A Resampling Technique for Relational Data Graphs

Hoda Eldardiry
Department of Computer Science
Purdue University
hdardiry@cs.purdue.edu

Jennifer Neville
Department of Computer Science and Statistics
Purdue University
neville@cs.purdue.edu

ABSTRACT

Resampling (a.k.a. bootstrapping) is a computationally-intensive statistical technique for estimating the sampling distribution of an estimator. Resampling is used in many machine learning algorithms, including ensemble methods, active learning, and feature selection. Resampling techniques generate *pseudosamples* from an underlying population by sampling with replacement from a single sample dataset. It is straightforward to sample with replacement from propositional data that are independent and identically distributed (i.i.d.). However, it is not clear how to sample with replacement from an interconnected relational data graph with dependencies among related instances. In this paper, we develop a novel method for resampling from relational data that uses a subgraph sampling approach to preserve the local relational dependencies while generating a pseudosample with sufficient global variance. We evaluate our approach on synthetic data, showing that compared to an i.i.d. resampling approach it results in significantly lower error when used to estimate the variance of feature scores. We also evaluate our approach on a real-world relational classification task, showing that it improves the accuracy of bagging when compared with i.i.d. resampling.

1. INTRODUCTION

Resampling is a statistical technique that approximates sampling from the true underlying population by sampling with replacement from a single dataset D to create a set of *pseudosamples* $\mathbf{D}'$. It can be used to estimate the sampling distribution of a statistic θ measured on D. The value of θ is calculated for each pseudosample and the resulting distribution of values is used as an approximation of the *sampling distribution* of θ. The approximate sampling distribution is useful for a wide variety of data mining and machine learning tasks. For example, resampling can be used to estimate the variance of a model and/or feature scores for model selection algorithms (see e.g., [13]). Also, resampling can be used in bagging techniques to estimate the mean prediction for an instance x—by learning an ensemble of models, one for each pseudosample, and applying each model to predict to a value $\hat{y}$ for x (see e.g., [1]).

Conventional approaches to resampling assume that the data consist of independent and identically distributed (i.i.d.) instances, $D = \{X_1, X_2, ..., X_n\}$, thus the algorithms simply sample instances with replacement independently from D. Relational data violates this assumption of independence—the data instances typically have dependencies both as a result of direct relations and through chaining multiple relations together. For example, in citation networks, there is correlation among the topics of two papers if one of the papers cites the other. In addition, there is correlation among the topics of papers that share a common coauthor.

Two common characteristics of relational data—concentrated linkage and relational autocorrelation—have been shown to reduce the *effective sample size* of relational datasets [7]. Concentrated linkage occurs when many objects are linked to the same neighbor and it is a common characteristic of relational data, which typically have skewed degree distributions. For example, many articles in the scientific literature are published in a small number of journals. Relational autocorrelation refers to correlation among the values of the same attribute (e.g., class label) on pairs of related instances. For example, two hyperlinked pages are more likely to share the same topic than two randomly selected web pages [15]. Concentrated linkage and relational autocorrelation combine to reduce the effective sample size of a data set by creating dependencies among a large set of linked instances. One can imagine this as having an urn filled with bunches of grapes—when you reach in to grab a single instance, you end up pulling out a set of interconnected instances instead.

Decreased effective sample size will increase the variance of parameters that are estimated from relational data. However, naive resampling techniques, that ignore the dependencies and link structure and sample independently from the instances, will not accurately capture this increased variance. Thus, i.i.d. resampling techniques should consistently underestimate the variance of sampling distributions in relational data. Our aim in this work is to develop a relational resampling technique that accurately estimates the variance of sampling distributions of statistics for heterogeneous, dependent data. The goal is to preserve the local relational dependencies (e.g., relational autocorrelation) and link structure, while introducing sufficient variance at a global level to model draws from the underlying population.

The 2nd SNA-KDD Workshop '08 (SNA-KDD'08), August 24, 2008 , Las Vegas, Nevada , USA .

In this paper, we present a novel relational subgraph resampling approach. The key idea of the approach is to sample *subgraphs* with replacement from the original data, thereby preserving the local link and attribute structure within the subgraphs. This is augmented with a procedure that links up the selected subgraphs in an attempt to match the global properties of the data without reproducing them exactly.

