Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media

Network Sampling Designs for Relational Classification

Nesreen K. Ahmed, Jennifer Neville, Ramana Kompella
Department of Computer Science
Purdue University
West Lafayette, IN 47907
{nkahmed, neville, kompella}@cs.purdue.edu

Abstract

Relational classification has been extensively studied recently due to its applications in social, biological, technological, and information networks. Much of the work in relational learning has focused on analyzing input data that comprise a *single* network. Although machine learning researchers have considered the issue of how to sample training and test sets from the input network (for evaluation), the mechanisms which are used to *construct* the input networks have largely been ignored. In most cases, the input network has itself been sampled from a larger target network (e.g., Facebook) and often the researcher is unaware of *how* the input network was constructed or what impact that may have on evaluation of the relational models. Since the goal in evaluating relational classification algorithms is to accurately assess their performance on the larger target network, it is critical to understand what impact the initial sampling method may have on our estimates of classification accuracy. In this paper, we present different sampling methods and systematically study their impact on evaluation of relational classification. Our results indicate that the choice of sampling method can impact classification performance, and thus consequently affects the accuracy of evaluation.

Introduction

Online social activity and interaction is becoming embedded into the fabric of our society. From electronic communication (e.g., email, IMS) to social media (e.g., Twitter) to online content sharing (e.g., Facebook, flicker, youtube)—we are currently undergoing an explosive growth in the manner and frequency in which people interact online, both with each other and with content. Modeling and analyzing the large-scale datasets that are collected from these traces of electronic activity has become increasingly important for many applications, such as identifying the behavior and interests of individuals, as well as the structure and dynamics of human-formed groups over time. Studying complex relational networks is nevertheless a challenging task due to their heterogeneous, dependent structure, large size, and evolution over the time. In addition to these complexities for data analysis, there is also often a data acquisition bottleneck (Choudhury et al. 2010), which makes it necessary to analyze smaller *sample* subgraphs that were collected from the full network. Typically *network sampling* designs are used to select a subset of the nodes/edges from the full network, with the goal of collecting a *representative* subgraph from the larger network.

Previously, researchers have studied how to collect sample subgraphs that closely match *topological* properties of the network (Leskovec and Faloutsos 2006; Hubler et al. 2008; Ahmed, Neville, and Kompella 2011). However, since the topological properties are never entirely preserved, it is also important to study how the sampling processes impact the performance of applications overlaid on the networks. One such study recently investigated the impact of sampling designs on the discovery of the information diffusion process (Choudhury et al. 2010). In this paper, we study the question of how the choice of the sampling design can impact the performance of relational classification algorithms. Network sampling can produce subgraphs with imbalance in class membership and bias in topological features (e.g., path lengths, clustering) due to missing nodes/edges—thus the sampling process can significantly impact the accuracy of relational classification. Bias may result from the size of the sample, the sampling method, or both. Most previous work in relational learning has focused on analyzing a single *input network* and research has considered how to further split the input network into training and testing networks for evaluation (Körner and Wrobel 2006; Macskassy and Provost 2007; Neville, Gallagher, and Eliassi-Rad 2009). However, the fact that the input network is often itself sampled from a larger target network has largely been ignored and there has been little focus on *how* the construction of the input networks may impact the evaluation of relational algorithms.

In this paper, we outline different network sampling methods and systematically study their impact on relational classification performance. We use the simple weighted-vote relational neighbor (wvRN) as our base classifier (Macskassy and Provost 2007). We consider wvRN for two reasons: (1) its performance is primarily due to the relational structure (i.e., the classifier assumes a network with sufficient linkage and homophily), thus, it provides a fair evaluation of the unique structure sampled using the various sampling designs, and (2) it is simple and efficient, which makes it practical for large-scale networks. In our experi-

ments, we show that classification performance can significantly change based on the sampling design used to collect the data. Our results show that both forest fire sampling (FFS, Leskovec *et al.* 2006) and edge sampling with graph induction (ES-i, Ahmed *et al.* 2011) need at least 30% of the larger network to get reasonable approximations of accuracy. However, ES-i maintains a relatively consistent performance across all datasets (compared to other algorithms). Moreover, sampling designs such as node and edge sampling produce very sparse samples that neither match the graph properties nor the classification accuracy, and they need at least 50% of the larger network to obtain reasonable approximations.

