RaSE: Random Subspace Ensemble Classification

Ye Tian† and Yang Feng‡

†Department of Statistics, Columbia University
‡Department of Biostatistics, School of Global Public Health, New York University

Introduction

High-dimensional classification and sparsity

We are dealing with the binary classification problem, where $\boldsymbol{x}|y=j\sim f^{(j)}, j=0,1.$ and $\boldsymbol{x}\in\mathbb{R}^p$. Suppose we have training data $\{\boldsymbol{x}_i,y_i\}_{i=1}^n$.

- In the high-dimensional problem, we often have $p\gg n$.
- Only a small feature subset S^* with cardinality $p^*\ll p$ contributes to the model, i.e. $y|\boldsymbol{x}\overset{d}{=}y|\boldsymbol{x}_{S^*}$.

Some examples: sparse linear discriminant analysis (LDA), sparse quadratic discriminant analysis (QDA), ...

Random subspace method

It was first applied in decision trees [2], then extended to various models including LDA, QDA, kNN, ... Its main idea is

- randomly generates some feature subsets
- train models within each subset
- merge these learners to get an ensemble learner

However, in high-dimensional problem, most random subspaces are useless!

RaSE Framework

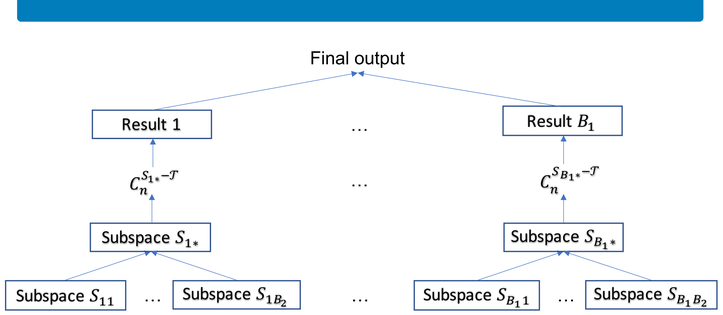


Fig. 1: A two-layer ensemble framework motivated by [1]

Vanilla RaSE algorithm

- For $b_1=1,\ldots,B_1$:
  - Independently generate B_2 subspaces $\{S_{b_1b_2}\}_{b_2=1}^{B_2}$, where $S_{b_1b_2}$ is sampled from *hierarchical uniform distribution*
    - Sample $d\sim\text{Unif}(\{1,\ldots,D\})$
    - $S_{b_1b_2}\leftarrow$ randomly choose d features with *equal* probability
  - Choose the best S_{b_1*} via some *criterion* $\mathcal{T}$
  - Trained classifier $C_n^{S_{b_1*}-\mathcal{T}}\leftarrow$ training data in S_{b_1*}
- Ensemble learner $C_n^{RaSE}(\boldsymbol{x})\leftarrow\mathbb{1}\left(B_1^{-1}\sum_{b_1=1}^{B_1}C_n^{S_{b_1*}-\mathcal{T}}(\boldsymbol{x})>\alpha\right)$

Criterion to evaluate subspaces

- Multiple choices, e.g. AIC, BIC, eBIC, cross-validation error, ...
- A new criterion, ratio information criterion (RIC) [4], is defined as

$$\text{RIC}(S)=-2\pi_0\text{KL}(f_S^{(0)}||f_S^{(1)})-2\pi_1\text{KL}(f_S^{(1)}||f_S^{(0)})+c_n\cdot\deg(S).$$

Iterative RaSE

Assigning sampling weights based on the selected frequency of each feature in the last round.

- Features in S^* can be more easily sampled out.
- It allows us to *rank importance of features* [3, 4].

Theoretical Properties

Notations

- $\mathbb{P}/\mathbf{P}/\text{P}$: probabilities w.r.t. randomness from samples/subspaces/all sources.
- $\mathbb{E}/\mathbf{E}/\text{E}$: expectations w.r.t. randomness from samples/subspaces/all sources.
- $\mathbf{Var}$: variance w.r.t. randomness from subspaces.
- Risk of classifier C: $R(C)=\mathbb{P}(C(\boldsymbol{x})\neq y)$.
- $\mathcal{S}_D^*=\{S:|S|\le D,S\supseteq S^*\}$, $\mathcal{S}_D^c=\{S:|S|\le D,S\not\supseteq S^*\}$

Impact of B_1

The following results hold except for finite values of α:

- $|\mathbf{E}[R(C_n^{RaSE})]-R(C_n^{RaSE*})|=O(\exp\{-C_\alpha B_1\})$, where

$$C_n^{RaSE*}(\boldsymbol{x})=\begin{cases}1, & \mu_n(\boldsymbol{x})>\alpha,\\ 0, & \mu_n(\boldsymbol{x})<\alpha,\\ \text{Bernoulli}\left(\frac{1}{2}\right), & \mu_n(\boldsymbol{x})=\alpha,\end{cases}$$

  and $\mu_n(\boldsymbol{x})=\mathbf{P}\left(C_n^{S_{1*}}(\boldsymbol{x})=1\right)$.
- $\mathbf{Var}[R(C_n^{RaSE})]=O(\exp\{-C_\alpha B_1\})$

Consistency of RIC

- (Weak) Under some conditions, $\mathbf{P}\left(\sup_{S\in\mathcal{S}_D^*}\text{RIC}_n(S^*)<\inf_{S\in\mathcal{S}_D^c}\text{RIC}_n(S)\right)\to1.$
- (Strong) With additional conditions, $\mathbf{P}\left(\text{RIC}_n(S^*)=\inf_{|S|\le D}\text{RIC}_n(S)\right)\to1.$

