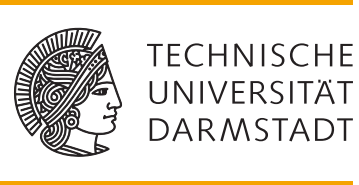
A Separate-and-Conquer Algorithm for Learning Multi-Label Head Rules



Master-Thesis by Michael Rapp

1. Referee: Prof. Dr. Johannes Fürnkranz
2. Referee: Dr. Eneldo Loza Mencía
3. Referee: Dr. Frederik Janssen

Contents

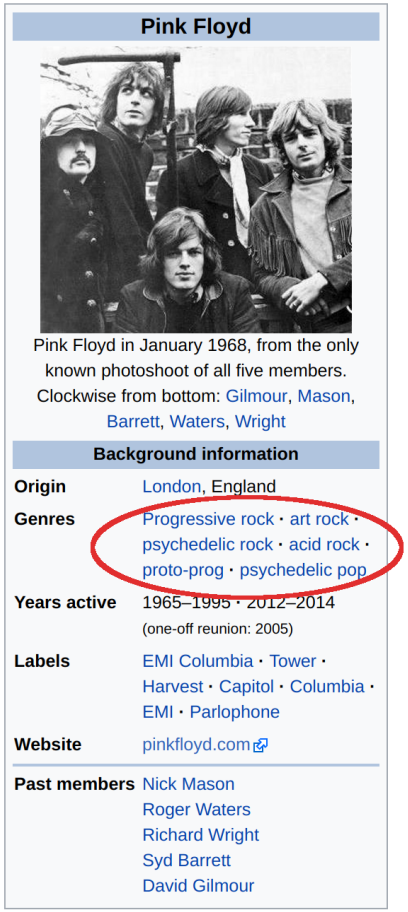
TECHNISCHE UNIVERSITÄT DARMSTADT

1 Foundations

1.1 Motivation

TECHNISCHE UNIVERSITÄT DARMSTADT

- **Multi-label classification** is the task of assigning a subset of (relevant) **labels** to an object. This is in contrast to
  - **Binary classification**, where each object is associated with one out of two classes
  - **Multi-class classification**, where each object is associated with one out of several classes
- **Typical application areas** are
  - Tagging books, news, music,... with topics, genres, authors,...
  - Detecting presence or absence of objects in images
  - Classifying function of genomes or proteins in biology
  - ...



1 Foundations

1.2 Label-dependencies

TECHNISCHE UNIVERSITÄT DARMSTADT

- Studies have shown, that the predictive performance of classification approaches may benefit from exploiting correlations between labels[1].
- Possible correlations are for example
  - **Exclusions:** The presence of certain labels might imply, that other labels are irrelevant.
    e.g. the music genres "progressive rock" and "easy listening" are contrary
  - **Subsumptions:** Labels can be ordered hierarchical.
    e.g. "progressive rock" is a sub-genre of "rock"

[1] Krzysztof Dembczyński, Willem Waegeman, Weiwei Cheng, and Eyke Hüllermeier. On label dependence and loss minimization in multi-label classification. *Machine learning*, 88(1-2):5–45, 2012.

1 Foundations

1.3 Example data set

TECHNISCHE UNIVERSITÄT DARMSTADT

Attributes				Labels				
outlook	temp.	humidity	windy	play	dontplay	eaticecream	drinktee	
sunny	85	85	FALSE	0	1	1	0	Training examples
sunny	80	90	TRUE	0	1	1	0	
overcast	83	86	FALSE	1	0	0	0	
rainy	70	96	FALSE	1	0	0	0	
rainy	68	80	FALSE	1	0	0	0	
rainy	65	70	TRUE	0	1	0	1	
overcast	64	65	TRUE	1	0	0	0	
sunny	72	95	FALSE	0	1	1	0	
sunny	69	70	FALSE	1	0	0	0	
rainy	75	80	FALSE	1	0	0	0	
sunny	75	70	TRUE	?	?	?	?	Test examples
overcast	72	90	TRUE	?	?	?	?	
overcast	81	75	FALSE	?	?	?	?	
rainy	71	91	TRUE	?	?	?	?	

1 Foundations

1.4 Types of Rules

TECHNISCHE UNIVERSITÄT DARMSTADT

- Rules contain conditions in their bodies and label assignments in their heads. They are notated as

$$head \leftarrow body$$

- **Single-label head rules** predict exactly one label, e.g.

$$dontplay = 1 \leftarrow outlook = rainy \wedge windy = TRUE$$

- **Label-dependent** rules are able to model correlations between labels, e.g.

$$dontplay = 1 \leftarrow play = 0$$

- **Multi-label head rules** may predict multiple labels at once, e.g.

$$play = 1, dontplay = 0 \leftarrow temperature \geq 66.5$$

Contents

TECHNISCHE UNIVERSITÄT DARMSTADT

2 Learning Single-Label Head Rules

2.1 Basic Idea

TECHNISCHE UNIVERSITÄT DARMSTADT

- Loza Mencía and Janssen proposed a **separate-and-conquer algorithm** for learning single-label head rules[2].
- New rules are learned iteratively. They are added to a **decision list** in the order of their induction.
- After each iteration, training examples, whose labels are predicted to some extent, are removed from the training data set.

[2] Eneldo Loza Mencía and Frederik Janssen. Learning rules for multi-label classification: A stacking and a separate-and-conquer approach. *Knowledge Engineering Group, Technische Universität Darmstadt, Germany, 2015.*

2 Learning Single-Label Head Rules

2.2 Structure of the Separate-and-Conquer Algorithm

TECHNISCHE UNIVERSITÄT DARMSTADT

Require: Training data set T, evaluation function δ

1 $R = \emptyset$ ▷ Initialize empty decision list
2 $T_{current} = T$ with all label vectors set to (?, ..., ?)
3 **while** examples are left in T **do** ▷ e.g. until only 5% are left
4 r = FINDBESTGLOBALRULE(T, $T_{current}$, δ) ▷ See next slide
5 $R = R \cup r$ ▷ Add rule to decision list
6 apply head of r to $T_{current}$
7 remove covered examples from T ▷ If enough of their labels are predicted
8 **return** decision list R

- The copied data set $T_{current}$ is used to to keep track of the labels, which are predicted by already learned rules.

- These labels can be used to induce label-dependent rules in later iterations.

2 Learning Single-Label Head Rules

2.3 Algorithm FINDBESTGLOBALRULE

TECHNISCHE UNIVERSITÄT DARMSTADT

Require: Training data sets T and $T_{current}$, evaluation function δ

1 $r_{best} = \emptyset \leftarrow \emptyset$
2 **for each** label y_i **do** ▷ Find best rule for each label
3 $T_i = T \setminus$ examples for which label y_i is set in $T_{current}$
4 $r =$ FINDBESTRULE(T_i, y_i, δ)
5 **if** r outperforms r_{best} according to δ **then**
6 $r_{best} = r$
7 **return** best rule r_{best}

- The subroutine FINDBESTRULE can be implemented by using any rule learner for solving binary classification problems.

