Supplementary Material for Hybrid Rule Set

Tong Wang [1]

1. Proofs

Proof 1 *(of Theorem 1) To prove an optimal model does not contain any rules with support smaller than a threshold, we prove that if there's a model that contains such a rule z, removing it will always decrease the objective value, thus violating the optimality assumption.*

We start with positive rules. For a rule $r \in \mathcal{R}^*_+$, we define

$$\mathcal{R}^*_{\setminus r} = \{z \in \mathcal{R}^*_+, z \neq r\}. \tag{1}$$

We want to find conditions on the support such that the following inequality always holds.

$$\Lambda(\mathcal{R}^*_{\setminus r}) \leq \Lambda(\mathcal{R}^*), \tag{2}$$

where

$$\begin{aligned}\Lambda(\mathcal{R}^*) =&\ell(\langle\mathcal{R}^*, f_b\rangle, \mathcal{D}) + \theta_1 \cdot \textit{size}(\mathcal{R}^*)\\ &- \theta_2 \cdot \frac{\textit{support}(\mathcal{R}^*)}{N} \qquad (3)\\ \Lambda(\mathcal{R}^*_{\setminus r}) =&\ell(\langle\mathcal{R}^*_{\setminus r}, f_b\rangle, \mathcal{D}) + \theta_1 \cdot (\textit{size}(\mathcal{R}^*) - 1)\\ &- \theta_2 \cdot \frac{\textit{support}(\mathcal{R}^*_{\setminus r})}{N}. \qquad (4)\end{aligned}$$

Since we want inequality (2) to hold for any $r \in \mathcal{R}^*$, we upper bound $\Lambda(\mathcal{R}^*_{\setminus r})$ by upper bounding $\ell(\langle\mathcal{R}^*_{\setminus r}, f_b\rangle, \mathcal{D})$ and support($\mathcal{R}^*_{\setminus r}$).

$$\ell(\langle\mathcal{R}^*_{\setminus r}, f_b\rangle, \mathcal{D}) \leq \ell(\langle\mathcal{R}^*, f_b\rangle, \mathcal{D}) + \frac{\textit{support}_\epsilon(r\mathcal{R}^*_+)}{N}, \tag{5}$$

with the minimum achieved when instances originally covered by r are all incorrectly classified after removing r.

$$\textit{support}(\mathcal{R}^*_{\setminus r}) \leq \textit{support}(\mathcal{R}^*) \tag{6}$$

with the minimum achieved when all instances originally covered by r are now covered by $\mathcal{R}^*_-$, therefore does not change the coverage of $\mathcal{R}^*$ overall.

Plugging formula (5) and (12) into equation (4) and combine it with (3) and (2) yields

$$\textit{support} \leq N\theta_1. \tag{7}$$

Thus, if support(r) $\leq N\theta_1$, removing it from a $\mathcal{R}^*$ will produce a better model. Therefore, such rules do not exist in an optimal model $\mathcal{R}^*$.

Then we follow the similar steps to prove for negative rules. We define $\mathcal{R}^*_{\setminus r}$ as a set of rules where r is removed from $\mathcal{R}^*_-$. The proofs here use the same steps from inequality (1) to (5). The effective coverage, however, equals to support($\mathcal{R}^*$) $-$ support$_\epsilon$($r\mathcal{R}^*_-$). Thus

$$\textit{support} \leq \frac{N\theta_1}{1-\theta_2}. \tag{8}$$

To summarize, $\mathcal{R}^*_+$ does not contain any rules with support $\leq N\theta_1$ and $\mathcal{R}^*_-$ does not contain any rules with support $\leq \frac{N\theta_1}{1-\theta_2}$.

Proof 2 *(of Theorem 2) We choose the optimal model found till time t to be the benchmark to compare with $\mathcal{R}^*$. Since $\mathcal{R}^* \in \min \Lambda(\mathcal{R})$,*

$$\lambda^*_{[t]} \geq \Lambda(\mathcal{R}^*), \tag{9}$$

Now we lower bound $\Lambda(\mathcal{R}^*)$, following equation (2)

$$\Lambda(\mathcal{R}^*) \geq 0 + \theta_1\Omega(\mathcal{R}^*) - \theta_2 \tag{10}$$

Combining inequality (12) and (10) yields

$$\Omega(\mathcal{R}^*) \leq \frac{\lambda^*_{[t]} + \theta_2}{\theta_1}. \tag{11}$$

Proof 3 *(of Theorem 3) We again compare $\mathcal{R}^*$ with the best model we found till time t and*

$$\lambda^{[t]} \geq \Lambda(\mathcal{R}^*), \tag{12}$$

Then we lower bound $\Lambda(\mathcal{R}^*)$,

$$\Lambda(\mathcal{R}^*) \geq 0 + \theta_1\Omega(\mathcal{R}^*) - \theta_2\frac{\textit{support}(\mathcal{R}^*)}{N} \tag{13}$$

If $\mathcal{R}^* \neq \emptyset$, then $\Omega(\mathcal{R}^*) \geq 1$, then

$$\lambda^{[t]} \geq \theta_1 - \theta_2\frac{\textit{support}(\mathcal{R}^*)}{N}, \tag{14}$$

Thus

$$\textit{support}(\mathcal{R}^*) \geq \frac{N(\theta_1 - \lambda^{[t]})}{\theta_2} \tag{15}$$

[1] Department of Business Analytics, University of Iowa, Iowa, USA. Correspondence to: Tong Wang <tong-wang@uiowa.edu>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019. Copyright 2019 by the author(s).

2. Tuning Interpretable Baselines

We use R and python packages (Hornik et al., 2007; Quinlan, 2004; Pedregosa et al., 2011) for baseline methods except for BRS which has the code publicly available. For C4.5 and C5.0, we tune the minimum number of samples in at least two of the splits. For BRS and SBRL, we set the maximum length of rules to 3. For BRS, there are parameters $\alpha_+, \beta_+, \alpha_-, \beta_-$ that govern the likelihood of the data. We set β_+, β_- to 1 and vary α_+, α_- from $\{100, 1000, 10000\}$. For SBRL, there are hyperparameters λ for the expected length of the rule list and η for the expected cardinality of the rules in the optimal rule list. We vary λ from $\{5, 10, 15, 20\}$ and η from $\{1, 2, 3, 4, 5\}$.

References

Hornik, K., Zeileis, A., Hothorn, T., and Buchta, C. Rweka: an r interface to weka. *R package version*, pp. 03–4, 2007.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830, 2011.

Quinlan, R. Data mining tools see5 and c5. 0. 2004.