Margin Distribution and Structural Diversity Guided Ensemble Pruning

Yi-Xiao He, Yu-Chang Wu, Chao Qian, Zhi-Hua Zhou

Contact

{heyx, wuyc, qianc, zhouzh}@lamda.nju.edu.cn

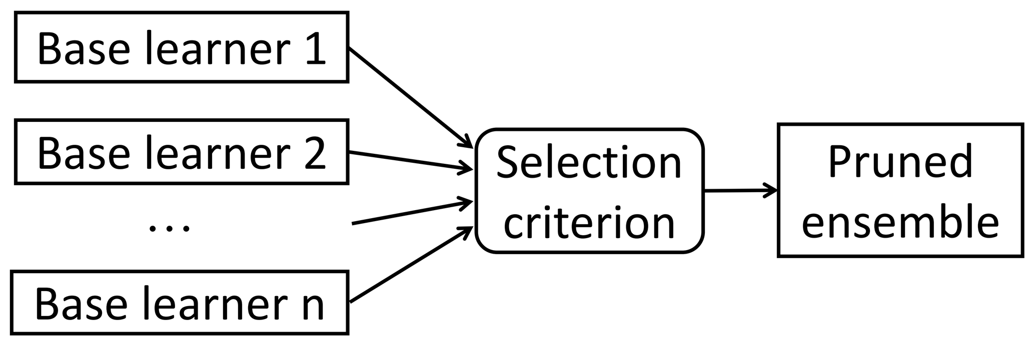
南京大學 NANJING UNIVERSITY 1902

LAMDA Learning And Mining from DatA http://www.lamda.nju.edu.cn

ACML 2023

Background

Ensemble pruning selects and combines a subset of base learners instead of combining them all.



Selection Criteria:
Validation error,
diversity + validation error,
margin or *margin distribution* + validation error, …

Before, *diversity* or *margin* was used **nested** with the validation error. This makes their benefits difficult to analyze.

Take *ordering-based ensemble pruning* as an example.
Diversity or margin: rank the learners.

