Semi-supervised Penalized Output Kernel Regression for Link Prediction

Céline Brouard¹, Florence d’Alché-Buc¹, Marie Szafranski¹,²
celine.brouard@ibisc.fr, florence.dalche@ibisc.fr, marie.szafranski@ibisc.fr

www.amis.ibisc.fr

¹IBISC EA 4526, Université d’Évry Val d’Essonne, F-91025 Évry cedex, France
²ÉNSIIE, F–91025 Évry cedex, France
Supervised link prediction

- $\mathcal{O}$: set of descriptions of the nodes (individuals, proteins, authors)
- Learning a predictor $f: \mathcal{O} \times \mathcal{O} \to \{0, 1\}$ from $\mathcal{G}_\ell = (\mathcal{O}_\ell, A_\ell)$ - with $\mathcal{O}_\ell \subseteq \mathcal{O}$ and $A_\ell$ the adjacency matrix of the known subgraph: $A_\ell(i, j) = f(o_i, o_j)$ for $i, j = 1, \ldots, \ell$

Existing works: Pairwise SVM [Ben-Hur and Noble, 2005], mixture of feature experts [Qi, 2008], KCCA [Yamanishi et al., 2004], metric learning [Yamanishi and Vert, 2005], output kernel regression tree [Geurts et al., 2006;2007]
Semi-supervised link prediction

Develop methods that exploit unlabeled data

- Let $\mathcal{O}_{\ell+u} = \{o_1, \ldots, o_{\ell+u}\}$ be a sample: $\ell$ fully labeled nodes, $u$ unlabeled nodes
- Learning a predictor $f: \mathcal{O} \times \mathcal{O} \rightarrow \{0, 1\}$ from $\mathcal{O}_{\ell+u}$ and $A_\ell$

Existing works: Kernel Matrix completion using EM [Tsuda et al., 2003], [Kato et al., 2005], Link Propagation [Kashima et al., 2009]
Outline

1. Link prediction with Output Kernel Regression
2. Semi-supervised learning with operator-valued kernels
3. Experiments
Building a classifier $f$ by learning a similarity $\kappa_y$

- $\kappa_y$: similarity between two "objects" as nodes in the known graph

**Similarity-based model:**

$$f_\theta(o, o') = \text{sgn}(\hat{\kappa}_y(o, o') - \theta)$$

- Learning a proxy of $\kappa_y$ and choosing $\theta = \text{learning the classifier } f_\theta$
Building the classifier $f$ by learning an output kernel $\kappa_y$

Additional assumption: an output kernel

Let $\kappa_y : \mathcal{O} \times \mathcal{O} \to \mathbb{R}$ be a positive semi-definite kernel. Then there exists a Hilbert space $\mathcal{F}_y$ with scalar product $\langle \cdot, \cdot \rangle_{\mathcal{F}_y}$ and an associated feature map $y : \mathcal{O} \to \mathcal{F}_y$ such that:

$$\forall (o, o') \in \mathcal{O} \times \mathcal{O}, \kappa_y(o, o') = \langle y(o), y(o') \rangle_{\mathcal{F}_y}$$
Building the classifier $f$ by learning an output kernel $\kappa_{y}$

**Additional assumption: an output kernel**

Let $\kappa_{y} : \mathcal{O} \times \mathcal{O} \rightarrow \mathbb{R}$ be a positive semi-definite kernel. Then there exists a Hilbert space $\mathcal{F}_{y}$ with scalar product $\langle \cdot, \cdot \rangle_{\mathcal{F}_{y}}$ and an associated feature map $y : \mathcal{O} \rightarrow \mathcal{F}_{y}$ such that:

$$\forall (o, o') \in \mathcal{O} \times \mathcal{O}, \kappa_{y}(o, o') = \langle y(o), y(o') \rangle_{\mathcal{F}_{y}}$$

- $K_{Y_{\ell}}$: Gram matrix associated to $\kappa_{y}$
- In this work we use the **diffusion kernel** [Kondor & Lafferty, 2002]) :

$$K_{Y_{\ell}} = \exp(-\beta L_{Y_{\ell}}),$$

where the graph Laplacian is defined by : $L_{Y_{\ell}} = D_{\ell} - A_{\ell}$, with $D_{\ell}$ the degree matrix.
Building the classifier $f$ by learning an output kernel $k_y$

\[ \forall (o, o') \in O \times O, \kappa_y(o, o') = \langle y(o), y(o') \rangle_{\mathcal{F}_y} \]

- Let us learn to predict $y$ with a function $h : O \rightarrow \mathcal{F}_y$
- Then we will get: $\hat{\kappa}_y(o, o') = \langle h(o), h(o') \rangle_{\mathcal{F}_y}$
- learning $h$ : output kernel regression

**OK3** : Output Kernel Regression Trees [Geurts et al., 2006]

=> instead of learning a pairwise classifier, we learn a single variable function with output values in a Hilbert space
1. Link prediction with Output Kernel Regression

2. Semi-supervised learning with operator-valued kernels

3. Experiments
RKHS with operator-valued kernels

Which RKHS theory can be used here for functions whose outputs are in some Hilbert space $\mathcal{F}_y$?

