Bias/Variance Analysis for Network Data

Jennifer Neville and David Jensen
Knowledge Discovery Laboratory
University of Massachusetts Amherst
Collective inference

Apply models to *collectively* infer class labels throughout network

Exploit autocorrelation to improve model performance

Collective SRL models

- Probabilistic relational models (e.g., RBNs, RDNs, RMNs)
- Probabilistic logic models (e.g., BLPs, MLNs)
- Adhoc collective models (e.g., pRNs, LBC)
Comparing collective models

![Diagram showing AUC vs. Proportion Labeled for WebKB dataset. The red line represents latent group models, and the blue line represents relational dependency networks. Error bars indicate variability.](image-url)
Comparing collective models

Why do RDNs perform poorly when few instances are labeled in test set?
Understanding RDN performance

Hypothesis
- High autocorrelation $\rightarrow$ features selection chooses class label rather than observed attributes
- Few labeled test set instances $\rightarrow$ identifiability problem
- Gibbs sampling $\rightarrow$ increased variance

How to evaluate hypothesis?
- Variance is due to collective inference procedure
- Need an analysis framework that can differentiate model errors due to learning and inference
Bias/variance analysis

Conventional bias/variance analysis
- Decomposes errors due to learning alone
- Assumes no variation due to inference

Relational bias/variance analysis
- Collective inference introduces new source of error
- SRL models exhibit different types of errors
- Network characteristics affect performance
Conventional bias/variance framework
Conventional bias/variance framework

- Expected error per instance
- Decompose into model bias/variance

Model predictions
Bias/variance framework for relational data

- Training Set
- Samples
- Models
- Fully labeled Test Set
- Model predictions
Bias/variance framework for relational data

- Measure *learning* bias and variance with full labeling
Bias/variance framework for relational data

Training Set

Samples

Models

Test Set

Inference Runs

Model predictions
Bias/variance framework for relational data

- Measure total bias and variance
  - Expectation over training and test sets
Bias/variance framework for relational data

- Measure *learning* bias and variance with full labeling
- Measure *total* bias and variance
  - Expectation over training and test sets
- Difference: *inference* bias and variance
Synthetic data experiments

Vary group size, linkage, autocorrelation

Compare LGMs, RDNs, RMNs

Preliminary findings

- LGMs: high learning bias when algorithm cannot identify underlying group structure
- RDNs: high inference variance when little information seeding inference process
- RMNs: high inference bias when network is densely connected or tightly clustered
Feature selection increases RDN inference variance
Feature selection increases RDN inference variance
Modified inference decreases variance

**Synthetic Data**

- Squared Loss
  - RDN
  - RDN-mod

- Variance
  - RDN-total
  - RDN-learn
  - RDN-mod-total
  - RDN-mod-learn
Improved performance on real data
Conclusions

Framework can be used to explain mechanisms behind SRL model performance
- Improves understanding of model behavior
- Suggests algorithmic modifications to increase performance

Future work
- Extend framework (e.g., loss functions, joint estimation)
- Investigate interaction effects between learning and inference errors
- Real data experiments to evaluate design choices
Further information:

jneville@cs.umass.edu
kdl.cs.umass.edu