Classification du cancer via PSO

Ahmed Bir-jmel[1], Sidi Mohamed Douiri[1], Souad Elbernoussi[1]

[1]Laboratory of Mathematics, Computing and Applications, Faculty of Sciences, University of Mohammed-V, Rabat, Morocco.

{douirisidimohamed@gmail.com}@gmail.com

Mots-clés *: Classification des données de puces à ADN, réduction de la dimensionnalité, métaheuristiques.*

1 Introduction

Nous vivons dans l'âge de l'information, les données s'accumulent et leur stockage n'est plus coûteux. Notamment dans le domaine médical lors d'examen des patients où il intervient la notion des données des biopuces (puces à ADN). Ces données peuvent être utilisées comme support de décision médicale (aide au diagnostic). Les biopuces (puces à ADN) permettent aux chercheurs de mesurer simultanément les niveaux d'expression de plusieurs milliers de gènes. Ces niveaux d'expression sont très importants dans la classification de différents types de tumeurs (reconnaissance tissu sain/tissu cancéreux ou distinction entre différents types de cancers). Pour cette tâche de classification, on dispose d'un faible nombre d'échantillons alors que chaque échantillon est décrit par un très grand nombre de gènes. Le traitement de ces données nécessite donc de réduire le nombre de gènes pour proposer un sous-ensemble de gènes pertinents et de construire un classifieur prédisant le type de tumeur qui caractérise un échantillon cellulaire. La sélection d'un sous ensemble de gènes dans une puce à ADN est un problème NP-difficile qui peut être résolu par les métaheuristiques. Dans ce travail, nous proposons un algorithme de sélection de sous ensemble de gènes pertinents à l'aide de l'algorithme ReliefF, la métaheuristique d'Optimisation par essaim de particules et un classifieur (algorithme de plus proche voisin 1-PPV) pour construire un modèle d'apprentissage robuste.

2 Méthode

Dans la littérature on trouve deux grandes approches pour résoudre le problème de sélection de gènes :
Méthodes Filtres : consistent à évaluer chaque gène, pour lui assigner un score (une mesure) de pertinence, ce score permet un classement des gènes. La sélection de gènes se fait par le choix des gènes les mieux classés (ReliefF comme exemple).
Méthodes Enveloppe : consistent à générer des sous ensembles candidats et de les évaluer grâce à un algorithme de classification, cette évaluation est faite par le calcul d'un score (taux de classification). L'originalité de notre approche provient de la combinaison de ces deux dernières méthodes.

2.1 ReliefF

Cet algorithme de filtrage, introduit sous le nom de Relief puis amélioré et adapté au cas multi-classes sous le nom de ReliefF, il ne se contente pas d'éliminer la redondance mais définit un critère de pertinence. Ce critère mesure la capacité de chaque caractéristique à regrouper les données de même étiquetée et discriminer celles ayant des étiquettes différentes. Le poids d'un gène est d'autant plus grand que les données issues de la même classe ont des valeurs proches et que les données issues de classes différentes sont bien séparées.

2.2 Méthode des k plus proches voisins

L'algorithme des k plus proches voisins (noté k-PPV) est une méthode basée sur la notion de proximité (voisinage) entre exemples, et sur l'idée de raisonner à partir des cas similaires pour prendre une décision. Le principe est le suivant : on note x un nouvel exemple décrit par un vecteur de N gènes . On trouve alors, parmi l'ensemble d'exemples d'apprentissage, les k plus proches voisins de x et on associe à x la classe majoritaire parmi ses k voisins lui ressemblant le plus dans la base d'apprentissage.

4 Approche proposée

Nous proposons une nouvelle approche de sélection de gènes basée sur la combinaison de deux approches de sélection : l'approche Filtre à travers l'algorithme de ReliefF et l'approche Enveloppe « wrapper » à travers un algorithme d'optimisation par essaim de particules binaire (BPSO) couplé à un classifieur. Notre choix est porté sur le classifieur de plus proche voisin 1-PPV pour ses performances pour les données de grande taille.
L'originalité de notre méthode provient de la combinaison de deux approches différentes, pour sélectionner un sous ensemble de gènes de petite taille et qui fournit de bonnes performances en classification.

4 Conclusion

L'approche proposée puise son efficacité dans l'amélioration remarquable du taux de précision de la classification dans tous les jeux de données. En effet « RBPSO » réduit le nombre de gènes comparé au nombre original de gènes des jeux de données, il s'agit d'une sélection qui assure des meilleures précisions.
En se basant sur ces résultats issus des expériences que nous avons réalisé, nous pouvons affirmer que notre approche (RBPSO) de sélection de gènes est très bien fondée. En effet, sur les quatre ensembles de données utilisés, l'algorithme RBPSO est parvenu à améliorer la qualité de classification. De plus, il a donné un taux de précision supérieur ou égale à 94 % pour tous les jeux de données, avec une classification parfaite 100 % pour (Leukimia1, Leukimia2, DLBC) et cela en utilisant moins de 54 gènes.