**Queen Mary University of London**
**School of Electronic Engineering and Computer Science**

# EFFECT OF COGNITIVE BIASES ON HUMAN UNDERSTANDING OF RULE-BASED MACHINE LEARNING MODELS

**Tomas Kliegr**

Submitted in partial fulfillment
of the requirements of the Degree of
Doctor of Philosophy

October 2017

# Statement of originality

I, Tomas Kliegr, confirm that the research included within this thesis is my own work or that where it has been carried out in collaboration with, or supported by others, that this is duly acknowledged below and my contribution indicated. Previously published material is also acknowledged below.

I attest that I have exercised reasonable care to ensure that the work is original, and does not to the best of my knowledge break any UK law, infringe any third party's copyright or other Intellectual Property Right, or contain any confidential material.

I accept that the College has the right to use plagiarism detection software to check the electronic version of the thesis.

I confirm that this thesis has not been previously submitted for the award of a degree by this or any other university.

The copyright of this thesis rests with the author and no quotation from it or information derived from it may be published without the prior written consent of the author.

Signature:

Date: October 24, 2017

*Details of collaboration and publications*: Part I of the thesis benefited from frequent discussions with Prof. Johannes Fürnkranz and Prof. Heiko Paulheim during my research stay at TU Darmstadt. In particular, it was the idea of Prof. Fürnkranz to study the influence of rule length on user preference, he also steered me towards executing experiments on the UCI Mushroom datasets. Prof. Paulheim provided the other three experimental datasets (Movies, Quality and Traffic) including the precomputed values of PageRank. It was also his idea to control performance of the crowdsourcing annotators by randomly swapping literals in the rules. Julius Stecher provided rules from the inverted heuristics learner. Instrumental to the design of the empirical study was experience gained with human annotation and crowdsourcing within the work on Linked Hypernyms Dataset [Kliegr, 2015, Kliegr and Zamazal, 2016]. Part II benefited from feedback I obtained from my two supervisors, Prof. Ebroul Izquierdo and Dr. Christopher Tyson. An earlier version of the text in Section 7.2, which gives account of related work in the area of association rule classification was written by myself and presented in collaborative papers Kliegr et al. [2014],

Fürnkranz and Kliegr [2015]. Similarly, an earlier version of the text in Section 7.3 covering the UTA method was written by myself and presented in collaborative paper Eckhardt and Kliegr [2012]. Other publications loosely topically relating to this thesis involve those covering the InBeat system: Kuchař and Kliegr [2013], Kliegr and Kuchař [2014], Kuchař and Kliegr [2014], Kliegr and Kuchař [2015], Kuchař and Kliegr [2017].

# Abstract

This thesis investigates to what extent do cognitive biases affect human understanding of interpretable machine learning models, in particular of rules discovered from data. Twenty cognitive biases (illusions, effects) are analysed in detail, including identification of possibly effective debiasing techniques that can be adopted by designers of machine learning algorithms and software. This qualitative research is complemented by multiple experiments aimed to verify, whether, and to what extent, do selected cognitive biases influence human understanding of actual rule learning results. Two experiments were performed, one focused on eliciting plausibility judgments for pairs of inductively learned rules, second experiment involved replication of the Linda experiment with crowdsourcing and two of its modifications. Altogether nearly 3.000 human judgments were collected. We obtained empirical evidence for the insensitivity to sample size effect. There is also limited evidence for the disjunction fallacy, misunderstanding of "and", weak evidence effect and availability heuristic.

While there seems no universal approach for eliminating all the identified cognitive biases, it follows from our analysis that the effect of many biases can be ameliorated by making rule-based models more concise. To this end, in the second part of thesis we propose a novel machine learning framework which postprocesses rules on the output of the seminal association rule classification algorithm CBA [Liu et al, 1998]. The framework uses original – undiscretized – numerical attributes to optimize the discovered association rules, refining the boundaries of literals in the antecedent of the rules produced by CBA. Some rules as well as literals from the rules can consequently be removed, which makes the resulting classifier smaller. Benchmark of our approach on 22 UCI datasets shows average 53% decrease in the total size of the model as measured by the total number of conditions in all rules. Model accuracy remains on the same level as for CBA.

# Contents and Chapter Headings

# List of Figures

# List of Tables

*Both the content and direction of biases can be predicted theoretically and explained by optimality when viewed through the long lens of evolutionary theory. Thus, the human mind shows good design, although it is design for fitness maximization, not truth preservation.*

Haselton and Nettle [2006]

*to Lucia*

# Acknowledgments

I am grateful to Prof. Ebroul Izquierdo, who made this thesis possible and was always there to assist me throughout the long course of PhD study. Also, I must acknowledge my additional supervisor Dr. Christopher Tyson for his constructive feedback at all stages of the work and my second supervisor Dr. Ioannis Patras. I would also like to thank my colleagues and friends at Queen Mary for their encouragement and advice, in particular Dr. Tomáš Piatrik and Dr. Krishna Chandramouli. I would also like to thank my employer, the Faculty of Informatics and Statistics at the University of Economics in Prague for supporting my sabbatical with Prof. Johannes Fürnkranz at TU Darmstadt, which contributed to the scope and quality of the thesis. TU Darmstadt also covered most of the costs of the performed experiments. My UEP colleague Prof. Vojtěch Svátek provided the first contact with Prof. Ebroul Izquierdo, which eventually led to this thesis. Dr. Magda Osman and Prof. Ke Chen provided detailed feedback and comments that helped to improve many aspects of the revised version of the thesis. Above all, I would like to thank my family for their support.

**License for reprinted work.** The right to use excerpt from articles Kliegr and Kuchař [2015], Fürnkranz and Kliegr [2015] in this thesis has been granted by Springer Int. Publishing (License Number 4023040104898,4183080648393).

## List of Acronyms

**ARC** Association Rule Classification (class of algorithms)

**CBA** Classification By Associations (classification algorithm)

**CMAR** Classification based on Multiple Association Rules (classification algorithm)

**CPAR** Classification based on Predictive Association Rules (classification algorithm)

**DROP** Default Rule Overlap Pruning (algorithmic step)

**FARC-HD** Fuzzy Association Rule-Based Classification Model for High-Dimensional Problems (classification algorithm)

**FURIA** Fuzzy Unordered Rule Induction (algorithm)

**LOD** Linked Open Data

**MARC** Monotonicity-exploiting Association Rule Classification (working name for contributed algorithmic framework)

**MDLP** Minimum Description Length Principle (preprocessing algorithm)

**QCBA** Quantitative Association Rule Classification (public name for the MARC algorithmic framework)

**TTB** Take-The-Best (algorithm)

**UTA-NM** UTilités Additives Non Monotonic (decision-support algorithm)

**UTA** UTilités Additives (decision-support algorithm)

# 1. Introduction

This thesis aims to investigate the effect of cognitive biases on human understanding of machine learning models, in particular inductively learned rules. We use the term cognitive bias as a representative term for various related cognitive phenomena (heuristics, effects, illusions and constraints) that demonstrate as seemingly irrational reasoning patterns that are thought to allow humans to make fast and risk averse decisions. The fundamental questions we will seek answer to is: How can cognitive biases affect understanding of rule-based models? Can we improve the interpretability of machine learning models by considering cognitive biases?

First, we study systematic distortions in human understanding of inductively learned rules. Suppressing the cognitive bias, or "debiasing" the human consumers of machine learning models, is a prerequisite to machine learning models being properly interpreted. For example, in article entitled "Psychology of Prediction" Kahneman and Tversky [1973] warned that cognitive biases can lead to violations of the Bayes theorem when people make fact-based predictions under uncertainty. These results also apply to plausibility of inductively learned rules, since these are associated with measures such as confidence and support expressing the (un)certainty of the prediction they make. Following the "cognitive biases and heuristics" research program started by Tversky and Kahneman in the 1970s over 50 cognitive biases have been discovered to date [Pohl, 2017]. Their cumulative effect on human reasoning should not be underestimated as already the early work showed that "cognitive biases seem reliable, systematic, and difficult to eliminate" [Kahneman and Tversky, 1972]. The effect of some cognitive biases is more pronounced when people do not have well-articulated preferences [Tversky and Simonson, 1993], which is often the case in explorative machine learning.

Previous works have analysed the impact of cognitive biases on multiple types of human behaviour and decision making. A specific example is the seminal book "Social cognition" by Kunda [1999], which is concerned with their impact on social interaction. Another, more recent work by Serfas [2011] is focused on the context of capital investment. This thesis deals with the impact of cognitive biases on human understanding of rule learning results.

The first contribution of our study is analysis of twenty cognitive biases that can distort interpretation of rule learning results. We include biases classified into judgment and thinking categories according to systematization proposed by Pohl [2017]. To validate whether selected biases manifest when people interpret the actual results of rule learning, we designed several experiments focused on assessing plausibility – or believability – of machine learning models. The second contribution of the thesis is – the first of its kind –

empirical study of the effect of selected cognitive biases on human understanding of machine learning results.

Based on the literature review and our experimental results, we propose a model describing which cognitive biases are triggered when humans assess plausibility of inductively learned rules and whether they influence plausibility in positive or negative way. The third contribution of the thesis is the resulting graphical model of plausibility of rules. This diagram summarizes our research findings and can be used to raise awareness about the effect of cognitive biases on perception of rule learning results among the designers of machine learning algorithms and software. This model is complemented with a list of practical recommendations for the same audience.

While there seems no easy way for eliminating all the identified cognitive biases, it follows from our analysis that the effect of many biases can be ameliorated by making rule-based models more concise. The fourth contribution of the thesis is a novel machine learning framework which reduces the size of the classifier as measured by the number of conditions in the rules composing the produced models. This framework postprocesses output of the state-of-the-art association rule classification algorithm CBA [Liu et al., 1998].

**Thesis organization.** The text is divided into three parts. Part I covers the interdisciplinary study of cognitive biases and heuristics. Part II describes the machine learning framework. Finally, Part III contains the conclusions.

Part I is organized as follows. Chapter 2 covers the limited prior work related to comprehensibility of machine learning models. Chapter 3 defines the problem addressed in the first part of the thesis. Chapter 4 reviews applicable research on cognitive biases and heuristics, and presents analysis if implications they may have on comprehension of inductively learned rules. Chapter 5 describes our crowdsourcing experiments with elicitation of plausibility for rules learned from data. Finally, Chapter 6 summarizes the findings into a qualitative visual model and a list of recommendations.

Part II is organized as follows. Chapter 7 reviews related learning algorithms. Chapter 8 defines the problem that we address in this part of the thesis. Chapter 9 describes the software framework. Finally, Chapter 10 reports on the experimental evaluation.

Part III is organized as follows. Chapter 11 provides a summary of contributions for both preceding parts. Chapter 12 presents limitations of the research performed and outlook for future work.

The Appendix contains a list of software and data accompanying the thesis.

# Part I.

# Effect of Cognitive Biases on Interpretation of Rules Discovered from Data

# 2. Related Work

The main factor driving evolution of machine learning algorithms is arguably measurable improvement of accuracy on basket of reference datasets. This chapter focuses on the limited amount of research that also reflects comprehensibility of the produced models. In order to track progress in comprehensibility, machine learning research adopted the assumption that model length can be used as its direct measure, making link to the well-known Occam's razor principle. As our review shows, there are also several studies that address other types of comprehension, such as "justifiability" and compliance with domain knowledge. The discussion of various proposed notions of comprehensibility is instrumental for selection of a specific metric used to measure comprehension of rules for further analysis within this thesis. We are concerned with model comprehension by actual people who use machine learning results to get new insights and make decisions. While the emphasis of this chapter is on research performed in machine learning, some results from cognitive science, which is concerned with the study of how humans process and perceive information, are also included.

**Chapter organization.** Section 2.1 presents the notions of plausibility and comprehensibility. Section 2.2 is devoted to the Occam's razor principle. There is a paucity of research on the specific topic of rule plausibility in the machine learning literature, nevertheless there has been work in the related area of interpretability of symbolic machine learning models, especially trees, which we review in Section 2.3. Section 2.4 reviews the limited available work on semantic comprehensibility and plausibility of machine learning models. Section 2.5 relates our work to prior research done in cognitive science specifically for rules. Results from cognitive science that apply to specific cognitive biases are covered in the following chapter.

## 2.1. Plausibility and Three Levels of Comprehensibility

In this section, we introduce several definitions of comprehensibility. While the fact that machine learning models should aim for human comprehensibility is largely undisputed, the exact definition of comprehensibility that should be optimized is as of writing a subject of debate [Lipton, 2016].

### 2.1.1. Syntactic Comprehensibility

Recently there has been resurgence of work on comprehensibility of output of symbolic machine learning algorithms, in particular of rule learners. These algorithms either directly

optimize some comprehensibility metric (e.g approach of Lakkaraju et al. [2016]) or have some inherent property that is considered to produce more comprehensible results (e.g. Stecher et al. [2016]). The notion of comprehensibility optimized in these algorithms is the size of the model, which corresponds to what is referred to in philosophy as *linguistic simplicity* – the length of the message conveying the hypothesis in given language. We will refer to this type of comprehensibility as *syntactic comprehensibility*.

### 2.1.2. Semantic Comprehensibility

A second level of comprehensibility corresponds to human ability to semantically interpret the model. According to the related notion of *semantic simplicity*, this would allow the analyst to come up with examples that would falsify the model [Post, 1960]. In inductive learning it is rarely the case that it is required that a hypothesis is consistent with all data, therefore a single conflicting example will rarely falsify a learned rule. Nevertheless, scores such as confidence and support that are associated with rules directly relate to the validity of these rules given the data, corresponding to the *ability to falsify* requirement posed by Popper [1935, 2005].[1] As a component of semantic comprehensibility we consider the ability of the analyst to relate the meaning of the individual conditions to the prediction made by the rule, which allows falsification through utilizing analyst's prior knowledge on the subject.

### 2.1.3. Pragmatic Comprehensibility

Third, final level of comprehensibility relates to human ability to relate the inductively learned hypothesis to prior hypotheses applicable to the subject. This corresponds to the *pragmatic simplicity* defined by Post [1960].

### 2.1.4. Plausibility

We introduce plausibility (persuasiveness, believability) of a machine learning model as a distinct notion from comprehensibility. A prerequisite for plausibility is the ability of the analyst to interpret, or comprehend, the model on all three levels.

Syntactic comprehensibility deals with correct "mechanical understanding" of the model, while semantic and pragmatic comprehension involves domain knowledge: on the level of individual conditions for semantic comprehension and on the level of the entire hypothesis for pragmatic comprehension. Mental product of pragmatic comprehension is the subjectively assigned level of *plausibility* of the hypothesis. Symbolically, the relation between plausibility and comprehensibility is as follows: *syntactic comprehension* $\rightarrow$ *semantic comprehension* $\rightarrow$ *pragmatic comprehension* $\rightarrow$ *plausibility*. The distinction between semantic and pragmatic comprehensibility is not, in practice, made in the scope of our work.

[1] Reference Popper [1935] is for original publication in German.

## 2.2. Occam's Razor

Occam's razor principle has been used as inductive bias in machine learning algorithms under the assumption that the simplest model will perform best. This principle is interpreted in machine learning generally as "choose the shortest explanation for the observed data" [Mitchell, 1997]. This principle is attributed to English philosopher and theologian William of Ockham (c. 1287-1347).

In inductive learning context, Michalski [1983] states that inductive learning algorithms need to incorporate a preference criterion for selecting hypotheses to address the problem of the possibly unlimited number of hypotheses, and that this criterion is typically *simplicity*, referring to recent philosophical works on simplicity of scientific theories by Kemeny [1953] and Post [1960], which refine the initial postulate attributed to Ockham. Occam's razor is also used in some algorithms under the assumption that more concise models are more comprehensible. "Simplicity first" methodology is one of the basic principles of the ID3 decision tree learner [Elomaa, 1994, Quinlan, 1986], one of the most influential machine learning algorithms.

Central to investigating the validity of Occam's razor as inductive bias is the definition of simplicity. According to Post [1960] judgment of simplicity should not be made "solely on the linguistic form of the theory".[2] This type of simplicity is referred to as *linguistic simplicity*. A related notion of *semantic simplicity* is described through the falsifiability criterion [Popper, 1935, 2005]: the theory is the simpler, the more easily it can be falsified. Third, Post [1960] introduces *pragmatic simplicity* which relates to the degree to which the hypothesis can be fitted into a wider context.

In the following, we will review work on the relation of Occam's razor to model accuracy and comprehensibility.

### 2.2.1. Model Accuracy

A particular implementation of Occam's razor in machine learning is the minimum description length principle, which selects models based on the number of bits needed to encode them. In his critical analysis Domingos [1999] shows that it is provably and empirically false to favour the simpler of two models with the same training-set error on the grounds that this would lead to lower generalization error. In other words, it is not true that model simplicity leads to greater accuracy.

### 2.2.2. Comprehensibility

Occam's razor is typically transposed to the rule learning research as fewer conditions in a rule model are preferred to more conditions [Domingos, 1999]. It follows that syntactically simpler and more concise representations complying to the Occam's razor principle maximize the syntactic comprehensibility metrics as discussed above. For example, empirical

[2] According to an example included in Kemeny [1953], when comparing solar system theories, the Tycho Brahe's system is linguistically simpler than Copernicus' system because of the convenient choice of the co-ordinate system associated with heliocentristic system.

result of Lakkaraju et al. [2016] is that syntactically simpler decision sets are better interpretable than decision lists, which are nested and can contain negations. Similarly, Piltaver et al. [2014] found that trees with smaller number of levels and smaller branching factor are more comprehensible than more complex trees. It should be noted that these results apply to comparing representations, or languages for conveying models. Here, simplicity relates to the syntax or "intuitiveness" of the representation, rather than to the amount of information the model conveys.

The evidence for "shorter is more comprehensible" is not unanimous. Other empirical results that focus on comparing models with the same syntax that we cover in Sections 2.3.3 and 2.4 show that actually models or hypotheses with more information can be considered as more comprehensible [Allahyari and Lavesson, 2011] and plausible [Lavrač, 1998].

Domingos concludes that the principle of Occam Razor is still relevant for machine learning but should be interpreted as a preference for more *comprehensible* (rather than *simple*) model. Here, the term *comprehensible* clearly does not refer to syntactic length.

## 2.3. Syntactic Comprehensibility (Interpretability)

When a machine learning text refers to comprehensibility, it most likely refers to syntactic comprehensibility, synonymous terms for which are interpretability and understandability [Piltaver et al., 2014].[3]

### 2.3.1. Measurement

The primary criterion in assessing interpretability of models is their syntactic form (rules, trees, ...) as well as the related metrics. For rule-based models there are two metrics: number of conditions in the rule and number of rules in the model [García et al., 2009, Lakkaraju et al., 2016]. For decision trees, the metrics are number of leaves, branching factor, number of nodes in a branch, number of instances belonging to the leaf, the weighted sum of the depths of leaves and the weighted sum of the branching factors on the paths from the root to the leaves [Piltaver et al., 2014, 2016].

According to Bibal and Frénay [2016] measuring comprehensibility by the size of the model can be justified psychologically by the theory of Miller [1956], which states that humans can cope with 5 to 9 entities at a time. Another opinion is expressed by Otero and Freitas [2013] who argue that syntactic size of the model is not a good measure of comprehensibility and instead propose the *prediction-explanation size*, which measures the simplicity of ordered rule list by number of conditions that need to be checked to apply the model to classify an instance.

Piltaver et al. [2014] designed a survey methodology for assessing interpretability of decision trees based among others, on the human ability to use the model to classify a new instance, explain classification and rate tree comprehensibility. The survey consisted of five types of questions asking participants to i) *classify* an instance using a visualization

[3] For further discussion of the terminology refer to Bibal and Frénay [2016].

of a decision tree, ii) *explain* which attributes values must be changed/retained in order to classify instance into another class, iii) *validate* whether a statement is confirmed or rejected according to the tree, iv) *discover* task asking the subject to find an attribute-value pair that is unusual for instances from given class, v) *rate* comprehensibility of a classification tree. The level of task completion was measured with *human accuracy* and *time on the task*. The primary result of this study is the validation of methodology used, rather than any nontrivial new findings.

The way that interpretability is measured in Lakkaraju et al. [2016] is quite similar to the earlier approach of Piltaver et al. [2014].[4] The first evaluation criterion used are decision boundaries, which roughly corresponds to the *classify* task mentioned above. The second type of evaluation asked the participants to provide descriptions of classes based on the discovered rules. This task is somewhat similar to the *explain* task, however, the correctness of answers cannot be determined automatically, but needs to be assessed by human judges. The results were measured using the same criteria as Piltaver et al. [2014] used, that is in terms of i) human accuracy and b) time spent. Additionally, the authors used the average number of words comprising the textual descriptions.

### 2.3.2. Which Representation is Most Comprehensible?

The comparative study of comprehensibility of various representations received some attention in the past years. An overview of the four mostly studied symbolic representations is given in the Example box on page 25.

One of the first such studies based on empirical results from experiments involving human subjects was performed by Huysmans et al. [2011]. The conclusion is that decision tables are better in terms of comprehensibility than decision trees or textually presented rules.

| Forecast | Sunny | Sunny | Cloudy | Rain | Rain |
|---|---|---|---|---|---|
| Humidity | High | Normal | – | – | – |
| Wind | – | – | – | Strong | Weak |
| Play Yes | | x | x | | x |
| Play No | x | | | x | |

Table 2.1.: Decision table example

[4] Lakkaraju et al. [2016] seem to be unaware of the prior work of Piltaver et al. [2014].

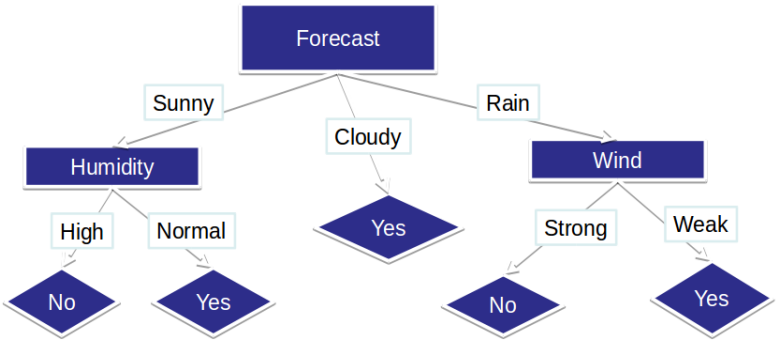



Figure 2.1.: Decision tree example

**Example. (Decision tree, decision list, decision set and decision table)**
An example of a decision tree is depicted in Figure 2.1. Decision tree consists of inner decision nodes (for example Forecast), attribute values are attached to edges (Rain), and the final decision is in the leaf node (Yes). Further discussion of decision trees can be found in the seminal work of Quinlan [1993].
The tree in Figure 2.1 corresponds to the following rules:

- IF Forecast=Sunny AND Humidity=High THEN Play(No)
- IF Forecast=Sunny AND Humidity=Normal THEN Play(Yes)
- IF Forecast=Cloudy THEN Play(Yes)
- IF Forecast=Rain AND Wind=Strong THEN Play(No)
- IF Forecast=Rain AND Wind=Weak THEN Play(Yes)

A *decision table* corresponding to these rules is depicted in Table 2.1. The decision table is divided into four quadrants delimited by the bold horizontal and vertical lines. The horizontal line divides the table into the condition part (top) and action part (bottom). The vertical line delimits attributes (left) from their values (right). The dash symbol denotes that the value is no relevant. The 'x' symbol in the action part denotes the class assigned when the conditions are met [Huysmans et al., 2011].
If the rules given above are associated with some rule quality metrics or information regarding the distribution of classes in the training data, they could also be interpreted as an *unordered rule set* [Fürnkranz et al., 2012, p. 26]. To evaluate unordered rule set all rules are tried and all firing rules can take part on assigning the class (subject to conflict resolution). Rule fires when its antecedent matches the description of the instance.
In a *decision list* rules are tried from the first to the last and the first rule which fires is used to classify the instance [Fürnkranz et al., 2012, p. 26]. Converting our example to a rule list would require adding a default rule to the end. A default rule assigns a class without checking any conditions.

According to another user study by Allahyari and Lavesson [2011] decision trees are

more comprehensible than rule sets. Most recently, Lakkaraju et al. [2016] performed a user study showing that decision sets are more interpretable than decision lists. In addition to user studies, we have identified one qualitative comparison. Freitas [2014] in his position paper discusses the comprehensibility of decision trees, classification rules, decision tables, nearest neighbours and Bayesian network classifiers. This research does not, however, include any original empirical results.

What is relevant to our research are indications that better syntactic comprehensibility of a particular type of representation (such as decision trees or rules) is tied with representation's adherence to the Occam's razor principle.

### 2.3.3. Relation between Comprehensibility and Model Size

Results on the relation between the size of representation and comprehensibility are limited and conflicting.

#### 2.3.3.1. Larger Models are Less Comprehensible

The relation between the size of representation and comprehensibility was according to our literature review for the first time empirically studied by Huysmans et al. [2011].[5] This paper confirmed the generally accepted knowledge in machine learning that "larger representations result in a decrease in answer accuracy, an increase in answer time, and a decrease in confidence."[6]

While Huysmans et al. [2011] obtained the result that larger models are less comprehensible, the study does not report on the domain knowledge the participants of their study had relating to the data used. It cannot thus be ruled out that the result they obtained was caused by lack of domain knowledge as hypothesized by Allahyari and Lavesson [2011], as discussed in the following.

#### 2.3.3.2. Larger Models are More Comprehensible

A direct elicitation of the perceived understandability of classification models has been performed by Allahyari and Lavesson [2011]. Somewhat similarly to our survey design presented in Chapter 5, these authors have performed elicitation of preferences on pairs of models which were generated from two UCI datasets: Labor and Contact Lenses. What is unique to this study is that the analysis took into account the estimated domain knowledge of the participants on each of the datasets. On the Labor dataset participants were expected to have good domain knowledge but not for the Contact Lenses dataset. The study

[5] This primacy is also directly claimed by Huysmans et al. [2011]: "These findings are consistent with a generally accepted assumption by the data mining community that people find smaller models more comprehensible. However, this assumption was until now not tested empirically in this context [Pazzani, 2000]."

[6] In Otero and Freitas [2013] the outcome of this study is interpreted exactly in the opposite way: "The findings of this study [by Huysmans et al. [2011]] indicate that the comprehensibility of the model from a user perspective tends to increase in line with the size of the model. " However, such interpretation directly contradicts with our quotation from [Huysmans et al., 2011].

was performed with 100 student subjects and involved several decision tree induction algorithms (J48, RIDOR, ID3) as well as rule learners (PRISM, Rep, JRip). Larger models were considered as more comprehensible than smaller models on Labor dataset. Allahyari and Lavesson [2011] provide the following rationale: "... the larger or more complex classifiers did not diminish the understanding of the decision process, but may have even increased it through providing more steps and including more attributes for each decision step."

The result is that more complex is more comprehensible on Labor dataset, but not on the Contact Lenses dataset. Allahyari and Lavesson [2011] explain the discrepancy as follows: "the provided data in Contact lenses fit in to a more specific area of knowledge (e.g., healthcare or medicine). Many participants have probably never heard about any of the medical terms if they have not experience an eyesight deficiency problem."

## 2.4. Semantic and Pragmatic Comprehensibility, Plausibility

Consider use case for a rule-based machine learning model, where the domain expert is presented with a list of rules learned from data and asked to filter out spurious or untrustworthy rules [Kliegr et al., 2011]. In this task, the domain expert is assumed to apply her previously acquired *domain knowledge* to assess *plausibility* of individual rules. It is implied that the domain expert is able to understand (interpret) the model.

Plausibility is closely related to the term *justifiability*, which requires the expert to assess that the model is in line with existing domain knowledge. In the recent review of interpretability definitions by Bibal and Frénay [2016], the term plausibility is not explicitly covered, but justifiability is stated to depend on the interpretability. Justifiability is defined by Martens et al. [2011] as "intuitively correct and in accordance with domain knowledge".

There is a paucity of prior research in this area. We identified the following observations relating to semantic/pragmatic comprehensibility and plausibility of machine learning models:

- Semantic coherence of literals in the rules improves (semantic) understandability [Gabriel et al., 2014].
- Violation of monotonicity constraints negatively affects plausibility [Freitas, 2014].
- Too simple hypotheses are not plausible [Freitas, 2014, Elomaa, 1994, Lavrač, 1998].
- Domain knowledge elicited from experts can be used to filter out uninteresting patterns [Pazzani, 1997, Rauch, 2009].

These results are in greater detail elaborated in the following.

### 2.4.1. Semantic Coherence

Gabriel et al. [2014] presented an approach to increasing understandability of a rule learning model by learning rules that are semantically coherent. This algorithm proceeds as follows.

The attribute label is parsed for occurrence of words that can be linked to WordNet [Fellbaum, 1998] synsets (sets of synonyms). Next, Lin's similarity [Lin, 1998] is computed between each pair of synsets. Given that each word or token can be linked to multiple synsets, the maximum similarity is used to represent the pair. Finally, the average across all pair-wise scores is used to denote the final coherence score of the rule. The experiments reported in the paper show that semantic coherence can be increased without significantly impacting accuracy of the classifier.

### 2.4.2. Too Simple not Plausible

According to Freitas [2014] and Elomaa [1994], experts are opposed to oversimplistic models. A detailed empirical account indicating that *smaller does not mean more plausible* was obtained in a study involving a small number experts is reported by Lavrač [1998]: "Frequently physicians dislike such [pruned] trees since two few parameters are taken into account and the trees describes the patients too poorly (not sufficiently detailed) to provide reliable decisions." Similar observation is reported by Lavrač [1998] for models produced by the CN2 rule learner.

### 2.4.3. Monotonicity Constraint

One of the few works that closely relates to plausibility of classification models is Freitas [2014]. It follows from this position paper that what we call plausibility[7] depends on the degree to which domain-specific constraints on monotonicity of attributes are followed.

An example of a monotonicity constraint for a numerical attribute given in Martens et al. [2011] is that increasing the weight of a newly designed car, keeping all other variables equal, should result in increased predicted fuel consumption. For nominal attributes, the monotonicity is defined by Freitas [2014] differently: a specific value will be globally predictive of a specific target class, therefore the learning algorithm should refrain from inserting this specific value to rules predicting other classes.

Martens et al. [2011] present a numerical measure for justifiability which reflects to what degree monotonicity constraints are complied with. Feelders [2000] showed on an example of real housing data and expert knowledge from real estate agents that decision tree models complying to monotonicity constraints were only slightly worse than unconstrained models in terms of classification performance, but they are much simpler.

### 2.4.4. Domain Knowledge used to Filter Uninteresting Patterns

Empirical evidence showing that domain experts do not find rules that contain conditions violating prior domain knowledge as plausible was provided already by Pazzani [1997]. This result is based on a user study, where classification rules predicting dementia were presented to neurologists.

[7] Freitas [2014] uses terms user "trust" and "acceptance".

One of the first systematic approaches to elicitation and utilization of domain constraints was done by Rauch [2009]. This paper introduced an intuitive arrow notation for elicitation of domain knowledge from experts. Elicited domain knowledge is subsequently used to filter out discovered rules.

## 2.5. Rule Learning in Cognitive Science

In the following, we will discuss the results obtained in cognitive science for studying inductive rule learning and their relevance for the problem of plausibility of inductively learned rules.

### 2.5.1. Cognitive Basis of Inductive Machine Learning

First approaches for inductive learning from data appeared in the 1950s [Michalski, 1969]. The AQ algorithm [Michalski, 1969] was the first inductive *rule learning* algorithm to use separate-and-conquer strategy [Fürnkranz, 1999], which is applied in current state-of-the-art rule learners, such as FURIA [Hühn and Hüllermeier, 2009].

While the original research paper on the AQ algorithm [Michalski, 1969] did not refer to any cognitive science research, in a follow-up work Michalski [1983] developed a theoretical framework for inductive learning which stresses the links with cognitive science. This work includes a *comprehensibility postulate* according to which "descriptions generated by inductive inference bear similarity to human knowledge representations". According to Michalski [1983] adherence to the comprehensibility postulate in rule learning is considered as "crucial".

Despite the early emphasis on the psychological dimension of comprehensibility in the early and influential articles by Michalski [1969, 1983], there is according to our review of machine learning literature no follow-up work on the transfer of results from cognitive science to the design of classification machine learning algorithms. However, considerable crossfertilization between cognitive science and machine learning occurred in the area of reinforcement learning following the publication of the Q-learning algorithm [Watkins and Dayan, 1992].

### 2.5.2. Rule as Object of Study in Cognitive Science

A hypothesis – a rule inductively learned from data – is a very specific form of alternative. Psychological research specifically on hypothesis testing in rule discovery task has been performed in cognitive science at least since the 1960's. The seminal article by Wason [1960] introduced what is widely referred to as *Wason's 2-4-6* task. Participants are given the sequence of numbers 2, 4 and 6 and asked to come up with a rule that generates this sequence. In search for the hypothesized rule they can ask the experimenter other sequences of numbers, such as 3-5-7 that are either supposed to conform to the rule or not. The experimenter answers yes or no. While the target rule is simple "ascending sequence", people find it difficult to discover this specific rule, because they apply the *confirmation*

*bias* [Nickerson, 1998], human tendency to focus on evidence confirming the hypothesis at hand [Nickerson, 1998].

One of the later works in this area is entitled "Strategies of rule discovery in an inference task" [Tweney et al., 1980]. While the title could suggest that this work is highly relevant to our machine learning problem, it is actually a psychological study of the inference processes (that is "meta-rules" people use in the reasoning process), which does not directly relate to the notion of "rules" used in machine learning, which typically correspond to particular pattern in data in a specific domain. Of similar limited relevance are follow-up works of Rossi et al. [2001], Vallée-Tourangeau and Payton [2008].

Another related field are cognitive theories of human decision making, which study how humans combine multiple pieces of evidence, which in our case correspond to conditions (literals) in a rule. The contribution of individual conditions to the overall plausibility of the rule is an important part of our research problem, but there is a paucity of directly applicable research in cognitive science. Most of this research that we identified in our brief survey is based on Bayesian reasoning (studies by Gopnik and Tenenbaum [2007], Griffiths et al. [2010]), rather than rule induction.

If we consider an individual rule (hypothesis) as one of alternatives between which the user decides, we can apply research of human decision-making processes. Most notably, this includes results of research program on cognitive heuristics and biases started by Amos Tversky and Daniel Kahneman in the 1970's. In our work, we draw heavily from this intensely studied area of human cognition. A review of applicable work from cognitive science continues in Chapter 4.

# 3. Problem Statement

The vision of finding nuggets of valuable knowledge has been one of the key drivers of development of the field of Knowledge Discovery in Databases [Fayyad et al., 1996]. The merits on which the results of interpretable machine learning models are evaluated is the analysts' appreciation of the *plausibility* (persuasiveness) of the discovered nuggets. Our work focuses on the relation between the length of a rule, as measured by the number of literals, and its plausibility.

**Chapter organization.** Section 3.1 justifies our choice of rule as the object of our study. Section 3.2 defines cognitive bias. Finally, Section 3.3 presents the research questions.

## 3.1. Inductively Learned Rule

We selected the *individual rule* on the output of rule learning as the object of our study. Focusing on simple artefacts – individual rules – as opposed to entire models such as rule sets or rule lists allows deeper, more focused analysis since rule is a small self-contained item of knowledge. Making a small change in one rule, such as adding a new condition, allows to test the effect of an individual factor that can influence perception of rule plausibility. In this section, we will shortly introduce the inductively learned rule in the machine learning and cognitive science contexts.

### 3.1.1. Decision Rules in Machine Learning

The type of inductively learned decision rule which we consider is depicted in Figure 3.1 and follows the notation used in Fürnkranz et al. [2012, page 25], an authoritative monograph on rule learning, where *confidence* and *support* are two selected rule evaluation measures.

```
IF A AND B THEN C,   confidence=c and support=s
IF veil is white AND odour is foul THEN mushroom is poisonous
 confidence = 90%, support = 5%
```

Figure 3.1.: Inductively learned rule

In this simple example, $A, B, C$ are *literals*, which are composed of attribute name (veil) and its value (white). Literal is sometimes referred to as *condition* throughout the text. The conjunction of conditions on the left side of the rule is called *antecedent*, the single literal predicted by the rule is called *consequent*.

While rule definition used in Fürnkranz et al. [2012, page 25] is restricted to conjunctive rules, other definitions, e.g. the formal definition given by Slowinski et al. [2006, page 2] allows also negation and disjunction as connectives. In the practical part of the thesis (crowdsourcing experiments in Chapter 5 and software framework in Part II) we adhere to the conjunctive definition. The analytical part (Part I) discusses also the implications that the inclusion of disjunction and negation would have on the comprehensibility of rules.

Rules on the output of rule learning algorithms are most commonly characterized by the following two parameters: *confidence* and *support*. The confidence of a rule is defined as $a/(a+b)$, where $a$ is the number of correctly classified objects, i.e. those matching rule antecedent as well as rule consequent, and $b$ is the number of misclassified objects, i.e. those matching the antecedent, but not the consequent. The support of a rule is defined either as $a/n$, where $n$ is the number of all objects (relative support), or simply as $a$ (absolute support).

Some rule learning frameworks, in particular association rule learning [Fürnkranz et al., 2012, page 14], which we build upon in Part II of the thesis, require the user to set thresholds for minimum confidence and support. Only rules with confidence and support values meeting or exceeding these thresholds are included on the output of rule learning and presented to the user.

### 3.1.2. Inductively Learnt Rules vs Rules in Cognitive Science

Inductively learned rules are a commonly embraced model of human reasoning in cognitive science [Smith et al., 1992, Nisbett, 1993, Pinker, 2015]. Rules also closely relate to Bayesian inference, another frequently used model of human reasoning.

Inductively learnt rule "IF A AND B THEN C" can be interpreted as a hypothesis corresponding to the logical implication $A \wedge B \Rightarrow C$. We can express the plausibility of such hypothesis in terms of Bayesian inference as the conditional probability $P(C|A,B)$. This corresponds to the confidence of the rule, a term used in rule learning, and to *strength of evidence*, a term used by cognitive scientists [Tversky and Kahneman, 1974].[1]

Given that $P(C|A,B)$ is a probability estimate computed on a sample, another relevant piece of information for determining the plausibility of the hypothesis is the robustness of this estimate. This will correspond to the number of observed instances for which the rule is true. The size of the sample (typically expressed as ratio) is known as rule support in machine learning and as *weight of the evidence* in cognitive science [Tversky and Kahneman, 1974].

[1] In the terminology used within the scope of cognitive science [Griffin and Tversky, 1992], confidence corresponds to the *strength* of the evidence and support to the *weight* of the evidence. Interestingly, this problem was already mentioned by Keynes [1922] (according to Camerer and Weber [1992]) who drew attention to the problem of balancing the likelihood of the judgment and the weight of the evidence in the assessed likelihood.

## 3.2. Cognitive Bias

In the introduction, we loosely defined cognitive biases as seemingly irrational reasoning patterns that allow humans to make fast and risk averse decisions. In this section, we further elaborate the definition of this pivotal concept.

#### 3.2.0.1. Cognitive Bias (Illusion)

According to the Encyclopedia of Human Behavior [Mata, 2012], the term cognitive bias was introduced in the 1970s by Amos Tversky and Daniel Kahneman, and is defined as:

> Systematic error in judgment and decision-making common to all human beings which can be due to cognitive limitations, motivational factors, and/or adaptations to natural environments.

The research on cognitive biases and heuristics is considered as the most important psychological research done in the past 40 years [Mata, 2012].

The narrow initial definition of cognitive bias as a shortcoming of human judgment was criticized by German psychologist Gerd Gigerenzer, who started in the late 1990s the "Fast and frugal heuristic" program to emphasize ecological rationality (validity) of cognitive biases. If cognitive bias is applied in the right environment, it results in "frugal" rather than "erroneous" judgment.

As for terminology, the concept of cognitive bias includes many cognitive phenomena, multiple of which are not called "biases" but instead *heuristics* (e.g. Representativeness heuristic), effects (e.g. Mere exposure effect), fallacies (e.g. Conjunction fallacy), illusions (e.g. Illusionary correlation) or otherwise.

Three types of cognitive biases are recognized in the recent authoritative work of Pohl [2017]: those relating to *thinking*, *judgment* and *memory*. Within this thesis, we focus on cognitive biases[2] falling into the *thinking* and *judgment* categories.

## 3.3. Objectives and Contribution

According to the seminal work of Michalski [1983] the goal of inductive rule learning is to create models that are comprehensible to humans. Previous efforts to improve comprehensibility of machine learning results have concentrated on making models smaller, mainly following the proposition that smaller models take less human time to process and provide less opportunities for errors. The problem of cognitive biases affecting comprehension of rule-based models has not yet been studied. However, there are several successful studies of the effect of cognitive biases on human interaction with other types of information [Kunda, 1999, Serfas, 2011].

In the following, we will introduce the goals of the present research:

[2] Called cognitive illusions in [Pohl, 2017].

**1) Analyze results of prior research to determine effect of selected cognitive biases on interpretability of rules discovered from data (Chapter 4).** Our objective is to identify applicable research from cognitive psychology and relate it to the problem of human interaction with rule learning results. Since model length is one of the few generic parameters that a particular learning algorithm can influence, and has thus been subject of most previous studies of comprehensibility of machine learning models, our research will also focus on model length. Since we decided to work with rule-based algorithms, we will specifically study how the effect of cognitive biases will vary depending on rule length.

**2) Experimentally determine effect of selected cognitive biases on interpretability of rules discovered from data (Chapter 5).** The aim of our experiments is to verify the effect of selected cognitive biases when actual human users are faced with rules discovered from real data. This is achieved by performing user study. Based on statistical analysis of the results we determine which cognitive biases affect interpretation of rules learnt from data. Particular attention will be paid to the problem of length of rules.

**3) Create concise model for effect of cognitive biases on interpretability of rules discovered from data (Chapter 6).** The last objective of the first part of the thesis is to merge the results of the qualitative analysis and obtained empirical results into one model. The outcome of this research is a list of implementable recommendations for designers of machine learning algorithms and interfaces.

**4) Propose machine learning algorithm that learns more comprehensible rule models (Part II).** The conclusion of research related to the previous objectives was that indeed shorter length of the model provides less opportunities for various cognitive biases to be triggered. The final objective of the thesis is to propose a machine learning algorithm that will reflect this finding to improve understanding of rule-based classification models discovered from data. A detailed definition of this objective is deferred to the first chapter of Part II of the thesis (Chapter 8).

Note that the fulfillment of these objectives leading to the list of contributions is summarized in the Conclusions (Chapter 11).

# 4. Analysis of Cognitive Science Literature

This chapter reviews selected cognitive biases (illusions) described in the field of cognitive science and applies them to the problem of human comprehension of rules discovered from data. The aim of this work is to understand the deviations between how the rule learning algorithm "means" a specific rule on its output and how humans perceive it.

**Focus of the analytical work.** The first objective of this thesis is to evaluate the effects of selected cognitive biases on interpretation of rules mined from data. Since these can demonstrate in multiple ways, we chose a narrow problem of the effect on the perceived plausibility of rules depending on their length, while we also pay attention to other effects. The advantages of focusing on rule length are:

- Rule length is one of few directly measurable properties common to all rules.
- Through selection of appropriate learning heuristic, the rule induction learning algorithm can influence the length of rules on its output. The experiments presented by Stecher et al. [2014] show that the length of rules is inversely related to the number of rules on the classifier output. It is open question what is more understandable to users: whether a smaller number of longer rules or larger number of shorter rules.
- Change in comprehension of rules depending on their length is measurable (this property will be used in the following chapter devoted to empirical analysis).
- According to our preliminary analysis, the effect of most biases is related to the length of the rules: increased length of the rule means extra stimuli the subject is exposed to.

While we mainly focus on the relation of the effect of biases to rule length, we also investigate the triggering of selected biases by other properties of rules, especially confidence and support (cf. the example in the box on page 36).

**Chapter organization.** Section 4.1 discusses the functions and environmental validity of cognitive biases and the way in which we work with them in our analysis. Section 4.2 specifies the selection criteria for inclusion of cognitive biases into our analysis and enumerates applicable cognitive biases and heuristics. The list of selected cognitive biases along with a short summary is presented in Section 4.3. The detail analysis of the implications of twenty

individual cognitive biases and heuristics for plausibility of inductively learned rules is split into three sections: Section 4.4 (biases relating to thinking), Section 4.5 (biases relating to judgment) and Section 4.6 (not categorized biases). Finally, Section 4.7 summarizes the results for all three types of biases and provides link to the following chapter focused on empirical evaluation.

**Example. (The significance of insensitivity to sample size effect for interpretation of rules discovered from data)**
Consider the following two rules:

- `IF a film is released in 2006 AND the language of the film is English THEN Rating is good, confidence = 80%, support = 10%.`
- `IF a film is released in 2006 AND the director was John Smith THEN Rating is good, confidence = 90%, support = 1%.`

It is well known in machine learning that rules with high confidence can appear on output of rule learning by chance [Azevedo and Jorge, 2007]. For this reason, the rule learning process typically outputs both confidence and support for the analyst to make an informed choice about merits of each rule. In our example both rules are associated with values of confidence and support to inform about the strength and weight of evidence for both rules. While the first rule is less strong (80% vs 95% correct), its weight of the evidence is ten times higher than of the second rule.
According to the *insensitivity to sample size effect* [Tversky and Kahneman, 1974] there is a systematic bias in human thinking that makes humans put higher weight on the strength of evidence (confidence) than on the weight of evidence (support). It has been shown that this bias is applicable also to statistically sophisticated psychologists [Tversky and Kahneman, 1971] and thus can be applicable to the widening number of professions that are using rule learning to obtain insights from data.
The second bias that we consider is *base rate fallacy*, according to which people are unable to correctly process conditional probabilities. The conditional probability in our example is the confidence value, which is a probability of good rating on the condition of release in 2006 and the film being in English.
The analysis of relevant literature from cognitive science not only reveals applicable biases, but also provides in some cases methods for limiting their effect (*debiasing*). The common way to express rule confidence and support metrics is to use ratios, as in our example. Extensive research has shown that if natural numbers are used instead then the number of errors in judgment drops [Gigerenzer and Goldstein, 1996, Gigerenzer and Hoffrage, 1995]. Reflecting these suggestions, the first rule in our example could be presented as follows:

- `IF a film is released in 2006 AND the language of the film is English THEN Rating is good. In our data, there are 100 movies which match the conditions of this rule. Out of these, 80 are predicted correctly as having good rating.`

## 4.1. Functions and Environmental Validity

In the introduction, we briefly characterized cognitive biases as "seemingly irrational reasoning patterns that are thought to allow humans to make fast and risk averse decisions." In fact, the function of cognitive biases is subject of scientific debate. According to the review of functional views in Pohl [2017], there are three fundamental positions among researchers. First group considers them as dysfunctional errors of the system, second group as faulty by-products of otherwise functional processes and the third group c) as adaptive and thus functional responses. According to Pohl [2017] most researchers are in the second group, where cognitive biases (illusions) are consider to be "built-in errors of the human information-processing systems".

It follows that biases and heuristics that humans apply are not generally interpreted as a lack of rational reasoning. Instead, these are considered as strategies that evolved to improve the fitness and chances of survival of the individual in particular situations. This stand in defense of biases is succinctly expressed by the influential work of Haselton and Nettle [2006]: "Both the content and direction of biases can be predicted theoretically and explained by optimality when viewed through the long lens of evolutionary theory. Thus, the human mind shows good design, although it is design for fitness maximization, not truth preservation." In our analysis of cognitive biases, we adopt this "built-in error" view and we try to identify measures that can correct the error by proposing debiasing techniques.

Empirical evidence also shows that cognitive biases are triggered or their effect strengthened by environmental cues and context [Haselton and Nettle, 2006]. However, as also follows from the quotation above, the application of cognitive biases comes at the cost of distorted perception of truth.

The interpretation of machine learning results and statistical hypotheses represents a new type of environment to which humans had only tiny bit of evolutionary time to adapt. It is therefore natural that when interpreting machine learning results, the human mind applies many of the heuristics and biases inappropriately. It follows that cognitive biases will not demonstrate in all humans and in all situations equally (or many times even at all). For this reason, in our analysis we put special attention to identifying groups of people who are (not)susceptible to the specific bias where this information is available. Also, we report the success rates – the number of people committing a fallacy corresponding to a specific bias in an experiment – if the information is available.

## 4.2. Selection Criteria and Limitations

Number of cognitive biases have been discovered. As Pohl [2017] in a recent authoritative book on cognitive illusions states: "There is a plethora of phenomena showing that we deviate in our thinking, judgment and memory from some objective and arguably correct standard." As a response to the high number of biases being discovered, several categorizations for their organization were proposed [Stanovich, 2009, Pohl, 2017]. In this chapter,

we opted to adhere to the three categories (thinking, judgment, memory) used by Pohl [2017], the most comprehensive review of cognitive biases to date. Furthermore, we focus on biases relating to judgment and thinking. While biases relating to memory are applicable to rule learning, they are mostly general phenomena with little specific bond to rules as subject of our study.

There are at least 51 different biases falling into the thinking and judgment categories [Evans et al., 2007, Pohl, 2017]. To select twenty biases into our review the author used his prior experience gained in the domain of designing machine learning algorithms [Kliegr, 2009, Kliegr et al., 2014], elicitation and handling of domain knowledge from experts for the purpose of improving machine learning results [Kliegr et al., 2011], interfaces for rule learning algorithms [Škrabal et al., 2012], designing systems for implicit learning of human preferences [Kuchař and Kliegr, 2013, Kliegr and Kuchař, 2014, Kuchař and Kliegr, 2014, Kliegr and Kuchař, 2015, Kuchař and Kliegr, 2017] and their empirical validation [Leroy et al., 2014].

