Optimizing the AUC with Rule Learning

TECHNISCHE UNIVERSITÄT DARMSTADT

Julius Stecher

Prof. Johannes Fürnkranz
Knowledge Engineering Group

Table of Contents

- **Separate-and-Conquer Rule Learning**
  - Heuristic Rule Learning
  - Basic algorithm
- **Optimization approach**
  - Modification of the basic algorithm
  - Specialized refinement heuristics
- **Experiments and Analysis**
  - Accuracy on 19 datasets
  - AUC on 7 binary-class datasets
- **Concluding remarks**

Separate-and-Conquer Rule Learning
Rule Learning

- **Belongs to machine learning field**
- **Classification Problem: Given training and testing data**
  - Algorithmically find rules based on training data
  - Rules can then be applied to new unlabeled testing data
  - Rules are of the form R: <class label> := {$cond_1$,$cond_2$, … ,$cond_n$}
  - Rule *fires* when conditions apply to example's attributes
- **Multiple ways to build a theory**
  - Decision list: Check rules in a set order, apply first one that fires
  - Rule set: Combine all available rules for classification
  - Here: *decision lists*

Separate-and-Conquer Rule Learning
Top-Down Rule Learning

- **Algorithm used is *Top-Down Hill-Climbing* Rule Learner**
- **General Procedure**
  - Start with the universal rule <majority class> := {} and empty theory T
  - Create set of possible refinements
    - Refinements consist of one single condition, e.g. „age <= 22“ or „color = red“
    - Adding refinements *specializes* the rule successively
    - Decrease *coverage*, increase *consistency* (ideally)
  - Evaluate refinements according to the heuristic used
  - Add best condition, proceed to refine if applicable
  - Add the best known rule to the theory T according to the heuristic used
    - Else go back to the refining step

Separate-and-Conquer Rule Learning
Separate-and Conquer Rule Learning

TECHNISCHE UNIVERSITÄT DARMSTADT

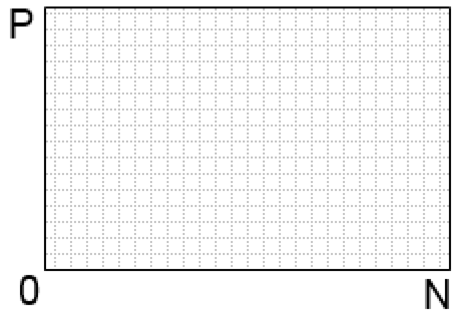
- **Idea:**
  - *Conquer* groups of training examples rule after rule...
  - By *separating* already conquered rules...
    - Into groups of rules that can be explained by one single rule
    - Successively adding rules to a decision list
    - Until we are satisfied with the theory learned
- **Greedy approach**
  - Requires on-the-fly performance estimates
- **Driven by *rule learning heuristics***
- **Term coined by Pagallo / Haussler (1990)**
  - a.k.a. „covering strategy"

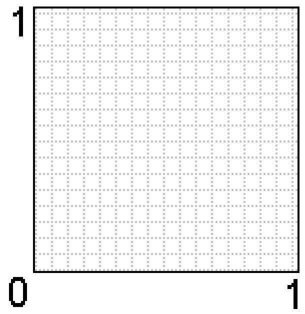
Separate-and-Conquer Rule Learning
Heuristic Rule Learning

- **Evaluating refinements and comparing whole rules:**
  - Requires on-the-fly performance assessment
  - Solution: rule learning heuristics
- **Generalized definition of heuristics**
  - h: Rule → [0,1]
  - Rules provide statistics in the form of a confusion matrix

	Classified positive	Classified negative	
+	**true positives**	**false negatives**	**P**
–	**false positives**	**true negatives**	**N**
			P+N

- **Given a confusion matrix, the following visualization is applicable:**



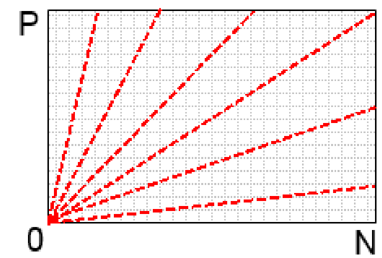


- **ROC space is normalized**
  - false positive rate *(fpr)* on x-axis
  - true positive rate *(tpr)* on y-axis

