Robust optimal classification trees under noisy labels

Efstratios Reppas

Objective

- Understand the theoretical background of this paper
- Reproduce its experimental results

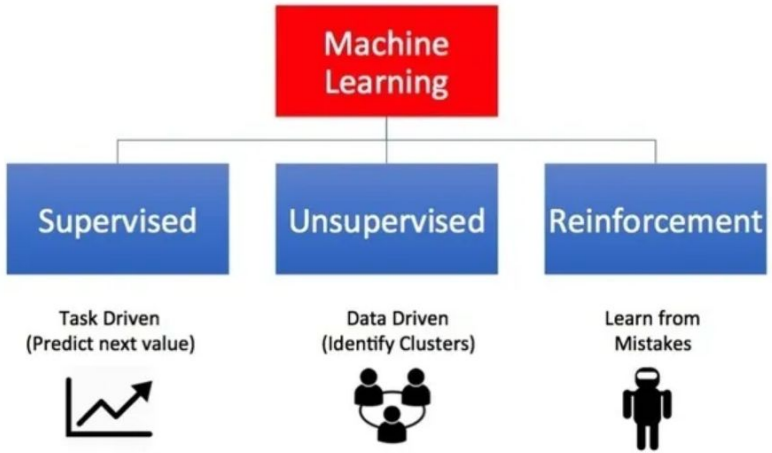
Robust optimal classification trees under noisy labels

- Victor Blanco, Alberto Japon and Justo Puerto - 2020
- Binary classification problem
- Combination of Optimal Decision Trees with SVM
- Focus on robustness by considering noisy data

Theoretical Background

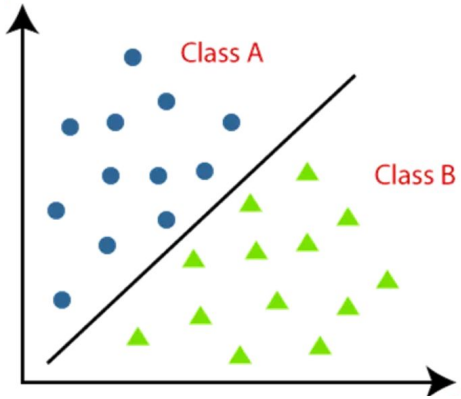
Supervised Learning

- One of the three main paradigms of machine learning
- Learn a function $h : X \rightarrow Y$ that maps inputs X (**features**) to outputs Y (**labels**).
- Labels are known for a subset of X



Binary Classification

- Subset of Supervised Learning
- The labels take binary values of $y \in \{0, 1\}$ or $\boldsymbol{y} \in \{-1, +1\}$



Mathematical Programming

- Involves problems that can expressed as minimization/maximization of a function, called the **objective function**, given set of constraints.

- General formulation:

$$\min_x f(x) \quad \text{s.t.} \quad g_i(x) \leq 0, \forall i, \quad h_j(x) = 0, \forall j$$

where f: objective function, g_i: inequality constraints, h_j: equality constraints

- Such problems can be often solved by several algorithms, off-the-shelf in many programming languages.

Categorization of Problems

- **Linear Programming**: All functions linear, variables continuous → $\in$ P

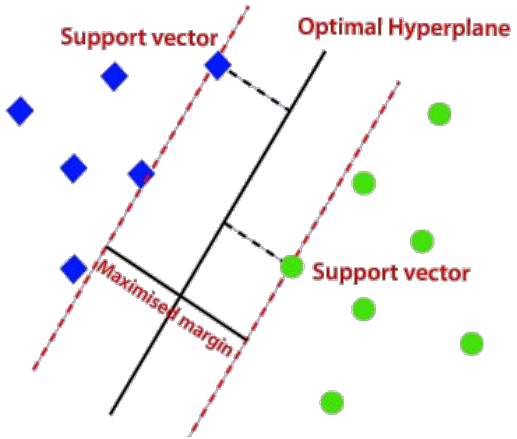
- **Mixed-Integer Programming**: Variables can be integers → $\in$ NP

- **Non-Linear Programming**: Functions can be non-linear → harder problem, when there are non-convexity issues

Support Vector Machines (SVMs)