2 Learning Single-Label Head Rules

2.4 Using decision lists for classification

TECHNISCHE UNIVERSITÄT DARMSTADT

Decision list:

1	$dontplay = 1 \leftarrow humidity >= 82.5 \wedge outlook = sunny$
2	$eaticecream = 1 \leftarrow dontplay = 1$
3	$play = 1 \leftarrow windy = FALSE$
4	$dontplay = 1 \leftarrow outlook = rainy$
5	$drinktee = 1 \leftarrow dontplay = 1$
6	$play = 1 \leftarrow \emptyset$

Test example to classify:

Attributes				Labels			
outlook	temperature	humidity	windy	play	dontplay	eaticecream	drinktea
rainy	71	91	TRUE	?	?	?	?

All labels are initially unset

2 Learning Single-Label Head Rules

2.4 Using decision lists for classification

TECHNISCHE UNIVERSITÄT DARMSTADT

Decision list:

1	$dontplay = 1 \leftarrow humidity >= 82.5 \wedge outlook = sunny$
2	$eaticecream = 1 \leftarrow dontplay = 1$
3	$play = 1 \leftarrow windy = FALSE$
4	$dontplay = 1 \leftarrow outlook = rainy$
5	$drinktee = 1 \leftarrow dontplay = 1$
6	$play = 1 \leftarrow \emptyset$

does not cover

Test example to classify:

Attributes				Labels			
outlook	temperature	humidity	windy	play	dontplay	eaticecream	drinktea
rainy	71	91	TRUE	?	?	?	?

2 Learning Single-Label Head Rules

2.4 Using decision lists for classification

TECHNISCHE UNIVERSITÄT DARMSTADT

Decision list:

1	$dontplay = 1 \leftarrow humidity >= 82.5 \wedge outlook = sunny$	does not cover
2	$eaticecream = 1 \leftarrow dontplay = 1$	does not cover
3	$play = 1 \leftarrow windy = FALSE$	
4	$dontplay = 1 \leftarrow outlook = rainy$	
5	$drinktee = 1 \leftarrow dontplay = 1$	
6	$play = 1 \leftarrow \emptyset$	

Test example to classify:

Attributes				Labels			
outlook	temperature	humidity	windy	play	dontplay	eaticecream	drinktea
rainy	71	91	TRUE	?	?	?	?

2 Learning Single-Label Head Rules

2.4 Using decision lists for classification

TECHNISCHE UNIVERSITÄT DARMSTADT

Decision list:

1	$dontplay = 1 \leftarrow humidity >= 82.5 \wedge outlook = sunny$	does not cover
2	$eaticecream = 1 \leftarrow dontplay = 1$	does not cover
3	$play = 1 \leftarrow windy = FALSE$	does not cover
4	$dontplay = 1 \leftarrow outlook = rainy$	
5	$drinktee = 1 \leftarrow dontplay = 1$	
6	$play = 1 \leftarrow \emptyset$	

Test example to classify:

Attributes				Labels			
outlook	temperature	humidity	windy	play	dontplay	eaticecream	drinktea
rainy	71	91	TRUE	?	?	?	?

2 Learning Single-Label Head Rules

2.4 Using decision lists for classification

TECHNISCHE UNIVERSITÄT DARMSTADT

Decision list:

1	$dontplay = 1 \leftarrow humidity >= 82.5 \wedge outlook = sunny$	does not cover
2	$eaticecream = 1 \leftarrow dontplay = 1$	does not cover
3	$play = 1 \leftarrow windy = FALSE$	does not cover
4	$dontplay = 1 \leftarrow outlook = rainy$	covers
5	$drinktee = 1 \leftarrow dontplay = 1$	
6	$play = 1 \leftarrow \emptyset$	

Test example to classify:

Attributes				Labels			
outlook	temperature	humidity	windy	play	dontplay	eaticecream	drinktea
rainy	71	91	TRUE	?	**1**	?	?

2 Learning Single-Label Head Rules

2.4 Using decision lists for classification

TECHNISCHE UNIVERSITÄT DARMSTADT

Decision list:

1	$dontplay = 1 \leftarrow humidity >= 82.5 \wedge outlook = sunny$	does not cover
2	$eaticecream = 1 \leftarrow dontplay = 1$	does not cover
3	$play = 1 \leftarrow windy = FALSE$	does not cover
4	$dontplay = 1 \leftarrow outlook = rainy$	covers
5	$drinktee = 1 \leftarrow dontplay = 1$	covers
6	$play = 1 \leftarrow \emptyset$	

Test example to classify:

Attributes				Labels			
outlook	temperature	humidity	windy	play	dontplay	eaticecream	drinktea
rainy	71	91	TRUE	?	**1**	?	**1**

2 Learning Single-Label Head Rules

2.4 Using decision lists for classification

TECHNISCHE UNIVERSITÄT DARMSTADT

Decision list:

1	$dontplay = 1 \leftarrow humidity >= 82.5 \wedge outlook = sunny$	does not cover
2	$eaticecream = 1 \leftarrow dontplay = 1$	does not cover
3	$play = 1 \leftarrow windy = FALSE$	does not cover
4	$dontplay = 1 \leftarrow outlook = rainy$	covers
5	$drinktee = 1 \leftarrow dontplay = 1$	covers
6	$play = 1 \leftarrow \emptyset$	covers

Test example to classify:

Attributes				Labels			
outlook	temperature	humidity	windy	play	dontplay	eaticecream	drinktea
rainy	71	91	TRUE	**1**	**1**	?	**1**

2 Learning Single-Label Head Rules

2.4 Using decision lists for classification

TECHNISCHE UNIVERSITÄT DARMSTADT

Decision list:

1	$dontplay = 1 \leftarrow humidity >= 82.5 \wedge outlook = sunny$	does not cover
2	$eaticecream = 1 \leftarrow dontplay = 1$	does not cover
3	$play = 1 \leftarrow windy = FALSE$	does not cover
4	$dontplay = 1 \leftarrow outlook = rainy$	covers
5	$drinktee = 1 \leftarrow dontplay = 1$	covers
6	$play = 1 \leftarrow \emptyset$	covers

Test example to classify:

Attributes				Labels			
outlook	temperature	humidity	windy	play	dontplay	eaticecream	drinktea
rainy	71	91	TRUE	**1**	**1**	**0**	**1**

Assume remaining labels to be irrelevant!

Contents

TECHNISCHE UNIVERSITÄT DARMSTADT

3 Learning Multi-Label Head Rules

3.1 Basic Idea

- The structure of the separate-and-conquer algorithm – including the removal of training examples – remains unchanged.
- For classification, the rules of a decision list are successively applied as before. However, several labels might be set by multi-label head rules.
- To induce multi-label head rules, the algorithm FINDBESTGLOBALRULE must be modified, because rules cannot be induced for each label individually.

3 Learning Multi-Label Head Rules

3.2 Algorithm FINDBESTGLOBALRULE

TECHNISCHE UNIVERSITÄT DARMSTADT

Require: Training data sets T and $T_{current}$, evaluation function δ

1 $r_{best} = \emptyset \leftarrow \emptyset$
2 improved = **true**
3 **while** improved **do** ▷ Until no refinements possible
4 r = REFINERULE(T, $T_{current}$, r_{best}, δ) ▷ See next slide
5 **if** r outperforms r_{best} according to δ **then**
6 $r_{best} = r$
7 **else**
8 improved = **false**
9 **return** best rule r_{best}

- Starting with the most-general rule, rules are specialized by adding additional conditions to their bodies.