One limitation of our analysis is that we do not consider the correlation between individual cognitive biases. For example, it is known that a number of cognitive biases (such as conjunction fallacy, base rate neglect, insensitivity to sample size, confusion of the inverse) can all be attributed to a more general phenomenon called representativeness heuristic [Ballin et al., 2008]. To our knowledge, the correlation between cognitive biases has not yet been systematically studied. Since in our analysis we built upon prior research in cognitive science, we decided also not to address this problem. In our review, we thus include multiple biases even though they may be correlated.

The list of the cognitive biases that we identified as important for interpretation of rule learning results follows.

## 4.3. Selected Cognitive Biases

This section contains a list of selected cognitive biases along with short description grouped by categories presented in Pohl [2017]. Not all of the included biases are assigned a category (Thinking, Judgment, Memory) by Pohl [2017]. These are listed under the "Other" category as the author does not have the confidence to assign these biases to any specific category.

### Thinking.

- *Base rate neglect [Kahneman and Tversky, 1973, Bar-Hillel, 1980].* Insensitivity to the prior probability of the outcome, violating the principles of probabilistic reasoning, especially Bayes' theorem.
- *Confirmation bias and positive test strategy [Nickerson, 1998].* Seeking or interpretation of evidence so that it conforms to existing beliefs, expectations, or a hypothesis in hand.
- *Conjunction fallacy and representativeness heuristic [Tversky and Kahneman, 1983].* Conjunction fallacy occurs when a person assumes that a specific condition is more

probable than a single general condition in case the specific condition seems as more representative of the problem at hand.

**Judgment.**

- *Availability heuristic [Tversky and Kahneman, 1973].* The easier it is to recall a piece of information, the greater the importance of the information.
- *Effect of difficulty [Griffin and Tversky, 1992].* If two mutually exclusive alternative hypotheses are evaluated and telling which one is better is difficult – both are nearly equally probable – people will grossly overestimate the confidence associated with their choice. This effect is also sometimes referred to as overconfidence effect [Pohl, 2017].
- *Mere-exposure effect [Zajonc, 1968].* Repeated encounter results in increased preference.

**Other.**

- *Ambiguity aversion [Ellsberg, 1961].* People tend to favour options for which the probability of a favourable outcome is known over options where the probability of favourable outcome is unknown. Some evidence suggests that ambiguity aversion has a genetic basis [Chew et al., 2012].
- *Averaging heuristic [Fantino et al., 1997].* Joint probability of two events is estimated as an average of probabilities of the component events. This fallacy corresponds to believing that $P(A, B) = \frac{P(A)+P(B)}{2}$ instead of $P(A, B) = P(A) * P(B)$.
- *Confusion of the inverse [Plous, 1993].* Conditional probability is equivocated with its inverse. This fallacy corresponds to believing that $P(A|B) = P(B|A)$.
- *Context and tradeoff contrast [Tversky and Simonson, 1993].* The tendency to prefer alternative $x$ over alternative $y$ is influenced by the context – other available alternatives.
- *Disjunction fallacy [Bar-Hillel and Neter, 1993].* People tend to find it as more likely for an object to belong to a more characteristic subgroup than to its supergroup.
- *Information bias [Baron et al., 1988].* Distorted evaluation of information: believing that more information the better, even if the extra information is irrelevant for the decision.
- *Insensitivity to sample size [Tversky and Kahneman, 1974].* Neglect of the following two principles: a) more variance is likely to occur in smaller samples, b) larger samples provide less variance and better evidence.

- *Recognition heuristic [Goldstein and Gigerenzer, 1999].* If one of two objects is recognized and the other is not, then infer that the recognized object has the higher value with respect to the criterion.
- *Negativity bias [Kanouse and Hanson Jr, 1987].* People weigh negative aspects of an object more heavily than positive ones.
- *Primacy effect [Thorndike, 1927].* This effect can be characterized by words of Edward Thorndike (1874-1949), one of the founders of modern education psychology, as follows: "other things being equal the association first formed will prevail" [Thorndike, 1927].
- *Reiteration effect [Hasher et al., 1977].* Frequency of occurrence is a criterion used to establish validity of a statement.
- *Unit bias [Geier et al., 2006].* People tend to consider each condition as a unit of equal weight at the expense of detailed scrutiny of the actual weight of the condition.
- *Weak evidence effect [Fernbach et al., 2011].* Presenting weak, but supportive evidence makes people less confident in predicting a particular outcome than presenting no evidence at all.

The cognitive biases listed above are in detail analysed with respect to their effect on interpretability of rule learning results in the following Sections 4.4-4.6. For all cognitive biases, we include a short description and a paragraph, which quantifies the effect of the cognitive bias and, where this information was available, we state the proportion of subjects committing it. As noted earlier, the particular problem that we study is perceived plausibility of rules depending on their length. For all cognitive biases we suggest a debiasing technique that could be effective in the rule learning context. The suggestions are based on empirical results obtained by psychologists, we indicate when these are our conjectures that are in need of further validation.

## 4.4. Thinking

This section discusses the cognitive biases in the Thinking category.

### 4.4.1. Base-rate Fallacy

The *base-rate fallacy* indicates that people are unable to correctly process conditional probabilities.

#### 4.4.1.1. Success Rates

In the original experiment reported in Kahneman and Tversky [1973] more than 95% of psychology graduate students committed the fallacy.

#### 4.4.1.2. Implications for Rule Learning

The application of the base rate fallacy suggests that when presented with two otherwise identical rules with different values of confidence and support metrics, analyst's preferences will be primarily shaped by the confidence of the rule.

It follows that by preferring higher confidence, base-rate fallacy will generally contribute to positive relation between rule length and plausibility, since the longer rule can better fit a particular group in data and thus have a higher confidence than a more general shorter rule. At the same time, the more specific longer rule will have smaller value of support.

#### 4.4.1.3. Debiasing Techniques

Literature review provides a valuable input for addressing base-rate fallacy. Gigerenzer and Hoffrage [1995] show that representations in terms of natural frequencies, rather than conditional probabilities, facilitate the computation of cause's probability. To the author's knowledge, confidence is exclusively presented as ration in current software systems. The support rule quality metric is sometimes presented as a ratio and sometimes as a natural number. It would foster correct understanding if analysts are consistently presented with natural frequencies in addition to ratios.

### 4.4.2. Confirmation Bias and Positive Test Strategy

Confirmation bias is the best known and most widely accepted notion of inferential error of human reasoning [Evans, 1989, p. 552].[1] This bias refers to the notion that people tend to look for evidence supporting the current hypothesis, disregarding conflicting evidence. Research suggests that even neutral or unfavourable evidence can be interpreted to support existing beliefs, or as Trope et al. [1997, p. 115-116] put it "the same evidence can be constructed and reconstructed in different and even opposite ways, depending on the perceiver's hypothesis."

A closely related heuristic is the *Positive Test Strategy* (PTS) proposed by Klayman and Ha [1987]. This heuristic suggests that when trying to test a specific hypothesis, people examine cases which they expect to confirm the hypothesis rather than the cases which have the best chance of falsifying it. The difference between PTS and confirmation bias is that PTS is applied to test a "candidate" hypothesis while the true confirmation bias tests hypotheses that are already established [Pohl, 2004, p. 93].

Regarding PTS, experimental results of Klayman and Ha [1987] show that under realistic conditions it can be a very good heuristic for determining whether a hypothesis is true or false, but it can also lead to systematic errors if applied to an inappropriate task.

Finally, it should be noted that according to review performed by Klayman and Ha [1987] this heuristic is used as a "general default heuristic" in situations where either there is absence of specific information that identifies some tests as more relevant than others or when the cognitive demands of the task prevent a more careful strategy.

[1] Cited according to Nickerson [1998].

#### 4.4.2.1. Success Rates

According to Mynatt et al. [1977, p. 404] in 70% of cases subjects did not abandon falsified hypotheses in an experiment that simulated research environment.[2] This success rate is particularly relevant for the problem of comprehending rule learning results as the simulated research environment is close to our target domain of analysts interpreting discovered rules.

#### 4.4.2.2. Implications for Rule Learning

This bias can have significant impact depending on the purpose for which the rule learning results are used. If the analyst had some prior hypothesis before she obtained the rule learning results, according to the confirmation bias she will "cherry pick" the rules confirming this prior hypothesis, disregarding rules that contradict it. Given that output of some rule learners contains contradicting rules, the analyst can select only the rules conforming to the hypothesis disregarding applicable rules with opposite conclusion, which would otherwise be considered as more relevant.

Using evidence gathered using MRI brain scans Westen et al. [2006] explain confirmation bias by emotions related to the favoured hypothesis. The evidence that challenges this hypothesis is suppressed. The experiments in this study were conducted by presenting information that challenged the moral integrity of the politician that the subject favoured. While it could be argued that data analysts interpreting the rule learning results are free of emotional bond with the problem and additionally trained to correctly interpret machine learning results, the confirmation bias may still apply to them:

- Stanovich et al. [2013] show that incidence of myside bias, which is closely related to confirmation bias, is surprisingly not related to intelligence. This suggests that even highly intelligent analysts can be affected.
- Some research can be interpreted so that data analysts can be even more susceptible to the myside bias than the general population. Experiment reported by Wolfe and Britt [2008] shows that subjects who defined good arguments as those that can be "proved by facts" (this stance, we assume, would also apply to many data analysts) were more prone towards the myside bias.[3]

#### 4.4.2.3. Debiasing Techniques

Tweney et al. [1980] successfully tested a modification of the Wason's 2,4,6 task. In the original Wason's 2,4,6 task participants try to "discover" the rule according to which the sequence 2,4,6 was created (for details cf. Section 2.5.2). The correct answer is ascending sequence of numbers. In the modification by Tweney et al. [1980] participants were asked

[2] Result for experiment performed in "complex" environment.

[3] This tendency is explained as follows: "For people with this belief, facts and support are treated uncritically. The intended audience is not part of the schema and thus ignored. More importantly, arguments and information that may support another side are not part of the schema and are also ignored."

to search for two rules ("any ascending sequence of numbers" and "all other sequences") instead of one rule ("ascending sequence of numbers"). Following this, the response format was changed from positive and negative to whether the rule belongs to the first category "DAX" or the second category "MED", which improved performance in the task. This relabeling of categories from positive and negative to something more neutral can possibly help to debias the analysts interpreting rule learning results in binary rule learning tasks.

Albarracín and Mitchell [2004] suggest that the susceptibility to the confirmation bias can dependent on one's personality traits. This publication also presents a diagnostic tool called "defense confidence scale" that can identify individuals prone to confirmational strategies.

Wolfe and Britt [2008] successfully experimented with providing the subjects with explicit guidelines for considering evidence both for and against the hypothesis. While this research is not directly related to hypothesis testing, providing explicit guidance combined with modifications of the user interface of the system presenting the rule learning results could prove as an effective debiasing technique.

### 4.4.3. Conjunction Fallacy and Representativeness Heuristic

Human-perceived plausibility of hypotheses has been extensively studied in cognitive science. The most well-known cognitive phenomenon related to our focus area of the influence of the number of conditions in a rule on its plausibility is the *Conjuctive fallacy*. This fallacy falls into the research program on cognitive biases and heuristics carried out by Amos Tversky and Daniel Kahneman since approximately 1970s'. The outcome of this research program can be succinctly summarized by a quotation from Kahneman's Nobel Prize lecture which was delivered at Stockholm University on December 8, 2002 [Kahneman, 2003]:

> ..., it is safe to assume that similarity is more accessible than probability, that changes are more accessible than absolute values, that averages are more accessible than sums, and that the accessibility of a rule of logic or statistics can be temporarily increased by a reminder.

This heuristic relates to the tendency to make judgments based on similarity, based on rule "like goes with like", which is typically used to determine whether an object belongs to a specific category. According to Gilovich and Savitsky [2002] representativeness heuristic can be held accountable for number of widely held false and pseudoscientific beliefs, including those in astrology or graphology.[4] It can also inhibit valid beliefs that do not meet the requirements of resemblance.

[4] Gilovich and Savitsky [2002] give the following example: resemblance of the physical appearance of the sign, such as crab, is related in astrology with personal traits, such as appearing tough on the outside. For graphology, the following example is given: handwriting to the left is used to indicate that the person is holding something back.

#### 4.4.3.1. Linda Problem

Conjunctive fallacy is in the literature often defined through the "Linda" problem [Tversky and Kahneman, 1983, page 299], which was first used to demonstrate it.[5]

```
Linda is 31 years old, single, outspoken, and very bright.
She majored in philosophy. As a student, she was deeply
concerned with issues of discrimination and social justice,
and also participated in anti-nuclear demonstrations.

Which is more probable?

(a) Linda is a bank teller.
(b) Linda is a bank teller and is active in the
    feminist movement.
```

Figure 4.1.: Linda problem

In the Linda problem (Figure 4.1), subjects are asked to compare conditional probabilities $P(F\&B|L)$ and $P(B|L)$, where $B$ refers to "bank teller", F to "active in feminist movement" and $L$ to the description of Linda [Bar-Hillel, 1991].

Multiple studies have shown that humans tend to consistently select the second, longer hypothesis, which is in conflict with the elementary law of probability: *the probability of a conjunction,* $P(A\&B)$*, cannot exceed the probability of its constituents,* $P(A)$ *and* $P(B)$ [Tversky and Kahneman, 1983]. In other words, it always holds for the Linda problem that

$$P(F\&B|L) \leq P(B|L).$$

Preference for alternative $F\&B$ (option b in Figure 4.1) is thus always a logical fallacy.[6]

Conjunction fallacy has been shown to hold across multiple settings (hypothetical scenarios, real-life domains), as well as for various kinds of subjects (university students, children, experts, as well as statistically sophisticated individuals) [Tentori and Crupi, 2012].

#### 4.4.3.2. Cause of Conjunctive Fallacy - Representativeness Heuristic and Other Explanations

The results of the conjunctive fallacy experiments manifest according to Tversky and Kahneman [1983] that *conjunction can be more representative than one of its constituents.* Conjunctive fallacy is a symptom of a more general phenomenon, in which people have a

[5] Note that the paper [Tversky and Kahneman, 1983] contains also different set of eight answer options for the Linda problem on page 297. The two option version on page 299 is prevalently used as a canonical version of the Linda problem in subsequent research (cf. the seminal paper of Gigerenzer [1996, page 592]), and is referred to by Daniel Kahneman as the "more direct version" of the Linda problem [Kahneman, 2003, page 712].

[6] $P(F\&B|L)$ is a notation used in cognitive science (e.g. Bar-Hillel [1991]). In computer science a corresponding notation would be $P(F, B|L)$.

tendency to overestimate the probabilities of representative events and underestimate those of less representative ones. The reason is attributed to the application of the *representativeness heuristic* [Tversky and Kahneman, 1983]. This heuristic provides humans with means for assessing a probability of an uncertain event. This is used to answer questions such as "What is the probability that object A belongs to class B? What is the probability that event A originates from process B?" According to the representativeness heuristic, *probabilities are evaluated by the degree to which A is representative of B, that is by the degree to which A resembles B* [Tversky and Kahneman, 1974].

Representativeness heuristic is not the only explanation for the results of the conjunctive fallacy experiments. Hertwig et al. [2008] hypothesized that the reason is caused by "a misunderstanding about conjunction", in other words by a different interpretation of "probability" and "and" by the subjects than assumed by the experimenters. The validity of this alternate hypothesis has been subject to criticism [Tentori and Crupi, 2012], nevertheless the problem of correct understanding of "and" exists and is of particular importance to machine learning. Another proposed hypothesis for explaining the conjunctive fallacy is the averaging heuristic [Fantino et al., 1997].

#### 4.4.3.3. Success Rates

Tversky and Kahneman [1983] report that 85% of the subjects indicate (b) as the more probable option for the Linda problem, which is defined in Figure 4.1. It should be noted that the actual proportion may vary, 83% are reported when the experiment was replicated by Hertwig and Gigerenzer [1999], and 58% when replicated by Charness et al. [2010].

#### 4.4.3.4. Implications for Rule Learning

Rules are not composed only of conditions, but also of an outcome (value of a target variable). A higher number of conditions generally allows the rule to filter a purer set of objects with respect to the value of the target variable than a smaller number of conditions. This means that conjunctive fallacy does not directly manifest when interpreting rule learning results since it cannot be stated that selection of a longer rule is a reasoning error in the rule learning context, even in cases when the set of conditions of the longer rule subsumes the set of conditions of the shorter rule. Nevertheless, application of representativeness heuristic can affect human perception of rule plausibility.

#### 4.4.3.5. Debiasing Techniques

Number of factors that decrease the ratio of subjects exhibiting conjunctive fallacy – as an undesired consequence of representativeness heuristic when its application is not rational – has been identified:

- Charness et al. [2010] found that the number of committed fallacies is reduced under monetary incentive. Addition of a monetary incentive is reported to drop the fallacy

rate to 33 %. The observed rate under a monetary incentive better hints at smaller importance of this problem for real-life decisions.

- Zizzo et al. [2000] found that unless the decision problem is simplified neither monetary incentive nor feedback ameliorate the fallacy rate. Reducing task complexity is a precondition for monetary incentives and feedback to be effective.
- Stolarz-Fantino et al. [1996] observed that the number of fallacies is reduced but still strongly present when the subjects receive training in logics.
- Gigerenzer and Goldstein [1996], Gigerenzer and Hoffrage [1995] show that the number of fallacies can be reduced or even eliminated by presenting the problems in terms of frequency rather than probability.

## 4.5. Judgment

This section discusses the cognitive biases in the Judgment category.

### 4.5.1. Availability Heuristic

The availability heuristic is a judgmental heuristic in which a person evaluates the frequency of classes or the probability of events by the ease with which relevant instances come to mind. This heuristic is explained by its discovers, Tversky and Kahneman [1973], as follows: "That associative bonds are strengthened by repetition is perhaps the oldest law of memory known to man. The availability heuristic exploits the inverse form of this law, that is, it uses the strength of the association as a basis for the judgment of frequency."

To determine availability, it is sufficient to assess the ease with which instances or associations could be brought to mind – it is not necessary to perform the actual operations of retrieval or construction. An illustration of this phenomenon by Tversky and Kahneman [1973] is: "One may estimate the probability that a politician will lose an election by considering the various ways he may lose support."

#### 4.5.1.1. Success Rates

Success rates for availability heuristics are very varied and depend greatly on the experiment setup. Among other factors, they depend on the ease of recall [Schwarz et al., 1991]. In one of the original experiments (judgment of word frequency) presented by Tversky and Kahneman [1973], the number of wrong judgments was 105 out of 152 (70%). The task was to estimate whether letter "R" appears more frequently on first or third position in English texts. The reason why most subjects incorrectly assumed the first position is that it is easier to recall words starting with R than words with R on the third position.

#### 4.5.1.2. Implications for Rule Learning

The application of availability heuristic is based on the perceived association between the literals in the antecedent and the consequent of the rule. The stronger this perceived association, the higher the perceived confidence of the rule. It is our opinion this heuristic will favour longer rules, since they have higher chance to contain a literal which the analyst perceives as associated with the predicted label.

It is true that the longer rule is also more likely to contain literals not perceived as associated. It can be argued that while the remaining weakly associated literals will decrease the preference for the longer rule, this effect can be attributed to the weak evidence heuristic rather than the availability heuristic. However, according to our literature review, the availability heuristic can have only positive effect on preference.

#### 4.5.1.3. Debiasing Techniques

Our initial review did not reveal any debiasing strategies. From the broader perspective, availability is associated with the associative System 1, which can be corrected by the rule-based System 2 [Kahneman, 2003]. Therefore, inducing conditions known to trigger engagement of System 2 could be effective.

### 4.5.2. Effect of Difficulty

When an analyst is supposed to give a preference judgment between two competing hypotheses, one of the factors used in the decision making process is the difficulty of the problem and the corresponding confidence that is related to the judgment.

Griffin and Tversky [1992] developed a model that combines the strength of evidence with its weight (credibility). Their main research finding is that people tend to combine strength with weight in suboptimal ways, resulting in the decision maker being too much or too little confident about the hypothesis at hand than would be normatively appropriate given the information available. This discrepancy between the normative confidence and the decision maker's confidence is called overconfidence or underconfidence. Research has revealed systematic patterns in overconfidence and underconfidence:

- If the estimated difference between the two hypotheses is large, it is easy to say which one is better, then there is a pattern of underconfidence.
- As the degree of difficulty rises (the difference between the normative confidence of two competing hypotheses is decreasing), there is a strengthening pattern of overconfidence.

People use the provided data to assess the hypothesis at hand but they insufficiently regard the quality of the data. Griffin and Tversky [1992] illustrate this manifestation of bounded rationality as follows: "If people focus primarily on the warmth of the recommendation with insufficient regard for the credibility of the writer, or the correlation between the predictor and the criterion, they will be overconfident when they encounter a glowing letter based on

casual contact, and they will be underconfident when they encounter a moderately positive letter from a highly knowledgeable source."

#### 4.5.2.1. Success Rates

Griffin and Tversky [1992] used regression to analyze the relation between the strength of evidence and weight of evidence. The conclusion was that the regression coefficient for strength was larger than the regression coefficient for weight for 30 out of 35 subjects, which was found statistically significant. The median ratio of these coefficients was established to be 2.2 to 1 in favour of strength.

#### 4.5.2.2. Implications for Rule Learning

The strongest overconfidence was recorded for problems where the *weight of evidence is low and the strength of evidence is high*. This directly applies to rules with high value of confidence and low value of support. These are typically the longer rules. The empirical results related to the effect of difficulty therefore suggest that the predictive ability of such rules will be substantially overrated by analysts.

#### 4.5.2.3. Debiasing Techniques

We conjecture that this effect can be ameliorated by filtering out rules that do not pass a statistical significance test from the output and informing the users on the value and meaning of the value of statistical significance.

### 4.5.3. Mere Exposure Effect

According to this heuristic (effect), repeated exposure to an object results in an increased preference for that object.

Mere exposure effect and recognition heuristic are according to Pachur et al. [2011] two separate phenomena, because unlike the former the mere exposure effect does not "require that the object is recognized as having been seen before".

#### 4.5.3.1. Success Rates

As with other biases, the success rates for mere exposure effect are very varied and depend greatly on the experiment setup. Among other factors, they depend on whether the stimuli the subject is exposed to is exactly the same as prior exposure or similar to it [Monahan et al., 2000]. Instead of selecting one particular success rate from a specific experiment, we can refer to the well-established finding that when a concrete stimulus is repeatedly exposed, preference for that concrete stimulus is increased logarithmically as a function of the number of exposures [Bornstein, 1989].

#### 4.5.3.2. Implications for Rule Learning

Already the initial research of Zajonc [1968] included experimental evidence on the correlation between word frequency and affective connotation of the word. From this it follows that a longer rule – as measured by word length rather than the number of conditions – will have a greater chance of containing a word that the analyst had been strongly exposed to. Additionally, the exposure effects of individual words may possibly be combined. This leads to the conclusion that mere exposure effect will increase plausibility of longer rules.

#### 4.5.3.3. Debiasing Techniques

While our limited literature review did not reveal any debiasing techniques, we conjecture that similarly to the related recognition heuristic the knowledge of the criterion variable could ameliorate the mere exposure effect: presenting information on the semantics of the literal as well as on its covariance with other literals may suppress the heuristic.

## 4.6. Other

This section discusses the cognitive biases not included in the categorization of cognitive biases present in Pohl [2017].

### 4.6.1. Ambiguity Aversion

Ambiguity aversion corresponds to the finding that humans tend to prefer known risks over unknown risks.

Ambiguity aversion is not a reasoning error. Consider the following comparison with conjunctive fallacy. When a typical subject is explained the conjunctive fallacy they will recognize their reasoning as an "error", and as Al-Najjar and Weinstein [2009] put it the subjects "feel embarrassed" for the *irrational* choice. This contrasts with the ambiguity aversion, as for example demonstrated by the Ellsberg paradox [Ellsberg, 1961].[7]

As follows from the research of Camerer and Weber [1992], ambiguity aversion is related to the information bias: demand for information in cases when it has no effect on decision can be explained by the aversion to ambiguity – "people dislike not having missing information".

#### 4.6.1.1. Success Rates

As noted by Camerer and Weber [1992], Ellsberg did not perform careful experiments. According to the same paper, follow-up empirical work can be divided into three categories: replication of the Ellsberg's experiment, determination of psychological causes of ambiguity and study of ambiguity in applied setting. The most relevant to the focus of our work are

[7] Ellsberg paradox: Humans tend to systematically prefer a bet with known albeit very small probability of winning over a bet with not precisely known probability of winning, even if it would in practice mean a near guarantee of winning.

experiments focusing on the applied setting. Curley et al. [1984] describe experiment in medical domain where 20% of subjects avoided ambiguous treatments.

#### 4.6.1.2. Implications for Rule Learning

The ambiguity aversion may have profound implications for rule learning. The typical data mining task will contain a number of attributes the analyst has no or very limited knowledge of. The ambiguity aversion will manifest by preference for rules that do not contain ambiguous (unknown) attributes or literals.

Ambiguity aversion steers the analyst to shorter rules as these can be expected to have lower chance of containing an ambiguous literal.

#### 4.6.1.3. Debiasing Techniques

We conjecture that this bias would be alleviated if textual description of the meaning of all the literals is made easily accessible to the analyst.

### 4.6.2. Averaging Heuristic

While the representativeness heuristic is the most commonly associated heuristic with the conjunctive fallacy, the averaging heuristics provides an alternate explanation: people evaluate the probability of a conjuncted event as the average of probabilities of the component events [Fantino et al., 1997].

The implications for conjunctive fallacy are succinctly summarized in Zizzo et al. [2000]: "Assume that P(A) = 20% and P(B) = 30%. Then the average is 25%, and any slight random deviation (by 5%) may make the subject not commit the fallacy (by choosing 20% or less). Assume now that P(A) = 20% but P(B) = 80%. Then the average is 50%, and one requires a random tremble of 30% in order for the conjunction fallacy not to occur."

#### 4.6.2.1. Success Rates

As reported by Zizzo et al. [2000]: "approximately 49% of variance in subjects' conjunctions could be accounted for by a model that simply averaged the separate component likelihoods that constituted a particular conjunction." This high success rate suggests that the averaging heuristic may be an important subject of further study within machine learning.

#### 4.6.2.2. Implications for Rule Learning

Applied to the rule learning scenario, if the antecedent consists of conditions rated about equally likely, then the incidence of misinterpretation is low, since the average is close (or identical) to the correct computation of joint probability by multiplying the probabilities. However, should the conditions have substantially different probability, the application of the averaging heuristic can lead to larger divergence of the perceived probability and true probability.

The averaging heuristic can be interpreted to increase preference for longer rules. The reason is that longer rules are more likely to contain literals with low probability. Due to the application of the averaging heuristic the analyst may not fully realize the consequences of the presence of a low-probability literal for the overall likelihood of the set of conditions in the antecedent of the rule.

#### 4.6.2.3. Debiasing Techniques

Experiments presented in Zizzo et al. [2000] showed that prior knowledge of probability theory, and a direct reminder of how are probabilities combined, are effective tools for decreasing the incidence of conjunctive fallacy, which is the hypothesized consequence of the averaging heuristic.

### 4.6.3. Confusion of the Inverse

This effect corresponds to confusing the difference between the confidence of the rule – corresponding to $P(consequent|antecedent)$ – with $P(antecedent|consequent)$. This confusion may manifest itself strongest in the area of association rule learning, where an attribute can be of interest to the analyst both in the antecedent and consequent of a rule.

#### 4.6.3.1. Success Rates

In a study referenced from Plous [1993] this fallacy was committed by 95% of physicians involved.

#### 4.6.3.2. Implications for Rule Learning

We do not see a systematic bias the confusion of the inverse would pose for perceived plausibility of rules depending on their length.

#### 4.6.3.3. Debiasing Techniques

Edgell et al. [2004] studied the influence of the effect of training of analysts in probabilistic theory with the conclusion that it is not effective in addressing the confusion of the inverse fallacy. Our literature review did not reveal any other applicable work.

### 4.6.4. Context and Tradeoff Contrast

Tversky and Simonson [1993] developed a theory that combines *background* context defined by prior options with *local* context which is given by the choice problem at hand. The contributions of both types of context are additive. While additivity is considered as not essential for the model, it is included because it "provides a good approximation in many situations and because it permits a more parsimonious representation". The analyst adjusts the relative weights of attributes in the light of tradeoffs implied by the background.

The reference application scenario for tradeoff contrast is that selection of one of the available alternatives, such as products or job candidates, can be manipulated by the

addition or deletion of alternatives that are otherwise irrelevant. Tversky and Simonson [1993] attribute the tradeoff effect to the fact that "people often do not have a global preference order and, as a result, they use the context to identify the most 'attractive' option."

#### 4.6.4.1. Success Rates

In one of the experiments described by Tversky and Simonson [1993], subjects were asked to choose between two microwave ovens (Panasonic priced 180 USD and Emerson priced 110 USD), both third off the regular price. The number of subjects who chose Emerson was 57% and 43% chose Panasonic. Another group of subjects was presented the same problem with the following manipulation: A more expensive Panasonic valued at 200 USD (10% off the regular price) was added to the list of possible options. The newly added device was described to look as inferior to the other Panasonic, but not to the Emerson device. After this manipulation, only 13% chose the more expensive Panasonic, but the number of subjects choosing the less expensive Panasonic rose from 43% to 60%.

It should be noted that according to Tversky and Simonson [1993] if people have well-articulated preferences, the background context has no effect on the decision.

#### 4.6.4.2. Implications for Rule Learning

In rule learning context manipulation will not likely be deliberate but a systematic result of the algorithmic process. It will manifest by presence of redundant rules or attributes within rules on the output.

The influence of context can be manifested by preference towards longer rules. The reason is that if a rule contains a literal with unknown predictive power and multiple other literals with known (positive) predictive power for the consequent of the rule, these known literals create a context which may make the analyst believe that also the unknown literal has positive predictive power. By doing so, the context provided by the longer rule can soften the effects of ambiguity aversion, which would otherwise have made the analyst prefer the shorter rule (cf. Subsection 4.6.1), and through the information bias (cf. Subsection 4.6.6) further increase the preference for the longer rule.

#### 4.6.4.3. Debiasing Techniques

We conjecture that similarly to other effects, the influence of context can be suppressed by reducing the number of rules the analyst is presented and removal of irrelevant literals from the remaining rules.

### 4.6.5. Disjunction Fallacy

Disjunction fallacy is demonstrated by assessing probability $P(X)$ as higher than probability $P(Z)$, where $Z$ is a union of event X with another event Y. Bar-Hillel and Neter

[1993] explain the disjunction fallacy with preference for the narrower possibility over the broader one. In case the narrower category is unlikely, the broader possibility is preferred.

In experiments reported by Bar-Hillel and Neter [1993], $X$ and $Z$ were nested pairs of categories, such as Brazil and Latin America. Subjects were assigned problems such as: "Writes letter home describing a country with snowy wild mountains, clean streets, and flower decked porches. Where was the letter written?" It follows that since Latin America contains Brazil, the normative answer is Latin America. However, Brazil was the most likely answer.

#### 4.6.5.1. Success Rates

The rate of the disjunction fallacy in experiment presented by Bar-Hillel and Neter [1993] averaged 64%. The authors offer two explanations for why this is a lower fallacy rate than for the conjunction fallacy. The first one is that the disjunction rule is more compelling than the conjunction rule. The second favoured explanation is that the Linda experiments in Tversky and Kahneman [1983] used highly non-representative categories (bank teller), while in [Bar-Hillel and Neter, 1993] both levels of categories (Brazil and Latin America) were representative.

#### 4.6.5.2. Implications for Rule Learning

In data mining context, it can be the case that the feature space is hierarchically ordered. The analyst can thus be confronted with rules containing attributes (literals) on multiple levels of granularity. Following the disjunction fallacy, the analyst will generally prefer rules containing more specific attributes, which can result in preference for rules with fewer backing instances and thus weaker statistical validity.

The disjunction fallacy can be generally expected to bias the analysts towards longer rules since these have a higher chance of containing a literal corresponding to a narrower category.

#### 4.6.5.3. Debiasing Techniques

We conjecture that disjunction fallacy could be alleviated by making the analysts aware of the taxonomical relation of the individual attributes and educating them on the benefits of larger supporting sample, which is associated with more general attributes.

### 4.6.6. Information Bias

Information bias relates to the tendency of people to consider more available information to improve the perceived validity of a statement even if the additional information is not relevant. The typical manifestation of the information bias is evaluating questions as worth asking even when the answer cannot affect the hypothesis that will be accepted [Baron et al., 1988].

#### 4.6.6.1. Success Rates

Experiments 1-4 performed by Baron et al. [1988] show the effect of information bias. For example, in their Experiment 4 subjects were asked to assess to what degree a medical test is suitable for deciding which of the three diseases to treat using scale 0 to 100. The test detected a chemical "Tutone", which was with certain given probability associated with each of the three diseases. This probability was varied across the cases. There were ten cases evaluated, the test could normatively help only in two of those (no. 2 and 9) – the correct answer for the remaining eight was thus 0. For example, in case no. 1 and 10 the probability of Tutone being associated with all three diseases was equal – the knowledge of Tutone presence had no value for distinguishing between the three diseases – and the normative answer was 0. Even for these simple cases, the mean rating was 21 and 9 instead of 0. The normative answer for cases 2 and 9 was equal at 24, while the subjects assigned 61 and 75 respectively.

#### 4.6.6.2. Implications for Rule Learning

Rules often contain redundant, or nearly redundant conditions. By redundant it is meant that the knowledge of the particular piece of information represented by the additional condition (literal) has no or very small effect on rule quality. According to information bias, rule containing the additional (redundant) literal will be – following the information bias – considered as more preferred to a rule not containing this literal. The information bias clearly steers the analyst towards longer rules.

#### 4.6.6.3. Debiasing Techniques

We conjecture that this bias would be alleviated by a visualization of the information value (e.g. by predictive strength) of individual conditions in the rule.

### 4.6.7. Insensitivity to Sample Size

This effect implies that analysts are unable to appreciate the increased reliability of the confidence estimate with increasing value of support.

Unlike base-rate fallacy, this effect assumes that the size of the sample is understood: base-rate fallacy deals with the more complex case when people are presented with probabilistic information, but are unable to understand it correctly. Insensitivity to sample size is a connecting problem that relates to people underestimating the increased benefit of higher robustness of estimate made on a larger sample.

Another bias to which insensitivity to sample size is connected is the *frequency illusion*, which relates to overestimation of the base rate of an event as a result of selective attention and confirmation bias.

#### 4.6.7.1. Success Rates

When the insensitivity to sample size effect was introduced by Tversky and Kahneman [1974], it was supported by experimental results from the Hospital problem. This problem is formulated so that the subjects are asked which hospital is more likely to record more days when more than 60 percent of the babies born are boys. The options are larger hospital, smaller hospital or about the same. The correct expected answer – the smaller hospital – was chosen only by 22% of subjects, the fallacy rate is thus 78%. The experimental subjects were 95 undergraduate students.

#### 4.6.7.2. Implications for Rule Learning

Given that longer rules can fit specific regions of data, they can be higher on confidence and lower on support. This implies that if confronted with two rules, one of them will have slightly higher confidence and the second rule higher support, the analyst will according to this cognitive bias prefer the longer rule with higher confidence (all other factors equal).

#### 4.6.7.3. Debiasing Techniques

In our opinion, one possible approach for mitigation of this bias in rule learning research is using the value of support to compute confidence (reliability) intervals for the value of confidence. This confidence interval might be better understood than the original "raw" value of support.

### 4.6.8. Recognition Heuristic

Definition according to Pachur et al. [2011] is: "For two-alternative choice tasks, where one has to decide which of two objects scores higher on a criterion, the heuristic can be stated as follows: If one object is recognized, but not the other, then infer that the recognized object has a higher value on the criterion."

The recognition heuristic can be differentiated from the availability heuristic as follows: "To make an inference, one version of the availability heuristic retrieves instances of the target event categories, such as the number of people one knows who have cancer compared to the number of people who have suffered from a stroke [Hertwig et al., 2005]. The recognition heuristic, by contrast, bases the inference simply on the ability (or lack thereof) to recognize the names of the event categories." [Pachur et al., 2011].

#### 4.6.8.1. Success Rates

An experiment performed by Goldstein and Gigerenzer [1999] focused on estimating which of the two cities in the presented pair is more populated. The estimates were analysed with respect to the recognition of the cities by subjects. The median proportion of judgments complying to the recognition heuristic was 93%. It should be noted that the application of this heuristic is in this case ecologically justified since recognition will be related to how

many times the city appeared in a newspaper report, which in turn is related to the city size [Beaman et al., 2006].

#### 4.6.8.2. Implications for Rule Learning

The recognition heuristic can manifest itself by preference for rules containing a recognized literal or attribute in the antecedent of the rule. Since the odds that a literal will be recognized increase with the length of the rule, the recognition heuristic generally increases the preference for longer rules.

One could argue that for longer rules, the odds of occurrence of an unrecognized literal will also increase. The counterargument is the empirical finding that – under time pressure – people assign a higher value to recognized objects than to unrecognized objects. This happens also in situations when recognition is a poor cue [Pachur and Hertwig, 2006].

#### 4.6.8.3. Debiasing Techniques

As to the alleviation of effects of recognition heuristic in situations where it is ecologically unsuitable, Pachur and Hertwig [2006] note that suspension of the heuristic requires additional time or the direct knowledge of the "criterion variable". This coincides with the intuition that the interpretation of rule learning results by experts should be less prone to recognition heuristic. However, in typical real-world machine learning tasks the data can include a high number of attributes that even subject-matter experts are not acquainted with in detail. When these recognized – but not understood – attributes are present in the rule model even the experts are liable to the recognition heuristic. We therefore conjecture that the experts can strongly benefit from easily accessible information on the meaning of individual attributes and literals.

### 4.6.9. Negativity Bias

According to this bias[8] negative evidence (or things) have greater effect than neutral or positive evidence of equal intensity.

#### 4.6.9.1. Success Rates

Extensive experimental evidence for negativity bias was summarized by Rozin and Royzman [2001] for a range of domains. The most relevant to our focus appears to be the domain of attention and salience. In the experiments reported by Pratto and John [2005], it was investigated whether the valence of a word (desirable or undesirable trait) has effect on the time required to identify the color in which the word appears on the screen. The result was that the subjects took 29 ms longer to name color of undesirable word than for desirable word (679 vs 650 ms). As for the number of subjects affected, for 9 out of the 11 subjects the mean latency was higher for desirable words. The fact that the response time

[8] Sometimes referred to as "effect".

was higher for undesirable words is explained by Pratto and John [2005] by the undesirable trait obtaining more attention.

#### 4.6.9.2. Implications for Rule Learning

There are two types of effects that we discuss in the following: 1) effect of negated literal in the antecedent and 2) effect of negative class in the consequent.

1. Some rule learning algorithms are capable of generating rules containing negated literals. For example, male gender can be represented as `not(female)`. According to the negativity bias, the negative formulation of the same information will be given higher weight.

2. Considering a binary classification task, when one class is viewed as "positive" and the other class as "negative", the rule model may contain a mix of rules with the positive and negative class in the consequent. According to the negativity bias rules with the negative class in the consequent will be given higher weight. This bias can also manifest in the multiclass setting, when one or more classes can be considered as "negative". This effect can manifest also in the subsequent decision making based on the discovered and presented rules, because according to the principle of negative potency [Rozin and Royzman, 2001] and prospect theory [Kahneman and Tversky, 1979] people are more concerned with the potential losses than gains.

An interesting discovery applicable to both negation in antecedent and consequent shows that negativity is an "attention magnet" [Fiske, 1980, Ohira et al., 1998]. This implies that a rule predicting a negative class will obtain more attention than a rule predicting a positive class, which may also apply to appearance of negated literals in the antecedent. Also, research suggests that negative information is better memorized and subsequently recognized [Robinson-Riegler and Winton, 1996, Ohira et al., 1998].

We hypothesize that negativity bias will result in greater preference for longer rules, since there is higher odds that a longer rule will contain a negation than a shorter rule.

#### 4.6.9.3. Debiasing Techniques

We conjecture that rule learning systems can mitigate the effects of the negativity bias by avoiding the use of negation: use `gender=male` instead of `not(gender=female)`.

### 4.6.10. Primacy Effect

Once humans form initial assessment of plausibility (favourability) toward an option, subsequent evaluations of this option will favour the initial disposition.

#### 4.6.10.1. Success Rates

Bond et al. [2007] investigated to what extent changing the order of information which is presented to a potential buyer affects the propensity to buy. If the positive information

(product description) was presented as first, the number of participants indicating they would buy the product was 48%. When the negative information (price) was presented first, this number decreased to 22%. Participants were 118 undergraduate students.

Additional experimental evidence was provided by Shteingart et al. [2013].

#### 4.6.10.2. Implications for Rule Learning

Following the primacy effect the analyst will favour rules that are presented as first in the rule model.

#### 4.6.10.3. Debiasing Techniques

A machine learning application can take advantage of the primacy effect by presenting rules that are considered as most plausible based on observed data as first in the resulting rule model. Some rule learning algorithms, such as CBA [Liu et al., 1998], are natively capable of taking advantage of the primacy effect, since they naturally create rule models that contain rules sorted by their strength.

### 4.6.11. Reiteration Effect

The reiteration effect describes the phenomenon, which makes repeated statements more believable [Hertwig et al., 1997, Pachur et al., 2011].

#### 4.6.11.1. Success Rates

The experiment performed by Hasher et al. [1977] presented subjects with general statements and asked them to asses their validity on 7-point scale. The experiment was performed in several sessions, in subsequent sessions some of the statements repeated. Part of the statements were false and part were true. The average validity of repeated true statements rose between Session 1 and Session 3 from 4.52 to 4.80, while for non-repeated statements it dropped slightly. Similarly, for false statements, the validity rose from 4.18 to 4.67 for repeated statements and dropped for non-repeated statements. In this case repeating of false statements increased the subjectively-perceived validity by 11%.

#### 4.6.11.2. Implications for Rule Learning

In the rule learning context, "the repeated statement which becomes more believable" corresponds to the entire rule or possibly a "sub rule" consisting of the consequent of the rule and a subset of conditions in its antecedent. A typical rule learning result contains multiple rules that are substantially overlapping. If the analyst is exposed to multiple similar statements, the reiteration effect will increase the analyst's belief in the *repeating* "sub rule". In the rule learning context the bias behind the reiteration effect may not be justified. Especially in the area of association rule learning, a very large set of redundant rules – covering the same, or nearly same set of examples – is routinely included in the output.

When considering the influence of the reiteration effect on preference depending on rule length, our conclusion is that the reiteration effect will increase the preference for longer rules. The reason is that the longer rule is more likely to contain more "sub rules" of other rules in the rule model.

#### 4.6.11.3. Debiasing Techniques

A possible remedy for the reiteration effect can be performed already on algorithmic level by ensuring that rule learning output does not contain redundant rules. This can be achieved by pruning algorithms [Fürnkranz, 1997]. We also conjecture that this effect can be alleviated by explaining the redundancy on rule learning output to the analyst, for example by clustering rules.

### 4.6.12. Misunderstanding of "and"

The misunderstanding of "and" is a phenomenon affecting syntactic comprehensibility of the logical connective "and". As discussed by Hertwig et al. [2008], "and" in natural language can express several relationships, including temporal order, causal relationship, and most importantly, can also indicate a collection of sets instead of their intersection.[9]

#### 4.6.12.1. Success Rates

According to the two experiments reported in Hertwig et al. [2008], the conjunction "bank tellers and active feminists" used in the Linda problem (cf. Section 4.4.3) was found by about half of the subjects as ambiguous – they explicitly asked the experimenter how "and" is to be understood. The experiment involved determining understanding of "and" based on shading of Venn diagrams. The results indicate that 45 subjects interpreted "and" as intersection and 14 subjects as a union. The fallacy rate is thus 23%. Two thirds of subjects were university students and one third of subjects were professionals.

#### 4.6.12.2. Implications for Rule Learning

This effect will increase the preference of longer rules for reasons similar to those discussed for the conjunctive fallacy (cf. Subsection 4.4.3).

#### 4.6.12.3. Debiasing Techniques

According to Sides et al. [2002] "and" ceases to be ambiguous when it is used to connect propositions rather than categories. The authors give the following example of a sentence which is not prone to misunderstanding: "IBM stock will rise tomorrow and Disney stock will fall tomorrow." Similar wording of rule learning results may be, despite its verbosity, preferred. We further conjecture that representations that visually express the semantics of "and" such as decision trees may be preferred over rules, which do not provide such visual guidance.

[9] As in "He invited friends and colleagues to the party"

### 4.6.13. Weak Evidence Effect

According to this effect presenting weak evidence in favour of an outcome can actually decrease the probability that a person assigns to it. In an experiment in the area of forensic science reported by Martire et al. [2013], it was shown that participants presented with evidence weakly supporting guilt tended to "invert" the evidence, thereby counterintuitively reducing their belief in the guilt of the accused.

#### 4.6.13.1. Success Rates

Martire et al. [2013] performed an experiment in the judicial domain. When the presented evidence provided by the expert was weak, but positive, the number of responses incongruent with the evidence provided was 62%. When the strength of evidence was moderate or high the corresponding average was 13%. The subjects were undergraduate psychology students and Amazon Mechanical Turk workers (altogether over 600 participants).

#### 4.6.13.2. Implications for Rule Learning

The weak evidence effect can be directly applied on rules: the evidence is represented by rule antecedent; the consequent corresponds to the outcome. The analyst can intuitively interpret each of the conditions in the antecedent as a piece of evidence in favour of the outcome. Typical of many machine learning problems is the uneven contribution of individual attributes to the prediction. Let us assume that the analyst is aware of the prediction strength of the individual attributes. If the analyst is to choose from a shorter rule containing only the strong predictor and a longer rule containing a strong predictor and a weak (weak enough to trigger this effect) predictor, according to the weak evidence effect the analyst should choose the shorter rule.

#### 4.6.13.3. Debiasing Techniques

Our review did not reveal any debiasing strategies. This is related to the fact that the weak evidence effect is a relatively recent discovery. Our conjecture is that this effect can be alleviated by intentional omission of weak predictors from rules either directly by the rule learner or as part of feature selection.

### 4.6.14. Unit Bias

This cognitive bias manifests by humans tending to consider each condition as a unit of equal weight at the expense of detailed scrutiny of the actual effect of the condition [Geier et al., 2006].

#### 4.6.14.1. Success Rates

The effect of this bias was evaluated in Geier et al. [2006] on three food items: Tootsie Rolls, pretzels and M&Ms. These food items were offered in two sizes/scoops (on different

days) and it was observed how this will affect consumption. For Tootsie Rolls and M&Ms the larger unit size was 4x the smaller one and for pretzels 2x the smaller one. It follows from the figure included in [Geier et al., 2006] that increasing the size of the unit had about 50% effect on the amount consumed.

We were unable to find other experiment more closely related to the rule learning domain.

#### 4.6.14.2. Implications for Rule Learning

From the technical perspective, the number of conditions (literals) in rules is not important. What matters is the actual discriminatory power of the individual conditions, which can vary substantially. However, following the application of unit bias, the number of conditions has effect on subjective perception of discriminatory power of the antecedent as a whole. Under the assumption that the analyst will favour rule with higher discriminatory power, this heuristic will clearly contribute to preference for longer rules, since these contain more literals considered as "units".

Unlike other modes of communication humans are used to, rules resulting from algorithmic analysis of data do not provide clues relating to the importance of individual conditions, since rules often place conditions of vastly different importance side-by-side, not even maintaining the order from the most important to the least important. Such computer-generated rules violate conversational rules or "maxims", because they contain conditions which are not informative or *relevant.*[10]

In summary, the application of the unit bias in the context of rule learning can result in gross errors in interpretation. When domain knowledge on the meaning of the literals in the rule is absent, the unit bias can manifest particularly strongly.

#### 4.6.14.3. Debiasing Techniques

We conjecture that informing analysts about the discriminatory power of the individual conditions (literals) may alleviate unit bias. Such indicator can possibly be generated automatically by listing the number of instances in the entire dataset that meet the condition. Second, rule learning algorithms should ensure that literals are present in the rules in the order of significance, complying to human conversational maxims.

## 4.7. Summary

The conclusions of our analysis are presented in Table 4.1. Most heuristics and biases that we reviewed suggest that longer rules can be considered as more plausible. In summary, it could be concluded that the longer length of the rule gives more opportunities for various cognitive biases to be triggered. For example, more conditions present in the rule can make the rule more representative of the target class. Out of the biases and heuristics included in

[10] The relevance maxim is one of four conversation maxims proposed by philosopher Paul Grice, which was brought to relation with the conjunctive fallacy in the work of Gigerenzer and Hoffrage [1999] (see also [Mosconi and Macchi, 2001]).

our review, only the weak evidence effect and ambiguity aversion point in the direction of the shorter rule. Out of these, weak evidence effect is the most recently discovered bias in our review and requires further verification. There are two biases (confusion of the inverse and primacy effect) that do not seem to have direct relevance to rule length.

Interestingly, there are also biases independent of the content of the antecedent of the rule, which can bias the analyst towards the longer rule. These are base rate fallacy, insensitivity to sample size and effect of difficulty. These biases are triggered by imbalance between strength and weight of evidence, which correspond to rule confidence and support.

