Régression semi-supervisée à sortie noyau pour la prédiction de liens

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

Régression semi-supervisée à sortie noyau pour la prédiction de liens

Céline Brouard[1], Florence d'Alché-Buc[1], 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

Résumé : Nous abordons le problème de la prédiction de liens comme une tâche d'apprentissage d'un noyau de sortie en introduisant une méthode de régression semi-supervisée à sortie noyau. En se plaçant dans le cadre de la théorie des espaces de Hilbert à noyau autoreproduisant à valeurs opérateurs pour des fonctions à valeurs vectorielles, nous établissons un nouveau théorème de représentation dédié à la régression semi-supervisée pour un critère des moindres carrés pénalisé. Nous définissons ensuite un noyau à valeur opérateur à partir d'un noyau d'entrée à valeurs scalaires et nous construisons un espace de Hilbert avec celui-ci comme noyau autoreproduisant. La minimisation des moindres carrés pénalisés dans ce cadre conduit à une solution analytique comme dans le cas de la régression ridge. La pertinence de cette nouvelle approche semi-supervisée est étudiée dans le cadre de la prédiction de liens transductive. Des jeux de données artificiels puis deux applications réelles sont traitées en utilisant un très faible pourcentage de données étiquetées.

Mots-clés : **régression à sortie noyau, théorie RKHS, apprentissage semi-supervisé, méthodes à noyaux, prédiction de liens**.

1 Introduction

L'inférence de réseaux, qu'ils soient sociaux ou biologiques, suscite depuis quelques années l'intérêt de la communauté scientifique de l'apprentissage automatique. La prédiction de liens (Huynen *et al.*, 2003; Liben-Nowell & Kleinberg, 2007), définie comme une tâche supervisée, vise à construire des classifieurs capables de prédire si deux objets intéragissent entre eux à partir d'un ensemble de paires d'objets étiquetées. L'hypothèse sous-jacente est que des propriétés en entrée relatives à chaque objet d'une paire peuvent

fournir des informations concernant la présence ou l'absence de lien. Les principales approches dédiées à cette tâche se décomposent en deux familles : les modèles graphiques probabilistes (Miller *et al.*, 2009; Taskar *et al.*, 2003) fournissent des probabilités a posteriori de la présence de liens tandis que les méthodes à noyaux tirent bénéfice de la polyvalence des noyaux à encoder diverses connaissances structurées dans l'espace d'entrée, comme dans l'espace de sortie (Yamanishi *et al.*, 2004; Ben-Hur & Noble, 2005; Geurts *et al.*, 2006, 2007a,b).

Cependant, dans de nombreux domaines, des informations supplémentaires sur les objets, qu'ils soient ou non étiquetés, sont également disponibles. En biologie par exemple, il est relativement facile d'obtenir une description détaillée des propriétés d'une protéine, alors que la détection expérimentale d'interactions physiques entre protéines reste longue et coûteuse. L'utilisation de l'apprentissage semi-supervisé apparaît donc pertinent pour le problème de la prédiction de liens.

L'objectif de ce papier est de développer des méthodes capables d'exploiter des données non étiquetées. Pour cela, nous avons choisi de convertir le problème de classification binaire à partir de paires d'objets en un problème d'apprentissage de noyau de sortie comme Geurts *et al.* (2006, 2007b). L'objectif de l'apprentissage est d'approcher un noyau de sortie cible, supposé encoder la similarité des données en tant que noeuds dans le graphe, en utilisant des descripteurs appropriés en entrée. L'utilisation de l'astuce du noyau dans l'espace de sortie permet ainsi de réduire le problème d'apprentissage à partir de paires d'objets à celui d'apprentissage d'une fonction d'une seule variable à valeurs dans l'espace caractéristique de sortie. Cette tâche de régression supervisée est appelée régression à sortie noyau. Une fois le noyau de sortie appris, la tâche de prédiction de liens est réalisée en seuillant la valeur du noyau pour une paire d'entrées. Des méthodes d'arbre ont été développées pour résoudre la régression à sortie noyau et ont été appliquées à l'inférence supervisée de réseaux biologiques (Geurts *et al.*, 2007a).

Afin de tirer partie des données non étiquetées, nous avons besoin d'étendre la régression à sortie noyau à l'apprentissage semi-supervisé. Une approche existante pour la régression semi-supervisée consiste à forcer la régularité de la fonction de prédiction sur le graphe décrivant les similarités entre entrées, afin de contraindre des objets proches en entrée à l'être également en sortie (Zhou *et al.*, 2004; Belkin & Niyogi, 2004). Belkin *et al.* (2006) ont proposé d'intégrer cette approche dans le cadre de la théorie des Espaces de Hilbert à Noyau Autoreproduisant (EHNA) de fonctions à valeurs réelles, permettant

ainsi de bénéficier d'un théorème de représentation dédié à l'apprentissage semi-supervisé.

Dans le cas de la régression à sortie noyau, la fonction que l'on cherche à apprendre n'est pas à valeurs réelles, mais à valeurs vectorielles dans un espace de Hilbert. Nous avons donc besoin de nous tourner vers la théorie des EHNA correspondante introduite dans (Senkene & Tempel'man, 1973; Micchelli & Pontil, 2005). Au sein de cette théorie, les noyaux sont à valeurs matricielles (ou valeurs "opérateurs") et s'appliquent à des vecteurs de l'espace de Hilbert considéré. Comme dans le cas scalaire, un EHNA peut être construit à partir d'un noyau défini positif et des théorèmes de représentation peuvent être prouvés (Micchelli & Pontil, 2005).

Caponnetto *et al.* (2008) et Argyriou *et al.* (2009) ont développé cette théorie pour résoudre des problèmes d'apprentissage multi-tâches et cette théorie a également été utilisée par Kadri *et al.* (2010) dans le cadre de la régression fonctionnelle.

