Cognitive biases and interpretability of inductively learnt rules

Tomáš Kliegr

Department of Information and Knowledge Engineering
University of Economics, Prague

Some parts are based on joint papers with prof. J Fürnkranz, prof. H Paulheim, prof. E Izquierdo, Dr. S Bahník and Dr. S Vojíř.
This presentation covers work in progress.

KEG Seminar, Nov 8, 2018

Outline

- Introducing Inductive and Cognitive Bias
    - AI and cognitive biases
- Problem
    - Background
- Study of Plausibility of Machine Learning Models
    - Methodology
    - Setup
    - Main Experiments
- More Experiments
    - Semantic coherence vs diversity
    - Linda experiments
    - Summary of results
- Incorporating Selected Cognitive Bias to Classification Algorithm
    - Overview of cognitive bias-inspired learning algorithms
    - Motivation
    - QCBA: Quantitative Classification based on Associations
    - Experiments
- Conclusions

Research background

- Around 2010, I set out to investigate how can we transfer cognitive biases (originally monotonicity constraint) into a machine learning algorithm.
- It turned out that the relation between cognitive and inductive biases is virtually unstudied.
- The most direct area to explore was effect of cognitive biases on perception of results of existing machine learning algorithms
- → we added studying the effect of cognitive biases on comprehensibility of machine learning models among research objectives
- Transfer of selected cognitive bias to a machine learning algorithm remained secondary objective.

Goals

1) Study semantic and pragmatic comprehension of machine learning models.

2) Verify validity of Occam's razor principle for interpretation of machine learning models.

3) Incorporate selected cognitive bias into a classification algorithm.

As a particular machine learning model to study we selected the *inductively-learned rule*.

Inductive bias (machine learning) I

> *Set of (explicit or implicit) assumptions made by a learning algorithm in order to perform induction, that is, to generalize a finite set of observation (training data) into a general model of the domain. Without a bias of that kind, induction would not be possible, since the observations can normally be generalized in many ways.*

[Hüllermeier et al., 2013]

Inductive bias (cognitive science)

> *Factors that lead a learner to favor one hypothesis over another that are independent of the observed data. When two hypotheses fit the data equally well, inductive biases are the only basis for deciding between them. In a Bayesian model, these inductive biases are expressed through the prior distribution over hypotheses.*

[Griffiths et al., 2010]

Cognitive bias (initial definition)

> *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. [Mata, 2012]*

Systematic study of cognitive biases was started in 1970's by Amos Tversky and Daniel Kahneman. It currently encompasses several dozens of cognitive phenomena.

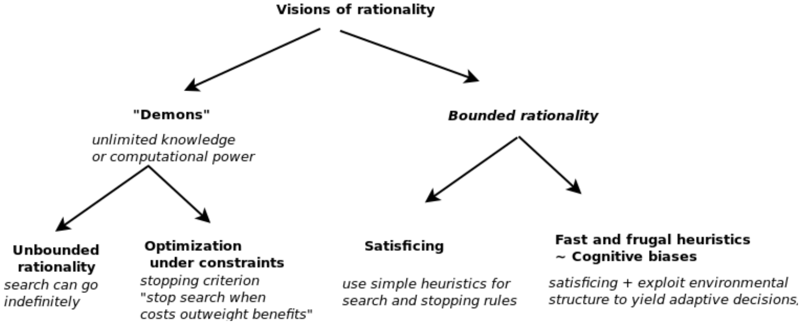
Cognitive bias (examples)

- *Base rate neglect.* Insensitivity to the prior probability of the outcome, violating the principles of probabilistic reasoning, especially Bayes' theorem.
- *Averaging heuristic.* Joint probability of two independent 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 $P(A, B) = P(A) * P(B)$.
- *Insensitivity to sample size.* 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.

Cognitive bias (Fast & Frugal revision)

- The narrow initial definition of cognitive bias as a shortcoming of human judgment was criticized – human judgment should not be compared with laws of logic and probability but rather with its performance in real world (e.g. Gigerenzer and Goldstein [1999, p. 22]).
- Gerd Gigerenzer started in the late 1990s the *Fast and frugal* heuristic program, which emphasizes ecological rationality (validity) of cognitive biases.
- If cognitive bias is applied in the right environment, it results in “frugal” rather than “erroneous” judgment.

Cognitive bias (Fast & Frugal revision)


Visions of rationality
"Demons"
unlimited knowledge
or computational power
Bounded rationality
Unbounded rationality
search can go indefinitely
Optimization under constraints
stopping criterion
"stop search when costs outweight benefits"
Satisficing
use simple heuristics for search and stopping rules
Fast and frugal heuristics
~ Cognitive biases
satisficing + exploit environmental structure to yield adaptive decisions,

Adapted from Gigerenzer and Goldstein [1999]

Why should AI study cognitive biases?

- **No free lunch theorem** [Wolpert et al., 1995]

  > *All algorithms that search for an extremum of a cost function perform exactly the same, when averaged over all possible cost functions.*

- Cognitive biases reflect reasoning patterns that the evolution has coded into the human mind to help the human species survive and **address real world problems.**

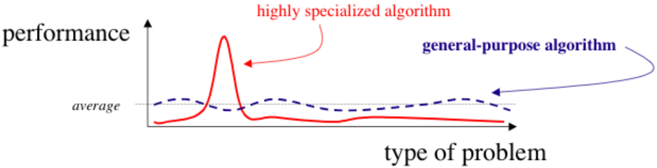
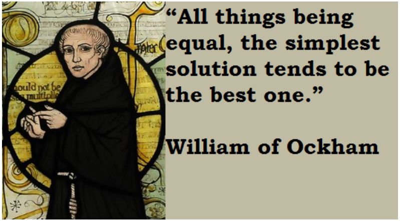


Image source: Wikipedia

Occam's razor as link between cognitive and inductive biases

- Occam's razor principle has been used as **inductive bias** in machine learning algorithms under the assumption that the simplest model will perform best.
- Are there cognitive biases that support the Occam's razor principle?


"All things being equal, the simplest solution tends to be the best one."
William of Ockham

English philosopher William of Ockham (c. 1287-1347).

In machine learning:
"Choose the shortest explanation for the observed data"
[Mitchell, 1997]

Outline

Introducing Inductive and Cognitive Bias

AI and cognitive biases

Problem

Background

Study of Plausibility of Machine Learning Models

Methodology

Setup

Main Experiments

More Experiments

Semantic coherence vs diversity

Linda experiments

Summary of results

Incorporating Selected Cognitive Bias to Classification Algorithm

Overview of cognitive bias-inspired learning algorithms

Motivation

QCBA: Quantitative Classification based on Associations

Experiments

Conclusions

Goals

1) **Study semantic and pragmatic comprehension of machine learning models.**

2) **Verify validity of Occam's razor principle for interpretation of machine learning models.**

3) Incorporate selected cognitive bias into a classification algorithm.

As a particular machine learning model to study we selected the *inductively-learned rule*.

Inductively-learned rule

Example:

```
 IF veil is white AND odour is foul THEN mushroom is
poisonous confidence = 90%, support = 5%
```

- *confidence(r)* $= a/(a+b)$, where a is number of objects 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.
- *support(r)*$= a/n$, where n is the number of all objects.

Why study rules?

- Inductively learned rules are a commonly embraced model of human reasoning in cognitive science [Smith et al., 1992, Nisbett, 1993, Pinker, 2015].
- Rule can be interpreted as a hypothesis corresponding to the logical implication $A \wedge B \Rightarrow C$.
  - rule confidence $\Leftrightarrow$ *strength of evidence* (cognitive science) $\Leftrightarrow$ conditional probability $P(C|A, B)$ (Bayesian inference)
  - rule support (machine learning) $\Leftrightarrow$ *weight of the evidence* (cognitive science)

Focusing on simple artefacts – individual rules – as opposed to entire models such as rule sets or decision trees allows deeper, more focused analysis since rule is a small self-contained item of knowledge

Comprehensibility of machine learning models

- Results on comparing representations: decision tables are better in terms of comprehensibility than decision trees or textually presented rules.
- Results on model comprehension depending on model size - mixed results:
  - Occam Razor based intuition – larger models are less comprehensible
  - Supported in some studies ([Huysmans et al., 2011]) contradicting evidence in others

> *... 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. [Allahyari and Lavesson, 2011]*

Domain constraints in machine learning models

- Plausibility of model depends on domain-specific constraints on monotonicity of attributes are followed [Freitas, 2014]

> Increasing the weight of a newly designed car, keeping all other variables equal, should result in increased predicted fuel consumption [Martens et al., 2011]

- Feelders [2000] showed on an example of real housing data and expert knowledge that decision tree models complying to monotonicity constraints were only slightly worse than unconstrained models, but they are much simpler.

