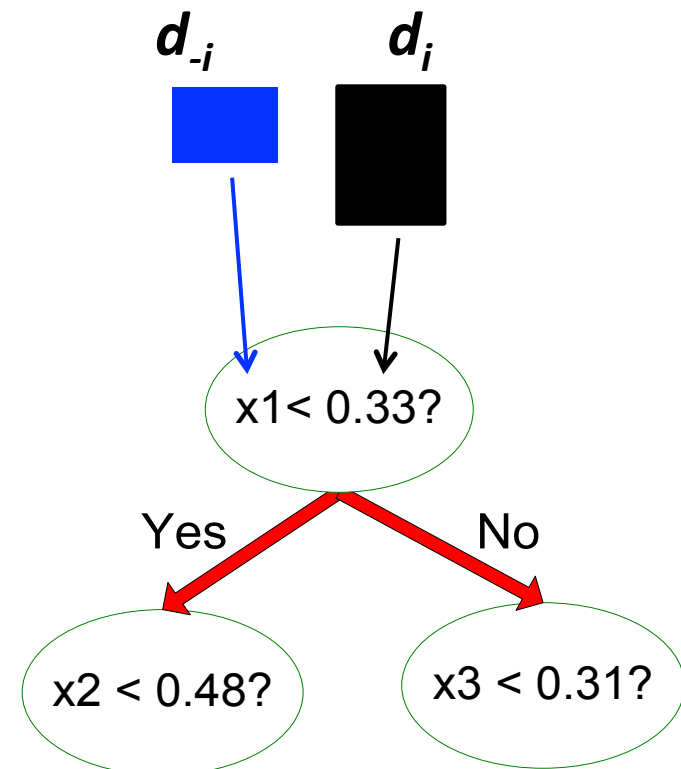
- Random Forest is greedy in that: \mathbf{d}_i is used for selecting the splitting variable at each node, but also used for selecting the splitting value of that variable.
- OOB Forest is try to make the RF less greedy by using \mathbf{d}_{-i} to decide the splitting variable and using \mathbf{d}_i to decide the splitting value of that variable

Variable Ranking

Solution 1: OOB Forest



Random Forest



OOB Forest

Variable Ranking

Solution 2: pForest [H. Deng, G. Runger 2011]

- Intuition:

- An informative variable X_i is expected to have a larger RF importance score than its randomly permuted version Z_i
- However, if Z_i is a completely permuted version, any X_i can easily beat Z_i . And thus partial permutation of X_i is used in this approach: permute δ (0%-100%) of the rows of X_i

- Procedure:

- Permute δ (percentage) rows of variable X_i to Z_i
- Augment the original data set $[X,Y]$ to $[X,Z,Y]$, where X and Z are the predictor variables, and Y is still the response variable
- Build R replicates of random forests on $[X,Z,Y]$; and calculate the importance of X_i as:

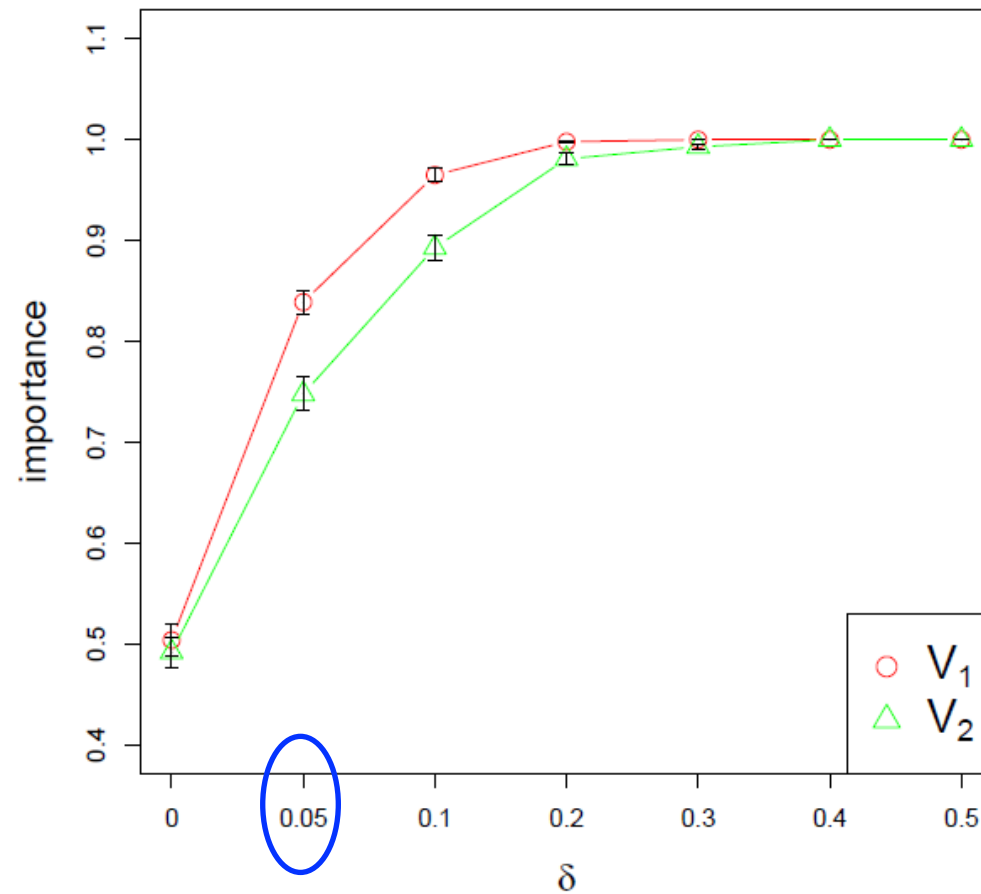
$$\text{Imp}(X_k) = \frac{1}{R} \sum_{r=1}^R I\{RF^r(X_k) > RF^r(Z_k)\}$$

