Robust rules for prediction and description

Manuel Proenca, H.

Robust rules for prediction and description

Hugo Manuel Proença

Keywords machine learning · data mining · rule lists · subgroup lists · subgroup discovery · pattern mining · interpretability · the Minimum Description Length (MDL) principle · Bayesian statistics.

SIKS Dissertation Series No. 2021-23
The research reported in this thesis has been carried out under the auspices of SIKS, the Dutch Research School for Information and Knowledge Systems.

ISBN: 978-94-6332-792-3
This work is part of the research programme Indo-Dutch Joint Research Programme for ICT 2014 with project number 629.002.201, SAPPAO, which is (partly) financed by the Netherlands Organisation for Scientific Research (NWO), in collaboration with IIT Roorkee and GE Global Research Bangalore.

Printed by: GVO Drukkers & Vormgevers B.V.
Cover: Kimber McLaughlin @pixelatedpeach

Typeset using LaTeX, diagrams generated using MATPLOTLIB and SEABORN.

Robust rules for prediction and description

Proefschrift

ter verkrijging van
de graad van doctor aan de Universiteit Leiden,
op gezag van rector magnificus prof.dr.ir. H. Bijl,
volgens besluit van het college voor promoties
te verdedigen op dinsdag 26 oktober 2021
klokke 10.00 uur

door

Hugo Manuel Proença

geboren te Hong Kong
in 1990

Promotiecommissie

Promotor:	Prof.dr. T.H.W. Bäck
Co-promotor:	Dr. M. van Leeuwen
Overige leden:	Prof.dr. A. Plaat
	Prof.dr.ir. N. Mentens
	Prof.dr. P.D. Grünwald (CWI, The Netherlands)
	Prof.dr. A.P.J.M. Siebes (Universiteit Utrecht, The Netherlands)
	Prof.dr. J. Vreeken (CISPA Helmholtz Center for Information Security, Germany)

aos meus pais

"You look at where you're going and where you are and it never makes sense, but then you look back at where you've been and a pattern seems to emerge. And if you project forward from that pattern, then sometimes you can come up with something."

Robert M. Pirsig in *Zen and the Art of Motorcycle Maintenance*

Contents

List of symbols

Supervised Dataset

D	Labelled dataset.
$\mathbf{X}$	Dataset of explanatory variables of D.
X	An explanatory variable of $\mathbf{X}$.
$\mathcal{X}$	Domain of X.
$\mathbf{x}$	A explanatory variables sample of $\mathbf{X}$.
x	The value of sample $\mathbf{x}$ for variable X.
$\mathbf{Y}$	Dataset of target variables of D.
Y	An target variable of $\mathbf{Y}$.
$\mathcal{Y}$	Domain of Y.
$\mathbf{y}$	A target variables sample of $\mathbf{Y}$.
y	The value of sample $\mathbf{y}$ for variable Y.
i	Index for subsetting by row.
j	Index for subsetting by column.
v	A generic explanatory variable.

k Number of classes of a nominal target variable.

n Number of examples in dataset D.

m Number of explanatory variables.

t Number of target variables.

d Subscript associated with dataset distribution or defaul rule.

Model classes

M Generic model (either a rule list or a subgroup list).

RL Rule list model (including rules R and default rule).

R Rules in model RL.

r A rule.

SL Subgroup list model (including subgroups S and default rule).

S Subgroups in model SL.

s A subgroup.

ω Number of rules/subgroups in M.

a Description of a subgroup s or a rule r.

a_i Description of the i^{th} rule/subgroup in model M.

$D^a = \{\mathbf{X}^a, \mathbf{Y}^a\}$ Samples of dataset D covered by description a.

$D^i = \{\mathbf{X}^i, \mathbf{Y}^i\}$ Samples of dataset D covered by the i^{th} description in model M.

$Dist(\Theta)$ Generic probability distribution with parameters Θ.

