Parcimonie et corrélation pour la sélection de noyaux

Abir Zribi Besbes[1], Maxime Berar[1], Alain Rakotomamonjy[1],

[1]Laboratoire d'Informatique de Traitement d'Information et des Systèmes (LITIS)
INSA/Université de Rouen, Avenue de l'Université - BP 8, 76801 Saint-Étienne-du-Rouvray Cedex, France
abir.zribi@insa-rouen.fr, maxime.berar@univ-rouen.fr,
alain.rakoto@insa-rouen.fr

Résumé − Dans ce papier nous étudions l'impact de termes régularisants basés sur la parcimonie et la corrélation sur les performances d'un problème d'apprentissage des noyaux multiples. L'apprentissage par noyaux multiples (MKL) est une approche d'apprentissage qui vise à déterminer la combinaison linéaire la plus efficace des fonctions noyaux à partir des données pour les méthodes d'apprentissage supervisé comme par exemple les séparateurs à vaste marge (SVM). Nous nous somme basé sur l'extension d'une version récente du MKL-SVM optimisant le poids des noyaux en deux étapes. La première étape consiste à reformuler et à résoudre un problème d'optimisation autour des poids associés à chaque noyaux. Durant la deuxième étape, le noyau appris est utilisé pour déterminer les paramètres optimaux du classifieur associé. Notre apport consiste à proposer trois termes régularisants associés à la résolution du problème d'apprentissage des poids d'une combinaison linéaire de noyaux, problème reformulé en un problème de classification vaste marge dans l'espace des couples. Le premier terme régularisant assure une sélection parcimonieuse à l'issue de la résolution du problème d'apprentissage, alors que les deuxième et troisième termes régularisants prennent en compte la similarité entre les noyaux via une métrique basée sur la corrélation. Suite à l'évaluation des effets des termes régularisants proposés sur trois bases de données bio-informatiques décrivant un problème de localisation de protéines bactériennes, nous avons obtenu des résultats de même ordre de grandeur que la méthode de référence en utilisant moins de fonctions noyaux. Les résultats obtenus prouvent la performance promise de la technique proposée comparée à d'autre méthodes compétitives.

Abstract − In this paper, we propose and discuss new regularization terms used to improve the performance of a Multiple Kernels Learning (MKL) problem. The multiple kernel learning (MKL) problem is a learning approach that ensure to find automaticaly the positive combination of kernels which performs the best in regards to the learning problem. In this paper we extend a recent version of MKL where the problem of learning is decomposed and solved in two steps. The first step consists in solving an optimization of the kernels weights problem defined as a large margin problem on the couple of the learning set. In the second step, the learned kernel is used to obtain the optimal parameters of the classifier (for example a Support Vector Machine). Our contribution is to study the influence of three regularization terms during the optimization of the kernels weights problem. In order to improve the efficiency of MKL with fewer kernels, we seek a balance between performance, sparsity and training time complexity. The first regularization term ensures that the kernel selection is sparse. While the second and the third terms introduce the concept of kernels similarity by using a correlation measure. Experiments on multiple bioinformatic data sets show a promising performance of our method compared to states of art methods.

1 Introduction

Ces dernières années, les méthodes à noyaux se sont montrées efficaces pour la résolution d'un grand nombre de problèmes d'apprentissage [1, 2]. Cependant la performance de ce type de méthode est fortement dépendante du choix du noyau par l'utilisateur. Plusieurs travaux ont donc été réalisés dans le but de rendre automatique ce choix en utilisant une combinaison de plusieurs noyaux. L'apprentissage de noyaux multiples (*Multi-Kernel Learning*) est une approche d'apprentissage qui vise à déterminer la combinaison linéaire la plus efficace des fonctions noyaux au regard du problème d'apprentissage à résoudre. Dans ce cadre, Lanckriet et al [3] ont proposé d'optimiser conjoitement les poids des noyaux et les paramètres du Séparateur à Vaste Marge (SVM) dans un seul problème d'optimisation. Une autre version de l'apprentissage multi-noyaux consiste plutôt à chercher les poids des noyaux et les paramètres du problème d'apprentissage en deux phases successives [4, 5, 6]. La première étape consiste alors à formuler et résoudre un premier problème d'optimisation autour des poids associés à chaque noyau. La somme des noyaux ponderés par les poids déterminés à l'issue de cette étape forme alors un nouveau noyau utilisé tel quel lors de la seconde étape. Le nouveau problème est alors un problème classique d'apprentissage par exemple la détermination des paramètres optimaux du classifieur SVM.

