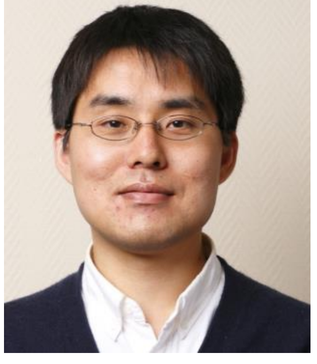
Average Sensitivity of Decision Tree Learning

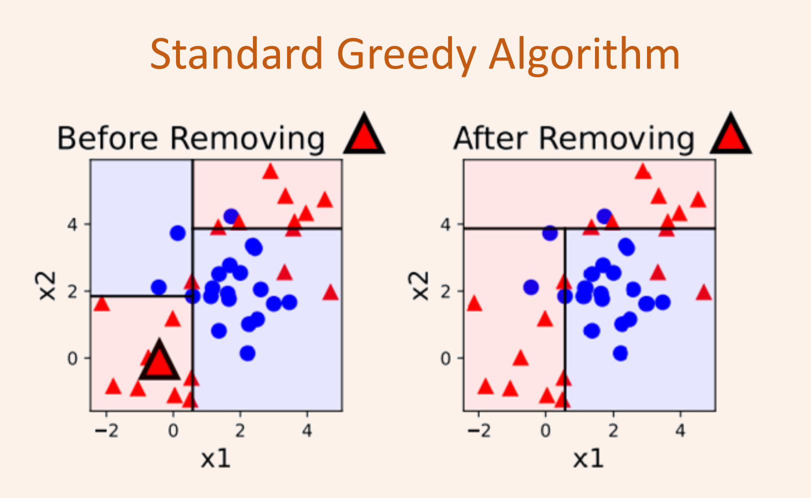
Satoshi Hara
(Osaka Univ., Japan)

Yuichi Yoshida
(NII, Japan)



Instability of Decision Tree (DT) Learning

- When one data point is removed



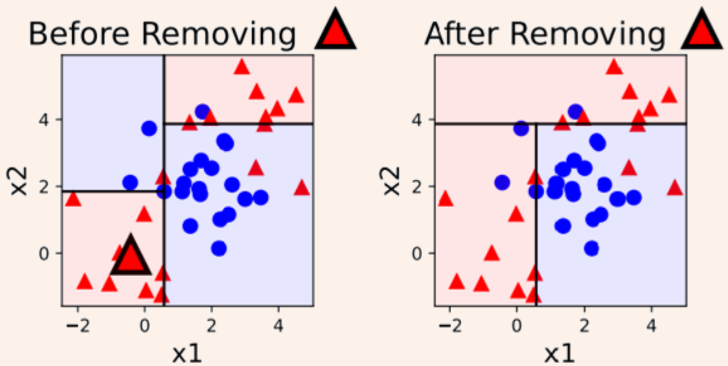
Removal of ▲ induces
a <u>completely different</u> tree.
→ "Unstable" Learning Algorithm

Not intuitive
≈ Less reliable

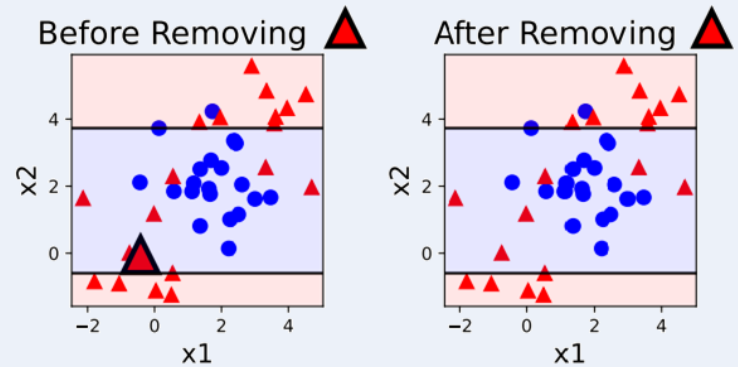
Instability of Decision Tree (DT) Learning

- When one data point is removed

Standard Greedy Algorithm



Proposed Stable Algorithm



Removal of ▲ induces
a <u>completely different</u> tree.
→ "Unstable" Learning Algorithm

