When characteristic rule-based models should be preferred over discriminative ones

Florian Beck[1], Johannes Fürnkranz[1] and Van Quoc Phuong Huynh[1]

[1]Johannes Kepler University Linz, LIT Artificial Intelligence Lab / Institute for Application-oriented Knowledge Processing (FAW), Altenberger Straße 66b/69, 4040 Linz, Austria

Abstract

In recent years, the interpretability of machine learning models has gained interest. White-box approaches like rule-based models serve as an interpretable alternative or as surrogate models of black-box approaches. Among these, more compact rule-based models are considered easier to interpret. In addition, they often generalize better and thus provide higher predictive accuracies than their overfitting complex counterparts. In this paper, we argue that more complex, "characteristic" rule-based models are a genuine alternative to more compact, "discriminative" ones. We discuss why characteristic models should not be considered as less interpretable, and that more included features can actually strengthen the model both in terms of robustness and predictive accuracy. For this, we evaluate the effects on the decision boundary for models of different complexity, and also modify a recently developed Boolean pattern tree learner to compare a characteristic and a discriminative version on five UCI data sets. We show that the more complex models are indeed more robust to missing data, and that they sometimes even improve the predictive accuracy on the original data.

Keywords

characteristic rules, discriminative rules, decision boundaries, interpretability, robustness

1. Introduction

With the rise of neural network models in many machine learning applications, the need has grown to actually understand what these black-box approaches learn. This has brought rule-based models back into the spotlight which can be used as interpretable surrogates of neural network approaches, e.g., by extracting rules from the whole network [1] or with the focus on explaining decision boundaries [2].

Independent of whether rule-based models are used as surrogates of neural networks or as a stand-alone model, usually the principle of Occam's Razor [3] is followed, which can be loosely translated as that the simplest explanation is the best one. Consequently, **discriminative rules** which discriminate an object of one category from objects of other categories are preferred over **characteristic rules** which try to capture all properties that are common to the objects of the target class [4]. This principle is also supported by the observation that longer explanations tend to overfit the training data, leading to worse performances on test data. Hence, most rule learner use some kind of pruning policy [5], resulting in learning short discriminative rules instead of longer characteristic ones.

However, there is a fine line between avoiding overfitting and learning too general theories. Consider the sample dataset in Table 1 consisting of six countries, three belonging to the class *Europe* and three to the class *South America*. For each country, the value for the three numeric attributes *Size*, *Age* and CO_2 is provided.

Table 1
A small country dataset with three numeric attributes *size* (in 1,000 km²), *age* (median; in years) and CO_2 (emission per capita and year; in tons)[1]. It is split into six training examples (three for each of the classes *Europe* and *South America*) and four test examples of unknown class.

	Size	Age	CO_2	Class
Austria	84	42.8	6.9	Europe
Bolivia	1099	23.9	1.8	South America
Brazil	8515	32.8	2.2	South America
Czechia	79	42.6	9.3	Europe
Ecuador	284	27.6	2.3	South America
Slovakia	49	40.6	6.1	Europe
Albania	29	37.3	1.7	?
Germany	357	44.9	8.0	?
Kosovo	11	30.5	4.8	?
Uruguay	176	35.2	2.3	?

Traditional rule learners like, e.g., Ripper [6] strive for discriminative rules, i.e., rules that minimize the number of used attributes when describing the classes. In this case, such a perfect, minimal description of the training data could be learned with a single rule r_1 only considering the first attribute *Size*, and the corresponding default rule r_0 for the other class[2]:

fbeck@faw.jku.at (F. Beck); juffi@faw.jku.at (J. Fürnkranz); vqphuynh@faw.jku.at (V. Q. P. Huynh)

CEUR Workshop Proceedings (CEUR-WS.org)

[2]In the following, $class = europe$ is abbreviated as $c = e$ and $class = south\_america$ as $c = sa$

$$r_1 : c = e \leftarrow size < 184$$
$$(r_0 : c = sa \leftarrow \top). \tag{1}$$