The following chapter is devoted to the empirical analysis of selected biases, focusing on their effect triggered by three types of stimuli – rule length, confidence and support interest values and negation.

| phenomenon | implications for rule-learning | debiasing technique |
|---|---|---|
| | biases that increase plausibility with rule length | |
| Availability Heuristic | Predictive strength of literal is based on association between the literals in the antecedent and the consequent of the rule | Trigger System 2 |
| Averaging Heuristic | Probability of antecedent as the average of probabilities of literals | Reminder of probability theory |
| Base-rate Fallacy | Emphasis on confidence, neglect for support | Express confidence and support in natural frequencies |
| Confirmation Bias | Rules confirming their prior hypothesis are "cherry picked" | i) Explicit guidance to consider evidence for and against hypothesis, ii) Screen analysts for susceptibility using Defense Confidence Scale questionnaire |
| Disjunction Fallacy | Prefer more specific literals over less specific | Inform on taxonomical relation between literals, explain benefits of higher support* |
| Effect of Difficulty | Rules with small support and high confidence are "overrated" | Filter rules that do not pass a statistical significance test, explain benefits of higher support* |
| Information Bias | More literals increase preference | Visualize value of literals, sort by predictive value* |
| Insensitivity to Sample Size | Analyst does not realize the increased reliability of confidence estimate with increasing value of support | Use support to compute confidence (reliability) intervals for the value of confidence |
| Mere Exposure Effect | Repeated exposure to literal results in increased preference | Extra time or knowledge of literals* |
| Misunderstanding of "and" | "and" is understood as disjunction | Express literal as proposition rather than as category |
| Negativity Bias | Negated or negative literals are considered as more important | Avoid the use of negation* |
| Recognition Heuristic | Recognition of literals increases preference | Extra time or knowledge of literals |
| Reiteration Effect | Same literal present in multiple rules increases preference | Pruning algorithms* |
| Representativeness Heuristic | Overestimate the probability of literal representative of target | Use natural frequencies instead of ratios |
| Tradeoff Contrast | Preference for literal is influenced by other literals in the rule or in other rules | i) Pruning discovered rules, ii) Information on semantics of the literal and covariance with other literals* |
| Unit Bias | Literals are perceived to have same weight | Inform on discriminatory power of literals* |
| | biases that decrease plausibility with rule length | |
| Ambiguity Aversion | Prefer known literal over unknown literal | Textual description of literals* |
| Weak Evidence Effect | Literal only weakly perceived as predictive of target decreases plausibility | Omission of weak predictors from antecedent |
| | effect independent of rule length | |
| Confusion of the Inverse | Confusing the difference between the confidence of the rule *P(consequent/antecedent)* with $P(antecedent\|consequent)$ | NA |
| Primacy Effect | Rules that are presented as first in the rule model are more preferred | Sort rules from strongest to weakest* |

Table 4.1.: Summary of analysis of cognitive biases. * conjecture, its absence means derived from empirical observation.

# 5. Empirical Analysis – Crowdsourcing Experiments

Based on the survey of related research it emerged that there are results from cognitive psychology applicable to semantic understandability and plausibility of inductively-learned rules. In this chapter, we report on our experimental evaluation of the effect of selected cognitive biases on two domains: a) inductive rule learning and b) Linda problem and its modifications.

The aim of our experiments is to verify the effect of selected cognitive biases when actual human users process rules discovered from data. In the previous chapter, we established that multiple individual cognitive biases should have effect on preference (plausibility) of rules depending on their length. Within this chapter we empirically test this proposition.

To limit the influence of a particular selection of target domain on the results, we decided to include rules mined from three different datasets. Our primary experiment performed on authentic results of rule learning is complemented by an experiment more tightly conforming to the problem formulation and setup established in cognitive science.

**Chapter organization.** Section 5.1 presents the research propositions. Section 5.2 introduces the crowdsourcing platform that we used to conduct the experiments. This section also covers the datasets from which we extracted rules, attributes and literals used in the instructions and questions that the subjects received. Section 5.3 describes the acquisition of proxy variables. Section 5.4 describes the experiment with pairs of inductively learnt rules. The description of experiments performed on the Linda problem and its modifications is confined to Section 5.5. Section 5.6 contains further discussion of selected results.

The following Chapter 6 presents a concise model that integrates the conclusions drawn from the meta-analysis of prior research in Chapter 4 with empirical results obtained in this chapter.

## 5.1. Research Propositions

A qualitative analysis of the effect of cognitive biases on rule length was performed in the previous chapter. Within the review we identified multiple cognitive biases that can impact interpretation of inductively learned rules. We identified three major types of stimuli that can trigger the biases: rule length, confidence and support values and negation in rules. We concluded that 16 biases out of 20 in our review steer the analyst to favour the longer rule and only two have the opposite effect.

The first hypothesis that we empirically test is whether humans indeed prefer longer rules. Then, to account preference to specific biases, we selected several biases for detailed empirical evaluation. The criteria used to select cognitive biases for empirical validation included our ability to devise a proxy variable [Clinton, 2004] that can represent the bias and our subjective estimate of the size of the bias.

The empirical analysis will focus on the following list of research propositions and problems:

- P 1: Longer rules are more plausible than shorter rules
- P 2: Higher Plausibility of Longer Rules is Caused by Misunderstanding of "and"
- P 3: Accounting relation of plausibility and rule length to cognitive biases
- P 4: Availability heuristic
- P 5: Weak evidence effect
- P 6: Insensitivity to sample size
- P 7: Disjunction fallacy
- P 8: Mere exposure Effect
- P 9: Replication of Linda Experiment
- P 10: Effect of negation on representativeness heuristic
- P 11: Triggering information bias by unknown value

Propositions 1–8 are tested within Experiment 1 and proposition 9–11 within Experiment 2. An overview of these experiments is presented in the following.

### 5.1.1. Motivation for Experiment 1 on Rules discovered from data

Subjects within Experiment 1 were presented with rules and asked to rate their plausibility. The purpose of the experiment was to attribute the elicited preference judgments to particular cognitive biases. A motivational example is included below.

**Example. (Motivation for Experiment 1)**
Consider the following rule:
`if the movie falls into all of the following group(s) (simultaneously)`
`American LGBT-related films and Englishlanguage Films`
`then the movie is rated as bad.`

Applicable cognitive biases and their hypothesized effects on perceived plausibility of this rule follow:

- Availability heuristic: what is the strength of association between genre and movie rating?
- Recognition heuristic: does the subject recognize the specific genre LGBT?
- Disjunction fallacy: which condition is more specific English-language Films or American LGBT-related films?
- Mere-exposure effect: what is the preference of the subject for "American LGBT-related films" and "Englishlanguage Films" based on the number of previous exposures to this type of movies?
- Weak evidence effect: is "American LGBT-related films" and "English-language Films" considered as strong-enough evidence for bad rating?

| variable | proxy for | primary bias |
|---|---|---|
| rule support – | sample size | insensitivity to sample size |
| PageRank (avg,max) ↑ | number of exposures | mere exposure effect |
| PageRank (min) ↓ | specificity of the concept | disjunction fallacy |
| attribute Relevance** ↑ | strength of association between predictor and target | availability |
| literal relevance (min*) ↓ | low strength of evidence | weak evidence effect |

Table 5.1.: Hypothesized links between explanatory variables and cognitive biases. ↑ indicates positive influence on plausibility with increasing value, ↓ indicates negative influence, – indicates no effect. * We model weak evidence effect via the minimum relevance associated with any literal in the rule. ** Even when controlled for literal relevance.

The proxy variables that we need to obtain to measure the effect of these biases are summarized in Table 5.1. Rule support is one of the direct products of rule learning. Values of PageRank [Page et al., 1998] can also be computed from data. Attribute relevance corresponds to human perception of the ability of a specific attribute in rule antecedent ("movie language") to predict values of the attribute in rule consequent ("movie rating"). Literal relevance goes one step further than attribute relevance by measuring human perception of the ability of a specific literal (attribute-value pair such as "language is English")

to predict specific value of the attribute in rule consequent ("rating is good"). The values of literal and attribute relevance are subjectively perceived and we elicit them empirically from human subjects. These values are elicited in experiments with different groups of subjects than the main experiments used to verify our research propositions P 1 – P 11.

### 5.1.2. Motivation for Experiment 2 with Variations on the Linda Problem

#### 5.1.2.1. Motivation for Replicating Linda

As follows from the data reported within the "Success Rates" sections in Chapter 4, most results in cognitive science that we used in our analysis were obtained in experiments where subjects were university students. In our empirical study we use crowdsourcing. Also, the experimental setting is different, because our questionnaire is distributed electronically, while most prior experiments were directly administered by experimenters in laboratory or classroom setting. Finally, the questions (rule pairs to assess) in Experiment 1 are generated algorithmically rather than manually as is the norm in psychological research. There are thus multiple elements that make the setup of Experiment 1 differ from related experiments performed in cognitive science research.

It has been shown that while results obtained with crowdsourcing contain noise, they can be replicated in the controlled laboratory setting [Brown et al., 2014]. In particular, the Linda problem has been previously replicated on the Amazon Mechanical Turk platform by Paolacci et al. [2010] and the authors found the fallacy rate to be within bounds of what has been previously published for the controlled laboratory setting. Overall, performing psychological experiments on the Amazon Mechanical Turk has been relatively extensively studied and no major differences compared to results obtained in the controlled laboratory setting were found [Paolacci and Chandler, 2014].

Since the Amazon Mechanical Turk platform is not available in Europe, we opted for another similar service, `CrowdFlower.com`. It is unclear to what extent the results of studies performed on Amazon Mechanical Turk are applicable to this platform. To justify our choice of crowdsourcing, and in particular of CrowdFlower as the experimental platform, we replicated the Linda experiment within Experiment 2.

#### 5.1.2.2. Motivation for Evaluating Effect of Negation

While considerable amount of attention has been paid to the problem of the interpretation of "and", less research has focused on the significance of negation ("not") as another important, yet unary, logical connective. Under a particular interpretation of the representativeness heuristic it is conceivable that a condition which is in its unnegated form considered as representative could continue to be representative even after it has been negated, and its meaning thus reversed. Many rule learning algorithms support negation, and the correct interpretation of rules containing negation is thus vital for correct understanding of rule learning output. The instructions administered to the second group of subjects provide data on how many people will "overlook" negation, evaluating the option the same way as if negation was not there.

#### 5.1.2.3. Motivation for Evaluating Effect of Unknown Value

Finally, we will use the Linda problem for analysis of the impact of the information bias. According to this bias, inclusion of a piece of information with no relevance can increase plausibility of a hypothesis. It is characteristic for the output of many – if not all – rule induction algorithms to involve "redundant" conditions – the same literal is included in multiple rules. Since whether a condition is redundant or not may be subject to personal opinion, we decided to evaluate the information bias on a particular example of a redundant condition – inclusion of a literal with unknown value. We use two modifications of the Linda problem to test this proposition.

### 5.1.3. Relation to Results of Chapter 4

Given the lack of previous research, we are primarily concerned whether cognitive biases selected from those included in Chapter 4 demonstrate when rules learned from data are interpreted. In addition to identification of biases relevant for rule learning, Chapter 4 also presents debiasing techniques. Out of these, within Experiment 1 (Proposition 2) we evaluate the effect of clarifying "and". Empirical validation of remaining suggested debiasing techniques is left for future work (cf. Section 12.2).

## 5.2. Experimental Platform and Subject Domains

In this section, we briefly introduce the crowdsourcing platform used for the experiments.

### 5.2.1. CrowdFlower

As the experimental platform, we used the CrowdFlower[1] crowdsourcing service. Similar to the better-known Amazon Mechanical Turk, CrowdFlower allows to distribute questionnaires to subjects in multiple countries around the world, who complete them for payment. The selection of crowdsourcing as a means of acquiring data allows us to gather thousands of responses in a manageable time frame while at the same time ensuring our results can be easily replicated.

#### 5.2.1.1. Overview of Experiment Setup in CrowdFlower

This subsection will briefly introduce the workflow of cognitive science experiment performed in CrowdFlower, covering also differences in terminology from the traditional laboratory-based experiment.

Subjects involved in experiments executed over crowdsourcing platforms are typically called *workers*, which is a term originally used in the Amazon Mechanical Turk platform [Paolacci and Chandler, 2014]. What is called an experiment in psychology corresponds to a *task* in CrowdFlower. Task in CrowdFlower is administered in *rows*, where one row corresponds to a question. A minimum amount of work one subject can complete is one *page*.

[1] `www.crowdflower.com`

The number of *rows per page* is set by the experimenter. An important difference between experiments previously performed in psychology and the typical CrowdFlower setup is that the number of subjects is not known before the crowdsourcing task finishes, since each participant can opt to finish different number of pages. What the experimenter sets instead of the number of participants in the CrowdFlower system is the number of *judgments per row*. CrowdFlower workers are always remunerated for their work. In CrowdFlower, the experimenter sets *price per page*, from which a price per judgment follows.

An integral part of the crowdsourcing task is ensuring that only qualifying subjects take part in the task and that they understand the instructions and stay motivated throughout the task. In CrowdFlower, these goals are met by selecting the level of participating subjects, the countries they are located at, by enabling *test questions* – administering a quiz that the subject is required to pass to qualify for the task, and by placing hidden quiz questions in the "work mode". The number of test questions presented is automatically managed by the platform. Test questions have the same structure as ordinary questions but additionally contain the expected correct answer (or answers) as well as explanation for the answer. The correct answer and explanation is shown after the subject had answered the question. Respondents can continue to the main task only if they pass the quiz with at least predefined level of accuracy. CrowdFlower also allows to specify *speed trap* – minimum time it should take the subject to complete a page of work. On the positive side, CrowdFlower gives the opportunity to reward the subjects with extra credit (*bonus*) after they have finished their work.

To allow replicability of all experiments, we report the following CrowdFlower-specific information in the *Methods* sections, included separately for Experiment 1 and 2 in Subsection 5.4.1 and 5.5.1:

- number of rows per page,
- number of judgments per row,
- price per judgment,
- test questions and minimum threshold to pass (optional),
- bonus (optional),
- maximum number of judgments per worker within task,
- *speed trap*,
- *country of residence*,
- *skill level.*

Note that the last three parameters (in italics) were set in the same way in all experiments as described in the following Subsection 5.2.1.2.

#### 5.2.1.2. Common Setup

In this section we describe the setup, which was common to both our experiments. CrowdFlower divides available workforce into three levels depending on the accuracy they obtained on earlier tasks. As the level of the CrowdFlower workers we chose Level 2, which is described as follows: "Contributors in Level 2 have completed over a hundred Test Questions across a large set of Job types, and have an extremely high overall Accuracy."[2]

Since the language of the assignment was English, we restricted the geographic eligibility of the task to subjects residing in U.S., Canada and United Kingdom. As the threshold for speed trap we used 180 seconds. If less than this amount of time to complete a page was taken, the subject was removed from the job. Maximum time required to complete the assignment was not specified (the CrowdFlower platform did not allow this).

Concerning the appropriateness of the remuneration, Schnoebelen and Kuperman [2010] give half-a-penny per question as the rule of thumb for payment on crowdsourcing services, which our remuneration exceeded in all experiments. To further ensure that the pay is appropriate, we checked the satisfaction scores reported in the final questionnaire by the subjects. On a 1-5 Likert scale (1 worst, 5 is best), the average subject rating of their remuneration for individual tasks was between 2.8 to 4.3. None of the jobs had pay rating in the red band.[3]

#### 5.2.1.3. Typical Procedure of Experiment Run with CrowdFlower

Crowdsourcing task performed in CrowdFlower consists of a sequence of the following steps:

1. The CrowdFlower platform recruits subjects for the task from among the cohort of its workers, who match the level and geographic requirements set by the experimenter. The workers decide to participate in the task based on the payment offered and the description of the task.

2. Subjects are presented assignment containing an illustrative example.

3. If task contains test questions, each subject has to pass a quiz session with test questions. Subjects learn about the correct answer after they pass the quiz mode. Subjects have the option to contest the correct answer if they consider it incorrect.

4. Subject proceed to the *work mode*, where they complete the task assigned by the experimenter. The task typically has a form of a questionnaire. If test questions were defined by the experimenter, the CrowdFlower platform randomly inserts test questions into the questionnaire. Failing predefined proportion of hidden test questions results in removal of the subject from the task. Failing the initial quiz or failing a task can also reduce subjects' accuracy on the CrowdFlower platform. Based on the average accuracy, subjects can reach one of the three levels. Higher level means access to additional – possibly better paying – tasks.

[2] `http://crowdflowercommunity.tumblr.com/post/80598014542/introducing-contributor-performance-levels`
[3] The CrowdFlower platform assigns three colour codes to the final scores (red, orange and green) to help interpreting the questionnaire results.

5. Subjects can leave the experiment at any time. To obtain payment for their work, subjects need to submit at least one page of work. After completing each page of work, the subject can opt to start another page of work. The maximum number of pages that subject can complete is set by the experimenter.

6. If bonus was promised, qualifying subjects receive extra credit.

It is one of the characteristic traits of CrowdFlower that the number of judgments per subject varies. The minimum number of judgments the subject had to provide corresponds to number of rows per page, a parameter set by experimenter. Note that in rare cases, the CrowdFlower platform may have asked the subject to provide lower number of judgments than five in order to obtain the total number of judgments ordered.

#### 5.2.1.4. The Crowdsourcing Worker

There is a number of differences between crowdsourcing and the controlled laboratory environment previously used to run psychological experiments. The central question is to what extent do the cognitive abilities and motivation of subjects differ between the crowdsourcing cohort and the controlled laboratory environment.

There is a paucity of data on the population of the CrowdFlower platform, which we use in our research. In this subsection, we present data related to another platform, specifically Amazon mechanical Turk, under the assumption that the descriptions of the populations will not differ substantially.[4]

The population of crowdsourcing workers is a subset of population of Internet users. This population is described in a recent meta study by Paolacci and Chandler [2014] as follows: "Workers tend to be younger (about 30 years old), overeducated, underemployed, less religious, and more liberal than the general population." While there is limited research on workers' cognitive abilities, Paolacci et al. [2010] found "no difference between workers, undergraduates, and other Internet users on a self-report measure of numeracy that correlates highly with actual quantitative abilities." According to a more recent study by Crump et al. [2013], workers learn more slowly than university students and may have difficulties with complex tasks. Possibly the most important observation related to the focus of our study is that according to Paolacci et al. [2010, page 417] crowdsourcing workers "exhibit the classic heuristics and biases and pay attention to directions at least as much as subjects from traditional sources."

### 5.2.2. Datasets

For Experiment 1 we used three datasets generated from the Linked Open Data (LOD) cloud and one dataset from the UCI repository. Overview of the LOD datasets (Traffic, Movies, Quality) is presented in Table 5.2 and additional details can be found in Ristoski

[4] This assumption is supported by the fact that until about 2014, CrowdFlower platform involved Amazon Mechanical Turk (AMT) workers. As of 2017, these workers are no longer involved, because according to CrowdFlower, the AMT channel was both slower and less accurate than other channels used by the CrowdFlower platform [Harris, 2014].

et al. [2016]. The dataset originating from the UCI repository is the Mushroom dataset, which contains mushroom records drawn from Field Guide to North American Mushrooms [Lincoff, 1981]. It is possibly the most frequently used dataset in the rule learning research, its main advantage are understandable attributes.

| dataset | source | # rows | # attributes | target |
|---|---|---|---|---|
| Traffic | LOD | 146 | 210 | rate of traffic accidents in a country |
| Movies | LOD | 2000 | 1770 | movie rating |
| Quality | LOD | 230 | 679 | quality of living in a country |
| Mushroom | UCI | 8124 | 23 | mushroom poisonous/edible |

Table 5.2.: Overview of experimental datasets. #rows denotes number of rows (instances) and #attributes the number of attributes describing each instance

#### 5.2.2.1. Relation between Experiments and Datasets

The workers in the CrowdFlower platform were invited to participate in individual tasks. There were multiple tasks sharing the same instructions, since it was desirable to collect preference data across multiple domains (datasets). In Experiment 1, which involved multiple datasets, each participant received questions related only to one dataset within the task. The CrowdFlower platform did not offer means for experimenter to prevent one subject to participate in multiple similar tasks. However, it is unlikely that this happens at a larger scale. At the time our experiments were run, the CrowdFlower platform was reported to have 5 million workers. The tasks were run at different times of day over the course of several months.

## 5.3. Acquisition of Proxy Variables

While we cannot directly observe the effect of individual biases, there are several variables that we can measure and use to indirectly estimate their influence: the relevance of the values and attributes appearing in the condition part of the rule for the prediction the rule makes. In order to acquire these ratings, we had to perform several crowdsourcing experiments. Additionally, we describe the role of PageRank which we use to model how well-known a particular literal is.

### 5.3.1. Rule Plausibility

In Section 2.1, we analysed the concept of "model comprehensibility" as understood in current research in the field of machine learning. We found that it relates to the concept of syntactic simplicity.

Consequently, we introduced semantic comprehensibility, which relates to human ability to use domain knowledge to invent examples that would falsify the model. Finally, pragmatic comprehensibility allows the decision maker to relate the hypothesis to prior

hypotheses applicable to the subject. Plausibility is a result of pragmatic comprehension of a hypothesis. Symbolically, we assume the following relation between plausibility and the three levels of comprehensibility: *syntactic comprehension* → *semantic comprehension* → *pragmatic comprehension* → *plausibility*.

Our main proposition is that when comprehensibility of models produced by a specific machine learning algorithm is tuned, the optimization criterion should be the highest, pragmatic[5], level of comprehension. This is a result of interplay of applicable domain knowledge, human cognitive biases and judgmental heuristics.

While prior work has repeatedly noted that the syntactic size of the model is inadequate to measure comprehensibility, alternate proposals remained also at the syntactic level. For example, a frequently cited notion is the *prediction-explanation size*, which relates to the number of conditions that need to be checked to predict a class value. In our experiments, we elicit *rule plausibility* as a proxy variable for semantic comprehensibility of the rule. Plausibility reflects the syntactic level, as well as subject's domain knowledge on both semantic and pragmatic level.

### 5.3.2. PageRank

We use PageRank [Page et al., 1998] to measure the general level of how well the given concept is important or known. In this way, PageRank can be linked to the *mere-exposure effect* as increasing value of PageRank can be used as a proxy variable for the number of exposures to the concept. The examples depicted in Table 5.3 show that PageRank is linked with how well-known the concept is, which can act as a proxy for how many times the person has been exposed to the concept.

Also, lower value of PageRank can act as a proxy for the specificity of the concept, thus modelling the *disjunction fallacy*. Referring to the example in Table 5.3, it is clear that "English language films" is a broad category and "Horror films from 1990" very narrow category. This illustrates how PageRank can represent the specificity of a concept.

It should be noted that our analysis is limited in that we neglect other heuristics and biases that can correlate with PageRank, such as the recognition heuristic: the higher the number of exposures, the higher the chance that the concept will be recognized (as a binary event). Also, PageRank can be negatively correlated with the occurrence of ambiguity aversion. Within the meaning of this heuristic, ambiguity is closely linked to the fear of the unknown, and higher PageRank implies better knowledge of the concept in question. Finally, it can be expected that specificity is inversely related to the number of exposures.

The PageRank values used in our analyses were precomputed for the experimental datasets by Prof. Heiko Paulheim (cf. Statement of originality on page 2).

[5] In order to limit the scope of our research, we decided not to make distinction between pragmatic and semantic comprehension.

Table 5.3.: Illustration of value distribution for explanatory variables. For the reported literal relevance values, , the highest corresponds to high quality (Quality), low accidents (Traffic), good rating (Movies), edible (Mushroom).

| dataset | lowest | highest |
|---|---|---|
| | PageRank | |
| Traffic | MemberStatesOfMercosur | MemberStatesOfTheUnitedNations |
| Quality | PopulatedPlacesInGabon | PopulatedCoastalPlacesInJapan |
| Movies | X1990sHorrorFilms | EnglishlanguageFilms |
| | literal relevance | |
| Traffic | SouthAsianCountries | UkrainianspeakingCountriesAndTerritories |
| Quality | RegionsOfNiger | CoastalCitiesInAustralia |
| Movies | X2000sHorrorFilms | FilmsFeaturingABestSupportingActor-AcademyAwardWinningPerformance |
| Mushr. | foul | anise |
| | attribute relevance | |
| Traffic | When the country was established | Level of development |
| Mushr. | population | veil-color |

### 5.3.3. Attribute Relevance

Attribute relevance serves us as a proxy variable for measuring the *availability heuristic* as it captures the strength of association between explanatory and target attribute.

Referring to Figure 5.1, the second rule in the example can be expected to have lower minimum attribute relevance than the first rule provided that date of release of a film is less relevant for determining film's quality than its language.

Attribute relevance does not only reflect availability heuristic, but also a level of recognition of the explanatory attribute, which is a prerequisite to determining the level of association with the target attribute. Example of a specific attribute that may not be recognized is "Sound Mix" for a movie rating problem. This would contrast with attributes such as "Oscar winner" or "year of release" – these are equally well recognized, but clearly associated to a different degree with the target attribute in rule consequent.

Attribute relevance judgment were collected using crowdsourcing experiment described in the following.

#### 5.3.3.1. Method

**CrowdFlower Setup.** For one judgment relating to one attribute we paid 0.04 USD. The number of judgments per row for this experiment was 5. The number of rows per page was set to 5. Test questions were not used. The quality bonus was provided after the completion of the task if the reason for the answer was longer than 10 characters; this

```
Rule 1: if the movie falls into all of the following group(s)
  (simultaneously)

    Englishlanguage Films

then the movie is rated as bad

Rule 2: if the movie falls into all of the following group(s)
 (simultaneously)

    Englishlanguage Films and
    Films Released In 2005

then the movie is rated as bad

Which of the rules do you find as more plausible?
```

Figure 5.1.: Example rule pair included in our experiments

criterion was however not revealed to subjects.

The common setting for the CrowdFlower platform described in Section 5.2.1.2 applied: the task was available to Level 2 workers residing in U.S., Canada and United Kingdom, maximum number of judgments per contributor was not limited and as the threshold for speed trap we used 180 seconds. Maximum time required to complete the assignment was not specified.

**Material.** Example wording of the attribute relevance elicitation task for the Mushroom dataset is included in Figure 5.2.

The instructions for the attribute relevance experiment were prepared for two groups of subjects – one obtained assignment generated for the Mushroom dataset, and the second group assignment for the Traffic dataset.

The text of the instructions followed the same template, but was customized according to the underlying dataset. Unfortunately, for the Traffic (LOD) dataset it was not possible to automatically construct meaningful names of attributes. For example, if the literal was "English speaking countries" we manually derived attribute "Language". For some literals, however, this choice was not as straightforward. After encountering these problems, we decided not to perform attribute relevance experiments for the remaining LOD datasets. The relevance judgments were collected in 1 to 10 range, where 1 meant that the attribute is *completely irrelevant* for determining the value of the class attribute in the given dataset and 10 meant *very relevant*.

The attribute relevance was the only type of experiment where we did not use test questions at all, since we did not find a way to design test questions that would not have the potential to bias the subjects. Instead, we provided 100% bonus for "quality". The only information subjects had available about the bonus is in Figure 5.2.

```
We kindly ask you to assist us in an experiment that will help
researchers understand which properties influence  mushroom
being considered as poisonous/edible.

Example task follows:

Property: Cap shape

Possible values: bell, conical, convex, flat, knobbed, sunken

What is the relevance of the property given above for
determining whether a mushroom is edible or poisonous?

Give a judgement on a 10 point scale, where:

    1 = Completely irrelevant
    10 = Very relevant

Obtaining further information
If the meaning of one of the properties is not clear, you can
try looking it up in Wikipedia.

BONUS

High quality responses will be rewarded by 100% of the original credit.

Thank you for your assistance !
```

Figure 5.2.: Mushroom attribute relevance experiment

Complete list of the questions (units) is presented in the supplementary material referenced from the Appendix.

**Procedure.** Subjects were faced with a web-based user interface, which presented a questionnaire consisting of questions, where one question related to one attribute. The responses to mushroom and traffic datasets were collected in separate tasks. An example task assignment is in Figure 5.3.

#### 5.3.3.2. Results

Elementary statistics describing the result of the attribute relevance experiment are shown in Table 5.4. We collected five judgments for each of the 24 attributes. The last column of Table 5.4 shows that the attributes in the Mushroom dataset were considered on average as 53% more relevant than the attributes in the Traffic dataset. This result can be at least partly accounted to the fact that for the Traffic dataset attribute names were manually

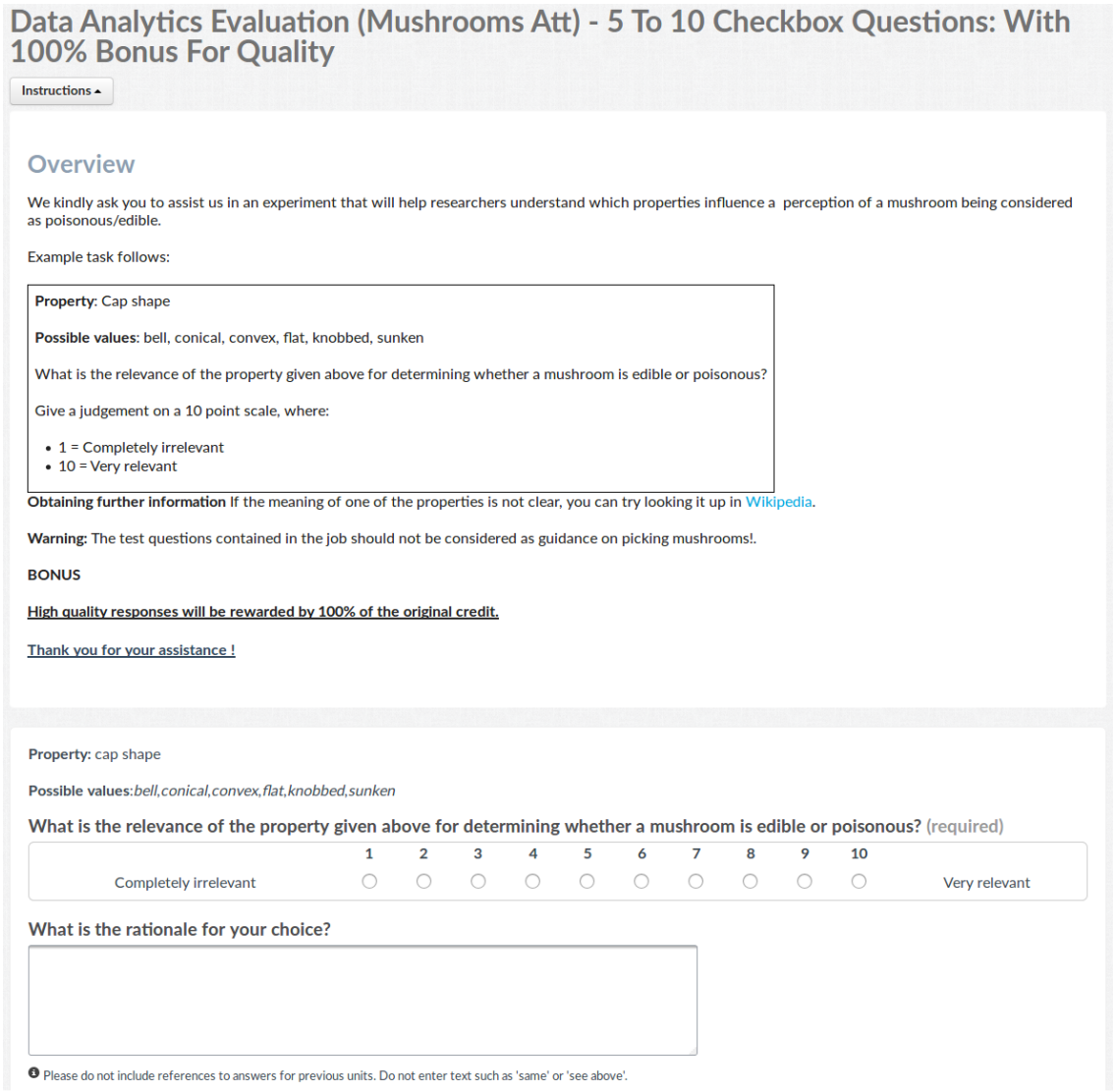



Figure 5.3.: Attribute relevance experiment assignment. The bottom part of the figure shows one unit (question), four other units from the same dataset were displayed below, within the web page.

derived by the author, while for the Mushroom dataset they were already present in the input data.

Complete raw data obtained within this experiment are present in the supplementary material referenced from the Appendix.

### 5.3.4. Literal Relevance

We use literal relevance as a measure of the *strength of evidence* represented by the literal. This can be used to model the *weak evidence effect*. Referring again to examples in Figure 5.1, literal relevance corresponds to the human-perceived predictive power of condition "English-language films" for "bad" movie rating.

It should be noted that we consider the literal relevance to also embed the value of the availability heuristic, since the literal ("film released in 2001") conveys also the attribute

Table 5.4.: Attribute relevance experiment result. 0 is completely irrelevant, 10 is very relevant. The avg. score attribute refers to the mean score assigned to attributes in the dataset listed in the first column in each row.

| dataset | distinct att. | example att. | judgments | avg. score |
|---|---|---|---|---|
| Mushroom | 14 | cap-shape | 92 | 5.5 |
| Traffic | 10 | form of government | 35 | 3.6 |

("year of release"). In addition to the attribute name, literal also conveys a specific value, which may not be due to its specificity recognized. This raises the problem of recognition as a prerequisite to association.

#### 5.3.4.1. Method

**CrowdFlower Setup.** For one judgment relating to one literal we paid 0.07 USD. The number of judgments per row for this experiment was 5. The number of rows per page was set to 5. Test questions were used (details are given in the Material section below). Only subjects achieving at least 70% accuracy on test questions could proceed to the main task. The quality bonus was not provided. The common setting for the CrowdFlower platform described in Section 5.2.1.2 applied: the task was available to Level 2 workers residing in U.S., Canada and United Kingdom, maximum number of judgments per contributor was not limited and as the threshold for speed trap we used was 180 seconds. Maximum time required to complete the assignment was not specified.

**Material.** Separate instructions were generated for all four datasets. Example wording of the literal relevance elicitation task for the Mushroom dataset is included in Figure 5.4. Figure 5.5 contains a sample test question; the correct answer is given in the caption of the figure.

The text of the instructions followed the same template, but differed based on the underlying dataset. It should be noted that there was a small difference in instructions for the three LOD datasets and the Mushroom dataset. The former instructions did contain links to Wikipedia for individual literals as these were naturally available from the underlying dataset. For Mushroom dataset, no such links were available.

The instructions asked for judgments on a five-point nominal scale: strong negative, weak negative, no influence, weak positive, strong positive. Unlike the attribute relevance experiment, the range of literal relevance was centered around 0 (No influence).

Complete list of the test questions as well as the main questions (units) is presented in the supplementary material referenced from the Appendix.

**Procedure.** Subjects were faced with a web-based user interface, which presented a questionnaire consisting of questions, where one question related to one literal. The responses

to individual datasets were collected in separate tasks. An example task assignment and one question is in Figure 5.4.

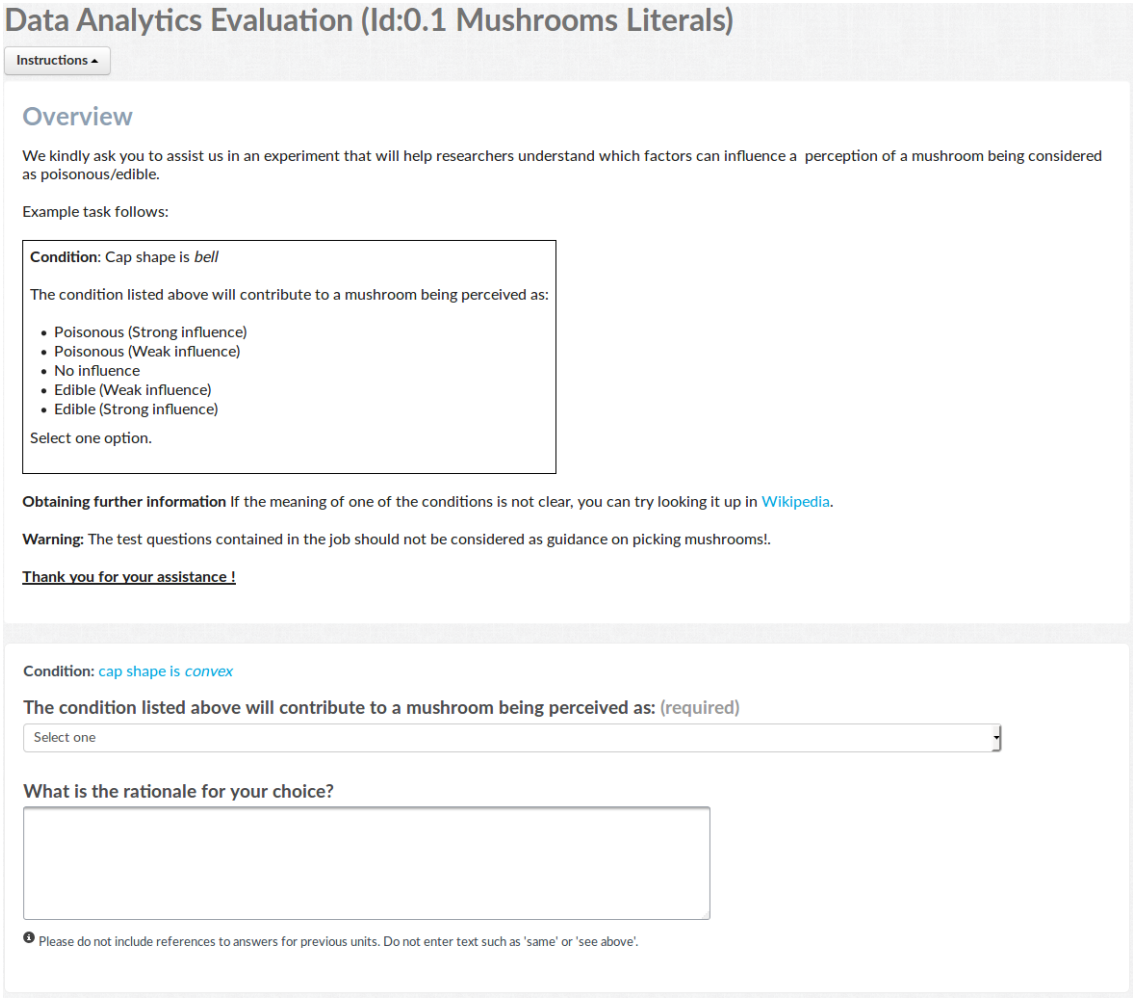



Figure 5.4.: Literal relevance experiment assignment. The bottom part of the figure shows one unit (question), four other units from the same dataset were displayed below, within the web page .

#### 5.3.4.2. Result

Elementary statistics describing the result of the literal relevance experiment are in Table 5.5. The last column in Table 5.5 reports the average relevance of literals in the individual datasets. Most relevant literals are in the Quality and Mushroom datasets.

The nominal response values were coded as integers in range $[-2; 2]$. The meaning of the codes was aligned with the target value of the rule pair, in which the literal appeared. For example, if the rule predicted good movie rating, relevance of literals in the antecedent of the rule was coded so that value 2 corresponded to good rating. If the same literal appeared in a rule with bad rating in the consequent, strong positive influence on movie rating was coded as -2.[6]

[6]For the Quality dataset, there are five distinct target values: rules with highest, high and medium label

```
We kindly ask you to assist us in an experiment that will help
researchers understand  which factors can influence  movie
ratings.

Example task follows:

Condition: Academy Award Winner or Nominee

The condition listed above will contribute to a movie being
rated as:

    Good (Strong influence)
    Good (Weak influence)
    No influence
    Bad (Weak influence)
    Bad (Strong influence)

Select one option.


Obtaining further information

If the meaning of one of the conditions is not clear, you can
click on the condition to see explanation in Wikipedia.

For example, consider condition "Obtaining XYZ award."
If you are not sure what exactly award XYZ is, you should
click on the link to consult the Wikipedia article.

Thank you for your assistance !
```

Figure 5.5.: Movie rating literal relevance test question. Expected correct answers were both Good (strong) or Good (weak).

Complete raw data obtained within this experiment are present in the supplementary material referenced from the Appendix.

## 5.4. Experiment 1: Rule Learning

Following the analysis of applicable biases and heuristics in Chapter 4 we reached the conclusion that longer rules will be considered as more plausible. This experiment aims to provide empirical verification of this proposition.

were considered as positive and rules predicting low and lowest quality as negative. Example of a rule pair where the negative version would apply is shown in Figure 5.1 on page 75.

Table 5.5.: Literal relevance experiment result, * positive or negative, the average (avg) is computed by coding both positive and negative strong as 2, positive/negative weak as 1, no influence as 0. The avg (average) column contains the mean score assigned to literals in the dataset listed in the first column in each row.

| dataset | literals | strong* | weak* | no influence | total | avg |
|---|---|---|---|---|---|---|
| Traffic | 58 | 16 | 47 | 227 | 290 | 0.27 |
| Quality | 33 | 20 | 73 | 72 | 165 | 0.68 |
| Movies | 30 | 7 | 34 | 109 | 150 | 0.32 |
| Mushroom | 34 | 14 | 34 | 102 | 150 | 0.41 |

### 5.4.1. Method

#### 5.4.1.1. CrowdFlower Setup

The workers in the CrowdFlower platform were invited to participate in individual tasks. For one judgment relating to one rule we paid 0.07 USD. The number of judgments per row for this experiment was 5. The number of rows per page was set to 5. Test questions were used (details are given in the Material section below). Only subjects achieving at least 70% accuracy on test questions could proceed to the main task. The quality bonus was not provided. The common setting for the CrowdFlower platform described in Section 5.2.1.2 applied: the task was available to Level 2 workers residing in U.S., Canada and United Kingdom, maximum number of judgments per contributor was not limited and as the threshold for speed trap we used 180 seconds. Maximum time required to complete the assignment was not specified. The maximum number of judgments per worker was not limited.

#### 5.4.1.2. Material

The list of versions of instructions is presented in Table 5.6. In total, instructions were prepared for 8 groups of subjects. The text of the instructions followed the same template, but differed based on the underlying dataset. In the following we describe the procedure that was used for all datasets to automatically generate the content (questions) that depended on the dataset. This section describes in detail the working of software that we designed to generate rules and instructions for the experiments. The other possibility is to use the software that we designed and used, which has been made openly available (cf. Appendix for details).

**Rule-pair Generation.** The basic element in the instruction is a pair of rules for which preference is elicited. We based the rule pair generation on several principles:

- All rules are result of rule learning on given dataset.
- There is at least one instance in the dataset covered by both rules in the pair.
- Rules in the pair have the same consequent (they predict the same class).

| group | dataset | instructions | min judgments | max judgments |
|---|---|---|---|---|
| G1mo | Movies | V1 | 5 | 32 |
| G1qu | Quality | V1 | 5 | 36 |
| G1tr | Traffic | V1 | 5 | 80 |
| G2mo | Movies | V2 | 5 | 32 |
| G2qu | Quality | V2 | 5 | 36 |
| G2tr | Traffic | V2 | 5 | 80 |
| G2mu | Mushroom | V2 | 5 | 10 |
| G3mo | Movies | V3 | 5 | 32 |

Table 5.6.: Versions of instructions – Experiment 1

- All literals in the rule are resolvable to a Wikipedia URL (applicable to LOD datasets only).

As input, we used the three LOD datasets described in Section 5.2.2. Before rule learning, we performed only minimum preprocessing. The three LOD datasets originally contained higher number of attributes than given in Table 5.2 on page 72, but these were pruned so that only specific types according to the YAGO ontology[7] were preserved. For the Mushroom dataset we used all available attributes. For the Quality dataset, we binned the numerical quality of living rank into five categories.[8]

The rules were generated for the LOD datasets by a standard implementation of the Apriori algorithm for association rule learning [Agrawal et al., 1993, Hahsler et al., 2011] and with inverted heuristic rule learner [Stecher et al., 2016]. For LOD datasets, the consequent was constrained to contain only the class attribute. The minimum confidence threshold was set to 0.5, minimum length to 1 and maximum length to 5. The rules for the Mushroom dataset were generated only by the inverted heuristic learner. The scripts used for rule mining and data preprocessing as well as the rules produced are included in the supplementary material (cf. Appendix).

**Choosing Rules to Form Rule Pairs.** For Traffic dataset we generated several approximately equally-sized groups of pairs of rules as depicted in Table 5.8. The "balancing" was possible only on the Traffic dataset, where there was sufficiently large number of rule pairs to choose from. For Quality and Movies datasets all rule pairs were used. For the Mushroom dataset, we selected rule pairs so that every difference in length is represented (one to five).

The scripts used as well as the rule pairs produced are included in the supplementary material (cf. Appendix). The size of the rule-pairs generated for the individual datasets is given in Table 5.7.

For all datasets, the order of the rules in the rule pair was randomized.

[7] `https://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/`

[8] "Highest" for rank smaller than 46, "High" for rank smaller than 92, "medium" for rank smaller than 138, "low" for rank smaller than 184, and "lowest" for the remainder.

Table 5.7.: Overview of final rule-pair per dataset, * excluding test questions

| dataset | number of rule pairs* |
|---|---|
| Mushroom | 10 |
| LOD datasets | 148 |
| - Traffic | 80 |
| - Quality | 36 |
| - Movies | 32 |

Table 5.8.: Traffic dataset rule selection groups

| |
|---|
| **subsuming** |
| different-length rules, either antecedent of rule 1 is subset of antecedent of rule 2, or antecedent of rule 2 is subset of antecedent of rule 1 |
| **different length rules with disjunct attributes** |
| different-length rules, the antecedent of rule 1 is disjunct with antecedent of rule 2 |
| **same length rules non disjunct attributes** |
| same-length rules, antecedent of rule 1 is not disjunct with antecedent of rule 2 |
| **same length rules disjunct attributes** |
| same-length rules, antecedent of rule 1 is disjunct with antecedent of rule 2 |
| **different length rules neither disjunct nor subsuming attributes** |
| different-length rules, the antecedent of rule 1 is not disjunct with antecedent of rule 2, antecedent of rule 1 is not subset of antecedent of rule 2, antecedent of rule 2 is not subset of antecedent of rule 1 |
| **large difference in rule length** |
| the difference between the lengths of the rules is at least 2 (selected only from inverted heuristic pairs) |
| **one difference in rule length** |
| the difference between the lengths of the rules is exactly 1 (selected only from inverted heuristic pairs) |

**Translation of Rules into Human-friendly Form.** All rule pairs were automatically translated into human friendly HTML-formatted text. Example rules for all four datasets are depicted in Figure 5.6.

The computer scripts used to generate the pairs for instructions as well as the translated rules are included in the supplementary material (cf. Appendix).

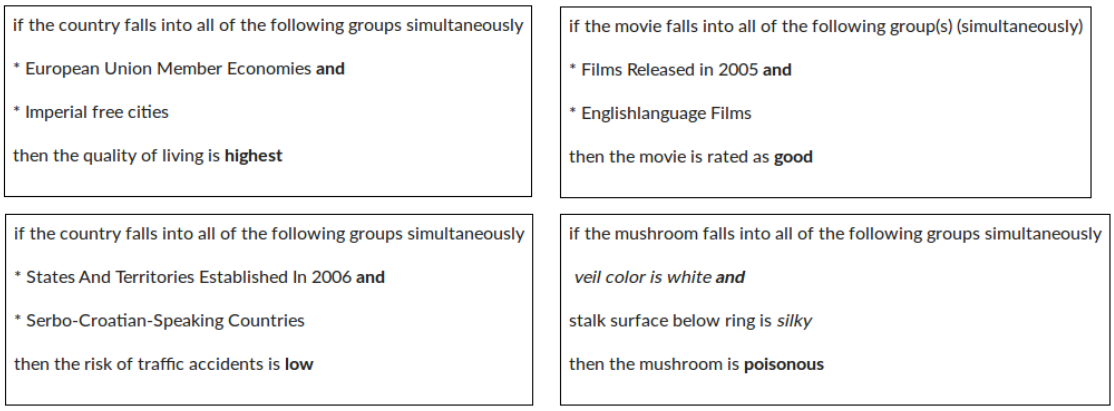



Figure 5.6.: Example translated rules for the four datasets

**Test Questions.** In CrowdFlower, test questions look the same as real questions but they also contain the correct answer and explanation. We devised two types of test questions: *swap* and *intersection* test question.

The *swap test question* contained two nearly identical rules. The only difference was a swapped order of literals in the antecedent so that the rules looked different. The subject was expected to answer "no preference" to these test rule pairs. The purpose was to filter out subjects who were not paying attention to the task. Example swap test question is depicted in Figure 5.7. Another example of a swap test question for different dataset is shown in Figure 5.8.

The *intersection test questions* were more difficult. We put these test questions in place to address the possible ambiguity of the *and* conjunction, which has been hypothesized to cause the conjunctive fallacy by Hertwig et al. [2008]. In order to correctly answer the intersection test question, the subject had to realize that the antecedent of one of the rules contains mutually exclusive conditions. The correct answer was a weak or strong preference for rule which did not contain the mutually exclusive conditions. Example of an intersection test question is shown in Figure 5.9.

The test questions used are included in the supplementary material (cf. Appendix).

