

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 sensibilité 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, \dots, D_n , par application de la technique de type « leave-one-out »
- Une méthode M de construction de motifs appliquée sur D d'un côté et sur D_1, D_2, \dots, D_n d'un autre côté et générant respectivement les ensembles de motifs EM pour D et EM_1, EM_2, \dots, EM_n pour D_1, D_2, \dots, 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 :
$$\frac{\text{Nombre de } EM_i \text{ tel que } x \in EM_i}{n} \geq \tau, \text{ avec } i = 1..n.$$

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 :

$$TMS = \frac{\text{Nombre de motifs stables}}{\left| \bigcup_{i=1}^n EM_i \right|}$$

Définition 3 : Sensibilité d'une méthode. Une méthode M_1 est plus sensible qu'une autre méthode M_2 si et seulement si pour les mêmes changements dans le même jeu de données, le taux de motifs stables de M_1 est plus faible que celui de M_2 . La sensibilité S d'une méthode est donc inversement proportionnelle à son taux de motifs stable : $S = 1 - TMS$

Définition 4 : Intérêt d'un ensemble de motifs stables. Un ensemble de motifs stables EMS est considéré intéressant si son utilisation dans une tâche de fouille de données T permet de conserver les valeurs de métriques de qualité de cette tâche par rapport aux valeurs qu'on peut trouver si on utilise l'ensemble de motifs EM généré à partir du jeu de données original D . Cette conservation C peut être mesurée par : $C = 1 - \frac{Mtr^T(EM) - Mtr^T(EMS)}{Mtr^T(EM)}$

Cependant, on ne peut juger de cet intérêt que si la méthode est déjà sensible. L'intérêt I est donc mesuré par le produit de la sensibilité S et la conservation C : $I = S * C$

Quatre échantillons de données, comportant 1327 séquences protéiques, extraits de la base biologique Swiss-Prot, sont utilisés dans notre étude expérimentale. Nous comparons quatre méthodes d'extraction de motifs reportées dans (Saidi et al., 2007), à savoir n-grammes NG, motifs actifs MA, descripteurs discriminants DD et descripteurs discriminants avec matrice de substitution DDSM. Dans nos expérimentations nous utilisons les mêmes paramètres par défaut tels que dans (Saidi et al., 2007).

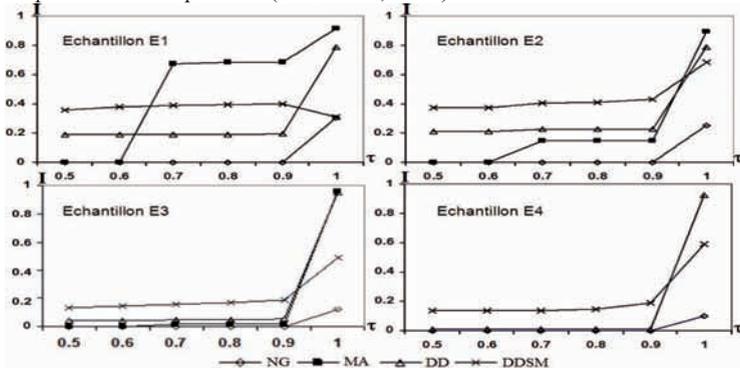


FIG. 1– Courbes représentant l'intérêt de motifs stables I en fonction du seuil de stabilité τ .

Les résultats expérimentaux décrits par la figure 1 montrent l'intérêt de motifs stables I (voir définition 4) en fonction du seuil de stabilité τ pour chaque échantillon de données. Dans notre cas, la métrique utilisée pour le calcul de I est le taux de classification. Généralement, la méthode DDSM présente l'intérêt le plus élevé par rapport aux autres méthodes, un intérêt qui tend à être constant.

Références

- Pavel, K., Josef, K., and Václav, H. (2007). Improving stability of feature selection methods. *CAIP*, Springer Berlin Heidelberg; 929-936.
- Saidi, R., Maddouri, M., and Mephu Nguifo, E. (2007). Biological Sequences Encoding for Supervised Classification (2007). *Intl. conf. on Bioinformatics Research and Development (BIRD)*, TU Berlin, March 12-14, Springer-Verlag, LNBI 4414; 224-238.