2010

Correcting Bias in Statistical Tests for Network Classifier Evaluation

Jennifer Neville
Purdue University, neville@cs.purdue.edu

Tao Wang
Purdue University, taowang@cs.purdue.edu

Brian Gallagher
Purdue University, bgallagher@llnl.gov

Tina Eliassi-Rad
Purdue University, eliassi@llnl.gov

Report Number:
10-012

Correcting Bias in Statistical Tests for Network Classifier Evaluation

Jennifer Neville, Tao Wang
CS Dept, Purdue University
West Lafayette, IN
{neville, taowang}@cs.purdue.edu

Brian Gallagher, Tina Eliassi-Rad
Lawrence Livermore National Laboratory
Livermore, CA
{bgallagher, eliassi}@llnl.gov

Abstract

It is difficult to directly apply conventional significant tests to compare the performance of network classification models because network data instances are not independent and identically distributed. Recent work [12] has shown that paired t-tests applied to overlapping network samples will result in unacceptably high levels (e.g., up to 50%) of Type I error (i.e., the tests lead to incorrect conclusions that models are different, when they are not). Thus, we need new strategies to accurately evaluate network classifiers. In this paper, we analyze the sources of bias (e.g. dependencies among network data instances) theoretically and propose analytical corrections to standard significant tests to reduce the Type I error rate to more acceptable levels, while maintaining reasonable levels of statistical power to detect true performance differences. We validate the effectiveness of the corrections empirically on both synthetic and real networks.

1 Introduction

A central methodological question in machine learning research is how to compare the empirical performance of two learning algorithms and assess the *significance* of observed performance differences. However, evaluating algorithms becomes more challenging in relational learning where data instances are not independent. In particular, two characteristics of relational learning and collective classification [14] can complicate the application of conventional statistical tests for comparing classification performance: (1) dependence between related instances leads to correlated errors and (2) dependence between training and test set samples leads to correlated test sets.

Most work on evaluating algorithms has mainly focused on data with independent and identically distributed (i.i.d.) instances. Dietterich [4] showed that some statistical tests in widespread use had a high probability of Type I error (i.e., concluding that there is a significant difference between algorithms when there is none). Other work has shown that the choice of training/test sets can lead to underestimation of variance in the cross-validation estimator of the generalization error [11, 2].

There have been some recent works on learning and generalization bounds from non-i.i.d. observations [9, 19, 5, 13, 7, 6, 18, 1, 3, 16, 15, 10, 12]. However, most of them do not address the problems mentioned above of dependent instances and dependent training and test sets. Usunier et al. [17] proposed a new framework to study the generalization properties of classifiers over data which can exhibit a suitable dependency structure. However, their focus was solely on dependent training sets. Neville et. al [12] investigated potential sources of bias empirically and showed that a commonly-used form of evaluation (paired t-test on overlapping network samples) often results in unacceptably high levels of Type I error (e.g., as much as 50%), and proposed a novel sampling method called *network cross-validation* (NCV), which uses overlapping *inference* sets but disjoint *test* sets. This approach results in more acceptable levels of Type I error, but at the expense of decreased statistical power.

In this paper, we present a theoretical study that formalizes the empirical work of [12]. We consider the problem of *within-network* relational learning, where there are dependencies among data instances and the goal is transductive network learning—models are learned on a partially labeled

network and then applied to *collectively* predict the class labels in the remainder of the network (i.e., the unlabeled portion). Within this setting, we demonstrate how the aforementioned network data characteristics contribute to increased Type I error. Our analysis shows that both error correlation and overlapping samples lead to misestimation of the variance that is used in statistical tests. Based on our analysis, we propose an analytical correction to the observed variance which can be used to adjust for the bias and reduce Type I error rates, while maintaining reasonable statistical power. We demonstrate the effectiveness of the correction on both synthetic and real world data, with simulated and real classifiers. Although we evaluate the properties of the corrected significance tests for within-network classification, the findings are also applicable to other learning tasks, which may have overlap but no error correlation.

2 Network classifier evaluation

When comparing the empirical performance of machine learning algorithms, there are two primary methodological choices: First, the *sampling procedure* dictates how the available data is partitioned into training and test sets for estimation of algorithm performance. Second, the *significance test* takes a set of performance measurements (e.g., accuracy) from the various sampling trials and makes a determination as to whether observed differences reflect a true difference in classifier performance or whether it is likely to have occurred by chance alone.

Sampling procedures: Given a fully labeled network S of size m, we consider two sampling procedures to generate training (labeled S_L) and test (unlabeled S_U) sets to evaluate within-network classification algorithms. The first method is *random resampling* (RS). It involves repeated random draws *without replacement* from the sample population (i.e., S) to generate the training/test splits (S_L, S_U); and, therefore, produces overlapping test sets. This method has been used extensively in past work on relational learning algorithms (see the survey in [12] for more detail).

