Tree Ensemble Compression for Interpretability

Laurens Devos, Deniz Can Oruç, Wannes Meert,
Hendrik Blockeel, and Jesse Davis

KU Leuven

Interpretability and Tree Ensembles

A function is interpretable if it is **human-simulatable** [1]. Eg. sparse-linear models, small decision trees, nearest neighbors etc.

Tree ensemble methods combines potentially **overlapping rules** [2, 3].

This increases the predictive performance, but it causes a **trade-off with interpretability** [4].

Compressing ensambles **reduces the number of nodes**, hence the used rules for evaluation, and increases the interpretability of the model

Idea **Compress** large models to obtain smaller models that are **more interpretable**

- Remove full trees and subtrees from the model
- Refit the leaf values
- Use logistic regression and L1 regularization to fit:

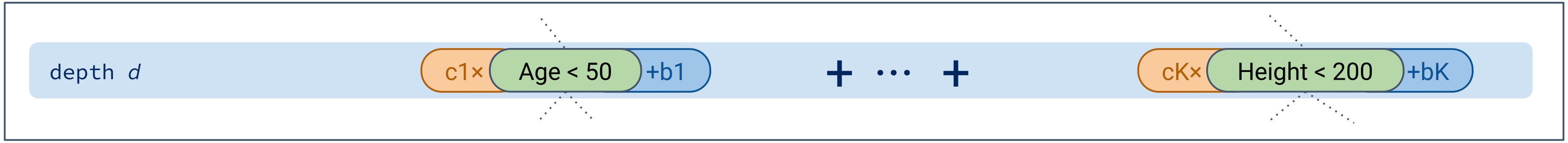
coefficient × subtree + bias

- Lossy compression → non-equivalence preserving
- Cut out unnecessary parts of the model
- Allow no more than X% performance loss on validation set (e.g., 5%)

$c \times f_n(x) + b$

How to fit ... ***d* ?**

Logistic regression + L1 regularization



$$\operatorname{expit}\left(\begin{bmatrix} \boxed{\begin{matrix} f_{n_1}(x_1) \\ f_{n_1}(x_2) \\ \vdots \\ f_{n_1}(x_N) \end{matrix}} & \boxed{\begin{matrix} \mathbb{1}_{n_1}(x_1) \\ \mathbb{1}_{n_1}(x_2) \\ \vdots \\ \mathbb{1}_{n_1}(x_N) \end{matrix}} & \begin{matrix} \cdots \\ \cdots \\ \\ \cdots \end{matrix} & \boxed{\begin{matrix} f_{n_K}(x_1) \\ f_{n_K}(x_2) \\ \vdots \\ f_{n_K}(x_N) \end{matrix}} & \boxed{\begin{matrix} \mathbb{1}_{n_K}(x_1) \\ \mathbb{1}_{n_K}(x_2) \\ \vdots \\ \mathbb{1}_{n_K}(x_N) \end{matrix}} \end{bmatrix} \cdot \begin{bmatrix} \boxed{c_1} \\ \boxed{b_1} \\ \vdots \\ \boxed{c_K} \\ \boxed{b_K} \end{bmatrix} + b_0\right)$$

