Cost-sensitive learning of classification trees, with application to imbalanced datasets

Magzhan Gabidolla
Dept. CSE, UC Merced
mgabidolla@ucmerced.edu

Arman Zharmagambetov
Meta AI (FAIR)
armanz@meta.com

Miguel Á. Carreira-Perpiñán
Dept. CSE, UC Merced
mcarreira-perpinan@ucmerced.edu

Many important practical applications involve a binary classification problem with imbalanced classes (e.g. few positives and many negatives) or asymmetric costs (e.g. a false positive is much more costly than a false negative), where the positives and negatives are suitably defined in each case. Examples are fraud or spam detection or churn prediction. Although many types of classifiers may be used, we focus on **classification trees**, which are widely recognized as among the most interpretable models. We consider the traditional axis-aligned trees (where each decision node uses a single feature) and also sparse oblique trees (where each decision node uses a linear combination of a small subset of features). The latter are far more powerful and subsume as particular cases axis-aligned trees (such as CART or C5.0) and linear classifiers (such as logistic regression or linear SVMs). In both cases each leaf node outputs a constant label.

In these types of problems, optimizing the raw accuracy does not work well because it can largely ignore the low-cost or infrequent class. It is desirable to have control on the number of false positives (FPs) or true positives (TPs). A popular way to achieve this is through the ROC curve, *but this results in suboptimal classifiers* because it does not explicitly optimize for accuracy or true positives while controlling the false positive rate.

Our paper has two contributions. Firstly, we formally propose the concept of *cost-optimal curve (COC)*. This defines a set of optimal-accuracy classifiers as a function of the false positive level. We give an equivalent, penalized formulation which has the form of **a weighted 0/1 loss** and (with our new algorithm) is more amenable to optimization, although still NP-hard in general. Although this idea is straightforward, the COC properties and its optimization appear not to have been explored before. Note this is different from using a weighted surrogate loss such as the cross-entropy [4, 6]; while this (being differentiable) can be easily optimized, its optimum can be quite far from the true one. Second, *we propose the first algorithm that directly tries to optimize this problem for classification trees, with the guarantee that the weighted 0/1 loss decreases monotonically at each iteration.* Our experiments confirm the algorithm provides a good approximation to the ideal COC curve and is generally much better than using the ROC curve or approaches based on over- or undersampling, cost-sensitive surrogates or weighted purity criteria for constructing the tree.

The suboptimality of the ROC curve The ROC curve is a fast, simple way to achieve a classifier with lower overall accuracy but a more desirable FP rate, particularly in cases with imbalanced classes or asymmetric class costs. However, it is clear that (except possibly for the base classifier) *this does not produce a classifier that, having the desired FP rate, is optimal within its model class*. To obtain this we need to solve a different optimization problem as a function of the FP rate, as we describe next.

Cost-Optimal Curve (COC) Assume a training set $\{(\mathbf{x}_n, y_n)\}_{n=1}^N$ where $\mathbf{x}_n \in \mathcal{X}$ and $y_n \in \{-1, +1\}$ for $n = 1, \dots, N$, and **a decision tree** $T(\cdot; \boldsymbol{\theta})$: $\mathcal{X} \to \{-1, +1\}$ with parameters $\boldsymbol{\theta}$. We want to optimize the number of false negatives (FN) subject to a constraint on the number of false positives (FP):

$$\min_{\boldsymbol{\theta}} \ \nu(\boldsymbol{\theta}) \quad \text{s.t.} \quad \pi(\boldsymbol{\theta}) \leq p \qquad \text{with } \nu(\boldsymbol{\theta}) = \sum_{n:\ y_n = +1}^{N} L(y_n, T(\mathbf{x}_n; \boldsymbol{\theta})),\ \pi(\boldsymbol{\theta}) = \sum_{n:\ y_n = -1}^{N} L(y_n, T(\mathbf{x}_n; \boldsymbol{\theta})) \tag{1}$$

Preprint. Under review.

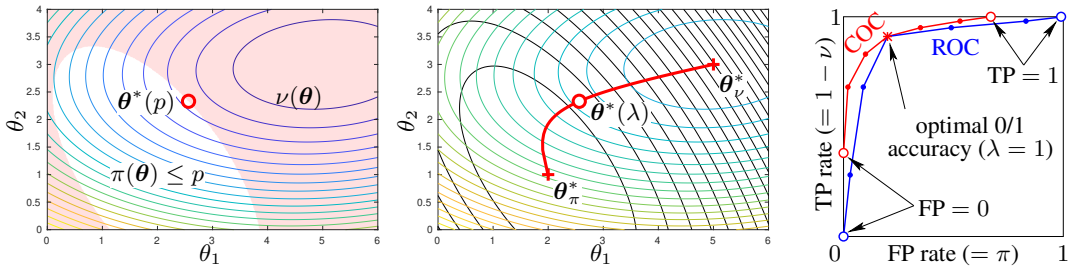

Figure 1: Illustration of the COC curve for a classifier with parameters $\boldsymbol{\theta} \in \mathbb{R}^2$, FP rate $\pi(\cdot)$ and FN rate $\nu(\cdot)$. *Left*: $\boldsymbol{\theta}^*(p)$ is an optimal classifier (minimizing ν) with an FP rate of at most p. The infeasible set is in pink. *Middle*: the optimal classifier path $\boldsymbol{\theta}^*(\lambda)$ over the cost λ, i.e., minimizing $\nu + \lambda\pi$, from $\boldsymbol{\theta}^*_\nu = \boldsymbol{\theta}^*(0)$ to $\boldsymbol{\theta}^*_\pi = \boldsymbol{\theta}^*(\infty)$. The contours of ν and π are in color and black, respectively. *Right*: the COC curve corresponding to the optimal classifier path and the ROC curve (assuming as base classifier that for $\lambda = 1$).

where $L(y, y') = 0$ if $y = y'$ and 1 if $y \neq y'$ is the 0/1 loss, and ν and π are the FN and FP rate, resp. Note $\nu(\boldsymbol{\theta}) + \pi(\boldsymbol{\theta}) = \sum_{n=1}^{N} L(y_n, T(\mathbf{x}_n; \boldsymbol{\theta}))$ is the 0/1 loss on the whole training set. This is a constrained optimization problem whose feasible set are those classifiers having a FP rate of at most p (fig. 1 (left)). By solving this for all $p \in [0, N]$ we obtain a finite collection of at most $N + 1$ classifiers, each for a different FP and FN (hence TP) rate, which defines the COC curve.