The second method is NCV, a new sampling approach proposed by [12]. NCV samples for k disjoint test sets that will be used for *evaluation*. When the target training set size is less than the size of the $k-1$ merged folds, this will leave a set of unlabeled nodes that are neither in the test set nor the training set. Since these unlabeled instances will likely be connected to nodes in the test set, collective inference is run over the full set of unlabeled nodes (the *inference* set), but model performance is only evaluated on the nodes assigned to the test set. Since NCV only *evaluates* model performance using disjoint test set instances, it eliminates much of the dependency due to overlapping test sets and will not suffer the same level of bias. Indeed NCV has been shown to be more robust to Type I error when compared to conventional resampling [12].

Significance tests: In within-network learning, after a sampling procedure has been chosen to create training/test splits within a network, the models are learned from each training set and the learned models are applied for collective inference over the appropriate test set (i.e., unlabeled portion of the network). The predictions on the test set nodes are evaluated to estimate algorithm performance (e.g., accuracy). This results in a set of performance measurements, one for each training/test split, for each algorithm and a significance test is then used to determine whether the observed performance differences are *significantly* different than would be expected if the performance measures were drawn from the same underlying distribution (i.e., the algorithms perform equivalently).

In this work, we considered both paired and unpaired t-tests for assessments of significance. We are interested in two characteristics of these tests: (1) *Type I error*: the probability of rejecting a *true* null hypothesis, and (2) *Power*: the probability of rejecting a *false* null hypothesis (i.e., 1-Type II error). If a statistical test has elevated levels of Type I error (i.e., greater than the chosen significance level α), that implies that many of the conclusions we draw from the test may be incorrect (e.g., algorithms that appear to be different may in fact have equivalent performance). In contrast, if a statistical test has low statistical power, that implies that legitimate performance differences may not be detected as significant.

3 Theoretical Analysis

Here we show theoretically how error correlation and random sampling (i.e., without replacement) from a network affects the variance of average network classification error. To do this, we model the node-level classification errors as Bernoulli random variables and analytically calculate the observed mean and variance of the average error over repeated samples from the same network. Specifically:

- The input population is a set of m random variables X (i.e., network size=m).
- The population consists of two types of random variables. There are pm random variables of type 1 (i.e., likely errors), which are Bernoulli distributed: $X_i^1 \sim$Bernoulli(q). There are $(1-p)m$ instances of type 0 (i.e., likely correct), which again are Bernoulli distributed: $X_i^0 \sim$Bernoulli($\frac{p}{(1-p)}(1-q)$).
- In the population, there are $|E|$ pairs of "linked" random variables that are correlated. Let ρ be the average correlation between the linked pairs ($(X_i, X_j) \in E$), otherwise we assume that the X_i are independent.
- We sample n random variables $\{X_i\}_{i=1}^n$ without replacement from the population. Since the sampling is without replacement, the random variables X_is are not independent.
- Let $E_k = \frac{1}{n}\sum_{i=1}^n X_i$ be the average value of the r.v.'s in sample k. We are then interested in the mean and variance of the random variable E_k.

The parameters of the Bernoulli variables are designed to keep the overall expected value E_k to be p (i.e., the average error), while allowing individual variation of the random variables across multiple samples: $E(E_k) = E\left(\frac{1}{n}\sum_{i=1}^n X_i\right) = E\left(pX_i^1 + (1-p)X_i^0\right) = pq + (1-p)\frac{p}{(1-p)}(1-q) = p$. Note that if $q = 1$, then the random variables have exactly the same values across all samples (if selected) so this would correspond to sampling from a hypergeometric distribution with pm 1s.

Given this setup, we can now show the effect of correlation and sampling without replacement on the variance of E_k.

Theorem 1. ***Correlated variables increase variance***
Let $\mathbf{X}$ *be an infinite population of Bernoulli(p) random variables. Assume that a sample of* n *variables are drawn randomly from the population. Let* ρ *be the average correlation between the* X_i *that are "linked", where the probability of linkage is* $\frac{|E|}{n(n-1)}$[1]*, and assume that otherwise the* X_i *are independent. Then the variance of* E_k *is* $Var_{corr}(E_k) = \frac{1}{n}p(1-p)\left[1 + \rho\frac{|E|}{n}\right]$.

Proof.

$$\begin{aligned} Var_{corr}(E_k) = Var\left(\frac{1}{n}\sum_{i=1}^n X_i\right) &= \frac{1}{n^2}\left(\sum_{i=1}^n Var(X_i) + \sum_{i=1}^n\sum_{j\neq i}^n Cov(X_i, X_j)\right) \\ &= \frac{1}{n^2}\left(n\cdot p(1-p) + |E|\rho\cdot p(1-p)\right) = \frac{1}{n}p(1-p)\left[1 + \rho\frac{|E|}{n}\right] \end{aligned}$$

□

Thus, as ρ or $|E|$ (i.e., number of correlated pairs) increase, the variance of the average E_k also increases.