Dans ce travail, la théorie des EHNA pour des fonctions à valeurs vectorielles nous fournit un cadre général pour la régression à sortie noyau. En partant des résultats existants dans le cas supervisé (Micchelli & Pontil, 2005), nous montrons qu'avec un choix approprié de noyau à valeurs matricielles défini à partir un noyau scalaire, nous retrouvons directement l'extension de la régression ridge aux noyaux de sortie proposée par Cortes *et al.* (2005). Nous proposons un nouveau théorème de représentation dédié à l'apprentissage semi-supervisé qui nous amène à définir un nouveau modèle, exprimé par une solution sous forme close. Nous utilisons le modèle obtenu pour approcher le noyau de sortie dans plusieurs tâches : un ensemble de réseaux artificiels, un réseau de co-publications NIPS et un réseau d'interactions protéine-protéine de la levure.

Dans le reste de ce papier, nous commençons par introduire le cadre existant de la régression à sortie noyau pour la prédiction de liens. Dans la section 3, nous rappelons brièvement la théorie des EHNA dédiée aux fonctions à valeurs dans un espace de Hilbert. La section 4 est consacrée à l'apprentissage supervisé et à la dérivation d'une solution sous forme close dans le cas du critère des moindres carrés pénalisé en choisissant un noyau d'entrée à valeurs matricielles adéquat. La section 5 présente les résultats principaux de l'article : un nouveau théorème de représentation consacré à l'apprentissage semi-supervisé dans le cas des fonctions à valeurs vectorielles et un nouveau modèle en résultant, toujours exprimé sous forme close. Dans la section 6, nous présentons des résultats expérimentaux dans le cadre transductif et dans

la section 7, nous tirons des conclusions et traçons quelques perspectives.

2 Régression à sortie noyau pour la prédiction de liens

Soit $\mathcal{O}$ un ensemble permettant de décrire des objets (personnes, protéines, auteurs). Nous considérons que ces objets sont les nœuds d'un graphe, et nous cherchons à estimer une relation $f_{target} : \mathcal{O} \times \mathcal{O} \to \{0, 1\}$ qui caractérise la présence ou l'absence d'interactions entre ces objets. Ces interactions peuvent être de nature physique si les nœuds représentent des protéines, ou de nature sociale lorsque les nœuds représentent des personnes (par exemple, une relation d'amitié, la publication d'ouvrages communs, etc.).

Lors de la phase d'apprentissage, nous disposons de $\mathcal{G}_\ell = (\mathcal{O}_\ell, A_\ell)$, un graphe non orienté défini par le sous-ensemble $\mathcal{O}_\ell \subseteq \mathcal{O}$ et la matrice d'adjacence A_ℓ de taille $\ell \times \ell$, telle que $A_\ell(i, j) = f_{target}(o_i, o_j)$. Dans le cadre supervisé, la prédiction de liens peut se traduire par l'apprentissage d'un classifeur binaire sur des paires d'objets : un classifieur $f : \mathcal{O} \times \mathcal{O} \to \{0, 1\}$ prédit ainsi si deux objets interagissent ou non à partir des données d'apprentissage $(\mathcal{G}_\ell)$.

Dans ces travaux, nous appréhendons le problème de la prédiction de liens par le biais de méthodes de régression à sortie noyau (Geurts *et al.*, 2006). Il est en effet possible de définir à partir de A_ℓ une matrice de Gram K_{Y_ℓ} permettant de coder la similarité des nœuds d'un graphe. Nous supposons que cette matrice est issue d'une fonction noyau définie positive $\kappa_y : \mathcal{O} \times \mathcal{O} \to \mathbb{R}$, telle que $\forall i, j \leq \ell$, $K_{Y_\ell}(i, j) = \kappa_y(o_i, o_j)$ [1], et qu'il existe un espace de Hilbert $\mathcal{F}_y$, appelé espace des caractéristiques, et une application $y : \mathcal{O} \to \mathcal{F}_y$ telle que $\forall (o, o') \in \mathcal{O}$, $\kappa_y(o, o') = \langle y(o), y(o') \rangle_{\mathcal{F}_y}$.

L'idée sous-jacente à la notion d'apprentissage à sortie noyau consiste à supposer qu'une approximation $\widehat{\kappa_y}$ de la fonction noyau κ_y fournit une information sur la proximité des objets de $\mathcal{O}$ dans le graphe partiellement connu. Étant donnée cette hypothèse, un classifieur f_θ est défini en seuillant les valeurs de l'approximation $\widehat{\kappa_y}$:

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

Cette approximation est construite à partir du produit scalaire entre les sorties d'une fonction d'une variable $h : \mathcal{O} \to \mathcal{F}_y$, ce qui en fait un noyau par

1. En particulier, nous utiliserons ici le noyau de diffusion $K_{Y_\ell} = \exp(-\beta L_{Y_\ell})$, où $L_{Y_\ell} = D_\ell - A_\ell$ et où D_ℓ représente la matrice diagonale des degrés (Kondor & Lafferty, 2002).

construction :

$$\widehat{\kappa_y}(o, o') = \langle h(o), h(o') \rangle_{\mathcal{F}_y} .$$

Ainsi, au lieu d'apprendre un classifieur sur des paires d'objets, nous devons apprendre une fonction d'une variable dont la sortie prend ses valeurs dans un espace de Hilbert (l'espace des caractéristiques des sorties $\mathcal{F}_y$).

Dans la suite de cet article, nous verrons comment la théorie des Espaces de Hilbert à Noyaux Autoreproduisants (EHNA) de fonctions à valeurs vectorielles s'applique aux cadres de l'apprentissage supervisé et semi-supervisé.

3 Théorie des EHNA et régression supervisée à noyau

3.1 EHNA et fonctions à valeurs vectorielles

La théorie des EHNA revêt une importance particulière pour les méthodes d'apprentissage régularisées. Dans le cas des fonctions à valeurs réelles, de nombreux modèles tels que la régression ridge ou les séparateurs à vaste marge sont issus de l'application du théorème de représentation à différentes fonctions de coût. L'extension de cette théorie aux fonctions à valeurs vectorielles permet d'utiliser l'astuce du noyau dans l'espace de sortie. Nous rappelons ici brièvement les éléments de cette théorie en se focalisant plus particulièrement sur la fonction de coût des moindres-carrés pénalisés.