Rule r_1 *covers* the three examples *Austria*, *Czechia* and *Slovakia* because these examples fulfill the condition $size < 184$. *Bolivia*, *Brazil* and *Ecuador* are not covered by r_1 but only by the most general rule r_0, thus classified as *South America*. While these rules perfectly describe the training examples, they fail to correctly classify the test example *Germany*, which is not covered by r_1 and hence misclassified as *South America* by r_0. Vice versa, *Uruguay* is covered by r_1 and hence misclassified as *Europe* . Note that these misclassifications could have been avoided if a different feature would have been picked, such as, e.g., in rules r_2 and r_3:

$$r_2 : c = e \leftarrow age \geq 36.7$$
$$r_3 : c = e \leftarrow CO_2 \geq 4.2. \tag{2}$$

However, r_2 does not cover the test example *Kosovo*, and r_3 does not cover *Albania*, so that neither of the three rules would be sufficient to classify all four test examples correctly, but only a combined rule set of r_2 and r_3 would do so. Similarly, the three suggested features $size < 184$, $age \geq 36.7$ and $CO_2 \geq 4.2$ can also be connected by conjunctions to a single rule r_e for class *Europe*, while the respectively contrasting features form rule r_s for class *South America*:

$$r_e : c = e \leftarrow size < 184 \wedge age \geq 36.7 \wedge CO_2 \geq 4.2$$
$$r_s : c = sa \leftarrow size \geq 184 \wedge age < 36.7 \wedge CO_2 < 4.2. \tag{3}$$

While none of the two rules covers any of the test examples, a slight modification of their semantics allows us to use them as reliable classifiers. Instead of requiring that all conditions of a rule need to be satisfied, we instead assign an example to its closest rule, a method that is reminiscent of *rule stretching* [7] or *nearest hyperrectangle classification* [8]. In our example, the first three test examples are assigned to class *Europe*, since for each of them two out of three conditions of r_e are satisfied and only one out of three of r_s. Analogously, test example *Uruguay* is correctly classified as *South America*.

Independent of using conjunctions or disjunctions as the connector, we notice that more characteristic rule theories in Equations 2 and 3 are able to classify all four test examples correct, while the discriminative rule theory in Equation 1 is not able to do so. Moreover, the inclusion of more features in the characteristic concepts might not only lead to a better performance but also arguably provide more interesting and interpretable models, stating that European and South American countries do not only differ in size, but also in median age and CO_2 emissions.

The rest of the paper is organized as follows: Section 2 further specifies the problem of finding good decision boundaries and presents characteristic models of non-rule-based classifiers as an inspiration for adaption in the rule-based setting, presented in Section 3. We modify a rule-based learner in Section 4 accordingly and evaluate a discriminative and characteristic version in Section 5 in terms of predictive accuracy and robustness. Section 6 concludes the results and takes a brief look at the remaining challenges.

2. Decision boundaries

As depicted in the introduction, in contrast to long characteristic rules being prone to overfitting, short discriminative rules come with the risk of providing too simplistic theories that overgeneralize. This can also be illustrated by the decision boundary of the country dataset rules, see Figure 1. For a better visualization, we omit the third attribute CO_2 to obtain a two-dimensional feature space using the attribute *Size* in logarithmic scale on the x-axis and *Age* on the y-axis. The raw data are shown in Figure 1a; the six training examples as points and the four test examples as circles, colored in blue for class *Europe* and in red for class *South America*, respectively.