I : indicator function

X2	Partial permute $\delta = 60\%$	Z2
5	→	2
2		4
4		5
1		1
7		7

Variable Ranking

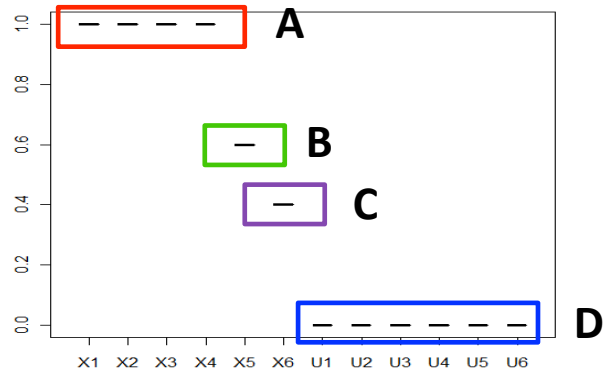
Selection of δ of pForest



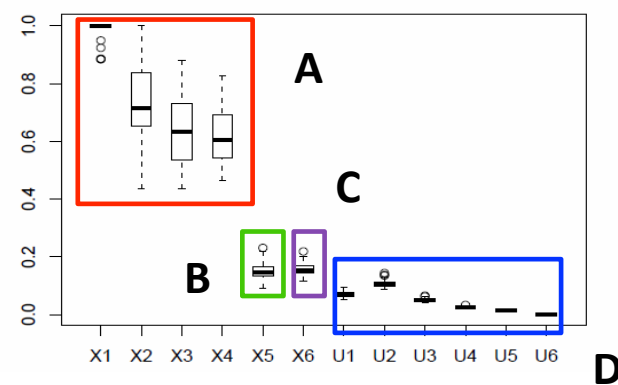
Variable Ranking

Experiments and Results

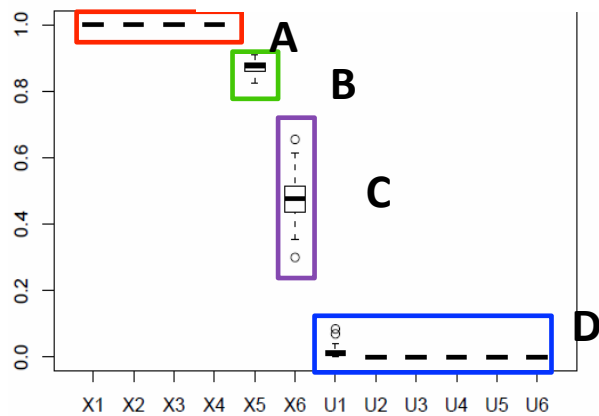
Expected Pattern



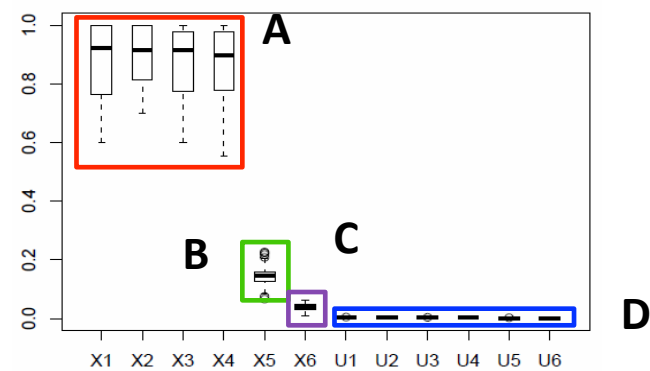
Random Forest



pForest



OOB Forest



Variable Selection

Variable selection: Markov blanket criterion

Let F be a full set of variables. Given a target variable T , let $MB(T) \subset F$ and $T \notin MB(T)$, $MB(T)$ is said to be a Markov Blanket (MB) for T if $T \perp (F - MB) | MB$. That is, T is conditionally independent of other features given MB .

- Objective: find a minimum set of variables that contain all the information about predicting Y .
- In practice, empirical errors from a classifier are often used as the evaluation criterion:
less error rate \rightarrow more information in the variables
- 2^P combinations of variable subsets.

Variable Selection

Relevancy and redundancy

- **Relevancy**: contribution of a predictor variable regarding predicting the response variable
- **Redundancy**: similarity between two predictor variables regarding predicting the response variable

Variable Ranking

X_1
 X_4 (similar to X_1)
 X_3
 X_{10}
 X_8
 X_{12} (unimportant)

Variable Selection

X_1
 X_3
 X_{10}
 X_8

Variable Selection

Filter methods

- Independent of a particular learner
- Information-based algorithms:
 - Relevancy: $I(Y, X_k)$; Redundancy: $I(X_i, X_k)$
 - Many algorithms use a combination of $I(Y, X_k)$ and $I(X_i, X_k)$ as the objective function to search the best subset of variables
 - CFS: correlation-based feature selection [M. Hall, 2000]
 - FCBF: fast correlation-based filter [L. Yu, etc. 2003]
 - *mRMR*: minimum-redundancy maximum-relevance *feature selection* [H. Peng, etc. 2003]
 - Disadvantages: can capture **only two-way interactions** between variables. e.g. Relationship(X1,X2) or Relationship(Y,X1). But they are not able to capture the relationship like: $Y = \text{XOR}(X1, X2)$, in which neither X1 nor X2 individually is predictive, but X1 and X2 together can correctly determine Y.