Separate-and-Conquer Rule Learning
Heuristics and Isometrics

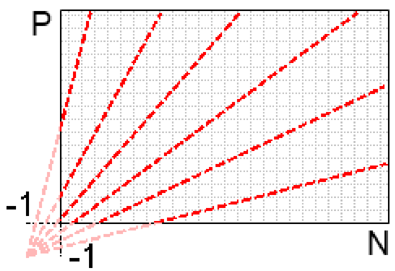
- **Precision :**

$$h_{prec}(p,n) = \frac{p}{p+n}$$



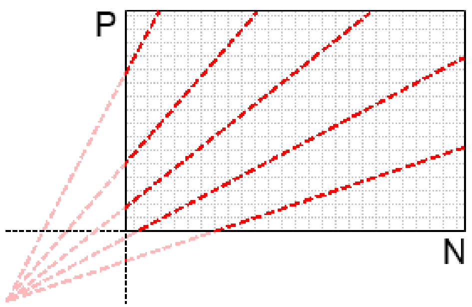
- **Laplace**

$$h_{lap}(p,n) = \frac{p+1}{p+n+2}$$



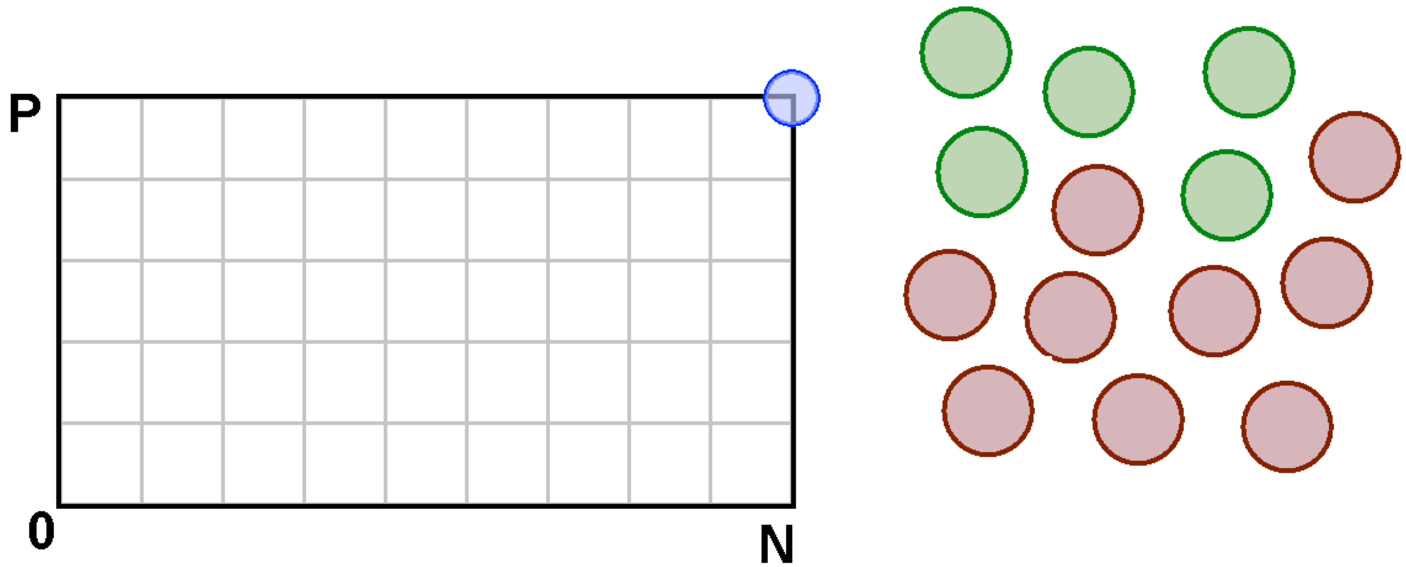
- **m- Estimate:**

$$h_{mest}(p,n) = \frac{p+m\cdot\frac{P}{P+N}}{p+n+m}$$



Separate-and-Conquer Rule Learning
Basic Algorithm

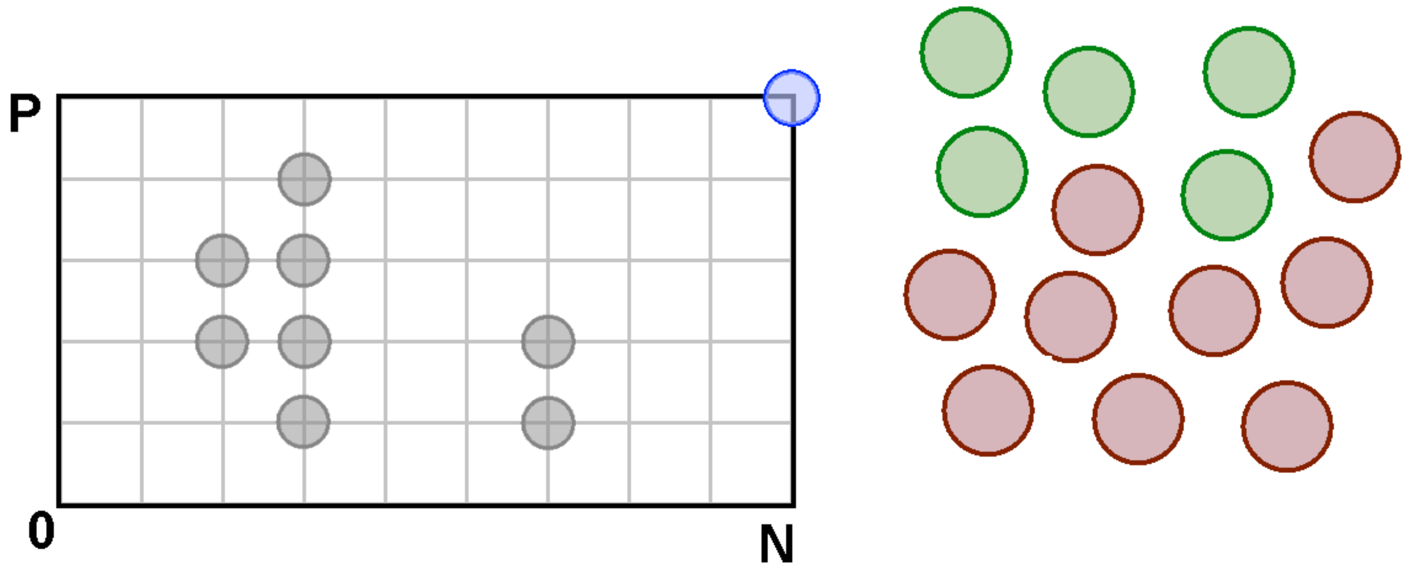
- **Short 14 instances example** *(weather.nominal.arff dataset)*



Top-Down Learner: begin with refining universal rule

Separate-and-Conquer Rule Learning
Basic Algorithm

- **Short 14 instances example** *(weather.nominal.arff dataset)*

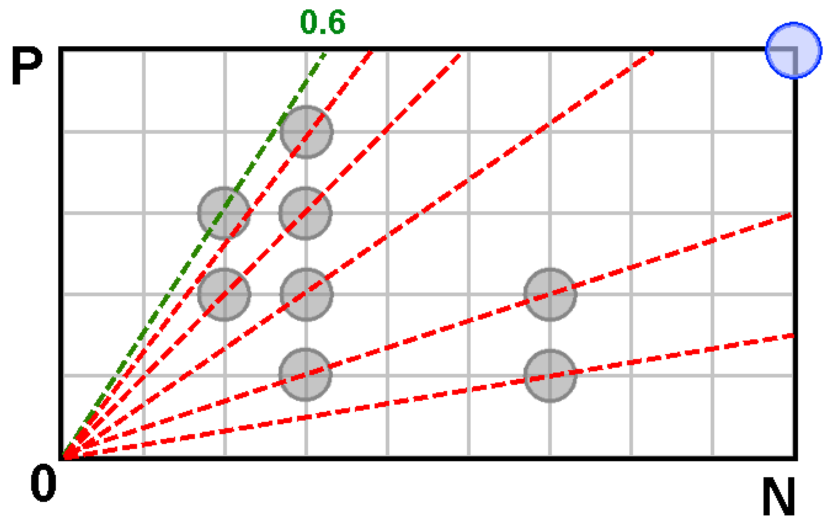


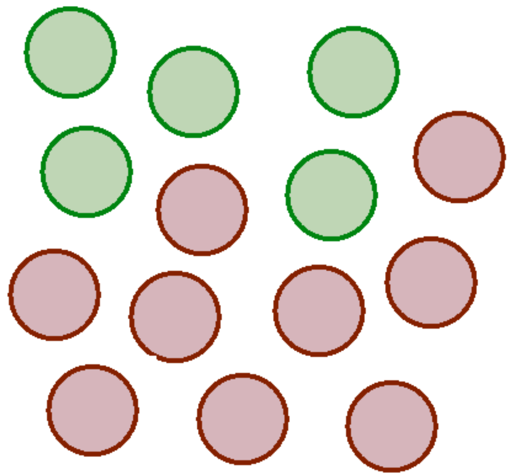
Top-Down Learner: begin with refining universal rule

