Learning Semantically Coherent Rules

Presentation of the Bachelor thesis of Alexander Gabriel
Overview

- Motivation
- Idea
- Implementation
- Experiments & Results
- Conclusion & Ideas for Further Research
Rule Learning
(a very inaccurate reminder)

- Given: A set of attributes and example variable realizations
- Goal: A rule that assigns the right examples to the target class
- Iterative process
  - Adding conditions one by one to the rule
  - Using heuristics to decide which conditions to add
- Removing covered examples
- Continue on reduced example set
Interpretability of Rules

- Rules should 'make sense'
- Attribute labels should be semantically related

- Rule Learning heuristics disregard attribute labels
- Rule Learning algorithms have no bias towards semantically related rules

- Few semantic relations between the attribute labels of a rule
Semantic Coherence

- An approximation of semantic relatedness
- Two concepts can be semantically similar
- Three and more can be semantically coherent

Semantically coherent rules should have attributes that are more related than semantically incoherent rules
Combining Heuristics

- 1 classic rule learning heuristic
- 1 semantic rule learning heuristic
- Combined using a weighted sum
  - The weight of the semantic part is called influence in the following

- Should have a bias towards semantically coherent rules
- Should increase semantic quality
- Should decrease modeling quality
WordNet

- Model of the structure of the English language
- Consists of
  - Synsets (sets of synonyms)
  - Semantic relations between synsets
    - e.g. part-whole, class-subclass, ...
- Free to use
- Online and offline versions available
The LIN metric & Information Content (IC)

- Distance metric on WN
- Works with nouns and verbs

\[ IC(c) = -\log(p(c)) \]

\[
\text{lin}(\text{syn1}, \text{syn2}) = 2 \times \frac{IC(lcs)}{IC(\text{synset1}) + IC(\text{synset2})}
\]
A Semantic Heuristic

- Compares pairs of concepts based on WordNet distance (LIN metric)
- One similarity score for each combination of attribute labels in a rule
- Optional tokenization of attribute labels
- Different similarity scores are combined to a single coherence score using a statistical method
Semantic Heuristic

1. Split rule into words and get synsets for each word
2. Compare synset pairs using LIN metric
3. Choose the maximum similarity value of each synset combination for each pair of words
4. Calculate the mean of the word pair similarity scores for each pair of conditions
5. Calculate the statistic value for the set of condition pair similarity scores
6. Return statistic value
Different Statistics

- **Minimum**
  - Returns the lowest similarity score
  - Discourages adding conditions that decrease the minimum

- **Mean**
  - Returns the mean of the similarity scores
  - Encourages adding conditions that increase the mean
  - Discourages adding conditions that decrease the mean

- **Maximum**
  - Returns the highest similarity score
  - Encourages adding conditions that increase the maximum
The SeCo-Framework

- Experimentation framework
- Modular
- Modify all the parts of the rule learning process
- Comes with reference implementations
- Features tools for evaluation
- Comprehensive summary of experiment results
Datasets

31 unmodified

- 100-1000 samples, 4-69 attributes
  0-100% labels found in WordNet

• 1 modified dataset

- 15 custom named attributes from 3 domains
  including compound attribute labels
Semantic and Modeling Quality over all Datasets using 10% Semantic Influence

