

Explanations Using Defeasible Rule Sets

Checking a rule based classifier's explanations with formal defeasible reasoners

BACKGROUND

Exception Closed Conjunctive Rule Sets (ECCRS)

IF-THEN rules over feature values. When several rules fire, the Most Specifics one Wins (MSW) and becomes the explanation.

Defeasible reasoning

Defeasible reasoning extends propositional logic by allowing statements and inferences that are typically true to hold and later be retracted given new information.

FROM ECCRS TO KLM

ECCRS rule set

r1: IF odour = none THEN edible
r2: IF odour = none $\wedge$ spore = green
THEN poisonous
r3: IF odour = foul THEN poisonous

translate

label 1 $\rightarrow$ m
label 0 $\rightarrow$ $\neg$ m

Defeasible knowledge base

none $\sim$ m
none $\wedge$ green $\sim$ $\neg$ m
foul $\sim$ $\neg$ m

The workbench first checks the three assumptions, then runs MSW and keeps an explanation trace for every instance. It translates the rule set: each body becomes an antecedent and each label becomes m (edible) or $\neg$ m (poisonous). r2 is an exception to r1, so the KLM ranking puts it above r1, just as MSW lets it win. For each instance F it asks whether $F \sim$ m or $F \sim$ $\neg$ m and compares the two verdicts.

RATIONAL CLOSURE

Dan Schlesinger

Ranks statements from general to specific. For an instance it discards the most general ranks until what is left is consistent with the instance, then reads off the label.

LEXICOGRAPHIC CLOSURE

Augustine Mochoeneng

Refines Rational Closure. Instead of throwing away a whole rank, it keeps as many of its statements as possible, so more knowledge survives each query.

SYSTEM W

Antony Baker

Compares possible worlds by sets of falsifying statements, layer by layer, using set inclusion. It refines Rational Closure while staying cautious about what it concludes.

RESULTS

Table 1: Agreement between MSW and each closure, per instance with rule sets that satisfies the three assumptions.

Instances	RC	LC	SW
13 509	100%	100%	100%

Table 2 showing assumption violations under LC. A conflict occurs when both labels are entailed or an abstention when no label is entailed.

Assumption Violated	Agree	Abstain	Conflict	Wrong label
Closure	13 395	144	0	0
Sanity	12 019	0	1 490	0
Override	13 374	0	135	0

CONCLUSIONS

- On all 13 509 instances, the label that ECCRS predicts is the same label that all three closures entail. Thus, ECCRS predictions under the given structural assumptions are equivalent to RC, LC and SW.
- If assumptions are violated in a ruleset, the reasoner does not confidently produce a label different to ECCRS. The reasoner either abstains or entails both label which is registered as a conflict.