Definition of an Operator-valued kernel

For a given Hilbert space $\mathcal{F}_y$:
- $L(\mathcal{F}_y)$ is the set of all bounded linear operators from $\mathcal{F}_y$ to itself.
- Given $A \in L(\mathcal{F}_y)$, $A^*$ denotes the adjoint of $A$.

Operator-valued kernel:
(Senkene & Tempel’man, 1973 ; Caponnetto et al., 2008)

$\mathcal{K}_x : \mathcal{O} \times \mathcal{O} \rightarrow L(\mathcal{F}_y)$ is an operator-valued kernel if:
- $\forall (o, o') \in \mathcal{O} \times \mathcal{O}$, $\mathcal{K}_x(o, o') = \mathcal{K}_x(o, o')^*$
- $\forall m \in \mathbb{N}$, $\forall \{(o_i, y_i)\}_{i=1}^m \subseteq \mathcal{O} \times \mathcal{F}_y$,
  $$\sum_{j=1}^m \langle y_i, \mathcal{K}_x(o_i, o_j)y_j \rangle_{\mathcal{F}_y} \geq 0.$$
Building a RKHS from an operator-valued kernel

Theorem (Senkene & Tempel’man, 1973 ; Micchelli & Pontil, 2005)

If $\mathcal{K}_x : \mathcal{O} \times \mathcal{O} \rightarrow \mathcal{L}(\mathcal{F}_y)$ is an operator-valued kernel, then there exists a unique RKHS $\mathcal{H}$ which admits $\mathcal{K}_x$ as the reproducing kernel, that is

$$\forall o \in \mathcal{O}, \forall y \in \mathcal{F}_y, \langle h, \mathcal{K}_x(o, \cdot)y \rangle_{\mathcal{H}} = \langle h(o), y \rangle_{\mathcal{F}_y}$$

$\mathcal{H}$ is built from functions of the form: $\sum_i \mathcal{K}_x(o_i, \cdot)a_i$, defining an inner product of functions $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ using the reproducing property and by completing this space by limits of Cauchy sequences.
Theorem in the semi-supervised case (1)

Adding a smoothness constraint: extension of [Belkin et al., 2006]

Theorem (Brouard, d’Alché-Buc and Szafranski)

Given a set of labeled examples \( S_\ell = \{(o_i, y_i)\}_{i=1}^{\ell} \subseteq O \times F_y \), a set of unlabeled examples \( S_u = \{o_i\}_{i=\ell+1}^{\ell+u} \subseteq O \), a RKHS with reproducing kernel \( K_x : O \times O \rightarrow L(F_y) \), and a matrix \( W \) measuring the similarity of objects in the input space, the minimizer \( \hat{h} \) of:

\[
\arg \min_{h \in \mathcal{H}} J(h) = \sum_{i=1}^{\ell} \| h(o_i) - y_i \|_F^2 + \lambda_1 \| h \|_H^2 + \lambda_2 \sum_{i,j=1}^{\ell+u} W_{ij} \| h(o_i) - h(o_j) \|_F^2,
\]

with \( \lambda_1 \) and \( \lambda_2 > 0 \),
Theorem in the semi-supervised case (2)

\( \hat{h} \) admits an expansion:

\[
\hat{h}(\cdot) = \sum_{j=1}^{\ell+u} \mathcal{K}_x(o_j, \cdot) c_j,
\]

where the vectors \( c_j \in \mathcal{F}_y, j = \{1, \cdots, \ell + u\} \) satisfy the equations:

\[
V_j y_j = V_j \sum_{i=1}^{\ell+u} \mathcal{K}_x(o_i, o_j)c_i + \lambda_1 c_j + 2\lambda_2 \sum_{i=1}^{\ell+u} L_{ij} \sum_{m=1}^{\ell+u} \mathcal{K}_x(o_m, o_i)c_m,
\]

where

- 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 \( \ell < j \leq \ell + u \),
- \( L = D - W \), where \( D \) is a diagonal matrix of general term \( D_{ii} = \sum_{j=1}^{\ell+u} W_{ij} \).
We define $\mathcal{K}_x$ as follows:

$$\mathcal{K}_x : \mathcal{O} \times \mathcal{O} \rightarrow \mathcal{L}(\mathcal{F}_y)$$

$$(o, o') \mapsto \kappa_x(o, o') \times I_{\mathcal{F}_y},$$

with $I_{\mathcal{F}_y}$, the identity matrix of dimensions $dim(\mathcal{F}_y) \times dim(\mathcal{F}_y)$.