Une des formulations les plus connues du premier problème a été suggérée par Cortes et al [6] et repose sur

l'utilisation d'un critère basé sur l'alignement des noyaux. Il s'agit d'aligner les noyaux utilisés à un noyau idéal. Tout récemment [7], un autre critère a été proposé permettant de définir un problème de classification à large marge à partir des données exprimées dans un nouvel espace de couples. Cette nouvelle méthode présente deux limites principales. Premièrement, l'utilisation de la norme ℓ_2 ne permet pas une sélection parcimonieuse des noyaux. Deuxièmement, cette méthode ne tient pas compte des liens qui peuvent exister entre les noyaux.

Dans ce papier, nous apportons une extension à cette méthode en intégrant la parcimonie dans le problème d'optimisation. Afin de contourner la deuxième limite, nous introduisons une mesure de similarité basée sur la corrélation entre les noyaux dans le problème d'optimisation. Dans la suite, nous présenterons notre formalisme et les résultats obtenus sur trois bases de données bioinformatiques.

2 Méthode

Soit $\{x_i, y_i\}_{i=1}^{l}$ nos données d'apprentissage pour la tâche de classification, où x_i appartient à l'espace d'entrée $\mathcal{X}$ et y_i appartient à l'ensemble fini discret des labels $\mathcal{Y}$. Étant donné un ensemble p de fonctions noyaux semi-definis positifs $\{k_m\}_{m=1}^{p}$ de $\mathcal{X}\times\mathcal{X}$ dans $\mathbb{R}$, notre but est d'apprendre la bonne combinaison linéaire de ces fonctions afin d'aboutir à une fonction noyau k_μ ainsi définie :

$$\forall x \in \mathcal{X}, \forall x' \in \mathcal{X} \quad k_\mu(x, x') = \sum_{m=1}^{p} \mu_m k_m(x, x'),$$

qui permet de bien classer nos données d'apprentissage dans l'espace initial. Pour cela, nous définisons un problème d'apprentissage binaire à travers un nouvel espace défini par les paires des données d'apprentissage et les noyaux [7]. Les instances de notre nouveau problème sont obtenues en concaténant l'évaluation des fonctions noyaux sur chaque paire de nos données d'apprentissage :

$$\mathbf{z}_{x,x'} = (k_1(x, x'), \cdots, k_p(x, x')).$$

Les labels correspondants sont obtenus en comparant les classes des deux points de chaque paire :

$$t_{y,y'} = 2 \cdot \mathbf{1}\{y = y'\} - 1.$$

En se basant sur le fait qu'une bonne combinaison de fonctions noyaux doit discriminer les exemples de classes différentes des exemples de même classe, notre nouveau problème de classification binaire prend la forme d'un problème contraint à large marge qui permet de séparer les instances correspondant à des couples de labels différents de celles correspondant à des couples de même label. Le problème d'optimisation utilisé est le suivant [7] :

$$\min_{\mu \geqslant 0} \frac{\lambda}{2} R(\mu) + \frac{1}{\binom{2}{n} + n} \sum_{1 \leq i \leq j \leq l} [1 - t_{ij} \mu \cdot \mathbf{z}_{x_i, x_j}]_+, \qquad (1)$$

où la perte charnière (*hinge*) est donnée par :

