Etude de stabilité de méthodes de sélection de motifs à partir des séquences protéiques

Rabie Saidi*,**, Sabeur Aridhi*,**
Mondher Maddouri***, Engelbert Mephu Nguifo*

*LIMOS – CNRS UMR 6158 Université de Clermont Ferrand 2, France
** URPAH / FSJ – Université de Jendouba, Tunisie
*** URPAH / FSG – Université de Gafsa, Tunisie
{saidi,aridhi,mephu}@isima.fr mondher.maddouri@fst.rnu.tn

Nous évaluons la robustesse des méthodes de sélection de motifs et nous étudions leur stabilité suite à des variations dans les données d'entrée (Pavel et al., 2007). Nous considérons les deux hypothèses suivantes :

Hypothèse 1. Une méthode de sélection de motifs permet une description fiable de données d'entrée si toute variation dans ces données a une incidence sur l'ensemble de motifs générés. C'est-à-dire qu'elle choisit d'éliminer certains motifs et de garder d'autres.

Hypothèse 2. Lors des variations de l'ensemble de motifs générés, les motifs gardés doivent être intéressants.

En se basant sur l'hypothèse 1, on introduit la notion de *sensibilité* (à ne pas confondre avec la métrique de sensitivité dans la classification supervisée). Cette notion reflète la capacité de produire un ensemble de motifs différent, donc une description différente, à chaque fois que l'on apporte une variation sur le jeu de données d'entrée. Ce critère de sensibilité peut être étudié à travers les motifs conservés appelés *motifs stables*. Il est aussi intéressant de vérifier l'hypothèse 2, c'est-à-dire la qualité des motifs stables, à travers l'étude de leur apport dans une tâche d'apprentissage artificiel.

Ci-après nous définissons les termes utilisés dans ce travail. Soient les éléments suivants :

- Un jeu de données D de taille n, décomposé en n sous-ensembles D_1, D_2,.., D_n, par application de la technique de type « leave-one-out »
- Une méthode M de construction de motifs appliquée sur D d'un coté et sur D_1, D_2,.., D_n d'un autre coté et générant respectivement les ensembles de motifs EM pour D et EM_1, EM_2,.., EM_n pour D_1, D_2,.., D_n.
- Une tâche de fouille de donnée T et Mtr une métrique de qualité de T. On note $Mtr^T(E)$ pour désigner la valeur de la métrique obtenue si on effectue T avec l'ensemble de motifs E comme espace de variables.

Définition 1 : Stabilité d'un motif. Un motif x est dit stable si et seulement si son taux d'apparition dans les EM_i, i = 1..n, est supérieur à un seuil τ. Ce taux d'apparition est tout simplement le rapport du nombre de EM_i, i = 1..n, où le motif x apparaît par le nombre n. Formellement : $\frac{Nombre\ de\ EM_i\ tel\ que\ x \in EM_i}{n} \geq \tau \quad , avec\ i = 1..n.$

Définition 2 : Taux de motifs stables. Le taux de motifs stables (TMS) d'une méthode M est égal au rapport de son nombre de motifs stables par le nombre de motifs distincts des EM_i, i = 1..n. Formellement : $TMS = \frac{Nombre\ de\ motifs\ stables}{\left|\bigcup_{i=1}^{n} EM_i\right|}$