Indice de complexité pour le tri et la comparaison de séquences catégorielles

Alexis Gabadinho, Gilbert Ritschard, Matthias Studer
et Nicolas S. Müller

Institut d'études démographiques et des parcours de vie, Université de Genève
{alexis.gabadinho,nicolas.muller,gilbert.ritschard,matthias.studer}@unige.ch
http://mephisto.unige.ch/traminer/

Résumé. Cet article[1] propose un nouvel indice de la complexité de séquences catégorielles. Bien que conçu pour des séquences représentant des trajectoires biographiques telles que celles rencontrées dans les sciences sociales, il s'applique à tous types de listes ordonnées d'états. L'indice prend en compte deux aspects distincts, soit la complexité induite par l'ordonnancement des états successifs qui est mesurée par le nombre de transitions (changements d'état) et la complexité liée à la distribution des états dont rend compte l'entropie.

1 Introduction

Dans les sciences sociales, les séquences catégorielles sont des listes ordonnées d'états ou d'événements décrivant typiquement des trajectoires ou parcours de vie, familiale ou professionnelle par exemple. Dans ce contexte, il importe de pouvoir distinguer les trajectoires selon leur complexité. A cette fin, deux mesures ont été considérées dans la littérature, d'une part l'*entropie* de la distribution des divers états dans une séquence (Gabadinho et al., 2009; Widmer et Ritschard, 2009) et d'autre part la *turbulence* (Elzinga et Liefbroer, 2007). La première mesure présente l'inconvénient de ne pas tenir compte du séquencement des états. La seconde, sensible à cet ordonnancement, reste difficile à interpréter car s'appuyant sur le concept peu intuitif du nombre de sous-séquences distinctes qui peuvent être extraites de la séquence et sur la variance des durées de séjour dans chacun des états. Nous proposons ici un nouvel indice qui combine l'entropie avec un indicateur de la complexité de l'ordonnancement des états. Bien que conçu pour des séquences décrivant des parcours de vie, l'indice caractérise utilement tout type de séquence d'états.

2 Données et définitions

Une séquence de longueur ℓ est une liste ordonnée de ℓ éléments choisis successivement dans un ensemble fini A de taille $|A|$ appelé *alphabet*. Pour illustrer notre propos et comparer le comportement de l'indice proposé nous considérons un jeu de séquences type, 3 jeux de

[1]Travail réalisé avec le soutien financier du Fonds national suisse, subside FN-100015-122230.