Mondrian Forests with Label Guided Splits

Ismaël Koné[1] Lahsen Boulmane
2MIA Research Group, LEM2A Lab
Faculté des Sciences, Université Moulay Ismail
Meknès, Morocco
[1]Email: ismael.from.kone@gmail.com

Objective

- Studying the effect of incorporating a label guided split scheme in the recent novel class of Random Forests, so-called *Mondrian Forests* which have attractive properties.

Context

- Mondrian Forest (MF): A novel a class of Random Forest (RF) introduced by Balaji [1].
- Mondrian Tree (MT): random hierarchical binary partition of $\mathbb{R}^D$(distribution over kd-tree data structures).

1. Each tree is bayesian →high uncertainty to data far from training data (bounding box) where splits only occur. There is no split or decision boundary outside thus the model is uncertain (see fig.1 and 2.).
2. Splits are done independently of labels.
3. Online version (no re-training!) matches the batch version.

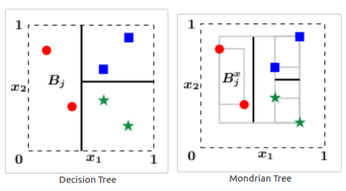

Figure 1: Difference between Decision Tree (DT) and Mondrian Tree (MT). Dots, square and stars are 2D training data and their colors represent their classes. The MT splits only within the bounding box of the training data points. Thus data outside these bounding boxes receive high uncertainty as no split occurs there. Conversely DT is confident even far from the training data as it splits the entire feature space $[0, 1]^2$.

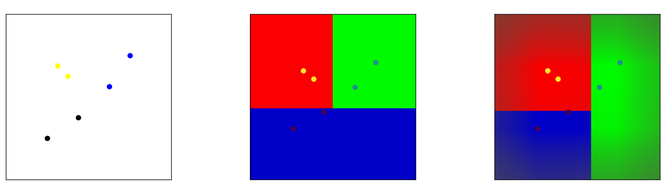

Figure 2: On the left, are represented training data with the color indicating their classes. Middle and right figures present respectively DT and MT predictions of the whole space $[0, 1]^2$ after training where each data point in any region shares the same class as the training data lying in that region. The color brightness of any region indicates the prediction confidence level, the brighter the more confident and inversely. The MT(right figure) is less confident as we move far from the training data while the DT (middle figure) is confident everywhere.

Problem

- Mondrian Forests still perform less than the standard Random Forests[1]
- Performs worsely when data contains many irrelevant features (No label used for splits).
- What can we do to make MF better ?

We propose simply to exploit labels to guide splits.

Method

Random Node Optimization using the Information Gain (IG)[2]

$$IG(\delta, \xi, \mathcal{D}) = H(\mathcal{D}) - \frac{1}{|\mathcal{D}|} \sum_{i \in \{R,L\}} |\mathcal{D}_i| H(\mathcal{D}_i) \tag{1}$$

with $\mathcal{D}_R = \{x \in \mathcal{D} | x_\delta \geq \xi\}$ and $\mathcal{D}_L = \{x \in \mathcal{D} | x_\delta < \xi\}$

$$H(\mathcal{D}) = -\sum_{c \in \mathcal{C}} p(c) log(c) \tag{2}$$

Selection of the best one among $T \times Q$ splits (δ, ξ) at each node

$$\delta^*, \xi^* = \underset{\delta, \xi \in \{\delta_1, ..., \delta_T\} \times \{\xi_1, ..., \xi_Q\}}{\operatorname{argmax}} IG(\mathcal{D}, \delta, \xi) \tag{3}$$

The dimension or feature δ_i is sampled with probability proportional to $(X_d)_{max} - (X_d)_{min}$ for any $d = 1, ..., D$. The split position ξ_j is sampled uniformly from the extent of feature δ_i: $[(X_{\delta_i})_{min}, (X_{\delta_i})_{max}]$.

- **Remark**: when $T = Q = 1$, we fall back to the original Mondrian Forest.

Experiments

We conducted experiments in with the following settings :

1. **Datasets**: usps[3], satimages[4], letter[4] and dna[4] datasets.
2. **split parameters**: $T = \sqrt{D}$ (D is the dataset dimension) as usual in RF and $Q = 6$.
3. **Number of trees**: 1 and 50.
4. **Mode**: batch and online with 10 mini-batches.

Results

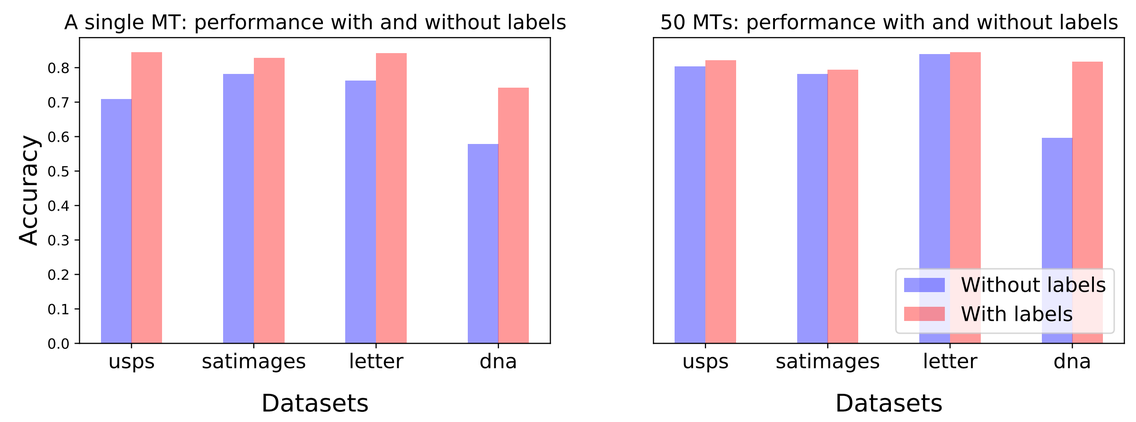

Figure 3: Comparison between the original MF and our modified version in **batch** mode on four datasets.

