Margin Distribution and Structural Diversity
Guided Ensemble Pruning

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

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.

Our Method: Decoupled Ensemble Pruning (DEP)

Framework

Stage 1: Distribution optimization
- n learners
- Margin optimization
- Structural diversity

Stage 2: Validation-error-based pruning
- n/2 learners
- Validation error
- Pruned ensemble

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

Training set \( T_r = \{ (x_i, y_i) \}_{i=1}^m \)
Generate base learners \( \{ h_i \}_{i=1}^n \)
Validation set \( V = \{ (x_i, y_i) \}_{i=1}^m \)
Select one combination out of all combinations based on validation error

Optimization

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.

A Novel Structural Diversity for Decision Tree Ensemble

Feature contribution diversity

Defined to be the variation of feature contribution vectors.

\( \Delta \sigma(x) = [0.5, 0.5] \)
\( \sigma(x) = 1 \)

\( \Delta \sigma(x) = [0.25, 0.25] \)
\( \sigma(x) = 1 \)

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.

Effectiveness of optimizing 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.

Experiments