An Emprirical Investigation of the Trade-Off Between Consistency and Coverage in Rule Learning Heuristics

Frederik Janssen
Johannes Fürnkranz

TECHNISCHE UNIVERSITÄT DARMSTADT

Outline

1. Motivation
2. Separate-and-Conquer Rule Learning
3. Rule Learning Heuristics
   - Basic Heuristics
   - Parametrized Heuristics
4. Experimental Setup
5. The Search Strategy
6. Results
   - Macro-averaged accuracy over parameter values
   - Isometrics
   - Accuracies
   - Statistical significance
7. Discussion

1. Motivation

- Open questions in Rule Learning:
  - selection of an appropriate heuristic
  - how to adjust the parameter of parametrized heuristics
    - trade-off between Consistency and Coverage
    - so far this trade-off is often fixed
- no visualization of the parametrized heuristics
- no exhaustive study about the behavior of many different heuristics on many different datasets

2. Separate-and-Conquer Rule Learning

In the experiments we used a simple SeCo Rule Learner with the following properties:

- allows the usage of different heuristics and uses a Top-Down Hill Climbing Search
- employs ordered class binarization
- classification is done by a decision list of rules
- does not perform pruning
- but performs implicit pruning when selecting the best rule along a refinement process
- this work focuses on heuristics not on sophisticated pruning methods

3. Rule Learning Heuristics

- a heuristic is a function of the form $h(p, n, P, N)$
- usually a good heuristic should optimize two criteria:
  - **Coverage:** the number of positive examples that are covered by the rule (p) should be maximized and
  - **Consistency:** the number of negative examples that are covered (n) should be minimized
- heuristics could be visualized in Coverage Spaces (un-normalized ROC spaces) following J. Fürnkranz and P. Flach (2005)

3. Rule Learning Heuristics

Basic Heuristics

	heuristic	formula
Consistency	Precision	$h_{Precision} = \frac{p}{p+n}$
	MinNeg	$h_{MinNeg} = -n$
	Rel. MinNeg	$h_{relMinNeg} = -\frac{n}{N}$
Coverage	Full Coverage	$h_{Coverage} = \frac{p+n}{P+N}$
	Weighted Relative Accuracy	$h_{WRA} = \frac{p}{P} - \frac{n}{N}$
	Recall (Rel. MaxPos)	$h_{Recall} = \frac{p}{P}$
	MaxPos	$h_{MaxPos} = p$

Parametrized Heuristics

heuristic	formula
Cost Measure	$h_{cost} = c \cdot p - (1 - c) \cdot n$
Relative Cost Measure	$h_{rcost} = c_r \cdot \frac{p}{P} - (1 - c_r) \cdot \frac{n}{N}$
F-Measure	$h_{F-Measure} = \frac{(\beta^2+1) \cdot h_{Precision} \cdot h_{Recall}}{\beta^2 \cdot h_{Precision} + h_{Recall}}$
m-Estimate	$h_{m-Estimate} = \frac{p + m \cdot \frac{P}{P+N}}{p+n+m}$
Klösgen	$h_{Kloesgen} = (h_{Coverage})^{\omega} \cdot \left(h_{Precision} - \frac{P}{P+N}\right)$

4. Experimental Setup

TECHNISCHE UNIVERSITÄT DARMSTADT

- 27 tuning datasets and 30 validation datasets (all from the UCI Repository)
- datasets are selected to cover a broad spectrum of different domains (i.e., different ratios of nominal to numeric attributed, different number of instances/classes)
- macro/micro-average accuracy of 10-fold stratified CV on m datasets
  - macro: $\frac{1}{m}\sum_{i=1}^{m}\frac{p_i+(N_i-n_i)}{P_i+N_i}$
  - micro: $\frac{\sum_{i=1}^{m}(p_i+N_i-n_i)}{\sum_{i=1}^{m}(P_i+N_i)}$
- averaged ranking of the heuristics on all datasets
- to test for significance we used a Friedman test along with a Nemenyi test as suggested by J. Demsar (2006)

5. The Search Strategy

- expectation: an inverse convex U-shape curve (x-axis: parameter, y-axis: macro-averaged accuracy)
- idea: binary search
  - record the accuracy of 10 (intuitively) parametrizations on all tuning sets
  - pick the parameter with highest accuracy
  - narrow down the bounds/the increment ($lowerbound \leftarrow p_{best} - \frac{i}{2}$, $upperbound \leftarrow p_{best} + \frac{i}{2}$ and $i \leftarrow \frac{i}{10}$) and
  - record the accuracies again
- greedy search algorithm for narrowing down the region of interest
- algorithm stores 3 candidate parameters (to avoid local optima)

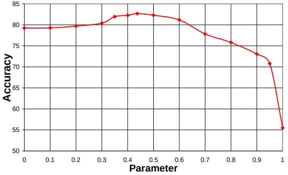
Run	set which has to be searched	increment	best parameter	Accuracy
1	{0.1, ..., 1.0}	0.1	0.4	84.5658
2	{0.35, ..., 0.45}	0.01	0.42	84.6852
3	{0.415, ..., 0.425}	0.001	0.418	84.7015
4	{0.4175, ..., 0.4185}	0.0001	0.4176	84.7045
5	{0.41755, ..., 0.41765}	0.00001	0.4176	84.7045

6. Results

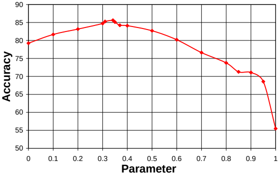
Macro-averaged accuracy over parameter values

