Réduction de la complexité spatiale et temporelle du Compact Prediction Tree pour la prédiction de séquences

Ted Gueniche*, Philippe Fournier-Viger*

*Département d'informatique, Université de Moncton
18 Antonine-Maillet, Moncton, NB E1A 3E9
ted.gueniche@gmail.com, philippe-fournier-viger@umoncton.ca

Résumé. La prédiction de séquences de symboles est une tâche ayant de multiples applications. Plusieurs modèles de prédiction ont été proposés tels que DG, All-k-order markov et PPM. Récemment, il a été montré qu'un nouveau modèle nommé Compact Prediction Tree (CPT) utilisant une structure en arbre et un algorithme de prédiction plus complexe, offre des prédictions plus exactes que plusieurs approches de la littérature. Néanmoins, une limite importante de CPT est sa complexité temporelle et spatiale élevée. Dans cet article, nous pallions ce problème en proposant trois stratégies pour réduire la taille et le temps de prédiction de CPT. Les résultats expérimentaux sur 7 jeux de données réels montrent que le modèle résultant nommé CPT+ est jusqu'à 98 fois plus compact et est 4.5 fois plus rapide que CPT, tout en conservant une exactitude très élevée par rapport à All-K-order Markov, DG, Lz78, PPM et TDAG.

1 Introduction

Le problème de prédiction de séquences est un problème important en fouille de données, défini de la façon suivante. Soit un alphabet $Z = \{e_1, e_2, ..., e_m\}$ contenant un ensemble d'éléments (symboles). Une séquence est une suite d'éléments totalement ordonnée $s = \langle i_1, i_2, ...i_n \rangle$, où $i_k \in Z$ $(1 \leq k \leq n)$. Un modèle de prédiction M est un modèle entraîné avec un ensemble de séquences d'entraînement. Une fois entraîné, le modèle peut être utilisé pour effectuer des prédictions. Une prédiction consiste, à prédire le prochain élément i_{n+1} d'une séquence $\langle i_1, i_2, ...i_n \rangle$ en utilisant le modèle M. La prédiction de séquences a des applications importantes dans une multitude de domaines tels que le préchargement de pages Web (Deshpande et Karypis, 2004; Padmanabhan et Mogul, 1996), la recommandation de produits de consommation, la prévision météorologique et la prédiction des tendances du marché boursier.

Un grand nombre de modèles de prédictions ont été proposés pour la prédiction de séquences. Un des modèle les plus connus est PPM (Prediction by Partial Matching)(Cleary et Witten, 1984). Ce modèle, basé sur la propriété de Markov, a engendré une multitude d'approches dérivées telles que Dependancy Graph (DG)(Padmanabhan et Mogul, 1996), All-k-order-Markov (Pitkow et Pirolli, 1999) et Transition Directed Acyclic Graph (TDAG) (Laird et Saul, 1994). Bien que des propositions ont été faites pour réduire la complexité temporelle