List all possible refinements

Separate-and-Conquer Rule Learning
Basic Algorithm

- **Short 14 instances example** *(weather.nominal.arff dataset)*





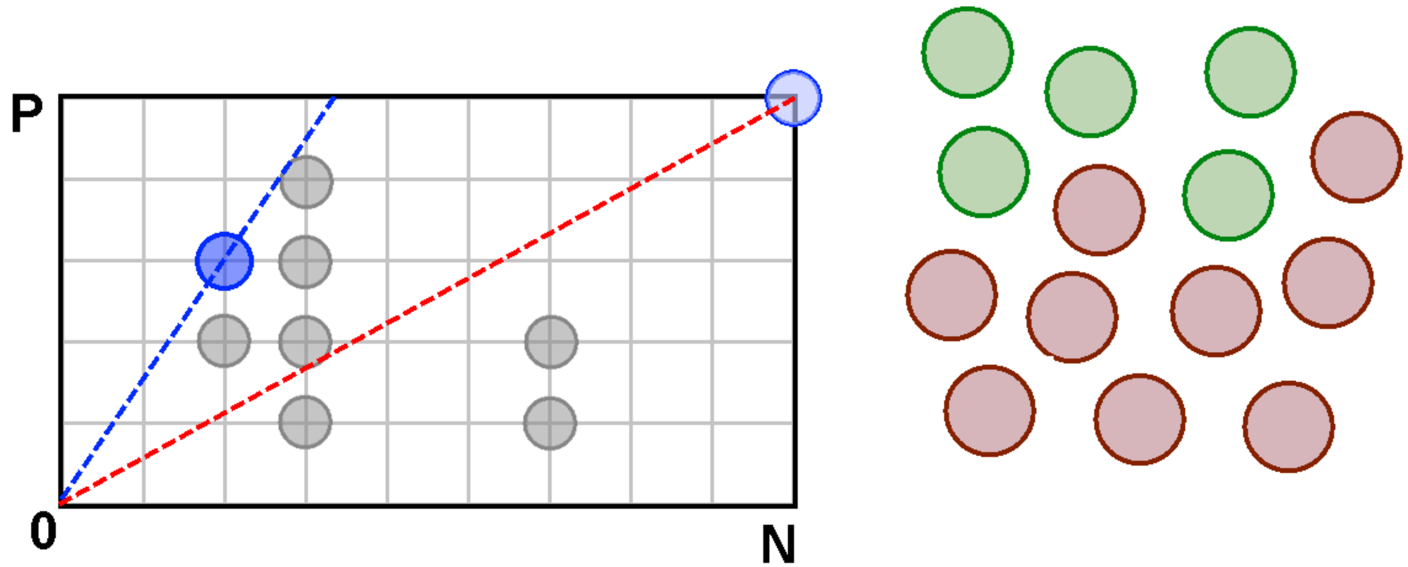
Top-Down Learner: begin with refining universal rule

