07/27/2011

cosmetic
abstract
motivation

Bayes Optimal Shuffled Graph Classification: Applications in Statistical Connectomics

Joshua T. Vogelstein and Carey E. Priebe

Abstract—Graph classification algorithms often do not incorporate vertex label information in their classifiers. In this work, we investigate the extent to which discarding vertex labels can hinder classification performance, and for which random graph models it would be expected to matter. Via theory we demonstrate a collection of results. Specifically, if one "shuffles" the graphs prior to classification, the vertex label information is irretrievably lost, which can degrade misclassification performance (and does whenever the vertex labels have class-conditional signal). not quite (Thm 2 disc.) Thus, while one cannot hope to recover the labels, trying to recover the labels actually X results in a consistent estimate of the optimal graph invariant. ? This approach therefore solves the question of "which invariant to use" for any graph classification problem, at least asymptotically. Via simulation we demonstrate that a finite (and small) number of training samples can be sufficient to achieve this bound. Finally, we apply this approach to a "connectome" classification problem (a connectome is the complete set of connections within a brain). Unshuffling the graphs indeed improves performance, although not over ? the best performance achievable composing a number of graph invariant and machine learning tools. Thus, given any unlabeled graph classification problem, the relative performance of an unshuffling approach might be difficult to predict with small sample sizes. ?

Index Terms—statistical inference, graph theory, network theory, structural pattern recognition, connectome.

1 Introduction

THIS work addresses graph classification in the presence of vertex label shuffling. A (labeled) graph $G = (\mathcal{V}, \mathcal{E})$ consists of a vertex set, $\mathcal{V} = [n]$, where $n < \infty$ is the number of vertices and $[n] = \{1, \ldots, n\}$, and an edge set $\mathcal{E} \subseteq \binom{[n]}{2}$. Vertex labels may or may not be observed. In the latter case, vertex v in one graph cannot be assumed to correspond to vertex v in another graph.

MOTIVATION

2 Graph Classification Models

2.1 A labeled graph classification model

Let $\mathbb{G}\colon \Omega \to \mathcal{G}_n$ be a graph-valued random variable taking values $G \in \mathcal{G}_n$, where $\mathcal{G}_n$ is the set of graphs on n vertices, and $|\mathcal{G}_n| = 2^{\binom{n}{2}} = d_n$. Let Y be a categorical random variable, $Y\colon \Omega \to \mathcal{Y} = \{y_1, \ldots, y_c\}$, where $c < \infty$. Assume the existence of a joint distribution, $\mathbb{P}_{\mathbb{G},Y}$ which can be decomposed into the product of a class-conditional distribution (likelihood) $\mathbb{P}_{\mathbb{G}|Y}$ and ~~a~~ class prior π_Y. Because n is finite, the class-conditional ~~distribution~~ distributions $\mathbb{P}_{\mathbb{G}|Y=y} = \mathbb{P}_{\mathbb{G}|y}$ can be considered discrete distributions $\text{Discrete}(G; \boldsymbol{\theta}_y)$, where $\boldsymbol{\theta}_y \in \boldsymbol{\Theta}_{d_n}$ are d_n-dimensional vectors with entries satisfying $\theta_{G|y} \geq 0$ $\forall G \in \mathcal{G}_n$ and $\sum_{G \in \mathcal{G}_n} \theta_{G|y} = 1$.

unit simplex Δ_{d_n}

2.2 A shuffled graph classification model

In the above, it was implicitly assumed that the vertex labels were observed. However, in certain situations (such as the motivating connectomics example presented in Section 1), this assumption is unwarranted. To proceed, we define two graphs $G, G' \in \mathcal{G}_n$ to be isomorphic if and only if there exists a vertex permutation (shuffle) function $Q\colon \mathcal{G}_n \to \mathcal{G}_n$ such that $Q(G) = G'$. Let $\mathbb{Q}$ be a permutation-valued random variable, $\mathbb{Q}\colon \Omega \to \mathcal{Q}_n$, where $\mathcal{Q}_n$ is the space of vertex permutation functions on n vertices so that $|\mathcal{Q}_n| = n!$. Extending the model to include this vertex shuffling distribution yields $\mathbb{P}_{\mathbb{Q},\mathbb{G},Y}$. We assume throughout this work (with loss of generality) that the shuffling distribution is both *class independent* and *graph independent*; therefore, this joint model can be decomposed as

$$\mathbb{P}_{\mathbb{Q},\mathbb{G},Y} = \mathbb{P}_{\mathbb{Q}}\mathbb{P}_{\mathbb{G},Y} = \mathbb{P}_{\mathbb{Q}}\mathbb{P}_{\mathbb{G}|Y}\pi_Y. \tag{1}$$