Pour un espace de Hilbert $\mathcal{F}_y$, on note $\mathcal{L}(\mathcal{F}_y)$ l'ensemble de tous les opérateurs linéaires bornés de $\mathcal{F}_y$ sur lui-même. L'adjoint de $A \in \mathcal{L}(\mathcal{F}_y)$ est noté A^*.

Définition 1 (Noyau à valeurs "opérateurs" (Senkene & Tempel'man, 1973; Caponnetto *et al.*, 2008))**.** Soient $\mathcal{O}$ un ensemble et $\mathcal{F}_y$ un espace de Hilbert. $\mathcal{K}_x : \mathcal{O} \times \mathcal{O} \to \mathcal{L}(\mathcal{F}_y)$ est un noyau si :

- $\forall (o, o') \in \mathcal{O} \times \mathcal{O}, \mathcal{K}_x(o, o') = \mathcal{K}_x(o, o')^*$,
- $\forall m \in \mathbb{N}, \forall \{(o_i, \mathbf{y}_i)\}_{i=1}^m \subseteq \mathcal{O} \times \mathcal{F}_y, \sum_{j=1}^m \langle \mathbf{y}_i, \mathcal{K}_x(o_i, o_j)\mathbf{y}_j \rangle_{\mathcal{F}_y} \geq 0$.

Le théorème 1 stipule qu'étant donné un noyau à valeurs matricielles, il est possible de construire l'EHNA correspondant.

Théorème 1 (Senkene & Tempel'man (1973); Micchelli & Pontil (2005))**.** *Soient $\mathcal{O}$ un ensemble et $\mathcal{F}_y$ un espace de Hilbert. Si $\mathcal{K}_x : \mathcal{O} \times \mathcal{O} \to \mathcal{L}(\mathcal{F}_y)$ est un noyau à valeurs matricielles, il existe un EHNA unique $\mathcal{H}_{\mathcal{K}_x}$ admettant $\mathcal{K}_x$ comme noyau reproduisant, c'est-à-dire qui vérifie :*

$$\forall o \in \mathcal{O}, \forall \mathbf{y} \in \mathcal{F}_y, \ \langle h, \mathcal{K}_x(o, \cdot)\mathbf{y} \rangle_{\mathcal{H}_{\mathcal{K}_x}} = \langle h(o), \mathbf{y} \rangle_{\mathcal{F}_y} . \tag{1}$$

Pour alléger les notations, nous utiliserons à présent $\mathcal{H} = \mathcal{H}_{\mathcal{K}_x}$. Par le biais du théorème de représentation, les EHNA fournissent un cadre théorique élégant pour les problèmes d'apprentissage régularisés. Le théorème énoncé ci-dessous s'applique au coût des moindres-carrés pénalisés.

Théorème 2 (Micchelli & Pontil (2005))**.** *Soient $\mathcal{O}$ un ensemble et $\mathcal{F}_y$ un espace de Hilbert. Étant donnés un ensemble d'exemples étiquetés $S_\ell = \{(o_i, \mathbf{y}_i)\}_{i=1}^{\ell} \subseteq \mathcal{O} \times \mathcal{F}_y$ et un EHNA $\mathcal{H}$ de noyau reproduisant $\mathcal{K}_x : \mathcal{O} \times \mathcal{O} \to \mathcal{L}(\mathcal{F}_y)$. Pour $\lambda_1 > 0$, la fonction $\hat{h}$ minimisant le problème d'optimisation :*

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

est telle que

$$\hat{h}(\cdot) = \sum_{j=1}^{\ell} \mathcal{K}_x(o_j, \cdot)\mathbf{c}_j \;, \tag{3}$$

où les vecteurs $\mathbf{c}_j \in \mathcal{F}_y, j = \{1, \cdots, \ell\}$ sont solutions des équations :

$$\mathbf{y}_j = \sum_{i=1}^{\ell} (\mathcal{K}_x(o_i, o_j) + \lambda_1 \delta_{ij})\mathbf{c}_i \;, \tag{4}$$

avec $\delta_{ii} = 1$ et $\forall j \neq i$, $\delta_{ij} = 0$.

Afin de tirer partie de cette théorie pour la régression à sortie noyau, nous devons définir un noyau d'entrée à valeurs matricielles. Nous modifions donc légèrement le cadre de la régression à sortie noyau en faisant l'hypothèse que nous disposons d'un noyau d'entrée à valeurs réelles à partir duquel nous pouvons définir un noyau à valeurs matricielles.

4 Théorie des EHNA pour la régression à entrée et sortie noyau

4.1 Noyau d'entrée à valeurs réelles

Le cadre de la régression à sortie noyau est désormais étendu aux données pouvant être décrites par un noyau d'entrée. L'ensemble d'apprentissage est défini par une matrice de Gram K_{X_ℓ} associée aux objets de l'ensemble d'apprentissage $\mathcal{O}_\ell$. Nous supposons que les coefficients de la matrice de Gram sont issus d'une fonction noyau définie positive $\kappa_x : \mathcal{O} \times \mathcal{O} \to \mathbb{R}$, avec

$\forall i, j \leq \ell, K_{X_{\ell}(i,j)} = \kappa_x(o_i, o_j)$. Étant donné κ_x, il existe un espace de Hilbert $\mathcal{F}_x$ et une application $x : \mathcal{O} \to \mathcal{F}_x$, telle que $\forall (o, o') \in \mathcal{O} \times \mathcal{O}$, $\kappa_x(o, o') = \langle x(o), x(o') \rangle_{\mathcal{F}_x}$. Cependant, contrairement à la fonction κ_y associée au noyau de sortie, la fonction κ_x est supposée connue. Notons que cette extension de la régression à sortie noyau correspond à la première étape de la reformulation de KDE (Kernel Dependency Estimation) de Cortes *et al.* (2005) [2].