Cognitive biases in machine learning

- Michalski [1983] includes a *comprehensibility postulate* according to which descriptions generated by inductive inference bear similarity to human knowledge representations
- Follow-up work on the transfer of results from cognitive science to the design of classification machine learning algorithms is, according to our review of machine learning literature, practically non-existent.
- This transfer occurred in other machine learning disciplines (e.g. in reinforcement learning)

Cognitive biases in psychological literature

- Human-perceived plausibility of hypotheses has been extensively studied in cognitive science.
- Research program on cognitive biases and heuristics was carried out by Amos Tversky and Daniel Kahneman since approximately 1970s'.

> *. . ., 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.*

The essence of cognitive biases according to Kahneman's Nobel Prize lecture (Stockholm University 2002)

[Kahneman, 2003].

Cognitive biases relevant to research goals

By analyzing psychological literature, we identified twenty relevant cognitive biases. For each of these biases, we performed:

- Justification why the bias is relevant
- The magnitude and direction of effect (increase/decrease preference for longer rules)
- Review of existing debiasing techniques, proposal of new ones.

Kliegr, Tomas, Stepan Bahnik, and Johannes Furnkranz. "A review of possible effects of cognitive biases on interpretation of rule-based machine learning models." arXiv preprint arXiv:1804.02969 (2018).

Example cognitive bias: representativeness heuristic

This heuristic relates to the tendency to make judgments based on similarity, based on rule “like goes with like”

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.

Representativeness heuristic – Linda problem

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.

Source:Tversky and Kahneman [1983]

Representativeness heuristic – Linda problem

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.

85% of people answer (b)

Conjunctive fallacy – prevalence

- 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 $P(B)$
- 85% of people answer (b) Tversky and Kahneman [1983] (83% in Hertwig and Gigerenzer [1999], and 58% in Charness et al. [2010a])
- Conjunction fallacy has been shown to hold across multiple settings (hypothetical scenarios, real-life domains), as well as for various kinds of respondents (university students, children, experts, as well as statistically sophisticated individuals) [Tentori and Crupi, 2012].

Example problem

Rule 1:

if mushroom odour is foul then the mushroom is poisonous

Rule 2:
if veil color is white and gill spacing is close and mushroom does not have bruises and has one ring and stalk surface below ring is silky then the mushroom is poisonous

Which of the rules do you find as more plausible?

- *Comprehensibility of machine learning models:* Additional conditions in rules allow the rule to appear more representative, which suggests that longer rules will be considered as more plausible than shorter rules.

Representativeness heuristic – debiasing techniques

- Charness et al. [2010a] found that the number of committed fallacies is reduced under **monetary incentive**.
- 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 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.

Outline

Introducing Inductive and Cognitive Bias

AI and cognitive biases

Problem

Background

Study of Plausibility of Machine Learning Models

Methodology

Setup

Main Experiments

More Experiments

Semantic coherence vs diversity

Linda experiments

Summary of results

Incorporating Selected Cognitive Bias to Classification Algorithm

Overview of cognitive bias-inspired learning algorithms

Motivation

QCBA: Quantitative Classification based on Associations

Experiments

Conclusions

Measuring model comprehensibility

syntactical comprehension → *semantical comprehension* → *pragmatic comprehension* → **plausibility**.

- Study of comprehensibility of machine learning models is limited to syntactic comprehensibility (size of model)
- We decided to measure comprehensibility by eliciting model **plausibility**.

For more on these definitions cf.: *Furnkranz, Johannes, Tomas Kliegr, and Heiko Paulheim. "On Cognitive Preferences and the Interpretability of Rule-based Models." arXiv preprint arXiv:1803.01316 (2018).*

Plausibility

In our experiments, we elicited preferences for rules. As a measure of preference we opted for "plausibility". To make the notion of plausibility more concrete, the respondents were provided with three dictionary definitions of plausibility:

- (Of an argument or statement) seeming reasonable or probable *(Oxford Dictionary)*
- Seeming likely to be true, or able to be believed *(Cambridge Dictionary)*
- Possibly true; able to be believed *(Cambridge Dictionary - American English)*

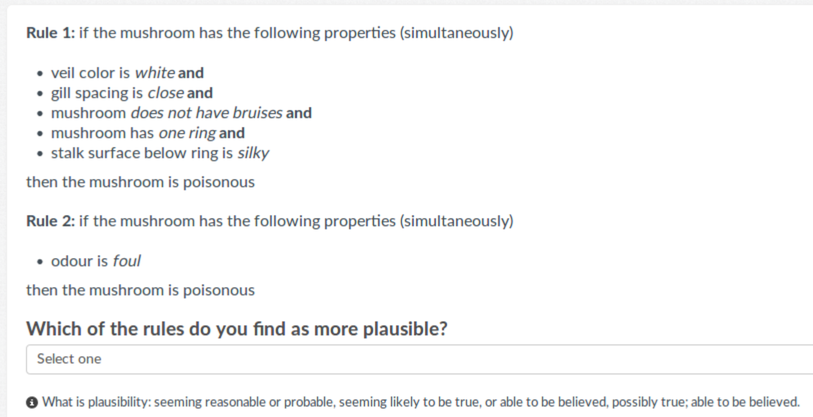
Goals

- Relevant research in cognitive science largely focuses on experiments demonstrating whether a specific bias occurs or not.
- We aim to quantify the strength of the bias as well as attribute it to specific variables.

Methodology

- Generate pairs of equally good alternatives, and ask the respondent to indicate strong/weak preference for one of the alternatives, answering "no preference" is also possible.
- Alternatives are described by observable quantitative proxy variables for cognitive biases and heuristics.
- Proxies should be ideally selected so that under perfectly rational reasoning they would have no effect on the preference.
- We analyse the effect of individual variables controlling for the effect of other variables.

Motivating example



Rule 1: if the mushroom has the following properties (simultaneously)

- veil color is *white* **and**
- gill spacing is *close* **and**
- mushroom *does not have bruises* **and**
- mushroom has *one ring* **and**
- stalk surface below ring is *silky*

then the mushroom is poisonous

Rule 2: if the mushroom has the following properties (simultaneously)

- odour is *foul*

then the mushroom is poisonous

Which of the rules do you find as more plausible?

Select one

ℹ What is plausibility: seeming reasonable or probable, seeming likely to be true, or able to be believed, possibly true; able to be believed.

Research questions

- E 1: Are longer rules more plausible than shorter rules?
- E 2: Is higher plausibility of longer rules caused by misunderstanding of “and”?
- E 3: Confidence but not support influence plausibility?
- E 4: Attribute and literal relevance as proxies?
- E 5: PageRank as a proxy for mere exposure effect?

Additional experiments (unpublished, in progress,...):

- E 6: Semantic coherence
- L 1: Can we replicate Linda experiments with crowdsourcing?
- L 2: Do people pay attention to negation?
- L 3: What is the influence of information bias?

Example problem I

Rule 1:

if mushroom odour is foul then the mushroom is poisonous

Rule 2:
if veil color is white and gill spacing is close and mushroom does not have bruises and has one ring and stalk surface below ring is silky then the mushroom is poisonous

Which of the rules do you find as more plausible?

Example problem II

Rule 1:
if mushroom odour is **creosote** then the mushroom is poisonous

Note that the bold font was not used in the original experiment.

Rule 2:
if veil color is white and gill spacing is close and mushroom does not have bruises and has one ring and stalk surface below ring is silky then the mushroom is poisonous

Which of the rules do you find as more plausible?

Initial hypotheses (later refined)

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

Hypothesized links between explanatory variables and cognitive biases. ↑ positive influence on plausibility with increasing value, ↓ negative influence, – no effect.

Example: Mere exposure effect

Rule 1
English-language Films → Rating=high

Rule 2
Horror films from 2000 → Rating=high

Because “English-language Films” have higher PageRank than Horror films from 2000, the assumptions are that:

- Through the mere exposure effect the R1 will be considered as more plausible.
- We will be able to measure the strength of correlation between maximum Pagerank and plausibility.

Example: Literal relevance – strength of evidence

Rule 2
Level of development = low → Accidents = high

Most people would likely accept that low level of development is predictive of high number of accidents.

Data elicitation

We used the CrowdFlower (`www.crowdflower.com`) to allow full reproducibility of results

CrowdFlower Products Company Blog Docs Logi

The World's Largest Workforce

Instantly hire millions of people to collect, filter, and enhance your data.

RTFM
Real Time Foto Moderator
Crowdsourced image moderation with a simple real-time API.

Senti
Sentiment Analysis
Fast, accurate human review of user-generated social media content.

