RaSE: Random Subspace Ensemble

Yang Feng

New York University

Oct 12 @ TEDS SEMINAR

Collaborator

Ye Tian

Zoe Zhu

Outline

Outline

Binary classification

- Features $\mathbf{x} \in \mathcal{X} \subset \mathbb{R}^p$
  - e.g., features of a credit card transaction.

Binary classification

- Features $\mathbf{x} \in \mathcal{X} \subset \mathbb{R}^p$
  - e.g., features of a credit card transaction.
- Class label $y \in \{0, 1\}$
  - e.g., fraud status (YES or NO) of a transaction.

Binary classification

- Features $\mathbf{x} \in \mathcal{X} \subset \mathbb{R}^p$
  - e.g., features of a credit card transaction.
- Class label $y \in \{0, 1\}$
  - e.g., fraud status (YES or NO) of a transaction.
- A classifier is a binary function $C : \mathcal{X} \to \{0, 1\}$.

Binary classification

- Features $\mathbf{x} \in \mathcal{X} \subset \mathbb{R}^p$
  - e.g., features of a credit card transaction.
- Class label $y \in \{0, 1\}$
  - e.g., fraud status (YES or NO) of a transaction.
- A classifier is a binary function $C : \mathcal{X} \to \{0, 1\}$.
- Training data $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$.

Binary classification

- Features $\mathbf{x} \in \mathcal{X} \subset \mathbb{R}^p$
  - e.g., features of a credit card transaction.
- Class label $y \in \{0, 1\}$
  - e.g., fraud status (YES or NO) of a transaction.
- A classifier is a binary function $C : \mathcal{X} \to \{0, 1\}$.
- Training data $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$.
- The risk of classification is a function of C:

$$R(C) = \mathsf{E}[\mathbb{1}(C(\mathbf{x}) \neq y)] = \mathsf{P}(C(\mathbf{x}) \neq y)).$$

Ensemble classification

Suppose we have trained B_1 classifiers $\{C_j\}_{j=1}^{B_1}$ based on the training data, then they can be aggregated in a simple average to get the final decision function

$$C^{En}(\mathbf{x}) = \mathbb{1}\left(\frac{1}{B_1}\sum_{j=1}^{B_1}\mathbb{1}(C_j(\mathbf{x})=1) > \alpha\right).$$

Ensemble classification

Suppose we have trained B_1 classifiers $\{C_j\}_{j=1}^{B_1}$ based on the training data, then they can be aggregated in a simple average to get the final decision function

$$C^{En}(\mathbf{x}) = \mathbb{1}\left(\frac{1}{B_1}\sum_{j=1}^{B_1}\mathbb{1}(C_j(\mathbf{x})=1) > \alpha\right).$$

- Previous work on ensemble classification:
  - Breiman (1996): Bagging
  - Breiman (2001): Random forest
  - Freund and Schapire (1995): Boosting
  - Ho (1998): Random subspace method
  - Ahn et al. (2007): Random partition ensemble classification
  - Blaser and Fryzlewicz (2016): Random rotation
  - Cannings and Samworth (2017): Random projection

Random subspace method (Ho, 1998)

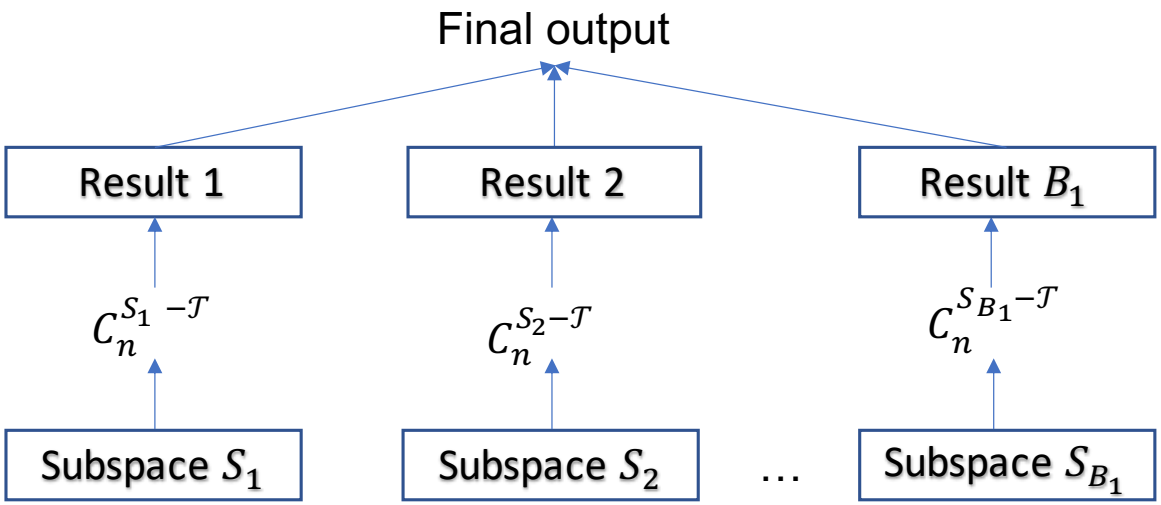
- It's a model-free framework. Suppose we generate B_1 random subspaces $\{S_j\}_{j=1}^{B_1}$ and construct j-th weak learner $C_n^{S_j-\mathcal{T}}$ of type $\mathcal{T}$, then the ensemble classifier is

$$C^{RSM}(\mathbf{x}) = \mathbb{1}\left(\frac{1}{B_1}\sum_{j=1}^{B_1}\mathbb{1}(C_n^{S_j-\mathcal{T}}(\mathbf{x})=1) > \alpha\right).$$

Random subspace method (Ho, 1998)

- It's a model-free framework. Suppose we generate B_1 random subspaces $\{S_j\}_{j=1}^{B_1}$ and construct j-th weak learner $C_n^{S_j-\mathcal{T}}$ of type $\mathcal{T}$, then the ensemble classifier is

$$C^{RSM}(\mathbf{x}) = \mathbb{1}\left(\frac{1}{B_1}\sum_{j=1}^{B_1}\mathbb{1}(C_n^{S_j-\mathcal{T}}(\mathbf{x})=1) > \alpha\right).$$



Sparse classification problems

In many problems, only p^* of p features are signals, where $p^* \ll p$.
Consider the mixture model

$$\mathbf{x} \sim \pi_0 f^{(0)} + \pi_1 f^{(1)},$$

where $f^{(0)}, f^{(1)}$ are the conditional densities, inducing measures $\mathsf{P}^{(0)}, \mathsf{P}^{(1)}$.

Sparse classification problems

In many problems, only p^* of p features are signals, where $p^* \ll p$.
Consider the mixture model

$$\mathbf{x} \sim \pi_0 f^{(0)} + \pi_1 f^{(1)},$$

where $f^{(0)}, f^{(1)}$ are the conditional densities, inducing measures $\mathsf{P}^{(0)}, \mathsf{P}^{(1)}$.

- Discriminative set S: $\boldsymbol{y}|\boldsymbol{x}_S$ is independent with $\boldsymbol{x}_{S^c}$. (Zhang and Wang, 2011; Kohavi et al., 1997; Mai et al., 2012)

Sparse classification problems

In many problems, only p^* of p features are signals, where $p^* \ll p$.
Consider the mixture model

$$\mathbf{x} \sim \pi_0 f^{(0)} + \pi_1 f^{(1)},$$

where $f^{(0)}, f^{(1)}$ are the conditional densities, inducing measures $\mathsf{P}^{(0)}, \mathsf{P}^{(1)}$.

- Discriminative set S: $\boldsymbol{y}|\boldsymbol{x}_S$ is independent with $\boldsymbol{x}_{S^c}$. (Zhang and Wang, 2011; Kohavi et al., 1997; Mai et al., 2012)
- An equivalent definition: There exists a function $h : \mathbb{R}^{|S|} \to [0, +\infty]$ such that

$$\frac{f^{(1)}(\boldsymbol{x})}{f^{(0)}(\boldsymbol{x})} = h(\boldsymbol{x}_S)$$

almost surely with respect to $\mathsf{P}^X = \pi_0 \mathsf{P}^{(0)} + \pi_1 \mathsf{P}^{(1)}$. (Tian and Feng, 2021)

Sparse classification problems

In many problems, only p^* of p features are signals, where $p^* \ll p$.
Consider the mixture model

$$\mathbf{x} \sim \pi_0 f^{(0)} + \pi_1 f^{(1)},$$

where $f^{(0)}, f^{(1)}$ are the conditional densities, inducing measures $\mathsf{P}^{(0)}, \mathsf{P}^{(1)}$.

- Discriminative set S: $\boldsymbol{y}|\boldsymbol{x}_S$ is independent with $\boldsymbol{x}_{S^c}$. (Zhang and Wang, 2011; Kohavi et al., 1997; Mai et al., 2012)
- An equivalent definition: There exists a function $h : \mathbb{R}^{|S|} \to [0, +\infty]$ such that

$$\frac{f^{(1)}(\boldsymbol{x})}{f^{(0)}(\boldsymbol{x})} = h(\boldsymbol{x}_S)$$

  almost surely with respect to $\mathsf{P}^X = \pi_0 \mathsf{P}^{(0)} + \pi_1 \mathsf{P}^{(1)}$. (Tian and Feng, 2021)
- Minimal discriminative set S^* : the discriminative set with minimal cardinality. (Tian and Feng, 2021)

Sparse classification problems

Example (Discriminant analysis):

$$\boldsymbol{x} \sim \pi_0 N(\boldsymbol{\mu}_{p\times 1}^{(0)}, \Sigma_{p\times p}^{(0)}) + \pi_1 N(\boldsymbol{\mu}_{p\times 1}^{(1)}, \Sigma_{p\times p}^{(1)}), \pi_0 + \pi_1 = 1.$$

Sparse classification problems

Example (Discriminant analysis):

$\boldsymbol{x} \sim \pi_0 N(\boldsymbol{\mu}^{(0)}_{p\times 1}, \Sigma^{(0)}_{p\times p}) + \pi_1 N(\boldsymbol{\mu}^{(1)}_{p\times 1}, \Sigma^{(1)}_{p\times p}), \pi_0 + \pi_1 = 1.$

- **LDA**: $\Sigma^{(0)} = \Sigma^{(1)} = \Sigma$. We have

$$\log\left(\frac{f^{(0)}(\boldsymbol{x})}{f^{(1)}(\boldsymbol{x})}\right) = (\boldsymbol{\mu}^{(0)} - \boldsymbol{\mu}^{(1)})^T \Sigma^{-1} \boldsymbol{x} + C.$$

where C is a constant unrelated to $\boldsymbol{x}$. We have $S^* = \{j : (\Sigma^{-1}(\boldsymbol{\mu}^{(1)} - \boldsymbol{\mu}^{(0)}))_j \neq 0\}$.