Theorem 2. ***Sampling without replacement decreases variance***
Let $\mathbf{X}$ *be a population of* m *Bernoulli random variables as described above, with* pm X^1 *variables (i.e., type 1) and* $(1-p)m$ X^0 *variables (i.e., type 0), where all the* X_i *are independent. Assume that a sample of* n *variables are drawn randomly from the population. Then the variance of* E_k *is* $Var_{rs}(E_k) = \frac{1}{n}p(1-p)\left[1 - \frac{(n-1)}{(m-1)}\left(\frac{q-p}{1-p}\right)^2\right]$.

Proof. First we consider the joint probability of two instances, based on sampling without replacement:

$$\begin{aligned} P(X_i{=}1 \wedge X_j{=}1) = {} & P(X_i \in X^1 \wedge X_i = 1)P(X_j \in X^1 \wedge X_j = 1 | X_i \in X^1) + \\ & P(X_i \in X^1 \wedge X_i = 1)P(X_j \in X^0 \wedge X_j = 1 | X_i \in X^1) + \\ & P(X_i \in X^0 \wedge X_i = 1)P(X_j \in X^1 \wedge X_j = 1 | X_i \in X^0) + \\ & P(X_i \in X^0 \wedge X_i = 1)P(X_j \in X^0 \wedge X_j = 1 | X_i \in X^0) \\ = {} & \left[\left(\frac{pm}{m}q\right)\left(\frac{pm-1}{m-1}q\right)\right] + \left[\left(\frac{pm}{m}q\right)\left(\frac{(1-p)m}{m-1}\frac{p}{1-p}(1-q)\right)\right] + \\ & \left[\left(\frac{(1-p)m}{m}\frac{p}{1-p}(1-q)\right)\left(\frac{pm}{m-1}q\right)\right] + \end{aligned}$$

[1]Note that $n(n-1)$ is the number of possible directed edges in a network of n nodes.

$$\left[\left(\frac{(1-p)m}{m}\frac{p}{1-p}(1-q)\right)\left(\frac{(1-p)m-1}{m-1}\frac{p}{1-p}(1-q)\right)\right]$$
$$=\frac{p}{(m-1)}\left[pm-q^2-\frac{p(1-q)^2}{(1-p)}\right]$$

Now consider the covariance of two instances, based on sampling without replacement:

$$Cov(X_i,X_j)=E(X_iX_j)-E(X_i)E(X_j)=P(X_i=1\wedge X_j=1)-p\cdot p$$
$$=\frac{p}{(m-1)}\left[pm-q^2-\frac{p(1-q)^2}{(1-p)}\right]-p^2\;=\;-\frac{p(1-p)}{(m-1)}\left[\frac{(q-p)^2}{(1-p)^2}\right]$$

With the covariance, we can compute the overall variance based on sampling without replacement:

$$Var_{rs}(E_k)=Var\left(\frac{1}{n}\sum_{i=1}^{n}X_i\right)=\frac{1}{n^2}\left[\sum_{i=1}^{n}Var\left(X_i\right)+\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}Cov(X_i,X_j)\right]$$
$$=\frac{1}{n}\left[p(1-p)-(n-1)\frac{p(1-p)}{(m-1)}\left[\frac{(q-p)^2}{(1-p)^2}\right]\right]\;=\;\frac{1}{n}p(1-p)\left[1-\frac{(n-1)}{(m-1)}\left(\frac{q-p}{1-p}\right)^2\right]$$

□

Note that when $q=p$, the variables correspond to independent Bernoullis across samples and the overall variance reduces to the case when each sample is independent: $Var(E_k)=\frac{1}{n}p(1-p)$. When $q=1$, the random variables have exactly the same value across different samples and the variance corresponds to sampling from a Hypergeometric distribution: $Var(E_k)=\frac{1}{n}p(1-p)\left[\frac{m-n}{m-1}\right]$.

We can extend the results of Theorem 2, to show the joint effect of correlation and sampling without replacement on the variance of E_k.

Theorem 3. ***Variance with correlation and sampling without replacement***
Let $\mathbf{X}$ *be a population of* m *Bernoulli random variables as described above, with* pm X^1 *variables (i.e., type 1) and* $(1-p)m$ X^0 *variables (i.e., type 0). Let* ρ *be the average correlation between the* X_i *that are "linked", where the probability of linkage is* $\frac{|E|}{n(n-1)}$*, and assume otherwise the* X_i *are independent. Assume that a sample of* n *variables are drawn randomly from the population. Let* $c=\sqrt{1-2p+pq}$*. Then the variance of* E_k *is* $Var_{obs}(E_k)=\frac{1}{n}p(1-p)\left[1-\frac{(n-1)}{(m-1)}\left(\frac{q-p}{1-p}\right)^2+\frac{|E|\rho}{n(m-1)}\left(\frac{1-q}{1-p}\right)\left[pmq-q+2mc\sqrt{pq}+mc^2-\frac{c^2}{(1-p)}\right]\right]$.

