Does the Model Think As We Expect? Exploring ML Model Logic Through Decision Rules

Natalia Andrienko, Gennady Andrienko, Bahavathy Kathirgamanathan

Fraunhofer Institute IAIS (Intelligent Analysis and Information Systems), Germany

Lamarr Institute for Machine Learning and Artificial Intelligence, Germany

Research Area: Human-Centred AI

INTRODUCTION

- **Key question:** ML models can be accurate, but do they reason as we expect?
- **Why This Matters:**
  - Trust in ML models is not just about accuracy – it's about understanding *why* they make decisions.
  - A model may produce correct predictions while relying on reasoning that differs from human logic.
  - This misalignment can affect model adoption, interpretation, and decision-making in critical applications.
- **Additional question:** Can we use a ML model to understand the data and phenomena it captures?
- **Our goal:** support exploring model's internal workings and logic.
  - We consider models represented by systems of decision rules.

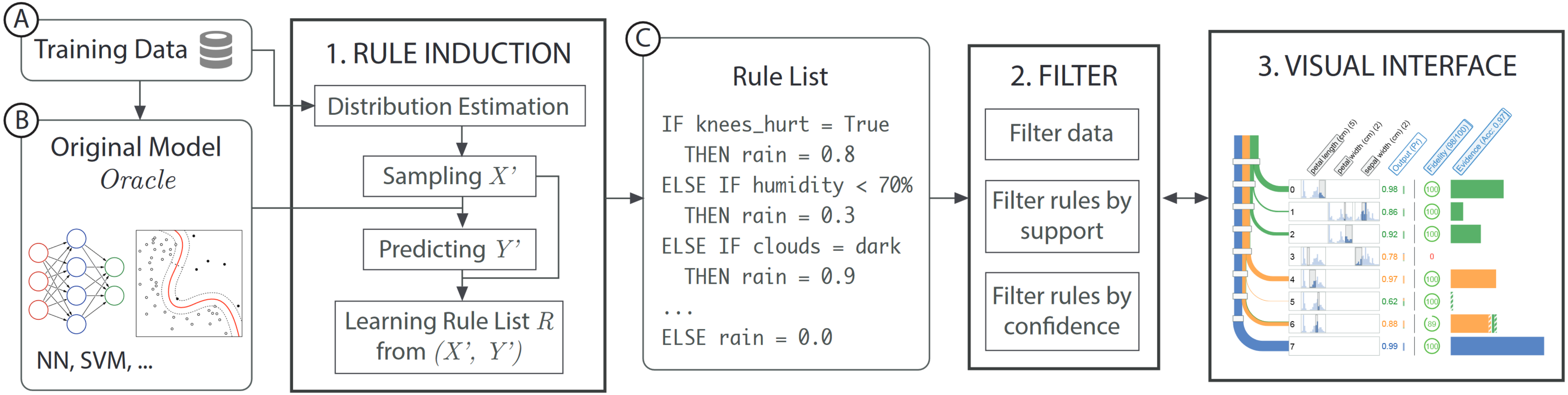
INTERPRETABILITY OF ML MODELS

- A model is **interpretable** if a person can understand its internal mechanics and capture relevant knowledge concerning relationships between inputs and outputs.
- Interpretability is essential for trust, transparency, debugging, and domain insight.

Inherently Interpretable Models (contrasted with "black box" models):

- **Linear Models** – simple mathematical expressions with coefficients showing direct feature influence.
- **Decision Trees** – visualize decisions as a sequence of understandable splits.
- **Rule-Based Models** – express decisions as explicit IF–THEN rules.
  - Decision trees can be transformed to equivalent rule-based models.
- A common approach to explain black-box models is to approximate their behaviour with an interpretable surrogate, such as a decision tree or rule set.

VISUALISATION OF A RULE-BASED MODEL: RULEMATRIX

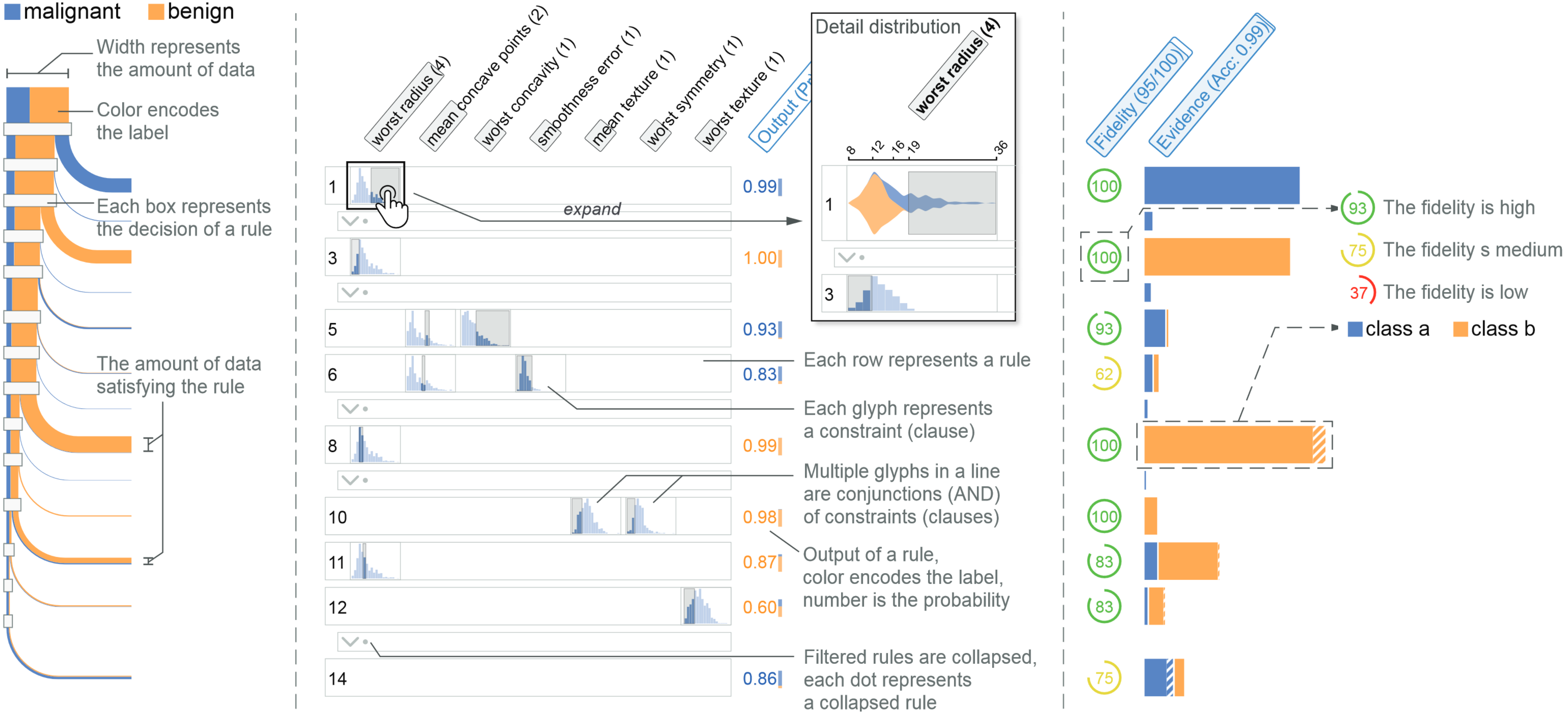


```
IF knees_hurt = True
  THEN rain = 0.8
ELSE IF humidity < 70%
  THEN rain = 0.3
ELSE IF clouds = dark
  THEN rain = 0.9
...
ELSE rain = 0.0
```

Fig. 2. The pipeline for creating a rule-based explanation interface. The rule induction step (1) takes (A) the training data and (B) the model to be explained as input, and produces (C) a rule list that approximates the original model. Then the rule list is filtered (2) according to user-specified thresholds of support and confidence. The rule list is visualized as RuleMatrix (3) to help users navigate and analyze the rules.

Ming, Y., Qu, H., & Bertini, E. (2018).
RuleMatrix: Visualizing and Understanding Classifiers with Rules.
IEEE Transactions on Visualization and Computer Graphics, 25, 342-352.

RULEMATRIX



Width represents the amount of data

Color encodes the label

Each box represents the decision of a rule

The amount of data satisfying the rule

expand

Each row represents a rule

Each glyph represents a constraint (clause)

Multiple glyphs in a line are conjunctions (AND) of constraints (clauses)

Output of a rule, color encodes the label, number is the probability

Filtered rules are collapsed, each dot represents a collapsed rule

The fidelity is high

The fidelity s medium

The fidelity is low

Our previous work:

EXPLAINING RULE-BASED MODEL'S LOGIC USING A SIMPLIFIED DESCRIPTIVE MODEL

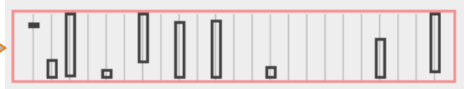
Adilova, L., Kamp, M., Andrienko, G. and Andrienko, N.
"Re-interpreting rules interpretability".
International Journal of Data Science and Analytics (2023).
doi:10.1007/s41060-023-00398-5.

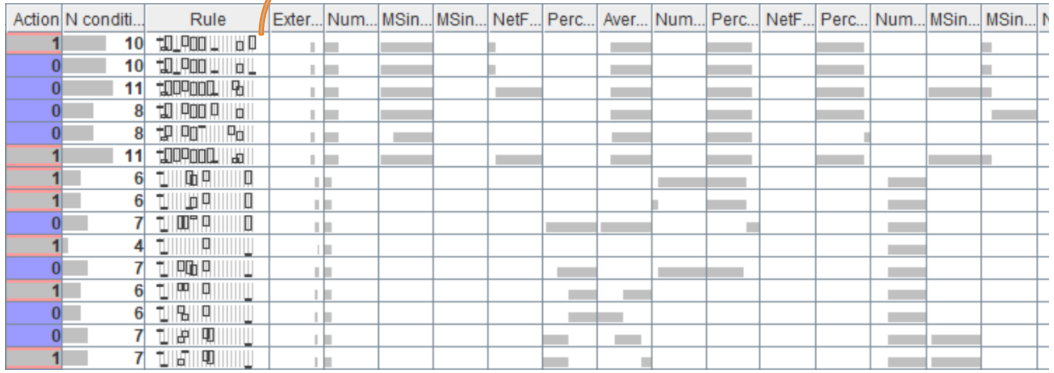
PROBLEM STATEMENT

- Given: a rule-based model with a large number of decision rules
- Task: facilitate human's comprehension of the logic of the entire model
  - Challenge: although decision rules and decision trees are considered "inherently interpretable", comprehension of a large system of rules or decision tree may be beyond human perceptual and cognitive capacity.
  - Aspects of complexity:
    - Number of rules
    - Number of conditions in a rule

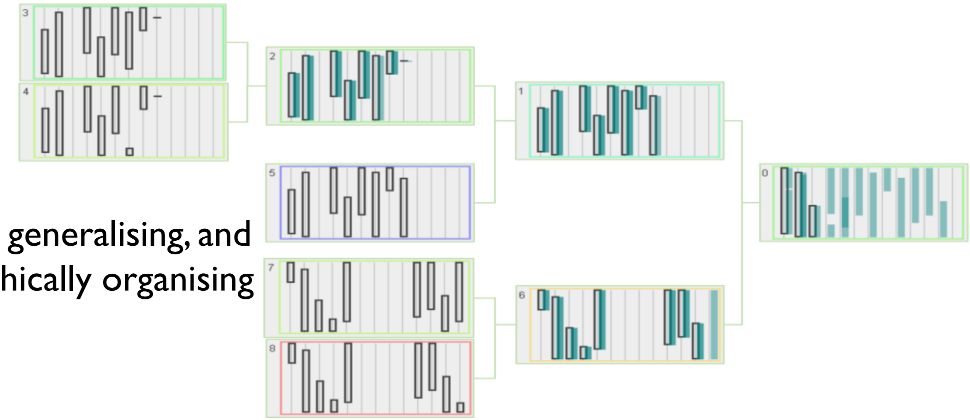
APPROACH: AGGREGATE – GENERALISE – CREATE A SIMPLER DESCRIPTIVE MODEL

Visualisation of a rule-based model



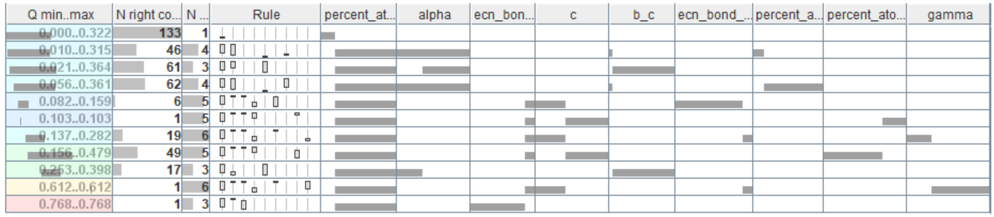


Action	N conditi...	Rule	Exter...	Num...	MSin...	MSin...	NetF...	Perc...	Aver...	Num...	Perc...	NetF...	Perc...	Num...	MSin...	MSin...	N
1	10																
0	10																
0	11																
0	8																
0	8																
1	11																
1	6																
1	6																
0	7																
1	4																
0	7																
1	6																
0	6																
0	7																
1	7																

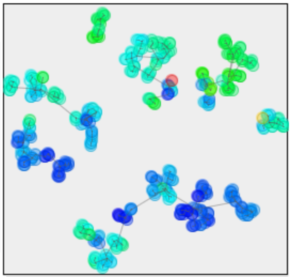


Joining, generalising, and hierarchically organising rules

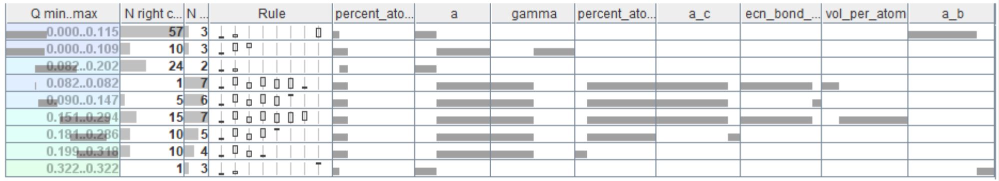
Aggregated rules (11), N conditions (45), Total uses (402); obtained with coherence threshold 1.000 and max Q difference 0.35000



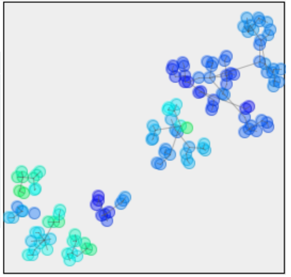
Q min..max	N right co...	N ...	Rule	percent_at...	alpha	ecn_bon...	c	b_c	ecn_bond_...	percent_a...	percent_ato...	gamma
0.000..0.322	133	1										
0.010..0.315	46	4										
0.021..0.364	61	3										
0.056..0.361	62	4										
0.082..0.159	6	5										
0.103..0.103	1	5										
0.137..0.282	19	6										
0.156..0.479	49	5										
0.253..0.398	17	3										
0.612..0.612	1	6										
0.768..0.768	1	3										



