Validating Collective Classification Using Cohorts

Eric Bax[*] **James Li**[†] **Abdullah Sonmez**[‡] **and Zehra Cataltepe**[§]

Abstract

Many networks grow by adding successive cohorts – layers of nodes. Often, the nodes in each layer are selected independently of each other, but from a distribution that can depend on which nodes were selected for previous cohorts. For example, successive waves of friends invite their friends to join social networks. We present error bounds for collective classification over these networks.

1 Introduction

Networks play fundamental and increasingly important roles in our lives. Networks of gene activation direct the growth of our bodies; networks of chemical and electrical signals control the processes that keep us alive. Networks influence how we understand the natural world, from food webs in ecology to the cycles of energy and chemistry that drive weather and environmental change. We build networks for transportation and power transmission. Social and professional networks, internet-enabled and otherwise, influence how we live and work.

Network analysis has a rich history in a variety of fields [1]. In the social sciences, the work spans from Zachary's karate club study [2] in 1977 to studies of online communication like Gopal's AIDS blog network study [3] and insights into the study of social networks [4]. General analysis of network data started with a statistical approach based on random graphs [5]. Work on network generation processes produced results about small-world graphs [6, 7] and graphs with other interesting properties. In recent years, there has been growing interest in statistical approaches to deduce properties of the graph generation process from graphs themselves and their histories.

Networks are about relationships, connections, and dependencies. This makes it a challenge to apply classical machine learning approaches directly to validation of collective classification, because those methods rely on examples being drawn i.i.d. from an underlying distribution. But graphs do not usually grow by selecting all nodes i.i.d. from an underlying distribution. Instead they often grow by accretion – with new nodes drawn from a distribution that depends on the nodes already in the network. In this work, we develop probably approximately correct (PAC) error bounds that take advantage of information about how the network grew. (For more on PAC error bounds, refer to [8, 9, 10, 11].) The bounds apply to the transductive setting [12], where we have the unlabeled nodes we wish to classify in the network.

Collective classification is a very active field. Some work of note includes an overview [13], a comparison of some different approaches [14], a general toolkit [15], and a book that covers several aspects in depth [16]. Recent work on validation of collective classifiers includes several approaches to the challenge that nodes are not drawn i.i.d., including methods based on partitioning into i.i.d. sub-networks [17], relying on multiple networks [18] or on locality [19] – that distant nodes usually have only small effects in collective classification – and a technique based on worst likely assignments [20] that is effective for small problems but does not scale to very large ones.

[*] ebax@yahoo-inc.com, Yahoo Labs, 3333 Empire Blvd, Burbank CA, 91504

[†] jamesyili@gmail.com, Statistics Department, Cornell University, Ithaca, NY

[‡] abdullah.sonmez@gmail.com, Istanbul technical University

[§] cataltepe@itu.edu.tr, Istanbul Technical University

Each validation technique has its strengths and limitations; the one introduced here is not intended to replace the others. The new technique does not require that sub-networks be drawn i.i.d., does not require multiple in-sample networks for statistical bounds or impose conditions on the form of the classification algorithm to ensure locality, and does scale without sacrificing bound strength. The new technique requires us to identify sets of in-sample nodes that are just as likely to have been left unlabeled as the set of nodes we wish to classify – we call the union of these sets a *cohort*. In addition, the new technique also requires that the process that selected which nodes to label operated independently over nodes in the cohort. We discuss these requirements in more detail later.

2 Results

In this section, we present PAC bounds for network classifiers based on cohorts. First we prove a PAC bound based on probabilities over different subsets of a cohort being selected for labeling (Theorem 1). Then we present a bound based on probabilities over randomly generated cohorts, similar to the usual setting for PAC bounds (Corollary 5).

Our approach is to validate a network classifier in two steps. First, we withhold from the network a portion of the labeled nodes that are in the same cohort as the nodes we wish to label. We perform collective classification on the withheld nodes and use their labels to evaluate the accuracy of collective classification for the network without the withheld nodes. Second, we evaluate the rate of disagreement between collective classification with and without the withheld nodes. This gives a bound on the difference in error rates between the validated withheld classifier and the full classifier based on all nodes.

