Link Enrichment for Strengthening Diffusion Based Kernels

Dinh Tran Van[1], Alessandro Sperduti[1] and Fabrizio Costa[2]

1- Department of Mathematics, Padova University
Trieste, 63, 35121 Padova, Italy

2- Bioinformatics Group, Department of Computer Science, Freiburg University
Georges-Kohler-Allee 106, 79110 Freiburg, Germany

Abstract. The similarity measurement between genes is one of the key points that determines the performance of disease-gene association discovering systems. Genes and their relations are commonly encoded into graphs in which nodes represent for genes and edges characterize the gene relations. Although information diffusion based graph node kernels emerge as the most promissing paradigms used to define gene similarity, they only present promising performance on the graphs which are dense and connected. Here, we propose a link enrichment based method to strengthen diffusion-based kernels on sparse graphs or graphs whose disconnected components. An empirical evaluation on different biological datasets shows that our propose method improves the performance of diffusion-based kernels.

1 Introduction and Related Work

Predictive systems for gene-disease associations are often based on a notion of similarity between genes. A common strategy is to encode relations between genes as a network and use graph based techniques to make useful inferences. In the last decades, a number of graph kernel methods have been proposed that directly exploit transitive properties in biological networks. The prototypical method is the Diffusion kernel (DK) [3] inspired by the heat diffusion phenomenon. The key idea is to allow a given amount of *heat* placed on nodes to *diffuse* through the edges. The similarity between two nodes v_i, v_j is then defined as the amount of heat starting from v_i and reaching v_j within a given time interval. In DK the heat flow is proportional to the number of paths connecting two nodes, which introduces an undesired bias that penalize peripheral nodes w.r.t. central ones. This problem is tackled by a modified version of DK called Markov exponential diffusion kernel (MED) [4] where a Markov matrix replaces the Laplacian matrix. Another kernel called Markov diffusion kernel (MD) [5], exploits instead the notion of *diffusion distances*, a measure of similarity between patterns of heat diffusion. The Regularized Laplacian kernel (RL) [6] represents instead a normalized version of the random walk with restart model and defines the node similarity as the number of paths connecting two nodes with different lengths. All these approaches can be applied to dense networks with high degree nodes. A drawback of these approaches is however their relatively low discriminative capacity. This is in part due to the fact that information is processed

in an additive and independent fashion which prevents them from accurately modeling the configuration of each gene's context. To address this issue here we propose to employ a *decompositional* graph kernel (DGK) [2] technique. To exploit its higher discriminative capacity we first decompose the network in a collection of connected sparse graphs and then we develop a suitable kernel, that we call Conjunctive Disjunctive Node Kernel (CDNK).

2 Method

2.1 Graph Node Kernels

Laplacian exponential diffusion kernel
One of the most well-known kernels for graphs is the Laplacian exponential diffusion kernel $\mathbf{K}_{LED}$, as it is widely used for exploiting discrete structures in general and graphs in particular. On the basis of the heat diffusion dynamics, Kondor and Lafferty proposed $\mathbf{K}_{LED}$ in [3]: imagine to initialize each vertex with a given amount of heat and let it flow through the edges until an arbitrary instant of time. The similarity between any vertex couple v_i, v_j is the amount of heat starting from v_i and reaching v_j within the given time. Therefore, $\mathbf{K}_{LED}$ can capture the long range relationship between vertices of a graph to define the global similarities. Below is the formula to compute $\mathbf{K}_{LED}$ values:

$$\mathbf{K}_{LED} = e^{-\beta \mathbf{L}} = \mathbf{I} - \beta \mathbf{L} + \frac{\beta \mathbf{L}^2}{2!} - ... \tag{1}$$

where β is the diffusion parameter and is used to control the rate of diffusion and $\mathbf{I}$ is the identity matrix. Choosing a consistent value for β is very important: on the one side, if β is too small, the local information cannot be diffused effectively and, on the other side, if it is too large, the local information will be lost. $\mathbf{K}_{LED}$ is positive semi-definite as proved in [3].

Markov exponential diffusion kernel
In $\mathbf{K}_{LED}$, similarity values between high degree vertices is generally higher compared to that 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. In order to make the strength of individual vertices comparable, a modified version of $\mathbf{K}_{LED}$ was introduced by Chen et al in [**?**], called Markov exponential diffusion kernel $\mathbf{K}_{MED}$ and given by the following formula:

$$\mathbf{K}_{MED} = e^{-\beta \mathbf{M}} \tag{2}$$

The difference with respect to the Laplacian diffusion kernel 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 graph. The role of β is the same as for $\mathbf{K}_{LED}$.