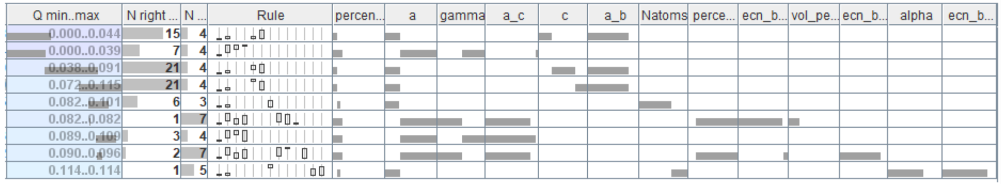
Aggregated rules (9), N conditions (40), Total uses (138); obtained with coherence threshold 1.000 and max Q difference 0.15000



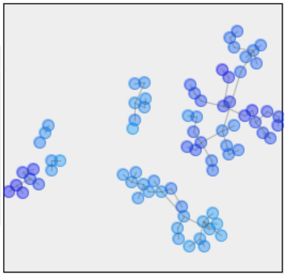
Q min..max	N right c...	N ...	Rule	percent_ato...	a	gamma	percent_ato...	a_c	ecn_bond_...	vol_per_atom	a_b
0.000..0.115	57	3									
0.000..0.109	10	3									
0.082..0.202	24	2									
0.082..0.082	1	7									
0.090..0.147	5	6									
0.151..0.294	15	7									
0.181..0.286	10	5									
0.199..0.318	10	4									
0.322..0.322	1	3									



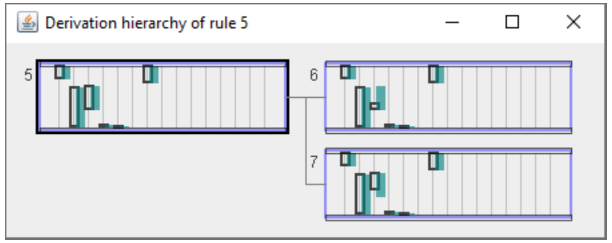
Aggregated rules (9), N conditions (42), Total uses (81); obtained with coherence threshold 1.000 and max Q difference 0.07500

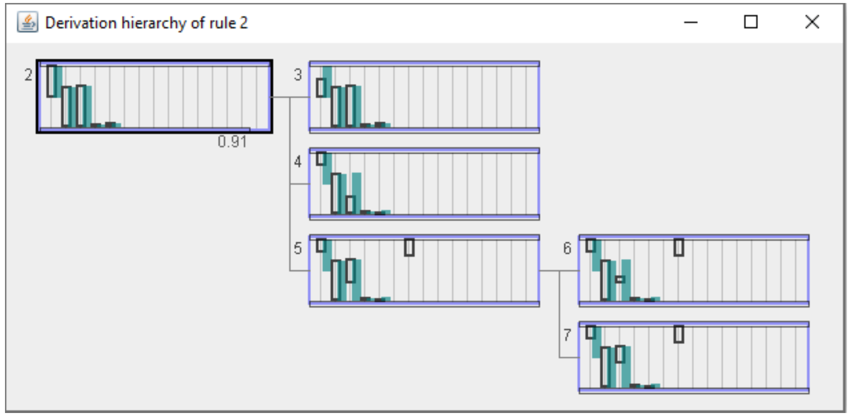


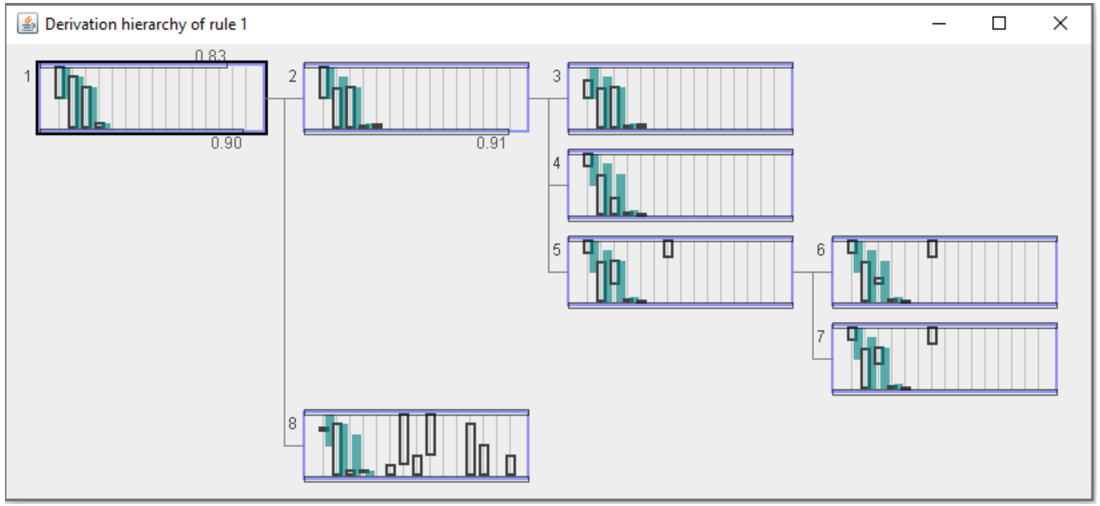
Q min..max	N right ...	N ...	Rule	percen...	a	gamma	a_c	c	a_b	Natoms	perce...	ecn_b...	vol_pe...	ecn_b...	alpha	ecn_b...
0.000..0.044	15	4														
0.000..0.039	7	4														
0.038..0.091	21	4														
0.072..0.115	21	4														
0.082..0.101	6	3														
0.082..0.082	1	7														
0.089..0.109	3	4														
0.090..0.096	2	7														
0.114..0.114	1	5														

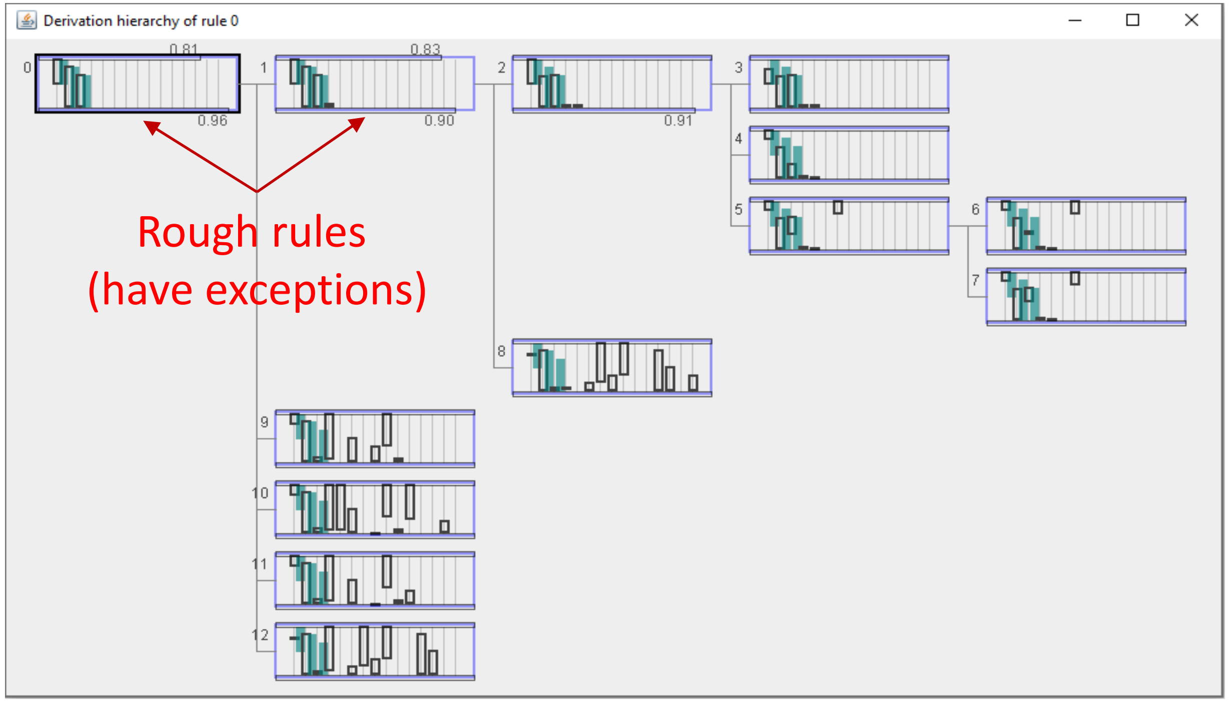


KEY IDEA: AGGREGATE AND GENERALISE SIMILAR RULES





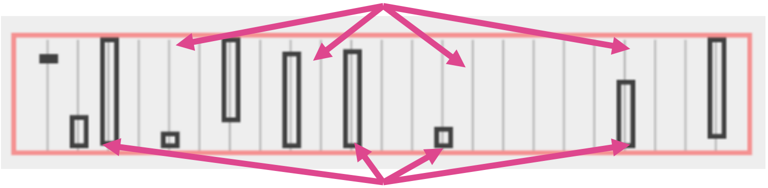




Rough rules
(have exceptions)

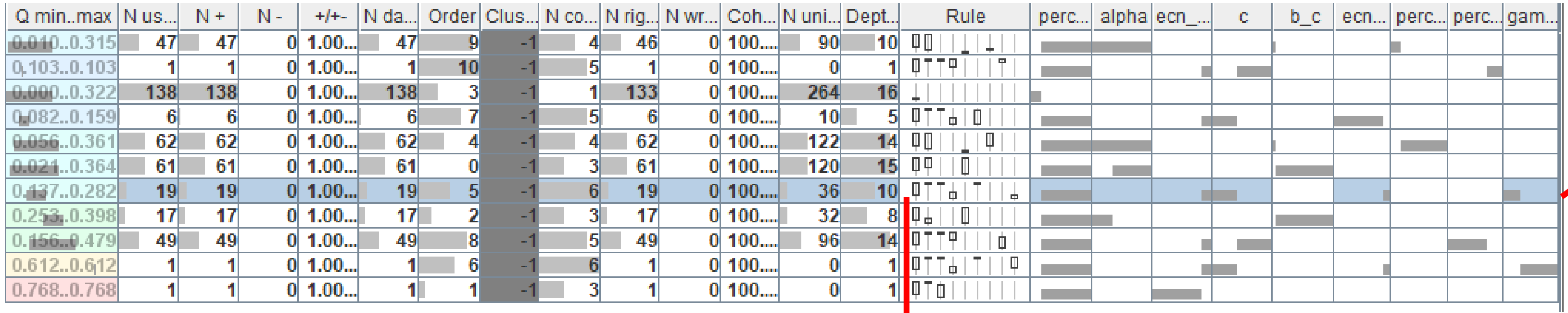
Visual representation of one rule

Features → vertical axes

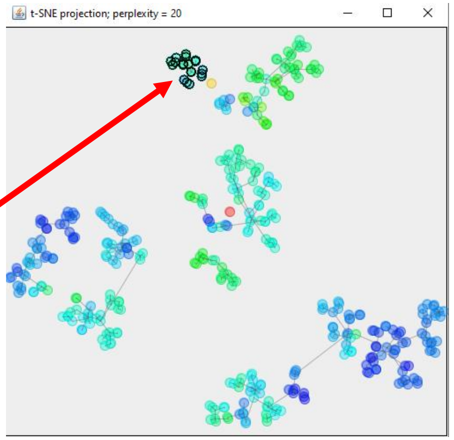


Intervals of feature values → bars

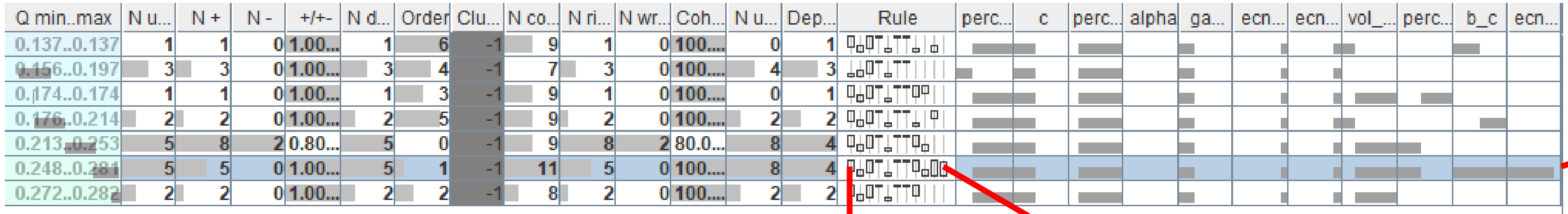
Highly generalised rule set



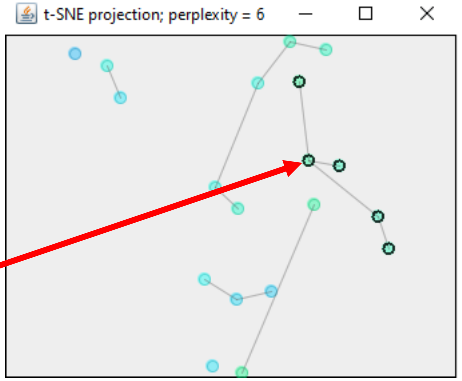
Q min..max	N us...	N +	N -	+/+-	N da...	Order	Clus...	N co...	N rig...	N wr...	Coh...	N uni...	Dept...	Rule	perc...	alpha	ecn_...	c	b_c	ecn...	perc...	perc...	gam...
0.010..0.315	47	47	0	1.00...	47	9	-1	4	46	0	100....	90	10										
0.103..0.103	1	1	0	1.00...	1	10	-1	5	1	0	100....	0	1										
0.000..0.322	138	138	0	1.00...	138	3	-1	1	133	0	100....	264	16										
0.082..0.159	6	6	0	1.00...	6	7	-1	5	6	0	100....	10	5										
0.056..0.361	62	62	0	1.00...	62	4	-1	4	62	0	100....	122	14										
0.021..0.364	61	61	0	1.00...	61	0	-1	3	61	0	100....	120	15										
0.137..0.282	19	19	0	1.00...	19	5	-1	6	19	0	100....	36	10										
0.253..0.398	17	17	0	1.00...	17	2	-1	3	17	0	100....	32	8										
0.156..0.479	49	49	0	1.00...	49	8	-1	5	49	0	100....	96	14										
0.612..0.612	1	1	0	1.00...	1	6	-1	6	1	0	100....	0	1										
0.768..0.768	1	1	0	1.00...	1	1	-1	3	1	0	100....	0	1										