<table>
<thead>
<tr>
<th>Statistic</th>
<th>No semantic heuristic</th>
<th>Minimum</th>
<th>Mean</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>$m$-Estimate</td>
<td>11.827%</td>
<td>16.128%</td>
<td><strong>16.599%</strong></td>
<td>16.387%</td>
</tr>
<tr>
<td>Laplace Estimate</td>
<td>11.011%</td>
<td>15.006%</td>
<td>13.333%</td>
<td><strong>15.095%</strong></td>
</tr>
<tr>
<td>Accuracy</td>
<td>11.980%</td>
<td>17.845%</td>
<td><strong>18.107%</strong></td>
<td>16.481%</td>
</tr>
<tr>
<td>Overall</td>
<td>11.606%</td>
<td><strong>16.326%</strong></td>
<td>16.013%</td>
<td>15.988%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Statistic</th>
<th>No semantic heuristic</th>
<th>Minimum</th>
<th>Mean</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>$m$-Estimate</td>
<td>76.728%</td>
<td>76.671%</td>
<td>76.163%</td>
<td>76.540%</td>
</tr>
<tr>
<td>Laplace Estimate</td>
<td>75.064%</td>
<td>74.722%</td>
<td>74.877%</td>
<td>74.691%</td>
</tr>
<tr>
<td>Accuracy</td>
<td>74.067%</td>
<td>73.480%</td>
<td>74.248%</td>
<td>73.771%</td>
</tr>
<tr>
<td>Overall</td>
<td>75.286%</td>
<td>74.958%</td>
<td>75.096%</td>
<td>75.001%</td>
</tr>
</tbody>
</table>
Semantic Quality on the Modified Dataset using 10% Semantic and 90% m-Estimate with and without Tokenization

<table>
<thead>
<tr>
<th>Configuration</th>
<th>Coherence Score</th>
<th>Average rule length</th>
<th>Number of rules</th>
</tr>
</thead>
<tbody>
<tr>
<td>Without Semantic Heuristic</td>
<td>25.3%</td>
<td>3.60</td>
<td>5</td>
</tr>
<tr>
<td>Using the Minimum Statistic</td>
<td>34.2%</td>
<td>3.50</td>
<td>6</td>
</tr>
<tr>
<td>Using the Mean Statistic</td>
<td>45.0%</td>
<td>3.50</td>
<td>6</td>
</tr>
<tr>
<td>Using the Maximum Statistic</td>
<td>32.9%</td>
<td>4.64</td>
<td>11</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Configuration</th>
<th>Coherence Score</th>
<th>Average rule length</th>
<th>Number of rules</th>
</tr>
</thead>
<tbody>
<tr>
<td>Without Semantic Heuristic</td>
<td>25.3%</td>
<td>3.60</td>
<td>5</td>
</tr>
<tr>
<td>Using the Minimum Statistic</td>
<td>34.2%</td>
<td>3.50</td>
<td>6</td>
</tr>
<tr>
<td>Using the Mean Statistic</td>
<td>46.7%</td>
<td>3.17</td>
<td>6</td>
</tr>
<tr>
<td>Using the Maximum Statistic</td>
<td>32.9%</td>
<td>4.64</td>
<td>11</td>
</tr>
</tbody>
</table>
Ruleset of the Modified Dataset using 10% Semantic and 90% m-Estimate

without semantic heuristic
Class 1: bush in, newspaper in
Class 2: flower in, blue train
Class 3: tree in, red ship

statistic: min | tokenization: off
Class 1: bush in, flower in
Class 2: blue train
Class 3: tree in, red ship

statistic: mean | tokenization: off
Class 1: bush in, flower in
Class 2: blue train
Class 3: tree in, red ship

statistic: max | tokenization: off
Class 1: bush in, flower in
Class 2: blue train
Class 3: tree in, red ship

statistic: min | tokenization: on
Class 1: bush in, flower in
Class 2: blue train
Class 3: tree in, red ship

statistic: mean | tokenization: on
Class 1: bush in, flower in
Class 2: blue train
Class 3: tree in, red ship

statistic: max | tokenization: on
Class 1: bush in, flower in
Class 2: blue train
Class 3: tree in, red ship

Look at your handout →
Decrease in Modeling Quality with Increasing Semantic Influence on Datasets with 30-60% Attribute Labels Found in WN