Cost-sensitive 0/1 loss Consider now the following unconstrained optimization problem:

$$\min_{\boldsymbol{\theta}} \ \nu(\boldsymbol{\theta}) + \lambda\,\pi(\boldsymbol{\theta}) \tag{2}$$

where $\lambda \geq 0$ and $\boldsymbol{\theta}$ defines the parameters of a decision tree (both axis-aligned and oblique). This objective function is **a weighted 0/1 loss**: it gives a misclassification cost of 1 to a FN and of λ to a FP. The usual misclassification error (unweighted 0/1 loss) results for $\lambda = 1$. Problems (1) and (2) have the same set of solutions, i.e., solving (1) for all $p \in [0, N]$ produces the same set of classifiers as solving (2) for all $\lambda \geq 0$ and hence the same COC curve, but solving (2) is easier than (1) in our case). Note that, if the base classifier used to construct the ROC curve was trained to minimize the 0/1 loss (i.e., eq. (2) with $\lambda = 1$), then the ROC curve touches the COC curve for $\lambda = 1$, but the latter otherwise dominates. See fig. 1 (middle and right).

Optimizing a cost-sensitive 0/1 loss over a classification tree. To realize the advantages of the COC curve one needs to optimize (2), which uses a weighted 0/1 loss, as best as possible. Traditionally, as in CART and C5.0, decision trees have been trained in a heuristic way by greedy recursive partitioning. This sets a decision node's split by optimizing a local "purity" criterion (such as the Gini index or information gain), but the overall procedure does not optimize any loss over the entire tree [2, 5, 3]. Recently, an algorithm has been proposed (*Tree Alternating Optimization (TAO)*) [1] which does optimize a global loss over a parametric tree (axis-aligned or oblique). In this work, we extend TAO to handle a weighted 0/1 loss objective such as that in (2).

Our experimental results show in general the dominance of COC curves over the whole TP-rate vs FP-rate space for oblique decision trees trained with TAO compared with other approximate methods such as cost supportive CART or C5.0 or other oblique decision trees such as OC1 with sampling. Fig. 2 depicts some subset of results: COC consistently achieves better results than the baselines.

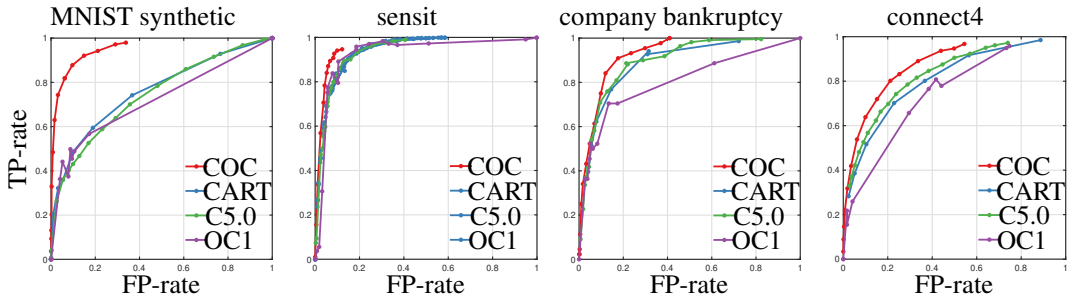

Figure 2: Some experiment results on imbalanced datasets.

Acknowledgments: Work funded in part by NSF award IIS–2007147.

References

[1] M. Á. Carreira-Perpiñán and P. Tavallali. Alternating optimization of decision trees, with application to learning sparse oblique trees. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems (NEURIPS)*, volume 31, pages 1211–1221. MIT Press, Cambridge, MA, 2018.

[2] C. Drummond and R. C. Holte. Exploiting the cost (in)sensitivity of decision tree splitting criteria. In *Proceedings of the Seventeenth International Conference on Machine Learning*, ICML '00, page 239–246, San Francisco, CA, USA, 2000. Morgan Kaufmann Publishers Inc. ISBN 1558607072.

[3] T. J. Hastie, R. J. Tibshirani, and J. H. Friedman. *The Elements of Statistical Learning—Data Mining, Inference and Prediction*. Springer Series in Statistics. Springer-Verlag, second edition, 2009.

[4] S. Höppner, B. Baesens, W. Verbeke, and T. Verdonck. Instance-dependent cost-sensitive learning for detecting transfer fraud. *Eur. J. Operational Research*, 297(1):291–300, Feb. 16 2022.

[5] R. F. Raubertas, L. E. Rodewald, S. G. Humiston, and P. G. Szilagyi. Roc curves for classification trees. *Medical Decision Making*, 14(2):169–174, 1994. URL `https://doi.org/10.1177/0272989X9401400209`.

[6] T. Vanderschueren, T. Verdonck, B. Baesens, and W. Verbeke. Predict-then-optimize or predict-and-optimize? an empirical evaluation of cost-sensitive learning strategies. *Information Sciences*, 594:400–415, May 2022.