Un algorithme d'optimisation à haute dimension pour la fouille de données : méthode et application en onco-pharmacogénomique

Vincent Gardeux[1,2], René Natowicz[3], Rachid Chelouah[1],
Roman Rouzier[4], Antônio Braga Padua[5], Patrick Siarry[2]

[1] L@ris ; EISTI ; Avenue du Parc ; 95011 Cergy, France
[2] LiSSi ; Université Paris-Est, Créteil, France ;
[3] Université Paris-Est ; ESIEE-Paris, Noisy-le-Grand, France ;
[4] Hôpital Tenon, UPRES EA 4053, Université P&M Curie, Paris, France
[5] Departamento de Engenharia Eletrônica, Universidade Federal de Minas Gerais, Brazil
vincent.gardeux@eisti.eu

Mots-clés : *line search, prédiction, sélection de gènes, onco-pharmacogénomique.*

1 Méthode, algorithme et application

Nous proposons un algorithme de sélection de caractéristiques (*feature selection*) à haute dimension et son application en onco-pharmacogénomique pour le cancer du sein. Dans cette application, nous devons sélectionner un ensemble de gènes dont les niveaux d'expression permettent une prédiction efficace de la réponse des patientes à un traitement de chimiothérapie préopératoire. Les données proviennent d'un essai clinique portant sur un total de 133 cas de patientes.

Pour chaque cas nous disposons, d'une part, des niveaux d'expression de 22283 sondes à ADN mesurés sur tissu tumoral à l'aide de puces à ADN Affymetrix U133A et, d'autre part, de la réponse au traitement, la patiente pouvant être répondeuse (PCR : *pathologic complete response*) ou non répondeuse (No-PCR). Afin d'éviter l'écueil du sur-apprentissage, les ensembles de gènes sélectionnés doivent être de faible cardinal par rapport au nombre de cas d'apprentissage. L'algorithme de sélection de caractéristiques que nous proposons est issu d'une heuristique d'optimisation de type *line search* [1], développée pour les problèmes à grande dimensionnalité. Cet algorithme développé pour des espaces continus est ici transposé aux espaces discrets et appliqué à la sélection de sous-ensembles des 22283 sondes à ADN. Deux objectifs concurrents sont à optimiser : la distance inter-classe est à maximiser (distance euclidienne entre les vecteurs des centres de gravité des deux classes, vecteurs relatifs au sous-ensemble de sondes sélectionné), et la taille du sous-ensemble de gènes sélectionné est à minimiser. Ces deux objectifs concurrents sont combinés dans une fonction bi-objectif :

$$F_w(S) = w \times D(S) + (1 - w) \times (1 - |S|) \tag{1}$$

où $w \in [0, 1]$ est le paramètre de la combinaison linéaire convexe des deux objectifs (distance $D(S)$ et cardinal $|S|$). Nous cherchons une solution optimale $S^*(w) = \arg\max_{S \in 2^{\mathcal{P}}} F_w(S)$, pour w fixé, où $\mathcal{P}$ est l'ensemble de toutes les sondes à ADN et $2^{\mathcal{P}}$ l'ensemble des sous-ensembles de sondes.

L'algorithme d'optimisation discret sélectionne les gènes par maximisation locale itérée. Une sonde σ est ajoutée au sous-ensemble de sondes courant S si $F_w(S \cup \{\sigma\}) > F_w(S)$ et, réciproquement, la sonde σ est retirée du sous-ensemble S. Le point de départ de l'itération est l'ensemble $\mathcal{P}$ de toutes les sondes. L'arrêt de l'itération est obtenu par atteinte d'un point fixe $S^*(w)$, ensemble pour lequel la valeur de la fonction bi-objectif ne peut plus être améliorée par ajout ou retrait d'une sonde.

2 Résultats et conclusion

Les niveaux d'expression des sondes à ADN sélectionnés par notre algorithme sont pondérés par des coefficients calculés par une analyse discriminante, conformément à [2]. Le Tableau 1 compare les résultats obtenus avec ceux de [2] et [3]. Les colonnes "LS DLDA 31" et "LS DLDA 11" sont les performances des prédicteurs calculés par l'algorithme que nous proposons, pour des valeurs de paramètre w retournant des sous-ensembles de 31 et 11 sondes. La colonne "t-test DLDA 31" indique les performances du meilleur prédicteur de [2], dont les 31 sondes sélectionnées étaient celles de plus faible p-value à un t-test et dont les niveaux d'expression avaient été pondérés par des coefficients calculés par DLDA. La colonne "BI Majorité 30" indique les performances du prédicteur calculé dans [3]. Ce prédicteur avait été obtenu par sélection des 30 sondes bi-informatives de plus grande valuation et prédiction par vote majoritaire non pondéré. Une validation croisée 3-*fold* sur l'ensemble des 133 données a été effectuée. Elle montre que notre méthode (sélection d'un ensemble de sondes par maximisation de F_w puis analyse discriminante linéaire) est robuste : les performances obtenues ne sont pas la conséquence d'un sur-apprentissage des données.

	LS DLDA 31	LS DLDA 11	t-test DLDA 31	BI Majorité 30
Distance inter-classes	5175,45	**3670,45**	1383,30	3210,28
Précision	0,863	**0,882**	0,765	0,863
Sensibilité	0,846	**0,923**	0,923	0,923
Spécificité	0,868	**0,868**	0,711	0,842
VPP	0,688	**0,706**	0,522	0,667
VPN	0,943	**0,971**	0,964	0,970

TAB. 1 – Performances comparées des différents prédicteurs (VPP et VPN sont les valeurs prédictives positives et négatives)

Conclusion. Notre méthode de sélection de sondes par algorithme *line search* discret a permis de construire des prédicteurs de la réponse à la chimiothérapie préopératoire pour le cancer du sein dont les performances sont aussi bonnes que celles des meilleurs prédicteurs publiés à ce jour pour les mêmes données. Notre résultat principal est que nos obtenons ces performances avec trois fois moins de sondes (et deux fois moins de sondes que [4]). Ce résultat est important pour la robustesse des prédicteurs, condition nécessaire à une possible utilisation en routine clinique. Par ailleurs, tout comme dans [1] pour les espaces continus, nous avons constaté que la méthode d'optimisation discrète convergeait rapidement et que les solutions calculées étaient indépendantes du point de départ et du mode d'itération de la maximisation locale.

Références

[1] Vincent Gardeux et al. A line search based algorithm for solving high-dimensional continuous non-linear optimization problems. *Soft Computing - A Fusion of Foundations, Methodologies and Applications*, 2011.

[2] K. R. Hess et al. Pharmacogenomic predictor of sensitivity to preoperative chemotherapy with paclitaxel and fluorouracil, doxorubicin, and cyclophosphamide in breast cancer. *Journal of Clinical Oncology*, 24 :4236–4244, 2006.

[3] René Natowicz et al. Prediction of the outcome of preoperative chemotherapy in breast cancer using dna probes that provide information on both complete and incomplete responses. *BMC Bioinformatics*, 9, 2008.

[4] René Natowicz et al. Prediction of chemotherapy outcomes by optimal gene subsets selected by dynamic programming. In *Proceedings of the Cancer Bioinformatics Workshop*, Cambridge, UK, 2nd - 4th September 2010.