Let F be the full set of nodes in a network, with some nodes having known labels and others unknown. Define a cohort to be a subset of F for which whether labels are known or unknown was determined at random, independently and with the same probability for each node. Let $C \subset F$ be a cohort. Define validation set V to be the nodes in C with known labels, and define working set W to be the nodes in C with unknown labels.

Our goal is to bound the error rate of collective classification over W. To do this, we will bound the error rate of classification based on nodes in $F - (V \cup W)$. Then we will bound the rate of disagreement between classification based on $F - (V \cup W)$ and classification based on the full network.

Let g^* be the output of classification based on the full network. The classification method may use any available information: the graph structure of the network, node inputs, link inputs, and any known node labels. Let $y \in \{0, 1\}$ be the label for a node. Then our goal is to bound

$$Pr_W\{g^* \neq y\}, \tag{1}$$

where the probability is over the specific nodes in W (not over random draws of W).

Let $g_\emptyset$ be the output of holdout classification of a node $z \in V \cup W$: classification based on $F - (V \cup W) \cup \{z\}$. The classification method may not use information about nodes in $(V \cup W) - \{z\}$ or links with those nodes, but it may use the other available network information.

Theorem 1 *Let $v = |V|$ and $w = |W|$. Let d be the number of disagreements over nodes in W between holdout classification and classification based on the full network:*

$$d = Pr_W\{g^* \neq g_\emptyset\}. \tag{2}$$

Let ϵ be the difference between the (unknown) error rate of the full classifier over unlabeled nodes and the holdout classification error rate over validation set nodes:

$$\epsilon = Pr_W\{g^* \neq y\} - Pr_V\{g_\emptyset \neq y\}, \tag{3}$$

where the probabilities are over the specific nodes in W and V, respectively. Then for $\delta > 0$,

$$Pr_{\{V|W\}}\{\epsilon > (1 + \frac{v}{w})\sqrt{\frac{-\ln \delta}{2v}} + \frac{d}{w}\} \leq \delta, \tag{4}$$

where the probability is over partitions of C into a size $|V|$ validation set and a size $|W|$ working set.

To prove the theorem, we will use two lemmas. The first states that the error rate of full-network classification is at most the sum of the error rate of held-out classification and the rate of disagreement between the full-network and held-out classification:

Lemma 2

$$Pr_W\{g^* \neq y\} \leq Pr_W\{g_\emptyset \neq y\} + Pr_W\{g_\emptyset \neq g^*\}. \tag{5}$$

Proof of Lemma 2. Note that

$$Pr_W\{g^* \neq y\} = Pr_W\{g^* \neq y \wedge g_\emptyset = g^*\} + Pr_W\{g^* \neq y \wedge g_\emptyset \neq g^*\}. \tag{6}$$

Since

$$(g^* \neq y \wedge g_\emptyset = g^*) \Rightarrow (g_\emptyset \neq y) \tag{7}$$

and

$$(g^* \neq y \wedge g_\emptyset \neq g^*) \Rightarrow (g_\emptyset \neq g^*), \tag{8}$$

$$Pr_W\{g^* \neq y\} \leq Pr_W\{g_\emptyset \neq y\} + Pr_W\{g_\emptyset \neq g^*\}. \square \tag{9}$$

The second probability on the RHS is the rate of disagreement between $g_\emptyset$ and g^* over W, which we can compute directly. The first probability on the RHS is the error rate of $g_\emptyset$ over W, which we can bound using V, according to the next lemma:

Lemma 3 *Recall that $v = |V|$ and $w = |W|$. Let $r_W = Pr_W\{g_\emptyset \neq y\}$ and $r_V = Pr_V\{g_\emptyset \neq y\}$. Then for $\delta > 0$,*

$$Pr\{r_W > r_V + (1 + \frac{v}{w})\sqrt{\frac{-\ln \delta}{2v}}\} \leq \delta. \tag{10}$$

Proof of Lemma 3. Let $r_{V \cup W} = Pr_{V \cup W}\{g_\emptyset \neq y\}$, i.e., the error rate of $g_\emptyset$ over the cohort. Let i be the number of classification errors by $g_\emptyset$ over V. Select a bound failure probability δ. Let j^* be the minimum number of classification errors by $g_\emptyset$ over W needed to make the probability of observing at most i errors over V be δ or less:

$$j^* = \min\{j : Pr_{\{V|W\}}\{r_V \leq \frac{i}{v} | r_{V \cup W} = \frac{i+j}{v+w}\} \leq \delta\}. \tag{11}$$