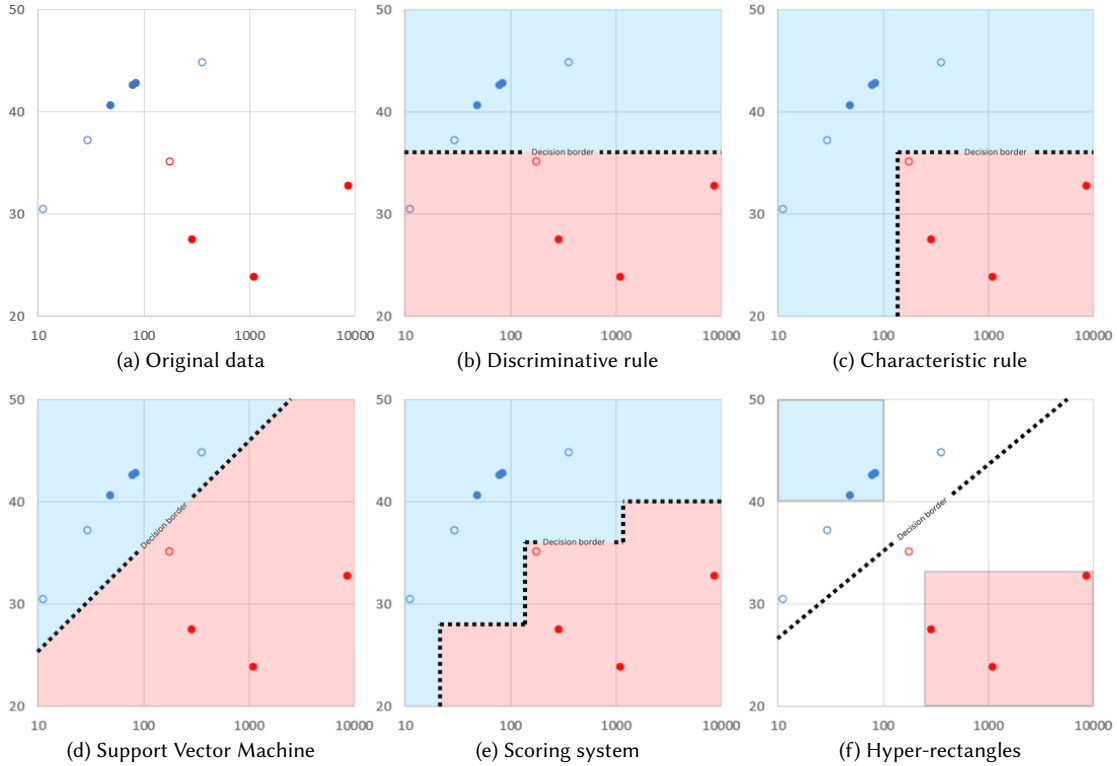
We see that the training examples are quite easily separable from each other, while the test examples complicate finding a good decision boundary. Figure 1b shows a discriminative rule $c = sa \leftarrow age < 36$, covering all four South American countries along with one European in the light-red area of the feature space. The light-blue area (classified by the default rule $c = e \leftarrow \top$) contains five true negatives. By adding the condition $size > 140$, the rule can be defined more characteristic, leading to a perfect classification of all examples, see Figure 1c.

Still, the provided decision boundary in Figure 1c can be considered suboptimal when compared with non-rule-based models. Figure 1d illustrates an arguably better decision boundary which other methods like, e.g., support vector machines [9], logistic regression [10] and naive Bayes [11] can find. All approaches have in common that they usually consider all attributes in the feature space and rely on continuous coefficients to build their models; in this case:

$$c = sa \leftarrow age - 10 \cdot \log_{10}(size) \leq 15.$$

In comparison to the methods just mentioned, conventional rule learners only use combinations of attribute-value-combinations for the splits of their classes. As a consequence, one of the main limitations of rule learning is arguably its restriction to axis-parallel decision boundaries. Though, the last two subfigures show two

[2]Retrieved 2024/07/04 from https://ourworldindata.org/age-structure and https://ourworldindata.org/co2-and-greenhouse-gas-emissions.



(a) Original data (b) Discriminative rule (c) Characteristic rule

(d) Support Vector Machine (e) Scoring system (f) Hyper-rectangles

Figure 1: Different decision boundaries of various learning approaches for the country dataset reduced to the attributes *Size* (x-axis, logarithmic) and *Age* (y-axis). (a) shows the six training examples as points and the four test examples as circles, colored in blue for class *Europe* and in red for class *South America*. The remaining subfigures (b)-(f) add a dotted decision boundary for various learners which show predictions as *Europe* in light-blue and as *South America* in light-red.

ways how rule-based methods can still mimic decision boundaries like in Figure 1d.

