Layerwise Training of Deep Rule-Based Networks Using Stacking

Bachelor-Thesis by Daniel Jung

1. Review: Prof. Dr. Johannes Fürnkranz
2. Review: Dr. Eneldo Loza Mencía
3. Review: Michael Rapp

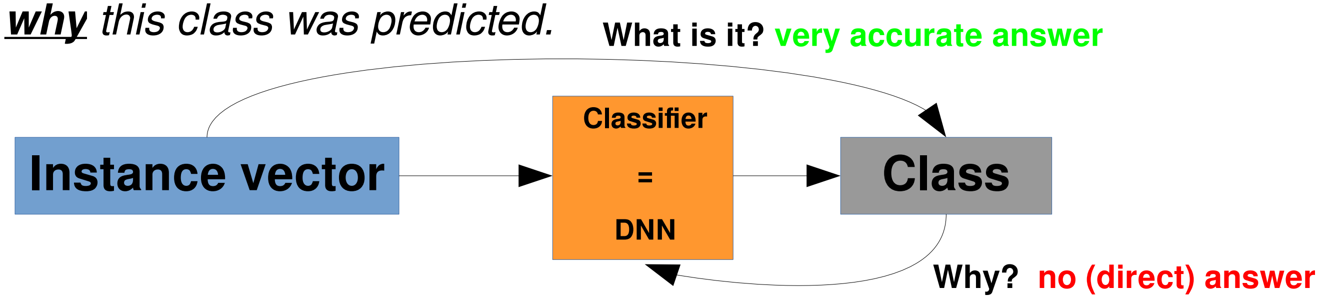
Contents

1 Foundations – Networks and Rules

1.1 Motivation by Deep Neural Networks DNN

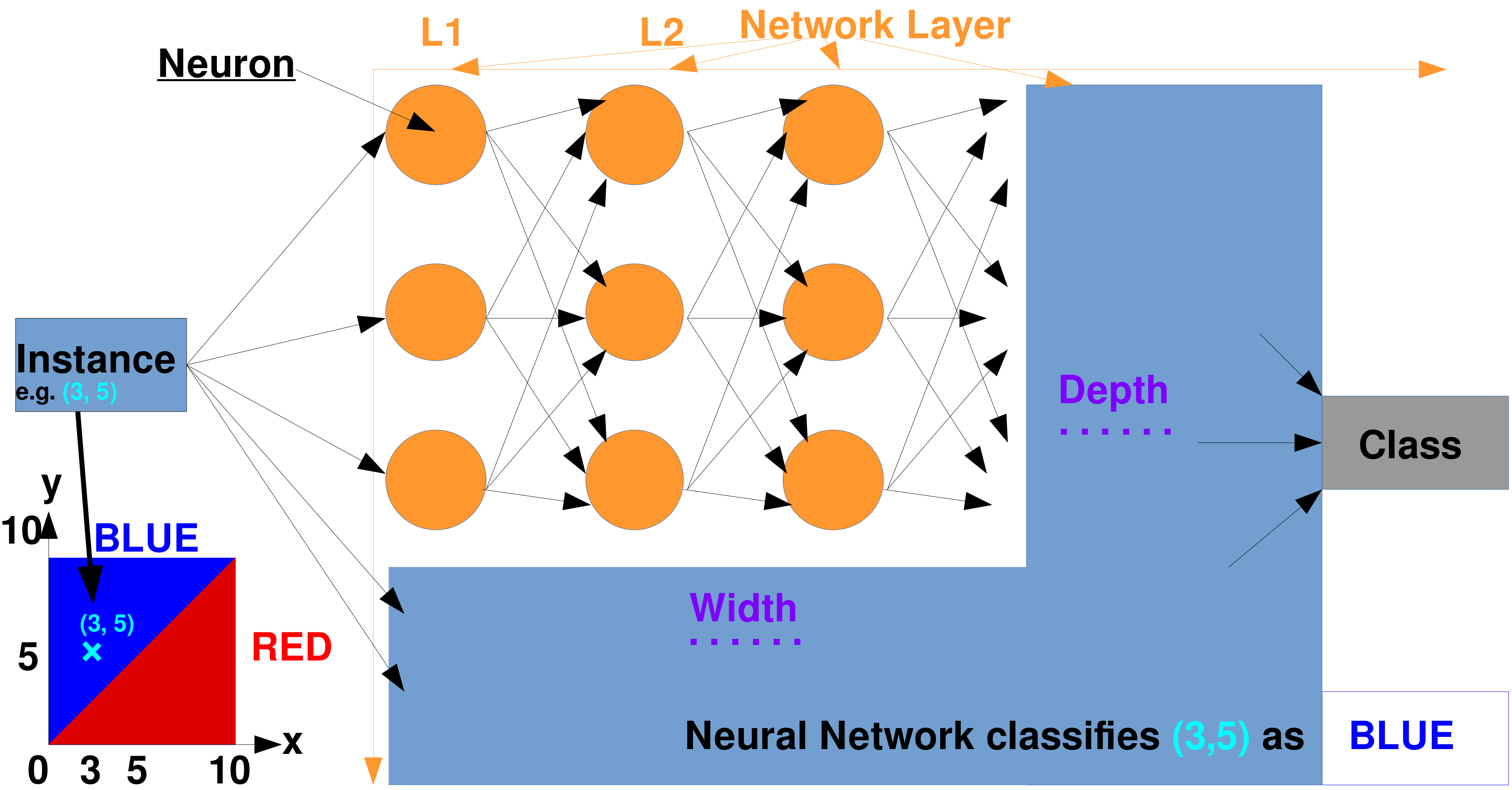
TECHNISCHE UNIVERSITÄT DARMSTADT

- Deep Neural Networks **DNN** *(Artificial Neural Networks ANN's with more than one hidden layer)* used as classifiers
  - have a **competitive prediction accuracy** *especially for high-dimensional data, e. g. classification of objects in images,*
  - but the network structure is generally **not descriptive**, *i. e. an observer of the data transformations from layer to layer cannot conclude **why** this class was predicted.*