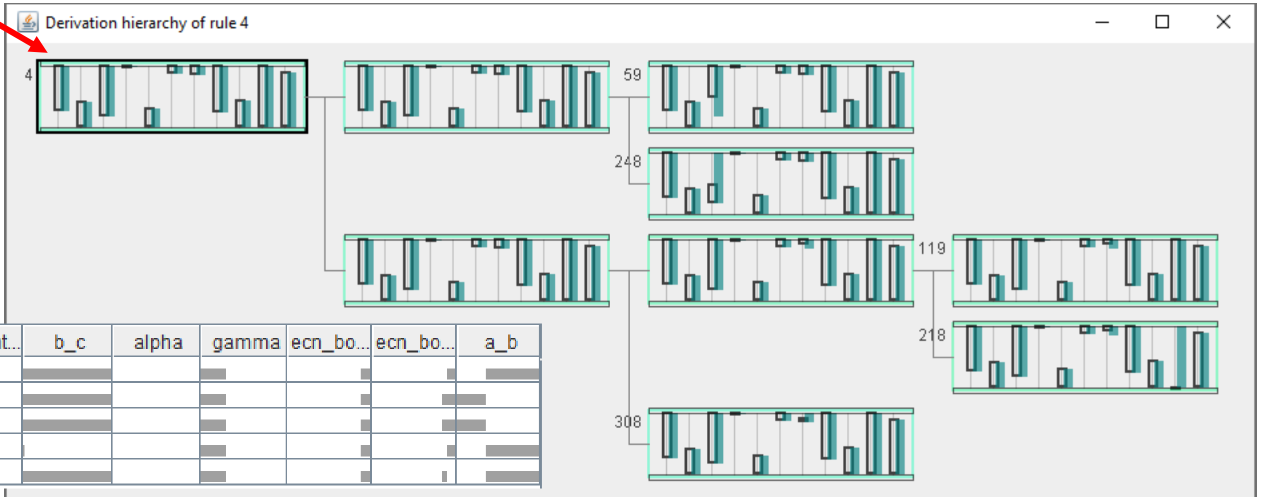


Refinement of a selected generalised rule

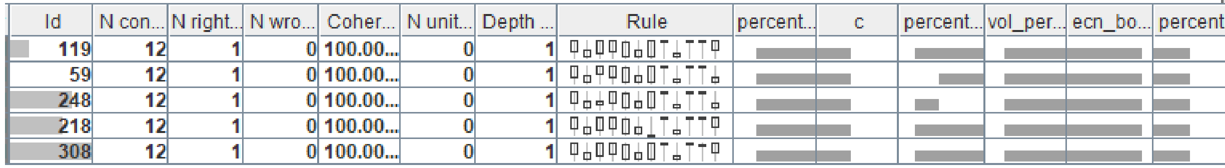


Q min..max	N u...	N +	N -	+/+-	N d...	Order	Clu...	N co...	N ri...	N wr...	Coh...	N u...	Dep...	Rule	perc...	c	perc...	alpha	ga...	ecn...	ecn...	vol_...	perc...	b_c	ecn...
0.137..0.137	1	1	0	1.00...	1	6	-1	9	1	0	100....	0	1												
0.156..0.197	3	3	0	1.00...	3	4	-1	7	3	0	100....	4	3												
0.174..0.174	1	1	0	1.00...	1	3	-1	9	1	0	100....	0	1												
0.176..0.214	2	2	0	1.00...	2	5	-1	9	2	0	100....	2	2												
0.213..0.253	5	8	2	0.80...	5	0	-1	9	8	2	80.0...	8	4												
0.248..0.281	5	5	0	1.00...	5	1	-1	11	5	0	100....	8	4												
0.272..0.282	2	2	0	1.00...	2	2	-1	8	2	0	100....	2	2												





Viewing original rules included in the selected rule



Id	N con...	N right...	N wro...	Coher...	N unit...	Depth ...	Rule	percent...	c	percent...	vol_per...	ecn_bo...	percent...	b_c	alpha	gamma	ecn_bo...	ecn_bo...	a_b
119	12	1	0	100.00...	0	1													
59	12	1	0	100.00...	0	1													
248	12	1	0	100.00...	0	1													
218	12	1	0	100.00...	0	1													
308	12	1	0	100.00...	0	1													

SUMMARY

- Problem: practical incomprehensibility of theoretically interpretable models due to large size and complexity
- Goal: facilitate comprehension by creating a smaller and simpler *descriptive model*
- Approach: iterative aggregation and generalisation
  - Rough rules with exceptions
  - Approximate rules for representing regression models

👎 Limited degree of simplification for models optimised for compactness

☝ Required: domain semantics-based grouping and aggregation of features

Our current work:

SEEKING MORE SCALABLE APPROACHES TO SUPPORTING INTERACTIVE VISUAL EXPLORATION OF LARGE RULE-BASED ML MODELS

Such as rules extracted from Random Forest models

RANDOM FOREST

- Ensemble learning method that creates many decision trees.
- Based on the principle of "wisdom of the crowd" – ***multiple weak learners*** form a strong predictor.

<u>How it works</u>:

- Each tree is trained on a random subset of the data (bagging).
- At each split, a random subset of features is considered.
- Final prediction is made by **majority vote** (classification) or **average** (regression).

Key features:

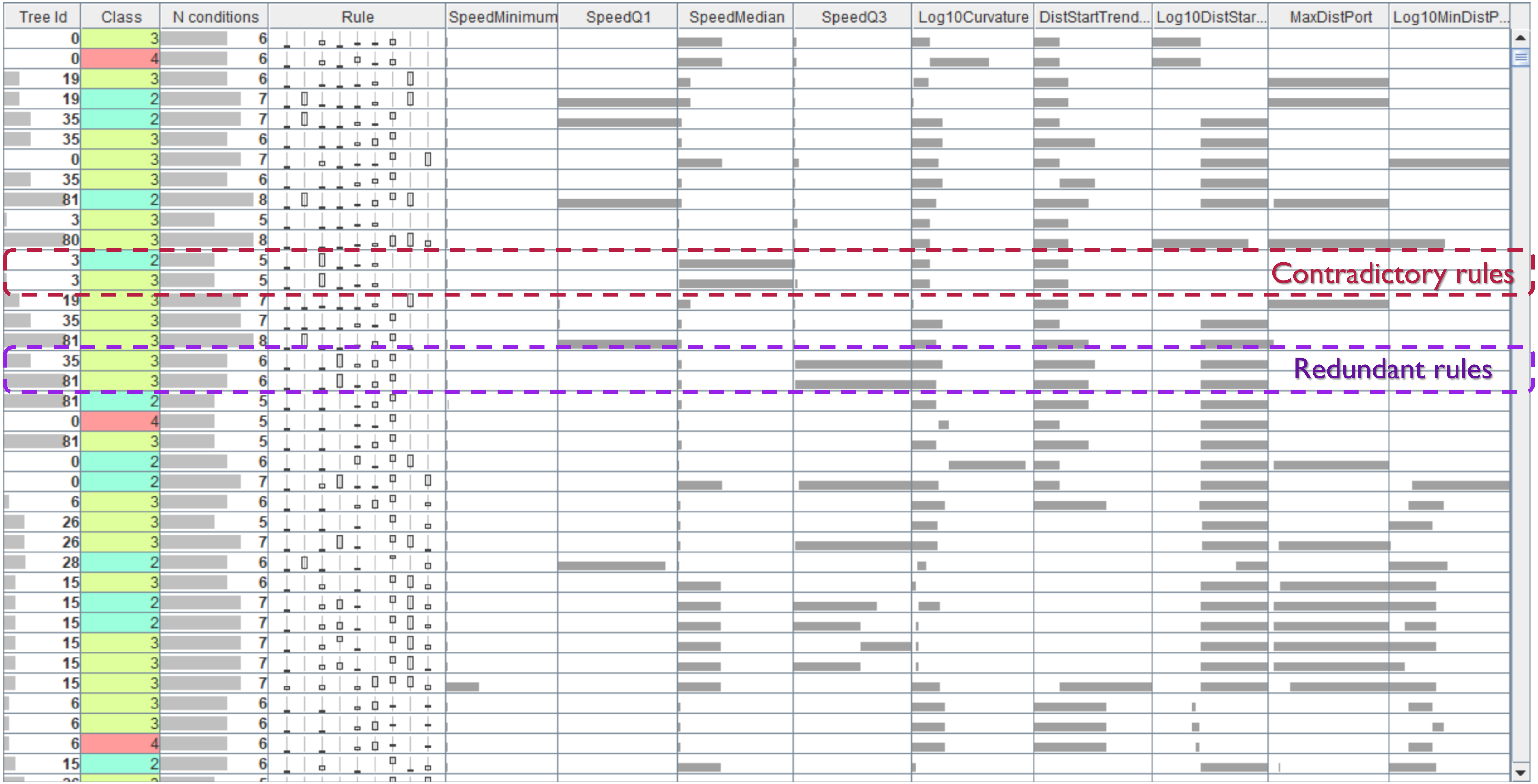
👍 Handles non-linear relationships and high-dimensional data well.

👍 Robust to overfitting due to randomization and averaging.

👎 Still a **black-box model** – hard to interpret due to large size, redundancy, and possible inconsistencies.

RUNNING EXAMPLE: VESSEL MOVEMENT PATTERN RECOGNITION

- **Goal:** Classify vessel movement segments into behavior types (class 1 – *Forward movement*, class 2 – *Trawling*, class 3 – *Port enter/exit*, class 4 – *Anchoring*)
- **Data:** Segments of vessel trajectories derived from AIS (Automatic Identification System) records and described by engineered *time interval-based* **features**:
  - *SpeedMinimum, SpeedQ1, SpeedMedian, SpeedQ3* – represent speed distribution over time interval.
  - *Log10Curvature* – logarithm of the curvature of the time series of the vessel's distance from the starting point computed as the ratio between the sum of absolute consecutive changes and the amplitude of values.
  - *DistStartTrendAngle, Log10DistStartTrendDevAmplitude* – angle of the linear trend fitted to the time series of the vessel's distance from the starting point and logarithm of the amplitude of deviations from the trend line.
  - *MaxDistPort, Log10MinDistPort* –maximum and log-transformed minimum distance from the nearest port.
- **Model:** Random Forest classifier transformed into a **rule-based model** for interpretability
  - **100** decision trees transformed to **9,939** rules with **56,838** conditions in total

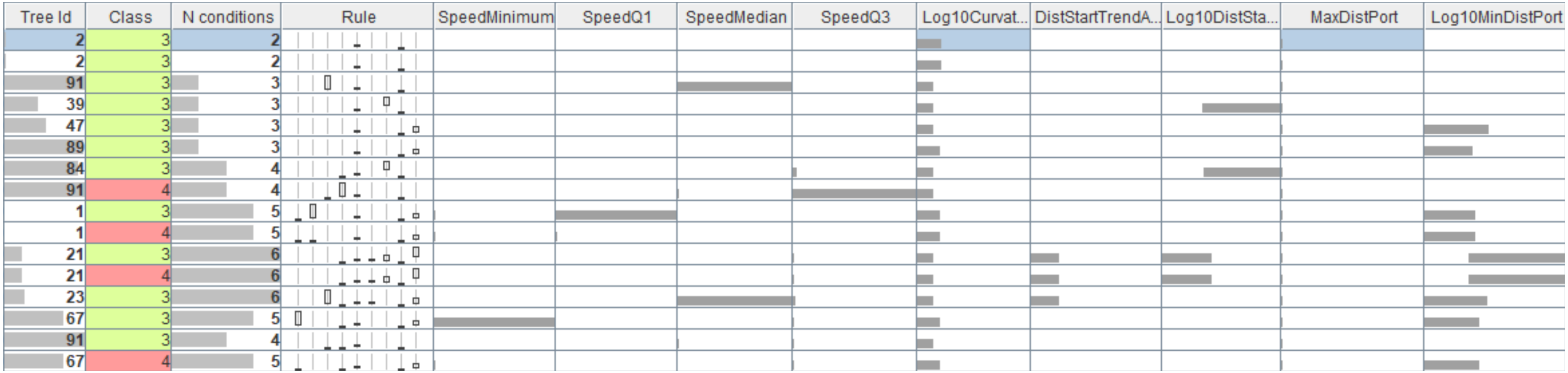


Tree Id	Class	N conditions	Rule	SpeedMinimum	SpeedQ1	SpeedMedian	SpeedQ3	Log10Curvature	DistStartTrend...	Log10DistStar...	MaxDistPort	Log10MinDistP...
0	3	6										
0	4	6										
19	3	6										
19	2	7										
35	2	7										
35	3	6										
0	3	7										
35	3	6										
81	2	8										
3	3	5										
80	3	8										
3	2	5										
3	3	5										
19	3	7										
35	3	7										
81	3	8										
35	3	6										
81	3	6										
81	2	5										
0	4	5										
81	3	5										
0	2	6										
0	2	7										
6	3	6										
26	3	5										
26	3	7										
28	2	6										
15	3	6										
15	2	7										
15	2	7										
15	3	7										
15	3	7										
15	3	7										
6	3	6										
6	3	6										
6	4	6										
15	2	6										

Contradictory rules

Redundant rules

AN EXAMPLE OF A CONTRADICTORY RULE



Tree Id	Class	N conditions	Rule	SpeedMinimum	SpeedQ1	SpeedMedian	SpeedQ3	Log10Curvat...	DistStartTrendA...	Log10DistSta...	MaxDistPort	Log10MinDistPort
2	3	2										
2	3	2										
91	3	3										
39	3	3										
47	3	3										
89	3	3										
84	3	4										
91	4	4										
1	3	5										
1	4	5										
21	3	6										
21	4	6										
23	3	6										
67	3	5										
91	3	4										
67	4	5										

The combination of the conditions of the topmost rule is more general than in the remaining rules shown. Some of the remaining rules predict another class than the topmost rule.

INITIAL CLEANING

- Automatically detect and remove contradictory rules – 113 rules removed
- Automatically detect and remove redundant rules (either same as or fully covered by other rules) – 311 rules removed

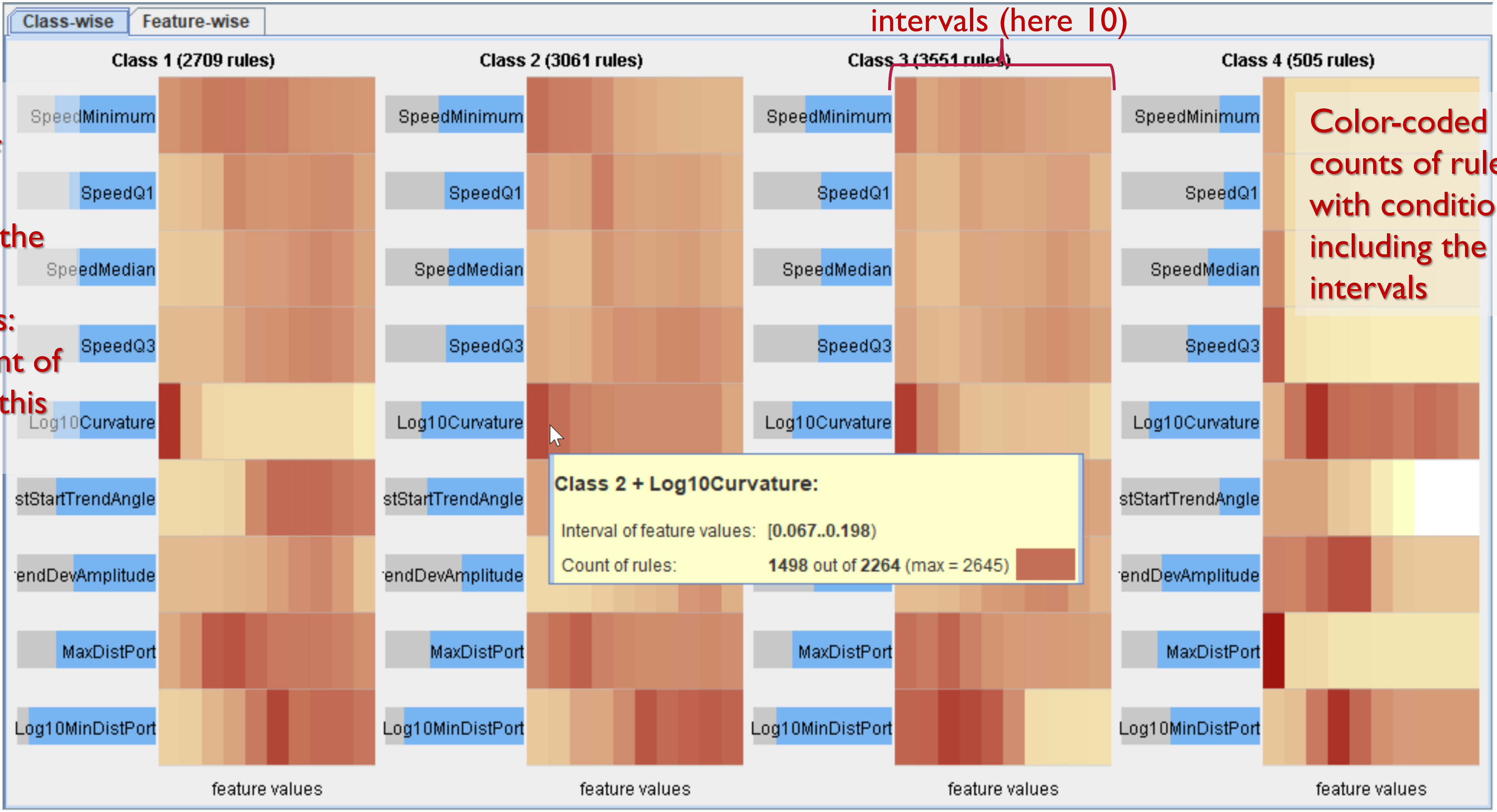
DISTRIBUTIONS OF FEATURES AND THEIR VALUE INTERVALS

Chosen number of intervals (here 10)