Removal of ▲ induces
an <u>almost same</u> tree.
→ "Stable" Learning Algorithm

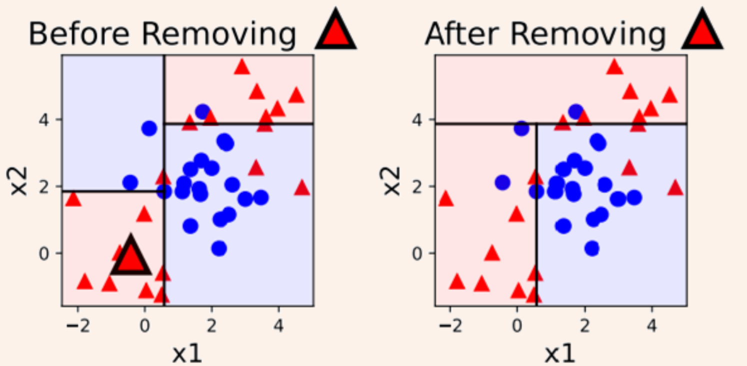
Not intuitive
≈ Less reliable

Intuitive
≈ Reliable

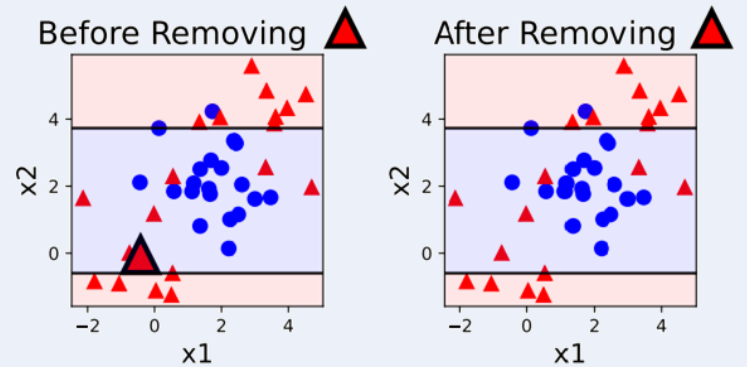
Instability of Decision Tree (DT) Learning

- When one data point is removed

Standard Greedy Algorithm



Proposed Stable Algorithm



Removal of ▲ induces a <u>completely different</u> tree.
→ "Unstable" Learning Algorithm

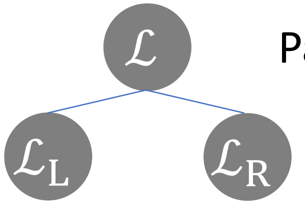
Removal of ▲ induces an <u>almost same</u> tree.
→ "Stable" Learning Algorithm

Contributions

- 1. Stable DT Learning Algorithm
- 2. Stability Guarantee

Contribution 1.

Stable DT Learning Algorithm



Dataset $\mathcal{L}$

x_1	x_2	y
5.2	1.1	1
-1.3	-1.8	0
0.9	0.3	1
0.1	2.5	1
3.2	-0.3	0

Partition by $\omega = [x_1 \geq 1]$

$\mathcal{L}_\mathrm{L}$

x_1	x_2	y
-1.3	-1.8	0
0.9	0.3	1
0.1	2.5	1

$\mathcal{L}_\mathrm{R}$

x_1	x_2	y
5.2	1.1	1
3.2	-0.3	0

$\mathrm{Accuracy}(\omega) = 3/5$

$\omega = [x_2 \geq 1]$

$\mathcal{L}_\mathrm{L}$

x_1	x_2	y
-1.3	-1.8	0
0.9	0.3	1

$\mathcal{L}_\mathrm{R}$

x_1	x_2	y
5.2	1.1	1
0.9	0.3	1
3.2	-0.3	0

$\mathrm{Accuracy}(\omega) = 3/5$

$\omega = [x_2 \geq 0]$

$\mathcal{L}_\mathrm{L}$

x_1	x_2	y
-1.3	-1.8	0
3.2	-0.3	0

$\mathcal{L}_\mathrm{R}$

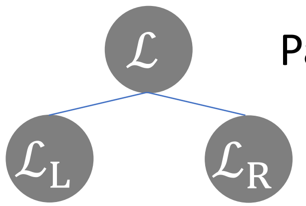
x_1	x_2	y
5.2	1.1	1
0.9	0.3	1
0.1	2.5	1

$\mathrm{Accuracy}(\omega) = 5/5$

Standard Greedy Alg. $\hat{\omega} = \mathrm{argmax}_\omega \, \mathrm{Accuracy}(\omega)$

Contribution 1.

