SEQTREE, un outil de fouille de données séquentielles par visualisation

Christine Largeron

Université Jean Monnet de Saint-Etienne
EURISE
23, rue du docteur Paul Michelon
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 non seulement d'un point de vue algorithmique mais aussi d'un point de vue visuel.

1 Introduction

Avec l'émergence de la fouille de données visuelle [Card et al., 1999, Spence, 2001, Keim, 2002, Davidson and Soukup, 2002, Poulet, 2004], les techniques de visualisation permettent de mieux appréhender les données et d'impliquer davantage l'utilisateur dans le processus d'extraction de connaissances. C'est dans cette perspective, que nous avons développé un logiciel de représentation et de comparaison de séquences par visualisation. Par séquence, nous entendons une suite de valeurs observées dans le temps. Il peut s'agir par exemple en climatologie du temps observé quotidiennement dans une région pendant une période donnée, en bioinformatique de séquences d'ADN. Si on suppose que le phénomène étudié présente une dépendance temporelle ; ce qui signifie que la valeur observée à un instant dépend des valeurs observées antérieurement ou du moins de certaines d'entre elles, on peut avoir recours à un modèle de Markov d'ordre variable [Rissanen, 1983]. Un modèle de Markov d'ordre variable peut être représenté par un arbre. Cet arbre, construit à partir de séquences d'apprentissage, peut être utilisé ensuite pour classer de nouvelles séquences ou pour prédire le caractère suivant dans une séquence. Cependant, à notre connaissance, les possibilités offertes par ce mode de représentation arborescent n'ont pas été exploitées dans une perspective de fouille de données visuelle. C'est la raison pour laquelle, nous avons conçu un outil de visualisation et de comparaison de séquences reposant sur ce modèle. Cet outil sera décrit dans la troisième section ; la suivante étant consacrée au modèle de Markov d'ordre variable et à sa représentation sous forme de PST.