- A very popular (binary) classification model
- Calculates a hyperplane that maximizes the margin between itself and the two classes
- Calculated as a **Quadratic Optimization Problem**:

$$\begin{aligned} \underset{\mathbf{w},\, b}{\text{minimize}} \quad & \frac{1}{2}\|\mathbf{w}\|^2 \\ \text{subject to} \quad & y_i(\mathbf{w}^\top \mathbf{x}_i - b) \geq 1 \quad \forall i \in \{1, \ldots, n\} \end{aligned}$$



RE-SVM

- Robust in noisy datasets - improved performance over standard SVM
- Allows for flipping labels during optimization
- Formulation:

$$\min_{w,w_0,e,\xi} \frac{1}{2}\|w\|^2 + c_1 \sum_{i=1}^{n} e_i + c_2 \sum_{i=1}^{n} \xi_i$$

subject to:

$$(1 - 2\xi_i) y_i (w^\top x_i + w) \geq 1 - e_i, \quad \forall i = 1, \ldots, n,$$

$$e_i \geq 0, \quad \xi_i \in \{0, 1\}, \quad \forall i = 1, \ldots, n.$$

e_i: the misclassification error for the i-th observation

ξ_i: indicates whether observation is relabeled ($\xi i = 1$) or not ($\xi i = 0$).

c_1, c_2 : cost parameters

RE-SVM

- Robust in noisy datasets - improved performance over standard SVM
- Allows for flipping labels during optimization
- Formulation:

$$\min_{w,w_0,e,\xi} \frac{1}{2}\|w\|^2 + c_1 \sum_{i=1}^{n} e_i + c_2 \sum_{i=1}^{n} \xi_i$$

subject to:

$$(1-2\xi_i)y_i(w^\top x_i + w) \geq 1 - e_i, \quad \forall i = 1,\ldots,n,$$

$$e_i \geq 0, \quad \xi_i \in \{0,1\}, \quad \forall i = 1,\ldots,n.$$

e_i: the misclassification error for the i-th observation

ξ_i: indicates whether observation is relabeled ($\xi i = 1$) or not ($\xi i = 0$).

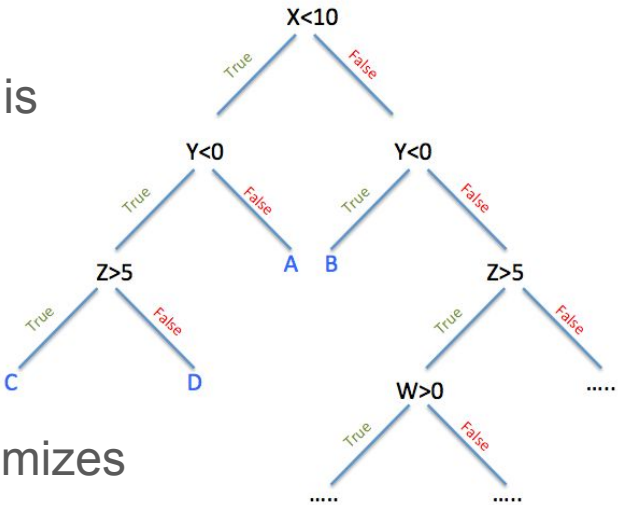
c_1, c_2 : cost parameters

$$e_i = \begin{cases} \max\{0, 1 - y_i(\omega^\top x_i + \omega_0)\}, & \text{if } \xi_i = 0, \\ \max\{0, 1 + y_i(\omega^\top x_i + \omega_0)\}, & \text{if } \xi_i = 1. \end{cases}$$

Decision Trees

- Hierarchical models that partition the feature space recursively to make decisions based on the input data.

- At each level of the tree, a binary decision rule is implemented based on an input feature. The classification is made in the leaf nodes



- CART, one of the baseline algorithms, uses a greedy approach to split the data at each node, by selecting the feature that minimizes the gini index:

$$Gini = 1 - \sum_j p_j^2$$

Optimal Decision Trees (OCTs)

- CART is an efficient algorithm, but produces suboptimal trees
- **OCTs** calculate trees that are optimal in some sense → minimize misclassification errors (L_t) and the number of splits (d_t)
- Their objective function:

$$\min \sum_{t \in \mathcal{L}} L_t + \alpha \sum_{t \in \mathcal{B}} d_t,$$

  where *a* controls the tradeoff between accuracy and tree depth, $\mathcal{L}$ is the set of leaf nodes and $\mathcal{B}$ is the set of intermediate nodes.