3 Learning Multi-Label Head Rules

3.3 Algorithm REFINERULE

TECHNISCHE UNIVERSITÄT DARMSTADT

Require: Training data sets T and $T_{current}$, rule r, evaluation function δ

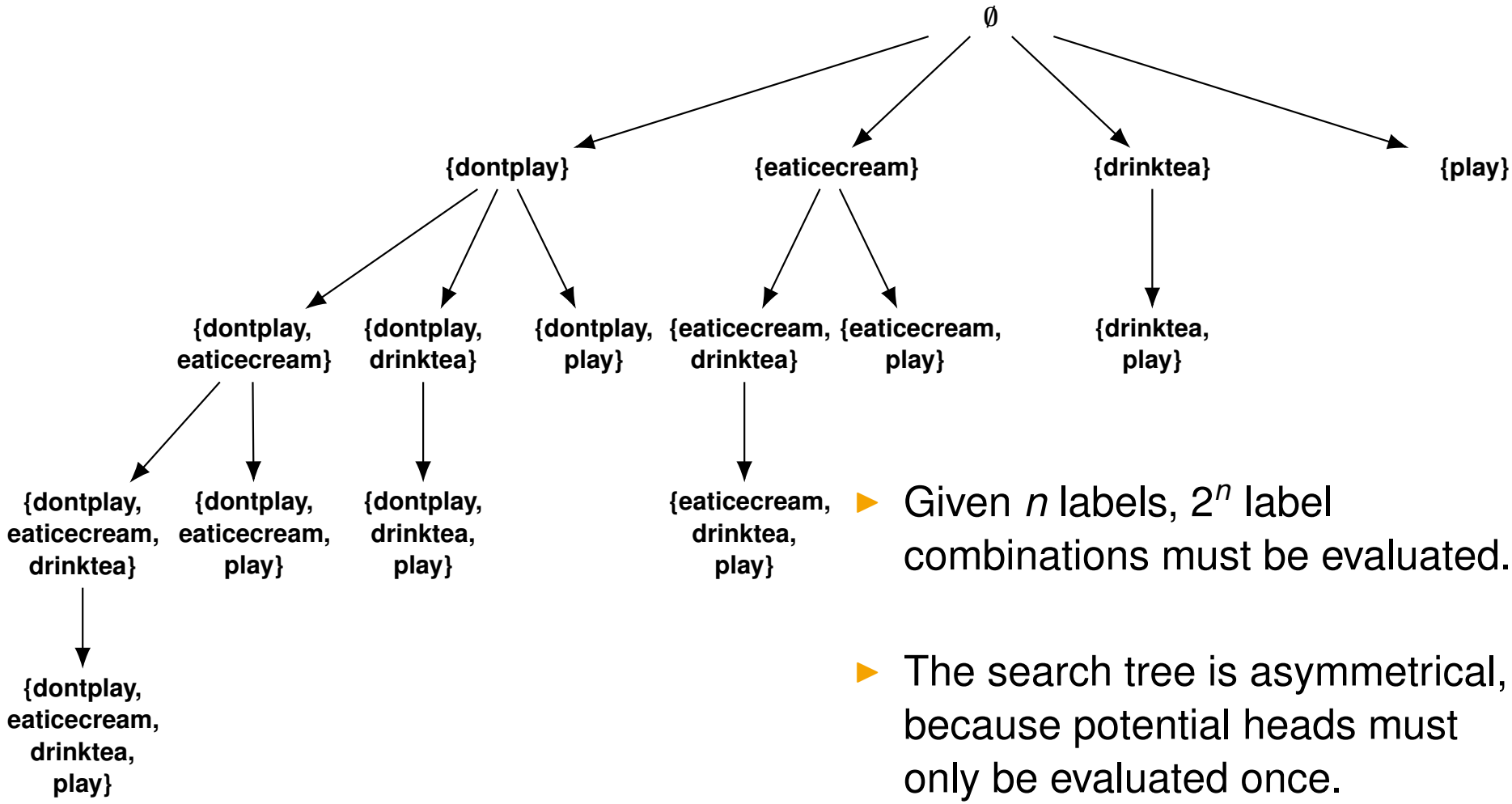
1 $r_{best} = r$
2 **for each** possible condition c not already in r **do**
3 $\quad r_{refined} = r$
4 $\quad$ add c to body of $r_{refined}$
5 $\quad r_{refined} = \text{FINDBESTHEAD}(T, T_{current}, r_{refined}, \delta)$ ▷ See next slide
6 $\quad$ **if** $r_{refined}$ outperforms r_{best} according to δ **then**
7 $\quad\quad r_{best} = r_{refined}$
8 **return** best refined rule r_{best}

- For each potential refinement, the best multi-label head must be found.
- The algorithm FINDBESTHEAD performs (pruned) searches through the label space.

3 Learning Multi-Label Head Rules

3.4 Exhaustive Search through the Label Space

TECHNISCHE UNIVERSITÄT DARMSTADT



- Given n labels, 2^n label combinations must be evaluated.
- The search tree is asymmetrical, because potential heads must only be evaluated once.

3 Learning Multi-Label Head Rules

3.5 Evaluation Functions (1/2)

TECHNISCHE UNIVERSITÄT DARMSTADT

- **Precision:** Percentage of correctly predicted labels among all predicted labels

$$\delta_{prec} = \frac{TP}{TP + FP}$$

- **Recall:** Percentage of correctly predicted labels among all relevant labels

$$\delta_{rec} = \frac{TP}{TP + FN}$$

- **Hamming Accuracy:** Percentage of correctly predicted labels among all labels

$$\delta_{hamm} = \frac{TP + TN}{TP + TN + FP + FN}$$

3 Learning Multi-Label Head Rules

3.5 Evaluation Functions (2/2)

TECHNISCHE UNIVERSITÄT DARMSTADT

- **F-Measure:** Weighted harmonic mean between precision and recall

$$\delta_f = \frac{\left(\beta^2 + 1\right) \cdot \delta_{prec} \cdot \delta_{rec}}{\beta^2 \cdot \delta_{prec} + \delta_{rec}} \text{ , with } \beta \in [0, \infty]$$

- **Subset Accuracy:** Percentage of perfectly predicted label vectors

$$\delta_{acc} = \frac{1}{m} \sum_{j=1}^{m} \left[Y_j = \hat{Y}_j \right] \text{ , with } [x] = \begin{cases} 1, & \text{if } x \text{ is true} \\ 0, & \text{otherwise} \end{cases}$$

3 Learning Multi-Label Head Rules

3.6 Confusion matrices

TECHNISCHE UNIVERSITÄT DARMSTADT

- **True positives:** Correctly predicted labels of covered examples.
- **False positives:** Incorrectly predicted labels of covered examples.
- **True negatives:** Irrelevant labels of uncovered examples.
- **False negatives:** Relevant labels of uncovered examples.