Framework

We define a *population* network as a graph $G = (V, E)$, such that $|V|$ is the number of nodes, and $|E|$ is the number of edges (links) in the network. A population (i.e., target) network G is usually a large network that is difficult to completely access and/or collect. Therefore, we consider a sampling algorithm S that procedurally selects a subnetwork $G_s = (V_s, E_s)$ according to a particular sampling design, where $V_s \subset V$ and $E_s \subset E$, such that $|V_s| = \phi \cdot |V|$. We refer to ϕ as the sampling fraction. We then consider a relational classifier R that takes G_s as input. A classifier R uses a portion of G_s with known class labels as the *training set* to learn the model. Then R collectively classifies the remainder of G_s as the *test set* and evaluates the performance based on the accuracy of the predicted class labels.

Our goal in this paper is to evaluate the quality of the sampled graph G_s by comparing the accuracy of a classifier R on G_s to the accuracy of R on G. We consider four different sampling algorithms, node sampling (NS), edge sampling (ES), forest fire sampling (FFS), and edge sampling with graph induction (ES-i). We use the weighted-vote relational neighbor classifier (wvRN) as our base classifier (Macskassy and Provost 2007). We evaluate the sampling algorithms based on: (1) topological graph properties (degree, path length, and clustering coefficient), and (2) classification accuracy (using area under the ROC curve).

Classes of Sampling Designs

Current sampling designs can be broadly classified as node-based, edge-based, and topology-based methods.

Node sampling (NS). In classic node sampling, nodes are chosen independently and uniformly at random from the original graph for inclusion in the sampled graph. For a target fraction ϕ of nodes required, each node is simply sampled with a probability of ϕ. Once the nodes are selected, the sampled graph consists of the *induced subgraph* over the selected nodes, *i.e.*, all edges among the sampled nodes are added to form the sampled graph. Sampled subgraphs produced by node sampling can be further refined using the Metropolis algorithms proposed in (Hubler et al. 2008). The key idea is to replace sampled nodes with other potential nodes that will better match the original degree distribution (or other metrics). Another way of node sampling is to sample the node with probability proportional to its PageRank weight or to its degree. However, the work in (Lee, Kim, and Jeong 2006) shows that it does not accurately capture properties for graphs with power-law degree distributions and the original level of connectivity is not likely to be preserved.

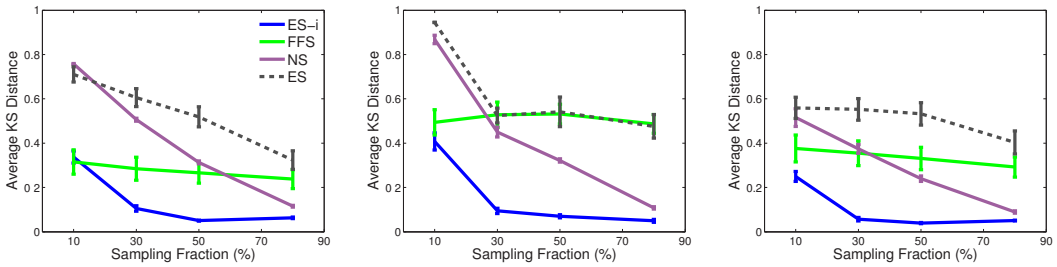
Edge sampling (ES). Edge sampling focuses on the selection of edges rather than nodes to populate the sample. Thus, edge sampling algorithm proceeds by randomly selecting edges, and including both nodes when a particular edge is sampled. Since ES samples edges independently, the resulting sparse subgraphs do not preserve clustering and connectivity. However, sampling based on edges results in a bias towards high degree nodes, thus ES may be able to preserve the connectivity of the graph if it can collect additional edges among the sampled nodes. Ahmed *et al.* proposed a simple algorithm called edge sampling with graph induction (ES-i), which randomly selects edges from the graph (similar to ES), then adds additional edges among the set of sampled nodes (Ahmed, Neville, and Kompella 2011).