We evaluate our resampling approach in two contexts. In the first set of experiments, we use synthetic data to show that our method produces more accurate variance estimates than naive i.i.d. resampling, for a feature score calculation task. In the second set of experiments, we compare our resampling approach to i.i.d. resampling using ensemble methods for classification of relational data. We demonstrate that bagging with our approach results in significant improvements in accuracy for both synthetic data and real-world relational datasets.

2. RESAMPLING

Resampling is a computationally-intensive statistical technique used when the observed sample is drawn from a population about which no other information is available. It works by generating multiple pseudosamples by drawing with replacement from the original data as if it were the population [10]. Each pseudosample contains as many instances as the original data set. Some instances in the original data set will occur multiple times in a given pseudosample, and others will not occur at all. The basic idea of resampling is that, in the absence of any other information about the population, the observed sample contains the best available information about the underlying population. Thus, resampling from the sample is the best way to approximate draws from the population.

2.1 Applications

Resampling is used for estimating the sampling distribution of a statistic θ empirically. In practice, it is used to assess a wide variety of statistics including: the generalization accuracy of models, feature scores, predicted class labels, and model parameter estimates.

Machine learning techniques that use resampling are generally concerned with estimating the mean and/or variance of sampling distributions. For example, model selection techniques may use resampling to estimate the mean generalization error of different models in order to identify the model with lowest average error [13]. Alternatively, feature selection techniques may use resampling to estimate the standard errors of feature scores or model coefficients in an effort to identify which features are most relevant to the task [4].

Resampling techniques are also used in bagging (bootstrap aggregation) methods, which learn ensembles of classifiers from sets of pseudosamples [1]. Classification algorithms use ensemble techniques to reduce variance and improve the stability of predictions, thus improving model accuracy. Bagging methods construct a number of different training sets by resampling from the original training set, then a model is learned from each new training set. Since the resampled training sets contain different combinations of the instances in the original dataset, models learned from the datasets will vary substantially if the model is unstable (i.e., small changes in the training set result in differences in the learned models). Each learned model is applied to the test set for classification and the predictions for an instance x are averaged to produce the final prediction for x. This reduces the variance of the classification model, which can often improve prediction accuracy.

Another application of resampling is active learning. The goal of active learning is to learn an accurate model with as few labeled instances as possible. Many criteria have been proposed to determine the most valuable instance for labeling. In particular, some methods have proposed selecting the instances whose prediction have highest variance, which is determined by resampling (see e.g., [11]).

2.2 Methods

2.2.1 Resampling IID Data

Assuming an independent and identically distributed (i.i.d.) sample, the pseudosamples are constructed by independent random sampling, with replacement. We will refer to this approach as IID resampling.

IID Resampling ($D = \{X_1, .., X_n\}$)

1. Let $D_{PS} = \{\}$
2. for $j = 1..n$
3. Randomly select X_s from D
4. D_{PS} += $\{X_s\}$

To estimate the sampling distribution of a statistic from a set of i.i.d. data, IID resampling is applied m times to create m pseudosamples (D_{PS}) of the data (D). 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.

2.2.2 Resampling Dependent Data

It is difficult to resample dependent data because the instances are interconnected in complex ways, hence the i.i.d. assumption is violated. When the data instance are interdependent, pseudosamples generated by IID resampling are likely to exhibit less variance than the underlying population distribution. Dependencies among instances reduce the *effective* sample size of the data and thus increase the variance of statistics estimated from those data [7]. Resampling techniques that ignore the dependencies and sample independently from the instances will be replicating the actual sample size, not the effective sample size, and thus they are likely to underestimate the variance of statistics calculated from the data.

Previous work in spatial statistics has investigated graph-based *reuse sampling* techniques for lattice graphs, which use small, overlapping subgraphs as pseudosamples [2, 14, 6]. A statistic is repeatedly calculated on smaller subgraphs to estimate the variance of its sampling distribution. This estimate is then rescaled to reflect the number of instances in the original data sample. For example, consider a regular lattice graph with degree four, we could use contiguous subgraphs of length four (i.e., 4×4 squares) as the pseudosamples and then scale the estimate of variance to approximate the original sample size.

In spatial and temporal datasets, where the link structure is generally homogeneous (either a line graph or a lattice of fixed degree), the choice of scaling factor is relatively straightforward. In relational data it will be difficult to determine the effective sample size of a relational data set analytically due to heterogeneous link structure.