Since the nodes in $V \cup W$ are in the same cohort, each partition of $V \cup W$ into a size $|V|$ validation set and a size $|W|$ working set is equally likely. So

$$Pr_{\{V|W\}}\{r_W > \frac{j^*}{w}\} \leq \delta. \tag{12}$$

Now consider how to compute j^*. If $r_W = \frac{i+j}{v+w}$, then there are $i + j$ errors among the $v + w$ nodes in $V \cup W$. The probability of drawing i or fewer of the errors while drawing v nodes without replacement is the tail of a hypergeometric distribution:

$$Pr_{\{V|W\}}\{r_W \leq \frac{i}{v} | r_{V \cup W} = \frac{i+j}{v+w}\} = \sum_{k=0}^{i} \frac{\binom{i+j}{k}\binom{(v+w)-(i+j)}{v-k}}{\binom{v+w}{v}}. \tag{13}$$

Apply a tail bound from Chvátal [21] and Hoeffding [22]:

$$Pr_{\{V|W\}}\{r_V \leq \frac{i}{v} | r_{V \cup W} = \frac{i+j}{v+w}\} \leq e^{-2\epsilon^2 v}, \tag{14}$$

where

$$\epsilon = \frac{i+j}{v+w} - \frac{i}{v}. \tag{15}$$

To use this bound, set $\delta = e^{-2\epsilon^2 v}$ and solve for $\frac{j}{w}$:

$$\frac{j}{w} = \frac{i}{v} + (1 + \frac{v}{w})\sqrt{\frac{-\ln \delta}{2v}}. \tag{16}$$

Substitute into Inequality 12 to get a bound:

$$Pr_{\{V|W\}}\{r_W > \frac{i}{v} + (1 + \frac{v}{w})\sqrt{\frac{-\ln \delta}{2v}}\} \leq \delta. \square \tag{17}$$

Proof of Theorem 1. Recall that

$$\epsilon = Pr_W\{g^* \neq y\} - Pr_V\{g_\emptyset \neq y\}. \tag{18}$$

By Lemma 2,

$$\epsilon \leq Pr_W\{g_\emptyset \neq y\} + Pr_W\{g_\emptyset \neq g^*\} - Pr_V\{g_\emptyset \neq y\} \tag{19}$$

$$= (Pr_W\{g_\emptyset \neq y\} - Pr_V\{g_\emptyset \neq y\}) + \frac{d}{w}. \tag{20}$$

Now apply Lemma 3 to $Pr_W\{g_\emptyset \neq y\} - Pr_V\{g_\emptyset \neq y\}$, recalling that $r_W = Pr_W\{g_\emptyset \neq y\}$ and $r_V = Pr_V\{g_\emptyset \neq y\}$. $\square$

We can easily extend Theorem 1 to cover cases where V and W result from a random partition of $V \cup W$:

Corollary 4 *If we define a cohort to be a set $C = V \cup W$ such that partition $C \to V|W$ was selected uniformly at random among partitions of C into a size $|V|$ and size $|V|$ subsets of C, then Theorem 1 still holds.*

Proof of Corollary 4. We used the definition of a cohort in the logic before Inequality 12. That logic holds directly if $V|W$ is the result of a random partition of C into size $|V|$ and size $|W|$ subsets. $\square$

In the following corollary, we extend Theorem 1 to produce a PAC error bound based on random i.i.d. draws of validation and working nodes to form a cohort. This is more like the usual setting for PAC bounds in machine learning, where training and working examples are assumed to be drawn i.i.d., except in this case only cohort nodes are assumed to be drawn i.i.d., and they can be drawn from a distribution that depends on other nodes in the network.