In Figure 1e, we see multiple steps in the decision boundary. Trivially, this behavior can be achieved by learning one rule for each step. While this is straightforward in this example, it is too hard to maintain in a high-dimensional feature space with an exponentially increasing number of combinations. Scoring systems [12] scale better by assigning low integer scores to attribute-value combinations, hereby providing a trade-off between rules and linear models. In the special case that all weights are binary, the scoring system converts into an m-of-n concept. With the scores being assigned by the following scheme and a threshold of 4 for class *South America*, all examples are classified correctly while providing a more customized decision boundary compared to Figure 1c:

$$age: \begin{cases} 3 & \text{if } < 28 \\ 2 & \text{if } < 36 \\ 1 & \text{if } < 40 \\ 0 & \text{else} \end{cases} \qquad size: \begin{cases} 3 & \text{if } \geq 1100 \\ 2 & \text{if } \geq 140 \\ 1 & \text{if } \geq 20 \\ 0 & \text{else.} \end{cases}$$

Finally, Figure 1f is the illustration of two characteristic rules similar to Equation 3: We describe both classes *Europe* and *South America* without using a default rule. Obviously, the learned rules of the two classes can overlap or — as in this case — leave wide areas of the feature space uncovered, so that a pure Boolean evaluation of the rules is not sufficient anymore. One way to handle these uncovered areas are nearest hyper-rectangles [13]. The decision boundary between the two classes can be shaped arbitrarily if enough hyper-rectangles, i.e., rules, are learned (and is actually neither quite straight in Figure 1f). Obviously, distances for nominal attributes can not be defined as straightforward as for numerical attributes, as is discussed in the following section.

In this work, we aim to expand rule-based approaches to reach this stronger expressiveness shown in the last three subfigures while still retaining the properties making them interpretable, i.e., without including all features instead of interactions and, most notably, without continuous coefficients like in SVMs, logistic regression or naive Bayes, for what characteristic rules are preferable.

3. Characteristic rule learning

So far we discussed why characteristic rules can be beneficial both in terms of interpretability and performance but observed as well that almost no rule-based methods learn such concepts. To understand why conventional rule learners prefer discriminative rules, we first briefly introduce the coverage space and related heuristics. Subsequently, we reveal potential issues with the latter and identify properties which should be taken into consideration when developing a characteristic rule-based learner.

3.1. Coverage space and heuristics

Traditionally, rules are gradually refined by adding individual conditions, whereby conjunctive refinements specialize a rule (afterwards it can never cover more examples than before the refinement), whereas disjunctive refinements generalize a rule (afterwards it can never cover fewer examples than before the refinement). This can be visualized in coverage space, a non-normalized ROC space, where the x-axis shows the covered negative and the y-axis the covered positive examples [14]. For example, Figure 2 shows a path that gradually refines an initially universal rule (covering all P positive and N negative examples, upper right corner of the coverage space) into the rule $+ \leftarrow c \wedge b$.

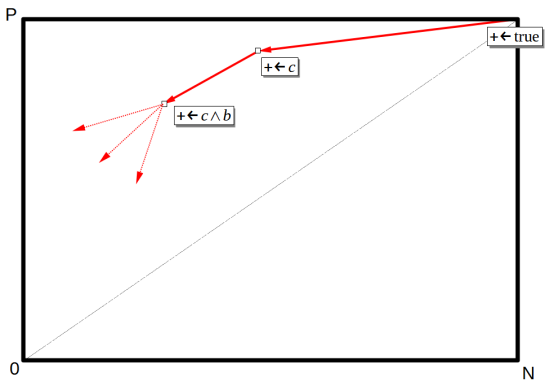


Figure 2: Rule refinement in coverage space

Apparently, a rule refined to the upper left corner can be considered perfect, since it covers only positive examples and no negatives. In most scenarios such a rule can not be found, so that a trade-off must be found between the importance of covering all positives (completeness) and not covering any negatives (consistency). For this purpose, heuristics are defined as functions $\mathrm{h}(p, n)$, where $0 \leq p \leq P$ $(0 \leq n \leq N)$ is the number of positive (negative) examples covered by a rule [14].

In previous studies it was found that most regular heuristics (in particular those striving for consistency) lead to the learning of discriminative rules, so that in this context, so-called inverted heuristics $\mathrm{u}(p, n)$ are suggested which better reflect the top-down nature of the rule refinement process in theory by originating from the other side of the coverage space [15]. Because of its typical focus on completeness, inverted heuristic can often "delay" the choice of too specific features, hence resulting in characteristic rules built of multiple more general features.