For example, consider a bipartite graph with 1000 objects X connected to 100 objects Y. There is a binary class label on the objects X and a binary attribute on the objects Y. When calculating feature scores concerning X, the actual sample size is $N_X = 1000$. However, if the class labels are perfectly autocorrelated through the objects Y (i.e., all X connected to the same object Y_i share the same class label value), then the *effective* sample size is $N_Y = 100$. Again one can think of this as having an urn filled with bunches of grapes—when you reach in to grab a single X you end up pulling out a single Y and all of its neighbors $\mathbf{X}$.

In practice, when the level of autocorrelation is somewhere between 0 and 1, the effective sample size N_{ESS} will be between the number of coordinating objects and the number of instances (i.e., $N_Y \leq N_{ESS} \leq N_X$). The goal of this work is thus to develop a relational resampling technique that accurately preserves the effective sample size of the data, thus producing more accurate estimates of the sampling distributions of statistics for heterogeneous, dependent data.

In order to maintain the dependencies among related data instances, we propose to use a two-phase relational subgraph resampling technique. First, we use subgraph sampling to identify and sample sets of interconnected instances with each selection. Then, our approach links up the selected subgraphs in an attempt to preserve various relational properties throughout the sample. We refer to our method as relational subgraph (RS) resampling.

2.2.3 Relational Subgraph Resampling

Relational subgraph (RS) resampling is a novel approach for resampling relational data. The first phase of the algorithm selects subgraphs based on snowball sampling [5]. It repeatedly selects a subgraph of size b via breadth-first search from a randomly selected seed node. The second phase then links up the selected subgraphs. The aim is to preserve the local relational dependencies among instances in the subgraph, while generating a pseudosample with sufficient global variance by linking up the set of selected subgraphs. The key idea behind our approach is that when autocorrelation is high, the effective sample size is determined by the number of underlying groups in the data (e.g., the bunches of grapes). As such, our approach attempts to sample these groups instead of single instances, thus preserving the effective sample size of the data.

One challenge is how to link up the subgraphs into a single relational data graph. Due to the varied link structure of relational data, there will be a large number of nodes on the periphery of the selected subgraphs. If the peripheral nodes are missing a significant portion of their neighbors, this could bias the properties of the sample. The potential for bias due to peripheral nodes is much greater in relational data with varied link structure than temporal or spatial data with regular link structure. Consider a lattice subgraph where each interior node has four neighbors. The peripheral nodes each have three neighbors, except for the four corners which have two. Each peripheral node is missing at most 50% of its neighbors. However, in relational data with concentrated linkage, if the peripheral nodes in the sample are hub nodes with high degree from the original data, they may be missing almost all their neighbors (i.e., ≈100%). To deal with this issue, we outline a procedure to link up the peripheral nodes in the selected subgraphs, which attempts to maintain the global graph properties and attribute dependencies of the original data. More specifically, the relational autocorrelation is maintained by maximizing attribute similarity between nodes as they are linked, while the link structure is maintained by considering the neighborhood similarity when selecting nodes to link.

We outline our modified resampling procedure in pseudocode below. Given a sample relational data graph $G = (V, E)$, it returns a pseudosample data graph $G_{PS} = (V_{PS}, E_{PS})$. The first phase samples a set of $N_S = \lceil \frac{|V|}{b} \rceil$ subgraphs of size b from G, using breadth-first search from N_S randomly selected seed nodes. We sample with replacement from the graph, so a node may appear in multiple subgraphs, one subgraph, or none.

The pseudosample node set (V_{PS}) consists of all the nodes selected in the subgraphs (suitably relabeled so multiple copies of the same original node are distinguishable). The pseudosample edge set (E_{PS}) initially consists of all the edges within the selected subgraphs. This is augmented by a process that links up the peripheral nodes across subgraphs, choosing the links that are most *similar* to the links that were broken by the subgraph selection process. For example, if a peripheral node v_p linked to node v_m in the original dataset but v_m was not selected as a member of v_p's subgraph, we will find a node similar to v_m in another subgraph and link it to v_p. We first attempt to link v_p to a copy of v_m in another subgraph in the pseudosample. If there are multiple copies, we choose the copy with the shortest path length to v_p and with the greatest number of missing neighbors. Then we create links for any nodes with neighbors still missing after the first pass. For example, if there were no copies of v_m selected for the pseudosample, then we would not create a corresponding link for v_p in the first pass. The second pass looks for the node in the pseudosample that is most similar to v_m. We calculate node similarity based on both the attributes of the nodes and on their link structure (i.e., the number of neighbors they have in common in the original data). Again, if there are multiple nodes with the same (maximum) similarity to v_m, we choose the node with the the shortest path length to v_p and with the greatest number of missing neighbors.