Proof. To combine the covariance based on error correlation with the covariance based on overlap, we need to determine the effect of the correlation on the conditional probability of a linked instance, i.e., $P(X_j=1|X_i=1,e_{ij}\in E)$. We can derive this from the relationship between correlation and covariance:

$$Cov(X_i,X_j|e_{ij}\in E)=Corr(X_i,X_j|e_{ij})Var(X_i)^{\frac{1}{2}}Var(X_j)^{\frac{1}{2}}$$
$$E(X_iX_j|e_{ij}\in E)-E(X_i)E(X_j)=\rho\cdot Var(X_i)^{\frac{1}{2}}Var(X_j)^{\frac{1}{2}}$$
$$P(X_j|X_i,e_{ij}\in E)=E(X_j)+\frac{\rho\cdot Var(X_i)^{\frac{1}{2}}Var(X_j)^{\frac{1}{2}}}{E(X_i)}$$

We can then enumerate the conditional probabilities for each of the four possible worlds for (X_i,X_j):

$$P(X_j^1|X_i^1)=E(X_j^1)+\frac{\rho Var(X_i^1)^{\frac{1}{2}}Var(X_j^1)^{\frac{1}{2}}}{E(X_i^1)}=q+\rho(1-q)$$
$$P(X_j^0|X_i^1)=E(X_j^0)+\frac{\rho Var(X_i^1)^{\frac{1}{2}}Var(X_j^0)^{\frac{1}{2}}}{E(X_i^1)}=\frac{p(1-q)}{1-p}+\rho\frac{(1-q)}{(1-p)}\sqrt{\frac{p(1-2p+pq)}{q}}$$
$$P(X_j^1|X_i^0)=E(X_j^1)+\frac{\rho Var(X_i^0)^{\frac{1}{2}}Var(X_j^1)^{\frac{1}{2}}}{E(X_i^0)}=q+\rho\sqrt{\frac{q(1-2p+pq)}{p}}$$
$$P(X_j^0|X_i^0)=E(X_j^0)+\frac{\rho Var(X_i^0)^{\frac{1}{2}}Var(X_j^0)^{\frac{1}{2}}}{E(X_i^0)}=\frac{p(1-q)}{1-p}+\rho\left(\frac{1-2p+pq}{1-p}\right)$$

Now we can incorporate these conditional probabilities into the calculation of $P(X_i,X_j)$ and $Cov(X_i,X_j)$,

incorporating both correlation and sampling without replacement. Let $c = \sqrt{1-2p+pq}$, then:

$$\begin{aligned}
P(X_i{=}1 \wedge X_j{=}1){=} & \left[\left(\frac{pm}{m}q\right)\left(\frac{pm-1}{m-1}[q+\frac{|E|}{n(n-1)}\rho(1-q)]\right)\right] + \\
& \left[\left(\frac{pm}{m}q\right)\left(\frac{(1-p)m}{m-1}\left[\frac{p}{1-p}(1-q)+\frac{|E|}{n(n-1)}\rho\frac{(1-q)}{(1-p)}c\sqrt{\frac{p}{q}}\right]\right)\right] + \\
& \left[\left(\frac{(1-p)m}{m}\frac{p(1-q)}{1-p}\right)\left(\frac{pm}{m-1}\left[q+\frac{|E|}{n(n-1)}\rho c\sqrt{\frac{q}{p}}\right]\right)\right] + \\
& \left[\left(\frac{(1-p)m}{m}\frac{p(1-q)}{1-p}\right)\left(\frac{(1-p)m-1}{m-1}\left[\frac{p(1-q)}{1-p}+\frac{|E|}{n(n-1)}\rho\left(\frac{c^2}{1-p}\right)\right]\right)\right] \\
= & \frac{p}{(m-1)}\left[pm-q^2-\frac{p(1-q)^2}{(1-p)}\right]+\frac{|E|}{n(n-1)}\left(\frac{pq(1-q)\rho}{m-1}\right)\left[pm-1+2mc\sqrt{\frac{p}{q}}+\frac{mc^2}{q}-\frac{c^2}{q(1-p)}\right]
\end{aligned}$$

$$\begin{aligned}
Cov(X_i,X_j) &= E(X_iX_j)-E(X_i)E(X_j) = P(X_i=1\wedge X_j=1)-p\cdot p \\
&= \frac{p}{(m-1)}\left[pm-q^2-\frac{p(1-q)^2}{(1-p)}\right]-p^2+\frac{|E|}{n(n-1)}\left(\frac{pq(1-q)\rho}{m-1}\right)\left[pm-1+2mc\sqrt{\frac{p}{q}}+\frac{mc^2}{q}-\frac{c^2}{q(1-p)}\right] \\
&= \frac{p(1-p)}{(m-1)}\left[-\left(\frac{q-p}{1-p}\right)^2+\frac{|E|\rho}{n(n-1)}\left(\frac{1-q}{1-p}\right)\left[pmq-q+2mc\sqrt{pq}+mc^2-\frac{c^2}{(1-p)}\right]\right]
\end{aligned}$$