Sparse classification problems

Example (Discriminant analysis):

$\boldsymbol{x} \sim \pi_0 N(\boldsymbol{\mu}^{(0)}_{p\times 1}, \Sigma^{(0)}_{p\times p}) + \pi_1 N(\boldsymbol{\mu}^{(1)}_{p\times 1}, \Sigma^{(1)}_{p\times p}), \pi_0 + \pi_1 = 1.$

- **LDA**: $\Sigma^{(0)} = \Sigma^{(1)} = \Sigma$. We have

$$\log\left(\frac{f^{(0)}(\boldsymbol{x})}{f^{(1)}(\boldsymbol{x})}\right) = (\boldsymbol{\mu}^{(0)} - \boldsymbol{\mu}^{(1)})^T \Sigma^{-1} \boldsymbol{x} + C.$$

  where C is a constant unrelated to $\boldsymbol{x}$. We have $S^* = \{j : (\Sigma^{-1}(\boldsymbol{\mu}^{(1)} - \boldsymbol{\mu}^{(0)}))_j \neq 0\}$.

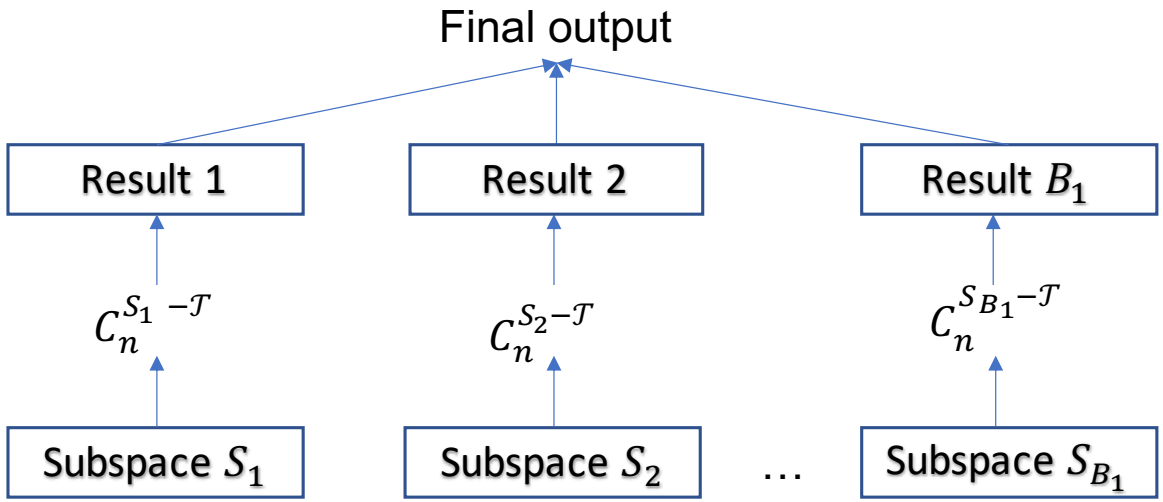
- **QDA**: $\Sigma^{(0)}$ and $\Sigma^{(1)}$ can be different. Since

$$\log\left(\frac{f^{(0)}(\boldsymbol{x})}{f^{(1)}(\boldsymbol{x})}\right) = \frac{1}{2}\boldsymbol{x}^T \Omega \boldsymbol{x} + \boldsymbol{\delta}^T \boldsymbol{x} + C.$$

  where C is a constant unrelated to $\boldsymbol{x}$, and $\Omega = (\Sigma^{(1)})^{-1} - (\Sigma^{(0)})^{-1}$, $\boldsymbol{\delta} = (\Sigma^{(0)})^{-1}\boldsymbol{\mu}^{(0)} - (\Sigma^{(1)})^{-1}\boldsymbol{\mu}^{(1)}$, we have $S^* = \{j : \Omega_{ij} \neq 0, \exists i\}$ $\cup \{j : \boldsymbol{\delta}_j \neq 0\}$.

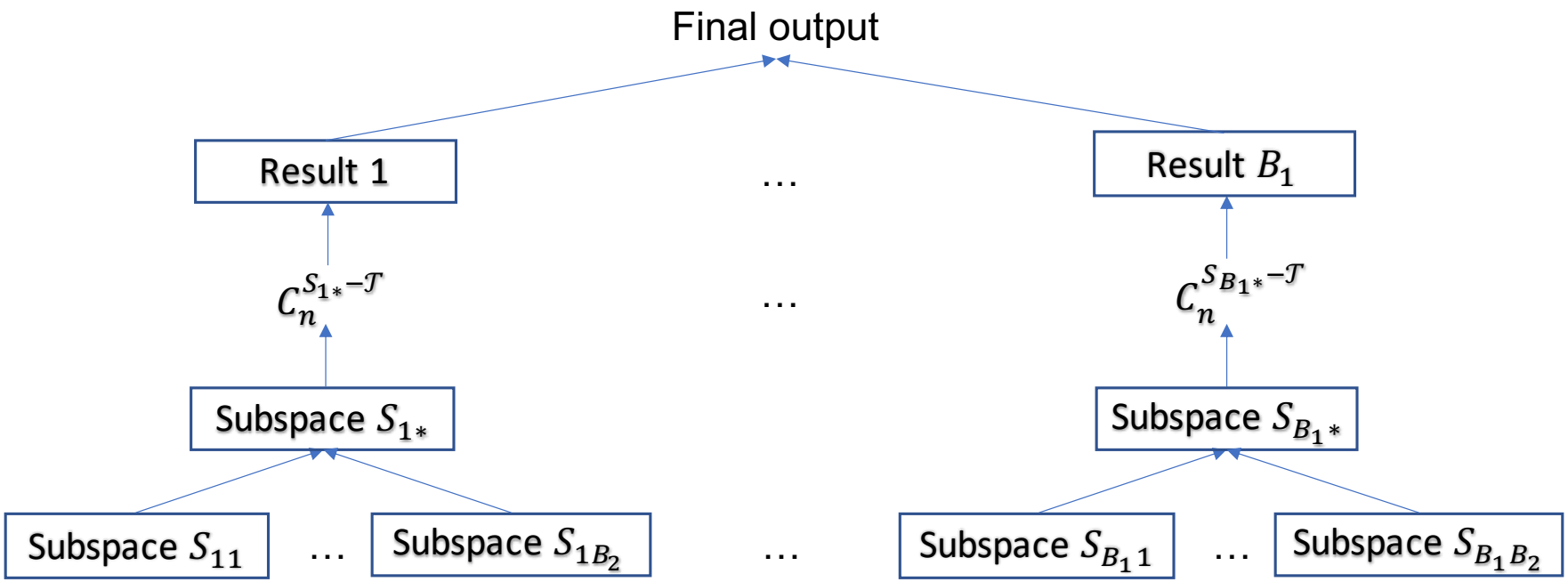
Classical aggregation framework

Let's recall the classical aggregation framework.

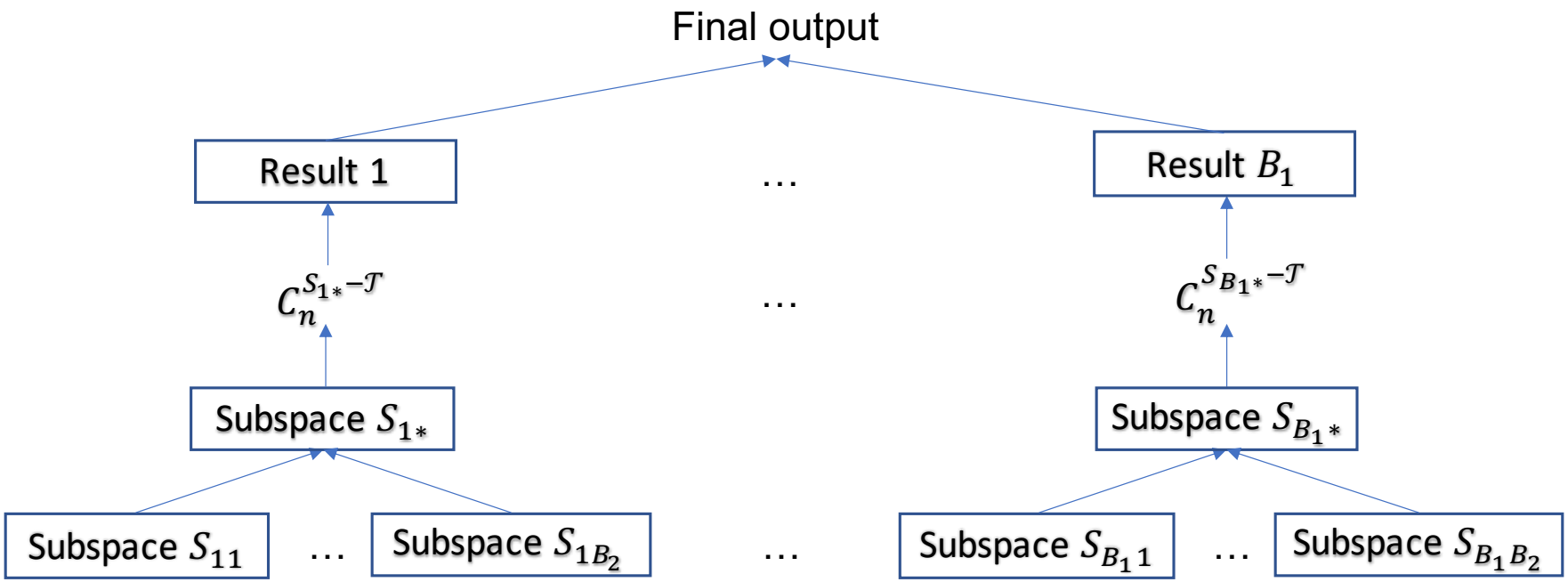


◦ For high-dimensional sparse problems, only a few of S_j can cover S^*.