We use the following similarity function to compare nodes based on both attributes and links:

$$\text{Sim}(v_i, v_j) = \alpha * \text{aSim}(v_i, v_j) + (1 - \alpha) * \text{lSim}(v_i, v_j)$$

where the attribute similarity is defined as $\text{aSim}(v_i, v_j) =$ # shared attribute values between v_i and v_j, and the link similarity is defined as $\text{lSim}(v_i, v_j) =$ # common neighbors between v_i and v_j. In the experiments reported in this paper, we set $\alpha = 0.15$ to upweight the importance of matching on link structure.

RS Resampling $(G = (V, E), b)$

1. Let $V_{PS} = \emptyset$
2. Let $E_{PS} = \emptyset$
3. for s in $1..\lceil \frac{|V|}{b} \rceil$
4. choose a seed node v_s randomly from V
5. construct V^S by selecting $b-1$ nodes around v_s using breadth-first search
6. Let $E^S = \{e_{ij} \in E \ s.t. \ v_i, v_j \in V^S\}$
7. $V_{PS} \mathrel{+}= V^S$
8. $E_{PS} \mathrel{+}= E^S$
9. for each V^S
10. for each $v_i \in V^S$
11. $N_i^S = \{v_j \ s.t. \ e_{ij} \in E \wedge v_j \notin V^S\}$
12. while *true*
13. $update = false$
14. for each *node* $v_i \in V_{PS}$
15. if $|N_i^S| > 0$
16. let v_j be a random select from N_i^S
17. let $C_j = \{v_k : v_k \equiv v_j \wedge v_k \in V^{S' \neq S} \wedge v_i \in N_k^S$
18. Select $v_m \in C_j \ s.t. \ v_m = argmin \operatorname{Path}(v_m, v_i)$, break ties by maximizing $|N_m|$
19. if $v_m \neq null$
20. $N_i^S = N_i^S - \{v_j\}; \ \ N_m^{S'} = N_m^{S'} - \{v_i\}$
21. $E_{PS} = E_{PS} + \{e_{im}\}$
22. $update = true$
23. break if $update = false$
24. while *true*
25. $update = false$
26. for each *node* $v_i \in V_{PS}$
27. if $|N_i^S| > 0$
28. let v_j be a random select from N_i^S
29. let $C_j = \{v_k : |N_k| > 0 \wedge v_k \in V^{S' \neq S}$
30. Select $v_m \in C_j \ s.t \ v_m = argmax \operatorname{Sim}(v_m, v_j)$, break ties by $\operatorname{argmin} \operatorname{Path}(v_m, v_i), \operatorname{argmax} |N_m|$
31. if $v_m \neq null$
32. $N_i^S = N_i^S - \{v_j\}; \ \ N_m^{S'} = N_m^{S'} - \{v_i\}$
33. $E_{PS} = E_{PS} + \{e_{im}\}$
34. $update = true$
35. break if $update = false$
36. return $G_{PS} = (V_{PS}, E_{PS})$

3. EXPERIMENTAL EVALUATION

To evaluate our resampling methodology, we applied it in two different relational settings. First, to estimate a sampling distribution of feature scores on synthetic data and calculate an accurate estimate of the variance of the feature score distribution. Second, to improve the accuracy of bagging on both real-world and synthetic relational classification tasks.

3.1 Variance Estimation

For synthetic relational datasets that exhibit relational autocorrelation and concentrated linkage, we use RS resampling to calculate an approximation of the unknown sampling distribution of features scores and estimate the variance of their distribution. We compare to IID resampling and show that RS resampling results in more accurate variance estimates on both correlated and random attributes.