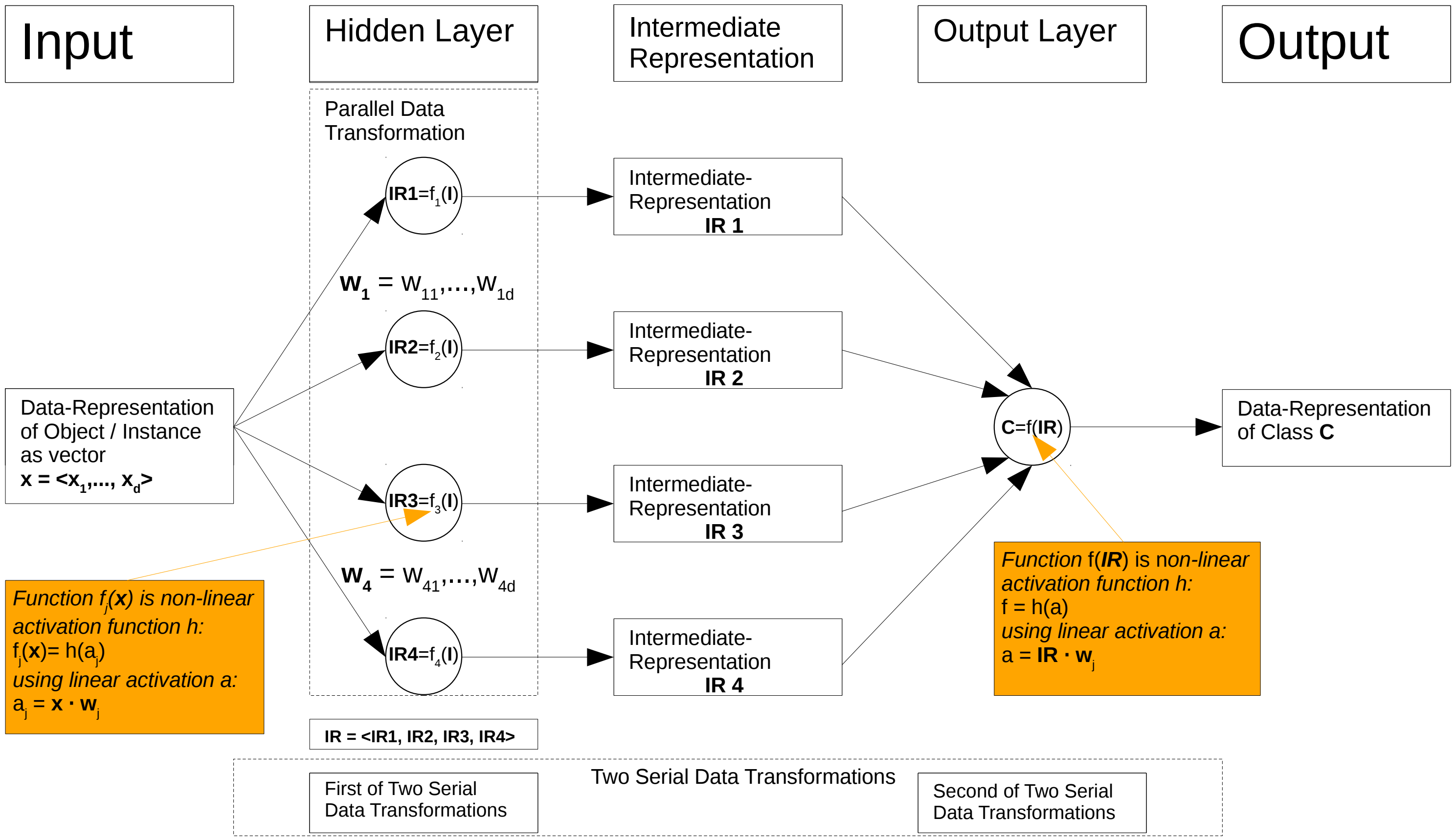
1 Foundations – Networks and Rules

1.2 Deep Neural Network Classifiers

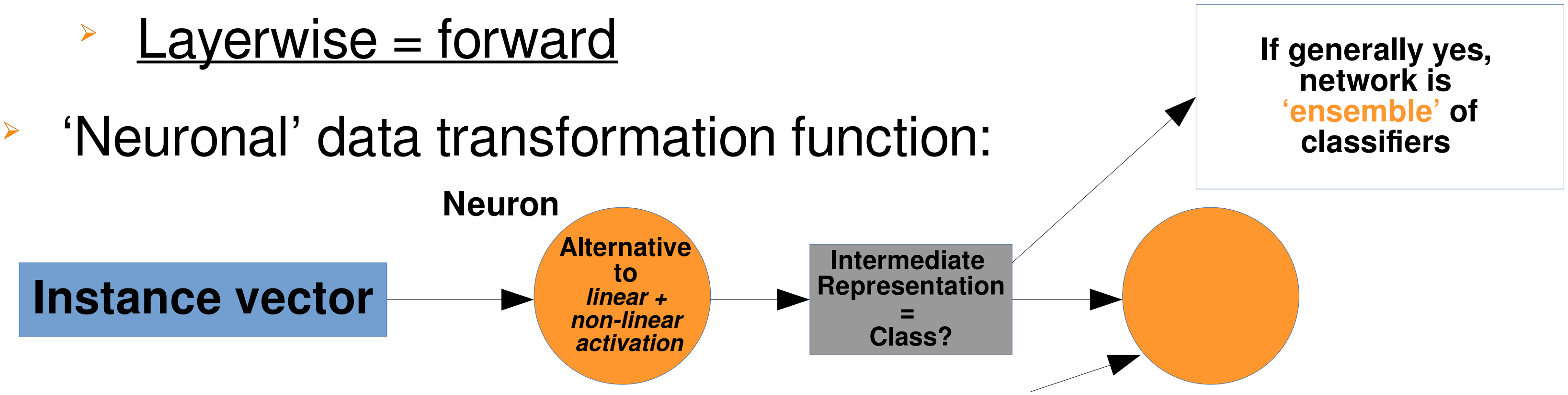


1 Foundations – Networks and Rules

1.3 Sub-Modules in ANN / DNN



- Training:
  - No Backpropagation
  - Layerwise = forward
- 'Neuronal' data transformation function:



  - Decision Tree ('*ForwardThinking* Deep Random Forest')
  - Random Forest ('*gc* Forest')
  - Rule (explored here)