Real-time Crowd Labor

— judgments/sec
current velocity

515,107,994
total judgments

Datasets

Overview of the datasets used for generating rule pairs

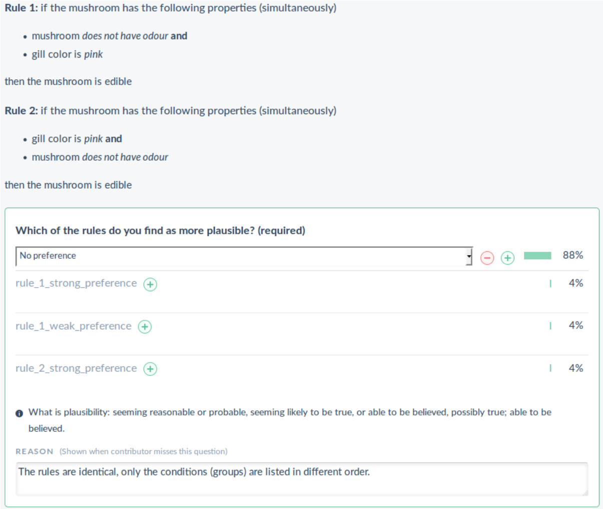
# pairs	dataset	data source	# rows	# attr.	target
80	Traffic	LOD	146	210	rate of traffic accidents in
36	Quality	LOD	230	679	quality of living in a city
32	Movies	LOD	2000	1770	movie rating
10	Mushroom	UCI	8124	23	mushroom poisonous/edib

Examples for individual datasets later on.

Quality assurance

- Level 2 contributors: “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”
- U.S., Canada and United Kingdom
- Initial quiz
- Hidden quiz questions

Example swap test question (mushrooms)



Rule 1: if the mushroom has the following properties (simultaneously)

- mushroom *does not have odour* **and**
- gill color is *pink*

then the mushroom is edible

Rule 2: if the mushroom has the following properties (simultaneously)

- gill color is *pink* **and**
- mushroom *does not have odour*

then the mushroom is edible

Which of the rules do you find as more plausible? (required)

No preference	88%
rule_1_strong_preference	4%
rule_1_weak_preference	4%
rule_2_strong_preference	4%

What is plausibility: seeming reasonable or probable, seeming likely to be true, or able to be believed, possibly true; able to be believed.

REASON (Shown when contributor misses this question)

The rules are identical, only the conditions (groups) are listed in different order.

Example swap test question (movie rating)

```
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?
```

Versions of the experiment setup

group	test questions	q*	reason
1	intersection, swap	no	baseline
2	swap	no	exclude effect of misinterpreted "and"
3	swap	yes	investigate effect of revealed conf. and supp.

* rule quality metrics shown to respondents

Data elicited

Rule-length experiment statistics. *pairs* refers to the distinct number of rule pairs, *judg* to the number of judgments, *qfr* to the quiz failure rate – the percentage of participants that did not pass the initial quiz as reported by the CrowdFlower dashboard, *part* to the number of distinct survey participants (workers), τ and ρ to the observed correlation values with p-values in parentheses.

	pairs	judg	qfr	part	Kendall's τ		Spearman's ρ	
Traffic	80	408	11	93	0.05	(0.226)	0.06	(0.230)
Quality	36	184	11	41	**0.20**	(0.002)	**0.23**	(0.002)
Movies	32	160	5	40	-0.01	(0.837)	-0.02	(0.828)
Mushrooms	10	250	13	84	**0.37**	(0.000)	**0.45**	(0.000)
total	158	1002	11	258				

Statistical methods

Rank correlation

- Kendall τ – primary measure of rank correlation
- Spearman ρ – less reliable than confidence intervals [Gibbons and Kendall, 1990]

For some experiments, we need to adjust the model for the effect of selected variables. Semipartials, $r(y|z, x)$, remove the effect of a control variable x (proxy for a specific bias) from

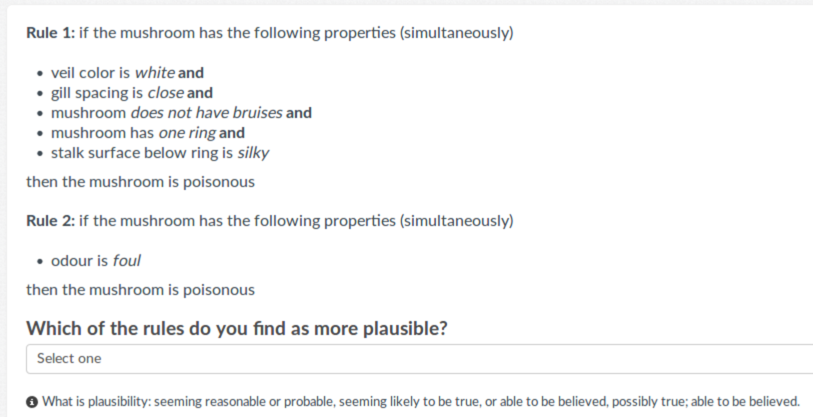
- the independent variable z (rule length Δ)
- but not from the dependent variable y (plausibility).

Exp 1: Are Shorter Rules More Plausible?

- Kendall's rank correlation coefficient τ 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).
- τ is strongest on the Mushroom dataset, $\tau =$0.37 ($p < 0.0001$) and $\rho = 0.45$ ($p < 0.0001$).
- We can reject the null hypothesis that length and plausibility are uncorrelated on two datasets (Mushroom and Quality), but not on the remaining two (Movies and Traffic).

Whether plausibility relates to rule length depends on the characteristics of the dataset.

Motivating example



Rule 1: if the mushroom has the following properties (simultaneously)

- veil color is *white* **and**
- gill spacing is *close* **and**
- mushroom *does not have bruises* **and**
- mushroom has *one ring* **and**
- stalk surface below ring is *silky*

then the mushroom is poisonous

Rule 2: if the mushroom has the following properties (simultaneously)

- odour is *foul*

then the mushroom is poisonous

Which of the rules do you find as more plausible?

ℹ What is plausibility: seeming reasonable or probable, seeming likely to be true, or able to be believed, possibly true; able to be believed.

Plausibility **increases** with rule length.

Exp 2: Misunderstanding of "and"?

- "and" possesses semantic and pragmatic properties that are foreign to $\wedge$ [Tentori et al., 2004]
- "He invited friends and colleagues to the party" ($\vee$ instead of $\wedge$) Hertwig et al. [2008]
- Measure effect: Group 1 included intersection test questions that Group 2 did not get
- Observe difference in preference for longer rules between Group 1 and Group 2.

Example intersection test question

```
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?
```

Exp 2: Misunderstanding of “and”?

Effect of intersection test questions that are meant to ensure that participants understand the logical semantics of ”and”.

		Group 1: w/o int. test questions					Group 2: with int. test questions				
dataset	pairs	judg	qfr	part	Kendall's τ		judg	qfr	part	Kendall's τ	
Quality	36	184	11	41	**0.20**	(0.002)	180	31	45	-0.03	(0.624)
Mushroom	10	250	13	84	**0.37**	(0.000)	150	44	54	**0.28**	(0.000)

Correlation between rule length Δ and plausibility Δ, p-value in parenthesis.

- The results show that misunderstanding of “and” affects plausibility on all datasets.
- On the Mushroom dataset it is not sufficient to explain the correlation between rule length and plausibility.

Exp 3: Confidence but not support influence plausibility

A certain town is served by two hospitals. In the larger hospital about 45 babies are born each day, and in the smaller hospital about 15 babies are born each day. As you know, about 50% of all babies are boys. However, the exact percentage varies from day to day. Sometimes it may be higher than 50%, sometimes lower.
For a period of 1 year, each hospital recorded the days on which more than 60% of the babies born were boys. Which hospital do you think recorded more such days?

1. The larger hospital
2. The smaller hospital
3. About the same (that is, within 5% of each other)

Most subject choose 3, while 1 is correct according to the sampling theory [Tversky and Kahneman, 1974].

P3: Experiment design (V3)

```
If movie falls into all of the following group(s) (simultaneously)

* Films Released in 2005 and
* Englishlanguage Films

then the movie is rated as good

Additional information: In our data, there are 76 movies which match
the conditions of this rule. Out of these 72 are predicted correctly
as having good rating. The confidence of the rule is 95%.
In other words, out of the 76 movies that match all the conditions
of the rule, the number of movies that are rated as good as predicted
by the rule is 72. The rule thus predicts correctly the
rating in 72/76=95 percent of cases.
```

Exp 3: Confidence but not support influence plausibility

Kendall's τ on the Movies dataset with and without additional information about the number of covered good and bad examples.

				Group 1					Group 3	
				Without information					With information	
measure	pairs	judg	qfr	part	Kendall's τ		judg	qfr	part	Kendal
Support	2*32	2*160	2*5	2*40	-0.07	(0.402)	2*160	2*5	2*40	-0.08
Confidence					0.00	(0.938)				**0.24**

Exp 3: Confidence, but not support, influences plausibility

- Insensitivity to sample size effect
- We stated the following proposition: *When both confidence and support are explicitly revealed, confidence but not support will positively affect rule plausibility.*
- The results for Movies with additional information show that the plausibility is related to confidence ($\tau = 0.24$, $p < 0.0001$) but not to support ($p = 0.36$).