Blue bars counts of rules involving the features

Grey bars: total count of rules for this class

Color-coded counts of rules with conditions including the intervals

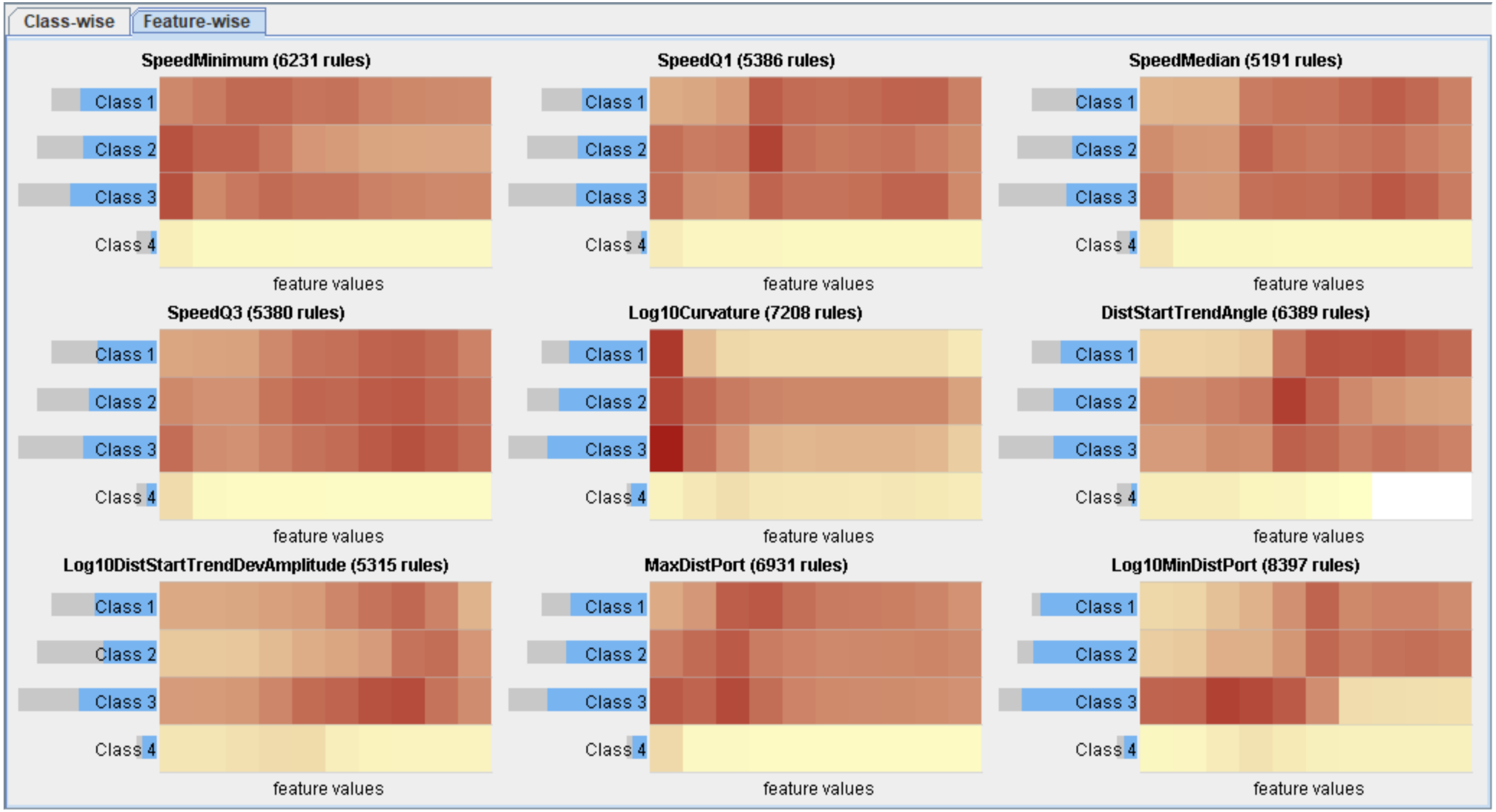


Class 2 + Log10Curvature:

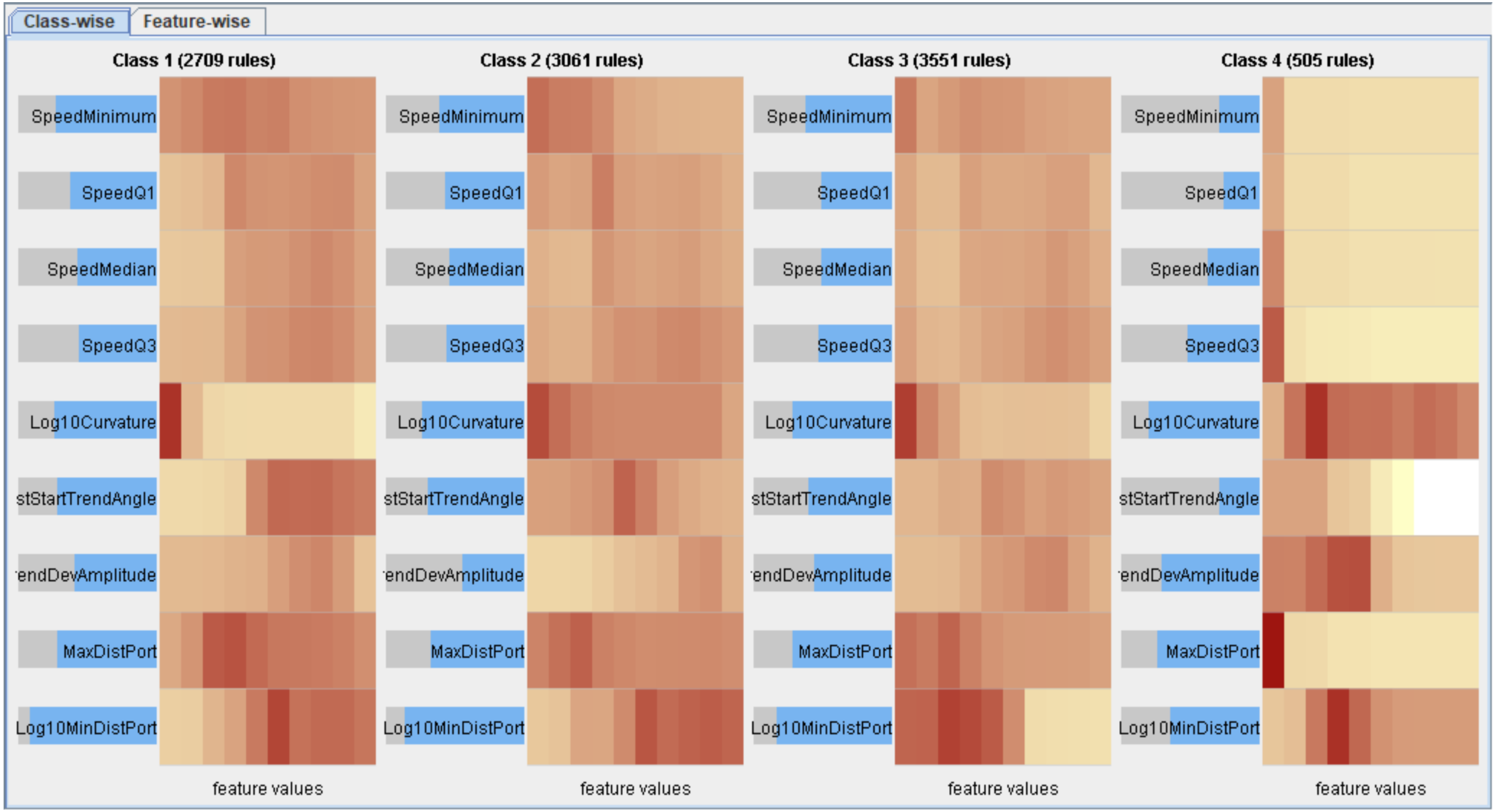
Interval of feature values: **[0.067..0.198)**

Count of rules: **1498** out of **2264** (max = 2645)

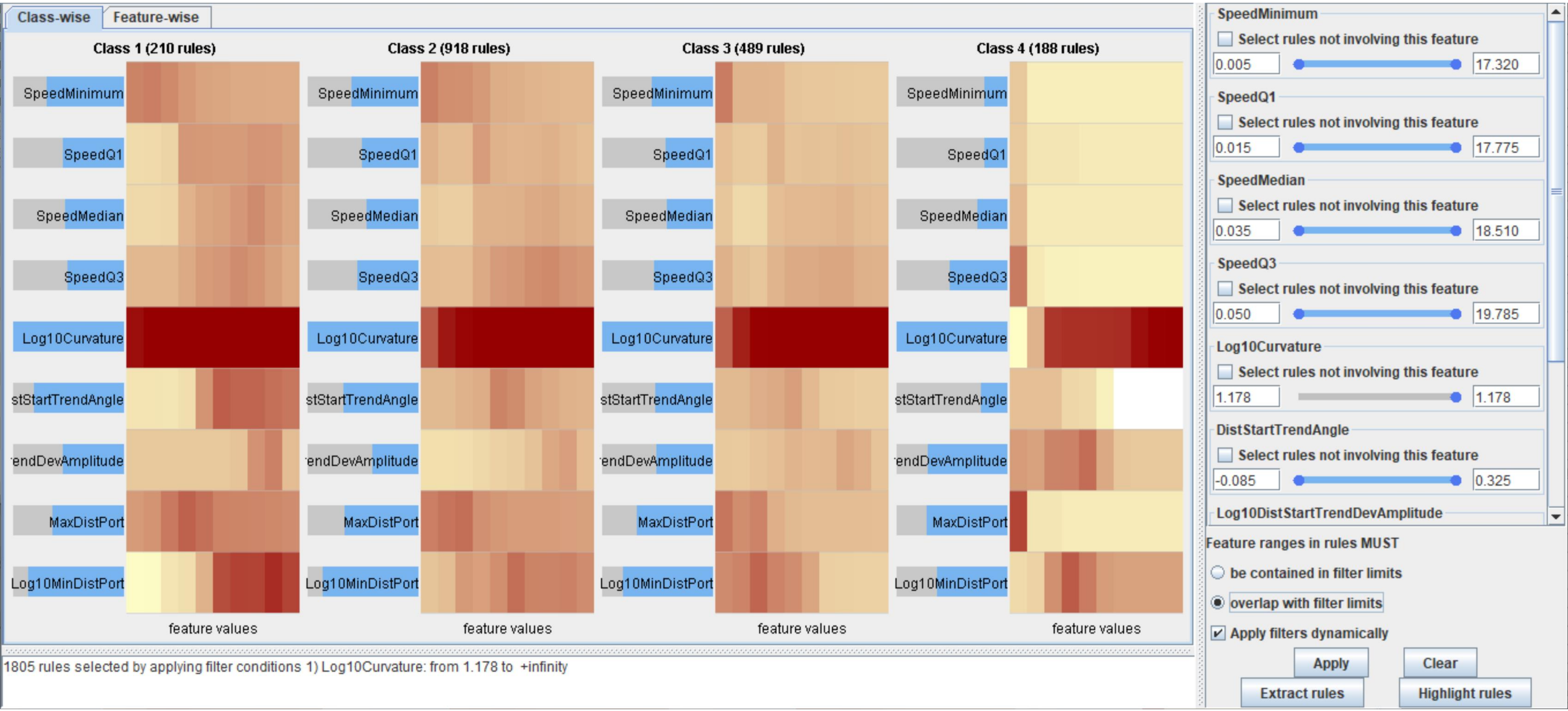
HOW DOES THE MODEL USE THE FEATURES?



DO THE DISTRIBUTIONS ALIGN WITH OUR EXPECTATIONS?



Controls for interactive filtering



1805 rules selected by applying filter conditions 1) Log10Curvature: from 1.178 to +infinity

SpeedMinimum

□ Select rules not involving this feature

0.005 – 17.320

SpeedQ1

□ Select rules not involving this feature

0.015 – 17.775

SpeedMedian

□ Select rules not involving this feature

0.035 – 18.510

SpeedQ3

□ Select rules not involving this feature

0.050 – 19.785

Log10Curvature

□ Select rules not involving this feature

1.178 – 1.178

DistStartTrendAngle

□ Select rules not involving this feature

-0.085 – 0.325

Log10DistStartTrendDevAmplitude

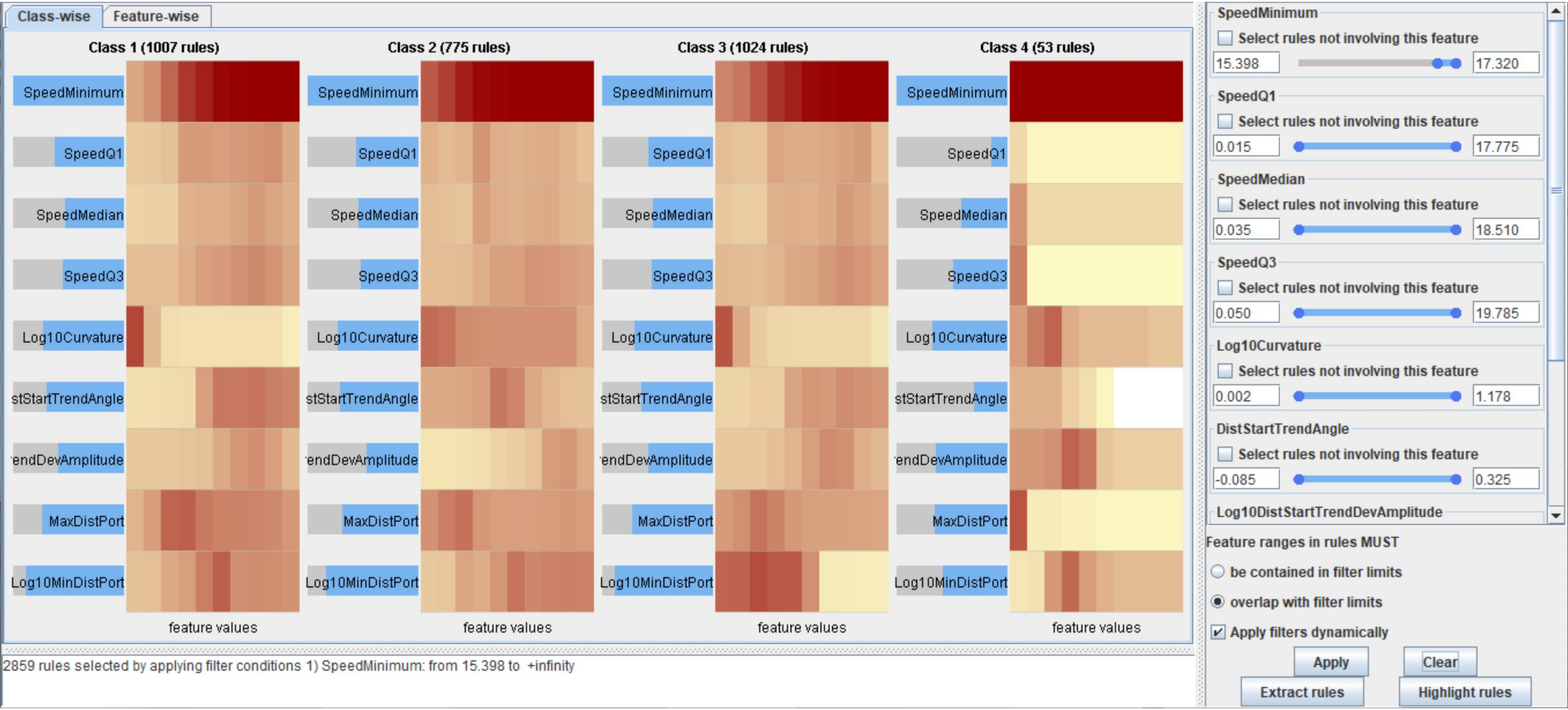
Feature ranges in rules MUST

○ be contained in filter limits

● overlap with filter limits

☑ Apply filters dynamically

There are rules allowing high curvature for forward movement patterns (class 1)