List all possible refinements

Evaluate refinements and choose best via heuristic

Separate-and-Conquer Rule Learning
Basic Algorithm

- **Short 14 instances example** *(weather.nominal.arff dataset)*



Top-Down Learner: begin with refining universal rule

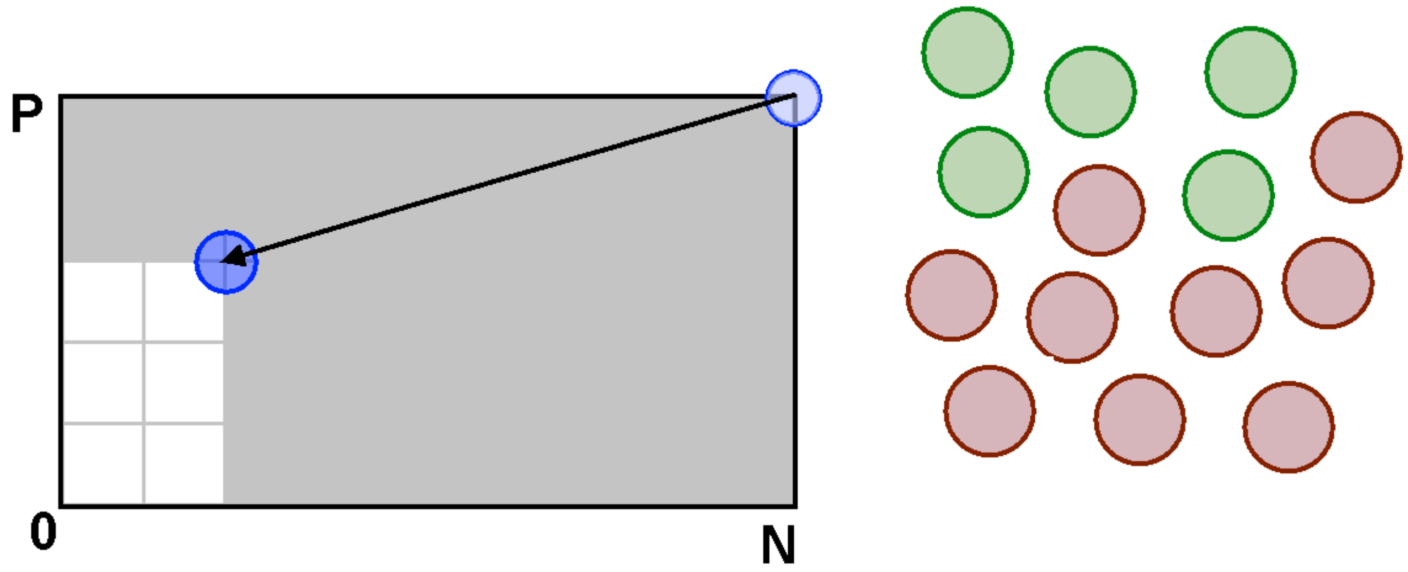
List all possible refinements

Evaluate refinements and choose best via heuristic

Compare rules and choose best via heuristc

Separate-and-Conquer Rule Learning
Basic Algorithm

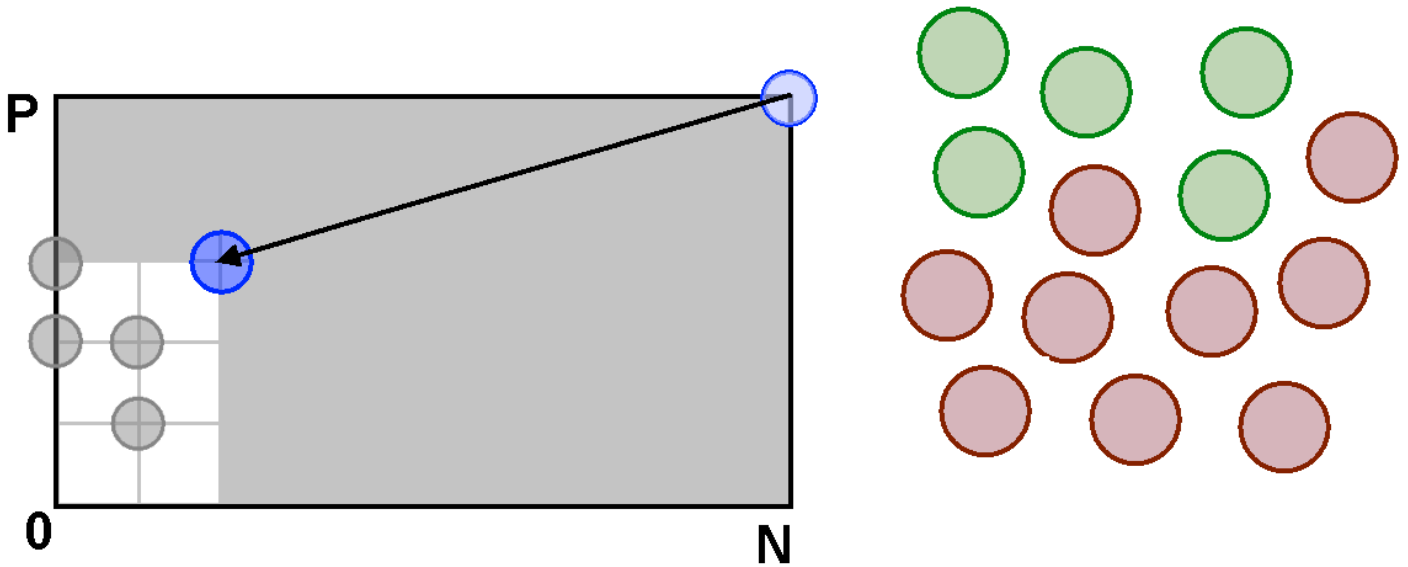
- **Short 14 instances example** *(weather.nominal.arff dataset)*