$\mathcal{N}(\mu; \sigma)$ Normal probability distribution with parameters μ and σ.

$Cat(p_1, \cdots, p_k)$ Categorical probability distribution with p_i probability per category.

μ Mean value parameter.

σ Standard deviation parameter.

δ Effect size (ratio of μ and σ).

$\hat{\theta}$ Maximum likelihood estimation of parameter θ.

Subgroup Discovery

$q(a)$ Subgroup discovery quality measure.

$Q(S)$ Subgroup set discovery quality measure.

$f(\hat{\Theta}^a, \hat{\Theta}^d)$ Function of differences between distribution $\hat{\Theta}^a$ and $\hat{\Theta}^d$.

α Tradeoff between subgroup coverage and distribution difference.

KL Kullback-Leibler divergence general form.

KL_{Cat} Kullback-Leibler divergence for categorical distributions.

KL_{μ} Kullback-Leibler divergence for location distributions.

$KL_{\mu,\sigma}$ Kullback-Leibler divergence for normal distributions.

WKL Weighted Kullback-Leibler divergence general form.

SWKL Sum of Weighted Kullback-Leibler divergences.

MDL

$L(\cdots)$ Length of encoding.

ℓ Log-likelihood.

$L_{\mathbb{N}}$ Universal code of integers.

$L_{NML}(Y_j^i)$ Normalized Maximum Likelihood length of encoding of data Y_j^i.

$\mathcal{C}(n_a, k)$ Multinomial distribution complexity with n_a points and k categories.

L_{Bayes} Bayesian length of encoding with improper priors.

$Y^{i|2}$ The two points that make the Bayesian encoding proper.

$L_{Bayes2.0}$ Bayesian length of encoding made proper with first 2 points.

$\Gamma(n)$ Gamma function, the extension of the factorial to real numbers.

Algorithm

$\Delta_\beta L(D, M \oplus a)$ Compression gain of adding description a to model M.

β Level of normalization of the compression gain.

ζ Set of all items (possible single conditions) in $\mathbf{X}$.

$stats$ Statistics of a subgroup.

d_{max} Beam search maximum depth of search.

w_b Beam search beam width.

n_{cut} Number of cut points for numeric discretization.

Acronyms

AUC Area Under the receiver operator Curve.

BIC Bayesian Information Criterion.

BTS Bureau of Transportation Statistics.

CART Classification And Regression Trees.

CORELS Certifiably Optimal RulE ListS.

CRS Computerized Reservation System.

DSSD Diverse Subgroup Set Discovery.

EDA Exploratory Data Analysis.

EWR NEWaRk liberty international.

FN False Negative.

FP False Positive.

FPR False Positive Rate.

FSSD Fast and efficient algorithm for Subgroup Set Discovery.

FURIA Fuzzy Unordered Rule Induction Algorithm.

IDS Interpretable decision sets.

KDD Knowledge Discovery from Data.

KL Kullback-Leibler divergence.

MCMC Markov Chain Monte Carlo.

MCTS Monte Carlo Tree Search.

MCTS4DM Monte Carlo Tree Search for Data Mining.

MDL Minimum Description Length.

NML Normalized Maximum Likelihood.

RIPPER Repeated Incremental Pruning to Produce Error Reduction.

RSD Robust Subgroup Discoverer.

RSS Residual Sum of Squares.

SaC Separate and Conquer.

SAPPAO a Systems APproach towards data mining and Prediction in Airlines Operations.

SBRL Scalable Bayesian Rule Lists.

SD Subgroup Discovery.

SISD Subjectively Insteresting Subgroup Discovery.

SSD Subgroup Set Discovery.

SVM Support Vector Machine.

SWKL Sum of Weighted Kullback-Leibler divergences.

TN True Negative.

TP True Positive.

TPR True Positive Rate.

UA United Airlines.

WKL Weighted Kullback-Leibler divergence.

WRAcc Weighted Relative Accuracy.