Another way for aggregation

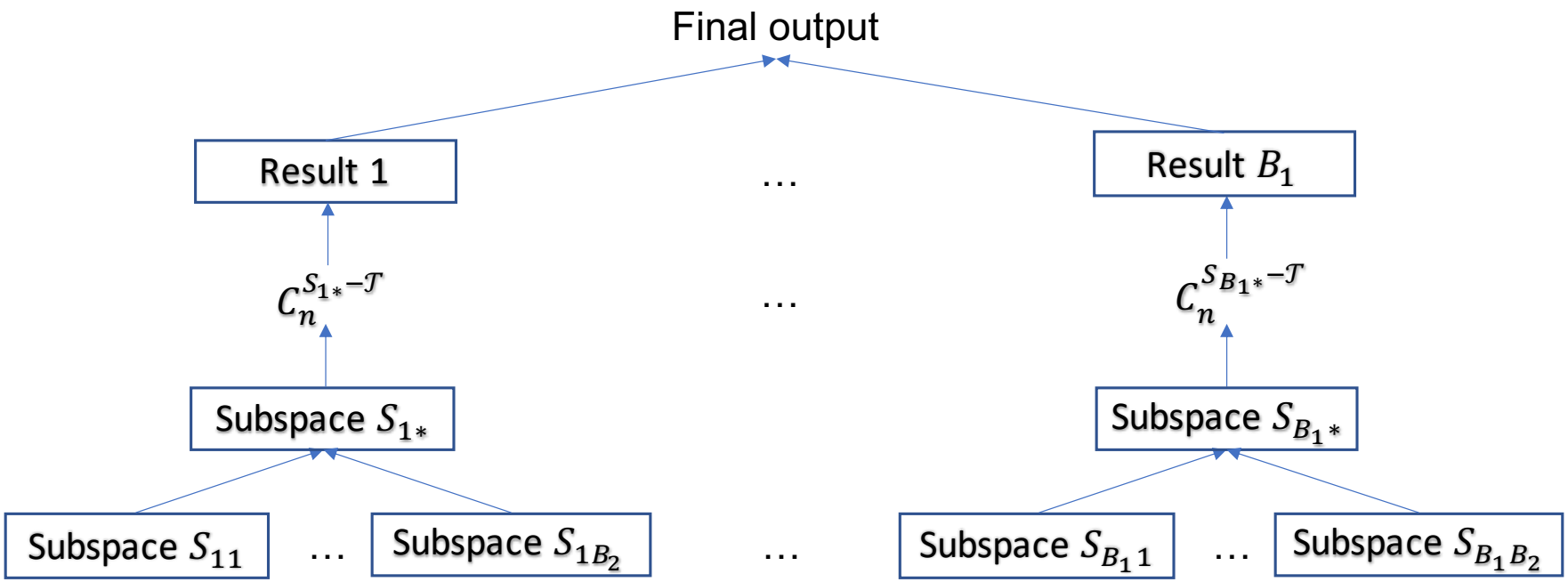


Another way for aggregation



- This scheme was also used by Cannings and Samworth (2017) for random projection ensemble.

Another way for aggregation



- This scheme was also used by Cannings and Samworth (2017) for random projection ensemble.
- The random subspace method works better for sparse classification problems.

Outline

Vanilla RaSE

Algorithm 1: Random subspace ensemble classification (RaSE)

Input: training data $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$, new data $\boldsymbol{x}$, subspace distribution $\mathcal{D}$, criterion $\mathcal{C}$, integers B_1 and B_2, type of base classifier $\mathcal{T}$

Output: the predicted label $C_n^{RaSE}(\mathbf{x})$, the selected proportion of each feature $\boldsymbol{\eta}$

Independently generate random subspaces $S_{jk} \sim \mathcal{D}, 1 \le j \le B_1, 1 \le k \le B_2$

for $j \leftarrow 1$ **to** B_1 **do**

  Select the optimal subspace S_{j*} from $\{S_{jk}\}_{k=1}^{B_2}$ according to $\mathcal{C}$

  Train $C_n^{S_{j*}-\mathcal{T}}$ in S_{j*}

end

Construct the ensemble decision function $\nu_n(\boldsymbol{x}) = \frac{1}{B_1}\sum_{j=1}^{B_1} C_n^{S_{j*}-\mathcal{T}}(\boldsymbol{x})$

Set the threshold $\hat{\alpha}$ according to (1)

Output $C_n^{RaSE}(\mathbf{x}) = \mathbb{1}(\nu_n(\boldsymbol{x}) > \hat{\alpha})$ and $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_p)^T$, where $\eta_l = B_1^{-1}\sum_{j=1}^{B_1} \mathbb{1}(l \in S_{j*}), l = 1, \ldots, p.$

Vanilla RaSE

- Distribution $\mathcal{D}$ for random subspaces: hierarchical uniform distribution
  - Set D
  - Draw d_{jk} i.i.d. from Uniform($\{1, 2, \ldots, D\}$)
  - Draw S_{jk} from Uniform($\{S \subseteq \{1, \ldots, p\} : |S| = d_{jk}\}$)

Vanilla RaSE

- Distribution $\mathcal{D}$ for random subspaces: hierarchical uniform distribution
  - Set D
  - Draw d_{jk} i.i.d. from Uniform$(\{1, 2, \ldots, D\})$
  - Draw S_{jk} from Uniform$(\{S \subseteq \{1, \ldots, p\} : |S| = d_{jk}\})$
- Threshold

$$\hat{\alpha} = \underset{\alpha \in (0,1)}{\operatorname{argmin}} \left[\text{training error of } C_n^{RaSE} \text{ based on } \alpha\right].$$

Vanilla RaSE

- Distribution $\mathcal{D}$ for random subspaces: hierarchical uniform distribution
  - Set D
  - Draw d_{jk} i.i.d. from Uniform$(\{1, 2, \ldots, D\})$
  - Draw S_{jk} from Uniform$(\{S \subseteq \{1, \ldots, p\} : |S| = d_{jk}\})$
- Threshold

$$\hat{\alpha} = \underset{\alpha \in (0,1)}{\operatorname{argmin}} \left[\text{training error of } C_n^{RaSE} \text{ based on } \alpha\right].$$

- The selected proportion of each feature $\boldsymbol{\eta}$ can be used to rank significance of features.

Vanilla RaSE

- Distribution $\mathcal{D}$ for random subspaces: hierarchical uniform distribution
  - Set D
  - Draw d_{jk} i.i.d. from Uniform$(\{1, 2, \ldots, D\})$
  - Draw S_{jk} from Uniform$(\{S \subseteq \{1, \ldots, p\} : |S| = d_{jk}\})$
- Threshold

$$\hat{\alpha} = \underset{\alpha \in (0,1)}{\operatorname{argmin}} \left[\text{training error of } C_n^{RaSE} \text{ based on } \alpha\right].$$

- The selected proportion of each feature $\boldsymbol{\eta}$ can be used to rank significance of features.
- Possible base classifier $\mathcal{T}$:
  - Parametric: LDA, QDA, SVM, logistic regression, ...
  - Non-parametric: kNN, trees, random forest, neural network, ...

Vanilla RaSE

- Distribution $\mathcal{D}$ for random subspaces: hierarchical uniform distribution
  - Set D
  - Draw d_{jk} i.i.d. from Uniform$(\{1, 2, \ldots, D\})$
  - Draw S_{jk} from Uniform$(\{S \subseteq \{1, \ldots, p\} : |S| = d_{jk}\})$
- Threshold

$$\hat{\alpha} = \underset{\alpha \in (0,1)}{\operatorname{argmin}} \left[\text{training error of } C_n^{RaSE} \text{ based on } \alpha\right].$$

- The selected proportion of each feature $\boldsymbol{\eta}$ can be used to rank significance of features.
- Possible base classifier $\mathcal{T}$:
  - Parametric: LDA, QDA, SVM, logistic regression, ...
  - Non-parametric: kNN, trees, random forest, neural network, ...
- Criterion $\mathcal{C}$: multiple choices (to discuss later).

Iterative RaSE

- **Restrictive multinomial distribution** $Rmultin(p, d, \boldsymbol{\eta})$: $J = (J_1, ..., J_p)^T, \sum_l J_l = d, J_l \in \{0, 1\}$ and each J_l has marginal probability $\boldsymbol{\eta}$.