Variable Selection

Embedded methods

- LASSO (least absolute shrinkage and selection operator) [R. Tibshirani, 1996] :
 - Advantage: one-pass, i.e. only need to build one model
 - Disadvantage: LASSO is not among the most accurate classifiers in general, and thus the variables selected may be not the most informative sub set
- Recursive Feature Elimination Framework
 - Consists of multiple iterations. At each iteration, drop the least important variable.
 - Strong classifiers can fit into this framework:
 - Support vector machine (drop the variable with the least coefficient of the learned hyperplane) [I. Guyon, etc. 2002]
 - **varSelRF**(drop the variable with the least importance score) [R. Uriarte, etc. 2006]
 - Disadvantage: computationally expensive: potentially need to build ***P*** models for ***P*** predictor variables.

Variable Selection

Regularized random forest (RRF) [H. Deng, G. Runger 2012]

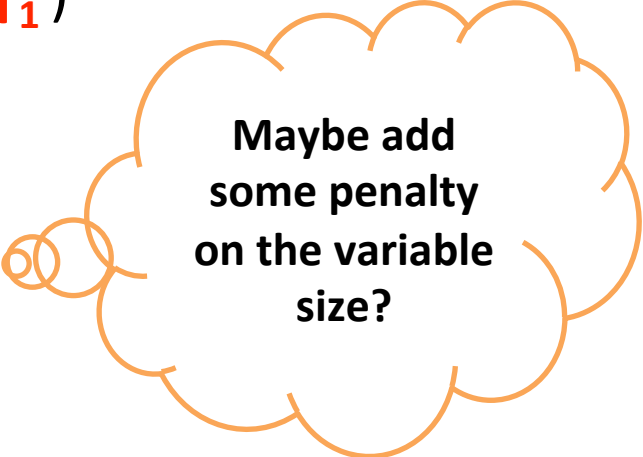
- A one-pass embedded method based on a strong classifier is very desirable

- RRF was inspired by LASSO:

$$\text{MIN} (\text{empirical loss} + \lambda ||\beta||_1)$$

- Ordinary Random Forest:

$$\text{MAX} (\text{information gain})$$



Maybe add
some penalty
on the variable
size?

Variable Selection

Regularized Random Forest (RRF)

- Intuition
 - Avoid selecting a new variable that is similar to the variables already selected in previous tree nodes.
- Procedure
 - A regularized information gain is considered:

$$gain_R(X_j) = \begin{cases} \lambda \cdot gain(X_j) & X_j \notin F \\ gain(X_j) & X_j \in F \end{cases}$$

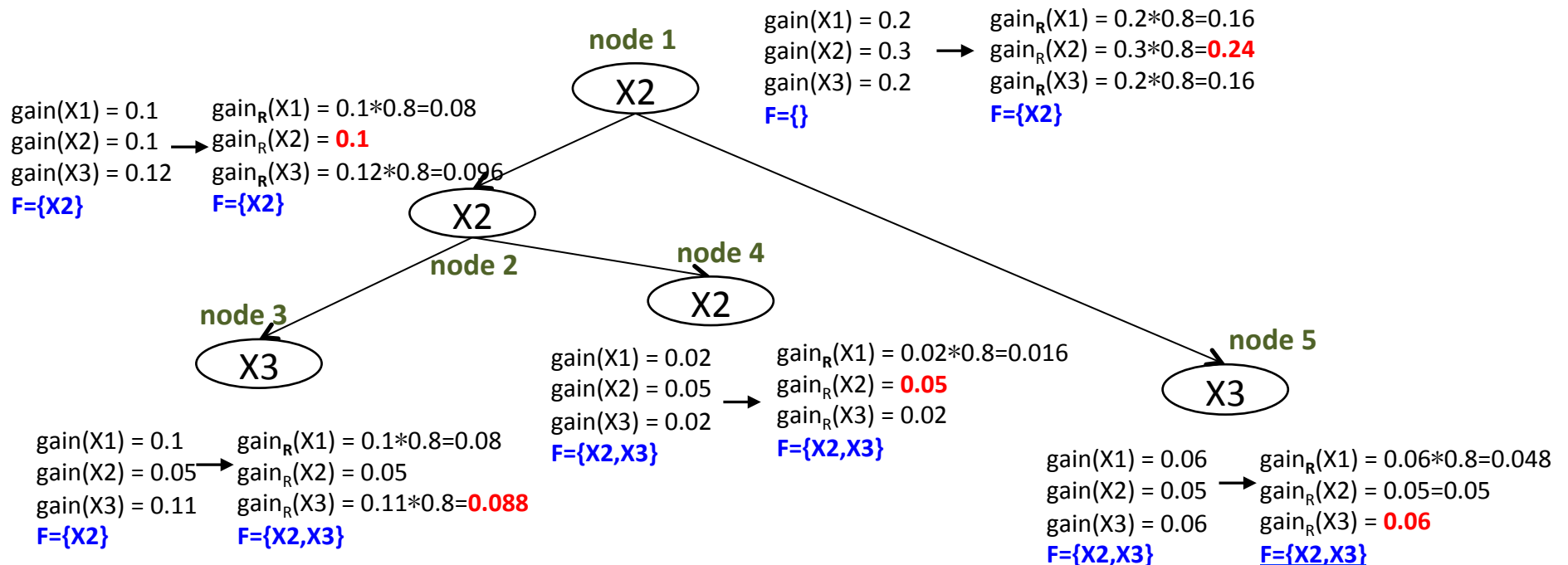
where F is the variable set selected in previous nodes, $gain(X)$ is the information gain measure in ordinary trees, $\lambda \in [0, 1]$ is called the coefficient. A smaller λ leads to a larger penalty and thus fewer variables are expected to be selected.