3.2. Limitations

Even though it has been shown empirically for some datasets that inverted heuristics result in characteristic rules [15], it is not inherent that they lead to characteristic rules. As a counterexample, consider learning a rule for the class *Europe* using all examples of the country dataset except of *Brazil*. The best single condition is $age \geq 29.1$ covering all six examples of class *Europe* as well as *Uruguay*. This false positive can not be excluded by further (single-cut) conditions on *Size* or CO_2 without losing coverage of at least one true positive, so that the inverted heuristic stops with a rule consisting of a single condition. Interestingly enough, in this case, regular heuristics would even learn longer rules than inverted heuristics, since they typically prefer this trade of removing a false positive at the cost of a false negative.

Most importantly though, traditional rule learners have a severe limitation of focusing only on the coverage of the learned rules but not how (well) they cover the examples. We already noticed in Table 1 that rule r_1 in Equation 1 can be expanded to r_e in Equation 3 by features considering the *age* and CO_2 of a country without covering more positive or less negative examples. Hence, both r_1 and r_e correspond to the same point in the coverage space in the top left corner, covering all positive and no negative training examples. As a consequence, independent of the chosen heuristic, conventional rule learners are not able to learn r_e if a refinement requires improving the heuristic.

Even if the heuristic improves by adding a new condition to the original rule a similar issue can occur. Assume a new rule r_4 learned on all ten examples in Table 1, which focuses on covering the example *Germany* based on the condition $size \geq 316$. This rule still covers *Bolivia* and *Brazil* as well and could therefore be refined to rules r_5 and r_6, both considering the *Age*-attribute:

$$\begin{aligned} r_4 &: c = e \leftarrow size \geq 316 \\ r_5 &: c = e \leftarrow size \geq 316 \wedge age \geq 36.3 \\ r_6 &: c = e \leftarrow size \geq 316 \wedge age \geq 44.5. \end{aligned} \tag{4}$$

While r_5 and r_6 both correspond to the same point in the coverage space (covering one positive example and no negatives), their coverage on unseen examples

might vary crucially since they cover different areas in the feature space. Arguably, r_5 should be preferred over r_6 because the added condition $age \geq 36.3$ covers four additional positive examples (and still no negative) compared to $age \geq 44.5$. So to say, while having the same "global" concept, we should choose the rule with the better "local" condition. Note that this is not limited to numeric attributes.

To summarize, characteristic rules are usually not learned because the learners rely on heuristics that only take the number of covered positive and negative examples into account instead of separating positive and negative examples with a variety of rules and conditions. Particularly, adding a condition without changing the covered examples results in the same heuristic value, in which case so far the shorter explanation is used, and the search usually stops. Additionally, the mere focus on the global coverage can lead to suboptimal "local" conditions if ties are not handled appropriately.

4. Boolean Pattern Trees

For the experimental comparison of deterministic and characteristic concepts, we use two versions of an *alternating Boolean pattern tree (ABPT)* learner recently developed in our group [16]. The task of learning an ABPT is quite similar to learning a rule: For every specific class $y_j \in \mathbb{Y}$, a tree $t : y_j \leftarrow B$ is learned, where B is a logical expression defined over the input features, which can be much more flexible than for rules. In contrast to rule learners which use either conjunctions or disjunctions, ABPTs can connect binary features by conjunctions and disjunctions in any arbitrary order. This also complicates the iterative learning of B, having multiple insertion options per feature. E.g., inserting a disjunction with feature c in $B = a \wedge b$ can result in $c \vee (a \wedge b)$, $(a \vee c) \wedge b$ or $a \wedge (b \vee c)$, which are all logically different. These insertions are repeated until the maximum number of iterations k (= number of features in the pattern tree) is reached. We refer to [16] for further details of the algorithm and focus on two adjustments for the experiments in the following.