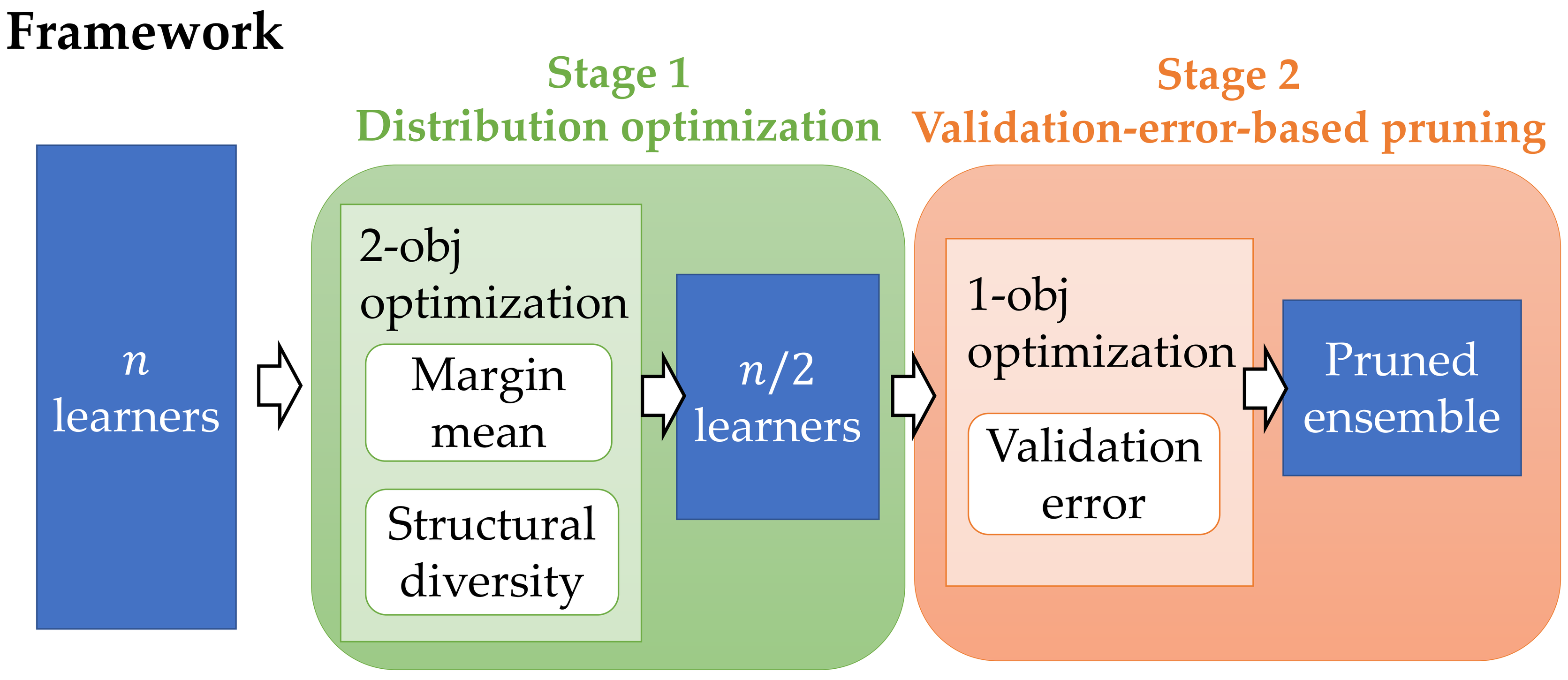
□ □ □ □ □ □ □ …

Validation error: which learner(s) to add.

Too much focus on diversity/margin leads to poor accuracy. Too much focus on validation error leads to overfitting.

Our Method: Decoupled Ensemble Pruning (DEP)

Framework



Analysis of validation-error-based pruning (stage 2)

Training set
$Tr = \{(\mathbf{x}_i, y_i)\}_{i=1}^{m_1}$

Generate base learners $\{h_t\}_{t=1}^{n}$

Validation set
$V = \{(\mathbf{x}_i, y_i)\}_{i=1}^{m_2}$

Select one combination out of all combinations based on validation error

Test set

Report generalization performance

$$\underset{\substack{V\sim\mathcal{D}^{m_2},\\ T\sim P(T|V)}}{\mathbb{E}} \mu_T \le \underset{\substack{V\sim\mathcal{D}^{m_2},\\ T\sim P(T|V)}}{\mathbb{E}} \phi_T + \sqrt{\frac{-\underset{\substack{V\sim\mathcal{D}^{m_2},\\ T\sim P(T|V)}}{\mathbb{E}} \log_2 \boxed{\mathrm{P}(E=\phi_T \mid V)}}{2m_2}}$$

Generalization error

Validation error of the selected combination

Selection bias

Combination distribution (the distribution of the validation error of all the possible combinations of base learners)

Design of distribution optimization (stage 1)

Requirements

Corollary: A combination distribution that is *heavier on the low error region* leads to better generalization performance.

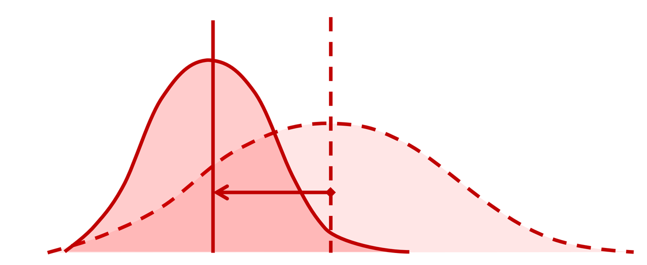
Key challenges

◊ Optimizing the mean of combination distribution usually results in a narrower spread of the distribution, because good learners are more similar. So we need to *maximize the variance* while *minimizing the mean*.

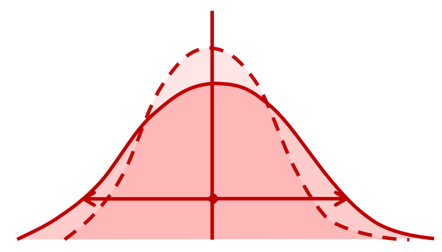
◊ In order to change the combination distribution, we should not leak the information of the validation set in the distribution optimization step, so we can only *optimize on the training set*. Predictions can often be perfect on the training set, so we need to incorporate *more information*.

Bi-objective formulation

◊ Maximizing *margin mean* for full ensemble is minimizing the average validation error of all combinations.