4.2 Noyau à valeurs matricielles

Définissons $\mathcal{K}_x$ comme :

$$\begin{aligned} \mathcal{K}_x : \quad & \mathcal{O} \times \mathcal{O} \to \mathcal{L}(\mathcal{F}_y) \\ & (o, o') \mapsto \kappa_x(o, o') \times I_{\mathcal{F}_y} , \end{aligned} \tag{5}$$

où $I_{\mathcal{F}_y}$ est la matrice identité de taille $dim(\mathcal{F}_y) \times dim(\mathcal{F}_y)$. Nous pouvons rapidement montrer que $\mathcal{K}_x$ est un noyau non-négatif : il est symétrique et donc hermitien. De plus, la semi positivité de la fonction κ_x entraîne celle de $\mathcal{K}_x$: $\forall m \in \mathbb{N}, \forall \{(o_i, \mathbf{y}_i)\}_{i=1}^m \subseteq \mathcal{O} \times \mathcal{F}_y$,

$$\sum_{i,j=1}^{m} \langle \mathbf{y}_i, \mathcal{K}_x(o_i, o_j)\mathbf{y}_j \rangle_{\mathcal{F}_y} = \sum_{i,j=1}^{m} \kappa_x(o_i, o_j) \langle \mathbf{y}_i, \mathbf{y}_j \rangle_{\mathcal{F}_y} \geq 0.$$

Le théorème 1 garantit qu'un EHNA $\mathcal{H}_{\mathcal{K}_x}$ peut être construit à partir du noyau à valeurs matricielles défini en 5 en complétant l'ensemble des fonctions s'écrivant : $\sum_{j=1}^{J} \mathcal{K}_x(p_j, \cdot)\mathbf{d}_j$, avec $J > 1$, $p_j \in \mathcal{O}$, $\mathbf{d}_j \in \mathcal{F}_y$, et en le dotant d'un produit scalaire approprié. Le lecteur peut se référer à la preuve du théorème 1 (Senkene & Tempel'man, 1973; Micchelli & Pontil, 2005) pour la construction classique de ce EHNA. Le théorème 2 nous permet alors d'obtenir la solution analytique présentée dans la proposition suivante.

Proposition 3. Lorsque $\mathcal{K}_x$ est défini par l'application (5), la solution du problème (2) s'écrit

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

où $Y_{\ell} = (\mathbf{y}_1^T, \cdots, \mathbf{y}_{\ell}^T)$ est une matrice de dimension $dim(\mathcal{F}_y) \times \ell$. $K_{X_{\ell}}$ est la matrice de Gram de taille $\ell \times \ell$ associée à la fonction noyau κ_x. Enfin, I_{ℓ} est une matrice identité de taille ℓ.

2. La seconde étape consistant à résoudre le problème de pré-image.

Ainsi, la proposition 3 conduit au modèle h :

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

où $X_\ell = (x(o_1)^T, \cdots, x(o_\ell)^T)$ est une matrice de dimension $dim(\mathcal{F}_x) \times \ell$. De plus, soulignons que le théorème 2 et la proposition 3 permettent de retomber sur le modèle linéaire proposé par Cortes *et al.* (2005) dans le cadre de la reformulation de la méthode KDE.

5 Régression semi-supervisée à sortie noyau

Belkin *et al.* (2006, Theorem 2) ont proposé un théorème de représentation pour des fonctions à valeurs réelles dans le cadre de l'apprentissage semi-supervisé. Dans celui-ci, les auteurs ajoutent un terme de régularisation à leur fonction de coût, ayant pour effet de contraindre des objets proches en entrée à l'être également en sortie.

Dans cette section, nous établissons le théorème 4 qui étend simultanément le théorème de Belkin *et al.* (2006) aux fonctions à valeurs vectorielles et le théorème 2 au cadre de l'apprentissage semi-supervisé.

Théorème 4. *Soient $\mathcal{O}$ un ensemble et $\mathcal{F}_y$ espace de Hilbert. Étant donnés un ensemble d'exemples étiquetés $S_\ell = \{(o_i, \mathbf{y}_i)\}_{i=1}^{\ell} \subseteq \mathcal{O} \times \mathcal{F}_y$, un ensemble d'exemples non étiquetés $S_u = \{o_i\}_{i=\ell+1}^{\ell+u} \subseteq \mathcal{O}$, un EHNA $\mathcal{H}$ de noyau reproduisant $\mathcal{K}_x : \mathcal{O} \times \mathcal{O} \to \mathcal{L}(\mathcal{F}_y)$ et W, une matrice à valeurs positives, mesurant la similarité des objets dans l'espace d'entrée*[3]*. Pour $\lambda_1, \lambda_2 > 0$, la fonction $\hat{h}$ minimisant le problème d'optimisation :*

$$\underset{h \in \mathcal{H}}{\operatorname{argmin}} J(h) = \sum_{i=1}^{\ell} \|h(o_i) - \mathbf{y}_i\|_{\mathcal{F}_y}^2 + \lambda_1 \|h\|_{\mathcal{H}}^2 \tag{6a}$$

$$+\lambda_2 \sum_{i,j=1}^{\ell+u} W_{ij} \|h(o_i) - h(o_j)\|_{\mathcal{F}_y}^2 \ , \tag{6b}$$

est telle que

$$\hat{h}(\cdot) = \sum_{j=1}^{\ell+u} \mathcal{K}_x(o_j, \cdot)\mathbf{c}_j \ , \tag{7}$$