Continue: refine the current best rule

Separate-and-Conquer Rule Learning
Basic Algorithm

- **Short 14 instances example** *(weather.nominal.arff dataset)*



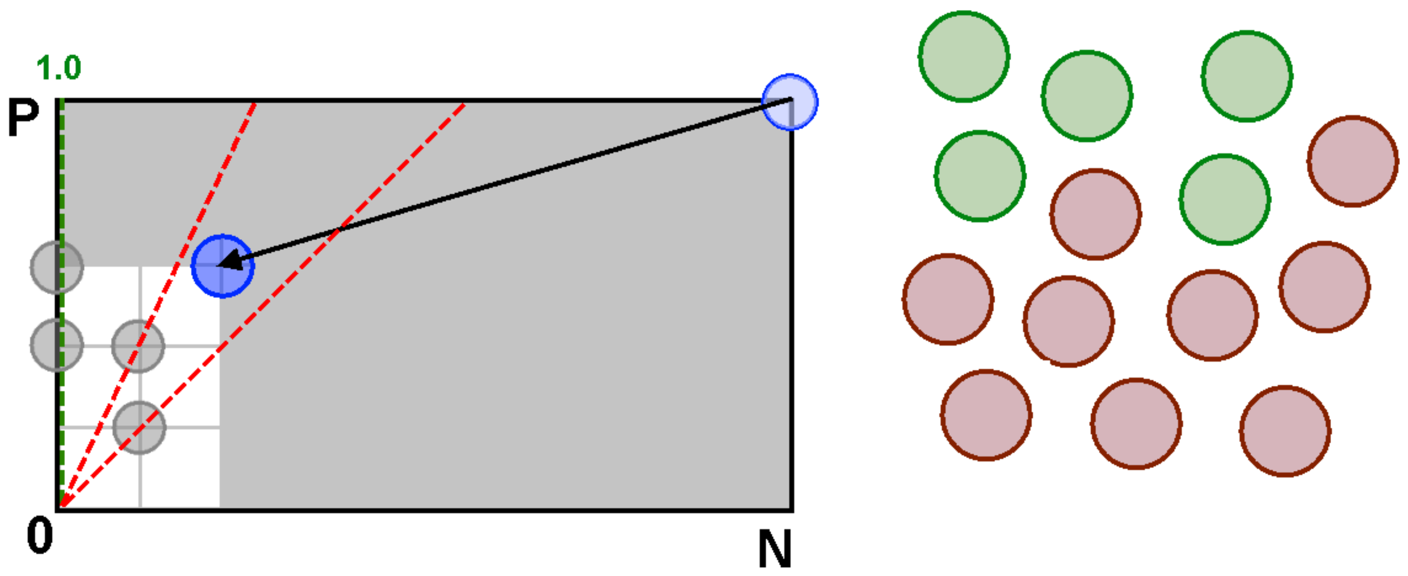
Continue: refine the current best rule

List all possible refinements

Separate-and-Conquer Rule Learning
Basic Algorithm

TECHNISCHE UNIVERSITÄT DARMSTADT

- **Short 14 instances example** *(weather.nominal.arff dataset)*



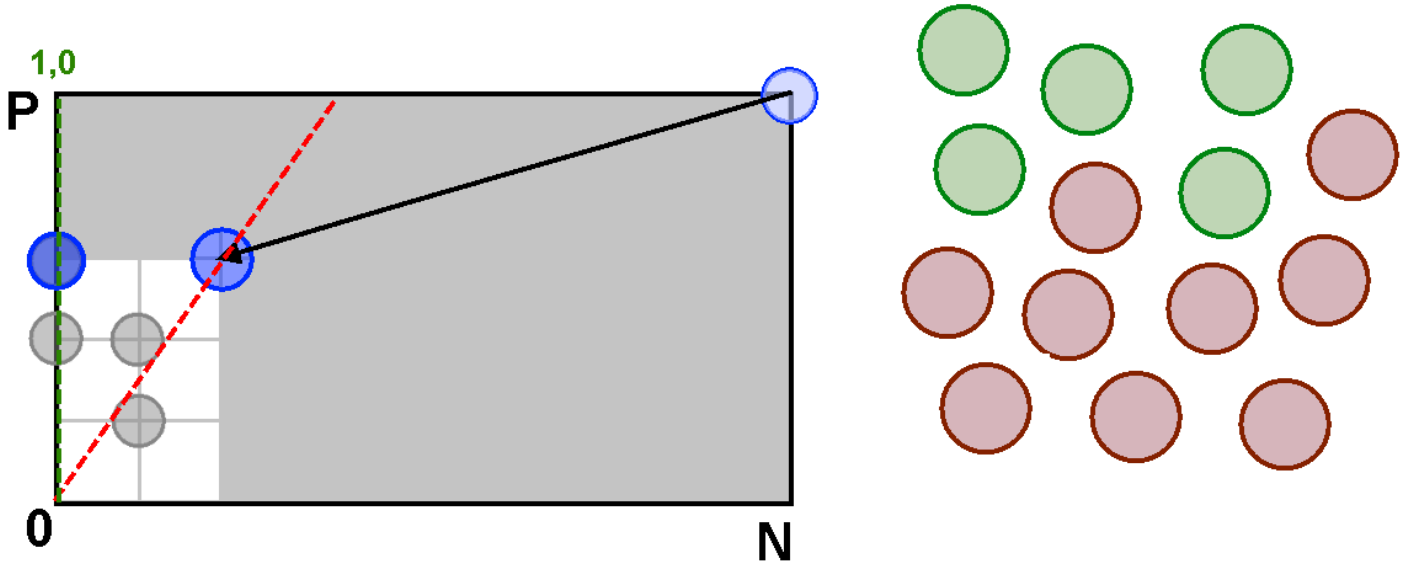
Continue: refine the current best rule

List all possible refinements

Evaluate refinements and choose best via heuristic

Separate-and-Conquer Rule Learning
Basic Algorithm

- **Short 14 instances example** *(weather.nominal.arff dataset)*



Continue: refine the current best rule

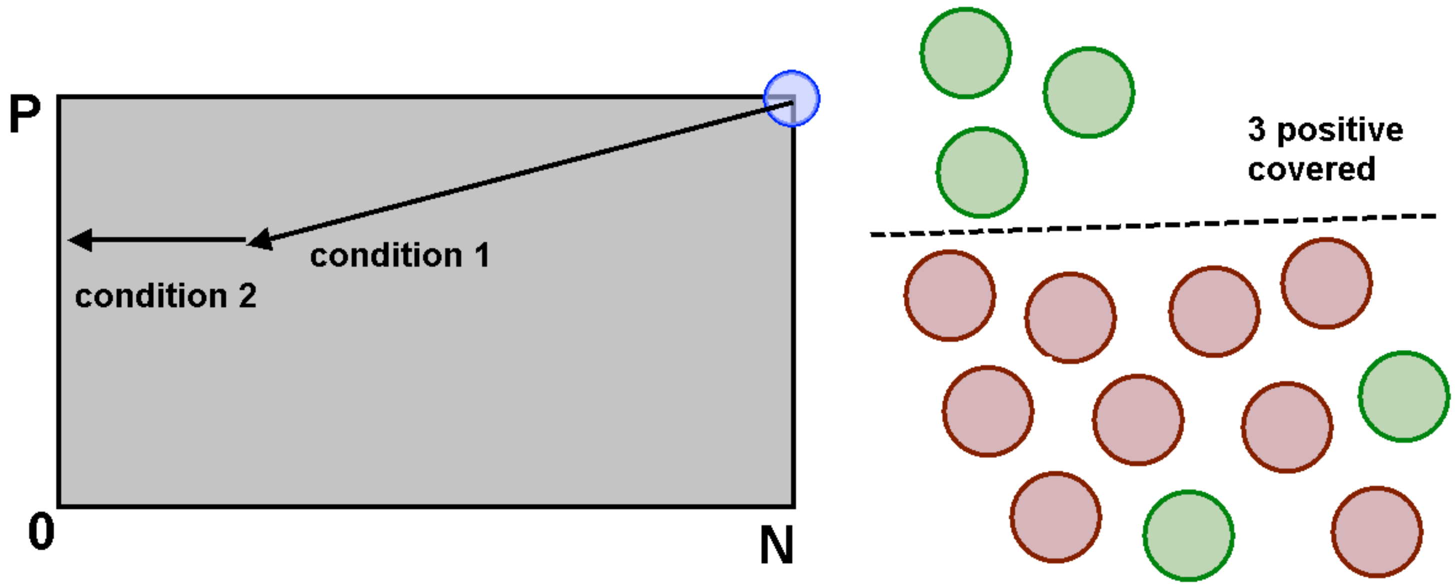
List all possible refinements

Evaluate refinements and choose best via heuristic

Compare rules and choose best via heuristc

Separate-and-Conquer Rule Learning
Basic Algorithm

- **Short 14 instances example** *(weather.nominal.arff dataset)*



Finished learning the rule, adding rule to theory
Conquering group of examples
Proceed to learn another rule on the rest

positives left --> learn another rule

Optimization Approach

- **Outline:**
  - Change the way rule refinements are evaluated
  - Use a secondary heuristic specifically for rule refinement
  - Keep the heuristic used for rule comparison
- **Goal:**
  - Select the best refinement based on minimal loss of positives
  - Try to build rules that explain a lot of data (coverage)
    - Preferably mostly positive data (consistency)
    - Coverage Space progression: go from n=N to n=0 in few meaningful steps
    - Do not „loose" too many positives in the process (keep height on p axis)