Insensitivity to sample size effect is applicable to interpretation of inductively learned rules

Exp 4: Attribute and literal relevance

- Attribute relevance corresponds to human perception of the ability of a specific attribute to predict values of the attribute in rule consequent.
- Literal relevance goes one step further than attribute relevance by measuring human perception of the ability of a specific condition to predict a specific value of the attribute in the rule consequent.

Elicited with crowdsourcing experiments.

Attribute relevance

```
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.
```

Literal relevance

```
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.
```

Exp 4:Attribute and literal relevance – Results

Attribute and Literal Relevance (Group 1, Kendall's τ). Column *att* refers to number of distinct attributes, *lit* to number of distinct literals (attribute-value pairs), *excl* refers to the percentage of excluded participants on the basis of reason given shorter than 11 characters (this criterion was used in Attribute relevance experiments instead of test questions)

					Attribute relevance					
Dataset	att	judg	excl	part	Min		Avg		Max	
Traffic	14	35	70	6	−0.01	(0.745)	0.01	(0.757)	0.00	(0.983)
Mushroom	10	92	66	31	**0.30**	(0.000)	−0.11	(0.018)	**0.27**	(0.000)
					Literal relevance					
Dataset	lit	judg	qfr	part	Min		Avg		Max	
Quality	33	165	40	45	−**0.24**	(0.000)	**0.29**	(0.000)	**0.31**	(0.000)
Movies	30	150	19	40	−0.11	(0.072)	0.15	(0.012)	**0.22**	(0.000)
Traffic	58	290	40	75	−0.04	(0.377)	0.04	(0.311)	0.01	(0.797)
Mushroom	34	170	16	42	**0.22**	(0.000)	−**0.19**	(0.000)	0.11	(0.037)

Exp 4:Attribute and literal relevance – Results

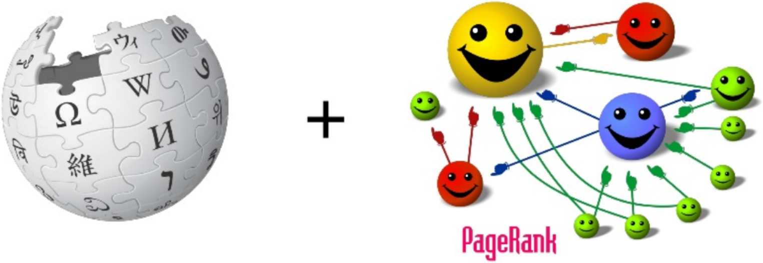
- Literal relevance has a strong correlation with the judgment of the plausibility of a rule
- Effect is strongest for the maximum relevance, which means that it is not necessary that all the literals are deemed important, but it suffices if a few (or even a single) condition is considered to be relevant

Exp 5: Modeling Recognition Heuristic using PageRank

- Recognition heuristic [Goldstein and Gigerenzer, 1999] is one of most studied fast and frugal heuristics.
- It essentially states that when you compare two objects according to some criterion that you cannot directly evaluate, and "*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.*"
- For example, if asked whether Hong Kong or Chongqing is the larger city, people are more likely to pick Hong Kong because it is better known (but Chongqing has 4x as many inhabitants).

PageRank as Proxy for Number of Exposures

In three of our datasets, the literals correspond to Wikipedia articles, which allowed us to use PageRank computed from the Wikipedia connection graph.



Adapted from slides for *Thalhammer, Andreas, and Achim Rettinger. "PageRank on Wikipedia: towards general importance scores for entities." International Semantic Web Conference. Springer, Cham, 2016.*

Modeling Recognition Heuristic using PageRank - Results

Correlation of PageRank in the knowledge graph with interpretability (plausibility) - results for Group 1.

dataset	lit	judg	qfr	part	Min		Avg		Max	
Quality	33	165	40	45	0.11	(0.048)	0.01	(0.882)	0.07	(0.213)
Movies	30	150	19%	40	**0.22**	(0.000)	−0.12	(0.051)	−0.07	(0.275)
Traffic	58	290	40%	75	−0.03	(0.471)	0.03	(0.533)	0.05	(0.195)

- To our knowledge, this is the first experiment that used PageRank to model recognition
- More research to establish the degree of actual recognition and PageRank values is needed.

Outline

Introducing Inductive and Cognitive Bias

AI and cognitive biases

Problem

Background

Study of Plausibility of Machine Learning Models

Methodology

Setup

Main Experiments

More Experiments

Semantic coherence vs diversity

Linda experiments

Summary of results

Incorporating Selected Cognitive Bias to Classification Algorithm

Overview of cognitive bias-inspired learning algorithms

Motivation

QCBA: Quantitative Classification based on Associations

Experiments

Conclusions

Example problem

Rule 1:
area > 6720, population > 607430, latitude <= 44.1281
=>Unemployment = low

Rule 2:
area > 6720, population > 607430 =>Unemployment = low

Which of the rules do you find as more *understandable*?
Which of the rules do you find as more *plausible*?

Semantic coherence

Alexander Gabriel, Heiko Paulheim, and Frederik Janssen. 2014. Learning semantically coherent rules. In Proceedings of the 1st International Conference on Interactions between Data Mining and Natural Language Processing - Volume 1202 (DMNLP'14)

Will coherent rules be better understandable?
→ Probably YES – Semantic coherence

Will they be more plausible?
→ ?? – Semantic coherence, diversity principle

Empirical studies needed

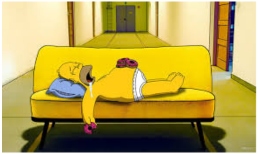
Support for semantic coherence hypothesis

SALT DEEP FOAM vs DREAM BALL BOOK
coherent triad vs *incoherence triad*

Example adapted from: Topolinski and Strack [2009]

- Semantic coherence induces *fluency* – easy cognitive processing [Topolinski and Strack, 2009]
- Perceptual fluency induces liking (preference) [Reber et al., 1998]

Backed by extensive empirical research.



Support for diversity hypothesis

> *hypotheses are better supported by varied than by uniform evidence [Tentori et al., 2016]*

```
 1) Hippopotamuses require Vitamin K for the liver to
function.
Rhinoceroses require Vitamin K for the liver to function.
---------------------------------------------------------
All mammals require Vitamin K for the liver to function.
```

```
(2) Hippopotamuses require Vitamin K for the liver to
function.
Hamsters require Vitamin K for the liver to function.
-----------------------------------------------------
All mammals require Vitamin K for the liver to function
```

Subjects judged arguments like (2) to be stronger. [Osherson et al., 1990, Heit et al., 2005]

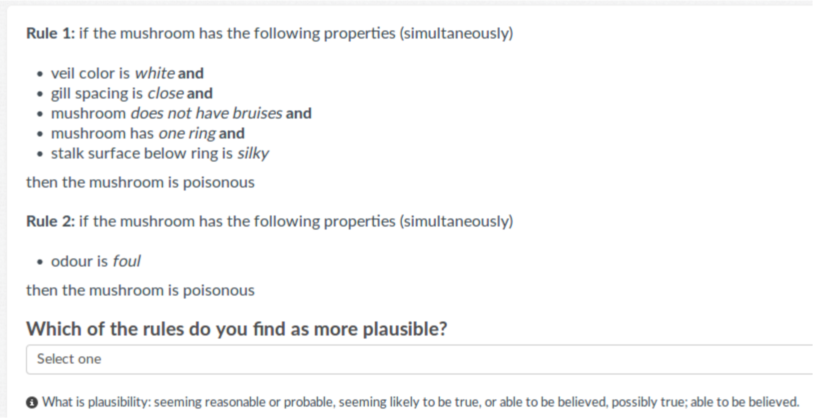
Experimental validation

Gabriel, Alexander, Heiko Paulheim, and Frederik Janssen. "Learning Semantically Coherent Rules." DMNLP@ PKDD/ECML. 2014.

- Eight UCI datasets: autos, baloons, bridges, flag, glass, hepatitis, primary-tumor, and zoo
- Use Lin similarity to compute semantic coherence of rule
- Goal was to create a rule learner respecting semantic coherence (an assumption)

Our goal: experimentally validate whether semantic coherence leads to better understandability or plausibility.

