A new theoretical angle to semi-supervised output kernel regression for protein-protein interaction network inference

Celine Brouard, Florence d'Alché-Buc, Marie Szafranski

A New Theoretical Angle to Semi-supervised Output Kernel Regression for Protein-protein Interaction Network Inference

Céline Brouard[1], Florence d'Alché-Buc[1], and Marie Szafranski[2,1]

[1] IBISC, EA 4526, Université d'Évry Val d'Essonne, F–91025 Évry cedex, France
{celine.brouard, florence.dalche, marie.szafranski}@ibisc.fr
[2] ÉNSIIE, F–91025 Évry cedex, France

1 Background

Recent years have witnessed a surge of interest for network inference in biological networks. *In silico* prediction of protein-protein interaction (PPI) networks is motivated by the cost and the difficulty to experimentally detect physical interactions between proteins. The underlying hypothesis is that some input features relative to the proteins provide valuable information about the presence or the absence of a physical interaction. The main approaches devoted to this task fall into two families: supervised approaches, which aim at building pairwise classifiers able to predict if two proteins interact, from a dataset of labeled pairs of proteins [1–5], and matrix completion approaches that fits into an unsupervised setting with some constraints [6, 7] or directly into a semi-supervised framework [8, 9].

Let us define $\mathcal{O}$ the set of descriptions of the proteins we are interested in. In this paper, we have chosen to convert the binary pairwise classification task into an output kernel learning task as in [3, 4]. This is made possible by noticing that a Gram matrix K_{Y_ℓ} on the training data $\mathcal{O}_\ell$ can be defined from the adjacency matrix using any kernel that encodes the proximities of proteins in the network (for instance a diffusion kernel [10]). We assume that a positive definite kernel κ_y: $\mathcal{O} \times \mathcal{O} \to \mathbb{R}$ underlies this Gram matrix such that $\forall i, j \leq \ell, K_{Y_\ell}(i, j) = \kappa_y(o_i, o_j)$. Moreover, there exists an Hilbert space $\mathcal{F}_y$, called the feature space, and a feature map $y : \mathcal{O} \to \mathcal{F}_y$ such that $\forall (o, o') \in \mathcal{O}, \kappa_y(o, o') = \langle y(o), y(o') \rangle_{\mathcal{F}_y}$.
The assumption underlying output kernel learning is that an approximation of κ_y will provide valuable information about the proximity of proteins in terms of nodes in the interaction graph. This approximation is built from the inner product between the outputs of a single variable function $h : \mathcal{O} \to \mathcal{F}_y$: $\widehat{\kappa_y}(o, o') = \langle h(o), h(o') \rangle_{\mathcal{F}_y}$. This allows one to reduce the problem of learning from pairs to learning a single variable function with values in the output feature space. This supervised regression task is referred to as Output Kernel Regression (OKR). Once the output kernel is learnt, a classifier f_θ is defined from the approximation $\widehat{\kappa_y}$ by thresholding its output values:

$$f_\theta(o, o') = \mathrm{sgn}(\widehat{\kappa_y}(o, o') - \theta) \, .$$

2 RKHS for vector-valued functions for supervised and semi-supervised OKR

In the case of OKR, the function to be learnt is not real-valued but vector-valued in the output Hilbert space. If we want to benefit from the theoretical framework of Reproducing Hilbert Space theory (RKHS), well appropriate to regularization, we need to turn to the proper RKHS theory, devoted to vector-valued functions, which was introduced in [13] and developed in [14]. In this theory, kernels are operator-valued and applied to vectors in the given output Hilbert space. While being very powerful, this theory is still underused.

