Sélection d'attributs de puce à ADN par essaim de particules

E-G. Talbi, L. Jourdan, J. Nieto, E. Alba

Dans ce travail, nous hybridons un algorithme à essaim de particules avec l'algorithme de classification SVM (Support Vector Machines SVM) pour permettre la sélection d'attributs sur des données issues de puces à ADN.
Ces données se caractérisent par un grand nombre de gènes à étudier simultanément mais avec peu d'exérimentations.
La sélection d'attributs permet de réduire l'ensemble de gènes à étudier et d'améliorer la classification des instances du jeu de données. La sélection d'attributs peut se modéliser comme un problème d'optimisation combinatoire: soit un ensemble d'attributs $F = \{f_1, ..., f_i, ..., f_n\}$, l'objectif est de trouver $F' \subseteq F$ qui maximise une fonction de score $\Theta : \Gamma \rightarrow G$ telle que

$$F' = argmax_{G \subset \Gamma}\{\Theta(G)\}, \quad (1)$$

où Γ est l'espace de tous les sous-ensembles d'attributs possibles de F et G un sous ensemble de Γ.
Il est classique dans la modélisation de la sélection d'attributs de considérer qu'une solution est un vecteur de bits où un '1' indique que l'attribut fait parti du sous ensemble d'attributs selectionnés et un '0' indique que l'attribut n'en fait pas parti.

Afin d'utiliser les PSO pour réaliser la sélection d'attributs, nous avons donc utilisé un encodage binaire et avons adapté l'algorithme de PSO continu à cet encodage discret.

Une première contribution montre que l'algorithme PSO_{SVM} est capable de trouver des gènes d'intérêt et d'améliorer la classification de manière significative. Une comparaison de l'approche avec différentes méthodes de la littérature a été réalisé sur six jeux de données différents de puces à ADN traitant du cancer (leukemia, breast, colon, ovarian, prostate, and lung) disponibles sur le web et nous a montré les très bonnes performances de la méthode.