**Manipulations – Versions of Instructions.** The experiment was run in three versions of instructions (not counting customizations for individual datasets), overview of which is provided by Table 5.9. An important observation related to this table is that not all instructions were applied in conjunction with all datasets. Example of task assignment for the Mushroom dataset for Version 1, 2 is shown in Figure 5.10 on page 87. Figure 5.11

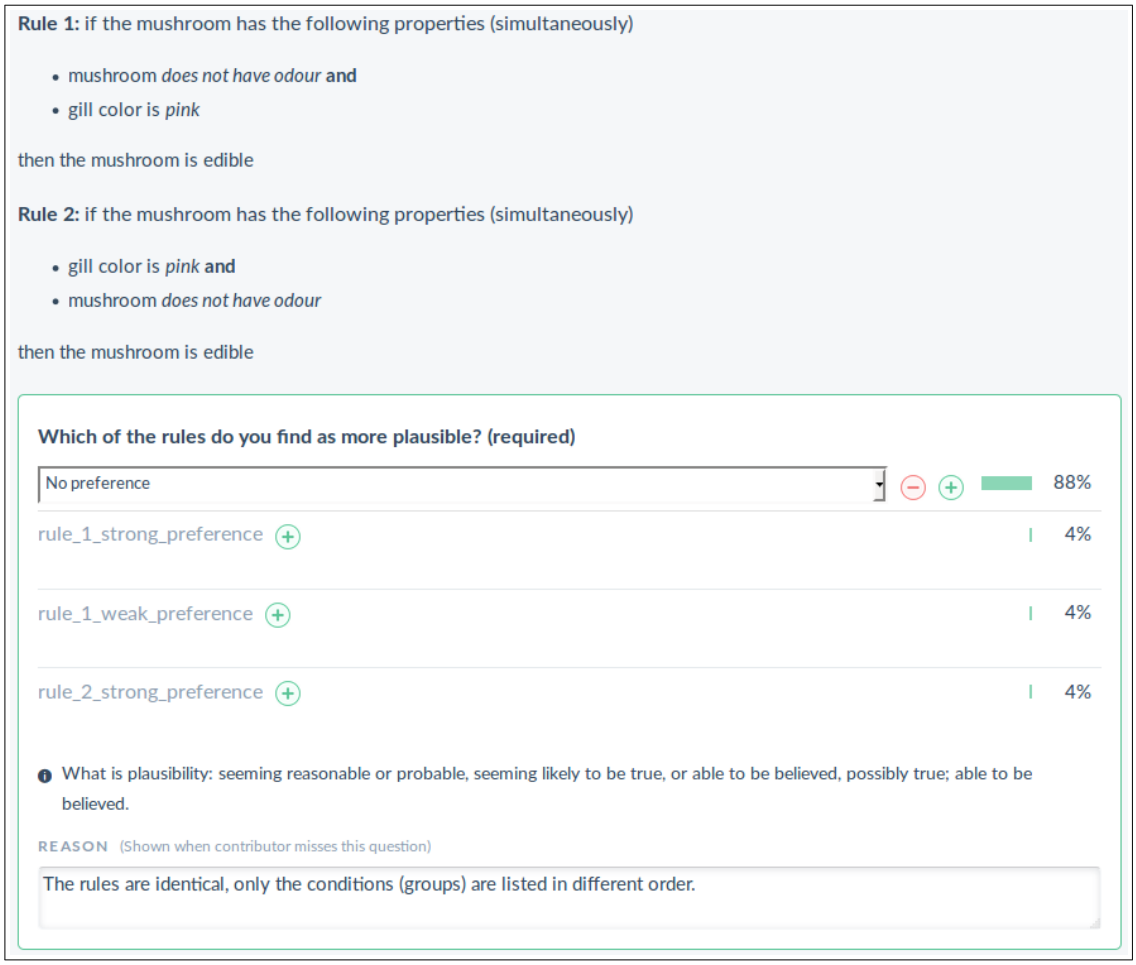



Figure 5.7.: Example test question - Mushroom dataset (administrator view including percentage of correct and incorrect answers)

shows an example task assignment (Movies dataset) for Version 3.

| version | test questions | quality shown | datasets |
|---|---|---|---|
| 1 | intersection, swap | no | Movies, Quality, Traffic |
| 2 | swap | no | Movies, Quality, Traffic, Mushroom |
| 3 | swap | yes | Movies |

Table 5.9.: Versions of the rule-length instructions

**Version 1** removed subjects misinterpreting "and": there were two types of test questions, easy *swap* test question were used to ensure that subjects pay attention to the task, and more difficult *intersection* test questions were used to remove subjects not understanding "and".

**Version 2** kept subjects misinterpreting "and": the intersection test questions were not used.

**Version 3** aimed at investigating the effect of explicitly revealed confidence and support: to suppress the effect of base-rate fallacy, we presented support as a natural number in our

```
Rule 1: if the movie falls into all of the following group(s)
(simultaneously)

    Englishlanguage Films and
    Serial Killer Films and
    Thriller Films Released In 2000s

then the movie is rated as bad

Rule 2: if the movie falls into all of the following group(s)
(simultaneously)

    Serial Killer Films and
    Englishlanguage Films and
    Thriller Films Released In 2000s

then the movie is rated as bad

Which of the rules do you find as more plausible?
```

Figure 5.8.: Example swap test question (Movies dataset), the assumed correct answer is no preference

```
Rule 1: if the movie falls into all of the following group(s)
(simultaneously)

    Religious Horror Films and
    Films Based On Children's Books

then the movie is rated as good

Rule 2: if the movie falls into all of the following group(s)
(simultaneously)

    American LGBTrelated Films and
    Englishlanguage Films

then the movie is rated as good

Which of the rules do you find as more plausible?
```

Figure 5.9.: Example intersection test question, the assumed correct answer is a weak or strong preference for Rule 2

experiment.

The difference between V1 and V2 was one manipulation. Instructions V2 did not contain the intersection test questions present in V1. Likewise, the difference between V2 and V3 was one manipulation. Instructions additionally contained the values of confidence and support not present in V2 (or V1). This version was prepared only for the Movies dataset. The reason was that for the other three datasets, the values of confidence and support within the rule pairs were either identical or very similar.

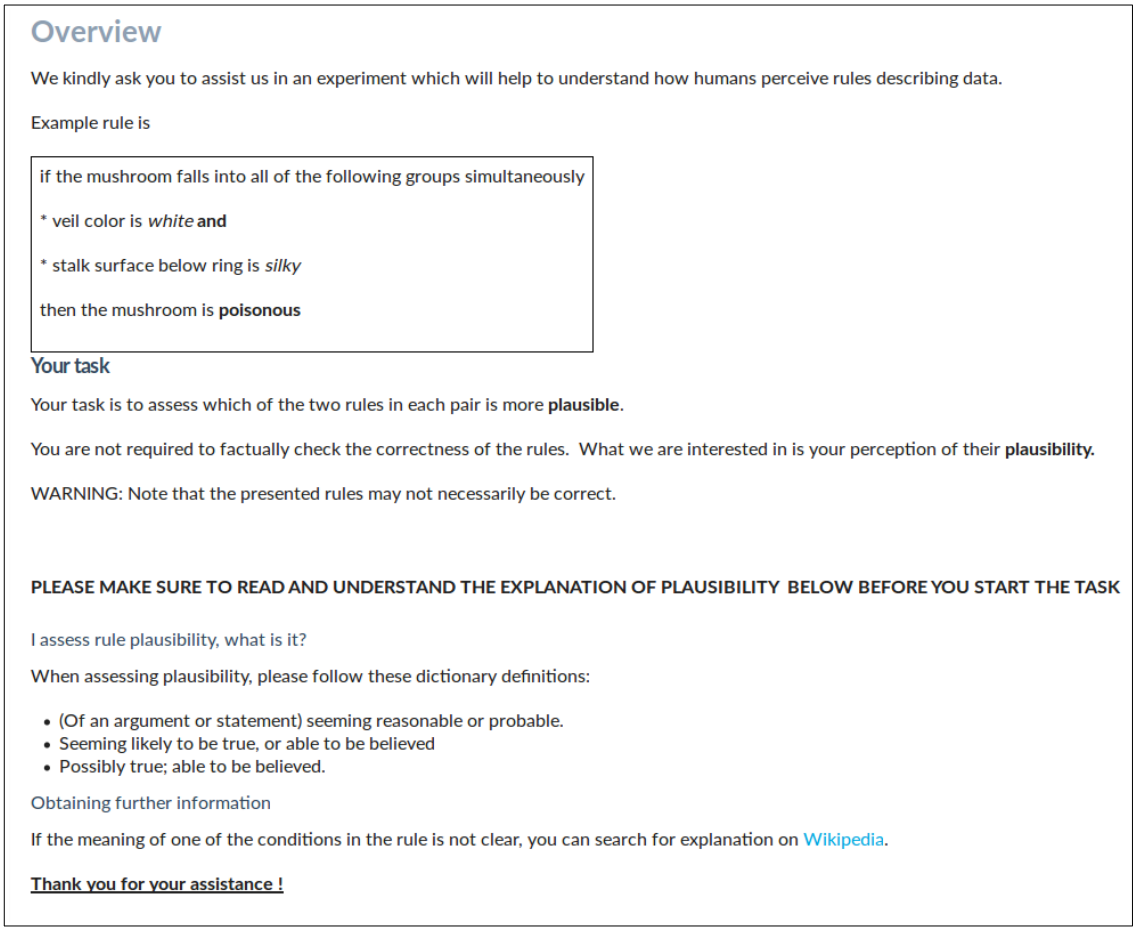



Figure 5.10.: Example of a task assignment for the Mushroom dataset (V1, V2)

#### 5.4.1.3. Procedure

The subjects were faced with a web-based user interface, which presented a questionnaire consisting of questions, where one question related to pairs of rules inductively mined from data. For each pair, the subjects were asked to give a) judgment which rule in each pair is more preferred and b) optionally a textual explanation for the judgment. Example of a complete task assignment is depicted in Figure 5.10.

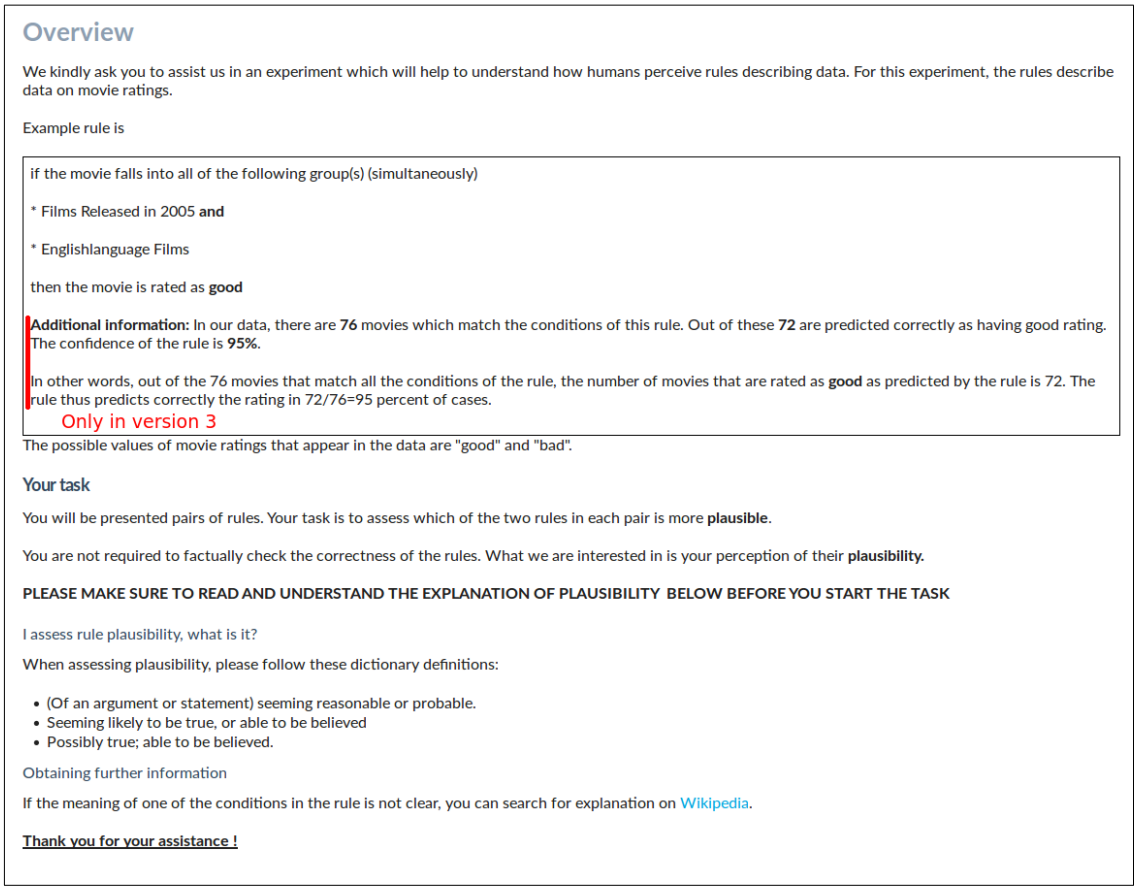



Figure 5.11.: Task assignment for the Movies dataset (Setup Version 3). Version 1 and 2 did not contain the "Additional information" block.

## 5.4.2. Results

### 5.4.2.1. Description of Cohort

While some basic parameters of the subject cohort can be set by the experimenter, the exact number of participants, their geographical location, as well as the average time required to complete the task is available only after the crowdsourcing task finishes. An overview of these characteristics is provided in Table 5.10.

### 5.4.2.2. Data Preprocessing

Once we collected the results from crowdsourcing, we faced a choice whether to aggregate the data before analysis or leave the data as is. We decided to work with the unaggregated data: each preference judgment for a rule pair constituted one data point. The alternate option was that each rule pair would appear just once in the dataset with the target variable being the aggregation (average) of all judgments relating to the pair. By performing the analysis at the level of individual judgments, also called micro-level, we avoided the possible loss of information as well as the aggregation bias [Clark and Avery, 1976]. Also, as shown for example by Robinson [1950] the ecological (macro-level) correlations are generally larger than the micro-level correlations, therefore by performing the analysis on the individual

Table 5.10.: Participant cohort in Experiment 1. Each line corresponds to one group of subjects. The *version* column denotes the version of the instructions this group received, *judg* the number of judgments collected, *workers* the total number of unique subjects, *usa*/*gbr*/*can* the number of judgments from subjects from United States of America/Great Britain/Canada, *avg dur* the average duration the subject took to finish one page of work in minutes and seconds and *reasons* the number of textual reasons longer than 10 characters.

| dataset | version | judg | workers | usa | gbr | can | avg dur | reasons |
|---|---|---|---|---|---|---|---|---|
| Traffic | 1 | 400 | 101 | 212 | 116 | 72 | 06:33 | 339 |
| Quality | 1 | 180 | 45 | 96 | 40 | 44 | 07:00 | 150 |
| Movies | 1 | 164 | 46 | 76 | 30 | 58 | 04:14 | 144 |
| Mushroom | 1 | 180 | 64 | 116 | 26 | 38 | 05:45 | 162 |
| Traffic | 2 | 408 | 93 | 204 | 120 | 84 | 06:27 | 339 |
| Quality | 2 | 184 | 41 | 68 | 64 | 52 | 0 06:41 | 145 |
| Movies | 2 | 160 | 40 | 80 | 52 | 28 | 06:03 | 104 |
| Mushroom | 2 | 300 | 105 | 126 | 99 | 75 | 05:57 | 263 |
| Movies | 3 | 160 | 40 | 84 | 44 | 32 | 07:26 | 125 |
| *total* | | 2136 | 575 | 1062 | 591 | 483 | na | 1771 |

level we obtain more conservative results.

#### 5.4.2.3. Choice of Statistical Methods

We performed statistical analysis of the crowdsourcing results using the following methods: rank correlation (Kendall's $\tau$, Spearman's $\rho$) and semi-partial correlation.[9] For all these methods we tested whether the coefficients are significantly different from zero. We will refer to the values of Kendall $\tau$ as the primary measure of rank correlation, since according to Gibbons and Kendall [1990] (cited according to Newson [2002]) the confidence intervals for Spearman's $\rho$ are less reliable than confidence intervals for Kendall's $\tau$.

For all obtained correlation coefficients we compute the *p value*, which is the probability of obtaining a correlation coefficient at least as extreme as the one that was actually observed assuming that the null hypothesis is true. The typical cutoff value for rejecting the null hypothesis is $\alpha = 0.05$.

Finally, in line with the current recommendations of the American Statistical Association on the use of p-values for supporting research hypotheses [Wasserstein and Lazar, 2016], we aim for full reporting and transparency by providing an overview of all the major hypotheses, including computed correlation coefficients and p-values, that we investigated.

#### 5.4.2.4. Response to Test Questions

Elementary statistics describing the result of the crowdsourcing experiment are in Table 5.11.

[9] To evaluate the predictive power, we also did significance testing on logistic regression coefficients, but this is not reported.

Table 5.11.: Rule-length experiment statistics. *units* distinct number of rule pairs, *judg* refers to the number of judgments, *qfr* refers to the quiz failure rate – the percentage of subjects that did not pass the initial quiz.

| ver. | | V1 | | V2 | | V3 | |
|---|---|---|---|---|---|---|---|
| dataset | units | judg | qfr [%] | judg | qfr [%] | judg | qfr [%] |
| Traffic | 80 | 419 | 55 | 412 | 12 | | |
| Quality | 36 | 180 | 31 | 184 | 11 | | |
| Movies | 32 | 176 | 15 | 156 | 14 | 160 | 5 |
| Mushroom | 10 | 150 | 40 | 250 | 14 | | |
| *total* | 158 | 955 | 40 | 962 | 13 | 160 | 5 |

Version 1 of the instructions that included the "intersection" test questions was more difficult with average 40% of candidate subjects not passing the initial quiz. After the intersection test questions were removed in Version 2 of the instructions, the failure rate dropped to 13%. For Version 3 there was a further drop in the failure rate, which can be attributed to the fact that in Version 3 the subjects got an additional clue that the rules in "swap" test questions are identical: both rules had the same values of confidence and support.

The plausibility judgments were coded as integers between -2 (strong preference for Rule 2) to 2 (strong preference for Rule 1). As Table 5.11 shows we collected on average five judgments (955 total) for each of the 158 rule pairs. Since we decided to perform micro-level analysis, each judgment corresponded to one data point.

#### 5.4.2.5. Enrichment of Data with Proxy Variables

The data were enriched with additional variables (Table 5.12). The values of these variables were computed as the value of the corresponding metric for rule 1 minus the value for rule 2. It should be noted that for the minimum variables, for example minimum literal relevance, the order of the rules was reversed: value for rule 1 was subtracted from value for rule 2.

We also experimented with computing ratios instead of deltas of the values, but due to methodological problems associated with the use of ratios in correlation and regression analysis [Kuh and Meyer, 1955, Tu et al., 2004] we decided not to include ratios into the final version.

#### 5.4.2.6. P 1: Longer Rules are More Plausible than Shorter Rules

The proposition that we tested is: "*Humans find longer rules (i.e. rules with more conditions) as more plausible than shorter rules, all other things being equal.*" The corresponding null hypothesis is that there is no correlation between rule length and perceived plausibility. To investigate this proposition, we need to refer to results obtained for groups that received V2 instructions on all four rule datasets. Recall that the V2 instructions included swap test questions that verified whether the subjects were paying attention to the task. The intersection test questions were not included. Kendall's rank correlation coefficient $\tau$

Table 5.12.: Independent (explanatory) variables

| variable | computation | meaning |
|---|---|---|
| basic rule statistics (all datasets) | | |
| LenDelta | r1-r2 | Rule length Δ |
| SuppDelta | r1-r2 | Rule support Δ |
| ConfDelta | r1-r2 | Rule confidence Δ |
| literal relevance (all datasets) | | |
| LitMax | r1-r2 | Max Literal relevance Δ |
| LitAvg | r1-r2 | Avg Literal relevance Δ |
| LitMin | r2-r1 | Min Literal relevance Δ |
| attribute relevance (Mushroom and Traffic) | | |
| AttMax | r1-r2 | Max Attribute relevance Δ |
| AttAvg | r1-r2 | Avg Attribute relevance Δ |
| AttMin | r2-r1 | Min Attribute relevance Δ |
| literal PageRank (LOD datasets) | | |
| PRMax | r1-r2 | Max PageRank Δ |
| PRAvg | r1-r2 | Avg PageRank Δ |

is used to measure ordinal association between the difference in length of rules in the pairs and the difference in the level of preference (plausibility).

The relevant results are in the first two rows of Table 5.13 (groups that received V2 instructions). On the LOD datasets with V2 instructions there is a small but statistically significant positive correlation between rule length and plausibility. The value of the Kendall's $\tau$ is 0.06 ($p < 0.05$). However, individually on V2 Movies and V2 Traffic datasets the correlation coefficient is not statistically different from zero. Overall, the correlation is strongest on the Mushroom dataset, where the Kendall's $\tau$ reaches 0.37 ($p < 0.0001$) and the Spearman's $\rho$ even 0.45 ($p < 0.0001$).

Table 5.13.: Correlation coefficients between rule length $\Delta$ and plausibility $\Delta$, p-value is given in parentheses.

| | V1 LOD | V2 LOD | V1 Quality | V2 Quality | V1 Movies | V2 Movies | V3 Movies | V1 Traffic | V2 Traffic | V1 Mushr. | V2 Mushr. |
|---|---|---|---|---|---|---|---|---|---|---|---|
| corr. | simple correlation | | | | | | | | | | |
| Ke. $\tau$ | -0.05 (0.139) | 0.06 (0.038) | -0.03 (0.624) | 0.2 (0.002) | -0.03 (0.689) | -0.01 (0.837) | 0.1 (0.14) | -0.03 (0.444) | 0.05 (0.226) | 0.28 (0) | 0.37 (0) |
| Sp. $\rho$ | -0.05 (0.154) | 0.08 (0.039) | -0.04 (0.627) | 0.23 (0.002) | -0.03 (0.711) | -0.02 (0.828) | 0.12 (0.132) | -0.04 (0.466) | 0.06 (0.23) | 0.34 (0) | 0.45 (0) |
| control | semi-partial correlation (single control), Kendall's $\tau$ | | | | | | | | | | |
| Literal Relevance $\Delta$ | | | | | | | | | | | |
| Max | -0.05 (0.029) | 0.05 (0.052) | -0.05 (0.295) | 0.1 (0.056) | -0.05 (0.391) | -0.05 (0.324) | 0.07 (0.173) | -0.03 (0.347) | 0.05 (0.128) | 0.27 (0) | 0.34 (0) |
| Avg | -0.04 (0.137) | 0.08 (0.001) | -0.04 (0.455) | 0.18 (0) | 0 (0.944) | 0.03 (0.61) | 0.12 (0.025) | -0.02 (0.534) | 0.06 (0.091) | 0.23 (0) | 0.28 (0) |
| Min | -0.02 (0.347) | 0.1 (0) | -0.01 (0.796) | 0.25 (0) | 0.01 (0.797) | 0.05 (0.357) | 0.11 (0.042) | -0.01 (0.742) | 0.06 (0.068) | 0.21 (0) | 0.26 (0) |
| Attribute Relevance $\Delta$ | | | | | | | | | | | |
| Max | -0.01 (0.673) | 0.05 (0.133) | na | na | na | na | na | -0.01 (0.673) | 0.05 (0.133) | 0.2 (0) | 0.22 (0) |
| Avg | -0.03 (0.41) | 0.05 (0.143) | na | na | na | na | na | -0.03 (0.41) | 0.05 (0.143) | 0.22 (0) | 0.31 (0) |
| Min | -0.03 (0.351) | 0.05 (0.122) | na | na | na | na | na | -0.03 (0.351) | 0.05 (0.122) | 0.12 (0.026) | 0.14 (0) |
| PageRank $\Delta$ | | | | | | | | | | | |
| Max | -0.02 (0.374) | 0.07 (0.004) | -0.02 (0.63) | 0.16 (0.002) | 0 (0.95) | 0.03 (0.559) | 0.05 (0.331) | -0.02 (0.523) | 0.05 (0.193) | na | na |
| Avg | -0.03 (0.219) | 0.11 (0) | -0.03 (0.615) | 0.19 (0) | -0.01 (0.87) | 0.04 (0.412) | 0.08 (0.119) | -0.04 (0.251) | 0.09 (0.017) | na | na |
| Min | -0.05 (0.032) | 0.06 (0.011) | -0.03 (0.494) | 0.19 (0) | -0.02 (0.644) | -0.01 (0.817) | 0.1 (0.063) | -0.07 (0.063) | 0.06 (0.067) | na | na |
| control | semi-partial correlation (multiple control variables), Kendall's $\tau$ | | | | | | | | | | |
| Attribute Relevance $\Delta$, Literal Relevance $\Delta$ | | | | | | | | | | | |
| min | -0.01 (0.74) | 0.06 (0.068) | na | na | na | na | na | -0.01 (0.74) | 0.06 (0.068) | 0.12 (0.028) | 0.13 (0.002) |
| max | -0.01 (0.668) | 0.05 (0.099) | na | na | na | na | na | -0.01 (0.668) | 0.05 (0.099) | 0.2 (0) | 0.27 (0) |
| avg | -0.01 (0.716) | 0.05 (0.129) | na | na | na | na | na | -0.01 (0.716) | 0.05 (0.129) | 0.2 (0) | 0.22 (0) |
| Attribute Relevance $\Delta$, Literal Relevance $\Delta$ and Pagerank $\Delta$ | | | | | | | | | | | |
| min | -0.05 (0.199) | 0.08 (0.028) | na | na | na | na | na | -0.05 (0.199) | 0.08 (0.028) | na | na |
| max | -0.03 (0.384) | 0.1 (0.008) | na | na | na | na | na | -0.03 (0.384) | 0.1 (0.008) | na | na |
| avg | -0.01 (0.76) | 0.05 (0.151) | na | na | na | na | na | -0.01 (0.76) | 0.05 (0.151) | na | na |

Table 5.14.: Correlations between additional variables and plausibility Δ (Kendall's $\tau$)

| | V1 LOD | V2 LOD | V1 Quality | V2 Quality | V1 Movies | V2 Movies | V3 Movies | V1 Traffic | V2 Traffic | V1 Mush. | V2 Mush. |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Literal Relevance Δ | | | | | | | | | | | |
| Max | 0.08 (0.009) | 0.16 (0) | 0.06 (0.277) | 0.31 (0) | 0.11 (0.063) | 0.22 (0) | 0.15 (0.014) | 0.05 (0.181) | 0.01 (0.797) | -0.04 (0.511) | -0.11 (0.037) |
| Avg | 0.07 (0.009) | 0.14 (0) | 0.08 (0.153) | 0.29 (0) | 0.08 (0.192) | 0.15 (0.012) | 0.08 (0.201) | 0.08 (0.031) | 0.04 (0.311) | -0.13 (0.046) | -0.19 (0) |
| Min | -0.07 (0.017) | -0.11 (0) | -0.09 (0.113) | -0.24 (0) | -0.07 (0.255) | -0.11 (0.072) | -0.01 (0.832) | -0.08 (0.066) | -0.04 (0.377) | 0.15 (0.017) | 0.22 (0) |
| Attribute Relevance Δ | | | | | | | | | | | |
| Max | -0.04 (0.325) | 0 (0.983) | na | na | na | na | na | -0.04 (0.325) | 0 (0.983) | 0.16 (0.013) | 0.27 (0) |
| Avg | -0.03 (0.429) | 0.01 (0.757) | na | na | na | na | na | -0.03 (0.429) | 0.01 (0.757) | -0.13 (0.047) | -0.11 (0.018) |
| Min | -0.01 (0.814) | -0.01 (0.745) | na | na | na | na | na | -0.01 (0.814) | -0.01 (0.745) | 0.21 (0.001) | 0.3 (0) |
| PageRank Δ | | | | | | | | | | | |
| Max | -0.05 (0.1) | 0 (0.949) | -0.01 (0.829) | 0.07 (0.213) | -0.03 (0.571) | -0.07 (0.275) | 0.07 (0.257) | -0.04 (0.281) | 0.05 (0.195) | na | na |
| Avg | -0.06 (0.027) | -0.05 (0.086) | -0.02 (0.772) | 0.01 (0.882) | -0.04 (0.543) | -0.12 (0.051) | 0.03 (0.6) | -0.08 (0.048) | 0.03 (0.533) | na | na |
| Min | 0.11 (0) | 0.07 (0.017) | 0.03 (0.55) | 0.11 (0.048) | 0.13 (0.03) | 0.22 (0) | 0.06 (0.346) | 0.14 (0.001) | -0.03 (0.471) | na | na |
| Rule Quality Δ | | | | | | | | | | | |
| Sup | -0.03 (0.465) | -0.07 (0.059) | -0.02 (0.822) | 0.04 (0.634) | -0.07 (0.402) | -0.18 (0.027) | -0.08 (0.361) | na | na | na | na |
| Conf | 0.03 (0.336) | 0.01 (0.765) | 0.02 (0.729) | 0 (0.954) | 0 (0.938) | 0.05 (0.418) | 0.24 (0) | 0.1 (0.029) | -0.09 (0.038) | -0.1 (0.161) | -0.15 (0.006) |

Based on these results we can reject the null hypothesis that rule length and plausibility are uncorrelated on two datasets (Mushroom and Quality), but not on the remaining two (Movies and Traffic). There is a marked difference between the absolute value of correlation on V2 Quality and Mushroom datasets (0.2 vs 0.37), which we analyse within Proposition 2.

#### 5.4.2.7. P 2: Higher Plausibility of Longer Rules Is Caused by Misunderstanding of "and"

In order to gauge the effect of *misunderstanding of "and"*, we carried out a separate set of experiments with V1 instructions. The only difference between V1 and V2 instructions were the intersection test questions included in V1, but not included in V2. These test questions were intended to ensure that all subjects in V1 version understand the *and* conjunction the same way it is defined in the probability calculus. The analysis is focused on the Mushroom and Quality datasets, because only on these two datasets we have observed higher plausibility for longer rules.

The results for V1 setup, also depicted in the first two rows of Table 5.13, show that the correlation coefficient is statistically significantly different from zero for the Mushroom dataset with Kendall's $\tau$ at 0.28 ($p < 0.0001$), but not for the Quality dataset which has $\tau$ not different from zero at $p < 0.05$. The remaining two datasets are not relevant, because as follows from the results of Proposition 1, rule length and plausibility are not correlated on these datasets even when misunderstanding of "and" is not controlled for (results of V2 instructions).

On the Quality dataset, we cannot reject the null hypothesis that the correlation coefficient between rule length and plausibility is equal to zero. On the Mushroom dataset, the correlation dropped but still remained statistically significant, therefore we can reject the null hypothesis. This suggests that on the Mushroom datasets there are other factors apart from misunderstanding of "and" that cause longer rules to be perceived as more plausible. Which factors these are is analysed within Proposition 3.

#### 5.4.2.8. P 3: Accounting Plausibility of Longer Rules to Cognitive Biases

Our results for propositions P 1 and P 2 presented above indicate that preference for longer rules is present with statistical significance after the misunderstanding of "and" has been controlled for only on the Mushroom dataset. Results of basic correlation analysis are in the first two rows of Table 5.13. Here, we investigate whether this preference can be explained by the following control variables: literal relevance, attribute relevance and PageRank [Page et al., 1998]

We need to control for the effect of selected variables. We compute *semi-partial correlations* between rule length and plausibility controlling for the effect of the various biases. Semipartials[10], denoted as $r(y|z, x)$ remove the effect of control variable $x$ (proxy for a spe-

[10] Why semi-partial correlation and not partial correlation? We decided for semi-partial correlations because we wanted to measure the incremental variance rule length adds into rule preference (plausibility)

cific bias) from the independent variable $z$ (Rule length $\Delta$), but not from the dependent variable $y$ (plausibility).

The analysis is performed on results for the V 1 instructions, which contain the "swap" test questions ensuring that subjects pay attention to the task and also the intersection test questions that ensure expected understanding of the "and" conjunction.

Controlling for the proxy variables on the Mushroom dataset decreases the value of $\tau$. The effect of the controls can be observed in the V1 Mushr. column of Table 5.13. The strongest effect is recorded for minimum attribute relevance, a proxy for the availability heuristics, which drops the value of $\tau$ to 0.12 ($p = 0.026$). Controlling individually for any of the proxies does not reduce the correlation between plausibility and rule length to zero. As the results in the bottom of the V1 Mush. column show, even controlling for multiple biases does not decrease the correlation coefficient below 0.12. The conclusion is that we were able to account part of the plausibility of longer rules to the availability heuristic.

#### 5.4.2.9. Methodological Considerations Relating to Results for P4 – P8

The validation of proposition P3 has shown that the availability heuristic can account for part of higher plausibility of longer rules. This result was obtained with Version 1 instructions, which remove the effect of misunderstood "and". However, the results in Table 5.14 show a marked difference in correlations between Version 1 and Version 2. For example, Kendall's $\tau$ for maximum literal relevance on the Quality dataset is at 0.31 for Version 2 statistically significant, and for Version 1 with $\tau = 0.06$ statistically insignificant. Similar drop in correlations can be observed for nearly all datasets and variables.

We are reluctant to explain this change in correlations solely based on the misunderstanding of "and", therefore it cannot be ruled out that the intersection test questions in Version 1 might have introduced some unforeseen side-effect. For this reason, we opted to perform the analysis in this section using Version 2 instructions, however, for comparison we also report results for Version 1 in Table 5.14. An alternative analysis of the affect of availability heuristic on Version 2 of the instructions is performed within P4. Version 2 will also be used to evaluate the effect of the remaining cognitive biases (P5 - P7). The limitation of basing the analysis on Version 2 instruction is that the results can be confounded by the effect of misunderstanding of "and". P 8 will be evaluated on Version 3 instructions, which were specifically designed to evaluate the effect of confidence and support.

#### 5.4.2.10. P 4: Availability Heuristic

Proposition P 4 is focused on the availability heuristic, which uses attribute relevance as a proxy. For this proxy variable, we have data from two datasets: Mushroom and Traffic. As justified in Subsection 5.4.2.9, the analysis is performed using V2 instructions.

above and beyond the effect of the other variables – the proxies of cognitive biases and heuristics. In contrast, the partial correlation would reflect the situation when all observations would have the same value of the controlled variable. It should be noted that the value of the semi-partial correlation is always lower or equal to the value of the partial correlation. We used the implementation available in the R `ppcor` package [Kim, 2015]

As follows from Table 5.14, the Kendall's $\tau$ between maximum attribute relevance and plausibility reaches of 0.27 for the Mushroom dataset. This supports our hypothesis that availability heuristic contributes to plausibility.

On the Traffic dataset the correlation coefficients between attribute relevance and plausibility are not significantly different from zero. This can be explained by a design issue in the corresponding experiment, since we attempted to construct the attributes for the Traffic dataset manually as they were not readily available in the data. This explanation is supported by data in Table 5.4 on page 78, which shows that attributes for the traffic dataset were considered as less relevant than attributes of the Mushroom dataset.

#### 5.4.2.11. P 5: Weak Evidence Effect

For the analysis of weak evidence effect we use literal relevance proxy variable. This variable is available for all LOD datasets. As justified in Subsection 5.4.2.9, the analysis is performed using V2 instructions.

We expected to obtain positive correlation between plausibility and average and maximum literal relevance. Such outcome would confirm the normative way or reasoning: increased strength of evidence results in increased plausibility. The weak evidence effect should demonstrate through negative correlation with minimum literal relevance: a literal with low relevance triggers the weak evidence effect.

Correlation values between literal relevance and plausibility as reported in Table 5.14 were congruent with our hypothesis on two LOD datasets: Quality and Movies. According to our expectations, we obtained negative correlation between minimum literal relevance and plausibility on all three LOD datasets. On the Quality dataset, this correlation was statistically significant ($p \leq 5\%$).

On the Mushroom dataset we obtained a positive correlation. This surprising result can be explained above by the fact that on this dataset the shortest rules contained the strongest literals. To shed some light on this phenomenon we included the actual rule pairs in the Mushrooms dataset into Table 5.15. Out of the 10 rule pairs, there were five pairs consisting of a discriminative rule of length 1 and a longer rule. The rules of length 1 consist of a single literal with relevance higher or equal to the maximum of the literal relevance in the longer rule.

Our explanation for the unexpected values of correlation for maximum and average literal relevance is that there were two effects in opposite direction, which largely neutralized each other:

- Some subjects preferred shorter rules, because they contained the most relevant literal.
- Some subjects opted for longer rules because they considered the otherwise appealing shorter rule as "oversimplistic". Such explanation is congruent with other results and opinions reported in the literature [Lavrač, 1998, Freitas, 2014, Elomaa, 1994].

Literal relevance turned out to have the highest correlation of all proxy variables with

Table 5.15.: Rule pairs in the Mushroom dataset. cons. stands for consequent (edible or poisonous), r1len for length of rule 1, r1c stands for confidence of rule 1, r1s for support of rule 1, r1MLR for maximum literal relevance in rule 1. Similarly for r2len, etc.

| id | r1 antecedent | r2 antecedent | cons. | r1len | r2len | r1s | r2s | r1c | r2c | r1MLR | r2MLR |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | veil-color=w,gill-spacing=c,population=v,stalk-shape=e,stalk-root=b,spore-print-color=r | spore-print-color=r | p | 6 | 1 | 0.009 | 0.009 | 1 | 1 | 0.6 | 0.6 |
| 2 | gill-size=b,stalk-surface-above-ring=s,ring-number=o,spore-print-color=n | odor=l | e | 4 | 1 | 0.179 | 0.049 | 1 | 1 | 0.4 | 0.6 |
| 3 | odor=n,ring-number=o,stalk-surface-above-ring=s,cap-shape=x | odor=n,stalk-surface-below-ring=f | e | 4 | 2 | 0.144 | 0.056 | 1 | 1 | 0.4 | 0.4 |
| 4 | gill-size=n,spore-print-color=w,stalk-color-below-ring=w | veil-color=w,ring-number=o,gill-size=n,spore-print-color=w,stalk-color-below-ring=w | p | 3 | 5 | 0.107 | 0.107 | 1 | 1 | 0 | 0.4 |
| 5 | veil-color=w,gill-spacing=c,bruises?=f,ring-number=o,stalk-surface-below-ring=k | odor=f | p | 5 | 1 | 0.266 | 0.266 | 1 | 1 | 0.4 | 1.4 |
| 6 | spore-print-color=r | veil-color=w,gill-spacing=c,population=v,cap-surface=s | p | 1 | 4 | 0.009 | 0.139 | 1 | 0.993 | 0.6 | 0 |
| 7 | odor=a | gill-size=b,stalk-surface-above-ring=s,ring-number=o,stalk-surface-below-ring=s,spore-print-color=k | e | 1 | 5 | 0.049 | 0.15 | 1 | 1 | -0.6 | 0.2 |
| 8 | odor=n,population=y | gill-size=b,odor=n,cap-shape=x | e | 2 | 3 | 0.119 | 0.182 | 1 | 1 | 0.4 | 0.4 |
| 9 | odor=n,cap-color=n,stalk-surface-below-ring=s | odor=n,ring-number=o,stalk-surface-above-ring=s,population=v | e | 3 | 4 | 0.125 | 0.13 | 1 | 1 | 0.4 | 0.4 |
| 10 | odor=c | veil-color=w,gill-spacing=c,bruises?=f,ring-number=o,population=v,stalk-root=b | p | 1 | 6 | 0.024 | 0.086 | 1 | 1 | 0.8 | 0.6 |

plausibility. The maximum $\tau$ at 0.31 ($p < 0.001$) was reached on the Quality dataset. Overall, we obtained limited evidence in favour of the weak evidence effect. As further elaborated in Section 5.6.2, the contradicting evidence obtained on the Mushroom dataset can be attributed to explanations other than the weak evidence effect not demonstrating.

#### 5.4.2.12. P 6: Insensitivity to Sample Size

In order to gauge the effect of confidence and support, we refer to results obtained on Version 3 of instructions (Figure 5.11, page 88), which included explicitly stated confidence and support. The only dataset where V3 instructions were applied is the Movies dataset, since the differences in confidence and support between the rules in the pairs were not large enough for the remaining datasets.

Table 5.14 (bottom) presents the correlations between confidence, support and rule plausibility. The results show that the plausibility is related to confidence ($\tau = 0.24$, $p < 0.0001$) but not to support ($\tau = -0.08$, $p < 0.36$). The results also show that the relationship between revealed rule confidence and plausibility is causal. This follows from confidence not being correlated with plausibility on the V2 instructions, which differed from V3 only by the absence of explicitly revealed rule quality. Overall, we found unanimous evidence in favour of the insensitivity to sample size effect. The limitation of this finding is that this result is based only on one dataset.

#### 5.4.2.13. P 7: Disjunction Fallacy

We hypothesized that minimum PageRank associated with any literal in the antecedent of a rule can serve as a proxy for the specificity of the antecedent and thus can be used to model the disjunction fallacy – the higher the value of the proxy variable the stronger the disjunction fallacy. According to disjunction fallacy, the subjects will prefer rules with more specific literals – literals with lower minimum PageRank. As justified in Subsection 5.4.2.9, the analysis is performed using V2 instructions.

Referring to Table 5.14, the correlation between minimum PageRank and plausibility was found to be statistically significantly positive ($p < 0.05$) for Quality and Movies. The highest Kendall's $\tau$ of 0.22 was reached on the Movies dataset. The correlation coefficient for the Traffic dataset was not different from zero at $p = 5\%$. Possible reasons are subject of discussion in Section 5.6.2.

Overall, we found evidence for disjunction fallacy on two datasets and no evidence on one dataset.

#### 5.4.2.14. P 8: Mere Exposure Effect

We hypothesized that average and maximum PageRank associated with any literal in the antecedent of a rule can serve as a proxy for the number of prior exposures and thus can be used to model the mere exposure effect – the higher the value of the proxy variables the stronger the mere exposure effect and the higher the preference for the rule. As justified in Subsection 5.4.2.9, the analysis is performed using V2 instructions.

None of the correlation coefficients for PageRankDelta (Table 5.14) is statistically significantly different from zero at $\alpha = 0.05$. We thus found no evidence for the mere exposure effect.

## 5.5. Experiment 2: Variations on Linda

In addition to Experiment 1 performed directly on inductive rule learning data, we also prepared an experiment on the Linda problem [Tversky and Kahneman, 1983], which was introduced in Subsection 4.4.3.1 and Figure 4.1 on page 44. The motivation for this experiment is described in Subsection 5.1.2 on page 67. Three version of instructions were prepared.

The first group of subjects in Experiment 2 received the Linda problem, which is likely the most studied problem in the cognitive biases domain. The comparison of our results with those that were previously published for the controlled laboratory and setting and for Amazon Mechanical Turk can help judge credibility of our main empirical results obtained within Experiment 1.

The second group of subjects received a modified version of the Linda problem to test for the effect of negation. Finally, two groups of subjects (there were two variation of instructions) participated in investigation of the effect of unknown value on the information bias. It was not possible to test these effects within Experiment 1, since this was performed on rules authentically discovered with rule learning algorithms that do not support negation. Similarly, incorporation of unknown value would require changes to the data used within this experiment. We found this therefore natural to use manually designed modifications of the Linda problem to test for these effects.

### 5.5.1. Method

#### 5.5.1.1. CrowdFlower Setup

The workers in the CrowdFlower platform were invited to participate in individual tasks. The subjects were remunerated for their answer by 10 US cents (without bonus), which exceeds the minimum amount of half-a-penny given in Schnoebelen and Kuperman [2010]. To further ensure that the pay is appropriate, we checked the satisfaction scores reported in the final questionnaire by the subjects. On a 1-5 Likert scale (1 worst, 5 is best), the average subject rating of their remuneration was between 3.6 to 4.2. This experiment consisted of only one question ("row"). Answers were collected from 150 distinct subjects. There were no test questions. Instead, we offered 50% bonus for quality to subjects who provided reason for their answer longer than 10 characters. The subjects were informed that 50% bonus for quality will be awarded but it was not conveyed that quality will be measured by the length of the answer. For analysis, we used all data including the answers with no or short reasons. Since the language of the assignment was English, we restricted the geographic eligibility of the task to subjects residing in U.S., Canada and United Kingdom. As the level of the CrowdFlower workers, we chose Level 2. For each of

the four versions of instructions we collected 150 judgments. Participants were recruited anew for each version of the instructions from among the cohort of CrowdFlower workers.

#### 5.5.1.2. Material

We prepared four versions of instructions, with the following purpose:

1. Replicate the original results of Tversky and Kahneman [1983] using crowdsourcing.
2. Determine the effect of negated condition.

3A. Determine the effect of information bias related to inclusion of a condition with unknown value.

3B. Determine the effect of information bias (a variation).

We prepared four versions of the experiment that differed only in the wording of the options (the text between "Which is more probable?" and "BONUS) in the following:

```
Overview

We kindly ask you to assist us in an experiment which will help to
understand how humans perceive rules describing data.

Consider the following description of a person.

Jenny is 32 years old, single, outspoken, and very bright. She majored
in philosophy. As a student, she was deeply concerned with issues of
discrimination and social justice, and also participated in
anti-nuclear demonstrations.

Which is more probable?

    Jenny is a bank teller.
    Jenny is a bank teller and is active in the feminist movement.

BONUS

High quality responses will be rewarded by 50% of the original credit.
Thank you for your assistance !
```

In all our experiments, we replaced the name Linda used in the original paper with Jenny. The reason for this modification is to make it somewhat more difficult for subjects to quickly find the solution to the problem using a search engine. Due to the crowdsourcing

setting, there was no restriction to the access to Internet. Otherwise, the wording of the description of Linda is the same as given by Tversky and Kahneman [1983, page 297].

**Jenny Version 1.** The first version of instructions aimed to replicate the Linda experiment. The options were thus the same as described by Tversky and Kahneman [1983, page 299], except that the answers were not marked by "T" and "T&F", but the subjects were choosing the preferred answer by making a selection from a dropdown box with the following options:

```
Jenny is a bank teller.
Jenny is a bank teller and is  active in the feminist movement.
```

**Jenny Version 2.** The second version of instructions aimed to test the effect of negation. In contrast to the original formulation of the Linda problem, the version used for this experiment includes negation in the second answer option. The second option became: "Jenny is a bank teller and is *not* active in the feminist movement."

```
Jenny is a bank teller.
Jenny is a bank teller and is not active in the feminist movement.
```

**Jenny Version 3A.** The third version of instructions aimed to test the effect of unknown value. We use a modified version of the Linda problem to investigate the influence of information bias on plausibility. The last option in these instructions involves a relevant condition with unknown value ("it is not known if she is active in feminist movement").

```
Jenny works as a cashier in a bank.
Jenny is not active in feminist movement.
Jenny is a bank teller and it is not known if  she is active in the
 feminist  movement.
```

Both version A and B contain a filler option "Jenny is not active in feminist movement."

**Jenny Version 3B.** Tests the effect of unknown value (a variation). The order of option 1 and 3 is reversed and the terms cashier and bank teller are swapped.

```
Jenny works as a cashier in a bank and it is not known  if she is
 active in the feminist movement.
Jenny is not active in the feminist movement.
```

#### 5.5.1.3. Procedure

Subjects were faced with a web-based user interface, which presented description of Linda, and asked the subjects to select one of the statements. The number and content of statements depended on the version of instructions.

The responses to individual versions of instructions were collected in separate tasks. Example instructions as shown to subjects in the CrowdFlower system are in Figure 5.12.

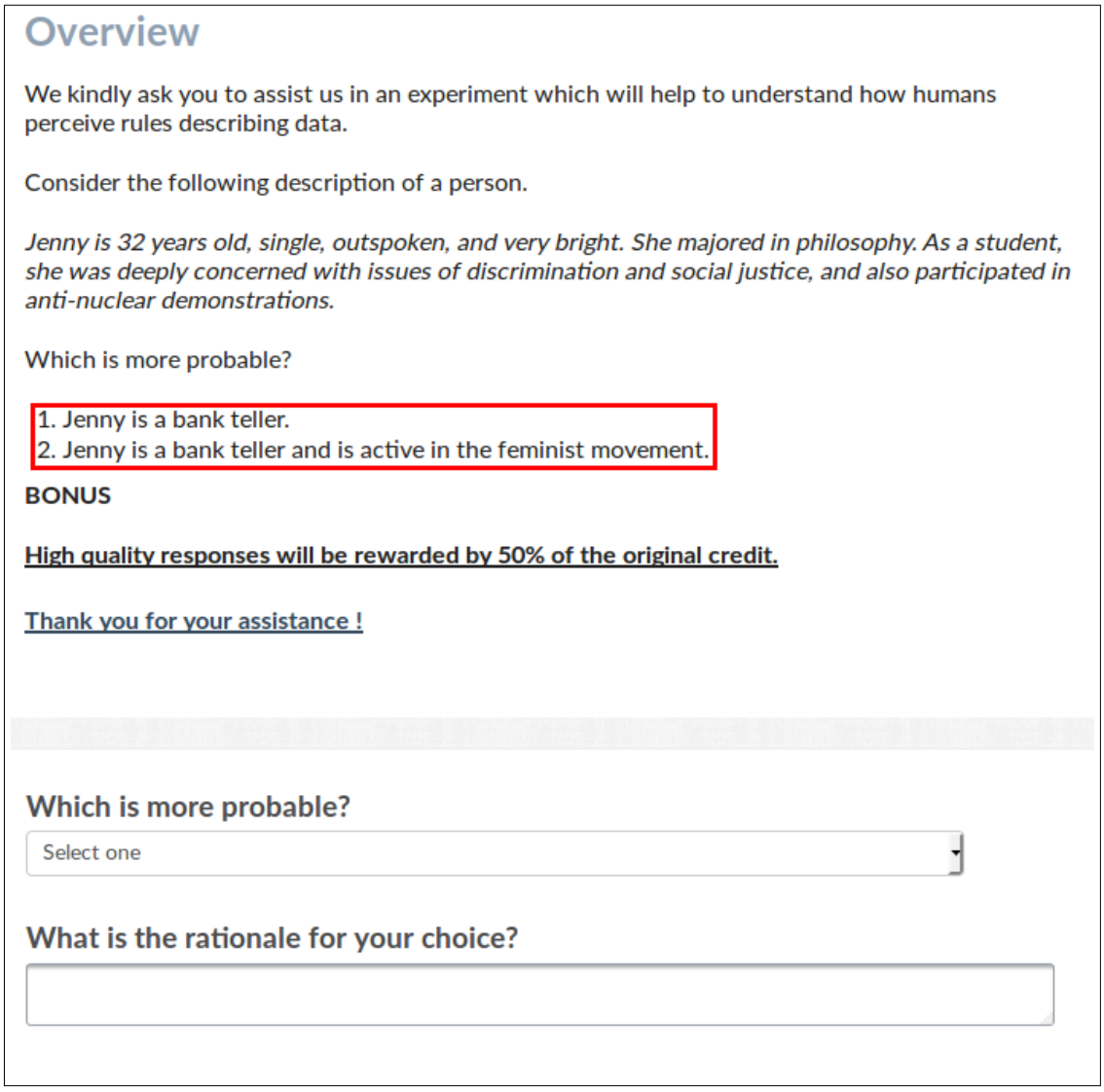



Figure 5.12.: Linda (Jenny) V1 Experiment Instructions. The text in the red rectangle was varied across instructions (groups of subjects). Note that the red rectangle was not part of the instructions.

