Robust rules for prediction and description

Manuel Proenca, H.

6 Conclusions

As machine learning and data mining permeate everyday life, the questions sought should be as much about algorithms as they should be about society itself. Algorithms increasingly affect the lives of individuals everywhere; thus, the pertinent questions are not only purely algorithmic but also about how they can help society solve systematic issues such as discrimination and social inequality.

For this reason, in recent years, we can find research on both classic topics such as algorithmic efficiency and statistical guarantees, and on newer issues such as privacy, fairness, and accountability. It is at the intersection of several of these topics, namely, algorithmic efficiency, statistical guarantees, and accountability, that we pose our main research question: "*How to learn robust and interpretable rule-based models from data for machine learning and data mining, and define their optimality?*".

In an honest attempt to answer it, we selected rule lists as models and the MDL principle as model selection theory. The former confers interpretability by design as humans can easily understand rule lists. At the same time, the latter allows for an objective formulation of learning rule lists from data that combines the performance and complexity of the model in one. Together, these allowed us to propose efficient algorithms that approximate our optimal formulation and achieve state-of-the-art performance while finding simpler models.

Nonetheless, this dissertation is just one step further in answering this research question. Its main limitation is our assumption that interpretability is associated with simplicity, but this is not always the case in reality. Interpretability is subjective, and it depends on the human that will act on or be acted upon by the model. Although

our MDL-based formulation attempts to have the least amount of assumptions, in some cases, it is necessary to add extra input from the human user to guarantee true interpretability.
Moreover, in Section 6.1 we present an overview of the main conclusions by chapter. Then, in Section 6.2 we discuss the strong and weak points of our proposal. Finally, in Section 6.3 we show possible directions of future work.

6.1 Summary

Chapter 1 introduced the scientific background and motivation for this dissertation.

Chapter 2 introduced the necessary mathematical background. In particular, it formally presented the tasks of rule-based prediction, subgroup discovery, and subgroup set discovery. Then, association rules, the standard component of these tasks, is promptly defined. Based on the previous definitions, we presented rule lists, i.e., the model class made of an ordered set of association rules. Furthermore, we distinguish predictive rule lists for machine learning and subgroup lists for data mining. Finally, it shows how to measure model quality in the classification and subgroup discovery setting.

Chapter 3 proposed an optimal formulation of predictive rule lists and subgroup lists for univariate and multivariate, nominal, and numeric target variables based on the Minimum Description Length (MDL) principle. Three new optimal data encodings for models that partition the data—rule lists, trees, clusters, etc.—are presented. In specific, these codes are: 1) the prequential plug-in code for nominal variables; 2) the Normalize Maximum Likelihood (NML) code for nominal variables; and 3) an objective Bayesian code with improper priors for numeric variables. We show that MDL-based subgroup lists with one subgroup are equivalent to top-1 subgroup discovery with weighted Kullback-Leibler divergence as a quality measure, thus validating subgroup lists as a valid generalization of subgroup discovery. Moreover, the best subgroup to add according to the MDL criteria maximizes an MDL equivalent to a Bayesian proportion, multinomial, or t-test plus a multiple hypothesis testing. In the end, we show the difference between predictive rules and subgroups through our MDL formulation of both problems.

Chapter 4 proposed CLASSY, a heuristic algorithm based on the MDL formulation of predictive rule lists for multiclass classification. Experiments show that it finds good predictive models that are also compact without hyperparameter tuning. CLASSY is composed of a frequent pattern mining algorithm to pre-mine all candidate rules and then iteratively adds one rule at a time to the rule list. It effectively only has one hyperparameter, the pre-mined set of candidate rules. If this set is made large enough to accommodate all possible rules in the data, it can find good models independently of any hyperparameters—at the expense of computational budget. The empirical tests show state-of-the-art performance on classification, interpretability, and overfitting.

Chapter 5 proposed the *Robust Subgroup Discoverer* (RSD), a heuristic algorithm based on our MDL formulation that finds good subgroup lists for univariate and multivariate nominal and numeric target variables. Experiments over 54 datasets show that it outperforms state-of-the-art subgroup set discovery algorithms regarding the quality of sets found, especially for numeric targets. The algorithm iteratively uses a beam search to find candidates and then adds the one that locally minimizes the MDL optimal formulation. This approximation is equivalent to a Bayesian test (factor) between subgroup and dataset marginal target distributions plus a penalty for multiple hypothesis testing. Thus, we guarantee the statistical robustness of each subgroup in the list.

Chapters 4 and 5. The algorithms of both chapters share the greedy adding of rules, although CLASSY pre-mines all association rules, and RSD uses beam search at each iteration. The algorithms can be interchanged for both tasks in practice, although given their historical development, they do not overlap. Nonetheless, CLASSY reflects the intention of having few hyperparameters that we aimed for classification, and RSD demonstrates the flexibility necessary for data exploration, making them appropriate for their respective chapters.

6.2 Discussion

In this section we discuss the advantages and disadvantages of our proposals. To make this section consistent with the previous, we organize the discussion per chapter that proposes new work, i.e., Chapters 2, 3, 4, and 5.

Chapter 2. The only new proposal of this chapter is the *subgroup list* model class. Sequential subgroup set discovery was always defined heuristically, and each subgroup was interpreted individually without considering the previously found ones. We propose the first *global* dataset formulation of the problem of subgroup set selection that is equivalent to top-1 subgroup discovery in the case of a subgroup list with only one subgroup, and that also fits some of the previous heuristic definitions of sequential selection. Its main limitation is that subgroups far down in the list are hard to interpret for large lists as they require considering previous ones. Also, this is not the only possible generalization of subgroup discovery to sets.

Chapter 3. In this chapter, we presented the MDL formulation of predictive rule lists and subgroup lists.