Iterative RaSE

- **Restrictive multinomial distribution** $Rmultin(p, d, \boldsymbol{\eta})$: $J = (J_1, ..., J_p)^T, \sum_l J_l = d, J_l \in \{0, 1\}$ and each J_l has marginal probability $\boldsymbol{\eta}$.
- How this helps to generate each subspace
  - Draw d from Uniform($\{1, 2, \ldots, D\}$).
  - Draw $J = (J_1, ..., J_p)^T \sim Rmultin(p, d, \boldsymbol{\eta})$, where $J_l = \mathbb{1}(l \in S)$, $l = 1, \ldots, p$.
  - $J \longleftrightarrow$ a subspace S of size d

Iterative RaSE

- **Restrictive multinomial distribution** $Rmultin(p, d, \boldsymbol{\eta})$: $J = (J_1, ..., J_p)^T, \sum_l J_l = d, J_l \in \{0, 1\}$ and each J_l has marginal probability $\boldsymbol{\eta}$.
- How this helps to generate each subspace
  - Draw d from Uniform($\{1, 2, \ldots, D\}$).
  - Draw $J = (J_1, ..., J_p)^T \sim Rmultin(p, d, \boldsymbol{\eta})$, where $J_l = \mathbb{1}(l \in S)$, $l = 1, \ldots, p$.
  - $J \longleftrightarrow$ a subspace S of size d
- How about updating subspace distribution $\mathcal{D}$ using $\boldsymbol{\eta}$?

Iterative RaSE

Algorithm 2: Iterative RaSE (RaSE$_T$)

Input: training data $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$, new data $\boldsymbol{x}$, initial subspace distribution $\mathcal{D}^{(0)}$, criterion $\mathcal{C}$, integers B_1 and B_2, the type of base classifier $\mathcal{T}$, the number of iterations T

Output: the predicted label $C_n^{RaSE}(\mathbf{x})$, the proportion of each feature $\boldsymbol{\eta}^{(T)}$

for $t \leftarrow 0$ **to** T **do** `// The iteration step`

Independently generate random subspaces $S_{jk}^{(t)} \sim \mathcal{D}^{(t)}, 1 \le j \le B_1, 1 \le k \le B_2$

for $j \leftarrow 1$ **to** B_1 **do**

Select the optimal subspace $S_{j*}^{(t)}$ from $\{S_{jk}^{(t)}\}_{k=1}^{B_2}$ according to $\mathcal{C}$ and $\mathcal{T}$

Train $C_n^{S_{j*}-\mathcal{T}}$ in $S_{j*}^{(t)}$

end

Update $\boldsymbol{\eta}^{(t)}$ where $\eta_l^{(t)} = B_1^{-1} \sum_{j=1}^{B_1} \mathbb{1}(l \in S_{j*}^{(t)}), l = 1, \ldots, p$

Update $\mathcal{D}^{(t)} \leftarrow$ restrictive multinomial distribution with parameter $(p, d, \tilde{\boldsymbol{\eta}}^{(t)})$, where $\tilde{\eta}_l^{(t)} = \eta_l^{(t)} \mathbb{1}(\eta_l^{(t)} > C_0/\log p) + \frac{C_0}{p} \mathbb{1}(\eta_l^{(t)} \le C_0/\log p)$

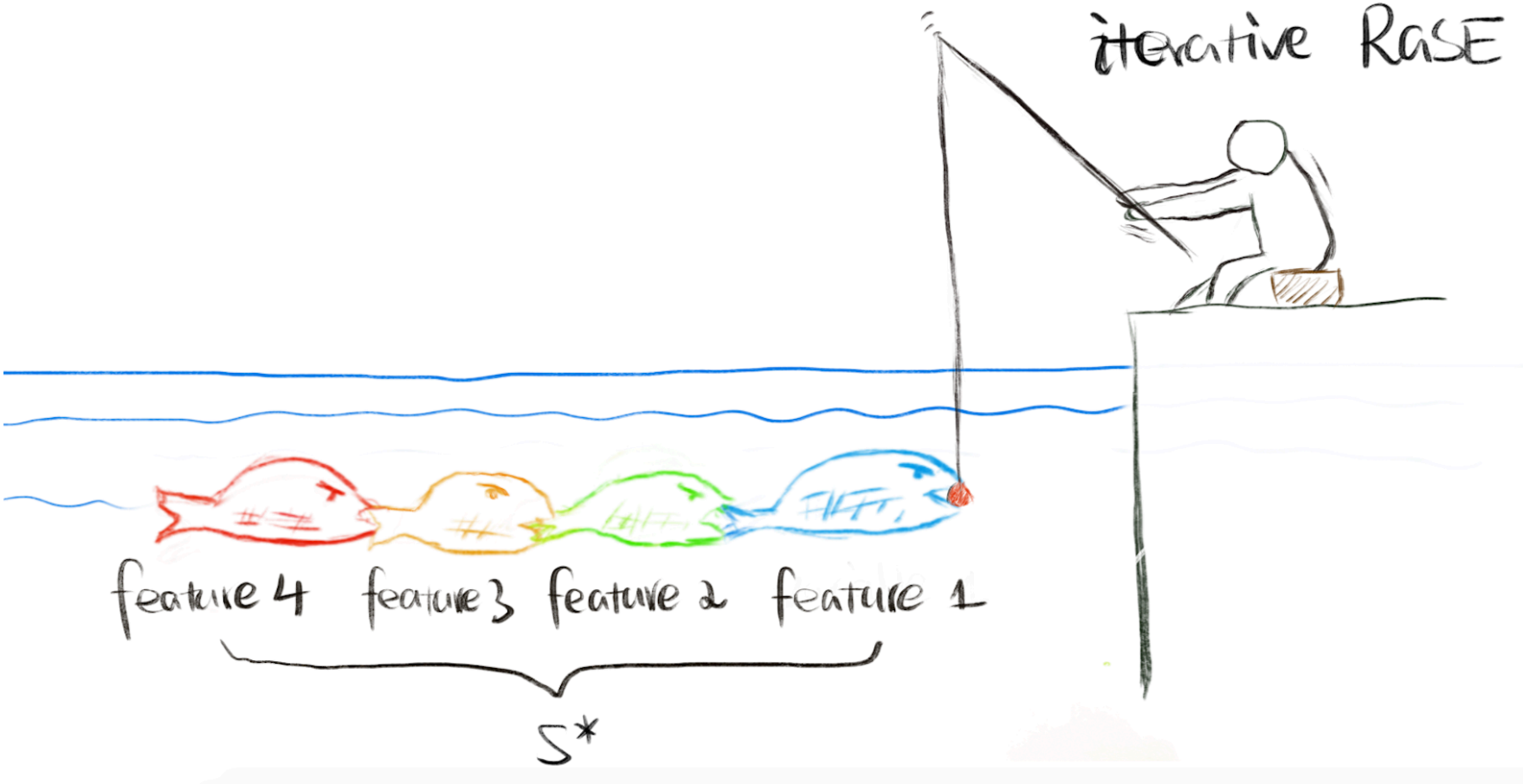
end

Construct the ensemble decision function $\nu_n(\boldsymbol{x}) = \frac{1}{B_1} \sum_{j=1}^{B_1} C_n^{S_{j*}-\mathcal{T}}(\boldsymbol{x})$

Set the threshold $\hat{\alpha}$ according to (1)

Output the predicted label $C_n^{RaSE}(\mathbf{x}) = \mathbb{1}(\nu_n(\boldsymbol{x}) > \hat{\alpha})$ and $\boldsymbol{\eta}^{(T)}$

Fishing Signals Using Iterative RaSE



Multiple choices for criterion $\mathcal{C}$

Other choices:

- Minimizing training error (Cannings and Samworth, 2017; Bryll et al., 2003)
- Minimizing validation or cross-validation error (Cannings and Samworth, 2017; Bryll et al., 2003)
- Minimizing other information criterion, like AIC, BIC and their generalizations (Akaike, 1973; Schwarz et al., 1978; Chen and Chen, 2008, 2012; Fan and Tang, 2013)

Misclassification rate of RaSE classifier

General misclassification rate (Tian and Feng (2021))

For RaSE classifier with threshold α and any criterion to choose optimal subspaces, it holds that

$$\mathbb{E}\{\mathbf{E}[R(C_n^{RaSE})-R(C_{Bayes})]\} \leq \frac{\mathbb{E}\sup_{\substack{S:S\supseteq S^* \\ |S|\leq D}}[R(C_n^S)-R(C_{Bayes})]+\mathbb{P}(S_{1*}\not\supseteq S^*)}{\min(\alpha,1-\alpha)}.$$

- $\mathbb{E}\sup_{\substack{S:S\supseteq S^* \\ |S|\leq D}}[R(C_n^S)-R(C_{Bayes})]$: the discrepancy between finite sample classifier and the oracle. It is shown to converge to zero when $\mathcal{T}$ is LDA or QDA under some conditions. (Efron, 1975; Li and Shao, 2015; Hall et al., 2008; Samworth et al., 2012)
- $\mathrm{P}(S_{1*}\not\supseteq S^*)$: the accuracy of subspace selection, which converges to zero.