Stable DT Learning Algorithm



Dataset $\mathcal{L}$

x_1	x_2	y
5.2	1.1	1
-1.3	-1.8	0
0.9	0.3	1
0.1	2.5	1
3.2	-0.3	0

Partition by $\omega = [x_1 \geq 1]$

$\mathcal{L}_\mathrm{L}$

x_1	x_2	y
-1.3	-1.8	0
0.9	0.3	1
0.1	2.5	1

$\mathcal{L}_\mathrm{R}$

x_1	x_2	y
5.2	1.1	1
3.2	-0.3	0

$\mathrm{Accuracy}(\omega) = 3/5$

$\omega = [x_2 \geq 1]$

$\mathcal{L}_\mathrm{L}$

x_1	x_2	y
-1.3	-1.8	0
0.9	0.3	1

$\mathcal{L}_\mathrm{R}$

x_1	x_2	y
5.2	1.1	1
0.9	0.3	1
3.2	-0.3	0

$\mathrm{Accuracy}(\omega) = 3/5$

$\omega = [x_2 \geq 0]$

$\mathcal{L}_\mathrm{L}$

x_1	x_2	y
-1.3	-1.8	0
3.2	-0.3	0

$\mathcal{L}_\mathrm{R}$

x_1	x_2	y
5.2	1.1	1
0.9	0.3	1
0.1	2.5	1

$\mathrm{Accuracy}(\omega) = 5/5$

Standard Greedy Alg. $\hat{\omega} = \mathrm{argmax}_\omega \, \mathrm{Accuracy}(\omega)$

Proposed Stable Alg. $\hat{\omega} \sim \Pr[\omega] \propto \exp\left(\frac{C}{\epsilon}\mathrm{Accuracy}(\omega)\right)$

Contribution 2.

Stability Guarantee

- **Average Sensitivity** [Varma & Yoshida, SODA' 21]
  - The average difference of the learned trees before/after one data point removal.

$$\frac{1}{n}\sum_{i=1}^{n} d_{\mathrm{DT}}\left(\mathrm{DT}_{\mathcal{L}}, \mathrm{DT}_{\mathcal{L}\setminus\{i\}}\right)$$

- $\mathcal{L}$: the training set $\mathcal{L} = \{x_i, y_i\}_{i=1}^{n}$
- n: the size of $\mathcal{L}$

Contribution 2.

Stability Guarantee

- Average Sensitivity [Varma & Yoshida, SODA' 21]
  - The average difference of the learned trees before/after one data point removal.

DT trained using the whole training set $\mathcal{L}$

$$\frac{1}{n}\sum_{i=1}^{n} d_{\mathrm{DT}}\left(\mathrm{DT}_{\mathcal{L}}, \mathrm{DT}_{\mathcal{L}\setminus\{i\}}\right)$$

- $\mathcal{L}$: the training set $\mathcal{L} = \{x_i, y_i\}_{i=1}^{n}$
- n: the size of $\mathcal{L}$

Contribution 2.

Stability Guarantee

- Average Sensitivity [Varma & Yoshida, SODA'21]
  - The average difference of the learned trees before/after one data point removal.

DT trained using the whole training set $\mathcal{L}$

DT trained without the i-th data point

$$\frac{1}{n}\sum_{i=1}^{n} d_{\mathrm{DT}}\left(\mathrm{DT}_{\mathcal{L}}, \mathrm{DT}_{\mathcal{L}\setminus\{i\}}\right)$$

- $\mathcal{L}$: the training set $\mathcal{L} = \{x_i, y_i\}_{i=1}^{n}$
- n: the size of $\mathcal{L}$

Contribution 2.

Stability Guarantee

- Average Sensitivity [Varma & Yoshida, SODA' 21]
  - The average difference of the learned trees before/after one data point removal.