- The tree regularization framework can be easily applied to many tree models.

Variable Selection

Regularized Random Forest (RRF)

- Illustration of the tree regularization (for simplicity, random sampling features is not considered here):
X1, X2, X3, with $\lambda=0.8$, a **depth-first** tree (leaf nodes not shown): node 1 \rightarrow node 2 \rightarrow node 3 \rightarrow node 4 \rightarrow node 5



Variable Selection

Guided Regularized Random Forest (GRRF) [H. Deng, G. Runger 2012]

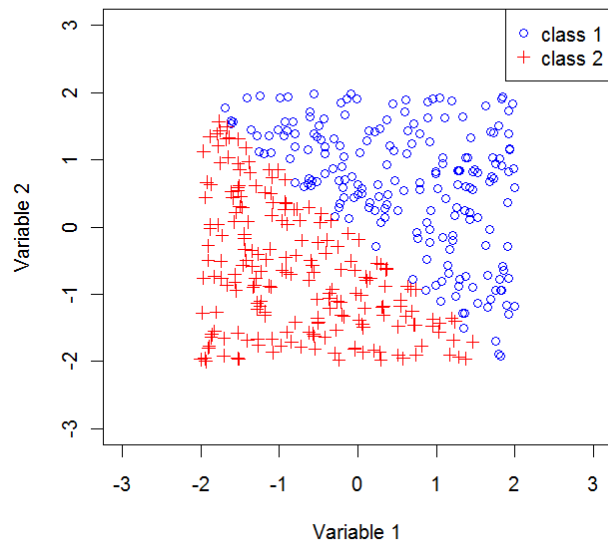
- Motivation: the variable selection process is greedy in that variables are evaluated at a local node.
- Intuition: Use a less-greedy importance score to give the important variables advantages
- Procedure:
 - Build an ordinary random forest and get the importance score for each variable: $imp_1^0, imp_2^0, imp_3^0, \dots, imp_p^0$
 - Normalize the importance score: $imp_i = \frac{imp_i^0}{\max_{j=1}^P imp_j^0}$
 - Calculate the guided coefficient: $\lambda_i = (1 - \gamma) * \lambda_0 + \gamma * Imp_i$
 - The regularized information gain: $Gain_R(X_i) = \begin{cases} \lambda_i \cdot Gain(X_i) & X_i \notin F \\ Gain(X_i) & X_i \in F \end{cases}$