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

2 Preliminaries

In this chapter predictive rule lists and subgroup lists are presented. To that end, we give a gentle introduction to association rules, what constitutes a rule-based classifier and subgroup discovery, and how to measure the quality of a rule-based model in classification and subgroup discovery.

This chapter is divided as follows. First, in Section 2.1 we give a high-level introduction of association rules, rule-based classifiers, subgroup discovery, and subgroup set discovery. Next, in Section 2.2 the notation for supervised structured *i.i.d.* data is presented together with a formal definition of prediction, subgroup discovery, and subgroup set discovery tasks. Then, in Section 2.3 association rules and their characteristics are introduced. After that, in Section 2.4 the rule list model class is defined in general plus the specific case of the predictive rule lists and the subgroup list. Then, in Sections 2.5, 2.6, and 2.7 performance measures for classification, quality measures for subgroup discovery, and quality measures for subgroup set discovery are introduced, respectively.

2.1 Introduction to rules

The main topics of this thesis, rule-based classification and subgroup discovery, are two paradigms arising from related fields, machine learning and data mining, respectively. Both topics share the fact that they are supervised tasks on structured data that resort to *association rules* to construct their models. Thus, we will now informally introduce what each of these tasks encompasses, starting from what they have in common, and finalizing with their differences.

Association rule. An association rule $a \mapsto b$ is an assertion of a possible relationship between the antecedent a and consequent b, which can be read in the form of "*If* a appears in the data *then* b usually also appears" with a certain level of confidence [2]. A classic example from market basket analysis is that people who buy bread and butter (antecedent), usually, also buy milk (consequent) [2]. In this case, the association rule takes the form of: $\{bread = yes\}\&\{butter = yes\} \mapsto milk = yes$. A probabilistic extension of these rules, deemed a probabilistic association rule [81], associates a parametric probability distribution to the consequent, thus, instead of having a crisp decision, it has a probability associated with each possible case:

$$a \mapsto b \sim Dist(\Theta), \tag{2.1}$$

where Θ are the parameters of the distribution $Dist$ that describe the consequent. In the case of the previous example, where the consequent is a binary variable, this could take the form of: $\{bread = yes\}\&\{butter = yes\} \mapsto milk \sim Bernouilli(p_{yes} = 0.60; p_{no} = 0.40)$; where p_{yes} is the probability of having bought milk, and $p_{no} = 1 - p_{yes}$ the probability of not having bought it. A rule is said to be active in a region of the data D if for a data instance $\mathbf{x} \in D$ its antecedent is present, such as in our example $\{bread = yes\}\&\{butter = yes\}$.

Rule-based classifiers. Classification is the task of predicting an unseen outcome y of a discrete target variable from an instance of explanatory variables $\mathbf{x}$ [36]. In order to learn the relationship between the variables, a classification model is learned from a supervised dataset $D = \{\mathbf{X}, Y\}$, which is composed of paired examples $(\mathbf{x}, y)$. Note that we only talk about rule-based classifiers and not regressors because, to the best of our knowledge, there are no competitive rule-based models for regression.

Rule-based classifiers aggregate several rules together in order to perform classification. Combining rules in different ways leads to different rule-based models, of which two stand out [39]: $1)$ *rule list or sequential activation* [109, 85]—the activation of the rules for prediction follows a pre-determined order of the form *if rule 1 then* $Dist(\Theta^1)$*... else if rule 2 then* $Dist(\Theta^2)$, finishing with a default *else* $Dist(\Theta^m)$ that captures all the data not covered by any of the previous rules; $2)$ *rule set or overlapping rules* [21]—an unordered set of rules where several individual rules can be activated at the same time, overlapping. The key difference between both models is that rule lists are ordered and only one rule is active for one data sample $\mathbf{x}$, while rule sets are unordered and multiple rules can be active for one data sample $\mathbf{x}$.
The objective of a rule-based classifier is to maximize a performance measure, thus each association rule should contribute to that *global* goal, i.e., each rule should

