Probabilistic Rule Sets Ready for Interactive Machine Learning

Lincen Yang[*†], **Matthijs van Leeuwen**

LIACS, Leiden University
{l.yang, m.van.leeuwen}@liacs.leidenuniv.nl

Extended Abstract

Introduction As data-driven decisions have become widely used in society, including in fields with very limited tolerance for mistakes such as medicine, rule-based machine learning methods have witnessed a renaissance. Three categories of rule learning methods exist: rules induced from decision trees, (ordered) rule lists, and (unordered) rule sets, among which rule sets are arguably the most interpretable and hence are the easiest to work with for domain experts.

We argue, however, that rule sets induced by existing methods are not ready for interactive machine learning (i.e., to allow domain experts to give feedback for model training and/or prediction), because of the issues caused by overlapping rules (i.e., one instance is covered by multiple rules). Most existing methods propose separate schemes to resolve prediction conflicts caused by overlaps (Clark and Boswell 1991; Boström 2004; Zhang and Gionis 2020; Lakkaraju, Bach, and Leskovec 2016), and they often assign the instances covered by multiple rules to one of these rules using a separate criterion (e.g., to the most accurate rule). However, they all suffer from two issues: 1) they are not probabilistic; 2) rules become implicitly dependent on each other: each rule cannot be examined by domain experts as an independent piece of knowledge, as the instances covered by one rule may be assigned to another rule for prediction.

These disadvantages make interactive rule learning difficult for the following reasons: 1) without proper probabilistic modeling for overlaps, the likelihood of the whole dataset cannot be calculated, which makes it difficult to inject prior preferences on rules as a prior distribution under the Bayesian framework; 2) while examining individual rules is the first step for domain experts giving feedback in many cases, separate schemes that assign instances to one of the overlapping rules potentially create many more rules to be examined by the domain experts; we will further elaborate on this later.

To solve these problems, we propose a probabilistic formalization of rule set models. Preliminary experiment results on benchmark datasets for classification tasks are surprisingly promising, which can be regarded as an empirical validation of our formalization.

Probabilistic rule sets Consider a random vector $X = (X_1, \ldots, X_d)$ and a categorical variable Y, a probabilistic rule is in the form of $(X_1 \in V_1 \wedge X_2 \in V_2 \wedge \ldots) \rightarrow P_Y$, where each of the $V_1, V_2, ...$ is either an interval or a set of categorical levels. A probabilistic rule essentially represents a subset S of the full sample space of X, such that for any $x \in S$, the conditional distribution $P(Y|X = x)$ is approximated by a single probability distribution denoted as P_Y, which does not depend on the specific value of $x \in S$. Further, a probabilistic rule set is a set of such rules that aims for describing the whole dataset accurately.

Approach While some existing methods treat overlaps in a rule set as a nuisance and explicitly aim for minimizing overlaps (Zhang and Gionis 2020; Lakkaraju, Bach, and Leskovec 2016), we consider overlap useful in the following two situations: overlap because of *uncertainty* and overlap representing *exceptions*.

First, hypothetically, imagine that we have enough data to infer the probability of flu given cough or fever, *respectively*, but not the probability of flu given both cough and fever. Hence, we may reliably induce the following two rules: "Fever $\rightarrow P(flu) = a$" and "Cough $\rightarrow P(flu) = b$" (where a and b are the probability estimates), but not a rule corresponding to the condition "Fever $\wedge$ Cough", as the probability estimate of the target variable must have substantial certainty. When the uncertainty is so large that the probability estimator is not significantly different from either a or b, the probabilistic modeling for the instances covered by this overlap becomes robust no matter which rule is used. In this situation, we argue that the overlap should be kept, as it is useful to 1) keep the rules compact, and 2) express the uncertainty of the rule set, indicating that prior knowledge from domain experts might be useful here.

Second, consider a different situation where we hypothetically induce two rules from the data: "Fever $\rightarrow P(flu) = a$" and "Fever $\wedge$ Cough $\rightarrow P(flu) = c$". The overlap between these two rules is different than in the previous case, as the first rule fully "contains" the second rule. If

[*]Accepted at the AAAI-22 Workshop on Interactive Machine Learning (IML@AAAI'22).

[†]This work is part of the research programme 'Human Guided Data Science by Interactive Model Selection' with project number 612.001.804, which is (partly) financed by the Dutch Research Council (NWO).

$a \approx P(flu|\text{Fever} \wedge \neg\text{Cough})$, we argue that the overlap should be kept to represent exceptions, i.e., the second rule is an exception of the first rule. In practice, one general rule may have many exceptions, so not allowing overlaps representing exceptions may lead to overly redundant rule sets.

Following these intuitions, we propose a principled probabilistic model enabling such overlaps, and we hence consider the rule set learning task as a probabilistic model selection problem, which we tackle by the minimum description length (MDL) principle (Grünwald and Roos 2019).

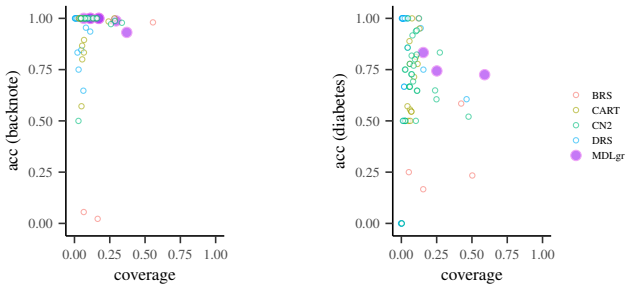

Figure 1: Individual rules qualities: coverage versus accuracy on the test sets.

Towards interactive rule set learning We now briefly discuss how our formalization can help domain experts digest rules and give feedback to a rule set model.