1 Foundations – Networks and Rules

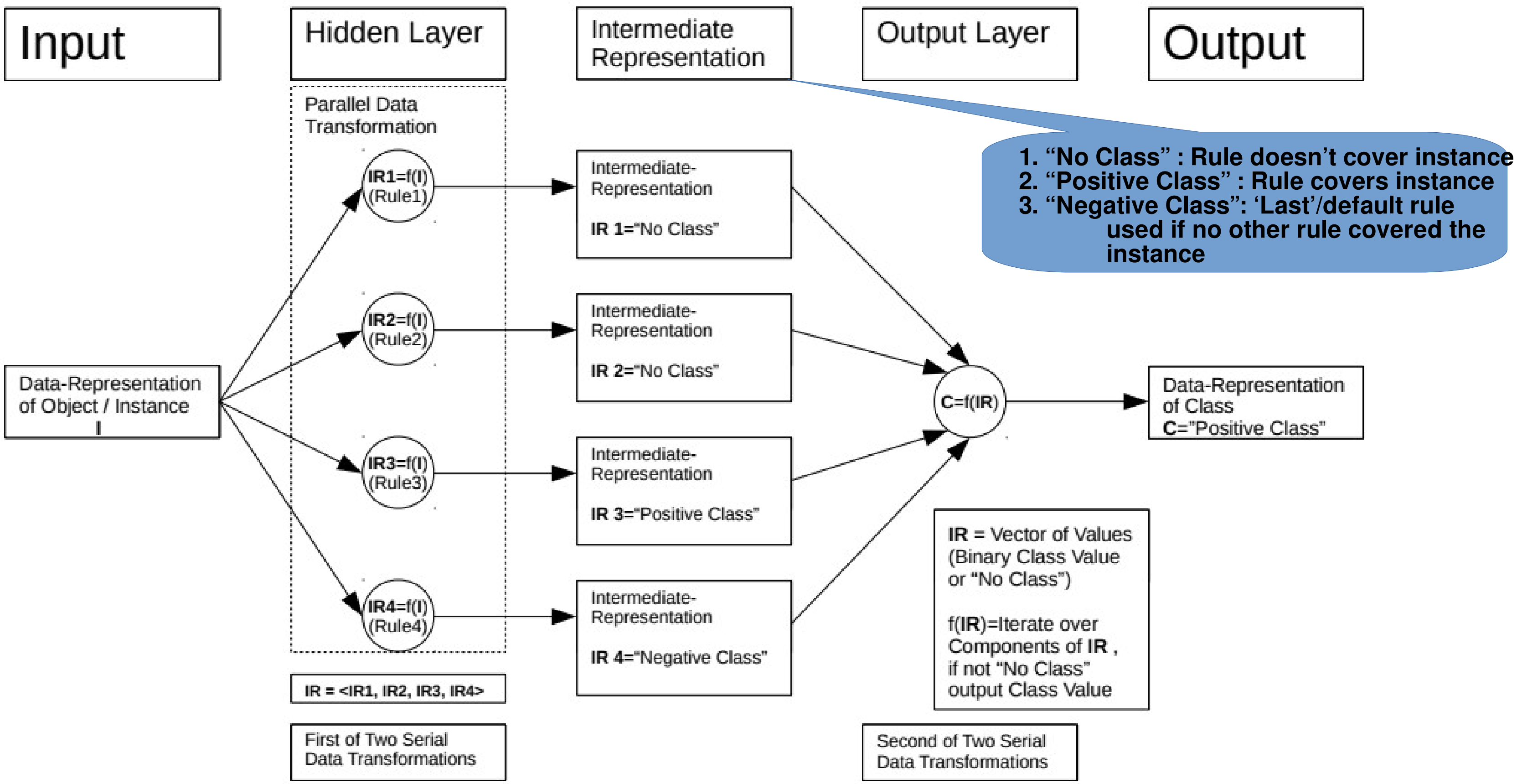
1.5 Rule-Based Classification - Decision List

TECHNISCHE UNIVERSITÄT DARMSTADT

- Rule-based *decision list* classifiers (trained by SeCo algorithms like Ripper)**,**
  - **high** predictive **accuracy**
  - **descriptive**,
    - exposing patterns relevant for the prediction.
    - because the instances of the instance space prior to the classification are grouped by eventually meaningful features different from the actual classes.
- Decision lists are '**ensembles' of rules** and have a network structure
  - SeCo produces '**diverse**' rules (classifiers)
  - Predictions are **combined** by the (predefined) 'rule':
    - First rule in list that covers the instance
    - determines the prediction of the decision list classifier