SpeedMinimum
□ Select rules not involving this feature
15.398 – 17.320

SpeedQ1
□ Select rules not involving this feature
0.015 – 17.775

SpeedMedian
□ Select rules not involving this feature
0.035 – 18.510

SpeedQ3
□ Select rules not involving this feature
0.050 – 19.785

Log10Curvature
□ Select rules not involving this feature
0.002 – 1.178

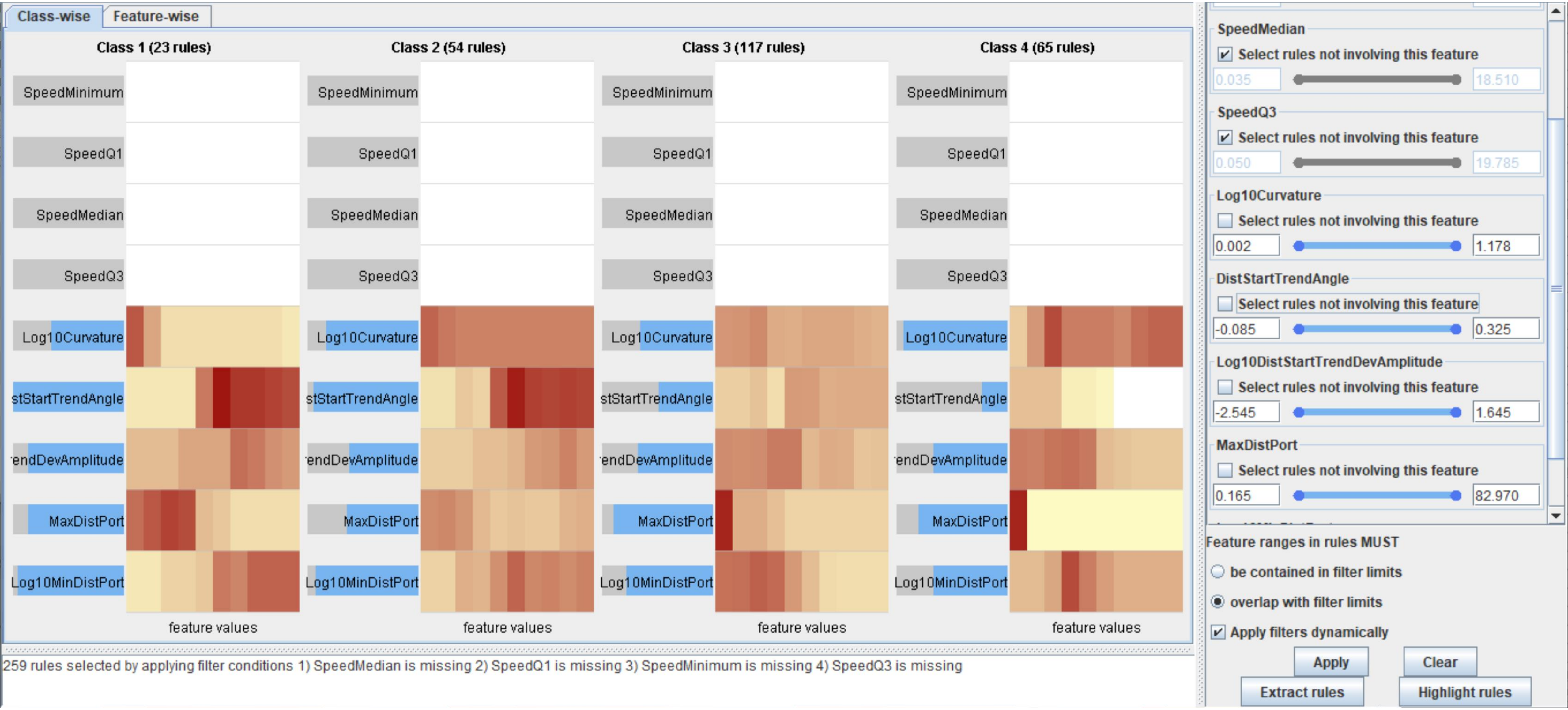
DistStartTrendAngle
□ Select rules not involving this feature
-0.085 – 0.325

Log10DistStartTrendDevAmplitude

Feature ranges in rules MUST
○ be contained in filter limits
● overlap with filter limits
☑ Apply filters dynamically

2859 rules selected by applying filter conditions 1) SpeedMinimum: from 15.398 to +infinity

There are rules allowing high minimal speed for anchoring patterns (class 4)



SpeedMedian

☑ Select rules not involving this feature

0.035 – 18.510

SpeedQ3

☑ Select rules not involving this feature

0.050 – 19.785

Log10Curvature

☐ Select rules not involving this feature

0.002 – 1.178

DistStartTrendAngle

☐ Select rules not involving this feature

-0.085 – 0.325

Log10DistStartTrendDevAmplitude

☐ Select rules not involving this feature

-2.545 – 1.645

MaxDistPort

☐ Select rules not involving this feature

0.165 – 82.970

Feature ranges in rules MUST

○ be contained in filter limits

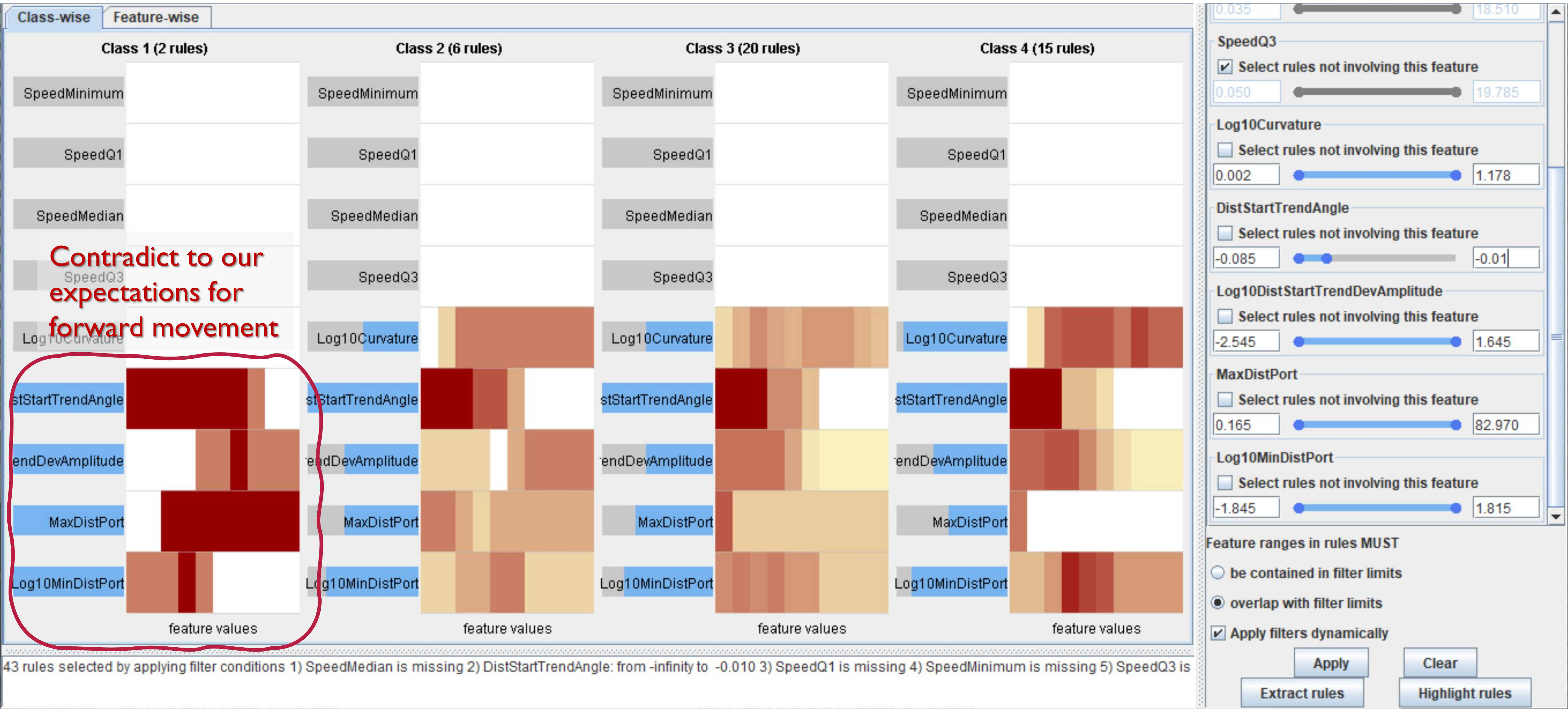
● overlap with filter limits

☑ Apply filters dynamically

259 rules selected by applying filter conditions 1) SpeedMedian is missing 2) SpeedQ1 is missing 3) SpeedMinimum is missing 4) SpeedQ3 is missing

There are rules ignoring the speed distribution features

Further filtering: rules with negative trend angle of the distances from the start



Contradict to our expectations for forward movement

43 rules selected by applying filter conditions 1) SpeedMedian is missing 2) DistStartTrendAngle: from -infinity to -0.010 3) SpeedQ1 is missing 4) SpeedMinimum is missing 5) SpeedQ3 is

0.035 18.510

SpeedQ3
☑ Select rules not involving this feature
0.050 19.785

Log10Curvature
□ Select rules not involving this feature
0.002 1.178

DistStartTrendAngle
□ Select rules not involving this feature
-0.085 -0.01

Log10DistStartTrendDevAmplitude
□ Select rules not involving this feature
-2.545 1.645

MaxDistPort
□ Select rules not involving this feature
0.165 82.970

Log10MinDistPort
□ Select rules not involving this feature
-1.845 1.815

Feature ranges in rules MUST
○ be contained in filter limits
● overlap with filter limits
☑ Apply filters dynamically

INTERACTIVE CLEANING THROUGH FILTERING AND TESTING ON LABELLED DATA

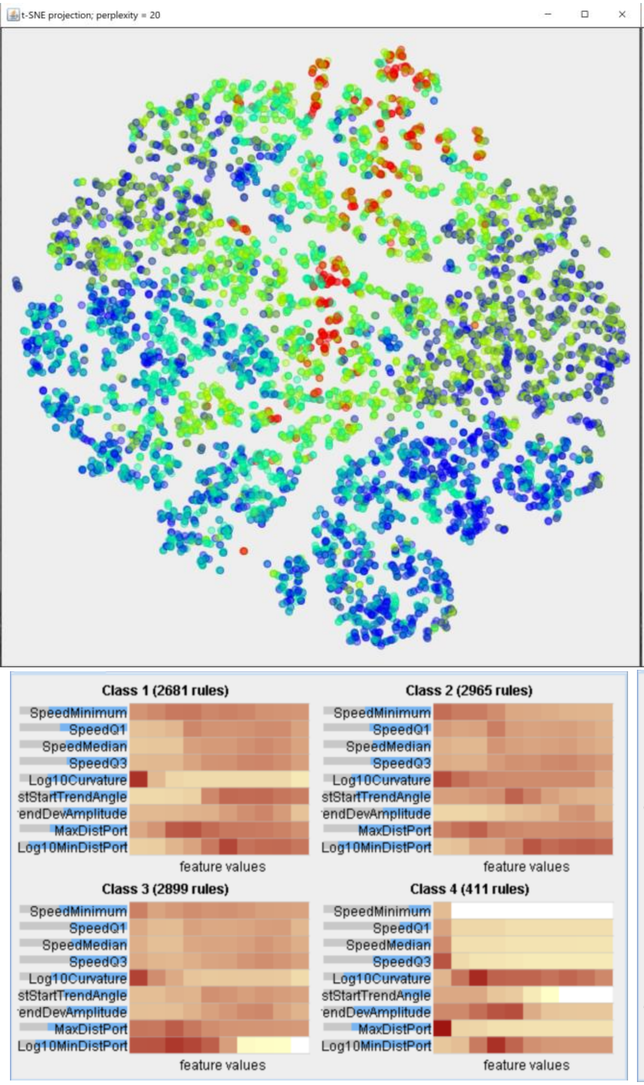
- 2 rules for class 1 (forward movement) ignoring speed distribution and allowing negative distance trend
- 7 rules for class 2 (trawling) that allow high values of *SpeedMinimum*
- 144 rules for class 3 (port entering or exiting) ignoring both *MaxDistPort* and *Log10MinDistPort*
- 353 rules for class 3 (port entering or exiting) not restricting *Log10MinDistPort*
- 53 rules for class 4 (anchoring) that do not limit *SpeedMinimum*