- **OCT-H** builds upon this paradigm, allowing for decision rules to be drawn based on a linear combination (hyperplane) of the features (instead of using only one at each node).

OCTSVM

OCTSVM

- The idea comes from the fact that OCTs do not include any notion of optimality regarding class separation.
- The hyperplanes of OCT-H can be drawn **anywhere between the classes**
- The authors fix this by incorporating the SVM principle of maximum margins between the hyperplane and the classes (**OCT + SVM**)
- No need for **leaf-nodes**: classification occurs at each node(*D-1* depth for same splits)
- Also allows **relabelling** for robustness
- The model is formulated as **Mixed Integer Non Linear Program (MINLP)**

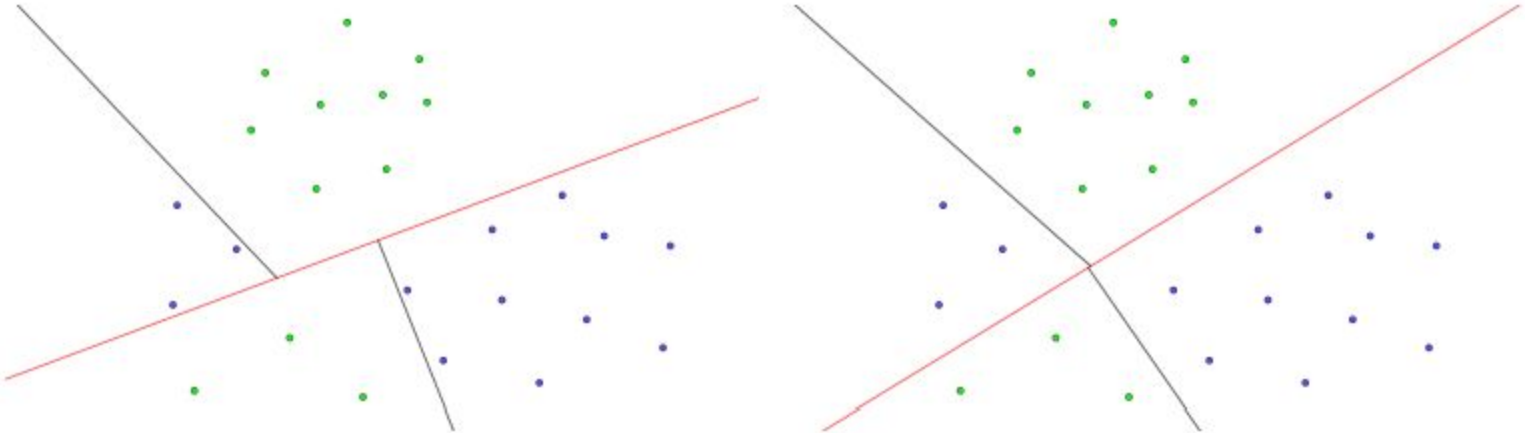


FIGURE 2. Optimal solutions for OCT-H with $D = 2$ (left) and OCTSVM with $D = 1$ (right).

Formulation as MINLP

SVM objective function

$$\frac{1}{2}||\omega_t||_2 \leq \delta \qquad \forall t = 1, \ldots, T.$$

Inverse of margin minimized ⇒ maximization of margin for each node t of the tree

minimized in the objective function

Relabelling

tradeoff parameters

$$c_1 \sum_{i=1}^{n} \sum_{t=1}^{T} e_{it} + c_2 \sum_{i=1}^{n} \sum_{t=1}^{T} \xi_{it}.$$

misclassification error

Binary {0,1} variable - indicates label flip

Same as Re-SVM, but now because we consider each observation i at each node t of the tree, we sum in their entirety.

To be minimized in the objective function.