$$[1 - s]_+ = \max\{0, 1 - s\}$$

et R est un terme régularisant appliqué à μ qui peut prendre différentes formes telles que la norme ℓ_2 [7] :

$$R_2(\mu) = \|\mu\|_2^2.$$

La contrainte $\mu \geqslant 0$ permet d'assurer que le noyau appris soit semi-défini positif. Souvent, on a un nombre élevé de fonctions noyaux à combiner, pour cette raison nous proposons un premier terme régularisant de norme ℓ_1 permettant une sélection parcimonieuse des noyaux :

$$R_1(\mu) = \sum_{m=1}^{p} \mu_m.$$

Cependant, le lien entre les noyaux n'a pas été pris en compte avec ce type de terme régularisant. Récemment un autre type de terme régularisant est apparu [9], sa particularité est qu'il tient compte de la corrélation entre les noyaux exprimée empiriquement ici à partir de matrices de Gramm K et K'

$$cor(k, k') = \frac{tr\left(K^\top K'\right)}{\sqrt{tr\left(K^\top K\right) tr\left(K'^\top K'\right)}}.$$

En introduisant une matrice Q définie par :

$$Q_{ij} = cor(k_i, k_j)$$

dans les termes régularisants R_2 et R_1, on définit R_2^Q par :

$$R_2^Q(\mu) = \mu^\top Q \mu$$

et R_1^Q par :

$$R_1^Q(\mu) = \mu^\top \sqrt{Q}\mathbf{1}.$$

Les termes régularisants R_1 et R_2 sont en réalité deux cas particuliers du terme régularisant $R_2^Q(\mu)$. Ainsi, si les noyaux sont totalement décorrélés, on aura alors :

$$R_2^Q(\mu) = R_2(\mu).$$

Au contraire si les noyaux sont totalement corrélés, tous les élèments de la matrice Q seront égaux à 1 et on aura donc :

$$R_2^Q(\mu) = R_1(\mu).$$

Enfin R_1^Q correspond exactement à R_1 si les noyaux sont totalement décorrélés et en cas de corrélation correspond à chercher la parcimonie dans l'espace de variables latentes décorrélées.

Quelque soit le choix du terme régularisant, l'inconvénient principal de cette approche est l'aspect quadratique des paires d'instances qui transforme une petite base de données en une grande base de données. Comme dans l'article [7] nous avons donc recours à un algorithme de gradient stochastique pour calculer la solution du problème définit par l'équation (1). Un autre inconvénient est le débalancement des classes introduit par le passage par un

espace de couple. Ainsi pour un problème multiclasse, la classe correspondant à des couples de classes différentes dans l'espace d'origine sera sur-représentée, ce qui peut perturber le bon fonctionnement du gradient stochastique.

Une fois le premier problème résolu, le noyau global k_μ obtenu combine les différents noyaux chacun associé à son poids correspondant. Dans le cas des expérimentations réalisées, ce noyau sera utilisé en entrée d'un classifieur SVM dont les paramètres sont obtenus en résolvant le problème d'optimisation classique donné ici sous sa forme duale avec $\alpha \in \mathbb{R}_+^N$ le vecteur des variables duales :

$$\max_{\alpha} \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j k_\mu(x_i, x_j) \qquad (2)$$

$$\text{s.c.} \quad C \geq \alpha_i \geq 0, \ \sum_{i=1}^{N} \alpha_i y_i = 0$$

3 Résultats expérimentaux

Les effets des termes régularisants proposés ont été évalués sur trois bases de données bio-informatiques (Psortpos, Psortneg et Plant). Les bases Psortpos (541 exemples, 4 classes) et Psortneg (1444 exemples, 5 classes) décrivent un problème de localisation de protéines bactériennes [8]. La base Plant (940 exemples, 4 classes) est une base de protéines végétales [9]. Afin de confronter nos résultats à ceux de la littérature, nous avons repris le dispositif expérimental décrit dans [7, 8]. Les même noyaux, les même schémas de validation et les même mesures de performance ont été utilisés. Les performances ont été évaluées pour toutes les classes selon le score F_1 moyen sur les bases Psortpos et Psortneg, et selon le coefficient de corrélation de Matthews (MCC) moyen sur la base Plant, dont voici la définition :