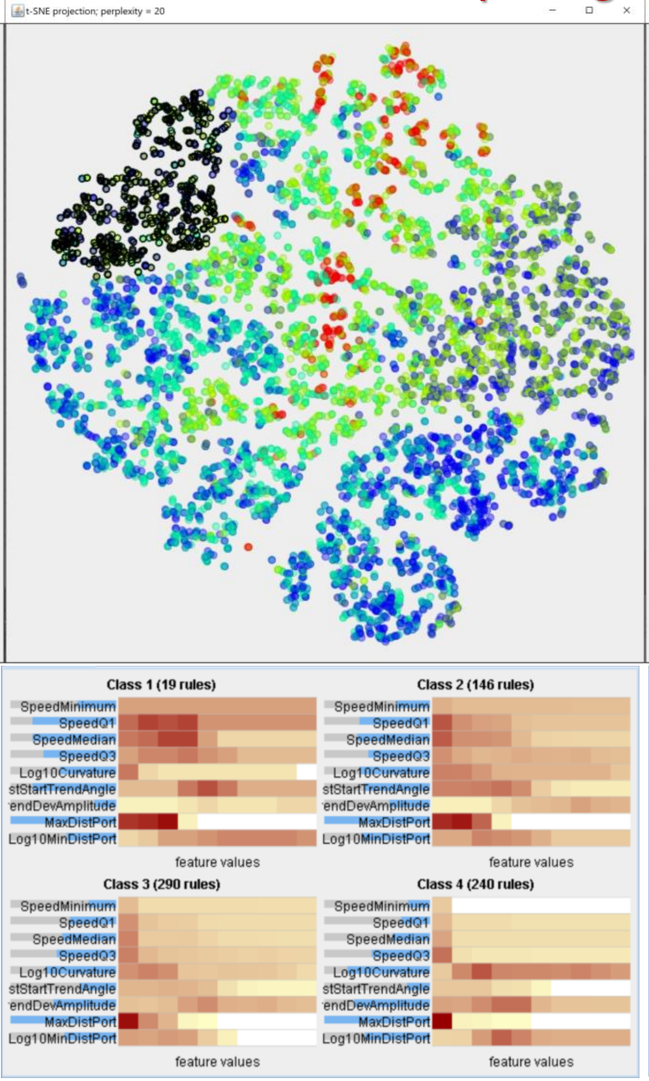
Result:

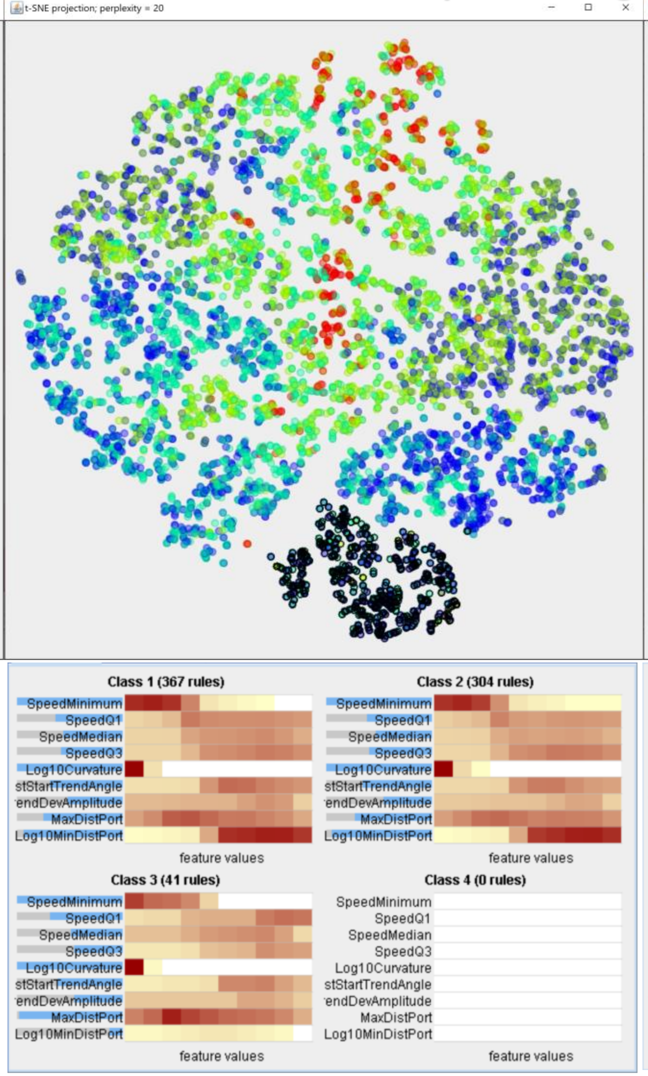
- 8,956 rules, no loss of accuracy

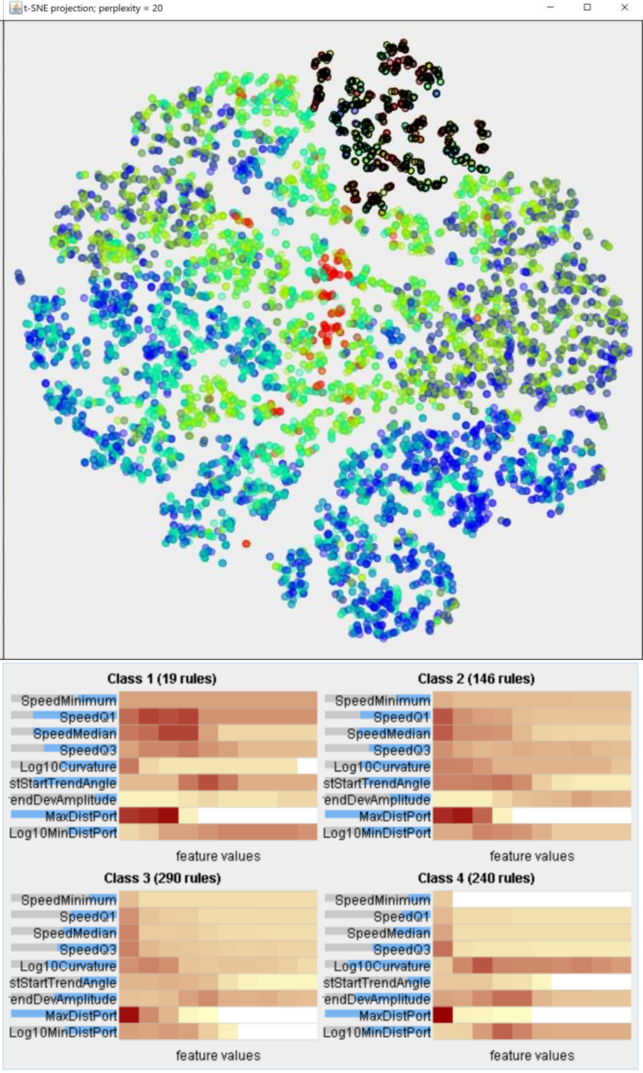
2D PROJECTION OF THE RULES BASED ON CONDITIONS SIMILARITY

Exploring subsets of similar rules predicting distinct classes









SUMMARY: INTERACTIVE RULE EXPLORATION

What we saw:

- Filtering and testing rules reveals inconsistencies that can be removed to improve model logic
- Better alignment with domain knowledge can be achieved without hurting accuracy

But …

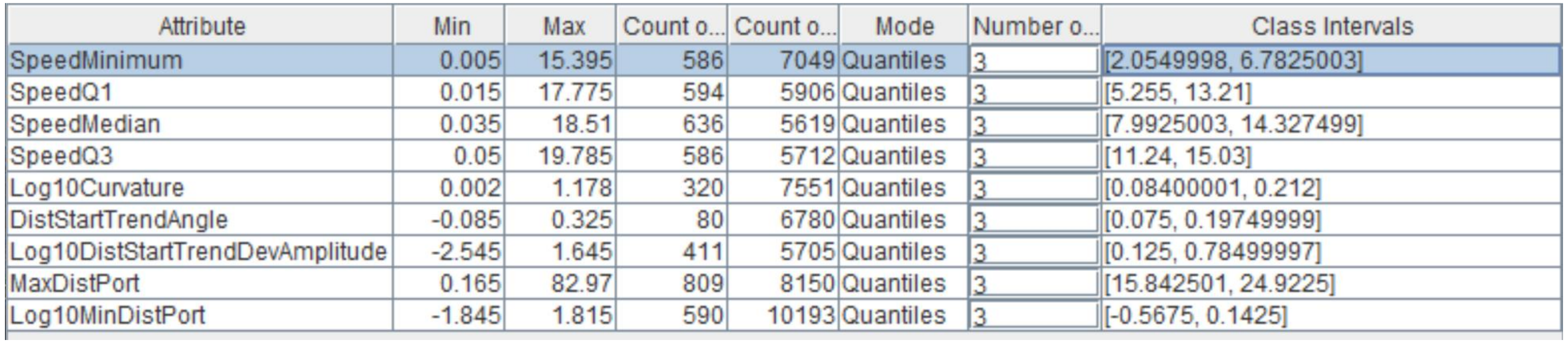
- Interpretability gains come at the cost of expert time and effort

Possible future research direction:

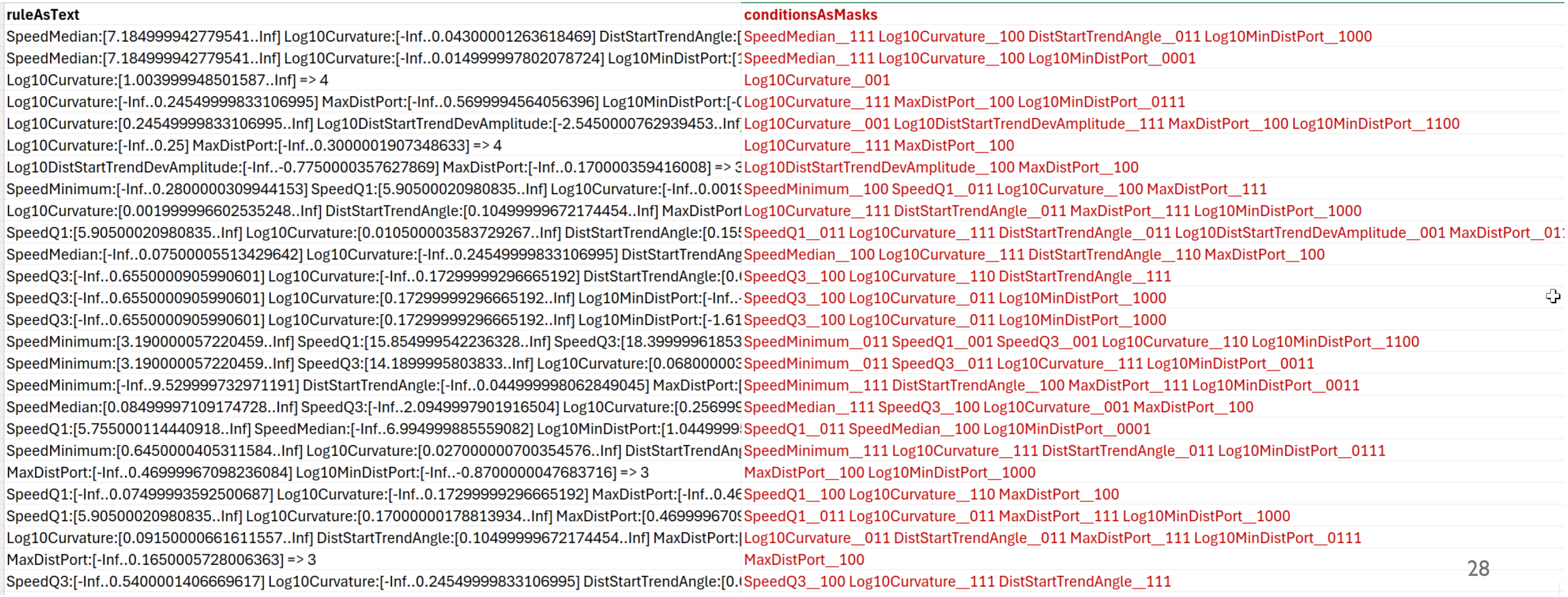
- Develop a smart expert UI to:
  - Define domain constraints
  - Automatically flag rule violations
  - Support semi-automated model refinement

TOPIC MODELLING TO REVEAL FEATURE INTERACTIONS

Encoding the rules:

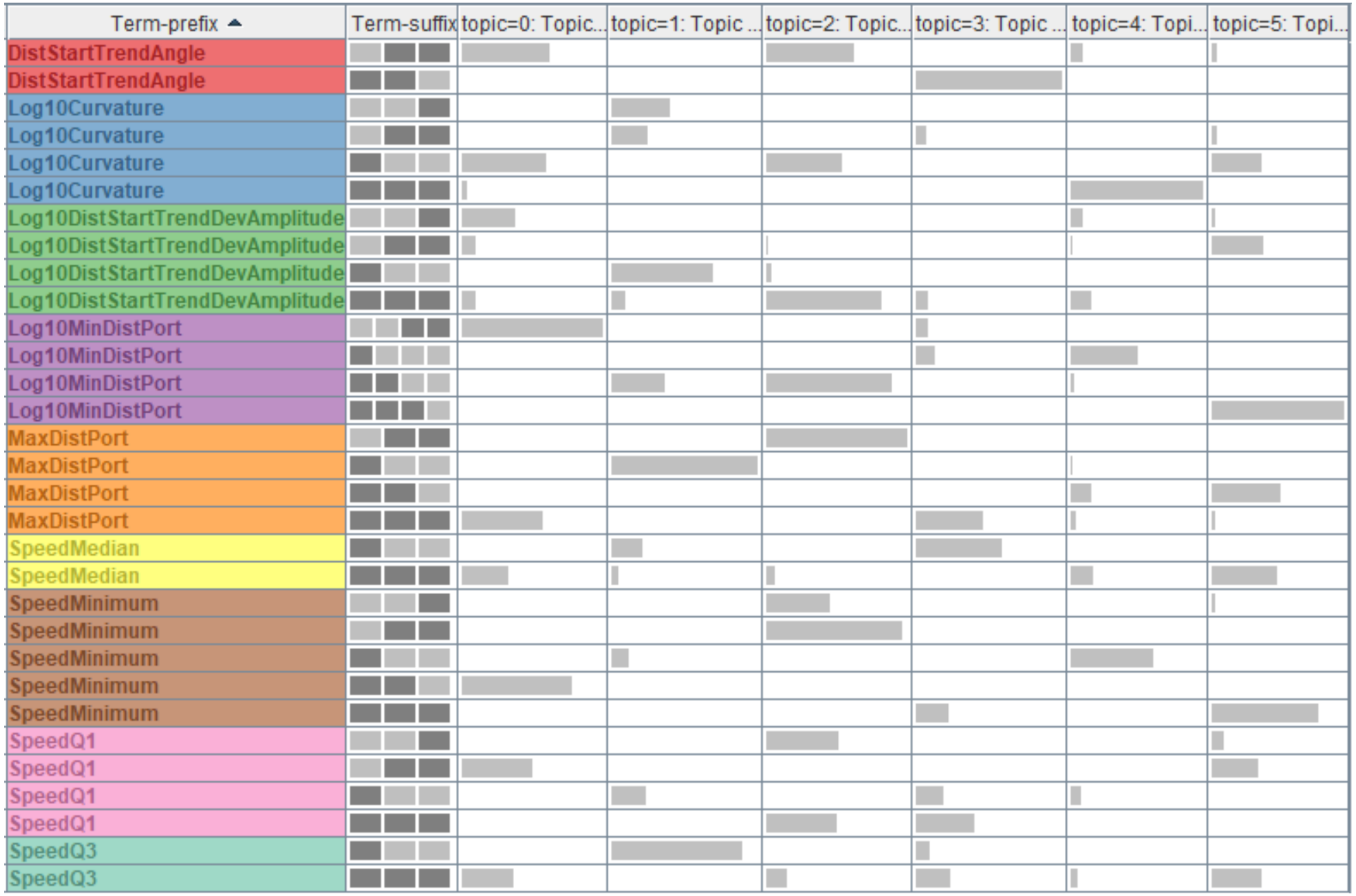


Attribute	Min	Max	Count o...	Count o...	Mode	Number o...	Class Intervals
SpeedMinimum	0.005	15.395	586	7049	Quantiles	3	[2.0549998, 6.7825003]
SpeedQ1	0.015	17.775	594	5906	Quantiles	3	[5.255, 13.21]
SpeedMedian	0.035	18.51	636	5619	Quantiles	3	[7.9925003, 14.327499]
SpeedQ3	0.05	19.785	586	5712	Quantiles	3	[11.24, 15.03]
Log10Curvature	0.002	1.178	320	7551	Quantiles	3	[0.08400001, 0.212]
DistStartTrendAngle	-0.085	0.325	80	6780	Quantiles	3	[0.075, 0.19749999]
Log10DistStartTrendDevAmplitude	-2.545	1.645	411	5705	Quantiles	3	[0.125, 0.78499997]
MaxDistPort	0.165	82.97	809	8150	Quantiles	3	[15.842501, 24.9225]
Log10MinDistPort	-1.845	1.815	590	10193	Quantiles	3	[-0.5675, 0.1425]