Definition of e_{it} and ξ_{it}

Binary variable → 0 if observation i not used in node t

$$y_i(\omega_t' x_i + \omega_{t0}) - 2y_i(\beta_t' x_i + \beta_{it0}) \geq 1 - e_{it} - M(1 - z_{it}), \quad \begin{cases} \forall i = 1, \ldots, N, \\ t = 1, \ldots, T, \end{cases} \tag{3}$$

$$\beta_{itj} = \xi_{it}\omega_{tj}, \quad \forall i = 1, \ldots, N, t = 0, \ldots, T, \ j = 0, \ldots, p. \tag{4}$$

$z_{it} = 0$

$f(\omega, \xi, e) \geq -\infty$

Always holds

$z_{it} = 1, \xi_{it} = 0$

$e_{it} \geq 1 - y_i(w_t x_i + w_{t0})$

Same as RE-SVM constraint

$z_{it} = 1, \xi_{it} = 1$

$e_{it} \geq 1 + y_i(w_t x_i + w_{t0})$

Applies RE-SVM in each node of the tree

OCT constraint

- Adding the minimal split objective as in OCTs (the misclassification error has already been implemented as part of RE-SVM), we have the following in the objective function:

$$c_3 \sum_{t=1}^{T} d_t.$$

- d_t is defined as a binary variable that expresses whether a split is applied at node t (0 else).

Definition of d_t

- $\|\omega_t\|_2 \leq M d_t$

  $d_t = 0 \rightarrow \|\omega_t\|_2 \leq 0 \Rightarrow \|\omega_t\|_2 = 0 \ \ \forall t = 1, \ldots, T.$

  $d_t = 1 \rightarrow \|\omega_t\|_2 \leq \infty \ \ \forall t = 1, \ldots, T.$ always holds

Is 1 only if weights are non zero

- $d_t \leq d_{p(t)}$

Parent node of t

If a node doesn't have a split, then all of its children must also have no splits (implements reduction of nodes)

Final Objective Function

- Adding the minimal split objective as in OCTs (the misclassification error has already been implemented as part of RE-SVM), we have the following:

$$\delta + c_1 \sum_{i=1}^{n} \sum_{t=1}^{T} e_{it} + c_2 \sum_{i=1}^{n} \sum_{t=1}^{T} \xi_{it} + c_3 \sum_{t=1}^{T} d_t.$$

Sanity constraints

(recall) Binary variable that indicates that the observation i is used in node t

Set of tree levels

- $\sum_{t \in u} z_{it} = 1, \qquad \forall i = 1, \ldots, N, \ u \in U.$

An observation is used exactly once in all levels

- $z_{it} \leq z_{ip(t)}, \qquad \forall i = 1, \ldots, N, t = 2, \ldots, T.$

If an observation i is not in a node t, then it cannot be in its successors

Sanity constraints

Binary variable that indicates in which half-space the observation i lies regarding the hyperplane at node t

$$\omega_t' x_i + \omega_{t0} \geq -M(1 - \theta_{it}), \quad \forall i = 1, \ldots, N, t = 1, \ldots, T,$$

$$\omega_t' x_i + \omega_{t0} \leq M\theta_{it}, \quad \forall i = 1, \ldots, N, t = 1, \ldots, T.$$

$$w_t x_i + w_{t0} \geq 0 \Rightarrow \theta_{it} = 1$$

Positive half-space

$$w_t x_i + w_{t0} \leq 0 \Rightarrow \theta_{it} = 0$$

Negative half-space

Sanity Constraints

Set of nodes from left splits (even indexed)

Set of nodes from right splits (odd indexed)

$$z_{ip(t)} - z_{it} \leq \theta_{ip(t)} \quad \forall i = 1, \ldots, N, t \in \tau_{bl}$$

$$z_{ip(t)} - z_{it} \leq 1 - \theta_{ip(t)} \quad \forall i = 1, \ldots, N, t \in \tau_{br}$$

respectively

$$\theta_{ip(t)} = 0, z_{ip(t)} = 1 \Rightarrow z_{it} = 1$$

Observation i is inherited to the left child if it lies on the negative half-space

