ABEUS : un algorithme d'optimisation discret appliqué à la sélection de variables sur des jeux de données de transcriptomique.

Vincent Gardeux[1,2], René Natowicz[3], Rachid Chelouah[1], Patrick Siarry[2]

[1] L@ris, EISTI, Avenue du Parc, 95011 Cergy, France
[2] LiSSi (EA 3956), Université Paris-Est, Créteil, France
[3] Université Paris-Est, ESIEE-Paris, Noisy-le-Grand, France
research@gardeux-vincent.eu

Mots-clés : *oncologie, sélection de variables, grande dimension, recherche linéaire.*

1 Introduction

Nous proposons l'algorithme de sélection de variables *Adapted Binary Enhanced Unidirectional Search (ABEUS)*, dérivé de l'algorithme *EUS* d'optimisation de problèmes continus en grandes dimensions [1]. Nous l'appliquons ici à des données d'expression de puces à ADN en oncologie, pour différents types de cancer.
Dans cette application, nous devons sélectionner un ensemble de gènes dont les niveaux d'expression permettent une prédiction efficace de la classe de l'échantillon testé. Les données d'expression de puces à ADN proviennent de 6 jeux à deux classes, librement accessibles sur le web (cf tableau 1). Afin d'éviter l'écueil du sur-apprentissage, les ensembles de gènes sélectionnés

Données	Références	Nb exemples	Nb sondes (n)
Colon	(Alon 99)	62	2000
Lymphome	(Shipp 02)	77	5469
Leucémie	(Golub 99)	72	7129
Prostate	(Singh 02)	102	10509
Cerveau	(Pomeroy 02)	60	7129
Ovaires	(Berchuck 05)	54	22283

TAB. 1 – Description des différents jeux de données du *benchmark*.

doivent être de faible cardinalité par rapport au nombre de cas d'apprentissage. L'algorithme de sélection de variables que nous proposons utilise une heuristique d'optimisation de type *line search*, qui a procuré d'excellents résultats pour les problèmes de grande dimension [2]. Cet algorithme, initialement développé pour des espaces continus, est ici adapté aux espaces discrets et appliqué à la sélection de sous-ensembles de sondes à ADN. Notre méthode est de type *wrapper* : nous optimisons un critère dépendant du modèle de classification, ici l'analyse discriminante linéaire diagonale (DLDA). Deux objectifs sont à optimiser simultanément : maximiser la *précision* du classifieur (ratio de prédictions correctes) et minimiser la taille du sous-ensemble de gènes sélectionnés. Pour éviter tout biais [3], la sélection de variables et le calcul de prédiction se font sur l'ensemble d'apprentissage seulement. De plus, pour simplifier l'optimisation, nous avons intégré le deuxième objectif dans la procédure d'optimisation ABEUS (une solution n'est remplacée que si la précision est meilleure ou si, à précision égale, le nombre de gènes est inférieur). La fonction objectif $F(S) = \text{précision}(S)$ dont l'argument S est un sous-ensemble de sondes, est à optimiser sur l'ensemble des sous-ensembles de sondes à ADN (de taille 2^n).
ABEUS sélectionne les sondes à ADN par parcours successif de chaque dimension, et opère une "oscillation stratégique" jusqu'à obtenir un optimum local, pour lequel la valeur de la fonction objectif ne peut plus être améliorée par ajout ou retrait d'une sonde. Une sonde est ajoutée au sous-ensemble de sondes courant si cet ajout augmente la précision du prédicteur, sinon elle est retirée du sous-ensemble. Le choix de l'ensemble de sondes initial est aléatoire.

2 Résultats et discussion

Le tableau 2 compare les résultats obtenus par validation croisée *leave-one-out* avec ceux de travaux récents sur les mêmes jeux de données (Prec. = précision calculée sur l'ensemble de test, Nb = nombre de sondes sélectionnées sur l'ensemble d'apprentissage). Dans la première partie, nous donnons les résultats obtenus et publiés par d'autres auteurs avec d'autres méthodes, et des protocoles d'apprentissage qui s'avèrent biaisés. Nous comparons ces résultats avec ceux de notre méthode ABEUS (dernière ligne), obtenus en appliquant respectivement les mêmes biais d'apprentissage. Ces résultats sont malheureusement optimistes et non représentatifs de la robustesse de la méthode. Dans la deuxième partie du tableau, nous avons donc repris les différents jeux de données en supprimant les biais d'apprentissage, et nous avons comparé nos résultats à ceux d'autres méthodes avec le même protocole.

Données	Colon		Prostate		Lymphome		Ovaires		Leucémie		Cerveau	
	Préc.	Nb	Préc.	Nb	Préc.	Nb	Préc.	Nb	Préc.	Nb	Préc.	Nb
Protocole biaisé												
(Berchuck 05)	-	-	-	-	-	-	85,20	186	-	-	-	-
(Chuang 08)	-	-	92,16	1294	100,00	1042	-	-	-	-	-	-
(Ghattas 08)												
F-test	87,81	3	96,29	315	-	-	-	-	-	-	-	-
GLMPath	93,60	2	100,00	3	-	-	-	-	-	-	-	-
Forêts Al.	90,38	55	94,46	7	-	-	-	-	-	-	-	-
ABEUS	95,20	8	99,02	7	100,00	6	100,00	9	100,00	5	90,00	10
Protocole non biaisé												
(Ramaswamy02)	-	-	-	-	-	-	-	-	-	-	60,00	21
(Deutsch03)	-	-	-	-	83,33	6	-	-	-	-	-	-
(Pochet04)	82,03	-	91,22	-	-	-	-	-	94,40	-	-	-
(Ghattas08)												
F-test	84,05	15,1	91,18	126,4	-	-	-	-	-	-	-	-
GLMPath	81,91	1,3	94,09	1,6	-	-	-	-	-	-	-	-
Forêts Al.	89,40	49,8	94,10	81	-	-	-	-	-	-	-	-
ABEUS	83,90	7,00	87,30	7,29	90,90	9,31	68,50	8,76	97,20	6,74	70,00	14,03

TAB. 2 – Performances de différentes méthodes sur notre *benchmark* (la précision est en %).

3 Conclusion

Notre méthode de sélection de variables a permis de construire des prédicteurs efficaces pour six problèmes de bi-partition supervisée de données d'expression de puces à ADN en oncologie. Les performances obtenues sont aussi bonnes, voire meilleures, que celles des meilleurs prédicteurs publiés à ce jour pour les mêmes données. Notre principale contribution est que nous obtenons ces performances avec très peu de sondes. Cette caractéristique est importante pour la robustesse de nos prédicteurs, condition nécessaire à une possible utilisation en routine clinique. Par ailleurs, tout comme dans [2], nous avons constaté que la méthode ABEUS convergeait très rapidement, dans la lignée des méthodes de *line search* dont elle est issue.

Références

[1] V. Gardeux, R. Chelouah, P. Siarry, and F. Glover. Unidimensional search for solving continuous high-dimensional optimization problems. In *Proceedings of the 2009 Ninth International Conference on Intelligent Systems Design and Applications*, pages 1096–1101, Pisa, Italy, November 30 - December 2, 2009. IEEE Computer Society.

[2] V. Gardeux, R. Chelouah, P. Siarry, and F. Glover. EM323 : A line search based algorithm for solving high-dimensional continuous non-linear optimization problems. *Soft Computing, A Fusion of Foundations, Methodologies and Applications*, 15(11) :2275–2285, 2011.

[3] R. Simon, M. D. Radmacher, K. Dobbin, and L. M. McShane. Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification. *Journal of the National Cancer Institute*, 95(1) :14–18, 2003.