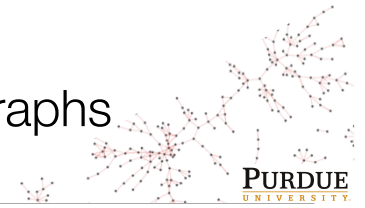


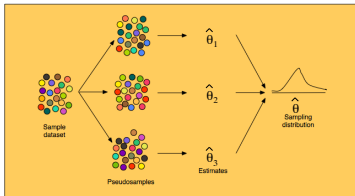
A Resampling Technique for Relational Data Graphs

Hoda Eldardiry and Jennifer Neville, CS Department, Purdue University



Introduction

- Resampling (a.k.a. bootstrapping) is a computationally-intensive statistical technique for estimating the sampling distribution of an estimator.
 - Resampling techniques generate *pseudosamples* from an underlying population by sampling with replacement from a single sample dataset.
- Resampling is used in many machine learning algorithms, including ensemble methods, active learning, and feature selection.



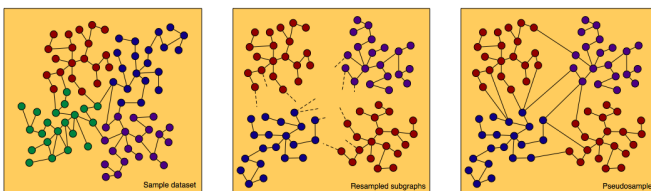
Motivation

- It is straightforward to sample with replacement from propositional data that are independent and identically distributed (IID).
- However, it is not clear how to sample with replacement from an interconnected relational data graph with dependencies among related instances.
 - In a relational dataset, dependencies among groups of linked instances reduce the effective sample size of the dataset... which increases the variance of statistics estimated from the dataset.
 - IID resampling ignores dependencies among data and thus will underestimate the variance of sampling distributions.
- Goal:
 - Develop a relational resampling technique to accurately estimate variance of sampling distributions of heterogeneous, dependent, relational data.

Background: Resampling IID Data

- Pseudosamples are constructed by independent random sampling with replacement.
- To estimate the sampling distribution of a statistic from a set of i.i.d. data, this approach is applied m times to create m pseudosamples of the data. The statistic is then calculated on each pseudosample and the empirical distribution of values is returned as an approximation of the statistic's sampling distribution.
- Assumes data is iid. If the assumption is violated, and there is correlation among the instances, then the variance of sampling distribution will be underestimated.

Approach: Relational Subgraph Resampling

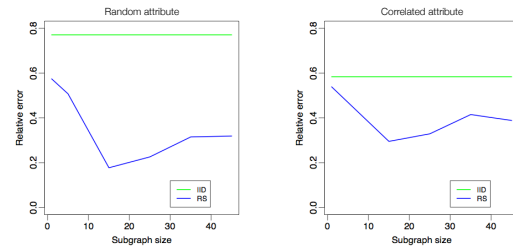


- Novel two-phase approach for resampling relational data.
- Phase 1:
 - Key idea: Resample subgraphs instead of single instances.
 - Preserves local relational dependencies among instances in the subgraph.
 - When autocorrelation and linkage are high, the effective sample size will be approximated by the number of underlying groups.
 - Approach:
 - Repeatedly select a subgraph of size b via breadth-first search from a randomly selected seed node.
- Phase 2:
 - Key idea: Link within the subgraphs to match the global properties of the data.
 - Maximizing attribute similarity across links aims to maintain relational autocorrelation.
 - Maximizing link similarity among neighbors aims to maintain the link structure.
 - Pseudosamples are generated with sufficient overall variance.
 - Approach:
 - Peripheral nodes from various subgraphs are linked to either:
 - a missing neighbor, if found in another subgraph, or
 - if not in sample, then to the node that is most similar to its missing neighbor, where similarity is determined by both attributes and linkage:

$$\text{Similarity}(v_i, v_j) = \alpha * \text{attributeSim}(v_i, v_j) + (1 - \alpha) * \text{linkSim}(v_i, v_j)$$

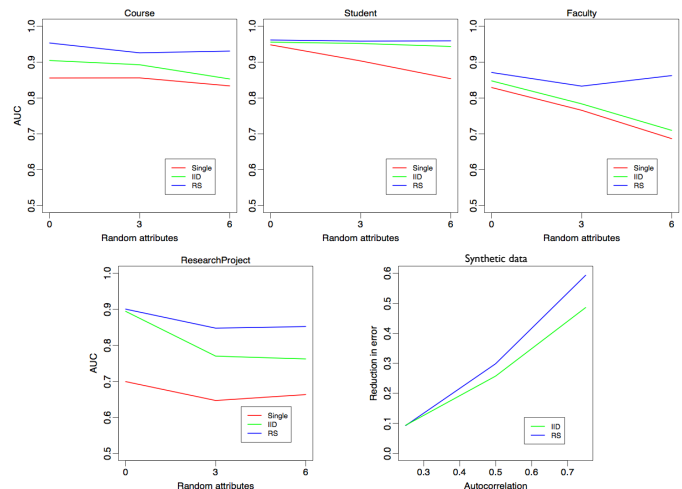
Experiments: Variance Estimation

- Data:
 - Synthetic datasets with relational autocorrelation and concentrated linkage generated with a latent group model (Neville and Jensen '05)
- Methodology:
 - Population variance of feature scores is estimated empirically from 100 datasets
 - From a single test dataset, resampling methods are used to generate pseudosamples and estimate the sampling distribution of feature scores
 - Evaluation: Compare resampling variance estimate to (empirical) population variance
- Results:
 - Relational subgraph (RS) resampling produces more accurate variance estimate than IID resampling, improvement is more pronounced for random feature
 - RS resampling performed best when subgraph size is equal to underlying group size.



Experiments: Bagging

- Synthetic data: as described above
- Real world data: WebKB (web pages from 4 CS departments, labeled with the categories: course, faculty, staff, student, research project, or other)
- Methodology:
 - Learn relational probability tree (Neville et al. '03) to predict the class label.
 - Use bagging to improve model accuracy (by reducing prediction variance).
 - Measure reduction in AUC error achieved by bagging over just a single model.
- Results:
 - Bagging using RS resampling outperforms bagging using IID resampling.
 - Difference in error reduction increases as autocorrelation level increases or number of random attributes increases.



Conclusions

- In contrast to IID resampling, the proposed RS resampling technique accounts for the link structure and attribute dependencies in the data.
 - It maintains the local autocorrelation dependencies while allowing the global structure to vary as if we were sampling from the population.
 - It avoids overestimating the effective sample size and thus is able to be used for accurate variance estimation.
- Future work
 - Use subsampling to calculate statistics on smaller subgraphs and then scaling the estimates to produce a valid estimate of the sampling distributions on the full graph.
 - Develop model selection and active learning techniques that exploit the increased accuracy afforded by RS resampling.