Observation i is inherited to the right child if it lies on the positive half-space

$$\min \delta + c_1 \sum_{i=1}^{n} \sum_{t=1}^{T} e_{it} + c_2 \sum_{i=1}^{n} \sum_{t=1}^{T} \xi_{it} + c_3 \sum_{t=1}^{T} d_t \qquad \text{(OCTSVM)}$$

$$\text{s.t. } \frac{1}{2}||\omega_t||_2 \leq \delta, \quad \forall t = 1, \ldots, T,$$

$$y_i(\omega_t' x_i + \omega_{t0}) - 2y_i(\beta_t' x_i + \beta_{it0}) \geq 1 - e_{it} - M(1 - z_{it}), \qquad \forall i = 1, \ldots, N, t = 1, \ldots, T,$$

$$\beta_{itj} = \xi_{it}\omega_{tj}, \qquad \forall i = 1, \ldots, N, t = 0, \ldots, T, \ j = 0, \ldots, p,$$

$$\|\omega_t\|_2 \leq M d_t, \qquad \forall t = 1, \ldots, T,$$

$$d_t \leq d_{p(t)}, \qquad \forall t = 1, \ldots, T,$$

$$\sum_{t \in u} z_{it} = 1, \qquad \forall i = 1, \ldots, N, \ u \in U,$$

$$z_{it} \leq z_{ip(t)}, \qquad \forall i = 1, \ldots, N, t = 2, \ldots, T,$$

$$\omega_t' x_i + \omega_{t0} \geq -M(1 - \theta_{it}), \qquad \forall i = 1, \ldots, N, t = 1, \ldots, T,$$

$$\omega_t' x_i + \omega_{t0} \leq M\theta_{it}, \qquad \forall i = 1, \ldots, N, t = 1, \ldots, T,$$

$$z_{ip(t)} - z_{it} \leq \theta_{ip(t)}, \qquad \forall i = 1, \ldots, N, t \in \tau_{bl},$$

$$z_{ip(t)} - z_{it} \leq 1 - \theta_{ip(t)}, \qquad \forall i = 1, \ldots, N, t \in \tau_{br},$$

$$e_{it} \in \mathbb{R}^+, \beta_{it} \in \mathbb{R}^p, \beta_{it0} \in \mathbb{R}, \xi_{it}, z_{it}, \theta_{it} \in \{0, 1\}, \forall i = 1, \ldots, N, t = 1, \ldots, T,$$

$$\omega_t \in \mathbb{R}^p, \omega_{t0} \in \mathbb{R}, d_t \in \{0, 1\}, \forall t = 1, \ldots, T.$$

Experimental Section

Overview

- The authors test their model alongside other aforementioned paradigms: **CART, OCT, OCT-H**
- They test on 8 different UCI Machine Learning Repository datasets
- To test for noise robustness, they apply different amounts of noise (label flips) in these datasets
- They compare the models in terms of average accuracy, after 10 evaluations using cross-validation

Algorithms

- **CART**: implemented off-the-shelf in sklearn python library (*DecisionTreeClassifier(criterion='gini'))*
- **OCT**: implemented here, based on Bertsimas et al. 2017
- **OCT-H**: a less constrained version of OCT, so we implement it by removing the constraint of using a single variable per node from OCT - **NOT included**
- **OCTSVM**: implemented by coding in the IMNLP formulation using the Gurobi library (code in my Github)

Experimental Setup

- 4 different noise levels / dataset: 0%, 20%, 30%, 40%
- 4-fold cross validation per experiment
- Results include the best average accuracy across the different hyperparameter setups
- Computed over 10 different hyperparameter values (10^i, $i \in [-5, 5]$ $\forall$ cost parameter)
- Experiments executed on an Intel i71165G7 CPU with 16GB RAM - both shared with other programs
- 30 second limit / experiment

Notes

- Due time constraints, the optimal hyperparameter search space for the OCTSVM was limited to i $\in$ [-4,1) for c_1 & c_2 and [-2,1) for c_3, in contrast to the paper's [-5, 5] and [-2,2] respectively.

- Although to the best of my knowledge I managed to locate the paper's datasets, minor discrepancies in size and features might be present due to old.

- Ultimately, the different processor, different hyperparameter tuning in the case of OCTSVM and discrepancies in the datasets could have resulted in the slightly different results.