Outline

Iterative RaSE Screening

Algorithm 3: Iterative RaSE screening (RaSE$_T$)

Input: training data $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$, initial subspace distribution $\mathcal{D}^{[0]}$, criterion function $\mathcal{C}_n$, integers B_1 and B_2, the number of iterations T, positive constant C_0, number of variables N to select

Output: the selected proportion of each feature $\hat{\boldsymbol{\eta}}^{[T]}$, the selected subset $\hat{S}$

for $t \leftarrow 0$ **to** T **do**

Independently generate random subspaces $S_{b_1 b_2}^{[t]} \sim \mathcal{D}^{[t]}, 1 \le b_1 \le B_1,\ 1 \le b_2 \le B_2$

for $b_1 \leftarrow 1$ **to** B_1 **do**

Select the optimal subspace $S_{b_1 *}^{[t]} = S_{b_1 b_2^*}^{[t]}$, where $b_2^* = \arg\min_{1 \le b_2 \le B_2} \mathcal{C}_n(S_{b_1 b_2}^{[t]})$

end

Update $\hat{\boldsymbol{\eta}}^{[t]}$ where $\hat{\eta}_j^{[t]} = B_1^{-1} \sum_{b_1=1}^{B_1} \mathbb{1}(j \in S_{b_1 *}^{[t]}), j = 1, \ldots, p$

Update $\mathcal{D}^{[t+1]} \leftarrow$ hierarchical restrictive multinomial distribution $\mathcal{R}(\mathcal{U}_0, p, \tilde{\boldsymbol{\eta}}^{[t]})$, where $\tilde{\eta}_j^{[t]} \propto [\hat{\eta}_j^{[t]} \mathbb{1}(\hat{\eta}_j^{[t]} > C_0/\log p) + \frac{C_0}{p} \mathbb{1}(\hat{\eta}_j^{[t]} \le C_0/\log p)]$ and $\sum_{j=1}^p \tilde{\eta}_j^{[t]} = 1$

end

Output the selected proportion of each feature $\hat{\boldsymbol{\eta}}^{[T]}$

Output $\hat{S} = \{1 \le j \le p : \hat{\eta}_j^{[T]}$ is among the N largest of all$\}$

Sure Screening Property

Sure screening property

For any $\alpha > 1$, let $\hat{S}_\alpha = \{1 \le j \le p : \hat{\eta}_j$ is among the $[\alpha D/c_{2n}]$ largest of all$\}$. Under certain conditions, when $B_1 \gg \log p^*$ and $n \to \infty$, we have

- $\mathbb{P}(S^* \subseteq \hat{S}_\alpha) \ge 1 - p^* \exp\left\{-2B_1 c_{2n}^2 \left(1 - \frac{1}{\alpha}\right)^2\right\} \to 1$;
- The selected model size $|\hat{S}_\alpha| \lesssim D$.

Outline

Motivation

- RaSE algorithm needs to pair with a base classifier and it could fail to work well if the base classifier is not properly set.
- We relax this requirement by replacing a single base classifier with *a collection of base classifiers*. For example, $\mathcal{T} = \{\text{LDA, QDA, KNN}\}$.

We call the new ensemble classification framework the *Super Random Subspace Ensemble (Super RaSE)*.

Algorithm 4: Super Random Subspace Ensemble classification (SRaSE)

Input: training data $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$, new data $\boldsymbol{x}$, subspace distribution $\mathcal{D}$, integers B_1 and B_2, the candidate base classifier set $\mathcal{T}$, base classifier distribution $\mathbb{D}$

Output: predicted label $C_n^{RaSE}(\boldsymbol{x})$, the selected proportion of each base classifier $\boldsymbol{\zeta}$, and for the base classifier $T_i \in \mathcal{T}$ where $i \in \{1, \cdots, M\}$, the selected proportion of each feature $\boldsymbol{\eta}_i$

Independently generate base classifiers $T_{jk} \sim \mathbb{D}, 1 \le j \le B_1, 1 \le k \le B_2$

Independently generate random subspaces $S_{jk} \sim \mathcal{D}, 1 \le j \le B_1, 1 \le k \le B_2$

for $j \leftarrow 1$ **to** B_1 **do**

  Select the optimal **subspace and base classifier pair** (T_{j*}, S_{j*}) from $\{(T_{jk}, S_{jk})\}_{k=1}^{B_2}$ using 5-fold cross-validation.

end

Construct the ensemble decision function $\nu_n(\boldsymbol{x}) = B_1^{-1} \sum_{j=1}^{B_1} C_n^{T_{j*}-S_{j*}}(\boldsymbol{x})$

Set the threshold $\hat{\alpha}$ according to (1)

Compute the selected proportion of each method $\boldsymbol{\zeta} = (\zeta_1, \cdots, \zeta_M)^T$, where $\zeta_i = B_1^{-1} \sum_{j=1}^{B_1} \mathbb{1}(i \in T_{j*})$

For each method $T_i, i = 1, \cdots, M$, compute the selected proportion of each feature $\boldsymbol{\eta}_i = (\eta_{i1}, \cdots, \eta_{ip})^T$, where $\eta_{il} = (\zeta_i B_1)^{-1} \sum_{j=1}^{B_1} \mathbb{1}(i \in T_{j*}) \mathbb{1}(l \in S_{j*}), l = 1, \cdots, p$

Output the predicted label $C_n^{RaSE}(\mathbf{x}) = \mathbb{1}(\nu_n(\boldsymbol{x}) > \hat{\alpha})$, the selected proportion of each method $\boldsymbol{\zeta} = (\zeta_1, \cdots, \zeta_M)^T$, and the selected proportion of each feature for each method $\boldsymbol{\eta}_i = (\eta_{i1}, \cdots, \eta_{ip})^T$

Iterative Super RaSE

Main Idea: update the base classification distribution, as well as the subspace distribution for each base classifer.

- Set $\mathbb{D}^{(t+1)}$ to be a discrete distribution over the candidate base classifier set $\mathcal{T}$, where for each base classifier $T_i \in \mathcal{T}$, $P(T_i) = \zeta_i^{(t)}$, where $\zeta_i^{(t)} = B_1^{-1} \sum_{j=1}^{B_1} \mathbb{1}(i \in T_{j*}^{(t)})$
- For each method $T_i, i = 1, \cdots, M$, compute $\eta_{il}^{(t)} = (\zeta_i^{(t)} B_1)^{-1} \sum_{j=1}^{B_1} \mathbb{1}(i \in T_{j*}^{(t)}) \mathbb{1}(l \in S_{j*}^{(t)}), l = 1, \cdots, p$
- Set $\mathcal{D}^{(t+1)}$ to be a restrictive multinomial distribution with parameter $(p, d, \tilde{\boldsymbol{\eta}}_i^{(t)})$, where $\tilde{\eta}_{il}^{(t)} = \eta_{il}^{(t)} \mathbb{1}(\eta_{il}^{(t)} > C_0 / \log p) + \frac{C_0}{p} \mathbb{1}(\eta_{il}^{(t)} \leq C_0 / \log p)$ and d is sampled from the uniform distribution over $\{1, \cdots, D\}$

Outline

QDA

Recall: $\boldsymbol{x} \sim \pi_0 N(\boldsymbol{\mu}^{(0)}_{p\times 1}, \Sigma^{(0)}_{p\times p}) + \pi_1 N(\boldsymbol{\mu}^{(1)}_{p\times 1}, \Sigma^{(1)}_{p\times p}), \pi_0 + \pi_1 = 1.$

$$\log\left(\frac{f^{(0)}(\boldsymbol{x})}{f^{(1)}(\boldsymbol{x})}\right) = \frac{1}{2}\boldsymbol{x}^T\Omega\boldsymbol{x} + \boldsymbol{\delta}^T\boldsymbol{x} + C,$$

where C is a constant unrelated to $\boldsymbol{x}$, and $\Omega = (\Sigma^{(1)})^{-1} - (\Sigma^{(0)})^{-1}$, $\boldsymbol{\delta} = (\Sigma^{(0)})^{-1}\boldsymbol{\mu}^{(0)} - (\Sigma^{(1)})^{-1}\boldsymbol{\mu}^{(1)}$.

QDA

Recall: $\boldsymbol{x} \sim \pi_0 N(\boldsymbol{\mu}_{p\times 1}^{(0)}, \Sigma_{p\times p}^{(0)}) + \pi_1 N(\boldsymbol{\mu}_{p\times 1}^{(1)}, \Sigma_{p\times p}^{(1)}), \pi_0 + \pi_1 = 1.$

$$\log\left(\frac{f^{(0)}(\boldsymbol{x})}{f^{(1)}(\boldsymbol{x})}\right) = \frac{1}{2}\boldsymbol{x}^T\Omega\boldsymbol{x} + \boldsymbol{\delta}^T\boldsymbol{x} + C,$$

where C is a constant unrelated to $\boldsymbol{x}$, and $\Omega = (\Sigma^{(1)})^{-1} - (\Sigma^{(0)})^{-1}$, $\boldsymbol{\delta} = (\Sigma^{(0)})^{-1}\boldsymbol{\mu}^{(0)} - (\Sigma^{(1)})^{-1}\boldsymbol{\mu}^{(1)}$.

- Setting (Fan et al., 2015): $\Omega^{(0)} = (\Sigma^{(0)})^{-1}$ is a $p \times p$ band matrix with $(\Omega^{(0)})_{ii} = 1$ and $(\Omega^{(0)})_{ik} = 0.3$ for $|i-k| = 1$. $\pi_0 = \pi_1 = 0.5$. Ω is a 200×200 sparse symmetric matrix with $\Omega_{10,10} = -0.3758$, $\Omega_{10,30} = 0.0616, \Omega_{10,50} = 0.2037, \Omega_{30,30} = -0.5482, \Omega_{30,50} = 0.0286, \Omega_{50,50} = -0.4614$. $\boldsymbol{\delta} = (0.6, 0.8, \mathbf{0}_{198})^T$.

QDA

Recall: $\boldsymbol{x} \sim \pi_0 N(\boldsymbol{\mu}_{p\times 1}^{(0)}, \Sigma_{p\times p}^{(0)}) + \pi_1 N(\boldsymbol{\mu}_{p\times 1}^{(1)}, \Sigma_{p\times p}^{(1)}), \pi_0 + \pi_1 = 1.$

$$\log\left(\frac{f^{(0)}(\boldsymbol{x})}{f^{(1)}(\boldsymbol{x})}\right) = \frac{1}{2}\boldsymbol{x}^T\Omega\boldsymbol{x} + \boldsymbol{\delta}^T\boldsymbol{x} + C,$$

where C is a constant unrelated to $\boldsymbol{x}$, and $\Omega = (\Sigma^{(1)})^{-1} - (\Sigma^{(0)})^{-1}$, $\boldsymbol{\delta} = (\Sigma^{(0)})^{-1}\boldsymbol{\mu}^{(0)} - (\Sigma^{(1)})^{-1}\boldsymbol{\mu}^{(1)}$.