As in the labeled case, the shuffled graph class-conditional distributions $\mathbb{P}_{\mathbb{Q}(\mathbb{G})|y}$ can be represented by discrete distributions $\text{Discrete}(G; \boldsymbol{\theta}'_y)$, where again $\boldsymbol{\theta}'_y \in \boldsymbol{\Theta}_d$. When $\mathbb{P}_{\mathbb{Q}}$ is uniform on $\mathcal{Q}_n$, all shuffled graphs within the same isomorphism set are equally likely; that is $\{\theta'_{G_i|y} = \theta'_{G_j|y} \, \forall G_i, G_j\colon Q(G_i) = G_j \text{ for some } Q \in \mathcal{Q}_n\}$.

the isomorphism function

2.3 An unlabeled graph classification model

Let $\widetilde{\mathcal{G}}_n$ be the collection of isomorphism sets. An *unlabeled graph* $\widetilde{G}$ is an element of $\widetilde{\mathcal{G}}_n$. The number of unlabeled graphs on n vertices is $|\widetilde{\mathcal{G}}_n| = \widetilde{d}_n \approx d/n!$ (see [?] and references therein). [1] A *unlabeling function* $U\colon \mathcal{G}_n \to \widetilde{\mathcal{G}}_n$ is a function that takes as input a graph and outputs the corresponding unlabeled graph. Let $\widetilde{\mathbb{G}}\colon \Omega \to \widetilde{\mathcal{G}}_n$ be a unlabeled graph-valued random variable taking values $\widetilde{G} \in \widetilde{\mathcal{G}}_n$. The joint distribution over unlabeled graphs and classes is therefore $\mathbb{P}_{\widetilde{\mathbb{G}},Y} = \mathbb{P}_{U(\mathbb{G}),Y} =$

an

- *J.T. Vogelstein and C.E. Priebe are with the Department of Applied Mathematics and Statistics, Johns Hopkins University, Baltimore, MD 21218. E-mail: joshuav@jhu.edu*

cep

$\mathbb{P}_{U(\mathbb{Q}(\mathbb{G})),Y}$, which decomposes as $\mathbb{P}_{\widetilde{\mathbb{G}}|Y}\pi_Y$. The class-conditional distributions $\mathbb{P}_{\widetilde{\mathbb{G}}|y}$ over isomorphism sets (unlabeled graphs) can also be thought of as discrete distributions Discrete$(\widetilde{G}; \widetilde{\boldsymbol{\theta}}_y)$ where $\widetilde{\boldsymbol{\theta}}_y \in \Theta_{\widetilde{d}_n}$ are $\widetilde{d}_n$-dimensional vectors in the d_n-simplex. Comparing shuffling and unlabeling for the independent and uniform shuffle distribution $\mathbb{P}_\mathbb{Q}$, we have $\{\theta'_{G|y} = \widetilde{\theta}_{\widetilde{G}|y}/|\widetilde{G}| \, \forall\colon G \in \widetilde{G}\}$. for all

3 Bayes Optimal Graph Classifiers

three scenarios described above: labeled, shuffled, and unlabeled.

We consider classification in the above three scenarios. To proceed, in each scenario we define three mathematical objects: (i) a classifier, (ii) the Bayes optimal classifier, and (iii) the Bayes risk.

3.1 Bayes Optimal Labeled Graph Classifiers

labeled

A *labeled graph classifier* $h\colon \mathcal{G}_n \to \mathcal{Y}$ is any function that maps from graph space to class space. The risk of a labeled graph classifier under $0-1$ loss is the expected misclassification rate $L(h) = \mathbb{E}[h(\mathbb{G}) \neq Y]$, where the expectation is taken against $\mathbb{P}_{\mathbb{G},Y}$.

The *labeled graph Bayes optimal classifier* is given by

$$h_* = \operatorname*{argmin}_{h \in \mathcal{H}} L(h), \tag{2}$$

where $\mathcal{H}$ is the set of possible labeled graph classifiers.

The *labeled graph Bayes risk* is given by

$$L_* = \min_{h \in \mathcal{H}} L(h), \tag{3}$$

where L_* implicitly depends on $\mathbb{P}_{\mathbb{G},Y}$.