1 Foundations – Networks and Rules

1.6 Sub-Modules in a Decision List



1 Foundations – Networks and Rules

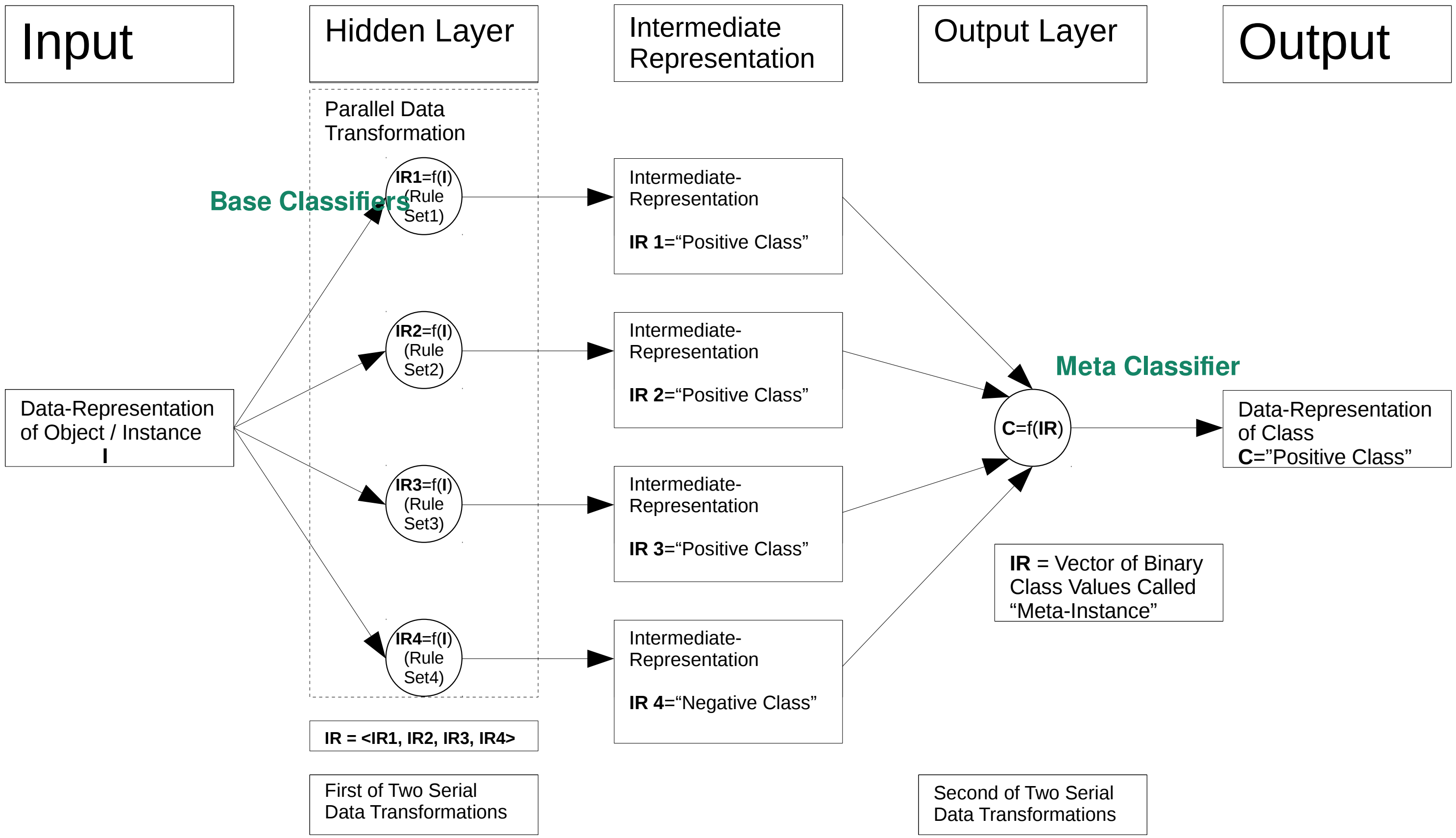
1.7 Ensembles of Rule-Based Classifiers

TECHNISCHE UNIVERSITÄT DARMSTADT

- Ensembles work well typically with different types of classifiers
- Stacking
  - Level 0 model: Base classifier
  - Level 1 model: Meta classifier
  - Rules (Rule set with one element)
    - can be used as base classifiers and
    - as the meta classifier
  - This model can be characterized as a network

1 Foundations – Networks and Rules

1.8 Sub-Modules – Stacking of Rule Sets



Contents

2 Network of Rules

2.1 Motivation

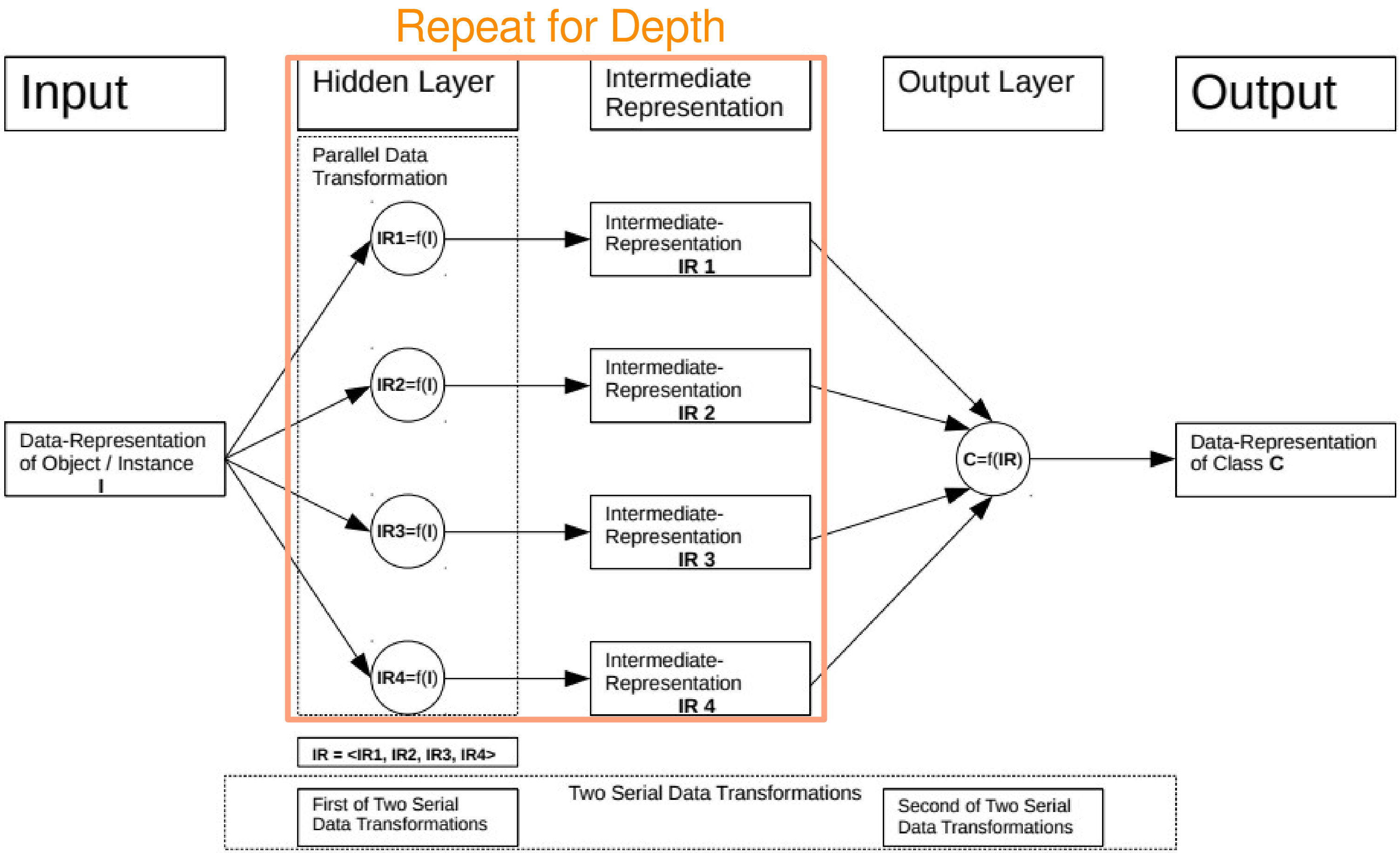
TECHNISCHE UNIVERSITÄT DARMSTADT

- Idea:

  Network of stacked layers of diverse single rule classifiers
- Prediction accuracy and descriptiveness of classifiers could profit from a feed-forward **network structure**.
  - *Width* : *I*ncrease **diversity** of **single rule classifiers** that form each layer
  - *Depth* : Additional layers consist of **single rules as meta classifiers**
    - that profit from the diversity of *preceding layers* and
    - serve the *succeeding layer* as **advanced** diverse set of base classifiers