take into account other rules to maximize the overall quality of the classifier. Looking back at our example, we see that $\{bread = yes\}\&\{butter = yes\} \mapsto milk \sim Bernouilli(p_{yes} = 0.60; p_{no} = 0.40)$ does not seem particularly good for prediction as it does not distinguish very well between both classes. On the other hand, the rule $\{yoghurt = yes\} \mapsto milk \sim Bernouilli(p_{yes} = 0.90; p_{no} = 0.10)$ seems well suited for prediction. A note should be made in relation to decision lists, which have the same format as rule lists, but instead of combining probabilistic association rules, they are composed of decision rules, with a crisp decision as in our example $\{bread = yes\}\&\{butter = yes\} \mapsto milk = yes$, and appeared first in the literature [109].

Subgroup Discovery (SD). Subgroup discovery is the data mining task of finding subgroups that stand out with respect to some given target variable(s). The definition of standing out, also known as interestingness, is quantified by a quality measure, which depends on the task at hand [123, 63]. In general, these measures quantify quality by how different the target variable distribution of a subgroup is from what is defined as 'normal' behavior in a dataset. In the case of structured data, a subgroup generally takes the form of an association rule, and the 'normal' behavior is usually measured by the average behavior of the target variable of that dataset [8]. Going back to the market basket analysis example, let us consider a dataset made up of the shopping baskets of different clients, and that has as target variable if a client bought milk (or not). 'Normal' behavior can be given by the percentage of people that buy milk over the whole dataset, and let us assume that this value is 90%. Thus, the subgroup defined by $\{bread = yes\}\&\{butter = yes\} \mapsto milk \sim Bernouilli(p_{yes} = 0.60; p_{no} = 0.40)$ seems interesting, as compared with normal behavior, people that buy bread and butter buy milk 33% times fewer times than an average client. This is in clear contrast with rules for prediction, as subgroups that are interesting do not have to divide well between classes: they need to stand out with respect to what is 'normal' behavior in the data. Sometimes, depending on the dataset and task at hand, a good subgroup will also be a good predictive rule, but both tasks arise from different goals and should thus not be confused. In its standard form, subgroup discovery is called top-k mining, as the goal is to find the k top subgroups that maximize a user-defined quality measure. As the quality measures only quantify the individual quality of a subgroup, top-k mining is a *local* paradigm, as it is only concerned with the independent performance of the k subgroups on the respective data covered by each of their descriptions. Top-k subgroup discovery usually finds subgroups that cover the same region of the data, hence it returns redundant subgroups for many datasets. As a solution to this, *subgroup set discovery* was proposed.

Subgroup Set Discovery (SSD). The task of finding a non-redundant set of subgroups that are individually and collectively interesting at the same time is called Subgroup Set Discovery (SSD) [75]. Contrary to a predictive paradigm, the objective is that the subgroups still abide by the standard subgroup discovery principle of *locally* standing out with respect to the 'normal' behavior, while at the same time, *globally* describing different regions of the dataset. To extend subgroup discovery to its set form, two main models exist: 1) *subgroup lists or ordered sets* [71]—a set of subgroups that should be interpreted sequentially and where no subgroup is allowed to overlap in the same region of data as another, take the form of *if subgroup 1 then* $Dist(\Theta^1)$... *else if subgroup 2 then* $Dist(\Theta^2)$, etc.; and 2) *subgroup sets or overlapping sets* [74]—a set of subgroups where each subgroup can be interpreted individually and overlap is allowed according to a definition of overlap interaction. Both extensions have their advantages and disadvantages: while subgroup lists are less interpretable, they have the advantage of a clear definition of the relevance of each subgroup and which subgroup explains each data point. On the other hand, subgroup sets allow for a (semi)independent interpretation of the subgroups, but they require an extra definition that favors non-redundant sets together with a definition of the interaction of subgroups in the region where they overlap, e.g., as a mixture model.