- Setting (Fan et al., 2015): $\Omega^{(0)} = (\Sigma^{(0)})^{-1}$ is a $p \times p$ band matrix with $(\Omega^{(0)})_{ii} = 1$ and $(\Omega^{(0)})_{ik} = 0.3$ for $|i-k| = 1$. $\pi_0 = \pi_1 = 0.5$. Ω is a 200×200 sparse symmetric matrix with $\Omega_{10,10} = -0.3758$, $\Omega_{10,30} = 0.0616, \Omega_{10,50} = 0.2037, \Omega_{30,30} = -0.5482, \Omega_{30,50} = 0.0286, \Omega_{50,50} = -0.4614$. $\boldsymbol{\delta} = (0.6, 0.8, \mathbf{0}_{198})^T$.
- Minimal discriminative set $S^* = \{j : \boldsymbol{\delta}_j \neq 0\} \cup \{j : \Omega_{ij} \neq 0, \exists i\}$ $= \{1, 2, 10, 30, 50\}$.

QDA

Recall: $\boldsymbol{x} \sim \pi_0 N(\boldsymbol{\mu}_{p\times 1}^{(0)}, \Sigma_{p\times p}^{(0)}) + \pi_1 N(\boldsymbol{\mu}_{p\times 1}^{(1)}, \Sigma_{p\times p}^{(1)}), \pi_0 + \pi_1 = 1.$

$$\log\left(\frac{f^{(0)}(\boldsymbol{x})}{f^{(1)}(\boldsymbol{x})}\right) = \frac{1}{2}\boldsymbol{x}^T\Omega\boldsymbol{x} + \boldsymbol{\delta}^T\boldsymbol{x} + C,$$

where C is a constant unrelated to $\boldsymbol{x}$, and $\Omega = (\Sigma^{(1)})^{-1} - (\Sigma^{(0)})^{-1}$, $\boldsymbol{\delta} = (\Sigma^{(0)})^{-1}\boldsymbol{\mu}^{(0)} - (\Sigma^{(1)})^{-1}\boldsymbol{\mu}^{(1)}$.

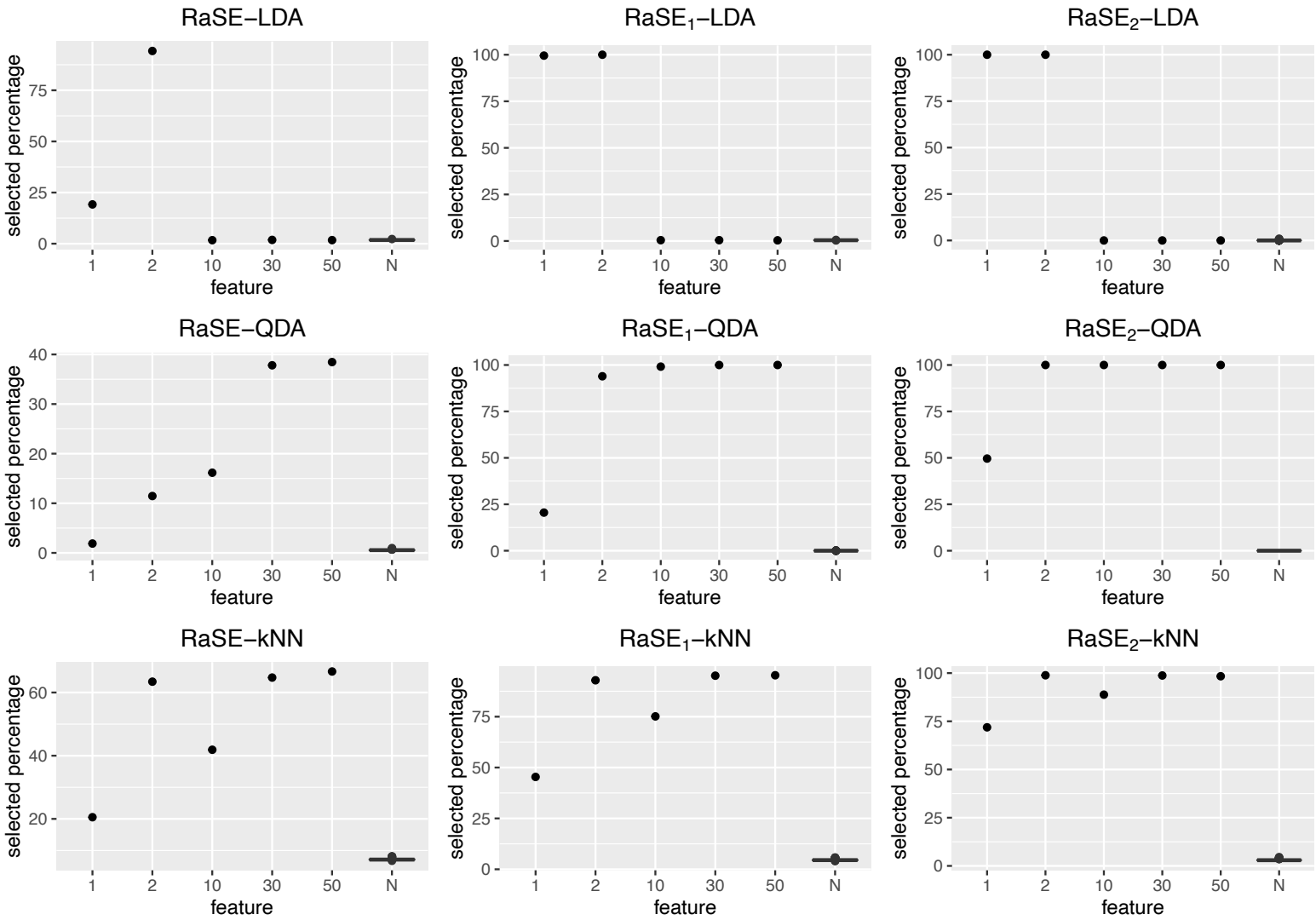
- Setting (Fan et al., 2015): $\Omega^{(0)} = (\Sigma^{(0)})^{-1}$ is a $p \times p$ band matrix with $(\Omega^{(0)})_{ii} = 1$ and $(\Omega^{(0)})_{ik} = 0.3$ for $|i-k| = 1$. $\pi_0 = \pi_1 = 0.5$. Ω is a 200×200 sparse symmetric matrix with $\Omega_{10,10} = -0.3758$, $\Omega_{10,30} = 0.0616, \Omega_{10,50} = 0.2037, \Omega_{30,30} = -0.5482, \Omega_{30,50} = 0.0286, \Omega_{50,50} = -0.4614$. $\boldsymbol{\delta} = (0.6, 0.8, \mathbf{0}_{198})^T$.
- Minimal discriminative set $S^* = \{j : \boldsymbol{\delta}_j \neq 0\} \cup \{j : \Omega_{ij} \neq 0, \exists i\}$ $= \{1, 2, 10, 30, 50\}$.
- Training data size $n \in \{200, 400, 1000\}$. Test data size is 1000. Repeat for 200 times.

Test error

Table: Summary of test classification error rates for each classifier under various sample sizes over 200 repetitions. The results are presented as mean values with the standard deviations in parentheses

	$n = 200$	$n = 400$	$n = 1000$
SRaSE	30.82(3.29)	28.68(3.23)	26.12(2.66)
SRaSE_1	*27.58*(2.33)	*24.64*(1.85)	*23.03*(1.38)
SRaSE_2	*27.36*(2.67)	**24.04**(1.74)	**22.63**(1.41)
RaSE-LDA	37.3(3.17)	36.11(1.97)	35.67(1.73)
RaSE-QDA	32.52(2.90)	30.44(2.60)	29(1.97)
RaSE-KNN	31.1(3.23)	27.83(2.41)	25.22(1.56)
RaSE_1-LDA	36.09(2.87)	32.82(1.74)	32.68(1.49)
RaSE_1-QDA	**26.83**(2.47)	*25.07*(1.89)	*23.53*(1.50)
RaSE_1-KNN	*28.76*(2.60)	25.88(1.98)	24.18(1.47)
RaSE_2-LDA	38.09(2.48)	33.69(1.83)	32.71(1.55)
RaSE_2-QDA	*26.99*(2.68)	*24.87*(1.99)	*23.11*(1.60)
RaSE_2-KNN	*28.73*(2.56)	*25.46*(1.82)	*23.76*(1.54)
LDA	49.03(1.94)	42.88(1.82)	38.68(1.70)
QDA	NA	NA	45.13(1.58)
KNN	45.67(1.78)	44.63(2.02)	43.43(1.63)
RF	37.34(2.91)	31.61(2.19)	27.42(1.60)