◊ Maximizing *structural diversity* is to maintain the spread of the combination distribution.



Optimization

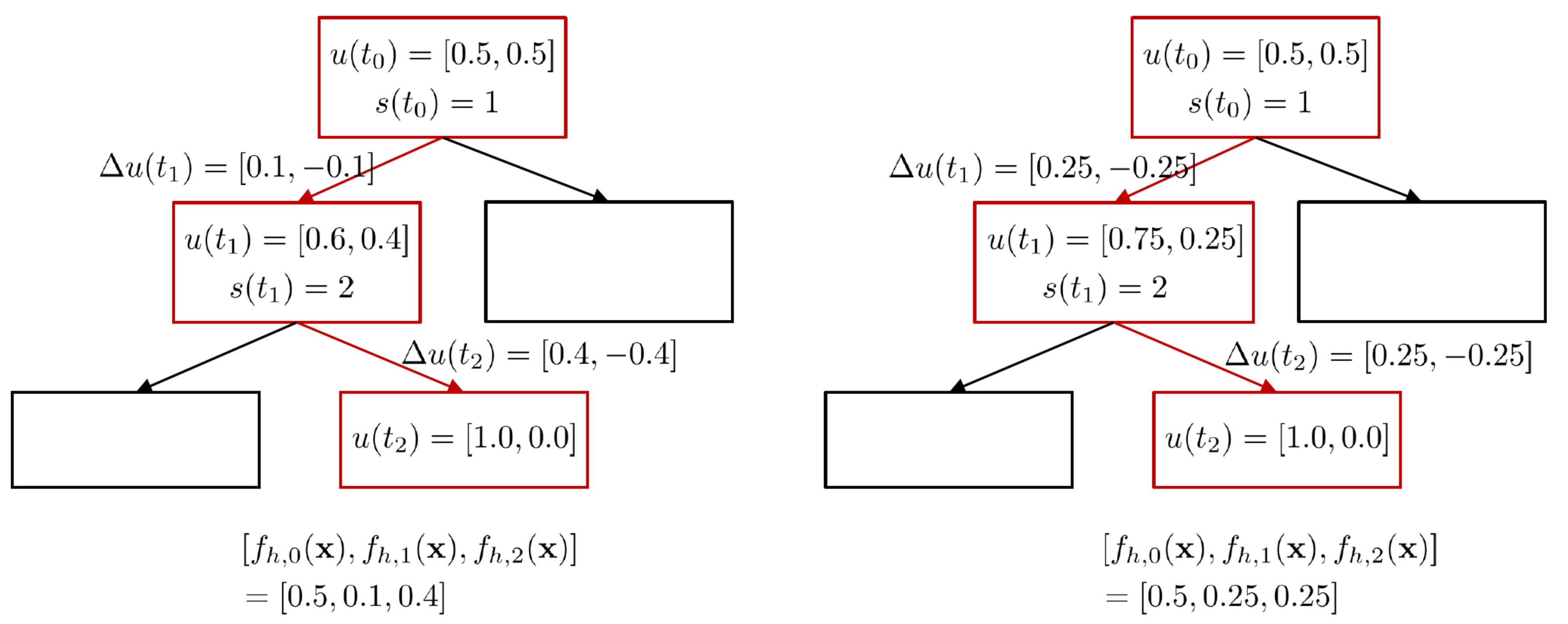
◊ Stage 1: Use evolutionary pareto optimization algorithm to solve the bi-objective optimization problem.

◊ Stage 2: Use single-objective evolutionary algorithm to solve.

A Novel Structural Diversity for Decision Tree Ensemble

Feature contribution diversity

Defined to be the variation of *feature contribution vectors*.