$$\text{MCC} = \frac{(\text{VP} \times \text{VN}) - (\text{FP} \times \text{FN})}{\sqrt{(\text{VP} + \text{FP})(\text{VP} + \text{FN})(\text{VN} + \text{FP})(\text{VN} + \text{FN})}}, \qquad (3)$$

avec VP le taux de vrais positifs, VN le taux de vrais négatifs, FP le taux de faux positifs et FN le taux de faux négatifs. La précision Pr est le nombre d'éléments bien classés appartenant à une classe donnée sur le nombre total d'éléments appartenant à cette même classe. Le rappel Ra est le nombre d'éléments bien classés appartenant à une classe donnée sur le nombre totale d' éléments attribués à cette même classe. Pour une classe donnée, le score F_1 est donné par

$$F_1 = 2 \times \frac{Pr \times Ra}{Pr + Ra}, \qquad (4)$$

Pour les trois bases, les données que nous avons utilisé se présentent sous la forme de 69 matrices de Gramm. Parmi ces matrices, 64 sont calculées à partir d'un noyau de suite de motifs, 3 sont calculées à partir d'un noyau sur les BLAST E-values et 2 sont obtenues en utilisant un noyau d'arbres phylogénétiques [8]. Toutes les matrices noyaux sont centrées et réduites afin de ne pas perturber le choix des noyaux lors de la première étape. En effet, de trop gros désiquilibres des ordres de grandeurs de chaque matrice pourraient tempérer l'influence de certaines.

Afin de déterminer la solution de la première étape du MKL, nous avons fixé un maximum de 20 000 itérations pour l'algorithme de descente de gradient, avec un seuil d'arrêt basé sur la norme du gradient. Le nombre d'époques du gradient stochastique est 100.

Pour compenser le débalancement des classes, nous avons utilisé une méthode de validation avec re-échantillonnage pour optimiser les valeurs de λ (équation 1). Pour sélectionner ce paramètre, un seul découpage aléatoire 80%/20% de l'ensemble d'apprentissage est utilisé avec comme objectif la plus petite perte hinge en validation. Le tableau 1 présente pour chacune des méthodes la précision moyenne (F_1 ou MCC) et le nombre moyen de noyaux sélectionnés ($\mu \neq 0$) sur 10 essais ainsi que les écarts types correspondants. Les résultats obtenus sur les 3 bases montrent des performances comparables avec des différences peu significatives par rapport à la méthode de référence R_2. Les nombres de noyaux sélectionnés donnés dans le tableau 1 montrent que les méthodes aux termes régularisants parcimonieux R_1 et R_1^Q ont permis de sélectionner moins de noyaux. Cette propriété est particulièrement marquée pour la base PsortPos. Nous remarquons que les écarts types associés aux nombre de noyaux sélectionnés sont particulièrement élevés. Cela signifie que les contraintes de parcimonie ne sont pas effectives pour tous les essais. Nous pensons que cela provient d'un nombre insuffisant d'itérations (20 000 itérations) empêchant la bonne convergence de l'algorithme. Le gradient stochastique est connu pour converger lentement.

La méthode R_2^Q, introduisant la notion de similarité entre les noyaux avec la norme ℓ_2, obtient des performances similaires à R_2 avec un nombre inférieur de noyaux sélectionnés. Ceci est probablement dû à l'effet de groupe qu'engendre la corrélation. Cet effet est moins observable si on compare les résultats donnés par R_1 et R_1^Q. Nous pensons que l'intégration d'un nouveau terme régularisant combinant sous une autre forme la parcimonie et la corrélation pourrait mener à de meilleures performances.