The options and instructions in the Linda experiment were carried out as described by Tversky and Kahneman [1983] with the following changes:

1. Quality control: subjects were promised a small monetary benefit for "quality" answers. The incorporation of the quality control is justified by the observation reported

by Grether [1992] according to which absence of financial incentive results in increase of incoherent or nonsense responses by factor of three.[11]

2. In order to identify quality answers, we asked participants to give a short justification of their choice. We conjecture that this modification could slightly decrease the fallacy rate, because explicitly formulating the reason could make some subjects realize their reasoning error.

3. We renamed Linda to Jenny to make it more difficult to quickly find the correct answer using a search engine.

### 5.5.2. Results

An overview of characteristics of the participant cohort recruited with crowdsourcing is provided in Table 5.16.

Table 5.16.: Overview of cohort involved in Experiments 2. Each line corresponds to one group of subjects. The *v* column denotes the version of the instructions this group received, *judg* to the number of judgments collected, *workers* to the total number of unique subjects, *usa*, *gbr*, *can* to number of judgments from the United States of America, Great Britain and Canada, *avg dur* to the average duration the subject took to finish one page of work in minutes and seconds and *reasons* number of textual reasons longer than 10 characters.

| dataset | v | judg | workers | usa | gbr | can | avg dur | reasons |
|---|---|---|---|---|---|---|---|---|
| Linda | 1 | 150 | 150 | 69 | 42 | 39 | 03:37 | 137 |
| Linda | 2 | 150 | 150 | 75 | 50 | 25 | 04:00 | 141 |
| Linda | 3a | 150 | 150 | 77 | 40 | 33 | 04:30 | 139 |
| Linda | 3b | 150 | 150 | 77 | 41 | 32 | 04:20 | 135 |
| *total* | | 600 | 600 | 298 | 173 | 129 | na | 552 |

The frequency of responses for all versions of Jenny Experiments are given in Table 5.17.

#### 5.5.2.1. P 9: Replicating Linda Experiment

The experiment $V_L1$ resulted in fallacy rate of 68%. The Linda problem (the same version with two possible answers) was replicated using Amazon Mechanical Turk by Paolacci et al. [2010]. The fallacy rate of 72% that they report is slightly higher than our 68%. This result confirms our methodological assumption that we can use CrowdFlower instead of Amazon Mechanical Turk without experiencing a strong difference in results. As for the relative decrease of fallacy rate by 5.5% that we obtained – we hypothesize that this can be accounted to the same reason used by Harris [2014] to explain the 3.5% improvement in task accuracy on CrowdFlower as opposed to Amazon Mechanical Turk, which is that CrowdFlower workers outperform Amazon Mechanical Turk in terms of accuracy. One

[11] This paper also states that data from scientific experiments with no financial incentives should be treated as "possibly contaminated".

Table 5.17.: Frequency of responses in Jenny (Linda) experiments. The first column indicates the version and option number within the version.

| v/o | text | freq |
|---|---|---|
| $V_L1/1$ | Jenny is a bank teller | 48 |
| $V_L1/2$ | Jenny is a bank teller and is active in the feminist movement | 102 |
| $V_L2/1$ | Jenny is a bank teller | 118 |
| $V_L2/2$ | Jenny is a bank teller and is not active in the feminist movement | 32 |
| $V_L3a/1$ | Jenny works as a cashier in a bank | 37 |
| $V_L3a/2$ | Jenny is not active in feminist movement | 38 |
| $V_L3a/3$ | Jenny is a bank teller and it is not known if she is active in feminist movement | 75 |
| $V_L3b/1$ | Jenny works as a cashier in a bank and it is not known if she is active in feminist movement. | 65 |
| $V_L3b/2$ | Jenny is not active in feminist movement | 44 |
| $V_L3b/3$ | Jenny is a bank teller | 41 |

reservation regarding this conclusion is that the Amazon Mechanical Turk experiment setup is not reported in sufficient detail by Paolacci et al. [2010] to rule out other causes of the higher fallacy rate, such as absence of quality control in the setup of Paolacci et al. [2010], the optional elicitation of reasons for judgment in textual form in our setup, or different level of participating workers.

Additional discussion of reasons for different results obtained with crowdsourcing compared to the standard laboratory environment is presented in Subsection 5.6.4.1.

#### 5.5.2.2. P 10: Effect of Negation on Representativeness Heuristic

We used a modified version of the Linda problem to answer the question to what extent do people semantically interpret negations. We expected a (considerable) proportion of subjects to "overlook" negation and instead determine the answer based on the words that have a direct connection to the description of Linda. In other words, we expected that some subjects will associate the "feminist" word representative in answer option $V_L2/2$ with Linda/Jenny description even though this word is negated.

Out of the 150 subjects, only 21% (32) preferred the longer option with negation as opposed to 68% (102) for the longer "positive" option in the baseline experiment. The difference in proportion is statistically significant at $p < 0.0001$. The results show that the negation was semantically interpreted and the term "not active in the feminist movement" was understood as not representative of Linda description by most subjects. However, the results also show that a considerable percentage of subjects (32%) chose the negated option not representative of Linda, which may suggest that, at least for some of these subjects, negation did not inhibit the representative effect of the property that followed the negation.

#### 5.5.2.3. P 11: Triggering Information Bias by Unknown Value

The design of the third variation of Linda instructions was motivated by the following problem: we hypothesized that if rule includes a condition with unknown value, the perceived plausibility of this rule compared to the same rule without this condition will increase. We investigated this phenomenon using a more general formulation based on Linda instructions, rather than by designing a rule-based experiment. We hypothesized that making a salient characteristic of Linda unknown in one of the answer options will, following the representativeness heuristic, make this option more preferred. Similarly to the effect of negation (P10) we hypothesized that a considerable proportion of subjects will "overlook" the expression of missing value "not known if she is feminist" and will consider the answer options containing the missing values as more representative of Linda than a shorter answer option with equivalent meaning, but without the explicit "not known".

Referring to Table 5.17, the corresponding null hypotheses are that a) proportion of answer option $V_L3a/1$ is not different from the proportion of answer $V_L3a/3$, b) proportion of answer no. $V_L3b/1$ is not different from the proportion of answer $V_L3b/3$. In both versions of instructions, the option containing relevant condition with the unknown value has the highest frequency. In variation $V_L3a$, the frequency of option 3 is 107% higher than the frequency of the baseline option 1, which is 37. In variation $V_L3b$, the corresponding increase is 59% (65 vs 41). In both cases, the difference in proportion is statistically significant at $p < 0.001$.

## 5.6. Summary of Results and Discussion

This section summary of results of for Experiment 1 and Experiment 2. The results are organized according to the level of evidence that we obtained: from unspurious evidence for insensitivity to sample size to no evidence for the mere exposure effect.

### 5.6.1. Biases with Unspurious Evidence

#### 5.6.1.1. Insensitivity to sample size (P6)

We aimed to evaluate the effect of explicitly revealed confidence (strength) and support (weight) on rule preference. The *insensitivity to sample size* effect suggests confidence should be given preference over support. In real situations, rules on the output of inductive rule learning have varying quality, which is communicated mainly by the values of confidence and support. Results obtained in cognitive science for the strength and weight of evidence suggest that weight of evidence is systematically undervalued while the strength of evidence is overvalued. This result is congruent with the hypothesis that insensitivity to sample size effect is applicable to interpretation of inductively learned rules. In other words, when both confidence and support is stated, confidence positively affects plausibility and support is ignored.

### 5.6.2. Biases with Limited Evidence

#### 5.6.2.1. Availability Heuristic (P4)

The evidence for the availability heuristic is mixed. We obtained results supporting its occurrence for the Mushroom dataset, but not for the Traffic dataset (only for these two datasets our proxy heuristic was available). The fact that it did not demonstrate on the Traffic dataset can be explained by a design issue in creating the proxy values for this dataset, therefore the results on this dataset can be – with caution – disregarded. Consequently, there are results on one dataset which support the availability heuristic.

#### 5.6.2.2. Weak Evidence Effect (P5)

As noted in the results section, we found limited confirmation for the weak evidence effect on the Quality dataset. The correlation between plausibility and minimum literal relevance (the designated proxy for this effect) was also negative on the Movies dataset, but the size of the correlation coefficient was only marginally statistically significant at $\alpha = 0.05$ (p-value $= 0.072$).

Literal relevance was not correlated with plausibility on the Traffic dataset. We attribute it to the fact that the relevance of literals in the Traffic dataset is the lowest from all datasets as follows from Table 5.5. The results for the Mushroom dataset were completely contrary to our expectations: the correlation between maximum and average literal relevance and plausibility negative, and the correlation with minimum literal relevance positive.

To conclude, we found evidence for the weak evidence effect on one dataset on interpretation of rule learning results, but further research is required to verify our findings.

#### 5.6.2.3. Disjunction Fallacy (P7)

The results supporting the incidence of disjunction fallacy were obtained on two of the three datasets (Movies and Quality). On the Traffic dataset the correlation between the proxy for the bias and plausibility was not statistically significantly different from zero.

The reason why the effect of PageRank was most pronounced on the Movies dataset can be attributed to the fact that on this dataset, the values of PageRank in our opinion best represented the subjects' perception of the literal as no background geographical knowledge was required (in contrast to the Traffic dataset). While some geographical knowledge was required for the Quality dataset, this was, we believe, lower than for the Traffic dataset (cf. Figure 5.6 on page 84 for examples drawn from all three datasets).

To illustrate the relation between minimum PageRank and plausibility on the Traffic dataset let us look at the following example.

**Example. (Disjunction fallacy)**

- r1: *RegionalCapitalsInSenegal=1.0 => Label=low.* PageRankMin=1.15
- r2: *PortCitiesInAfrica=1.0 AND CapitalsInAfrica=1.0 => Label=low.* PageRankMin=3.65

The first rule is more specific, has lower minimum PageRank, and therefore should be found according to the disjunction fallacy as more preferred. The PageRankMinDelta for this pair is 2.5.

Overall, the results support the hypothesis that disjunction fallacy affects interpretation of rule learning results. The triggering of the fallacy is contingent on subjects possessing the domain knowledge that allows them to evaluate specificity of the concepts included in the rules.

### 5.6.3. Biases with No Evidence

#### 5.6.3.1. Mere Exposure Effect (P8)

We found no evidence for the mere exposure effect using the designated proxy variables derived from PageRank.

### 5.6.4. Linda Experiments

#### 5.6.4.1. Replicating Linda Experiment with Crowdsourcing (P9)

The first version of our Jenny instructions aimed to verify that the crowdsourcing setting yields similar results to those reported earlier for Linda in the literature. We found agreement with prior results obtained with crowdsourcing. However, the fallacy rate that we obtained is different from what was obtained in the traditional laboratory environment. In the following, we will try to identify the possible reasons.

The original fallacy reported by Tversky and Kahneman [1983] is based on experiment, where answers were elicited from 88 students of University of British Columbia and no incentives were provided. Charness et al. [2010] reported that providing an incentive decreased the fallacy rate to 33% (94 total subjects) and without incentive they report fallacy rate of 58% (68 subjects). Charness et al. [2010] elicited answers from students of University of California, Santa Barbara. Based on these reports, we can expect the fallacy rate to be between 33% and 85%.

In our experiment $V_L1$ the fallacy rate was 68%, which is significantly different from the original fallacy rate of 85% at $p < 0.01$ (test for equality of proportions). If we limit the analysis to the outcome of the statistical test, we have not replicated the original results.

A critical analysis of our replication yields to the objection why we did not more strictly adhere to the original setup of the experiment by Tversky and Kahneman [1983]. In other words, was it necessary to provide incentives? In our opinion, the crowdsourcing environment does not allow for verbatim replication of the original experiment, which did not

provide any incentives to subjects. This position is supported by number of observations of prior experimenters as summarized in Oleson et al. [2011]: "All authors agree that crowdsourcing requires explicit quality assurance mechanism to deal with scammers, insufficient attention and incompetent workers."

Another possible criticism is that if we decided to provide incentives, then why not compare the result with Charness et al. [2010], who provided incentives, rather than with Tversky and Kahneman [1983], who did not. In this respect, Charness et al. [2010] provided incentive of 4 dollars for correct answer, while we provided 5 cents for "quality answer". As follows from Schnoebelen and Kuperman [2010], such a high reward would be unusual for crowdsourcing.

Our conclusion is that the fallacy rate that we obtained with crowdsourcing for Linda problem with a small incentive is in the range reported in the literature for experiments in the controlled laboratory setting and very closely matches previous results obtained on another crowdsourcing platform [Paolacci et al., 2010].

#### 5.6.4.2. Effect of Negation on Representativeness Heuristic (P10)

We obtained convincing experimental evidence showing that negation is semantically interpreted and affects the application of the representativeness heuristic. The results also show that a considerable percentage of subjects (32%) chose the negated option not representative of Linda, which may suggest that, at least for some of these subjects, negation did not inhibit the representative effect of the property that followed the negation.

#### 5.6.4.3. Triggering Information Bias by Unknown Value (P11)

The response rates that we obtained for Linda experiments suggest that inclusion of a condition with no information value has effect on increasing plausibility of an option. However, further analysis shows that the underlying reason is not entirely caused by representativeness heuristic triggered by "not known if she is active in feminist movement", but instead a different interpretation of the short answer option "Jenny is a bank teller." From reasons given in the textual answers[12] (justifications input the subjects) it follows that some subjects chose the option with the unknown value because they interpreted "Jenny is a bank teller" as "Jenny is a bank teller and NOT active in feminist movement." Second, the option "Jenny is a bank teller and it is not known if she is active in feminist movement" was interpreted as "Jenny is a bank teller and she MIGHT be active in feminist movement."

This roughly corresponds to observation of Hilton [1995], who analysed conjunction errors for the answer option in Figure 5.13 reported by Tversky and Kahneman [1983]. According to Hilton [1995, p. 260], this option was interpreted as "Linda is a bank teller even if she is a feminist."

Considering our results, given that not being active in the feminist movement is not representative of Jenny, while possibly being active is, the option with the unknown value was considered as more representative.

[12] Refer also to point 8 of Subsection 6.1.2, where specific textual reasons given by subjects are presented.

Linda is a bank teller whether or not she is active in the feminist movement.

Figure 5.13.: Related answer option used by Tversky and Kahneman [1983] to the one tested in our Linda V3 instructions.

Number of earlier results (cf. Sides et al. [2002] for a reference list) showed that in presence of alternative "$B \wedge F$", alternative "$B$" is interpreted as "$B \wedge \neg F$" (B denotes Bank teller and F feminist). In other words, "Linda is a bank teller" is interpreted as "Linda is a bank teller and not a feminist." Our results provide further confirmation for the observation that such interpretation is applied by some subjects to alternative "$B$" in presence of alternative "$B \wedge F$ *is unknown*".

This type of misinterpretation can occur in rule learning context when the analyst considers two rules, one not containing a literal and the second rule containing the literal (possibly with unknown value). Rules thus should be presented in such a way that it is made clear that the absence of a condition does not mean negation. The meaning of the unknown value should also be clarified.

# 6. Conveying Effect of Cognitive Biases on Interpretation of Rules

Based on literature review and our experimental results presented in the previous chapters, we propose a visual model of plausibility of rules describing which cognitive biases are triggered when humans assess plausibility of inductively learned rules and whether they influence plausibility in positive or negative way. This model is intended to raise awareness about the effect of cognitive biases on perception of rule learning results among the designers of machine learning algorithms and software.[1] This visual model is complemented with a list of practical recommendations for the same audience.

**Chapter organization.** Section 6.1 presents the visual model. Section 6.2 presents the recommendations.

## 6.1. Visual Qualitative Model

In this section, we propose a visual model that summarizes the effect selected cognitive biases have on human assessment of plausibility of rules discovered from data.

The model consists of two decision trees, which are presented in Figure 6.1. The first tree captures the contributions of individual literals in the antecedent of the rule towards increase or decrease of human perceived plausibility of the rule. The second tree shows how the individual literal contributions are combined to perception of plausibility of the complete rule. Subsection 6.1.1 describes the sources of evidence that we used to construct the plausibility model. The first tree in the model is covered by Subsection 6.1.2 and Subsection 6.1.3 covers the second tree.

### 6.1.1. Evidence

All nodes in the trees are numbered and described in the text under the corresponding number. Additionally, for leaf nodes the following pieces of information are provided:

- EVIDENCE: a justification for the value (positive or negative preference). We distinguish between insights obtained by qualitatively analysing the textual justifications of the answers and statistical analysis.
- BIAS: responsible cognitive biases, heuristics, fallacies or similar effects described in the literature.

[1] The prospective human users of rule learning models are called "analysts" in this chapter.

- STRENGTH: our indicative estimate of the strength of the contribution.

#### 6.1.1.1. Analysis of Textual Responses

As part of all experiments reported in Chapter 5, we asked subjects to provide a short explanation for their answer. For some groups of subjects, these explanations were processed to provide subjects who completed the explanation with a bonus. Beyond this automatic processing, we decided not to perform more detailed analysis within Chapter 5, which draws its conclusions based on statistical analysis of results as is the norm in prior research of cognitive biases in psychology. The explanations provided by subjects are too varied to allow for quantitative analysis. Nevertheless, analysis of content of the textual responses can reveal reasons why subjects committed a specific fallacy, which is important for the qualitative model described within this chapter. Here we use qualitative analysis of the explanations to support our qualitative model.

**Methodology.** For selected elements (decision nodes) in the qualitative model, the author went through all textual explanations and tried to identify those that support or refute it. If any such responses were found, then several representative quotes were selected and included as support evidence into the model. Such pieces of evidence are marked as *Evidence (textual explanations)*.

**Amount of processed data.** Only textual explanations from groups included in the Experiment 1 were processed. This amounts to 1771 responses longer than 10 characters. The breakdown of the number of responses elicited from individual groups of subjects is included in Table 5.10 on page 89.

**Supplementary material.** The responses for all participants are available in supplementary material (cf. Appendix). The explanations cited in this chapter are featured in a minimally edited form, including the original spelling errors, etc.

#### 6.1.1.2. Statistical Analysis

The results that we include within *Evidence (statistical analysis)* have been in part sourced from cognitive science research papers covered in Chapter 4. Our empirical results obtained in Chapter 5 were used to provide evidence for nodes 4, 5, 7, 8, 10 of the first tree and for nodes 16 and 17 in the second tree.

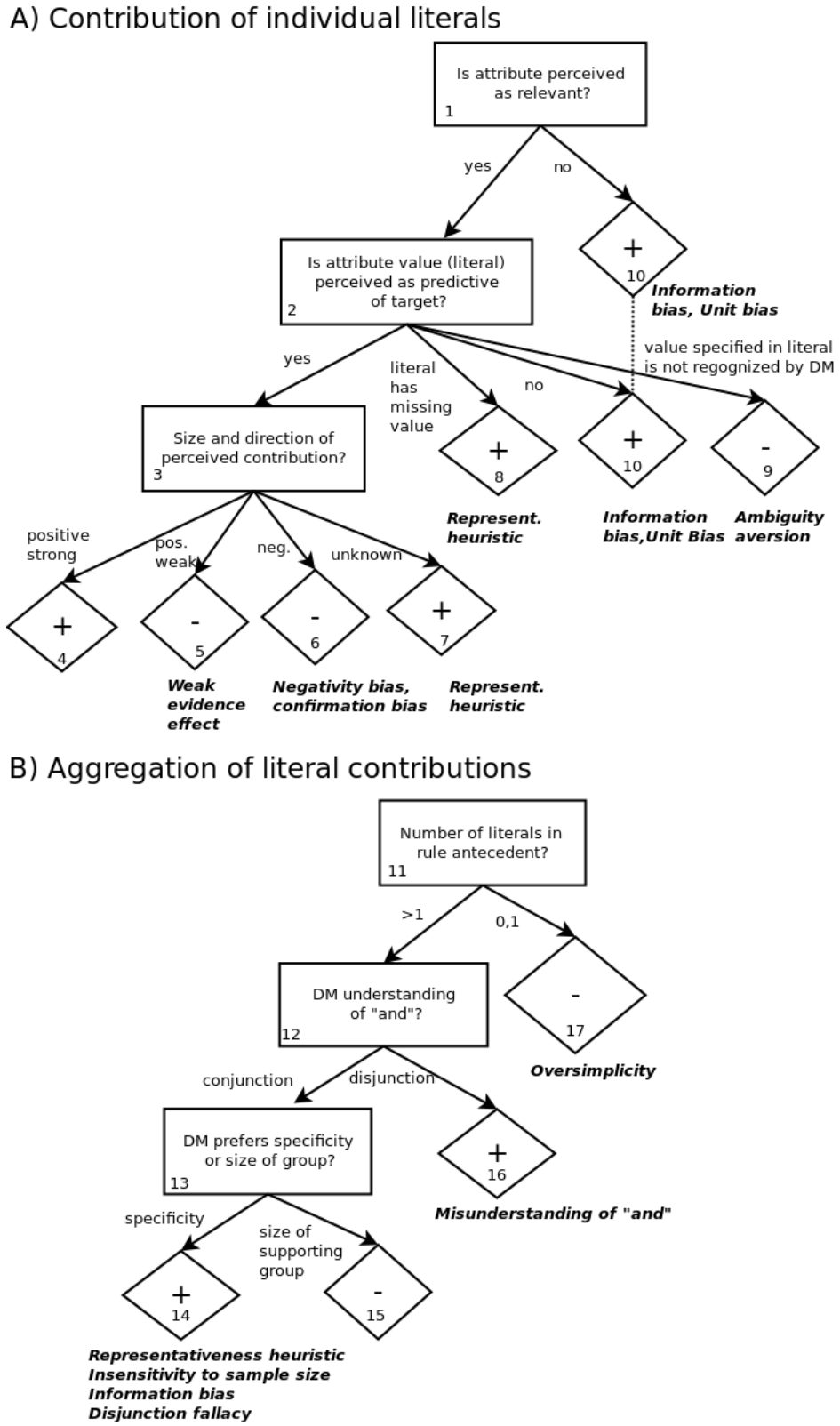



Figure 6.1.: Proposed qualitative model for determining human-perceived plausibility of inductively learned rules. "+" means increase, "-" decrease of plausibility. Some leaf nodes are associated with hypothesized list of effective biases and heuristics (in bold). Explanation of numbered nodes is in the text.

### 6.1.2. Contribution of Individual Literals

Tree depicted in Figure 6.1A models a decision process that applies to individual literals in the rule. This process largely depends on the domain knowledge of the analyst.

**1: Is attribute perceived as relevant?** The analyst evaluates if the attribute is considered as relevant for the problem at hand. In order to assess attribute relevancy, analyst does not need to possess knowledge on the specific values of the attribute or on their link to the target variable.

*Example (no):* "There is nothing to determine whether the cap shape has any link to poison." *Example (yes):* /different subject about cap shape/ "Mushrooms with caps that are wide and look like an umbrella are poisonous and should be avoided."

**2: Is attribute value perceived as predictive of target?** The analyst determines plausibility based on the perceived correlation between the attribute value (literal) and the target class.

*Example (yes):* "The smell of almond indicates poison (i.e. cyanide)." *Example (no):* "there are many different mushrooms with a brown cap colour this does not mean they are poisonous", "Without a book, I would have no idea" (the subject may consider gill size as predictive, but does possess the specific knowledge). The literal can also have a *missing value*, such as "IF odour=cinnamon and bruises=unknown THEN mushroom is poisonous". There is no real supporting answer of a subject, since we did not have such example in our dataset.

**3: Positive or negative correlation?** If the attribute value is perceived as correlated with the target class, the effect on plausibility will relate to the sign and strength of the correlation. *Example (weak):* "Many mushrooms with narrow gills are edible and can be found in the supermarket, but that is not to say that all mushrooms with narrow gills are fit to eat." *Example (strong):* "The colour red in mushrooms normally denoted poison." *Example (unknown):* "Without a book, I would have no idea" (relating to broad gill size).

**4: Perceived positive contribution increases plausibility if strength exceeds threshold.**

- *Evidence (statistical analysis)*: Literal relevance is positively correlated with plausibility on the LOD datasets.[2]
- *Evidence (textual answers)*: For the mushroom dataset there is a number of responses where the anise smell is explicitly given as reason for positive preference on the grounds of prior experience: "Anise smells nice and usually nice smelling things aren't poisonous", "personal experience". Similarly, for the Quality dataset: "Cities in Switzerland tend to be nice".

[2]For Mushroom dataset there is a negative correlation due to the artefact in discovered rules discussed in Section 5.14.

- *Bias*: None
- *Strength*: High. Literal relevance has the strongest correlation with rule preference from all the available variables.

**5: Perceived positive contribution decreases plausibility if strength below threshold.**

- *Evidence (statistical analysis)*: Negative correlation between *minimum* literal relevance and plausibility on all LOD datasets.
- *Evidence (textual answers)*: The literals were automatically added to the rule by the rule learner on the basis of data and all were thus generally positive evidence in favour of the predicted class. Despite this, there were responses associated with preference for the rule not containing the weak evidence, such as: "logic be damned, I know of no critically acclaimed LGBT films at all", "The addition of western Asian countries weakens the plausibility".
- *Bias*: Weak evidence effect
- *Strength*: High

**6: Perceived negative contribution decreases plausibility.**

- *Evidence (statistical analysis)*: not performed
- *Evidence (text)*: We have not found any actual answers that use negative evidence such as "this rule is not preferred because literal xyz actually gives contrary evidence". However, the absence of such answers has a limited importance, because we have not designed our experiments to test for the effect of negative evidence.
- *Biases*: negativity bias, confirmation bias.
- *Strength*: Weaker than positive evidence. While based on the application of the *negativity bias* the effect of negative evidence on plausibility should be stronger than the effect of positive evidence, there is also the opposite effect of the *confirmation bias*, according to which people tend to look for evidence supporting the hypothesis at hand, disregarding conflicting evidence. Since in the textual responses we have not observed cases where a decision would be made in favour of the alternate hypothesis based on negative evidence, we hypothesize that the confirmation bias is stronger than the negativity bias. This proposition should be taken with caution, because as noted at other places we have not generally designed our experiments to include negative evidence. On the contrary, the rule learners in general included the literal in the rule only if it provided evidence for the class predicted by the rule.

**7: Unknown relation between attribute value and target class increases plausibility.**

- *Evidence (statistical analysis)*: The results are spurious due to artefacts in the Mushroom dataset.
- *Evidence (text)*: Indicative analysis of the textual answers for the Mushroom dataset reveals that the majority of answers refers to general properties of rules (number of literals) rather than to specific values. These are responses such as "more evidence in rule 1", "more describing factors make identification easier". At the same time, the Mushroom dataset had the strongest correlation between rule length and plausibility.
- *Strength:* ?
- *Biases:* Representativeness heuristic

**8: Relevant attribute with missing value increases plausibility.**

- *Evidence (statistical analysis).* Result of Linda V3 experiment: addition of a new condition with unknown value "Jenny is a bank teller and it is not known if she is active in feminist movement" increased preference over baseline option "Jenny is a bank teller" (cf. Table 5.17).
- *Evidence (text)*: The reasons given by subjects selecting the option with the unknown value included the following: "It might be assumed that Jenny is active in the feminist movement based on the information given about her. However, it is not known for sure that she is, so I chose the 3rd option.", "Jenny probably has a good job, and it's reasonable that she may work in a bank. She takes part in demonstrations and is concerned about discrimination issues, so she may be involved in a feminist movement, but there is no information that would tell us either way, whether she was or wasn't.".

  It follows from the following response that the absence of condition was interpreted as negation: "Out of the three this is the most likely she might well be active in a feminist movement given her background. She could be a bank cashier and if she is she may well not want it to be known that she is active in a feminist movement. To me this makes the third option more likely the reality is that she is likely to be active in a feminist movement or at least an activist in another organisation. she is unlikely to have left her past completely behind."
- *Strength*: ?
- *Biases*: Representativeness heuristic

**9: Unknown/unrecognized/implausible value decreases preference.**

- *Evidence (statistical analysis)*: Not performed.

- *Evidence (text)*: On the Mushroom dataset there are several pairs where the other rule is preferred giving reason that the subject does not know the value "anise" or "never heard of a mushroom smelling like anise." Similarly, there were similar responses such as "Could a mushroom smell of creosote?", "It's hard to tell what creosote smells like.", "Rule 1 doesn't sound likely because I do not recall seeing green mushrooms." These arguments were given as reasons for selecting the alternative rule not containing unknown value as more plausible.
- *Strength:* ?
- *Biases:* Ambiguity aversion.

**10: Irrelevant attributes/values have positive effect on preference.** By irrelevant attribute we mean an attribute the value of which is considered as not predictive of the target. If an attribute is considered as irrelevant, then any literals (attribute-value pairs) derived from the attribute are also irrelevant. Also, this situation covers attributes which are considered as relevant, but their particular value is not considered as relevant.

- *Evidence (statistical analysis)*: The Traffic dataset had lowest literal relevance (cf. Table 5.5). We expected the weak evidence effect to manifest via negative correlation between rule length and preference on this dataset, but instead we obtained near zero, but positive, correlation for Traffic V2 ($\tau = 0.05$ in Table 5.13).
- *Evidence (text)*: Textual answers for a selected rule pair are in Table 6.1.
- *Strength:* Weak.
- *Biases:* Information bias, Unit bias.

| count | preference | reason |
|---|---|---|
| 2 | No preference | "All European Countries are members of NATO", "most European countries are members of NATO anyway" |
| 2 | R 2 weak | "Rule 2 includes more people", "more to select from" |
| 1 | R 1 weak | "Rule 1 seems more plausible because it only covers one group" |

Table 6.1.: Analysis of responses for rule pair selected from the Traffic dataset: R1: NATO Member → low risk of accidents, R2: NATO Member AND EU Member → low risk of accidents.

### 6.1.3. Aggregation of Literal Contributions

Tree depicted in Figure 6.1B models the decision process that applies to aggregation of plausibility values obtained for literals. This process largely depends on generic strategies that the analyst applies when evaluating plausibility and these in turn depend on the ability of the analyst to combine evidence rationally, suppressing the various cognitive biases.

**11: Number of literals?** If rule has multiple literals, the preference values for individual literals need to be combined. The second branch applies to rules with 0 or 1 literals.

**12: Analyst's understanding of "and"?** The way one interprets rules with two or more attributes depends on the interpretation of the "and" connective joining the literals.

**13: Analyst prefers specificity or size of the group?** It follows from the textual responses that subjects had two fundamental types of preferences. In the absence of other deciding factors, one group of subjects preferred longer rules on the grounds that these are considered as more "descriptive", "specific", "less broad" or having "more data". The second group of subjects preferred shorter rules, giving reasons such as "less categories", "less options".

**14: Preference for specificity increases plausibility.** If the analyst prefers specificity and at the same time is not liable to misunderstanding of "and", she will tend to prefer the longer rules, since more conditions make them more specific. This personal attitude could demonstrate via preference for situation depicted in Figure 6.2B over situation depicted in Figure 6.2A.

- *Evidence (statistical analysis)*: not performed
- *Evidence (text)*: Reasons given for higher plausibility assigned to a longer rule included: "There's nothing that stands out as an "obvious" indicator of toxicity, so I've gone for a weak preference for Rule 2 as it's describing a smaller number of species than Rule 1 and thus likely to be the more accurate of the two.", "more describing factors make identification easier", "more indicators means more certainty", "You need to have more information so should be more accurate.". "Rule 1 has a much tighter definition of what would constitute a poisonous mushroom with 5 conditions as compared to rule 2 which only contains just 1 condition for the same result so rule 1 is a much higher plausibility of being believable".

  Some responses also exhibited misunderstanding of the task: "it's a narrower group so it has fewer drivers and thus fewer accidents".
- *Strength*: Medium. On the Mushroom dataset, there is a number of responses justifying the preference for specificity, while on other datasets this ratio is lower. It seems that the activation of this aggregation is conditioned by the absence of a literal with known relation to the target that can serve as a deciding factor.
- *Biases*: Representativeness, information bias, disjunction fallacy, insensitivity to sample size.

**15: Preference for size of supporting group.** Some subjects are aware of the fact that small group size implies statistical unreliability of the rule. Indicative analysis of the frequency of responses in both groups of subjects in our experiment suggests that the concern

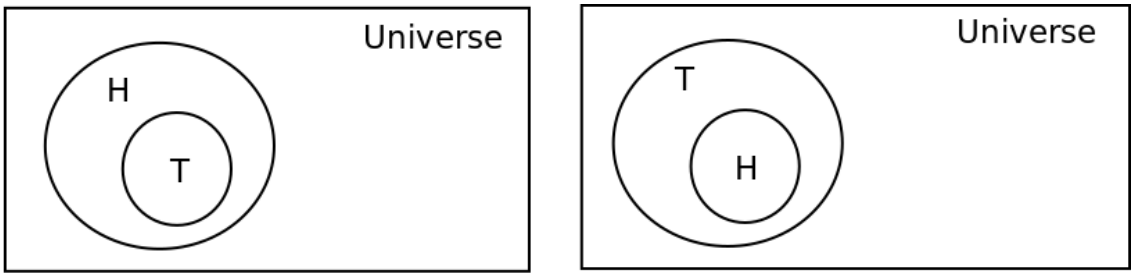



Figure 6.2.: Illustration of the relation between rule scope and hypothesis space, A) Hypothesized rule H embeds the correct rule T. B) Correct rule embeds the hypothesized rule. Adapted from Klayman and Ha [1987].

for the size of the supporting group is less prevalent than preference for specificity. This personal attitude could demonstrate via preference for situation depicted in Figure 6.2A over situation depicted in Figure 6.2B.

- *Evidence (statistical analysis)*: Not performed.
- *Evidence (text)*: Responses such as "Less describing factors means more likely" or "Rule 2 is clear and unambiguous. Rule 1 is too detailed to apply in practice.", "Rule 2 demands more specificity - both a capital city AND a port city, so weak preference to rule 1.", "Rule 1 would seem far broader, and thus more plausible. To expect a European country to be simultaneously a member of the Mediterranean Union is pretty specific."
- *Strength*: Medium. Results on Version 3 instructions (Movies dataset) show that explicitly conveyed size of group (rule support) has no effect on plausibility if communicated along with rule confidence.

**16: Analyst understands "and" as disjunction.** Majority of analysts who misunderstand "and" prefer longer rules. Analysis of responses show that the there are two fundamental paths from misunderstood "and" to the preference of longer rules: i) if the analyst prefers specificity, the longer rule is more specific and thus more preferred, ii) if the analyst prefers size of supporting group, the longer rule has more supporting examples if "and" is interpreted as "or", thus making the rule more preferred.

- *Evidence (statistical analysis)*: Inclusion of intersection test questions in V1 instructions reduced preference for longer rules on all datasets. This indicates that some subjects could understand "and" as disjunction.
- *Evidence (text)*: Example responses: "Rule one contains twice as many properties as rule 2 does for determining the edibility of a mushroom so that makes it statistically twice as plausible, hence much higher probability of being believable", "An extra group increases the likelihood."

- *Strength*: ?
- *Biases*: Misunderstanding of "and" (if it can be considered as a bias)

**17: Penalty for "oversimplicity".** If the antecedent contains only one or no literal, this can make the rule appear untrustworthy, since it will be viewed as "too simple". Remarkably, we have not observed any reason in data that would disqualify a rule based on its excess complexity. Oversimplicity was generally mentioned only for rules of length 1 according to our analysis of textual responses. We assume oversimplicity would also apply to rules with empty antecedent[3], although we have not tested this empirically.

- *Evidence (statistical analysis)*: Reverse than expected values for correlation between plausibility and literal relevance on the Mushroom dataset. This dataset contained a number of rule pairs in which the antecedent of the shorter rule with length 1 contained a strong predictor.
- *Evidence (text)*: For example, for Movies dataset there were responses such as: "You cannot blanket an entire genre as bad", "All English language movies can't be bad.", "You cannot blanket an entire decade of movies as bad".

  For Mushroom dataset, consider responses to rule pair no. 10 in Table 5.15, where the short Rule 1 can be summarized as "if odour is creosote then poisonous" and the alternative long Rule 2 as "if veil colour is white and gill spacing is close and does not have bruises and has one ring and population is several and cap surface is smooth then poisonous". Most subjects have considered Rule 2 as more plausible, giving reasons such as: "much more specific than just odor".
- *Strength*: ?
- *Biases*: "Oversimplicity bias" was not encountered in our review of cognitive science literature, but it was earlier reported in scope of machine learning research that "extreme simplicity is not acceptable for users" [Freitas, 2014]. A decision tree containing only one node is reported to be unacceptable for medical doctors by Elomaa [1994].

## 6.2. Practical Recommendations for Design of Machine Learning Software

This section provides a concise list of recommendations that is aimed to help machine learning practitioners to suppress effect of cognitive biases on comprehension of rule-based models. We expect part of the recommendations to be useful also for other symbolic machine learning models, such as decision trees. The list of recommendations follows.

[3]The rule with empty antecedent always corresponds to the case depicted in Figure 6.2A.

1. *Remove near-redundant rules and near-redundant literals from rules.* Rule models often incorporate output which is considered as marginally relevant. This can take form of (near) redundant rules or (near) redundant literals in the rule. Our analysis shows that these redundancies can induce a number of biases. For example, a frequently occurring but otherwise not very important literal can – by the virtue of the mere exposure effect – be perceived as more important than would be appropriate given the data.

2. *Represent rule quality measures as frequencies not ratios.* Currently, rule interest measures such as confidence and support are typically represented as ratios. Extensive research has shown that natural frequencies are better understood.

3. *Make "and" conjunction unambiguous.* There are several cognitive studies indicating "and" is often misunderstood. The results of our experiments also support this conclusion. Machine learning software should thus make sure that the meaning of *and* in presented rules is communicated unambiguously.

4. *Present confidence interval for rule confidence.* The tendency of humans to ignore base-rates and sample sizes (which closely relate to rule support) is a well-established fact in cognitive science, results of our experiments on inductively learned rules also provide evidence for this conclusion. Our proposition is that this effect can be addressed by computing confidence (reliability) interval for confidence. In this way, the "weight of evidence" will effectively be communicated through confidence.

5. *Avoid the use of negated literals as well as positive/negative class labels.* It is an established fact in cognitive science that negative information receives more attention and is associated with higher weight than positive information. There is research indicating that recasting a yes/no attribute to two "neutral" categories (such as "DAX/MED") can improve human understanding.

6. *Sort rules as well as literals in the rules from strongest to weakest.* People have the tendency to put higher emphasis to information they are exposed to first. By presenting the important information as first, machine learning software can also conform to these human *conversational maxims.* The output could also visually delimit literals in the rules based on their significance, which would again correspond to humans using various non-verbal clues to convey significance in the spoken word.

7. *Provide explanation for literals in rules.* There is a number of cognitive phenomena that result from the lack of domain knowledge on literals in the rules. Some examples include ambiguity aversion or unit bias. Providing the analyst with easily accessible information on literals in the rules including their predictive power can prove as an effective debiasing technique.

8. *Explain difference between negation and absence of a condition.* Prior results in cognitive science as well as our experimental results show that absence of a condition

can be misinterpreted as negation if the omitted condition is present in in other rules. Consider Rule 1: *IF bankteller=yes THEN class=A* and Rule 2: *IF bankteller=yes AND feminist=yes then class=B*. In presence of Rule 2, Rule 1 can be read as *IF bankteller=yes AND feminist=no THEN class=A*.

9. *Elicit and respect monotonicity constraints.* Research has shown that if monotonicity constraints – such as that fuel consumption increases with increasing car weight – are observed, the plausibility of the rule model increases.

10. *Educate and assess human analysts.* One perhaps surprising result related to confirmation (myside) bias is that its incidence is not related to intelligence. Some research even suggests that analysts, who think that good arguments are those that can be "proved by facts", are even more susceptible to myside bias than the general population. There is a psychological test that can reveal the susceptibility of a person to myside bias. Several studies have shown that providing explicit guidance and education on formal logics, hypothesis testing and critical assessment of information can reduce fallacy rates in some tasks.

These proposals are based on the results of our qualitative analysis summarized in Table 4.1 on page 63, on our empirical results presented in Chapter 5, as well as on the textual answers analysed within this chapter.

# Part II.

# Software Framework

## Less Bias through Smaller Number of Shorter Rules

# 7. Related Algorithms

This chapter sets our work in the context of two groups of related machine learning algorithms. The first group involves association rule-based models, which are basis of our final approach, and the second group utility-based methods, that we initially worked on.

**Chapter organization.** Section 7.1 introduces association rule learning and the apriori algorithm. A brief overview of association rule-based classification, the basis of our framework, is presented in Section 7.2. Utility models, which served as the initial inspiration for the proposed algorithmic framework, are covered in Section 7.3.

## 7.1. Apriori and Association Rule Mining

Apriori algorithm [Agrawal et al., 1993] is an association rule learning approach well known for its scalability. For example, it has been reported to be successfully used to mine massive commercial databases with thousands of features and tens of million rows of data [Hastie et al., 2001]. The Apriori algorithm has two stages. In the first stage, all one itemsets with minimum coverage (support) are generated. Using this intermediate result two-itemsets are generated etc. The generated one-, two-, etc. itemsets, also called frequent itemsets, are a form of a frequent pattern [Toivonen, 2010]. For example, a frequent itemset of length two can be "veil color=white, odor=foul". If this itemset has support 2% it means that two percent of all instances (rows) in the input dataset have both white veil color and foul odor. In the second stage, rules are created from these itemsets. Only rules satisfying the minimum confidence threshold set by the user are generated. A concise description of the algorithm is presented in Fürnkranz and Kliegr [2015].

#### 7.1.0.1. Apriori Successors

A number of alternative algorithms, such as Eclat [Zaki and Gouda, 2003] or FP-Growth [Han et al., 2004], have been proposed to speed up frequent itemset discovery originally proposed within the Apriori algorithm. Mining for *closed* frequent itemsets proposed by Pasquier et al. [1999] is another optimization. A frequent itemset $P$ is closed if $P$ is included in no other itemset that has the same support as $P$.

In recent years there was a growing interest in approaches that support parallel execution of frequent itemset mining in order to harness modern multi-core architectures. PLCM [Negrevergne et al., 2010] and MT-Closed [Lucchese et al., 2007] are parallel implementations of two fastest algorithms LCMv2 [Uno et al., 2004] and DCI Closed [Lucchese, 2004]

according to the FIMI'04 workshop[1], which provided a benchmark of submitted frequent itemset mining implementations. The recently proposed ParaMiner [Negrevergne et al., 2014] algorithm yields comparable execution times to PLCM and MT-Closed, while it allows to mine not only for closed frequent itemsets, but also for additional types of patterns such as connected relational graphs and gradual itemsets.

For surveys of frequent set mining and association rule discovery we refer the reader to [Aggarwal and Han, 2014, Fürnkranz et al., 2012].

## 7.2. Association Rule-based Classifiers

Separate-and-conquer strategy described is common approach used in rule learning, which can be, according to Fürnkranz [1999], traced back to the AQ algorithm [Michalski, 1969]. This strategy provides a basis for the seminal RIPPER algorithm [Cohen, 1995] as well as for the state-of-the-art FURIA algorithm [Hühn and Hüllermeier, 2009]. Association rule learning [Agrawal et al., 1993] is algorithmically different approach, which was originally designed to discover interesting patterns in very large and sparse instance spaces. Association rule learning yields a set of rules that correspond to high density regions in the data. Unlike in most separate-and-conquer approaches, cardinal features need to be discretized prior to the execution of association rule learning. The resulting rules correspond to hypercubes in the transformed space. This transformation impairs precision of the rules, but greatly reduces the combinatorial complexity thus allowing the algorithm to process even large datasets.

Association rule learning was several years after its conception adopted also for classification. The first Association Rule Classification (ARC) algorithm dubbed CBA (Classification based on Associations) was introduced by Liu et al. [1998]. While there were multiple follow-up algorithms providing marginal or small improvements in classification performance (e.g. CPAR [Yin and Han, 2003], CMAR [Li et al., 2001b] and MMAC [Thabtah et al., 2006]), the structure of most ARC algorithms follows that of CBA [Vanhoof and Depaire, 2010]:

1. learn classification association rules,
2. prune rules,
3. classify new objects.

In the following, this process will be discussed in detail for the CBA algorithm.

### 7.2.1. Classification-based on Associations (CBA)

One-rule classification and crisp rules make CBA classification models possibly most comprehensible among all association rule classification algorithms. In this section, we will describe the building of the CBA classifier from the conceptual perspective. For algorithmic details, please refer to Liu et al. [1998].

[1] `http://fimi.ua.ac.be/fimi04/`

#### 7.2.1.1. Learning Class Association Rules

In the first step of CBA a modified version of *Apriori* [Agrawal et al., 1993] is used to learn conjunctive classification rules from data. The mining setup is constrained so that only the target class values can occur in the consequent of the rules. The output of association rule learning algorithms is determined by two parameters: minimum confidence and support thresholds on the training data. Let us briefly remind the definition of these two metrics. The confidence of a rule is defined as $a/(a+b)$, where $a$ is the number of correctly classified objects, i.e. those matching rule antecedent as well as rule consequent, and $b$ is the number of misclassified objects, i.e. those matching the antecedent, but not the consequent. The support of a rule is defined as $a/n$, where $n$ is the number of all objects (relative support), or simply as $a$ (absolute support).

The main obstacles for straightforward use of the discovered association rules as a classifier are the excessive number of rules discovered even on small datasets, and the fact that contradicting rules are generated. ARC algorithms provide several extensions over association rule learning algorithms which address exactly these issues. These algorithms contain a rule pruning step, which significantly reduces the number of rules, and define a conflict resolution strategy for cases when one object is matched by multiple rules.

#### 7.2.1.2. Pruning Candidate Rules

CBA adopts *data coverage pruning* [Vanhoof and Depaire, 2010]. This type of pruning processes the rules in the order of their strength, removing transactions (instances, objects) that the rule matches from the database. If rule does not correctly cover at least one instance, it is deleted (pruned). Additionally, CBA incorporates optional *pessimistic pruning step*, which is used in the first CBA phase when candidate association rules are generated. In CBA, data coverage pruning is combined with "default rule pruning": the algorithm replaces all rules below the current rule with default rule if a default rule inserted at that place would reduce the number of errors. Default rule is a rule with empty antecedent, which ensures that a query instance is always classified even if it is not matched by any other rule in the classifier.

The effect of pruning on the size of the rule model is reported by Liu et al. [1998], who present evaluation on 26 UCI datasets. To illustrate the effect of pruning using data coverage pruning in the CBA algorithm, the average number of rules per dataset without pruning was 35,140, with pruning the average number of rules was reduced to 69 without effectively impacting accuracy. Pessimistic pruning was found to have no effect on classifier accuracy (on average across multiple datasets).

#### 7.2.1.3. Classification

The original CBA algorithm performs "one rule" classification. First, rules are sorted according to their *strength*, which is determined based on the following criteria:

- confidence,

- support,
- rule length (shorter rule is placed higher).

Instance is assigned to the class in the consequent of the first rule with antecedent matching the instance in the ordered list of rules. The advantage of one rule classification is that it is easily understandable, which provides advantages in some applications, such as business rule learning [Kliegr et al., 2014].

### 7.2.2. Comparison of CBA and its Successors

The main benefit of using a rule-based classifier, as opposed to state-of-the-art sub-symbolic method such as a deep neural network, should ideally be i) comprehensibility of the rule-based model, ii) fast execution on large and sparse datasets, iii) accuracy comparable to state-of-the-art "black-box" classification models.

Individual ARC algorithms meet these aspirations to a different degree. Table 7.1 presents a comparison between ten most well-known ARC algorithms (and closely related approaches) in terms of key comprehensibility metrics, accuracy and performance.

We selected CBA as a basis for our framework, since as follows from Table 7.1 it produces more comprehensible models than any of its successors while maintaining high accuracy and fast execution times. The difference between accuracy of CBA model and accuracy of state-of-the-art ARC algorithms such as FARC-HD is very small. In terms of accuracy, CBA is outperformed only by FARC-HD (by 4%) and CPAR (by 2%). However, CPAR has 4x times more rules on the output and less comprehensible multi-rule classification. FARC-HD outperforms CBA in terms of accuracy, and even more so in its evolved version FARC-HD-OVO [Elkano et al., 2015]. On the other hand, this algorithm is more than 100x slower than CBA and produces less comprehensible fuzzy rules.

In addition to criteria in Table 7.1, CBA has also the advantage that it uses standard (constrained) association rule learning in the first step. This makes work on postprocessing CBA output "future-proof", since the performance of CBA can be improved by replacing Apriori with another association rule learning algorithm. For example, FP-Growth is reported to be faster on most problems than the Apriori algorithm [Goethals and Zaki, 2003].

## 7.3. Utility-based Algorithms

As a representative of utility learning, in this section we will cover the UTA (UTilités Additives) method [Jacquet-Lagreze and Siskos, 1982]. UTA is a time-tested method, which is widely used as a basis for many recent utility-based preference learning algorithms (e.g. Greco et al. [2007]).

