1 Summary

The manuscript proposes a Bayesian model for classifying graphs in which the number of nodes n is fixed and known a priori. Due to this property, the problem boils down to associating with each class a subset of all $n(n-1)/2$ possible edges between n nodes. The structure of the graph can therefore be be encoded in $n(n-1)/2$ binary random variables which puts the model in the context of conventional Bayesian networks.

The main contribution is to assume that the probability whether an edge is present or not is class-dependent only for a subset of edges. This subset is introduced as a random variable into the model and is constrained a priori to have either a fixed number of edges or a fixed number of nodes to which all edges are incident. This makes for an interesting higher-order prior that is motivated by the application: estimating correlation graphs from MR brain images.

Except for the interesting assumption that only some edge probabilities are class-conditional, the model is simple in the sense that it makes strong conditional independence assumptions (see Fig. 1 below) which are clearly violated in practice and "almost certainly nonsense for real connectome data".

2 Novelty and Importance

The proposed model is novel not because it encodes the structure of graphs differently than other models but because it introduces a likelihood term that makes only part of the edges class dependent, a random variable that selects these edges and an application-motivated prior on this variable.

Learning boils down to estimating the individual terms by standard methods (χ^2 or absolute difference testing and empirical averaging) which is nice because the estimates are consistent, but conventional.

The model itself is a conceptual contribution. To be able to judge whether it is important enough to qualify for publication in this prestigious journal, more experimental evidence for its usefulness needs to be provided. The experiments with synthetic data suggest that the model was implemented correctly. The application to real data does not include a comparison with known truth or a gold standard and no comparison with other state-of-the art graph classification methods [b]. Based on the results in the manuscript, this reviewer cannot judge the importance of the model.

3 Required Changes

In summary, the following issues need to be addressed (see also more specific comments below): 1. Clarification of the technical correctness of the probabilistic model. 2. Experimental confirmation that the proposed model is useful. 3. Discussion of related work. 4. Improvement of the presentation, focusing on the original contributions and limiting the use of mathematical rigor to parts which are novel.

3.1 Technical Correctness

To discuss the technical correctness of the manuscript, it is essential to validate that my understanding of the model is correct. Therefore, I rephrase the model in the following and ask the authors to point out any misunderstandings I might have:

Under consideration is a triple (A, S, Y) of random variables. Realizations of A and S are simple undirected graphs with a fixed number $n \in \mathbb{N}$ of nodes. These can be encoded by indicator vectors $a, s \in \{0,1\}^m$ whose entries are associated with the $m = n(n-1)/2$ possible edges. Realizations of Y are class labels. The proposed probabilistic model takes the form

$$P(A, Y, S|\theta) = \frac{P(S)\ P(Y)}{P(\theta)} \prod_{e \in E} P(A_e|\theta_e) P(\theta_e|S_e, Y) \ . \tag{1}$$

that can be derived as follows from the conditional independence assumptions stated graphically and explicitly in Fig. 1:

$$\begin{aligned}
P(A, Y, S|\theta) &= P(A|Y, S, \theta)\ P(Y, S|\theta) && (2)\\
&= P(A|\theta)\ P(S, Y|\theta) && (3)\\
&= P(S, Y|\theta) \prod_{e \in E} P(A_e|\theta_e) && (4)\\
&= \frac{P(\theta|S, Y)\ P(S, Y)}{P(\theta)} \prod_{e \in E} P(A_e|\theta_e) && (5)\\
&= \frac{P(\theta|S, Y)\ P(S)\ P(Y)}{P(\theta)} \prod_{e \in E} P(A_e|\theta_e) && (6)\\
&= \frac{P(S)\ P(Y)}{P(\theta)} \prod_{e \in E} P(A_e|\theta_e) P(\theta_e|S_e, Y) \ . && (7)
\end{aligned}$$

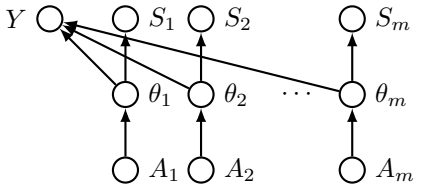
The two contributions of the manuscript are i. to define $P(\theta_e|S_e, Y)$ such that θ_e depends on Y only if $s_e = 1$, i.e. to make θ_e class-conditional only for a subset of edges, and ii. to introduce a higher-order prior $P(S)$ that constrains the class-conditional subgraph to have certain properties, either a maximum number $q \in \mathbb{N}$ of edges,

$$p(s) = P(S = s) = \begin{cases} \text{const.} & \text{if } \sum_{j=1}^{m} s_j \leq q \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

or a maximum number $r \in \mathbb{N}$ of nodes to which all selected edges are incident,

$$p(s) = P(S = s) = \begin{cases} \text{const.} & \text{if } \left|\cup_{j=1}^{m} e_j\right| \leq r \\ 0 & \text{otherwise} \end{cases} \ . \tag{9}$$