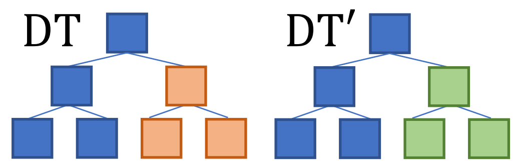
$$\frac{1}{n}\sum_{i=1}^{n} d_{\mathrm{DT}}\left(\mathrm{DT}_{\mathcal{L}}, \mathrm{DT}_{\mathcal{L}\setminus\{i\}}\right)$$

DT trained using the whole training set $\mathcal{L}$

DT trained without the i-th data point

the number of different nodes between trees

- $\mathcal{L}$: the training set $\mathcal{L} = \{x_i, y_i\}_{i=1}^{n}$
- n: the size of $\mathcal{L}$



$d_{\mathrm{DT}}(\mathrm{DT}, \mathrm{DT}') = 6$

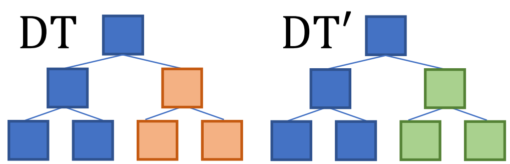
Stability Guarantee

- **Average Sensitivity** [Varma & Yoshida, SODA' 21]
  - The average difference of the learned trees before/after one data point removal.

$$\frac{1}{n}\sum_{i=1}^{n} d_{\mathrm{DT}}\left(\mathrm{DT}_{\mathcal{L}}, \mathrm{DT}_{\mathcal{L}\setminus\{i\}}\right)$$

- $\mathrm{DT}_{\mathcal{L}}$: DT trained using the whole training set $\mathcal{L}$
- $\mathrm{DT}_{\mathcal{L}\setminus\{i\}}$: DT trained without the i-th data point
- d_{DT}: the number of different nodes between trees

- $\mathcal{L}$: the training set $\mathcal{L} = \{x_i, y_i\}_{i=1}^{n}$
- n : the size of $\mathcal{L}$



$d_{\mathrm{DT}}(\mathrm{DT}, \mathrm{DT}') = 6$

Main Result

$$\text{Average Sensitivity of Proposed Alg.} = O\left(\frac{B2^{B}\log|\Omega|}{\epsilon n}\right)$$

- B : the depth of tree
- Ω : the set of splitting rules ω

Contribution 2.

Stability Guarantee

■ Average Sensitivity [Varma & Yoshida, SODA'21]

- The average difference of the learned trees before/after one data point removal.

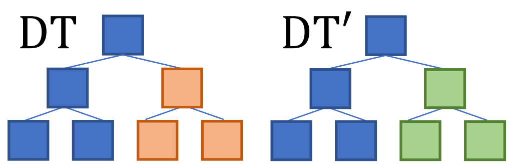
$$\frac{1}{n}\sum_{i=1}^{n} d_{\mathrm{DT}}\left(\mathrm{DT}_{\mathcal{L}}, \mathrm{DT}_{\mathcal{L}\setminus\{i\}}\right)$$

$\mathrm{DT}_{\mathcal{L}}$: DT trained using the whole training set $\mathcal{L}$

$\mathrm{DT}_{\mathcal{L}\setminus\{i\}}$: DT trained without the i-th data point

d_{DT}: the number of different nodes between trees

- $\mathcal{L}$: the training set $\mathcal{L} = \{x_i, y_i\}_{i=1}^{n}$
- n: the size of $\mathcal{L}$



$d_{\mathrm{DT}}(\mathrm{DT}, \mathrm{DT}') = 6$

■ Main Result

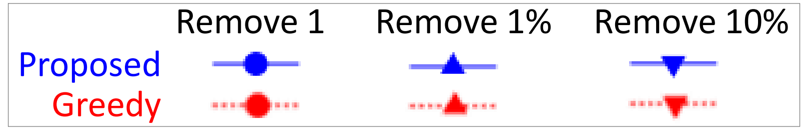
$$\text{Average Sensitivity of Proposed Alg.} = O\left(\frac{B2^B \log|\Omega|}{\epsilon n}\right)$$