Rule-based classifiers *versus* Subgroup Set Discovery. As was shown throughout this section, predictive rules and subgroups share a lot of the same characteristics. Rule-based classifiers aggregate association rules to maximize a *global* objective of a good overall classification, while subgroup sets balance both a *local* definition of quality with respect to the 'normal' behavior of the dataset and a *global* objective of covering different regions of the data. It is natural that for some datasets good subgroups will be good predictive rules and vice versa, but this is not always the case and it should be distinguished. Throughout this work, we will be referencing them separately to emphasize the different paradigms: *1) predictive rule* will refer to an association rule as used in rule-based models for classification in machine learning; *2) subgroup* to descriptive rules in subgroup discovery; and *3) association rule* or just *rule* to an association rule in general, i.e., when it refers to either a predictive rule or a subgroup. All their idiosyncrasies may not be apparent yet, but as we progress we will continue to emphasize their similarities and differences.

2.2 Supervised data

Consider a dataset $D = (\mathbf{X}, \mathbf{Y}) = \{(\mathbf{x}^1, \mathbf{y}^1), (\mathbf{x}^2, \mathbf{y}^2), ..., (\mathbf{x}^n, \mathbf{y}^n)\}$ of n *i.i.d.* instances. Each instance $(\mathbf{x}^i, \mathbf{y}^i)$ is composed of a vector of explanatory variable values $\mathbf{x}^i$ and a

vector of target variable values $\mathbf{y}^i$.
Each observed explanatory vector has m values $\mathbf{x} = [x_1, ..., x_m]$, one for each variable $X_1, ..., X_m$. The domain of a variable X_j, denoted $\mathcal{X}_j$, can be one of two types: nominal or numeric. Similarly, each observed target vector is composed of t values $\mathbf{y} = [y_1, ..., y_t]$, one for each target variable $Y_1, ..., Y_t$, with associated domains $\mathcal{Y}_j$. The target variables can be one of two types: nominal, or numeric. In the nominal case it is $\mathcal{Y}_j = \{1, \cdots, k\}$, with $\mathcal{Y}_j$ the set of k classes/categories of variable Y_j, and in the numeric, the domain is $\mathcal{Y}_j = \mathbb{R}$.
Note that we use subscripts on the dataset variables ($D, \mathbf{X}, \mathbf{Y}, X, Y, x, y$) to indicate column subsets and overscripts to subset over rows. In the case of other notation, such as number of elements n or statistics μ, σ we will not use the superscript as it can be confused with the exponentiation of that value. Also, X_i (resp. Y_i) refers to both the properties of the i^{th} explanatory (resp. target) variable and to all the values of this variable for a specific column. When the dataset only contains one target variable $\mathbf{Y}$ is substituted by Y.

Prediction In statistical learning, the task of prediction is to infer unseen values of a target variable from a set of explanatory variables through the use of past evidence that shows the relationship between target and explanatory variables [36]. Formally, this means that we want to find the best mapping g, from a space of possible hypotheses $\mathcal{G}$, between explanatory data $\mathbf{X}$ to target data Y (in the univariate case and without loss of generality). This mapping can be summarized as $g : \mathcal{X}_1 \times \cdots \times \mathcal{X}_m \rightarrow \mathcal{Y}$; and in the case of a probabilistic predictor, such as ours, this mapping is just a conditional probability $g(x) = \Pr(y \mid \mathbf{x} = x)$, and by abuse of notation $g(\mathbf{X}) = \{g(x^1), \cdots, g(x^n)\}$. Assuming that we are dealing with probabilistic mappings, we can now start making predictions $\hat{y}$ for the target variable values for each instance $\mathbf{x}$, by returning the outcome with the largest probability