Corollary 5 *Assume each node in V, including any node data, the node label, and any links to nodes in $F - (V \cup W)$, was drawn i.i.d. from a joint input-label-neighbor distribution, which may depend on nodes in $F - (V \cup W)$. Assume the same for the nodes in W. Then for $\delta > 0$,*

$$Pr\{\epsilon > (1 + \frac{v}{w})\sqrt{\frac{-\ln \delta}{2v}} + \frac{d}{w}\} \leq \delta, \tag{21}$$

where the probability is over random draws of $|V|$ nodes to form V and $|W|$ nodes to form W.

Proof of Corollary 5. Let $f(V \cup W)$ be the pdf of each $V \cup W$, and let $d(V \cup W)$ be an infinitesimal around $V \cup W$. Then

$$Pr\{\epsilon > (1 + \frac{v}{w})\sqrt{\frac{-\ln \delta}{2v}} + \frac{d}{w}\} \tag{22}$$

$$= \int_{V \cup W} Pr_{\{V|W\}}\{\epsilon > (1 + \frac{v}{w})\sqrt{\frac{-\ln \delta}{2v}} + \frac{d}{w}|V \cup W\} f(V \cup W) d(V \cup W). \tag{23}$$

Since each node in $V \cup W$ is drawn i.i.d., each partition of each $V \cup W$ into a size $|V|$ validation set and a size $|W|$ working set is equally likely. So we can apply Corollary 4 to the probability conditioned on each $V \cup W$:

$$\leq \int_{V \cup W} \delta f(V \cup W) d(V \cup W) = \delta \int_{V \cup W} f(V \cup W) d(V \cup W) = \delta. \square \tag{24}$$

3 Analysis, Application Requirements, and Extensions

In this section, we analyze the error bound, outline requirements to apply it, and discuss methods to meet the requirements. To analyze the error bound from Theorem 1 and Corollary 5, note that the range of the bound:

$$(1+\frac{v}{w})\sqrt{\frac{-\ln\delta}{2v}} = \sqrt{\frac{-\ln\delta}{2v}} + \sqrt{\frac{v}{w}}\sqrt{\frac{-\ln\delta}{2w}}. \tag{25}$$

The first term on the RHS is the standard Hoeffding bound [22] for the difference between expected error rate over random validation sets drawn i.i.d. according to the cohort distribution and the empirical error rate on the specific validation set in our network. The second term accounts for the difference between expected error rate over random working sets and the empirical error rate on our specific working set. (It goes to zero as w increases.)

The bound range is minimized with respect to v when $v = w$. (Take the derivative of the RHS with respect to v, set to zero, and solve.) However, this is just an artifact of using the exponential form of Chvátal's bound. To improve the bound as v increases beyond w, bound the RHS of Equation 13 using Inequality 1 from Chvatal [21], which is stronger than the exponential inequality. Alternatively, compute the RHS of Equation 13 directly, using the built-in hypergeometric function in R or using BigDecimal in Java, or compute a close approximation by applying Loader's method [23].

To apply the bound from Theorem 1, we need a process to identify a cohort that includes the working examples. The random partitioning argument in the proof is valid if the process would identify the same cohort regardless of which subset of it was the working set. For example, identifying all nodes that joined the network during the same weeks as the working nodes would meet this requirement. The bound also requires that revelation of node labels is independent and with identical probabilities over the cohort nodes. In some applications, we can design a process to make this condition hold. For example, if nodes represent people, then we can select cohort nodes at random to be validation nodes and elicit labels from the corresponding people if they have not supplied them already. (If we first select a sample size and then select that number of people at random for the validation set, then we can apply Corollary 4 in place of Theorem 1.)

In these applications, we cannot ignore missing or incorrect labels. For missing labels, we can sometimes follow up with extra efforts to elicit labels; then, for validation nodes still unlabeled, we must assume worst-case validation error rates (i.e. that they are incorrectly classified). Incorrect labels can be more challenging. If nodes represent people, then we can use the technique introduced by Coffman et al. [24] with a sample to estimate the rate of incorrect labeling, then assume worst-case validation error rates for incorrect labels. (For more on the challenge of eliciting labels from people, refer to [25].)

In some applications, the working examples will not all belong to the same cohort. In these cases, we can produce separate error bounds for the working nodes in different cohorts and average those bounds to get a single bound for the whole working set. (For some background on averaging bounds, refer to [26].) A challenge for the future is to understand how to optimally weight bound parameters, or take weighted averages of bound ranges, to optimize bounds for these cases.

