Graph-based sampling of diverse sub-forests

Project Supervisor: Dr. Christel Sirocchi

Random forests are widely used machine learning models due to their ability to capture complex relationships, handle high-dimensional and noisy data, and accommodate missing values. They achieve strong predictive performance with minimal preprocessing and can even outperform deep learning models in certain settings. While primarily applied to supervised tasks, random forests have also been adapted for unsupervised applications such as clustering and anomaly detection, making them a versatile tool across applications.

Their predictive strength arises from combining many decision trees. However, research has shown that a carefully selected subset of a few diverse trees can retain most of the predictive performance of the full forest while reducing computational costs. Existing methods for sampling such sub-forests typically rely on vector representations of the tree structure or predictions and have mainly been studied in supervised settings.

Building on recent work that introduces a versatile method for deriving graph-based representations of decision trees, this project will investigate whether such representations can be used to select a diverse subset of trees from a forest. A novel graph-based selection method will be developed, compared with existing vector-based approaches, and evaluated in both supervised and unsupervised learning tasks.

Methodological Steps. The project will proceed through the following steps:

1. Generate graph-based representations of trees from trained random forests.
2. Compute pairwise graph similarities and select n trees that are maximally diverse.
3. Evaluate the predictive performance retained by the reduced forest.
4. Compare results to tree selection based on vector representations.
5. Explore combinations of graph- and vector-based representation.
6. Test the approach across various tasks and learning strategies (classification, regression, clustering, and anomaly detection).

Requirements

To complete this project, the following prerequisites are recommended:

- Knowledge of R or Python.
- Familiarity with UNIX environments and command line usage.
- Willingness to independently dive into the ML literature and test new approaches.

Depending on the results, continuation of the project in the context of a MSc thesis is possible.

References

Bayir, M. A., Shamsi, K., Kaynak, H., and Akcora, C. G. (2022). Topological forest. *IEEE Access*, 10:131711–131721.

Sirocchi, C., Urschler, M., and Pfeifer, B. (2025). Feature graphs for interpretable unsupervised tree ensembles: centrality, interaction, and application in disease subtyping. *BioData Mining*, 18(1):15.

Zhang, H. and Wang, M. (2009). Search for the smallest random forest. *Statistics and its Interface*, 2(3):381.