Now we can compute the overall variance of E_k, with correlation as well as sampling without replacement:

$$\begin{aligned}
Var_{obs}(E_k){=}Var\left(\frac{1}{n}\sum_{i=1}^{n}X_i\right) &= \frac{1}{n^2}\left[\sum_{i=1}^{n}Var\left(X_i\right)+\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}Cov(X_i,X_j)\right] \\
&= \frac{1}{n}\left[p(1-p)-\frac{n(n-1)}{n}\frac{p(1-p)}{(m-1)}\left[\left(\frac{q-p}{1-p}\right)^2-\frac{|E|\rho}{n(n-1)}\left(\frac{1-q}{1-p}\right)\left[pmq-q+2mc\sqrt{pq}+mc^2-\frac{c^2}{(1-p)}\right]\right]\right] \\
&= \frac{1}{n}p(1-p)\left[1-\frac{(n-1)}{(m-1)}\left(\frac{q-p}{1-p}\right)^2+\frac{|E|\rho}{n(m-1)}\left(\frac{1-q}{1-p}\right)\left[pmq-q+2mc\sqrt{pq}+mc^2-\frac{c^2}{(1-p)}\right]\right]
\end{aligned}$$

□

Finally, we can use these results to show these two effects combine together to bias conventional statistical tests for network domains.

Theorem 4. ***Sampling without replacement and error correlation increase Type I error***
Let algorithm A and algorithm B have equal error rates of p on network datasets drawn from the same domain D. Let X_i be the classification error for node i and assume that $X_{i.A}$ and $X_{i.B}$ (the error made by algorithm A and B respectively) are Bernoulli distributed as described above, i.e., with probability p, $X_{i.A/B}$ is of type 1 and with probability $(1-p)$, X_i is of type 0. Let ρ be the average correlation between the X_i, X_j that are linked (i.e., $e_{ij} \in E$) and assume that otherwise the X_i are independent. Assume that a test set of size n is drawn from the network of m nodes.

Let $\mathbf{E^A} = \{E_1^A, E_2^A, ...E_k^A\}$ and $\mathbf{E^B} = \{E_1^B, E_2^B, ...E_k^B\}$ be the set of average test set errors ($E_k = \frac{1}{n}\sum_n X_i$) for $j = [1,k]$ repeated samples drawn from a network of size m from the domain D. Let $c = \sqrt{1-2p+pq}$. Then an unpaired t-test over $\mathbf{E^A}$ and $\mathbf{E^B}$ will underestimate the variance of the null distribution by: $\frac{1}{n}p(1-p)\left[\frac{(n-1)}{(m-1)}\left(\frac{q-p}{1-p}\right)^2+\rho\frac{|E|}{n}\left[1-\frac{1}{(m-1)}\left(\frac{1-q}{1-p}\right)\left[pmq-q+2mc\sqrt{pq}+mc^2-\frac{c^2}{(1-p)}\right]\right]\right]$.

Proof. The unpaired t-test uses the average (i.e., *pooled*) variance of E^A and E^B for the null distribution. Since the error distribution of A and B are equal, the average is equal to the variance of a single algorithm. When the nodes are repeatedly sampled without replacement, we know from Theorem 3 that the observed variance of E_k will be the following: $Var_{obs}(E_k) = \frac{1}{n}p(1-p)\left[1-\frac{(n-1)}{(m-1)}\left(\frac{q-p}{1-p}\right)^2+\frac{|E|\rho}{n(m-1)}\left(\frac{1-q}{1-p}\right)\left[pmq-q+2mc\sqrt{pq}+mc^2-\frac{c^2}{(1-p)}\right]\right]$, where $c = \sqrt{1-2p+pq}$. However, when there is error correlation ρ among the instances in the data, from Theorem 1 we know that the variance of E_k with independent samples is the following:

$Var_{corr}(E_k) = \frac{1}{n}p(1-p)\left[1+\rho\frac{|E|}{n}\right]$. Since the t-test assumes independent samples, the variance of the null distribution should correspond to the variance without repeated sampling $Var_{corr}(E_k)$. If the observed variance $Var_{obs}(E_k)$ is used in the t-test, it will result in an underestimate of Δ:

$$\begin{aligned}
\Delta &= Var_{corr}(E_k) - Var_{obs}(E_k) \\
&= \frac{1}{n}p(1-p)\left[1+\rho\frac{|E|}{n}\right] \\
&\quad - \frac{1}{n}p(1-p)\left[1 - \frac{(n-1)}{(m-1)}\left(\frac{q-p}{1-p}\right)^2 + \frac{|E|\rho}{n(m-1)}\left(\frac{1-q}{1-p}\right)\left[pmq - q + 2mc\sqrt{pq} + mc^2 - \frac{c^2}{(1-p)}\right]\right] \\
&= \frac{1}{n}p(1-p)\left[\frac{(n-1)}{(m-1)}\left(\frac{q-p}{1-p}\right)^2 + \rho\frac{|E|}{n}\left[1 - \frac{1}{(m-1)}\left(\frac{1-q}{1-p}\right)\left[pmq - q + 2mc\sqrt{pq} + mc^2 - \frac{c^2}{(1-p)}\right]\right]\right]
\end{aligned}$$

□

As ρ (the amount of error correlation) or q (the correlation of node error across samples) increases, the amount of underestimation (i.e., Δ) increases. This increases the probability of a Type I error in the following way. For unpaired tests, the t-statistic is: $\hat{t} = \frac{\bar{E^A} - \bar{E^B}}{\sqrt{Var(E_{A/B})}\cdot\sqrt{\frac{2}{k}}}$. where $Var(E_{A/B})$ is the pooled variance. Since $Var_{obs}(E_k) < Var_{corr}(E_k)$, the result will be that $\hat{t}_{obs} > \hat{t}_{corr}$ and thus $P(\hat{t}_{obs}|T) < P(\hat{t}_{corr}|T)$, where T is the appropriate t distribution with $dof = k$. Thus using $Var_{obs}(E_k)$ instead of $Var_{corr}(E_k)$, it is more likely that the null hypothesis will be rejected even when it holds, and as such Type I error will increase.

