# Sélection d'attributs par Learning Tabu Search

Lucien Mousin[1,2], Laetitia Jourdan[1,2], Marie-Eléonore Marmion[1,2], Clarisse Dhaenens[1,2]

[1] Université de Lille, CNRS, Centrale Lille, UMR 9189 - CRIStAL, F-59000 Lille, France
[2] INRIA Lille - Nord Europe, Équipe projet DOLPHIN France
{lucien.mousin}@ed.univ-lille1.fr
{laetitia.jourdan, marie-eleonore.marmion, clarisse.dhaenens}@univ-lille1.fr

**Mots-clés** : *Sélection d'attributs, méta-heuristique, apprentissage, optimisation*

## 1 Problématique

Le nombre de données disponibles étant en constante augmentation, les algorithmes d'extraction de connaissances ont de plus en plus de mal à extraire les informations pertinentes. En effet, dans cette masse de données, beaucoup sont non-significatives ou redondantes, ce qui crée du bruit pour les algorithmes d'extraction de connaissances et rend ainsi les modèles incapables de faire de la prédiction sur de nouvelles données. Pour remédier à ce problème, nous utilisons la sélection d'attributs comme une phase préliminaire aux algorithmes d'apprentissage qui va réduire la taille des données afin de permettre de découvrir de meilleurs modèles. Le problème de sélection d'attributs consiste donc à choisir parmi un ensemble d'attributs de très grande taille, un sous-ensemble plus petit d'attributs qui sont significatifs pour le problème étudié. De nombreuses méthodes de résolution pour la sélection d'attributs ont été proposées [2, 6].

Ce problème étant un problème combinatoire, nous avons choisi de le modéliser comme un problème d'optimisation combinatoire. La qualité du sous-ensemble d'attributs choisi est mesurée en appliquant une méthode de classification sur ces données et en calculant le pourcentage de bonne classification des données. Ce procédé d'évaluation d'un sous-ensemble devient rapidement coûteux lorsque la taille de données est grande. De ce fait, si nous voulons avoir une chance d'avoir une bonne solution, nous devons utiliser une méta-heuristique efficace. C'est pourquoi nous avons choisi pour notre problème d'utiliser un Learning Tabu Search (LTS), une recherche taboue avec un mécanisme d'apprentissage pour guider la recherche.

## 2 Learning Tabu Search pour la Sélection d'attributs

Le Learning Tabu Search (LTS) mis au point par D. Schindl et N. Zufferey [7] est construit suivant cette idée : si certaines combinaisons de caractéristiques appartiennent souvent à de bonnes solutions au cours de la recherche, alors ces combinaisons doivent être privilégiées lors de la génération des nouvelles solutions.

Dans une recherche locale, lors de l'exploration du voisinage (ajout/suppression d'une caractéristique dans une solution), l'algorithme va estimer la qualité potentielle de chaque voisin, au lieu de faire une évaluation complète du voisinage comme dans une recherche taboue. Grâce à cette estimation, au lieu d'évaluer l'ensemble du voisinage, on peut n'évaluer que les $q$ meilleurs voisins suivant l'estimation. Ceci permet d'éviter d'évaluer le voisinage d'une solution en entier. De plus, D. Schindl et N. Zufferey ont montré que sélectionner $q$ solutions à évaluer par la qualité de l'estimation était plus performant que choisir $q$ solutions aléatoirement.

Dans notre problème, ce sont les combinaisons d'attributs qui appartiennent souvent à de bonnes solutions que nous souhaitons privilégier. Pour cela nous nous appuyons sur une mémoire mesurant la pertinence des associations d'attributs. En se basant sur cette pertinence, il est alors possible d'estimer la qualité potentielle des voisins et alors de sélectionner les meilleurs à évaluer complètement. La construction de cette mémoire et son utilisation seront détaillées.

## 3 Résultats

Pour tester notre algorithme, nous avons travaillé avec six jeux de données de la littérature. Pour évaluer la qualité du modèle et tester la résistance au sur-apprentissage, nous avons testé la solution de l'algorithme sur de nouvelles données non connues pendant la recherche.

Nous nous sommes ensuite comparés à une recherche taboue sans apprentissage afin de tester l'apport de la qualité d'un tel mécanisme et à un HillClimbing itéré, c'est-à-dire un HillClimbing redémarrant d'une solution aléatoire lorsqu'un optimum local est atteint, afin de se comparer à une méta-heuristique au comportement différent.

| **Données** | **# Att.** | **# Ind.** | **HillClimbing** | **Tabu Search** | **LTS** |
|---|---|---|---|---|---|
| Colon [8] | 2000 | 62 | **96.88** | 82.81 | 92.70 |
| Breast [8] | 24481 | 62 | 50 | 47.12 | **50.96** |
| Schizophrenia [1] | 410 | 56 | 61.94 | 53.33 | **63.33** |
| DNA [3] | 180 | 1400 | 93.61 | 93.48 | **94.74** |
| Madelon [5] | 500 | 2000 | 58.26 | 56.94 | **59.31** |
| Arcene [4] | 10000 | 100 | 71.14 | 71.81 | **74.81** |

TAB. 1 – Pourcentage moyen de bonne classification sur l'ensemble de validation

## 4 Conclusions et perspectives

Les résultats montrent l'intérêt d'utiliser une méthode d'intégration de connaissances dans une méta-heuristique pour améliorer les performances de cette dernière. Par exemple, il est ici aisé de voir le gain obtenu entre un tabu search sans intégration de connaissances, et celui avec intégration de connaissance. Cet apport est d'autant plus vrai lorsque l'évaluation est coûteuse. Les tests statistiques ont confirmé ces résultats.

L'intégration de connaissances dans une méta-heuristique semble prometteuse et il serait intéressant d'adapter ce mécanisme à d'autres méta-heuristiques afin d'appuyer l'intérêt d'une telle méthode. Notamment dans les algorithmes multi-objectifs où de nombreuses évaluations sont nécessaires, et si elles sont coûteuses, le mécanisme d'apprentissage permettra d'aller beaucoup plus vite.

## Références


[1] Vince Calhoun. Mlsp 2014 schizophrenia classification challenge. `https://www.kaggle.com/c/mlsp-2014-mri`, 2014.

[2] V. Sudha George and V. Cyril Raj. Review on feature selection techniques and the impact of SVM for cancer classification using gene expression profile. *CoRR*, abs/1109.1062, 2011.

[3] Cesar Guerra-Salcedo and L. Darrell Whitley. Genetic approach to feature selection for ensemble creation. In *GECCO 1999*, pages 236–243, 1999.

[4] Isabelle Guyon, Steve Gunn, Masoud Nikravesh, and Lofti A Zadeh. *Feature extraction : foundations and applications*, volume 207. Springer, 2008.

[5] Isabelle Guyon, Steve R. Gunn, Asa Ben-Hur, and Gideon Dror. Result analysis of the NIPS 2003 feature selection challenge. In *NIPS 17*, pages 545–552, 2004.

[6] Y. Saeys, I. Inza, and P. Larrañaga. A review of feature selection techniques in bioinformatics. *Bioinformatics*, 23(19) :2507–2517, 2007.

[7] D. Schindl and N. Zufferey. Solution methods for fuel supply of trains. *INFOR*, 51(1) :23–30, 2013.

[8] Zexuan Zhu, Yew-Soon Ong, and Manoranjan Dash. Markov blanket-embedded genetic algorithm for gene selection. *Pattern Recognition*, 40(11) :3236–3248, 2007.