Selected percentage of features in RaSE



Selected percentage of base classifier in Super RaSE

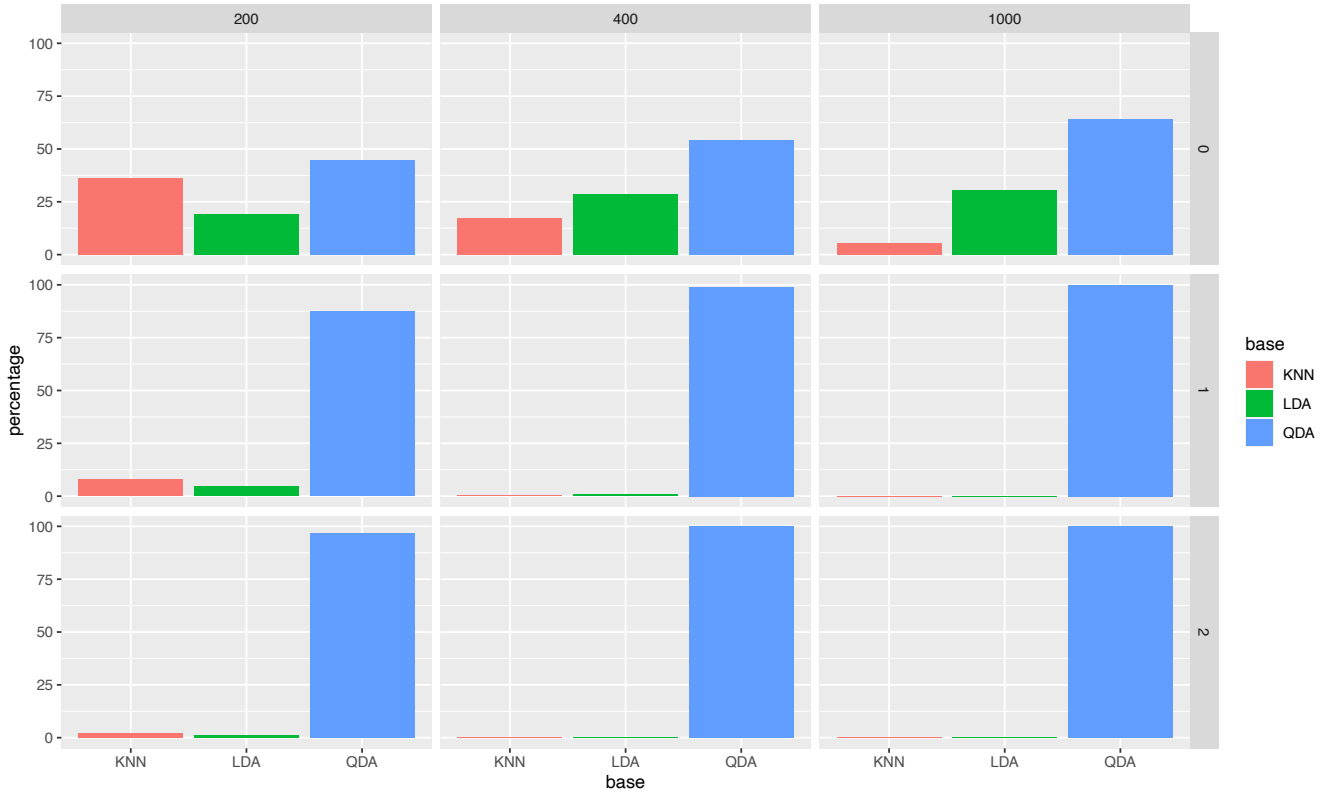


Figure: The average selected proportion for each base method for different sample sizes (corresponding to each column) and iteration number (corresponding to each row) in Model 2 (QDA).

Madelon

- An artificial dataset from NIPS 2003 feature selection challenge. (Guyon et al., 2005)
- It is generated from 32 clusters (placed on vertices of hypercube), which are assigned class 0, 1 randomly.
- 20 out of 500 features are signals.
- Total number of observations: 2000 = 1000 (class 0) + 1000 (class 1)
- Training data size: $n \in \{200, 500, 1000\}$.
- The remained data is used as test data. Repeat the random split 200 times.

Test error

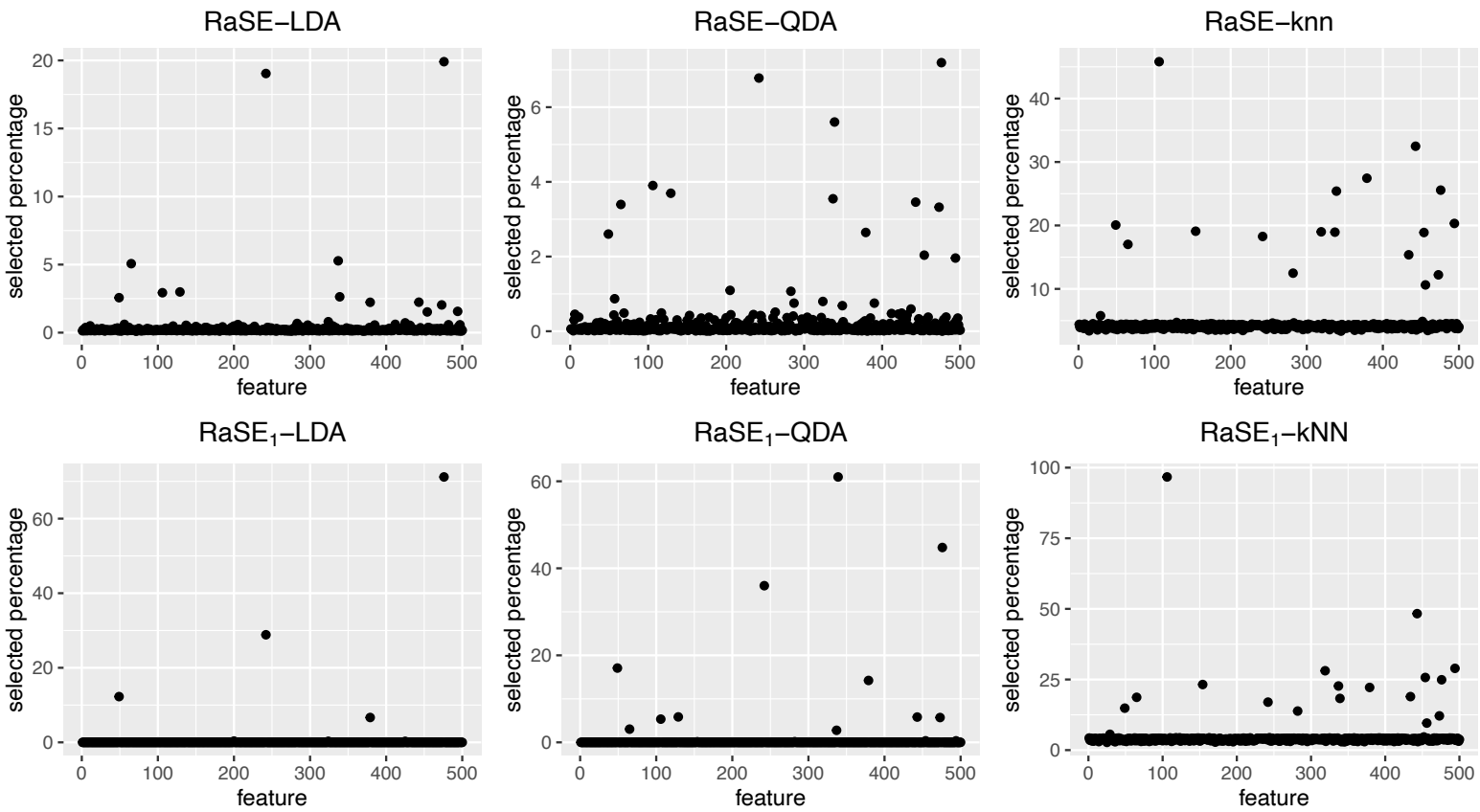
Method	$n = 200$	$n = 500$	$n = 1000$
RaSE-LDA	$44.13_{3.73}$	$39.69_{1.60}$	$39.16_{1.27}$
RaSE-QDA	$44.55_{3.50}$	$40.45_{1.71}$	$39.89_{1.51}$
RaSE-kNN	$34.89_{3.10}$	$26.49_{2.71}$	$21.35_{1.94}$
RaSE$_1$-LDA	$45.98_{3.00}$	$40.16_{2.39}$	$38.69_{1.11}$
RaSE$_1$-QDA	$45.01_{5.32}$	$37.73_{3.28}$	$34.22_{2.24}$
RaSE$_1$-kNN	$31.71_{4.10}$	$21.09_{2.53}$	$18.97_{1.73}$
RP-LDA	$41.34_{1.84}$	$39.85_{1.14}$	$39.53_{1.32}$
RP-QDA	$40.03_{1.63}$	$39.31_{1.59}$	$38.94_{1.61}$
RP-kNN	$40.15_{1.79}$	$39.07_{1.42}$	$38.54_{1.46}$
LDA	---†	$49.71_{1.37}$	$47.46_{1.37}$
QDA	---†	---†	---†
kNN	$36.23_{1.72}$	$31.68_{1.43}$	$28.61_{1.37}$
sLDA	$43.18_{3.30}$	$40.66_{2.18}$	$39.50_{1.29}$
RAMP	$48.86_{3.67}$	$42.33_{5.30}$	$38.56_{1.12}$
NSC	$42.07_{2.83}$	$40.21_{1.22}$	$40.10_{1.24}$
RF	$44.23_{2.90}$	$38.52_{1.58}$	$34.46_{1.48}$

*: the best classifier

**: the one within 1 sd

---†: not applicable

Selected percentage of features



Outline

Summary

- We introduced a new ensemble classification framework RaSE, which enjoys the following properties:

Summary

- We introduced a new ensemble classification framework RaSE, which enjoys the following properties:
  - It can be coupled with any base classifier.

Summary

- We introduced a new ensemble classification framework RaSE, which enjoys the following properties:
  - It can be coupled with any base classifier.
  - It provides an intuitive way for feature ranking and screening.

Summary

- We introduced a new ensemble classification framework RaSE, which enjoys the following properties:
  - It can be coupled with any base classifier.
  - It provides an intuitive way for feature ranking and screening.
- We connected random subspace method and the sparse classification problems, and studied the theoretical properties of RaSE.

Summary

- We introduced a new ensemble classification framework RaSE, which enjoys the following properties:
  - It can be coupled with any base classifier.
  - It provides an intuitive way for feature ranking and screening.
- We connected random subspace method and the sparse classification problems, and studied the theoretical properties of RaSE.
- The effectiveness of RaSE was verified via extensive numerical experiments.
- We also proposed the Super RaSE, which can work with a base classifier set.

Thanks!

- R package `RaSEn` is available on CRAN: `https://cran.r-project.org/web/packages/RaSEn/`
- Tian, Y. & Feng, Y. (2021). RaSE: Random Subspace Ensemble Classification. Journal of Machine Learning Research.
- Tian, Y. & Feng, Y. (2021). RaSE: A variable screening framework via random subspace ensembles. Journal of the American Statistical Association.
- Zhu, J. & Feng, Y. (2021). Super RaSE: Super Random Subspace Ensemble Classification. Manuscript.