First, we notice that in the standard version of ABPT already multiple heuristics for the evaluation of a tree extension are used, focusing on consistency and completeness in different search branches. By using various cost ratios in the *linear cost* metric ($h_{lc}(t) = c \cdot p - (1-c) \cdot n$), ABPT is capable to learn both models preferred of regular and inverted heuristics. Though, Section 3 presented as well problems that could not be fixed solely by the heuristic. To choose characteristic models instead of discriminative ones in case of ties, the learner picks the tree learned in a later iteration, i.e., using more conditions (and vice versa). This way we do not stop in local optima, but always use all k iterations. Furthermore, all conditions are sorted based on the *accuracy* metric ($h_{acc}(t) = \frac{p+N-n}{P+N}$) in the first iteration, so that in subsequent iterations always the best "local" condition can be picked, as discussed in Section 3.

Second, the handling of multiple pattern trees is crucial for the decision boundary. In the original ABPT classifier, one pattern tree for each class $y \in \mathbb{Y}$ is learned. Since in a Boolean context, the output of the Boolean expression represented by the pattern tree can only be true or false for the features of the test example, ties occur if a test example is matched by multiple pattern trees, which can be broken by a fixed order of the pattern trees in a decision list.

An alternative that is used in fuzzy pattern tree classifiers [17] is evaluating all pattern trees in a probabilistic way, whereby the highest probability decides about the class prediction. A straightforward way to achieve this behavior in ABPT is using a constant uncertainty factor u, resulting in probabilities $p(f) = 1 - u$ for fulfilled features and $p(f) = u$ else, which are then aggregated bottom-up over the respective child nodes C as $p(n) = \prod_{i \in C} p(i)$ for conjunctive and $p(n) = 1 - \prod_{i \in C}(1 - p(i))$ for disjunctive nodes. However, this comes with two inconveniences of (a) weighing all child nodes the same — independent of their importance and (b) penalizing all conjunctive conditions (always decreasing $p(n)$) and rewarding all disjunctive conditions (always increasing $p(n)$) independent of their quality, which particularly affects characteristic models negatively. To address (a), we determine $p(f)$ flexibly in the range $[u, 1-u]$ as

$$p(f) = u + (1 - 2u) \cdot \frac{p}{P}$$

for fulfilled features ($\simeq$ probability a positive example fulfills the feature) and

$$p(f) = u + (1 - 2u) \cdot \frac{P - p}{P - p + N - n}$$

else ($\simeq$ probability a positive example not fulfilling the feature is negative). Additionally, for (b) we relax $p(n)$ for the interior tree nodes as

$$p(n) = \frac{1}{2} \cdot \left(\frac{1}{|C|} \cdot \sum_{i \in C} p(i) + \min_{i \in C} p(i)\right)$$

for conjunctive nodes and as

$$p(n) = \frac{1}{2} \cdot \left(\frac{1}{|C|} \cdot \sum_{i \in C} p(i) + \max_{i \in C} p(i)\right)$$

for disjunctive nodes, resulting in a probability less dependent on the number of child nodes, so that models with different numbers of conjunctive and disjunctive nodes can be compared better.

Table 2
Predictive accuracies of the ABPT learner on five UCI datasets for six different settings using 10-fold-cross-validation. In the first row a Boolean evaluation and in the second row a probabilistic evaluation of the pattern tree is used. The first column shows results on the original dataset, the second on an incomplete version of the dataset where 30% of the values are replaced by missing values, and the third a combination using the original data for training and the incomplete data for testing.

	discr.	char.
labor	87.72	84.21
mushroom	100.00	100.00
soybean	92.68	92.53
vote	94.48	94.71
zoo	89.11	86.14

(a) Original + Boolean

	discr.	char.
labor	77.19	78.95
mushroom	96.91	96.91
soybean	66.03	66.33
vote	88.28	88.74
zoo	76.24	76.24

(b) Incomplete + Boolean

	discr.	char.
labor	78.95	80.70
mushroom	84.00	88.00
soybean	19.77	46.71
vote	78.39	78.16
zoo	41.58	65.35

(c) Mixed + Boolean

	discr.	char.
labor	66.67	64.91
mushroom	83.83	87.64
soybean	90.04	90.48
vote	95.40	95.40
zoo	89.11	92.08

(d) Original + Probabilistic

	discr.	char.
labor	64.91	64.91
mushroom	49.93	49.93
soybean	67.64	68.52
vote	88.51	88.51
zoo	77.23	75.25

(e) Incomplete + Probabilistic

	discr.	char.
labor	64.91	64.91
mushroom	79.60	75.49
soybean	52.86	57.25
vote	87.13	87.13
zoo	62.38	85.15

(f) Mixed + Probabilistic

5. Experiments

In the experiments we analyze two different aspects of the ABPT learner — using the default configuration of $k = 20$ iterations and accuracy and seven different values for the linear cost as metrics. First, we compare a discriminative version preferring smaller trees in case of tied heuristics and a characteristic version preferring bigger trees. Second, we evaluate both versions not only in a Boolean setting but also in a probabilistic setting, as suggested in the end of Section 4.