Finally, the techniques developed in this paper can be applied beyond the transductive setting, in settings where the goal is to bound error rate of the classifier based on nodes in the network over nodes that have not yet been added to the network. Call these nodes the test nodes. In these cases, the challenge is to estimate the rate of disagreement between the held-out classifier and the full classifier over the test nodes. To do this, we can use labeled or unlabeled nodes in the network that are from the same cohort as the test nodes. Alternatively, if we know the distribution of the test nodes (but not necessarily their labels), then we can generate a random sample to estimate the rate of disagreement. If we use the same nodes for validation and estimating the rate of disagreement, then we need to use a sum bound on the union of probabilities of mis-estimating error rate for the held-out classifier and of mis-estimating the rate of disagreement between the held-out classifier and the classifier based on all examples.

References

[1] E. D. Kolaczyk. *Statistical Analysis of Network Data*. Springer, 2010.

[2] W. Zachary. An information flow model for conflict and fission in small groups. *Journal of Anthropological Research*, 33(4):452–473, 1977.

[3] S. Gopal. The evolving social geography of blogs. In H. Miller, editor, *Societies and Cities in the Age of Instant Access*, pages 275–294. Springer, 2007.

[4] C. R. Shalizi and A. C. Thomas. Homophily and contagion are generically confounded in observational social network studies. *Sociological Methods and Research*, 40(2):211–239, 2011.

[5] B. Bollobas. *Random Graphs, Second Edition*. Cambridge University Press, 2001.

[6] D. Watts and S. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442, 1998.

[7] D. Watts. *Six Degrees: The Science of a Connected Age*. Norton & Company, 2003.

[8] L.G. Valiant. A theory of the learnable. *Commun. ACM*, 27(11):1134–1142, 1984.

[9] L. Devroye, L. Gyorfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer-Verlag, New York, 1996.

[10] J. Langford. Tutorial on practical prediction theory for classification. *Journal of Machine Learning Research*, 6:273–306, 2005.

[11] E. Bax and A. Callejas. An error bound based on a worst likely assignment. *Journal of Machine Learning Research*, 9:581–613, 2008.

[12] V. Vapnik. *Statistical Learning Theory*. John Wiley & Sons, 1998.

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

[14] P. Sen and L. Getoor. Empirical comparison of approximate inference algorithms for networked data. In *ICML workshop on Open Problems in Statistical Relational Learning (SRL2006)*, 2006.

[15] Sofus A. Macskassy and Foster Provost. Classification in networked data: A toolkit and a univariate case study. *Journal of Machine Learning Research*, 8:935–983, May 2007.

[16] Lise Getoor and Ben Taskar. *Introduction to Statistical Relational Learning*. The MIT Press, 2007.

[17] Amit Dhurandhar and Alin Dobra. Distribution-free bounds for relational classification.

[18] Ben London, Bert Huang, and Lise Getoor. Improved generalization bounds for large-scale structured prediction. In *NIPS Workshop on Algorithmic and Statistical Approaches for Large Social Networks*, 2012.

[19] B. London, B. Huang, B. Taskar, and L. Getoor. Collective stability in structured prediction: generalization from one example. *Proceedings of the 30th International Conference on Machine Learning*, 2013.

[20] J. Li, A. Sonmez, Z. Cataltepe, and E. Bax. Validation of network classifiers. *Structural, Syntactic, and Statistical Pattern Recognition Lecture Notes in Computer Science*, 7626:448–457, 2012.

[21] Chvátal. The tail of the hypergeometric distribution. *Discrete Mathematics*, 25(3):285–287, 1979.

[22] W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

[23] Catherine Loader. Fast and accurate computation of binomial probabilities. 2000.

[24] K. B. Coffman, L. C. Coffman, and K. M. Marzilli Ericson. The size of the lgbt population and the magnitude of anti-gay sentiment are substantially underestimated. *NBER Working Paper No. 19508*, 2013.

[25] A. Blume, E. K. Lai, and W. Lim. Eliciting private information with noise: the case of randomized response. *Working Paper*, 2013.

[26] E. Bax. Validation of average error rate over classifiers. *Pattern Recognition Letters*, pages 127–132, 1998.