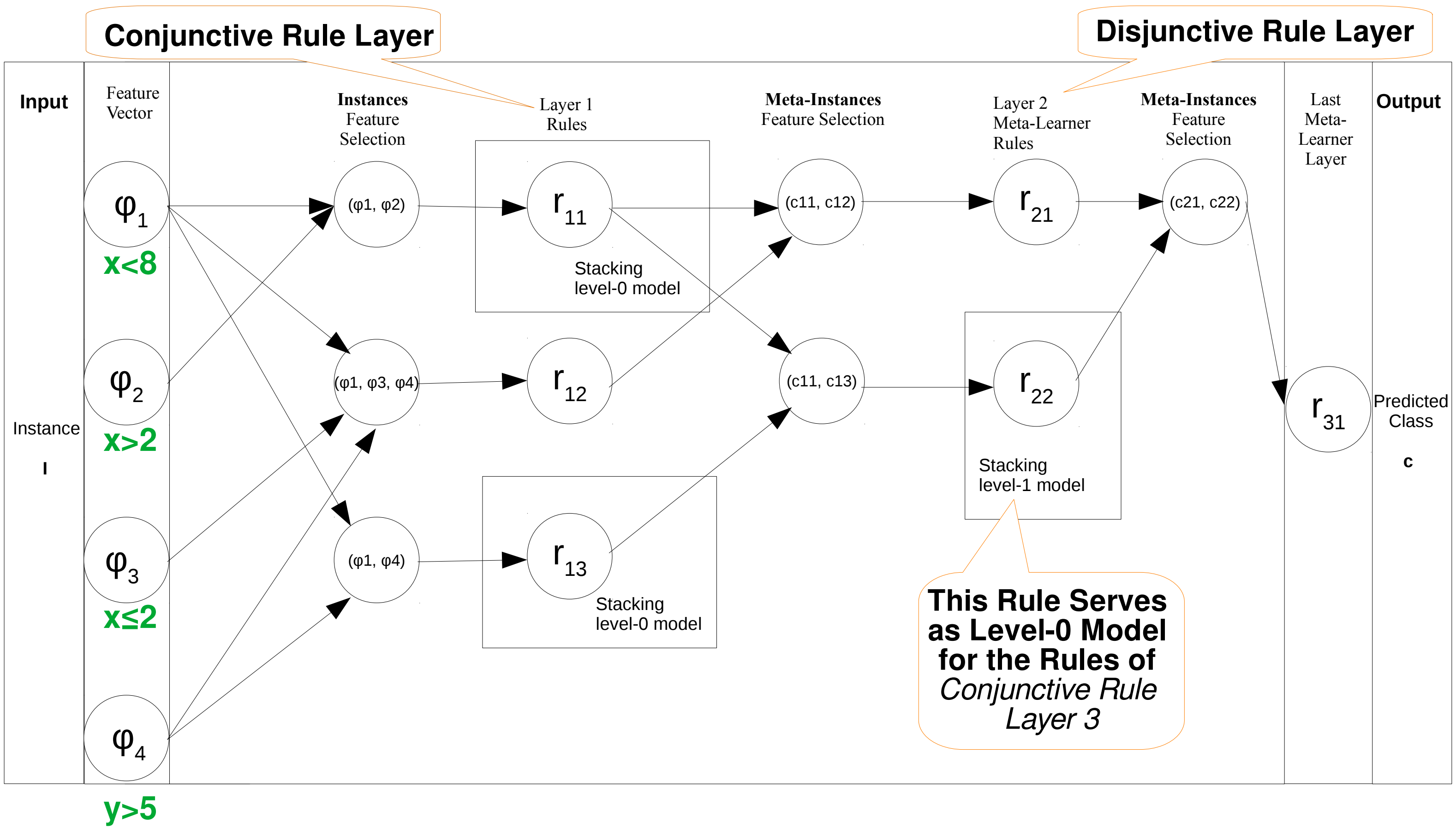
2 Network of Rules

2.2 Classification Using Network Structure



2 Network of Rules

2.3 Network of Stacked Rule Classifiers



2 Network of Rules

2.4 Rule Induction - Diversity

TECHNISCHE UNIVERSITÄT DARMSTADT

- Goal: Diversity of Rules in Each Layer
- Possible approach:
  - Separate and Conquer – SeCo
    - Pro: **diverse rules**
    - Contra: **limited number of different rules**
  - Weighted Covering
    - Pro: **higher number of different rules**
    - Contra: **decreasing difference between rules**
  - Bagging
    - Individual: **each rule induced on different bootstrap sample data sets**
    - Per SeCo/Weighted Covering Cycle: **Use same sample data set per set of rules**
      - Pro: **for SeCo the separation of instances remains consistent**
  - Random Attribute Subset Selection
    - Pro: **diverse rules**
    - Contra: **not ‘compatible’ with SeCo**

2 Network of Rules

2.5 Rule Induction – Disjunctive Rules

TECHNISCHE UNIVERSITÄT DARMSTADT

- Goal: High Expressiveness of Final Hypothesis
  - Sets of ordinary **conjunctive rules** are usually interpreted as **disjunctions**
  - effectively creating a hypothesis in **DNF** for binary classification
- The final meta classifier:
  - single rule of the last layer of the network
  - would be a conjunctive rule since it would use predictions of conjunctive rules recursively
- Solution:
  - Conjunctive and disjunctive rules
  - alternate layerwise
- Disjunctive rule:
  - empty rule covers no instance
  - adding alternatives generalizes the rule

2 Network of Rules

2.6 Rule Induction – Heuristic

- Confusion Matrix CM $CM = \begin{bmatrix} TP & FN \\ FP & TN \end{bmatrix}$ $Recall(CM_{rule}) = TP/(TP+FN)$
  $Precision(CM_{rule}) = TP/(TP+FP)$
- F-Measure $F\!-\!measure(CM_{rule}) = \frac{(\beta^2+1)\cdot Precision(CM_{rule})\cdot Recall(CM_{rule})}{\beta^2\cdot Precision(CM_{rule})+Recall(CM_{rule})}$
  - to find trade off between recall and precision
  - Optimize β for
    - rules in conjunctive layers
    - rules in disjunctive layers

Contents

3 Rule Types

3.1 Types of Rules

- Used Rules:

  All Rules that Form the Prediction Model

  - Copy Rules:

    Used Rules that is equivalent to a rule from it’s preceding layer

  - Feature Rules:

    Used Rules that combine more than one rule from preceding layer

  - Birth Rules:

    Used Rules that has no condition, i. e. is not a combination of any rule in the preceding layer

- Unused Rules:

