Optimizing the AUC with Rule Learning
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 $R: \text{<class label>} := \{\text{cond}_1, \text{cond}_2, \ldots, \text{cond}_n\}$
  - 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

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
  - \( h: \text{Rule} \rightarrow [0,1] \)
  - Rules provide statistics in the form of a confusion matrix

<table>
<thead>
<tr>
<th></th>
<th>Classified positive</th>
<th>Classified negative</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>true positives</strong></td>
<td></td>
<td>false negatives</td>
<td>P</td>
</tr>
<tr>
<td><strong>false positives</strong></td>
<td></td>
<td>true negatives</td>
<td>N</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>P+N</td>
</tr>
</tbody>
</table>
Separate-and-Conquer Rule Learning
Coverage Spaces and ROC Space

- 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 heuristic
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

- 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 heuristic
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
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)
General Procedure

- Start with the universal rule $\text{<majority class>} := \{\}$ 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

Optimization Approach
Modification of the Basic Algorithm
Separate-and-Conquer Rule Learning
Specialized Refinement Heuristics

- 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:

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

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Rule refining:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Precision</td>
<td>Mod. Precision</td>
<td>Mod. Laplace</td>
<td>Mod. M-Est.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>breast-cancer.arff</td>
<td>68.53</td>
<td>72.38</td>
<td>72.03</td>
<td>73.43</td>
<td>69.58</td>
<td>70.83</td>
<td>71.33</td>
<td>72.73</td>
<td>71.33</td>
<td>72.03</td>
<td>72.38</td>
</tr>
<tr>
<td>car.arff</td>
<td>90.1</td>
<td>90.34</td>
<td>90.51</td>
<td>88.66</td>
<td>90.45</td>
<td>91.2</td>
<td>91.73</td>
<td>91.2</td>
<td>89.64</td>
<td>90.45</td>
<td>90.28</td>
</tr>
<tr>
<td>contact-lenses.arff</td>
<td>79.17</td>
<td>87.5</td>
<td>87.5</td>
<td>83.33</td>
<td>79.17</td>
<td>87.5</td>
<td>87.5</td>
<td>83.33</td>
<td>87.5</td>
<td>87.5</td>
<td>87.5</td>
</tr>
<tr>
<td>futebol.arff</td>
<td>28.57</td>
<td>64.29</td>
<td>57.14</td>
<td>42.88</td>
<td>28.57</td>
<td>64.29</td>
<td>57.14</td>
<td>42.88</td>
<td>50</td>
<td>64.29</td>
<td>57.14</td>
</tr>
<tr>
<td>glass.arff</td>
<td>56.54</td>
<td>65.89</td>
<td>68.69</td>
<td>62.15</td>
<td>61.22</td>
<td>65.89</td>
<td>68.69</td>
<td>62.15</td>
<td>69.16</td>
<td>67.29</td>
<td>71.5</td>
</tr>
<tr>
<td>hepatitis.arff</td>
<td>78.07</td>