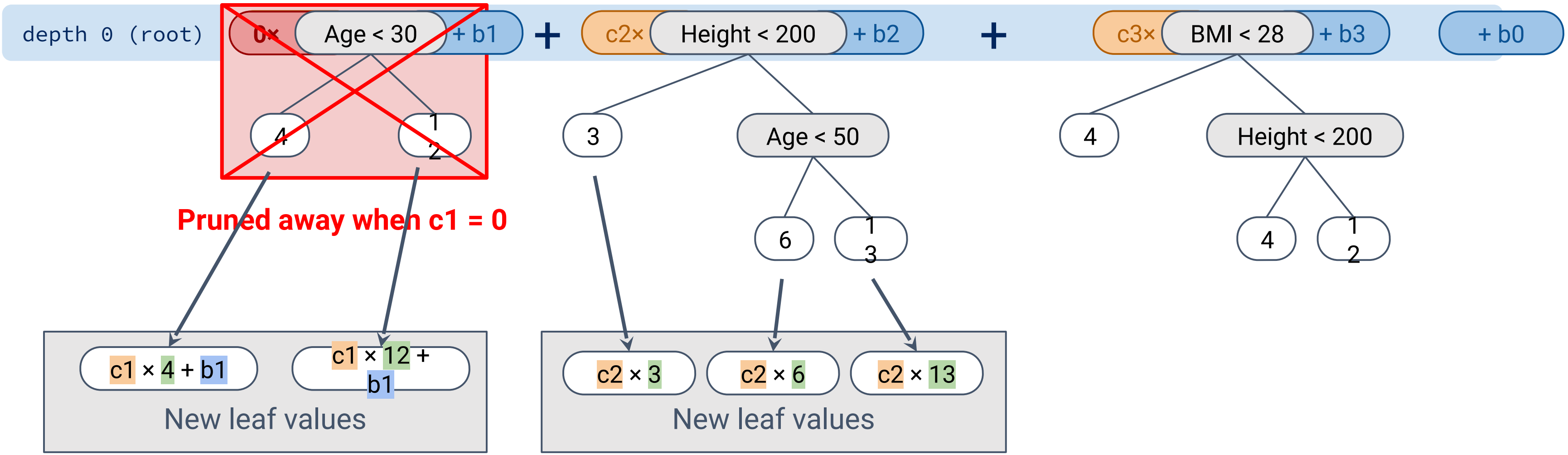
leaf value reached by $\mathbf{x}_i$ in subtree if $\mathbf{1|_{nk}(x_i)=1}$,else 0

1 if $\mathbf{x}_i$ passes through $\mathbf{n}_k$, else 0

$$c_1 f_{n_1}(x_i) + b_1 \mathbb{1}_{n_1}(x_i) + \cdots + c_K f_{n_K}(x_i) + b_K \mathbb{1}_{n_K}(x_i)$$

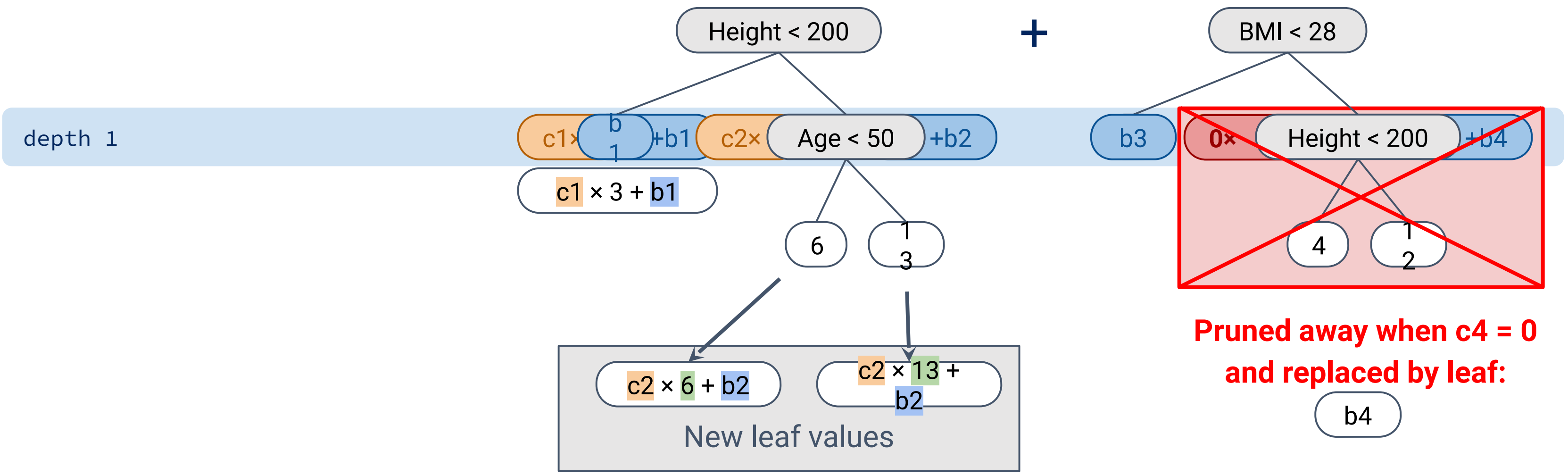
Tree ensemble compression: Top-down tree pruning