Candidate rule:

$$play = 0, dontplay = 1 \leftarrow outlook = sunny$$

Training examples:

Attributes				Labels				
outlook	temp.	humidity	windy	play	dontplay	eaticecream	drinktee	
sunny	72	80	TRUE	0	1	1	0	covered
sunny	68	72	FALSE	0	1	1	0	covered
overcast	65	83	FALSE	1	0	0	0	not covered
rainy	64	75	FALSE	1	0	0	0	not covered

3 Learning Multi-Label Head Rules

3.7 Different Averaging Strategies (1/4)

TECHNISCHE UNIVERSITÄT DARMSTADT

- **Micro-Averaging:** Evaluation function is applied to a global confusion matrix.

$$h = \delta \left(\begin{array}{c|c} 6 & 2 \\ \hline 2 & 6 \end{array} \right)$$

Attributes				Labels				
outlook	temp.	humidity	windy	play	dontplay	eaticecream	drinktee	
sunny	72	80	TRUE	0	1	1	0	covered
sunny	68	72	FALSE	0	1	1	0	
overcast	65	83	FALSE	1	0	0	0	not covered
rainy	64	75	FALSE	1	0	0	0	

3 Learning Multi-Label Head Rules

3.7 Different Averaging Strategies (2/4)

TECHNISCHE UNIVERSITÄT DARMSTADT

- **Label-based Averaging:** A confusion matrix is created per label. The evaluation function is applied to each one and the results are averaged.

$$h = \frac{1}{4} \cdot \left(\underbrace{\delta \begin{pmatrix} 2 & 2 \\ 0 & 0 \end{pmatrix}}_{\text{play}} + \underbrace{\delta \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}}_{\text{dontplay}} + \underbrace{\delta \begin{pmatrix} 0 & 0 \\ 2 & 2 \end{pmatrix}}_{\text{eaticecream}} + \underbrace{\delta \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}}_{\text{drinktee}} \right)$$

Attributes				Labels				
outlook	temp.	humidity	windy	play	dontplay	eaticecream	drinktee	
sunny	72	80	TRUE	0	1	1	0	covered
sunny	68	72	FALSE	0	1	1	0	
overcast	65	83	FALSE	1	0	0	0	not covered
rainy	64	75	FALSE	1	0	0	0	

3 Learning Multi-Label Head Rules

3.7 Different Averaging Strategies (3/4)

TECHNISCHE UNIVERSITÄT DARMSTADT

- **Example-based Averaging:** A confusion matrix is created per example. The evaluation function is applied to each one and the results are averaged.

$$h = \frac{1}{4} \cdot \left(\underbrace{\delta \begin{pmatrix} 3 & 0 \\ 1 & 0 \end{pmatrix}}_{\text{1st example}} + \underbrace{\delta \begin{pmatrix} 3 & 0 \\ 1 & 0 \end{pmatrix}}_{\text{2nd example}} + \underbrace{\delta \begin{pmatrix} 0 & 1 \\ 0 & 3 \end{pmatrix}}_{\text{3rd example}} + \underbrace{\delta \begin{pmatrix} 0 & 1 \\ 0 & 3 \end{pmatrix}}_{\text{4th example}} \right)$$

Attributes				Labels				
outlook	temp.	humidity	windy	play	dontplay	eaticecream	drinktee	
sunny	72	80	TRUE	0	1	1	0	covered
sunny	68	72	FALSE	0	1	1	0	
overcast	65	83	FALSE	1	0	0	0	not covered
rainy	64	75	FALSE	1	0	0	0	

3 Learning Multi-Label Head Rules

3.7 Different Averaging Strategies (4/4)

TECHNISCHE UNIVERSITÄT DARMSTADT

- **Macro-Averaging:** A confusion matrix is created per example and label. The evaluation function is applied to each one and the results are averaged example- and label-wise.

$$h = \frac{1}{16} \cdot \left(\underbrace{\delta\left(\begin{array}{c|c} 1 & 0 \\ \hline 0 & 0 \end{array}\right)}_{\text{1st example / play}} + \underbrace{\delta\left(\begin{array}{c|c} 1 & 0 \\ \hline 0 & 0 \end{array}\right)}_{\text{2nd example / play}} + \ldots + \underbrace{\delta\left(\begin{array}{c|c} 0 & 0 \\ \hline 0 & 1 \end{array}\right)}_{\text{4th example / drinktee}} \right)$$

Attributes				Labels				
outlook	temp.	humidity	windy	play	dontplay	eaticecream	drinktee	
sunny	72	80	TRUE	0	1	1	0	covered
sunny	68	72	FALSE	0	1	1	0	
overcast	65	83	FALSE	1	0	0	0	not covered
rainy	64	75	FALSE	1	0	0	0	

3 Learning Multi-Label Head Rules

3.8 Rule-dependent vs. Rule-independent Evaluation

TECHNISCHE UNIVERSITÄT DARMSTADT

- **Rule-dependent:** Only labels, which are predicted by a rule, are taken into account.
- **Rule-independent:** All labels are taken into account. Labels, which are not contained in the head, are considered as unpredicted.

Candidate rule:

$$play = 0, dontplay = 1 \leftarrow outlook = sunny$$

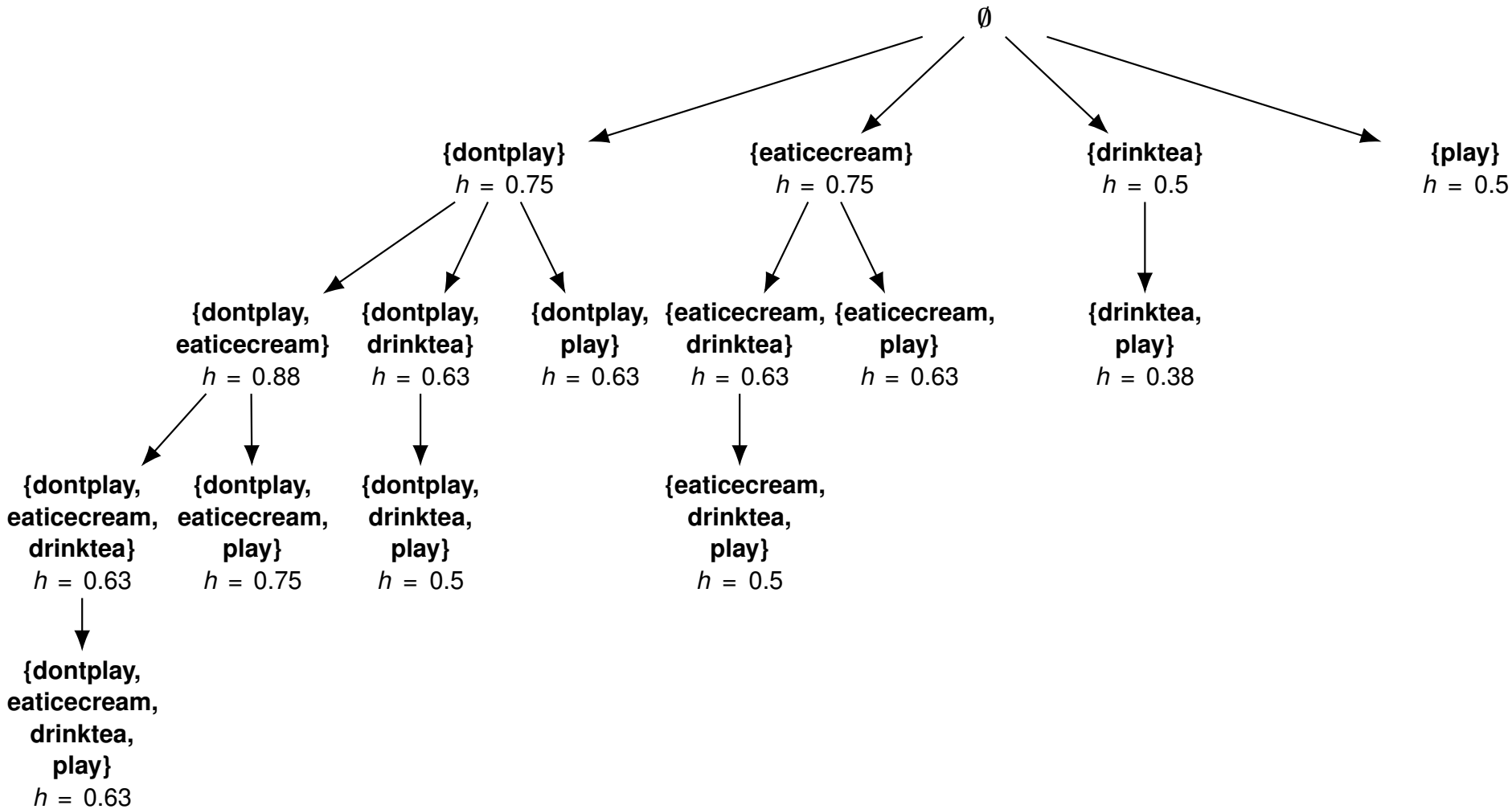
Training examples:

Attributes				Labels				
outlook	temp.	humidity	windy	play	dontplay	eaticecream	drinktee	
sunny	80	90	TRUE	0	1	1	0	covered
sunny	72	95	FALSE	0	1	1	0	
overcast	83	86	FALSE	1	0	0	0	not covered
rainy	75	80	FALSE	1	0	0	0	

3 Learning Multi-Label Head Rules

3.9 Heuristic Values of Potential Heads

TECHNISCHE UNIVERSITÄT DARMSTADT



3 Learning Multi-Label Head Rules

3.10 Pruning according to Anti-Monotonicity (1/2)

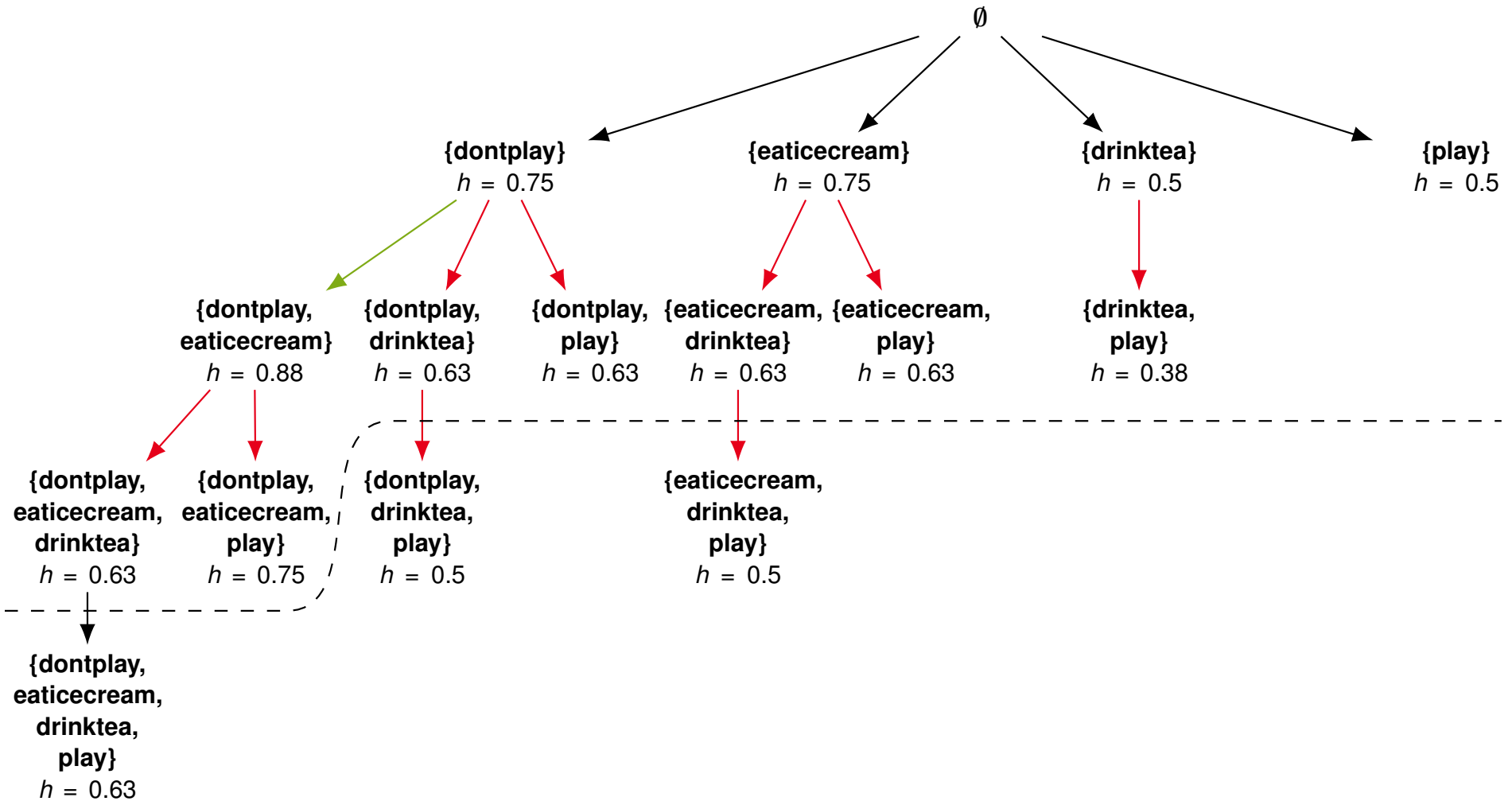
TECHNISCHE UNIVERSITÄT DARMSTADT

- **Definition "Anti-Monotonicity":**
  If adding a label to a rule's head causes the performance to decrease, by adding more labels the best possible performance cannot be reached anymore.