UTA learns an additive piece-wise linear utility model. UTA takes a set of alternatives ordered according to user's preferences, and it learns utility functions for each attribute. Using these functions, the utility for individual attribute values are combined into the overall utility for a given object.

| algorithm | year | single | crisp | det | assoc | acc | rules | time |
|---|---|---|---|---|---|---|---|---|
| CBA Liu et al. [1998] | 1998 | yes | yes | yes | yes | .80 | 185 | 35s |
| CBA 2 Liu et al. [2001] | 2001 | yes | yes | yes | yes | .79 | 184 | 2 m |
| 2SLAVE González and Pérez [2001] | 2001 | no? | no | no | no | .77 | 16 | 22m |
| CMAR Li et al. [2001a] | 2001 | no | yes | yes | yes | .79 | 1419 | 6m |
| CPAR Yin and Han [2003] | 2003 | no | yes | yes | yes | .82 | 788 | 11s |
| LAFAR Hu et al. [2003] | 2003 | no | no | no | yes | .75* | 47* | 5h* |
| FH-GBML Ishibuchi et al. [2005] | 2005 | no | no | no | no | .77 | 11 | 3h |
| CFAR Chen and Chen [2008] | 2008 | yes | no | yes | yes | .71* | 47* | 17m* |
| SGERD Mansoori et al. [2008] | 2008 | no? | no | no | no | .74 | 7 | 3s |
| FARC-HD Alcala-Fdez et al. [2011] | 2011 | no? | no | no | yes | .84 | 39 | 1h 20m |

Table 7.1.: Comparison between CBA and other association rule (or closely related) classifiers. *single* refers to single rule classification, *crisp* to whether the rules comprising the classifier are crisp (as opposed to fuzzy), *det.* to whether the algorithm is deterministic with no random element such as genetic optimization, *assoc* corresponds to whether the method is based on association rules, *acc*, *rules* and *time* is average accuracy, average rule count and average run time across 26 datasets as reported by Alcala-Fdez et al. [2011]. * indicates that the algorithm did not process all datasets

UTA aims at inferring one or more additive value functions from a given ranking (weak ordering) on a reference set of objects. Each object is described by $N$ attributes. The method uses linear programming to find such $N$ partial value functions $u_i$ that best explain given preferences. The overall preference rating for an object $\mathbf{o}$ is computed as an average of utility values for all attributes: $u(\mathbf{o}) = \sum_{i=1}^{N} u_i(o_i)$, where $u_i$ are non-decreasing value functions and $o_i$ are attribute values.

The method expects that the input attributes are monotone with respect to preferences, which not only requires manual input for each attribute, but also limits the applicability of the method. The utility function for each attribute is either marked as *cost* (utility does not increase with rising attribute value) or *gain* (utility does not decrease with rising value). Since UTA method was originally developed as an algorithm for Multi-Criteria Decision Making (MCDM), manually specifying whether an attribute is cost or gain was not an issue. Also, in the MCDM setting it is possible to ensure that there are no ambivalent attributes, which cannot be unanimously classified to the cost or gain category for all decision makers. If UTA is to be applied in wider machine learning context, such manual approach to enforcing monotonicity is not feasible.

In previous work [Kliegr, 2009] we proposed a non-monotonic extension of UTA (UTA-NM), which allows $u_i$ to change direction from ascending to descending. Every change of the direction from gain to cost or vice versa within one partial utility function is penalized to ensure that the resulting model is not overly complex and over-fitted to the data. An illustration of the principle applied in non-monotone UTA is depicted at the following example.

**Example. (Worker comfort)** Consider the following preference learning problem: determine worker utility (comfort) on 4 points scale of a worker based on temperature and humidity of the environment. Figure 7.1A depicts the utility curve for the value of the temperature attribute if humidity is 50%. At humidity 100%, the worker's utility function will be different (Figure 7.1B). However, it is unlikely that the worker's utility function will look like depicted on Figure 7.1C at any humidity level: too many changes of shape of the utility function are thus penalized in UTA-NM,

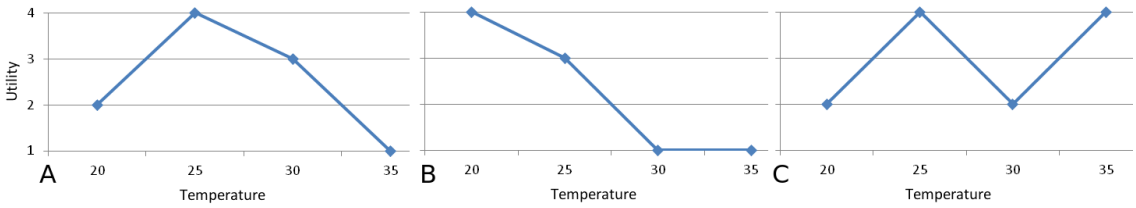



Figure 7.1.: Class label values associated with the Temperature attribute. A] humidity 50%, B] humidity 100%, C] improbable shape at any humidity level

Another option for dealing with non-monotonicity, based on a transformation of arbitrary input attributes into a monotone space was presented in Eckhardt and Kliegr [2012].

Contrary to the initial plan, we decided not to base our framework on the extension of the UTA method. The reason is that UTA and its derivatives impose too strong inductive bias – the individual partial value functions are not only monotonic, piece-wise linear, but also unconditionally additive: the total utility from an alternative is given by sum of partial utilities. In UTA, the utility function relating to the temperature attribute is independent of the value of the humidity attribute.

Also, learning an UTA model can be slow on large data due to the fact that the model is a solution to a linear programming optimization problem.[2] When working on UTA-NM extension we also observed that the time complexity is very sensitive to the inclusion of additional constraints to the optimization problem.

[2]Recently, Ghaderi et al. [2017] proposed modification to our UTA-NM algorithm that result in improvement of execution times.

# 8. Problem Statement

The results presented in the previous part of the thesis show that multiple cognitive biases affect understanding of rules discovered from data. While these have different causes and effects, there is a clear pattern of "less is more" – the more concise the model presented to the human user is, the less opportunities there will be for cognitive biases to be triggered.

In this part, we will focus on making a selected rule learning algorithm more comprehensible by implementing the "less is more" debiasing strategy.

### Formal Problem Definition

Let $\mathcal{C}$ be a classifier composed of association rules. Let the classifier adopt the *one rule* classification strategy: each training instance is covered by the rule with the highest precedence among the rules that can cover the instance. The quality of the classifier is evaluated using two metrics: $accuracy(\mathcal{C})$, as measured by the number of correctly classified test instances divided by the number of all test instances, and the size of the classifier $size(\mathcal{C})$ as measured by the count of conditions in the antecedents of all rules in $\mathcal{C}$.

The goal of this part of the thesis is to present a new postprocessing algorithm, which takes on its input classifier $\mathcal{C}$ and outputs a processed classifier $\widehat{\mathcal{C}}$, so that the following holds:

- $\mathcal{C}$ adopts *one rule* classification strategy
- $accuracy(\widehat{\mathcal{C}}) \geq accuracy(\mathcal{C})$
- $size(\widehat{\mathcal{C}}) \leq size(\mathcal{C})$

According to the Law of Conservation for Generalization Performance (LCG) [Schaffer, 1994] it holds that when taken across all learning tasks, the generalization performance of any learner sums to 0. In light of this theorem, it is impossible to construct a learning algorithm that has better accuracy than some other learning algorithm on all possible problems. However, as discussed by Giraud-Carrier and Provost [2005], LCG does not reflect the higher probability with which certain learning tasks (functions to learn) occur in the real world than other tasks. Therefore, we will focus on evaluating performance of the proposed algorithm on a diverse set of real world datasets.

# 9. Monotonicity-exploiting Association Rule Classification

Monotonicity-exploiting Association Rule Classification (MARC) is a postprocessing algorithm for association rule classification algorithm CBA [Liu et al., 1998]. It uses original, undiscretized numerical attributes to optimize the discovered association rules, refining the boundaries of literals in the antecedent of the rules produced by CBA. Some rules as well as literals from the rules can consequently be removed, which makes the resulting classifier smaller. One-rule classification and crisp rules make CBA classification models possibly most comprehensible among all association rule classification algorithms. These viable properties are retained by MARC (QCBA).[1] The postprocessing is conceptually fast, because it is performed on a relatively small number of rules that passed data coverage pruning in CBA.

**Chapter organization.** An overview of the framework is provided by Section 9.1. Section 9.2 introduces key concepts and notation related to association rule learning, which are referenced from the descriptions of the algorithms in the subsequent sections. The first step of the framework – building of the classifier – is covered by Section 9.3. The proposed framework consists of a succession of the following optimization steps that postprocess the CBA model: i) refit, ii) literal pruning, iii) trimming, iv) extension, v) postpruning and vi) default rule overlap pruning. These are described in this order in Sections 9.4 – 9.9.

## 9.1. Overview and Motivation

Current rule learning approaches can be divided into two categories depending on how they learn rules and process numerical attributes: inductive rule learning, typically based on a variation of separate-and-conquer approach natively supporting numerical attributes, and association rule-based classification approaches, which work only on nominal data. Largely owing to this restriction, association rule-based algorithms can be very fast on datasets with many instances and high dimensions.

The proposed framework is based on Association Rule-based Classification (ARC). Current mainstream ARC approaches can be applied on data with numerical attributes, but only if these are discretized prior to mining. The disconnection between discretization and model building is a source of inefficiencies in the resulting classifier – rule boundaries

[1] The software package with implementation of MARC used for evaluations in Chapter 10 is distributed under the name "Quantitative CBA", which we consider as more comprehensible to the target machine learning audience.

are not fit to the original continuous data, resulting in loss of accuracy and redundancies. While recently several ARC approaches that support numerical data have been proposed, these produce fuzzy association rules, containing fuzzy itemsets that deteriorate comprehensibility of the model.

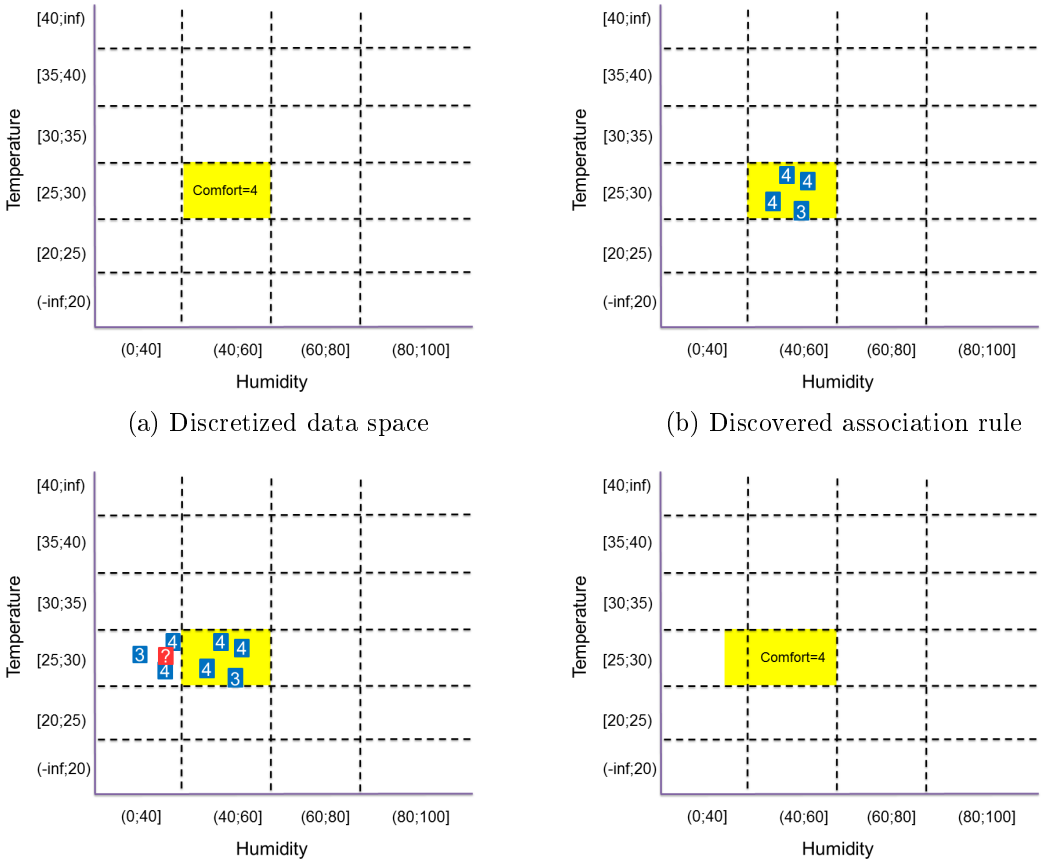
### 9.1.1. CBA as Base Learner

The framework is conceived as a postprocessing algorithm for ARC classification algorithm CBA [Liu et al., 1998], which reverts to the original attribute space to "edit" discovered association rules, refining the scope of literals (conditions) in the antecedent of the rules. As a consequence, the fit of the individual rules to data improves, rendering some of the rules and attributes redundant. These are removed, making the resulting classifier smaller. CBA models are rule lists with several properties that make them comprehensible, such as one-rule classification and crisp rules. Our preprocessing retains these favourable properties. The postprocessing is conceptually fast, because it is performed on a relatively small number of rules that passed the data coverage pruning in CBA.

### 9.1.2. Inspiration by Monotonicity Constraint

A key optimization step in our algorithm is inspired by the monotonicity constraint, which was already mentioned within our review of prior research on comprehensibility in machine learning in Section 2.4.3. In many classification and regression problems it is often the case that a domain of a predictor attribute has a – ceteris paribus – monotone relationship with the target class label. An example of such monotonicity constraint is that within a certain range subjective comfort level will be increasing with room temperature, all other variables being equal.

Unlike our initial approach [Kliegr, 2009], which was based on the UTA method (cf. Section 7.3), we do not aim to impose monotonicity as a hard or even soft constraint. This would result in multi-objective optimization: a drop in standard rule quality metrics such as confidence will be accepted as long as monotonicity is ensured or at least improved. Instead we readjust association rule output to reflect monotonicity without adversely affecting these metrics: association rule learning and classification operates on prediscretized data, which results in a learned rule often covering a narrower region than it could. We apply the monotonicity constraint when readjusting the rules to better fit the raw data, detaching them from the multidimensional grid, which is the result of the discretization. The intuition behind our approach is briefly exemplified in two dimensions in the following.




(a) Discretized data space

(b) Discovered association rule

(c) Query instance, data for extension

(d) Rule after monotonic extension

Figure 9.1.: Illustration of monotonic literal extension

**Example. (Monotonic rule extension)**
Figure 9.1a describes a discretized instance space corresponding to two originally continuous attributes (Temperature and Humidity) and a class label (Comfort, integer from 1 to 4). After association rule learning with minimum confidence threshold of 75% and minimum support of 3 instances, one of the discovered rules is depicted as the yellow region in Figure 9.1a:
*IF Humidity = [40,60) AND Temp = [25;30) THEN Comfort = 4, conf = 75%, supp = 3.*
As Figure 9.1b shows there are four instances matched by the antecedent of the rule, three of which indeed have class "4". The confidence of this rule is thus $3/4 = 75\%$ and the number of correctly covered instances is 3 (absolute support).
Figure 9.1c shows a query instance '?' with $Humidity = 38$ and $Temperature = 28$. This data point is not covered by our rule, however, applying the monotonicity constraint we can expect the class label to remain 4 or to decrease to 3 in case humidity of 38 feels to dry. A jump to comfort value of 2 or 1 would seem as unlikely.
Figure 9.1d shows a new rule exploiting this intuition about monotone relation of comfort and humidity. The rule was created by lowering the boundary on humidity to 38. The confidence of the rule improves to 5/6:
*IF Humidity = [38,60) AND Temp = [25;30) THEN Comfort = 4, conf = 83%, supp = 6.*

In the simple example above we demonstrated the way in which our framework exploits the non-decreasing/non-increasing parts of a monotone relationship (*piece-wise monotonicity assumption*) between a predictor and target to adjust boundaries of conditions in association rules. From this follows the working name "Monotonicity-exploiting Association Rule Classification" (MARC). It should be noted that MARC does not take advantage of ordinal structure of the class label even if it is available. On the other hand, the framework incorporates a number of other optimizations to reduce the size of the model.

### 9.1.3. Workflow

Classification workflow involving MARC consists of two main components. The first component is a CBA implementation. The core of the framework is the second component, which post-processes the discovered rules comprising the CBA classifier. This second component has two phases. In the first phase, individual rules are optimized. This increases coverage of individual rules and reduces their length by removing redundant attributes. In the second phase, three types of rule pruning are performed to reduce the number of rules. A conceptual description of the algorithm follows:

1. **Standard CBA classifier building**.
    a) **Discretization of numeric fields in the dataset**. Any discretization technique can be used. CBA is typically used with discretization based on the Minimum Description Length Principle (MDLP) [Fayyad and Irani, 1993], which selects number of cut-points with the highest entropy gains.

b) **Discovery of candidate association rules.** Class association rules can be discovered using any base association rule learner (Apriori, FP-Growth, etc.)

c) **Rules are sorted to create initial rule list**. The sort criteria are confidence, support and rule length.

d) **Data coverage pruning**. Any of the CBA variant (M1 or M2 introduced in Liu et al. [1998]) can be used.

2. **Tuning of individual rules.**

a) **Refitting rules to value grid.** Literals originally aligned to borders of the discretized regions are refit to finer grid with steps corresponding to all unique attribute values appearing in the training data.

b) **Literal pruning**: Removal of redundant literals (attribute-value pairs) from rules. Literal is considered redundant if its removal does not decrease rule confidence.

c) **Trimming.** Boundaries of literals in discovered rules are trimmed so that their borders do not contain regions covering no training instances.

d) **Extension.** Ranges of literals in the body of each rule are extended. The extension is accepted only if it does not deteriorate rule confidence.

3. **Pruning of the optimized rule list.** Rules are resorted and since the regions they cover could change, another iteration of pruning is performed to remove rules made newly redundant:

a) **Data coverage pruning.** Once the rules have been extended, they match more objects, which can make some of the rules redundant, therefore data coverage pruning is performed to remove some of the newly redundant rules.

b) **Default rule pruning.** All rules below the current rule are replaced by default rule if this reduces the number of errors on training data. Note that to preserve more rules for MARC to work with, this pruning is skipped in CBA and is performed at later point, after MARC optimized the rule list.

c) **Default rule overlap pruning.** Some rules that classify into the same class as the default rule in the end of the classifier can be removed.

Algorithm 1 depicts the succession of optimization steps in MARC and provides pointers to individual algorithms described in detail in the following subsections.

**Algorithm 1** MARC *marc()*

**Require:** *rules* – input rule list generated by CBA
**Ensure:** optimized *rules*

1: *rules* ← remove any rules with empty antecedent from *rules*. {CBA includes at least one such rule as default rule in the end of the list.}
2: **for** *rule* ∈ *rules* **do** {process rules in the CBA sort order (Fig. 9.8)}
3: *rule* ← *refit*(*rule*) {cf. Alg. 2}
4: *rule* ← *pruneLiterals*(*rule*){cf. Alg. 3}
5: *rule* ← *trim*(*rule*){cf. Alg. 4}
6: *rule* ← *extendRule*(*rule*) {cf. Alg. 5}
7: **end for**
8: *rules* ← *postprune*(*rules*) {cf. Alg. 8, postpruning adds a new default rule, if postpruning is disabled, MARC ensures that default rule is added at this point.}
9: *rules* ← *drop*(*rules*) {cf. Alg. 9 (transaction-based, Alg. 10 (range-based) version of default rule overlap pruning}
10: **return** *rules*

## 9.2. Preliminaries

In the following we define the main concepts relating to association rule learning. We also introduce several extensions to the commonly used notation in order to preserve the link between discretized and original data, which is required by the MARC framework.

**Definition 1.** ***(training dataset)*** *A training dataset* $\mathbf{T}$ *is a set of objects* $\{o\}$*, each object described by vector* $\langle o_1, \ldots, o_n, c\rangle \in \mathbf{A}_1 \times \ldots \times \mathbf{A}_n \times \mathbf{C}$*, where* $\mathbf{A}_i$ *is the domain of attribute* $A_i$ *and* $\mathbf{C}$ *the domain of the target attribute (class label)* $C$*.*

The training dataset is comprised of objects (instances) which are described by attributes. We distinguish between two types of attributes: nominal and cardinal. An ordinal attribute is not considered as a separate type, since it can be converted to a cardinal attribute.

**Definition 2.** ***(attribute in training dataset)*** *A domain of an attribute* $A$*, denoted as* $\mathbf{A}$ *is a set of all distinct values* $\mathbf{A} = \{v\}$ *of attribute* $A$ *in the training dataset* $T$*. Attribute can be either of a "nominal" or "cardinal" type.*

1. *If* $A$ *is a nominal attribute then* $\mathbf{A}$*, then every two different values* $v_1, v_2 \in \mathbf{A}$ *are incomparable.*

2. *If* $A$ *is a cardinal attribute then* $\mathbf{A}$*, then for every two different values* $v_1, v_2 \in \mathbf{A}$ *it either holds that* $v_1 > v_2$ *or* $v_2 > v_1$*.*

A typical association rule learning setup involves discretization of all cardinal attributes in the training dataset into bins. Nominal attributes with many distinct values can also be binned, if a distance function is known. The result of preprocessing is a modified training dataset $\overline{\mathbf{T}}$, which contains a smaller number of distinct values. Assuming that an attribute $A$ with domain $\mathbf{A} = \{v\}$ was discretized and the result is attribute $\bar{A}$ with domain $\bar{\mathbf{A}} = \{\bar{v}\}$, then each value $\bar{v} \in \bar{\mathbf{A}}$ can be mapped to one or more values $v \in \mathbf{A}$.

| $A_1$ | | $A'_1$ | $A_2$ | | $A'_2$ |
|---|---|---|---|---|---|
| $a_{1,1}, a_{1,2}$ | $\rightarrow$ | $a'_{1,1}$ | $a_{2,1}, a_{2,3}$ | $\rightarrow$ | $a'_{2,1}$ |
| $a_{1,3}, a_{1,4}$ | $\rightarrow$ | $a'_{1,2}$ | $a_{2,2}, a_{2,4}$ | $\rightarrow$ | $a'_{2,2}$ |
| $a_{1,5}, a_{1,6}, a_{1,7}$ | $\rightarrow$ | $a'_{1,3}$ | | | |

Table 9.1.: Example of the discretization process

**Definition 3.** ***(preprocessed training dataset)*** *Let $f_i : \mathbf{A_i} \rightarrow \bar{\mathbf{A}}_i$ be a preprocessing function for attribute $A_i$. Training set $\overline{\mathbf{T}}$ consists of objects $\bar{o} = \langle f_1(o_1), \ldots, f_n(o_n), c\rangle$, where $o = \langle o_1 \ldots o_n, c\rangle \in \mathbf{T}$.*

In the following we assume that the rules are learned on the preprocessed dataset $\overline{\mathbf{T}}$ in the attribute space $\bar{\mathbf{A}}_1 \times \ldots \times \bar{\mathbf{A}}_n \times C$. Each preprocessed value $\bar{v}$ appearing in the discovered rule, can be represented using one or more values from the original attribute space. In the following, we will thus refer to the original attribute space and training set $\mathbf{T}$ unless explicitly noted otherwise.

**Example. (rule in original and preprocessed space)** Consider the following rule $r_1$: $A'_1 = a'_{1,2} \wedge A'_2 = a'_{2,1} \rightarrow C = c_1$. The rule contains two literals defined over two attributes: $A'_1 = \{a'_{1,1}, a'_{1,2}, a'_{1,3}\}$ and $A'_2 = \{a'_{2,1}, a'_{2,2}\}$. Attribute $A'_1$ is created by performing discretization of cardinal attribute $A_1 = \langle a_{1,1}, a_{1,2}, a_{1,3}, a_{1,4}, a_{1,5}, a_{1,6}, a_{1,7}\rangle$ in the original dataset, which is depicted in Table 9.1. The attribute $A'_2$ is nominal, and the bins were created by user-defined value merging.
In a dataset dealing with preferences of second-hand car buyers, $A_1 = \{1 \ldots 7\}$ could correspond to the age of the car in years and $A_2 = \{yellow, brown, white, black\}$ to colour of the car, $A'_1 = \langle [1,3), [3,5), [5,7]\rangle$, $A'_2 = \{light, dark\}$. The class $c_1$ expresses the value "highly preferred".

MARC assumes that on the input it obtains association rules, which are composed of literals.

**Definition 4.** ***(literal)*** *A literal $l = (A, V)$ is an association of attribute $A$ with value range $V$.*

Algorithm *Apriori*, the de facto standard in association rule learning, outputs rules in which one literal corresponds to one value (or as originally called, with some simplification, an *item*). In our framework, we assume that a literal may be associated with a value range, which is a disjunction of multiple attribute values. The primary reason for extending the definition of literal, as present in Yin and Han [2003], is that multi-value literals are on the output of the MARC Extension procedure.

**Definition 5.** ***(value range of literal defined on nominal attribute)*** *Let $l = (A, V)$ be literal defined over a nominal attribute $A$. The value range $V$ is a set of $m \geq 1$ values: $V = \{v_i\}, v_i \in \mathbf{A}$, where $\mathbf{A}$ is the domain of attribute $A$ in the training set.*

For literals created over a cardinal attribute, the literal values are a subsequence of the domain of the attribute.

**Definition 6.** ***(value range of literal defined on cardinal attribute)*** *Let $l = (A, V)$ be literal defined over a cardinal attribute $A$. The value range $V$ is a sequence of $m \geq 1$ values: $\langle x_1, \ldots, x_i, \ldots, x_m \rangle, \forall_{i=1\ldots m} x_i \in \mathbf{A}$, where $\mathbf{A}$ is the domain of attribute $A$ in the training dataset $T$. The sequence $V$ has the property that among each two its consecutive elements $x_j$, $x_{j+1}$ there is no element $y \in \mathbf{A}$ for which it would hold that $x_j < y < x_{j+1}$.*

When a candidate rule is created during rule learning or when the rule is applied on test objects, it is necessary to verify which objects match individual literals in the rule.

**Definition 7.** ***(satisfaction of literal by object)*** *An object $o$ satisfies a literal $l = (A_i, V)$ if and only if the value of the object in attribute $A_i$, denoted as $o_i$, meets one of the following conditions:*

1. *$A_i$ is a nominal or cardinal attribute and $\exists v_i \in V : v_i = o_i$*
2. *$A_i$ is a cardinal attribute and $\exists v_u, v_l \in V : v_u \geq o_i \wedge v_l \leq o_i$*

Condition (2) is applicable only in the test phase, since as follows from Definition 6, the value range $\langle v_l, \ldots, v_u \rangle$ of a literal created over a cardinal attribute contains all values in the domain of attribute A in the training dataset within this range, hence the check with condition (1) is satisfactory.

**Definition 8.** ***(rule)*** *A rule $r$ takes the form $l_1 \wedge l_2, \wedge \ldots \wedge l_m \rightarrow c$. The body of the rule, denoted as body($r$), consists of conjunction of literals $l_1, l_2, \ldots, l_m$, $m \geq 0$. There are no two literals $l_i, l_j$ in body($r$) which are associated with the same attribute $A_k$. The consequent of the rule consists of literal c, which is the class label (denoted as class($r$)) of the rule. Rule is assigned confidence (conf($r$) $\in [0; 1]$) and relative support (denoted as supp($r$) $\in [0; 1]$).*

Our definition of rule is compatible with rules that are learned with the Apriori algorithm, on the condition that the head of the rule is constrained to contain only literals created from the target (class) attribute. The values of confidence and support for a rule are computed by the rule learning algorithm from the training data.

**Definition 9.** ***(satisfaction of rule body by object)*** *An object $o$ satisfies body($r$) if and only if it satisfies every literal in body($r$). If body($r$) contains zero literals, any object satisfies it. If an object satisfies body($r$), $r$ predicts that the object is of class($r$).*

**Definition 10.** ***(confidence and support of a rule)*** *Let $r$ be a rule $l_1 \ldots \wedge l_m \rightarrow c$, $\mathbf{T}$ a training dataset. Let $\mathbf{S}$ denote the set of all objects $x \in \mathbf{T}$ for which $x$ satisfies body($r$).*

*Confidence is computed as:*

$$\text{conf}(r) = \frac{|o \in \mathbf{S} : o \text{ has class label } c|}{|\mathbf{S}|}. \tag{9.1}$$

*Support as:*

$$supp(r) = \frac{|o \in \mathbf{S} : o \text{ has class label } c|}{|\mathbf{T}|}. \tag{9.2}$$

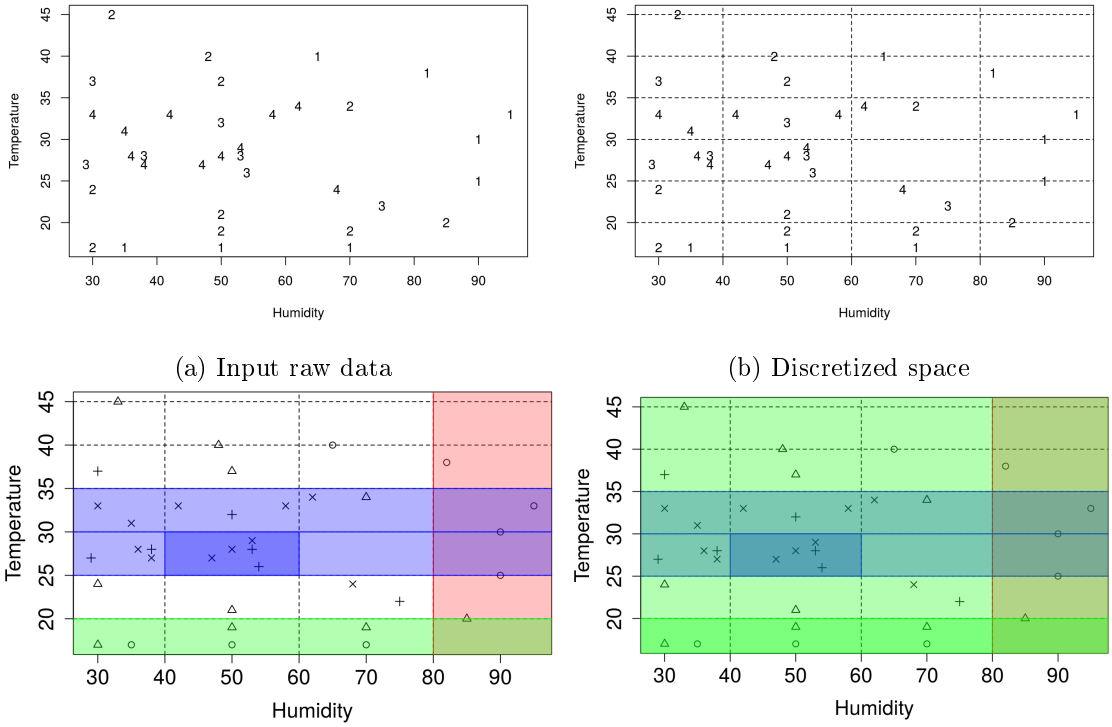
Class association rule learning is executed on the training dataset $\mathbf{T}$ with $C$ as the target attribute. The output is a set of all rules that meet the predefined minimum support and minimum confidence thresholds (and possibly some other constraints and settings).

**Definition 11.** ***(rule list)*** *The output of class association rule learning is an ordered sequence of rules* $R = \langle r_1, \ldots, r_m \rangle$.

## 9.3. CBA Model Building

The first step in building a CBA classifier is learning a list of classification association rules. Association rule learning requires discretized data. Discretization splits the data space into hypercubes. The discovered rule delimits one of the hypercubes and assigns a class label to it. The CBA algorithm prunes the discovered rules and adds a default rule to the end of the rule list.

**Example (CBA model building).** Let's introduce the "HumTemp" sample data for this chapter (Figure 9.2a). There are two explanatory attributes (Temperature and Humidity). The target attribute is preference (e.g. subjective comfort level). The data were discretized using equidistant binning (Figure 9.2b). Given that there are two attributes, the discovered rule can only correspond to a rectangular region with borders aligned to the grid (Figure 9.2c). The color correspond to the class which is predicted by the rule (red = 1, green =2, blue = 4). The rules are processed by CBA to form a classifier (Figure 9.2d). In this case, CBA only added a default rule (green background) that ensures that all instances are classified.




(a) Input raw data

(b) Discretized space

(c) Discovered rules in discretized space.

(d) CBA model

Figure 9.2.: Illustration of CBA model building (HumTemp dataset)

## 9.4. Refit

As a first step, MARC *refits* rule boundaries to a finer grid, which corresponds to all unique attribute values actually appearing in the data (Algorithm 2). The refit operation is inspired by the way the C4.5 decision tree learning algorithm selects splitting points for numerical attributes – it also searches for a value that is present in the training data [Quinlan, 1993].

**Algorithm 2** Refit rule *refit()*

**Require:** $r$ – input rule learnt on discretized training data
**Ensure:** rule $r$ with refit literals

1: **for** $literal = (A, V) \in antecedent(r)$ **do** {V is a value range, e.g. interval [20;30) and A is attribute, e.g. humidity}
2: $\mathcal{A} \leftarrow$ all unique values appearing in training data in attribute A.
3: $left \leftarrow min(\mathcal{A} \cap V)$
4: $right \leftarrow max(\mathcal{A} \cap V)$
5: $r \leftarrow$ replace $literal$ in $r$ with new *literal = (A,[left,right])*
6: **end for**
7: **return** $r$

## 9.5. Literal Pruning

The literal pruning step removes redundant attributes (attribute-value pairs) from rules. Literal is considered redundant if its removal does not decrease rule confidence. Literal pruning is depicted in Algorithm 3.

**Example (Literal pruning).** To demonstrate literal pruning, we need to use dataset with higher number of attributes than the HumTemp dataset has. Let's use the well-known Iris dataset [Fisher, 1936]. The model built without literal pruning is in Figure 9.3 and the model with literal pruning is in Figure 9.4. As follows from comparison of the individual rules, removal of the literals did not affect rule confidence:

- From Rule 1 literal `Petal.Length=[-Inf;Inf]` was removed since it had no discriminatory power due to boundaries [-Inf;Inf], which were result of the refit operation.
- From Rule 2 literal `Sepal.Length=[6.1;Inf]` was removed. Note that the boundaries of the remaining literal created over `Petal.Length` attribute differ. The reason is that since literal pruning is performed as a second step in QCBA after refitting, the result of extension can affect the final boundaries. This is also the reason why the support of the rule after literal pruning has decreased, although a literal was removed.
- Rule is unaffected.

```
## id  rule                                                                  support confidence
## 1 {Petal.Length=[-Inf;Inf],Petal.Width=[-Inf;0.6]}   => {Species=setosa}     0.32       1.00
## 2 {Sepal.Length=[6.1;Inf],Petal.Length=[5.1;Inf]}    => {Species=virginica}  0.27       1.00
## 3 {Sepal.Width=[-Inf;3.1],Petal.Width=[1.8;Inf]}     => {Species=virginica}  0.20       1.00
## 4 {}                                                 => {Species=versicolor} 0.36       0.36
```

Figure 9.3.: Example QCBA model - no literal pruning (Iris Dataset)

```
## id  rule                                                                 support confidence
## 1 {Petal.Width=[-Inf;0.6]} => {Species=setosa}                             0.32       1.00
## 2 {Petal.Length=[5.2;Inf]} => {Species=virginica}                          0.25       1.00
## 3 {Sepal.Width=[-Inf;3.1],Petal.Width=[1.8;Inf]} => {Species=virginica}    0.20       1.00
## 4 {} => {Species=versicolor}                                               0.36       0.36
```

Figure 9.4.: Example QCBA model - with literal pruning (Iris Dataset)

**Algorithm 3** Literal pruning *pruneLiterals()*

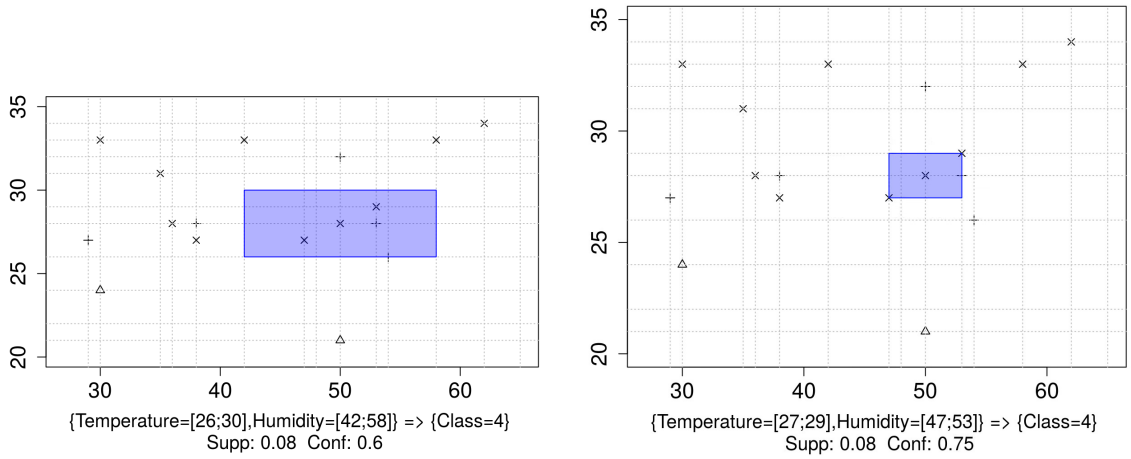
**Require:** $r$ – input rule
**Ensure:** rule $r$ with redundant attributes (literals) removed

1: $attrRemoved \leftarrow$ false
2: **repeat**
3:  **for** $literal \in antecedent(r)$ **do** {Literals are iterated in arbitrary order}
4:  $r' \leftarrow$ remove $literal$ from $r$
5:  **if** $confidence(r') \geq confidence(r)$ **then**
6:  $r \leftarrow r'$
7:  $attrRemoved \leftarrow$ true
8:  **break**
9:  **else**
10:  $attrRemoved \leftarrow$ false
11:  **end if**
12:  **end for**
13: **until** $attrRemoved =$ false
14: **return** $r$

## 9.6. Trimming

The trimming operation processes all rules. Literal boundaries in the given rule are shaved of any values that belong solely to instances that are misclassified by this rule (Algorithm 4).

**Example (Trimming).** Figure 9.5b shows the result of trimming of rule in Figure 9.5a: rule is shaved of one misclassified data point, confidence rises from 0.6 to 0.75.




(a) Original rule after refit

(b) Trimming applied

Figure 9.5.: Illustration of trimming algorithm in QCBA model building (HumTemp dataset)

**Algorithm 4** Rule trimming *trim()*

**Require:** $r$ – input rule
**Ensure:** rule $r$ with trimmed literals

1: $corrCovByR \leftarrow$ training instances covered and correctly classified by $r$
2: **for** $literal = (A, V) \in antecedent(r)$ **do** {Literals are iterated in arbitrary order}
3: $corrCovByL \leftarrow$ training instances covered by $literal$
4: $distValsL \leftarrow$ distinct values training instances have in attribute $A$
5: **if** $size(distValsL) \leq 1$ **then**
6: **continue**
7: **end if**
8: $distValsLinR \leftarrow$ distinct values of attribute $A$ in $corrCovByR$
9: $V' \leftarrow [min(distValsLinR), max(distValsLinR)]$
10: $r \leftarrow$ replace $literal$ in $r$ with new $literal = (A, V')$
11: **end for**
12: **return** $r$

## 9.7. Extension

The extension process is depicted in Algorithm 5. The ranges of literals in the body of each rule are attempted to be enlarged. The range of each literal is increased one literal and one boundary at a time. The extension is generally accepted only if it improves rule confidence. To overcome local minima, the extension process can provisionally accept drop in confidence of intermediate result of extension compared to the seed rule.

**Algorithm 5** Rule Extension *extendRule()*

**Require:** $train = \{x_l | l = 1, \ldots, n\} - n$ train objects, defined over $m$ attributes: $\{A_i | i = 1, \ldots, m\}$, $r$ – input rule, $minImprovement \in (-1, 1)$ with 0 as default, $minCondImprovement \in (-1, 0)$ with -1 as default
**Ensure:** extended rule $r$

```
1: curBest ← r
2: directExtensions ← getExtensions(r,train)
3: repeat
4:   extensionSuccessful ← false
5:   for cand ∈ directExtensions {Iteration in order according to criteria in Figure 9.8} do
6:     Δ_conf ← conf(cand) - conf(curBest), Δ_supp ← sup(cand) - sup(curBest)
7:     if crispAccept(Δ_conf, Δ_supp, minImprovement) then
8:       curBest ← cand, extensionSuccessful ← true
9:       break
10:    else if conditionalAccept(Δ_conf, minCondImprovement) then
11:      enlgmnt ← cand
12:      loop
13:        enlgmnt ← getBeamExtension(enlgmnt)
14:        if enlgmnt=∅ then
15:          break
16:        end if
17:        Δ_conf ← conf(enlgmnt) - conf(curBest), Δ_supp ← sup(enlgmnt) - sup(curBest)
18:        if crispAccept(Δ_conf, Δ_supp, minImprovement) then
19:          curBest ← enlgmnt, extensionSuccessful ← true
20:          break
21:        else if conditionalAccept(Δ_conf, minCondImprovement) then
22:          continue {Extension in conditional accept band}
23:        else
24:          break
25:        end if
26:      end loop
27:      if extensionSuccessful = true then
28:        break
29:      end if
30:    else
31:      continue {Improvement below conditional threshold, going to next candidate}
32:    end if
33:  end for
34: until extensionSuccessful = false
35: return curBest
```

First, for given seed rule $r$ the algorithm retrieves all possible extensions. These are generated by Algorithm 6 on page 145. The algorithm refers to the notion of direct extension presented in Definition 12 on page 146. Candidate extensions are sorted according to criteria applied by CBA, which are depicted in Figure 9.8 on page 148.

**Algorithm 6** Get Extensions *getExtensions()*

**Require:** $train = \{x_l | l = 1, \ldots, n\} - n$ train objects, defined over $m$ attributes: $\{A_i | i = 1, \ldots, m\}$, rule $r$
**Ensure:** up to two extensions of rule $r$

1: $extendedRules \leftarrow \emptyset$
2: **for** $literal \in body(r)$ **do**
3: **if** *type(literal)*=nominal **then** {Nominal attributes are skipped}
4: **continue**
5: **end if**
6: $neighbourhood \leftarrow$ direct extension of $literal$ in $train$ {See Def. 12}
7: **for** $extendedLiteral \in neighbourhood$ **do**
8: $extRule \leftarrow$ replace $literal$ in $r$ with $extendedLiteral$
9: $extendedRules \leftarrow extendedRules \cup extRule$
10: **end for**
11: **end for**
12: **return** $extendedRules$

**Algorithm 7** Beam Rule Extension *getBeamExtension()*

**Require:** $rule$ $\{r\}$
**Ensure:** extended rule $r$ or null

1: $literal$, $extendType \leftarrow$ let $r'$ be a rule from which $r$ was created by direct extension of $extendType$=$\{higher, lower, nominal\}$ by replacing $l \in R$ by $literal$
2: **if** $extendType$ = nominal **then**
3: **return** $\emptyset$ {Extension is not applicable on nominal attributes}
4: **else**
5: $extendedLiteral \leftarrow$ direct extension of $literal$ of type $extendType$
6: **if** $extendedLiteral = \emptyset$ **then**
7: **return** $\emptyset$ {Direct extension not found}
8: **end if**
9: $extRule \leftarrow$ replace $literal$ in $r$ with $extendedLiteral$
10: **end if**
11: **return** $extRule$

**Example (Extension).** Figures 9.6a-9.7d show on eight plots the progress of the extension operation.

1.-6. Rule coverage is extended. Confidence and support is unaffected.

7. Left boundary on Humidity conditionally extends - one additional misclassified instance and one correctly classified instance are covered. Confidence drops to 0.67.

8. Left boundary on Humidity conditionally extends - one additional correctly classified instance is covered. Confidence increases to 0.71, but is still below 0.75.

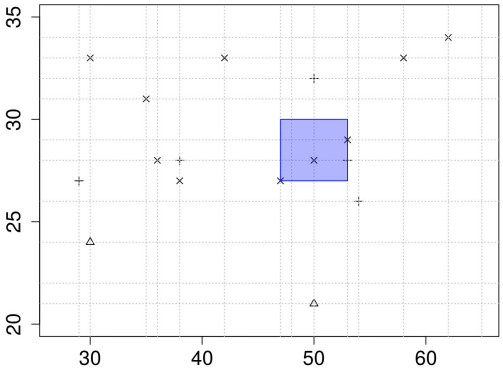
9. Left boundary on Humidity extends - one additional correctly classified instance is covered. Confidence returns 0.75.

10. Left boundary on Humidity extends. Confidence and support is unaffected. Final rule.

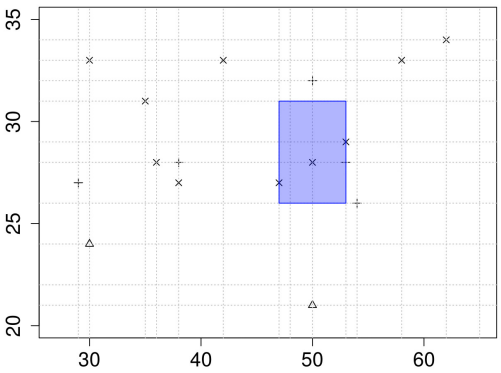
**Definition 12.** ***(direct extension of literal defined over cardinal attribute)*** *Let $l = (A, V)$ be a cardinal literal, and $V = \langle x_i, \ldots, x_j, \ldots, x_k \rangle$ a value range. A direct extension of $l$ is a set $E_l$ of up to two literals derived from $l$: higher direct extension and lower direct extension. A higher direct extension of $l$ is a literal $l_H = (A, V')$, where $V' = \langle x_i, \ldots, x_j, \ldots, x_k, x_{k+1} \rangle$. A lower direct extension of $l$ is a literal $l_L = (A, V')$, where $V' = \langle x_{i-1}, x_i, \ldots, x_j, \ldots, x_k \rangle$. If both higher and lower extension exist, $E_l$ has two elements, if only one exists, $E_l$ has one element, if none of these extensions exists, $E_l$ is empty.*

On line 7 of Algorithm 5, an extension is accepted if it meets criteria for *crisp accept* (Figure 9.9). If the extension does not meet one of these conditions, it can still be conditionally accepted on line 18 (Figure 9.10). The conditional accept sets a direction, on which the algorithm "locks" the beam extension, and on lines 12-26 verifies, whether this direction will yield an unconditional accept, or not. The *getBeamExtension* procedure is depicted in Algorithm 7. Note that the extension can also be accepted if the extension does not cover any additional training instance, which results in confidence as well as support remaining unchanged after extension.

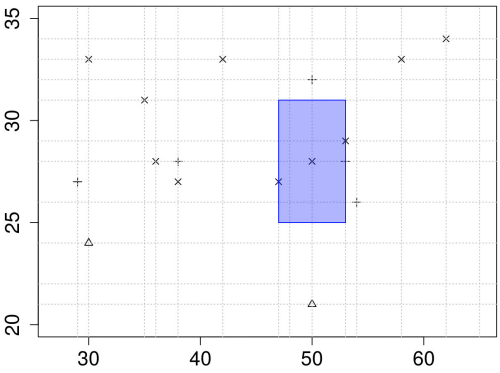
By default, extend is accepted if it does not deteriorate rule confidence, which corresponds to setting of the *minImprovement* meta-parameter to 0 by the user. This value can be increased if the user desires to reduce the number of extensions tried, resulting in performance improvement. As for the conditional accept process, by default all extensions in given direction are tried until all values are exhausted, which corresponds to $minCondImprovement = -1$. The user may wish to decrease this value to obtain faster failure of the conditional extension process, improving performance on datasets with many distinct values (cf. Subsection 10.3.3).