3. En particulier, on peut choisir $W_{ij} = \kappa_x(o_i, o_j)$.

où les vecteurs $\mathbf{c}_j \in \mathcal{F}_y, j = \{1, \cdots, (\ell+u)\}$ *sont solutions des équations :*

$$V_j \, \mathbf{y}_j = V_j \sum_{i=1}^{\ell+u} \mathcal{K}_x(o_i, o_j)\mathbf{c}_i \; + \; \lambda_1 \mathbf{c}_j + \; 2\lambda_2 \sum_{i=1}^{\ell+u} L_{ij} \sum_{m=1}^{\ell+u} \mathcal{K}_x(o_m, o_i)\mathbf{c}_m \; , \quad (8)$$

où V_j *de dimension* $dim(\mathcal{F}_y) \times dim(\mathcal{F}_y)$ *est la matrice identité si* $j \leq \ell$ *et la matrice nulle si* $\ell < j \leq (\ell+u)$*, et où* L *est le Laplacien défini par* $L = D - W$*, avec* D *la matrice diagonale de terme général* $D_{ii} = \sum_{j=1}^{\ell+u} W_{ij}$*.*

Éléments de preuve. Le problème (6) admet une solution unique $\hat{h}$ donnée par (7), si pour tout $h \in \mathcal{H}$, $J(h) \geq J(\hat{h})$. Pour montrer cela, nous définissons une fonction $g = h - \hat{h}$ et établissons que $J(h) = J(g + \hat{h}) = J(\hat{h}) + c$, avec $c \geq 0$ [4]. □

Si nous utilisons maintenant le noyau à valeurs matricielles $\mathcal{K}_x$ défini par (5), nous pouvons de nouveau appliquer le théorème 1 et construire l'EHNA correspondant :

Proposition 5. Lorsque $\mathcal{K}_x$ est défini par (5), la solution du problème (6) s'écrit

$$C = Y_\ell U (K_{X_{\ell+u}} U^T U + \lambda_1 I_{\ell+u} + 2\lambda_2 K_{X_{\ell+u}} L)^{-1}, \quad (9)$$

où C est une matrice de dimension $\dim(\mathcal{F}_y) \times (\ell+u)$ regroupant les vecteurs $\mathbf{c}_j$ de (7), $Y_\ell = (\mathbf{y}_1^T, \cdots, \mathbf{y}_\ell^T)$ est une matrice de dimension $dim(\mathcal{F}_y) \times \ell$. U est une matrice de dimension $\ell \times (\ell+u)$ contenant une matrice identité de taille $\ell \times \ell$ sur la partie gauche et une matrice nulle $\ell \times u$ sur la partie droite. $K_{X_{\ell+u}}$ est la matrice de Gram de taille $(\ell+u) \times (\ell+u)$ associée au noyau κ_x. Enfin, $I_{\ell+u}$ est une matrice identité de taille $(\ell+u)$.

L'équation (9) conduit donc au modèle h :

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

où $X_{\ell+u} = (x(o_1)^T, \cdots, x(o_{\ell+u})^T)$ est une matrice identité de dimension $dim(\mathcal{F}_x) \times (\ell+u)$. Ainsi, le calcul de cette solution requiert principalement l'inversion d'une matrice de taille $(\ell+u) \times (\ell+u)$.

4. La preuve complète est disponible sur `http://amis-group.fr/?q=node/379`.

6 Expériences

6.1 Prédiction de liens dans le cadre transductif

Le problème de la régression semi-supervisée à sortie noyau étant résolu, nous revenons maintenant au problème de la prédiction de liens en construisant le classifieur suivant, comme annoncé dans la section 2 :

$$\forall (o, o') \in \mathcal{O} \times \mathcal{O}, \hat{f}_\theta(o, o') = \operatorname{sgn}(\langle \hat{h}(o), \hat{h}(o') \rangle_{\mathcal{F}_y} - \theta) \ . \tag{10}$$

En utilisant le noyau à valeurs matricielles $\mathcal{K}_x$ défini en (5) et l'approximation $\hat{h}$ obtenue en (9), nous avons :

$$\begin{aligned} \forall (o, o') \in \mathcal{O} \times \mathcal{O}, \langle \hat{h}(o), \hat{h}(o') \rangle_{\mathcal{F}_y} &= \left\langle C X_{\ell+u}^T x(o), C X_{\ell+u}^T x(o') \right\rangle_{\mathcal{F}_y} \\ &= x(o)^T X_{\ell+u} B^T K_{Y_\ell} B X_{\ell+u}^T x(o') \ , \end{aligned}$$

avec $B = U(K_{X_{\ell+u}} U^T U + \lambda_1 I_{\ell+u} + 2\lambda_2 K_{X_{\ell+u}} L)^{-1}$, et où $X_{\ell+u}$, $K_{X_{\ell+u}}$, U et $I_{\ell+u}$ sont définis dans la proposition 5.
La fonction noyau en entrée κ_x étant supposée connue, ainsi que les valeurs de κ_y sur $\mathcal{O}_\ell \times \mathcal{O}_\ell$ (c'est à dire, K_{Y_ℓ}), nous pouvons apprendre le classifieur et effectuer des prédictions pour de nouvelles données. En faisant varier le seuil θ dans (10), des courbes ROC et Précision-Rappel peuvent être construites.

Dans les expériences, nous avons évalué la méthode de Régression Régularisée à Sortie Noyau (RRSN) dans un cadre transductif : nous avons ainsi supposé que tous les noeuds étaient connus au début de la phase d'apprentissage et que seul un graphe défini pour un sous-ensemble de noeuds était donné [5].

6.2 Protocole expérimental

Nous avons réalisé des expériences sur un ensemble de jeux de données artificielles, un réseau de co-publications et un réseau d'interactions protéine-protéine (IPP). Pour différentes valeurs de ℓ, le nombre de noeuds étiquetés, nous avons tiré aléatoirement 10 fois un sous-ensemble d'exemples d'apprentissage et utilisé les exemples restants comme exemples de test. Les interactions étiquetées correspondent aux interactions entre deux noeuds de l'ensemble d'apprentissage et l'objectif est de compléter la matrice d'interactions.

