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)

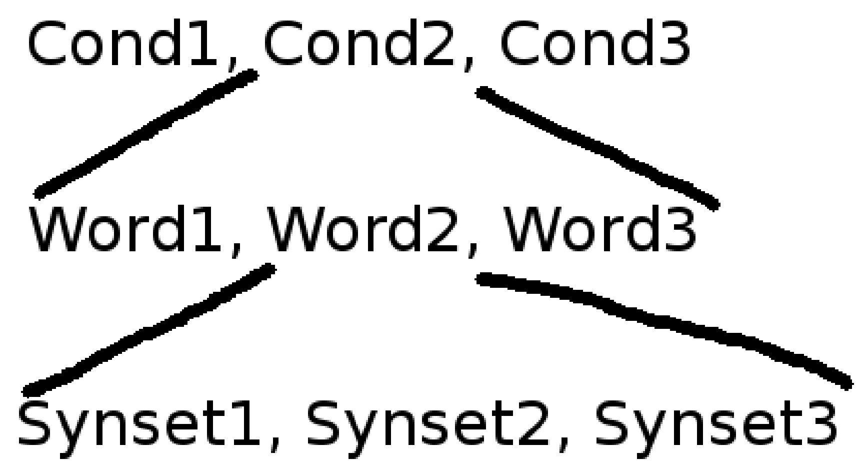
- Distance metric on WN
- Works with nouns and verbs

$$IC(c) = -\log(p(c))$$

$$lin(syn1, syn2) = 2 * \frac{IC(lcs)}{IC(synset1) + IC(synset2)}$$

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

Statistic	No semantic heuristic	Minimum	Mean	Maximum
m-Estimate	11.827%	16.128%	**16.599%**	16.387%
Laplace Estimate	11.011%	15.006%	13.333%	**15.095%**
Accuracy	11.980%	17.845%	**18.107%**	16.481%
Overall	11.606%	**16.326%**	16.013%	15.988%

Statistic	No semantic heuristic	Minimum	Mean	Maximum
m-Estimate	76.728%	76.671%	76.163%	76.540%
Laplace Estimate	75.064%	74.722%	74.877%	74.691%
Accuracy	74.067%	73.480%	74.248%	73.771%
Overall	75.286%	74.958%	75.096%	75.001%

Semantic Quality on the Modified Dataset using 10% Semantic and 90% m-Estimate with and without Tokenization

Configuration	Coherence Score	Average rule length	Number of rules
Without Semantic Heuristic	25.3%	3.60	5
Using the Minimum Statistic	34.2%	3.50	6
Using the Mean Statistic	45.0%	3.50	6
Using the Maximum Statistic	32.9%	4.64	11

Configuration	Coherence Score	Average rule length	Number of rules
Without Semantic Heuristic	25.3%	3.60	5
Using the Minimum Statistic	34.2%	3.50	6
Using the Mean Statistic	46.7%	3.17	6
Using the Maximum Statistic	32.9%	4.64	11

Ruleset of the Modified Dataset using 10% Semantic and 90% m-Estimate

Look at your handout →

without semantic heuristic

```
Class =r :- bush =n, newspaper =n, radio =n, red_ship =n, tree =n.  [89|0] Val: 0.876
Class =r :- bush =n, blue_train =y, television =n.  [39|8] Val: 0.635
Class =r :- flower =y, newspaper =n, blue_train =y.  [15|1] Val: 0.468
Class =r :- bush =n, red_ship =n, orange_bus =y.  [7|3] Val: 0.277
Class =r :- blue_train =y, tree =y, radio =n, book =y.  [7|3] Val: 0.261
Class =d.  [252|11]
```

statistic: min | tokenization: off

```
Class =r :- bush =n, flower =y.  [121|12] Val: 0.823
Class =r :- bush =n, tree =n, plant =y.  [12|3] Val: 0.454
Class =r :- blue_train =y, newspaper =n, yellow_bicycle =y, book =y.  [18|3] Val: 0.429
Class =d.  [249|17]
```

statistic: mean | tokenization: off

```
Class =r :- bush =n, flower =y, plant =y, newspaper =n.  [96|3] Val: 0.828
Class =r :- bush =n, tree =n.  [33|12] Val: 0.59
Class =r :- blue_train =y, newspaper =n, tree =y.  [18|2] Val: 0.451
Class =r :- flower =y, plant =y, yellow_bicycle =y, book =y.  [9|6] Val: 0.281
Class =d.  [244|12]
```