Layer-per-layer, fit coefficient + bias: $c \times f_n(x) + b$ using **L1**



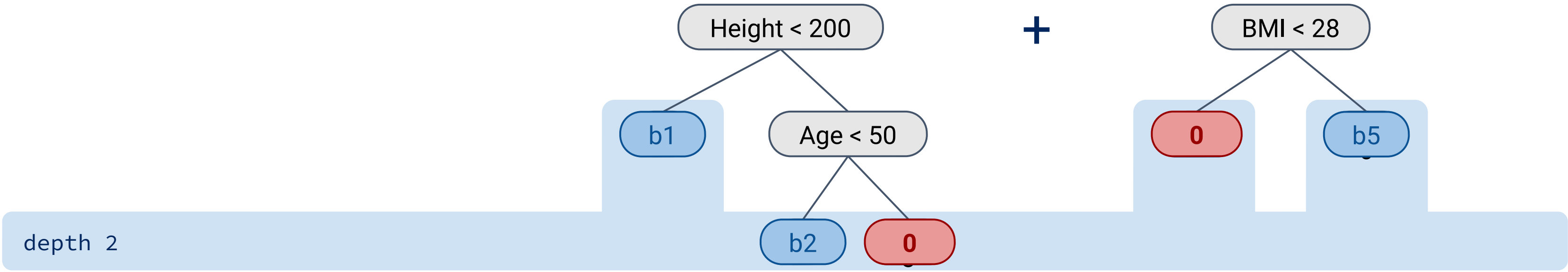
Tree ensemble compression: Top-down tree pruning

Layer-per-layer, fit coefficient + bias: $c \times f_n(x) + b$ using **L1**



Tree ensemble compression: Top-down tree pruning

Layer-per-layer, fit coefficient + bias: $\boxed{c \times f_n(x) + b}$ using **L1**



We count:
5: number of remaining leaves
3: number of non-zero leaves (nnz-leaves)

Experimental Questions

- What is the performance in terms of compression and the effect of compression on predictive performance?
- What is the trade-off between model size and predictive performance?
- What is the computational complexity to produce the smaller ensembles?

Experimental Setting

- Experiments are with **XGBoost**
- Train models on **15 binary classification datasets** using a different selection of **160 hyper-param settings** in the grid.
- Select a set of up to **20 good parameter settings** from the subset of hyper-parameter settings that is **Pareto optimal** in at least one fold among **5-folds.**