3.2 Bayes Optimal Shuffled Graph Classifiers

A *shuffled graph classifier* is also any function $h\colon \mathcal{G}_n \to \mathcal{Y}$. However, by virtue of the input being a shuffled graph as opposed to a labeled graph, the shuffled risk under $0-1$ loss is given by $L'(h) = \mathbb{E}[h(\mathbb{Q}(\mathbb{G})) \neq Y]$, where the expectation is taken against $\mathbb{P}_{\mathbb{Q}(\mathbb{G}),Y}$.

The *shuffled graph Bayes optimal graph classifier* is given by

$$h'_* = \operatorname*{argmin}_{h \in \mathcal{H}} L'(h). \tag{4}$$

The *shuffled Bayes risk* is given by

$$L'_* = \min_{h \in \mathcal{H}} L'(h), \tag{5}$$

where L'_* implicitly depends on $\mathbb{P}_{\mathbb{Q}(\mathbb{G}),Y}$.

3.3 Bayes Optimal Unlabeled Graph Classifiers

An *unlabeled* graph classifier $\widetilde{h}\colon \widetilde{\mathcal{G}}_n \to \mathcal{Y}$ is any function that maps from unlabeled graph space to class space. The risk under $0-1$ loss is given by $\widetilde{L}(\widetilde{h}) = \mathbb{E}[\widetilde{h}(\widetilde{\mathbb{G}}) \neq Y]$, where the expectation is taken against $\mathbb{P}_{\widetilde{\mathbb{G}},Y}$.

The *unlabeled graph Bayes optimal classifier* is given by

$$\widetilde{h}_* = \operatorname*{argmin}_{\widetilde{h} \in \widetilde{\mathcal{H}}} L(\widetilde{h}), \tag{6}$$

The *unlabeled Bayes risk* is given by

$$\widetilde{L}_* = \min_{\widetilde{h} \in \widetilde{\mathcal{H}}} L(\widetilde{h}), \tag{7}$$

where $\widetilde{\mathcal{H}}$ is the set of possible unlabeled graph classifiers and $\widetilde{L}_*$ implicitly depends on $\mathbb{P}_{\widetilde{\mathbb{G}},Y}$.

3.4 Parametric Classifiers

The three Bayes optimal graph classifiers can be written explicitly in terms of their model parameters:

$$h_* = \operatorname*{argmax}_{y \in \mathcal{Y}} \theta_{G|y}\pi_y, \tag{8}$$

$$h'_* = \operatorname*{argmax}_{y \in \mathcal{Y}} \theta'_{G|y}\pi_y, \tag{9}$$

$$\widetilde{h}_* = \operatorname*{argmax}_{y \in \mathcal{Y}} \widetilde{\theta}_{\widetilde{G}|y}\pi_y. \tag{10}$$

4 Shuffling Can Degrade Optimal Performance

The result of either shuffling or unlabeling a graph can only degrade, but not improve Bayes risk. This is a restatement of the data processing lemma for this scenario. Specifically, [2] shows that the data processing lemma indicates that in the classification domain, $L^*_X \leq L^*_{T(X)}$, for any ~~statistic~~ transformation T and data X. In our setting, this becomes:

Theorem 1. $L_* \leq \widetilde{L}_* = L'_*$.

for simplicity

Proof: Assume $|\mathcal{Y}| = 2$ and $\pi_0 = \pi_1 = 1/2$.

$$\begin{aligned} \widetilde{L}_* &= \sum_{\widetilde{G} \in \widetilde{\mathcal{G}}_n} \min_y \widetilde{\theta}_{\widetilde{G}|y} = \sum_{\widetilde{G} \in \widetilde{\mathcal{G}}_n} \min_y \sum_{G \in \widetilde{G}} \theta'_{G|y} = L'_* \\ &= \sum_{\widetilde{G} \in \widetilde{\mathcal{G}}_n} \min_y \sum_{G \in \widetilde{G}} \theta_{G|y} \geq \sum_{\widetilde{G} \in \widetilde{\mathcal{G}}_n} \sum_{G \in \widetilde{G}} \min_y \theta_{G|y} = L_*. \end{aligned} \tag{11}$$

□

An immediate consequence of the above proof is that the inequality in the statement of Theorem 1 strict whenever the inequality in Eq. 11 is strict:

Theorem 2. $L_* < \widetilde{L}_* = L'_*$ *if and only if there exists* $\widetilde{G}$ *such that*

$$\min_y \widetilde{\theta}_{\widetilde{G}|y} > \sum_{G \in \widetilde{G}} \min_y \theta_{G|y}.$$

OK!