This effect will impact paired t-tests in a similar way, as the decrease in observed variance of E^A and E^B will also result in an underestimate of the *difference variance* $Var(E^A - E^B)$, which is used instead of the pooled variance.

4 Analytical correction for bias

Based on the theoretical analysis in Section 3, we propose an analytical adjustment to correct for the bias due to repeated sampling without replacement. We would like to remove the effects of resampling, and adjust the observed variance $Var_{obs}(E_k)$ to make it equal to the variance we would expect just due to correlation: $Var_{corr}(E_k) = \frac{1}{n}p(1-p)[1+\rho\frac{|E|}{n}]$. To achieve this, we simply add in the correction factor Δ from Theorem 4 above: $Var_{new}(E_k) = Var_{obs}(E_k) + \Delta = Var_{corr}(E_k)$.

Correction for unpaired t-test: The correction can be used in an unpaired t-test in the following manner. We estimate model error (for each model) in the conventional manner, recording average performance over multiple test sets. After computing the variance of the average performances for a particular model (i.e., $Var_{obs}(E_k)$), we compute the appropriate Δ from above and use it to scale the observed variance. Then the corrected variance $Var_{new}(E_k)$ is used in place of the observed variance in the standard formulation of the unpaired t-test.

Correction for paired t-test: For the paired t-test, we can use the correction to rescale each observed value before computing the differences and variance. The idea is to compute the standardized value with the original variance (Var_{obs}) and then *unstandardize* using the corrected variance (Var_{new}). Let x^A be an observed error value for algorithm A. Let μ^A be the mean (observed) error for algorithm A. Let $\sigma^A_{obs} = (Var^A_{obs})^{\frac{1}{2}}$ be the observed standard deviation of the average performance of algorithm A. Let $\sigma^A_{new} = (Var^A_{new})^{\frac{1}{2}}$ be the corrected standard deviation of algorithm A. Then the adjustment for each measured performance value x^A is the following: $x^A_c = \left[\left(\frac{x^A - \mu^A}{\sigma^A_{obs}}\right)\cdot\sigma^A_{new}\right] + \mu^A$ $= \left(\frac{\sigma^A_{new}}{\sigma^A_{obs}}\right)x^A + \left(1 - \frac{\sigma^A_{new}}{\sigma^A_{obs}}\right)\mu^A$ The same adjustment is then applied to errors for algorithm B, with appropriate mean and variances. Once all the observed errors are adjusted, we can then compute the paired t-test in the standard way.

The correction Δ requires values for the parameters: $n, m, p, q, \rho, |E|$. We can easily calculate $n, m, |E|$ from the properties of the training/test networks used in a particular evaluation. Also, p, q, ρ can be estimated from the training/test network evaluations. For the experiments below, we use the mean zero-one loss for p, the average misclassification across test sets for q, and for ρ we use the ϕ coefficient to measure the correlation of errors for linked instances in the network. In the

following sections we report results for paired tests only. Experiments with unpaired tests yielded qualitatively similar results.

5 Experimental results

To investigate the effectiveness of our proposed correction, RS-C, for significance tests of network classifiers, we conducted experiments with both simulated and real relational classifiers under varying data characteristics, using synthetic data and data from the Internet Movie Database (imdb.com).

We compare the Type I error rates and statistical power of RS, NCV, and RS-C using paired t-tests. In all the experiments, both Type I error rates and statistical power rates were averaged over 500 (simulated) or 50 (synthetic/real) trials. For a given dataset, in each trial we *sample* from the network, either using random sampling (RS) or using network cross-validation (NCV), to create 10 training/test splits (subnetworks). Then we learn classifiers (using two competing algorithms A and B) on the training subnetwork and apply the learned classifiers on its corresponding test subnetwork to measure its performance (e.g. average error rate). To compare performance, we conducted significant tests to either accept or reject the null hypothesis that the performance of algorithm A and B are equivalent. When the experiments are designed so that two learned classifiers have equivalent error rates, any rejection of the null hypothesis corresponds to a Type I error (i.e., false positive identification of a difference when it does not exist). However, when the two classifiers perform differently, a rejection of the null hypothesis represents the *statistical power* of the test (i.e., true positive identification of a difference when it exists). We calculate and report the proportion of trials for which the null hypothesis was rejected (i.e. Type I error or power in its corresponding experimental setup).