ruleAsText	conditionsAsMasks
SpeedMedian:[7.184999942779541..Inf] Log10Curvature:[-Inf..0.04300001263618469] DistStartTrendAngle:[	SpeedMedian__111 Log10Curvature__100 DistStartTrendAngle__011 Log10MinDistPort__1000
SpeedMedian:[7.184999942779541..Inf] Log10Curvature:[-Inf..0.014999997802078724] Log10MinDistPort:[1	SpeedMedian__111 Log10Curvature__100 Log10MinDistPort__0001
Log10Curvature:[1.003999948501587..Inf] => 4	Log10Curvature__001
Log10Curvature:[-Inf..0.24549999833106995] MaxDistPort:[-Inf..0.5699994564056396] Log10MinDistPort:[-(	Log10Curvature__111 MaxDistPort__100 Log10MinDistPort__0111
Log10Curvature:[0.24549999833106995..Inf] Log10DistStartTrendDevAmplitude:[-2.5450000762939453..Inf	Log10Curvature__001 Log10DistStartTrendDevAmplitude__111 MaxDistPort__100 Log10MinDistPort__1100
Log10Curvature:[-Inf..0.25] MaxDistPort:[-Inf..0.3000001907348633] => 4	Log10Curvature__111 MaxDistPort__100
Log10DistStartTrendDevAmplitude:[-Inf..-0.7750000357627869] MaxDistPort:[-Inf..0.170000359416008] => 3	Log10DistStartTrendDevAmplitude__100 MaxDistPort__100
SpeedMinimum:[-Inf..0.2800000309944153] SpeedQ1:[5.90500020980835..Inf] Log10Curvature:[-Inf..0.0019	SpeedMinimum__100 SpeedQ1__011 Log10Curvature__100 MaxDistPort__111
Log10Curvature:[0.0019999996602535248..Inf] DistStartTrendAngle:[0.10499999672174454..Inf] MaxDistPort	Log10Curvature__111 DistStartTrendAngle__011 MaxDistPort__111 Log10MinDistPort__1000
SpeedQ1:[5.90500020980835..Inf] Log10Curvature:[0.010500003583729267..Inf] DistStartTrendAngle:[0.155	SpeedQ1__011 Log10Curvature__111 DistStartTrendAngle__011 Log10DistStartTrendDevAmplitude__001 MaxDistPort__01
SpeedMedian:[-Inf..0.07500005513429642] Log10Curvature:[-Inf..0.24549999833106995] DistStartTrendAng	SpeedMedian__100 Log10Curvature__111 DistStartTrendAngle__110 MaxDistPort__100
SpeedQ3:[-Inf..0.6550000905990601] Log10Curvature:[-Inf..0.17299999296665192] DistStartTrendAngle:[0.(	SpeedQ3__100 Log10Curvature__110 DistStartTrendAngle__111
SpeedQ3:[-Inf..0.6550000905990601] Log10Curvature:[0.17299999296665192..Inf] Log10MinDistPort:[-Inf..-	SpeedQ3__100 Log10Curvature__011 Log10MinDistPort__1000
SpeedQ3:[-Inf..0.6550000905990601] Log10Curvature:[0.17299999296665192..Inf] Log10MinDistPort:[-1.61	SpeedQ3__100 Log10Curvature__011 Log10MinDistPort__1000
SpeedMinimum:[3.190000057220459..Inf] SpeedQ1:[15.854999542236328..Inf] SpeedQ3:[18.39999961853	SpeedMinimum__011 SpeedQ1__001 SpeedQ3__001 Log10Curvature__110 Log10MinDistPort__1100
SpeedMinimum:[3.190000057220459..Inf] SpeedQ3:[14.1899995803833..Inf] Log10Curvature:[0.068000003	SpeedMinimum__011 SpeedQ3__011 Log10Curvature__111 Log10MinDistPort__0011
SpeedMinimum:[-Inf..9.529999732971191] DistStartTrendAngle:[-Inf..0.044999998062849045] MaxDistPort:[	SpeedMinimum__111 DistStartTrendAngle__100 MaxDistPort__111 Log10MinDistPort__0011
SpeedMedian:[0.08499997109174728..Inf] SpeedQ3:[-Inf..2.0949997901916504] Log10Curvature:[0.256999	SpeedMedian__111 SpeedQ3__100 Log10Curvature__001 MaxDistPort__100
SpeedQ1:[5.755000114440918..Inf] SpeedMedian:[-Inf..6.994999885559082] Log10MinDistPort:[1.0449999	SpeedQ1__011 SpeedMedian__100 Log10MinDistPort__0001
SpeedMinimum:[0.6450000405311584..Inf] Log10Curvature:[0.027000000700354576..Inf] DistStartTrendAn	SpeedMinimum__111 Log10Curvature__111 DistStartTrendAngle__011 Log10MinDistPort__0111
MaxDistPort:[-Inf..0.46999967098236084] Log10MinDistPort:[-Inf..-0.8700000047683716] => 3	MaxDistPort__100 Log10MinDistPort__1000
SpeedQ1:[-Inf..0.07499993592500687] Log10Curvature:[-Inf..0.17299999296665192] MaxDistPort:[-Inf..0.46	SpeedQ1__100 Log10Curvature__110 MaxDistPort__100
SpeedQ1:[5.90500020980835..Inf] Log10Curvature:[0.17000000178813934..Inf] MaxDistPort:[0.469999670	SpeedQ1__011 Log10Curvature__011 MaxDistPort__111 Log10MinDistPort__1000
Log10Curvature:[0.09150000661611557..Inf] DistStartTrendAngle:[0.10499999672174454..Inf] MaxDistPort:[	Log10Curvature__011 DistStartTrendAngle__011 MaxDistPort__111 Log10MinDistPort__0111
MaxDistPort:[-Inf..0.1650005728006363] => 3	MaxDistPort__100
SpeedQ3:[-Inf..0.5400001406669617] Log10Curvature:[-Inf..0.24549999833106995] DistStartTrendAngle:[0.(	SpeedQ3__100 Log10Curvature__111 DistStartTrendAngle__111

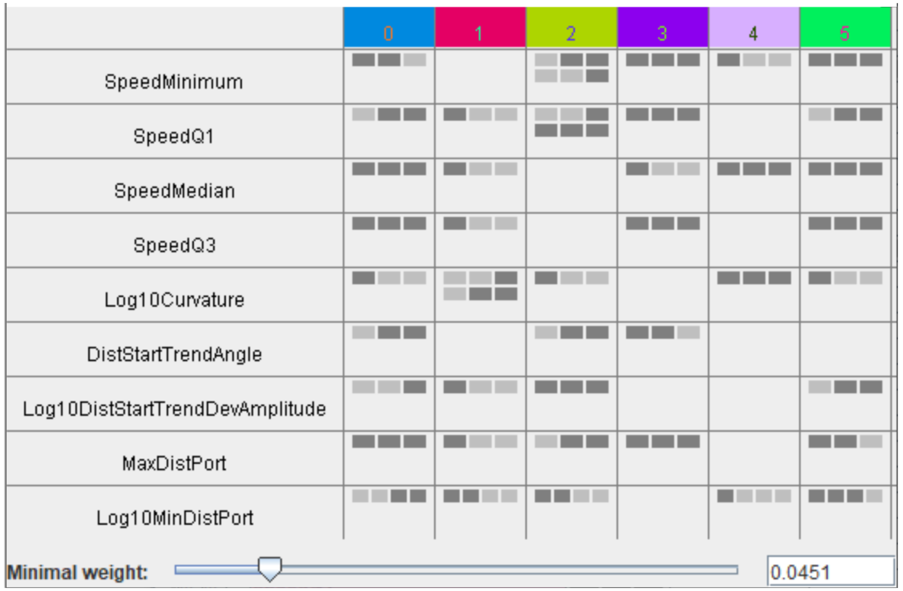
TOPICS



Term-prefix ▲	Term-suffix	topic=0: Topic...	topic=1: Topic ...	topic=2: Topic...	topic=3: Topic ...	topic=4: Topi...	topic=5: Topi...
DistStartTrendAngle							
DistStartTrendAngle							
Log10Curvature							
Log10Curvature							
Log10Curvature							
Log10Curvature							
Log10DistStartTrendDevAmplitude							
Log10DistStartTrendDevAmplitude							
Log10DistStartTrendDevAmplitude							
Log10DistStartTrendDevAmplitude							
Log10MinDistPort							
Log10MinDistPort							
Log10MinDistPort							
Log10MinDistPort							
MaxDistPort							
MaxDistPort							
MaxDistPort							
MaxDistPort							
SpeedMedian							
SpeedMedian							
SpeedMinimum							
SpeedMinimum							
SpeedMinimum							
SpeedMinimum							
SpeedMinimum							
SpeedQ1							
SpeedQ1							
SpeedQ1							
SpeedQ1							
SpeedQ3							
SpeedQ3							

Topics defined as vectors of weights of the "terms", i.e., encoded conditions

Most significant "terms" defining the topics:

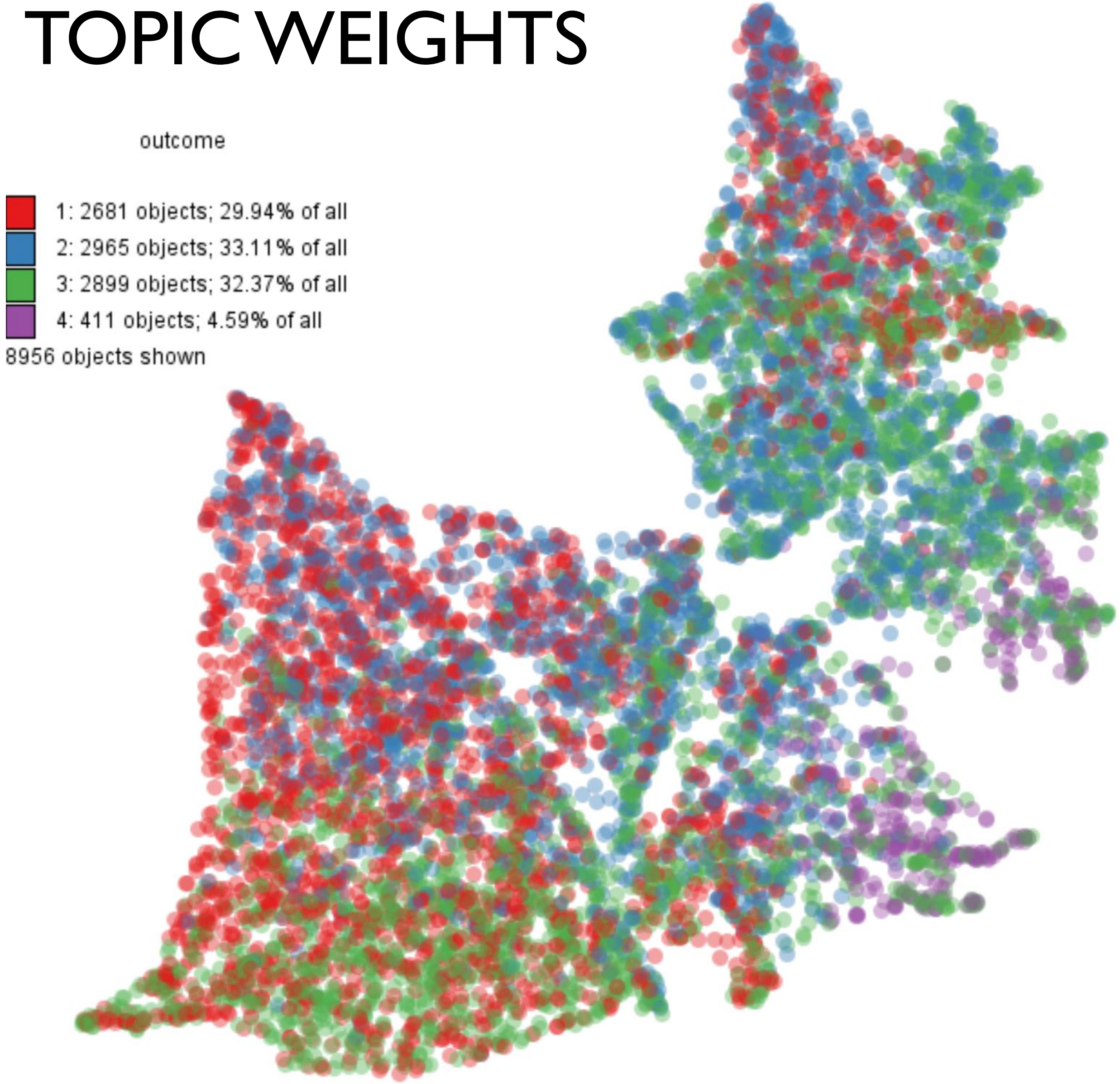


	0	1	2	3	4	5
SpeedMinimum						
SpeedQ1						
SpeedMedian						
SpeedQ3						
Log10Curvature						
DistStartTrendAngle						
Log10DistStartTrendDevAmplitude						
MaxDistPort						
Log10MinDistPort						