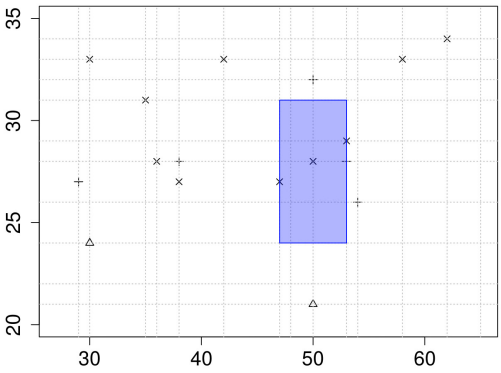

(a) Extension process – step 1



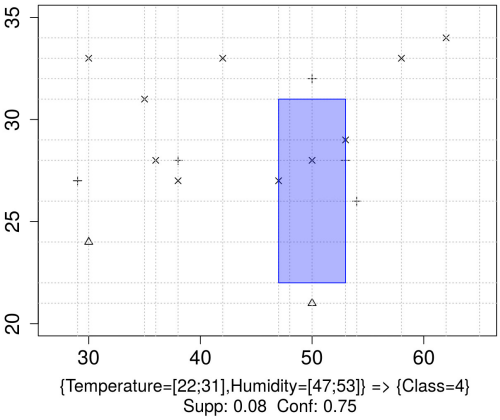

(b) Extension process – step 2



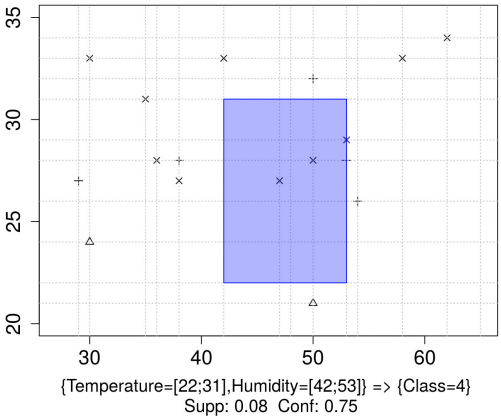

(c) Extension process – step 3




(d) Extension process – step 4

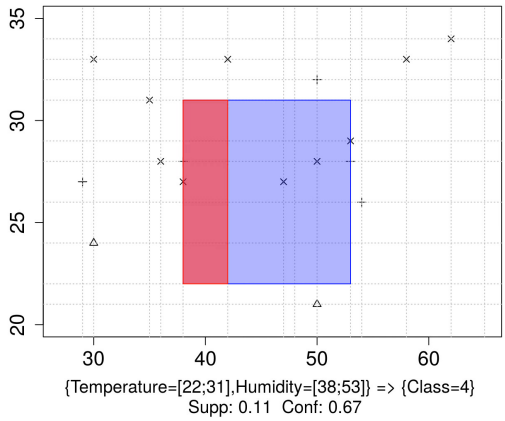



(e) Extension process – step 5

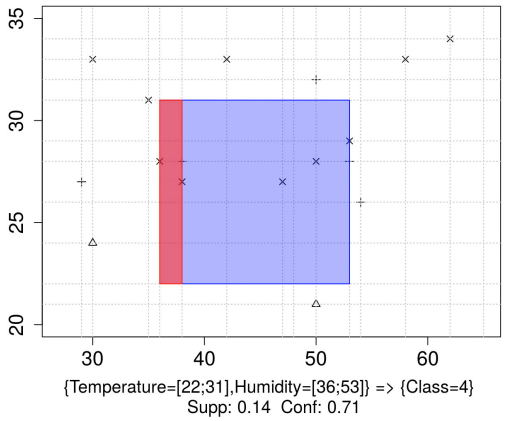



(f) Extension process – step 6

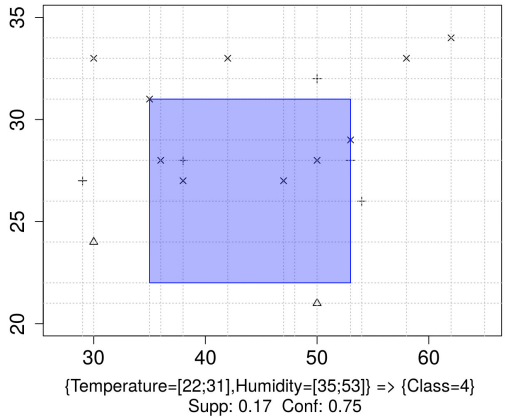
Figure 9.6.: Illustration of extension algorithm (HumTemp dataset)



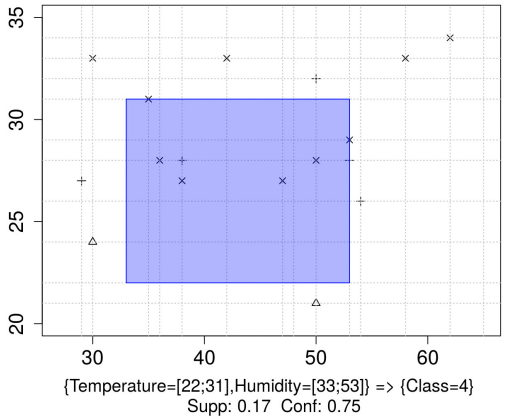

(a) Extension process – step 7




(b) Extension process – step 8




(c) Extension process – step 9




(d) Extension process – step 10

Figure 9.7.: Illustration of conditional extension algorithm (HumTemp dataset)

1. rule $A$ is ranked higher if confidence of rule $A$ is greater than that of rule $B$,
2. rule $A$ is ranked higher if confidence of rule $A$ is the same as confidence of rule $B$, but support of rule $A$ is greater than that of rule $B$,
3. rule $A$ is ranked higher if rule $A$ has shorter antecedent (fewer conditions) than rule $B$.

Figure 9.8.: Rule ranking criteria

**Crisp accept**: rule is accepted if i) the support of the candidate does not drop below the original rule and ii) confidence improves at least by the predefined threshold.

1. **if** $\Delta_{conf} \geq minImprovement$ **and** $\Delta_{supp} \geq 0$ **then** true
2. **else** false

Figure 9.9.: $crispAccept(\Delta_{conf}, \Delta_{supp}, minImprovement)$

**Conditional accept**: rule is conditionally accepted if confidence improves at least by the predefined threshold.

1. **if** $\Delta_{conf} \geq minCondImprovement$ **then** true
2. **else** false

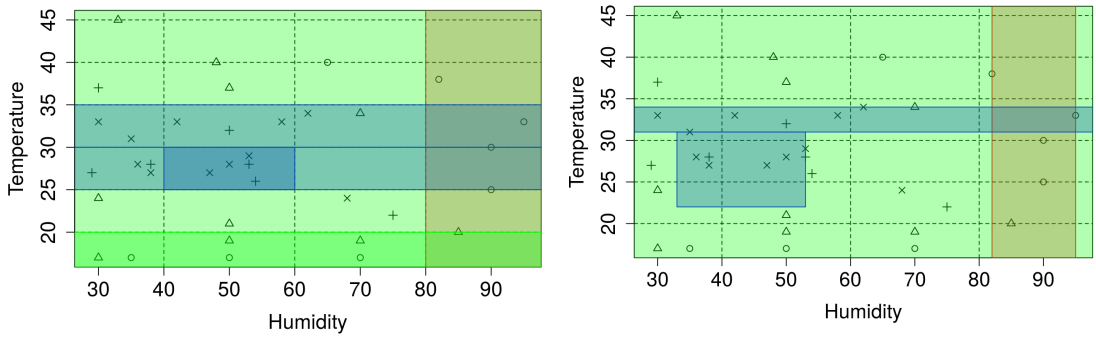
Figure 9.10.: $conditionalAccept(\Delta_{conf}, minCondImprovement)$

## 9.8. Postpruning

The previous steps affected individual rules, changing their coverage. The number of rules can now be reduced using adaptation of CBA's data coverage and default rule pruning (Algorithm 8). This will also add a default rule to the end of the rule list. We refer to this second iteration[2] of CBA as *postpruning*. Each rule is matched against the training data. If a rule does not correctly classify any object, it is discarded. Otherwise, the rule is kept. In any case, objects matching the rule are discarded. The data coverage pruning is combined with default rule pruning, which determines the rule with the lowest number of errors on training data if rules below it are replaced by a default rule, and performs this replace.

**Example (Postpruning).** Figure 9.11a shows a CBA model. After refit and extension have been performed, postpruning in MARC removed two rules from the CBA model (9.11b). The default rule was recomputed, but still classifies to green (Class 2).

[2]The result of the first iteration of data coverage pruning is the rule list on the input of MARC. We obtained better results if default rule pruning is not performed during the first iteration (within CBA), since in this way MARC is left with more rules to optimize.




(a) Postpruning example (b) Postpruning example

Figure 9.11.: Illustration of postpruning algorithm (HumTemp dataset)

**Algorithm 8** Postpruning *postPruning()*

**Require:** rules – output of *extendRuleList()*, set of training instances $T$
**Ensure:** pruned *rules* (some elements of input rule list removed, default rule added)

```
1:
2: cutoffRule ← ∅
3: cutoffClass ← most frequent class in T
4: lowestTotalError ← |T| − |t ∈ T : class(t) = cutoffClass|
5: totalErrorsWithoutDefault ← 0
6: rules ← sort rules according to criteria in Fig. 9.8
7: defClass ← most frequent class in T
8: {Data coverage pruning}
9: for all r ∈ rules do
10:    covered ← instances in T matched by antecedent(r)
11:    corrCovered ← instances in T matched by antecedent(r) and consequent(r)
12:    T := T \ covered {Remove instances covered by r from training data}
13:    if corrCovered=∅ then
14:       rules ← rules \ r {remove r from rules}
15:    else
16:       misclassified ← covered - corrCovered
17:       totalErrorsWithoutDefault ← totalErrorsWithoutDefault + misclassified
18:       defClass ← most frequent class in T
19:       defaultRuleError ← |T| − |t ∈ T : class(t) = defClass|
20:       totalErrorWithDefault ← defaultRuleError + totalErrorsWithoutDefault
21:       if totalErrorWithDefault<lowestTotalError then
22:          cutoffRule, lowestTotalError, cutoffClass ← r, totalErrorWithDefault, defClass
23:       end if
24:    end if
25: end for
26: {Default rule pruning}
27: rules ← remove all rules below cutoffRule from rules
28: rules ← append new default rule "{} → cutoffClass' at the end of rules
29: return rules
```

```
## id   rule                                                    support  confidence
## 1  {Temperature=[21;45],Humidity=[82;95]} => {Class=1}     0.11       1.00
## 2  {Temperature=[34;45],Humidity=[33;58]} => {Class=2}     0.08       1.00
## 3  {Temperature=[29;34],Humidity=[29;48]} => {Class=4}     0.08       1.00
## 4  {Temperature=[19;26],Humidity=[29;53]} => {Class=2}     0.08       1.00
## 5  {Temperature=[37;45],Humidity=[53;95]} => {Class=1}     0.06       1.00
## 6  {Temperature=[34;40],Humidity=[29;47]} => {Class=3}     0.03       1.00
## 7  {Temperature=[17;24],Humidity=[82;95]} => {Class=2}     0.03       1.00
## 8  {Temperature=[29;34],Humidity=[29;68]} => {Class=4}     0.17       0.86
## 9  {Temperature=[22;31],Humidity=[33;70]} => {Class=4}     0.19       0.70
## 10 {Temperature=[17;21]}                  => {Class=2}     0.14       0.63
## 11 {}                                     => {Class=3}     0.20       0.19
```

Figure 9.12.: Default rule overlap pruning example: initial rule list

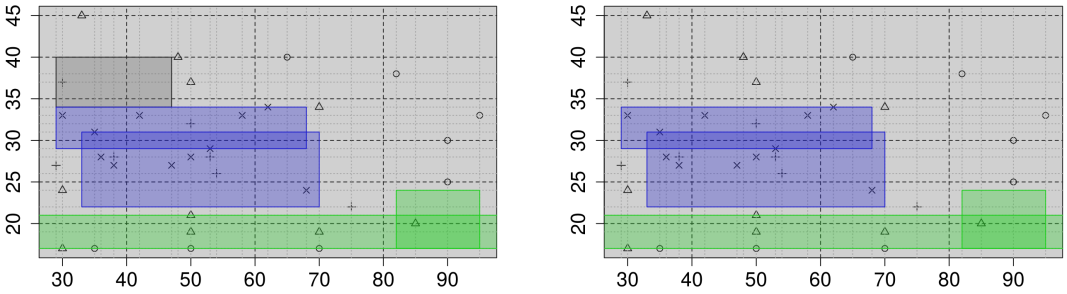
## 9.9. Default Rule Overlap Pruning

Default rule overlap pruning (*drop* for short) iterates through all rules classifying into the same class as the default rule. These rules all overlap with the default rule both in terms of coverage and class assigned and are thus candidates for pruning. They can be removed only if their removal will not change classification of those instances *in training data/in the entire instance space* that these rules correctly classify. This change in classification would be caused by rules that are between these rules and the default rule. We consider two versions of drop: *transaction-based* and *range-based*.

The *transaction-based* version, depicted in Algorithm 9, removes rule if there is no transaction in the training data, which would be misclassified as a result of removing this rule.

The *range-based* version analyzes overlaps in the range of literals in the pruning candidate with respect to ranges in the potentially clashing rules (rules classifying to different class) below it. The pruning is confirmed only if the potentially clashing rules cover different "geometric" regions (Algorithm 10). Range-based pruning thus guarantees a solution that generalizes beyond the training data. Its potential disadvantage is that it removes less rules, since it is stronger than the transaction-based pruning.

**Example (Default Rule Overlap Pruning - Transaction-based).** Let's us create a different classifier from the HumTemp dataset by lowering the minimum support threshold and disabling trimming (Figure 9.12). The default rule overlap pruning removes rule #6. This rule assigns into Class 3 - the same class as the default rule in the end of the classifier (grey background). If we look at rules between #6 and the default rule #11, we can observe that there is no rule below #6 that would prevent the training instances covered by #6 from being classified by the default rule (which has the same class as #6). Figure 9.13a depicts rule #6 to #11 in the original classifier with default rule overlap pruning disabled. When default rule overlap pruning is activated, rules such as #6 are removed and the area is left for classification to the default rule, which comes as last (Figure 9.13b).




(a) Initial rule list

(b) Rule #6 removed

Figure 9.13.: Illustration of default rule overlap pruning algorithm (HumTemp dataset). Class 2 is assigned green color, Class 3 grey, and Class 4 blue.

**Algorithm 9** Default Rule Overlap Pruning (Transaction-based) *drop-tr()*

**Require:** *rules*, set of training instances $T$
**Ensure:** pruned *rules* (some elements of *rules* removed)

*defRule* ← default (last) rule in *rules*
**for all** *prunCand* ∈ *rules* **do**
  **if** *consequent(prunCand)* ≠ *class(defRule)* **or** *prunCand* = *defRule* **then**
    **continue**
  **end if**
  *corrCovered* ← instances in $T$ correctly classified by *prunCand*
  *nonEmptyIntersection* ← FALSE
  **for all** *candClash* ∈ rules below *prunCand* in *rules* **do**
    **if** *consequent(candClash)* = *class(defRule)* **then**
      **continue**
    **end if**
    *candClashCovered* ← instances in $T$ matching *antecedent(candClash)*
    **if** *candClashCovered* ∩ *corrCovered* ≠ ∅ **then**
      *nonEmptyIntersection* ← TRUE
      **break**
    **end if**
  **end for**
  **if** *nonEmptyIntersection* = FALSE **then**
    *rules* ← *rules* \ *prunCand*
  **end if**
**end for**
**return** *rules*

**Algorithm 10** Default Rule Overlap Pruning (Range-based) *drop-ra()*

**Require:** *rules*, set of training instances $T$
**Ensure:** pruned *rules* (some elements of *rules* removed)

```
1: defRule ← default (last) rule in rules
2: for all prunCand ∈ rules do
3:   if consequent(prunCand) ≠ class(defRule) or prunCand = defRule then
4:     continue
5:   end if
6:   literals ← literals in antecedent(prunCand)
7:   attributes ← attributes appearing in antecedent(prunCand)
8:   clashingRuleFound ← FALSE
9:   for all candClash ∈ rules below prunCand in rules do
10:    if consequent(candClash) = class(defRule) then
11:      continue
12:    end if
13:    sharedAttributes ← attributes∩ attributes in antecedent(candClash)
14:    if sharedAttributes = ∅ then
15:      clashingRuleFound ← TRUE {No shared attribute with potentially disjunct ranges
   results in the two rules overlapping on a subset of the data space}
16:      break
17:    end if
18:    literalsInClashOnSharedAtt ← literals in antecedent(candClash) defined over attributes
   in sharedAttributes
19:    {if there is NO intersection on at least one of the shared attributes we have no CLASH}
20:    attLeastOneAttDisjunct ← FALSE
21:    for all literalCC = (A, V) ∈ literalsInClashOnSharedAtt do
22:      literal ← get literal in antecedent(prunCand), which is defined over attribute A
23:      if V has empty intersection with value range of literal then
24:        attLeastOneAttDisjunct ← TRUE
25:        break
26:      end if
27:    end for
28:    if attLeastOneAttDisjunct = FALSE then
29:      clashingRuleFound ← TRUE
30:    end if
31:  end for
32:  if clashingRuleFound = FALSE then
33:    rules ← rules \ prunCand
34:  end if
35: end for
36: return rules
```

# 10. Experiments

In this section, we present evaluation of the MARC (QCBA)[1] framework on a number of standard datasets. The evaluation focuses on comparison with CBA in terms of accuracy, classifier size and runtime. In order to verify the correctness of our CBA implementation, the last subsection is devoted to comparison of results we obtained and those reported by the CBA authors.

**Chapter organization.** Section 10.1 describes the datasets used for evaluation. Section 10.2 presents the evaluation of various MARC (QCBA) setups. Section 10.3 presents the results. Section 10.4 verifies our reimplementation of CBA against earlier published results. Finally, Section 10.5 explores the connection between the results obtained from the experiments with results and models covered in Part I of the thesis.

## 10.1. Datasets

University of California provides at `https://archive.ics.uci.edu` a set of publicly available datasets, which are commonly used for benchmarking machine learning algorithms. Several datasets come from visual information processing or signal processing domains (ionosphere, letter, segment, sonar). The second strongly represented domain are medical datasets (colic, breast-w, diabetes, heart-statlog, lymph). Eleven datasets are binary classification problems, nine datasets are multi-class and two datasets have ordinal class attribute (autos and labour).

### 10.1.1. Selection Criteria

We chose 22 datasets to perform the evaluation. The selection criteria were a) at least one numerical predictor attribute, b) the dataset being previously used in evaluation of symbolic learning algorithms in one of the following seminal papers: Alcala-Fdez et al. [2011], Hühn and Hüllermeier [2009], Liu et al. [1998], Quinlan [1996]. Details of the selected datasets are given in Table 10.1.

### 10.1.2. Preprocessing, Missing Value Treatment

Missing values in numerical attributes were replaced by mean. Missing values for categorical attributes were not imputed. All datasets contain cardinal attributes, which needed to be pre-discretized for CBA and QCBA. All numeric explanatory attributes with three

[1] As noted earlier, the MARC framework implementation, which we used for evaluation, is called QCBA.

| dataset | att. | inst. | miss. | class | description |
|---|---|---|---|---|---|
| anneal | 39 | 898 | Y | nominal (6) | NA |
| australian | 15 | 690 | N | binary | credit card applications |
| autos | 26 | 205 | Y | ordinal (7) | riskiness of second hand cars |
| breast-w | 10 | 699 | Y | binary | breast cancer |
| colic | 23 | 368 | Y | binary | horse colic (surgical or not) |
| credit-a | 16 | 690 | Y | binary | credit approval |
| credit-g | 21 | 1000 | N | binary | credit risk |
| diabetes | 9 | 768 | N | binary | diabetes |
| glass | 10 | 214 | N | nominal (6) | types of glass |
| heart-statlog | 14 | 270 | N | binary | diagnosis of heart disease |
| hepatitis | 20 | 155 | Y | binary | hepatitis prognosis (die/live) |
| hypothyroid | 30 | 3772 | Y | nominal (3) | NA |
| ionosphere | 35 | 351 | N | binary | radar data |
| iris | 5 | 150 | N | nominal (3) | types of irises (flowers) |
| labor | 17 | 57 | Y | ordinal (3) | employer's contribution to health plan |
| letter | 17 | 20000 | N | nominal (26) | letter recognition |
| lymph | 19 | 148 | N | nominal (4) | lymphography domain |
| segment | 20 | 2310 | N | nominal (7) | image segment classification |
| sonar | 61 | 208 | N | binary | determine object based on sonar signal |
| spambase | 58 | 4601 | N | binary | spam detection |
| vehicle | 19 | 846 | N | nominal (4) | object type based on silhouette |
| vowel | 13 | 990 | N | nominal (11) | NA |

Table 10.1.: Overview of datasets involved in the benchmark. att. denotes number of attributes, inst. number of instances (objects), miss. whether or not the dataset contains missing observations.

or more distinct values were subject to discretization using the MDLP algorithm. Other algorithms involved in the benchmark did not require prediscretization. The evaluation was performed using a 10-fold stratified cross-validation. To ensure that all algorithm runs will use exactly the same folds, the folds were materialized.

## 10.2. Experiment Setup

The CBA algorithm has three hyperparameters – minimum confidence threshold, minimum support threshold and the total number of candidate rules. In Liu et al. [1998] it is recommended to use 50% as minimum confidence, 1% as minimum support. For our experiments, we used these thresholds. In Liu et al. [1998] 80.000 was used as the total number of rules, however it was noted that the performance starts to stabilize already around 60.000 rules. According to our experiments, there is virtually no difference between 80.000 and 50.000 threshold apart from the higher computation time for the former, therefore we used 50.000.[2] We also limited the maximum number of items per itemset to 5.

All results were obtained using open source CBA and MARC (QCBA) implementations

[2]In our best setup, we observed less than 0.1% improvement in average accuracy and 5% increase in average rule count when the maximum number of rules was increased from 50.000 to 80.000.

available at `https://cran.r-project.org/web/packages/arc/` and `https://github.com/kliegr/qcba`. All experiments can be directly replicated using the evaluation framework published at `https://github.com/kliegr/arc`.

The MARC (QCBA) framework does not have any mandatory thresholds. The extension process (Algorithm 5) contains two numeric parameters, which were left set to their default values *minImprovement=0* and *minCondImprovement=-1*. These default values have natural explanations (cf. Subsection 9.7) and the tuning of these thresholds can be generally recommended only for improving runtime on larger datasets.

It should be noted that MARC (QCBA) takes on the result of CBA execution with default rule pruning not performed, a variation of CBA not reported to be evaluated in Liu et al. [1998] or in other prior research. Our CBA implementation, available in the `arc` package, was thus adapted to allow deactivation of default rule pruning.

### 10.2.1. Evaluation Methodology

We evaluated several variations of the MARC (QCBA) setup. As a baseline, we use a CBA run with default parameters. The purpose of this evaluation is to show the effect on classification performance and the size of the model.

Classification performance is measured by accuracy, which is computed as $correct/N$, where *correct* is the number of correct predictions and $N$ the total number of objects. All results are reported using ten-fold cross validation with macro averaging. The average accuracy for all 22 datasets is reported as an indicative comparison measure. For a more reliable comparison, we included the won-tie-loss matrix, which compares two classifiers by reporting the number of datasets where the reference classifier wins, loses or the two classifiers perform equally well. We include p-value for the Wilcoxon signed test, which is recommended for comparison of classifiers over multiple datasets in the authoritative work of Demšar [2006].

We use three metrics to measure the size of the model: average antecedent length (number of conditions in the rule), number of rules per model and average number of conditions per model computed as number of rules times average antecedent length.

Despite our implementation not being optimized for speed, we decided to include a benchmark indicating how much processing power the postprocessing by MARC (QCBA) requires. The build times reported were computed as an average of classifier learning time for 220 models (10 folds for each of the 22 datasets).

## 10.3. Results

Summary of results is presented in Table 10.2. The table presents baseline results for CBA and then seven different configurations for MARC (QCBA), which allow us to demonstrate the effect of all individual postprocessing steps comprising MARC (QCBA). Configuration #1 corresponds to refit optimization being performed on top of CBA, configuration #2 to refit optimization and literal pruning, etc. Configuration #6 and #7 correspond to the full

MARC (QCBA), where (#6) used transaction-based default rule overlap pruning (drop) and (#7) range-based version of default rule overlap pruning.

### 10.3.1. Accuracy

The MARC (QCBA) setup that produces the highest accuracy while achieving maximum reduction in the size of the classifier is configuration #7, which includes all optimization steps, very closely followed by #5, which excludes default rule overlap pruning. These configurations have the same average accuracy as CBA and surpass CBA what concerns the won-tie-loss metric: they win on 14 datasets while CBA wins on 7 datasets and there is a draw[3] on 1 dataset. The p-value for the Wilcoxon signed rank test indicates that the change in the won-tie-loss matrix is not significantly different compared to CBA for any of the MARC (QCBA) configurations. It should be noted that the p-value of 0.12 of MARC (QCBA) configurations #5 and #7 is close to the 10% significance level, marginally improving on CBA results. The configuration number #6 performs exactly equally well as CBA winning on 11 datasets and loosing also on 11 datasets.

### 10.3.2. Classifier Size

The models produced by the best-performing MARC (QCBA) configuration #7 are smaller than CBA models: there is a reduction of 21% in the number of rules and 18% in the average number of conditions. These reductions combined amount to 35% reduction in model size in terms of total number of conditions. Further reduction in model size can be achieved by the transaction variant of default rule overlap pruning (#6), which reduces the size of the model by 53% to the average of 133 conditions from 285 conditions in the original CBA model. This additional improvement is offset by incurring 1% drop in the average accuracy. Overall, in terms of the won-tie-loss record, which is 11-0-11, the configuration #7 performs equally well as CBA.

As follows from comparison of #5 and #7, the range-based pruning was ineffective on this collection of datasets.

### 10.3.3. Runtime

The results for runtime are reported in the last two rows of Table 10.2. It can be seen that the refit, literal pruning and trimming optimizations take together roughly as much time on average as learning a CBA model. The most computationally intensive operation is extension. If we look at the median build times, we see that MARC (QCBA) prolongs CBA execution by factor of 3.5.

The discrepancy between median and average build times for MARC (QCBA) can be explained by several datasets for which MARC (QCBA) extension step takes excessive time to complete, which increases the average runtime and leaves median run time unaffected. Extension is particularly slow when there is a large number of distinct values in the dataset.

[3]Draw occurs when the accuracies on given dataset rounded to 0.1 percent match.

| configuration | cba | #1 | #2 | #3 | #4 | #5 | #6 | #7 |
|---|---|---|---|---|---|---|---|---|
| refit | na | Y | Y | Y | Y | Y | Y | Y |
| literal pruning | na | - | Y | Y | Y | Y | Y | Y |
| trimming | na | - | - | Y | Y | Y | Y | Y |
| extension | na | - | - | - | Y | Y | Y | Y |
| postpruning | na | - | - | - | - | Y | Y | Y |
| def. rule overlap (tran.) | na | - | - | - | - | - | Y | - |
| def. rule overlap (range) | na | - | - | - | - | - | - | Y |
| wins/ties/losses vs CBA | na | 14-1-7 | 15-0-7 | 12-0-10 | 11-0-11 | 14-1-7 | 11-0-11 | 14-1-7 |
| P-value (Wilcoxon) | na | 34% | 57% | 73% | 61% | 12% | 32% | 12% |
| accuracy (macro average) | 81% | 81% | 81% | 81% | 81% | 81% | 80% | 81% |
| avg conditions / rule | 3.4 | 3.4 | 2.8 | 2.8 | 2.8 | 2.8 | 2.8 | 2.8 |
| avg number of rules | 84 | 92 | 92 | 92 | 92 | 66 | 48 | 65 |
| avg conditions / model | 285 | 311 | 260 | 260 | 260 | 184 | 133 | 184 |
| build time [s] (median) | 12.3 | 23.8 | 19.9 | 19.8 | 42.7 | 42.9 | 43.2 | 42.9 |
| - normalized | 1.0 | 1.9 | 1.6 | 1.6 | 3.5 | 3.5 | 3.5 | 3.5 |
| build time [s] (average) | 18.4 | 34.7 | 37.1 | 37.6 | 319.2 | 317.1 | 318.2 | 319.3 |
| - normalized | 1.0 | 1.9 | 2.0 | 2.0 | 17.4 | 17.3 | 17.3 | 17.4 |

Table 10.2.: MARC (QCBA) evaluation – aggregate results for 22 UCI datasets

| | CBA (baseline) | | | CBA (Liu) | | MARC (#5) | | | MARC (#7) | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | acc | rules | con | acc | rules | acc | rules | con | acc | rules | con |
| anneal | .96 | 27 | 3.0 | .98 | 34 | .99 | 25 | 2.3 | .99 | 25 | 2.3 |
| australian | .85 | 109 | 4.0 | .87 | 148 | .82 | 42 | 3.8 | .87 | 74 | 3.8 |
| autos | .79 | 57 | 3.0 | .79 | 54 | .79 | 44 | 2.5 | .78 | 50 | 2.5 |
| breast-w | .95 | 51 | 2.8 | .96 | 49 | .95 | 20 | 2.7 | .95 | 31 | 2.7 |
| diabetes | .75 | 51 | 3.9 | .75 | 57 | .76 | 30 | 2.9 | .77 | 40 | 3.0 |
| glass | .71 | 28 | 3.9 | .73 | 27 | .69 | 22 | 2.8 | .69 | 24 | 2.8 |
| hepatitis | .79 | 32 | 3.9 | .85 | 23 | .82 | 22 | 3.0 | .82 | 28 | 3.0 |
| hypothyroid | .98 | 29 | 3.1 | .98 | 35 | .98 | 15 | 2.4 | .99 | 16 | 2.5 |
| ionosphere | .92 | 53 | 2.5 | .92 | 45 | .86 | 22 | 1.9 | .88 | 40 | 1.9 |
| iris | .92 | 6 | 2.0 | .93 | 5 | .93 | 4 | 1.1 | .93 | 5 | 1.2 |
| labor | .84 | 11 | 3.6 | .83 | 12 | .86 | 8 | 1.8 | .88 | 11 | 1.6 |
| lymph | .81 | 38 | 3.7 | .80 | 36 | .79 | 37 | 2.9 | .79 | 37 | 2.9 |
| sonar | .74 | 44 | 2.9 | .76 | 37 | .72 | 19 | 2.7 | .77 | 35 | 2.8 |
| vehicle | .69 | 147 | 3.9 | .69 | 125 | .69 | 79 | 3.6 | .71 | 106 | 3.7 |
| *average* | .84 | 49 | 3.3 | .84 | 49 | .83 | 28 | 2.6 | .84 | 37 | 2.6 |

Table 10.3.: Comparison of our results (included as *baseline* in the table) with Liu et al. [1998] (*Liu*). *acc* denotes accuracy, *rules* number of rules in the classifier, *con* number of conditions in rule antecedent

The slowest datasets were *segment*, *letter* and *spambase*. The *segment* and *letter* datasets contain various image metrics and *spambase* word frequency attributes. Such datasets are not typical representatives of use cases, where interpretable machine learning models are required. Nevertheless, the evaluation of the runtime indicates that the computational optimization of the extension algorithm is one of the most important areas for further work.

## 10.4. Verification of Results

The official implementation by authors of CBA [Liu et al., 1998] is not publicly available. We used our implementation to obtain the baseline results for CBA to support the central assertion that MARC (QCBA) reduces model size of CBA classifiers while keeping accuracy unaffected. In order to verify that our implementation is correct, we compared accuracy and number of rules[4] reported in Liu et al. [1998] for 14 datasets with results that we obtained. As noted earlier, we used nearly identical CBA setting as Liu et al. [1998] reports.

Detailed results are present in Table 10.3. While there are small variations for individual datasets, the overall average accuracy for our CBA implementation and the official one is equal at 84%. Regarding number of rules, original CBA has on average 49 rules, which is also exactly the same number as for our baseline implementation.

The precise match of the results came as a surprise, because our implementation of CBA does not include the optional pessimistic pruning step used in Liu et al. [1998], which is used in the first CBA phase when candidate association rules are generated.[5] According to results reported in Liu et al. [1998], the absence of pessimistic pruning has no effect on classifier accuracy, which is congruent with our results. While our results also indicate that this pruning has no effect on the number of rules in the classifier, the account of its effects in Liu et al. [1998] suggests that pessimistic pruning could be still an effective technique for reducing the time required to build the model.

In summary, comparison of our results with those reported by Liu et al. [1998] confirms the conclusion that MARC (QCBA) reduces the number of rules in CBA-built model while not negatively affecting the accuracy of the classifier.

## 10.5. Debiasing properties of MARC (QCBA) models

This section succinctly summarizes the connection between the results reported above and the analysis presented in Chapters 2 and 4.

The conclusions drawn from the crowdsourcing experiments presented in Chapter 5 are that several biases indeed occur during interpretation of the output of rule learning. As shown in Chapter 2, current research focusing on improving comprehensibility of machine

[4] The average length of the rules was unfortunately not reported in Liu et al. [1998].

[5] Note that Liu et al. [1998] explicitly marks pessimistic pruning step in CBA as optional. We could thus use standard association rule learner to obtain candidate rules.

learning models aims at reducing the size of the models. The results of the review of applicable cognitive biases in Chapter 4 showed that there are many specific effects that influence plausibility of rules requiring the designers of machine learning algorithms and software to adopt multiple specific debiasing strategies. However, our review also showed that the reduction of the size of the model presented to the analyst can help mitigate most biases, especially those linked to the representativeness heuristic. Each additional information in the model can trigger a range of biases, it is therefore essential to keep the models concise.

The CBA algorithm mitigates cognitive biases in the following ways:

- Rules do not contain negation. This can help prevent the negativity bias.
- Rules are sorted from the strongest to weakest. This ameliorates the primacy effect.
- Only the first rule in sort order is applied to score an instance.

One-rule classification and crisp rules make CBA classification models possibly most comprehensible among all association rule classification algorithms.

The postprocessing of CBA models with MARC (QCBA) framework further improves comprehensibility by reducing model size by approximately 50%. Removal of literals or complete rules from the rule list can help reduce the effect of multiple biases (cf. Table 4.1 on page 63). The favourable properties of CBA models are retained after MARC (QCBA) postprocessing. Model accuracy remains on the same level as for CBA.

# Part III.

# Conclusions

# 11. Summary of Contributions

The analysis presented in Chapter 4 is to the author's knowledge first systematic study of the impact of cognitive biases on interpretation of machine learning results. Based on a novel methodology for assessing model plausibility, we performed an empirical study using crowdsourcing. The results, presented in Chapter 5, validate the manifestation of selected biases in the rule learning domain. The results are integrated into a qualitative model of plausibility presented in Chapter 6. A practical result of our research are recommendations for the design of rule learning software.

Second, we aimed to reflect selected findings in the design of chosen machine learning algorithm in order to mitigate the effect of cognitive biases when models produced by this algorithm are interpreted. As a basis of our framework we chose the association rule classification algorithm CBA. Compared to most other machine learning models, rules have the advantage of being intuitively well understandable. They are also a commonly used knowledge representation in cognitive science. The association rule learning approach was selected over the more common separate and conquer approach for its ability to scale to large data. The proposed algorithm is described in Chapter 9. Empirical validation presented in Chapter 10 showed that the proposed framework fulfills the expectations – it reduces the number of rules in the classifier while maintaining accuracy. This improves model understandability and decreases possibilities for triggering of multiple cognitive biases derived from the representativeness heuristic.

**Chapter organization.** This chapter discusses the contributions in the order in which they appear in the thesis. In Section 11.1 we highlight the results of our literature review on cognitive biases. Section 11.2 presents the main contributions of the experimental study of the effect of cognitive biases on interpretation of rules discovered from data. Section 11.3 presents the proposed qualitative model of plausibility and summarizes the practical consequences of our study for the design of rule learning applications. Section 11.4 presents the main contributions of the proposed software framework.

## 11.1. Literature Review and Analysis (Chapters 2,4)

In the following, we summarize the main results of our literature review, focusing on findings that we consider as novel or not very well known in machine learning.

### 11.1.1. Review of Syntactic Comprehensibility in Machine Learning Research

In prior works in the area of machine learning, model comprehensibility relates to human understanding of the model in terms of its syntactic representation. We contrast this syntactic comprehensibility with semantic and pragmatic comprehensibility, which is a deeper level of understanding where people align the result of machine learning with their domain knowledge and reasoning patterns.

Emerging research moves in the direction of adjusting learning algorithms so that the syntactic comprehensibility of learned models is improved. Prior research has mostly focused on comparing entire representations (such as decision trees vs. decision rules) or on effects of hyperparameters (branching factor of a decision tree) on comprehensibility. Regarding domain knowledge, there is work on the importance of conformation to domain constraints and importance of semantic cohesion of attributes in inductively learned rules.

Our contribution is a review of prior research relating to syntactic comprehensibility in machine learning. We found that semantic comprehensibility, including the effect of cognitive biases, has not yet been studied.

### 11.1.2. Identification of Factors Reported To Affect Model Plausibility

The results pertaining to plausibility were scattered in articles dealing with other topics. To our knowledge, plausibility of machine learning models has not yet been systematically studied. Our contribution is identification of the following two factors that have been reported to affect plausibility of machine learning models:

1. **Oversimplicity avoidance.** Several authors have mentioned that domain experts have not trusted very simple machine learning models, such as a decision tree with a single inner node.

2. **Observation of domain constraints**. Perceived model plausibility depends on whether the model complies to domain constraints. The most studied are monotonicity constraints. An example of a monotonicity constraint on a numerical attribute is that the likelihood of whether a person will buy a product (target attribute) will increase with increasing product price. There is empirical evidence showing that domain experts do not find rules that contain conditions violating prior domain knowledge as plausible.

### 11.1.3. Analysis of Twenty Cognitive Biases

To our knowledge, cognitive biases have not yet been discussed in relation to interpretation of machine learning results. We thus initiated the review of research published in cognitive science with the intent to give psychological basis to changes in inductive rule learning algorithms which would foster better understanding of learning results. Our review identified twenty cognitive biases, heuristics and effects that can give rise to systematic errors

when inductively learned rules are interpreted. We propose to divide these heuristics and biases into two groups:

- Triggered by domain knowledge related to attributes and values in the rules. An example is aversion to ambiguous information.
- Generic strategies applied when evaluating alternatives. An example is insensitivity to sample size, which implies that rule confidence is considered as more important than rule support.

For most biases and heuristics involved in our study, psychologists have proposed "debiasing" measures. Our contribution is that we related these to machine learning, proposing how they could be applied to improve understanding of inductively learned rules.

### 11.1.4. Smaller Models Suppress Cognitive Biases

"Smaller is better" theories in machine learning are based on the Occam's razor principle. Minimizing the size of the model communicated to the analyst helps to improve interpretability and understanding of the model. However, in our review of literature from cognitive science, we did not identify results that would support this view. The only practical constraint are human cognitive capabilities – humans can process only 3-7 pieces of information at a time.

What our analysis did reveal is a number of cognitive phenomena that would make longer rules (or generally descriptions) more likely to trigger various cognitive biases than would shorter rules (descriptions). An example of such bias is the information bias – preference for more information even if it does not help to address the problem at hand. Overall, the analysis showed that most of the identified cognitive biases increase plausibility of longer rules (or generally descriptions). To summarize our contribution, we found indirect support in psychology for the "Smaller is better" paradigm used in many machine learning algorithms. While small models may not necessarily be found as more plausible by humans than larger models, smaller models provide less opportunities for cognitive biases to be triggered, leading to better, more truthful, comprehension. Additional discussion relating to limitations of these conclusions are present in Section 12.4.

## 11.2. Overview of Experimental Results (Chapter 5)

Given that we performed a first empirical validation of its kind, we had to invent a methodology for evaluating effect of cognitive biases on interpretation of rule learning results. We performed two crowdsourcing experiments. The first experiment directly targeted authentic rules discovered from data, relying on the proposed methodology. The second experiment was based on modifications of the Linda problem, which is frequently used as a basis for studying effects of cognitive biases related to the representativeness heuristic.

### 11.2.1. Measuring Effect of Cognitive Biases on Interpretation of Rules

Relevant research in cognitive science largely focuses on experiments demonstrating whether a specific bias occurs or not, quantifying the the proportion of subjects committing the bias, and describing the conditions that induce it. In our work, we proposed and validated a new methodology that allows to quantify the strength of the bias as well as attribute it to specific variables.

First, we generate pairs of equally good alternatives, and ask the subject to indicate strong/weak preference for one of the alternatives, answering "no preference" is also possible. These alternatives are described by observable quantitative proxy variables for cognitive biases and heuristics. For our study of plausibility, we used the proxies summarized in Table 5.1 on page 66. In the last step, we analysed the effect of individual variables controlling for the effect of other variables using semipartial rank correlation (Kendall's $\tau$). Correlation coefficient significantly different from zero between given proxy variable and elicited preference indicates an evidence for bias or heuristic linked to that proxy variable.

To the best of our knowledge, our contribution is the first methodology for measuring quantitative impact of cognitive biases on interpretation of machine learning results. The limitations of the proposed methodology are summarized in Section 12.2.

### 11.2.2. Cognitive Biases Supported by our Empirical Results

The main results based on the empirical experiments performed:

- *Shorter rules are not considered as more plausible.* We obtained positive correlation between rule length and plausibility on some datasets. We have not observed statistically significant negative correlation between plausibility and rule length on any of the datasets.
- *Misunderstanding of "and".* Our results support the conjecture that misunderstanding of "and" connective in inductively learned rules affects interpretation of rule learning results.
- *Insensitivity to sample size effect.* Our results show that when both confidence and support are stated, confidence positively affects plausibility and support is ignored.

We obtained also some preliminary results in support of the following propositions:

- *Weak evidence effect*
- *Availability heuristic*
- *Disjunction fallacy (preference for specificity)*

We have not identified any prior work that would do purposeful empirical investigation of the effect of cognitive biases on interpretation of machine learning results.

### 11.2.3. Contributions of Experiments with Linda and its Variations

So called Linda problem is a seminal experiment performed in cognitive science for demonstration of conjunctive fallacy.

In addition to replication of Linda, we performed two other experiments with the modifications of the original problem. Our contribution are the following findings:

- The fallacy rate of 68% that we obtained for the Linda problem using the CrowdFlower crowdsoucing platform is very close to earlier replication on the Amazon Mechanical Turk platform (fallacy rate of 72% in Paolacci et al. [2010]). This supports the choice of CrowdFlower for execution of our main empirical experiments.
- We obtained convincing experimental evidence showing that negation is mostly semantically interpreted and suppresses the application of the representativeness heuristic.
- Number of earlier results showed that in presence of alternative "$B \wedge F$", alternative "$B$" is interpreted as "$B \wedge \neg F$". Our results confirm that a similar pragmatic interpretation is applied to alternative "$B$" in presence of alternative "$B \wedge F$ *is unknown*". This may affect understanding of rule learning results, since it is common for discovered rules to contain rules containing literals with unknown value.

## 11.3. Plausibility Model and Recommendations (Chapter 6)

The first part of the thesis culminates with a proposal for a qualitative model for assessing plausibility of rules. The model is based on our analysis of literature in machine learning and cognitive science, empirical results, as well as on our prior experience with designing rule learning algorithms and applications.

Our literature review identified twenty cognitive biases and heuristics that have the potential to distort the understanding of inductively learned rules. Application of prior empirical results obtained in cognitive science allowed us to propose several methods that could be effective in suppressing these cognitive phenomena when machine learning models are interpreted. These proposals are in full covered in Chapter 4 with Table 4.1 on page 63 providing a quick overview, here we summarize the most important contributions.

### 11.3.1. Visual Model of Effect of Cognitive Biases on Rule Interpretability

The qualitative model that we designed is depicted on page 112. Figure 6.1A of the model shows that the effect of the individual literals in the rule largely depends on the domain knowledge of the person inspecting the model. In contrast, the way the contribution of the literals is aggregated into a final plausibility score in Figure 6.1B depends on the general information processing style and preferences of the person. While domain knowledge may be difficult to change, systematic errors in reasoning can be often avoided. One example is making the person aware of the fact that low rule support influences the reliability of the rule confidence estimate.

To our knowledge, our contribution is the first model for describing the effect of cognitive biases on interpretation of rules discovered from data.

### 11.3.2. Practical Recommendations for Rule Learning Software

Our literature review identified twenty cognitive biases and heuristics that have the potential to distort the understanding of inductively learned rules. Application of prior empirical results obtained in cognitive science allowed us to propose several methods that could be effective in suppressing these cognitive phenomena when machine learning models are interpreted.

As follows from the survey of machine learning literature presented in Chapter 2, we contributed the first set of guidelines for reflecting cognitive biases in the design of machine learning algorithms and software.

## 11.4. Software Framework (Chapters 7 - 10)

### 11.4.1. First non-fuzzy ARC Approach for Numerical Data

The presented framework ameliorates one of the major drawbacks of association rules, the adherence of rules comprising the classifier to the multidimensional grid created by discretization of numerical attributes. The novel aspect in MARC (QCBA) compared to other CBA-inspired association rule classification algorithms is that our framework reverts to the original attribute space to "edit" the discovered association rules, extending the scope of rule literals. To the best of our knowledge, all other association rule classification algorithms treat the rules discovered in their rule learning phase as atomic.

All previous adaptations of association rule classification for numerical data known to the author were fuzzy approaches. For example, the state-of-the-art FARC-HD association rule classifier outputs rules with fuzzy regions. This makes FARC-HD rules less comprehensible than the crisp rules output by CBA. MARC (QCBA) is, to the author's knowledge, the first non-fuzzy association rule classification algorithm supporting quantitative attributes. MARC (QCBA) reuses some of the central concepts in CBA, such as data coverage pruning, but introduces several new optimization and pruning steps (especially trimming, extension, default rule overlap pruning). While similar algorithms may have been used in other symbolic learning algorithms, their use in the context of association rule learning and classification is – as to the author's knowledge – novel. MARC (QCBA) design avoids introduction of new data-specific thresholds for the user to set or optimize, which constitutes a certain advancement over previous quantitative association rule learning approaches such as QuantMiner or NAR-Discovery. MARC (QCBA) does, however, contain several parameters that can be changed to speed up model building.

### 11.4.2. Smaller CBA models

While CBA [Liu et al., 1998] is the first association rule classification approach, it is according to our review still the best association rule-based classification algorithm what

concerns balance between comprehensibility of the model, predictive power and scalability. Numerous enhancements to CBA have been proposed since the seminal paper of Liu et al. [1998]. According to our review in Chapter 7, the modifications in all the succeeding association rule classification approaches negatively affect comprehensibility of the resulting rule-based model, yielding none or very small improvement in accuracy.

Our conclusions suggest that reduction of the size of the model presented to the person can help mitigate most biases, especially those linked to the representativeness heuristic. Each additional information in the model can trigger a range of biases, it is therefore essential to keep the models concise. To achieve this, our framework postprocesses CBA models in order to reduce their size.

Benchmark of our MARC (QCBA) approach on 22 UCI datasets shows average 53% decrease in the total size of the postprocessed CBA model as measured by the total number of conditions in all rules. Model accuracy remains on the same level as for CBA.

# 12. Limitations and Future Work

This chapter presents limitations of our approach and an outlook for future work.

**Chapter organization.** The discussion is organized according to chapters of the thesis. Section 12.1 covers our review and analysis of cognitive biases, Section 12.2 their empirical analysis and Section 12.3 the final plausibility model. Finally, Section 12.4 presents the proposed software framework including the experiments performed.

## 12.1. Chapters 2,4: Review and Analysis of Cognitive Biases

### 12.1.1. Limited Backing For Some Debiasing Techniques

We aimed to validate whether cognitive biases affect interpretation of machine learning models and propose remedies if they do. Since this field is untapped from the machine learning perspective, we tried to approach this problem holistically. Our work yielded a number of partial contributions, rather than a single profound result. We mapped applicable cognitive biases, identified prior works on their suppression and proposed how these could be transfered to machine learning. All the shortcomings of human judgment pertaining to interpretation of inductively learned rules that we have identified are based on empirical cognitive science research. To relate them to machine learning, for each bias we had to provide a justification of how the bias would relate to machine learning. Due to absence of applicable prior research, this justification is subjective and mostly based on prior experience of the author in the field of machine learning.

Also, due to paucity of prior research, there is a lack in understanding between computer scientists of how human ways of thinking and judgment can affect interpretation of machine learning results. We felt compelled to help address this by providing an actionable result from our analysis – a list of "debiasing techniques". Those recommendations in this list that are drawn based on very limited evidence are marked as so.

### 12.1.2. Incorporating Additional Biases

There are about 24 cognitive biases covered in the authoritative overview of cognitive biases by Pohl [2017] and even 51 different biases are covered by Evans et al. [2007]. While doing the initial selection of cognitive biases to study we tried to identify those most salient for machine learning research. This is the reason why we included the *weak evidence effect*, which has been discovered only recently and is not yet included into the latest edition of Cognitive Illusions [Pohl, 2017]. In the end, our review focused on a selection of 20

cognitive biases (effects, illusions). Future work might focus on expanding the review with additional relevant biases, such as labelling and overshadowing effects [Pohl, 2017, page 373].

### 12.1.3. Applicability of Results on Wason's 2-4-6 problem

In order to study semantic comprehension of models, we relied on research performed in cognitive science over last five decades. According to our review, the results obtained in cognitive science have only exceptionally been integrated or aligned with research done in machine learning. As our review also showed, there is a number of interesting and applicable results. Remarkably, since 1960 there is a consistent line of work done by psychologists on the problem of studying cognitive processes related to rule induction, which is centred around the so called *Wason's 2-4-6 problem*.

To our knowledge[1], cognitive science research on rule induction in humans has been so far completely unnoticed in the rule learning subfield of machine learning. It was out of the scope of the objectives of this thesis to perform analysis of the significance of results obtained for the Wason's 2-4-6 problem for rule learning, nevertheless we belief such investigation could bring interesting insights for cognitively-inspired design of rule learning algorithms.

## 12.2. Chapter 5: Empirical Analysis – Crowdsourcing Experiments

### 12.2.1. Incomplete Explanation of Higher Plausibility of Longer Rules

Our results suggest that whether plausibility relates to rule length depends on the characteristics of the dataset. Misunderstanding of "and" seems to affect plausibility on all datasets involved in our research, generally increasing preference for longer rules. However, the data indicate that there are also other factors – the individual biases and heuristics – that also mostly favour the longer rule. It should be noted that we cannot rule out that the drop in correlation between our V1 and V2 instructions (for Experiment 1) was not caused by other effect than misunderstanding of "and". This other confounding effect can be attributed to the inclusion of intersection test questions in the V1 instructions. It is a matter of further independent verification to exclude other causes.

While we have not observed higher preference for longer rules on all datasets, on the Mushroom dataset we did – even after misunderstanding of "and" was eliminated by using V1 instructions. We were unable to explain higher preference for longer rules on the Mushroom dataset by cognitive biases for which we had proxies available. There might be many possible causes, including:

- High variance in attribute and literal relevance values, since these were computed from a small number of responses.

[1] Based on our analysis of cited reference search in Google Scholar for [Wason, 1960].

- Restriction of our analysis to only several biases.
- Not robust enough estimates of literal and attribute relevance as these were computed from relatively small samples of responses.
- Lack of account for the varying level of domain knowledge that subjects possessed in relation to the datasets.

Our conjecture is that the last point is of particular relevance, since indeed from the analysis of textual answers we observed that the subjects had the required domain knowledge for LOD datasets but not for the Mushroom dataset. Such explanation would also be congruent with observation reported by Allahyari and Lavesson [2011] that preference for more complex models depends on the availability of domain knowledge.

### 12.2.2. Evaluating Debiasing Techniques

The empirical validation has primarily focused on determining whether selected cognitive biases demonstrate when rule learning results are interpreted. Out of the debiasing techniques proposed in Chapter 4 only misunderstanding of "and" was evaluated. It is a matter of future work to evaluate remaining proposed debiasing techniques directly on rule models.

### 12.2.3. Limited Strength of Evidence

Chapter 5 empirically investigated a number of cognitive biases. For one bias, our experiments provided unspurious result — the insensitivity to sample size effect. For several biases – availability heuristic, weak evidence effect and disjunction fallacy we collected some, yet limited, evidence. In these cases, additional research is required to confirm our hypothesis that they affect interpretation of rule learning results. Finally, the empirical data that we collected for the mere exposure effect do not support the hypothesis that it affects rule learning results.