Dataset	% Flipped	CART		OCT		OCTSVM	
		Paper	Ours	Paper	Ours	Paper	Ours
Australian	0	82.99	**85.51**	85.44	83.48	**86.34**	84.5
	20	74.85	83.47	83.45	**85.5**	**84.55**	78
	30	66.87	**82.03**	79.15	70.72	**80.24**	72.08
	40	56.93	**76.67**	71.34	68.68	**73.89**	55.94
BreastCancer	0	92.22	94.56	93.18	94.71	**96.25**	**95.57**
	20	90.47	93	90.92	88.13	**91.57**	**95.28**
	30	83.29	93.42	**90.87**	89.56	87.98	**92.41**
	40	77.75	87.69	**86.50**	86.82	81.92	**92.27**
Heart	0	73.66	77.37	73.88	77.01	**84.13**	**84.84**
	20	72.98	67.78	71.63	71.09	**82.61**	**81.1**
	30	68.22	65.57	70.18	68.87	**80.51**	**79.98**
	40	62.36	66.62	64.52	63.77	**76.19**	**68.54**
Ionosphere	0	83.08	**89.45**	81.80	82.62	**85.51**	89.16
	20	75.65	**81.2**	79.59	78.91	**80.51**	74.95
	30	70.02	73.53	75.20	**76.08**	**77.79**	50.12
	40	60.17	64.1	70.65	**75.81**	**75.71**	43.29
MONK's	0	60.67	78.01	59.03	79.63	**61.26**	**85.18**
	20	57.11	75	59.59	75.46	**60.48**	**78.7**
	30	57.19	76.15	57.87	**77.55**	**60.09**	73.14
	40	54.01	64.81	55.25	**75**	**60.12**	74.07
Sonar	0	65.36	71.63	66.90	69.23	**74.76**	**79.33**
	20	57.41	60.58	60.38	**67.3**	**69.92**	65.38
	30	54.64	59.13	58.97	**62.98**	**67.30**	62.02
	40	55.16	55.77	57.77	**62.01**	**63.84**	56.25
Wholesale	0	**90.28**	**91.59**	90.01	91.14	89.75	88.41
	20	83.79	88.86	87.84	**90.23**	82.06	87.72
	30	78.54	84.09	**84.58**	84.77	78.13	**85.91**
	40	69.52	67.72	**75.79**	76.14	71.72	**85.23**

Table 1. Accuracy (%) comparison of the models between [1] and our experiments (excl. averages - bold is best).

Key observations

- The main observation made by the paper's authors still stands:

  OCTSVM usually exhibits the greatest accuracy

- However the performance of OCTSVM isn't as robust under the noisy datasets as presented in the original paper. In fact, OCT seems to outperform it in several cases. This might be due to the restrained parameter search space we used in these simulations.

- Also, CART exhibits better performance in these calculations in respect to the paper's.

Final remarks

- A significant fluctuation in performance can be observed using the OCTSVM model with the same parameters and dataset, by simply using different training and testing subsets. That potentially points out the model's volatility and high sensibility in factors like runtime.

- In many cases, the OCTSVM model used only the first node of the tree, essentially **reducing** the model **to a RE-SVM.** While that is probably due to the tree depth cost parameter, c_3, this was kept within the lower bounds set by the paper. Perhaps higher runtime limits, more complex datasets, lower c_3 parameter values and higher max depth would make use of extra tree levels.

- The time difference in the runtime of these models (for reference, CART ran all the experiments within a few seconds, while OCT was a bit faster than OCTSVM in that it didn't use the full 30s in some cases - both required hours, if not days to complete their experiments) perhaps isn't worth the slightly improved accuracy in most applications.

Thank you!

References

[1] V. Blanco, A. Jap´on, and J. Puerto, “Robust optimal classification trees under noisy labels,” arXiv:2012.08560 [cs.LG], Dec. 2020.

[2] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, Classification and Regression Trees, 1st ed. Belmont, CA, USA: Wadsworth, 1984.

[3] V. Vapnik, The Nature of Statistical Learning Theory, 1st ed. New York,NY, USA: Springer, 1995.

[4] V. Blanco, A. Jap´on, and J. Puerto, “A mathematical programming approach to SVM-based classification with label noise,” Computers & Industrial Engineering, vol. 172, p. 108611, 2022.

References

[5] H. A. Taha, Operations Research: An Introduction, 10th ed. Upper Saddle River, NJ, USA: Pearson, 2017.

[6] T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd ed. New York, NY, USA: Springer, 2009.

[7] D. Bertsimas and J. Dunn, “Optimal classification trees,” Machine Learning, vol. 106, no. 7, pp. 1039–1082, 2017