1 Introduction

Rules are an essential part of what makes us humans. They are prime methods of information storage and sharing, employed every day to assimilate complex ideas into more manageable chunks of information. Their use can be found everywhere, from the mental note “if I do not put the alarm, then I will not wake up” to the intricate system of clinical diagnosis rules employed by physicians. A simple example is a clinical rule for diagnosing the flu, given by “if a person has either a fever or sore throat (between others), then she has the flu.”.

A rule does not need to be always correct, but in most cases, it should; otherwise, it would not contain the essential information about the problem. Nonetheless, their most compelling property is that they are easy to understand, i.e., interpretable. Thus, it should not come as a surprise that researchers have long used them to describe the world, be it in machine learning or data mining.

While machine learning is concerned with finding a representation from data that can predict current and future events, data mining is concerned with extracting interesting information from data. Even though both tasks are intertwined and rule utilization is extensive in both fields, they stem from different intentions; hence, they arrive at different outcomes. In machine learning, the combination of several rules forms a model that makes predictions about future events. In data mining, sets of rules describe patterns in data that are worth seeing.

In the wake of deep learning successes, one can question if rule-based models still have a place; after all, deep learning models seem to make better predictions. Nevertheless, the complex nature of deep learning from which its success stems is also its main limitation as its models are inscrutable and not accountable. For this reason, it is imperative to find algorithms that can learn high-quality rule-based models from data that are competitive in prediction while at the same time interpretable.

Even though research on rule-based models started more than half a century ago, many questions remain to be answered, such as: What is an optimal set of rules? What is the relationship between rules in data mining and machine learning? Can we guarantee that the models are statistically robust before seeing future data?
To answer such questions, we apply the Minimum Description Length (MDL) principle to rule-based models, which objectively quantifies the quality of models and guarantees statistical robustness. Based on information theory, the principle states that the best model is the simplest that describes the data well. This idea is a formal restatement of Occam's Razor, the law of parsimony that directly relates to the notion that a good rule, but now a set of rules, should only describe what matters most in the data for a particular task.

More specifically, we focus on ordered rule sets, i.e., rule lists. These are the first rule-based model invented, and—compared to their unordered counterparts—they have appealing mathematical properties that allow for a suitable formulation according to the MDL principle. This dissertation establishes a better understanding of rules and rule lists in machine learning and data mining. To distinguish between both, rule lists are called predictive rule lists in machine learning and subgroup lists in data mining. Our focus in machine learning is on supervised learning and in data mining on subgroup discovery. For the less acquainted with the last topic, subgroup discovery is the task of finding descriptions of data subsets—rules in tabular data—that deviate from "normal behaviour" for a target variable. In both cases, the MDL principle formalizes their optimality for a given dataset.

Motivation. The research conducted in this dissertation was in part motivated by the real-world problem of flight delays, and in specific by the SAPPAO (a Systems APproach towards data mining and Prediction in Airlines Operations) project. Its objective was to integrate flight delay predictions in optimizing airplane and crew schedules to reduce fuel and crew costs and decrease unnecessary CO2 emissions. In our part of the project, we focused on the characterization of subgroups of flights with above-average delays. We show an example of utilizing our theory and algorithms in publicly available datasets in Section 5.5.

1.1 Predictive rule lists

Interpretable machine learning has recently witnessed a strong increase in attention [26], 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 [81, 68] and societal problems [67, 126].

While it is of interest to investigate how existing 'black-box' machine learning models can be made transparent [104], 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 [95], generalized additive models [84], decisions sets [69, 122], and predictive rule lists [81, 125]. Restricting our focus to classification, we make two important observations. First, we observe that state-of-the-art algorithms [69, 122, 81, 125, 5] are designed for binary classification; no interpretable methods specifically aimed at multiclass classification have been proposed, despite 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 [5] or Bayesian rule lists [125] for binary classification. Our second observation is that although current methods based on rules [81, 125] and decision sets [69, 122] are 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 hyperparameter tuning, which makes the application of our approach very straightforward and the resulting models both adequate classifiers and easy to interpret.

We use probabilistic rule lists, as both the antecedent of a rule (i.e., a *pattern*) and its consequent (i.e., a probability distribution) is interpretable [81]. 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. Note 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 form an interpretable rule list.