Markov diffusion kernel

The original Markov diffusion kernel $\mathbf{K}_{MD}$ was introduced by Fouss et al. [5] and exploits the idea of diffusion distance, which is a measure of how similar the pattern of heat diffusion is among a pair of initialized nodes. In other words, it expresses how much nodes "influence" each other in a similar fashion. If their diffusion ways are alike, the similarity will be high and, vice versa, it will be low if they diffuse differently. This kernel is computed starting from the transition matrix $\mathbf{P}$ and by defining $\mathbf{Z}(t) = \frac{1}{t}\sum_{\tau=1}^{t}\mathbf{P}^{\tau}$, as follows:

$$\mathbf{K}_{MD} = \mathbf{Z}(t)\mathbf{Z}^{\top}(t) \tag{3}$$

Regularized Laplacian kernel
Another popular graph node kernel function used in graph mining is the regularized Laplacian kernel $\mathbf{K}_{RL}$. This kernel function was introduced by Chebotarev and Shamis in [6] and represents a normalized version of the random walk with restart model. It is defined as follows:

$$\mathbf{K}_{RL} = \sum_{n=0}^{\infty}\beta^{n}(-\mathbf{L})^{n} = (\mathbf{I} + \beta\mathbf{L})^{-1} \tag{4}$$

where the parameter β is again the diffusion parameter. $\mathbf{K}_{RL}$ counts the paths connecting two nodes on the graph induced by taking $-\mathbf{L}$ as the adjacency matrix, regardless of the path length. Thus, a non-zero value is assigned to any couple of nodes as long as they are connected by any indirect path. $\mathbf{K}_{RL}$ remains a relatedness measure even when diffusion factor is large, by virtue of the negative weights assigned to self-loops.

2.2 Link Prediction

2.3 Conjunctive Disjunctive Node Kenrel

2.4 Link Enrinchment for Strengthening Diffusion-based Kernel

3 Evaluation

We perform an empirical evaluation of the predictive performance of several kernel based methods on two of the databases used in [4].

BioGPS: a gene co-expression network is constructed from BioGPS dataset, which contains 79 tissues, measured with the Affymetrix U133A array. Edges are inserted when the pairwise Pearson correlation coefficient (PCC) between genes is larger than 0.5.

HPRD: a database of curated proteomic information pertaining to human proteins. It is derived from [1] with 9465 vertices and 37039 edges. We enploy the HPRD version used in [4] in which they remove some vertices to have 7311 vertices at the end.

To evaluate the performance of graph node kernels we analyze the *gene prioritization*, i.e. given a set of genes known to be associated to a given disease,

gene prioritization is the task to rank the candidate genes based on their probabilities to be related to that disease. Similar to the evaluation process used in [4], we choose 12 diseases with at least 30 confirmed genes. For each disease, we construct a positive set $\mathcal{P}$ with all confirmed disease genes, and a negative set $\mathcal{N}$ which contains random genes associated at least to one disease class which is not related to the class that is defining the positive set. In [4] the ratio between the dataset sizes is chosen as $|\mathcal{N}| = \frac{1}{2}|\mathcal{P}|$. The predictive performance of each method is evaluated via a leave-one-out cross validation: one gene is kept out in turn and the rest are used to train an SVM model. We compute a decision score q_i for the test gene g_i as the top percentage value of score s_i among all candidate gene scores. We collect all decision scores for every gene in the training set to form a global decision score list on which we compute the AUC ROC.

Model Selection. The hyper parameters of the various methods are set using a k-fold on a dataset set that is then never used in the predictive performance estimate. We try the values for diffusion parameter in DK and MED in $\{10^{-3}, 10^{-3}, 10^{-2}, 10^{-1}\}$, time steps in MD in $\{1, 10, 100\}$ and RL parameter in $\{1, 4, 7\}$. For CDNK, we try for the degree threshold values in $\{10,\ 15,\ 20\}$, clique size threshold in $\{4,\ 5\}$, maximum radius in $\{1,\ 2\}$, maximum distance in $\{2,\ 3,\ 4\}$. Finally, the C of SVM is searched in $\{10^{-5},\ 10^{-4},\ 10^{-3},\ 10^{-2},\ 10^{-1},\ 1,\ 10,\ 10^{2}\}$.

4 Results and Discussion

Table 2 shows the AUC performance of the models trained by using different graph node kernels on 11 gene-disease association problems using the BioGPS and Pathways datasets to materialize the gene relation network. In the table, the best result on each disease is marked in bold. We note that CDNK ranks first in 7 out of 11 cases using both networks. CDNK is the top performant kernel also when considering the average AUC ROC and the average rank with 73.3/2.0, 76.5/1.8, with a difference w.r.t. the second best of 5.5% and 1% on BioGPS and Pathways, respectively. MED and RL show similar and moderate results with small gap between them. DK and MD exhibit modest performance on average and are ranked last in many cases: 7 times out of 11 for MD in BioGPS and 10 out of 11 for DK in Pathways. Finally note that although CDNK has state of the art performance, there are cases that would lead to a significant decrease in the quality of the results, namely when networks have high *average* node degree as this would lead to very sparse and fragmented decompositions.

5 Conclusions

We have shown how decomposing a network in a set of connected sparse graphs allows us to take advantage of the discriminative power of CDNK, a novel decomposition kernel, to achieve state-of-the-art results. In future work we will investigate how to 1) decompose networks in a data driven way and 2) extend the CDNK approach to gene-disease association problems exploiting multiple

heterogeneous information sources.

