Revision notes for paper “The Conjunctive Disjunctive Graph Node Kernel for Disease Gene Prioritization”

Dinh Tran Van, Alessandro Sperduti and Fabrizio Costa

November 17, 2017

We thank the reviewer for the reviews and helpful comments on our manuscript “The Conjunctive Disjunctive Graph Node Kernel for Disease Gene Prioritization”. In the following, we provide a description of how each comment (in *italic*) has been addressed in this version of the paper.

1 Reviewer 1

1

- *lines 10-15: if the first claim (bias toward high degree nodes for the heat kernel) is understandable, the others are not and I don’t get how the diffusion kernel (for instance) solves this issue, since it is very similar to the previous one. An evaluation of how the kernels address this property should be included in the experimental section or the authors should cite a paper already studying this property on different kernels;*

  Answer: we explain in detail in “Section Graph Node Kernels” of the revised version as follows.

  In DK, the similarity values between high degree vertices are generally higher compared to those between low degree ones. Intuitively, the more paths connect two vertices, the more heat can flow between them. This could be problematic since peripheral nodes (i.e., low degree nodes) have unbalanced similarities with respect to central nodes (i.e., high degree nodes). In order to make the strength of individual vertices comparable, a modified version of DK is Markov exponential diffusion kernel (MED) and given by the following formula:

$$\mathbf{K} = e^{-\beta \mathbf{M}}. \tag{1}$$

The difference with respect to the DK is the replacement of $\mathbf{L}$ by the matrix $\mathbf{M} = (\mathbf{D} - \mathbf{A} - n\mathbf{I})/n$ where n is the total number of vertices in the graph. In particular, the introduced modification of the Laplacian is designed to compensate the problem described above via a sort of normalization of the degree of vertices. The role of β is the same as for DK.

- *lines 20-25: what does "process in an additive and independent fashion" mean in this context. Similarly as before, please justify your claim about the poor discriminative power of these kernels;*

  Answer: we explain in more intuitive way in the new version of the paper.

  The state-of-the-art graph node kernels used to measure node similarity are based on the notion of information diffusion, which mainly depends on the number of paths connecting two nodes in the graph. These graph node kernels often show relatively low discriminative capacity, especially when working with sparse graphs (i.e., graphs with a low number of links), because of the following limitations. First, they are defined using a heat diffusion dynamics which sums up the contributions of all paths from one node to another one, disregarding the local topological context of the nodes and considering the single contribution of one path as independent with respect to the contributions of other paths. Second, they do not take into account additional information (i.e. properties of a single node) eventually associated to nodes of graphs, when available.

- *lines 146-150: The first claim is misleading. The problem is not that the degree of some nodes is high it is that the degree distribution is highly skewed (and you must include a reference stating this fact for biological networks, as well as describe the own networks on which you are working for this property). Also, the rest of the argument is very strange: how do you get that the probability to find matches decreases exponentially. If there is a mathematical argument behind it, please, detail it;*

Answer: The existance of high degree nodes will lead to the big neighborhood subgraphs. Since the proposed kernel is based on the exact matching neighborhood subgraphs (isomorphic neighborhood subgraphs), there will be a low probability to find any couple of big neighborhood subgraphs which are isomorphic. This probabability decrease shapely w.r.t the increase of neighborhood subgraphs size or node degree.

2

The paper is oriented exclusively toward an given application in biology. However, it is not visible from the title of the article and methodological choices are seldom justified in this context. Gene labelling should be evaluated: for instance, it would have been helpful to have a descriptive evaluation of the results of those labels in the experiment section. For the topological labelling, what is the biological meaning of nodes with similar degree being more similar? (give a reference) I don't understand if the topological labels are considered as a discrete or numerical label? An example of the way it partitions the network would be useful. For the functional labelling, evaluate your clusters in a way. Do they somehow relate to the graph structure or not? Do they correspond to clusters enriched in some functions? How robust is this clustering to very incomplete information as found in GO databases? In a similar fashion, can you provide an evaluation of your decomposition on the networks you are working with?

Answer: We agree with the suggestion of the reviewer about the title of the kernel that it should be toward to disease gene prioritization. Therefore, we have modified it as "The conjunctive disjunctive graph node kernel for disease gene prioritization" in the new version of the paper.

In general, our kernel takes labeled graphs as the input. Given a graph, users can apply any proper node labeling function to the graph. The labeling functions proposed in the paper are examples to evaluate the performance of the disease gene prioritization system with the use of our kernel. Regarding the proposed labeling functions, we consider topological labels as discrete labels. The GO datasets do not provide complete information. However, they still provide additional information for the kernel. We strongly agree that the intensitive evaluation for labeling functions is necessary to improve the performance of the proposed kernel. However, in this paper, we would like to have an attempt to see how the CDNK performs compared to other graph node kernels.

3

The paper is extremely badly organized, with some notions (like conjunctive and disjunctive edges) used before they are defined. I suggest the following re-organization of the paper:

1. *Introduction (that must include a discussion about the problem at stake, gene prioritization)*

   Answer: we have added in the new version of the paper.