For the experiments, we choose five UCI [18] datasets where most features are not used in the discriminative models and therefore the characteristic models could differ remarkably: LABOR, MUSHROOM SOYBEAN, VOTE and ZOO. On some datasets (LABOR, SOYBEAN, ZOO) the otherwise inferior naive Bayes classifier even outperforms rule learners like RIPPER, indicating potential for improvement of the decision boundary in the probabilistic setting.

The learners are not only applied to the original datasets but also to an "imcomplete" version of the dataset, where 30% of the values are replaced by missing values, and finally a "mixed" version, where only the test data is "incomplete". This way we can analyze the robustness of the learned models, where characteristic models might be expected to perform better, since additional features are used as a fallback option.

The predictive accuracies of a 10-fold-cross-validation are shown in Table 2. Each table shows a head-to-head comparison of the discriminative and characteristic learner in a given setting of dataset type and used evaluation. Overall, we see that the discriminative and characteristic learner perform roughly equally well, while the effects of missing values vary considerably between the datasets. Except few cases, the harder the setting (from left to right), the lower the predictive accuracy.

Independent of the dataset setting, the predictive accuracy drops drastically for LABOR and MUSHROOM when changing from the Boolean to the probabilistic setting. Though, it often increases for the other three datasets — in particular, in the "mixed" setting, the accuracies can be improved drastically using a probabilistic evaluation, indicating that the adjusted decision boundaries can make the model more robust to incomplete training data.

```
c1 ← eggs=false.
c1 ← hair=true.
c2 ← toothed=true.
c2 ← catsize=true.
c2 ← legs=(1.0:4.0].
c3 ← backbone=true.
c3 ← airborne=false.
c3 ← aquatic=false.
c3 ← fins=false.
c3 ← tail=true.
c3 ← domestic=false.
c3 ← predator=false.
c3 ← predator=true.
c3 ← domestic=true.
c3 ← tail=false.
c3 ← catsize=false.
c3 ← fins=true.
c ← milk=true ∧ c1 ∧ c2 ∧ breathes=true∧
    feathers=false ∧ c3.
```

Figure 3: Model learned by characteristic ABPT on the ZOO dataset for $class = mammal$ when being transformed into a set of conjunctive rules.

As an example, consider the characteristic model for the zoo dataset in Figure 3. In Boolean evaluation, the missing values result in too many examples being not covered by the model, However, the numerous conditions still indicate the correct class with a probabilistic evaluation. We also notice a big gap between the performances of the discriminative and characteristic model here, indicating that the small trees with only up to four conditions of the discriminative learner (e.g., for `class=mammal` it only uses the condition `mlik=true`) are not as robust as the trees of the characteristic counterpart.

As Figure 3 also shows, the characteristic models are usually considerably larger. From an interpretability perspective, this can be preferred, since in this case we do not only discover that mammals yield milk but also breathe, do not wear feathers, either do not lay eggs or have hair and either are toothed, catsized or have 1-4 legs. We also notice that the last concept `c3` (which was also the last to be added to the model) is not helpful at all and indeed can be reduced to `true` because of tautologies. This indicates that in a characteristic setting stopping criteria or pruning techniques are needed as well to preserve interpretability.

6. Conclusion

In this paper, we look at the possible advantages that characteristic models, which are rarely learned in conventional rule learning algorithms, can provide. While previous work on characteristic rules usually focused on the interpretability aspects, we have shown that the inclusion of additional features, both via conjunction and disjunction, can additionally help to find better decision boundaries, resulting in more robust models. We also discussed that for learning characteristic rules a mere focus on coverage is insufficient so that both regular and inverted heuristics can not guarantee learning characteristic rules. Finding a suitable distance metric to separate positive and negative examples remains as an open question.