The experiment with negation that we performed by manipulating the Linda instructions had a modest and predictable outcome of verifying that negation will inhibit representativeness heuristic. Other effects of negation remain to be investigated. Of particular importance in the rule learning context is the empirical result obtained by Pratto and John [2005] that negation increases attention. To which direction and to what extend negation would effect plausibility of rules is a viable direction of future work.

While we obtained convincing results for the information bias on the Linda problem, the analysis of the textual responses showed that the results are partly affected by the answer option with missing information affecting understanding of other options. As an alternative to the analysis of textual answers, which poses methodological challenges, we suggest that any future variation on this experiment directly elicits confidence judgments.

Additionally, what would be interesting to investigate is the effect of ambiguity aversion. According to this bias, missing information should result in decreased plausibility. While our experiment was not designed to evaluate ambiguity aversion, and the results do not

suggest that the ambiguity aversion occurred at larger scale, it is conceivable that in a different setting literal with missing information will trigger the ambiguity aversion.

### 12.2.4. Transferability of our Crowdsourced Results to other Populations

According to our assumptions the observed deviation from normatively correct judgments should be attributed to specific cognitive biases and effects. Both the degree of deviation from normatively correct judgments as well as the distribution of reasons (inferences) that lead to the judgments are specific to our cohort of crowdsourcing workers.

Most data scientists have a university degree; therefore the traditional sample of university students might be representative of this occupation. According to results presented in Paolacci et al. [2010] crowdsourcing users are not less able to handle quantitative tasks than university students. Most importantly, this study found the fallacy rate for Linda problem to not differ strongly between crowdsourcing and the standard student cohort. It follows that there is no evidence that our results should not be applicable for the data science occupation.

The *European Union's new General Data Protection Regulation*, which is proposed to take effect as law across the EU in 2018, which will effectively create a "right to explanation" that will allow the users of services relying on machine learning to request explanation of decisions that were made about them [Goodman and Flaxman, 2016]. The second group for which our results could be relevant is thus general computer literate population in Europe. We were unable to find any research that would analyze the differences between crowdsoucing and general population in terms of cognitive abilities in quantitative tasks and fallacy rates of cognitive biases.

### 12.2.5. Learning Within Crowdsourcing Task

All subjects had to pass the quiz ensuring basic understanding of the task before they started working on the main questions within our Experiment 1. The crowdsourcing platform also put hidden test questions within the "work mode" to ensure high consistency of the answers. Given these two layers of test questions, there was relatively low space for learning within the task. We therefore decided not to perform the analysis of the change in rates of errors over time. This means that we do not expect learning to notably affect error rates, not that learning did not occur. It is possible that subjects who decided to provide judgment for a higher number of rule pairs in Experiment 1 exhibited a different response pattern in terms of the preference for longer rules in their final responses as opposed to their initial ones. As a recommendation for further experiments we would suggest that the maximum number of preference judgments elicited from a particular subject is decreased substantially, possible even to one.

## 12.3. Chapter 6: Plausibility Model

Chapter 6 introduced a qualitative model of plausibility of rules, which provides a possible explanation of cognitive processes that are triggered and aggregated when humans assess plausibility of rules discovered from data.

### 12.3.1. Strength of Evidence

The purpose of our qualitative model is to empower machine learning practitioners with a schematic view of various biases and their interaction. While each node in the model is justified, the amount and strength of evidence vary. While some nodes have certain statistical backing, other nodes are based on a small number of textual responses. Nevertheless, we believe that the qualitative model can be used as a useful heuristic that can warn machine learning practitioners against possible misinterpretations that rule-based models are liable to.

### 12.3.2. Expanding to Quantitative Model

The qualitative model was designed specifically for evaluating plausibility of rules, other types of knowledge representations are not considered. What the qualitative model also does not provide is a quantitative prediction of plausibility of a specific rule.

As for future work, it is conceivable to expand the model to handle more general decision-making problems. Another direction is further elaboration of the existing model in terms of transition to a *quantitative* model, which would be able to numerically predict plausibility level within certain interval.

## 12.4. Chapters 9, 10: Software Framework

### 12.4.1. Contesting Views for "Smaller is More Comprehensible"

The elementary assumption for our enhancements of CBA models is that *smaller models are more comprehensible*. As our review in Chapter 2 shows this view is largely accepted in the machine learning community. However, there are also several works that contest this. In particular, there is the notion of *characteristic rules*, which is used in descriptive data mining for *concept characterization*. A characteristic rule should provide "concise and succinct" summary of a concept [Han et al., 2012, p. 16, 166]. In contrast, the evaluation metrics of classification algorithms, such as CBA or MARC (QCBA), are concerned with how good the rules are in terms of *concept comparison*.

While any representative empirical study showing that longer rules are more understandable than shorter rules has not yet been to our knowledge done, Stecher et al. [2016] in article entitled "Shorter Rules Are Better, Aren't They?" argue that "longer rules may in many cases be more understandable than shorter rules" and they back up their assertion with several examples. Gabriel et al. [2014] and Stecher et al. [2016] present algorithms for learning classification models composed of characteristic rules. The implications of this

line of research for our work on making association rule classification models smaller is that by *decreasing the average rule length* in MARC (QCBA) models, the rules in the models may become *less comprehensible* in terms of their ability to characterize the concepts they predict.

### 12.4.2. Future Work on Algorithmic Framework

The most imminent future work is improvement of the proposed algorithms in terms of scalability. The evaluation of the runtime indicates that the "extension" algorithm in MARC (QCBA) can be slow on datasets containing attributes with many distinct values. Improvements can include incorporation of the pessimistic pruning, using the M2 version of data coverage pruning proposed in Liu et al. [1998] instead of the M1 version and optimization of the extension algorithm, which is according to the results of the runtime benchmark the biggest bottleneck on some datasets.

Currently, the framework processes only numerical attributes, however, it would be relatively straightforward to extend it to ordinal attributes. Eckhardt [2010] described *representants* algorithm for ordering the domain of a nominal attribute based on training data. Another related algorithm for performing this type of attribute transformation is the Value Difference Metric, described e.g. by Wilson and Martinez [1996]. Experiments performed with the UTA method reported in [Eckhardt and Kliegr, 2012, Eckhardt, 2010] suggest that preprocessing with the *representants* algorithm could be effective.

The core optimization in MARC is rule extension. Analogously, one could perform rule shrinkage, which would remove values from the value bins if it would improve accuracy. Another promising area of future development would be using Bayesian confirmation measures instead of confidence and support as the interestingness (rule quality) measures as proposed in Brzezinski et al. [2016]. This could improve descriptive qualities of the resulting classifier, since Bayesian confirmation measures have been shown to have number of desirable properties to this effect [Brzezinski et al., 2016].

### 12.4.3. Expanding Experimental Evaluation

The empirical validation of the proposed MARC (QCBA) framework has been performed on 22 standard datasets. The purpose of this benchmark was to show that MARC (QCBA) provides and improvement over CBA, which this validation, in our opinion, achieved. For practical use cases, it might be desirable to extend this benchmark so that other symbolic algorithms are included. Such work would be particularly useful if it is complemented by analysis of reasons why one algorithm performs on a given dataset better than another one. This could help deciding which algorithm is most suitable for given type of data.

Another viable line of research might involve an empirical study focused on evaluation of comprehensibility of the learned models. Are indeed the more succinct models learned by MARC (QCBA) more comprehensible than those created by CBA? Such research could, with some adaptations, reuse methodology proposed for evaluation of comprehensibility of decision trees in Piltaver et al. [2016].

# Appendix

This thesis is accompanied by the following software and data repositories.

- `https://github.com/kliegr/rule-length-project`. Data and software related to crowdsourcing experiments reported in Chapter 5. Release version 1.0.
- `https://github.com/kliegr/arc`. The arc R package with M1 implementation of the CBA rule learner and several enhancements. Also available at CRAN as `https://cran.r-project.org/web/packages/arc/`. Release version 1.1.2.
- `https://github.com/kliegr/qcba` MARC (QCBA) framework installable as R package qCBA. Release version 0.2.
- `https://github.com/kliegr/arcbench`. Benchmark suite for the MARC (QCBA) framework used for evaluations reported in Chapter 10. Release version 1.1.2.

Additional details regarding the arc and qCBA packages follow.

### arc and qCBA R Packages

This package uses a discretization algorithm based on the minimum description length principle (MDLP) [Fayyad and Irani, 1993] as implemented in the R's *discretize* package [Kim, 2012]. All numeric explanatory attributes with three or more distinct values are by default subject to discretization.

Association rule learning step in CBA is handled by the C implementation of Apriori [Agrawal and Srikant, 1994], which is accessed via *arules* package [Hahsler et al., 2011]. The package implements the M1 version of the CBA-CB algorithm [Liu et al., 1998]. The pruning steps in M1 were implemented in a novel way using multiplication of sparse matrices exposed by the *arules* package. The computationally intensive operations are performed using C code in the Matrix package [Bates and Maechler, 2017]. Out of the two pruning algorithms (data coverage pruning and pessimistic pruning) described in Liu et al. [1998], the *arc* package implements only the data coverage pruning. According to evaluations presented in Liu et al. [1998] and the discussion in Section 10.4 pessimistic pruning provides only limited benefits.

The *arc* package can be used as a standalone implementation of the CBA algorithm. The package can be easily integrated to R workflows, because it supports the standard `predict()` interface. The package documentation is available in Kliegr [2016] and additional information is available at `https://github.com/kliegr/arc`.

MARC (QCBA) method is implemented in Java 8 in the *qCBA* R package. This provides a wrapper for the R ecosystem and integrates this package with the *arc* package.

# Bibliography


Charu C Aggarwal and Jiawei Han. *Frequent pattern mining*. Springer, 2014.

Rakesh Agrawal and Ramakrishnan Srikant. Fast algorithms for mining association rules in large databases. In *Proceedings of the 20th International Conference on Very Large Data Bases*, VLDB '94, pages 487–499, San Francisco, CA, USA, 1994. Morgan Kaufmann Publishers Inc. ISBN 1-55860-153-8.

Rakesh Agrawal, Tomasz Imielinski, and Arun N. Swami. Mining association rules between sets of items in large databases. In *SIGMOD*, pages 207–216. ACM Press, 1993.

Nabil I Al-Najjar and Jonathan Weinstein. The ambiguity aversion literature: a critical assessment. *Economics and Philosophy*, 25(03):249–284, 2009.

Dolores Albarracín and Amy L Mitchell. The role of defensive confidence in preference for proattitudinal information: How believing that one is strong can sometimes be a defensive weakness. *Personality and Social Psychology Bulletin*, 30(12):1565–1584, 2004.

Jesús Alcala-Fdez, Rafael Alcala, and Francisco Herrera. A fuzzy association rule-based classification model for high-dimensional problems with genetic rule selection and lateral tuning. *IEEE Transactions on Fuzzy Systems*, 19(5):857–872, 2011.

Hiva Allahyari and Niklas Lavesson. User-oriented assessment of classification model understandability. In *11th Scandinavian Conference on Artificial Intelligence*. IOS Press, 2011.

Paulo J Azevedo and Alípio Mário Jorge. Comparing rule measures for predictive association rules. In *Ecml*, volume 7, pages 510–517. Springer, 2007.

Marco Ballin, Riccardo Carbini, Maria Francesca Loporcaro, Massimo Lori, Roberto Moro, Valeria Olivieri, and Mauro Scanu. The use of information from experts for agricultural official statistics. In *European Conference on Quality in Official Statistics (Q2008)*, 2008.

Maya Bar-Hillel. The base-rate fallacy in probability judgments. *Acta Psychologica*, 44(3): 211–233, 1980.

Maya Bar-Hillel. Commentary on Wolford, Taylor, and Beck: The conjunction fallacy? *Memory & cognition*, 19(4):412–414, 1991.

Maya Bar-Hillel and Efrat Neter. How alike is it versus how likely is it: A disjunction fallacy in probability judgments. *Journal of Personality and Social Psychology*, 65(6): 1119, 1993.

Jonathan Baron, Jane Beattie, and John C Hershey. Heuristics and biases in diagnostic reasoning: II. congruence, information, and certainty. *Organizational Behavior and Human Decision Processes*, 42(1):88–110, 1988.

Douglas Bates and Martin Maechler. *Matrix: Sparse and Dense Matrix Classes and Methods*, 2017. URL `https://CRAN.R-project.org/package=Matrix`. R package version 1.2-8.

C Philip Beaman, Rachel McCloy, and Philip T Smith. When does ignorance make us smart? additional factors guiding heuristic inference. In *Proceedings of the Cognitive Science Society*, volume 28, 2006.

Adrien Bibal and Benoît Frénay. Interpretability of machine learning models and representations: an introduction. In *Proceedings of the 24th European Symposium on Artificial Neural Networks (ESANN)*, pages 77–82, 2016.

Samuel D Bond, Kurt A Carlson, Margaret G Meloy, J Edward Russo, and Robin J Tanner. Information distortion in the evaluation of a single option. *Organizational Behavior and Human Decision Processes*, 102(2):240–254, 2007.

Robert F Bornstein. Exposure and affect: overview and meta-analysis of research, 1968–1987. *Psychological Bulletin*, 2(106):265–289, 1989.

Harriet R Brown, Peter Zeidman, Peter Smittenaar, Rick A Adams, Fiona McNab, Robb B Rutledge, and Raymond J Dolan. Crowdsourcing for cognitive science–the utility of smartphones. *PloS one*, 9(7):e100662, 2014.

Dariusz Brzezinski, Zbigniew Grudziński, and Izabela Szczęch. Bayesian confirmation measures in rule-based classification. In *International Workshop on New Frontiers in Mining Complex Patterns*, pages 39–53. Springer, 2016.

Colin Camerer and Martin Weber. Recent developments in modeling preferences: Uncertainty and ambiguity. *Journal of Risk and Uncertainty*, 5(4):325–370, 1992. ISSN 1573-0476.

Gary Charness, Edi Karni, and Dan Levin. On the conjunction fallacy in probability judgment: New experimental evidence regarding Linda. *Games and Economic Behavior*, 68(2):551 – 556, 2010. ISSN 0899-8256.

Zuoliang Chen and Guoqing Chen. Building an associative classifier based on fuzzy association rules. *International Journal of Computational Intelligence Systems*, 1(3):262–273, 2008.

Soo Hong Chew, Richard P. Ebstein, and Songfa Zhong. Ambiguity aversion and familiarity bias: Evidence from behavioral and gene association studies. *Journal of Risk and Uncertainty*, 44(1):1–18, 2012. ISSN 1573-0476.

William AV Clark and Karen L Avery. The effects of data aggregation in statistical analysis. *Geographical Analysis*, 8(4):428–438, 1976.

Joshua D Clinton. Proxy variable. *The SAGE Encyclopedia of Social Science Research Methods. PLACE: Sage*, 2004.

William W. Cohen. Fast effective rule induction. In *Proceedings of the Twelfth International Conference on Machine Learning*, pages 115–123. Morgan Kaufmann, 1995.

Matthew JC Crump, John V McDonnell, and Todd M Gureckis. Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research. *PloS one*, 8(3):e57410, 2013.

Shawn P Curley, Stephen A Eraker, and J Frank Yates. An investigation of patient's reactions to therapeutic uncertainty. *Medical Decision Making*, 4(4):501–511, 1984.

Janez Demšar. Statistical comparisons of classifiers over multiple data sets. *J. Mach. Learn. Res.*, 7:1–30, December 2006. ISSN 1532-4435. URL `http://dl.acm.org/citation.cfm?id=1248547.1248548`.

Pedro Domingos. The role of Occam's razor in knowledge discovery. *Data mining and knowledge discovery*, 3(4):409–425, 1999.

Alan Eckhardt. *Induction of User Preferences For Semantic Web*. Mathematic-Physics Faculty of the Charles University, 2010. PhD Dissertation.

Alan Eckhardt and Tomáš Kliegr. Preprocessing algorithm for handling non-monotone attributes in the UTA method. In *Preference Learning: Problems and Applications in AI (PL-12)*, 2012.

Stephen E Edgell, J Harbison, William P Neace, Irwin D Nahinsky, and A Scott Lajoie. What is learned from experience in a probabilistic environment? *Journal of Behavioral Decision Making*, 17(3):213–229, 2004.

M. Elkano, M. Galar, J. A. Sanz, A. Fernández, E. Barrenechea, F. Herrera, and H. Bustince. Enhancing multiclass classification in farc-hd fuzzy classifier: On the synergy between $n$-dimensional overlap functions and decomposition strategies. *IEEE Transactions on Fuzzy Systems*, 23(5):1562–1580, Oct 2015. ISSN 1063-6706. doi: 10.1109/TFUZZ.2014.2370677.

Daniel Ellsberg. Risk, ambiguity, and the Savage axioms. *The Quarterly Journal of Economics*, 75(4):643–669, 1961. ISSN 00335533, 15314650. URL `http://www.jstor.org/stable/1884324`.

Tapio Elomaa. In defense of C4.5: Notes on learning one-level decision trees. In *Proceedings of the 11th International conference on machine learning*, volume 254, pages 62–69. Morgan Kaufmann, 1994.

Jonathan St BT Evans. *Bias in human reasoning: Causes and consequences*. Lawrence Erlbaum Associates, Inc, 1989.

Jonathan St BT Evans et al. *Hypothetical thinking: Dual processes in reasoning and judgement*, volume 3. Psychology Press, 2007.

Edmund Fantino, James Kulik, Stephanie Stolarz-Fantino, and William Wright. The conjunction fallacy: A test of averaging hypotheses. *Psychonomic Bulletin & Review*, 4(1): 96–101, 1997.

U. M. Fayyad and K. B. Irani. Multi-Interval Discretization of Continuous-Valued Attributes for Classification Learning. In *13th International Joint Conference on Uncertainly in Artificial Intelligence (IJCAI93)*, pages 1022–1029, 1993.

Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. The KDD process for extracting useful knowledge from volumes of data. *Commun. ACM*, 39(11):27–34, November 1996. ISSN 0001-0782. doi: 10.1145/240455.240464.

Ad J Feelders. Prior knowledge in economic applications of data mining. In *European Conference on Principles of Data Mining and Knowledge Discovery*, pages 395–400. Springer, 2000.

Christiane Fellbaum, editor. *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*. The MIT Press, May 1998. ISBN 026206197X.

Philip M Fernbach, Adam Darlow, and Steven A Sloman. When good evidence goes bad: The weak evidence effect in judgment and decision-making. *Cognition*, 119(3):459–467, 2011.

Ronald A Fisher. The use of multiple measurements in taxonomic problems. *Annals of human genetics*, 7(2):179–188, 1936.

Susan T Fiske. Attention and weight in person perception: The impact of negative and extreme behavior. *Journal of personality and Social Psychology*, 38(6):889, 1980.

Alex A Freitas. Comprehensible classification models: a position paper. *ACM SIGKDD explorations newsletter*, 15(1):1–10, 2014.

Johannes Fürnkranz. Pruning algorithms for rule learning. *Machine Learning*, 27(2): 139–172, 1997.

Johannes Fürnkranz. Separate-and-conquer rule learning. *Artificial Intelligence Review*, 13(1):3–54, 1999.

Johannes Fürnkranz and Tomáš Kliegr. A brief overview of rule learning. In *International Symposium on Rules and Rule Markup Languages for the Semantic Web*, pages 54–69. Springer, 2015.

Johannes Fürnkranz, Dragan Gamberger, and Nada Lavrač. *Foundations of Rule Learning*. Springer-Verlag, 2012. ISBN 978-3-540-75196-0.

Alexander Gabriel, Heiko Paulheim, and Frederik Janssen. Learning semantically coherent rules. In *Proceedings of the 1st International Workshop on Interactions between Data Mining and Natural Language Processing co-located with The European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (DMNLP@ PKDD/ECML)*, pages 49–63, Nancy, France, 2014. CEUR Workshop Proceedings.

Salvador García, Alberto Fernández, Julián Luengo, and Francisco Herrera. A study of statistical techniques and performance measures for genetics-based machine learning: accuracy and interpretability. *Soft Computing*, 13(10):959–977, 2009.

Andrew B Geier, Paul Rozin, and Gheorghe Doros. Unit bias a new heuristic that helps explain the effect of portion size on food intake. *Psychological Science*, 17(6):521–525, 2006.

Mohammad Ghaderi, Francisco Ruiz, and Núria Agell. A linear programming approach for learning non-monotonic additive value functions in multiple criteria decision aiding. *European Journal of Operational Research*, 259(3):1073–1084, 2017.

Jean D Gibbons and MG Kendall. Rank correlation methods. *Edward Arnold*, 1990.

Gerd Gigerenzer. On narrow norms and vague heuristics: A reply to Kahneman and Tversky. *Psychological Review*, pages 592–596, 1996.

Gerd Gigerenzer and Daniel G Goldstein. Reasoning the fast and frugal way: models of bounded rationality. *Psychological review*, 103(4):650, 1996.

Gerd Gigerenzer and Ulrich Hoffrage. How to improve Bayesian reasoning without instruction: frequency formats. *Psychological review*, 102(4):684, 1995.

Gerd Gigerenzer and Ulrich Hoffrage. Overcoming difficulties in Bayesian reasoning: A reply to Lewis and Keren (1999) and Mellers and McGraw (1999). *Psychological Review*, (106):425–430, 1999.

Thomas Gilovich and Kenneth Savitsky. *Like goes with like: The role of representativeness in erroneous and pseudo-scientific beliefs*. Cambridge University Press, 2002.

Christophe Giraud-Carrier and Foster Provost. Toward a justification of meta-learning: Is the no free lunch theorem a show-stopper. In *Proceedings of the ICML-2005 Workshop on Meta-learning*, pages 12–19, 2005.

Bart Goethals and Mohammed J Zaki. FIMI'03: Workshop on frequent itemset mining implementations. In *3rd IEEE ICDM Workshop on Frequent Itemset Mining Implementations*, pages 1–13, 2003.

Daniel G Goldstein and Gerd Gigerenzer. The recognition heuristic: How ignorance makes us smart. In *Simple heuristics that make us smart*, pages 37–58. Oxford University Press, 1999.

Antonio González and Raúl Pérez. Selection of relevant features in a fuzzy genetic learning algorithm. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 31(3):417–425, 2001.

Bryce Goodman and Seth Flaxman. European union regulations on algorithmic decision-making and a" right to explanation". *arXiv preprint arXiv:1606.08813*, 2016.

Alison Gopnik and Joshua B Tenenbaum. Bayesian networks, Bayesian learning and cognitive development. *Developmental science*, 10(3):281–287, 2007.

Salvatore Greco, Vincent Mousseau, and Roman Slowinski. Inductive models of user preferences for semantic web. *European Journal of Operational Research*, (191):416–436, 2007.

David M Grether. Testing Bayes rule and the representativeness heuristic: Some experimental evidence. *Journal of Economic Behavior & Organization*, 17(1):31–57, 1992.

Dale Griffin and Amos Tversky. The weighing of evidence and the determinants of confidence. *Cognitive psychology*, 24(3):411–435, 1992.

Thomas L Griffiths, Nick Chater, Charles Kemp, Amy Perfors, and Joshua B Tenenbaum. Probabilistic models of cognition: Exploring representations and inductive biases. *Trends in cognitive sciences*, 14(8):357–364, 2010.

Michael Hahsler, Sudheer Chelluboina, Kurt Hornik, and Christian Buchta. The arules r-package ecosystem: analyzing interesting patterns from large transaction data sets. *Journal of Machine Learning Research*, 12(Jun):2021–2025, 2011.

Jiawei Han, Jian Pei, Yiwen Yin, and Runying Mao. Mining frequent patterns without candidate generation: A frequent-pattern tree approach. *Data Mining and Knowledge Discovery*, 8(1):53–87, January 2004. ISSN 1384-5810.

Jiawei Han, Jian Pei, and Micheline Kamber. *Data mining: concepts and techniques*. Elsevier, 3rd edition, 2012.

Ashleigh Harris. Dropping mechanical turk helps our customers get the best results, 2014. URL `https://www.crowdflower.com/crowdflower-drops-mechanical-turk-to-ensure-the-best-results-for-its-customers/`.

Martie G Haselton and Daniel Nettle. The paranoid optimist: An integrative evolutionary model of cognitive biases. *Personality and social psychology Review*, 10(1):47–66, 2006.

Lynn Hasher, David Goldstein, and Thomas Toppino. Frequency and the conference of referential validity. *Journal of Verbal Learning and Verbal Behavior*, 16(1):107–112, 1977.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA, 2001.

Ralph Hertwig and Gerd Gigerenzer. The "conjunction fallacy" revisited: How intelligent inferences look like reasoning errors. *Journal of Behavioral Decision Making*, 12(4): 275–305, 1999.

Ralph Hertwig, Gerd Gigerenzer, and Ulrich Hoffrage. The reiteration effect in hindsight bias. *Psychological Review*, 104(1):194, 1997.

Ralph Hertwig, Thorsten Pachur, and Stephanie Kurzenhäuser. Judgments of risk frequencies: tests of possible cognitive mechanisms. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 31(4):621, 2005.

Ralph Hertwig, Björn Benz, and Stefan Krauss. The conjunction fallacy and the many meanings of and. *Cognition*, 108(3):740–753, 2008.

Denis J Hilton. The social context of reasoning: Conversational inference and rational judgment. *Psychological Bulletin*, 118(2):248, 1995.

Yi-Chung Hu, Ruey-Shun Chen, and Gwo-Hshiung Tzeng. Finding fuzzy classification rules using data mining techniques. *Pattern Recognition Letters*, 24(1):509–519, 2003.

Jens Hühn and Eyke Hüllermeier. FURIA: an algorithm for unordered fuzzy rule induction. *Data Mining and Knowledge Discovery*, 19(3):293–319, 2009.

Johan Huysmans, Karel Dejaeger, Christophe Mues, Jan Vanthienen, and Bart Baesens. An empirical evaluation of the comprehensibility of decision table, tree and rule based predictive models. *Decision Support Systems*, 51(1):141–154, 2011.

Hisao Ishibuchi, Takashi Yamamoto, and Tomoharu Nakashima. Hybridization of fuzzy GBML approaches for pattern classification problems. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 35(2):359–365, 2005.

E. Jacquet-Lagreze and J. Siskos. Assessing a set of additive utility functions for multicriteria decision-making, the uta method. *European Journal of Operational Research*, 10 (2):151–164, June 1982.

Daniel Kahneman. A perspective on judgment and choice. *American Psychologist*, 58, 2003.

Daniel Kahneman and Amos Tversky. Subjective probability: A judgment of representativeness. *Cognitive psychology*, 3(3):430–454, 1972.

Daniel Kahneman and Amos Tversky. On the psychology of prediction. *Psychological Review*, 80(4):237 – 251, 1973. ISSN 0033-295X.

Daniel Kahneman and Amos Tversky. Prospect theory: An analysis of decision under risk. *Econometrica: Journal of the econometric society*, pages 263–291, 1979.

David E Kanouse and L Reid Hanson Jr. Negativity in evaluations. In *Attribution: Perceiving the causes of behavior*. Lawrence Erlbaum Associates, Inc, 1987.

John G Kemeny. The use of simplicity in induction. *The Philosophical Review*, 62(3): 391–408, 1953.

John Maynard Keynes. *A Treatise on Probability*. Macmillan & Co, 1922.

HyunJi Kim. *discretization: Data preprocessing, discretization for classification*, 2012. URL `https://CRAN.R-project.org/package=discretization`. R package version 1.0-1.

Seongho Kim. ppcor: An R package for a fast calculation to semi-partial correlation coefficients. *Communications for Statistical Applications and Methods*, 22(6):665–674, 2015.

Joshua Klayman and Young-Won Ha. Confirmation, disconfirmation, and information in hypothesis testing. *Psychological review*, 94(2):211, 1987.

Tomás Kliegr. UTA - NM: Explaining stated preferences with additive non-monotonic utility functions. In *Proceedings of the ECML'09 Preference Learning Workshop*, 2009.

Tomáš Kliegr, Jaroslav Kuchař, Davide Sottara, and Stanislav Vojíř. Learning business rules with association rule classifiers. In Antonis Bikakis, Paul Fodor, and Dumitru Roman, editors, *Rules on the Web. From Theory to Applications: 8th International Symposium, RuleML 2014, Co-located with the 21st European Conference on Artificial Intelligence, ECAI 2014, Prague, Czech Republic, August 18-20, 2014. Proceedings*, pages 236–250, Cham, 2014. Springer International Publishing. ISBN 978-3-319-09870-8.

Tomáš Kliegr. Linked hypernyms: Enriching DBpedia with targeted hypernym discovery. *Web Semantics: Science, Services and Agents on the World Wide Web*, 31:59 – 69, 2015. ISSN 1570-8268.

Tomáš Kliegr. *Association Rule Classification*, 2016. URL `https://CRAN.R-project.org/package=arc`. R package version 1.1.

Tomáš Kliegr and Jaroslav Kuchař. Orwellian Eye: Video recommendation with Microsoft Kinect. In *Conference on Prestigious Applications of Intelligent Systems (PAIS'14) collocated with European Conference on Artificial Intelligence (ECAI'14)*. IOS Press, August 2014.

Tomáš Kliegr and Jaroslav Kuchař. Benchmark of rule-based classifiers in the news recommendation task. In Josiane Mothe, Jacques Savoy, Jaap Kamps, Karen Pinel-Sauvagnat, Gareth J. F. Jones, Eric SanJuan, Linda Cappellato, and Nicola Ferro, editors, *Experimental IR Meets Multilinguality, Multimodality, and Interaction - 6th International Conference of the CLEF Association, CLEF 2015, Toulouse, France, September 8-11, 2015, Proceedings*, volume 9283 of *Lecture Notes in Computer Science*, pages 130–141. Springer, 2015. ISBN 978-3-319-24026-8.

Tomáš Kliegr, Vojtěch Svátek, Milan Šimunek, and Martin Ralbovský. Semantic analytical reports: A framework for post-processing of data mining results. *Journal of Intelligent Information Systems*, 37(3):371–395, 2011.

Tomáš Kliegr and Ondřej Zamazal. LHD 2.0: A text mining approach to typing entities in knowledge graphs. *Web Semantics: Science, Services and Agents on the World Wide Web*, 39(0), 2016. ISSN 1570-8268.

Jaroslav Kuchař and Tomáš Kliegr. GAIN: web service for user tracking and preference learning - a SMART TV use case. In *7th ACM Conference on Recommender Systems, RecSys '13, Hong Kong, China, October 12-16, 2013*, 2013.

Jaroslav Kuchař and Tomáš Kliegr. InBeat: News recommender system as a service @ CLEF NEWSREEL'14. In *CEUR-WS Proceedings Vol-1180: CLEF-2014*, 2014.

Jaroslav Kuchař and Tomáš Kliegr. Inbeat: Javascript recommender system supporting sensor input and linked data. *Knowledge-Based Systems*, 135:40 – 43, 2017. ISSN 0950-7051. doi: https://doi.org/10.1016/j.knosys.2017.07.026. URL `http://www.sciencedirect.com/science/article/pii/S0950705117303428`.

Edwin Kuh and John R Meyer. Correlation and regression estimates when the data are ratios. *Econometric, Journal of the Econometric Society*, pages 400–416, 1955.

Ziva Kunda. *Social cognition: Making sense of people*. MIT press, 1999.

Himabindu Lakkaraju, Stephen H. Bach, and Jure Leskovec. Interpretable decision sets: A joint framework for description and prediction. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 1675–1684, New York, NY, USA, 2016. ACM. ISBN 978-1-4503-4232-2.

Nada Lavrač. Data mining in medicine: Selected techniques and applications. *Artificial Intelligence in Medicine*, 16:3–23, 1998.

Julien Leroy, Francois Rocca, Matei Mancas, Radhwan Ben Madhkour, Fabien c Grisard, Tomáš Kliegr, Jaroslav Kuchař, Jakub Vit, Ivan Pirner, and Petr Zimmermann. Innovative and creative developments in multimodal interaction systems. In Yves Rybarczyk, Tiago Cardoso, and Joao Rosas, editors, *KINterestTV - Can we measure in a non-invasive way, the interest that a user has in front of his television displaying its content?* Springer, 2014.

Wenmin Li, Jiawei Han, and Jian Pei. CMAR: Accurate and efficient classification based on multiple class-association rules. In *ICDM '01 Proceedings*, pages 369–376, Washington, DC, USA, 2001a. IEEE. ISBN 0-7695-1119-8.

Wenmin Li, Jiawei Han, and Jian Pei. Cmar: Accurate and efficient classification based on multiple class-association rules. In *Proceedings of the 2001 IEEE International Conference on Data Mining*, ICDM '01, pages 369–376, Washington, DC, USA, 2001b. IEEE Computer Society. ISBN 0-7695-1119-8.

Dekang Lin. An information-theoretic definition of similarity. In *Proceedings of the 15th International Conference on Machine Learning*, pages 296–304. Morgan Kaufmann, 1998.

Gary H Lincoff. *The Audubon society field guide to North American mushrooms*. Knopf, 1981.

Zachary Lipton. The mythos of model interpretability. *arXiv preprint arXiv:1606.03490*, 2016. Presented at 2016 ICML Workshop on Human Interpretability in Machine Learning (WHI 2016), New York, NY.

Bing Liu, Wynne Hsu, and Yiming Ma. Integrating classification and association rule mining. In *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining*, KDD'98, pages 80–86. AAAI Press, 1998.

Bing Liu, Yiming Ma, and Ching-Kian Wong. Classification using association rules: weaknesses and enhancements. *Data mining for scientific applications*, 591, 2001.

C. Lucchese, S. Orlando, and R. Perego. Parallel mining of frequent closed patterns: Harnessing modern computer architectures. In *Proc. of the 7th IEEE International Conference on Data Mining*, ICDM 2007, pages 242–251, 2007.

Claudio Lucchese. DCI Closed: A fast and memory efficient algorithm to mine frequent closed itemsets. In *Proc. of the IEEE ICDM 2004 Workshop on Frequent Itemset Mining Implementations (FIMI'04)*, 2004.

Eghbal G Mansoori, Mansoor J Zolghadri, and Seraj D Katebi. Sgerd: A steady-state genetic algorithm for extracting fuzzy classification rules from data. *IEEE Transactions on Fuzzy Systems*, 16(4):1061–1071, 2008.

David Martens, Jan Vanthienen, Wouter Verbeke, and Bart Baesens. Performance of classification models from a user perspective. *Decision Support Systems*, 51(4):782–793, 2011.

Kristy A Martire, Richard I Kemp, Ian Watkins, Malindi A Sayle, and Ben R Newell. The expression and interpretation of uncertain forensic science evidence: verbal equivalence, evidence strength, and the weak evidence effect. *Law and human behavior*, 37(3):197, 2013.

R Mata. Cognitive bias. *Encyclopedia of human behaviour*, 1:531–535, 2012.

Ryszard S Michalski. On the quasi-minimal solution of the general covering problem. In *Proceedings of the V International Symposium on Information Processing (FCIP 69)(Switching Circuits)*, pages 125–128, Yugoslavia, Bled, 1969.

Ryszard S Michalski. A theory and methodology of inductive learning. In *Machine learning*, pages 83–134. Springer, 1983.

George A Miller. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological review*, 63(2):81, 1956.

Tom M Mitchell. Machine learning. 1997. *Burr Ridge, IL: McGraw Hill*, 45:37, 1997.

Jennifer L Monahan, Sheila T Murphy, and Robert B Zajonc. Subliminal mere exposure: Specific, general, and diffuse effects. *Psychological Science*, 11(6):462–466, 2000.

Giuseppe Mosconi and Laura Macchi. The role of pragmatic rules in the conjunction fallacy. *Mind & Society*, 2(1):31–57, 2001.

Clifford R Mynatt, Michael E Doherty, and Ryan D Tweney. Confirmation bias in a simulated research environment: An experimental study of scientific inference. *The quarterly journal of experimental psychology*, 29(1):85–95, 1977.

Benjamin Negrevergne, Alexandre Termier, Marie-Christine Rousset, Jean-François Méhaut, and Takeaki Uno. Discovering closed frequent itemsets on multicore: Parallelizing computations and optimizing memory accesses. In *Proc. of the International Conference on High Performance Computing and Simulation*, HPCS 2010, pages 521–528, 2010.

Benjamin Negrevergne, Alexandre Termier, Marie-Christine Rousset, and Jean-François Méhaut. Para miner: a generic pattern mining algorithm for multi-core architectures. *Data Mining and Knowledge Discovery*, 28(3):593–633, 2014. ISSN 1384-5810.

Roger Newson. Parameters behind "nonparametric" statistics: Kendall's tau, Somers' d and median differences. *Stata Journal*, 2, 2002.

Raymond S Nickerson. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of general psychology*, 2(2):175, 1998.

Richard E Nisbett. *Rules for reasoning*. Psychology Press, 1993.

Hideki Ohira, Ward M Winton, and Makiko Oyama. Effects of stimulus valence on recognition memory and endogenous eyeblinks: Further evidence for positive-negative asymmetry. *Personality and Social Psychology Bulletin*, 24(9):986–993, 1998.

David Oleson, Alexander Sorokin, Greg P Laughlin, Vaughn Hester, John Le, and Lukas Biewald. Programmatic gold: Targeted and scalable quality assurance in crowdsourcing. *Human computation*, 11(11), 2011.

Fernando EB Otero and Alex A Freitas. Improving the interpretability of classification rules discovered by an ant colony algorithm. In *Proceedings of the 15th annual conference on Genetic and evolutionary computation*, pages 73–80. ACM, 2013.

Thorsten Pachur and Ralph Hertwig. On the psychology of the recognition heuristic: Retrieval primacy as a key determinant of its use. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 32(5):983, 2006.

Thorsten Pachur, Peter M Todd, Gerd Gigerenzer, Lael Schooler, and Daniel G Goldstein. The recognition heuristic: A review of theory and tests. *Frontiers in psychology*, 2:147, 2011.

Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The PageRank citation ranking: Bringing order to the web. Technical report, Stanford Digital Library Technologies Project, 1998.

Gabriele Paolacci and Jesse Chandler. Inside the Turk: Understanding Mechanical Turk as a participant pool. *Current Directions in Psychological Science*, 23(3):184–188, 2014.

Gabriele Paolacci, Jesse Chandler, and Panagiotis G Ipeirotis. Running experiments on Amazon Mechanical Turk. *Judgment and Decision Making*, 5(5), 2010.

Nicolas Pasquier, Yves Bastide, Rafik Taouil, and Lotfi Lakhal. Discovering frequent closed itemsets for association rules. In Catriel Beeri and Peter Buneman, editors, *Inf. Proc. of the 7th International Conference on Database Theory*, volume 1540 of *ICDT'99*, pages 398–416. Springer, 1999. ISBN 978-3-540-65452-0.

M Pazzani. Comprehensible knowledge discovery: gaining insight from data. In *First Federal Data Mining Conference and Exposition*, pages 73–82, Washington, DC, 1997.

M. J. Pazzani. Knowledge discovery from data? *IEEE Intelligent Systems and their Applications*, 15(2):10–12, March 2000. ISSN 1094-7167.

Rok Piltaver, Mitja Luštrek, Matjaž Gams, and Sanda Martinčić-Ipšić. Comprehensibility of classification trees – survey design validation. In *6th International Conference on Information Technologies and Information Society - ITIS2014*, 2014.

Rok Piltaver, Mitja Lustrek, Matjaz Gams, and Sanda Martincic-Ipsic. What makes classification trees comprehensible? *Expert Systems with Applications*, 62:333 – 346, 2016. ISSN 0957-4174.

Steven Pinker. *Words and rules: The ingredients of language*. Basic Books, 2015.

Scott Plous. *The psychology of judgment and decision making*. McGraw-Hill Book Company, 1993.

Rüdiger Pohl. *Cognitive illusions: A handbook on fallacies and biases in thinking, judgement and memory*. Psychology Press, 2004.

Rüdiger Pohl. *Cognitive illusions: A handbook on fallacies and biases in thinking, judgement and memory*. Psychology Press, 2017. 2nd ed.

Karl Popper. *The logic of scientific discovery*. Routledge, 2005.

Karl Raimund Popper. *Logik der Forschung: zur Erkenntnistheorie der moderner Naturwissenschaft*. Verlag von Julius Springer, 1935.

HR Post. Simplicity in scientific theories. *The British Journal for the Philosophy of Science*, 11(41):32–41, 1960.

Felicia Pratto and Oliver P John. Automatic vigilance: The attention-grabbing power of negative social information. *Social cognition: key readings*, 250, 2005.

J. Ross Quinlan. Induction of decision trees. *Machine learning*, 1(1):81–106, 1986.

J. Ross Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, 1993. ISBN 1-55860-238-0.

J Ross Quinlan. Improved use of continuous attributes in C4. 5. *Journal of artificial intelligence research*, 4:77–90, 1996.

Jan Rauch. *Considerations on Logical Calculi for Dealing with Knowledge in Data Mining*, pages 177–199. Springer Berlin Heidelberg, Berlin, Heidelberg, 2009. ISBN 978-3-642-02190-9.

Petar Ristoski, Gerben Klaas Dirk de Vries, and Heiko Paulheim. *A Collection of Benchmark Datasets for Systematic Evaluations of Machine Learning on the Semantic Web*, pages 186–194. Springer International Publishing, Cham, 2016. ISBN 978-3-319-46547-0. doi: 10.1007/978-3-319-46547-0_20. URL `https://doi.org/10.1007/978-3-319-46547-0_20`.

William S Robinson. Ecological correlations and the behavior of individuals. *American Sociological Review*, 15(3):351–337, 1950.

Gregory L Robinson-Riegler and Ward M Winton. The role of conscious recollection in recognition of affective material: Evidence for positive-negative asymmetry. *The Journal of General Psychology*, 123(2):93–104, 1996.

Sandrine Rossi, Jean Paul Caverni, and Vittorio Girotto. Hypothesis testing in a rule discovery problem: When a focused procedure is effective. *The Quarterly Journal of Experimental Psychology: Section A*, 54(1):263–267, 2001.

Paul Rozin and Edward B Royzman. Negativity bias, negativity dominance, and contagion. *Personality and social psychology review*, 5(4):296–320, 2001.

Cullen Schaffer. A conservation law for generalization performance. In *Proceedings of the 11th international conference on machine learning*, pages 259–265, 1994.

Tyler Schnoebelen and Victor Kuperman. Using Amazon Mechanical Turk for linguistic research. *Psihologija*, 43(4):441–464, 2010.

Norbert Schwarz, Herbert Bless, Fritz Strack, Gisela Klumpp, Helga Rittenauer-Schatka, and Annette Simons. Ease of retrieval as information: Another look at the availability heuristic. *Journal of Personality and Social psychology*, 61(2):195, 1991.

Sebastian Serfas. Cognitive biases in the capital investment context. In *Cognitive Biases in the Capital Investment Context*, pages 95–189. Springer, 2011.

Hanan Shteingart, Tal Neiman, and Yonatan Loewenstein. The role of first impression in operant learning. *Journal of Experimental Psychology: General*, 142(2):476, 2013.

Ashley Sides, Daniel Osherson, Nicolao Bonini, and Riccardo Viale. On the reality of the conjunction fallacy. *Memory & Cognition*, 30(2):191–198, 2002.

Roman Slowinski, Izabela Brzezinska, and Salvatore Greco. Application of bayesian confirmation measures for mining rules from support-confidence pareto-optimal set. *Artificial Intelligence and Soft Computing–ICAISC 2006*, pages 1018–1026, 2006.

Edward E Smith, Christopher Langston, and Richard E Nisbett. The case for rules in reasoning. *Cognitive science*, 16(1):1–40, 1992.

Keith E Stanovich. Distinguishing the reflective, algorithmic, and autonomous minds: Is it time for a tri-process theory. *In two minds: Dual processes and beyond*, pages 55–88, 2009.

Keith E Stanovich, Richard F West, and Maggie E Toplak. Myside bias, rational thinking, and intelligence. *Current Directions in Psychological Science*, 22(4):259–264, 2013.

Julius Stecher, Frederik Janssen, and Johannes Fürnkranz. Separating rule refinement and rule selection heuristics in inductive rule learning. In *Machine Learning and Knowledge Discovery in Databases*, pages 114–129. Springer Berlin Heidelberg, 2014.

Julius Stecher, Frederik Janssen, and Johannes Fürnkranz. Shorter rules are better, aren't they? In *International Conference on Discovery Science*, pages 279–294. Springer, 2016.

Stephanie Stolarz-Fantino, Edmund Fantino, and James Kulik. The conjunction fallacy: Differential incidence as a function of descriptive frames and educational context. *Contemporary Educational Psychology*, 21(2):208–218, 1996.

Katya Tentori and Vincenzo Crupi. On the conjunction fallacy and the meaning of and, yet again: A reply to Hertwig, Benz, and Krauss (2008). *Cognition*, 122(2):123–134, 2012.

Fadi Thabtah, Peter Cowling, and Yonghong Peng. Multiple labels associative classification. *Knowledge and Information Systems*, 9(1):109–129, 2006. ISSN 0219-1377.

Edward L Thorndike. The influence of primacy. *Journal of Experimental Psychology*, 10 (1):18, 1927.

Hannu Toivonen. *Frequent Itemset*, pages 418–418. Springer US, Boston, MA, 2010. ISBN 978-0-387-30164-8. doi: 10.1007/978-0-387-30164-8_317. URL `https://doi.org/10.1007/978-0-387-30164-8_317`.

Yaacov Trope, Benjamin Gervey, and Nira Liberman. Wishful thinking from a pragmatic hypothesis-testing perspective. *The mythomanias: The nature of deception and self-deception*, pages 105–31, 1997.

Yu-Kang Tu, Valerie Clerehugh, and Mark S. Gilthorpe. Ratio variables in regression analysis can give rise to spurious results: illustration from two studies in periodontology. *Journal of Dentistry*, 32(2):143 – 151, 2004. ISSN 0300-5712.

Amos Tversky and Daniel Kahneman. Belief in the law of small numbers. *Psychological bulletin*, 76(2):105, 1971.

Amos Tversky and Daniel Kahneman. Availability: A heuristic for judging frequency and probability. *Cognitive psychology*, 5(2):207–232, 1973.

Amos Tversky and Daniel Kahneman. Judgment under uncertainty: Heuristics and biases. *Science*, 185(4157):1124–1131, 1974.

Amos Tversky and Daniel Kahneman. Extensional versus intuitive reasoning: the conjunction fallacy in probability judgment. *Psychological review*, 90(4):293, 1983.

Amos Tversky and Itamar Simonson. Context-dependent preference. *Management science*, 39(10):1179–1189, 1993.

Ryan D Tweney, Michael E Doherty, Winifred J Worner, Daniel B Pliske, Clifford R Mynatt, Kimberly A Gross, and Daniel L Arkkelin. Strategies of rule discovery in an inference task. *Quarterly Journal of Experimental Psychology*, 32(1):109–123, 1980.

Takeaki Uno, Masashi Kiyomi, and Hiroki Arimura. LCM ver. 2: Efficient mining algorithms for frequent/closed/maximal itemsets. In *Proc. of the IEEE ICDM 2004 Workshop on Frequent Itemset Mining Implementations (FIMI'04)*, 2004.

Frédéric Vallée-Tourangeau and Teresa Payton. Goal-driven hypothesis testing in a rule discovery task. In *Proceedings of the 30th Annual Conference of the Cognitive Science Society*, pages 2122–2127. Cognitive Science Society Austin, TX, 2008.

K. Vanhoof and B. Depaire. Structure of association rule classifiers: a review. In *2010 International Conference on Intelligent Systems and Knowledge Engineering (ISKE)*, pages 9–12, November 2010.

Radek Škrabal, Milan Šimůnek, Stanislav Vojíř, Andrej Hazucha, Tomáš Marek, David Chudán, and Tomáš Kliegr. Association rule mining following the web search paradigm. In Peter A. Flach, Tijl Bie, and Nello Cristianini, editors, *Machine Learning and Knowledge Discovery in Databases*, volume 7524 of *Lecture Notes in Computer Science*, pages 808–811. Springer Berlin Heidelberg, 2012. ISBN 978-3-642-33485-6.

Peter C Wason. On the failure to eliminate hypotheses in a conceptual task. *Quarterly journal of experimental psychology*, 12(3):129–140, 1960.

Ronald L Wasserstein and Nicole A Lazar. The ASA's statement on p-values: context, process, and purpose. *The American Statistician*, 2016.

Christopher JCH Watkins and Peter Dayan. Q-learning. *Machine learning*, 8(3-4):279–292, 1992.

Drew Westen, Pavel S Blagov, Keith Harenski, Clint Kilts, and Stephan Hamann. Neural bases of motivated reasoning: An fMRI study of emotional constraints on partisan political judgment in the 2004 US presidential election. *Journal of cognitive neuroscience*, 18(11):1947–1958, 2006.

D Randall Wilson and Tony R Martinez. Value difference metrics for continuously valued attributes. In *Proceedings of the International Conference on Artificial Intelligence, Expert Systems and Neural Networks*, pages 11–14, 1996.

Christopher R Wolfe and M Anne Britt. The locus of the myside bias in written argumentation. *Thinking & Reasoning*, 14(1):1–27, 2008.

Xiaoxin Yin and Jiawei Han. CPAR: Classification based on predictive association rules. In *Proceedings of the SIAM International Conference on Data Mining*, pages 369–376, San Franciso, 2003. SIAM Press.

Robert B Zajonc. Attitudinal effects of mere exposure. *Journal of personality and social psychology*, 9(2, Pt. 2):1, 1968.

Mohammed J Zaki and Karam Gouda. Fast vertical mining using diffsets. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 326–335. ACM, 2003.

Daniel John Zizzo, Stephanie Stolarz-Fantino, Julie Wen, and Edmund Fantino. A violation of the monotonicity axiom: Experimental evidence on the conjunction fallacy. *Journal of Economic Behavior & Organization*, 41(3):263–276, 2000.