Topology-based sampling. Due to the known limitations of NS and ES, researchers have also considered many other topology-based sampling methods. One example is snowball sampling, which selects nodes using breadth-first search from a randomly selected seed node. Snowball sampling accurately maintains the network connectivity within the snowball, however it suffers from a *boundary bias* in that many peripheral nodes (*i.e.*, those sampled on the last round) will be missing a large number of neighbors (Lee, Kim, and Jeong 2006). Random-walk (RW) based sampling methods are another class of topology-based sampling methods. In random walk sampling, we start with a random seed node v. At each iteration of the algorithm, the next hop node u is selected uniformly at random from the neighbors of the current node v. Leskovec *et al.* proposed a Forest Fire Sampling (FFS) method. It starts by picking a node uniformly at random then 'burns' a fraction of its outgoing links with the nodes attached to them. This fraction is a randomly drawn from a geometric distribution with mean $p_f/(1-p_f)$), ($p_f = 0.7$). The process is recursively repeated for each burnt neighbor until no new node is selected, and a new random node is chosen to start the process until we obtain the desired sample size.

Relational Classification

Conventional classification algorithms focus on the problem of identifying the unknown class (e.g., group) to which an entity (e.g., person) belongs. Classification models are learned from a training set of (disjoint) entities, which are assumed to be independent and identically distributed (i.i.d.) and drawn from the underlying population of instances. However, relational learning problems differs from this conventional view in that entities violate the i.i.d. assumption. In relational data, entities (e.g. users in social networks) can exhibit complex dependencies. For example, friends often share similar interests (e.g. political views).

Recently, there have been a great deal of research in relational learning and classification. For example, (Friedman et al. 1999) and (Taskar, Segal, and Koller 2001) are probabilistic relational learning algorithms that search the space for relational attributes and structures of neighbors



(a) Degree (b) Path Length (c) Clustering Coefficient

Figure 1: Average KS Distance across the three datasets.

to improve the classification accuracy. Macskassy proposed a simple relational neighbor classifier (weighted-vote relational neighbor wvRN) that requires no learning and iteratively classifying the entities of a relational network based only on the relational structure (Macskassy and Provost 2007). Macskassy showed that wvRN performs competitively to other relational learning algorithms.

The wvRN is a simple classifier that classifies a particular entity using only the class labels of known related entities. It defines the class membership probability of an entity e belonging to class c as: $P(c|e) = \frac{1}{Z}\sum_{e_j \in D_e} w(e, e_j) * P(c|e_j)$, where D_e is the set of entities that are linked to e, $w(e, e_j)$ is the weight of the link, and $Z = \sum_{e_j \in D_e} w(e, e_j)$.

Experiments

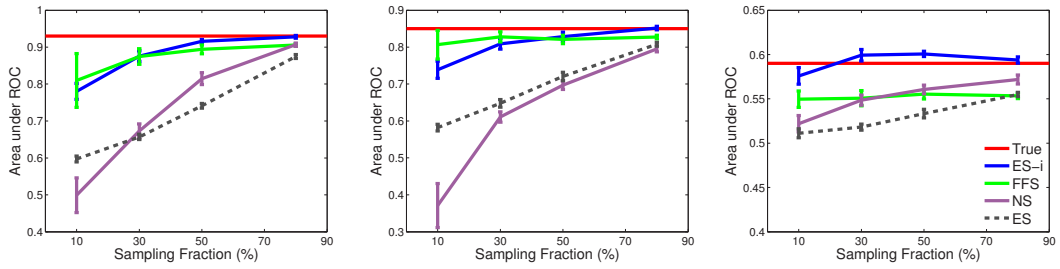
We consider three real networks: two citation networks (CoRA, and Citeseer) with 2708 and 3312 nodes respectively (Sen et al. 2008), and a social media network (Facebook) with 7315 users (Xiang, Neville, and Rogati 2010). While these three datasets are themselves subnetworks of other larger networks, we use them as examples of population (target) networks for evaluation. We collect a sample G_s such that the sample size is between $10\% - 80\%$ of the population network G. For each sample size, we run the sampling algorithms S ten different runs. In each run, we use the sampled network G_s as input to the relational classifier (wvRN) and we vary the proportion of nodes in G_s for which the class labels are initially known (again $10\% - 80\%$) by selecting randomly from the graph, and for each of these settings we use 5-fold cross validation. We repeat the same setup on the population network G. We compare the classification performance on G_s to the classification performance on the population network G.

