An improved random forest based on the classification accuracy and correlation measurement of decision trees

DISCLAIMER: Summarized by AI

Problem they are trying to solve / Purpose of method

The authors aim to address two key issues in traditional random forests:

- **Low classification accuracy** of some decision trees (CARTs) due to randomness in training data and features.
- **High correlation (low diversity)** between CARTs, leading to decision redundancy and reduced generalization.

They propose a method to:

- Retain CARTs with high classification accuracy.
- Reduce correlation among CARTs using a quantifiable similarity measure.

Why is the method introduced/needed?

- Existing methods either improve decision tree performance or their diversity, but **seldom both simultaneously**.
- Current evaluation methods (like OOB accuracy) are unstable.
- Most approaches don't **quantify correlation** between trees, making diversity control indirect and less effective.

How does it differ from other methods?

- **Dual focus**: Simultaneously considers both classification accuracy and correlation.
- **Improved evaluation**: Uses three reserved test sets for robust classification accuracy estimation, instead of relying on OOB.
- **Quantified correlation**: Introduces a modified **dot product method** to measure cosine similarity between CARTs based on their feature subsets.
- **Selective pruning**: CARTs with high correlation and low accuracy are selectively removed before ensemble construction.

How the method works

Overview:

1. Generate more CARTs than needed using Bagging.
2. Evaluate each CART's classification accuracy using **three independent validation sets**.
3. Measure pairwise **correlation** between CARTs using the **improved dot product method**.

4. Apply a **grid search** to find the optimal correlation threshold.
5. Prune CARTs with high correlation and lower accuracy until the desired number of trees is reached.
6. Build the final random forest from the selected CARTs.

In Detail:

- **Classification Accuracy**: CARTs are tested on three different data subsets; their average accuracy is used for ranking.
- **Correlation Measurement**: The dot product between feature vectors of two CARTs estimates their angle (correlation). Smaller angles = higher similarity.
- **Pruning**: CART pairs with correlation above the threshold are pruned by removing the less accurate one.
- **Final Ensemble**: Top-N CARTs (high accuracy, low correlation) are combined using majority voting.