AOLGK-A MODIFICATION OF NSPDK FOR GRAPH NODE SIMILARITY IN THE DENSE GRAPHS

1. Motivation

In genetic data-based predictive systems, the definition of gene-gene similarity is one of the key aspects that determines the performance of the systems. A popular approach is to use a domain specific notion of relatedness between genes and materialize a graph where nodes are genes and edges are relationships that satisfy a stringent criterion (e.g. if the products of the two genes interact in a known pathway). Once the graph is materialized, one can define the node similarities by exploiting graph properties.

Concerning graph node similarity measurement, graph node kernels (kernels) are known as the most effective paradigm to use. In the last decade, a number of kernels have been proposed. Most kernels are based on the natural transitive property (i.e. connecting two nodes via a path). For instance, diffusion kernel takes into account all the paths connecting two nodes to compute the similarity. On the positive side, these approaches can be employed on networks that: 1) are dense, 2) have nodes with high maximal degrees, 3) have cliques. On the negative side, these approaches do not have a high discriminative capacity. There are two reasons for this: 1) they are based on considering in an additive and independent fashion the contribution of each information source and, 2) they employ a representation that is aware of the distance between nodes but cannot model the precise configuration of the context (i.e. the neighboring sub-graph).

In order to increase the modelling discriminative capacity we propose to employ decomposition graph kernels (DGK). On one side, these approaches are computationally efficient (quasi linear in the network size) and can adequately represent contextual information. On the other side, they can be employed on networks that are sparse and have a small (3-5) maximal node degree. In this paper, we employ the idea from NSPDK [1] to propose a new kernel named AOLGK that can work with dense graph meanwhile it still preserves the advantage of DGK.

2. AOLGK

3. Evaluation

In this section, we aim at evaluating the performance of AOLGK and comparing with some state of the art kernels.

3.1 Data source

For the first attempt, we choose the dataset named BioGPS mentioned in [7]. The constructed graph contains 7311 nodes and 911,294 edges. It is a dense graph and includes cliques, high average node degree.

3.2 Results

Disease	AOLGK	LEDK	MDK	RLK	MEDK
Cardiovascular	0.70/3	0.62/5	0.63/4	0.72/1.5	0.72/1.5
Connective	0.60/1	0.58/2	0.57/4	0.57/4	0.57/4
Dermatological	0.76/2	0.68/5	0.79/1	0.72/3	0.71/4
Developmental	0.89/1	0.88/2	0.74/5	0.84/3.5	0.84/3.5
Endocrine	0.65/5	0.73/2	0.74/1	0.70/3	0.69/4
Hematological	0.87/1	0.84/4	0.83/5	0.85/2.5	0.85/2.5
Immunological	0.76/5	0.78/3	0.77/4	0.80/1.5	0.80/1.5
Metabolic	0.79/2	0.76/5	0.78/4	0.79/2	0.79/2
Muscular	0.69/5	0.74/4	0.78/1	0.77/2	0.76/3
Ophthalmological	0.79/1	0.75/4	0.75/4	0.77/2.5	0.77/2.5
Renal	0.76/1	0.70/5	0.75/2	0.72/4	0.73/3
Skeletal	0.71/3	0.75/2	0.77/1	0.66/4.5	0.66/4.5
Average	0.75/2.5	0.73/3.6	0.74/3.0	0.74/2.8	0.74/3.0

Table 1: Performance of kernels on different genetic diseases

In which:

- LEDK: Laplacian exponential diffusion kernel
- MDK: Markov diffusion kernel
- RLK: Regularized Laplacian kernel
- MEDK: Markov exponential diffusion kernel

References

[1] Smalter A, Lei SF, Chen XW. Human disease-gene classification with integrative sequence-based and topological features of protein-protein interaction networks. In Bioinformatics and Biomedicine, 2007. BIBM 2007. IEEE International Conference on (pp. 209-216). IEEE.

[2] Yang P, Li X, Chua HN, Kwoh CK, Ng SK. Ensemble Positive Unlabeled Learning for Disease Gene Identification. PLoS ONE 2014, 9(5): e97079. doi:10.1371/journal.pone.0097079.

[3] Chen B, Li M, Wang J, Shang X, Wu FX. A fast and high performance multiple data integration algorithm for identifying human disease genes. BMC Medical Genomics 2015, 8(Suppl 3):S2.

[4] Khler S, Bauer S, Horn D, Robinson PN. Walking the interactome for prioritization of candidate disease genes. The American Journal of Human Genetics 82.4 (2008): 949-958.

[5] Chen Y et al. In silico gene prioritization by integrating multiple data sources. PloS one 6.6 (2011): e21137.

[6] Mordelet F, Vert JP. ProDiGe: Prioritization Of Disease Genes with multitask machine learning from positive and unlabeled examples. BMC Bioinformatics 2011, 12(1): 389.

[7] Chen BL et al. Disease gene identification by using graph kernels and Markov random fields. Science China Life Sciences, 2014.

[8] McKusick VA. Mendelian Inheritance in Man and its online version, OMIM. The American Journal of Human Genetics 80.4 (2007): 588-604.

[9] Costa et al. Fast neighborhood subgraph pairwise distance kernel. Proceedings of the 26th International Conference on Machine Learning. Omnipress, 2010.

[10] Aiolli F, Donini M. EasyMKL: a scalable multiple kernel learning algorithm. Neurocomputing 2015, http://dx.doi.org/10.1016/j.neucom.2014.11.078.

[11] Aiolli F, Da San Martino G, Sperduti A. A kernel method for the optimization of the margin distribution. Proceedings of the ICANN Conference, Praga 2008.