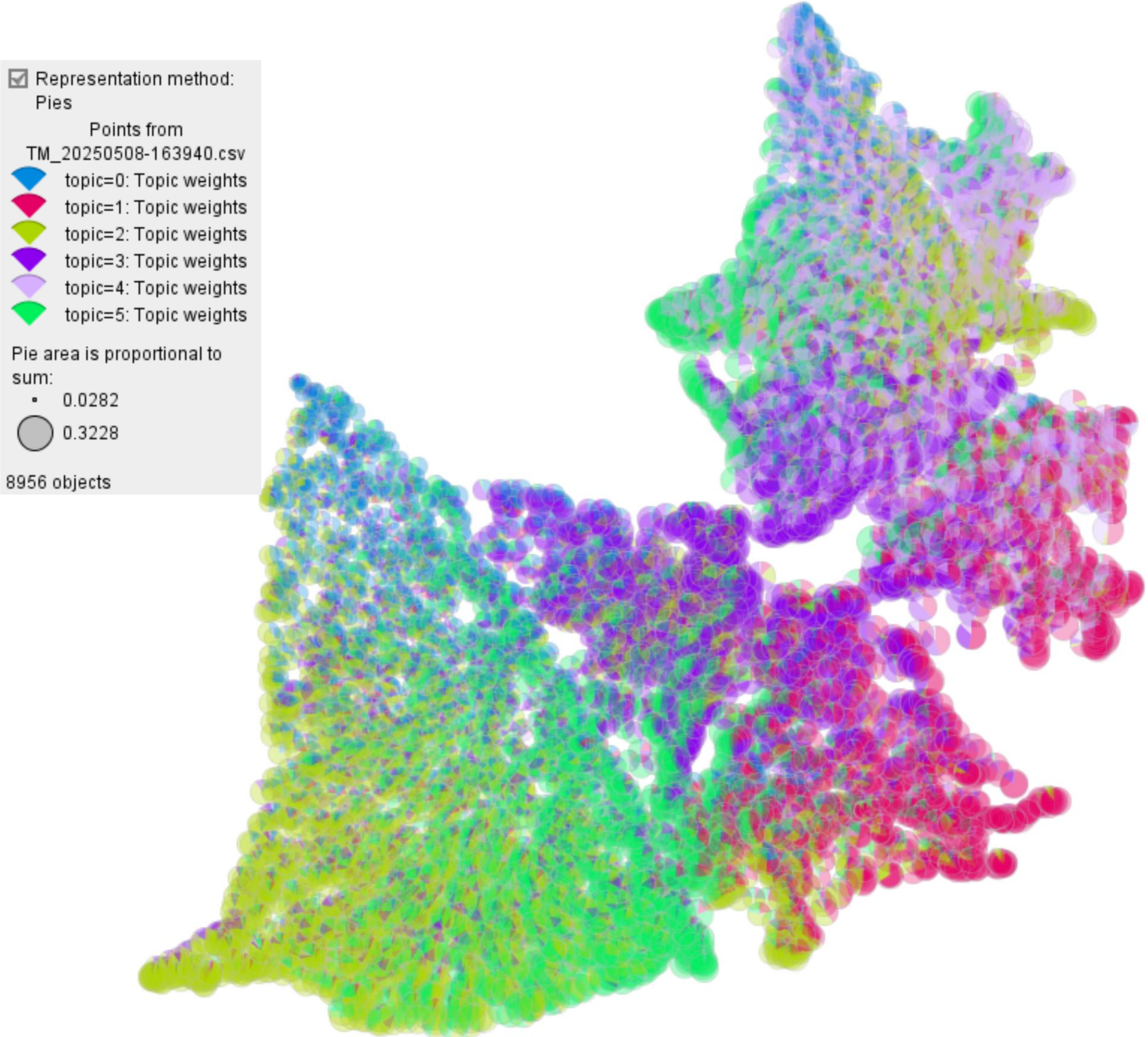
Minimal weight: 0.0451

Topics represent re-occurring combinations of similar conditions

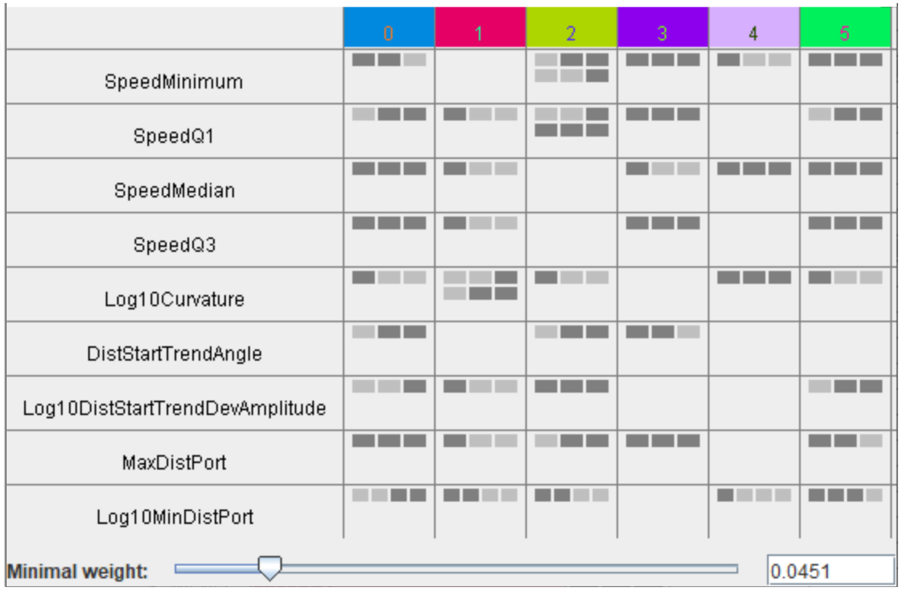
APPLYING DIMENSIONALITY REDUCTION TO VECTORS OF TOPIC WEIGHTS



- Each rule receives a multidimensional vector of topic weights.
- We apply dimensionality reduction (e.g., UMAP) to obtain a 2D projection.
- Each rule is represented by a point in the projection space.
- We represent the rule outcomes (predicted classes) by colours of dot marks.
- We see that the classes are not separated by the topic weights.



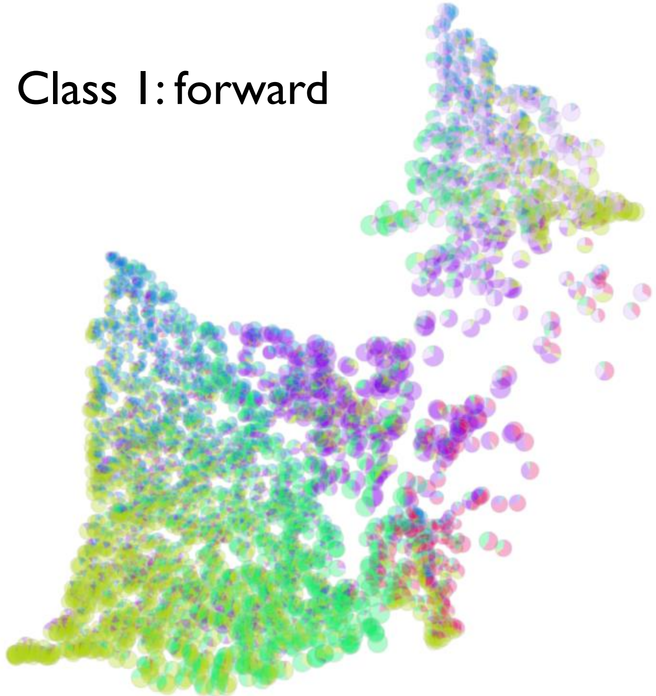
- Pie charts represent compositions of topics characterizing the rules.
- In the projection map we see areas (= groups of rules) dominated by specific topics.



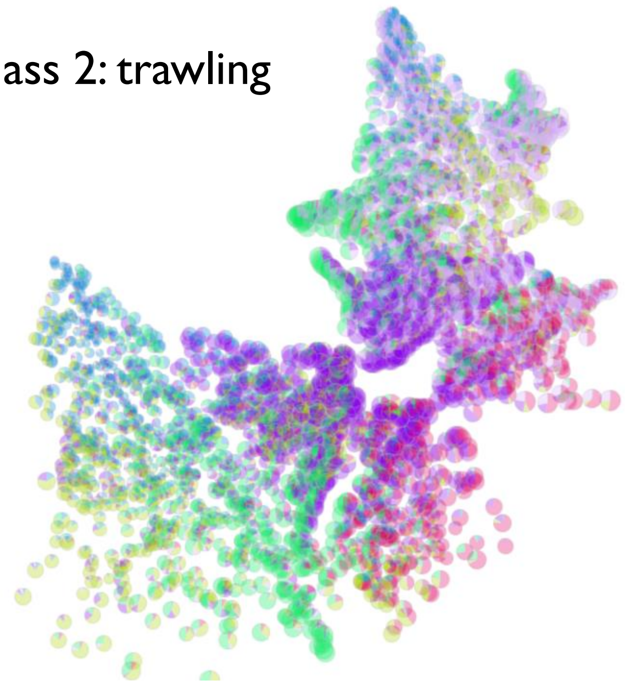
	0	1	2	3	4	5
SpeedMinimum	■■□		□■■ □□■	■■■	■□□	■■■
SpeedQ1	□■■	■□□	□□■ ■■■	■■■		□■■
SpeedMedian	■■■	■□□		■□□	■■■	■■■
SpeedQ3	■■■	■□□		■■■		■■■
Log10Curvature	■□□	□□■ □■■	■□□		■■■	■□□
DistStartTrendAngle	□■■		□■■	■■□		
Log10DistStartTrendDevAmplitude	□□■	■□□	■■■			□■■
MaxDistPort	■■■	■□□	□■■	■■■		■■□
Log10MinDistPort	□□■■	■■□□	■■□□		■□□□	■■■□

Minimal weight: 0.0451

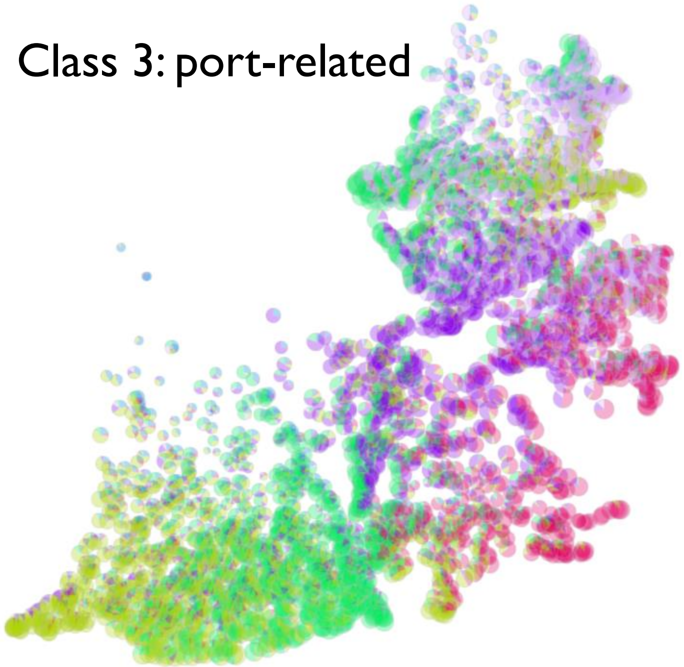
Class 1: forward



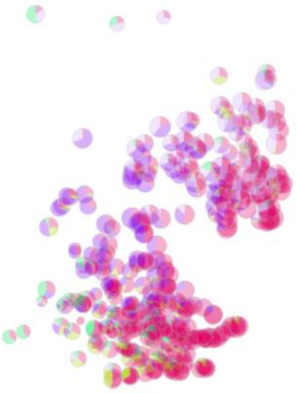
Class 2: trawling



Class 3: port-related

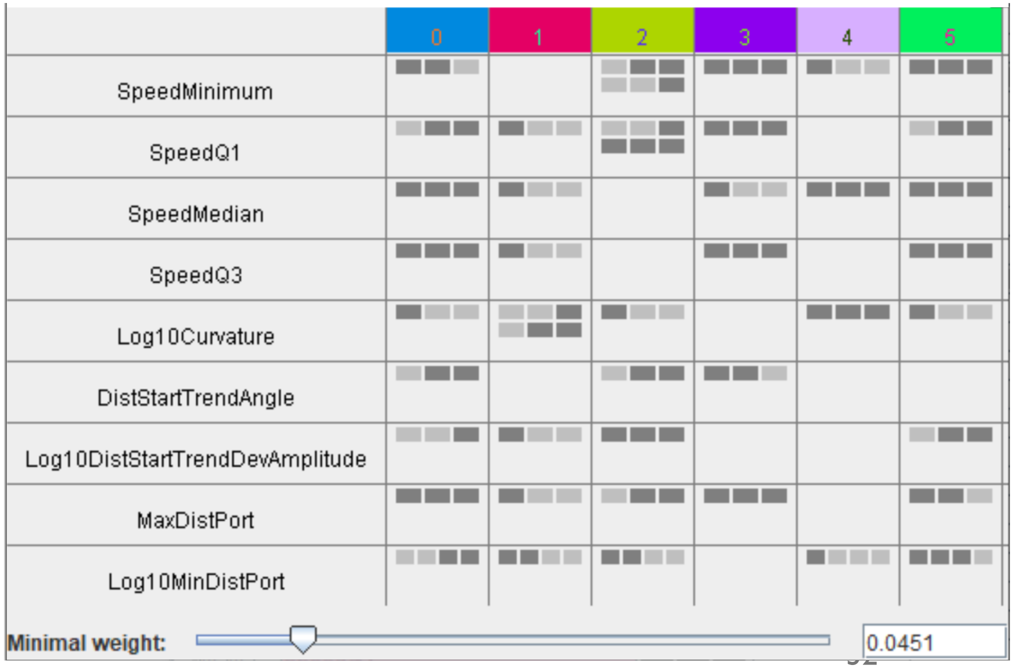


Class 4: anchoring



Topic association tendencies:

- Topic 0: forward movement and trawling
- Topic 1: high association with anchoring, weak association with trawling and port-related
- Topic 2: forward movement and port-related
- Topic 3: medium association with trawling but also co-occurs with the other classes
- Topic 4: medium association with trawling but co-occurs with the others
- Topic 5: medium association with port-related and forward movement, less with trawling



	0	1	2	3	4	5
SpeedMinimum	■■□		□■■ □□■	■■■	■□□	■■■
SpeedQ1	□■■	■□□	□□■ ■■■	■■■		□■■
SpeedMedian	■■■	■□□		■□□	■■■	■■■
SpeedQ3	■■■	■□□		■■■		■■■
Log10Curvature	■□□	□□■ □■■	■□□		■■■	■□□
DistStartTrendAngle	□■■		□■■	■■□		
Log10DistStartTrendDevAmplitude	□□■	■□□	■■■			□■■
MaxDistPort	■■■	■□□	□■■	■■■		■■□
Log10MinDistPort	□□■■	■■□□	■■□□		■□□□	■■■□

Minimal weight: 0.0451

SUMMARY: TOPIC MODELING FOR RULE ANALYSIS

- **No simple pattern**: Except for Class 4, classes are not defined by distinct recurring rule conditions
  - Random forests lack interpretable, consistent class definitions
- What Topic Modelling Adds:
  - Reveals feature interdependencies and co-occurring conditions
  - Helps to see key feature interactions in rule subsets
  - Provides a common space for comparing all rules enabling 2D projections, clustering, and identification of subgroups
- **Takeaway**:
  - Topic modelling complements visual filtering and rule inspection
  - Supports higher-level understanding beyond individual rules

CONCLUSIONS: KEY INSIGHTS & CONTRIBUTIONS

- **Focus on logical consistency** of rule-based models with respect to human reasoning and domain knowledge, not just accuracy or performance.
- **Exposing model’s internal workings**: synoptic and detailed views for navigating complex rule sets.
- **Feature interdependency analysis**: topic modelling and similarity metrics reveal collective effects.
- **Logic-focused refinement**: tools for detecting and testing rule inconsistencies and model cleaning.
- **Domain knowledge integration**: supporting expert-driven improvements that enhance model transparency and reasoning.
- **Main limitation:** high reliance on expert judgment – manual and time-intensive.
- **Direction for future work**: automation of domain constraint enforcement.

CLOSING REFLECTIONS & OPEN QUESTIONS

- Trustworthiness is not just about accuracy—it's about ***understanding why*** the model makes decisions.
- How can we *incorporate human logic* into the interpretation and explanation of data-driven models?
- Should ML models be adjusted to better *reflect human reasoning*, even if accuracy slightly decreases?
- Can domain knowledge be integrated *during model development*, rather than only in post hoc analysis? Can *visual analytics* help to achieve this?
- How can we scale interpretability work via *semi-automated expert-guided interfaces*?