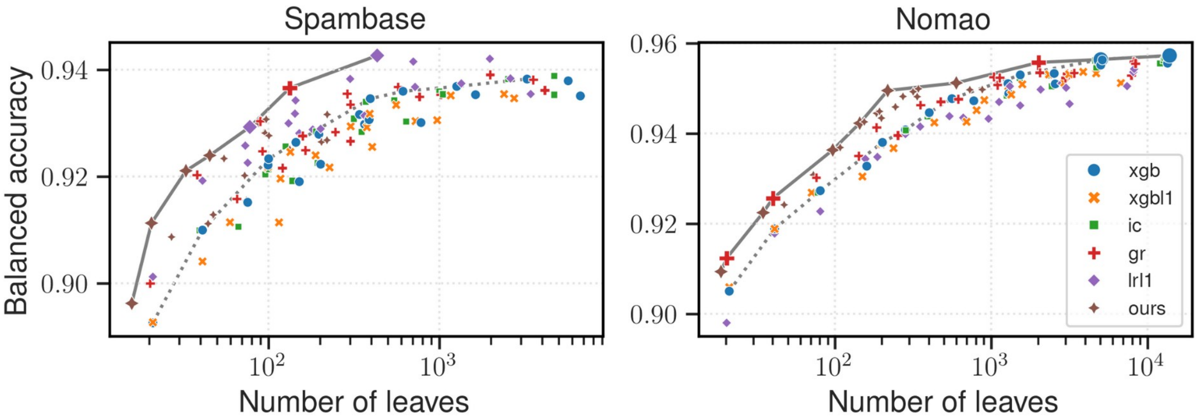
Baselines

- **xgblrl1:** Retrained XGBoost model with L1 regularization.
- **gr:** Global refinement [5] combines leaf refinement with L2 regularization and a simple pruning strategy. Operates at the *finest extreme*: it only considers the leaf level.
- **ic:** Individual contribution [6] is a standard technique for pruning trees from tree ensembles. It represents the other *coarsest extreme*: it only operates at the tree level.
- **lrl1:** Combines leaf refinement and ensemble pruning with L1 regularization [7]. *Combines the two extremes* (coarse tree level + finest leaf level), but does not work at the subtree level as our method does.

Q1: Compression quality: compression ratio and difference in predictive performance.

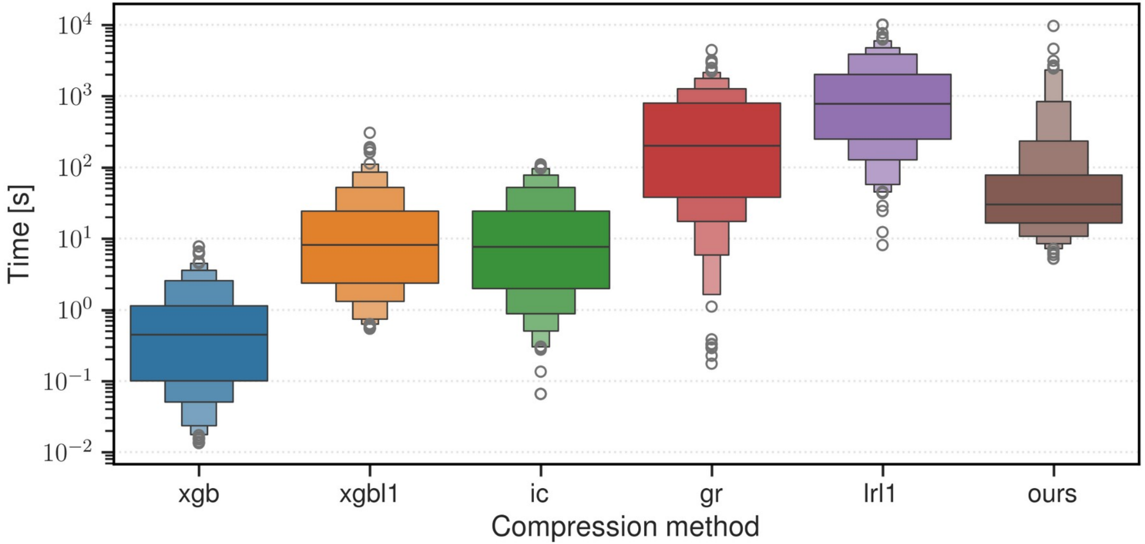
	#Leaf	Compress. ratio (×)					Bal.Acc.	Bal.Acc. diff. (%)				
	xgb	xgbl1	ic	gr	lrl1	**ours**	xgb	xgbl1	ic	gr	lrl1	**ours**
Compas	289	2.6	1.2	1.4	2.2	5.1	67.2	-0.1	-0.1	-0.3	-0.5	-0.4
Vehicle	453	1.3	1.5	1.2	1.2	3.0	93.7	-0.8	-0.6	-0.2	0.5	-1.5
Spambase	1120	1.5	1.1	1.3	1.4	7.9	92.7	-0.3	-0.1	0.2	0.4	-0.5
Phoneme	2138	1.2	1.2	1.2	1.1	4.9	83.2	-0.0	-0.1	0.5	0.1	-0.3
Nomao	3387	1.6	1.0	1.3	1.3	8.5	94.5	-0.3	-0.0	0.1	-0.4	-0.4
Adult	1955	3.0	1.2	1.5	1.3	3.3	76.3	-0.0	0.1	1.2	-0.9	-0.1
Ijcnn1	4162	1.3	1.0	1.2	1.1	3.1	91.5	-0.2	-0.0	1.1	-0.0	-0.0
Mnist	321	1.4	1.1	1.3	1.6	4.4	97.9	-0.3	0.0	-0.1	0.1	-0.5
DryBean	1866	2.0	1.2	1.6	1.2	43.2	90.9	-0.3	-0.0	0.7	-0.1	0.1
Volkert	1973	2.4	1.1	1.4	2.0	18.0	98.3	-0.3	-0.0	-0.1	-0.3	-0.5
Credit	546	1.5	1.1	1.2	1.5	3.2	77.2	-0.3	-0.1	-0.2	-0.5	-0.4
California	2664	1.2	1.0	1.3	1.3	5.9	88.8	-0.0	-0.0	0.0	-0.2	-0.4
MiniBooNE	3172	1.5	1.0	1.4	1.3	5.1	92.1	-0.1	-0.0	0.2	-0.3	-0.4
Electricity	3970	1.2	1.0	1.1	1.1	2.9	86.1	-0.1	-0.0	0.1	-0.2	-0.3
Jannis	3157	1.4	1.1	1.2	1.2	2.7	77.1	-0.3	-0.0	-0.1	-0.9	-0.4
average	2078	1.7	1.1	1.3	1.4	8.1	87.2	-0.2	-0.1	0.2	-0.2	-0.4