5.1 Experiments with simulated classifiers

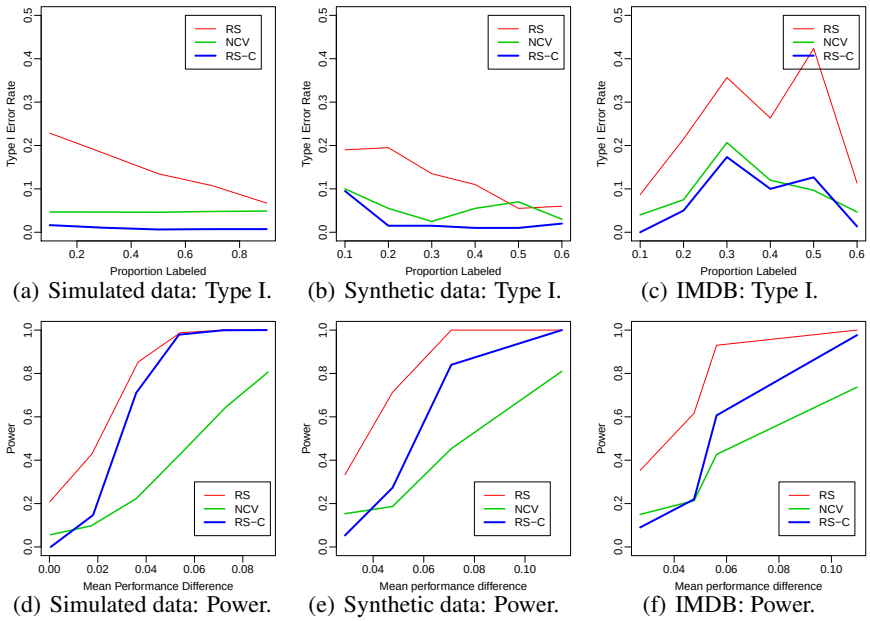
Here we replicate the experiments of [12] to analyze test characteristics with simulated classifiers. We simulate the correlated errors observed in real network classifiers by dividing data instances into disjoint groups and assigning "classification errors" such that errors are correlated among the instances within a group. We simulate two group-based classifiers A and B, ensuring that A and B have the same error rate (p) while still making different kinds of errors (i.e., A misclassifies different groups from B). Each trial utilizes datasets with default parameters $m = 300$, $p = 0.1$, and $q = 0.9$.

Figure 1(a) shows the effects of varying the proportion of labeled data for training. In these experiments, algorithms A and B have equal error rates of $p = 0.1$ so rejecting the null hypothesis corresponds to a Type I error. For RS, the Type I error rate increases as $propLabeled$ decreases. This result is expected since the degree of overlap between test sets increases as the number of unlabeled instances increases. Since NCV disallows overlapping test sets by design, it is not susceptible to this problem, achieving uniformly low Type I error rates. The corrected test, RS-C, exhibits a further reduction in type I error over NCV since it accounts for error correlation as well as test set overlap.

Figure 1(d) shows the statistical power of the tests when the difference in error rates between A and B is varied ($propLabeled$ = 0.3). In this case, since the algorithm error rates are different, a rejection of the null hypothesis corresponds to a true positive. RS has the highest statistical power overall, but when its high Type I error rates are taken into account, RS has little practical utility. RS-C, on the other hand, is able to maintain low Type I error while achieving a reasonable amount of statistical power. For example when there is a 4% difference in error rates, RS-C will be able to detect it 80% of the time. NCV has substantially lower statistical power—it will only be able to detect a 4% difference 20% of the time.

5.2 Experiments with real classifiers

To further investigate RS-C, we compare the collective classification models used in [12]: weighted-vote relational neighbor (wvRN) [8] and network-only Bayes classifier (nBC) [8]. For both models, we use relaxation labeling for collective inference. To estimate Type I error, we handicap the *better* performing model (wvRN) until the performance difference between the models is equivalent (i.e., ≤ 0.005). This is achieved by randomly selecting $b\%$ of the wvRN's predictions and perturbing those probabilities toward the opposite class. We searched for a value of b that resulted in a performance difference of ≤ 0.005 between the two models on a separate set of *calibration* networks. To estimate statistical power, we handicap the *worse* performing model (nBC) to increase the perfor-



(a) Simulated data: Type I. (b) Synthetic data: Type I. (c) IMDB: Type I.

(d) Simulated data: Power. (e) Synthetic data: Power. (f) IMDB: Power.

Figure 1: Type I error and power experiments on simulated and real classifiers on various data.

mance difference between the two models. We used $b = [0.025, 0.075, 0.15, 0.3]$ and measured the resulting performance difference, which is reported in Figure 1(e) and 1(f).

Results on synthetic data: In this set of experiments, we use synthetic datasets as described in [12]. The generated networks have size $m = 300$ with average autocorrelation$= 0.40$ and class prior $P(+) = 0.70$. The data is designed so that wvRN and nBC will make *different* classification errors on different nodes. To measure Type I error rates and power of the statistical tests, we used four synthetic networks (in addition a set of 50 calibration networks).