TECHNISCHE UNIVERSITÄT DARMSTADT

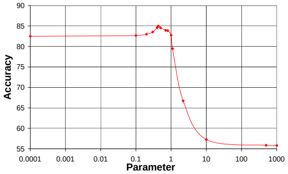
Cost Measure



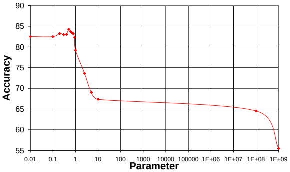
Rel. Cost Measure



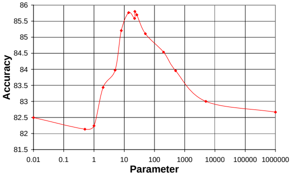
Klösgen Measure



F-Measure

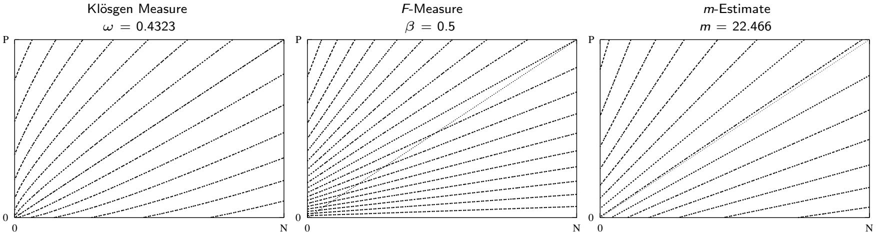


m-Estimate



6. Results

Isometrics



- Cost Measures are just parallel lines with a slope corresponding to the best setting
  - best parameter of Cost Measure: $c = 0.437$
  - best parameter of Relative Cost Measure: $c_r = 0.342$

6. Results

Accuracies

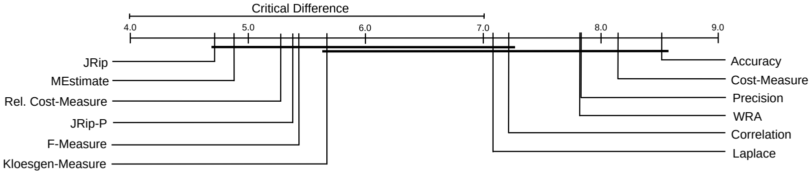
Heuristic	average accuracy		average	
	Macro	Micro	Rank	Size
$m = 22.466$	85.87	93.87 (1)	4.54 (1)	36.85 (4)
$c_r = 0.342$	85.61	92.50 (6)	5.54 (4)	26.11 (3)
$\omega = 0.4323$	84.82	93.62 (3)	5.28 (3)	48.26 (8)
JRip	84.45	93.80 (2)	5.12 (2)	16.93 (2)
$\beta = 0.5$	84.14	92.94 (5)	5.72 (5)	41.78 (6)
JRip-P	83.88	93.55 (4)	6.28 (6)	45.52 (7)
Correlation	83.68	92.39 (7)	7.17 (7)	37.48 (5)
WRA	82.87	90.43 (12)	7.80 (10)	14.22 (1)
$c = 0.437$	82.60	91.09 (11)	7.30 (8)	106.30 (12)
Precision	82.36	92.21 (9)	7.80 (10)	101.63 (11)
Laplace	82.28	92.26 (8)	7.31 (9)	91.81 (10)
Accuracy	82.24	91.31 (10)	8.11 (12)	85.93 (9)

Heuristic	average accuracy		average	
	Macro	Micro	Rank	Size
JRip	78.98	82.42 (1)	4.72 (1)	12.20 (2)
$c_r = 0.342$	78.87	81.80 (3)	5.28 (3)	25.30 (3)
$m = 22.466$	78.67	81.72 (4)	4.88 (2)	46.33 (4)
JRip-P	78.50	82.04 (2)	5.38 (4)	49.80 (6)
$\omega = 0.4323$	78.46	81.33 (6)	5.67 (6)	61.83 (8)
$\beta = 0.5$	78.12	81.52 (5)	5.43 (5)	51.57 (7)
Correlation	77.55	80.91 (7)	7.23 (8)	47.33 (5)
Laplace	76.87	79.76 (8)	7.08 (7)	117.00 (10)
Precision	76.22	79.53 (9)	7.83 (10)	128.37 (12)
$c = 0.437$	76.11	78.93 (11)	8.15 (11)	122.87 (11)
WRA	75.82	79.35 (10)	7.82 (9)	12.00 (1)
Accuracy	75.65	78.47 (12)	8.52 (12)	99.13 (9)

- m-Estimate performs best on the tuning sets (85.87%)
- JRip was the best algorithm on the validation sets (78.98%)
- Ranking has not changed much
- some evidence for robustness of the best performing parameters

6. Results

Statistical significance (validation sets)



- for $p = 0.05$ the null hypothesis of the Friedmann Test was rejected
- only the Klösgen Measure is not significantly better than the Accuracy heuristic
- noticable gap between the tuned and the basic heuristics

7. Discussion

TECHNISCHE UNIVERSITÄT DARMSTADT

- we have determined suitable parameter settings for 5 parametrized heuristics
- taking the class distribution into account is mandatory
- rating the true positive rate more heavily than the false positive rate yields good overall performance among all parametrized heuristics
- isometrics of the best settings showed strong similarities
- this work yields a very exhaustive experimental comparison of different heuristics

References

- J. Fürnkranz and P. Flach (2005): J. Fürnkranz and P. Flach. ROC'n'Rule Learning - Towards a Better Understanding of Covering Algorithms. *Machine Learning*, 58(1):39-77, January 2005. ISSN 0885-6125.
- J. Demsar (2006): J. Demsar. Statistical comparisons of classifiers over multiple datasets. *Machine Learning Research*, (7):1-30, 2006.