Optimization Approach
Modification of the Basic Algorithm

General Procedure

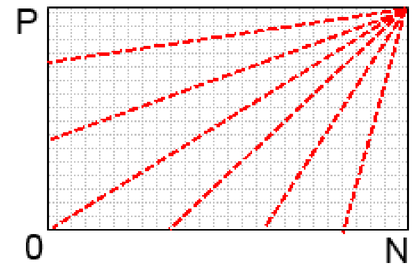
- Start with the universal rule <majority class> := {} and empty theory T
- Create set of possible refinements
  - Refinements consist of one single condition, e.g. „age <= 22“ or „color = red“
  - Adding refinements *specializes* the rule successively
  - Decrease *coverage*, increase *consistency* (ideally)
- Evaluate refinements according to the *rule refinement heuristic*
- Add best condition, proceed to refine if applicable
- Add the best known rule to the theory T according to the *rule selection heuristic*
  - Else go back to the refining step

Specialized Refinement Heuristics

TECHNISCHE UNIVERSITÄT DARMSTADT

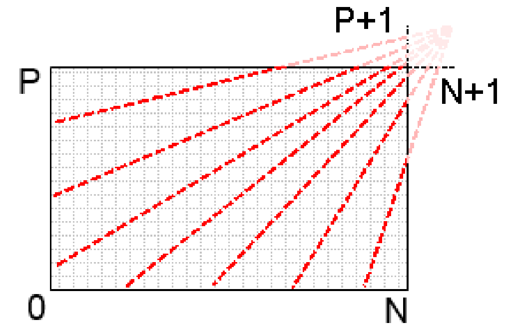
- **Modified precision :**

$$h'_{prec}(p,n,P,N) = \frac{N-n}{(P+N)-(p+n)}$$



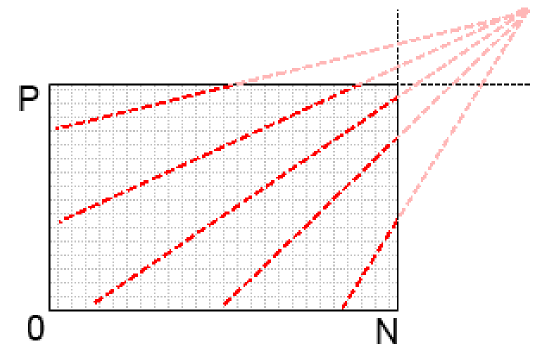
- **Modified laplace:**

$$h'_{lap}(p,n,P,N) = \frac{N-n+1}{(P+N)-(p+n-2)}$$



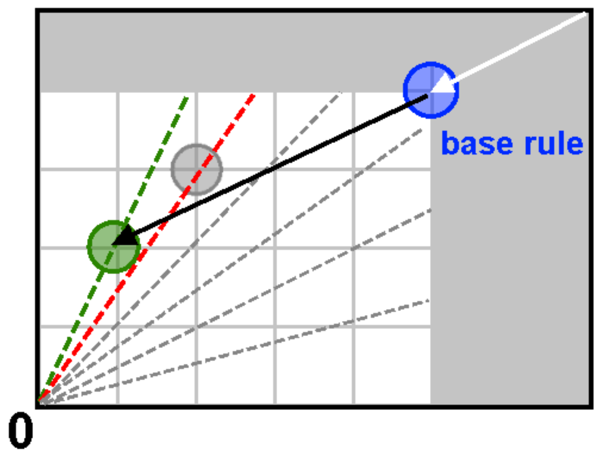
- **Modified m- Estimate:**

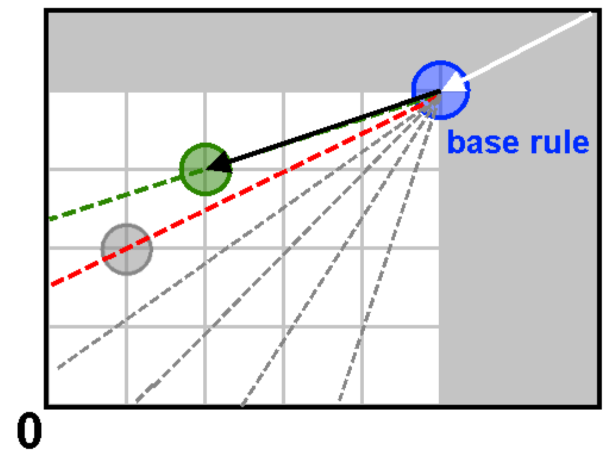
$$h'_{mest}(p,n,P,N) = \frac{N-n+m\cdot\frac{P}{P+N}}{(P+N)-(p+n-m)}$$



Separate-and-Conquer Rule Learning
Specialized Refinement Heuristics

- Example of the isometrics w.r.t. rule refinement (here: Precision) follows





$$h_{prec}(p,n) = \frac{p}{p+n}$$

$$h'_{prec}(p,n,P,N) = \frac{N-n}{(P+N)-(p+n)}$$

The two refinement heuristics choose different refinements in this example.