- $\kappa_x : \mathcal{O} \times \mathcal{O} \rightarrow \mathbb{R}$: input scalar kernel
- Then there exists an Hilbert space $\mathcal{F}_x$ and an associated feature map $x : \mathcal{O} \rightarrow \mathcal{F}_x$ such that:

$$\forall (o, o') \in \mathcal{O} \times \mathcal{O}, \kappa_x(o, o') = \langle x(o), x(o') \rangle_{\mathcal{F}_x}$$
Application of the Representer theorem

Proposition (semi-supervised case):
When $\mathcal{K}_x$ is defined by mapping (4), the solution reads

$$\forall o \in \mathcal{O}, \hat{h}(o) = CX_{\ell+u}^T x(o),$$

with

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

- $Y_\ell = (y(o_1), \ldots, y(o_\ell))^T$, $C = (c_1, \ldots, c_{\ell+u})^T$
- $X_{\ell+u} = (x(o_1), \ldots, x(o_{\ell+u}))^T$
- $K_{X_{\ell+u}} = X_{\ell+u}^T X_{\ell+u}$ is the Gram matrix associated to $\kappa_x$
- $I_{\ell+u}$ is the identity matrix of size $(\ell + u)$
- $U$ denotes a matrix of dimension $\ell \times (\ell + u)$: $U = [I_\ell, 0]$
∀(o, o′) ∈ O × O,

\[ f_\theta(o, o′) = \text{sgn}(\langle \hat{h}(o), \hat{h}(o′) \rangle_{\mathcal{F}_y} - \theta) \, . \]

\[ \langle \hat{h}(o), \hat{h}(o′) \rangle_{\mathcal{F}_y} = \langle CX_{\ell+u}^T x(o), CX_{\ell+u}^T x(o′) \rangle_{\mathcal{F}_y} \]

\[ = x(o)^T X_{\ell+u} B^T K_{Y_{\ell}} BX_{\ell+u}^T x(o′) \, , \]

with

\[ B = U(K_{X_{\ell+u}} U^T U + \lambda_1 I_{\ell+u} + 2\lambda_2 K_{X_{\ell+u}} L)^{-1} \]
Outline

1. Link prediction with Output Kernel Regression
2. Semi-supervised learning with operator-valued kernels
3. Experiments
Experimental protocol

- For different percentages of labeled nodes, we have randomly picked 10 times a subsample of training examples and used the remaining as testing examples:

  ![Diagram showing labeled and unlabeled data split]

  **Note that a 10% selection of labeled nodes actually corresponds to only 1% of labeled interactions.**
Nips co-authorship network

- Reconstruction of the co-authorship network for the authors with a minimum of two links in the network: 2026 authors with an empirical link density of 0.002.
- Input features: frequency of words used by the authors in their papers.
Yeast protein-protein interaction network (1)
[Kato et al., 2005]

- Protein-protein interaction network: 984 proteins, 2478 edges
- Input features: gene expression data

**Supervised setting:**
- Comparison with the results in [Bleakley et al., 2007] using the same procedure of 5−CV

<table>
<thead>
<tr>
<th>Methods</th>
<th>Auc-Roc</th>
</tr>
</thead>
<tbody>
<tr>
<td>em</td>
<td>80.6 ± 1.1</td>
</tr>
<tr>
<td>Pkernel</td>
<td>83.8 ± 1.4</td>
</tr>
<tr>
<td>local</td>
<td>78.1 ± 1.1</td>
</tr>
<tr>
<td>OK3+ET</td>
<td><strong>84.6 ± 1.4</strong></td>
</tr>
<tr>
<td>POKR</td>
<td>83.3 ± 2.1</td>
</tr>
</tbody>
</table>

- Local methods only apply to predict interactions between learning set (LS) and test set (TS) but the other methods can complete both.
Yeast protein-protein interaction network (2)

- **Transductive setting:**

![Graph showing Auc-Roc vs Percentage of labeled proteins](image)

- **Graph Details:**
  - X-axis: Percentage of labeled proteins
  - Y-axis: Auc-Roc
  - Lines represent different methods:
    - Supervised
    - Transductive

*Image credit: Brouard et al. (IBISC)*
Conclusion and Perspectives

- An original way to solve the link prediction problem
- An extension of RKHS theory with operator-valued kernel and a new application
- A contribution to kernel learning literature with output kernel regression (OKR)
- Use other definitions of operator-valued kernels
- Use and learn multiple input kernels
- Extend theoretical results to other cost function and to other application of OKR