Expected risk upper bound

$$\mathbb{E}\{\mathbf{E}[R(C_n^{RaSE})-R(C_{Bayes})]\}\le\frac{\mathbb{E}\sup_{S\in\mathcal{S}_D^*}[R(C_n^S)-R(C_{Bayes})]+\text{P}(S_{1*}\not\supseteq S^*)}{\min(\alpha,1-\alpha)}.$$

- $\mathbb{E}\sup_{S\in\mathcal{S}_D^*}[R(C_n^S)-R(C_{Bayes})]$: caused by limited samples/noisy features
- $\text{P}(S_{1*}\not\supseteq S^*)$: caused by inaccurate subspace selection

Benefits of iterative RaSE

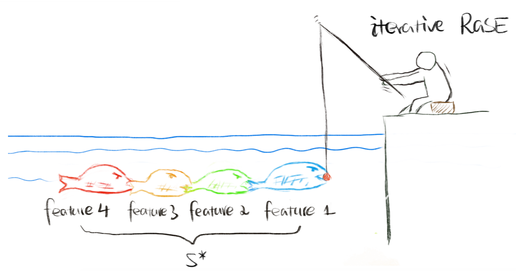


Fig. 2: Iterative RaSE is "fishing" S^* step by step.

- When p is large (e.g. $p>100$), it's hard to sample a subspace covering the whole S^*. (thinking about $1/\binom{100}{5}\approx10^{-8}$). $\Rightarrow$ stringent requirement of B_2
- But it's still possible to cover partial S^*. ($\binom{95}{4}\binom{5}{1}/\binom{100}{5}\approx0.21$)
- Under some *stepwise detectable conditions*, after several iterations, $\text{P}(S_{1*}\supseteq S^*)\to1$ holds with moderate B_2 settings.

A Simulated QDA Example

Model set-up

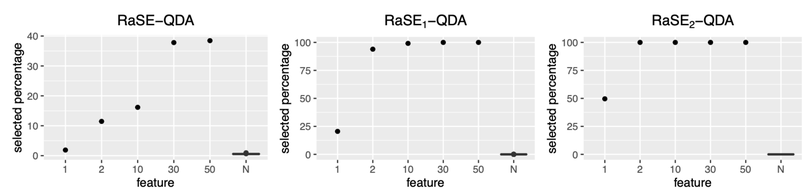
- $\boldsymbol{x}|y=r\sim f^{(r)}=N(\boldsymbol{\mu}^{(r)},\Sigma^{(r)}), r=0,1.$ And $p=200$.
- $S^*=\{1,2,10,30,50\}$, where $\{1,2\}$ contributes to the linear part while $\{10,30,50\}$ contributes to the quadratic part.

Test error rates

Method	$n=200$	$n=400$	$n=1000$
RaSE-LDA	$37.26_{2.86}$	$36.08_{1.99}$	$35.67_{1.66}$
RaSE-QDA	$32.19_{2.87}$	$30.57_{2.82}$	$29.05_{1.91}$
RaSE-kNN	$30.92_{2.92}$	$27.72_{2.41}$	$25.28_{1.75}$
RaSE$_1$-LDA	$35.81_{2.97}$	$33.36_{2.13}$	$32.81_{1.64}$
RaSE$_1$-QDA	$27.18_{2.69}$	$25.19_{1.97}$	$24.20_{1.44}$
RaSE$_1$-kNN	$29.44_{3.15}$	$27.05_{2.30}$	$25.47_{1.58}$
RaSE$_2$-LDA	$36.77_{2.42}$	$33.67_{1.79}$	$32.70_{1.49}$
RaSE$_2$-QDA	$27.12_{3.04}$	$24.78_{1.95}$	$24.09_{1.38}$
RaSE$_2$-kNN	$30.34_{3.48}$	$26.95_{2.46}$	$24.76_{1.59}$
RP-LDA	$44.80_{1.84}$	$43.03_{1.89}$	$40.20_{1.78}$
RP-QDA	$43.15_{2.03}$	$40.26_{2.03}$	$36.35_{1.77}$
RP-kNN	$44.13_{1.79}$	$42.74_{1.71}$	$40.79_{2.10}$
LDA	$49.07_{2.22}$	$43.13_{1.88}$	$38.55_{1.82}$
QDA	—†	—†	$45.19_{1.75}$
kNN	$45.35_{1.81}$	$44.45_{1.91}$	$43.23_{1.70}$
sLDA	$36.77_{3.34}$	$34.05_{2.13}$	$33.13_{1.55}$
RAMP	$37.53_{6.25}$	$33.03_{2.04}$	$32.47_{1.80}$
NSC	$41.76_{4.29}$	$37.93_{3.68}$	$35.41_{2.32}$
RF	$37.40_{3.15}$	$31.74_{2.36}$	$27.46_{1.57}$
Sig-QDA	$23.46_{1.52}$	$22.75_{1.41}$	$22.38_{1.29}$

*: the best classifier
**: the one within 1 sd
—†: not applicable

Selected percentages of each feature



References

[1] Timothy I Cannings, Richard J Samworth, et al. "Random-projection ensemble classification". In: *Journal of the Royal Statistical Society Series B* 79.4 (2017), pp. 959–1035.

[2] Tin Kam Ho. "The random subspace method for constructing decision forests". In: *IEEE transactions on pattern analysis and machine intelligence* 20.8 (1998), pp. 832–844.

[3] Ye Tian and Yang Feng. "RaSE: A variable screening framework via random subspace ensembles". In: *arXiv preprint arXiv:2102.03892* (2021).

[4] Ye Tian and Yang Feng. "RaSE: Random subspace ensemble classification". In: *Journal of Machine Learning Research* 22.45 (2021), pp. 1–93.