Supervised setting

In this work, the RKHS theory devoted to functions with values in a Hilbert space provides us with a general framework for OKR. Let $\mathcal{F}_y$ be an Hilbert space. Let $S_\ell = \{(o_i, \mathbf{y}_i)\}_{i=1}^{\ell} \subseteq \mathcal{O} \times \mathcal{F}_y$ be a set of labeled examples, and $\mathcal{H}$ be a RKHS with reproducing kernel $\mathcal{K}_x$. We focus here on the to penalized least square cost in the case of vector-valued functions:

$$\underset{h \in \mathcal{H}}{\operatorname{argmin}} J(h) = \sum_{i=1}^{\ell} \|h(o_i) - \mathbf{y}_i\|_{\mathcal{F}_y}^2 + \lambda_1 \|h\|_{\mathcal{H}}^2 \text{ , with } \lambda_1 > 0. \tag{1}$$

Michelli & Pontil [14] have shown that the minimizer of this problem admits an expansion $\hat{h}(\cdot) = \sum_{j=1}^{\ell} \mathcal{K}_x(o_j, \cdot)\mathbf{c}_j$, where the vectors $\mathbf{c}_j \in \mathcal{F}_y, j = \{1, \cdots, \ell\}$, satisfy the equations:

$$\mathbf{y}_j = \sum_{i=1}^{\ell} \mathcal{K}_x(o_i, o_j)\mathbf{c}_i + \lambda_1 \mathbf{c}_j \ . \tag{2}$$

To benefit from this theory, we must define a suitable input operator-valued kernel. OKR is extended to data described by some input scalar kernel. The training input set is now defined by an input Gram matrix K_{X_ℓ}, which encodes for the properties of the training objects $\mathcal{O}_\ell$. As in the output case, the coefficients of the Gram matrix are supposed to be defined from a positive definite input kernel function $\kappa_x : \mathcal{O} \times \mathcal{O} \to \mathbb{R}$, with $\forall i, j \leq \ell, K_{X_\ell}(i, j) = \kappa_x(o_i, o_j)$. We define an operator-valued kernel $\mathcal{K}_x$ from this scalar kernel:

$$\mathcal{K}_x(o, o') = \kappa_x(o, o') \times I_{\mathcal{F}_y} \ , \tag{3}$$

with $I_{\mathcal{F}_y}$, the identity matrix of size $dim(\mathcal{F}_y) \times dim(\mathcal{F}_y)$. The theorem from [13, 14] ensures that a RKHS can be built from it. Starting from the results existing in the supervised case for the penalized least-square cost, we show that with this choice of the operator-valued kernel, we can derive a closed-form solution.

Proposition 1. *When $\mathcal{K}_x$ is defined by mapping* (3)*, the solution of Problem* (1) *reads*

$$C = Y_\ell (K_{X_\ell} + \lambda_1 I_\ell)^{-1} \ , \tag{4}$$

where $Y_\ell = (\mathbf{y}_1, \cdots, \mathbf{y}_\ell)$, $C = (\mathbf{c}_1, \cdots, \mathbf{c}_\ell)$, and I_ℓ is the $\ell \times \ell$ identity matrix.

It is worth noting that we directly retrieve the extension of kernel ridge regression to output kernels proposed by [15].

Semi-supervised setting

In biology, it is much easier to get a detailed description of the properties of a protein compared to the cost of experimental methods used to detect physical interactions between two proteins. To benefit from the usually large amount of unlabeled data, we need to extend OKR to semi-supervised learning. A powerful approach is based on graph-based regularization that forces the prediction function to be smooth on the graph describing similarities between inputs. Enforcing smoothness of the function permits to propagate output labels over close inputs as shown in [11, 12]. [12] have proposed to explicitly embed such ideas into the framework of regularization within RKHS for real-valued functions.

Let $S_\ell = \{(o_i, \mathbf{y}_i)\}_{i=1}^{\ell}$ be a set of labeled examples and $S_u = \{o_i\}_{i=\ell+1}^{\ell+u}$ a set of unlabeled examples. Let $\mathcal{H}$ be a RKHS with reproducing kernel $\mathcal{K}_x$, and a symmetric matrix W with positive values measuring the similarity of objects in the input space. We consider the following optimization problem:

$$\underset{h\in\mathcal{H}}{\operatorname{argmin}}\, J(h) = \sum_{i=1}^{\ell} \|h(o_i) - \mathbf{y}_i\|^2_{\mathcal{F}_y} + \lambda_1 \|h\|^2_{\mathcal{H}} + \lambda_2 \sum_{i,j=1}^{\ell+u} W_{ij} \|h(o_i) - h(o_j)\|^2_{\mathcal{F}_y} \ , \quad (5)$$

with λ_1 and $\lambda_2 > 0$.