2. *A new kernel (or whatever title you want more informative than "Material and methods" that is a title better suited for bioinformatics journals) with A. Notations B. Node graph kernels + their problems C. Network decomposition D. Your kernel (with the case of discrete and numerical labels separated)*

   Answer: we have re-organized the structure of the paper based on the reviewer's comments.

3. *Examples of node labelling in your context (to my opinion, Section 2.2.2 does not need the labels to be known. Just explain the kernel in Section 2 and then, in a separate section afterwards, explain your proposals for labelling with a better illustration of their usefulness in the biological framework)*

   Answer: we have modified the paper based on the reviewer's comment.

4. *Experiments (since the discussion section is very short, I don't think that it is necessary or even relevant to have it separated from the rest of the paper)*

   Answer: we have modified the paper based on the reviewer's comment.

4

The experimental part is very light. I would have expected to see a smart evaluation on a simulated dataset illustrating in which theoretical cases the approach is relevant. I think that an evaluation of the sensitivity of the method to some parameters is missing. For instance, why use a gene co-expression network with a correlation threshold of 0.5? (To my own experience, this is a very low threshold that should give a very dense network and *is not the standard choices in this area). Also, why have twice less negative examples than positive ones (this is very far from being realistic since positive genes are usually very rare in this kind of real-life applications)?*

Answer: We would like to compare our node kernel diffusion-based graph node kernels on relatively dense networks, since diffusion-based kernels often show good performance with dense networks. That is why we borrow BioGPS and Pathways which are used in [2].

In disease prioritization, given a genetic disease, we are often given a small set of positive genes and a large set of un-labeled genes. It is difficult to have a proper negative set of genes. Therefore, in [2] they choose the number of postive genes two times greater than the number of negative ones. The negative genes are choosen from genes that are associated with at least one disease class, but are not related to the class that is defining the positive set. We assume that they are less likely to be associated to the considered class, since disease genes are generally more studied and a potential association has more chances to be identified.

2 Reviewer 2

- *I have found the paper to be hard to follow. This is mostly due to the large number of "sub-methods" that are involved in the proposed apparach, and to the too simplified description of the "global picture" of what the authors are proposing. It would be better to have an initial section or a table or a figure to summarize and easily introduce what the authors are going to propose/describe in the rest of the paper.*

  Answer: we have modified the structure of the paper in order to be easier to follow.

- *Related to the previous comment, the notion of conjunctive and disjuctive edges seems central, however they are shortly mentioned in Section 2.2.2 and then shorlty re-defined (?) in Section 2.4. In the first case, the description of these edges must be expanced and clarified, togheter with some more details about Figure 1 (caption and text). Maybe it can be anticipated in the "Definitions" paragraph (page 3), or in the "initial decription" that I introduced in the previous comment.*

  Answer: We have added a more detailed explanation in the new version of the paper based on the reviewer's comment.

- *Overall, there is very large number of parameters involved in the proposed approach. The authors describe their model validation procedure in Page 9/10, however the space of possible combinations of parameters is huge + the choices over other thresholds described in the paper. This seems to be an important drawback of the method, and some comments on this point should be added.*

  Answer: We have added a new section named "Hyper-parameter" to explain why the parameter space for the proposed kernel is not high as follows.

  Our kernel consists of five hyper-parameters: the threshold degree D, the clique size threshold C, the maximal radius r^*, the maximal distance d^*, and the number of clusters P. In order to choose the optimal tuple of values for kernel hyper-parameters in a specific setting, a model selection procedure is normally adopted. The hyper-parameter space of our kernel is potentially large due to the relatively high number of hyper-parameters. However, all hyper-parameters are integer variables and it makes sense to use only small values for them. For example, the value of the threshold degree D should neither be too high nor too low: if it is high, the neighborhood subgraphs are going to be too big and the isomorphism check becomes inefficient; vice versa, if it is too low, the decomposed graph contains connected components which are too sparse to allow learning to be effective. Moreover, it is reasonable to keep the clique size threshold C within a narrow range of values in order to avoid the removal of too many edges. Concerning the maximal radius r^*, too high values will lead again to big neighborhood subgraphs, especially in graphs with a high node degree average. The maximal distance d^* can be set with small values since the average shortest distance length between nodes of biological networks is relatively low. Finally, if the value assigned to P is too low, all nodes basically get the same information attached to them and thus the labels are not informative; vice versa, if a high value is used, there will be very few matches and the kernel will be extremely sparse, making learning ineffective.

- *Page 4, Functional label and real valued vector information encoding: The K-means algorithm is exploited, using the classical Euclidean distance, and the same distance is used in the vec.info.encoding. The authors should motivate this choice.*

Answer: We have added a more detailed explanation in the new version of the paper.

- *Page 6, Section 2.3. A few words about the hash code computation can be helpful.*

  Answer: We have added a more detailed explanation in the new version of the paper.