Figure 1(b) plots the Type I error rates for three statistical tests. Notably, the level of Type I error exhibited by RS-C is significantly lower than that of RS ($> 50\%$ reduction in error). RS-C Type I error is also slightly lower than that of NCV. Figure 1(e) plots the power of each statistical test on networks with 30% labeled nodes. Here we observe, that RS-C again achieves much higher power than NCV. This is due to its use of larger test sets sizes—after correcting for overlap, the *effective* sample size is still larger than the disjoint sets used in NCV.

Results on real data: In the second set of experiments, we use data from the Internet Movie Database (IMDB). We collected a sample of 1,543 movies released in the United States between 2003 and 2007, with their associated producers and studios. To create six disjoint network samples, using stratified sampling by studios. Within each partition, we created links among movies with a common producer. The resulting networks have an average size of 257 nodes and the movies have average degree of 16. The classification task is to predict whether the movie will make more than $\$60mil$ in total box office receipts. The average autocorrelation in these networks is 0.35.

Figure 1(c) and 1(f) show the Type I error and statistical power for each test respectively. As expected, the statistical tests exhibit similar behavior on the real network data as on the synthetic data. RS-C has Type I error rates comparable to NCV and significantly lower than RS. Again RS-C has much higher power than NCV for detecting the algorithm differences in real network data.

6 Conclusion

We investigated two biases present in statistical tests for within-network classification algorithms: (1) correlated errors among related instances and (2) overlap between samples. These biases increase the Type I error to unacceptably high-levels. To account for these biases, we introduced analytical corrections to the empirical estimates of variance. Experiments on real and synthetic data, using real and simulated classifiers demonstrate that our corrections reduce the Type I error while maintaining good statistical power. Compared to the network cross-validation, our corrections result in a significant increase in statistical power.

References

[1] Amit Dhurandhar amd Alin Dobra. Study of classification models and model selection measures based on moment analysis. In *NIPS'08 Workshop on New Challenges in Theoretical Machine Learning: Learning with Data-dependent Concept Spaces*, 2008.

[2] Y. Bengio and Y. Grandvalet. No unbiased estimator of the variance of k-fold cross-validation. *JMLR*, 5:1089–1105, 2004.

[3] Amit Dhurandhar and Alin Dobra. Probabilistic characterization of random decision trees. *JMLR*, pages 2321–2348, 2008.

[4] T. Dietterich. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, 10:1895–1923, 1998.

[5] Murat Dundar, Balaji Krishnapuram, Jinbo Bi, and R. Bharat Rao. Learning classifiers when the training data is not iid. In *IJCAI*, 2007.

[6] Mark Herbster, Guy Lever, and Massimiliano Pontil. Exploiting cluster-structure to predict the labeling of a graph. In *NIPS'08 Workshop on New Challenges in Theoretical Machine Learning: Learning with Data-dependent Concept Spaces*, 2008.

[7] Mark Herbster, Guy Lever, and Massimiliano Pontil. Online prediction on large diameter graphs. In *NIPS'08 Workshop on New Challenges in Theoretical Machine Learning: Learning with Data-dependent Concept Spaces*, 2008.

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

[9] Mehryar Mohri and Afshin Rostamizadeh. Stability bounds for non-i.i.d. processes. In *NIPS*, 2007.

[10] Mehryar Mohri and Afshin Rostamizadeh. Stability bounds for stationary ϕ-mixing and β-mixing processes. *JMLR*, 11:789–814, 2010.

[11] C. Nadeau and Y. Bengio. Inference for the generalization error. *MLJ*, 52(3):239–281, 2003.

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

[13] Liva Ralaivola, Marie Szafranski, and Guillaume Stempfel. Chromatic pac-bayes bounds for non-iid data. In *NIPS'08 Workshop on New Challenges in Theoretical Machine Learning: Learning with Data-dependent Concept Spaces*, 2008.

[14] Prithviraj Sen, Galileo Namata, Mustafa Bilgic, Lise Getoor, Brian Gallagher, and Tina Eliassi-Rad. Collective classification in network data. *AI Magazine*, 29(3):93–106, 2008.

[15] Ingo Steinwart and Andreas Christmann. Fast learning from non-i.i.d. observations. In *NIPS*, 2009.

[16] Ben Taskar. Structured prediction cascades. In *NIPS'09 Workshop on Approximate Learning of Large Scale Graphical Models: Theory and Applications*, 2009.

[17] Nicolas Usunier, Massih-Reza Amini, and Patrick Gallinari. Generalization error bounds for classifiers trained with interdependent data. In *NIPS*, 2005.

[18] Fabio Vitale, Nicolò Cesa-Bianchi, and Claudio Gentile. Online graph predictionwith random trees. In *NIPS'08 Workshop on New Challenges in Theoretical Machine Learning: Learning with Data-dependent Concept Spaces*, 2008.

[19] Xinhua Zhang, Le Song, Arthur Gretton, and Alex Smola. Kernel measures of independence for non-iid. In *NIPS*, 2009.