We state and prove a new representer theorem devoted to semi-supervised learning in RKHS with vector-valued functions:

Theorem 1. *The minimizer $\hat{h}$ of the optimization problem* (5) *admits an expansion $\hat{h}(\cdot) = \sum_{j=1}^{\ell+u} \mathcal{K}_x(o_j, \cdot)\mathbf{c}_j$, where the vectors $\mathbf{c}_j \in \mathcal{F}_y, j = \{1, \cdots, (\ell+u)\}$ satisfy the equations:*

$$V_j \mathbf{y}_j = V_j \sum_{i=1}^{\ell+u} \mathcal{K}_x(o_i, o_j)\mathbf{c}_i + \lambda_1 \mathbf{c}_j + 2\lambda_2 \sum_{i=1}^{\ell+u} L_{ij} \sum_{m=1}^{\ell+u} \mathcal{K}_x(o_m, o_i)\mathbf{c}_m \ . \quad (6)$$

The matrix V_j of dimension $dim(\mathcal{F}_y) \times dim(\mathcal{F}_y)$ is the identity matrix if $j \leq \ell$ and the null matrix if $\ell < j \leq (\ell + u)$. L is the $(\ell+u)\times(\ell+u)$ Laplacian matrix, given by $L = D - W$, where D is a diagonal matrix such that $D_{ii} = \sum_{j=1}^{\ell+u} W_{ij}$.

Using the operator-valued kernel defined previously leads us to define a new model, expressed as a closed-form solution.

Proposition 2. *When $\mathcal{K}_x$ is defined by mapping* (3)*, the solution of Problem* (5) *reads*

$$C = Y_\ell U (K_{X_{\ell+u}} U^T U + \lambda_1 I_{\ell+u} + 2\lambda_2 K_{X_{\ell+u}} L)^{-1}, \quad (7)$$

where $Y_\ell = (\mathbf{y}_1, \cdots, \mathbf{y}_\ell)$, $C = (\mathbf{c}_1, \cdots, \mathbf{c}_{\ell+u})$. *$U$ denotes a* $\ell \times (\ell+u)$ *matrix that contains an identity matrix of size* $\ell \times \ell$ *on the left hand side and a zero matrix of size* $\ell \times u$ *on the right hand side.* $K_{X_{\ell+u}}$ *is the Gram matrix of size* $(\ell+u)\times(\ell+u)$ *associated to kernel* κ_x. *Finally,* $I_{\ell+u}$ *is the identity matrix of size* $(\ell+u)$.

3 Experiments

We extensively studied the behavior of the provided models on transductive link prediction using artificial data and a protein-protein interaction network dataset.

Synthetic networks We illustrate our method on synthetic networks in order to measure the improvement brought by the semi-supervised method in extreme cases (i.e. for low percentage of labeled proteins) when the input kernel is a very good approximation of the output kernel. We produce the data by sampling random graphs from a Erdős-Renyi law with different probabilities of presence of edges. The input feature vectors have been obtained by applying Kernel PCA on the diffusion kernel associated with the graph. Finally, we use the components that capture 95% of the variance to define the input features. We observe from the results obtained that the semi-supervised approach improves upon the supervised one on Auc-Roc and Auc-Pr, especially for a small percentage of labeled data (up to 10%). Based on these results one can formulate the hypothesis that supervised link prediction is harder in the case of more dense networks and that the contribution of unlabeled data seems more helpful in this case. One can also assume that using unlabeled data increases the AUCs for low percentage of labeled data. But when enough information can be found in the labeled data, semi-supervised learning does not improve the performance.

Protein-protein interaction network We illustrate our method on a PPI network of the yeast *Saccharomyces Cerevisiae* composed of 984 proteins linked by 2438 interactions. To reconstruct the PPI network, we deal with usual input features that are gene expression data, phylogenetic profiles, protein localization and protein interaction data derived from yeast two-hybrid (see for instance [2–6] for a more complete description).

Table 1. Auc-roc and Auc-pr obtained for the reconstruction of the PPI network from the gene expression data in the supervised and the semi-supervised settings. The percentage values correspond to the proportions of labeled proteins.