Interpretable rule list discovery. Informally, the problem of finding interpretable rule lists for prediction is: how to select a *compact* set of rules that together define a predictive rule list that is *accurate* yet it does not *overfit*. Overfitting is not only important to ensure generalizability beyond the observed data, but it also aligns with keeping the models as compact as possible: larger models are harder to interpret by a human analyst [56] and more prone to overfit. Another layer of interpretability that we consider is that the algorithm used to find these rule lists does not have many hyperparameters, and thus does not require much human intervention to obtain good and reliable models.

Recent optimization [69] and Bayesian [125] approaches to obtain interpretable rule lists for classification heavily rely on hyperparameters to achieve this, but those need to be tuned by the analyst and we specifically aim to avoid this.
To accomplish this, the solution that we propose is based on the MDL principle [107, 48].

1.2 Subgroup lists

Exploratory Data Analysis (EDA) [118] aims at enhancing its practitioner's natural ability to recognize patterns in the data being studied. The more she explores the more she discovers, but also the higher the risk of finding interesting results arising out of coincidences, as, e.g., spurious relations between variables that have no connection in the real world. Intuitively this corresponds to testing multiple hypothesis without realizing it. This duality of EDA requires a thorough analysis of results and highlights the need for statistically robust techniques that allow us to explore the data in a responsible way. While EDA encompasses all techniques referring to data exploration, *Subgroup Discovery* (SD) [63, 8] is the subfield concerned with discovering interpretable descriptions of subsets of the data that stand out with respect to a given target variable, i.e., *subgroups*. In this dissertation, we aim at improving the discovery of subgroup lists, i.e., ordered sets of subsets, that describe different regions of the data while being statistically robust at an individual level and as a whole.

Subgroup discovery (SD) can be seen as the exploratory counterpart to rule learning or association rule mining, where the targets/consequent of the rules are fixed, and rules are ranked according to quality measures combining subgroup size and deviation of the target variable(s) with respect to the overall distribution in the data. In its traditional form, subgroup discovery is also referred to as top-k subgroup mining [8], which entails mining the k top-ranking subgroups according to a *local* quality measure and a number k selected by the user. Since its conception, subgroup discovery has been developed for various types of data and targets, e.g., nominal, numeric [45], and multi-label [72] targets. SD has been applied in a wide range of different domains [52, 8], such as identifying the properties of materials [43], unusual consumption patterns in smart grids [60], identifying the characteristics of delayed flights [98], and understanding the influence of pace in long-distance running [23].

Even though SD appeals to several domains, top-k mining traditionally suffers from three main issues that make it impractical for many applications: $1)$ poor efficiency of exhaustive search for more relevant quality measures [12]; $2)$ *redundancy* of mined subgroups, i.e., the fact that subsets with the highest deviation according to a certain *local* quality measure tend to cover the same region of the dataset with slight variations in their description of the subset [75]; $3)$ lack of generalization or statistical robustness of mined subgroups [77]. In this dissertation, we focus on the last two issues together: lowering *redundancy* by finding small lists of subgroups that describe the differences in the data well; and obtaining *statistically robust* subgroups. First, we define what an optimal subgroup list is using the MDL principle. Second, we propose a greedy algorithm that finds good subgroup lists using a *local* objective that is equivalent to maximizing Bayesian one-sample proportions, multinomial or t-test between each subgroup's distribution and the dataset marginal distribution, for binary, nominal or numeric data, respectively, plus a penalty for multiple hypothesis testing.

In recent years both issues have been partially addressed, mostly independent of each other; we next briefly discuss recent advances and limitations.

In terms of *redundancy*, the first main limitation of existing works is their focus on one type of target variables, such as binary targets [14, 10], nominal targets [71], or numeric targets [83], where only DSSD focuses on univariate and multivariate nominal and numeric targets [75]. The second main limitation is the lack of an optimality criterion for subgroup sets or lists, where the only exception is FSSD [10]. It is important to emphasize that some works aim to find sequential subgroups or subgroup *lists*, while others aim to find unordered sets or subgroup *sets*. Subgroup lists are akin to predictive rule lists [96] in the sense that each subgroup needs to be interpreted sequentially and they are not allowed to overlap, while subgroup sets are allowed to overlap. In this chapter, we focus solely on subgroup lists, and although previous works often did not use this term, we retroactively rename those models that are in

fact subgroup lists.
In terms of *statistical robustness*, most existing approaches consider first mining the top-k subgroups and then post-processing them in terms of a statistical test to find if the discovered subgroups are statistically significant [30, 77].