5. Remarque : une comparaison avec le cadre de la *link propagation* proposé par (Kashima *et al.* (2009)) ne serait pas appropriée du fait que les interactions étiquetées peuvent être considérées de façon arbitraire, alors que le cadre de la Régression Régularisée à Sortie Noyau (RRSN) nécessite de connaître un sous-graphe d'interactions connues.

Ainsi une proportion de noeuds étiquetés de 10% correspond en fait à une proportion d'interactions étiquetées de 1% seulement. Les performances ont été évaluées en calculant les aires sous les courbes ROC et Précision-Rappel (notées respectivement Auc-roc et Auc-pr) moyennées pour 10 choix aléatoires de l'ensemble d'apprentissage.

Le noyau d'entrée $K_{X_{\ell+u}}$ a été construit en utilisant un noyau gaussien, pour lequel l'hyperparamètre σ a été choisi de sorte à maximiser un critère d'information ($\sigma = 5.7$). Le noyau de sortie est un noyau de diffusion de paramètre β. Un autre noyau de diffusion de paramètre β_2 est également utilisé dans la contrainte de continuité au lieu du Laplacien du graphe : $\exp(-\beta_2 L) = \sum_{i=0}^{\infty} \frac{(-\beta_2 L)^i}{i!}$. Nous avons fixé $W = K_{X_{\ell+u}}$. Les hyperparamètres λ_1, λ_2, β, et β_2 ont été sélectionnés par une procédure de validation croisée en 5 parties sur l'ensemble d'apprentissage.

6.3 Réseaux synthétiques

Nous avons illustré notre méthode sur des réseaux synthétiques, afin de mesurer l'amélioration apportée par la méthode semi-supervisée pour un faible pourcentage de noeuds étiquetés, et cela lorsque le noyau d'entrée est une très bonne approximation du noyau de sortie. Pour produire ces données, nous avons échantillonné des graphes aléatoires à partir d'une loi de Erdős-Renyi. Ces graphes contiennent 700 noeuds et leurs densités [6] ont été respectivement fixées à 0.01, 0.02 and 0.03. Les vecteurs caractéristiques en entrée ont été obtenus en appliquant l'ACP à noyau sur le noyau de diffusion associé au graphe, dont le paramètre de diffusion a été choisi de sorte à maximiser un critère d'information. Enfin, nous avons utilisé les composantes permettant de capturer 95% de la variance pour définir les vecteurs caractéristiques en entrée.

La figure 1 reporte les AUC moyennes et les écarts-types obtenus pour différentes densités de réseaux. On peut observer que l'approche semi-supervisée améliore les performances par rapport à l'approche supervisée sur les deux types d'AUC, et plus particulièrement pour un faible pourcentage de données étiquetées (jusqu'à 10%). A partir de ces résultats, on peut formuler l'hypothèse que la prédiction de lien supervisée est plus difficile dans le cas de réseaux plus denses et que la contribution des données non étiquetées semble être plus utile dans ce cas. On peut également faire la supposition que l'uti-

6. La densité d'un graphe correspond à la probabilité de la présence d'arcs dans le graphe.

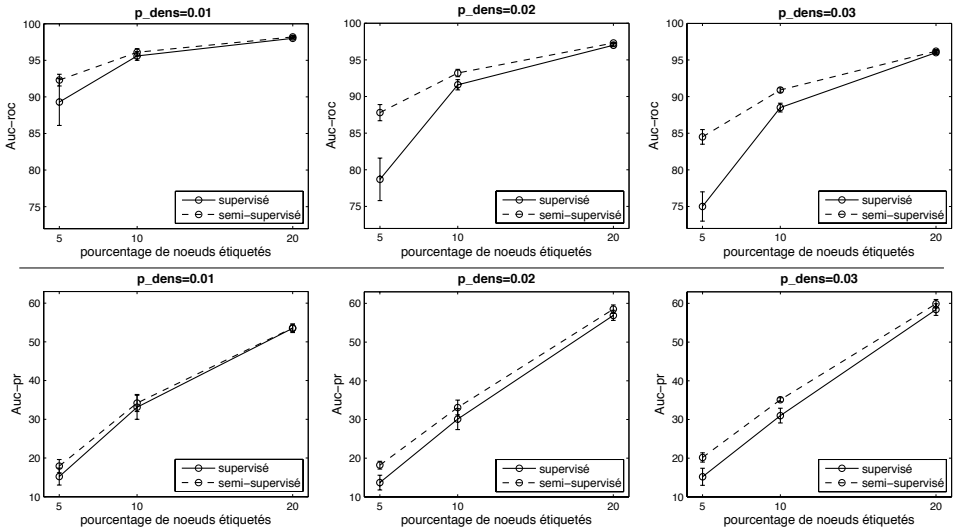

FIGURE 1: Valeurs moyennes et écart-types des Auc-roc (ligne du haut) et des Auc-pr (ligne du bas) pour la reconstruction de trois réseaux synthétiques, étant donné un pourcentage de noeuds étiquetés de 5%, 10% et 20%. Chaque colonne correspond à différentes densités de réseaux (notée p_dens) : 0.01, 0.02 et 0.03.

lisation de données non étiquetées améliore les AUCs pour de faibles pourcentages de données étiquetées, mais que lorsque les données étiquetées présentent suffisamment d'information, l'apprentissage semi-supervisé ne permet plus d'améliorer les performances.

6.4 Réseau de co-publications NIPS

Nous avons appliqué notre méthode à un jeu de données (Globerson *et al.*, 2007) contenant des informations sur les publications des conférences NIPS de 1988 à 2003. Les noeuds du réseau d'interactions représentent ici des auteurs de publications et un arc relie deux auteurs entre eux s'ils ont co-écrit au moins une publication. Parmi les 2865 auteurs, seuls ceux ayant au moins deux liens dans le réseau ont été pris en compte et nous avons ainsi considéré un réseau contenant 2026 auteurs avec une densité de lien empirique de 0.002. Chaque auteur est décrit par un vecteur de 14036 valeurs, qui correspondent aux fréquences avec lesquelles il a utilisé dans ses publications chaque mot

du corpus considéré.