3.1.1 Data

Our synthetic datasets are generated with a latent group model [8]. The relational data graphs are homogeneous (i.e., single object type); each object has a boolean class label C (that is determined by the type of group to which it belongs), and two boolean attributes X_0 and X_1 . We generated datasets with 270 objects and groups of size 15. The class label C has an autocorrelation level of 0.5 and the probabilities of intra- and inter-group linkage are 0.4 and 0.004 respectively. The attribute X_0 is correlated with C, and X_1 has no dependencies (i.e., it is random).

3.1.2 Methodology

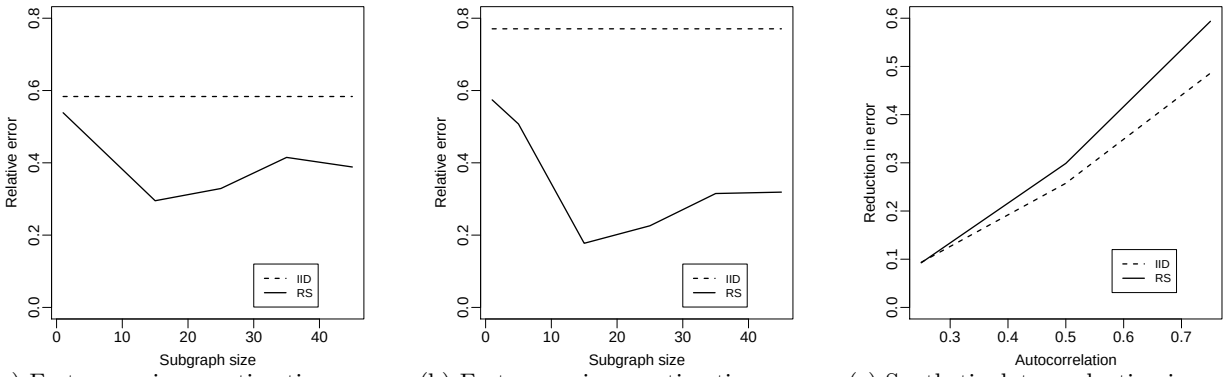
We use both IID and RS resampling to estimate the variance of relational feature scores in our synthetic datasets. To calculate variance, we create 20 pseudosamples, calculate a feature score for each sample, then we calculate the variance (Var_{est}) of the distribution of the 20 feature scores. We consider two relational features: one that is correlated with the class (i.e., MODE(linked.X_0)) and one that is random (i.e., MODE(linked.X_1)). The feature score calculation assesses the correlation of the feature values with the class labels C using Pearson's corrected contingency coefficient [12].

To evaluate the accuracy of the feature score variance estimates, we compare to the empirical variance of the feature scores in the synthetic datasets. We estimate the population variance Var_{pop} of the features by generating 100 different datasets and calculating the variance from the empirical distribution of features scores in those datasets. We use relative error as a measure of accuracy: $\frac{(Var_{pop} - Var_{est})}{Var_{pop}}$.

3.1.3 Results

We calculated the feature score variances using RS resampling and IID resampling and measured the relative error for both approaches. We report the average relative error over 10 trials. For RS resampling, we evaluate performance on subgraphs of varying sizes: {1,5,15,25,35,45}. Since our algorithm aims to exploit the underlying groups structure, we expect it to outperform the IID resampling most significantly when the subgraph size is the same as the average group size (15) of the generated data.

Figure 1(a) and 1(b) graphs the average relative error in variance for both IID resampling and RS resampling using different subgraph sizes. Figure 1(a) graphs the results for the correlated feature and Figure 1(b) graphs the results for the random feature. Both plots show that RS resampling results in lower error than IID resampling. Furthermore, the RS resampling estimates of variance increase in accuracy as the subgraph size approaches the underlying group size (15). Notice also that RS resampling shows a more significant reduction in estimation error for the feature formed from the random attribute (Figure 1(b)).



(a) Feature variance estimation error for correlated attribute

(b) Feature variance estimation error for random attribute

(c) Synthetic data: reduction in error for bagging approaches

Figure 1: Experimental results on synthetic data.

Accurately estimating the variance of random (or irrelevant) features is likely to impact model learning more significantly than accurate estimation for real features, since reduction in effective sample size increases the risk of Type I errors. Improved resampling techniques can be used to develop more accurate feature selection models, reducing the risk that random features are selected for inclusion in relational models when the data exhibit linkage and autocorrelation.