Q2: The model-size and predictive performance trade-off



	Pareto (Q2)	
	#on front	#wins
xgb	12	0
xgbl1	6	0
ic	9	0
gr	29	2
lrl1	12	2
ours	69	11

Q3: Computational cost: how long does it take to compress an ensemble



	Run time [s] (Q3)	
	Mean	Median
xgb	0.986	0.451
xgbl1	23.1	8.14
ic	19.9	7.69
gr	522	202
lrl1	1580	783
ours	233	30.1

Conclusion

- We proposed a novel technique for compression that is much more effective at compressing models than existing approaches.

- Moreover, each compressed model performs similarly to its uncompressed counterpart.

- Compression techniques are helpful when exploring a model size vs. performance trade-off.

- Often the final epsilon improvement comes at the cost of substantially larger models. This has important implications for interpretability.

References

[1] Lipton, Z.: The mythos of model interpretability. Communications of the ACM 61 (10 2016)

[2] Bühlmann, P., Yu, B.: Analyzing bagging. Annals of Statistics 30 (2002)

[3] Bühlmann, P.: Bagging, boosting and ensemble methods. Handbook of Computational Statistics (2012)

[4] Nalenz, M., Augustin, T.: Compressed rule ensemble learning. In: Proceedings of The 25th International Conference on Artificial Intelligence and Statistics. Proceedings of Machine Learning Research, vol. 151, pp. 9998–10014 (2022)

[5] Ren, S., Cao, X., Wei, Y., Sun, J.: Global refinement of random forest. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 723–730 (2015)

[6] Lu, Z., Wu, X., Zhu, X., Bongard, J.: Ensemble pruning via individual contribution ordering. In: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. p. 871–880. KDD '10 (2010)

[7] Buschjäger, S., Morik, K.: Joint leaf-refinement and ensemble pruning through l1 regularization. Data Min. Knowl. Discov. 37(3), 1230–1261 (2023)