The above result demonstrates that even when the labels *do* carry some class-conditional signal, it may be the case that shuffling or unlabeling does not degrade performance. In other words, to state that labels contain information is equivalent to stating that some graphs within an isomorphism set are class-conditionally more likely than others: $\exists\, \theta_{G_i|y} \neq \theta_{G_j|y}$ where $Q(G_i) = G_j$ for some $G_i, G_j \in \mathcal{G}_n$, $Q \in \mathcal{Q}_n$, and $y \in \mathcal{Y}$. Shuffling has the effect of "flattening" likelihoods within isomorphism sets, from $\boldsymbol{\theta}_y$ to $\boldsymbol{\theta}'_y$, so that $\boldsymbol{\theta}'_y$ satisfies $\{\theta'_{G|y} = \widetilde{\theta}_{\widetilde{G}|y}/|\widetilde{G}| \, \forall\colon G \in \widetilde{G}\}$. But just because the shuffling changes class-conditional likelihoods does *not* mean that Bayes risk must also change. This result follows for generalized (non-equal) class priors,

immediately upon realizing that posteriors can change without classification performance changing. The above results hold ~~in the absence of equal priors~~, and are easily generalized to c-class classification problems. To see this, ignoring ties, simply replace each minimum with a sum over all non-maxima:

maps to

$$\min_y \theta_{G|y} \mapsto \sum_{y \in \mathcal{Y}'} \theta_{G|y} \text{ where } \mathcal{Y}' = \{y : y \neq \operatorname{argmax}_y \theta_{G|y}\}.$$

5 Bayes Optimal Graph Invariant-Based Classification After Shuffling

A graph invariant on $\mathcal{G}_n$ is any function ψ such that $\psi(G) = \psi(Q(G))$ for all $G \in \mathcal{G}_n$ and $Q \in \mathcal{Q}_n$. A graph invariant based classifier is a composition of a vector-based classifier with an invariant function, $h^\psi = f^\psi \circ \psi$, where $f^\psi : \mathbb{R}^{d'} \to \mathcal{Y}$. The Bayes optimal graph invariant classifier minimizes risk over all invariants:

$$h_*^\psi = \operatorname*{argmin}_{\psi \in \Psi, f^\psi \in \mathcal{F}^\psi} \mathbb{E}[f(\psi(\mathbb{G})) \neq Y], \tag{12}$$

composable with ?

where Ψ is the space of all possible invariants and $\mathcal{F}^\psi$ is the space of classifiers ~~using~~ applicable to invariant ψ. The expectation in Eq. (12) is taken against $\mathbb{P}_{\mathbb{G},Y}$ or equivalently $\mathbb{P}_{\mathbb{Q}(\mathbb{G}),Y}$, since invariants are invariant. Let L_*^ψ denote the Bayes invariant risk.

Theorem 3. $\widetilde{L}_* = L_*^\psi$.

Proof: Let ψ indicate in which equivalence set G resides; that is, $\psi(G) = \widetilde{G}$ if and only if $G \in \widetilde{G}$. Then

$$h_*^\psi = \operatorname*{argmax}_{y \in \mathcal{Y}} \widetilde{\theta}_{\psi(G)|y}\pi_y = \operatorname*{argmax}_{y \in \mathcal{Y}} \widetilde{\theta}_{\widetilde{G}|y}\pi_y = \widetilde{h}_*. \tag{13}$$

□

6 A Consistent and Efficient Unshuffling-Based Classifier

Section 4 shows that one cannot fruitfully "unshuffle" graphs: once they have been shuffled by a uniform shuffler, the label information is lost. Section 5 shows that if graphs have been uniformly shuffled, there is a relatively straightforward algorithm for optimal classification. However, that classifier depends on knowing the parameters, $\widetilde{\boldsymbol{\theta}} = \{\widetilde{\theta}_y\}_{y \in \mathcal{Y}}$ and $\boldsymbol{\pi} = \{\pi_y\}_{y \in \mathcal{Y}}$. Instead we consider $(\mathbb{Q}_i, \mathbb{G}_i, Y_i) \overset{iid}{\sim} \mathbb{P}_{\mathbb{Q},\mathbb{G},Y}$. For shuffled graph classification we observe only the *training data* $\mathcal{T}_s = \{\mathbb{G}'_i, Y_i\}_{i \in [s]}$, where $\mathbb{G}'_i = \mathbb{Q}_i(\mathbb{G}_i)$, and are thusly unable to observe useful vertex labels. Moreover, $\mathbb{P}_{\mathbb{Q}}$ is uniform, so that all label information is both unavailable and irrecoverable. Our task is to utilize training data to induce a classifier $\hat{h}_s : \mathcal{G}_n \times (\mathcal{G}_n \times \mathcal{Y})^s \to \mathcal{Y}$ that approximates $\widetilde{h}_*$ as closely as possible.