3 Learning Multi-Label Head Rules

3.10 Pruning according to Anti-Monotonicity (2/2)

TECHNISCHE UNIVERSITÄT DARMSTADT



3 Learning Multi-Label Head Rules

3.11 Pruning according to Decomposability (1/2)

TECHNISCHE UNIVERSITÄT DARMSTADT

- Imagine the following multi-label head rule:

$$\{y_1, ..., y_n\} \leftarrow body$$

- For each label a corresponding single-label head rule can be created:

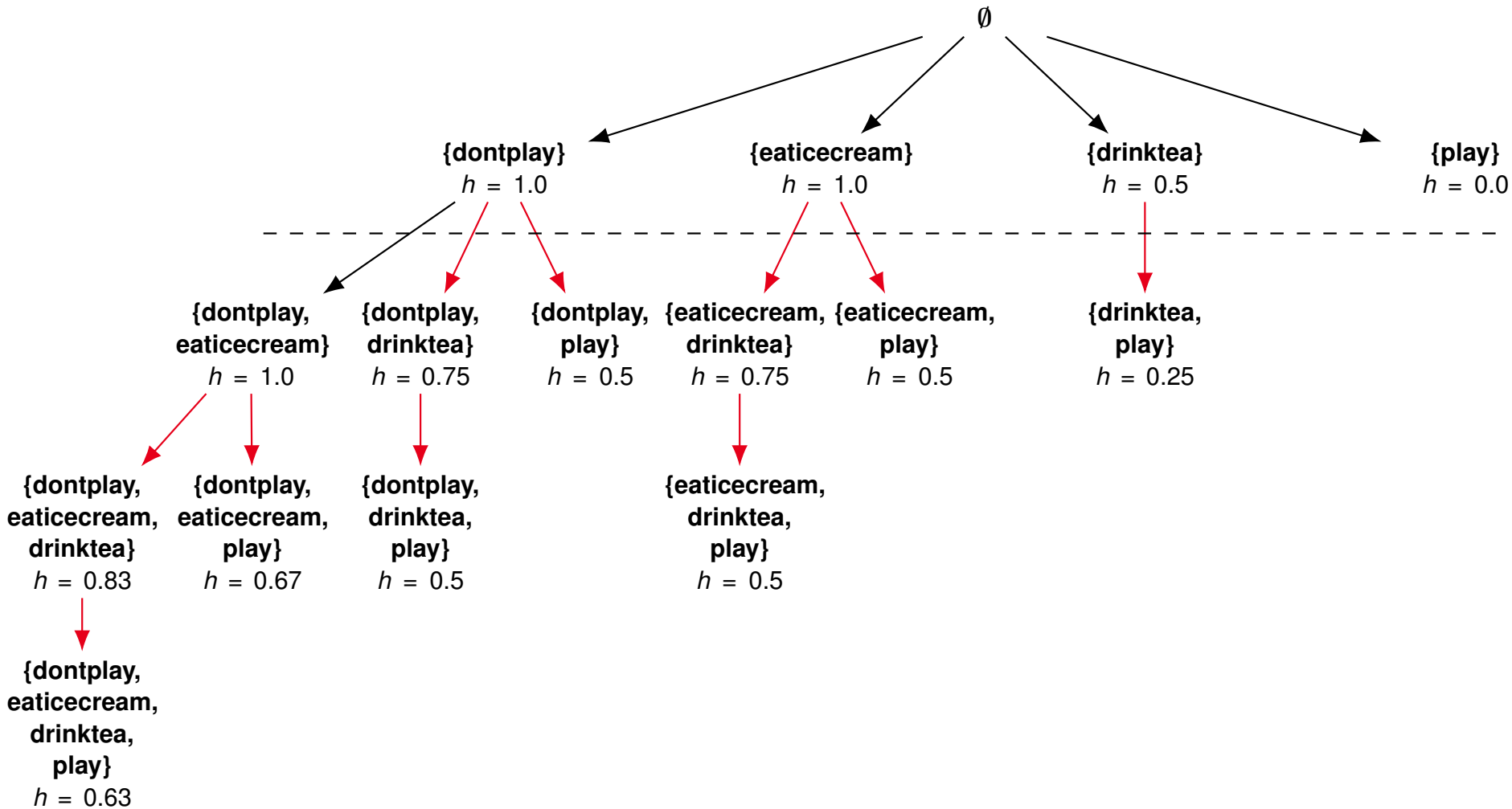
$$y_1 \leftarrow body, ..., y_n \leftarrow body$$

- **Definition "Decomposability"**:
  1. If all single-label head rules reach an equal performance, the corresponding multi-label head rule reaches that performance as well.
  2. If the performance of at least one single-label head rule is less than the best possible performance, the corresponding multi-label head rule does not reach that performance either.

3 Learning Multi-Label Head Rules

3.11 Pruning according to Decomposability (2/2)

TECHNISCHE UNIVERSITÄT DARMSTADT



3 Learning Multi-Label Head Rules

3.12 Results of Formal Examinations (1/2)

TECHNISCHE UNIVERSITÄT DARMSTADT

Evaluation Function	Evaluation Strategy	Averaging Strategy	Anti-Monotonicity	Decomposability
Precision	Rule-dependent	Micro-Averaging	Yes	Yes
		Label-based Averaging	Yes	Yes
		Example-based Averaging	Yes	Yes
		Macro-Averaging	Yes	Yes
	Rule-independent	Micro-Averaging	Yes	-
		Label-based Averaging	Yes	-
		Example-based Averaging	Yes	-
		Macro-Averaging	Yes	-
Recall	Rule-dependent	Micro-Averaging	Yes	Yes
		Label-based Averaging	Yes	Yes
		Example-based Averaging	-	-
		Macro-Averaging	Yes	Yes
	Rule-independent	Micro-Averaging	Yes	-
		Label-based Averaging	Yes	-
		Example-based Averaging	-	-
		Macro-Averaging	Yes	-

3 Learning Multi-Label Head Rules

3.12 Results of Formal Examinations (2/2)

TECHNISCHE UNIVERSITÄT DARMSTADT

Evaluation Function	Evaluation Strategy	Averaging Strategy	Anti-Monotonicity	Decomposability
Hamming Accuracy	Rule-dependent	Micro-Averaging	Yes	Yes
		Label-based Averaging	Yes	Yes
		Example-based Averaging	Yes	Yes
		Macro-Averaging	Yes	Yes
	Rule-independent	Micro-Averaging	Yes	-
		Label-based Averaging	Yes	-
		Example-based Averaging	Yes	-
		Macro-Averaging	Yes	-
F-Measure	Rule-dependent	Micro-Averaging	Yes	Yes
		Label-based Averaging	Yes	Yes
		Example-based Averaging	Yes	Yes
		Macro-Averaging	Yes	Yes
	Rule-independent	Micro-Averaging	Yes	-
		Label-based Averaging	Yes	-
		Example-based Averaging	Yes	-
		Macro-Averaging	Yes	-
Subset Accuracy	Rule-dependent	Example-based Averaging	Yes	-
	Rule-independent		-	-

Contents

TECHNISCHE UNIVERSITÄT DARMSTADT

4 Evaluation

4.1 Overview of Statistical Experiments

TECHNISCHE UNIVERSITÄT DARMSTADT

- Comparison with the original algorithm and a "Binary Relevance" approach.
- Evaluation on 5 different data sets using
  - Precision, Hamming Accuracy, $F_{0.5}$-Measure and Subset Accuracy
  - Rule-dependent and Rule-independent Evaluation
  - Different Averaging Strategies