Robust subgroup discovery. Informally the problem of robust subgroup discovery is to define and find the *globally* optimal set or list (i.e., an ordered set) of non-redundant subgroups that together explain the most relevant *local* deviations in the data with respect to specified target variables. As finding the optimal set or list will typically be practically infeasible, the secondary problem is to construct an algorithm that efficiently mines "good" subgroup sets or lists from the data that retains as much from the *global* formulation's statistical properties as possible.

In this dissertation we restrict our focus to finding *subgroup lists*, because 1) they were one of the first model classes proposed for subgroup set discovery [71]; 2) they allow for an optimal formulation based on the MDL principle due to its property of unambiguously partitioning the data into non-overlapping parts; and 3) finally, they allow an ordered interpretation of the subgroups, i.e., from most to least relevant discovered subgroup.

1.3 Research question and contributions

This dissertation attempts to answer one overarching research question:

How to learn robust and interpretable rule-based models from data for machine learning and data mining, and define their optimality

In pursuit of valid answers to this question, this dissertation presents contributions on five topics: 1) predictive rule lists; 2) subgroup lists; 3) MDL learning theory; and 4) the difference between predictive rules and subgroup discovery rules.

Our contributions on **predictive rule lists** and **machine learning** are the following:

1. *Interpretable predictive rule lists using MDL* (Chapter 3) – We define optimal predictive rule lists for single- and multi-target classification and regression using the MDL principle. For classification, we derive two optimal encodings: the prequential plug-in; and the Normalized Maximum Likelihood (NML) (Section 3.4.3). For regression we use a Bayesian encoding with non-informative priors (Section 3.5.3).

2. CLASSY *algorithm* (Chapter 4) – We propose a heuristic algorithm for finding good predictive rule list for multiclass classification. The algorithm combines a frequent pattern mining algorithm to mine all the candidate rules with a greedy search to sequentially add rules to a list. Technically, CLASSY only has one hyperparameter, the candidate rules taken as input to find the rule list. It is empirically shown that Classy outperforms RIPPER, C5.0, CART, and Scalable Bayesian Rule Lists (SBRL) [125] when it comes to the combination of classification performance and interpretability.

Our contributions on **subgroup lists** and **subgroup discovery** are the following:

3. *Subgroup list model class* (Chapter 2) – We define the subgroup list model class over a tabular dataset in general, providing a *global* probabilistic formulation for the problem of sequential subgroup mining, and in particular for univariate and multivariate, nominal and numeric targets.

4. *Robust subgroup lists using MDL* (Chapter 3) – We define optimal subgroup lists using the MDL principle, where we resort to the optimal Normalized Maximum Likelihood (NML) encoding for nominal targets (Section 3.4) and the Bayesian encoding with non-informative priors for numeric targets (Section 3.5). Notably, we show that this problem formalization is equivalent to the standard definition of top-1 subgroup discovery with Weighted Kullback-Leibler (WKL) divergence as quality measure for the case of a subgroup list with one subgroup (Section 3.4.3 for nominal targets and Section 3.5.3 for numeric targets).

5. *RSD algorithm* (Chapter 5) – We propose the *Robust Subgroup Discoverer* (RSD) algorithm, which combines beam search to find subgroups with greedy search to iteratively add the best found subgroup to the subgroup list (Section 5.2). We show that the greedy objective is equivalent to a one-sample Bayes proportions, multinomial, or t-test (for binary, nominal or numeric targets, respectively) plus a penalty to compensate for multiple hypothesis testing (Section 3.4.4 for binary and nominal targets, Section 3.5.4 for numeric targets, and Section 5.2.3 for the greedy objective of RSD).

The contributions on **MDL learning theory** are the following:

6. *Prequential plug-in code for partition models* – Derivation of the prequential plug-in asymptotically optimal encoding, a refined MDL encoding, for model classes that partition the data for nominal target variables—subgroup lists, rule lists, trees, etc. (presentation in Section 3.4 and full derivation in Appendix B).