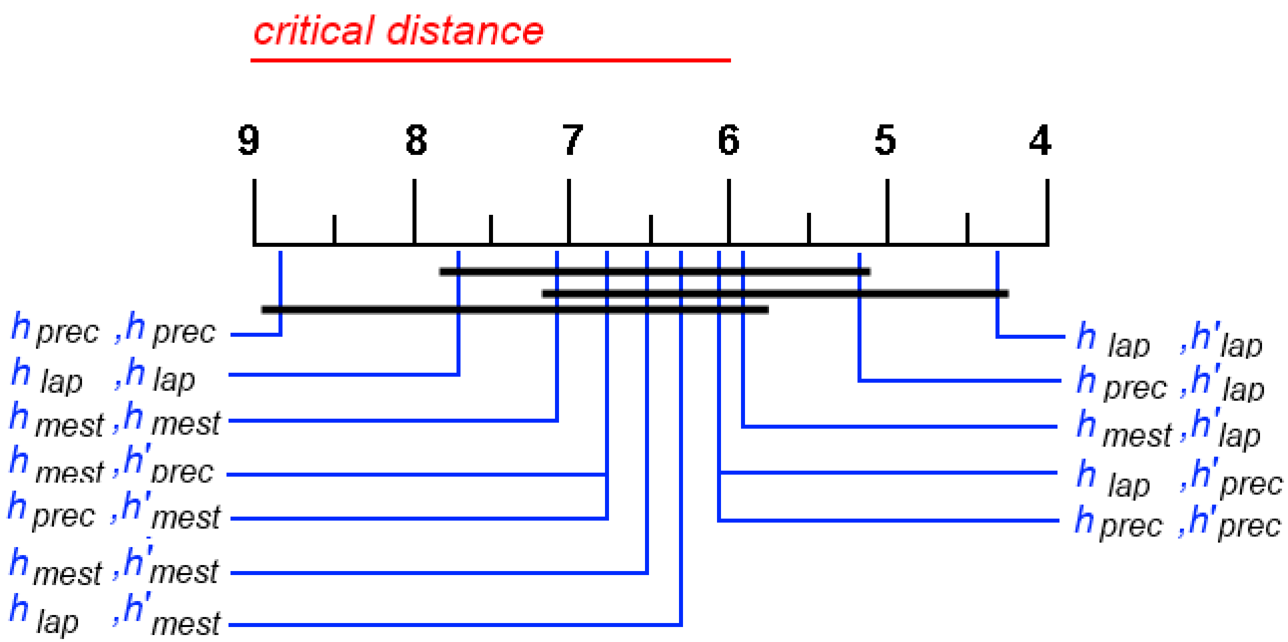
- Rule selection: no changes

Experiments
Accuracy on 19 datasets

Rule selection:		Precision	Precision	Precision		Laplace	Laplace	Laplace		M-Est.	M-Est.	M-Est.
Rule refining:	**Precision**	**Mod. Precision**	**Mod. Laplace**	**Mod. M-Est.**	**Laplace**	**Mod. Precision**	**Mod. Laplace**	**Mod. M-Est.**	**M-Est.**	**Mod. Precision**	**Mod. Laplace**	**Mod. M-Est.**
breast-cancer.arff	68,53	72,38	72,03	73,43	69,58	70,63	71,33	72,73	71,33	72,03	72,38	73,78
car.arff	90,1	90,34	90,51	88,66	90,45	91,2	91,73	91,2	89,64	90,45	90,28	87,91
contact-lenses.arff	79,17	87,5	87,5	83,33	79,17	87,5	87,5	83,33	87,5	87,5	87,5	83,33
futebol.arff	28,57	64,29	57,14	42,88	28,57	64,29	57,14	42,88	50	64,29	57,14	42,86
glass.arff	56,54	65,89	68,69	62,15	61,22	65,89	68,69	62,15	69,16	67,29	71,5	63,55
hepatitis.arff	78,07	79,36	80	76,77	78,71	79,36	80	76,74	78,07	79,36	80	76,77
hypothyroid.arff	98,23	98,61	98,74	98,83	98,39	98,61	98,74	98,83	98,8	98,61	98,74	98,83
horse-colic.arff	72,01	79,35	79,35	77,99	70,65	79,35	80,16	77,99	77,45	79,35	78,8	77,99
idh.arff	62,07	82,76	75,86	75,86	62,07	82,76	75,86	75,86	68,97	82,76	75,86	75,86
iris.arff	92,67	93,33	95,33	94,67	94	93,33	95,33	94,67	94	93,33	95,33	94,67
ionosphere.arff	95,16	82,62	83,19	89,46	94,87	82,62	93,19	89,46	91,74	82,91	83,19	91,17
labor.arff	91,23	80,7	82,46	89,47	91,23	80,7	82,46	89,47	85,97	80,7	82,46	89,47
lymphography.arff	83,78	77,7	84,46	83,11	85,14	77,7	84,46	83,11	75	76,35	81,08	83,78
mushroom.arff	100	100	100	100	100	100	100	100	100	100	100	100
monk3.arff	87,71	82,79	82,79	84,43	88,53	85,25	84,43	86,89	81,15	79,51	81,15	82,79
primary-tumor.arff	33,63	39,23	35,1	30,97	32,45	39,23	35,99	30,38	33,92	37,76	34,51	30,68
soybean.arff	90,04	91,51	92,24	91,36	90,34	91,8	92,39	90,63	91,51	90,92	90,48	91,36
tic-tac-toe.arff	97,39	98,02	97,6	97,81	97,6	98,02	97,6	97,81	98,12	98,02	97,6	97,81
vote.arff	94,94	93,56	94,25	94,48	95,4	94,25	94,25	94,94	93,33	93,56	94,71	96,09
zoo.arff	84,16	88,12	92,08	90,1	86,14	88,12	92,08	90,1	89,11	88,12	92,08	90,1
Treffer	**2**	**4**	**4**	**1**	**3**	**3**	**7**	**1**	**3**	**3**	**4**	**2**

Experiments

Accuracy on 19 datasets – Nemenyi Test



Experiments
#Rules / #Conditions for selected Algorithms

Rule selection:		m-Estimate	m-Estimate	m-Estimate
Rule refining:	m-Estimate	Mod. Precision	Mod. Laplace	Mod. m-Estimate
breast-cancer.arff	34/158	33/189	39/179	20/66
car.arff	161/846	161/833	162/834	165/845
contact-lenses.arff	3/8	3/9	3/8	4/13
futebol.arff	2/4	2/9	2/5	4/7
glass.arff	17/55	15/241	15/90	28/84
hepatitis.arff	8/30	6/60	7/46	6/24
hypothyroid.arff	10/52	11/285	9/69	15/80
horse-colic.arff	23/114	18/163	19/111	31/111
idh.arff	3 /4	2/9	2/5	2/5
iris.arff	5/15	5/28	5/17	6/15
ionosphere.arff	9/21	7/111	8/42	12/40
labor.arff	3 /4	3/22	3/12	3/5
lymphography.arff	13/46	10/97	10/49	16/49
mushroom.arff	11/13	7/44	7/35	7/29
monk3.arff	14/44	14/50	14/45	14/40
primary-tumor.arff	77/521	81/1001	79/563	74/298
soybean.arff	46/151	43/516	44/192	53/163
tic-tac-toe.arff	15/64	16/74	16/69	25/93
vote.arff	12/63	12/69	12/59	7/25
zoo.arff	11/15	6/48	6/14	12/14