	Psortpos	
	F_1 (std)	$\mu \neq 0$ (std)
R_2	87.5 (3.6)	69 (0)
R_1	86.8 (2.5)	34 (22)
R_2^Q	89.0 (3.7)	48 (26)
R_1^Q	86.2 (3.4)	31 (21)
	Psortneg	
	F_1 (std)	$\mu \neq 0$ (std)
R_2	89.4 (2.2)	59 (5)
R_1	88.0 (2.1)	45 (24)
R_2^Q	88.9 (1.9)	58 (10)
R_1^Q	87.6 (1.4)	45 (8)
	Plant	
	MCC (std)	$\mu \neq 0$ (std)
R_2	88.3 (2.4)	68 (1)
R_1	85.6 (2.8)	45(11)
R_2^Q	85.5 (2.7)	50 (13)
R_1^Q	87.4 (3.6)	44 (20)

TABLE 1 – Mesure de la précision moyenne (écart-type) via 10 essais pour les bases de données Psortpos, Psortneg et Plant.

4 Conclusion

Dans ce papier, nous avons étudié l'impact des termes régularisants sur un problème d'apprentissage de noyaux multiples défini comme un problème de classification vaste marge des données exprimées dans l'espace des couples. Notre contribution a consisté à proposer trois termes régularisants.

La formulation du premier terme vise la sélection parcimonieuse des noyaux les plus utiles par le biais d'une norme ℓ_1. Les deuxième et troisième termes introduisent la matrice de corrélation inter-noyau et permettant d'intégrer ainsi une connaissance relative à leur similarité.
Les résultats expérimentaux obtenus confrontés à ceux de la littérature montrent que notre approche permet de réduire le nombre de noyaux utilisés, sans pour autant dégrader les performances quel que soit le terme régularisant choisi.

Une première perspective de notre travail sera d'améliorer la résolution du premier problème d'optimisation en affinant l'implémentation et l'usage de l'algorithme de gradient stochastique. A plus long terme, l'étude d'autres termes plus complexes de régularisation et leur application à diffèrents problèmes de noyaux multiples constitue une piste à explorer .

Références

[1] B. SchÖlkopf and A. Smola. *Learning with Kernels.* MIT Press : Cambridge, MA, 2002.

[2] J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis.* Cambridge Univer- sity Press, 2004.

[3] R.G. Lanckriet, N. Cristianini, L. El Ghaoui, P. Bratlett et M.I. .Jordan. *Learning the kernel matrix with semidefinite programming.* In Proceedings of the 19th International Conference on Machine Learining, 2002.

[4] T. Joachims, N. Cristianini et J. Shawe-taylor. *Composite kernels for hypertext categorisation.* In Proceedings of the 18th International Conference on Machine Learining, 2001.

[5] J. Kandola, J. Shawe-taylor et N. Cristianini. *Optimizing kernel alignment over combinations of kernels.* In Proceedings of the 19th International Conference on Machine Learining, 2002.

[6] C. Cortes, M. Mohri et A. Rostamizadeh. *Learning non-linear combinations of kernels.* In Advances in Neural Information Processing Systems 22, 2010b.

[7] A. Kumar, A. Niculescu, K. Kavukcoglu et H. Daumé III. *A binary classification framework for two-stage multiple kernel learning.* In Proceedings of the 29th International Conference on Machine Learining, 2012 .

[8] A. Zien et C.S. Ong. *Multiclass Multiple Kernel Learning.* In International Conference on Machine Learining, 2007.

[9] C. Hinrichs, V. Singh J. Peng S. C. Johnson. *Q-MKL : Matrix-induced Regularization in Multi-Kernel Learning with Applications to Neuroimaging.* The Neural Information Processing Systems (NIPS), 2012.

[10] O. Emanuelsson, H. Nielsen, S. Brunak2 and G. von Heijne1. *Predicting Subcellular Localization of Proteins Based on their N-terminal Amino Acid Sequence.* Journal of Molecular Biology, 300 :1005-1016, 2000.