To begin with, introducing a separate scheme to assign the instances covered by overlapping rules to one of these rules will entangle the rules in the rule set. That is, when two rules overlap and the rule with condition A ranks higher than the rule with condition B due to the scheme, the scheme implicitly creates a new rule with condition $(\neg A \wedge B)$. As a result, even interactions in simple forms (e.g., users giving like/dislike feedback to rules) become burdensome, as the number of (implicit) rules that domain experts potentially need to examine can grow substantially. In contrast, with our formalization, each rule can be regarded as an independent piece of knowledge, which is more interpretable to domain experts.

Further, our formalization—with explicit interpretations for overlaps—is useful for domain experts giving feedback in the following two situations. First, it is common that a domain expert does not agree with a certain prediction by an individual rule because it neglects "important" variables; i.e., the rule is too general. Based on our formalization, domain experts in this case can directly ask for a more refined rule containing these important variables, which represents an exception of the more general rule.

Second, when domain experts would like to directly correct or modify rules induced from data, the rules they input can also have overlaps; i.e., our formalization allows domain experts to be uncertain about the interactive effect of multiple rules, which is common in practice.

Preliminary results for empirical validation We next validate our probabilistic formalization empirically. Together with the model selection criterion based on the formalization, we apply a baseline rule learning algorithm to benchmark datasets for classification tasks. The baseline algorithm iteratively adds the "best" individual rule to the rule set and optimizes our model selection criterion greedily.

dataset	algorithm	acc	auc	# rules	rule length
backnote	BRS	**0.97**	0.96	7.65	2.81
backnote	CART	0.94	0.98	11	3.54
backnote	CN2	0.92	**0.99**	13.7	2.88
backnote	DRS	0.95	0.95	20.9	2.66
backnote	**MDLgreedy**	**0.97**	0.98	**6**	**1.83**
diabetes	BRS	0.72	0.69	3.95	2.85
diabetes	CART	0.73	**0.78**	13	3.85
diabetes	CN2	0.64	0.67	30	4.41
diabetes	DRS	0.65	0.67	23.7	4.2
diabetes	**MDLgreedy**	**0.74**	0.73	**2.3**	**1.89**
magic	BRS	**0.83**	0.79	7.55	**3**
magic	CART	0.82	**0.87**	15	4.47
magic	CN2	0.59	0.61	419.8	4.81
magic	DRS	0.61	0.68	12	3.94
magic	**MDLgreedy**	0.8	0.85	**6.15**	3.18
anuran	CART	**0.88**	0.92	11	3.64
anuran	CN2	0.84	**0.93**	59.85	4.03
anuran	DRS	0.8	0.86	42.6	9.05
anuran	**MDLgreedy**	0.85	0.75	**4.05**	**3.5**
avila	CART	0.61	0.87	16	5
avila	CN2	**0.81**	**0.9**	360.1	**3.52**
avila	DRS	0.18	0.58	31.1	6.26
avila	**MDLgreedy**	0.54	0.85	**6.1**	3.58
contraceptive	CART	**0.53**	**0.68**	14	4.57
contraceptive	CN2	0.41	0.58	54.45	4.58
contraceptive	DRS	0.37	0.53	15.7	4.92
contraceptive	**MDLgreedy**	0.46	0.6	**2.1**	**1.83**
iris	CART	**0.93**	**0.98**	5	2.8
iris	CN2	0.91	0.98	4	2.08
iris	DRS	0.92	0.94	11.3	2.36
iris	**MDLgreedy**	0.91	0.95	**2**	**1**

Table 1: Results on UCI datasets, averaged over 20 random training/testing splits (80% versus 20%): test accuracy, ROC-AUC, number of rules, and average rule lengths.

We compare the performance of this "MDL-based greedy" algorithm with both classic and recently proposed rule learning algorithms. We use the same seven datasets as DRS (Zhang and Gionis 2020) used for its empirical evaluation, which is the most recently proposed (unordered) rule set learning algorithm. We compare with DRS, BRS (Wang et al. 2017) (only for binary classification), CART (Breiman et al. 1984), and CN2 (Clark and Boswell 1991).

As shown in Table 1, we obtain competitive results in several aspects based on our novel formalization that deals with overlaps of rules in a principled way. First, the prediction accuracy is close to the best, although AUC is slightly weaker. Second, our method always produces the smallest number of rules and almost always the smallest average rule length.

Besides, we show in Figure 1 that our method produces high-quality rules with both high accuracy and large coverage. These results demonstrate that our probabilistic formalization has great potential and hence more advanced algorithms will be developed specifically for our formalization as future work.

References

Boström, H. 2004. Pruning and exclusion criteria for unordered incremental reduced error pruning.

Breiman, L.; Friedman, J.; Stone, C. J.; and Olshen, R. A. 1984. *Classification and regression trees*. CRC press.

Clark, P.; and Boswell, R. 1991. Rule induction with CN2: Some recent improvements. In *European Working Session on Learning*, 151–163. Springer.

Grünwald, P.; and Roos, T. 2019. Minimum description length revisited. *International journal of mathematics for industry*, 11(01): 1930001.

Lakkaraju, H.; Bach, S. H.; and Leskovec, J. 2016. Interpretable decision sets: A joint framework for description and prediction. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 1675–1684.

Wang, T.; Rudin, C.; Doshi-Velez, F.; Liu, Y.; Klampfl, E.; and MacNeille, P. 2017. A bayesian framework for learning rule sets for interpretable classification. *The Journal of Machine Learning Research*, 18(1).

Zhang, G.; and Gionis, A. 2020. Diverse Rule Sets. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 1532–1541.