<td>79.36</td>
<td>80</td>
<td>76.77</td>
<td>78.71</td>
<td>79.36</td>
<td>80</td>
<td>76.74</td>
<td>78.07</td>
<td>79.36</td>
<td>80</td>
</tr>
<tr>
<td>hypothyroid.arff</td>
<td>98.23</td>
<td>98.61</td>
<td>98.74</td>
<td>98.83</td>
<td>98.39</td>
<td>98.61</td>
<td>98.74</td>
<td>98.83</td>
<td>98.8</td>
<td>98.61</td>
<td>98.74</td>
</tr>
<tr>
<td>horse-colic.arff</td>
<td>72.01</td>
<td>79.35</td>
<td>79.35</td>
<td>77.99</td>
<td>70.65</td>
<td>79.35</td>
<td>80.16</td>
<td>77.99</td>
<td>77.45</td>
<td>79.35</td>
<td>78.8</td>
</tr>
<tr>
<td>idh.arff</td>
<td>62.07</td>
<td>82.76</td>
<td>75.86</td>
<td>75.86</td>
<td>62.07</td>
<td>82.76</td>
<td>75.86</td>
<td>75.86</td>
<td>68.97</td>
<td>82.76</td>
<td>75.86</td>
</tr>
<tr>
<td>iris.arff</td>
<td>92.67</td>
<td>93.33</td>
<td>95.33</td>
<td>94.67</td>
<td>94</td>
<td>93.33</td>
<td>95.33</td>
<td>94.67</td>
<td>94</td>
<td>93.33</td>
<td>95.33</td>
</tr>
<tr>
<td>ionosphere.arff</td>
<td>95.16</td>
<td>82.62</td>
<td>83.19</td>
<td>89.46</td>
<td>94.87</td>
<td>82.62</td>
<td>93.19</td>
<td>89.46</td>
<td>91.74</td>
<td>82.91</td>
<td>83.19</td>
</tr>
<tr>
<td>labor.arff</td>
<td>91.23</td>
<td>80.7</td>
<td>82.46</td>
<td>89.47</td>
<td>91.23</td>
<td>80.7</td>
<td>82.46</td>
<td>89.47</td>
<td>85.97</td>
<td>80.7</td>
<td>82.46</td>
</tr>
<tr>
<td>lymphography.arff</td>
<td>83.78</td>
<td>77.7</td>
<td>84.46</td>
<td>83.11</td>
<td>85.14</td>
<td>77.7</td>
<td>84.46</td>
<td>83.11</td>
<td>75</td>
<td>76.35</td>
<td>81.08</td>
</tr>
<tr>
<td>mushroom.arff</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>monk3.arff</td>
<td>87.71</td>
<td>82.79</td>
<td>82.79</td>
<td>84.43</td>
<td>88.53</td>
<td>85.25</td>
<td>84.43</td>
<td>86.89</td>
<td>81.15</td>
<td>79.51</td>
<td>81.15</td>
</tr>
<tr>
<td>primary-tumor.arff</td>
<td>33.63</td>
<td>39.23</td>
<td>35.1</td>
<td>30.97</td>
<td>32.45</td>
<td>39.23</td>
<td>35.99</td>
<td>30.38</td>
<td>33.92</td>
<td>37.76</td>
<td>34.51</td>
</tr>
<tr>
<td>soybean.arff</td>
<td>90.04</td>
<td>91.51</td>
<td>92.24</td>
<td>91.36</td>
<td>90.34</td>
<td>91.8</td>
<td>92.39</td>
<td>90.63</td>
<td>91.51</td>
<td>90.92</td>
<td>90.48</td>
</tr>
<tr>
<td>tic-tac-toe.arff</td>
<td>97.39</td>
<td>98.02</td>
<td>97.6</td>
<td>97.81</td>
<td>97.6</td>
<td>98.02</td>
<td>97.6</td>
<td>97.81</td>
<td>98.12</td>
<td>98.02</td>
<td>97.6</td>
</tr>
<tr>
<td>vote.arff</td>
<td>94.94</td>
<td>93.56</td>
<td>94.25</td>
<td>94.48</td>
<td>95.4</td>
<td>94.25</td>
<td>94.25</td>
<td>94.94</td>
<td>93.33</td>
<td>93.56</td>
<td>94.71</td>
</tr>
<tr>
<td>zoo.arff</td>
<td>84.16</td>
<td>88.12</td>
<td>92.08</td>
<td>90.1</td>
<td>86.14</td>
<td>88.12</td>
<td>92.08</td>
<td>90.1</td>
<td>89.11</td>
<td>88.12</td>
<td>92.08</td>
</tr>
</tbody>
</table>