Approach	Hamm. Acc.	Subset Acc.	Ex.-based Prec.	Ex.-based Rec.	Ex.-based F1	Mi. Prec.	Mi. Rec	Mi. F1	Avg. Rank
F-Measure									
BR	93.91% 7	41.49% 9	56.87% 5	56.04% 9	55.22% 6	40.56% 5	41.85% 10	41.85% 8	4
Single+$_{stop}$	89.93% 14	2.48% 16	14.27% 16	19.39% 21	15.56% 16	22.12% 12	38.66% 17	38.66% 15	19
Single+	84.41% 17	0.31% 18	9.43% 17	23.47% 20	12.73% 17	15.96% 15	48.24% 7	48.24% 5	18
Multi+$_{stop,mn}$	93.48% 12.5	39.94% 11.5	54.23% 8.5	55.74% 10.5	53.49% 8.5	37.24% 10.5	40.58% 13.5	40.58% 11.5	12.5
Multi+$_{stop,mM}$	93.48% 12.5	39.94% 11.5	54.23% 8.5	55.74% 10.5	53.49% 8.5	37.24% 10.5	40.58% 13.5	40.58% 11.5	12.5
Multi+$_{stop,Mn}$	5.39% 20.5	0.31% 18	5.39% 20.5	53.25% 12.5	9.41% 20.5	5.12% 18.5	100.00% 1.5	9.37% 16.5	20.5
Multi+$_{stop,MM}$	5.39% 20.5	0.31% 18	5.39% 20.5	53.25% 12.5	9.41% 20.5	5.12% 18.5	100.00% 1.5	9.37% 16.5	20.5
Multi+$_{mn}$	86.18% 15.5	32.51% 14.5	42.34% 14.5	60.53% 1.5	45.92% 14.5	20.51% 13.5	59.42% 5.5	59.42% 3.5	10.5
Multi+$_{mM}$	86.18% 15.5	32.51% 14.5	42.34% 14.5	60.53% 1.5	45.92% 14.5	20.51% 13.5	59.42% 5.5	59.42% 3.5	10.5
Multi+$_{Mn}$	32.69% 18.5	0.00% 20.5	6.72% 18.5	49.07% 14.5	11.49% 18.5	6.63% 16.5	93.29% 3.5	93.29% 1.5	15.5
Multi+$_{MM}$	32.69% 18.5	0.00% 20.5	6.72% 18.5	49.07% 14.5	11.49% 18.5	6.63% 16.5	93.29% 3.5	93.29% 1.5	15.5
Single±$_{stop}$	94.02% 6	39.32% 13	55.46% 7	56.79% 8	54.57% 7	41.35% 4	41.21% 11.5	41.21% 9.5	5
Single±	94.10% 5	41.18% 10	56.79% 6	57.56% 7	55.65% 5	42.02% 3	41.21% 11.5	41.21% 9.5	3
Multi±$_{stop,mn}$	93.87% 8.5	43.03% 6.5	59.58% 1.5	59.33% 3.5	57.90% 1.5	40.54% 6.5	43.13% 8.5	43.13% 6.5	1.5
Multi±$_{stop,mM}$	93.87% 8.5	43.03% 6.5	59.58% 1.5	59.33% 3.5	57.90% 1.5	40.54% 6.5	43.13% 8.5	43.13% 6.5	1.5
Multi±$_{stop,Mn}$	94.92% 1.5	47.06% 1.5	47.37% 10.5	47.21% 16.5	47.27% 10.5	66.67% 1.5	0.64% 18.5	1.27% 18.5	9
Multi±$_{stop,MM}$	94.92% 1.5	47.06% 1.5	47.37% 10.5	47.21% 16.5	47.27% 10.5	66.67% 1.5	0.64% 18.5	1.27% 18.5	8
Multi±$_{mn}$	93.74% 10.5	43.03% 6.5	57.60% 3.5	58.11% 5.5	56.44% 3.5	38.73% 8.5	38.98% 15.5	38.98% 13.5	6.5
Multi±$_{mM}$	93.74% 10.5	43.03% 6.5	57.60% 3.5	58.11% 5.5	56.44% 3.5	38.73% 8.5	38.98% 15.5	38.98% 13.5	6.5
Multi±$_{Mn}$	94.90% 3.5	46.75% 3.5	46.75% 12.5	46.75% 18.5	46.75% 12.5	0.00% 20.5	0.00% 20.5	0.00% 20.5	15.5
Multi±$_{MM}$	94.90% 3.5	46.75% 3.5	46.75% 12.5	46.75% 18.5	46.75% 12.5	0.00% 20.5	0.00% 20.5	0.00% 20.5	15.5
Hamming Accuracy									
BR	93.16% 15	39.01% 11	52.59% 11	53.51% 15	51.66% 11	51.66% 11	39.62% 7	37.13% 5	11
Single+$_{stop}$	67.05% 21	1.55% 20	12.41% 20	28.78% 20	15.06% 20	15.06% 20	56.23% 1	14.83% 21	20
Single+	85.25% 20	0.00% 21	10.84% 21	25.03% 21	14.42% 21	14.42% 21	47.28% 6	24.65% 20	21
Multi+$_{stop,mn}$	93.29% 12.5	37.46% 13.5	51.68% 13.5	51.34% 17.5	50.16% 13.5	50.16% 13.5	35.14% 9.5	34.81% 17.5	17.5
Multi+$_{stop,mM}$	93.29% 12.5	37.46% 13.5	51.68% 13.5	51.34% 17.5	50.16% 13.5	50.16% 13.5	35.14% 9.5	34.81% 17.5	17.5
Multi+$_{stop,Mn}$	93.29% 12.5	37.46% 13.5	51.68% 13.5	51.34% 17.5	50.16% 13.5	50.16% 13.5	35.14% 9.5	34.81% 17.5	17.5
Multi+$_{stop,MM}$	93.29% 12.5	37.46% 13.5	51.68% 13.5	51.34% 17.5	50.16% 13.5	50.16% 13.5	35.14% 9.5	34.81% 17.5	17.5
Multi+$_{mn}$	92.10% 17.5	34.06% 17.5	49.35% 17.5	55.00% 12.5	50.06% 17.5	50.06% 17.5	47.92% 3.5	38.22% 2.5	13.5
Multi+$_{mM}$	92.10% 17.5	34.06% 17.5	49.35% 17.5	55.00% 12.5	50.06% 17.5	50.06% 17.5	47.92% 3.5	38.22% 2.5	13.5
Multi+$_{Mn}$	92.10% 17.5	34.06% 17.5	49.35% 17.5	55.00% 12.5	50.06% 17.5	50.06% 17.5	47.92% 3.5	38.22% 2.5	13.5
Multi+$_{MM}$	92.10% 17.5	34.06% 17.5	49.35% 17.5	55.00% 12.5	50.06% 17.5	50.06% 17.5	47.92% 3.5	38.22% 2.5	13.5
Single±$_{stop}$	94.25% 1.5	46.44% 1.5	59.83% 1.5	56.37% 1.5	56.72% 1.5	56.72% 1.5	31.95% 20.5	36.17% 14.5	1.5
Single±	94.25% 1.5	46.44% 1.5	59.83% 1.5	56.37% 1.5	56.72% 1.5	56.72% 1.5	31.95% 20.5	36.17% 14.5	1.5
Multi±$_{stop,mn}$	94.18% 6.5	44.27% 6.5	58.46% 6.5	55.52% 6.5	55.58% 6.5	55.58% 6.5	32.91% 15.5	36.59% 9.5	6.5
Multi±$_{stop,mM}$	94.18% 6.5	44.27% 6.5	58.46% 6.5	55.52% 6.5	55.58% 6.5	55.58% 6.5	32.91% 15.5	36.59% 9.5	6.5
Multi±$_{stop,Mn}$	94.18% 6.5	44.27% 6.5	58.46% 6.5	55.52% 6.5	55.58% 6.5	55.58% 6.5	32.91% 15.5	36.59% 9.5	6.5
Multi±$_{stop,MM}$	94.18% 6.5	44.27% 6.5	58.46% 6.5	55.52% 6.5	55.58% 6.5	55.58% 6.5	32.91% 15.5	36.59% 9.5	6.5