  Rules that are induced by any layer but it’s not used by the succeeding

3 Rule Types

3.2 Hypothesis Examples of Rule Networks

TECHNISCHE UNIVERSITÄT DARMSTADT

Set of Example Features = $\{\phi_1 = x < 8, \phi_2 = x > 2, \neg\phi_2 = x \leq 2, \phi_3 = y > 5\}$		
Number of Layers	Set of Rules	Hypothesis Representation
2	$\{r_{11} = \phi_1 \wedge \phi_2, r_{12} = \neg\phi_2 \wedge \phi_3, r_{13} = \phi_3, r_{21} = \hat{c}_{11} \vee \hat{c}_{12}\}$	$r_{21} = \phi_1 \wedge \phi_2 \vee \neg\phi_2 \wedge \phi_3$
4	$\{r_{11} = \phi_1 \wedge \phi_2, r_{12} = \neg\phi_2 \wedge \phi_3, r_{13} = \phi_3, r_{21} = \hat{c}_{11} \vee \hat{c}_{12}, r_{22} = \hat{c}_{13}, r_{32} = \hat{c}_{21}, r_{34} = \hat{c}_{22}, r_{41} = \hat{c}_{32} \vee \hat{c}_{34}$	$r_{41} = (\phi_1 \wedge \phi_2 \vee \neg\phi_2 \wedge \phi_3) \vee \phi_3$

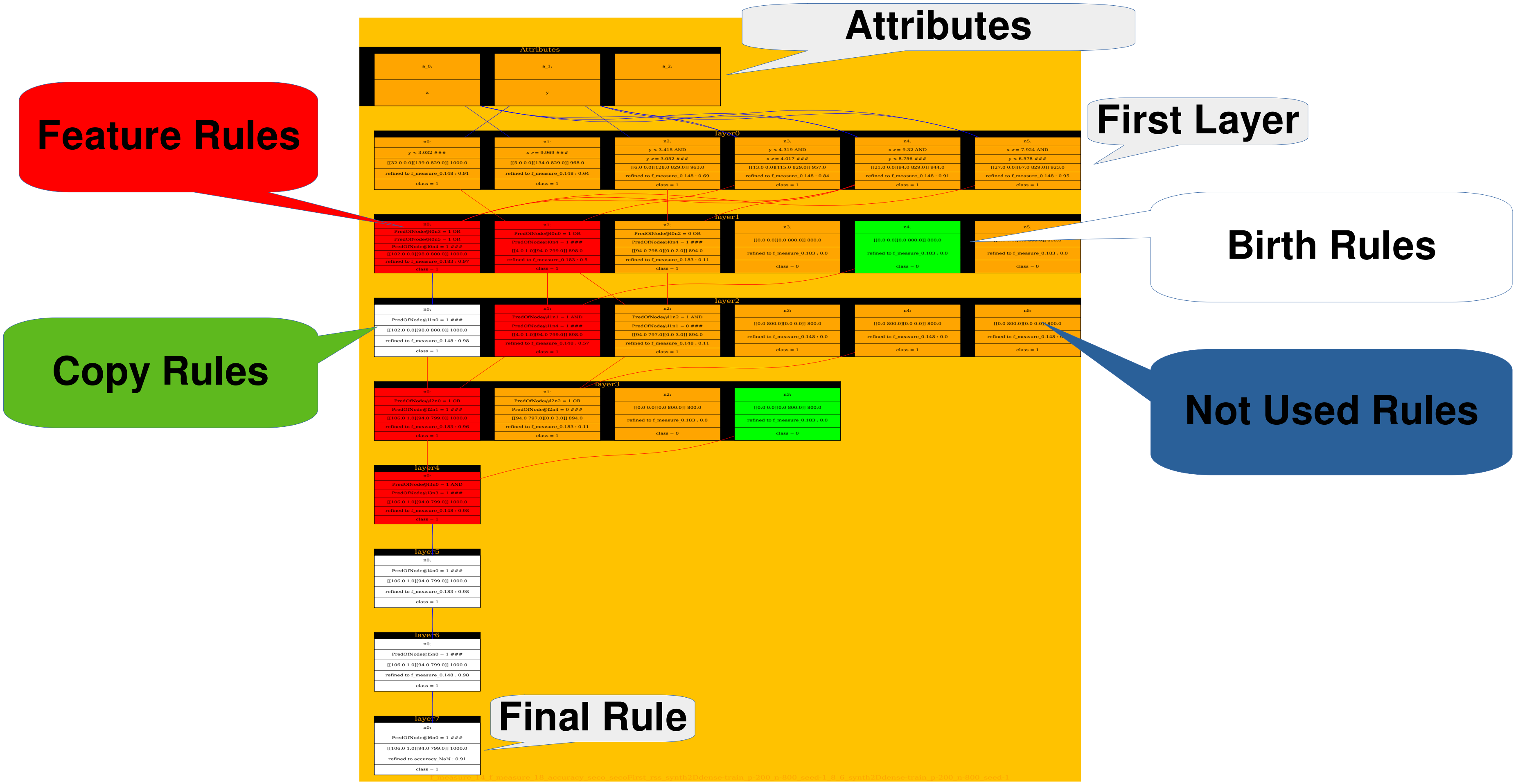
Feature (final) rule r41
in layer 4
predicts class
if
r32 (rule of layer3)
predicts class
or
r34 predicts class

Copy rule r32
in layer 3
predicts class
if
r21 predicts class
(this would be a
conjunction if it
was a feature rule)

Feature rule r21
in layer 2
predicts class
if
r11 predicts class
or
r12 predicts class

3 Rule Types

3.3 Network Example



3 Rule Types

3.4 Feature Rules

TECHNISCHE UNIVERSITÄT DARMSTADT

- High number of feature rules indicates
  - learned hypothesis could have profited from the network structure
  - and improved from layer to layer
- If rules of the first layer (regular conjunctive rules) are disjunctively combined by a rule of the second layer (disjunctive layer)
  - and this rule *performs already well,*
  - it is *challenging to improve* such a rule combining it conjunctively or disjunctively with other rules,
  - so it will likely be propagated to the last layer by copy rules

Contents

4 Evaluation

4.1 Experimental Setup

TECHNISCHE UNIVERSITÄT DARMSTADT

Goal: Find Correlations between parameters and accuracy to optimize parameters

Hyper-parameters for rule induction – (random selection)

- Network size:
  - Number of layers: l
  - Maximum number of rules per layer: n
- Trade off between recall and precision
  - f-measure parameter for conjunctive rules β-con
  - f-measure parameter for disjunctive rules β-dis
- Diversification of rules

  (separate for *first* layer and *all* consecutive layers):
  - SeCo (with cyclic bagging): SeCoFirst / SeCoAll
  - Weighted covering (with cyclic bagging): WeightedFirst / WeightedAll
  - Bagging (sample for each rule): BaggingFirst / BaggingAll
  - Random attribute subset selection RssFirst / RssAll
    - Number of attributes: k
  - random enforcement of feature rules (enforcing a minimum of two conditions per rule): enforceTwoCond