Preserving Graph Properties. Our evaluation of how different sampling algorithms preserve graph properties is primarily based on a set of topological graph properties (degree, path length, and clustering coefficient) that capture both the local and global structural properties of the graph. These properties were first used by Leskovec *et al.* to evaluate the quality of a sample. We compute the Kolmogorov-Smirnov (KS) statistic to measure the distance between the distribution of the sampled graph G_s and the distribution of the population graph G for the degree, path length, and clustering coefficient. The KS statistic is computed as the maximum vertical distance between the cumulative distribution functions (CDF) of the two distributions, where x represents the range of the random variable and F_1 and F_2 represent the two CDFs: $KS(F_1, F_2) = max_x|F_1(x) - F_2(x)|$.

Figures 1(a)–1(c) show the average KS distance across the three datasets for each of the graph properties. We observe that edge sampling with graph induction (ES-i) outperforms the other methods across all three graph properties. Forest Fire (FFS) sampling performs better than NS and ES when the sample is less than 50%. These results implies that both edge sampling with graph induction (ES-i) and forest fire sampling (FFS) can produce a sampled network G_s with enough linkage and connectivity to closely match the full network. We also observe that edge sampling with graph induction performs consistently well when the sample size is larger than 10% of the population network across the three measures. Further, we observe that both node sampling and edge sampling produce sparse graphs with poor linkage. However, node sampling is better than edge sampling and its performance improves as the sample size increases.

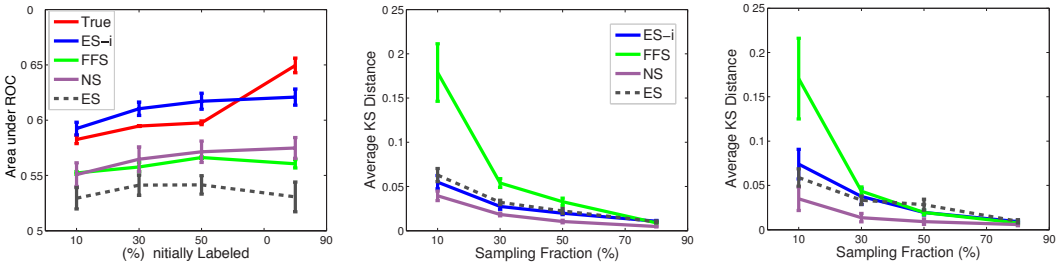
Classification Accuracy. We follow the common methodology used in (Macskassy and Provost 2007) to evaluate classification accuracy. As we mentioned before, we use wvRN classifier as our base classifier. For each sample network G_s, we vary the proportion of initially labeled nodes from $10\% - 80\%$; and we run 5-fold cross validation. We repeat the same methodology for the population network G. In each setting, we calculate the area under the ROC curve (AUC) using the class membership probabilities produced by wvRN. Note that the AUC is calculated for the most prevalent class.

