Soft classification trees: a decomposition training algorithm with ILP-based sparsity and pruning

E. Amaldi[1], A. Consolo[2], and F. Gandini[1]

[1]DEIB, Politecnico di Milano, Milano, Italy , ✉ edoardo.amaldi@polimi.it ✉ filippo.gandini@mail.polimi.it
[2]DISCo, Universitá degli Studi di Milano-Bicocca , ✉ antonio.consolo@unimib.it

Decision trees are popular supervised Machine Learning (ML) methods for classification and regression tasks. They are widely used in a variety of application fields ranging from Business Analytics to Medicine. The success of decision trees lies mainly in their interpretability and good accuracy. Unlike most black-box ML models, they reveal the decisions leading to the tree response for any input vector. Interpretability is of particular importance in applications where the ML models support domain experts decisions and justifiable predictions are required, such as in medical diagnosis and criminal justice.

A decision tree is a directed binary tree with a set of internal nodes, referred to as branch nodes, including the root, a set of leaf nodes, and two outgoing arcs (branches) for each branch node associated to a splitting rule applied to the input space (feature space). Once values are assigned to the tree parameters (those associated to the branch nodes and those associated to the leaf nodes), any input vector is routed from the root along the tree according to the splitting rules at the branch nodes, eventually falling into a leaf node where the output (the linear prediction or the class label) is determined. Hard or soft splitting rules can be considered at branch nodes. In hard (deterministic) splits, the left branch is followed if a single feature (univariate case) or a linear combination of the features (multivariate case) exceeds a given threshold value. In soft splits, both left and right branches are followed with complementary probabilities given by applying a continuous sigmoid function to a linear combination of the features.

Starting from the seminal CART method for building univariate classification and regression trees [7], early methods were based on greedy approaches where branch nodes are optimized one at a time and often followed by a pruning phase to reduce overfitting. Since training decision trees is known to be NP-hard, it is a challenging problem to globally optimize them over all the tree parameters.

During the last decade, growing attention has been devoted to globally optimized (trained) decision trees with deterministic or soft splitting rules at branch nodes, leveraging on the remarkable advances in mixed integer linear programming (MILP) and nonlinear continuous programming (NLP). For MILP formulations and local search methods to train deterministic classifications trees, see for instance [1,3,4] and the references therein. For NLP formulations and a first decomposition method to train soft classification trees, see [5,6] and [2] respectively.

In this work, we propose a new variant of soft multivariate classification trees (SCTs) with single leaf node predictions. For every input vector, the predicted class label is the one assigned to the single leaf node obtained by routing the input vector from the root, following at each branch node the branch with higher probability. Such SCTs satisfy the conditional computation property and are suitable for decomposition. We devise a node-based decomposition algorithm including an initialization procedure and an ad hoc heuristic for rerouting the data points along the tree. To obtain sparse and compact SCTs, we associate with the resulting soft classification tree an equivalent deterministic classification tree and exploit the power of ILP for pruning and inducing sparsity. We aim for the best of the two approaches: improved scalability via NLP decomposition and improved sparsity and pruning via ILP formulations.

The performance of the decomposition training algorithm and the impact of ILP-based sparsity and pruning is assessed on 20 datasets from the literature in terms of classification accuracy and tree size. The results are compared with those obtained with the extensively used Random Forest ensemble method.

References

[1] S. Aghaei, A. Gómez, and P. Vayanos. Strong Optimal Classification Trees. *Operations Research*, 73(4):2223-2241, 2025.

[2] E. Amaldi, A. Consolo, and A. Manno. On multivariate randomized classification trees: l_0-based sparsity, VC-dimension and decomposition methods. *Computers & Operations Research*, 151:106058, 2023.

[3] J. Dunn. Optimal trees for prediction and prescription. Ph.D. thesis, Massachusetts Institute of Technology, 2018.

[4] D. Bertsimas and J. Dunn. Machine learning under a modern optimization lens. Dynamic Ideas LLC, 2019.

[5] R. Blanquero, E. Carrizosa, C. Molero-Rio, and D.R. Morales. Sparsity in optimal randomized classification trees. *European Journal of Operational Research*, 284(1):255-272, 2020.

[6] R. Blanquero, E. Carrizosa, C. Molero-Rio, and D.R. Morales. Optimal randomized classification trees. *Computers & Operations Research*, 132:105281, 2021.

[7] L. Breiman L, J. Friedman, R.A. Olshen, and C.J. Stone. Classification and regression trees. CRC press, 1984.