To analyze the effects of characteristic rule-based models empirically, we implemented a characteristic version of the ABPT learner and compared it with the original discriminative version on five UCI datasets. The experiments did not show a clear advantage for any of the learners in terms of predictive accuracy, indicating that smaller models are not necessarily overgeneralizing and larger models not inevitably lead to overfitting. In a robustness check using incomplete test data, characteristic models outperformed discriminative models. Furthermore, characteristic models slightly outperformed discriminative models when combined with probabilistic evaluation.

We also see multiple paths to further develop the potential of characteristic models in future work: Most importantly, the decision boundary artificially moved by a probabilistic evaluation (which certainly provides room for improvement as well) should not only be considered during classification but also in the learning phase. In this regard, the definition of heuristics not only considering coverage but also the "quality" of the coverage, connected to the distance between the example and the decision boundary, is crucial. This way rule learners could not only deliver a prediction but also determine how certain the prediction is, and optionally abstain from making a prediction.

References

[1] R. Andrews, J. Diederich, A. B. Tickle, Survey and critique of techniques for extracting rules from trained artificial neural networks, Knowledge-based systems 8 (1995) 373–389.

[2] R. Guidotti, A. Monreale, F. Giannotti, D. Pedreschi, S. Ruggieri, F. Turini, Factual and counterfactual explanations for black box decision making, IEEE Intelligent Systems 34 (2019) 14–23.

[3] A. Blumer, A. Ehrenfeucht, D. Haussler, M. K. Warmuth, Occam's razor, Information processing letters 24 (1987) 377–380.

[4] R. S. Michalski, A theory and methodology of inductive learning, in: Machine learning, Elsevier, 1983, pp. 83–134.

[5] J. Fürnkranz, Pruning algorithms for rule learning, Machine learning 27 (1997) 139–172.

[6] W. W. Cohen, Fast effective rule induction, in: Machine learning proceedings 1995, Elsevier, 1995, pp. 115–123.

[7] M. Eineborg, H. Boström, Classifying uncovered examples by rule stretching, in: C. Rouveirol, M. Sebag (Eds.), Proceedings of the Eleventh International Conference on Inductive Logic Programming (ILP-01), Springer Verlag, Strasbourg, France, 2001, pp. 41–50.

[8] S. Salzberg, A nearest hyperrectangle learning method, Machine Learning 6 (1991) 251–276.

[9] N. Cristianini, J. Shawe-Taylor, An introduction to support vector machines and other kernel-based learning methods, Cambridge university press, 2000.

[10] J. S. Cramer, The origins of logistic regression (2002).

[11] I. Rish, et al., An empirical study of the naive bayes classifier, in: IJCAI 2001 workshop on empirical methods in artificial intelligence, volume 3, Seattle, WA, USA;, 2001, pp. 41–46.

[12] C. Rudin, B. Ustun, Optimized scoring systems:

Toward trust in machine learning for healthcare and criminal justice, Interfaces 48 (2018) 449–466.
[13] S. Salzberg, A nearest hyperrectangle learning method, Machine learning 6 (1991) 251–276.
[14] J. Fürnkranz, P. A. Flach, Roc'n'rule learning—towards a better understanding of covering algorithms, Machine learning 58 (2005) 39–77.
[15] J. Stecher, F. Janssen, J. Fürnkranz, Shorter rules are better, aren't they?, in: Discovery Science: 19th International Conference, DS 2016, Bari, Italy, October 19–21, 2016, Proceedings 19, Springer, 2016, pp. 279–294.
[16] F. Beck, J. Fürnkranz, V. Q. P. Huynh, Learning deep rule concepts as alternating boolean pattern trees, in: Discovery Science: 27th International Conference, DS 2024, Pisa, Italy, October 14–16, 2024, Proceedings 27, Springer, 2024.
[17] R. Senge, E. Hüllermeier, Top-down induction of fuzzy pattern trees, IEEE Transactions on Fuzzy Systems 19 (2010) 241–252.
[18] M. Kelly, R. Longjohn, K. Nottingham, The UCI machine learning repository, 2024. URL: https://archive.ics.uci.edu.