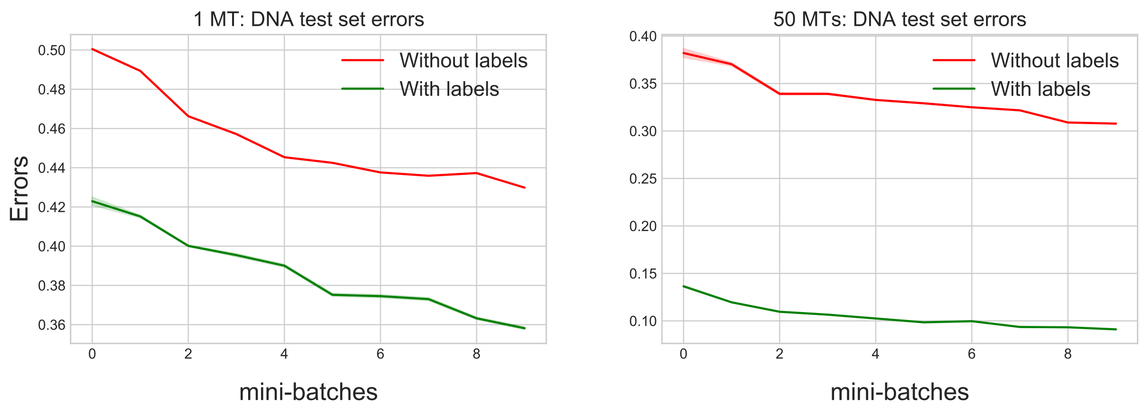

Figure 4: Comparison between the original MF and our modified version in **online** mode on DNA dataset (containing irrelevant features).

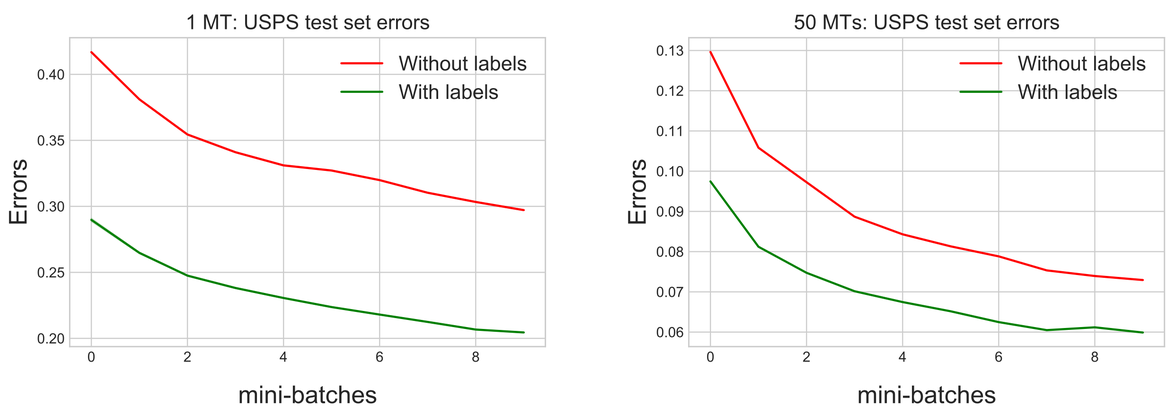

Figure 5: Comparison between the original MF and our modified version in **online** mode on USPS dataset.

Analysis

In batch mode:

- 1 MT: accuracy increases by +10.6%.
- 50 MTs: accuracy increases by 25.3% for the dna dataset (containing irrelevant features) and similar otherwise (1.3%)

In online mode:

- DNA dataset: same increase as in batch mode 25%, the gap gets bigger as more trees.
- USPS: accuracy increases slightly as in batch mode, 1.2%.

- **Remark**: In the online mode, using labels for splits breaks the guarantee that the batch mode matches the online one.
- Our approach increases the running time as we compute the Information Gain $T \times Q$ times.

Conclusion and Perspectives

- The performance increases considerably when the dataset contains irrelevant features.
- It is less noticeable otherwise.
- We project to perform a theoretical analysis of the induced time complexity and bias-variance analysis in both cases with and without label guided splits.

References

[1] D. M. Roy B. Lakshminarayanan and Y. W. Teh. Mondrian forests: Efficient online random forests. NIPS, 2014.

[2] C. Antonio S. Jamie., and K. Ender. Decision forests: A unified framework for classification, regression, density estimation, mani- fold learning and semi-supervised learning. Founda- tions and Trends in Computer Graphics and Vision, 7(2-3):81–227, 2012.

[3] J. J. Hull. A database for handwritten text recognition research. IEEE Transactions on Pattern Analysis and Machine Intelligence, 16(5):550-554, May 1994.

[4] Feng,C., Sutherland,A., King,S., Muggleton,S. Henery,R.(1993). Comparison of Machine Learning Classifiers to Statistics and Neural Networks. AI Stats Conf. 93.