4 Evaluation

4.2 Accuracy - Synthetic Data Set

TECHNISCHE UNIVERSITÄT DARMSTADT

Beta Con	Beta Dis	secoFirstLayer	secoAll	baggingFirstLayer	baggingAll	rssAll	weightedFirstLayer	weightedAll	rssK	tpCovering	enforceTwoCond	accuracy	accuracyClass
0.705	0.875	+	+	+	-	-	-	-	-1	+	0.511	0.959	VeryGood
0.345	1.25	+	-	+	+	-	-	+	-1	+	0.11	0.956	VeryGood
0.344	0.311	+	-	-	+	-	-	+	-1	-	0.178	0.954	VeryGood
0.521	0.386	+	-	+	-	-	-	+	-1	+	0.716	0.954	VeryGood
0.404	0.052	+	+	+	-	+	-	-	8	-	0.077	0.952	VeryGood
0.141	0.375	+	-	-	+	-	-	+	-1	+	0.703	0.947	Good
0.194	0.945	+	+	-	+	-	-	-	-1	-	0.427	0.946	Good
0.471	0.154	+	-	-	+	-	-	+	-1	+	0.184	0.946	Good
0.056	0.087	+	+	+	+	-	-	-	-1	-	0.31	0.945	Good
0.065	0.273	+	+	+	+	-	-	-	-1	-	0.954	0.945	Good
0.242	0.47	+	-	-	-	-	-	+	-1	-	0.836	0.942	Good
0.226	0.072	-	-	-	-	-	+	+	-1	-	0.276	0.941	Good
0.678	0.333	-	-	+	-	-	+	+	-1	+	0.885	0.941	Good
0.074	0.171	+	+	+	-	-	-	-	-1	-	0.637	0.936	Good
0.372	0.162	-	-	-	+	-	+	+	-1	+	0.367	0.935	Good
1.167	0.279	-	-	+	+	-	+	+	-1	+	0.321	0.934	Good
0.848	1.062	-	-	+	-	+	+	+	8	+	0.051	0.929	Good
0.033	0.084	+	+	+	-	-	-	-	-1	+	0.794	0.928	Good
0.032	0.038	+	+	+	+	-	-	-	-1	-	0.68	0.926	Good
0.072	0.132	+	+	-	-	-	-	-	-1	+	0.413	0.925	Good
0.223	0.068	+	+	+	-	-	-	-	-1	-	0.258	0.917	Good
1.12	1.659	-	+	-	-	+	+	-	7	+	0.971	0.916	Good
0.765	0.052	-	-	-	+	+	+	+	8	+	0.806	0.916	Good
1.145	0.048	-	-	-	-	-	+	+	-1	+	0.748	0.916	Good
0.403	0.04	+	-	+	-	-	-	+	-1	+	0.081	0.915	Good
0.999	0.767	-	+	+	+	+	+	-	9	+	0.769	0.913	Good
0.804	0.16	+	+	+	-	-	-	-	-1	+	0.385	0.91	Good
0.111	0.165	-	-	-	-	-	+	+	-1	-	0.422	0.908	Good
0.148	0.183	+	+	+	-	+	-	-	6	+	0.707	0.902	Good
0.065	0.052	+	-	+	-	-	-	+	-1	+	0.161	0.902	Good
0.445	0.053	-	-	+	-	+	+	+	9	-	0.797	0.902	Good
0.04	0.152	+	-	-	-	+	-	+	8	+	0.443	0.895	Bad
0.394	0.087	+	+	+	+	+	-	-	8	-	0.93	0.889	Bad
1.188	0.252	-	+	+	+	+	+	-	1	-	0.907	0.886	Bad
2.273	0.357	-	-	+	-	-	+	+	-1	-	0.284	0.884	Bad
0.931	0.204	-	-	+	-	+	+	+	4	+	0.549	0.884	Bad
0.269	0.07	+	-	+	+	+	-	+	8	+	0.757	0.88	Bad
0.342	0.499	-	+	-	-	-	+	-	-1	-	0.396	0.88	Bad
0.221	0.077	-	-	+	+	-	+	+	-1	-	0.156	0.878	Bad
0.273	0.215	-	+	+	-	-	+	-	-1	+	0.034	0.878	Bad
1.718	0.583	-	-	+	-	-	+	+	-1	+	0.769	0.875	Bad
0.133	0.523	+	-	-	+	+	-	+	5	+	0.436	0.874	Bad
1.654	0.355	+	-	-	-	+	-	+	3	-	0.674	0.874	Bad
2.177	0.039	-	-	-	-	-	+	+	-1	-	0.175	0.873	Bad
0.09	0.807	+	-	-	+	+	-	+	5	+	0.502	0.872	Bad