3.2 Bagging

In classification, bagging is used to improve model accuracy by reducing prediction variance. In bagging, multiple training sets are generated by resampling from the original training set, then a model is learned on each pseudosample. Each of the learned models is then applied to the test set, producing a set of predictions for each instance, which are then aggregated.

To evaluate our resampling method, we compared the classification accuracy of three modeling techniques: a single model, bagging using IID resampling, and bagging using RS resampling. We compare the performance of these three models on the WebKB dataset and synthetic datasets. The ensemble methods construct five pseudosamples and learn an ensemble of five models. For the WebKB experiment, RS resampling uses a subgraph size of 50.

3.2.1 Data

The synthetic data we used for these experiments was the same as described previously, except that we generated datasets of size 120 for training and 255 for testing.

The real-world relational dataset we used for model evaluation was collected by the WebKB Project [3]. The data consists of a set of 4,135 web pages from four computer science departments, labeled with the categories: course, faculty, staff, student, research project, or other. The classification task was to predict page category. As in previous work on this dataset, we do not try to predict the category "other".

3.2.2 Methodology

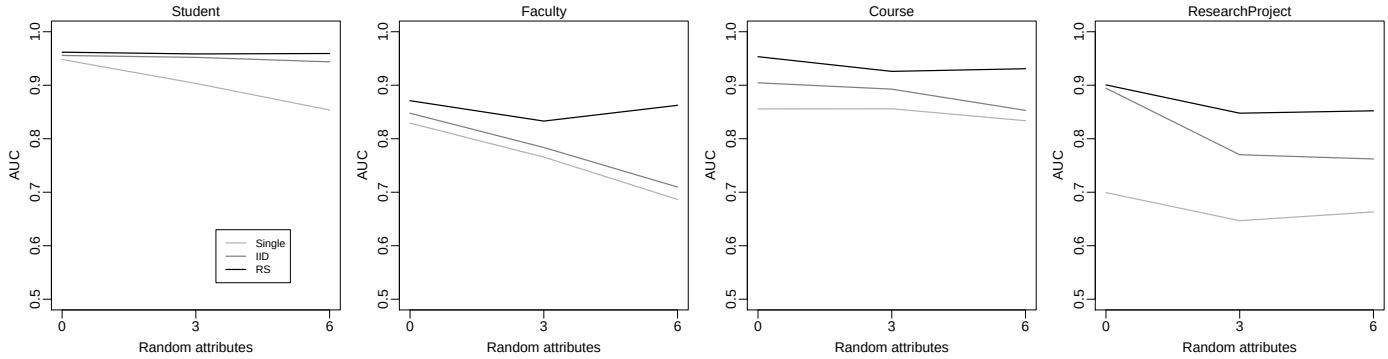
For the synthetic data experiments, we generated four training and testing sets of sizes 120 and 255 respectively, for a total of 16 training-test pairs. We learned relational probability trees (RPTs) [9] to predict C, using MODE, COUNT, and PROPORTION as the aggregation functions in feature construction. We measured the area under the ROC curve (AUC) of each type of model and then measured the error reduction of each bagging approach compared to the single model. We evaluate performance on datasets with increasing levels of autocorrelation $\{0.25,0.50,0.75\}$ to test the hypothesis that as autocorrelation increases the improvement of RS resampling over IID resampling should increase as well (due to a lower effective sample size).

In the WebKB experiments we learned RPTs to predict the page type using MODE features. For each of the three models, we used 12 training-testing pairs based on the four disjoint websites in WebKB. We compared the performance by measuring the AUC for each class label value separately. We also evaluated model robustness by adding random attributes to the data. We present the results for 0, 3 and 6 random attributes. This is to test the hypothesis that RS resampling will be more accurate at determining which features are irrelevant in relational data.

3.2.3 Results

Figure 1(c) presents the results for synthetic data experiments where we varied the level of autocorrelation in the data. We graph the reduction in AUC error achieved by each of the bagging models over the single model. Notice that as autocorrelation increases, the difference between RS and IID resampling approaches increases. These synthetic data experiments were conducted with relatively simple relational datasets. We expect that the performance difference between the two approaches will only increase on complex, real-world relational datasets.

