Learning kernels from biological networks by maximizing entropy

Koji Tsuda, [1] **William Stafford Noble** [2]

Keywords: function prediction, protein interaction networks, metabolic pathways, support vector machines, diffusion kernels

1 Introduction

When predicting the functions of unannoted proteins based on a protein network, one relies on some notions of "closeness" or "distance" among the nodes. However, inferring closeness among the nodes is an extremely ill-posed problem, because the proximity information provided by the edges is only local. Moreover, it is preferable that the resulting similarity matrix be a valid *kernel matrix* so that function prediction can be done by support vector machines (SVMs) or other high-performance kernel classifiers [2]. Maximum entropy methods have been proven to be effective for solving general ill-posed problems. However, these methods are concerned with the estimation of a probability distribution, not a kernel matrix. In this work, we generalize the maximum entropy framework to estimate a positive definite kernel matrix.

We found that the *diffusion kernel* [1], which has been used successfully for making predictions from biological networks (e.g. [3]), can be derived from this framework. However, one drawback inherent in the diffusion kernel is that, in the feature space, the distances between connected samples have high variance. As a result, some of the samples are *outliers*, which should be avoided for reliable statistical inference. Our new kernel based on local constraints resolves this problem and thereby shows better accuracy in yeast function prediction.

2 Locally Constrained Diffusion Kernels

SVMs work by embedding samples into a vector space called a *feature space*, and searching for a linear discriminant function in such a space [2]. If we have an undirected graph with n nodes and m edges, the n nodes in a graph are mapped to n points in the feature space $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n \in \mathcal{F}$. The embedding is defined implicitly by specifying an inner product via a positive definite kernel matrix $K_{ij} = \boldsymbol{x}_i^\top \boldsymbol{x}_j, i, j = 1, \cdots, n$. Because the discriminant function is solely represented by inner products, we do not need to have an explicit representation of $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n$. Once a kernel matrix is determined, the (squared) Euclidean distance between two points can also be computed as $D_{ij} := \|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2 = K_{ii} + K_{jj} - 2K_{ij}$.

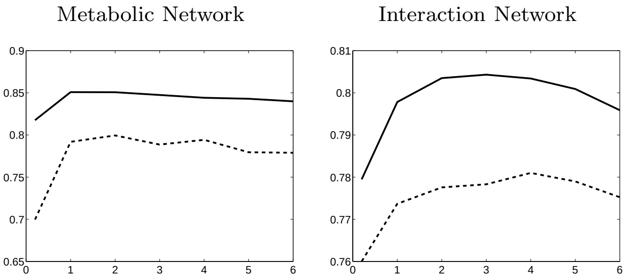
We have found that the matrix of the diffusion kernel [1] can be derived as the optimal solution of the following maximum entropy problem:

$$\min_K \operatorname{tr}(K \log K), \quad \operatorname{tr}(K) = 1, \operatorname{tr}(KL) \leq c,$$

where log denotes the matrix logarithm operation, c is a positive constant, and L is the graph Laplacian matrix [1]. Let $\{s_j, t_j\}_{j=1}^m$ denote the node pairs connected by m edges.

[1] MPI for Biological Cybernetics, Spemannstr. 38, 72076 Tübingen, Germany, and AIST CBRC, Tokyo, Japan. E-mail: koji.tsuda@tuebingen.mpg.de

[2] Dept. of Genome Sciences and Dept. of Computer Science, University of Washington, 1705 NE Pacific St., Seattle, WA 98109, USA E-mail: noble@gs.washington.edu


Metabolic Network

Interaction Network

Figure 1: **Mean ROC score as a function of the diffusion parameter.** The plots show the mean ROC scores computed across the set of CYGD categories, using (A) the metabolic network and (B) the protein-protein interaction network. The solid and broken lines correspond to the new and conventional kernels, respectively.

The quantity $\mathrm{tr}(KL)$ equals the sum of Euclidean distances between connected samples: $\mathrm{tr}(KL) = \sum_{j=1}^{m} \|\boldsymbol{x}_{s_j} - \boldsymbol{x}_{t_j}\|^2$. The objective function corresponds to the (negative) von Neuman entropy. In order to impose a more uniform network structure, we consider the following *local constraints*:

$$\min_K \mathrm{tr}(K \log K), \quad \mathrm{tr}(K) = 1, \mathrm{tr}(KV_j) \leq \gamma, \quad j = 1, \cdots, m, \tag{1}$$

where $\mathrm{tr}(KV_j) = \|\boldsymbol{x}_{s_j} - \boldsymbol{x}_{t_j}\|^2$ corresponds to the Euclidean distance between each pair of connected samples.

3 Experiments

We computed kernels from two different types of yeast biological networks. The first network was derived by [3] from the LIGAND database of chemical reactions in biological pathways. The second network was created by [4] from protein-protein interactions. We tested the kernels' utility in the context of an SVM classification task. We used as a gold standard the functional categories of the MIPS Comprehensive Yeast Genome Database. We selected all functional categories containing at least 30 positive examples resulting in 36 categories for the metabolic network and 76 categories for the protein-protein interaction network. Figure 1 compares the classification performance of SVMs. The figure shows that, for both types of network, our new kernel out-performs the conventional diffusion kernel.

References

[1] I. Kondor and J. Lafferty. Diffusion kernels on graphs and other discrete structures. In *Proceedings of ICML 2002)*, pages 315–322. Morgan Kaufmann, 2002.

[2] B. Schölkopf and A. J. Smola. *Learning with Kernels.* MIT Press, Cambridge, MA, 2002.

[3] J.P. Vert and M. Kanehisa. Graph-driven features extraction from microarray data using diffusion kernels and kernel CCA. In *Advances in Neural Information Processing Systems 15*, volume 15, pages 1425–1432. MIT Press, 2003.

[4] C. von Mering, *et al.* Comparative assessment of large-scale data sets of protein-protein interactions. *Nature*, 417:399–403, 2002.