An unlabeled graph Bayes *plugin* classifier estimates the likelihood and prior terms and plugs them in to Eq. (10):

$$\hat{h}_s = \operatorname*{argmax}_{y \in \mathcal{Y}} \hat{\widetilde{\theta}}_{\widetilde{G}|y}\hat{\pi}_y. \tag{14}$$

nearly ~~[illegible]~~ the everywhere else you did NOT use parens?

Let $\hat{L}_s = L(\hat{h}_s)$ be the risk of the induced classifier.

Use max likelihood to obtain the plugin estimates for Eq (14)

Theorem 4. $\hat{L}_s \to \widetilde{L}^*$ *as* $s \to \infty$.

Proof: Because $\widetilde{\mathcal{G}}_n$ and $\mathcal{Y}$ are both finite, their respective maximum likelihood estimates are guaranteed consistent by the law of large numbers. Hence, the Bayes unlabeled graph plugin classifier is also consistent [2]. □

7 A Practical Approach to Shuffled Graph Classification

graph matching? graph isomorph?

Although Eq. (14) yields consistency from Theorem 4, utilizing this is practically hopeless as it requires solving s computationally difficult problems, and acceptable performance will typically require $s \gg \widetilde{d}_n$. Specifically, using Eq. (14) requires first enumerating all $\widetilde{d}_n$ isomorphism sets, then determining in which isomorphism set the to-be-classified graph and each of the training graphs resides. There are no known polynomial time solvers for graph isomorphism. This approach is therefore, in general, impractical on two levels: (i) the number of parameters to estimate, $\widetilde{d}_n$, is too large, and (ii) exact graph isomorphism is too computationally taxing. We therefore consider a modified approach.

to $\widetilde{L}_*$

A k_s nearest-neighbor (kNN) classifier using Frobenius norm is universally consistent for labeled graph classification as long as $k_s \to \infty$ with $k_s/s \to 0$ as $s \to \infty$ [?]. This non-parametric approach circumvents the need to estimate $\widetilde{d}_n$ parameters. We use a graph-matched Frobenius norm as the distance function,

for shuffled,

$$\delta(G_i, G_j) = \operatorname*{argmin}_{Q \in \mathcal{Q}_n} \|Q(G_i) - G_j\|_F^2. \tag{15}$$

universal ~~but yields consistency~~ to $\widetilde{L}_*$

Eq. (15) requires solving a graph matching problem, which is NP-hard. Therefore, we instead use an inexact graph matching approach based on the quadratic assignment formulation described in [3] which is only cubic in n. Note that while the kNN approach with exact matching maintains universal consistency, to $\widetilde{L}_*$ kNN with inexact graph matching approximation may not be consistent.

The kNN classifier for shuffled graphs proceeds as follows. First, compute the graph-matched Frobenius norm distance between the test graph and all training graphs, $\{\delta_i = \delta(G, G_i)\}_{i \in [s]}$. Second, order the graph/class pairs according to their distances, $\delta_{(1)} \leq \cdots \leq \delta_{(s)}$. Finally, let the estimated class be the plurality class of the k_s closest graphs; that is, $\hat{y} = \operatorname{argmax}_{y \in \mathcal{Y}} \sum_{i \in [k_s]} \mathbb{I}\{y_{(i)} = y\}$.

8 Simulated Experiment

To demonstrate the practically of an isomorphism-based unlabeled graph classifier, we conduct the following simulated experiment. Sample $s+1$ triplets identically and independently from the joint shuffler/graph/class model, $(\mathbb{Q}_i, \mathbb{G}_i, Y_i) \overset{iid}{=} \mathbb{P}' = \mathbb{P}_{\mathbb{Q}}\mathbb{P}_{\mathbb{G}|Y}\pi_Y$, where $\mathbb{P}_{\mathbb{Q}}$ is uniform, and π_Y is Bernoulli so $\pi_0 = \pi_1 = 1/2$. The edges are independent so the likelihood factorizes as