Treffen | 2 | 4 | 4 | 1 | 3 | 3 | 7 | 1 | 3 | 3 | 4 | 2 |
Experiments

Accuracy on 19 datasets – Nemenyi Test

*critical distance*
## Experiments

### #Rules / #Conditions for selected Algorithms

<table>
<thead>
<tr>
<th>Rule selection:</th>
<th>m-Estimate</th>
<th>m-Estimate</th>
<th>m-Estimate</th>
<th>m-Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Mod. Precision</td>
<td>Mod. Laplace</td>
<td>Mod. m-Estimate</td>
</tr>
<tr>
<td>breast-cancer.arff</td>
<td>34/158</td>
<td>33/189</td>
<td>39/179</td>
<td>20/66</td>
</tr>
<tr>
<td>car.arff</td>
<td>161/846</td>
<td>161/833</td>
<td>162/834</td>
<td>165/845</td>
</tr>
<tr>
<td>futebol.arff</td>
<td>2/4</td>
<td>2/9</td>
<td>2/5</td>
<td>4/7</td>
</tr>
<tr>
<td>glass.arff</td>
<td>17/55</td>
<td>15/241</td>
<td>15/90</td>
<td>28/84</td>
</tr>
<tr>
<td>hepatitis.arff</td>
<td>8/30</td>
<td>6/60</td>
<td>7/46</td>
<td>6/24</td>
</tr>
<tr>
<td>hypothyroid.arff</td>
<td>10/82</td>
<td>11/285</td>
<td>9/69</td>
<td>15/80</td>
</tr>
<tr>
<td>horse-colc.arff</td>
<td>23/114</td>
<td>18/163</td>
<td>19/111</td>
<td>31/111</td>
</tr>
<tr>
<td>idh.arff</td>
<td>3/4</td>
<td>2/9</td>
<td>2/5</td>
<td>2/5</td>
</tr>
<tr>
<td>iris.arff</td>
<td>5/15</td>
<td>5/28</td>
<td>5/17</td>
<td>6/15</td>
</tr>
<tr>
<td>ionosphere.arff</td>
<td>9/21</td>
<td>7/111</td>
<td>8/42</td>
<td>12/40</td>
</tr>
<tr>
<td>labor.arff</td>
<td>3/4</td>
<td>3/22</td>
<td>3/12</td>
<td>3/5</td>
</tr>
<tr>
<td>lymphography.arff</td>
<td>13/46</td>
<td>10/97</td>
<td>10/49</td>
<td>16/49</td>
</tr>
<tr>
<td>mushroom.arff</td>
<td>11/13</td>
<td>7/44</td>
<td>7/35</td>
<td>7/29</td>
</tr>
<tr>
<td>monk3.arff</td>
<td>14/44</td>
<td>14/50</td>
<td>14/45</td>
<td>14/40</td>
</tr>
<tr>
<td>primary-tumor.arff</td>
<td>77/521</td>
<td>81/1001</td>
<td>79/563</td>
<td>74/298</td>
</tr>
<tr>
<td>soybean.arff</td>
<td>46/151</td>
<td>43/516</td>
<td>44/192</td>
<td>53/163</td>
</tr>
<tr>
<td>tic-tac-toe.arff</td>
<td>15/64</td>
<td>16/74</td>
<td>16/69</td>
<td>25/93</td>
</tr>
<tr>
<td>vote.arff</td>
<td>12/63</td>
<td>12/69</td>
<td>12/59</td>
<td>7/25</td>
</tr>
<tr>
<td>zoo.arff</td>
<td>11/15</td>
<td>6/48</td>
<td>6/14</td>
<td>12/14</td>
</tr>
</tbody>
</table>
## Experiments

**AUC on 7 datasets**

<table>
<thead>
<tr>
<th>Rule Selection</th>
<th>Precision</th>
<th>Laplace</th>
<th>M-Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>breast-cancer (AUC)</td>
<td>0.605</td>
<td>0.617</td>
<td>0.626</td>
</tr>
<tr>
<td>hepatitis (AUC)</td>
<td>0.685</td>
<td>0.670</td>
<td>0.668</td>
</tr>
<tr>
<td>tic-tac-toe (AUC)</td>
<td>0.981</td>
<td>0.980</td>
<td>0.982</td>
</tr>
<tr>
<td>vote (AUC)</td>
<td>0.949</td>
<td>0.937</td>
<td>0.938</td>
</tr>
<tr>
<td>horse-colic (AUC)</td>
<td>0.747</td>
<td>0.782</td>
<td>0.783</td>
</tr>
<tr>
<td>monk3 (AUC)</td>
<td>0.886</td>
<td>0.947</td>
<td>0.850</td>
</tr>
<tr>
<td>kr-vs-kp (AUC)</td>
<td>0.995</td>
<td>0.990</td>
<td>0.993</td>
</tr>
</tbody>
</table>
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:

\[
\begin{align*}
h(\text{ref}1) & = h(\text{ref}2) \\
h(\text{ref}3) & < h(\text{ref}1) \\
h(\text{ref}3) & < h(\text{ref}2)
\end{align*}
\]

![Diagram showing the relationship between different references and base rule.](image)
Concluding Remarks
Modified Laplace vs. Precision and m-Estimate

- **Modified m-Estimate**: Parameter $m \approx 22.5$ [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 \( \geq m \)
    - Large for \( m = 22.5 \)

- Possible further research?