Experiments
AUC on 7 datasets

Rule Selection	Precision				Laplace				M-Estimate			
Rule Refining		**mod.P**	**mod.L**	**mod.M**		**mod.P**	**mod.L**	**mod.M**		**mod.P**	**mod.L**	**mod.M**
breast-cancer	69,76	70,73	70,63	73,64	70,42	70,42	69,65	72,80	68,46	70,04	70,21	73,18
(AUC)	***0,605***	***0,617***	***0,626***	***0,639***	***0,601***	***0,617***	***0,619***	***0,634***	***0,606***	***0,611***	***0,620***	***0,635***
hepatitis	76,39	80,84	78,84	77,68	78,52	80,84	78,84	77,68	79,16	80,84	78,84	77,68
(AUC)	***0,685***	***0,670***	***0,668***	***0,639***	***0,704***	***0,670***	***0,668***	***0,639***	***0,685***	***0,670***	***0,668***	***0,639***
tic-tac-toe	97,39	98,18	98,01	97,86	97,61	98,18	98,01	97,86	98,00	98,18	97,99	97,86
(AUC)	***0,981***	***0,980***	***0,982***	***0,976***	***0,978***	***0,980***	***0,982***	***0,976***	***0,984***	***0,980***	***0,982***	***0,976***
vote	94,55	93,29	92,97	94,21	94,69	93,61	93,08	94,12	93,33	93,22	93,61	95,10
(AUC)	***0,949***	***0,937***	***0,938***	***0,948***	***0,955***	***0,941***	***0,939***	***0,947***	***0,940***	***0,934***	***0,943***	***0,955***
horse-colic	75,11	79,21	78,15	77,77	72,42	79,21	78,21	77,77	78,91	79,19	78,42	77,80
(AUC)	***0,747***	***0,782***	***0,783***	***0,796***	***0,737***	***0,782***	***0,783***	***0,796***	***0,785***	***0,783***	***0,789***	***0,797***
monk3	86,15	83,77	82,95	85,16	86,97	83,93	82,79	84,84	80,90	80,66	80,90	82,54
(AUC)	***0,886***	***0,847***	***0,850***	***0,862***	***0,893***	***0,846***	***0,848***	***0,856***	***0,793***	***0,785***	***0,795***	***0,807***
kr-vs-kp	99,07	98,78	99,01	98,79	99,14	98,72	98,98	98,84	99,49	98,78	99,01	98,81
(AUC)	***0,995***	***0,990***	***0,993***	***0,993***	***0,996***	***0,989***	***0,993***	***0,994***	***0,997***	***0,990***	***0,993***	***0,993***

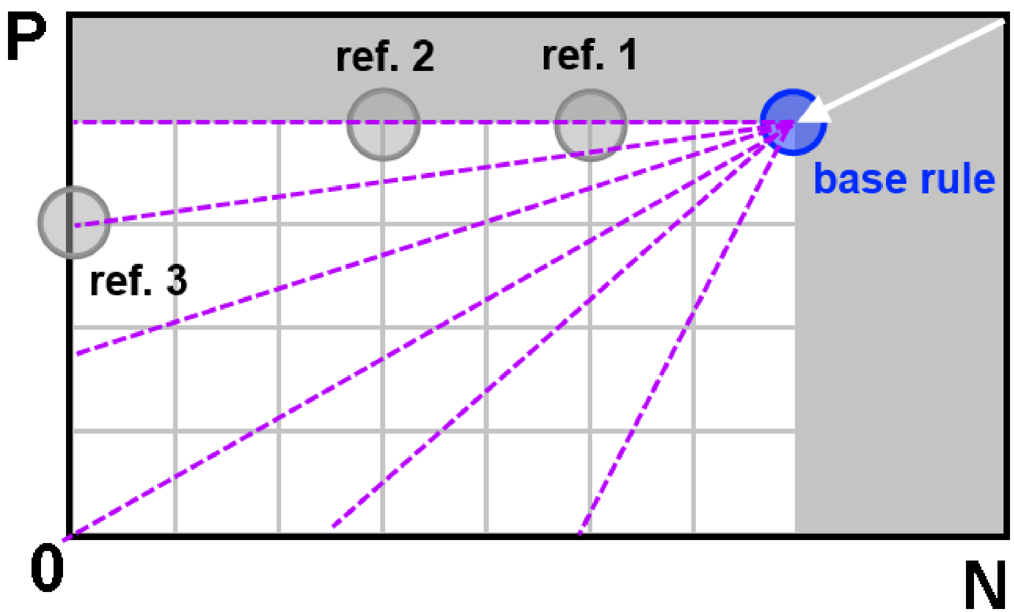
Concluding Remarks
General

- **Experiments w.r.t. the AUC suffer from certain problems**
  - Small testing folds
  - Examples always grouped
  - Small datasets
- **Experiments w.r.t. Accuracy: some notable properties (next page)**
  - Modified Laplace appears to perform better than Precision or the m-Estimate

  With the same rule selection heuristic applied

Concluding Remarks
Modified Laplace vs. Precision and m-Estimate

- Modified Precision causes very long rules (# of conditions)
- Mostly small steps in coverage space while learning rules
  - Tends to overfit on the training data set
  - Assessing refinements in a fictional example:



h(ref1) = h(ref2)
h(ref3) < h(ref1)
h(ref3) < h(ref2)

Concluding Remarks
Modified Laplace vs. Precision and m-Estimate

- **Modified m- Estimate: Parameter m ~= 22,5 [Janssen/Fürnkranz 2010]**
  - Possibly no longer optimal in this case?
- **Isometrics with m approaching infinity equal *weighted relative accuracy***
  - *WRA* tends to over-generalize [Janssen 2012]
- Possible explanation for following m-Estimate result properties:
  - Short rules
  - More rules needed to reach stopping criterion (no positive examples left)

Concluding Remarks
Modified Laplace vs. Precision and m-Estimate

- **Distance of isometrics origin from (P,N):**
  - For precision: 0
  - For laplace: sqrt(2)
  - For the m-Estimate: Depending on P/N, but >= m
    - Large for m = 22,5
- **Possible further research?**