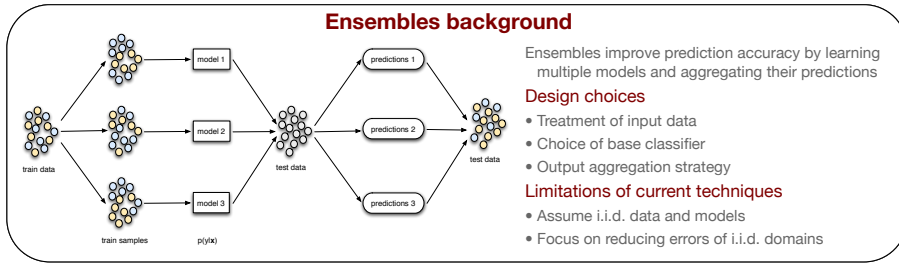
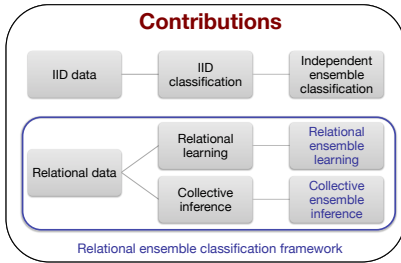


Ensemble classification for relational domains

Hoda Eldardiry, *Purdue University*, (Advisor: Jennifer Neville)

Problem: "Relational learning can benefit from ensembles. But, characteristics of relational data and relational models must be considered in ensemble design."



Relational ensemble learning

Challenges

- Bagging generates datasets by *Independent* bootstrap sampling to learn models, then aggregate their predictions
- To improve classification of relational data, bagging can be used, but the typical independent sampling approach is problematic when applied to relational data

Independent sampling from relational data

- ⊗ Destroys link structure which can otherwise be exploited to improve classification
- ⊗ Underestimates population variance in relational data, so bagging fails to fully reduce prediction variance

Relational Subgraph Resampling (RSR) (SNA-KDD'08)

Key insight

- Effective sample size is approximated by number of groups of interdependent instances

Approach

- Sample subgraphs with replacement by BFS from random seeds
- Link subgraphs by feature & structure similarities

Advantages

- ⊙ Accurately captures population variance
- ⊙ Improves bagging accuracy
- ⊙ Preserves dependencies among instances
- ⊙ Improves base classifier accuracy

Evaluation

RSR improves bagging over independent sampling

- Task: predict webpage category on WebKB data
- ⊙ Result: bagging using RSR significantly outperforms bagging using independent sampling

RSR improves bagging due to better population variance estimation

- Task: estimate population variance of feature scores on synthetic data
- ⊙ Result: RSR estimates population variance more accurately than independent sampling

Collective ensemble inference

Challenges

- Collective classification models use joint inference of linked nodes to improve predictions
- To improve ensemble classification of relational data, a collective inference base classifier can be used, but this raises two challenges

Using a collective inference base classifier on network data

- ⊗ Collective inference models introduce inference error, but typical ensembles only reduce errors in learning
- ⊗ Collective inference exploits linkage to improve predictions, what if there are multiple link graphs, can the ensemble allow exploiting all of them?

Collective Ensemble Classification (CEC) (AAAI'11)

Key insight

- Unique opportunity to aggregate predictions across models during collective inference process

Approach

- Learn a model per link graph
- Apply models simultaneously for collective inference, aggregating inferences across models *during* inference.
- Combine final models' predictions after inference.

Advantages

- ⊙ Each base model utilizes one link structure
- ⊙ The ensemble utilizes all link structures
- ⊙ Treating link graphs separately for learning reduces learning variance
- ⊙ Cross-model inference aggregation reduces inference variance

Evaluation

CEC maximizes label propagation by exploiting all link sources

- Task: predict Facebook user political view
- ⊙ Result: CEC significantly outperforms all models even on sparsely labeled data

CEC effectively utilizes additional link sources

- Task: predict binary class label on synthetic data
- ⊙ Result: CEC significantly outperforms all models as more link graphs considered

CEC improves accuracy due to inference variance reduction

- Task: increase #models aggregated during inference
- ⊙ Result: increase in accuracy coincides with decrease in inference variance

Contributions and Future Work

- Relational ensemble learning can be combined with collective ensemble inference into a full ensemble framework that works for both single- and multi- graph settings, and reduces errors due to variance in both learning and inference
- Have shown empirically that (1) relational ensembles improve classification accuracy, and (2) the accuracy improvement is due to reduction in variance
- Current work: Theoretical analysis to show how, why, and when ensembles improve relational classification