Variable Selection

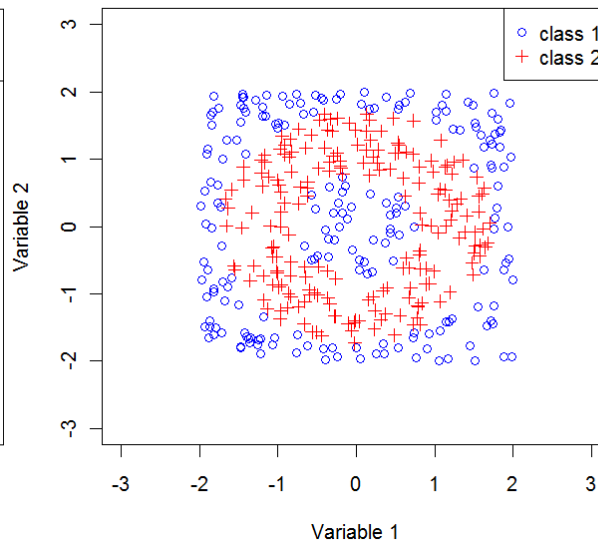
Experiments: simulated data

- Consider LASSO, CFS (correlation variables selection), RRF (regularized random forest), RF-RFE (random forest with RFE)
- In all data sets, only **2** out of **100** variables are needed for classification.

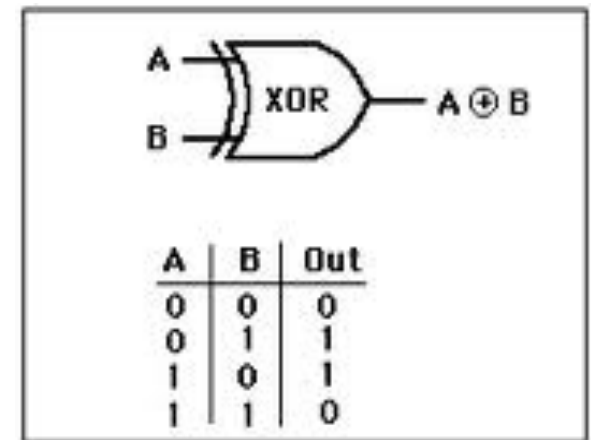
Linear Separable
LASSO, CFS, RF-RFE, RRF



