Information Sciences xxx (xxxx) xxx

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Interpretable multiclass classification by MDL-based rule lists

Hugo M. Proença*, Matthijs van Leeuwen

LIACS, Leiden University, the Netherlands

ARTICLE INFO

Article history:
Received 9 May 2019
Revised 10 August 2019
Accepted 25 October 2019
Available online xxx

Keywords:
Rule lists
Minimum Description Length principle
Interpretable models
Classification

ABSTRACT

Interpretable classifiers have recently witnessed an increase in attention from the data mining community because they are inherently easier to understand and explain than their more complex counterparts. Examples of interpretable classification models include decision trees, rule sets, and rule lists. Learning such models often involves optimizing hyperparameters, which typically requires substantial amounts of data and may result in relatively large models.

In this paper, we consider the problem of learning compact yet accurate probabilistic rule lists for multiclass classification. Specifically, we propose a novel formalization based on probabilistic rule lists and the minimum description length (MDL) principle. This results in virtually parameter-free model selection that naturally allows to trade-off model complexity with goodness of fit, by which overfitting and the need for hyperparameter tuning are effectively avoided. Finally, we introduce the Classy algorithm, which greedily finds rule lists according to the proposed criterion.

We empirically demonstrate that Classy selects small probabilistic rule lists that outperform state-of-the-art classifiers when it comes to the combination of predictive performance and interpretability. We show that Classy is insensitive to its only parameter, i.e., the candidate set, and that compression on the training set correlates with classification performance, validating our MDL-based selection criterion.

1. Introduction

Interpretable machine learning has recently witnessed a strong increase in attention [12], both within and outside the scientific community, driven by the increased use of machine learning in industry and society. This is especially true for applications domains where decision making is crucial and requires transparency, such as in health care [27,30] and societal problems [26,47].

While it is of interest to investigate how existing 'black-box' machine learning models can be made transparent [38], the trend towards interpretability also offers opportunities for data mining, or *Knowledge Discovery from Data* (KDD), as this field traditionally has a stronger emphasis on intelligibility.

In recent years several interpretable approaches have been proposed for supervised learning tasks, such as classification and regression. Those include approaches based on prototype vector machines [35], generalized additive models [32], decisions sets [25,44], and rule lists [30,46]. Restricting our focus to classification, we make two important observations. First,

* Corresponding author.
E-mail addresses: h.manuel.proenca@liacs.leidenuniv.nl (H.M. Proença), m.van.leeuwen@liacs.leidenuniv.nl (M. van Leeuwen).

https://doi.org/10.1016/j.ins.2019.10.050
0020-0255/

Rule		antecedent		consequent	usage
1	IF	$\{backbone = no\}$	THEN	$\Pr(invertebr.) = 0.55$	10
				$\Pr(bug) = 0.45$	8
2	ELSE IF	$\{breathes = no\}$	THEN	$\Pr(fish) = 0.93$	13
				$\Pr(reptile) = 0.07$	1
3	ELSE IF	$\{feathers = yes\}$	THEN	$\Pr(bird) = 1.00$	20
4	ELSE IF	$\{milk = no\}$	THEN	$\Pr(reptile) = 0.50$	4
				$\Pr(amphibian) = 0.50$	4
∅	ELSE	ELSE	THEN	$\Pr(mammal) = 1.00$	41

Fig. 1. Example of a Probabilistic Rule List (PRL) obtained by CLASSY on the *zoo* dataset, without the need for any parameter tuning. Test accuracy: 87%. The dataset contains 7 classes, 101 examples, and 35 binary variables. Usage refers to the number of examples covered by a certain rule and class label. Note that we did not apply Laplace smoothing here for clarity of presentation; Eq. (5) defines the actual probability estimates.

we observe that state-of-the-art algorithms [3,25,30,44,46] are designed for binary classification; no interpretable methods specifically aimed at multiclass classification have been proposed, in spite of being a common scenario in practice. Multiclass classification is more challenging because of 1) the increased complexity in model search, due to the uncertain consequences of favouring one class over the others, and 2) the lack of possibilities to prune the search such as commonly used when finding, e.g., decision lists [3] or Bayesian rule lists [46] for binary classification. Our second observation is that although recent methods based on rules [30,46] and decision sets [25,44] have been shown to be effective, they tend to have 1) a fair number of hyperparameters that need to be fine-tuned, and 2) limited scalability. Especially the need for hyperparameter tuning can be problematic in practice, as it requires significant amounts of computation power and data (i.e., not all data can be used for training, as a substantial part has to be reserved for validation).

To address these shortcomings, *we introduce a novel approach to finding interpretable, probabilistic multiclass classifiers that requires very few hyperparameters and results in compact yet accurate classifiers*. In particular, we will show that our method naturally provides a desirable trade-off between model complexity and classification performance without the need for parameter tuning, which makes the application of our approach very straightforward and the resulting models both adequate classifiers and easy to interpret.

We will use probabilistic rule lists, as both the antecedent of a rule (i.e., a *pattern*) and its consequent (i.e., a probability distribution) are interpretable [30]. Using a probabilistic model has the additional advantage that one cannot only provide a crisp prediction, but also make a statement about the (un)certainty of that prediction.

We show that, given a set of ordered patterns, we can trivially estimate the corresponding consequent probability distributions from the data. The remaining question, then, is how to select a set of patterns that together define a probabilistic rule list that is accurate yet does not overfit. This is not only important to ensure generalizability beyond the observed data, but also to keep the models as compact as possible: larger models are harder to interpret by a human analyst [21]. Recent optimization [25] and Bayesian [46] approaches heavily rely on hyperparameters to achieve this, but those need to be tuned by the analyst and we specifically aim to avoid this.

The solution that we propose is based on the minimum description length (MDL) principle [18,40], which has been successfully used to select small sets of patterns that summarize the data in the context of exploratory data mining [9,29,42]. The MDL principle can be paraphrased as "*induction by compression*" and roughly states that the best model is the one that best compresses the data. Advantages of the MDL principle include that it has solid theoretical foundations, avoids the need for hyperparameters, and automatically protects against overfitting by balancing model complexity with goodness of fit.

Our first main contribution is the formalization of the problem of selecting the optimal probabilistic rule list using the minimum description length principle. Although the MDL principle has been used for pattern-based classification before [42], we are the first to introduce a MDL-based problem formulation aimed at selecting rule lists for multiclass classification. Technically, our approach includes the use of the prequential plug-in code, a form of *refined MDL* that has only been used once in pattern-based modelling [9]. One advantage of our approach is that the resulting problem formulation is completely parameter-free.

Our second main contribution is CLASSY *, a heuristic algorithm for finding good probabilistic rule lists.* Inspired by the KRIMP algorithm [42], we select a good set of rules from a set of candidate patterns. We empirically demonstrate, by means of a variety of experiments, that CLASSY outperforms RIPPER, C5.0, CART, and Scalable Bayesian Rule Lists (SBRL) [46] when it comes to the combination of classification performance and interpretability, in particular when taking into account that it has much fewer hyperparameters.

Please cite this article as: H.M. Proença and M. van Leeuwen, Interpretable multiclass classification by MDL-based rule lists, Information Sciences, https://doi.org/10.1016/j.ins.2019.10.050