HPRD Dataset								
	LEDK		**MEDK**		**MDK**		**RLK**	
Disease	No_Link	Link	No_Link	Link	No_Link	Link	No_Link	Link
0	73/2	74/1	71/1	71/2	71/1	69/2	73/2	73/1
1	68/2	69/1	54/2	55/1	82/2	82/1	80/2	80/1
2	77/1	75/2	72/1	70/2	79/1	77/2	81/1	76/2
3	60/2	62/1	62/2	63/1	63/2	66/1	65/2	67/1
4	67/2	68/1	65/2	67/1	68/2	70/1	68/1	68/2
5	67/1	67/2	69/1	68/2	65/2	68/1	66/2	70/1
6	87/2	88/1	87/2	87/1	87/2	89/1	87/2	88/1
7	79/2	80/1	78/2	79/1	81/1	78/2	79/1	77/2
8	78/2	79/1	72/2	72/1	79/1	75/2	80/2	80/1
9	73/2	76/1	71/2	73/1	70/2	72/1	70/2	77/1
10	80/1	79/2	82/1	81/2	77/1	75/2	80/2	80/1
11	76/1	75/2	75/2	75/1	83/2	87/1	84/1	83/2
AUC	74/1.67	74/1.33	71/1.67	72/1.33	75/1.58	76/1.42	76/1.67	77/1.33
		0.76		0.73		0.76		0.77

Table 1: *Predictive performance on 12 gene-disease associations using network induced by the HPRD. We report the AUC-ROC (%) and the rank for each kernel method.*

HPRD Dataset								
	LEDK		**MEDK**		**MDK**		**RLK**	
Disease	No_Link	Link	No_Link	Link	No_Link	Link	No_Link	Link
0	73/2	76/1	76/2	79/1	70/1	69/2	70/1	67/2
1	60/1	61/2	59/1	58/2	60/2	61/1	60/2	72/1
2	85/1	84/2	85/1	84/2	82/1	82/2	82/2	84/1
3	66/2	67/1	56/2	60/1	66/2	69/1	66/2	67/1
4	61/2	62/1	63/1	62/2	58/2	58/1	58/2	58/1
5	70/1	69/2	70/1	69/2	67/2	68/1	67/2	68/1
6	73/1	71/2	68/1	68/2	69/1	69/2	69/2	76/1
7	69/2	69/1	69/2	69/1	67/2	67/1	67/1	65/2
8	73/2	74/1	70/2	71/1	65/2	68/1	65/2	67/1
9	69/1	67/2	65/2	66/1	64/2	65/1	64/1	64/2
10	58/2	60/1	52/2	56/1	60/2	61/1	60/2	70/1
11	69/2	73/1	70/2	70/1	60/2	61/1	60/2	62/1
$\frac{\overline{AUC}}{\overline{Rank}}$	69/1.58	69/1.42	67/1.58	68/1.42	66/1.75	67/1.25	66/1.75	68/1.25
		72		69		68		71

Table 2: *Predictive performance on 12 gene-disease associations using network induced by the BioGPS. We report the AUC-ROC (%) and the rank for each kernel method.*

References

[1] Prasad, TS Keshava, et al. Human protein reference database-2009 update. Nucleic acids research 37. suppl 1 (2009): D767-D772.

[2] Haussler, David. Convolution kernels on discrete structures. Vol. 646. Technical report, Department of Computer Science, University of California at Santa Cruz, 1999.

[3] Kondor, Risi Imre, and John Lafferty, Diffusion kernels on graphs and other discrete input spaces. *ICML*. Vol. 2. 2002.

[4] Chen. B, et al. Disease gene identification by using graph kernels and Markov random fields. Science China Life Sciences 57.11 (2014): 1054-1063.

[5] Fouss F, et al. An experimental investigation of graph kernels on a collaborative recommendation task. Proceedings of the 6th international conference on data mining 2006. ICDM 2006, 863-868.

[6] Chebotarev P and Shamis E. The matrix forest therem and measuring relations in small social groups. Automation and Remote Control 1997, 58(9):1505-1514.

[7] C. Fabrizio, and K. De Grave, Fast neighborhood subgraph pairwise distance kernel. Proceedings of the 26th, International Conference on Machine Learning. Omnipress, 2010.

[8] Goh KI et al, The human disease network. Proceedings of the National Academy of Sciences 104.21 (2007): 8685-8690.

[9] A. Hamelin, et al, K-core decomposition: A tool for the visualization of large scale networks. arXiv preprint cs/0504107 (2005).

[10] R. E. Tarjan, Decomposition by clique separators, Discrete Math, vol. 55, no. 2, pp. 221-232, July. 1985.

[11] Gene Ontology Consortium. The Gene Ontology (GO) database and informatics resource. Nucleic acids research,(2004) 32(suppl 1), D258-D261.