Figure 2 shows the results for the WebKB data, plotting the AUC values for each class label value: Student, Faculty, Course and Research Project. Bagging with IID resampling produces higher accuracy than the single model. However, bagging with RS resampling is not only significantly better than the single model, it also achieves equivalent or better performance compared to IID resampling for all datasets. As more random attributes are included in the learning process, the single model and the IID bagging model both experience



(a) Student (b) Faculty (c) Course (d) Research Project

Figure 2: Experimental results on WebKB.

a degradation in performance while bagging using subgraph resampling is more robust.

4. CONCLUSIONS

Accurate resampling methods are important for many machine learning algorithms, including ensemble methods, active learning, and feature selection. Although it is straightforward to sample with replacement from IID data, it is more difficult to sample with replacement from an interconnected relational data graph in a manner that preserves the link structure and relational attribute dependencies.

In this paper, we present a novel method for resampling from relational data, which accounts for the link structure and attribute dependencies of the data. Resampling in this manner maintains the local autocorrelation dependencies while allowing the global structure to vary as if we were sampling from the population.

Since RS resampling explicitly accounts for the local structure in the data, it avoids overestimating the effective sample size and thus is able to be used for accurate variance estimation. To our knowledge, this is the first estimation algorithm that can effectively estimate sampling distributions in data with autocorrelation and heterogeneous link structure.

We evaluate our approach on synthetic data, showing that compared to an IID approach, RS resampling results in significantly lower error when used to estimate the variance of feature scores. We also evaluate our methodology on a real-world relational classification task, showing that it improves the accuracy of bagging when compared to IID resampling.

In future work we plan to investigate a more efficient subgraph resampling approach that uses subsampling to calculate statistics on smaller subgraphs and then scales the estimates to produce a valid estimate of the sampling distributions on the full graph. This approach should scale more effectively but may suffer from larger approximation errors and boundary effects. In addition, we are developing model selection and active learning techniques that exploit the increased accuracy afforded by RS resampling.

Acknowledgments

This material is based on research sponsored by DARPA, AFRL, and IARPA under grants HR0011-07-1-0018 and FA8750-07-2-0158. The views and conclusion contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of DARPA, AFRL, IARPA, or the U.S. Government.

5. REFERENCES

[1] L. Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996.

[2] E. Carlstein. The use of subseries values for estimating the variance of a general statistic from a stationary sequence. *The Annals of Statistics*, 14:1171–1179, 1986.

[3] M. Craven, D. DiPasquo, D. Freitag, A. McCallum, T. Mitchell, K. Nigam, and S. Slattery. Learning to extract symbolic knowledge from the World Wide Web. In *Proceedings of the 15th National Conference on Artificial Intelligence*, pages 509–516, 1998.

[4] D. Freedman and S. Peters. Bootstrapping a regression equation: Some empirical results. *Journal of the American Statistical Association*, 79(385), 1984.

[5] L. Goodman. Snowball sampling. *Annals of Mathematical Statistics*, 32:148–170, 1961.

[6] P. Hall and B. Jing. On sample reuse methods for dependent data. *Journal of the Royal Statistical Society, Series B*, 58:727–737, 1996.

[7] D. Jensen and J. Neville. Linkage and autocorrelation cause feature selection bias in relational learning. In *Proceedings of the 19th International Conference on Machine Learning*, pages 259–266, 2002.

[8] J. Neville and D. Jensen. Leveraging relational autocorrelation with latent group models. In *Proceedings of the 5th IEEE International Conference on Data Mining*, pages 322–329, 2005.

[9] J. Neville, D. Jensen, L. Friedland, and M. Hay. Learning relational probability trees. In *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 625–630, 2003.

[10] E. Noreen. *Computer Intensive Methods for Testing Hypotheses*. Wiley, 1989.

[11] M. Saar-Tsechansky and F. Provost. Active sampling for class probability estimation and ranking. *Machine Learning*, 54(2):153–178, 2004.

[12] L. Sachs. *Applied Statistics*. Springer-Verlag, 1992.

[13] J. Shao. Bootstrap model selection. *Journal of the American Statistical Association*, 91, 1996.

[14] M. Sherman. On sample reuse methods for dependent data. *Journal of the Royal Statistical Society, Series B*, 58:509–523, 1996.

[15] B. Taskar, P. Abbeel, and D. Koller. Discriminative probabilistic models for relational data. In *Proceedings of the 18th Conference on Uncertainty in Artificial Intelligence*, pages 485–492, 2002.