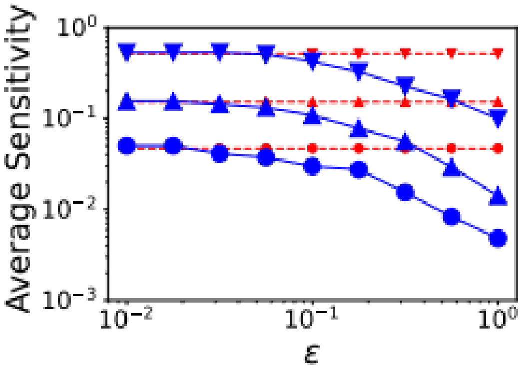
- B: the depth of tree
- Ω: the set of splitting rules ω

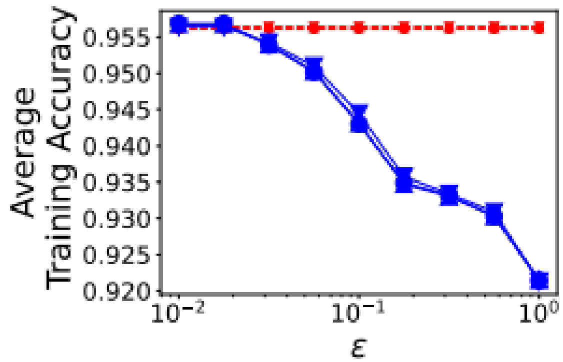
- Typically, $2^B \ll n$ to avoid overfitting.
- Stable when $\epsilon \sim 1$.

Example Results (more in paper)

- Data: breast-cancer





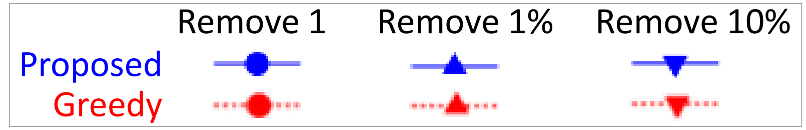


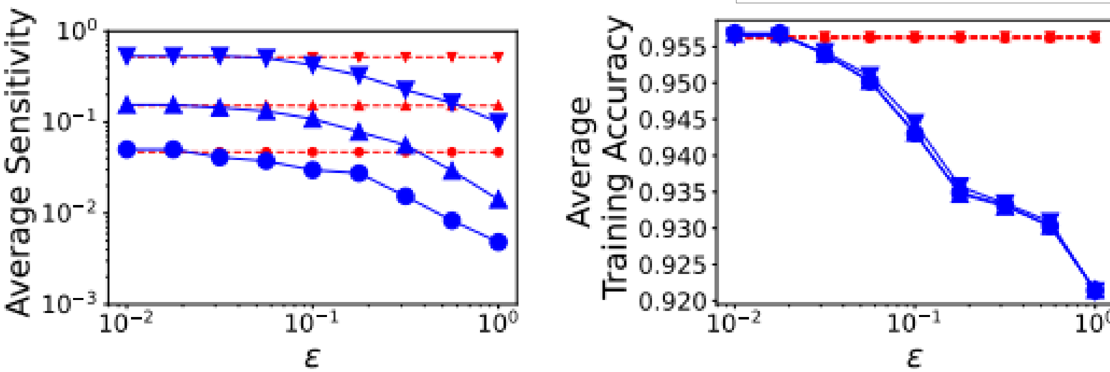
Small ϵ
- Proposed Alg. ~ Greedy
- Less stable
- Higher accuracy

Large ϵ
- Proposed Alg. ~ Random
- More stable
- Lower accuracy

Example Results (more in paper)