Old “Questionnaire-based” approach



Rule 1: if the mushroom has the following properties (simultaneously)

- veil color is *white* **and**
- gill spacing is *close* **and**
- mushroom *does not have bruises* **and**
- mushroom has *one ring* **and**
- stalk surface below ring is *silky*

then the mushroom is poisonous

Rule 2: if the mushroom has the following properties (simultaneously)

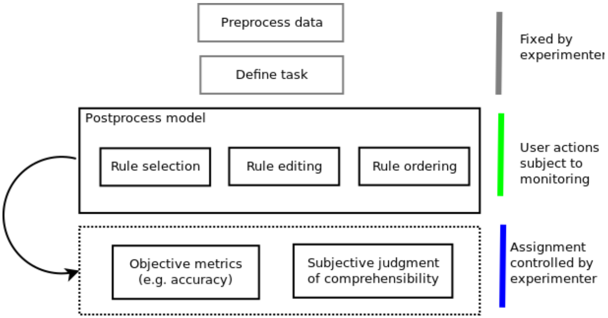
- odour is *foul*

then the mushroom is poisonous

Which of the rules do you find as more plausible?

ℹ What is plausibility: seeming reasonable or probable, seeming likely to be true, or able to be believed, possibly true; able to be believed.

Exp 6: Semantic coherence experiments


Preprocess data
Define task
Fixed by experimenter
Postprocess model
Rule selection
Rule editing
Rule ordering
User actions subject to monitoring
Objective metrics (e.g. accuracy)
Subjective judgment of comprehensibility
Assignment controlled by experimenter

Instructions

Version A *After you create the model, proceed to the rule editor and modify the model so that it exhibits a good ratio between accuracy and convincingness (plausibility). For the purpose of this task, accuracy has the same importance as convincingness. There are no other criteria or indications available for what is an acceptable value of model accuracy, or how model convincingness should be assessed.*

Version B *After you create the model, proceed to the rule editor and modify the model so that its accuracy is improved.*

Semantic coherence experiments – Results

dataset	attributes		rules		coherence	
	orig	mod	orig	mod	orig	mod
			Version A			
zoo	13	14	8	7	0.14	0.14
			Version B			
autos	138	106	54	43	0.16	0.17
glass	130	130	53	53	0.38	0.38
glass	130	121	53	53	0.38	0.39
hepatitis	47	43	18	16	0.03	0.03
primary-tumor	180	119	46	42	0.14	0.10
flag	141	33	52	18	0.18	0.16
zoo	13	15	8	8	0.14	0.16
average (for B)	111	81	41	33	0.20	0.20

Limitations

Overall, the results have shown differences among the individual datasets, which we were unable to fully explain by the selected cognitive biases. There might be many possible causes, including:

- High variance in attribute and literal relevance values, since their values were based on small number of responses.
- 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 respondents possessed in relation to individual datasets.

Replicating and extending Linda experiments

1. Replicate the original results of Tversky and Kahneman [1983] using crowdsourcing.
2. Determine the effect of negated condition.
3. Determine the effect of information bias related to inclusion of a condition with unknown value.

Setup

1. We replaced the name Linda used in the original paper with Jenny
2. There were no test questions. Instead, we offered 50% bonus for quality to respondents who provided reason for their answer longer than 10 characters
3. For analysis we used all data including the answers with no or short reasons.

Tasks and responses

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

The numbers are frequencies of responses

Exp L1: Replicating Linda

- The proportion of subjects committing fallacy in the original paper by Tversky and Kahneman [1983] was 85%.
- In our experiment V_L1 this percentage is 68%, which is significantly different from 85% at $p < 0.01$ (test for equality of proportions).
- Charness et al. [2010b] reported that providing an incentive dropped the fallacy rate to 33% (94 total respondents) and without incentive they report fallacy rate of 58% (68 respondents)
- **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 where the participants are approached directly.**

Exp L2: Effect of negated condition

v/o	text	freq
$V_L2/1$	Jenny is a bank teller	118
$V_L2/2$	Jenny is a bank teller and is not active in the feminist movement	32

- Out of the 150 respondents, 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 $p < 0.0001$.
- **We obtained convincing experimental evidence showing that negation is semantically interpreted and affects the application of the representativeness heuristic.**
- ... which is a scientific confirmation of an obvious thing.

Exp L3: Will relevant condition with unknown value increase plausibility?

v/o	text	freq
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

- 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 $p < 0.001$.
- What does this show?

Unknown value – discussion

v/o	text	freq
$V_L3a/1$	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

- Assumed reason: representativeness heuristic triggered by “not known if she is active in feminist movement”.
- Real reason: “Jenny works as a cashier in a bank” was interpreted as “Jenny works as a cashier in a bank and NOT active in feminist movement”.
- Not a new discovery. Sides et al. [2002] showed that in presence of alternative “$B \wedge F$”, alternative “B” is interpreted as “$B \wedge \neg F$”

Qualitative model of plausibility

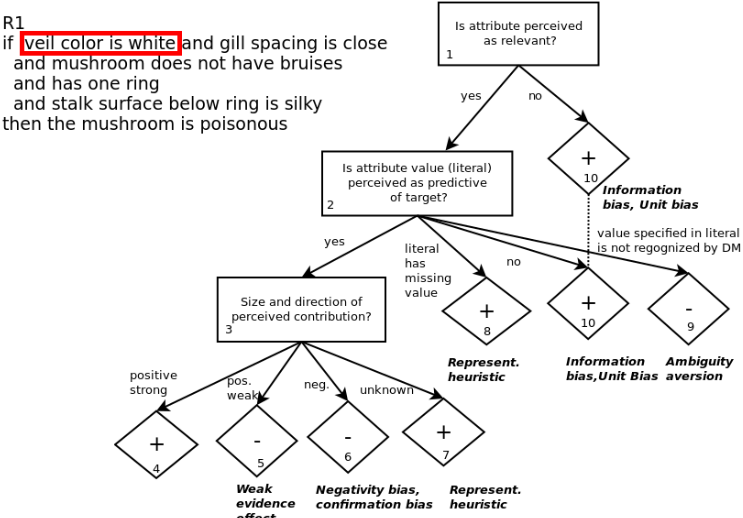
We created a qualitative model for plausibility of inductively learned rules based on:

- Results reported in cognitive science literature
- Quantitative analysis of our results
- Qualitative analysis of answers

> “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”

Example justification for response

Individual contributions of literals

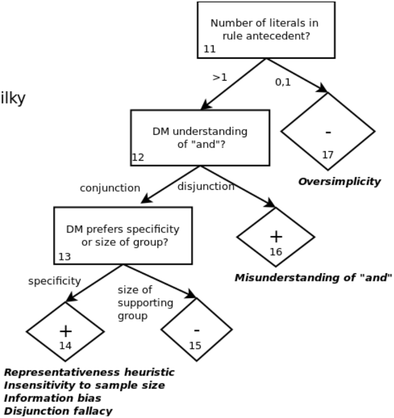


R1
if veil color is white and gill spacing is close
and mushroom does not have bruises
and has one ring
and stalk surface below ring is silky
then the mushroom is poisonous

Is attribute perceived as relevant?
1
yes
no
+
10
Information bias, Unit bias
Is attribute value (literal) perceived as predictive of target?
2
yes
literal has missing value
no
value specified in literal is not regognized by DM
+
8
Represent. heuristic
+
10
Information bias,Unit Bias
-
9
Ambiguity aversion
Size and direction of perceived contribution?
3
positive strong
pos. weak
neg.
unknown
+
4
-
5
Weak evidence effect
-
6
Negativity bias, confirmation bias
+
7
Represent. heuristic

Aggregation of literal contributions

if veil color is white and gill spacing is close
and mushroom does not have bruises
and has one ring and stalk surface below ring is silky
then the mushroom is poisonous


Number of literals in rule antecedent?
11
>1
0,1
DM understanding of "and"?
12
-
17
Oversimplicity
conjunction
disjunction
DM prefers specificity or size of group?
13
+
16
Misunderstanding of "and"
specificity
size of supporting group
+
14
-
15
Representativeness heuristic
Insensitivity to sample size
Information bias
Disjunction fallacy

(Work-in-progress)

Outline

Introducing Inductive and Cognitive Bias

AI and cognitive biases

Problem

Background

Study of Plausibility of Machine Learning Models

Methodology

Setup

Main Experiments

More Experiments

Semantic coherence vs diversity

Linda experiments

Summary of results

Incorporating Selected Cognitive Bias to Classification Algorithm

Overview of cognitive bias-inspired learning algorithms

Motivation

QCBA: Quantitative Classification based on Associations

Experiments

Conclusions

Goals

1) Study semantic and pragmatic comprehension of machine learning models.

2) Verify validity of Occam's razor principle for interpretation of machine learning models.

3) Incorporate selected cognitive bias into a classification algorithm.

Cognitive bias-inspired learning algorithms

Algorithms developed in psychology with explicit grounding in cognitive biases or processes:

- Weighted K-Nearest neighbour (Nosofsky [1990])
- Take-the-best (Gigerenzer and Goldstein [1996])
- MINERVA-Decision Making (Dougherty et al. [1999])
- PROBabilities from EXemplars (PROBEX) (Juslin and Persson [2002])

I did not find many other recent theories that met inclusion criteria (citations).