Methods	**Auc-roc**			**Auc-pr**		
	5%	**10**%	**20**%	**5**%	**10**%	**20**%
Supervised	76.9 ± 4.3	80.3 ± 0.9	82.1 ± 0.6	5.4 ± 1.6	7.1 ± 1.1	8.1 ± 0.7
Semi-supervised	79.6 ± 0.9	80.7 ± 1.0	81.9 ± 0.7	6.6 ± 1.1	7.6 ± 0.8	8.4 ± 0.5

We experimented with our method in the semi-supervised setting and compared the results with those obtained in the supervised setting. For different

values of ℓ, that is the number of labeled proteins, we randomly sub-sampled a training set of proteins and considered all the remaining proteins for the test set. The interaction assumed to be known are those between two proteins from the training set. We ran each experiment ten times and tuned the hyperparameters by 5-fold cross-validation on the training set. Averaged and standard deviations of the Auc-roc and Auc-pr values when using gene expression data as input features are summarized in Table 1. It is worth noting that the semi-supervised method reaches better performances when the number of labeled proteins is small, which is usually the case in PPI network inference problems.

References

1. Ben-Hur, A., Noble, W.S.: Kernel methods for predicting protein–protein interactions. Bioinformatics, vol. 21, pp. 38–46 (2005)
2. Yamanishi, Y., Vert, J.-P., Kanehisa, M.: Protein network inference from multiple genomic data: a supervised approach. Bioinformatics, vol. 20, pp. 363–370 (2004)
3. Geurts, P., Wehenkel, L., d'Alché-Buc F.: Kernelizing the output of tree-based methods. In: Proc. of the 23th Intl. Conf. on Machine learning (2006)
4. Geurts, P., Touleimat, N., Dutreix, M., d'Alché-Buc, F.: Inferring biological networks with output kernel trees. BMC Bioinformatics, vol. 8 (2007)
5. Bleakley, K., Biau, G., Vert, J.-P.: Supervised reconstruction of biological networks with local models. Bioinformatics, vol. 23, pp. i57–i65 (2007)
6. Kato, T., Tsuda, K., Asai, K.: Selective integration of multiple biological data for supervised network inference. Bioinformatics, vol. 21, pp. 2488–2495 (2005)
7. Tsuda, K., Noble, W. S.: Learning kernels from biological networks by maximizing entropy. Bioinformatics, vol. 20, pp. 326–333 (2004)
8. Kashima, H., Yamanishi, Y., Kato, Ts., Sugiyama, M., Tsuda, K.: Simultaneous inference of biological networks of multiple species from genome-wide data and evolutionary information. Bioinformatics, vol. 25, pp. 2962–2968 (2009)
9. Yip, K. Y., Gerstein, M.: Training set expansion: an approach to improving the reconstruction of biological networks from limited and uneven reliable interactions. Bioinformatics, vol. 25, pp. 243–250 (2009)
10. Kondor, R. I., Lafferty, J. D.: Diffusion Kernels on Graphs and Other Discrete Input Spaces. In: Proc. of the 19th Intl. Conf. on Machine Learning (2002)
11. Zhou, D., Bousquet, O., Lal, T.N., Weston, J., Schölkopf, B.: Learning with Local and Global Consistency. In: Adv. in Neural Information Processing Systems 16 (2004)
12. Belkin, M., Niyogi, P., Sindhwani, V.: Manifold Regularization: A Geometric Framework for Learning from Labeled and Unlabeled Examples. J. Mach. Learn. Res., vol. 7, pp. 2399–2434 (2006)
13. Senkene, E., Tempel'man, A.: Hilbert Spaces of operator-valued functions. Lithuanian Mathematical Journal, vol. 13, pp. 665–670 (1973)
14. Micchelli, C.A., Pontil, M. A.: On Learning Vector-Valued Functions. Neural Computation, vol. 17, pp. 177–204 (2005)
15. Cortes, C.,Mohri, M., Weston, J.: A general regression technique for learning transductions. In: Proc. of the 22nd Intl. Conf. on Machine Learning, pp 153–160 (2005)