Figures 2(a)–2(c) show the AUC for sample sizes $10\% - 80\%$ when 10% of the class labels are provided to seed the collective classification process. For CoRA and Citeseer, we observe that for sample sizes $10\% - 30\%$, the AUC is underestimated for all the sampling methods. However, edge sampling with graph induction (ES-i) and forest fire sampling (FFS) (unlike ES and NS) produce estimates of AUC that are close to the "True" AUC on G. However, in the Facebook data, it is clear that ES-i performs better than the other sampling methods and converges to the "True" AUC on the larger network. Figure 3(a) shows the AUC on Facebook samples averaged over the sampling sizes $10\% - 80\%$, as the proportion of known class labels is varied between $10\% - 80\%$. We omit the graphs for CoRA and Citeseer due to the limited space, however they show similar behavior.



(b) Citeseer (c) Facebook

] FS, NS, and ES with 10% initially labeled nodes for classification.



(a) Facebook (b) CoRA (c) Citeseer

Figure 3: (a): Accuracy (AUC) For Facebook Network. (b-c): Average KS Distance for Class Distribution

Class Distribution. We finally compare the class distribution of the sampled networks G_s to the class distribution of the population networks. We compute the KS distance between the two class label distributions and plot the results in Figures 3(b)–3(c). We observe that FFS produces a high bias especially for 10% sample size. While this might not affect the performance of FFS sampled subgraphs when wvRN is used as the base classifier (since there is no learning in wvRNs), we conjecture that would have a much larger effect if we learn a model based on the sample and test the model on a hold-out test set. We will study this in future work.

Conclusion and Future Work

In this paper, we investigated the effect of network sampling on estimates of relational classification performance. We outline different network sampling methods and systematically study their impact on relational classification performance. Our results show that the performance of wvRN classifiers can significantly change for different sampling methods. ES-i and FFS need at least to collect 30% of the larger network to get reasonable accurate estimates of performance on the larger (full) target network. However, ES-i maintains a relatively consistent performance across all the datasets compared to other algorithms. We aim to extend this study to other relational classifiers that use attribute information as well as relational information to predict the unknown class labels. Further, we aim to analyze the impact of sampling bias on collective classifiers theoretically.

Acknowledgements

This research is supported by ARO, NSF under contract number(s) W911NF-08-1-0238, IIS-1017898, IIS-0916686, IIS-1149789. The views and conclusions 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 ARO, NSF, or the U.S. Government.

References

Ahmed, N.; Neville, J.; and Kompella, R. 2011. Network sampling via edge-based node selection with graph induction. In *Purdue University, CSD TR #11-016*, 1–10.

Choudhury, M.; Lin, Y.; Sundaram, H.; Candan, K.; Xie, L.; and Kelliher, A. 2010. How does the data sampling strategy impact the discovery of information diffusion in social media. In *ICWSM*, 34–41.

Friedman, N.; Getoor, L.; Koller, D.; and Pfeffer, A. 1999. Learning probabilistic relational models. In *IJCAI*, 1300–1309.

Hubler, C.; Kriegel, H.; Borgwardt, K.; and Ghahramani, Z. 2008. Metropolis algorithms for representative subgraph sampling. In *ICDM*, 283–292.

Körner, C., and Wrobel, S. 2006. Bias-free hypothesis evaluation in multirelational domains. In *PAKDD'*, 668–672.

Lee, S.; Kim, P.; and Jeong, H. 2006. Statistical properties of sampled networks. *Physical Review E* 73(1):016102.

Leskovec, J., and Faloutsos, C. 2006. Sampling from large graphs. In *KDD*, 631–636.

Macskassy, S., and Provost, F. 2007. Classification in networked data: A toolkit and a univariate case study. *JMLR* 8(May):935–983.

Neville, J.; Gallagher, B.; and Eliassi-Rad, T. 2009. Evaluating statistical tests for within-network classifiers of relational data. In *ICDM*.

Sen, P.; Namata, G.; Bilgic, M.; Getoor, L.; Galligher, B.; and Eliassi-Rad, T. 2008. Collective classification in network data. *AI magazine* 29(3):93.

Taskar, B.; Segal, E.; and Koller, D. 2001. Probabilistic classification and clustering in relational data. In *IJCAI*, 870–878.

Xiang, R.; Neville, J.; and Rogati, M. 2010. Modeling relationship strength in online social networks. In *WWW*.