Représentation et comparaison de séquences par visualisation

Christine Largeron (*), Cedric Dreissia (**)

Université Jean Monnet de Saint-Etienne
(*) EURISE
23, rue du docteur Paul Michelon
(**) CREUSET
6, rue Basse des Rives
42023 Saint-Etienne Cedex 2
Christine.Largeron@univ-st-etienne.fr

Résumé. Dans cet article, nous présentons un outil de visualisation de séquences modélisées par des arbres de suffixes probabilistes (Prediction suffix trees - PST). Ce type d'arbre permet de représenter une chaîne de Markov d'ordre variable. Dans différentes applications, il s'est avéré plus efficace qu'une chaîne de Markov d'ordre fixe, avec un coût calculatoire moindre. Pour ces raisons, il nous a paru intéressant d'exploiter le caractère arborescent de ce mode de représentation des séquences, non seulement d'un point de vue algorithmique, mais aussi d'un point de vue visuel. Le logiciel que nous avons développé dans ce but fournit une représentation graphique d'un PST appris à partir de séquences et, il permet de le comparer à un autre. Dans un contexte de classement supervisé d'une nouvelle séquence, il apporte une information complémentaire par rapport au PST en mettant en évidence les sous-séquences qui n'ont pas été observées dans la nouvelle séquence bien qu'elles soient caractéristiques du modèle sous-jacent à sa classe d'affectation. Ainsi, il permet de mieux appréhender la structure des séquences et d'améliorer le processus de fouille de données par leur visualisation.

1 Introduction

Les travaux précurseurs en fouille visuelle de données (Visual Data Mining) remontent à Bertin ou encore à Tufte [Bertin, 1977, Tufte, 1983]. Ils portaient sur la représentation graphique de données. Jusqu'à un passé proche, les techniques de visualisation étaient principalement employées dans deux étapes lors du processus de traitement de données :

- au début de la chaîne du traitement, dans une phase exploratoire des données brutes,
- à la fin du traitement, dans une phase de présentation des résultats sous une forme souvent plus synthétique.

Avec l'émergence de la fouille visuelle de données [Card *et al.*, 1999, Spence, 2001, Keim, 2002, Davidson et Soukup, 2002, Poulet, 2004], elles interviennent dans la phase principale du processus de fouille, afin d'impliquer plus directement l'utilisateur