What about models developed in machine learning?

Griffiths et al. [2010] discusses the relation between inductive biases and cognitive science suggesting that the knowledge representations used in machine learning, such as rules or trees, can be useful for explaining human inferences.

Neural networks

- "little is known concerning how these structured representations [probabilistic models]can be implemented in neural systems". Griffiths et al. [2010]

Rules

- Cognitive scientist seem to shift towards exemplar-based models: *Platzer, Christine, and Arndt Bröder. "When the rule is ruled out: Exemplars and rules in decisions from memory." Journal of Behavioral Decision Making 26.5 (2013): 429-441.*

Further, we will focus only on models developed in psychology.

German Cities Problem (Gigerenzer and Goldstein [1996])

Which city has a larger population?
(a) Darmstadt
(b) Paderborn

Benchmark task - Seed data

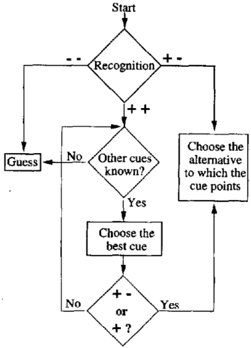
Which city has a larger population? (a) Darmstadt (b) Paderborn

- Nine explanatory attributes, numerical target (population), 83 cities → 3,403 city pairs.

City	Population	Soccer	State capital	E Germany	Uni
Darmstadt	138920	-	-	-	+
Paderborn	120680	-	-	-	+
Leipzig	511079	-	-	-	+

Take-the-best

Gigerenzer, Gerd, and Daniel G. Goldstein. "Reasoning the fast and frugal way: models of bounded rationality." Psychological review 103.4 (1996): 650.



Phase 1: Recognition principle

	Paderborn	Darmstadt	Leipzig
Recognition	+	+	-
Soccer team	+	?	?
State capital	+	+	?
E Germany	?	?	?
Industrial belt	?	+	?
Licence plate	+	+	?
Intercity	+	+	?
Exposition site	+	?	?
National capital	+	+	?
University	+	+	?

Phase 2: Search for attribute values

Identify attributes with known values for both alternatives

	Paderborn	Darmstadt	Eco validity
Soccer team	+	?	
State capital	+	+	
E Germany	?	?	
Industrial belt	?	+	
License plate	**+**	**+**	**0.77**
Intercity	**+**	**+**	**0.78**
Exposition site	+	?	
National capital	**+**	**+**	**1**
University	**+**	**+**	**0.71**

ecological validity: relative frequency with which the attribute predicts the target within the pair if it discriminates.

Phase 3: Discrimination rule

Step	Attribute	Discriminates
1	National capital	No

Attribute	Paderborn	Darmstadt	Eco validity
National capital	-	-	1
Intercity	-	+	0.78
License plate	+	+	0.77
University	+	+	0.71

Phase 4: Cue substitution

- National capital does not discriminate, search for next cue.

Step	Attribute	Discriminates
2	Intercity	Yes

Satisficing: TTB does not attempt to integrate information, but uses substitution.

Attribute	Paderborn	Darmstadt	Eco validity
National capital	-	-	1
Intercity	-	+	0.78
License plate	+	+	0.77
University	+	+	0.71

Phase 5: Maximizing rule for choice

Choose Darmstadt as the larger city.

Data in Gigerenzer and Goldstein [1996]

Paderborn	Darmstadt	Target
138k	120k	Population

Current data (Wikipedia)

Paderborn	Darmstadt	Target
145k	151k	Population

Cognitive biases in Take-the-best

- Less-is-more effect
  - Experimentally confirmed for TTB by Gigerenzer and Goldstein [1996], Lee [2015].
  - U.S. students are more correct about German city populations than about U.S. cities
  - German students are more correct about U.S. city populations than about German cities
- Confidence-frequency effect
- Overconfidence bias, hard-easy effect
- Recognition heuristic (principle)

According to Gigerenzer and Goldstein [1996].

Exemplar-based methods

Example: Nearest neighbour, weighted K-NN with $K =$*dataset size* is known in psychology as *General Context Model*:

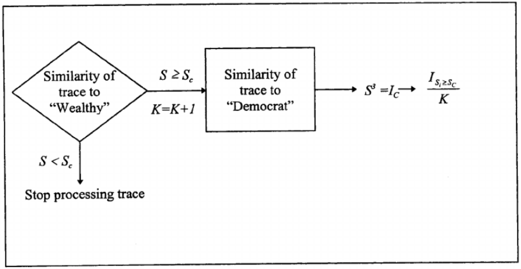
> *Nosofsky, Robert M. "Relations between exemplar -similarity and likelihood models of classification." Journal of Mathematical psychology 34.4 (1990): 393-418.*

Psychological justification [Chater et al., 2003]:

- Previously used in psychological models of categorization and memory
- Used in the *MINERVA model* of memory and generalization
- Used in model of the processes underlying probability judgments
- PROBEX model of probabilistic inference

MINERVA-Decision Making

Dougherty, Michael RP, Charles F. Gettys, and Eve E. Ogden. "MINERVA-DM: A memory processes model for judgments of likelihood." Psychological Review 106.1 (1999): 180.


Similarity of trace to "Wealthy"
$S \geq S_c$
$K=K+1$
Similarity of trace to "Democrat"
$S^3 = I_C$
$\frac{I_{S_i \geq S_C}}{K}$
$S < S_c$
Stop processing trace

Source: Dougherty et al. [1999]

Cognitive biases and MINERVA-DM

The authors of MINERVA-DM claim that the method explains:

- *Base rate neglect.* Insensitivity to the prior probability of the outcome, violating the principles of probabilistic reasoning, especially Bayes' theorem.
- *Insensitivity to sample size.* 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.
- *Conservatism, Overconfidence, Hindsight, Availability, Representativeness, Conjunction fallacy,...*

Whether MINERVA-DM explains base rate neglect is contested by Juslin and Persson [2002, p 601].

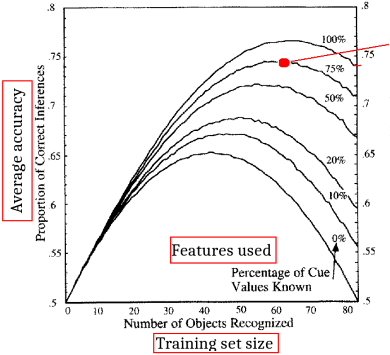
PROBEX

Juslin, Peter, and Magnus Persson. "PROBabilities from EXemplars (PROBEX): A "lazy" algorithm for probabilistic inference from generic knowledge." Cognitive science 26.5 (2002): 563-607.

- Similar to MINERVA-DM, but implements the "fast and frugal exemplar model": accurate judgments with less demands on psychological computation demands
- Probability judgments are made by comparisons between the probe and retrieved exemplars
- The judgment reflects the similarity of the retrieved example and the probe (classified instance)
- Complete version of PROBEX includes sequential sampling and dampening.

Benchmark setup

This shows that recognition bias can lead to better accuracy than using all information.



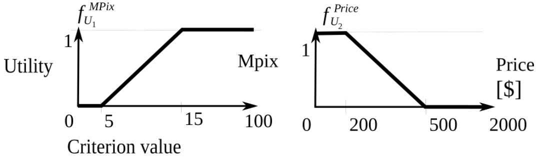
The accuracy of 74% is based on 500 randomly generated datasets (individuals), each with 3,403 comparisons (city pairs).

Within one dataset:
60 out of 80 cities are recognized
For each recognized city,
values 75% of the 9 features are known.

Adapted from Gigerenzer and Goldstein [1996] (red boxes were added).

Cognitive bias (monotonicity constraint)

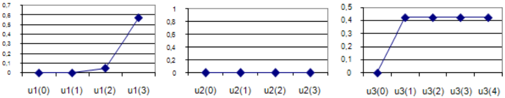
- Monotonicity "More is preferred to less" is a basic assumption relating to analysis of preferences in economics [Becker, 2007].
- In machine learning algorithms used for preference modeling, such as UTA, monotonicity is interpreted as higher value on a given criterion of an alternative results in greater or equal utility.
- As cognitive bias: a heuristic used by humans in preference problems such as product choice.



Adapted from Eckhardt and Kliegr [2012]

UTA method

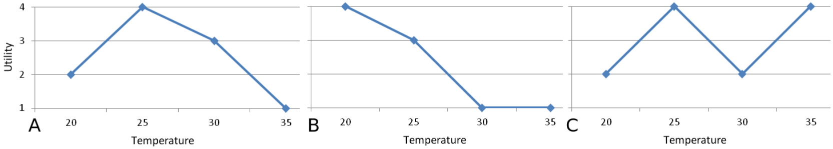
- UTA (UTilités Additives) method learns an additive piece-wise linear utility model
- The overall preference rating for an object **o** is computed as an average of utility values for all attributes: $u(\mathbf{o}) = \sum_{i=1}^{N} u_i(o_i)$, where u_i are non-decreasing value functions and o_i are its attribute values.
- The method expects that the input attributes are monotone with respect to preferences



Inside temperature Outside temperature Environment Quitness

Allowing non-monotone utility

Example. (Worker comfort) Consider the following preference learning problem: determine the utility (comfort) on 4 points scale of a worker based on temperature and humidity of the environment.