An example of the feature contribution vectors for the same instance in two decision trees

Compare to other diversity measures

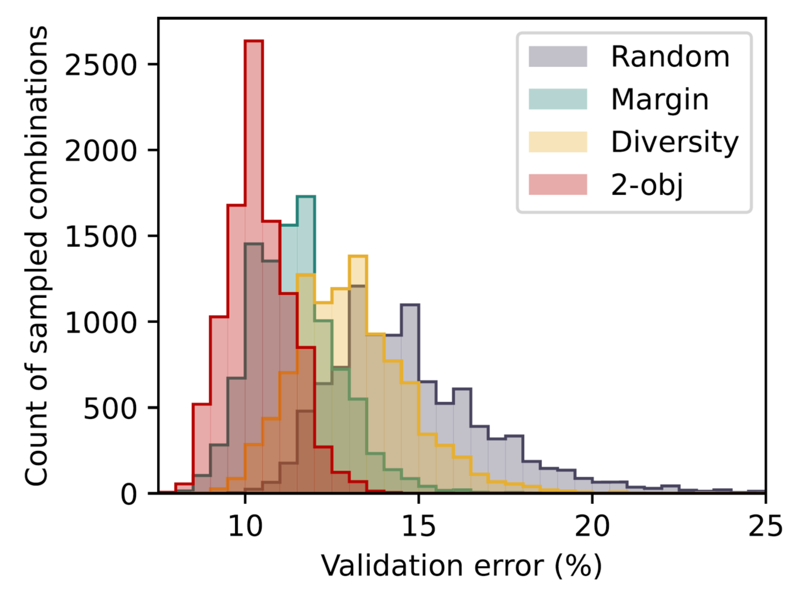
We can tell the difference between two trees even when other methods fail.

		Interpolation regime		Non-interpolation regime	
		Different tree structure	Same tree structure	Same splitting features Different splitting points	Different tree structures
Behavior diversity	Kappa (Margineantu and Dietterich, 1997; Martínez-Muñoz et al, 2008)	×	√	√	√
	Disagreement (Li et al, 2012)	×	√	√	√
	Complementarity (Martínez-Muñoz and Suárez, 2004)	×	√	√	√
Structural diversity	Tree matching distance (Sun and Zhou, 2018)	√	×	×	√
	Feature contribution (ours)	√	√	√	√

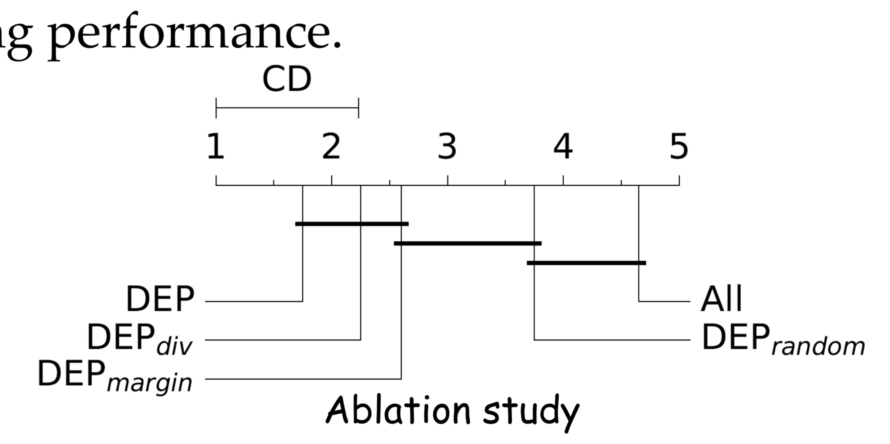
Experiments

Effectiveness of optimizing combination distribution

◊ Optimizing margin mean and structural diversity brings about better combination distribution.



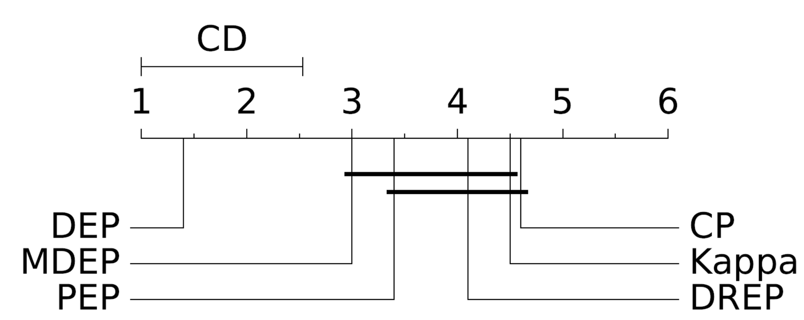
◊ better combination distribution leads to better ensemble pruning performance.



Ablation study

DEP vs. state-of-the-art pruning methods

◊ DEP has a *significant* advantage in generalization performance over other methods.



DEP compared to SOTA methods