Multi±$_{mn}$	94.18% 6.5	44.27% 6.5	58.46% 6.5	55.52% 6.5	55.58% 6.5	55.58% 6.5	32.91% 15.5	36.59% 9.5	6.5
Multi±$_{mM}$	94.18% 6.5	44.27% 6.5	58.46% 6.5	55.52% 6.5	55.58% 6.5	55.58% 6.5	32.91% 15.5	36.59% 9.5	6.5
Multi±$_{Mn}$	94.18% 6.5	44.27% 6.5	58.46% 6.5	55.52% 6.5	55.58% 6.5	55.58% 6.5	32.91% 15.5	36.59% 9.5	6.5
Multi±$_{MM}$	94.18% 6.5	44.27% 6.5	58.46% 6.5	55.52% 6.5	55.58% 6.5	55.58% 6.5	32.91% 15.5	36.59% 9.5	6.5
Precision									
BR	92.50% 11	38.70% 11	51.02% 7	52.87% 11	50.35% 7	30.81% 9	37.70% 8.5	37.70% 8.5	7
Single+$_{stop}$	89.23% 14	2.48% 16	14.83% 16	20.06% 21	15.88% 16	20.21% 12	37.70% 8.5	37.70% 8.5	17
Single+	74.71% 17	0.31% 18	7.20% 17	31.42% 20	11.08% 19	11.97% 15	62.30% 5	62.30% 5	21
Multi+$_{stop,mn}$	92.29% 12.5	34.67% 12.5	48.53% 8.5	50.17% 12.5	47.68% 8.5	28.49% 10.5	33.87% 10.5	33.87% 10.5	10.5
Multi+$_{stop,mM}$	92.29% 12.5	34.67% 12.5	48.53% 8.5	50.17% 12.5	47.68% 8.5	28.49% 10.5	33.87% 10.5	33.87% 10.5	10.5
Multi+$_{stop,Mn}$	5.39% 20.5	0.31% 18	5.39% 20.5	53.25% 9.5	9.41% 20.5	5.12% 18.5	100.00% 1.5	100.00% 1.5	14.5
Multi+$_{stop,MM}$	5.39% 20.5	0.31% 18	5.39% 20.5	53.25% 9.5	9.41% 20.5	5.12% 18.5	100.00% 1.5	100.00% 1.5	14.5
Multi+$_{mn}$	85.92% 15.5	33.75% 14.5	42.78% 14.5	57.08% 6.5	45.63% 14.5	17.63% 13.5	47.92% 6.5	47.92% 6.5	12.5
Multi+$_{mM}$	85.92% 15.5	33.75% 14.5	42.78% 14.5	57.08% 6.5	45.63% 14.5	17.63% 13.5	47.92% 6.5	47.92% 6.5	12.5
Multi+$_{Mn}$	32.69% 18.5	0.00% 20.5	6.72% 18.5	49.07% 14.5	11.49% 17.5	6.63% 16.5	93.29% 3.5	93.29% 3.5	19.5
Multi+$_{MM}$	32.69% 18.5	0.00% 20.5	6.72% 18.5	49.07% 14.5	11.49% 17.5	6.63% 16.5	93.29% 3.5	93.29% 3.5	19.5
Single±$_{stop}$	94.15% 6	45.51% 8.5	58.23% 6	56.68% 8	56.26% 6	40.80% 4	32.59% 17	32.59% 17	6
Single±	94.43% 5	45.51% 8.5	59.29% 1	57.62% 5	57.12% 3	43.93% 3	33.55% 13	33.55% 13	1
Multi±$_{stop,mn}$	93.32% 9.5	45.51% 8.5	59.28% 2.5	58.47% 3.5	56.93% 4.5	34.10% 7.5	33.23% 15.5	33.23% 15.5	4.5
Multi±$_{stop,mM}$	93.32% 9.5	45.51% 8.5	59.28% 2.5	58.47% 3.5	56.93% 4.5	34.10% 7.5	33.23% 15.5	33.23% 15.5	4.5
Multi±$_{stop,Mn}$	94.92% 1.5	47.06% 1.5	47.37% 10.5	47.21% 16.5	47.27% 10.5	66.67% 1.5	0.64% 18.5	0.64% 18.5	8.5
Multi±$_{stop,MM}$	94.92% 1.5	47.06% 1.5	47.37% 10.5	47.21% 16.5	47.27% 10.5	66.67% 1.5	0.64% 18.5	0.64% 18.5	8.5
Multi±$_{mn}$	93.79% 7.5	46.44% 5.5	58.24% 4.5	59.34% 1.5	57.21% 1.5	37.77% 5.5	33.55% 13	33.55% 13	2.5
Multi±$_{mM}$	93.79% 7.5	46.44% 5.5	58.24% 4.5	59.34% 1.5	57.21% 1.5	37.77% 5.5	33.55% 13	33.55% 13	2.5
Multi±$_{Mn}$	94.90% 3.5	46.75% 3.5	46.75% 12.5	46.75% 18.5	46.75% 12.5	0.00% 20.5	0.00% 20.5	0.00% 20.5	17
Multi±$_{MM}$	94.90% 3.5	46.75% 3.5	46.75% 12.5	46.75% 18.5	46.75% 12.5	0.00% 20.5	0.00% 20.5	0.00% 20.5	17
Subset Accuracy									
Multi+$_{stop,Mn}$	93.30% 3	38.39% 3	52.51% 3	53.13% 3	51.46% 3	35.24% 3	37.38% 2	36.28% 4	3
Multi+$_{Mn}$	91.49% 4	30.65% 4	46.03% 4	52.42% 4	46.93% 4	29.63% 4	48.56% 1	36.80% 1	4
Multi±$_{stop,Mn}$	94.18% 1.5	44.27% 1.5	58.46% 1.5	55.52% 1.5	55.58% 1.5	41.20% 1.5	32.91% 3.5	36.59% 2.5	1.5
Multi±$_{Mn}$	94.18% 1.5	44.27% 1.5	58.46% 1.5	55.52% 1.5	55.58% 1.5	41.20% 1.5	32.91% 3.5	36.59% 2.5	1.5

TECHNISCHE UNIVERSITÄT DARMSTADT

- Using **micro- or label-based averaging** results in the best performance.
- When using micro- or label-based averaging, the algorithm tends to benefit from **predicting irrelevant labels**.
- Using the **rule-dependent evaluation** is usually a better choice than using the rule-independent evaluation.
- The algorithm always **outperforms the "Binary Relevance" approach**, if label dependencies are given.
- The algorithm **is able to compete with the original approach**. In some cases it is even ranked higher.

TECHNISCHE UNIVERSITÄT DARMSTADT

- Using **macro- or example-based averaging** results in less rules being learned.
- **Predicting irrelevant labels** causes the number of label-dependent rules to increase.
- The amount of **label-dependent rules** is usually similar to the original approach. Multi-label head rules can be considered as an alternative.
- Significantly more multi-label head rules are learned when using the precision metric.

Thank You!

TECHNISCHE
UNIVERSITÄT
DARMSTADT

Any Questions?