Assumption of strictly monotonic relation is unrealistic for many domains.

UTA - Non Monotonic method

Our initial point of attack was adjusting the UTA linear program formulation to penalize, rather than forbid non-monotonicity. This approach was published in Kliegr [2009] (details upon request).

Limitations of UTA-based methods:

- 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.

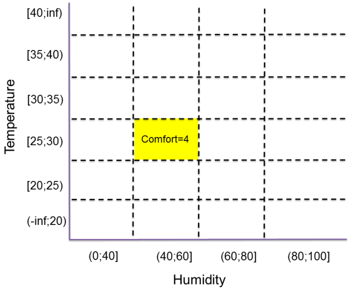
  > The utility function relating to the temperature attribute is completely independent of the value of the humidity attribute.

- Learning an UTA model can be slow on large data, relaxing motonicity further increases complexity of the LP

Selected based approach – association rule classification

Classification Based on Associations (CBA) introduced by Liu et al. [1998] and successor algorithms (CPAR, CMAR, ...).

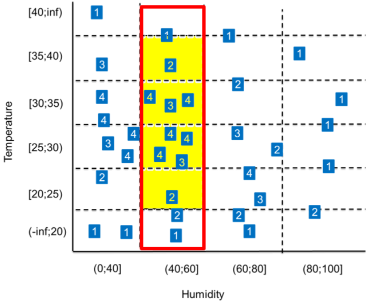
- Rules correspond to high density regions in the data
- Cardinal features need to be discretized prior to execution
  - Reduces the combinatorial complexity
  - Impairs precision of the rules



Rule in the figure):
IF Humidity = [40,60) AND Temperature = [25;30) THEN Comfort = 4
confidence = 75%, support = 4

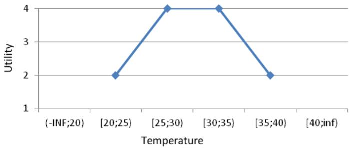
Limitations of CBA

- Association rules identify only the high density regions in the data, which have a strong presence of one target class.
- The definition of “high density” is controlled by the minimum support parameter, and the definition of strong presence by the minimum confidence parameter.



Humidity=(40;60] & Temperature=[20;25) => Utility=2
Humidity=(40;60] & Temperature=[25;30) => Utility=4
Humidity=(40;60] & Temperature=[30;35) => Utility=4
Humidity=(40;60] & Temperature=[35;40) => Utility=2

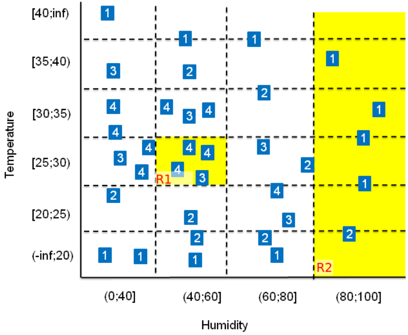
Corresponding
Conditional utility model



Ceteris paribus: Humidity = (40;60]

Challenges for association rule learning

- Ignores regions in the data with small density (otherwise combinatorial explosion).
- Limited to hypercube (rectangle) regions: The problem is further aggravated by the fact that learning is performed on transformed feature space (cardinal features are discretized to bins).
- Does not incorporate the monotonicity assumption
- Prediction is crisp rather



Approach

The standard way to incorporate domain constraints into the learning algorithm is

- → 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.

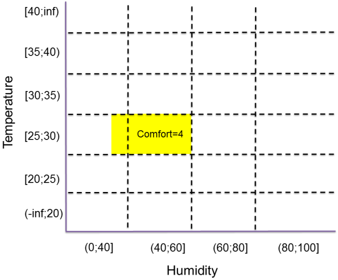
What we do:

- Readjust association rule output to reflect monotonicity *without adversely affecting confidence and support*

Win-win?

“The discretization trick”

- 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



Rule after monotonic extension

Overview of the MARC (QCBA) framework

Monotonicity Exploiting Association Rule Classification

- Learn association rules
- Postprocess the rules to incorporate the monotonicity assumption
- Annotate the rules with probability density functions (optional)

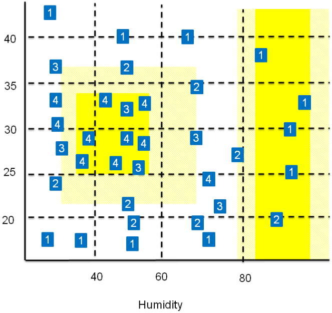
Procedures:

- Association rule learning and pruning (standard algorithms)
- Rule Extension – the core procedure implementing the mon. assump.
- Rule Fuzzification - further extending rule coverage
- Rule Annotation with probability density functions
- Rule Mixture/one rule classification

Interactive demonstration

`https://nb.vse.cz/~klit01/qcba/tutorial.html`

Rule fuzzification



The coverage of each literal created over a cardinal attribute in the body of a rule is extended by appending a value adjacent to the lowest and highest values.

Setup – Datasets (22)

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

Ablation study

QCBA evaluation and ablation study – aggregate results for 22 UCI datasets

configuration	cba	#1	#2	#3	#4	#5	#6	#7
refit		Y	Y	Y	Y	Y	Y	Y
literal pruning		-	Y	Y	Y	Y	Y	Y
trimming		-	-	Y	Y	Y	Y	Y
extension		-	-	-	Y	Y	Y	Y
postpruning		-	-	-	-	Y	Y	Y
def. rule overlap - tran.		-	-	-	-	-	Y	-
def. rule overlap - range		-	-	-	-	-	-	Y
wins/ties/losses vs CBA		14-1-7	15-0-7	12-0-10	11-0-11	14-1-7	11-0-11	14-1-7
P-value (Wilcoxon)		.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	24	20	20	43	43	43	43
build time normalized	1.0	1.9	2.0	2.0	17.4	17.3	17.3	17.4

Benchmark

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.

	CBA (baseline)			CBA (Liu)		QCBA (#5)			QCBA (#6)		
	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	.87	76	3.8	.82	42	3.8
autos	.79	57	3.0	.79	54	.78	50	2.5	.79	44	2.5
breast-w	.95	51	2.8	.96	49	.95	31	2.7	.95	20	2.7
diabetes	.75	51	3.9	.75	57	.77	41	2.9	.76	30	2.9
glass	.71	28	3.9	.73	27	.69	24	2.8	.69	22	2.8
hepatitis	.79	32	3.9	.85	23	.82	29	3.0	.82	22	3.0
hypothyroid	.98	29	3.1	.98	35	.99	16	2.4	.98	15	2.4
ionosphere	.92	53	2.5	.92	45	.88	40	1.9	.86	22	1.9
iris	.92	6	2.0	.93	5	.93	5	1.1	.93	4	1.1
labor	.84	11	3.6	.83	12	.88	11	1.8	.86	8	1.8
lymph	.81	38	3.7	.80	36	.79	37	2.9	.79	37	2.9
sonar	.74	44	2.9	.76	37	.77	35	2.7	.72	19	2.7
vehicle	.69	147	3.9	.69	125	.71	107	3.6	.70	79	3.6
	.84	49	3.3	.84	49	.84	38	2.6	.83	28	2.6
average	.84	49	3.3	.84	49	.83	28	2.6	.84	37	2.6

Benchmark

- Symbolic learners: C4.5, FURIA, PART, RIPPER
- we used the implementations available in the Weka framework
- For CBA and QCBA we used our implementations

Counts of wins, losses and ties for QCBA (#5)

dataset	QCBA won	tie	loss	omitted	p
J48 auto	12	1	9	18	0.46510
PART auto	8	5	8	17	0.71514
RIPPER auto	12	4	6	18	0.15787
FURIA auto	5	4	12	0	0.13963
CBA	16	2	4	0	0.00450

Outline

- Introducing Inductive and Cognitive Bias
  - AI and cognitive biases
- Problem
  - Background
- Study of Plausibility of Machine Learning Models
  - Methodology
  - Setup
  - Main Experiments
- More Experiments
  - Semantic coherence vs diversity
  - Linda experiments
  - Summary of results
- Incorporating Selected Cognitive Bias to Classification Algorithm
  - Overview of cognitive bias-inspired learning algorithms
  - Motivation
  - QCBA: Quantitative Classification based on Associations
  - Experiments
- Conclusions

Factors affecting plausibility of ML models

We identified only two factors that are 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**. There is empirical evidence showing that domain experts do not find rules that contain conditions violating prior domain knowledge (such as monotonicity) as plausible.

