Gene Function Prediction via Discriminative Graph Embedding

Canh Hao Nguyen  
canhhao@kuicr.kyoto-u.ac.jp  
Hiroshi Mamitsuka  
mami@kuicr.kyoto-u.ac.jp

1 Bioinformatics Center, Institute of Chemical Research, Kyoto University, Gokasho, Uji, Kyoto, 611-0011, Japan.

Keywords: Gene Function Prediction, Discriminative Graph Embedding, Spectral Transform

1 Introduction

Gene function has been a subject of interest but it is far from fully understood. It is known that some genes have certain functions but it is not clear whether those are all the functions they have. It is a recent trend to use different means to predict gene functions; one of them is to use computational methods on large data sets. Different types of information are used in computational methods, each coming with different hypotheses and applicable to different part of genomes.

We regard this problem as a prediction problem on networks. The idea is to use gene network information with a hypothesis that a link in the network means a highly probable similarity in gene functions. This is formulated as a node classification problem on graph with smooth label assignment with respect to the graph structure. We propose a method to embed the graph into a Euclidean space considering label information. The embedding is supposed to trade off the graph structure preservation to have a good discriminative ability. Then, label inference can be done via many machine learning techniques, in which we formulate a large margin spectral transform, known for its robustness. We apply the method to predict gene functions on gene network [1] and protein-protein interactions network [2]. Our method gives higher prediction ability than traditional ways of embedding graphs.

2 Method and Result

There are many methods to embed graphs into Euclidean spaces for learning algorithms to be used to infer node labels in graphs. However, most methods embed the graph faithfully regardless of the problem in hand. Therefore, the resulting embedding distribution respects strictly graph structures but does not necessarily contain the very information needed for the task, namely predicting node labels. There are also methods for estimating node labels directly, but due to their inflexibility, they cannot use robust prediction models to achieve the highest possible result.

We propose a method to embed a graph into a Euclidean space keeping the discriminative information needed for prediction later on. This embedding step has two objectives. One is the label smoothness on a graph, which is similar to Laplacian eigenmaps [3]. Another is to be discriminative, which we formulate to separate the two classes’ means as far as possible in the embedding space. Each of these objectives is formulated as a generalized eigenvalue problem. We combine these two objectives by weighted combination of the two eigenvalue problem. The final formulation is a generalized eigenvalue problem, which has efficient solutions. We pick only some eigenvectors with smallest eigenvalues to represent graphs in a Euclidean spaces for learning algorithms to be used later. This embedding is supposed to have both graph structure information (label smoothness on the graph) and discriminative information (large distance between classes’ means).
We formulate an efficient way to combine features (as eigenvectors) for a discriminative purpose (classification). As it is used to transform eigenvalues, it becomes a (nonparametric) spectral transform method. The formulation is robust because we use the large margin objective function. The problem is then formulated into a multiple kernel learning (MKL) problem [4], which is well-studied and has efficient solutions.

We used two graphs to infer gene functions. One was the gene network from the KEGG pathway database [1]. Another was a protein-protein interactions network [2]. We could use only a connected part of the whole graph due to the limitation of the method (using only graph structures). We compared our method with the method without discriminative information in the embedding process. The result, measured in Area under the ROC curve, over 50 runs of train/test split, is presented in Table 1.

<table>
<thead>
<tr>
<th>Method</th>
<th>Gene Network</th>
<th></th>
<th></th>
<th>PPI Network</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>AminoAcid</td>
<td>Carbohydrate</td>
<td>Nucleotide</td>
<td>AminoAcid</td>
<td>Carbohydrate</td>
<td>Nucleotide</td>
</tr>
<tr>
<td>L.E.</td>
<td>0.87 (±0.06)</td>
<td>0.75 (±0.08)</td>
<td>0.89 (±0.07)</td>
<td>0.54 (±0.09)</td>
<td>0.69 (±0.10)</td>
<td>0.52 (±0.11)</td>
</tr>
<tr>
<td>Ours</td>
<td>0.98 (±0.01)</td>
<td>0.99 (±0.02)</td>
<td>1.00 (±0.00)</td>
<td>0.86 (±0.06)</td>
<td>0.86 (±0.07)</td>
<td>0.80 (±0.08)</td>
</tr>
</tbody>
</table>

Table 1: Average AUC scores for predicting gene function categories with two networks. The Laplacian eigenmaps (L.E.) embedding respects the graph structures but does not contain discriminative information, therefore performs poorly in classification tasks. Our embedding method gives a significantly better performance in the two networks, across all function classification tasks.

3 Discussion

The result shows that this embedding is more suitable for a commonly used Machine Learning technique to classify nodes on graphs, meaning that the embedding makes the learning technique more effective for label propagation. Our method is recommended to use for the parts of genomes where dense enough networks of function similarity can be obtained.

4 Acknowledgements

This work is partially supported by BIRD (JST) and JSPS fellowship.

References