Nonlinear
CFS, RF-RFE, RRF



XOR data
RRF, RF-RFE



Variable Selection

Experiments: high-dimensional data

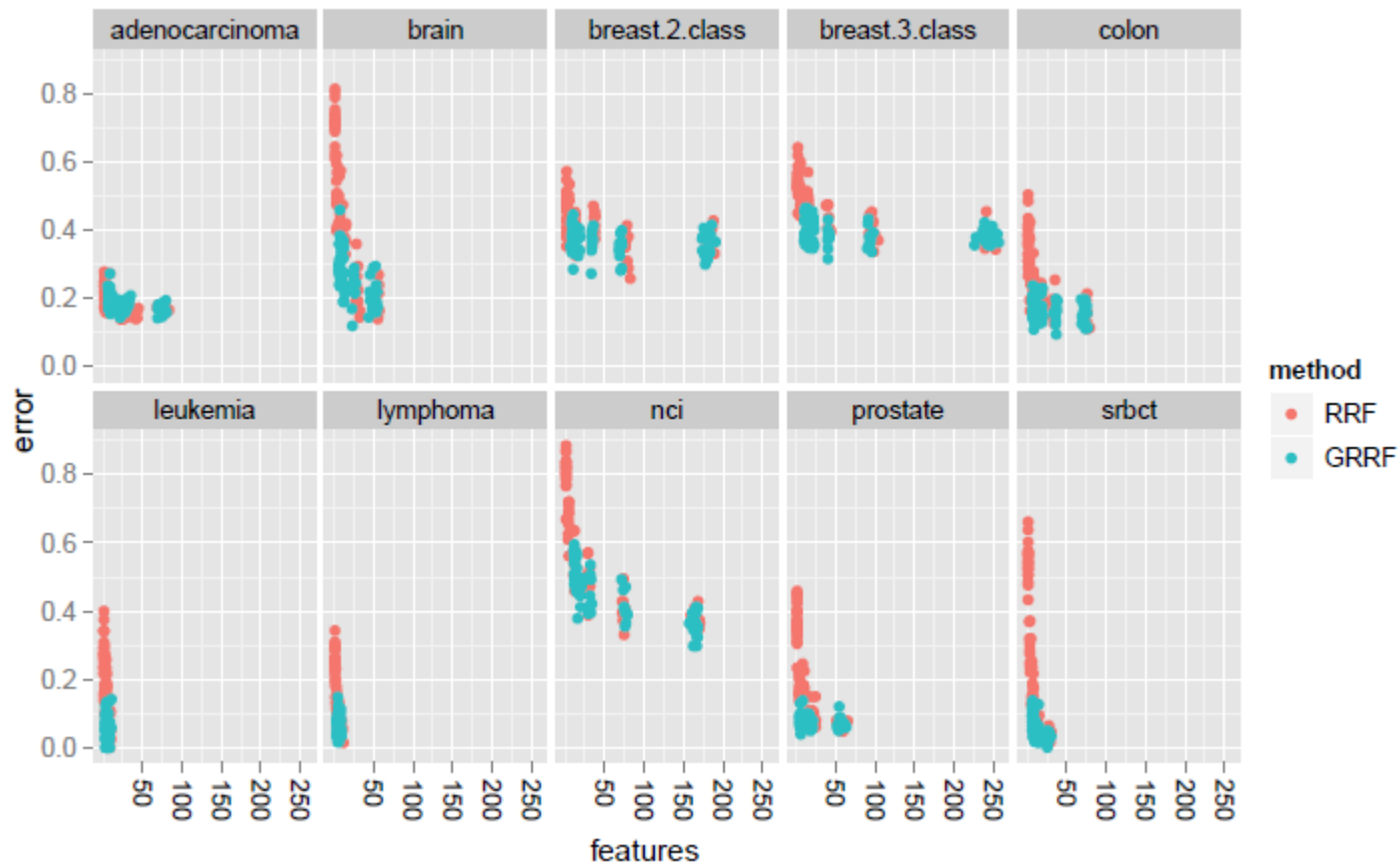
- 10 Gene data sets
- For GRRF, use $\lambda_0=1$ and use 10-fold CV to select the best γ .

Data set	Reference	# Examples	# Features	# classes
Adenocarcinoma	[14]	76	9868	2
Brain	[15]	42	5597	5
Breast.2.class	[16]	77	4869	2
Breast.3.class	[16]	95	4869	3
Colon	[17]	62	2000	2
Leukemia	[18]	38	3051	2
Lymphoma	[19]	62	4026	3
Nci60	[20]	61	5244	8
Prostate	[21]	102	6033	2
Srbct	[22]	63	2308	4

Variable Selection

GRRF is more accurate and more stable than RRF

Given similar number of variables, GRRF produces less error rates



Summary

- Random Forests have many advantages in variable selection and ranking
- The ordinary random forest is biased towards categorical variables with more values. OOB Forest and pForest are two solutions.
- The regularized tree framework can be easily applied to many tree models
- Guided regularized random forest produces better results than regularized random forest

References

- Houtao Deng, George Runger, "Gene Selection with Guided Regularized Random Forest", technical report, 2012.
- Houtao Deng, George Runger, "Feature Selection via Regularized Trees", Proceedings of the 2012 International Joint Conference on Neural Networks (IJCNN), IEEE, 2012.
- Houtao Deng, George Runger, Eugene Tuv, "Bias of Importance Measures for Multi-Valued Attributes and Solutions", Proceedings of the 21st International Conference on Artificial Neural Networks (ICANN 2011), pp 293-300.