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

Dinh Tran Van, Alessandro Sperduti and Fabrizio Costa

October 22, 2017

We thank the reviewer for the reviews and helpful comments on our manuscript “The Conjunctive Disjunctive Graph Node Kernel”. 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: in LEDK, the similarity values between high degree nodes 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 have unbalanced similarities with respect to central nodes. MEDK adapt LEDK by adding a normalization in order to make the strength of individual vertices comparable.

- *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:

- *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 kernel extracts features as pairwise neighiborhood subgraphs. If the degree of nodes is high, it leads to big extracted neighborhood subgraphs, and it is difficult to find big graphs which are isomorphic. Therefore, we state that the probability to find matches exponentially decreases w.r.t the 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 have modified the title of the paper toward to disease gene prioritization.
- In the method, we consider the topological labeling as the discrete label
- The number of clusters used in the functional labeling is chosen from a set of predefined values through the model selection.

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: the organization of the paper is modified 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 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 already modified based on the reviewer's comment.

4

The experimental part is very light. I would have expected to see an 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 follow the experimental setting, including dataset BioGPS and Pathways, performed in [2].

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 based on the reviewer's comment.

- *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 modified 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 Parameter Space with the content as follow.

  The kernel consists of five parameters: the threshold degree D, clique size threshold C, maximal radius r^*, maximal distance d^* and number of clusters P clusters in which their values are in $\mathbb{N}^*$. In order to choose the optimal tuple of parameters for kernel in a specific setting, a model selection procedure is normally adopted from a determined subset of parameter space. The parameter space of our kernel seems large due to the relatively high number of parameters. However, most parameter values are supposed to be in limited natural ranges. Therefore, it is actually not too big. In particular, the clique size threshold

C is recommended to be in $\{4, 5\}$. The maximal radius r^* is set with a value smaller than or equal to 3 since it will lead to big neighborhood subgraphs. The maximal distance d^* is assigned with values less than or equal to 4 because the value of the maximal shortest distance between nodes in a connected component is often not too high. The values of the threshold degree D should neither be too high and too low. If it is high, we have to face with high degree node problem, and if it is too low, the obtained graph contains of too sparse connected components. Therefore, we suggest the value for D in $[6, 20]$

- *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.