7. *Normalized Maximum Likelihood for partition models* – Derivation of the Normalized Maximum Likelihood (NML) optimal encoding, a refined MDL encoding,

for model classes that partition the data for nominal target variables—subgroup lists, rule lists, trees, etc. (presentation in Section 3.4 and full derivation in Appendix C).

8. *Bayesian encoding of normal distributions for partition models* – Derivation of a Bayesian optimal encoding of normal distributions with non-informative priors for numeric targets (presentation in Section 3.5 and full derivation in Appendix D). It is shown that for large number of instances it converges to the BIC (Appendix E). Similarly to the prequential and NML encodings, it can be used by any model class that unambiguously partitions the data, such as subgroup lists, rule lists, trees, etc.

9. *Greedy MDL algorithms maximize local statistical test* (Chapter 5) – We show that the greedy gain commonly used in the MDL for pattern mining literature can be interpreted as an MDL equivalent to a *local* Bayesian hypothesis test, a.k.a. Bayesian factor, on the likelihood of the data being better fitted by the greedy extended model versus the current model, plus a penalty for the extra model complexity (Section 5.2.3).

Finally, our contribution on the difference between **predictive rules** and **subgroup discovery rules**:

10. *Subgroups discovery versus rule-based prediction* – We demonstrate the difference between the formal objectives for subgroup discovery and predictive rule models, such as classification rule lists, from the perspective of our MDL-based approach (Section 3.7).

1.4 Outline of this dissertation

The structure of this dissertation is as follows. Chapter 2 presents the fundamental problem definitions and mathematical notation necessary to understand later chapters. It starts with a gentle introduction of association rules in rule-based classifiers, subgroup discovery, and subgroup set discovery, posteriorly formalizing these tasks for supervised data. Moreover, it presents the rule list model class and specializes this generic model class to the predictive rule list in machine learning and the subgroup list in subgroup discovery. Then, it shows how to empirically measure the quality of classification models, subgroups, and subgroup sets.

Chapter 3 presents how to encode rules, predictive rule lists, and subgroup lists using the MDL principle for univariate and multivariate nominal and numeric target variables. Then, it proceeds to prove the equivalence of MDL-based subgroup lists with one subgroup and the standard definition of subgroup discovery—top-1 mining—with the Weighted Kullback-Leibler (WKL) divergence as a quality measure. Finally, we use our MDL formulation of predictive rule lists and subgroup lists to find the similarities and differences between rule-based prediction and subgroup discovery.

Chapter 4 introduces CLASSY, a heuristic algorithm based on the MDL principle to find good predictive rule lists for multiclass classification. Then, extensive empirical comparisons validate our proposed MDL formulation and algorithm in terms of classification performance, interpretability, overfitting, and runtime. They show that CLASSY is competitive in classification performance against state-of-the-art algorithms that produce rule-based models (or trees) while usually finding simpler models.

In Chapter 5, we propose the Robust Subgroup Discoverer (RSD) algorithm finds good subgroup lists based on our MDL formulation. It combines beam search for candidate generation with greedy search to add one subgroup at a time. Moreover, this greedy gain equals an MDL equivalent of Bayesian testing. Then, our MDL formulation and RSD show that they obtain high-quality subgroup lists on 54 datasets compared to state-of-the-art algorithms. In the end, we conduct three case studies to show how RSD works on real-world problems.

Finally, Chapter 6 presents the main conclusions of this dissertation and possible future work directions.

1.5 Publications

The chapters of this thesis are based on the following publications:

- H. M. Proença and M. van Leeuwen. Interpretable multiclass classification by mdl-based rule lists. *Information Sciences*, 512:1372–1393, 2020

- H. M. Proença, P. Grünwald, T. Bäck, and M. van Leeuwen. Discovering outstanding subgroup lists for numeric targets using mdl. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 19–35. Springer, 2020

- H. M. Proença, T. Bäck, and M. van Leeuwen. Robust subgroup discovery. *Data Mining and Knowledge Discovery (preprint available in arXiv:2103.13686)*, submitted

Other publications

- H. M. Proença, R. Klijn, T. Bäck, and M. van Leeuwen. Identifying flight delay patterns using diverse subgroup discovery. In *2018 IEEE SSCI*, pages 60–67. IEEE, 2018