statistic: max | tokenization: off

```
Class =r :- bush =n, flower =y, newspaper =n, plant =y, radio =n, tree =n, red_ship =n.  [79|0] Val: 0.87
Class =r :- bush =n, flower =y, blue_train =y.  [41|9] Val: 0.652
Class =r :- bush =n, tree =n, blue_train =y, plant =y.  [12|3] Val: 0.465
Class =r :- blue_train =y, newspaper =n, radio =n, book =y.  [14|2] Val: 0.47
Class =r :- flower =y, plant =y, newspaper =n, orange_bus =y, radio =y.  [6|0] Val: 0.331
Class =r :- bush =n, plant =n, red_ship =n.  [4|1] Val: 0.257
Class =r :- blue_train =y, tree =y, bush =y, journal =y, yellow_bicycle =y, book =y.  [4|1] Val: 0.25
Class =d.  [251|8]
```

statistic: min | tokenization: on

```
Class =r :- bush =n, flower =y.  [121|12] Val: 0.823
Class =r :- bush =n, tree =n, plant =y.  [12|3] Val: 0.454
Class =r :- blue_train =y, newspaper =n, yellow_bicycle =y, book =y.  [18|3] Val: 0.442
Class =d.  [249|17]
```

statistic: mean | tokenization: on

```
Class =r :- bush =n, flower =y, plant =y, newspaper =n.  [96|3] Val: 0.828
Class =r :- bush =n, tree =n.  [33|12] Val: 0.59
Class =r :- blue_train =y, newspaper =n, book =y, yellow_bicycle =y.  [21|3] Val: 0.492
Class =r :- bush =n, plant =y, orange_bus =y.  [4|1] Val: 0.227
Class =d.  [248|14]
```

statistic: max | tokenization: on

```
Class =r :- bush =n, flower =y, newspaper =n, plant =y, radio =n, tree =n, red_ship =n.  [79|0] Val: 0.87
Class =r :- bush =n, flower =y, blue_train =y.  [41|9] Val: 0.652
Class =r :- bush =n, blue_train =y, plant =y, radio =n.  [16|4] Val: 0.496
Class =r :- blue_train =y, newspaper =n, book =y, yellow_bicycle =y, journal =y.  [13|1] Val: 0.467
Class =r :- flower =y, tree =y, blue_train =y.  [5|1] Val: 0.288
Class =r :- bush =n, plant =n, red_ship =n.  [3|1] Val: 0.224
Class =d.  [251|11]
```

Decrease in Modeling Quality with Increasing Semantic Influence on Datasets with 30-60% Attribute Labels Found in WN

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

Heuristic	influence										
	0	10	20	30	40	50	60	70	80	90	100
m-Estimate	66.73	67.20	67.48	67.94	67.86	67.70	67.69	67.44	67.39	67.35	46.41
Accuracy	64.89	65.93	65.71	65.93	65.98	66.03	65.77	65.99	65.99	65.72	45.65
Laplace Est.	66.89	67.03	66.70	66.66	66.61	66.56	67.11	67.50	67.35	67.49	45.78

Increase in Semantic Quality with Increasing Semantic Influence on Datasets with 30-60% Attribute Labels Found in WN

	0%	10%	20%	30%
Average Semantic Coherence	6.111%	10.493%	12.806%	14.346%
Average Rule length	3.34	2.20	3.13	3.03
Number of Rules	17.0	15.0	14.0	15.0

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

- Princeton University. About WordNet., 2010.
  URL http://wordnet.princeton.edu/.
- Kevin Bache and Moshe Lichman. UCI Machine Learning Repository, 2013.
  URL http://archive.ics.uci.edu/ml.
- Philip Resnik. Using Information Content to Evaluate Semantic Similarity in a Taxonomy.
  In Proceedings of the 14th International Joint Conference on Artificial Intelligence, volume 1, 1995.
- Dekang Lin. An Information-Theoretic Definition of Similarity.
  In ICML, pages 296–304, 1989.
- Frederik Janssen and Markus Zopf. The SeCo-Framework for Rule Learning.
  In Proceedings of the German Workshop on Lernen, Wissen, Adaptivität - LWA2012, 2012.