- Data: breast-cancer

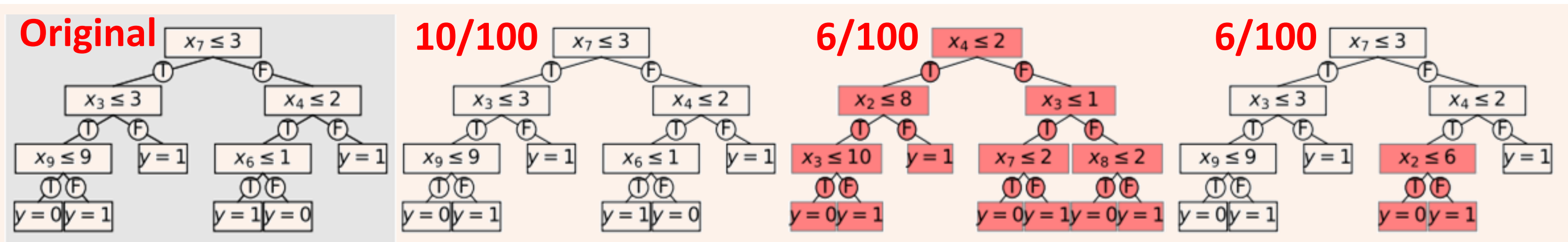




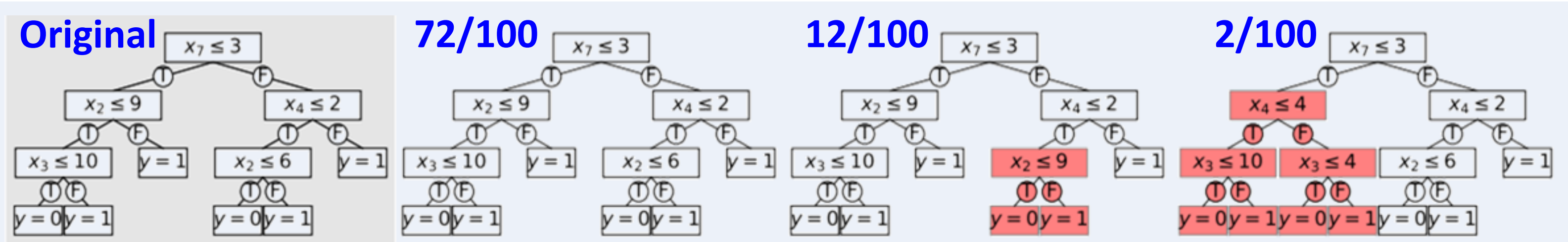
Small ϵ	• Proposed Alg. ~ Greedy • Less stable • Higher accuracy
Large ϵ	• Proposed Alg. ~ Random • More stable • Lower accuracy

- Most frequent tree patterns (100 trials w/ Remove 10%)

Greedy (Unstable)



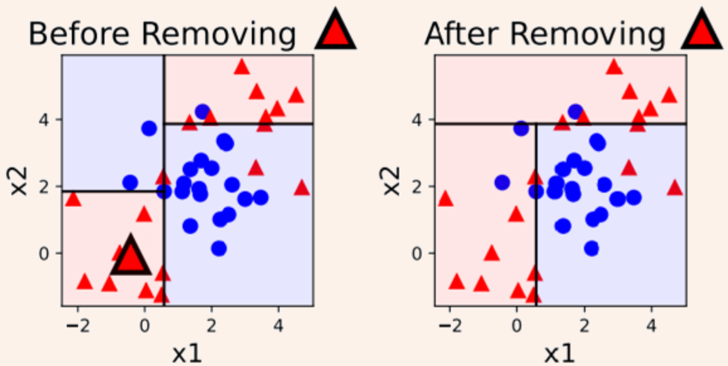
Proposed (Stable)



Instability of Decision Tree (DT) Learning

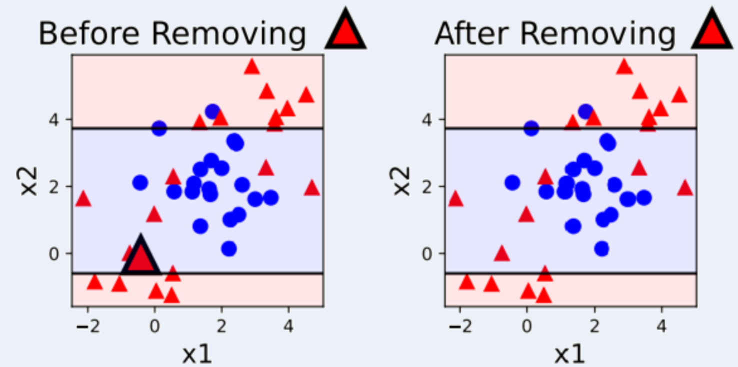
- When one data point is removed

Standard Greedy Algorithm



Removal of ▲ induces
a <u>completely different</u> tree.
→ "Unstable" Learning Algorithm

Proposed Stable Algorithm



Removal of ▲ induces
an <u>almost same</u> tree.
→ "Stable" Learning Algorithm

Contributions

- 1. Stable DT Learning Algorithm
- 2. Stability Guarantee