The results pertaining to plausibility in the list above were scattered in articles dealing with other topics

Twenty cognitive biases related to ML

Our review identified twenty cognitive biases, heuristics and effects that can give rise to systematic errors when inductively learned rules are interpreted. They can be divided into two groups:

- Triggered by domain knowledge related to attributes and values in the rules. Example: aversion to ambiguous information.
- Generic strategies applied when evaluating alternatives. Example: insensitivity to sample size (confidence more important as support).

For most biases and heuristics involved in our study, psychologists have proposed "debiasing" measures. We related these to machine learning.

Occam's razor

- "Smaller is better" theories in machine learning are based on the Occam's razor principle.
- In our review of literature from cognitive science, we have not identified 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.
- Surprising result: reports of "oversimplicity" avoidance

While Occam's razor is a generally accepted principle, to our knowledge the problem whether it is used as a "built-in" heuristic or cognitive bias in human reasoning has not yet been to our knowledge studied.

Take-the-best

According to our review, the only machine learning algorithm inspired by cognitive biases.

- Select the best solution based on the first discriminatory feature.
- For small training sample sizes it is reported to outperform standard ML models (KNN, NN, DT).
- The algorithm is based on the satisficing behavioural strategy, which corresponds to the notion of overfitting avoidance in machine learning.

While TTB has already been presented at machine learning venues, it does not, in our opinion, obtained the level of attention it would deserve.

Relation between cognitive and inductive biases

We identified the following correspondences between the two notions:

- Take-The-Best is a particular example of a reasoning heuristic and an effective inductive bias.
- Both cognitive biases and inductive biases have certain scope, set of problems, for which they are suitable – ecologically valid – and for other problems they result in errors.
- Knowledge representations used in machine learning, such as rules or trees, are accepted by some cognitive scientists for explaining human inferences.

Our contribution: Methodology for measuring strength of cognitive biases

Practical recommendations for ML Software I

- Remove near-redundant rules and near-redundant literals from rules
- Represent rule quality measures as frequencies not ratios
- Make “and” conjunction unambiguous
- Present confidence interval for rule confidence
- Avoid the use of negated literals as well as positive/negative class labels

Practical recommendations for ML Software II

- Sort rules as well as literals in the rules from strongest to weakest
- Provide explanation for literals in rules
- Explain difference between negation and absence of a condition
- Elicit and respect monotonicity constraints
- Educate and assess human analysts

Software & Data

- data & analysis software for rule length experiments at `https://github.com/kliegr/rule-length-project`, `https://github.com/kliegr/rule_interpretability_analysis`
- R packages:*arc* package[1], *qcba* package[2]

All open source license.

[1] https://cran.r-project.org/web/packages/arc/

[2] https://cran.r-project.org/web/packages/qCBA/

Thank you for attention!

Some of the earliest and most influential learning algorithms were developed by psychologists.

Elements of machine learning (Langley, 1996, p.383)

Literature I

Eyke Hüllermeier, Thomas Fober, and Marco Mernberger. *Inductive Bias*, pages 1018–1018. Springer New York, New York, NY, 2013. ISBN 978-1-4419-9863-7. doi: 10.1007/978-1-4419-9863-7_927. URL http://dx.doi.org/10.1007/978-1-4419-9863-7_927.

Tom M Mitchell. Machine learning. 1997. *Burr Ridge, IL: McGraw Hill*, 45:37, 1997.

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.

Literature II

R Mata. Cognitive bias. *Encyclopedia of human behaviour*, 1: 531–535, 2012.

Gerd Gigerenzer and Daniel G Goldstein. Fast and frugal heuristics. In *Simple heuristics that make us smart*, pages 75–95. Oxford University Press, 1999.

David H Wolpert, William G Macready, et al. No free lunch theorems for search. Technical report, Technical Report SFI-TR-95-02-010, Santa Fe Institute, 1995.

Edward E Smith, Christopher Langston, and Richard E Nisbett. The case for rules in reasoning. *Cognitive science*, 16(1):1–40, 1992.

Richard E Nisbett. *Rules for reasoning*. Psychology Press, 1993.

Literature III

Steven Pinker. *Words and rules: The ingredients of language*. Basic Books, 2015.

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.

Hiva Allahyari and Niklas Lavesson. User-oriented assessment of classification model understandability. In *11th Scandinavian Conference on Artificial Intelligence*. IOS Press, 2011.

Alex A Freitas. Comprehensible classification models: a position paper. *ACM SIGKDD explorations newsletter*, 15(1):1–10, 2014.

Literature IV

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.

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.

Ryszard S Michalski. A theory and methodology of inductive learning. In *Machine learning*, pages 83–134. Springer, 1983.

Daniel Kahneman. A perspective on judgment and choice. *American Psychologist*, 58, 2003.

Amos Tversky and Daniel Kahneman. Extensional versus intuitive reasoning: the conjunction fallacy in probability judgment. *Psychological review*, 90(4):293, 1983.

Literature V

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.

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, 2010a. ISSN 0899-8256.

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.

Literature VI

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.

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.

Gerd Gigerenzer and Daniel G Goldstein. Reasoning the fast and frugal way: models of bounded rationality. *Psychological review*, 103(4):650, 1996.

Literature VII

Gerd Gigerenzer and Ulrich Hoffrage. How to improve Bayesian reasoning without instruction: frequency formats. *Psychological review*, 102(4):684, 1995.

Jean D Gibbons and MG Kendall. Rank correlation methods. *Edward Arnold*, 1990.

Katya Tentori, Nicolao Bonini, and Daniel Osherson. The conjunction fallacy: a misunderstanding about conjunction? *Cognitive Science*, 28(3):467–477, 2004.

Ralph Hertwig, Björn Benz, and Stefan Krauss. The conjunction fallacy and the many meanings of and. *Cognition*, 108(3): 740–753, 2008.

Amos Tversky and Daniel Kahneman. Judgment under uncertainty: Heuristics and biases. *Science*, 185(4157):1124–1131, 1974.

Literature VIII

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.

Sascha Topolinski and Fritz Strack. Scanning the “fringe” of consciousness: What is felt and what is not felt in intuitions about semantic coherence. *Consciousness and Cognition*, 18(3): 608–618, 2009.

Rolf Reber, Piotr Winkielman, and Norbert Schwarz. Effects of perceptual fluency on affective judgments. *Psychological science*, 9(1):45–48, 1998.

Literature IX

Katya Tentori, Nick Chater, and Vincenzo Crupi. Judging the probability of hypotheses versus the impact of evidence: Which form of inductive inference is more accurate and time-consistent? *Cognitive science*, 40(3):758–778, 2016.

Daniel N Osherson, Edward E Smith, Ormond Wilkie, Alejandro Lopez, and Eldar Shafir. Category-based induction. *Psychological review*, 97(2):185, 1990.

Evan Heit, Ulrike Hahn, and Aidan Feeney. Defending diversity. *Categorization inside and outside of the laboratory: Essays in honor of Douglas L. Medin*, pages 87–99, 2005.

Literature X

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, 2010b.

Ashley Sides, Daniel Osherson, Nicolao Bonini, and Riccardo Viale. On the reality of the conjunction fallacy. *Memory & Cognition*, 30(2):191–198, 2002.

Robert M Nosofsky. Relations between exemplar-similarity and likelihood models of classification. *Journal of Mathematical psychology*, 34(4):393–418, 1990.

Michael RP Dougherty, Charles F Gettys, and Eve E Ogden. Minerva-dm: A memory processes model for judgments of likelihood. *Psychological Review*, 106(1):180, 1999.

Literature XI

Peter Juslin and Magnus Persson. Probabilities from exemplars (probex): A "lazy" algorithm for probabilistic inference from generic knowledge. *Cognitive science*, 26(5):563–607, 2002.

Michael D Lee. Evidence for and against a simple interpretation of the less-is-more effect. *Judgment and Decision Making*, 10(1): 18, 2015.

Henry Brighton. Robust inference with simple cognitive models. In *AAAI Spring Symposium: Between a Rock and a Hard Place: Cognitive Science Principles Meet AI-Hard Problems*, pages 17–22, 2006.

Literature XII

Nick Chater, Mike Oaksford, Ramin Nakisa, and Martin Redington. Fast, frugal, and rational: How rational norms explain behavior. *Organizational behavior and human decision processes*, 90(1): 63–86, 2003.

Gary Stanley Becker. *Economic theory*. Transaction Publishers, 2007.

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.

Tomás Kliegr. UTA - NM: Explaining stated preferences with additive non-monotonic utility functions. In *Proceedings of the ECML'09 Preference Learning Workshop*, 2009.

Literature XIII

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.