Méthodes	**Auc-roc**			**Auc-pr**		
	$\mathbf{5\%}$	$\mathbf{10\%}$	$\mathbf{20\%}$	$\mathbf{5\%}$	$\mathbf{10\%}$	$\mathbf{20\%}$
Supervisée	71.1 ± 1.1	77.5 ± 0.7	82.2 ± 0.8	7.7 ± 1.8	15.9 ± 1.2	29.3 ± 2.7
Semi-supervisée	75.9 ± 1.3	80.7 ± 0.8	84.7 ± 0.3	8.3 ± 1.3	17.2 ± 1.2	29.0 ± 0.9

TABLE 1: Reconstruction du réseau de co-publications NIPS pour les auteurs ayant au moins deux liens dans le réseau. Les valeurs de pourcentage correspondent aux proportions d'auteurs étiquetés. Les Auc-roc et Auc-pr moyennes sont reportées pour la méthode RRSN, ainsi que les écarts-types, dans les cadres supervisé et transductif.

Les résultats moyens des AUCs dans les cadres supervisé et transductif sont montrés dans la Table 1. Comme précédemment, on peut observer que la méthode semi-supervisée améliore les performances par rapport à la méthode supervisée. Pour un pourcentage d'auteurs étiquetés de 5 %, cette amélioration atteint 4.8 en Auc-roc et 0.6 en Auc-pr.

6.5 Réseau d'interactions protéine-protéine

Nous avons également réalisé des expériences sur un réseau d'interactions protéine-protéine chez la levure *S. Cerevisiae* composé de 984 protéines reliées entre elles par 2438 interactions. Plusieurs descripteurs en entrée ont déjà été utilisés dans le cadre de ce réseau : expressions de gène, profils phylogénétiques, localisation et interactions protéiques dérivées d'expériences double hybride chez la levure (Yamanishi *et al.*, 2004; Kato *et al.*, 2005; Geurts *et al.*, 2006; Bleakley *et al.*, 2007). Dans les expériences, nous avons utilisé les expressions de gènes comme descripteur en entrée, car ces données apparaissent comme la source d'information la plus importante pour cette tâche.

Cadre supervisé.

Plusieurs méthodes d'inférence de réseaux supervisées ont utilisé ce réseau biologique comme benchmark. Nous avons complété la comparaison réalisée par Bleakley *et al.* (2007) avec notre méthode ($RRSN$) et les arbres à sortie noyau avec la méthode des extra-trees ($OK3 + ET$) (Geurts *et al.*, 2007a). Le protocole décrit dans (Bleakley *et al.*, 2007) a été utilisé : chaque méthode est évaluée par une procédure de validation croisée en 5 parties et les hyper-paramètres sont sélectionnés sur les ensembles d'apprentissage. Les Auc-roc

et Auc-pr ont été calculées uniquement pour les interactions possibles entre des protéines de l'ensemble de test et des protéines de l'ensemble d'apprentissage.

Méthodes	**Auc-Roc**	**Auc-Pr**
em	80.6 ± 1.1	6.3 ± 1.2
Pkernel	83.8 ± 1.4	7.6 ± 1.0
locale	78.1 ± 1.1	2.6 ± 0.4
OK3+ET	$\mathbf{84.6 \pm 1.4}$	11.2 ± 3.3
RRSN	83.3 ± 2.1	$\mathbf{13.7 \pm 4.4}$

TABLE 2: Valeurs des Auc-roc et Auc-pr estimées par 5-cv pour la reconstruction du réseau d'IPP de la levure à partir de données d'expressions de gènes dans le cadre supervisé. Les trois premières lignes proviennent de (Bleakley *et al.*, 2007) : *em* désigne la méthode de projection em, *Pkernel* la méthode SVM avec un noyau tensoriel défini sur des paires d'objets et *locale* la méthode SVM avec un modèle local. Les résultats obtenus pour $OK3 + ET$ (Geurts *et al.*, 2007a) et RRSN sont également donnés.

La Table 2 reporte les résultats de (Bleakley *et al.*, 2007) présentant les meilleures Auc-roc et Auc-pr, ainsi que les résultats pour les méthodes OK3+ET et RRSN. En termes d'Auc-roc, la méthode RRSN se comporte aussi bien que les méthodes OK3+ET et Pkernel, avec un léger avantage pour OK3+ET. Concernant l'Auc-pr, la méthode RRSN réalise des performances plutôt bonnes comparé aux autres méthodes.

Cadre transductif.

Nous présentons maintenant les expériences réalisées sur la méthode RRSN dans le cadre transductif en utilisant le protocole expérimental décrit dans 6.2.

Méthodes	**Auc-roc**			**Auc-pr**		
	5%	**10%**	**20%**	**5%**	**10%**	**20%**
Supervisée	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-supervisée	79.6 ± 0.9	80.7 ± 1.0	81.9 ± 0.7	6.6 ± 1.1	7.6 ± 0.8	8.4 ± 0.5

TABLE 3: Reconstruction du réseau d'IPP à partir de données d'expressions de gènes. Les valeurs de pourcentage correspondent aux proportions de protéines étiquetées. Les Auc-roc et Auc-pr sont reportées pour la méthode RRSN dans les cadres supervisé et transductif

Les valeurs moyennes et les écarts-types des Auc-roc et Auc-pr obtenues sont résumées dans la Table 3. Il est intéressant de noter que les problèmes d'inférence de réseaux d'IPP sont caractérisés par un petit nombre de protéines étiquetées, et on peut observer que la méthode semi-supervisée permet une légère amélioration dans ce cas.