In the case of predictive rule lists, it allows using a single measure—the MDL score—to measure the bias-variance trade-off, one of the core problems in learning models from data. Even though predictive rule lists have long been used in machine learning, they have solely focused on classification and mostly on univariate target classification. We have proposed a theoretical formulation for classification and multi-target classification and regression. Compared with its Bayesian counterpart for rule lists for classification, our MDL formulation tries to make fewer assumptions, which we believe makes it more robust against overfitting. The main limitation of our formulation is that it assumes that a parsimonious model is interpretable, which is not always the case [26].

In the case of subgroup lists, it formulates a *global* perspective of sequential subgroup discovery. It generalizes the original problem of subgroup discovery to lists and gives it a balance between the complexity of the list and the quality of the descriptions.

Chapter 4. CLASSY is a heuristic algorithm with very few hyperameters that is competitive against state-of-the-art algorithms. It finds models with similar classification performance that are more compact and overfit less. It can also be made independent of its hyperparameters at a computational expense. The main drawback of our approach is that it is limited to binary input variables and single-target multiclass problems. Also, compared with RSD of Chapter 5 it does not provide local statistical guarantees for each of the added rules except that it improves the global score.

Chapter 5. RSD is a heurisitic algorithm that can find subgroup lists for univariate and multivariate nominal and numeric targets. Contrary to CLASSY it can deal with both nominal and numeric input variables. In the case of numeric targets, it uses a dispersion-aware measure to find subgroups with smaller standard deviations in the target values. Its normalization hyperparameter can change the granularity of the search from very specific to more general subgroups. One of its disadvantages is that we do not know how close we are to the global optimum. Also, when the normalization is used to the maximum (normalized gain), the algorithm is susceptible to noise in the data, i.e., it would find a different model if small variations are added to the data.

6.3 Future Work

This dissertation focuses on predictive rule lists for machine learning and subgroup lists for subgroup discovery based on the MDL principle. We decided to divide future work into technical developments that can be achieved soon —short and medium-term research—and the vision of the role rule-based models can take in machine learning and data mining— long-term research.

6.3.1 Short and medium-term research

Given the main topics of this dissertation, we will divide the technical advances into five different lines of research: 1) the MDL formulation of rule lists; 2) predictive rule lists; 3) subgroup lists; 4) search algorithms, and 5) rule sets.

1) The MDL formulation of rule lists can be extended to different types of target variables or distributions. First, it is straightforward to combine nominal and numeric targets through independent categorical and normal distributions using the MDL principle. It gives us an objective measure of both in bits. Then, instead of assuming independence between target variables, one can accommodate dependencies using multivariate numeric distributions. Finally, other types of distributions that can be more appropriate in different scenarios can be used, such as a Poisson distribution.

2) Predictive rule lists were only empirically tested for multiclass classification. Chapter 3 already defines the optimal predictive rule list for regression and multi-target classification and regression; thus, only the algorithm would need to be extended to these cases.

3) Subgroup lists, and similarly to the MDL future work, could accommodate non-independent distributions for multivariate targets and propose extensions to RSD find them.

4) Search algorithms. At the moment, only a greedy separate-and-conquer search is proposed. It would be essential to test the feasibility of optimal search algorithms such as branch-and-bound or Markov Chain Monte Carlo (MCMC).

5) Rule sets. In this dissertation, we only study rule lists (ordered rule sets). Extending the MDL theory and algorithms to overlapping rule sets would be a considerable development.

6.3.2 Long-term research

In terms of long-term research, the main directions we envision are related to better approximations of the real problems with fewer assumptions about the ideal behavior of the data. We will divide the topics into four groups: 1) interpretability; 2) rule-based models for sequential data; 3) rule-based models for image data; and 4) causal analysis.

1) Interpretability. As mentioned before, interpretability is subjective, and one cannot expect to have one universal formulation that works for everyone. Also, every person has a unique background that will make, e.g., some variables in the data easier to understand than others. For this reason, it is necessary to insert the human in the learning loop by having an algorithm that takes into consideration both objective concepts and the subjective nature of each individual. A path towards this end would be to start with an MDL formulation of a problem similar to ours, representing a *tabula rasa* or the minimum level of assumptions possible. Then, build upon the subjective characteristics of the user. This last part is crucial, and there are many options for it. It can be done at the beginning in the form of something similar to Bayesian priors or iteratively by presenting the user with a model and querying her about what they prefer (or not). At no point should the model overfit the data, and for that, the tabula rasa represents a baseline of the best formulation with minimum assumptions.

2) Rule-based models for sequential data. Even though there is already research on this direction, it tends to be composed of heuristics that lack statistical robustness. It would be interesting to conjugate the MDL principle with rule-based models for sequential data that formally consider dynamic learning and concept drift.

3) Rule-based models for image data. Rule-based models are shallow learners because they take the input variables as they come and do not transform them into more complex features. For this reason, to make rule-based models appropriate for image analysis, it is necessary that they either make their transformations or that they couple with other tools, such as classic computer vision techniques or neural networks that return human-understandable macro structures. A specific case of interest would be to use an image segmentation tool coupled with subgroup discovery to identify regions in the data that stand out with respect to a particular target, e.g., to describe areas in satellite image data with more pollution than the average.

4) Causal analysis. In supervised learning, it is usually assumed that any variable present in the data can be used to predict or describe the target variable. However, not all variables have the same relationship with the target, as some can cause the target, be caused by it, or be independent. Also, input variables can be caused by other input variables. If one pays attention to the causal relationships when learning predictive models, it allows one to find robust models that generalize to distributions different from the training data. Also, considering the causal relationships in the data allows asking questions about counterfactuals or “What if something would have happened differently than we see on the data?”.