References I

Ahn, H., Moon, H., Fazzari, M. J., Lim, N., Chen, J. J., and Kodell, R. L. (2007). Classification by ensembles from random partitions of high-dimensional data. *Computational Statistics & Data Analysis*, 51(12):6166--6179.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. Proceeding of IEEE international symposium on information theory.

Blaser, R. and Fryzlewicz, P. (2016). Random rotation ensembles. *The Journal of Machine Learning Research*, 17(1):126--151.

Breiman, L. (1996). Bagging predictors. *Machine learning*, 24(2):123--140.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5--32.

Bryll, R., Gutierrez-Osuna, R., and Quek, F. (2003). Attribute bagging: improving accuracy of classifier ensembles by using random feature subsets. *Pattern recognition*, 36(6):1291--1302.

References II

Cannings, T. I. and Samworth, R. J. (2017). Random-projection ensemble classification. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(4):959--1035.

Chen, J. and Chen, Z. (2008). Extended bayesian information criteria for model selection with large model spaces. *Biometrika*, 95(3):759--771.

Chen, J. and Chen, Z. (2012). Extended bic for small-n-large-p sparse glm. *Statistica Sinica*, pages 555--574.

Efron, B. (1975). The efficiency of logistic regression compared to normal discriminant analysis. *Journal of the American Statistical Association*, 70(352):892--898.

Fan, Y., Kong, Y., Li, D., Zheng, Z., et al. (2015). Innovated interaction screening for high-dimensional nonlinear classification. *The Annals of Statistics*, 43(3):1243--1272.

References III

Fan, Y. and Tang, C. Y. (2013). Tuning parameter selection in high dimensional penalized likelihood. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(3):531--552.

Freund, Y. and Schapire, R. E. (1995). A desicion-theoretic generalization of on-line learning and an application to boosting. In *European conference on computational learning theory*, pages 23--37. Springer.

Guyon, I., Gunn, S., Ben-Hur, A., and Dror, G. (2005). Result analysis of the nips 2003 feature selection challenge. In *Advances in neural information processing systems*, pages 545--552.

Hall, P., Park, B. U., Samworth, R. J., et al. (2008). Choice of neighbor order in nearest-neighbor classification. *The Annals of Statistics*, 36(5):2135--2152.

Ho, T. K. (1998). The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(8):832--844.

References IV

Kohavi, R., John, G. H., et al. (1997). Wrappers for feature subset selection. *Artificial intelligence*, 97(1-2):273--324.

Li, Q. and Shao, J. (2015). Sparse quadratic discriminant analysis for high dimensional data. *Statistica Sinica*, pages 457--473.

Mai, Q., Zou, H., and Yuan, M. (2012). A direct approach to sparse discriminant analysis in ultra-high dimensions. *Biometrika*, 99(1):29--42.

Samworth, R. J. et al. (2012). Optimal weighted nearest neighbour classifiers. *The Annals of Statistics*, 40(5):2733--2763.

Schwarz, G. et al. (1978). Estimating the dimension of a model. *The annals of statistics*, 6(2):461--464.

Tian, Y. and Feng, Y. (2021). Rase: Random subspace ensemble classification. *J. Mach. Learn. Res.*, 22:45--1.

Zhang, Q. and Wang, H. (2011). On bic's selection consistency for discriminant analysis. *Statistica Sinica*, pages 731--740.

Assumptions for QDA

Recall QDA model:

$$\boldsymbol{x} \sim \pi_0 N(\boldsymbol{\mu}_{p\times 1}^{(0)}, \Sigma_{p\times p}^{(0)}) + \pi_1 N(\boldsymbol{\mu}_{p\times 1}^{(1)}, \Sigma_{p\times p}^{(1)}), \pi_0 + \pi_1 = 1.$$

Assumptions for QDA

Recall QDA model:
$\boldsymbol{x} \sim \pi_0 N(\boldsymbol{\mu}^{(0)}_{p\times 1}, \Sigma^{(0)}_{p\times p}) + \pi_1 N(\boldsymbol{\mu}^{(1)}_{p\times 1}, \Sigma^{(1)}_{p\times p}), \pi_0 + \pi_1 = 1.$

Suppose the following conditions are satisfied, where m, M, M' are constants, and denote $\Omega^{(r)}_{S,S} = (\Sigma^{(r)}_{S,S})^{-1}$:

- **Condition 1**: $\lambda_{\min}(\Sigma^{(r)}) \geq m > 0, \lambda_{\max}(\Sigma^{(r)}) \leq M < \infty, r = 0, 1;$
- **Condition 2**: $\|\boldsymbol{\mu}^{(1)} - \boldsymbol{\mu}^{(0)}\|_\infty \leq M' < \infty;$
- **Condition 3**: Denote $\gamma_l = \inf_j \left|(\Omega^{(1)}_{S,S}\boldsymbol{\mu}^{(1)}_S - \Omega^{(0)}_{S,S}\boldsymbol{\mu}^{(0)}_S)_j\right| > 0, \gamma_q = \inf_i \sup_j \left|(\Omega^{(1)}_{S^*_q,S^*_q} - \Omega^{(0)}_{S^*_q,S^*_q})_{ij}\right| > 0$, then

$$\min\{\gamma_l^2, \gamma_q^2, \gamma_q\} \gg D^2\sqrt{\frac{\log p}{n}} = o(1).$$

Consistency

QDA consistency of RIC (Tian and Feng (2021))

For QDA model, under Assumptions for LDA, we have
(i) If $Dc_n/\gamma^2 = o(1)$, then the following screening consistency holds for RIC: If $D^2 c_n / \min\{\gamma_l^2, \gamma_q^2, \gamma_q\} = o(1)$, then RIC is screening consistent:

$$\mathbb{P}\left(\sup_{\substack{S: S \supseteq S^* \\ |S| \le D}} \mathsf{RIC}_n(S) < \inf_{\substack{S: S \not\supseteq S^* \\ |S| \le D}} \mathsf{RIC}_n(S)\right) \ge 1 - O\left(p^2 \exp\left\{-Cn\left(\frac{\min\{\gamma_l^2, \gamma_q^2, \gamma_q\}}{D^2}\right)^2\right\}\right)$$

$$\to 1.$$

(ii) Further, if $c_n \gg D^2\sqrt{\frac{\log p}{n}}$, then RIC is weakly consistent:

$$\mathbb{P}\left(\mathsf{RIC}_n(S^*) = \inf_{S: |S| \le D} \mathsf{RIC}_n(S)\right) \ge 1 - O\left(p^2 \exp\left\{-Cn\left(\frac{c_n}{D^2}\right)^2\right\}\right) \to 1.$$

In practice, we set $c_n = n^{-1/2} \log\log n$.

Consistency

- Under some conditions, similar results can be extended to more general setting.

The requirement of B_2 in vanilla RaSE

Corollary (Tian and Feng, 2021):
By using RIC (or any other criterion), we have

$$\mathbb{P}(S_{1*} \supseteq S^*) \geq \underbrace{\mathbb{P}\left(\sup_{\substack{S:S\supseteq S^* \\ |S|\leq D}} \mathrm{RIC}_n(S) < \inf_{\substack{S:S\not\supseteq S^* \\ |S|\leq D}} \mathrm{RIC}_n(S)\right)}_{\to 1 \text{ by screening consistency}} \cdot \mathbf{P}\left(\bigcup_{j=1}^{B_2}\{S_{1j} \supseteq S^*\}\right),$$

where

$$\mathbf{P}\left(\bigcup_{j=1}^{B_2}\{S_{1j} \supseteq S^*\}\right) = 1-(1-p_{S^*})^{B_2} \geq 1 - O\left(\exp\left\{-B_2 p_{S^*}\right\}\right).$$

Here $p_{S^*} = \mathbf{P}(S_{11} \supseteq S^*) = \frac{1}{D}\sum_{p^*\leq d\leq D} \frac{\binom{p-p^*}{d-p^*}}{\binom{p}{d}}$. We hope there holds

$$B_2 p_{S^*} \gg 1 \quad \Rightarrow \quad B_2 \gg \left(\frac{p-p^*+1}{D}\right)^{p^*}.$$

Relax the requirement of B_2 using iterative RaSE

Under a set of conditions, where we assume $D \log\log p \ll \log p$, we have the following result. $\bar{p}^*$ is a positive integer smaller than p (defined in the conditions, omitted here).

Sure coverage by iterative RaSE (Tian and Feng (2021))

For Algorithm 2, the B_2 in the first step is set as

$$Dp^{\bar{p}^*} \lesssim B_2 \ll \left(\frac{p}{\bar{p}^* D}\right)^{\bar{p}^*+1},$$

and B_2 in the following steps is set as

$$(D + C_0)^D p^{\bar{p}^*} (\log p)^{p^*} \lesssim B_2 \ll \left(\frac{p}{\bar{p}^* D}\right)^{\bar{p}^*+1}.$$

Set B_1 such that $B_1 \gg \log p^*$. After $T \geq \lceil \frac{p^*}{\bar{p}^*} \rceil$ iterations, as $n, B_2 \to \infty$ there holds

$$\mathbb{P}(S_{1*}^{(T)} \not\supseteq S^*) \to 0.$$

Relax the requirement of B_2 using iterative RaSE

- Now suppose D, p^* are all fixed constants.

Relax the requirement of B_2 using iterative RaSE

- Now suppose D, p^* are all fixed constants.
- A sufficient condition for B_2 in iterative RaSE to achieve sure coverage is

$$B_2 \gtrsim p^{\bar{p}^*} (\log p)^{p^*}.$$

When $\bar{p}^* < p$, this is much weaker than the requirement

$$B_2 \gg p^{p^*}$$

for vanilla RaSE implied by the corollary.