4 Evaluation

4.3 Rule Characteristics

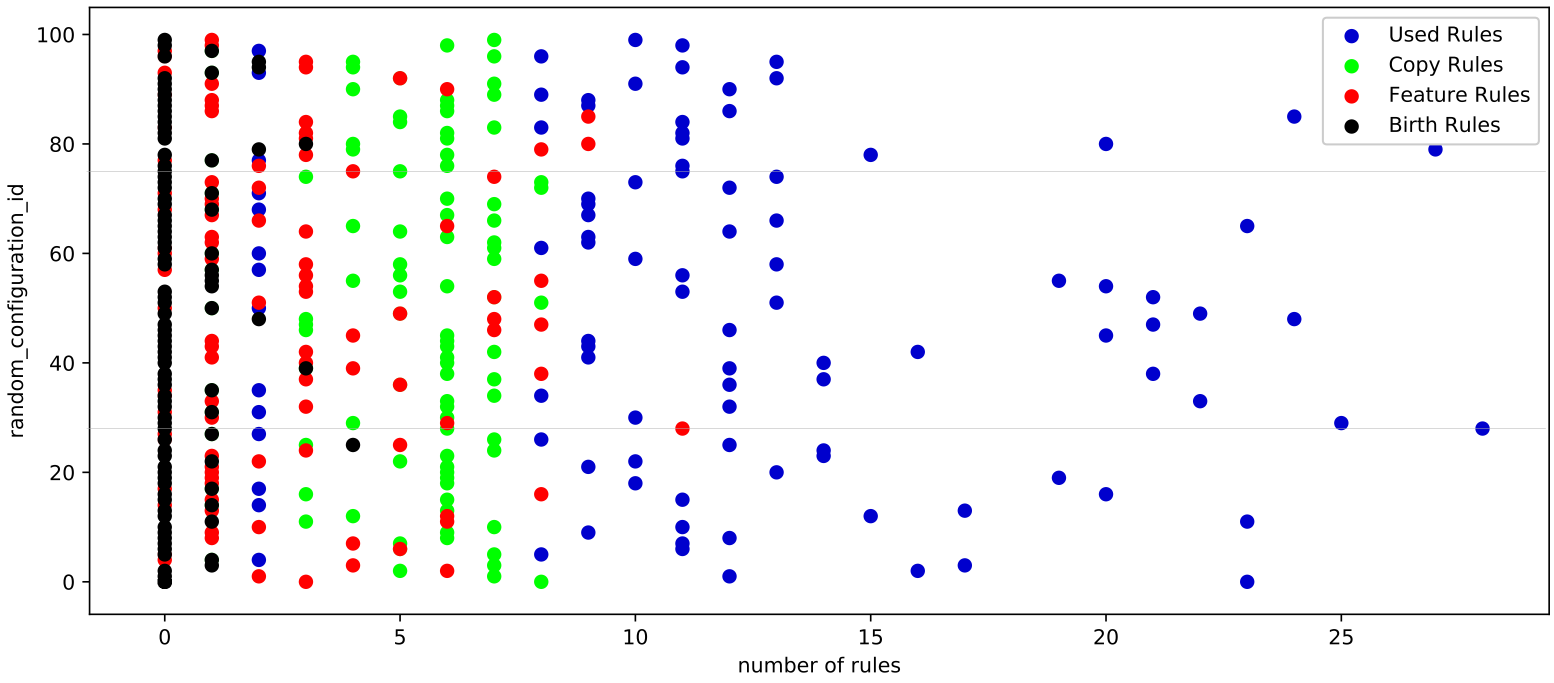
TECHNISCHE UNIVERSITÄT DARMSTADT

The rule characteristics show

- how the number of feature rules is related to the number of mere copy rules
- how many of the rules used for the hypothesis (used rules) are
  - first layer rules = neither copy nor feature rule
  - copy rules
  - feature rules
- number of birth rules indicating weak preceding rules

4 Evaluation

4.4 Rule Characteristics - Vote



4 Evaluation

4.5 Rule Characteristics

- Preferable configurations that encourage
  - feature rules over
  - copy rules
- Could be observed in preliminary experiments where the first layer
  - uses random subset selection
  - but not SeCo or weighted covering to induce rules
  - led to lower accuracy in comparison with SeCo
- Random Hyper-parameter experiments could not reveal other correlations

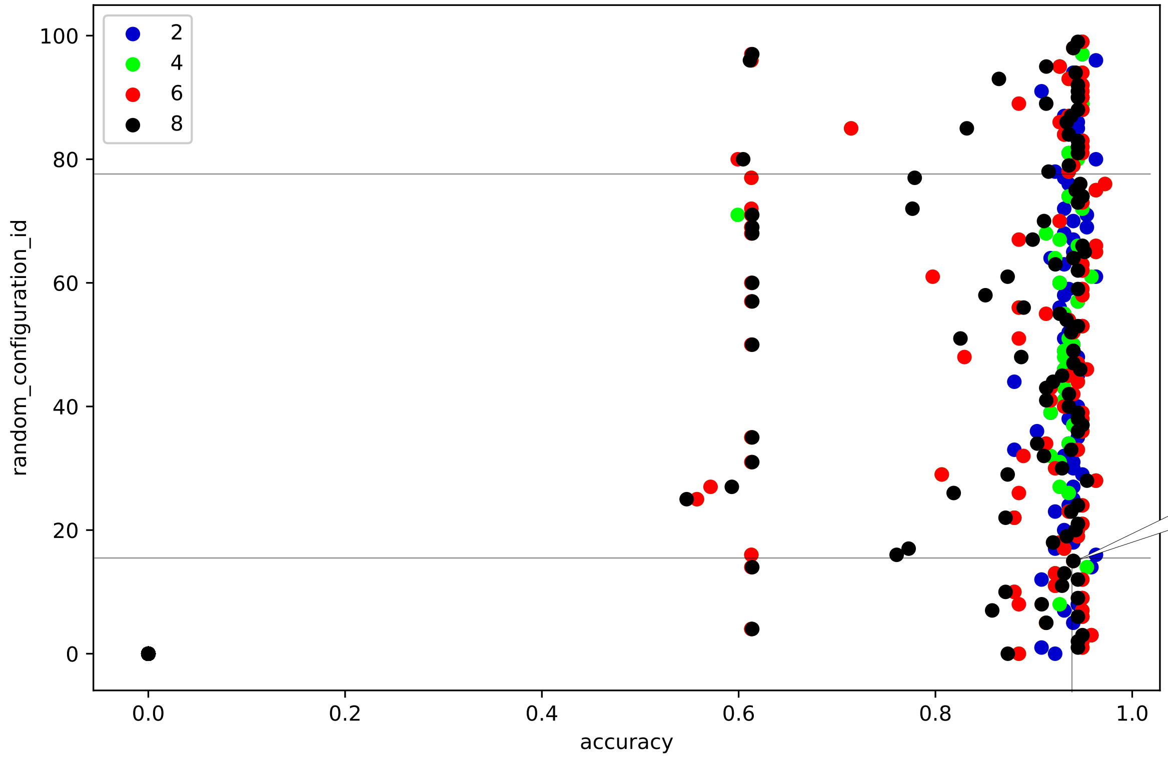
4 Evaluation

4.6 Layerwise Accuracy

- Expected increase in predictive accuracy with additional layers
- Decrease possible, e. g. in case of enforced feature rules if one condition is optimal

4 Evaluation

4.7 Layerwise Accuracy - Vote



Due to copy rules No change with increasing layers

4 Evaluation

4.8 Layerwise Accuracy

- For the applied diversity strategy and the dimension of the network (up to 100 rules per layer)
  - there could be no pattern observed that would increase the accuracy by adding layers
- If the performance of the rules in the first layers is kept low (e.g. by random attribute subset selection without SeCo in the First Layer)
  - an increase can be observed
  - but not beyond the accuracy that occurs if higher performance in first layers is encouraged

Conclusion

TECHNISCHE UNIVERSITÄT DARMSTADT

- If SeCo in first layer

  highest accuracy typically
  - already within 2 layers
  - copy rules propagate result to

    available deeper layers
- No considered configuration exceeds this accuracy
  - additional layers can increase accuracy
  - but only from a lower accuracy in the first 2 layers (Random subset selection without SecO/Weighted Covering)
- Future Work:
  - Random subset selection per SeCo cycle
    - for high number of cycles – high number of rules per layer
    - In case of layerwise increase – experiments with more layers
  - Increase of number of random hyper-parameter configurations
    - to find correlations that allow parameter optimization

TECHNISCHE
UNIVERSITÄT
DARMSTADT

Thank You!

Questions?