If my understanding of the model is correct, I suggest to adopt the formulation above because it makes clear that S is a random variable that can be



$$S \perp Y \tag{10}$$
$$A \perp Y \mid \theta \tag{11}$$
$$A \perp S \mid \theta \tag{12}$$

For all $e, f \in E$ such that $e \neq f$:

$$A_e \perp \theta_f \mid \theta_e \tag{13}$$
$$A_e \perp A_f \mid \theta \tag{14}$$
$$\theta_e \perp S_f \mid Y \tag{15}$$
$$\theta_e \perp \theta_f \mid Y \tag{16}$$

Figure 1: Without a prior on S, the probabilistic model would be a Bayesian network (left) with the conditional independence assumptions on the right. The higher-order prior $P(S)$ makes $S_1, \ldots, S_m$ mutually dependent which is consistent with (10)–(16).

estimated, it exposes the higher-order prior on S that is an original contribution and it makes the conditional independence assumptions explicit.

Due to the introduction of the random variable S and the prior $P(S)$, this model is more specific than the framework of hybrid generative-discriminative models in [a].

3.2 Related Work

It would be appropriate to devote one section to related work, e.g. some of the methods described in [b]. Without any references except [1,2] the following statements remain unjustified:

> This graph-model based approach is qualitatively different from most previous approaches which utilize only vertex labels or graph structure
>
> Nonetheless, the classifiers described below still significantly outperform the benchmarks.

Due to the existence of the encoding mentioned above, the following statement should be avoided.

> This graph-model based approach is qualitatively different from most previous approaches which utilize only vertex labels or graph structure. In the former case, simply representing the adjacency matrix with a vector and applying standard machine learning techniques ignores graph structure [...]

3.3 Notation, Presentation and Style

While the article is understandable and conveys a message, readability needs to be improved. I suggest

- to get rid of the passages that recite textbook knowledge and instead focus on the original contributions. The additional space should be devoted to an informal, intuitive introduction of these contributions and a discussion of related work.
- to drastically reduce the amount of mathematical rigor and to use a more common mathematical notation where necessary, e.g. only (1) from this review and possibly the graphical model in Fig. 1.
- to get rid of Fig. 1 and shorten Section 2.4 as far as it refers to this figure. It is sufficient to say that the number of all possible graphs with n nodes, even with the constraining properties, is too large to be explored by exhaustive search.
- to start with a more pronounced and intuitive motivation for graph classification in the context of MR connectomics.
- to use the term *class-conditional subgraph* consistently instead of *signal subgraph*. The latter is potentially misleading because it does not refer to a signal as in *signal processing* as some readers of this journal will assume but to *signalling neurons*.
- to replace the term *coherent graph* that has many connotations unrelated to its meaning here and should therefore be avoided.

Moreover, it is sufficient to say,

- instead of Section 2.4.1.1 (Section 2.4.1.2) that a threshold c is selected minimally such that $>= s$ edges have a significance level $< c$ ($>= s$ nodes exists such that each has an indicent edge with a significance level $< c$).
- instead of 3.1 that the standard methods for estimating the terms of the model are consistent.

Examples of textbook style:

> A model defines the set of distributions under consideration.
>
> Two standard approaches for tackling a classification problem are (i) the generative approach and (ii) the discriminative approach. In a generative strategy, one decomposes the joint distribution into a product of a likelihood term and a prior term: [...]. In a discriminative strategy, one decomposes the joint distribution into a posterior term and a marginal term: [...].
>
> An estimator is a function that maps from the multiple-sample space to the parameter space. [...] The output of this function is called the estimate.

3.4 Minor Remarks

Since the model is defined for undirected graphs, the edge set should not be defined as a Cartesian product but as a set of 2-elementary subsets, e.g. using the notation $\binom{V}{2}$.

4 References

[a] Lasserre J.A., Bishop C.M. and Minka, T.P. Principled Hybrids of Generative and Discriminative Models. CVPR 2006

[b] Ketkar, N.S.; Holder, L.B.; Cook, D.J.; , "Empirical comparison of graph classification algorithms," Computational Intelligence and Data Mining, 2009. CIDM '09. IEEE Symposium on , vol., no., pp.259-266, March 30 2009-April 2 2009 doi: 10.1109/CIDM.2009.4938658