<table>
<thead>
<tr>
<th>Heuristic</th>
<th>0</th>
<th>10</th>
<th>20</th>
<th>30</th>
<th>40</th>
<th>50</th>
<th>60</th>
<th>70</th>
<th>80</th>
<th>90</th>
<th>100</th>
</tr>
</thead>
<tbody>
<tr>
<td>m-Estimate</td>
<td>87.65</td>
<td>82.57</td>
<td>81.74</td>
<td>82.21</td>
<td>81.56</td>
<td>82.26</td>
<td>82.08</td>
<td>82.13</td>
<td>82.08</td>
<td>81.70</td>
<td>46.68</td>
</tr>
<tr>
<td>Accuracy</td>
<td>90.69</td>
<td>83.24</td>
<td>82.70</td>
<td>83.08</td>
<td>83.93</td>
<td>83.93</td>
<td>83.87</td>
<td>83.65</td>
<td>83.60</td>
<td>82.94</td>
<td>44.64</td>
</tr>
<tr>
<td>Laplace Est.</td>
<td>94.17</td>
<td>88.31</td>
<td>88.46</td>
<td>90.95</td>
<td>90.85</td>
<td>89.92</td>
<td>88.98</td>
<td>85.81</td>
<td>85.12</td>
<td>84.78</td>
<td>44.64</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Heuristic</th>
<th>0</th>
<th>10</th>
<th>20</th>
<th>30</th>
<th>40</th>
<th>50</th>
<th>60</th>
<th>70</th>
<th>80</th>
<th>90</th>
<th>100</th>
</tr>
</thead>
<tbody>
<tr>
<td>m-Estimate</td>
<td>66.73</td>
<td>67.20</td>
<td>67.48</td>
<td>67.94</td>
<td>67.86</td>
<td>67.70</td>
<td>67.69</td>
<td>67.44</td>
<td>67.39</td>
<td>67.35</td>
<td>46.41</td>
</tr>
<tr>
<td>Accuracy</td>
<td>64.89</td>
<td>65.93</td>
<td>65.71</td>
<td>65.93</td>
<td>65.98</td>
<td>66.03</td>
<td>65.77</td>
<td>65.99</td>
<td>65.99</td>
<td>65.72</td>
<td>45.65</td>
</tr>
<tr>
<td>Laplace Est.</td>
<td>66.89</td>
<td>67.03</td>
<td>66.70</td>
<td>66.66</td>
<td>66.61</td>
<td>66.56</td>
<td>67.11</td>
<td>67.50</td>
<td>67.35</td>
<td>67.49</td>
<td>45.78</td>
</tr>
</tbody>
</table>
### Increase in Semantic Quality with Increasing Semantic Influence on Datasets with 30-60% Attribute Labels Found in WN

<table>
<thead>
<tr>
<th></th>
<th>0%</th>
<th>10%</th>
<th>20%</th>
<th>30%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average Semantic Coherence</td>
<td>6.111%</td>
<td>10.493%</td>
<td>12.806%</td>
<td>14.346%</td>
</tr>
<tr>
<td>Average Rule length</td>
<td>3.34</td>
<td>2.20</td>
<td>3.13</td>
<td>3.03</td>
</tr>
<tr>
<td>Number of Rules</td>
<td>17.0</td>
<td>15.0</td>
<td>14.0</td>
<td>15.0</td>
</tr>
</tbody>
</table>
General Conclusions

- Use of the semantic heuristic generally increases semantic coherence
- Use of the semantic heuristic often leads to shorter rules
- Even a small amount of semantic influence can improve the semantic quality noticeably
- Large amounts of semantic influence do not generally result in drastic loss of modeling performance
Conclusions about Statistics

- The mean statistic has a more continuous and balanced influence
- The minimum statistic discourages the addition of conditions that create a new minimum similar condition pair
- The maximum statistic encourages the addition of conditions that create a new maximum similar condition pair
General Conclusions

• The semantic heuristic should fit to the domain of the attribute labels
• Attributes should be labeled with semantically expressive titles

Otherwise the influence of the semantic heuristic is both weaker and less equally spread
Ideas for Future Research

- Other semantic heuristics
  - e.g. heuristics fitting the domain of the attribute labels
- Other WordNet distance metrics
  - e.g. metrics that incorporate other semantic relations
- Other classical heuristics
  - e.g. heuristics that use pruning
- Rule quality evaluation by humans
Thank you for your attention
References