$$\hat{y} = \arg\max_{y \in \mathcal{Y}} \Pr(y \mid \mathbf{x}) \tag{2.2}$$

The characteristics of a good mapping are: 1) capture the properties in $\mathbf{X}$ that allow predicting Y; and 2) generalize well on previously unseen data $D_{new} = \{\mathbf{X}_{new}, Y_{new}\}$. In order to choose the best possible mapping, we need to introduce a performance measure $meas$ that empirically quantifies the quality of our mappings for a given dataset, formally $meas : \mathcal{Y}^n \times \mathcal{Y}^n \rightarrow \mathbb{R}_{\geq 0}$. Thus the problem of finding the best mapping g in a dataset $D = \{\mathbf{X}, Y\}$ reduces to:

$$g^* = \arg\max_{g \in \mathcal{G}} meas(Y, g(\mathbf{X})), \tag{2.3}$$

but then another, D_{new} is required for evaluation, as this takes into account generalization and avoids overfitting. Some examples of measures $meas$ for classification

are the accuracy or the AUC, described in Section 2.5, or the Mean Squared Error for regression.
Several variations exist, such as using only predictions $\hat{y}$ instead of $g(\mathbf{x})$ or structural measures that add an extra term to $meas$ to penalize for the structural complexity of the mapping [119]. E.g., in the case of nested mappings such as a polynomial regression, the use of higher-order polynomials is "more complex" than lower-order ones, as they have extra terms. The Minimum Description Length (MDL) principle used throughout this dissertation, is a type of probabilistic structural error minimization principle and this mapping g is called a model M or point hypothesis in it [47].

Subgroup discovery Subgroup discovery is the data mining task of discovering *unkown* patterns in the data that stand out with respect to a target variable [116]. In mathematical terms the objective is to find a mapping between descriptions a of the explanatory data $\mathbf{X}$ and the target variable Y (for the univariate case without loss of generality) that stand out in relation to the 'normal' behavior of the target variable Y. Formally, a description is a function $a : \mathcal{X}_1 \times \cdots \times \mathcal{X}_m \mapsto \{false, true\}$. And in our specific case, a description a is a conjunction of conditions on $\mathbf{X}$, each specifying a value or interval on a variable. The domain of possible conditions depends on the type of a variable: numeric variables support *greater and less than* $\{\geq, \leq\}$; nominal support *equal to* $\{=\}$. E.g., from Figure 2.2, where for the *Car import* dataset, a description can be "weight = heavy & consumption-city ≤ 8 km/L", where the variable weight is conditioned to one value (norminal variable) and $consumption - city$ is conditioned to one interval (numeric variable). As the dataset is made of pairs $(\mathbf{x}^i, y^i)$, for each description a there is an associated subset of data $D^a = \{\mathbf{X}^a, Y^a\}$ with $n_a = D^a$ instances, and an associated empirical parameter distribution of the target Y^a given by $\hat{\Theta}^a$—where the parameters depend on the distribution selected by the user. Thus, in the case of *i.i.d.* data, a subgroup is an association rule $s : a \mapsto y \sim Dist(\hat{\Theta}^a)$.
To quantify how interesting a subgroup s with description a is, we need to define a quality measure $q(n_a, \hat{\Theta}^a, \hat{\Theta}^d)$ that is a function of the subgroup empirical distribution $\hat{\Theta}^a$ and the dataset empirical marginal distribution $\hat{\Theta}^d$—'normal' behavior of the dataset.
Formally the best subgroup, or top-1 subgroup, is given by

$$s^* = \underset{s=(a,\hat{\Theta}^a)\in\mathcal{A}}{\arg\max}\ q(n_a, \hat{\Theta}^a, \hat{\Theta}^d), \tag{2.4}$$

where in the case of top-k subgroup discovery, we return the k top ranking subgroups that maximize q. An example of a quality measure for binary targets in the Weighted Relative Accuracy (WRAcc) or the Weighted Kullback-Leibler (WKL) divergence presented in Section 2.6. Contrary to *prediction*, SD does not aim at performing