7 Conclusion

Nous avons présenté une nouvelle méthode pour la prédiction de liens dans le cadre de l'apprentissage semi-supervisé ou transductif qui est fondée sur l'apprentissage d'une fonction d'une seule variable à valeurs vectorielles.

Afin de tirer partie d'informations non étiquetées, nous avons utilisé un terme de régularisation connu pour être compétitif dans la régression semi-supervisée (Zhou *et al.*, 2004; Belkin *et al.*, 2006). À partir de la théorie des EHNA avec des noyaux à valeurs matricielles introduite par (Senkene & Tempel'man, 1973), nous avons établi et prouvé un théorème de représentation dédié à l'apprentissage semi-supervisé, pour un coût quadratique pénalisé. Puis, nous avons défini un noyau à valeurs matricielles qui permet l'application du théorème de représentation et conduit finalement à une solution analytique qui étend la reformulation de la méthode KDE proposée par Cortes *et al.* (2005) dans le cadre de la régression supervisée.

Nous avons étudié le comportement des modèles résultants dans le cadre transductif, sur des données artificielles et deux jeux de données réelles : la complétion d'un réseau d'interactions protéine-protéine et d'un réseau de co-publications. Ces expériences ont montré qu'utiliser des données non étiquetées permettait d'améliorer les performances lorsque seul un faible pourcentage de noeuds est étiqueté.

Nous envisageons dans la suite d'étudier la régression à sortie noyau pour d'autres choix de noyaux à valeurs matricielles ainsi que d'autres fonctions de coût. Notons que l'application de la théorie des EHNA proposée ici est assez différente des applications existantes telles que l'apprentissage multi-tâche (Argyriou *et al.*, 2009) et la régression fonctionnelle (Kadri *et al.*, 2010). Cette théorie ouvre la voie à l'exploitation de données complexes et structurées en sortie.

Références

ARGYRIOU A., MICCHELLI C. A. & PONTIL M. (2009). When is there a representer theorem ? Vector vs matrix regularizers. *J. Mach. Learn. Res.*, **10**.

BELKIN M. & NIYOGI P. (2004). Semi-supervised learning on riemannian manifolds. *Machine Learning*, **56**(1-3).

BELKIN M., NIYOGI P. & SINDHWANI V. (2006). Manifold regularization : A geometric framework for learning from labeled and unlabeled examples. *J. Mach. Learn. Res.*, **7**.

BEN-HUR A. & NOBLE W. S. (2005). Kernel methods for predicting protein–protein interactions. *Bioinformatics*, **21**(1), 38–46.

BLEAKLEY K., BIAU G. & VERT J.-P. (2007). Supervised reconstruction of biological networks with local models. *Bioinformatics*, **23**.

CAPONNETTO A., MICCHELLI C. A., PONTIL M. & YING Y. (2008). Universal multitask kernels. *J. Mach. Learn. Res.*, **9**.

CORTES C., MOHRI M. & WESTON J. (2005). A general regression technique for learning transductions. In *Proc. of the 22nd Intl. Conf. on Machine Learning*.

GEURTS P., TOULEIMAT N., DUTREIX M. & D'ALCHÉ-BUC F. (2007a). Inferring biological networks with output kernel trees. *BMC Bioinformatics (PMSB06 special issue)*, **8**(Suppl 2), S4.

GEURTS P., WEHENKEL L. & D'ALCHÉ-BUC F. (2007b). Gradient boosting for kernelized output spaces. In *Proc. of the 24th Intl. Conf. on Machine Learning*.

GEURTS P., WEHENKEL L. & D'ALCHÉ BUC F. (2006). Kernelizing the output of tree-based methods. In *Proc. of the 23th Intl. Conf. on Machine learning*.

GLOBERSON A., CHECHIK G., PEREIRA F. & TISHBY N. (2007). Euclidean embedding of co-occurrence data. *J. Mach. Learn. Res.*, **8**.

HUYNEN M. A., VON MERING C. & BORK P. (2003). Function prediction and protein networks. *Current Opinion in Cell Biology*, **15**(2).

KADRI H., DUFLOS E., PREUX P., CANU S. & DAVY M. (2010). Nonlinear functional regression : a functional rkhs approach. In *JMLR Proc. of Intl. Conf. on Artificial Intelligence and Statistics*, volume 9.

KASHIMA H., KATO T., YAMANISHI Y., SUGIYAMA M. & TSUDA K. (2009). Link propagation : A fast semi-supervised learning algorithm for link prediction. In *Proc. of the 9th SIAM Intl. Conf. on Data Mining*.

KATO T., TSUDA K. & ASAI K. (2005). Selective integration of multiple biological data for supervised network inference. *Bioinformatics*, **21**.

KONDOR R. I. & LAFFERTY J. D. (2002). Diffusion kernels on graphs and other discrete input spaces. In *Proc. of the 19th Intl. Conf. on Machine Learning*.

LIBEN-NOWELL D. & KLEINBERG J. (2007). The link-prediction problem for social networks. *J. Am. Soc. Inf. Sci. Technol.*, **58**(7).

MICCHELLI C. A. & PONTIL M. A. (2005). On learning vector-valued functions. *Neural Computation*, **17**.

MILLER K., GRIFFITHS T. & JORDAN M. (2009). Nonparametric latent feature models for link prediction. In *Advances in Neural Information Processing Systems 22*.

SENKENE E. & TEMPEL'MAN A. (1973). Hilbert spaces of operator-valued functions. *Lithuanian Mathematical Journal*, **13**(4).

TASKAR B., WONG M., ABBEEL P. & KOLLER D. (2003). Link prediction in relational data. In *Advances in Neural Information Processing Systems 15*.

YAMANISHI Y., VERT J.-P. & KANEHISA M. (2004). Protein network inference from multiple genomic data : a supervised approach. *Bioinformatics*, **20**.

ZHOU D., BOUSQUET O., LAL T. N., WESTON J. & SCHÖLKOPF B. (2004). Learning with local and global consistency. In *Advances in Neural Information Processing Systems 16*.